Multivariate Analysis of Mineral Profile in Paprika with Protected Designation of Origin

TFM - Màster en Bioinformàtica i Bioestadística
Universiat Oberta de Catalunya

Autor

Óscar López Pascual

Directora

Dra. Nuria Pérez Álvarez

Coordinador de l'assignatura

Dr. Carles Ventura Royo

Àrea 2
Subàrea 2: Anàlisis de dades

Juny 2020
UOC-UB

Llicència Aquest treball s'ha dut a terme al laboratori de Química de Silliker Ibérica S.A.U (Merieux NutriSciences) a Barcelona.

Figure 1: Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

Agraïments

En primer lugar, me gustaría dar las gracias, "obviously", a mi colega Marta Parada. No sólo me ha ayudado de forma impagable en la revisión del texto en inglés, sino que en su momento me hizo ver que se podía ser más valiente y plantear un proyecto más ambicioso. (Gracias por tu energía y tu entusiasmo)

Gracias también a la tutora del proyecto, Nuria Pérez, por su clarividencia en la aplicación de la técnica de "random forest", sus buenos consejos y su contagioso optimismo.

Este trabajo no habría sido posible sin la exquisita profesionalidad del equipo "heavy metal" de Silliker Ibérica (Enric, David y Gemma), y la confianza depositada en el proyecto desde el primer momento por parte de Pascal y Roger. Gràcies.

Y por último y sí, más importante, gracias a Mónica, que siempre sufre en primera persona los "daños colaterales" de mis aventuras, y a pesar de ello siempre me anima a seguir adelante. Es una gran suerte para mí.

Cluster: *Cúmulo estelar de las Pléyades (Messier 45). A simple vista se identifican como siete estrellas agrupadas entre tantas otras visibles en el cielo nocturno. Un análisis más detallado revela una estructura más rica y compleja, compuesta de estrellas jóvenes y polvo estelar, unidas entre sí por la gravedad. Fotografía: Wikimedia (NASA, ESA, AURA/Caltech, Palomar Observatory)*

Fitxa del treball final

Títol del treball: *Multivariate Analysis of Mineral Profile in Paprika with Protected Designation of Origin*

Nom de l' autor: Óscar López Pascual

Nom de la consultora: Nuria Pérez Alvarez

Nom del PRA: Carles Ventura Royo

Data de lliurament: 06/2020

Titulació: Màster en Bioinformàtica i Bioestadística. UOC-UB.

Àrea del treball final: Àrea 2 - Subàrea 2 - Anàlisis de dades.

Idioma del treball: Anglès.

Paraules clau: random forest, ICP-MS, food fraud

Resum:
En un entorn cada cop més globalitzat, el frau alimentari és una preocupació creixent per part tant de productors i distribuïdors d'aliments, com de consumidors, i en aquest entorn els laboratoris de control alimentari juguen un paper clau en la seva detecció. Un dels possibles tipus de fraus és aquell vinculat amb l'incorrecte etiquetat de productes amb Denominació d'Origen Protegida. Des del punt de vista analític, una de les tècniques de detecció d'origen més modernes és l'anàlisis multivariant no selectiu, com el perfil mineral, seguit d'un tractament estadístic adient de les dades generades que permeti extreure conclusions relatives a l'origen del producte. Concretament en aquest treball s'han aplicat models d'aprenentatge automàtic sobre dades de perfil mineral de pebre vermell amb denominació d'origen, obtenint resultats de classificació que informin de l'origen dels productes. Les tècniques emprades han estat l'anàlisis de components principals, l'anàlisis de conglomerats, l'anàlisis lineal discriminant i, per primer cop en aquest tipus de producte, el mètode del bosc aleatori, que ha estat capaç de classificar correctament segons l'origen geogràfic la totalitat dels productes analitzats. S'ha

desenvolupat un mètode que inclou l'adquisició del perfil mineral mitjançant l'espectrometria de masses i l'algoritme de processat de dades, en llenguatge R, que podrà ser aplicat tant a altres denominacions de pebre vermell com presumiblement a d'altres tipus de productes amb denominació d'origen.

Abstract:
In an increasingly globalized world, food fraud has brought a growing concern in food producers, distributors and also consumers. In this context, food control laboratories play a key role in fraud detection. One possible type of fraud is related to incorrect labelling of products with a Protected Designation of Origin specification. From the analytical point of view, one of the main modern techniques for origin detection is the untargeted multivariate analysis, like mineral profile, followed by an appropriate statistical treatment of the produced data, which its aim is drawing conclusions regarding the origin of the product. In this work machine learning models have been applied to paprika's mineral profile data with a protected designation of origin for the obtention of classificatory results in terms of the origin of the products. The applied techniques have been Principal Components Analysis, Cluster Analysis, Discriminant Linear Analysis and, for the first time in this type of product, Random Forest method, which has been able to correctly classify all the analyzed samples according to the geographical origin. A global method has been developed including the acquisition of the mineral profile by means of Mass Spectrometry, and the data processing algorithm, in R language, which can be applied to other paprika designations and presumably to other types of products with Protected Designation of Origin.

Contents

1 Glossary

ANN Artificial Neural Networks
ANOVA Analysis of variance
AOAC Association of Official Analytical Chemists
CA Cluster Analysis
CART Classification And Regression Trees
GC-MS Gas Chromatography - Mass Spectrometry
ICP-MS Induced Coupled Plasma - Mass Spectrometry
ICP-OES Induced Coupled Plasma - Optical Emission Spectroscopy
kNN k-Nearest Neighbors
LC-MS Liquid Chromatography - Mass Spectrometry
LDA Linear Discriminant Analysis
NIR Near Infrared Spectroscopy
MS Mass Spectrometry
PC Principal Component
PCA Principal Component Analysis
P.D.O. Protected Designation of Origin
PLS Partial Least Squares regression
ppb parts per billion
ppm parts per million
RF Random Forest
SIMCA Soft Independent Modellin of Class Analogy
SVM Support Vector Machine
UV-vis UltraViolet-visible spectroscopy

sample[1]

[1] In the context of present work, "sample" is used as "observation" or "individuals from one population".

2 Project information

2.1 Context

The main goal of the present project is the development of a food fraud detection method in products from a specific Protected Designation of Origin (P.D.O.). Paprika "de la Vera" was the matrix chosen for the current study. The procedure proposed is based on the analysis of the mineral profile of paprika samples through multielement techniques like ICP-MS and the succeeding data analysis with multivariate models for the detection of possible fraud referring to the P.D.O. paprika "de la Vera" label.

The multivariate models applied include Principal Component Analysis (PCA) with Cluster Analysis (CA) together with artificial intelligence algorithms (Random Forest -RF- and Linear Discriminant analysis-LDA-).

2.2 Objectives

- Development of an ICP-MS method for the study of mineral profiles of paprika samples from the P.D.O.
- Apply multivariate data analysis techniques to be able to identify paprika from the P.D.O. and possible frauds in widespread paprika by means of supervised (PCA and CA) and unsupervised (RF and LDA) models.

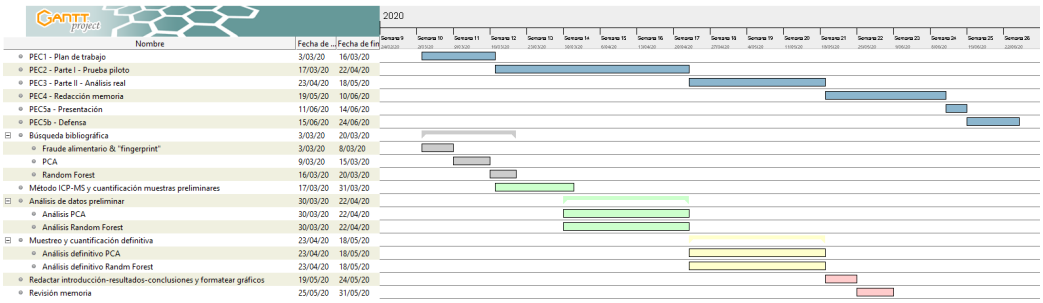
2.2.1 Specific objectives

- Select the elements for analyze from the periodic table that bring the most information in terms of geographical origins of the product.
- Create the acquisition and quantification method with the ICP-MS.
- Analyze the chosen samples with the ICP-MS for the obtention of the mineral profile.
- Design the algorithm for the data analysis (R code).
- Explore the obtained data (including hypothesis test).
- Application of unsupervised analysis (PCA, CA).
- Performance of supervised models (RF and LDA).
- Evaluate the obtained results and the method applicability for the desire

goal.

2.3 Project planning

The excuted temporal planning and the main tasks that have conformed the project are presented in a Gantt diagram. Real data was obtained as planned, and all tasks were performed following the scheduled plan.
The main source of uncertainty in the timeline was the derived from the pandemic COVID-19 crisis, which affected specially to sampling process.
Before the end of the project, two extra weeks were proposed by the University.



2.4 Sumary of results

Results of the project must include:
- The thesis, which includes the detailed bibliography that applies to the project, the obtained results and the drawn conclusions.
- The R code used for the data analysis.
- Intermediate progress reports (PACS from 1 to 3).
- The final oral presentation that will be delivered (PAC 4B).

Furthermore, if the obtained results are satisfactory, the project can be extended to the services of *Silliker Ibérica (Merieux NutriSciences)*, and also published in a scientific paper.

2.5 Sumary of chapters

After Glossary and Project Information, chapter 3 (Introduction) is an overview of Food Safety and Food Fraud state of the art. This chapter also covers geographical P.D.O."de la Vera" overview.

Chapter 4 is fosused in experimental data acquisition and first data exploration. Mineral fingerprint in Food Analysis and ICP-MS applied method are dicussed. Obtained dataset is described and data pipeline is presented.

Chapter 5 covers unsupervised machine learning models. After an introductory explanation and state of the art description, Principal Components Analysis and Cluster Analysis are discussed.

Chapter 6 repeats the same scheme as chapter 5, with supervised models. Linear Discrimant Analysis and Random Forest are presented.

To enlarge sample size, resampling technique of Bootstrap is covered in chapter 7. Principal Components Analysis and Random Forest are applied to the re-sampled subset.

Chapter 8 covers a project discussion, comparing the results from the different methods, and the final conclusions and future work.

Two addendum chapters are included. Addendum I shows the experimental dataset. Adendum II is the data pipeline. Rcode is presented, in Markdown format. This Addendum is presented in an attached document (pdf format).

3 Introduction

3.1 Food safety in a globalized world

According to the World Health Organization, "Food safety, Nutrition and Food security are inextricably linked. Unsafe food creates a vicious cycle of disease and malnutrition, particularly affecting infants, young children, elderly and the sick."[2]

In addition, food supply chains are crossing multiple borders, therefore control measures are nowadays mandatory in order to keep food safety and people's health. Analytical control is a key point for food companies and retailers, in order to make sure that their products are safe for the consumers. The more distance and borders the ingredients and products have to cross , the more difficult that control becomes.

Food safety is not the only factor in the "Food equation". Food is not only about providing energy and nutrients, or about people's health, but also about culture and human being's pleasure. In an increasingly worldwide connected planet, culture globalization means also food globalization, in the sense of local ingredients traveling abroad, based on their widespread popularity. Italian or Japanese cuisines, as an example, can be found all over the world (pizza and sushi are eaten in all continents). Therefore, the local produced cheese "Parmgiano" is exported around the globe, so the brand value might be highly affected by any related food safety issue or any commited food fraud. Consequently food control is not only affecting consumers but also producers.

