N ° d'ordre 42571 Année 2017

Université de Lille - Sciences et Technologies

Ecole Doctorale des Sciences de la matière, du Rayonnement et de l'Environnement

THESE

Présentée pour l'obtention du titre de

DOCTEUR DE L'Université de Lille

Discipline : Chimie théorique, physique, analytique

Par

Damian SIEPKA

Development of multidimensional spectral data processing procedures for analysis of composition and mixing state of aerosol particles by Raman and FTIR spectroscopy

Soutenue à l'Institut des Sciences Moléculaires, Bordeaux, le 20 décembre 2017
devant la commission d'examen :

Rapporteurs:

Pr. Laurent Servant, Professeur de l'Université de Bordeaux, Institut des Sciences Moléculaires, UMR 5255

Pr Nathalie Dupuy – Professeur de l'Université d'Aix Marseille, Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE)

Examinateurs:

Dr Jean Marc Bonmatin, chercheur CNRS, Centre de Biophysique Moléculaire, UPR 4301, Orléans.

Dr Aurélien Moncomble, MCF de l'Université de Lille, Laboratoire de Spectrochimie Infrarouge et Raman UMR 8516

Directrices de thèse :

Dr Sophie Sobanska, Chercheur CNRS – HDR, Institut des Sciences Moléculaires, UMR 5255, Bordeaux.

Pr hab. Elzbieta A. Stefaniak, Professeur de l'Université Catholique de Lublin Jean-Paul II, Centre de Recherche Interdisciplinaire, Lublin, Pologne.

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

Development of multidimensional spectral data processing procedures for analysis of composition and mixing state of aerosol particles by Raman and FTIR spectroscopy

Abstract

Sufficiently adjusted, multivariate data processing methods and procedures can significantly improve the process for obtaining knowledge of a sample composition. Spectroscopic techniques have nowadays capabilities for fast analysis of various samples. Indeed, much type of measurements was developed for research and industrial purposes and offer a huge possibility for advanced molecular analysis of complex samples, where atmospheric aerosol particles are a perfect example. Airborne particles affect air quality and successively, human and ecosystem condition, playing an important role in the Earth's climate system. The purpose of this thesis is twofold. On an analytical level, the functional algorithm for specification of quantitative composition of atmospheric particles by Raman microspectrocopy (RMS) from single particle analysis was established. On a constructive level, the readily accessible analytical system for Raman and FTIR data processing was developed. Considering these aims in more detail: Firstly, the potential of single particle analysis by RMS has been exploited by application of the designed analytical algorithm for an efficient description of chemical mixing of aerosol particles. The algorithm was applied to experimental data, exceeding the limitations in trace constituent detection and quantitative analysis, as well as providing a new way of sample description. Secondly, the new software which includes the described algorithm and several easy-to access, powerful data processing techniques was developed. Moreover, the created software features were applied for some challenging aspects of pattern recognition in the scope of Raman and FTIR spectroscopy.

© 2017 Tous droits réservés.

Élaboration de procédures de traitement des données spectrales multidimensionnelles pour l'analyse de la composition et de l'état de mélange d'aérosols atmosphériques par spectroscopie Raman et IFTR

Résumé

Les méthodologies de traitement de données multidimensionnelles peuvent considérablement améliorer la connaissance des échantillons. Les techniques spectroscopiques permettent l'analyse moléculaire avancée d'échantillons variés et complexes. La combinaison des techniques spectroscopiques aux méthodes de chimiométrie trouve des applications dans de nombreux domaines. Les particules atmosphériques affectent la qualité de l'air, la santé humaine, les écosystèmes et jouent un rôle important dans le processus de changement climatique. L'objectif de cette thèse a été de développer des outils de chimiométrie, simples d'utilisation, permettant de traiter un grand nombre de données spectrales provenant de l'analyse d'échantillons complexes par microspectrométrie Raman (RMS) et spectroscopie d'absorption IRTF. Dans un premier temps, nous avons développé une méthodologie combinant les méthodes de résolution de courbes et d'analyse multivariée afin de déterminer la composition chimique d'échantillons de particules analysées par RMS. Cette méthode appliquée à l'analyse de particules collectées dans les mines en Bolivie, a ouvert une nouvelle voie de description des échantillons. Dans un second temps, nous avons conçu un logiciel facilement accessible pour le traitement des données IRTF et Raman. Ce logiciel inclue plusieurs algorithmes de prétraitement ainsi que les méthodes d'analyse multivariées adaptées à la spectroscopie vibrationnelle. Il a été appliqué avec succès pour le traitement de données spectrales enregistrées pour divers échantillons (particules de mines de charbon, particules biogéniques, pigments organiques).

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

Abbreviations and symbols

ALS	Alternating Least Square
AsLS	Asymmetric Least Squares Baseline Correction
CA	Cluster Analysis
EPMA	Electron Probe X-ray Microanalysis
EDS	Energy-Dispersive X-ray Spectroscopy
FTIR	Fourier-Transform Infrared Spectroscopy
HCA	Hierarchical Cluster Analysis
MCR	Multivariate Curve Resolution
MSC	Multiplicative Scatter Correction
MLR	Multiple Linear Regression
MR	Mixing Rank
PARAFAC	Parallel Factor Analysis
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares Regression
RMS	Raman Microspectroscopy
RRSSQ	Relative Root of Sum of Square Difference
SEM	Scanning Electron Microscopy
SERS	Surface-enhanced Raman Spectroscopy
SIMPLISMA	Simple-to-use Interactive Self-Modelling Mixture Analysis
SNV	Standard Normal Variate
SOA	Secondary Organic Aerosol
SPA	Single Particle Analysis
TERS	Tip-enhanced Raman Spectroscopy

© 2017 Tous droits réservés.

Acknowledgments:

Firstly, I would like to express my sincere gratitude to my advisor Dr. Sophie Sobanska for the continuous support of my Ph.D. study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Laurent Servant, Prof. Nathalie Dupuy, Dr J.M. Bonmatin and Dr. Aurelien Moncomble, to accept to evaluate this work, which incited me to widen my research from various perspectives.

My sincere thanks also go to Prof. Elżbieta Anna Stefaniak, who provided me an opportunity to join her team, and who gave access to the laboratory and research facilities. Without her precious support, it would not be possible to conduct this research.

Last but not the least, I would like to thank my fiancé, my family: my parents and to my brother for supporting me spiritually throughout writing this thesis and my life in general.

© 2017 Tous droits réservés.

Andrzej Sapkowski

"It's better to die than to live in the knowledge that you've done something that needs forgiveness." The Witcher, Blood of Elves.

Louis Pasteur

"*La science ne connaît pas de pays, parce que la connaissance appartient à l'humanité, et elle est la torche qui illumine le monde.*"

Confucius

"*Our greatest glory is not in never falling, but in rising every time we fall.*"

Carlos Ruíz Zafón

"People tend to complicate their own lives, as if living wasn't already complicated enough." The Shadow of the Wind

Antoni Kępiński

"*Nauka jest przeciwstawna mądrości, mądrość bowiem nie dąży do władzy, ale do większego zbliżenia się i tym samym lepszego zrozumienia świata otaczającego, wczucia się w jego tajemny rytm.*"

© 2017 Tous droits réservés.

Table of contents

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

INTRODUCTION

© 2017 Tous droits réservés.

The application of optimal mathematical and statistical data processing in chemistry is an integral part of chemometrics. Sufficiently adjusted chemometric methods and procedures can significantly improve the process of resolving a sample composition (Geladi 2003a). The amount of data collected during chemical analyses has outgrown the prime conception of the analytical chemistry (Ziegel 2004). From a single run, it is now possible in even few seconds to obtain large amounts of data, which generally require further processing. Predominantly, those data can consist of two constituents: informative and non-informative data. The acquired data do not have any value without an appropriate processing, only a prerequisite for information (Massart et al. 1997). Spectroscopic techniques have nowadays capabilities for fast analysis of various samples. Indeed, many types of measurements – including on line, *in situ*, spatially resolved, time resolved, imaging, portable, etc. – were developed for research and industrial purposes and offer a huge possibility for advanced molecular analysis of complex samples, such as atmospheric aerosol particles (Ault & Axson 2017; Cochran et al. 2017; Sobanska et al. 2014), live cells (Smith et al. 2016; Bergholt et al. 2017; Li et al. 2015), food and pharmaceutical nanomaterials (Li & Church 2014; Zong et al. 2013; Hong et al. 2010), pharmaceutical products (Y. Li et al. 2016; Dymińska 2015; Craig et al. 2015), environmental nanoparticles (Guo et al. 2017; Alessi et al. 2013; Tang & Lo 2013), works of art and archeological artifacts (Bersani et al. 2016; Otero et al. 2014; Leona et al. 2011).

Atmospheric aerosols affect air quality and – successively – humans and ecosystem,, playing an important role in the Earth's climate system (Hartmann et al. 2013; Mcneill 2017; IPCC 2014). Moreover, the European Commission's Thematic Strategy on Air Pollution (CEP Thematic Strategy on Air Pollution), which was developed as a long-term, strategic and integrated policy, advises to protect against significant negative effects of atmospheric aerosol particles on human health and the environment (European·Union 2005). The long- and short-term exposure to ambient aerosol particles is associated with an increase in mortality. In addition, there is a strong evidence that ambient aerosol particles impact respiratory and cardiovascular health

© 2017 Tous droits réservés.

and contribute to a lung cancer risk (Kim et al. 2015). Various satellite remote sensing instruments have been extensively used to study aerosol properties in global or regional coverage (Mcneill 2017). Despite their global-scale impacts, there are still many analytical challenges towards understanding their molecular composition, surface chemistry, heterogeneous reactivity and optical properties (Buajarern et al. 2007; Ault & Axson 2017). Knowledge about the chemical composition, morphology, size and internal structure of the particles can provide insight into their physical, chemical and optical properties which, in turn, is crucial to evaluate their main adverse effects (Sobanska et al. 2012; Carvalho-Oliveira et al. 2015; Jimoda 2012). For this purpose, the laboratory analysis of collected particles was considered as appropriate (Laskin et al. 2016).

The approved and well-received method for determination of chemical composition is bulk analysis of filter-collected particles (McMurry 2002; Kulkarni et al. 2011). The main advantages of this type of analysis are: identification of main components (chemical elements, compounds or ions), a fast and validated analysis, a user-friendly statistical data treatment, etc (Li et al. 2016). Many efforts have been made for determination of chemical mixing and internal structure composition of individual particles as recently reviewed by Ault & Axson (2017). Indeed, single particle analysis techniques were developed and introduced to study chemical mixing and surface properties of individual particles. These analytical methods specifically focus on compositions and sizes of single aerosol particles. For this purpose, offline spectroscopic techniques coupled with microscopy (optical or electronic), have a great potential (Li et al. 2016). The atmospheric aerosol is currently a subject of extensive research because unravelling physical and chemical properties of aerosols requires an advanced study at the individual particle scale (Krieger et al. 2012), although there are some meaningful limitations in the field of single particle analysis by microspectroscopic techniques (Laskin et al. 2016).

© 2017 Tous droits réservés.

Much of the information acquired from the single particle analysis techniques is essential for the specification of the particle molecular composition, as well as processes where chemical mixing is important, such as heterogeneous reactivity, liquid-liquid phase separations, water uptake/hygroscopic growth and ice nucleation (Liu et al. 2008; Ciobanu et al. 2009; Baustian et al. 2012; Ault et al. 2014; Laskin et al. 2015). However, due to a complex nature of aerosol particles, and an amount of acquired data, an advanced chemometric analysis is required (Ofner, Katharina A Kamilli, et al. 2015). The efficient quantitative analysis of atmospheric aerosol particles by vibrational microspectroscopic techniques is limited due to a lack of dedicated statistical tools. There is a significant scientific gap in the field of analytical algorithms dedicated to single particle analysis, which are able to exceed such limitations. In the field of SPA, there is a constant necessity for a readily accessible, open-source, integrated data analysis system composed of well-established chemometric methods for fast and reproducible processing, especially dedicated to single particle analysis.

In this context, the purpose of this thesis is twofold. On an analytical level, the functional algorithm for specification of quantitative composition of atmospheric particles from single particle analysis was established. On a constructive level, the readily accessible analytical system for Raman and FTIR data processing was investigated.

To discuss these aims in detail, it must be emphasized that:

1) The potential of single particle analysis by Raman microspectroscopy has been exploited by application of the originally designed analytical algorithm for an efficient description of chemical mixing of aerosol particles. The algorithm was applied to experimental data, exceeding the limitations in trace constituent detection and quantitative analysis, as well as providing a new way of a sample description. Additionally, the important aspect of suitable measurement conditions, such as particle collecting substrate, was evaluated.

© 2017 Tous droits réservés.

2) A software, which includes the described algorithm and several easy-to access, powerful chemometric techniques, was developed to facilitate a reproducibility of the data processing. Moreover, the methodology was applied to some aspects of pattern recognition in the scope of Raman and FTIR spectroscopy.

The thesis outline is presented as follow:

Chapter 1 gives a general background on particle in the atmosphere, single particle analysis with a special emphasis on multivariate statistics and a blind source separation problem in the analysis of individual aerosol particles. A brief review of chemometric methodology applied for spectroscopy is given.

Chapter 2 is dedicated to the description of samples and analytical methods used in this work.

Chapter 3 is focused on the characterization of individual particles by combination of Raman microspectrometry and chemometric methods including the evaluation of substrate for imaging of collected particles and the designed analytical algorithm for processing of Raman spectra collected by single particle analysis to recognize chemical mixing and quantitative results of aerosol particles. The analytical algorithm is applied to the analysis of aerosol particles collected in the mining environment.

Chapter 4 provides a description of the integrated software system for processing, analyzing, and clustering of Raman and FTIR spectra, named Spectronomy.

Chapter 5 is related to the application of Spectronomy system for data processing of Raman and FTIR spectra, especially in preprocessing, pattern recognition, sparsity boosting and dimensionality reduction. The focus of the application is on the noteworthy paradigms from the field of industrial and biogenic aerosol particle analysis, as well as non-destructive microanalysis of cultural heritage materials.

Conclusions and perspectives are discussed in the last part.

© 2017 Tous droits réservés.

CHAPTER 1 – STATE OF THE ART

© 2017 Tous droits réservés.

This chapter is divided in two parts. In the first part, a brief overview about single particle analysis techniques is presented, while the second part is focused on common chemometric methods applied to a spectral data treatment.

1.1. Particles in the atmosphere

Atmospheric aerosols are generally considered to be particles that range in size from a few nanometers (nm) to tens of micrometers (µm) in diameter. Particles may be either directly emitted into the atmosphere or formed there by the oxidation of precursor gases, such as sulphur dioxide, nitrogen oxides and volatile organic compounds (VOCs), where the resulting oxidation products nucleate to form new particles or condense on pre-existing ones. Particles formed through these two routes are referred to as primary and secondary particles, respectively (Finlayson-Pitts and Pitts, 1997; Seinfeld and Pandis, 1998). Particles in the atmosphere arise from natural sources as well as anthropogenic activities [Seinfeld and Pandis, 1998]. The former source includes windborne dust, sea spray, volcanic activities and biomass burning, while emissions of particles attributable to the activities of humans arise primarily from four source categories: fuel combustion, industrial processes, nonindustrial fugitive sources (e.g. construction work), and transportation sources (e.g. automobiles). Natural aerosols are usually 4 to 5 times larger than anthropogenic ones on a global scale, but regional variations in man-made pollution may change this ratio significantly in certain areas, particularly in the industrialized Northern Hemisphere (Seinfeld and Pandis, 1998).

The data on chemical composition, size, morphology, internal mixing and physical states of particles obtained by offline analytical methods are crucial for understanding aerosol formation and reaction mechanisms, their atmospheric evolution, their impacts and source apportionment (Prather et al. 2008; Claudio et al. 2017; Mcneill 2017).

© 2017 Tous droits réservés.

Aerosol particles cover a wide size range of more than five orders of magnitude, with diameters ranging from $5{\cdot}10^{-3}$ to 2.5 µm for fine particles and greater than 2.5 µm for coarse particles (Hinds 1999). The fine particles include both (i) the Aitken nuclei, mainly ranging from $5{\cdot}10^{-3}$ to $5 \cdot 10^{-2}$ µm, and (ii) the accumulation mode particles ranging from $5 \cdot 10^{-2}$ to about 2 µm. In this classification, it is worth mentioning that (i) the nuclei constitute the most important part of the ultrafine particles ($<10^{-1}$ µm) and (ii) the accumulation mode particles are mainly generated through coagulation of small particles from the nuclei class and condensation of vapors on existing particles. Consequently, the number of particles within the presented size subrange increases, and the accumulation mode becomes gradually more evident (Fig. 1).

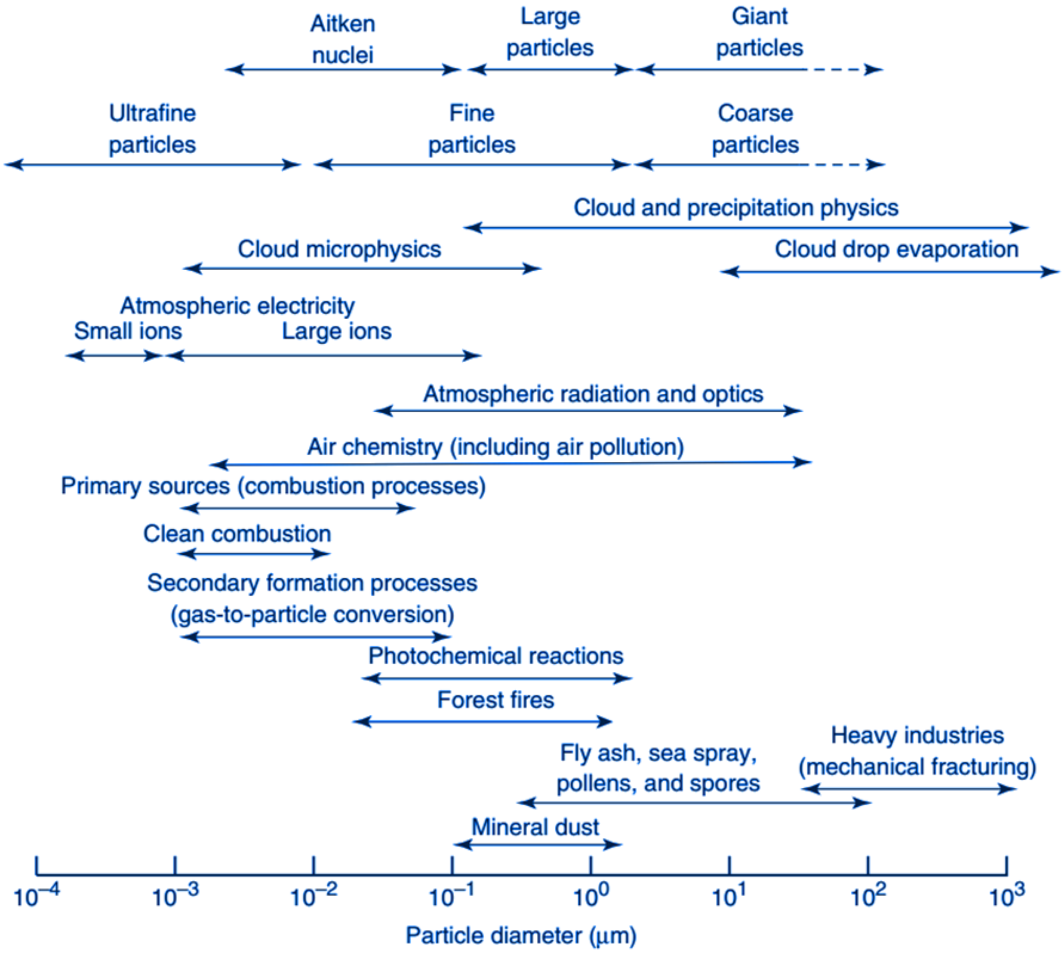


Fig. 1. Size range of aerosol particles in the atmosphere and their role in atmospheric physics and chemistry.

© 2017 Tous droits réservés.

Therefore, such particles have longer residence times than the nuclei, and their number concentration tends to increase (Claudio et al. 2017). It should be noted, that the particles which are formed in the secondary processes are in general smaller in size comparing to the primary emitted particles (Ault & Axson 2017). Different processes govern the behavior and fate of aerosol particles in the atmosphere and depend on their size and chemical composition. Contemporary research showed that the values of relative humidity (RH) associated with the hysteresis between deliquescence and efflorescence (dissolution and crystallization as a function of RH) are different for 100 nm versus 6 µm particles (Laskin et al. 2015). This example proves a necessity of studying composition of different fractions of aerosol particles with a special attention to a fine mode.

All aerosol characteristics and aerosol effects are controlled by the properties of individual particles. A volume of air contains many particles in any given particle size interval, which can be expected to have individual properties, because of the multitude of possible source and transformation processes. The chemical composition is a major factor that controls the atmospheric effects of aerosol particles. The many possible source processes and source types coupled with physical and chemical atmospheric transformation processes lead to a large variability in aerosol composition. The connections between the microphysical state and chemical composition of a particle with the transformation processes are summarized in Fig. 2. The chemical composition and microphysical state influence the response of the particle to environmental changes.

© 2017 Tous droits réservés.

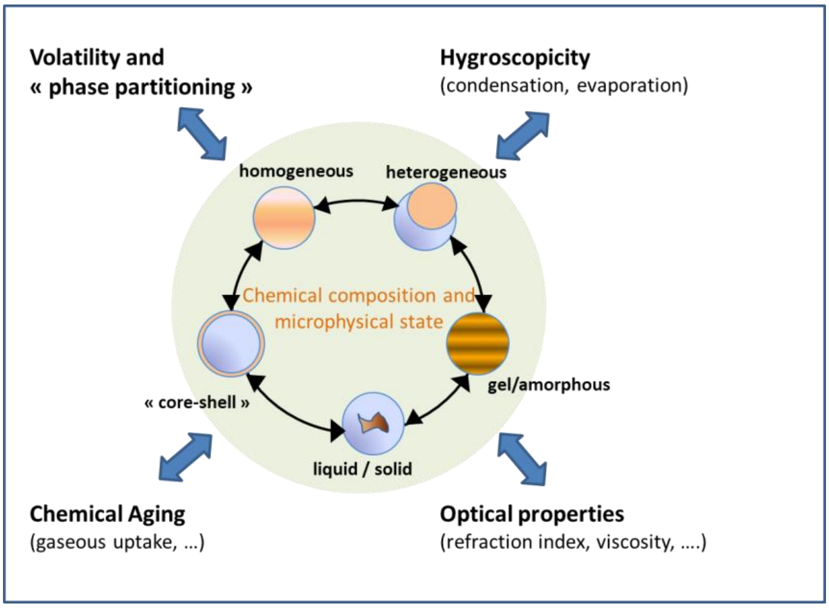


Fig. 2. Scheme of chemical composition and microphysical state modification of aerosol particles (Adapted from Krieger et al., 2012)

1.2. Single Particle Analysis of ambient aerosol particles

Spectroscopic and microscopic analysis has a long history in the field of aerosol particles' characterization. However, acquiring comprehensive information on a chemical composition of atmospheric particles is challenging because no single analytical chemistry technique can provide all the required information (Navel et al. 2015; Jung et al. 2014; Sobanska et al. 2012; Stefaniak et al. 2009; Ault et al. 2010). Single particle analysis (SPA) approach is presently recognized as a powerful tool to reveal detailed information, inaccessible by bulk techniques, concerning the particle origin, formation, reactivity, transformation reactions and their environmental impact (Ebben et al. 2013; Fitzgerald et al. 2015; Craig et al. 2017; Bondy et al. 2017; Andrew P Ault et al. 2012; Sobanska et al. 2012; Sun et al. 2016; Li & Shao 2010b; Moffet et al. 2016; Axson et al. 2016). Initially, SPA was commonly applied to an identification of the filter-collected particles and asbestos fibers, mostly determined by optical and electron microscopies (Fletcher et al. 2011). At an early stage of development, due to

© 2017 Tous droits réservés.

technological limitations, both techniques were restricted in their ability to provide chemical information of a sample composition. A combination of microscopic and spectroscopic techniques into a unique system made a breakthrough in SPA, simultaneously enabling a chemical analysis and determination of morphological features within a single particle. For the last 20 years, a use of these microanalytical methods has dramatically increased, mostly owing to technical improvements in chemical information that can be obtained and a need, with respect to aerosols, to improve understanding of a mixing state (Fitzgerald et al. 2015; Li et al. 2010; Adachi et al. 2014; Moffet et al. 2016; Laskin et al. 2015; Sobanska et al. 2012). Coupling of spectroscopy and microscopy, particularly for the analysis of single particles, has facilitated a more thorough understanding of a physicochemical mixing state (Ault & Axson 2017). Currently, a variety of microscopy, microprobe, spectroscopy, and mass spectrometry techniques are commonly applied to a complex characterization of aerosol particles collected in field campaigns and laboratory studies (Jonić et al. 2008; Prather et al. 2008; Hartonen et al. 2011; Hoffmann et al. 2011; Bzdek et al. 2012; Krieger et al. 2012; Nozière et al. 2015; Laskin et al. 2016)

In the literature, there are several important analytical instruments for single particle analysis, such as scanning electron microscopy (SEM) and transmission electron microscopy (TEM) with their equivalents within the coupled microanalytical systems, e.g. computer-controlled SEM (CCSEM)/energy- dispersed X-ray detector (EDS), environmental SEM(ESEM), focused ion beam (FIB)/SEM, high-resolution transmission electron microscopy (HRTEM)/electron energy loss spectroscopy (EELS), scanning transmission X-ray microscopy (STXM)/near edge X-ray absorption fine structure spectroscopy (NEXAFS). These techniques provide morphological features specification (diameter size, shape etc.), elemental composition and phase composition. However, the analysis of individual aerosol particles being complex environmental individuals is much broader in its scope and requires the use of techniques that provide complementary results. Therefore, other techniques were introduced in the field of SPA, such as atomic force microscopy (AFM) which enables

© 2017 Tous droits réservés.

detailed morphological analysis of submicron particles, time of flight secondary ion mass spectrometer (TOF-SIMS) and X-ray photoelectron spectroscopy (XPS) for surface analysis (Song & Peng 2009), Raman microspectroscopy (RMS) and FTIR microscopectroscopy (μ-FTIR) for molecular composition (Ault et al. 2013; Cheng et al. 2013; Sobanska et al. 2014; Hritz et al. 2016). The general scheme of the above mentioned techniques is presented below (Fig. 3).

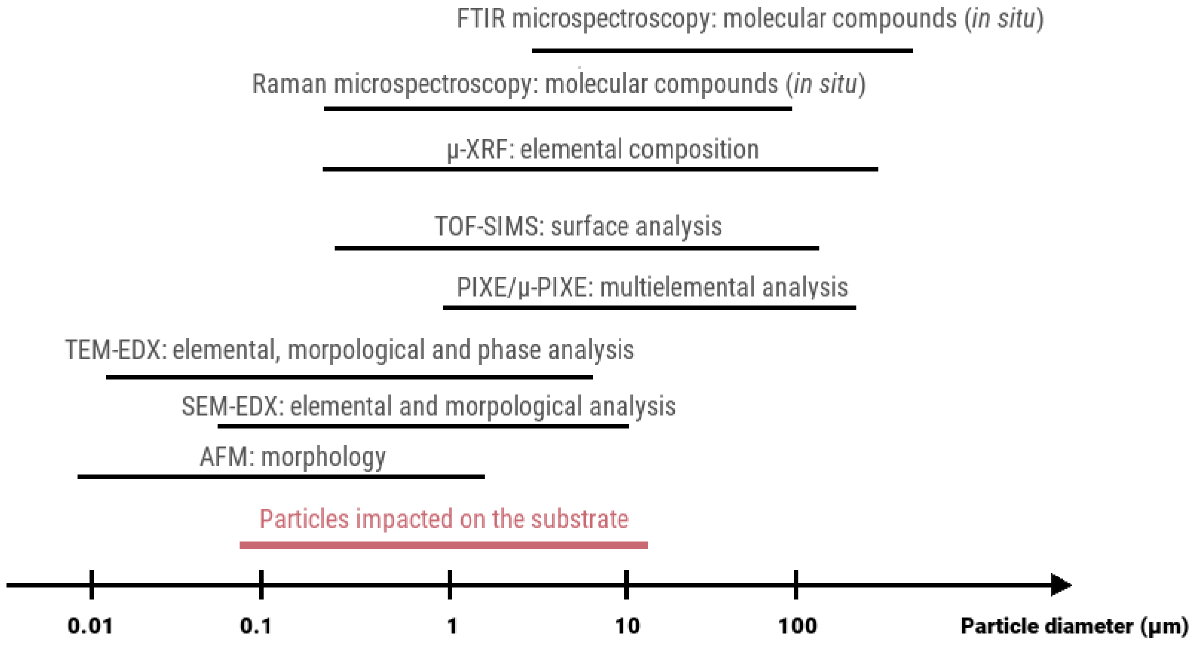

FTIR microspectroscopy: molecular compounds (*in situ*)
Raman microspectroscopy: molecular compounds (*in situ*)
μ-XRF: elemental composition
TOF-SIMS: surface analysis
PIXE/μ-PIXE: multielemental analysis
TEM-EDX: elemental, morpological and phase analysis
SEM-EDX: elemental and morpological analysis
AFM: morphology
Particles impacted on the substrate
0.01
0.1
1
10
100
Particle diameter (μm)

Fig. 3. Main techniques for single particle analysis of individual aerosol particles.

Historically, single particle analysis began with the electron microscopy-based techniques. For electron microscopy, a spatial resolutions is very high (<1−5 nm), which allows to obtain detailed morphology, even of fine aerosol particles (20-30 nm). The weakness of high vacuum electron microscopy is the loss of semi-volatile components. However, through an application of the improved methods like environmental scanning electron microscopy (ESEM) this process can be limited. The second disadvantage is related to the high-energy electrons, which may damage sensitive material, such as particles composed of organic carbon and/or ammonium nitrate. The popular SEM/EDX and EPMA techniques detect elemental (qualitative) composition and morphology of analyzing objects, recognizing a signal from a small

© 2017 Tous droits réservés.

interaction volume. The specialized software allows to create a semi-/fully-automated procedure dedicated to a rapid and nondestructive analysis of a large group of objects such as aerosol particles (Ro et al. 1999; Van Grieken et al. 2000; Osán et al. 2000; Ro et al. 2004; De Hoog et al. 2005; Ebben et al. 2013). There are some limitations related to quantifying of elemental contents, which requires adapted methods for determining the chemical composition of individual particles. Quantification procedure based on Monte-Carlo simulations was developed for the suitable measurements of low-Z elements. On the other hand, the gunshot residue (GSR) software enables detection of particles with high-Z elements. Based on the elemental composition and morphology, the particles can be classified into different groups related to the chemical composition of the samples. A significant number of publications reporting a successful application of computer-controlled SEM/EPMA measurements of a large number of particles (Sobanska et al. 2000; De Bock et al. 2000; Ro et al. 2001; Andrew P. Ault et al. 2012) followed by quantification and appropriate data classification, prove the need for fast and reliable tools in the field of SPA.

In the case of optical microspectrometry, the size limit for individual particle analysis has typically been set at 1 µm, due to the Abbe diffraction limit. Many studies were focused on the analysis of larger aerosol particles (1-10 µm diameter) providing valuable information on internal particle processes (Ciobanu et al. 2009; Brunamonti et al. 2015; Dallemagne et al. 2016). With technological improvements, the lower limits of optical microscopy are approaching the diffraction limit allowing spectral analysis of fine aerosol particles, however resolving submicron particles still remains challenging (Offroy et al. 2015; Bzdek et al. 2012; Ault & Axson 2017; Sun et al. 2016; Brunamonti et al. 2015).

A remarkable example of the techniques that are currently used in the analysis of fine aerosol particles is Raman microspectroscopy (RMS) and its complementary counterpart: Fourier Transform Infra-Red spectroscopy coupled with a microscope

© 2017 Tous droits réservés.

(FTIR microspectroscopy), providing molecular information in response to a wide range of radiation wavelengths. Infrared (IR) absorption and Raman scattering are both commonly used to study and identify substances using the compounds characteristic internal vibrations. IR spectroscopy is an absorption process, measuring the fraction of the light absorbed as the wavelength of the light is varied. The incident light is absorbed when the energy of the light closely matches the energy of a vibrational transition in the sample. A tiny proportion (approximately 1 in 10^9) of the incident photons interacts with vibrations in a sample and is scattered at higher or lower energy (Raman scattering). Additionally, by using a microscopic tool in both cases it is possible to observe morphological features, i.e. to map a surface of individual particles. The ability of Raman spectrometers to employ visible laser excitation and high quality microscope objectives results in a diffraction-limited laser spot of < 1 µm. IR systems employ longer wavelengths and hence the theoretical diffraction limit is much larger - around 20 µm. Moreover, IR spectrometers also suffer from less efficient objectives, typically giving an illuminated spot of around 100 µm. Apertures can be used to improve the resolution by spatially filtering the collected light, but at the expense of lower signal intensity. These differences crucially affect an ability of the instrument to resolve the components of inhomogeneous mixtures. Vibrational spectroscopy can operate at ambient pressure, and thus ambient RH, which avoids a potential loss of water or semi-volatile compounds. The application of automated microanalytical techniques for RMS and FTIR microspectroscopy was used in analysis of atmospheric aerosols (Sobanska et al. 2006; Song et al. 2010; Ivleva et al. 2013, Jentzsch et al. 2012, Jentzsch & Bolanz et al. 2012). Moreover, a combination of Raman microscopy and diffuse reflectance Fourier transform infrared spectroscopy (FTIR) has been used to characterize a micron-size tropospheric aerosols particles (Gaffney et al. 2015; Jung et al. 2014). By using the specially designed sample holders and flow reactor assemblies, a water uptake by particles and their subsequent phase transformations and ice nucleation can be quantified (Schill & Tolbert 2014; Laskin et al. 2015). In addition, some significant

© 2017 Tous droits réservés.

improvements of Raman microspectroscopy can find a measurable impact on analysis of aerosol particles. Regardless the analytical mode used in Raman or FTIR microspectroscopy – point analysis or molecular imaging – the appropriate chemometrics tools can significantly improve the process of obtaining knowledge from the collected data. Moreover, such a complex multidimensional data require a general statistical approach to evaluate the obtained spectra and allocate spectral features to size- and time-dependent properties in the atmosphere (Gautam et al. 2015). Moreover, pioneering work of Ofner research group (Ofner et al. 2016) show, that SPA like tip-enhanced Raman spectroscopy (TERS) opens an access to a deeper understanding of aerosol nanoparticles, which play a major role in many atmospheric processes and in particular in the global climate system. Sustaining this scientific trend is also reflected in the work of Craig et al. (2015). The promising results of the first application of atomic force microscopy with infrared spectroscopy (AFM-IR) to detect trace organic and inorganic species and probe intraparticle chemical variation in individual particles down to 150 nm was made (Craig et al. 2015, Bondy et al. 2017). However, it should be emphasized that these techniques are still under development for single aerosol particles analysis.

Transmission electron microscopy (TEM) and scanning electron microscopy (SEM) yield detailed images of a physical structure of individual particles. In TEM, a focused electron beam is transmitted through a specimen to form an image, while in SEM a focused electron beam scans a specimen's surface to create an image. Coupling spectroscopic and microscopic methods provides elemental and molecular composition of individual particles. TEM and SEM coupled to Energy Dispersive X-ray (EDX) spectrometry are commonly used for decades to analyze particle morphology, size, elemental composition, and internal structures with micrometer (SEM) and nanometer (TEM) lateral resolution (Pósfai & Buseck 2010; De Bock et al. 2000; Hoornaert et al. 2004; De Hoog et al. 2005; Stefaniak et al. 2006; Potgieter-Vermaak et al. 2005; Van Grieken et al. 2000; Worobiec et al. 2007; Darchuk et al. 2010; W. Li et al. 2016; Sun et al. 2016; Li & Shao 2010a). TEM is generally used for an analysis of an

© 2017 Tous droits réservés.

internal composition and a mixing state of single aerosol particles in a non-automatic mode (Moffet et al. 2016; Laskin et al. 2016; Li et al. 2010; Fu et al. 2012; Adachi et al. 2014). Electron Energy Loss Spectroscopy (EELS) coupled to TEM enables assessment of chemical bonding for selected elements within individual particles. A crystalline structure of particles can be determined through the analysis of the selected-area electron diffraction. In turn, SEM in a computer-controlled mode (CCSEM) permits a routine analysis of hundreds-to-thousands of particles deposited on substrates and provides statistically significant data on particle-type populations (Cprek et al. 2007; Yu et al. 2007; Bernstein et al. 2008; R.E. O'Brien et al. 2015; Laskin et al. 2016). Synchrotron-based X-ray microscopes enable chemical imaging of particles with an advanced speciation of carbon bonding and chemical characterization of different forms of carbon-rich particles (Shakya et al. 2013). Scanning transmission X-ray microscopy coupled with near edge X-ray fine structure spectroscopy (STXM/NEXAFS) has an advantage of providing quantitative measurements of low-Z (atomic number) elements (C, N, and O), as well as some heavier elements with L-shell absorption edges in the same energy range,e.g. K and Ca (Fraund et al. 2017). Moreover, STXM has a lower lateral resolution (>20 nm) than SEM and TEM (Laskin et al. 2016), but its higher chemical specificity has made it an instrument of choice for analysis of carbon-rich and mixed carbonaceous/inorganic particles (Kelly et al. 2013; Moffet et al. 2016).

Raman microscopectroscopy applied to individual aerosol particles has unique strengths and weaknesses in comparison with other vibrational techniques, such as Fourier transform infrared (FTIR) spectroscopy. Beside a different spatial resolution related to an irradiation wavelength, both Raman and FTIR microscopectroscopies provide complementary data.

© 2017 Tous droits réservés.

1.3. Chemometrics methodology applied to Single Particle Analysis

Over the past 40 years, a scope of chemometrics, across both academic research and industrial applications, has grown, evolved and been refined (Ziegel 2004). Nowadays, chemometric techniques, such as signal processing, multivariate data processing, pattern recognition, experimental design/optimization, have been widely investigated by scientists or adopted in industrial applications (Lavine & Workman 2013). This sub-discipline is rather new, but it had already a huge impact on the field of microanalysis. For single particle analysis (SPA), a combination of analytical techniques and chemometrics has become an important challenge in the last few decades (Gautam et al. 2015; Krammer et al. 2016; Äijälä et al. 2016). Nowadays, a large variety of chemometric methods has been used for different purposes of data analysis. In general, chemometrics came out from chemistry and introduced new methods capable of dealing with the large amounts of chemical data by means of multivariate data analysis. This state of affairs is of particular importance in the data processing. To analyze and visualize large data sets, sophisticated multivariate statistical analysis tools are necessary to reduce the data and extract components of interest. The data processing applied prior to multivariate analysis is known as preprocessing. Preprocessing for spectroscopic data is required to eliminate effects of unwanted signals such as fluorescence, Mie scattering, detector noise, calibration errors, cosmic rays, laser power fluctuations, signals from the cell media or glass substrate, etc. The multivariate analysis methods are applied in order to obtain information about sample composition, intercorrelation between particles as well as specification of the main particle groups. The description of particles on the basis of just a few representatives facilitates a possibility of monitoring their chemical composition in relation to particular variables, i.e. meteorological conditions or aerodynamic diameter (Yotova et al. 2016). Moreover, in the field of molecular imaging, the chemometric methods are an integral part of an analytical protocol, even for pushing back the analytical limits (Offroy et al. 2015). Several important methods, widely used

© 2017 Tous droits réservés.

in the data analysis, are presented below with examples in the field of the SPA for aerosol particles.

1.3.1. Asymmetric Least Squares Baseline Correction (AsLS)

Asymmetric least squares smoothing is attractive for baseline estimation of Raman and FTIR spectra, owing to few important advantages: 1) this way of a baseline correction is fast, even for large signals (large number of variables); 2) the flexibility in the baseline adaptation by only one parameter; 3) the flexibility in the position of the baseline adaptation by only one parameter. Given the two parameters, the computations are completely reproducible (Eilers & Boelens 2005). Unfortunately, there is no single, versatile recipe for an automatic choice of the parameters for arbitrary signals, so the judgment of a user is always needed. The asymmetric least squares method combines a smoother with an asymmetric weighting of deviations from a smooth trend to form an effective baseline estimation method (Eilers & Boelens 2005). However, the limitation of this algorithm is that only a smoothness constraint with a second derivative is considered. In practice, the method requires that the baseline fits the raw data well, and that the first derivative is very close.

Based on the Whittaker smoother, the asymmetric least squares (AsLS) method was proposed for background removal by Eilers (Eilers & Boelens 2005). A given vector *y* = *{y_1, y_2,..., y_i}* is defined as *i*, the observed frequency domain spectral intensities. The smoothing series *z* = *{z_1, z_2,..., z_i}* is faithful to *y*. Then, the penalized least squares function is minimized:

$$F = \sum_i (y_i - z_i)^2 + \lambda \sum_i \left(\Delta^2 z_i\right)^2 \qquad (1)$$

with $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$, i ꞓ $[1, 2, 3, ..., m]$, Δ is a second-order differential operator. The parameter *λ* is introduced to tune the balance between the

© 2017 Tous droits réservés.

smoothness and fitness. Finally, the vector *w* is defined as the weights of fitness and the minimized function is introduced as follows:

$$F = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i \left(\Delta^2 z_i\right)^2 \tag{2}$$

The minimization of the equation above can lead to the following equation:

$$(W + \lambda D^T D)z = Wy \tag{3}$$

with *W = diag(w)*, *W* is the diagonal matrix for vector *w*, *T* describes the transpose of a matrix, and *D* is the second order differential matrix: $Dz = \Delta^2 z$.

Generally, a lighter smoothing is capable of removing the noise, otherwise, a stronger smoothing can eliminate the true signal. In order to estimate the true background, much more attention should be paid to the deviations in the positive direction for the baseline correction. However, the weights of both negative and positive residuals *y–z* are the same when using the Whittaker smoother. Therefore, a key parameter of the asymmetric least squares for the baseline correction, p $(0 < p < 1)$, is introduced and computed as follows: $w_i = p$ if $y_i > z_i$ and $w_i = 1 - p$ otherwise.

The AsLS background correction was used in estimation of fluorescence background from 32,718 Raman spectra in the automatic mode, which significantly improve a process of a semi-continuous automated detection of airborne bioagents based on the Raman spectra of single particles (Doughty & Hill 2017).

1.3.2. Scatter corrections

Under scatter-correction methods, three preprocessing concepts: Multi Scatter Correction (*MSC)*, Standard Normal Variate (*SNV)* and normalization were considered. These techniques are designed to reduce the variability between samples. All three also adjust for baseline shifts between them.

© 2017 Tous droits réservés.

Multiplicative Scatter (or, in general, Signal) Correction (MSC) is probably the most widely used preprocessing technique for IR spectra, closely followed by SNV (Rinnan et al. 2009). The concept behind MSC is that artifacts or imperfections (e.g. undesirable scatter effect) will be removed from a data matrix prior to a data modeling (Chen & Thennadil 2012). MSC comprises two steps:

1. Estimation of correction coefficients (additive and multiplicative contributions).
2. Correcting a recorded spectrum.

In most applications, an average spectrum of a calibration set is used as a reference spectrum. However, a generic reference spectrum can also be applied. In the original paper about MSC (Geladi et al. 1985), it was suggested to use only those parts of the spectral axis that do not include relevant information (baseline). While this makes good spectroscopic sense, it is difficult to determine such spectral regions in practice. This is the reason why, in most cases, the entire spectrum is used to find the scalar correction parameters in MSC.

The basic form of MSC has been expanded into more elaborate augmentations (Martens et al. 2003; Xu et al. 2008) commonly known as an extended multiplicative signal correction (EMSC). This expansion includes both a second-order polynomial fitting to the reference spectrum, a fitting of a baseline on a wavelength axis and a use of a priori knowledge from the spectra of interest or spectral interferents. The application of EMSC was evaluated in the field of the SPA for analysis of aeroallergens by FTIR spectroscopy (Zimmermann et al. 2015). The analysis of the EMSC parameters indicates that the FTIR methodology offers an indirect estimation of morphology of pollen and spores. Thus, the study has shown that identification of principal aeroallergen bioparticles can be based on FTIR methodology (Zimmermann et al. 2015).

Standard Normal Variate (SNV) preprocessing is probably the second most applied method for a scatter correction of spectral data (Barnes et al. 1989). The signal-

© 2017 Tous droits réservés.

correction concepts behind SNV and Normalization are the same as for MSC except that a common reference signal is not required. Instead, each observation is processed on its own, isolated from the remainder of the set. A lack of need for a common reference might be a practical advantage. Since SNV and normalization do not involve a least squares fitting in their parameter estimation, they can be sensitive to noisy entries in the spectrum (Bi et al. 2016). Instead of using an average and a standard deviation as the correction parameters, one might consider using more robust equivalents of these statistical moments. Guo et al. (Guo et al. 1999) suggested using a median or a mean of an inner quartile range and a standard deviation of an inner quartile as estimates. This would be especially appropriate for noisy spectra (e.g., in Near Infrared applications). As an example, the SNV correction was an important part of the data preprocessing in specification of the influence of optical substrates on micro-FTIR analysis of single mammalian cells (Wehbe et al. 2013).

1.3.3. Savitzky-Golay smoothing and derivation

Savtizky and Golay (SG) (Savitzky & Golay 1964) popularized a method for a numerical derivation of a vector that includes a smoothing step. In order to find a derivative at a center point *i*, a polynomial is fitted in a symmetric window on the raw data. When the parameters for this polynomial are calculated, the derivative of any order of this function can easily be found and this value is subsequently used as the derivative estimate for this center point. This operation is applied sequentially to all points in the spectra. The number of points used to calculate the polynomial (window size) and the degree of the fitted polynomial are both decisions that need to be made. The highest derivative that can be determined depends on the degree of the polynomial used during the fitting (i.e. a third-order polynomial can be used to estimate up to the third-order derivative). There is an intrinsic redundancy in the hierarchy of SG derivation. For each derivation, two subsequent polynomial fits will give the same estimate of the coefficients. For the first derivative, a first-degree

© 2017 Tous droits réservés.

polynomial and a second-degree polynomial will give the same answer (as will the third and fourth degrees). For the second derivative, a second and third-degree polynomial will give the same answer (as will the fourth and fifth degrees), etc. When this method was first introduced by Savitzky and Golay (Savitzky & Golay 1964), it was still computationally cumbersome to calculate the parameters in estimating the derivative. For that reason, the authors reported a set of tabulated values for several different types of derivatives and polynomial combination. However, errors were introduced in their first article, Steinier et al. (Steinier et al. 1972) published a corrected and expanded version of the original tables. The tables were later even further expanded by Madden (Madden 1978). However, with modern computers, there is no longer any real need for these tables. The original forms of SG derivation use a symmetric window smoothing, requiring a number of data points on each side of a center point to be the same. Therefore, the technique neglects a number of points at each end of the spectrum during the preprocessing. For SG derivation, the number of points lost equals the number of points used for smoothing minus one. If the spectral vector is long (i.e. more than 500 points), this issue is not important, but, for shorter, this loss of wavelengths can be important. Proctor and Peter (Proctor & Peter 1980) and later, Gorry (Gorry 1990) suggested a solution that involves using a fitted polynomial based on an asymmetric window for the end-points. In practice, this means that the m first points of the spectra are estimated from the $2m+1$ first points in the spectra, and a similar estimate for the last m points. However, such a solution will evidently introduce artifacts, as the accuracy of the derivative decreases with the distance from the center point ($m+1$). Furthermore, the estimation of the end-points does not possess the inherent redundancy mentioned for SG: no two subsequent polynomial order fittings will give the same estimates. In addition to this, the estimate of the d^{th} derivative will be equal for all the end-points if the spectrum is smoothed by a d^{th}-order polynomial. The SG derivation uses common filtering techniques to estimate the derivative spectra, and, instead of using the finite-difference approach, fits a polynomial through a number of points.

