A chemometric approach to investigating South African wine behaviour using chemical and sensory markers

by

Mpho Mafata

Dissertation presented for the degree of
Doctor of Philosophy (Agricultural Sciences)

at
Stellenbosch University
Department of viticulture and Oenology, Faculty of AgriSciences

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Supervisor: Dr. Astrid Buica
Co-supervisors: Dr. Jeanne Brand and Prof. Andrei V. Medvedovici

March 2021

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated) that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2021

Summary

The aim of this dissertation was to demonstrate the value of comprehensive narratives and elucidate critical steps in data handling in Oenology, while highlighting some common misconceptions and misinterpretations related to the process. This compilation was a journey through different stages of dealing with oenological data, with increasing complexity in both the strategies and the techniques used (sensory, chemistry, and statistics).

To achieve this aim, different strategies and multivariate tools were used under two prime objectives. Firstly, several multivariate descriptive approaches were used to investigate two oenological problems and lay out the contextual foundations for the statistics-focused work (Chapters 3 and 5). Secondly, in increasing levels of complexity, statistical strategies of constructing comprehensive data fusion as well as pattern recognition models were investigated (Chapters 4 and 6).

A comprehensive literature review (Chapter 2) examined and addressed common misconceptions in the different stages of data handling Oenology.

The first oenological problem, described in Chapter 3, investigated the evolution of the sensory perception of aroma, as well as the antioxidant-related parameters and volatile compound composition of Sauvignon Blanc and Chenin Blanc wines stored under different conditions and durations. The study applied an appropriate sensory method for this research question, namely, Pivot©Profiling. The study was able to show the evolution of Sauvignon Blanc from 'fruity' and 'herbaceous' and of Chenin Blanc from 'fruity' and 'tropical' both towards 'toasted', 'oak', and 'honey' attributes. Chemically, the volatile composition did not show any trends. However, wines stored at higher temperatures for longer periods had relatively higher UV-Vis absorbance, colour density as well as higher b* (yellow) values and lower clarity in terms of L* index, compared to the control.

The second oenological problem, described in Chapter 5, investigated the typicality of South African old vine Chenin Blanc perceptually and conceptually using a typicality rating and a flexible sorting task. The sensory methodology followed published strategies for investigating typicality. This study did not find a unique sensory space of the old vine Chenin Blanc due to a lack of perceptual consensus among the industry professionals for the wines included in the study. However, it did find that the industry professionals had unified ideas about the attributes of an ideal old vine Chenin Blanc wine.

The first of the statistics-focused studies, described in Chapter 4, explored data fusion at low and mid-level using principal component analysis - PCA (low and mid-level) and multiple factor analysis - MFA (mid-level). The study looked at data pre-processing and matrix compatibility, which are important data handling stages for data fusion. Like the contextual chapters (Chapter 3 and 5), and keeping with the aim of this compilation, this chapter gave a detailed descriptive narrative of the data handling. Through detailed examination of the process, the study found that MFA was the most appropriate data fusion strategy. The second statistics-focused study, described in Chapter 6, continued to exploit the multiple advantages of multiblock approach of MFA. Additionally, this chapter showed the reliability of fuzzy k-means clustering compared to agglomerative hierarchical clustering (AHC).

Opsomming

Die doel van hierdie proefskrif was om die waarde van omvattende vertellings te demonstreer en om kritiese stappe in die hantering van data in die wynkunde toe te lig, terwyl enkele algemene wanopvattings en verkeerde interpretasies in verband met die proses uitgelig word. Hierdie samestelling was 'n reis deur verskillende stadiums van die hantering van wynkundige data, met toenemende kompleksiteit in beide die strategieë en die gebruikte tegnieke (sensoriese, chemie en statistieke).

Om hierdie doel te bereik, is verskillende strategieë en meerveranderlike instrumente onder twee hoofdoelstellings gebruik. Eerstens is verskeie multivariate beskrywingsbenaderings gebruik om twee oenologiese probleme te ondersoek en die kontekstuele grondslae vir die statistiekgerigte werk uit te lê (hoofstukke 3 en 5). Tweedens, in toenemende vlakke van kompleksiteit, is statistiese strategieë vir die konstruering van omvattende datafusie sowel as patroonherkenningsmodelle ondersoek (hoofstukke 4 en 6).

'N Omvattende literatuuroorsig (hoofstuk 2) het algemene misverstande in die verskillende stadiums van datahantering van wynkunde ondersoek en behandel.

Die eerste wynprobleem, wat in hoofstuk 3 beskryf word, het die evolusie van die sintuiglike waarneming van aroma ondersoek, asook die antioksidant-verwante parameters en die vlugtige samestelling van Sauvignon Blanc- en Chenin Blanc-wyne wat onder verskillende toestande en duur gestoor is. Die studie het 'n toepaslike sensoriese metode vir hierdie navorsingsvraag toegepas, naamlik Pivot©Profiling. Die studie kon die evolusie van Sauvignon Blanc van 'vrugtige' en 'kruidagtige' en van Chenin Blanc van 'vrugtige' en 'tropiese' sowel as 'geroosterde', 'eikehout' en 'heuning'-eienskappe aantoon. Chemies het die vlugtige samestelling geen neigings getoon nie. Wyne wat vir langer tydperke by hoër temperature gestoor is, het egter relatief hoër UV-Vis-absorbansie, kleurdigtheid sowel as hoër b * (geel) waardes en laer helderheid in terme van L * -indeks, vergeleke met die kontrole.

Die tweede wynprobleem, wat in hoofstuk 5 beskryf word, het die tipiesheid van die Suid-Afrikaanse ou wingerdstok Chenin Blanc perseptueel en konseptueel ondersoek met behulp van 'n tipiese klassifikasie en 'n buigsame sorteertaak. Die sensoriese metodologie het gepubliseerde strategieë vir die ondersoek na tipiesheid gevolg. Hierdie studie het nie 'n unieke sensoriese ruimte vir die ou wingerdstok Chenin Blanc gevind nie, omdat daar 'n gebrek aan konseptuele konsensus tussen die professionele persone vir die wyne wat in die studie opgeneem is, was. Dit het egter gevind dat professionele persone in die bedryf eenvormige idees gehad het oor die eienskappe van 'n ideale ou wynstok Chenin Blanc-wyn.

Die eerste van die statistiekgerigte studies, wat in hoofstuk 4 beskryf word, het datafusie op lae en middelvlak ondersoek met hoofkomponentanalise - PCA (lae en middelvlak) en meervoudige faktorontleding - MFA (middelvlak). Die studie het gekyk na die voorverwerking van data en matriksversoenbaarheid, wat belangrike stadiums vir die hantering van data is vir die versmelting van data. Net soos die kontekstuele hoofstukke (Hoofstuk 3 en 5), en in ooreenstemming met die doel van hierdie samestelling, het hierdie hoofstuk 'n gedetailleerde beskrywende vertelling van die datahantering gegee. Deur middel van 'n uitvoerige ondersoek van die proses, het die studie bevind dat MFA die mees geskikte strategie vir data-fusie was. Die tweede statistiekgerigte studie, wat in hoofstuk 6 beskryf word, het voortgegaan om die veelvuldige voordele van multiblokke benadering van MFA te benut. Verder het hierdie hoofstuk die betroubaarheid van fuzzy k-middelgroepering vergeleke met agglomeratiewe hiërargiese groepering (AHC) getoon.

This dissertation is dedicated to

My loving sister, Boitumelo Mafata. To my mother, Mme Kelapile Florence Mafata (1957 - 2008), and my brother Lebogang John Makati (1977-2014), may their souls rest in peace.

Biographical sketch

Mpho Mafata was born in the town of Mogwase in the North West province on 14 August 1990. She matriculated at JM. Ntsime High school, in the same town, in 2008. In 2009 she enrolled at the University of Cape Town and obtained her B.Sc (Chemistry) degree in 2012 and B.Sc (Chemistry) Honours in 2013. She completed her MSc in Wine Biotechnology in 2017 at the Institute for Wine Biotechnology at the University of Stellenbosch whilst under the employment of the Agricultural Research Council. She then enrolled for her doctoral studies in 2017 under the supervision of Dr. Astrid Buica.

Acknowledgements

I wish to express my sincere gratitude and appreciation to the following persons and institutions:

- **Dr. Astrid Buica** for being my thinking and reasoning partner, my writing as well as my travel partner. Thank you for the mentorship.

- **Prof. Martin Kidd** for the best, most professional, expert statistical analyses.

- **Dr. Jeanne Brand** and **Valeria Panzeri** for expertise in sensory science.

- **Dr. Maritjie Stander of the Central analytical facility (CAF), Mass Spectrometry (MS) Unit,** for expertise in analytical chemistry.

- **Herman Redelinghuys of the Centre for Research on Evaluation, Science and Technology (CREST),** for expertise in machine learning techniques.

- **Dr. Jaco Brand of the Central analytical facility (CAF), Nuclear Magnetic Resonance (NMR) Unit,** for expert experimental designs and analyses.

- **Ms. Elsa Malherber of the Central analytical facility (CAF), Nuclear Magnetic Roesonance (NMR) Unit,** for efficient administration and analyses.

- **The South African Old Vines Project and their members** for their insight and participation in sensory evaluations.

- **Participating cellars** for the supply of Chenin Blanc and Sauvignon Blanc.

- **Participating vineyards** for donating Old Vine Chenin Blanc grapes.

- **Ms Marissa Nell and Ms Freda Aléta Bruwer)** for the winemaking.

- **National research foundation (NRF) of South Africa** for funding the research.

Preface

This dissertation is presented as a compilation of **seven chapters**. The results chapters have been published, or provisionally accepted for publication in several peer-reviewed journals. To facilitate ease of reading for the reader and for the sake of continuity, this compilation is written according to the style of the **South African Journal of Enology and Viticulture**.

Chapter 1 **General Introduction and project aim**

Chapter 2 **Literature review**
Oenological applications of chemometric and sensometric techniques

Chapter 3 **Research results**
A multivariate approach to evaluating the chemical and sensorial evolution of South African Sauvignon Blanc and Chenin Blanc wines under different bottle storage conditions

Chapter 4 **Research results**
Exploration of data fusion strategies using Principal Component Analysis (PCA) and Multiple Factor Analysis (MFA)

Chapter 5 **Research results**
Investigating the concept of South African old vine Chenin Blanc

Chapter 6 **Research results**
Data fusion using Multiple Factor Analysis coupled with non-linear pattern recognition (fuzzy k-means)

Chapter 7 **General discussion and conclusions**

List of outputs

The work presented in this dissertation was submitted for publication to peer review scientific journals, presented at scientific conferences and communicated through publication of popular articles.

Scientific articles

1. Mafata, M., Brand, J., Panzeri, V., and Buica, A., 2020. Investigating the conceptual and perceptual sensory space of South African old vine Chenin Blanc using typicality rating, sorting, and free word association. *South African Journal of Enology and Viticulture,* 41 (2), 168-183. DOI:https://doi.org/10.21548/41-2-4018
2. Mafata, M., Brand, J., Panzeri, V., Kidd, M., and Buica, A., 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. doi:10.1016/j.foodres.2019.108515
3. Mafata, M., Stander, M. A., Thomachot, B., and Buica, A., 2018. Measuring Thiols in Single Cultivar South African Red Wines Using 4,4-Dithiodipyridine (DTDP) Derivatization and Ultraperformance Convergence Chromatography-Tandem Mass Spectrometry. *Foods*, 7(9), 138. doi:10.3390/foods7090138

Conference participation

Workshops

Mafata, M., Brand, J., Panzeri, V. and Buica, A. Elucidating South African Old Vine Chenin Blanc character: Chemically and sensorially. South African Society for Enology and Viticulture conference, (October 2018), Somerset West, South Africa.

Oral presentations

Buica, A., Mafata, M., Brand, J., Panzeri, V., Kidd, M., and H., Redelinghuys. Strategies for data fusion: Sensometric and chemometric perspective, 3rd International Fragrance and Flavors Conference, (October 2019), Vina del Mar, Chile

Posters

1. Panzeri, V., and Brand, J. Toward demonstrating the concept of “old vine character” for South African Chenin Blanc wines. Chenin Blanc international conference, (July 2019), Angers, France.
2. Mafata, M., Brand, J., Panzeri, V., Kidd, M., and Buica, A. A multivariate approach to evaluating the chemical and sensorial evolution of South African Sauvignon and Chenin Blanc wines under different bottle storage conditions. Oeno/In Vino Analytica Scientia (IVAS)2019, (June 2019), Bordeaux, France.

3. Mafata, M., Brand, J., Panzeri, V. and Buica, A. Elucidating South African Old Vine Chenin Blanc character: Chemically and sensorially. South African Society for Enology and Viticulture conference, (October 2018), Somerset West, South Africa.
4. Mafata, M., Brand, J., Panzeri, V. and Buica, A. Investigating thiol stability in commercial Chenin Blanc wines using chemical analysis and sensory evaluation: a comparison with Sauvignon Blanc. Macrowine (May/June 2018), Zaragoza, Spain.

Popular articles

1. Buica, A., Mafata, M., Panzeri, V., and Brand, J. Dec., 2019. Stability of young Chenin blanc and Sauvignon blanc wine during storage (Part 2): Sensory aspects. *Oenology research, Winetech Technical.*
2. Buica, A., Mafata, M., Panzeri, V., and Brand, J. Nov., 2019. Stability of young Chenin blanc and Sauvignon blanc wine during storage (Part 1): Chemical evaluation. *Oenology research, Winetech Technical.*
3. Buica, A., Mafata, M., and Stander, M. Dec, 2018. Thiol profiles of single cultivar red wines. *Oenology research, Winetech Technical*.

Table of Contents

Chapter 1

General introduction and project aims

Chapter 1: General introduction and project aims

1.1 Introduction

Background: Wine has a complex and dynamic chemical composition; this is the reason why there are so many different chemical techniques and sensory methods for evaluating its behaviour (Stevenson, 2005). This very often results in large amounts of data collected. A look at current trends in literature shows a growing interest to advance the informational value of large data sets through the use of advanced statistical modelling tools in sensory (Cariou & Qannari, 2018; Valente *et al.*, 2018; Cariou *et al.*, 2019) and chemistry (Biancolillo *et al.*, 2019). Bioinformatics, metabolomics, chemometrics, and sensometrics are all forms of statistical data handling in their respective fields (McKillup, 2012). There is field specificity and certain sets of rules when it comes to how statistics are applied in each of these fields. Regardless, every field follows the same process when it comes to handling the data: data collection/capturing, cleaning/pre-processing, modelling, and interpreting the data (Salkind. J. & Kristin. R., 2007; McKillup, 2012; Cocchi, 2019a).

Contextualization of key terms and concepts: Data collection, when done intelligently, will be planned through the use of a design of experiments (DOE) (Kreutz & Timmer, 2009; Yu *et al.*, 2018; Ferreira, 2019). Oenological studies collect various chemical and sensory data to answer the research question and address the demands of the DOE. Nowadays, chemical data mostly involves automated collection and capturing, which can be in the form of targeted or untargeted measurements. Sensory data collection is based on the various types of methods, which including ordinal, intensity-based, and frequency of citation data. Capturing of sensory data can be automated or done manually, depending on the availability of the specialised software.

The collected data may require some clean-up and/or data pre-processing, depending on the nature (type) of the data. For instance, targeted chemical data usually only needs simple factor or scale or conversions while untargeted data may require complex mathematical pre-processing such as Fourier transformations (*e.g.* Infrared/IR and Nuclear Magnetic Resonance/NMR) and scaling such as multiple scatter correction for IR (Rinnan *et al.*, 2009; Engel *et al.*, 2013). Sensory pre-processing can be done through manual and statistical consolidation (McKillup, 2012). This step is critical for the handling of sensory data since only what is captured and is consolidated can be modelled. Descriptor consolidation mainly includes lemmatization, linguistic and semantic consolidation (Deneulin & Bavaud, 2016).

Data modelling can be only exploratory (unsupervised) or include elements for prediction, classification or discrimination (supervised) depending on the research question and DOE (Sohail & Arif, 2020). Generally, exploratory techniques are used for hypothesis-forming purposes while supervised techniques are used for hypothesis testing. The commonly used unsupervised multivariate techniques in Oenology include Cluster Analysis (CA), Multidimensional Scaling (MDS), Multifactorial Analysis (MFA), and Principal Component Analysis (PCA). As to when and how to use which technique, it depends on the type of data matrix captured (*e.g.* ordinal, correlation, co-occurrence) and on the research question. Commonly used supervised techniques include variants of Partial Least Squares (PLS *vs.* PLS-Discriminant Analysis or PLS-DA, Orthogonal-PLS or O-PLS, *etc.*) and Linear Discriminant Analysis (LDA) (Seisonen *et al.*, 2016; De Carvalho Rocha *et al.*, 2020).

When it comes to making sense and assessing the significance of the outcomes of these models, it is necessary to apply both statistical and contextual interpretation. In doing so, there are various model performance parameters as well as visualization aids (graphs and illustrations) available. This manner of interpretation can help minimize misinterpretation or confirmation bias. It is at this point that critical thinking must be applied since, when using different approaches different types of information can be extracted from a single data set. This is especially critical when working with multi-way/multi-modal Oenological data that can be considered an information bank from which different currencies can be withdrawn (*i.e.* data of different informational value and scale). Visual aids are used for knowledge compression to aid in interpretation, but caution must be taken when applying them since they can alter perceptions.

Motivation: Researchers need to understand that oenological data is generally multi-way/multi-modal and it needs to be treated as such to solve issues of absolute *vs* relative significance. By highlighting gaps in the communication of the data handling for oenologists and pointing out the critical steps, it can be shown that this process is not a “black-box”. Both theoretical and executional limitations in data handling can be addressed by examining the process and the philosophy rather than simply focusing on the input and output elements. This means creating approaches that emphasize exploration of the problem by aligning multiple perspectives, rather than approaches which focus on perfecting the answer to a single problem.

Problem statement: In Oenology, there are certain misconceptions about data handling due to the lack of articulation of the process. This includes misconceptions about the way in which the data should be handled as well as how the process and the results should be communicated. This proliferates low confidence in handling and interpreting data in a critical manner. The lack of confidence and the misconceptions make it difficult to develop on the repertoire of data handling techniques in Oenology and move towards the age of artificial intelligence. Multi-way problems mean evaluating the relative importance of different data. This means being specifically considerate of which data sets can and should be combined, and how. Applications of data fusion methods, which combine and integrate data sets, appropriately address the issue of relative importance between data sets by separately scaling them according to their variation (Cocchi, 2019b). This critical thinking is important when trying to address questions related to combining sensory with chemistry data (Alañón *et al.*, 2015; Seisonen *et al.*, 2016; Cariou *et al.*, 2019; Bokade *et al.*, 2021). By developing a more "self-aware" approach to data handling process in Oenology, perhaps we can start asking the correct questions of the data and becoming more accustomed to both hypothesis-testing and hypothesis-forming results.

1.2 Aim and objectives

The aims of this dissertation were to demonstrate the value of a comprehensive narrative of the process of data analysis in Oenology and to elucidate critical steps in data handling while highlighting some common misconceptions and misinterpretations. This work is a journey through different stages, with increasing complexity, of dealing with Oenological data.

To achieve the aims, different strategies and multivariate tools were used. Firstly, several multivariate descriptive approaches were used to investigate two oenological problems. Then, in increasing the complexity, strategies of constructing comprehensive data fusion models were investigated. As such, the objectives can be grouped under:

1. **Evolution of wine throughout different storage conditions**
   a) To show that the evolution of wine aroma attributes, volatile and antioxidant compounds can be modelled and compared using unsupervised multivariate analysis on the experimental set-up of South African Chenin Blanc and Sauvignon Blanc wines stored at different temperatures for different periods
   b) To compare the use of PCA (low-level and mid-level) and MFA (mid-level) models for fusing multimodal data. The purpose was building comprehensive and representative data fusion models, while exploring critical troubleshooting.
2. **Investigation of the typicality of Chenin Blanc old vine wines**
   a) To establish the concept and perception of old vine Chenin Blanc among industry experts using typicality rating, sorting, and free word association.
   b) To investigate the potential use of artificial intelligence (AI) strategies for pattern recognition. Classical multivariate statistical tools (by MFA) were used for creating representative data fusion models and classical clustering (agglomerative hierarchical clustering - AHC) and machine learning tools of fuzzy k-means cluster were explored.

References

Alañón, M., Pérez-Coello, M. & Marina, M., 2015. Wine science in the metabolomics era. Trends Anal. Chem. 74 1–20.

Biancolillo, A., Boqué, R., Cocchi, M. & Marini, F., 2019. Data Fusion Strategies in Food Analysis. In: Data Handl. Sci. Technol. Vol. 31. Elsevier Ltd 271–310.

Bokade, R., Navato, A., Ouyang, R., Jin, X., Chou, C.-A., Ostadabbas, S. & Mueller, A. V, 2021. A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. Expert Syst. Appl. 165 113885.

Cariou, V. & Qannari, E.M., 2018. Statistical treatment of free sorting data by means of correspondence and cluster analyses.

Cariou, V., Jouan-Rimbaud Bouveresse, D., Qannari, E.M. & Rutledge, D.N., 2019. ComDim Methods for the Analysis of Multiblock Data in a Data Fusion Perspective. In: Data Handl. Sci. Technol. Vol. 31. Elsevier Ltd 179–204.

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

Cocchi, M., 2019a. Introduction: Ways and Means to Deal With Data From Multiple Sources. In: Data Handl. Sci. Technol. Vol. 31. Elsevier Ltd 1–26.

Cocchi, M., 2019b. Data fusion methodology and applications. Vol. 31.

Deneulin, P. & Bavaud, F., 2016. Analyses of open-ended questions by renormalized associativities and textual networks: A study of perception of minerality in wine. Food Qual. Prefer. 47 34–44.

Engel, J., Gerretzen, J., Szyman´ska, E., Szyman´ska, S., Jansen, J.J., Downey, G., Blanchet, L. & Buydens, M.C., 2013. Breaking with trends in pre-processing? Trends Anal. Chem. 50 96–106.

Ferreira, S.L.C., 2019. Chemometrics and statistics | experimental design. In: Encycl. Anal. Sci. Elsevier 420–424.

Kreutz, C. & Timmer, J., 2009. Systems biology: Experimental design. FEBS J. 276(4), 923–942.

McKillup, S., 2012. Statistics explained : an introductory guide for life scientists. (2nd ed.). Cambridge University Press.

Rinnan, Å., Berg, F. van den & Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28(10), 1201–1222.

Salkind. J. & Kristin. R., 2007. Encyclopidia of Measurement and Statistics. Sage.

Seisonen, S., Vene, K. & Koppel, K., 2016. The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. Food Chem. 210 530–540.

Sohail, A. & Arif, F., 2020. Supervised and unsupervised algorithms for bioinformatics and data science. Prog. Biophys. Mol. Biol. 151 14–22.

Stevenson, T., 2005. The-New-Sothebys-Wine-Encyclopedia. (Fourth ed.). Dorling Kindersley Limited.

Valente, C.C., Bauer, F.F., Venter, F., Watson, B. & Nieuwoudt, H.H., 2018. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. Sci. Rep. 8(1),.

Yu, P., Low, M.Y. & Zhou, W., 2018. Design of experiments and regression modelling in food flavour and sensory analysis: A review. Trends Food Sci. Technol. 71 202–215.

Chapter 2

Literature review

Oenological data analysis: applications of chemometric and sensometric techniques

Chapter 2: Oenological data analysis: applications of chemometric and sensometric techniques

2.1 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, Dijksterhuis, Qannari, *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, 2012). In such cases, multivariate data analysis tools that reduce the dimensionality of large data in order to highlight and visualize the important features that describe the overall relationships are needed (Granato, de Araújo Calado & Jarvis, 2014).

Data fusion (defined as combining and integrating different data sets) is important when working with complex systems such as natural products (Cocchi, 2019; White, 1991). Data fusion systems provide holistic and comprehensive data models (Handling & Science, 2019). 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, Ferré, Boqué, *et al.*, 2015; Handling & Science, 2019).

Oenological evaluations look at a wine's behaviour throughout the winemaking process under different stimuli such as temperature (Mafata, Brand, Panzeri, *et al.*, 2019; Mafata, Buica, du Toit & van Jaarsveld, 2018; Serra-Cayuela, Jourdes, Riu-Aumatell, *et al.*, 2014; Du Toit & Piquet, 2014) and temporal changes (Coetzee, Van Wyngaard, Šuklje, *et al.*, 2016; Pereira, Carvalho, Miranda, *et al.*, 2016; Pereira, Reis, Saraiva, *et al.*, 2011). The field has advanced to use holistic measurements that capture various sensory and chemical responses to the stimuli, 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 profile, but some methods may be redundant in the information they provide. It is thus important to use techniques that are compatible and information-rich (Borràs *et al.*, 2015; Cocchi, 2019). 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), modelling (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 modelling. The modelling is dependent on the research question and the type of data acquired. The output refers to tables of calculations of model parameters and related figures of merit. 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 analyse 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-centred 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 Oenology, a descriptive bibliometric search was performed including documents published in the past decade (2010 - 2020). The analysis used two credible academic databases, namely: Scopus (citation) and Commonwealth Agricultural Bureaux Index or Centre 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 oenological application often refer to it as ‘*omics*’ (‘*metabolomics*’ or ‘*wineomics*’) (Alañón, Pérez-Coello & Marina, 2015; Moyano, Serratosa, Marquez, *et al.*, 2018). For sensory investigations in Oenology, most publications use the term ‘*multivariate analysis*’; only some of the sensory investigations refer to the data handling as ‘*sensometrics*’ (Brand, 2019; Cariou, Qannari, Rutledge, *et al.*, 2018; Guld, Nyitrainé Sárdy, Gere, *et al.*, 2020). 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 Oenology can focus either on a specific application or on methodology development. In most reported cases, both use statistical analysis for hypothesis testing (Granato *et al.*, 2014; Granato & Ares, 2013). 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 Oenology. 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 Oenology. 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.

2.2 Evaluation in Oenology and rationale behind the movement towards multivariate statistical analysis (MVA)

2.2.1 Categories of Chemistry and Sensory methods

The major modes of evaluation in Oenology are chemical and sensorial. Chemical methods can be broadly categorised under targeted and untargeted (Alañón *et al.*, 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, Fang, Humpfer, *et al.*, 2013). Chemical data used in Oenology can be further sub-categorized into volatile and non-volatile 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 (Borràs *et al.*, 2015; Lapalus, Wessel & Du Toit, 2016) 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 methods are widely used for authentication applications (Alañón *et al.*, 2015; Borràs *et al.*, 2015; Ríos-Reina, Callejón, Savorani, *et al.*, 2019). Untargeted techniques use supervised data models for prediction or classification of samples (Versari, Laurie, Ricci, *et al.*, 2014) and few have attempted to use untargeted techniques to predict sensory data (Niimi, Tomic, Næs, *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, Chollet, Lelièvre, *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, Ballester, Langlois, *et al.*, 2009; Murray, Delahunty & Baxter, 2001; Torrens, Rlu-Aumatell, Vichi, *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 (Ares, Deliza, Barreiro, *et al.*, 2010; Fleming, Ziegler & Hayes, 2015) and non-verbal methods such as rating, which measures a single sensory character of each sample (Ballester, Patris, Symoneaux, *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 (Ballester *et al.*, 2008; Mafata, Buica, du Toit & van Jaarsveld, 2018; Valentin *et al.*, 2012).

2.2.2 Statistical approaches taken in evaluating oenological experiments

Advances in statistical data handling techniques have naturally progressed to analyse more variables simultaneously from univariate, bivariate, multivariate, to what is sometimes called megavariate data analysis (Eriksson, Johansson, Kettaneh-Wold, Trygg, Wikstr, *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 *et al.*, 2014).

Megavariate is often used for advanced multivariate techniques that use multiple sets of data acquired from different sources, requiring specialised statistical treatment (Eriksson, Johansson, Kettaneh-Wold, Trygg, Wikstr, *et al.*, 2006). The multiple data sets are designated as blocks and used in multiblock and data fusion approaches (Cocchi, 2019). 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 oenological approaches mainly due to an increase in the number methods from both Chemistry and Sensory (Alañón *et al.*, 2015). Chemistry methods have increased in numbers and sophistication, in accordance with advances in technological and computing power (Alañón *et al.*, 2015; Borràs *et al.*, 2015; Gagolewski, 2012). 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 (Valentin *et al.*, 2012; Varela & Ares, 2014). 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, Mihnea, Peyron, *et al.*, 2013; Hayward, Jantzi, Smith, *et al.*, 2020).

From an applied perspective, multivariate statistical analyses can be categorised under supervised and unsupervised methods (Sohail & 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 & Arif, 2020). From a theoretical perspective, multivariate methods can be categorized as parametric (classical approach) and non-parametric (non-classical/advanced approach) (Härdle & 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 generalised correspondence analyses). Non-parametric analyses such as machine learning techniques do not assume normal distribution or a fixed average (Härdle & Simar, 2015).

Research in Oenology frequently seeks to understand what drives/contributes to predictions and classification, and hence use supervised data analyses to find the discriminating markers (Brand, Panzeri & Buica, 2020). Advanced data handling techniques such as *k*-nearest neighbours (kNN) have provided a good starting point to dig deeper into these types questions (De Carvalho Rocha, Do Prado & Blonder, 2020). In Oenology and Sensory research, Artificial Intelligence (AI) applications have been used in supervised strategies (De Carvalho Rocha *et al.*, 2020; Valente, Bauer, Venter, *et al.*, 2018). 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 (De Carvalho Rocha *et al.*, 2020).

2.3 Model input

Only what has been captured can be modelled; 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 Oenology. Before the data captured can be modelled, several decision steps concerning the pre-modelling processes of the data and the modelling specifications must be taken. Several standardised pre-modelling 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 oenological data perspective. There is an imbalance of greater detailing of the strategy behind the method compared to the data modelling. This section will cover how to convey important specifications and create a complete methodology based on important aspects of the data input stage.

2.3.1 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 (McKillup, 2005; Yu, Low & Zhou, 2018). Several experimental designs have been development from a statistical perspective (Ferreira, 2019), as well as for natural sciences perspective, including Chemistry (Kreutz & Timmer, 2009). Recently, design of experiments (DOEs) that are particularly sensitive to the structure and premise of sensory methods have been reviewed (Yu *et al.*, 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 modelling, 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 *et al.*, 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 analyser in order to monitor a single m/z channel) or total ion current (TIC - resulting from the Full Scan exploitation of the mass analyser) 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, Papi, Kyriakidis, *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 made clear and unambiguous to collect relevant data which is compatible with the experimental design. Some methods may have verbal and non-verbal aspects to them; 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 & Abdi, 2005; Faye, Courcoux, Giboreau, *et al.*, 2013; Makhotkina, Pineau & 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.

2.3.2 Pre-modelling 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 biases 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 modelling (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 (Deneulin & Bavaud, 2016; Salkind. J. & Kristin. R., 2007). Targeted analyses tend to produce data sets with smaller dimensions/variables than untargeted data sets and generally are not pre-processed (Engel, Gerretzen, Szyman´ska, *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 colour density (OIV, 2006; Ribereau-Gayon, Glories, Maujean, *et al.*, 2006; Waterhouse, 2002). Untargeted data sets often have associated pre-processing methods such as those developed for IR, NMR, Raman spectroscopy, and UV-Vis (Campos & Reis, 2020; Rinnan, Berg & Engelsen, 2009). Untargeted data sets have inherent issues related to their acquisition, and the nature of the sample for which the pre-processing is done to address these issues such as baseline offset and noise and saturated peaks often seen in NMR, IR, and UV-Vis spectra (Engel *et al.*, 2013; Rinnan *et al.*, 2009).

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 *et al.*, 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 colour/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 (Gawel, Oberholster & Leigh. Francis, 2000; Lawless & Civille, 2013; Pickering & Demiglio, 2008).

In practice, the approach is from the lowest level upwards or a bottom-up approach (synonyms, lemmatisation, 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, Maller, Kapsalis, *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, Valentin, Sulmont-Rossé, *et al.*, 2005; Deneulin & 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 cut-off 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, emphasising 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.

2.4 Data modelling 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 that is generated. The main aspects of choosing which data modelling 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 modelling. Depending on the type of data available, some modelling 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 modelling that can be done. Large data sets with high sample number (distinguishable samples, not including repeats), high sample variability, large number and diverse nature of measurements, can be modelled 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 modelling tools.

As previously mentioned, the algorithms that are used to model data can be either supervised or unsupervised (Sohail & Arif, 2019). The mathematical aspects related to these models will not be covered in this review, which will take a process-centred look at the aspects of modelling from an application point-of-view. 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.

2.4.1 Matrix compatibility

Chemical data generally has standardized outputs in compatible matrices, making various data modelling 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 RT in the FLD chromatogram and the RT in the MS chromatogram). The distribution of data in a discreet data set *vs* a continuous data set are different, making it difficult to combine the two. Since the data is scaled before modelling, the assumption in statistical context is that the distribution of the two are 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 developed with the statistics as part of the design of experiments (Valentin *et al.*, 2012; Yu *et al.*, 2018). As previously discussed, the execution has implications on the data analysis. The sensory matrix captured is dependent on the method, including co-ordinates (*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. The 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, Panzeri, Brand, *et al.*, 2020; Hayward *et al.*, 2020). The implication is that the matrices are then modelled differently based on the different types of matrices captured.

2.4.2 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 modelling or data fusion (Gagolewski, 2015; Handling & Science, 2019; Lahat, Adali & Jutten, 2015; Vera, Aceña, Aceña, *et al.*, 2011).

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

2.4.3 Supervised modelling

Supervised models are used for classification, discrimination, or prediction (Sohail & Arif, 2020). These models are based on a measurable trend/regression which distinguishes one set of samples or variables from another (Sohail & 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 discrimination and classification models are qualitative. These models are used in Oenology to classify samples according to regionality, cultivar, and wine styles among others (Cuadros-Inostroza, Giavalisco, Hummel, *et al.*, 2010; Edelmann, Diewok, Schuster, *et al.*, 2001; Makris, Kallithraka & Mamalos, 2006).

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 (Engel *et al.*, 2013; Moyano *et al.*, 2018; Petrovic, Aleixandre-Tudo & Buica, 2019). 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 which represent a certain phenomenon such as predicting total antioxidant capacity (TAC) (Versari, Parpinello, Scazzina, *et al.*, 2010) or yeast assimilable nitrogen (YAN) (Petrovic *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 modelling in Oenology uses least squares for classification (Borràs et al., 2016; Silvestri et al., 2014; Vera, Aceña, Aceña, et al., 2011) and discrimination (Vera, Aceña, Guasch, et al., 2011). Some prediction models in Oenology 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 towards non-invasive sample preparations, untargeted chemical analysis, and data fusion strategies for coupling and ultimately predicting sensory perception from chemistry data (Brand *et al.*, 2020; Seisonen, Vene & Koppel, 2016). Additionally, to address this shortcoming, studies have advocated for the use of advanced techniques such as artificial intelligence and machine learning (Seisonen *et al.*, 2016).

2.4.4 Performance parameters and model optimisation

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 & Simar, 2015; Salkind. J. & Kristin. R., 2007). This section will address the parameters most often reported in Oenological 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, and 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, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; Härdle & Simar, 2015; McKillup, 2005; Salkind. J. & Kristin. R., 2007; Wheelock, 2002).

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 (Härdle & Simar, 2015; Kruskal, 1977; McKillup, 2005; Robinson, Boss, Solomon, *et al.*, 2014; Salkind. J. & Kristin. R., 2007). The stress is a relative measure of the total explained variation in the model (McKillup, 2005; Salkind. J. & Kristin. R., 2007). 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 Oenology (Alañón *et al.*, 2015; Valente *et al.*, 2018; Valentin *et al.*, 2012). 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 RARA have %EV similar to chemistry data because, similar to chemistry, their data is based on intensity (Brand, 2019; Valentin *et al.*, 2012). 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 (Brand, 2019; Valentin *et al.*, 2012).

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 *et al.*, 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 (Borràs *et al.*, 2015; Lahat *et al.*, 2015). When the %EV for the first two dimensions are low, the efficiency of the model can be communicated by looking at the first three dimensions (Parr, Ballester, Peyron, *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, Brand, Panzeri, *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 & 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, Benton, Rinehart, *et al.*, 2015; Kruskal, 1977; Naumann, Lasch, Diem, *et al.*, 2007). 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 (Härdle & Simar, 2015; Myhre, Mikalsen, Løkse, *et al.*, 2018). 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 (Edelmann *et al.*, 2001; Myhre *et al.*, 2018; Radovanovic, Jovancicevic, Arsic, *et al.*, 2016).

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 (Eriksson, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; Wheelock, 2002). Effective use of latent variables in pre-processing steps to improve model performance has been considered from an applications perspective (Iorgulescu, Voicu, Sârbu, *et al.*, 2016) and a statistical method perspective (Engel *et al.*, 2013). Considering the previous data handling steps discussed in this review, ensuring improved model performance requires attention to detail from both perspectives (Campos & Reis, 2020; Gerretzen, Szymańska, Jansen, *et al.*, 2015).

Model optimization for unsupervised and supervised models can be done similarly from statistical and application perspectives (Iorgulescu *et al.*, 2016). In the Oenology context, it is especially necessary to have both of these 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 (Fleming *et al.*, 2015; Valentin *et al.*, 2012). 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 *et al.*, 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 (Engel *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-modelling optimization which requires variable selection can be done in supervised modelling 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 (Engel *et al.*, 2013; Lahat *et al.*, 2015; Rinnan *et al.*, 2009). Supervised models are more often optimised compared to unsupervised; this goes hand in hand with more applications using supervised than unsupervised modelling.

These principles for optimisation applied in Oenology include variable selection, feature selection, and using latent variables as pre-processing techniques coupled to supervised strategies such as PLS (Guld *et al.*, 2020; Larsen, van den Berg & Engelsen, 2006; Pereira *et al.*, 2016; Petrovic, 2018; Seisonen *et al.*, 2016). Variable selection has been used for choosing certain wavenumbers in IR modelling *a priori* (before the modelling 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, Quintelas, Mesquita, *et al.*, 2018). Feature selection has been done on similar data using IR, NMR, and UV-Vis for the selection of principal components (Borràs *et al.*, 2015; Pereira *et al.*, 2016) and/or the use of latent variables for optimizing untargeted spectral data (Brand *et al.*, 2020; Cuadros-Inostroza *et al.*, 2010; Godelmann *et al.*, 2013).

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 coefficients, RV, for clustering).

2.5 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 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, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; McKillup, 2005; Wheelock, 2002). 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. 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 (Eriksson, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; McKillup, 2005; Wheelock, 2002). In oenological 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 2D projections or as a 3D graph (Ballester, Dacremont, Fur, *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 & 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 & 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 (Brand, 2019; Pagés & Husson, 2005).

Cluster analysis can be applied to the Cartesian plots, containing as many dimensions as needed for pattern recognition, visualised 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, 2007b,a; Kruskal, 1977). These matrices can be visualised 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 visualised 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 (Brand *et al.*, 2020; Mafata *et al.*, 2019). 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, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; Wheelock, 2002).

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 (Cariou & Qannari, 2018; Hayward *et al.*, 2020; Mafata, Buica, du Toit, Panzeri, *et al.*, 2018; Valentin *et al.*, 2012).

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 modelling 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 *et al.*, 2018). Additionally, the manner (*i.e*. the intuitiveness/level of difficulty) and order of execution of the tasks may influence the success of the modelling (Brand, 2019; Valentin *et al.*, 2012). It can happen that the judges are better at executing the secondary task, in which case the contextual interpretation of the results have to take this into account.

2.6 Data fusion and advanced data modelling in Oenology

The most recent trends in data modelling for Oenology are towards the use of artificial intelligence (AI) (Garrido-Delgado, Arce, Guamán, *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 (Handling & Science, 2019). Data fusion approaches can use algorithms from both classical multivariate modelling and AI at different levels of complexity using either supervised or unsupervised techniques (Cocchi, 2019).

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 modelled 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.

2.6.1 Data fusion frameworks

Data fusion is classified under low, medium, and high level (Figure 2.1) according to increasing levels of complexity (Borràs *et al.*, 2015; Handling & Science, 2019; Lahat *et al.*, 2015), taking both statistical (Handling & Science, 2019) and strategic approach (Lahat *et al.*, 2015). Oenological data fusion strategies used for these levels have been reviewed by Borràs *et al*. (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, 2019). It is for this reason that it is often not called data fusion but rather data aggregation or concatenation (Borràs *et al.*, 2015; Cocchi, 2019). The implications of data concatenation are that the data sets are dependent and vary similarly in scale and distribution (Härdle & Simar, 2015). In Oenology, low-level data fusion is commonly done on targeted measurements but keep 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 *et al.*, 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 & 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 & Heymann, 2012; Seisonen *et al.*, 2016). One study has also attempted to use descriptive analysis profile of wine to predict typicality with good success (Coulon-Leroy, Poulzagues, Cayla, *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 (Figure 2.1) (Borràs *et al.*, 2015; Cocchi, 2019; de Juan, Gowen, Duponchel, *et al.*, 2019). 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.*) (Bro, 1997; Niimi, Boss & Bastian, 2018; Silvestri, Elia, Bertelli, *et al.*, 2014). Pre-processing for matrix compatibility also includes mathematical transformations (rating converted to frequency data) and the use of exploratory modelling for scaling (Campos & Reis, 2020; Engel *et al.*, 2013; Rinnan *et al.*, 2009). Although supervised data fusion approaches are more common in Oenology, unsupervised approaches are gaining popularity.

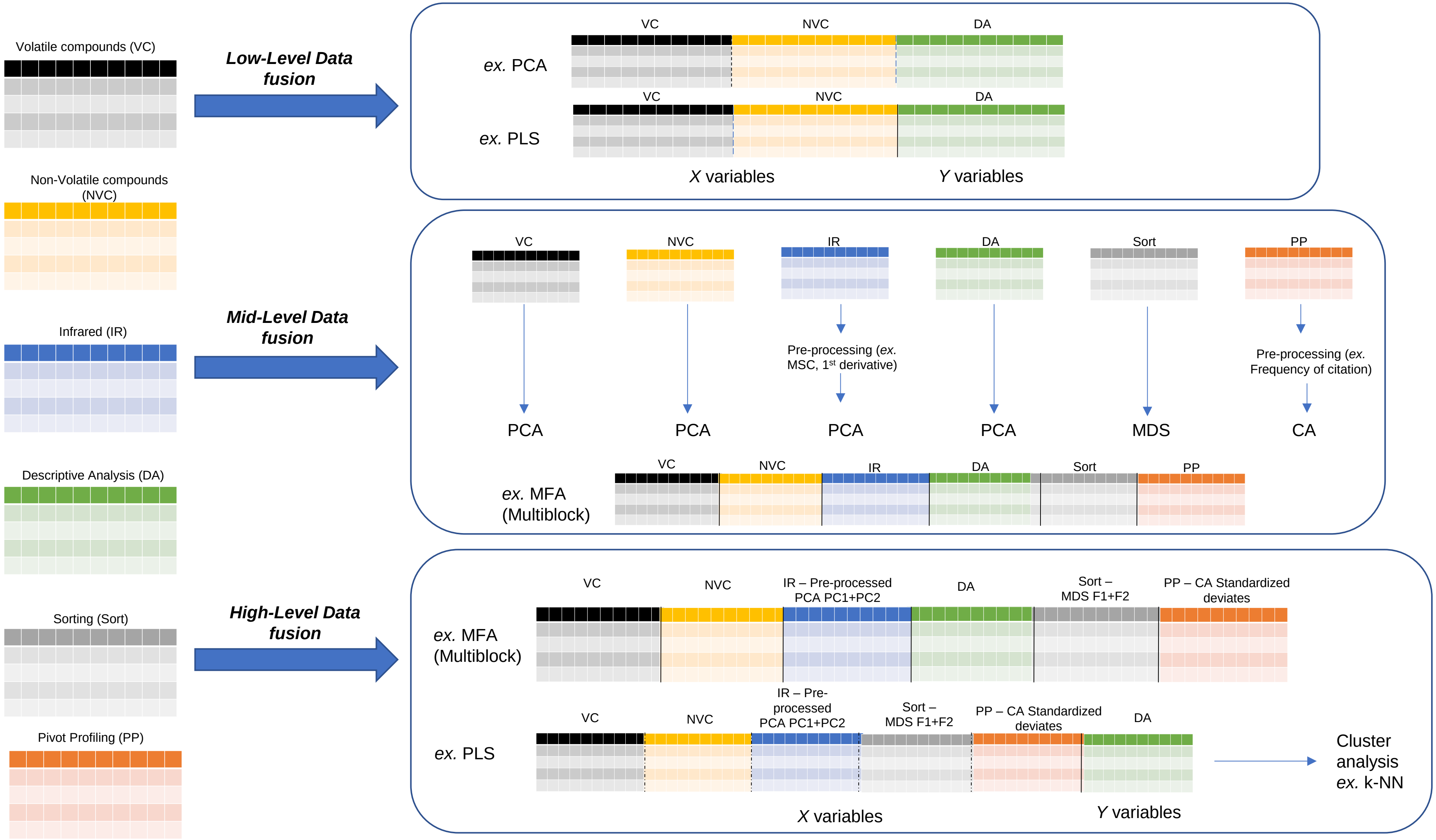


Figure 2.1: Example of theoretical framework for data fusion. Dotted line designates a 'soft boundary' between data sets (concatenation), while a full line designates a 'hard boundary' (multiblock). MFA – multiple factor analysis; PLS – partial least squares; CA – correlation analysis; k-NN – k-nearest neighbours; MDS – multidimensional scaling.

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, Bauer, Panzeri, *et al.*, 2019) as well as furanmethanethiol (FMT) and coffee aroma in Pinotage wines (Garrido-Bañuelos & Buica, 2020). Supervised mid-level data fusion approaches have been of relevance to Oenology 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 & Cantos-Villar, 2017; Cozzolino, Smyth, Lattey, *et al.*, 2005; Culbert, Cozzolino, Ristic, *et al.*, 2015; Fudge, Wilkinson, Ristic, *et al.*, 2013; Gambetta, Cozzolino, Bastian, *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 (Borràs *et al.*, 2015; Handling & Science, 2019). Also called decision-level data fusion, these approaches maximize informational value, precision, and accuracy (Borràs *et al.*, 2015; Cocchi, 2019). 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 *et al.*, 2019). This means that model performance and optimization are very important aspects in these strategies. In Oenology, modelling 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 *et al.*, 2015; Borràs *et al.*, 2015). These high-level strategies involve process technology for acquisition, monitoring, and modelling process outcomes (Borràs *et al.*, 2015; Cocchi, 2019; Ríos-Reina, Azcarate, Cami, *et al.*, 2020). Examples include the use of infrared spectroscopy for accurate predictions of oenological parameters such as yeast assimilable nitrogen (YAN) (Petrovic *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 (Alañón *et al.*, 2015; Borràs *et al.*, 2015; Cavaglia, Schorn-García, Giussani, *et al.*, 2020; Fourie, Luis Aleixandre-Tudo, Mihnea, *et al.*, 2020).

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 Oenological applications. This is especially true for cases that have used advanced modelling 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 modelling strategies, the full process to the approach, which may contain elements of other levels of data fusion, are not always communicated. This can create misconceptions about the level of difficulty in fusing 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 modelling types. There are so many modelling 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.

2.6.2 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 artificial intelligence (AI) systems. Using classical statistics, supervised modelling advancements have worked towards increasing the calibration and discriminative power of models (Eriksson, Johansson, Kettaneh-Wold, Trygg, Wikstrom, *et al.*, 2006; Härdle & Simar, 2015; McKillup, 2005). Both supervised and unsupervised modelling are advancing towards 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 (De Carvalho Rocha *et al.*, 2020).

Classical multivariate analyses derive linear relationships and linear regression algorithms based on normal parametric distribution (Härdle & Simar, 2015; McKillup, 2005). 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 analysing big data. Artificial intelligence is more nuanced in that it accommodates non-binary (*i.e.* classifications) and non-linear (*i.e.* calibrations) relationships (De Carvalho Rocha *et al.*, 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 (Arbara, De Andrade, De Gois, *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 (De Carvalho Rocha *et al.*, 2020). The review narrated some important behavioural 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 behaviour 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 modelling as pre-processing steps. The strategy has been to use classical MVA followed by AI analysis for pattern recognition (Figure 2.1), *i.e.* the data is first normalised using classical MVA and then AI is applied (Härdle & Simar, 2015; Myhre *et al.*, 2018). 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 (Myhre *et al.*, 2018).

The most common documented uses of AI in Oenology include support vector machines (SVM), self-organising maps (SOM), and k-nearest neighbours (k-NN) or k-means clustering (De Carvalho Rocha *et al.*, 2020). Both generally and in Oenology, 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 (De Carvalho Rocha *et al.*, 2020). In Oenology, 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 (Borràs *et al.*, 2015; Gómez-Meire, Campos, Falqué, *et al.*, 2014; Latorre, García-Jares, Médina, *et al.*, 1994). SOM has 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 modelling 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 Oenology. They could be capable of elucidating answers to big questions in Oenology such as sensory and chemistry markers of wine quality, as well as wine authenticity.

