Combining isotopic signatures of $n(^{87}Sr)/n(^{86}Sr)$ and light stable elements (C, N, O, S) with multi-elemental profiling for the authentication of provenance of European cereal samples

[3]Daniel GOITOM ASFAHA, [3]Christophe R. QUETEL*, [4]Freddy THOMAS, [1]Micha HORACEK, [1]Bernhard WIMMER, [1]Gerhard HEISS, [2]Christian DEKANT, [2]Peter DETERS-ITZELSBERGER, [2]Stefan HOELZL, [2]Susanne RUMMEL, [3]Christophe BRACH-PAPA, [3]Marleen VAN BOCXSTAELE, [4]Eric JAMIN, [5]Malcolm BAXTER, [5]Katharina HEINRICH, [6]Daniela BERTOLDI, [6]Luana BONTEMPO, [6]Federica CAMIN, [6]Roberto LARCHER, [6]Matteo PERINI, [5,7]Simon KELLY, [8]Andreas ROSSMANN, [9]Antje SCHELLENBERG, [9]Claus SCHLICHT, [10]Heinz FROESCHL, [11]Jurian HOOGEWERFF, [11]Henriette UECKERMANN

*corresponding author:
Christophe Quétel
Institute for Reference Materials and Measurements (Joint Research Centre- European Commission).
111 Retieseweg. B-2440 Geel, Belgium.
e-mail: Christophe.Quetel@ec.europa.eu
Tel.: +32-14-57-1658
Fax: +32-14-57-1863

[1] AIT (Austrian Institute of Technology), Seibersdorf, Austria
[2] BSPG (Bayerische Staatssammlung für Paläontologie und Geologie), München, Germany
[3] EC-JRC-IRMM (European Commission – Joint Research Centre – Institute for Reference Materials and Measurements), Geel, Belgium
[4] Eurofins Scientific Analytics, Nantes, France
[5] FERA (Food and Environment Research Agency), York, United Kingdom
[6] IASMA (Fondazione E. Mach – Istituto Agrario di San Michele all'Adige) San Michele all'Adige, Italy
[7] IFR (Institute of Food Research), Norwich, United Kingdom
[8] IGmbH (Isolab GmbH), Schweitenkirchen, Germany
[9] LGL (Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit), Oberschleißheim, Germany
[10] Seibersdorf Labor GmbH, Seibersdorf, Austria
[11] UEA (University of East Anglia), Norwich, United Kingdom

Keywords: cereal; authentication of geographical origin; isotopic tracers; multi-elemental profiling

Abbreviations: CRM - Certified Reference Materials; **IRMS** - Isotope ratio mass spectrometry; **MC-TIMS** - Multi Collector Thermal Ionisation Mass Spectrometer; **MC-ICPMS** - Multi Collector Inductively Coupled Plasma Mass Spectrometer; **PCA** - Principal Component Analysis; **PLS-DA** - Partial Least Square Discriminant Analysis

Abstract

The aim of this work (from the FP6 project TRACE) was to develop methods based on the use of geochemical markers for the authentication of the geographical origin of cereal samples in Europe (cf. EC regulations 2081/92 and 1898/06). For the first time the potential usefulness of combining $n(^{87}Sr)/n(^{86}Sr)$ and $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{34}S$ isotopic signatures, alone or with key element concentrations ([Na], [K], [Ca], [Cu] and [Rb], progressively identified out of 31 sets of results), was investigated through multiple step multivariate statistics for more than 500 cereal samples collected over 2 years from 17 sampling sites across Europe representing an extensive range of geographical and environmental characteristics.

Both models compared involved three sample classification categories (north/south; proximity to the Atlantic Ocean/to the Mediterranean Sea/to else; bed rock geologies). The first two categorisations were the most efficient, particularly when using the ten variables selected together (with, in some instances, element concentrations making a greater impact than the isotopic tracers). Validation of models included external prediction tests on 20% of the data randomly selected and, rarely done, a study on the robustness of these multivariate data treatments to uncertainties on measurement results. With the models tested it was possible to individualise 15 of the sampling sites.

I. Introduction

In Europe, guaranteeing the geographical origin of a food product (cf. EC regulations 2081/92 and 1898/06) is regarded as an assurance of quality and safety. The FP6 project TRACE funded by the European Commission was launched in 2005, aiming to deliver tools "that will enhance consumer confidence in the authenticity of food" (FP6-TRACE project-website, 9th April 2010). An important component of this project was to study how geochemical markers and the relationships between these markers could be used for the characterisation of the region of origin of food products. For instance, it was shown that variations in H, C, N and S isotopic composition and element concentrations are useful for the differentiation between samples coming from some different regions in Europe in the case of lamb meat, honey and olive oils (Camin et al., 2007, 2010; Schellenberg et al., 2010). The present study also reports research findings from the TRACE project but on a broader scope level and with regards to wheat. For the first time the potential usefulness of C, N, O, S and Sr isotopic signatures combined to multi-element profiles was investigated for over 500 samples originating from sites in Europe representing an extensive range of geographical and environmental characteristics (climate, distance to sea/ocean, geology).

The $\delta^{13}C$ values of plant compounds depend on photosynthetic pathways (discrimination between C3 and C4 plants), the plant age and level of maturation (Farquhar et al., 2003; Smith and Epstein, 1971). $\delta^{13}C$ is also affected, although to a lesser extent, by several environmental factors such as relative humidity, temperature, amount of precipitation and water stress (O'Leary, 1995). $\delta^{15}N$ values depend on the botanical type of the plant and on the bacterial activity associated to its growth (Farquhar et al., 2003), and can also be affected by temperature and agricultural practices (e.g. type of fertilizers used) (Bateman et al., 2005; Martin and Martin, 2003). The $\delta^{18}O$ of plant material reflects the water taken up (linked to temperature, latitude, elevation, distance from the sea and amount of precipitation) and exchanged (evaporative and diffusion effects during transpiration), and also biosynthetic pathways including the isotopic exchange between organic molecules and plant water (Bréas et al., 1998). Variations in the $\delta^{34}S$ ratio in plants have been reported to be linked with the soil geology (Thode, 1991), the distance from the sea and proximity to industrial activities (Camin et al., 2007). The $n(^{87}Sr)/n(^{86}Sr)$ ratio in soils may vary depending on the [Rb]/[Sr] concentration ratio in (and age of) the surrounding rock or mineral and, thus, has the potential to reflect the nature of the underground geology (Capo et al., 1998). It is expected that the $n(^{87}Sr)/n(^{86}Sr)$ in plants should also be a good indicator of the rocks and soil conditions during plant growth. Similarly, it is expected that element concentrations in food commodities such as olive oils (Camin et al., 2010), potatoes (Anderson et. al., 1999), garlic (Smith, 2005), orange juice (Simpkins et al. 2000) and coffee (Anderson et. al., 2002) are mainly related to the geological and

pedoclimatic characteristics of the site of growth and plant farming practices. A discussion on reasons why the elemental composition in plant may provide unique markers of the geographical origin can be found in Kelly et. al. (2005).

