Journal of Food Research; Vol. 6, No. 3; 2017
ISSN 1927-0887 E-ISSN 1927-0895
Published by Canadian Center of Science and Education

Digital Fingerprinting of Coffee Blending by Sensitive Crystallization

Pietro Guccione[1], Antonella Galati[2], Emanuela Pusceddu[3] & Rocco Caliandro[2]

[1]Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, via Orabona 4, 70125, Bari, Italy

[2]Institute of Crystallography, CNR, via Amendola, 122/o, 70126, Bari, Italy

[3]Institute of Biometeorology, CNR, via Caproni 8, 50145 Firenze, Italy

Correspondence: Pietro Guccione, Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, via Orabona 4, 70125, Bari, Italy. Tel: 39-080-596-3925. E-mail: pietro.guccione@poliba.it

Received: January 17, 2017 Accepted: February 13, 2017 Online Published: March 22, 2017

doi:10.5539/jfr.v6n3p21 URL: https://doi.org/10.5539/jfr.v6n3p21

Abstract

The authenticity and quality of productions is an area of priority interest that involves safety of consumers and potential economic damages deriving from frauds on origin, adulteration and labeling of products. Several investigation techniques are currently used to characterize food matrices from physical-chemical-biological point of view using different methods in order to limit possible adulterations. In this work, we have developed an experimental and computational framework to improve the potentialities of sensitive crystallization: an experimental technique known since 1936, but never used for quantitative assessment of food quality. As a test case, it has been applied to investigate the geographical traceability and quality of coffee samples. An extensive statistical analysis associated with a careful choice of advanced image descriptors allows gathering quantitative information about the samples, which can constitute a digital fingerprint of their composition. With this new tool we are able to distinguish with blind tests high-quality coffee brands from low-quality mixtures, different coffee species, green from toasted condition of beans and, to a lesser extent, the macro-geographical provenience. A powder X-ray diffraction analysis reinforces the results obtained by sensitive crystallization for the case where crystalline domains are present in the coffee sample.

Keywords: sensitive crystallization, food analysis, food digital fingerprint, multivariate statistical analysis, coffee

1. Introduction

The analysis of agricultural products with the purpose of finding quality markers, product origin, processing conditions or presence of pollutants or adulteration, has become more and more important in the last years, especially in the context of a global market, with the introduction of alien species or with the difficulty in protecting areas of excellence for specific products.

The scientific community developed different approaches to fight frauds in the agro-food sector, focusing on the characterization of wine, wheat, rice, olive oil. Many experimental techniques are currently available to characterize food composition and for the detection of food adulteration and contamination. They include mass spectrometry, nuclear magnetic resonance, infrared and Raman spectroscopy, X-ray powder diffraction, thermo-gravimetry and differential scanning calorimetry (Wanga S., Wanga J., Yub J. and Wanga S., 2016; Ellis et al., 2012; Anderson, Moore, Tarczynski & Walker, 2001).

The aspects related to geographical origin of food products are in the spotlight of scientific investigations (Augagneur, Médina, Szpunar & Lobiński, 1996; González, Armenta and De la Guardia, 2011; Garrett et al., 2013; Zhao, Zhang S. and Zhang Z., 2017; Beltrán, Sánchez-Astudillo, Aparicio and García-González, 2015; Geana et al., 2014). For instance, the analysis by mass spectrometer inductively coupled plasma (ICP-MS) has been used (Greenough, Longerich and Jackson, 1996), demonstrating that the composition of multiple elements in wine is strongly influenced by the solubility of inorganic compounds found into the soil. Neutron activation analysis showed that the distribution of lanthanides in soil is reproduced in the roots, leaves, and finally in the grapes and in grape-must with no significant changes. In particular, the distribution of concentration of lanthanides atoms in both soil and in the wine has been successfully tested with ICP-MS (Taylor, Longerich and Greenough, 2003; Geana et al., 2014), although highlighting the limits of this method in accounting for the modifications in elemental composition of matrices that follows food transformation processes. In the last few years, attention has been given to authentication of the geographical origin of olive oil assessing the profiles of volatile compounds (Vichi, Pizzale, Conte, Buxaderas and López-Tamames, 2007; Romero, Saavedra, Tapia, Sepùlveda and Aparicio, 2015; Bajouba et al., 2017), sterols (Alves, Cunha, Amaral, Pereira and Oliveira, 2005; Merchaka et al., 2017), with nuclear magnetic resonance method (Rezzi et al., 2005) and stable isotope ratios and elemental mineral content. However, none of these is able to unambiguously determine the entire food sample properties.

Sensitive crystallization, originally introduced by E. Pfeiffer in the 1930ies, is based on the idea of dissolving the substances to be analyzed in salted solutions, and to reveal them by evaporating the liquids. Salt crystals are thus generated with reproducible crystal patterns, which are influenced by the presence of additives in the crystallization solution (Kleber and Steinike-Hartung, 1959). This method has been applied to differentiate organic and conventional food samples (Cocude, 1998), to study the effect of different stages of freshness and degradation on carrot quality (Le Gia, Teisseron, Michel and Cauffet, 1996; Andersen 2001; Andersen, Kaack, Nielsen, Thorup-Kristensen and Labouriau, 2001), and the effect of mineral N fertilization and light intensity on the barley properties (Andersen, 2001; Andersen, Henriksen, Laursen and Nielsen, 1999), as well as used as a diagnostic tool by comparing the human blood from healthy and unhealthy patients (Shibata, Tanaka, Kogure, Iguchi and Ogawa, 1998; Piva et al., 1994; Piva, 1998).

Being crystal growth influenced by interactions of surface salt molecules with additive molecules, sensitive crystallization is a method for investigating substances at molecular level, and it is in principle able to detect the presence of even trace elements in foods. Besides sensitivity, sensitive crystallization is very easy to be implemented, does not need expensive instruments, and can be carried out in a high-throughput way, with many samples analyzed at the same time.

On the other hand, the main difficulty arises from the interpretation of the crystal morphology and in defining its relationships with the properties of the substance used as crystallization additive. Most of the analysis is carried out by qualitative visual inspection of the crystal pattern formed on glass plates (Baumgartner, Doesburg, Scherr and Andersen, 2012), and few attempts have been made to perform quantitative analysis by using image-processing methods (Andersen, Henriksen, Laursen and Nielsen, 1999; Meelusarn 2006).

In this paper we demonstrate that sensitive crystallization can be used for food fingerprinting, if complemented by an extensive computational effort, which includes the extraction of quantitative indicators from digital crystal images and a multivariate analysis to link them with food properties. We considered coffee as a case study: it is considered as sensible product due to its large manipulation and importance in the international market. Two main Coffea (Rubiaceae) species were considered: C. arabica L. (Arabica coffee), which provides more than 95% of the world's coffee (Vega, Rosenquist and Collins, 2003), and C. canephora (Robusta coffee) (Davis, Govaerts, Bridson and Stoffelen, 2006).

The paper is organized as follows. Section 2 includes the description of the samples, of the protocol followed to implement sensitive crystallization and of the multivariate analytical processing. In Section 3 the results of the analysis are reported and discussed, and conclusions are drawn in Section 4.

2. Method

2.1 Coffee Samples

Coffee samples from different geographical areas have been provided as green or toasted beans and then crushed to get powder. Some samples were provided directly in powder. Samples have been grouped in brands (9 brands) or for the status of the bean: green/toasted (2 groups), as shown in Table 1. For each brand and status, 8 repetitions have been prepared, to validate the statistical analysis by accounting for the natural within-group variability. Therefore, a total of 120 samples have been prepared. In literature, two main species of coffee are well described: Arabica (C. arabica) and Robusta (C. canephora) (Bicho, Lidon, Ramalho and Leitao, 2013; Cagliani, Pellegrino, Giugno and Consonni, 2013). The peculiarity of the first one is the elongated shape of the beans; on the contrary, Robusta beans are more rounded (Figure 1). Regardless of the country of origin, the two species have been identified thanks to information provided by the suppliers of the samples; a careful visual inspection of the beans shape has confirmed the information. This allowed the following classification:

Peru, Colombia, Costa Rica and Ethiopia beans are Arabica species from different areas of each country;

for Brazil and Guatemala, we have beans belonging to both species and both of them have been used for crystallization experiments.

In Table 1, the identified species has been reported in the fourth column. The brands named 'Powder1', 'Powder2' and 'Powder3' have been provided in powder, since taken from specific commercial brands, so no information on the beans species was available. For such samples, we may hypothetize that a blend of the two species has been prepared, as often occurs for many cheap commercial brands in the coffee market.

