Food Control 73 (2017) 164–174

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Discriminant analysis for unveiling the origin of roasted coffee samples: A tool for quality control of coffee related products

Paulo R.A.B. de Toledo [a], Marcelo M.R. de Melo [b], Helena R. Pezza [a], Aline T. Toci [c], Leonardo Pezza [a], Carlos M. Silva [b, *]

[a] Institute of Chemistry, State University of São Paulo – UNESP, 14800-060, Araraquara, SP, Brazil
[b] CICECO – Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Aveiro, 3810-193, Portugal
[c] Latin American Institute of Science of Life and Nature, Federal University of Latin American Integration – UNILA, 85867-970, Foz do Iguaçú, PR, Brazil

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form
1 August 2016
Accepted 1 August 2016
Available online 2 August 2016

Keywords:
Chemical markers
Coffee quality
Discriminant analysis
Geographic origin
Volatiles composition

ABSTRACT

Coffee quality is highly dependent on geographical factors. Based on the chemical characterization of 25 coffee samples from worldwide provenances and same roasting degree, Discriminant Analysis (DA) was employed to develop models that are able to identify the continental or country (Brazil) provenance of blind coffee samples. These models are based on coffee composition, particularly on several key compounds either with or without significant impact on aroma, such as 2,3-butanedione, 2,3-pentanedione, 2-methylbutanal and 2-ethyl-6-methylpyrazine. All models were validated with new and independent data from literature, and also through cross validation and permutation tests. Furthermore, the robustness of the proposed models in case of incomplete characterization data was also tested, being concluded that missing data is supportable by the models. In the whole, this article provides compelling arguments for the development of DA-based tools with the purpose of controlling the quality of coffee in terms of their continental and/or national origins.

1. Introduction

Coffee is one of the most ubiquitous edible products consumed around the world, playing a central economic role in several countries where it is produced and exported. From its ancestral origins in Africa, coffee cultivation wandered east and west, eventually forming a belt roughly bounded by the Tropics of Cancer and Capricorn (Smith AW, 1985). Nowadays, the top ten coffee-producing countries are Brazil, Vietnam, Colombia, Indonesia, Ethiopia, Honduras, India, Uganda, Mexico and Guatemala. For the season of 2014/2015, Brazil was responsible for more than a third of the overall world-scale coffee production, followed by Vietnam with 19.3% share (see Fig. 1). In the whole, a group of more than twenty countries produce coffee on a regular and sizeable basis.

In light of the diversified offer and to the fact that consumers started to value products with label of origin, the confirmation of coffee authenticity through chemical/physical analysis is of great relevance. Stakeholders such as importers or sellers are interested in the development of analytical methods able to demonstrate that the imported coffee had not been adulterated along the commercial chain, or really matches the expected origin and quality specifications. Such challenges represent investigation opportunities within the coffee research field.

Several countries adopted the certification known as Protected Designation of Origin (PDO) in order to protect and control the quality and provenience of their coffee as well as to boost their added value. This certification links the product to the specific culture methods, and operating and atmospheric conditions, as well as to the raw materials. While, for the consumers, PDO products are expected to have distinctive organoleptic features (characteristic of a given provenance), the sensorial spectrum that defines the flavor of coffee may be rather complex and subjective, which complicates a clear confirmation of samples origin. Such difficulties can only be circumvented if reliable and robust analytical based methods are developed for assessing quality parameters of coffee.

The most effective way to keep track of coffee quality and provenience is through the analysis of its volatile composition, which may be directly linked or not to the final aroma experienced by the customer. The definition of quality is thus not a simple task. An

* Corresponding author.
E-mail address: carlos.manuel@ua.pt (C.M. Silva).

http://dx.doi.org/10.1016/j.foodcont.2016.08.001
0956-7135/© 2016 Elsevier Ltd. All rights reserved.

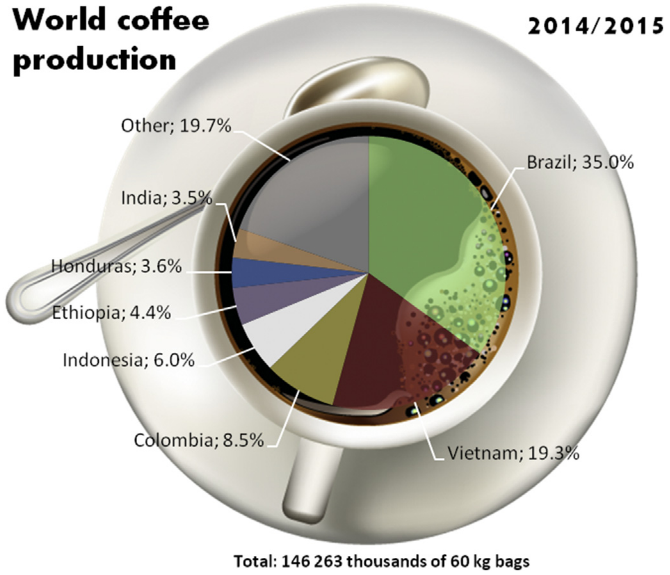

Fig. 1. World coffee production in 2014/2015 [USDA, 2015].

official definition provided by the International Organization for Standardization (ISO) refers that quality should be understood as follows: "*the extent to which a group of intrinsic features (physical, sensorial, behavioral, temporal, ergonomic, functional, etc.) satisfies the requirements, where requirement means need or expectation, which may be explicit, generally implicit, or binding*" (NBR ISO 9001, 2000). Remarkably, roasted coffee is one of the most aromatic food products, and is mainly consumed for the pleasure provided by its volatile components. The concentration of aromatic compounds in roasted coffee can reach 1 g/kg (Flament & Bessière-Thomas, 2002; Toci & Farah, 2014), and their characterization has been extensively studied over the years (Costa Freitas et al., 2001; Rocha, Maeztu, Barros, Cid, & Coimbra, 2004; Mondello et al., 2005; Yener et al., 2014).

The aroma of coffee is intrinsically related to the chemical composition of the beans, which undergo innumerable chemical modifications during roasting, generating a wide variety of volatile compounds. On the other hand, the chemical composition of the beans, and consequently their quality, is directly affected by a wide range of parameters, including species and variety of coffee, climate, soil, bean quality, blend type, post-harvest processing, type of roast, and storage. More than 800 volatile compounds have been identified in roasted coffee so far. These can be divided into different classes, including (in order of abundance) furans, pyrazines, ketones, pyrroles, phenols, hydrocarbons, acids and anhydrides, aldehydes, esters, alcohols, sulfur compounds, and others (Flament & Bessière-Thomas, 2002). Nonetheless, the desirable coffee aroma is produced by a delicate balance in the composition of volatiles, and it is believed that only about 5% of these compounds are actually odorous and capable of impacting coffee flavor (Yeretzian, Jordan, & Lindinger, 2003) (see Table 1). Among these compounds, pyrazines stand out, followed by furans, aldehydes, ketones, phenols, and sulfur compounds, among others (Akiyama et al., 2005; Czerny, Mayer, & Grosch, 1999; Maeztu et al., 2001; Sanz, Czerny, Cid, & Schieberle, 2002).

Arabica coffee is known to have a favorable growth at medium to high altitudes (1000–2100 m) and daily average temperatures around 18–22 °C, typical of equatorial regions. In addition, annual rainfall levels of 1500–2500 mm seem to favor this variety (Illy, 2005). On the other hand, Bertrand et al. (2012) noticed that pluviometric indices ranging from 807 to 1918 mm/year led to higher levels of volatile compounds known to impart negative notes to the coffee, such as 2-ethylhexan-1-ol (heavy, earthy, and slightly floral), and 3-methyl-2-butenoate (overripe fruity/ethereal) (Flament & Bessière-Thomas, 2002). Nevertheless, for coffees grown at high altitudes under annual rainfall levels of 1500–2500 mm, an increase of 5-methyldihydrofuran-2(3H)-one (γ-valerolactone) was noticed, which confers positive sweet/vanilla notes.