2.7 Conclusions

The aim of this review was to examine the different stages of the data handling process in Oenology 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 Oenology, this review identified the key decision-making aspects during the data input stage (capturing and pre-processing), the modelling, and the model output (visualisation/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 modelled. 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 rationalised the intertwining of statistical and applied reasoning for interpretation of modelling 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 Oenology towards 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

[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

strategies since this has critical contribution to the philosophy and progression of science and research.

References

Abdi, H. 2007a. Metric Multidimensional Scaling In: Encyclopedia of Measurement and Statistics.

Abdi, H. 2007b. RV Coefficient and Congruence Coefficient. in *Encyclopedia of Measurement andStatistics*. [Online], Available: http://www.utd.edu/ [2018, November 23].

Abdi, H. & Valentin, D. 2007. Multiple Correspondence Analysis. in *Encyclopedia of Measurement and Statistics* Thousand Oaks (CA): Sage. 651–657. [Online], Available: http://www.utd.edu/ [2020, September 25].

Alañón, M., Pérez-Coello, M. & Marina, M. 2015. Wine science in the metabolomics era. *Trends in Analytical chemistry*. 74:1–20.

Arbara, B., De Andrade, M., De Gois, J.S., Xavier, V.L. & Luna, A.S. 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.

Ares, G., Deliza, R., Barreiro, C., Giménez, A. & Gámbaro, A. 2010. Comparison of two sensory profiling techniques based on consumer perception. *Food Quality and Preference*. 21(4):417–426.

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):461–469.

Ballester, J., Dacremont, C., Fur, Y. Le & Etiévant, P. 2005. The role of olfaction in the elaboration and use of the Chardonnay wine concept. *Food Quality and Preference*. 16(4):351–359.

Ballester, J., Patris, B., Symoneaux, R. & Valentin, D. 2008. Conceptual vs. perceptual wine spaces: Does expertise matter? *Food Quality and Preference*. 19(3):267–276.

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

Boccard, J. & Rutledge, D.N. 2014. Iterative weighting of multiblock data in the orthogonal partial least squares framework. *Analytica Chimica Acta*. 813:25–34.

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

Brand, J. 2019. Rapid sensory profiling methods for wine : Workflow optimisation for research and industry applications. Stellenbosch University.

Brand, J., Panzeri, V. & Buica, A. 2020. Wine quality drivers: A case study on South African chenin blanc and pinotage wines. *Foods*. 9(6):1–17.

Bro, R. 1997. Chemometrics and intelligent laboratory systems Tutorial PARAFAC. Tutorial and applications. *Chemomemcs and Intelligent Laboratory Systems*. 38:149–171. [Online], Available: https://www.cs.cmu.edu/~pmuthuku/mlsp_page/lectures/Parafac.pdf [2018, November 13].

Campo, E., Ballester, J., Langlois, J., Dacremont, C. & Valentin, D. 2009. Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines.

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

Cardello, A. V., Maller, O., Kapsalis, J.G., SEGARS, R.A., SAWYER, F.M., MURPHY, C. & MOSKOWITZ, H.R. 1982. Perception of Texture by Trained and Consumer Panelists. *Journal of Food Science*. 47(4):1186–1197.

Cariou, V. & Qannari, E.M. 2018. Statistical treatment of free sorting data by means of correspondence and cluster analyses.

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

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

Cavaglia, J., Schorn-García, D., Giussani, B., Ferr, J., Busto, O., Ace, L., Mestres, M., Boqu, R., et al. 2020. Monitoring wine fermentation deviations using an ATR-MIR spectrometer and MSPC charts.

Cayuela, J.A., Puertas, B. & Cantos-Villar, E. 2017. Assessing wine sensory attributes using Vis/NIR. *European Food Research and Technology*. 243:941–953.

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

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

Cocchi, M. 2019. Introduction: Ways and Means to Deal With Data From Multiple Sources. in *Data Handling in Science and Technology* Vol. 31. Elsevier Ltd. 1–26.

Coetzee, C., Van Wyngaard, E., Šuklje, K., Silva Ferreira, A.C. & Du Toit, W.J. 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):7979–7993.

Coulon-Leroy, C., Poulzagues, N., Cayla, L., Symoneaux, R. & Masson, G. 2018. Is the typicality of “provence Rosé wines” only a matter of color? *Oeno One*. 52(4):1–15.

Cozzolino, D., Smyth, H.E., Lattey, K.A., Cynkar, W., Janik, L., Dambergs, R.G., Francis, I.L. & Gishen, M. 2005. Relationship between sensory analysis and near infrared spectroscopy in Australian Riesling and Chardonnay wines. *Analytica Chimica Acta*. 539:341–348.

Cuadros-Inostroza, A., Giavalisco, P., Hummel, J., Eckardt, A., Willmitzer, L. & Peña-Cortés, H. 2010. Discrimination of wine attributes by metabolome analysis. *Analytical Chemistry*. 82(9):3573–3580.

Culbert, J., Cozzolino, D., Ristic, R. & Wilkinson, K. 2015. Classification of sparkling wine style and quality by MIR spectroscopy. *Molecules*. 20(5):8341–8356.

Deneulin, P. & Bavaud, F. 2016. Analyses of open-ended questions by renormalized associativities and textual networks: A study of perception of minerality in wine. *Food Quality and Preference*. 47:34–44.

Le Dien, S. & Pagès, J. 2003. Hierarchical Multiple Factor Analysis: application to the comparison of sensory profiles. *Food Quality and Preference*. 14(5–6):397–403.

Edelmann, A., Diewok, J., Schuster, K.C. & Lendl, B. 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):1139–1145.

Engel, J., Gerretzen, J., Szyman´ska, E., Szyman´ska, S., Jansen, J.J., Downey, G., Blanchet, L. & Buydens, M.C. 2013. Breaking with trends in pre-processing? *Trends in Analytical chemistry*. 50:96–106.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstr, C. & Wold, S. 2006. Multi- and Megavariate Data Analysis. Part I Basic Principles and Applications. Second revised and enlarged edition. *Ume Sweden: MKS Umetrics AB*. (January, 1):1–103.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C. & Wold, S. 2006. *Multivariate and Megavariate Data Analysis Basic Principles and Applications (Part I)*. Vol. 16.

Faye, P., Courcoux, P., Giboreau, A. & Qannari, E.M. 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*.

Ferreira, S.L.C. 2019. Chemometrics and statistics | experimental design. in *Encyclopedia of Analytical Science* Elsevier. 420–424.

Fleming, E.E., Ziegler, G.R. & Hayes, J.E. 2015. Check-all-that-apply (CATA), sorting, and polarized sensory positioning (PSP) with astringent stimuli. *Food Quality and Preference*. 45:41–49.

Fourie, E., Luis Aleixandre-Tudo, J., Mihnea, M. & Du Toit, W. 2020. Partial least squares calibrations and batch statistical process control to monitor phenolic extraction in red wine fermentations under different maceration conditions.

Fudge, A.L., Wilkinson, K.L., Ristic, R. & Cozzolino, D. 2013. Synchronous two-dimensional MIR correlation spectroscopy (2D-COS) as a novel method for screening smoke tainted wine. *Food Chemistry*. 139:115–119.

Gagolewski, M. 2012. Data fusion. in O. Hryniewicz, J. Mielniczuk, W. Penczek, & J. Waniewski (eds.) O. Hryniewicz, J. Mielniczuk, W. Penczek, & J. Waniewski (eds.). 69–70.

Gagolewski, M. 2015. *Data fusion. Theory, methods, and applications*. O. Hryniewicz, J. Mielniczuk, W. Penczek, & J. Waniewski (eds.).

Gambetta, J.M., Cozzolino, D., Bastian, S.E.P. & Jeffery, D.W. 2019. Classification of Chardonnay Grapes According to Geographical Indication and Quality Grade Using Attenuated Total Reflectance Mid-infrared Spectroscopy. *Food Analytical Methods*. 12(1):239–245.

Garrido-Bañuelos, G. & Buica, A. 2020. Is There a Link Between Coffee Aroma and the Level of Furanmethanethiol (FMT) in Pinotage Wines ? *South African Journal Enology and Viticulture*. 41(2):245–250.

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

Garrido-Bañuelos, G., Panzeri, V., Brand, J. & Buica, A. 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).

Gawel, R., Oberholster, A. & Leigh. Francis, I. 2000. A ‘Mouth-feel Wheel’: terminology for communicating the mouth-feel characteristics of red wine. *Australian Journal of Grape and Wine Research*. 6(3):203–207.

Genisheva, Z., Quintelas, C., Mesquita, D.P., Ferreira, E.C., Oliveira, J.M. & Amaral, A.L. 2018. New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR).

Food Chemistry. 246:172–178.

Gerretzen, J., Szymańska, E.S., Jansen, J.J., Bart, J., Van Manen, H.-J., Van Den Heuvel, E.R. & Buydens, L.M.C. 2015. Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Anal. Chem*. 87:12096–12103.

Godelmann, R., Fang, F., Humpfer, E., Schütz, B., Bansbach, M., Schäfer, H. & Spraul, M. 2013. Targeted and nontargeted wine analysis by1H NMR spectroscopy combined with multivariate statistical analysis. differentiation of important parameters: Grape variety, geographical origin, year of vintage. *Journal of Agricultural and Food Chemistry*. 61(23):5610–5619.

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F. & Fdez-Riverola, F. 2014. Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International*. 50:230–240.

Granato, D. & Ares, G. 2013. *Mathematical and Statistical Methods in Food Science and Technology*. wiley.

Granato, D., de Araújo Calado, V.M. & Jarvis, B. 2014. Observations on the use of statistical methods in Food Science and Technology. *Food Research International*. 55:137–149.

Guld, Z., Nyitrainé Sárdy, D., Gere, A. & Rácz, A. 2020. Comparison of sensory evaluation techniques for Hungarian wines. *Journal of Chemometrics*. 34(4):1–15.

Handling, D. & Science, I.N. 2019. *Data Fusion Methodology and Applications*. Vol. 31. M. Cocchi (ed.).

Härdle, W.K. & Simar, L. 2015. *Applied multivariate statistical analysis, fourth edition*.

Hayward, L., Jantzi, H., Smith, A. & McSweeney, M.B. 2020. How do consumers describe cool climate wines using projective mapping and ultra-flash profile? *Food Quality and Preference*. 86:104026.

Hopfer, H., Ebeler, S.E. & Heymann, H. 2012. The Combined Effects of Storage Temperature and Packaging Type on the Sensory and Chemical Properties of Chardonnay. *Journal of Agricultural and Food Chemistry*. 60:10743–10754.

Hunter, E.A., Dijksterhuis, G.B., Qannari, E.M. & Macfie, H.J.H. 1995. Second Sensometrics Meeting-Edinburgh, 16-18 September 1994: introduction on behalf of the organising committee. *Food Quality and Preference*. 6:215–216.

Iorgulescu, E., Voicu, V.A., Sârbu, C., Tache, F., Albu, F. & Medvedovici, A. 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 extrac. *Talanta*. 155:133–144.

Ivanisevic, J., Benton, H.P., Rinehart, D., Epstein, A., Kurczy, M.E., Boska, M.D., Gendelman, H.E. & Siuzdak, G. 2015. An interactive cluster heat map to visualize and explore multidimensional metabolomic data. *Metabolomics*. 11(4):1029–1034.

de Juan, A., Gowen, A., Duponchel, L. & Ruckebusch, C. 2019. Image Fusion. in *Data Handling in Science and Technology* Vol. 31. Elsevier Ltd. 311–344.

Kowalski, B.R. 1980. Chemometrics. *Analytical Chemistry*. 52(5):112–122. [Online], Available: https://pubs.acs.org/sharingguidelines.

Kreutz, C. & Timmer, J. 2009. Systems biology: Experimental design. *FEBS Journal*. 276(4):923–942.

Kruskal, J. 1977. The Relationship between Multidimensional Scaling and Clustering. in *Classification and Clustering* Elsevier. 17–44.

Lahat, D., Adali, T. & Jutten, C. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*. 103(9):1449–1477.

Lapalus, E., Wessel, P. & Du Toit, J. 2016. Linking sensory attributes to selected aroma compounds in South African Cabernet Sauvignon wines. (March).

Larsen, F.H., van den Berg, F. & Engelsen, S.B. 2006. An exploratory chemometric study of1H NMR spectra of table wines. *Journal of Chemometrics*. 20(5):198–208.

Latorre, M.J., García-Jares, C., Médina, B. & Herrero, C. 1994. *Pattern Recognition Analysis Applied to Classification of Wines from Galicia (Northwestern Spain) with Certified Brand of Origin*. [Online], Available: https://pubs.acs.org/sharingguidelines [2020, September 28].

Lawless, L.J.R. & Civille, G. V. 2013.

Mafata, M., Buica, A., du Toit, W.J. & van Jaarsveld, F.P. 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):141–148.

Mafata, M., Buica, A., du Toit, W., Panzeri, V. & van Jaarsveld, F.P. 2018. The effect of grape temperature on the sensory perception of Méthode Cap Classique wines. *South African Journal of Enology and Viticulture*. 39(1):132–140.

Mafata, M., Brand, J., Panzeri, V., Kidd, M. & Buica, A. 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(February):108515.

Mafata, M., Brand, J., Panzeri, V. & Buica, A. 2020. Investigating the Concept of South African Old Vine Chenin Blanc Investigating the Concept of South African Old Vine Chenin Blanc. *South African Journal of Enology & Viticulture*. 41(2):168–182.

Makhotkina, O., Pineau, B. & Kilmartin, P.A. 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):91–99.

Makris, D.P., Kallithraka, S. & Mamalos, A. 2006. Differentiation of young red wines based on cultivar and geographical origin with application of chemometrics of principal polyphenolic constituents. *Talanta*. 70(5):1143–1152.

McKay, M., Bauer, F.F., Panzeri, V., Mokwena, L. & Buica, A. 2019. Profiling potentially smoke tainted red wines: Volatile phenols and aroma attributes. *South African Journal of Enology and Viticulture*. 40(2):1–16.

McKillup, S. 2005. *Statistics explained: An introductory guide for life scientists*. Cambridge University Press.

McKillup, S. 2012. *Statistics explained : an introductory guide for life scientists*. 2nd ed. Cambridge University Press.

Moyano, L., Serratosa, M.P., Marquez, A. & Zea, L. 2018. Optimization and validation of a DHS-TD-GC-MS method to wineomics studies.

Murray, J.., Delahunty, C.. & Baxter, I.. 2001.

Myhre, J.N., Mikalsen, K.Ø., Løkse, S. & Jenssen, R. 2018. Robust clustering using a kNN mode seeking ensemble R. *Pattern Recognition*. 76:491–505.

Naumann, D., Lasch, P., Diem, M. & Ha, W. 2007. Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. *Journal of Chemometrics*. (November):209–220.

Niimi, J., Tomic, O., Næs, T., Jeffery, D.W., Bastian, S.E.P. & Boss, P.K. 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(November 2017):195–202.

Niimi, J., Boss, P.K. & Bastian, S.E.P. 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*. 106:304–316.

OIV. 2006. Determination of chromatic characteristics according to CIELab. *Compendium of International Analysis of Methods*. (Chromatic Characteristics):1–16. [Online], Available: http://www.oiv.int/oiv/files/6 - Domaines scientifiques/6 - 4 Methodes d analyses/6-4-1/EN/OIV-MA-AS2-11.pdf.

Pagés, J. & Husson, F. 2005. Multiple factor analysis with confidence ellipses: A methodology to study the relationships between sensory and instrumental data. *Journal of Chemometrics*. 19(3):138–144.

Pagès, J. 2004. Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadistica*. 27(1):1–26.

Parr, W. V., Ballester, J., Peyron, D., Grose, C. & Valentin, D. 2015. Perceived minerality in Sauvignon wines: Influence of culture and perception mode. *Food Quality and Preference*. 41:121–132.

Pereira, A.C., Reis, M.S., Saraiva, P.M. & Marques, J.C. 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.

Pereira, A.C., Carvalho, M.J., Miranda, A., Leça, J.M., Pereira, V., Albuquerque, F., Marques, J.C. & Reis, M.S. 2016. Modelling the ageing process: A novel strategy to analyze the wine evolution towards the expected features. *Chemometrics and Intelligent Laboratory Systems*. 154:176–184.

Petrovic, G. 2018. A survey of the YAN status of South African grape juices and exploration of multivariate data analysis techniques for spectrometric calibration and cultivar discrimination purposes.

Petrovic, G., Aleixandre-Tudo, J.L. & Buica, A. 2019. Unravelling the complexities of wine: A big data approach to yeast assimilable nitrogen using InfraRed spectroscopy and chemometrics. *Oeno One*. 53(2):107–127.

Pickering, G.J. & Demiglio, P. 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.

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

Ribereau-Gayon, P., Glories, Y., Maujean, A. & Dubourdieu, D. 2006. *Handbook of enology*. 2nd ed. P. Ribereau-Gayon (ed.). Wiley Online Library.

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

Ríos-Reina, R., Callejón, R.M., Savorani, F., Amigo, J.M. & Cocchi, M. 2019. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta*. 198:560–572.

Ríos-Reina, R., Azcarate, S.M., Cami, J.M. & ector Goicoechea, H.C. 2020. Multi-level data fusion strategies for modeling three-way electrophoresis capillary and fluorescence arrays enhancing geographical and grape variety classification of wines.

Robinson, A.L., Boss, P.K., Solomon, P.S., Trengove, R.D., Heymann, H. & Ebeler, S.E. 2014. Origins of Grape and Wine Aroma. Part 2. Chemical and Sensory Analysis. *Am. J. Enol. Vitic*. 65(1).

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

Seisonen, S., Vene, K. & Koppel, K. 2016. The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. *Food Chemistry*. 210:530–540.

Serra-Cayuela, A., Jourdes, M., Riu-Aumatell, M., Buxaderas, S., Teissedre, P.L. & López-Tamames, E. 2014. Kinetics of browning, phenolics, and 5-hydroxymethylfurfural in commercial sparkling wines. *Journal of Agricultural and Food Chemistry*.

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

Sohail, A. & Arif, F. 2019. Supervised and unsupervised algorithms for bioinformatics and data science.

Sohail, A. & Arif, F. 2020. Supervised and unsupervised algorithms for bioinformatics and data science. *Progress in Biophysics and Molecular Biology*. 151:14–22.

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

Du Toit, W.J. & Piquet, C. 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*.

Torrens, J., Rlu-Aumatell, M., Vichi, S., López-Tamames, E. & Buxaderas, S. 2010. Assessment of volatlle and sensory profiles between base and sparkling wines. *Journal of Agricultural and Food Chemistry*.

Trikas, E.D., Papi, R.M., Kyriakidis, D.A. & Zachariadis, G.A. 2016. A sensitive LC-MS method for anthocyanins and comparison of byproducts and equivalent wine content. *Separations*. 3(2).

Valente, C.C., Bauer, F.F., Venter, F., Watson, B. & Nieuwoudt, H.H. 2018. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. *Scientific Reports*. 8(1).

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

Varela, P. & Ares, G. 2014. Novel Techniques in Sensory Characterization and Consumer Profiling. in.

Vera, L., Aceña, L., Aceña, A., Guasch, J., Boqué, R., Mestres, M. & Busto, O. 2011. Discrimination and sensory description of beers through data fusion. *Talanta*. 87:136–142.

Versari, A., Parpinello, G.P., Scazzina, F. & Rio, D. Del. 2010. Prediction of total antioxidant capacity of red wine by Fourier transform infrared spectroscopy. *Food Control*. 21(5):786–789.

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

Waterhouse, A.L. 2002. Wine phenolics. *Annals of the New York Academy of Sciences*.

Wheelock, C. 2002. Multivariate Data Analysis and Modelling. *Umetrics*.

White, F.E. 1991. *Data fusion lexicon*. San Diego, CA.

Yu, P., Low, M.Y. & Zhou, W. 2018. Design of experiments and regression modelling in food flavour and sensory analysis: A review. *Trends in Food Science and Technology*. 71:202–215.

Chapter 3

Research results

A multivariate approach to evaluating the chemical and sensorial evolution of South African Sauvignon Blanc and Chenin Blanc wines under different bottle storage conditions

This manuscript was published in the peer-reviewed journal of **Food Research International**[2]

[2] Mafata, M., Brand, J., Panzeri, V., Kidd, M., and Buica, A., 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. doi:10.1016/j.foodres.2019.108515

Chapter 3: A multivariate approach to evaluating the chemical and sensorial evolution of South African Sauvignon Blanc and Chenin Blanc wines under different bottle storage conditions

Abstract

Volatile compound composition contributes to the aroma profile of wine and is susceptible to change due to oxidation which may occur during storage and transportation, especially at high temperatures. Changes in sensory attributes may also occur, altering the sensory profile of wine. Classical univariate analysis only looks at the deviations for one factor at a time and may overlook the overall effect of treatments. In this study, changes in South African Sauvignon Blanc and Chenin Blanc wine sensory profile, volatile and antioxidant-related parameters resulting from storage under different temperatures (room temperature, 15 °C and 25 °C) and durations (0, 3 and 9 months) were investigated using a multivariate approach. Bottled, unwooded wines of both cultivars from six wineries were used. As expected, the chemical evolution of the wines was characterised by increases in absorbance at 420 nm (browning), colour density and hue with prolonged storage at high temperatures. To be able to compare the evolution of the sample sets regardless of the initial (T0/control) wine profile and composition, multivariate regression analysis in the form of regression vector (RV) coefficients were used to assess the correlations in the sensory and chemical changes relative to the control in each set. Using Pivot© Profile for the first time in this type of stability assessment and applying a new algorithm for data handling in addition to the classical one, this study showed that prolonged exposure to higher temperatures resulted in the change from fruity to toasted aroma attributes**.**

3.1 Introduction

Wine matrix can be easily susceptible to change due to several factors resulting from influences originating from viticultural practices to storage and transportation of finished wines. The prescribed conditions of storage for white wines is to be chilled and refrigerated. Studies on New Zealand Sauvignon Blanc have shown that there is a significant change in the sensory attributes as well as an evolution in the colour and volatile compound composition with sub-optimal storage conditions (Herbst-Johnstone, Nicolau & Kilmartin, 2011; Makhotkina, Pineau & Kilmartin, 2012).

The non-volatile matrix can be affected by improper storage conditions, which may lead to redox reactions that affect antioxidant related compounds such as phenolics (e.g. phenolic acids, hydroxycinnamic acids and flavanols), resulting in changes in colour causing it to brown (Waterhouse, 2002). These reactions are exacerbated by oxygenation, high temperatures, and temperature fluctuations (Waterhouse, 2002). The visual differences between samples can be calculated based on their UV-Vis absorption. CIELab parameters are an approximation of the visual perception of colour through the human eye (OIV, 2006; Pérez-Caballero, Ayala, Federico Echávarri, *et al.*, 2003). The colour can be approximated using the parameters a* (a*>0 red; a*<0 green) and b* (b*>0 yellow, b*<0 blue) and the clarity of the wine can be approximated using the L* parameter (L* = 0 black; L* = 100 colourless).

Varietal thiols, methoxypyrazines, and major volatile compounds (organic esters, acids, acetates and higher alcohols) contribute to fruity, floral, and herbaceous/vegetative aroma attributes. Oxygen intake of wine during bottling has been linked to decreases in varietal thiols throughout storage and consequently in fruity aromas (Coetzee, Van Wyngaard, Šuklje, *et al.*, 2016). Sub-optimal storage temperatures have also been shown to lead to decreased levels of volatile compounds in white wines, which lead to decreases in floral and fruity aromas as well as increases in ripe and toasted aromas (Pérez-Coello, González-Viñas, García-Romero, *et al.*, 2003).

The evolution of wine throughout storage can be evaluated using different sensory methods. Oenological studies commonly use descriptive and/or quantitative methods such as DA (Herbst-Johnstone *et al.*, 2011) and rating to evaluate wines (Pérez-Coello *et al.*, 2003). These methods profile and/or quantify perceived attributes for each individual wine sample. When it comes to studies on the sensorial evolution of wine, methods that comparatively assess wines against one another should be better suited for the task. Methods such as polarized sensory positioning (PSP), projective mapping (PM), sorting, and triangle test compare samples to one another in either a directed or undirected manner (Valentin, Chollet, Lelièvre, *et al.*, 2012). Using comparative rapid profiling methods such as sorting can elucidate the overall effect of treatments in an efficient manner without the need to individually profile wines. Several rapid profiling methods have been investigated against their effectiveness relative to DA (Lelièvre-Desmas, Valentin & Chollet, 2017). The latest method, Pivot©Profile (PP), comparatively assesses the sensory attributes of wine samples relative to a reference sample (the pivot). This frequency-based method uses free description to profile wines, with each sample assessed one at a time against the pivot. This results in positive and negative frequencies which are translated to positive cumulative frequencies only (Thuillier, Valentin, Marchal, *et al.*, 2015). Correspondence analysis is performed on the data to produce a sensory map or heatmaps to obtain a direct comparison of the sample to the pivot. The potential of this method to profile wines in studies related to their evolution is very promising.

In order to assess storage effects, the classical statistical approach has often taken a more targeted experimental design. Using several repeats of targeted chemical measurements, with univariate statistical approaches such as ANOVA and/or multivariate approach such as MANOVA being used to handle the data (Granato, de Araújo Calado & Jarvis, 2014; Murray, Delahunty &

Baxter, 2001). This approach comes from the need to measure variations in natural systems, hence biological repeats are used to factor the deviations. In the case of oenological experiments, this often results in a focus on only one cultivar and/or a few geographical samples in order to accommodate biological repeats and vintages.

Multivariate analysis has previously been used to investigate several contributing factors to variation in an experimental set-up for issues such as cultivar discrimination and for authentication (Borràs, Ferré, Boqué, *et al.*, 2015; Silvestri, Elia, Bertelli, *et al.*, 2014). Soft, unsupervised multivariate modelling methods such as principal component analysis (PCA) and cluster analysis have often been used for differentiation between treatments but still relied on univariate analysis for tracking evolution (Ugliano, Kwiatkowski, Vidal, *et al.*, 2011). Recently, multivariate regression analysis such as partial least squares (PLS) have gained popularity in modelling the evolution of wine chemical properties. OPLS (orthogonal PLS) is able to give additional information regarding class discrimination when taking into account multiple factors such as cultivar, vintage, and geographical location (Hopfer, Ebeler & Heymann, 2012).

Unlike the targeted univariate approach, multivariate regression analysis can assess trends and/or groupings among treatments taking into account multiple factors as well as multiple variables whilst assessing the inter-relatedness of the relationships between sample groups for all given variables. It can also calculate the contribution the measured variables have to the overall effect in a given experiment. Multivariate approaches do not make assumptions based on the deviations between discreet samples and/or variables and can thus reveal subtle, inherent relationships concerning the behaviour of wine.

Using South African Chenin Blanc and Sauvignon Blanc wines stored at different temperatures for different periods, this study aimed to show that the evolution of wine aroma attributes, volatile, and antioxidant compounds can be modelled using multivariate regression analysis.

3.2 Materials and methods

3.2.1 Wines and treatment

Unwooded Chenin Blanc and Sauvignon Blanc wines were sourced from six wineries in 2016. The wines were stored at room temperature (RT), 15°C and 25°C for 3 months and 9 months (T3 and T9), after which they were transferred into a -4°C cooling room until analysis. The control for each winery (T0) was stored at -4°C until analysis. Therefore, each experimental set consisted of seven samples (T0, T3/15, T3/25, T3/RT, T9/15, T9/25, T9/RT) per winery (AVN, CDB, DTK, FRV, KZC, PDB) for each cultivar (CB and SB), twelve sets in total.

3.2.2 Sensory evaluation

Pivot©Profile (PP) was performed in August 2017, according to the method by Thuillier, *et al.* (2015). A panel of 15 expert judges was used. The cultivars were tested separately. Three repeats were tasted in separate sessions, on a different day. One session consisted of three flights of samples belonging to different wineries. Each flight was an experimental set and additionally included the control (T0) as a blind duplicate. The evaluation was done against the respective T0 as the pivot. The panellists took a 10-minute break between flights. The samples were randomised across judges and were presented according to William Latin Square design. Samples were coded with unique three-digit numbers. Different codes were assigned for all flights including the repeats. Judges were instructed as shown in Supplementary Figure 3.1.

3.2.3 Chemical analysis

3.2.3.1 Oenological parameters

The pH, titratable acidity (TA), total (TSO_2) and free (FSO_2) sulphur dioxide were measured on a Metrohm 862 compact titrosampler (Herisau, Switzerland) using chemicals (sodium hydroxide (NaOH), potassium iodide/ potassium iodate (KI/KIO_3) and sodium thiosulfate ($Na_2S_2O_3$)) purchased from Cameron chemical consultants (Cape Town, South Africa).

3.2.3.2 Thiol analysis

Thiol analysis was performed according to the method by Mafata *et al.* (2018). The following compounds were measured (followed by abbreviations and codes): 3-mercapto-1-hexanol (3MH, C34), 3-mercaptohexyl acetate (3MHA, C35), 4-mercapto-4-methylpentan-2-one (4MMP, C36). The method is based on the derivatization of the thiols with DTDP (4,4′-Dithiodipyridine), followed by sample clean-up by SPE and injection. Quantitative analysis was performed on a Waters Acquity UPC^2 using a Waters Viridis BEH 2EP Column (130 Å, 1.7 µm, 3 mm X 100 mm, 1/pkg) and quantitative mass spectrometric detection was carried out using a Xevo TQ-S triple quadrupole mass spectrometer (Waters, Milford, USA). Data collection and analysis were performed using MassLynx 4.1 (Waters Corporation).

3.2.3.3 Glutathione

Analysis of glutathione (GSH) was performed according to the method by Kritzinger *et al.* (Kritzinger, Stander & Du Toit, 2013). Direct injection of the samples was done on a Waters Acquity UPLC fitted to a Waters Xevo triple-quadrupole mass detection in positive mode (Milford, MA, USA).

3.2.3.4 Major volatiles

The determination of 32 volatile compounds was performed according to Louw (2009). The compounds measured were (followed by codes): ethyl_acetate (C1), methanol (C2), ethyl-2-methyl-propanoate (C3), ethyl_butyrate (C4), propanol (C5), isobutanol (C6), isoamyl_acetate (C7), butanol (C8), isoamyl_alcohol (C9), ethyl_hexanoate (C10), pentanol (C11), hexyl_acetate (C12), acetoin (C13), 3-methyl-1-pentanol (C14), ethyl_lactate (C15), hexanol (C16), 3-ethoxy-1-propanol (C17), ethyl_caprylate (C18), acetic_acid (C19), ethyl-3-hydroxybutanoate (C20), propionic_acid (C21), isobutyric_acid (C22), butyric_acid (C23), ethyl_caprate (C24), isovaleric_acid (C25), diethyl_succinate (C26), valeric_acid (C27), ethyl_phenethylacetate (C28), 2-phenylacetate (C29), hexanoic_acid (C30), 2-phenylethanol (C31), octanoic_acid (C32), decanoic_acid (C33). Wine samples were extracted with diethyl ether and the organic layer was dried over anhydrous sodium sulphate prior to analysis by GC-FID (HP 6890, Hewlett Packard, Palo Alto, California, United States).

3.2.3.5 Measurements of colour

Spectrophotometric measurements were performed in triplicate from 280 nm to 780 nm on a Thermo Scientific Multiskan GO 1510-02586 microplate spectrophotometer. Colour intensity (CI, A_{520} + A_{420}), colour hue (CH, A_{520} / A_{420}), total phenolics (A_{280}), hydroxycinnamic acids (A_{320}) and browning (A_{420}) were determined. Calculation of chromatic characteristics was done according to the method by Pérez-Caballero *et al*., (2003) which originated from the Commission Internationale de l'Eclairage (CIELab) method (OIV, 2006) and optimized for white wines.

3.2.4 Statistical analysis

Multivariate analysis was performed separately for each winery and each cultivar, in sets of seven samples. Correspondence Analysis (CA) was performed on sensory data and Principal Component Analysis (PCA) on the scaled and centred chemical data. The generated scores and loadings were submitted to additional statistical analysis. In order to assess the configurational similarity between sample sets, pair-wise regression vector (RV) coefficients were calculated separately from the first two dimensions of the CA results of the sensory data (scores) and the first two dimensions of the PCA results of the chemistry data (scores), based on the generalized Pearson correlation coefficient. Three dimensional representations of Multidimensional Scaling (MDS) plots were generated based on the RV coefficients. Statistical calculations were performed using Statistica™ 13 (TIBCO, Dell software, Inc., Texas, United States) and R version 3.4.0 (www.R-project.org) using personally tailored "R" scripts.

3.3 Results and discussion

Since the sensory evaluation was performed separately for each winery and each cultivar, each experimental set has gone through statistical analysis separately. The same strategy was used for the chemical analysis results. Hence, for each cultivar, six CA and six PCA were performed. The results of this unsupervised statistical approach contributed to the descriptive part of the work. For the initial step of the evaluation, in addition to the CA, heatmaps were generated from the sensory data. In the next stage, in order to quantify the similarity of the patterns between the CA/PCA plots, multivariate regression analysis was performed separately for each cultivar and pair-wise RV coefficients generated (Supplementary Tables 3.1 to 3.5). This constituted the modelling part of the work, in which the patterns of evolution of the sample sets were the focus. The discussion of the results follows the same steps.

3.3.1 Sensory evaluation

As highlighted in the Introduction, PP data can be analysed using two different approaches: translating all frequencies in the contingency table to have only positive values (Thuillier *et al.*, 2015), on which CA can be performed, or leaving the values as they are in which case heatmaps can be used to visualise and analyse the data (Brand, 2019). The data analysis began with the capture of 180 forms for SB and 180 for CB. 240 Attributes were generated with little redundancies. Using semantic grouping led to 170 attributes from which cumulative frequencies were calculated for CB and SB separately. Attributes containing less than 95% “zero” citation were selected, which resulted in 29 attributes for CB and 33 for SB and used for the CA plots (Supplementary Figure 3.2) and heatmaps (Figure 3.1). Heatmaps are descriptive and intuitive (Figure 3.1). Horizontally, the profile of a sample is displayed relative to the rest of the samples; vertically, each attribute’s relative intensity is presented for each sample. Additionally, the dendrogram shows how the samples are related to each other using cluster analysis.

The results showed that judges perceived the aroma of the control (T0) to be consistently different from the treatments in all wineries and for both cultivars. The controls were generally (as expected) most different from the extreme treatment (T9/25). The control samples, described mostly with ‘fruity’ and ‘floral’ attributes, were different from the T3 samples (described mostly with green attributes) and the T9 samples (described mostly as ‘toasted’, ‘oaky’ and ‘woody’ attributes). Regardless of the similarity between the individual attributes used for the wineries, the dendrograms of both SB and CB indicate that the frequency of citation for samples stored at lower temperatures for shorter periods of time are more similar to attributes used to describe the control and are different from attributes used for samples stored at higher temperatures for prolonged periods of time (Figure 3.1). The evolution of ‘floral’ and ‘fruity’ notes in the controls were more subtle in SB than CB wines. CB control wines, for most of the wineries, experienced a sharper decline in floral and fresh notes over time and at increased temperatures.

SB control wines were more ‘fruity’, ‘floral’ and ‘tropical’ than wines stored at elevated temperatures for longer which were more ‘toasty’, ‘oaky’ and ‘spicy’. This is similar to the results on New Zealand SB found by Makhotkina, et al. (2012). KZC SB sample set was an exception as the control had particularly higher ‘pineapple’, ‘green’ and less ‘fruity’ notes compared to the other wineries. ‘Fresh’ remained relatively unchanged, with only some wineries experiencing a slight decrease with elevated temperature and prolonged exposure.

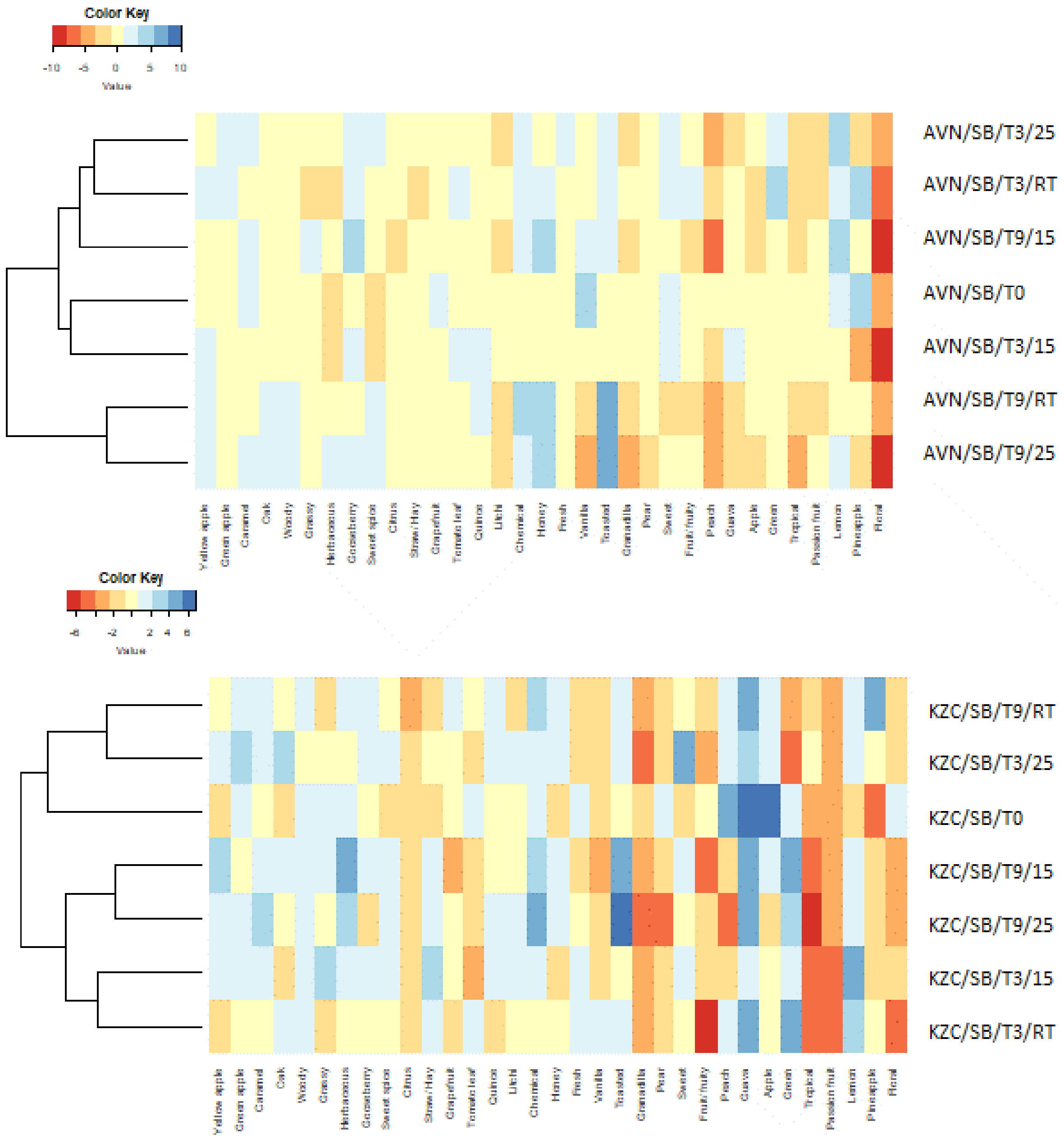


Figure 3.1: Heatmaps of the results from the Pivot©Profile of SB wines from AVN and KZC wineries. Wines were stored under the conditions: T0 = control/ pivot at -4°C; T3/RT= 3 months at room temperature; T3/15 = 3 months at 15°C; T3/25 = 3 months at 25°C; T9/RT= 9 months at room temperature; T9/15 = 9 months at 15°C; T9/25 = 9 months at 25°C.

Even though the use of the CA means that the frequencies of citation have to be normalized to avoid the use of negative values, there are some advantages. For example, CA shows the value of the inertia for the dimensions considered for the analysis. The first two dimensions explained between 60 and 69 for SB and between 66 and 77 for CB. It was evident from the CA score plots that there was a gradual change from the T0 to T3 to T9 samples (Supplementary Figure 3.2). Similar to the heatmap approach, CA also includes associated dendrograms that show how the samples within a set are related to each other. The interpretation of the CA biplot (samples and attributes) though is more difficult than in the case of a heatmap, where each sample is presented with its own attribute profile. Overall, the results in this study are comparable to the findings on commercial New Zealand SB wines (Makhotkina *et al.*, 2012) and on Spanish white wines (Pérez-Coello *et al.*, 2003). Both studies found significant increases in buttery, ripe and spicy attributes and decreases in fruity, fresh and floral attributes with higher storage temperatures.

Even though there is a wealth of information that can be extracted, the same issue arises for both CA and heatmaps when working with more than one sample set. Careful inspection can lead to observing trends between sets looking at the sample configuration and attributes. However, the sets are dealt with separately so one can only notice trends and exceptions, not statistically measure similarities between sets. In this case, an additional statistical analysis such as pair-wise regression vector (RV) coefficients were calculated between each of the six sample sets for each cultivar, followed by MDS representation. In principle, RV coefficients can be calculated for both scores (samples) and loadings (attributes) from the CA results, as long as the new variables (samples and attributes) are the same between the sets. However, since the wines were not described using the same attributes, this step could only be applied to the scores (samples).

For most of the wineries similar patterns of evolution were observed, with correlations greater than 50% (RV≥0.50 at p ≤ 0.05), with the exception of a few wineries. CB data sets had generally higher RV values (0.71±0.14) than SB (0.64±0.14). For both cultivars, one of the wineries' (KZC) evolution pattern generally differed from the rest, as reflected in the RV coefficients (Supplementary Table 3.1). For KZC CB, RV coefficients were between 0.46 and 0.55, while for KZC SB between 0.36 and 0.66.

The addition of RV coefficient results to the descriptive heatmaps and the CA plots made it possible to see that for each cultivar and regardless of the initial (T0) wine profiles, the storage conditions had similar effects on the evolution of the wines with the exception of KZC CB and SB. This can be seen in the MDS plot for both CB and SB wines (Figure 3.2), showing KZC further placed from the rest of the wineries.

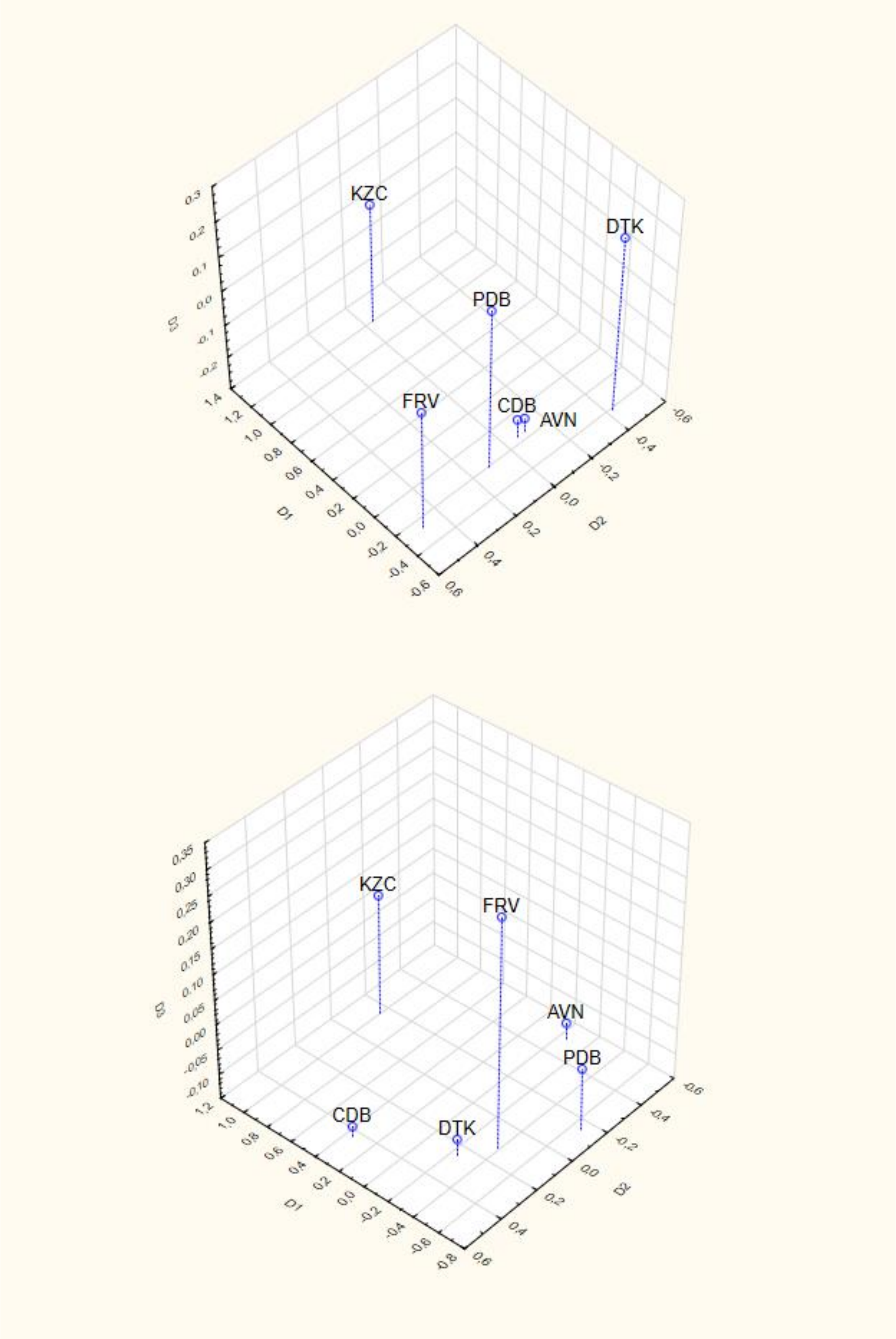


Figure 3.2: MDS from the RV coefficients of the six wineries, generated from the Pivot©Profile of CB (top) and SB (bottom) stored under different conditions.

3.3.2 Chemical evaluation

3.3.2.1 Volatile compounds

The variables used in the PCA featured 34 volatile compounds comprised of varietal thiols (3-MH, 3-MHA and 4-MMP), esters, organic acids, and higher alcohols (Supplementary Tables 3.6 to 3.11). The first two PCs contained 73 to 80% for the explained variance for SB and 72 to 83% for CB sample sets. Similar to the sensory results, T0 samples were consistently different from the other treatments and a gradual trend was observed from T0 to T3 and T9 samples.

Among the volatile compounds measured, the varietal thiols concentrations were associated with storage time/temperature combination for both SB and CB (Figure 3.3). The controls had higher 3-MHA concentrations compared to the extreme treatment (T9/25), which has also been previously shown in New Zealand Sauvignon Blanc (Herbst-Johnstone *et al.*, 2011). Samples stored at high temperatures for longer (T9/25 as the extreme) had higher 3-MH and 4-MMP which in previous studies was found not to have changed throughout storage (Makhotkina *et al.*, 2012).

Major volatile composition was different between wineries and cultivars, in both levels and profile. A study on the evolution of Spanish white wines (Airén, Viura, and Macabeo) over a four year period did not show changes in volatile acids at the end of the first year of storage at sub-optimal conditions (Pérez-Coello *et al.*, 2003) which may explain the results found in the current study, since the storage was only over nine months.

The RV coefficients calculated from the PCA scores for both SB (0.49±0.15; range 0.25 to 0.83) and CB (0.49±0.12; 0.24 to 0.69) indicated that they had different patterns in evolution (Supplementary Table 3.2). This may be an indication that the treatment did not affect the volatile composition of the wines similarly. Additionally, the list of volatile compounds measured might not have been comprehensive enough to model the effect of the treatment, in which case an untargeted approach may be more appropriate.

The RV coefficients for the loadings for both CB (0.36±0.17; range to 0.08 to 0.66) and SB (0.33±0.16; from 0.07 to 0.65) were very low (Supplementary Table 3.3). This may suggest another reason why the evolution was different, namely that the volatile composition of the controls was initially so different between the sets that their evolution through storage also differed. The MDS of the scores and the loadings for both the SB and CB show most of the six wineries sitting far apart from one another further illustrating differences in evolution (Figure 3.4).