© 2017 Tous droits réservés.

The application of the SG smoothing and derivatization was crucial to improve an identification of a single bacteria cell or strain by Raman microspectroscopy. More precisely, a smoothed signal and a first derivative were calculated by Savitzky-Golay polynomial filters, that constituted a powerful, simple and fast method to obtain the desired signal without loss of intensity (Strola et al. 2014).

1.3.4. Principal Component Analysis

One of the fundamental aspects to be considered in linear transformations is the correlations between variables. By using them, a transformation of an original set into another space can be made, in which new variables are uncorrelated or statistically independent. Principal component analysis (PCA) is a statistical method that defines a linear transformation which is able to support description of a stationary stochastic process given in the form of a set of *N*-dimensional vectors with reduced dimensions to lower *K*-dimensional set (*N*>*K*). This transformation takes place through the *W matrix* of dimensions *K*×*N* where in such a way; the output space *y* of the reduced dimension retains the most important information about the original process. In other words, PCA replaces a high amount of information contained in a mutually correlated input into a set of statistically independent components, according to their importance. PCA is a well-known chemometric method for a decomposition of two-way matrices that are schemed in Fig. 4 (Bro & Smilde 2014). The steps in PCA are as follows: (i) the *X*-space (where *X* corresponds to the *X* data matrix) is given a coordinate system where each variable gets an axis which length corresponds to its scaling; (ii) each observation in this space is represented by a point; (iii) the average of each variable is then calculated and subtracted (mean centering) - this is equivalent to moving the swarm of points to the center of the coordinate system; (iv) a function is fitted to the data that describes as closely as possible the variance of the observations in the *X*-space. By projecting each point down to the line (Euclidian distance) and measuring the distance between the center point and the projection

© 2017 Tous droits réservés.

point, the *score value* (*t*) of each observation is obtained. The angle between the line and each variable axis determines the influence of each variable, the loading value (*p*). One loading value is given for each variable in the data set. When the first Principal Component (PC) has been calculated, the remaining unexplained variance is left in the residual matrix, *E*:

$$X = TP' + E \tag{4}$$

where, *T* is score matrix and *P'* is transposed loading matrix.

The second PC is orthogonal to the first. More PCs can be calculated as long as unexplained information is left. The significant number of principal components can be estimated by different methods, of which cross validation is an often-used method. The variance of a principal component is described by the eigenvalue, which is proportional to the variance explained by a PC. Although PCA can be calculated using different algorithms, the two most common methods are non-linear iterative partial least squares (NIPALS) (Wold et al. 1987) and singular value decomposition (SVD) (Jackson 2005). A basic rationale in PCA is that the informative rank of the data is less than the number of original variables. Hence, it is possible to replace the original number of variables with principal components and gain a number of benefits. The influence of noise is minimized as the original variables are replaced with weighted averages, and the interpretation and visualization is greatly aided by having a simpler view to all the variations. Furthermore, the compression of the variation into fewer components can yield statistical benefits in further modeling with the data. In addition, it is quite common to use PCA as a preprocessing step in order to get a compact representation of a dataset. For example, the scores may be used for building a classification model using linear discriminant analysis (Hair et al. 2010). PCA is frequently used for identifying pollutant sources affecting the air quality (Adams 1994; Cusack et al. 2013; Genga et al. 2012). This procedure is advantageous because detailed information regarding atmospheric chemistry and meteorology is not required. The application of PCA for single particle analysis has a rich history. For

© 2017 Tous droits réservés.

example, by using PCA the variability of the concentrations of diversified sources of particles, such as: soil derived dust and particles related to biological processes (forest fires or agricultural burning) was discovered (Adams 1994). In another work, PCA was performed in order to study the correlation among the particles collected in yard, urban and rural sampling sites, to obtain information on the pollution sources and to investigate the differences among them (Genga et al. 2012). In the case of a large data set processing, PCA has been used for the results from both single particle (SEM/EDS) and bulk (X-ray fluorescence) analysis results of a combined set of approximately 25,000 individual particles collected over Lake Balaton in Hungary (Osán et al. 2001). Such data treatment was applied to determine potential sources of the collected aerosol particles. PCA is also an important element of many single particle analysis algorithms. A noticeable example is an automated data analysis method for atmospheric particles using scanning transmission X-ray microscopy coupled with near edge X-ray fine structure spectroscopy (STXM/NEXAFS). This method was applied to solve a structure of complex internally mixed submicrometer particles containing organic and inorganic material (Moffet et al. 2010).

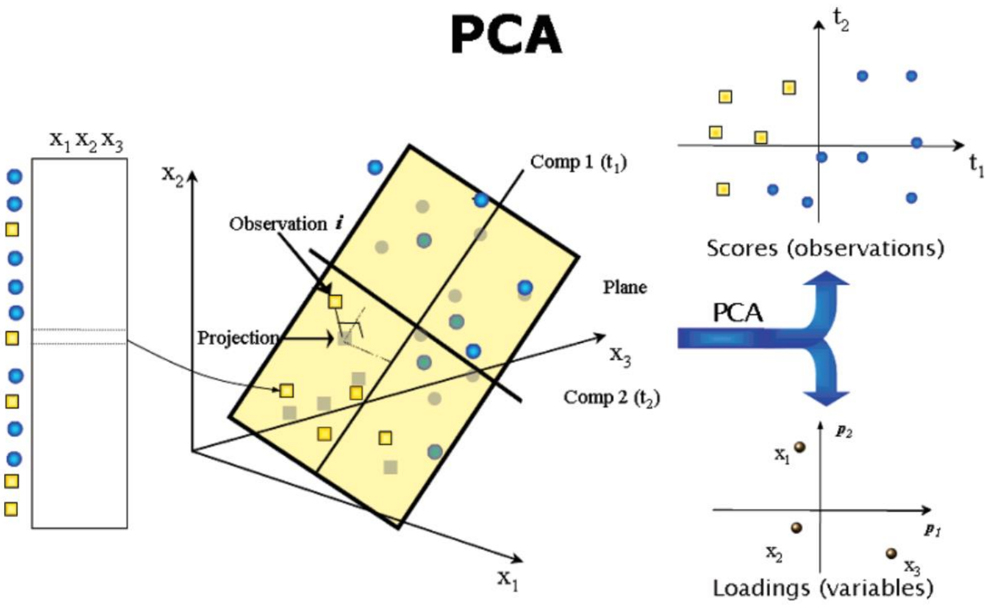


Fig. 4. Principal component analysis (PCA) model approximates variation in data table by low dimensional model plane (Hair et al. 2010).

© 2017 Tous droits réservés.

1.3.5. Cluster Analysis

Clustering data into meaningful groups is an important task of chemometrics (Ziegel 2004). Clustering is considered to be an unsupervised classification of data. A number of clustering algorithms have been described in the literature (Everitt et al. 2011). Some of them are capable of discovering proper clustering of data only when a number of clusters is known in advance. Other algorithms are capable of discovering clusters of particular shapes only. There are also algorithms that are able to identify noise data. In general terms, the task of clustering can be formulated as follows: given a set of *N* elements, find its partition into *K* clusters, such that the elements within groups are more similar to each other than the elements that belong to different groups. Such classes are meant to express the structure of the data, not given *a priori*. The choice of the clustering algorithm cannot be made separately from the context of a particular data set and expectations about the structure of clusters. When a given sample is taken as a point in the space defined by variables, the clustering algorithm can calculate the distance between this point and all the other points, thereby establishing a matrix that describes the proximity between all the samples studied. There are several ways of calculating the distance between two points, the best known and most often used is the Euclidean distance (da Silva Torres et al. 2006). Different clustering methods can be categorized into three types (Jain & Dubes 1988). The following list explains these categories:

1) Agglomerative vs divisive. The former begin by treating each spectrum (sample) as a cluster and successively merge them until a stopping criterion is met (the bottom-up style); the latter begin by placing all spectra (samples) in a single group and perform splitting until a stopping criterion is met (the top-down style). The details about these methods can be found elsewhere (Everitt et al. 2011)

© 2017 Tous droits réservés.

2) Hierarchical vs partitional. This aspect relates to the structure of the clusters that are produced. The former algorithms form a hierarchy of clusters: clusters at lower levels are nested to upper level clusters (Everitt et al. 2011).

3) Hard vs fuzzy. This aspect concerns a cluster membership. The former method allocates each spectrum (sample) to a single cluster while the latter predicts its degree of membership for multiple clusters. A fuzzy method can be converted to a hard one by assigning a spectrum (sample) to a cluster that has the highest degree of membership. Details about the Hard and Fuzzy clustering can be found elsewhere (Gosain & Dahiya 2016; Pedrycz 2005).

Although clustering algorithms are not the main focus of this thesis, they are important for a fair evaluation of different representation models and similarity measures. In order to compare the main concept-based spectral clustering methods, two popular clustering algorithms are briefly presented: the agglomerative hierarchical clustering algorithm (HCA) and the partitional k-means algorithm.

In the agglomerative hierarchical clustering, the data are not partitioned into a particular number of clusters at a single step (Everitt et al. 2011). Instead, the clustering consists of a series of partitions, which run from *n* clusters containing a single individual, to single cluster containing all individuals. Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, an investigator wishes to have a solution with an optimal number of clusters specification. This problem of deciding on the correct number of clusters is discussed below. Hierarchical clustering may be represented by a two-dimensional diagram known as a *dendrogram* (Fig. 5), which illustrates the fusions made at each stage of the analysis. HCA has been demonstrated to be a powerful tool to classify aerosol particles (Gabey et al. 2011; Robinson et al. 2013; Crawford et al. 2015) however, the available toolkits are limited by heavy computational burdens, making the analysis of large data sets problematic (Crawford et al. 2015). In another work, the hierarchical cluster analysis in combination with principal component analysis of

© 2017 Tous droits réservés.

fused hyperspectral data cubes allowed a detailed and well-grounded assignment of chemical species and their relationship to each other in ambient aerosol particles (Ofner, Katharina A. Kamilli, et al. 2015).

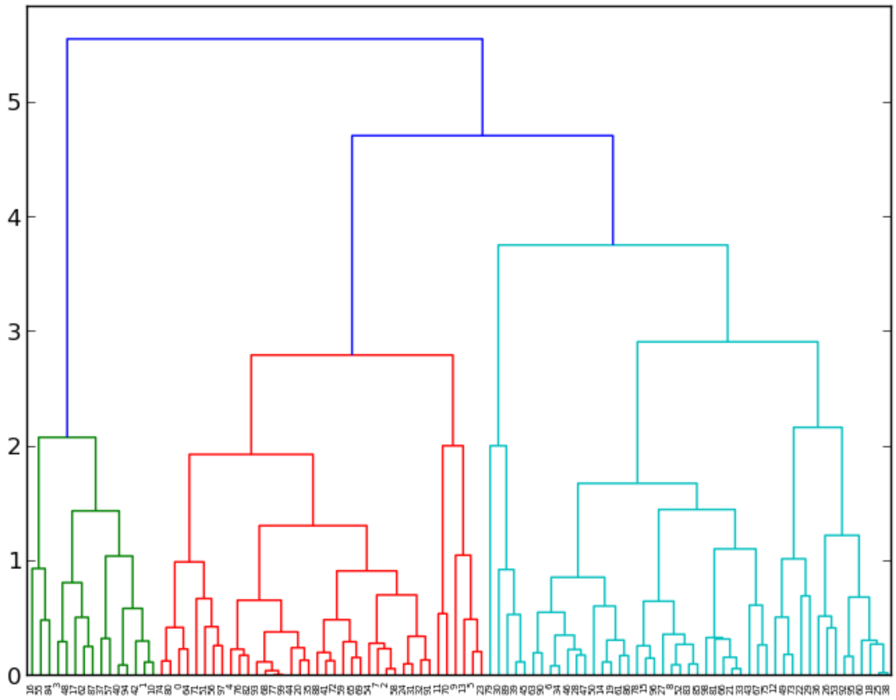


Fig. 5. Dendrogram plot generated by Ward's HCA algorithm (https://docs.scipy.org).

Partitional clustering methods find clusters by optimizing a certain objective function that defines the optimal solution (Hartigan 1975). For instance, the k-means algorithm minimizes the squared error in the resulting cluster structure, by assigning each point to its closest cluster in each iteration. It must be noted that an exhaustive search through all possible partitions for the optimal solution is computationally prohibitive. It is common to approximate this by running the algorithms multiple times with different initialization, each time generating a different partition of the data set, and then use the best clustering result (Jain 2010). Due to this approximation, partitional methods are usually efficient, therefore they are favored

© 2017 Tous droits réservés.

for handling large data sets (Butler et al. 2016). The k-means algorithm is one of the most commonly used partitional clustering methods in chemometrics. As an example, the k-means cluster analysis was performed for ~5000 individual aerosol particles analyzed by CCSEM/EDX and RMS techniques (Moffet et al. 2012). By application of k-means clustering, Fe^{2+} was found to be an insignificant enrichment in ambient anthropogenic particles beside to standard Asian mineral dust. In another work, k-means clustering was applied to the relative atomic abundances from the CCSEM-EDX spectra collected for the 0.7−5.0 µm size range of ambient and laboratory-generated particles (Axson et al. 2016). This allowed mathematical grouping of spectra to determine the types of aerosols present in each sample without human bias during sorting.

Different clustering algorithms usually lead to distinctive data partitioning. Even for the same algorithm, the selection of particular parameters may greatly affect the final clustering results (Charrad et al. 2014). Thus, effective evaluation standards and criteria are critically important in the cluster analysis. At the same time, these assessments also provide some meaningful insights on how many clusters are hidden in the data. As such, numerous indexes for determining the number of clusters in a data set have been proposed (Duda & Hart 1973; Hubert & Levin 1976; Krzanowski & Lai 1988). Moreover, Milligan and Cooper (Milligan & Cooper 1985) presented a complex work about a comparison of 30 internal validity indexes for hierarchical clustering algorithms, whereas a systematic study of 16 external validation measures for K-means clustering is given in the literature (Wu, Xiong, et al. 2009; Wu, Chen, et al. 2009). The external indexes are based mainly on prior information of the data. In practice, such information is often unavailable, and calculation of internal indexes is therefore more useful. In the case of individual aerosol particle analysis, the calculation of such indexes is not a common practice. This fact is dictated by a complex structure of data which processing is necessary to be supervised at each step. However, in the case of automatic data processing, the step of optimal clustering number specification may greatly improve the analytical protocol

© 2017 Tous droits réservés.

effectiveness. In our work (see chapter 4), two internal indexes for specification of the main groups of aerosol particles was made. By application of the Dindex (Baccini 2010) and Hubert's index (Hubert & Levin 1976) the designation of an optimal number of clusters for HCA was possible and the calculated criteria were complementary with the standard dendrogram inspection methods.

1.3.6. Multivariate Curve Resolution

The large heterogeneity at the level of individual particles in environmental samples generates a severe overlap of spectral information to obtain pure component spectra and concentrations. For cases where spectral mixture data are available without either pure component spectra or concentration profiles of pure components, a wide variety of self-modeling mixture analysis tools is available in Multi Curve Resolution methods (Hamilton et al. 1990, Windig 1992).

The Simplisma (Simple-to-use interactive self-modeling mixture analysis) approach is different in that it is not based on PCA and it is interactive (Windig & Guilment 1991, Windig et al. 1992)

The interactivity is important to resolve data sets dealing with environmental chemistry where it is not possible to obtain replicate samples when trouble-shooting: user interaction based on their spectroscopic knowledge is necessary to avoid meaningless problems. When highly overlapping spectral features and/or baselines are present in the spectra, second derivative spectra can be used to resolve the data properly (Windig & Stephenson 1992, Windig 1994, Windig & Merkel 1993, Guilment et al. 1994).

This approach uses pure spectra as a first estimate and derives pure variables from the resolved contribution profiles.

SIMPLISMA method (Windig & Guilment 1991; Windig 1997) is based on evaluating the relative standard deviation of the column *n, pn*, defined from equation:

© 2017 Tous droits réservés.

$$p_n = \frac{s_n}{\overline{X}_n + \delta} \tag{5}$$

where s_n is the standard deviation of the column *n*, $\overline{X}$ is the average of the column *n* and δ is a correction factor that is added to avoid columns with a low average value (generally associated with noise) being the purest variables. A large relative standard deviation (p_n) indicates a high purity of the column. The process involves, in the first step, finding a column with the highest value of a relative standard deviation and then normalizing this column. The variable with the second highest purity, as well as having a high relative standard deviation, must have the least correlation with the first pure variable. A weight factor, w_n, is therefore calculated as follows:

$$w_n = \det(Y_{n^T} Y_n) \tag{6}$$

where *Y* is a matrix made up of the pure variables found and each *n^th* column of the data matrix that has not yet been selected. The value calculated by the determinant will be proportional to the independence between the pure variables and the *n^th* row, which has been used to build the matrix Y_n.

To calculate a new pure variable p_i, therefore, the weight factor will be applied.

$$p_n = w_n \left(\frac{s_n}{\overline{X}_n + \delta} \right) \tag{7}$$

The algorithm selects the maximum value of p_i, which corresponds to the variable of the greatest purity, and so on until all the pure variables are found. With the purest variables, the columns (the purest spectra) are obtained. SIMPLISMA was implemented for Raman imaging as SIMPLISMAX by adding a derivative tool leading to distinction between wide and narrow Raman bands and then extracting properly the fluorescence spectra from the Raman ones (Windig et al. 2002)

Simplisma has been described for a variety of applications based on Raman and FTIR microspectroscopy (Windig 1994, Windig & Merkel 1993, Guilment et al. 1994, Smith & Kramer 1999).

© 2017 Tous droits réservés.

lilliad.univ-lille.fr

Additionally, the alternative least square can be applied to Multi Curve Resolution method (MCR-ALS). The aim of the alternating least squares (ALS) method (Tauler et al. 1993) is to obtain, from an initial estimate using e.g. SIMPLISMA, a result with a chemical significance that corresponds, in a satisfactory way, to the observed experimental behavior. ALS allows several constraints to be imposed (De Juan & Tauler 2003). This method imposes a linear model on experimental data and, when working with spectroscopic responses, transforms the spectral concentration values in the non-negative mode.

In the first step of MCR-ALS process, PCA is applied to determine how many significant principal components or sources of variation are present in the matrix. In the second step, an initial estimation is calculated, of either the concentration profiles or the spectra of pure products, from the number of significant principal components previously found. This initial estimation can be made using PCA, e.g. the evolving factor analysis (EFA) or by other techniques, such as independent component analysis, and simple-to-use interactive self-modeling mixture analysis (SIMPLISMA), based on finding pure variables. From the initial estimation and the number of significant principal components selected in principal component analysis, alternating least squares (ALS) is applied to obtain a non-negative matrix of concentration profiles for instance (Fig. 6). This can also be powerful as a refinement method in the extraction process of very similar Raman spectra (Sobanska et al., 2006).

© 2017 Tous droits réservés.

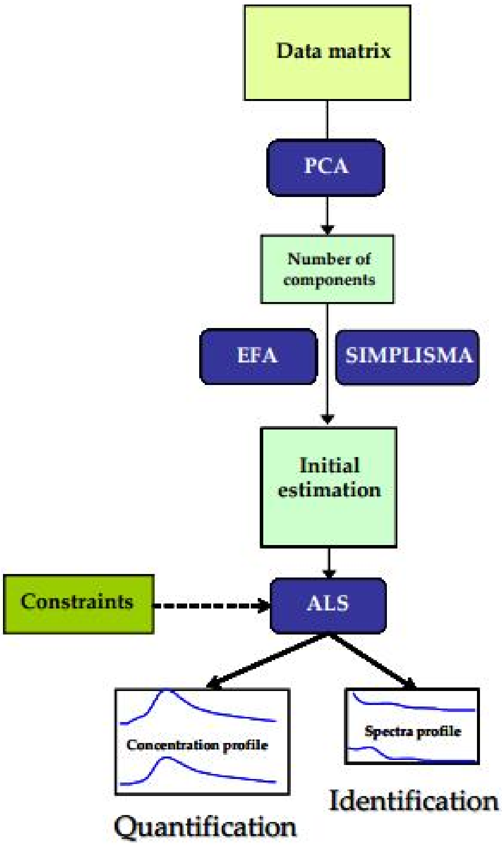


Fig. 6. Scheme of MCR-ALS.

There is a rich repertory of scientific publications about an application of the MCR methods (or MCR-ALS) in single particle analysis - usually for the mapping. The effectiveness of a combined use of a computer-controlled Raman microspectrometry mapping and MCR methods for Raman images to determine chemical and heterogeneous characteristics of individual tropospheric aerosol particles was successfully presented in many publications (Batonneau et al. 2001; Batonneau et al. 2003; Batonneau et al. 2006; Sobanska et al. 2006; Sobanska et al. 2014)

In another work, the Raman mapping and multivariate curve resolution were applied to the chemical changes occurring at the interface between single particles, creating the reactive interface (Falgayrac et al. 2006; 2012). The laboratory analysis using in situ Raman imaging combined with MCR-ALS approach indicated a fundamental role

© 2017 Tous droits réservés.

of small amounts of liquid water in initiating the chemical reaction of $ZnSO_4{\cdot}7H_2O$ microparticles in contact with a $CaCO_3$ surface (Falgayrac et al. 2014). The same group, based on the data from Raman microspectroscopy with MCR-ALS approach, as well as TOF-SIMS spectrometry, elaborated heterogeneous microchemistry between $CdSO_4$ and $CaCO_3$ particles under humidity and liquid water (Falgayrac et al. 2013).

© 2017 Tous droits réservés.

CHAPTER 2: MATERIALS AND METHODS

© 2017 Tous droits réservés.

2.1. Description of the samples

In this work, the results from the analysis of inorganic aerosol particles (coal mine dust), as well as organic (biogenic) particles in the form of pellets were presented. The results were supplemented by specification of the deposited organic pigments on the paper and characterization of inorganic aerosol particles deposited on various substrates. The selection and differentiation of the samples was intended to present various aspect of the use of multivariate statistics methods to noteworthy paradigms from the field of industrial and biogenic aerosol particle analysis, as well as non-destructive microanalysis of cultural heritage materials and specification of suitable condition for Raman imaging of aerosol particles.

2.1.1 Mining particles

Oruro - Bolivia

The southwest of Bolivia consists of the Altiplano, a longitudinal basin flanked on the west by the Cordillera Occidental and on the east by the Cordillera Oriental. In turn, the Oruro area located in the Altiplano basin is regularly rich in polymetallic deposits. Thus, the mining and metallurgical activity plays an important role in the economy of this region. The critical impact of mining caused a significant degradation of the local environment (Banks et al. 2002; Jacobsen 2011; Tapia et al. 2012; Rötting et al. 2014). One of the specific harmful factors is the inhalation of metal-rich particles that may cause serious health consequences to the exposed population (miners, inhabitants of the area) (Pavilonis et al. 2017). The particles were sampled in the galleries of San Jose Mine (150 m, underground), located in the Oruro area, Bolivia (17°46'0"S – 67°28'60"W – 3,674 MAMSL). The particle collection was performed using a personal cascade impactor (SIOUTAS, 3 l.min-1), allowing simultaneous walking along underground passages and sampling of four size fractions of particles, i.e. PM 10, PM 2.5 , PM 1 and PM 0.5 corresponding to 10-2.5µm, 2.5-1µm, 1-0.5µm and < 0.5µm

© 2017 Tous droits réservés.

aerodynamic diameter, respectively. The particles were collected on TEM grids mounted in the impaction plates.

Bogdanka coal mine

The world leaders of hard coal productions are China and USA; according to the data published by International Energy Association IEA statistics (http://www.iea.org), Poland is within the top ten coal producers in the world, with its average annual production of ca 140 Mt. Most of the global production of coal is dedicated to the local state consumption; only around 15% is destined for the international coal market (http://www.worldcoal.org). Sampling of the coal dust was conducted in the Bogdanka - underground coal mine located in the Eastern Poland. Sampling sites were distributed near an outtake shaft with a diameter φ 7.5 m and 996 m depth, in the coal seam. According to Philpott (Philpott 2002), the coal from this seam shows the following parameters: the ash content 10.02% to 38.47% (the average 21.71%) and the sulphur content varies from 0.82% to 2.16% (the average 1.27%). The seam shows a changeable morphology because of the coalbed thickness and mullock interlayers. Variability of the ash content is strongly linked with the non-coal rock interlayers, since their presence causes a decrease of the coal calorific value and increase of after-burning residue. Sulphur content also varies quite substantially, but it is not related to the presence of worthless material (mullock); it is likely to originate from sulphides such as pyrite being present in the coal exploited in these coal mine (Sawlowicz et al. 2005).

Dust samples were collected from gravitational deposition along the main gallery, beginning with a spot near the shaft exit and moving gradually closer to the longwall.

© 2017 Tous droits réservés.

2.1.2. Biogenic particles

Pollen grains is the natural source of proteins, lipids, vitamins and mineral salts for bees, being the only source of nitrogenated food available for bee larvae, and its absence may result in the hive extinction. Pollen is stored in pollen baskets on posterior legs of the bees and brought to a hive. To make pollen stick together, the bees add some saliva and nectar. In the hive, it is stored in honey combs and used as food for the bees. The pollen pellets were collected directly from beehives. Pollen pellets were purchased from Percie du Sert (Saint-Hilaire de Lusignan, France). For the purpose of this subsection, the two types of pollen pellets were analysed: (i) contaminated by pesticides (i.e. imidacloprid) and (ii) with pesticide traces (i.e. imidacloprid). Both two pollen pellets after collection were transferred to numbered plastic vials, which were hermetically closed and kept in a refrigerator at 4 °C until use. The pollen pellets without any preparation were analysed by means of a confocal Raman microspectrometer.

2.1.3. Organic pigments

The scientific study of artworks on paper shares common objectives with technical studies of any work of art. Artifacts are examined in order to answer historical questions about their origin, namely, where, when, and by whom the artwork was created. Scientific examinations seeking to answer these questions generally require identification of the materials and working methods used to craft the object. Other studies seek to answer basic questions about the care of the artifact: its physical and chemical condition, causes for deterioration, and vulnerability to storage or exhibition conditions. The most common investigation for paintings or colored prints on paper involves identification of the pigments. Due to this requirement, the set of 6 natural powder pigments: turmeric, dragon's blood, indigo, safflower, cochineal, gamboge (Kremer Pigmente) and binding medium (rice starch) were prepared. As a deposition substrate, 4 different papers were tested i.e. Whatman, K14, K78 and M20. For the

© 2017 Tous droits réservés.

purpose of the current methodology, the specification of each paper is not necessary. Pigments were mixed with water and then deposited by wooden spatula on the paper. Finally, the paper with deposited pigments was left to dry for at least 12 hours.

2.2. Analysis of the samples

The analysis of the above mentioned samples was made by application of the means of techniques enabling the analysis of individual particles in a statistically significant group, as well as bulk analysis. A brief description of the main techniques used in this work is outlined below.

2.2.1. Raman microspectrometry

The Raman effect has been known and exploited for many years, and the physics behind it is very well described (Ferraro et al. 2003). The first theoretical predictions of inelastic scattering light were made by Smekal (1923), followed by the first experimental observation in 1928 by Raman and Krishnan (Raman & Krishnan 1928) who observed a frequency shift in the spectrum of scattered light compared to incident light. This frequency shift is known today as the Raman shift and can be calculated by the formula:

$$\Delta(cm^{-1}) = 10^{-7}\left(\frac{1}{\lambda_{excitation}} - \frac{1}{\lambda_{Raman}}\right) \qquad (8)$$

Where Δ is the Raman shift, $\lambda_{excitation}$ is the wavelength of the excitation source and λ_{Raman} is the corresponding Raman wavelength. In classical interpretation, Raman effect can be explained by the interaction of incident radiation of the electric field $\vec{E}$ with a molecule. The incident electromagnetic field induces an electric dipole moment $\vec{P}$:

$$\vec{P} = \bar{\bar{\alpha}}\,\vec{E} \qquad (9)$$

© 2017 Tous droits réservés.

Where $\bar{\bar{\alpha}}$ is the polarizability tensor (matrix of 3rd order) of the molecule and $\vec{E}$ is the amplitude of electrical field corresponding to the incident electromagnetic wave. The polarizability represents an intrinsic property of the molecule and depends on the electronic structure and the nature of the chemical bonds. For non-isotropic molecules, the polarizability may vary with position and interatomic distances, and depends on the molecule symmetry.

A simple qualitative description of the Raman scattering can be obtained using the classical electromagnetic theory. Consider an electromagnetic wave defined by the formula (10):

$$E = E_0 \cos(2\pi\nu_0 t) \tag{10}$$

Where ν_0 is the frequency. From equations (8) and (9), the time dependent induced electric dipole moment is:

$$P = \alpha E_0 \cos(2\pi\nu_0 t) \tag{11}$$

For the next step the assumption of single model molecule that is free to vibrate, but not rotate is needed. The molecule is fixed in space in its equilibrium position and nuclei can vibrate around their equilibrium positions. Any disturbance in the electronic cloud caused by an incident electromagnetic wave will induce changes in the molecule polarizability. This variation of the polarizability during the vibrations of the molecule can be expressed by expanding the polarizability α in a Taylor series (Goodwillie 2003) with respect to the coordinates x_i of vibration:

$$\alpha = \alpha_0 + \frac{\partial \alpha}{\partial x_i} x_i \tag{12}$$

The coordinate of vibration x_i can be written as a sinusoidal function in terms of the frequency of the vibration ν_i ; the characteristic frequency of ith normal vibrational mode and time *t*:

$$x_i = x_i^0 \cos(2\pi\nu_i t) \tag{13}$$

Combining equation (12) with equation (13) yields:

© 2017 Tous droits réservés.

$$\alpha = \alpha_0 + \alpha_1 x_1^0 \cos(2\pi\nu_i t) \qquad (14)$$

Where $\alpha_1 = \frac{\partial\alpha}{\partial xi}$ and α_0 is the initial polarizability.

The induced electric dipole moment can be expressed as:

$$P = \alpha_0 E_0 co\,s(2\pi\nu_0 t) + \alpha_1 E_0\; co\,s(2\pi\nu_0 t) cos(2\pi\nu_i t) \qquad (15)$$

Equation 15 can be rearranged using the trigonometric identity:

$$\cos a \cos b = \frac{\cos(a+b)+\cos(a-b)}{2} \qquad (16)$$

and:

$$P = \alpha_0 E_0 co\,s(2\pi\nu_0 t) + \alpha_1 E_0 \frac{cos2\pi(\nu_0+\nu_i)t + cos2\pi(\nu_0-\nu_i)t}{2} \qquad (17)$$

Although equation 17 was obtained using the classical electromagnetic theory, it describes several important properties of Raman scattering processes. First, the polarization and scattering intensity have linear dependence of the laser intensity. It is also apparent that only vibrations that change the polarizability of the molecule are Raman active $\frac{\partial\alpha}{\partial x_i} \neq 0$. The changes in frequency, also known as Raman shift can be positive or negative in respect to the laser frequency. Because$\alpha_1 \ll \alpha_0$, Raman scattering is much weaker than Rayleigh scattering. Equation 15 shows that light will be scattered by the molecule at three frequencies. The first term represents the Rayleigh scattering, the second term contains waves with frequencies ν_0 +ν_i and is known as anti-Stokes Raman scattering (Ozaki & Šašić 2007) and relates the outgoing scattered photons with an increase in frequency by an amount ν_i and finally the third term ν_0 -ν_i, called Stokes Raman scattering is associated with a decrease in frequency of the resulting scattered photon.

© 2017 Tous droits réservés.

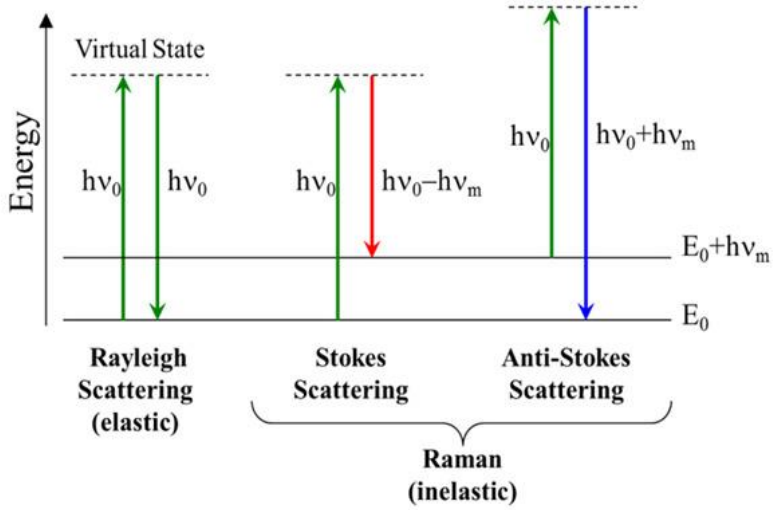


Fig. 7. Energy diagram for Raman scattering (Wikipedia 2016).

The Stokes and anti-Stokes Raman scattering can also be explained using the energy diagram (Fig. 7). Only a small of part (10^{-6} of scattered photons) of incident of incident photon will suffer inelastic scattering. The origin of inelastic scattering can be explained in terms of energy transfer between incident radiation and scattering molecule.

The ratio between intensities of the Stokes and anti-Stokes scattered light depends on the population of the vibrational ground and excited states and can be calculated using Boltzmann's equation (McCreery 2000):

$$\frac{I_{Stokes}}{I_{anti-Stokes}} = \left(\frac{\nu_0-\nu_1}{\nu_0+\nu_1}\right)^4 \frac{h\nu_1}{e^{k_b T}} \tag{18}$$

Where T is the absolute temperature, k_B is the Boltzmann constant.

Equation 18 highlights the proportionality of the Raman intensity with fourth power of the frequency. In general the Raman scattering intensity can be expressed as (Jestel 2010):

$$I = KI_0\alpha^2(\nu_0 \pm \nu_i)^4 \tag{19}$$

© 2017 Tous droits réservés.

Where *K* represents a series of constants, I_0 is the intensity of the incident radiation. It should be noted, that for full explanation of Raman scattering the quantum theory of Raman scattering explains shortcomings of classical Raman scattering.

A Raman microspectrometer (RMS) consists of a specially designed Raman spectrometer integrated with an optical microscope (Fig. 8). Confocal Raman microscopy was developed in 70's by Delhaye and Dhamelincourt (Delhaye & Dhamelincourt 1975).

This microanalytical technique allows acquiring Raman spectra of microscopic samples or microscopic areas of macroscopic samples. Raman microspectrometry combines capabilities of Raman scattering and the spatial resolution of optical microscopy is well adopted for obtaining direct molecular information on individual micron size aerosol particles under ambient conditions (Ault et al. 2014; Jung et al. 2014; Offroy et al. 2015; Sobanska et al. 2014; Sobanska et al. 2012; Brunamonti et al. 2015). Currently automated Raman systems are available for acquiring two-dimensional molecular images with a lateral resolution limited by light diffraction.

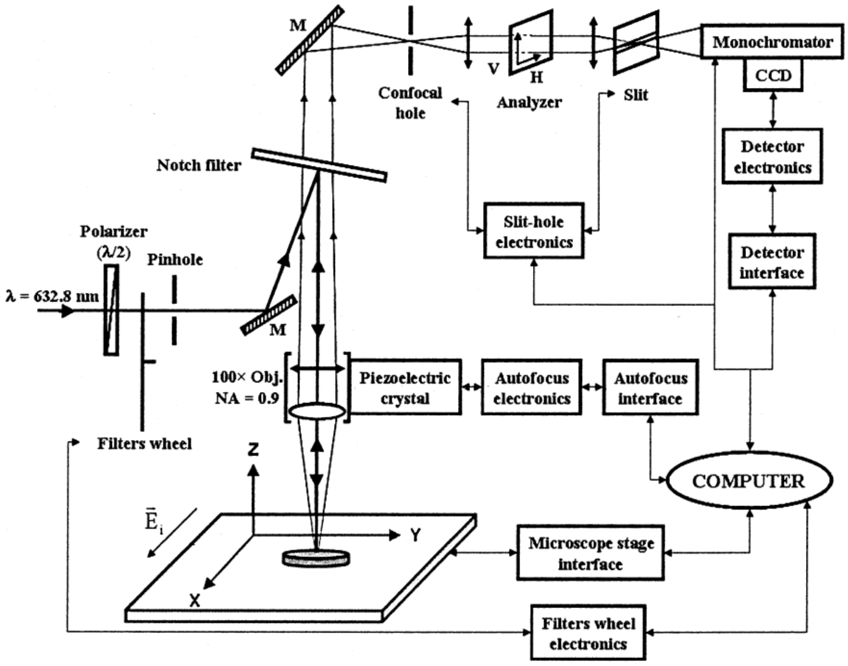


Fig. 8. Raman microspectrometer scheme.

© 2017 Tous droits réservés.

2.2.2. FTIR

Infrared spectroscopy is based on the interaction between matter and radiation in the IR range of the electromagnetic spectrum and is used to study fundamental vibrations and associated rotational-vibrational structure. A vibrational mode is active in infrared absorption spectroscopy if the derivative of the molecular dipole moment (p) with respect to the normal coordinate is nonzero: dp/dQ≠0.

Fourier transform infrared (FTIR) spectroscopy is a measurement technique for collecting infrared spectra using a Michelson's interferometer. The Fourier transform (FT) changes a signal (or any data) from the time domain to the frequency domain (and back again through the inverse FT) where *f*(*t*) is the signal in the time domain and *F*(*ω*) is the signal in the frequency domain:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,\partial t \qquad (20)$$

Any signal in the time domain is a sum of numbers at discrete points in time. Instead of summation, it is common to let the terms approach zero and do an integral to reveal a convolution of the delta function that returns *f*(*t*).

$$\sum f(t_n)\partial(t-t_n) \rightarrow \int f(t_n)\partial(t-t_n) \qquad (21)$$

As opposed to the summing up all these points, wave functions are used to cover all points in time using Euler's formula.

$$e^{ix} = \cos(x) + isin(x) \qquad (22)$$

Wave packets have magnitude *M* with an exponential term e^{iwt}. *Acos(ωt)* is the real part; proportional to cosine with amplitude *A* and imaginary part proportional to the sin with amplitude *Bsin (ωt)*. Any time domain signal can be represented by a sum of all possible combinations of sinusoidal waves with ω_n, frequencies and A_n, B_n amplitudes that stretch over all time at the right magnitude.

$$Me^{i\omega t} = Acos(\omega t) + iBsin(\omega t) \qquad (23)$$

© 2017 Tous droits réservés.

In simplistic terms, eq. 20 states the amount of signal with frequency (ω) in *f*(*t*) is equal to:

$$\frac{f(t)}{e^{iwt}} = f(t)e^{-iwt} \tag{24}$$

The above equation states as *f*(*t*) is a single number and does not have frequency components. More correctly, the amount of signal *F(ω)* is calculated by integrating over all values of *t*:

$$F(\omega) = \frac{f(t)}{e^{iwt}} = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\, \partial t \tag{25}$$

For each frequency, the amplitudes of the real and imaginary parts as a function of omega *F(ω)*. The output from the Discrete Fourier transform (DFT) produces a complex number where the magnitude is the real part of the signal and the phase, which is the initial angle of the wave. In Fourier transform spectroscopy, all wavelengths of light enter in parallel, simultaneously producing interference patterns for each one. The multiplex (Fellgett) principle states this as an advantage of reduced measurement time in comparison with a continuous wave spectrometer that observes only a single wavelength at a time (Griffiths & De Haseth 2007). There are no requirements for entrance and exit slits as wavelengths are being measured in parallel so the interferometer's output light intensity is almost equal to the input intensity, which makes signal detection easier (Griffiths & De Haseth 2007).

For a given wavelength or frequencies of IR radiation striking a sample, these two interactions are inversely related through the following equation:

$$A = log1/T \tag{26}$$

Where: *A*= absorbance and *T*= transmittance *(%T/100*).

IR spectral region of the electromagnetic spectrum extends from the red end of the visible spectrum to the microwave region; it includes radiation with wavenumbers ranging from about 14,000 to 20 cm^{-1}, (wavelengths from 0.7 to 500 µm). Because of

© 2017 Tous droits réservés.

application and instrumentation reasons, it is convenient to divide the IR region into the near (NIR), middle (mid-IR), and far (FIR) subregions.

The near-IR (NIR, NIRS) domain extends from the visible region at 14,000 cm^{-1} (0.7 µm) to the mid-IR region at 4000 cm^{-1} (2.5 µm). Spectra generated in the near-IR region consist of many overtones and combinations of the mid-IR region fundamental vibration modes. Since all organic species absorb in the NIR and produce many overlapping bands, single band spectroscopy and qualitative band assignments are nearly impossible. NIR is useful for quantitative work, including *in situ* monitoring of reactions (Fontalvo-Gómez et al. 2013).

The spectral range of greatest use for chemical analysis is the mid-IR (MIR) region (Lewandowski et al. 2015). It covers the frequency range from 4000 to 500 cm^{-1} (2.5-20 µm). This region can be subdivided into the group frequency region, 4000-1300 cm^{-1} (2.5-8.0 µm) and the fingerprint region, 1300-500 cm^{-1} (8.0-20 µm). The absorption bands in the fingerprint region of the spectrum are the results of single-bond as well as skeletal vibrations of polyatomic systems. Multiple absorptions in this region make it difficult to assign individual bands, but the overall combined pattern is very characteristic, reproducible, and useful for material identification when it is matched to reference spectra (Huber et al. 2007).

The far-IR (FIR) region is generally designated as 500-20 cm^{-1} (20-500 µm). In this region, the entire molecule is involved in low frequency bending and torsional motions, such as lattice vibrations in crystals. These molecular vibrations are particularly sensitive to changes in the overall structure of the molecule that are difficult to detect in the mid-IR region. For example, the far-IR bands of amino acids can often be differentiated (Matei et al. 2005). FTIR is also useful in the identification and differentiation of minerals (Brusentsova et al. 2010).

© 2017 Tous droits réservés.

2.2.3. Scanning electron microscope with energy dispersive X-ray spectroscopy (SEM/EDS)

SEM/EDS provides an image with a high spatial resolution and a deep field of view owing to interaction of an electron beam with an observed object (surface). As a result of matter exposure to high-energy electrons, X-ray radiation is emitted and recorded by an X-ray detector (based on energy or wavelength dispersion – EDS or WDS, respectively) coupled with SEM. EDS detectors are more efficient an faster, compared to WDS, although at the expense of spectral resolution, mich more efficient in case of WDS Over the past decades, EDS has become firmly established in the aerosol scientific community as a powerful technique for specification of the elemental composition of substrate-collected particles. Scanning electron microscopy coupled with energy dispersive spectroscopy (SEM/EDS) yields images with ≥3 μm lateral resolution where the EDS provides elemental analysis with an accuracy of 0.1 - 1 at%. In most studies on atmospheric particles, the main objective is a general characterization of the aerosol. In the conventional system, the sample is measured in vacuum and thus dehydratation of particles occurs. In addition, the specimen morphology may change, what in conjunction with the previous assumption makes potentially difficult to observe the hygroscopic behaviour of the particles and analysis of semi-volatile compounds. These problems have been partly overcome by the recent developments of electron microscopes in which a sample can be studied in low-vacuum conditions. The environmental SEM (ESEM) is now an established tool in the study of atmospheric particles (Zimmermann et al. 2007; Rachel E. O'Brien et al. 2015; Chen et al. 2013).

Straightforwardly, when the electron beam hits a sample, there is a high probability that X-rays will be generated. The produced X-ray escapes the sample and reaches the detector generating a charge pulse. This short-lived current is then converted into a voltage pulse with amplitude reflecting the energy of the detected X-ray. Finally,

© 2017 Tous droits réservés.

this voltage pulse is converted to a digital signal and one more count is added to the corresponding energy channel. Once the measurement is completed, the accumulated counts produce a typical X-ray spectrum with the major peaks superimposed on the background. However, high-energy electrons can interact with samples' atoms in many different ways. Some signals are used for imaging (secondary electrons, BSEs, transmitted electrons, etc.), but the X-ray signal will be discussed here. In the case of the X-rays emission, an electron (from the beam) strikes an atom, which ejects an electron originally positioned in an inner shell (K shell) (Fig. 9). When an inner shell electron is displaced by collision with a primary electron, an outer shell electron may fall into the inner shell to re-establish the proper charge balance in its orbitals following an ionization event. Thus, by the emission of an X-ray photon, the ionized atom returns to ground state. Therefore, the energy released (expressed in eV) is exactly equal to the energy difference between the two levels (Fig. 9). In addition to the characteristic X-ray peaks, a continuous background is generated through the deceleration of high-energy electrons as they interact with the electron cloud and with the nuclei of atoms in the sample. This component refers to the Bremsstrahlung or Continuum X-ray signal. This constitutes a background noise and is usually stripped from the spectrum before analysis although it contains information that is essential to the proper understanding and quantification of the X-ray spectrum.

© 2017 Tous droits réservés.

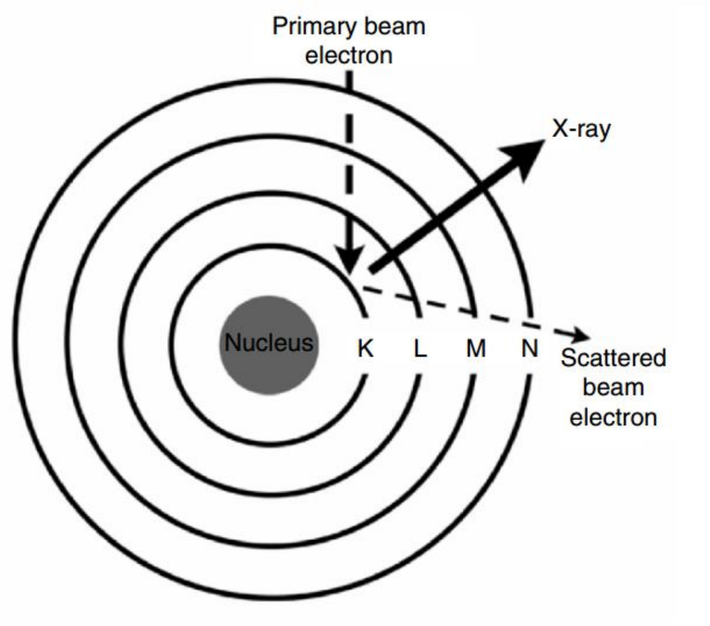


Fig. 9. X-ray generation in a sample from the interaction of high-energy electrons in an electron microscope.