By evaluating the effect of temperature, Bertrand et al. (2012) noticed that, in comparison to coffee samples grown at lower temperatures, those cultivated under hot conditions evidenced notable increases of the concentrations of certain alcohols, such as 2-butoxyethanol, 2,3-butanediol and 1,3-butanediol. The last two compounds, which impart earthy and green flavors, have been associated with lower aromatic quality of coffee (Flament and Bessière-Thomas, 2002). In contrast, molecules like 2-methylfuran (caramel/nutty notes), 2-butanone (raspberry ketone/sweet-fruity odor) and methylthiomethane (dimethylsulfide-cabbage, sulfurous) suffer significant concentration reduction as temperature is increased (Flament and Bessière-Thomas, 2002). As a general statement, it is admitted the quality of Arabica coffee can be improved under fresher climatic conditions. In turn, Robusta coffee benefits from a hot and humid climate, lower altitudes (100–1000 m), and an average daily temperature of 22–26 °C, found in tropical regions (Illy, 2005).

The present article proposes the discrimination of coffee samples (mostly *Coffea arabica)* from different countries and continents, setting their volatiles composition as assessment criterion. Upon application of Discriminant Analysis (DA) to data, valid equations for provenance labeling are sought as tools to validate coffee samples origin. By compiling and using a database containing 25 coffee samples, this is the first attempt in the literature to reach such comprehensive classification through DA methods.

The document is structured in the following way: Section 2 is devoted to modeling; Section 3 comprises the database information, namely coffee samples (3.1) and coffee characterization (3.2). Results are presented in Section 4, in the following sequence: preliminary normalization of results (4.1), discrimination of samples according to their geographical origins (4.2), specific differentiation of Brazilian coffee samples (4.3), and geographic and environmental factors and coffee aromas (4.4). Finally, the main conclusions are drawn in Section 5.

2. Modeling

In view of the large amount and variety of volatile compounds found in roasted coffee samples, statistical approaches may be of special usefulness to treat and interpret experimental data. Within this context, multivariate analysis is a powerful tool, since not only considers individual direct impact of the volatile compounds, but also takes into account eventual correlations between them. Two of the most popular methods are Principal Component Analysis (PCA) and Discriminant Analysis (DA).

As far as PCA is concerned, it comprises a technique for shortening the size of a given collection of data without loss of their variance. For this the number of variables is reduced to a minimum, called principal components, that keep the information of the original data set. One particularity of PCA relies on the fact that it is a fully automated method that identifies itself the principal components without human specification of the groups (components) that should be formed. In this sense, PCA is rather suitable for the analysis of multidimensional data where crossed correlation (redundancy) may be present (Jackson, 1991).

Taking into account the data set of this article is not

Table 1
Chemical compounds with sensorial notes found in roasted coffee. Adapted from Toci (2010).

No.	Compound	Sensorial notes	References*
1	2,3-butanedione	Buttery	1, 2, 3, 4
2	2,3-pentanedione	Oily-buttery	3, 4
3	1-octen-3-one	Mushroom	1
4	2-hydroxy-3-methyl-2-ciclopenten-1-one	Sweet/caramel	2
5	Propanal	Roasted/fruity	3, 4
6	2-methylpropanal	Malty/fruity	3, 4
7	3-methylpropanal	Roasted cocoa	3
8	2- e 4-methylbutanal	Buttery	2, 3, 4
9	Hexanal	Butter rancid	3
10	(E)-2-nonenal	Buttery	2
11	Methyonal	Cooked potato	1
12	Methanothiol	Cooked potato	3, 4
13	4-methyl-2-buteno-1-thiol	Smoke/Roasted	2, 4
14	2-methyl-4-furanthiol	Meat	1, 4
15	5-dimethyl-trisulfide	Sulfur	1, 4
16	2-furfurilthiol	Roasted	1, 4
17	2-furanmethanethiol	Smoke/roasted	2
18	2-(methylthiol)propanal	Soy sauce	2
19	2-(methylthio-methyl)furan	Smoke/Roasted	2
20	3,5-dihydro-4(2H)-thiophenone	Smoke/Roasted	2
21	2-acetyl-2-thiazoline	Roasted	1
22	4-methylbutanoic acid	Sweet/acid	1, 2
23	(E)-1-(2,6,6-Trimethyl-1-cyclohexa-1,3-dienyl)but-2-en-1-one (β-damascen)	Cooked apple/Sweet/fruity	1,2, 4
24	3-hydroxy-2,5-dimethyl-4(2H)-furanone (sotolon)(s(so(furaneol)	Caramel/Sweet	1, 2, 4
25	2-ethyl-furaneol	Caramel	1
26	3-hydroxy-4,5-dimethyl-2(5H)-furanone (sotolon)	Spicy	1, 4
27	5-ethyl-4-hydroxy-4-methyl-2(5H)-furanone (abhexon)	Spicy	1
28	2-ethyl-4-hydroxy-5-methyl-4(5H)-furanone	Sweet/Caramel	2
29	2-methoxyphenol (guaiacol)	Phenolic/Roasted	1, 2, 3, 4
30	4-methoxyphenol	Phenolic	1, 2
31	4-ethyl-2-methoxyphenol (4-ethyl-guaiacol)	Phenolic	1, 4
32	4-vinil-2-methoxyphenol (4-vinil-guaiacol)	Cravo	1
33	4-ethenyl-2-methoxyphenol (4-ethenyl-guaiacol)	Phenolic	2
34	3-methylindole	Coconut	1
35	4-hydroxy-3-methoxybenzaldehyde (Vaniline)	Vanilla	1, 4
36	2,3-dimethylpyrazine	Hazelnut/Roasted	2
37	2,5-dimethylpyrazine	Hazelnut/Roasted	2
38	2-ethylpyrazine	Peanuts/Roasted	3
39	2-ethyl-6-methylpyrazine	Peanuts/Roasted	3
40	2,3-diethyl-5-methylpyrazine	Hazelnut/Roasted	1, 2, 4
41	2-ethyl-3,5-dimethylpyrazine	Earth/Hazelnut/Roasted	1, 2, 3, 4
42	3-ethyl-2,5-dimethylpyrazine	Earth	1
43	3-isopropyl-2-methoxypyrazine	Earth	1
44	3-isobutyl-2-methoxypyrazine	Earth	1
45	2-etenyl-3,5-dimethylpyrazine	Earth	1, 4
46	2-etenyl-3-ethyl-5-methylpyrazine	Earth	1, 4
47	6,7-dihydro-5*H*-ciclopentapyrazine	Hazelnut/Roasted	2
48	6,7-dihydro-5-methyl-5*H*-ciclopentapyrazine	Hazelnut/Roasted	2
49	3-mercapto-3-methylbutyl formate	Cat/Green/cassis	1, 2, 4
50	3-mercapto-3-methylbutanol	Hazelnut/Roasted	2

* 1- (Sanz et al., 2002);> 2-(Akiyama et al., 2005); 3-(Maeztu et al., 2001); 4-(Czerny et al., 1999).

multidimensional (only peak area ratios are considered) and that a deliberated grouping of samples is sought (e.g., geographic regions or countries), DA was adopted for the statistical modeling. Likewise PCA this method is able to reduce data redundancy (McLachlan, 2004) through the generation of discriminant functions that sort the distribution of each data point within the chosen grouping (classes). Hence, in the DA approach, the transformation process is said to be human guided and class dependent, which provides models that might not be the absolute best considering other grouping solutions, but are those that best matches the user grouping expectations. However, by allowing human knowledge to be present in the mathematical computations, improved classification performances can be attained through DA.

In the whole, the application of DA in this work aimed at identifying the most relevant compounds, whose concentration variations between coffee samples from different geographic origins are clear enough to be further used to judge/confirm such origins, enabling thus a quality control procedure. Examples of research studies on coffee samples that have employed DA technique were published by Maeztu et al. (2001), Murota (1993), and Powers and Keith (1968). With this purpose, linear equations combining the statistically relevant compounds (factors) were generated from experimental data, as follows:

$$Y = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \beta_3 V_3 + \ldots + \beta_n V_n \tag{1}$$

where Y is the discriminant function, β_0,β_1,…,β_n are the linear discriminant coefficients, and V_1,V_2,...,V_n are the correspondent abundance ratios of the volatiles.

The DA statistical modeling was performed using the software SPSSv.23 and Matlab v.7.8.0. For this a data set comprising chromatograms of 25 coffee samples submitted to Medium roasting degree was employed, which implied extracting information from several publications (see Table 2). Since more than 800 compounds

Table 2
Summary of coffee samples database compiled for the discriminant analysis.

