CHAPTER 5

Strategies for Accurate Food Data Mining and Optimizing Information Generation

MPHO MAFATA, JEANNE BRAND AND ASTRID BUICA*

South African Grape and Wine Research Institute, Department of Viticulture and Oenology, Faculty of AgriSciences, Stellenbosch University, Private Bag X1, 7602 Matieland, South Africa
*Email: astrid.buica@gmail.com

5.1 Introduction

Foodstuffs generally have a complex and dynamic chemical composition. This is the reason why there are so many different chemical techniques and sensory methods for evaluating their quality, often resulting in the collection of large amounts of related data. A look at current trends in the food research literature shows a growing interest in advancing the informational value of large food evaluation datasets through the use of advanced statistical modelling tools in sensory and chemistry.[1–4] Bioinformatics, metabolomics, chemometrics and sensometrics are all forms of statistical data handling within their respective fields.[5] Hence, there is high field specificity and certain sets of rules when it comes to how statistics are applied to each of these data. Regardless, every field follows a similar workflow when it comes to handling data: collection/capturing, cleaning/pre-processing, modelling and, finally, interpretation.[5–7]

There can be a considerable gap between generating data and obtaining information, and even more between identifying a problem and then finding

Food Chemistry, Function and Analysis No. 43
Sensory Evaluation and Consumer Acceptance of New Food Products: Principles and Applications
Edited by Ana Isabel de Almeida Costa, Maria João P. Monteiro and Elsa Lamy

Published by the Royal Society of Chemistry, www.rsc.org

its solution. Strategic thinking implies that questions are formulated taking the context into account. Possible ways of answering them will consider the entire research flow, from resources (logistics, time, skills, among others) to data interpretation and dissemination; keeping the entire process ("the big picture") in mind dictates choices with implications for the final outcomes.

The planning of smart data analysis starts with a suitable study design, as only captured data can be modelled. Data modelling can be merely exploratory or include elements for the prediction, classification or discrimination of phenomena, depending on the original research question.[8] Generally speaking, exploratory techniques are used for hypothesis-forming purposes while supervised techniques are used for hypothesis testing. The choice of technique (exploratory *vs.* confirmatory) and the details of its implementation (when and how) depend on the formulated research question and the type of data required to answer it. Still, it is sometimes possible to expand the scope of statistical modelling activities from hypothesis forming to testing, and *vice versa*, if enough data have been collected or are otherwise available.

Researchers should understand that foodstuff data are generally multi-way/multi-modal and thus need to be treated as such to solve issues regarding the absolute *vs.* relative significance of the parameters and effects to be estimated. By highlighting gaps in the communication of data handling activities and pointing out critical steps, it can be shown that this is not a "black-box" process. Both theoretical and execution limitations in data handling across fields can be addressed by examining each field's process and the philosophy of analysis, rather than simply focusing on its inputs and outputs. This means creating approaches that emphasize the exploration of the problem by aligning multiple perspectives, rather than approaches which focus on perfecting the answer to a single problem or modality.

Despite the amount of literature available, there are still certain misconceptions about data handling due to the lack of articulation of this process. This includes misconceptions about how data should be handled as well as how the process and results should be communicated. In turn, this fosters a lack of confidence in handling and interpreting data in a critical manner. Such misconceptions and lack of confidence may impede the development of a new repertoire of data handling techniques and progress towards the age of artificial intelligence tools for data analysis. Multi-way problems mean evaluating the relative importance of different data sets, with particular consideration given to which data sets can and should be combined, and in what manner. Applications of data fusion methods – which not only combine but also integrate data sets – appropriately address the issue of relative importance between data sets by separately scaling them according to their variation.[7] By fostering a greater "self-awareness" in our approach to data handling processes, perhaps we can start asking the correct questions about the data and become more accustomed to both hypothesis-testing and hypothesis-forming results.

In this context, the aim of this chapter is to demonstrate the importance of a comprehensive narrative of the process of data analysis, from its generation using the most recent developments, to data fusion, and the knowledge it can provide.

5.2 Instrumental Analysis

This section deals with instrumental data acquisition, data accessibility and data treatment. State-of-the-art instruments can acquire large and sophisticated data sets, but these data sets can be difficult to read and treat due to their particularities. Instrument developers continuously work on hardware and software to improve the interface between acquisition and data treatment, but these interfaces are often limited for large and/or complex data.

In an effort to extract relevant information in a more efficient manner, the developments in instrumentation go hand in hand with data handling. Generating raw data from analyses is happening now at a higher rate than ever before. Even though not always reaching the level to be considered as big data, data obtained from analytical systems do meet some (if not all) of the related criteria: volume, velocity, variety, veracity and value (with visualization and variability sometimes added).[9] If some of the aspects are application or analysis-related, value is a criterion that can be re-defined and re-evaluated in a broader context.

Omics techniques aim to collectively characterize and measure pools of molecules that translate into biological structure, function and dynamics. They are thus focused on extracting value from (large) data sets that would otherwise be too difficult to exploit. Omics have been proven to be a powerful tool (and strategy) to study a number of food-related issues, such as quality control,[10,11] authenticity/adulteration,[10,12–15] safety[14] or traceability,[16] among others. Indeed, nowadays there are numerous highly specialised omics sub-fields, some bearing more creative names than others, *e.g.*, metabolomics, proteomics, lipidomics, volatilomics,[17] foodomics,[18–20] nutriproteomics,[14,21] wineomics,[22] and beeromics.[23] Irrespectively, the omics field can be broadly divided into targeted and untargeted techniques. In the former, compounds are first identified and measured, and the statistical analysis follows; in the latter (also called profiling or fingerprinting), raw data are first subjected to statistical analysis and the outcomes are then used to identify and measure the compounds of interest.[24]

5.2.1 Acquisition: State-of-art in Instrumentation and Method Optimization

Due to the fairly complex and variable matrix of foods, food chemical analysis often involves a large number of compounds, either due to their intrinsic interest or because they interfere in other relevant processes. Such complexity and variability can be the result of natural (*e.g.*, fresh produce characteristics) and human action (*e.g.*, fermentation, distillation, and

cooking), or, often, of both. The need remains to reliably quantify compounds and consistently characterise products. From this perspective, it is understandable that the food science and technology fields are increasingly engaging with the use of smart ways of optimizing methods, complex analytical techniques and systems, by making use of available statistical power and moving towards data integration. In such settings, the use of "information-rich" techniques has become paramount, with method development aiming at providing several, highly specialised measurements (*e.g.*, detecting compounds present in complex samples at very low concentrations) preferably within a single analysis, (*i.e.*, a "shotgun" approach).

