University of Modena and Reggio Emilia

Research Doctorate School in Agri-Food Sciences, Technologies and Bio-Technologies

XXXVII cycle

Chemometric Strategies for Food Products Optimization and Control

Ph.D. Candidate: Pier Lorenzo Rolando

Tutor:	Prof. Alessandro Ulrici
Co-tutors:	Prof. Rosalba Calvini
	Prof. Giorgia Foca
Dean of Ph.D. School:	Prof. Fabio Licciardello

2021 – 2024

“For every complex problem there is a solution which is clear, simple and wrong.”

H. L. Mencken

Abstract

This thesis explores different applications of chemometric approaches for the optimization and control of food products.

The first part of this thesis focuses on the optimization of food products, in particular considering color stability of Strawberry Yogurt Purée (SYP), used as a case study conducted in collaboration with a food company. SYP is a semi-finished product used by other industries in the production of strawberry-flavored yoghurts, where color stability is a crucial factor influencing consumer appeal for final food products. The objective was to assess how different SYP formulations affect color and its degradation over time due to browning phenomena. A combined approach using I-optimal mixture design and multivariate analysis of RGB images was employed. Strawberry purée, sugar, lemon juice, and two types of thickeners were mixed in varying proportions to create 44 SYP formulations. These samples were subjected to light and temperature stress for five weeks, during which the corresponding RGB images were captured, along with the images of the corresponding control samples. The images were analyzed using two approaches: first, Principal Component Analysis (PCA) was applied to colourgrams to explore the main sources of variability, and second, median values of color-related parameters were calculated, from which Response Surfaces and Partial Least Squares - Discriminant Analysis (PLS-DA) classification models were built. The analysis revealed that relative green (i.e., the ratio of green to lightness) and red values are key indicators of color degradation, with formulations containing a higher percentage of strawberry purée showing greater color variation. Results also revealed that the type of thickener may influence the degradation kinetics.

The second part of this thesis focuses on food control through the analysis of the Processing Factor (PF) database, a critical tool in European risk assessment conducted by the European Food Safety Authority (EFSA). This publicly available database harmonizes the collection and evaluation of pesticide residue data for processed foods. This information is condensed into the PF parameter, which corresponds to the ratio of the residue of a specific pesticide in the processed product to that in the corresponding unprocessed product. The PF database includes more than 250 chemical substances across 120 processed commodities derived from 70 different food matrices. To analyze this dataset, extensive preprocessing was conducted, including the encoding of categorical features into binary data. To obtain valuable insights on pesticides residues behavior, data analysis and visualization was performed using both the classical tools available to display categorical data (e.g., treemap and alluvial plots) and an innovative multivariate approach based on PCA. Since PCA itself cannot be applied to categorical data, the proposed approach consists in transforming the matrix of categorical data into the corresponding distance matrix according to Jaccard index, and then apply PCA to the Jaccard distance matrix. The results demonstrated the added value of a multivariate chemometric strategy, even when applied to datasets with categorical features, and revealed potential trends to enhance food safety and control measures in food processing.

Through the application of chemometric techniques, this research offers new perspectives on both the optimization and formulation of food products on one side, and the enhancement of control systems on the other, contributing to safer and higher quality food products.

Riassunto

Questa tesi esplora diverse applicazioni di approcci chemiometrici per l'ottimizzazione ed il controllo dei prodotti alimentari.

La prima parte del lavoro si focalizza sull'ottimizzazione dei prodotti alimentari, in particolare attraverso lo studio, condotto in collaborazione con un'azienda, volto ad indagare la stabilità del colore di un semilavorato alla fragola (Strawberry Yoghurt Purée, SYP) destinato all'industria dello yogurt. Il colore è da sempre un fattore cruciale che influenza l'attrattiva del consumatore finale per i prodotti alimentari. L'obiettivo era valutare come le diverse formulazioni di SYP influenzino il colore e la sua degradazione nel tempo a causa di fenomeni di imbrunimento. A tal fine, è stato utilizzato un approccio combinato di disegno di miscela I-ottimale e analisi multivariata di immagini RGB. Purea di fragole, zucchero, succo di limone e due tipologie di addensante sono stati miscelati in proporzioni variabili per ottenere 44 formulazioni di SYP. Queste sono state sottoposte a condizioni di luce e temperatura elevate per cinque settimane, durante le quali sono state acquisite le immagini RGB in parallelo con quelle dei corrispondenti campioni di controllo. Le immagini sono state analizzate utilizzando due approcci: i) analisi delle componenti principali (PCA) sui colorigrammi ottenuti dalle immagini, al fine di esplorare le principali fonti di variabilità; ii) calcolo dei valori mediani dei parametri di colore per la creazione di superfici di risposta e di modelli di classificazione multivariata mediante Partial Least Squares-Discriminant Analysis (PLS-DA). L'analisi ha individuato il verde relativo (ovvero il rapporto tra verde e luminosità) e il rosso come indicatori chiave della degradazione del colore, con una maggiore variazione cromatica per le formulazioni a più alto contenuto di fragola; inoltre, è stato evidenziato che il tipo di addensante può influenzare la cinetica di degradazione.

La seconda parte del lavoro è focalizzata sul controllo alimentare, attraverso l'analisi del database dei valori di Processing Factor (PF), uno strumento essenziale nelle valutazioni sulla sicurezza alimentare condotte dall'Autorità Europea per la Sicurezza Alimentare (EFSA). Questo database, accessibile pubblicamente, armonizza la raccolta e la valutazione dei residui di pesticidi negli alimenti trasformati. Questa informazione viene condensata nel parametro PF, dato dal rapporto tra il residuo di un pesticida nel prodotto trasformato e quello nella materia prima agricola. Il database dei PF include più di 250 sostanze chimiche, in 120 prodotti trasformati derivati da 70 matrici alimentari diverse. Per analizzare questo dataset si sono resi necessari molti step di pretrattamento, inclusa la codifica delle categorie in dati binari. Per ottenere informazioni utili sul comportamento dei pesticidi, sono state utilizzate sia tecniche più classiche di analisi e visualizzazione dei dati categorici (e.g., treemap e alluvial plot) sia un approccio multivariato innovativo basato su PCA. Dato che PCA non può essere usata direttamente su dati categorici, l'approccio proposto prevede la loro trasformazione in valori di distanze di Jaccard, quindi l'applicazione di PCA sulla risultante matrice delle distanze di Jaccard. I risultati hanno dimostrato il valore aggiunto di una strategia chemiometrica multivariata, persino quando applicata a dataset di questo tipo, rivelando potenziali miglioramenti nei processi di trasformazione per la sicurezza e il controllo dei prodotti alimentari.

Attraverso l'applicazione di tecniche chemiometriche, questa ricerca offre nuove prospettive sia sull'ottimizzazione e formulazione dei prodotti, sia sul miglioramento dei sistemi di controllo, contribuendo all'aumento della sicurezza e della qualità dei prodotti alimentari.

Table of Contents

Chapter 1: General Overview 1

Chapter 2: Theory 2

2.1 Design of Experiments and Mixture Design 2

2.2 Multivariate Data Analysis 6

2.2.1 Principal Component Analysis (PCA) 6

2.2.2 Partial Least Squares-Discriminant Analysis (PLS-DA) 9

2.3 RGB Imaging for Food Analysis 11

2.3.1 Multivariate Image Analysis 13

2.4 Graphical Representation of Categorical Data 15

2.4.1 Treemap 15

2.4.2 Alluvial Plot 18

2.4.3 PCA on Jaccard Distance Matrix 19

Chapter 3: Food Optimization 23

3.1 Introduction 24

3.2 Materials and Methods 26

3.2.1 Mixture Design for Strawberry Yoghurt Samples Preparation 26

3.2.2 Experimental Set-Up for Stressing the Samples 27

3.2.3 Image Acquisition System 28

3.2.4 Data Dimensionality Reduction of RGB Images Dataset 30

3.2.5 Exploratory Analysis of *Colourgrams* by PCA 31

3.2.6 Modelling of Median Red Parameter 31

3.2.7 Evaluation of Color Variation by PLS-DA 32

3.3 Results and Discussion 33

3.3.1 PCA of *Colourgrams* 33

3.3.2 Mixture Design Models on Median Red Parameter 34

3.3.3 PLS-DA Model on Median Values 37

3.4 Conclusions 42

Chapter 4: Food Control 43

4.1 Introduction 43

4.1.1 European Legislation on Pesticides Maximum Residues Levels 44

4.1.2 Processing Factor 46

4.2 Materials and Methods 48

4.2.1 European Database of Processing Factors ... 48
4.2.2 Data Filtering – Aggregation – Enrichment ... 49
4.2.3 Treemap and Alluvial Plot ... 52
4.2.4 J-PCA ... 53
4.3 Results and Discussion ... 54
4.3.1 Raw Agricultural Commodity ... 54
4.3.1.1 Citrus Fruits ... 55
4.3.1.2 Berries and Small Fruits ... 64
4.3.1.3 Fruits (fresh or frozen) and Tree Nuts ... 71
4.3.1.4 Cereals ... 80
4.3.2 Process and Processed Commodity ... 87
4.3.2.1 Fermentation-Distillation ... 87
4.3.2.2 Fruit Juice ... 98
4.3.3 All Processing Factors Database ... 106
4.4 Conclusions ... 114
Chapter 5: Final Considerations ... 115
Appendices ... 116
Appendix I – RAC Groups and Subgroups ... 116
Appendix II – Processing Techniques (with OECD groups) ... 118
Appendix III – Processed Commodities Groups ... 120
Appendix IV – Active Substances with WHO Risk Class and Use ... 123
Appendix V – White Wine Process Flowchart (V 001) ... 124
Appendix VI – Red Wine Process Flowchart (V 002) ... 125
Appendix VII – Beer Process Flowchart (V 005 – V 006) ... 126
Appendix VIII – Citrus Juice Flowchart (II 001) ... 127
Appendix IX – Pome Juice Flowchart (II 002) ... 128
Appendix X – Grape Juice Flowchart (II 003) ... 129
References ... 130

List of Publications

Rolando, P. L., Calvini, R., Foca, G., & Ulrici, A. (2023). Mixture design and multivariate image analysis to monitor the colour of strawberry yoghurt purée. Microchemical Journal, 194, 109222. https://doi.org/10.1016/j.microc.2023.109222.

List of Abbreviations and Notation

AS	active substance[1]
DoE	design of experiments
EC	European Community
EFSA	European Food Safety Authority
EU	European Union
LV	latent variable
MIA	multivariate image analysis
MRL	maximum residue level
OECD	Organization for Economic Co-operation and Development
PC	principal component[2], processed commodity[3]
PCA	principal component analysis[4]
PF	processing factor
PLS	partial least squares
PLS-DA	partial least squares-discriminant analysis
RAC	raw agricultural commodity
RGB	red, green, blue
RSM	response surface methods
SYP	strawberry yoghurt purée
WHO	World Health Organization

[1] In the context of this thesis the term "pesticide" is also used to refer to active substances (*see* Section 4.1.1).
[2] When referring to PCA. In score plots description combined with a number (e.g., PC1, PC2).
[3] When referring to processing factors database structure. In treemap and alluvial plots description also in combination with word "group" (e.g., PC Group = processed commodity group).
[4] When preceded with J (J-PCA), refers to PCA applied on Jaccard distance matrix.

Chapter 1: General Overview

Food safety and quality are fundamental priorities in modern food production, reflecting consumer demands for transparency, regulatory compliance, and public health protection. As food systems evolve to meet these expectations, the fields of food optimization and control emerge as critical areas for scientific and technological advancement. This thesis focuses on the application of chemometric strategies to address challenges in these domains, showcasing their utility in enhancing both food product development and food safety monitoring.

Food optimization focuses on improving product attributes (such as sensory qualities, nutritional value, and stability) while ensuring cost-effective production. In the context of this work, optimizing the color stability of semi-finished strawberry yogurt purée (SYP) serves as a representative case study. Since color is a key factor influencing consumer acceptance, its stability under stress conditions is vital for maintaining product appeal.

Conversely, food control involves ensuring the integrity of food products throughout the supply chain. This is exemplified in this thesis by the analysis of the European Database of Processing Factors (PF), a vital tool for assessing pesticide residue levels during food processing. The PF database not only informs regulatory decisions on maximum residue levels (MRLs) but also supports comprehensive dietary risk assessments.

Chemometric techniques offer a powerful toolkit for handling the complexity of food datasets, enabling the extraction of meaningful patterns and insights. These methods are used in this thesis for:

- **Food Optimization**: utilizing Multivariate Image Analysis (MIA) and Design of Experiments (DoE) to model and predict quality parameters;
- **Food Control**: adapting multivariate exploratory data analysis techniques, such as Principal Component Analysis (PCA), to datasets primarily composed of categorical features, like the European Database of Processing Factors (PF), to enhance applications of exploratory data analysis.

In addition to the present introductory chapter, the thesis is organized into four additional chapters.

Chapter 2 reviews the theoretical foundations of key methods, including design of experiments, multivariate data analysis, and graphical tools to visualize categorical data.

Chapter 3 focuses on the optimization of SYP formulations using multivariate image analysis and mixture design to assess color stability.

Chapter 4 explores the analysis of the PF database, detailing preprocessing steps and applying innovative visualization and multivariate techniques.

Chapter 5 summarizes the findings, emphasizing the implications for food safety and quality, and outlines future research directions.

Through the application of chemometric strategies, this thesis is aimed at contributing to both the enhancement of food product development and the reinforcement of safety protocols, aligning with contemporary demands for sustainable and transparent food systems.

Chapter 2: Theory

2.1 Design of Experiments and Mixture Design

Design of Experiments (DoE), also known as experimental design, is a multivariate approach used to efficiently explore an experimental domain defined by a set of experimental factors (x_1, x_2, …, x_n) and their relationship with system responses (y_1, y_2, …, y_n). First introduced in 1935 by Ronald Fisher, a renowned statistician, biologist, and mathematician, DoE is now widely applied across numerous scientific fields. Its primary objective is to maximize the ratio between the quality of extracted information and the required experimental effort (Benedetti et al., 2020). By strategically planning and selecting experiments, DoE enables researchers to obtain robust insights with a reduced number of trials. DoE is considered foundational in chemometrics as it lays the groundwork for mathematical modeling and algorithm calibration.

Any experimental problem can be framed in terms of relationships between controllable experimental factors (x) and responses (y). Controllable experimental factors are independent variables that can be controlled and set to specific values, whereas responses are measurable, factor-dependent quantities that provide insight into how individual factors and their interactions influence the system. The initial and most critical step in DoE is to define which quantitative and/or qualitative factors, either independently or in combination, are likely to influence the response values of interest. This identification is based on prior empirical knowledge or existing scientific literature. Then, identified factors are assigned to fixed levels within a specified range of variation[5] and this defines the experimental domain containing the experimental conditions. Finally, in the *n*-dimensional experimental domain defined by the *n*-factors, response values can be investigated through Response Surface Methodology (RSM). This approach enables researchers to analyze how variables influence the response and to construct a mathematical model capable of predicting the response at any point within the experimental domain, including points that were not explicitly tested, with a known level of precision. Benefits of DoE techniques are clear, since they permit simultaneous investigation of the effect of more factors and to detect the effect of their interactions, to predict the response in every point of the domain and to rationalize the experimental effort. In summary, a good DoE is applied when: i) a final empirical model which can describe the response (or responses) as a function of the experimental factors is obtained; ii) system responses in domain points that were not previously tested can be predicted with a solid statistical accuracy; iii) the chosen design is the one that meets the experimenter needs and best fits the data, providing at the same time the simplest mathematical approximation of the system under study (Benedetti et al., 2020).

What is described above is the classical *Response Surface Design*, but other types of DoE exist. *Screening Designs* are used when too many variables are initially identified: their scope is to select the most relevant ones to simplify the subsequent modelization. *Response Surface Designs* and *Screening Designs* are generally employed for the study of independent variables (i.e., characterized by null or minimal covariance among each other). On the other hand, *Mixture Designs* are used in the case of non-independent variables, like components of a mixture, since their final value depends on the amount of each other component. In *Mixture Designs*, the following conditions apply:

[5] It is worth mentioning that not all factors can be set on experimenter desire: the effect of uncontrollable factors like weather conditions or biological variability, must be minimized with adequate blocking and randomization strategies.

$$x_i \geq 0, \qquad i = 1,2,\dots,q \qquad\qquad \sum_{i=1}^{q} x_i = x_1 + x_2 + \cdots + x_q = 1 \qquad \textbf{(1)}^{6}$$

where q is the number of mixture components and x_i is the proportion of the i^{th} component in the mixture.

According to Equation 1, it is possible that a single component could make up the entire mixture: this is called a pure component or a single-component mixture. More realistic situations involve $q > 1$ components, with $q-1$ establishing the dimensions of the geometric space containing all possible mixture points, i.e., of the mixture domain. This space is called simplex. For $q = 2$ components, the simplex factor space is a segment and each blend of the two components is represented by a point on the segment. Points on extremes of this segment represent the single-component mixtures. With three components ($q = 3$), the simplex factor space is an equilateral triangle, and for $q = 4$ the simplex space is a tetrahedron. Whatever the simplex region, RSMs are then used to explore response(s): the choice of a proper model with suitable design and adequate testing are key factors for a good surface modelization and interpretation (Shewhart et al., 2002).

The main scope of mixture designs is the determination of the exact mathematical equation that adequately represents:

$$\eta = \phi\left(x_1, x_2, \dots, x_q\right) \qquad \textbf{(2)}$$

where η is the response, $x_1, x_2, \dots, x_q$ are the proportions of mixture components and Φ is the relation. In most cases, *Scheffé* first or second-degree polynomials are used.

$$\eta = \sum_{i=1}^{q} \beta_i \; x_i \qquad\qquad \eta = \sum_{i=1}^{q} \beta_i \; x_i + \sum_{i=1}^{q} \sum_{j>i}^{q} \beta_{ij} \; x_i \; x_j \qquad \textbf{(3)}$$

Sometimes even the 3-factors interactions may be relevant for the response: in such cases special-cubic or full-cubic designs must be applied. In any case, final properties of the polynomial used to estimate the response function primarily depend on the specific program of experiments applied, on a case-by-case basis (Shewhart et al., 2002).

[6] It may happen however that the sum of the component proportions is less than unity. Such cases include so-called "constant mixture factors": fixed value for one or more components that are assumed non-significant.

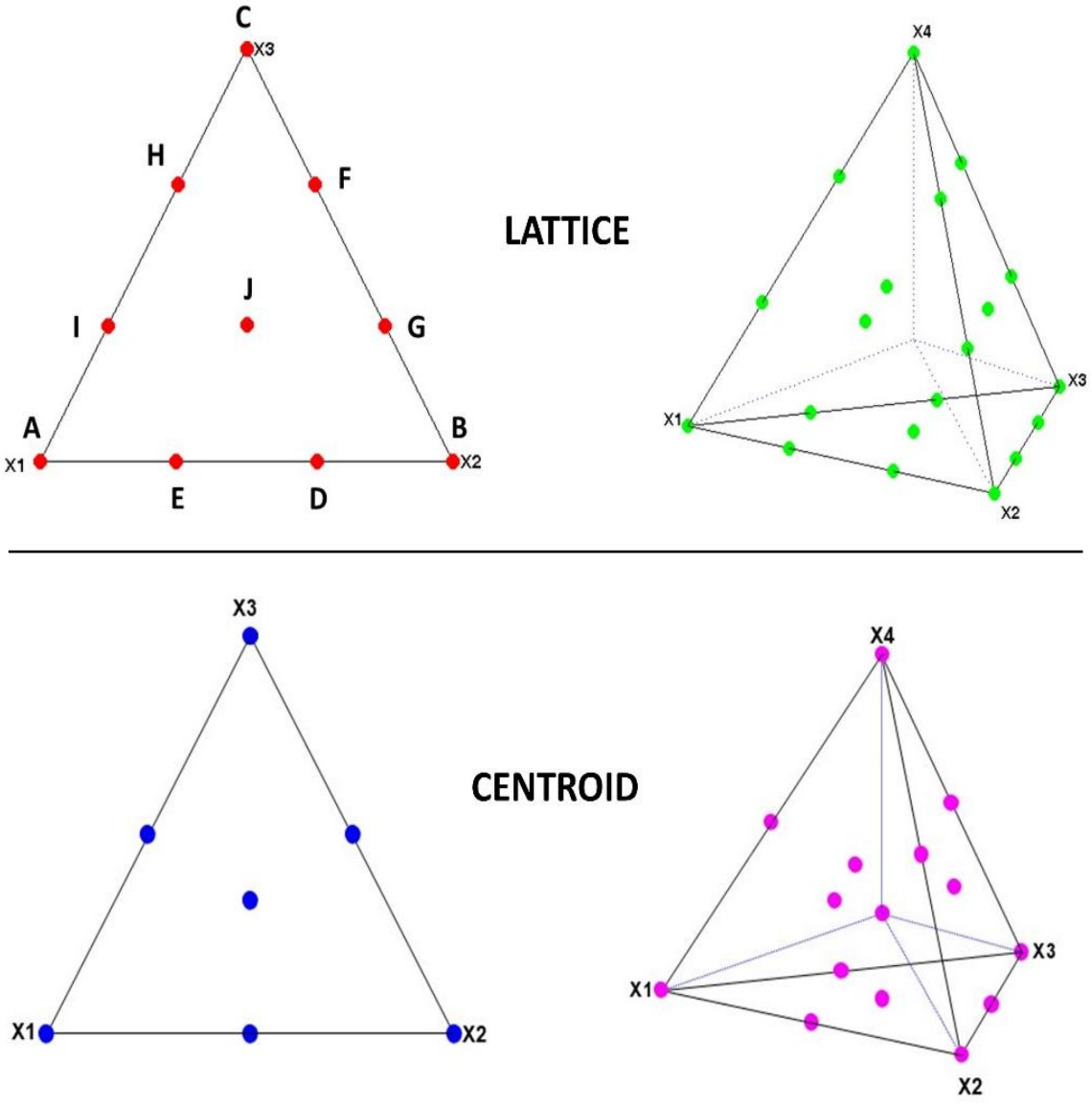


Figure 2-1 Representation of full simplex domain for lattice (above) and centroid (below) designs with evidence of the experimental data points. Three-components (x = 3) designs are displayed on the left while correspondent four-components (x = 4) on the right.

The experimental design also defines the range of component proportions. This is because it may cover the entire simplex factor space *(Simplex Lattice Design* and *Simplex Centroid Design* as seen in Fig. 2-1) or only a subregion of it. In this situation, additional constraints in the form of upper and/or lower bounds are placed on the component proportions. From a geometric perspective, setting lower and/or upper bounds can be seen as "cutting" the original factor space into smaller subregions. This cutting operation occurs in parallel to the opposite edge (three-components system) or to the opposite surface (four-component system) of the considered component, in correspondence to lower/upper bounds values. In the particular case when only lower bounds are present on each component, the shape of the subregion is not distorted, and it retains the shape of a regular simplex (Fig. 2-2 A).

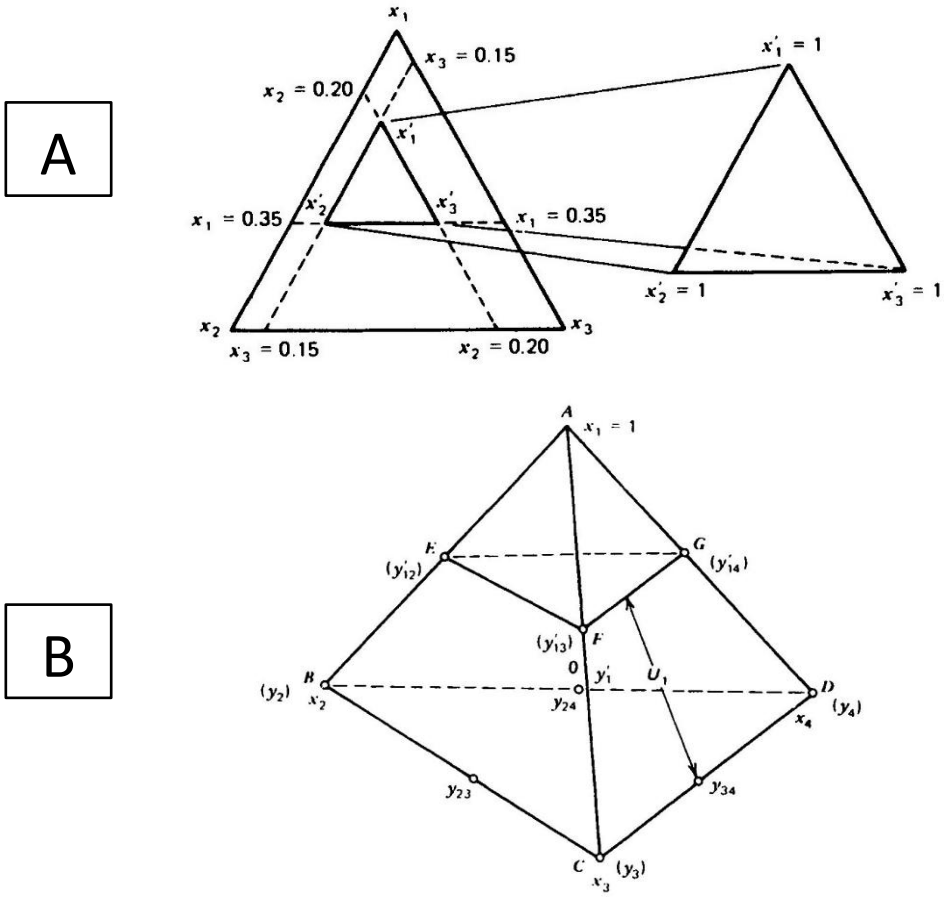


Figure 2-2 Coming from (Shewhart et al., 2002), in A) only lower bounds application results in a regular sub-simplex of the original one. In B) an upper bound is placed on component x1 (A vertex of the tetrahedron), simplex sub-region becomes irregular.

When one or more upper bounds are defined or if both upper and lower bounds are placed (as for the food product optimization application in this work) the resulting region usually does not assume the shape of a regular simplex and optimization algorithms are needed to select the best experimental points for further modelization (Fig. 2-2 B). *D-optimality* and *I-optimality* criteria are the most applied, but many others exist. All these methods work in a similar manner: the user defines the number of experimental points and starting from a list of candidate points, the final experimental points are selected based on different mathematical quality criteria (e.g., minimizing variance in prediction or minimizing the determinant of the variance matrix).

2.2 Multivariate Data Analysis

2.2.1 Principal Component Analysis (PCA)

Among the various multivariate statistical techniques, Principal Component Analysis (PCA) is one of the most widely used methods for data exploration, dimensionality reduction and pattern recognition in complex dataset (Bro & Smilde, 2014). PCA is a mathematical approach for reorganizing and simplifying the information within a multivariate dataset. It analyzes all the data simultaneously describing both objects' structure, variables' structure and correlations between objects and variables. While PCA can be applied to datasets with a small number of variables, its utility becomes particularly apparent when dealing with large datasets, such as those generated from spectroscopic techniques. PCA calculates a new set of variables, called Principal Components (PCs), which describe the directions of maximum variance in the data. This enables the description of the relevant data structure using considerably fewer variables than the original ones (Davies & Fearn, 2004). PCA is a multivariate exploratory data analysis technique, which simplifies the original *m*-dimensional variables space in a simpler *a*-dimensional one (with $a < m$ and a = number of *PCs* selected) on which samples are projected. This enables easier and faster exploratory data analysis through the interpretation of few bi- or tri-dimensional plots.

Rules for computing PCs are quite simple: they are based on the criteria of maximum variance and orthogonality. The first PC is the direction that explains the highest variance in the data. PC2 must be orthogonal to PC1 and describes the maximum amount of the residual variance, i.e., of the variance not already described by PC1 and so on, until a proper number of PCs are defined (Fig. 2-3).

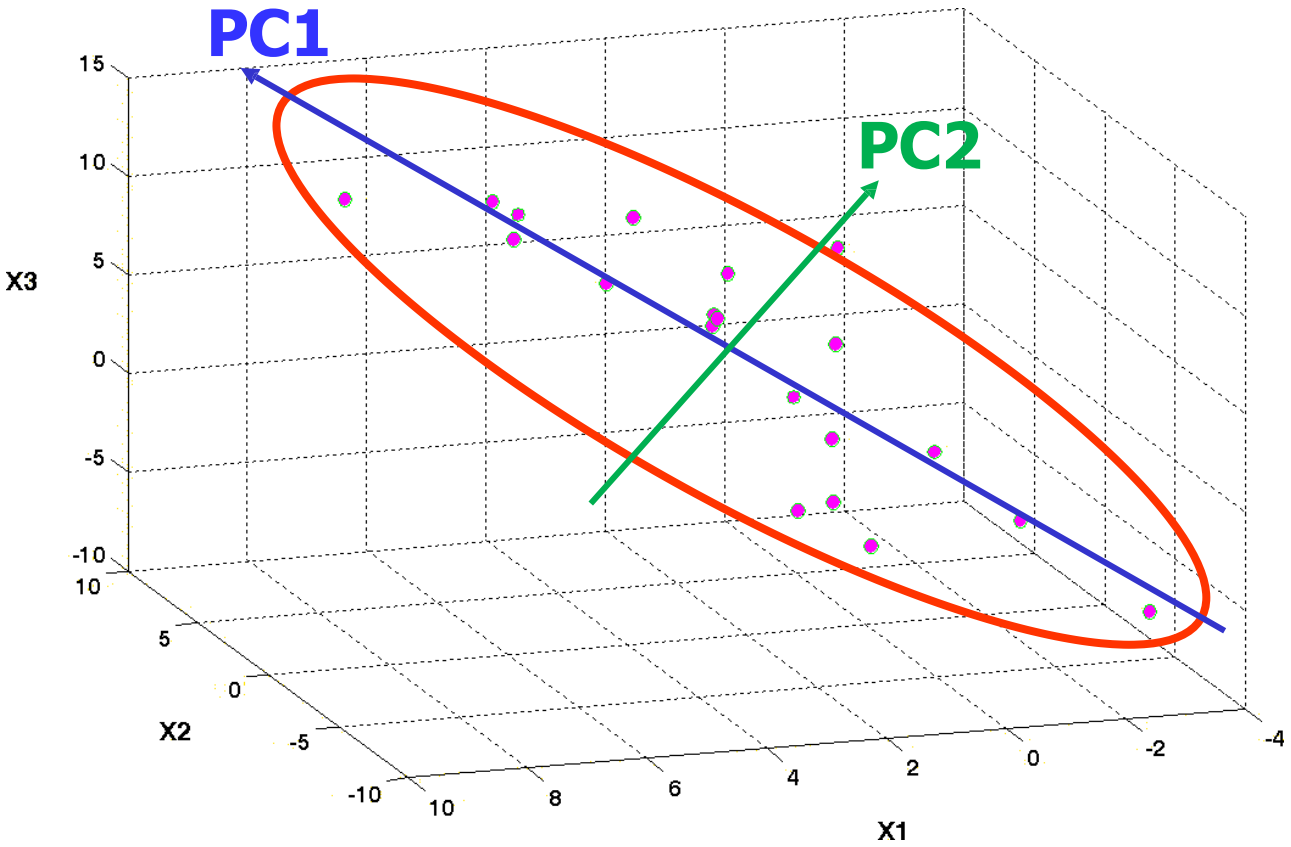


Figure 2-3 Example of a PC1-PC2 space projection of n samples described by three initial variables (X1, X2 and X3). PC1 (blue) follows the direction of maximum variance of the n samples. PC2 (green) is orthogonal to PC1 and follows the direction of maximum residual variance.

Samples can now be observed in PCs space, and they assume a new set of values called **scores**. Scores simply represent orthogonal projections of objects along a given PC. A score vector is the collection of all object projections for that PC. Objects score values along PCs can be positive or negative: this is because PCs are characterized by a zero-origin corresponding to the centroid of (at least) mean-centered data. Indeed, before performing PCA, basic column-wise data pre-treatments are necessary:

- **Mean-centering**: this is done by calculating the mean value of each variable and then subtracting its value from each measurement so that the variables vary around zero; mean centered variables have mean equal to zero and retain the standard deviation of the original data;
- **Autoscaling**: for datasets with variables on different scales, mean-centered data should be divided by their standard deviation to avoid scale effects; autoscaled variables have mean equal to zero and unit standard deviation.

Further and more advanced row-wise data pre-treatments are possible and suggested when dealing with specific data type (e.g., spectroscopic data and in general signal-like data), like e.g., smoothing, derivatives, standard normal variate, detrending, multiplicative scatter correction. Data pre-treatments can be applied alone (e.g., only mean-centering) or in combination with each other: the choice of right pre-treatments depends on the specific dataset and must be determined on a case-by-case basis.

As previously highlighted, the position of the samples in the new PCs space is described by the corresponding score values. Plotting PC score vectors against each other yields a **score plot** (Fig. 2-4), which is particularly useful to visualize and analyze the data structure.

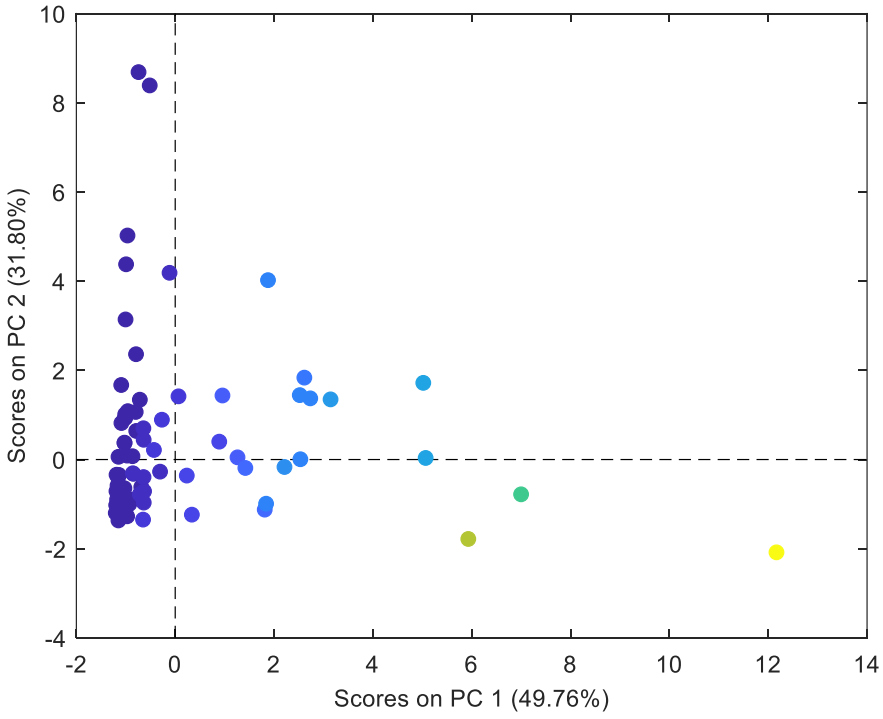


Figure 2-4 Example of a score plot. Whatever feature describes samples color scale (from dark blue to yellow), it can be observed that yellow feature increases with positive PC1 score values, while dark blue feature with negative PC1 score values. Within the dark blue feature, PC2 describes some variation along its score values.

PCA also describes variables structure through **loading vectors** and **loading plots**. Similar to score values, which represent the influence of each object on a particular PC, the loading vectors represent the influence of original variables on PC directions. The size of their influence depends on absolute values in the PC loading vector; mathematically, these values are calculated according to the following equation:

$$\alpha\left(\mathrm{x_j}\right)\ \cos\left(\widehat{\mathrm{x_j\ PC}_a}\right) \tag{4}$$

with $\alpha(x_j)$ being a correction term, proportional to the scale of each original variable, and $\cos\left(\widehat{\mathrm{x_j\ PC}_a}\right)$ the cosine of the angle between the *j-th* variable (x_j) and the *a*-th principal component (PC_a). If variables have been autoscaled, they all have the same scale, so $\alpha(x_j) = 1$. Therefore, in general, high absolute values in a PC loading vector for a given original variable correspond to high influence of the considered variable on the PC direction. Loading plots have very different shapes depending on variables number: with a limited number of original variables, the values of the loading vectors of the PCs can be reported as scatter plots (as for the score values, *see* Fig. 2-5). When the number of original variables is higher, for example when the dataset is constituted by signals (as for the food product optimization application in this work), loading plots can be obtained by reporting the loading vectors as a function of the original variables' domain, obtaining a signal-like chart. This representation allows for the quick identification of regions that are more relevant for the considered PC(s).

The information from loading plots and score plots can be combined in a **biplot**. In this visualization, both scores and loadings are represented in the same scatter plot, improving the interpretation of data structure.

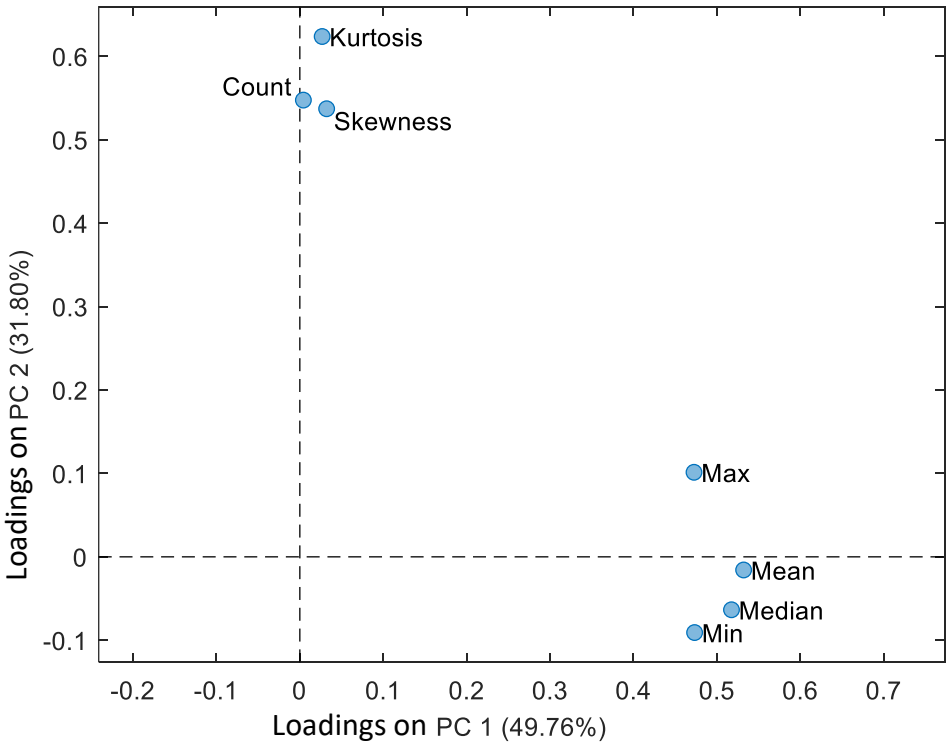


Figure 2-5 Example of a loading plot. It can be observed how two main clusters of variables are present, one at positive PC1 loading values (including Max, Mean, Median and Min) and the other at positive PC2 loading values (with Kurtosis, Count and Skewness).

The final aim of PCA is essentially to decompose the variance of the *X* data matrix in:

- one part describing objects variation (scores);
- one part describing variables variation (loadings);
- one part containing noise (residuals);

which can be expressed using matrix notation as follows:

$$X = T \cdot P + E \quad \textbf{(5)}$$

where:

- *X* is the mean centered data matrix with size {*n, m*} (*n* objects, *m* variables).
- *T* is the score matrix with size *{n, A}* (*n* objects, *A* PCs).
- *P* is the loading matrix with size *{A, m}* (*A* PCs, *m* variables).
- *E* is the residuals matrix with size *{n, m}.*

Starting from its mathematical definition, PCA works with continuous numeric variables and cannot handle categorical data directly. However, alternative PCA-based approaches have been developed and used in areas where it is necessary to explore categorical data with a multivariate approach (e.g., genetic data). One of these approaches, also applied in the food control application of this work, will be explained in 2.4.

2.2.2 Partial Least Squares-Discriminant Analysis (PLS-DA)

Partial Least Squares-Discriminant Analysis (PLS-DA) is an adaptation of the Partial Least Squares (PLS) algorithm, designed specifically for classification tasks (Barker & Rayens, 2003). This method operates by constructing a PLS regression model considering a "dummy" ***Y*** matrix, wherein binary column vectors encode the class membership of each sample. For a given class, the corresponding column of ***Y*** assigns a value of 1 to samples belonging to that class and 0 to those that do not. Once the PLS model is calculated, Bayesian statistics are applied to estimate the probability of class membership for each sample. Thresholds for these probabilities are then defined to determine class assignments, enabling the discrimination between samples belonging to different classes.

To assess the performances of a classification model like PLS-DA, several metrics can be derived from the confusion matrix (Ballabio & Consonni, 2013; Ballabio & Todeschini, 2009). This matrix represents the relationship between actual and predicted classes, with diagonal elements indicating correctly classified samples and off-diagonal elements representing misclassifications (Tab. 2-1).

		Predicted Class		
		Class 1	**Class 2**	**Class 3**
Actual Class	**Class 1**	70	2	3
	Class 2	0	50	0
	Class 3	8	0	22

Table 2-1 Example of a confusion matrix. Green diagonal cells are correctly classified samples. Red cells correspond to misclassified samples.

Classification metrics that can be calculated from the confusion matrix include:

- **Accuracy (ACC):** the overall percentage of correctly classified samples across all classes;
- **Sensitivity (SENS):** the percentage of samples correctly assigned to a specific class;
- **Specificity (SPEC):** the percentage of samples belonging to other classes correctly rejected by a specific class;
- **Efficiency (EFF):** the geometric mean of SENS and SPEC, offering a balanced view of classification performance;

While SENS, SPEC and EFF are class-specific metrics, ACC provides an overall evaluation of the classification model.

A critical step in both PLS and PLS-DA is determining the optimal number of *Latent Variables* (LVs) to include in the model. This implies capturing as much useful information as possible from the covariance between X and Y blocks while minimizing noise. The common approach for this is *Cross-Validation*. In *Cross-Validation,* the **training set** is divided into multiple deletion groups according to a pre-defined cross-validation scheme (e.g., *Leave One-Out*, *Venetian Blinds, Contiguous Blocks*) or a user-defined one. PLS-DA models are iteratively built by excluding one group at a time, and the excluded samples are used for class prediction. Classification performance metrics are calculated for each iteration. This process is repeated *n* times, with *n* = models with an increasing number of LVs, from 1 until a maximum number of LVs (LV_{MAX}) that depends on the size of the initial matrix ***X***. This approach allows for evaluating the variation in cross-validation error as a function of the number of LVs included in the model. The optimal model dimensionality is selected by minimizing the classification error in cross-validation. When classification performance values are very similar, the most parsimonious model (fewer LVs) is preferred.

Subsequent crucial step is validation of model predictive performance. This is assessed through *External Validation* using the **test set,** an independent set of samples. This set is built by either randomly selecting one-third of the initial available samples or considering measurements performed at a later stage than those in the training set. Whatever the data origin, these samples must not be used during model construction. The same metrics mentioned above can be calculated to assess the model classification performance in prediction.

PLS-DA is particularly powerful for analyzing complex datasets with high-dimensional features, enabling simultaneous consideration of numerous variables and their interactions. However, the success of this approach depends on careful selection of LVs, robust validation, and proper dataset preparation.

2.3 RGB Imaging for Food Analysis

RGB images (based on the RGB color model and named from the initials of the three additive primary colors: red, green, and blue) are characterized by some key features which make them an interesting tool to analyze food color:

- they can be acquired very quickly and using cheap instrumentations such as cameras, flatbed scanners, smartphones or webcams (Calvini et al., 2020);
- the photosensitive elements of these devices are designed to mimic the response of human vision to color, thus allowing to simulate human color perception;
- thanks to technology development, current systems can depict heterogeneous surfaces with a high degree of color accuracy and detail;
- RGB images can store a vast amount of information, since they allow us to measure the color properties of the analyzed sample for each individual pixel of the image.

An RGB image can be visualized as a grid of pixels, where each single pixel's color is determined by the combination of values from the three primary color channels (red, green and blue), each ranging from 0 to 255. In other words, a RGB image is like a superposition of three matrices, each corresponding to a specific color channel, with a unique value for each pixel (Fig. 2-6).

Values from 0 to 255 reflect the amount of light captured by device sensor in the red ($\lambda \approx 700$ nm), green ($\lambda \approx 546$ nm) and blue ($\lambda \approx 436$ nm) regions of the visible spectrum (Ohta & Robertson, 2005). Since light reflection and absorption phenomena can be linked to physicochemical sample properties, image pixel values are a powerful source of information. For this reason, most Multivariate Image Analysis (MIA) strategies employ a pixel-level approach, which consists in considering each single pixel of the analyzed image as a separate object (Calvini et al., 2020). This kind of approach can extract a lot of information on the subject under study, but it becomes computationally intensive when comparing multiple images at the pixel level. For example, a single RGB image with a size of 100 × 100 pixels generates 10000 objects, each characterized by 3 variables, for a total of 30000 data. This would not be a problem for modern computational power, however practical applications often require evaluating the variability of large sets of samples, which implies the need to consider large datasets of images. In these situations, hundreds or even thousands of images must be compared simultaneously, thus the analysis at the pixel-level would require strong computation efforts and may become impractical (Calvini et al., 2020). To overcome this issue, image-level data reduction methods such as the **colourgrams** approach can be applied. This technique converts each three-dimensional RGB image into a one-dimensional signal, called *colourgram* (Antonelli et al., 2004). This approach will be explained in detail in Section 2.3.1.

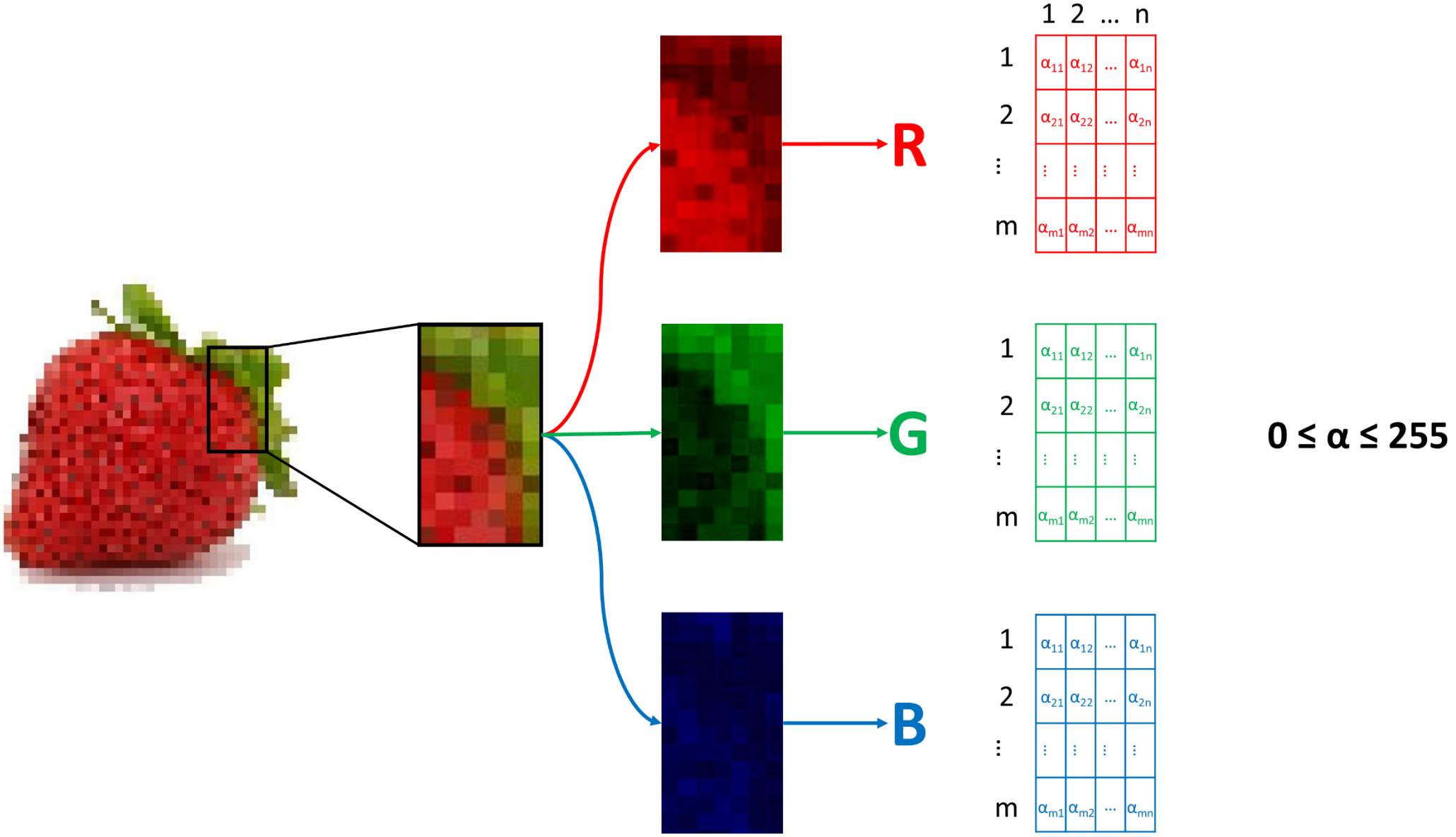


Figure 2-6 Representation of a digital image decomposition into three channel images, one in red (R), one in green (G) and one in blue (B). Each specific channel image can be seen then as a matrix where each α_{mn} value of the matrix codifies for a specific pixel value in the respective channel.

2.3.1 Multivariate Image Analysis

Multivariate Image Analysis (MIA) is a branch of multivariate analysis that applies multivariate algorithms, such as PCA, to extract, transform and analyze data from digital images. In this work, MIA was applied to the analysis of RGB images for a food color optimization study on strawberry yoghurt purée (SYP), following image-level approaches.

Indeed, when many images have to be simultaneously analyzed, data dimensionality reduction is generally performed for computational efficiency. Image-level data dimensionality reduction consists in converting each image into a feature vector acting like a fingerprint of the color-related properties of the corresponding image. Then, these feature vectors can be collected into a data matrix and elaborated using common chemometric methods. A straightforward method for data dimensionality reduction consists in computing mean or median values of R, G and B channels and/or other color-related parameters derived from RGB values; these features can be used to evaluate the color properties of the corresponding samples.

To account for spatial color variability within images, an alternative data dimensionality reduction method, i.e., the *colourgram* approach, has also been applied in the food color optimization study reported in this thesis (*see* Chapter 3). Technically speaking, *colourgrams* are obtained by merging in sequence the frequency distribution curves of the R, G and B channels of the original image, along with additional color parameters derived from the R, G and B values (Calvini et al., 2020). Such a transformation simplifies the analysis of a dataset of images, that can be converted into a matrix of signals, where each row corresponds to the information coming from a specific image.

The one-dimensional matrix of *colourgrams* can then be analyzed by means of multivariate techniques, to visualize the overall dataset structure, to highlight clusters of similar samples, to identify the presence of outliers, to calculate calibration or classification models, and to select the color features relevant to a specific problem. A major advantage of the *colourgrams* approach is its ability to provide a comprehensive investigation of all the color aspects of an image, without requiring any *a priori* assumption about the color parameters of interest for the problem under investigation. In this way, the best color features for classification and/or calibration purposes are found in real time, tailored to the specific problem as analyses proceed and tendencies emerge. In the literature, several *colourgrams* applications on food matrices can be found, like (Borin et al., 2007), (Foca et al., 2011), (Ulrici et al., 2012), (Giraudo et al., 2018), (Orlandi et al., 2019), (Menozzi et al., 2023). Despite food matrices differences, the steps to obtain *colourgrams* are the same, and they can be summarized as follows (Fig. 2-7):

1. At first, the RGB image, with size *{r, c, 3}* (where *r* is the number of pixel rows, *c* is the number of pixels columns and 3 corresponds to the R,G and B channels), is unfolded into a two-dimensional matrix with as many rows as the number of pixels ($r \times c$) and 3 columns corresponding to the R, G and B values (Calvini et al., 2020).
2. Afterwards, the unfolded RGB matrix is augmented by calculating additional color-related parameters for each pixel. These parameters include: lightness (L), calculated as the sum of R, G and B values; the relative colors (rR, rG and rB), calculated as the ratio between each RGB value and lightness; the hue (H), saturation (S) and intensity (I) values, obtained by converting the RGB color space into the HSI color space according to equations (6), (7) and (8) respectively. Furthermore, the score vectors derived from three PCA models calculated from the raw (SC1R, SC2R, SC3R), mean centered (SC1M, SC2M, SC3M) and autoscaled (SC1A, SC2A, SC3A) RGB data matrix, are also considered. This process results in a matrix containing 19 variables for each pixel.

$$H = \frac{1}{6} \times \begin{cases} \dfrac{G-B}{R-min(R,G,B)} & if\ R = max(R,G,B) \\ 2 + \dfrac{B-R}{G-min(R,G,B)} & if\ G = max(R,G,B) \\ 4 + \dfrac{R-G}{B-min(R,G,B)} & if\ B = max(R,G,B) \end{cases} \quad \textbf{(6)}$$

$$S = \frac{max(R,G,B) - min(R,G,B)}{max(R,G,B)} \quad \textbf{(7)}$$

$$I = max(R,G,B) \quad \textbf{(8)}$$

3. Then, for each variable, a 256 bins-long frequency distribution vector is calculated, and the 19 frequency distribution vectors are merged in sequence.
4. Finally, the loading vectors and the eigenvalues of the PCA models are added at the end of the signal.