The way these tracers could be used to enable the authentication of provenance of cereal samples was investigated through multiple steps multivariate statistical analysis. The method validation scheme we designed included a study on the robustness of these multivariate data treatments to uncertainties in measurement results.

II. Experimental

The measurands in cereal samples are hereby referred to as *variables*. They consist of Sr isotope ratios (noted $n(^{87}Sr)/n(^{86}Sr)$) and 31 element concentrations (noted [E]) in the bulk, as well as δ-scale values for the CNOS group (noted $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{34}S$) in the defatted fraction.

$\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{34}S$ measurements were carried out in seven laboratories (Eurofins Scientific Analytics, Nantes, France; FERA, York, UK; AIT, Seibersdorf, Austria; LGL, Oberschleißheim, Germany; Isolab, Schweitenkirchen, Germany; IASMA, San Michele all'Adige, Italy; IFR, Norwich, UK). $n(^{87}Sr)/n(^{86}Sr)$ measurements were carried out in four laboratories (EC-JRC-IRMM, Geel, Belgium; FERA; BSPG, München, Germany; UEA, Norwich, UK). Element concentration measurements were carried out in four laboratories (Seibersdorf Labor GmbH, Seibersdorf, Austria; FERA; IASMA; UEA). The scientists responsible for producing these experimental data are co-authoring this paper.

a) Sampling strategy and sample collection

Cereal samples were collected from multiple farms or local producers in 19 sampling sites in Europe between summer 2005 and summer 2007. The names and average geographical coordinates of these sampling sites are provided in Table 1.

Although wheat (of different varieties, including winter wheat, Durum wheat, Emmer wheat, Epeautre and Spring wheat) was the main target, barley, rye, triticale, oat (all C3 plants) and four corn (C4 plant) samples were also collected from areas where wheat was not produced. The sampling plan was to collect four samples from five different producers or fields within each of the predefined sampling sites per year, and this was repeated for a second year around the same time. The ideal scheme of 40 cereal samples per sampling site could not always be strictly followed, due to practical limitations, and a total of 557 cereal samples were obtained, as shown in Table 1. Algarve and Barcelona samples (incl. the four corns) were excluded for this study because they

were purchased on markets and their claimed origin could not be verified. Samples from Galicia were collected for one season only, and only three samples could be obtained from Iceland.

b) Sample preparation, spectrometric measurements and validation of measurement methods

i) For $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{34}S$ results

$\delta^{13}C$, $\delta^{15}N$ and $\delta^{18}O$ were measured using a range of Isotope Ratio Mass Spectrometers (Delta Plus XL, Delta Plus XP, Delta V, Delta S, ThermoFinnigan, Bremen, Germany; Isoprime, AP2003, Optima, GV Instruments Ltd., Manchester, UK), IRMS. They were connected to pyrolisers (TC/EA, ThermoFinnigan; PyrOH and EA3000, Eurovector, GV Instruments Ltd.) for $^{18}O/^{16}O$ or an elemental analysers (Flash EA 1112, 1110, 1108 ThermoFinnigan; Costech ECS4010; NA2100 Proteins, Carloerba, Milan, Italy; Vario EL III, Elementar Analysensysteme GmbH, Hanau, Germany) for $^{13}C/^{12}C$ and $^{15}N/^{14}N$. Samples contained very low amounts of sulphur (usually less than 0.1 % according to Sieper et al., 2006) and $\delta^{34}S$ values were obtained from only one instrument (Vario EL III, Elementar Analysensysteme GmbH). IRMS instrumental conditions for each isotope ratio were reported by the same authors in a previous paper (Camin et al., 2007).

The isotopic values were expressed in δ‰ versus V-PDB (Vienna - Pee Dee Belemnite) for $\delta^{13}C$, versus AIR for $\delta^{15}N$, versus V-SMOW (Vienna – Standard Mean Ocean Water) for $\delta^{18}O$, and versus V-CDT (Vienna Canyon-Diablo-Troilite) for $\delta^{34}S$ according to the following formula: $[(R_s - R_{std})/R_{std}] \times 1000$, where R_s is the isotope ratio measured for the sample and R_{std} is the isotope ratio of the international standard. The values were calculated against in-house standards, which were themselves calibrated against international reference materials: fuel oil NBS-22 (IAEA) and sugar IAEA-CH-6 (IAEA) for $^{13}C/^{12}C$, USGS-40 and IAEA-N-1 (IAEA) for $^{15}N/^{14}N$, Benzoic Acid IAEA-601 (IAEA) and IAEA-CH-6 (IAEA) for $^{18}O/^{16}O$ and IAEA-S-1 for $^{34}S/^{32}S$.

The measurement repeatability was ≤ 0.2‰ for $\delta^{13}C$, ≤ 0.5‰ for $\delta^{18}O$, ≤ 0.1 ‰ for $\delta^{15}N$ and ≤ 0.3‰ for $\delta^{34}S$. An inter comparison was organized between the seven laboratories in charge of IRMS measurements for validation purposes. The test material for the comparison (wheat flour) was also used periodically in each laboratory as quality control, to monitor deviations over time. The dispersions of results between laboratories during the comparison were 0.2‰ for $\delta^{13}C$, 0.9‰ for $\delta^{18}O$, 0.8‰ for $\delta^{15}N$ and 1.2‰ for $\delta^{34}S$.

ii) For $n(^{87}Sr)/n(^{86}Sr)$ results

The instrumentation and laboratories involved in the Sr isotopic measurements were the same as those of the study on mineral water samples published recently (Voerkelius et al., 2010). There were one Multi Collector Thermal Ionisation Mass Spectrometer (MC-TIMS – MAT 261/262, Thermo Finnigan) and three Multi Collector Inductively Coupled Plasma Mass Spectrometers (MC-

ICPMS – Nu Plasma, Nu Instruments; Axiom, ex- VG Instruments; Isoprobe, ex-Micromass). MC-TIMS measurements were carried out on tungsten single filaments. For MC-ICPMS measurements, typically, the instrumentation was equipped with a minicyclonic jacketed cinnabar spray chamber or a porous membrane based desolvation unit, a 0.2 ml min^{-1} microconcentric nebuliser and Ni or Pt sampler and skimmer cones.

Sample uptake was 0.1 - 0.3 g for MC-TIMS measurements, and 0.4 - 1 g for MC-ICPMS measurements. Samples were dry-ashed for MC-TIMS measurements, or acid digested in a microwave oven for MC-ICPMS measurements. Residuals were dissolved in concentrated HNO_3 to allow separation of Sr from other elements such as Ca, Ba or Rb by ion chromatography on a Sr specific crown ether resin (Sr-spec®). After separation the Sr concentration in solution was 10-500 ng g^{-1} in about 2.5 g of 2-3% HNO_3.