Currently, one of the most interesting approaches in method development and optimization is the application of the analytical quality by design (AQbD) framework and principles. Initially used in pharmaceutical applications for optimization and validation (especially for testing robustness), AQbD is making its way into food analysis, even if in an adapted form.[25–28] Sometimes called "multivariate optimization"[25] or "multi-response statistical techniques"[29] in food applications, its goal remains the same: to evaluate multiple factors in a systematic manner rather than using the classical 'one at a time' approach.

AQbD is based on insights from design of experiments (DoE), the branch of applied statistics that supports the design of systematic data collection and analysis plans, in order to determine the factors controlling the value of a parameter or parameters. As such, it can be an effective way for method optimization since it works with multiple parameters and levels at a time. Its implementation entails three main steps: (1) the establishment of an analytical target profile, (2) the determination of critical method attributes or criteria, often related to selectivity, sensitivity, and accuracy, and (3) the identification of critical method parameters.[30] Care must be taken when considering the relative importance of the latter for a method; for example, the importance of the type of stationary phase *vs.* mobile phase *vs.* column temperature *vs.* flow rate in liquid chromatography. Screening experiments are often necessary to identify critical method parameters and their relative importance in the context of the application.

Admittedly, AQbD may never be as appropriate in food analysis as it is in pharmaceutical applications. When setting up an untargeted method linked to an omics strategy, or fingerprinting, factors such as the analytical target profile or critical method attributes may pose difficulties in their definition. Still, AQbD can make a substantial contribution to targeted analyses of multiple compounds, in addition to the standard tools of analytical researchers. For the fingerprinting of complex samples, the highest resolution, or separation, power is desirable. In terms of instrumental development, the highest resolution power is currently provided by hyphenated techniques (*i.e.*, separation combined with spectroscopy); however, there is a shift towards multidimensional techniques combined with spectroscopy. Despite being relatively recent, this field has experienced an incredible development, as demonstrated by the numerous applications, various instrumental setups used, and even the number of reviews published.[31–37]

Multidimensional chromatographic techniques work on the principle of orthogonality between two separations, with fractions eluting from the first separation dimension (column) being injected in to the second separation dimension (column), resulting in increased resolution. This can be carried out either off-line or on-line, for all or some fractions, and targeting either the entire sample or a specific region in the first dimension.[38] The resolution power combined with various operational modes would make comprehensive chromatography the most powerful tool for untargeted analyses; yet, very few combined instruments are commercially available for routine (high throughput) applications. Another aspect that is worth considering is the amount and type of data generated, which has further implications for data analysis.[39]

One of the strongest techniques for fingerprinting and omics, alongside mass spectrometry (or rivalling with it), is nuclear magnetic resonance spectrometry.[40,41] However, despite offering higher reproducibility, it is still not able to achieve the sensitivity of mass spectrometry techniques. This is one of the reasons why efforts are dedicated to increasing this aspect of instrumental and analysis performance, along with other parameters. Another important shortcoming of nuclear magnetic resonance spectrometry applications is their cost, especially for equipment maintenance.[42,43]

Non-invasive or non-destructive techniques are also of interest in food analysis.[42,44–46] Their strength lies in the rapid generation of raw data; coupled with powerful chemometric tools, these techniques can be used on-line or in-line to monitor processes and make decisions in real time.[45,47] Even though some of these techniques are still used off-line, their non-destructive nature means that the sample can be recovered, which is relevant for small volume experiments.

Regardless of the nature of the analytical technique employed, there is a trend towards creating "pipelines" or workflows integrating multiple analytical steps, from sample preparation to data capturing, in a single appropriate format for statistical analysis, with statistical functions sometimes being already included in the set-up. Some of these workflows are technique-based[48–51] while other are application-based, for example, for food authentication.[12,52–55]

5.2.2 Accessibility: Reading the Data

Finding the informational value requires solving issues surrounding accessibility (reading data) and data handling (modelling and making sense of data). Accessibility, or the technological ability to read the acquired data, is no small problem. The instrumentation is set up to control the analysis parameters and capture the data generated, although not always in a format compatible with easy data export. There are many formats in which large array data are saved and usually the larger the data, the more complex the format. Sometimes this creates a bottleneck between generating the raw data and transferring it to a program for further (statistical) analysis.

Software coupled to hyphenated instruments may capture the responses in independent channels and/or in a conjugated matrix. For example, in liquid chromatography coupled with mass spectrometry, the data can be extracted as a chromatogram or a matrix.[56] The chromatogram can be extracted in two modes: selected ion monitoring (ion extracted chromatogram, for multiple *m*/*z* channels, or selected ion monitoring, for a single *m*/*z* channel) or total ion current (for full scan exploitation of the mass analyser). Both are two-dimensional representations of retention time (RT) *vs.* ion abundance. The matrix is extracted as RT_mass-to-charge pair (RT_*m*/*z*) *vs.* ion abundance for each *m*/*z* channel. The software generates automated outputs that, even in the case of hyphenated techniques, can provide the user with options regarding which information to capture. The hyphenated instruments with multiple detectors are set-up in such a way that there is a single output (multi array), where the different channels are captured as a single matrix aligned across a common array/dimension, usually the retention time. Depending on whether the eluent is split, or the detectors follow each other, there can be a delay between the time axes of the detectors. An example of such an instrumental setup is of the type separation – detector 1 – detector 2, where detector 1 must not be destructive (see Figure 5.1A). Logically, there will be a constant delay between the signal in the first and second detector due to the flow rate of the eluent. On the other hand, if the detectors are in parallel, there should be no time delay between the signals (see Figure 5.1B).

If a "regular" hyphenated method can generate 3D data (*e.g.*, retention time *vs.* ion abundance *vs. m*/*z* for one dimension chromatography coupled with mass spectrometry), multidimensional chromatography coupled with mass spectrometry will result in a 3D sample map (retention time dimension 1 *vs.* retention time dimension 2 *vs.* ion abundance) plus a fourth dimension, which is the spectra at each elution point. These data have to be

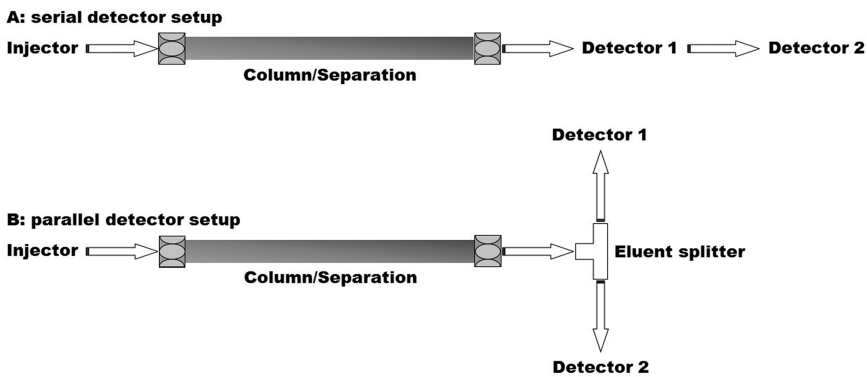

Figure 5.1 Possible instrumental configurations allowing signal generation with detectors (A) in series, resulting in a time lag between the two detectors' signals; and (B) in parallel, without time lag. The arrows indicate the direction of the flow and interfaces between the modules.

pre-processed, aligned, and corrected before proceeding to peak detection and profiling.[39]

There are some common problems related to reading multi-array data. Some of these issues can be handled only before data acquisition and some can be dealt with afterwards. Important pre-acquisition problems include the addition of internal standards (off-line or on-line) and baseline pre-settings. The former aid in intensity regulation, while the latter help correct spectral or chromatographic drifts. Internal standards cannot be corrected (added) post-acquisition, but some baseline shifts can be corrected post-acquisition. When it comes to single acquisition, these issues may not always be of concern, but when trying to consolidate two data sets or multiple samples, irregularities in baseline can compromise the data treatment and further statistical modelling.