This procedure converts the original RGB image into a one-dimensional signal with size equal to 4900 points, encoding the color content of the image (= 256 bins × 19 color related parameters + 3 PCA models × 3 PCs × (3 loading coefficients + 1 eigenvalue)).

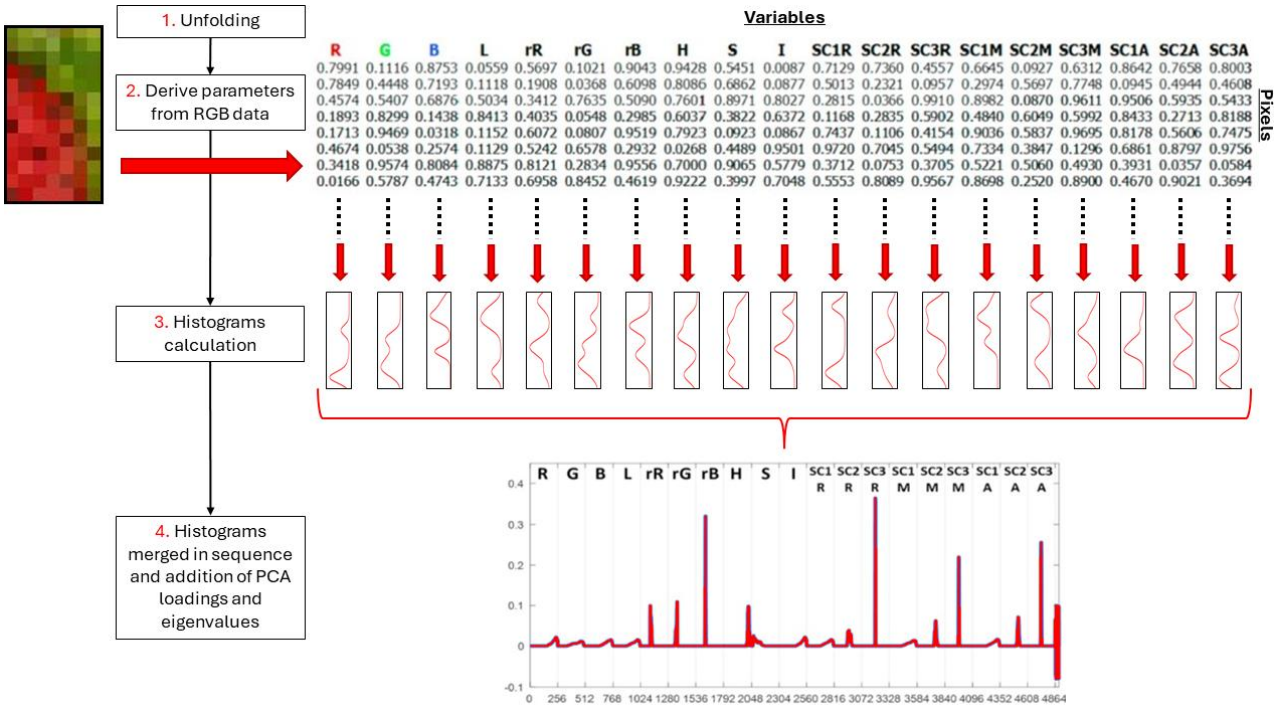


Figure 2-7 Scheme of the procedure to convert a digital image into a colourgram.

2.4 Graphical Representation of Categorical Data

Samples in a dataset can be described using different types of variables, such as numeric and/or categorical variables. Numeric variables represent quantitative measurements or quantities using numerical values, while categorical variables represent qualitative properties that group samples into distinct categories. More in detail, a *categorical variable* (also referred to as a *class set*) is a type of variable that takes on a limited number of distinct values, representing different categories or groups. These variables are typically non-numeric, although they can sometimes be encoded numerically in the form of zeros and ones for computational purposes. For example, a categorical variable might represent the "Type of Fruit" with categories such as "Apple" "Grapes" and "Orange".

Throughout the text, the terms *categorical variable* and *class set* are used interchangeably to describe a specific column in the dataset that holds categorical information. Within such a categorical variable, the distinct possible values it can take are commonly referred to as *classes* or *categories*. For example, in the "Type of Fruit" variable, the classes or categories would be "Apple", "Grapes", and "Orange".

To ensure clarity and consistency, it should be noted that the term *level* is also used synonymously with class or category in this text. For instance, when referring to the "Type of Fruit" variable, the individual levels of this variable are "Apple" "Grapes" and "Orange". These interchangeable terms—*class*, *category*, and *level*—are intended to capture the same concept: the distinct possible values within a categorical variable.

By explicitly defining and aligning these terms, this text aims to provide a coherent and intuitive framework for discussing categorical data and its representation in datasets. Indeed, the exploratory analysis of datasets containing both numeric and categorical variables is generally conducted by considering only the numeric variables as input of multivariate statistical methods such as PCA (*see* Section 2.2.1). Then, the categorical variables are used *a-posteriori* to verify the presence of clusters based on the classes of interest. Another possibility is to convert the categorical variables into binary variables with zeros and ones encoding for class belonging and analyze such variables together with the original numeric variables. However, in the Food Control case study considered in Chapter 4 the analyzed dataset consisted only of categorical variables, making exploratory methods like PCA are not suitable to be applied to the dataset as-is. For these reasons, we explored traditional approaches for the graphical representation of categorical data, such as treemaps (*see* Section 2.4.1) and alluvial plots (*see* Section 2.4.2), and we proposed an innovative approach based on Jaccard distance and PCA (*see* Section 2.4.3).

2.4.1 Treemap

Treemaps are versatile visualization tools that represent hierarchical or categorical data using nested rectangles, where the size and color of each rectangle correspond to specific numerical and categorical variables, respectively. This method is particularly effective for providing a comprehensive overview of relationships in datasets with multiple levels of hierarchy or grouped variables. To create treemaps, datasets must be arranged with a column storing counts of occurrences for each nested hierarchical combination that has to be displayed. Each rectangle represents a category or element within the dataset, with its area proportional to this quantitative measure, while colors can add a further dimension, often encoding categorical or additional numerical information. Treemaps enable the visualization of parent-child relationships, offering insights into the structure and proportions within the data (*see* Fig. 2-8).

Treemaps are particularly advantageous when combining size and color dimensions to display multiple variables simultaneously and conducting comparative analyses to highlight proportional differences between categories. However, challenges such as readability issues with small or tightly packed rectangles, overlapping

labels, and the need for effective color scaling require careful consideration. Furthermore, while treemaps can effectively handle one or two levels of nesting, introducing more than three nested levels often results in visualization difficulties. This is because the increasing subdivision of space leads to progressively smaller rectangles, making it harder to distinguish individual elements, interpret patterns, or allocate labels clearly. As a result, when dealing with highly complex hierarchies, alternative visualization techniques or simplified data representations may be more appropriate to maintain clarity and interpretability.

To conclude, while treemaps effectively visualize hierarchical and proportional relationships, they are less suited for tracking pathways or transitions over time, where alternative methods like alluvial plots (*see* Section 2.4.2) may be more appropriate. Nevertheless, the flexibility and depth of insight offered by treemaps make them a powerful tool for exploring and presenting data.

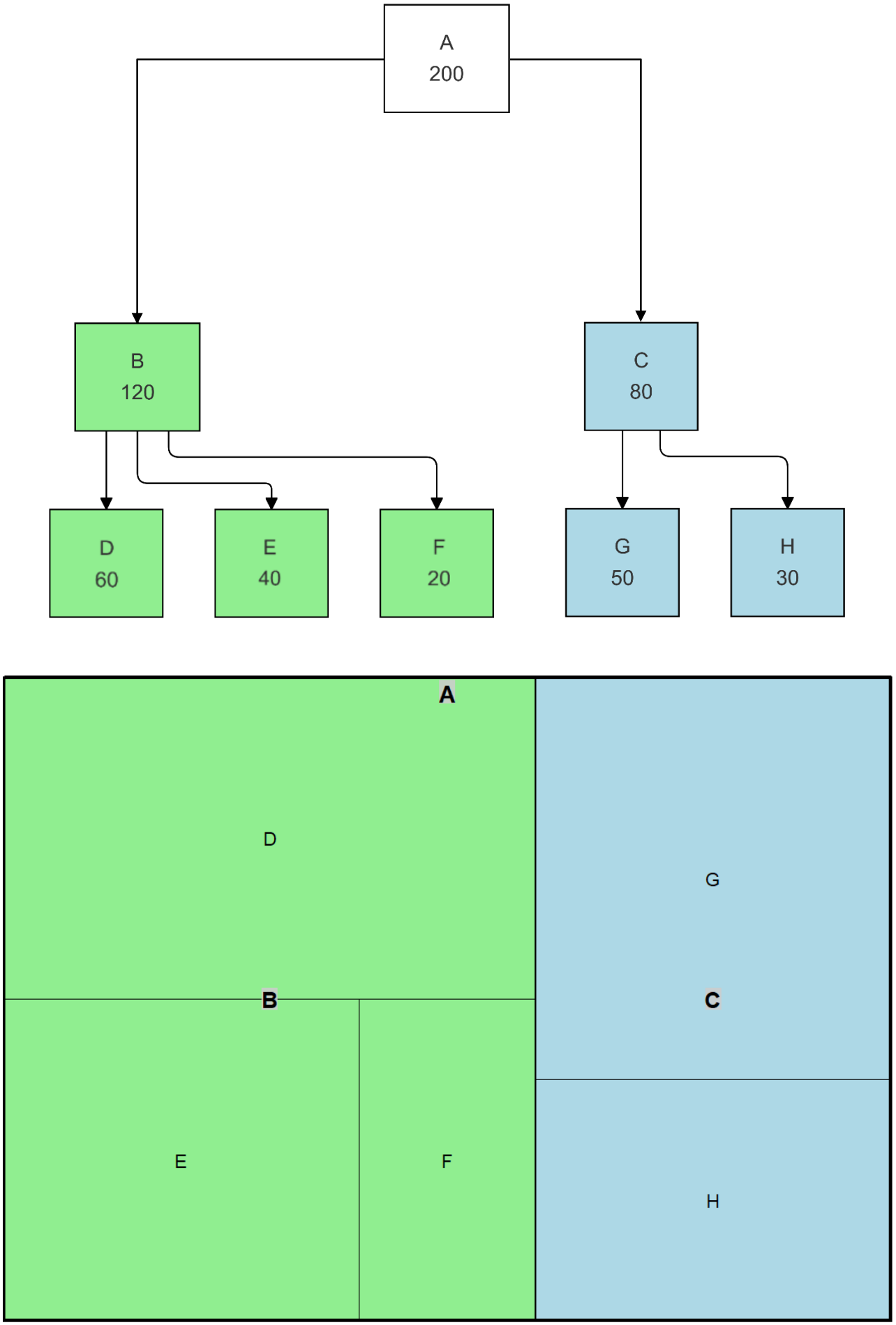


Figure 2-8 Schematic representation of an example treemap with three nested levels. The upper part of the figure reports a schematic representation of the dataset composed by 200 samples characterized by two levels of nested categorical variables, while the lower part of the figure reports the corresponding treemap.

2.4.2 Alluvial Plot

Alluvial plots are versatile visualization tools designed to depict the flow and transition of samples categories across multiple categorical variables, offering an effective way to analyze complex relationships in the dataset. These plots are particularly valuable for tracing connections, proportions, and distributions within grouped data, making them a key asset in exploratory data analysis across various scientific disciplines. Their structured design, characterized by components such as *axes*, *strata*, *flows*, *alluvia*, and *lodes*, allows them to capture both local and global relationships within datasets (Fig. 2-9).

Axes serve as vertical partitions in an alluvial plot, representing discrete states, variables, or stages within the dataset. In other words, axes represent the categorical variables used to describe the samples. Along each axis, *strata* represent the individual categories or classes within the corresponding categorical variable. *Flows*, represented by ribbons of varying width, connect strata across adjacent axes, illustrating the movement or transitions of categories. An *alluvium* represents the continuous movement of a single entity (sample) across all axes, providing a global perspective on its progression through the dataset. *Lodes*, on the other hand, denote specific intersections where an alluvium crosses a *stratum*, offering a localized view of how individual trajectories contribute to or interact with particular categories.

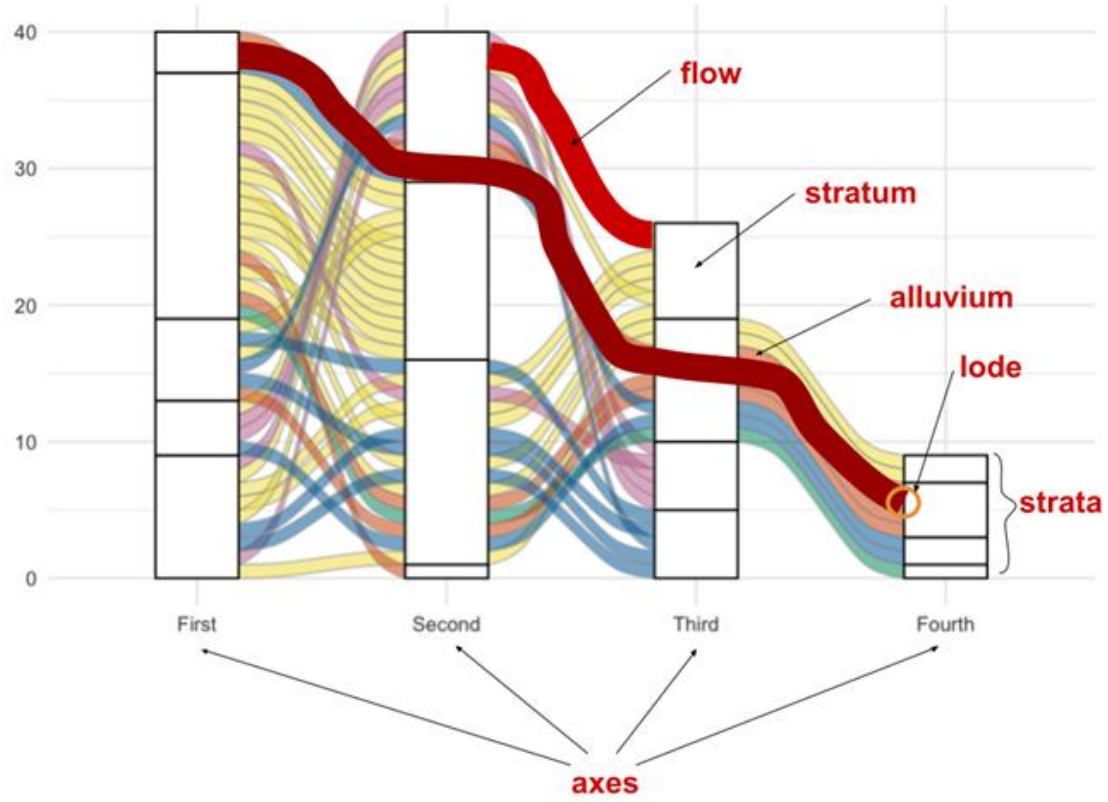


Figure 2-9 Schematic representation of an alluvial plot with main terminology (R Consortium, 2021; Robbins, 2021).

The design of alluvial plots enables them to illustrate relationships across hierarchical, temporal, or categorical datasets effectively. They can highlight how categories are redistributed between stages, reveal the relative size of flows, and uncover transition patterns. For instance, in a temporal dataset, an alluvial plot might show how individuals shift between demographic groups over time, while in a hierarchical dataset, it could visualize how classifications branch into smaller categories.

One of the key advantages of alluvial plots is their ability to condense high-dimensional data into an accessible, intuitive format. By aligning flows with their associated categories and scaling them proportionally, these plots provide an immediate understanding of complex data structures. This is particularly useful for datasets with numerous variables or classifications, where traditional tables or simpler charts may obscure critical insights. However, as the number of axes, strata, and flows increases, the visualization can become cluttered, with overlapping pathways and label congestion reducing readability. Effective customization is essential to mitigate these challenges and preserve interpretability.

While alluvial plots excel at summarizing complex relationships, their clarity can diminish when applied to highly intricate datasets with many axes and strata. Overlapping flows, dense connections, and label congestion are common challenges. To address these, it is advisable to limit the number of axes and strata, strategically use color and spacing to highlight key patterns, and prioritize clear labeling of critical components.

In summary, alluvial plots are powerful tools for visualizing flows and transitions within data categories. Their adaptability makes them suitable for a wide range of applications, from tracking categorical changes to analyzing hierarchical and temporal data. Despite the challenges posed by complexity, careful design and customization allow alluvial plots to remain a good exploratory data analysis tool, offering users a detailed yet comprehensible view of their data.

2.4.3 PCA on Jaccard Distance Matrix

The multivariate exploratory capabilities of PCA are inherently limited when applied to categorical data, even when such data are transformed into a binary format of zeros and ones. This limitation arises because PCA is fundamentally designed to work with continuous numerical variables, which are necessary for computing the covariance matrix and performing eigenvalue decomposition. As a result, PCA becomes less effective for datasets that include many categorical features or that are entirely composed by categorical variables, such as the one used in this work for food control application (*see* Chapter 4). By comparing a classical PCA score plot (*see* Fig. 2-4) with a treemap or alluvial plot, the potential of using a PCA-like representation for categorical data becomes evident due to the following advantages:

i. Unlike treemaps or alluvial plots, which rely on graphical constructs such as rectangle areas or strata to represent groupings, PCA's native scatter-plot visualization directly represents each sample as a point in the principal components space. This maintains a high level of detail at the individual sample level, ensuring clarity and precision in the representation (except in cases of dense samples overlap).
ii. Multivariate relationships are typically captured within 2–4 principal components, offering a concise representation of complex data. In contrast, treemaps require additional nested levels, and alluvial plots require more axes to accommodate multivariate relationships, which can complicate visualization and interpretation.

This highlights the need to develop PCA-based methods to explore and display categorical data. The conceptual approach proposed in this work draws inspiration from fields such as bioinformatics and computational biology, where PCA is extensively used to explore genetic diversity and evolutionary relationships by analyzing sequence data, including DNA sequences or protein amino acid chains (Konishi et al., 2019). In these applications, sequence data—represented as strings encoding nucleotide bases or amino acids (e.g., GATCAGGTC...)—are transformed into distance matrices (using specific distance-similarity mathematical definitions), which are then used as input for PCA to produce exploratory visualizations. Following an analogous approach, the algorithm proposed in this thesis to represent datasets composed of

categorical variables is based on calculating distances metrics (i.e., continuous numeric values usually ranging from 0 to 1) representing the (dis)similarity between pair of observations of the original dataset. This can be achieved through the following steps:

1. **Encoding of each categorical variable into binary class vectors with 0-1 values.** Table 2-2 reports an example of a small categorical dataset with 5 samples (A, B, C, D and E) and 3 categorical variables: Fruit with 3 classes (Orange, Apple and Grapes), Process with 2 classes (Juice Extraction and Cooking) and Product with 2 classes (Juice and Jam). Each categorical variable is converted into a binary matrix with as many columns as the number of classes. Class belonging is encoded in binary, with 1s and 0s indicating whether a specific sample belongs to a class or not, respectively. Table 2-3 reports the outcome of this procedure starting from the example matrix in Table 2-2.

		Categorical Variables		
		Fruit	**Process**	**Product**
Samples	**A**	Orange	Juice Extraction	Juice
	B	Orange	Cooking	Jam
	C	Apple	Juice Extraction	Juice
	D	Grapes	Juice Extraction	Juice
	E	Apple	Cooking	Jam

Table 2-2 Example representation of a dataset with categorical features (Fruit, Process, Product) with different class values for each observation.

		Encoded Categorical Variables						
		Fruit			**Process**		**Product**	
		Orange	**Apple**	**Grapes**	**Juice Extraction**	**Cooking**	**Juice**	**Jam**
Samples	**A**	1	0	0	1	0	1	0
	B	1	0	0	0	1	0	1
	C	0	1	0	1	0	1	0
	D	0	0	1	1	0	1	0
	E	0	1	0	0	1	0	1

Table 2-3 Encoding of categorical variables in Tab. 2-2 into binary class vectors (0 = not belonging, 1 = belonging).

2. **Calculation of *Jaccard* similarity coefficient.** When dealing with binary data, a wide variety of distance-similarity metrics are available. According to Todeschini and coauthors (2012, 2020), around 50 similarity coefficients may be relevant for chemometric or cheminformatic applications. For the method developed in this work, the *Jaccard* similarity coefficient (also known as *Jaccard-Tanimoto* similarity coefficient) was selected, as it is one of the most commonly used similarity indices for binary data. Given the two A and B row vectors from Tab. 2-3, *Jaccard* similarity coefficient, also referred to as *Intersection over Union* (IoU), is defined as:

$$S_{A,B}^{\mathrm{JT}} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{1}{5} = 0.2 \qquad \textbf{(9)}$$

where $|A \cap B|$ represents the intersection of A and B sample vectors, i.e., the number of common entries with a value equal to one, while $|A \cup B|$ corresponds to the union of A and B vectors, i.e., the overall number of ones in either row vectors. For example, considering A and B row vectors in Table 2-3, there is only one occurrence of ones for both vectors, corresponding to Orange column. This corresponds to the intersection of A and B vectors. The total number of entries equal to 1 of either vectors is equal to 5, corresponding to union of the two vectors. Therefore, the *Jaccard* similarity index for A and B samples is equal to 1/5, corresponding to 0.2 (*see* Eq. 9). *Jaccard* similarity coefficient is classified as an asymmetric similarity measure because it does not consider the absence of the feature in both vectors. This asymmetry makes it particularly suited for datasets where the presence of certain features carries more significance than their mutual absence. The *Jaccard* index is commonly referred to simply as *Jaccard* index and this term will be used in this thesis.

3. **Conversion of Jaccard similarity index into Jaccard distance.** The *Jaccard* similarity coefficients obtained in Step 2 are converted into distances values between pairs of observations in the original dataset. This is done, in the case of *Jaccard* similarity, thanks to the following relation between similarity and distance measures:

$$D_{A,B}^{\mathrm{JT}} = 1 - S_{A,B}^{\mathrm{JT}} = 1 - 0.2 = 0.8 \qquad \textbf{(10)}$$

where $D_{A,B}^{\mathrm{JT}}$ is the Jaccard distance between A and B samples, while $S_{A,B}^{\mathrm{JT}}$ is the corresponding Jaccard similarity coefficient. Considering the example of A and B sample vectors in Table 2-3, their Jaccard distance is equal to 0.8, corresponding to 1 – 0.2. Steps 1 – 3 convert the *n,m* data matrix (with *n* = *rows/observations*, and *m=columns/categorical variables*) into an *n,n* matrix, where diagonal elements reflect to the distance each observation from itself, while off-diagonal elements represent the distances between pairs of observations in the original data matrix.

Samples

Samples	**A**	**B**	**C**	**D**	**E**
A	0	0.8	0.5	0.5	1
B	0.8	0	1	1	0.5
C	0.5	1	0	0.5	0.8
D	0.5	1	0.5	0	1
E	1	0.5	0.8	1	0

Table 2-4 Jaccard distance matrix derived from encoded matrix in Tab. 2-3.

Therefore, Jaccard distances are continuous numerical values ranging from 0 to 1. Furthermore, as noted by Todeschini et al. (2020), it is interesting to note that Jaccard distance is intrinsically related to Euclidean distance since the square form of Jaccard can be viewed as normalized squared Euclidean distance.

4. **PCA analysis on Jaccard distance matrix.** The Jaccard distance matrix represented in Tab. 2-4 is analyzed by PCA to perform a multivariate data exploration of the relationships existing between the samples in the multidimensional features space described by the original categorical variables. This can be simply achieved by the observation of the corresponding score plots (Fig. 2-10).

Methodology described in Steps 1 – 4 will be referred to as *J-PCA* from this moment onwards.

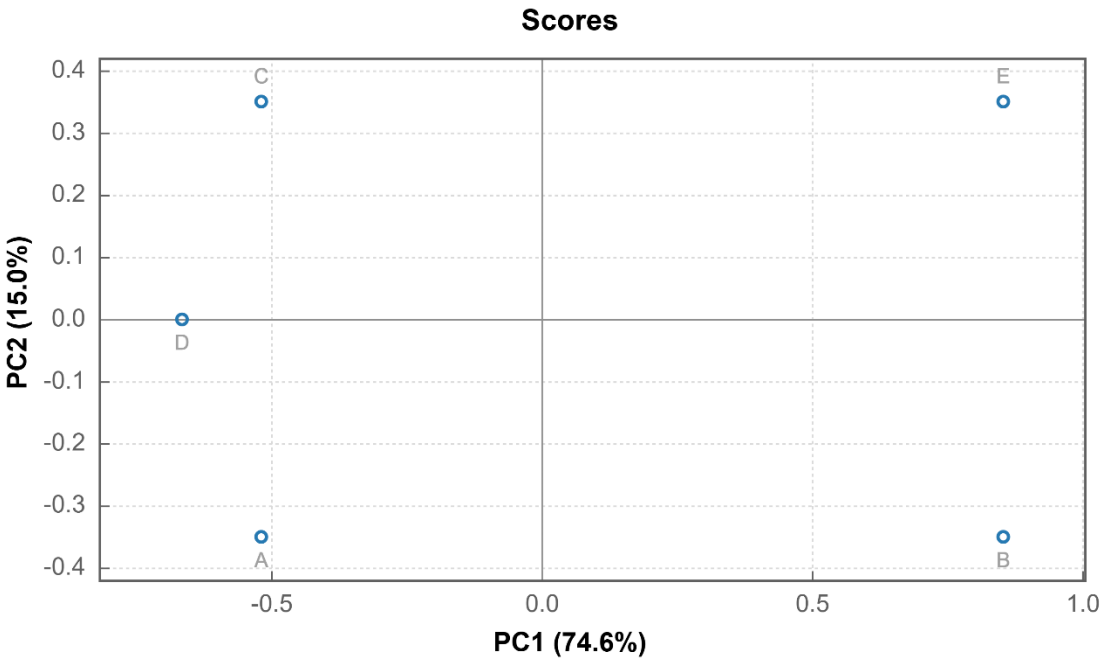


Figure 2-10 PCA performed on Jaccard distance matrix in Tab. 2-4 (J-PCA).

Chapter 3: Food Optimization

What follows is the integral content of Rolando, P. L., Calvini, R., Foca, G., & Ulrici, A. (2023). Mixture design and multivariate image analysis to monitor the colour of strawberry yoghurt purée. *Microchemical Journal*, 194, 109222. https://doi.org/10.1016/j.microc.2023.109222.

Abstract

Food color is a commercial added value, since it represents the first appealing factor for consumers. In this context, this study was aimed at evaluating the effect of strawberry yoghurt purée (SYP) formulation on the corresponding color and on its variation over time, which is mainly due to degradation and browning phenomena. To this aim, a combined approach was used that included mixture design and multivariate analysis of RGB images. Strawberry purée, sugar, lemon juice and two types of thickener were mixed in different proportions by I-optimal mixture design to obtain 44 SYP formulations. The samples were subjected to light and temperature stress conditions for five weeks; during this time the RGB images of the samples were acquired using a flatbed scanner, along with the images of the corresponding control samples. The dimensionality of the acquired images was reduced by two different approaches: i) the conversion of images into signals, namely *colourgrams*, which can be seen as the color fingerprint of the imaged samples, and ii) the calculation of the median values of various color-related parameters. The *colourgrams* dataset was then subjected to exploratory data analysis using Principal Component Analysis, while the median values of color-related parameters were analyzed using Response Surface Methodology and Partial Least Squares-Discriminant Analysis. The aim of data analysis was both to find the best color parameters to describe color variability over time, and to investigate the cause-effect relationship between mixture proportions and color response. The results highlighted that, among the considered color parameters, relative green (i.e., the ratio of green to lightness) and red could be used to monitor color changes. Color variation due to stress conditions was more pronounced for samples with a high percentage of strawberry purée, and the type of thickener also affected the color degradation kinetics.

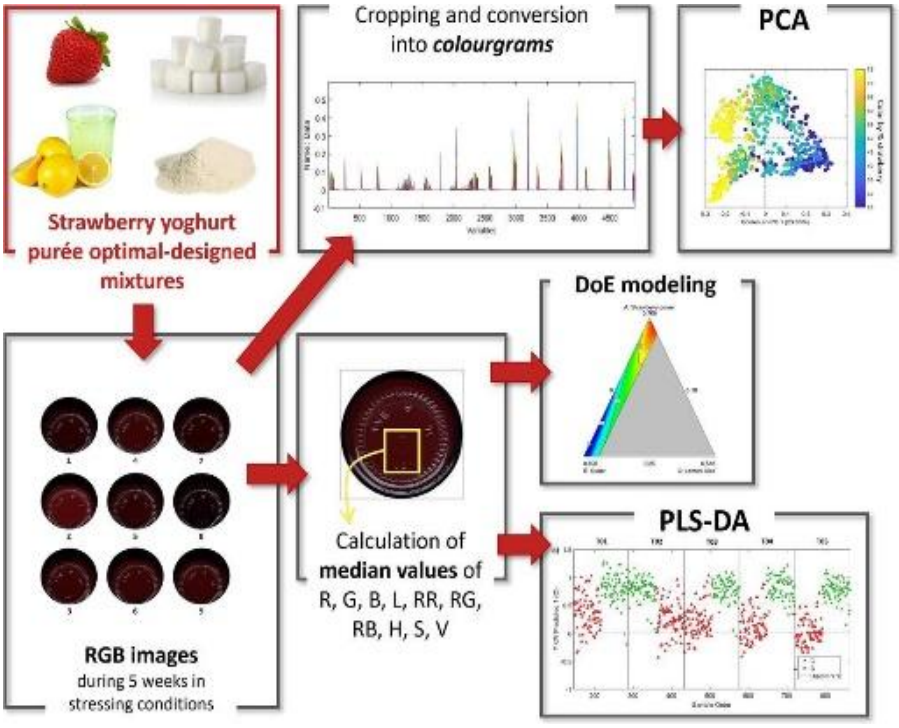


Figure 3-1 Graphical abstract representing all steps of analysis conducted in Rolando et al., 2023.

3.1 Introduction

Strawberry yoghurt purée (SYP) is a specific formulation – composed of strawberry, sucrose, lemon juice, thickener, natural strawberry flavor, and water – to be used as a semi-finished product in the manufacturing of flavored yoghurts. In particular, SYP is added during yoghurt production in a percentage ranging from 15% to 20% w/w of the entire end product.

In order to satisfy consumers' expectations, a good SYP formulation has to transmit to the final product a balanced and recognizable strawberry flavor/taste and a pronounced and pleasant strawberry-like red color. Focusing on composition variables only, the optimal taste strongly depends on the right balance between strawberry and sugar content in the formulation, as well as on the addition of strawberry food flavoring. On the other hand, color dependencies are way more complex to foresee and understand: multiple interactions between all the components may come into play for the determination of an optimal color. Furthermore, the color is not always stable over time, but it is subjected to changes that depend on different factors. In this context, a topic of great concern for SYP manufacturers is the identification of the optimal formulation that ensures the desired color appearance and its stability over time, since these two properties can greatly affect purchases of the final product.

Anthocyanins are the natural pigments which are mainly responsible for strawberry color. Despite many studies showed a notable variability on anthocyanins profile based on strawberry cultivar and maturation stage (Da Silva et al., 2007; Dzhanfezova et al., 2020) scientific literature agrees on pelargonidin-3-glucoside as the most abundant compound, which ranges between 50% and 90% of total anthocyanins content (Bursać Kovačević et al., 2015; Da Silva et al., 2007; Dzhanfezova et al., 2020; Ertan et al., 2020). Color properties of these compounds originate from their resonating structures, which also confer them an intrinsic instability (Delgado-Vargas et al., 2000) that must be taken into consideration when trying to optimize and control the color of food products containing anthocyanins. Many factors have a great influence on their color and stability, including pH, temperature, light exposure, time, oxidase enzyme activity, water activity and total soluble content, among others.

While pH-dependent structural changes have been largely documented and studied (Delgado-Vargas et al., 2000), the impact of temperature as the second most crucial parameter on anthocyanin stability is not sufficiently understood from a chemical point of view (Sadilova et al., 2007). However, it is well known that anthocyanins are thermolabile, leading to the so-called browning effects. These phenomena are reported in two different studies, in which pilot productions of strawberry jams (Martinsen et al., 2020) and juices (Hartmann et al., 2008) were put under observation to evaluate which process caused higher losses in anthocyanins content. In addition, Martinsen et al., 2020 compared different levels of thermal treatment and observed a first order reaction kinetics. Browning processes also occur when strawberries are mashed into pulp, due to the effect of oxidizing enzymes such as polyphenol oxidase. In this case, a balanced blanching operation can inactivate these enzymes and prevent undesired color variations in the final product (Sulaiman et Silva, 2013).

Given the complex relationships between the various factors underlying SYP color and its stability, in this work we have focused on the effect of compositional factors. A quaternary mixture design was elaborated, considering the I-optimality criterion (Goos et al., 2016) for the selection of the experimental conditions, i.e., of the recipes. This allowed to study the color characteristics of SYP according to the recipe used for its production and how the color changes over time under controlled storage conditions. Indeed, ingredients proportions in the recipe influence the properties of SYP, such as final pH of the solution, total soluble solids content, water activity (Holzwarth, Korhummel, et al., 2013, where the authors also studied the effect of

different thickeners), and citric acid content (Holzwarth, Wittig, et al., 2013) among others. In turn these properties affect anthocyanins stability and therefore the final color of the product.

In order to accelerate degradation and browning phenomena we considered extreme stress conditions, i.e., continuous exposure to intense light at a temperature of about 35 °C. Any other factor potentially affecting anthocyanins stability was kept constant.

To evaluate the color of the SYP samples and its variation over time, we considered Red-Green-Blue (RGB) imaging due to its several advantages. Indeed, this tool allows to objectively assess color related information about the investigated samples using affordable devices in a fast and non-destructive manner, resulting also a green approach.

Compared to traditional tristimulus colorimeters, RGB images can be acquired by commonly used devices such as digital cameras, smartphones or flatbed scanners, allowing also to evaluate color variability within the sample surface (Wu & Sun, 2013). RGB images are complex data arrays, and it is of utmost importance to apply proper image analysis strategies to extract the useful information from such data (Prats-Montalbán et al., 2011).

When multiple images have to be analyzed altogether, data-dimensionality reduction is generally performed, which consists in converting each image into a feature vector acting like a fingerprint of the sample in terms of color. Then, the feature vectors obtained from each acquired image can be collected into a data matrix and elaborated by multivariate data analysis to extract the useful information and relevant color characteristics of the images. As a very straightforward method for data-dimensionality reduction, it is possible to calculate average or median values of R, G and B values and/or other color parameters derived from RGB data, using then these descriptors to evaluate the color properties of the considered samples (Santos et al., 2012), (Solana-Altabella et al., 2018), (Pagnin et al., 2020). In order to preserve also the information related to color variability within the images, some of us proposed an alternative approach for data dimensionality reduction, which consists in calculating the frequency distribution curves of a set of color descriptors and merging them in sequence to obtain a signal named *colourgram* (Antonelli et al., 2004).

The main idea behind the present study is to combine mixture design and multivariate RGB image analysis to evaluate how color properties of SYP are affected by compositional factors and stress conditions. More in detail, a mixture domain of the semi-finished strawberry product was defined considering four ingredients, i.e., strawberry purée, sugar, lemon juice and thickener. Within this domain, 44 different mixtures selected using the I-optimality criterion were prototyped and placed under observation over a five-week period, during which RGB images of the samples put under stress conditions and of the corresponding control samples were acquired every week using a flatbed scanner. The best parameters to describe color variability were investigated by multivariate image analysis, also trying to establish a cause-effect relationship between mixture proportions and color response.

3.2 Materials and Methods

3.2.1 Mixture Design for Strawberry Yoghurt Samples Preparation

In order to model the variation of different color-related properties, strawberry yoghurt purée samples preparation was carried out following the design of experiments (DoE) approach by using the Design Expert ver. 10 software (Stat-Ease Inc., USA). A double-constrained quaternary mixture design was applied, in which components variation ranges were derived from initial business recipe based on a number of considerations, including commercial interest for a strawberry formulation of up to 75% w/w, company know-how on the feasibility of some mixtures, legal/technical limits for some ingredients, such as the thickener, and finally the need to obtain a fairly pronounced color variation for better modelling. The resulting ranges for each component are summarized in Table 3-1; the total sum of the four ingredients percentage was equal to 95% w/w, since a constant quantity of 5% w/w of water was added to each mixture. As it can be noticed, natural strawberry flavor was not added to samples, due to its small amount in the recipe (<0.1 % w/w) and its null effect on mixture color; furthermore, no one of the above referred publications listed flavoring as a potential influencing factor.

Lower Limit	**Component**			**Upper Limit**
0.25	≤	**A**: Strawberry purée	≤	0.75
0.18	≤	**B**: Sugar	≤	0.696
0	≤	**C**: Lemon juice	≤	0.06
0.004	≤	**D**: Thickener	≤	0.01
		A+B+C+D	=	0.95

Table 3-1 Lower and upper limits for mixture components expressed as weight fractions (w/w); for each mixture the sum of the four components is equal to 95% w/w, since a fixed quantity of water (5% w/w) was added.

Table 3-2 shows the design matrix consisting of 22 experiments, chosen based on the I-optimality criterion to fit the following quadratic polynomial model:

$$y = \sum_{i=1}^{m} b_i \; x_i + \sum_{i=1}^{m} \sum_{j>i}^{m} b_{ij} \; x_i \; x_j \qquad \textbf{(11)}$$

where *y* is the color parameter to be modelled, *m* is the number of mixture components, b_i is the one-factor term of the *i*-th component with proportion x_i and b_{ij} is the interaction term for components *i* and *j* (with $i \neq j$).

All the mixtures resulting from DoE were prepared with two different thickeners, pectin (PEC) and locust bean gum (LBG), following the random order reported in Table 3-2 and obtaining a total of 44 SYP samples (=22 experiments × 2 thickener types).

Each mixture was prepared in a 1 kg batch using a Bimby TM21 robot (Vorwerk & Co. KG, Wuppertal, Germany). The complete cooking cycle for each mixture took in general between 20 and 25 min, considering ingredients loading, temperature rising till 93 °C and thermal treatment at 93 °C ± 5 °C for 15 min. A unique

batch of strawberry purée was picked from company's stock. It was obtained from strawberries of a single cultivar, which were subjected to a blanching process (78 °C for 3 min) and then mashed. This batch of strawberry purée was used for the preparation of all the samples in combination with other ingredients. After the cooking process, each mixture was divided into 6 aliquots of about 165 g and the aliquots were transferred into 105 ml jars, subsequently vacuum-sealed and labelled. The jars were filled until the edge with the mixtures still hot; in this manner no significant headspace was left for air oxygen, which is a factor that could affect the color stability of the samples. Two jars of each mixture were used for the experimental part of this study, while the remainder jars served as backup.

Run order	PEC samples name	LBG samples name	A: Strawberry Purée (%)	B: Sugar (%)	C: Lemon Juice (%)	D: Thickener (%)
1	PEC1	LBG1	70.60	18.00	6.00	0.40
2	PEC2	LBG2	54.12	37.76	2.72	0.40
3	PEC3	LBG3	25.00	69.60	0.00	0.40
4	PEC4	LBG4	72.50	18.00	3.50	1.00
5	PEC5	LBG5	25.00	63.60	6.00	0.40
6	PEC6	LBG6	75.00	19.40	0.00	0.60
7	PEC7	LBG7	39.53	53.59	1.33	0.55
8	PEC8	LBG8	54.12	37.76	2.72	0.40
9	PEC9	LBG9	55.20	33.00	6.00	0.80
10	PEC10	LBG10	75.00	19.40	0.00	0.60
11	PEC11	LBG11	25.00	67.20	2.00	0.80
12	PEC12	LBG12	58.33	35.67	0.00	1.00
13	PEC13	LBG13	25.00	63.00	6.00	1.00
14	PEC14	LBG14	72.50	18.00	3.50	1.00
15	PEC15	LBG15	54.06	37.58	2.66	0.70
16	PEC16	LBG16	50.00	44.00	0.00	1.00
17	PEC17	LBG17	40.00	48.00	6.00	1.00
18	PEC18	LBG18	47.65	40.65	6.00	0.70
19	PEC19	LBG19	64.53	28.59	1.33	0.55
20	PEC20	LBG20	54.06	37.58	2.66	0.70
21	PEC21	LBG21	25.00	67.20	2.00	0.80
22	PEC22	LBG22	54.06	37.58	2.66	0.70

Table 3-2 Design matrix for both PEC and LBG series of samples.

3.2.2 Experimental Set-Up for Stressing the Samples

The 44 mixtures were randomly divided into 4 blocks of 11 samples each, paying attention to distribute as much uniformly as possible among the different blocks the various mixture compositions.

In order to accelerate oxidative phenomena that normally could take several months to occur, one jar for each one of the 44 mixtures (labelled as "S" series) was stored under stress conditions for approximately five weeks. More in detail, the jars were flipped over and stored in a closed polystyrene box containing a 50 W lamp (2 bulbs, full spectrum 380–720 nm). In this manner, the bottom of the jars was exposed to light for the entire period of five weeks (690 h of light exposure overall). In order to minimize the effect of heterogeneous lighting conditions inside the box, the internal area of the polystyrene box was divided in four areas, one for each block of samples, and once a week the jars were moved to the adjacent area (Figure 3-2 on next page).

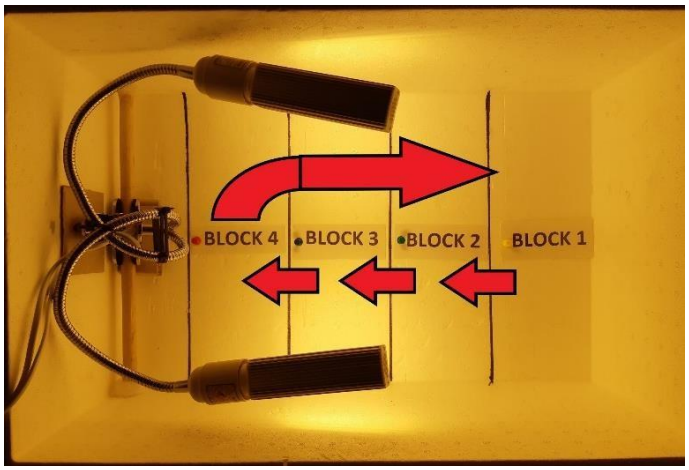


Figure 3-2 Polystyrene chamber used to put S samples under light and temperature stress conditions; the red arrows indicate how the blocks of jars were moved during the experimentation.

During this procedure each sample was inspected and shaken in order to prevent the possible formation of different density layers, even if this effect has never been observed over the five weeks of the experiment.

Light was not the only stress factor inside the box. In fact, due to the Joule effect, lamp bulbs warmed up the environment at a temperature of 34.5 °C ± 0.9 °C. Temperature was monitored with an Inkbird ITC-308 digital thermal regulator. The combination of light and temperature stress factors visibly accelerated oxidative reactions in samples, leading red anthocyanins towards more brownish compounds.

Simultaneously, the second jar of each one of the 44 mixtures was used as control sample (labelled as “C” series) and kept in the dark at a refrigerated temperature between 0 °C and 4 °C for the same period of five weeks.

3.2.3 Image Acquisition System

The evolution of samples color during time in both stressing and controlled conditions was monitored weekly using an image acquisition system. The system was composed by a flatbed scanner (Epson Perfection V39), a white cardboard with 9 slots for the jars containing the mixtures, a carton box to cover the flatbed scanner during image acquisition and a computer to save and store images for further elaboration. Fig. 3-3 shows an example of an image acquired with this system.

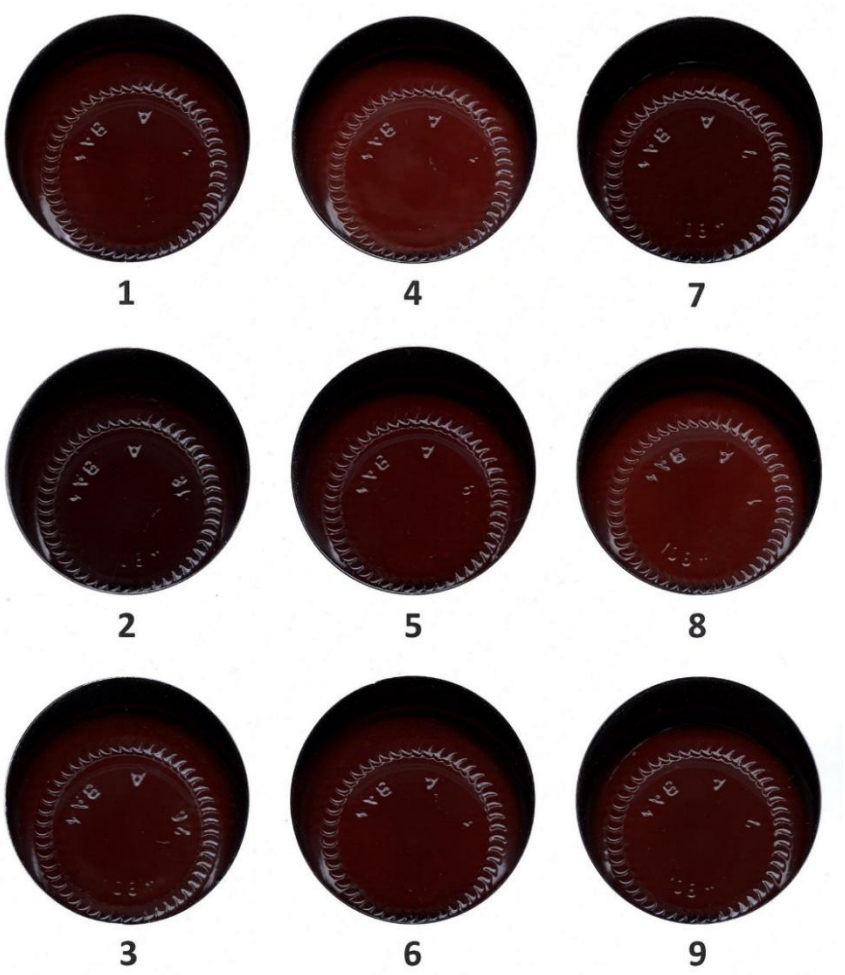


Figure 3-3 Example of an image acquired with the flatbed scanner on a set of 9 samples. Following the numerical order (from 1 to 9) the samples are: PEC8, PEC7, LBG2, LBG1, LBG22, PEC18, LBG16, PEC6 and LBG12. All the samples belong to the stressed (S) series acquired at T04.

Once a week for five weeks, both the S and the C samples series were taken from their respective storage places for images acquisition. Since each block was composed of 11 samples but only 9 slots were available in the white cardboard mask, two scans for each block were necessary. For this reason, the 11 samples of each block were divided into two sub-blocks, one with 5 samples and the other one with the remaining 6 samples. The two scans for each block were acquired as follows:

- the first scan contained the 5 samples of the first sub-block and 4 samples randomly selected from the second sub-block;
- the second scan contained the 6 samples of the second sub-block and 3 samples randomly selected from the first sub-block.

The position of the samples in the image scene was randomized for each image acquisition, as well as the subdivision of the samples of each block into the two sub-blocks.

Images of SYP mixtures were acquired at 6 different acquisition times from T00 to T05, where T00 corresponds to the day in which the mixtures were prepared and T01-T05 correspond to the weekly acquisitions performed from the first until the fifth storage weeks. Therefore, the final images dataset was composed of 96 images (=2 scans × 4 blocks × 2 series × 6 acquisition times).

For all the images, the following acquisition parameters were applied directly from Epson scanner software (EPSON Scan Ver. 3.9.4.7IT) and kept constant for all acquisition times: lightness + 20, contrast + 20, saturation + 60.

All the images were saved in .JPG format, with 24-bit depth and a spatial resolution of 9359 row pixels × 6800 column pixels. Considering a file size approximately equal to 191 MB per image, the overall dataset size was about 18 GB.

3.2.4 Data Dimensionality Reduction of RGB Images Dataset

Before further analysis, additional image pre-processing steps were necessary in order to crop the image of each single sample. These operations were automatically carried out by means of an image cropping algorithm written in MATLAB language (ver. 9.3, The Mathworks Inc., USA), obtaining on the whole 864 (=96 × 9) images of single samples.

This large dataset of images needed to be managed through proper data dimensionality reduction methods in order to retrieve useful information. In this study two different data dimensionality reduction approaches were considered. Initially, the *colourgrams* approach was used as a first strategy to gain a preliminary overview of the entire dataset of images and evaluate differences over time between stressed and control samples. *Colourgrams*, proposed the first time by Antonelli et al., 2004 and later used in other successful applications (Borin et al., 2007; Caramês et al., 2021; Foca et al., 2011; Giraudo et al., 2018; Liu et al., 2021; Masino et al., 2008; Orlandi, Calvini, Foca, et al., 2018; Orlandi, Calvini, Pigani, et al., 2018; Ulrici et al., 2012), allow to codify the color information contained in the three-dimensional array of each RGB image by reducing it to a signal (vector). Basically, the frequency distribution curves of the R, G and B channels of the original image and of other parameters derived from the values of R, G and B are obtained; then, the *colourgram* is built by merging these frequency distribution curves in sequence. In addition to R, G and B, the color parameters considered for *colourgrams* computation include lightness (L = R + G + B), relative red (RR = R / L), relative green (RG = G / L), relative blue (RB = B / L), hue (H), saturation (S), intensity (I) and the three score vectors of each one of three PCA models calculated on raw, mean centered and autoscaled RGB data, respectively. For further details about *colourgrams* computation and the considered color parameters, the reader is referred to (Antonelli et al., 2004) and (Calvini et al., 2020)

Such a transformation simplifies the analysis of a dataset of images that can be converted into a matrix of signals, where each row corresponds to a signal codifying the color properties of a specific image in the dataset. Furthermore, the inclusion in the *colourgram* of more color parameters in addition to the RGB channels allows to simultaneously evaluate different color-related features of the imaged samples and to identify which of them are the more relevant for the considered application.

An easy-to-use graphical user interface developed by some of the authors (Calvini et al., 2020) for the creation of *colourgrams* is freely downloadable from the web (*Colourgrams GUI – DOWNLOADS – Chimslab*). In this case, before calculating the *colourgrams*, background removal was performed by selecting a threshold limit: only the pixels with values of the blue channel < 200 have been included in *colourgrams* elaboration.

Subsequently, to simplify the identification of the color parameters mainly influenced by mixture composition and stress conditions, an alternative approach of data dimensionality reduction was considered. At first, an additional cropping procedure was necessary to restrict sample images to the center of the jars in order to focus on mixtures color and eliminate image noise deriving from shadows and reflections (Figure 3-4 on next page).