3.2 Food fraud

The costs related to food fraud for the global food industry in the EU are estimated to rise every year up to 30 billion euros. [3]
Moreover, some fraudulent practises have drawn worldwide attention, arousing great impact within consumers' confidence. Some popular fraudulent

[2]"http://who.int/news-room/fact-sheets/detail/food-safety"
[3]"http://ec.europa.eu/food/safety/food-fraud"

episodes were:

- "Rapessed oil", Spain 1981. More than 600 people were poisoned and many others irreversibly affected due to the consumption of adulterated rapeseed oil. The mentioned vegetable oil was mixed with mineral oil, not allowed as food ingredient.

- "Dioxin crisis", Belgium 1999. Dioxins from a factory unintentionally contaminated the food production chain, which caused massive economic losses.

- "Melamine issue", China 2008. Melamine was introduced in milk and infant formula to fraudulently increase the nitrogen content of the product, and thus, simulate higher protein concentrations (products with elevated protein content are more valuable). The outcome of the fraud was bigger than expected, since the use of melamine as an ingredient originated kindness issues in infants.

- "The horse meat scandal", Europe 2013. Horse meat (not allowed for human comsuption) wwas mixed in processed meat products and sold as bovine (hamburgers, filled pasta, etc.)

- "Sudan dyes", Europe 2005. Sudan are azo dyes which have been shown to cause liver cancer in animal tests. Their use as food ingredients is forbiden, but they were found in paprika samples, to adulterate poor quality products.

Fraud may or may not be committed intentionally. Regardless of the purpose, surveillance from National and International agencies (related to food safety) plays a key role, by constantly monitoring food products. RASFF (*Rapid Alert System for Food and Feed*)(RASFF, 2019) and EFSA (*European Food Safety Agency*) are examples of anti-fraud supervision organizations at a European level.

Different classifications can be found to describe fraudulent actions based on diverse criterion. From a composition approach, two types of fraud are defined. The first group englobes those actions related to the addition, substitution or falsification of ingredients. The second class refers to labeling

issues: labels declaration and misleading ingredients, product characteristics or benefits, including false geographical origins.

3.3 Geographical origin. Paprikra *de la Vera*

Products authenticity is becoming nowadays an emerging topic: consumers from developed countries seek out high quality and locally produced goods, among other characteristics. Some reasons are encouraged by environmental-friendly ideas since pollution linked to transport is reduced and therefore the product footprint becomes lower.

Local manufacturing that follows defined procedures of production, like Protected Designation of Origins (P.D.O.), has earned reputation towards customers, based on higher quality products. This is the case of Paprika "de la Vera".

P.D.O paprika "de la Vera" refers the product obtained from the grinding of red fruits of the varieties "Ocaleas" , (Jaranda, Jariza, Jeromín), and "Bola" variety, belonging to the species *Capsicum annum L* and *Capsicum longum L*. These fruits are dried with oak firewood, following the traditional system of La Vera (one specific production area located in the South-West of Spain). (D.O.P.Vera, 2006) Therefore, species and variety, production process, and geographic production area are specific for the product labeled as Paprika "de la Vera".

4 Data adquisition

4.1 ICP-MS and Mineral fingerprint

4.1.1 Mineral fingerprint. State of the art

From an analytical point of view, two main approaches can be used to monitor food authenticity. First one is focused on detecting known targeted components that must comply with products' specifications. Going back paprika "de la Vera" example, the product specification, among others, for

ASTA color (an spectrophotometric absorbance value, performed at an specific wavelight) must be higher than one previously established value of 90 (D.O.P.Vera, 2006). The global compliance of all single criteria from the specification sheet is accepted as the product conformity, and generally recognized by producers and retailers.

The second approach is based on untargeted components analysis. Chromatograpic, spectrometric techniques (UV-Vis, NIR, MS) and their combinations in coupled techniques like GC-MS, LC-MS and ICP-MS, are also used to obtain non-specific data from the samples. These data, that could go from a sample spectrum to a multivariate data record, can be used as the sample "fingerprint": the data properly managed can be a descriptor of the sample in a really individual way, allowing the detection of similarities and differences between samples.

Inductively coupled plasma mass spectrometry (ICP-MS) is a coupled technique that combines mass spectrometry and inductively coupled plasma that ionizes the sample at atomic level. It creates atomic and small polyatomic ions, which are isolated and detected in the mass spectrometer. This technique is known as a powerful tool in terms of sensitivity, selectivity and high dynamic range. It can detect almost all the elements from the periodic table and also different isotopes of the same element, which makes it a versatile tool for isotopic labeling. Based on these characteristics, ICP-MS is one of the best choices for authentication studies, and especially for accurate geographical origin verification. The mineral profile and the relative abundance of natural isotopes is related to local conditions and may therefore provide information about the origin of food products. (Picó, 2015)

ICP-MS mineral fingerprint has been proven to be a reliable technique to identify the provenance in all types of food. Selected examples are the following:

-Lead and strontium isotopic ratios can be used for wine authentication if their contents in the wines are compared to their contents in the soil samples from where the grapes were cultivated. (Dehelean & Voica, 2012)

-Multielemental analysis by ICP-OES and exploratory data using PCA showed that the elemental composition of spices is influenced by the country

of origin, also allowing discrimination between countries.[4]

- Nineteen spices from the same country were classified into their different types and brands by ICP-MS followed by PCA and CA data analysis.(Tokalıoğlu, Çiçek, İnanç, Zararsız, & Öztürk, 2018)

Some elements and isotopes found in food, specially light elements like nitrogen, oxygen and sulphur, are strongly influenced by chemical, physical and biological phenomena. In addition, these elements are non metal, therefore non easily ionizable by ICP-MS. On the contrary, heavier elements (and easier to analyze by ICP-MS) are not that strongly influenced by biological phenomena. Elements like strontium or lead, for example, remain more constant because they are not subject to relevant seasonal variability or biological cycles. Once established into rocks, those metals are maintained unaltered in the passage from soil to food, which makes them good markers in terms of fingerprint for the determination of geographic location: finding fingerprint differences between paprika samples would allow the determination of the provenance location.

Several papers from recent literature demonstrate the good performance of mineral profile followed by different chemometrics analyses to investigate vegetable origins, although it is still a novel approach, therefore there are not officially recognized methods yet. A growing number of studies and publications is expected for the coming years, which indicates the great effort and the motivation of the international scientific community to improve food quality worldwide by using this multivariate approach.(Picó, 2015). Unfortunately, at the time this project is being written, not much information is found regarding mineral fingerprint analysis followed by chemometrics data analysis in paprika samples.

Recently, Ordog et al (Ördög et al., 2018) found differences between hot and sweet paprika of the Szeged region, Hungary, by multi-elemental ICP-MS followed by PCA analysis. Closer to the scope of the present study, Palacios-Morillo et al (Palacios-Morillo, Jurado, Alcázar, & de Pablos, 2014) have performed geographical characterization of paprika from the two paprika P.D.O. in Spain (de la Vera and Murcia), by using multi-elemental

[4]https://foodqualityandsafety.com/article/authentic-spices-identifying-country-origin

ICP-OES plus multivariate analysis. Good classification was obtained.

As a consequence of the novelty of these techniques applied to origin certification, and the lack of legislation related to its application within P.D.O.'s, every author is proposing different mineral selection for testing. Also, different chemometric approaches are described.

4.1.2 Selected elements and ICP-MS analysis

The chemical analyses for the present work have been conducted at Chemistry laboratory facilities of Silliker Ibérica (Merieux NutriSciences). Merieux NutriSciences is an international company dedicated to protecting consumers' health by delivering a wide range of test and consultancy services to the food and nutrition industries.

In Spain, the main laboratory is located in Barcelona. Throughout more than 40 years, the laboratory owns deep experience in nutritional testing, including minerals and metals analyses in food. The mineral tests are conducted by a dedicated and experienced team, and their results in food commodities are under the ISO17025 accreditation.

The used equipment includes an Agilent ICP-MS 7800 and a microwave oven Milestone Ultrawave. All paprika samples were firstly well mixed, and then a small and representative portion of every sample was heated in the microwave oven at high pressure, in an acidic-oxidant medium. During the digestion procedure all the organic content was removed, and the final sample extract was diluted and analyzed in the ICP-MS, where mineral content was quantified.

The aim in the selection of the elements for this project was to get a general view, as wide as possible, of the periodic table. Hence, light and heavy elements were included; also elements from different atomic groups: group 1 (alkali metals), group 2 (alkali earth metals), transition metals, rare earth metals and heavy metals.

In the metals selected there were included some multi-isotopes acquisi-

Element	Symbol	Atomic weight
Boron	B	11
Sodium	Na	23
Magnesium	Mg	24
Aluminum	Al	27
Phosphorus	P	31
Sulfur	S	32
Potassium	K	39
Calcium	Ca	44
Vanadium	V	51
Chromium	Cr	52
Manganese	Mn	55
Iron	Fe	56
Cobalt	Co	59
Nickel	Ni	60
Copper	Cu	63
Zinc	Zn	66
Strontium	Sr	86-87-88
Molibdenum	Mo	95
Cadmium	Cd	111
Tin	Sn	118
Antimony	Sb	121
Barium	Ba	137
Europium	Eu	153
Lead	Pb	208
Uranium	U	238

Table 1: Minerals

tions. Throughout the scientific literature it has been proven that geographic determination by mineral fingerprint using lead (Pb) and strontium (Sr) isotopic pattern has provided successful results. Sr, alkaline-earth metal, has four stables naturally occurring isotopes: 84Sr, 86Sr, 87Sr and 88Sr. Only 87Sr is radiogenic, and gradually increases in minerals due to the radioactive decay of 87Rb (rubidium). Differences in the absolute proportion of 87Sr vary with the geological ages and consequently with the geographical locations.(Dehelean & Voica, 2012) Therefore, the 87Sr/86Sr ratio can provide information regarding vegetables sample's origin grown in different types of soil.

Certified standards were used for quantitation purposes. An internal calibration method was used for the elements most analyzed in food: sodium, calcium, iron, lead,...For those elements where certified standards were not available, the quantitation software provides a semi-quant method, based on a default response of every element of the periodic table previously introduced by the manufacturer(Zhao et al., 2018). This semi-quant method is not as good, in terms of accuracy, as the internal calibration method with certified standards, therefore some bias can be introduced for these elements. This error component in the concentration value will be removed after the initial data transformation in last section of this chapter.

The high dynamic range of this technique allows to quantify from the ppt's to the ppm's range without any extra dilution, so all results per sample could be collected in a single run.

4.1.3 Sampling

In every analytical method, including those focused in food testing, sampling is a critical step in order to assure reliable results, in terms of precision and accuracy. Correct sampling is one of the most important and challenging steps in food fraud analysis.

An important aspect of sampling in fraud detection analysis is knowing the actual size of the population under study. In P.D.O. samples, as paprika "de la Vera", the population can be measured in terms of the number of certified producing companies/factories. Currently, 16 manufacturers are listed

in the P.D.O. The regulatory council P.D.O. "de la vera" certifies that all the paprika products labeled with "de la vera" comply the specification. Three main varieties of paprika are produced (sweet, spicy and bittersweet) in a specific smoking process used to dry the pepper, which contributes to the characteristic flavor of the product.