The basis for elemental analysis with EDS is the Moseley's Law (Eq. 27).

$$E = C_1(Z - C_2)^2 \tag{27}$$

Where: E = energy of the emission line for a given X-ray series (e.g. Kα), Z = atomic number of the emitter C_1 and C_2 are constants.

The energy of the characteristic radiation within a given series of lines varies monotonically with the atomic number.

In the SEM/EDS microanalysis the net peak intensity, that is, the intensity of the characteristic X-ray signal above the background signal is measured. However, it should be noted that counting error in any measurement of peak intensity might occur. In evaluating the intensity of a spectral peak, that is, the number of X-ray counts, a source of random error is current. More precisely, the emission and subsequent detection of a characteristic X-ray can be regarded as a statistically independent event, which has a fixed probability of occurring within each faint time interval ***δt***. Under conditions such as these, the number ***n*** of X-rays detected during any finite time interval is governed by the Poisson law (Eq. 28):

© 2017 Tous droits réservés.

$$P(n) = \frac{e^{-\bar{n}}\,\bar{n}^{n}}{n!} \tag{28}$$

Where *P(n)* is the probability of detecting exactly ***n*** X-rays and $\bar{n}$ is the mean number of X-rays counted during a large number of such trials. The confidence in the accuracy can be no greater than that indicated by the extent of the Poisson distribution plot of *P(n)* versus *n*. This unavoidable error is called the counting error.

The background signal itself is responsive to counting error. Therefore, the confrontation with the problem of distinguishing between random fluctuations in the background and real peaks is needed. Furthermore, a confidence level must be established as well as maintained in any assertion that an element is present at the minimum detection limit. For example, a 95% confidence level would be consistent with the statement that, in a large number of observations, 95% of the observations indicating the presence of an element at the minimum detection limit reflect the actual presence of that element, whereas 5% of such observations reflect only random fluctuations in background counting rate. In practice, minimum detection limits are influenced by a number of experimental factors including instrument stability, spectral peak overlaps, and interactions within the sample matrix. For routine EDS analysis, the generic detection limit is about 1000 ppm or 0.1 wt%.

An important aspect which must be faced in the SEM/EDS analysis is the determination of low-Z elements such as carbon, nitrogen and oxygen, with comparable analytical abilities for heavier elements (Z≥11) observed by the conventional technique. The low-Z elements are very important because they form the major mass of aerosols (Prather et al. 2008; Bzdek et al. 2012; E A Stefaniak et al. 2009). By the application of the SEM/EDS technique, which employs either a windowless or thin-window EDS detector (Van Grieken et al. 2000; Ro et al. 1999), chemical compositions, including the low-Z components, of individual particles can be at least semi-quantitatively elucidated. For the last two decades, there has been an extensive progress in the field of single particle analysis by SEM/EDS. It began with instrumental developments, followed by the design of computer software facilitating

© 2017 Tous droits réservés.

an automated mode of particle detection based on backscattered electron images and then development of the software for estimation of element weight concentrations (based on CASINO simulations) in each recognized particle in semi-quantitative analysis (Osán et al. 2000; Ro et al. 2004). The development in this matter was due mostly to the quantification methods in the EDS. The classical ZAF and $\varphi(\rho z)$-based procedures aim to correct for matrix and geometric effects which are even more pronounced for light-element X-rays. The most reliable and widely used quantification method for microparticles is the so-called particle-ZAF algorithm developed by Armstrong et al. (Armstrong & Buseck 1985). The particle-ZAF methods based on bulk standards introduce large errors for the light elements, mostly because of the large absorption correction needed, and the difference between the behaviour of bulk samples and single particles under electron bombardment. Also, when the average atomic number of the substrate differs significantly from that of the particle, the side-scattering correction (Armstrong & Buseck 1985) of the $\varphi(\rho z)$ function is reasonable only if the electron excitation volume is smaller than the particle itself. The schematic illustration of interaction volumes for various electron-specimen interactions is presented on Fig. 10.

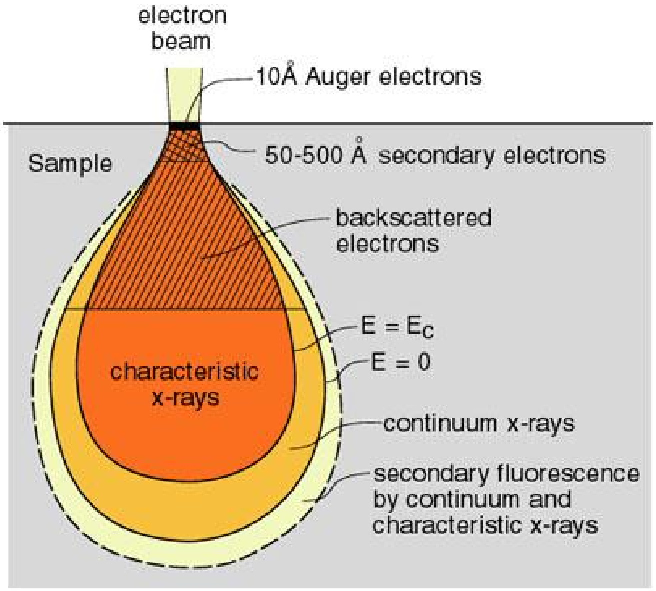


Fig. 10. Schematic illustration of interaction volumes for various electron-specimen interactions (https://nau.edu/cefns/).

© 2017 Tous droits réservés.

The interaction volume is dependent on the mean atomic number of the sample and the operating conditions of the SEM. The quantification procedure for low-Z elements is based on a modified version of the single scattering CASINO Monte Carlo program (Gauvin et al. 1995; Hovington et al. 1997), which was designed for low-energy beam interaction to generate X-ray and electron signals. The modified version of the CASINO program allows the simulation of electron trajectories in spherical, hemispherical and hexahedral particles located on a flat substrate (Ro et al. 1999). The simulation procedure also determines the characteristic and continuous X-ray intensity emitted from the substrate material (Osán et al. 2000). Using a combination of simulations and successive approximation, a reverse Monte Carlo quantitative program was developed for standardless concentration determination from characteristic X-ray intensities obtained from SEM/EDS of individual particles (Ro et al. 2003). To sum up, by means of SEM/EDS it is possible to semi-quantitatively determine the concentrations of low-Z elements such as C, N and O by application of the quantitative programs based on Monte Carlo simulation, as well as higher-Z elements that can be analysed by conventional energy-dispersive electron probe X-ray microanalysis (Maskey & Ro 2011). The morphology and the composition of environmental microparticles are heterogeneous, and therefore the detailed information about both major and trace constituent elements is required. Due to this assumption, SEM/EDS is an appropriate technique for aerosol particles analysis.

2.3. Chemometric methods for spectral data treatment

PCA

In this work, PCA was performed in order to study the correlation among the mine dust particles, their composition and chemical mixing. In the case of a large data set processing, PCA was used for the results from single particle (analysis (SEM/EDS and

© 2017 Tous droits réservés.

Raman). Such data treatment was applied to determine pattern formation after application of the authorial algorithm for Raman spectra. PCA was also important for separation of the FTIR spectra of organic pigments after preprocessing. PCA was also an element of exploratory data analysis in order to support specification of optimal number of clusters for clustering methods.

Clustering

The several clustering algorithms (HCA, k-means, fuzzy-c-means) were applied for the results from almost all samples presented in this work. The application of such a proceeding was made due to the specification of the main groups of objects (e.g. Raman spectra, chemical mixing membership) with relatively high similarity. The cluster analysis was intended to present the main groups of the aerosol particles and organic pigments (based on the collected spectra).

MCR

The large heterogeneity at the level of individual particles in environmental samples generates severe overlap of spectral information to obtain pure component spectra and concentrations. The application of multivariate curve resolution algorithm was made due to the specification of pure compounds from mixed Raman spectra. This algorithm is also an integral part of the authorial data analysis algorithm for specification of the chemical mixing of aerosol particles. Moreover, the results from the MCR procedure are the core of the results for the specification of the most suitable single aerosol particles imaging substrate for Raman microspectroscopy.

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

CHAPTER 3: CHARACTERIZATION OF INDIVIDUAL PARTICLES BY COMBINATION OF RAMAN MICROSPECTROMETRY AND CHEMOMETRY

© 2017 Tous droits réservés.

3.1. The evaluation of the influence of collecting substrate on the Raman mapping of aerosol particles

Since SPA techniques are off-line techniques, particles must be collected on suitable substrates for analysis. Thus, the choice of an analytical substrate for single particle analysis has to be made wisely. Obviously, the substrate should be characterized by optimal contrast adjustment, if optical images are considered, and by chemical inertness, to avoid any modification of the chemical composition and morphology of the particles. The substrate may have a large signal contribution compared to the relevant information in the sample and thus impair final results. This is particularly crucial when the particle size is lower than the beam spot size that is typically encountered for atmospheric micron-sized particles. In this chapter, we present the evaluation of the common particle-collecting substrates in the context of the MCR approach application for Raman spectra. Laboratory-generated single-component particles of calcite ($CaCO_3$) and mixed particles of calcite ($CaCO_3$), nitratine ($NaNO_3$), hematite (Fe_2O_3) and anglesite ($PbSO_4$) were deposited by cascade impaction on: Ag, In, Si, SiO_2, microscope slide and TEM-grid substrates and analysed by RMS. The evaluation of the spectral contribution exported by the MCR was made in reference to the specification of the optimal analytical substrate for the RMS mapping of aerosol particles.

© 2017 Tous droits réservés.

Analytica Chimica Acta 1014 (2018) 41–49

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Influence of collecting substrate on the Raman imaging of micron-sized particles

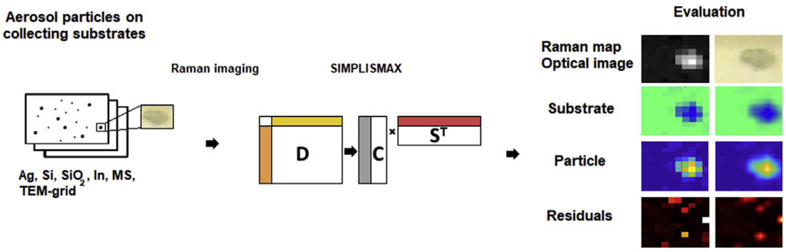
Guillaume Falgayrac [a, *], Damian Siepka [b, d, c], Elżbieta A. Stefaniak [c], Guillaume Penel [a], Sophie Sobanska [b, d, **]

[a] Univ. Lille, Univ. Littoral Côte d'Opale, EA 4490 - PMOI - Physiopathologie des Maladies Osseuses Inflammatoires, F-59000 Lille, France
[b] Laboratoire de Spectrochimie Infrarouge et Raman, UMR CNRS 8516, Lille 1 University – Science and Technology, Bat. C5, 59655 Villeneuve d'Ascq Cedex, France
[c] Laboratory of Composite and Biomimetic Materials, Centre for Interdisciplinary Research, The John Paul II Catholic University of Lublin, Konstantynów 1J, 20-708 Lublin, Poland
[d] Institut des Sciences Moléculaires, UMR CNRS 5255, University of Bordeaux, 351 cours de la Libération, 33405 Talence, France

HIGHLIGHTS

- Evaluation of 6 substrates for Raman imaging and multivariate curve resolution.
- Substrate contribution is superior to micron-sized aerosol particle contribution.
- TEM Grid is the suitable substrate for micron-sized aerosol particle.
- Resolved spectra of compounds are always impaired by Si substrate contribution.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 September 2017
Received in revised form
2 February 2018
Accepted 4 February 2018
Available online 14 February 2018

Keywords:
Raman spectroscopy
Micron-sized
Particles
Multivariate curve resolution
Substrates
Imaging

ABSTRACT

The influence of six common substrates on the Raman imaging of micron-sized inorganic aerosol particles was examined. Laboratory-generated single-component particles of calcite ($CaCO_3$) and mixed particles of calcite ($CaCO_3$), nitratine ($NaNO_3$), hematite (Fe_2O_3) and anglesite ($PbSO_4$) were deposited by cascade impaction on Ag, In, Si, SiO_2, microscope slide and TEM-grid substrates. The spectral contribution of substrates to Raman images of the deposited particles was evaluated by Multivariate Curve Resolution. The shape and intensity of the substrate spectra affect the effectiveness capability of the spectral deconvolution. The substrates were characterized and compared with respect to their effect on the reconstruction of Raman images of aerosol particles. The TEM-grid substrate yielded spatially stable sample measurements with a homogeneous spectral contribution, satisfactory Raman map reconstruction and the potential for application in other techniques (e.g., SEM-EDX).

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Atmospheric aerosol particles suspended in the air, with their wide variety of sources and physico-chemical properties, strongly contribute to environmental quality [1]. They may scatter or absorb light and may affect the radiative balance of the atmosphere [2]. It

* Corresponding author. Univ. Lille, EA 4490 - PMOI - Physiologie des Maladies Osseuses Inflammatoires, F-59000 Lille, France.
** Corresponding author. Institut des Sciences Moléculaires, UMR CNRS 5255, University of Bordeaux, 351 cours de la Libération, F-33405 Talence, France.
E-mail addresses: guillaume.falgayrac@univ-lille2.fr (G. Falgayrac), sophie.sobanska@u-bordeaux.fr (S. Sobanska).

https://doi.org/10.1016/j.aca.2018.02.015
0003-2670/© 2018 Elsevier B.V. All rights reserved.

has been shown that selected aerosol particles act as cloud condensation nuclei, affecting the lifetime of clouds [3]. Aerosol particles might have a substantial impact on human health [4]. Both the atmospheric and health-related impacts of aerosol particles are related to their chemical composition, morphological features, size and sources [5]. Furthermore, the chemical heterogeneity of ambient aerosol particles is an essential parameter, as the particle size and elemental concentration are not sufficient to properly estimate the environmental impact of aerosol particles, namely, bioaccessibility [6], toxicological effects [7], reactivity [8] and optical properties [9]. Several microanalytical techniques can be used to determine chemical heterogeneity at the single-particle scale because their lateral resolution is comparable to the size of particles collected from ambient air [10]. Among the single-particle analysis (SPA) techniques, Raman microspectroscopy (RMS) is a powerful technique that provides both the molecular composition and imaging of micron-sized aerosol particles [11–13]. RMS has been used to resolve the molecular composition and heterogeneity of individual aerosol particles [12,14–17]. Raman spectra may be complex due to the contribution of multiple compounds within one particle. Thus, chemometric methods such as Multivariate Curve Resolution (MCR) have found an application in the separation of spectral contributions from chemical compounds within aerosol particles and can significantly improve Raman images [13,14,17,18], even beyond resolution limits [19]. Since SPA techniques are off-line techniques, particles must be collected on suitable substrates for analysis. Thus, the choice of an analytical substrate for single particle analysis has to be made wisely. Obviously, the substrate should be characterized by optimal contrast adjustment, if optical images are considered, and by chemical inertness, to avoid any modification of the chemical composition and morphology of the particles. The substrate may have a large signal contribution compared to the relevant information in the sample and thus impair final results. This is particularly crucial when the particle size is lower than the beam spot size that is typically encountered for atmospheric micron-sized particles. The issue of substrate selection has already been considered for several single particle analysis techniques [20–22]. Only one publication has referred to the evaluation of substrate for both RMS and sequential electron probe X-ray microanalysis using thin-window energy-dispersive X ray detection (TW-EDX-EPMA) for analysis of aerosol particles [23]. Several substrates were examined in the study: carbon tape, nucleopore filter, silicon wafer, beryllium disc, TEM-grid, aluminium wafer and silver wafer. The study focuses on the features of the substrate, such as roughness, as well as the influence of measurement parameters, i.e., laser wavelength (514 nm and 785 nm), laser power density, objective (× 100 or × 50) and finally the size of the single particle. In this study, the effect of substrate contribution on spectral intensity, and thus on Raman imaging, was not investigated.

One reason for the application of MCR methods is the ability to separate the contribution of each compound and substrate, then reconstruct their respective spatial distributions. The use of an inadequate substrate may impair the accuracy of the calculated spatial distribution. The use of a suitable substrate for Raman microspectrometry imaging combined with MCR methods has not yet been evaluated. Developing interest in molecular imaging in nanoscience, including the environmental fate of produced micro- and nano-objects, has necessitated improvements in submicrometric analysis, including the selection of a suitable substrate. The substrate's spectral contribution influences analysis, its contribution might increase while the size of the object analysed decreases. Therefore, the choice of substrate cannot be neglected when MCR methodology is applied. This work concerns the evaluation of 6 substrates commonly used for particle collection: Si-wafer, Ag layer, In layer, TEM-grid, SiO_2 and microscope slide (MS). We have investigated for the first time the influence of substrate contribution on the spectral characterization of single-component and mixed micro-sized particles by using both RMS and MCR.

2. Material and methods

2.1. Substrates

Six substrates that fulfil the criteria for application in RMS analysis were used: (i) silver (thickness ~ 1 mm) (named Ag) and (ii) indium (thickness ~ 1 mm) layers (named In), each separately sputtered on a microscope slide (10 × 10 mm); (iii) grid for transmission electron microscopy (TEM-grid, AGAR Scientific F1 type G2761C, diameter = 3.05 mm), which was covered with a thin formvar film and a nanometric layer of carbon (named TEM-grid); (iv) Si-wafer (10 × 10 mm, Interuniversity Micro-electronic Centre, Belgium) (named Si); (v) SiO_2 (10 × 10 mm, silica slide, optical quality, from Alfa Aesar) (named SiO_2); and (vi) standard microscope slide, 10 × 10 mm (named MS).

The selection criteria were as follows: chemical homogeneity and inertness, substrates inert to laser excitation, a flat surface, and compatibility with the impaction sampling system.

2.2. Materials

Calcite ($CaCO_3$), hematite (Fe_2O_3), nitratine ($NaNO_3$) and anglesite ($PbSO_4$), fine powders with a purity of 99.99%, were used as model compounds usually found in natural and industrial aerosols [24]. The particle size was below 10 μm.

2.3. Preparation of the samples

The substrates were first ultrasonically cleaned in a mixture of ethanol and deionized water (50/50 vol) for 15 min to remove potential contamination by indoor particles. In this paper, "clean substrates" refers to substrates without any impacted particles. "Impacted substrates" refers to substrates with impacted particles.

Aerosolized calcite particles were generated in a homemade turbulent airflow reactor [15]. A mass of 0.125 g of calcite powder was introduced into the reactor to generate particles. The particles were collected by an inertial cascade impactor (PM10 Dekati). The substrates were mounted on one impaction plates corresponding to particles with an aerodynamic diameter ranging from 10 to 2.5 μm. The collection time was set for 3 min to avoid substrate overloading. A similar protocol was used for mixed aerosol particles composed of calcite, hematite, nitratine and anglesite. The content of each compound was fixed to 25% (w/w) (0.125 g), and the resulting mixture was then introduced into the airflow reactor. As described previously, the particles were collected on the 10–2.5 μm stage for 3 min. This methodology allows the production of mixed aggregates, as demonstrated in our previous work [15]. Both clean and impacted substrates were analysed by Raman microspectroscopy without further preparation. In this study, the stage 10–2.5 μm was used for the Raman analysis. Particles with a diameter of 5 μm were selected for the sake of the comparison between all substrates.

2.4. Raman microspectroscopy

Each substrate (with and without aerosol particles) was analysed by means of a Labram confocal Raman microspectrometer (Horiba, Jobin-Yvon) equipped with a 100 × , 0.9 numerical aperture Olympus objective. Raman scattering was excited with the 632.8 nm wavelength of a He–Ne laser. Acquisitions were performed at 10 mW. The spot diameter of the laser beam at the sample was measured at 1 μm. The applied system uses a high-precision piezo translator and feedback signal to automatically track and adjust the laser focus on the sample to ensure a perfect focus for each measurement. XYZ computer-controlled Raman mapping recorded spectra in a point-by-point XY scanning mode (y rows, x points per row) with a 1-μm step and 30-s integration time. Raman mapping generates a three-dimensional data set (X × Y × λ), i.e., X × Y spectra, each containing $\lambda = 2040$ spectral elements, corresponding to a spectral window of 200–1200 cm^{-1} with a spectral resolution of 4 cm^{-1}. Raman maps were acquired successively on clean substrates, substrates impacted with calcite and substrates impacted with a mixture of calcite, hematite, nitratine and anglesite. For clean substrates, the size of the Raman map was fixed at 3 × 3 pixels, with a 1-μm step for all acquisitions, i.e., a total of 9 spectra for one map and, thus, 9 spectra in the data matrix. For substrates impacted with particles, large maps of 10 × 10 pixels for the Ag, In, TEM-grid, SiO_2 and MS substrates and 15 × 15 pixels for the Si substrate were acquired with 1-μm steps. For MCR data treatment and comparison with clean substrate results, an area of 3 × 3 pixels, giving a total data matrix of 9 spectra, was selected from the large maps at the centre of the particle (or aggregate). The particles (or aggregates) were selected to have an apparent geometric diameter of 5 μm, based on their optical image. Ten maps were acquired for each sample to insure the reproducibility of the results. The results presented in this study are representative of the results obtained for the 10 images per substrate and per condition (clean and impacted substrates). The LABSPEC software was used for spectral acquisition and cosmic spike removal, while the data pre-processing and processing were performed using PLS-Toolbox (Eigenvector Research Inc.).

2.5. Data processing

The collected spectra in the Raman maps were normalized (Block Variance Scaling), smoothed (Savitzky-Golay smoothing with filter width = 5) and then decomposed by the SIMPLe-to-use Interactive Self-modelling Mixture Analysis approach (PURITY in PLS-Toolbox) [25]. Briefly, the matrix D_{exp} (experimental data set) is decomposed into the product of three matrices: C, S^T and E:

$$D_{exp} = C \cdot S^T + E$$

where matrix C contains the spectral contributions of the k resolved pure compounds per pixel, S^T is the transposed S matrix of the pure compounds' Raman spectra and E represents the matrix of the residuals not explained by the model.

The experimental data set D_{exp} and the calculated data set D_{calc} ($= C \cdot S^T$) should be very similar. The equation for the difference between D_{exp} and D_{calc}, i.e., the relative root sum of squares (RRSSQ), is defined as follows:

$$RRSSQ = \sqrt{\frac{\sum\sum\left(d^{exp}_{i,j,\lambda} - d^{calc}_{i,j,\lambda}\right)^2}{\sum\sum\left(d^{exp}_{i,j,\lambda}\right)^2}}$$

where $d_{i,j,\lambda}$ is the i × jth row and λ^{th} column element of D and $d^{calc}_{i,j,\lambda}$ is the i × jth row and λ^{th} column element of D_{calc}.

RRSSQ charts the relative difference between the original and reconstructed data sets. The lower the RRSSQ, the closer the reconstructed data set is to the experimental one. It should be emphasized that an RRSSQ lower than 5% provides a realistic overview of the components in the data set [26].

In practice, the application of PURITY begins with the selection of a characteristic wavenumber of a pure compound. Then, the algorithm extracts the spectral contribution from each pure compound spectrum for each pixel. Wavenumber selection is obvious for the data sets of clean substrates such as Si and Ag, due to their thin and well-defined Raman bands. The wavenumber selection from the data sets of In, MS, SiO_2 and TEM-grid is less clear because their Raman bands are broader. First, the optimal wavenumber was identified by comparing the RRSSQs as a function of the wavenumber range. The optimal wavenumber in this paper is identified as the wavenumber corresponding to the lowest RRSSQ for the substrate.

Second, the homogeneity of the Raman signals of clean substrates was estimated by comparing the spectral contributions for each pixel. Ideally, the spectral contribution for each pixel should be equal to 100%. However, this outcome was not observed due to the residuals remaining after decomposition. The residuals represent spectral components that are not resolved into a pure spectrum by PURITY. These variations result from contributing factors that influence the spectrum, such as focus volume, laser intensity drift, instrument optics, and dark current. The residual value was exported for each pixel in the map. The residual contribution was calculated as the difference between the sum of the spectral contributions of resolved compounds and a normalized value of 100%.

The same approach was performed for impacted substrates. The residuals were exported for both clean and impacted substrates and then compared among the different substrates. The abundance of residuals was used as comparative parameter to evaluate the substrates for Raman imaging combined with MCR analysis.

3. Results and discussion

3.1. Analysis of clean substrates

Raman spectral maps were acquired for each clean substrate as described in the experimental section. Si and Ag are characterized by thin bands at 520.7 cm^{-1} and 180 cm^{-1}, respectively. The Raman spectra of In, MS, SiO_2 and TEM-grid have broad bands. In and SiO_2 have a broad band in the 180-650 cm^{-1} region, whereas the microscope slide (MS) has a broad and low intense band in the 500-650 cm^{-1} region. The TEM-grid has broad and intense bands at 1300 and 1600 cm^{-1}, which are characteristic of its amorphous carbon support membrane (Fig. 1).

Raman maps of clean substrates were acquired, then the data were used for PURITY analysis. The first step was to identify the

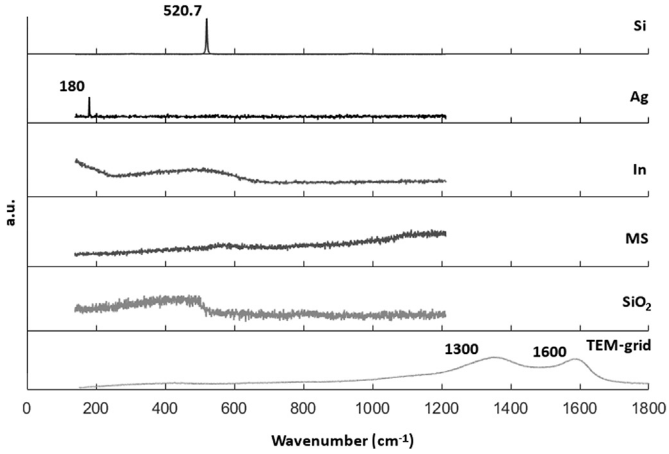


Fig. 1. Reference Raman spectra of the 6 substrates: Si, Ag, In, MS (Microscope Slide), SiO_2 and TEM-grid.

optimal wavenumber for pure variable selection. For this purpose, the calculated RRSSQ values were plotted (Y variable) against the wavenumber range (X variable) (Fig. S1 in SI). As expected, the optimal wavenumbers are 180 and 520 cm^{-1} for Ag and Si substrates, respectively, while the values for TEM-grid, In, SiO_2 and MS are 1200, 250, 256 and 551 cm^{-1}, respectively. These wavenumbers were chosen to have RRSSQ values less than 5%, except for SiO_2, for which the minimum RRSSQ was 6.4%.

The homogeneity of Raman spectra collected from clean substrates was estimated by comparing the spectral contribution (C matrix) of the analysed substrate with the remaining residuals (Fig. 2). The spectral contribution values were normalized as a part of the sum for each pixel, and thus the sum of normalized contributions was equal to 100%. The spectral contributions of the substrate and residuals were calculated by the procedure in the PLS-Toolbox for collected Raman maps. The remaining residuals in all Raman maps have a contribution lower than 1%, which indicates the suitability of the decomposition process, due to the satisfactory homogeneity of the data. The In, Si, and MS substrates have the lowest residual contributions, not exceeding 0.05%. In the case of Ag and SiO_2, the residual values are also low: 0.1% and 0.08%, respectively. The residual contribution for TEM-grid is approximately 1% and represents the highest value in this comparison. This could be explained by the variation in background among points analysed on the amorphous carbon membrane. Moreover, the extracted spectra are all correctly resolved and correspond to their reference spectra (Fig. 3).

In this comparison, the results showed that except for TEM-grid, all substrates are suitable for multivariate curve resolution due to the small residual values relative to spectral contributions from defined compounds.

3.2. Analysis of substrates with impacted $CaCO_3$ particles

The analytical protocol described previously was applied to substrates with impacted $CaCO_3$ particles. An area of 3 × 3 pixels (1-μm step) was selected from the centre of the calcite particle. The main goal of this process was to evaluate the spectral contributions of the substrates and $CaCO_3$, based on the residuals remaining after PURITY analysis. As mentioned previously, 10 images were recorded for each substrate to evaluate reproducibility. MCR data treatment was applied to each spectral data set, and the standard deviation was calculated for each spectral contribution. For all substrates, particles with a similar diameter size (~5 μm) were used to achieve better comparative conditions. During this procedure, 3 spectra were resolved for each data set, corresponding to the substrate, $CaCO_3$ and the residuals (Fig. 4). The RRSSQ was below 5%, attesting to the correct extraction of pure spectra by MCR. The spectral contributions were normalized. The quality of extracted spectra from the PURITY approach was evaluated for all substrates. The quality of extracted spectra in the S^T matrix in this work refers to an absence of signal from the substrate in the particle spectrum and the apparent signal-to-noise ratio. A low-quality extracted calcite spectrum is clearly observed in the case of the Si substrate (Fig. 4), where an intense artefact peak at 520.7 cm^{-1} remains after the extraction. As a consequence, the procedure did not resolve the Raman spectrum of calcite properly due to the intense signal from the Si substrate. For the other substrates, the quality of the extracted spectra is satisfactory for Raman imaging (Fig. 4) because there is no contribution from the substrate. Moreover, we noticed a better signal-to-noise ratio in the extracted calcite spectrum for the Ag, MS and TEM-grid substrates than for In.

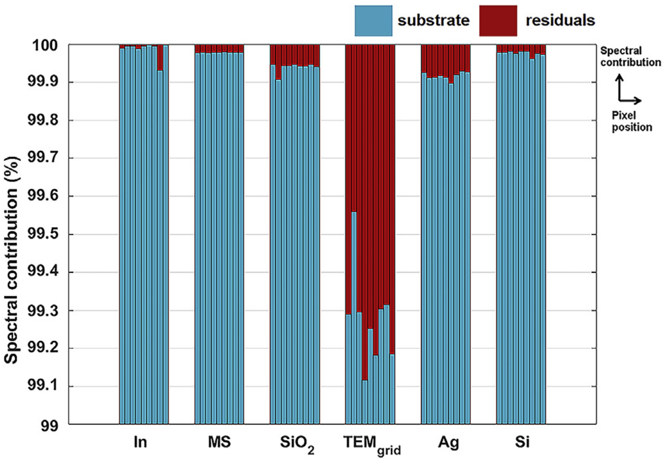


Fig. 2. Spectral contribution and residuals for each clean substrate as function of the position of the pixel. Each column is divided into 9 bars corresponding to the 9 pixels. Each bar represents the spectral contributions from the substrate and the residual within a pixel.

The spectral contribution (C matrix) of calcite depends on the substrate type (Fig. 5). The highest contribution from the calcite particle signal was found for the Ag and Si substrates, while the mean spectral contribution from $CaCO_3$ was less than 0.1% for the other substrates. An apparent subset distinction of substrates with corresponding spectral contribution values can be observed. This distinction is related to the pixel position, i.e., border versus centre of the particle. The average contributions of the substrate, $CaCO_3$ and residuals are reported in Table S1 (in SI files). The average contribution of the In substrate was the highest (98.38%) compared with the other substrates, with the lowest standard deviation (0.94%). The contribution of the calcite particle was relatively low (1.57%), with a low contribution of residuals (0.05%). This situation is also observed in the case of TEM-grid, for which the contribution of the substrate, calcite and residuals are 96.99%, 2.99% and 0.03%, respectively. In the second subset, the SiO_2 and MS substrates have spectral contributions of 93.36% and 90.99%, respectively, while the impacted $CaCO_3$ particle has spectral contributions of 6.56% and 8.97%, respectively. The residual values for both SiO_2 (0.08%) and MS (0.04%) are low. The Ag and Si substrates were placed in the last subset, due to their significantly lower mean spectral contribution relative to an increased contribution from calcite. The spectral contributions from Si and Ag were 28.67% and 35.88%, respectively, whereas the calcite particle contributions were 70.41% and 64.07%. Only for the Si and Ag substrates, their spectral contribution was lower than that from the particle. However, the standard deviation (SD) values calculated for the calcite particle spectral contributions differed significantly among all substrates. The most noticeable example was the standard deviations for In (0.94%) and Ag (22.47%). The SDs for $CaCO_3$ fell within a similar range, with values of 0.99% and 22.47% for In and Ag, respectively. Such a difference in SD values highlighted the variability of spectral contributions for each pixel (Fig. 5). A significant variability in substrate contribution can be observed for each pixel of the Ag layer and Si wafer, while for the In layer, TEM-grid, MS and SiO_2 substrates, values were more homogeneous. This variability was observed in all Raman images for all substrates and reflects the non-reproducibility of the procedure for Ag and Si substrates.

The raw Raman maps and reconstructed Raman maps were both evaluated (Fig. S2 and Fig. S3). Raw Raman maps of calcite and substrates were reconstructed from the total integrated area of each typical bands. No pre-processing was applied on spectra. The reconstructed Raman maps were reconstructed based on spectral contributions as extracted by the PURITY approach (Fig. S3 in SI). The reconstructed shape of the particle in the Raman map is

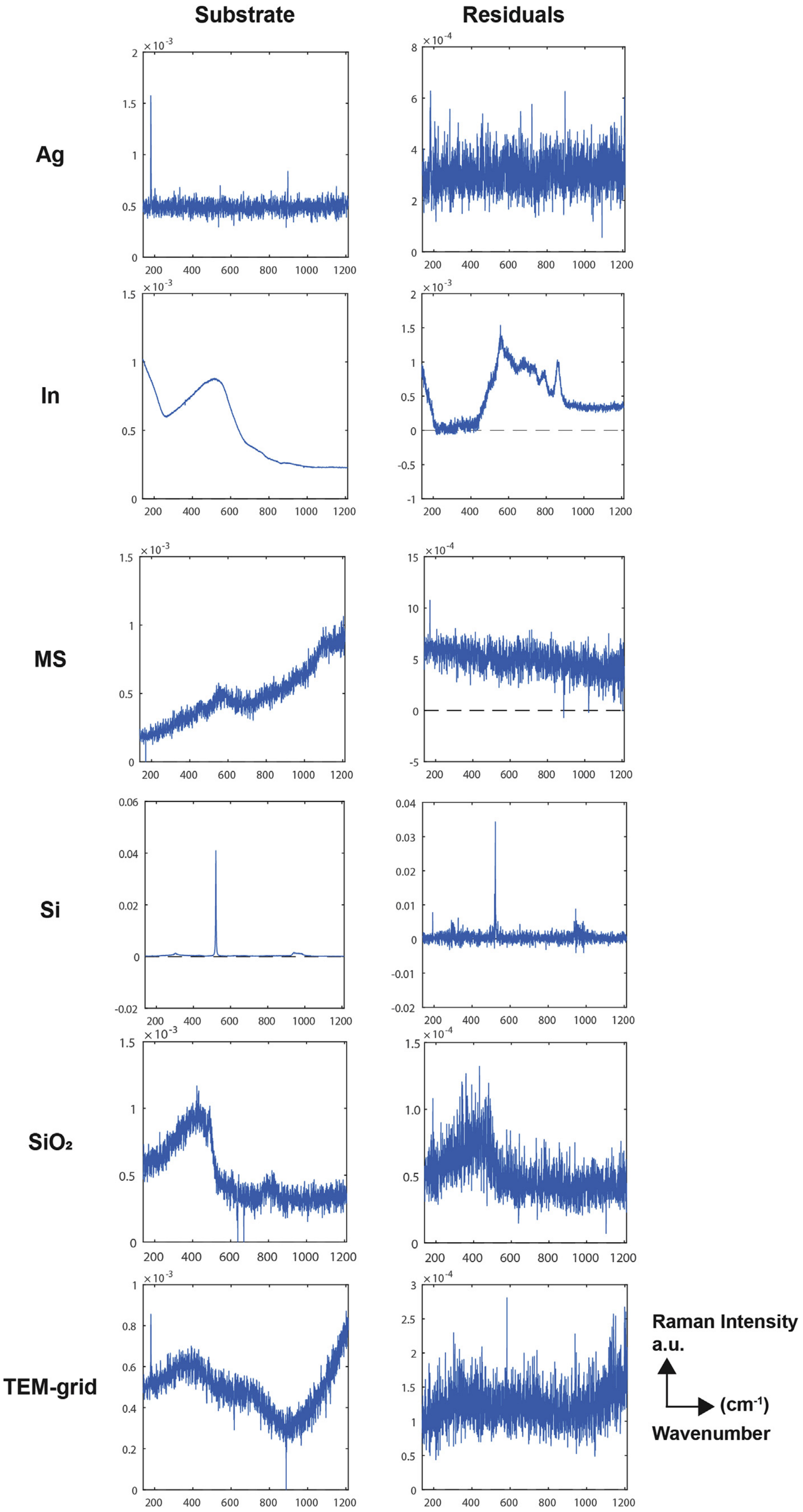


Fig. 3. Raman spectra of the substrate and residuals in S^T matrix from PURITY processing. (a.u.: arbitrary units).

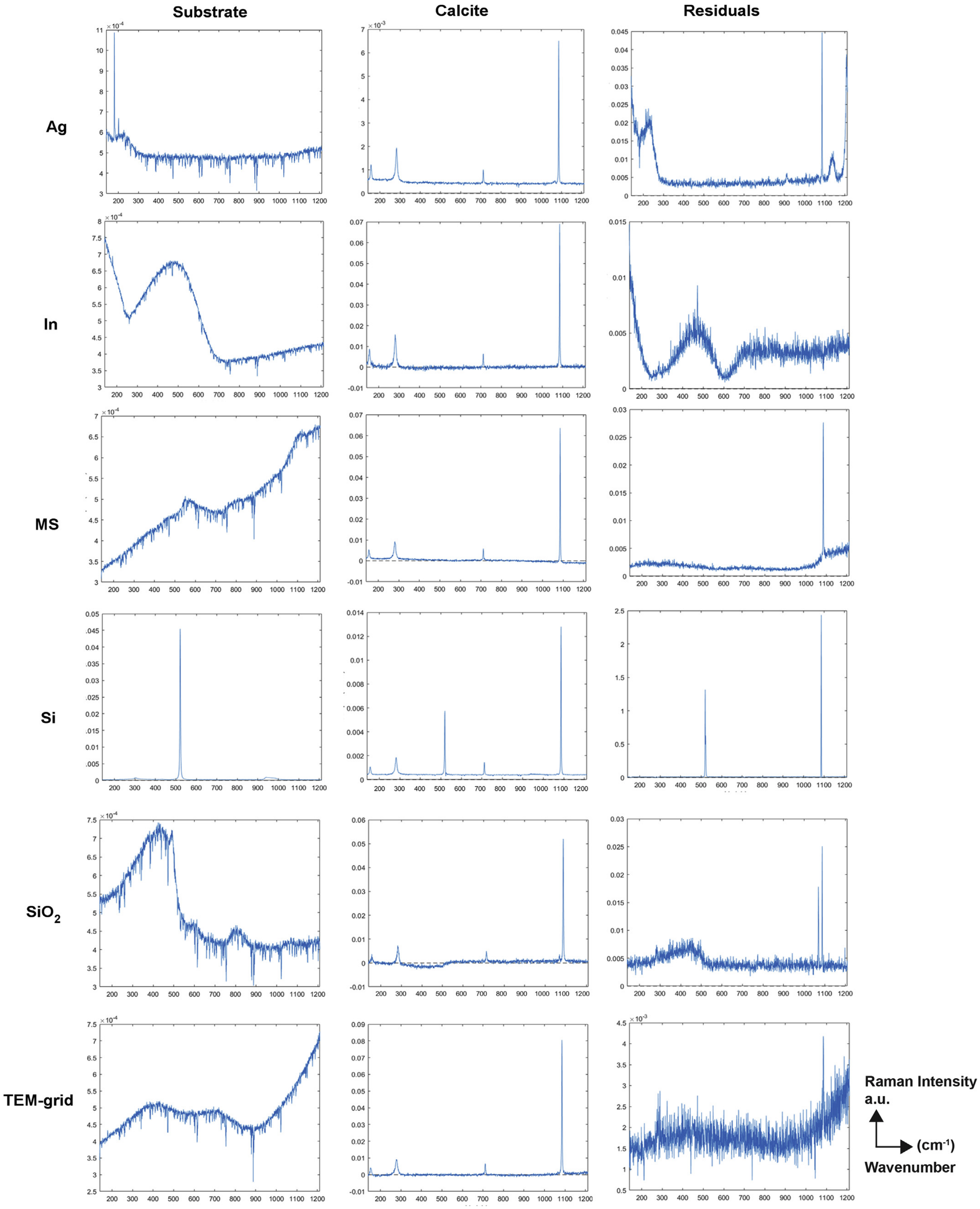


Fig. 4. Raman spectra of the substrate, calcite and residuals in S^T matrix from PURITY processing. (a.u.: arbitrary units).

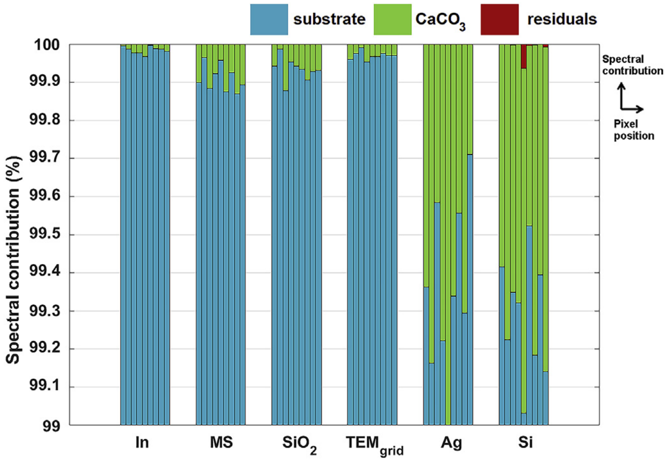


Fig. 5. Spectral contributions of $CaCO_3$, the substrate and the residual as a function of pixel position. Each column is divided into 9 bars, corresponding to the 9 pixels. Each bar represents the spectral contribution from the $CaCO_3$, substrate and the residual within a pixel.

satisfactory in the case of In, Ag, Si and TEM-grid, whereas for MS and SiO_2, significant deviation from the optical image can be observed. Deviation refers to a perturbation of the particle shape in the reconstructed map compared with the optical image, which is considered a benchmark. Such inconsistencies may be attributed to the intensity of the measured calcite signal, as well as the shape of the extracted Raman spectrum of the substrate itself. Due to the noticeable deviation of their reconstructed Raman maps, MS and SiO_2 are not the most appropriate substrates for molecular image reconstruction. However, the suitably resolved spectrum of $CaCO_3$ obtained from the PURITY procedure allows the correct identification of impacted particles on the substrates. In addition, the quality of the substrate used for the analysis is crucial. Godoi et al. [27] investigated different substrates while testing various energy densities of 514 nm and 785 nm lasers. Their work revealed oxidation of the Ag substrate, which should be evaluated before using Ag as a substrate for aerosol particles. Under our experimental conditions (laser wavelength 632.8 nm and maximal power density), Ag was not oxidized during acquisition. Si remained stable during the acquisitions, which is in agreement with Godoi et al. [27]. However, Si is not suitable as a substrate due to its strong band at 520.7 cm^{-1} which is still evident in the resolved $CaCO_3$ spectrum. MS and SiO_2 are the most popular substrates employed for RMS analysis of particles [14–17], since they are flat and chemically nonreactive, but our results demonstrated that other than the robust extraction of pure compounds, the reconstructed molecular images from the MCR procedure reveal the under-evaluation of the spectral contribution, which can induce image deformation. The TEM-grid appears to be a valuable substrate for analysis and Raman imaging of aerosol particles [13,28]. In this work, satisfactory image reconstruction based on spectral contribution values is demonstrated despite the poor spectral contribution obtained from the clean substrate.

3.3. Analysis of substrates with impacted $CaCO_3$, $PbSO_4$, $NaNO_3$ and Fe_2O_3 particles

The Ag, Si, TEM-grid, In, SiO_2 and MS substrates were impacted with $PbSO_4$, Fe_2O_3, $NaNO_3$ and $CaCO_3$ particles. Raman maps with a dimension of 10 × 10 pixels (1-μm step) and 15 × 15 pixels (for Si) were collected, but in order to compare the results with those for clean substrates and substrates impacted with calcite particles, smaller areas of the particle-impacted maps (3 × 3 pixels) were selected. The total size of each spot for each Raman map with impacted mixed particles was equal to 9 pixels. The mixed particles were an aggregate of the 4 compounds. Mixed particles used for analysis were selected to have a similar shape and size to the previously analysed $CaCO_3$ particles for the sake of comparison. Note that aggregate composition differs from one substrate to another since it is difficult to produce exactly the same particle type. During this procedure, 6 spectra were extracted, each of which corresponds to a substrate, 4 compounds and residuals (Fig. 6). Substrate spectral extraction was performed using optimal wavenumbers, as defined in section 3.1. The resolved spectra of individual particle species were obtained using their main Raman bands, i.e., 1085 cm^{-1}, 220 cm^{-1}, 974 cm^{-1}, and 1068 cm^{-1} for $CaCO_3$, Fe_2O_3, $PbSO_4$, and $NaNO_3$, respectively. In the case of the multi-composition aggregate, the extraction of each pure Raman spectrum was more complex, as is evident in the quality of the resolved spectra.

We observed that the choice of substrate mainly influences the noise level of resolved Raman spectra. This is apparent when comparing the Si and SiO_2 substrates. The Raman spectrum of Si has a flat baseline with low background; thus, the resolved compound spectra are characterized by a low signal-to-noise ratio. However, SiO_2 has a noisy, rough background, and hence the resolved spectra also have a noisy background with rough baseline. In the case of the heterogeneous aggregate, the amount of each compound within the 9 μm^2 area was very low. The PURITY procedure could not readily resolve the spectra of pure compounds, consequently, produce the corresponding contribution. Finally, as observed previously, for Si, a band at 520.7 cm^{-1} appeared on the resolved Raman spectrum of each species, which affected the real contribution of species after MCR extraction (Fig. 6). The spectral contributions of the signals were normalized and are presented in Fig. 7.