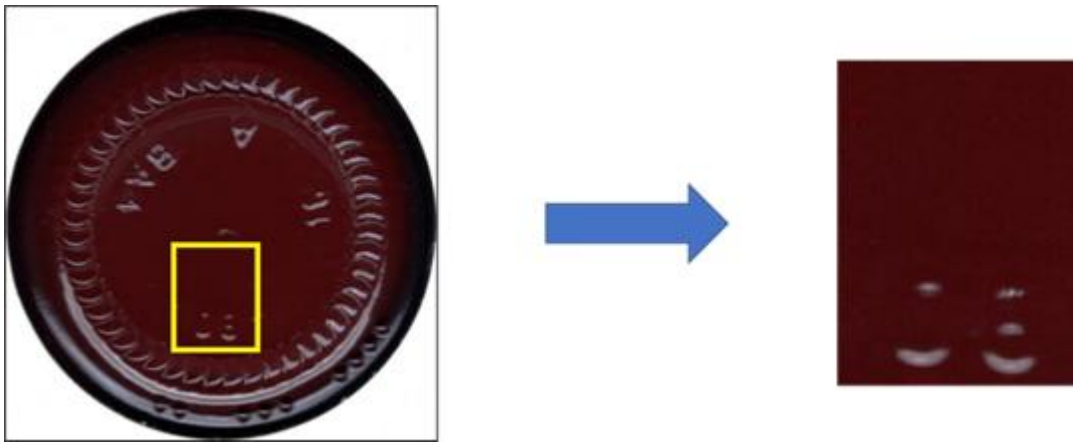


Figure 3-4 Additional image cropping to reduce the effect of shadows and reflections; the obtained images were used to calculate the median value of the considered color parameters.

From these re-cropped images, median values of red, green, blue, lightness, relative red, relative green, relative blue, hue, saturation and intensity were obtained, which correspond to the color-related parameters considered in the calculation of *colourgrams*. Median values were calculated considering only the pixels with blue values lower than 50; this threshold was set to exclude remaining reflections due to signs of the jar glass. In this manner, each re-cropped image was converted into a feature vector of 10 elements, corresponding to the median values of the considered color-related parameters.

3.2.5 Exploratory Analysis of *Colourgrams* by PCA

The mean centered matrix of *colourgrams* was analyzed by Principal Component Analysis (PCA), by using the PLS Toolbox Version 8.8.1 (Eigenvector Inc.) running in MATLAB environment (ver. 9.3, The Mathworks Inc., USA). PCA performed on the *colourgrams* matrix allowed to visualize the overall structure of the dataset composed of 864 RGB images, to highlight clusters of similar images, and to identify the presence of outlier images.

3.2.6 Modelling of Median Red Parameter

In order to study the effect of mixtures composition and of time on the SYP color of the S series of samples, i.e., the series for which the time effect is expected to be much more relevant, the median values of different color parameters were considered for possible modelling by DoE. Since median red was identified as the parameter showing better correlation with the percentage of strawberry content, for each thickener type the corresponding response surfaces were calculated considering the images acquired at the beginning (T00) and at the end of the stress time interval (T05). It should be noticed that, for the samples that were imaged twice in the same image acquisition session, the mean value of the two corresponding median red values was considered for model calculation.

Analysis of variance (ANOVA) was used to verify the statistical significance of the model and of the lack of fit; the comparison of the variation sources was based on the Fisher distribution ($P < 0.05$). The values of the coefficient of determination (R^2), of the adjusted R^2 (R^2 Adj) and of the predicted R^2 (R^2 Pred, estimated by leave-one-out crossvalidation) were considered to express the models performance, while response surfaces were obtained to represent the variation of median red with mixture composition at the different image acquisition times.

3.2.7 Evaluation of Color Variation by PLS-DA

In order to highlight color variations caused by stress conditions over time and to identify the color parameters more suitable to monitor this aspect, the dataset of median values of the different color parameters described in Section 3.2.4 was analyzed using a multivariate classification algorithm, namely Partial Least Squares-Discriminant Analysis (PLS-DA). In particular, a PLS-DA classification model was calculated to discriminate between stressed and control samples. In this case, only the samples belonging to the acquisition times from T01 to T05 were considered, while the T00 samples were excluded.

The dataset of the median values was preprocessed using autoscaling, and the optimal number of latent variables was selected in cross-validation considering 4 deletion groups, corresponding to the 4 randomized sample blocks described in Section 3.2.2.

The classification performances were evaluated considering sensitivity (SENS), specificity (SPEC) and classification efficiency (EFF), calculated in calibration and cross-validation. SENS is the percentage of samples of the modelled class correctly accepted by the class model, SPEC is the percentage of objects of the other classes correctly rejected by the class model, and EFF is the geometric mean of SENS and SPEC (Ballabio et al., 2018).

It has to be highlighted that in this study the PLS-DA model was not calculated for prediction purposes, but only to identify one or more color descriptors that are more related to color variations due to stress conditions. To this aim, the Variable Importance in Projection (VIP) scores were used to point out color parameters with higher relevance in the discrimination between control and stressed samples. Indeed, VIP scores provide a measure of the relevance of each variable in the definition of the PLS-DA model; as a general rule, variables with VIP score greater than 1 can be considered as significant (Gosselin et al., 2010).

3.3 Results and Discussion

3.3.1 PCA of *Colourgrams*

An initial exploratory analysis was performed on the entire *colourgrams* dataset, considering all acquisition times (from T00 to T05) and both control and stressed samples. The analysis of PC1-PC2 score plot revealed interesting trends, as reported in Fig. 3-5, where the same samples are colored according to stressed and control samples (A), and according to the percentage of strawberry of the corresponding mixtures (B).

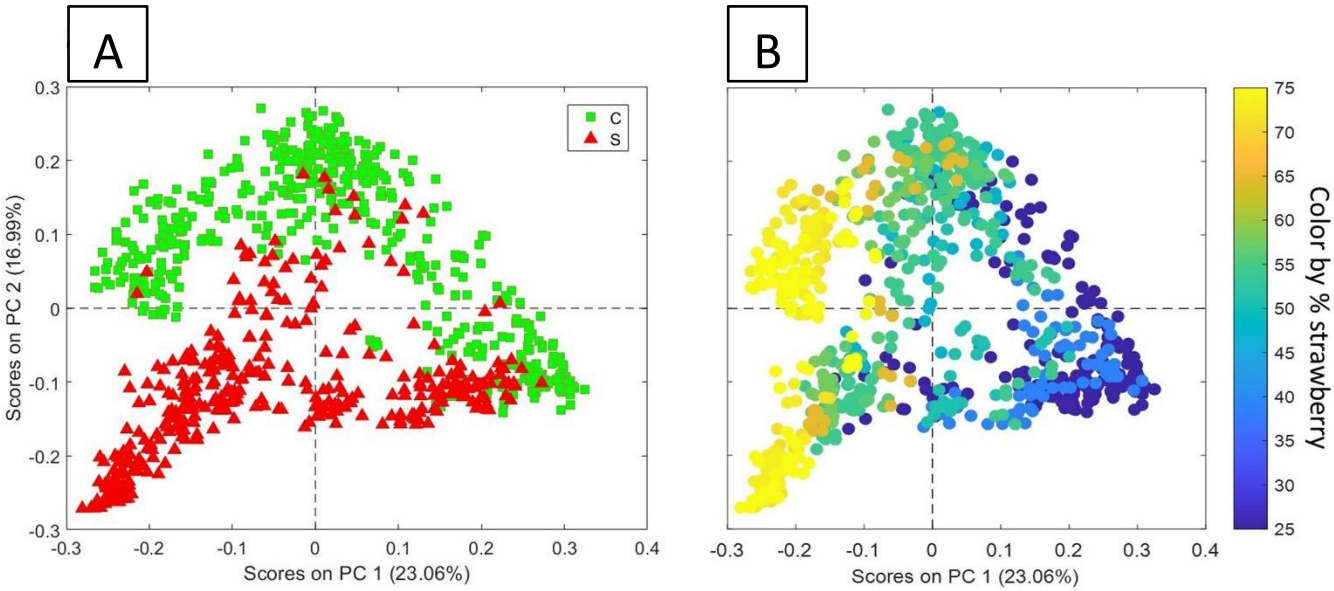


Figure 3-5 PC1-PC2 score plot of the colourgrams matrix; the objects are colored according to stressed and control samples (A) and according to strawberry percentage of the corresponding mixtures (B).

PC1 describes the color differences of the SYP samples according to the percentage of strawberry in the mixture (Fig. 3-5 B), where the samples prepared with the highest amount of strawberry are located at negative PC1 score values, while the samples with the lowest strawberry percentage are located at positive PC1 score values. Indeed, considering the corresponding RGB images, it is possible to observe that the mixtures prepared with a higher amount of strawberry have a lighter and reddish color, while the mixtures with a lower amount of strawberry have a darker color.

On the other hand, PC2 highlights the differences between stressed and control samples (Fig. 3-5 A), which are more evident for mixtures with high strawberry percentage (Fig. 3-5 B).

Even if these simple observations could have been done by visually comparing the RGB images of some representative samples, it has to be considered that it is not feasible to simultaneously evaluate all the 864 images of the dataset with the naked eye. Conversely, the *colourgrams* approach coupled with PCA allowed to gain a general overview of the color properties of the considered samples and to highlight color variation trends common to all the images of the dataset.

In addition, PCA results showed that the stress conditions considered in this study resulted to be appropriate to promote color variations in the samples compared to the corresponding control samples, as well as that mixtures composition is responsible for a higher color variability than the one due to stressing factors. It must be also highlighted that stress conditions and sample composition affected sample color in different directions in the score plot, suggesting that different color parameters should be used to describe composition-dependent and stress-dependent color variation.

In particular, this PCA model allowed to highlight two relevant points: the color of mixtures is mainly influenced by strawberry percentage, and the stress conditions led to a color modification which is more pronounced for samples with a high strawberry percentage. However, the analysis of the corresponding PC1 and PC2 loading vectors did not allow to easily identify the specific color parameters mainly influenced by mixture composition and/or stress conditions. For this reason, the subsequent modelling steps were carried out considering a simpler approach based on the use of median values of the color parameters.

3.3.2 Mixture Design Models on Median Red Parameter

The analysis of the mixture design models calculated for different color parameters revealed that the median red value was the most relevant parameter to describe color variation of S samples according to mixture composition. A preliminary visual inspection of the scatter plot of strawberry percentage vs. median red at T00 and T05 (Fig. 3-6 A and Fig. 3-6 B, respectively) has however revealed an unexpected behavior for some LBG samples (namely: LBG10 and LBG13 samples at T00; LBG3, LBG10 and LBG13 samples at T05), suggesting a possible anomaly in their color. In particular, LBG13 was characterized by a higher median red value at both T00 and T05 compared to other samples prepared with the same amount of strawberry, which corresponds to the lowest limit, i.e., 25%. At T05 LBG3 showed the same anomalous behavior of LBG13. On the other hand, mixture LBG10, containing the highest amount of strawberry (75%), showed a median red value much lower than the one measured for the other samples with very similar strawberry percentage.

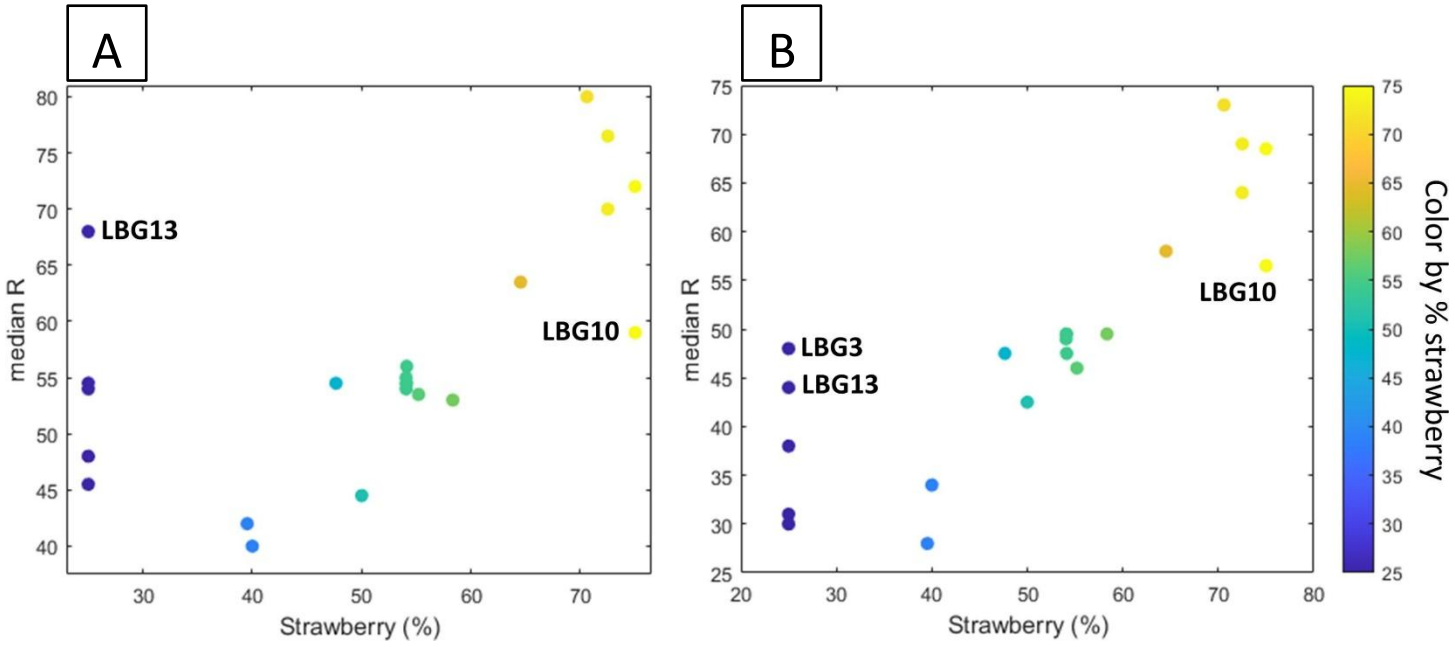


Figure 3-6 Variation of median R with increasing strawberry percentage in the formulation containing LBG as thickener at T00 (A) and T05 (B).

To further confirm this trend, the images corresponding to the anomalous samples were visually compared with those of the most similar samples in terms of composition, in particular with regard to strawberry percentage; the reader can view these images and compare them with the naked eye: Fig. 3-7 reports in subplot a) the images of the mixtures with a low percentage of strawberry (25%) and in subplot b) the images of the mixtures with a high percentage of strawberry (70.6–75%). In the figure, the anomalous samples are boxed in blue and the replicate mixtures have the name label in the same color. As it can be seen, in Fig. 3-7 A anomalous samples appear lighter than the similar mixtures, while in Fig. 3-7 B they appear darker than the similar mixtures. LBG13 (Fig. 3-7 A) and LBG10 (Fig. 3-7 B) are anomalous from the beginning, i.e., at both

T00 and T05. Moreover, it can be noticed that LBG10 is markedly different from its replicate mixture LBG6. The most likely cause for this behavior is ascribable to problems that arose during the cooking process. LBG3 instead changed its color in an anomalous way over time and its final color appears also less homogeneous than that of the other samples.

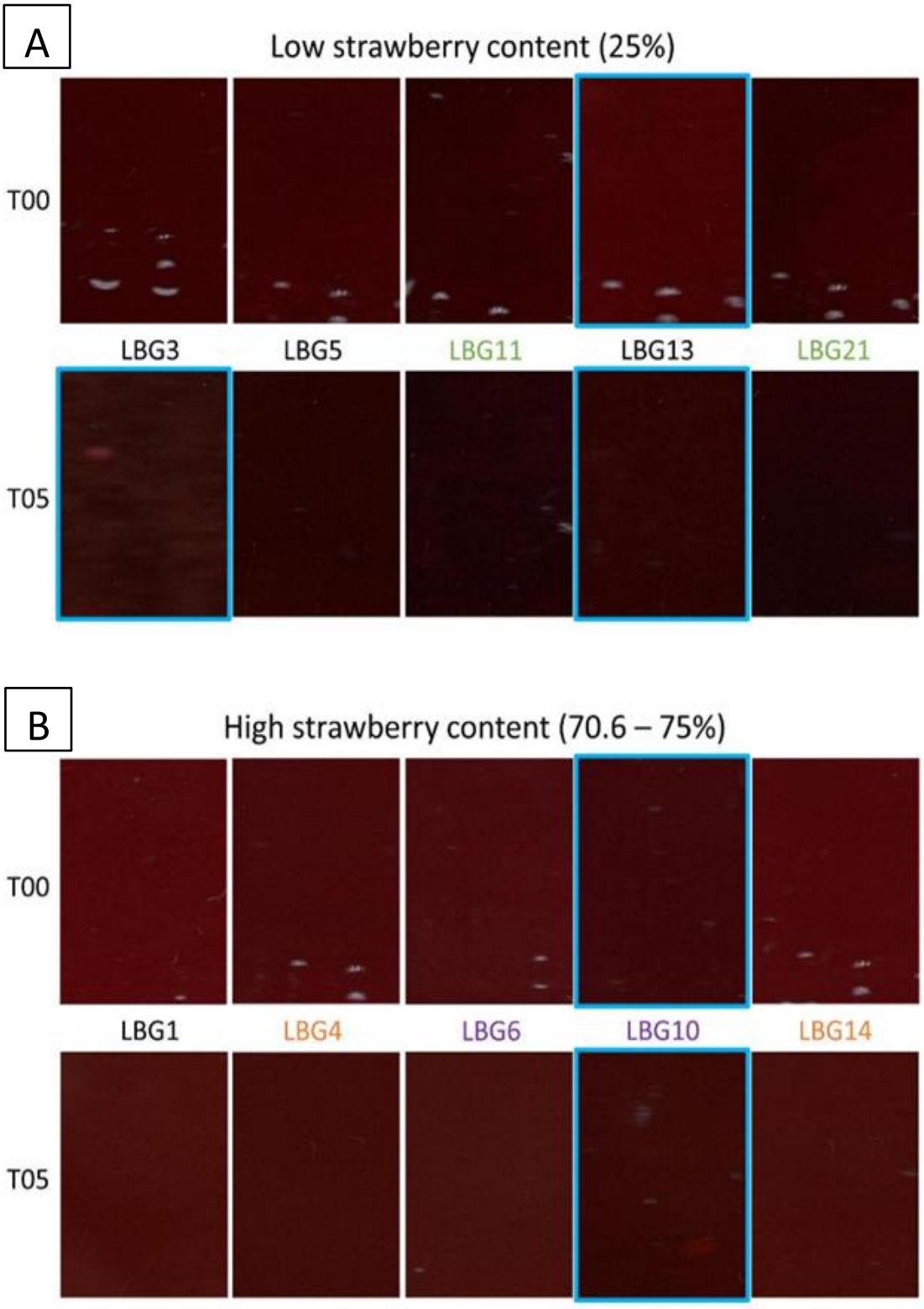


Figure 3-7 Original cropped images of mixtures with low percentage (A) and high percentage (B) of strawberry (anomalous samples boxed in blue; same label color for replicated mixtures).

These considerations led us to eliminate the anomalous samples from the dataset before calculating the mixture design models. In particular, the model calculated on T00 samples was obtained after excluding LBG10 and LBG13, while in the model calculated on T05 samples also LBG3 was excluded in addition to LBG10 and LBG13. Indeed, these samples resulted to be anomalous since they did not show the same relationship between sample color, expressed in terms of median red value, and sample composition as the other mixtures and their inclusion in the mixture design models would have inevitably made them less reliable. However, for the sake of completeness, the performances of the models calculated including also the anomalous samples are reported in Table 3-3.

	LBG series	
Stressing time	**T00**	**T05**
Model	Significant	Significant
Lack of fit	Significant	Significant
R^2	0.44	0.72
R^2 Adj	0.35	0.67
R^2 Pred	0.04	0.50
Model terms	Linear mixture: A, B, C, D	Linear mixture: A, B, C, D

Table 3-3 Statistics of the mixture models for the median red parameter obtained for the LBG thickener at T00 and T05 when the anomalous samples were maintained in the models.

Table 3-4 reports the significant terms of the models calculated for both PEC and LBG thickener types at T00 and T05, and the corresponding performances. It should be noted that before accepting each model it was verified that the errors had a normal distribution and that the error variance was constant for any value of the independent variable.

	PEC series		LBG series	
Stressing time	**T00**	**T05**	**T00**	**T05**
Model	Significant	Significant	Significant	Significant
Lack of fit	Not significant	Not significant	Not significant	Not significant
R^2	0.84	0.85	0.96	0.99
R^2 Adj	0.81	0.82	0.93	0.98
R^2 Pred	0.73	0.73	0.64	0.90
Model terms	Linear mixture: A, B, C, D	Linear mixture: A, B, C, D	Linear mixture: A, B, C, D Quadratic terms: AB, AC, AD, BC, BD, CD	Linear mixture: A, B, C, D Quadratic terms: AB, AC, AD, BC, BD, CD
Excluded samples			#10, #13	#3, #10, #13

Table 3-4 Statistics of the mixture models for the median red parameter obtained for the PEC and LBG thickeners at T00 and T05.

Overall, the models were satisfactory, with acceptable R^2 Pred values. The median red parameter was adequately fitted using a linear model when the mixture was prepared using PEC as a thickener, while a quadratic model was found to be more adequate to fit median red of the mixtures containing LBG.

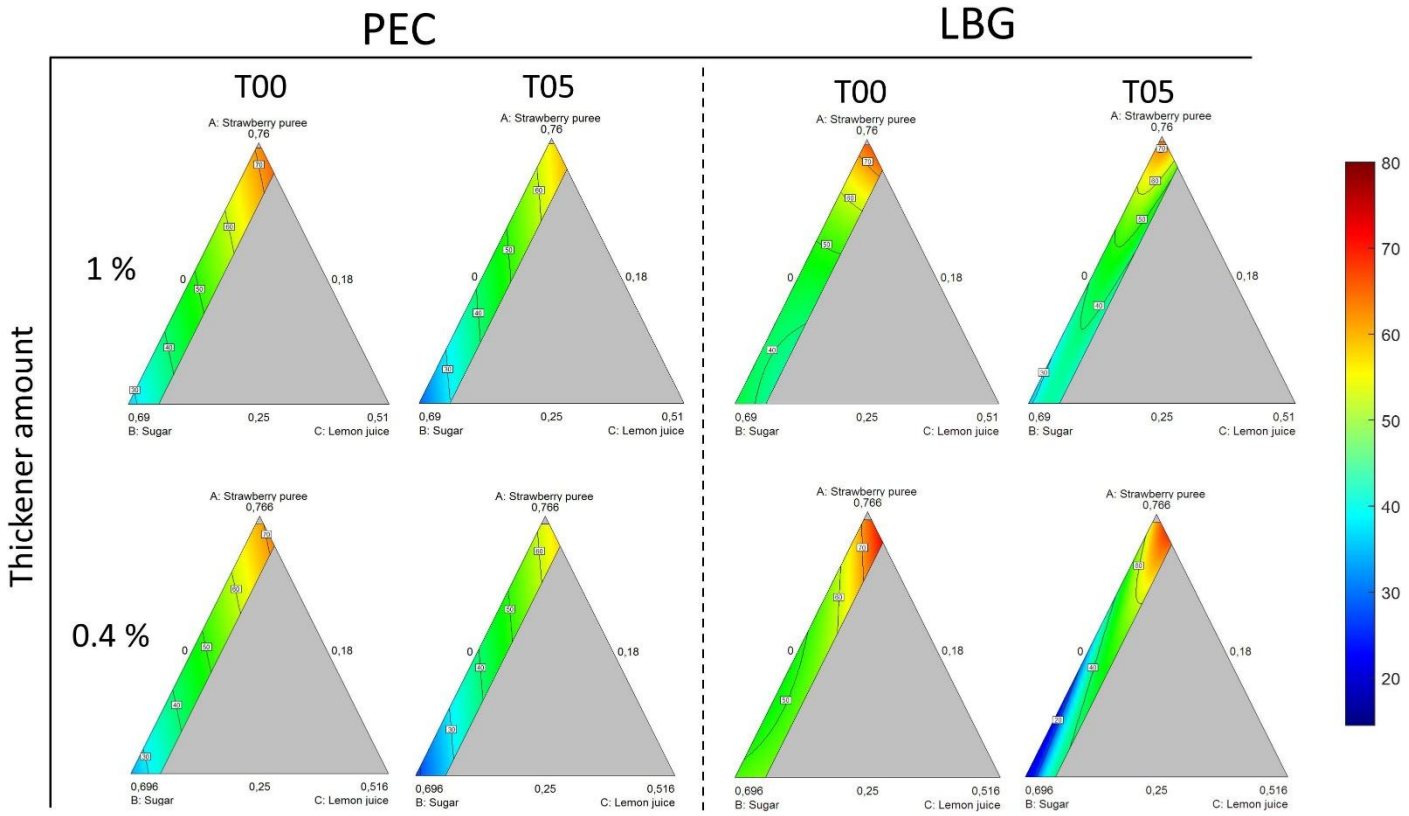


Figure 3-8 Response surfaces for median R calculated from SYP images at initial (T00) and final (T05) acquisition times (samples containing PEC as a thickener on the left, samples containing LBG on the right; thickener amount decreases top-down).

Fig. 3-8 shows the response surfaces of the models, in the triangular domain defined by the strawberry purée, sugar and lemon juice components. The lower (0.40 %) and higher (1.00 %) levels of the fourth component, i.e., the thickener, are shown below and above, respectively. It is interesting to observe that all surfaces show a similar trend as regards the variation of the median red parameter with the composition of the mixture and with stressing time. In general, the median red values increase as the amount of strawberry puree increases, as expected, and they tend to decrease with stressing time. Regarding the PEC response surfaces, the decrease in median red observed with time is almost independent of the amount of thickener in the mixture; a slight positive effect of lemon juice on median red values can be observed both at T00 and at T05. The LBG response surfaces show that, considering the same proportions of the other ingredients, at T00 the median red value is generally lower with a higher amount of thickener, and in this case it seems that the addition of lemon juice does not lead to a significant effect. At T05 the decrease in median red values is more pronounced for mixtures containing a low amount of strawberry, lemon juice and thickener. However, the comparison between the response surfaces at T00 and T05 with LBG = 0.4% also suggests that the maximum amount of lemon juice combined with the highest possible level of strawberry leads to the highest and most stable median red levels, which could therefore correspond to the optimal condition.

3.3.3 PLS-DA Model on Median Values

Since LBG3, LBG10 and LBG13 samples resulted to be outlier at T00 and/or at T05, they were also excluded from the calculation of the PLS-DA model to discriminate between stressed and control samples considering all the acquisition times from T01 to T05.

The PLS-DA model was calculated with 4 latent variables selected according to cross-validation (*see* Section 3.2.7) and the corresponding results are reported in Table 3-5. The satisfactory results expressed in terms of SENS, SPEC and EFF values in calibration (Cal) and cross-validation (CV) confirmed a detectable color difference between stressed and control samples. Fig. 3-9 A shows the Y values in cross-validation (Y CV) for control class versus the acquisition order, where the samples are colored according to control and stressed classes. In particular, the samples with a Y CV predicted value higher than the discriminant threshold (dashed red line) are classified as control samples, while the other samples are classified as stressed samples. As expected, the color difference between control and stressed samples becomes progressively more pronounced over time; indeed, the number of misclassified samples is much lower at T05 than at T01.

	SENS	**SPEC**	**EFF**
Cal	93.2 %	86.2 %	89.6 %
CV	91.7 %	85.0 %	88.3 %

Table 3-5 Results of the PLS-DA model calculated on the dataset of median colour values; the SENS and SPEC values are referred to the control (C) class.

Furthermore, Fig. 3-9 B shows the Y CV predicted values for control class against strawberry percentage in the mixtures. In this case it is possible to observe that the samples with a strawberry percentage equal to 55% or higher are better classified compared to the samples with lower strawberry amount. This finding confirms that the higher the strawberry percentage, the more pronounced is color variation due to stress conditions.

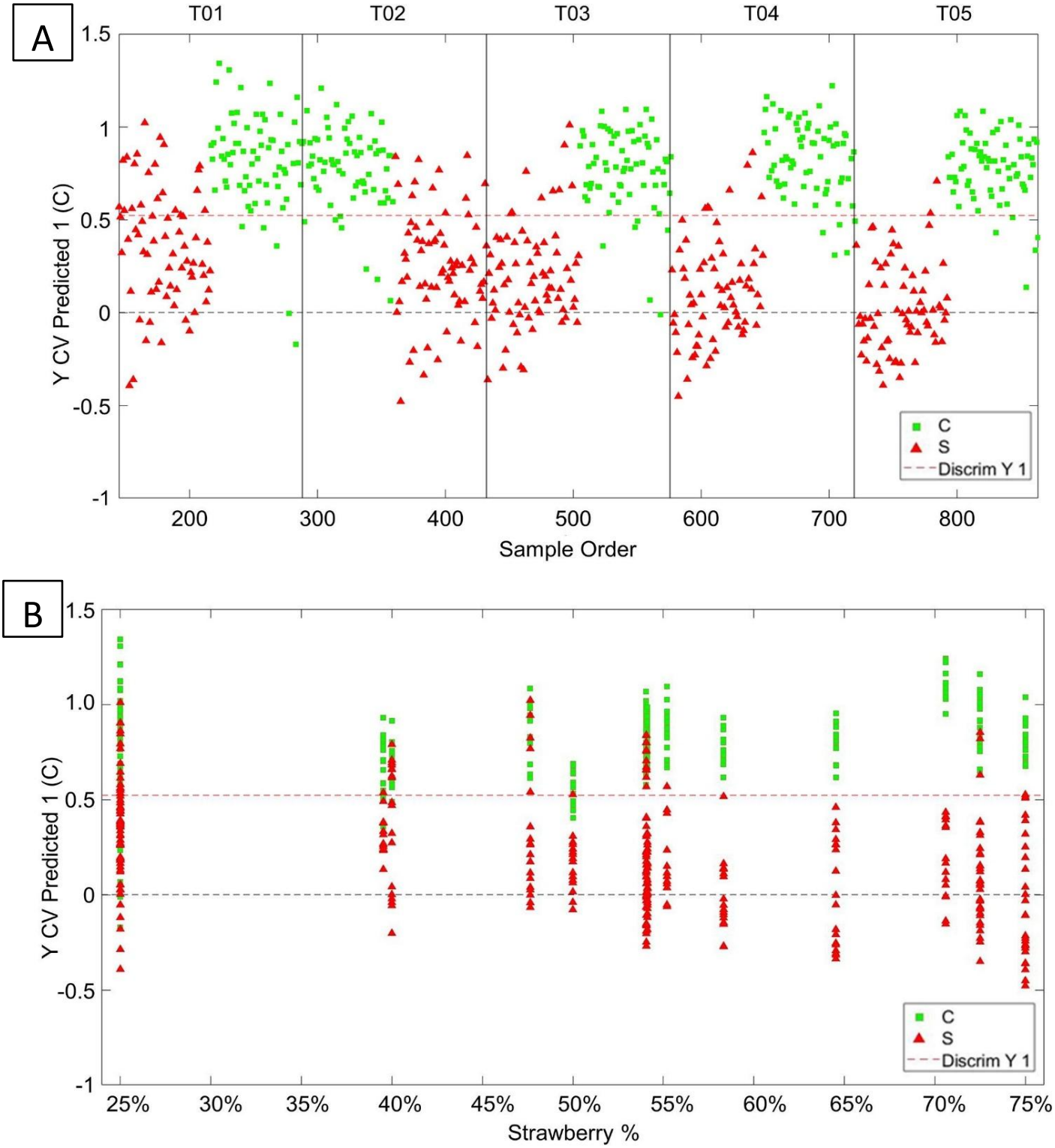


Figure 3-9 Y CV predicted for C class vs. acquisition order (A) and vs. strawberry percentage of the corresponding mixtures (B) obtained from the PLS-DA model calculated on the dataset of median color values.

As previously stated in Section 3.2.7, this PLS-DA model has not been calculated for prediction purposes, i.e., to predict unknown samples assignation to stressed or control classes, but to obtain a latent variables space oriented toward the maximum separation of the samples according to storage conditions. In this manner it was possible to focus on color variability related to degradation phenomena due to the considered stressing factors and observe how this variability is related with mixtures composition and storage time, as previously discussed in the description of Fig. 3-9.

In addition, by observing the VIP scores and the regression vector of stressed class of the PLS-DA model it is possible to identify the color parameters that are more relevant to describe color variations due to stress conditions and evaluate how they vary from control to stressed samples. More in detail, Fig. 3-10 A shows the VIP scores of the PLS-DA model, where it is possible to observe that median values of relative red (RR), relative green (RG), saturation (S) and hue (H) have VIP score values higher than one (Fig. 3-10 A), thus resulting to be the more relevant variables for the discrimination. In particular, the median value of RG parameter shows the highest VIP scores value, suggesting that this parameter is the most relevant to characterize color differences due to stress conditions. Furthermore, considering the PLS-DA regression vector of stressed class (Fig. 3-10 B), it is possible to observe that median RG has a positive value of the regression coefficient, therefore RG values of the SYP samples tend to increase for stressed samples over time.

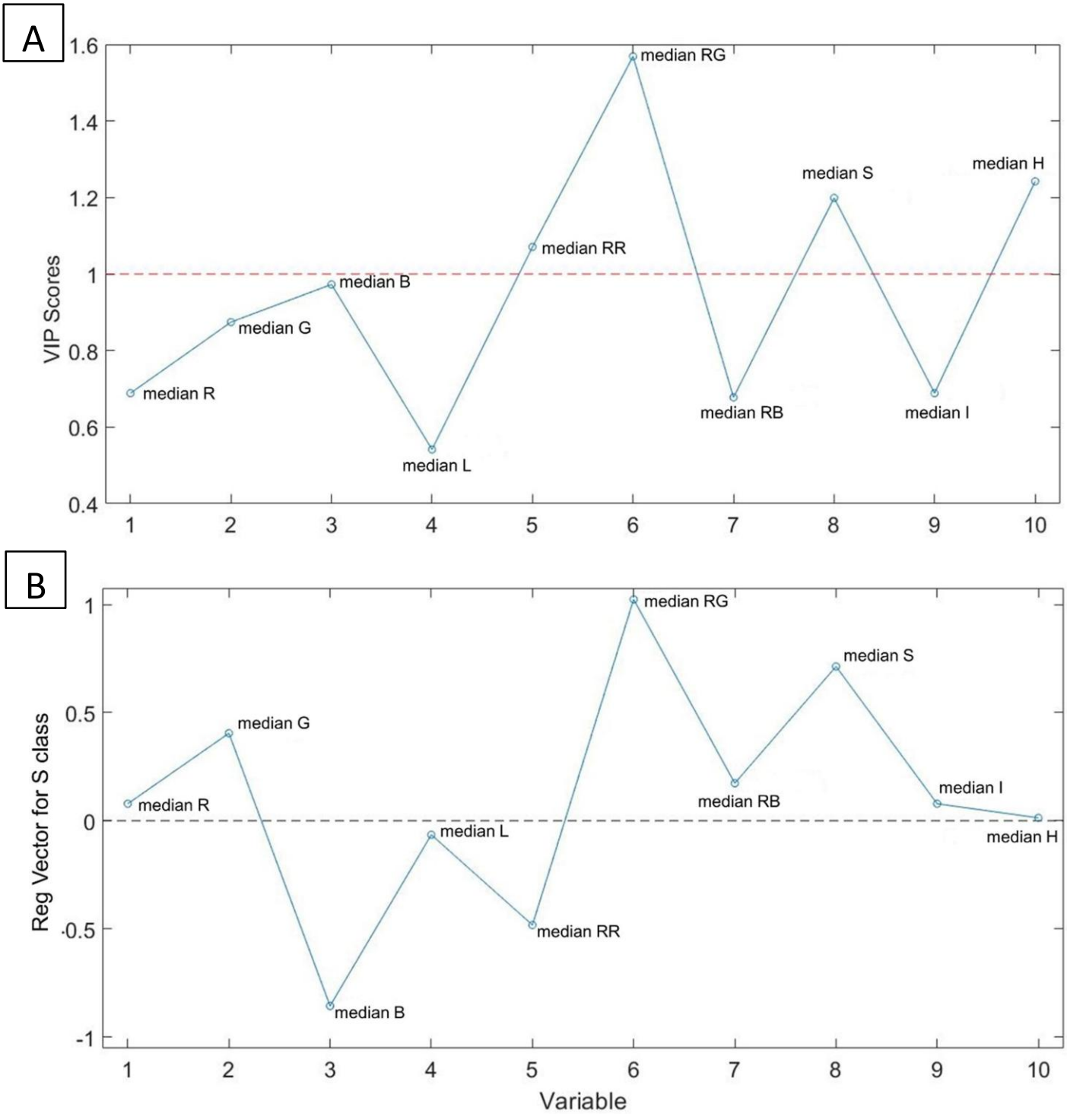


Figure 3-10 PLS-DA model calculated on the dataset of median color values: VIP scores (A) and regression vector for S class (B).

3.4 Conclusions

In this work we investigated how the ingredients proportions influence color and color stability of a strawberry-based preparation widely used in industry, namely strawberry yoghurt purée (SYP). The natural pigments mainly responsible for the color of this food product are anthocyanins, whose color may vary due to a multitude of factors related to food composition and storage conditions, including pH, temperature, light exposure, oxidase enzyme activity, water activity, and total soluble content among others.

To tackle these problems, we have chosen to investigate how color depends on SYP composition, by analyzing a series of SYP samples whose composition varied according to an I-optimal mixture design. Furthermore, we focused on the effect of exposure to intense light at a temperature of 35 °C, measuring by RGB imaging how the color of the product changed during a period of 5 weeks, compared to the color of control samples with the same composition stored for the same period in the dark at a temperature between 0 and 4 °C.

The RGB images of the samples were analyzed using different strategies, including multivariate methods that allowed to consider various color parameters deriving from the RGB ones in order to explore different color features of the samples. The results highlighted that the optimal parameter to monitor SYP color is the median value of the red channel (R) of the RGB images, which as expected depends strongly on the amount of strawberry purée, but also on other factors such as the amount of lemon juice and the type of thickener, although this latter component is present in small amounts.

Furthermore, the comparison between the samples stored under stress conditions with the control samples showed a progressive browning over time, which is more pronounced the more the amount of strawberry purée is high, since anthocyanins are the thermolabile components of the mixture. Results also showed that the type of thickener had an effect on the color degradation kinetics of the product. Finally, it was observed that, among the color parameters extracted from the images, relative green tends to increase for stressed samples over time. This color parameter, together with red, could be used as an index of the degradation process of SYP color.

These results are only a first step in the understanding of such complex phenomena. Future applications of these preliminary results will be focused on the optimization and validation of the components proportions, in order to find the best compromise between a bright red color of the fresh product and color stability over time. In turn, this goal will be achieved through the development of targeted image analysis strategies aimed at building efficient and robust predictive models. The same approach could then be used on other strawberry-based products for the rapid and non-destructive evaluation of color related properties on an industrial scale in view of the implementation of green and eco-friendly strategies to monitor product quality.

Chapter 4: Food Control

Analysis performed for food control application involved the European Database of Processing Factors (PF). This database is an enormous collection of studies (around 17000 studies in the second update analyzed in this work) on pesticides residues in various food matrices, from raw agricultural commodity (RAC) to final processed commodity (PC). This kind of information is expressed using the Processing Factor (PF) parameter, i.e., the ratio of the concentration of a specific pesticide residue in the processed product to that in the corresponding unprocessed commodity. The PF database includes more than 250 active substances, across 120 processed commodities derived from 70 different food matrices.

This dataset required extensive preprocessing, as it was structured in categorical features that are not directly suitable for classical chemometric algorithms like PCA. After initial data wrangling operations, a custom and intuitive three-level risk system has been implemented based on minimum and maximum PF values. The next step involved transforming data into distance metrics, according to Jaccard definition. State-of-the-art (e.g., treemap and alluvial plots on categorical features dataset) and multivariate (PCA on the Jaccard distance matrix) visualization techniques were applied to compare their graphical representation and highlight the added value of multivariate approaches, even in a categorical-features system. The application of multivariate analysis techniques could lead to the unraveling of additional layers of information, providing interesting insights for safety and control in food processing.

4.1 Introduction

Food control is a cornerstone of food safety, encompassing procedures and operations defined to ensure that food products remain safe for final consumers through the entire food chain. The European Union places a strong emphasis on food safety starting from the early 2000's: Regulation (EC) N° 178/2002 is considered the founding legislative act in this direction, laying down food law principles in Europe. The renewed approach toward food safety contained in Regulation (EC) N° 178/2002 came after a period in which consumers' trust in food products was severely undermined by bovine spongiform encephalopathy (BSE) outbreak, a food safety crisis that occurred between the 1980s and 1990s, resulting in the death of millions of cattle and hundreds of people.

Despite the years that have passed since that episode, food safety remains one of the most significant factors influencing Europeans' food choices. The European Food Safety Authority (EFSA) emphasizes that 46% of respondents prioritize food safety when purchasing food, reflecting high public awareness and concern about potential risks (EFSA, 2022). Among the 15 food safety topics surveyed, pesticides residues were among the top concerns, with 65% of Europeans acknowledging awareness of these residues, and 40% ranking them as a primary food safety concern. The report also underscores the complexity of public perceptions, where confidence in regulatory systems coexists with anxiety about foodborne risks and contaminants. It highlights the need for more accessible tools and clearer communication strategies to overcome significant barriers to engagement, such as the technical nature of food safety information.

Given the growing complexity of food safety challenges and the increasing demand for transparency, enhancing the accessibility of PF database information through chemometric strategies can reveal patterns and relationships, offering deeper insights into how pesticide residues behave across different food matrices and processing methods.

4.1.1 European Legislation on Pesticides Maximum Residues Levels

The European framework for regulating pesticide residues in food and feed is primarily governed by Regulation (EC) N° 396/2005 which harmonizes the setting and enforcement of Maximum Residue Levels (MRLs) across the European Union. Before digging into the details of this piece of legislation, some specific terms used broadly throughout the text need to be clarified.

- **Active Substance** (AS): defined in *Article 3(2)(a)* of 396/2005 as:

  "Substances or micro-organisms, including viruses, having general or specific action against harmful organisms or on plants, parts of plants or plant products"

  This term refers specifically to the biologically active component of a pesticide product that achieves the intended pest control effect.

- **Pesticides**: the term "pesticides" is not directly defined in 396/2005, but it is commonly used to refer to **Plant Protection Products** defined in Article 2 of Regulation (EC) N° 1107/2009 as:

  "This Regulation shall apply to products, in the form in which they are supplied to the user, consisting of or containing active substances, safeners or synergists, and intended for one of the following uses:

  (a) protecting plants or plant products against all harmful organisms or preventing the action of such organisms, unless the main purpose of these products is considered to be for reasons of hygiene rather than for the protection of plants or plant products;

  (b) influencing the life processes of plants, such as substances influencing their growth, other than as a nutrient;

  (c) preserving plant products, in so far as such substances or products are not subject to special Community provisions on preservatives;

  (d) destroying undesired plants or parts of plants, except algae unless the products are applied on soil or water to protect plants;

  (e) checking or preventing undesired growth of plants, except algae unless the products are applied on soil or water to protect plants.

  These products are referred to as 'plant protection products'."

Therefore, the active substance is the core chemical or microorganism responsible for pest control in a pesticide product. Pesticides refer to the broader category[7] encompassing the full formulation[8] authorized and designed for a specific use in agriculture or along the food chain.

[7] Including plant protection products but also other products such as biocides, which are formulations intended for pest control in non-plant uses (e.g., insects, rats and mice).
[8] Active substance(s) plus adjuvants, solvents, and other inert ingredients making up the authorized formulation.

This distinction is critical in the regulation, as MRLs are established for “pesticide residues”, which are the traces that a pesticide (i.e., plant protection product) leaves in treated products. These traces usually include the active substance but also its metabolites and sometimes additional compounds (like degradation products and reaction products) when they are relevant for risk assessment. For the scope of the present work, and in particular for results discussion in Section 4.3, the term “pesticide” is used as synonym of “active substance”.

Coming back to the rationale behind Regulation (EC) No. 396/2005, it aims at ensuring a high level of consumer protection while facilitating the smooth functioning of the internal market and trade with third countries. It supersedes and consolidates earlier directives, creating a single comprehensive legal instrument. The regulation applies to all products of plant and animal origin intended for human or animal consumption, as well as composite and processed products. It establishes the concept of MRLs as the maximum concentration of each pesticide residue of food safety concern legally permitted in or on food and feed when pesticides are applied according to good agricultural practices (GAP). These levels are set to be safe for all population groups, with a particular focus on most vulnerable ones like children, pregnant women and vegetarians.

MRLs are established or modified based on scientific risk assessments coordinated by the European Food Safety Authority (EFSA). Applications for MRLs can be submitted by Member States, stakeholders, or commercial parties. These applications must include comprehensive data on pesticide use, residues levels, toxicology, and analytical methods. EFSA evaluates these applications and provides a reasoned opinion, which includes an assessment of consumer and environmental risks. The European Commission subsequently adopts MRLs through a regulatory process involving Member States. Temporary MRLs may be set in specific situations, such as unauthorized pesticide use resulting from environmental contamination, newly introduced products, or minor dietary components. These MRLs are reassessed periodically and adjusted based on new data or safety concerns.

Member States are responsible for monitoring and enforcing compliance with MRLs through official control programs. Sampling and analysis are conducted to ensure that pesticide residues in food and feed do not exceed legal limits. The regulation also prohibits mixing or processing non-compliant products to bring them into conformity. Member States are required to publish annual monitoring results online, providing detailed data on sampling, residue levels, and non-compliant products. This transparency promotes consumer trust and accountability among producers and traders.

The regulation emphasizes consumer safety, particularly for high-risk groups, and aligns with international standards like those set by the Codex Alimentarius. It also accounts for differences in agricultural practices and pesticide use in third countries by setting import tolerances when safety criteria are met. Regulation (EC) No. 396/2005 is a cornerstone of EU food safety legislation, ensuring the effective management of pesticide residues in food and feed to protect consumer health and support fair trade practices. Its comprehensive approach — encompassing risk assessment, MRLs setting, monitoring, and transparency — provides a robust framework for addressing the challenges of pesticide residues management in a globalized market.

For processed and composite products, the regulation considers changes in pesticide residues levels (or nature) caused by processing or mixing. Specific dilution or concentration factors may be applied to reflect these changes accurately, ensuring that derived products remain compliant with MRL requirements. In this context, the *Processing Factor* concept becomes relevant.

4.1.2 Processing Factor

The *Processing Factor* (PF) represents the ratio of pesticide residue in a processed commodity (PC) to that in the raw agricultural commodity (RAC). This metric is crucial for evaluating the impact of food processing on pesticide residue, such as dilution, concentration, or chemical transformation during processes like cooking, fermentation, or dehydration. The PF allows for refined dietary exposure assessments, monitoring compliance with MRLs, and estimating pesticide levels in animal feed derived from processed products. The PF is calculated as:

$$PF = \frac{Residue\ in\ the\ processed\ commodity\ (\frac{mg}{kg})}{Residue\ in\ the\ raw\ commodity\ (\frac{mg}{kg})} \quad \textbf{(12)}$$

with PF values greater than 1 indicating concentration, while lower than 1 indicating reduction during processing.

PFs are determined through processing studies, which simulate typical industrial or domestic processes using field treated RACs with quantifiable residues. The OECD (Organisation for Economic Co-operation and Development) provides detailed guidance (OECD, 2008a, 2008b) on designing, conducting, and reporting processing studies. These studies are an integral part of ensuring food safety and supporting regulatory decisions on the management of pesticide residues in processed foods. At least two independent trials are required to ensure consistency. When significant procedural differences exist (e.g., red vs. white wine production), separate trials must be conducted for each variation. Residues behavior during processing depends on factors like water solubility, fat affinity, or heat stability, which influence residues distribution in the final food matrix (e.g., pesticides concentrating in oils or juices). The results of processing studies can sometimes be extrapolated to similar commodities and processes within the same category (e.g., orange juice PF applied to other citrus fruits like mandarins and lemons). However, these extrapolations require careful validation to ensure reliability. When large variability exists between trials, additional studies may be needed to refine the PF. Similarly, extreme concentration factors highlight the importance of accurate residues assessment for high-risk scenarios.

PFs are used to:

i. Determine refined dietary exposure for consumers.

ii. Calculate dietary burdens in animal feeding studies, impacting residues in animal-derived products like milk or meat.

iii. Set MRLs for processed foods.

iv. Monitor adherence to MRLs established for raw agricultural commodities.

During food processing, pesticide residues in RACs can undergo significant changes, which must be carefully accounted for when calculating the PF. When the residue definitions differ between the RAC and the PC, the PF must be adapted to include not only the active substance but also any relevant metabolites or transformation products formed during processing. This adjustment is essential because chemical transformations, such as degradation, enzymatic activity, or heat-induced reactions, can alter residue composition, leading to the formation of new compounds that may contribute to dietary exposure or toxicity.

To ensure accuracy in PF calculations under these circumstances, the residue levels of transformation products are often expressed as parent-equivalent residues, typically through molecular weight adjustments or other standard recalibration methods. This approach ensures that the PF reflects the total residue contribution in the processed product, aligning with the residue definition used for dietary exposure assessment and maximum residue level (MRL) setting. For example, a pesticide undergoing chemical breakdown during juice extraction or thermal processing may produce additional metabolites that were not present in the RAC but must still be quantified and included in the PF calculation.

Overall, aligning PF calculations with residue definitions tailored to the specific processing scenario is critical for accurately assessing pesticide concentrations in processed foods. This approach provides a reliable foundation for dietary risk assessment, regulatory compliance, and MRL establishment by accounting for both the behavior of the parent compound and any relevant transformation products formed during processing.

4.2 Materials and Methods

4.2.1 European Database of Processing Factors

The EU Database of Processing Factors for pesticide residues was developed to support consumer risk assessment by providing a structured, harmonized approach to evaluate how food processing affects pesticide residues. This database, established under an EFSA-funded project started in 2016, aligns with the FoodEx2 classification system, ensuring compatibility with existing EU food monitoring frameworks (for further information *see* EFSA, 2015). The database compiles processing factors derived from studies documenting the changes in pesticide residue levels as RACs are processed into food and feed products (processed commodities or PCs in short). These studies adhere to uniform quality criteria and are evaluated for their representativeness of typical food processing techniques, as outlined in (Scholz et al., 2022). The database facilitates the identification of residue behavior patterns, supports MRL setting, and aids dietary exposure assessment. However, uncertainties may arise from missing storage stability data, variations in processing methods, or limited availability of PFs for certain commodity-process combinations. The database should thus be viewed as an approximation reflecting typical processes rather than definitive values.

Initially published in 2018 (Scholz et al., 2018), the database is constantly updated to expand its scope and improve its applicability in cumulative exposure assessments. Key updates in version two of the database (Zincke et al., 2022) – which is the one considered in this work - have included the integration of additional processing studies, expansion of processed commodity categories, and improved methods for deriving PFs for composite foods. The database's scope, however, remains limited to plant-based commodities, as no studies on the effects of processing on pesticide residues in animal-derived products are available.

Structurally, the database is presented as a MS Excel file with four spreadsheets:

- **ReadMe**: it serves as a comprehensive guide, detailing the context, structure, and fields of the dataset. It provides essential background information on the database, outlining its purpose, sources, and methodology.
- **ProcStudies Evaluation**: it collects the evaluation aspects and results of all included processing studies. Each record corresponds to a specific study, providing comprehensive details on the individual trial, including active substance and corresponding residue definition analyzed, raw and processed commodity involved, process and storage conditions applied, final PF value obtained and its acceptability according to studies quality parameters. The data contained in this spreadsheet were analyzed in this Thesis and elaborated as reported in Sections 4.2.2 – 4.2.4.
- **List Median PF**: it summarizes the content of the previous ProcStudies Evaluation part calculating median PF values for unique combinations of active substances, raw agricultural and processed commodities. For each combination, the overall number of trials considered from the ProcStudies Evaluation spreadsheet is provided together with a reliability assessment of the median PF values obtained. When a single active substance is evaluated not only as itself but also considering a metabolite of concern, studies aggregation is performed separately for each chemical compound linked to the original active substance.
- **List References**: it includes bibliographic information about the studies listed in ProcStudies Evaluation, such as publication year, study title, and report number, ensuring traceability and transparency of data reported.

Database's design allows exploration across various stages of the food production chain, facilitating the analysis of pesticide residue behavior from the raw material to the final processed product. Its systematic

organization is a first step towards a clear understanding of how different factors interact during food processing and their collective impact on residue levels through the final PFs calculation (*see* Section 4.1.2).