Table 1. List of coffee samples

Sample name	**# samples from green beans**	**# samples from toasted beans / powder**	**Species**
Peru	8	8	Arabica
Brazil	8	8	Robusta/Arabica
Colombia	8	8	Arabica
Costa Rica	8	8	Arabica
Ethiopia (Sidemo)	8	8	Arabica
Guatemala	8	8	Robusta/Arabica
Powder1	0	8	Blend
Powder2	0	8	Blend
Powder3	0	8	Blend

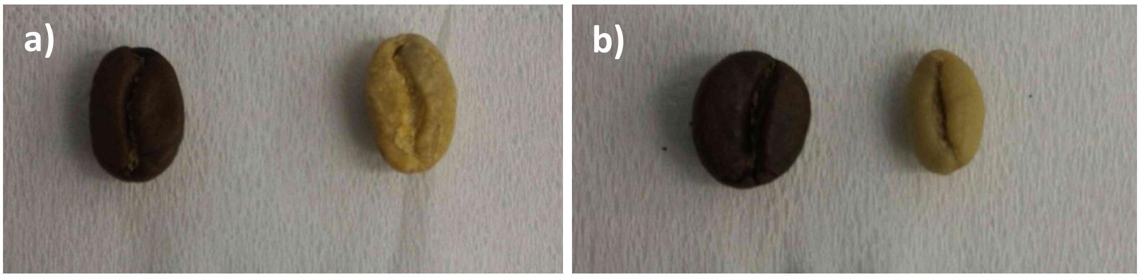


Figure 1. Toasted (left) and green (right) coffee beans belonging to the Arabica (a) and Robusta (b) species.

2.2 Sensitive Crystallization Protocol

Copper (II) chlorydedihydrate (CuCl2 *2H2O) from Sigma Aldrich was used as salt for sensitive crystallization. It was solubilized by using distilled water obtained commercially. Coffee beans were used as additives. They were milled, with a mean particles size of approximately 0.8 mm, and suspended in distilled water for 2h 45 min. The additive solution was then filtered by a 0.45 µm membrane. The crystallization solution for each sample is composed by 132 µl of additive solution, 121 µl of distilled water and 147 µl of an aqueous solution of CuCl2*2H2O (0.3 mol*l-1), in order to obtain a global concentration of 0.11 mol*l-1. Crystallization plates from Hampton Research containing 24 wells of 1.5 cm diameter, typically used for protein crystallization, were used to implement the sensitive crystallization. They were preliminarily sterilized through several passages: an immersion of 15 minutes in a commercial disinfectant (Amuchina), rinsed repeatedly under cold, then hot distilled water and immersed for 10 minutes in boiled distilled water. After this treatment, the plates were dried in an oven at 100 ℃ and the crystallization solutions were promptly pipetted in a crystallization plate. The plates were stored in a room at controlled temperature of 21 ℃ and covered from dust; crystals were obtained in one week.

2.3 Crystal Images Acquisition

After crystals of CuCl2 have appeared, optical microscopic images have been taken by a Nikon A100 stereomicroscope. An image for each of the 120 samples has been taken with the same magnification (10x) and by using polarized light. Images have been digitalized by using an image acquisition system connected to the stereomicroscope. As a visual example of the crystal growth, just three samples are shown in Figure 2.

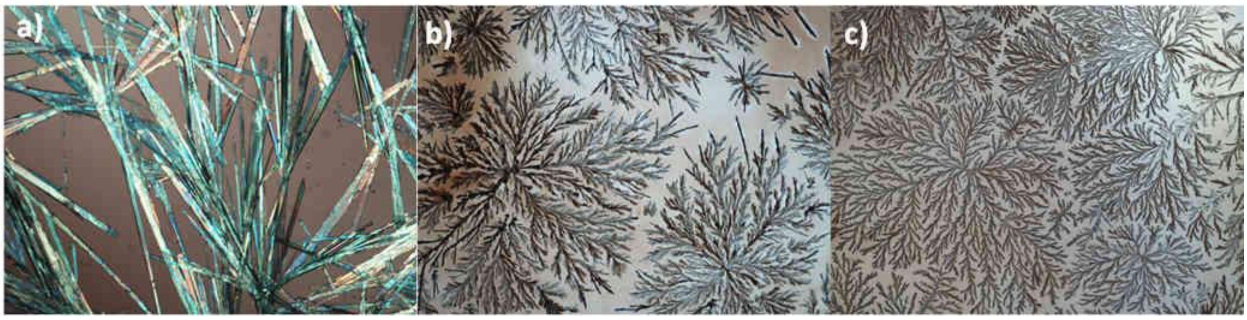


Figure 2 Comparison of different $CuCl_2$ crystals obtained by our sensitive crystallization procedure. Crystal of $CuCl_2$ without additive (a); crystals of $CuCl_2$ obtained by using *Powder2* (b) and *Guatemala* (toasted beans) (c) samples as additive. The magnification is 10x. The different color of the background is due to the different polarized light conditions and does not affect the acquired features.

2.4 Digital Data Processing

Figure 2 shows that crystals may have different characteristics due to the specificity of the additive: the variation throughout the image of the intensity signal, the texture (i.e. the microscopic structure), the density of occupation of the crystal in the slide, the crystal growth and branching (somewhat connected to the degree of fractality of the image) and many others. In literature, it has been stated that these characteristics may be captured using some statistical tools applied on the gray-level image (Andersen, Henriksen, Laursen and Nielsen, 1999; Meelusarn, 2006). The general idea of the method is then to exploit such features to characterize the samples and enlighten the differences, according to a given degree of statistical significance. To this purpose, the repetition of the same sample is used to account for the statistical variability within the same group.

In detail, the following steps are performed:

1. Pre-processing, which includes:
    a. Conversion of images from color to gray level. Images are captured in full color.
    b. Detection and selection of the Region of Interest (ROI) within the image. Not all the parts of the image are equally interesting and the background (i.e. the plate) may be a nuisance factor, especially if polarized light produces some artefacts. Identifying the region with just crystals, avoiding changes in illumination may be difficult and requires some care.
    c. Equalization of the gray level histogram. This step can be performed to enhance the contrast. This is a general step described also in (Meelusarn, 2006) but in this case it has not been applied.
    d. Application of a set of linear and nonlinear filters, to detect peculiar characteristics of the image and to have different version of the same image. Among these: Gaussian filter, Laplacian filter, Laplacian of Gaussian, Sobel filter and horizontal and vertical Prewitt filters. Many of these filters aim at enhancing the contour (high-pass filters as the Sobel filter) or reduce the effect of details and smooth contours (as the Gaussian filter). This is a general step described also in (Meelusarn, 2006) but, in this case, it has not been applied since it has been found that characteristics extracted from filtered images did not add relevant information.
2. Feature extraction, which includes:
    a. Extraction of the first-order statistics, i.e. estimation of statistics that can be inferred from the estimated probability density function of the ROI-selected gray level image. Among these, we have: the mean, the variance, the skewness, the entropy, the level of energy and so on.
    b. Extraction of the second order-statistics, i.e. estimation of the statistics that can be inferred from the Gray Level Co-occurrence Matrix (GLCM). This matrix is defined over an image to be the distribution of co-occurring values at a given offset. In order to estimate the optimal offset and the offset direction, a semi-variogram of the image is estimated. The semi-variogram is used to decide: (i) the step (offset) at which the gray level co-occurrence matrix must be computed. In principle, it is meaningless to compute the GLCM at a too short distance (too much correlation between a sample and its neighbor may exist if the image has been captured with a too high resolution); at

the same time, it is meaningless to compute the GLCM using a too far 'neighbor' (no correlation at all for too distant pixels). (ii) The presence of anisotropy in the image. In case of anisotropy, the GLCM is computed for different directions and the statistics retrieved from the GLCM are then averaged in a single value.

c. Extraction of high-order statistics, i.e. not based on the first or second order image characteristics. The level of fractality in an image is one of these characteristics.

3. Multivariate analysis which includes:

   a. Analysis of Variance (ANOVA). A number of ANOVA tests are carried out, to investigate the various properties of the samples. ANOVA aims at investigating whether difference exists in a dependent variable because of a difference among groups (the so-called factors). The result of the statistical tests is to restrict the successive unsupervised analysis just to few, selected features, the ones relevant to distinguish among groups.

   b. Outlier samples removal. Whether removing outliers before or after feature selection is a debated question. We opted to make this process using all the available features.

   c. Unsupervised clustering. Clustering is carried out by using the selected set of features and evaluating the level of clusterization. Different clustering solutions have been tested according to specific strategies on the kind of factors to investigate on (high quality brands vs. low quality brands, or geographic origin are two possible examples) and adopting a measure of cluster similarity, the Fowlkes-Mallows index (Fowlkes and Mallows, 1983), to evaluate the goodness of the clustering achieved.

The entire above analysis has been implemented by means of in-house MATLAB scripts.

2.5 X-ray Diffraction Analysis

A subset of coffee samples of both toasted and green beans were analyzed by X-ray diffraction. Single green beans were grinded; powders from both toasted and green beans were put in 1 mm capillaries. X-ray powder diffraction measurements were performed by a diffractometer RIGAKU RINT 2500 X'Pert PRO Philips, Eindhoven, Netherlands, operating at 50kV and 200mA, with CuKα radiation (λ=1.5405 Å). X-ray diffraction patterns were collected at room temperature in 2θ/θ geometry, by continuously scanning a 2θ range of 10–70 ° with a step size of 0.02 ° and a scan rate of 0.5 °/min. The capillary containing the sample was continuously rotated during data collection. The diffraction patterns were compared by means of the RootProf package (Caliandro and Belviso, 2014): a background-subtraction step was followed by principal component analysis.