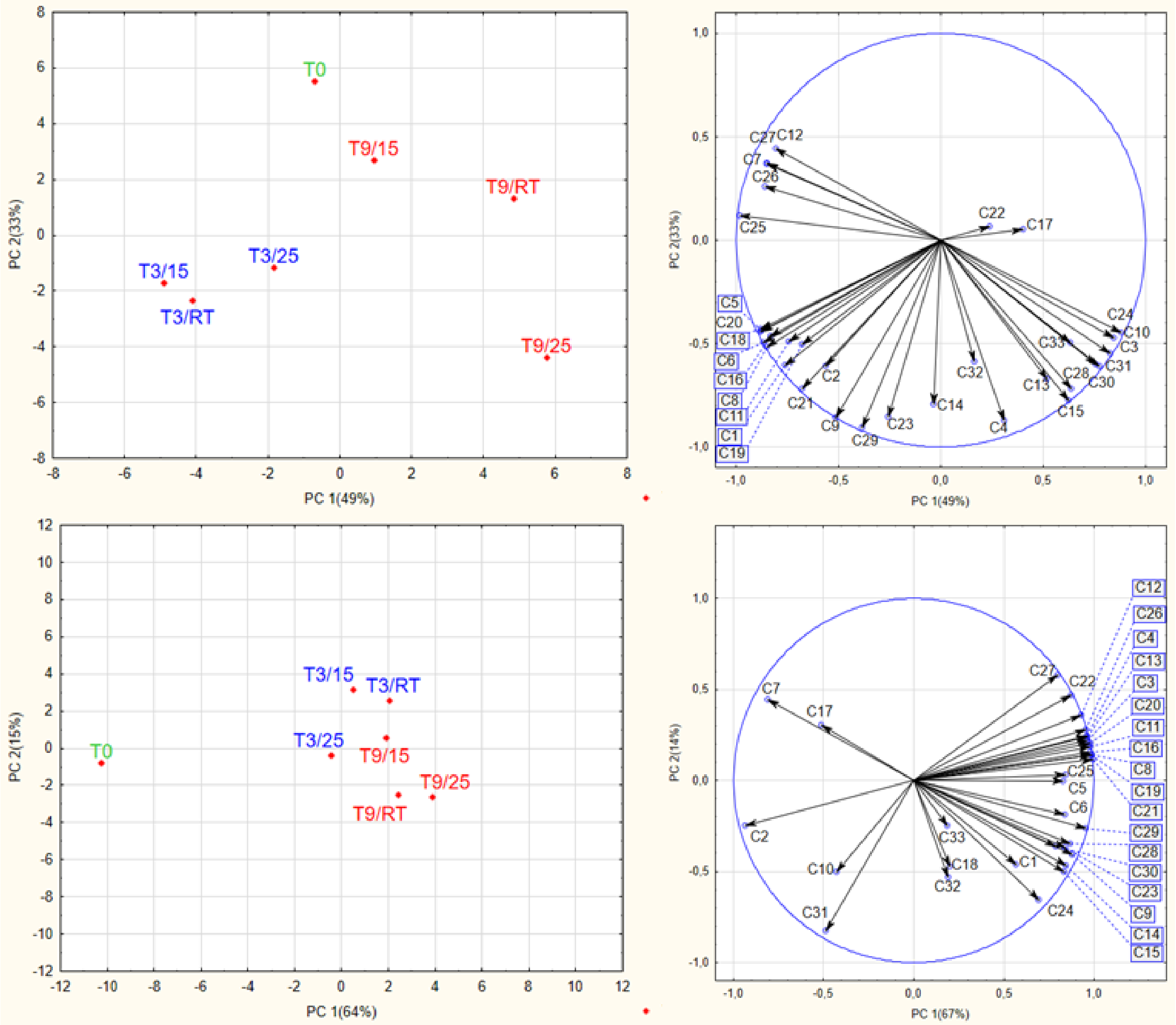


Figure 3.3: PCA (scores and loadings) based on the *v*olatile compound results for CB (top graphs) and SB (bottom graphs) wines from AVN winery, stored under different conditions.

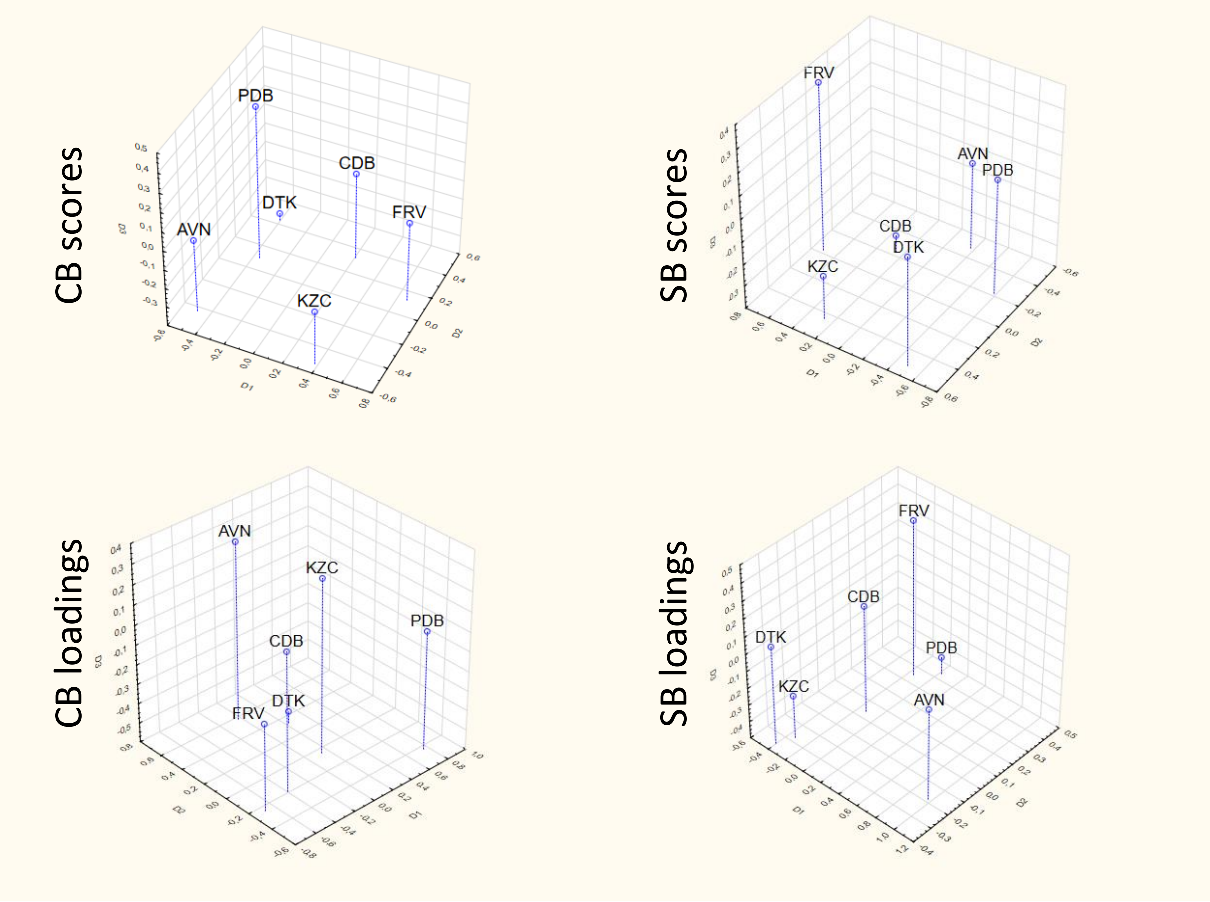


Figure 3.4: MDS from the RV coefficients generated from scores and loadings of volatile compounds results of CB (left) and SB (right) wines of the six wineries stored under different conditions.

3.3.2.2 Antioxidant-related parameters

Variables used for the PCA (Figure 3.5) of antioxidant-related compounds/parameters consisted of 11 measurements including reduced glutathione (GSH), total and free SO_2 content (TSO_2 and FSO_2), spectrophotometric analysis at discrete absorption wavelengths (A_{520}, A_{420}, A_{320}, A_{280}, colour density and colour hue), as well as CIELab measurements of colour (L*, a*, b*, Cab* and hab*)(Supplementary Tables 3.12 and 3.13). The PCAs (Figure 3.5) showed a gradual distribution of samples according to duration of storage (from T0 to T3 and T9) with the first two PCs explaining 75 to 88% of the variation for SB and 80 to 94% for CB.

Wines stored at higher temperatures for longer showed increased absorbance at 420 nm and higher b* (yellow). These wines also generally had higher UV-Vis absorption (A_{280}, A_{320} and A_{520}) and colour density. The clarity (L*) of these samples was lower compared to the control. Oxidation of samples stored at higher temperatures for longer was indicated by the decrease in reduced glutathione, total and free SO_2. The results in this study are similar to that on New Zealand Sauvignon Blanc wines found a decrease in GSH and FSO_2 within a few months after bottling and increases in absorbance at 420 nm after 7-months after bottling (Herbst-Johnstone *et al.*, 2011).

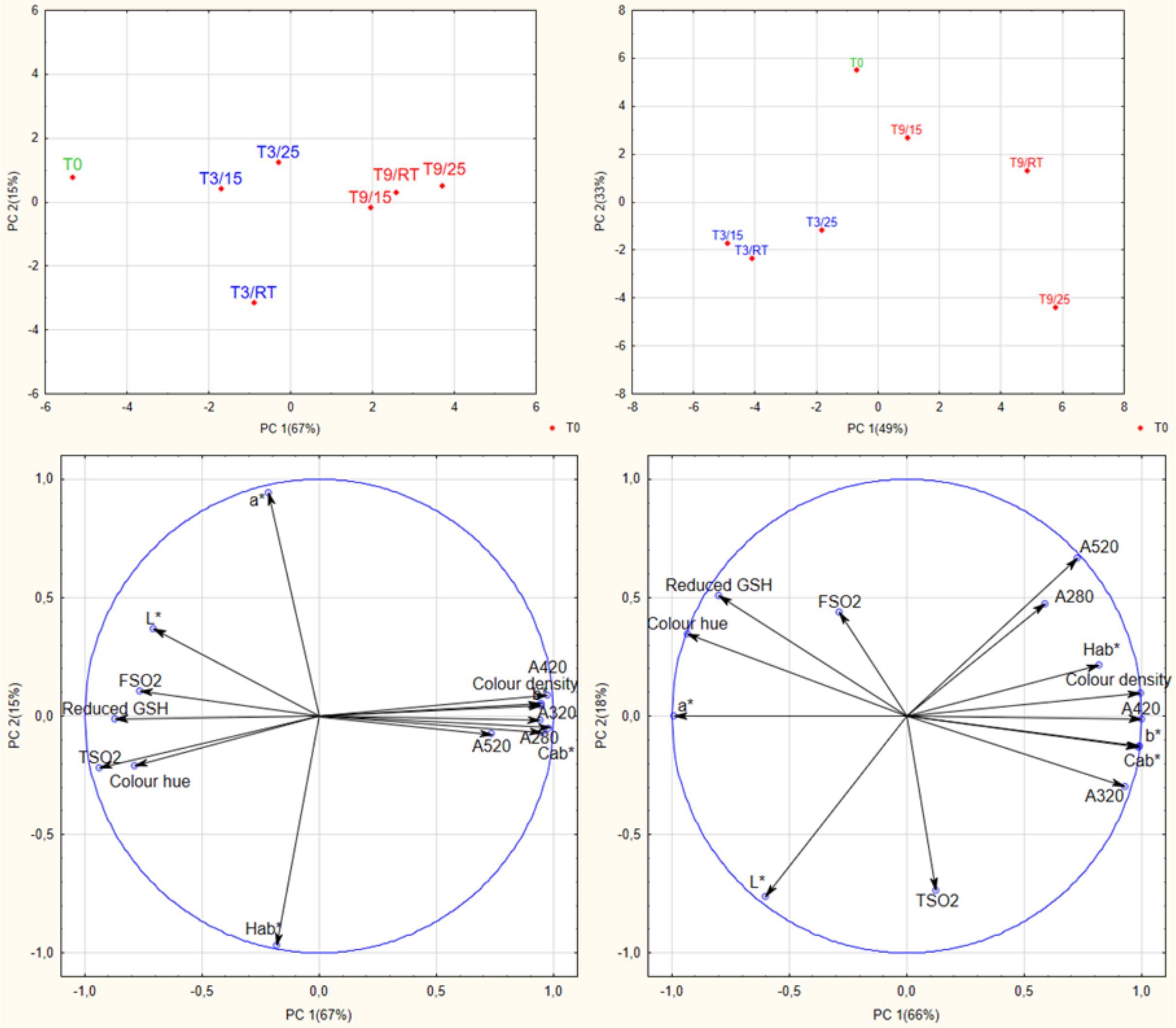


Figure 3.5: PCA (scores- top and loadings-bottom) based on antioxidant-related parameters of SB (left) and CB (right) wines for AVN winery stored under different conditions.

The similarity in the pattern of evolution of the sample sets (scores) throughout storage was generally observed in both cultivars. SB had an average RV coefficient of 0.75±0.13 (from 0.52 to 0.91) excluding CDB with an RV ranging from 0.08 to 0.57 and an average RV of 0.2 (Supplementary Table 3.4). CB had an average RV coefficient of 0.90±0.04 (from 0.81 to 0.97), indicating a more consistent response to the storage conditions in CB compared to SB. This is similar to the sensory response of CB wines discussed in section 3.3.1.

The RV coefficients for the antioxidant-related parameters (loadings, Supplementary Table 3.5) was higher for CB (0.78±0.14; from 0.61 to 0.93) with FRV and PDB pair having the lowest correlations with an RV of 0.43. In contrast, SB wines had low RV coefficients (0.31±0.16; from 0.03 to 0.63). The possibility that the differences in chemical matrix of the control wines resulted in very different patterns of change throughout time and temperature is still valid. The CB MDS (Figure 3.6) showed very close relatedness in the pattern of evolution of the samples with that of the loadings being captured within the first two dimensions. The opposite was true for SB seeing as the CDB winery had such very low RV coefficients to the rest of the wineries (Figure 3.6).

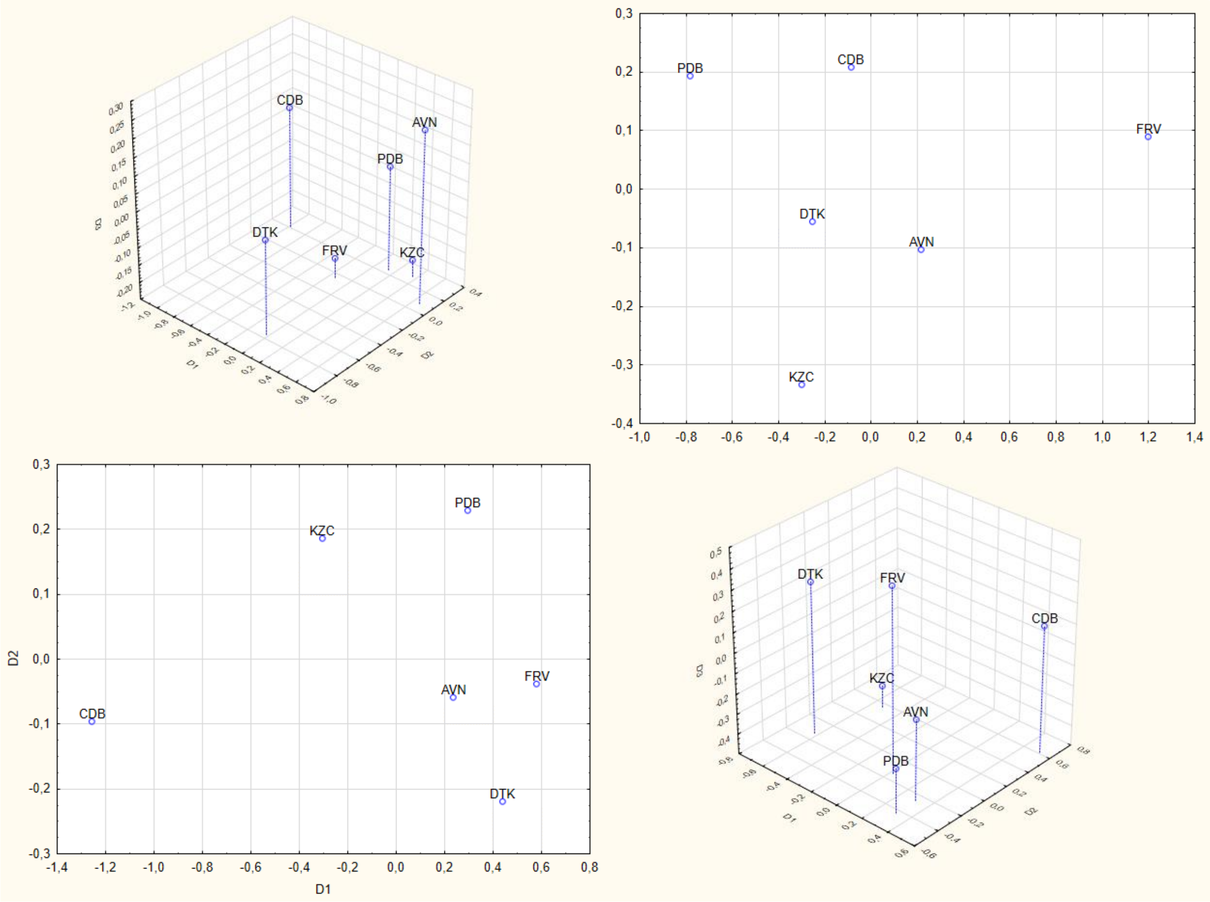


Figure 3.6: MDS from the RV coefficients generated from scores (left) and loadings (right) of antioxidant-related results of CB (top) and SB (bottom) wines for all six wineries stored under different conditions.

3.4 Conclusion

Using a multivariate regression approach, the evolution of the sensory perception of aroma, as well as the volatile and antioxidant-related composition of Sauvignon Blanc and Chenin Blanc wines under different storage conditions and durations was investigated. The wines investigated showed an aroma evolution from ‘fruity’ and ‘herbaceous’ for Sauvignon Blanc and from ‘fruity’ and ‘tropical’ for Chenin Blanc to ‘toasted’, ‘oak’, and ‘honey’. RV coefficients for the scores (samples) showed a significant correlation in the observed evolution among the six wineries. CB wines had higher RV coefficients, indicating that the evolution was more consistent across wineries compared to SB.

The volatile compound analysis showed very little correlation between the patterns of evolution across wineries as measured by RV coefficients for both scores (samples) and loadings (compounds). This indicates a need to either increase the number and diversity of compounds measured or to perhaps take an untargeted analysis approach.

The response of the antioxidant-related parameters to the treatment was similar to that observed for the sensory evaluation, with high RV coefficients between wineries. Samples stored at higher temperatures for longer periods correlated with higher UV-Vis absorbance, colour density as well as higher b* (yellow) values and lower clarity in terms of L* index. Chenin Blanc control wines had very similar aroma and antioxidant-related profile which may have resulted in their uniform response to the treatments (high RV coefficients for both scores and loadings) as compared to Sauvignon Blanc. The large differences in chemical makeup of the wines may be as a result of the grapes themselves (clonal differences, ripening status at harvest, climatic conditions during growth) or winemaking practices (varied between the cellars).

Although the multivariate analysis used in this study was able to elucidate the evolution pattern of the wines, it could not measure statistical significance in the evolution. The issue of statistical significance could be addressed by including biological repeats in the design. The would be the expansion of the sample sets to the level that the chemistry becomes extremely laborious and the sensory evaluation impossible.

References

Brand, J. (2019). Rapid sensory profiling methods for wine: workflow optimization for research and industry applications. PhD dissertation, Stellenbosch University.

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

Coetzee, C., Van Wyngaard, E., Šuklje, K., Silva Ferreira, A.C. & Du Toit, W.J. 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):7979–7993.

Granato, D., de Araújo Calado, V.M. & Jarvis, B. 2014. Observations on the use of statistical methods in Food Science and Technology. *Food Research International*. 55:137–149.

Herbst-Johnstone, M., Nicolau, L. & Kilmartin, P.A. 2011. Stability of varietal thiols in commercial sauvignon blanc wines. *American Journal of Enology and Viticulture*. 62(4):495–502.

Hopfer, H., Ebeler, S.E. & Heymann, H. 2012. The Combined Effects of Storage Temperature and Packaging Type on the Sensory and Chemical Properties of Chardonnay. *Journal of Agricultural and Food Chemistry*. 60:10743–10754.

Kritzinger, E.C., Stander, M.A. & Du Toit, W.J. 2013. Assessment of glutathione levels in model solution and grape ferments supplemented with glutathione-enriched inactive dry yeast preparations using a novel UPLC-MS/MS method. *Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment*. 30(1):80–92.

Lelièvre-Desmas, M., Valentin, D. & Chollet, S. 2017. Pivot profile method: What is the influence of the pivot and product space? *Food Quality and Preference*. 61(May):6–14.

Louw, L., Roux, K., Tredoux, A., Tomic, O., Naes, T., El`ene, H.´, El`ene, E., Nieuwoudt, H., et al. 2009. Characterization of Selected South African Young Cultivar Wines Using FTMIR Spectroscopy, Gas Chromatography, and Multivariate Data Analysis. *Journal of Agricultural and Food Chemistry*. 57:2623–2632.

Mafata, M., Stander, M., Thomachot, B., Buica, A., Mafata, M., Stander, M.A., Thomachot, B. & Buica, A. 2018. Measuring Thiols in Single Cultivar South African Red Wines Using 4,4-Dithiodipyridine (DTDP) Derivatization and Ultraperformance Convergence Chromatography-Tandem Mass Spectrometry. *Foods*. 7(9):138.

Makhotkina, O., Pineau, B. & Kilmartin, P.A. 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):91–99.

Murray, J.., Delahunty, C.. & Baxter, I.. 2001.

OIV. 2006. Determination of chromatic characteristics according to CIELab. *Compendium of International Analysis of Methods*. (Chromatic Characteristics):1–16. [Online], Available: http://www.oiv.int/oiv/files/6 - Domaines scientifiques/6 - 4 Methodes d analyses/6-4-1/EN/OIV-MA-AS2-11.pdf.

Pérez-Caballero, V., Ayala, F., Federico Echávarri, J. & Negueruela, A.I. 2003. *Determination of Chromatic*

Characteristics of Wine-59 Proposal for a New Standard OIV Method for Determination of Chromatic Characteristics of Wine. [Online], Available: http://www.ajevonline.org/content/ajev/54/1/59.full.pdf [2018, September 06].

Pérez-Coello, M.S., González-Viñas, M.A., García-Romero, E., Díaz-Maroto, M.C. & Cabezudo, M.D. 2003. Influence of storage temperature on the volatile compounds of young white wines. *Food Control*. 14(5):301–306.

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

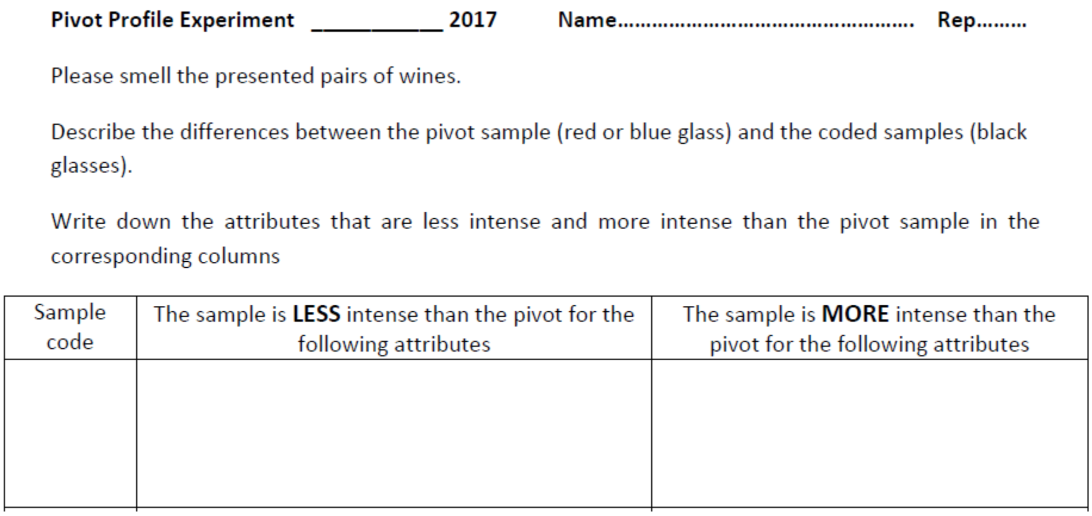
Thuillier, B., Valentin, D., Marchal, R. & Dacremont, C. 2015.

Ugliano, M., Kwiatkowski, M., Vidal, S., Capone, D., Siebert, T., Dieval, J.B., Aagaard, O. & Waters, E.J. 2011. Evolution of 3-mercaptohexanol, hydrogen sulfide, and methyl mercaptan during bottle storage of Sauvignon blanc wines. Effect of glutathione, copper, oxygen exposure, and closure-derived oxygen. *Journal of Agricultural and Food Chemistry*. 59(6):2564–2572.

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

Waterhouse, A.L. 2002. Wine phenolics. *Annals of the New York Academy of Sciences*.

Chapter 3
Supplementary



Pivot Profile Experiment **___________ 2017** **Name..** **Rep.........**

Please smell the presented pairs of wines.

Describe the differences between the pivot sample (red or blue glass) and the coded samples (black glasses).

Write down the attributes that are less intense and more intense than the pivot sample in the corresponding columns

Sample code	The sample is **LESS** intense than the pivot for the following attributes	The sample is **MORE** intense than the pivot for the following attributes

Figure 3.1. Instructions to panellists for the Pivot©Profile of Sauvignon and Chenin Blanc wines stored under different conditions.

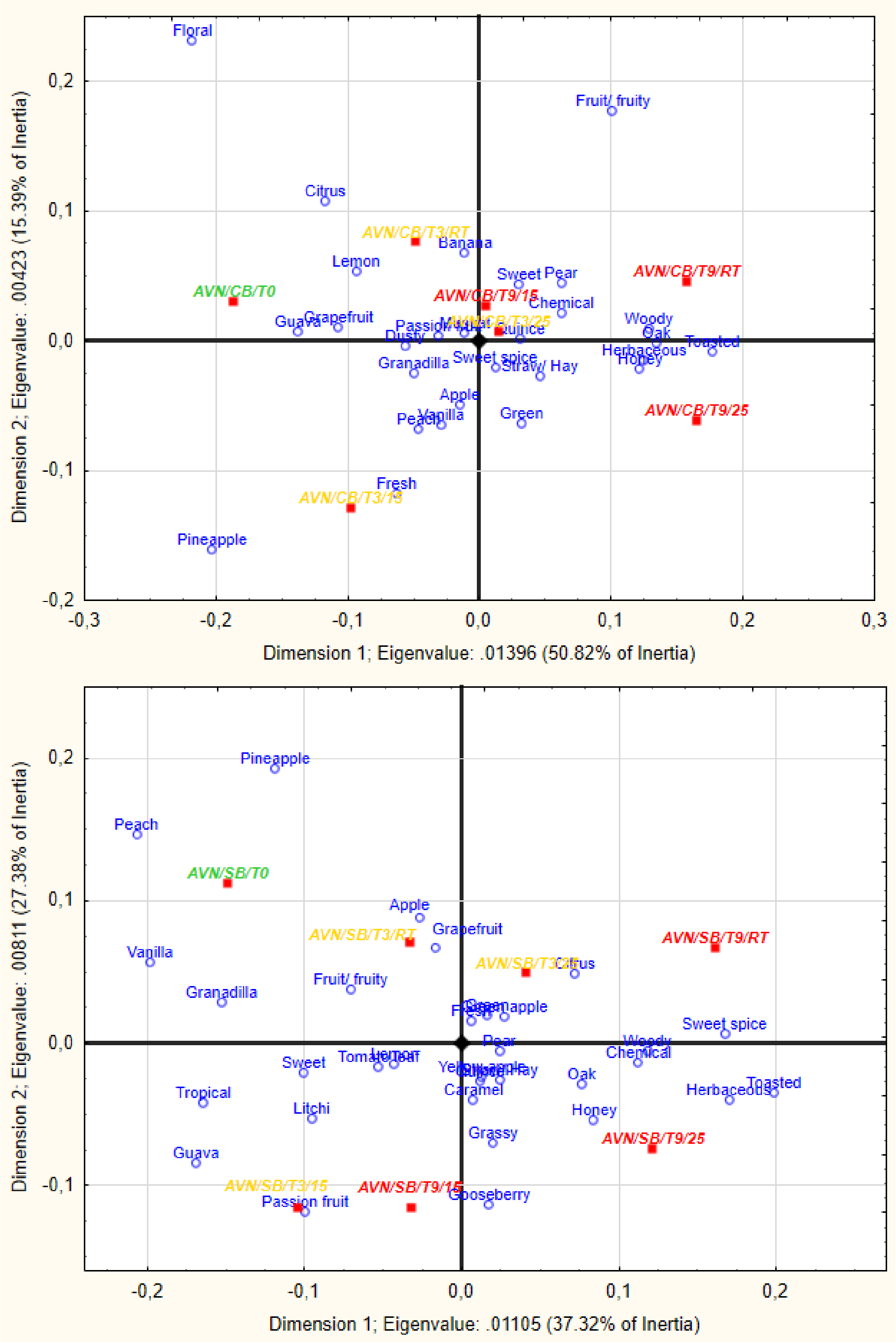


Figure 3.2: CA plots of the results on the Pivot©Profile of the aroma of CB and SB wines from AVN winery, stored under different conditions. T0 = control at -4°C; T3/RT= 3 months at room temperature; T3/15 = 3 months at 15°C; T3/25 = 3 months at 25°C; T9/RT= 9 months at room temperature; T9/15 = 9 months at 15°C; T9/25 = 9 months at 25°C.

Table 3.1: RV coefficients of the scores from the Pivot©Profile (PP) on the aroma of CB and SB wines stored at different conditions.

Chenin Blanc				Sauvignon Blanc			
plot 1	plot 2	RV coefficient	p-value	plot 1	plot 2	RV coefficient	p-value
KZC	PDB	0.54	0.07	KZC	PDB	0.34	0.35
KZC	AVN	0.55	0.06	KZC	AVN	0.58	0.05
KZC	CDB	0.55	0.06	KZC	CDB	0.66	0.02
KZC	DTK	0.46	0.14	KZC	DTK	0.48	0.12
KZC	FRV	0.50	0.09	KZC	FRV	0.46	0.16
PDB	AVN	0.83	0.00	PDB	AVN	0.74	0.01
PDB	CDB	0.84	0.01	PDB	CDB	0.56	0.07
PDB	DTK	0.81	0.00	PDB	DTK	0.78	0.01
PDB	FRV	0.86	0.00	PDB	FRV	0.81	0.01
AVN	CDB	0.87	0.00	AVN	CDB	0.60	0.05
AVN	DTK	0.78	0.01	AVN	DTK	0.70	0.01
AVN	FRV	0.77	0.00	AVN	FRV	0.67	0.02
CDB	DTK	0.78	0.01	CDB	DTK	0.77	0.01
CDB	FRV	0.78	0.00	CDB	FRV	0.65	0.03
DTK	FRV	0.70	0.02	DTK	FRV	0.83	0.00
	mean	0.71	0.03		mean	0.64	0.06
	dev	0.14	0.04		dev	0.14	0.09

Figures in red indicate RV coefficients corresponding to low similarity between paired sets (i.e. RV coefficients below 0.5 and/or p values greater than 0.05).

Table 3.2: RV coefficients of the scores of the volatile compounds of SB and CB wines stored at different conditions.

Chenin Blanc				Sauvignon Blanc			
plot 1	plot 2	RV coefficient	p-value	plot 1	plot 2	RV coefficient	p-value
AVN	CDB	0.48	0.15	AVN	CDB	0.51	0.10
AVN	DTK	0.49	0.12	AVN	DTK	0.50	0.09
AVN	FRV	0.24	0.59	AVN	FRV	0.42	0.20
AVN	KZC	0.56	0.06	AVN	KZC	0.64	0.03
AVN	PDB	0.60	0.04	AVN	PDB	0.33	0.30
CDB	DTK	0.52	0.10	CDB	DTK	0.44	0.22
CDB	FRV	0.69	0.02	CDB	FRV	0.83	0.00
CDB	KZC	0.43	0.20	CDB	KZC	0.76	0.01
CDB	PDB	0.52	0.10	CDB	PDB	0.34	0.40
DTK	FRV	0.39	0.24	DTK	FRV	0.39	0.31
DTK	KZC	0.32	0.40	DTK	KZC	0.54	0.07
DTK	PDB	0.56	0.05	DTK	PDB	0.41	0.23
FRV	KZC	0.66	0.02	FRV	KZC	0.41	0.23
FRV	PDB	0.44	0.16	FRV	PDB	0.25	0.57
KZC	PDB	0.40	0.22	KZC	PDB	0.56	0.06
	mean	0.49	0.17		mean	0.49	0.19
	dev	0.12	0.15		dev	0.15	0.16

Figures in red indicate RV coefficients corresponding to low similarity between paired sets (i.e. RV coefficients below 0.5 and/or p values greater than 0.5).

Table 3.3: RV coefficients of the loadings of the volatile compounds of SB and CB wines stored at different conditions.

		Chenin Blanc				Sauvignon Blanc	
plot 1	plot 2	RV coefficient	p-value	plot 1	plot 2	RV coefficient	p-value
AVN	CDB	0.52	0.00	AVN	CDB	0.19	0.01
AVN	DTK	0.24	0.00	AVN	DTK	0.07	0.23
AVN	FRV	0.26	0.00	AVN	FRV	0.15	0.04
AVN	KZC	0.28	0.00	AVN	KZC	0.12	0.07
AVN	PDB	0.66	0.00	AVN	PDB	0.20	0.01
CDB	DTK	0.22	0.01	CDB	DTK	0.48	0.00
CDB	FRV	0.41	0.00	CDB	FRV	0.50	0.00
CDB	KZC	0.44	0.00	CDB	KZC	0.54	0.00
CDB	PDB	0.32	0.00	CDB	PDB	0.45	0.00
DTK	FRV	0.07	0.27	DTK	FRV	0.42	0.00
DTK	KZC	0.39	0.00	DTK	KZC	0.65	0.00
DTK	PDB	0.49	0.00	DTK	PDB	0.35	0.00
FRV	KZC	0.39	0.00	FRV	KZC	0.43	0.00
FRV	PDB	0.08	0.22	FRV	PDB	0.44	0.00
KZC	PDB	0.21	0.01	KZC	PDB	0.43	0.00
	mean	0.33	0.03		mean	0.36	0.02
	dev	0.16	0.0836		dev	0.17	0.06

Figures in red indicate RV coefficients corresponding to low similarity between paired sets (i.e. RV coefficients below 0.5 and/or p values greater than 0.5).

Table 3.4: RV coefficients of the scores of the non-volatile compounds of SB and CB wines stored at different conditions.

Chenin Blanc				Sauvignon Blanc			
plot 1	plot 2	RV coefficient	p-value	plot 1	plot 2	RV coefficient	p-value
AVN	CDB	0.81	0.01	AVN	CDB	0.22	0.51
AVN	DTK	0.88	0.00	AVN	DTK	0.91	0.00
AVN	FRV	0.90	0.00	AVN	FRV	0.81	0.01
AVN	KZC	0.93	0.00	AVN	KZC	0.73	0.01
AVN	PDB	0.95	0.00	AVN	PDB	0.86	0.01
CDB	DTK	0.85	0.00	CDB	DTK	0.13	0.71
CDB	FRV	0.90	0.00	CDB	FRV	0.08	0.88
CDB	KZC	0.84	0.01	CDB	KZC	0.51	0.11
CDB	PDB	0.89	0.00	CDB	PDB	0.17	0.67
DTK	FRV	0.92	0.00	DTK	FRV	0.89	0.00
DTK	KZC	0.88	0.01	DTK	KZC	0.54	0.06
DTK	PDB	0.89	0.00	DTK	PDB	0.74	0.01
FRV	KZC	0.94	0.00	FRV	KZC	0.52	0.08
FRV	PDB	0.95	0.00	FRV	PDB	0.81	0.01
KZC	PDB	0.97	0.00	KZC	PDB	0.71	0.01
	mean	0.90	0.00		mean	0.57	0.21
	dev	0.04	0.0012		dev	0.28	0.30

Figures in red indicate RV coefficients corresponding to low similarity between paired sets (i.e. RV coefficients below 0.5 and/or p values greater than 0.5).

Table 3.5: RV coefficients of the loadings of the non-volatile compounds of SB and CB wines stored at different conditions.

		Chenin Blanc				Sauvignon Blanc	
plot 1	plot 2	RV coefficient	p-value	plot 1	plot 2	RV coefficient	p-value
AVN	CDB	0.89	0.00	AVN	CDB	0.31	0.04
AVN	DTK	0.89	0.00	AVN	DTK	0.26	0.07
AVN	FRV	0.74	0.00	AVN	FRV	0.52	0.00
AVN	KZC	0.85	0.00	AVN	KZC	0.22	0.15
AVN	PDB	0.71	0.00	AVN	PDB	0.63	0.00
CDB	DTK	0.92	0.00	CDB	DTK	0.03	1.00
CDB	FRV	0.72	0.00	CDB	FRV	0.33	0.04
CDB	KZC	0.84	0.00	CDB	KZC	0.16	0.36
CDB	PDB	0.88	0.00	CDB	PDB	0.19	0.18
DTK	FRV	0.62	0.00	DTK	FRV	0.52	0.00
DTK	KZC	0.93	0.00	DTK	KZC	0.42	0.01
DTK	PDB	0.86	0.00	DTK	PDB	0.20	0.14
FRV	KZC	0.61	0.00	FRV	KZC	0.24	0.13
FRV	PDB	0.43	0.01	FRV	PDB	0.42	0.01
KZC	PDB	0.82	0.00	KZC	PDB	0.17	0.25
	mean	0.78	0.00		mean	0.31	0.16
	dev	0.14	0.0035		dev	0.16	0.25

Figures in red indicate RV coefficients corresponding to low similarity between paired sets (i.e. RV coefficients below 0.5 and/or p values greater than 0.5).

Table 3.6: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from AVN winery.

Sample ID	Sauvignon Blanc AVN							Chenin Blanc AVN						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	106.70	121.92	115.63	98.33	119.66	110.99	123.66	82.14	96.05	96.37	91.28	83.63	86.18	85.44
Methanol	344.21	115.16	105.86	82.26	89.64	113.29	103.95	105.06	138.72	145.24	144.53	111.40	135.34	125.06
Ethyl-2-Methyl-Propanoate	0.00	2.19	2.16	2.19	2.24	2.19	2.25	2.10	2.12	2.12	2.13	2.16	2.14	2.19
Ethyl_Butyrate	0.00	0.68	0.66	0.66	0.66	0.64	0.67	0.65	0.66	0.67	0.66	0.66	0.65	0.69
Propanol	45.96	66.96	64.98	49.39	62.28	64.00	64.77	100.82	131.63	130.93	121.77	96.22	109.95	99.85
Isobutanol	27.77	34.98	33.73	28.73	33.67	32.98	35.48	28.99	34.69	34.77	32.66	27.54	29.84	29.10
Isoamyl_Acetate	6.92	6.09	6.22	5.74	5.01	5.39	4.74	6.34	5.86	6.08	5.64	5.07	5.35	4.88
Butanol	0.00	1.11	1.06	0.92	1.09	1.06	1.15	1.28	1.51	1.51	1.43	1.23	1.32	1.29
Isoamyl_Alcohol	208.77	242.87	229.53	220.34	242.55	233.82	253.49	195.33	220.71	222.01	214.01	197.48	198.19	213.18
Ethyl_Hexanoate	1.72	1.71	1.64	1.68	1.70	1.65	1.68	1.36	1.38	1.38	1.39	1.44	1.43	1.54
Pentanol	0.00	0.08	0.07	0.07	0.07	0.08	0.08	0.07	0.07	0.07	0.07	0.06	0.07	0.07
Hexyl_Acetate	0.00	0.67	0.68	0.65	0.61	0.63	0.60	0.65	0.64	0.64	0.63	0.61	0.62	0.60
Acetoin	0.00	5.74	0.00	0.00	0.00	5.19	0.00	0.00	5.14	0.00	4.78	4.19	0.00	4.20
3-Methyl-1-Pentanol	0.00	0.52	0.51	0.51	0.52	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.52
Ethyl_Lactate	20.53	29.89	27.33	23.96	33.03	32.32	37.56	18.74	28.97	27.04	28.42	25.62	27.34	27.92
Hexanol	1.13	1.28	1.18	1.25	1.34	1.28	1.37	1.04	1.13	1.12	1.14	1.17	1.09	1.27
3-Ethoxy-1-Propanol	0.00	7.29	6.91	5.23	6.66	7.29	7.23	6.66	8.90	9.01	8.32	6.65	7.68	6.83
Ethyl_Caprylate	0.83	0.84	0.82	0.77	0.78	0.77	0.75	0.63	0.56	0.62	0.51	0.55	0.77	0.73
Acetic_Acid	751.87	785.35	748.61	600.93	716.04	773.25	832.09	475.08	601.43	600.01	560.08	450.88	514.23	469.36
Ethyl-3-hydroxybutanoate	0.00	0.00	0.00	0.56	0.67	0.68	0.71	0.81	0.96	0.91	0.90	0.77	0.84	0.79
Propionic_Acid	0.00	2.65	2.51	2.07	2.26	2.65	2.89	2.54	3.21	3.42	3.15	2.75	3.08	2.76
Isobutyric_Acid	0.00	1.75	1.65	1.51	1.65	1.69	1.71	1.05	1.17	1.16	1.11	0.97	1.02	1.03
Butyric_Acid	0.00	1.79	1.69	1.59	1.84	1.78	1.93	1.48	1.66	1.66	1.60	1.43	1.47	1.56
Ethyl_Caprate	0.00	0.18	0.19	0.15	0.15	0.16	0.13	0.16	0.15	0.17	0.14	0.14	0.24	0.20
Isovaleric_Acid	1.23	1.37	1.28	1.27	1.34	1.32	1.36	1.19	1.26	1.26	1.23	1.18	1.14	1.27
Diethyl_Succinate	0.97	2.27	1.70	2.53	4.74	3.24	6.05	1.96	3.59	3.09	3.95	6.88	5.12	8.85
Valeric_Acid	0.00	0.14	0.15	0.09	0.08	0.11	0.20	0.20	0.21	0.23	0.18	0.10	0.15	0.08
Ethyl_Phenethylacetate	0.00	1.03	1.04	1.00	1.00	1.01	1.00	1.11	1.08	1.10	1.05	1.01	1.03	1.01
2-Phenylacetate	0.00	0.39	0.39	0.36	0.28	0.32	0.24	0.40	0.36	0.37	0.34	0.28	0.31	0.27
Hexanoic_Acid	4.31	4.96	4.54	4.78	5.00	4.77	5.02	2.93	3.13	3.11	3.16	3.29	2.93	3.62
2-Phenylethanol	21.02	26.34	24.76	23.85	26.56	25.73	27.55	21.32	23.71	23.62	23.09	21.52	21.33	23.38
Octanoic_Acid	5.74	6.56	5.94	6.42	6.64	6.23	6.60	3.33	3.76	3.49	3.92	4.26	3.48	4.79
Decanoic_Acid	2.55	2.34	2.18	2.26	2.41	2.21	2.45	1.22	1.35	1.28	1.44	1.48	1.36	1.71
3-Mercapto-1-hexanol (3-MH)	574.37	112.19	373.91	775.83	1057.09	821.07	942.56	477.34	468.67	530.73	410.51	433.64	357.89	603.21
3-mercaptohexyl acetate (3MHA)	212.77	75.28	84.48	109.05	52.40	106.72	40.69	74.92	54.11	38.96	33.94	6.88	14.67	0.59
4-mercapto-4-methylpentan-2-one (4MMP)	27.97	6.74	25.88	83.70	95.56	28.27	51.05	6.51	8.29	5.23	5.06	7.64	5.35	12.12

Table 3.7: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from CDB winery.

Sample ID	Sauvignon Blanc CDB							Chenin Blanc CDB						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	105.72	117.80	109.29	122.88	120.76	123.08	118.06	76.39	84.29	90.12	79.40	79.55	82.89	64.02
Methanol	133.31	105.17	112.01	112.04	119.16	117.40	120.80	86.44	105.13	113.06	85.17	98.81	98.63	85.18
Ethyl-2-Methyl-Propanoate	0.00	2.09	2.09	2.11	2.13	2.12	2.13	0.00	0.00	0.00	0.00	2.10	0.00	2.11
Ethyl_Butyrate	0.73	0.73	0.73	0.77	0.73	0.72	0.66	0.75	0.77	0.77	0.73	0.76	0.77	0.73
Propanol	39.73	39.96	38.62	40.33	40.42	42.28	41.71	49.99	64.27	68.81	58.32	64.60	63.45	47.60
Isobutanol	26.64	28.58	26.65	29.98	28.87	29.24	29.24	14.17	16.92	17.98	15.95	17.51	17.24	13.81
Isoamyl_Acetate	9.96	8.91	9.32	8.93	7.11	7.87	6.07	8.69	7.82	8.23	7.13	6.16	7.09	5.67
Butanol	0.83	0.92	0.85	0.95	0.99	0.97	1.02	1.25	1.53	1.61	1.42	1.60	1.57	1.25
Isoamyl_Alcohol	176.35	190.39	181.72	206.30	197.85	190.89	188.38	124.51	143.84	149.30	135.42	152.06	148.86	130.73
Ethyl_Hexanoate	1.75	1.75	1.75	1.82	1.75	1.69	1.56	1.65	1.62	1.64	1.57	1.62	1.65	1.62
Pentanol	0.10	0.11	0.11	0.12	0.11	0.11	0.11	0.07	0.07	0.08	0.07	0.07	0.08	0.07
Hexyl_Acetate	0.70	0.68	0.68	0.68	0.63	0.65	0.61	0.76	0.72	0.74	0.70	0.65	0.69	0.63
Acetoin	4.91	4.58	0.00	4.86	0.00	4.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-Methyl-1-Pentanol	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.50	0.51	0.50	0.50	0.51	0.51	0.50
Ethyl_Lactate	6.12	8.94	7.82	10.20	13.55	11.85	15.69	7.07	11.94	11.39	12.06	16.57	14.15	13.38
Hexanol	0.73	0.79	0.78	0.86	0.87	0.80	0.78	0.85	0.95	0.94	0.91	1.05	1.03	1.02
3-Ethoxy-1-Propanol	5.81	5.29	5.36	5.37	5.67	5.67	5.97	5.22	6.65	7.08	5.89	6.76	6.64	4.99
Ethyl_Caprylate	0.79	0.85	0.79	0.84	0.82	0.72	0.62	0.66	0.54	0.64	0.52	0.56	0.65	0.56
Acetic_Acid	732.49	709.71	700.21	747.45	744.81	761.70	0.00	354.35	440.27	463.06	390.23	442.50	429.38	343.51
Ethyl-3-hydroxybutanoate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.70	0.00	0.68	0.69	0.56
Propionic_Acid	1.94	1.61	1.89	1.86	1.79	2.13	1.82	1.95	2.18	2.39	1.85	2.12	2.16	1.84
Isobutyric_Acid	1.25	1.32	1.24	1.42	1.30	1.29	1.29	0.72	0.82	0.84	0.77	0.82	0.81	0.69
Butyric_Acid	1.59	1.72	1.63	1.85	1.80	1.74	1.83	1.84	2.15	2.22	2.05	2.29	2.22	1.90
Ethyl_Caprate	0.18	0.19	0.17	0.20	0.20	0.17	0.16	0.13	0.12	0.15	0.12	0.10	0.17	0.14
Isovaleric_Acid	0.96	0.98	0.97	1.07	1.01	0.97	0.93	0.74	0.79	0.81	0.74	0.82	0.82	0.74
Diethyl_Succinate	0.07	0.28	0.20	0.38	0.83	0.51	0.95	0.20	0.49	0.39	0.57	1.43	0.95	1.61
Valeric_Acid	0.06	0.13	0.15	0.13	0.10	0.13	0.09	0.12	0.13	0.14	0.10	0.08	0.10	0.05
Ethyl_Phenethylacetate	1.11	1.05	1.06	1.03	0.99	1.02	0.98	1.05	1.03	1.05	1.01	0.99	1.00	0.98
2-Phenylacetate	0.57	0.51	0.53	0.52	0.40	0.44	0.32	0.37	0.33	0.35	0.31	0.27	0.31	0.26
Hexanoic_Acid	5.24	5.43	5.47	5.98	5.79	5.44	4.99	4.24	4.62	4.59	4.26	4.76	4.84	4.61
2-Phenylethanol	13.57	14.58	14.17	15.80	15.07	14.52	14.53	8.85	9.95	10.20	9.41	10.40	10.24	9.25
Octanoic_Acid	5.58	5.54	5.68	6.29	6.21	5.77	5.27	4.39	5.13	4.91	4.71	5.03	5.33	5.46
Decanoic_Acid	1.71	1.40	1.38	1.65	1.65	1.58	1.52	1.27	1.58	1.48	1.51	1.47	1.60	1.75
3-Mercapto-1-hexanol (3-MH)	4596.26	6662.81	633.87	42.02	1495.19	655.69	1864.26	379.45	472.83	379.93	569.16	909.06	462.46	835.06
3-mercaptohexyl acetate (3MHA)	1255.15	2043.78	180.91	7.79	187.98	260.98	210.54	164.72	128.03	131.29	166.44	161.88	103.16	105.42
4-mercapto-4-methylpentan-2-one (4MMP)	57.63	80.40	7.20	0.29	16.90	13.42	19.24	4.54	1.72	2.57	2.17	1.32	1.68	0.77

Table 3.8: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from DTK winery.