By centralizing and harmonizing data on pesticide residue changes during food processing, the EU PF database serves as a critical resource for regulatory authorities, providing a reliable foundation for setting Maximum Residue Levels (MRLs) and assessing dietary risks across the European Union. Database consultation by individual users is possible either directly through the MS Excel file[9] or via an online tool developed by the German Federal Institute for Risk Assessment (BfR) as part of the project[10]. However, direct use of the Excel file can be challenging due to its large size (over 17000 records in version 2), while the online tool has limitations, as it operates on a one-active-substance-per-query logic, which can be time-consuming and restrict broader insights.

To address these challenges, a novel approach was implemented in this work, aimed at achieving a multivariate and comprehensive exploration of the data using J-PCA method (*see* Section 2.4.3). The methodology's performance is benchmarked against other state-of-the-art techniques used for the exploratory assessment of qualitative data, such as treemaps and alluvial plots. While these latter methods enable the simultaneous visualization of different categorical variables, their interpretability becomes limited when too many variables are considered at the same time. This drawback restricts their capacity to uncover deeper and more complex multivariate relationships compared to the proposed J-PCA-based approach.

4.2.2 Data Filtering – Aggregation – Enrichment

To create a representative and analyzable dataset suitable for exploratory data analysis, extensive data wrangling was carried out on the initial dataset, i.e., on the data contained in the ProcStudies Evaluation spreadsheet of the original MS Excel file. The pre-processing workflow was systematically organized into three main stages: data filtering, aggregation, and enrichment. These steps ensured data quality, relevance, and utility for the intended analyses.

Among the 35 columns originally present in ProcStudies Evaluation spreadsheet, 6 of them were selected to proceed with these data wrangling operations: *Active Substance*, *Raw Primary (Agricultural) Commodity*, *Processed Commodity*, *Processing (Process) Code*, *Individual PF* (numeric value), *PF Acceptable?* (yes, indicative, no).

Filtering

Filtering focused on retaining only high-quality records (i.e., PF studies) to enhance the reliability of subsequent analyses. Key filtering criteria included:

i. **Quality standards**: only records labeled with "yes" or "indicative" for the *PF Acceptable?* column in the original dataset were kept, ensuring that the considered studies met quality benchmarks for PF calculation.
ii. **Sufficient study coverage**: pesticides with fewer than 100 studies were excluded to maintain a robust statistical foundation.

The original dataset was composed of 17258 studies, while the data table obtained after the two filtering criteria contained 5646 entries.

[9] Accessible at https://zenodo.org/records/6827098
[10] Accessible at https://knimehpc.bfr.berlin/knime/webportal/space/EFSA_Processing

Aggregation

The filtered dataset was aggregated into groups based on unique combinations of three key features: *Active Substance* (AS), *Raw Agricultural Commodity* (RAC), and *Processed Commodity* (PC). This aggregation structure was inspired by the aggregation logic used in the **List Median PF** spreadsheet in the original MS Excel file. However, we decided to consider in the same record the active substance and its different metabolites when present, in order to simplify the dataset for exploratory analysis, focusing on uncovering patterns and relationships rather than performing detailed risk assessment on each pesticide residue and its derivatives. By streamlining residue definitions in this manner, the analysis remained focused on broader trends while maintaining alignment with PF rationale. This is a first approach for data exploration of PF database: in the future, analysis could be extended at a high granularity level considering in separate records the different compounds originating from the same active substance. However, it is also worth mentioning that, out of 35 active substances included in the final dataset, only 5 of them have different residues definitions in addition to the starting active substance (Bupirimate, Flupyradifurone, Mancozeb, Metiram and Thiophanate-methyl).

In the aggregation step, the filtered dataset was transformed into a new dataset with 1327 rows, each one corresponding to a unique combination of active substance, raw agricultural commodity and processed commodity. From here onwards each row of the aggregated dataset will be considered as an object (also referred to as a "sample").

While aggregating data, basic statistical parameters were obtained across the *Individual PF* values of each unique combination and they were added as new features. These parameters provided a comprehensive summary of the variability and distribution of PF values for each unique combination considered during aggregation. Therefore, columns of the aggregated dataset at this stage correspond to: AS, RAC, PC, *Process Code*, *Count of Studies*, *Median PF*, *Minimum PF*, *Maximum PF* and *Standard Deviation*, for a total of 9 columns.

These statistical parameters served also for further filtering refinements: samples with a count of studies less than or equal to two were excluded to prevent unreliable median values derived from insufficient data points. Then, records with a median PF greater than or equal to 10 were removed to mitigate the influence of extreme outlier values. Finally, samples showing no variability in PF values, identified by a standard deviation equal to zero, were also excluded.

After this further filtering procedure, the final aggregated dataset accounted for 791 samples (equivalent to 4745 individual studies of the original PF database) and 9 features.

Enrichment

On the aggregated data, several new features were added to enable detailed analysis and visualization. In particular, new hierarchical groupings were added:

- **RAC group and subgroup**: retrieved from relevant literature (Zincke et al., 2022), these classifications grouped RACs into broader categories for better interpretability. For example, *cotton seeds*, *peanuts*, *rapeseeds*, *soyabeans for oil* and *sunflower seeds* are the RACs belonging to the RAC subgroup *oilseeds* which, together with the RAC subgroup *oil fruits*, constitutes the RAC group *oilseeds and oil fruits*. *See* Appendix I for a complete overview of RACs, RAC groups and subgroups.

- **OECD process group**: extracted from the Latin number part of the original Process Code and also available in OECD, 2008b, 2008a. For example, processes *Citrus juice citrus fruits (code **II** 001), Pome juice pome fruits (apples pears) (code **II** 002), Grape juice berries and small fruits (currants) (code **II** 003), Grape juice grapes (code **II** 003)* and *Stone fruit juice stone fruits (**II** 004)* all belong to the OECD process group ***II** Fruit juice*. *See* Appendix II for a complete overview of process codes included in this work.
- **PC group**: developed specifically within this thesis, using food technology expertise to cluster processed commodities logically. For example, *beer, must, sake, spirit, red, rosé and white wine* are all part of the processed commodity group *alcoholic beverages* (AB). *See* Appendix III for a complete list of processed commodities and corresponding processed commodity groups.
- **PF risk level**: a column representing PF risk level through a traffic light color-coded system for visual clarity. Green (G) class was assigned to samples with maximum PF value (calculated across all the studies corresponding to that specific combination of AS, RAC and PC) ≤ 1, indicating a decrease in pesticide residue during processing. Red (R) class for samples with minimum PF value ≥ 1, indicating concentration of pesticide residue during processing. For both green and red classes all the studies related to a unique combination of AS, RAC and PC reported consistent results with PF values lower or higher than one, respectively. However, some unique combinations of the aggregated dataset reported minimum PF value < 1 and maximum PF value > 1, representing a variable behavior between different studies. These samples were assigned to the Yellow (Y) class.

After data enrichment, the dataset was composed by 791 samples and 14 variables retaining the essential information required to investigate PF behavior while ensuring a balanced representation of the diverse combinations of active substances, processes, and commodities from the original database. The 14 variables considered at this stage include 9 categorical variables (AS, RAC, RAC subgroup, RAC group, PC, PC group, Process Code, OECD process group, PF risk level) and 5 numerical variables (number of studies, median PF, minimum PF, maximum PF, standard deviation).

Figure 4-1 summarizes all the above preprocessing steps.

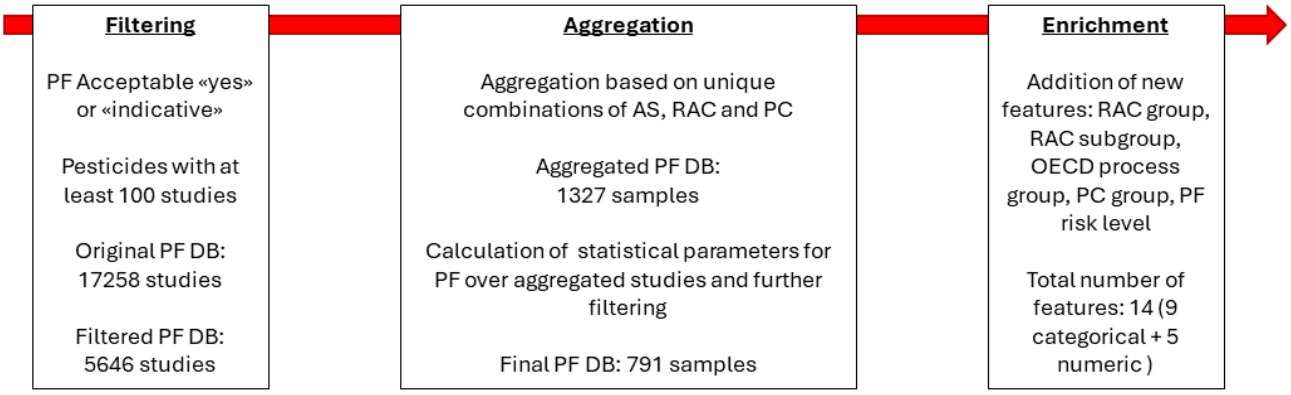


Figure 4-1 Preprocessing steps performed on European Database of Processing Factors version 2 prior to exploratory data analysis. From the original downloaded dataset (17258 records), a filtered version of 5646 records was obtained. Then row aggregation was performed to get to 1327 samples and finally, after further filtering and enrichment operations, the 791 samples x 14 variables dataset was obtained.

4.2.3 Treemap and Alluvial Plot

Treemap and alluvial representations (*see* Sections 2.4.1 and 2.4.2) of PF data were generated through direct analysis of the 791 samples × 14 variables dataset using *R* version 4.4.0 (2024-04-24). Visualizations were created employing treemap (v. 2.4-4) and ggalluvial (v. 0.12.5) packages.

Treemap

The *treemap* package in *R* provides robust tools for creating customizable treemaps. Key functionalities include dynamic hierarchy definition through the *index* parameter, rectangle sizing with *vSize* argument and advanced color coding with *vColor*, which supports schemas such as categorical, value-based, or depth-based coloring. Layout algorithms like "squarified" and "pivotSize" optimize rectangle arrangement to improve aspect ratios and logical ordering. Additional customization options, such as *fontsize.labels* and *align.labels*, address label placement and styling to enhance readability (Tennekes & Ellis, 2023). In current application:

- The *index* parameter was configured to define a three-level hierarchy, aligning with the general guidance outlined in Section 2.4.1. PF class was used as outermost level, while the innermost two levels correspond to RACs and PCs, with varying levels of granularity depending on the number of records included in the analysis. This hierarchical structure was chosen to facilitate a nested visualization, grouping processed commodities under their respective primary sub-groups and organizing them by overarching PF levels.
- The *vSize* argument was set to "count," allowing the size of each rectangle to represent the frequency of observations. This proportional representation visually emphasized the most prevalent categories, supporting a clear interpretation of the data distribution.
- The *vColor* parameter was based on the PF level, ensuring that each category (e.g., green, yellow, red) was distinctly highlighted. The categorical type setting (type = "categorical") ensured discrete color mapping, while a custom color palette provided consistency and alignment with the dataset's interpretative context.
- The "pivotSize" algorithm was employed to arrange rectangles efficiently, prioritizing logical proportions and aspect ratios. This choice enhanced readability by ensuring a well-organized layout that visually balanced the hierarchical structure and category sizes.
- Rectangles were sorted by frequency using the *sortID* parameter, emphasizing the most significant categories at each level of the hierarchy. This ensured that higher-frequency items appeared prominently, facilitating the identification of major trends.

These features, applied in combination, ensured that the treemaps effectively visualized complex categorical relationships while maintaining proportional accuracy and interpretability.

Alluvial

The *ggalluvial* package in *R* integrates the functionality of alluvial plots into the *ggplot2* framework, enabling seamless creation of these diagrams from *tidy* datasets. As an extension of *ggplot2*, it allows users to harness the layered grammar of graphics to build and customize plots. Central to the package is the *geom_alluvium()* function, which maps flows between strata, with options to control curve styles (e.g., linear, cubic, quintic, sigmoid) to suit different visualization needs. Additional layers, such as *geom_stratum()*, can emphasize individual strata, while *geom_flow()* and *geom_lode()* provide greater granularity for detailed analysis (Brunson & Read, 2023).

Key features of the *ggalluvial* package utilized for the visualization of PF dataset include:

- The *geom_alluvium()* layer to define flows between strata, with each flow corresponding to transitions across hierarchical axes. The axes represent specific categorical variables that could vary in number depending on the analysis level. In any case, the common goal was the visualization of relationships from raw agricultural commodities (RACs) to active substances (AS) while tracking intermediate processes and PF levels.
- The *geom_stratum()* function was applied to delineate the strata at each axis clearly. The alignment and size of strata were optimized for proportional representation with class occurrences for each categorical variable, aiding in the interpretability of categorical contributions at each stage.
- The use of *fill* aesthetics for the PF_level variable provided a consistent color scheme (green, yellow, red), corresponding to the three PF classes. This color-coding enabled immediate recognition of critical patterns in residue behavior across the axes.

This structured approach effectively showcased the power of *ggalluvial* in handling complex categorical datasets, offering a balance of detail and readability.

4.2.4 J-PCA

PF dataset was also analyzed by means of J-PCA. As outlined in Section 2.4.3, J-PCA algorithm consists in several steps needed to convert categorical data into quantitative variables that can be analyzed with PCA.

Firstly, from the final PF dataset obtained in Section 4.2.2, 8 categorical variables were selected: Active Substance, RAC group, RAC subgroup, RAC, Process Code, PC group, PC, and PF risk level. These variables were converted into a binary data matrix with size 791 × 308. Each column of the so obtained binary data matrix corresponds to a single class of the 8 categorical variables of PF dataset and binary coding is used to indicate sample class belonging. Redundancies due to identical binary vectors were then resolved by removing 51 duplicate variables, resulting in 257 distinct binary variables.

Jaccard similarity index was then calculated as per equation (9), and then Jaccard distance matrix was subsequently derived as per equation (10). This transformation was conducted using *MATLAB* version 9.10, producing a 791 × 791 Jaccard distance matrix, ready to be analyzed by PCA using *PLS Toolbox* version *8.8.1*.

4.3 Results and Discussion

The analysis of the Processing Factor database, pretreated as described in Section 4.2, was conducted from the particular to the general, applying various filters to the initial dataset to gather insights from different perspectives. The database was explored following the steps of the food production chain, starting with raw agricultural commodities in Section 4.3.1 and then moving to the processing phase and resulting processed commodities in Section 4.3.2. Based on the hierarchical group-subgroup structure, the most relevant and largest partitions were selected for observation and comparison. For each subgroup and group of samples, the reader is presented with examples of classical approaches for visualizing categorical data (e.g., treemaps and alluvial plots), alongside multivariate approaches using local J-PCA models. In Section 4.3.3, these methods are compared in the context of representing the entire dataset (i.e., 791 aggregated samples). Finally, Section 4.4 summarizes the main differences, advantages, and drawbacks of each approach.

4.3.1 Raw Agricultural Commodity

From the raw agricultural commodity (RAC) perspective, the following key questions should be considered:

i. To what extent does the RAC type influence the PF level?
ii. Is there a relationship between PF levels and the RACs, the type of process they undergo, and/or the final processed commodities? If so, how?
iii. If the relation outlined in point ii exists, which processes and final commodities determine a higher occurrence of yellow and red PF levels?

Appropriate data visualization can help to address these questions, while also preparing the ground for the process and processed commodity analysis. Among the 8 RAC groups available in the Processing Factor database, we focused on *Fruits* and *Cereals* groups, since they are the largest ones and involve very different processes and processed commodities. Furthermore, within the *Fruits* RAC group, *Citrus Fruits* and *Berries and Small Fruits* subgroups were analyzed separately. In Tab. 4-1 three-level RAC partitioning structure is shown, with specific groups and subgroups analyzed in the following chapters.

RAC GROUP	RAC SUBGROUP	RAC
FRUITS (FRESH OR FROZEN) AND TREE NUTS (*4.3.1.3*)	CITRUS FRUITS (*4.3.1.1*)	ORANGES
		LEMONS
		MANDARINS AND SIMILAR
	BERRIES AND SMALL FRUITS (*4.3.1.2*)	CURRANTS (BLACK, RED AND WHITE)
		STRAWBERRIES
		TABLE GRAPES
		WINE GRAPES
	MISCELLANEOUS FRUITS WITH EDIBLE PEEL SMALL (MISC. FRUITS ED. PEEL)	TABLE OLIVES
	MISCELLANEOUS FRUITS WITH INEDIBLE PEEL LARGE (MISC. FRUITS INED. PEEL L)	COMMON BANANA
		MANGOES
		PINEAPPLES
	POME FRUITS	APPLES
		PEARS
	STONE FRUITS	APRICOTS

		CHERRIES (SWEET)
		PEACHES AND SIMILAR
		PLUMS
CEREALS (*4.3.1.4*)	CEREALS	BARLEY GRAINS
		COMMON WHEAT GRAIN
		MAIZE GRAIN
		OAT GRAIN
		RICE GRAIN
		RYE GRAIN

Table 4-1 RAC groups-subgroups dependencies analyzed in 4.3.1. For a complete view of RAC groups and subgroups see Appendix I.

4.3.1.1 Citrus Fruits

Analysis starts at RAC subgroup level and with *Citrus Fruits* subgroup. All data sources were filtered accordingly, including only rows (in case of treemaps and alluvial plots) or only rows and columns (in case of Jaccard distance matrix for J-PCA) including this specific subgroup value. Filtered datasets account for 105 records.

Treemap

Treemap visualization in Fig. 4-2 accounts for unique combinations of PF levels, RAC and processed commodity for *Citrus Fruits* RAC subgroup. Combinations counts are represented by rectangles' sizes, which are then ordered from top-left to bottom-right corner. At this level of detail, treemap returns quite informative visualizations. PF levels, RAC and processed commodities have been included as three partitions. RACs included in citrus fruits subgroup are immediately recognizable: oranges, mandarins and similar, lemons. This is evident in the second partition level, based on RACs. It's also easy to identify the most abundant RAC: oranges. Being far more numerous than the other two RACs, oranges dominate across all PF levels (first partition level). At the third and innermost partition level, focused on processed commodities, greater variation emerges across RACs and PF levels. Processed commodities like marmalade, pulp, juice, pasteurized juice, and canned fruit dominate the green area. In contrast, the red and yellow areas are highly populated by peel keyword.

However, a strong limitation of treemaps is the number of partitions that can be reasonably included so that the final graph does not become unreadable. The Processing Factors database is structured around highly complex and detailed group-subgroup relationships between samples, with the number of connections increasing significantly when moving from specific to general levels (compare Fig. 4-2 with Fig. 4-10). In total, PF database contains eight different categorical variables while on treemap in Fig. 4-2 only three of them are included (i.e., PF level, RAC, processed commodity). This is because a fourth level of partitions on the graph (with processes or pesticides) would have been too much to handle and interpret. The selection of only a few class sets for simultaneous visualization on the chart creates a challenge in identifying potential relationships between RACs, processes, and processed commodities on one side, and the final PF levels on the other. Additionally, even in the citrus subgroup context, some labels are difficult to read, such as 'marmalade' within the *Mandarins and similar* RAC for green PF level.

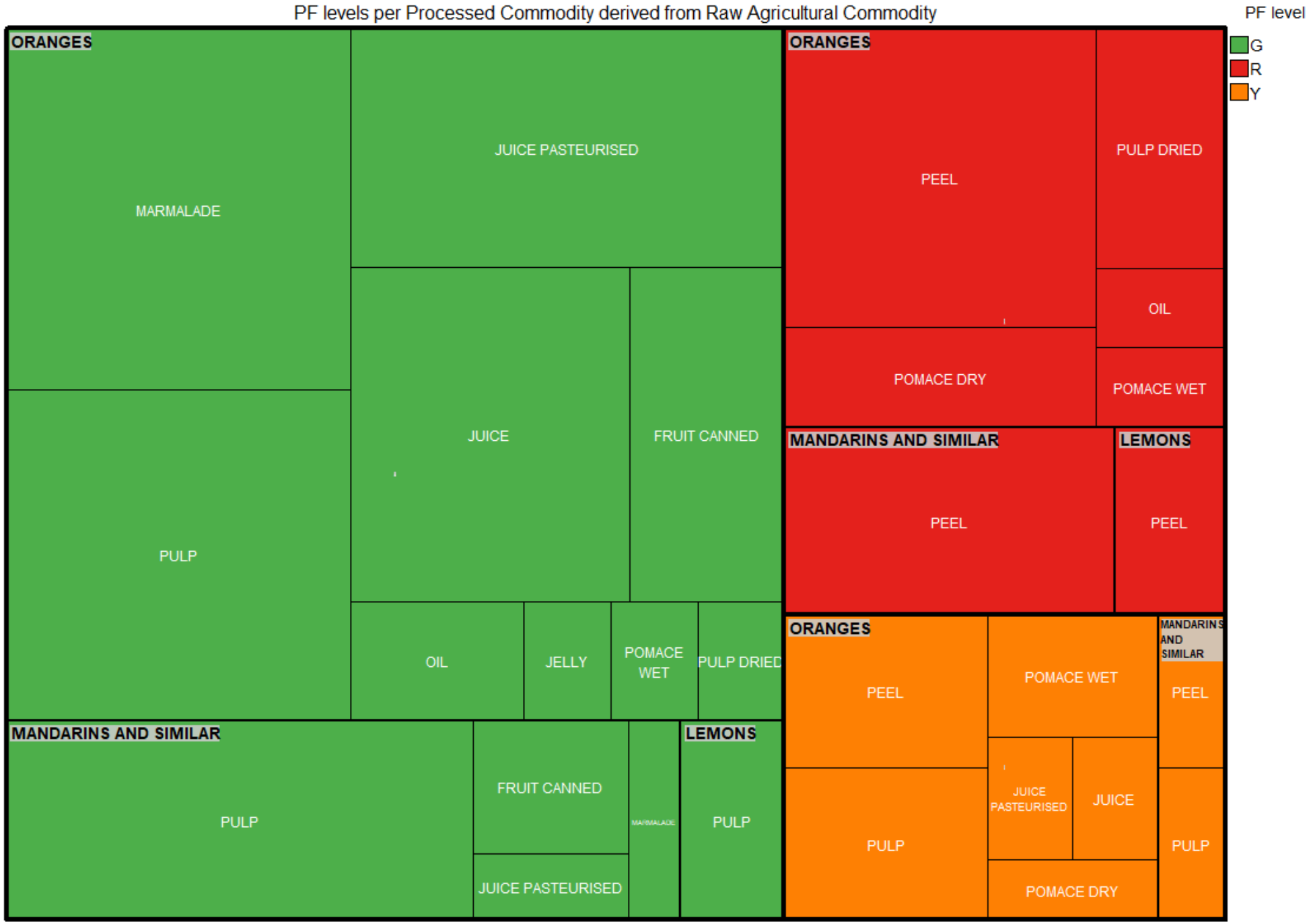


Figure 4-2 Citrus fruits treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RACs and processed commodities.

Alluvial

Alluvial plot allows for the inclusion of all groups and subgroups, logically connected by streams that represent single samples. At this level, six axes are necessary to represent all the categorical variables of PF database that characterize the samples (RAC group and subgroup can be skipped since the filtering applied is on specific RAC subgroup). The alluvial plot in Fig. 4-3 provides a clear overview of the *Citrus* RAC subgroup. Unlike the treemap in Fig. 4-2, the alluvial plot allows for the inclusion of all the categorical variables of interest simultaneously, offering additional layers of information, from specific RACs to pesticide levels, and all other categorizations in the database. On the first axis (RAC), oranges are clearly identifiable as the main commodity, represented by the largest stratum. From this point, the major red, yellow, and green alluvia branch out. The underlying numerical difference between oranges, mandarins and lemons samples is substantial, making it easy to distinguish even though the red and yellow PF levels are more dispersed compared to their grouping in the treemap.

At the process code axis, there is a polarization of red and yellow PF levels into two main processes: no process at all (No) and fruit juice processing (II 001). The By-product (BP) axis collects almost all red and yellow PF levels, derived from peel samples, as also observed in the treemap plot. Other marginal contributions to the red and yellow PF levels are pomace wet, pomace dry (both part of by-product umbrella), pulp dried and essential oil.

One limitation of the alluvial plot is the inability to determine the exact numerical significance of individual alluvia, for instance, whether the red essential oil corresponds to a single sample or multiple ones. We can only compare the relative sizes of the alluvia, which is one of the drawbacks of this approach.

After the PF level axis (equivalent to the first partition level in the treemap), there is a mixed distribution into pesticides. Only a few pesticides are associated with a single PF level: cyhalothrin lambda (CYA), fludioxonil (FLUD), hexythiazox (HEX), and thiophanate-methyl (THIO) are all in the green category, while fosetyl-aluminium (FOS-AL) is entirely in the yellow category.

The alluvial plot provides a comprehensive view of the *Citrus Fruits* subgroup, allowing for the comparison of each variable and offering an overall perspective on PF level trends. However, this broader view comes at the expense of detail for individual alluvia (equivalent to single samples). The bending, crossing, separation, and rejoining of the alluvia create visual noise and can lead to confusion, making it nearly impossible to trace a single alluvium from the first to the last axis.

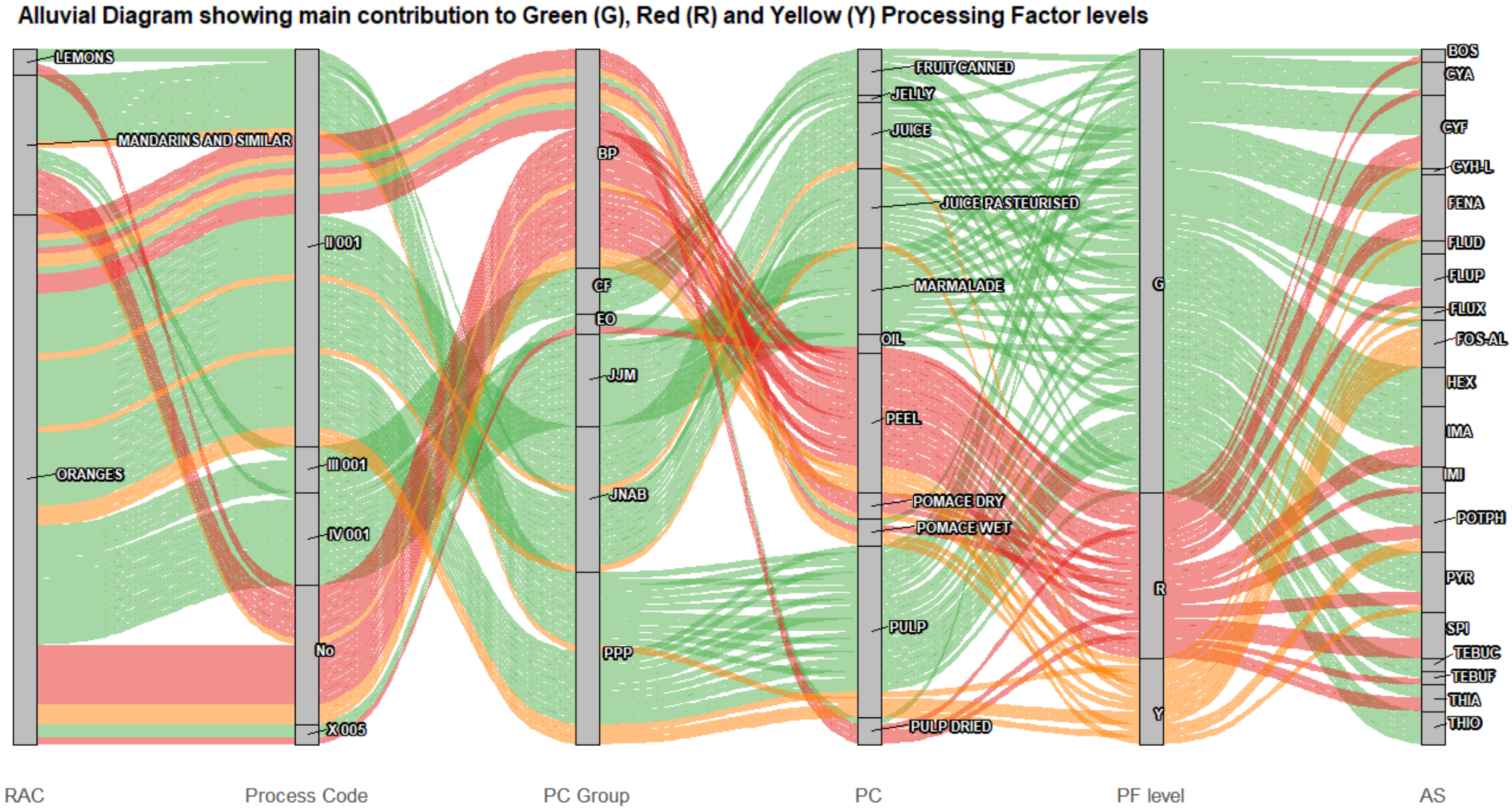


Figure 4-3 Citrus fruits alluvial plot: visualization according to RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

Taking a closer look at the *Citrus Fruits* subgroup using a multivariate approach (Fig. 4-4), a J-PCA model with two principal components was calculated on the corresponding Jaccard distance matrix. The data were pre-treated using mean centering, and the model explained 65% of total data variance. In this analysis, sample visualization is performed in the J-PCA model space using combinations of colors and labels, representing different categories at the individual sample level. The PC1-PC2 space primarily separates pulp samples (positive PC1 and PC2 scores) from peel samples (negative PC1 scores). Pulp samples and juices are notably separated along PC2, while most marmalade samples are positioned at the farthest negative PC2 scores. The remaining processed commodities are more mixed and concentrated in the central region of the score plot. No particular pattern was highlighted with specific RACs.

The group of peel samples at negative PC1 scores corresponds to pesticide residues measured directly on fresh citrus peels. This forms the red/yellow PF hotspot within this RAC subgroup, followed by other samples always included in the by-products categorization, such as pomace (wet and dry) and dried pulp. These latter are derived from fruit juice processing (compare Fig. 4-4 A and B). This pattern was already evident from both the treemap and alluvial plots.

Combining class color information with labels clarifies the key distinction between cluster of peel samples and other processed commodities. Citrus peels are indeed reasonably included in other processed commodities, like for example the well-defined cluster of marmalade samples at negative PC2 scores. Nevertheless, all marmalade samples end up being in green PF level, while all peel samples are in red PF level. Labels unveil the difference in processing for these two classes of samples: marmalade is produced through the "IV Other fruit products" process, which is also used for jams and jellies, while group of peels undergo no processing at all (labeled "No process"). This information is crucial to extend scope of questions presented in Section 4.3.1: it is not only about comparing process A versus process B for PF level categorization, but also about considering the total absence of a transformation process, which seems to cause higher occurrence of yellow and red PF levels.

The higher occurrence of red PF values in untreated citrus peel samples may be attributed to the well-documented phenomenon of higher concentration of pesticides residues in fruit peel in comparison to pulp, as reported in several studies: Cabras et al., 2000 on grapes, Calvaruso et al., 2020 on citrus fruits, Han et al., 2015 on pears. In general, this trend is also confirmed by Scholz et al., 2022.

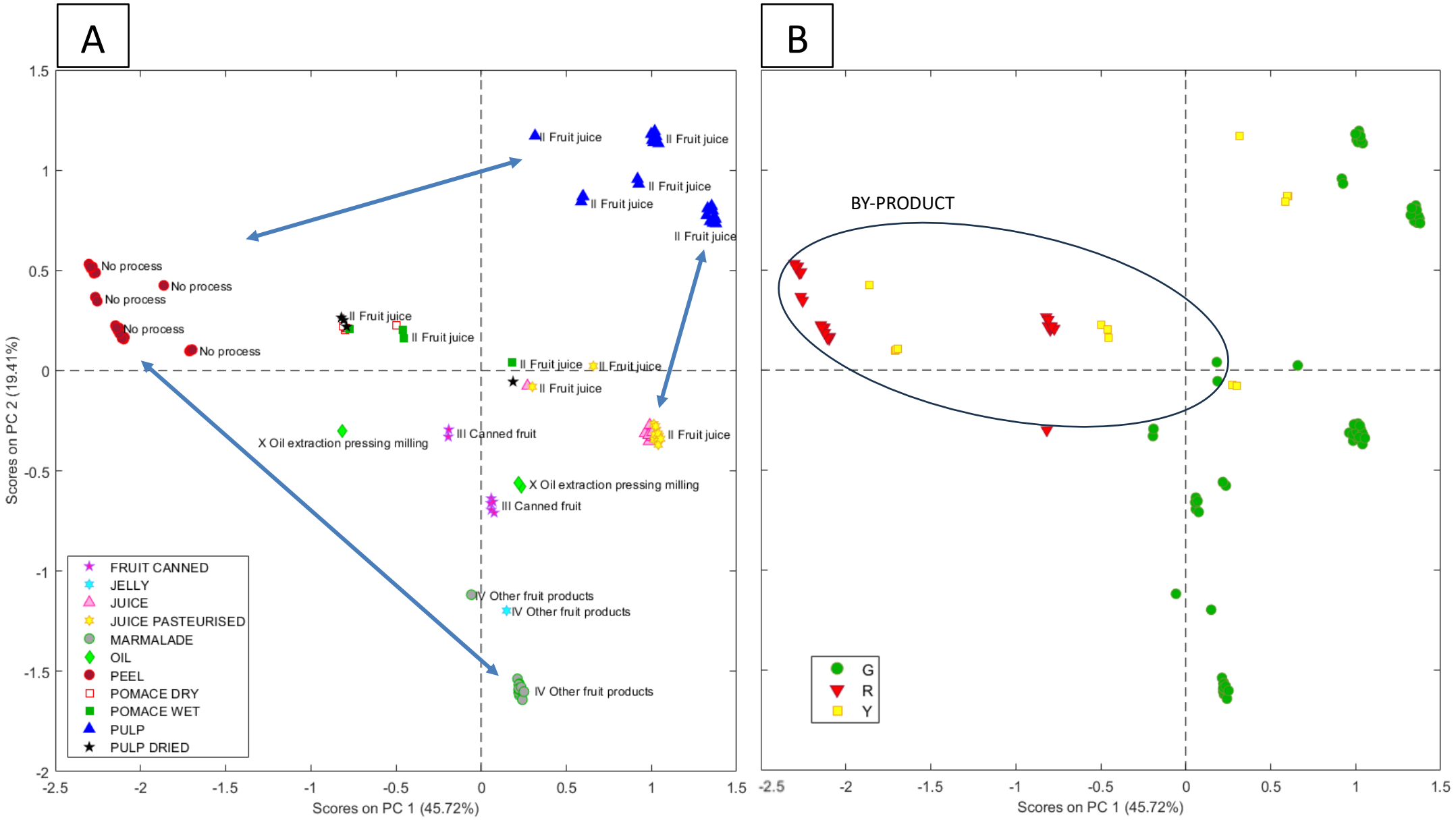


Figure 4-4 Citrus Fruits J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the processed commodity and labels indicate the process code, while in B) samples are colored according to the PF level (G - green, R - red, Y - yellow).

In Fig. 4-5 samples are colored according to their actual PF median values within the same PC1-PC2 space. Since PF levels are derived from these median values, it is unsurprising that this perspective aligns with Fig. 4-4 B. However, this visualization is particularly useful to identify the sample with the highest PF median value, which is an essential oil sample (circled in Fig. 4-5). This layer of detail would have been more difficult to extract from a treemap or alluvial plot. An interesting observation is that citrus essential oils are extracted from fruit peel. This highlights the importance of domain knowledge when interpreting data, as it directly connects the highest PF median value with the citrus peel samples identified earlier.

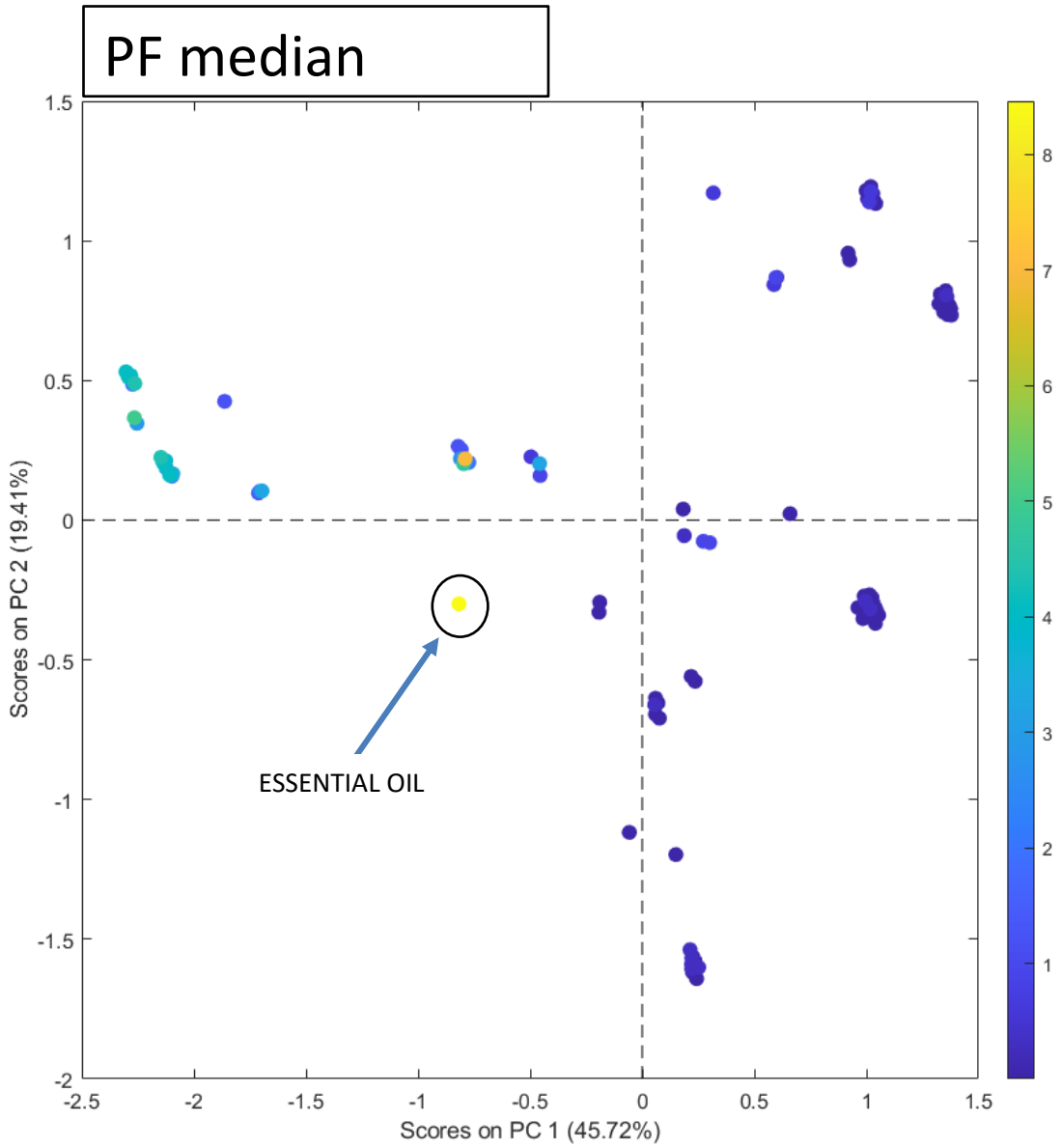


Figure 4-5 Citrus fruits J-PCA: PC1 and PC2 score plot. Samples are colored by median PF value.

At this level of granularity, J-PCA provides little additional information compared to the treemap, or even more so, the alluvial plot. However, median PF gradient visualization proved helpful in identifying the sample with highest PF median value.

Furthermore, an additional advantage of J-PCA is that it retains the single-sample detail level, and this enables a detailed tracking of all categories. As an example, details of pesticide residues for yellow PF samples in pulp and juice processed commodities are shown in Fig. 4-6 A and B, respectively. These figures highlight

the persistence of fosetyl-aluminum (FOS-AL) in yellow PF level both for pulp and juice samples, both in oranges and mandarins. On the other hand, it also brings evidence of how two pesticides, fenazaquin (FENA) and potassium phosphonates (POTPH), pass from yellow PF level in pulp orange samples to green PF level in juice samples. Tracing back the samples to the specific pesticide residue is possible with J-PCA while it is very challenging with treemap or alluvial plots.

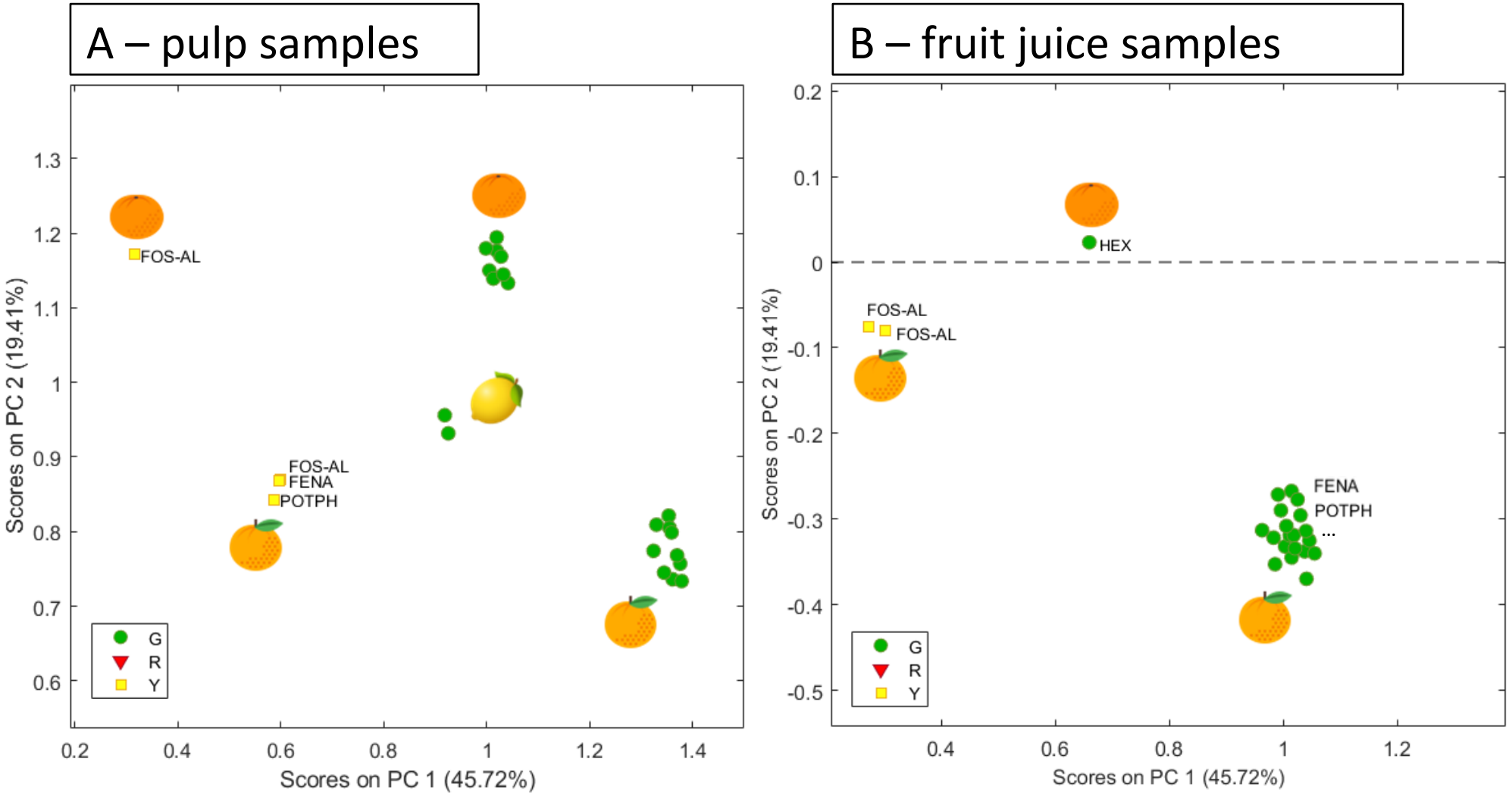


Figure 4-6 Citrus Fruits J-PCA: detail of PC1 and PC2 score plot for pulp samples (A) and fruit juice samples (B). Samples are colored according to the PF level (G - green, R - red, Y - yellow) while labels indicate the pesticide and icons (from Icons8) indicate the RAC.

4.3.1.2 Berries and Small Fruits

Berries and Small Fruits form a RAC subgroup, like the previously explored *Citrus Fruits* subgroup. The filtered dataset is comparable in terms of number of records included in the analysis, with 118 samples.

Treemap

Treemap visualization in Fig. 4-7 accounts for unique combinations of PF levels, RAC and processed commodity for *Berries and Small Fruits* RAC subgroup. Combination counts are represented by rectangles' sizes, which are then ordered from top-left to bottom-right corner. In this treemap, yellow represents the predominant PF level, with wine grapes as the main RAC, along with all related winery products (red, white, and rosé wine) and intermediates (must), which are classified as processed commodities. Wine grapes and their derivatives are also present in the green PF level, and when comparing the two, the green rectangle is larger, suggesting a higher amount of wine grapes records falling in the green PF level rather than in the yellow one. Including the red PF quota for wine grapes along with the yellow introduces some uncertainty when trying to compare the total of these two with the green. This highlights one of the drawbacks of treemap visualizations: it becomes difficult to visually sum and compare multiple areas when they are similar in size. The same issue applies to table grapes and strawberries RACs, especially since their areas have significantly different width-to-height ratios across the three PF levels.

For the red PF level, the largest RAC is table grapes as the raisins processed commodity. Raisins are also the largest yellow-processed commodity within the table grapes' yellow PF quota. Notably, no raisin samples are recorded in the green PF level.

For the RAC currants, it's clear that the combined yellow and red quota is larger than the green. However, this is primarily because the yellow portion is significantly larger than both the red and green.

The general limitations noted for the citrus fruits subgroup regarding the maximum number of variables that can be included are present here as well. Additionally, some labels are difficult to read also in this case (e.g., juice pasteurized from currants in yellow PF level).

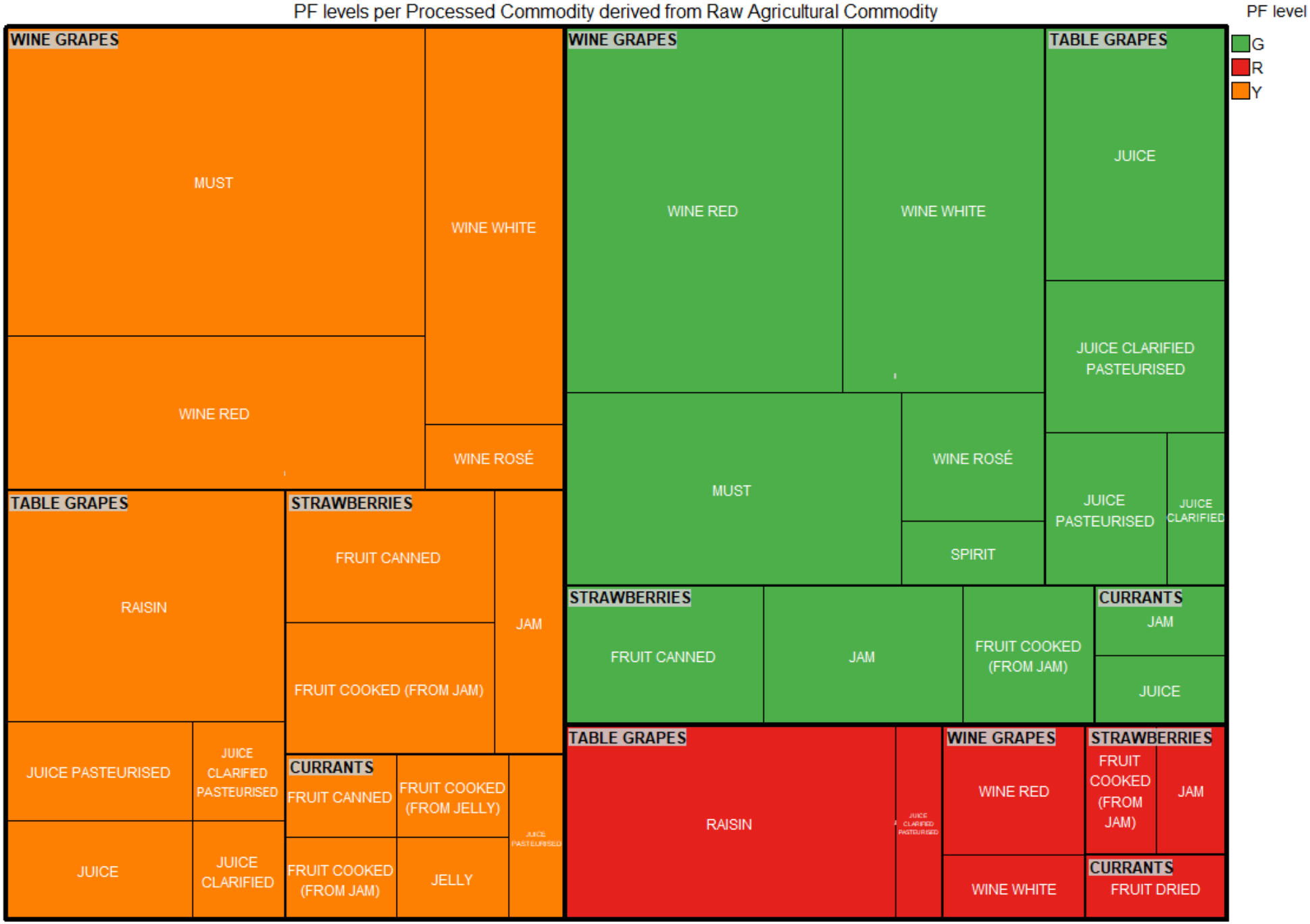


Figure 4-7 Berries and Small Fruits treemap: visualization according to PF levels, RACs and processed commodities.

Alluvial

Alluvial plot allows for the inclusion of all groups and subgroups, logically connected by streams that represent initial samples. At this level six axes are necessary (as for *Citrus Fruits* case, RAC group and subgroup cluster can be skipped since are known a priori). The alluvial plot in Fig. 4-8 illustrates the dominant trends in *Berries and Small Fruits* RAC subgroup. The largest red alluvium originates from table grapes that, together with a small red contribution from currants, flows into the XVI 001 fruit dehydration process stratum, resulting in dried food, particularly the raisin processed commodity. Although this same alluvium has a considerable yellow portion, it is much smaller compared to the three large yellow alluvia originating from the wine grapes RAC. These flow through fermentation processes (primarily V 001 White wine production and V 002 Red wine production) resulting in must, red, and white wine, which represent the top three steps on the yellow PF podium, all under the alcoholic beverages (AB) processed commodity group.

Canned food from strawberries shows a relatively balanced distribution of yellow and green PFs: since the process is common too (*III 001 Canned fruits berries and small fruits*), the difference between yellow and green cases should rely on different pesticides residues, which are difficult to identify in the final alluvial axis. The situation in the jam, jelly, marmalade, and juices/non-alcoholic beverages groups is more variable in terms of PF levels.

Since yellow PF is very prevalent within this RAC subgroup, several pesticides show a high occurrence at this level. Notably, fosetyl-aluminum (FOS-AL) stands out, as it was also found exclusively in the yellow PF category in the citrus fruits subgroup. In contrast, there is only one pesticide that appears entirely in the green PF category: pyraclostrobin (PYR).

The same drawbacks already mentioned for *Citrus Fruits* subgroup are applicable to this alluvial plot as well.