As observed in section 3.2, the spectral contributions of compounds in mixed particles depend on the substrate (Fig. 7). Indeed, the variation in the spectral contribution of each species observed for each pixel reflects the heterogeneous composition of the aggregate. The average contributions of the substrate, particle compounds, and residuals, as well as the standard deviation (SD) for the complete data set, are detailed in Table S2 (in SI files). The average contribution of the SiO_2 substrate was the highest (98.17%) compared with the other substrates, with relatively low SD (1.06%). As a consequence, the contribution of each compound was relatively low, with values varying between 0.07 and 1.17%, and a low contribution of the residual was observed (0.03%). For the In and MS substrates, their signals accounted for 94.39% and 93.76%, respectively. The spectral contribution of compounds for In was in the range 1.09–2.51%, while for MS the values were 0.07–4.14%. In the case of TEM-grid, a slightly greater compound contribution was observed, with spectral contributions from 0.23 to 7% and a contribution of 85.23% for the substrate. In the last subset, the Ag and Si substrates had significantly lower mean spectral contributions, with a larger contribution from the mixed particles. The spectral contributions for Si and Ag were 28.71% and 43.11%, respectively. The mixed particles' contributions ranged from 2.51 to 47.34% and 0.19–28.12% for Si and Ag, respectively. The values of the remaining residuals for all substrates were less than 0.7%. Differences in standard deviation (SD) were observed. The SDs for the In, MS and SiO_2 substrates are 2.2%, 2.3% and 1.06%, respectively. The SD of the TEM-grid substrate is 8.02%. The Si and Ag substrates differ the most from the other substrates: the SD for Si is 35.18%, and that for Ag is 40.18%. The high SD values for the Si and Ag substrates result from substantial differences in spectral contribution for individual pixels (Fig. 7). For $CaCO_3$ particles, significant variability in substrate contribution can be observed among all

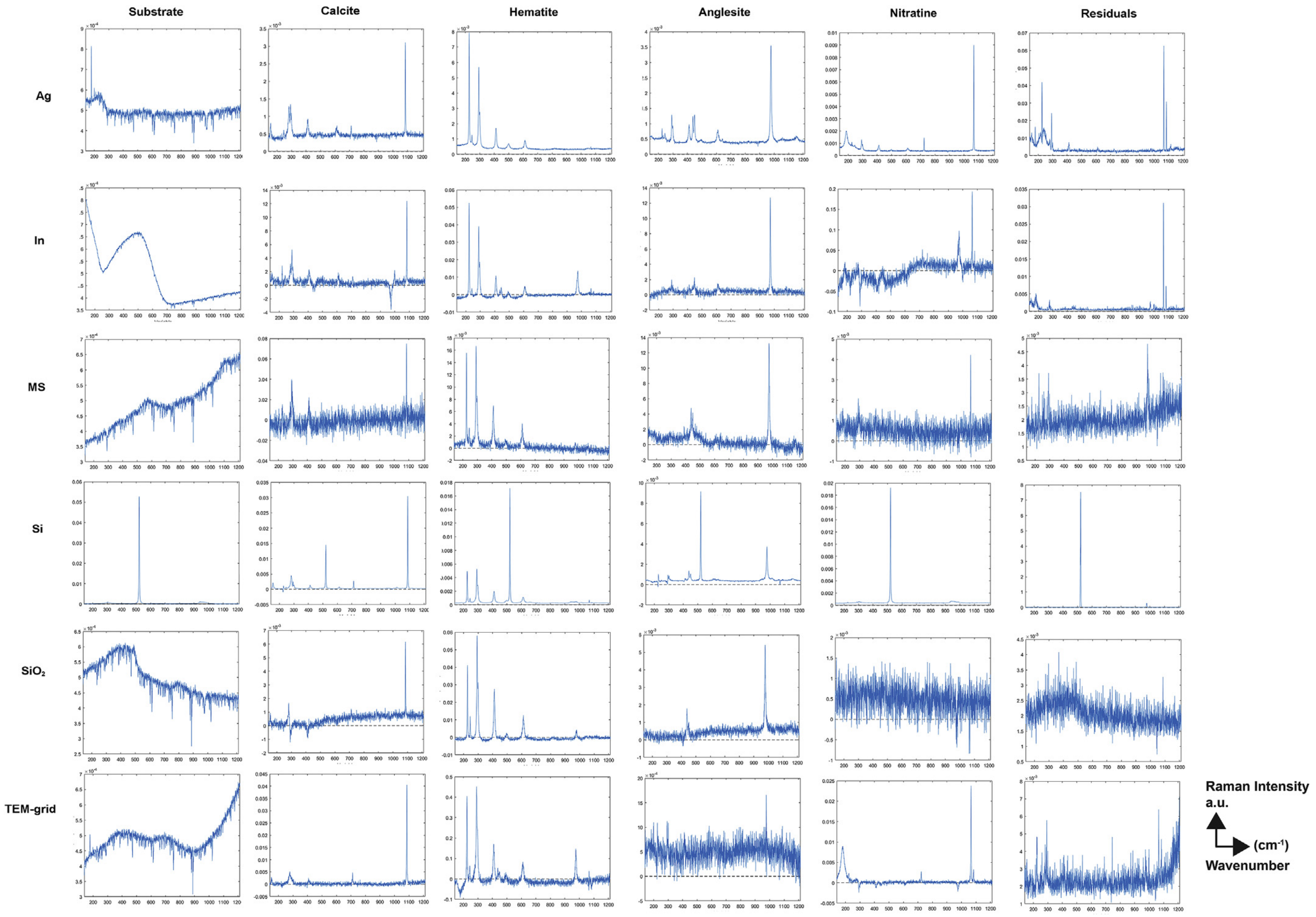


Fig. 6. Raman spectra of the substrate, calcite, hematite, nitratine, anglesite and residuals in S^T matrix from PURITY processing. (a.u. arbitrary units).

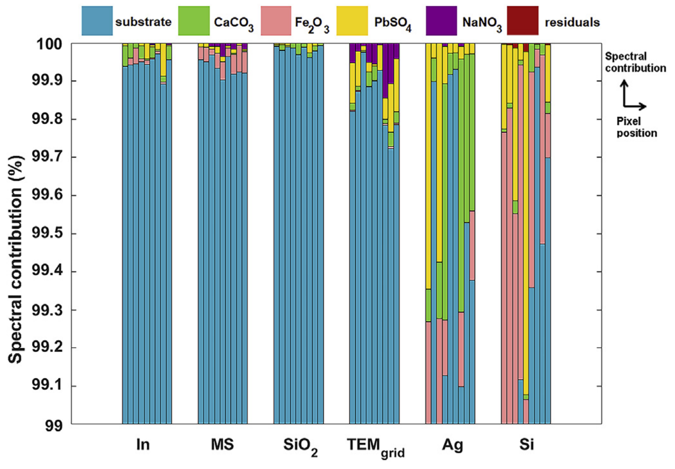


Fig. 7. Spectral contributions of the substrate, $PbSO_4$, Fe_2O_3, $NaNO_3$, $CaCO_3$ and the residual, as a function of pixel position. Each column is divided into 9 bars, corresponding to the 9 pixels. Each bar represents the spectral contribution of the $CaCO_3$, Fe_2O_3, $PbSO_4$, $NaNO_3$, substrate and residual within a pixel.

pixels for Ag and Si, while for the In, MS and SiO_2 substrates, values are more homogeneous. This confirms that the heterogeneity is mainly due to the substrate's contribution itself instead of the spectral contribution from the particle, i.e., the particle composition and thus Raman cross section of the species, or MCR extraction. Whatever the substrate, the contributions of the compounds are within the same order of magnitude compared to the standard deviation. Surprisingly, the contributions of the Ag and Si substrates are almost equal to their respective standard deviations, even with a flat baseline and a thin Raman band. The contributions of TEM-grid, SiO_2, MS and In are more uniform in comparison with Ag and Si.

Finally, raw Raman maps and reconstructed Raman maps were evaluated (Fig. S4 and Fig S5 in SI). It should be noted that the reconstructed shape of the particle in the Raman map corresponds to the optical image in the case of Ag, Si and TEM-grid. For In, MS and SiO_2 notable deviation of the Raman maps from the optical image can be observed, as was previously observed for $CaCO_3$-only particles.

4. Conclusions

Among six investigated substrates commonly used in Raman imaging, four (In, SiO_2, TEM-grid, and MS) show a significant signal contribution in Raman images of deposited particles. Conversely, the contribution from particles is low (inferior to 7%). The other two substrates (Si and Ag) appear more useful because the spectral contribution from the particles is higher than that from the substrate. However, the high standard deviations indicate the non-representativeness of the reconstructed image. The extracted spectra of aerosol compounds showed that the Si wafer was unsuitable for Raman imaging of the particles, due to the presence of a strong Raman band at 520.7 cm^{-1} in resolved spectra. Thus, an Ag layer would be preferred, but special care should be taken during application of an Ag substrate due to the oxidation process, which can significantly affect the homogeneity of the substrate surface. The extraction and identification of the aerosol compounds for TEM-grid, SiO_2, MS and In were satisfactory. However, the lower spectral contribution of the compounds may affect the effectiveness of the MCR (PURITY) method for extraction of particle compounds. This is particularly critical for MS and SiO_2, for which Raman images may be deformed after MCR reconstruction. We have demonstrated that this effect is worsened when the particle composition is complex. Finally, an In-coated substrate compromised reconstruction of the $CaCO_3$ spectrum in the region 900-1000 cm^{-1}. The counterpoint for substrate selection would be TEM-grid, which represents stable sample measurements with a homogeneous spectral contribution, satisfactory Raman map reconstruction and the potential for application in other single particle analysis techniques (e.g., SEM-EDX or TEM).

Conflicts of interest

The authors declare no conflict of interest.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2018.02.015.

References

[1] P.R. Buseck, M. Pósfai, Airborne minerals and related aerosol particles: effects on climate and the environment, Proc. Natl. Acad. Sci. Unit. States Am. 96 (1999) 3372–3379.

[2] D.K. Farmer, C.D. Cappa, S.M. Kreidenweis, Atmospheric processes and their controlling influence on cloud condensation nuclei activity, Chem. Rev. 115 (2015) 4199–4217.

[3] E. Montilla, S. Mogo, V. Cachorro, J. Lopez, A. de Frutos, Absorption, scattering and single scattering albedo of aerosols obtained from in situ measurements in the subarctic coastal region of Norway, Atmos. Chem. Phys. Discuss. 2011 (2011) 2161–2182.

[4] V. Bollati, B. Marinelli, P. Apostoli, M. Bonzini, F. Nordio, M. Hoxha, V. Pegoraro, V. Motta, L. Tarantini, L. Cantone, J. Schwartz, P.A. Bertazzi, A. Baccarelli, Exposure to metal-rich particulate matter modifies the expression of candidate microRNAs in peripheral blood leukocytes, Environ. Health Perspect. 118 (2010) 763–768.

[5] F.J. Kelly, J.C. Fussell, Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter, Atmos. Environ. 60 (2012) 504–526.

[6] F. Liang, G. Zhang, M. Tan, C. Yan, X. Li, Y. Li, Y. Li, Y. Zhang, Z. Shan, Lead in Children's blood is mainly caused by coal-fired ash after phasing out of leaded gasoline in Shanghai, Environ. Sci. Technol. 44 (2010) 4760–4765.

[7] P.E. Tolbert, Invited commentary: heterogeneity of particulate matter health risks, Am. J. Epidemiol. 166 (2007) 889–891.

[8] O.S. Ryder, A.P. Ault, J.F. Cahill, T.L. Guasco, T.P. Riedel, L.A. Cuadra-Rodriguez, C.J. Gaston, E. Fitzgerald, C. Lee, K.A. Prather, T.H. Bertram, On the role of particle inorganic mixing state in the reactive uptake of N2O5 to ambient aerosol particles, Environ. Sci. Technol. 48 (2014) 1618–1627.

[9] R. Arimoto, Y.J. Kim, Y.P. Kim, P.K. Quinn, T.S. Bates, T.L. Anderson, S. Gong, I. Uno, M. Chin, B.J. Huebert, A.D. Clarke, Y. Shinozuka, R.J. Weber, J.R. Anderson, S.A. Guazzotti, R.C. Sullivan, D.A. Sodeman, K.A. Prather, I.N. Sokolik, Characterization of asian dust during ACE-asia, Global Planet. Change 52 (2006) 23–56.

[10] A.P. Ault, J.L. Axson, Atmospheric aerosol chemistry: Spectroscopic and microscopic advances, Anal. Chem. 89 (2017) 430–452.

[11] A.P. Ault, T.L. Guasco, J. Baltrusaitis, O.S. Ryder, J.V. Trueblood, D.B. Collins, M.J. Ruppel, L.A. Cuadra-Rodriguez, K.A. Prather, V.H. Grassian, Heterogeneous reactivity of nitric acid with nascent Sea Spray aerosol: large differences observed between and within individual particles, J. Phys. Chem. Lett. 5 (2014) 2493–2500.

[12] S. Sobanska, G. Falgayrac, J. Rimetz-Planchon, E. Perdrix, C. Brémard, J. Barbillat, Resolving the internal structure of individual atmospheric aerosol particle by the combination of Atomic Force Microscopy, ESEM–EDX, Raman and ToF–SIMS imaging, Microchem. J. 114 (2014) 89–98.

[13] S. Sobanska, H. Hwang, M. Choël, H.-J. Jung, H.-J. Eom, H. Kim, J. Barbillat, C.-U. Ro, Investigation of the chemical mixing state of individual asian dust particles by the combined use of electron probe X-ray microanalysis and Raman microspectrometry, Anal. Chem. 84 (2012) 3145–3154.

[14] Y. Batonneau, S. Sobanska, J. Laureyns, C. Bremard, Confocal microprobe Raman imaging of urban tropospheric aerosol particles, Environ. Sci. Technol. 40 (2006) 1300.

[15] G. Falgayrac, S. Sobanska, C. Bremard, Particle-particle chemistry between micrometer-sized $PbSO_4$ and $CaCO_3$ particles in turbulent flow initiated by liquid water, J. Phys. Chem. 116 (2012) 7386–7396.

[16] G. Falgayrac, S. Sobanska, C. Bremard, Heterogeneous microchemistry between $CdSO_4$ and $CaCO_3$ particles under humidity and liquid water, J. Hazard Mater. 248–249 (2013) 415–423.

[17] G. Falgayrac, S. Sobanska, C. Bremard, Raman diagnostic of the reactivity between $ZnSO_4$ and $CaCO_3$ particles in humid air relevant to heterogeneous zinc chemistry in atmosphere, Atmos. Environ. 85 (2014) 83–91.

[18] J. Ofner, K.A. Kamilli, E. Eitenberger, G. Friedbacher, B. Lendl, A. Held, H. Lohninger, Chemometric analysis of multisensor hyperspectral images of precipitated atmospheric particulate matter, Anal. Chem. 87 (2015) 9413–9420.

[19] M. Offroy, M. Moreau, S. Sobanska, P. Milanfar, L. Duponchel, Pushing back the limits of Raman imaging by coupling super-resolution and chemometrics for aerosols characterization, Sci. Rep. 5 (2015) 12303.

[20] M. Choël, K. Deboudt, J. Osán, P. Flament, R. Van Grieken, Quantitative determination of low-z elements in single atmospheric particles on boron substrates by automated scanning electron Microscopy–Energy-dispersive X-ray spectrometry, Anal. Chem. 77 (2005) 5686–5692.

[21] C. Font Palma, G.J. Evans, R.N.S. Sodhi, Imaging of aerosols using time of flight secondary ion mass spectrometry, Appl. Surf. Sci. 253 (2007) 5951–5956.

[22] S. Maskey, M. Choël, S. Kang, H. Hwang, H. Kim, C.-U. Ro, The influence of collecting substrates on the single-particle characterization of real atmospheric aerosols, Anal. Chim. Acta 658 (2010) 120–127.

[23] I. Szalóki, J. Osán, A. Worobiec, J. de Hoog, R. Van Grieken, Optimization of experimental conditions of thin-window EPMA for light-element analysis of individual environmental particles, X Ray Spectrom. 30 (2001) 143–155.

[24] D.P. Veghte, J.E. Moore, L. Jensen, M.A. Freedman, Influence of shape on the optical properties of hematite aerosol, J. Geophys. Res.: Atmosphere 120 (2015) 7025–7039.

[25] W. Windig, B. Antalek, J.L. Lippert, Y. Batonneau, C. Brémard, Combined use of conventional and second-derivative data in the SIMPLISMA Self-modeling mixture analysis approach, Anal. Chem. 74 (2002) 1371–1379.

[26] G. Falgayrac, S. Sobanska, J. Laureyns, C. Bremard, Heterogeneous chemistry between PbSO4 and calcite microparticles using Raman microimaging, Spectrochim. Acta Mol. Biomol. Spectrosc. 64 (2006) 1095–1101.

[27] R.H.M. Godoi, S. Potgieter-Vermaak, J. De Hoog, R. Kaegi, R. Grieken, Substrate selection for optimum qualitative and quantitative single atmospheric particles analysis using nano-manipulation, sequential thin-window electron probe X-ray microanalysis and micro-Raman spectrometry, Spectrochim. Acta B 61 (2006) 375–388.

[28] A.K. Lee, T.Y. Ling, C.K. Chan, Understanding hygroscopic growth and phase transformation of aerosols using single particle Raman spectroscopy in an electrodynamic balance, Faraday Discuss 137 (2008) 245–263 discussion 297–318.

Influence of Collecting Substrates on the Raman Imaging of Aerosol Particles – Supporting Information

The following supporting information contains 5 figures and 2 tables.

Fig. S 1 represents the optimal wavenumber for the pure variable selection as function of the substrate. Fig. S 2 represents optical images and raw Raman maps of substrates with impacted $CaCO_3$ particles based on the area of the peak of interest. Fig. S 3 represents optical images and reconstructed Raman maps of substrates with impacted $CaCO_3$ particles based on spectral contribution C matrix after PURITY approach. Fig. S 4 represents optical images and raw Raman maps of substrates with impacted particles based on the area of the peak of interest. Fig. S 5 represents optical images and reconstructed Raman maps of substrates with impacted mixed particles based on spectral contribution matrix after Purity approach. Table S1 contains the average contribution (%) of a substrate, calcite, residuals, and RRSSQ after PURITY extraction. Table S2 contains average contribution (%) of a substrate, particles, residue and RRSSQ after PURITY extraction.

© 2017 Tous droits réservés.

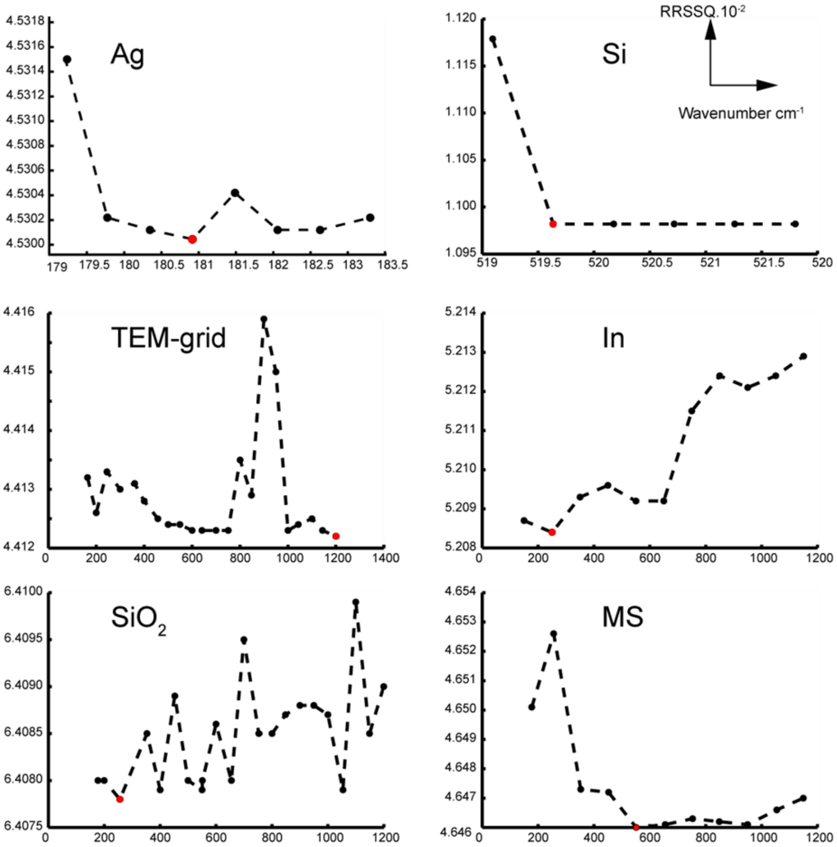


Fig. S 1. Relative Root Sum of Square (RRSSQ) as function of the wavenumber of extraction for the sixth clean substrates. The red dot corresponds to the lowest RRSSQ. Abbreviation: MS = Microscope Slide

© 2017 Tous droits réservés.

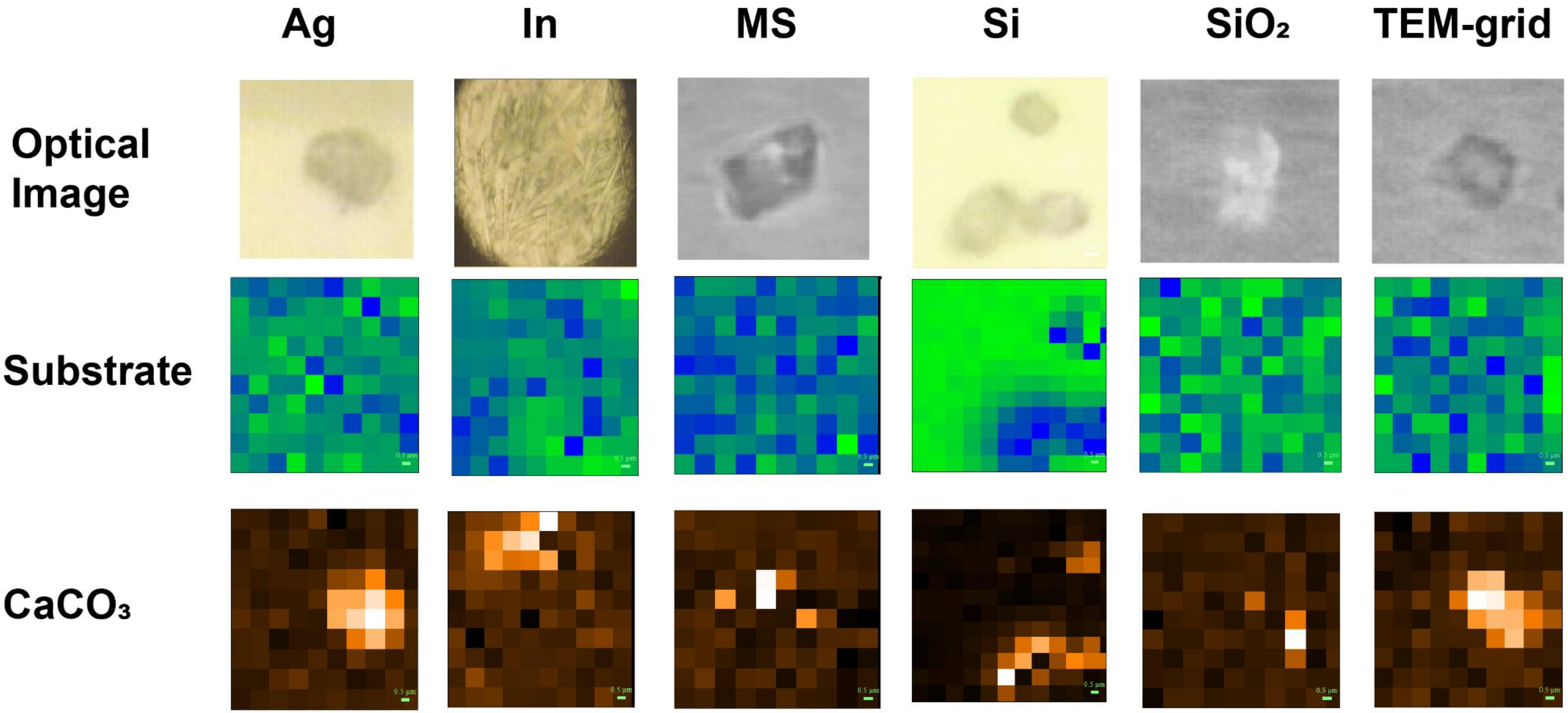


Fig. S 2: Optical images and raw Raman maps of substrates with impacted $CaCO_3$ particles based on the area of the peak of interest $CaCO_3$ 1085 cm^{-1}. The peaks of interest chosen for the substrates are the same which were selected in the Fig. S 1. (1 pixel=1 µm). No pre-processing is applied.

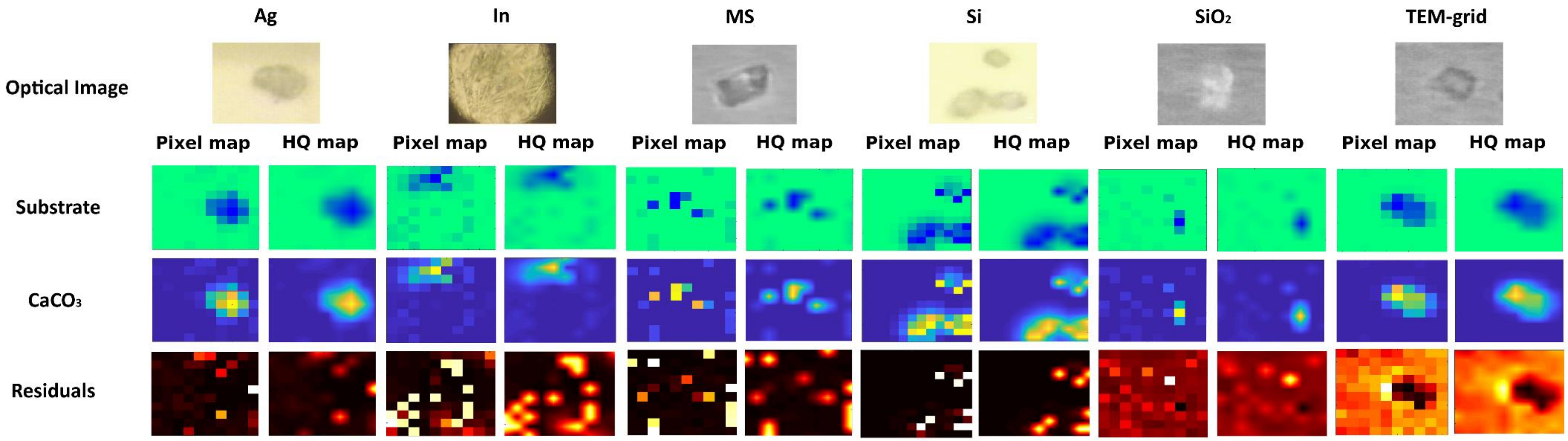


Fig. S 3. Optical images and reconstructed Raman maps of substrates with impacted $CaCO_3$ particles based on spectral contribution C matrix after PURITY approach (1 pixel=1 µm). HQ map is the Pixel map after smoothing procedure.

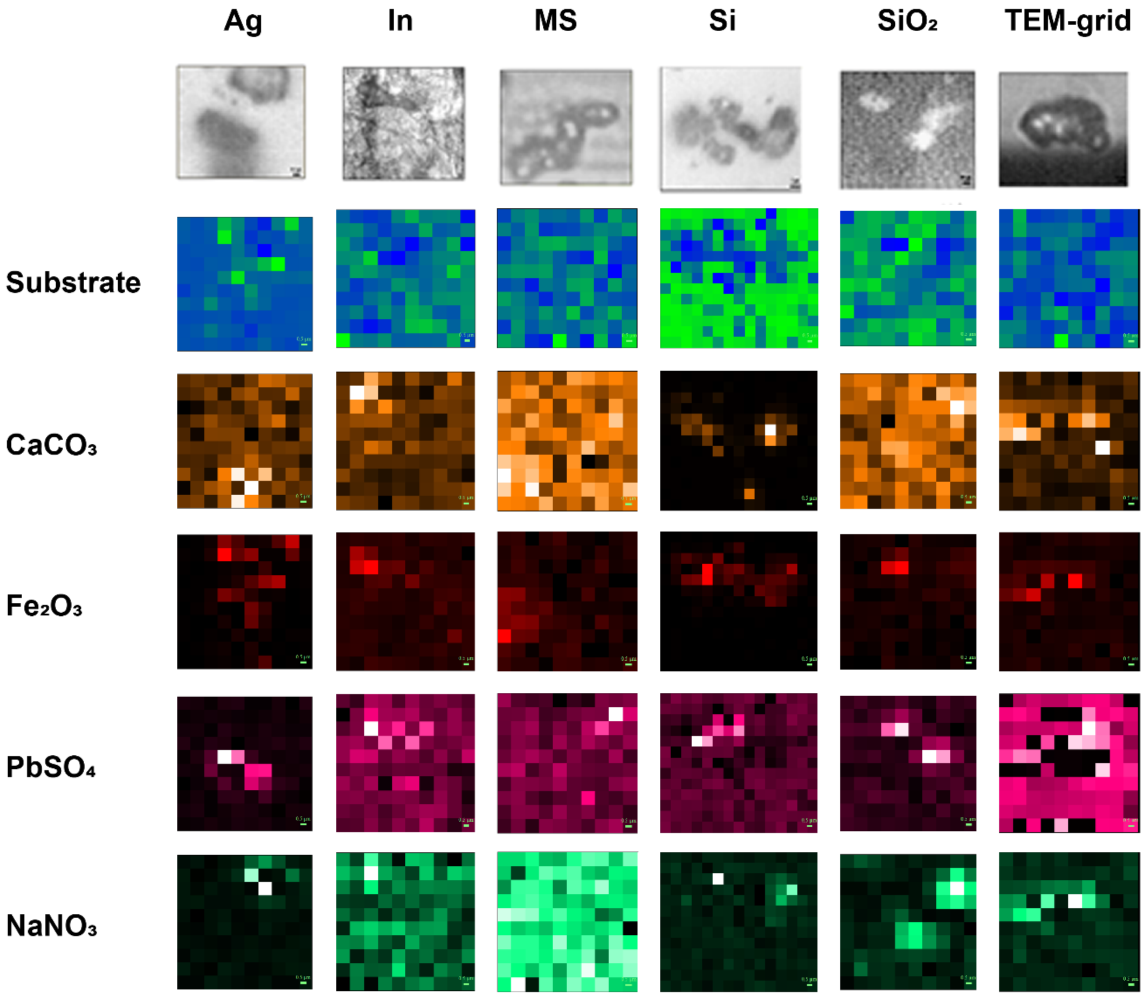


Fig. S 4. Optical images and raw Raman maps of substrates with impacted particles based on the area of the peak of interest $CaCO_3$ 1085 cm^{-1}, Fe_2O_3 220 cm^{-1}, $PbSO_4$ 974 cm^{-1}, $NaNO_3$ 1068 cm^{-1}. The peaks of interest chosen for the substrates are the same which were found in the Fig. S 1. (1 pixel=1 µm). No pre-processing is applied.

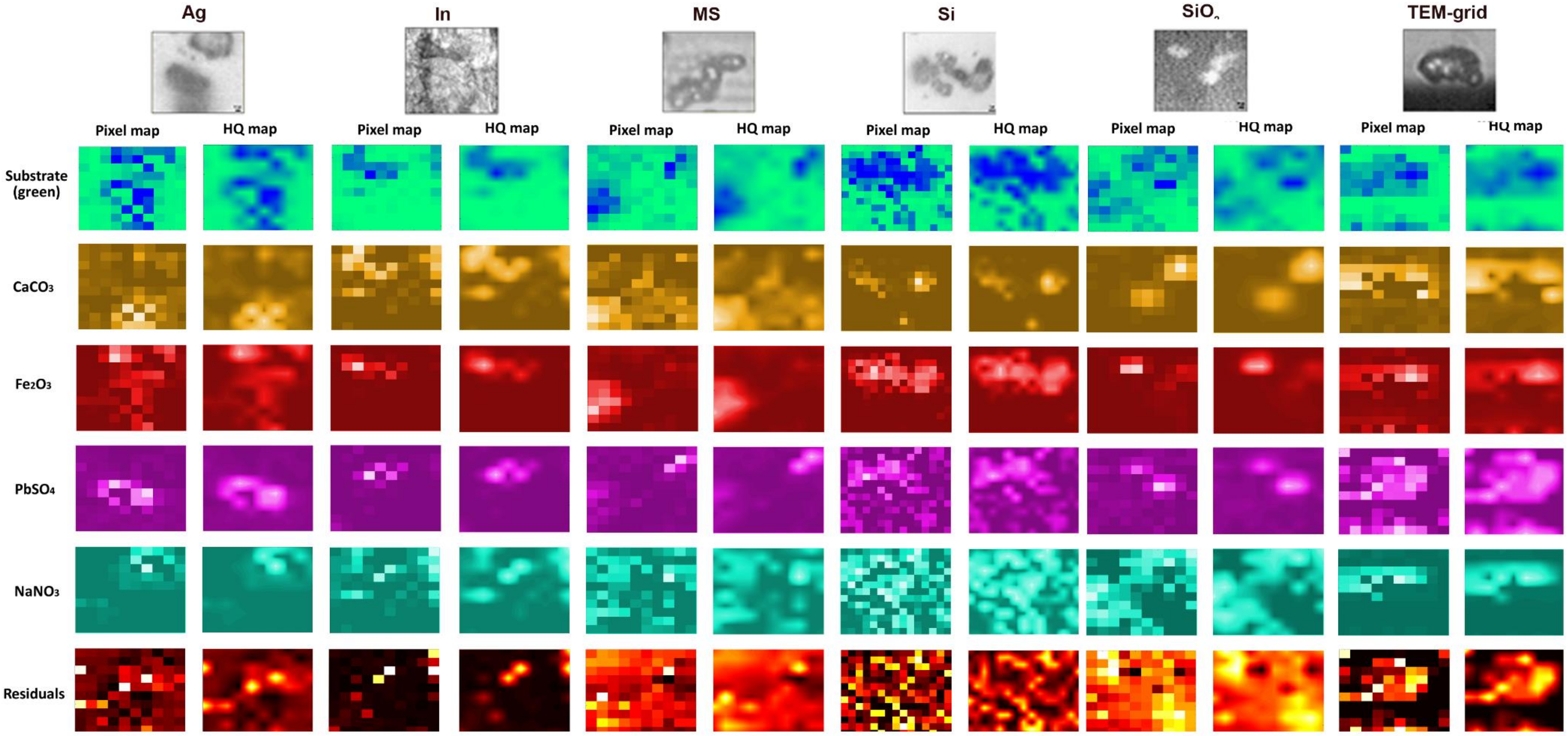


Fig. S 5. Optical images and reconstructed Raman maps of substrates with impacted mixed particles based on spectral contribution matrix after Purity approach (1 pixel=1 µm). HQ map is the Pixel map after smoothing procedure.

Table S1. Average contribution (%) of a substrate, calcite, residuals, and RRSSQ after PURITY extraction

Name	*Substrate ±SD*	*$CaCO_3$ ±SD*	*Residuals ±SD*	*RRSSQ*
In	98.38 ± 0.94	1.57 ± 0.99	0.05 ± 0.09	0.03
MS	90.99 ± 3.48	8.97 ± 3.49	0.04 ± 0.02	0.04
SiO_2	93.36 ± 2.99	6.56 ± 3.01	0.08 ± 0.02	0.05
TEM-grid	96.99 ± 1.06	2.99 ± 1.06	0.03 ± 0.01	0.04
Ag	35.88 ± 22.47	64.07 ± 22.47	0.05 ± 0.01	0.05
Si	28.67 ± 15.41	70.41 ± 14.14	0.92 ± 2.04	0.03

Abbreviation: SD = standard deviation

© 2017 Tous droits réservés.

Table S2. Average contribution (%) of a substrate, particles, residue and RRSSQ after PURITY extraction

Name	*Sub.±SD*	*Fe_2O_3 ±SD*	*$CaCO_3$ ±SD*	*$PbSO_4$ ±SD*	*$NaNO_3$ ±SD*	*Res. ±SD*	*RRSSQ*
In	94.39 ± 2.2	1.09 ± 1.29	2.51 ± 1.7	1.85 ± 2.87	0.14 ± 0.15	0.04 ± 0.06	0.02
MS	93.76 ± 2.3	4.14 ± 1.69	0.07 ± 0.08	0.95 ± 0.51	1.03 ± 1.07	0.05 ± 0.01	0.05
SiO_2	98.17 ± 1.06	0.07 ± 0.05	1.17 ± 0.76	0.56 ± 0.8	ND	0.03 ± 0.01	0.03
TEM-grid	85.23 ± 8.02	0.23 ± 0.17	2.1 ± 1.52	7.00 ± 4.94	5.35 ± 4.58	0.04 ± 0.02	0.05
Ag	43.11 ± 40.18	11.86 ± 11.96	28.12 ± 25.41	16.66 ± 25.28	0.19 ± 0.38	0.06 ± 0.02	0.06
Si	28.71 ± 35.18	47.34 ± 32.16	2.51 ± 2.07	20.8 ± 29.12	ND	0.65 ± 0.72	0.03

Abbreviations: SD = standard deviation. ND=Not Detected, Sub = Substrate, Res. = Residuals

© 2017 Tous droits réservés.

3.2. The analytical algorithm for designation of chemical mixing and quantitative results from RMS of aerosol particles.

In this part, we present the effective analytical algorithm for processing of Raman spectra collected by single particle analysis in designation of chemical mixing and quantitative composition of aerosol particles. The focus of the application is the analysis of metal-rich, mine dust particles collected in the Oruro (Bolivia) mining environment. The particle collection was performed using a personal cascade impactor that allowed sampling of three particle size ranges (in µm), i.e. $PM_{10-2.5}$ (sample A), $PM_{2.5-1}$ (sample B), $PM_{1-0.5}$ (sample C) and additionally size fraction $PM_{0.5}$ (<0.5 µm) (sample D). In addition, particles from personal filters from the miner's protection mask were also collected (sample FM). The optimization and verification of the algorithm were performed based on the analysis of sample B. The large number of spectra in this fraction dictated this approach. The results of this procedure appeared in the publication which is an integral part of this paragraph. We present the potential of single particle analysis by Raman microspectroscopy for an efficient description of chemical mixing of mine dust particles. The application of the algorithm for experimental data, proves the possibility of extending the limitations in trace component detection and quantitative analysis, as well as provides a new way of sample description.

3.2.1. Combining Raman microspectrometry and chemometrics for determining quantitative molecular composition and mixing state of atmospheric aerosol particles

The presented results give an opportunity to indicate the quantitative composition of particles by comparing their composition (with the level of mixing) with respect to the whole particle population. The methodology was published as a publication in Microchemical Journal and presented in the following subsection.

© 2017 Tous droits réservés.

The proposed methodology consists of analysing of particles through automated RMS measurements. Structuration of the collected data from the single particle - RMS measurements is made by chemometrics methods. Through the application of the MCR approach on the Raman spectra data set, the designation of pure compounds was possible. The methodology is constructed for the purpose to distinguish particles composed of only one species (pure molecular compounds) from those which molecular composition is comprised of more than one species (mixed particles) and to define particle types according to their chemical similarity. Finally, quantification of each group of particles according to their composition and chemical mixing in order to describe the chemical composition and heterogeneity was presented. For that purpose, multivariate curve resolution and unsupervised multivariate analysis techniques were used in combination. This algorithm is an attempt to combine Raman spectroscopy with chemometric methods to obtain new information on molecular composition of particles as well as to obtain information on the level of mixing of aerosol particles.

© 2017 Tous droits réservés.

Microchemical Journal 137 (2018) 119–130

Contents lists available at ScienceDirect

Microchemical Journal

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

Combining Raman microspectrometry and chemometrics for determining quantitative molecular composition and mixing state of atmospheric aerosol particles

Damian Siepka [a,b,d], Gaëlle Uzu [c], Elżbieta A. Stefaniak [a], Sophie Sobanska [d,*]

[a] Laboratory of Composite and Biomimetic Materials, Center for Interdisciplinary Research, The John Paul II Catholic University of Lublin, Konstantynów 1J, 20-708 Lublin, Poland
[b] Laboratoire de Spectrochimie Infrarouge et Raman, UMR CNRS 8516, , Université de Lille 1, Bat. C5, 59655 Villeneuve d'Ascq Cedex, France
[c] Université Grenoble Alpes, CeNRS, IRD, IGE, F-38000 Grenoble, France
[d] Université de Bordeaux, Institut des Sciences Moléculaires, CNRS UMR 5255, 351 cours de la Libération, 33405 Talence cedex, France

ARTICLE INFO

Article history:
Received 17 July 2017
Received in revised form 9 October 2017
Accepted 9 October 2017
Available online 18 October 2017

Keywords:
Raman microspectroscopy
Aerosol particles
Multivariate Curve Resolution
Multivariate data analysis
Quantitative Raman analysis

ABSTRACT

Determining quantitative molecular composition of atmospheric particles is required for assessing their environmental and health impacts. The presented algorithm was designed to analyse numerous Raman spectra of metal-rich atmospheric particles. Multivariate curve resolution-alternating least squares procedure (MCR-ALS) has been applied to resolve complex data from Raman microanalysis by means of a computer-assisted analytical procedure called Single Particle Analysis (SPA). The SPA – contrary to Raman mapping – provides data in which each single particle is assigned to a single spectrum, in the group with a statistically significant size. During the procedure, the relative contributions of individual compounds in the recorded Raman spectra have been specified. Grouping and relationship determination of the collected data have been performed by hierarchical cluster analysis (HCA) and principal component analysis (PCA). A new methodology is proposed to quantitatively determine the molecular composition and chemical mixing of single airborne particles based on the data from the automated Raman microspectrometry measurements.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Single Particle Analysis (SPA) applied to atmospheric aerosol composition is an approach to analyse particulate matter by measuring individual particles [1–10]. Determining the composition of particles collected in ambient air using SPA, as opposed to bulk analysis, presents essential advantages that have been demonstrated in numerous works [11]. The analytical techniques appropriate for SPA must show a lateral resolution adapted to the particulate matter size. They must also show sensitivity suitable to detect both major and minor compounds within the particles. The SPA approach requires large quantities of data in order to obtain statistical significance for the description of collected particles. This assumption bears two major issues.First off, an automation of measurements is needed to achieve a sample size above the level of statistical significance. The second issue concers development of robust data treatment since the large data matrix entails reducing the number of objects into the groups. In the case of atmospheric particles, it leads to detailed study of a few representatives only. It also helps provide reliable spatial and temporal chemical evolution of particle composition.

Among SPA techniques, Scanning Electron Microscopy (SEM) equipped with an energy dispersive X-ray detection (EDX) and Raman microspectrometry (RMS) have been recognized as powerful methods for studying particulate matter of different origin (airborne particles, sediments, soils, etc.) [3,6,7,9,12–21]. Both techniques belong to micro-analytical methods where the source of information is a result of a beam interaction (i.e. electrons or photons) with micro-size objects, producing elemental and molecular spectra, respectively. Automation of SEM/EDX applied to particulate matter was implemented more than twenty years ago and has been successfully developed ever since. However,there are some limitations related to the elemental quantification. Calculation of particle elemental composition requires a fit-for-purpose quantification methodology based of X-ray spectra. Quantification procedure based on Monte-Carlo simulations was developed for suitable measurements of low-Z elements [3–9], while the gun-shot residue (GSR) software enables detection of particles with high-Z elements [22]. Based on the elemental composition and morphology, the particles can be classified into different groups related to their chemical characteristics. Nevertheless, a significant number of publications reporting a successful application of computer-controlled SEM/EDX measurements of a large number of particles, followed by quantification and appropriate data classification, prove the need for fast and reliable tools in the field of SPA [3–9] [23–31].

[*] Corresponding author.
E-mail address: sophie.sobanska@u-bordeaux.fr (S. Sobanska).

https://doi.org/10.1016/j.microc.2017.10.005
0026-265X/© 2017 Elsevier B.V. All rights reserved.

Confocal Raman microspectrometry (RMS) is another SPA technique providing molecular composition (speciation) of atmospheric particles, with a capability for automation of measurements. A Raman spectrometer coupled with an optical microscope and an automated XY stage, allows for computer-controlled measurements. The manual spot mode has been used for decades for characterizing the molecular composition of individual particles collected in ambient air [33,34]. In an imaging mode, RMS is able to describe chemical heterogeneity of micro-sized aerosol particles [20–38]. Considering a complex molecular composition resulting from heterogeneity of airborne particles, a Raman spectrum of one particle often reflects a mixture of several compounds. The multivariate curve resolution approach, such as SIMPLISMA, has substantially improved chemical description of particles by resolving the contribution of pure variables in the mixed spectra [13–21] [35–37]. Confocal Raman microspectroscopy has been considered as less appropriate compared to SEM-EDX because of the three main issues: (i) identification of molecular species can be assessed using RMS but quantitative data is not accessible; this is because the Raman band intensities are not indicative for the proportion of species in solid and heterogeneous particles (ii) as mentioned previously, the analysis of particles results in a Raman spectrum of mixed species that requires automated procedure for unmixing in order to analyse a large number of particles, (iii) clustering of particles based on their Raman spectra requires automation of measurements and preprocessing of the Raman spectra. As a consequence, an application of RMS has only recently been extended to a large number of particles. Reisner et al. [39] described an integrated software for processing, analyzing, and classifying of multiple Raman spectra. The developed system was equipped with multiple preprocessing methods, including: a median filter to reduce noise and remove spikes due to cosmic rays, a wavelet filter for further noise reduction, an automated background fluorescence subtraction [40], normalization of spectra and subtraction of artifacts. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were implemented as the two main multivariate analysis methods. Additionally, some classification methods, such as an artificial neural network classifier and a support vector machine classifier were implemented. Another significant feature was made by implementation of a relational model database for storing of any kind of information about the spectra, such as a date of data collection or a type of the specimen. Reisner et al. [39] made a compelling step to integrate tools for multiple Raman spectra processing, management and classification into a freely available, cross-platform system. Despite the versatility of the described system, there is a lack of spectral unmixing as well as clustering methods which are crucial in handling of the data from the environmental sample analysis [39]. Finally, Reisner et al. [39] evaluated the described system using the data from biological specimens. The results demonstrated the vast potential of the system in processing and classification of the spectra, based on well specified parameters. Ofner et al. [41] presented multisensor hyperspectral images of atmospheric particulate matter coupled with chemometrics analysis of an extensive chemical description of the collected particles. The authors provided a comprehensive image-based analysis of aerosol particles with a sectional description of samples. The combination of RMS and SEM-EDX used in multisensor hyperspectral data analysis made a way for detailed and well-grounded assignment of chemical species in the sample and the relationships among them. Notwithstanding, the described methodology is dedicated to the data from hyperspectral imaging, which significantly reduces the number of particles during a single run, compared to SPA. The structuration of the data was based on the pixel relationship taken from hyperspectral images which may not be directly translated into the description of the particle population. Additionally, operations on hyperspectral images are less efficient due to a large number of spectra collected during a single analysis, from which only few correspond to atmospheric particles. Finally, the mixing chemical state of particles is not considered. Jentzsch et al. [42] studied classification of mean Raman spectra of model particles composed of the most representative salts expected to be produced in the atmosphere. The chemometrics data analysis was used to distinguish the potential of the Raman spectra classification from a single particle analysis. Jentzsch et al. [42] did not include any spectral unmixing algorithm, which might be a limitation for analysis of ambient aerosol particles [42]. Furthermore, Craig et al. [43] applied computer controlled Raman microscpectroscopy (CC-Raman) for a single particleanalysis of a large number of both laboratory-generated and real-life particles. The authors conducted comprehensive characterization and clustering of aerosol particles based on the distinct features of their Raman spectra. Moreover, Craig et al. [43] juxtaposed the results from Raman microspectroscopy with SEM/EDX and a condensation particle counter (CPC), which equalled in similar cluster formulation. Nonetheless, due to the lack of spectral unmixing algorithm, the specification of particular compounds and their chemical mixing in the ambient aerosol particles was limited.

Regarding the previous work, an application of RMS for a chemical description of particles collected in ambient air is not complete yet, since the quantitative composition of the samples, i.e. the relative abundance of particle types defined from their molecular composition, has not been provided. The multivariate curve resolution methods (MCR) applied to a Raman data set offers a considerable advantage as mentioned above. The mathematical principle of MCR is based on the presence of pure variables. In terms of spectroscopy, a pure variable (e.g. a wavenumber of a Raman spectrum) is a variable which has an intensity contribution from only one component of a mixture. Once the pure variables of every component are known, their content in the mixture spectra is calculated. Reliability of the contribution profiles may be used as calibration process as previously demonstrated [15]. Alternatively, resolved contributions may be considered to be indicative for a presence or an absence of chemical species within particles and used for determining a relative abundance of particle types. It should be noted that the application of the multivariate analysis – mostly clustering – after the matrix decomposition procedure, has already been described by other authors [44,45]. For instance, Liu et al. 2003 [44] proposed to initially perform principal component analysis in order to obtain a matrix of scores for further clustering. Furthermore, Tamayo et al. 2007 [45], suggested to apply the matrix decomposition by the non-negative matrix factorization on high-dimensional data for clustering of a basis matrix only.