Inter-lot and inter-harvest variations can also increase the population of the product for sale. At the pre-planning phase,contacting with P.D.O. was considered in order to collaborate, exchanging samples and information. Due to the circumstances occurred during this project (global pandemic), this contact was never made, therefore the info from this source is not available.

Other types of paprika from all over the world, including fraudulent samples, were also analyzed, belonging to a second population (not "de la vera" paprika). Since there is an unquestionable impossibility of knowing the whole population, a "supermarket sampling" approach was performed: the laboratory was provided with as much paprika samples as possible (considering mobility restrictions during pandemic crisis in Spain).

Fraudulent samples were not known as fraudulent in advance, so the project target was focused on finding statistical difference between "de la Vera" labelled samples and the rest of paprika samples.
From the 27 samples analyzed, 12 were from paprika "de la Vera" population (labelled), the other 15, without the distinction logo, were randomly picked from local markets. To get a general idea of studies carried in this field: other research with paprika analyzed a number of samples in the range of 100-150.(Palacios Morillo, 2015). A fraud study in almonds (López, Trullols, Callao, & Ruisánchez, 2014) was conducted with 28 samples.

Since during the development of the project local market samples were the only available, one of this work prospects would be the enlargement of the sample size (specially if the P.D.O's involvement can bring material for study). The greatest concern during the sampling period was to find more samples than variables for study, which was a must in order to develop a feasible and global method and data analysis. Same procedure, data treatment techniques and data pipeline can be applied to a larger dataset, therefore, even the actual size is big enough to test the model, it can be updated and

ID	Paprika Type	Vera: Y/N
1	Sweet	Yes
2	Sweet	No
3	Sweet	No
4	Sweet	Yes
5	Sweet	Yes
6	Hot	Yes
7	Sweet	No
8	Sweet	No
9	Bittersweet	Yes
10	Sweet	No
11	Sweet	Yes
12	Sweet	Yes
13	Sweet	No
14	Sweet	No
15	Hot	Yes
16	Sweet	Yes
17	Sweet	No
18	Bittersweet	Yes
19	Sweet	No
20	Sweet	No
21	Hot	No
22	Hot	Yes
23	Sweet	No
24	Sweet	No
25	Sweet	Yes
26	Sweet	No
27	Hot	No

Table 2: Samples (form different producers and lot numbers)

enlarged anytime without important changes in the pipeline and analysis.

4.2 Dataset

27 experiments were performed (paprika samples) and 28 variables were studied: sample ID, levels factor (Vera-NoVera) and 26 elements concentration in each sample expressed in μg/kg (variables from 3 to 28). The data set layout consisted of rows defining the samples of the study and the columns containing the information of the variables.

Not all the minerals were in the same range of concentration in every sample, some metals content were considerably higher than others. Results were all expressed in micrograms of analyte per kilogram of sample (ppb), hence a wide range of values can be found in the dataset: phosphorus mean content is about 9.5e6 (0.95g/100g) ppb while antimony mean concentration is 25ppb (0.0000025g/100g).

id	vera	11.B	23.Na	24.Mg (...)
1	1	35306.93	315708.74	5510893.34
2	0	21764.12	909690.43	7178715.61
3	0	21163.09	949565.85	7346523.88
4	1	32823.16	598800.01	4999496.65
5	1	38335.73	354324.15	5752216.70
6	1	34660.70	476751.36	5324170.98
(...)				

Table 3: Dataset (head)

In order to fit all the information (metals concentration from the raw data) in a narrower and comparable range of values, the data was normalized to perform the algorithm: all the values for each variable were divided by its mean, obtaining a more cententered and unitless dataset, while variability was preserved.

All data analyses have been conducted with R (R Core Team, 2019). R code and data can be found in the addendum.

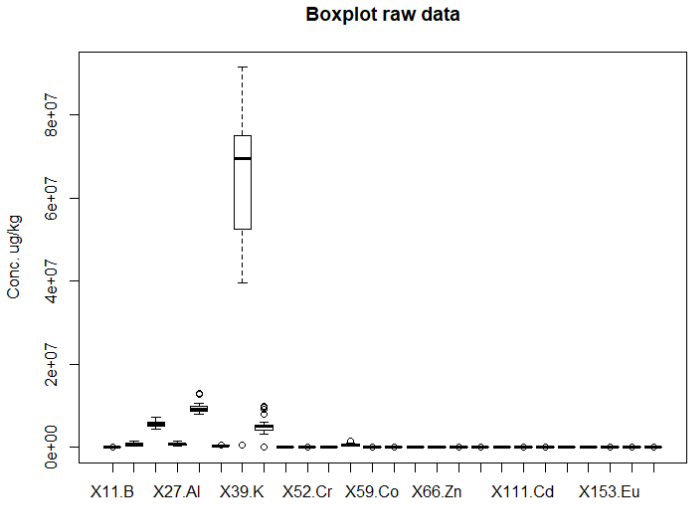


Figure 2: Boxplots of raw data

4.2.1 Statistical analysis

After the data acquisition in the ICP-MS the data analysis was performed. The proposed data workflow (the data pipeline, or global algorithm), included the statistical analyses and models are described here below.

First performed step was data importation to the statistical software used in this project: R. All the data analysis was performed with different R libraries and every detailed step is shown in the Pipeline and Rcode section (Addendum II).

After importation, data exploration and normalization was performed with graphical and summary functions from R. Methods for outliers detection were not used since precision and accuracy was proven and considered as acceptable during ICP-MS validation (previous to this work). [5]

[5]based in ISO17025, every analytical technique must be validated before using it, re-

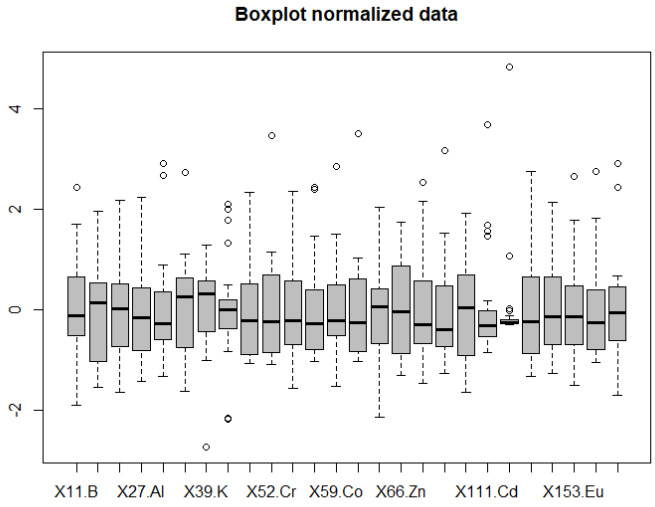


Figure 3: Boxplots of normalized data.

First approach to the data was based on the study of multivariate/univariate data distribution, homogeneity of variances and homogeneity of means.

Then two types of data analysis models were applied to the dataset: first unsupervised models (Principal Component Analysis and Hierarchical Cluster Analysis), secondly supervised models (Linear Discriminant Analysis and Random Forest). Unsupervised models were used in this case with viewing purposes, in other words, displaying multivariate data. For the application of supervised models the original dataset was divided into two subsets (training set and test set): the models were trained for classification purposes (Vera-NoVera), and then validated with the test set in order to calculate the error generated from the application of the model.

Finally, due to the size of the dataset, Bootstrapping was applied to simulate a dataset with a higher number of observations. In this work, this step was performed to check the models working with a more likely size dataset. Not all the applied models were used with the bootstrapped dataset, only

porting evidences of its analytical performance

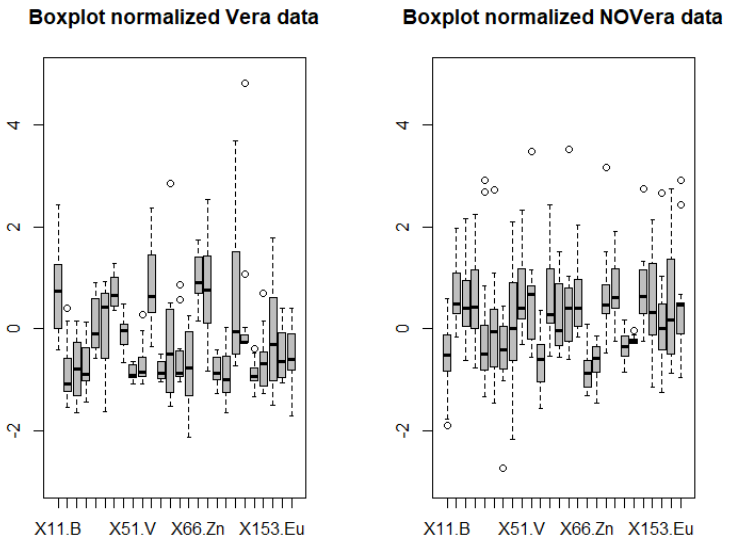


Figure 4: Boxplots of normalized Vera and NoVera samples.

PCA and RF were performed (one unsupervised and one supervised model).

4.2.2 Data exploration

Regular data exploration commonly includes data viewing, data distributions study (checking for normal distribution), variance study (homoscedasticity) and mean comparison (ANOVA). The first exploration performed was based on dividing the set of data in the two levels of the variable defined as a factor: values classified with Vera and NoVera (data represented was also normalized).

Continuing with the set separated by the factor Vera/NoVera, the following graphs show some differences between both subsets for some selected minerals.

Depending on the mineral, different behaviour between Vera and NoVera samples is shown. Data comparison will be performed in coming sections.