Sample	Origin	Species	Analytical technique	No. of compounds	References
1	El salvador	*C. arabica*	HS-SPME-GC-qMS	73	Mondello et al. (2005)
2	Costa Rica	*C. arabica*	HS-SPME-GC-qMS		
3	Brazil	*C. arabica*	HS-SPME-GC-qMS		
4	Togo	*C. robusta*	HS-SPME-GC-qMS		
5	India	*C. robusta*	HS-SPME-GC-qMS		
6	Vietnam	*C. robusta*	HS-SPME-GC-qMS		
7	Ethiopia	*C. arabica*	HS-SPME-GC/MS	80	Akiyama et al. (2008)
8	Tanzania	*C. arabica*	HS-SPME-GC/MS		
9	Guatemala	*C. arabica*	HS-SPME-GC/MS		
10	Ethiopia	*C. arabica*	HS-SPME-GC/MS	53	Moon and Shibamoto (2009)
11	Nicaragua	*C. arabica*	HS-SPME-GC/MS		
12	Sumatra	*C. arabica*	HS-SPME-GC/MS		
13	Yemen	*C. arabica*	HS-SPME-GC/MS	17	Murota, (1993)
14	Indonesia	*C. arabica*	HS-SPME-GC/MS		
15	Tanzania	*C. arabica*	HS-SPME-GC/MS		
16	Colombia	*C. arabica*	HS-SPME-GC/MS		
17	Brazil	*C. arabica*	HS-SPME-GC/MS		
18	Guatemala	*C. arabica*	HS-SPME-GC/MS		
19	Thailand	*C. arabica*	HS-SPME-GC/MS	62	Cheong et al. (2013)
20	Indonesia	*C. arabica*	HS-SPME-GC/MS		
21	China	*C. arabica*	HS-SPME-GC/MS		
22	Indonesia	*C. arabica*	HS-SPME-GC/MS		
23	Brazil	*C. arabica*	HS-PTR-ToF-MS	73	Yener et al. (2014)
24	Ethiopia	*C. arabica*	HS-PTR-ToF-MS		
25	Guatemala	*C. arabica*	HS-PTR-ToF-MS		

have been identified in coffee matrices (Flament & Bessière-Thomas, 2002), the ideal model would be one that relies on the minimum number of compounds that ensure a correct distinction between samples. Despite the vast range of molecules found in coffee samples, our work was based on the GC-MS characterization reported in six studies from the literature, which comprise only up to 80 compounds. Nonetheless, several molecules already acknowledged and known to impart a decisive influence upon the aroma of the coffee were considered (see Table 1), being thus studied for a different role: geographic markers.

Firstly, a preliminary treatment of data was performed to make possible a comparison between the chromatographic data of different articles. Hence the normalization of the individual peaks through ratios with a compound common to all the studies was attempted aiming at finding the reference molecule that could lead to the best discrimination results and functions. In this effort, the following molecules were tested for the normalization: 2,3-pentanedione, pyridine, 2-methylpyrazine, furfural and 2-methoxyphenol. They were chosen due to being reported by most of the research works considered for the analysis.

By ensuring a robust comparison basis, four different geographic regions were chosen for the DA: Central America, South America, Africa and Asia (see Fig. 2). These embodied all database samples, and the final goal was then the development of reliable discriminant functions for labeling unknown samples within the previous four categories. In addition, a special DA was performed to distinguish the three Brazilian samples, which is justified in view of the fact this country holds the lion's share in terms of world production of this raw material. All models were cross validated by means of rotating coffee samples between the training and validation subsets, in aid of checking the reliability and dependence of the discriminant models to the training datasets.

With reference to the validation of the DA models, the assessment of the prediction capability of the produced models was also checked through the inclusion of new data. Moreover, in order to demonstrate that the proposed models do not lead to good classification performances by pure chance, and that modeling does not suffer from overfitting, permutation tests were carried out following the method described by Westerhuis et al. (2008), being the DA functions permuted 10,000 times for each case. The classification errors were assessed in terms of the number of misclassification (NoM) and through the scoring of individual prediction error measure (Q^2) values, which are defined as follows:

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (2)$$

where y_i is the discriminant function score of the sample i, $\hat{y}_i$ is the predicted value of class membership for sample i, and $\bar{y}$ refers to the mean value of all samples. In summary, the farther Q^2 departs from 1, the worse the class prediction will be in relation to the class label expectations.

3. Compiled database

3.1. Coffee samples

As systematized in Table 2, 25 coffee samples from 15 producing countries across the globe were picked up from the literature in this study. With exception of three coffee samples that are of Robusta species (taken from Mondello et al., 2005), all of them comprise *C. arabica* coffee. The criteria for building the database encompassed three requirements: i) roasted coffee samples with clearly identified country provenances; ii) coffee volatiles sampling through headspace solid-phase microextraction (HS-SPME); iii) analytical results obtained by mass spectrometry.

The coffee samples of Mondello et al. (2005) were provided by Mauro Caffè S.p.A. (Reggio Calabria, Italy) and the roasting process, in all cases, was carried out by the company that supplied the sample under identical conditions.

The *C. arabica* samples of Akiyama et al. (2008) were provided by Unicafe Inc. (Tokyo, Japan), roasted to Medium degree (L value 26) using a Probat G-12 roaster. The mentioned degree was represented as an L value, which was determined by measuring ground-roasted coffee (particle size < 500 μm) using a colorimeter ZE-2000 (Nippon Denshoku Industries Co. Ltd., Tokyo, Japan).

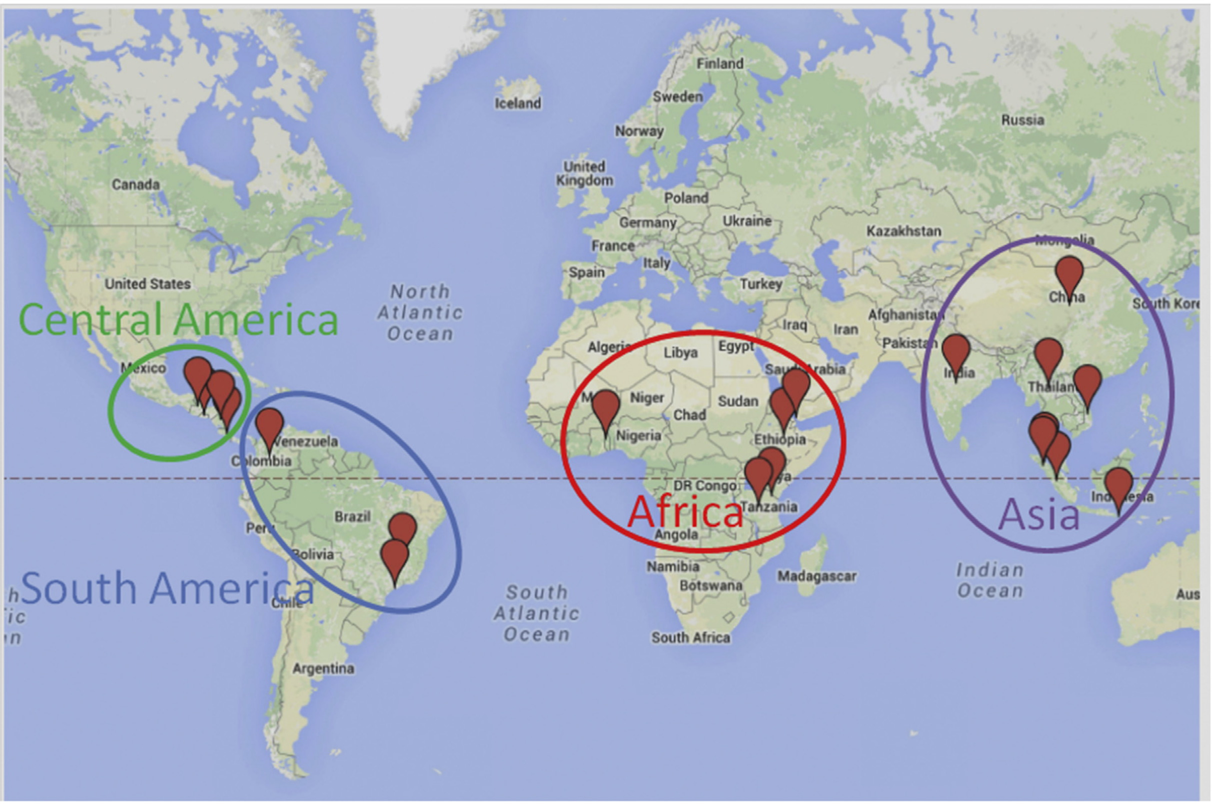

Fig. 2. Geographic regions (colored circles) used for grouping coffee samples data (red icons) in the Discriminant Analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The *C. arabica* samples studied by Moon and Shibamoto (2009) were roasted in a Gene Cafe coffee bean roaster (Fresh Beans Inc., Park City, UT) under different conditions. Only results from Medium roasting (240 °C for 14 min) were considered for this work. After roasting, the coffee beans were grounded with a Starbucks Barista coffee grinder (Seattle, WA).