The present study aims to provide a quantitative chemical description of collected airborne particles. The proposed methodology consists of analyzing a large number of particles through automated measurements. Grouping of the collected data is made by chemometrics methods through the application of the MCR approach on the Raman spectra data matrix. Firstly, the methodology requires first to distinguish particles composed of only one species (pure molecular compounds) from those with molecular composition comprised of more than one species (mixed particles). The second step is to define particle types according to their chemical similarity. The final part involves quantification of each group of the particles according to their composition and chemical mixing in order to describe the chemical composition and heterogeneity of the collected particles. For that purpose a multivariate curve resolution and an unsupervised multivariate analysis are used in combination for description of real particle samples collected in the mining environment.

2. Experimental

2.1. Sample description

The particles were sampled in the galleries of San Jose Mine (150 m, underground),located in the Oruro area, Bolivia (17°46′0″S – 67°28′60″ W – 3674 MAMSL). The particle collection was performed using a personal cascade impactor (SIOUTAS, 3 l.min-1), allowing simultaneous walking along underground passages and sampling of four size fractions of particles, i.e. PM_{10}, $PM_{2.5}$, PM_1 and $PM_{0.5}$ corresponding to 10–2.5 μm, 2.5-1 μm, 1–0.5 μm and <0.5 μm aerodynamic diameter, respectively. The particles were collected on TEM grids mounted in the impaction plates [46]. One stage of the impactor corresponding to the $PM_{2.5}$ fraction was selected for this methodological study. The elemental composition of this particulate matter fraction, measured by ICP-MS, is given in the Appendix, Table A1. $PM_{2.5}$ sample mainly contains Fe and secondary metals as Fe> > Sb = Pb > Cu > Sn > Zn.

2.2. Raman microspectrometry measurements

The particles were analyzed using a Labram confocal Raman microspectrometer (HR800, Horiba, SA) equipped with a 100× (N.A. 0.9) Olympus objective. A video camera provided an optical image of the samples. Raman scattering was excited along the optical axis of a microscope objective (defined as *Z*-axis) with a 632.8 nm wavelength He—Cd laser beam. The Labram instrument was equipped with a front-illuminated LN2-cooled charge-coupled device (CCD) detector. The laser power delivered to the sample was about 8 mW and could be attenuated by a set of neutral density filters. The substrates with particles were mounted on the automated XY stages of the microscope without further sample preparation. The XY computer-controlled Raman mapping consists of recording one spectrum for each particle, 20 s of integration time and one accumulation. Raman automated spot mode generates a two-dimensional data matrix $\boldsymbol{D}$ (n × λ), i.e. n spectra, each containing λ = 2040 spectral elements corresponding to a spectral range of approximately 1000 cm^{-1} with a spectral resolution of 4 cm^{-1}. In order to provide a statistical significance in this methodology, we acquired 700 spectra (Fig. A1 in Appendix) from 700 individual particles of $PM_{2.5}$. LABSPEC 5.1 software (HORIBA) was used for spectral acquisition.

3. Processing of Raman spectra data matrix

As mentioned above, the preprocessing step of a Raman data matrix is crucial for suitable statistical treatment of the data. Current methodology for the data treatment is summarized in the Fig. A2 in the Appendix.

3.1. Data normalization

The experimental data matrix (***D***) was filtered for cosmic spikes removal in Labspec 5.1 software (HORIBA) and the data was normalized by the intensity value, generating the ***D'*** matrix.

3.2. Multi-Curve Resolution methodology (MCR-ALS)

Considering the complex sample composition resulting from particles heterogeneity, the procedure is affected by application of a pure variable approach to resolve mixed spectra. Thus, in the first step, the Multivariate Curve Resolution procedure was applied to the ***D'*** Raman data matrix. The methodology consists of (i) determining the number of pure variables (components) and (ii) proceeding to the MCR processing.

The determination of the number of components in terms of application of the MCR-ALS approach is the first assignment. An incorrect choice can lead to information loss (underestimation) or noise component inclusion (overestimation) [47]. Many methods have been proposed to determine the number of components [48–50]. Furthermore, in the case of analysis of environmental particles no a priori knowledge about the number of components is available. Due to that, singular value decomposition (SVD) can be used as a first estimation [51].

The detailed description of the MCR procedure can be found elsewhere [52,53], and is briefly described below.

The ***D'*** matrix was decomposed using multivariate curve resolution with an alternating least squares algorithm (MCR-ALS) in MCR-ALS GUI 2.0 developed for Matlab computing environment [54]. From the ***D'*** matrix, the extracted spectra and their spectral contribution were obtained, as:

$$\boldsymbol{D}' = C' \times S\prime^{T} + E' \tag{3}$$

where ***D'*** is the data matrix of collected Raman spectra after normalization; C' is a spectral contribution matrix, S'^{T} is a transposed matrix with exported spectra and E' is an error matrix including an apparatus function.

The experimental data matrix ***D'*** and the reconstructed data matrix $\boldsymbol{D'^{rec}}$ should be very similar when the quality of the model is high, where:

$$\boldsymbol{D'^{rec}} = C' \times S'^{T} \tag{4}$$

For an equation of the difference, providing the general error value, we used the relative root of sum of square differences (RRSSQ) expressed as follows:

$$RRSSQ = \sqrt{\frac{\sum_{i=1}^{n_{spec}} \sum_{j=1}^{n_{var}} \left(d\prime_{i,j,\lambda} d_{i,j,\lambda}^{rec}\right)^2}{\sum_{i=1}^{n_{spec}} \sum_{j=1}^{n_{var}} d\prime_{i,j,\lambda}^{2}}} \tag{5}$$

where $d'_{i,j,\lambda}$ is the i × j^{th} row and i^{th} column element of ***D'***, $d^{rec}_{i,j,\lambda}$, the i × j^{th} row and i^{th} column element of $\boldsymbol{D}^{rec}$; n_{spec} is the number of mixture spectra; and n_{var} is the number of recorded intensities.

The resolved spectra show difference between original and reconstructed data set lower than 5% RRSSQ.

Application of an alternating least squares algorithm for the initial estimators allows to generate a non-negative spectral contribution matrix (***C'***). The produced ***C'*** matrix does not contain any negative values of spectral contributions which may be difficult to interpret.

As a result, the obtained pure spectra ($\boldsymbol{S'^{T}}$) and spectral contribution profiles (***C'***) describe the sample as its variables. Chemical composition of the sample is identified thanks to the extracted pure spectra ($\boldsymbol{S'^{T}}$) when the matrix of spectral contributions (***C'***) is being used for quantification and characterization of particle chemical mixing (i.e. chemical heterogeneity) through chemometrics methods.

3.3. Identification of compounds within particles

Identification of the components was performed through comparison between the extracted pure Raman spectra ($\boldsymbol{S'^{T}}$ matrix) from MCR-ALS procedure, with the Raman spectra (band positions and relative intensities) in the well-established Raman databases [55–57]. In addition to these well-known Raman databases, we have also collected reference Raman spectra of pure compounds relevant to the speciation of metals in minerals and inorganic species with the same analytical setup. The identified spectra from the $\boldsymbol{S'^{T}}$ matrix were then classified into two groups: (1) the Raman spectra and (2) the other signals, including the non-Raman spectra (i.e. non-active Raman species) and a broad signal obtained either from the substrate or a luminescence background.

3.4. Determination and quantification of particle types within the sample

Calculating the number of particle types within the sample based on their molecular composition is the final goal of this work. The spectral contribution ***C'*** is related to either the presence or the absence of the species within the particle. The proposed methodology consists of an application of the ***C'*** matrix to determine species within the particles followed by multivariate analysis. This methodology corresponds to the previously published works [44,45]. Three steps involved in pretreatment of the ***C'*** matrix are required before the application of clustering and the classification of the particles: (i) removal of contribution profiles related to non-Raman signals (ii) application of an automatic threshold for determination of spectral contribution limit values (iii) binarization of spectral contributions for a principal component analysis and hierarchical clustering.

(i) Based on an identification of the extracted pure Raman spectra, the $\boldsymbol{C'}$ matrix was filtered to eliminate all data related to non-Raman signals, i.e. luminescence or signal from species which are not Raman active. The columns corresponding to the spectral contribution of these signals were removed from the $\boldsymbol{C'}$ matrix. Then the remaining matrix with only Raman spectra corresponds to the $\boldsymbol{C_{comp}}$ matrix in this methodology.

(ii) An application of the threshold resulted from the determination of a boundary value which clearly proves that a compound occurs in a particle. The threshold was implemented in Rstudio 0.99 (R programming language) for spectral contributions of extracted Raman spectra ($\boldsymbol{C_{comp}}$). The threshold was set for a value that satisfies the following eq. (6):

$$x_t = \mu - 3\sigma \qquad (6)$$

where x_t is the filtering boundary value, μ is a mean of the spectral contribution values, and 3σ is a three standard deviation value of the spectral contributions.

(iii) The values below and above the threshold were replaced by 0 and 1, respectively. The matrix after binarization consists of a combination of the previously mentioned values 0 and 1 which correspond to the absence or the presence of the identified compound, respectively. The binarization was important due to the description of chemical mixing of the particles in the sample.

All particles with the data corresponding to the following eq. (7) were removed:

$$\sum_{i=1}^{n} c_i = 0 \qquad (7)$$

where c_i is the *ith* compound value from the $\boldsymbol{C_{comp}}$ matrix after binarization; n is the determined number of identified components in the $\boldsymbol{C_{comp}}$ matrix. Transformation of the $\boldsymbol{C_{comp}}$ matrix with the eq. (7) generates a new binary matrix $\boldsymbol{C_{final}}$ The procedure of the $\boldsymbol{C_{final}}$ matrix extraction is presented in Fig. 1.

The dominant groups of the $PM_{2.5}$ fraction were distinguished by the application of an unsupervised multivariate analysis (no prior knowledge about composition is necessary). The essence of the appropriate sample description in the current methodology is based on two assumptions. The first one considers that it is possible to fully-describe the sample only by a description of all existing combinations of the specified compounds in the matrix. The combinations in this methodology correspond to the possible chemical mixing of the compounds in each particle. The second one assumes that the dominant group designation is needed for a general overview of the sample. This assumption might be even more crucial for the comparison of particles collected from different fractions, sampling sites, etc.

In the current methodology the implementation of PCA is caused by the two main requirements. The first one was to reduce *d*-dimensional space of the input data to a smaller *k*-dimensional subspace, which helps to identify patterns, based on the correlations between features. The second one was the presence of binary data in the $\boldsymbol{C_{final}}$ matrix, which should be avoided in terms of application of an Euclidean distance during clustering. After the PCA approach, scores of the first three principal components were used as new features, generating the $\boldsymbol{C_{pca}}$ data matrix, in which the data was no longer binary. The PCA algorithm used in this methodology can be found elsewhere [58]. The principal component analysis was employed in Rstudio 0.99 (R programming language). It needs to be highlighted that the $\boldsymbol{C_{pca}}$ matrix was created only for HCA approach to obtain the clustering vectors, then the results were specified based on the $\boldsymbol{C_{final}}$ matrix with the vectors obtained from HCA.

Partitional clustering (e.g. k-means) used in this approach seems more reliable because of the more efficient run-time wise as compared to agglomerative clustering, such as Ward's HCA. However, the k-means clustering works well only for clusters which are round shaped and of roughly equal sizes/density. It fails for clusters with non-convex shapes, as well as those with different densities. Thus, in this methodology Hierarchical Clustering Analysis was used.

The HCA approach was implemented in Rstudio 0.99 by application of the Ward's algorithm [59] to the $\boldsymbol{C_{pca}}$ matrix. The chemical mixing of the compounds in each particle in the sample was specified in the $\boldsymbol{C_{final}}$ matrix, where clear information about either the presence or the absence of each compound was included. Therefore, PCA in this particular case was implemented to prepare the data for HCA, but the clustering

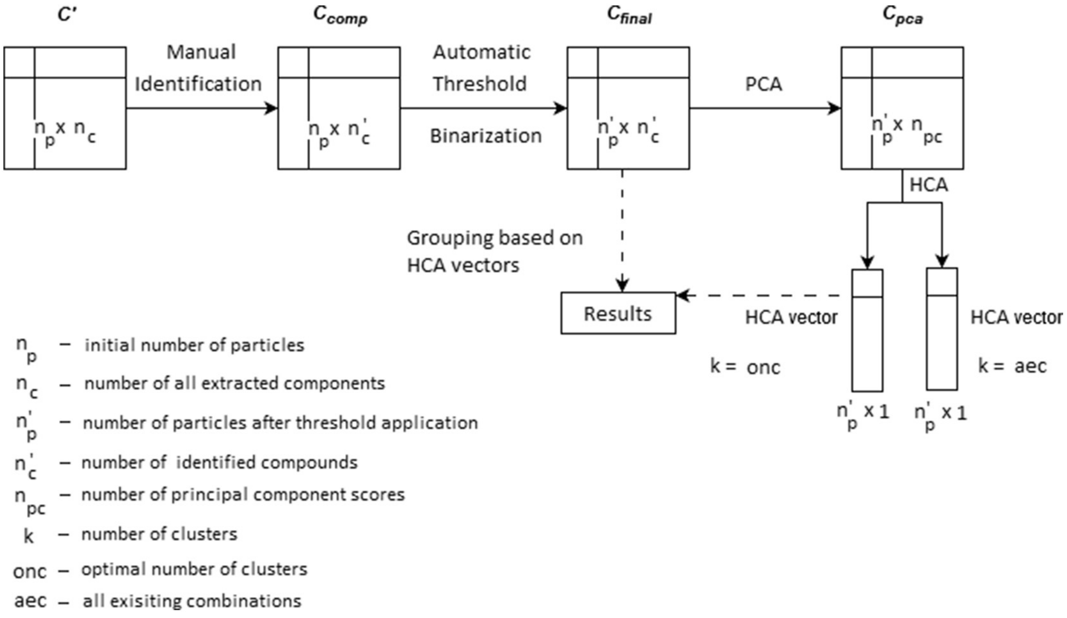

C'
C_comp
C_final
C_pca
Manual Identification
Automatic Threshold
Binarization
PCA
n_p x n_c
n_p x n'_c
n'_p x n'_c
n'_p x n_pc
HCA
Grouping based on HCA vectors
Results
HCA vector
k = onc
HCA vector
k = aec
n'_p x 1
n'_p x 1
n_p – initial number of particles
n_c – number of all extracted components
n'_p – number of particles after threshold application
n'_c – number of identified compounds
n_pc – number of principal component scores
k – number of clusters
onc – optimal number of clusters
aec – all exisiting combinations

Fig. 1. The general scheme of the procedure to obtain the $\boldsymbol{C_{final}}$ matrix.

results were transferred to the $\boldsymbol{C_{final}}$ matrix. Originally, the cut-off point was distinguished based on a visual inspection of the dendrogram (Euclidian distance method), as well as by calculation of Dindex by the Nbclust package [60] in Rstudio 0.99. After the clustering process, the clustering vector was obtained. Based on that, the particles were sorted by the cluster index and then labelled. The labeling was based on the identified compounds content in the $\boldsymbol{C_{final}}$ matrix.

4. Results and discussion

4.1. Identification of compounds within the sample

By application of SVD-based processing, the optimal number of components in the **D'** matrix for MCR-ALS approach was specified. The starting value of the components (determined by SVD) was gradually increased to the final stand of 14 compounds. The final balanced point of this processing was based on the assumption that increasing the number of compounds in the model does not provide any new Raman spectrum in the $\boldsymbol{S'^T}$ matrix [61]. From the MCR-ALS procedure, 14 signals were extracted with a RRSSQ of 2.5%. After the examination of the $\boldsymbol{S'^T}$ matrix, only 6 pure components correspond to the Raman spectra (Fig. A3 in the Appendix), where 5 of them refer to the pure spectra and one to the spectrum of a mixture. Consequently, 9 components in the $\boldsymbol{S'^T}$ are related to the non-Raman signals, i.e. background, broad bands due to substrate effect, luminescence signals, non-Raman active species. The Raman spectra of the individual compounds were assigned to Fe_2O_3, PbO, CuS, FeS_2 and Na_2SO_4 (Fig. A3. in the Appendix). We characterized the mixed spectrum as: $Sn^{2+}{}_3O_2(OH)_2 + CuO + CaSO_4$, described in this work as MS (Mixed Spectrum). The metal rich and the sulfur rich species are typical for particulate matter from a mining environment [62]. Unsurprisingly, some mineral compounds related to the soil composition were not identified since most of clay minerals generate strong fluorescence background [20], which by application of our approach was removed during the identification process. The RMS results were in agreement with the particles' elemental composition ($ng \cdot m^{-3}$) analized by ICP-MS (Appendix, Table A1). In fact, all speciations identified by the RMS spectra were also identified as the main elements by ICP-MS analysis. Very high levels of trace elements were observed for sampling in the tunnels of a multi-elemental mine. Iron, aluminium and sodium, were the most abundant crustal elements. Among trace elements, high levels of lead, copper, antimony and tin were observed. Aluminium was not identified by RMS but could have been segregated by clay minerals as already explained.

4.2. Evaluation of the C′ matrix processing

The **C′** matrix after the identification step charts the contribution of the extracted Raman spectra (**S′T** matrix) in the corresponding particles. The particles corresponding to the identified non-Raman signals were removed from **C′** to fix a new data matrix ($\boldsymbol{C_{comp}}$). From this step 188 spectra of the particles were removed from the matrix, which represent 22% of the sample particle population. At this point the composition was specified for 78% of the sample population, i.e. 512 particles. To assess the composition of the sample, a spectral contribution threshold and binarization of the contribution matrix was necessary (see experimental section). The principal advantage is an evidence about a compound being present in a particle, which was listed in the $\boldsymbol{C_{final}}$ matrix.

4.3. Particles' heterogeneity description

The heterogeneity of the particles is defined here as the occurrence of two or more species within the same particle. To determine the heterogeneity of the particles, the following eq. (10) was applied for each particle in the $\boldsymbol{C_{final}}$ matrix:

$$MR = \sum_{i=1}^{n} c_i \tag{10}$$

Where MR is the particle mixing rank and c_i is the *ith* component from the $\boldsymbol{C_{final}}$ matrix; n is the total number of components in the $\boldsymbol{C_{final}}$.

Based on the MR parameter, the heterogeneity of particles was described, then the results were visualized in the Fig. 2, i.e. the particle with MR = 1 was characterized as a single-component particle (homogeneous), MR = 2 as a binary particle (two components, heterogeneous), MR = 3 as a ternary particle (three components, heterogeneous), etc.

>59% of particles in the sample contain more than one component and can be assigned as mixed particles, where the remaining 41% is corresponding to single-component particles. The majority of them are binary (30%) and ternary (19%). In addition, quaternary particles (9%) were classified as a considerable part of the sample. The vast minority is a group of quinary particles with a contribution of only 1% of the sample population. These results substantiate the application of MCR-ALS processing to experimental data matrix **D′**. However, it should be emphasized, that the mixed spectrum (MS) was treated as a single component in the $\boldsymbol{C_{final}}$ matrix. Notwithstanding, despite that not all the Raman spectra in the **S′T** matrix correspond to a single compound,

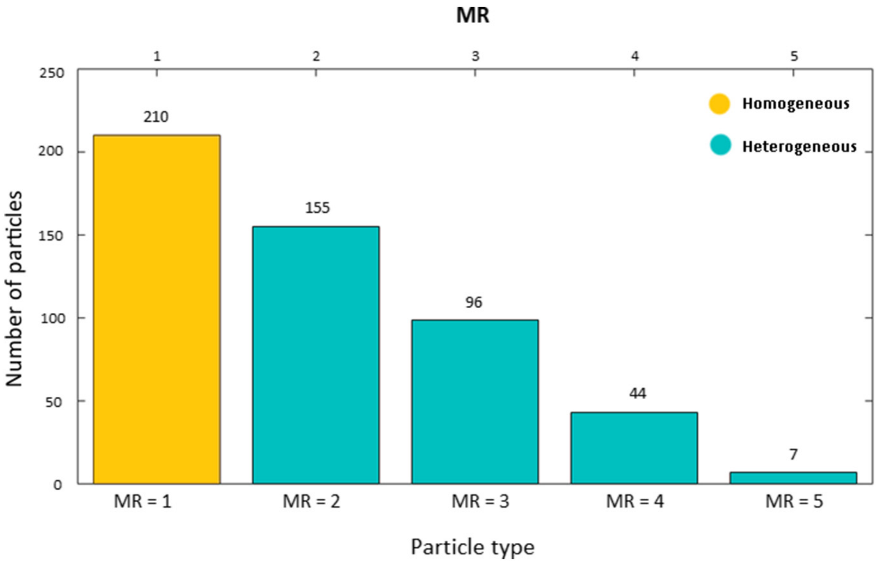


Fig. 2. Classification of the particles based on the MR parameter (with the particle number for each MR).

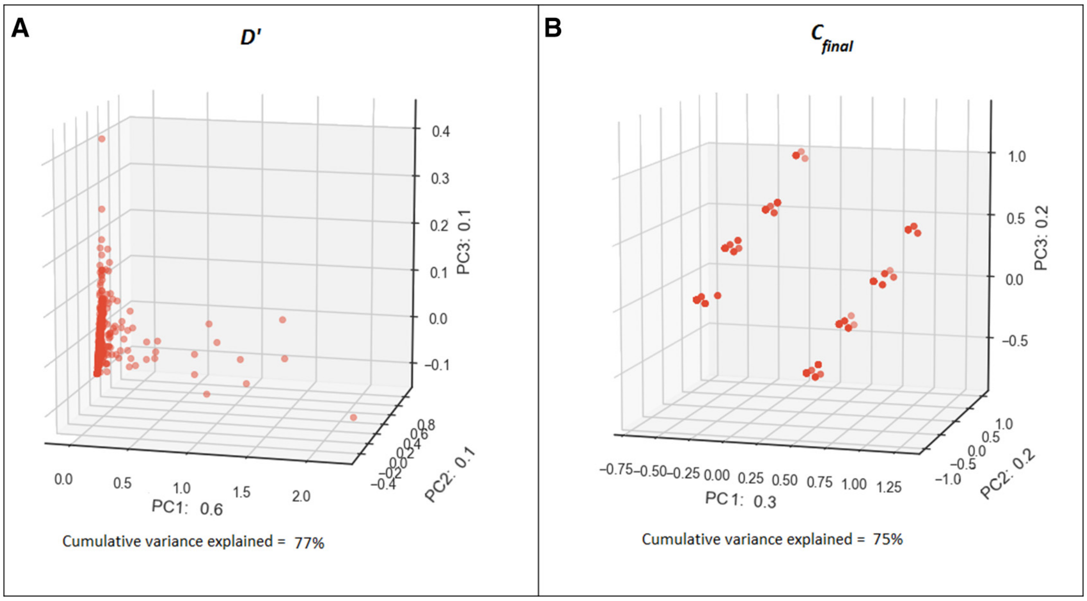


Fig. 3. Three-dimensional scatter plots of ***D'*** (A) and ***C_{final}*** (B) data matrices. PC – principal component.

chemical composition of the sample can be described. The combination with the clustering procedure presented below will provide the chemical composition of mixed particle groups. This underlines the importance of experimental data matrix handling for current procedure.

4.4. Determination of particle groups

The determination of particle types based on their molecular composition was achieved using Ward's hierarchical cluster analysis (HCA) applied on the ***C_{pca}*** matrix. The specification of major clusters in the sample was made as described in the experimental section.

However, initially we estimated the influence of the MCR-ALS approach on the data structure by the application of PCA and HCA (Fig. A4 in the Appendix) to the experimental data matrix (***D'***). The principal component analysis (PCA) was used to visualize two multidimensional data sets ***C_{final}*** (6 variables) and ***D'*** (1024 variables). The first three principal components were used to generate the three-dimensional scatter plots. The cumulative variance explained is around 77% and 75%, for ***D'*** and ***C_{final}*** data matrix, respectively (Fig.3).

The most remarkable result of the PCA is pattern formation in the ***C_{final}*** matrix, comparing to unstructured data in the ***D'*** matrix. The scores of ***D'*** matrix (Fig. 3. A) are cumulated around one corresponding direction with some visible outliers. In the second scatterplot (Fig.3. B), 8 groups of scores are well recognizable, which may provide information about the number of major clusters in the sample confirming the necessity of MCR-ALS procedure for a better description of the sample.

The Ward's hierarchical clustering was applied to ***C_{pca}*** matrix. By examination of the lowest level of the dendrogram, generated by the HCA, the *cut-off* point of the dendrogram was set above the first agglomeration level. Then the number of the branches below this point was used as the number of all existing combinations in the ***C_{final}*** matrix (Fig. A5 in the Appendix). Subsequently, the data vector (column vector) with all existing combinations was exported. Based on the exported clustering vector the values from ***C_{final}*** were classified and listed (Table A1, the Appendix). This step was tested manually by examining of the ***C_{final}*** matrix and manual specification of the combinations.

The cluster-tree graph (dendrogram) generated from the ***C_{pca}*** matrix is shown in the Fig.4.

As expected, the significant gap between the pelting nodes is noticeable; it is due to the data structure after the PCA process. In addition, the calculation of the dindex for an optimal number of clusters was made by Nbclust package in Rstudio 0.99 (Fig. A6 in the Appendix).

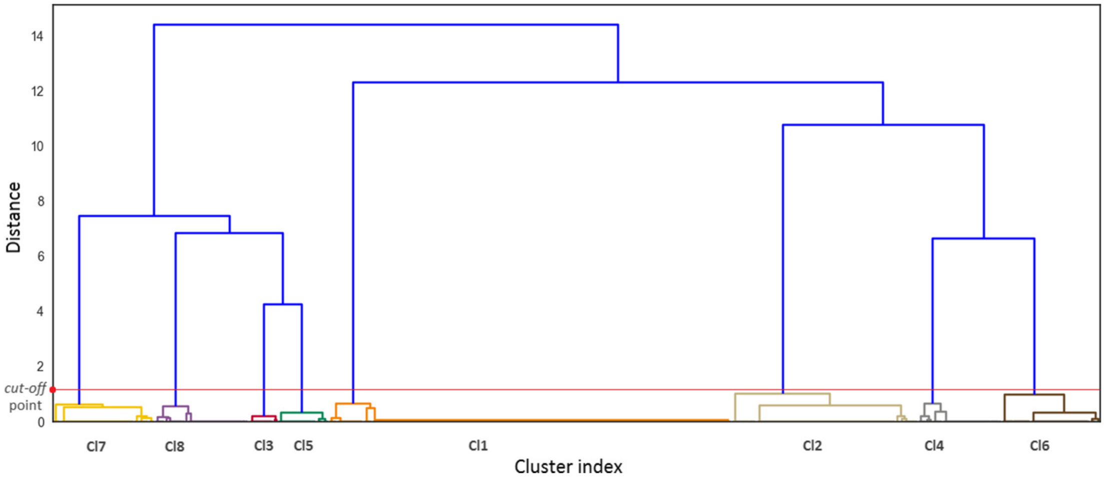


Fig. 4. The dendrogram of ***C_{final}*** after the PCA data set with a marked cut-off point for major particle groups in the sample.

Table 1
The main group of particles and their percentage in the sample.

Cluster number	Type of particles	No.	Sample %
1	Fe_2O_3	197	38%
2	Fe_2O_3 + FeS_2	91	18%
3	Fe_2O_3 + FeS_2 + MS	46	9%
4	Fe_2O_3 + FeS_2 + CuS	41	8%
5	Fe_2O_3 + FeS_2 + CuS + MS	25	5%
6	Fe_2O_3 + CuS	48	9%
7	Fe_2O_3 + MS	50	10%
8	Fe_2O_3 + CuS + MS	14	3%

Abbreviations: No., number of particles;

Table 2
Comparision between the number of particles in the main HCA groups and selfsame mixing combinations.

Cluster number	Type of particles	No.	Corr. No.	Contr.
1	Fe_2O_3	197	176	89%
2	Fe_2O_3 + FeS_2	91	68	75%
3	Fe_2O_3 + FeS_2 + MS	46	30	65%
4	Fe_2O_3 + FeS_2 + CuS	41	29	71%
5	Fe_2O_3 + FeS_2 + CuS + MS	25	19	76%
6	Fe_2O_3 + CuS	48	29	65%
7	Fe_2O_3 + MS	50	36	72%
8	Fe_2O_3 + CuS + MS	14	12	99%

Abbreviations: No., number of particles; Contr., Percentage contribution of cluster in the sample; Corr. No., corrected number of particles;

Both, the dindex values and the visual inspection of the dendrogram indicate the cut-off point of the main clusters in the sample for HCA of the $\boldsymbol{C_{pca}}$ matrix above 8 clusters. This is consistent with the PCA results shown in the Fig. 3B.

Based on the HCA, the particle number for each cluster and the composition of the cluster was assigned (Table 1). Thus, 8 particle types can be distinguished in the $PM_{2.5}$ sample with only three main clusters encountering for >10%. Each specified cluster contains Fe_2O_3. The major particle type corresponds to the single-compound, Fe_2O_3 particles (38% of the population). The second particle type (encountered for 18%) is related to Fe_2O_3 and FeS_2. The next 3 groups (3rd, 4th and 5th) correspond to the particles containing Fe_2O_3, FeS_2 as well as MS or CuS accounting for 9%, 8% and 5%, respectively. The 5th group represents quaternary particles with all the previously mentioned compounds (Fe_2O_3, FeS_2, MS and CuS). The particles that contain both Fe_2O_3 and FeS_2 (as well as the other compounds such as MS, CuS) represent 40% of the sample population. The groups 6th and 8th contain particles with Fe_2O_3 and CuS, while the group 8th also contains MS. Those particles constitute 12% of the sample population. Domination of Fe_2O_3 is correlated with elemental composition from ICP-MS, pointing out Fe (31% of the total reconstructed mass) as the main component of this sample after Al (44%). Al was not identified by RMS because it is mainly included in clay minerals, which are known for the fluorescence signal and organic matter presence in this phase. The mixed spectra (MS) identified the main trace elements Pb and Cu (2.3% and 1%, respectiviely of the total content). We can point out the complementarity of both methods since ICP-MS brings quantification, whereas RMS allows a remarkable identification of speciations.

Regarding the possible component combinations presented in Table A1, it is clear that the HCA based on the $\boldsymbol{C_{pca}}$ analysis provided the mains groups of particles. The corrected number of particles corresponding to all the types of existing chemical mixing combinations in the sample was accounted.

The chemical mixing combinations with the same labels as the groups specified by the HCA were transferred to Table 2 in order to compare the contribution of the particles with exactly the same chemical mixing type as it was specified by cluster labeling, after the HCA. As it can be noticed, the values of the particle quantity in the corresponding clusters are slightly different. In our methodology the labeling procedure was based on the overview of the chemical mixing combination of the particles in the cluster. The most abundant chemical mixing was selected as a label of the cluster. This fact is related to the nature of the binary data in the $\boldsymbol{C_{final}}$ matrix. This incompatibility of particle clustering was corrected based on searching of the analogous combination with the same label as the cluster. Subsequently, the number of particles in the cluster was corrected by the number of particles found in the combination.

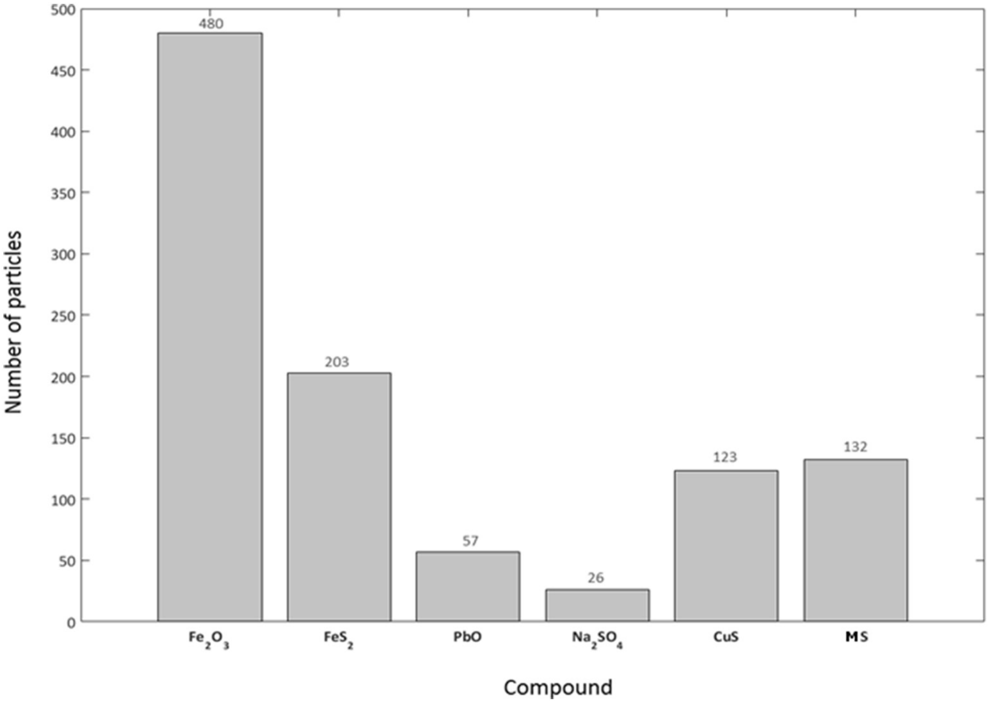


Fig. 5. Contribution of particles containing identified compounds from MCR-ALS.

Finally, the cluster analysis generally tends to characterize the sample under consideration. The HCA was applied for specification of the major clusters, which gives the ability to compare different samples depending on the same category (major groups of particles). Additionally, it is the first factor to classify the sample, e.g. $PM_{2.5}$ fraction can be related to Fe_2O_3-only particles.

Due to the HCA procedure, only 8 main groups of particles were specified for the general sample description. However, for a detailed characterization of the particles based on the collected data, the crucial step is to define all existing chemical mixing combinations in the sample. If the demonstration of the results is applied only on the HCA results, then the presence of other compounds existing in the sample will be eclipsed. It should be noted that due to the results from the MCR-ALS approach, the pure PbO spectrum was extracted. In this example the Pb- containing particles might be important for environmental assessment and due to this fact, all existing combinations of the compounds in the sample were specified. The characterization of the particle composition along with the identified compounds is presented in Fig. 5.

Approximately 96% of the particles contain Fe_2O_3, and almost 40% include FeS_2. These are the two major compounds which are resent in the sample composition. However, due to specific aspects of aerosol chemistry, the content of the secondary compounds such as PbO, may be even more important, e.g. when considering a potential impact of the collected particles on the environment or human health. Slightly >11% of the particles contain PbO. The importance of this information may be crucial for further (more detailed) analysis. It should be emphasized that the impact of Pb-rich airborne particles is broadly described in the literature. What is essentially important is that lead, even at a trace level, can affect human health, e.g. a child's growth and intelligence [63]. The remaining groups are corresponding to CuS-containing (24%), Na_2SO_4-containing (5%) and MS-containing (26%) particles. These results show that classification according to the specific chemical mixing or components is possible.

The labels of clusters were conceptualized based on the most frequently occurring compounds in the cluster. Notwithstanding, it does not mean that all particles in the cluster have exactly the same composition as it was specified in the label and as explained previously (See Table 2). Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) were chosen as the main data analysis algorithms to identify the various particle types and groups [64]. Osan et al. [65] have used PCA for the results from both single particle (SEM/EDX) and bulk (X-ray fluorescence) analysis results of a combined set of approximately 25,000 individual particles collected over Lake Balaton in Hungary. Such data treatment was applied to determine potential sources of the collected aerosol particles.

5. Conclusions

This work provides an account of the data analysis procedure for numerous Raman spectra collected in an automated single particle analysis mode. We have applied an original approach based on MCR-ALS treatment of the $\boldsymbol{D}'$ matrix and multivariate analysis of the resulting contribution matrix ($\boldsymbol{C}'$). The clustering based on the data from MCR-ALS can provide much more relevant information contrary to statistical data treatment of raw Raman spectra. Moreover, multivariate data analysis on a spectral contribution matrix can provide all possible chemical combinations of compounds in the particles. It gives an original and effective way to describe molecular composition of aerosol particles which is complementary to elemental composition provided by ICP-MS. The results obtained from the analysis of the collected particles suggest that the following procedure may have an application for obtaining the molecular composition of particles analyzed by RMS in an automated mode. Moreover, sample characterization based on the results from well-known unsupervised multivariate analysis methods may give a general and detailed description of the collected samples. Almost 78% of the particle population was taken under consideration in this methodology which indicates that 22% of the particles were removed from the data matrix. A major source of unreliability of the algorithm for these particles is linked to the application of RMS, as a unique SPA method. Due to the limitations of RMS the application of another, complementary analytical method for SPA might provide more reliable results of particle composition (e.g. SEM-EDX\RMS system). The composition of the collected particles from Oruro mining environment corresponds to the characteristics of the area. We have obtained rewarding results of the sample speciation and chemical mixing of the particles which can be useful for assessing their environmental and health impact. We believe it could be a starting point for coupling various SPA methods into a unique system based on automatic data analysis.

Acknowledgements

Damian Siepka acknowledges the French government fellowship (BGF) in the frame of the cotutelle PhD programme. LASIR is involved in CaPPA project (Chemical and Physical Properties of the Atmosphere) funded by the Frenech National Research Agency (ANR) through the PIA (Programme d'Investissement d'Avenir) under contract ANR -11- LABX -0005-01.

Appendix A. Appendix

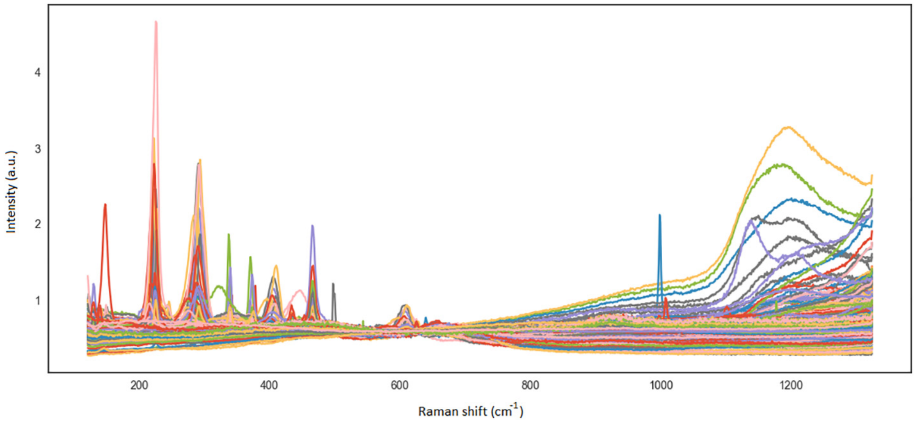


Fig. A1. Raman spectra from the experimental data matrix $\boldsymbol{D}$.

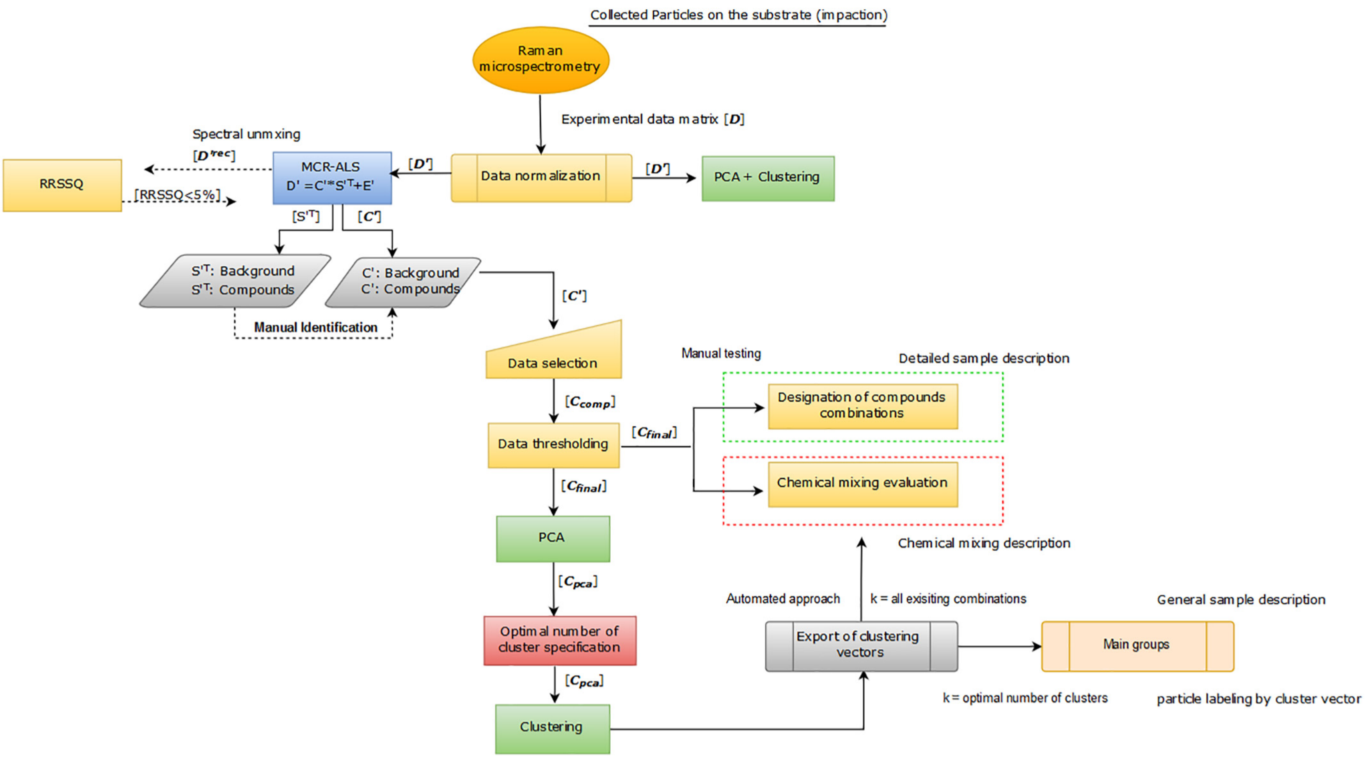

Collected Particles on the substrate (impaction)
Raman microspectrometry
Experimental data matrix [D]
Data normalization
[D']
PCA + Clustering
[D']
MCR-ALS
D' =C'*S'T+E'
Spectral unmixing
[D'rec]
RRSSQ
[RRSSQ<5%]
[S'T]
[C']
S'T: Background
S'T: Compounds
C': Background
C': Compounds
Manual Identification
[C']
Data selection
[Ccomp]
Data thresholding
[Cfinal]
[Cfinal]
Manual testing
Detailed sample description
Designation of compounds combinations
Chemical mixing evaluation
Chemical mixing description
PCA
[Cpca]
Optimal number of cluster specification
[Cpca]
Clustering
Automated approach
k = all exisiting combinations
Export of clustering vectors
k = optimal number of clusters
General sample description
Main groups
particle labeling by cluster vector

Fig. A2. The flow graph of the data matrix during the data analysis.

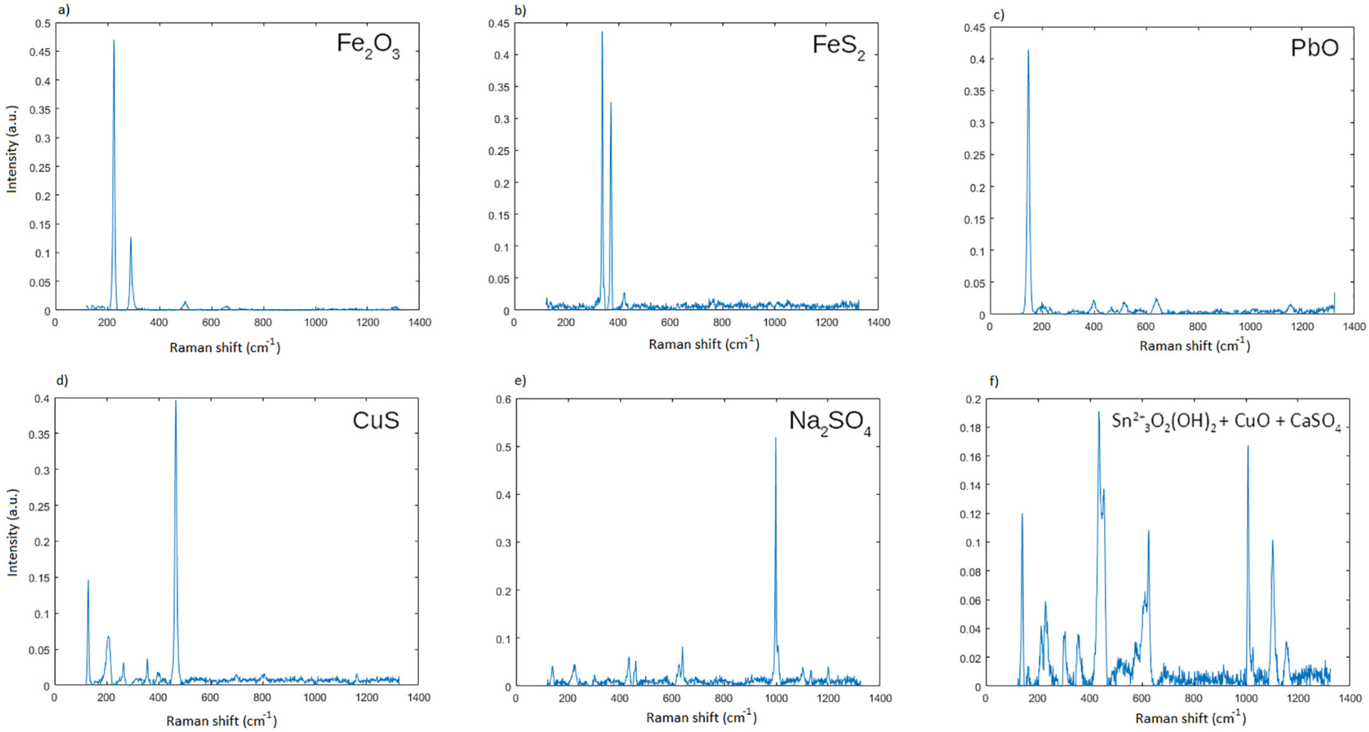


Fig. A3. Raman spectra identified in the ***S'T*** matrix from MCR-ALS processing.

