Food Chemistry 464 (2025) 141564

Contents lists available at ScienceDirect

Food Chemistry

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

Data fusion of HS-SPME-GCMS, NIRS, and fluorescence, using chemometrics, has the potential to explore the geographical origin of gentian rhizomes

Céline Lafarge [a], Laurence Dujourdy [b,c], Gilles Figueredo [d], Stéphanie Flahaut [e], Christophe Poix [f], Laurent Rios [f], Elias Bou-Maroun [a,*], Christian Coelho [f,*]

[a] *Université Bourgogne Franche-Comté, Institut Agro, Université Bourgogne, INRAE, UMR PAM 1517, 21000 Dijon, France*
[b] *Institut Agro Dijon, Direction Scientifique, Cellule d'Appui à la Recherche en sciences des données, 21000 Dijon, France*
[c] *LIB, Laboratoire d'Informatique de Bourgogne, 21000 Dijon, France*
[d] *LEXVA Analytique, 7 rue Henri Mondor, Biopole Clermont Limagne, 63360 Saint Beauzire, France*
[e] *CPPARM, ZA Les Quintrands, Route de Volx, 04100 Manosque, France*
[f] *Université Clermont Auvergne, INRAE, VetAgro Sup campus agronomique de Lempdes, UMR F, 15000 Aurillac, France*

ARTICLE INFO

Keywords:
Gentiana Lutea
Data fusion
Near-infrared spectroscopy
Fluorescence
SPME-GCMS
Geographical origin

ABSTRACT

Gentiana lutea rhizomes are known for their bitter tasting properties conferred by its unique biochemical content. They are currently of interest in phytotherapy, animal nutrition, food processing, cosmetic applications and agroecology. In this study, a NIRS, fluorescence and HS-SPME-GCMS dataset of 55 rhizomes from four different French mountains (Alpes, Jura, Massif Central and Pyrénées) was collected with the aim of assessing the variability of *Gentiana lutea* composition at different scales. The feasibility of data fusion strategies was demonstrated to be effective in distinguishing the geographical origin of *Gentiana lutea* roots over a wide area. The results suggest that data fusion methods have the potential to be more effective in the quality of separation of studied sites of *Gentiana lutea* roots than individual decisions obtained from individual analytical tools. However, to guarantee the geographical origin of *Gentiana lutea* roots within a single massif using these techniques, environmental factors must be considered.

1. Introduction

The yellow gentian, *Gentiana Lutea*, grows annually during the spring season. It is native to the mountains of central and southern Europe, preferring calcareous soils, and grows naturally on uncultivated land in France, Spain and the Balkans. The plant is protected in Europe and is mainly cultivated in Germany and France (European Medicines Agency (EMA), 2009). A sustainable management program of *Gentiana Lutea* is being carried out in Auvergne (France) in order to safeguard *Gentiana lutea*, develop and promote *Gentiana lutea* and its products. *Gentiana lutea* is spread in the mountains of seed dispersal at the end of the flowering season and by vegetative dispersal through its underground rhizomes during the dormant season (Ando et al., 2007; Arberas et al., 1995; Toriumi et al., 2003). *Gentiana Lutea* rhizomes accumulate primary and secondary metabolites whose concentrations are influenced by environmental and developmental factors (Coelho et al., 2022; Marković et al., 2019). Among them, seco-iridoids are the most important secondary metabolite components of *Gentiana lutea* rhizomes, reaching 6 % to 12 % of the rhizome dry weight, which contributes to its attractiveness for industrial transformations in the field of food, beverages, pharmaceuticals, cosmetics and phytopharmaceuticals (Berthon et al., 2023; Biehlmann et al., 2020; Coelho et al., 2023; Mirzaee et al., 2017; Mustafa et al., 2015; Ponticelli et al., 2023). Recent analytical developments have revealed the complex composition of rhizomes in terms of compounds volatile (Aberham et al., 2011; Ando et al., 2007; Gibitz-Eisath et al., 2022; Xu et al., 2017). Novel phytochemical compounds with potential bioactivity belonging to the family of iridoids, terpenes, xanthones, flavones and their glycosylated forms have been

* Corresponding authors.
E-mail addresses: celine.lafarge@agrosupdijon.fr (C. Lafarge), laurence.dujourdy@agrosupdijon.fr (L. Dujourdy), g.figueredo@lexva-analytique.com (G. Figueredo), stephanie.flahaut@cpparm.org (S. Flahaut), christophe.poix@vetagro-sup.fr (C. Poix), laurent.rios@vetagro-sup.fr (L. Rios), elias.bou-maroun@institut-agro.fr (E. Bou-Maroun), christian.coelho@vetagro-sup.fr (C. Coelho).

https://doi.org/10.1016/j.foodchem.2024.141564
Received 3 May 2024; Received in revised form 18 September 2024; Accepted 5 October 2024
Available online 9 October 2024

discovered on *Gentiana Lutea* rhizomes for pharmaceutical purposes (Aberham et al., 2011; Enders et al., 2021; Toriumi et al., 2003). Some studies on foods flavoured with gentian also indicate putative chemical biomarkers (amarogentine, loganic acid, barium, aluminium, 2-methoxy-sec-butylpyrazine, phenyl acetate) of mountain origin in the composition of these gentian-based products.(Biehlmann et al., 2020; Coelho et al., 2023).

Spectroscopic techniques using fluorescence emission and infrared absorption combined with chemometrics offered the possibility to study the chemical composition of botanical plants with unsupervised vision (Mazina et al., 2015; Obeidat et al., 2007; Y. Wang et al., 2018). In most cases, chemometric tools using multivariate analysis are used for molecular profiling and authentication purposes (Abraham & Kellogg, 2021). Spectral analysis of different botanical species of *Gentiana* in the near-infrared (NIR: 1100–2500 nm) and mid-infrared (MIR: 2500–25,000 nm) regions allow the assessment of species differences, geographical origin and chemical composition of *Gentiana lutea* rhizomes. (Coelho et al., 2022; Shen et al., 2020).

Even if these analytical methods seemed to be useful for classification and predictive abilities of botanical origin at different scale levels, some inaccuracies in spatial predictive models are often mentioned due to multiple environmental factors affecting the growth of herbs. Moreover, some genetic variations and historical biogeography have been shown to modify the phenotypic characteristics and chemical composition of *Gentiana lutea* populations in a wide area of occurrence (González-López et al., 2014; Veiga et al., 2016). Improving this accuracy could be done in two different ways: (i) selecting appropriate models carried out on the most relevant samples (Coelho et al., 2022; J. Li, 2019), or combining multispectral analysis with a data fusion approach on multi-source datasets (Y. Li et al., 2018; Pérez-Ràfols et al., 2023; Ríos-Reina et al., 2019; Q.-Q. Wang et al., 2019; Y. Zhang & Wang, 2023). Such an analytical strategy has already been applied to assess the geographical origin of aerial and subterranean parts of *Gentiana rigescens* (Shen et al., 2019) and to our knowledge, never for *Gentiana lutea*. Recent studies have shown that data fusion, coupled with chemometric approaches, can effectively assess and classify food quality and identify geographical origin. (Drivelos et al., 2014; Márquez et al., 2016; Ottavian et al., 2014; Robert et al., 2021; Schwolow et al., 2019).

To the best of our knowledge, the combination of chromatographic, fluorescence and spectroscopy techniques by data fusion on the exploration of the geographical origin of gentian rhizomes has not been described. Data fusion is an approach where the data from multiple sources of different nature are combined and analyzed jointly in order to take advantage of their features and improve the representation of information compared to the respective sources separately (Castanedo, 2013). Chromatographic, fluorescence and spectroscopy data can be fused using a mid-level data fusion approach. Mid-level data fusion is a systematic approach comprised of intermediate steps between the raw data and the final model (Mafata et al., 2022). Many chemometric tools can be introduced as data projection linear methods (Brereton, 2003), which compress raw data, uncover hidden correlations, and separate useful information from noise. Projection methods provide a very intuitive and visual approach for data analysis and Principal Component Analysis (PCA) is the main tool for this purpose. The most commonly used unsupervised data fusion methods are principal component analysis (PCA) and multiple factor analysis (MFA) (Mafata et al., 2022). The subspace identified by PCA constitutes the most faithful dimensional approximation of the original data. This allows compression of the data dimensionality and at the same time a minimal loss of information (Cordella, 2012). PCA was used to explore chemical profiles in more depth and to determine the existence of specific groups. On the other hand, multiple factor analysis (MFA) is a technique for multivariate data analysis designed to simplify and present intricate data tables wherein individuals are characterized by multiple sets of variables, whether quantitative or qualitative, grouped together (Husson et al., 2017). It considers the influence of all active variable groups when assessing distances between individuals. While the number and type of variables can vary across groups, it is crucial that variables within a specific group share the same nature (Abdi & Williams, 2010). In the MFA framework, the simultaneous consideration of multiple variable sets requires balancing influences from each set. Consequently, a weighting process is applied to variables during the analysis. Variables within the same group are normalized using a common weighting value, which may vary across groups. MFA has diverse applications across fields, particularly in scenarios where variables are organized into groups (Pagès, 2002). MFA, where statistical techniques are used to integrate and analyze data from multiple sources or types to uncover underlying relationships and patterns, allows to gain a more comprehensive understanding of complex datasets by leveraging the complementary information provided by different data sources or variables (Cocchi & Reggio, 2019; Mafata et al., 2022). Moreover, this type of fusion enables an easy interpretation of the results, since the contribution of each individual block can be visualized (Ríos-Reina et al., 2019).

In this study, a data set of NIR-infrared spectra, excitation-emission matrices of fluorescence and HS-SPME-GCMS analysis of 55 rhizomes from four different French mountains (Alpes, Jura, Massif Central and Pyrénées) was collected with the aim of evaluating the variability of the composition of *Gentiana lutea* at different scales. In fact, Mustafa et al. (2015) already show differences in different populations of *Gentiana lutea* rhizomes. Moreover, in a previous study, Coelho et al. (2022) already show differences among the 55 rhizomes sampled in the four different French mountains (Alpes, Jura, Massif Central and Pyrénées) that the seco-iridoids contents ranged between 6 and 12 % in dry weight according to the geographical origin and plant-growing conditions.

The current work aims to compare the feature of individual of each of the analytical techniques (NIR-infrared spectroscopy, excitation-emission matrices of fluorescence and HS-SPME-GCMS analysis) of those obtained by combining three analytical techniques to go deeper in the characterization of *Gentiana lutea* roots. This, in this study, each analytical technique (NIR-infrared spectroscopy, fluorescence excitation-emission matrices and HS-SPME-GCMS analysis) was first analyzed individually to evaluate the capabilities of each analytical technique for the characterization of *Gentiana lutea* roots. These techniques provided complementary information that allowed a mid-level data fusion approach thus demonstrating the potential and benefits of data fusion. By adopting such a strategy, we hope to gain a more holistic view of the geographical effects on the composition of *Gentiana lutea* rhizomes, and to assess the extent to which a mountainous environment can confer chemical typicity on the plant.

2. Material and methods

2.1. Gentian roots sample

The information regarding the sampling of the 55 *Gentiana lutea* rhizomes were previously reported in Coelho et al. (2022). In summary, a total of 55 *Gentiana lutea* roots were sampled entirely in July 2018 from ten different sites in four French mountains (Massif Central, Jura, Pyrénées, and Alpes). In each site, a minimum of five gentians were uprooted with an “evil fork” to consider the biochemical variability of the sampling site. Two of these sites were two gentian cultivars (CUL) ($n = 10$) and the eight others were wild sites (WIL) ($n = 45$). Table 1 summarizes the number of sampled gentian roots for each site in four French mountains used in this study. All roots were cleaned from their residual soil, manually sliced in 1–2 cm pieces, dried at 40 °C for two days and lastly ground to a fine powder and stored at 4 °C until physicochemical analyses.