Murota et al. (1993) analyzed six cultivars from *C. arabica*, which were roasted for 10–15 min under temperatures of 160–200 °C. Afterwards the samples were milled with a steel cut grinder (18–22 mesh).

Cheong et al. (2013) took different *C. arabica* samples from local villages of Thailand, Indonesia and China, roasted them in a home coffee roaster (Imex, Korea) for 12 min, and then cooled down by blowing in the roaster for 4 min. Grinding took then place in a coffee grinder (Braun KMM30, Germany).

Finally, three *C. arabica* samples were investigated by Yener et al. (2014), whose origins were Brazil, Ethiopia and Guatemala, commercialized by Illycaffè S.p.A, (Trieste, Italy). They were already in course powder and medially roasted.

3.2. Chemical characterization

Within the six articles from which experimental data were compiled, three analytical methods were employed for samples characterization: Headspace solid-phase microextraction and gas chromatography combined with quadrupole-mass spectrometry (HS-SPME-GC-qMS) (Mondello et al., 2005), Headspace solid-phase microextraction and gas chromatography combined with mass spectrometry (HS-SPME-GC-MS) (Murota et al., 1993, Akiyama et al., 2008; Moon & Shibamoto, 2009; Cheong et al., 2013), and proton-transfer-reaction time-of-flight mass spectrometry (PTR-ToF-MS) (Yener et al., 2014). Specific details such as material and diameter of SPME fibers, chromatographic conditions and others may be consulted in the original publications.

4. Results and discussion

4.1. Preliminary normalization of results

As referred in Section 2, a preliminary data treatment was required in order to ensure a comparison basis between the chromatogram peaks of different authors. To fully understand the importance of this procedure, it should be mentioned that while most of the compiled articles reports the quantification of coffee volatiles in terms of peak areas (Akiyama et al., 2008; Mondello et al., 2005; Moon & Shibamoto, 2009), others rely on mean content (Murota, 1993), concentration in ppm (Cheong et al., 2013), and concentration in parts per billion by volume (ppbv) (Yener et al., 2014). Hence the normalization of individual peaks through ratios with a common compound to all the chosen studies was attempted, aiming at finding the reference molecule that leads to the best discrimination results and functions.

Two criteria were adopted to judge the performance of each reference compound under analysis: the % of variance explained by the main two resulting discriminant functions, and the % of correct classification of the cases. In Table 3 the performance of the five compounds tested for subsequent normalization is summarized, these being: 2,3-pentanedione, pyridine, 2-methylpyrazine, furfural and 2-methoxyphenol.

From Table 3 it can be noticed that 2,3-pentanedione and furfural are the worst choices for database normalization, since not only their respective two functions only explain 85.2% and 86.8% of

Table 3
Compound suitability results for usage as reference for database normalization. Pyridine leads to the best results for the DA of geographic origins.

Reference compound	% of variance explained by discriminant functions			Correct classification
	Y_1	Y_2	Y_1+Y_2	
furfural	63.1%	22.1%	85.2%	68.4%
2,3-pentanedione	49.2%	37.6%	86.8%	76.5%
2-methylpyrazine	76.6%	13.5%	90.1%	90.9%
2-methoxyphenol	53.8%	34.8%	88.6%	93.5%
pyridine	82.7%	14.6%	97.3%	100.0%

the data variance, but also they are prone to greater classification mistakes (correct classification scores of 68.4% and 76.5%). In terms of medium ranked performances one can find 2-methoxyphenol which scored 93.5% in terms of correct classification despite a variance explanation of only 88.6%.

In the whole, 2-methylpyrazine and pyridine were the ones with most effective performances for usage as reference, particularly pyridine, that allowed a 100.0% correct classification. Remarkably, its two discriminant functions are able to explain the 97.3% of variance of the compiled data set. For this reason, pyridine was established as reference component to normalize all experimental results.

4.2. Discrimination of samples according to their geographical origins

Having chosen pyridine as reference compound for normalization, the DA was then carried out regarding the four continental regions of Fig. 2. The DA involved 80 compounds with the objective to identify the best ones in terms of differentiation of samples provenances. This approach took into account that some publications of Table 2 only report 17 compounds, which constrains the model in relation to others that have more molecules reported. As a result, three discriminant functions relying on a total of 18 compounds present in the coffee sample were adjusted to database (see Table 4), explaining 100% of the variance. These are identified in Table 5 as well as the coefficients (β_i) for each discriminant function: Y_1, Y_2 and Y_3.

The discrimination of coffee samples according to their geographical origins is graphed in Fig. 3 using only Y_1 and Y_2 functions, which together explain 97.3% (82.7% + 14.6%) of the data variance (see Table 4). For each geographical group considered, the centroids are plotted with filled marks, while the data points are graphed in their individual coordinates (non shaded symbols) but connected with a line to the respective group centroid. To emphasize the areas belonging to the different groups, colored circles were drawn around each geographic group.

As far as the differentiation of geographic groups is concerned, Asia clearly detaches from the remaining ones, while Africa, Central America, and South America share close regions in the graph, albeit without overlapping in the bidimensional representation. Nonetheless, while Y_1 function is enough to distinguish Asian coffee samples, the combination of Y_1 and Y_2 is necessary to clearly differentiate the two American classes and Africa.

The dispersion of data points within a given group is a key parameter for assessing the quality of the differentiation achieved through DA. In this respect, attention should be paid to the Central America category, where the highest relative dispersion between samples was observed (see Fig. 3), being the highest distance to the centroid provided by the sample from El Salvador. Such internal deviations would be less expectable in this group than in larger regions like Africa or Asia, which in turn exhibit a moderate dispersion of discriminant scores within each group. The dispersion of Central America approximates this group to Africa (above) and South America (below) classes. In this sense, from a DA understanding, coffee from that origin seems to possess intermediate compositions of the discriminant compounds in relation to the other two classes.

With regard to the dispersion of scores within Africa group (Fig. 3), it is worth noting the two samples with greater deviation to the centroid belong to Togo and Yemen. With relation to Togo coffee, two features should be underlined: it comprises the only sample that belongs to the Atlantic coast of Africa, and also it is one of Robusta coffee that was intentionally used in this study to verify if different coffee species could significantly disturb the model. With reference to the dispersion associated to the Yemen sample, it should be noted that this coffee was included in Africa category, while, rigorously, it belongs to Middle-East. Such inclusion is due to a lack of samples from identical origins that would justify an independent Middle-East category.

The dispersion observed inside Asia group (see Fig. 3) seems to be interlinked with the physical distance between countries like China, Indonesia, India and Thailand, which imply differences on environmental conditions between the coffees. On the other hand, despite two Robusta samples were present in this category, no evidence of clear outcasting from the other samples was noticed. Hence, the eventual dispersion caused by the coffee species seems to be absorbed by the dispersion directly due to geographic factors.

With reference to South America, the dispersion observed in Fig. 3 is caused by Brazilian samples rather than by Colombian coffee. While more data points are needed to confirm this observation, the available results suggest that samples from Brazil are of particular complexity within that group.

Table 6 summarizes the performance of the discriminant models regarding the classification of samples from the training database. Accordingly, an entirely correct classification was attained without cross validation (using all coffee samples), which may be attributed to the suitability of the compounds of the database to be used as discriminant factors, but also to the preliminary normalization that allowed the selection of a reliable standard molecule. Nonetheless, the application of a cross validation (CV) exposed some misclassifications involving Central America, Africa and Asia samples. On those classes the success of the classifications was reduced to 50.0% (Central America), 43.0% (Africa), and 62.5% (Asia) after cross validation. In the case of South America, 100% of success was attained.