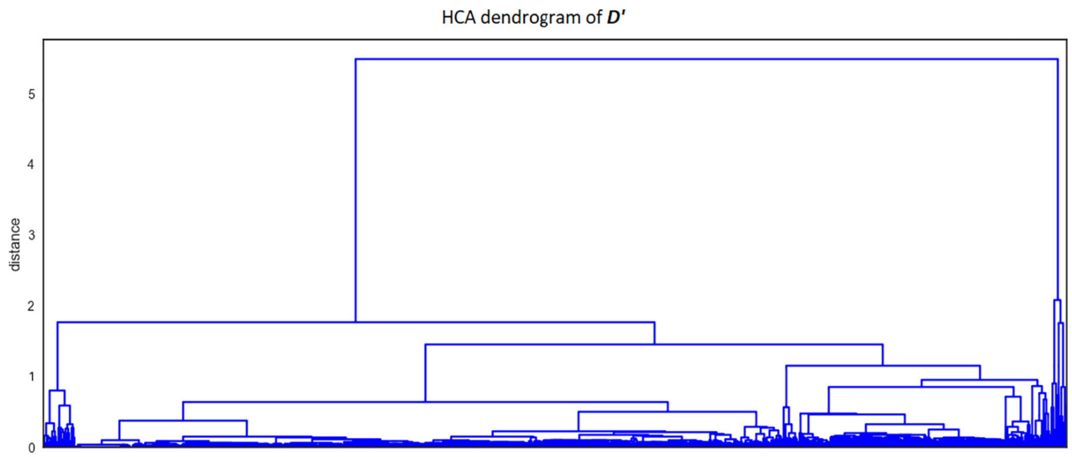


Fig. A4. HCA dendrogram of the ***D'*** spectral data matrix.

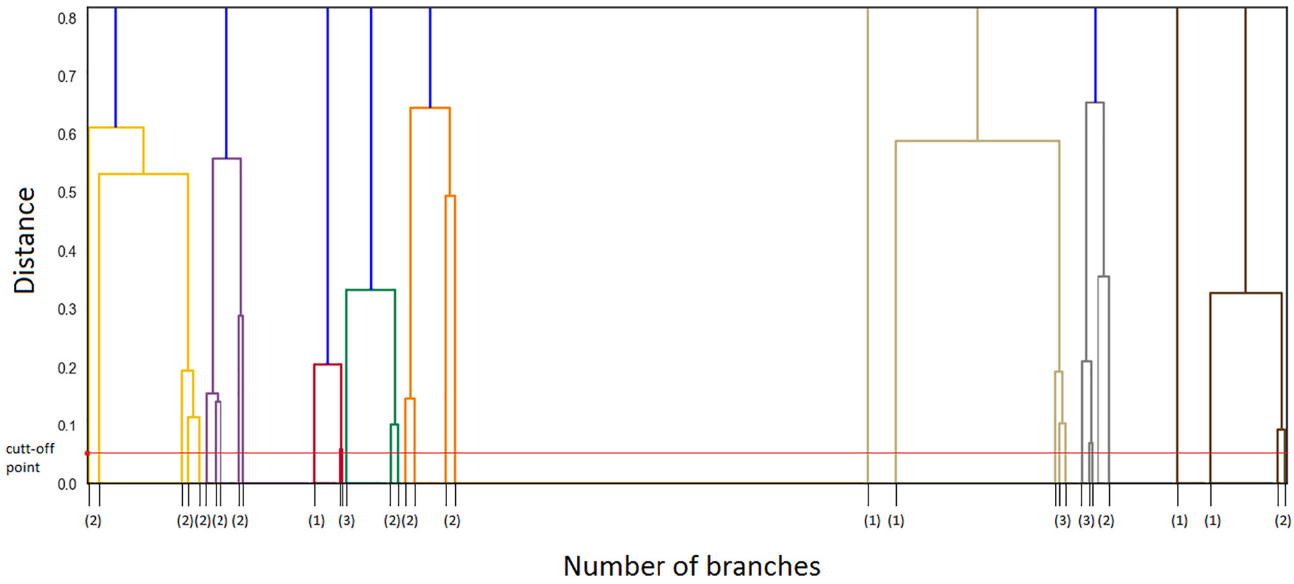


Fig. A5. The bottom-level of the dendrogram generated by Ward's HCA on the $\boldsymbol{C_{final}}$ data matrix. The number of branches corresponds to the number of permutations in the data matrix (34 branches = 34 combinations).

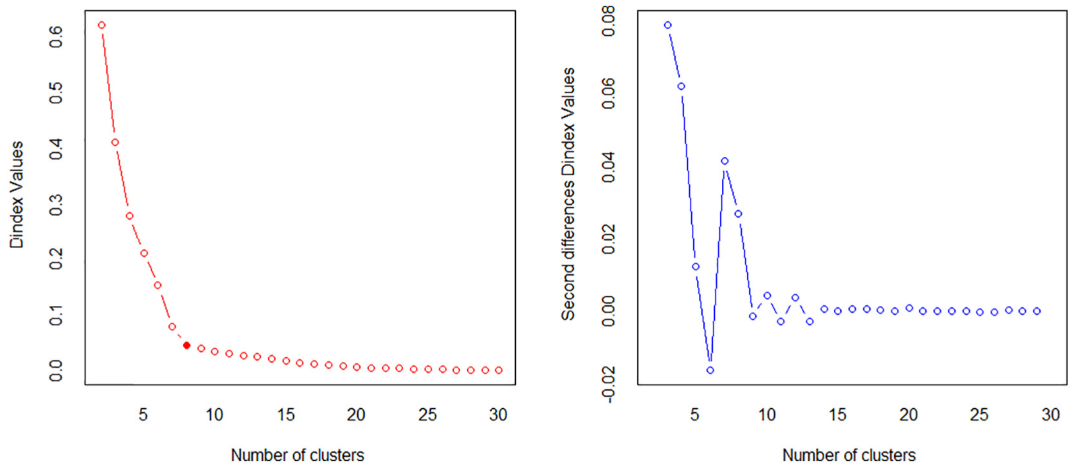


Fig. A6. The dindex (A) and the second differences dindex (B) values calculated for the $\boldsymbol{C_{pca}}$ data matrix.

Table A1
Elemental composition of the PM2.5 sample from the San José Mine galleries. The accuracy of ICP-MS is around ±5%.

element	Ag	Al	As	Ba	Ca	Cd	Co	Cu	Cr	Fe	Mn	Mo	Na	Ni	Pb	Sb	Sn	Zn
ng.m-3	36.5	4864.0	60	25.4	141.0	2.5	0.9	116.8	22.2	3460.0	8.5	1.8	1451.0	2.2	250.0	245.7	98.7	73.4

Table A2
All the existing compounds combinations in the $\boldsymbol{C_{final}}$ data matrix.

MC index	Compounds	Chemical mixing	Particles no.
1	Fe_2O_3	single-compound	176
2	$Fe_2O_3 + FeS_2$	binary	68
3	$Fe_2O_3 + FeS_2$ + MS	ternary	30
4	$Fe_2O_3 + FeS_2$ + PbO	ternary	7
5	$Fe_2O_3 + FeS_2$ + CuS	ternary	29
6	$Fe_2O_3 + FeS_2 + Na_2SO_4$	ternary	2
7	$Fe_2O_3 + FeS_2$ + PbO + MS	quaternary	9
8	$Fe_2O_3 + FeS_2$ + PbO + Na_2SO_4	quaternary	2
9	$Fe_2O_3 + FeS_2$ + PbO + CuS	quaternary	5
10	$Fe_2O_3 + FeS_2$ + CuS + MS	quaternary	19
11	$Fe_2O_3 + FeS_2 + Na_2SO_4$ + MS	quaternary	5
12	$Fe_2O_3 + FeS_2$ + PbO + Na_2SO_4 + MS	quinary	1
13	$Fe_2O_3 + FeS_2$ + PbO + CuS + MS	quinary	3
14	$Fe_2O_3 + FeS_2$ + CuS + Na_2SO_4 + MS	quinary	3
15	Fe_2O_3 + CuS	binary	29
16	Fe_2O_3 + CuS + MS	ternary	12
17	Fe_2O_3 + CuS + Na_2SO_4	ternary	2
18	Fe_2O_3 + CuS + Na_2SO_4 + MS	quaternary	1
19	Fe_2O_3 + PbO	binary	13
20	Fe_2O_3 + PbO + CuS	ternary	3
21	Fe_2O_3 + PbO + MS	ternary	5
22	Fe_2O_3 + PbO + CuS + MS	quaternary	1
23	Fe_2O_3 + PbO + Na_2SO_4 + MS	quaternary	2
24	$Fe_2O_3 + Na_2SO_4$	binary	4
25	$Fe_2O_3 + Na_2SO_4$ + MS	ternary	3
26	Fe_2O_3 + MS	binary	36
27	FeS_2	single-compound	12
28	FeS_2 + CuS	binary	4
29	FeS_2 + CuS + Na_2SO_4	ternary	1
30	FeS_2 + MS	binary	1
31	FeS_2 + PbO + CuS	ternary	2
32	CuS	single-compound	14
33	PbO	single-compound	4
34	MS	single-compound	4

Abbreviations: MC index, mixing combination index; Particle no., number of particles;

References

[1] H. Van Malderen, R. Van Grieken, T. Khodzher, V. Obolkin, V. Potemkin, Composition of individual aerosol particles above Lake Baikal, Siberia, Atmos. Environ. 30 (1996) 1453–1465.

[2] E. Ganor, Z. Levin, R. Van Grieken, Composition of individual aerosol particles above the Israelian Mediterranean coast in summer during the summertime, Atmos. Environ. 32 (1998) 1631–1642.

[3] C.-U. Ro, J. Osán, R. Van Grieken, Determination of low-Z elements in individual environmental particles using windowless EPMA, Anal. Chem. 71 (1999) 1521–1528.

[4] W. Jambers, V. Dekov, R. Van Grieken, Single particle characterisation of inorganic and organic North Sea suspension, Mar. Chem. 67 (1999) 17–32.

[5] B. Sitzmann, M. Kendall, J. Watt, I. Williams, Characterisation of airborne particles in London by computer-controlled scanning electron microscopy, Sci. Total Environ. 298 (1999) 131–145.

[6] C.-U. Ro, J. Osán, I. Szaloki, K.-Y. Oh, H. Kim, R. Van Grieken, Determination of chemical species in individual aerosol particles using ultrathin window EPMA, Environ. Sci. Technol. 34 (2000) 3023–3030.

[7] J. Osán, I. Szaloki, C.-U. Ro, R. Van Grieken, Light element analysis of individual microparticles using thin-window EPMA, Mikrochim. Acta 132 (2000) 349–355.

[8] W. Jambers, V. Dekov, R. Van Grieken, Single particle and inorganic characterization of rainwater collected above the North Sea, Sci. Total Environ. 256 (2000) 133–150.

[9] J. de Hoog, J. Osán, I. Szalóki, K. Eyckmans, A. Worobiec, C.-U. Ro, R. Van Grieken, Thin-window electron probe X-ray microanalysis of individual atmospheric particles above the North Sea, Atmos. Environ. 39 (2005) 3231–3242.

[10] A. Held, K.-P. Hinz, A. Trimborn, B. Spengler, O. Klemm, Chemical classes of atmospheric aerosol particles at a rural site in Central Europe during winter, J. Aerosol Sci. 33 (2002) 581–594.

[11] E.A. Stefaniak, A. Buczyńska, V. Novakovic, R. Kuduk, R. Van Grieken, Determination of chemical composition of individual airborne particles by SEM/EDX and micro-Raman spectrometry: a review, J. Phys. Conf. Ser. 162 (2009).

[12] Y. Batonneau, C. Bremard, J. Laureyns, J.C. Merlin, Microscopic and imaging Raman scattering study of PbS and its photo-oxidation products, J. Raman Spectrosc. 31 (2000) 1113–1119.

[13] Y. Batonneau, J. Laureyns, J.-C. Merlin, C. Bremard, Self-modeling mixture analysis of Raman microspectrometric investigations of dust emitted by lead and zinc smelters, Anal. Chim. Acta 446 (2001) 23–37.

[14] W. Windig, B. Antalek, J.L. Lippert, Y. Batonneau, C. Bremard, Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach, Anal. Chem. 74 (2002) 1371–1379.

[15] N. Dupuy, Y. Batonneau, Reliability of the contribution profiles obtained through the SIMPLISMA approach and used as reference in a calibration process application to Raman micro-analysis of dust particles, Anal. Chim. Acta 495 (2003) 205–215.

[16] Y. Batonneau, C. Bremard, J. Laureyns, J.-C. Merlin, W. Windig, Polarization effects of confocal Raman microspectrometry of crystal powders using interactive self-modeling analysis, J. Phys. Chem. B 107 (2003) 1502–1513.

[17] Y. Batonneau, C. Bremard, L. Gengembre, J. Laureyns, A. Le Maguer, D. Le Maguer, E. Perdrix, S. Sobanska, Speciation of PM10 sources of airborne nonferrous metals within the 3-km zone of lead/zinc smelters, Environ. Sci. Technol. 38 (2004) 5281–5289.

[18] G. Falgayrac, S. Sobanska, J. Laureyns, C. Bremard, Heterogeneous chemistry between PbSO4 and calcite microparticles using Raman microimaging, Spectrochim. Acta A 64 (2006) 1095–1101.

[19] S. Sobanska, G. Falgayrac, J. Laureyns, C. Bremard, Chemistry at level of individual aerosol particle using multivariate curve resolution of confocal Raman image, Spectrochim. Acta A 64 (2006) 1102–1109.

[20] Y. Batonneau, S. Sobanska, J. Laureyns, C. Bremard, Confocal microprobe Raman imaging of urban tropospheric aerosol particles, Environ. Sci. Technol. 40 (2006) 1300–1306.

[21] S. Sobanska, H. Hwang, M. Choël, H.-J. Jung, H.-J. Eom, H. Kim, J. Barbillat, C.-U. Ro, Investigation of the chemical mixing state of individual Asian dust particles by the combined use of electron probe X-ray microanalysis and Raman microspectrometry, Anal. Chem. 84 (2012) 3145–3154.

[22] F.S. Romolo, P. Margot, Identification of gunshot residue: a critical review, Forensic Sci. Int. 119 (2001) 195–211.

[23] J. Osan, B. Alfoldy, S. Torok, R. Van Grieken, Characterisation of wood combustion particles using electron probe microanalysis, Atmos. Environ. 36 (2002) 2207–2214.

[24] C.-U. Ro, H. Kim, K.-Y. Oh, S.K. Yea, C.B. Lee, M. Jang, R. Van Grieken, Single-particle characterization of urban aerosol particles collected in three Korean cities using low-Z electron probe X-ray microanalysis, Environ. Sci. Technol. 36 (2002) 4770–4776.

[25] J. Osan, S. Kurunczi, S. Torok, R. Van Grieken, X-ray analysis of riverbank sediment of the Tisza (Hungary): identification of particles from a mine pollution event, Spectrochim. Acta Part B 57 (2002) 413–422.

[26] S. Hoornaert, R.H.M. Godoi, R. Van Grieken, Single particle characterisation of the aerosol in the marine boundary layer and free troposphere over Tenerife, NE Atlantic, during ACE-2, J. Atmos. Chem. 46 (2003) 271–293.

[27] K. Eyckmans, J. de Hoog, L. Van der Auwera, R. Van Grieken, Speciation of aerosols by combining bulk ion chromatography and thin-window electron probe micro-analysis, Int. J. Environ. Anal. Chem. 83 (2003) 777–786.

[28] S. Hoornaert, R.H.M. Godoi, R. Van Grieken, Elemental and single particle aerosol characterisation at a Background Station in Kazakhstan, J. Atmos. Chem. 48 (2004) 301–315.

[29] A. Worobiec, I. Szaloki, J. Osan, W. Maenhaut, E.A. Stefaniak, R. Van Grieken, Characterisation of Amazon Basin aerosols at the individual particle level by X-ray microanalytical techniques, Atmos. Environ. 41 (2007) 9217–9230.

[30] A. Worobiec, L. Samek, Z. Spolnik, V. Kontozova, E. Stefaniak, R. Van Grieken, Study of the winter and summer changes of the air composition in the church of Szalowa, Poland, related to conservation, Microchim. Acta 156 (2007) 253–261.

[31] A. Worobiec, E.A. Stefaniak, S. Kiro, M. Oprya, A. Bekshaev, Z. Spolnik, S.S. Potgieter-Vermaak, A. Ennan, R. Van Grieken, Comprehensive microanalytical study of welding aerosols with x-ray and Raman based methods, X-Ray Spectrom. 36 (2007) 328–335.

[32] O. Popovicheva, E. Kireeva, N. Persiantseva, M. Timofeev, H. Bladt, N.P. Ivleva, R. Niessner, J. Moldanova, Microscopic characterization of individual particles from multicomponent ship exhaust, J. Environ. Monit. 14 (2012) 3101–3110.

[33] M. Delhaye, M. Dupeyrat, R. Dupeyrat, Y. Levy, An improvement in the Raman spectrometry of very thin films, J. Raman Spectrosc. 8 (6) (1979) 351–352.

[34] S. Sobanska, N. Ricq, A. Laboudigue, R. Guillermo, C. Brémard, J. Laureyns, J.C. Merlin, J.P. Wignacourt, Microchemical investigations of dust emitted by a lead smelter, Environ. Sci. Technol. 33 (9) (1999) 1334–1339.

[35] Y. Batonneau, C. Bremard, L. Gengembre, J. Laureyns, A. Le Maguer, D. Le Maguer, E. Perdrix, S. Sobanska, Speciation of PM10 sources of airborne nonferrous metals within the 3-km zone of lead/zinc smelters, Environ. Sci. Technol. 38 (20) (2004) 5281–5289.

[36] W. Windig, Mixture analysis of spectral methods by multivariate methods, Chemom. Intell. Lab. Syst. 9 (1990) 7–30.

[37] W. Windig, J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist, A.P. Snyder, Interactive self-modeling multivariate analysis, Chemom. Intell. Lab. Syst. 9 (1990) 7–30.

[38] S. Sobanska, G. Falgayrac, J. Rimetz-Planchon, E. Perdrix, C. Brémard, J. Barbillat, Resolving the internal structure of individual atmospheric aerosol particle by the combination of atomic force microscopy, ESEM-EDX, Raman and ToF-SIMS imaging, Microchem. J. 114 (2014) 89–98.

[39] L.A. Reisner, A. Cao, A.K. Pandya, An integrated software system for processing, analyzing, and classifying Raman spectra, Chemom. Intell. Lab. Syst. 105 (2011) 83–90.

[40] A. Cao, A.K. Pandya, G.K. Serhatkulu, R.E. Weber, H. Dai, J.S. Thakur, V.M. Naik, R. Naik, G.W. Auner, R. Rabah, D.C. Freeman, A robust method for automated background subtraction of tissue fluorescence, J. Raman Spectrosc. 38 (2007) 1199–1205.

[41] J. Ofner, K.A. Kamilli, E. Eitenberger, G. Friedbacher, B. Lendl, A. Held, H. Lohninger, Chemometric analysis of multisensor hyperspectral images of precipitated atmospheric particulate matter, Anal. Chem. 87 (2015) 9413–9420.

[42] P.V. Jentzsch, B. Kampe, V. Ciobota, P. Rösch, J. Popp, Inorganic salts in atmospheric particulate matter: Raman spectroscopy as an analytical tool, Spectrochim. Acta A Mol. Biomol. Spectrosc. 115 (2013) 697–708.

[43] R.L. Craig, A.L. Bondy, A.P. Ault, Computer-controlled Raman microspectroscopy (CC-Raman): a method for rapid characterization of individual atmospheric aerosol particles, Aerosol Sci. Technol. (2017) 1–14.

[44] J.S. Liu, J.L. Zhang, M.J. Palumbo, C.E. Lawrence, Bayesian clustering with variable and transformation selections, Bayesian Statistics 7 (2003) 249–275.

[45] P. Tamayo, D. Scanfeld, B.L. Ebert, M.A. Gillette, C.W.M. Roberts, J.P. Mesirov, Metagene projection fro cross-platform, cross-species characterization of global transcriptional states, PNAS 104 (2007) 5959–5964.

[46] S. Potgieter-Vermaak, R. Van Grieken, Preliminary evaluation of micro-Raman spectrometry for the characterization of individual aerosol particles, Appl. Spectrosc. 60 (1) (2006) 39–47.

[47] H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, M. Inoue, H. Toki, O. Minowa, T. Noda, J. Kikuchi, Identification of reliable components in multivariate curve resolution-alternating least squares (MCR-ALS): a data- driven approach across metabolic processes, Sci Rep 5 (2015), 15710. .

[48] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.

[49] R.B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (1966) 245–276.

[50] J.A. Horn, A. Rationale, Test for the number of factors in factor analysis, Psychometrika 30 (1965) 179–185.

[51] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. de Juan, A. Gorzsás, Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS), Nat. Protoc. 10 (2015) 217–240.

[52] W. Windig, Mixture analysis of spectral methods by multivariate methods, Chemom. Intell. Lab. Syst. 9 (1990) 7–30.

[53] W. Windig, J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist, A.P. Snyder, Interactive self-modeling multivariate analysis, Chemom. Intell. Lab. Syst. 9 (1990) 7–30.

[54] Jaumot, J., de Juan, A., Tauler, R., 2015. MCR-ALS GUI 2.0: new features and applications. Chemometrics and Intelligent Laboratory Systems 140 (2015), 1–12.

[55] Jobin-Yvon Horiba, Mineral Spectroscopy Server.

[56] California Institute of Technology, Pasadena, California, USA.

[57] Department of Physics and Earth Sciences from the University of Parma, Italy.

[58] Wall, et al., in: D.P. Dubitzky, W. Granzow (Eds.), A Practical Approach to Microarray Data Analysis, Kluwer Academic Publishers, New York, 2003.

[59] J.H. Ward Jr., Hierarchical Grouping to Optimize an Objective Function, J. Am. Stat. Assoc. 58 (301) (1963).

[60] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: an R package for determining the relevant number of clusters in a data set, J. Stat. Softw. 61 (2014) 1–36.

[61] F. de Souza Lins Borba, T. Jawhari, R.S. Honorato, A. de Juan, Confocal Raman imaging and chemometrics applied to solve forensic document examination involving crossed lines and obliteration cases by a depth profiling study, Analyst 142 (2017) 1106–1118.

[62] F. Moricz, I.F. Walder, F. Madai, Geochemical and Mineralogical Characterization of Waste Material from the Itos Sn-Ag Deposit, Bolivia. 8th International Conference on Acid Rock Drainage, 8th Intern, 2009 1–10.

[63] J. Wang, P. Guo, X. Li, J. Zhu, T. Reinert, J. Heitmann, D. Spemann, J. Vogt, R.-H. Flagmeyer, T. Butz, Source identification of lead pollution in the atmosphere of Shanghai City by analyzing single aerosol particles (SAP), Environ. Sci. Technol. 34 (2000) 1900–1905.

[64] C. Xhoffer, P. Bernard, R. Van Grieken, Chemical characterization and source apportionment of individual aerosol particles over the North Sea and the English Channel using multivariate techniques, Environ. Sci. Technol. 25 (1991) 1470–1478.

[65] J. Osán, B. Alföldy, S. Kurunczi, S. Török, L. Bozó, J. Injuk, A. Worobiec, R. Van Grieken, Characterization of atmospheric aerosol particles over Lake Balaton, Hungary, using X-ray emission methods, Idõjárás 105 (2001) 145–156.

3.2.2. Chemical mixing and molecular composition of particles collected within a gallery of polymetallic ore extraction in the Oruro area

The methodology developed previously has been applied for the characterization and the quantification of species within particle samples as well as the description of the chemical mixing. The aim of this study is to evaluate the chemical and mixing state evolution of the particles as function of the particle size in the context of mining environment. The detailed characterization of particles collected within a gallery of polymetallic ore extraction (Sn, Sb, Ag, Zn, Pb) located in the Oruro mine area., was investigated in regards to the relation between four samples ($PM_{10-2.5}$ - sample A, $PM_{2.5-1}$ - sample B, $PM_{1-0.5}$ - sample C and $PM_{0.5}$ - sample D) of collected particles. In addition, particles deposited on personal miner's filters were collected and included as sample FM. This part of the work demonstrates the potential of the presented algorithm in providing a relevant information on atmospheric chemistry and environmental risk assessment.

3.2.2.1. Sample description and analysis

As described in the paragraph 3.1.1, the particles were sampled inside the gallery of a mine (200m depth) located in Oruro area (17°46'0"S - 67°28'60"W). The particle collection was performed using personal cascade impactor (SIOUTAS) allowing sampling four size fractions of particles i.e. $PM_{10-2.5}$, $PM_{2.5-1}$, $PM_{1-0.5}$ and $PM_{0.5}$ at 9 $L.min^{-1}$. In addition, personal filters from the miner's protection mask were also collected. The particles were analysed using a confocal Raman microspectrometer as it was described previously. For each individual particle a spectrum was obtained. The numbers of collected spectra are as follows: sample A ($PM_{10-2.5}$) – 204 spectra corresponding to 204 particles, sample B ($PM_{2.5-1}$) – 700 spectra, sample C ($PM_{1-0.5}$) – 155 spectra, sample D ($PM_{0.5}$) – 258 spectra and sample FM – 250 spectra.

© 2017 Tous droits réservés.

3.2.2.2. Identification of species within each size fractions.

The acquired spectra were gathered in the matrices, where each spectrum was located in the single row of the spreadsheet and wavenumbers were located in columns. Such a data structure was required both for MCR-ALS analysis and for the multivariate analysis. The collected spectra for each fraction are presented in Fig. 11. Considering the complex sample composition resulting from particles heterogeneity, the procedure is affected by application of a pure variable approach to resolve mixed spectra. Thus, in the first step, the Multivariate Curve Resolution procedure was applied for getting extracted spectrum and C matrix as described in the procedure shown in the paragraph 3.2.1.

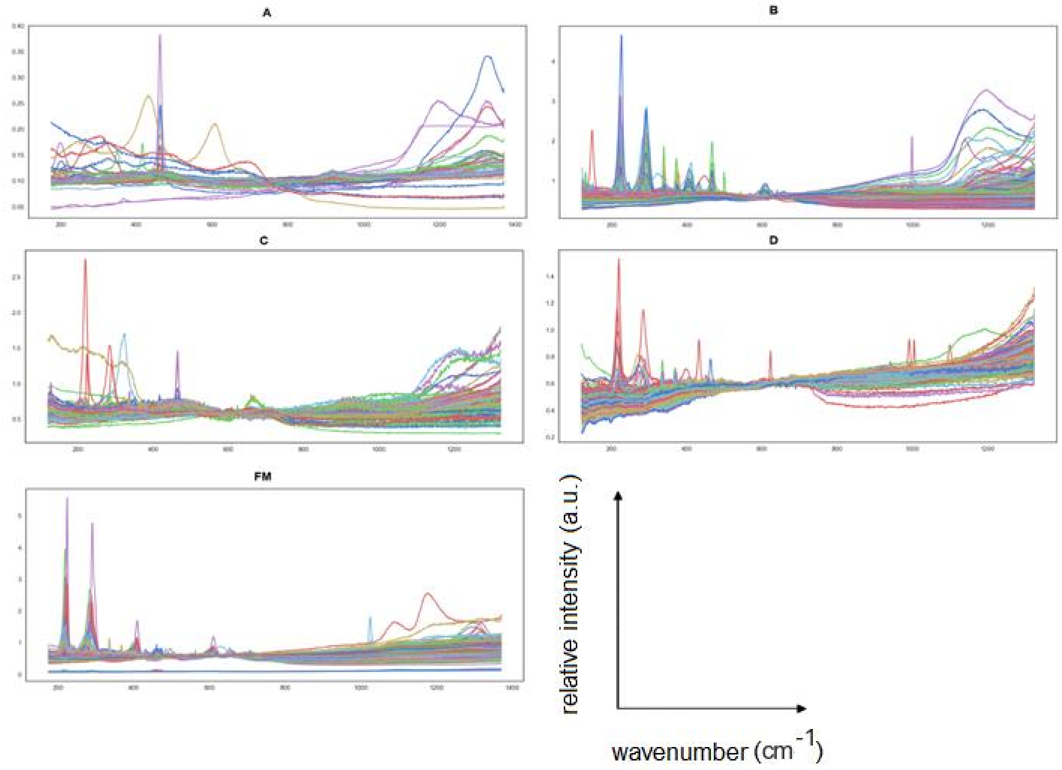


Fig. 11. Raman spectra of mine dust particles acquired from four different fractions: A ($PM_{10-2.5}$), B ($PM_{2.5-1}$), C (PM_1), D ($PM_{0.5}$) and FM (miner's filter).

© 2017 Tous droits réservés.

A unique C matrix was constructed from A, B, C, D and FM samples and used for extracting the pure spectra by MCR. The total number of extracted spectra was 23, where 56% of them (13 spectra) were classified as non-Raman spectra or/and signal from the background. The corresponding spectral contribution from the non-Raman spectra and signal from the background were removed from the C matrix.

The mean value of the removed signals' contribution was 62%. This value indicate that obtaining of all existing Raman spectra in the matrix requires the iterative approach of the MCR procedure, where the starting point is the value estimated by SVD (see the 3.2.1. subsection). For all MCR-ALS procedure, the RRSSQ value was found less than 5%.

The Raman spectra extracted by MCR were assigned for each sample. The Raman spectra extracted by MCR procedure from sample A are presented below (Fig. 12).

© 2017 Tous droits réservés.

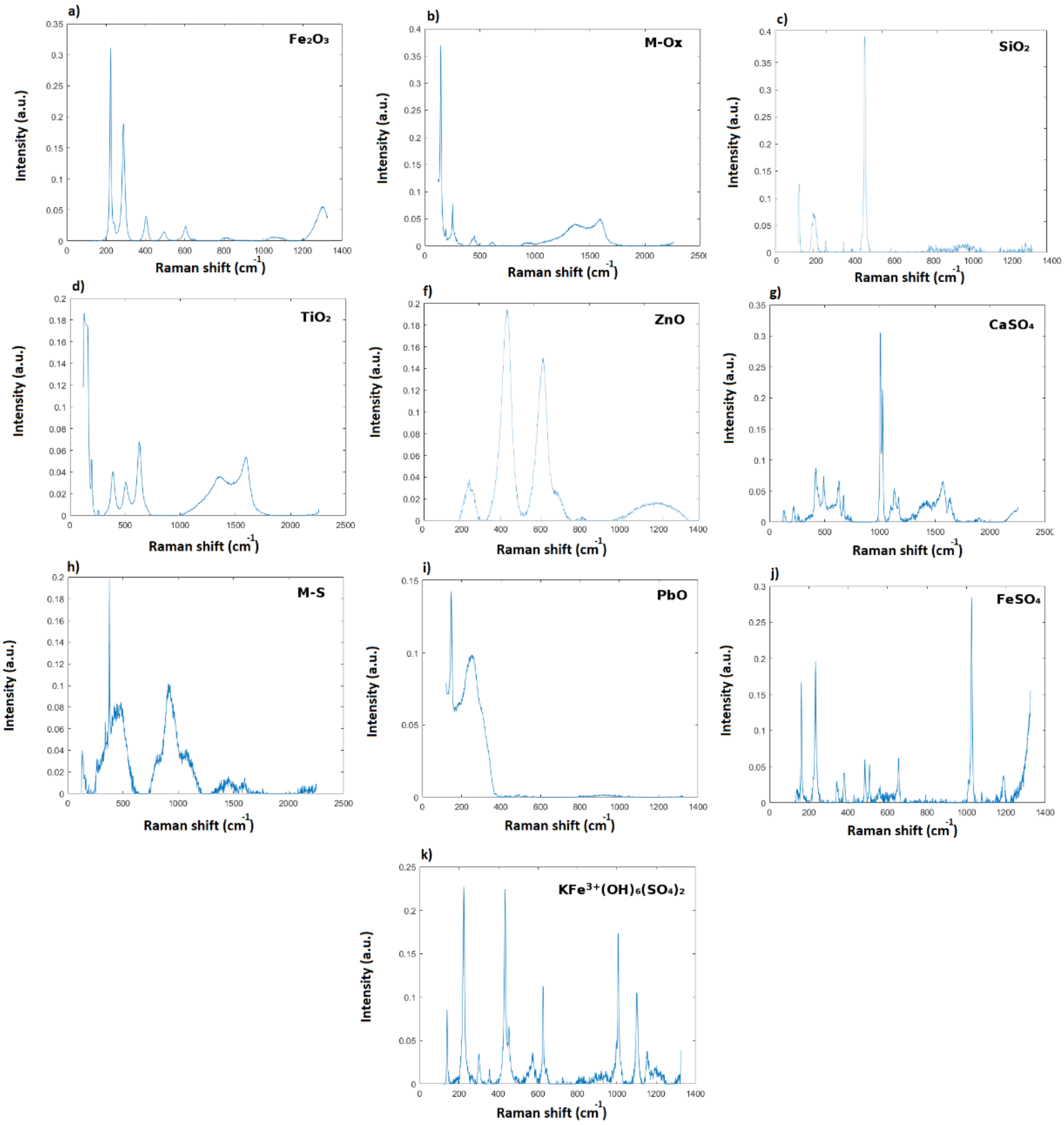


Fig. 12. “Pure” spectra exported into S^T matrix by MCR approach from PM_{10} fraction (sample A). M-Ox – metal oxide, M-S – metal sulphide.

The number of the identified Raman spectra in sample A was 10, and they were: Fe_2O_3, M-Ox (metal oxide), SiO_2, TiO_2, ZnO, $CaSO_4$, M-S (metal sulphide), PbO, $FeSO_4$, $KFe^{3+}(OH)_6(SO_4)_2$ (Jarosite). The identification of the components was performed through comparison between the extracted pure Raman spectra from MCR procedure, with the Raman spectra (band positions and relative intensities) in the well-established Raman databases (more in subsection 3.2.1).

© 2017 Tous droits réservés.

Similarly, the Raman spectra identified in sample B are presented below (Fig. 13).

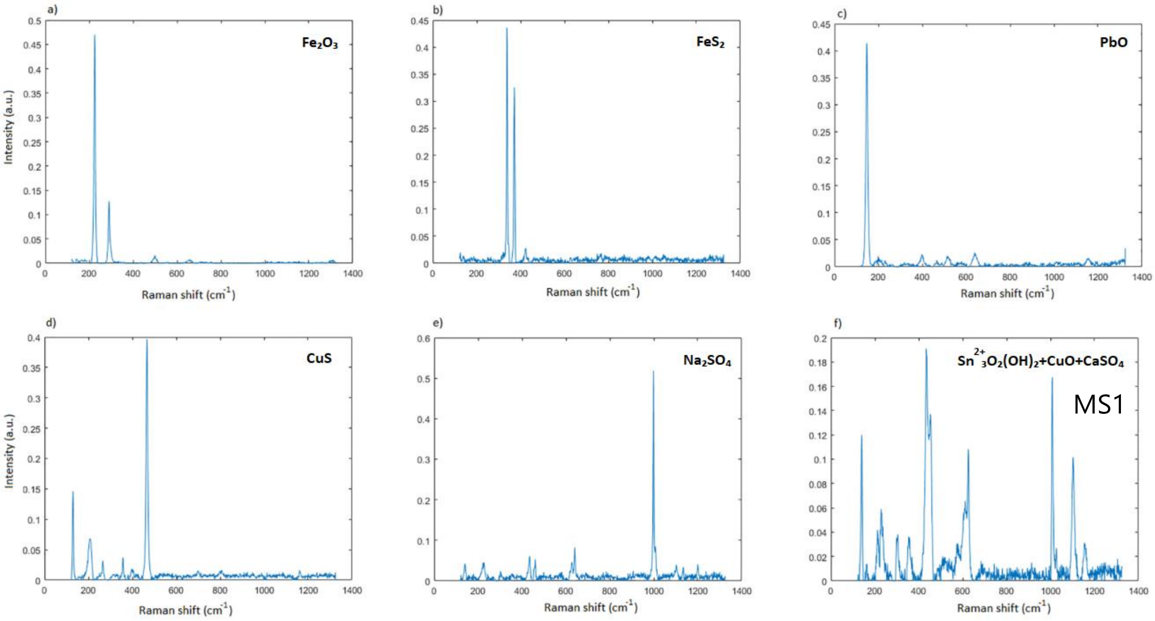


Fig. 13. "Pure" spectra exported into S^T matrix by MCR approach from $PM_{2.5}$ fraction (sample B).

From the MCR procedure, 14 spectra were extracted from the data collected from sample B. After the examination of the S^T matrix, only 6 pure components correspond to the Raman spectra (Fig. 13), where 5 of them refer to the pure spectra and one to the spectrum of a mixture. Consequently, 9 components are related to the non-Raman signals, i.e. background, broad bands caused by a substrate effect, luminescence signals, non-Raman active species, which in total is 64% of signals in the S^T matrix. The Raman spectra of the individual compounds were assigned to Fe_2O_3, PbO, CuS, FeS_2 and Na_2SO_4. We characterized the mixed spectrum as: $Sn^{2+}{}_3O_2(OH)_2$ + CuO + $CaSO_4$ (MS1 – mixed spectrum in this work).

The MCR approach was performed on sample C and produced 16 spectra. After the examination of the exported S^T matrix, only 6 was identified as Raman spectra, which is 38% of the exported signals. The Raman spectra were assigned to Fe_2O_3, Fe_3O_4, FeS_2, SiO_2, ZnS and mixed spectrum of CuO + $CaSO_4$ (MS2) (Fig. 14).

© 2017 Tous droits réservés.

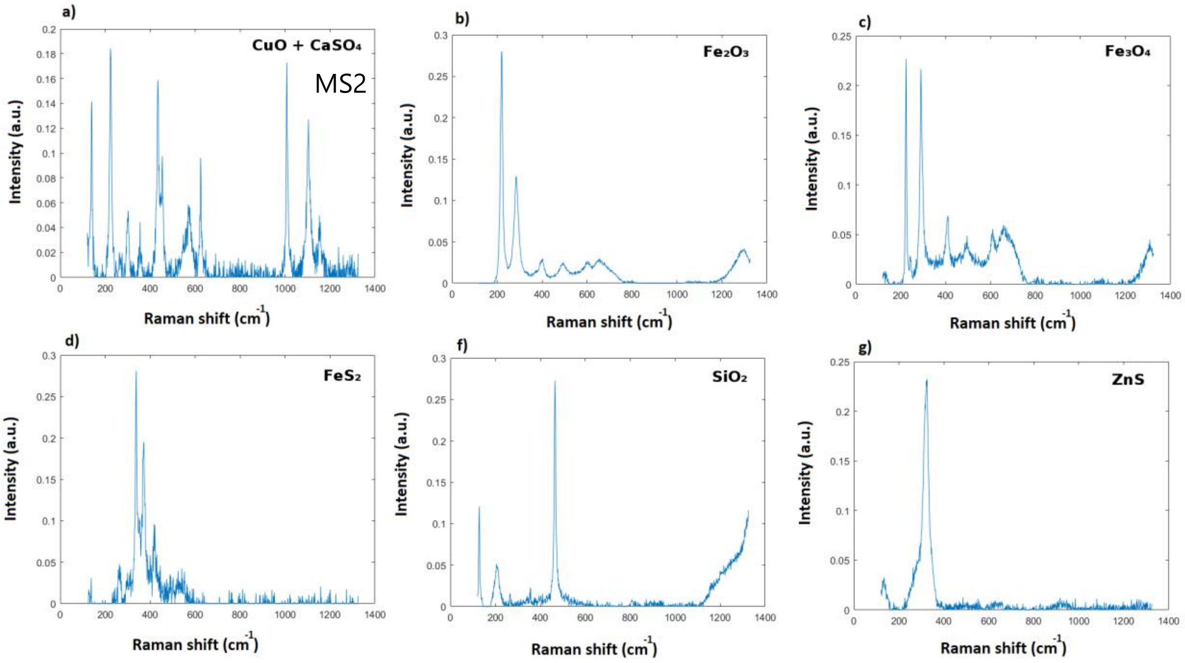


Fig. 14. “Pure” spectra exported into S^T matrix by MCR approach from PM_1 fraction (sample C).

The extracted spectra represent mainly Fe-rich and S-rich compounds, which is consistent with the results from the previous samples.

The Raman spectra from sample D are presented below (Fig. 15).

© 2017 Tous droits réservés.

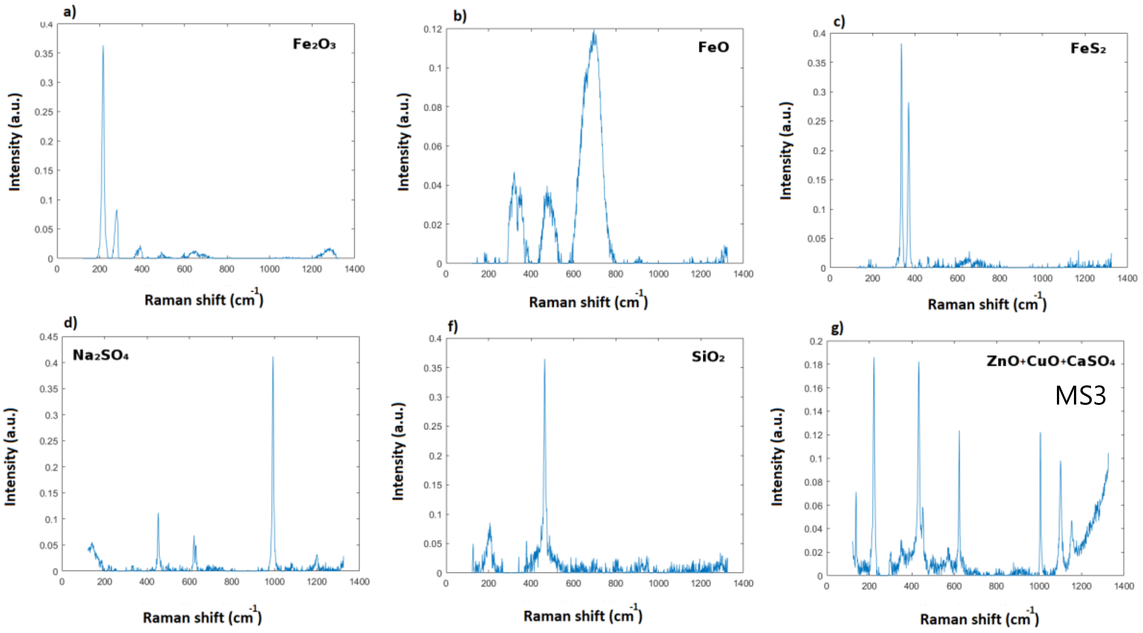


Fig. 15. “Pure” spectra exported into S^T matrix by MCR approach from $PM_{0.1}$ fraction (sample D).

By means of the MCR procedure, 18 spectra were extracted from sample D. After the examination of the S^T matrix, only 6 pure components correspond to the Raman spectra (Fig. 15), where 5 of them refer to the pure spectra and one to the spectrum of a mixture. Consequently, 12 components are related to the non-Raman signals, i.e. background or broad bands, caused by substrate effect, luminescence signals, non-Raman active species, which is 66% of the signals in the S^T matrix. The Raman spectra of the individual compounds were assigned to Fe_2O_3, FeO, FeS_2, Na_2SO_4 and SiO_2. The mixed spectrum was identified as ZnO + CuO + $CaSO_4$ (MS3).

© 2017 Tous droits réservés.

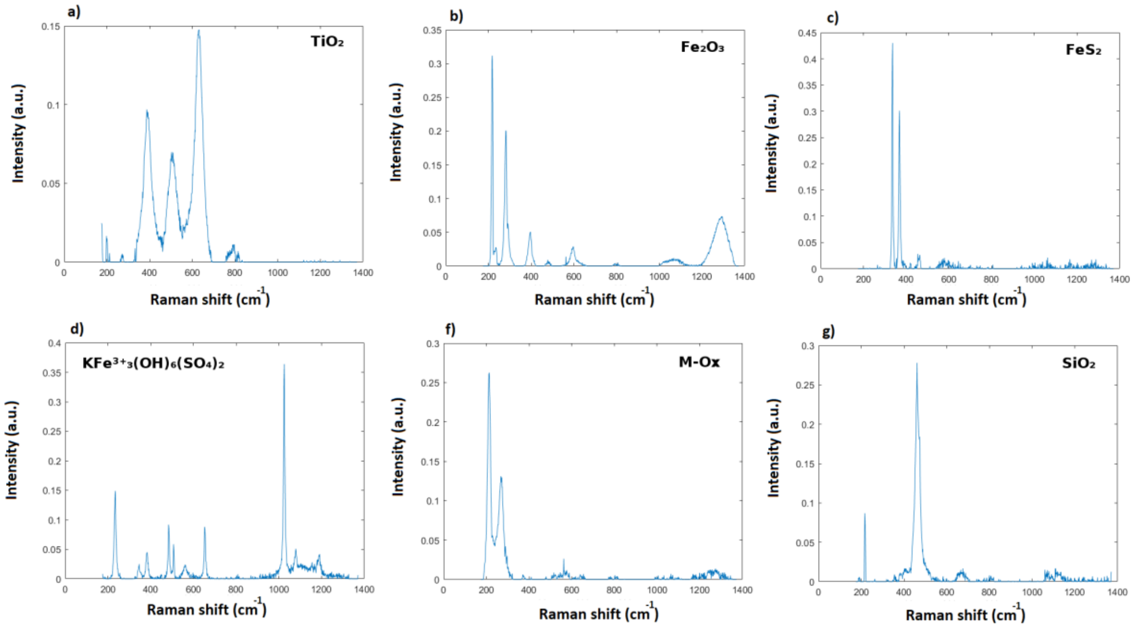
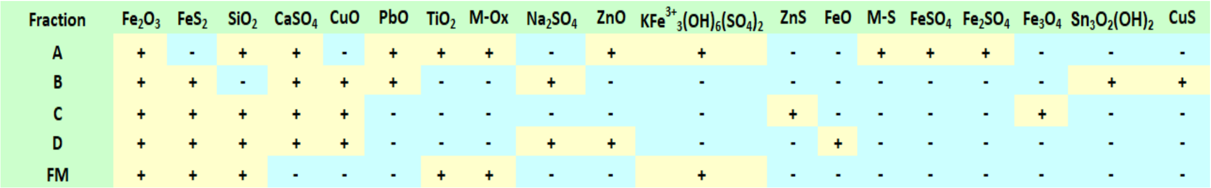


Fig. 16. "Pure" spectra exported into S^T matrix by MCR approach from miner's filter (sample FM).

The Raman spectra extracted by MCR approach from the miner's filter (sample FM) are: TiO_2, Fe_2O_3, FeS_2, $KFe^{3+}(OH)_6(SO_4)_2$, M-Ox (metal oxide) and SiO_2 (Fig. 16). The total number of signals in the S^T matrix was 16, where during identification 10 signals assigned to the non-Raman spectra and background signal were removed (63%).

The presence of the individual compounds extracted by the MCR approach from all the fractions was summarized in the Table 5.