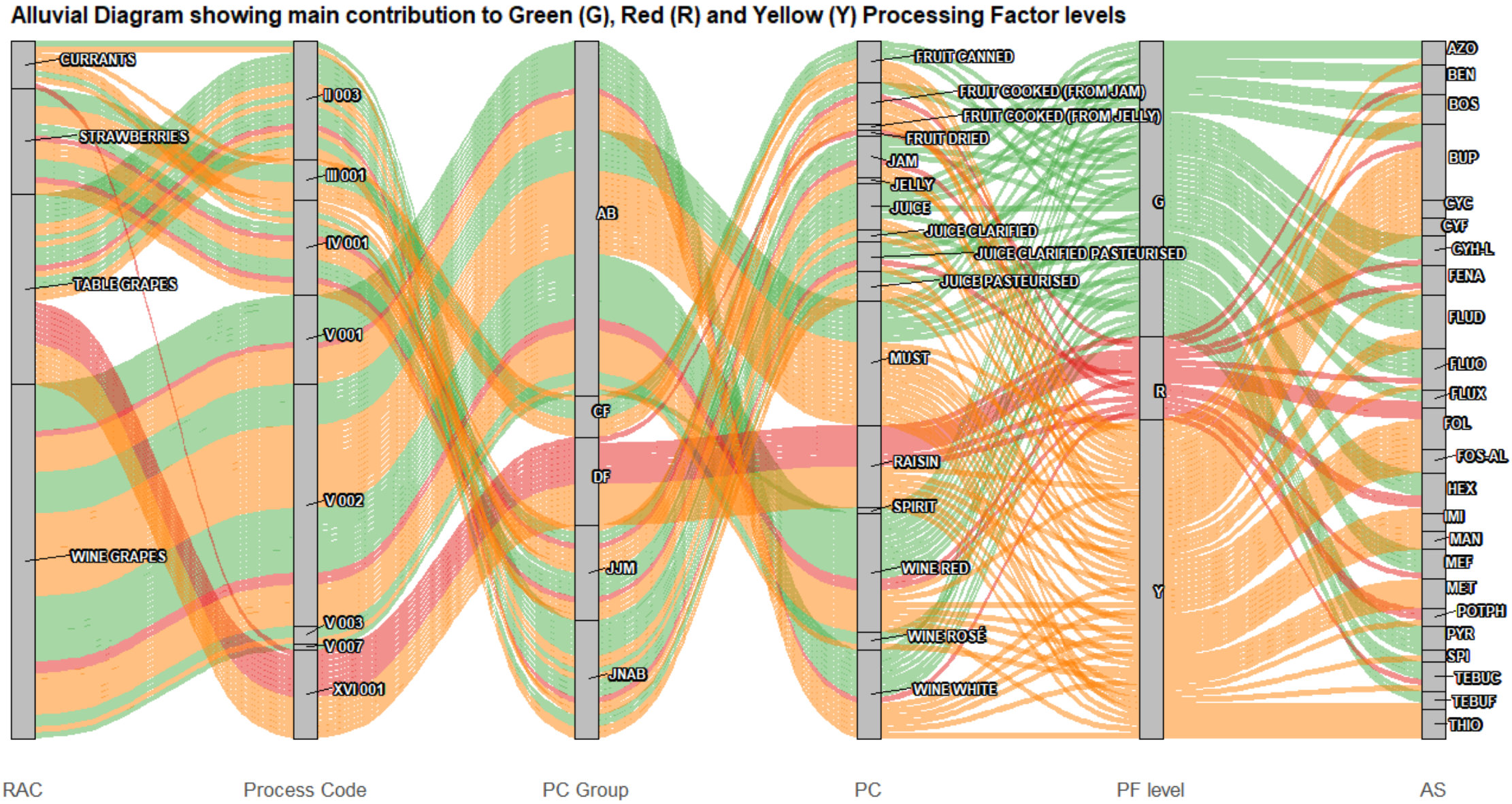


Figure 4-8 Berries and Small Fruits alluvial plot: visualization according to RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

A local J-PCA model with 2 principal components and mean centering as pre-treatment was calculated, explaining around 67% of total data variance. In the PC1-PC2 space (Fig. 4-9), processes that are highly divergent are well-separated. For instance, the V Fermentation process is positioned at negative PC1 scores, while the II Fruit Juice process is located on the opposite side. In contrast, processes that share similarities in their flowcharts and final products are grouped closer together. This is evident in the III Canned Food and IV Other Fruit Products processes, which produce canned fruit and jams/spreads/jellies, respectively.

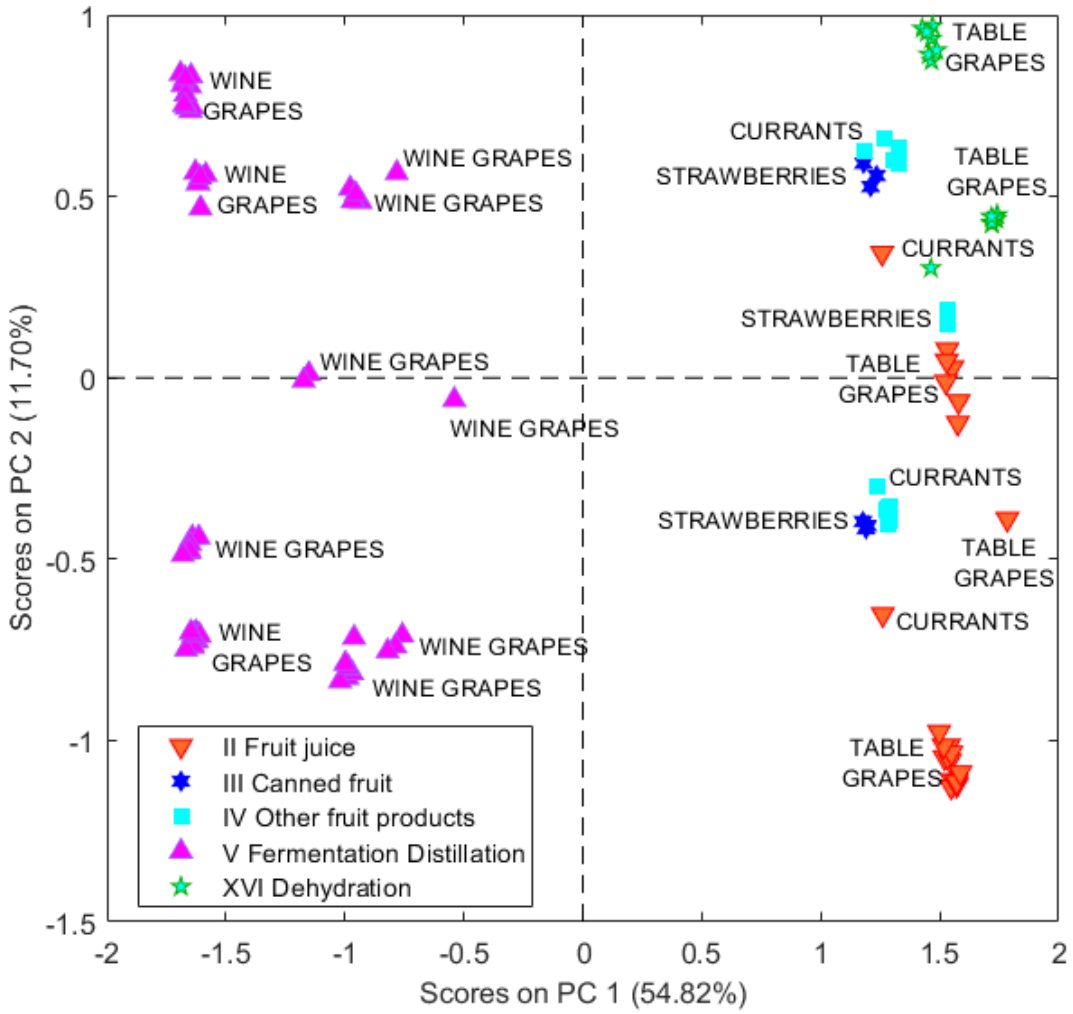


Figure 4-9 Berries and Small Fruits J-PCA: PC1 and PC2 score plot. Samples are colored according to the OECD process and labels indicate the raw agricultural commodity.

PF information is distributed along PC2 (Fig. 4-10 A), with values ranging from green at negative PC2 scores to yellow and red starting around zero and extending to positive PC2 scores. The fermentation process, previously discussed, begins with a single raw material, wine grapes, and results in a highly variable green/yellow/red PF distribution within the same RAC. Fig. 4-10 B illustrates that fermentation produces several different products, such as must and red, white, or rosé wine, but this does not explain the variability in PF levels either, as all processed commodities display the full range of PF values. Differences among these samples will be further analyzed in Section 4.3.2.1.

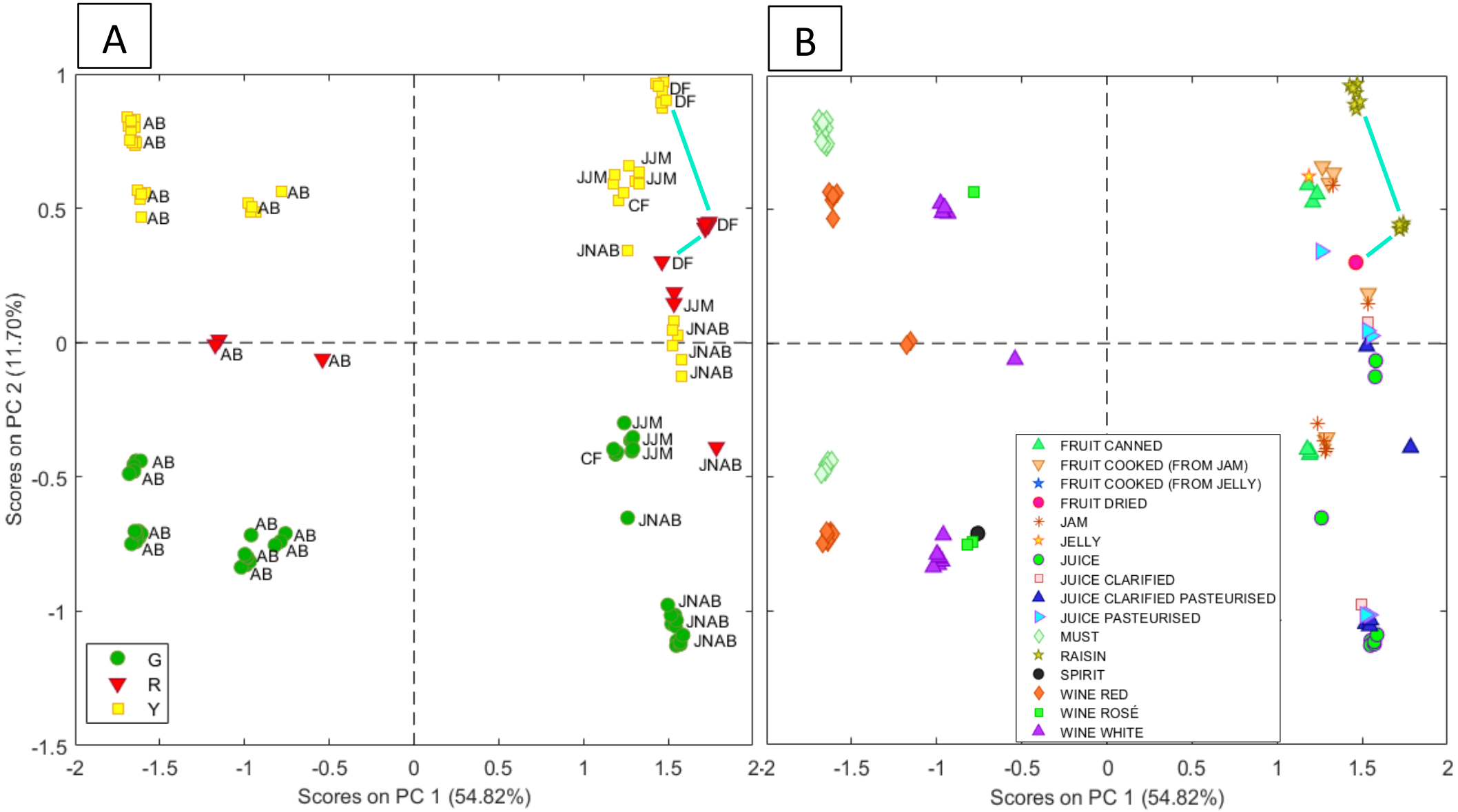


Figure 4-10 Berries and Small Fruits J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the PF level (G - green, R - red, Y - yellow) and labels indicate the processed commodity groups (see Appendix III) while in B) samples are colored according to the specific processed commodity.

The same interpretation applies to most samples at positive PC1 scores: the same RACs, processes, and final products appear in both the green and yellow/red regions of the PC1-PC2 space without a discernible pattern. However, one notable exception in this context is the XVI Dehydration process, which shows PF levels only concentrated in the yellow/red area, at positive PC1-PC2 values. This is represented by three clusters of samples connected by a line in Fig. 4-10 A and B. This effect persists even when the process is applied to different RACs, such as table grapes and currants, to produce different commodities, like raisins and dried fruit, all part of the dried food (DF) processed commodity group. In this case, it is clear how a specific transformation process can influence PF levels.

Also in this case, J-PCA provides little additional information compared to the treemap, or the alluvial plot previously showed. Nevertheless, we are starting to gather information from different points of view within the database to answer key questions posed in Section 4.3.1. The Citrus Fruits subgroup analysis previously discussed in Section 4.3.1.1 indicated that no specific process was particularly biased towards yellow and/or red PF levels; rather, it was primarily the final product or, better to say by-product, considered. At the same time, it revealed that even unprocessed foods could be problematic (e.g., the fresh raw citrus peel samples). In contrast, the berries analysis reveals that, probably, a problematic process exists across RACs and final processed commodities: dehydration. These assumptions need to be confirmed with a broader analysis.

4.3.1.3 Fruits (fresh or frozen) and Tree Nuts

Moving from the analysis at the RAC SUBGROUP level to the RAC GROUP level (*see* Table 4-1), the focus now shifts to the *Fruits (fresh or frozen) and Tree Nuts* RAC group. All data sources are filtered accordingly, including only relevant rows for treemap and alluvial approaches or both rows and columns for J-PCA. The filtered dataset contains 369 records.

Treemap

Treemap visualization in Fig. 4-11 accounts for unique combinations of PF levels, RAC subgroup and processed commodity for all *Fruits (fresh or frozen) and tree nuts* RAC group. Combination counts are represented by rectangles' sizes, which are then ordered from the upper left corner to the lower right corner. Focusing on outermost rectangles separation, the one representing three PF levels, we can notice that the majority of samples fall in the green level, since the thicker vertical split defining the green area occupies more than half of the external rectangle's base. We can also get general proportions with other PF levels: given the overall green area, yellow is about half of it and red one is about two-thirds of yellow. Looking at inner separation grid though, comparison between specific RAC subgroups (e.g., *Citrus Fruits*, *Berries and Small Fruits*) is more difficult since the same RAC subgroups appear in different PF categories. Furthermore, these rectangular areas are not adjacent like the outer ones, so they must be visually extracted from their position and mentally compared. While this is still feasible with one-to-one comparison of significantly different areas (e.g., for citrus fruits the green rectangle is larger than the red one), this becomes very difficult when trying to sum up two areas to compare with a third one, or when areas are very similar or, on the opposite, very different in their width/height ratio. For example, from a numerical point of view half of the records of stone fruits have green PF level. As a consequence, the sum of the corresponding green area in the treemap is equivalent to the sum of red and yellow areas of stone fruits. However, it is not straightforward to retrieve this information looking at the treemap. In principle counts (or percentages) of different partitions could be shown on the graph, but this would mean showing even more labels, making the visualization of the treemap more confusing since some labels are already unreadable at this stage (e.g., miscellaneous fruit with edible/inedible peel). Some kind of interactivity on the graph could help in this direction but this is out of scope, since here we are comparing static views for each different methodology.

In conclusion, at this level, treemap still allows to capture general trends with partial views, but overall connections with other groups and subgroups are lost due to the well-known readability constraints which do not allow to include all variables at the same time. Here, an attempt to address this limitation is made: multiple treemaps with different partitioning schemas are used simultaneously (one for processed commodities, as in Fig. 4-11, and another for processes, as in Fig. 4-12). However, the trade-off between informative content and noise is problematic, as the relationships between groups and subgroups remain unclear. This is because the partitioning schemas on the graphs depend solely on the counts of each unique combination. For example, process *II 001 Citrus juice, citrus fruit* in Fig. 4-12 is not positionally aligned with pulp, juice pasteurized, and juice processed commodities in Fig. 4-11, even though they are related. This last observation applies to all previous and future cases of treemap use.

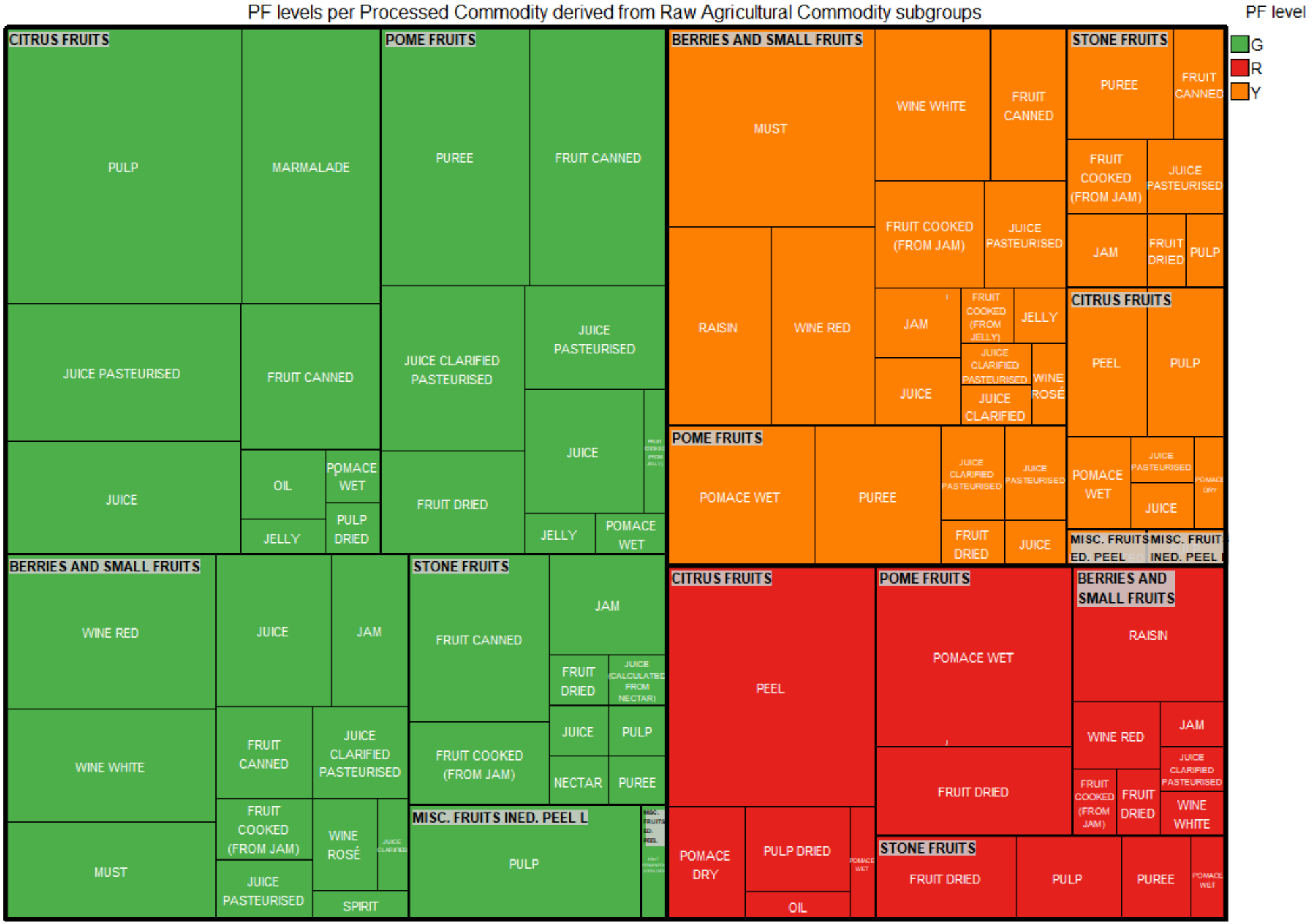


Figure 4-11 Fruits (fresh or frozen) and Tree Nuts treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RAC subgroups and processed commodities.

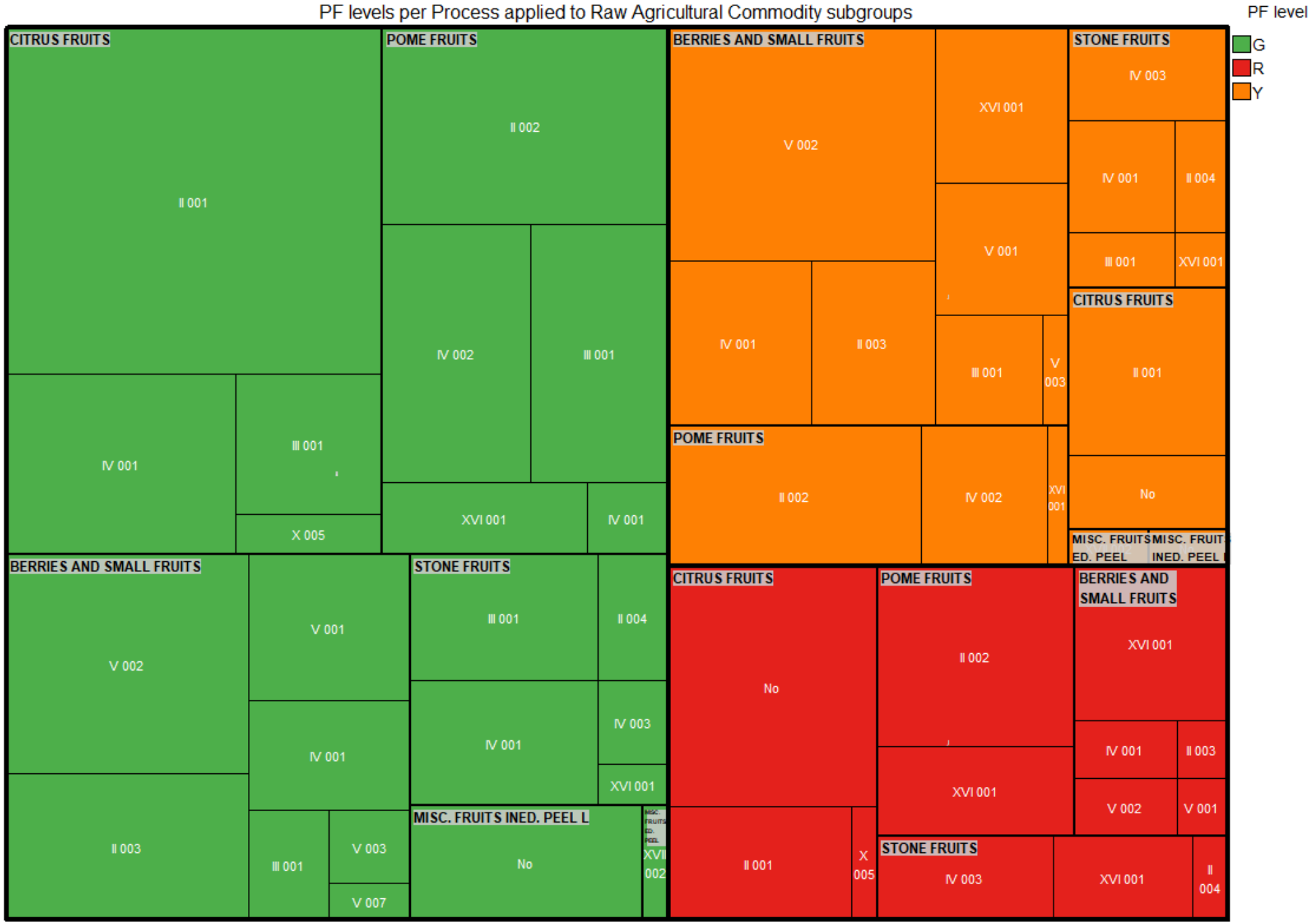


Figure 4-12 Fruits (fresh or frozen) and Tree Nuts treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RAC subgroups and process codes (see Appendix II).

Alluvial

Also in this case alluvial plot allows for the inclusion of all groups and subgroups, logically connected by streams that represent initial samples. At this level, an additional axis corresponding to RAC subgroup is required, bringing the total to seven and making the graph harder to follow. Furthermore, the increased number of samples to be displayed implies a higher number of alluvia in the plot, that has a negative effect on plot definition.

The representation in Fig. 4-13 effectively captures the internal relationships from RACs to final processed commodities, including processes and pesticides. The main streams representing different PF strata can be identified by first examining the PF axis and then tracing back to the PC and PC group axes on the left. The key streams feeding into the red PF level include fruit dried, raisin, peel, pomace dry, and pomace wet, which originate from the by-product (BP) and dried food (DF) processed commodity groups. However, as the eye moves further from the PF axis toward the RAC subgroup axis, it becomes difficult to maintain a clear perception of the streams' origins due to their subsequent splits, crossings, and reconnections at each axis. Additionally, the bending of streams distorts their sizes, and the color coding can only represent one grouping class at a time (in the case of Fig. 4-13, this is PF levels). As a result, it becomes difficult to distinguish the partitions of other groups and subgroups, with strata sizes being the only indicator. Also, some labels overlap with axes, increasing overall confusion and decreasing readability. In conclusion, the view on specific axes (e.g., RAC subgroups, RAC, pesticides) is cluttered with noise.

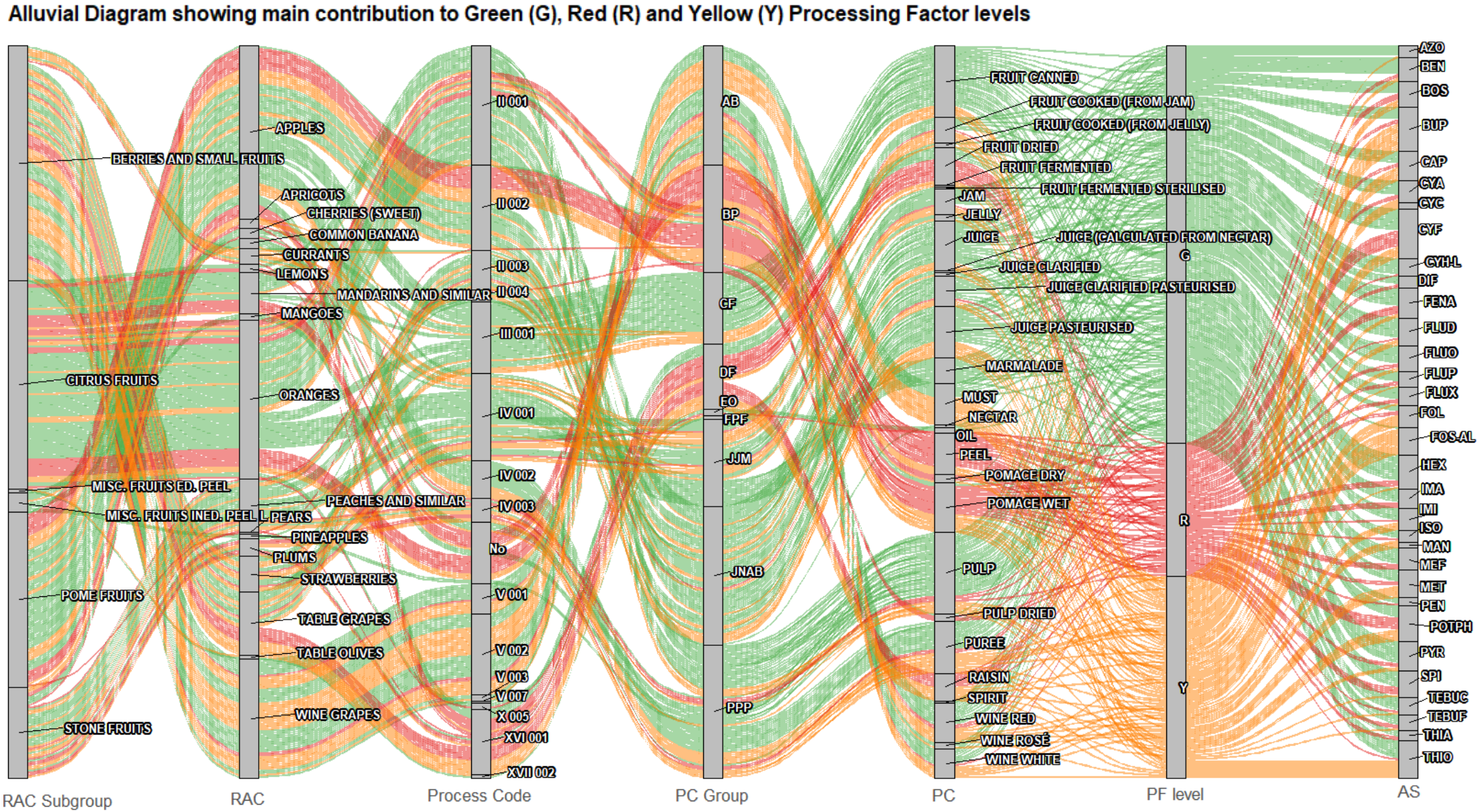


Figure 4-13 Fruits (fresh or frozen) and Tree Nuts alluvial plot: visualization according to RAC subgroups, RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

This final visualization approach for the fruits RAC group aims to overcome the limitations of the treemap and alluvial methods. The goal is to create a representation that preserves the level of detail required for specific groups and subgroups, while also offering insight into the overall relationships between them. This is achieved through a J-PCA model using three principal components, with mean centering as the pre-treatment, explaining approximately 60% of total data variance. By examining a combination of score plots, information is extracted in synergy, minimizing redundancy and noise.

Fig. 4-14 reports the PC1 and PC2 score plot where the objects are colored according to RAC subgroup (Fig. 4-14 A) and single RACs (Fig. 4-14 B). RAC subgroups can be clearly identified (Fig. 4-14 A) and within each subgroup, predominant raw agricultural commodities can be visualized (4-14 B). *Berries and Small Fruits* is the largest subgroup, with wine grape being the prevalent RAC. In particular, wine grapes samples lie at the most negative PC1 score values, separated from the other *Berries and Small Fruits* samples, as well as all other samples,: this reflects the very different process - processed commodity (V Fermentation – alcoholic beverages) effect already seen in Fig. 4-9 and 4-10 with J-PCA on specific *Berries and Small Fruits* subgroup. *Citrus Fruits* subgroup is the second largest subgroup, with oranges being the most abundant RAC, followed by *Pome Fruits* and *Stone Fruits* subgroups. More in detail, apples and peaches and similar are the largest RACs within *Pome Fruits* and *Stone Fruits* subgroups, respectively. This type of identification was much simpler using J-PCA visualization than with treemaps or alluvial. No PF information can be observed in PC1-PC2 space since its levels appear to be equally spread in RAC subgroups (Fig. 4-15), so at least for fruits, it can be said that type of RAC influence PF levels to a lesser extent than already observed in the analysis of *Citrus Fruits* and *Berries and Small Fruits*.

The third principal component comes into play in Fig. 4-16, where PC1 is reported against PC3. In Fig. 4-16 A, RAC subgroups are displayed with OECD process code labels, while in Fig. 4-16 B PF levels are reported with processed commodity group labels. Information on PF levels is now included in the picture, stretching samples along PC3: samples with green PF level are mainly located at negative PC3 values while samples with red or yellow PF levels generally have positive PC3 scores. This aligns with a higher occurrence of by-products (BP), derived from the II Fruit Juice process and the 'No Process' category, and dried food (DF), derived solely from the XVI Dehydration process, at positive PC3 scores. Such labels occur less frequently at negative PC3 scores, with only a few exceptions, such as some PF green samples of dried pome fruit and citrus by-products. Therefore, the assumptions made for the citrus and berries subgroups can indeed be extended to the entire fruits group: a relationship exists between the RAC, the type of process it undergoes, and/or the final processed commodity, and this relationship appears to influence PF levels in at least two distinct ways. First, process-wise with dehydration, and second, processed commodity-wise with by-products. This graph also highlights the set of alcoholic beverages (AB) samples in yellow and red PF classes (already discussed in Section 4.3.1.2 *Berries and Small Fruits*) but also a considerable yellow/red cluster of puree pulp paste samples (PPP) in stone fruits RAC subgroup, and another yellow cluster of PPP in pome fruits RAC subgroup, both derived from the OECD IV Other fruit process family.

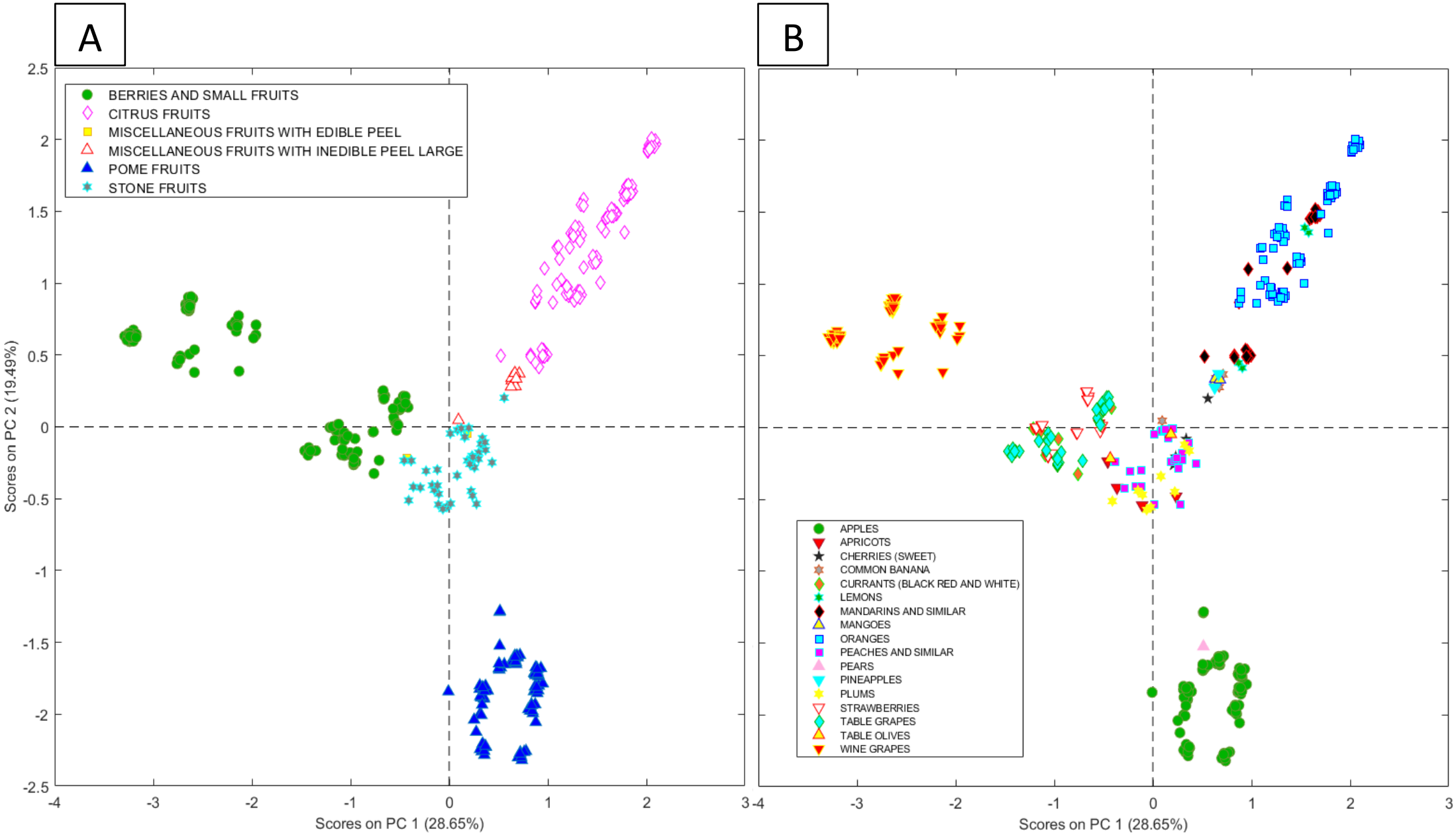


Figure 4-14 Fruits (fresh or frozen) and Tree Nuts J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the RAC subgroup while in B) samples are colored according to the specific RAC.

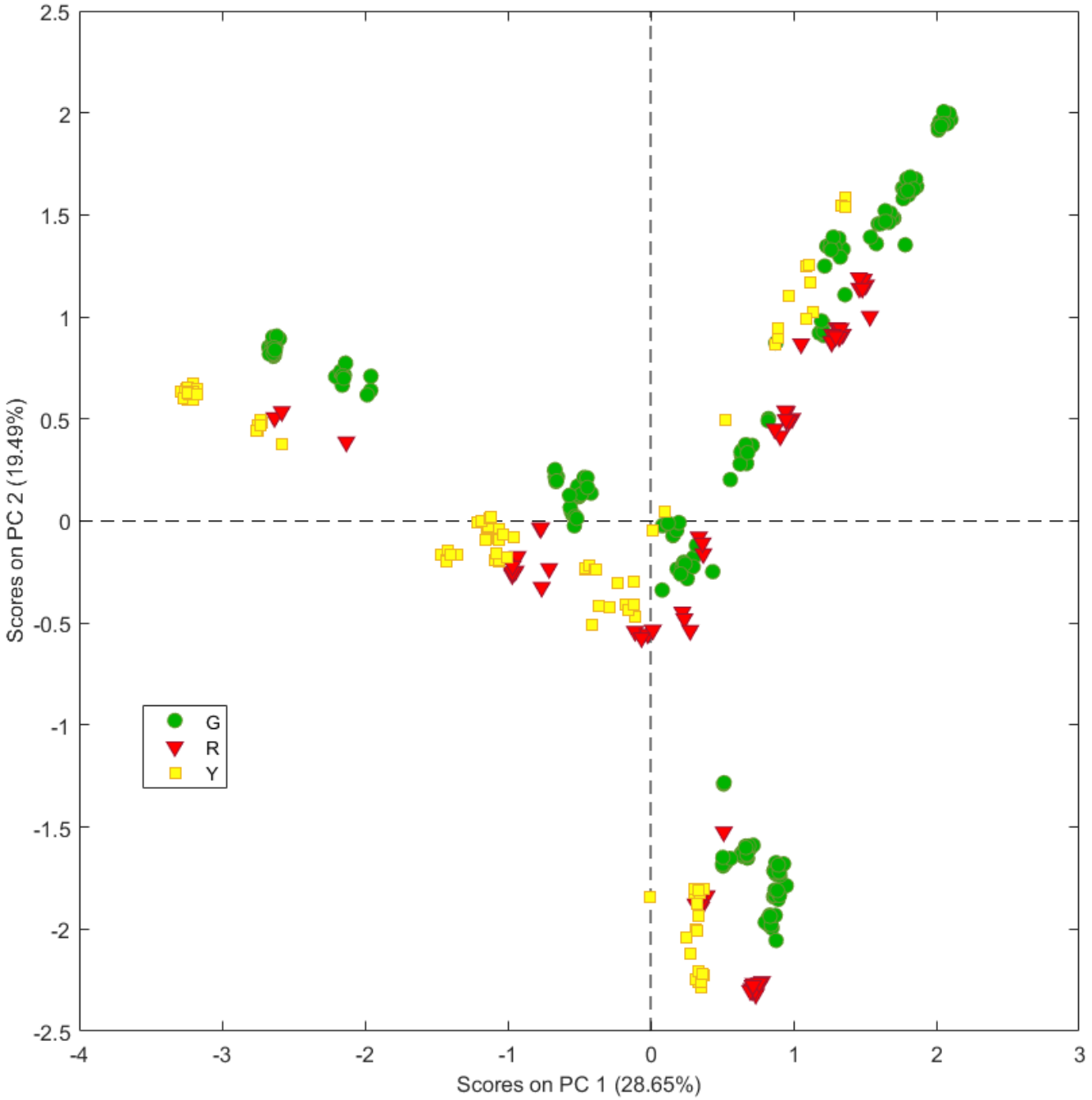


Figure 4-15 Fruits (fresh or frozen) and Tree Nuts J-PCA: PC1 and PC2 score plot. Samples are colored according to the PF level (G - green, R - red, Y - yellow).

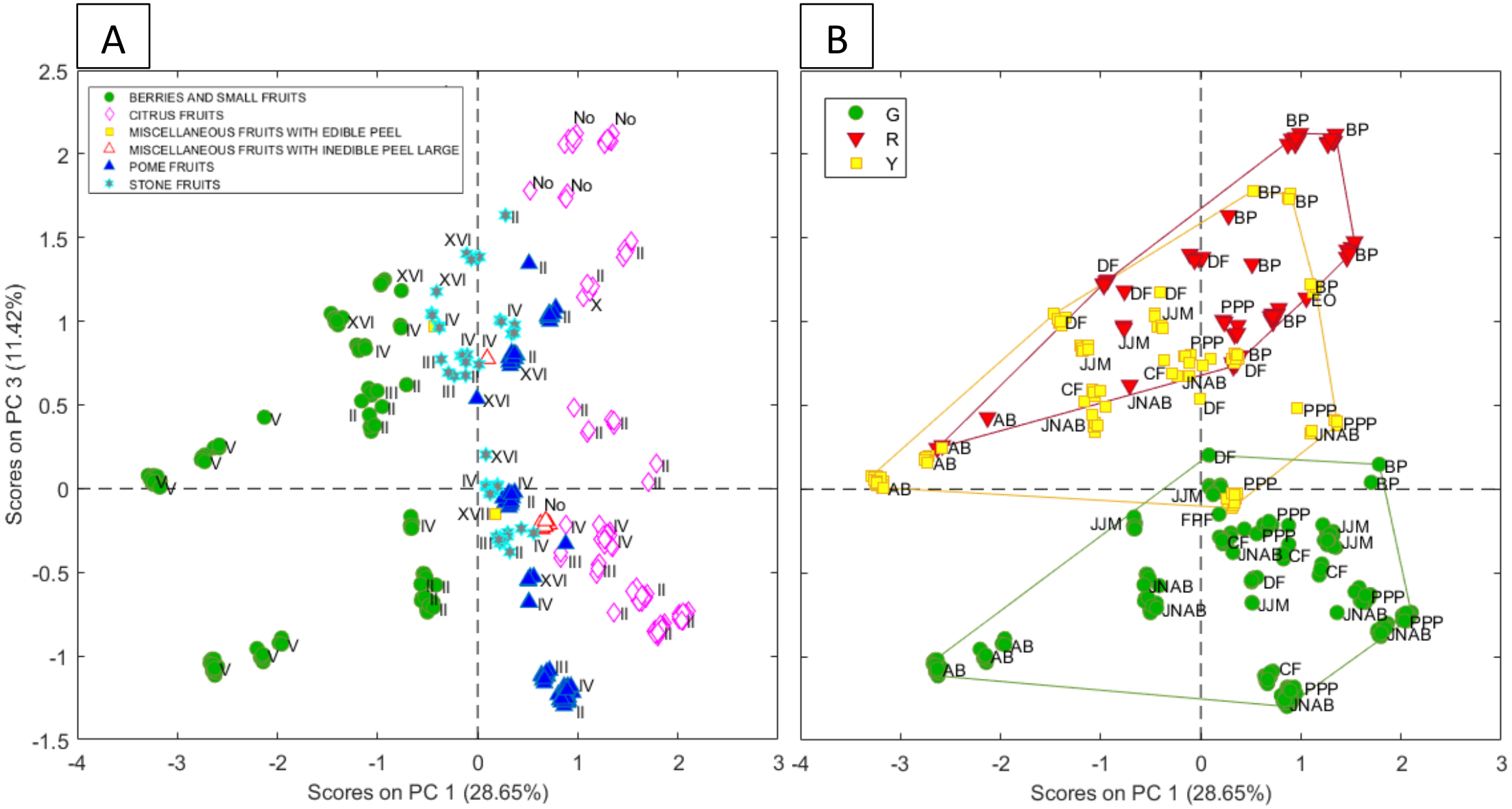


Figure 4-16 Fruits (fresh or frozen) and Tree Nuts J-PCA: PC1 and PC3 score plot. In A) samples are colored according to the RAC subgroup and labels indicate the OECD process (see Appendix II) while in B) samples are colored according to the PF level (G - green, R - red, Y - yellow) and labels indicate the processed commodity groups (see Appendix III).

4.3.1.4 Cereals

So far, only a portion of the entire dataset has been explored, and it is important to confirm the general trends found for the *Fruits (fresh or frozen) and Tree Nuts* group with other RAC groups, especially those that differ significantly in chemical and biological terms. Cereals emerge as a strong candidate for this purpose, given their size in the dataset (133 samples). In the following plots, all data sources are filtered accordingly.

Treemap

The largest areas of the treemap in Fig. 4-17 are occupied by processed commodities derived from common wheat and barley grains, underscoring their prevalence in this group. Common wheat is associated with a wide range of processed products, such as flour (white, wholemeal), gluten, and starch, which span multiple PF levels. Barley grains also occupy substantial space, linked to processed commodities like malt, beer, and flour.

The green PF level is primarily occupied by commodities such as flour, starch, and various gluten products derived from several RACs, indicating that these products are more numerous in this section. Barley and wheat products are notably present here, alongside a few from oats and maize. For instance, white flour, malt and beer frequently appear in the green PF zone. The yellow PF level includes processed products like bran, wholemeal flour, and germ. Although these commodities occupy smaller areas compared to the green section, they still represent a significant portion of the products derived from wheat, maize, and rye grains. In this section, barley shows an opposite trend compared to the green PF level: fewer occurrences of beer and a higher presence of malt sprouts. The red PF level is densely populated by-products like bran (from wheat, barley and rice), and hulls (from oat). These products exhibit the highest PF levels, covering a substantial portion of the right-hand side of the treemap. Bran is particularly prominent in this section, especially from common wheat and barley. When comparing areas across PF levels, the green zone dominates in terms of both variety and size, particularly for processed commodities like flour, malt and beer. In contrast, the red zone contains fewer products but larger rectangles for by-products like bran and germ, indicating that the concentration of pesticide residues during processing is more frequently observed in these items.

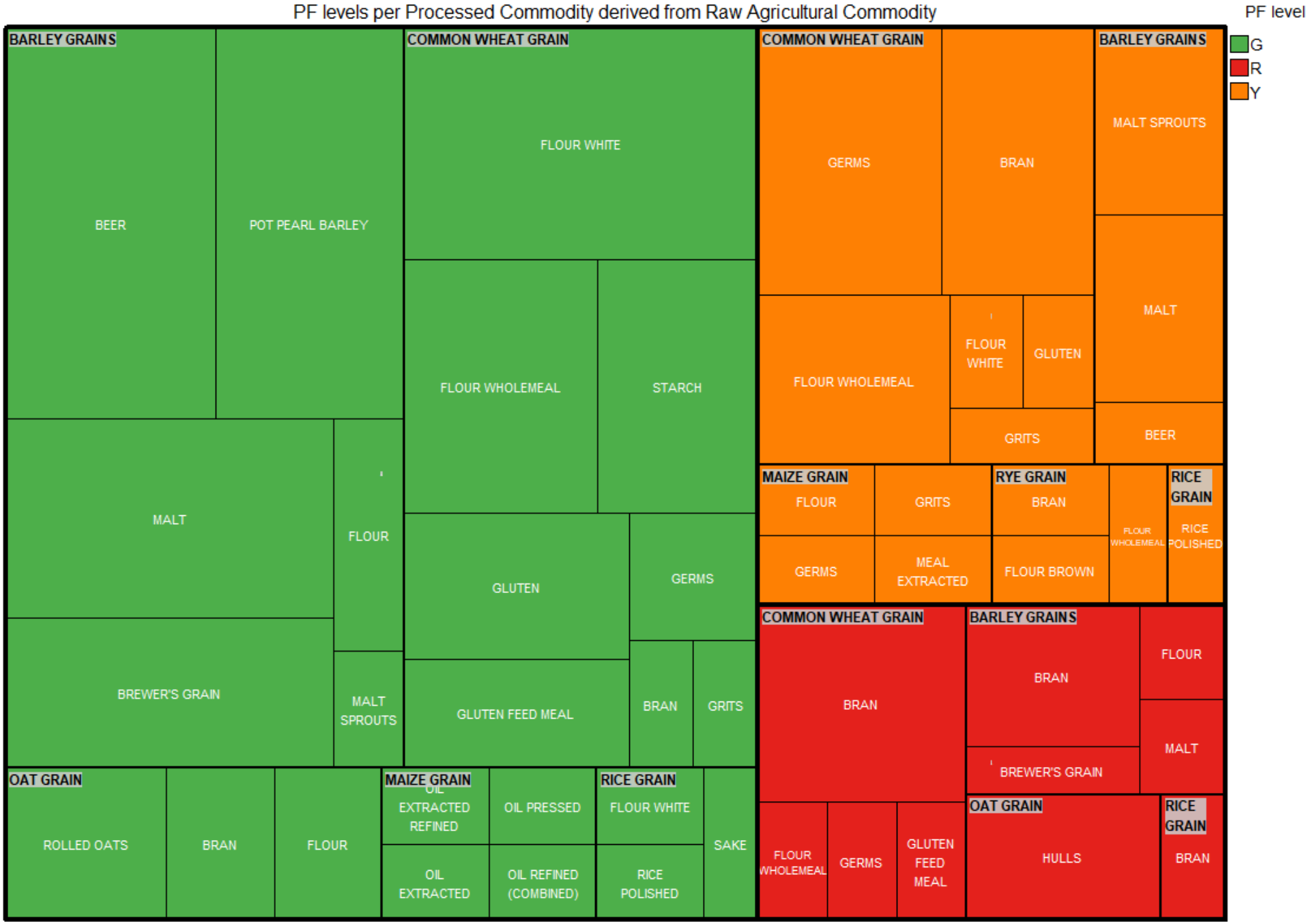


Figure 4-17 Cereals treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RACs and processed commodities.

Alluvial

Cereal is a particular case in the dataset, since there is no difference between RAC group and subgroup, therefore the alluvial diagram in Fig. 4-18 has again six axes and highlights key trends in cereal processing, with common wheat and barley grains dominating the chart. These grains, processed into commodities like flour, gluten, and malt, constitute the most significant flows across the processing chain. Smaller contributions come from maize, rice, oat, and rye grains, indicating a lower frequency of processed products from these RACs. Processing flows in the diagram show common wheat being associated with process family XI (Milling) while barley mainly allocated for process V (Fermentation) and, secondly, for milling. Common wheat supports a wider variety of products, including white and wholemeal flour, gluten, and starch. On the other hand, maize is more associated with oil production. PF levels vary across the diagram, with green flows dominating in products such as gluten, starch, beer, white flour, different extracted oil, pot pearl barley, rolled oats and sake. Meanwhile, yellow flows are commonly associated with by-products like bran and germs but there is also a consistent part of wholemeal flour with this PF level, showing that residues often remain at similar levels to those found in the raw grain. Red flows, though fewer, concentrate in by-products such as bran and hulls. In this case it is noticeable that when strata are small, the corresponding labels become compressed and difficult to read.

Overall, the diagram effectively captures the relationships between raw agricultural commodities, processing methods, and residue outcomes, providing a clear visual differentiation between products that reduce pesticide residues (green) and those that concentrate them (red). However, the complexity of overlapping lines and small labels makes it difficult to trace specific flows, especially for less abundant RACs and active substances.