Sample ID	Sauvignon Blanc DTK							Chenin Blanc DTK						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	43.23	36.85	40.72	42.20	36.65	47.62	38.50	62.51	62.91	53.49	57.96	57.98	59.97	52.51
Methanol	77.07	57.48	82.89	96.26	86.59	101.16	88.99	105.54	108.99	102.55	122.07	99.83	110.18	93.66
Ethyl-2-Methyl-Propanoate	2.10	2.11	2.10	2.11	2.13	2.12	2.14	2.10	2.11	2.10	2.12	2.13	2.12	2.14
Ethyl_Butyrate	0.76	0.72	0.72	0.69	0.65	0.69	0.66	0.95	0.86	0.79	0.83	0.78	0.80	0.75
Propanol	44.47	35.84	42.28	48.71	41.00	57.18	45.32	43.25	48.27	43.60	50.78	45.00	51.63	41.67
Isobutanol	21.09	18.41	20.86	23.04	19.32	26.00	21.17	20.04	21.21	18.73	21.02	20.21	21.74	19.08
Isoamyl_Acetate	6.72	5.61	5.86	5.23	4.57	4.97	4.36	7.47	6.19	5.75	5.41	4.69	5.09	4.39
Butanol	0.79	0.68	0.79	0.86	0.74	0.93	0.78	0.97	1.05	0.94	1.05	1.00	1.08	0.95
Isoamyl_Alcohol	154.76	142.78	151.82	161.59	143.81	170.91	153.38	161.65	157.16	138.63	154.56	151.05	155.16	148.26
Ethyl_Hexanoate	1.58	1.52	1.51	1.47	1.50	1.49	1.47	1.75	1.71	1.59	1.67	1.63	1.63	1.55
Pentanol	0.10	0.08	0.09	0.09	0.08	0.10	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08
Hexyl_Acetate	0.74	0.66	0.68	0.64	0.60	0.62	0.58	0.76	0.70	0.67	0.65	0.60	0.63	0.58
Acetoin	3.99	4.95	5.25	0.00	3.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-Methyl-1-Pentanol	0.49	0.49	0.49	0.49	0.49	0.50	0.49	0.50	0.50	0.49	0.50	0.50	0.50	0.50
Ethyl_Lactate	17.07	17.15	19.74	26.24	23.19	30.53	26.15	25.70	33.37	30.84	37.44	35.39	38.37	34.24
Hexanol	1.84	1.81	1.78	1.85	1.79	1.88	1.88	1.73	1.61	1.45	1.61	1.60	1.58	1.62
3-Ethoxy-1-Propanol	3.12	2.55	2.95	3.60	2.95	3.92	3.15	1.63	1.87	1.81	2.02	1.67	1.92	1.62
Ethyl_Caprylate	0.96	0.83	0.78	0.72	0.85	0.93	0.79	0.75	1.01	0.85	0.83	0.83	0.91	0.66
Acetic_Acid	230.10	188.94	229.79	266.05	213.90	304.42	238.05	289.91	329.24	311.87	349.16	308.12	346.39	291.58
Ethyl-3-hydroxybutanoate	0.57	0.50	0.58	0.60	0.52	0.00	0.56	0.57	0.00	0.00	0.65	0.00	0.65	0.62
Propionic_Acid	1.45	1.26	1.77	1.46	1.22	1.67	1.38	1.39	1.59	1.58	1.60	1.38	1.57	1.37
Isobutyric_Acid	0.88	0.83	0.87	0.92	0.79	0.98	0.83	0.85	0.84	0.75	0.81	0.78	0.81	0.76
Butyric_Acid	1.77	1.66	1.79	1.96	1.72	2.10	1.87	2.55	2.59	2.31	2.58	2.56	2.63	2.51
Ethyl_Caprate	0.20	0.16	0.15	0.16	0.17	0.17	0.18	0.16	0.24	0.20	0.17	0.18	0.21	0.13
Isovaleric_Acid	0.77	0.70	0.72	0.74	0.67	0.75	0.70	0.77	0.70	0.63	0.68	0.65	0.67	0.65
Diethyl_Succinate	0.53	0.99	0.83	1.27	2.12	1.68	2.69	0.63	0.93	0.87	1.28	1.99	1.56	2.52
Valeric_Acid	0.15	0.07	0.09	0.08	0.13	0.07	0.14	0.07	0.11	0.10	0.11	0.10	0.11	0.10
Ethyl_Phenethylacetate	1.01	0.98	0.99	0.98	0.97	0.98	0.97	1.01	1.00	0.99	0.99	0.98	0.98	0.98
2-Phenylacetate	0.27	0.23	0.24	0.22	0.19	0.21	0.18	0.27	0.23	0.22	0.21	0.19	0.20	0.00
Hexanoic_Acid	4.66	4.44	4.35	4.40	4.15	4.34	4.36	5.84	5.07	4.57	4.94	4.80	4.71	4.81
2-Phenylethanol	11.56	10.90	11.43	12.06	10.89	12.36	11.49	8.36	8.14	7.28	7.88	7.85	7.89	7.76
Octanoic_Acid	5.27	4.93	4.91	5.03	4.66	4.48	4.91	7.01	5.86	5.55	6.02	5.78	5.48	5.55
Decanoic_Acid	1.29	1.16	1.18	1.27	1.17	1.12	1.19	1.70	1.47	1.57	1.67	1.60	1.56	1.55
3-Mercapto-1-hexanol (3-MH)	394.21	343.69	351.73	339.89	360.48	770.82	817.09	175.44	163.48	121.91	225.70	370.30	173.89	1486.75
3-mercaptohexyl acetate (3MHA)	103.47	45.70	40.88	49.86	11.68	56.45	28.32	54.38	38.20	38.90	38.12	27.29	19.92	82.59
4-mercapto-4-methylpentan-2-one (4MMP)	1.33	0.86	1.34	0.92	1.21	2.25	2.64	1.13	1.15	0.97	0.80	1.54	2.00	2.45

Table 3.9: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from FRV winery.

Sample ID	Sauvignon Blanc FRV							Chenin Blanc FRV						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	73.22	82.95	83.80	67.97	79.55	85.17	84.74	63.99	73.27	72.26	69.87	76.04	51.18	68.47
Methanol	65.02	67.13	68.43	67.72	67.40	66.60	84.66	54.37	70.71	63.56	48.18	67.94	85.35	62.97
Ethyl-2-Methyl-Propanoate	2.11	2.14	2.12	2.13	2.15	2.15	2.17	2.08	2.12	2.11	2.13	2.14	0.00	2.14
Ethyl_Butyrate	0.66	0.68	0.66	0.64	0.62	0.64	0.64	0.75	0.77	0.78	0.77	0.74	0.62	0.71
Propanol	53.31	56.82	60.10	43.12	55.34	58.40	61.86	42.75	56.90	52.55	45.83	58.92	60.89	50.21
Isobutanol	22.44	23.50	24.40	19.02	22.63	23.80	24.68	19.70	24.29	22.94	22.02	25.13	25.91	22.08
Isoamyl_Acetate	6.40	5.63	5.79	5.27	4.55	4.98	4.42	7.20	6.58	6.88	6.19	5.23	5.24	4.90
Butanol	0.91	0.95	0.97	0.79	0.93	0.96	1.01	1.04	1.28	1.21	1.16	1.33	1.36	1.18
Isoamyl_Alcohol	165.17	173.06	169.63	153.17	161.79	167.68	178.99	145.82	173.62	163.95	162.19	171.10	178.06	157.59
Ethyl_Hexanoate	1.50	1.52	1.49	1.49	1.44	1.46	1.48	1.48	1.53	1.63	1.58	1.56	1.37	1.56
Pentanol	0.11	0.11	0.11	0.10	0.11	0.11	0.12	0.07	0.08	0.08	0.08	0.09	0.09	0.08
Hexyl_Acetate	0.74	0.68	0.71	0.66	0.60	0.64	0.59	0.69	0.66	0.68	0.66	0.61	0.62	0.61
Acetoin	0.00	0.00	0.00	0.00	0.00	0.00	3.64	6.63	8.58	7.74	6.02	8.81	9.79	7.83
3-Methyl-1-Pentanol	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.51	0.52	0.51	0.51	0.51	0.51	0.51
Ethyl_Lactate	17.71	21.19	21.01	17.86	24.25	23.55	28.05	11.16	17.93	15.85	15.65	23.21	21.55	20.07
Hexanol	2.38	2.60	2.38	2.46	2.38	2.42	2.63	1.34	1.54	1.47	1.47	1.51	1.56	1.47
3-Ethoxy-1-Propanol	7.73	7.82	8.11	6.29	7.86	7.95	9.21	3.23	4.43	4.14	3.31	4.60	4.74	3.86
Ethyl_Caprylate	0.45	0.45	0.38	0.46	0.48	0.45	0.45	0.42	0.39	0.59	0.55	0.49	0.34	0.48
Acetic_Acid	460.23	464.21	489.83	374.17	462.64	473.91	527.83	313.16	403.45	365.58	321.04	422.00	448.81	361.30
Ethyl-3-hydroxybutanoate	0.55	0.54	0.58	0.49	0.56	0.55	0.62	0.58	0.69	0.65	0.61	0.74	0.72	0.64
Propionic_Acid	1.68	1.63	1.60	1.44	1.55	1.55	1.79	1.71	1.90	1.75	1.41	1.92	2.15	1.71
Isobutyric_Acid	0.92	0.93	0.92	0.81	0.87	0.89	0.94	0.82	0.95	0.89	0.89	0.91	0.96	0.82
Butyric_Acid	1.49	1.55	1.55	1.38	1.56	1.55	1.69	1.70	2.02	1.89	1.90	2.08	2.13	1.86
Ethyl_Caprate	0.08	0.08	0.05	0.09	0.11	0.10	0.09	0.08	0.08	0.10	0.12	0.09	0.07	0.08
Isovaleric_Acid	0.80	0.82	0.77	0.74	0.74	0.75	0.79	0.76	0.85	0.80	0.79	0.81	0.87	0.74
Diethyl_Succinate	0.38	0.85	0.60	0.96	1.78	1.23	2.42	0.77	1.43	1.27	1.66	2.85	2.18	3.19
Valeric_Acid	0.12	0.09	0.10	0.06	0.09	0.06	0.10	0.10	0.11	0.11	0.07	0.07	0.10	0.09
Ethyl_Phenethylacetate	1.02	0.99	1.00	0.98	0.98	0.98	0.98	1.06	1.03	1.04	1.00	1.00	1.01	0.99
2-Phenylacetate	0.31	0.27	0.28	0.25	0.20	0.23	0.19	0.36	0.33	0.34	0.30	0.25	0.29	0.23
Hexanoic_Acid	4.38	4.74	4.25	4.46	4.21	4.26	4.57	4.54	5.09	4.77	4.75	4.71	5.05	4.51
2-Phenylethanol	12.41	13.03	12.66	11.69	12.56	12.57	13.43	12.28	14.15	13.54	13.37	13.94	14.54	12.79
Octanoic_Acid	5.11	5.59	5.17	5.54	5.16	5.19	5.69	4.55	5.24	4.63	4.99	4.88	5.29	4.89
Decanoic_Acid	1.31	1.34	1.32	1.58	1.55	1.47	1.71	1.20	1.21	1.02	1.39	1.26	1.33	1.34
3-Mercapto-1-hexanol (3-MH)	485.76	589.16	298.34	718.15	335.76	689.28	55.24	527.92	392.78	393.03	726.51	259.03	735.96	503.78
3-mercaptohexyl acetate (3MHA)	294.24	155.59	58.01	56.99	10.99	35.01	0.00	165.60	65.68	58.12	93.07	51.14	63.14	23.44
4-mercapto-4-methylpentan-2-one (4MMP)	3.22	2.51	1.59	3.82	1.36	3.65	0.13	0.75	0.79	0.41	0.84	1.55	1.79	1.10

Table 3.10: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from KZC winery.

Sample ID	Sauvignon Blanc KZC							Chenin Blanc KZC						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	73.52	69.27	70.40	75.62	73.34	77.35	78.33	120.16	116.67	119.79	99.18	83.65	86.77	97.82
Methanol	117.09	88.83	97.08	88.78	85.50	100.11	110.10	78.43	73.79	115.44	105.02	80.58	87.91	85.24
Ethyl-2-Methyl-Propanoate	0.00	2.09	2.08	0.00	2.10	2.09	2.11	0.00	0.00	0.00	0.00	2.10	2.09	2.11
Ethyl_Butyrate	0.69	0.66	0.68	0.63	0.62	0.66	0.63	0.66	0.67	0.65	0.62	0.60	0.58	0.61
Propanol	42.10	35.79	37.40	40.77	38.54	41.46	41.47	36.41	34.41	40.61	36.48	27.38	30.44	35.59
Isobutanol	22.22	19.01	20.60	21.05	20.39	21.85	21.58	29.04	28.86	31.47	28.99	23.92	24.48	28.87
Isoamyl_Acetate	7.09	6.02	6.39	5.58	4.94	5.50	4.79	10.04	9.03	9.12	7.84	6.45	6.90	5.99
Butanol	0.94	0.81	0.88	0.88	0.87	0.91	0.92	0.95	0.96	1.05	0.98	0.81	0.84	0.97
Isoamyl_Alcohol	144.75	125.01	140.81	127.25	125.96	138.71	136.74	175.84	179.93	192.53	176.26	160.44	152.79	184.87
Ethyl_Hexanoate	1.66	1.47	1.55	1.42	1.40	1.48	1.44	1.65	1.69	1.63	1.56	1.57	1.51	1.56
Pentanol	0.09	0.08	0.09	0.08	0.07	0.08	0.08	0.08	0.08	0.09	0.08	0.07	0.07	0.08
Hexyl_Acetate	0.78	0.68	0.70	0.65	0.61	0.64	0.60	0.81	0.76	0.76	0.71	0.65	0.66	0.64
Acetoin	0.00	0.00	0.00	0.00	0.00	4.59	4.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-Methyl-1-Pentanol	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.48	0.48	0.49	0.48	0.48	0.48	0.48
Ethyl_Lactate	10.68	10.79	11.54	13.50	15.19	15.10	17.44	9.60	11.99	13.56	15.18	13.92	13.91	18.88
Hexanol	1.15	1.09	1.19	1.05	1.08	1.19	1.20	1.19	1.26	1.31	1.21	1.24	1.12	1.35
3-Ethoxy-1-Propanol	3.52	2.79	3.22	3.06	2.96	3.32	3.17	2.15	1.98	2.51	2.32	1.81	2.02	2.31
Ethyl_Caprylate	1.06	0.60	0.71	0.60	0.57	0.65	0.54	0.61	0.75	0.76	0.60	0.80	0.81	0.58
Acetic_Acid	489.20	370.84	421.70	422.31	401.22	447.89	452.96	549.93	504.84	0.00	0.00	0.00	499.14	0.00
Ethyl-3-hydroxybutanoate	0.56	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58
Propionic_Acid	1.94	1.57	1.78	1.59	1.64	1.70	1.98	1.58	1.44	1.73	1.77	1.38	1.46	1.60
Isobutyric_Acid	0.88	0.73	0.82	0.73	0.70	0.80	0.79	0.93	0.90	1.00	0.89	0.78	0.77	0.89
Butyric_Acid	1.75	1.45	1.69	1.55	1.55	1.67	1.66	1.35	1.35	1.50	1.34	1.22	1.17	1.43
Ethyl_Caprate	0.36	0.15	0.18	0.17	0.14	0.17	0.15	0.14	0.19	0.23	0.16	0.20	0.18	0.15
Isovaleric_Acid	0.75	0.64	0.72	0.62	0.61	0.69	0.66	0.68	0.66	0.72	0.63	0.60	0.55	0.66
Diethyl_Succinate	0.57	1.04	1.03	1.29	2.28	1.88	3.06	0.17	0.37	0.34	0.44	0.85	0.63	1.11
Valeric_Acid	0.13	0.09	0.10	0.08	0.10	0.08	0.10	0.13	0.10	0.14	0.11	0.06	0.09	0.07
Ethyl_Phenethylacetate	1.13	1.03	1.06	1.03	1.00	1.02	0.99	1.10	1.03	1.07	1.02	0.98	1.00	0.98
2-Phenylacetate	0.39	0.31	0.34	0.28	0.24	0.28	0.23	0.74	0.66	0.68	0.55	0.43	0.46	0.39
Hexanoic_Acid	4.47	4.06	4.46	3.73	3.72	4.28	4.18	5.03	5.18	5.37	4.73	4.70	4.16	5.11
2-Phenylethanol	12.04	10.46	11.90	10.67	10.76	11.53	11.31	15.91	15.90	17.45	15.42	14.54	13.50	16.53
Octanoic_Acid	4.92	4.51	4.77	3.95	4.00	4.46	4.63	6.24	6.32	6.33	5.57	5.28	4.37	6.17
Decanoic_Acid	1.73	1.40	1.42	1.23	1.30	1.32	1.41	2.01	1.88	1.85	1.65	1.43	1.20	1.89
3-Mercapto-1-hexanol (3-MH)	376.06	547.03	458.22	342.76	317.11	483.72	611.21	1199.16	3731.88	1454.57	2009.82	2563.13	1829.98	2162.48
3-mercaptohexyl acetate (3MHA)	295.82	236.19	222.99	126.91	24.35	103.51	39.06	121.94	542.23	197.14	250.79	120.88	198.69	122.70
4-mercapto-4-methylpentan-2-one (4MMP)	7.75	11.52	8.88	8.66	5.66	8.44	10.29	0.93	3.07	1.07	1.46	2.12	1.51	2.44

Table 3.11: Volatile compound composition of Sauvignon Blanc and Chenin Blanc wines from PDB winery.

Sample ID	Sauvignon Blanc PDB							Chenin Blanc PDB						
Treatment	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25	Control	T3/RT	T3/15	T3/25	T9/RT	T9/15	T9/25
Storage Time (months)	0	3	3	3	9	9	9	0	3	3	3	9	9	9
Temperature (°C)	-4	RT	15	25	RT	15	25	-4	RT	15	25	RT	15	25
Ethyl_Acetate	83.33	113.85	89.44	96.98	99.07	96.35	99.08	98.46	90.38	100.12	103.03	73.68	95.90	86.38
Methanol	89.84	115.78	68.83	71.22	71.85	67.52	77.16	138.34	75.28	97.93	122.23	78.99	100.57	88.67
Ethyl-2-Methyl-Propanoate	2.12	2.15	2.14	2.17	2.19	2.17	2.20	0.00	2.12	0.00	2.12	2.11	2.12	2.14
Ethyl_Butyrate	0.63	0.65	0.65	0.67	0.64	0.64	0.61	0.61	0.63	0.61	0.62	0.55	0.60	0.58
Propanol	40.51	60.14	37.89	39.79	40.98	38.78	38.93	39.20	27.26	33.66	34.16	28.29	31.95	26.35
Isobutanol	20.88	29.07	20.92	21.93	21.92	21.65	21.10	24.67	19.62	22.70	23.23	19.68	21.76	18.79
Isoamyl_Acetate	5.95	5.46	5.71	5.32	4.71	5.02	4.47	6.46	6.08	6.07	5.67	4.74	5.39	4.62
Butanol	0.79	1.09	0.79	0.84	0.84	0.83	0.81	1.02	0.81	0.95	0.96	0.82	0.91	0.80
Isoamyl_Alcohol	141.38	181.41	150.81	157.41	153.11	151.46	145.52	148.10	140.12	143.09	151.57	132.71	143.28	133.10
Ethyl_Hexanoate	1.48	1.51	1.55	1.57	1.54	1.52	1.47	1.48	1.60	1.50	1.56	1.43	1.55	1.46
Pentanol	0.08	0.11	0.09	0.09	0.09	0.09	0.08	0.09	0.08	0.08	0.09	0.08	0.08	0.08
Hexyl_Acetate	0.67	0.63	0.65	0.63	0.60	0.61	0.59	0.68	0.66	0.66	0.64	0.60	0.63	0.59
Acetoin	5.37	6.38	0.00	4.25	4.63	4.28	5.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3-Methyl-1-Pentanol	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.00	0.48	0.00	0.48	0.48	0.48	0.48
Ethyl_Lactate	15.40	26.83	15.76	18.31	20.86	19.05	21.32	15.46	12.25	14.60	17.00	15.28	16.63	15.37
Hexanol	1.25	1.45	1.37	1.44	1.41	1.37	1.33	1.20	1.34	1.22	1.35	1.25	1.30	1.30
3-Ethoxy-1-Propanol	4.28	6.15	3.63	3.89	4.00	3.78	3.95	2.15	1.51	1.78	1.85	1.50	1.75	1.54
Ethyl_Caprylate	0.55	0.46	0.59	0.53	0.50	0.52	0.48	0.58	0.62	0.62	0.62	0.51	0.61	0.43
Acetic_Acid	575.65	0.00	515.99	540.40	540.07	526.77	547.57	0.00	0.00	673.77	0.00	0.00	0.00	0.00
Ethyl-3-hydroxybutanoate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Propionic_Acid	1.65	2.23	1.49	1.50	1.54	1.51	1.58	1.97	1.26	1.56	1.58	1.26	1.56	1.28
Isobutyric_Acid	1.16	1.51	1.19	1.23	1.18	1.17	1.11	1.21	1.08	1.15	1.19	1.03	1.12	1.00
Butyric_Acid	1.39	1.87	1.44	1.53	1.52	1.48	1.46	1.68	1.52	1.61	1.71	1.51	1.63	1.49
Ethyl_Caprate	0.11	0.07	0.11	0.10	0.09	0.10	0.09	0.14	0.13	0.14	0.13	0.12	0.14	0.10
Isovaleric_Acid	0.85	0.99	0.88	0.90	0.86	0.85	0.81	0.78	0.79	0.76	0.81	0.73	0.77	0.74
Diethyl_Succinate	0.58	1.25	0.96	1.44	2.37	1.71	2.80	0.30	0.62	0.46	0.76	1.24	0.90	1.58
Valeric_Acid	0.11	0.11	0.08	0.07	0.12	0.06	0.11	0.10	0.06	0.07	0.06	0.08	0.09	0.08
Ethyl_Phenethylacetate	1.04	1.04	1.01	1.00	0.99	0.99	0.98	1.08	1.01	1.03	1.01	0.99	1.00	0.98
2-Phenylacetate	0.31	0.28	0.29	0.27	0.22	0.24	0.20	0.35	0.34	0.33	0.31	0.25	0.29	0.23
Hexanoic_Acid	4.27	4.75	4.61	4.86	4.59	4.41	4.25	4.58	5.26	4.61	5.03	4.59	4.85	4.83
2-Phenylethanol	11.86	14.74	12.59	13.25	12.90	12.52	12.25	11.04	10.86	10.80	11.56	10.42	10.99	10.61
Octanoic_Acid	4.92	5.36	5.14	5.59	5.40	5.10	4.93	4.33	5.16	4.33	4.81	4.43	4.79	4.70
Decanoic_Acid	1.43	1.50	1.33	1.45	1.52	1.39	1.38	1.16	1.25	1.10	1.23	1.15	1.28	1.15
3-Mercapto-1-hexanol (3-MH)	197.27	180.95	218.04	229.83	199.55	320.48	235.23	330.53	279.22	267.17	313.42	247.77	209.05	234.71
3-mercaptohexyl acetate (3MHA)	86.42	72.62	78.76	75.65	71.75	nd	69.68	116.39	nd	93.18	92.07	76.48	78.63	74.00
4-mercapto-4-methylpentan-2-one (4MMP)	1.27	1.35	1.81	1.54	1.69	1.16	1.61	0.44	0.29	0.64	0.58	0.67	0.51	0.70

Table 3.12: Antioxidant-related compound composition of Sauvignon Blanc wines stored at different conditions.

Sample ID	Storage Time (months)	Temperature (°C)	FSO2 (pppm)	TSO2 (ppm)	Reduced GSH (ppm)	A520 (nm)	A420 (nm)	A320 (nm)	A280 (nm)	Colour density	Colour hue	L*	a*	b*	Cab*	hab*
AVN/SB/T0	0	-4	46	103	2.89	0.041	0.063	2.84	3.09	0.10	0.64	96.54	-0.31	1.92	1.95	-80.89
AVN/SB/T3/RT	3	RT	28	98	0.97	0.041	0.065	2.90	3.16	0.11	0.63	96.42	-0.47	2.06	2.11	-77.23
AVN/SB/T3/15	3	15	22	96	0.82	0.041	0.064	2.90	3.15	0.10	0.63	96.50	-0.32	1.98	2.00	-80.75
AVN/SB/T3/25	3	25	27	93	0.15	0.041	0.066	2.89	3.15	0.11	0.62	96.45	-0.30	2.07	2.09	-81.80
AVN/SB/T9/RT	9	RT	25	93	0.37	0.042	0.069	2.94	3.18	0.11	0.62	96.42	-0.35	2.28	2.30	-81.22
AVN/SB/T9/15	9	15	25	92	0.30	0.043	0.068	2.93	3.17	0.11	0.64	96.36	-0.34	2.15	2.18	-81.05
AVN/SB/T9/25	9	25	24	87	0.20	0.042	0.069	2.94	3.22	0.11	0.60	96.46	-0.37	2.33	2.36	-81.05
CDB/SB/T0	0	-4	52	89	2.40	0.040	0.060	2.80	3.25	0.10	0.67	96.48	-0.31	1.65	1.68	-79.41
CDB/SB/T3/RT	3	RT	51	89	8.12	0.072	0.093	2.92	3.34	0.17	0.77	94.21	-0.22	2.16	2.17	-84.09
CDB/SB/T3/15	3	15	52	101	4.23	0.046	0.067	2.91	3.33	0.11	0.69	96.04	-0.38	1.74	1.78	-77.64
CDB/SB/T3/25	3	25	50	99	6.74	0.046	0.065	2.94	3.33	0.11	0.71	96.27	-0.39	1.78	1.82	-77.70
CDB/SB/T9/RT	9	RT	47	93	0.31	0.040	0.066	2.90	3.34	0.11	0.61	96.44	-0.38	2.00	2.04	-79.30
CDB/SB/T9/15	9	15	27	95	3.12	0.041	0.063	2.86	3.29	0.10	0.65	96.51	-0.35	1.89	1.92	-79.44
CDB/SB/T9/25	9	25	51	89	0.53	0.041	0.065	2.83	3.29	0.11	0.63	96.57	-0.45	2.08	2.13	-77.77
DTK/SB/T0	0	-4	19	101	1.26	0.043	0.079	2.43	3.26	0.12	0.54	96.08	-0.55	2.71	2.76	-78.50
DTK/SB/T3/RT	3	RT	17	88	0.35	0.042	0.069	2.30	3.11	0.11	0.60	96.12	-0.52	1.75	1.83	-73.36
DTK/SB/T3/15	3	15	18	92	0.56	0.043	0.075	2.36	3.18	0.12	0.57	96.26	-0.51	2.56	2.61	-78.66
DTK/SB/T3/25	3	25	18	90	0.35	0.041	0.070	2.33	3.15	0.11	0.59	96.40	-0.41	2.38	2.42	-80.31
DTK/SB/T9/RT	9	RT	17	88	0.24	0.042	0.067	2.28	3.10	0.11	0.62	96.13	-0.21	1.94	1.95	-83.91
DTK/SB/T9/15	9	15	18	87	0.72	0.042	0.068	2.27	3.08	0.11	0.62	96.01	-0.25	1.89	1.90	-82.46
DTK/SB/T9/25	9	25	16	83	0.21	0.043	0.066	2.31	3.12	0.11	0.65	96.31	-0.26	2.02	2.04	-82.54
FRV/SB/T0	0	-4	18	77	0.27	0.044	0.072	1.92	2.66	0.12	0.61	96.23	-0.45	2.35	2.39	-79.11
FRV/SB/T3/RT	3	RT	17	67	0.64	0.040	0.063	1.83	2.51	0.10	0.65	96.30	-0.40	1.75	1.80	-77.25
FRV/SB/T3/15	3	15	16	67	0.38	0.047	0.074	1.90	2.61	0.12	0.63	96.03	-0.39	2.47	2.50	-81.08
FRV/SB/T3/25	3	25	17	68	0.72	0.040	0.062	1.84	2.53	0.10	0.65	96.20	-0.32	1.53	1.56	-78.30
FRV/SB/T9/RT	9	RT	15	64	1.26	0.040	0.061	1.81	2.49	0.10	0.66	96.47	-0.47	1.60	1.66	-73.75
FRV/SB/T9/15	9	15	15	65	0.96	0.041	0.062	1.80	2.48	0.10	0.65	96.46	-0.35	1.77	1.81	-78.80
FRV/SB/T9/25	9	25	15	63	1.83	0.042	0.062	1.76	2.44	0.10	0.68	96.09	-0.29	1.57	1.59	-79.47
KZC/SB/T0	0	-4	26	95	5.83	0.042	0.072	2.83	3.34	0.11	0.58	96.35	-0.42	2.53	2.56	-80.59
KZC/SB/T3/RT	3	RT	25	97	2.60	0.069	0.070	2.70	3.23	0.14	0.99	96.34	-0.60	1.89	1.98	-72.54
KZC/SB/T3/15	3	15	24	93	0.85	0.049	0.069	2.81	3.32	0.12	0.71	96.33	-0.79	2.12	2.26	-69.61
KZC/SB/T3/25	3	25	23	95	1.96	0.041	0.066	2.87	3.33	0.11	0.63	96.43	-0.45	2.08	2.13	-77.73
KZC/SB/T9/RT	9	RT	23	91	0.66	0.040	0.063	2.83	3.26	0.10	0.63	96.56	-0.31	2.03	2.06	-81.19
KZC/SB/T9/15	9	15	22	95	3.25	0.043	0.067	2.87	3.28	0.11	0.64	96.37	-0.34	2.15	2.18	-81.08
KZC/SB/T9/25	9	25	23	92	0.24	0.040	0.060	2.63	3.16	0.10	0.67	96.55	-0.37	1.63	1.67	-77.05
PDB/SB/T0	0	-4	22	90	1.42	0.042	0.067	2.43	2.93	0.11	0.63	96.39	-0.55	2.14	2.21	-75.58
PDB/SB/T3/RT	3	RT	22	85	0.83	0.041	0.068	2.51	3.00	0.11	0.61	96.37	-0.45	2.21	2.25	-78.45
PDB/SB/T3/15	3	15	23	84	1.11	0.042	0.067	2.53	3.02	0.11	0.62	96.33	-0.49	1.99	2.05	-76.28
PDB/SB/T3/25	3	25	22	84	0.67	0.043	0.070	2.55	3.04	0.11	0.61	96.43	-0.38	2.37	2.40	-80.83
PDB/SB/T9/RT	9	RT	22	84	0.34	0.043	0.073	2.56	3.05	0.12	0.59	96.42	-0.37	2.63	2.65	-82.01
PDB/SB/T9/15	9	15	21	87	0.69	0.042	0.071	2.55	3.04	0.11	0.60	96.43	-0.36	2.47	2.50	-81.70
PDB/SB/T9/25	9	25	19	88	0.43	0.043	0.075	2.63	3.13	0.12	0.57	96.43	-0.43	2.79	2.83	-81.18

Table 3.13: Antioxidant-related compound composition of Chenin Blanc wines stored at different conditions.

Sample ID	Storage Time (months)	Temperature (°C)	FSO2 (pppm)	TSO2 (ppm)	Reduced GSH (ppm)	A520 (nm)	A420 (nm)	A320 (nm)	A280 (nm)	Colour density	Colour hue	L*	a*	b*	Cab*	hab*
AVN/CB/T0	0	-4	22	71	3.20	0.047	0.084	3.02	3.55	0.13	0.56	96.09	-0.47	3.35	3.38	-81.98
AVN/CB/T3/RT	3	RT	19	76	1.78	0.045	0.087	3.09	3.53	0.13	0.52	96.23	-0.51	3.64	3.68	-82.07
AVN/CB/T3/15	3	15	21	87	1.77	0.045	0.084	3.06	3.54	0.13	0.54	96.21	-0.46	3.49	3.52	-82.52
AVN/CB/T3/25	3	25	22	84	1.24	0.046	0.088	3.11	3.55	0.13	0.52	96.19	-0.55	3.68	3.72	-81.52
AVN/CB/T9/RT	9	RT	21	80	0.40	0.047	0.092	3.12	3.55	0.14	0.50	96.12	-0.59	4.00	4.05	-81.59
AVN/CB/T9/15	9	15	21	82	0.56	0.046	0.088	3.08	3.53	0.13	0.52	96.16	-0.54	3.72	3.76	-81.75
AVN/CB/T9/25	9	25	20	80	0.34	0.048	0.097	3.17	3.57	0.14	0.49	96.05	-0.65	4.27	4.32	-81.35
CDB/CB/T0	0	-4	57	107	18.99	0.037	0.059	2.03	2.80	0.10	0.63	96.77	-0.43	1.74	1.80	-76.31
CDB/CB/T3/RT	3	RT	58	138	9.35	0.037	0.060	2.08	2.85	0.10	0.62	96.78	-0.46	1.83	1.89	-75.90
CDB/CB/T3/15	3	15	57	174	12.00	0.037	0.060	2.10	2.87	0.10	0.61	96.78	-0.44	1.84	1.89	-76.43
CDB/CB/T3/25	3	25	56	198	6.70	0.037	0.061	2.11	2.88	0.10	0.61	96.75	-0.45	1.88	1.94	-76.46
CDB/CB/T9/RT	9	RT	58	247	1.89	0.037	0.062	2.10	2.86	0.10	0.60	96.75	-0.49	1.97	2.03	-76.07
CDB/CB/T9/15	9	15	59	102	4.94	0.039	0.063	2.11	2.87	0.10	0.61	96.61	-0.47	1.92	1.98	-76.37
CDB/CB/T9/25	9	25	55	100	1.09	0.038	0.065	2.17	2.94	0.10	0.59	96.72	-0.52	2.11	2.17	-76.19
DTK/CB/T0	0	-4	33	106	3.83	0.039	0.069	2.43	3.26	0.11	0.57	96.60	-0.47	2.47	2.51	-79.29
DTK/CB/T3/RT	3	RT	33	100	2.24	0.040	0.072	2.47	3.28	0.11	0.55	96.53	-0.47	2.65	2.70	-79.94
DTK/CB/T3/15	3	15	33	102	2.57	0.041	0.072	2.48	3.29	0.11	0.56	96.52	-0.48	2.61	2.66	-79.55
DTK/CB/T3/25	3	25	26	102	1.32	0.043	0.078	2.52	3.32	0.12	0.56	96.33	-0.53	2.85	2.90	-79.51
DTK/CB/T9/RT	9	RT	29	107	0.59	0.041	0.079	2.57	3.35	0.12	0.53	96.45	-0.57	3.06	3.11	-79.46
DTK/CB/T9/15	9	15	40	112	1.07	0.040	0.075	2.52	3.33	0.11	0.54	96.52	-0.51	2.81	2.85	-79.61
DTK/CB/T9/25	9	25	34	93	0.40	0.042	0.082	2.60	3.39	0.12	0.52	96.40	-0.62	3.24	3.30	-79.11
FRV/CB/T0	0	-4	23	84	2.63	0.044	0.069	2.44	3.36	0.11	0.63	96.35	-0.21	2.20	2.21	-84.58
FRV/CB/T3/RT	3	RT	22	84	1.97	0.042	0.071	2.45	3.35	0.11	0.60	96.40	-0.36	2.38	2.40	-81.41
FRV/CB/T3/15	3	15	22	81	2.25	0.043	0.070	2.45	3.34	0.11	0.62	96.43	-0.33	2.30	2.33	-81.80
FRV/CB/T3/25	3	25	18	82	1.29	0.042	0.072	2.48	3.37	0.11	0.59	96.41	-0.40	2.48	2.51	-80.79
FRV/CB/T9/RT	9	RT	23	84	0.60	0.042	0.074	2.49	3.35	0.12	0.57	96.42	-0.47	2.64	2.68	-79.93
FRV/CB/T9/15	9	15	22	87	0.94	0.042	0.072	2.49	3.38	0.11	0.59	96.31	-0.39	2.42	2.45	-80.81
FRV/CB/T9/25	9	25	21	86	0.37	0.043	0.077	2.54	3.40	0.12	0.56	96.39	-0.47	2.84	2.88	-80.58
KZC/CB/T0	0	-4	25	96	1.80	0.044	0.070	2.11	3.17	0.11	0.62	96.41	-0.34	2.28	2.31	-81.42
KZC/CB/T3/RT	3	RT	24	95	0.96	0.044	0.076	2.14	3.20	0.12	0.58	96.24	-0.46	2.61	2.65	-79.94
KZC/CB/T3/15	3	15	25	95	1.26	0.043	0.072	2.11	3.17	0.11	0.59	96.41	-0.37	2.41	2.44	-81.15
KZC/CB/T3/25	3	25	24	88	0.79	0.043	0.077	2.17	3.22	0.12	0.56	96.34	-0.50	2.85	2.90	-79.98
KZC/CB/T9/RT	9	RT	22	90	0.32	0.045	0.084	2.24	3.30	0.13	0.53	96.25	-0.59	3.34	3.39	-79.97
KZC/CB/T9/15	9	15	24	87	0.45	0.043	0.079	2.17	3.22	0.12	0.54	96.35	-0.51	2.83	2.88	-79.71
KZC/CB/T9/25	9	25	22	83	0.16	0.045	0.089	2.27	3.34	0.13	0.50	96.25	-0.68	3.75	3.81	-79.80
PDB/CB/T0	0	-4	22	87	2.48	0.042	0.072	2.88	3.34	0.11	0.58	96.46	-0.41	2.43	2.47	-80.35
PDB/CB/T3/RT	3	RT	20	86	1.28	0.041	0.074	2.92	3.37	0.12	0.56	96.49	-0.45	2.68	2.72	-80.53
PDB/CB/T3/15	3	15	21	87	1.63	0.041	0.072	2.88	3.35	0.11	0.57	96.53	-0.42	2.56	2.59	-80.69
PDB/CB/T3/25	3	25	21	87	0.96	0.042	0.076	2.92	3.36	0.12	0.55	96.45	-0.46	2.79	2.83	-80.68
PDB/CB/T9/RT	9	RT	19	83	0.57	0.043	0.080	2.93	3.38	0.12	0.54	96.38	-0.50	3.01	3.05	-80.62
PDB/CB/T9/15	9	15	21	85	0.74	0.043	0.077	2.92	3.37	0.12	0.56	96.39	-0.45	2.82	2.85	-80.89
PDB/CB/T9/25	9	25	20	78	0.28	0.045	0.085	2.97	3.41	0.13	0.53	96.23	-0.53	3.29	3.33	-80.89

Chapter 4

Research results

Exploration of data fusion strategies using Principal Component Analysis (PCA) and Multiple Factor Analysis (MFA)

Chapter 4: Exploration of data fusion strategies using Principal Component Analysis (PCA) and Multiple Factor Analysis (MFA)

Abstract

In the field of oenology, statistical analyses are used for descriptive purposes, in the majority of cases sensory and chemistry data sets are kept separate. Cases that combine the different data sets are mostly supervised, usually seeking to optimize discrimination, classification or prediction power. Unsupervised methods are used as preliminary steps that work to refine the predictive/discriminant/classification power of supervised models. However, there is potential for unsupervised methods to combine different data sets into comprehensive, information-rich models. In this study, stepwise strategies for creating data fusion models at different levels of complexity were explored. Principal component analysis (PCA) and multiple factor analysis (MFA) were used to combine five data blocks (four chemistry – antioxidant related parameters, infrared, UV-Vis, volatile compound composition and one sensory – pivot profile). The efficiency of the models was evaluated using the explained variance, the slope of the eigenvalue exponential decay, while the configuration similarity between the models generated was evaluated using regression vector (RV) coefficients. At both low- and mid-level data fusion, the PCA approach resulted in a skewed sample configuration. The MFA models were less efficient than the PCA models, having a gradual distribution of the eigenvalue across the different model dimensions. As indicated by high RV coefficients between MFA and the individual blocks, the sample configurations resulting from the MFA were more representative than the PCA.

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

4.1 Introduction

The fields of metabolomics, engineering, and chemistry have a long history of working on data-orientated approaches to combining data sets from different sources, termed chemometric data fusion (Cocchi, 2019a; Gagolewski, 2015; Lahat, Adalı, & Jutten, 2015). Chemometric data fusion has a long history in other fields but is recent for agricultural sciences, and more so for the oenology field.

In order to compile a comprehensive account of the response of a wine to a certain phenomenon or influence, data from different sources is gathered; for example, wine can be profiled chemically and sensorially. Due to the complexity of sensory data matrices, the two are commonly discussed separately from one another and similarities are inferred. This has been the case for wine authenticity studies, whereby several measurements are taken and discussed separately (Arvanitoyannis, Katsota, Psarra, Soufleros, & Kallithraka, 1999). Although this works well for contained cases that have an application-based approach, cases that require collection of multiple responses across different stimuli or time require a data-orientated approach. Combining data sets from different sources creates a comprehensive profile of the behaviour of a product (in this case, wine) in response to said stimuli. This is in alignment with the motivation for the fourth industrial revolution which requires not just gathering large amounts of different data but looking at the data in smarter ways.

Putting together chemistry and sensory data has its own set of challenges. Data outputs for analytical chemistry instruments have made strides to develop standardized matrix arrangements for two to four-dimensional data (e.g. hyphenated techniques such as LC-MS/MS-TOF used in wine metabolomics, which uses multiple detectors) (Alañón, Pérez-Coello, & Marina, 2015). This required consolidation of statistical treatments (normalization of peak intensities for each peak) and alignment across the different detectors. On the contrary, the complex, and very often qualitative nature of sensory data is usually communicated through descriptive narratives. Although there are standardised statistical treatments for certain methods (Granato, de Araújo Calado, & Jarvis, 2014; Valentin et al., 2012), there is still a way to go to reach consensus on standardised matrix arrangements and outputs that encourage data consolidation. Due to the qualitative nature of many sensory evaluations, the assumptions made through statistical treatment of data are continuously debated and tend to be misconstrued as over-reaching or over-fitting (Valentin et al., 2012).

Data fusion is defined not simply as putting together, but rather as "integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source" (isif.org). Data fusion is classified under low-level, mid-level, and high-level, based on increasing complexity of the models and depending on the number of steps between the capturing of the raw data and the final fused model (Cocchi, 2019b). Low-level data fusion is the simplest form which usually uses the raw data with little pre-modelling processing.

Issues and challenges related to pre-modelling processing have previously been described for sensory (Brand, 2019; Valentin et al., 2012) and chemical analysis (López-Rituerto et al., 2012; Ragone et al., 2015; Rinnan, Berg, & Engelsen, 2009) in oenological applications. Low-level data fusion requires data sets to have compatible matrices, with compatible matrix order (2D, 3D, etc.) and at least one of the dimensional arrays (observations or variables) being the same (Cocchi, 2019b). Low-level data fusion models are often used as a pre-modelling step in mid- and high-level data fusion. From the low-level model, pre-modelling processing used include selection of variables or features, and the new matrix is then modelled. Mid-level data fusion is a systematic approach comprised of steps between the raw data and the final model. This may be due to differences in matrix dimensions, directed goals requiring feature selections, and/or pre-modelling processing. High-level, also called decision-level, data fusion, is the most complex and involves several directed steps. High-level data fusion strategies use both classical statistical analysis and machine learning techniques (Biancolillo, Boqué, Cocchi, & Marini, 2019). High-level supervised models have been used in oenology for prediction and calibration of oenological processes, such as the case of modelling ageing (Pereira et al., 2016). Recently, machine leaning techniques such as text-mining for qualitative sensory data (Valente, Bauer, Venter, Watson, & Nieuwoudt, 2018) and fuzzy logic (Ballabio, Todeschini, & Consonni, 2019; Silvestri et al., 2014) have been used for information mining in food applications.

In each of the three levels of data fusion, unsupervised modelling strategies in which the objective is data exploration may be used. These unsupervised modelling strategies look for patterns of grouping, similarity, or for the best-fit model. The objective of the data fusion may have a specific target in mind, in which case supervised modelling strategies are used. In the field of oenology, most reported cases of data fusion are supervised, with unsupervised methods being used as preliminary explorative steps that work to refine the final model (Biancolillo et al., 2019; Borràs et al., 2015). Supervised data fusion approaches are goal-orientated, by targeting and selecting only certain features from data blocks related to the phenomenon under investigation, reducing dimensionality and increasing predictive, discriminant, or classification power (Cocchi, 2019a). In trying to refine these supervised models, the data that does not contribute to increasing the regression coefficients is discarded. Conversely, unsupervised data fusion approaches retain most of the information captured whilst reducing the dimensionality.

The most commonly used unsupervised data fusion methods in oenology are principal component analysis (PCA) and multiple factor analysis (MFA) (Borràs et al., 2015; Pagés & Husson, 2005b). PCA is one of the most popular multivariate statistical tools in applied science (Salkind. J. & Kristin. R., 2007) which can be used for low or mid-level data fusion (by matrix concatenation), or as a pre-processing model. The focus of PCA is efficiency, accomplished by reducing the dimensions of a data set into more manageable dimensions called principal components, which make it easier to interpret complex data (McKillup, 2012). These principal components standardize the raw data to capture the essence of the correlations or covariance between the variable and the observations (the common vectors). It is because of these functions that PCA is an appropriate low-level data fusion model of choice in applied food science (Borràs et al., 2015). The disadvantages of PCA are its inability to handle data with high counts of zero or 'missing data' which can be an issue for certain sensory data and chemical instrument outputs (Borgognone, Bussi, & Hough, 2001; McKillup, 2012). In such cases, the raw data is revisited, and pre-processed manually or through statistical exclusions of some data before being modelled again. This rigorous approach can result in overfitting/ overcorrection that disregards the unsupervised intent of PCA modelling (Borràs et al., 2015).

MFA is another popular multivariate tool in applied food science, that goes beyond the simple matrix concatenation approach of PCA (Pagés & Husson, 2005a; J. J. Pagès, 2005). MFA has a multiblock data fusion approach that retains and standardizes each block before fusion, retaining the weight and contributions of the variables in each block to avoid any skewing by one data block (Abdi & Valentin, 2007). MFA is used for solving issues around the combining of sensory data, such as differences in variation between panels. Combinations of qualitative and quantitative data sets can thus be handled using pre-processing steps such as PCA and correspondence analysis (CA) before MFA modelling (J. Pagès, 2004). Rapid sensory methods capture data as ordinal, rating, or frequency matrices (Valentin et al., 2012), for which MFA is usually recommended to overcome the issues related to matrix compatibility (J. Pagès, 2004; Valentin et al., 2012).

The question still remains: which model is best? The performance of unsupervised data fusion models are evaluated comparatively and descriptively by looking at the distribution of the explained variance over different dimensions, grouping of samples when using cluster analysis or confidence ellipses (Le Dien & Pagès, 2003; J. Pagès, 2004; Pagés & Husson, 2005a). Recently, in order to compare the similarities between the sample configurations of different models, regression vector coefficients have been used (Abdi, 2007; Antúnez et al., 2015; Cadena et al., 2013; Fleming, Ziegler, & Hayes, 2015; Mafata, Brand, Panzeri, Kidd, & Buica, 2019).

This study explored data fusion strategies using low-level PCA and mid-level PCA and MFA models. The aim was to detail the rationale behind the different steps of data fusion, from data set curation to the evaluation of the final fused models. The data used in this study was based on the response of white wine to different storage conditions (Mafata et al., 2019). The data was captured and grouped under five blocks: antioxidant-related parameters (ARP), volatile compounds composition (VCC), UV-Vis spectrum, infra-red spectrum (IR), and sensory. The purpose of building these models was to create efficient, comprehensive, and representative data fusion models. The performance of the models was evaluated by looking at the distribution of the percentage explained variance (%EV) and the slope of the exponential decay of the eigenvalue across the different model dimensions, as a measures of information distribution. Comparisons between model sample configurations were evaluated using pair-wise regression vector (RV) coefficients. Issues surrounding model efficiency and redundancies between data blocks, and the representativeness of the data fusion model will be discussed.

4.2 Materials and Methods

4.2.1 Experimental design

The materials and methods related to the winemaking, wine treatments, sensory evaluation, and chemical analysis (oenological parameters, thiols, glutathione, major volatiles) have been previously published by Mafata, *et al.* (Mafata, Brand, Panzeri, *et al.*, 2019). In brief, the experiment focused on the stability of wines at various temperatures and for different time periods. Samples belonged to two cultivars (Chenin Blanc and Sauvignon Blanc) from six wineries each (twelve sample sets in total). Each sample set consisted of seven wines corresponding to the experimental storage conditions (*i.e.* no storage time/control, three- and nine-months storage; three temperatures: 15°C, 25°C and room temperature).

4.2.2 Sensory data methodology

The descriptive part of the sensory data methodology (panel parameters and instructions) was previously published in Mafata et al. (Mafata et al., 2019). For the purpose of the current study, some relevant aspects are described here. The sensory method chosen for this experiment was Pivot© Profile (PP) (Thuillier, Valentin, Marchal, & Dacremont, 2015). PP is a verbal, reference-based method that collects information about the attributes, per sample, relative to the pivot (Valentin et al., 2012), in this case the control sample. The data was captured as a rating of either +1 (more than pivot) or -1 (less than pivot) and for attributes that were not mentioned, a rating of zero was given. The raw data were captured per data set, with judges and repeats kept separate and not concatenated further (Lelièvre-Desmas, Valentin, & Chollet, 2017).