The same standard operating procedure (SOP, cf. details in report of FP6-TRACE Deliverable D15.9, 2009) applied to all laboratories to allow for comparison of measurement results. This SOP included recommendations on Rb/Sr separations, corrections for procedural blanks, the prescription of the $n(^{86}Sr)/n(^{88}Sr)$ and $n(^{87}Rb)/n(^{85}Rb)$ reference values used for corrections (0.1194 ± 0 and 0.38565 ± 0.00030, respectively). It also required running regularly (every 1 to 4 samples maximum) experiments on the NIST-987 Sr isotopic CRM for the purpose of contributing to the validation of each methods developed. The relative dispersion of results between laboratories for identical cereal test materials during four comparisons was up to approximately 0.9‰ depending on physical (flour/coarse grains, grains with or without husk etc.) and chemical ([Rb] and [Sr]/[Rb] concentration ratio) characteristics of samples.

iii) For [E] results

Measurements of element concentrations were performed on four quadrupole ICP-MS instruments (two Agilent 7500ce, Tokyo, Japan; one Elan 6100 and one Elan 6000, Perkin Elmer Sciex, Toronto, Canada). The first two were fitted with concentric Micromist nebulisers (Glass Expansion, Melbourne, Australia) and water-cooled Scott double-pass spray chambers, the third with a cross-flow nebuliser and the fourth with a concentric nebuliser and a cyclonic water-cooled spray-chamber. In all cases peristaltic pumps were used to regulate sample flow rates. RF power settings ranged from 1100 to 1500W on the four instruments. Pt cones were used on the Elan 6000, while Ni cones were used on the other instruments.

Microwave digestion was employed by three of the laboratories, digesting between 0.3 and 0.5 g cereal with distilled nitric acid, either with or without addition of ultrapure hydrochloric acid. The fourth laboratory used ultraviolet digestion of 0.5 g sample material with 4 ml nitric acid and 1 ml

192 distilled hydrochloric acid. All labware involved was cleaned beforehand with solutions of between
193 5% and 10% nitric acid. Results were corrected for cereal moisture content.
194 NIST1567a Wheat Flour was used for method validation purposes and measured with all sample
195 batches. There was agreement within uncertainty boundaries with certified values for all elements
196 of interest in this study.
197
198 ***c) Data set description and the multivariate data analysis tools applied***
199 In our study the $n(^{87}Sr)/n(^{86}Sr)$ and the CNOS isotopic signatures were treated alone and in
200 combination with [E] results. In the first case it was necessary to exclude an additional 13 samples
201 (of 32) from Jura Krakowska and 2 samples (of 9) from Jylland for which four out of five of the
202 isotopic data were not reported. Thus, the overall cereal data set consisted of 512 samples from 17
203 sampling sites. When isotopic and [E] results were combined all 527 samples could be considered
204 (e.g. including the 15 samples excluded in the previous case).
205 Prior to the application of multivariate statistical calculations some data pre-treatment was required
206 (Eriksson et al., 2001; Esbensen, 2006). For each variable unit variance scaling and mean-centering
207 of data was applied (division by the standard deviation followed by subtraction of the mean value).
208 In addition, [Na] and [Rb] data required log-transformation in order to normalise the distributions.
209 In this study *Principal Component Analysis* (PCA) and *Partial Least Square Discriminant Analysis*
210 (PLS-DA) modelling tools were used. PCA provides a more comprehensive overview of all data by
211 producing a few orthogonal (uncorrelated) *principal components* (PCs) which extract the main
212 information about the data set (Eriksson et al., 2001; Esbensen, 2006). PCA is normally applied at
213 early stage of multivariate data analysis as "exploratory tool". PLS-DA modelling is more suitable
214 for smaller numbers of defined classes and maximises the separation between them.
215 PCA and PLS-DA offer a number of useful parameters and diagnostic tools expressed graphically
216 and numerically. These include PC-score plots, R2X(cum), R2Y(cum), Q2Y(cum) and VIP
217 (*Variable Importance in the Projection*). The R2 parameters are a quantitative measure of the
218 "*goodness of fit*" of a given model (R2X for the "predictors", and R2Y for the "responses"),
219 whereas the Q2 parameters indicate the "*goodness of prediction*" (predictive ability) of a given
220 model. There exist some rules to help identifying the best balance between the predictive power and
221 a reasonable level of fitness of the model. According to Eriksson et al. (2001) "generally, a Q2 >
222 0.5 is regarded as good and a Q2 > 0.9 as excellent, but these guidelines are of course heavily
223 application dependent", and "differences between R2 and Q2 larger than 0.2-0.3 indicate the
224 presence of many irrelevant model terms or few outlying data points". For PLS-DA the relative
225 importance of variables is illustrated by VIP plots. Variables with a VIP value greater than unity

play the most important roles for the discrimination of the classes. For more elaborated explanations and mathematical expressions related to these diagnostic tools and parameters readers are referred to Eriksson et al. (2001).

In this work, all the multivariate computations were carried out using SIMCA-P Ver. 12 (Umetrics AB, Sweden) software. The Statistica Ver. 7 (Statsoft Inc., USA) software was used to produce the box-whisker plots for the variables.

Our validation scheme included *external prediction* tests on 20% of the data assigned as independent data sets (randomly selected). These data were compared (on class membership) against prediction results obtained from PLS-DA models built with the remaining 80% of the data set. Models were considered satisfactory when external prediction rates were $\geq$ 70%.

III Results and discussion

To have an overview of all 527 samples from 17 sites with 36 descriptor variables, a PCA was applied to the data set (see Figure ES1 provided as electronic supplementary material).

There were no clear trends or patterns to distinguish samples of different sampling sites. Thus we decided to apply PLS-DA, based on supervised classification, and to compare two models, using the five isotopic variables alone and in combination with [E] results.

The classification categories we chose (Table 1) were according to *latitude* (two classes: north and south of Europe), proximity *to a marine environment* (three classes: Atlantic, Mediterranean and Inland) and *geology of the underlying bed rock* (four classes: Shale, Acid magmatic, Limestone and Basaltic). The way we established these classification categories and the results obtained are discussed in the subsequent sections.

The five isotopic variables ($n(^{87}Sr)/n(^{86}Sr)$, $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{34}S$) were systematically considered as they were found alternately significant depending on the classes considered for the data, as shown later. Preliminary investigations (successive PLS-DA and VIP plots) were carried out to evaluate the significance of the 31 element concentration variables available. From the first list of nine apparently influential concentration variables identified, four ([Co], [Ga], [Cd] and [Cs]) were further eliminated due to low concentration in the majority of the samples and rather poor analytical figures of merit (measurement reproducibility amongst the laboratories for cereal samples was, from the TRACE quality assurance report (2009), $\leq$ 20% for [Cd], 20% to 50% for [Co] and over 50% for [Ga] and [Cs]). That left only [Na], [K], [Ca], [Cu] and [Rb]. Hereafter, the two methods of modelling will be referred to as the "*5-variable*" and "*10-variable*" (the 5 isotopic markers and [Na], [K], [Ca], [Cu] and [Rb]) models, respectively.

a) PLS discriminant analysis models

i) Classification according to latitude (north and south classes)

The motivation was the assumption of a relationship between climatic conditions and the latitudinal geographic position of each site where the cereal samples were harvested. After several model refining attempts, the optimal boundary line for the discrimination between north and south samples was found to be around latitude line 47° N through "central" Europe. In both cases PLS-DA computations with this classification resulted in two optimal significant PCs, and relatively high values of *goodness of prediction* Q2(cum) (figures not between brackets in Table 2). As shown in the score plots for PC1 and PC2 (Figure 1), the two classes were well sorted mainly in the direction of PC1. The dashed line was added to delineate the two groups visually.