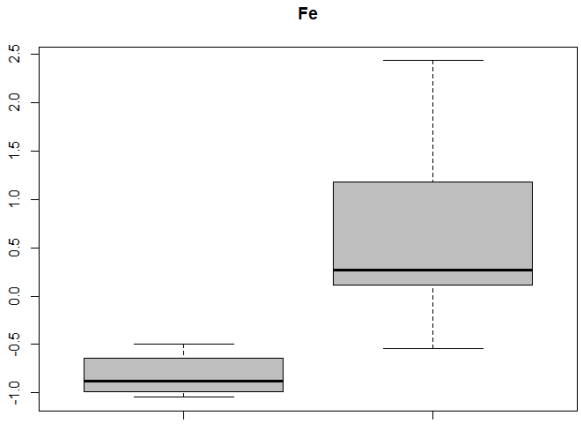


Figure 5: Iron.Left:Vera-Right:NoVera

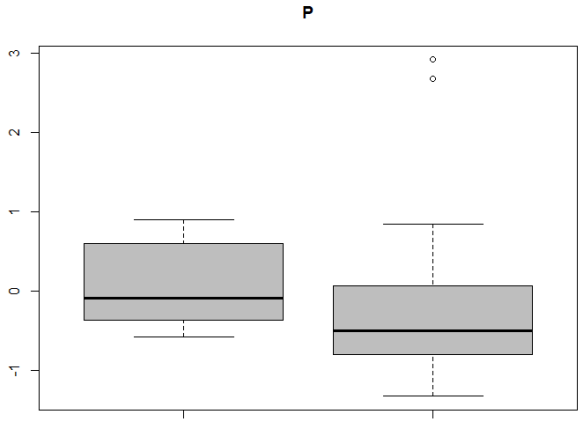


Figure 6: Phosphorus.Left:Vera-Right:NoVera

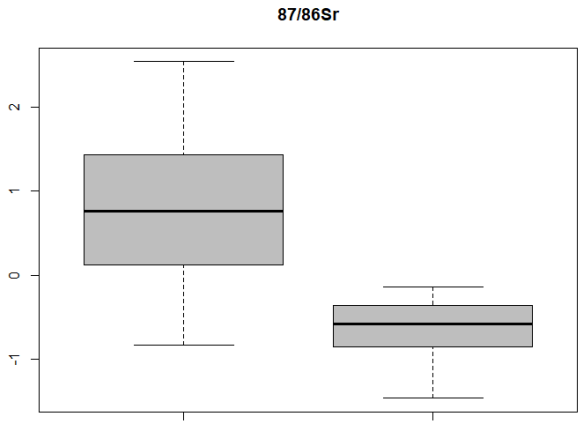


Figure 7: Strontium Isotopic ratio.Left:Vera-Right:NoVera

4.2.3 Normality

Second step, related to the study of the distribution of the data set, was performed: normality study. Due to the number of observations (12 Vera samples and 15 No Vera samples) and the high number of variables taken into account (26), multivariate test like Mardia test (mardiaTest MVN) cannot be directly applied, considering that the sample size is close to the dimension and only few methods can deal with this situation(Tan, Fang, Tian, & Wei, 2005). Furthermore, if the dataset is divided into Vera and NoVera subsets, sample size is actually lower than the number of dimensions, which implies that multivariate normality tests are discarded. Univariate tests can be used keeping in mind that a p-value correction might be needed in order to avoid errors from multiple comparisons (increase of false positives/negatives, errors known as type I and II).
Focusing on the set of data under study, Shapiro test (univariate normality test) was applied with exploratory purposes: 19 out of the 26 variables from the Vera subset obtained a p-value higher than 0.05. For the NoVera subset, 17 of the 26 variables have a p-value higher than 0.05. As an example, iron's QQ-plot Figure 5 shows the data distribution.
In the chapter "Bootstrap" normality will be checked again.

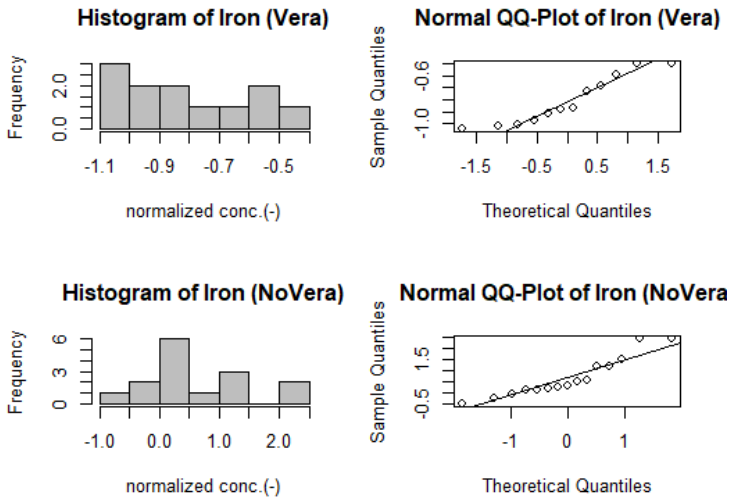


Figure 8: Iron data distribution in Vera and NoVera subsets

4.2.4 Homoscedasticity

To check homogeneity of variances, Leven's test was used (median formula was applied). Levene test assess the equality between two or more groups variables variances. Two groups were considered, Vera/No Vera, and for every variable (mineral) a p-value was obtained. P-values showed homoscedasticity of variances between Vera and No vera for 19 out of the 26 minerals observed (equality of variances was assumed). For the other seven minerals studied, the null hypothesis of equal variances was rejected (p-value < 0.05), therefore homoscedasticity for those metals between the two subsets was not proved.

Lack of homoscedasticity for some of the minerals could happen due to several reasons. Low size data can generate a wrong variance estimation because of a possible bias in the mineral data. Another reason can be the heterogeneity of NoVera samples. The only matter they have in common is that they are not produced in the P.D.O., so differences among NoVera observations when checking the content of some minerals are reasonable.

4.2.5 Comparison of means

After the study of the variance, mean comparison was performed, and despite the fact that homoscedasticity was not assumed for all variables, MANOVA test was applied. MANOVA is the multivariate generalization of ANOVA which infers the t-test beyond two means. Although in the analysis only two means were compared (Vera and NoVera) MANOVA was used due to its applicability in multivariate scenarios. Hotelling's T^2 test also applies to this particular case. Same conclusions were obtained with the two tests proposed: mean difference between groups (Vera- NoVera) is significative (p-value: 0.02188).

5 Data analysis by means of unsupervised models

5.1 Introduction and state of the art

Since the last century physicists have been looking for a "Theory of Everything"[6], a theoretical framework that fully explains and links together all physical aspects of the universe. This unsolved problem has already some candidate theories, like the "Strings Theory"[7]. According to this currently unproven theory, a multidimensional space, with more than three dimensions, has been proposed. But space with more than three dimensions (width, height, length) is not possible to imagine by the human being. One point is one dimensions object, one line is a two dimensions object, and one cube is a three dimensions object, but none can imagine a four dimensions object which is precisely the disadvantage of multivariate analysis: data with more than three variables cannot be plotted in a two-dimensions or three-dimensions graph. On the contrary, the advantage is that much more information can be obtained, compared to univariate data.

"A picture is worth a thousand words", therefore techniques applied to reduce data dimensions without losing much information, would allow us to perform two- or three-dimensional plotting, which can be very helpful to understand multivariate data. For instance, Principal Components Analysis (PCA) is widely used in multivariate experiments. PCA and others PCA-related techniques have been commonly used in food fraud detection(Callao, 2014). In addition, LDA, CA and PLS have been commonly used in several food products. Different factors may influence in the selection of the technique: number of classes, number of variables, type of data (discrete, continuous) and the aim of the analysis. (M. Forina & Oliveri, 2009)

In general terms, two main types of models are used in food fraud testing, moreover, in the whole data analysis field: unsupervised and supervised models.

Unsupervised models work with a set of observations of a variable (X)

[6]https://en.wikipedia.org/wiki/Theory_of_everything
[7]https://en.wikipedia.org/wiki/String_theory

without knowing the association with the other variable Y (Y=f(X)). The aim of this models is not the prediction (for instance giving response to Vera-No Vera), but to visualize the data and/or detect subgroups among the samples. (James, Witten, Hastie, & Tibshirani, 2013). Two of the main techniques of unsupervised models are PCA and CA (correspondence analysis). While PCA is used for data visualization or data pre-processing (before supervised techniques are applied), CA is applied for detection of subgroups in the data set.

It is important to keep in mind that there is no way to check the obtained results from the application of those methods because it is not possible to know the true answer. (James et al., 2013). Despite the fact that unsupervised methods do not have any universally accepted mechanism for validating results from an independent data set, they are extensively used in fingerprint analysis and fraud analysis in food, specially PCA before applying supervised models(Berrueta, Alonso-Salces, & Héberger, 2007)

Supervised models will be discussed in the next chapter.

5.2 Principal Components Analysis

"PCA is a multivariate technique with the central aim of reducing the dimensionality of a multivariate data set while accounting for as much of the original variation as possible present in the data set. This aim is achieved by transforming to a new set of variables, the principal components, that are linear combinations of the original variables". (Everitt & Hothorn, 2011)

These new "variables" called principal components (PC) are not correlated, which is an important attribute for fingerprint analysis. Correlation of the variables may involve redundant data. Moreover, principal components can be arranged based on their variation, from PC with higher variation associated to the lowest. In PCA analysis, variation of the variables is understood as the amount of information that the PC brings, so the method of analysis allows to keep most of the initial information using few variables: in most of the cases , two or three principal components account for more than 70-90% of the total variation from the original set. As a linear combination of the original variables, the PC's have no units, meaning that in case

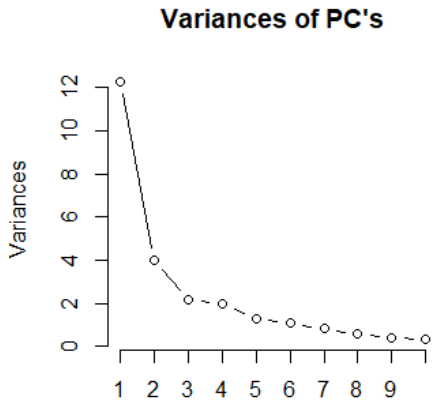


Figure 9: Variances vs PC

the variables were expressed in terms of concentration of the minerals, they would be dimensionless in a two or three dimensions PCA plot.

5.2.1 PCA results

PCA was applied to the normalized dataset. Calculations were performed by "prcompstats" function for R (R Core Team, 2019).Then ten principal components were plotted, principal components variation is shown in Figure 9. Three PCs are the maximum of variables that could be plotted. Furthermore, the "elbow" (change in the decreasing tendency), is located in the 3rd PC, which implies that from the 4th PC there is not much variation added.

The total cumulative variation of a PC (expressed as the percentage of the variance of the PC referred to the total dataset variation) is: 47% for the 1st PC, 63% for the 2nd PC, 71% for the 3rd, 79% for the 4th. 99% of the total variance is explained at PC14, and 100% at PC26.

Plotting two PC's it is shown 63% of the total variance of the data in a two-dimensions graph (Figure 10)
A three-dimensions graph shows 71% of the total variance (Figure 11).
In both graphs, samples from the two factors (Vera-NoVera) were high-

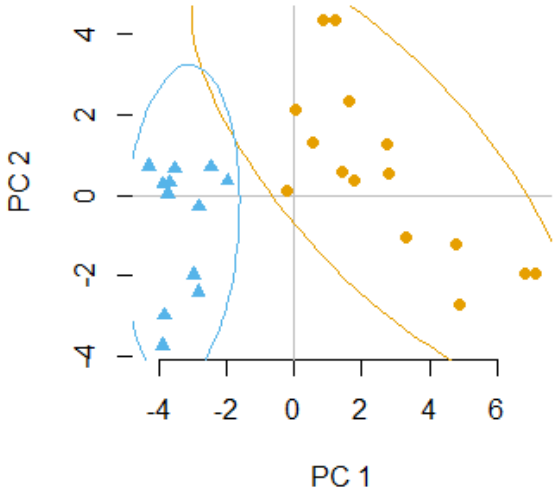


Figure 10: Blue triangles represent Vera samples, orange dots represent NoVera.

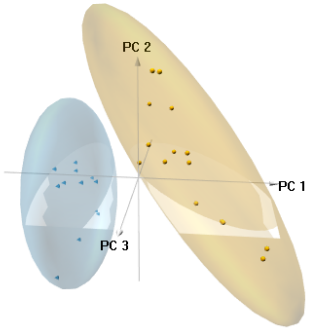


Figure 11: Blue triangles represent Vera samples, orange dots represent NoVera.

lighted with different colours, and ellipses show, respectively, the two- and three-dimensions 95% confidence intervals.

For a two-dimensions analysis, it is shown the coefficients applied to the first and second PC (plotted as coordinates). This is called the "variables factor map" and it is frequently shown together with PCA graph (Figure 12).