Software solutions are available for post-acquisition online baseline drift correction for analyses such as nuclear magnetic resonance and infrared spectroscopy. Although software for processing spectral and chromatographic shifts is also available, it is usually coupled online with instrumental acquisition and the parameters are set before acquisition.

Software developers work with instrumental developers, and field analysts, to find the best systems to read and solve some problems related to data acquisition.[57] For example, untargeted gas chromatography–mass spectrometry data have particularities that are continuously being investigated. A suite of software tools (PARAllel factor deconvolution and identification system – PARADise) has been developed to meet the multiple needs of researchers when it comes to reading the data, correcting irregularities and modelling the data within a single interface.[58,59] Since this is such a sophisticated problem, some studies have further explored streamlining gas spectrometry data processing using machine learning techniques.[60] Accessing this software requires using programming languages, such as Matlab, Python or R, which creates another bottleneck in the skills needed to work with such complex data sets.

Similarly, dealing with chromatographic RT shifts in large sample sets in untargeted high-performance liquid chromatography analyses is difficult. Scripps XCMS and MZmine 2 are automated online software packages available for reading these types for data that do not require programming skills.[61–63] They were created to access chromatographic–mass spectroscopic data. Since the software is automated, the data handling cannot be tailored to optimize resulting models. This is the major motivation for most researchers to rely on programming.

5.2.3 Treatment: Towards Finding the Informational Value

Once accessibility issues are solved, appropriate strategies for treating the data are necessary. The previously mentioned software packages (Scripps XCMS and MZmine 2) can process the data as integrated peaks or as continuous interval data, which has implications for the modelling

approach.[61–63] For peak selection (or "peak picking"), a RT or spectral window can be chosen.[64–66] Irregularities cannot be overcome when using continuous intervals ("binning"), where the interval intensity rather than the peak area is measured. The disadvantage of choosing peak integration is in finding the optimal RT or spectral window when there is a big overlap between peaks. Thus, it is easier to automate continuous data capturing compared to peak picking. Furthermore, continuous interval/binning leads to large amounts of data, expanding the range of modelling techniques that can be performed.

Deciding whether to process the data as continuous intervals or peak picking should be based on the purpose: testing or exploration. Often peak picking is used for hypothesis testing, to identify unknown peaks and further annotate them using databases such as the National Institute of Standards and Technology (NIST). Conversely, continuous-interval processing is used to find the unique features that create discrimination, which may not be correlated to a certain compound, but rather to a group or combination of signals. Similarly, the data modelling of these types of data has to be compatible with the processing method used. Peak picking is more supervised, whereas the continuous interval treatment is more unsupervised. Thus, one has to think from the beginning of the experiment what is needed from the data. Even if the purpose is to do data mining, devising a more data-dense processing method such as continuous-interval can be more advantageous.

5.3 Sensory Analysis

Analytical sensory analysis is distinguishable from sensory consumer science as it requires different levels of precision and accuracy. The concept of precision in sensory analysis is associated with repeatability and consistency.[67] Repeatability can be measured temporally from the perspective of both the panel and the samples. Accuracy is assessed through panel training, using sensory standards wherever possible. Since there are different types of panels (consumer, trained, and industry/experts) and methods, there are different expected levels of precision and accuracy for each.

Appropriate methods or their combinations have to be carefully chosen to answer the research question. Thus, a strategic approach is better, since *a priori* decisions are generally more important for sensory analysis than for chemistry. This is due to the nature of the information collected (based on the specific method), the way the raw data are captured, the limitations on the further manipulation of the data, among others.

5.3.1 Rapid Methods as Alternatives to Classic Descriptive Analysis

Several rapid methods have been developed to address needs such as product discrimination, similarity and different ways of product profiling.[67,68]

The methods are generally classified as verbal-based, similarity-based or reference-based (as the main task), according to the psychological processes the judges go through while evaluating products.[67]

The use of rapid sensory methods as alternatives to classic descriptive analysis for benchmarking and fit-for-purpose applications is currently more prevalent than ever. In terms of benchmarking, the statistical performance (*i.e.*, precision and accuracy) of rapid methods can be compared to that of classic approaches, but the compatibility of the methods with the research question is more contextually important. For sensory scientists, the importance of rapid methods lies in their fit-for-purpose nature, in addition to the original motivation for creating them, which was to minimize the time and costs associated to the application of classic techniques. Another aspect, which is seldom mentioned in association with the latter, is the relatively small number of attributes assessed. When complex products are evaluated based on a limited set of attributes, a loss of information can occur compared to methods where a larger number of attributes can be evaluated, such as some rapid techniques.[69]

The sensitivity of a method is related to the instrument (panel) used and the type of data generated (continuous or discreet). Expert, industry, and consumer panels are less sensitive to minor product differences compared to trained ("analytical") panels. This has been a continuous point of concern when benchmarking rapid methods. Another important contributor to perceived sensitivity (as measured statistically) is the type of data collected. Methods that collect categorical data, such as sorting, projective mapping or Pivot©Profile, generally result in lower statistical performance (*e.g.*, explained variance), compared to those that collect intensity-based data, like descriptive analysis or rate-all-that-apply, for instance.

The performance of the methods varies according to the purpose of the method and the product space. For example, the accuracy of reference-based methods such as Pivot©Profile can vary when method parameters (*e.g.*, the reference/pivot) are changed.[70] The ability to vary experimental conditions is an indication of the analytical nature introduced to sensory science through the benchmarking of these rapid methods.

It has been shown that descriptive analysis is more sensitive for finding small differences between products, while rapid methods are more appropriate for solving categorical, discrimination or comparative problems.[71] It is thus important to take strategic approaches to sensory investigations, especially when choosing a method.