2.2. Chromatographic and spectroscopic analysis

2.2.1. Analysis of volatile compounds by HS-SPME-GCMS

Whatever the sample, the extraction of volatile compounds was done

Table 1

: Characteristic of *Gentiana Lutea* roots samples collected from the four French mountains. GPS represent the Global Positioning Unit, n represent the sample size of *Gentiana Lutea* roots, WIL/CUL represent the wild vs cultivated growing practices (Biehlmann et al., 2020).

Geographical origin	Sites	GPS coordinates	Samples size (n)	Growing practices
Massif Central	A: Fraux	N: 45°02.971′ E: 2°52.45’	8	WIL
	B: Liorangues	N: 45°0.571′ E: 3°38.723’	5	CUL
	C: Malbo	N: 44°58.709′ E: 2°45.921’	5	WIL
	D: Nasbinals	N: 44°40.736′ E: 3°02.397’	5	WIL
	E: Gelles	N: 45°45.49′ E: 2°44.970’	5	CUL
	F: Pégrol	N: 45°36.586′ E: 3°51.645’	5	WIL
	G: Picherande	N: 45°28.512′ E: 2°50.89′	5	WIL
Jura	H: La Chapelle des Bois	N: 46°37.481′ E: 6°8.5’	5	WIL
Pyrénées	I: Bagnères de Luchon	N: 42°44.430′ E: 0°38.946’	6	WIL
Alpes	J: Samoens	N: 46°12.477′ E: 6°28.920’	6	WIL

by headspace solid-phase microextraction and analyzed by GC–MS (HS-SPME-GCMS).

This analysis was carried out using a three-phase fiber (divinylbenzene (DVB) / carboxen (CAR) / polydimethylsiloxane (PDMS), 50/30 μm, Supelco). Before use, the fiber was conditioned in accordance with the manufacturer's recommendations. A preliminary study was done to choose the best extraction parameters (Biehlmann et al., 2020). The selected extraction conditions were the following: 1 g of powder in a 20 mL vial was incubated in a water bath at 40 °C for 15 min. Then the fiber was exposed to the sample headspace for 15 min at 40 °C and was desorbed for 10 min into GC–MS inlet. The analyses were done in technical triplicate.

A mass spectrometer (Agilent 5975C-VLMSD, electron ionization at 70 eV) paired with an Agilent 7890 A gas chromatograph fitted with a split/splitless injector was used to perform this analysis. The chromatograph was equipped with a capillary column DB5ms of 30 m × 0.32 mm (J&W Scientific). Film thickness was 0.50 μm. Helium was used as carrier gas at a rate of 1.5 $mL.min^{-1}$ (average velocity of 44 $cm.s^{-1}$). The temperature of the oven was increased from 40 °C to 240 °C at 4 $°C.min^{-1}$ and maintained 5 min at 240 °C. The injection temperature was 240 °C and was done in splitless mode. The purge flow to split vent was 25 $mL.min^{-1}$ at 2 min.

The mass spectrometer was used in scan mode from *m/z* 29 to 400. The corresponding volatile compounds were tentatively identified by matching their spectral fragmentation with those provided by the mass spectral library of the National Institute of Standards and Technology (NIST) and the Wiley Registry (WILEY). In addition, for each volatile compound obtained (76 volatile compounds), linear retention index (LRI) was calculated using the retention times of a standard mixture of C8-C19 saturated alkanes (Sigma Aldrich) and compared with the LRI values published in the literature of columns with the same polarity. Moreover, Extract Ion Chromatograms (EIC) from total ion chromatograms (TIC) were used for the analysis. In an extracted-ion chromatogram (EIC), one or more *m/z* values representing one or more analytes of interest are extracted from the entire data set of total ions chromatograms (TIC). In order to identify gentian robust markers, volatile compounds systematically present in at least two of the three technical replicates and in the five biological replicates for at least one of the two geographical sites has been selected for the chemometric analysis. If a compound was absent in one of the five biological samples by geographical site, it was discarded for the chemometric analysis. Finally, ten major volatile compounds have been selected.

2.2.2. Fluorescence analysis

The sample preparation for fluorescence analysis was adapted from a previous published methodology (Biehlmann et al., 2020; Coelho et al., 2022). *Gentiana lutea* rhizomes powders were extracted in methanol (weight/volume) ratio of (10 g/1 L) during 12 h at 60 °C under heat reflux. Methanolic solutions were filtered with 0.45 μm nylon filters and diluted twenty times in ultrapure water prior being analyzed in a Horiba Aqualog® spectrofluorimiter using 1 cm-pathlength quartz cell. Excitation Emission Matrices (EEMs) were acquired in the range of excitation wavelengths from 600 to 225 nm (3 nm steps) and emission wavelengths from 211 to 617 nm (3.34 nm steps). All EEMs were corrected from inner filter effects and Rayleigh scattering and normalized to a 1 ppm quinine sulfate reference solution. PARAllel FACtor analysis of the 55 corrected EEMs was carried, using the drEEM tutorial (Murphy et al., 2013) on a Matlab console, in order to build a PARAFAC model fitting the entire variability of EEM datasets. The number of PARAFAC components was optimized by the CORe CONsistency DIAgnostic test and PARAFAC model was split half validated. After model validation, F_{max} values for each PARAFAC components are enabled to describe the original EEM datasets. For this study, four F_{max} values (named from Fmax1 to Fmax 4) and Fmax4/Fmax1 ratio were considered (five variables).

2.2.3. Near infra-red spectroscopy (NIRS)

Absorbance spectra were obtained directly on the gentian powder without sample pre-treatment using NIR spectrophotometer (Field Spec 4 Standard - Res, Bonsai Advanced Technologies, Madrid, Spain) equipped with optic fiber (diameter 600 μm; length 15 cm). The software used was RS3 by ASD Inc. Measurement was performed in reflectance mode between 350 and 2500 nm (2151 variables). Ten scans were recorded at room temperature and were averaged to give one spectrum. The spectra were recorded at intervals of 1 nm. To minimize sample error, all samples were analyzed five times. All spectra were considered in the data analysis.

2.3. Chemometric analysis

As data came from a multiplatform analysis of the same samples (NIRS, GC–MS, and Fluorescence), a multiblock dataset was built in which the data were not simply multivariate but also multimodal, i.e. multivariate and multisource. In this case, analytical profiles are multivariate (as the responses are acquired at several wavenumbers) and the modes are represented by the different analytical techniques. Multiblock data analysis accomplishes tasks similar to those undertaken with single-block chemometric techniques such as PCA, but they can achieve an enhanced understanding of the common and the distinct information present in the data acquired from different platforms (Mishra et al., 2021). To enhance the interpretation, we conducted a multiple factor analysis (MFA).

The preparation of near-infrared (NIR) spectral data through pre-processing has become a fundamental aspect of chemometrics modelling. The primary aim of this pre-processing is to eliminate inherent physical phenomena within the spectra, thereby enhancing the performance of subsequent multivariate regression, classification models, or exploratory analyses (Rinnan & Engelsen, 2009). Data from the spectral dataset were first pre-processed using of Savitky-Golay filter (window 11, polynomial order 2, 2nd derivative). The Savitzky-Golay filter is a digital signal processing technique used for smoothing and differentiation of data. It is particularly useful for reducing noise and extracting trends from signals. When applied to near-infrared (NIR) data, the Savitzky-Golay filter can help enhance the signal by removing unwanted fluctuations or noise.

Data analyses were carried out using the RStudio 2023.09.1 and R-4.2.1 (R Core Team, 2018), and specific packages: "rstatix" (Kassambara, 2024), "FactoMineR" (Lê et al., 2008). "rstatix" was used for ANOVA calculation, "FactoMineR" and "factoextra" were used for PCA and MFA. Quasar 1.9.0 was used for radviz plots for fluorescence data, preprocessing of NIR spectra and PCA on these data (part 1) (Demšar et al., 2013; He et al., 2021; Toplak et al., 2021) (Demšar et al., 2013; He et al., 2021; Toplak et al., 2021).

3. Results and discussion

3.1. Study of variability within a massif: The massif central (MC)

The first phase of our investigation focused on the variation between samples within a specific geographical unit, namely the Massif Central. This massif was selected for the study due to the availability of a considerable number of sites, which provided a substantial amount of data for comparison with other massifs. This approach was chosen to understand in depth the variability of the data generated by three analytical techniques: Fluorescence, NIRS and HS-SPME-GCMS. This step serves as a basic exploration aimed at elucidating the intricacies of variability in the designated region before proceeding with further analyses between geographical areas.

Fig. 1A shows an example of the EEM fluorescence obtained for the wild site A with the Fmax obtained. With an excitation wavelength of 250 nm, two PARAFAC components named Fmax 1 and Fmax 4 showed a maximum emission at 350 nm and 450 nm respectively. With an excitation wavelength of 300 nm and 350 nm, two other PARAFAC components, Fmax 2 and Fmax 3, respectively, showed a maximum emission at 450 nm. The EEM fluorescence was quantitatively separated into these four fluorescence components. All collected Fmax were plotted using Radial Coordinate Visualization (RadViz) and are shown in Fig. 1B. RadViz is a multivariate data visualization algorithm that draws each dimension (variable) uniformly around the circumference of a circle, and then draws points inside the circle in such a way that the point normalizes its values on the axes from the center to each arc. This mechanism makes it possible to represent as many dimensions as fit on a circle, greatly increasing the dimensionality of the visualization (Hoffman et al., 1999). No clear clusters were observed. The five wild area samples (A, C, D, F, G) are scattered and overlapping. To confirm this descriptive result, hypothesis tests (Kruskal-Wallis non-parametric test) were performed on the five variables. No significant difference was observed between the areas at 5 % risk. Fig. 1C shows the results of the Fmax 4/Fmax 1 ratio. The R-squared estimate corresponds to the percentage of variance in the dependent variable (Fmax 4/Fmax 1 ratio) explained by the independent variable (site). A value of 0.18 is obtained for the R-square estimate. The interpretation values commonly used in the literature are $0.01 < 0.06$ (small effect), $0.06 < 0.14$ (moderate effect) and > 0.14 (large effect) (Cohen, 1992; Kotrlik & Williams, 2003; Tomczak & Tomczak, 2014). In our study, a large effect is obtained, which means that the observation of no differences between sites must be taken with caution because of the small sample size (five per site).

NIRS of 25 samples (5 sites × 5 replicates) were collected. The raw spectral data are shown in Fig. 2A. The original spectra showed similar trends. After pre-processing of the spectral data set with a cut-off between 1822 and 1839 nm and an edge cut at 1000 nm and 2400 nm and application of a Savitzky-Golay filter, a Principal Component Analysis (PCA) was performed to provide a more accurate interpretation of the relationships between observations and variables. Fig. 2B corresponds to the scores plot in plan Dim 1-Dim 2 (representing 50 % and 15 % of the total variance, respectively), illustrating the differences between sites: A and F on the left, C and D on the right. Fig. 2C corresponds to the loading plot on Dim 1 and shows obvious high absorption bands at 1450 nm and 1930 nm. The first region (1450 nm) combines the first overtone of the O—H stretching vibrations (H2O water) and the R-OH stretching vibrations. The second region (1930 nm), as a fingerprint region, combines the second overtone of C═O stretching vibrations and combination bands of O—H stretching vibrations. These peaks are related to the absorption of water, cellulose and sugars. (Cevoli et al., 2024; Y. Li et al., 2018; Ozaki et al., 2006).