Table 1. The presence of individual chemical compounds in the studied samples



Fraction	Fe_2O_3	FeS_2	SiO_2	$CaSO_4$	CuO	PbO	TiO_2	M-Ox	Na_2SO_4	ZnO	$KFe^{3+}{}_3(OH)_6(SO_4)_2$	ZnS	FeO	M-S	$FeSO_4$	Fe_2SO_4	Fe_3O_4	$Sn_3O_2(OH)_2$	CuS
A	+	-	+	+	-	+	+	+	-	+	+	-	-	+	+	+	-	-	-
B	+	+	-	+	+	+	-	-	+	-	-	-	-	-	-	-	-	+	+
C	+	+	+	+	+	-	-	-	-	-	-	+	-	-	-	-	+	-	-
D	+	+	+	+	+	-	-	-	+	+	-	-	+	-	-	-	-	-	-
FM	+	+	+	-	-	-	+	+	-	-	+	-	-	-	-	-	-	-	-

In the Table 1, the four main compounds are Fe_2O_3, FeS_2, SiO_2 and $CaSO_4$. The collected particles are mainly containing Fe and S-rich compounds, which is

© 2017 Tous droits réservés.

complementary to the specification of the Oruro mining environment (Banks et al. 2002). The extracted compounds are expected for the dust particles collected in the polymetallic mining environment (U.S Department Of Health And Human Services 2003). The metal rich and the sulfur rich species are typical for particulate matter from a polymetallic mining environment (Moricz et al. 2009).

In addition, the common compound, which was identified in almost all fractions (except sample B), is SiO_2. It should be noted, that crystalline silica particles could result in the initiation and progression of interstitial lung disease. Pathogenesis is the consequence of damage to lung cells and resulting lung scarring associated with activation of fibrotic processes (Castranova 2000).

3.2.2.3. Relationships between the species in the samples

Another issue taken into account in this work was the relationships among the specified compounds based on hierarchical clustering for evaluation of the mixing state for each sample. In addition to a classical Euclidean distance measurement method, Dindex calculations were initially used to determine the effectiveness of this procedure and the usefulness of the data for a further procedure. All calculations (Dindex, PCA and HCA) were performed on the C matrix as described in the subsection 3.2.1. The results for Dindex calculation are presented in Fig. 17.

© 2017 Tous droits réservés.

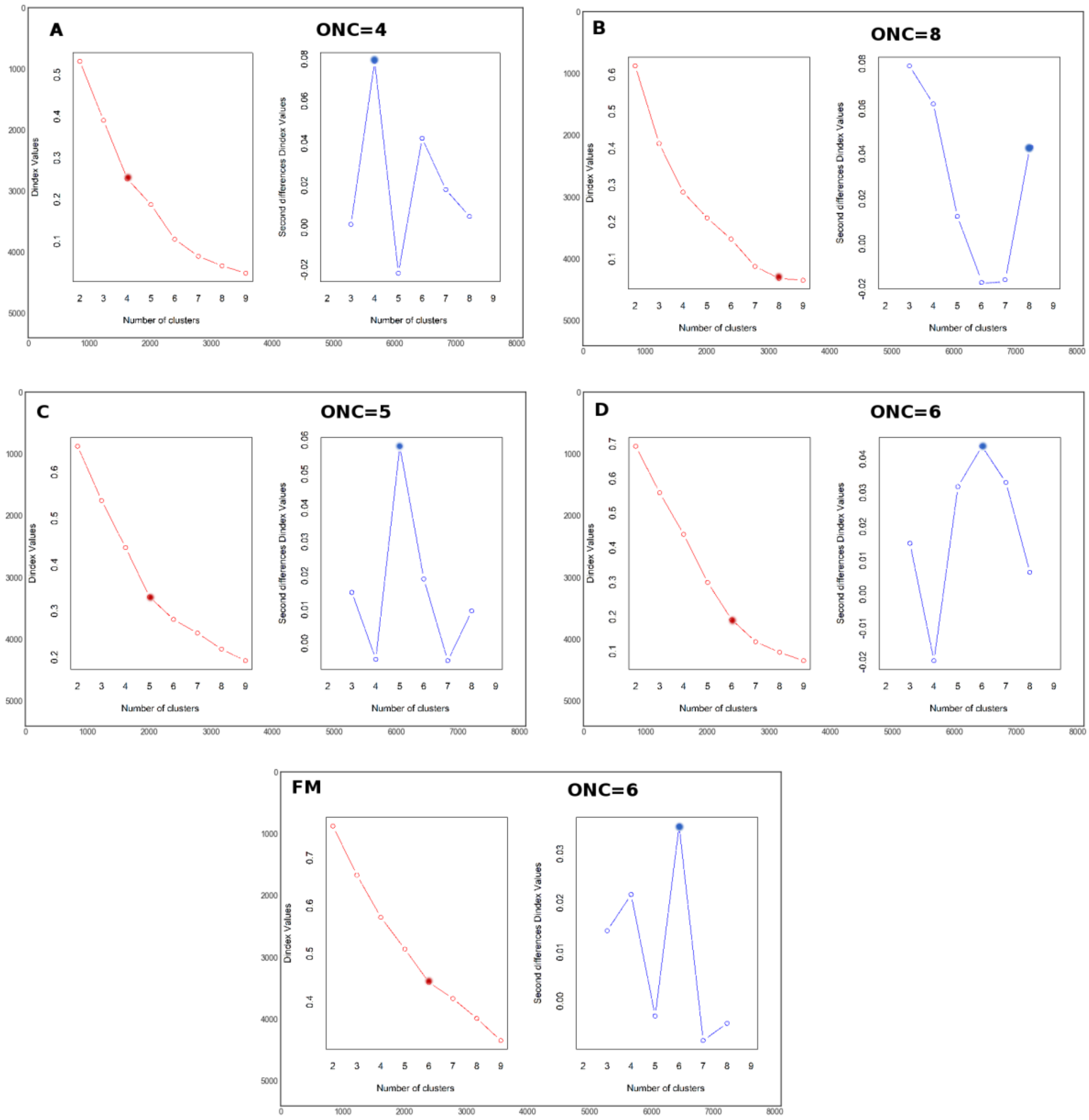


Fig. 17. Dindex values calculated for each fraction spectral contribution matrix exported from MCR-ALS approach. The optimal number of cluster (ONC) was specified for Ward's HCA. The accuracy of the Dindex is given in the subsection 3.2.1.

The calculated Dindex values indicate the optimal number of clusters (ONC) within the range 4-8 clusters. The lowest value among the main groups was calculated for sample A (4 clusters) where in turn the highest was specified for sample B (8 clusters). The consecutive fractions are as follows: sample C (5 clusters), sample D (6 clusters)

© 2017 Tous droits réservés.

and sample FM (6 clusters). The calculated values were correlated with the PCA and Ward's HCA results for each fraction.

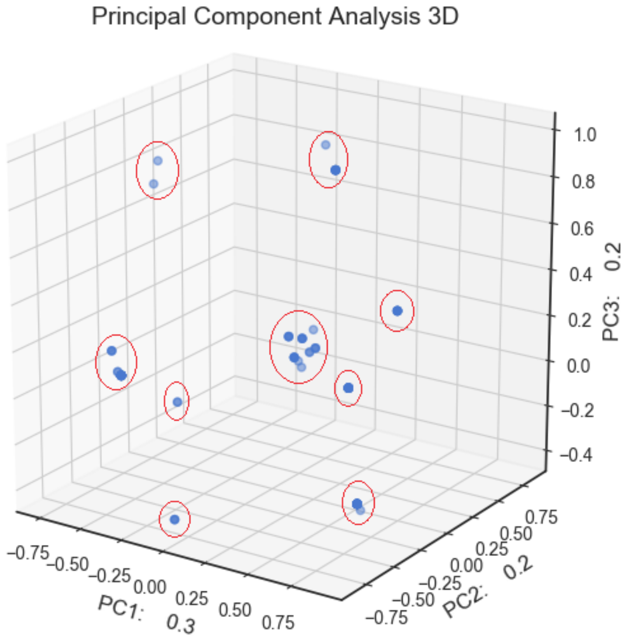


Fig. 18. 3D PCA scatter plot of sample A spectral contribution matrix.

The cumulative variance explained by 3 principal components is ~70%. In the 3D PCA scatterplot (Fig. 18), 9 groups can be specified. In comparison to the Dindex value (4 clusters) the more detailed separation of the scores can be observed. In order to specify the actual number of components in the data set, the Ward's HCA dendrogram was generated (Fig. 19).

© 2017 Tous droits réservés.

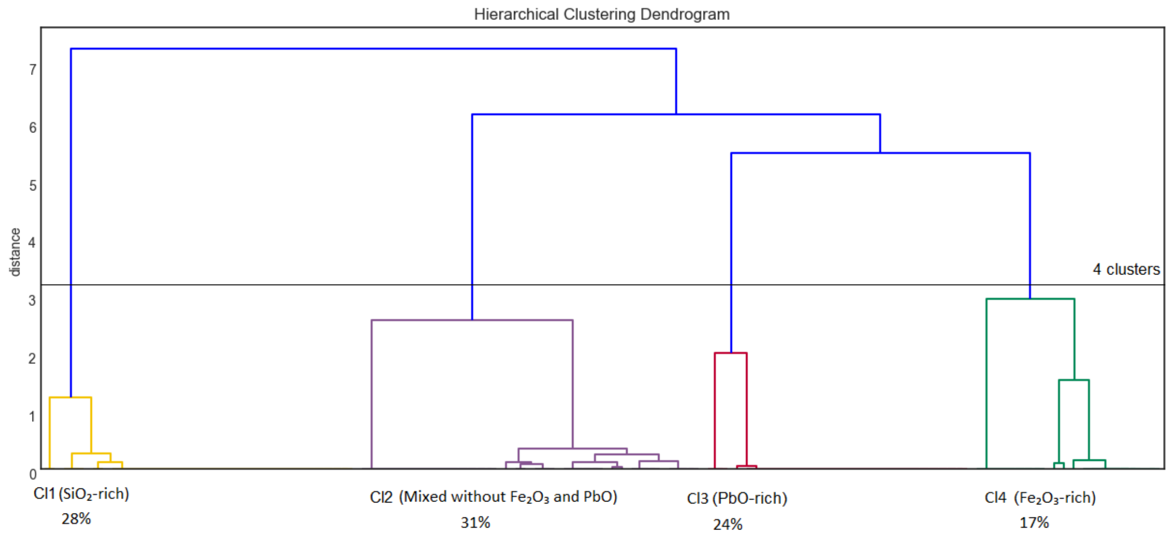


Fig. 19. Ward's HCA dendrogram with specified *cut-off* point of sample A contribution matrix.

The procedure of the Ward's HCA processing was described in detail in subsection 1.2.5., where the input matrix constitutes the PCA-reduced matrix generated by the algorithm to avoid operations on binary data. The number of 4 clusters was used for the data structuration. The *cut-off* point is corresponding to the Dindex value. The number of potential groups visible in the 3D scatter plot generated by PCA (Fig. 18) is not complementary with the data structure. The structuration based on the number of clusters specified by PCA causes an overestimation of the data where different clusters contain exactly the same particle types, what was tested manually during this procedure. In sample A ($PM_{10-2.5}$) cluster 1 contains SiO_2-rich particles and represents 28% of the particle population. Cluster 1 was placed in the separated node in the dendrogram (Fig. 19). Other clusters with a homogeneous particle composition are: cluster 3 (PbO-rich) and cluster 4 (Fe_2O_3-rich), which represent 24% and 17% of the particle population, respectively. The most abundant group of the particles is cluster 2 which contains mixed particles of compounds: M-Ox (metal oxide), SiO_2, TiO_2, ZnO, $CaSO_4$, M-S (metal sulphide), $FeSO_4$ and $KFe^{3+}(OH)_6(SO_4)_2$.

© 2017 Tous droits réservés.

This cluster is the only heterogeneous group of particles in sample A and represents 31% of the particle population. However, it should be noted that some homogeneous particles, which were not included in any other cluster, are a part of cluster 2.

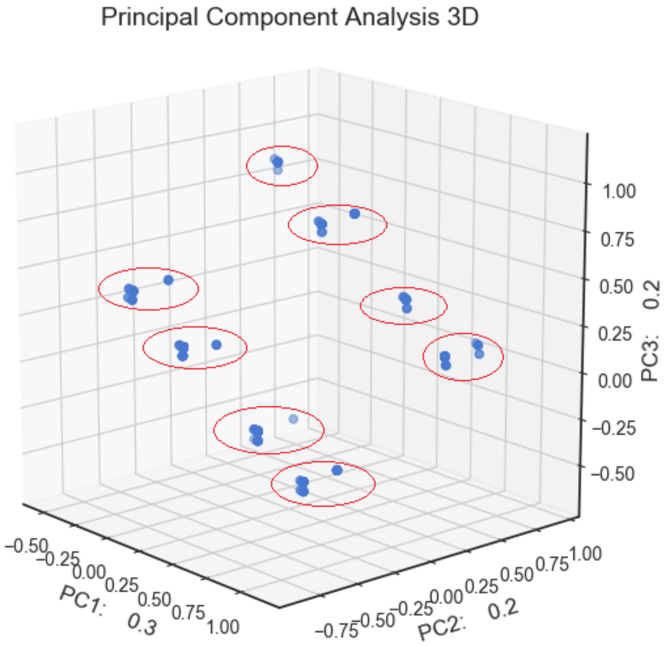


Fig. 20. 3D PCA scatter plot of sample B spectral contribution matrix.

The cumulative variance explained by 3 principal components is ~70% for the spectral contribution matrix of sample B. In the 3D PCA scatterplot (Fig. 20), 8 groups can be specified. This value is complementary to the Dindex value (8 clusters). To confirm the actual number of the clusters in the data set, the Ward's HCA dendrogram was generated (Fig. 21).

© 2017 Tous droits réservés.

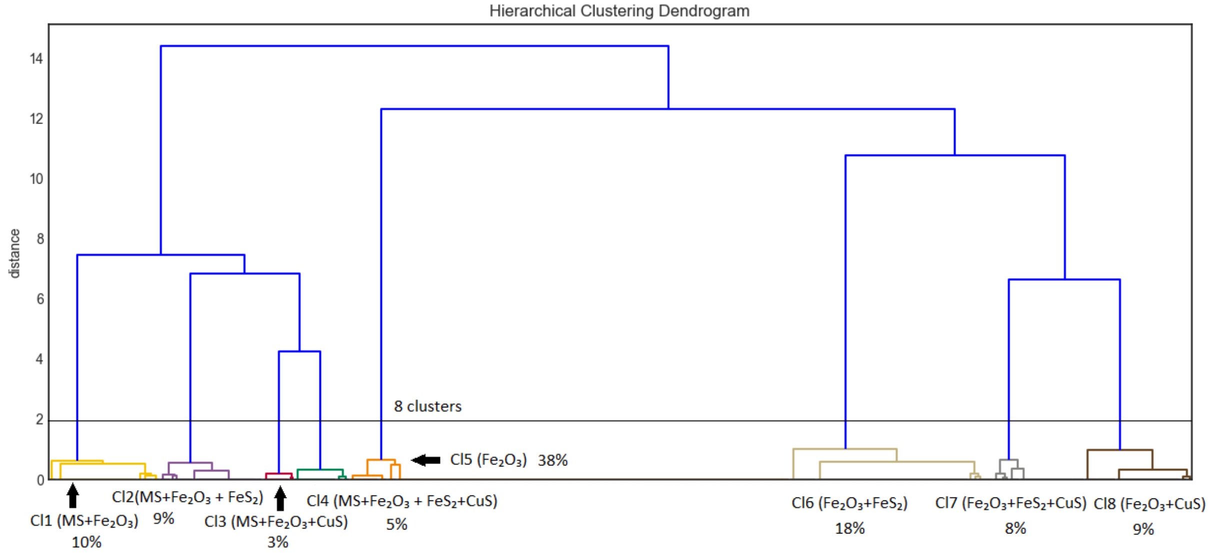


Fig. 21. Ward's dendrogram with specified *cut-off* point of sample B contribution matrix.

8 particle groups can be distinguished in the $PM_{2.5}$ fraction (sample B) with only three main clusters encountering for ≥10%. Each specified cluster contains Fe_2O_3. The major particle type corresponds to the single-compound Fe_2O_3 particles (38% of the population). The second particle type (encountered for 18%) is related to Fe_2O_3 and FeS_2. The next 3 groups (3rd, 4th and 5th) correspond to the particles containing Fe_2O_3, FeS_2 as well as MS or CuS accounting for 9%, 8% and 5%, respectively. The 5th group represents quaternary particles with all the previously mentioned compounds (Fe_2O_3, FeS_2, MS and CuS). The particles that contain both Fe_2O_3 and FeS_2 (as well as the other compounds such as MS, CuS) represent 40% of the sample population. The groups 6th and 8th contain particles with Fe_2O_3 and CuS, while the group 8th also contains MS. Na_2SO_4 and PbO spectra were extracted during MCR-ALS approach, nonetheless only by specification of all existing chemical mixings it is possible to estimate the number of particles containing these compounds (see more in 3.2.1. subsection).

© 2017 Tous droits réservés.

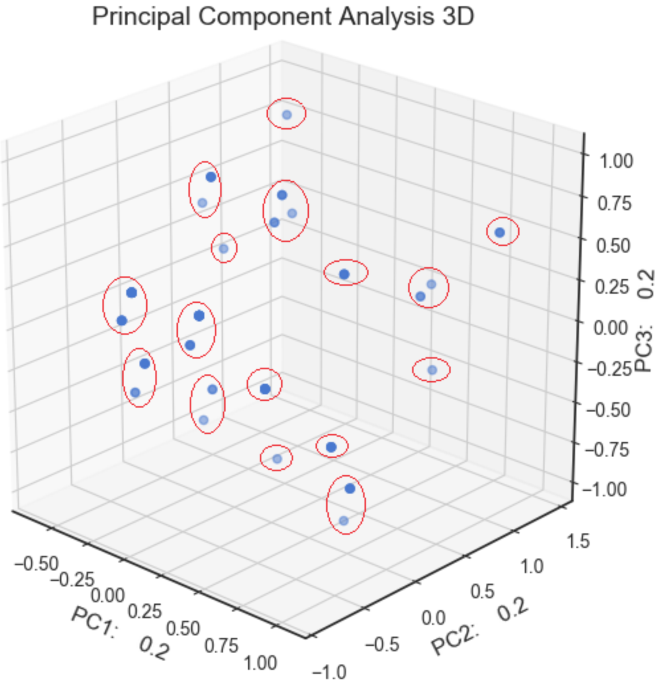


Fig. 22. 3D PCA scatter plot of sample C spectral contribution matrix.

The cumulative variance explained by 3 principal components is ~60%. In the 3D PCA scatterplot (Fig. 22), 16 groups can be specified. However, the designated number of groups is questionable. In comparison to the Dindex value (5 clusters) the detailed separation of the scores can be observed. On order to specify the actual number of components in the data set, the Ward's HCA dendrogram was generated (Fig. 23).

© 2017 Tous droits réservés.

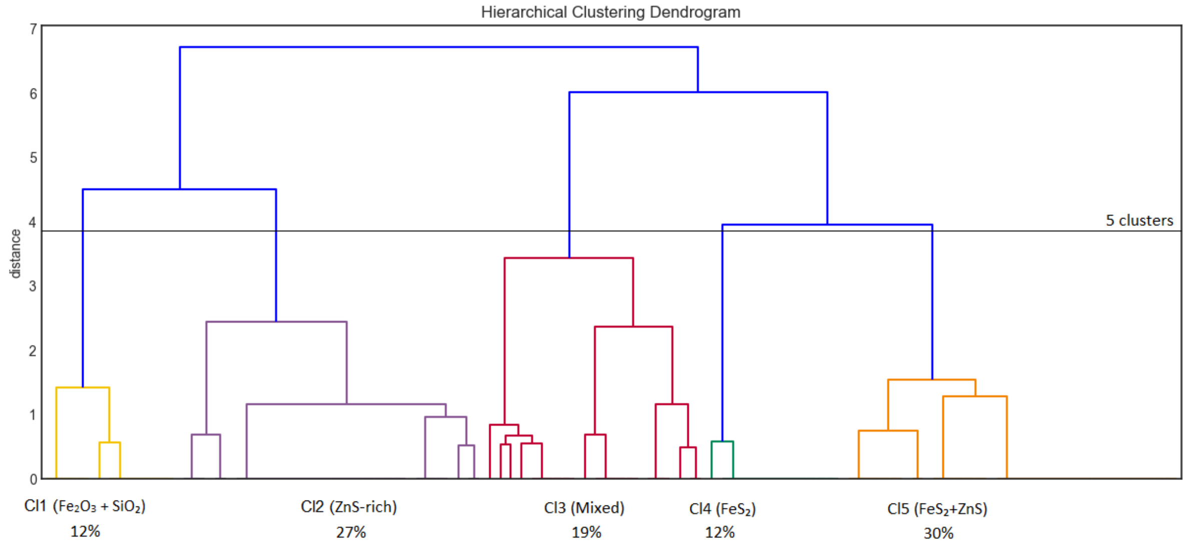


Fig. 23. Ward's dendrogram with specified *cut-off* point of sample C contribution matrix.

5 particle groups can be distinguished in the PM_1 fraction (sample C). The dominating compounds are FeS_2, ZnS and Fe_2O_3. It can therefore be assumed that sample C is rich in Fe-rich and S-rich compounds. The major particle type corresponds to the binary FeS_2 + ZnS particles (30% of the population) and constitutes cluster 5. The second particle type (encountered for 27%) is related to ZnS-rich particles (cluster 2). The next 3 groups (1st cluster, 3rd cluster and 4th cluster) correspond to the particles containing Fe_2O_3+ SiO_2 (12%), mixed particles (19%) andFeS_2 (12%), respectively.

© 2017 Tous droits réservés.

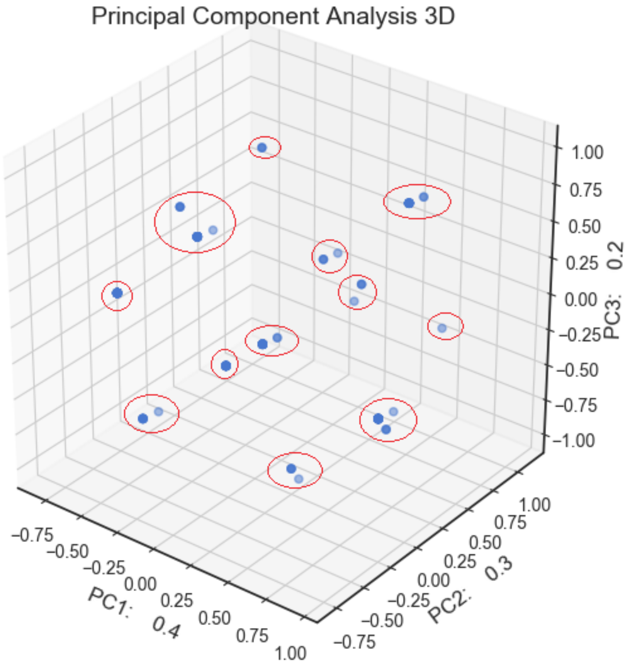


Fig. 24. 3D PCA scatter plot of sample D spectral contribution matrix.

The cumulative variance explained by 3 principal components is ~90% for sample D. In the 3D PCA scatterplot (Fig. 24), 12 groups can be specified. However, the designated number of groups is questionable. In comparison to the Dindex value (6 clusters) the detailed separation of the scores can be observed. In order to specify the actual number of components in the data set, the Ward's HCA dendrogram was generated (Fig. 25).

© 2017 Tous droits réservés.

lilliad.univ-lille.fr

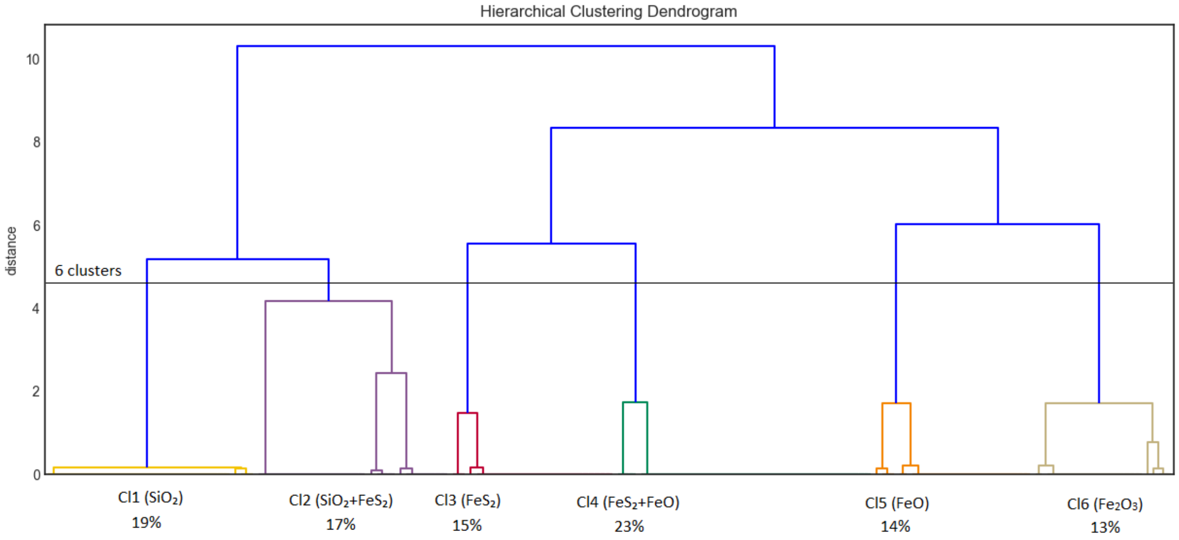


Fig. 25. Ward's dendrogram with specified *cut-off* point of sample D contribution matrix.

The structuration based on the number of clusters specified by PCA causes the overestimation of the data where different clusters contain exactly the same particle types, which was tested manually during this procedure. In sample D ($PM_{0.5}$) cluster 1 contains SiO_2-rich particles and represents 19% of the particle population. Cluster 1 was placed in the separated node with cluster 2 (SiO_2 + FeS_2 particles) in the dendrogram (Fig. 25). Cluster 2 represents 17% of the population. The clusters that have a homogeneous particle composition are: cluster 3 (FeS_2), cluster 5 (FeO) and cluster 6 (Fe_2O_3), which represent 15%, 14% and 13% of the particle population, respectively. The most abundant group of the particles is cluster 4, which contains FeS_2 + FeO particles and represents 23% of particle population. In general, sample D can be described as Fe-rich and SiO_2 rich particle fraction.

© 2017 Tous droits réservés.

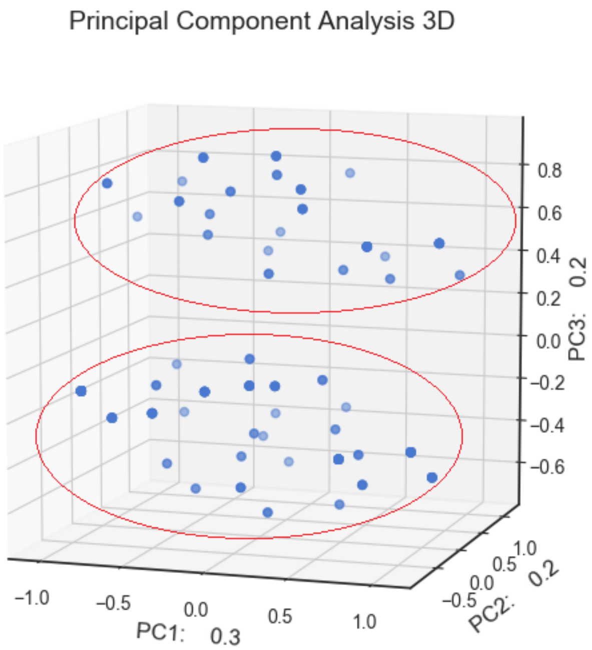


Fig. 26. 3D PCA scatter plot of sample FM spectral contribution matrix.

The cumulative variance explained by 3 principal components is ~70% for sample FM. In the 3D PCA scatterplot (Fig. 26), only 2 groups can be specified. The scores of the scatter plot (Fig. 26) are randomly distributed, which makes the matrix structuration difficult. In comparison to the Dindex value (6 clusters) the detailed separation of the scores can be observed. In order to specify the actual number of components in the data set, the Ward's HCA dendrogram was generated (Fig. 27).

© 2017 Tous droits réservés.

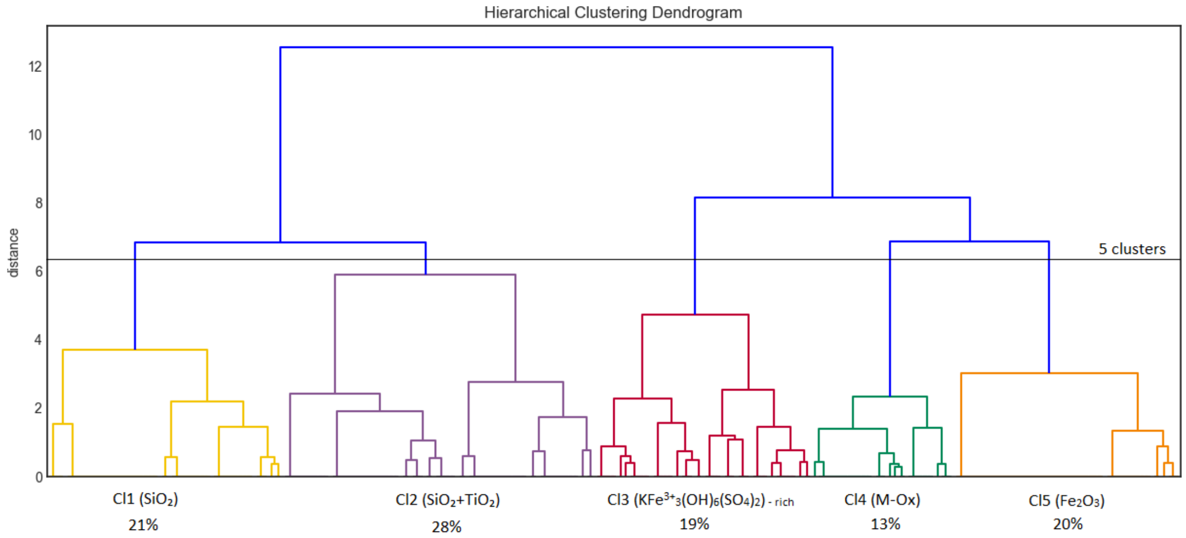


Fig. 27. Ward's dendrogram with specified *cut-off* point of sample FM contribution matrix.

5 particle groups can be distinguished in sample FM. This value is smaller than estimated by the Dindex, and it is due to the observation of the results from two Ward's HCA approaches. By specification of 6 clusters (calculated by Dindex), cluster 2 was separated into 2 individual groups which have exactly the same particle composition. However, it should be noted that such a situation was observed only for sample FM. The dominating compounds are Fe_2O_3, SiO_2 and $KFe^{3+}(OH)_6(SO_4)_2$. It can therefore be assumed that sample FM is rich in Fe and SiO_2 compounds. The major particle type corresponds to the binary SiO_2 + TiO_2 particles (28% of the population) and constitutes cluster 2. The second particle type (encountered for 21%) is related to SiO_2 particles (cluster 1). The next 3 groups (3rd cluster, 4th cluster and 5th cluster) correspond to the particles containing $KFe^{3+}(OH)_6(SO_4)_2$ (19%), metal oxides (13%) and Fe_2O_3 (20%) of particle population, respectively. The two groups, which can be distinguished in the 3D PCA, scatter plot (Fig. 26) are related to the SiO_2-rich particles (1st group) and the others (2nd group).

© 2017 Tous droits réservés.

3.2.2.4. Description of the mixing state of the particles

The number of homogeneous and heterogeneous particles i.e. composed of one or several species was determined for each fraction (Fig. 28). The highest number of homogeneous particles was identified in the $PM_{10-2.5}$ fraction. In sample A almost 80% of the particles are homogeneous. In turn, the richest in heterogeneous particles is sample FM, where heterogeneous particles constitute ~75% of the particle population.

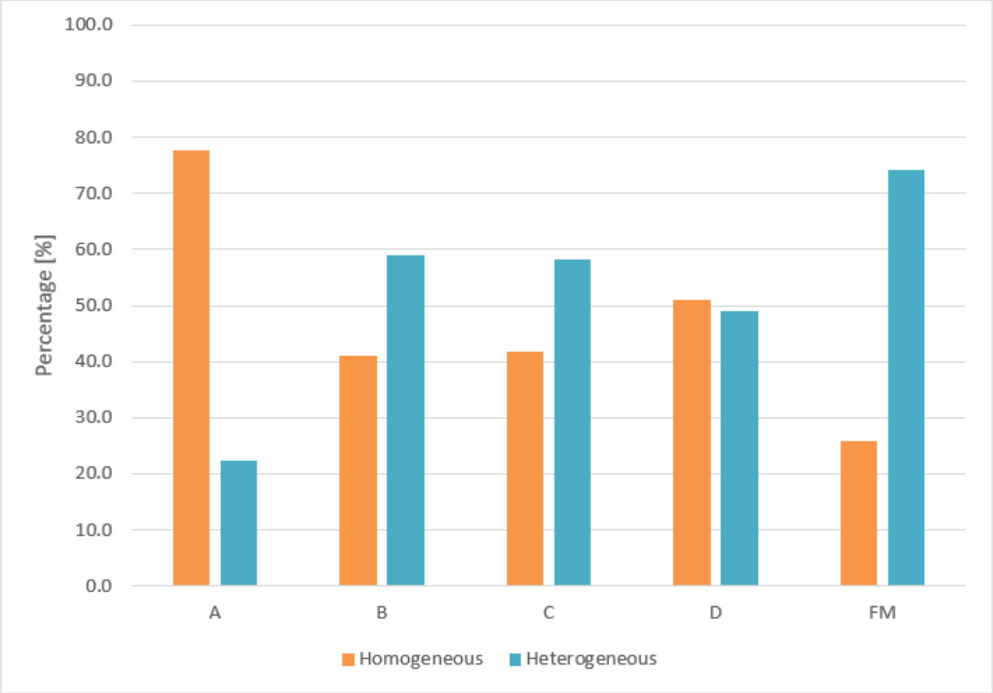


Fig. 28. Classification of particles for homogeneous and heterogeneous based on MR parameter.

The almost identical ratio of homogeneous to heterogeneous particles is observed for samples B and C. In these samples, the heterogeneous particles represent ~60% of the particle population. The small difference between these two classes of particles is observed in sample D. In the fraction $PM_{0.5}$, 51% of the particle population is represented by homogeneous particles where 49% by heterogeneous. These results are surprising since heterogeneous particles as aggregates would be expected in the coarsest fractions.

© 2017 Tous droits réservés.

The more specific mixing level i.e. mixing ratio (MR) is summarized in the Fig. 29, with MR1 related to single species, MR2 mixing of two compounds, MR3 mixing of 3 components and MR4 and MR5 mixing to 4 and 5 compounds, respectively

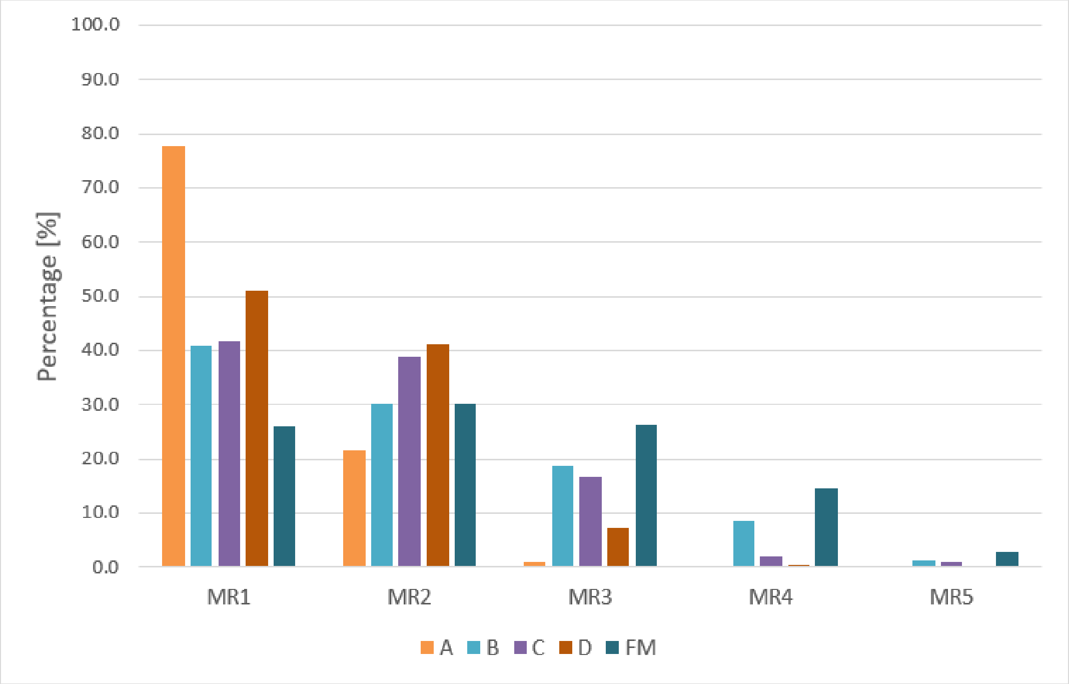


Fig. 29. Classification of particles based on MR parameter.

It can be observed in Fig. 29, that the amount of particles representing the level decreases as the level of mixing increases. The smallest chemical mixing variation was identified for sample A, were sample FM is just the opposite and contains the most heterogeneous particles from all of the fractions, as mentioned previously. In sample C the numbers of homogeneous and binary particles are almost the same, corresponding to ~40% of particle population. The ternary particles were found in the $PM_{2.5-1}$, $PM_{1-0.5}$, $PM_{0.5}$ and miner's filter samples. In sample D the lower level of chemical mixing is observed, comparing to the C and B samples. In the $PM_{0.5}$ no quaternary particles were found and almost 90% of the particle population can be described as homogeneous and binary particles. It can be explained by the size range of particles for which few aggregates can be observed. The classification of the

© 2017 Tous droits réservés.

particles by main particle groups was made (Fig. 30). The specification of the groups was performed based on the extracted compounds using the MCR approach. The SiO_2 group was formulated based only on the contribution from the SiO_2 spectrum. The Fe-rich group was formulated based on the compounds: Fe_2O_3, $FeSO_4$, FeS_2, Fe_3O_4, FeO and $KFe^{3+}(OH)_6(SO_4)_2$. In turn, metal-rich groups were designated by: M-Ox, TiO_2, ZnO, PbO, CuO, MS, CuO_$CaSO_4$ and ZnO_CuO_$CaSO_4$. It should be noted, that some compounds are assigned to 2 different groups occasionally. Nonetheless, this situation is dictated by the mixed spectra where "pure" (single) compounds were unable to be extracted by MCR approach. Furthermore, the S-rich particles contain $CaSO_4$, M-S, CuS, Na_2SO_4, MS, ZnS, CuO_$CaSO_4$ and ZnO_CuO_$CaSO_4$ compounds

© 2017 Tous droits réservés.

3.2.2.5. Chemical composition and evolution of the particles in polymetallique mining environment

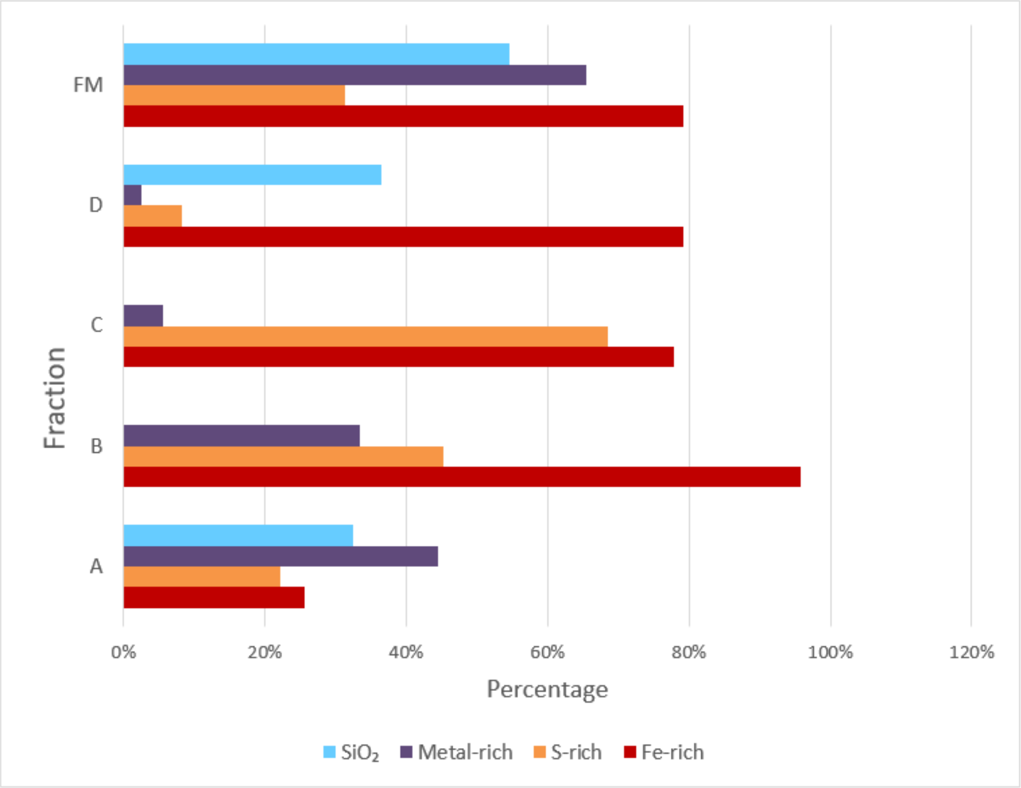


Fig. 30. Classification of particles based on the specified groups.

It can be seen in Fig. 30, that the Fe-rich particles dominate in samples B, C, D and FM, where they represent ~96%, ~78%, ~79% and ~79% of the particle population. The most homogeneous structure of the particles due to a contribution of the specified groups is characteristic for sample A. SiO_2 particles are located in the $PM_{10-2.5}$, $PM_{0.5}$ and miner's filter fractions. This type of particles was not found in samples B and C. The highest contribution of SiO_2 particles is observed in sample FM (~57%). The abundance of S-rich particles increase for samples A, B and C, starting from 22%-69%. The significant gap in the S-rich particles contribution can be observed in sample D, where in turn almost 38% particles containing SiO_2 can be observed. The highest amount of Metal-rich particles was identified in sample FM (68%).

© 2017 Tous droits réservés.

3.3. Conclusions

By applying the presented previously algorithm, it was possible to determine the particles chemical mixing level and molecular composition. Calculation of the Dindex was an appropriate approach to specify the number of main particle groups. The compounds typical for the mining environment activity were present in the particles from all of the fractions. The most homogeneous fraction, in terms of the presented chemical mixing, is the $PM_{10-2.5}$ fraction, where almost 80% of particles are homogeneous. The opposite situation is illustrated with sample FM (miner's filter) where relatively high heterogeneity was observed. The particles were classified by 4 main particle groups which was favourable for the specification of the changes in the composition of the particles via their size. The decreasing contribution of the Metal-rich particles can be observed along with the decreasing particle size. In turn, the increasing contribution of the S-rich particles was demonstrated for the $PM_{10-2.5}$, $PM_{2.5-1}$ and $PM_{1-0.5}$. These results confirm the usefulness of the presented algorithm to describe the chemical composition and chemical mixing of particles from different atmospheric aerosol fractions.

© 2017 Tous droits réservés.

CHAPTER 4: SPECTRONOMY:

A GRAPHICAL INTEGRATED SYSTEM FOR PROCESSING, ANALYZING, AND CLUSTERING OF RAMAN AND INFRARED SPECTRAL DATA SET

© 2017 Tous droits réservés.

In this chapter, we present an integrated, open-source software system for processing, analyzing, and clustering of Raman and FTIR spectral data sets. The Spectronomy graphical system was developed in collaboration with the Institute of Molecular Sciences (UMR CNRS 5255) from the University of Bordeaux.

Spectronomy is an application of several algorithms for spectral pre-processing (scale-based normalization, automated baseline correction, extended multiplicative signal correction, standard normal variate correction, Savitzky-Golay filtering), cluster analysis (k-means, HCA, fuzzy-C-means), unsupervised multivariate analysis (PCA) and optimal number of cluster calculation (D-index, Hubert's index) that are going to be described in detail in this chapter.

4.1. Software specifications and system architecture

The presenting system was built in the Python programming language on the Microsoft Windows operating system. The Spectronomy works in connection with the R language through an interface to benefit from optimal capabilities of the libraries of both languages. Python was chosen due to the open source license, which greatly facilitates the scientific community's active contribution to the development of the presented software. Despite the rich standard library, many specialized external packages have been used. Most of them are dedicated to scientific programming in which Python is particularly popular. Of particular note is the fact, that Python is easy to combine with other languages, such as Fortran, C++ or even MATLAB, which are widely used for scientific computations. However, the performance of interpreted languages, such as Python, for computation-intensive tasks are inferior in comparison to lower-level programming languages. Therefore, several external libraries were used in the Spectronomy to increase computational efficiency of the system. The external libraries such as *NumPy* and *SciPy* have been used for fast and vectorized operations on multidimensional arrays, where the Pandas library was applied for high-performance data structure analysis and

© 2017 Tous droits réservés.

manipulation. In addition, selected features of the integrated Python module for machine learning problems (Scikit-learn) was used (Pedregosa & Varoquaux 2011). The highest level-interface for visualization of spectra and graph plotting is provided by the Seaborn visualization library.

The system is implemented with a graphical user interface (GUI), shown in Fig. 31. Each menu item operates on a unique routine or toolset with integral subroutines for the data loaded in the system (e.g. FTIR dataset). The simple-to-use-graphical interface allows an effortless interaction with the features included in the system, with no need for any programming skills. The menu is designed for flexible feature operations with a very restricted data flow, designed only for subroutines (e.g. HCA). In that way, the trial and error method of solving the specific problem can be applied by casual users. Moreover, an instant observation of each step is available directly in the main window of the application, which allows to control each step of the spectra processing and analysis. The prime features are divided into separated groups in the menu by categories i.e. File, Factor Analysis, Clustering, Preprocessing, Matrix and Other. These features will be described in the following sections.

© 2017 Tous droits réservés.

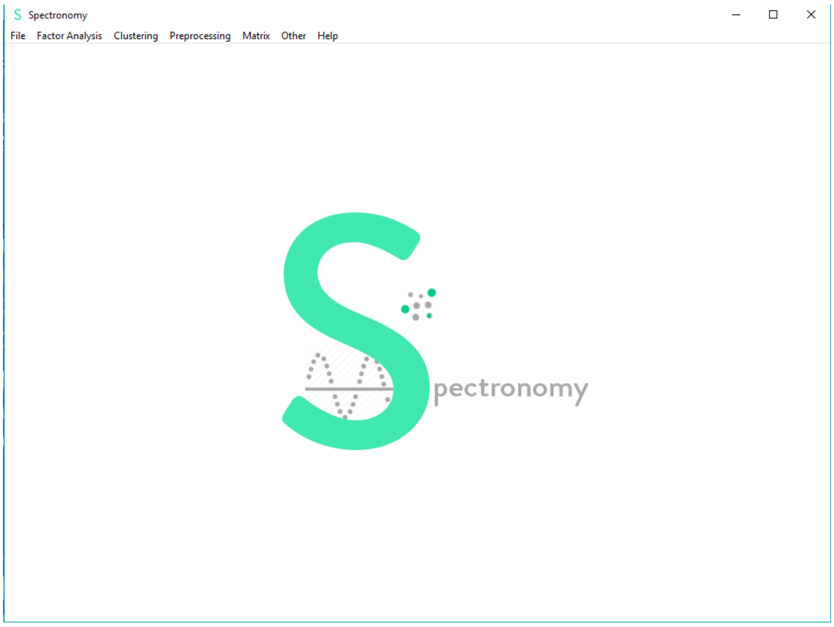

Spectronomy
File Factor Analysis Clustering Preprocessing Matrix Other Help
Spectronomy

Fig. 31. Welcome page for primary graphical user interface of the developed Spectronomy system.