In order to further check the reliability of the DA performed above, new data were collected from literature for accomplishing validation tests beyond the original database. Table 7 lists the samples used for validation, where it can be promptly noticed that

Table 4
Performance of the obtained discriminant functions for the “Four geographical regions”. Functions Y1+Y2 are sufficient to explain almost all variance (97.3%) of database.

Function	Variance explanation (%)	Cumulative variance explanation (%)	Canonical correlation
Y_1	82.7	82.7	0.990
Y_2	14.6	97.3	0.948
Y_3	2.7	100	0.787

Table 5
Non-standardized coefficients of the proposed discriminant functions for each of the three discrimination studies.

Compounds (V_i)	Function coefficients (β_i)				
	"Four geographical regions"			"Brazil *vs.* Others"	"Brazil *vs.* America"
	Y_1	Y_2	Y_3	Y_4	Y_5
(Constant)	−15.169	−1.024	−8.256	−3.429	−21.115
2-methylbutanal	−10.985	1.720	−3.751	−1.355	5.057
3-methylbutanal	−4.019	−18.606	0.230	6.952	13.738
2,3-butanedione	17.054	18.480	7.733	0.749	−28.552
2,3-pentanedione	−6.305	−0.294	−1.002	−0.200	2.368
2-methyl-1 H-pyrrole	10.233	−3.429	4.809	6.357	−31.846
pyrazine	75.998	45.763	54.104	11.083	489.508
2-methylpyrazine	−3.867	−2.519	7.057	2.881	−5.038
2,5-dimethylpyrazine	76.283	6.368	11.547	−2.179	22.431
2,6-dimethylpyrazine	−19.994	−17.754	−17.163	−1.709	–
2-ethylpyrazine	−64.957	−0.781	16.457	7.192	–
2,3-dimethylpyrazine	−10.179	−4.642	2.662	−6.017	–
1-hydroxy-2-butanone	12.534	20.875	6.569	7.400	–
2-ethyl-6-methylpyrazine	−1.438	−1.666	−44.529	−1.914	–
2-ethyl-3-methylpyrazine	−12.705	19.236	7.666	8.803	–
furfural	28.312	22.991	−1.534	4.266	–
5-methylfurfural	−22.335	−13.563	4.400	−4.208	–
2-furfuryl-5-methylsulfide	199.682	44.156	45.226	–	–
pyrrole-2-carboxaldehyde	−7.785	−7.261	−6.501	–	–
2-methoxyphenol	–	–	–	−10.069	–
2-acetylpyrrole	–	–	–	4.193	–

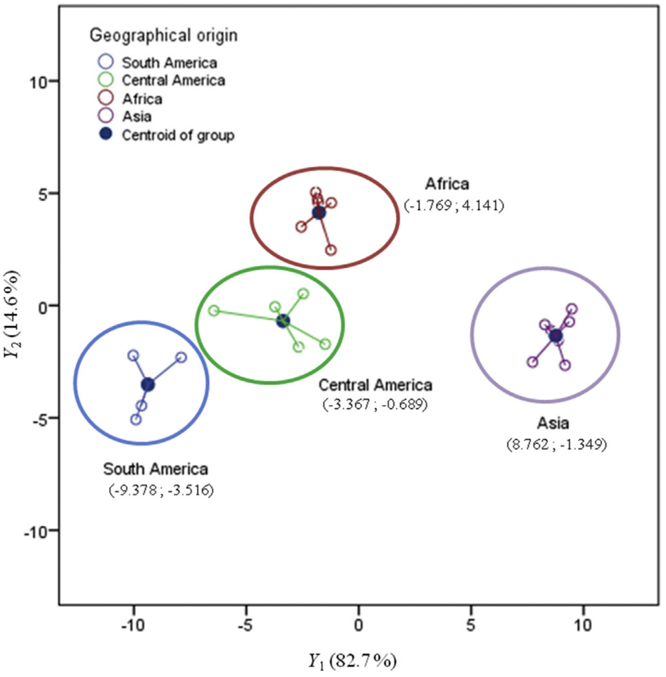

Fig. 3. Discriminant scores and centroid values of DA for the four geographic origins considered in this work. It is clear that Y_1 is enough to differentiate coffee samples from Asia groups, while Y1 and Y2 are needed to discriminate the remaining three regions.

samples from the four continental regions are presented for the validations tests. From a computation perspective, the procedure comprised the compilation of the chemical composition of the samples, normalizing by pyridine peak, and then applying the resulting coefficients of each discriminant function to the respective ratios, so that the final scores were obtained. The same table contains the scores of the validation tests for four geographical regions. In this respect, the criterion for the final classification is the distance between the discriminant scores of a given sample and the group centroid of each geographic region considered. The decision of the classification is based on the smallest length to the group centroid. For instance, in the case of the Vietnamese sample taken from Akiyama et al. (2008), the discriminant scores (Y_1 and Y_2) are such that it is 6.38 units far from Asia centroid, while 7.83, 10.23 and 12.78 units far from Central America, Africa, and South America centroids, respectively. Hence, the discriminant model states that this sample belongs to the Asia group, which is the exact classification. The validation tests were able to properly classify three of the four coffee samples used for this purpose. In fact, the sample from Colombia reported by Rocha et al. (2004) led to a dubious classification involving South America and Africa groups, since the distance to these centroids was 10.06 and 10.23, respectively. Such proximity of distances unables an undisputable declaration of sample origin, although allowing to discard correctly two other provenances.

A noteworthy aspect on the classification tests comprises the fact that not all the compounds needed for the discriminant functions are reported in the works used for the validation tests. This fact is responsible for samples where the distance to the centroid is particularly high (e.g. 20 units) such in the case of the dubious samples from Colombia. Nevertheless the results from Table 7 evidence that the discriminant models tend to maintain their correct classification capability despite eventual lack of information for key compounds.

Finally, a permutation test was applied to Y_1 in light of being the function that explains the majority of the variance of the data in the classification study (82.7%; Table 4). Fig. 4 presents the outcomes of this permutation test, whose results are represented by two parameters: the Number of Misclassifications (NoM: plot A1 in Fig. 4) and through the individual prediction error measure (Q^2: plot A2 in Fig. 4). The results are represented in terms of frequency histograms. The permutation results revealed that Y_1 scored out of the densest regions scored by the randomly built functions (see plots A1 and A2 in Fig. 4), up to the point of standing beyond the statistical boundaries (95% confidence level) where statistical significance is ensured. In the case of Q^2 this is achieved by a score of Y_1 that is significantly higher than the average value obtained by random fitting. With respect to NoM, Y_1 leads to very few

Table 6
Overall classification performance (count and %) of the DA model, with and without cross validation (CV), for the four continents differentiation study.

Original class	Predicted class (DA Model)							
	South America		Central America		Africa		Asia	
	Without CV	With CV	Without CV	With CV	Without CV	With CV	Without CV	With CV
South America	4 (100%)	4 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Central America	0 (0%)	1 (16.7%)	6 (100%)	3 (50.0%)	0 (0%)	1 (16.7%)	0 (0%)	1 (16.6%)
Africa	0 (0%)	2 (28.5%)	0 (0%)	2 (28.5%)	7 (100%)	3 (43.0%)	0 (0%)	0 (0%)
Asia	0 (0%)	0 (0%)	0 (0%)	2 (25.0%)	0 (0%)	1 (12.5%)	8 (100%)	5 (62.5%)

Table 7
Validation tests on the “Four geographical regions” (VT1-4), “Brazil *vs.* Others” (VT5-6) and “Brazil *vs.* America” (VT7-8) studies. (VTi = acronym of Validation Test number i). Successful classifications were obtained for all samples out of the original database. Values in bold correspond to the shortest distances to the centroids.