It can be seen on VIP plots (see Figure ES2 provided as electronic supplementary material) that the first two most influential variables for a north/south differentiation were $\delta^{13}C$ and $\delta^{18}O$, which is coherent with both being potential indicators of climatic conditions (Heaton et al., 2008; Kelly et al., 2002, 2005; Rossmann et al., 2000; Suzuki et al., 2008). More remarkably [Cu] played a significant discriminatory role in the case of the *10-variables* model. As shown by box-whisker plots (see Figure ES3 provided as electronic supplementary material) [Cu] was rather higher for samples collected in the south than in the north. It is reasonable to assume that the amount of copper in a cereal plant is proportional to the amount of Cu available from the soil where it was grown. What we observed for our cereal data might be explained by the highest Cu concentrations in topsoil and subsoil in Europe reported by Foregs maps (Salminen, 2005) for southern regions, around the Mediterranean basin and within the Iberian peninsula, and along the west coasts of France, UK and Norway. The distribution of copper in subsoil is mainly related to regional and local geology, and to mineralisation. According to Salminen (2005) the distribution of copper in topsoil may also be influenced by anthropogenic contamination (pollution by agricultural sewage enriched in Cu or the use of copper sulphate as a fungicide in fruit cultivation and vineyards).

PLS-DA model fittings for external prediction tests showed satisfactory results with Q2(cum) values similar to those of PLS-DA models generated using the entire data set. The external prediction rates described in column 3 and 4 of Table 3 (figures not between brackets, on 103 and 106 samples, respectively) were 100 % for 9-10 out of 17 sampling sites. The inclusion of [E] data did not make a significant difference overall: there were improvements for samples from Fraenkische Alb, Allgaeu, Chalkidiki, and Galicia, and degradations for samples from Muehlviertel, Limousin, Firenze, Jura Krakowska and Cornwall. The external prediction rates were less than 70

% only for Galicia, Chalkidiki (the *5-variables* model only) and Firenze (both models). These results indicate that both models proposed are robust/stable for the north/south discrimination.

ii) Classification according to proximity to a marine environment (Atlantic, Mediterranean and Inland classes)

We then examined whether the proximity to oceanic/sea conditions (via sea-spray deposition, unusual historical sedimentation conditions, and so on) could be characterised.

Three classes of samples were considered: "Atlantic" and "Mediterranean" for sampling sites less than 100 km away from the respective coasts, and "Inland" for the remaining samples (cf. Table 1).

PLS-DA with the *5*- and *10-variables* models resulted in three and four PCs, respectively. As illustrated in Figure 2 (score plots with the first two PCs) the choice of these three classes (delineated with dotted lines) for sorting samples was relevant. There was also a significant improvement of the modelisation when incorporating the five [E] variables, as confirmed by the increase of Q2(cum) values in Table 2 (from 0.49 to 0.66).

VIP values for both models are presented in Figure ES2. $\delta^{34}S$, $\delta^{18}O$ and $\delta^{13}C$ were the most discriminatory isotopic variables but, with the *10-variables* model, only after [Na] and just before [K].

Furthermore, box-whisker plots in Figure ES3 show that "Atlantic" class cereal samples can be distinguished from others as they exhibit, globally, higher values of [Na], $\delta^{34}S$ and [K] (associated with the emission of sea-spray and the deposition of sea-salt). Cereal samples close to the Mediterranean Sea can also be discriminated but essentially on the basis of tracers of climatic conditions (highest $\delta^{13}C$ and $\delta^{18}O$ values, globally), thus in this case a north/south issue in line with conclusions from the previous discussion.

External prediction tests were run in the same way as described before, and the new PLS-DA model parameters (PCs, R2, Q2) were almost identical to the original ones (Table 2).

The results in column 5 and 6 in Table 3 (figures not between brackets), with the *5*- or the *10-variables* models respectively, indicate 100% success rate for 5 and 9 sampling sites, and < 70% success rate for 4 sampling sites in both cases. For Jylland, Galway and Iceland though, the number of samples available for the external prediction tests was scarce and the results obtained (100% or 0%) must be interpreted cautiously. When considering sampling sites with ≥ 70% success rate (and apart from the 3 sites with ≤ 3 samples), the number of samples predicted correctly is either identical for both modelling approaches (for 7 sites) or better by 11% to 29% with the *10-variables* modelling (for Marchfeld, Gaeuboden, Allgaeu and Jura Krakowska). Therefore it was possible to conclude, first, that the proximity (or not) to a marine environment could be correctly predicted for a very large majority of cereal samples and, second, that the combination with [E] variables (and

particularly [Na] and [K]) significantly improved this prediction ability as compared to the use of isotopic variables alone. The sites with < 70% success rates to the external prediction tests for both types of models (also apart from the 3 sites with ≤ 3 samples) were Chalkidiki, Firenze and Galicia, similarly to what had been observed previously with the "Latitude" classification (*5-variables* model only).

iii) Classification according to bedrock geology (Acid Magmatic, Shale, Limestone and Basaltic classes)

For the third test we sorted cereal samples into four classes ("Shale/mudstone/clay/loess" incl. sandstones and other clastic sediments, "Acid Magmatic", "Limestone" and "Basaltic") according to the bedrock geology of their respective sampling sites, described in Table 1.

Application of PLS-DA to these four classes produced three and five PCs when using the *5-* and *10-variables* models, respectively. Q2(cum) values (Table 2) indicated that the predictive ability of these models had deteriorated in comparison to the previous two classifications. Score plots with the first two PCs are shown in Figure 3. The "Acid Magmatic" class was consistently better sorted than the other 3 classes for both modelling approaches. With only 3 samples the "Basaltic" class was not discriminated at all, although there was a slight difference whether the *5-variables* or the *10-variables* model was considered.