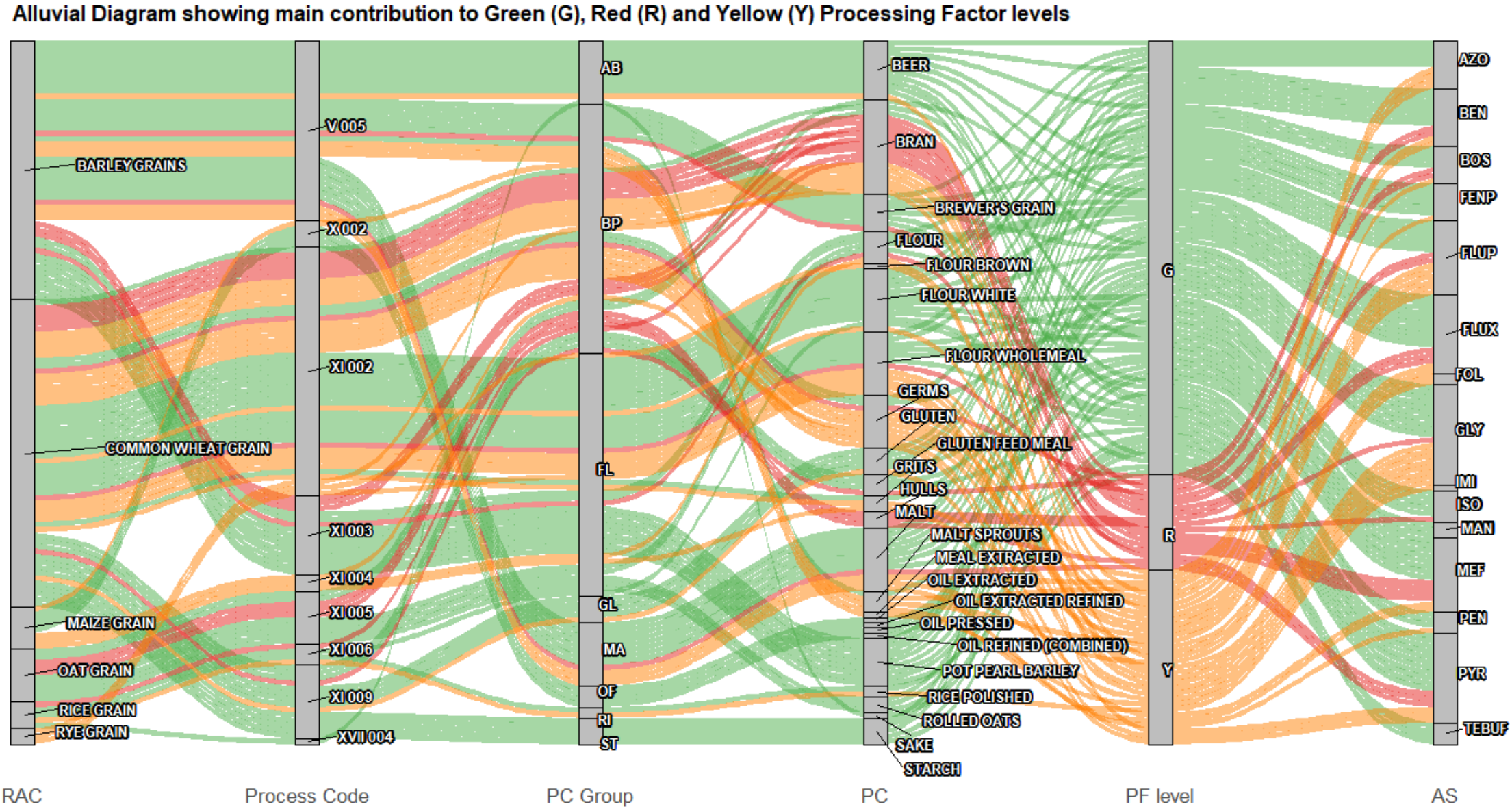


Figure 4-18 Cereals alluvial plot: visualization according to RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

The J-PCA model for *Cereals* RAC group was calculated with 2 principal components, accounting for 54% of total data variance. The corresponding PC1-PC2 score plot is reported in Fig. 4-19, showing that PC1 separates the samples according to RAC type (coded by samples color in Fig. 4-19 A) and processes (labels in Fig. 4-19 A). The cereals group is dominated by the XI process (Milling), which varies depending on RAC type. Other processes include the V process (Fermentation) for barley (fermentation into beer), as well as minor processes like rice fermentation (XVII Other Fermentation) and X process (Oil Extraction/Pressing Milling). PF levels, on the contrary, are more separated along PC2 (Fig. 4-19 B), showing an orthogonal trend with respect to processes or raw agricultural commodities. As a result, no clear pattern is distinguishable with these two categories. Looking for alignment with the final processed commodity (labels in Fig. 4-19 B), this reveals again many by-product samples at the red and yellow PF levels (circled cluster at positive PC2 scores) while the main product, i.e., flour, typically maintains a green PF level (circled cluster at negative PC2 scores). Therefore, also in the case of cereals, it seems that the final processed commodity (and the presence of by-products in particular) explains more about PF levels. Nevertheless, exceptions to this trend are more frequent than in the fruits group, leading to a higher occurrence of green by-products and red/yellow main products. To summarize, while the "by-product trend" is still evident, it is accompanied by more variability.

By-products in the cereal group are primarily food waste from the milling of kernels. This becomes clear by selecting only common wheat grain samples as an example (Fig. 4-20). This set of samples undergoes a common process (XI 002, milling flour wheat-rye process), to produce various types of flour (white, wholemeal, grits) as well as by-products like bran and germ. Conceptually, bran is similar to fruit peel, as it is the outermost kernel's layer. These two by-products are the primary contributors to yellow and red PF levels within this RAC. However, also in this specific case, exceptions are frequent, and several flour samples show yellow or red PF levels.

A possible explanation can be found by comparing the two sets of samples for white and wholemeal flour (Fig. 4-21 A and B). Although both processed commodities originate from the same RAC and undergo the same process, they consist of eleven samples each, corresponding to eleven pesticide residues tested in both processed commodities for processing factor definition. When examining the samples pesticide-by-pesticide, wholemeal flour is classified at the yellow or red PF level for five of them: flupyradifurone (FLUP), glyphosate (GLY), benzovindiflupyr (BEN), fluxapyroxad (FLUX), boscalid (BOS), whereas white flour only for flupyradifurone (FLUP). This means that 45% of the wholemeal samples fall into the yellow/red PF level, compared to less than 10% of the white flour samples. Wholemeal flour, as opposite to white flour, is subject to a less refined milling process that consequently implies more bran and germs in the final product (Scholz et al., 2022). Once again, this insight is possible due to specific food technology domain knowledge, which helps to explain the behavior of certain samples that initially appear to contradict the overall data trend. The only pesticide that results in yellow PF class for both wholemeal and white flour is flupyradifurone (FLUP), which is also the sole insecticide among the tested pesticides (*see* Appendix IV).

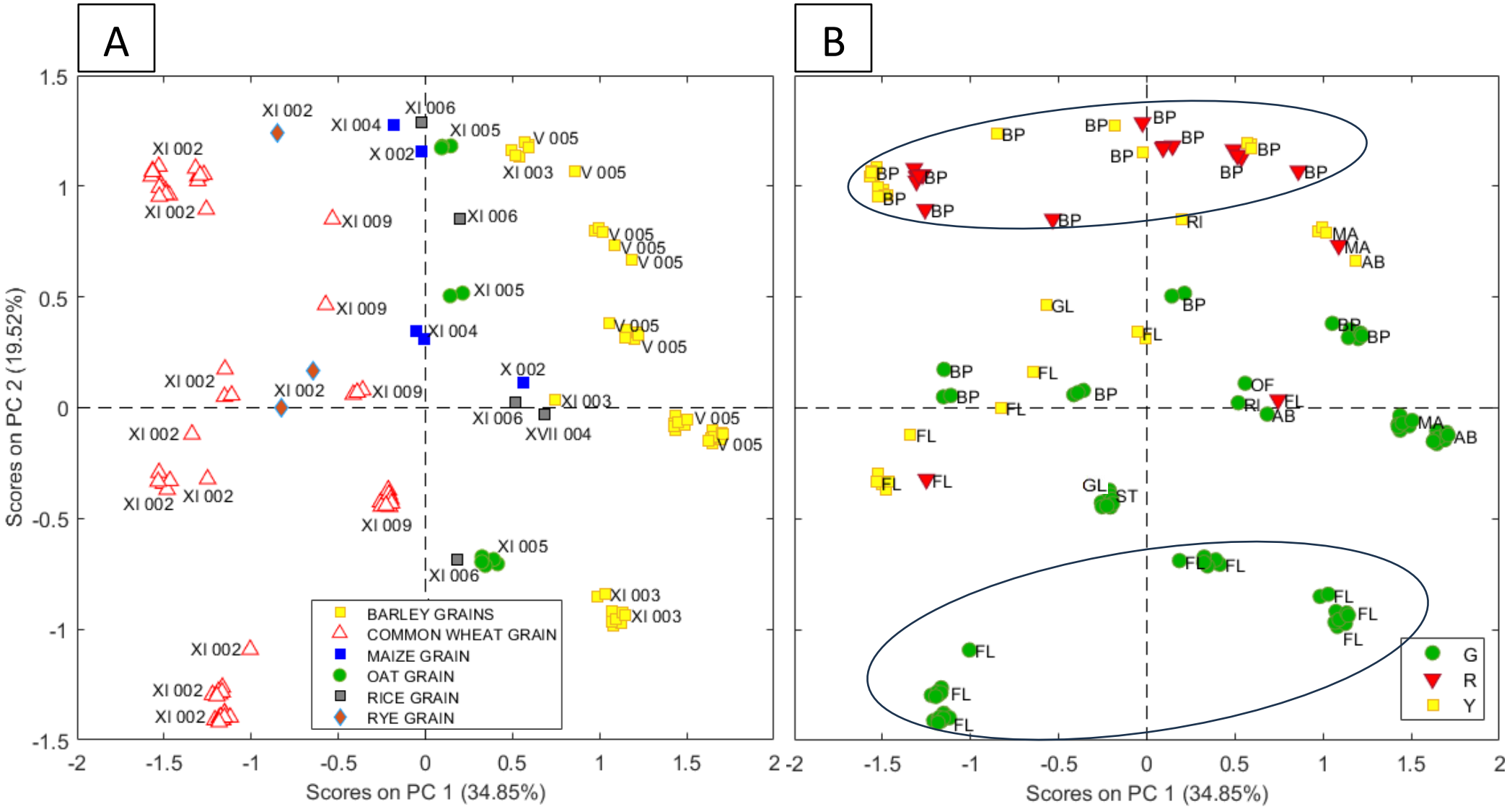


Figure 4-19 Cereals J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the specific RAC and labels indicate process codes (see Appendix II) while in B) samples are colored according to the PF level (G - green, R - red, Y - yellow) and labels indicate the processed commodity groups (see Appendix III).

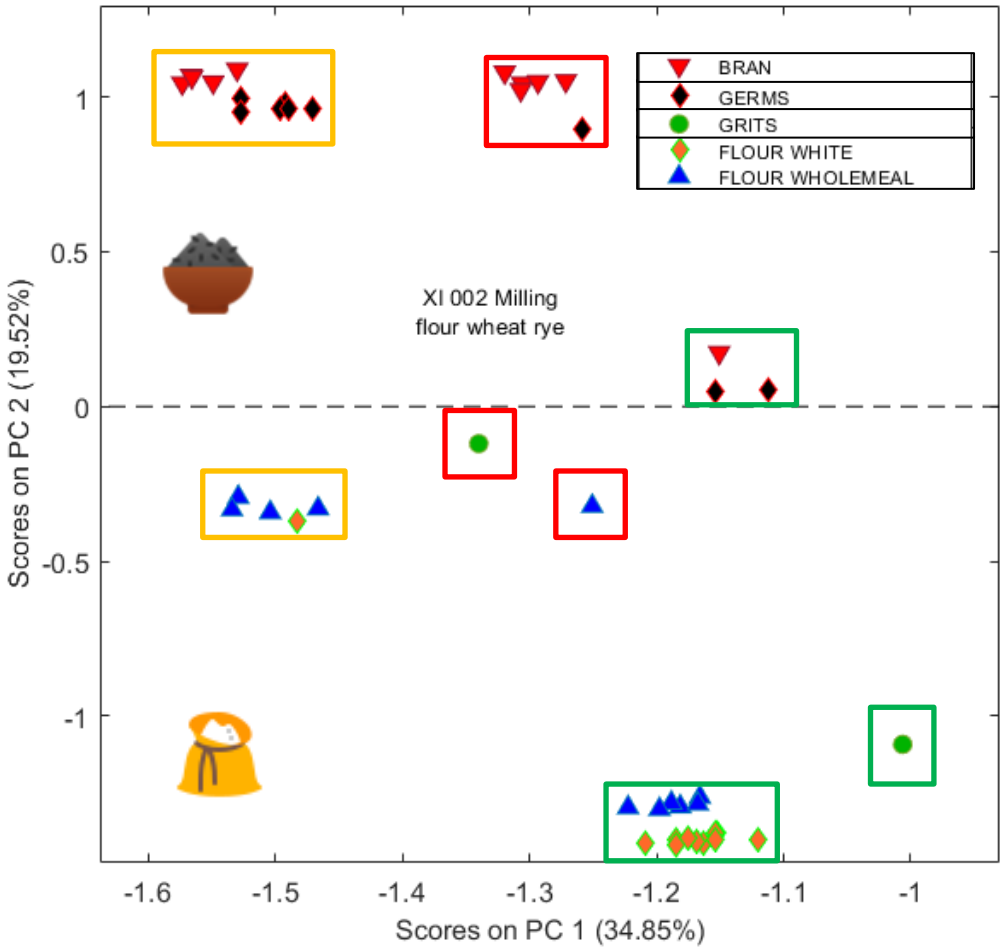
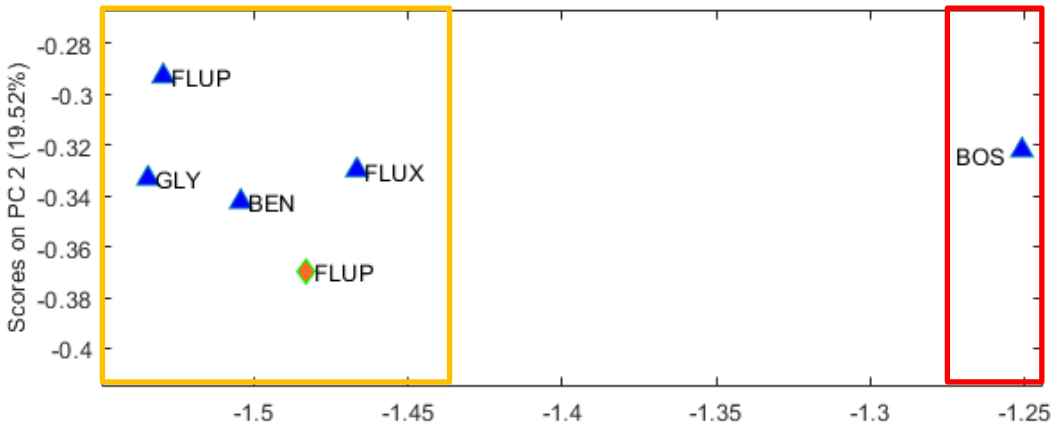


Figure 4-20 Cereals J-PCA: detail of PC1 and PC2 score plot for common wheat grain that undergoes milling process (XI 002). Samples are colored according to the specific PC, with icons (from Icons8) indicating the main separation flour/by-product. Samples are enclosed in colored clusters according to the PF level (G - green, R - red, Y - yellow).

A



B

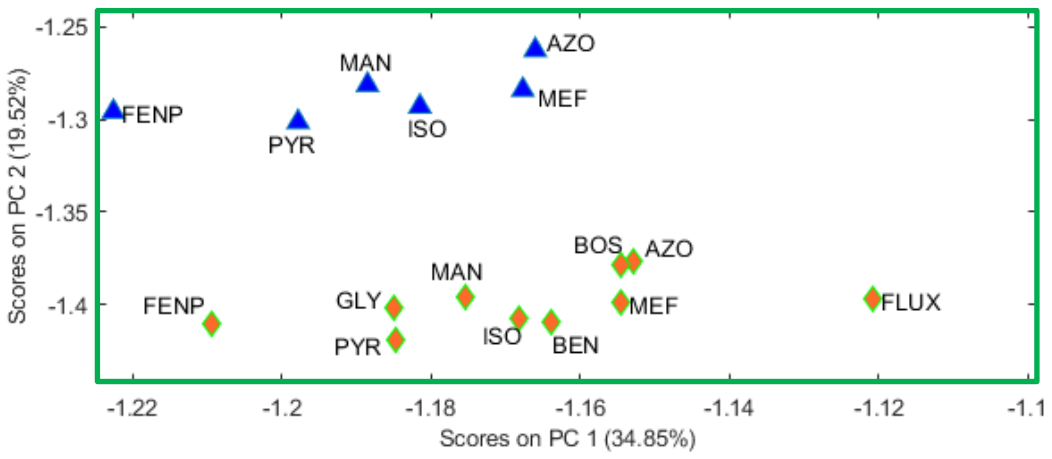


Figure 4-21 Cereals J-PCA: detail of PC1 and PC2 score plot for white (green-orange diamonds) and wholemeal (blue triangles) flour. In A) samples are all enclosed in a green PF cluster while in B) samples are enclosed in red and yellow clusters. Labels indicate pesticides (see Appendix IV).

4.3.2 Process and Processed Commodity

From a process-processed commodity perspective, it would be more valuable to investigate ambiguous processing techniques – those not clearly associated with high or low occurrences of yellow and red PF classes – rather than ones with a well-established impact, such as dehydration. During the exploration of raw agricultural commodities, certain ambiguous behaviors have already been identified. Additionally, it would be beneficial to focus on processes involving a diverse range of raw agricultural commodities or producing multiple distinct processed commodities, rather than one-to-one input-output processes.

Identifying processes with a one-to-many relationship to RACs and processed commodities in the database is challenging, as each RAC group typically has its own specific processes, with few processes crossing groups. However, two cases stand out as worthy of further exploration: *Fermentation Distillation* and *Fruit juice* processing.

OECD CODE	OECD PROCESS GROUP	PROCESS CODE	SPECIFIC PROCESS NAME	PROCESSED COMMODITY
V	Fermentation Distillation (*4.3.2.1*)	V 001	White wine production wine grapes	BEER BREWER'S GRAIN MALT MALT SPROUTS MUST SPENT HOPS SPIRIT WINE RED WINE ROSÉ WINE WHITE
		V 002	Red wine production wine grapes	
		V 003	Rosé wine production wine grapes	
		V 005	Beer brewing barley grain to malt	
			Beer brewing barley malt to beer	
		V 006	Beer brewing hops	
		V 007	Distillates wine grapes	
II	Fruit juice (*4.3.2.2*)	II 001	Citrus juice citrus fruits	JUICE JUICE (CALCULATED FROM NECTAR) JUICE CLARIFIED JUICE CLARIFIED-PASTEURISED JUICE PASTEURISED NECTAR POMACE DRY POMACE WET PULP PULP DRIED
		II 002	Pome juice pome fruits (apples pears)	
		II 003	Grape juice berries and small fruits (currants)	
			Grape juice grapes	
		II 004	Stone fruit juice stone fruits	

Table 4-2 OECD and specific processes dependencies analyzed in 4.3.2. Obtained processed commodities are also reported. For a complete view of processes and processed commodities see Appendices II and III, respectively.

4.3.2.1 Fermentation-Distillation

Recalling the analysis performed at both *Berries and small fruits* (4.3.1.2) and *Fruits (fresh or frozen) and Tree Nuts* (4.3.1.3) levels, fermentation showed a high variability in PF classes distribution, with many samples with a green PF level but also a significant number of occurrences with yellow or red PF levels. Furthermore, in this process two main different RAC groups are involved: *Fruits* and *Cereals*. In this way, it is possible to extrapolate valuable insights from PF database considering two very different processed commodities, wine and beer, linked by the core transformation step of alcoholic fermentation worked by yeasts metabolism.

As illustrated in Tab. 4-2, the dataset has been filtered by OECD process *V Fermentation Distillation* to include all the corresponding specific processes: this allowed for broader and more insightful comparisons in the analysis. The filtered dataset contains 104 records.

Treemap

The treemap in Fig. 4-22 effectively illustrates the overall distribution of green, yellow, and red PF classes. It is immediately apparent that most samples fall into the green PF class, followed by a substantial portion in the yellow PF class, with relatively few occurrences in the red PF class. However, when examining individual processed commodities (e.g., red wine, white wine), treemap visualization spreads related commodities across separate sections, making it more challenging to follow their specific contributions to PF levels. Despite these limitations, a clear distinction emerges between beer-related products (derived mainly from barley grains and hops) and wine-related products from wine grapes. Beer and barley-derived products are predominantly represented in the green PF class. In contrast, wine products are more dispersed across green, yellow (notably), and red PF classes, indicating greater variability in residue concentration.

While the treemap provides an efficient overview of PF level proportions, it lacks a detailed pathway clarity between RACs, processed commodities, and PF levels.

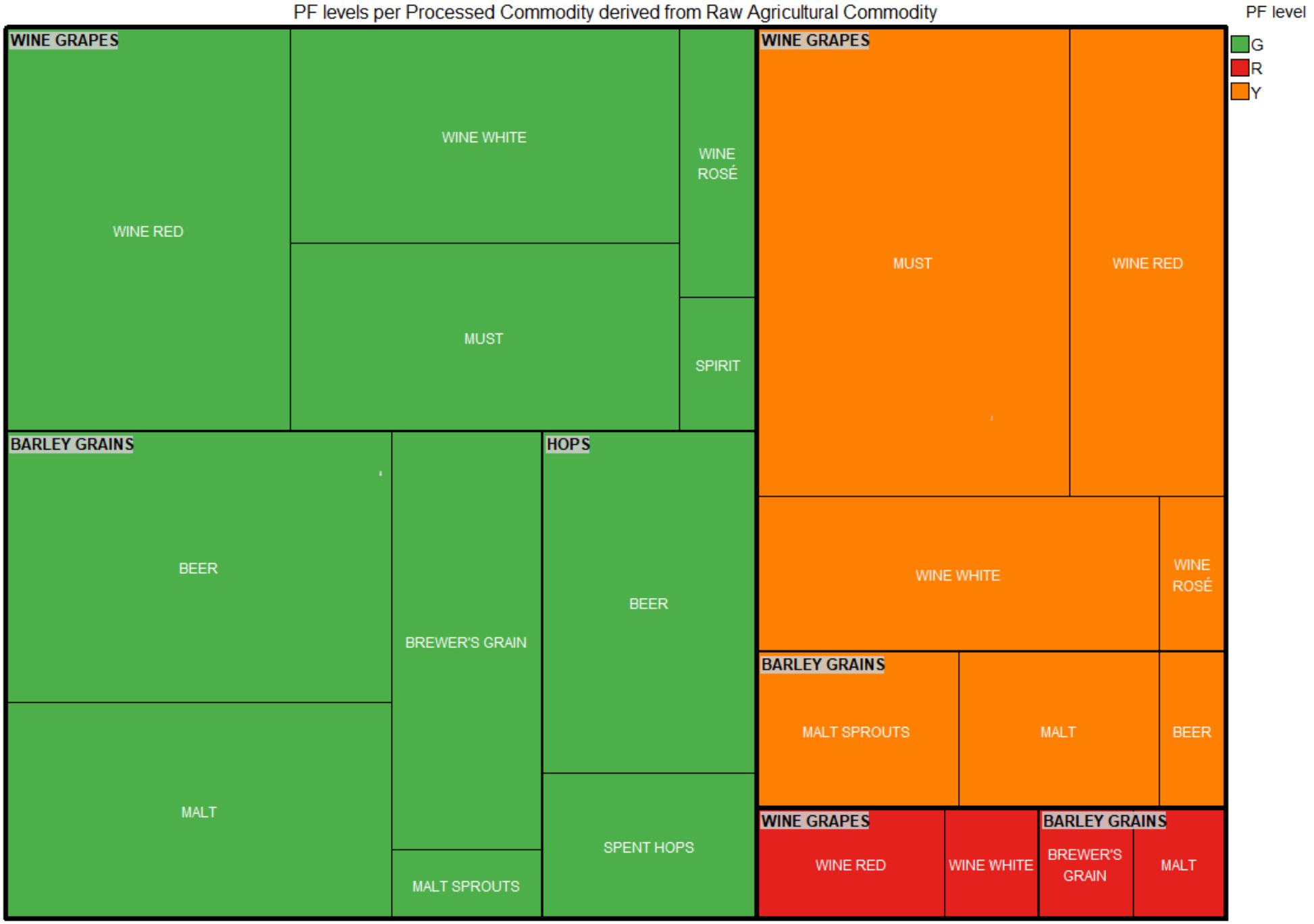


Figure 4-22 Fermentation-Distillation treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RACs and processed commodities.

Alluvial

In the alluvial plot (Fig. 4-23) the overall partitioning of green, yellow, and red PF levels is less evident as the color-coded PF class alluvia are intertwined throughout the graph. However, the overall distribution is clarified when looking from PF level axis backwards: green flows are dominant, especially among products like beer, brewer's grain, malt, and spent hops, all derived from barley grains and hops. This may suggest that beer processing generally reduces pesticide residue levels effectively. Yellow and red flows are more frequently associated with wine-related products (from wine grapes): this variability may stem from differences between beer and wine fermentation or from the inherent characteristics of wine grapes and associated pesticides residues, which are anyway difficult to identify in alluvial (it will be simpler in J-PCA).

Compared to treemap, the alluvial diagram provides clearer partitions on specific categorical features (e.g., processed commodity), making it easier to observe individual contributions to each PF level. For example, the diagram clearly shows the transitions from wine grapes to processed commodities like red wine, white wine, and rosé, along with their final PF levels. Such specific partitions related to processed commodities are less visually discernible in the treemap, where similar commodities are scattered across different sections.

It has to be noticed that filtering the PF dataset at the OECD process level requires the inclusion of the dedicated process axis in the alluvium plot, as well as all axes related to RAC grouping (3 axes, corresponding to RAC group, RAC subgroup ad single RAC) and to the processed commodity class sets (2 axes, corresponding to processed commodity group and processed commodity). The usual PF level and active substance (AS) axes are also included, resulting in a total of eight axes needed to provide a comprehensive view. This new complexity makes it even more challenging to follow flows from start to end.

Many pathways overlap, and several labels (particularly around the process code and active substance axes) are difficult to read due to space limitations. The high density of connections between axes, especially near the PF level and AS axes, complicates the tracking of specific flows and makes it harder to distinguish connections between RACs, processed commodities, and residue levels. Furthermore, labels overlap along multiple axes (e.g., RAC Subgroup and PF Level) further reduces readability, making it difficult to quickly identify specific categories or flows without close inspection.

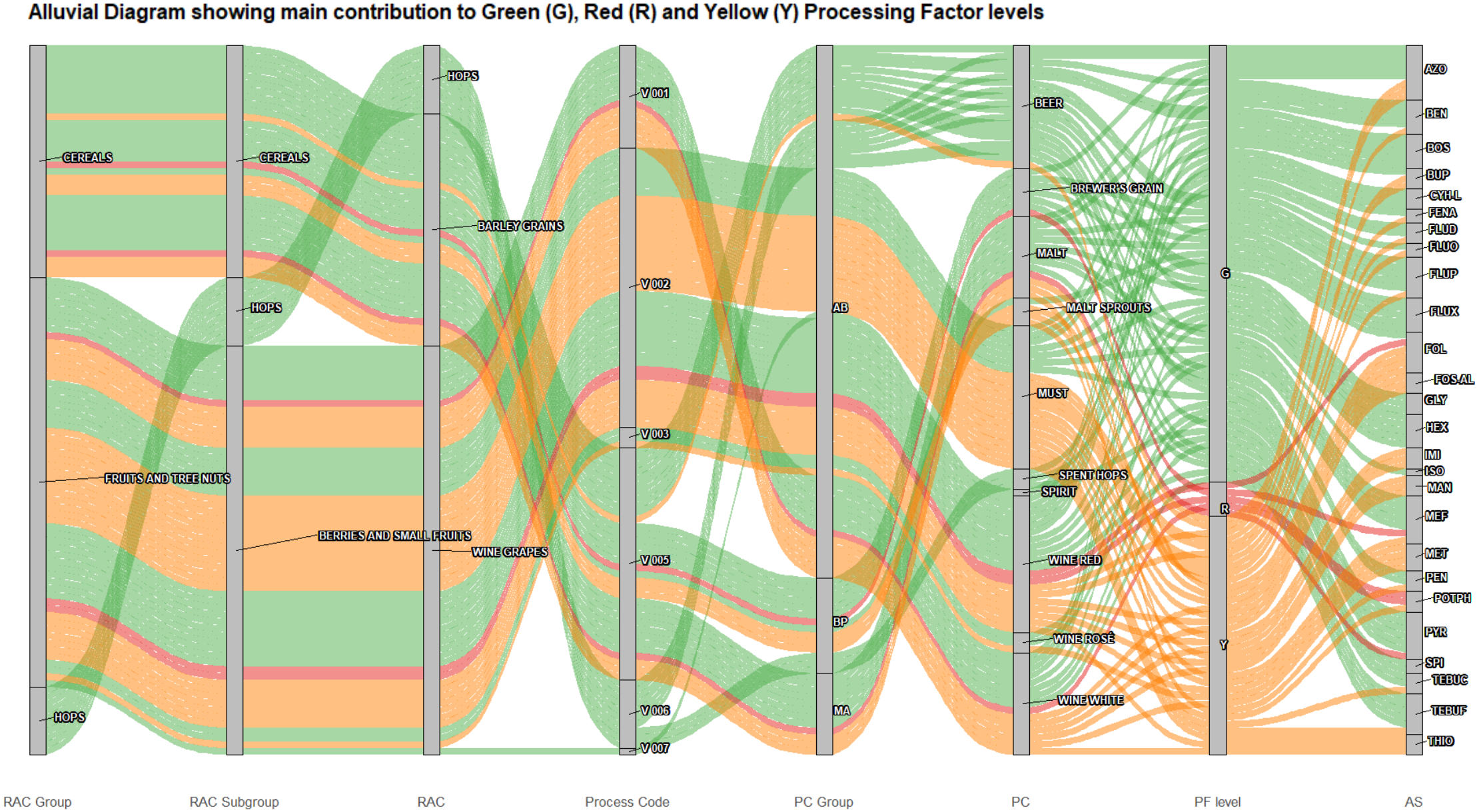


Figure 4-23 Fermentation-Distillation alluvial plot: visualization according to RAC groups, RAC subgroups, RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

The J-PCA score plot in Fig. 4-24 displays the distribution of winery and brewery products based on two principal components, which explain over 80% of total data variance. This score plot enables visual differentiation of samples based on the raw agricultural commodity from which they are produced (represented by wine grapes in one case, barley grains and hops in the other), as well as their process and processed commodity types. Of course, the main difference between winery and brewery samples captured by PC1 is the RAC (Fig. 4-24 A).

Wine grapes cluster distinctly at negative PC1 scores, indicating a unique profile separated from barley and hops. Barley grains are located at the right side of the plot, i.e., with positive PC1 scores. Barley samples are widely distributed along the PC2 axis, suggesting diversity within barley-derived samples across different processing outcomes. Hops are in the lower right quadrant, at negative PC2 scores and they are clustered in two groups. This spatial separation among RACs demonstrates clear differentiation in the profiles of these commodities based on the primary components.

In Fig. 4-24 B the samples are colored based on the processed commodity class set, including beer, brewer's grain, malt, malt sprouts, must, spent hops, spirit, and various types of wine (red, white, and rosé), and each sample is labeled according to specific processing methods. Beer samples lie at negative PC2 score values (except for one sample), aligning with barley grain and hops RACs, while malt, malt sprouts and brewer’s hops are mainly located at positive PC2 values. Considering wine grapes RAC, two clusters are evident along PC2, each one containing both must and wine (white, red, and rosé), while the spirit sample is only located in the cluster at negative PC2 values. The clustering behavior within the wine grapes RAC becomes clear when the objects in the PC1 and PC2 space are colored according to PF classes, as highlighted in Fig. 4-25. Indeed, the samples with negative PC2 score values are characterized by green PF levels, while the samples with yellow or red PF levels are located at positive PC2 values. This variability of PF levels can be explained considering that different pesticides may have different behaviors during the transformation and fermentation processes occurring during wine production. In contrast, all but one beer samples fall within the green PF class, suggesting that the specific processing method effectively reduces residue levels. These differences between winery and brewery products may result from the distinct underlying transformation techniques, despite both processes being fermentations (V), as well as from the fundamentally different raw agricultural commodities used (wine grapes versus barley grains), which are likely associated with distinct pesticide applications during farming. Nevertheless, it is evident that the occurrence of yellow and red PF classes in the main product beer is rare (with only one yellow sample, while the rest are malt, malt sprouts, and brewer’s grain samples), while it is more frequent in wine.

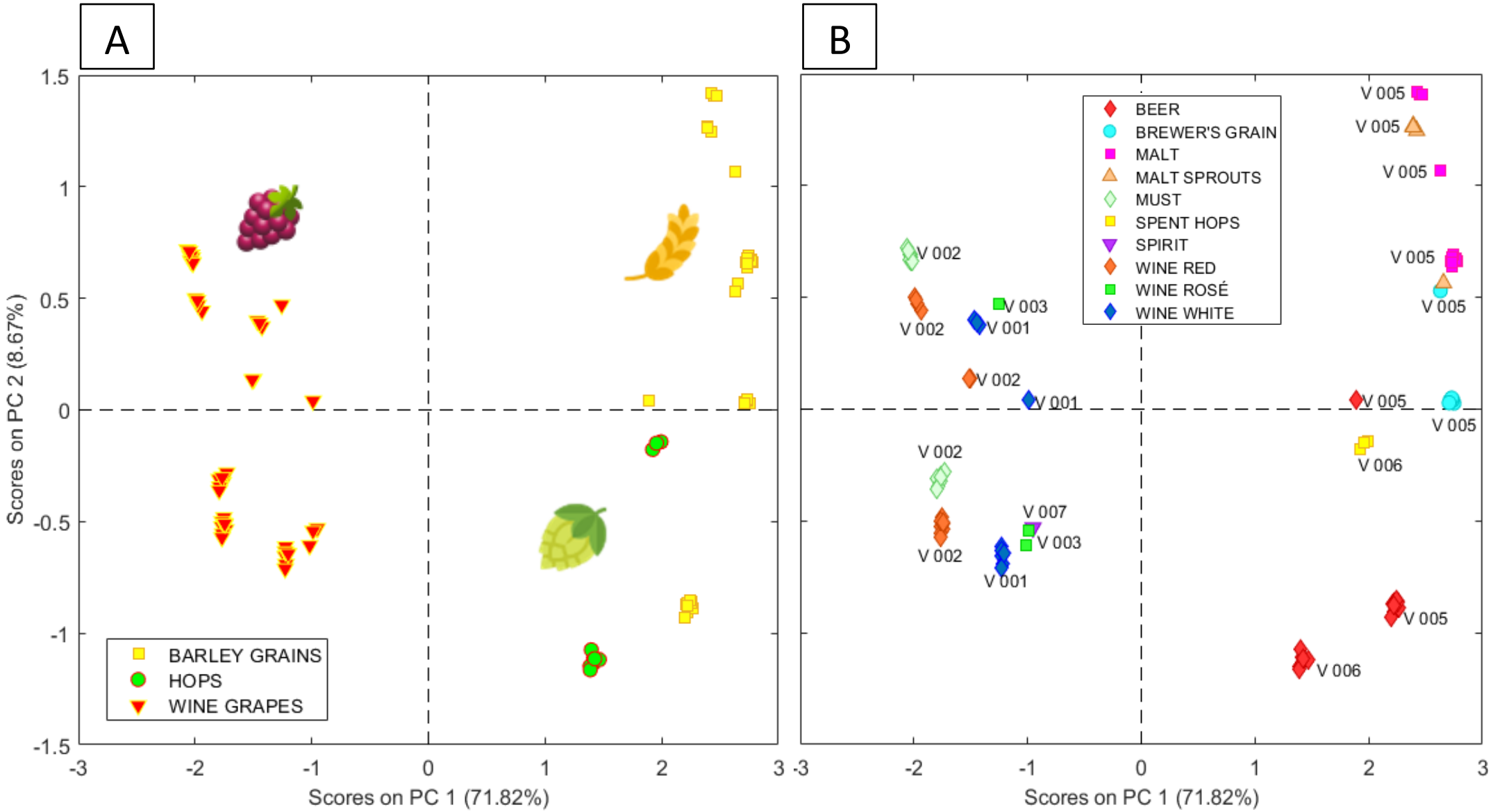


Figure 4-24 Fermentation-Distillation J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the specific RAC, with icons (from Icons8) while in B) samples are colored according to the processed commodity and labels indicate process codes (see Appendix II).

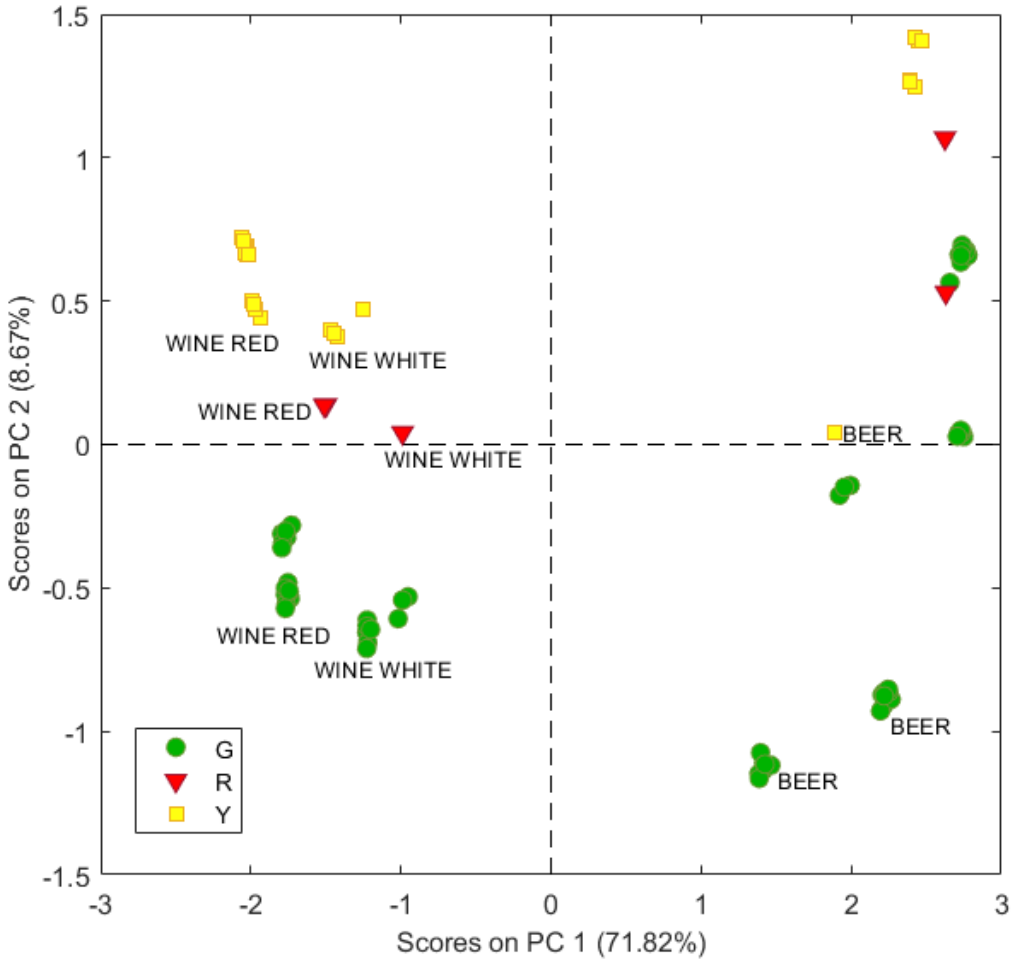


Figure 4-25 Fermentation-Distillation J-PCA: PC1 and PC2 score plot. Samples are colored according to the PF level (G - green, R - red, Y - yellow) and labels indicate the processed commodity.

To further investigate pesticides effect, Fig. 4-26 compares the profiles of red and white wine, revealing a high degree of similarity between the two processed commodities, with the same pesticides generally appearing in the same PF classes for both types of wine. A few exceptions can be found. Some pesticides are present only in red wine, such as azoxystrobin (AZO), fenazaquin (FENA) and fluopyram (FLUO), while other pesticides like bupirimate (BUP) appear in the yellow PF class for red wine but in the green PF class for white wine. This similarity in pesticide PF classes between red and white wine, both in terms of yellow/red class occurrence and type of residue in which class, is particularly noteworthy. Based on earlier observations in Section 4.3.1, pesticides tend to concentrate more in the outer layers of RACs, such as the peel in fruits or the bran in cereals. This generally results in a higher occurrence of yellow and red PF levels in these by-products, as well as in products that contain them (e.g., wholemeal flour) or, extending by analogy, in products that remain in contact with them for extended periods. Given the longer skin contact time in red wine vinification, one would expect red wine to show a more pronounced occurrence of yellow and red PF classes compared to white wine. However, this effect appears to be less significant than anticipated.

Two possible explanations for this discrepancy can be formulated. One explanation is the presence of a thinner and softer “peel” in grapes (which is more commonly called skin for this reason), which may allow pesticides to penetrate more readily into the fruit pulp, thus leading to a more uniform distribution of pesticides residues and less difference in red/white wine profiles. On this topic, the study conducted by Cabras et al. (2000) was already cited as an example showing higher presence of folpet (FOL) on grape skin, in alignment with other pesticides applications in fruits (citrus, pomace). Similar conclusions can also be found in Scholz et al. (2022), but as showed by Teixeira et al. (2004), some uncertainty remains regarding grapes, as pesticides residues stratification in peel/pulp can vary on a case-by-case basis. Another explanation is that

pesticides residues passage into liquid grape juice phase (and from there into final fermented wine) happens mostly at the fruit crushing stage already, when initial mash is obtained. This is a common production step in both red and white vinification, in which grapes skins and juice are mixed for a certain period (shorter in white, longer in red). Then mash is pressed to filter out pomace and skins and obtain must, on which white vinification is started, or kept as it is to start the red vinification. In this scenario, this additional skin contact time in red wine production may not be a crucial factor. For further details *see* respective process flowcharts in Appendices V and VI.

From a risk assessment perspective, Fig. 4-26 highlights common critical active substances in winery products. These are potassium-phosphonates (POTPH), showing red PF level in both types of wine, and folpet (FOL), red PF in red wine and yellow PF in white wine. For folpet, this pattern notably contrasts with findings of Čuš et al. (2010), where this pesticide is not even detected in final analyzed wine samples. While this individual study is certainly compliant with good laboratory practices (GLP), with statistically significant and reliable results, it underscores the importance of evaluating individual paper results within a broader research context. Comparing results across multiple studies can indeed provide a more comprehensive understanding, as isolated findings may not always align with general trends. In current analysis the two data points account for 16 studies on folpet in total (12 on white wine, 4 on red wine), all meeting GLP and quality criteria: this broader dataset suggests a moderate likelihood of folpet residue increase from grapes to wine, a trend that was not evident in Čuš et al. (2010). Thus, examining findings across a range of studies provides a more comprehensive view that can help avoid conclusions based solely on single studies; J-PCA on the PF dataset specifically enabled this broader scope. Additionally, the folpet case aligns with initial expectations on PF level for red (higher) and white (lower) wine, given the previously discussed background on respective processing conditions.

Fig. 4-27 displays beer pesticides profile. It shows some common pesticides with wine in the green PF class, but also many more that are specific to barley-hops-beer process. From specific beer flowchart in Appendix VII, it can be noted how the fermentation process starts after a cleaning step of barley grains, where presumably many by-products are separated from the RAC: a crucial aspect absent in wine fermentation, which may account for most of the differences in PF level expressions between these two commodities. Notably, folpet is the only pesticide in the yellow PF class for beer as well, raising further questions about this substance and its potential specific interaction with alcoholic fermentation process in general.

On a final note, it is important to reiterate that the in-depth insights provided by J-PCA would have been much more challenging to achieve with treemap or alluvial plots. While these visualizations are effective for providing a general overview – along with the discussed peculiarities, pros and cons of each method – they generally lack the ability to easily track samples across categories, from RAC to pesticide, underscoring the added value of J-PCA for detailed analysis.

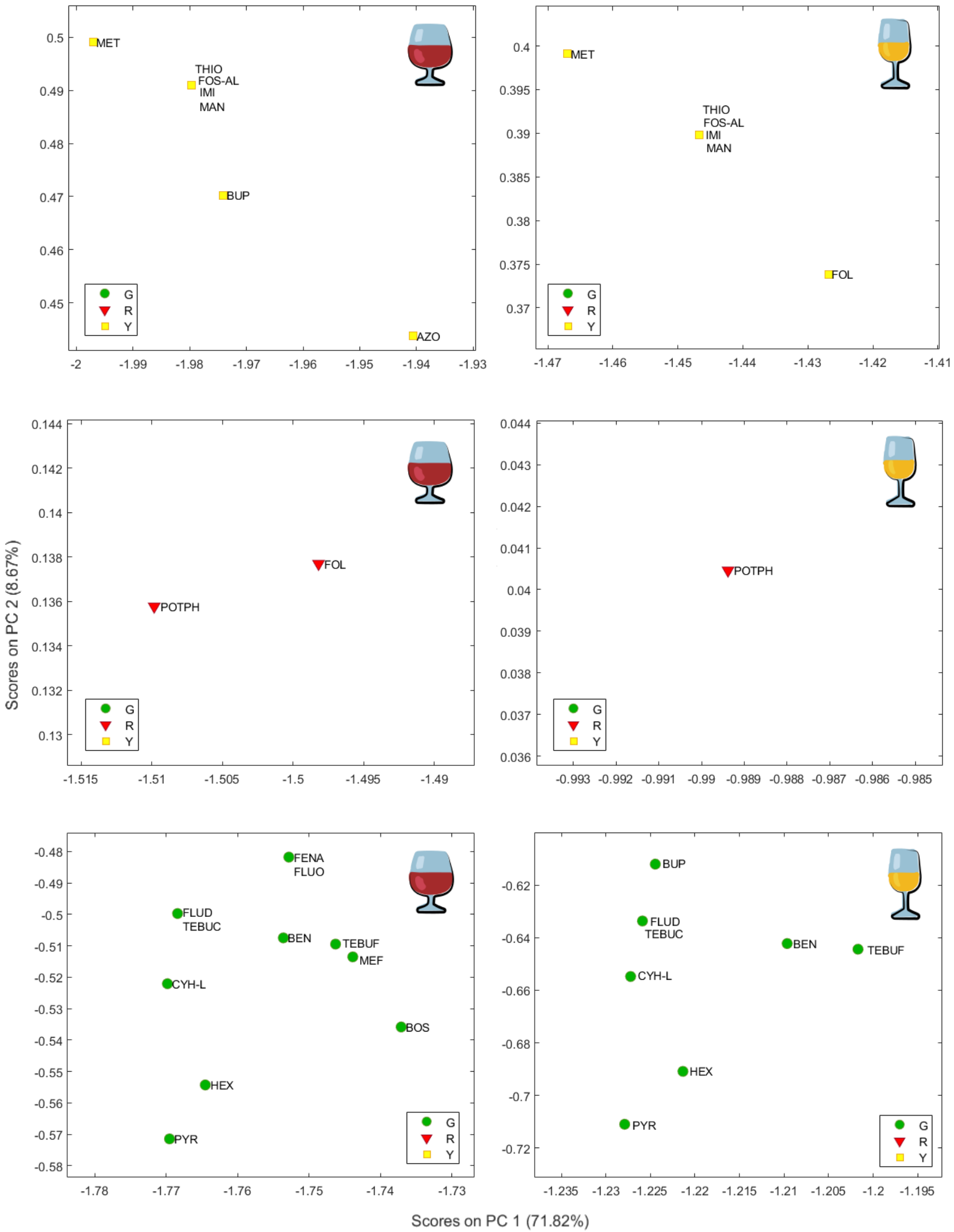


Figure 4-26 Fermentation-Distillation J-PCA: detail of PC1 and PC2 score plot for red and white wine samples (icons from Icons8), colored according to the PF level (G - green, R - red, Y - yellow) and with labels indicating the pesticide (see Appendix IV).

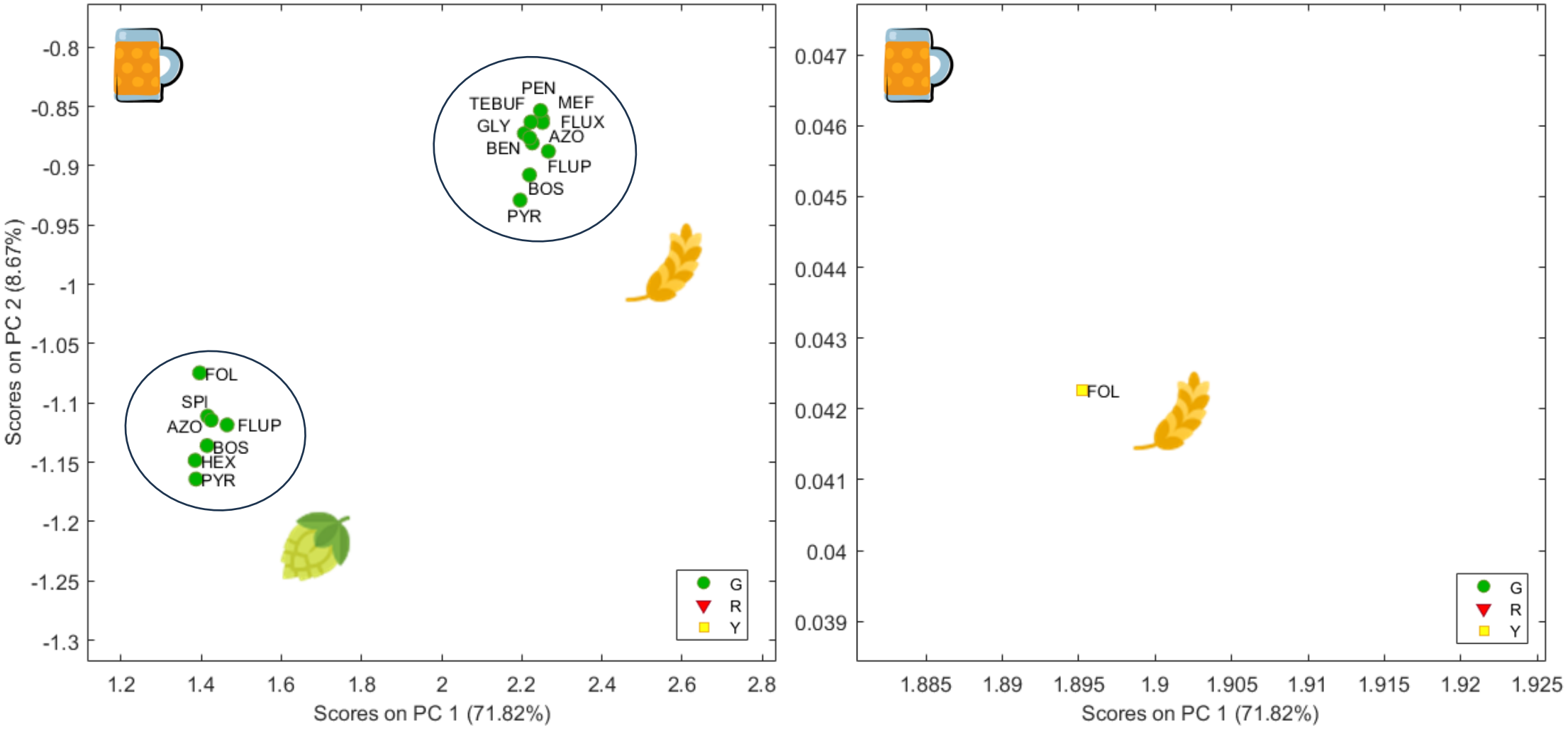


Figure 4-27 Fermentation-Distillation J-PCA: detail of PC1 and PC2 score plot for beer samples (icons from Icons8), colored according to the PF level (G - green, R - red, Y - yellow) and with labels indicating the pesticide (see Appendix IV).

4.3.2.2 Fruit Juice

The fruit juice process was selected due to its high variability in outputs, despite involving only a single RAC group as input. Indeed, the PF dataset includes distinct types of juice, each corresponding to further steps in the processing chain: juice, juice pasteurized, juice clarified, and juice that is both pasteurized and clarified. Is there any detectable difference in PF levels between these processing steps?

The filtered dataset contains 129 records.

Treemap

The treemap in Fig. 4-28 provides an overview of PF level distributions for processed commodities derived from various fruit RACs, including oranges, apples, table grapes, and other fruits like peaches, plums, and currants. The green PF class dominates the distribution, followed by a significant presence of yellow PF class, with fewer occurrences in the red PF class.

In the green PF area, prominence of products like pulp and various types of juice mainly from oranges and apples is shown, suggesting that the underlying process may effectively reduce pesticide residues across various fruit types. Green PF levels also prominently include juice and clarified juice derived from table grapes, further indicating this trend. The yellow PF level is more varied, but with a substantial representation of certain by-products, such as pomace wet, from apples and oranges. There is also a high occurrence of juice products for apples and table grapes while orange-derived juices, though present as well, seem to have a smaller entity. The red PF class, though smaller, is largely composed of by-products, specifically pomace wet and pomace dry from apples, oranges, and a few other fruits. For instance, pomace wet and pulp dried from oranges occupy significant sections within the red PF class, suggesting that these processing forms are more prone to residue concentration. Juice samples are nearly absent in this area, except for those derived from table grapes. However, it is unclear whether this represents one or multiple samples.

Overall, also this treemap provides on one hand an efficient and specific snapshot of PF level distributions across RACs and PCs, while lacking on the other hand the detailed connectivity between RACs, processing steps, and PF levels that would allow for clearer monitoring of general trends.