Linguistic and semantic reduction of terms were performed manually resulting in a total of 200 attributes. Statistical consolidation was then done for each sample set separately. Each attribute was summed across judges and repeats, translated into positive ratings, and zero-sum terms excluded. The positive translation was done to convert the data from rating to frequency so that the modelling could be done by CA (Thuillier et al., 2015). Terms with less than 5% citations were removed, resulting in 29 to 36 attributes per sample set.

4.2.3 Chemical data collection and capturing

The chemical data categorised under volatiles (VCC data set: thiols, major volatiles) and antioxidant-related parameters (ARP data set: Colour intensity (CI, A520 + A420), colour hue (CH, A520/A420), total phenolics (A280), hydroxycinnamic acids (A320) and browning (A420), CIElab parameters, glutathione, total and free sulphur dioxide) were previously discussed (Mafata et al., 2019). Ultraviolet-visible light spectrophotometric scans (UV-Vis data set) were run from 280 nm to 780 nm (in 1 nm increments) in triplicate on a Thermo Scientific Multiskan GO 1510-02586 microplate spectrophotometer. Infra-red spectra measurements (IR data set), in the mid-infrared range (4000-600 cm^{-1}) were collected on the Alpha-P ATR FT-MIR spectrometer (Bruker Optics, Ettlingen, Germany). Each sample was scanned at a resolution of 4 cm-1 and at a scanning velocity of 7.5 kHz, averaged over 64 scans to give a final reading. Instrumental control and data capturing were carried out using OPUS software (OPUS v. 7.0 for Microsoft, Bruker Optics, Ettlingen, Germany).

4.2.4 Statistical analysis

Multivariate analysis was performed separately for each winery and each cultivar, each sample set consisted of seven wines (2.1). The data were divided into five blocks based on properties and modality of acquisition: volatile compounds (VCC), antioxidant-related parameters (ARP), UV-Vis spectra (UV-Vis), infra-red spectra (IR), and sensory data (Table 4. 1). In other words, each block consisted of twelve data sets and each data set contributed to five blocks. The raw sensory data was submitted to Correspondence Analysis (CA) and the standardised deviates matrix was used for data fusion. All PCA analyses in this study were based on the generalized Pearson correlation coefficient with standard univariate scaling applied to all measurements before modelling. MFA was performed on correlation matrices of the chemistry data sets (observations vs. variables) and the latent variables of the sensory data. The data blocks were first standardised by PCA and then MFA was performed (Abdi & Valentin, 2007). For each model, an exponential decay curve was plotted using eigenvalues for each dimension and the slope calculated using Microsoft Excel (Excel Office 365, version 2002, Microsoft Corp., United States). Configurational similarities for all score plots were calculated using pair-wise regression vector (RV) coefficients (Abdi, 2007) and infer topology (iTOP) RV between the PCA and MFA data fusion models (Aben et al., 2018) (Figure 4.1). Statistical calculations and modelling were performed using Statistica™ 13 (TIBCO, Dell software, Inc., Teas, United States).

Table 4.2: Five block low-level and mid-level data fusion approaches using principal component analysis (PCA) and multifactorial analysis (MFA).

Level	Blocks	Input (raw data)			Pre-processing	Modelling		Model output			Interpretation		
		Description	Matrix type	Values		Modelled matrix	Model	New matrices	New matrix row	New matrix column	Evaluation parameter	Visualization aids	
Individual data blocks	Antioxidant-related parameters (ARP)	Discreet measurements	Correlation	Concentration	none	raw data	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
	Infra-red (IR)	Spectral	Continuous	Transmittance	MSC + 1st deriv transformations	raw data	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
	Ultra-violet visible light (UV-Vis)	Spectral	Continuous	Absorbance	none	raw data	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
	Volatile compound composition (VCC)	Discreet measurements	Correlation	Concentration	none	raw data	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
	Sensory	Pivot profile reference-based method	Rating	Rating	Conversion to frequency matrix	Positive FoC	CA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
								Standardized deviates	Observations	Variables			
Low-level	ARP + IR + UV-Vis + VCC	Data fusion	Mixed	Mixed	matrix concatenation	Concatenated matrix	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
Mid-level	ARP + IR + UV-Vis + VCC + sensory	Data fusion	Mixed	Mixed	matrix concatenation *Sensory latent variables (standardized deviates) from CA	Concatenated matrix	PCA	Scores	Observations	Principal components	Pair-wise RV coefficients of scores	Scores plot	Loadings plot
	ARP + IR + UV-Vis + VCC + sensory	Data fusion	Multiblock	Mixed	PCA per block *Sensory latent variables (standardized deviates) from CA	Multiblock standardized deviates from individual PCA	MFA	Scores	Observations	MFA dimensions	Pair-wise RV coefficients of scores	Scores Factor Map	Scores cluster plot
								Loadings	Blocks	MFA dimensions	Pair-wise RV coefficients of loadings	Block Factor Map	Block cluster plot
											iTop-RV coefficients between the data fusion models (PCA vs MFA)		

4.3. Results and discussions

The fusion of the five data blocks in this study (VCC, ARP, IR, UV-Vis, and sensory) was unsupervised and explorative, from low-level to mid-level data fusion strategies in increasing complexity (Table 4.1). This section is arranged according to both the complexity of the conceptualisation of the approach as well as the operational order taken in fusing the data blocks.

4.3.1 Curation of data blocks

4.3.1.1 Assessment of pre-modelling processing

It is important to first inspect the data blocks specifically for the purposes of data fusion since this will dictate which type and which level of fusion is needed; the decisions taken might be different to the ones when data fusion is not the purpose (Engel et al., 2013). When looking at pre-modelling processing methods in view of data fusion, two criteria were considered in this study, namely matrix compatibility and signal correction.

Matrix compatibility is an important eligibility criterion for low-level data fusion strategies (Smilde & Van Mechelen, 2019). If matrices are incompatible, then pre-modelling processing must be done. The chemistry data sets (ARP, VCC, UV-Vis, and IR) were captured as compatible correlation matrices (Table 4.1) and, thus, could be combined using either low-level or higher-level data fusion strategies. In contrast, in order to obtain a compatible matrix for the sensory data, the standardised deviates matrix was obtained from the CA model (Table 4.1). The raw sensory data was captured as rating data, the matrix of which consisted of 0, 1, and -1 ratings. These types of data sets cannot be modelled using PCA since they contain large counts of zero measurements (Salkind. J. & Kristin. R., 2007).

With regards to signal correction, spectral pre-processing is often considered for UV-Vis and IR spectral data blocks, and included as toolkits for most software (Gishen, Dambergs, & Cozzolino, 2005; Umetrics, 2012). In the case of the UV-Vis data block, high model efficiency (%EV) was taken as good indicator for proceeding with the raw UV-Vis data for the fusion without the need for pre-processing.

Since IR had lower %EV and pair-wise RV coefficients (i.e. between scores of the PCA models with vs. without pre-processing), pre-modelling processing was considered to try and better these model evaluation parameters. Infra-red spectral data are prone to spectral irregularities which are categorised under two phenomena, namely scattering and base line irregularities (Rinnan et al., 2009). The mathematical conversions done to correct these phenomena fall under these two categories. Infra-red data regularly use multiplicative scatter correction (MSC) for scatter, first derivative transformations for baseline corrections, and combinations of the two (Rinnan et al., 2009). In this section, the raw data, MSC, 1st derivative, and combinations of MSC with 1st derivative were investigated as potential methods of pre-processing infra-red data.

The impact of the transformations on the efficiency of the PCA models were evaluated by %EV (Table 4.2) and any effect on the sample set configuration was evaluated through pairwise RV coefficients between the PCA models of the raw and the transformed data (Table 4.1). The raw data produced PCA models with the highest efficiency, with an average cumulative %EV for the first two principal components 84±9 for CB and 70±6 for SB. All other pre-processing transformations lowered the efficiency of the models, with some exceptions; the MSC increased the efficiency of the PCA models of PDB and KZC CB sample sets by 7% and 4%, respectively. The KZC CB sample set model efficiency was increased by the pre-processing methods, except for the first derivative transformation. KZC had the second highest %EV of all the wineries; the 4% increase was thus relatively negligible, and inspection of the spectra showed no obvious faults.

Table 4.2: Cumulative percentage explained variance (%EV) for the first two principal components of infra-red spectral raw data and its mathematical transformations using multiplicative scatter correction (MSC) and first derivative (1st deriv), and their combinations.

		raw	1st deriv	MSC	1st deriv MSC	MSC 1st deriv
Chenin Blanc	AVN	82	52	73	51	53
	CDB	72	57	61	52	52
	DTK	97	62	97	72	73
	FRV	76	52	68	50	53
	KZC	96	79	100	100	100
	PDB	81	54	88	50	60
	average	84	59	81	63	65
	stdev	9	9	15	18	17
Sauvignon Blanc	AVN	72	43	55	40	39
	CDB	74	54	63	51	51
	DTK	63	45	46	39	40
	FRV	74	50	51	40	41
	KZC	62	43	51	38	39
	PDB	77	45	52	38	39
	average	70	47	53	41	42
	stdev	6	4	5	5	4
Overall	low	62	43	46	38	39
	high	97	79	100	100	100
	average	77	53	67	52	53
	stdev	10	10	18	17	17

RV coefficients showed high configurational similarities between the different pre-processed models vs. the raw data, except for the KZC CB sample set (Supplementary Table 4.1), meaning that generally the pre-processing had very little effect on the sample configuration. The raw data set had the lowest RV coefficients, ranging from 0.73 to 0.95 for CB (0.84±0.06, ave±SD) and 0.70 to 0.90 for SB (0.78±0.06). This means that the configurations of the transformed spectra were more similar to each other than to the raw data. However, this was a negligible difference in configurations, with a maximum 15% increase in RV coefficients on average.

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

4.3.1.2 Performance of individual block models

The chemistry data blocks had each a set number of variables (UV-Vis 501 wavelengths, ARP 14 parameters, VCC 34 compounds, and IR 879 wavenumbers); the sensory data had a varying number of variables since the number of attributes was different for each data set after pre-processing. A comparative exploration of the models' packing efficiency was done using the %EV (Supplementary Table 4.2) and the configurational similarity of the scores (seven samples per set) was calculated through pairwise RV coefficients between the data sets (Supplementary Table 4.3). Overall, the UV-Vis models were the most efficient, with cumulative %EV ranging from 78 to 99, and an average of 91±7 for the first two PCs. ARP was the second most efficient (75 to 94 %EV, 84±5) followed by IR (64 to 98%EV, 78 ±10) and VCC (72 to 83%EV, 77±3). Sensory had the lowest cumulative %EV (55 to 78%EV, 68 ±6) for the first two dimensions of the CA, which is an inherent characteristic of holistic techniques such as sensory analysis (Valentin et al., 2012).

The sample configurations of UV-Vis and ARP were the most similar, with RV coefficients ranging from 0.78 to 0.93 for CB and 0.73 to 0.93 for SB. This is understandable since compounds with antioxidant properties can absorb UV-Vis energy (Stevenson, 2005). Additionally, the CIE lab and other colour indices listed in the ARP data block were calculated from specific measurements in the UV-Vis spectrum. RV coefficients for ARP vs. VCC were the second highest, ranging from 0.55 to 0.83 for CB and 0.45 to 0.82 for SB. RV coefficients for UV-Vis vs. VCC were lower compared to those of ARP vs. VCC, ranging from 0.31 to 0.81 for CB and 0.33 to 0.62 for SB. RV coefficients were very low between IR and the other chemistry data blocks (UV-Vis, ARP, and VCC), ranging from 0.10 to 0.71 for CB and 0.21 to 0.79 for SB. RV coefficients between IR and sensory were higher, ranging from 0.38 to 0.86 for CB and 0.51 to 0.72 for SB. RV coefficients between sensory and UV-Vis were poor, ranging from 0.59 to 0.74 for CB and 0.43 to 0.79 for SB. RV coefficients were higher between sensory and VCC ranging from 0.60 to 0.87 for CB and 0.60 to 0.85 for SB. Since the sensory method evaluated only the aroma of the wines, it is understandable that it resulted in higher configurational similarity with the VCC data set.

4.3.2 Low-level data fusion

Low-level fusion involves the simple concatenation of raw data with compatible matrix dimensions (Cocchi, 2019b; Ríos-Reina, Callejón, Savorani, Amigo, & Cocchi, 2019). The ARP, VCC, UV-Vis and IR data blocks had compatible observations vs. variables correlation matrices, and thus could be fused using low-level strategies. In order to fuse the sensory with the chemistry data, a mid-level data fusion strategy had to be employed; this is explored in the next section. The four chemistry data blocks were first concatenated into one correlation matrix of seven observations (for each sample set) vs. 1428 variables (corresponding to the sum of variables for the chemistry data blocks) and modelled by PCA.

It has previously been shown that the individual models for the four chemistry data blocks were highly efficient, with most of the explained variance captured within the first two principal components (Section 4.3.1.2, Supplementary Table 4.2). Comparatively, the low-level PCA fusion model was less efficient (Table 4.3), hence a more in-depth exploration of the data distribution was needed to assess the model performance. The overall stress in the model and the slope of the exponential decay in the stress across the principal components (Table 4.3) were used to evaluate the model efficiency (Salkind. J. & Kristin. R., 2007).

The 1428 variables were fitted over six principal components and the stress onto an exponential curve with R2 of between 0.81 and 0.99. CB had more efficient models compared to SB as measured by the slope, which ranged from 0.44 to 0.88 for CB and 0.38 to 0.55 for SB (Table 4.3). Approximately 80% of the explained variance was achieved within the first three principal components, which was less efficient than the individual models (Supplementary Table 4.2). This is characteristic of multimodal data fusion, due to the increased number of variables and the different types of data sources (Cocchi, 2019a). KZC CB data set had the highest performance indicators again, with a slope of 0.87 (R2=0.95) and a cumulative %EV of 89 for the first two principal components (Table 4.3).

Table 4.3: Performance parameters and stress distribution for the low-level data fusion of ARP, VCC, UV-Vis, and IR chemical data by principal component analysis (PCA).

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

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, IR – infrared, PCA – principal component analysis, %EV – percentage explained variation.

Due to the concatenated (one matrix) nature of the PCA data fusion strategy, it is difficult to attribute the performance of the model to any one of the data blocks. In order to try and address issues of redundancy between the data blocks in this low-level strategy, the sample configurations resulting from the PCA on the concatenated data were compared to the individual data sets' PCAs using RV coefficients (Supplementary Table 4.4). Although previously the KZC CB sample set was an exception in the individual PCA models, the low-level PCA data fusion model is not since it has similar RV coefficients patterns described for the other sample sets.

It may be misconstrued that the concatenated model is likely to be skewed by the most variable dense data block, in this case the IR (879 variables); and, since this data block had the highest RV coefficients (IR vs. low-level PCA), the hypothesis seemed to have some support. IR vs. low-level PCA had RV coefficients ranging from 0.88 to 0.96 for CB and 0.83 to 0.95 for SB data sets. As previously discussed, the sample configuration of the IR data block was different from the other data blocks (Section 4.3.1.2, Supplementary Table 4.3). A look at the RV coefficients between the low-level PCA model and the other data blocks, showed that the sample configurations were mainly case-specific, no one-fits-all generalization of the patterns could be applied for VCC and ARP data sets. Although UV-Vis data block had the second highest number of variables (507), it did not always have the second the highest RV coefficient. This meant that the number of variables was not the most influential factor on the sample configuration of the fusion model, but rather the amount of information the technique carries. As previously discussed, IR is an information-rich technique and infra-red activity is a more general property of organic molecules than UV-Vis (Robinson, 2017).

4.3.3 Mid-level data fusion

4.3.3.1 Principal Component Analysis (PCA)

In order to incorporate the sensory results into fused models, the data had to be in a format compatible with the rest of the data blocks (Cocchi, 2019a; McKillup, 2012). To achieve this, the standardized deviates (standardized co-ordinates) from the CA model of the sensory data were used. These were added to chemistry data blocks by concatenation and the new matrix was modelled by (mid-level) PCA (Table 4.1). The distribution of the stress and performance indicators of the model are listed in Table 4.4. As expected, the increased dimensionality due to the addition of sensory data resulted in decreased model efficiency compared to both the individual data blocks and the low-level fusion PCA. The CB models were the most efficient, with the slope of the exponential decay curves ranging from 0.43 to 0.83 (R2 > 0.90) compared to those for SB ranging from 0.37 to 0.53 (R2 > 0.80). The KZC CB mid-level PCA model was the most efficient, with a slope of 0.82 (R2 = 0.94) and a cumulative %EV of 89 for the first two principal components.

In the data curation section (Section 4.3.1.2), it was noted that the sensory data model was the least efficient, having the lowest cumulative %EV of all the data blocks. However, the concatenation of the sensory data with the chemistry data sets did not lower the cumulative %EV compared to the low-level PCA. On average, the cumulative %EV across the model dimensions decreased by 1% from low-level to mid-level PCA. As the addition of the sensory data block is valuable to the overall information, a compromise in model efficiency was acceptable.

The similarity in sample configurations were again assessed using RV coefficients (Table 4.5). The addition of sensory data resulted in lower RV coefficients between the mid-level PCA vs. the PCA for individual blocks compared to the RV coefficients between the low-level data fusion PCA vs. individual data blocks.

Table 4.4: Performance indicators and stress distribution of the mid-level PCA data fusion model of infra-red, antioxidant-related, UV-Vis, volatile compounds, and sensory data sets.

						Cumulative %EV per PC					
	Winery	Observations	Total stress (eigenvalue)	Slope	R²	F1	F2	F3	F4	F5	F6
Chenin Blanc	AVN	1458	601	0.45	0.97	41	68	81	89	95	100
	CDB	1463	595	0.53	0.99	41	67	83	91	97	100
	DTK	1461	747	0.83	0.96	51	84	92	97	99	100
	FRV	1463	698	0.43	0.92	48	68	80	88	95	100
	KZC	1458	968	0.82	0.94	66	89	95	97	99	100
	PDB	1461	847	0.54	0.90	58	77	85	91	97	100
Sauvignon Blanc	AVN	1459	661	0.45	0.94	45	69	81	89	95	100
	CDB	1457	721	0.53	0.97	50	73	84	92	97	100
	DTK	1464	544	0.37	0.93	37	64	77	86	93	100
	FRV	1463	561	0.53	0.93	38	75	84	92	97	100
	KZC	1464	805	0.40	0.80	55	69	78	87	95	100
	PDB	1458	661	0.45	0.94	45	69	81	89	95	100
Range	min	1457	544	0.37	0.80	37	64	77	86	93	100
	max	1464	968	0.83	0.99	66	89	95	97	99	100

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, IR – infrared, PCA – principal component analysis, %EV – percentage explained variation.

Since the fusion model is a composition of different data blocks coming from measurements of the different properties of wine, a resulting model that has a unique sample configuration was expected. Although for a concatenated matrix the within-model redundancy cannot be calculated, the RV coefficients (mid-level PCA vs. individual blocks range 0.52-0.88) could be considered an indicator of relatively low redundancy. The exception was once more the IR data block. As previously discussed in section 4.3.1.2, the IR data block provided the most unique sample configuration pattern compared to the other data blocks, indicated by low RV coefficients (IR vs. other data blocks, Supplementary Table 4.3). The IR sample configuration was the most similar to that of the mid-level PCA fusion model, indicated by high RV coefficients (mid-level PCA vs. IR) ranging from 0.88 to 0.96 for CB and 0.83 to 0.96 for SB data sets. The pattern of RV coefficients between the mid-level PCA data fusion model and the other individual data blocks could not be generalized. The patterns were case-specific and unique for each data set. Much like the IR data block, due to it nature the sensory data contains a unique profile of the wine, but although the sensory experience is holistic, given to the method used the data captured was not. Pivot profile is a comparative method, and not a profiling method such as CATA, that includes a comprehensive list of attributes (Valentin et al., 2012). Compared to IR, which is unique and information-rich, sensory in this case in unique but not as information-rich.

Table 4.5: Pairwise regression vector coefficients (p ≤ 0.01) for the PCAs of the individual data vs the mid-level PCA fused model of all five data blocks.

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

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, PCA – principal component analysis. *Cumulative percentage explained variance of the first two dimensions of the correspondence analysis.

4.3.3.2 Multiple Factor Analysis (MFA)

Unlike the PCA which aims to reduce the dimensionality and produce the most efficient model, the MFA seeks to create/build the most representative model of the relationships between blocks of data (Abdi & Valentin, 2007). The figures of merit related to the performance of the MFA models are shown in Table 4.6. As a multiblock analysis, the stress calculated on the MFA is relative to the different blocks and not the individual variables within each block (Abdi & Valentin, 2007); as such, the eigenvalues are lower than those of the PCA data fusion models. The exponential decay curves had R2 ranging from 0.84 to 0.99, except for PDB SB that had an R2 of 0.71. Generally, the models had low efficiency; CB had higher efficiency as evaluated by the slopes (0.35 to 0.47) than SB (0.27 to 0.37). For all models, the stress was distributed gradually across the different dimensions with less than 80 %EV accumulated over the first three dimensions.

Table 4.6: Stress distribution over components in the MFA data fusion of multimodal data in sets of several for six different wineries.

					Cumulative %EV per PC					
	Winery	Total stress (eigenvalue)	Slope	R²	C1	C2	C3	C4	C5	C6
Chenin Blanc	AVN	9.8952	0.43	0.96	40	67	79	89	95	100
	CDB	9.7308	0.35	0.94	40	61	75	86	93	100
	DTK	9.0578	0.43	0.99	41	64	78	89	96	100
	FRV	9.7637	0.38	0.96	42	63	77	87	94	100
	KZC	8.9517	0.47	0.97	42	68	82	90	96	100
	PDB	9.0879	0.37	0.84	49	64	76	86	95	100
Sauvignon Blanc	AVN	9.3392	0.30	0.85	39	60	72	82	92	100
	CDB	10.6642	0.37	0.99	33	58	76	86	94	100
	DTK	10.3854	0.33	0.94	38	57	75	85	93	100
	FRV	9.3328	0.35	0.93	39	63	75	85	94	100
	KZC	10.7258	0.32	0.98	33	58	73	84	93	100
	PDB	9.21612	0.27	0.71	43	56	70	82	93	100
Range	min	8.9517	0.27	0.71	33	56	70	82	92	100
	max	10.7258	0.47	0.99	49	68	82	90	96	100

An MFA model generates new weights for the different data blocks, relative to each other, and can thus show the correlations between different data blocks. This means that the MFA sample configuration is most representative of all the data blocks and is not skewed by any individual data block (as might be the case with low/mid-level data fusion PCA). RV coefficient values can be calculated between the sample configurations of the data blocks after weighing (Supplementary Table 4.5). The RV coefficients for the MFA (vs. individual data blocks) were higher than those of the mid-level data fusion PCA (vs. individual data blocks), ranging from 0.52

to 0.95 for CB and 0.64 to 0.92 for SB. RV coefficients between MFA vs. IR data block (ranging from 0.55 to 0.88 for CB and 0.64 to 0.87 for SB) were lower compared to the other data blocks (ranging from 0.76 to 0.95 for CB and 0.69 to 0.92 for SB). This is unlike the results for the PCA data fusion models (low and mid-level) in which the RV coefficients between the PCA data fusion model vs. IR data block were the highest compared to the other data blocks (Tables 4.5 and supplementary Table 4.4). This is indicative of how the number and nature of the variables from the IR data block had a skewing effect on the PCA data fusion models. This means that in the concatenated matrices, the IR data block influenced the sample configuration the most. This could not be directly demonstrated in the case of PCA due to the nature of the statistical analysis.

The sample configurations of the mid-level PCA and MFA fusion models were calculated using the conventional RV coefficient and infer topology (iTOP) calculation of the RV (Table 4.7). The infer topology (iTOP) RV reportedly takes into account the redundancy between data blocks and skewing by any one data block (Aben et al., 2018). Although the iTOP RV coefficients were slightly lower than the conventional RV coefficient, they were similar. All RV coefficients were higher than 0.70 indicating very high similarity between the two approaches (iTOP vs. conventional) but since the two data fusion models contain the same original data this was expected. The burden now shifts to the 30% dissimilarity between the data fusion approaches.

Table 4.7: Pairwise regression vector coefficients (p ≤ 0.01) between PCA and MFA for the mid-level data fusion of five-modal data sets.

		RV coefficient	**iTOP (inferred topology) RV**
Chenin Blanc	AVN	0.82	0.70
	CDB	0.82	0.75
	DTK	0.80	0.62
	FRV	0.94	0.93
	KZC	0.85	0.80
	PDB	0.96	0.95
Sauvignon Blanc	AVN	0.78	0.77
	CDB	0.93	0.92
	DTK	0.81	0.70
	FRV	0.89	0.85
	KZC	0.84	0.79
	PDB	0.81	0.81

PCA – principal component analysis, MFA - multiple factor analysis, RV – regression vector.

4.4 General discussion

The case chosen to illustrate the stepwise approach to data fusion had its particularities originating from the type of sensory method that generated the data and the fact that data sets were first considered separately due to the original experimental design. However, these types of results are quite common in wine evaluation, where one or more analytical chemistry techniques are used in addition to (usually) one sensory method. Different steps and levels of data modelling for the purposes of data fusion have been presented, from individual data blocks, low-level, mid-level data fusion to multiblock data fusion. In assessing the different models, it is important to use multiple evaluation parameters that take into account different aspects of the models. In this study, the models' performance were evaluated by looking at the distribution of the data over dimensions and the slope of the exponential decay as indicators of model efficiency; the RV coefficients were used to evaluate the representativeness of the fusion models and evaluate redundancy in the cases where other parameters could not be used.

Low-level data fusion is generally appropriate for data blocks with only a small number of variables, since finding patterns in correlations between a large number of variables can be tedious and the visual aids offer very little assistance with the complex interpretations (McKillup, 2012). The low-level and mid-level PCA fusion models did not offer any information on the within-model correlations between data blocks. Although the models were highly efficient, they were not representative. Due to the incompatibility of the sensory data matrix with the four chemistry data blocks, low-level PCA data fusion was not as comprehensive as the mid-level strategies. Although the addition of the sensory data block resulted in slightly lower model efficiency, the sensory aspect is adding to the overall informational value and comprehensiveness of the data fusion model; thus, the compromise in model efficiency must be made. For cases where the model efficiency is drastically lowered by the inclusion of a data block, the influence of the additional block must be further investigated. This can be done by revisiting the pre-modelling processing to "clean" the data.

Mid-level PCA data fusion models were skewed by the information dense IR data block. This was revealed by lower RV coefficients between for PCA vs individual blocks compared to PCA vs IR data block. Mid-level PCA sample configuration was thus an unrepresentative data fusion model of all blocks. Mid-level MFA models were less efficient than PCA models but were more representative of the commonality between data blocks, indicated by high RV coefficients (the models had sample configurations more representative of all the data blocks). Although the PCA fusion models were highly efficient (high %EV and slope), this was rather indicative of overfitting of the data since the models were also found to be unrepresentative due to skewing by the information-rich IR data block. Hence, by comparison, MFA proved to be less biased and more representative of the individual data blocks.

4.5 Conclusion

The aim of this study was to explore low-level PCA and mid-level PCA, and MFA data fusion strategies. The study evaluated model efficiency (%EV and slope of the exponential decay in stress) and model representativeness (within-model and between-model pairwise RV coefficients). Using these parameters, issues of overfitting of data and redundancy between the different data blocks were inferred. Adding more data, especially data of a different nature, resulted in reduced model efficiency. Since the addition of more data of different variation is the motivation of data fusion, the model efficiency was found to an ineffective evaluation parameter for data fusion models. The RV coefficients were a more effective parameter for evaluating data fusion model performance. However, RV could not be used within low/mid-level PCA data fusion models; only the MFA multiblock strategy offered this feature.

It is for these reasons, that for large data sets such as those presented in this study, MFA should be considered a more appropriate unsupervised data fusion strategy.

References

Abdi, H. (2007). RV Coefficient and Congruence Coefficient. In Encyclopedia of Measurement and Statistics. Retrieved from http://www.utd.edu/

Abdi, H., & Valentin, D. (2007). Multiple Factor Analysis (MFA). In Encyclopedia of Measurement and Statistics. Retrieved from http://www.utd.edu/

Aben, N., Westerhuis, J. A., Song, Y., Kiers, H. A. L., Michaut, M., Smilde, A. K., & Wessels, L. F. A. (2018). iTOP : inferring the topology of omics data. 988–996. https://doi.org/10.1093/bioinformatics/bty636

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

Antúnez, L., Salvador, A., de Saldamando, L., Varela, P., Giménez, A., & Ares, G. (2015). Evaluation of Data Aggregation in Polarized Sensory Positioning. Journal of Sensory Studies, 30(1), 46–55. https://doi.org/10.1111/joss.12135

Arvanitoyannis, I. S., Katsota, M. N., Psarra, E. P., Soufleros, E. H., & Kallithraka, S. (1999). Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). Trends in Food Science & Technology, 10, 321–336.

Ballabio, D., Todeschini, R., & Consonni, V. (2019). Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data. In Data Handling in Science and Technology (Vol. 31, pp. 129–155). https://doi.org/10.1016/B978-0-444-63984-4.00005-3

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

Borgognone, M. G., Bussi, J., & Hough, G. (2001). Principal component analysis in sensory analysis: covariance or correlation matrix? Food Quality and Preference, 12, 323–326. https://doi.org/https://doi.org/10.1016/S0950-3293(01)00017-9

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

Brand, J. (2019). Rapid sensory profiling methods for wine: Workflow optimisation for research and industry applications. Stellenbosch University.

Cadena, R. S., Cruz, A. G., Netto, R. R., Castro, W. F., Faria, J. de A. F., & Bolini, H. M. A. (2013). Sensory profile and physicochemical characteristics of mango nectar sweetened with high intensity sweeteners throughout storage time. Food Research International, 54(2), 1670–1679. https://doi.org/10.1016/J.FOODRES.2013.10.012

Cocchi, M. (2019a). Data fusion methodology and applications (Vol. 31; M. Cocchi, Ed.).

Cocchi, M. (2019b). Introduction: Ways and Means to Deal With Data From Multiple Sources. In Data Handling in Science and Technology (Vol. 31, pp. 1–26). https://doi.org/10.1016/B978-0-444-63984-4.00001-6

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

Fleming, E. E., Ziegler, G. R., & Hayes, J. E. (2015). Check-all-that-apply (CATA), sorting, and polarized sensory positioning (PSP) with astringent stimuli. Food Quality and Preference, 45, 41–49. https://doi.org/10.1016/j.foodqual.2015.05.004

Gagolewski, M. (2015). Data fusion. Theory, methods, and applications (O. Hryniewicz, J. Mielniczuk, W. Penczek, & J. Waniewski, Eds.). https://doi.org/10.1109/isrcs.2012.6309295

Gishen, M., Dambergs, R. G., & Cozzolino, D. (2005). Grape and wine analysis - enhancing the power of spectroscopy with chemometrics. Australian Journal of Grape and Wine Research, 11(3), 296–305. https://doi.org/10.1111/j.1755-0238.2005.tb00029.x

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

Iorgulescu, E., Voicu, V. A., Sârbu, C., Tache, F., Albu, F., & Medvedovici, A. (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 extrac. Talanta, 155, 133–144. https://doi.org/10.1016/j.talanta.2016.04.042

Lahat, D., Adalı, T., & Jutten, C. (2015). Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. Institute of Electrical and Electronics Engineers, 103(9), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697ï

Le Dien, S., & Pagès, J. (2003). Hierarchical Multiple Factor Analysis: application to the comparison of sensory profiles. Food Quality and Preference, 14(5–6), 397–403. https://doi.org/10.1016/S0950-3293(03)00027-2

Lelièvre-Desmas, M., Valentin, D., & Chollet, S. (2017). Pivot profile method: What is the influence of the pivot and product space? Food Quality and Preference, 61, 6–14. https://doi.org/10.1016/j.foodqual.2017.05.002

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

Mafata, M., Brand, J., Panzeri, V., Kidd, M., & Buica, A. (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(February), 108515. https://doi.org/10.1016/j.foodres.2019.108515

McKillup, S. (2012). Statistics explained : an introductory guide for life scientists. Cambridge University Press.

Pagès, J. (2004). Multiple factor analysis: Main features and application to sensory data. Revista Colombiana de Estadistica, 27(1), 1–26.

Pagés, J., & Husson, F. (2005a). Multiple factor analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data. Journal of Chemometrics, 19(3), 138–144. https://doi.org/10.1002/cem.916

Pagés, J., & Husson, F. (2005b). Multiple factor analysis with confidence ellipses: A methodology to study the relationships between sensory and instrumental data. Journal of Chemometrics, 19(3), 138–144. https://doi.org/10.1002/cem.916

Pagès, J. J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. Food Quality and Preference, 16(7), 642–649. https://doi.org/10.1016/j.foodqual.2005.01.006

Pereira, A. C., Carvalho, M. J., Miranda, A., Leça, J. M., Pereira, V., Albuquerque, F., … Reis, M. S. (2016). Modelling the ageing process: A novel strategy to analyze the wine evolution towards the expected features. Chemometrics and Intelligent Laboratory Systems, 154, 176–184. https://doi.org/10.1016/J.CHEMOLAB.2016.03.030

Ragone, R., Crupi, P., Piccinonna, S., Bergamini, C., Mazzone, F., Fanizzi, F. P., … Antonacci, D. (2015). Classification and Chemometric Study of Southern Italy Monovarietal Wines Based on NMR and HPLC-DAD-MS. Food Sci. Biotechnol, 24(3), 817–826. https://doi.org/10.1007/s10068-015-0106-z

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

Ríos-Reina, R., Callejón, R. M., Savorani, F., Amigo, J. M., & Cocchi, M. (2019). Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. Talanta, 198, 560–572. https://doi.org/10.1016/j.talanta.2019.01.100

Robinson, J. W. (2017). Practical handbook of spectroscopy. In Practical Handbook of Spectroscopy. https://doi.org/10.1201/9780203742433

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

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

Smilde, A. K., & Van Mechelen, I. (2019). A Framework for Low-Level Data Fusion. In Data Handling in Science and Technology (Vol. 31, pp. 27–50). https://doi.org/10.1016/B978-0-444-63984-4.00002-8

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

Thuillier, B., Valentin, D., Marchal, R., & Dacremont, C. (2015). Pivot© profile: A new descriptive method based on free description. Food Quality and Preference, Vol. 42, pp. 66–77. https://doi.org/10.1016/j.foodqual.2015.01.012

Umetrics, M. (2012). User Guide to SIMCA 13. Umetrics, 13, 1–661. https://doi.org/10.1007/SpringerReference_27988

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

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

Chapter 4 Supplementary

Table 4.1: Pairwise regression vector coefficients ($p \leq 0.01$) for infra-red spectral raw data and its mathematical transformations using multiplicative scatter correction (MSC) and first derivative (1st deriv), and their combinations.

		Chenin Blanc					Sauvignon Blanc				
		1st deriv	1st deriv MSC	MSC	MSC 1st deriv	raw	1st deriv	1st deriv MSC	MSC	MSC 1st deriv	raw
	1st deriv	1	0.99	0.81	0.98	0.83	1	0.97	0.89	0.97	0.84
	1st deriv MSC	0.99	1	0.82	0.99	0.78	0.97	1	0.90	0.99	0.74
AVN	MSC	0.81	0.82	1	0.87	0.88	0.89	0.90	1	0.92	0.83
	MSC 1st deriv	0.98	0.99	0.87	1	0.82	0.97	0.99	0.92	1	0.74
	raw	0.83	0.78	0.88	0.82	1	0.84	0.74	0.83	0.74	1
	1st deriv	1	0.98	0.92	0.98	0.89	1	0.98	0.91	0.98	0.86
	1st deriv MSC	0.98	1	0.94	1.00	0.78	0.98	1	0.90	0.99	0.76
CDB	MSC	0.92	0.94	1	0.95	0.75	0.91	0.90	1	0.93	0.77
	MSC 1st deriv	0.98	1.00	0.95	1	0.78	0.98	0.99	0.93	1	0.76
	raw	0.89	0.78	0.75	0.78	1	0.86	0.76	0.77	0.76	1
	1st deriv	1	0.88	0.81	0.89	0.89	1	0.98	0.92	0.97	0.89
	1st deriv MSC	0.88	1	0.95	1.00	0.89	0.98	1	0.97	1.00	0.78
DTK	MSC	0.81	0.95	1	0.94	0.95	0.92	0.97	1	0.97	0.70
	MSC 1st deriv	0.89	1.00	0.94	1	0.88	0.97	1.00	0.97	1	0.76
	raw	0.89	0.89	0.95	0.88	1	0.89	0.78	0.70	0.76	1
	1st deriv	1	0.97	0.88	0.95	0.88	1	0.95	0.91	0.94	0.90
	1st deriv MSC	0.97	1	0.90	1.00	0.75	0.95	1	0.93	1.00	0.72
FRV	MSC	0.88	0.90	1	0.92	0.77	0.91	0.93	1	0.94	0.74
	MSC 1st deriv	0.95	1.00	0.92	1	0.73	0.94	1.00	0.94	1	0.71
	raw	0.88	0.75	0.77	0.73	1	0.90	0.72	0.74	0.71	1
	1st deriv	1	0.44	0.37	0.46	0.98	1	0.97	0.91	0.97	0.88
	1st deriv MSC	0.44	1	0.99	0.99	0.38	0.97	1	0.93	1.00	0.76
KZC	MSC	0.37	0.99	1	0.99	0.31	0.91	0.93	1	0.94	0.73
	MSC 1st deriv	0.46	0.99	0.99	1	0.41	0.97	1.00	0.94	1	0.76
	raw	0.98	0.38	0.31	0.41	1	0.88	0.76	0.73	0.76	1
	1st deriv	1	0.98	0.82	0.96	0.92	1	0.96	0.91	0.95	0.85
	1st deriv MSC	0.98	1	0.90	0.99	0.89	0.96	1	0.94	1.00	0.74
PDB	MSC	0.82	0.90	1	0.94	0.86	0.91	0.94	1	0.93	0.76
	MSC 1st deriv	0.96	0.99	0.94	1	0.90	0.95	1.00	0.93	1	0.73
	raw	0.92	0.89	0.86	0.90	1	0.85	0.74	0.76	0.73	1
average	overall	0.85	0.88	0.84	0.89	0.79	0.93	0.91	0.88	0.91	0.78
	without KZC	0.91	0.92	0.88	0.92	0.84	---	---	---	---	---
Stdev	Overall	0.17	0.16	0.16	0.15	0.17	0.04	0.10	0.08	0.10	0.06
	without KZC	0.06	0.08	0.06	0.07	0.06	---	---	---	---	---
Overall	min	0.37	0.38	0.31	0.41	0.31	0.84	0.72	0.70	0.71	0.70
	max	0.99	1.00	0.99	1.00	0.98	0.98	1.00	0.97	1.00	0.90
Without KZC	min	0.81	0.75	0.75	0.73	0.73	---	---	---	---	---
	max	0.99	1.00	0.95	1.00	0.95	---	---	---	---	---

Table 4.2: Cumulative percentage explained variance (%EV) of the first two principal components of the PCA (VCC, ARP, UV-Vis, and IR), first two dimensions of the CA (sensory).

Cultivar	Winery	VCC	ARP	UV-Vis	Infra-red	Sensory
Chenin Blanc	AVN	82	84	93	79	65
	CDB	83	80	91	74	78
	DTK	72	85	97	96	66
	FRV	74	83	78	76	71
	KZC	76	94	93	98	76
	PDB	74	92	95	83	72
	average	77	86	91	84	71
	stdev	4	5	6	9	5
Sauvignon Blanc	AVN	79	82	92	68	64
	CDB	73	75	97	71	68
	DTK	76	81	80	64	71
	FRV	76	86	99	80	66
	KZC	80	79	97	65	55
	PDB	74	88	82	77	62
	average	76	82	91	71	64
	stdev	2	4	8	6	5
Overall	low	72	75	78	64	55
	high	83	94	99	98	78
	average	77	84	91	78	68
	stdev	3	5	7	10	6

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, PCA – principal component analysis, CA – correspondence analysis, stdev – standard deviation. *Cumulative percentage explained variance of the first two dimensions of the correspondence analysis

Table 4.3: Pairwise regression vector coefficients (p ≤ 0.01) for the scores of the individual data blocks.

		Chenin Blanc					Sauvignon Blanc				
		ARP	UV-Vis	IR	Sensory	VCC	ARP	UV-Vis	IR	Sensory	VCC
AVN	ARP	1	0.80	0.38	0.85	0.83	1	0.73	0.31	0.80	0.82
	UV-Vis	0.80	1	0.39	0.73	0.81	0.73	1	0.46	0.63	0.50
	IR	0.38	0.39	1	0.51	0.56	0.31	0.46	1	0.51	0.27
	Sensory	0.85	0.73	0.51	1	0.81	0.80	0.63	0.51	1	0.66
	VCC	0.83	0.81	0.56	0.81	1	0.82	0.50	0.27	0.66	1
CDB	ARP	1	0.79	0.47	0.87	0.63	1	0.93	0.47	0.60	0.50
	UV-Vis	0.79	1	0.39	0.67	0.39	0.93	1	0.31	0.43	0.33
	IR	0.47	0.39	1	0.71	0.54	0.47	0.31	1	0.72	0.75
	Sensory	0.87	0.67	0.71	1	0.72	0.60	0.43	0.72	1	0.85
	VCC	0.63	0.39	0.54	0.72	1	0.50	0.33	0.75	0.85	1
DTK	ARP	1	0.82	0.20	0.79	0.66	1	0.90	0.40	0.82	0.62
	UV-Vis	0.82	1	0.10	0.70	0.48	0.90	1	0.62	0.79	0.62
	IR	0.20	0.10	1	0.38	0.45	0.40	0.62	1	0.55	0.51
	Sensory	0.79	0.70	0.38	1	0.79	0.82	0.79	0.55	1	0.73
	VCC	0.66	0.48	0.45	0.79	1	0.62	0.62	0.51	0.73	1
FRV	ARP	1	0.90	0.71	0.78	0.55	1	0.83	0.38	0.73	0.60
	UV-Vis	0.90	1	0.67	0.73	0.58	0.83	1	0.21	0.46	0.45
	IR	0.71	0.67	1	0.78	0.69	0.38	0.21	1	0.61	0.62
	Sensory	0.78	0.73	0.78	1	0.73	0.73	0.46	0.61	1	0.73
	VCC	0.55	0.58	0.69	0.73	1	0.60	0.45	0.62	0.73	1
KZC	ARP	1	0.78	0.33	0.62	0.57	1	0.72	0.74	0.69	0.58
	UV-Vis	0.78	1	0.56	0.74	0.52	0.72	1	0.79	0.43	0.34
	IR	0.33	0.56	1	0.47	0.41	0.74	0.79	1	0.66	0.46
	Sensory	0.62	0.74	0.47	1	0.60	0.69	0.43	0.66	1	0.66
	VCC	0.57	0.52	0.41	0.60	1	0.58	0.34	0.46	0.66	1
PDB	ARP	1	0.93	0.67	0.63	0.76	1	0.92	0.68	0.71	0.45
	UV-Vis	0.93	1	0.62	0.59	0.66	0.92	1	0.62	0.67	0.51
	IR	0.67	0.62	1	0.86	0.58	0.68	0.62	1	0.66	0.42
	Sensory	0.63	0.59	0.86	1	0.67	0.71	0.67	0.66	1	0.63
	VCC	0.76	0.66	0.58	0.67	1	0.45	0.51	0.42	0.63	1

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light.

Table 4.4: Pairwise regression vector coefficients (p ≤ 0.01) for the PCA scores of the chemistry data blocks and the low-level PCA fused model.

Winery	Data blocks	Chenin Blanc	Sauvignon Blanc
AVN	IR	0.90	0.88
	ARP	0.66	0.61
	VCC	0.80	0.47
	UV-Vis	0.75	0.82
CDB	IR	0.88	0.85
	ARP	0.74	0.83
	VCC	0.60	0.72
	UV-Vis	0.78	0.76
DTK	IR	0.88	0.94
	ARP	0.56	0.67
	VCC	0.64	0.63
	UV-Vis	0.56	0.85
FRV	IR	0.95	0.83
	ARP	0.85	0.75
	VCC	0.73	0.72
	UV-Vis	0.86	0.72
KZC	IR	0.96	0.95
	ARP	0.53	0.78
	VCC	0.51	0.46
	UV-Vis	0.77	0.94
PDB	IR	0.92	0.93
	ARP	0.88	0.86
	VCC	0.69	0.54
	UV-Vis	0.88	0.86

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, IR – infrared, PCA – principal component analysis, %EV – percentage explained variation.

Table 4.5: Pairwise regression vector coefficients ($p \leq 0.01$) between the multifactorial analysis (MFA) and individual principal component analysis (PCA) data blocks.

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

VCC – volatile compounds composition, ARP – antioxidant-related parameters, UV-Vis – ultraviolet visible light, IR – Infra-red.

Chapter 5

Research results

Investigating the Concept of South African Old Vine Chenin Blanc

This manuscript was published in the peer-reviewed journal of **South African Journal of Enology and Viticulture**[3]

[3] Mafata, M., Brand, J., Panzeri, V., and Buica, A., 2020. Investigating the conceptual and perceptual sensory space of South African old vine Chenin Blanc using typicality rating, sorting, and free word association. South African Journal of Enology and Viticulture, 41 (2), 168-183. DOI:https://doi.org/10.21548/41-2-4018

Chapter 5: Investigating the Concept of South African Old Vine Chenin Blanc

Abstract

Although South African vineyards are still young by European standards, there is a belief in the industry that vines aged 35 or more years produce grapes and wines with specific characteristics ("old vine wines"). The aim of this study was to investigate the existence of the concept of old vine Chenin Blanc wines using a typicality rating and sorting tasks. Chenin Blanc wines were made from grapes harvested from vines aged five to 45 years old. Winemaking was standardised, with no wood contact. Typicality rating and sorting tasks were performed on young (first-stage) and two-year bottle-aged (second-stage) wines. Principal component analysis (PCA) on rating data demonstrated judge consensus, but no correlation was found between vine age and typicality rating. Sorting results were submitted to agglomerative hierarchical clustering (AHC) performed on the correspondence analysis (CA) and multidimensional scaling (MDS) results for grouping and attributes resulting from the sorting task. The clusters were different for the young wines and two-year bottle-aged wines. The verbal aspect of the sorting demonstrated the judges' agreement on the concept of old vine Chenin Blanc, shown by the annotation of the old vine group as 'complex', 'balance', 'rich' and 'good mouthfeel'. However, because the judges did not sort the wines according to vine age, the perceptual aspect of the concept could not be confirmed, its features could not be tested further, and the sensory space could not be built.

Abbreviations: RV (regression vector); PCA (principal component analysis); MDS (multidimensional scaling); CA (correspondence analysis); AHC (agglomerative hierarchical clustering); DA (descriptive analysis); CATA (check all that apply)

5.1 Introduction

In comparison to the long history of European and Middle Eastern vines (Stevenson, 2005), South African vineyards are young, with the first vines planted in the 17th century. According to recent statistics, 64% of the Chenin Blanc planted (by area under vine) is less than 20 years old and 36% is older than 20 years (SAWIS, 2018). The “old vine” designation has been used as a heritage mark to support the conservation of these vines and was established by the South African Old Vine Project (OVP) in 2017. The OVP demarked South African “old vines” as being 35 years or older, based on information gathered from years of collaborative input from industry experts, including viticulturists and winemakers (Crous, 2016).

Old vines (vineyards, grapes and wines) tend to receive special treatment with regard to viticultural and winemaking practices, documented by several surveys and interviews with industry experts. This special treatment is actively encouraged by the OVP, as it is believed that it will harness the full potential of the old vine and impart the character to the resulting wine. Some of the guidelines include a “holistic approach to weed control”, “movement from inorganic fertilisers to organic fertilisers”, “a minimalistic approach towards winemaking” for the wines to “be given the chance to reflect their specific terroir”, etc. (Old Vine Project [OVP], n.d.). Worldwide, it has been shown that any special treatment of a product (wine or other foodstuff) creates an emotional attachment to the product, along with expectations (Schouteten *et al*., 2015; Niimi *et al*., 2019).

The agreement among experts, which is reinforced through the OVP and its experience, is that old vine wines are less intense in fruity attributes but have more complex sensory attributes focused on mouthfeel; additionally, the full potential of the wine is reached after some years in the bottle, with the wines not being released in the harvest year (SASEV, 2018). Anecdotal evidence collected by the authors concerning old vine character (SASEV 2018) has created an interest in substantiating these ideas. In defining and testing the concept of “old vine character”, evidence needs to be collected and hypotheses have to be formulated and tested.