2.6 Multivariate Analysis Methods

ANOVA (Analysis of Variance) is a statistical tool used to understand if meaningful differences (from a statistical point of view) exist between group means (more than two), according to one (or more) dependent variable. If the groups are two, ANOVA coincides with the traditional Student test. A strong assumption done on the dependent variable is that it changes according to a Gaussian random variable within the groups. The null hypothesis of ANOVA is that all the groups have the same mean and no statistical difference between groups exists. The alternative hypothesis states that at least one group has a different mean. ANOVA does not say which differ but, clearly, the result is trivial for two groups as in the case of the status (green/toasted).

ANOVA may be substituted by other tests specifically designed for not-Gaussian distributions. In case of not Gaussian distribution hypothesis, non-parametric tests are preferred as the Kruskal-Wallis test (the test mechanism remains the same). Results of such tests (the p-value) is the probability that the null hypothesis cannot be refused. An extremely low value (possibly, a cut-off significance value of the test can be assigned to decide how much low is) makes the alternative hypothesis extremely likely.

Clustering is an unsupervised analysis, i.e. hidden structures are searched in data with no prior knowledge of possible cause-effects between groups and variates. Although brands have been clearly defined (i.e. they could be considered labels), we do not deem to consider such study a classification problem, which is a supervised analysis, since we have no prior hypothesis on the real possibility to distinguish between brands. In this case, clustering may be a more viable approach to such a problem: we look at possible grouping of samples, inferring some properties of the dataset about closeness of some brands to others and about the discerning of brands according to specific combination of features. For this reason, we carried out the analysis by taking the results of the ANOVA tests, which helped to understand the most promising features to exploit. The number of cluster is fixed by using some prior working hypotheses, in which data are grouped basing on specific investigation criteria. The clustering achieved are finally compared with the ideal grouping using a quantitative index, the Fowlkes-Mallows index (FMI) (Fowlkes and Mallows, 1983). It is defined as:

$$FMI = \frac{TP}{\sqrt{TP+FP}\cdot\sqrt{TP+FN}} \tag{1}$$

being TP, FP, FN the total number of true positives, false positives and false negatives, respectively. FMI ranges within [0, 1], with higher values indicating a greater similarity between the clusters and the benchmark grouping.

3. Results

3.1 Dataset Definition

The initial dataset is composed by 120 images of CuCl2 crystals obtained by the sensitive crystallization protocol described in Section 2.2. The application of the pre-processing and of the features extraction steps allowed the generation of a matrix of 120x24 values (24 features achieved: 7 from the first-order statistical analysis, 14 from the second-order statistical analysis and 3 from the high-order statistical analysis).

An outlier removal procedure has been applied to this dataset.

3.2 Outliers Removal Procedure

Outliers may be removed before feature selection, since the selection may be conditioned by the presence of outliers, or after feature selection, in order to focus the anomaly behavior just on the selected features. The correct procedure is a debating matter in literature (Nguyen and Gopalkrishnan, 2010). We decided to perform outliers removal using all the extracted features, since we deem the feature quite independent among them and not prone to a dimensionality reduction as PCA. Outliers have been identified as the samples which: (i) have feature values significantly different from the average of the remaining samples; (ii) are unique in presenting this difference so, there is not a “cluster” of few samples significantly far from the others.

Because of the different dynamics of each feature, multivariate data are normalized (i.e. for each features mean is removed and then values are divided for their standard deviation). From a mathematical point of view, each sample can be considered a point in a multidimensional space, the space spanning all the features. Firstly, we compute the distance matrix, i.e. the distance of each point from any other point in the space. The Mahalanobis distance has been used, since we considered each sample part of a multivariate Gaussian distribution and so the distance is computed as the number of standard deviation it differs from the mean of the distribution. Let us consider the k-th sample in the multidimensional space $x_k = \left\{x_{k,1}, x_{k,2}, ..., x_{k,N_f}\right\}$, being N_f the number of features (named in Table 2). The distance of x_k from x_n is defined as:

$$d(k,n) = \sqrt{\left(x_k - x_n\right)^T C^{-1}\left(x_k - x_n\right)} \tag{2}$$

with C the sample covariance matrix of all the samples in the dataset.

Said D the matrix of distances (i.e. the matrix in which the *(k,n)*-th element is the distance of x_k from x_n), we performed the analysis sample by sample. For the k-th sample, $D(1:N,k)$ are the distances of x_k from all the remaining samples (*D(k,k)* is zero and is discarded). We compute then:

$$\gamma_k = \frac{\max\limits_{n\neq k}\left\{D(n,k)\right\}}{\frac{1}{N-1}\sum\limits_{\substack{n=1\\ n\neq k}}^{N} D(n,k)} \qquad k = 1,....,N \tag{3}$$

i.e. the ratio between the maximum of such distances divided by the mean. The higher is this ratio, the farther is the sample from all the others. To avoid samples that are far from the mean but forming clusters among them, we finally discard only the samples for which γ_k is the highest. In detail:

$$\gamma_{outlier} > \overline{\gamma} + 3std\left(\gamma\right) \tag{4}$$

After this analysis, 2 samples from the brand Peru have been discarded.

3.3 ANOVA

The main purpose of this study is to understand to what extent it is possible to distinguish between the different brands. In principle, cultivation in different part of the world should generate slight differences in the beans quality (due to latitude, altitude, sun illumination, rain, temperature, and so on), but it is questionable if such differences may be caught by sensitive crystallization.

From a statistical point of view, the dataset can be grouped by two factors: the brand (related to the geographic origin: up to 9 groups) and the status (the manufacturing of the bean: green/toasted = 2 groups). To answer the main question (if we can distinguish between brands), the status is clearly a nuisance factor. Since we are interested only in the different brands, regardless of the status of the bean (green or toasted bean), the first issue to be solved is to understand if differences exist between green beans and toasted beans, in order to discard such factor and consider it a nuisance, as supposed. Statistical analysis is an initial way to verify such assumption. We applied then a statistical test to investigate on possible difference between the two groups, regardless of the brand.

3.3.1 ANOVA Test on Bean Status: Green vs. Toasted Samples

ANOVA (parametric and non-parametric) tests have been performed on samples, grouping them for their status. Results of the p-test, separately for each feature, are reported in Table 2. The features in bold are the ones for which the tests (ANOVA and Kruskal-Wallis) have given a result under the significance level of 2%, refusing the null hypothesis. 11 of 24 features have the ability to distinguish between green vs toasted coffee samples. As it can be seen, a meaningful number of features are able to capture the difference between green and toasted beans crystals. This means that the manufacturing of the bean is an important factor in the brand analysis and should be excluded from successive processing.

Table 2. ANOVA analysis of toasted vs green beans samples (*status* factor)

Feature	**p-value (parametric)**	**p-value (non-parametric)**
mean	0.52	0.34
variance	0.10	0.26
coefficient of variation	0.68	0.57
skewness	0.31	0.0198
kurtosis	**1.7E-05**	**8.4E-05**
level of energy	0.78	0.72
entropy	0.18	0.19
Angular Second Moment	0.78	0.00083
Second order Entropy	**3.3E-05**	**1.7E-05**
Maximum probability	0.37	0.089
Autocorrelation	**1.4E-15**	**7.4E-13**
Diagonal moment	0.34	0.46
Difference energy	**3.0E-09**	**7.2E-09**
Difference entropy	**7.5E-15**	**2.3E-12**
Inertia	**1.6E-13**	**2.5E-14**
Inverse difference moment	**3.0E-05**	**4.4E-06**
Sum energy	0.82	0.79
Sum entropy	0.12	0.12
Sum Variance	0.013	0.047
Cluster shade	**0.0093**	**0.0038**
Cluster prominence	**0.00021**	**0.00058**
logabsFourier-horizontal slope	**1.2E-10**	**1.4E-09**
logabsFourier-vertical slope	**2.4E-07**	**9.9E-06**
correlation dimension	0.0072	0.2

Note. In bold, sensible features and the corresponding p-value (the lower, the better).

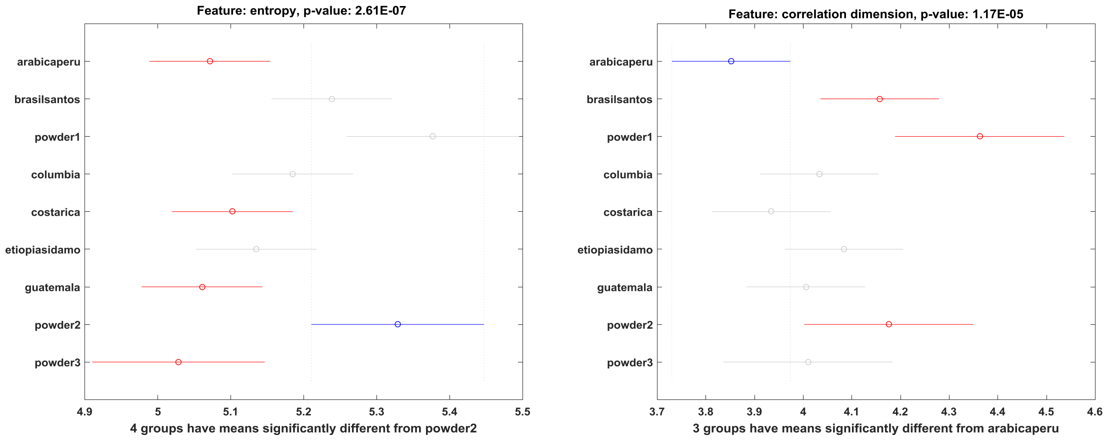
3.3.2 ANOVA Test on Brands

From the previous analysis, we found the toasted samples significantly different from green samples. We have two options to exclude such factor from the brand analysis:

1. Using ANCOVA. ANCOVA (Analysis of Co-variance) is a general linear model, which puts together ANOVA and regression. ANCOVA evaluates whether the population means of a dependent variable are equal across the levels of the factor (the categorical independent variable, in our case, the brand), while the effect of other variables, that are not of primary interest (the covariates or nuisance variables), are controlled.
2. Removing the green (i.e. not roasted) samples from the analysis.