The same 25 samples were analyzed by HS-SPME-GCMS. An example of the chromatogram obtained is shown in Fig. 3A. The components identified by HS-SPME-GCMS showed a remarkable chemical diversity belonging to different classes of compounds. The main volatile components were found in the first 20 min of the chromatogram. The top ten molecules were identified using the Extracted-Ion Chromatogram (EIC) and were mainly aldehydes and terpenes (Fig. 3A **and Table ST1**). This is in good agreement with literature (Biehlmann et al., 2020; Mustafa et al., 2016). Using these HS-SPME-GCMS-EIC results, an MFA was performed to show how the major volatile compounds contribute most to explaining the variations in the dataset (Fig. 3B). For a given site, the square corresponds to the center of gravity of the partial points of the individual. The sites with similar profiles are close to each other on the factor map. The first axis, which explains 62.9 % of the variance, is mainly between sites C, D (positive values) and sites A, F, G (negative values). The second axis, which explains 17 % of the variance, provides no additional information. By using only the main volatile compounds, we have chosen to deprive ourselves of the global information obtained from the whole signal. In order to obtain a more accurate interpretation of the relationships between observations and variables, and between the variables themselves, we therefore decided to carry out a second analysis, this time using the full chromatograms (Total Ion Current chromatogram TIC) to see if the response was the same. A Principal Component Analysis (PCA) was performed and is shown in Fig. 3C. The first two dimensions accounted for 46.3 % of the total variance. The quality of the representation of the variables on the factor map, called $\cos^2$ (square cosine), was indicated by a colour gradient. A high $\cos^2$ (close to 1) indicated a good representation of the principal component. Different groups could be observed from the Dim1-Dim2 plot. Dim1 modelled the difference between sites C, D, A (negative values) and sites F and G (positive values). Dim2 modelled the difference between site A (positive values) and sites D and C (negative values). This gives us 3 different groups of individuals: A, F-G and C—D. The main feature responsible for the difference between the sites was the presence of different volatile components. The retention times (4.76 Hexanal CAS N° [66–25-1] and 2.81 Pentanal CAS N° [110–62-3]) represented the site C and D. The retention times (14.97 1-(2,4-Dimethyl-furan-3-yl)-ethanone 61 CAS N° [032933–07-6], 1.82 Acetic acid CAS N° [64–19-7], 1.11 unknown molecule, 6.96 Allyl Isothiocyanate CAS N° [57–06-7]) represented the site A and the retention times (11.97 Nonane, 2,6-dimethyl- CAS N° [17302–28-2], 11.63 unknown molecule, 6.475 Octane 4-methyl- CAS N° [2216-34-4], 13.37 Undecane 4,6-dimethyl- CAS N° [17312–82-2], 9.87 Nonane, 2-methyl- CAS N° [871–83-0], 11.83 Nonane, 2,5-dimethyl- CAS N° [017302–27-1], 15.06 and 11.16 are unknown molecule) represented the site F and G. According to these results, for this study the discrimination between sites is more representative when working with Total Ion Current (TIC) chromatogram data than with Extracted Ion Chromatogram (EIC) data. This clearly shows that compounds other than the main volatiles are involved in the discrimination of samples. These results confirm the work of Reyrolle et al. on volatile fingerprints: non-targeted analysis improves the ability to detect unexpected compounds. (Reyrolle et al., 2022, 2023).

Each of the individual techniques used consecutively to analyze the same sample (SPME-GCMS, fluorescence, NIRS) provided information about the samples and made it possible to identify, with different levels of precision, which samples were similar and which were different. The fusion of all these data from complementary techniques can be a

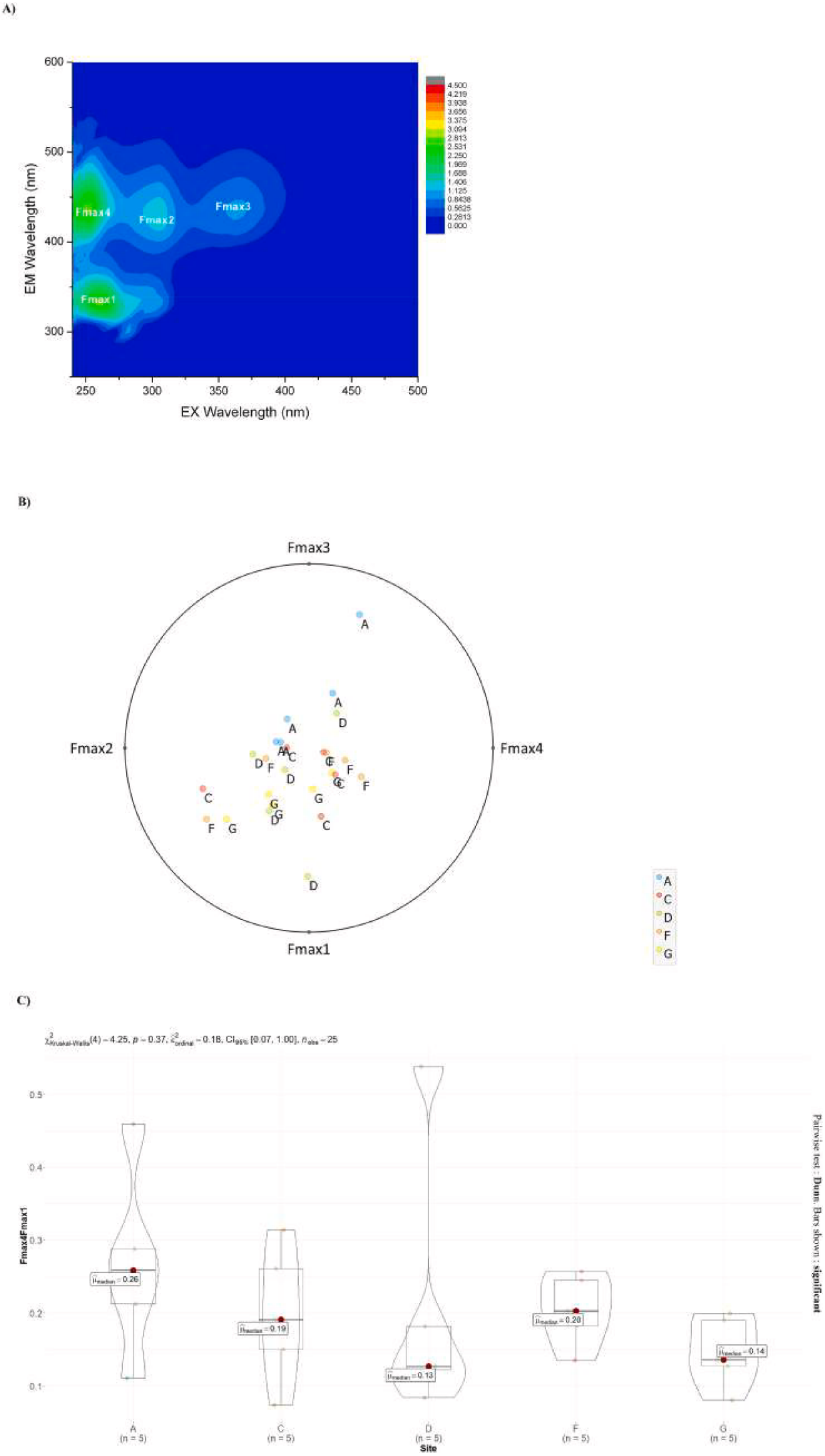


Fig. 1. **A)** Locations of the main EEM peaks and fluorescence indices obtained for the wild sites A of massif central, **B)** Radviz plot for Excitation Emission Matrices (EEMs) samples of wild sites from massif central: in blue site A, in red site C, in green site D, in orange site F, and in yellow site G, **C)** Violin plots of Fmax1/Fmax4 ratios for the five wild sites of massif central labelled A, C, D, F, G according to Kruskal-Wallis tests. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

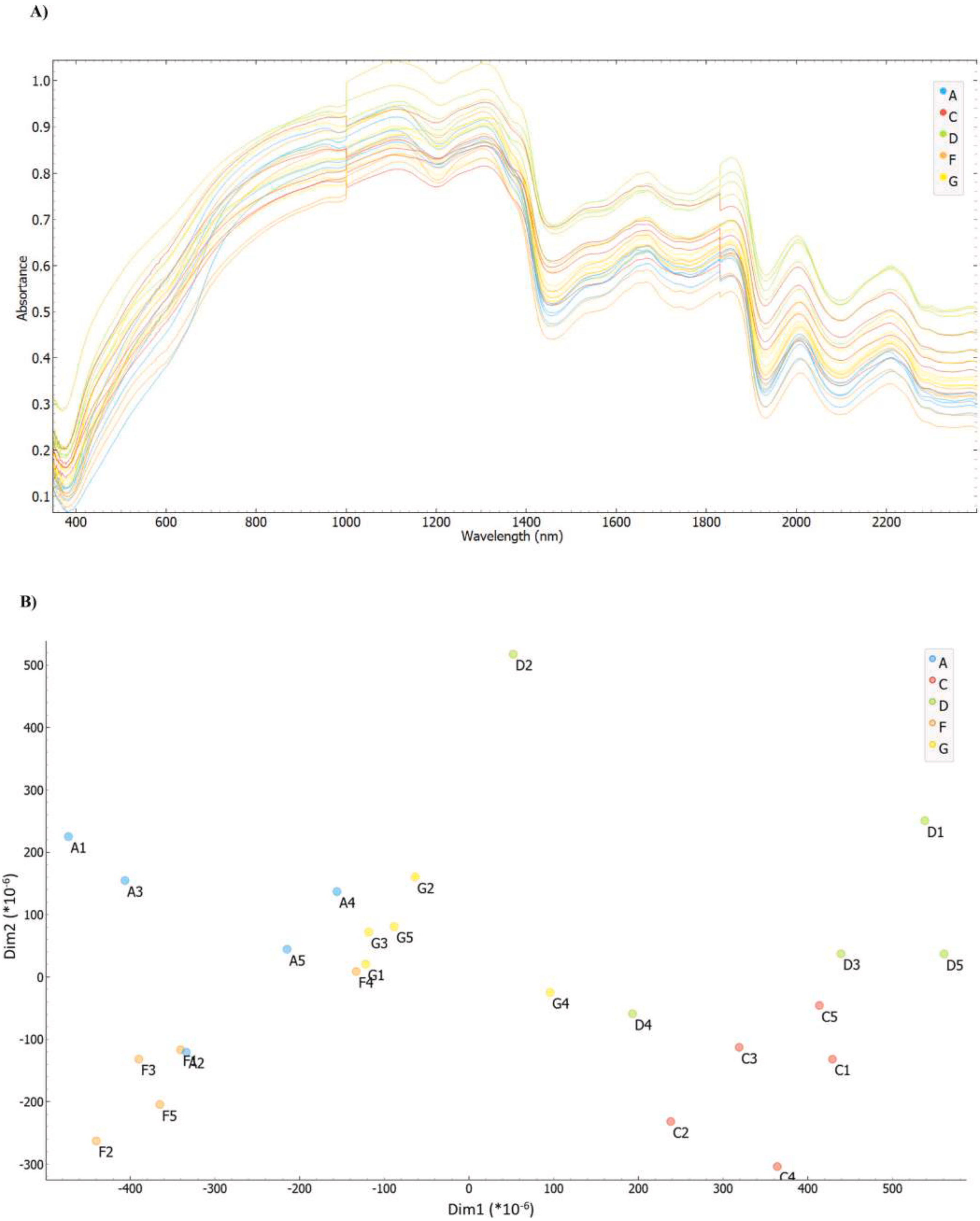


Fig. 2. **A)** NIRS Absorbance spectra of *Gentiana lutea* collected from all wild sites of massif central labelled A, C, D, F, G. **B)** PCA scores plot on Dim1-Dim2 plan obtained from NIRS Absorbance spectra of *Gentiana lutea* collected from all wild sites of massif central labelled A (in blue), C (in red), D (in green), F (in orange), G (in yellow). **C)** PCA loadings plot on Dim 1 obtained from NIRS Absorbance spectra of *Gentiana lutea* collected from all wild sites of massif central. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