Currently, there is little scientific support for the anecdotal evidence, as only one study profiled 16 Chenin Blanc wines from vines older than 40 years using descriptive analysis (Crous, 2016). The study evaluated multiple sensory modalities, namely odour and in-mouth sensations, with a focus on mouthfeel. It also used calibrated standards and, where standards were not available, conceptual consensus was established based on discussions among judges. The reasoning for the mouthfeel approach was based on the anecdotal evidence mentioned above (OVP, n.d.; SASEV, 2018). In the work by Crous (2016), when panellists described old vine Chenin Blanc wines, the terms *body*, *concentration*, *complexity*, *length*, *acidity*, *heat*, *balance* and *integration* featured prominently. Since the samples were commercial wines made using different protocols, Crous (2016) noted that the effects of winemaking outweighed any possible correlations with the vine age.

One approach to studying old vine character is through establishing its associated typicality features. Wine typicality refers to a group of sensory attributes that, together, become the defining features describing a concept; typicality may be categorised under cultivar, winemaking style, regionality (appellation) or, in this case, old vine character. In this context, typicality is defined as the level (or "degree of representativeness") of a sample to a category, measured against a prototype (Chrea *et al*., 2005). In the case of a sensory concept, the prototypes or "established references" (Perrin & Pagès, 2009) can be different for each assessor due to differences in experience and exposure; hence, typicality judgments may differ among experts. Consistency among assessors suggests the homogeneity of the prototypes, or even the existence of a common prototype and possible conditions for demonstrating a typicality concept (Casabianca *et al*., 2006). In practice, it was demonstrated that wines that are less representative of the prototype belong to neighbouring categories (Perrin & Pagès, 2009) and it is possible for instances of borders between categories to arise (Ballester *et al*., 2005).

There are four stages to testing concepts of typicality and, according to the methodology proposed by Perrin and Pagès (2009), these have to be followed in sequence. Firstly, *panel agreement* has to be established, followed by *conceptual agreement*, *perceptual agreement* and, finally, *measuring the feature/drivers* can be considered. Each step is dependent on the previous one. If at any point agreement is not achieved, the investigation cannot be continued and the methods or panels have to be revisited.

Typicality can be investigated sensorially in different ways using verbal and/or non-verbal methods (Perrin & Pagès, 2009). The reasoning behind this is that the differences between wines considered to be most and least representative of the concept under investigation should manifest both intuitively (as seen in non-verbal methods) and through verbal cues. It is important to understand when to use which type of method (verbal, non-verbal or a combination), how to choose the mode of assessment (gustatory, olfactory or global) and which type of panel to use (experts or trained). Elements to consider when making these decisions are whether or not the concept has been well established previously, whether there are known features that contribute to the concept, and whether these features have standards that can be used for calibration (Perrin & Pagès, 2009).

Verbal methods used for typicality studies include descriptive analysis (DA) for the colour of Provence Rosé wines (Coulon-Leroy *et al*., 2018) and check all that apply (CATA) for the minerality of Burgundy Chardonnay (Ballester *et al*., 2013). Non-verbal methods include sorting for demonstrating the existence of a Chardonnay wine concept (Ballester *et al*., 2005), typicality and hedonic rating for minerality in French vs New Zealand Sauvignon Blanc (Parr *et al*., 2015), and other various combinations.

As mentioned previously, the evaluation can be used to investigate the contributions of the features to the concept through gustatory, olfactory or global assessment. Studies have found the differences in the success of the mode of assessment to be based on the dominant features related to the concept. If, for example, the prominent features are known to manifest in the aroma, then the assessment will be on the olfactory stimuli. If, however, a concept has not previously been annotated with features, then a global assessment is used. This type of systematic investigation is illustrated by Ballester *et al*. (2008) in testing the concept of Chardonnay by both expert and consumer panels. The study found a clear distinction between Chardonnay wines and Melon de Bourgogne (used as a non-Chardonnay example to establish the borders of the concept) by an expert panel. The borders of representativeness were then tested in two ways using rating (to look at the degree of representativeness) and sorting (to look at the membership in the designated groups).

The use of trained and expert panels has also been investigated in the literature. If a concept has features that can be calibrated for using standards and/or definitions, a trained panel may be used (Ballester *et al*., 2008). Concepts that include features that could not be calibrated, and thus rely on experience, favour expert panels. In this case, it is possible that the conceptual agreement when defining terms and the perceptual agreement when consistently assessing the features in wine are not unified, as was the case with the minerality of Burgundy Chardonnay (Ballester *et al*., 2013); although the investigation achieved both panel consensus and conceptual agreement on minerality, perceptual agreement could not be reached and hence the features could not be verified.

In this context, the aim of the current study was to investigate the concept of old vine Chenin Blanc using typicality rating, sorting, and free word association. Compared to the previous study by Crous (2016), in which the intrinsic features of each wine were measured by DA using a bottom-up approach that is experimentally directed (Lindsay & Norman, 1977), the current work proposes a top-down approach in which the understanding of the concept is first developed before trying to measure its features (Lindsay & Norman, 1977; Brochet & Dubourdieu, 2001). A combination of non-verbal (rating and sorting) and verbal (the added annotation of sensory attributes in the sorting exercise) methods was used. The sensory panel was constituted of industry professionals. Since the previous study noted the potential influence of winemaking (Crous, 2016), the same winemaking protocol was used in this study for all the grapes sourced from vineyards aged five to 44 years. In addition, the wines were evaluated young (first evaluation stage approximately three months after bottling) and after two years of ageing in the bottle (second evaluation stage).

5.2 Materials and Methods

5.2.1 Grape sources and winemaking

Chenin Blanc grapes were sourced from 23 vineyards across the Western Cape province of South Africa. Grapes were harvested in 2017 at commercial maturity according to the growers, ranging from 23°Brix to 25°Brix, with two exceptions at 17.3°Brix (sample 765) and 19.2°Brix (sample 769). Twelve young vines (< 35 years old) and 11 old vines (≥ 35 years old) were included in the project; vine ages ranged from five to 45 years. Grapes were treated with 30 mg/L sulphur dioxide (SO_2) at crushing. The juice was settled overnight at 4°C, racked and allowed to come to room temperature. Juice was inoculated with Vin7 yeast (Zymasil, AEB Group SpA, Bologna, Italy) according to the manufacturer's instructions. The fermentation was allowed to proceed in a temperature-controlled room at 15°C to 18°C. The SO2 levels were adjusted to 50 mg/L post-alcoholic fermentation, and 50 mg/L bentonite was added before cold stabilisation, which took place over two weeks at -4°C. The wine was then racked and bottled without filtration in 750 mL screw cap green bottles (Consol, South Africa). The wines were stored in the vinoteque under controlled temperature and humidity conditions until their evaluation: first as young wines (three months after bottling), then as bottle-aged wines (two years after bottling). Grape juice and wine oenological parameters (Table 5.1) were measured on a Metrohm 862 compact titrosampler (Herisau, Switzerland) using chemicals (sodium hydroxide, potassium iodide/ potassium iodate and sodium thiosulfate) purchased from Cameron Chemical Consultants (Cape Town, South Africa).

5.2.2 Sensory evaluation

The approach used in this study is based on the methodology published by Ballester et al. (2008). The analysis was performed in a quiet, well-ventilated and odour-free room with the temperature set at 20 ± 2°C. Samples were presented in black ISO glasses, covered with a Petri dish and labelled with a three-digit code. Samples were randomised across judges prior to analysis according to a William's Latin square design. An expert panel of 32 judges in 2018 and 14 in 2019 assessed the 23 wines; the judges were industry professionals with more than five years' experience in the production and evaluation of old vine Chenin Blanc. The experimental design was done using Compusense cloud (Compusense, Guelph, Canada). Two sensory tasks, namely rating and sorting (Valentin et al., 2012), were performed in one session with a 15-minute break and a free word association exercise between them. The first task was a typicality rating on a 100 mm unstructured line scale, ranging from “very bad example” anchored at 0 to “very good example” anchored at 100 (Garrido-Bañuelos et al., 2020) and samples were presented monadically. The experts were instructed to rate each sample on the scale according to their judgement for an old vine Chenin Blanc wine. Before beginning the second task, judges were asked to list three to five words that came to mind when “typical old vine Chenin Blanc wine” was mentioned. The second task was a flexible sorting exercise with all 23 wines presented at once. This was considered a flexible sorting since the judges were instructed to sort the wines into two groups, namely “young vine CB” or “old vine CB”, but they were allowed to create a third group if the samples did not fit either of the two groups. Judges were also asked to give three to five attributes associated with each group. The terms generated during the sorting task were consolidated based on their semantic and synonymous relationship by agreement among the researchers.

Table 5.3: Oenological parameters for Chenin Blanc grapes (mass, NOPA, ammonium, YAN, and ˚Brix) harvested from old and young vines in 2017 and resulting wines (pH and TA).

Sample code	Vine age (years)	Class designation	Mass (kg)	NOPA (mg N/L)	NH_4 (mg N/L)	YAN (mg N/L)	˚Brix	pH	TA (mg/L)
YV751	29	Young	17	180	50	230	21.7	3.25	6.57
OV752	**n/s	Old	18	160	30	190	22.2	3.30	5.51
YV753	5	Young	20	200	60	260	21.8	3.35	5.66
OV754	**n/s	Old	20	170	30	200	23.6	3.34	6.56
OV755	**n/s	Old	22	180	60	240	22.4	3.42	4.59
OV756	39	Old	20	170	50	220	24.1	3.46	5.53
YV757	34	Young	19	130	30	160	24.6	3.34	6.24
YV758	34	Young	19	150	40	190	21.8	3.36	6.99
YV759	28	Young	34	-	-	-	20.0	3.41	6.06
OV760	39	Old	24	-	-	-	24.6	3.53	5.22
YV761	34	Young	18	120	30	150	24.2	3.50	6.53
YV762	**n/s	Young	18	150	50	200	23.8	3.64	4.71
YV763	6	Young	19	140	30	170	23.9	3.60	5.59
YV764	24	Young	20	130	50	180	23.0	3.43	5.70
OV765	39	Old	21	210	80	290	17.3	3.17	10.35
YV766	33	Young	17	160	50	210	21.6	3.57	4.38
OV767	37	Old	23	210	50	260	22.1	3.68	5.88
OV768	41	Old	20	190	150	340	21.5	3.75	4.34
YV769	31	Young	38	160	50	210	19.2	3.35	8.16
OV770	37	Old	18	150	40	190	22.7	3.46	5.03
OV771	35	Old	19	220	50	270	22.5	3.63	5.58
YV772	27	Young	21	140	40	180	24.5	3.55	5.03
OV773	44	Old	17	-	-	-	23.0	3.54	6.03

Young – vines 34 years and younger, Old – vines 35 years and older. Mass means the mass of grapes as measured before crushing. NOPA - nitrogen by o-phthaldialdehyde assay; NH_4 – chemical formula for ammonium, YAN - yeast assimilable nitrogen; TA – total acidity; n/s – not specified.

5.2.3 Statistical analysis

Rating data was captured as a judge vs wines correlation matrix. Principal component analysis (PCA) was performed on the correlation matrix to evaluate judge consensus (Perrin & Pagès, 2009). The data was averaged over the judges and PCA was performed on the resulting correlation matrix to investigate correlations between the different wines (Perrin & Pagès, 2009). Data groupings on the basis of the sorting were captured as a co-occurrence matrix and the attributes used to describe the groups were captured as a correlation matrix of wines and attributes. Multidimensional scaling (MDS) was performed on the co-occurrence matrix and correspondence analysis (CA) on the correlation matrix (Salkind, 2012). Regression vector (RV) coefficients were calculated among the CA and MDS biplot co-ordinates for each year, and between the young and the two-year bottle-aged wines (Abdi, 2007). Unweighted pair-average agglomerative hierarchical clustering (AHC), using a similarity-based, Pearson correlation coefficient, was performed on the MDS and on the CA for both the wines' and the attributes' correlation matrices. Statistical analyses were performed in XLSTAT2018 (Addinsoft, Paris, France).

5.3 Results

5.3.1 Judge consensus

In order to evaluate panel consensus, PCAs were conducted on the rating scores for both the young and bottle-aged wines (Figure 5.1). The results for the young wines show a cumulative explained variance of 16% for the first three dimensions. Full explained variance (100%) was achieved over 22 dimensions, with all dimensions contributing almost equally (from PC1 with 5.8% to PC22 with 3.7%). Results from the bottle-aged wines showed a cumulative explained variance of 17% for the first three dimensions of the PCA, with the full explained variance being achieved over 21 dimensions (from PC1 with 6.1% to PC21 with 3.7%).

Although the cumulative explained variance for both years of the evaluation was less than 20% for the first three dimensions (Figure 5.1), the linear correlation across the first dimension was an indicator of good consensus between the judges. The correlation between judges varied linearly along the first dimension, with judges 12 and 24 being the exception for the first evaluation stage (young wines) and judge 10 for the second (bottle aged). The judges who were not in consensus with the rest of the panel were not excluded from further analyses, because they were within the 95% confidence interval and thus not statistical outliers.

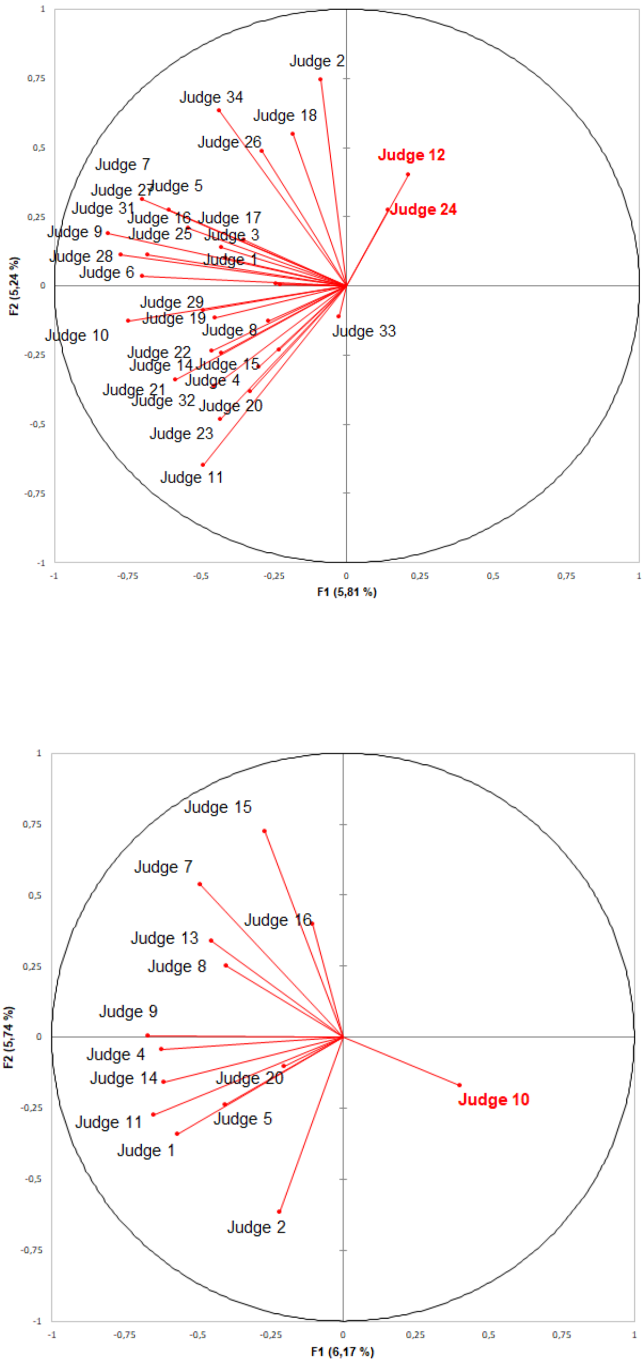


Figure 5.1: Principal component analysis (PCA) of rating data collected from young wines (top) and wines aged for two years in the bottle (bottom) based on experiments on old and young vine Chenin Blanc wines.

5.3.2 Non-verbal typicality assessments

5.3.2.1 Typicality rating

In order to see if there was a correlation between vine age and the typicality rating, the average scores per sample were plotted against the vine age. If the old vine concept was to be observed, the old vine wines should have been rated higher on the typicality scale than the young vine wines, according to their degree of representativeness of the concept. This was not the case, as linear regression analysis showed no correlation between the average rating score and the vine age for either young wines or bottle-aged wines.

The results for both evaluation stages show a wide distribution of the average typicality scores. Judges used the entire scale (from 0 to 100), with the average scores ranging from 20 to 66 for young wines and 29 to 67 for bottle-aged wines. This result indicates that the judges did not have a unified perception of the wine typicality with regard to the old vine status. Statistically, the score distribution of each sample was not always normal, as some samples had a bimodal distribution whereas others had a random distribution (Figure 5.2). For young wines, the wine rated the lowest was OV765, which was made from a 39-year-old vine. Surprisingly, the wine made from the oldest vines (OV773, 44 years old) and youngest vines (YV753, five years old) were rated similarly (56 and 49 for OV773 and YV753, respectively). For bottle-aged wines (second stage), the sample with the lowest rating was the wine from the oldest vines in the set, OV773, which was rated even lower than the wine made from the youngest vines in the sample group (YV753, five-year-old vines).

In order to investigate any relationship between the two years' results, the average scores for each year were plotted against each other. The regression coefficient (R2 = 0.5852) indicated only a trend between the young and bottle-aged wines. This means that any changes that occurred during ageing could neither be correlated with vine ageing nor typicality rating. Given the random distribution of samples, no borders could be imposed based on vine age, and thus no classifications could be made according to age. This means that there was no perceptual agreement between judges when it came to old vine South African Chenin Blanc typicality as measured by the rating task.

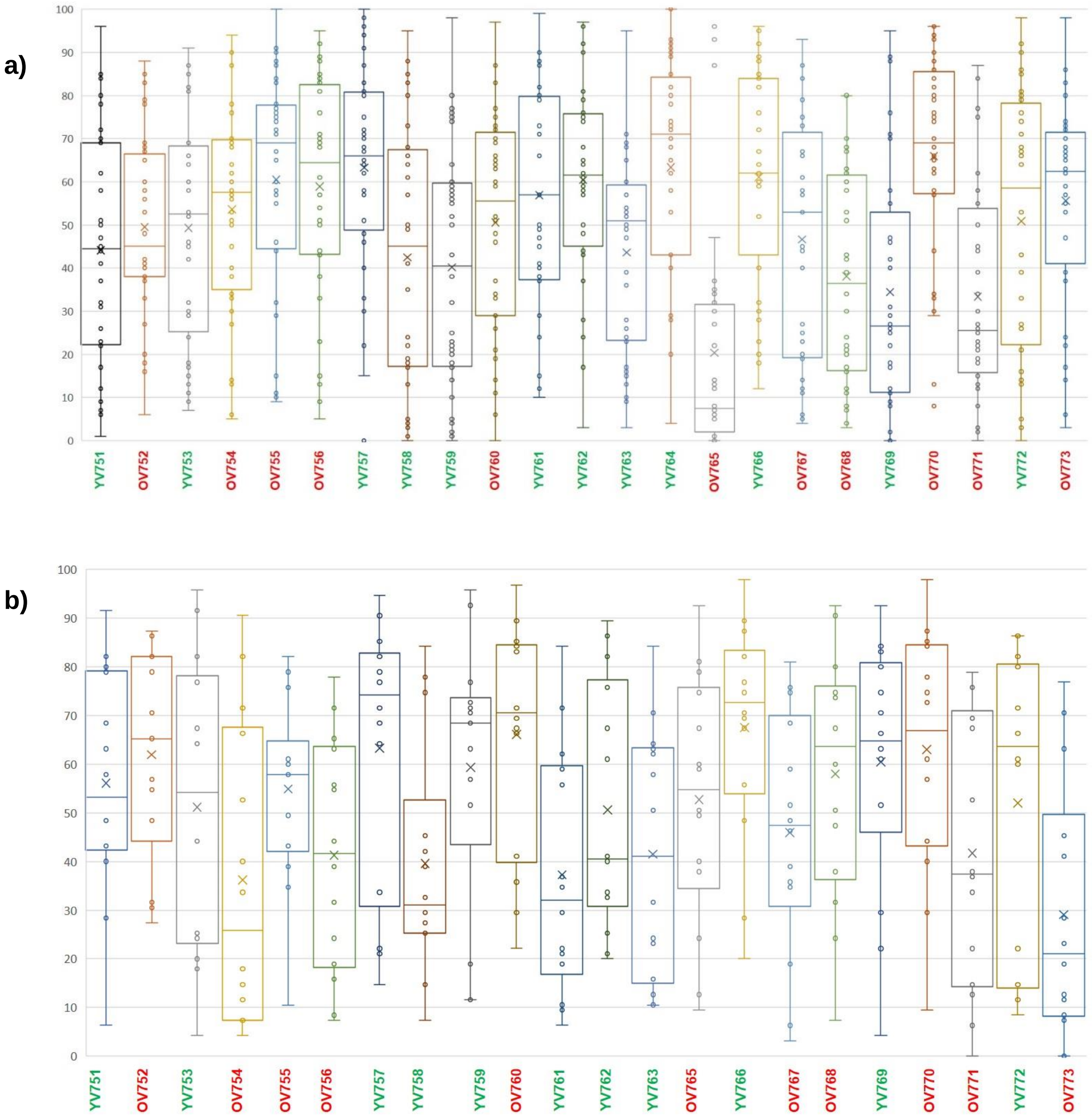


Figure 5.2: Box-and-whisker distribution plot of typicality rating scores for young wines (**a**) and two-year bottle-aged wines (**b**) of old vine Chenin Blanc grapes vines of different ages. Young vines coded with YV (green) and old vines with OV (red) before the unique three-digit code.

5.3.2.2 Multidimensional scaling (MDS) on typicality sorting data

The second non-verbal assessment of the typicality of old vine Chenin Blanc wine was the sorting task. Unlike the rating task, in which the presentation of the samples is monadic, in this second task wines were judged together and grouped according to their similarity under the groups *old vine* and *young vine*. The first three dimensions of the MDS were considered enough for assessing significant relationships between samples based on Kruskral's stress indices (results not shown) for both evaluation stages (young wines and two-year bottle-aged wines). MDS and agglomerative hierarchical clustering (AHC) were then performed on the first three dimensions, and the results are shown in Figs 3 and 4 for the two evaluation stages.

Cluster analysis of the MDS gave three main clusters and showed no grouping of samples according to vine age for either evaluation stage. The wine from the oldest vine (wine OV773, 44-year-old vine) and the youngest vine (YV753, five-year-old vine) were in two separate clusters. For both stages, the distribution within each cluster was random, the distances between the members of each cluster (i.e. samples or branches) was also random and not related to vine age. It can be concluded that clustering was related to neither the categories "old vine"/ "young vine" nor to any observable trends according to vine age.

5.3.2.3 Correspondence analysis (CA) on typicality sorting

Correspondence analysis of the sorting data provided a biplot that showed the correlation between samples (presented in this section) and between attributes (presented under Verbal assessments below). CA showed the distribution of the total inertia (0.327 and 0.494 for the first and second evaluation stage, respectively) over 22 and 21 dimensions, respectively. The first three dimensions had cumulative percentages of 61% and 64% of the inertia respectively for the two stages. AHC was done only on these first three dimensions (Figs 5 and 6). Three clusters were formed in each case; the clusters contained samples from different vine ages. The clustering of samples was related neither to the "old vine"/"young vine" categories, nor to vine age. Unlike in the MDS, the wine from the oldest vines (OV773, 44 years old) and the wine from the youngest vines (YV753, five years old) belonged to the same cluster for the first evaluation stage and to the same cluster for the second.

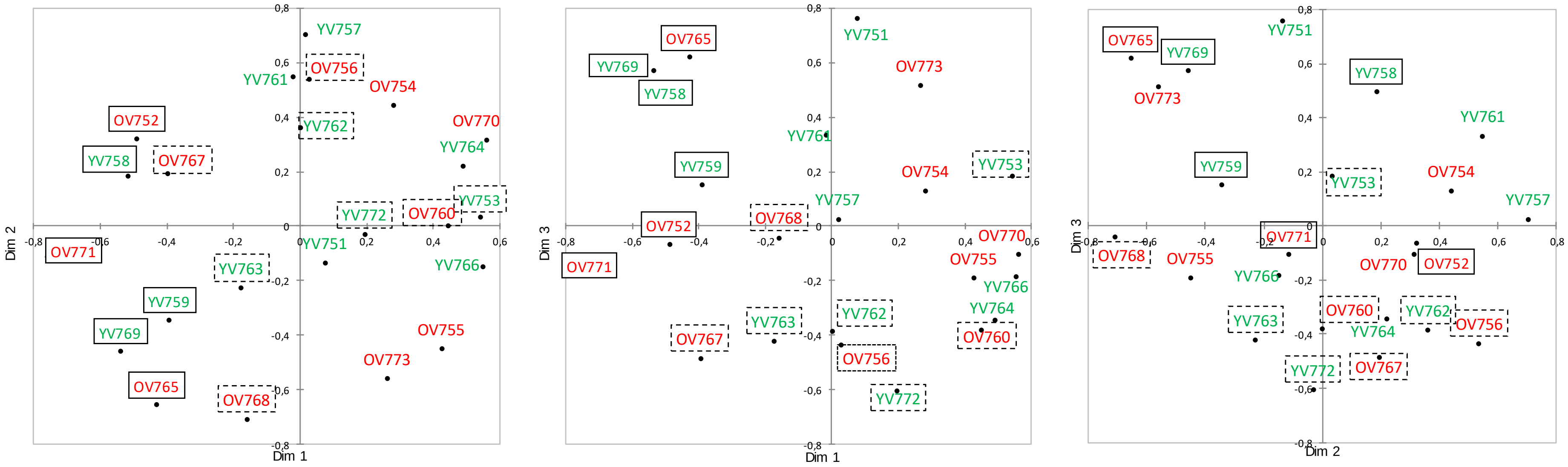


Figure 5.3: Multidimensional scaling (MDS) on sorting task of old (red) and young (green) vine Chenin Blanc wines analysed in the first year. Different shading indicates the groups according to agglomerative hierarchical clustering (AHC) performed on the first three dimensions of the MDS.

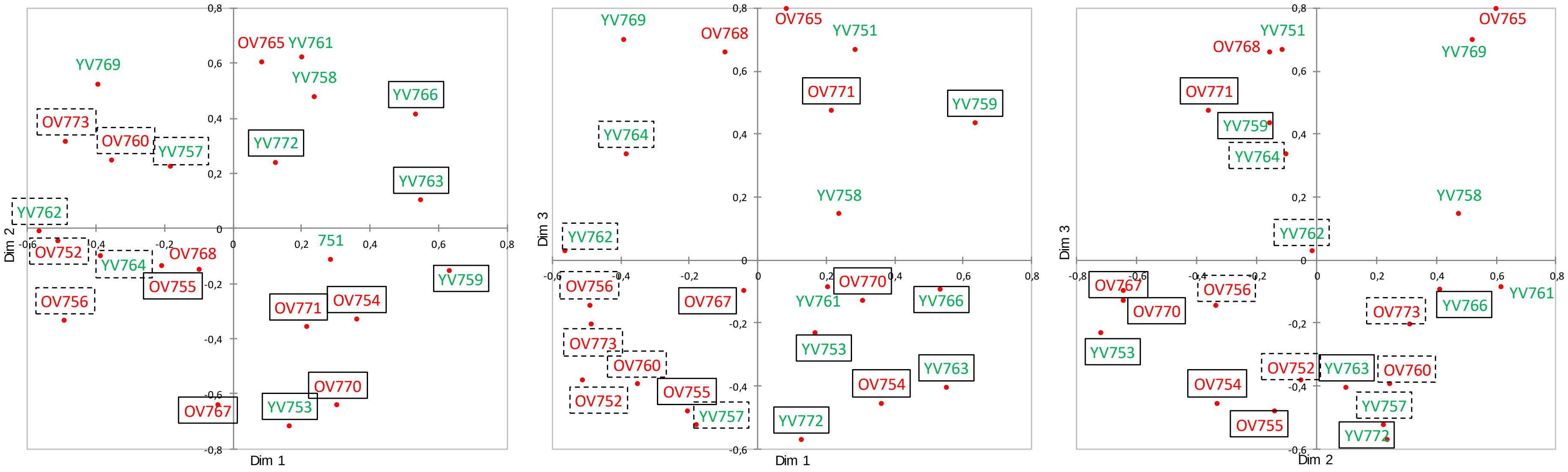


Figure 5.4: Multidimensional scaling (MDS) on sorting task of old (red) and young (green) vine Chenin Blanc wines analysed after two years of ageing in the bottle. Different shading indicates the groups according to agglomerative hierarchical clustering (AHC) performed on the first three dimensions of the MDS.

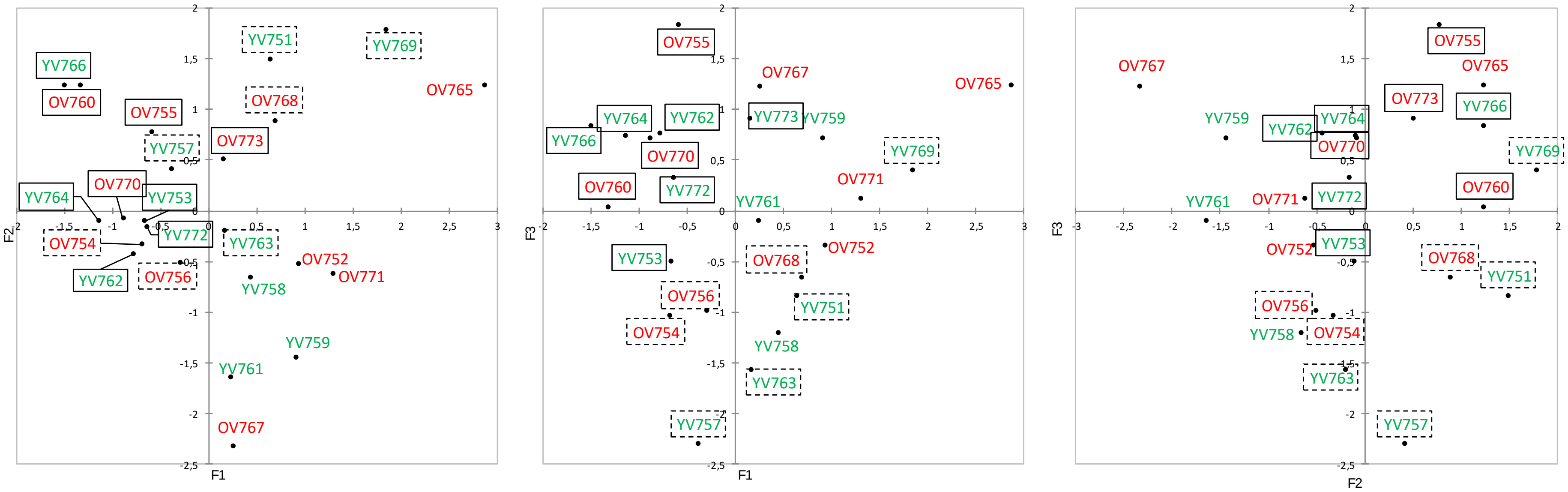


Figure 5.5: Correspondence analysis (CA) on sensory analysis from old (red) and young (green) vine Chenin Blanc wines analysed in the first year. Samples with the same box shading belong to the same cluster, analysed using agglomerative hierarchical clustering (AHC) on the first three dimensions.

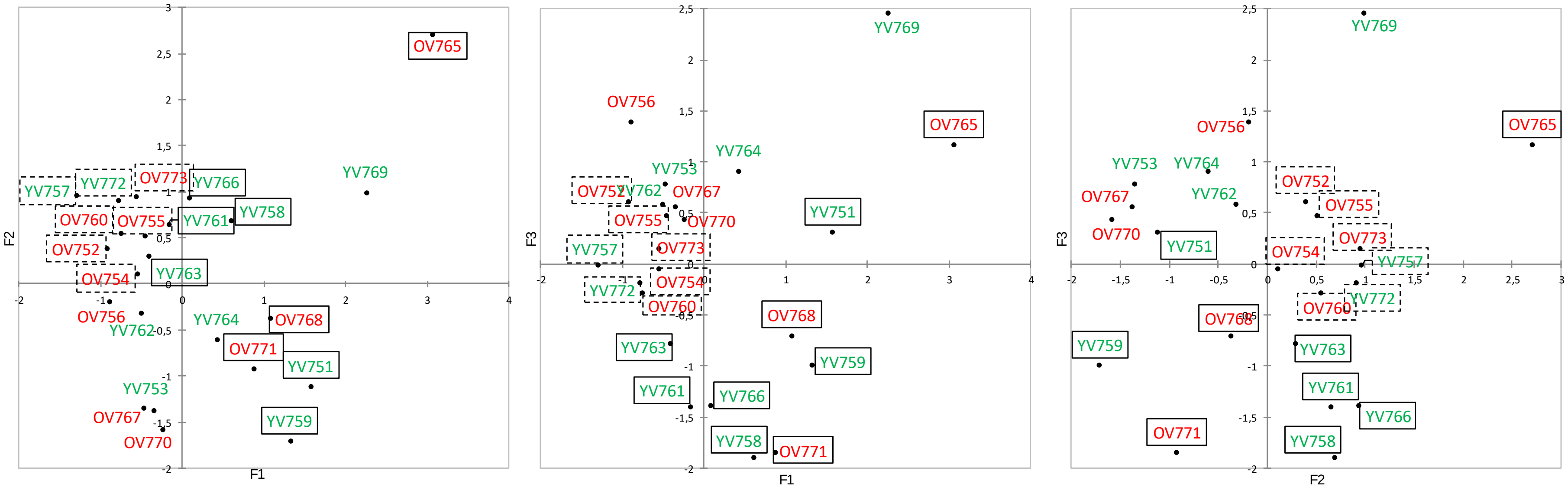


Figure 5.6: Correspondence analysis (CA) on sensory analysis from old (red) and young (green) vine Chenin Blanc wines analysed after two years of ageing in the bottle. Samples with the same box shading belong to the same cluster, analysed using agglomerative hierarchical clustering (AHC) on the first three dimensions.

5.3.2.4 Comparison of sample configurations

RV coefficients were calculated in order to assess any differences or similarities between sample configurations generated in the two stages through MDS and CA. The comparison was two-fold: within a stage, MDS to CA, and between the stages, CA to CA and MDS to MDS configurations. The data captured from the rating task also generated one PCA for each evaluation stage that contained sample configurations. However, as one of the samples was not included in the second-stage evaluation, RV coefficients could not be calculated for the rating results.

MDS and CA plots were generated for the verbal and non-verbal aspects of the sorting data. The main difference in these analyses is that the MDS relies only on the associations between samples, whereas the CA uses the attributes to generate the correlation between samples. Since these were done within one task, although looking at different aspects, they should result in a similar relationship between samples. As such, RV coefficients were used to measure the configurational similarity between the CA and MDS plots.

For sorting, the results for young wines showed CA vs MDS RV coefficients of 0.68 and 0.60 for the first two and three dimensions, respectively. The second stage (bottle-aged wines) results showed CA vs MDS RV coefficients of 0.68 and 0.71 for the first two and three dimensions, respectively. Looking at correlations between the two years of evaluation, RV coefficient were calculated for MDS vs MDS (0.37 and 0.34, first two and three dimensions, respectively) and CA vs CA (0.47 and 0.39, first two and three dimensions, respectively). These values were low, meaning that the samples were sorted differently for the different evaluation stages. Although three clusters were formed for both the evaluation stages, the members belonging to each of the clusters were different.

Looking for any similarity between the two datasets (rating and sorting), the configurational space was assessed using RV coefficients. The wine samples were considered observations in the rating data and modelled by PCA; the resulting configuration was used to generate the RV coefficients against the CA and MDS results.

In the case of the evaluation of the young wine, the results showed poor correlation between the configurations for rating by PCA and sorting by MDS (first two dimensions, RV = 0.44; first three dimensions, RV = 0.41) and between rating by PCA and sorting by CA (first two dimensions, RV = 0.52; first three dimensions, RV = 0.474). This could be because of the non-normal distribution of the rating scores for each sample, as discussed above. The membership of the same sample to different groups (*young vine* and *old vine*) in the sorting could also contribute to the differences in configurations (i.e. low RV coefficient values). Since sample 764 was excluded from the rating of the bottle-aged wines, the RV coefficients for the second evaluation stage could not be calculated.

5.3.3 Verbal assessment of typicality

5.3.3.1 Verbal aspects of the sorting task

The sorting resulted in three groups for both the young wines and the two-year bottle-aged wines. The groups *young vine* and *old vine* were allocated to them, but the judges collectively generated the *teenager* and *outlier* group identities for the first and second evaluation stages, respectively. The consolidation of attributes resulted in 46 terms for young wines and 68 for bottle-aged wines, which were used to generate the CA. The first three dimensions of the CA contained 61% and 64% of the explained variance for the two evaluation stages, respectively. AHC done on the three-dimensional space resulted in the formation of two main clusters (Figure 5.7). The members of each cluster, their weight and their contributions to the explained variance in the first three dimensions are listed in Tables 5.2 and 5.3. The *old vine* cluster had associated terms that are mouthfeel-related and support the findings of Crous (2016)*.* Some examples are ‘robust’, ‘texture, ‘good mouthfeel’ and ‘complex’ for the young wines, and ‘structure’, ‘dense palate’, ‘texture’ and ‘rich mouthfeel’ for the bottle-aged wines.

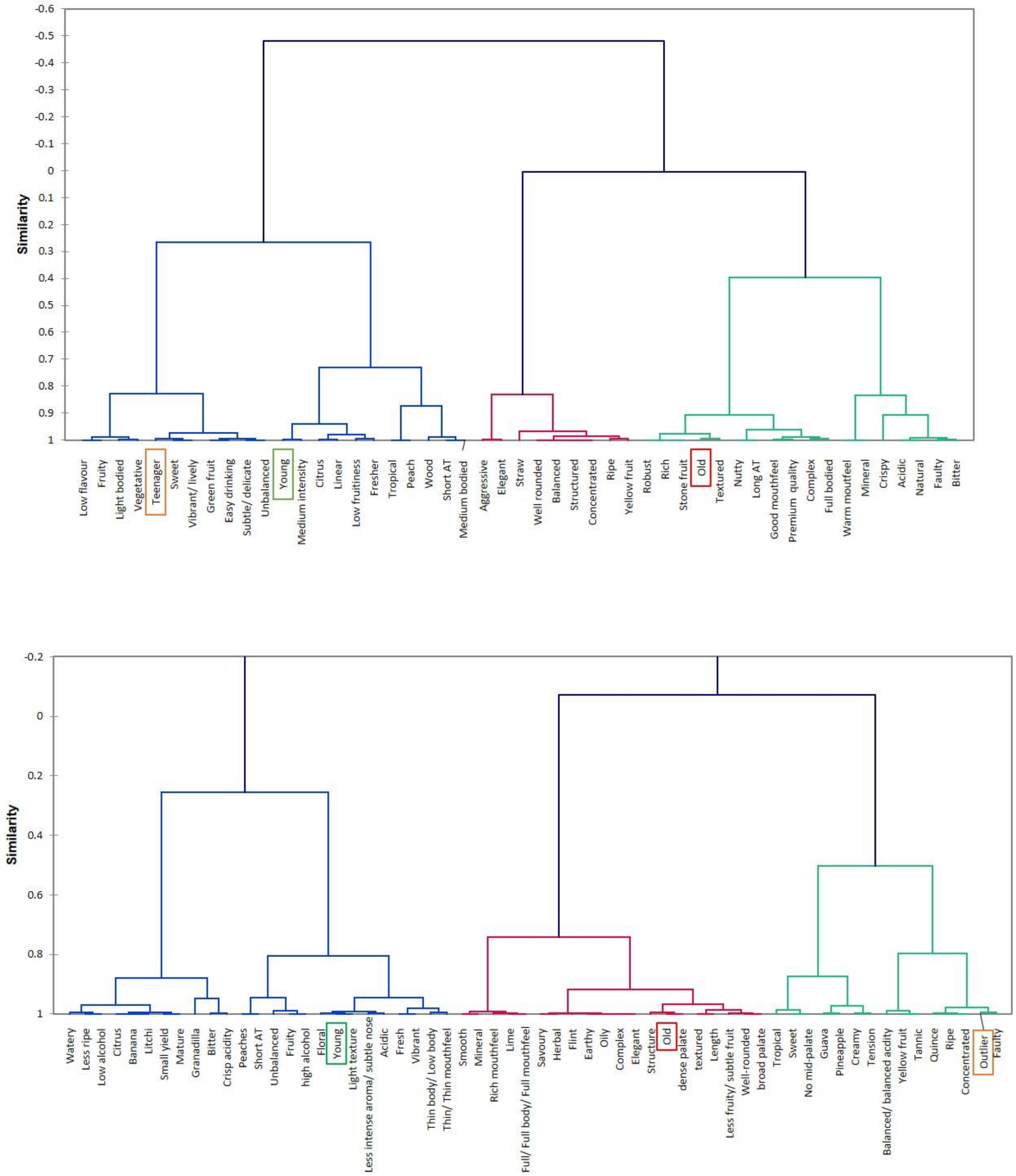


Figure 5.7: Agglomerative hierarchical clustering (AHC) on the first-year results on CA attributes for the first year (top), and two-year bottle aged (bottom) wines.

Table 5.2: AHC groups for the first three dimensions of the CA for the analysed during the first year.

CLUSTER 1

		41.88%	11.19%	8.46%	61.53%
Attribute	Weight (relative)	F1	F2	F3	Sum
Old	0.123	0.108	0.013	0.003	0.124
Textured	0.027	0.030	0.025	0.004	0.059
Robust	0.008	0.011	0.000	0.002	0.013
Rich	0.028	0.032	0.021	0.000	0.054
Nutty	0.016	0.020	0.031	0.020	0.070
Complex	0.027	0.010	0.000	0.005	0.014
Crispy	0.010	0.010	0.000	0.097	0.106
Stone fruit	0.010	0.012	0.016	0.000	0.027
Good mouthfeel	0.009	0.011	0.002	0.009	0.022
Warm mouthfeel	0.005	0.000	0.013	0.004	0.017
Long AT	0.040	0.036	0.002	0.036	0.075
Full bodied	0.019	0.012	0.000	0.005	0.017
Faulty	0.013	0.031	0.086	0.033	0.150
Mineral	0.029	0.000	0.059	0.024	0.084
Acidic	0.024	0.170	0.206	0.045	0.422
Bitter	0.005	0.030	0.053	0.001	0.083
Natural	0.004	0.003	0.012	0.073	0.089
Premium quality	0.004	0.005	0.001	0.004	0.011

CLUSTER 2

	2018	41.88%	11.19%	8.46%	61.53%
Attribute	Weight (relative)	F1	F2	F3	Sum
Young	0.105	0.092	0.020	0.014	0.126
Wood	0.002	0.001	0.039	0.006	0.046
Low fruitiness	0.009	0.005	0.019	0.001	0.026
Fresher	0.052	0.024	0.044	0.003	0.071
Medium intensity	0.003	0.003	0.076	0.010	0.089
Citrus	0.024	0.011	0.033	0.007	0.050
Tropical	0.034	0.002	0.006	0.001	0.009
Peach	0.013	0.000	0.024	0.121	0.145
Short AT	0.014	0.005	0.006	0.021	0.032
Linear	0.013	0.011	0.004	0.011	0.025
Medium bodied	0.003	0.000	0.021	0.005	0.026
Teenager	0.017	0.035	0.001	0.006	0.042
Low flavour	0.005	0.002	0.003	0.001	0.006
Fruity	0.057	0.027	0.030	0.013	0.070
Green fruit	0.007	0.017	0.005	0.057	0.078
Subtle/ delicate	0.019	0.000	0.007	0.016	0.023
Unbalanced	0.016	0.023	0.003	0.000	0.026
Sweet	0.017	0.021	0.038	0.045	0.104
Light bodied	0.042	0.044	0.001	0.006	0.051
Vegetative	0.004	0.002	0.041	0.013	0.056
Easy drinking	0.005	0.023	0.002	0.004	0.028
Vibrant/ lively	0.010	0.001	0.011	0.024	0.036

CLUSTER 3

		41.88%	11.19%	8.46%	61.53%
Attribute	Weight (relative)	F1	F2	F3	Sum
Structured	0.010	0.009	0.008	0.094	0.112
Ripe	0.034	0.014	0.002	0.024	0.039
Concentrated	0.014	0.009	0.002	0.001	0.011
Yellow fruit	0.008	0.023	0.000	0.041	0.065
Aggressive	0.004	0.001	0.000	0.000	0.001
Balanced	0.028	0.035	0.000	0.027	0.062
Well rounded	0.013	0.007	0.000	0.011	0.019
Straw	0.008	0.011	0.010	0.048	0.069
Elegant	0.013	0.013	0.005	0.005	0.023

Table 5.3: AHC groups for the first three dimensions of the CA wines aged for two years in the bottle.

CLUSTER 1						CLUSTER 2						CLUSTER 3					
		38.18%	15.37%	10.88%	64.43%			38.18%	15.37%	10.88%	64.43%			38.18%	15.37%	10.88%	64.43%
Attributes	Weight (relative)	F1	F2	F3	sum	Attributes	Weight (relative)	F1	F2	F3	sum	Attributes	Weight (relative)	F1	F2	F3	sum
Old	0.077	0.046	0.002	0.000	0.048	Young	0.074	0.073	0.020	0.010	0.104	Outlier	0.012	0.107	0.177	0.111	0.394
Less fruity/ subtle fruit	0.012	0.008	0.016	0.011	0.035	Less intense aroma/ subtle nose	0.020	0.020	0.001	0.001	0.022	Ripe	0.027	0.001	0.033	0.007	0.042
Lime	0.005	0.009	0.003	0.009	0.020	Fruity	0.056	0.002	0.042	0.010	0.053	Yellow fruit	0.008	0.008	0.004	0.000	0.012
textured	0.019	0.000	0.006	0.008	0.014	Fresh	0.011	0.004	0.008	0.014	0.026	Guava	0.021	0.014	0.042	0.019	0.075
Rich mouthfeel	0.014	0.017	0.009	0.013	0.039	Less ripe	0.002	0.044	0.015	0.050	0.110	Tropical	0.018	0.001	0.001	0.029	0.031
Full/ Full body/ Full mouthfeel	0.040	0.018	0.009	0.003	0.030	Banana	0.005	0.003	0.040	0.045	0.088	Quince	0.010	0.009	0.016	0.027	0.052
Well-rounded	0.014	0.013	0.009	0.007	0.028	Litchi	0.005	0.003	0.040	0.045	0.088	Pineapple	0.012	0.002	0.016	0.015	0.033
dense palate	0.012	0.007	0.011	0.004	0.022	Citrus	0.005	0.003	0.040	0.045	0.088	Sweet	0.030	0.001	0.003	0.052	0.056
broad palate	0.026	0.021	0.001	0.000	0.022	Peaches	0.032	0.003	0.011	0.000	0.014	Balanced/ balanced acidity	0.041	0.042	0.006	0.002	0.049
Smooth	0.011	0.013	0.002	0.016	0.031	Granadilla	0.007	0.003	0.005	0.005	0.012	Creamy	0.011	0.000	0.039	0.050	0.090
Length	0.051	0.036	0.003	0.007	0.047	Floral	0.023	0.000	0.001	0.001	0.002	Tannic	0.008	0.008	0.004	0.000	0.012
Structure	0.019	0.002	0.001	0.000	0.004	Bitter	0.007	0.002	0.010	0.002	0.014	No mid-palate	0.003	0.003	0.000	0.051	0.054
Complex	0.019	0.027	0.003	0.021	0.051	Crisp acidity	0.006	0.008	0.004	0.014	0.026	Concentrated	0.012	0.003	0.022	0.011	0.036
Savoury	0.006	0.008	0.001	0.011	0.020	Acidic	0.018	0.115	0.002	0.002	0.119	Tension	0.005	0.001	0.006	0.032	0.040
Herbal	0.006	0.008	0.001	0.011	0.020	Light texture	0.011	0.009	0.004	0.002	0.014	Faulty	0.006	0.018	0.072	0.016	0.107
Flint	0.014	0.001	0.004	0.009	0.014	Watery	0.012	0.013	0.065	0.018	0.095						
Mineral	0.016	0.020	0.003	0.016	0.038	Thin body/ Low body	0.026	0.035	0.006	0.060	0.101						
Earthy	0.004	0.007	0.000	0.002	0.009	Thin/ Thin mouthfeel	0.018	0.057	0.001	0.022	0.080						
Oily	0.010	0.013	0.000	0.004	0.017	Unbalanced	0.022	0.053	0.014	0.000	0.066						
Elegant	0.006	0.006	0.000	0.002	0.009	Short AT	0.004	0.001	0.050	0.003	0.055						
						Low alcohol	0.002	0.044	0.015	0.050	0.110						
						high alcohol	0.013	0.000	0.001	0.000	0.001						
						Small yield	0.005	0.000	0.038	0.009	0.047						
						Mature	0.005	0.000	0.038	0.009	0.047						
						Vibrant	0.005	0.007	0.003	0.006	0.016						

5.4 Discussion

The original idea of the project was to explore the sensory space typical of the OV Chenin Blanc wines. As required by the methodology used when testing a typicality concept, the process was laid out in steps in such a way that multiple checks were put in place. The systematic approach taken in establishing and understanding an oenological concept requires a reliable panel (judge consensus), as well as conceptual and perceptual agreement (Perrin & Pagès, 2009; Maitre *et al*., 2010). The establishment of a sensory space unique to a concept (in this case the OV Chenin Blanc) would constitute the final step in the process, which can be reached *only* once all the previous stages have been demonstrated.