The first kind of analysis requires the covariates to be continuous, in order to regress them out of the data. In our case, ANCOVA cannot be applied, since the status is a binary categorical variable (green/toasted) and regression would be a poor method. We then choose to remove the green samples from the brand processing, resulting in 72 samples (eight samples for each of the nine brands). ANOVA has been repeated, one feature at time, using the brands as factor on the toasted samples only. Multiple comparison post-hoc tests have been then performed.

Two examples of the multiple comparison post-hoc test (Tukey's range test) are illustrated with a graphical plot in Figure 3. The test can be read this way: taken one feature at a time and selecting a single brand (the one in blue), the method automatically provides the brands statistically different from the selected brand, as the ones (in red, if any) for which the segments do not overlap in horizontal (each segment length is a function of the data standard deviation and test significance level). We found that for some features the brands were not easily noticeable among them; for other features (the ones reported in the example figure), many brands were distinguished among them.



(a) (b)

Firure 3. (a) Entropy and (b) Correlation dimension post-hoc Tukey's range tests. In (a): 'Powder2' is compared with all the other brands (resulting in 4 meaningful differences, in red); in (b) 'Peru' is compared with all the other brands (resulting in three meaningful differences, in red).

In principle, the ideal solution would be the one for which all the brands can be distinguished among them (each one from all the others). However, we did not find any feature able to do this, neither for parametric nor for non-parametric tests. We then decided to select a subset of 'best' features, considering them for a successive clustering analysis. These features are the ones for which the number of significant differences is high, let's say above a given threshold. Moreover, since some of the differences between brands are considered especially meaningful, according also to the expected classification (as reported in the species column of Table 2), a supervised inspection of each interactive plot has been done to decide if a feature was worth to be included in the final set. After this analysis, we selected the features reported in Table 3 (6 for the parametric analysis, 5 for the non-parametric analysis).

Table 3. The set of selected features

Parametric analysis		Non-parametric analysis	
Feature	**p-value**	**Feature**	**p-value**
Coefficient of variation	9.3E-09	Coefficient of variation	1.8E-06
Entropy	2.6E-07	---	---
---	---	Angular Second Moment	6.0E-03
Second order entropy	2.3E-06	---	---
Inertia	7.3E-06	---	---
---	---	Sum energy	9.7E-06
Sum entropy	4.9E-08	Sum entropy	7.7E-07
Correlation dimension	1.2E-05	Correlation dimension	1.3E-04

3.4 Clustering Analysis

Clustering has been performed by using the set of features selected in Table 3. Different clusterization tests have been carried out by using the following working hypothesis:

1. A cluster test with N_{clust}=2. In this test, we suppose that green beans may be differently grouped from toasted beans (test code: G-T). The powder samples have been excluded from this test.
2. A cluster test with N_{clust}=2. In this test, we suppose that the beans (only toasted beans are used) may be differently grouped from powder samples, which are supposed to be blends (test code: B-P).
3. A cluster test with N_{clust}=2. In this test, we suppose that the Arabica samples may be differently grouped from the Robusta samples. The blended samples (Powder1, Powder2, Powder3) are left out (test code: A-R).
4. A cluster test with N_{clust}=2. In this test, we suppose that high quality samples, taken from a single plantation, may be differently grouped from low quality samples, where coffee beans belonging to different plantations are mixed. We have elements to suppose that just Powder1 and Powder2 are low quality samples (test code: H-L).
5. A cluster with N_{clust}=3. In this test, we suppose that brands may be grouped by large geographical locations, as described in Table 4 (test code: BRC).

For all the grouping strategies, all the possible combinations of features (taken from Table 3, first or third column) are used and the best combinations are then investigated. The quality of grouping is measured by using the FMI.

In Table 4, a summary of the grouping strategy is reported, for each test and each brand. It is worth remarking that in the A-R test, the Robusta species has been supposed for Brazil and Guatemala samples, although a mixing of beans of both the species were present there. Moreover, in the BRC test, we grouped the brands by using the mean latitude of the country as rationale (i.e. supposing the contribution of the sun illumination to be the main factor to discriminate among species, even coming from different countries).

Table 4. Summary of the applied grouping strategy

Brand name	Supposed species	Grouping strategy				
		G-T, N_{clust}=2	B-P, N_{clust}=2	A-R, N_{clust}=2	H-L, N_{clust}=2	BRC, N_{clust}= 3
Peru	Arabica	Green/toasted	Beans	Arabica	High Quality	South tropic
Brazil	Robusta/Arabica	Green/toasted	Beans	Robusta	High Quality	South tropic
Columbia	Arabica	Green/toasted	Beans	Arabica	High Quality	North tropic
Costa Rica	Arabica	Green/toasted	Beans	Arabica	High Quality	North tropic
Etiopia	Arabica	Green/toasted	Beans	Arabica	High Quality	North tropic
Guatemala	Robusta/Arabica	Green/toasted	Beans	Robusta	High Quality	North tropic
Powder1	Blend	Not used	Powder	Not used	Low Quality	Not used
Powder2	Blend	Not used	Powder	Not used	Low Quality	Not used
Powder3	Blend	Not used	Powder	Not used	High Quality	Not used

Taking the 6 features in the first column of Table 3, all the possible combinations (single features, couple of features, triple and so on) have been used to make clustering. For each specific combination of feature, 100 repetitions of the clustering are performed, to reduce the variability of the results (that is intrinsic in the clustering algorithms). In these experiments, we used the k-medoids clustering algorithm, a more robust to noise and outliers method than the k-means. It uses the samples nearest to the true centroids of the current clusters for the starting and iterating centroid estimation, reducing the effect of randomization intrinsic in k-means. We found that using the feature subset in the third column of Table 3 (the five features achieved after the application of non-parametric ANOVA tests) or using k-means, produced similar (although slightly worse) results.

For each grouping strategy, each specific combination of features and each repetition, the FMI is computed and then averaged over the 100 repetitions. The features combination with the highest average FMI is finally selected.

As an example of FMI calculation we consider the B-P test. The corresponding contingency table is shown in Table 5.

Table 5. Contingency Table achieved for B-P test

Group\Cluster	#1	#2	Total
Beans	39	7	46
Powder	6	18	24
Total	45	25	70

It reports, for each combination, the number of samples within that cluster and belonging to that group. In principle, if clustering were ideal, this table should be diagonal. Assigning the cluster with the highest value to that group (in the example above, cluster #1 can be assigned to Beans tag, cluster #2 to Powder tag), we define these values as True Positives (TP).; Ffor each TP, the other values in the same column are the False Positives (FP, i.e. the values that belonging to the same cluster, but do not belong to the same group) and the other values in the same row are the False Negatives (FN, i.e. the values that belonging to the same group, but have not been recognized in the same cluster).