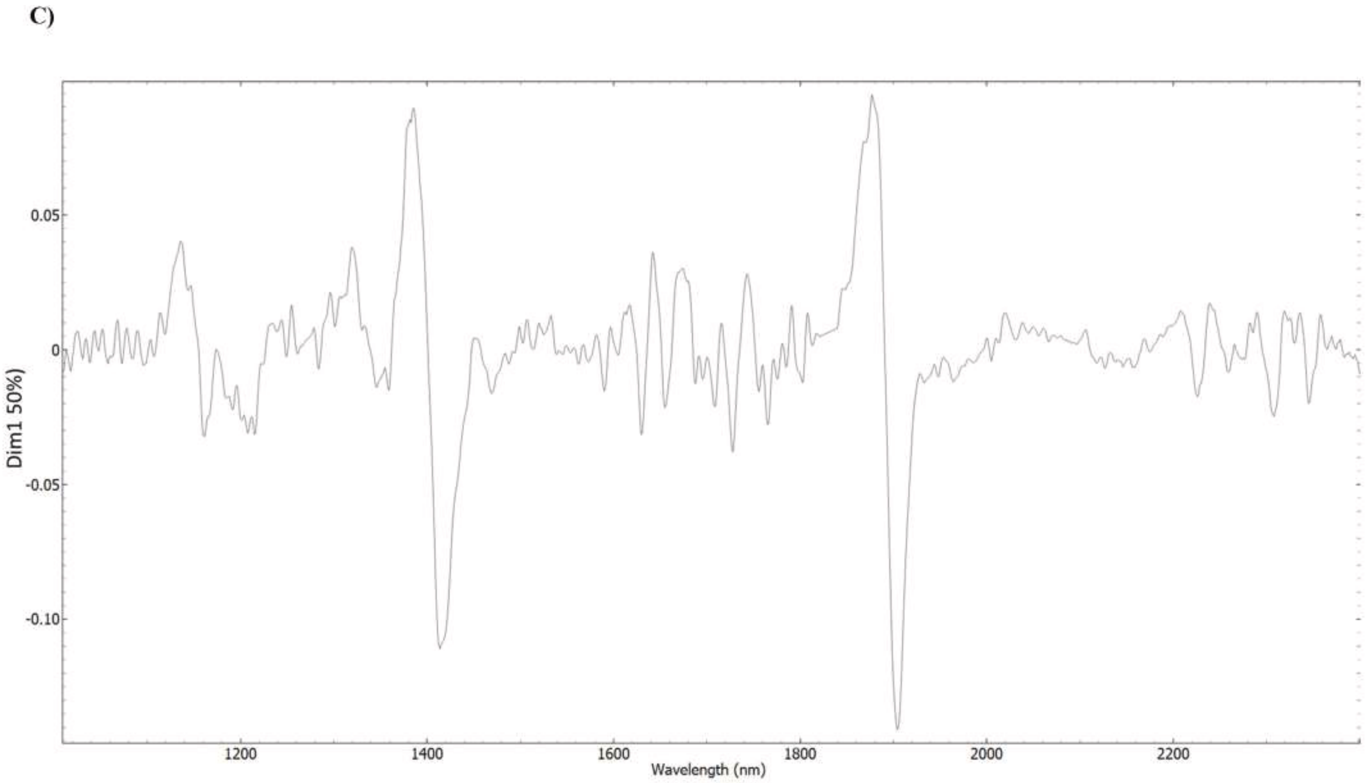


Fig. 2. (*continued*).

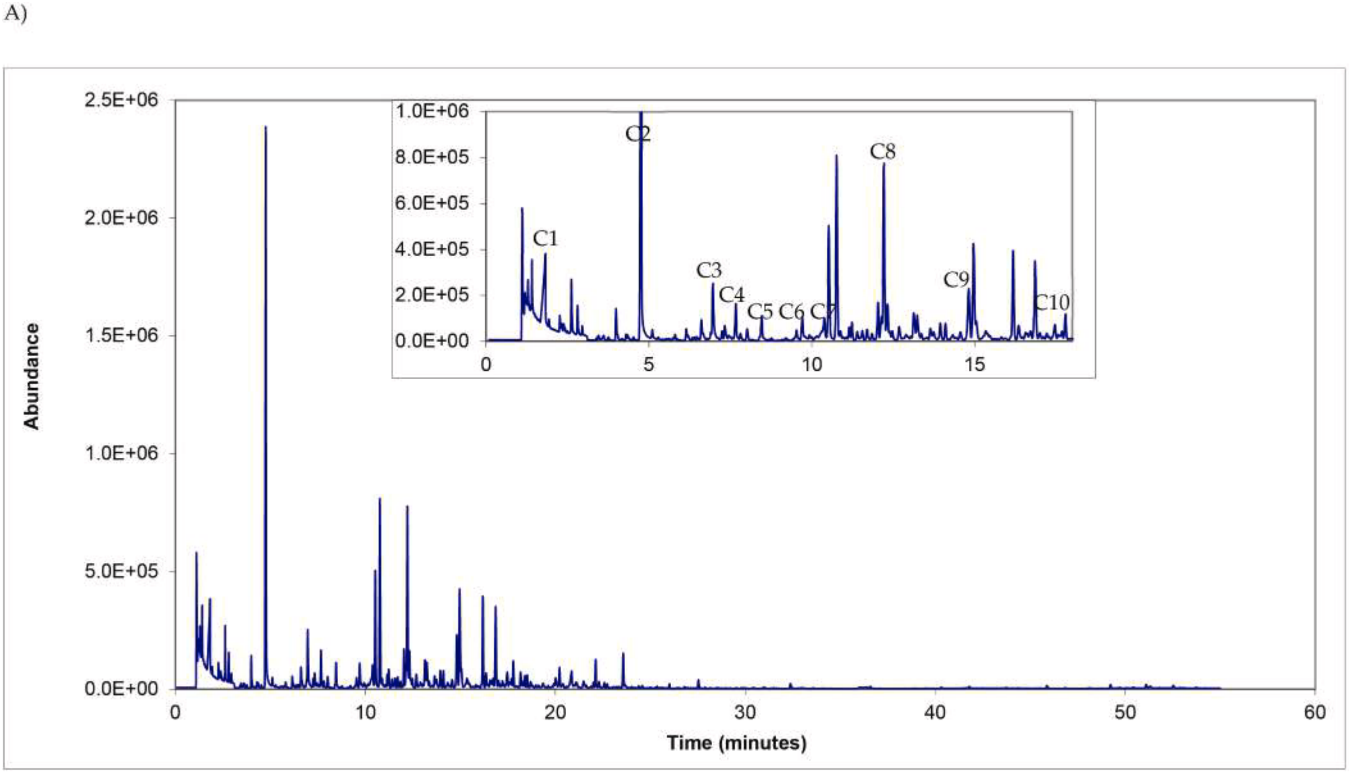
powerful tool to obtain more reliable results, providing additional information useful for a more precise understanding of the samples (Borràs et al., 2015). The purpose of using fusion strategies is to fully exploit the benefits of complementary information and overcome the shortcomings of individual techniques for exploration the geographical origin of gentian rhizomes. As shown in Fig. 4A, the first factor (Dim 1, 24 %) was highly related to NIRS data, to GCMS data and location as complementary data. The second factor (Dim 2, 18.4 %) was highly related to GCMS data and location as additional data. The contribution of NIRS data is the least useful group of variables for discriminating between samples on MFA Dim 2. The Fluo data point is in the middle of the map, this parameter contributes as much to Dim 1 as Dim 2, but at a lower level than the other two. Fig. 4B shows the score plot with 95 % confidence ellipses. It can be observed that there is a clear separation between the individual sites A, F, G and the merged group C and D. Fig. 4C shows each site considered by each group (Fluo, GCMS, NIRS) and its barycentre. For a given site, there are as many subpoints as there are groups of variables. The subpoints are given the colours used in the group plots. The red line represents how it is seen in terms of the "fluo" variable only. The green line shows how it is seen in relation to the "GCMS" variable only. This graph allows to see how the different groups influence the position of a given point.

These results showed that data fusion of HS-SPME-GCMS data, NIRS spectra and excitation-emission matrices of fluorescence could discriminate *Gentiana lutea* samples from different sites of the Massif Central, but this discrimination has its limitation because it is not possible to discriminate the geographical origin of sites C and D. This result is consistent with the literature (Shen et al., 2019) where the authors state that the Partial Least Squares Discriminant Analysis (PLS-DA) model could efficiently discriminate *Gentiana rigescens* from different geographical origins, but they could not be accurately determined for some samples. There are several possible explanations: (i) temperature, water and altitude factors are worth evaluating to systematically control plant quality (Mazina et al., 2015; Shen et al., 2019), (ii) chemical profiles of *Gentiana rigescens* were influenced by latitudinal gradients of production areas (Shen et al., 2019). We can assume that this is also true for *Gentiana lutea*, (iii) the environmental factors (geography, climate and soil) influence the content of bioactive compounds in *Gentiana* plants (J. Zhang et al., 2020).

To validate these conclusions, the experimental design was completed by adding the cultivated sites of the Massif Central, labelled B and E, to the wild sites previously used. The score plot of MFA with 95 % confidence ellipses is shown in Fig. 5. It can be seen that there is a clear separation between the two cultivated areas labelled B and E, and a clear separation between the wild areas A, F, G as before. On the other hand, samples C, D, E form a single group, which means that with the information available to us it is not possible to identify the origin of these three samples. As shown in **Fig. S1**, the samples from sites A, B, F and G are located in mountainous areas, while those from sites E and D are located in plains. This confirms that environmental factors (climate, altitude and soil type) must be taken into account to allow a certified geographical origin of gentian in a restricted area. Future work is needed on the effects of environmental factors and their interactions on the quality of *Gentiana lutea*, which could contribute to good cultivation practices and the conservation of wild populations.

3.2. Variability between massifs: Jura, Pyrénées, Alpes

In a limited area (only the Massif Central), according to our mid-level data fusion conditions, the geographical exploration of gentian roots is not certain. In order to validate whether this method is nevertheless promising, a new experimental design was studied on a broader scale, while maintaining the same objective: the exploration of the geographical origin of gentian rhizomes coming from different French massifs using mid-level data fusion of HS-SPME-GCMS data, NIRS spectra and excitation-emission matrices of fluorescence. The same methodology as described above was used. It included only samples representing the massifs of Jura, Alpes and Pyrénées. Due to the disproportionate number of gentian root sampling sites between the Massif Central and the other French massifs, the Massif Central was