VIP plots for the PLS-DA on bed rock geology classes (Figure ES2) show [Rb], $n(^{87}Sr)/n(^{86}Sr)$ and $\delta^{15}N$ as being the most influential variables. Since the VIP score of [Ca] and $\delta^{34}S$ were ≥ 1 within confidence intervals, these variables may also be considered relevant for this classification. As explained earlier $n(^{87}Sr)/n(^{86}Sr)$ values change depending on the [Rb]/[Sr] concentration ratio in the surrounding geological structure. Thus the major role played by [Rb] and $n(^{87}Sr)/n(^{86}Sr)$ in this classification is logical. Box-whisker plots in Figure ES3 show that with higher values globally for both variables the "Acid Magmatic" class differentiates from the other two major classes. The fact that [Rb] was significantly more influential than $n(^{87}Sr)/n(^{86}Sr)$ is noteworthy. Furthermore, $\delta^{15}N$ values were the highest globally for the "Shale/mudstone/clay/loess" class. $\delta^{15}N$ values may change depending on mineral origin (from -6 ‰ to +6 ‰) or organic origin (from +1 ‰ to 37 ‰) of the fertilisers (Bateman and Kelly, 2007). The trend observed here might indicate agricultural practices specific to farming territories associated with this geological class, although this statement is speculative since its verification was beyond the scope of this study.

As expected, PLS-DA computations for external predictions tests showed no improvements of the Q2(cum) values. According to results in column 7 and 8 in Table 3 the origin of samples was predicted 100% correctly only for Marchfeld (both models), for Limousin (*5-variables* model) and for Orkney and Carpentras (*10-variables* model). External prediction rates for Sicily (both models),

Firenze (*5-variables* model) and Allgaeu (*10-variables* model) were also satisfactory (≥ 70%), and 5 of these correctly predicted sites belonged to the "Shale/mudstone/clay/loess" class. The capability to identify correctly samples from Firenze was also a positive result (not the case with previous two classifications). Globally, external prediction rates were again slightly better when taking into account the five [E] variables than with the five isotopic variables alone.

Overall, samples from Galicia and Chalkidiki were the only ones (also with those from Firenze in the case of the *10-variables* approach) that could not be identified with the combination of classification categories investigated. The outcome for the "Bed rock geology" sorting was less useful than for the "Latitude" and "Marine/Inland" sorting. Several possible reasons can be envisaged to explain this difference. Geological backgrounds are often not homogeneous (for Galicia and Chalkidiki in particular it is more a patchwork than a uniform system, as visible from maps by Asch, 2005). Besides, there is not necessarily a straightforward relationship between bed rock geologies and the compositions of the mineral fractions available for collection by plants growing on the soil surface. Understanding the intimate interactions between soil and plants was outside the scope of TRACE but this supplementary dimension could be a meaningful inclusion in a follow-up project.

b) Robustness to uncertainties on measurement results as additional model validation

The validation approach described in the previous sections for our models was based on the calculation of Q2(cum) values and the run of external prediction tests. We also investigated a much more unusual way of validating these models by examining their robustness to inter-comparison variability. PLS-DA based models assume exactness of the input data and these mathematical tools are not designed to handle measurement uncertainties. We introduced changes to our data set based on the maximum dispersions of results '*d*' observed during our inter-comparisons (0.2‰ for $\delta^{13}C$, 0.9‰ for $\delta^{18}O$, 0.8‰ for $\delta^{15}N$, 1.2‰ for $\delta^{34}S$ and, in relative terms, 0.9‰ for $n(^{87}Sr)/n(^{86}Sr)$ and 40% for [E]). A component '*c*' was added to all original observations corresponding to the product of '*d*' and a randomly generated '*r*' ranging from -1 to +1 (equ. 1).

Equation 1 $c = r*d$

A new series of PLS-DA models for each classification category were produced from the simulated data set. The new prediction results (values between brackets in Tables 2 and 3) showed that such combinations of multivariate analysis were globally robust to the fluctuations imposed to the original data set although some degradation was observed. Q2(cum) values decreased by no more

than 0.05. Trends for the external prediction rates remained also quite similar. Transition from ≥ 70% to < 70% was observed with the *10-variables* "Latitude" model for Chalkidiki and Galicia, with the *10-variables* "Marine/Inland" model for Jura Krakowska, with the *5-variables* "bed rock geology" model for Firenze and Sicily and for the *10-variables* "bed rock geology" model for Algaeu and Carpentras. Transition from < 70% to ≥ 70% was observed with the *5-variables* "bed rock geology" model for Carpentras, with the *10-variables* "bed rock geology" model for Firenze and Gauboden. Overall, external prediction results for all the *10-* and *5- variables* models differed for not more than 11 samples (over 103 or 106 test samples) compared to results from tests with the original data.

These simulations with these dispersion ranges illustrate the potential sensitivity of the prediction tools investigated to the quality of the experimental data, and provide an indication of what could be the maximum experimental uncertainty tolerable for the models proposed to work.

IV Conclusions

The sequential approach described in this study was successful. Grouping sampling sites successively according to the *latitude* (north and south classes) and the *proximity to a marine environment* (Atlantic, Mediterranean and Inland classes) was particularly efficient, using ten carefully chosen variables (five isotopic tracers combined with [Na], [K], [Ca], [Cu] and [Rb] concentrations). Grouping sampling sites according to the *geology of the underlying bed rock* (Shale, Acid magmatic, Limestone and Basaltic classes) was less useful, probably due to the simplicity of the classification used and the fact that bed rock geologies alone may not be adequate to predict soil chemistry. Moreover, this combination of classification categories allowed the identification of 12 unique and generic 'identities' (column 6 in Table 1), thus providing multiple ways of describing geographical locations based on easy to use principles.

Another important lesson learned was that in some instances element concentrations made a greater impact as variables than the five isotopic tracers ([Na] versus $\delta^{34}S$ as a proxy to a marine environment and to the Atlantic ocean in particular; [Rb] versus $n(^{86}Sr)/n(^{88}Sr)$ as a proxy to a certain bed rock geology).

Although our validation scheme was quite extensive, it is natural to examine the question of the domains of applicability. Clearly, it cannot be claimed that the models proposed will always be able to predict/identify correctly the geographical origin of all possible cereal sample collected in Europe. The 17 sampling sites chosen do not cover all the variability on this continent in term of geological backgrounds, soil characteristics and climatic conditions. However, a study like this one,

because of its unique size (number of sampling sites, number of samples per site, number of variables assessed per sample), allows the design of a tool applicable to a much wider range of cereal samples in Europe than any previous comparable studies, and proposes a proof of concept applicable to other types of grains and food products than only cereals.

Our results also demonstrate the feasibility, the potential interest and also the limitations of such a large size study at the scale of a continent. Given the variety of parameters investigated and the great number of samples involved, this kind of project requires the collaboration of many organisations. A tight coordination between partners, including vast efforts in the field of data quality assurance, is mandatory. The quality of experimental data used for modelling purposes cannot be considered better than the difference between measurement results (on similar samples) observed for the different project partners. And our simulations have shown that these multivariate data treatments are not insensitive to such levels of uncertainties.

References

1. Anderson, K. A., Magnuson, B. A., Tschirgi, M. L., Smith, B., 1999. Determining the Geographic Origin of Potatoes with Trace Metal Analysis Using Statistical and Neural Network Classifiers. *Journal of Agricultural and Food Chemistry*, 47, 1568-1575.