5.2.2 PCA conclusions

Vera and NoVera samples can be distinguished applying PCA analysis. Using the first three PCs (71% of the total dataset variance) both factors are completely separated with 95% of confidence (if normal distribution assumed for Vera and NoVera subsets). The small surface for Vera samples in both graphics (Figures 10 and 11) in comparison with the area for NoVera can be explained in terms of geographical origins of the samples: The group of NoVera includes samples which the only thing they have in common is that the origins do not belong to "la Vera" location. On the contrary, all Vera

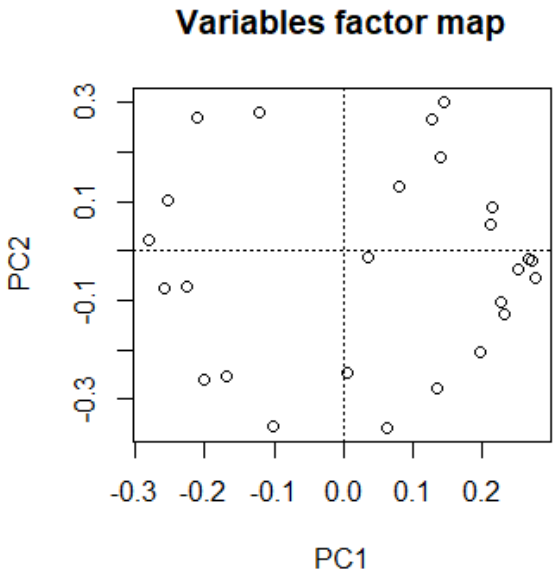


Figure 12:

samples have a common geographical origin.

PC's are composed by the linear combination of the "original" variables, i.e. analyzed elements. For instance,

$$PC1 = \alpha \cdot B + \beta \cdot Na + \gamma \cdot Mg + \cdots$$

Coefficients (α, β,γ,...) of the first and second PC's (mathematically known as eigenvectors) are shown on table2. They are the coordinates of Figure10, and describe the contribution of the original variable in the new one (the PC).

5.3 Cluster Analysis

"Cluster analysis is a generic term for a wide range of numerical methods with the common goal of [...] discovering groups or clusters of observations that are homogeneous and separated from other groups."(Everitt & Hothorn, 2011). It is included in the unsupervised models group because the response

variable Y=f(X) can be unknown. After PCA, CA has been the most used unsupervised technique in food analysis during the last decade(Berrueta et al., 2007)

There are different types of clustering methods, being k-means clustering and hierarchical clustering the two main examples. In this dataset, hierarchical clustering was applied.This technic classifies data from one single cluster (the whole data) to one individual cluster for each sample from the total data set, generating a "tree-shape" graph called dendrogram that shows different groups within the observations. This groups are build based on the distance between the individuals, which can be measured in different ways: Euclidean and Mahalanobis distance are the most common techniques. Euclidean distance is a generalization of the Pythagorean theorem for a multi-dimensional space. It is the straight-line distance between two points in an Euclidean space. With the data of the distance between all pairs of individuals, it is built the distance matrix. CA method uses this distance matrix to build the dendrogram, following the hierarchical clustering algorithm. It was the process used for this work.

Once the dendrogram is being plotted, the number of partitions of the data must be decided, since the dendrogram could be "cut" at any height. The number of partitions is known as k, and different methods are described to select which could be the best number of partitions to be applied in a specific dataset. In most cases, k cannot be easily decided with mathematical methods, and the opinion of the experts must be considered. (Irizarry & Love, 2016). Concerning the current dataset, two main groups (Vera-NoVera) were expected. However, as NoVera samples origin is unknown, more clusters inside NoVera samples can be found. The mathematical method proposed to get the best number of clusters k for the dataset is Average Silhouette, which represents the silhouette of the average value (mean of the similarity between de individual samples and the clusters they belong) for every value of k.

5.3.1 CA results and conclusions

Complete linkage and Euclidean distance methods were used with the R function *agnes(cluster)* for cluster analysis calculations. The obtained den-

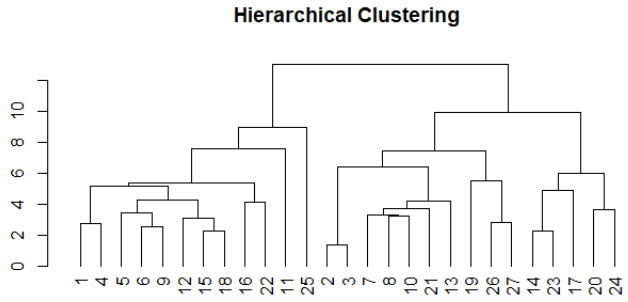


Figure 13: Samples on the left (1-4-...-25) are Vera samples. Non Vera samples are located on the right (2-3-...-24). See details in table 4

drogram is shown in figure 13.

	id	1	4	5	6	9	11	12	15	16	18	22	25		
	vera	1	1	1	1	1	1	1	1	1	1	1	1		
id	2	3	7	8	10	13	14	17	19	20	21	23	24	26	27
vera	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: Association of Hierarchical Clustering numbers (Fig.13) to the ID and factor of each sample to better results understanding.

The average silhouette graphic (Figure 14) indicates that the best average is achieved with 2 or 3 clusters.

We can see below the clustering graphics obtained with k=2 and k=3 (Figure 15)

From a non-mathematical point of view, with k=2 we can observe two groups perfectly define: Vera and NoVera. When k=3 , NoVera group was split in two subsets. With these analyses we concluded that CA and PCA results are concordant and furthermore those techniques are capable of differentiate Vera and NoVera samples.

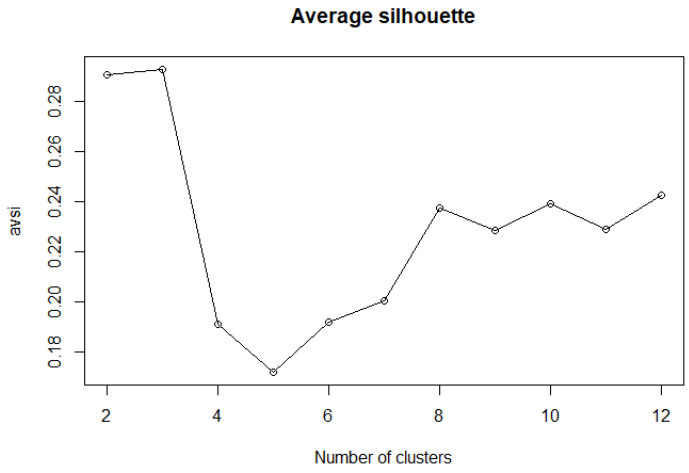


Figure 14:

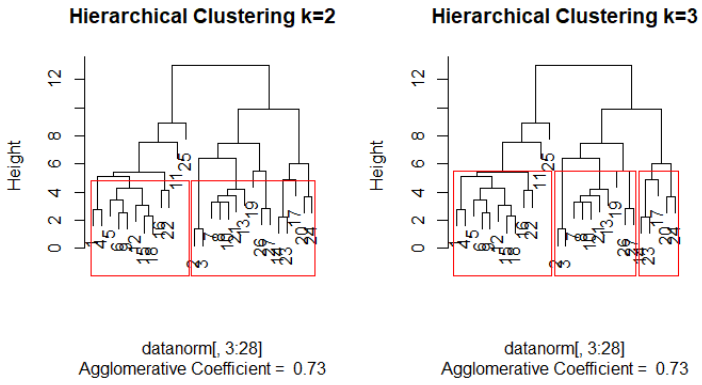


Figure 15: k=2 and k=3

6 Data analysis by means of supervised models

6.1 Introduction and state of the art

As we have discussed in previous chapters, the response variable is not needed to the application of the model: PCA results (coefficients, coordinates, plots) are the same whether Vera- No Vera factor is known or not. Supervised models are based on the knowledge of the response: training the model. These types of algorithms are also known as supervised machine learning.

Any machine learning algorithm includes the following steps: (1) data collection, (2) data exploration and normalization, (3) model training, (4) model evaluation and (5) model improvement. (Lantz, 2013).For step 3, a training set is needed: in this project, since the Vera-No Vera parameter was known for all samples, the training was performed with a random subset from the original dataset. Model evaluation or validation (step 4) was performed with another random subset, called test dataset. This set of data for testing was processed as unknown samples, and the responses obtained from the model were compared with the known information for that data. This procedure allows the model to evaluate itself, and to obtain quality data as false positive and negative errors. Usually the applied ratio for "training set:test set" is "80:20". This type of validation procedure is known as *cross-validation*.

There are a wide range of techniques included in supervised machine learning, and they have been applied in very varied fields: predicting results of elections, discovering genetic sequences linked to diseases, or forecasting of weather behaviour and long-term climate changes(Lantz, 2013) Even though supervised machine learning models have not been used in the food industry analysis as much as in other areas like economics or biostatistics, the last decade shows an increasing tendency on its application in this field, including multivariate analysis and supervise models. LDA is the most common supervised technique for classification purposes in food analysis. Other common applied techniques are k-nearest neighbours (kNN) , classification and regression trees (CART), artificial neural networks (ANN), partial least squares discriminant analysis (PLS-DA), soft independent modelling of class

analogy (SIMCA) or support vector machine (SVM).(Berrueta et al., 2007)

It is worth to note that it is expected for food industry to experience a great increase in supervised model applications in the multivariate food analysis in the coming years. The predictions skills of supervised machine learning models, together with the growing popularity of Mass Spectrometry and its multivariate data acquisition, have become promising fields for food laboratories and industries. The amount if information that is currently produced with modern instrumentation, combined with the ease of computing and predicting with supervised models, will bring loads of information, compared with classical targeted screening approaches.

6.2 Linear Discriminant Analysis

LDA is a technique based on finding the linear function that does the best separation in classes. It can be used in simple class separation, 2 factors (Vera-No Vera), or multiple class separation (with higher number of classes). LDA and PCA are very popular techniques in multivariate food analysis, both of them reducing data dimension and projecting to a lower dimension space: "PCA selects a direction that retains maximal structure among the data in a lower dimension, LDA selects a direction that achieves maximum separation among the given classes"(Berrueta et al., 2007).

The result of LDA is a contingency table, where known results are compared with the model output: as an example a 2x2 table can be built showing true positive and true negative values, false positive and false negative values, [8], which gives valuable information of the model quality and performance.

6.2.1 LDA results and conclusions

When model validation is performed using a cross-validation approach, the original data must be randomly divided into a training and a test set. In

[8]In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a "false positive" cocnlusion), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" conclusion. *https://en.wikipedia.org/wiki/Type_I_and_type_II_errors*

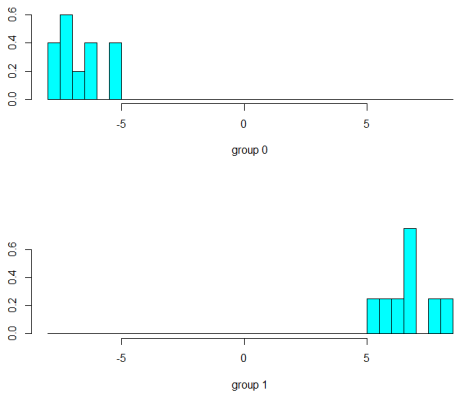


Figure 16: NoVera:group 0 - Vera:group 1

order to do that, a sufficient number of observations is recommended: the test set should be around 20% of the total size of the dataset. In this project dataset, considering the number of variables and available observations, finding a good compromise balance between training and test was necessary. The lowest test set size was fixed in nine observations. Considering that Vera samples were 44.4% of the dataset, a random test set will have 4 Vera-group observations.

Before LDA, dataset partition was performed with R function *createDataPartition(caret)*. LDA was applied with R function *lda(MASS)*. LDA function returns the group means for all variables and the coefficients of linear discriminants. The means were displayed in the histograms (Figure 16).