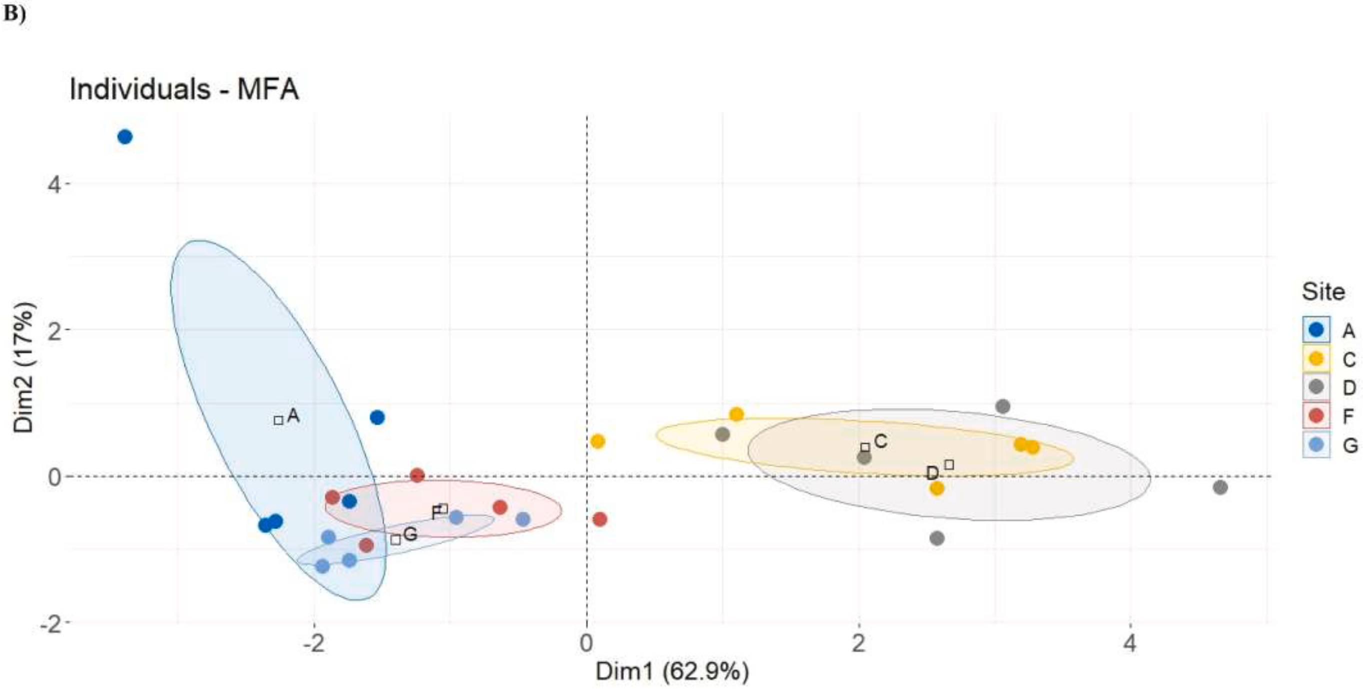


Fig. 3. A) An example of a chromatogram obtained by HS-SPME-GCMS from the wild site of massif central labelled A. C1 = acetic acid, C2 = 1-pentanol, C2 = hexanal, C3 = allyl isothiocyanate, C4 = heptanal, C5 = benzaldehyde, 2-methyl, C6 = benzaldehyde, C7 = hexanoic acid, C8 = limonene, C9 = linalool, C10 = 3-cyclohexen-1-ol, 4 methyl-1-(1-methylethyl)-; **B)** Individual factor map of the first two principal dimension (Dim1 – Dim 2) from the Multiple Factor Analysis (MFA) obtained from the Extracted-Ion Chromatogram (EIC) of all wild sites of massif central labelled A (in dark blue), C (in yellow), D (in grey), F (in red), G (in light blue) analyzed by GC–MS. **C)** Principal Component Analysis (PCA) obtained from the total ion current chromatogram (TIC) of all wild sites of massif central labelled A (in dark blue), C (in yellow), D (in grey), F (in red), G (in light blue) analyzed by GC–MS. PCA loading plot and score plot of the first two principal dimensions (Dim1 – Dim 2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

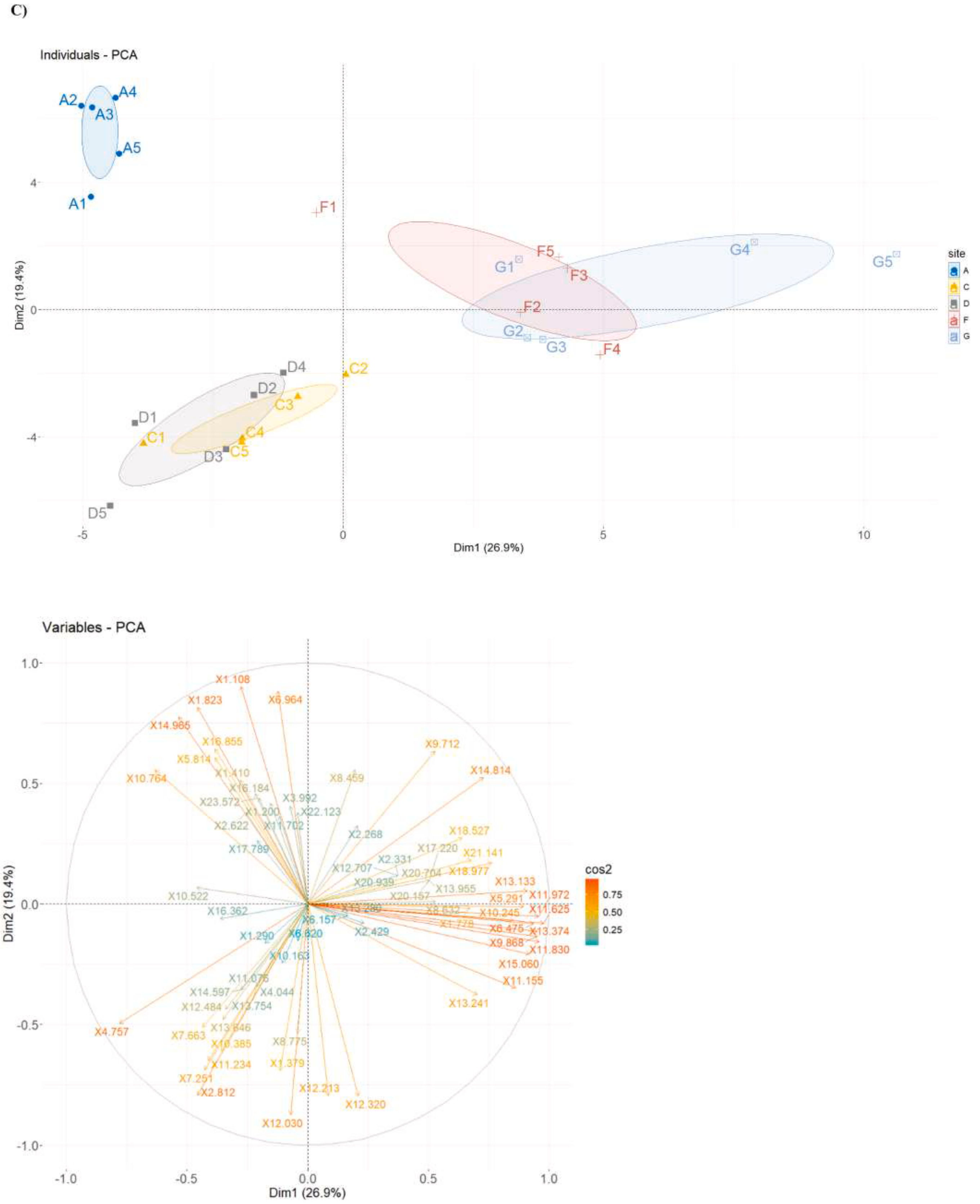


Fig. 3. (*continued*).

excluded from this new experimental design. In Fig. 6A the coordinates of the two active groups (NIRS and GCMS) on the first dimension are almost identical. This means that their contribution to the first dimension is similar (38.9 %). Regarding the second dimension (14.1 %), the fluo group has the highest contribution, indicating the highest contribution to the second dimension. Fig. 6B shows the score plot with

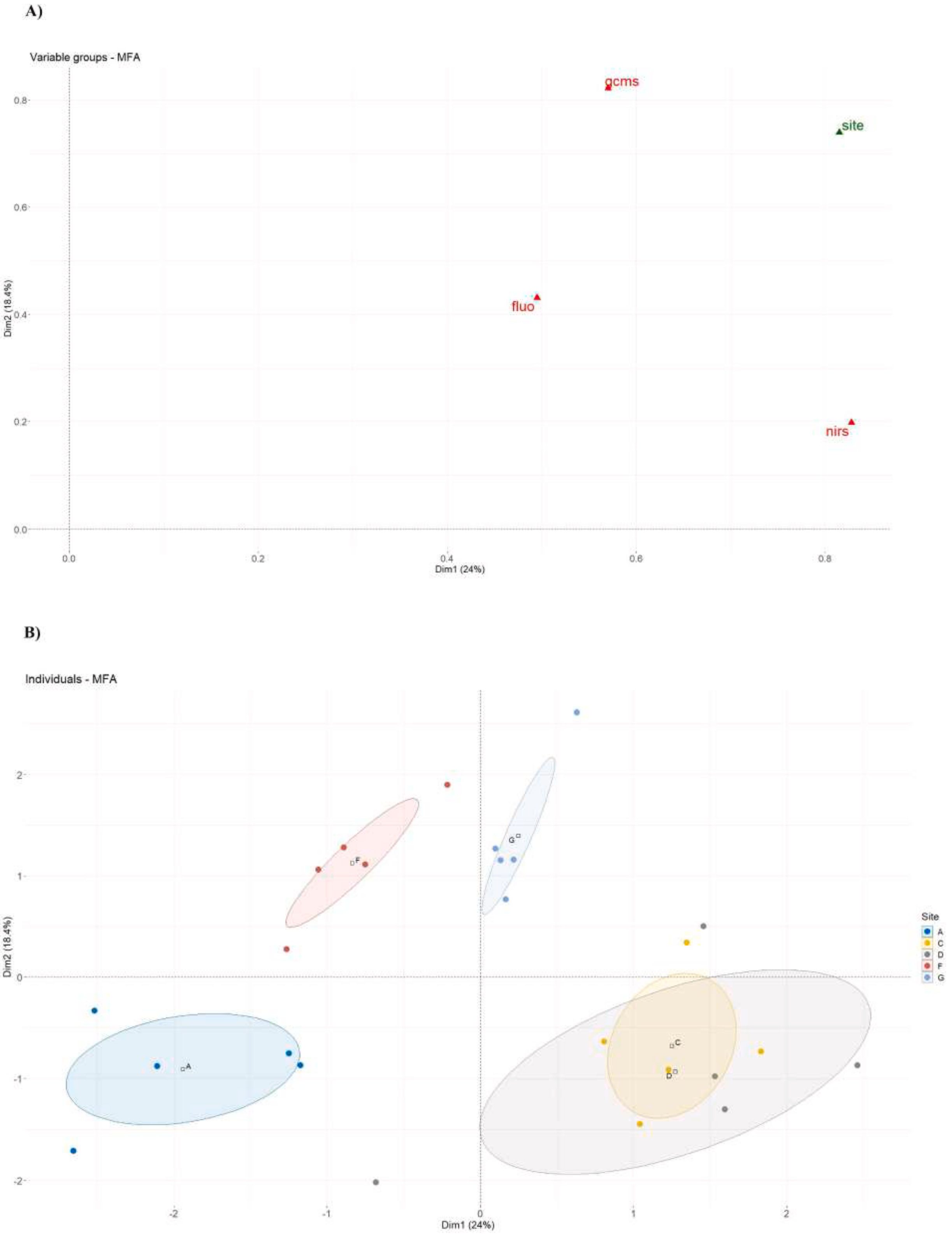


Fig. 4. Graphical representation of Multiple factors Analysis (MFA) of the first two principal dimensions (Dim1 – Dim 2). **A)** plots of the groups of variables. Correspondence of codes for group's representation: site (supplementary variable), fluo = Fluorescence analysis, gcms = HS-SPME-GCMS analysis, nirs = Near Infra-Red Spectroscopy analysis. **B)** Graph of individuals with 95 % confidence ellipse of all wild sites of massif central labelled A (in dark blue), C (in yellow), D (in grey), F (in red), G (in light blue). **C)** Graph of partial individuals seen by each variable. Correspondence of codes: fluo = Fluorescence analysis, gcms = HS-SPME-GCMS analysis, nirs = Near Infra-Red Spectroscopy analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