2. Anderson, K. A., Smith, B., 2002. Chemical Profiling to Differentiate Geographic Growing Origins of Coffee. *Journal of Agricultural and Food Chemistry*, 50, 2068-2075.

3. Asch, S., IGME-map. *IGME 5000*, 2005. *The 1:5 Million International Geological Map of Europe and Adjacent Areas*, BGR, Hannover.

4. Bateman, A. S., Kelly, S. D., Jickells, T. D., 2005. Nitrogen Isotope Relationships between Crops and Fertilizer: Implications for Using Nitrogen Isotope Analysis as an Indicator of Agricultural Regime. *Journal of Agricultural and Food Chemistry*, 53, 5760-5765.

5. Bateman, A. S., Kelly, S. D., 2007. Fertilizer nitrogen isotope signatures. *Isotopes in Environmental Health Studies*, 43, 237-247.

6. Bréas, O., Guillou, C., Reniero, F., Sada, E., Angerosa, F., 1998. Oxygen-18 measurement by continuous flow pyrolysis/isotope ratio mass spectrometry of vegetable oils. *Rapid Communications in Mass Spectrometry*, 12, 188-192.

7. Camin, F., Bontempo, L., Heinrich, K., Horacek, M., Kelly, S. D., Schlicht, C., Thomas, F., Monahan, F. J., Hoogewerff, J., Rossmann, A., 2007. Multi-element (H,C,N,S) stable isotope characteristics of lamb meat from different European regions. *Analytical and Bioanalytical Chemistry*, 389, 309-320.

8. Camin, F., Larcher, R., Nicolini, G., Bontempo, L., Bertoldi, D., Perini, M., Schlicht, C., Schellenberg, A., Thomas, F., Heinrich, K., Voerkelius, S., Horacek, M., Ueckermann, H., Froeschl, H., Wimmer, B., Heiss, G., Baxter, M., Rossmann, A., Hoogewerff, J., 2010. Isotopic and elemental data for tracing the origin of European olive oils. *Journal of Agricultural and Food Chemistry*, 58, 570-577.

9. Capo, R. C., Stewart, B. W., Chadwick, O. A., 1998. Strontium isotopes as tracers of ecosystem processes: theory and methods. *Geoderma*, 82, 197-225.

10. Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S., 2001. *Multi- and Megavariate Data Analysis: Principles and Applications*. Umetrics AB: Umeå.

11. Esbensen, K. H., 2006. *Multivariate Data Analysis - In Practice. An Introduction to Multivariate Data Analysis and Experimental Design*. Fifth ed., CAMO Software, Oslo.

12. Farquhar, G. D., Ehleringer, J. R., Hubick, K. T., 2003. Carbon Isotope Discrimination and Photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology*, 40, 503-537.

476 13. FP6-TRACE project-website, 9th April, 2010. *http://www.trace.eu.org/*.
477 14. Heaton, K., Kelly, S. D., Hoogewerff, J., Woolfe, M., 2008. Verifying the geographical origin
478 of beef: The application of multi-element isotope and trace element analysis. *Food Chemistry*, 107,
479 506-515.
480 15. Kelly, S., Baxter, M., Chapman, S., Rhodes, C., Dennis, J., Brereton, P., 2002. The application
481 of isotopic and elemental analysis to determine the geographical origin of premium long grain rice.
482 *European Food Research and Technology*, 214, 72-78.
483 16. Kelly, S., Heaton, K., Hoogewerff, J., 2005. Tracing the geographical origin of food: The
484 application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology*, 16,
485 555-567.
486 17. Martin, G. J., Martin, M. L., 2003. Climatic significance of isotope ratios. *Phytochemistry*
487 *Reviews*, 2, 179-190.
488 18. O'Leary, M. H., 1995. Environmental effects on carbon fractionation in terrestrial plants. In
489 *Stable Isotopes in the Biosphere*, Wada, E.; Yoneyama, T.; Minigawa, M.; Ando, T.; Fry, B. D.,
490 Eds. Kyoto University Press, Kyoto,; pp 78–91.
491 19. TRACE quality assurance report, 2009. In *Report of the FP6-TRACE Deliverable D15.9 "Final*
492 *WP15 QA Report"*; 16 September, 2009.
493 20. Rossmann, A., Haberhauer, G., Hölzl, S., Horn, P., Pichlmayer, F., Voerkelius, S., 2000. The
494 potential of multielement stable isotope analysis for regional origin assignment of butter. *European*
495 *Food Research and Technology*, 211, 32-40.
496 21. Salminen, R. (ed.). 2005. *Geochemical Atlas of Europe. Part 1: Background Information,*
497 *Methodology and Maps*. Geological Survey of Finland: Espoo.
498 22. Schellenberg, A., Chmielus, S., Schlicht, C., Camin, F., Perini, M.; Bontempo, L., Heinrich, K.,
499 Kelly, S. D., Rossmann, A., Thomas, F., Jamin, E., Horacek, M., 2010. Multielement stable isotope
500 ratios (H, C, N, S) of honey from different European regions. *Food Chemistry*121, 770–777.
501 23. Sieper, H.-P., Kupka, H.-J., Williams, T., Rossmann, A., Rummel, S., Tanz, N., Schmidt, H.-L.,
502 2006. A measuring system for the fast simultaneous isotope ratio and elemental analysis of carbon,
503 hydrogen, nitrogen and sulfur in food commodities and other biological material. *Rapid*
504 *Communications in Mass Spectrometry*, 20, 2521-2527.
505 24. Simpkins, W. A., Louie, H., Wu, M., Harrison, M., Goldberg, D., 2000. Trace elements in
506 Australian orange juice and other products. *Food Chemistry*, 71, 423-433
507 25. Smith, B. N., Epstein, S., 1971. Two Categories of 13C/12C Ratios for Higher Plants. *Plant*
508 *Physiology*, 47, 380-384.

509 26. Smith, R. G., 2005. Determination of the Country of Origin of Garlic (Allium sativum) Using
510 Trace Metal Profiling. *Journal of Agricultural and Food Chemistry*, 53, 4041-4045
511 27. Suzuki, Y., Chikaraishi, Y., Ogawa, N. O., Ohkouchi, N., Korenaga, T., 2008. Geographical
512 origin of polished rice based on multiple element and stable isotope analyses. *Food Chemistry*, 109,
513 470-475.
514 28. Thode, H., 1991, Sulfur isotopes in nature and the environment: an overview. In *Stable*
515 *Isotopes. Natural and anthropogenic sulfur in the environment* (SCOPE 43), Krouse, H.; Grinenko,
516 V., Eds. Wiley, Chichester,; pp. 2-25.
517 29. Voerkelius, S., Lorenz, G. D., Rummel, S., Quétel, C. R., Heiss, G., Baxter, M., Brach-Papa, C.,
518 Deters-Itzelsberger, P., Hoelzl, S., Hoogewerff, J., Ponzevera, E., Van Bocxstaele, M.,
519 Ueckermann, H., 2010. Strontium isotopic signatures of natural mineral waters, the reference to a
520 simple geological map and its potential for authentication of food. *Food Chemistry* 118, 933-940.
521