Test	Reference	Origin	Species	Distance to centroid				Classification
				South America	Central America	Africa	Asia	
VT1	Akiyama et al. (2008)	Vietnam	*C. robusta*	12.78	7.83	10.23	**6.38**	Asia
VT2	Akiyama et al. (2008)	Indonesia	*C. robusta*	16.62	13.47	16.61	**9.92**	Asia
VT3	Rocha et al. (2004)	Colombia	*C. arabica*	**10.06**	10.90	**10.23**	21.90	South America or Africa
VT4	Amstalden, Leite, and Menezes (2001)	Brazil	*C. arabica*	**4.03**	9.80	11.42	22.65	South America

Test	Reference	Origin	Species	Distance to centroid		Classification	Distance to centroid		Classification
				“Brazil vs. Others”			“Brazil vs. America”		
VT5	Akiyama et al. (2008)	Vietnam	*C. robusta*	6.21	**3.87**	Others	–	–	–
VT6	Akiyama et al. (2008)	Indonesia	*C. robusta*	2.78	**0.44**	Others	–	–	–
VT7	Rocha et al. (2004)	Colombia	*C. arabica*	3.88	**1.54**	Others	10.61	**4.21**	America
VT8	Amstalden et al. (2001)	Brazil	*C. arabica*	**0.39**	1.15	Brazil	**2.53**	12.29	Brazil

misclassifications and clearly remains on the left of the statistical threshold limit for p-value = 0.05. In the whole, the permutation test results demonstrate that the proposed discriminant function did not achieve a good classification performance at the expenses of pure chance and/or overfitting problems, which supports the validity of the developed model for the four regions differentiation.

4.3. Specific differentiation of Brazilian coffee samples

In light of its lion's share importance in the context of world coffee production (35%, see Fig. 1), a DA was specifically developed to discriminate Brazilian coffee samples from all the other origins listed in the database, as well as from continental “neighbors” from America. Accordingly our database was reprocessed again but for a grouping of just two categories: “Brazil” and “Others”. The results showed that only one discriminant function was needed to differentiate Brazilian coffee samples from samples with different proveniences. The resulting single function is able to explain 100% of the data and rely on 18 compounds, which are listed in Table 5. In comparison to the discriminant factors of Section 4.2, the differentiation “Brazil *vs.* Others” relies on the same compounds with exception of 2-methoxyphenol and 2-acetylpyrrole that were replaced by two new molecules: 2-ethyl-3-methylpyrazine and pyrrole-2-carboxaldehyde. Hence, the reformulation of the model to specifically distinguish Brazilian coffee samples involved in a greater extent a refitting of the previous model coefficients than a substantial replacement of discriminant compounds.

With regard to the discrimination of Brazilian coffee samples from those of the same continent, independently of belonging to Central America or South America group regions of Section 4.2, discriminant factors of the “Brazil *vs.* America” differentiation are also listed in Table 5. It was shown that one function was also enough to discriminate samples. Remarkably, the resulting discriminant model for this differentiation only needs 8 compounds: 2-methylbutanal, 3-methylbutanal, 2,3-pentanedione, 2,3-butanedione, methyl-1H-pyrrole, pyrazine, 2-methylpyrazine and 2,5-dimethylpyrazine. In fact, since fewer samples were used for this classification, having a discriminant function that depends on less than half of the chemical markers is in agreement with our expectations.

Fig. 5 illustrates the discriminant scores for both analyses, where it is clear the good classification performance of the models. Taking into account the more complex challenge of discriminating Brazilian coffee samples from a group of samples that include a vast source of origins, the gap between discriminant groups (Brazil and Others) becomes narrower, a feature evidenced by the absolute distance between group centroids: |-2.062−0.281| = 2.343 units in “Brazil *vs.* Others”, against 8.079 units in “Brazil *vs.* America”.

Likewise done in Section 4.2, the overall classification performance of the Brazilian samples models were assessed with and without cross validation. These results are presented in Tables 8 and 9, where it can be noticed that by performing a cross validation the correct assignment of the data is reduced to 62% in the case of the “Brazil *vs.* Others” model, and to 71% in the case of the “Brazil *vs.* America”.

With reference to the two differentiation studies for the Brazilian samples, Table 7 provides the validation tests results, where perfect classification performances were attained. The functions were able to correctly attribute the origin of the coffee sample from Colombia to the “Others”, and “America” groups rather than to “Brazil”.

Following the same approach of the four regions discrimination study (Section 4.2), permutation tests were carried out for functions Y_4 and Y_5, being the results presented in the plots B1 and B2, and C1 and C2, of Fig. 4, respectively. Both Y_4 and Y_5 scored out of the densest regions of randomly built functions, exhibiting thus statistical significance: regarding NoM, both Y_4 and Y_5 clearly lie on the left of the boundary p = 0.05; in the case of Q^2, they are

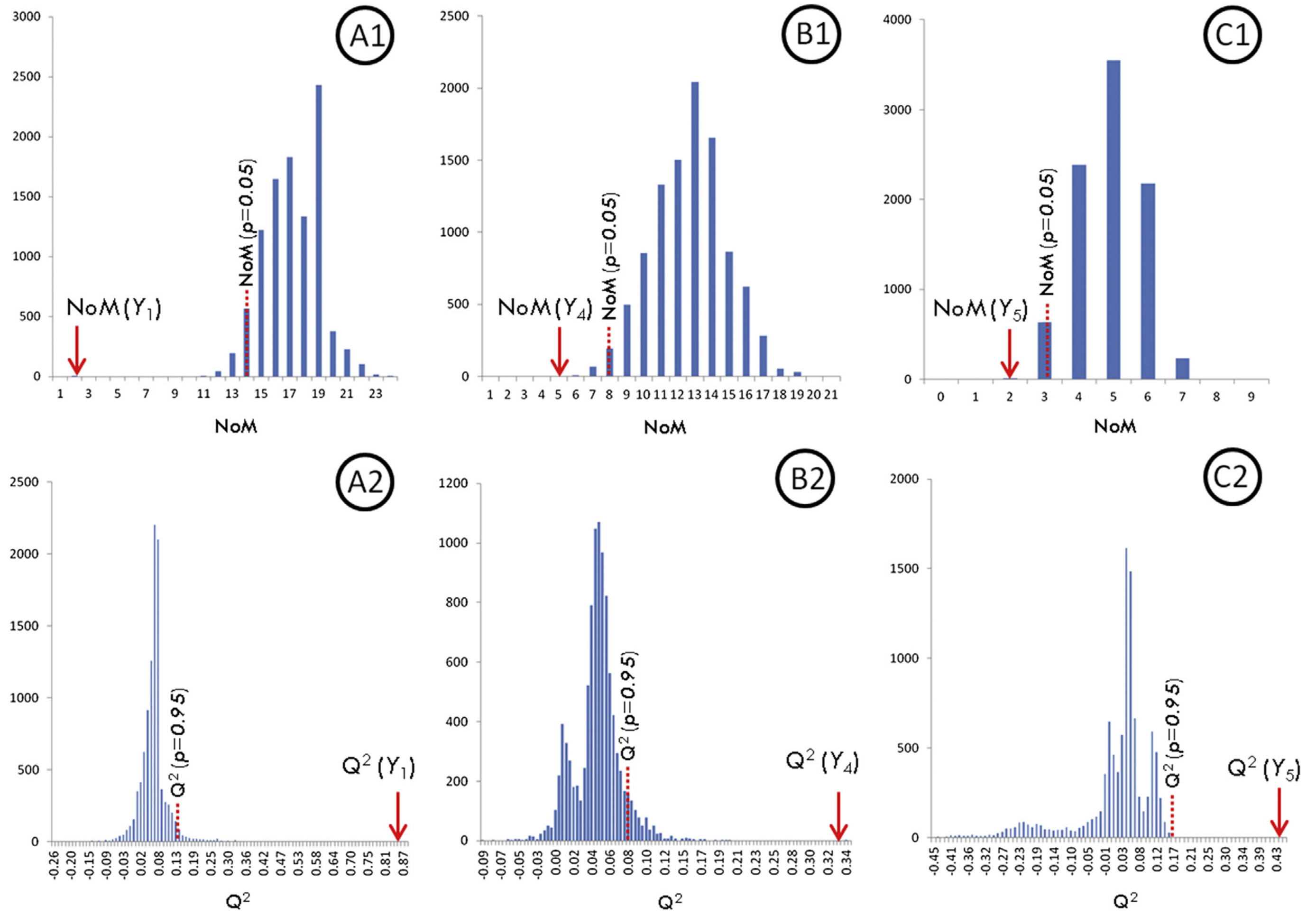

Fig. 4. Permutation tests results of a total of 10000 permutations of the data: A, B and C refer to four geographical regions, “Brazil *vs.* Others”, and “Brazil *vs.* America” discriminations. The upper plots (A1, B1, and C1) refer to the Number of Misclassifications (NoM), and the lower ones (A2, B2, C2) comprise the individual prediction error measures (Q^2). It is evident that the three DA functions developed in this work scored outside the 95% confidence bounds, and, consequently, are significant.

expressively located on the right of the reference limit $p = 0.95$. Hence, identical conclusions may be drawn from the attained results: the functions composing the DA models of the specific differentiation of Brazilian coffee samples are valid, as they were not produced by chance or taking advantage of overfitting.