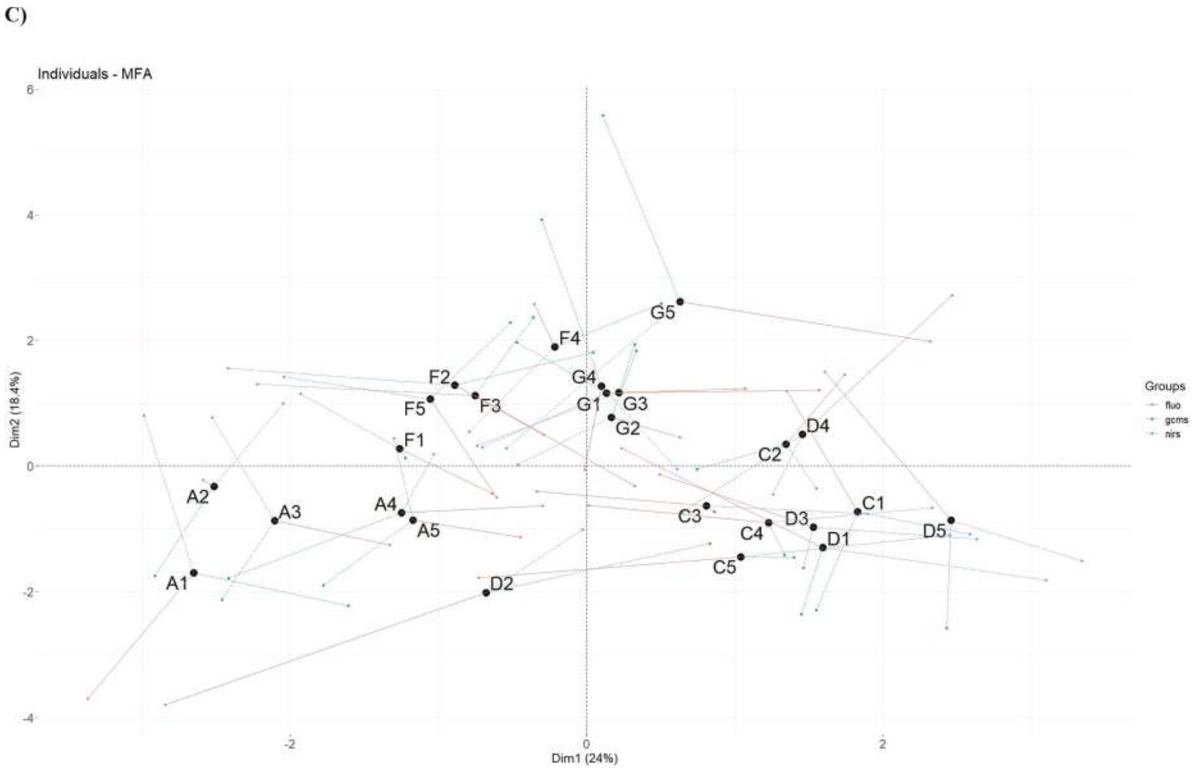


Fig. 4. (*continued*).

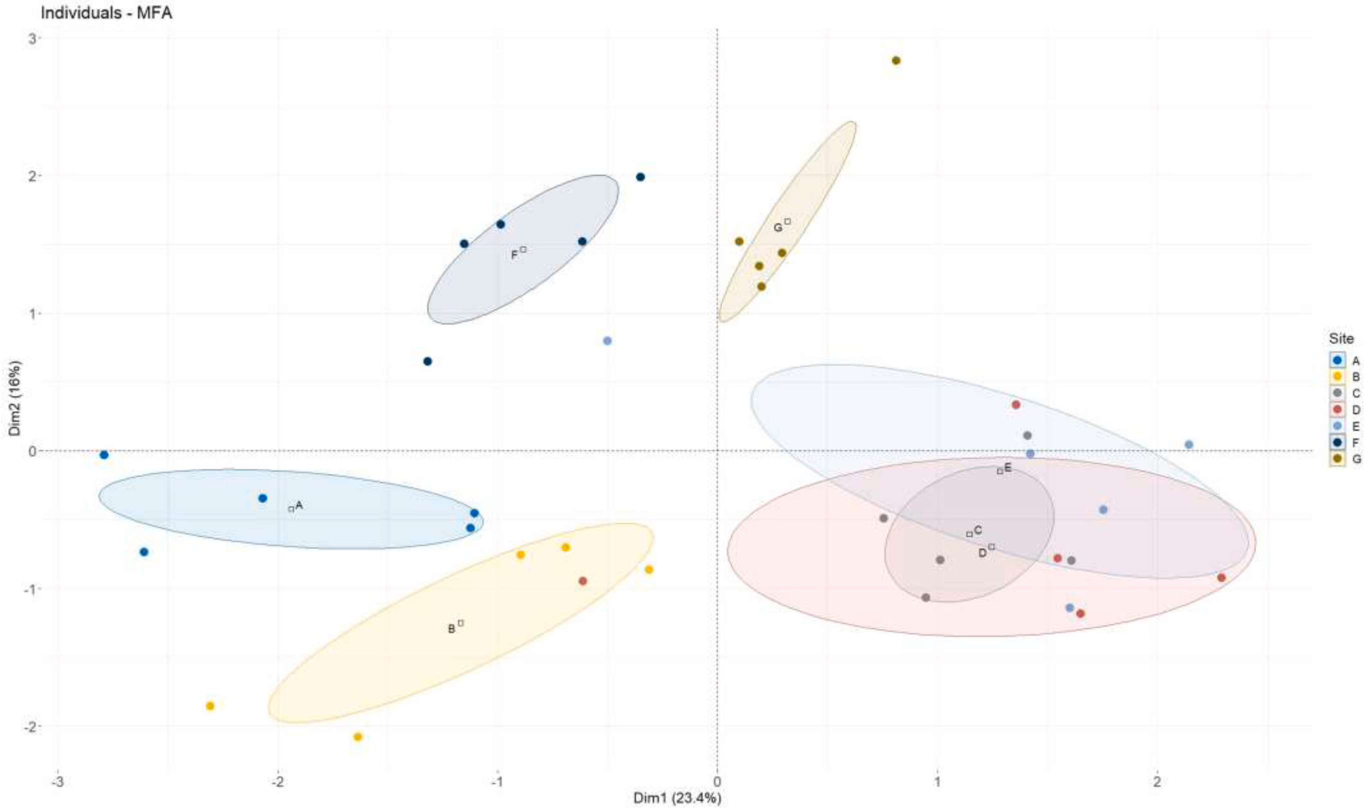


Fig. 5. Graphical representation of Multiple factors Analysis (MFA) of the first two principal dimensions (Dim1 – Dim 2): Graph of individuals with 95 % confidence ellipse of all sites of massif central (cultivated growing practices (B (in yellow), E (in light blue)) and wild growing A (in dark blue), C (in grey), D (in red), F (in black), G (in brown)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

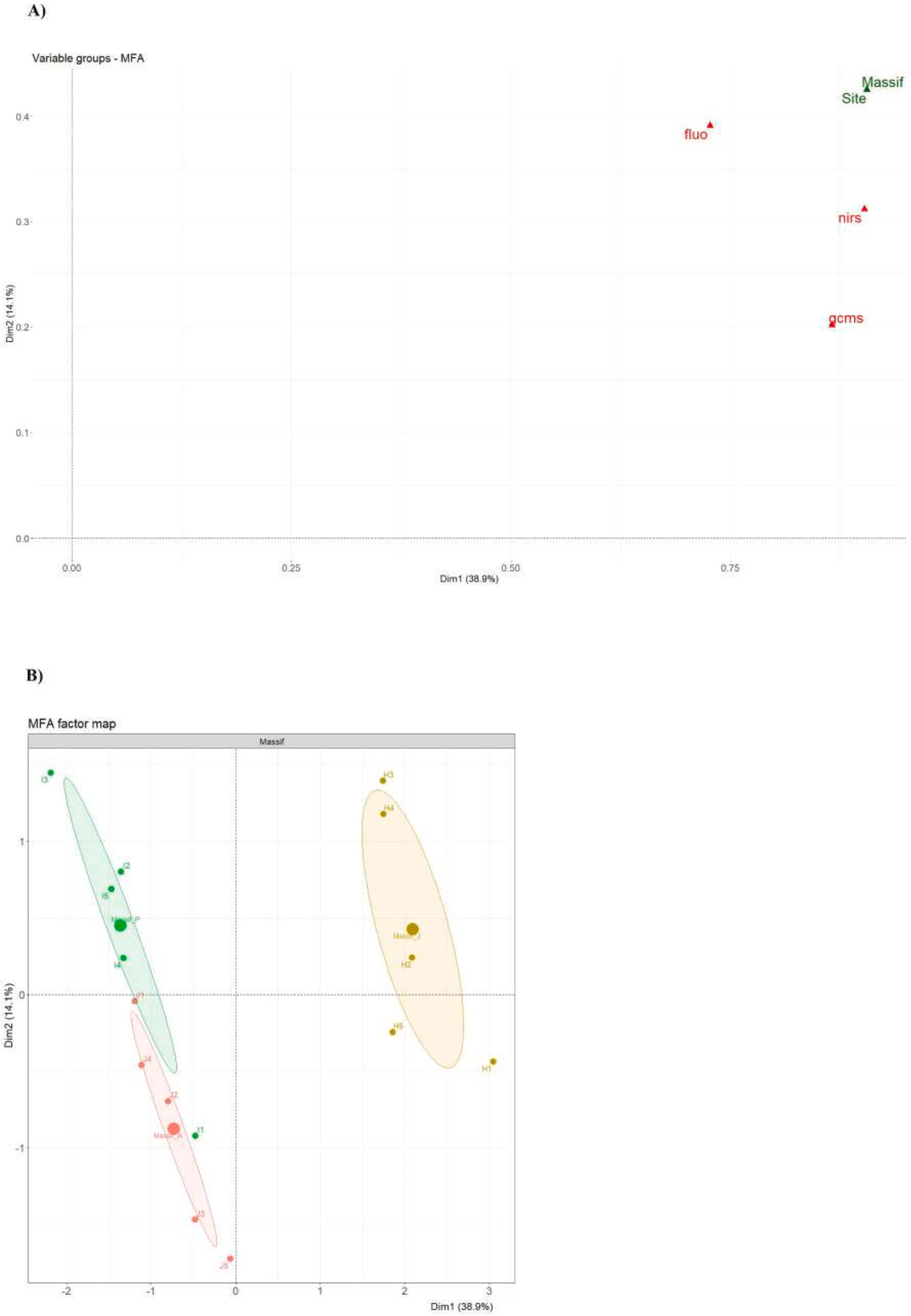


Fig. 6. Graphical representation of Multiple factors Analysis (MFA) of the first two principal dimensions (Dim1 – Dim 2). **A)** plots of the groups of variables. Correspondence of codes for group's representation: all sites of Jura, all sites of Alpes and all sites of Pyrénées (supplementary variable), fluo = Fluorescence analysis, gcms = HS-SPME-GCMS analysis, nirs = Near Infra-Red Spectroscopy analysis. **B)** Factor map according to the massif: Massif P = Pyrénées with wild site I (in green), Massif A = Alpes with wild site J (in yellow), Massif J = Jura with wild site H (in yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

95 % confidence ellipses. There is a clear separation between the massifs. The first axis contrasts the massifs of Jura (on the right) and Pyrénées, Alpes (on the left). On the MFA factor map, it is interesting to note that the separation is made according to the type of mountain: Alpes and Pyrénées are separate but close, while Jura is on the opposite side. Both are high mountains, while Jura is classified as medium mountains. The data fusion clearly shows that (i) it is possible to separate the different gentian samples according to the type of mountain (high or medium) by using a wider range and (ii) a mountainous environment can confer chemical typicity on the plant. This result confirms the previous conclusions: the environmental factors (geography, climate and soil) must be taken into account when analyzing samples from the same massif in a small area. According to the literature, the following factors should also be taken into account: genetic variability, life cycle, age of the plant (Mustafa et al., 2016).