Tables

Table 1

Site Name (country code)	Average GPS coordinates (deg., min. and sec.)	samples	Latitude	Marine/Inland	Bed rock geology	Combined 'identity'
Marchfeld (AT)	N48 13 43 E16 49 30	41	N	I	Sh	1
Gauboden (DE)	N48 49 05 E12 34 19	34	N	I		
Allgaeu (DE)	N48 03 37 E10 36 07	38	N	I		
Jylland (DK)	N56 18 43 E09 59 01	7/9*	N	A		2
Orkney (GB)	N58 58 07 W02 56 14	24	N	A		
Firenze (IT)	N43 57 29 E11 18 58	46	S	M		3
Sicily (IT)	N37 45 14 E14 36 00	40	S	M		
Muehlviertel (AT)	N48 27 54 E14 04 29	25	N	I	AM	4
Limousin (FR)	N45 59 01 E02 12 56	40	S	I		5
Cornwall (GB)	N50 04 33 W05 40 32	40	N	A		6
Galicia (ES)	N43 03 34 W08 04 44	20	S	A		7
Chalkidiki (GR)	N40 22 27 E23 36 41	40	S	M		8
Fraenkische Alb (DE)	N49 57 04 E11 06 35	40	N	I	L	9
Jura Krakowska (PL)	N50 10 00 E19 45 03	19/32*	N	I		
Galway (IE)	N53 09 06 W08 56 52	15	N	A		10
Carpentras (FR)	N44 12 00 E05 19 20	40	S	M		11
Iceland (IS)	N63 32 24 W19 39 31	3	N	A	B	12

**** Only with the 10-variables*** **model.**

Table 1. Sampling sites, average GPS coordinates, number of samples per site and classification categories: Latitude (North , **N**, and South, **S**), Marine/Inland (Inland, **I**; Atlantic, **A**; and Mediterranean, **M**) and Bed rock geology (Shale/mudstone/clay/loess incl. sandstone and other clastic sediments, **Sh**; Acid Magmatic, **AM**; Limestone, **L**; and Basaltic, **B**)

Table 2

Classification	Models	PCs	R2X(cum)	R2Y(cum)	Q2(cum)	R2Y(cum)-Q2(cum)
Latitude	*5V*	2 – (2)	0.52 – (0.51)	0.59 – (0.57)	0.58 – (0.57)	0.00 (0.01)
	10V	2 – (2)	0.33 – (0.32)	0.62 – (0.60)	0.62 – (0.59)	0.01 (0.01)
Marine/Inland	*5V*	3 – (3)	0.72 – (0.71)	0.50 – (0.46)	0.49 – (0.45)	0.00 (0.00)
	10V	4 – (4)	0.66 – (0.60)	0.67 – (0.65)	0.66 – (0.64)	0.01 (0.01)
Bed rock geology	*5V*	3 – (3)	0.73 – (0.73)	0.27 – (0.24)	0.26 – (0.23)	0.01 (0.01)
	10V	5 – (4)	0.77 – (0.56)	0.37 – (0.33)	0.35 – (0.30)	0.02 (0.04)

Table 2. Main PLS-DA parameters estimated for the *5-variables* model (5V) and the *10-variables* model (10V). PCs is the number of principal components Figures between brackets correspond to results obtained for the simulated data set (addition to all original data of component corresponding to twice the value of the stated standard uncertainties multiplied by a randomly generated number ranging from -1 to +1)

Table 3

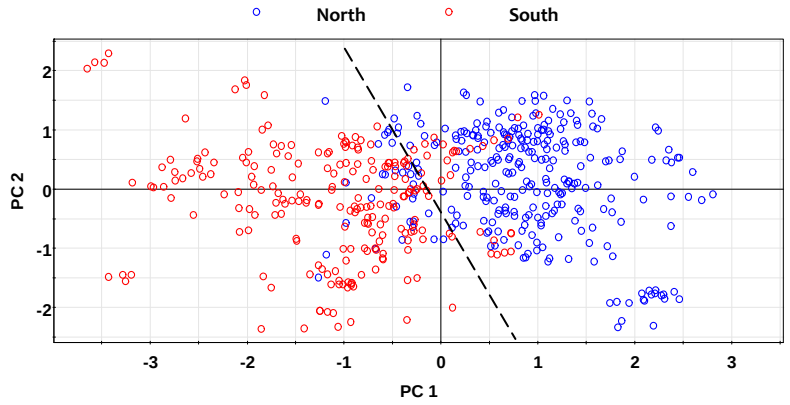
		Latitude		Marine/Inland		Bed rock geology	
Sample site	**n**	***5-variables* model**	***10-variables* model**	***5-variables* model**	***10-variables* model**	***5-variables* model**	***10-variables* model**
Marchfeld	8	100 - (100)	100 - (100)	88 - (100)	100 - (100)	100 - (88)	100 - (100)
Muehlviertel	5	100 - (100)	80 - (80)	100 - (100)	100 - (100)	40 - (20)	60 - (40)
Fraenkische Alb	8	75 - (75)	100 - (88)	75 - (75)	75 - (88)	0 - (0)	13 - (13)
Gauboden	7	100 - (100)	100 - (100)	71 - (86)	100 - (100)	14 - (0)	57 - (71)
Allgaeu	8	88 - (88)	100 - (100)	75 - (75)	100 - (88)	38 - (38)	88 - (63)
Jylland	1	100 - (100)	100 - (100)	100 - (100)	0 - (0)	0 - (0)	0 - (0)
Carpentras	8	100 - (100)	100 - (100)	100 - (100)	100 - (100)	63 - (75)	100 - (63)
Limousin	8	100 - (100)	88 - (88)	88 - (88)	88 - (88)	100 - (100)	88 - (88)
Chalkidiki	8	63 - (50)	75 - (63)	25 - (38)	13 - (25)	50 - (38)	50 - (50)
Galway	3	100 - (100)	100 - (100)	0 - (0)	100 - (100)	0 - (0)	0 - (0)
Firenze	9	56 - (56)	44 - (44)	44 - (44)	44 - (44)	78 - (67)	56 - (78)
Sicily	8	88 - (100)	88 - (88)	88 - (75)	88 - (88)	75 - (63)	88 - (75)
Jura Krakowska	7* / 4	75 - (75)	71 - (71)	75 - (75)	86 - (57)	0 - (0)	29 - (29)
Galicia	4	50 - (50)	75 - (50)	0 - (0)	50 - (25)	0 - (0)	50 - (25)
Cornwall	8	100 - (100)	75 - (75)	100 - (88)	100 - (100)	25 - (25)	63 - (63)
Orkney	5	100 - (100)	100 - (100)	100 - (100)	100 - (100)	80 - (80)	100 - (100)
Iceland	1	100 - (100)	100 - (100)	0 - (0)	100 - (100)	0 - (0)	0 - (0)