Globally, our results show that Brazilian coffee samples possess sufficiently distinct compositions, which makes possible to distinguish them solely based on a small group of key compounds using the proposed discriminant functions.

4.4. Geographic and environmental factors and coffee aromas

Up to now, several studies have demonstrated that the contents of various classes of volatile compounds of coffee are influenced by geographical and climatic factors, including altitude, longitude, latitude, daily temperature fluctuations, precipitation, and solar radiation. These factors have been found to affect the levels of odorant compounds including 2-furfurylthiol, 2,3-butanedione, 2,3-pentanedione, 2-methylbutanal, 2,3-dimethylpyrazine, 2-ethyl-6-methylpyrazine, and guaiacol (Bertrand et al., 2012; Cheong et al., 2013; Freitas & Mosca, 1999; Costa Freitas et al., 2001; Mondello et al., 2005; Risticevic, Carasek, & Pawliszyn, 2008; Zambonin & Balest, 2005). Remarkably, our study revealed that 2,3-butanedione, 2,3-pentanedione, 2-methylbutanal, 2-ethyl-6-methylpyrazine, 2,3-dimethylpyrazine are in fact statistically relevant to evaluate the provenance of coffee samples.

Moreover, worthwhile insights may be taken from the group of 18 compounds considered the most statistically suitable for discriminating coffee samples origin. For instance, seven of these compounds belong to an impacting class with high odour activity values (OAV) (De Maria, Moreira, & Trugo, 1999). It includes 2-methylbutanal, 2,3-butanedione and 2,3-pentanedione, known to impart oily and/or buttery aromas (Table 1); 2,5-dimethylpyrazine and 2,6 dimethylpyrazine, which induce roasted and hazelnut aromas; and finally 2-ethylpyrazine and 2-ethyl-6-methylpyrazine, whose aromatic features resemble those of roast and nuts. While all of the referred high OAV compounds are of positive organoleptic impact on coffee quality, the presence of 1-methylpyrrole reveals also that a negative aroma (known as a tracer of defective coffee grain) is in fact relevant to distinguish coffee samples. Hence, the DA models proposed in this work depend on discriminant molecules that impart both positive and negative notes to coffee.

Concerning the specific case of Brazilian samples differentiation, 2-ethyl-3-methylpyrazine and pyrrole-2-carboxaldehyde were included in the model in replacement of 2-methoxyphenol and 2-acetylpyrrole. While the new compounds are not of great odorant impact in comparison to the other two discarded, they are more relevant to distinguish Brazilian samples. In addition, the short model developed to differentiate samples in the “Brazil *vs.* America” assessment relies in great extent on previously mentioned

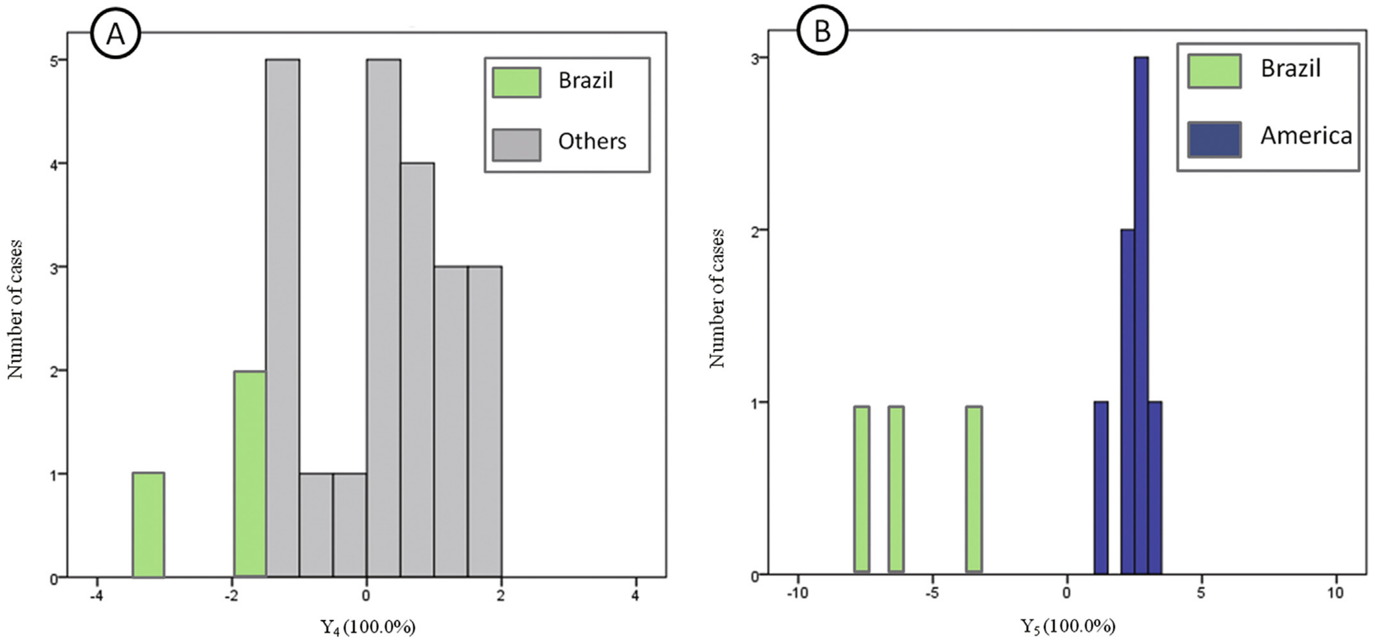

Fig. 5. Discriminant scores for the (A) “Brazil *vs*. Others”, and (B) “Brazil *vs*. America” analyses. Centroids of analysis (A) are −2.062 for Brazil and 0.281 for Others; centroids of analysis (B) are −5.655 for Brazil and 2.424 for America. The results show that 1D representation is enough to segregate the coffee samples from Brazil, particularly in case (B).

Table 8
Overall classification performance (count and %) of the DA model, with and without cross validation (CV), for the “Brazil *vs*. Others” study.

Original class	Predicted class (DA model)			
	Brazil		Others	
	Without CV	With CV	Without CV	With CV
Brazil	3 (100%)	3 (100%)	0 (0%)	0 (0%)
Others	0 (0%)	8 (38%)	21 (100%)	13 (62%)

Table 9
Overall classification performance (count and %) of the DA model, with and without cross validation (CV), for the “Brazil *vs*. America” study.

Original class	Predicted class (DA model)			
	Brazil		America	
	Without CV	With CV	Without CV	With CV
Brazil	3 (100%)	3 (100%)	0 (0%)	0 (0%)
America	0 (0%)	2 (29%)	7 (100%)	5 (71%)

compounds of oily and/or buttery aromas, namely 2-methylbutanal, 3-methylbutanal, 2,3-butanedione and 2,3-pentanedione, as well as molecules with roasted-like flavoring like pyrazine and methylpyrazine. The obtained results are in agreement with existing sensorial evaluations of American coffee samples, namely those stating that, due to its lower altitude (comparing to Colombia), Brazilian coffee exhibits sensorial notes of roast, bitter and soft (Sandalj & Eccardi, 2003). In contrast, coffees from Guatemala, for instance, evidence refined acidity, sweet aroma and rich taste (Murota, 1993).

As a final remark, it should be noted that not all compounds considered as highly impacting from a sensorial perspective (i.e., high OAV) were found statistically relevant to distinguish coffee samples with different origins. Hence, one may anticipate that sensorial analysis may be inefficient to sort accurately coffee samples according to their provenience, and thus the discriminant methodology developed in this work can advantageously aid stakeholders (importers and sellers) performing this task. On the other hand, DA revealed a good accuracy and hints to a deep chemical understanding of how to best differentiate coffee samples according to their respective origins.

5. Conclusion

In this work, Discriminant Analysis (DA) was applied to 25 coffee samples whose characterization (by HS-SPME-GC/MS and/or HS-PTR-ToF-MS) is reported in the literature, aiming at the identification of key compounds to differentiate these samples according to their continental or country (Brazil) provenance. A preliminary study on the identification/selection of a suitable compound to be used as standard for results normalization showed that pyridine is the best choice.