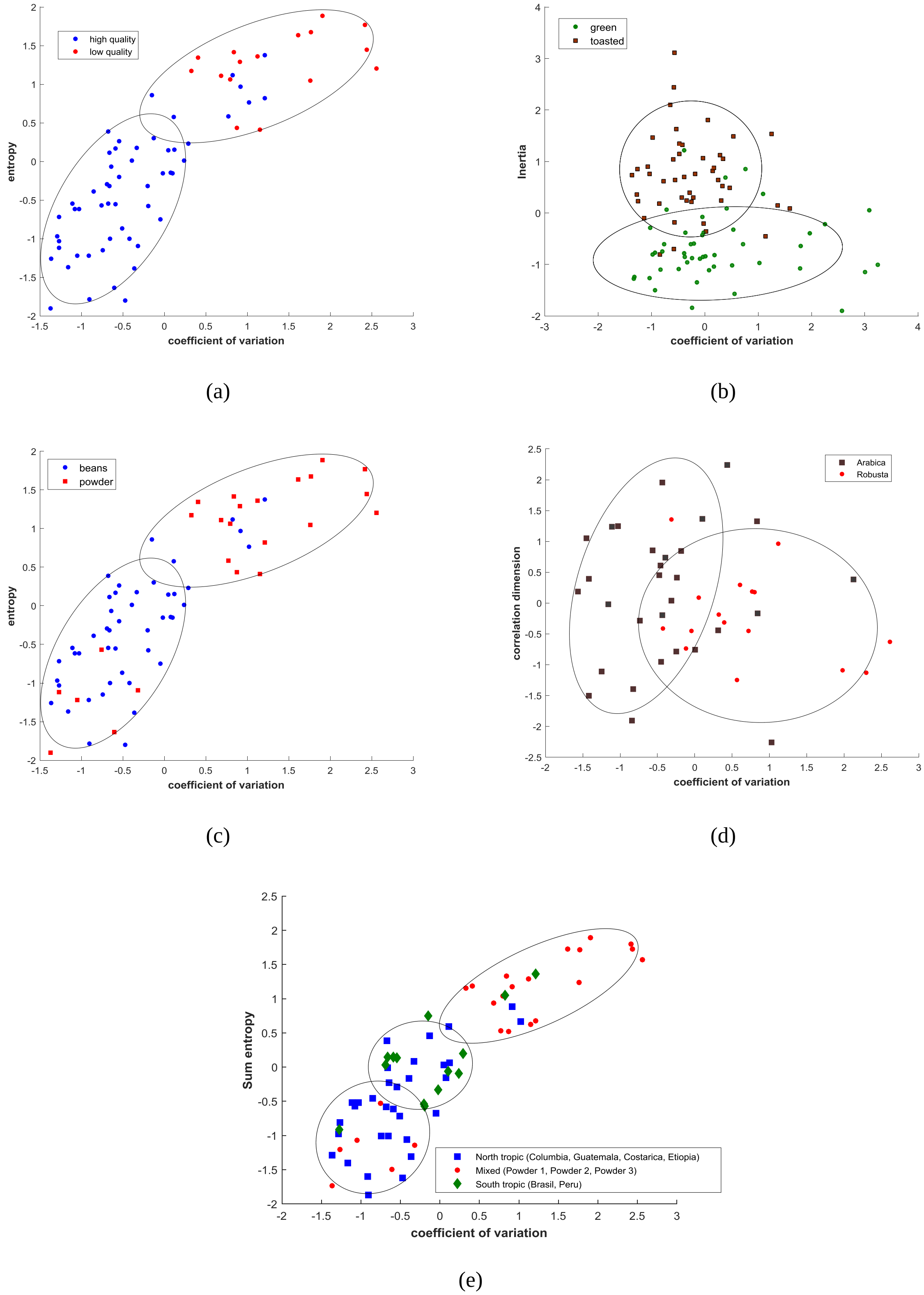
For each grouping strategy, in Table 6 the highest averaged FMI is reported, together with the corresponding combination of features used to achieve that clustering. As it can be seen, often the optimal combination of features includes the Coefficient of variation (the ratio between the standard deviation and the mean of the gray image), then the Sum of Entropy and the other features.

Table 6. Averaged FMI for each grouping strategy and the corresponding best combination of features

Grouping test	**N_{clust}**	**FMI**	**# features**	**Features description**
H-L	2	0.85	4	Entropy, Inertia, Sum Entropy, Correlation dimension
G-T	2	0.79	2	Coefficient of variation and Inertia
B-P	2	0.73	1	Coefficient of variation
A-R	2	0.65	1	Coefficient of variation
BRC	3	0.55	2	Coefficient of variation, Sum of Entropy

For a visual comparison of the clusters, we selected the first sub-optimal solution in which two features are taken. This way, the samples can be reported as dots in a 2D plane and the clusters can be visualized as ellipses (error ellipses at 2σ = 95% is drawn, with the hypothesis of Gaussian dispersion for the intra-cluster samples). In Figure 4 the grouping test cases are described. As it can be seen, the clusterization is able to clear discern high quality from low quality samples (for a further analysis of low quality samples, see also the X-ray diffraction analysis below), as well as green from toasted samples (Figure 4a and 4b, respectively). For the beans/powder samples, instead, we found that the disambiguation is still good, but for few samples, belonging to Powder3, that are within the cluster of beans, even if provided as powder (the red dots in the lower-left part of Figure 4c). Distinction of Arabica and Robusta is still quite good (Figure 4d).

Finally, the grouping by geographical location provides interesting cue. In Figure 4e it seems that the Powder1 and Powder2 (grouped within the Mixed tag) are well distinguished from clusters of defined geographical provenience. Instead, Powder3 (red dots in the lower-left part of the figure) seems to group well with North tropic group. Thus, based on such analysis, we may suppose that Powder3 is of high quality and has a very well defined provenience (North tropic); on the contrary, Powder1 and Powder2 samples seem to spread in the geographical provenience having a well distinct tag.



(a)

(b)

(c)

(d)

(e)

Figure 4. (a) High/Low quality grouping test. Optimal features are coefficient of variation and sum entropy, with an average FMI=0.8102. (b) Green-Toasted grouping test. Optimal features are coefficient of variation and Inertia, with an average FMI=0.7891. (c) Bean-Powder grouping test. Optimal features are coefficient of variation and entropy, with an average FMI=0.715. (d) Arabica-Robusta grouping test. Optimal features are coefficient of variation and correlation dimension, with an average FMI=0.6122. (e) Geographical location grouping test. Optimal features are coefficient of variation and sum of entropy, with an average FMI=0.5449

In Figure 5, finally, we grouped the best FMI for each grouping strategy in a single plot. Two further strategies (above not discussed) have been added: (i) A-R-M, in which we try to make a distinction among Arabica, Robusta and blended samples (the powder samples) and (ii) a geographical location test, in which Powder3 (supposed of high quality compared to Powder1 and Powder2) is a distinct group from mixed tag. As expected, the higher is the level of complexity required to the clustering analysis, the lower the discrimination capability.

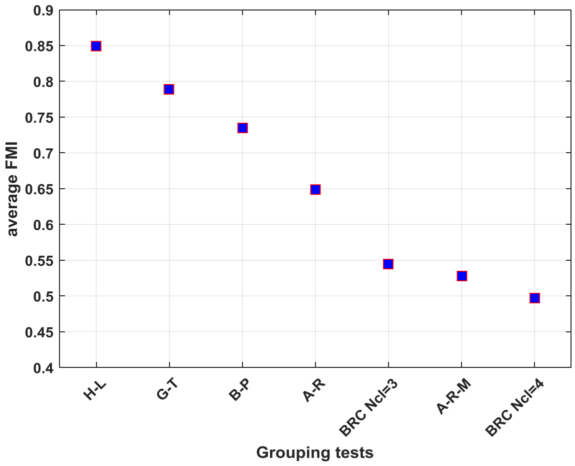


Figure 5. Average FMI for each grouping test discussed in Table 4. For each test, the FMI for the best combination of features is shown

3.4 X-ray Diffraction Analysis

Powder X-ray diffraction has been used as reference technique to complement and validate sensitive crystallization results. A subset of samples has been considered, their diffraction patterns being shown in Figure6a. They all have a huge background for 2θ values between 15° and 30° whose shape varies slightly from sample to sample, which can be ascribed to the amorphous content. Although the general features of the diffraction patterns are preserved in repeated measurements, made in a month distances (Peru patterns in Figure6a), we verified that such variability is not straightly linked to specific sample properties. Few samples, namely Powder1 and Brazil, green bean, have instead very narrow peaks emerging from background, which are due to the presence of minerals (highly crystalline compounds) in the micrometric fraction of the powders. By performing a background-subtraction step, diffraction patterns could be compared solely on the basis of their mineral content. The result of the application of principal component analysis (PCA) on background-subtracted patterns is shown in Figure6b: most of the samples, both from toasted and green beans, belong to the same cluster, while samples Powder1, and to a lesser extent Brazil, green bean, are discriminated.

The analysis of the PCA loadings shows that the first component, which explains 22.8% of the total data variability, discriminates on the presence of the sharp peaks of the Powder1 sample. The second component (10.1% of the total data variability explained) discriminates on the presence of the sharp peaks of the Brazil green sample. Based on its position in the score plot, we can argue that the sharp peaks of the Powder1 sample are different from those of the Brazil green sample.

In summary, mineral elements are present in samples Powder1 and Brazil green, which strongly characterize their X-ray diffraction properties. The mineral in Powder1 is present in large quantities, and it is different from that/those present in the Brazil, green bean sample. Any attempt to understand the nature of these minerals by using known diffraction profiles failed.