In the current study, the panel agreement was proven from the rating results, even if the explained variance was distributed almost equally over a large number of dimensions. Scalar data with a single measurement has an approximately equal distribution of the explained variance across the multiple dimensions of the PCA; in other words, all dimensions have an almost equal input into the distribution of data (Granato & Ares, 2014), as observed for the results of the current work. Conversely, even if the explained variance is high, the experiment stops if panel consensus is not reached. This was the case in the study by Ballester *et al*. (2013), in which no correlations were observed in the agreement between judges assessed by PCA; in that case, there was no consensus and the investigation did not proceed further.

Only after the reliability of the panel was confirmed could the perceptual agreement be tested. The borders of the perceptual agreement can be gradual, referred to as the "degree of representativeness", and are tested using rating tasks (Ballester *et al*., 2005; Chrea *et al*., 2005). These borders can also be categorical, referred to as membership in the concept group, and are tested using sorting (Ballester *et al*., 2005). This means that the samples selected to test the concept need to cover the range of representativeness, including their borders (Ballester *et al*., 2005; Chrea *et al*., 2005).

The focus of a sorting task is the grouping of samples according to the given criteria (Valentin *et al*., 2012), in this case *old vine*/*young vine*. The instruction to describe the groups provided a secondary (verbal) aspect to the task. The flexible sorting task, as designed in this study, had both bottom-up and top-down elements to it (Lindsay & Norman, 1977; Brochet & Dubourdieu, 2001). To decide whether a sample belonged to the *old vine* group, a judge had to think first of the characteristics that qualify the sample for that category (top-down thinking). To describe the group based on the samples included, the judge had to consider the attributes of the wines themselves (bottom-up thinking).

Since these two aspects are intertwined, both the grouping and the descriptors were used to give an indication of the conceptual space related to old vine Chenin Blanc typicality. The values of the RV coefficients supported the hypothesis that the verbal and non-verbal aspects of the sorting task were in agreement.

In line with the idea related to the origin of the old vine character coming from the grapes, this study covered sample variability in terms of vine age, but limited variability from a winemaking perspective. The wines were tested as young and bottle aged. Although the same number of clusters resulted from the analysis of the sorting results for both evaluation stages, the members belonging to each of the clusters were different. Using vine age as the single source of variability may have resulted in wines being too similar to each other for the judges to be able to distinguish between them. Unlike in this study, the previous study by Crous (2016) included variability in winemaking, but not in vine age. This may have created a greater variability between the wine samples but, as often seen, highly involved winemaking practices may outweigh other factors (in this case, vine age).

Conceptually, the experts agreed on the attributes associated with the OV concept. Perceptually, the experts could not agree on a set of wines whose only variable was vine age. At this point, the process could not be taken further.

It is only once the perceptual agreement and the borders are elucidated that the attributes associated with the concept can be tested (Perrin & Pagès, 2009). This would have resulted in building and describing a sensory space unique to OV Chenin Blanc wines. The correct samples have to be consistently associated with the attributes in order for them to be considered features of the tested concept. This was not the case in the current study, where the last stage in the investigation could not be carried out due to the lack of perceptual agreement. As such, the features and the drivers of the concept could not be identified. In addition to the possible lack of variation in the resulting wines coming from a standardised winemaking, one other possible cause for the lack of perceptual agreement could be linked to the “expertise” and “exposure” factors related to the expert judges, factors highlighted in the literature in similar cases of testing complex concepts (Chrea *et al*., 2005; Perrin & Pagès, 2009). Even though the industry professionals participating in this experiment were experts in the topic, their reference (or “prototype”, as described by Chrea *et al*. (2005)) most probably was built on repetitive exposure to a variety of old vine wines, with common but also very different characters. This aspect is one of the most difficult ones in relation to ensuring consistency in concepts, in contrast to attributes or features for which the researchers can use standards and calibrate analytical panels or even experts.

Previous studies have used predictive models, such as partial least squares (PLS) (Coulon-Leroy *et al*., 2018) and multiple linear regression (MLR) (Ballester *et al*., 2005; Parr *et al*., 2015), to explore the relationship between the rating and sorting data in the case of typicality. These models work when there is both panel consensus and perceptual consensus, so that the features of the typicality concept can be correlated or predicted. Since perceptual agreement on vine age or the categories of *old vine*/*young vine* was not reached in the current study, predictive or linear regressions could not be used.

5.5 Conclusion

The South African old vine Chenin Blanc typicality was tested perceptually and conceptually. The perception of a Chenin Blanc wine as having “old vine character” was evaluated using a typicality rating and a flexible sorting task. The conceptual understanding of old vine Chenin Blanc was investigated by allowing judges to describe the *old vine* and *young vine* sorted groups.

As shown by the results, a unique sensory space of the OV Chenin Blanc could not be demonstrated because the results indicated a lack of perceptual consensus among the industry professionals during the sorting task. However, the industry professionals did demonstrate a conceptual alignment/ agreement, as demonstrated by the rating results, which was the foundation on which the rest of the work was built.

If similar work were to be repeated with commercial wines (from YV and OV), the existence of a unique sensory space of commercial OV wines could be demonstrated. However, such an experiment would still not answer the question: where is this character coming from? Researchers could get closer to answering the question by finding the features/drivers of the concept and maybe backtrack them to the origin. However, the source of the OV character could be multiple – interactions between the vineyard conditions, winemaking techniques, and vineyard and cellar flora. Even if experiments were to be designed around these factors, excluding them one by one, the interaction aspect would be lost.

The sensory space characteristic of OV Chenin Blanc wines can also be re-created by better understanding the opinions of the wine industry professionals. Qualitative approaches such as interviews and surveys would be insightful.

These results show that, conceptually, the experts agreed on the attributes of old vine Chenin Blanc wines, although they could not align perceptually. Since variability in winemaking was factored out, the unique properties gained by the wine during winemaking and the inclusion of viticultural and microbiome elements (wild fermentations) have been lost. However, if the guidelines of the OVP to take the minimalistic approach are to be followed, it is put into perspective how the various approaches taken in winemaking practices influence the final product.

References

Abdi, H., 2007. RV coefficient and congruence coefficient. In Salkind, N. (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks, CA, Sage. 1 – 10.

Ballester, J., Dacremont, C., Le Fur, Y. & Etievant, P., 2005. The role of olfaction in the elaboration and use of the Chardonnay wine concept. Food Qual. Prefer. 16, 351-359.

Ballester, J., Mihnea, M., Peyron, D. & Valentin, D., 2013. Exploring minerality of Burgundy Chardonnay wines: A sensory approach with wine experts and trained panellists. Aust. J. Grape Wine Res. 19(2), 140-152.

Ballester, J., Patris, B., Symoneaux, R. & Valentin, D., 2008. Conceptual vs. perceptual wine spaces: Does expertise matter? Food Qual. Prefer. 19(3), 267-276.

Brochet, F. & Dubourdieu, D., 2001. Wine descriptive language supports cognitive specificity of chemical senses. Brain Lang. 77, 187-196.

Casabianca, F., Sylvander, B., Noël, Y., Béranger, C., Coulon, J.-B., Giraud, G., Flutet, G., Roncin, F. & Vincent, E., 2006. Terroir et Typicité: proposition de définitions pour deux notions essentielles à l'appréhention des Indications Géographiques et du dévelopement durable. In: VIth Int. Terroir Congr. Vol. 1, 544-551.

Chrea, C., Valentin, D., Sulmont-Rossé, C., Nguyen, D.H. & Abdi, H., 2005. Semantic, typicality and odor representation: A cross-cultural study. Chem. Senses 30(1), 37-49.

Coulon-Leroy, C., Poulzagues, N., Cayla, L., Symoneaux, R. & Masson, G., 2018. Is the typicality of "Provence Rosé wines" only a matter of color? Oeno One 52(4), 1-15.

Crous, R., 2016. The sensory characterisation of old-vine Chenin blanc wine: An exploratory study of the dimensions of quality. Thesis, Stellenbosch University, Private BagX1, 7602 Matieland (Stellenbosch), South Africa.

Granato, D. & Ares, G., 2014. Mathematical and statistical methods in food science and technology. John Wiley & Sons, Chichester, UK.

Lindsay, P.H. & Norman, D.A., 1977. Human information processing: An introduction to psychology. Academic Press, Cambridge, MA, USA.

Maitre, I., Symoneaux, R., Jourjon, F. & Mehinagic, E., 2010. Sensory typicality of wines: How scientists have recently dealt with this subject. Food Qual. Prefer. 21(7), 726-731.

Niimi, J., Danner, L. & Bastian, S.E., 2019. Wine leads us by our heart not our head: Emotions and the wine consumer. Curr. Opin. Food Sci. 27, 23-28.

Old Vine Project (OVP), n.d. Certification process. Retrieved 29 April, 2020, from http://oldvineproject.co.za/old-vine-project-certification-process/

Parr, W.V, Ballester, J., Peyron, D., Grose, C. & Valentin, D., 2015. Perceived minerality in Sauvignon wines: Influence of culture and perception mode. Food Qual. Prefer. 41, 121-132.

Perrin, L. & Pagès, J., 2009. A methodology for the analysis of sensory typicality judgments. J. Sens. Stud. 24(5), 749-773.

Salkind, N. (ed.), 2012. Encyclopedia of Measurement and Statistics. Thousand Oaks, CA, Sage.

SASEV, 2018. Unlocking value in South Africa's old vine resources. Workshop presented at SASEV-Winetech 41st International Conference, October 2018, Somerset West, South Africa.

SAWIS, 2018. Status of wine-grape vines as on 31 December 2018. Retrieved 31 August, 2020, from http://www.sawis.co.za/info/download/ Vineyards_2018_final.pdf

Schouteten, J.J., De Steur, H., De Pelsmaeker, S., Lagast, S., De Bourdeaudhuij, I. & Gellynck, X., 2015. An integrated method for the emotional conceptualization and sensory characterization of food products: The EmoSensory®Wheel. Food Res. Int. 78, 96-107.

Stevenson, T., 2005 (4th ed.). The new Sotheby's wine encyclopedia. Dorling Kindersley Limited, London, UK.

Valentin, D., Chollet, S., Lelievre, M. & Abdi, H., 2012. Quick and dirty but still pretty good: A review of new descriptive methods in food science. Int. J. Food Sci. Technol. 47(8), 1563-1578.

Chapter 6

Research results

Data fusion using Multiple Factor Analysis coupled with non-linear pattern recognition (fuzzy k-means)

Chapter 6: Data fusion using Multiple Factor Analysis coupled with non-linear pattern recognition (fuzzy k-means)

6.1 Introduction

Pattern recognition in complex natural systems such as wine requires gathering not just a large amount of data (and corresponding data variation) but diverse and informational rich data (and corresponding variability of measured parameters). Oenological studies hence include variation by measuring several categories of chemical (*e.g*. volatile and non-volatile composition) and sensory (verbal and non-verbal) data (Stevenson, 2005; Valentin *et al.*, 2012). Data variability can be included through the use of both targeted and untargeted chemistry data and/or the combination of sensory with chemistry results. Combining these different sets requires appropriate data integration using statistical methods of data fusion and intelligent strategic approaches (Borràs *et al.*, 2015; Biancolillo *et al.*, 2019; Cocchi, 2019). Methods of data fusion for Oenology are usually supervised, seeking to find optimal discrimination of samples based on cultivar, origin, age, among others (Borràs *et al.*, 2015; Biancolillo *et al.*, 2019; De Carvalho Rocha *et al.*, 2020). The previous chapter (Chapter 4 – *Exploration of data fusion strategies using Principal Component Analysis (PCA) and Multiple Factor Analysis (MFA)*) presented the different strategies and advantages of constructing unsupervised data fusion models. It is not only the methods (chemistry, sensory, and statistical) chosen that are important, but the strategy (data handling process) itself.

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

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

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

Parametric techniques used for pattern recognition include variants of cluster analysis (agglomerative hierarchical clustering/AHC, hierarchical cluster analysis/HCA,), linear regression analysis (linear discriminant analysis/LDA and multiple linear regression/MLR), and discriminant/ classification analysis (partial least squares/PLS –and discriminant analysis/DA) (McKillup, 2012). Non-parametric groups of methods for pattern recognition such as k-nearest neighbours (kNN), artificial neural networks (ANN), and support vector machines (SVM) consider non-binary relationships (Härdle & Simar, 2015; Radovanovic *et al.*, 2016; Myhre *et al.*, 2018; De Carvalho Rocha *et al.*, 2020).

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

Fuzzy k-means, originally referred to as fuzzy c-means (Bezdek, 1981), is a variant of k-NN that uses fuzzy algorithms for clustering observations (samples). The partitioning (grouping) of the clusters can be based on an average centroid, class membership (group designation), or random assignment of centroids (Bezdek, 1981). This is different from hard clustering techniques such as AHC and classification techniques such as PLS which impose strict partitioning (McKillup, 2012; Härdle & Simar, 2015). Of the available machine learning algorithms and based on the case studies by Myher *et al*., (2018) and studies on fuzzy c-means (Khang *et al.*, 2020), fuzzy k-means was the most appropriate technique for this study.

The current study explored an unsupervised strategy consisting of data fusion coupled with pattern recognition. It included variation in the form of combinations of data sets (*i.e.* chemistry and sensory) and information-rich techniques (*i.e.* fingerprinting by NMR and HRMS). MFA was used for the fusion of sensory (verbal and nonverbal flexible sorting) and chemistry (NMR and HRMS) data. Pattern recognition was done using AHC and fuzzy k-means. Fuzzy k-means was explored by varying coefficients of fuzziness and the number of clusters. The effectiveness of the different strategies of cluster analysis were evaluated using coefficients of variance in optimal classification (*i.e.* Wilks' Lambda and cophenetic correlation coefficient).

6.2 Materials and Methods

6.2.1 Experimental design: winemaking and sensory analysis

The winemaking, wine treatments as well as sensory evaluation have been previously published (Mafata *et al.*, 2020). In brief, the experimental design consisted of 23 wines made from grapes harvested from vines aged 5 to 45 years old and made using the same vinification protocol. The wines were evaluated twice: three months in the bottle (Year 1) and two years in the bottle (Year 2). Sensory analysis consisted of a flexible sorting task with a non-verbal (grouping) and a verbal (group description) aspects[4]. Full details are given in Chapter 5 – *Investigating the Concept of South African Old Vine Chenin Blanc.*

6.2.2 Chemical analysis

6.2.2.1 High resolution mass spectrometry (HRMS)

Sample preparation: The 23 wine samples were clarified through centrifugation at 3,000 rotations per minute (rpm) and sampled into a 2 mL HPLC vial.

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

Data processing: Using the MarkerLynx software (Waters Corporation, Milford, MA, USA) the data was integrated and extracted as a (RT_m/z, ion abundance) matrix.

6.2.2.2 Nuclear magnetic resonance (NMR) spectroscopy

Sample preparation: The wine pH was adjusted to 3 using 10M HCl and 10M NaOH solutions. In a 5 mm Wilmad NMR tube (Rototec-Spintec GmbH, Bad Wildbad, Germany), 50 µL of a 1M Tetramethylsilane standard (TMS in D_2O) and 500 µL of the pH adjusted wine were added. Nitrogen gas was passed over the headspace for preservation until analysis.

Instrumental acquisition: NMR was performed using a Varian Unity Inova 600MHz liquid state NMR Spectrometer and based on methods by López-Rituerto *et al.*, (2012) and Godelmann *et al.,* (2013). All spectra were acquired at 298.0 K. Manual tuning, matching, locking, and shimming were performed for each sample analysis. Two 1-H NMR analyses were performed, a pre-saturation for suppression of the water and ethanol signals followed by 1-D NOESY experiment at 256 scans per sample (Figure 6.1).

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

[4] The typicality rating results were excluded from this study because one of the samples from Year 2 was missing which made the data set incompatible with full data fusion.

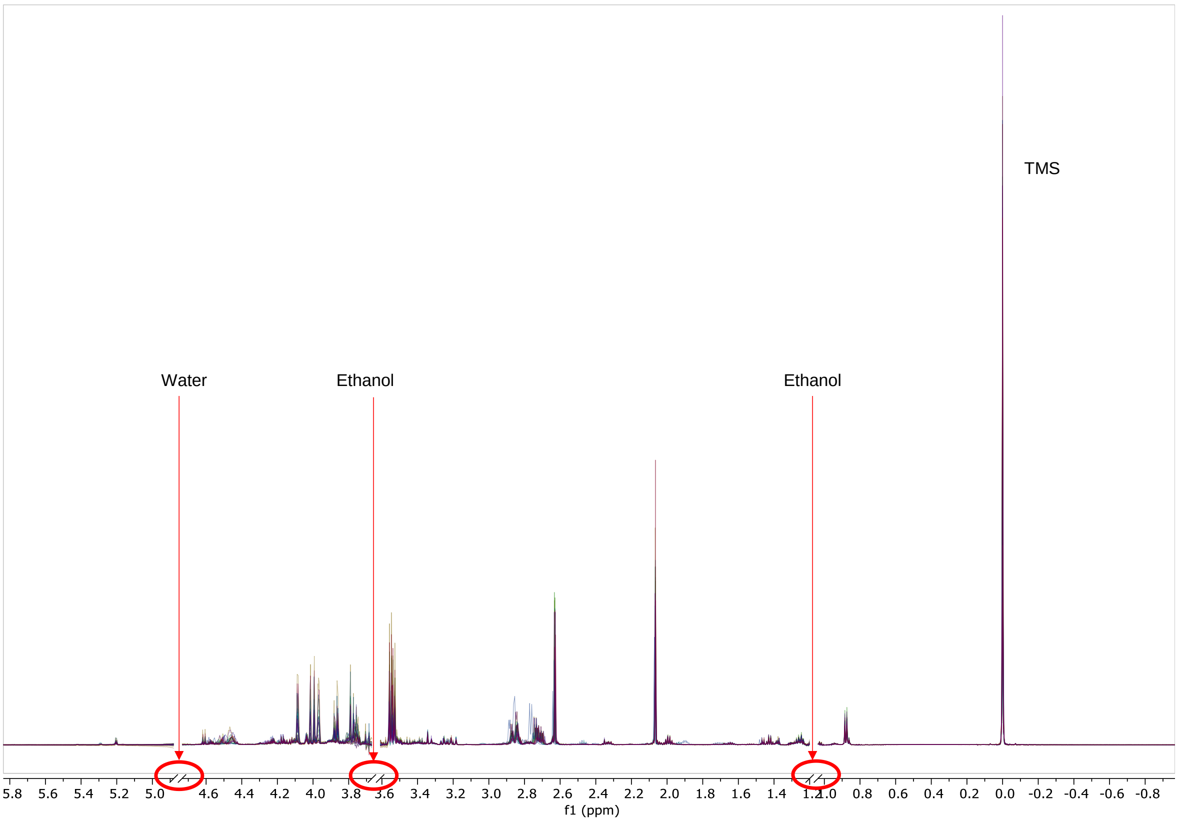


Figure 6.2: Nuclear Magnetic Resonance (NMR) spectral overlay for the 23 experimental wines analysed in Year 2.

6.2.3 Statistical analysis

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

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

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

Pattern recognition was done on the ordinal matrix from the MFA, by AHC and fuzzy k-means using XLSTAT software (version 2020, Addinsoft, New York, USA). AHC followed the Ward's method of agglomeration based on the Euclidean distance. Fuzzy k-means clustering was done using the Wilks' Lambda based on the Euclidean distance. Partitioning was done randomly with repetitions set at *n* = number of samples. Additional visualisations (graphical illustrations of the model output) were built using Statistica™ 13 (TIBCO, Dell software, Inc., Texas, United States).

6.3 Results and discussion

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

6.3.1 Exploratory data analysis

The sensory data sets (non-verbal and verbal aspects of sorting) were explored in a previous chapter (Chapter 5 – *Investigating the Concept of South African Old Vine Chenin Blanc*). Here, the performance of the models of the chemistry data sets will be discussed. This section will also include remarks on the application.

6.3.1.1 High resolution mass spectrometry (HRMS)

The HRMS fingerprint for wine is commonly acquired in the positive mode since more compounds present in wine can be ionised in positive mode and thus give an MS signal but metabolomic and untargeted studies use both ionisation modes (Alañón *et al.*, 2015). Thus, in this study, the fingerprint included acquisition in the negative mode since it too contains discriminating features, although fewer than the positive mode. The two modes contain important information that can be used to discriminate samples in supervised modelling strategies. A block PCA analysis which keeps the two modes separate can elucidate the relationship between the two.

The configuration of the samples (Figure 6.2) for the two modes can be used an indication of the different information the acquisition modes contain. The configurations derived from the positive and negative mode have an RV coefficient of 0.64 (Figure 6.2A1 *vs* 6.2A2) and 0.62 (Figure 6.2B1 vs 6.2B2) for Year 1 and Year 2, respectively.

The complete HRMS fingerprint was a combination of both MS ionisation modes (positive/Pos. and negative/Neg.). The full fingerprint consisted of 187 and 257 unique features for Year 1 and Year 2, respectively (Supplementary Figure 6.1). These were then treated as the new variables and modelled by PCA (Figure 6.3). The new combined models were configurationally more similar to each mode than they were to each other. This was indicated by high RV coefficient, 0.98 and 0.80 (combined *vs* Pos. and *vs* Neg.), and 0.99 and 0.73 (combined *vs* Pos. and *vs* Neg.) for Year 1 and Year 2, respectively.

The scree plot (Supplementary Figure 6.2) is an indication of the efficiency of the data model. A steep decline of the stress in Year 2 results implied greater efficiency, a consequence of the higher number of unique features compared to the Year 1 data set. The inflection point (the point at which the stress begins to plateau) was around the fifth dimension for both years, for which 61% and 69% of the total variation can be explained for each year, respectively.

In the context of the original application, although both the separate modes and the combined HRMS data sets provide unique fingerprints, neither were able to distinguish the old vine samples from the young vine samples. Due to this, there were no markers that were linked to or could be used to discriminate the samples according to vine age or class designation (young vine and old vine). As an exploratory analysis, the unsupervised PCA revealed no problematic issues with the HRMS data set and thus no corrective pre-processing was done. Therefore, the combined modes, without any feature selection will be used for the data fusion.

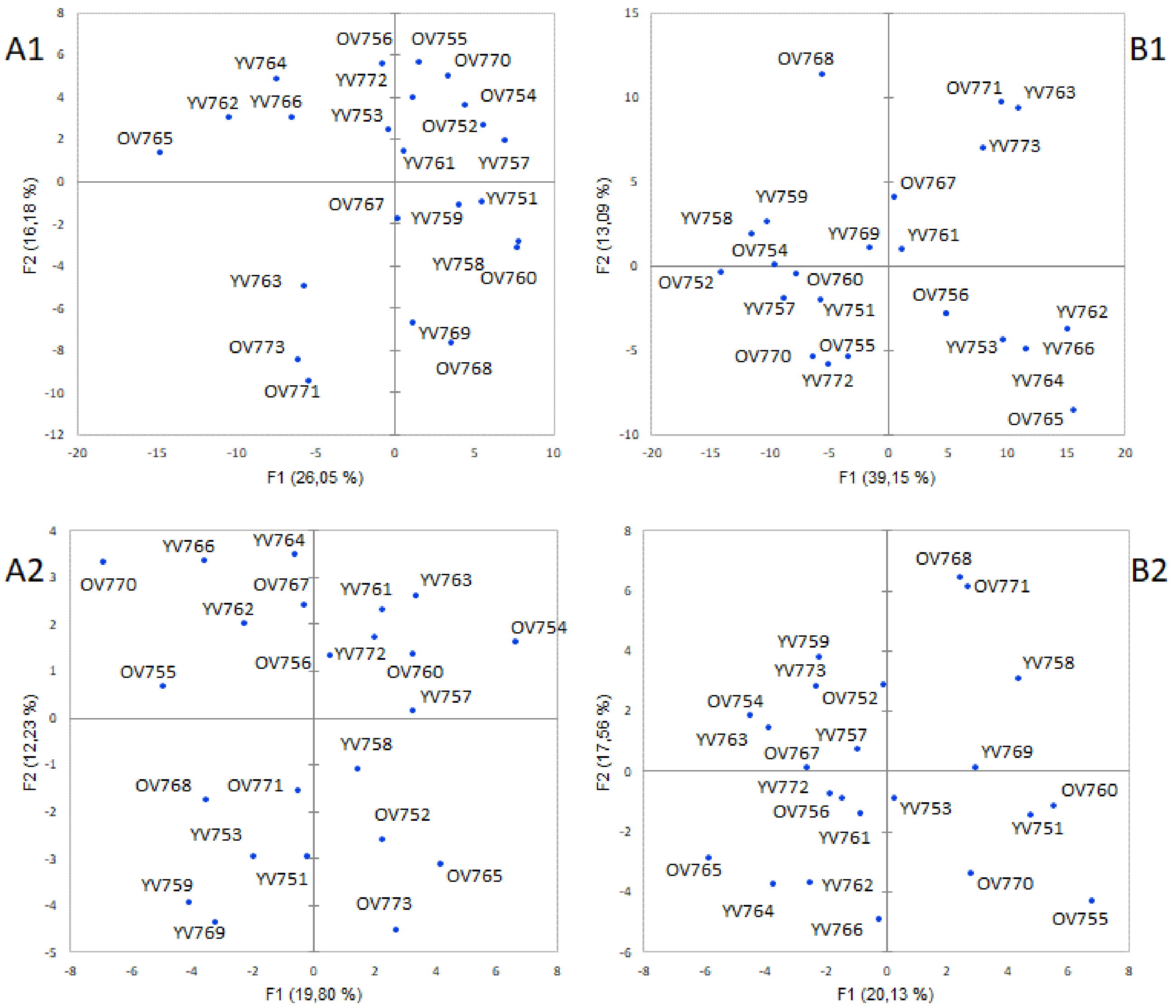


Figure 6.2: Block PCA analysis on HRMS data acquisition for Year 1 in positive mode (A1) and negative mode (A2), and Year 2 in positive mode (B1) and negative mode (B2).

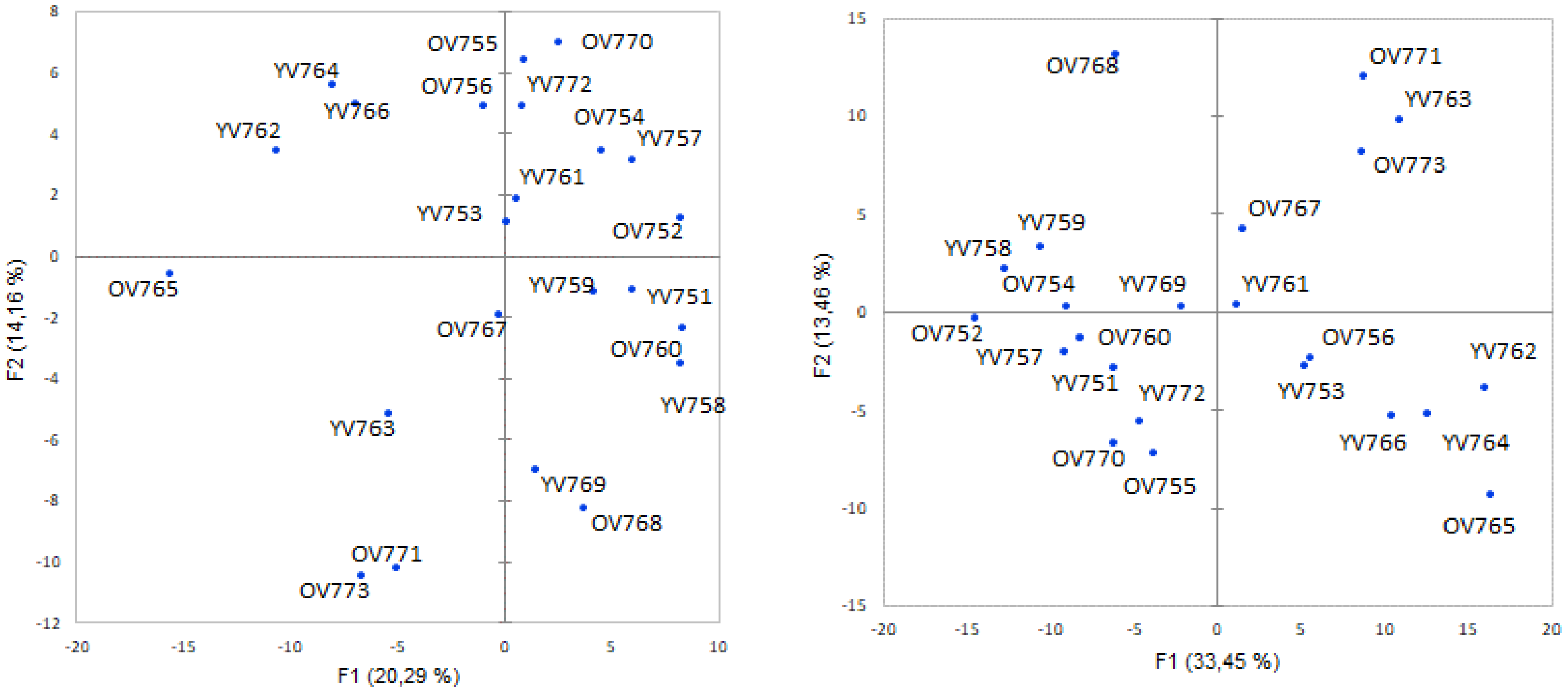


Figure 6.3: PCA scores of HRMS in combined (Pos. and Neg.) acquisition modes for Year 1 (top) and Year 2 (bottom).

6.3.1.2 Nuclear magnetic resonance (NMR) spectroscopy

NMR data was processed to extract the spectra as untargeted data in the form of new variables (“bins’’). This resulted in 390 and 392 variables for Year 1 and Year 2, respectively. These new variables were then modelled by PCA, and the efficiency of the model is illustrated by the scree plot in supplementary Figure 6.3. The plot shows high efficiency in the models, with the inflection at F3 corresponding to 90% of the cumulative variance for both years. The NMR configuration map of the samples (Figure 6.4A1 and 6.4B1) shows a particularly unique distribution for samples analysed during Year 1. With the exception of two samples (OV754 and OV756), there is an element of linearity diagonally across the first and second dimensions (at Y=X). The loadings indicate that the linearity may be due to intensity variations (McKillup, 2012). However, this phenomenon was not observed for Year 2.

To try and find reasons for the linear variation in Year 1 data set, and for further exploration of the NMR data, the spectra were submitted to block analysis. Three important spectral regions were identified and processed separately: the alkyl region (≤3.61 ppm), carbohydrate region (3.67 to 4.74 ppm), and aromatic region (≥4.88 ppm). PCA on the carbohydrate region for Year 1 (Figure 6.5) is the only region that exhibited the linearity seen in the full NMR spectra. Correlated with the linear trajectory are 6 bins: (2.998 to 3.038ppm, malic acid), (3.08 to 3.078, citric acid), (3.518 to 3.558, glycerol/fructose), (4.398 to 4.438, tartaric acid), (4.438 to 4.478, lactose), and (4.598 to 4.638, fucose) tentative assignment based on literature (López-Rituerto *et al.*, 2012; Mascellani *et al.*, 2021). The other regions (alkyl and aromatic) showed no discernible patterns in the spectra (observationally) or PCA (statistically). Year 2 results showed no discernible patterns, with high similarity between all three regions (RV ranging from 0.70 to 0.91).

Looking at the high RV coefficient values for the full NMR spectra *vs* the carbohydrate region for Year 1, an additional exploratory approach was considered – namely MFA. MFA is a multiblock approach that first builds individual PCAs for each block (*i.e.* NMR region) and then separately scales them according to their eigenvalues before building the final model. By doing this, the method avoids skewing by any one block (Abdi & Valentin, 2007). Since the data fusion strategy in this study is also MFA, the NMR scaling issue was explored in two ways, namely, doing data fusion with the regions separate and first fusing the regions by MFA and using the MFA as the representative NMR fingerprint. The latter resulted in increased RV coefficients (RV>0.85 regions *vs* MFA) indicating that the issue may have been differences in scale between the regions (Supplementary Table 6.1). For an unsupervised exploratory aim, the MFA approach would have been optimal; in order to advance the aims of the current study in devising a systematic approach to coupling data fusion with pattern recognition, the full spectra were used further without prior block analysis.

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

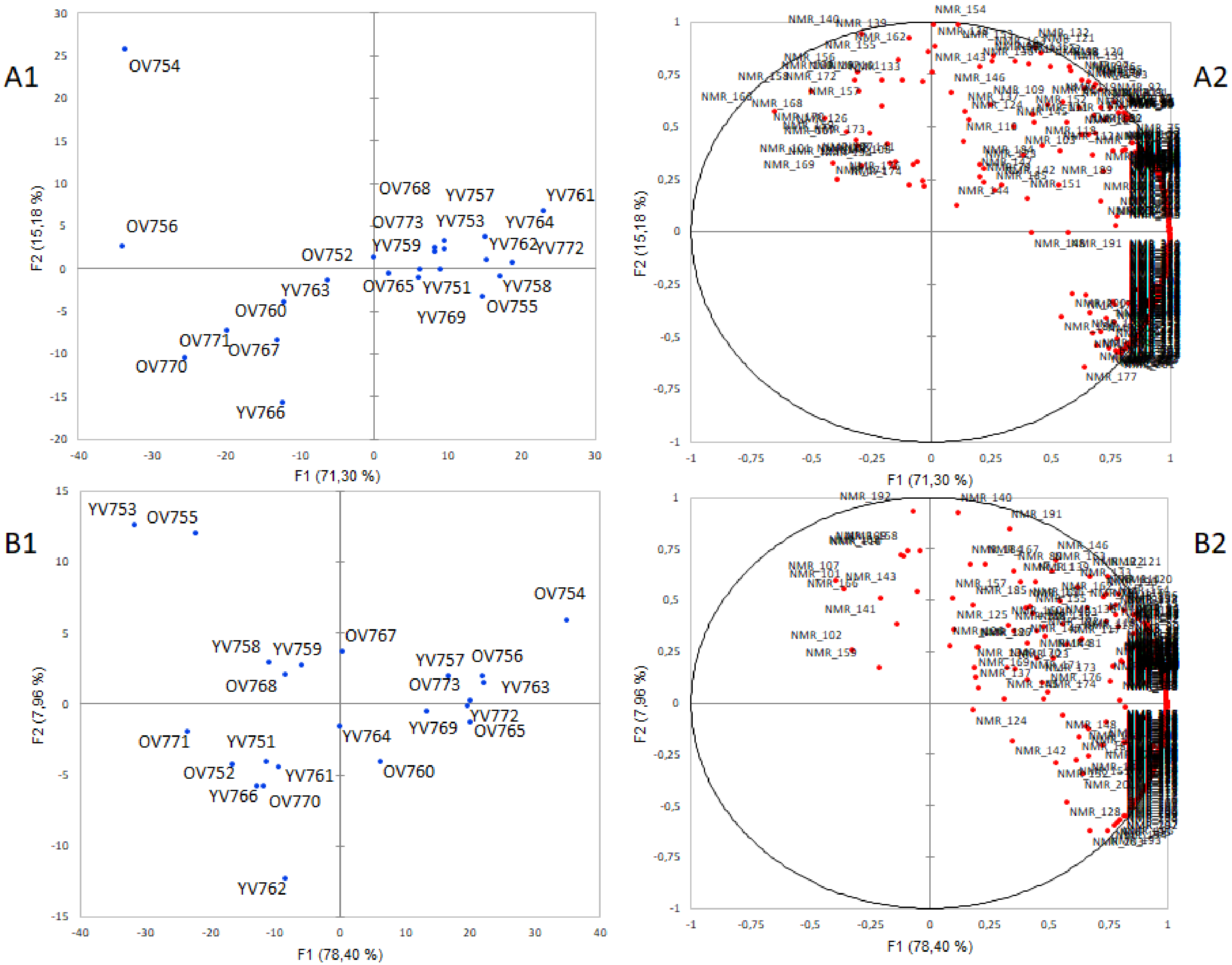


Figure 6.4: PCA models of NMR data for Year 1 (A1 – scores, A2 - loadings) and Year 2 (B1 – scores, B2 - loadings).

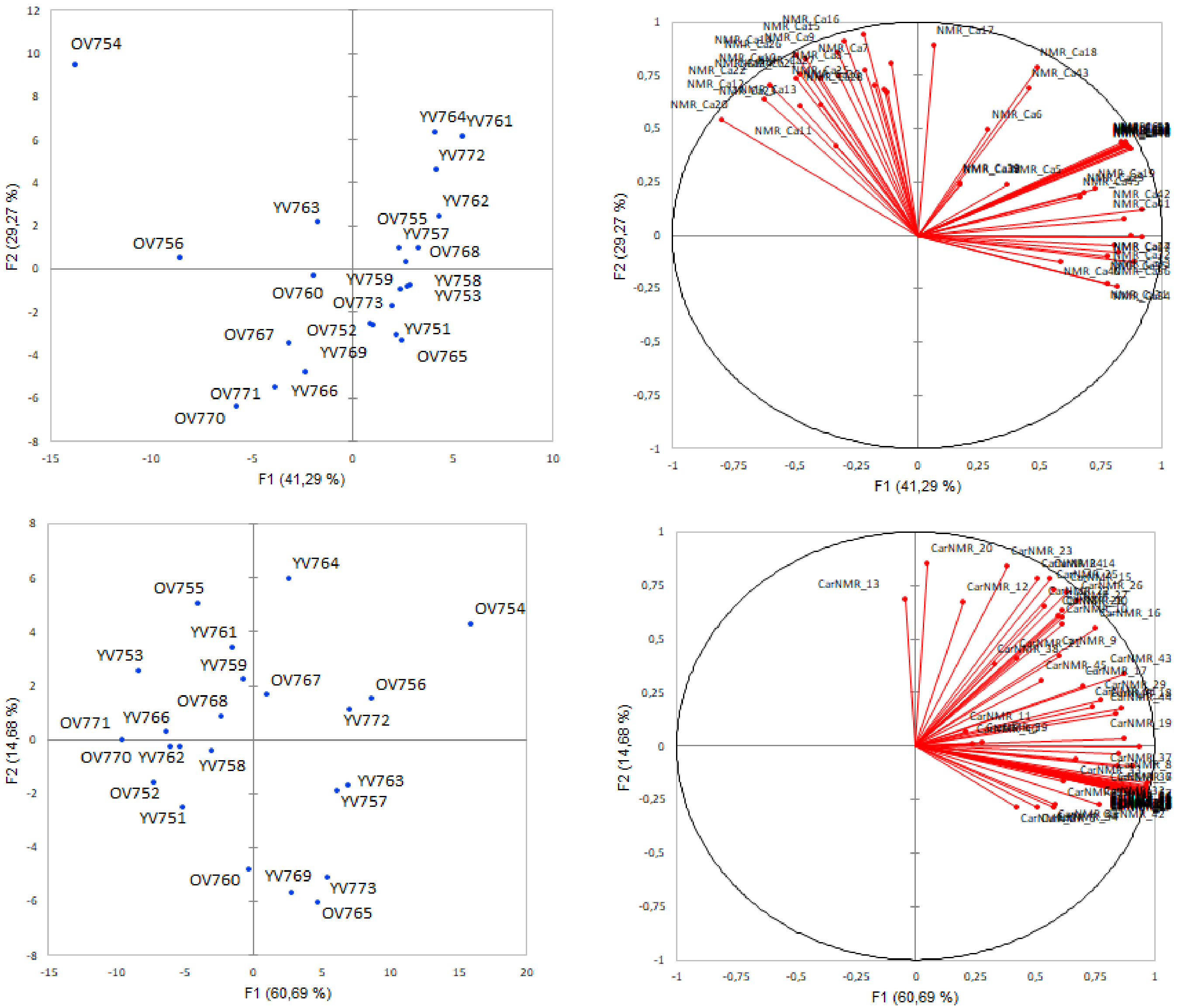


Figure 6.5: Carbohydrate region of the NMR for Year 1 (top) and Year 2 (bottom) showing scores (left) and loading plots (right).

6.3.2 Data fusion by Multiple factor analysis (MFA)

The performance of the data fusion models is indicated by the decay in stress, visualised using the scree plots (Supplementary Figure 6.4). Year 1 saw a total of 649 observations modelled, while Year 2 had 743 observations, due to more HRMS features and verbal sorting terms. Conversely, the Year 1 stress (2.54 eigenvalue) is relatively higher compared to Year 2 (2.18 eigenvalue). This means that there are factors influencing the efficiency of the data fusion models other than the obvious number of observations. The scree plots show that the inflection point for Year 1 (F8, corresponding to 53% cumulative variance) and Year 2 (F9, corresponding to 59% cumulative variance) are similar. Year 2 had better performance in terms of lower stress (eigenvalue) and higher efficiency (% cumulative variance) since 28% is packed in the first three dimension compared to 24% for Year 1.

As a multiblock approach, the performance of the MFA data fusion models is determined by the blocks and their relationships to one another. The relationship of the four blocks in this case (HRMS, NMR, verbal sorting/VSorting, and non-verbal sorting/NVSorting) can be seen in the three-dimensional representation of the group factor map (GFM) from the MFA (Figure 6.6). There was little overlap between the four data sets, indicating that they contain different information. Accompanying the GFM are pairwise RV coefficients (Table 6.1) calculated on the scores of each data set. Emphasising the point, the RV coefficients were low between the data sets, the highest being HRMS *vs* NVSorting (0.69) for Year 1 and NVSorting *vs* verbal sorting (0.60) for Year 2.

The low RV coefficients between the individual data sets indicate unique information contained in each data set and are a positive result for unsupervised data fusion. In data fusion it is preferable to have not only informational density but also informational variability. Informational variability guarantees low redundancies in the configurational maps of the samples and observations lowering incidences of false correlations. The MFA data fusion models were representative, indicated by high RV coefficients between the MFA and the individual data sets for both years (Table 6.1). However, NMR exhibited low RV coefficients *vs* the MFA for both Year 1 (0.42) and Year 2 (0.45), due to the scaling issues previously discussed in Section 6.3.1.2).

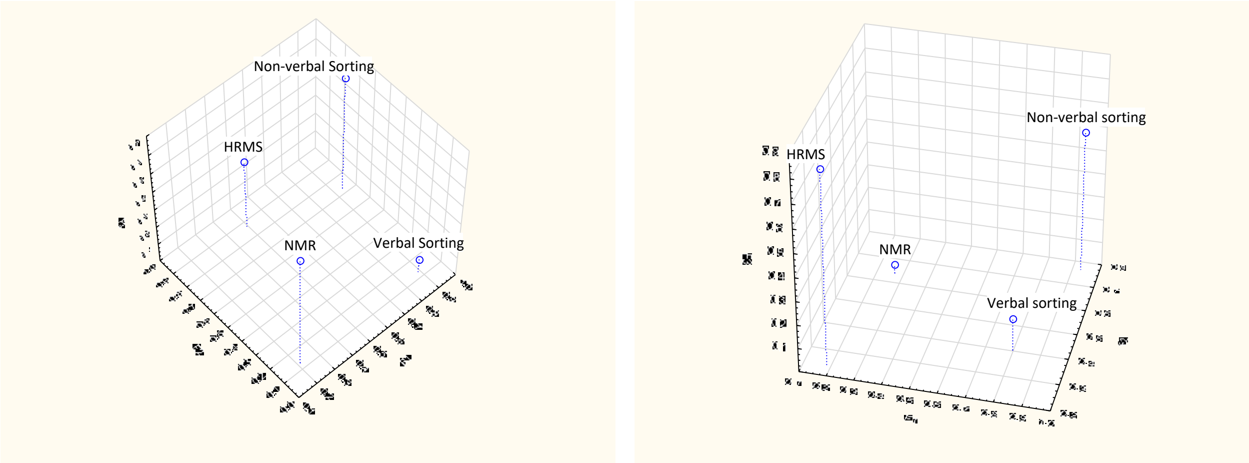


Figure 6.6: MFA group factor map for Year 1 (left) and Year 2 (right) depicted in the first three dimensions.

Table 6.4: Pairwise RV coefficients for MFA scores configurations between the four individual data sets and the MFA data fusion models for both Year 1 and Year 2 of the study.

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

The advantage of data integration is that inferences on discriminating or unique elements can be made using the various compilation of data sets, facilitating comprehensive problem identification and description. This can be exemplified by the identification of the unique samples OV765 (Year 1 and Year 2) and YV769 (Year 1) (Figure 6.7). The partial axes from the MFA data fusion model showed the data sets from which the difference comes. The graphs show that in both years, the sensory analysis (V and NVSorting) consistently discriminated the unique samples from the rest of the samples. The same differences were noted in the analysis of the sensory results (Chapter 5, Sections 5.3.2 and 5.3.3). The contextual meaning of these results was also discussed in those sections.

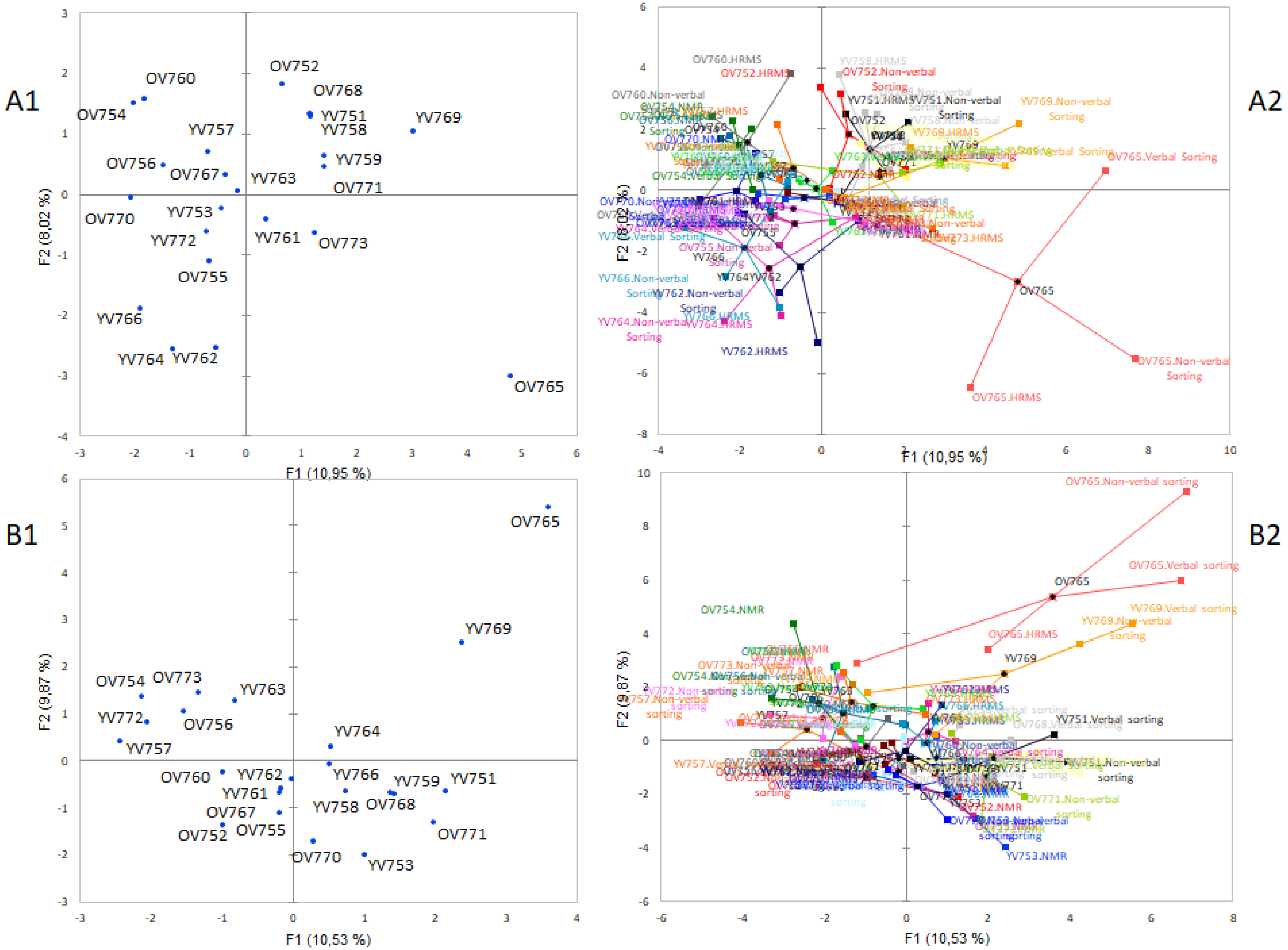


Figure 6.7: MFA results showing the scores (A1 and B1) and the coordinates for the projected points (A2 and B2) for Year 1 (top) and Year 2 (bottom).