A model comprising three discriminant functions based on 18 compounds was built to classify coffee samples according to African, Asian, South American or Central American origin. The model relies on molecules like 2,3-butanedione, 2,3-pentanedione, 2-methylbutanal and 2-ethyl-6-methylpyrazine, which are known to possess high odour activity values (OAV) and thus a great impact on coffee organoleptics, and also on compounds without significant odorant features, like furfural and 1-hydroxy-2-butanone.

A single function model based on 18 compounds was built to specifically differentiate Brazilian coffee samples from others of any origin. These compounds were almost the same as those included in the previous model, with just a rearrangement of their relative importance for this type of discrimination. Finally, a model to differentiate Brazilian samples from American samples was also developed, based solely on 8 of the previous compounds, with preference for high OAV molecules.

The proposed models were both cross-validated and validated with new and independent data from literature, and their accurate classification capability was demonstrated. In addition, the accomplishment of permutation tests provided a complementary statistical validation of the DA models. Furthermore, the robustness of the proposed discriminant functions in cases of insufficient characterization (lack of data) was also tested, being concluded that our statistical models tolerate missing data, being still able to correctly classify coffee samples provenience.

In view of the successful application of DA to databases of this size and variability, this article provides compelling arguments for the development of DA-based tools with the purpose of assessing the quality of coffee in terms of their continental and/or national origins.

Acknowledgements

The authors thank the Brazilian National Research Council (CNPq) and the Coordination for the Improvement of Higher Level Personnel (CAPES) for financial support. This work was developed within the scope of the project CICECO-Aveiro Institute of Materials, POCI-01-0145-FEDER-007679 (FCT Ref. UID/CTM/50011/2013), financed by national funds through the FCT/MEC and when appropriate co-financed by FEDER under the PT2020 Partnership Agreement.

References

Agriculture, U.S.O.D.. (2015). *Coffee: World Markets and Trade*.

Akiyama, M., Murakami, K., Hirano, Y., Ikeda, M., Iwatsuki, K., Wada, A., et al. (2008). Characterization of headspace aroma compounds of freshly brewed arabica coffees and studies on a characteristic aroma compound of Ethiopian coffee. *Journal of Food Science, 73*(5), 335–346.

Akiyama, M., Murakami, K., Ikeda, M., Iwatsuki, K., Kokubo, S., Wada, A., et al. (2005). Characterization of flavor compounds released during grinding of roasted Robusta coffee beans. *Food Science Technology Research, 11*, 298–307.

Amstalden, L., Leite, F., & Menezes, H. (2001). Identificação e quantificação de voláteis de café através de cromatografia gasosa de alta resolução/espectrometria de massas empregando um amostrador. *Ciência Tecnologia de Alimentos, 21*(1), 123–128.

Bertrand, B., Boulanger, R., Dussert, S., Ribeyre, F., Berthiot, L., Descroix, F., et al. (2012). Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. *Food Chemistry, 135*, 2575–2583.

Cheong, M. W., Tong, K. H., Ong, J. J. M., Liu, S. Q., Curran, P., & Yu, B. (2013). Volatile composition and antioxidant capacity of Arabica coffee. *Food Research International, 51*, 388–396.

Costa Freitas, A. M., Parreira, C., & Vilas-Boas, L. (2001). The use of an electronic aroma-sensing device to assess coffee Differentiation—Comparison with SPME gas chromatography–mass spectrometry aroma patterns. *Journal of Food Composition and Analysis, 14*, 513–522.

Czerny, M., Mayer, F., & Grosch, W. (1999). Sensory study on the character impact odorants of roasted Arabica coffee. *Journal of Agricultural Food Chemistry, 47*, 695–699.

De Maria, C. A. B., Moreira, R. F. A., & Trugo, L. C. (1999). Componentes voláteis do café torrado. Parte I: Compostos heterocíclicos. *Quimica Nova, 22*, 209–217.

Flament, I., & Bessière-Thomas, Y. (2002). *Coffee flavor chemistry*. Chichester: Wiley.

Freitas, A., & Mosca, A. (1999). Coffee geographic origin—an aid to coffee differentiation. *Food Research International, 32*(8), 565–573.

Illy, A. (2005). Quality. In A. Illy, & R. Viani (Eds.), *Espresso coffee: The science of quality* (2 ed., pp. 1–19). London, UK: Elsevier Academic Press.

Jackson, J. E. (1991). *A User's guide to principal components, 587* pp. 1–58). Chicago, IL: John Wiley & Sons.

Maeztu, L., Sanz, C., Andueza, S., Paz De Peña, M., Bello, J., & Cid, C. (2001). Characterization of espresso coffee aroma by static headspace GC-MS and sensory flavor profile. *Journal of Agricultural Food Chemistry, 49*(11), 5437–5444.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition* (vol. 544, pp. 168–211). Chicago, IL: John Wiley & Sons.

Mondello, L., Costa, R., Tranchida, P. Q., Dugo, P., Lo Presti, M., Festa, S., et al. (2005). Reliable characterization of coffee bean aroma profiles by automated headspace solid phase microextraction-gas chromatography-mass spectrometry with the support of a dual-filter mass spectra library. *Journal of Separation Science, 28*, 1101–1109.

Moon, J. K., & Shibamoto, T. (2009). Role of roasting conditions in the profile of volatile flavor chemicals formed from coffee beans. *Journal of Agricultural Food Chemistry, 57*, 5823–5831.

Murota, A. (1993). Canonical discriminant analysis applied to the headspace GC profiles of coffee cultivars. *Bioscience. Biotechnology and Biochemistry, 57*, 1043–1048.

NBR ISO 9001/2000. (2001). *Brazilian association of technical standards: Quality management system* (Rio de Janeiro).

Powers, J. J., & Keith, E. S. (1968). Stepwise discriminant analysis of gas chromatographic data as an aid in classifying the flavor quality of foods. *Joural of Food Science, 33*, 207–213.

Risticevic, S., Carasek, E., & Pawliszyn, J. (2008). Headspace solid-phase microextraction–gas chromatographic–time-of-flight mass spectrometric methodology for geographical origin verification of coffee. *Analytica Chimica Acta, 617*(1), 72–84.

Rocha, S., Maeztu, L., Barros, A., Cid, C., & Coimbra, M. A. (2004). Screening and distinction of coffee brews based on headspace solid phase microextraction/gas chromatography/principal component analysis. *Journal of Science Food and Agriculture, 84*, 43–51.

Sandalj, V., & Eccardi, F. (2003). *O café: Ambientes e diversidade* (Casa da Palavra, Rio de Janeiro, Brazil).

Sanz, C., Czerny, M., Cid, C., & Schieberle, P. (2002). Comparison of potent odorants in a filtered coffee brew and in an instant coffee beverage by aroma extract dilution analysis. *European Food Research Technology, 214*, 299–302.

Smith, A. W. (1985). *Coffee Chemistry, vol.1* (pp. 1–41). London: Elsevier.

Toci, A. T. (2010). Compostos Voláteis e Qualidade do Café. *Matéria Prima, Torrefação e Armazenamento*. Brazil: Universidade Federal do Rio de Janeiro.

Toci, A. T., & Farah, A. (2014). Volatile fingerprint of Brazilian defective coffee seeds: Corroboration of potential marker compounds and identification of new low quality indicators. *Food Chemistry, 153*, 298–314.

Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., et al. (2008). Assessment of PLS-DA cross validation. *Metabolomics, 4*, 81–89.

Yener, S., Romano, A., Cappellin, L., Märk, T. D., Sánchez Del Pulgar, J., Gasperi, F., et al. (2014). PTR-ToF-MS characterisation of roasted coffees (*C. arabica*) from different geographic origins. *Journal Mass Spectrometry, 49*, 929–935.

Yeretzian, C., Jordan, A., & Lindinger, W. (2003). Analysing the headspace of coffee by proton-transfer-reaction mass-spectrometry. *International Journal Mass Spectrometry, 223*, 115–139.

Zambonin, C., & Balest, L. (2005). Solid-phase microextraction–gas chromatography mass spectrometry and multivariate analysis for the characterization of roasted coffees. *Talanta, 66*(1), 261–265.