The versatility of rapid methods lies in the possibility of coupling them to address different needs for a research question. For example, projective mapping is frequently coupled with ultra flash profiling to develop an extensive product space verbally and non-verbally.[72] The reason for using such mixed method approaches is comparable to the motivations behind hyphenated instrumental techniques: they are used for solving complex problems. Mixed methods result in increased data generation and enhance the potential informational value. Since mixed methods consist of multiple

tasks, there are implications on method performance, given the different priorities and order of tasks. Again, due to their analytical nature, method parameters can sometimes be adjusted to increase performance. For example, for complex problems related to wine typicality, a mixed methodology is recommended.[73] In this case, the methodology prioritizes intrinsic tasks of typicality rating and sorting, over verbal tasks. This is because the methodology prioritizes discrimination, emphasizing the representation of typicality concepts, and puts discriminating profiles second. In this case, the primary task answers the research question while the secondary task annotates the problem space.

Still, a barrier to using mixed methods is the number of samples that can be evaluated in one flight or session, due to panel fatigue. A common approach to address is to conduct multiple sessions with different methods.

Most sensory methods can be considered static: the panellists have to rate either peak (highest) intensity or an average of a perceived attribute. However, the headspace or mouthfeel of a complex product can be dynamic. In this context, temporal methods are being revisited, also partly due to the availability of software that can overcome issues related to speed of data capturing and extrapolation of attribute trajectories with time. In addition to time intensity methods, examples are temporal dominance of sensations, multiple attribute time intensity and temporal check-all-that-apply. Unlike the first three, the fourth is frequency-based, where assessors can choose attributes varying from none to multiple times during an evaluation.[74]

Both the category of the method (*e.g.*, verbal, similarity, and mixed) and the manner of evaluation (*e.g.*, aroma, taste, and global) have to be considered when choosing a method. Since rapid methods generally use untrained panellists, the researcher has to consider both extrinsic and intrinsic aspects of the tasks. Taking these aspects into account helps avoid panel fatigue which can result in decreased sensitivity.

5.3.2 Alternative Sources of Data

Primary data collected in the laboratory (*i.e.*, during sit-down product evaluations, performed in a strictly controlled environment according to a pre-defined protocol) is not the only source of data available for the sensory characterisation of a product. Big data approaches and AI techniques are increasingly used to integrate secondary data (*i.e.*, the wealth of information generated by past studies and their evaluations) with data resulting from the evaluations of critics, producers and other experts; these approaches are particularly suitable for dealing with the volume and characteristics of the information generated.[75] Still, some sources of information are more credible than others; with this in mind, when approaching the available data, caution should be exercised.

Meta-analysis studies are more common in the fields of science where the topic and the protocols are subject to strict standardisation (for example, medicine or pharma). Meanwhile, systematic and critical reviews of available

peer-reviewed literature are more common in the field of sensory science. Their topics can vary from the application of a method or methodology to the use of statistical approaches,[76–78] and to the evaluation of a certain product using multiple methods and/or types of panels,[79–82] or even across disciplines.[83–85]

Other possible sources of data are technical notes, tasting notes, critics' reviews, scores, competition results, among others. Despite being freely available (or maybe because of that), these data should be treated with caution, as, unlike specialist literature, this type of information is, in most cases, written for the benefit of end consumers. Still, this source of information can be mined using appropriate strategies, with the help of designer software.[1] One important issue in this context is the lack of a standardised lexicon to create product descriptions, which can be overcome by using sensory wheels. These are structured collections of perceptual descriptors (typically aroma and flavour attributes) establishing a clear and common sensory vocabulary to describe a product or category of products, with the aim of standardizing training for evaluation and facilitating communication between trained panels, manufacturers, distributers and consumers.[86] Nonetheless, the elimination of non-standardised vocabulary poses the risk of eliminating descriptors associated with unique product characteristics. By applying consolidation rules based on standardised lexicons (for example, from sensory wheels), such unique descriptors may be eliminated. The sensory space will then reflect only common aspects and not outstanding ones, which is not a true reflection of the variety of products evaluated. In this sense, researchers must be careful not to replace data quality with data quantity.

Product competitions can be an interesting source of information, as they reflect the trends in a product's appreciation by professionals, which in turn can influence consumers and future trends. Most of the competitions do not require judges to deliver product profiles; in such instances, a re-evaluation of the products evaluated can take place outside the competition. In the case of wine for instance, some studies already presented methodologies for such re-evaluations, crossed-checked results with those obtained from the competition itself, and looked into ways of determining the drivers for the quality scores obtained.[87–89]

5.4 Data Analysis

The drive of advancements in both instrumental analysis and sensory evaluation aims to increase the informational value, consequently resulting in large data sets. Likewise, the motivation behind data fusion is to increase the informational value by combining and integrating several data sets from different sources. This is not always a simple task; often, some data pre-processing has to be performed to generate the necessary compatibility between data sets and/or data matrices and the desired data modelling technique. To ensure this, there needs to be alignment along at least one

dimension, usually the observations (samples). Once this is assured, pre-processing methods can be applied to standardize data.

5.4.1 Data Exploration, Clean-up and Modelling

Different data sets have different particularities that require certain treatments before modelling. This "clean-up" is carried out to maximize the informational value from a statistical point of view, but it should not lose the context of the application. Data cleaning is a reiterative cycle of pre-processing, modelling, exploration.

Pre-processing is needed because different data sets have different distributions and scales, and hence they cannot always be simply modelled or combined further. When data sets with different distributions (variable scales and distribution) are modelled together in a simple concatenation, there is a risk that the results will be skewed in such a way that this will result in a false representation of the correlations between variables and/or observations. Hence, principles of data fusion must first be applied to properly integrate data sets.

Data cleaning from the perspective of instrumental analysis has already been partially addressed in this chapter. However, from a data handling perspective, pre-processing has a different motivation. Continuous instrumental data have irregularities related to baseline drift and spectral scatter. For instance, irregularities in infrared data, a popular fingerprinting technique in food science, are corrected post-acquisition using mathematical transformations.[90] Spectral pre-processing of ultraviolet and visible light spectrometry data, for instance, is automated in some software packages, such as SIMCA.[91]

Meanwhile, intensity-based sensory methods, like descriptive analysis or rate-all-that-apply methods, do not require data cleaning and consolidation since the attributes chosen are selected by trained panels or sensory screening, respectively.[92,93] On the other hand, reference-based methods, such as sorting and Pivot©Profile, do require manual cleaning (consolidation) to reduce panel noise and enhance model performance. Since sensory data cannot always be captured automatically, it is important to maintain separate records of different elements (panels, sessions, flights, judges, samples, attributes and replicates) and the different aspects (verbal and non-verbal) of the analysis when capturing and consolidating data. This helps to keep a complete and accurate record of raw data.