It is interesting to note that the volatile compounds present differ according to the type of massif (high or medium mountain) and can be very reliable fingerprints (Fig. 7). The main retention times (10.764 heptane, 2,2,4,6,6-pentamethyl CAS N° [13475–82-6]) represented The Alpes and Pyrénées (high mountains). The common retention times

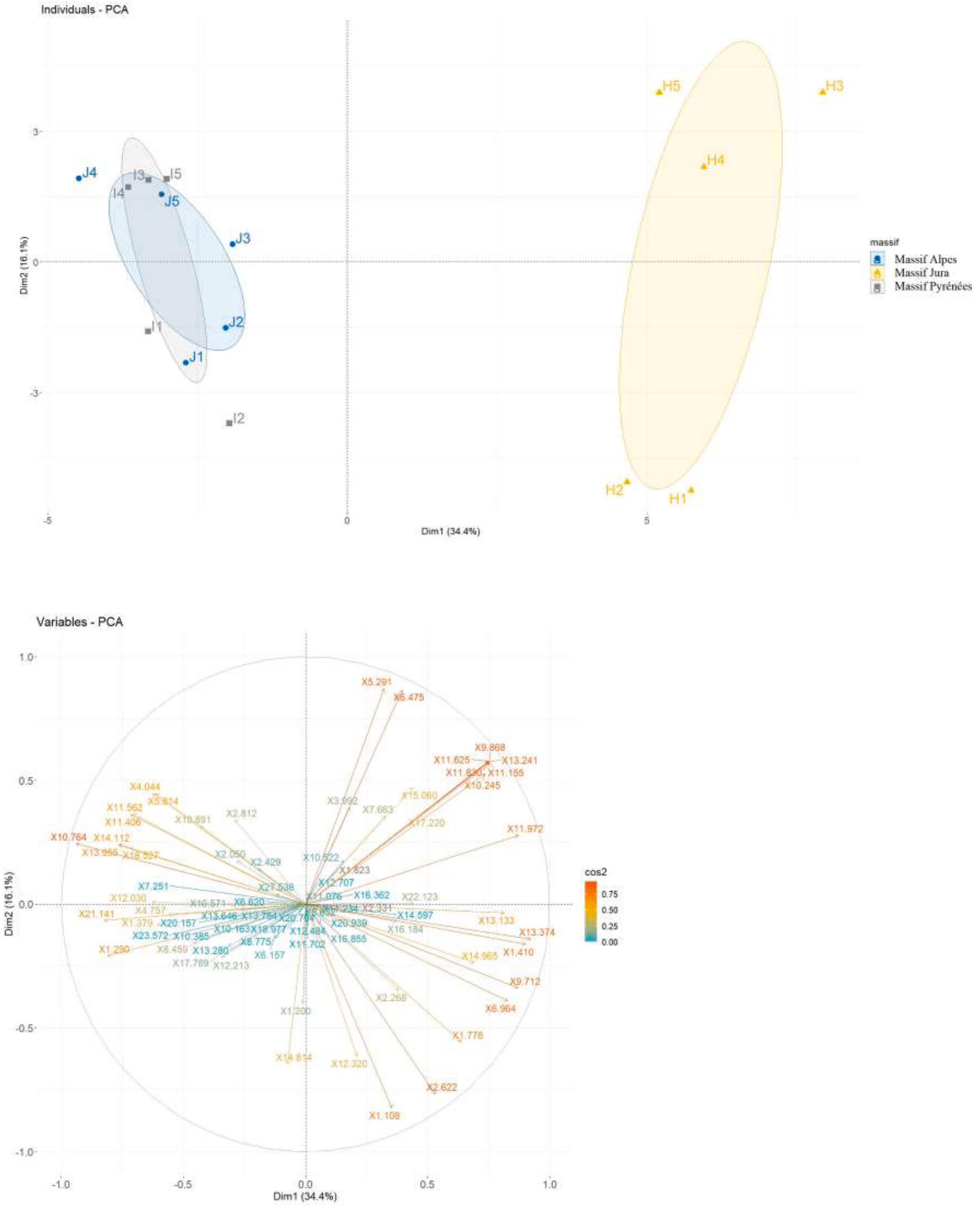


Fig. 7. Principal Component Analysis (PCA) obtained from the total ion current chromatogram (TIC) of all wild sites of the studies: wild site H of Jura (in yellow), wild site I of Pyrénées (in grey), wild site J of Alpes (in blue) analyzed by GC–MS. PCA loading plot and score plot of the first two principal dimensions (Dim1 – Dim 2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between Jura and Massif Central (middle mountains) are 1.11 min unknown molecule, 6.96 Allyl Isothiocyanate N° CAS [57–06-7], 11.97 Nonane, 2,6-dimethyl- CAS N° [17302–28-2], 11.63 unknown molecule, 6.475 Octane 4-methyl- CAS N° [2216-34-4], 13.37 Undecane 4,6-dimethyl- CAS N° [17312–82-2], 9.87 Nonane, 2-methyl- CAS N° [871–83-0], 11.83 Nonane, 2,5-dimethyl- CAS N° [017302–27-1], 11.16 unknown molecule. To go deeper, Jura and Massif Central have volatile compounds specific to their sites. The retention times (2.62 unknown molecule, 9.71 Benzaldehyde CAS N° [100–52-7], 1.41 Formic acid CAS N° [64–18-6], 13.24 Decane, 4-ethyl- CAS N° [1636-44-8], 5.29 Heptane, 2,4-dimethyl- CAS N° [2213−23−2]) represented the Jura. The retention times (1.82 Acetic acid CAS N° [64–19-7], 4.76 Hexanal CAS N° [66–25-1], 14.97 1-(2,4-Dimethyl-furan-3-yl)-ethanone 61 CAS N° [032933–07-6], 15.060 unknown molecule, 2.812 Pentanal CAS N° [110–62-3]) represented only the massif central (Fig. 3**)**. The results showed that the composition of the volatile fraction of *Gentiana lutea* is largely similar, but that each massif brings its own quite different specificity, its own volatile fingerprints.

4. Conclusion

In this study, the feasibility of combining information from HS-SPME-GCMS, NIRS and excitation-emission matrices of fluorescence was demonstrated using mid-level data fusion strategies to explore the geographical origin of *Gentiana lutea* roots over a wide area (between massifs). The results suggest that data fusion methods have the potential to be more effective in the quality of separation of studied sites of *Gentiana lutea* roots than individual decisions obtained from individual analytical tools. However, to ensure the geographical origin of *Gentiana lutea* root with these techniques at the massif scale, environmental factors (climate, altitude and soil type) must be taken into account to better ensure gentian authenticity. Exploratory data analysis showed that gentian volatile compounds, assessed by GCMS, can be used to fingerprint massif provenance. Gentian volatile markers were isolated and enabled to discriminate massif, and with a better consistency when data fusionning it with NIRS and fluorescence spectral data. Future work is needed on the effects of environmental factors and their interactions on the quality of *Gentiana lutea*, which could contribute to good growing practices and the conservation of wild populations. By highlighting the potential of mid-level data fusion techniques in this context, our research opens up avenues for further exploration to ensure the quality and conservation of other wild plant populations.

Funding statement

This work was supported by Agence Nationale de la Cohésion des Territoires (ANCT) and the Fonds national d'aménagement et de développement du territoire (FNADT).

CRediT authorship contribution statement

Céline Lafarge: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Laurence Dujourdy:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Gilles Figueredo:** Funding acquisition. **Stéphanie Flahaut:** Funding acquisition. **Laurent Rios:** Writing – review & editing. **Elias Bou-Maroun:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Christian Coelho:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors agreed with the published version of the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

We are very grateful to CPPARM and French *Gentiana lutea* association and to the farmers who allowed us to sample *Gentiana lutea* roots in their fields. We thank Vianney Bellegarde and Maxime Pacheco for their technical and analytical support. We thank the financial supports provided through the Fond National d'aménagement et de développement du territoire (FNADT) by the Agence Nationale de la Cohésion des Territoire (ANCT).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2024.141564.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs. Computational Statistics, 2*(4), 433–459. https://doi.org/10.1002/wics.101

Aberham, A., Pieri, V., Croom, E. M., Ellmerer, E., & Stuppner, H. (2011). Analysis of iridoids, secoiridoids and xanthones in Centaurium erythraea, Frasera caroliniensis and Gentiana lutea using LC-MS and RP-HPLC. *Journal of Pharmaceutical and Biomedical Analysis, 54*(3), 517–525. https://doi.org/10.1016/j.jpba.2010.09.030

Abraham, E. J., & Kellogg, J. J. (2021). Chemometric-guided approaches for profiling and authenticating botanical materials. *Frontiers in Nutrition, 8*, Article 780228. https://doi.org/10.3389/fnut.2021.780228

Ando, H., Hirai, Y., Fujii, M., Hori, Y., Fukumura, M., Niiho, Y., Nakajima, Y., Shibata, T., Toriizuka, K., & Ida, Y. (2007). The chemical constituents of fresh gentian root. *Journal of Natural Medicines, 61*(3), 269–279. https://doi.org/10.1007/s11418-007-0143-x

Arberas, I., Leiton, M. J., Domínguez, J. B., Bueno, J. M., Ariño, A., de Diego, E., … de Renobales, M. (1995). *The volatile flavor of fresh Gentiana lutea L. Roots. In G. Charalambous (Ed.), Developments in food science (Vol. 37, pp. 207–234). Elsevier.* https://doi.org/10.1016/S0167-4501(06)80158-5

Berthon, J.-Y., Cabannes, M., Bouton, C., Carre, M., Bridon, E., & Filaire, E. (2023). In vitro, ex vivo and clinical approaches to evaluate the potential effect of Gentiana lutea extract on skin. *International Journal of Cosmetic Science, 45*(5), 688–698. https://doi.org/10.1111/ics.12878

Biehlmann, M., Nazaryan, S., Krauss, E., Ardeza, M. I., Flahaut, S., Figueredo, G., … Coelho, C. (2020). How chemical and sensorial markers reflect gentian geographic origin in chardonnay wine macerated with Gentiana lutea roots? *Foods, 9*(8). *Article, 8.* https://doi.org/10.3390/foods9081061

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment – A review. *Analytica Chimica Acta, 891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Brereton, R. G. (2003). Signal Processing. In *Chemometrics* (pp. 119–181). John Wiley & Sons, Ltd.. https://doi.org/10.1002/0470863242.ch3

Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal, 2013*(1), Article 704504. https://doi.org/10.1155/2013/704504

Cevoli, C., Iaccheri, E., Fabbri, A., & Ragni, L. (2024). Data fusion of FT-NIR spectroscopy and Vis/NIR hyperspectral imaging to predict quality parameters of yellow flesh "Jintao" kiwifruit. *Biosystems Engineering, 237*, 157–169. https://doi.org/10.1016/j.biosystemseng.2023.12.011

Cocchi, M., & Reggio, E. (2019). Data fusion methodology and applications. https://shop.elsevier.com/books/data-fusion-methodology-and-applications/cocchi/978-0-444-63984-4.

Coelho, C., Bord, C., Fayolle, K., Bibang, C., & Flahaut, S. (2023). Development of a novel flavored goat cheese with Gentiana lutea rhizomes. *Foods, 12(3), Article 3.* https://doi.org/10.3390/foods12030468

Coelho, C., Figueredo, G., Lafarge, C., Bou-Maroun, E., & Flahaut, S. (2022). Mid-infrared spectroscopy combined with multivariate analysis and machine-learning: A powerful tool to simultaneously assess geographical origin, growing conditions and bitter content in *Gentiana lutea* roots. *Industrial Crops and Products, 187*, Article 115349. https://doi.org/10.1016/j.indcrop.2022.115349

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Cordella, C. B. Y. (2012). *PCA: The basic building block of Chemometrics*. In Analytical Chemistry: IntechOpen. https://doi.org/10.5772/51429

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*(71), 2349–2353.