* Only with the *10-variables* model

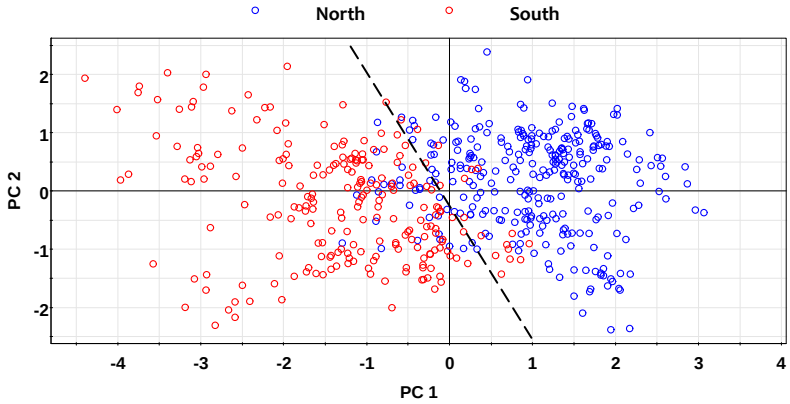
Table 3. External prediction results (%) for “Latitude”, “Marine/Inland” and “Bed rock geology” classification categories using 103 or 106 samples (for *5*- and *10-variables* models, respectively) as test data sets. External prediction results between brackets are for the data set simulated to estimate the impact of measurement uncertainty in these models.

Figures

Figure 1



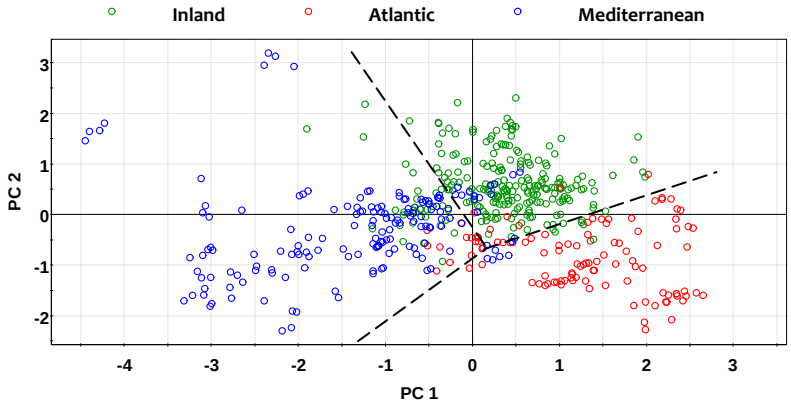
(a) *5-variables* model (*n*=512)



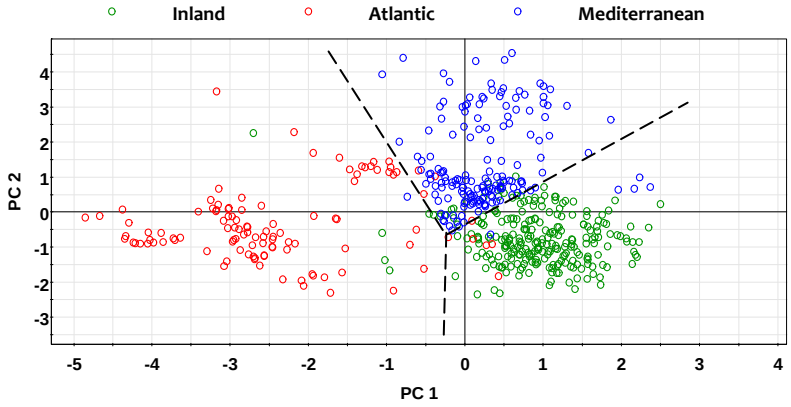
(b) *10-variables* model (*n*=527)

Figure 1. "Latitude" classification (classes North and South) PLS-DA score plots for the *5-variables* (a) and the *10-variables* (b) models

562 Figure 2



(a) *5-variables* model (*n*=512)

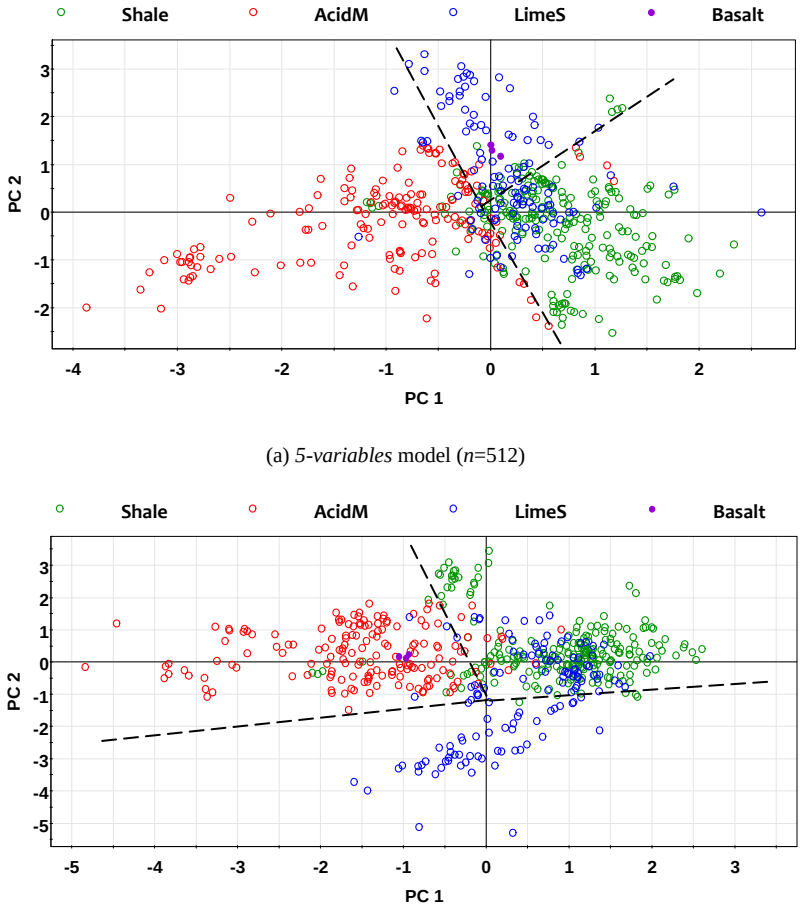


(b) *10-variables* model (*n*=527)

Figure 2. "Marine/Inland" classification (classes Inland, Atlantic and Mediterranean) PLS-DA score plots for the *5-variables* (a) and the *10-variables* (b) models

563
564
565
566
567
568
569
570
571
572
573
574
575
576

577 Figure 3



(a) *5-variables* model (*n*=512)

(b) *10-variables* model (*n*=527)

Figure 3. "Bed rock geology" classification (classes Shale, Acid Magmatic, Limestone and Basaltic) PLS-DA score plots for the *5-variables* (a) and the *10-variables* (b) models

578