Sensory data can be pre-processed using linguistic and semantic consolidation as well as statistical data cleaning. Semantic consolidation of sensory data is guided by sensory wheels for different products such as whiskey,[94] beer,[95] olive oil,[96] rooibos tea,[97] and different types of wine,[98] among others. However, the best practice often involves a bottom–up approach, allowing the generated attributes to guide the consolidation and thus retaining descriptors that represent unique product characteristics.

Once data are cleaned, modelling is done based on the compatibility of the matrix (discreet, continuous or frequency-based) with the technique as well as the intent of the analysis (unsupervised or supervised). The choice of modelling approach is equally dependent on the research question and experimental design. The most popular exploratory technique is principal component analysis, used for intensity-based data gathered in both discrete and continuous data sets. This technique can also be used in the modelling of intensity-based sensory data. But the diversity of data generated in sensory science is typically modelled in several different ways, depending on matrix compatibility. Frequently, categorical and frequency data are modelled through correspondence analysis, similarity data through multidimensional scaling and ordinal data through multiple factor analysis.[99] These exploratory techniques are also used for reducing the dimensionality of the data and scaling it before data fusion.

5.4.2 Data Fusion Strategies

Data fusion models are necessary when aiming to combine and integrate data from different sources.[100] The goal is to merge the data sets in such a way that the resulting model is information dense, representative and robust – these last two concepts are linked to model performance. Fusion can be relatively simple when data originate from the same domain, but cross-domain data fusion (*e.g.*, chemical and sensory) is notoriously difficult. In this respect, the fusion of instrumental data sets is easier than that of sensory data. The matrices of instrumental data are often compatible, and the scale conversions of the measurements are conventional and known. Meanwhile, sensory data can be continuous, discrete, categorical, ranked, among others. Hence, sensory data matrices and measurement scales vary, with conversions usually requiring standardisation by using results from multivariate models.[77]

The theoretical framework for data fusion procedures includes three levels of complexity – low, medium or high – depending on the difficulty and number of pre-processing steps between the raw data sets and the final model.[101] The simplest form of data fusion is concatenation, where different data sets are combined into a single matrix; a common technique for simple data fusion by concatenation is principal component analysis. Mid-level data fusion is distinguished by rigorous data pre-processing, use of multiple data sets of different modes and model optimization. Finally, high-level data fusion is distinguished by the use of machine learning techniques. In practice, differentiating between such theoretical frameworks is not always easy; 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, is not always communicated. This can create misconceptions about the level of difficulty in fusing different data sets, 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 techniques or strategies.

There are many modelling options available and interchangeable. Applications from a purely statistical approach can be simply based on the final method for analysis, but because applied sciences usually need to address the contextual interpretation of data and findings, communicating the rationale behind the methodological approach employed is very beneficial for progression in the field. Strategies for data fusion can be based on supervised or unsupervised intent. Recent reviews on data fusion in food science indicated that most studies made use of supervised techniques.[4,102,103]

The objective of clarifying the analytical rationale employed is not to distinguish between levels of analytical complexity, but rather to elucidate the sequence of decisions for pre-processing, exploration, optimization and choice of models for different data fusion motivations. It is a recommended good practice to create an analysis pipeline that always includes an unsupervised data exploration stage. Exploration is thus included as a step in data fusion that must be encouraged, regardless of whether the final intent is supervised or unsupervised. It is possible that pre-processing may once more be necessary, but this time in view of data.

Commonly, data modelling approaches chosen for exploration of the individual data sets are employed as pre-processing steps and for data exploration. From these exploration models, in order to optimize the final model, the standardized data are chosen. Referred to as "latent variables", these standardised measurements or model outputs are chosen as the representatives of the original (raw) data.[5,6,104] Most unsupervised data fusion models use the transformed/scaled variables as the representative, while supervised models may choose or delete features of the model. This is because unsupervised models retain all data while supervised models only seek to retain the information that will enhance the data model's performance.

The exploration of the data before trying to optimize a data fusion model is an important qualitative check that will decrease chances of overzealous data pre-processing, especially when it comes to feature or variable selection or removal. Careful examination and curation of the data allows for appropriate data integration without removal of vital elements from both statistical and contextual perspectives. This reduces the habit of thinking of the data purely statistically (as just numbers) and encourages a continued appreciation of the data as a subject in applied sciences.

Data fusion models can be evaluated for representativeness and robustness using a variety of performance parameters. Representativeness evaluates how similar the fused model is to the original data. Representativeness is used for both unsupervised and supervised approaches. A common measure of representativeness is the regression vector (RV) coefficient, which assesses the configurational similarity between data sets.[105] Representativeness is important

since the models should generate new information without compromising the integrity of the original data set. Therefore, continued data exploration is important, and optimization should never be aggressive.

Robustness evaluates how well-fitted the data are and how reliable a model is. Several goodness-of-fit criteria are used based on the techniques chosen (*e.g.*, R^2 and root mean square error). The reliability of the model relates to robustness when tested on new data. Data models are more representative of the range of variability and variation in the original data. Yet, they can sometimes be used for inference on unknown data within the same range. The range of variability/variation creates different degrees of localization of models. Generally, highly local models are more reliable than global models. For example, local models of single wine cultivar styles (for instance, green style Sauvignon Blanc wine) are more successful than global models for general cultivar or type (*i.e.*, white or red wines).

Various model performance parameters, and visualization aids (graphs and illustrations), are available for assessing the informational value of data fusion models. Using multiple visual representations helps minimize misinterpretation and/or confirmation bias, that is, limiting findings to those that were initially expected. Critical thinking must be applied, since different modelling approaches can result in different types of information extracted from a single data set. This is especially critical when working with multi-way/multi-modal and cross-domain data. Data fusion models can be considered an information bank from which different currencies can be withdrawn, that is, data of different informational value and scale. In such context, visual aids are used for knowledge compression to aid in interpretation, but caution must be taken when applying them since they can alter perceptions.

5.4.3 Data Fusion Coupled with Artificial Intelligence (AI)

Artificial intelligence (AI) and machine learning techniques are sophisticated classes of data analysis tools that increase information generation and interpretation of complex models.[106] Various data modelling techniques are available for working with both unstructured (raw) and structured (modelled) data. Data fusion coupled with machine learning techniques offers multiple avenues for data exploration, undertaking both supervised and unsupervised learning.[8,107] This makes these methods very versatile since they can be used for exploring data that have not been pre-processed. This implies minimal data treatment and increased representativeness of the models.

AI methods are generally more sensitive to minor differences between data points, meaning that they can discriminate better than classical supervised multivariate techniques. Furthermore, they can enhance differentiation between data points, thus enabling the handling of globalised models. Appropriate application of these models can result in more robust supervised models, which is one of the motivations for their applications. They are hence marketed as tools for handling big data, the idea behind it being that

they are in the high range of localization, containing high variability and variation.