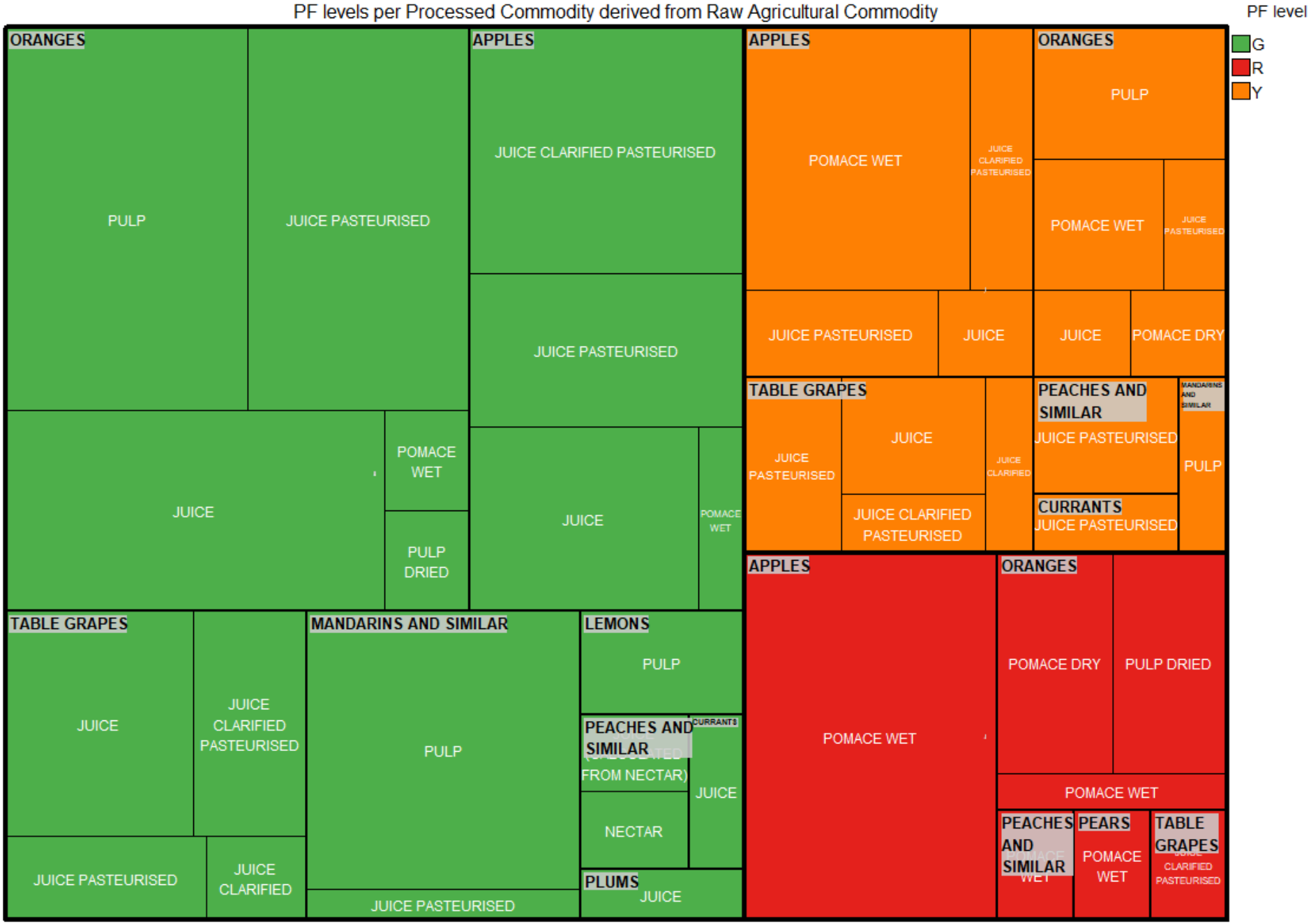


Figure 4-28 Fruit juice treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RACs and processed commodities.

Alluvial

This alluvial diagram (Fig. 4-29) provides an intricate visualization of residue distribution across fruit juice processing stages, from RAC groups to active substances. By tracing the color-coded PF levels, it becomes apparent that green flows are predominant, especially for clarified and pasteurized juice products. The RACs of these processed commodities are more difficult to deduce, but the previous treemap visualization indicates that oranges and apples are the main ones.

Yellow and red PF levels, in contrast, are more concentrated in by-products such as pomace wet and pomace dry, as well as in certain juice products from table grapes and stone fruits. Apple fruit juice samples with yellow PF level are more scattered and not easy to assess in comparison to treemap.

The alluvial format offers a more refined partitioning of individual contributions from specific RACs and processed commodities compared to a treemap, however, the inclusion of eight different axes also in this case adds considerable complexity to the visualization. Actually, in this case the RAC group axis could potentially be skipped since all samples undergoing fruit juice process come from *Fruits and tree nut*s RAC group. Some pathways overlap heavily, especially near the PF level and active substance (AS) axes, making it challenging to follow specific flows from start to end. Additionally, the high density of connections results in label overlap, particularly around process codes and active substances, reducing readability. While this diagram is effective in assessing overall trends, the visual density limits its usefulness for quickly grasping detailed pathways.

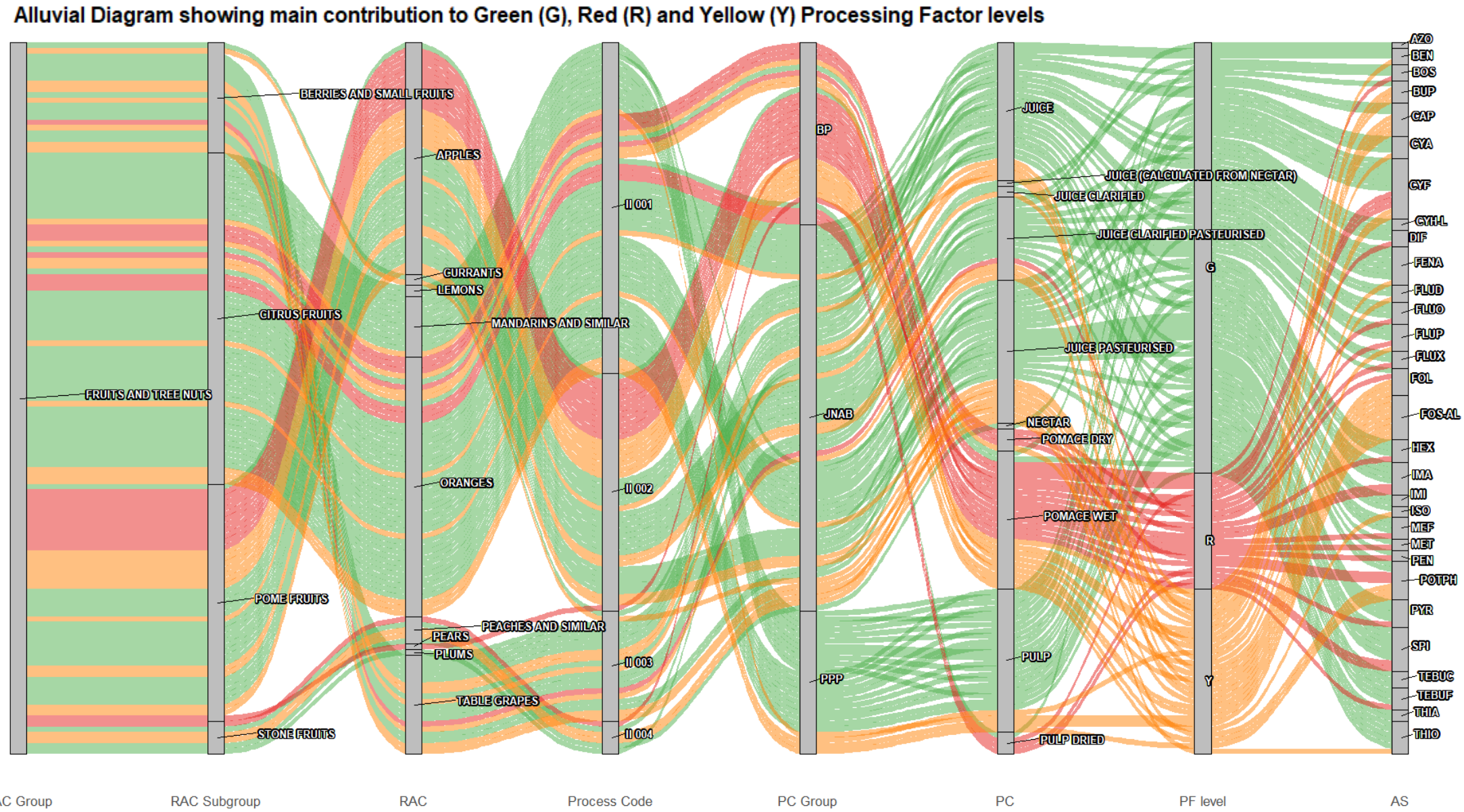


Figure 4-29 Fruit juice alluvial plot: visualization according to RAC groups, RAC subgroups, RACs, process codes, PC groups, specific PCs, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

J-PCA was performed on the Jaccard distance matrix related to *II Fruit Juice* OECD process, selecting two PCs that accounted for approximately 71% of total data variance. The corresponding PC1-PC2 score plot is reported in Fig. 4-30, revealing a clear separation between samples based on both RAC and processed commodity type. Along PC1 samples are grouped by RAC, with *Citrus Fruits* positioned at negative PC1 scores, *Pome Fruits* clustered at high positive PC1 values, and *Stone Fruits* along with *Berries and Small Fruits* occupying intermediate positions of PC1. In contrast, PC2 differentiates samples essentially into by-products (dry and wet pomace) and pulp at positive PC2 scores, while various juice products, including normal, pasteurized, clarified and with both treatments, are located at negative PC2 values. In accordance with previous explorations, and considering that *II Fruit juice* processing involves fruit samples, PF classes vary more PC2-processed commodity wise.

In Fig. 4-30 A samples are colored by processed commodity and labels indicate raw agricultural commodity, while in Fig. 4-30 B the color codes reflect the PF levels and processed commodity groups are reported as labels. No clear pattern can be observed RAC-wise. Instead, by-product (BP) trend is again quite clear. A consistent *Juices and Non-Alcoholic Beverages* (JNAB) group of samples can be observed in yellow-red PF classes for *Berries and Small Fruits*, followed by some yellow samples for *Pome Fruits* and only few yellow occurrences for *Citrus Fruits*. This different occurrence entity is probably due to peculiarities in the processes, like peel removal or not prior to juice pressing phase (*see* Appendices VIII, IX and X for further insights).

Additional juice processing steps, such as pasteurization and/or clarification, do not show a systematic effect in comparison to simple pressed juice samples residue levels, as also reported in Scholz et al. (2022).

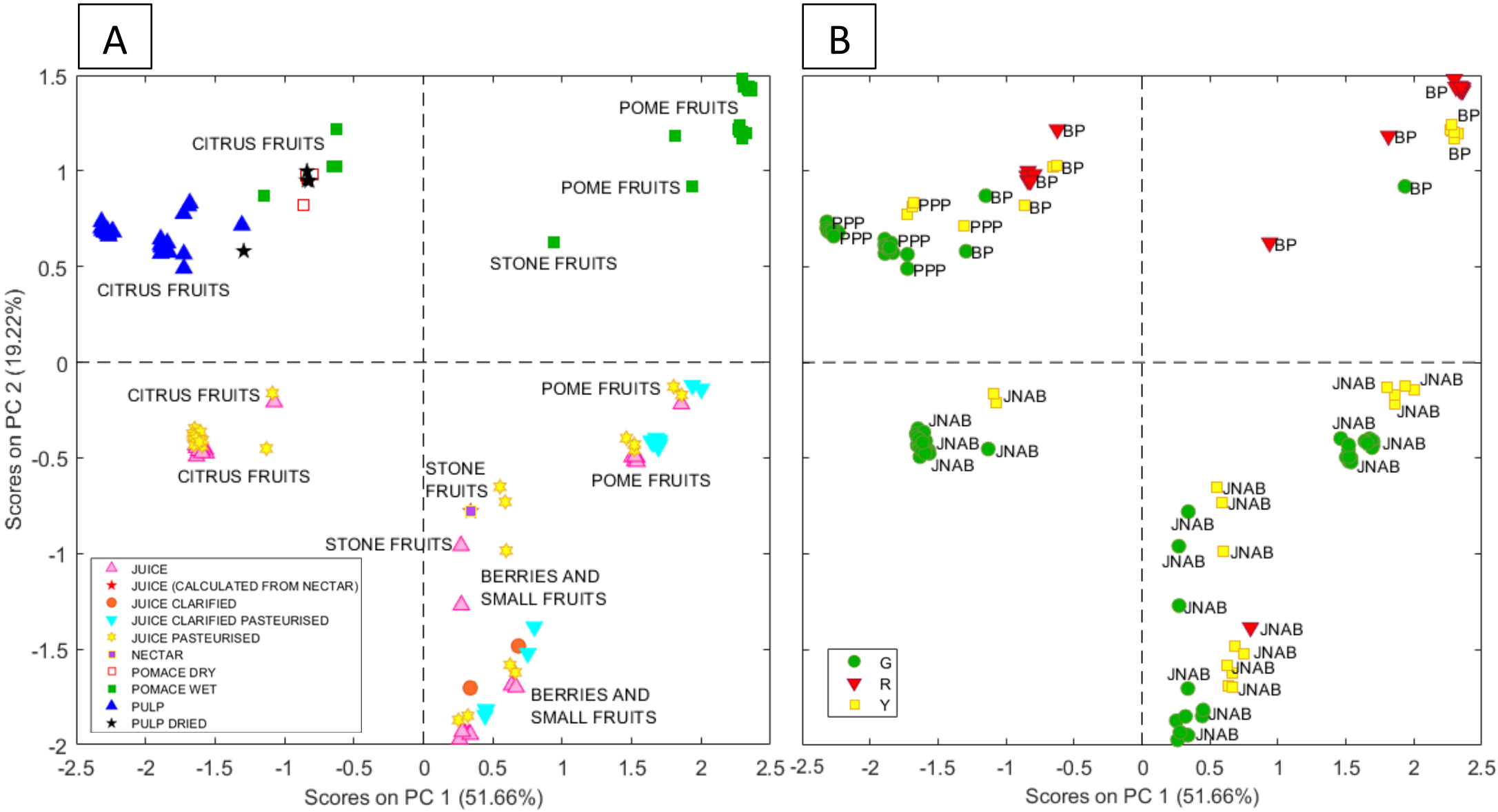


Figure 4-30 Fruit juice J-PCA: PC1 and PC2 score plot. In A) samples are colored according to the specific processed commodity and labels indicate RAC subgroups, while in B) samples are colored according to the PF level and labels indicate processed commodity groups (see Appendix III).

Fig. 4-31 presents the detail of two specific examples of fruit juice processing in the PC1-PC2 space: one for pome fruits (A) and the other for grapes (B). Consistent with the general findings in Fig. 4-30, no clear pattern in the PF level emerges regarding the impact of additional juice processing steps such as clarification and/or pasteurization. There is only one example of a pesticide residue that shifts from yellow PF class in simple pressed juice to green PF class when the juice is clarified and pasteurized: fludioxonil (FLUD) in grapes. However, this single case is insufficient to establish a widespread positive effect (i.e., reduction of PF value). Indeed, in the same case studies detailed in Fig. 4-31 other active substances have a different behavior. In pome fruits (Fig. 4-31 A), captan (CAP) consistently remains in the yellow PF class across juice, pasteurized juice, and clarified pasteurized juice, indicating no effect of these further processing steps. Conversely, in grapes (Fig. 4-31 B), folpet (FOL) PF values increase from the yellow PF class in juice and pasteurized juice to the red PF class in clarified pasteurized juice.

The substantial absence of effects on PF values of further treatments (thermal in particular) on captan in apple juice is consistent with Jankowska & Łozowicka (2022) work on processing factors for canning and pasteurization of apple pulp. Among the tested pesticides, captan consistently shows PF values greater than one and, in general, the highest values across both treatments. The authors note that the captan's thermal degradation point is 173 °C, which is significantly higher than the temperature used in apple juice pasteurization.

Folpet's behavior in grape juice is in accordance with observations made for the fermentation and distillation in winemaking (4.3.2.1). This residue consistently exhibits yellow and red PF levels after both processes, suggesting that neither alcoholic fermentation nor thermal processing effectively reduces folpet residues in grapes, regardless of whether they are intended for juice (table grapes) or wine production (wine grapes). This highlights the importance of adhering to Good Agricultural Practices for this RAC, particularly when folpet is used (but it is not the only case). The initial carryover effect into mash (a common step also in grape juice process, *see* Appendix X) has the same final consequences.

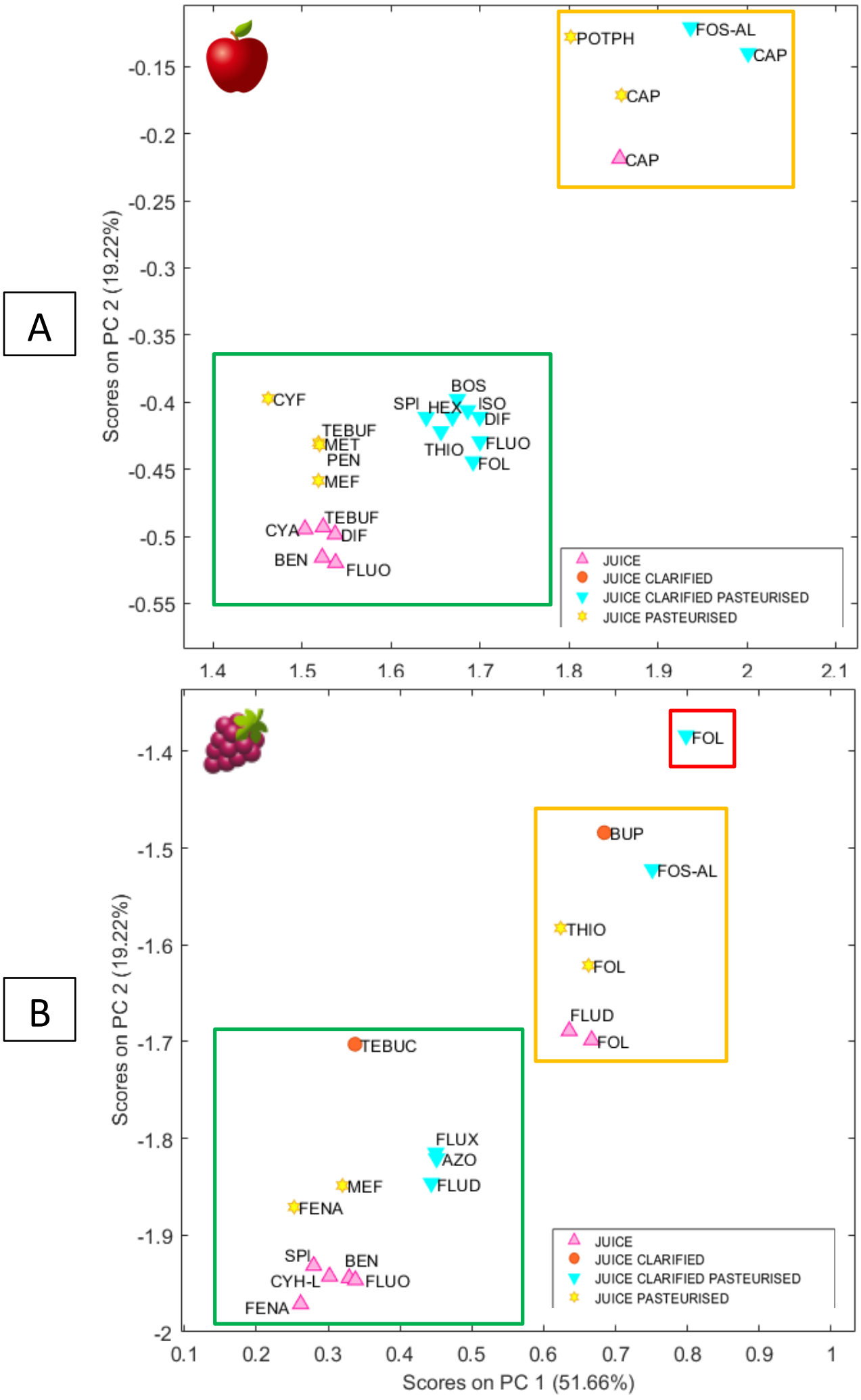


Figure 4-31 Fruit Juice J-PCA: detail of PC1 and PC2 score plot for apple samples (A) and table grapes (B). Samples are colored by the processed commodity and enclosed in clusters according to the PF level (G - green, R - red, Y - yellow). Labels indicate the pesticide. Icons from Icons8.

4.3.3 All Processing Factors Database

The Processing Factors database considered for this work includes 791 different samples, representing 4745 individual studies of the original database. Treemap, alluvial and J-PCA graphs are here compared to face and reveal visual complexity of this amount of data, further highlighting their profound differences, as well as advantages and limitations of each technique.

Treemap

To create a readable treemap visualization of the entire database, it is necessary to use broad RAC and processed commodity groupings, together with the PF level. Therefore, in Fig. 4-32, general RAC groups, such as fruits, vegetables, cereals, and oilseeds, are shown, organized into aggregated processed commodity categories like juices and non-alcoholic beverages, by-products, puree pulp paste and so on (*see* Appendix I and Appendix III for the complete list of RAC and processed commodities groups, respectively). The green PF class dominates the distribution, while the yellow PF class is less frequent, and the red PF class occupies the smallest area.

In the green PF area, products like juices, purees, and canned foods are prominent, particularly within the *Fruits (Fresh or Frozen) and Tree Nuts*, and the *Vegetables Fresh or Frozen* RAC groups. Flour and alcoholic beverages derived from *Cereals*, and oil-fat derived from *Oilseeds and Oil Fruits* are also represented primarily in the green PF area.

The yellow PF level displays more diversity, with a mixture of commodity groups, particularly by-products and dried foods within the *Fruits (Fresh or Frozen) and Tree Nuts*, and the *Cereals* families. This moderate PF level is evident across various processes, including juice production and some types of alcoholic beverages, which appear frequently in this section.

The red PF class, while less prominent, is primarily composed of by-products and dried food items, especially within the *Fruits (Fresh or Frozen) and Tree Nut*s, and the *Vegetables (fresh or frozen)* categories. Notably, items like canned foods and beverages are almost absent from the red PF class, suggesting that they are less prone to higher residue concentrations.

Overall, this treemap effectively illustrates PF level distributions across different groups of RACs and processed commodities. While it provides a clear snapshot of PF levels, it has limitations. The need to use broad RAC and processed commodity groupings for readability means that more specific partitions cannot be displayed, limiting the level of detail that can be displayed in this visualization. Additionally, overlapping labels in densely populated sections further reduce readability, making it challenging to distinguish specific processed commodities within larger RAC groups. As a result, the graph offers only a high-level view, without finer details on specific processing steps that could help clarify factors influencing PF variability across different commodity types.

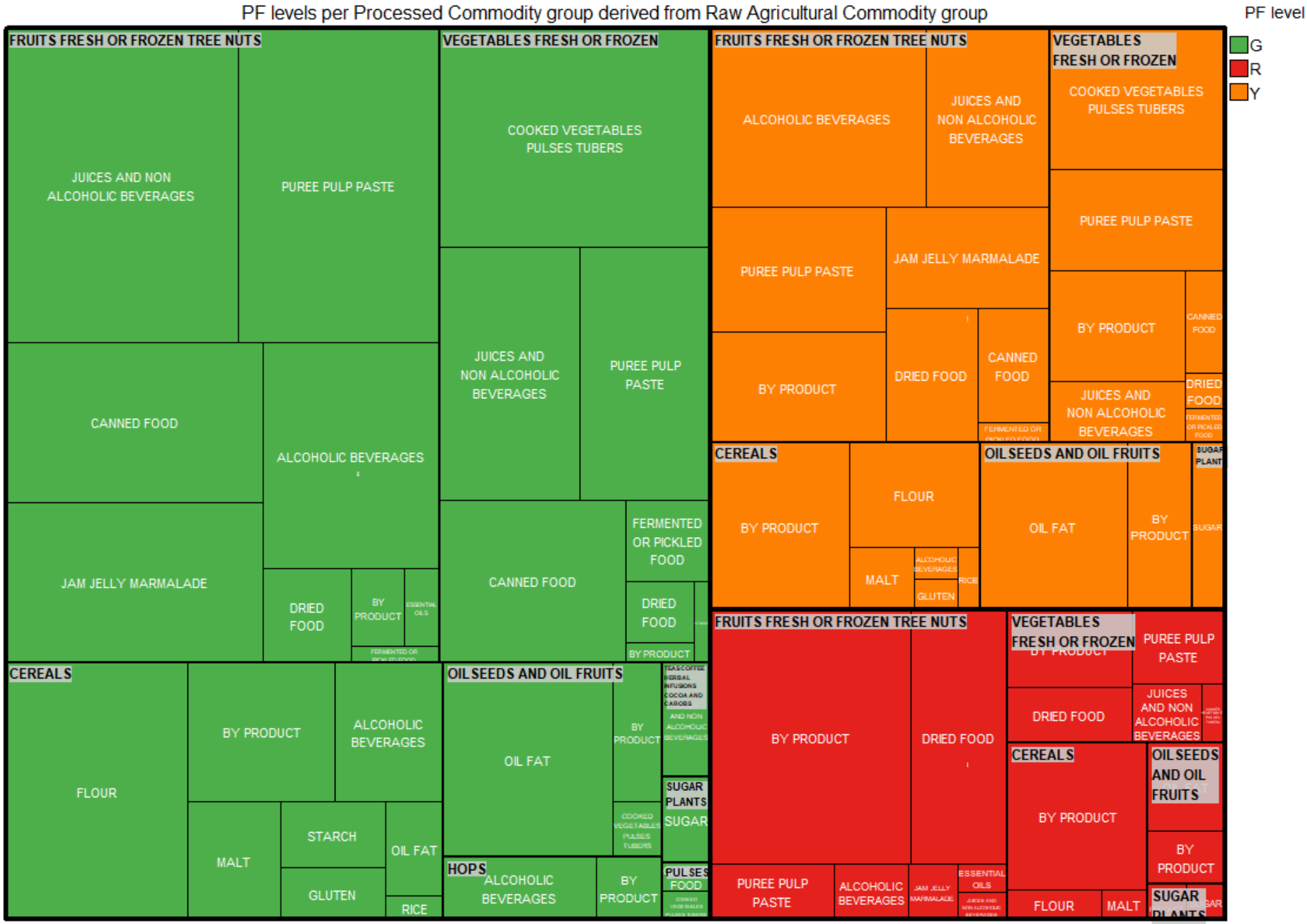


Figure 4-32 All PF database treemap: visualization according to PF levels (G - green, R - red, Y - yellow), RAC groups and PC groups.

Alluvial

Different attempts were made to obtain an alluvial diagram from the whole PF dataset, and the one reported in Fig. 4-33 is the most readable version that has been obtained. The diagram allows us to trace the journey of residues from RAC subgroups, through OECD process stages and PC groups, to PF levels and active substances. A total of five axes are employed.

Green flows are predominant throughout, indicating that many processes lead to low PF levels. This trend is especially evident in processes like cooking in water and canning (both fruit and vegetables), suggesting that these methods are associated with reduced residue concentrations. Conversely, red flows, indicating higher PF levels, are more concentrated in by-products and certain processing methods like fermentation and dehydration, suggesting that these techniques may retain or increase residue levels. The yellow flows are more dispersed across different processes and commodities, appearing in a variety of RAC subgroups and processed products, including both by-products and primary food products. The diversity in yellow flows highlights a moderate PF level variability across processes, without clear clustering.

While this alluvial format provides some level of partitioning of residue contributions across multiple processing stages and active substances, the high density of overlapping pathways, particularly around the PF level and active substance (AS) axes, makes it impossible to trace specific flows from start to end. Label overlap around process codes and active substances further reduces readability, as it becomes difficult to follow individual connections amid the densely packed labels.

Overall, this diagram offers a possible workaround to include more variables in the same visualization to identify how general trends impact PF levels, but the visual density limits its readability for quickly understanding detailed pathways of specific residues or processes.

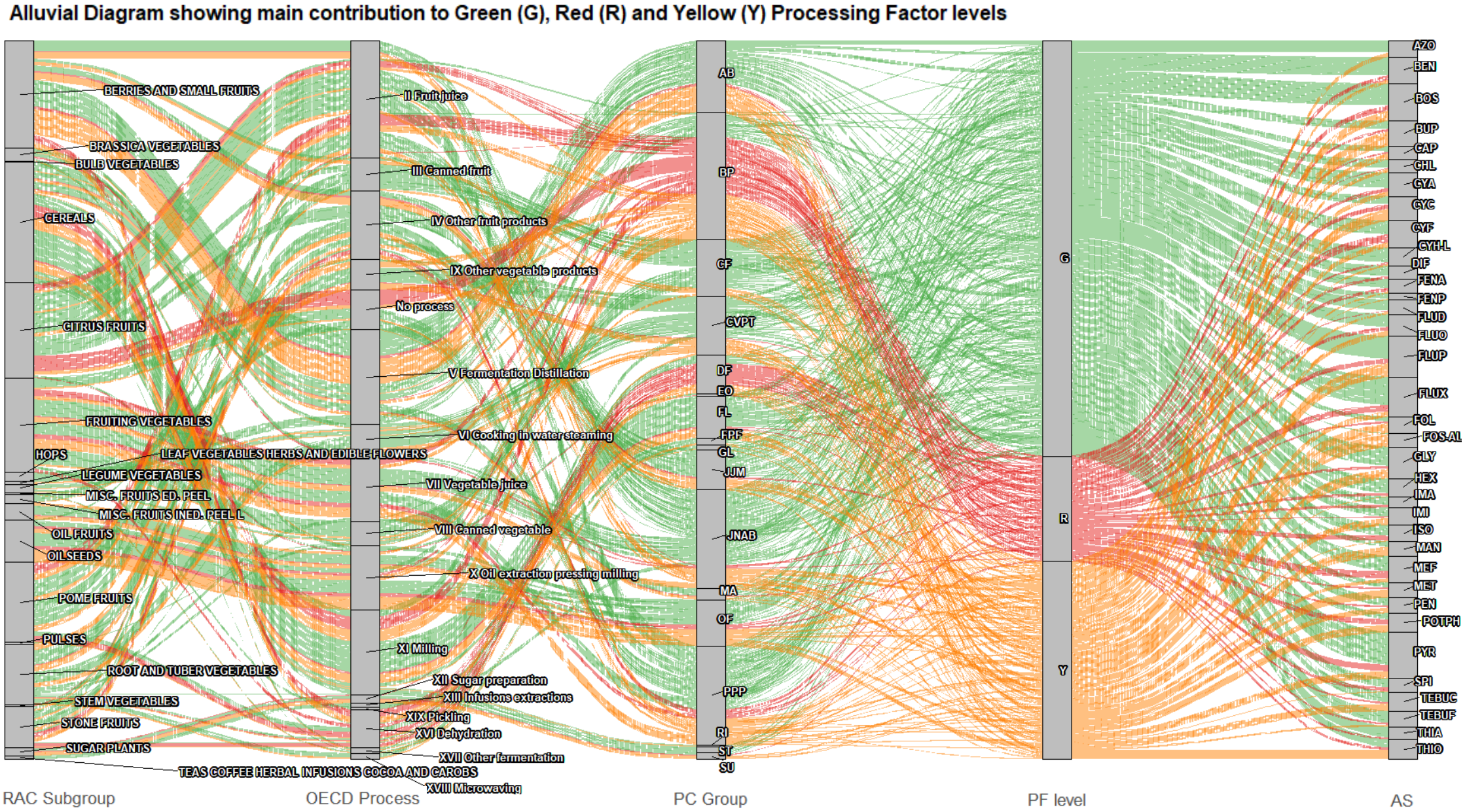


Figure 4-33 All PF database alluvial plot: visualization according to RAC subgroups, OECD processes, PC groups, PF levels (G - green, R - red, Y - yellow), and active substances. For explanation of process codes and other abbreviations see Appendices II, III and IV.

J-PCA

A J-PCA model incorporating four principal components was developed, explaining nearly 55% of total data variance. While the combination of PC1 and PC3 effectively differentiates between RAC groups (Fig. 4-34 A), the distribution of PF levels (Fig. 4-34 B) reveals no distinct pattern within individual RAC groups, as PF levels are broadly distributed across all categories. This suggests that, even when considering the full dataset, the different PF levels are not satisfactorily explained by RAC grouping alone. Instead, PF levels appear to depend more significantly on other factors within the processing itself, in line with the specific examples discussed in previous chapters.

The factors influencing PF levels can be more effectively explored using PC2 and PC4 , where samples are more distinctly clustered by PF classes (Fig. 4-35). In Fig. 4-35 B, J-PCA reveals that green PF samples are more concentrated at negative score values of both PC2 and PC4 while yellow and red PF samples at positive score values of both PCs. A practical rule-of-thumb for separating PF classes could be the bisector of the second and fourth quadrant of PC2 and PC4 score plot (i.e., PC4 = -PC2), reported as a black dashed line in Fig. 4-35 B. By comparing this rule-of-thumb structure with plot A of Fig. 4-35, it can be quickly noted that some relations with processed commodities exist.

These relations are better highlighted in the subsequent figure (Fig. 4-36), where two extreme PF behaviors are selected as examples. In plot A, the samples belonging to canned food group are highlighted in magenta color and it is possible to notice that they are primarily located below the black dashed line, with only sparse occurrences above. In contrast, by-product samples (highlighted in magenta color in Fig. 4-36 B) are mainly concentrated above the black dashed line, indeed showing a high occurrence of red and yellow PF samples. This approach for PF class occurrence evaluation can be expanded to all other processed commodities and processes, offering quick and intuitive visualizations in comparison to the treemap or alluvial diagrams presented before.

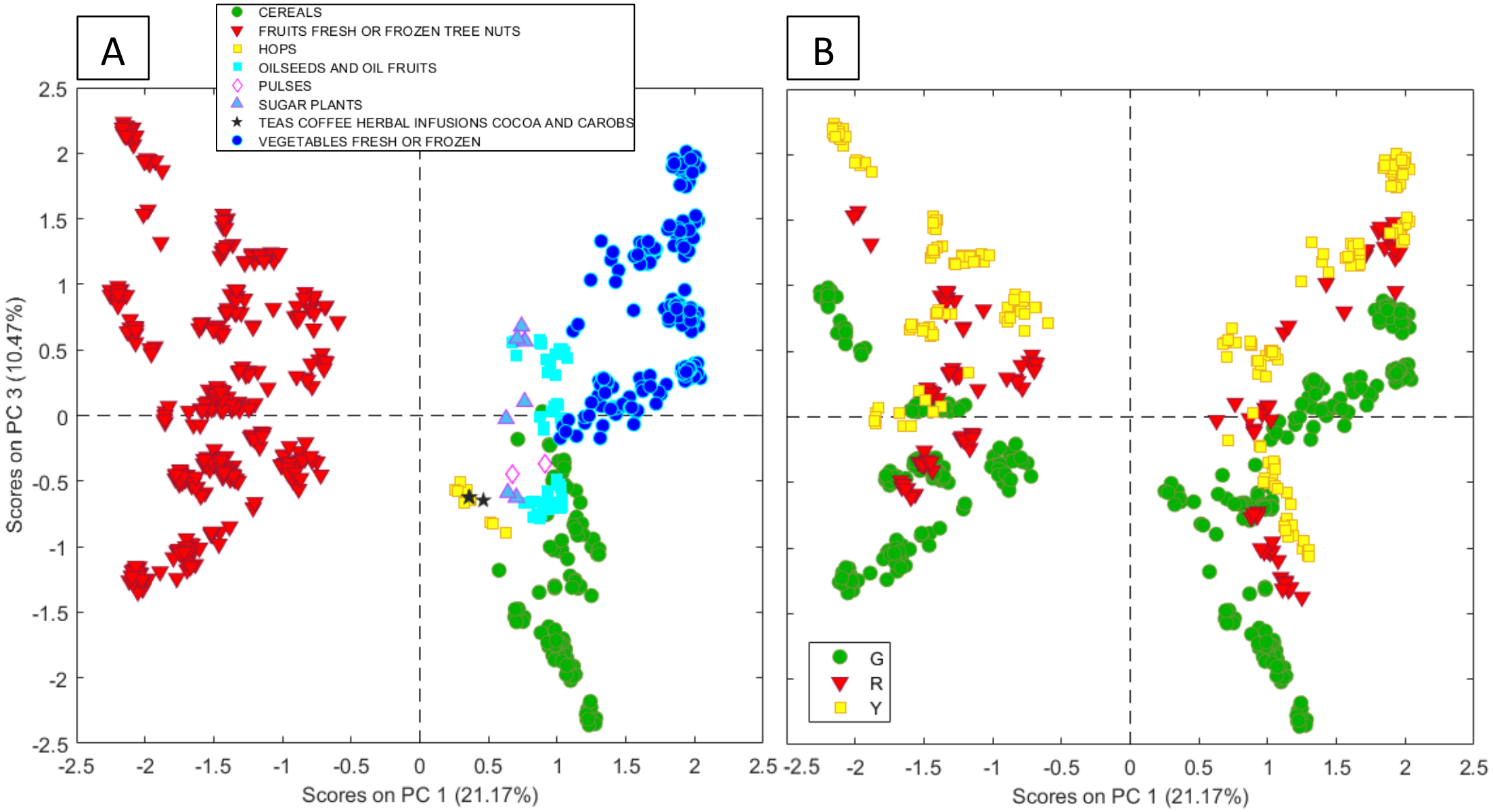


Figure 4-34 All PF database J-PCA: PC1 and PC3 score plot. In A) samples are colored according to RAC groups while in B) samples are colored according to the PF level (G - green, R - red, Y - yellow).

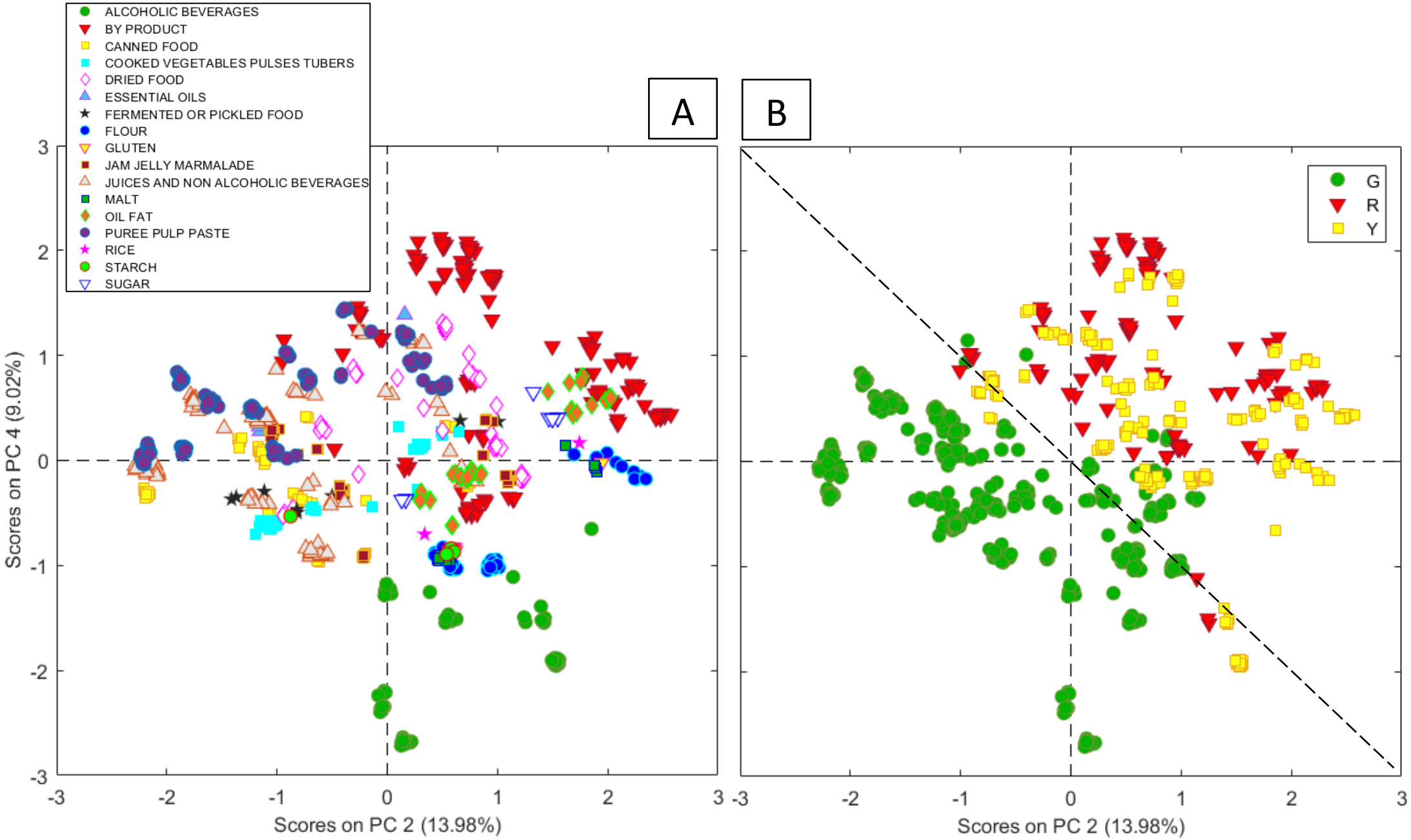


Figure 4-35 All PF database J-PCA: PC2 and PC4 score plot. In A) samples are colored according to PC groups while in B) samples are colored according to the PF level (G - green, R - red, Y - yellow).

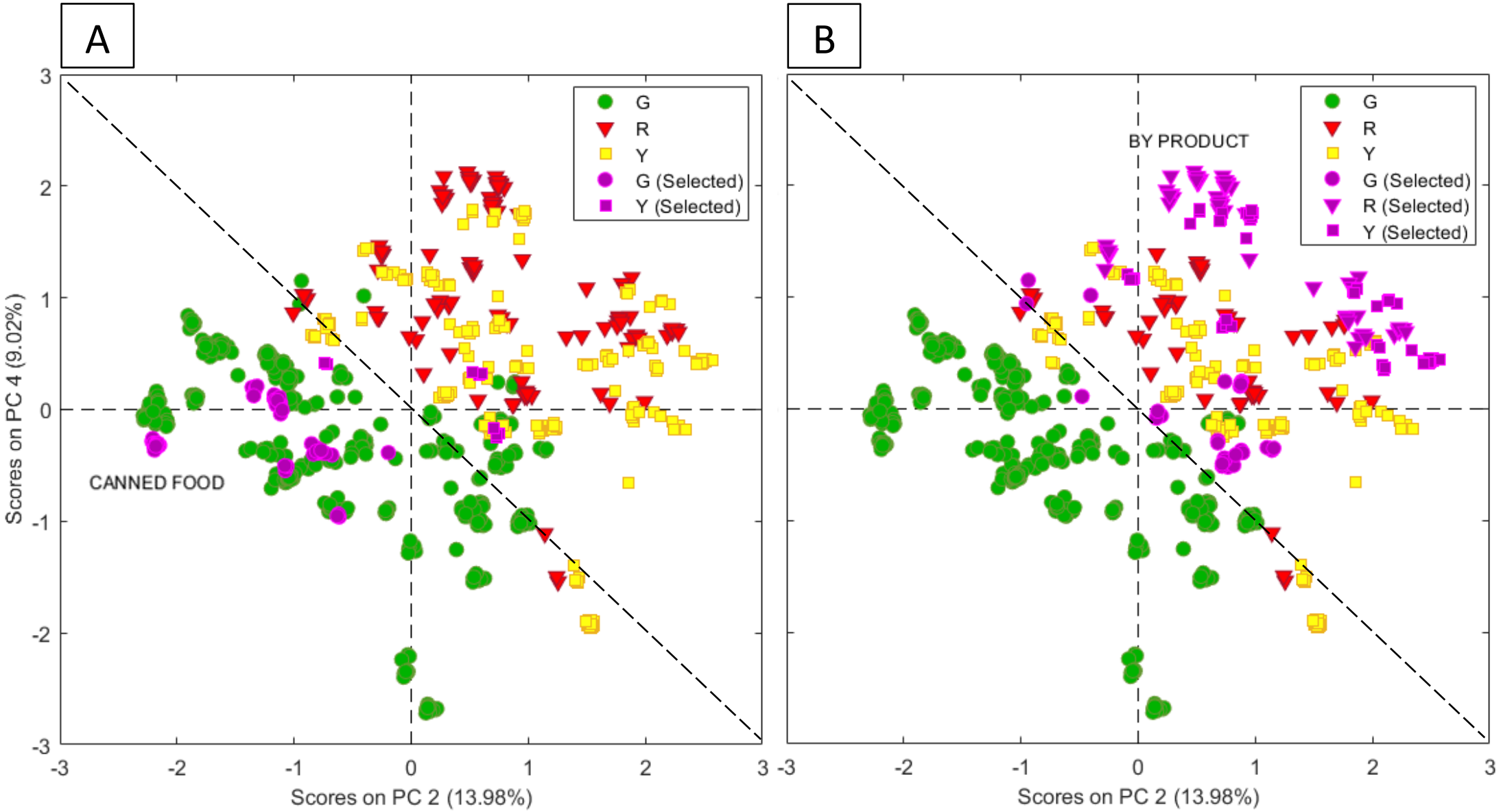


Figure 4-36 All PF database J-PCA: PC2 and PC4 score plot. In both plots samples are colored according to the PF level (G - green, R - red, Y - yellow). In A) canned food samples are selected while in B) by-product samples are selected.

4.4 Conclusions

In this Chapter, the European PF database has been extensively analyzed to evaluate existing relationships between raw agricultural commodities, the transformation product they undergo to obtain final processed commodities, and pesticides residues found in the end products. The analyzed dataset is composed of categorical variables encoding different information layers related to RACs, processes, processed commodities and PF level. To gain an exploratory evaluation of such dataset, different visualization approaches have been tested. At first, common tools available to display categorical data, i.e., treemaps and alluvial plots, were used. These methods were compared with a novel approach developed in this PhD Thesis, which is based on performing PCA on the Jaccard distance matrix obtained from the original dataset of categorical variables, referred throughout the Thesis as J-PCA.

Treemaps are inherently limited by the need to include only a few variables, which restricts their ability to convey complex information and makes it challenging to integrate insights from multiple treemaps. However, they are effective for providing focused visualizations of specific variables.

Alluvial diagrams lose effectiveness in representing detailed information as the number of axes and samples increases. They work well for illustrating overall trends, but the density of overlapping paths and labels can reduce readability.

J-PCA visualizations, on the other hand, provide stability across different levels of analysis. They support detailed inspection of individual samples while retaining general trends, offering a balance between specific detail and a broader overview.

The inclusion of **by-products** as a cross-cutting processed commodity group spanning multiple RACs and processes provides valuable insights for interpreting PF levels. When a process involving one or more RACs produces by-products as final outputs, PF levels (yellow or red) are often concentrated within these by-products, while main processed commodities tend to display a higher frequency of green PF levels. Consequently, processed commodities containing higher concentrations of components typically considered as by-products are more likely to exhibit high PF levels. This trend is due to the tendency of by-products to carry high pesticide residues, which persist in the final product if those by-product components remain included (e.g., the wholemeal flour case). However, exceptions do occur, often explained by domain-specific food technology knowledge or unique pesticide residues properties.

A notable exception to this pattern is the **dehydration** process, where the method itself significantly influences PF levels. Unlike other processes, where by-products generally carry high PF levels (yellow or red) while main products mainly show green PF levels, dehydration reverses this trend. The dehydration process concentrates pesticide residues, making the final dried product the riskier outcome with higher PF levels, rather than confining them to by-products. This illustrates that in specific processes like dehydration, the processing method is the primary driver of high PF levels, shifting the risk from by-products (as seen in other processes) to the main dried product itself. This finding emphasizes the importance of understanding process-specific factors when assessing PF levels and the resulting risks associated with different types of processed commodities.

Chapter 5: Final Considerations

This work explored the potential of chemometric strategies for multivariate data analysis, focusing on two key applications in the food domain: food optimization and food control. Leveraging innovative approaches and adapting established techniques, the research demonstrated the critical role of advanced data analysis in addressing contemporary challenges in food science.

The research on food product optimization underscored the power of chemometric tools in formulating products that meet consumer preferences and industry standards. By employing multivariate techniques for image analysis, experimental design, response surface modeling, PCA and PLS-DA, the study facilitated efficient exploration of formulation parameters. This approach not only minimized experimental efforts but also enabled the identification of optimal conditions for achieving desired product attributes. The integration of these techniques into product development workflows paves the way for creating high-quality, consumer-driven food products while maintaining operational efficiency. The demonstrated method for product optimization is directly applicable to industry practices, bridging the gap between academic research and real-world applications.

In the realm of food control, the thesis presented a pioneering application of multivariate techniques to the European Database of Processing Factors (PF Database). This database, a cornerstone for pesticides residues monitoring, was explored using novel adaptations of chemometric methods, including Jaccard-based Principal Component Analysis (J-PCA). By applying a multivariate perspective, the work revealed hidden patterns and relationships within the dataset, offering insights into pesticides residues behavior across diverse food matrices and processing methods. The study also benchmarked the J-PCA approach against state-of-the-art visualization techniques, such as treemaps and alluvial plots, highlighting the strengths and limitations of each method. While traditional visualization methods provided clear insights only when few categorical variables are simultaneously considered, J-PCA demonstrated superior capacity to capture and interpret complex relationships. The application of J-PCA to categorical data extends the scope of chemometric analysis, providing a new tool for exploratory data analysis in non-numerical datasets. Furthermore, by enhancing the interpretability and accessibility of the PF database, the research supports improved pesticide residue monitoring and dietary risk assessment.

The findings of this thesis open several avenues for further research:

- **Extension of J-PCA applications**: the methodology could be applied to other categorical datasets in food science, such as consumer preference surveys or supply chain analyses, to uncover hidden trends and dependencies.
- **Enhanced database utilization:** future work could focus on integrating machine learning algorithms with the PF database to predict residue behavior under untested scenarios, providing proactive risk assessment tools.
- **Broader chemometrics adoption**: encouraging wider adoption of chemometric strategies in the food industry could improve decision-making processes, from product development to regulatory compliance.

This thesis highlights the transformative potential of chemometric strategies in food science. By combining food optimization and food control applications, the research provides a foundation for advancing food safety, quality, and innovation. As the food industry continues to face evolving challenges, the approaches demonstrated here offer valuable tools for navigating complexity and driving progress.

Appendices

Appendix I – RAC Groups and Subgroups

Below is the full table of RAC groups, subgroups and raw agricultural commodities included in this work. More items are originally present in the second version of European database of processing factors for pesticides residues in food (Zincke et al., 2022).