Both groups (Vera-NoVera) were properly separated, and thus, apparently, the model has a clear discriminant decision area. After LDA was performed, the test set was introduced in the model , and the output was compared with the known information. The following contingency table was obtained (Table 5)

The number of false positives and false negatives is zero, therefore the model success is 100%. However, it should be noted that subset partition is a random process. If data partition is repeated again and again with random

	predicted	
test	Vera	NoVera
Vera	4	0
NoVera	0	5

Table 5: Contingency table

test sets, in some cases false positives and/or false negatives may appear. This is known as the test error. Test error can be calculated with resampling techniques that will be discussed in next sections (Random Forest and Bootstrapping).

6.3 Random Forest

The first algorithm for Random Forests (RF) was created and first published in 1995 by Tin Kam Ho, 3 years after she received her PhD degree in computer science from the State University of New York at Buffalo.(Ho, 1995). The term "Random Forest" was proposed by Leo Breiman and Adele Cutler in 2001.(Breiman, 2001) Since then, Random Forest has grown very fast, probably due to its capacity to improve the method?s accuracy for both training and test sets, comparing to other supervised methods. Due to their power, versatility, and ease of use, Random Forests is becoming one of the most popular machine learning methods. (Lantz, 2013)

However, RF is a young technique that has just arrived to some scientific areas, like food analysis. To get an idea of the actual situation, the search of "random forest" in the AOAC scientific journal (J.AOAC International) returns only one published paper, from 2019(Lim et al., 2018)

Random Forest is based on *decision trees algorithms*, which are supervised methods that can be applied for data classification (and regression). Decision trees are composed of nodes and shaped like trees, the trunk represents the whole data set, and successively the data is divided depending on the question located in the node generating brunches. The nodes represent division points based on conditionals, for instance if the data of one variable is higher or lower than a randomly selected value. Depending on the value,

the data continues in one or the other brunch, until the same procedure happens further down in the next node (child node), where the variable is processed in the same way. The process is repeated until the terminal node is reached, and a classification label is assigned. In our project , the label was the factor Vera-No Vera. Since the label and the actual factor for each sample was known, the tree response was evaluated, and depending on the success of the process (number of successful labels), the tree learned the best cutting-values and variables order for each node.

Random Forest goes one step forward from the techniques previously studied, takes the prediction of the decision tree, and produces hundreds of them. The average of all the predictions provides a more accurate result than a prediction obtained from a single model, therefore the testing error observed in LDA is reduced. Another strength is that RF fixes the issue of the correlation of the trees: when decision trees are built, if one of the variables is a much better predictor than other, there is the risk of many of the trees using the same variable. RF algorithm starts every time with randomly selected predictors, and thus correlation is reduced and so variance. The consequence of all of the above is the reduction of the test error, by applying a supervised machine learning model and a smart resampling and average method.

6.3.1 RF Results and conclusions

The present work applies for the first time random forest analysis to mineral fingerprint data in paprika samples (probably due to the still low incidence of RF techniques in the food field). The R function used is *randomForest(randomForest)*. Same training set and test set obtained for LDA was used. 500 were the number of trees used in the model with 5 variables tried at each node (default randomForest function value). All parameters are the randomForest function default values.

The output of the RF function returns the following information:
- The training test confusion matrix shows 0% of error (false positives and false negatives). 10 NoVera samples were returned by the model as NoVera samples, and 8 Vera samples were classified as Vera samples.
- Mean decrease Gini (Figure 17 - Table 6). This value is directly related to

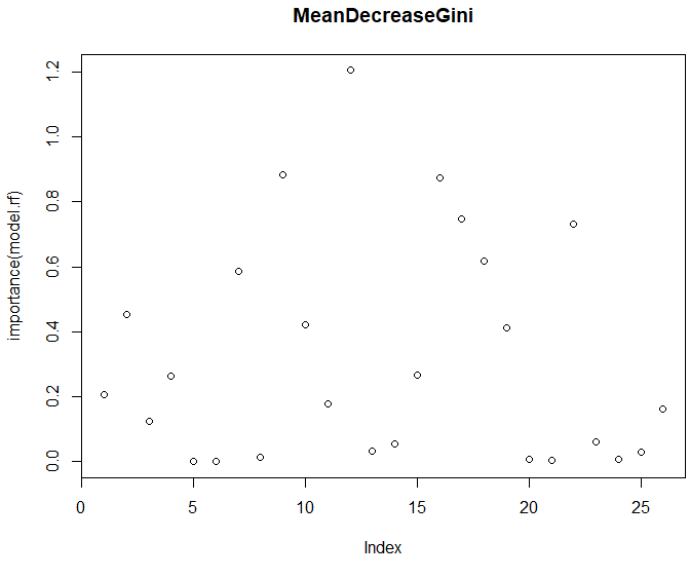


Figure 17:

the variable significance. When RF is applied as an unsupervised model because the value of the response is not known, the main information obtained is the Gini parameter.

	MeanDecreaseGini
11.B	0.21
23.Na	0.45
24.Mg	0.13
27.Al	0.26
31.P	0.00
32.S	0.00
39.K	0.59
44.Ca	0.01
51.V	0.88
52.Cr	0.42
55.Mn	0.18
56.Fe	1.20
59.Co	0.03
60.Ni	0.06
63.Cu	0.27
66.Zn	0.87
87.86.Sr	0.75
88.Sr	0.62
95.Mo	0.41
111.Cd	0.01
118.Sn	0.00
121.Sb	0.73
137.Ba	0.06
153.Eu	0.01
208.Pb	0.03
238.U	0.16

Table 6: MeanDecreaseGini

The most important predictors according to the RF model are iron (Gini:1.2), vanadium (Gini:0.88), zinc (Gini:0.87), 87/86Sr (Gini: 0.75) and antimony (Gini: 0.73)

Figure 17 shows all Gini values ordered by element weight. 26 elements are arranged from lighter to heavier atomic weight. Observing the graph we could tell that lighter and heavier elements provide less information to the

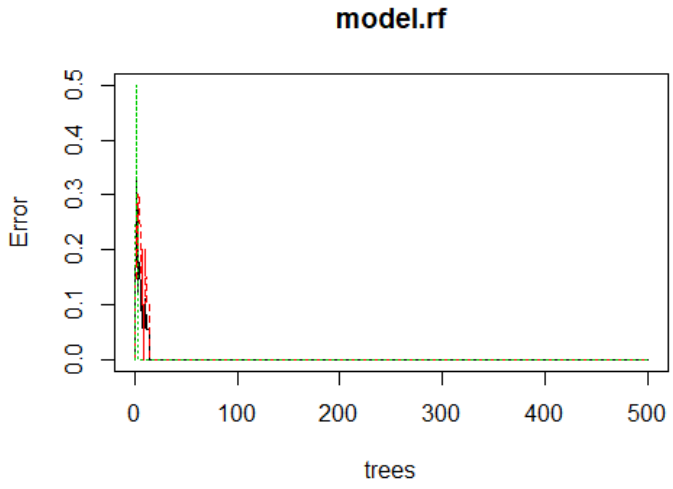


Figure 18: RF test error vs number of trees

model than "medium size" elements.

Obtained predictors from RF are aligned with PCA results, furthermore, results are concordant with literature of geographic determination by mineral fingerprint.

Concerning test error, as shown in Figure 18, it decreases with the number of averaged trees. With less than 50 averaged trees the RF test error is technically 0, which is a great strength of the model and the tree decorrelation.

7 Bootstrap

Cross-validation was applied in the present work for calculating the test error: the sample set was randomly divided into a training and a test set, then the model was checked with "blind-for-it" samples, and therefore the error could be calculated.

Another approach studied is what is called Bootstrap (or Bootstraping). Bootstrap is a resampling method with replacement: the initial dataset is randomly sampled to obtain a subset, then a function is performed in this subset. The next random extraction is performed again over the original dataset in order to obtain a new different subset. The process continues and system works over and over with subsets created every time from the whole initial data set. This technique is widely used for calculating errors or confidence intervals in data science.

Bootstrap is also used in order to increase the data size. It is a technique recommended for poor sample size sets that allows you, through inference of the original data set, to increase the volume of the data. However, it must be noted that the results may depend on the representativity of the dataset, and this could be the main inconvenient of Bootstrapping: if the initial dataset is not representative of the population, Bootstrapping application can generate a bias on the results. The name of this technique comes from the English expression "to lift oneself up by one's own bootstrap" and represents the impossible task of build information out of "nothing".

7.1 Bootstrap results and conclusions

Parametric Bootstrap was performed with R function *boot(boot)*. Resampling apply the "mean" function [9]. Applying Bootstrap to the global dataset was discarded. Based on the information known from the sampling, Vera samples have in common cultivation area, vegetal variety and preparation process, while NoVera samples are much more heterogeneous. Due to this fact, two groups (Vera - NoVera), with two assumed normal distributed population have been used for Bootstrap (although NoVera samples distribution

[9] see R code in the addendum for more details

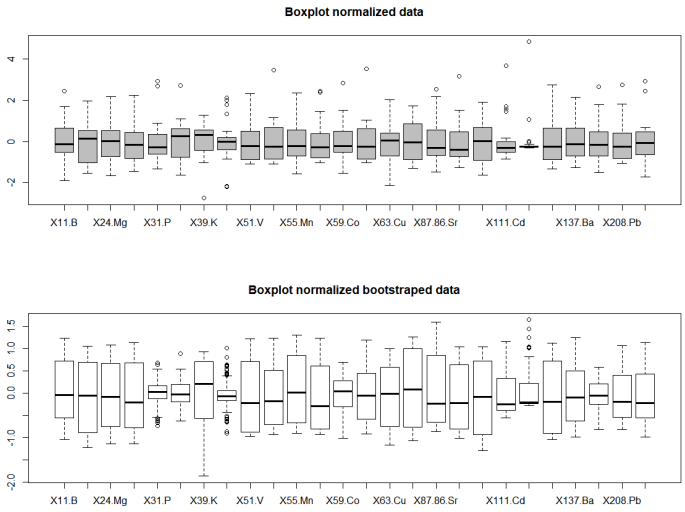


Figure 19: Boxplots comparision before and after Bootstrap

is actually unknown and probably samples from this group do not even belong to the same population). Bootstrap was applied in this work to check how results from normality test, PCA and RF applied in the original dataset would change with a bigger sample set (a desirable situation and hopefully achievable in the near future).

For the application of the Bootstrap the sample size was fixed in 200 samples, and even the Vera- NoVera ratio was 45:55 (approx.), the Bootstrapped dataset was Solomonically divided in 100 Vera and 100 No Vera samples in order to simplify calculations. 200 is the number of observations that has been found as a top range of number samples in scientific papers of food fingerprint and chemometrics analyses.

The new dataset is shown in Figure 19. As expected, variance has decreased due to the use of "mean" function.

Data from previous examples, liker iron data from Vera samples, can be plotted before and after the Bootstrapping was applied (Figure 20). The same comparison was done with iron from NoVera samples (Figure 21).