AI models can generate information on both small and large sections of data with greater confidence. The potential of these techniques is that they can be used along with factorial designs to create complex data fusion models. In turn, these methods can confidently and reliably generate databases showing the relationship between data as different sets and individual data points, considering hierarchical importance as well as relative influence. Few attempts have been made in coupling data fusion modelling with machine learning in food science.[102] Studies that attempted this were supervised and aimed to increase the performance of classical multivariate methods. Most studies used mid to high-level data fusion, encompassing extensive pre-processing and found that the AI models were more robust, but none tested the representativeness of these models. One of the limitations to using AI in food science has been the insufficient generation of big data. But because researchers are experimenting with data fusion and hyphenated/mixed methods in this field nowadays, the possibilities are not far away.

5.5 Conclusion and Trends

With so many tools at their disposal, researchers have the opportunity to try to test and determine which option suits them best. However, resources are better spent on working smart. Having a lot of data is not the same as having good data; data is not information, the same as information is not automatically knowledge. Intertwining statistical and applied (contextual) reasoning will improve the generation and interpretation of results.

Analytical development seems to turn to instrumentation that is more complex and, at the same time, to more straightforward ways of dealing with data; as instrumentation becomes more sophisticated, data aspects become more accessible. When choosing a method of analysis, researchers should try not to fall into the trap of using the newest technologies without first considering their suitability. Aspects like the experimental design and the research question in context become more relevant when resources are limited.

Sensory science is a field with a lot of potential and momentum. In addition to new methods addressing specific problems (*e.g.*, temporal methods for dynamic perception of attributes), some statistical aspects should be reconsidered. One of the advantages of working cross-disciplines is the ability to adapt solutions from other disciplines to solve own issues. The cross-over into AI and its particular tools and algorithms offers the opportunity to evaluate their suitability in new contexts, such as sensory science, since many user-friendly and accessible software now allows experimentation with these advanced techniques.

Food quality is one aspect sensory researchers are somewhat hesitant to approach. As a concept, it is multi-dimensional, ill-defined and changeable depending on the circumstances. Such flexibility can be a positive point,

meaning it is "adaptable" to the needs of the researcher, but also a negative one, as the absence of a standardised definition means people can misuse it. It is possible that the solution is not in finding a strict definition, but in elucidating its features through a systematic approach similar to that used for typicality. In other circumstances, knowing the features of "quality products" might be enough for consumers and producers.

It is important to remove uncertainties surrounding statistics and unpack this "black-box". Addressing the gap requires integrating the different disciplines through transdisciplinary approaches. However, a limitation to this is that being specialized in any of the related disciplines (*i.e.*, analytical chemistry, sensory science, and data science) is difficult. Hence, true multi-disciplinarity is even harder to accomplish since it would require integrating contextual and technical knowledge of all disciplines at the same time. The first step in achieving true transdisciplinary ability is to understand the sequence of stages for problem solving in each discipline and consolidate them.

In some instances, the contextual significance can be more important than the statistical significance of model outcomes. Contextual significance is not just based on the absolute values, but also on their relative importance and meaning. This is especially important when combining chemical with sensory data.[103] 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: in absolute values, sensory data are lower in dimensionality than chemistry data and, similarly, targeted chemistry data are lower in dimensionality than untargeted chemistry data. This may be a limitation since while many statistics/omics issues can be overcome by having a larger data set, this is not always possible, especially when dealing with sensory data.

The issue of "how much data is enough data" can be difficult to answer. Researchers have to be careful not to replace quality with quantity, especially as computing power and a multitude of statistical tools are becoming easier to access. The use of 'double-checks' is recommended by evaluating multiple performance parameters – for example, comparing a goodness-of-fit criterion (*e.g.*, explained variance) with different coefficients of fit (*e.g.*, R^2 and Wilks' *Lambda*), evaluating estimated parameters such as RV and cophenetic correlation coefficients – in order to enhance the interpretability of model outcomes. This means weighing the relative importance of every parameter used against the strategy pursued and the research question originally posed.

A less explored avenue is the unsupervised approach to data handling. Such an approach requires an open mind for the outcome and its interpretation, creating more opportunities for hypothesis formation. On the other hand, it comes with the need to understand not only the limits of statistical techniques, but also the context of the data generated. The main limitation to unsupervised approaches is the sacrifice of optimal coefficients in favour of optimal goodness-of-fit. It is preferable 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.

By developing descriptive narratives of the analytical steps undertaken, from formulating a question through to the data handling process and the statistical investigations, it can be demonstrated that there is no "black-box", but perhaps a gap in critical thinking and full engagement with data handling. Researchers need to keep the "long game" in mind, from the research question, design of experiments, data acquisition and statistical strategy, to the interpretation of models in context.

References

1. C. C. Valente, F. F. Bauer, F. Venter, B. Watson and H. H. Nieuwoudt, *Sci. Rep.*, 2018, **8**, 4987.
2. V. Cariou, D. J.-R. Bouveresse, E. M. Qannari and D. N. Rutledge, in *Data Handling in Science and Technology*, Elsevier, Amsterdam, 2019, pp. 179–204.
3. V. Cariou, E. M. Qannari, D. N. Rutledge and E. Vigneau, *Food Qual. Prefer.*, 2018, **67**, 27.
4. A. Biancolillo, R. Boqué, M. Cocchi and F. Marini, in *Data Handling in Science and Technology*, ed. M. Cocchi, Elsevier, Amsterdam, 2019, pp. 271–310.
5. S. McKillup, *Statistics explained: An Introductory Guide for Life Scientists*, Cambridge University Press, Cambridge, United Kingdom, 2nd edn, 2005, p. 280.
6. J. Salkind and R. Kristin, *Encyclopedia of Measurement and Statistics*, Sage Publications, Thousand Oaks, CA, 2007, p. 1416.
7. M. Cocchi, in *Data Handling in Science and Technology*, ed. M. Cocchi, Elsevier, Amsterdam, 2019, pp. 1–26.
8. A. Sohail and F. Arif, *Prog. Biophys. Mol. Biol.*, 2020, **151**, 14.
9. A. Gandomi and M. Haider, *Int. J. Inf. Manage.*, 2015, **35**, 137.
10. M. Amargianitaki and A. Spyros, *Chem. Biol. Technol. Agric.*, 2017, **4**, 9.
11. K. Böhme, P. Calo-Mata, J. Barros-Velázquez and I. Ortea, *TrAC, Trends Anal. Chem.*, 2019, **110**, 221.
12. G. P. Danezis, A. S. Tsagkaris, V. Brusic and C. A. Georgiou, *Curr. Opin. Food Sci.*, 2016, **10**, 22.
13. G. P. Danezis, A. S. Tsagkaris, F. Camin, V. Brusic and C. A. Georgiou, *TrAC, Trends Anal. Chem.*, 2016, **85**, 123.
14. R. Korte and J. Brockmeyer, *TrAC, Trends Anal. Chem.*, 2017, **96**, 99.
15. S. Medina, R. Perestrelo, P. Silva, J. A. M. Pereira and J. S. Câmara, *Trends Food Sci. Technol.*, 2019, **85**, 163.
16. M. E. Alañón, M. S. Pérez-Coello and M. L. Marina, *TrAC, Trends Anal. Chem.*, 2015, **74**, 1.
17. T. Majchrzak, W. Wojnowski, M. Rutkowska and A. Wasik, *Trends Plant Sci.*, 2020, **25**, 302.