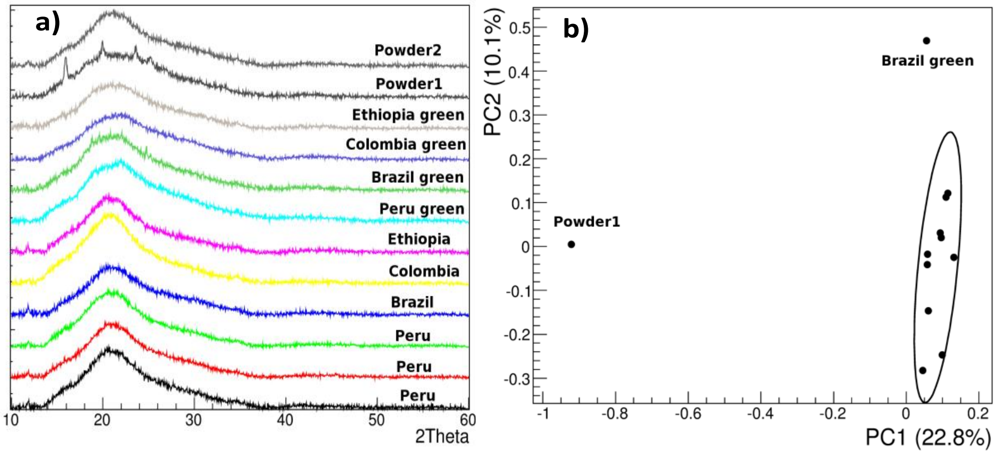


Figure 6. a) X-ray diffraction patterns of a subset of coffee samples and b) score plot based on the first two principal components (PC1 and PC2) obtained by applying principal component analysis to diffraction patterns of (a). Each point on the plot represents a diffraction pattern. The percentage of total variance explained by PC1 and PC2 is reported on the respective axes, and 85% confidence level ellipse is drawn as a result of hierarchical clustering of representative points

4. Discussion

Computational multivariate analysis has recently supported experimental techniques in food analysis, helping more automated objective decisions, especially in discriminating among classes of products: geographic origins (Corvucci, Nobili, Melucci and Grillenzoni, 2015), adulteration (L ópez, Trullols, Callao and Ruis ánchez, 2014) and so on. However, no unified protocol to make analytical processing a systematic tool still exists, since the purpose of the investigation may be various, as well as the nature of the samples. Furthermore, supervised classification (the most common approach) starts from the idea that any single sample may be labeled for sure and that the study aims at investigating the level of within-group variability with respect to the between-group variability only, in order to understand if the proposed analytical protocol may be used to systematically discriminate among the samples. The problem is by far more complicated if we do not have any basic hypothesis on samples regarding their within- / between- group variability, since labels (e.g. due to species or to geographic origin issue) may not be important or may not be detectable with any statistical method. To this aim, unsupervised analysis has been considered suitable to investigate on such problems, since the structure of the data is not prior known.

Sensitive crystallization, supported by unsupervised multivariate analysis, revealed that:

- Powder1 and Powder2 always form a cluster where the other brands don't appear, except rare cases where they cluster with samples of the Arabica species;
- Most of Brazil and Guatemala samples clusterize together and sometimes with samples of the Arabica species. This is an expected result only if we suppose that in the preparation of the powder for the crystallization process, Robusta samples but also Arabica beans have been randomly taken, in spite of the supposed bean species (Brazil and Guatemala were initially supposed to belong to Robusta species).
- The remaining brands, i.e. Peru, Colombia, Costa Rica, Ethiopia often clusterize together, confirming the hypothesis that all of them belong to the same species (Arabica)
- Powder3 always clusterizes with Colombia and Costa Rica samples, addressing a suggestion on the possible origins of the powder.

Crystallographic analysis explains why Powder1 should be separated by the other samples: it is unique in containing micrometric domains of highly crystalline material. Samples Brazil and is also discriminated by this analysis, since it contains minerals of different species than Powder1. However, this only occurs for the green bean sample. It can be argued that some minerals could be present in some green beans in low amounts, which are lost during the toasted process, or reduced in weight fraction by mixing different beans.

It is worth noting that X-ray diffraction is only sensitive to matter in the crystalline form, while sensitive crystallization is influenced by the presence of molecules in whichever form. This could explain why Powder2 is not discriminated by X-ray diffraction.

5. Conclusions

Food frauds cause serious economic damages to agri-food industries as well potential health problems to consumers. The development and implementation of geographical traceability methodologies (or protocols) is a tool that can be used against the food frauds and to define the link between agro-products, such as coffee, wine, saffron, etc. and their terroirs. The identity card (fingerprint) is a discriminating instrument with strong impacts for agri-food companies and terroirs, which is able to develop geographical certification systems and thus, to defend consumers from geographical origin frauds and adulteration risks.

Here, the potentialities of the sensitive crystallization technique to contribute to digital fingerprinting of coffee species are explored. This is the first time this technique has been used for such challenging application, and this was made possible by the use of advanced computational methods, which made objective and quantified the results of the sensitive crystallization experiments.

We demonstrate that this technique is able to discriminate with high confidence level high-quality coffee blends from low quality mixtures, and the Arabica from the Robusta species. The geographical origin is determined with less reliability, and can be only determined by specifying wide macro-regions, like north-tropic and south-tropic. When compared with powder X-ray diffraction, it emerges that this latter is sensitive to the presence of crystalline domains, which give rise to peculiar diffraction profiles. As a matter of fact, X-ray diffraction discriminates the sample Powder1 due to the presence of such domains, which are absent or in trace amounts in the others. This is a common characteristic of many commonly used techniques, which reveal particular features of the sample with very high sensitivity. Sensitive crystallization is instead sensitive to molecular interactions involved in the process of crystal grown and, as such, has a wide spectrum of sensitiveness, i.e. it can reveal a large number of chemical and physical features of the analyzed sample.

It can be envisaged that the use of sensitive crystallization combined with other experimental techniques and coupled with proper computational analysis tools, could further augment the degree of information stored in the digital fingerprint of food, allowing better discrimination, even of geographical nature.

Acknowledgments

We thank Giuseppe Chita for X-ray diffraction measurements, and Valentina Mirabelli for help in implementing the sensitive crystallization protocol. This study was supported by Ente Cassa di Risparmio di Firenze (ECR). The authors acknowledge Dr. Carlo Odello (International Institute of coffee tasters) for providing the coffee samples and precious suggestions. Finally, we are grateful to Dr Brunella Aresta and Giovanni Filograsso for administrative tasks.

References

Alves, M. R., Cunha, S. C., Amaral, J. S., Pereira, J. A., & Oliveira, M. B. (2005). Classification of PDO olive oils on the basis of their sterol composition by multivariate analysis. *Analytica Chimica Acta, 549*(1), 166-178. https://doi.org/10.1016/j.aca.2005.06.033

Andersen, J. O., Henriksen, C. B., Laursen, J., & Nielsen, A. A. (1999). Computerised image analysis of biocrystallograms originating from agricultural products. *Computers and Electronics in Agriculture, 22,* 51-69. https://doi.org/10.1016/S0168-1699(98)00043-X

Andersen, J.-O., Kaack, K. V., Nielsen, M., Thorup-Kristensen, K., & Labouriau, R. (2001). Comparative study between biocrystallization and chemical analyses of carrots (Daucus carota L.) grown organically using different levels of green manures. *Biological Agriculture and Horticulture, 19,* 29-48. https://doi.org/10.1080/01448765.2001.9754907.

Andersen, J-O. (2001). Development and application of the biocrystallization method. PhD.Thesis. Department of Agriculture Science/ Section for Organic Farming.The Royal Veterinary and Agricultural University, 127-139. Copenhagen, Denmark.

Anderson, J. E., Moore, S., Tarczynski, F., & Walker, D. (2001). Determination of the onset of crystallization of N 1-2-(thiazolyl) sulfanilamide (sulfathiazole) by UV-Vis and calorimetry using an automated reaction platform; subsequent characterization of polymorphic forms using dispersive Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 57*(9), 1793-1808. https://doi.org/10.1016/S1386-1425(01)00407-3.

Augagneur, S., Médina, B., Szpunar, J., & Lobiński, R. (1996). Determination of rare earth elements in wine by inductively coupled plasma mass spectrometry using a microconcentric nebulizer. *J. Anal. At. Spectrom., 11,* 713-721. https://doi.org/10.1039/JA9961100713

Bajouba, A., Medina-Rodrígueza, S., Gómez-Romerob, M., Ajalc, E. A., Bagur-Gonzáleza, M. G., Fernández-Gutiérreza, A., & Carrasco-Pancorboa, A. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry, 215,* 245-255. https://doi.org/10.1016/j.foodchem.2016.07.140.

Baumgartner, S., Doesburg, P., Scherr, C., & Andersen, J. O. (2012). Development of a Biocrystallisation Assay for Examining Effects of Homeopathic Preparations Using Cress Seedlings. *Evidence-based Complementary and Alternative Medicine, 40,* 125945. https://doi.org/10.1155/2012/125945.

Beltrán, M., Sánchez-Astudillo, M., Aparicio, R., & García-González, D. L. (2015). Geographical traceability of virgin olive oils from south-western Spain by their multi-elemental composition. *Food Chemistry, 169,* 350-357. https://doi.org/10.1016/j.foodchem.2014.07.104.

Bicho, N. C., Lidon, F. C., Ramalho, J. C., & Leitao, A. E. (2013). Quality assessment of Arabica and Robusta green and roasted coffees - A review. *Emirates Journal of Food and Agriculture, 25*(12), 945-950. https://doi.org/10.9755/ejfa.v25i12.

Cagliani, L. R., Pellegrino, G., Giugno, G., & Consonni, R. (2013). Quantification of Coffea arabica and Coffea canephora var. robusta in roasted and ground coffee blends. *Talanta, 106,* 169-173. https://doi.org/10.1016/j.talanta.2012.12.003.

Caliandro, R., & Belviso, D. B. (2014). RootProf: software for multivariate analysis of unidimensional profiles. *J. Appl. Cryst., 47*(3), 1087-1096. https://doi.org/10.1107/S1600576714005895.