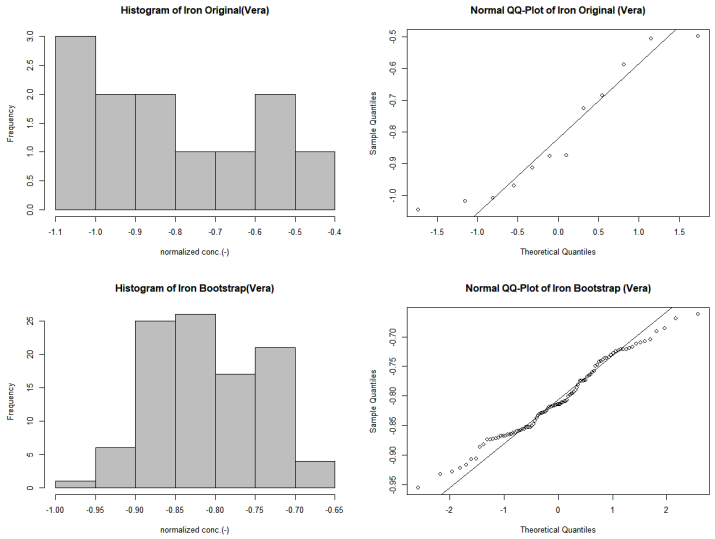


Figure 20: Iron of Vera samples before and after Bootstrap

From the obtained results and plots we can conclude that iron from NoVera samples were not normally distributed in one single population, probably two or more populations conform the entire NoVera set. These results are aligned with the fact that No Vera samples have differences in terms of origins.

7.1.1 Normality (Bootstrap)

The Bootstrapped dataset was checked for multivariate normality, which was not possible to do with the original data. Mardia and Shapiro-Wilk (univariate) tests were applied *mrv(MRV)*.

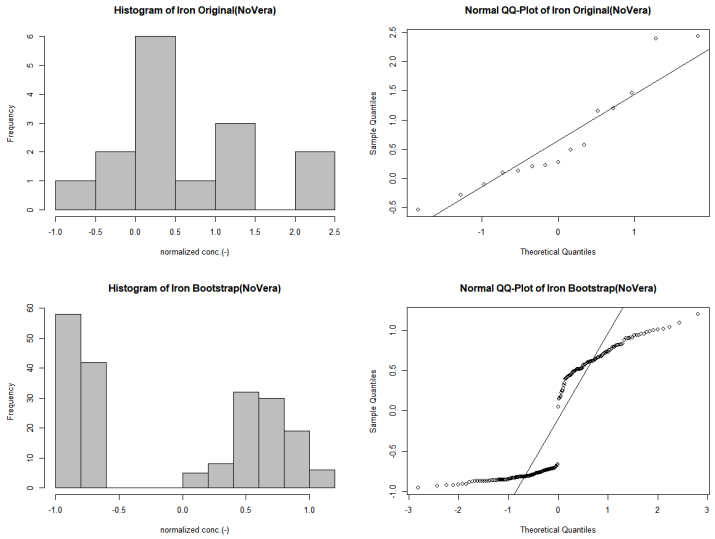


Figure 21: Iron of NoVera samples before and after Bootstrap

Multivariate normality was accepted in both cases.

7.1.2 PCA (Bootstrap)

PCA was performed with the new dataset and better results were obtained in terms of separation of the factor Vera-No Vera (Figure 22): the sampling size has impact on the PCA results. Moreover, bootstrapped dataset has less variance, and thus there is more distance between groups.

7.1.3 RF (Bootstrap)

Random Forest was applied using a training set of 120 samples and a test set of 80 samples. Once again, test and train error were cero (no false positives or false negatives were obtained).

The main difference observed in comparison with the RF applied on the initial dataset was the order of importance of the predictors (Mean Decrease Gini): the three principal predictors in this case are barium (Gini:4.66), 87/86 strontium (Gini:4.18) and sodium (4.05). Iron, which was the first

	Test	Statistic	p value	Result
1	Mardia Skewness	3179.95288670592	0.883000993981649	YES
2	Mardia Kurtosis	-1.9045360067442	0.0568404118809671	YES
3	MVN			YES

	Test	Variable	Statistic	p value	Normality
1	Shapiro-Wilk	11.B	0.9861	0.3795	YES
2	Shapiro-Wilk	23.Na	0.9892	0.5993	YES
3	Shapiro-Wilk	24.Mg	0.9820	0.1890	YES
4	Shapiro-Wilk	27.Al	0.9920	0.8233	YES
5	Shapiro-Wilk	31.P	0.9913	0.7655	YES
6	Shapiro-Wilk	32.S	0.9898	0.6511	YES
7	Shapiro-Wilk	39.K	0.9811	0.1611	YES
8	Shapiro-Wilk	44.Ca	0.9889	0.5781	YES
9	Shapiro-Wilk	51.V	0.9862	0.3888	YES
10	Shapiro-Wilk	52.Cr	0.9950	0.9761	YES
(...)					

Table 7: Vera Multivariate normality test results (Mardia) and Univariate (Shapiro-Wilk)-showing first 10 minerals

predictor in the initial analysis, after applying the Bootstrap, has become a medium size predictor (Gini:2.62).

The probable explanation is that initial dataset was biased for some elements, and Bootstrap application gains this bias. The best way to check this hypothesis would be enlarging the initial dataset with real samples, which is one of the main targets in the future.

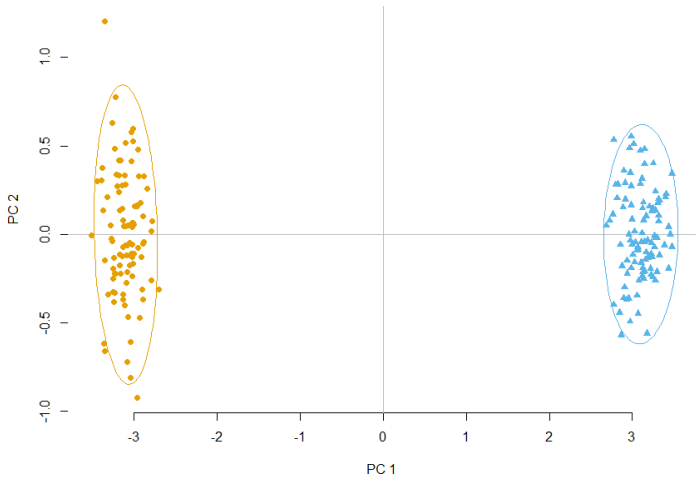


Figure 22: Blue triangles represent Vera samples, orange dots represent NoVera.

8 Final conclusions

8.1 Discussion: applied methods results

Individual results discussion for every model applied have been exposed in previous chapters. This section introduces a global analysis of the obtained results. The project pipeline can be divided into four steps:
- Data exploration, normalization and application of hypothesis tests for normality, variance and mean analysis .
- Unsupervised machine learning models application: PCA and CA.
- Supervised machine learning models application: LDA and RF.
- Bootstrap resampling and retest PCA and RF.

The four steps were successfully applied and consistent results were obtained in all cases. The conclusions concerning classification Vera-NoVera were comparable, regardless the used model. ANOVA test showed that differences in the mean value of Vera subset and NoVera subset were significant. Using three PC's the Principal Components Analysis showed a complete separation of both subsets with a confidence of 0.95. Cluster analysis with k=2 also obtained a complete separation of Vera-NoVera into two clusters.

After a random split of the dataset into training and test subsets, supervised machine learning methods were applied for classification. LDA obtained good results, although not in all randomly cases it was possible to obtain a 0% value of false positive and/or false negative errors. However, random forest has been proven to be the model that provides the best classification, with 0% mislabelling classification errors.

Predictor elements (Fe, V or 87/86Sr) obtained from the Random Forest application (Mean decrease Gini value) and PCA (main PC's coefficients) were in concordance, showing that from both techniques comparable results were obtained.

Resampling method of Bootstraping was used to enlarge the dataset. The obtained dataset provided the size needed to apply Multivariate normality tests, PCA and RF with a higher number of data. Bootstrap has simulated the usability of the proposed Rcode with a 200 observations dataset. However, Bootstraping has not been able to generate conclusive results for the NoVera subset, probably due to more than one population were conforming the subset.

Concerning RF, no issues related to the time of computing have been detected, using a regular personal computer. That shows that the algorithm proposed is capable to manage real datasets.

8.2 Final conlusions and future work

The objectives described at the beginning of the project have been achieved. A multivariate method for fingerprint analysis of paprika samples has been developed, showing that selected elements are able to provide the analytical information needed for further data analysis.
Proposed unsupervised machine learning methods have been able to achieve the objective of classifying the samples into Vera-NoVera groups. Proposed supervised methods have been trained and tested, providing good results in both cases (LDA and RF). For the first time, Random Forest technique has been used to classificate paprika samples based on a P.D.O. description. The method performance has obtained 0% of error type I and type II, proving

that it is a useful technique in this area of food fraud detection.
Validation of the models has been performed with cross-validation.
The designed algorithm (Rcode) is ready for further analyses, including expanding current dataset or processing new datasets from others P.D.O.'s.
Additional personal skills have been also developed, like using LaTex editor and JabRef for bibliography managing.

Future work This project has proven that coupling mineral fingerprint acquisition with ICP-MS together with data analysis with Random Forest supervised machine learning technique is a successful method for paprika with P.D.O. classification. However, the present work can be extended in different ways: first thing, sample size should be enlarged.

Sampling is a crucial step regarding the method ability to detect Vera mislabelling products (food fraud related with DPO labelling). The group of NoVera samples is heterogeneous and thus, expert opinion from P.D.O. regulatory council and paprika producers, plays a key role to ensure the best possible approach. Binary classification techniques success (specially in the case where only one of the two groups is located in one specific location) is based on a proper sampling process.

Future work is also related to variables selection. Despite the fact that the dataset can be enlarged in the future, variables under study can also be reduced. The models have indicated that some minerals have low prediction levels for classification purposes, probably due to collinearity of the variables, and/or same mineral content in both groups Vera-NoVera. A better selection of variables can be done in the future in order to optimize the method performance.

Variables selection can also include other food components, involving different analytical techniques, like amino acids or fatty acid profile. The quality of the product fingerprint can be expanded with a more holistic approach.

Furthermore, supervised techniques can be extended. Even RF has been proven as a powerful method, other supervised models can be applied: SIMCA, SVM, etc.

Data analysis was based on one variable factor (Vera-NoVera), however, other factors can be included in the method scope (hot/sweet/bitter paprika, harvest year and fumed/not fumed paprika).