18. A. Cifuentes, *J. Chromatogr. A*, 2009, **1216**, 7109.
19. D. Cozzolino, *Curr. Opin. Food Sci.*, 2015, **4**, 39.
20. D. I. Ellis, *Curr. Opin. Food Sci.*, 2019, **28**, v.
21. S. Sauer and T. Luge, *Proteomics*, 2015, **15**, 997.
22. Wine-Omics, *Nature*, 2008, **455**, 699.
23. C. A. Hughey, C. M. McMinn and J. Phung, *Metabolomics*, 2016, **12**, 11.
24. E. Gorrochategui, J. Jaumot, S. Lacorte and R. Tauler, *TrAC, Trends Anal. Chem.*, 2016, **82**, 425.
25. J. P. Coutinho, G. F. Barbero, H. T. Godoy, M. Palma and C. G. Barroso, *Anal. Methods*, 2016, **8**, 1659.
26. J. Freitas, P. Silva, P. Vaz-Pires and J. S. Câmara, *Foods*, 2020, **9**, 1321.
27. P. Silva, C. L. Silva, R. Perestrelo, F. M. Nunes and J. S. Câmara, *Food Anal. Methods*, 2020, **13**, 1634.
28. J. Kopp, F. B. Zauner, A. Pell, J. Hausjell, D. Humer and J. Ebner, *et al.*, *J. Pharm. Biomed. Anal.*, 2020, **188**, 113412.
29. M. Subhi, S. Clavijo, R. Suárez, H. Seddik and V. Cerdà, *Talanta*, 2017, **167**, 695.
30. T. Tome, N. Žigart, Z. Časar and A. Obreza, *Org. Process Res. Dev.*, 2019, **23**, 1784.
31. C. Kulsing, Y. Nolvachai and P. J. Marriott, *TrAC, Trends Anal. Chem.*, 2020, **130**, 115995.
32. P. J. Marriott, S. T. Chin and Y. Nolvachai, *J. Chromatogr. A*, 2021, **1636**, 461788.
33. F. A. Franchina, D. Zanella, L. M. Dubois and J. F. Focant, *J. Sep. Sci.*, 2021, **44**, 188.
34. W. Lv, X. Shi, S. Wang and G. Xu, *TrAC, Trends Anal. Chem.*, 2019, **120**, 115302.
35. M. S. S. Amaral and P. J. Marriott, *Molecules*, 2019, **24**, 2080.
36. Y. Nolvachai, C. Kulsing and P. J. Marriott, *TrAC, Trends Anal. Chem.*, 2017, **96**, 124.
37. F. Cacciola, P. Dugo and L. Mondello, *TrAC, Trends Anal. Chem.*, 2017, **96**, 116.
38. K. Arena, F. Mandolfino, F. Cacciola, P. Dugo and L. Mondello, *J. Sep. Sci.*, 2021, **44**, 17.
39. M. Navarro-Reig, C. Bedia, R. Tauler and J. Jaumot, *Proteomics*, 2018, **18**, 1700327.
40. E. Hatzakis, *Compr. Rev. Food Sci. Food Saf.*, 2019, **18**, 189.
41. F. Tang, M. Vasas, E. Hatzakis and A. Spyros, in *Annual Reports on NMR Spectroscopy*, Academic Press, Cambridge, MA, 2019, pp. 239–306.
42. D. S. Wishart, *J. Magn. Reson.*, 2019, **306**, 155.
43. K. Chandra, S. Al-harthi, S. Sukumaran, F. Almulhim, A. Emwas and H. S. Atreya, *et al.*, *RSC Adv.*, 2021, **11**, 8694.
44. K. Fan and M. Zhang, *Crit. Rev. Food Sci. Nutr.*, 2019, **59**, 2202.
45. S. Grassi and C. Alamprese, *Curr. Opin. Food Sci.*, 2018, **22**, 17.
46. J. U. Porep, D. R. Kammerer and R. Carle, *Trends Food Sci. Technol.*, 2015, **46**, 211.

47. A. R. Monforte, S. I. F. S. Martins and A. C. S. Ferreira, *15th Weurman Flavour Research Symposium*, Graz, Austria, 2017.
48. N. P. Kalogiouri and V. F. Samanidou, *Environ. Sci. Pollut. Res.*, 2020, 59150.
49. A. M. Knolhoff and T. R. Croley, *J. Chromatogr. A*, 2016, **1428**, 86.
50. L. Lacalle-Bergeron, D. Izquierdo-Sandoval, J. V. Sancho, F. J. López, F. Hernández and T. Portolés, *TrAC, Trends Anal. Chem.*, 2021, **135**, 116161.
51. F. Stilo, C. Bicchi, A. M. Jimenez-Carvelo, L. Cuadros-Rodriguez, S. E. Reichenbach and C. Cordero, *TrAC, Trends Anal. Chem.*, 2021, **134**, 116133.
52. Y. Xu, P. Zhong, A. Jiang, X. Shen, X. Li and Z. Xu, *et al.*, *TrAC, Trends Anal. Chem.*, 2020, **131**, 116017.
53. D. J. Beale, F. R. Pinu, K. A. Kouremenos, M. M. Poojary, V. K. Narayana and B. A. Boughton, *et al.*, *Metabolomics*, 2018, **14**, 152.
54. I. Ortea, G. O'Connor and A. Maquet, *J. Proteomics*, 2016, **147**, 212.
55. M. Gil, C. Reynes, G. Cazals, C. Enjalbal, R. Sabatier and C. Saucier, *Sci. Rep.*, 2020, **10**, 1170.
56. A. Versari, V. F. Laurie, A. Ricci, L. Laghi and G. P. Parpinello, *Food Res. Int.*, 2014, **60**, 2.
57. R. Spicer, R. M. Salek, P. Moreno, D. Cañueto and C. Steinbeck, *Metabolomics*, 2017, **13**, 106.
58. M. Bevilacqua, R. Bro, F. Marini, Å. Rinnan, M. A. Rasmussen and T. Skov, *TrAC, Trends Anal. Chem.*, 2017, **96**, 42.
59. L. G. Johnsen, P. B. Skou, B. Khakimov and R. Bro, *J. Chromatogr. A*, 2017, **1503**, 57.
60. K. Sirén, U. Fischer and J. Vestner, *Anal. Chim. Acta: X*, 2019, **1**, 100005.
61. R. Tautenhahn, G. J. Patti, D. Rinehart and G. Siuzdak, *Anal. Chem.*, 2012, **84**, 5035.
62. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779.
63. T. Pluskal, S. Castillo, A. Villar-Briones and M. Orešič, *BMC Bioinf.*, 2010, **11**, 395.
64. A. R. Monforte, S. I. F. S. Martins and A. C. S. Ferreira, *Food Chem.*, 2021, **352**, 129288.
65. O. A. Adebo, S. A. Oyeyinka, J. A. Adebiyi, X. Feng, J. D. Wilkin and Y. O. Kewuyemi, *et al.*, *Int. J. Food Sci. Technol.*, 2021, **56**, 1514.
66. N. Feizi, F. S. Hashemi-Nasab, F. Golpelichi, N. Saburouh and H. Parastar, *TrAC, Trends Anal. Chem.*, 2021, **138**, 116239.
67. D. Valentin, S. Chollet, M. Lelièvre and H. Abdi, *Int. J. Food Sci. Technol.*, 2012, **47**, 1563.
68. M. Bécue-Bertaut, *Food Qual. Prefer.*, 2014, **32**, 2.
69. E. Campo, J. Ballester, J. Langlois, C. Dacremont and D. Valentin, *Food Qual. Prefer.*, 2010, **21**, 44.
70. M. Lelièvre-Desmas, D. Valentin and S. Chollet, *Food Qual. Prefer.*, 2017, **61**, 6.
71. P. Varela and G. Ares, *Food Res. Int.*, 2012, **48**, 893.