6.3.3 Pattern recognition

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

A few patterns have already been highlighted in the previous sections of this chapter; they were based on exceptions and not on discernible differences/patterns between samples according to the vine age or class designation. In order to try to identify these patterns, an extensive pattern recognition strategy was evaluated next. As previously discussed, fuzzy clustering is best recommended for attempting to discriminate samples for data with high similarity due to its sensitivity (Radovanovic *et al.*, 2016; Myhre *et al.*, 2018).

Pattern recognition in the form of parametric (AHC) and non-parametric (fuzzy k-means) cluster analyses was done on the correlation matrix of the MFA samples. Figure 6.7 shows two-dimensional representations of the MFA factor maps of the samples. The stress distribution of the MFA was gradual, having 50%EV over eight factors for both years (Supplementary Figure 6.4). In order to get a fair representation of the relationship between samples, AHC dendrogram was plotted for all factors (Figure 6.8). From the dendrograms, it is evident that sample OV765 for year 1 and samples OV765 and YV769 for year 2 are outliers. Due to the sensitivity of the fuzzy k-means method and the fact that the strategy used in this study is unsupervised (using random partitioning instead of supervised clustering such as centroid or class membership), these outlying samples were removed before any further clustering was applied. The random partitioning was chosen because:

1. from the descriptive study (Chapter 5 – *Investigating the Concept of South African Old Vine Chenin Blanc*) it was shown that there was no prototype that could be used as the best average example for classification;
2. the exploratory stages (Section 6.3.1) showed no distinguishable pattern between the samples;
3. random partitioning was the most compatible approach to unsupervised clustering.

The outlying samples had a big effect on the clustering; removing the outliers resulted in the cophenetic correlation coefficient decreasing from 0.637 to 0.403 in year 1 and 0.813 to 0.464 in year 2. The percentage of the total variation within-class was 91% for year 1 and 95% for year 2. Both the high variation and the low cophenetic correlation coefficient mean that the assignment of members of each cluster was unreliable (Bezdek, 1981). Fuzzy k-means is more reliable than AHC at assigning cluster membership.

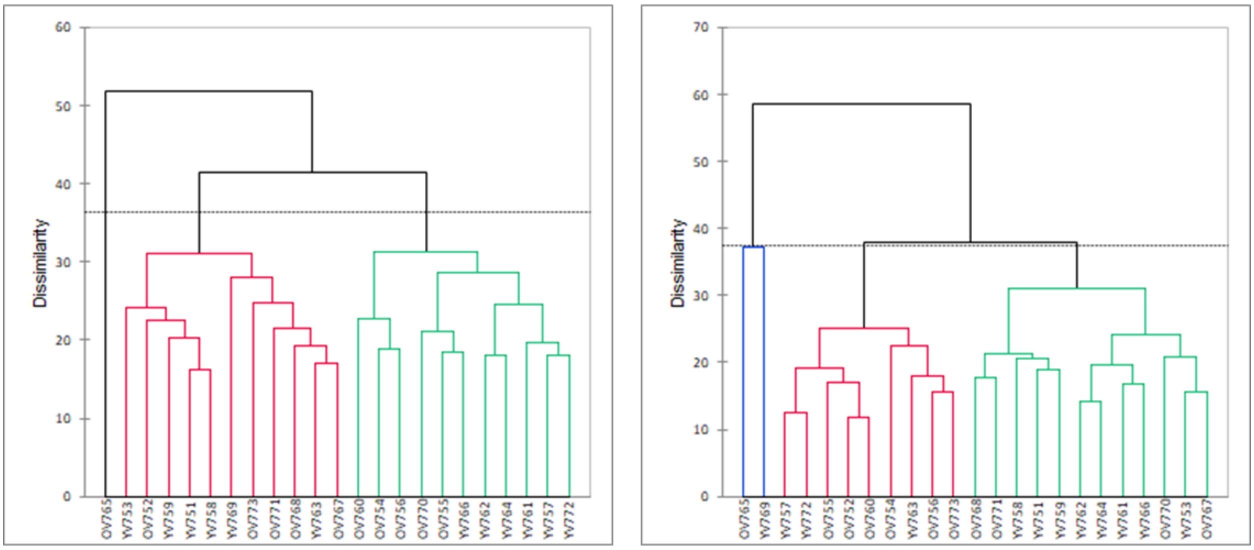


Figure 6.8: Agglomerative hierarchical clustering (AHC) dendrogram on the MFA of Year 1 (left) and Year 2 (right).

The fuzzy k-means was explored by varying the fuzzy partition coefficient (F_k) as well as the number of clusters (k). Since this study is using an unsupervised approach, reasonable F_k at the upper-limit (1.05) and lower-limit (1.001) were explored (Figure 6.9). Various degrees outside these parameters (*i.e.* F_k>1.05) were also investigated and these values were chosen because they illustrated the sensitivity of the fuzzy k-means technique (data not shown).

The objective function of the clustering, from which the goodness-of-fit criterion is based, is to maximize the discrimination between clusters based on Euclidean distance in this case (Bezdek, 1981). The goodness-of-fit criterion in this analysis is analogous to the R^2 used for goodness-of-fit used for PCAs, for instance (Härdle & Simar, 2015). Both Year 1 and Year 2 results showed high goodness-of-fit criterion at k<9 (Figure 6.9) but the Wilks' Lambda coefficient values were poor. This means that although the data was well fitted, the class membership was unreliable. The Wilks' Lambda coefficient was lower (*i.e.* λ<1) at k≥12 for both Year 1 and Year 2, and the percentage goodness-of-fit criterion increased (Figure 6.9). When F_k was set at 1.001 both the goodness-of-fit criterion and Wilks' Lambda coefficient improved (λ=0.085 for both years) (Supplementary Table 6.2). Hence, the optimal clustering was with twelve clusters at a fuzzy coefficient of 1.001 for both years. Although the λ values reported in this study were relatively high compared to the parametric AHC, the fuzzy k-means clustering was more reliable.

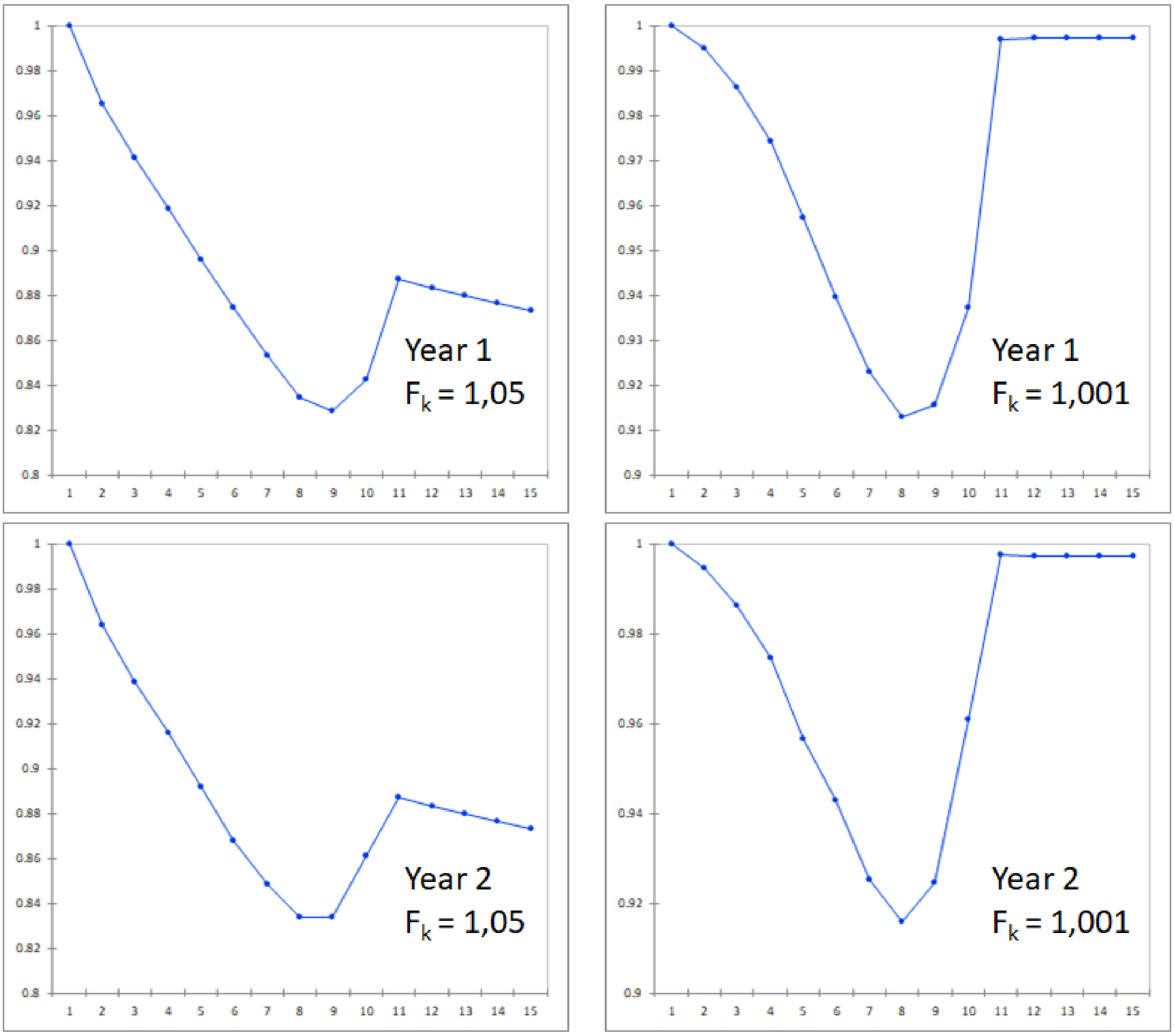


Figure 6.9: Evolution of the goodness-of-fit criterion in the fuzzy k-means analysis of Year 1 (top) and Year 2 (bottom) with F_k set at the upper-limit of 1.05 (left) and the lower-limit of 1.001 (right).

Correspondence analysis (CA) was done on the membership probabilities at the optimal clustering by treating them as categorical/ frequency-based data (Figure 6.10) (McKillup, 2005), analogous to the use of CA in sensory sorting tasks (Valentin *et al.*, 2012; Cariou & Qannari, 2018). The discrimination between samples was better at F_k set at 1.001 than at 1.05. Cluster analysis on the CA showed improved clustering compared to the parametric AHC without fuzzy k-means. Year 1 showed an increase in the cophenetic correlation coefficient, from 0.403 to 0.869 and the within-class variation decreased from 96% to 16% (Supplementary Table 6.3). For an unsupervised approach, these are markers of reliable clustering and discrimination between samples when using fuzzy k-means (Bezdek, 1981). Similarly, Year 2 showed an increase in the cophenetic correlation coefficient, from 0.464 to 0.835 but minimal decrease in the within-class variation (from 95% to 87%) (Supplementary Table 6.3).

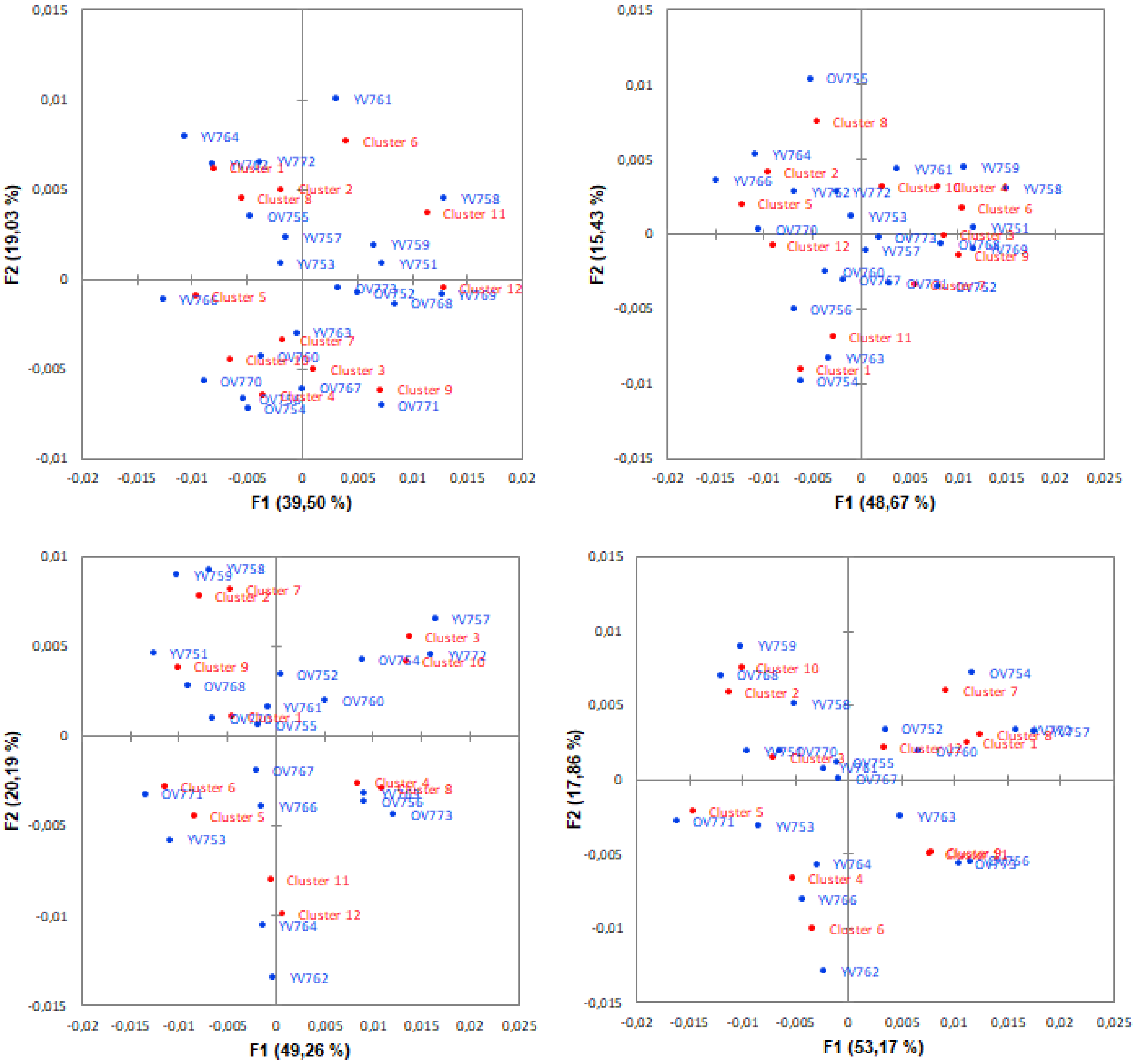


Figure 6.10: Correspondence analysis (CA) biplots on the membership probabilities at optimal clustering conditions for Year 1 (top) and Year 2 (bottom) with F_k set at 1.001 (left) vs 1.05 (right).

6.3.3.2 Contextual interpretation of the cluster analysis

The contextual sensory perspective elements related to this application were investigated based on the typicality of old vine South African Chenin Blanc from class designation and age of vines (Mafata *et al.*, 2020). The discussion of the results in the previous study showed no indication of suitable prototypes (typical of old vine wine) or distinguishable age-defined border between the classes (young vine wine *vs* old vine wine). Based on these findings, the fuzzy clustering approach could not use class membership or centroid partitioning. If a cluster analysis was to be performed for the four individual data sets used in the data fusion (NMR, HRMS, Non-verbal, and Verbal sorting), it would, much like the AHC, show the random distribution of samples that demonstrates the high similarity between samples.

Both statistical and contextual random effects of grouping were observed in this case. Statistically, by varying the different parameters in the fuzzy clustering (*i.e*. F_k and k), although the discrimination power was improved, there were no observable core centroids (average sample representative for each cluster). Hence, the algorithm could only exclude the most dissimilar sample at a time (Figure 6.11). The fuzzy k-means dendrograms show that the wines examined during Year 1 are more discriminable from one another than the aged wines of Year 2 (Figure

6.11).

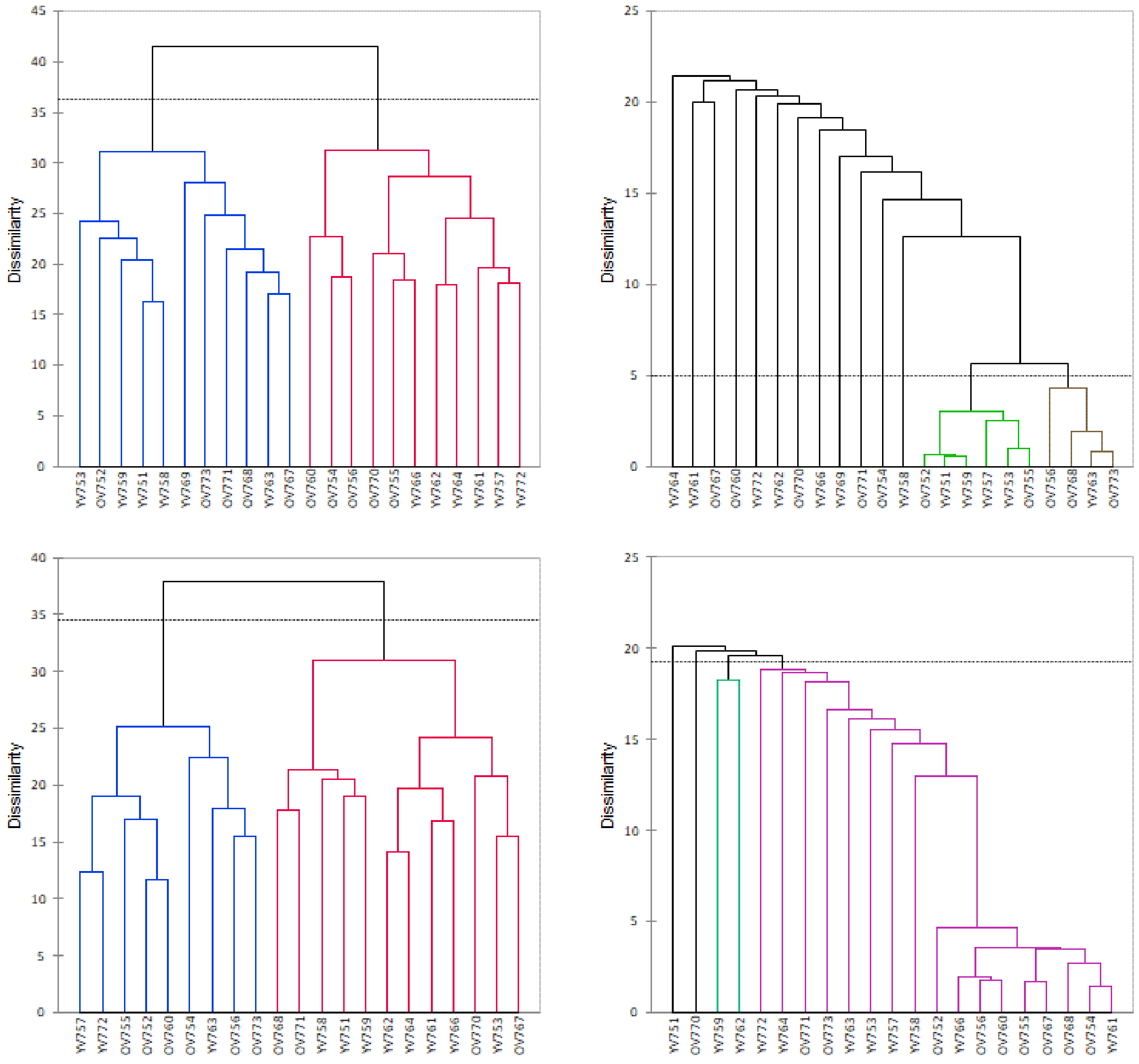


Figure 6.11: AHC dendrogram on the parametric clustering (left) and fuzzy k-means clustering (right) results for Year 1 (top) and Year 2 (bottom) with optimal clustering at 12 clusters and the fuzzy coefficient (F_k) set at 1.001.

The random associations inherent in this data was demonstrated at the individual data set explorations (Section 6.3.1) and most importantly in the data fusion models (Table 6.1) where Year 2 had higher RV coefficients between the individual data sets and the MFA model. Although not shown here, when partitioning at F_k>optimal (*i.e.* approximately two members per cluster), there was random assignment of memberships supporting the previously reported lack of a representative centroid sample. By varying the number of clusters, the analysis looks at how well the samples can be "pulled apart" and still be reliably closely associated. By varying the partition coefficient (F_k) the analysis looks at how wide the bandwidth around each centroid (using random assignment/partitioning) can be while retaining reliable clustering. The bandwidth would be comparable to borders in typicality assignment in sensory typicality experiments (Ballester *et al.*, 2013). Although the discrimination between samples and hence the clustering was improved by using the fuzzy algorithms, it carried little contextual applicability in this case. This strategy could, however, be applied to cases such as the different styles of South African Chenin Blanc to improve on the previous use of parametric statistical techniques (Lawrence, 2012; Buica *et al.*, 2017; Valente *et al.*, 2018).

6.4 Conclusion

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

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

References

Abdi, H. & Valentin, D., 2007. Multiple Factor Analysis (MFA). In: Encycl. Meas. Stat. Sage.

Alañón, M., Pérez-Coello, M. & Marina, M., 2015. Wine science in the metabolomics era. Trends Anal. Chem. 74 1–20.

Amargianitaki, M. & Spyros, A., 2017. NMR-based metabolomics in wine quality control and authentication. Chem. Biol. Technol. Agric. 4(1), 1–12.

Ballester, J., Mihnea, M., Peyron, D. & Valentin, D., 2013. Exploring minerality of Burgundy Chardonnay wines: A sensory approach with wine experts and trained panellists. Aust. J. Grape Wine Res. 19(2), 140–152.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. (First ed.). Vol. 66. Springer US.

Biancolillo, A., Boqué, R., Cocchi, M. & Marini, F., 2019. Data Fusion Strategies in Food Analysis. In: Data Handl. Sci. Technol. Vol. 31. Elsevier Ltd 271–310.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L. & Busto, O., 2015. Data fusion methodologies for food and beverage authentication and quality assessment - A review. Anal. Chim. Acta 891 1–14.

Buica, A., Brand, J., Wilson, C. & Stander, M., 2017. Evaluating South African Chenin Blanc wine styles using an LC-MS screening method. Stud. Univ. Babes-Bolyai Chem. 62(2Tom1), 113–123.

Cariou, V. & Qannari, E.M., 2018. Statistical treatment of free sorting data by means of correspondence and cluster analyses.

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

Cayuela, J.A., Puertas, B. & Cantos-Villar, E., 2017. Assessing wine sensory attributes using Vis/NIR. Eur. Food Res. Technol. 243 941–953.

Cocchi, M., 2019. Data fusion methodology and applications. Vol. 31.

Engel, J., Gerretzen, J., Szyman´ska, E., Szyman´ska, S., Jansen, J.J., Downey, G., Blanchet, L. & Buydens, M.C., 2013. Breaking with trends in pre-processing? Trends Anal. Chem. 50 96–106.

Figueiredo-González, M., Martínez-Carballo, E., Cancho-Grande, B., Santiago, J.L., Martínez, M.C. & Simal-Gándara, J., 2012. Pattern recognition of three Vitis vinifera L. red grapes varieties based on anthocyanin and flavonol profiles, with correlations between their biosynthesis pathways. Food Chem.

130(1), 9–19.

Garrido-Bañuelos, G., Buica, A., De Villiers, A. & Du Toit, W.J., 2019. Impact of Time , Oxygen and Different Anthocyanin to Tannin Ratios on the Precipitate and Extract Composition Using Liquid Chromatography-High Resolution Mass Spectrometry. South African J. Enol. Vitic. 40(1),.

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

Ghasemi, J.B., Heidari, Z. & Jabbari, A., 2013. Toward a continuous wavelet transform-based search method for feature selection for classification of spectroscopic data. Chemom. Intell. Lab. Syst. 127 185–194.

Godelmann, R., Fang, F., Humpfer, E., Schü Tz, B., Bansbach, M., Schä, H. & Spraul, M., 2013. Targeted and Nontargeted Wine Analysis by 1 H NMR Spectroscopy Combined with Multivariate Statistical Analysis. Differentiation of Important Parameters: Grape Variety, Geographical Origin, Year of Vintage. J. Agric. Food Chem 13 24.

Granato, D., de Araújo Calado, V.M. & Jarvis, B., 2014. Observations on the use of statistical methods in Food Science and Technology. Food Res. Int. 55 137–149.

Härdle, W.K. & Simar, L., 2015. Applied multivariate statistical analysis, fourth edition.

Khang, T.D., Vuong, N.D., Tran, M.K. & Fowler, M., 2020. Fuzzy C-means clustering algorithm with multiple fuzzification coefficients. Algorithms 13(13),.

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

López-Rituerto, E., Savorani, F., Avenoza, A., Busto, J.H., Peregrina, J.M. & Engelsen, S.B., 2012. Investigations of la Rioja terroir for wine production using 1H NMR metabolomics. J. Agric. Food Chem. 60(13), 3452–3461.

Mafata, M., Brand, J., Panzeri, V. & Buica, A., 2020. Investigating the Concept of South African Old Vine Chenin Blanc Investigating the Concept of South African Old Vine Chenin Blanc. South African J. Enol. Vitic. 41(2), 168–182.

Mascellani, A., Hoca, G., Babisz, M., Krska, P., Kloucek, P. & Havlik, J., 2021. H NMR chemometric models for classification of Czech wine type and variety. FOOD Chem. 339.

McKillup, S., 2005. Statistics explained: An introductory guide for life scientists. Cambridge University Press.

McKillup, S., 2012. Statistics explained : an introductory guide for life scientists. (2nd ed.). Cambridge University Press.

Myhre, J.N., Mikalsen, K.Ø., Løkse, S. & Jenssen, R., 2018. Robust clustering using a kNN mode seeking ensemble R. Pattern Recognit. 76 491–505.

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

Pereira, A.C., Reis, M.S., Saraiva, P.M. & Marques, J.C., 2010. Analysis and assessment of Madeira wine ageing over an extended time period through GC–MS and chemometric analysis. Anal. Chim. Acta 660(1–2), 8–21.

Radovanovic, A., Jovancicevic, B., Arsic, B., Radovanovic, B. & Bukarica, L.G., 2016. Application of non-supervised pattern recognition techniques to classify Cabernet Sauvignon wines from the Balkan region based on individual phenolic compounds. J. Food Compos. Anal. 49 42–48.

Rinnan, Å., Berg, F. van den & Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28(10), 1201–1222.

Salkind. J. & Kristin. R., 2007. Encyclopidia of Measurement and Statistics. Sage.

Seisonen, S., Vene, K. & Koppel, K., 2016. The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. Food Chem. 210 530–540.

Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Li Vigni, M., Marchetti, A. & Cocchi, M., 2014. A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. Chemom. Intell. Lab. Syst. 137 181–189.

Stevenson, T., 2005. The-New-Sothebys-Wine-Encyclopedia. (Fourth ed.). Dorling Kindersley Limited.

Valente, C.C., Bauer, F.F., Venter, F., Watson, B. & Nieuwoudt, H.H., 2018. Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns. Sci. Rep. 8(1),.

Valentin, D., Chollet, S., Lelièvre, M. & Abdi, H., 2012. Quick and dirty but still pretty good: a review of new descriptive methods in food science. Int. J. Food Sci. Technol. 47(8), 1563–1578.

Vannier, A., Brun, O.X. & Feinberg, M.H., 1999. Application of sensory analysis to champagne wine characterisation and discrimination. Food Qual. Prefer. 10 101–107.

Versari, A., Laurie, V.F., Ricci, A., Laghi, L. & Parpinello, G.P., 2014. Progress in authentication, typification and traceability of grapes and wines by chemometric approaches. Food Res. Int. 60 2–18.

Chapter 6
Supplementary

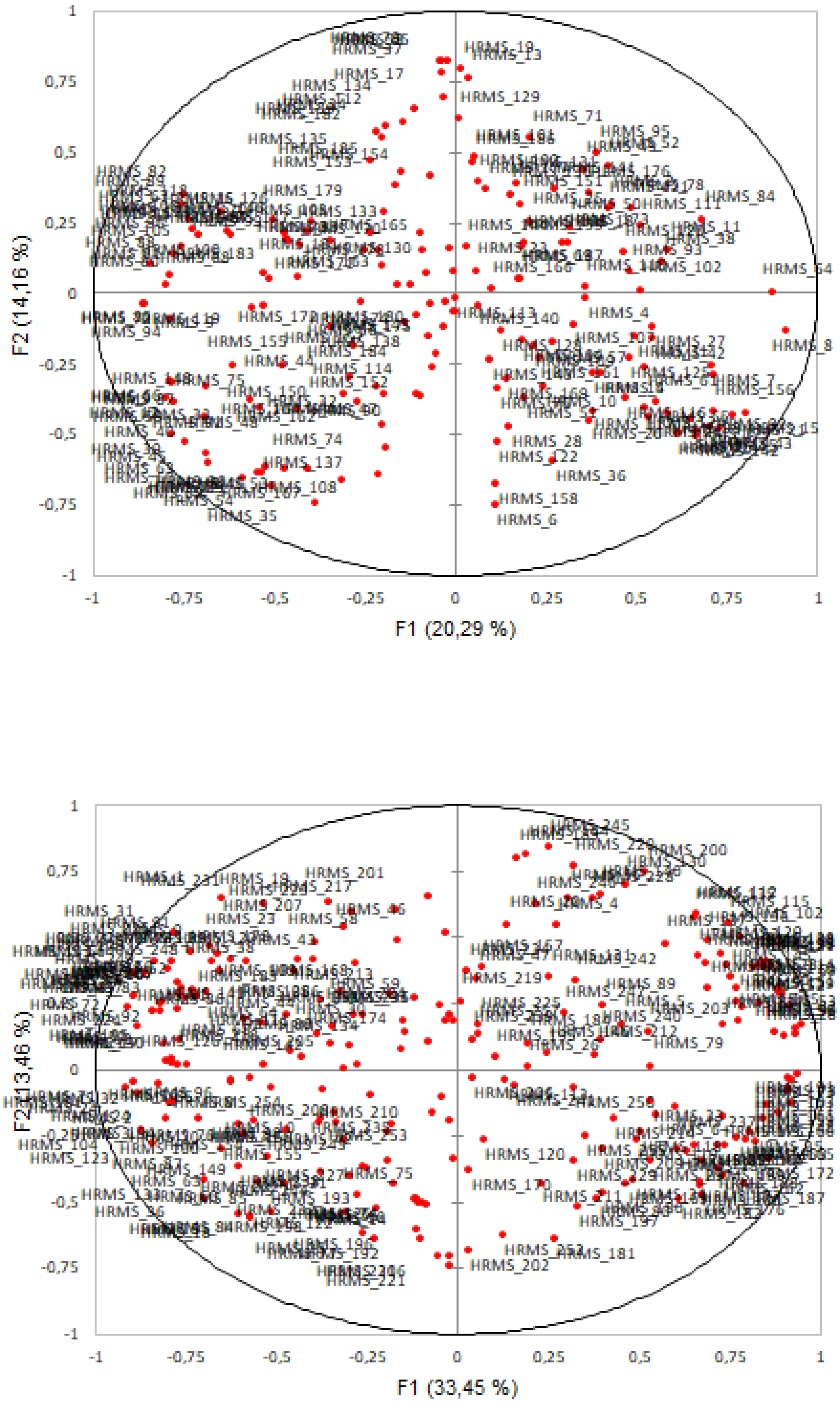


Figure 6.1: PCA loadings of HRMS in combined (Pos. and Neg.) acquisition modes for year 1 (top) and year 2 (bottom).

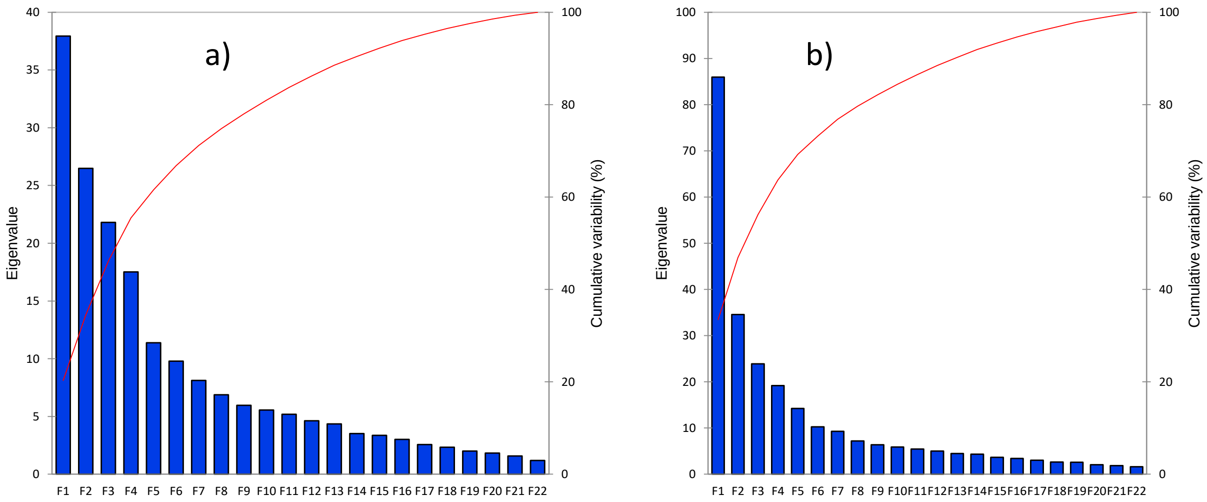


Figure 6.2: Scree plot for the PCA of the combined HRMS (Positive and negative) of Year 1 (a) and Year 2 (b) chemical data.

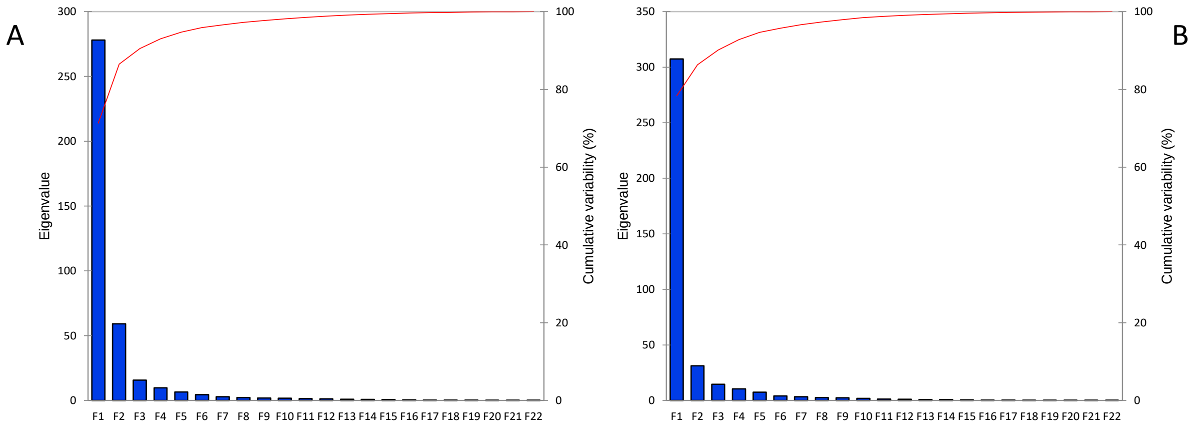


Figure 6.3: Scree plot of the PCA for Year 1 (A) and Year 2 (B) of NMR analysis.

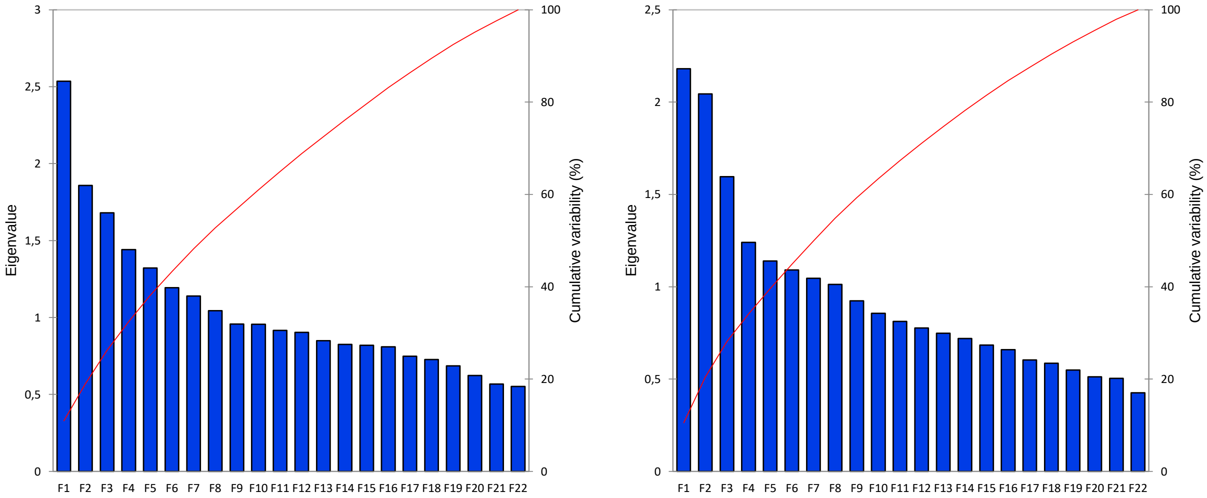


Figure 6.4: Scree plot for the MFA data fusion of Year 1 (left) and Year 2 (right)

Table 6.1: Pairwise RV coefficients between the three NMR regions (Arkyl, Aromatic and carbohydrate) for Year 1 and Year 2 data exploration.

Year 1					Year 2				
	Alkyl	Carbs	Aromatics	MFA		Alkyl	Carbs	Aromatics	MFA
Alkyl	1	0.716	0.554	0.852	Alkyl	1.000	0.813	0.807	0.924
Carbs	0.716	1	0.769	0.943	Carbs	0.813	1.000	0.914	0.960
Aromatics	0.554	0.769	1	0.864	Aromatics	0.807	0.914	1.000	0.956
MFA	0.852	0.943	0.864	1	MFA	0.924	0.960	0.956	1.000

Table 6.2: Fuzzy k-means clustering at F_k set at 1.001 and the number samples at 22 for Year and 21 fro Year 2

Year 1					Year 2				
Number of clusters	Criterion	Between-classes	Within-class variance	Wilks' Lambda test	Number of clusters	Criterion	Between-classes	Within-class variance	Wilks' Lambda test
1	1.000	0.000	487.658	1.000	1	1.000	0.000	399.889	1.000
2	0.996	218.181	269.478	0.553	2	0.995	177.728	222.160	0.556
3	0.988	284.711	202.947	0.416	3	0.986	234.489	165.399	0.414
4	0.976	316.788	170.871	0.350	4	0.975	260.357	139.532	0.349
5	0.961	336.322	151.337	0.310	5	0.957	274.137	125.751	0.314
6	0.945	352.778	134.880	0.277	6	0.943	289.119	110.769	0.277
7	0.932	369.635	118.023	0.242	7	0.925	300.846	99.043	0.248
8	0.922	385.240	102.418	0.210	8	0.916	315.501	84.387	0.211
9	0.923	402.286	85.372	**0.175**	9	0.924	329.859	70.030	0.175
10	0.944	416.330	71.329	0.146	10	0.961	340.511	59.378	0.148
11	**0.998**	423.323	64.335	**0.132**	11	0.998	362.560	37.329	0.093
12	**0.998**	446.013	41.646	**0.085**	12	0.998	365.707	34.181	0.085

Table 6.3: Variance decomposition for the optimal classification for Year and Year 2 clustering using parametric (Agglomerative hierarchical clustering - AHC) and non-parametric (Fuzzy k-means) methods.

	Year 1			Year 2		
		Absolute	Percent		Absolute	Percent
AHC with all samples	Within-class	22.308	90.95%	Within-class	19.96	88.57%
	Between-classes	2.219	9.05%	Between-classes	2.575	11.43%
	Total	24.528	100.00%	Total	22.535	100.00%
		Absolute	Percent		Absolute	Percent
AHC without outliers	Within-class	22.308	96.07%	Within-class	19.051	95.28%
	Between-classes	0.913	3.93%	Between-classes	0.944	4.72%
	Total	23.222	100.00%	Total	19.994	100.00%
		Absolute	Percent		Absolute	Percent
Fuzzy k-means clustering	Within-class	1.835	15.93%	Within-class	10.084	87.30%
	Between-classes	9.688	84.07%	Between-classes	1.466	12.70%
	Total	11.524	100.00%	Total	11.55	100.00%

Chapter 7

General discussion and conclusions

Chapter 7: General discussion and conclusions

General discussion and conclusions

Oenology is the study of the behaviour of wine, from the processing of grapes to the enjoyment of the final wine (Stevenson, 2005). Thus, Oenology is multidisciplinary, composed of different disciplines including winemaking/cellar technology and microbiology (creation of the wine), chemistry (chemical composition of wine), and sensory (sensorial composition and human perceptions/enjoyment of the wine). Studying the behaviour of wine requires a variety of chemical and sensory techniques (Stevenson, 2005). By applying a broad spectrum of these techniques, different measurements can be taken, generating large amounts of data. Making sense of all the data can be difficult, and this is where statistics are needed. In terms of nomenclature, applied statistics is referred to as: bioinformatics, chemometrics, or sensometrics (Hunter, Dijksterhuis, Qannari, *et al.*, 1995; Kowalski, 1980; McKillup, 2012; Sohail & Arif, 2019). Since there are already a host of disciplines involved in Oenology as well as various techniques, incorporating statistics is complicated and is often done collaboratively with statisticians/bioinformaticians. This can create a gap between experimental data acquisition and statistical data results, sometimes considered as a "black-box"(Cortez & Embrechts, 2011). It is important to remove uncertainties surrounding statistics in Oenology and unpack this "black-box". Addressing the gap requires integrating the different disciplines through transdisciplinary approaches in Oenology. However, a limitation to this is that being specialized in any of the three disciplines is difficult, and true transdisciplinarity is even harder since it would require integrated contextual and technical knowledge. The first step to achieving true transdisciplinary ability is to understand the sequence of stages for problem solving in each discipline and consolidate them.

The aim of this dissertation was to elucidate critical steps in data handling while highlighting some common misconceptions and misinterpretations, and to demonstrate the value of comprehensive narratives of the process of data analysis in Oenology. This compilation was a journey through different stages of dealing with oenological data, with increasing complexity in both the strategies and the techniques (sensory, chemistry, and statistics). The work devised several systematic approaches for solving complex oenological problems, which focused on creating strategies, rather than finding a particular statistical solution.

Overall, the statistics-focused work identified the key decision-making aspects during the data input (capturing and pre-processing) and the model output (visualisation and interpretation) stages. The pre-processing of the data was shown to affect the performance of models as measured by performance parameters such as the explained variance (%EV) and calibration coefficients (Engel, Gerretzen, Szyman´ska, *et al.*, 2013). Pre-processing was investigated for infrared data, since many options are available (Rinnan, Berg & Engelsen, 2009), but this was unnecessary for an unsupervised approach. The need for pre-processing was investigated again for nuclear magnetic resonance (NMR) data, where blocking of different NMR regions and statistical treatment by MFA was used. Owing to the reiterative nature of the data handling process, model optimization techniques such as variable/feature selection or exclusion were addressed, such as the exclusion of outlier samples, which resulted in better clustering (*i.e*. increased goodness-of-fit criterion and lower within-class variation).

Since this work exclusively used unsupervised techniques, models and their optimization were based on their compatibility with the strategies employed rather than trying to solve a single question. This was a novel approach for Oenology, since – to date – most studies use supervised techniques. Taking an unsupervised approach is open-ended and requires an open mind for the outcome and interpretation, creating more opportunities for hypothesis formation. On the other hand, an unsupervised approach also comes with the need to understand not only the limits of the statistical techniques, but also the context of the data generated.

Interpretation of model outcomes (statistical) and study outcomes (contextual) should never be compromised by misleading visual aids, which can perpetuate confirmation bias. Hence, this work discussed the impactful nature of visual aids and offered a rationale as to how to couple them with each other and with performance parameters. For example, this work used comprehensive descriptions of multivariate model distributions (model dimensionality, stress distribution and rate of decay, and inflection points) rather than the more standard description of only the first two dimensions in multivariate data model outputs. This work also habitually coupled model Cartesian plots with dendrograms from cluster analysis to avoid biased visualised perceptions of grouping.

Heatmaps were proposed as a novel approach to visualizing Pivot©Profile data. Pivot©Profile data is commonly analysed using correspondence analysis (CA) after translating the raw data into frequency data (Thuillier, Valentin, Marchal, *et al.*, 2015). This work showed that, by keeping the data as positive and negative ratings and understanding the different types of statistical analysis available, the data could be represented by heatmaps instead of scores and loadings plots from the CA. This was most appropriate since Pivot©Profile is a reference-based method and heatmaps show the relative change of attributes across samples, both negative and positive. By committing to the strategy of differentiating samples from the reference (pivot), an appropriate statistical method could be applied.

‘Double-checks’ were also used throughout by evaluating multiple performance parameters – for example, comparing the goodness-of-fit criterion (%EV) with coefficients of fit such as R^2 and Wilks’ λ and evaluation parameters such as regression vector (RV) and cophenetic correlation coefficients, in order to enhance the interpretability of model outcomes. This meant weighing the relative importance of every parameter used against the strategy and the research question. This was important since, without an appropriate strategy, it would be difficult to have a marker for optimal model outcomes for the unsupervised approaches used, in stark contrast to supervised models.

The main limitation to unsupervised approaches was the sacrifice of optimal coefficients in favour of optimal goodness-of-fit. This was the most appropriate approach since it is best to have confidence in the objective function for addressing the hypothesis rather than to optimize coefficients (*e.g*. discrimination, classification) for ill-fitted data. For example, it would be analogous to a futile attempt to optimize the p-value (error in calculation) for a poor R^2 (coefficient of goodness-of-fit) with little contextual meaning.

This work also discussed why the contextual significance can be more important than the statistical significance of the model outcomes. Contextual significance is not just based on the absolute values but their relative importance and meaning. This was important since the work indirectly addressed an important Oenological problem of combining chemical data with sensory data (Seisonen, Vene & Koppel, 2016). Methods of data fusion combine ***and*** integrate data sets to create comprehensive and representative models where appropriate. Data fusion methods address the scaling issue by making it relative, *i.e.* in absolute values sensory data is lower dimensionality than chemistry data, similarly, targeted chemistry data is lower in dimensionality than untargeted chemistry data. This is a limitation since many statistics/omics issues can be overcome by having a larger data set, but this is not always possible, especially in Sensory.

Another novelty in this work comprised of the first published study on the systematic evaluation of the South African old vine Chenin Blanc typicality concept. Although the concept was not confirmed, the study found success in the use of sensory and chemistry strategic approaches. The limitations were a lack of perceptual consensus on the typicality of old vine Chenin Blanc wines, possibly due to the wines included in the evaluation as the experimental wines were all made in a standardized manner. In future, the issue could be overcome by using these now-developed strategies on commercial wines.

Looking at the journey, through developing descriptive narratives for data handling process of the oenological problems and the statistical investigations, it was shown that there is no “black-box” but perhaps a gap in critical thinking and full engagement with the data handling. One needs to keep the ‘long game’ in mind. From the research question, design of experiments, data acquisition, statistical strategy, to the interpretation of models in context. This is especially true when using unsupervised approaches and/or in view of data fusion as the treatment is not the same, but dependent on the intent of the strategy.

For this reason, this study recommends the following:

○ Creating a descriptive narrative of the data handling process (from input to the interpretation of results) in order to minimize misinterpretations of data modelling and its results

○ Intertwining statistical and applied contextual reasoning for interpretation of modelling outcomes

○ Experimenting with the use of Artificial Intelligence/Machine Learning in Oenology, since many user-friendly and accessible software now offer the opportunities to experiment with these advanced techniques.

References

Cortez, P. & Embrechts, M.J. 2011. Opening black box Data Mining models using Sensitivity Analysis. in *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining* IEEE. 341–348.

Engel, J., Gerretzen, J., Szyman´ska, E., Szyman´ska, S., Jansen, J.J., Downey, G., Blanchet, L. & Buydens, M.C. 2013. Breaking with trends in pre-processing? *Trends in Analytical chemistry*. 50:96–106.

Hunter, E.A., Dijksterhuis, G.B., Qannari, E.M. & Macfie, H.J.H. 1995. Second Sensometrics Meeting-Edinburgh, 16-18 September 1994: introduction on behalf of the organising committee. *Food Quality and Preference*. 6:215–216.

Kowalski, B.R. 1980. Chemometrics. *Analytical Chemistry*. 52(5):112–122. [Online], Available: https://pubs.acs.org/sharingguidelines.

McKillup, S. 2012. *Statistics explained : an introductory guide for life scientists*. 2nd ed. Cambridge University Press.

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

Seisonen, S., Vene, K. & Koppel, K. 2016. The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data. *Food Chemistry*. 210:530–540.

Sohail, A. & Arif, F. 2019. Supervised and unsupervised algorithms for bioinformatics and data science.

Stevenson, T. 2005. *The-New-Sothebys-Wine-Encyclopedia*. Fourth ed. Dorling Kindersley Limited.

Thuillier, B., Valentin, D., Marchal, R. & Dacremont, C. 2015.