9 Addendum I: Dataset

	vera	X11.B	X23.Na	X24.Mg	X27.Al	X31.P	X32.S
1	1	35306.93	315708.74	5510893.34	395165.33	10506511.10	309419.58
2	0	21764.12	909690.43	7178715.61	898881.57	12778856.04	262841.75
3	0	21163.09	949565.85	7346523.88	905411.07	13066797.81	319880.03
4	1	32823.16	598800.01	4999496.65	384500.58	9782071.68	247563.64
5	1	38335.73	354324.15	5752216.70	507791.36	10585140.36	432825.87
6	1	34660.70	476751.36	5324170.98	797529.69	10360780.28	248019.32
7	0	30458.30	1151927.41	5952460.40	695671.49	10515416.51	335205.89
8	0	28873.51	1425866.73	5900383.04	948662.30	8872081.61	270037.92
9	1	34121.26	376228.90	5695737.00	582128.35	10068786.40	376686.49
10	0	30608.57	861183.81	6425203.09	858742.82	9693111.58	305433.08
11	1	42667.30	883795.82	5081970.64	391689.02	8771388.72	460466.86
12	1	28772.36	345297.04	4595474.12	301953.20	9016525.14	461260.84
13	0	27343.36	839246.64	5645044.86	800994.18	8526113.08	339299.54
14	0	28015.06	708514.28	5691832.14	1150870.30	9159517.76	478946.04
15	1	30872.90	393290.90	4614068.34	654923.04	9133780.50	400480.78
16	1	39017.40	791608.32	5039829.04	673577.00	9461618.06	483441.72
17	0	29966.52	1092508.26	5727815.66	1170529.44	7851135.38	421724.94
18	1	34287.02	306437.44	4973955.92	445736.20	9282323.88	441491.42
19	0	29647.48	1215901.50	5151240.22	479541.26	8228835.68	648752.66
20	0	27965.44	1077998.44	6345362.96	1540074.38	8704986.50	392198.34
21	0	26441.72	892286.58	5497194.68	707207.30	8373719.34	383341.06
22	1	30314.86	243970.28	4489568.24	422499.90	9046480.10	476485.60
23	0	29888.18	855588.02	6342480.54	1300396.54	8685186.10	426636.22
24	0	33552.08	1307024.60	6953305.62	1533610.20	8462668.08	499029.38
25	1	28570.06	205401.42	4328138.48	244513.64	8826574.38	449704.22
26	0	23911.52	683819.36	6138210.82	859695.18	9443194.42	424500.50
27	0	26505.56	782147.26	5498802.66	615812.58	8920924.04	439547.90

	X39.K	X44.Ca	X51.V	X52.Cr	X55.Mn	X56.Fe	X59.Co
1	86633630.41	4863045.79	388.23	929.83	72578.27	328019.93	722.55
2	40438442.68	3117754.78	1383.95	3410.70	50620.13	710801.38	477.89
3	39540822.91	3114072.88	1409.86	3283.58	51693.72	728834.78	486.37
4	83606224.73	3912475.29	431.88	562.63	74960.94	328924.93	493.43
5	85542041.51	6153482.24	430.33	1274.09	57133.45	295584.46	259.12
6	84469973.56	4995908.79	634.10	1166.71	81039.51	457375.59	374.25
7	68186192.77	4787818.61	1108.92	4522.91	45747.31	598752.56	626.06
8	72269245.19	5015690.92	1703.20	6939.58	41291.10	805480.96	953.26
9	91364816.10	5031548.51	611.16	4092.67	69403.48	459781.67	457.63
10	69572646.71	5232722.65	1852.42	6619.90	49722.45	833536.66	589.82
11	72708038.86	5508745.34	381.66	463.16	96384.42	315391.62	708.08
12	70619941.66	3491628.52	366.90	1587.38	66231.86	278432.78	265.58
13	62371951.20	4573688.22	1506.28	6676.76	40196.46	703444.26	543.28
14	51230452.54	8048215.74	2296.20	6636.98	55483.78	1051728.10	716.52
15	72076312.64	4378139.50	541.32	1250.98	73091.70	380290.62	323.74
16	77428500.52	5231963.58	616.94	959.28	97918.08	428255.80	497.56
17	52013164.94	6095513.40	2038.64	16327.64	58483.38	1037323.12	969.86
18	71288116.98	5223225.08	380.36	1285.20	54643.04	269001.84	223.82
19	64075399.98	4061236.16	920.58	2307.64	30733.16	446299.84	545.14
20	48952326.56	9815869.58	3129.16	8071.40	68854.12	1483413.58	855.60
21	58640982.46	5388900.06	1289.28	5580.96	40681.86	678035.10	532.86
22	73333143.82	4581945.54	340.68	5227.36	83952.06	394742.44	679.62
23	52879800.84	9596690.22	2385.32	7379.48	55576.94	1145664.26	794.38
24	57706342.90	9106261.60	2999.40	8251.16	57131.00	1468769.90	841.34
25	76797422.48	4224038.20	272.92	2975.96	108182.33	282157.50	1296.85
26	478414.74	65521.92	1421.91	3290.21	45141.67	667266.88	462.90
27	619820.46	42665.66	984.62	3638.71	36925.25	533873.19	501.00

	X60.Ni	X63.Cu	X66.Zn	X87.86.Sr	X88.Sr	X95.Mo	X111.Cd
1	2034.34	22558.36	83063.97	6.65	558.86	601.48	102.53
2	1806.88	28223.80	57182.22	3.97	1024.41	1050.20	45.73
3	1849.64	29491.43	58982.07	3.68	976.22	1081.28	45.93
4	1385.79	23649.52	72947.13	8.09	666.13	516.16	96.13
5	1277.07	21009.90	80468.89	5.83	958.80	574.55	59.48
6	1334.28	20926.17	68540.30	4.58	747.26	713.94	64.73
7	2693.65	24371.70	47045.97	3.96	1528.18	928.87	122.58
8	4015.31	23291.73	39669.70	3.16	1713.95	1167.78	67.84
9	2141.34	23362.77	77667.56	5.20	1009.21	858.28	69.77
10	4066.17	29022.18	47701.42	3.56	1949.26	1079.74	56.92
11	1118.84	16413.00	64393.68	3.20	991.70	546.42	303.38
12	1351.02	18906.36	70953.96	5.95	688.20	905.36	122.62
13	3417.38	24843.16	38988.48	4.33	1574.40	819.74	108.28
14	3453.04	24227.36	49052.16	3.61	1616.10	1320.30	155.74
15	1273.94	18997.88	59957.46	4.91	617.60	632.16	143.68
16	1501.62	20486.14	84110.28	4.44	861.50	396.50	293.42
17	8354.92	23236.20	50914.68	4.11	1414.90	1213.70	132.70
18	1133.10	19062.30	70763.70	5.76	892.30	508.50	136.82
19	2370.26	22678.74	44143.14	3.33	1529.00	852.28	98.86
20	4405.18	26745.94	45543.12	2.76	2088.10	1309.04	120.28
21	3385.96	23724.06	37812.60	3.62	1553.10	1128.18	99.28
22	3689.60	19538.98	71885.42	7.14	670.30	751.08	315.98
23	3786.98	24500.64	44242.62	3.11	2249.20	1478.82	122.02
24	4355.90	23098.46	40918.90	2.16	3303.90	1295.10	96.18
25	4133.34	23896.78	67799.80	8.71	470.40	433.74	529.48
26	2239.16	25504.59	44538.27	3.16	1597.60	988.88	89.88
27	2364.94	22627.44	39739.13	4.05	1309.00	1045.19	96.35

	X118.Sn	X121.Sb	X137.Ba	X153.Eu	X208.Pb	X238.U
1	46.77	17.27	7829.40	0.34	108.99	3.25
2	60.84	27.55	10031.64	1.38	171.20	3.84
3	201.66	24.17	10672.02	1.64	168.73	6.13
4	3981.65	9.04	6501.14	0.00	116.82	1.87
5	122.04	12.65	7100.88	0.33	141.98	3.59
6	158.92	10.06	11662.09	1.35	184.79	4.84
7	327.95	32.29	10516.69	2.36	298.21	2.81
8	104.39	35.37	13516.16	2.70	221.04	5.23
9	87.77	12.95	13865.11	1.71	147.33	3.54
10	270.77	41.53	14120.36	2.78	287.64	6.08
11	14789.12	9.24	6658.12	0.69	304.36	4.02
12	942.02	4.60	6475.98	1.24	200.96	2.99
13	410.00	30.28	11270.72	0.28	509.64	6.06
14	252.94	34.04	15322.56	1.65	866.78	6.07
15	130.28	9.58	9078.86	2.92	430.78	4.73
16	129.96	18.50	8792.58	1.35	420.72	5.69
17	176.42	42.48	17837.10	1.87	1042.86	6.54
18	141.74	7.30	6004.14	3.60	284.44	3.68
19	67.96	20.92	6556.28	0.73	468.50	4.32
20	110.08	65.78	19558.06	1.48	1346.80	11.71
21	539.86	34.50	10750.08	4.56	582.96	5.60
22	90.02	10.78	9022.08	0.96	524.16	5.91
23	85.52	41.92	17060.76	0.79	937.72	6.22
24	150.20	44.42	18556.32	1.59	931.88	10.58
25	153.76	11.71	9403.71	3.16	584.78	1.07
26	814.46	32.29	12365.84	1.01	622.74	5.81
27	54.82	20.79	8790.44	1.97	368.55	3.05

10 Addendum II: Pipeline and R Code

(see attached document)

References

Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of chromatography A*, *1158*(1-2), 196–214.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Callao, M. P. (2014). Multivariate experimental design in environmental analysis. *TrAC Trends in Analytical Chemistry*, *62*, 86–92.

Dehelean, A., & Voica, C. (2012). Determination of lead and strontium isotope ratios in wines by inductively coupled plasma mass spectrometry. *Romanian Journal of Physics*, *57*(7-8), 1194–1203.

D.O.P.Vera. (2006). (Vol. 17) (No. 510).

Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with r*. Springer Science & Business Media.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Irizarry, R. A., & Love, M. I. (2016). *Data analysis for the life sciences with r*. CRC Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Lantz, B. (2013). *Machine learning with r*. Packt Publishing Ltd.

Lim, D. K., Long, N. P., Mo, C., Dong, Z., Lim, J., & Kwon, S. W. (2018). Optimized mass spectrometry-based metabolite extraction and analysis for the geographical discrimination of white rice (oryza sativa l.): a method comparison study. *Journal of AOAC International*, *101*(2), 498–506.

López, M. I., Trullols, E., Callao, M. P., & Ruisánchez, I. (2014). Multivariate screening in food adulteration: untargeted versus targeted modelling. *Food chemistry*, *147*, 177–181.

M. Forina, M. C., & Oliveri, P. (2009). *Application of chemometrics to food chemistry (chapter in book)* (E. B.V., Ed.).

Ördög, A., Štajner, D., Popović, B., Poór, P., Bátori, Z., & Tari, I. (2018). Comparison of the mineral content of processed spice samples of sweet and hot paprika from the szeged region. *Journal of elementology*, *23*(2), 521–530.

Palacios Morillo, A. (2015). *Caracterización analítica y diferenciación geográfica de pimentón mediante técnicas de reconocimiento de patrones*

(Unpublished doctoral dissertation).

Palacios-Morillo, A., Jurado, J. M., Alcázar, Á., & de Pablos, F. (2014). Geographical characterization of spanish pdo paprika by multivariate analysis of multielemental content. *Talanta*, *128*, 15–22.

Picó, Y. (2015). *Advanced mass spectrometry for food safety and quality*. Elsevier.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

RASFF. (2019). *2018 annual report* (Tech. Rep.). The Rapid Alert System for Food and Feed.

Tan, M., Fang, H.-B., Tian, G.-L., & Wei, G. (2005). Testing multivariate normality in incomplete data of small sample size. *Journal of Multivariate Analysis*, *93*(1), 164–179.

Team, R. C., et al. (2013). R: A language and environment for statistical computing.

Tokalıoğlu, Ş., Çiçek, B., İnanç, N., Zararsız, G., & Öztürk, A. (2018). Multivariate statistical analysis of data and icp-ms determination of heavy metals in different brands of spices consumed in kayseri, turkey. *Food analytical methods*, *11*(9), 2407–2418.

Zhao, S., Cao, S., Luo, L., Zhang, Z., Yuan, G., Zhang, Y., ... others (2018). A preliminary investigation of metal element profiles in the serum of patients with bloodstream infections using inductively-coupled plasma mass spectrometry (icp-ms). *Clinica Chimica Acta*, *485*, 323–332.