Drivelos, S. A., Higgins, K., Kalivas, J. H., Haroutounian, S. A., & Georgiou, C. A. (2014). Data fusion for food authentication. Combining rare earth elements and trace metals to discriminate "fava Santorinis" from other yellow split peas using chemometric tools. *Food Chemistry, 165*, 316–322. https://doi.org/10.1016/j.foodchem.2014.03.083

Enders, A. A., North, N. M., Fensore, C. M., Velez-Alvarez, J., & Allen, H. C. (2021). Functional group identification for FTIR spectra using image-based machine learning models. *Analytical Chemistry, 93*(28), 9711–9718. https://doi.org/10.1021/acs.analchem.1c00867

European Medicines Agency (EMA). (2009). *Assessment report on Gentiana lutea L., radix. EMA/HMPC/607863/2017*.

Gibitz-Eisath, N., Seger, C., Schwaiger, S., Sturm, S., & Stuppner, H. (2022). Simultaneous quantitative analysis of the major bioactive compounds in Gentianae Radix and its beverages by UHPSFC–DAD. *Journal of Agricultural and Food Chemistry, 70*(24), 7586–7593. https://doi.org/10.1021/acs.jafc.2c01584

González-López, O., Polanco, C., György, Z., Pedryc, A., & Casquero, P. A. (2014). Genetic variation of the endangered Gentiana lutea L. var. Aurantiaca (Gentianaceae) in populations from the Northwest Iberian Peninsula. *International Journal of Molecular Sciences, 15(6), Article 6*. https://doi.org/10.3390/ijms150610052

He, H., Yan, S., Lyu, D., Xu, M., Ye, R., Zheng, P., Lu, X., Wang, L., & Ren, B. (2021). Deep learning for biospectroscopy and biospectral imaging: State-of-the-art and perspectives. *Analytical Chemistry, 93*(8), 3653–3665. https://doi.org/10.1021/acs.analchem.0c04671

Hoffman, P., Grinstein, G., & Pinkney, D. (1999). *Dimensional Anchors: A Graphic Primitive for Multidimensional Multivariate Information Visualizations., 16*. https://doi.org/10.1145/331770.331775

Husson, F., Le, S., & Pagès, J. (2017). Exploratory multivariate analysis by example using R (2nd ed.). *Chapman and Hall/CRC.*. https://doi.org/10.1201/b21874

Kassambara, A. (2024). *Rstatix: Pipe-friendly framework for basic statistical tests version 0.7.2 from CRAN* (0.7.2) [Computer software]. Retrieved March 18, 2024, from https://rdrr.io/cran/rstatix/.

Kotrlik, J. W., & Williams, H. A. (2003). *The incorporation of effect size in information technology*. Learning, and Performance Research.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*, 1–18. https://doi.org/10.18637/jss.v025.i01

Li, J. (2019). A critical review of spatial predictive modeling process in environmental sciences with reproducible examples in R. *Applied sciences, 9*(10). https://doi.org/10.3390/app9102048. article 10.

Li, Y., Zhang, J.-Y., & Wang, Y.-Z. (2018). FT-MIR and NIR spectral data fusion: A synergetic strategy for the geographical traceability of Panax notoginseng. *Analytical and Bioanalytical Chemistry, 410*(1), 91–103. https://doi.org/10.1007/s00216-017-0692-0

Mafata, M., Brand, J., Kidd, M., Medvedovici, A., & Buica, A. (2022). Exploration of data fusion strategies using principal component analysis and multiple factor analysis. *Beverages, 8*(4), Article 4. https://doi.org/10.3390/beverages8040066

Marković, T., Radanović, D., Nastasijević, B., Antić-Mladenović, S., Vasić, V., & Matković, A. (2019). Yield, quality and safety of yellow gentian roots produced under dry-farming conditions in various single basal fertilization and planting density models. *Industrial Crops and Products, 132*, 236–244. https://doi.org/10.1016/j.indcrop.2019.02.027

Márquez, C., López, M. I., Ruisánchez, I., & Callao, M. P. (2016). FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. *Talanta, 161*, 80–86. https://doi.org/10.1016/j.talanta.2016.08.003

Mazina, J., Vaher, M., Kuhtinskaja, M., Poryvkina, L., & Kaljurand, M. (2015). Fluorescence, electrophoretic and chromatographic fingerprints of herbal medicines and their comparative chemometric analysis. *Talanta, 139*, 233–246. https://doi.org/10.1016/j.talanta.2015.02.050

Mirzaee, F., Hosseini, A., Jouybari, H. B., Davoodi, A., & Azadbakht, M. (2017). Medicinal, biological and phytochemical properties of Gentiana species. *Journal of Traditional and Complementary Medicine, 7*(4), 400–408. https://doi.org/10.1016/j.jtcme.2016.12.013

Mishra, P., Roger, J.-M., Jouan-Rimbaud-Bouveresse, D., Biancolillo, A., Marini, F., Nordon, A., & Rutledge, D. N. (2021). Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends in Analytical Chemistry, 137*, Article 116206. https://doi.org/10.1016/j.trac.2021.116206

Mustafa, A. M., Caprioli, G., Maggi, F., Vittori, S., & Sagratini, G. (2016). Comparative analysis of the volatile profiles from wild, cultivated, and commercial roots of Gentiana lutea L. by headspace solid phase microextraction (HS–SPME) coupled to gas chromatography mass spectrometry (GC–MS). *Food Analytical Methods, 9*(2), 311–321. https://doi.org/10.1007/s12161-015-0196-5

Mustafa, A. M., Caprioli, G., Ricciutelli, M., Maggi, F., Marín, R., Vittori, S., & Sagratini, G. (2015). Comparative HPLC/ESI-MS and HPLC/DAD study of different populations of cultivated, wild and commercial Gentiana lutea L. *Food Chemistry, 174*, 426–433. https://doi.org/10.1016/j.foodchem.2014.11.089

Obeidat, S., Glasser, T., Landau, S., Anderson, D., & Rayson, G. (2007). Application of multi-way data analysis on excitation-emission spectra for plant identification. *Talanta, 72*(2). https://doi.org/10.1016/j.talanta.2006.11.045

Ottavian, M., Fasolato, L., Serva, L., Facco, P., & Barolo, M. (2014). Data fusion for food authentication: Fresh/frozen–thawed discrimination in west African goatfish (Pseudupeneus prayensis) fillets. *Food and Bioprocess Technology, 7*(4), 1025–1036. https://doi.org/10.1007/s11947-013-1157-x

Ozaki, Y., Morita, S., & Du, Y. (2006). Spectral analysis. In *Near-infrared spectroscopy in food science and technology* (pp. 47–72). John Wiley & Sons, Ltd.. https://doi.org/10.1002/9780470047705.ch3

Pagès, J. (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée, 50*(4), 5–37.

Pérez-Ràfols, C., Serrano, N., & Díaz-Cruz, J. M. (2023). Authentication of soothing herbs by UV–vis spectroscopic and chromatographic data fusion strategy. *Chemometrics and Intelligent Laboratory Systems, 235*, Article 104783. https://doi.org/10.1016/j.chemolab.2023.104783

Ponticelli, M., Lela, L., Moles, M., Mangieri, C., Bisaccia, D., Faraone, I., Falabella, R., & Milella, L. (2023). The healing bitterness of Gentiana lutea L., phytochemistry and biological activities: A systematic review. *Phytochemistry, 206*, Article 113518. https://doi.org/10.1016/j.phytochem.2022.113518

R Core Team. (2018). *R: A Language and Environment for Statistical Computing* [Computer software]. https://www.R-project.org.

Reyrolle, M., Bareille, G., Epova, E. N., Barre, J., Bérail, S., Pigot, T., … Le Bechec, M. (2023). Authenticating teas using multielement signatures, strontium isotope ratios, and volatile compound profiling. *Food Chemistry, 423*, Article 136271. https://doi.org/10.1016/j.foodchem.2023.136271

Reyrolle, M., Ghislain, M., Bru, N., Vallverdu, G., Pigot, T., Desauziers, V., & Le Bechec, M. (2022). Volatile fingerprint of food products with untargeted SIFT-MS data coupled with mixOmics methods for profile discrimination: Application case on cheese. *Food Chemistry, 369*, Article 130801. https://doi.org/10.1016/j.foodchem.2021.130801

Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry, 28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Ríos-Reina, R., Callejón, R. M., Savorani, F., Amigo, J. M., & Cocchi, M. (2019). Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta, 198*, 560–572. https://doi.org/10.1016/j.talanta.2019.01.100

Robert, C., Jessep, W., Sutton, J. J., Hicks, T. M., Loeffen, M., Farouk, M., … Gordon, K. C. (2021). Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat. *Food Chemistry, 361*, Article 130154. https://doi.org/10.1016/j.foodchem.2021.130154

Schwolow, S., Gerhardt, N., Rohn, S., & Weller, P. (2019). Data fusion of GC-IMS data and FT-MIR spectra for the authentication of olive oils and honeys-is it worth to go the extra mile? *Analytical and Bioanalytical Chemistry, 411*(23), 6005–6019. https://doi.org/10.1007/s00216-019-01978-w

Shen, T., Yu, H., & Wang, Y.-Z. (2019). Assessing geographical origin of Gentiana Rigescens using untargeted chromatographic fingerprint, data fusion and Chemometrics. *Molecules, 24*(14), Article 14. https://doi.org/10.3390/molecules24142562

Shen, T., Yu, H., & Wang, Y.-Z. (2020). Discrimination of Gentiana and its related species using IR spectroscopy combined with feature selection and stacked generalization. *Molecules, 25(6), Article 6*. https://doi.org/10.3390/molecules25061442

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences, 1*(21), 19–25. https://api.semanticscholar.org/CorpusID:73706075.

Toplak, M., Read, S. T., Sandt, C., & Borondics, F. (2021). Quasar: Easy machine learning for biospectroscopy. *Cells, 10*(9). https://doi.org/10.3390/cells10092300. article 9.

Toriumi, Y., Kakuda, R., Kikuchi, M., Yaoita, Y., & Kikuchi, M. (2003). New triterpenoids from Gentiana lutea. *Chemical & Pharmaceutical Bulletin, 51*(1), 89–91. https://doi.org/10.1248/cpb.51.89

Veiga, T., Guitián, J., Guitián, P., Guitián, J., Munilla, I., & Sobral, M. (2016). Flower colour variation in the montane plant Gentiana lutea L. (Gentianaceae) is unrelated to abiotic factors. *Plant Ecology & Diversity, 9*(1), 105–112. https://doi.org/10.1080/17550874.2015.1074626

Wang, Q.-Q., Huang, H.-Y., & Wang, Y.-Z. (2019). Geographical authentication of Macrohyporia cocos by a data fusion method combining ultra-fast liquid chromatography and Fourier transform infrared spectroscopy. *Molecules, 24(7), Article 7*. https://doi.org/10.3390/molecules24071320

Wang, Y., Huang, H.-Y., Zuo, Z.-T., & Wang, Y.-Z. (2018). Comprehensive quality assessment of Dendrubium officinale using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 205*, 637–648. https://doi.org/10.1016/j.saa.2018.07.086

Xu, Y., Li, Y., Maffucci, G., & K., Huang, L., & Zeng, R.. (2017). Analytical methods of phytochemicals from the genus Gentiana. *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry, 22*(12), 2080. https://doi.org/10.3390/molecules22122080

Zhang, J., Zhang, Z., Wang, Y., Zuo, Y., & Cai, C. (2020). Environmental impact on the variability in quality of *Gentiana rigescens*, a medicinal plant in Southwest China. *Global Ecology and Conservation, 24*, Article e01374. https://doi.org/10.1016/j.gecco.2020.e01374

Zhang, Y., & Wang, Y. (2023). Recent trends of machine learning applied to multi-source data of medicinal plants. *Journal of Pharmaceutical Analysis, 13*(12), 1388–1407. https://doi.org/10.1016/j.jpha.2023.07.012