Cocude, M. (1998). Introduction in crystallization workshop. In Cocude, M. (ed.), Crystallisation workshop Paris 06/22/1998.Commission for Scientific and Technical Research on Safety and Health in the Extractive Industries, French Ministry of the Economy, Finance and industry, 1-2. Paris.

Corvucci, F., Nobili, L., Melucci, D., & Grillenzoni, F. V. (2015). The discrimination of honey origin using melissopalynology and Raman spectroscopy techniques coupled with multivariate analysis. *Food Chem, 15*(169), 297-304. https://doi.org/10.1016/j.foodchem.2014.07.122.

Davis, A. P., Govaerts, R., Bridson, D. M., & Stoffelen, P. (2006). An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). *Botanical Journal of the Linnean Society, 152*(4), 465-512. https://doi.org/10.1111/j.1095-8339.2006.00584.x.

Ellis, D. I., Brewster, V. L., Dunn, W. B., Allwood, J. W, Golovanov, A. P., & Goodacrea, R. (2012). Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chem. Soc. Rev, 41,* 5706-5727. https://doi.org/10.1039/c2cs35138b.

Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association, 78*(383), 553. https://doi.org/10.1080/01621459.1983.10478008

Garrett, R., Schmidt, E. M., Pereira, L. F. P., Kitzberger, C. S. G., Scholz, M. B. S., Eberlin, M. N., & Rezende, C. M. (2013). Discrimination of arabica coffee cultivars by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and chemometrics. *Food Science and Technology, 50,* 496-502. https://doi.org/10.1016/j.lwt.2012.08.016.

Geana, E. I., Marinescu, A., Iordache, A. M., Sandru, C., Ionete, R. E., & Bala, C. (2014). Differentiation of Romanian wines on geographical origin and wine variety by elemental composition and phenolic components. *Food Analytical Methods, 7,* 2064-2074. https://doi.org/10.1007/s12161-014-9846-2.

Gonzálvez, A., Armenta, S., & De la Guardia, M. (2011). Geographical traceability of “Arròs de Valencia” rice grain based on mineral element composition. *Food Chem, 126*(3), 1254-1260. https://doi.org/10.1016/j.foodchem.2010.11.032.

Greenough, J. D., Longerich, H. P., & Jackson, S. E. (1996). Trace element concentrations in wines by ICP-MS: evidence for the role of solubility in determining uptake by plants. *Canadian Journal of Applied Spectroscopy, 41*(3), 76-80.

Kleber, W., & Steinike-Hartung, U. (1959). Ein Beitrag zur Kristallisation von Kupfer (II)-chlorid-Dihydrat aus Lösungen. *Zeitschrift für Kristallographie-Crystalline Materials, 111*(1-6), 213-234. https://doi.org/10.1524/zkri.1959.111.1-6.213

Le Gia, V., Teisseron, G., Michel, M. C., & Cauffet, G. (1996). Application of texture analysis for the crystallization image characterisation. In Grossmann et al. (eds). Proceeding from the third European research symposium, pp 1-9. Stockholm, Sweden 1995.

López, M. I., Trullols, E., Callao, M. P., Ruisánchez, I. (2014). Multivariate screening in food adulteration: untargeted versus targeted modelling. *Food Chem, 15*(147), 177-181. https://doi.org/10.1016/j.foodchem.2013.09.139.

Meelursam, A. (2006). Statistical evaluation of texture analysis from the biocrystallization method: Effect of image parameters to differentiate samples from different farming systems, PhD dissertation. University of Kassel, Germany.

Merchaka, N., El Bachaa, E., Khouzamc, R. B., Rizka, T., Akokab, S., & Bejjani, J. (2017). Geoclimatic, morphological, and temporal effects on Lebanese olive oils composition and classification: A 1H NMR metabolomic study. *Food Chemistry, 217,* 379-388. https://doi.org/10.1016/j.foodchem.2016.08.110

Nguyen, H. V., & Gopalkrishnan, V. (2010). Feature extraction for outlier detection in high-dimensional spaces. *Journal of Machine Learning Research, 10*(2), 252-262.

Piva, M-T. (1998). General principles-indication of risk. In Cocude, M. (ed.), Crystallisation workshop Paris.Commission for Scientific and Technical Research on Safety and Health in the Extractive Industries, French Ministry of the Economy, *Finance and industry*, 30-34. Paris.

Piva, M-T., Lumbroso, S., Sieso, V., Monnin, E., Mion, H., Blanc, F., & Mayde Bornier, B. (1994). Cupric chloride crystallization with human blood study of pictures obtained in different pathologies. *Elemente der Naturwissenschaft, 61*(2), 25-39.

Rezzi, S., Axelson, D. E., Héberger, K., Reniero, F., Mariani, C., & Guillou, C. (2005). Classification of olive oils using high throughput flow 1 H NMR fingerprinting with principal component analysis. *Analytica Chimica Acta, 552*(1), 13-24. https://doi.org/10.1016/j.aca.2005.07.057.

Romero, N., Saavedra, J., Tapia, F., Sepùlveda B., & Aparicio, R. (2015). Influence of agroclimatic parameters on phenolic and volatile compounds of Chilean virgin olive oils and characterization based on geographical origin, cultivar and ripening stage. *Journal of the Science of Food and Agriculture, 96*(2), 583-592. https://doi.org/10.1002/jsfa.7127.

Shibata, T., Tanaka, A., Kogure, M., Iguchi, T., & Ogawa, T. (1998). Crystallographic configurations of hydrated cupric chloride crystals grown from aqueous solutions with a small amount of diabetic blood added. In Cocude, M. (ed.), Crystallisation workshop Paris 06/22/1998.

Taylor, V. F., Longerich, H. P., & Greenough, J. D. (2003). Multielement Analysis of Canadian Wines by Inductively Coupled Plasma Mass Spectrometry and Multivariate Statistics. *J. Agric. Food Chem., 51*(4), 856-860. https://doi.org/10.1021/jf025761v.

Vega, F. E., Rosenquist, E., & Collins, W. (2003). Global project needed to tackle coffee crisis. *Nature, 425*(6956), 343-343. https://doi.org/10.1038/425343a.

Vichi, S., Pizzale, L., Conte, L. S., Buxaderas, S., & López-Tamames, E. (2007). The occurrence of volatile and semi-volatile aromatic hydrocarbons in virgin olive oils from north-eastern Italy. *Food Control*, *18*(10), 1204-1210. https://doi.org/10.1016/j.foodcont.2006.07.015.

Wanga, S., Wanga, J., Yub, J., & Wanga, S. (2016). Effect of fatty acids on functional properties of normal wheat and waxy wheat starches: A structural basis. *Food Chemistry, 190,* 285-292. https://doi.org/10.1016/j.foodchem.2015.05.086.

Zhao, H., Zhang, S., & Zhang, Z. (2017). Relationship between multi-element composition in tea leaves and in provenance soils for geographical traceability. *Food Control, 76,* 82-87. https://doi.org/10.1016/j.foodcont.2017.01.006

Appendix A

Preprocessing

The pre-processing includes two steps:

1. Grey level conversion, performed by means of usual RGB to gray conversion routines

2. ROI definition, which is based on representing the characteristics of the inverse Zipf model of each sub-image

by a point in a representation space in which a clustering process is performed (Caron, 2007).

At the end, the ROI is a binary image which approximates the following requirement:

$$ROI(x,y)=\begin{cases}1 & crystal\\ 0 & background\end{cases} \tag{A.1}$$

Appendix B

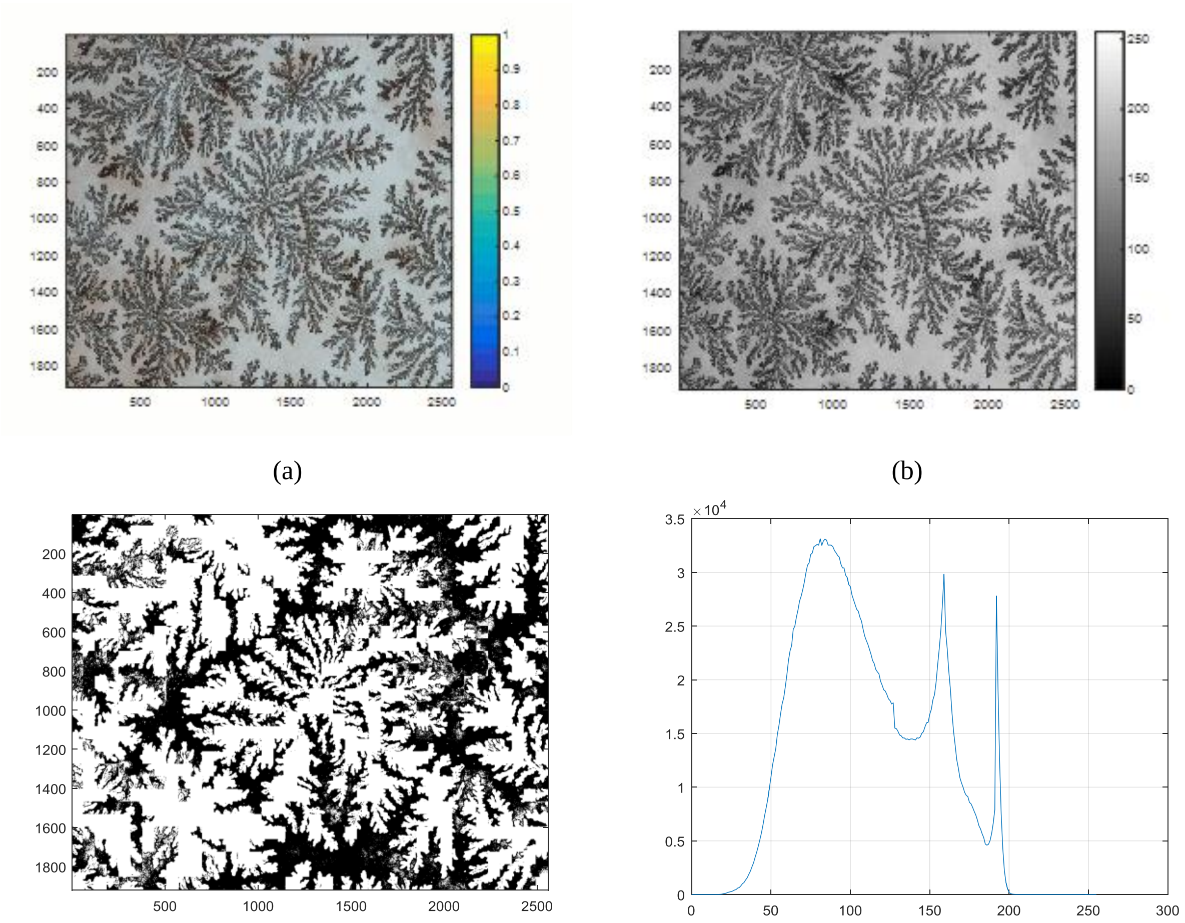
Feature extraction

First order statistics

First order statistics are the ones that can be retrieved from the histogram of the gray levels. Said $I(x,y)$ the digitalized image, after the application of the ROI mask $ROI(x,y)$,

$$I_{ROI}(x,y)=I(x,y)\cdot ROI(x,y) \tag{B.1}$$

(multiplication is pixel-wise), the image is vectorized, i.e. all the values different from 0 are put in a vector, regardless their position in the original image $v(z)=vect\{I_{ROI}(x,y)\}$. Since the image has been converted into a 8-bit gray image, 256 gray levels are expected. A typical histogram of such images in shown in Figure B.1, together with the main step of the pre-processing described above.



(a) (b)

(c) (d)

Figure B.1 (a) The original sample image of crystals. (b) The gray-level corresponding image. (c) The mask applied to select only crystal regions. (d) The 1D histogram achieved after the application of the ROI to the gray-level image

Such histogram can be considered an estimation of the first order probability density function (pdf) $p_V(v)$ of the stochastic process that model the image. The first order statistics extracts all the first order information from such pdf estimation:

- Mean: $\mu_V = \sum_v p_V(v) \cdot v$

- Variance: $\sigma_V^2 = \sum_v p_V(v) \cdot (v - \mu_V)^2$

- Coefficient of variation: $C_V = \frac{\sigma_V}{\mu_V}$

- Skewness: $s_V = \sum_v \frac{p_V(v) \cdot (v - \mu_V)^3}{\sigma_V}$

- Kurtosis: $k_V = \sum_v \frac{p_V(v) \cdot (v - \mu_V)^4}{\sigma_V^4} - 3$

- Level of energy: $E_V = \sum_v p_V^2(v)$

- Entropy: $H_V = -\sum_v p_V(v) \cdot \log p_V(v)$

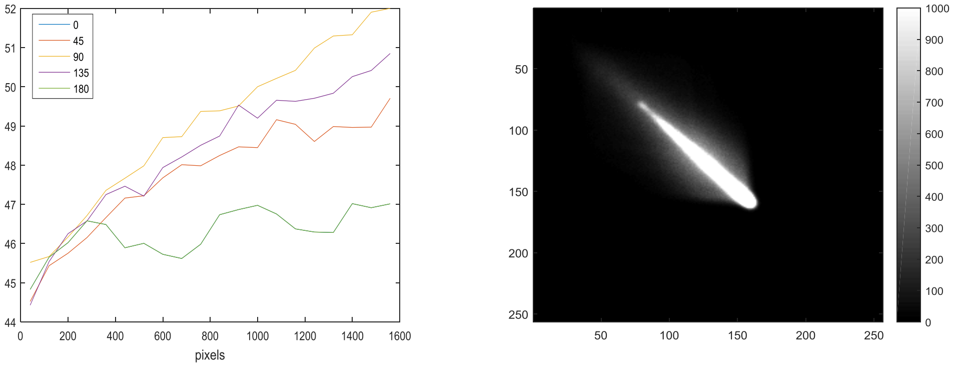
Second order statistics

Second order statistics are the ones that can be retrieved from the Gray Level Co-occurrence Matrix (GLCM). The GLCM is defined as follows:

$$C_{\Delta p,\Delta q}(x,y) = \sum_p \sum_q \begin{cases} 1 \quad if \quad I(p,q) = x \quad and\, I(p+\Delta p, q+\Delta q) = j \\ 0 \quad\quad\quad\quad\quad\quad otherwise \end{cases} \tag{B.2}$$

where x and y are the intensity values of the image I, p and q are the spatial positions in the image and the offset (Δp, Δq) depends on the direction and extent at which the matrix is computed.

GLCM is function of the chosen couple of offset (Δp, Δq). For isotropic images the GLCM is expected to remain unchanged, regardless the direction we take for the offset (Δp, Δq). In principle, since the crystal should not have any preferential direction of growth, an isotropic image is expected. However, to make more robust the estimation of the matrix, the offset can be taken in different directions and the GLCMs (or the computed second order parameters) can be then averaged. Another issue is the extent of the offset, i.e. the amount (Δp, Δq) must take. Too low value of (Δp, Δq), especially for a high-resolution image, should produce a poor estimate of GLCM, since the too high correlation of close pixels would prevent to estimate the real co-occurrence. On the other side, a too high value of (Δp, Δq) would produce opposite effects. The estimation of the anisotropic variogram (which describes the degree of spatial dependence of a spatial random field) may help is the selection of the position at which the decorrelation of pixels becomes important and then may help in the selection of the offset (Δp, Δq). In Figure B.2 an example of the variogram of a biocrystal gray-level image and the corresponding GLCM at direction 0 °are sketched.



(a) (b)

Figure B. 2. (a) Variogram of a biocrystal image after conversion to 8-bit gray (256 levels). Variogram along 0 ° and 180° are superimposed. After the variogram evaluation, a very low value of offset have been taken (Δp, Δq)=(5,5). (b) The corresponding GLCM at 0°.

The GLCM and the corresponding second order statistics are computed at 4 different direction (at 45 °, 135 °, 225 ° and 315 °, so to include shifts (+d,+d), (-d,+d), (-d,-d), (+d,-d)), with d=5 pixels. Second order statistics are then averaged over the different shifts. Formulation of the second order statistics have been given in (Meelusarn, 2006) and shall not be repeated here in details. Such parameters are:

-Angular Second Moment, which values are included between [Levels-2,1] (from Eq. (3.23) of Meelusarn, 2006)

-Second order Entropy, which values are included between [0, -log Levels-2] (from Eq. (3.24) of Meelusarn, 2006)

-Maximum probability, which values are included between [Levels-2,1] (from Eq. (3.25) of Meelusarn, 2006)

-Autocorrelation, which values are included between [-1,1] (from Eq. (3.26) Meelusarn, 2006)

-Diagonal moment, which is a difference in correlation (from Eq. (3.27) Meelusarn, 2006)

-Difference energy, which values are included between [Levels-1,1] (from Eq. (3.29) Meelusarn, 2006)

-Difference entropy, which values are included between [0, -log Levels-1] (from Eq. (3.30) Meelusarn, 2006)

-Inertia, or contrast or variogram (from Eq. (3.31) Meelusarn, 2006)

-Inverse difference moment (also known as local homogeneity) (from Eq. (3.32) Meelusarn, 2006)

-Sum of energy, which values are included between [(2Levels-1)-1,1] (from Eq. (3.35) Meelusarn, 2006)

-Sum entropy, which values are included between [0 log(2Levels-1)] (from Eq. (3.36) Meelusarn, 2006)

-Sum Variance (from Eq. (3.37) Meelusarn, 2006)

-Cluster shade (from Eq. (3.32) Meelusarn, 2006)

-Cluster prominence (from Eq. (3.32) Meelusarn, 2006)

Higher order statistics

High order statistics are the ones that cannot be explained using histogram of gray level or the GLCM. Usually, high order statistics are very difficult to compute, because of the high computational effort. We ground the idea to take higher order statistics from the observation that crystal growth resembles scale-invariant fractal structure. For this reason, we try to catch some properties such as: (i) the slope of the energy spectrum. In case of fractal image, it is well known that this slope is related to the fractal dimension; (ii) the correlation dimension of the image, which is a measure of the dimensionality of the space occupied by the crystal (with respect to the background) and is often called fractal dimension.

Copyrights