RAC GROUP	RAC SUBGROUP	RAC
CEREALS	CEREALS	BARLEY GRAINS
		COMMON WHEAT GRAIN
		MAIZE GRAIN
		OAT GRAIN
		RICE GRAIN
		RYE GRAIN
FRUITS (FRESH OR FROZEN) AND TREE NUTS	CITRUS FRUITS	LEMONS
		MANDARINS AND SIMILAR
		ORANGES
	BERRIES AND SMALL FRUITS	CURRANTS (BLACK, RED AND WHITE)
		STRAWBERRIES
		TABLE GRAPES
		WINE GRAPES
	MISCELLANEOUS FRUITS WITH EDIBLE PEEL SMALL (MISC. FRUITS ED. PEEL)	TABLE OLIVES
	MISCELLANEOUS FRUITS WITH INEDIBLE PEEL LARGE (MISC. FRUITS INED. PEEL L)	COMMON BANANA
		MANGOES
		PINEAPPLES
	POME FRUITS	APPLES
		PEARS
	STONE FRUITS	APRICOTS
		CHERRIES (SWEET)
		PEACHES AND SIMILAR
		PLUMS
HOPS	HOPS	HOPS
OILSEEDS AND OIL FRUITS	OILSEEDS	COTTON SEEDS
		PEANUTS
		RAPESEEDS
		SOYABEANS FOR OIL
		SUNFLOWER SEEDS
	OIL FRUITS	OLIVES FOR OIL PRODUCTION
PULSES	PULSES	PEAS (DRY) AND SIMILAR
SUGAR PLANTS	SUGAR PLANTS	SUGAR BEET ROOTS
		SUGAR CANES
		COFFEE BEANS GREEN

TEAS COFFEE HERBAL INFUSIONS COCOA AND CAROBS	TEAS COFFEE HERBAL INFUSIONS COCOA AND CAROBS	TEAS LEAVES DRY AND OR FERMENTED AND SIMILAR
VEGETABLES FRESH OR FROZEN	BRASSICA VEGETABLES	HEAD CABBAGES
	BULB VEGETABLES	ONIONS
	FRUITING VEGETABLES	CHILI PEPPERS
		GHERKINS
		MELONS
		TOMATOES
	LEAF VEGETABLES HERBS AND EDIBLE FLOWERS	SPINACHES
	LEGUME VEGETABLES	BEANS (FRESH SEEDS WITHOUT PODS) AND SIMILAR
		BEANS (WITH PODS) AND SIMILAR
		PEAS (WITHOUT PODS) AND SIMILAR
		SOYABEANS FOR CONSUMPTION (DRY)
	ROOT AND TUBER VEGETABLES	CARROTS
		MAIN CROP POTATOES
	STEM VEGETABLES	LEEKS

Appendix II – Processing Techniques (with OECD groups)

Below is the full table of codified processing techniques included in this work, according to OECD grouping (OECD, 2008b, 2008a) and Scholz et al. (2022). More items are included in second version of European database of processing factors for pesticides residues in food Zincke et al. (2022). Non-processed samples, such as those derived from raw pulp/peel trials in fruits, initially had an empty value in the process column of PF database. These were explicitly labeled with "No process" (process code “No”) to ensure clarity and to distinguish them from codified processes.

OECD CODE	OECD PROCESS GROUP	PROCESS CODE	SPECIFIC PROCESS NAME
II	Fruit juice	II 001	Citrus juice citrus fruits
		II 002	Pome juice pome fruits (apples pears)
		II 003	Grape juice berries and small fruits (currants)
			Grape juice grapes
		II 004	Stone fruit juice stone fruits
III	Canned fruit	III 001	Canned fruits berries and small fruits
			Canned fruits citrus fruits (mandarins oranges)
			Canned fruits pome fruits (apples pears)
			Canned fruits stone fruits
IV	Other fruit products	IV 001	Jelly apples
			Jelly berries and small fruits (grapes currants)
			Jam berries and small fruits (strawberries currants)
			Jam stone fruits (apricots cherries peaches plums)
			Marmalade citrus fruits (oranges mandarins)
		IV 002	Fruit purée pome fruits (apples pears)
		IV 003	Fruit purée stone fruits
V	Fermentation Distillation	V 001	White wine production wine grapes
		V 002	Red wine production wine grapes
		V 003	Rosé wine production wine grapes
		V 005	Beer brewing barley grain to malt
			Beer brewing barley malt to beer
		V 006	Beer brewing hops
		V 007	Distillates wine grapes
VI	Cooking in water steaming	VI 001	Cooking in water brassica vegetables
			Cooking in water carrots
			Cooking in water leaf vegetables (spinach)
			Cooking in water potatoes
			Cooking in water stem vegetables (leeks)
		VI 002	Cooking in water beans peas
			Cooking in water peas (with pod) peas (without pod)
		VI 003	Cooking in water pulses
		VI 005	Steaming potatoes
VII	Vegetable juice	VII 001	Paste tomatoes

			Vegetable purée tomatoes
		VII 002	Vegetable juices tomatoes
		VII 003	Vegetable juices carrots
		VII 004	Vegetable juices
VIII	Canned vegetable	VIII 001	Canned fruiting vegetables tomatoes
		VIII 002	Canned vegetables legume vegetables (peas beans)
			Canned vegetables pulses (peas beans)
			Canned vegetables root and tuber
IX	Other vegetable products	IX 001	Deep frying (chips French fries) potatoes
		IX 002	Deep frying (crisps) potatoes
		IX 003	(Pan)frying potatoes
		IX 004	Baking potatoes
		IX 005	Roasting peanuts
		IX 006	Soya drink and tofu soya beans
X	Oil extraction pressing milling	X 001	Oil production olives olives for oil production
		X 002	Oil production maize dry milling
		X 003	Oil production oil oilseeds
		X 005	Production of essential oils citrus fruits (oranges)
XI	Milling	XI 002	Milling flour wheat rye
		XI 003	Milling pearl barley flour Barley
		XI 004	Milling flour maize
		XI 005	Milling rolled oats oat
		XI 006	Milling rice processing rice
		XI 008	Starch production potatoes
		XI 009	Starch production cereals (wheat sorghum)
XII	Sugar preparation	XII 001	Sugar sugar beet roots
		XII 002	Sugar sugar canes
XIII	Infusions extractions	XIII 001	Roasting coffee beans
		XIII 003	Infusion tea
XVI	Dehydration	XVI 001	Drying berries and small fruits
			Drying pome fruits (apples pears)
			Drying stone fruits (plums apricots peaches)
			Drying table grapes
		XVI 002	Drying Fruiting vegetables (chili peppers)
			Drying Fruiting vegetables (tomatoes)
			Drying onions
			Drying potatoes
		XVI 004	Potato flakes granules potatoes
XVII	Other fermentation	XVII 001	Fermentation to sauerkraut head cabbages
		XVII 002	Fermentation of fruits table olives
		XVII 004	Rice wine processing
XVIII	Microwaving	XVIII 001	Microwaving potatoes
XIX	Pickling	XIX 001	Pickling of vegetables gherkins

Appendix III – Processed Commodities Groups

Below is the full table of processed commodities groups and processed commodities included in this work. More items are originally present in second version of European database of processing factors for pesticides residues in food (Zincke et al., 2022).

PROCESSED COMMODITY GROUP	PROCESSED COMMODITY
ALCOHOLIC BEVERAGES (AB)	BEER
	MUST
	SAKE
	SPIRIT
	WINE RED
	WINE ROSÉ
	WINE WHITE
BY PRODUCT (BP)	BRAN
	BREWER'S GRAIN
	GERMS
	GLUTEN FEED MEAL
	HULLS
	MALT SPROUTS
	MEAL EXTRACTED
	PEEL
	POMACE DRY
	POMACE WET
	PULP DRIED (if coming from all process other than XVI Dehydration)
	SPENT HOPS
CANNED FOOD (CF)	FRUIT CANNED
	POD CANNED
	ROOT BODY CANNED
	SEED CANNED
COOKED VEGETABLES PULSES TUBERS (CVPT)	CRISPS PEELED
	CRISPS PEELED (RECALCULATED FOR POTATO PART ONLY)
	CRISPS UNPEELED
	CRISPS UNPEELED (RECALCULATED FOR POTATO PART ONLY)
	HEAD COOKED
	LEAVES COOKED
	PEANUT BUTTER
	PEANUTS ROASTED
	POD COOKED
	ROOT BODY PEELED COOKED
	SEED COOKED
	TUBER BAKED UNPEELED
	TUBER COOKED PEELED

	TUBER COOKED UNPEELED
	TUBER DEEP FRIED
	TUBER DEEP FRIED (RECALCULATED FOR POTATO PART ONLY)
	TUBER DEEP FRIED PEELED
	TUBER DEEP FRIED PEELED (RECALCULATED FOR POTATO PART ONLY)
	TUBER MICROWAVE COOKED PEELED
	TUBER MICROWAVE COOKED UNPEELED
	TUBER PEELED
	TUBER PEELED COOKED
	TUBER PEELED FRIED
	TUBER PEELED FRIED (RECALCULATED FOR POTATO PART ONLY)
	TUBER PEELED MICROWAVE COOKED
	TUBER STEAMED UNPEELED
	VEGETABLE COOKED
DRIED FOOD (DF)	BULB DRIED
	FLAKES GRANULES
	FRUIT DRIED
	PULP DRIED (if coming from XVI Dehydration process)
	RAISIN
ESSENTIAL OILS (EO)	OIL
FERMENTED OR PICKLED FOOD (FPF)	FRUIT CANNED
	FRUIT FERMENTED
	FRUIT FERMENTED STERILISED
	SAUERKRAUT PASTEURISED
FLOUR (FL)	FLOUR
	FLOUR WHITE
	FLOUR WHOLEMEAL
	GRITS
	POT PEARL BARLEY
	ROLLED OATS
	FLOUR BROWN
GLUTEN (GL)	GLUTEN
JAM JELLY MARMALADE (JJM)	FRUIT COOKED (FROM JAM)
	FRUIT COOKED (FROM JELLY)
	JAM
	JELLY
	MARMALADE
JUICES AND NON ALCOHOLIC BEVERAGES (JNAB)	COFFEE BEAN ROASTED
	INSTANT COFFEE
	JUICE
	JUICE (CALCULATED FROM NECTAR)
	JUICE CLARIFIED
	JUICE CLARIFIED PASTEURISED

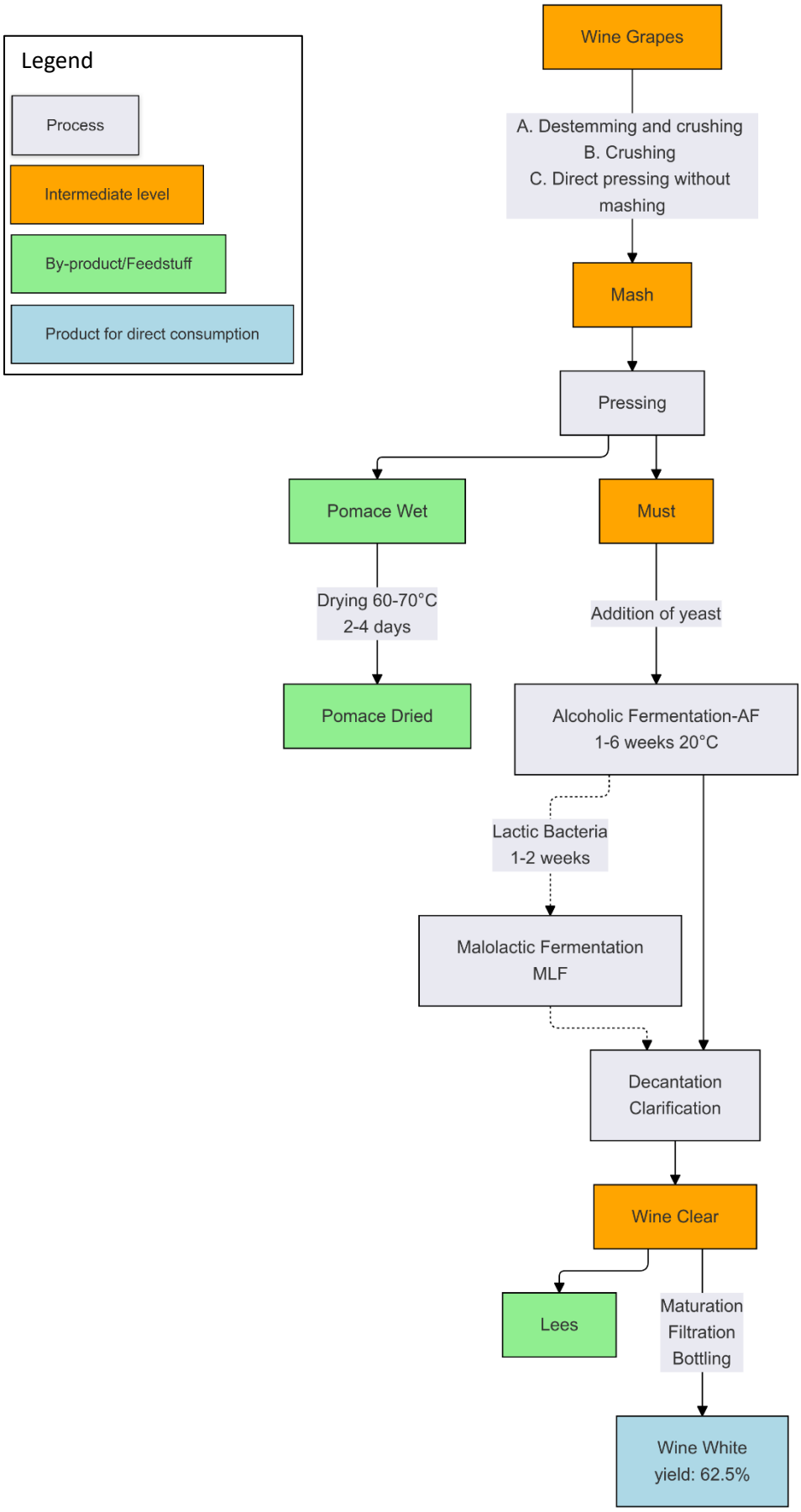
	JUICE PASTEURISED
	JUICE RAW
	JUICE STERILISED
	NECTAR
	SOYA DRINK
	TEA INFUSION
MALT (MA)	MALT
OIL FAT (OF)	OIL CRUDE (COMBINED)
	OIL EXTRACTED
	OIL EXTRACTED REFINED
	OIL NATIVE
	OIL PRESSED
	OIL PRESSED REFINED
	OIL REFINED (COMBINED)
PUREE PULP PASTE (PPP)	PASTE
	PULP
	PUREE
RICE (RI)	RICE POLISHED
STARCH (ST)	STARCH
SUGAR (SU)	MOLASSES
	SUGAR RAW
	SUGAR REFINED

Appendix IV – Active Substances with WHO Risk Class and Use

Full table of active substances included in this work. More items are present in second version of European database of processing factors for pesticides residues in food (Zincke et al., 2022). For WHO risk classification and main use *see*: *The WHO Recommended Classification of Pesticides by Hazard and Guidelines to Classification, 2019 Edition*.

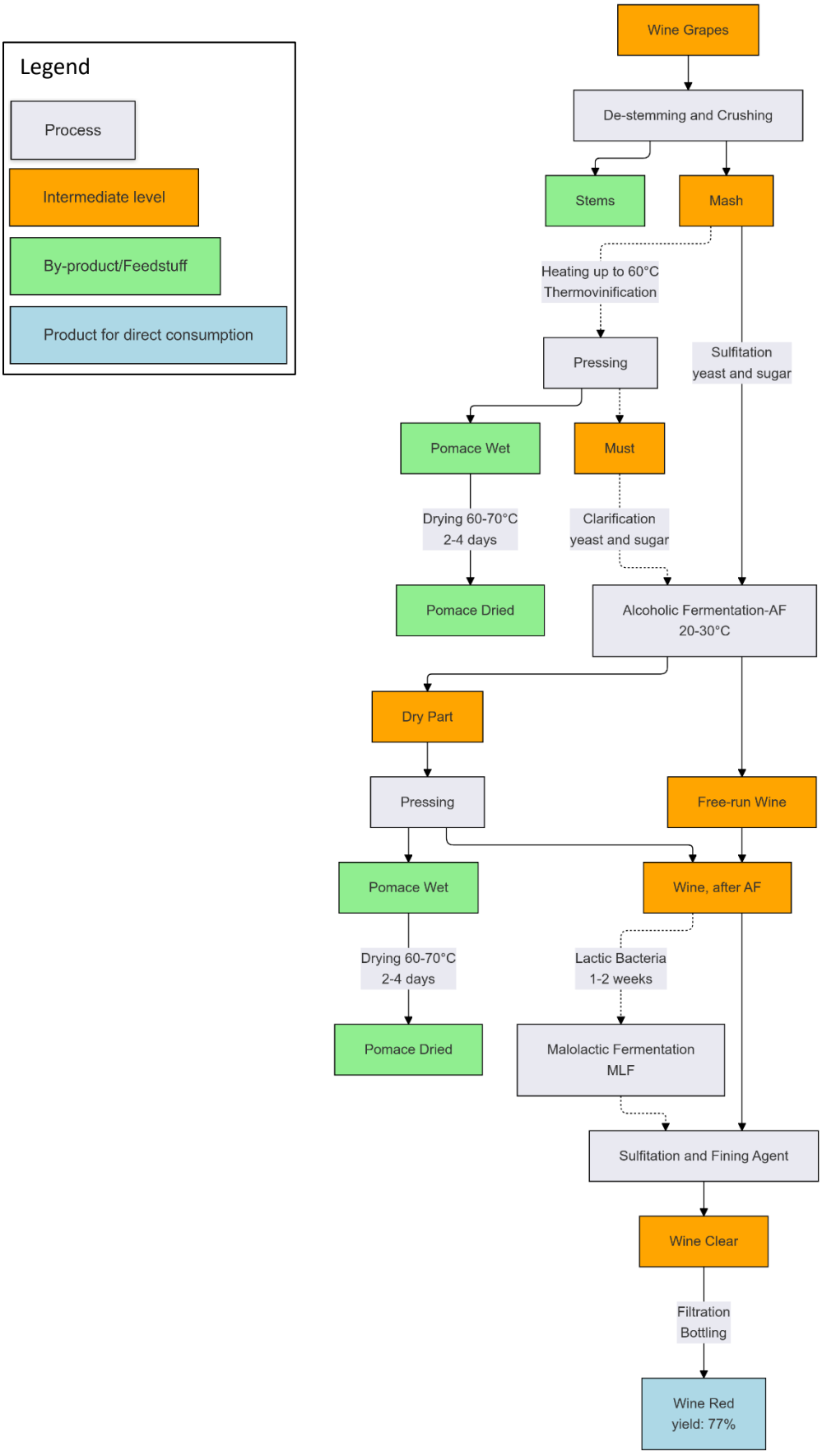
WHO RISK CLASS	**WHO MAIN USE** **AC** acaricide; **F** fungicide, other than for seed treatment; **H** herbicide; **I** insecticide; **PGR** plant growth regulator	**AS**	**ACTIVE SUBSTANCE**
II Moderately hazardous	F	BEN	BENZOVINDIFLUPYR
		DIF	DIFENOCONAZOLE
		IMA	IMAZALIL
		ISO	ISOPYRAZAM
		TEBUC	TEBUCONAZOLE
	I	CYH-L	CYHALOTHRIN LAMBDA
		FLUP	FLUPYRADIFURONE
		IMI	IMIDACLOPRID
	AC	FENA	FENAZAQUIN
III Slightly hazardous	F	BUP	BUPIRIMATE
		FENP	FENPICOXAMID
		FLUO	FLUOPYRAM
		FLUX	FLUXAPYROXAD
		MEF	MEFENTRIFLUCONAZOLE
		PEN	PENTHIOPYRAD
		PYR	PYRACLOSTROBIN
		THIA	THIABENDAZOLE
	AC	CYF	CYFLUMETOFEN
		SPI	SPIRODICLOFEN
	H	CYC	CYCLOXYDIM
		GLY	GLYPHOSATE
U Unlikely to present acute hazard	F	AZO	AZOXYSTROBIN
		BOS	BOSCALID
		CAP	CAPTAN
		FLUD	FLUDIOXONIL
		FOL	FOLPET
		FOS-AL	FOSETYL ALUMINIUM
		MAN	MANCOZEB
		MET	METIRAM
		POTPH	POTASSIUM PHOSPHONATES
		THIO	THIOPHANATE METHYL
	I	CYA	CYANTRANILIPROLE
		TEBUF	TEBUFENOZIDE
	AC	HEX	HEXYTHIAZOX
	PGR	CHL	CHLORPROPHAM

Appendix V – White Wine Process Flowchart (V 001)



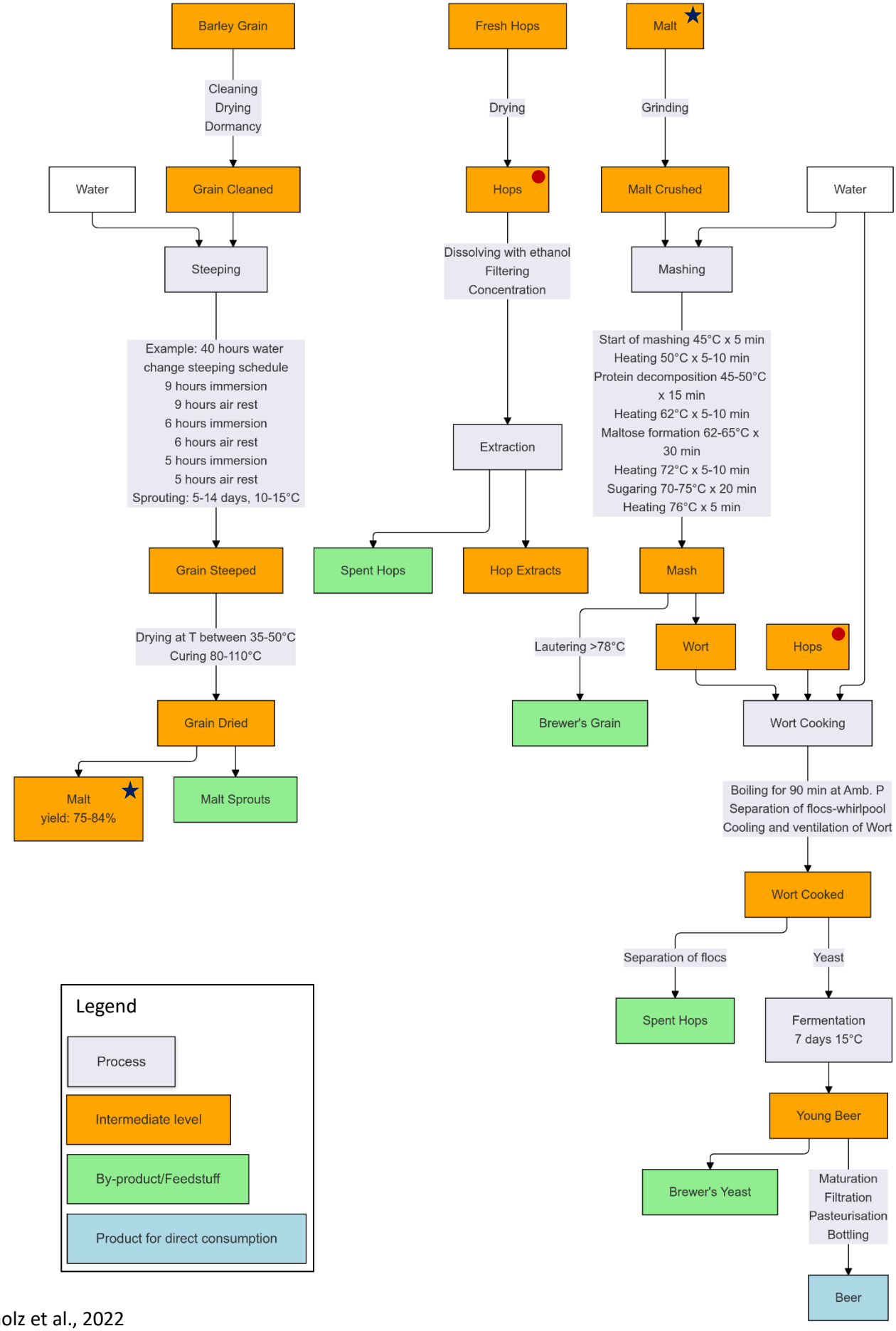
Scholz et al., 2022

Appendix VI – Red Wine Process Flowchart (V 002)



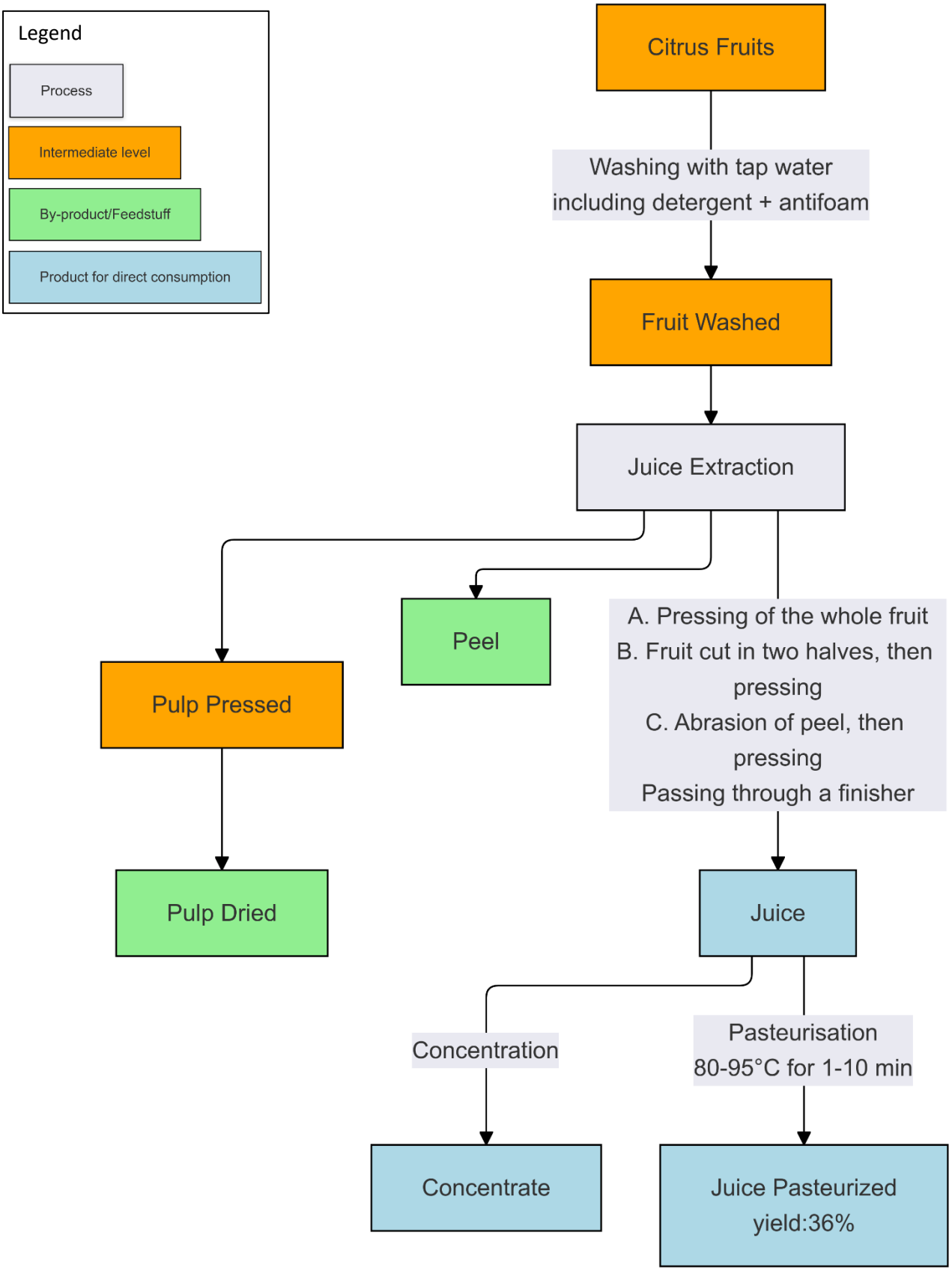
Scholz et al., 2022

Appendix VII – Beer Process Flowchart (V 005 – V 006)



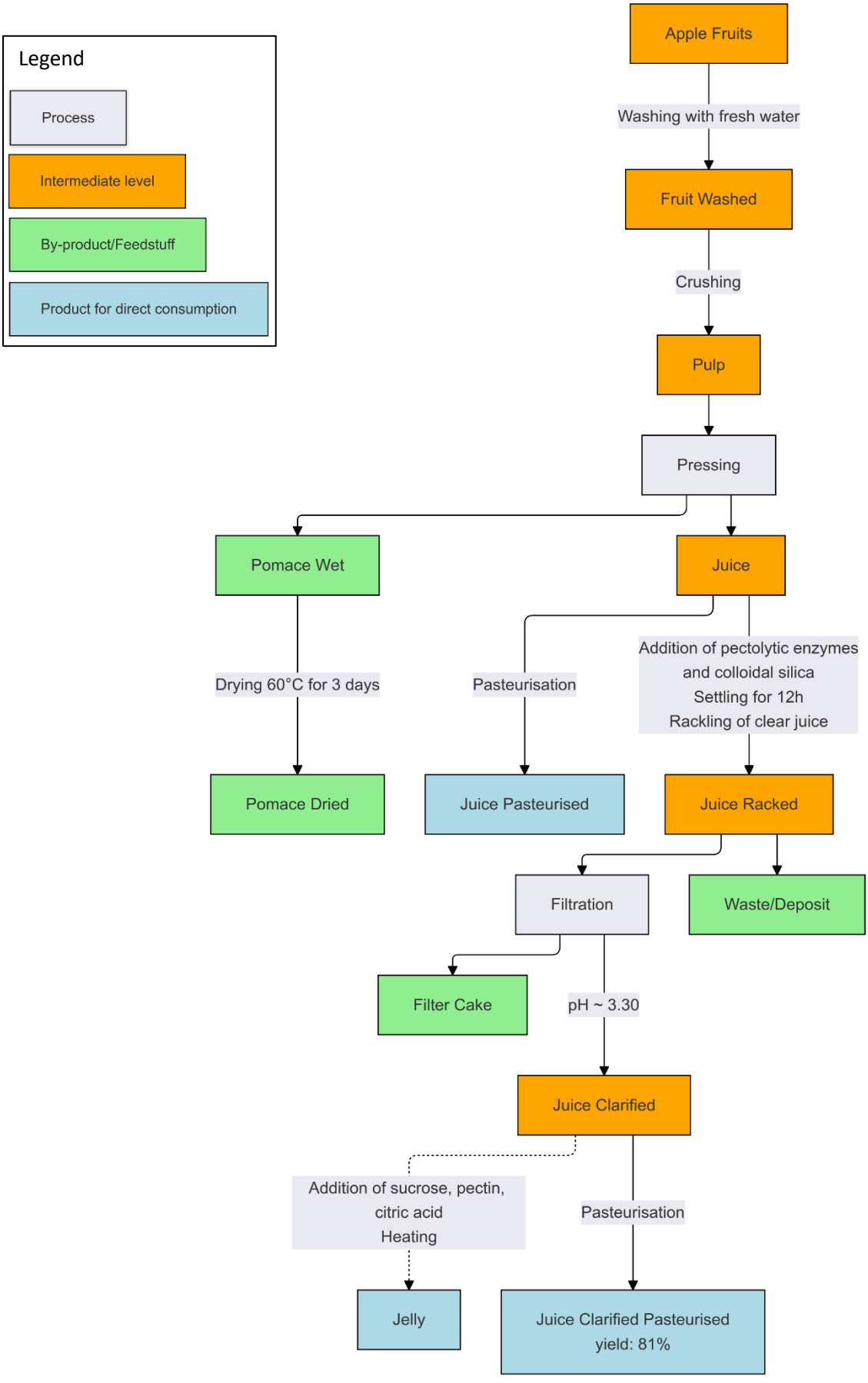
Scholz et al., 2022

Appendix VIII – Citrus Juice Flowchart (II 001)



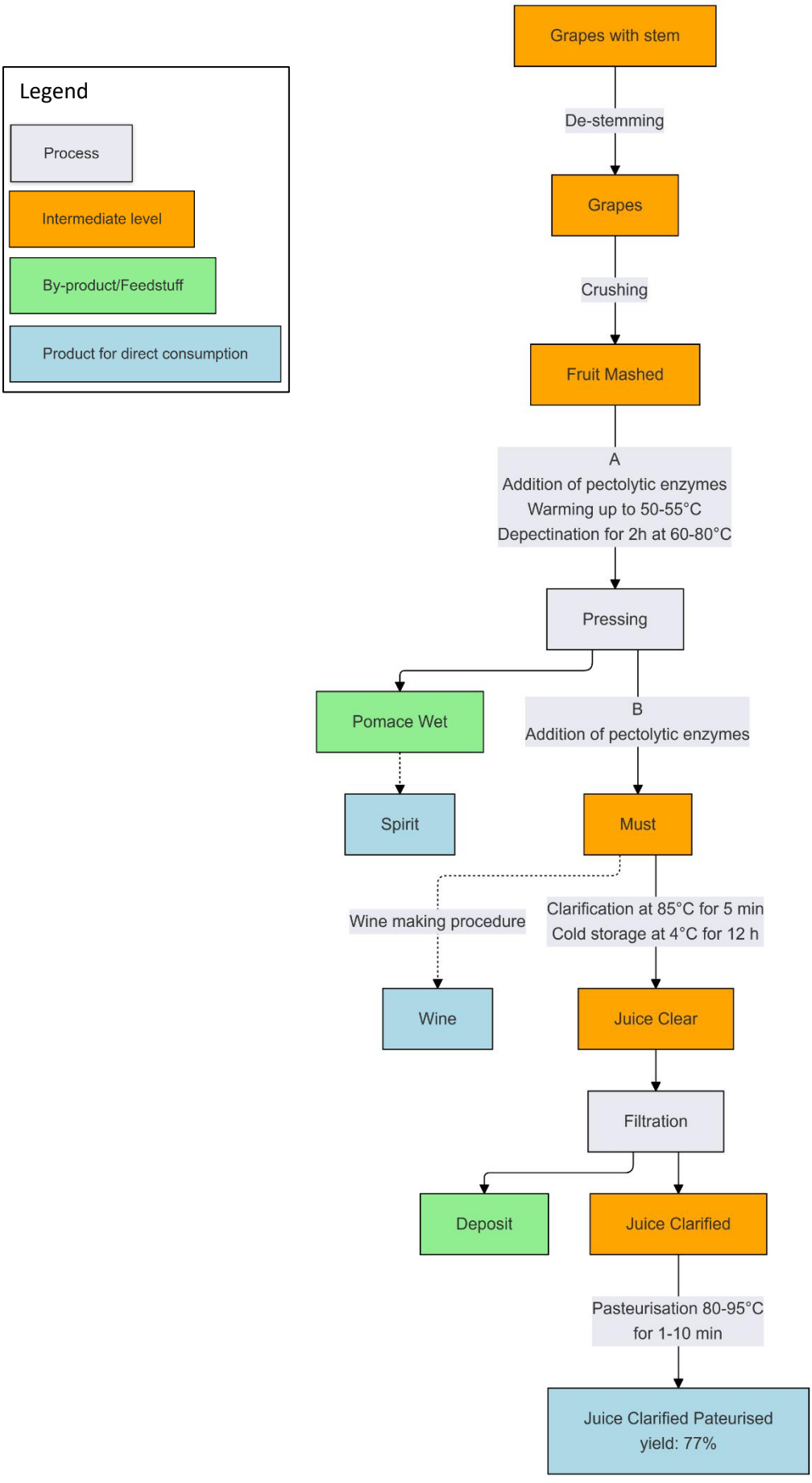
Scholz et al., 2022

Appendix IX – Pome Juice Flowchart (II 002)



Scholz et al., 2022

Appendix X – Grape Juice Flowchart (II 003)



Scholz et al., 2022

References

Antonelli, A., Cocchi, M., Fava, P., Foca, G., Franchini, G. C., Manzini, D., & Ulrici, A. (2004). Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Analytica Chimica Acta*, *515*(1), 3–13. https://doi.org/10.1016/j.aca.2004.01.005

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790–3798. https://doi.org/10.1039/C3AY40582F

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, *174*, 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004

Ballabio, D., & Todeschini, R. (2009). *Multivariate Classification for Qualitative Analysis*. NL. https://doi.org/10.1016/B978-0-12-374136-3.00004-3

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Benedetti, B., Caponigro, V., & Ardini, F. (2020). Experimental Design Step by Step: A Practical Guide for Beginners. *Critical Reviews in Analytical Chemistry*. https://doi.org/10.1080/10408347.2020.1848517

Borin, A., Ferrão, M. F., Mello, C., Cordi, L., Pataca, L. C. M., Durán, N., & Poppi, R. J. (2007). Quantification of Lactobacillus in fermented milk by multivariate image analysis with least-squares support-vector machines. *Analytical and Bioanalytical Chemistry*, *387*(3), 1105–1112. https://doi.org/10.1007/s00216-006-0971-7

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/C3AY41907J

Brunson, J. C., & Read, Q. D. (2023). *ggalluvial: Alluvial Plots in "ggplot2"* (Version 0.12.5) [Computer software]. https://cran.r-project.org/package=ggalluvial

Bursać Kovačević, D., Putnik, P., Dragović-Uzelac, V., Vahčić, N., Babojelić, M. S., & Levaj, B. (2015). Influences of organically and conventionally grown strawberry cultivars on anthocyanins content and color in purees and low-sugar jams. *Food Chemistry*, *181*, 94–100. https://doi.org/10.1016/j.foodchem.2015.02.063

Cabras, P., Angioni, A., Caboni, P., Garau, V. L., Melis, M., Pirisi, F. M., & Cabitza, F. (2000). Distribution of Folpet on the Grape Surface after Treatment. *Journal of Agricultural and Food Chemistry*, *48*(3), 915–916. https://doi.org/10.1021/jf990069u

Calvaruso, E., Cammilleri, G., Pulvirenti, A., Lo Dico, G. M., Lo Cascio, G., Giaccone, V., Vitale Badaco, V., Ciprì, V., Alessandra, M. M., Vella, A., Macaluso, A., Di Bella, C., & Ferrantelli, V. (2020). Residues of 165 pesticides in citrus fruits using LC-MS/MS: A study of the pesticides distribution from the peel to the pulp. *Natural Product Research*, *34*(1), 34–38. https://doi.org/10.1080/14786419.2018.1561682

Calvini, R., Orlandi, G., Foca, G., & Ulrici, A. (2020). Colourgrams GUI: A graphical user-friendly interface for the analysis of large datasets of RGB images. *Chemometrics and Intelligent Laboratory Systems*, *196*, 103915. https://doi.org/10.1016/j.chemolab.2019.103915

Caramês, E. T. dos S., Baqueta, M. R., Conceição, D. A., & Pallone, J. A. L. (2021). Near infrared spectroscopy and smartphone-based imaging as fast alternatives for the evaluation of the bioactive potential of freeze-dried açai. *Food Research International*, *140*, 109792. https://doi.org/10.1016/j.foodres.2020.109792

Colourgrams GUI – DOWNLOADS – Chimslab. (n.d.). https://www.chimslab.unimore.it/downloads/

Čuš, F., Česnik, H. B., Bolta, Š. V., & Gregorčič, A. (2010). Pesticide residues in grapes and during vinification process. *Food Control*, *21*(11), 1512–1518. https://doi.org/10.1016/j.foodcont.2010.04.024

Da Silva, F. L., Escribano-Bailón, M. T., Pérez Alonso, J. J., Rivas-Gonzalo, J. C., & Santos-Buelga, C. (2007). Anthocyanin pigments in strawberry. *LWT - Food Science and Technology*, *40*(2), 374–382. https://doi.org/10.1016/j.lwt.2005.09.018

Davies, A. M. C., & Fearn, T. (2004). Back to basics: The principles of principal component analysis. *Tony Davies Column*, 20–23.

Delgado-Vargas, F., Jiménez, A. R., & Paredes-López, O. (2000). Natural Pigments: Carotenoids, Anthocyanins, and Betalains — Characteristics, Biosynthesis, Processing, and Stability. *Critical Reviews in Food Science and Nutrition*, *40*(3), 173–289. https://doi.org/10.1080/10408690091189257

Dzhanfezova, T., Barba-Espín, G., Müller, R., Joernsgaard, B., Hegelund, J. N., Madsen, B., Larsen, D. H., Martínez Vega, M., & Toldam-Andersen, T. B. (2020). Anthocyanin profile, antioxidant activity and total phenolic content of a strawberry (Fragaria × ananassa Duch) genetic resource collection. *Food Bioscience*, *36*, 100620. https://doi.org/10.1016/j.fbio.2020.100620

Ertan, K., Türkyılmaz, M., & Özkan, M. (2020). Color and stability of anthocyanins in strawberry nectars containing various co-pigment sources and sweeteners. *Food Chemistry*, *310*, 125856. https://doi.org/10.1016/j.foodchem.2019.125856

European Food Safety Authority (EFSA). (2015). The food classification and description system FoodEx 2 (revision 2). *EFSA Supporting Publications*, *12*(5), 804E. https://doi.org/10.2903/sp.efsa.2015.EN-804

European Food Safety Authority (EFSA). (2022, September 28). *2022 Eurobarometer on Food Safety in the EU | EFSA*. https://www.efsa.europa.eu/en/corporate/pub/eurobarometer22

Foca, G., Masino, F., Antonelli, A., & Ulrici, A. (2011). Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques. *Analytica Chimica Acta*, *706*(2), 238–245. https://doi.org/10.1016/j.aca.2011.08.046

Giraudo, A., Calvini, R., Orlandi, G., Ulrici, A., Geobaldo, F., & Savorani, F. (2018). Development of an automated method for the identification of defective hazelnuts based on RGB image analysis and colourgrams. *Food Control*, *94*, 233–240. https://doi.org/10.1016/j.foodcont.2018.07.018

Goos, P., Jones, B., & Syafitri, U. (2016). I-Optimal Design of Mixture Experiments. *Journal of the American Statistical Association*, *111*(514), 899–911. https://doi.org/10.1080/01621459.2015.1136632

Gosselin, R., Rodrigue, D., & Duchesne, C. (2010). A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometrics and Intelligent Laboratory Systems*, *100*(1), 12–21. https://doi.org/10.1016/j.chemolab.2009.09.005

Han, J., Fang, P., Xu, X., Li-Zheng, X., Shen, H., & Ren, Y. (2015). Study of the pesticides distribution in peel, pulp and paper bag and the safety of pear bagging. *Food Control*, *54*, 338–346. https://doi.org/10.1016/j.foodcont.2015.02.021

Hartmann, A., Patz, C.-D., Andlauer, W., Dietrich, H., & Ludwig, M. (2008). Influence of Processing on Quality Parameters of Strawberries. *Journal of Agricultural and Food Chemistry*, *56*(20), 9484–9489. https://doi.org/10.1021/jf801555q

Holzwarth, M., Korhummel, S., Siekmann, T., Carle, R., & Kammerer, D. R. (2013). Influence of different pectins, process and storage conditions on anthocyanin and colour retention in strawberry jams and spreads. *LWT - Food Science and Technology*, *52*(2), 131–138. https://doi.org/10.1016/j.lwt.2012.05.020

Holzwarth, M., Wittig, J., Carle, R., & Kammerer, D. R. (2013). Influence of putative polyphenoloxidase (PPO) inhibitors on strawberry (Fragaria x ananassa Duch.) PPO, anthocyanin and color stability of stored purées. *LWT - Food Science and Technology*, *52*(2), 116–122. https://doi.org/10.1016/j.lwt.2012.10.025

Jankowska, M., & Łozowicka, B. (2022). The processing factors of canning and pasteurization for the most frequently occurring fungicides and insecticides in apples and their application into dietary risk assessment. *Food Chemistry*, *371*, 131179. https://doi.org/10.1016/j.foodchem.2021.131179

Konishi, T., Matsukuma, S., Fuji, H., Nakamura, D., Satou, N., & Okano, K. (2019). Principal Component Analysis applied directly to Sequence Matrix. *Scientific Reports*, *9*(1), 19297. https://doi.org/10.1038/s41598-019-55253-0

Liu, Z., Yang, S., Wang, Y., & Zhang, J. (2021). Discrimination of the fruits of *Amomum tsao-ko* according to geographical origin by 2DCOS image with RGB and Resnet image analysis techniques. *Microchemical Journal*, *169*, 106545. https://doi.org/10.1016/j.microc.2021.106545

Martinsen, B. K., Aaby, K., & Skrede, G. (2020). Effect of temperature on stability of anthocyanins, ascorbic acid and color in strawberry and raspberry jams. *Food Chemistry*, *316*, 126297. https://doi.org/10.1016/j.foodchem.2020.126297

Masino, F., Foca, G., Ulrici, A., Arru, L., & Antonelli, A. (2008). A chemometric study of pesto sauce appearance and of its relation to pigment concentration. *Journal of the Science of Food and Agriculture*, *88*(8), 1335–1343. https://doi.org/10.1002/jsfa.3221

Menozzi, C., Calvini, R., Nigro, G., Tessarin, P., Bossio, D., Calderisi, M., Ferrari, V., Foca, G., & Ulrici, A. (2023). Design and application of a smartphone-based device for in vineyard determination of anthocyanins content in red grapes. *Microchemical Journal*, *191*, 108811. https://doi.org/10.1016/j.microc.2023.108811

OECD. (2008a). *Guidance Document on Magnitude of Pesticide Residues in Processed Commodities*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/environment/guidance-document-on-magnitude-of-pesticide-residues-in-processed-commodities_e2b71e69-en

OECD. (2008b). *Test No. 508: Magnitude of the Pesticide Residues in Processed Commodities*. OECD. https://doi.org/10.1787/9789264067622-en

Ohta, N., & Robertson, A. R. (2005). *Colorimetry Fundamentals and Applications* (M. A. Kriss, A. C. Lowe, L. W. MacDonald, & Y. Miyake, Eds.). John Wiley & Sons, Ltd.

Orlandi, G., Calvini, R., Foca, G., Pigani, L., Simone, G. V., & Ulrici, A. (2019). Data fusion of electronic eye and electronic tongue signals to monitor grape ripening. *Talanta*, *195*, 181–189. https://doi.org/10.1016/j.talanta.2018.11.046

Orlandi, G., Calvini, R., Foca, G., & Ulrici, A. (2018). Automated quantification of defective maize kernels by means of Multivariate Image Analysis. *Food Control*, *85*, 259–268. https://doi.org/10.1016/j.foodcont.2017.10.008

Orlandi, G., Calvini, R., Pigani, L., Foca, G., Vasile Simone, G., Antonelli, A., & Ulrici, A. (2018). Electronic eye for the prediction of parameters related to grape ripening. *Talanta*, *186*, 381–388. https://doi.org/10.1016/j.talanta.2018.04.076

Pagnin, L., Calvini, R., Wiesinger, R., Weber, J., & Schreiner, M. (2020). Photodegradation Kinetics of Alkyd Paints: The Influence of Varying Amounts of Inorganic Pigments on the Stability of the Synthetic Binder. *Frontiers in Materials*, *7*. Scopus. https://doi.org/10.3389/fmats.2020.600887

Prats-Montalbán, J. M., De Juan, A., & Ferrer, A. (2011). Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 1–23. https://doi.org/10.1016/j.chemolab.2011.03.002

R Consortium (Director). (2021, August 8). *Graphing multivariate categorical data: The how, what and why of mosaic plots and alluvial diagrams* [Video recording]. https://www.youtube.com/watch?v=fIB-FlTHheI

Regulation (EC) No 396/2005 of the European Parliament and of the Council of 23 February 2005 on Maximum Residue Levels of Pesticides in or on Food and Feed of Plant and Animal Origin and Amending Council Directive 91/414/EEC (Consolidated Text Version of 14/10/2024), Pub. L. No. 396/2005 (2005). https://eur-lex.europa.eu/eli/reg/2005/396/2016-05-13

Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 Laying down the General Principles and Requirements of Food Law, Establishing the European Food Safety Authority and Laying down Procedures in Matters of Food Safety (Consolidated Text of 01/07/2024) (2002). http://data.europa.eu/eli/reg/2002/178/2024-07-01/eng

Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 Concerning the Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC (Consolidated Text Version of 21/11/2022) (2009). http://data.europa.eu/eli/reg/2009/1107/2022-11-21/eng

Robbins, J. (2021). *Jtr13/graphcat* [HTML]. https://github.com/jtr13/graphcat (Original work published 2021)

Rolando, P. L., Calvini, R., Foca, G., & Ulrici, A. (2023). Mixture design and multivariate image analysis to monitor the colour of strawberry yoghurt purée. *Microchemical Journal*, *194*, 109222. https://doi.org/10.1016/j.microc.2023.109222

Sadilova, E., Carle, R., & Stintzing, F. C. (2007). Thermal degradation of anthocyanins and its impact on color and *in vitro* antioxidant capacity. *Molecular Nutrition & Food Research*, *51*(12), 1461–1471. https://doi.org/10.1002/mnfr.200700179

Santos, P. M., Wentzell, P. D., & Pereira-Filho, E. R. (2012). Scanner Digital Images Combined with Color Parameters: A Case Study to Detect Adulterations in Liquid Cow's Milk. *Food Analytical Methods*, *5*(1), 89–95. https://doi.org/10.1007/s12161-011-9216-2

Scholz, R., Donkersgoed, G., Herrmann, M., Kittelmann, A., Kraus, C., Schledorn, M., Mahieu, C., Velde-Koerts, T., Anagnostopoulos, C., Bempelou, E., & Michalski, B. (2022). *Compendium of Representative Processing Techniques Investigated in Regulatory Studies for Pesticides*. https://doi.org/10.5281/zenodo.6564208

Scholz, R., Herrmann, M., Kittelmann, A., Schledorn, M., Zincke, F., Donkersgoed, G., Graven, C., Velde-Koerts, T., Anagnostopoulos, C., Bempelou, E., & Michalski, B. (2022). *Background Document on the EU Database of Processing Factors for Pesticide Residues*. https://doi.org/10.5281/zenodo.6564214

Scholz, R., van Donkersgoed, G., Herrmann, M., Kittelmann, A., von Schledorn, M., Graven, C., Mahieu, K., van der Velde-Koerts, T., Anagnostopoulos, C., Bempelou, E., & Michalski, B. (2018). Database of processing techniques and processing factors compatible with the EFSA food classification and description system FoodEx 2 Objective 3: European database of processing factors for pesticides in food. *EFSA Supporting Publications*, *15*(11), 1510E. https://doi.org/10.2903/sp.efsa.2018.EN-1510

Shewhart, W. A., Wilks, S. S., Bloomfield, P., Cressie, N. A. C., Fisher, N. I., Johnstone, M., Kadane, J. B., Ryan, L. M., Scott, D. W., Silverman, B. W., Smith, A. F. M., Teugels, J., Barnett, V., Hunter, J. S., & Kendall, D. G. (2002). *Experiments with Mixtures – Designs, Models, and the Analysis of Mixture Data, Third Edition*.

Solana-Altabella, A., Sánchez-Iranzo, M. H., Bueso-Bordils, J. I., Lahuerta-Zamora, L., & Mellado-Romero, A. M. (2018). Computer vision-based analytical chemistry applied to determining iron in commercial pharmaceutical formulations. *Talanta*, *188*, 349–355. https://doi.org/10.1016/j.talanta.2018.06.008

Sulaiman, A., & Silva, F. V. M. (2013). High pressure processing, thermal processing and freezing of 'Camarosa' strawberry for the inactivation of polyphenoloxidase and control of browning. *Food Control*, *33*(2), 424–428. https://doi.org/10.1016/j.foodcont.2013.03.008

Teixeira, M. J., Aguiar, A., Afonso, C. M. M., Alves, A., & Bastos, M. M. S. M. (2004). Comparison of pesticides levels in grape skin and in the whole grape by a new liquid chromatographic multiresidue methodology. *Analytica Chimica Acta*, *513*(1), 333–340. https://doi.org/10.1016/j.aca.2003.11.077

Tennekes, M., & Ellis, P. (2023). *treemap: Treemap Visualization* (Version 2.4-4) [Computer software]. https://cran.r-project.org/package=treemap

The WHO Recommended Classification of Pesticides by Hazard and guidelines to classification, 2019 edition. (n.d.). Retrieved May 13, 2024, from https://www.who.int/publications-detail-redirect/9789240005662

Todeschini, R., Ballabio, D., & Consonni, V. (2020). Distances and Similarity Measures in Chemometrics and Chemoinformatics. In *Encyclopedia of Analytical Chemistry* (pp. 1–40). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470027318.a9438.pub2

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., & Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, *52*(11), 2884–2901. https://doi.org/10.1021/ci300261r

Ulrici, A., Foca, G., Ielo, M. C., Volpelli, L. A., & Lo Fiego, D. P. (2012). Automated identification and visualization of food defects using RGB imaging: Application to the detection of red skin defect of raw hams. *Innovative Food Science & Emerging Technologies*, *16*, 417–426. https://doi.org/10.1016/j.ifset.2012.09.008

Wu, D., & Sun, D.-W. (2013). Colour measurements by computer vision for food quality control – A review. *Trends in Food Science & Technology*, *29*(1), 5–20. https://doi.org/10.1016/j.tifs.2012.08.004

Zincke, F., Fischer, A., Kittelmann, A., Kraus, C., Scholz, R., Michalski, B., & BfR (German Federal Institute for Risk Assessment). (2022). First update of the EU database of processing factors for pesticide residues. *EFSA Supporting Publications*, *19*(9). https://doi.org/10.2903/sp.efsa.2022.EN-7453