72. G. Garrido-Bañuelos, V. Panzeri, J. Brand and A. Buica, *J. Sens. Stud.*, 2020, **35**, e12575.
73. L. Perrin and J. Pagès, *J. Sens. Stud.*, 2009, **24**, 749.
74. J. C. Castura, L. Antúnez, A. Giménez and G. Ares, *Food Qual. Prefer.*, 2016, **47**, 79.
75. D. Tao, P. Yang and H. Feng, *Compr. Rev. Food Sci. Food Saf.*, 2020, **19**, 875.
76. P. Yu, M. Y. Low and W. Zhou, *Trends Food Sci. Technol.*, 2018, **71**, 202.
77. T. Naes, P. B. Brockhoff and O. Tomic, *Statistics for Sensory and Consumer Science*, John Wiley & Sons, Chichester, 2010, p. 304.
78. K. M. Carabante and W. Prinyawiwatkul, *J. Sens. Stud.*, 2018, **33**, e12435.
79. G. L. Marcazzan, C. Mucignat-Caretta, C. Marina Marchese and M. L. Piana, *J. Apic. Res.*, 2018, **57**, 75.
80. R. L. Heiniö, M. W. J. Noort, K. Katina, S. A. Alam, N. Sozer and H. L. de Kock, *et al.*, *Trends Food Sci. Technol.*, 2016, **47**, 25.
81. A. N. Schiano, W. S. Harwood and M. A. Drake, *J. Dairy Sci.*, 2017, **100**, 9966.
82. J. M. Ennis, B. Rousseau and D. M. Ennis, *J. Sens. Stud.*, 2014, **29**, 89.
83. A. Sarkar and E. M. Krop, *Curr. Opin. Food Sci.*, 2019, **27**, 64.
84. B. V. Humia, K. S. Santos, A. M. Barbosa, M. Sawata, M. C. Mendonça and F. F. Padilha, *Molecules*, 2019, **24**, 1568.
85. B. Piqueras-Fiszman and C. Spence, *Food Qual. Prefer.*, 2015, **40**, 165.
86. M. C. Meilgaard, C. E. Dalgliesh and J. F. Clapperton, *J. Inst. Brew.*, 1979, **85**, 38.
87. J. Brand, V. Panzeri and A. Buica, *Foods*, 2020, **9**, 805.
88. J. Brand, M. Kidd, L. van Antwerpen, D. Valentin, T. Naes and H. H. Nieuwoudt, *S. Afr. J. Enol. Vitic.*, 2018, **39**, 163.
89. H. Hopfer, J. Nelson, S. E. Ebeler and H. Heymann, *Molecules*, 2015, **20**, 8453.
90. Å. Rinnan, F. van den Berg and S. B. Engelsen, *TrAC, Trends Anal. Chem.*, 2009, **28**, 1201.
91. Å. M. Wheelock and C. E. Wheelock, *Mol. BioSyst.*, 2013, **9**, 2589.
92. S. Chollet, D. Valentin and H. Abdi, *Food Qual. Prefer.*, 2005, **16**, 13.
93. P. Faye, P. Courcoux, A. Giboreau and E. M. Qannari, *Food Qual. Prefer.*, 2013, **28**, 317.
94. K.-Y. M. Lee, A. Paterson, J. R. Piggott and G. D. Richardson, *J. Inst. Brew.*, 2001, **107**, 287.
95. M. C. Meilgaard, C. E. Dalgliesh and J. F. Clapperton, *J. Inst. Brew.*, 1979, **85**, 38.
96. J. Mojet and S. de Jong, *Grasas Aceites*, 1994, **45**, 42.
97. I. S. Koch, M. Muller, E. Joubert, M. van der Rijst and T. Næs, *Food Res. Int.*, 2012, **46**, 217.
98. G. J. Pickering and P. Demiglio, *J. Wine Res.*, 2008, **19**, 51.
99. D. Valentin, S. Chollet, M. Leliévre and H. Abdi, *Int. J. Food Sci. Technol.*, 2012, **47**, 1563.

100. M. Mafata, *A. Chemometric Approach to Investigating South African Wine Behaviour Using Chemical and Sensory Markers*, PhD thesis, Stellenbosch University, 2021, p. 163.
101. A. Smolinska, J. Engel, E. Szymanska, L. Buydens and L. Blanchet, in *Data Handling in Science and Technology*, ed. M. Cocchi, Elsevier, Amsterdam, 2019, pp. 51–79.
102. E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña and O. Busto, *Anal. Chim. Acta*, 2015, **891**, 1.
103. S. Seisonen, K. Vene and K. Koppel, *Food Chem.*, 2016, **210**, 530.
104. W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer, Berlin, Heidelberg, Germany, 2015.
105. H. Abdi, in *Encyclopedia of Measurement and Statistics*, ed. N. Salkind, Sage, Thousand Oaks, CA, 2007, p. 10.
106. T. Meng, X. Jing, Z. Yan and W. Pedrycz, *Inf. Fusion*, 2020, **57**, 115.
107. Y. Zheng, *IEEE Trans. Big Data*, 2015, **1**, 16.