4.2. Operating procedure

4.2.1. Data description, input and management

Data set description

The data set used for this software demonstration is formed of Mid-IR spectra, recorded on a FTIR spectrometer (Alpha FTIR Spectrometer from Bruker optic), equipped with a deuterated triglycine sulphate (DTGS) detector and a germanium beam splitter, interfaced to a computer with a Windows-based operating system, and connected to the software of an OPUS operating system (Version 7.0 Bruker optic). FTIR spectra were collected at the frequency regions of 6000–400 cm^{-1} by recording 10 scans with a resolution of 2 cm^{-1}. All spectra were corrected by subtracting a background of air recorded in the same conditions. 16 spectra were collected in order to create a data matrix with a dimension of 2747 × 16 (wavenumbers × number of spectrum).

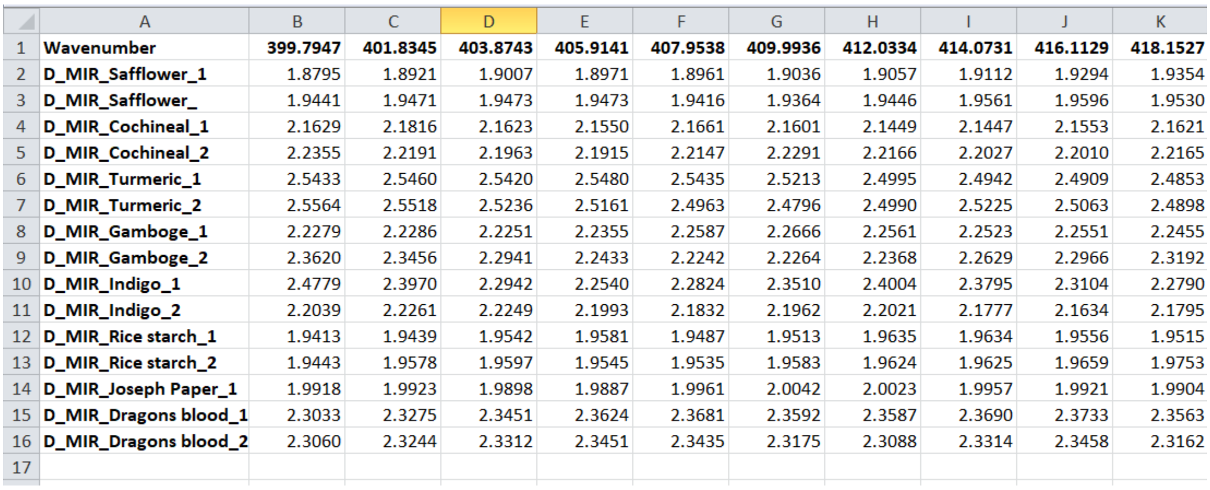
Data structure

The software is able to handle a 2-dimensional dataset, where each sample is defined

© 2017 Tous droits réservés.

by a series of discrete or continuous measurements. As in the most programming languages, data are stored in the single structure called a programming variable. The main data structure used in Spectronomy and stored as a programming variable is pandas' DataFrame. A DataFrame represents a tabular, spreadsheet-like data structure containing an ordered collection of columns, each of which can be a different value type – numeric, string, Boolean, etc. (Fig. 32). However, because of the purposes for which the program was designed (Raman and FTIR spectroscopy), its functionality is most of the time limited to operations on the numeric data (integers, float-point number and complex numbers), where the string type (sequence of characters) is dedicated for data labeling and the Boolean type – for control flow. It should be noted that the Spectronomy system creates and manipulates a large amount of programming variables during utilization, and they belong only to pandas' DataFrame structure. The variables contain multiple parameters of the system such as the information necessary for the exploratory analysis and clustering models, intermediate data, equation results, etc. The software is also equipped with a data serializing and deserializing system for saving the arbitrary data and then sending them to other processes to improve the software efficiency. The access to the program features is available through the graphical user interface menu. By clicking on the menu item, an associated routine or entire subroutine toolset is executed on a copy of the main data structure.

© 2017 Tous droits réservés.



	A	B	C	D	E	F	G	H	I	J	K
1	**Wavenumber**	**399.7947**	**401.8345**	**403.8743**	**405.9141**	**407.9538**	**409.9936**	**412.0334**	**414.0731**	**416.1129**	**418.1527**
2	**D_MIR_Safflower_1**	1.8795	1.8921	1.9007	1.8971	1.8961	1.9036	1.9057	1.9112	1.9294	1.9354
3	**D_MIR_Safflower_**	1.9441	1.9471	1.9473	1.9473	1.9416	1.9364	1.9446	1.9561	1.9596	1.9530
4	**D_MIR_Cochineal_1**	2.1629	2.1816	2.1623	2.1550	2.1661	2.1601	2.1449	2.1447	2.1553	2.1621
5	**D_MIR_Cochineal_2**	2.2355	2.2191	2.1963	2.1915	2.2147	2.2291	2.2166	2.2027	2.2010	2.2165
6	**D_MIR_Turmeric_1**	2.5433	2.5460	2.5420	2.5480	2.5435	2.5213	2.4995	2.4942	2.4909	2.4853
7	**D_MIR_Turmeric_2**	2.5564	2.5518	2.5236	2.5161	2.4963	2.4796	2.4990	2.5225	2.5063	2.4898
8	**D_MIR_Gamboge_1**	2.2279	2.2286	2.2251	2.2355	2.2587	2.2666	2.2561	2.2523	2.2551	2.2455
9	**D_MIR_Gamboge_2**	2.3620	2.3456	2.2941	2.2433	2.2242	2.2264	2.2368	2.2629	2.2966	2.3192
10	**D_MIR_Indigo_1**	2.4779	2.3970	2.2942	2.2540	2.2824	2.3510	2.4004	2.3795	2.3104	2.2790
11	**D_MIR_Indigo_2**	2.2039	2.2261	2.2249	2.1993	2.1832	2.1962	2.2021	2.1777	2.1634	2.1795
12	**D_MIR_Rice starch_1**	1.9413	1.9439	1.9542	1.9581	1.9487	1.9513	1.9635	1.9634	1.9556	1.9515
13	**D_MIR_Rice starch_2**	1.9443	1.9578	1.9597	1.9545	1.9535	1.9583	1.9624	1.9625	1.9659	1.9753
14	**D_MIR_Joseph Paper_1**	1.9918	1.9923	1.9898	1.9887	1.9961	2.0042	2.0023	1.9957	1.9921	1.9904
15	**D_MIR_Dragons blood_1**	2.3033	2.3275	2.3451	2.3624	2.3681	2.3592	2.3587	2.3690	2.3733	2.3563
16	**D_MIR_Dragons blood_2**	2.3060	2.3244	2.3312	2.3451	2.3435	2.3175	2.3088	2.3314	2.3458	2.3162
17											

Fig. 32. Print-screen of data organization in the input file (spectral data matrix) for Spectronomy software in the form of table.

Data input and management

The Spectronomy system is based on the assumption that a whole spectral data set has been previously collected and calibrated for the typical wavenumber and intensity variations caused by a spectrometer. Therefore, the spectral data need to be placed in a Microsoft Excel file (.xlsx file extension) in a form of column mode profiles with instrumental responses (Jaumot et al. 2015). In other words, the spectra are stored in rows (each spectrum in a row), where in turn the wavenumbers are stored in columns. The desirable practice is to place the x-axis into the first row of the data matrix, which will help avoid potential identification errors. The Spectronomy system has an ability to load two types of data matrices. The first one is a two-dimensional data matrix with single or multiple spectra. The second type is a row mode profiles matrix (Fig. 34), e.g. C matrix from Multivariate Curve Resolution (Jaumot et al. 2015) (see Fig. 33)

© 2017 Tous droits réservés.

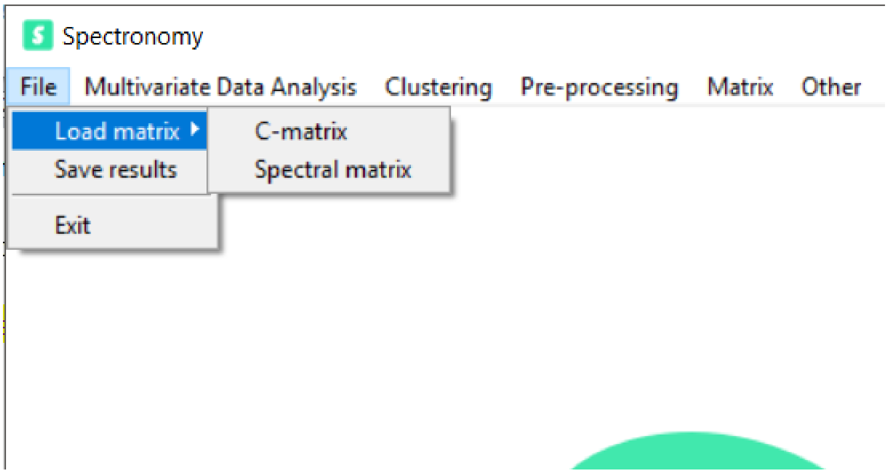
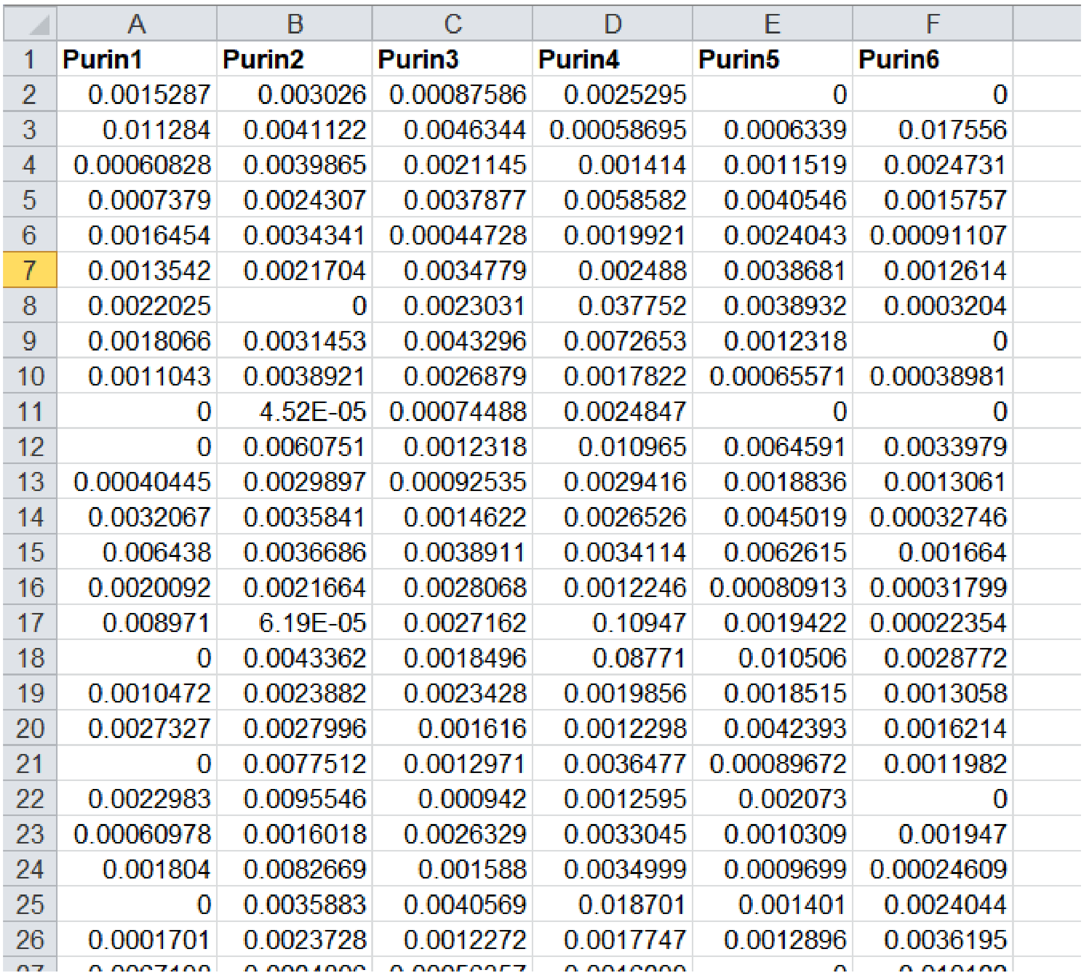


Fig. 33. Load function of two different matrices – spectral matrix and C-matrix.

Once the data are loaded, the system will ask about existing headers of the data matrix. The headers of the data frame are stored in the separate programming variables to avoid any errors during the analysis. Thus, exported headers can be used as a label of spectra for, e.g. hierarchical cluster analysis. One of the key applications of Raman and FTIR spectroscopy is to collect data from various sets or groups of samples to determine the similarities or differences among them. By specification of the row header (first column in the Excel file), the proper name of spectra will be used during the analysis. However, if a user does not specify the headers, then the program will automatically assign the numbers of each spectrum based on the order in which they are placed in the matrix. In addition, the system creates a copy of the loaded file in the temporary folder on the computer, which avoids a potential problem with a file overwrite.

© 2017 Tous droits réservés.



	A	B	C	D	E	F
1	Purin1	Purin2	Purin3	Purin4	Purin5	Purin6
2	0.0015287	0.003026	0.00087586	0.0025295	0	0
3	0.011284	0.0041122	0.0046344	0.00058695	0.0006339	0.017556
4	0.00060828	0.0039865	0.0021145	0.001414	0.0011519	0.0024731
5	0.0007379	0.0024307	0.0037877	0.0058582	0.0040546	0.0015757
6	0.0016454	0.0034341	0.00044728	0.0019921	0.0024043	0.00091107
7	0.0013542	0.0021704	0.0034779	0.002488	0.0038681	0.0012614
8	0.0022025	0	0.0023031	0.037752	0.0038932	0.0003204
9	0.0018066	0.0031453	0.0043296	0.0072653	0.0012318	0
10	0.0011043	0.0038921	0.0026879	0.0017822	0.00065571	0.00038981
11	0	4.52E-05	0.00074488	0.0024847	0	0
12	0	0.0060751	0.0012318	0.010965	0.0064591	0.0033979
13	0.00040445	0.0029897	0.00092535	0.0029416	0.0018836	0.0013061
14	0.0032067	0.0035841	0.0014622	0.0026526	0.0045019	0.00032746
15	0.006438	0.0036686	0.0038911	0.0034114	0.0062615	0.001664
16	0.0020092	0.0021664	0.0028068	0.0012246	0.00080913	0.00031799
17	0.008971	6.19E-05	0.0027162	0.10947	0.0019422	0.00022354
18	0	0.0043362	0.0018496	0.08771	0.010506	0.0028772
19	0.0010472	0.0023882	0.0023428	0.0019856	0.0018515	0.0013058
20	0.0027327	0.0027996	0.001616	0.0012298	0.0042393	0.0016214
21	0	0.0077512	0.0012971	0.0036477	0.00089672	0.0011982
22	0.0022983	0.0095546	0.000942	0.0012595	0.002073	0
23	0.00060978	0.0016018	0.0026329	0.0033045	0.0010309	0.001947
24	0.001804	0.0082669	0.001588	0.0034999	0.0009699	0.00024609
25	0	0.0035883	0.0040569	0.018701	0.001401	0.0024044
26	0.0001701	0.0023728	0.0012272	0.0017747	0.0012896	0.0036195

Fig. 34. print-screen of data organization in the input file (C-matrix) for Spectronomy software in the form of table.

4.2.2. Data visualization

The system has a variety of features to enable the data visualization. Firstly, a specific function is dedicated to displaying a loaded matrix (Fig. 35). Regardless the matrix type (column or row mode profiles) the system will automatically recognize it by a dimension. In the case of a spectral matrix, the number of rows should be lower than the number of columns, where in the case of C-matrix the situation is opposite. The data matrix structure should be properly defined for the purpose of an appropriate graph projection. The column mode profile matrix generates an individual XY graph with the imposed spectra, where the row mode profile matrix generates a stacked bar chart. The application of the data plotting function was implemented to examine

© 2017 Tous droits réservés.

a compliance of the loaded matrix with the software requirements, by giving a preview of the data before further analysis. Secondly, the system is able to plot the specific type of graph depending on a function (e.g. PCA biplot, HCA dendrogram, etc.), with various parameters provided during the analysis. Such an example is the explained variance value of each principal component in PCA, which provides a visual representation of the model. In addition, the visual inspection of the generated graphs is supplemented with an interactive navigation toolbar for a simple-to use manipulation of the image and a cursor positioning. The features included in the toolbar are practical for a detailed examination (e.g. zooming, elongating, determination of the peak position, etc.) of the graphs. In addition, for differentiation of the qualitative data (e.g. clusters in the dendrogram) the color palette based on Kelly's work (Christie et al. 2007) was implemented to obtain the best possible visualization of the data.

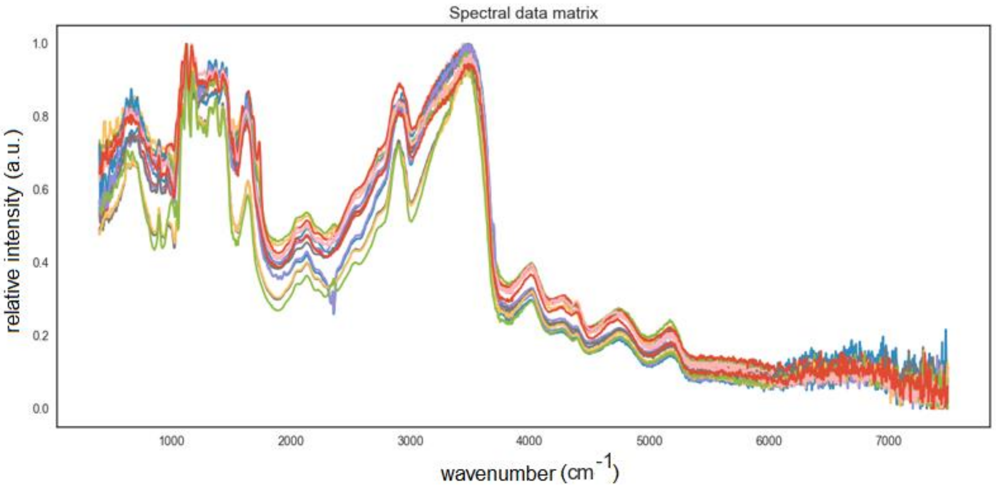


Fig. 35. Data visualization feature applied for FTIR spectra data set.

4.2.3. Spectral pre-processing

Pre-processing of a spectral data set has become an integral part of chemometrics (Roussel et al. 2014). The objective of the pre-processing is to improve the subsequent multivariate analysis. Several widely used pre-processing algorithms were implemented in the Spectronomy software (Fig. 36 with the different pre-processing

© 2017 Tous droits réservés.

choices visible).

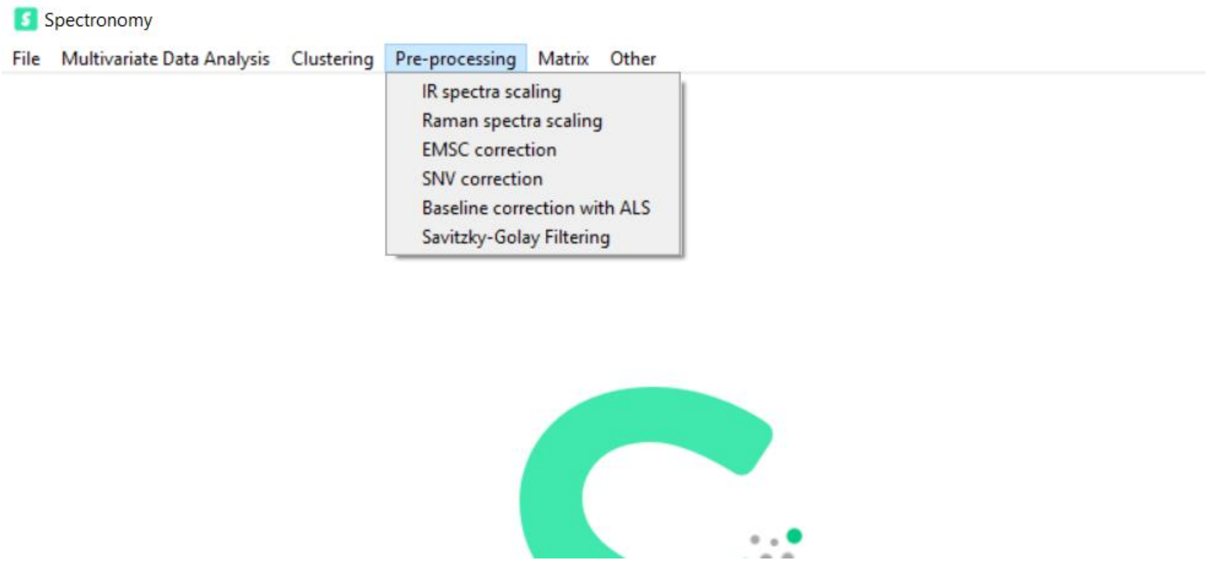


Fig. 36. Pre-processing functions included in the Spectronomy software.

The automatic baseline correction with asymmetric least squares (Eilers & Boelens 2005) was implemented to perform fast, simple and effective background subtraction. The algorithm elaborated by Eilers and Boelens (Eilers & Boelens 2005) was selected due to its uncomplicated adjustment of parameters for obtaining a satisfactory approximation to a real baseline. Another feature, which combines the advantages of spectra normalization and baseline correction is an extended multiplicative signal correction (EMSC). EMSC allows to separate and quantify the different types of chemical and physical variations in the spectra (Afseth & Kohler 2012). EMSC has particular application in the field of FTIR spectroscopy, owing to an effortless and flexible parameter optimization for scattering effect correction (Afseth & Kohler 2012). EMSC is useful in the correction of additive baseline effects, multiplicative scaling effects and interference effects (Afseth & Kohler 2012). Standard normal variate (SNV) pre-processing is probably one of the most popular method for scatter correction of Near-IR spectra (Rinnan et al. 2009). This algorithm included in the Spectronomy system was elaborated for reducing spectral noise and

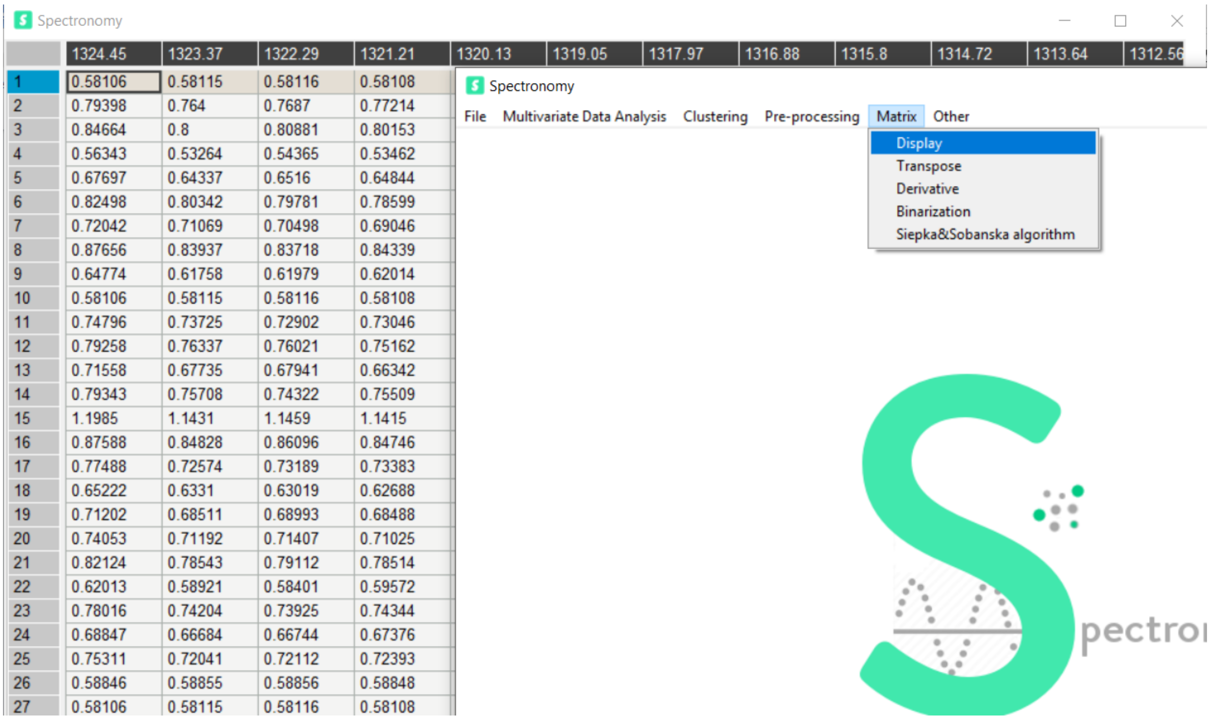
© 2017 Tous droits réservés.

eliminating background effects mainly for the Near-IR spectroscopy data. The digital filter for smoothing and differentiation based on the Savitzky-Golay algorithm (Savitzky & Golay 1964) was implemented and represents the last pre-processing feature included in the system. Basically, this smoothing algorithm consists of an elimination of the noise from the signal with the lowest possible signal distortion. In the described software, the Savitzky-Golay filtering function is also assembled with a spectral derivative estimation. The normalization of spectra by an intensity scaling (Randolph 2006) is the most favorable for Raman and FTIR spectra with different intensity values. By application of an intensity scaling procedure, such a disparity is compensated across the spectra under the same experimental conditions, where at the same time, the algorithm preserves the relative intensities of peaks within each spectrum. The spectra with normalized intensities can provide better performance with some algorithms, e.g. hierarchical cluster analysis (HCA).

In the presented software, a function for a matrix display is also included (Fig. 37). This feature is favorable for data monitoring after each step of the processing.

In addition, three practical operations on the data matrix are accessible. These features include: matrix transposition, binarization and displaying (Fig. 37). A substantial part of the described system is an easy-to use, well-developed pattern recognition module based on Scikit-learn Python package (Pedregosa & Varoquaux 2011).

© 2017 Tous droits réservés.



	1324.45	1323.37	1322.29	1321.21	1320.13	1319.05	1317.97	1316.88	1315.8	1314.72	1313.64	1312.56
1	0.58106	0.58115	0.58116	0.58108								
2	0.79398	0.764	0.7687	0.77214								
3	0.84664	0.8	0.80881	0.80153								
4	0.56343	0.53264	0.54365	0.53462								
5	0.67697	0.64337	0.6516	0.64844								
6	0.82498	0.80342	0.79781	0.78599								
7	0.72042	0.71069	0.70498	0.69046								
8	0.87656	0.83937	0.83718	0.84339								
9	0.64774	0.61758	0.61979	0.62014								
10	0.58106	0.58115	0.58116	0.58108								
11	0.74796	0.73725	0.72902	0.73046								
12	0.79258	0.76337	0.76021	0.75162								
13	0.71558	0.67735	0.67941	0.66342								
14	0.79343	0.75708	0.74322	0.75509								
15	1.1985	1.1431	1.1459	1.1415								
16	0.87588	0.84828	0.86096	0.84746								
17	0.77488	0.72574	0.73189	0.73383								
18	0.65222	0.6331	0.63019	0.62688								
19	0.71202	0.68511	0.68993	0.68488								
20	0.74053	0.71192	0.71407	0.71025								
21	0.82124	0.78543	0.79112	0.78514								
22	0.62013	0.58921	0.58401	0.59572								
23	0.78016	0.74204	0.73925	0.74344								
24	0.68847	0.66684	0.66744	0.67376								
25	0.75311	0.72041	0.72112	0.72393								
26	0.58846	0.58855	0.58856	0.58848								
27	0.58106	0.58115	0.58116	0.58108								

Fig. 37. Print-screen of matrix manipulation features included in the Spectronomy software.

4.2.4. Clustering

Description of methods

The most common method of an unsupervised pattern recognition is a cluster analysis (CA), widely used to describe the relationship within the dataset (Forina et al. 2008). With Spectronomy, three types of clustering are possible i.e. Hierarchical Cluster Analysis, k-means and Fuzzy Clustering (Fig. 38). The k-means algorithm is one of the most commonly used partitional clustering methods in chemometrics and usually efficient for handling large datasets (Butler et al. 2016). In turn, the HCA clustering is an effective way of presenting the hierarchy of data in a readable graphical form called a dendrogram. Finally, by fuzzy clustering algorithm each data point can belong to more than one cluster, which can favors an association of samples within specified clusters.

© 2017 Tous droits réservés.

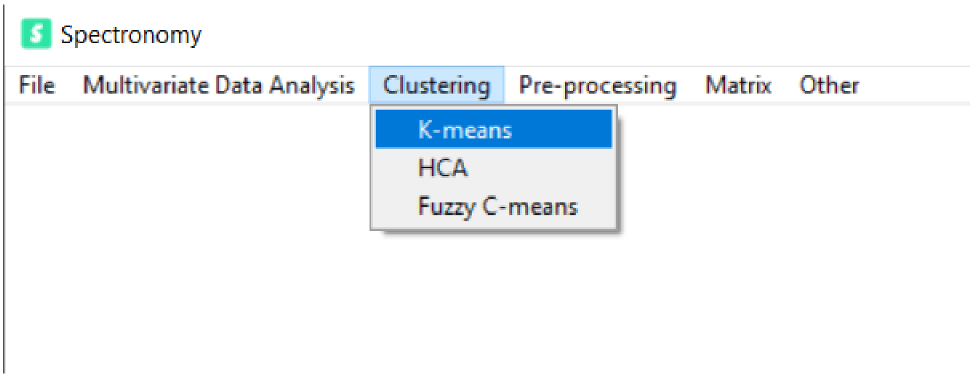


Fig. 38. Clustering functions included in the Spectronomy software.

Hierarchical cluster analysis (HCA) is an unsupervised pattern recognition technique that determines a grouping structure in a dataset by a nested tree graphical representation, called a dendrogram. In a dendrogram, the spectra are gradually associated according to their similarities. To build a dendrogram based on an agglomerative hierarchical cluster algorithm, at first each spectrum is considered as a cluster. At each agglomerative step, two clusters are merged with respect to a cluster distance until a final cluster is obtained (de Souza Lins Borba et al. 2015). The Spectronomy software includes several hierarchical clustering algorithms for computing a distance between clusters, such as the minimum variance algorithm (*Ward*), weighted center of mass distance (*median*), centroid distance (*centroid*), weighted average distance (*weighted*), shortest distance algorithm (*single*), unweighted average distance (*average*) and the furthest distance algorithm (*complete*) (Everitt et al. 2011). The elaborated system for a dendrogram projection was applied. Beyond a projection of a classical dendrogram, Spectronomy can generate a limited-type cluster-tree (a truncated dendrogram) based on the denotation of a nodes' number of from a user. Furthermore, a cut-off point for a dendrogram is easy to specify based on the specification of a distance value by a user. The labeling of samples is also possible, if the proper text values were included in the loaded data matrix (see section 2.2.1.) (Fig. 39).

© 2017 Tous droits réservés.

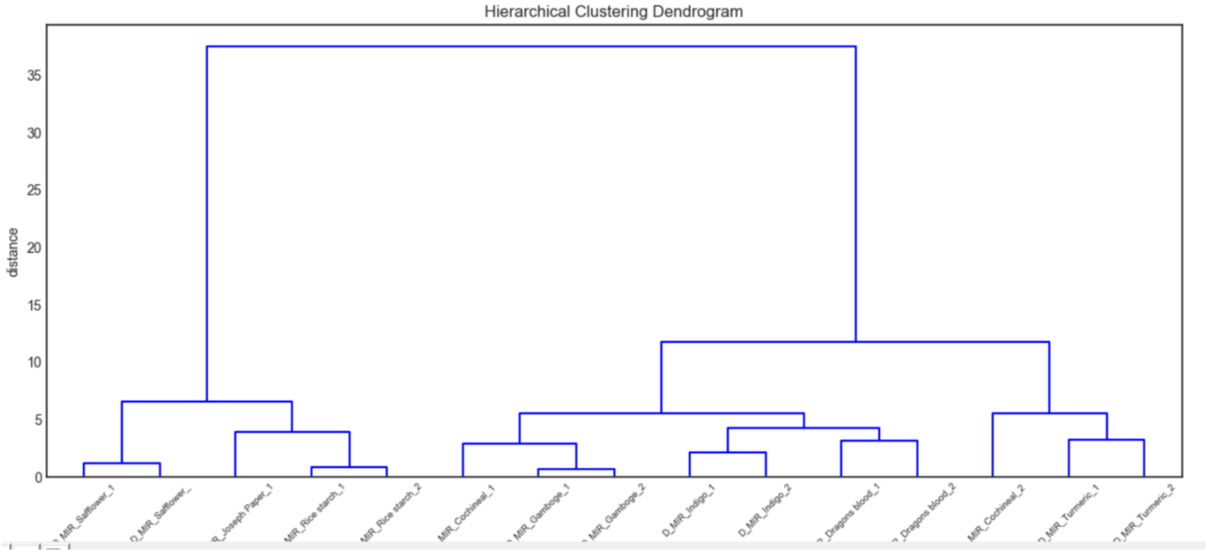


Fig. 39. Dendrogram generated by the Ward's HCA function with sample labels.

Another implementation is the k-means clustering, which represents a partitional clustering method. Partitioning clustering algorithms have widely been applied because of its effectiveness and applicability for the large data sets (Reddy & Jana 2012). In the k-means clustering a membership of each spectrum is initially assigned randomly to an a priori specified number of clusters. Then, a centroid of each of these clusters is calculated, where for each spectrum, a distance to previously specified cluster centroids is determined. If a spectrum is not associated with the closest centroid, it will be transferred into the closest cluster. The centroid positions are recalculated every time the spectrum has changed its membership. In general, the k-means clustering represents a better calculation performance beside the HCA, but a number of clusters need to be specified before the clustering procedure. In the described software, the results can be exported in the form of the Voronoi diagram, cluster membership vector and centroid position vector (Fig. 40).

© 2017 Tous droits réservés.

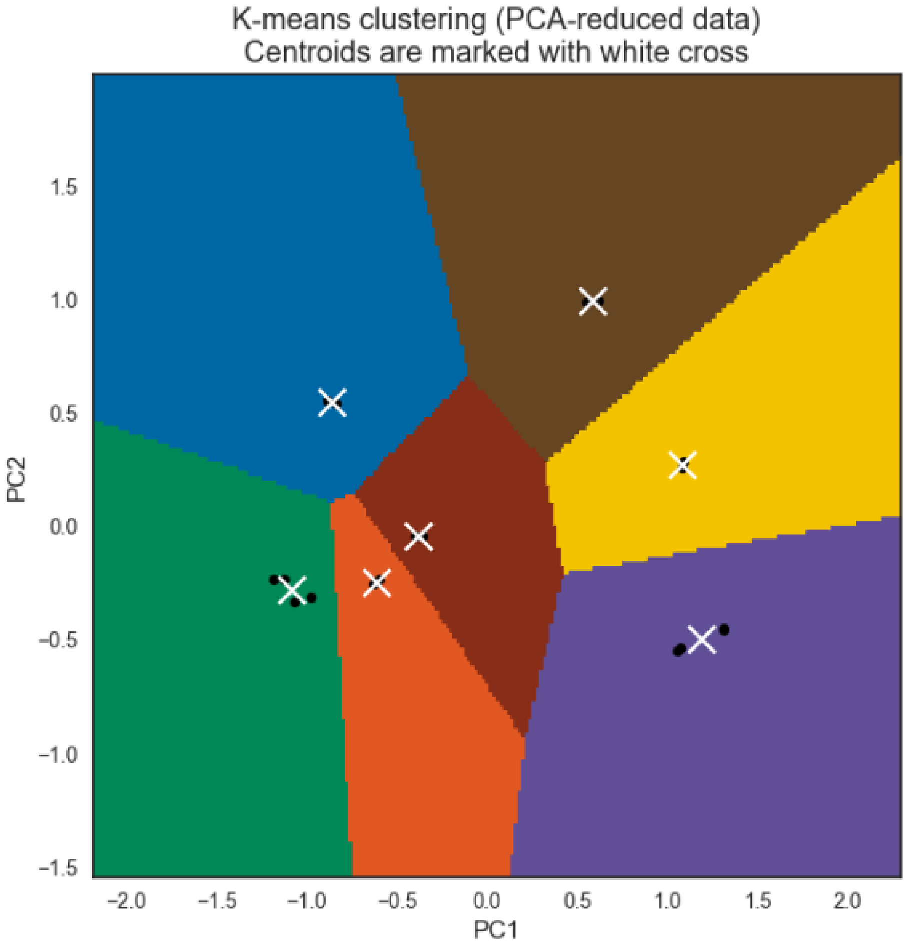


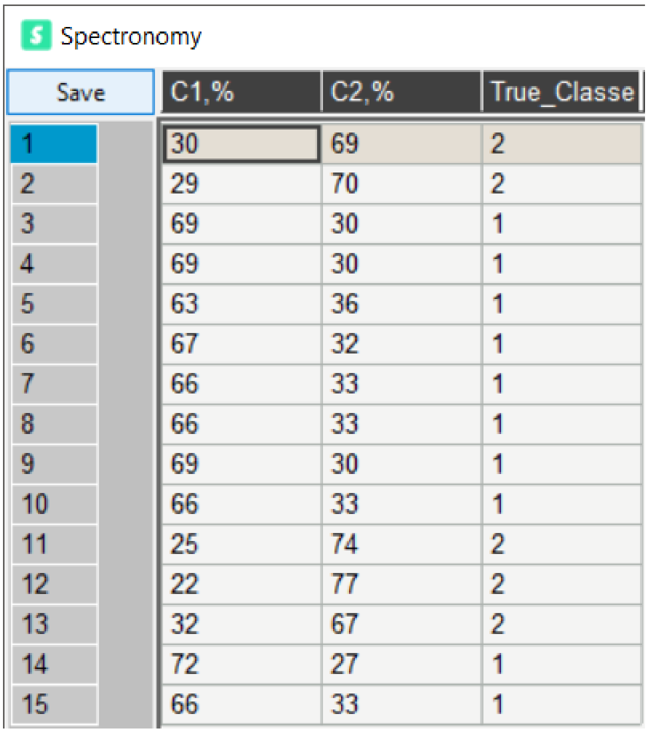
Fig. 40. Voronoi diagram generated by application of k-means clustering algorithm on PCA-reduced data with marked centroids.

The Voronoi diagram of the points is calculated using the initially calculated centroids. Each segment in the Voronoi diagram is a separate cluster. The centroids are updated to the mean of each segment. The algorithm is then repeated until a stopping criterion is fulfilled. Usually, the algorithm stops when a relative decrease in the objective function between iterations is less than a given tolerance value. This is not the case in this implementation: the iterations stop when centroids move less than the tolerance. The Voronoi diagram visualizes the k-means partitioning with the borders of each cluster positioned on the PCA score plot plane (see section 2.5.). This type of a graph simplifies an identification of the cluster centroid distances, as well as the spectra, which lie between clusters.

© 2017 Tous droits réservés.

The last, unsupervised pattern recognition method is fuzzy-C-means clustering. Fuzzy clustering was originally developed in 1969 by Ruspini based on a fuzzy set theory (Ruspini 1969). One of the major differences between fuzzy clustering and hard clustering (e.g. k-means, HCA) is that fuzzy clustering allows each pattern to belong to more than one cluster with varying degrees of certainty, based on their distance to the cluster centers. The fuzzy C-means algorithm is one of the most popular fuzzy clustering algorithms. It was first developed by Dunn in 1973 (Dunn 1973) and was subsequently improved by Bezdek research group (Khalilia et al. 2014). The Spectronomy system is equipped with a special function dedicated to specify the best value for fuzzy-C-means partitioning. The fuzzy partitioning coefficient (FPC) is defined in the range from 0 to 1. It is a metric, which determines clustering model conformity. When the FPC is maximized, then the data are described in the best way by fuzzy-C-means for the corresponding number of clusters. The Spectronomy system has a feature to export the fuzzy-C-means results in the form of an Excel spreadsheet file (Fig. 41). In each column, the values in percentage correspond to the cluster affiliation of each spectrum.

© 2017 Tous droits réservés.



	C1,%	C2,%	True_Classe
1	30	69	2
2	29	70	2
3	69	30	1
4	69	30	1
5	63	36	1
6	67	32	1
7	66	33	1
8	66	33	1
9	69	30	1
10	66	33	1
11	25	74	2
12	22	77	2
13	32	67	2
14	72	27	1
15	66	33	1

Fig. 41. Print-screen of results generated by fuzzy-c-means algorithm with a "save to file" button in the form of table.

Optimal number of cluster calculation

Various clustering algorithms commonly generate a distinctive set of groups. Even for the same algorithm, the modification of parameters or the data description order can significantly affect the final clustering results. Thus, an effective evaluation is crucial to provide relevant information about internal structures, which occur in the data. The Spectronomy system is equipped with two main features for a calculation of an optimal number of clusters for each data matrix. The Dindex is based on the clustering gain on intra-cluster inertia, which measures a degree of homogeneity between the data associated with a cluster (Charrad et al. 2014). It calculates their distances compared to a reference point representing a cluster profile. In turn, the Hubert's index is a point serial correlation coefficient between two matrices. High values of the normalized Hubert's index indicate an existence of compact clusters (Charrad et al. 2014). Both, the Dindex and Hubert's indexes are delivered and visualized in the form of an XY plot by NbClust package (Charrad et al. 2014). In

© 2017 Tous droits réservés.

these plots, the Dindex and the normalized Hubert's index versus the number of clusters are presented. By designation of a distinctive knee, an optimal number of clusters can be specified. Both, the Dindex and Hubert's indexes can be applied for k-means clustering and HCA (Fig. 42).

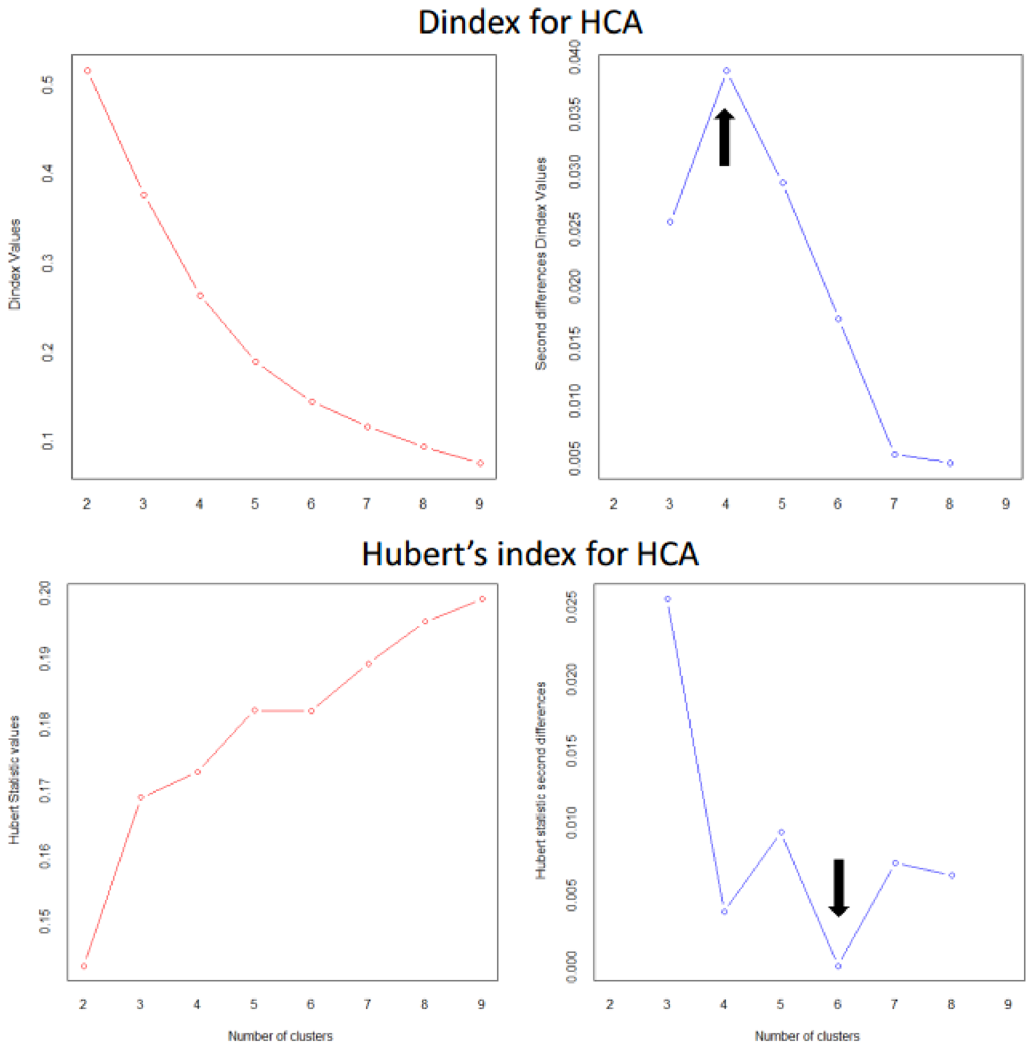


Fig. 42. Dindex and Hubert's index graphs presenting calculated optimal number of cluster value marked by an arrow.

4.2.5. Multivariate Data Analysis

The Spectronomy system presently implements several multivariate data analysis and clustering techniques, such as principal component analysis (PCA), kernel principal

© 2017 Tous droits réservés.

component analysis (kPCA) (Fig. 18), hierarchical cluster analysis (HCA), k-means clustering and fuzzy-C-means clustering (Fig. 43).

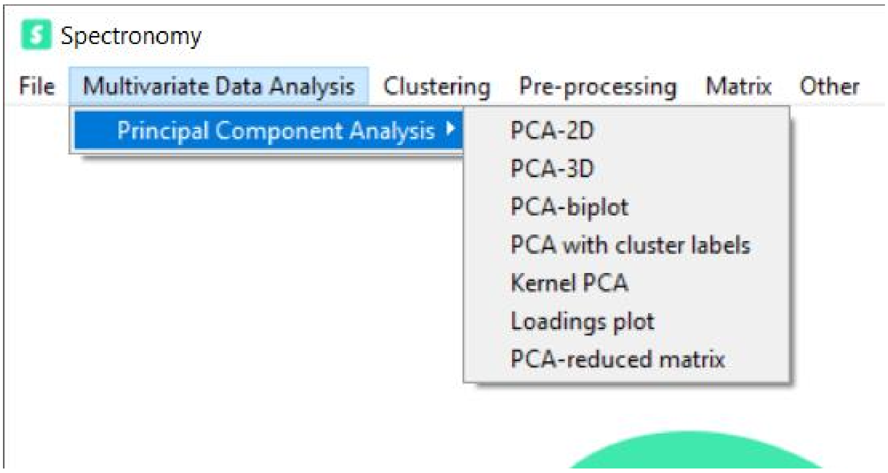
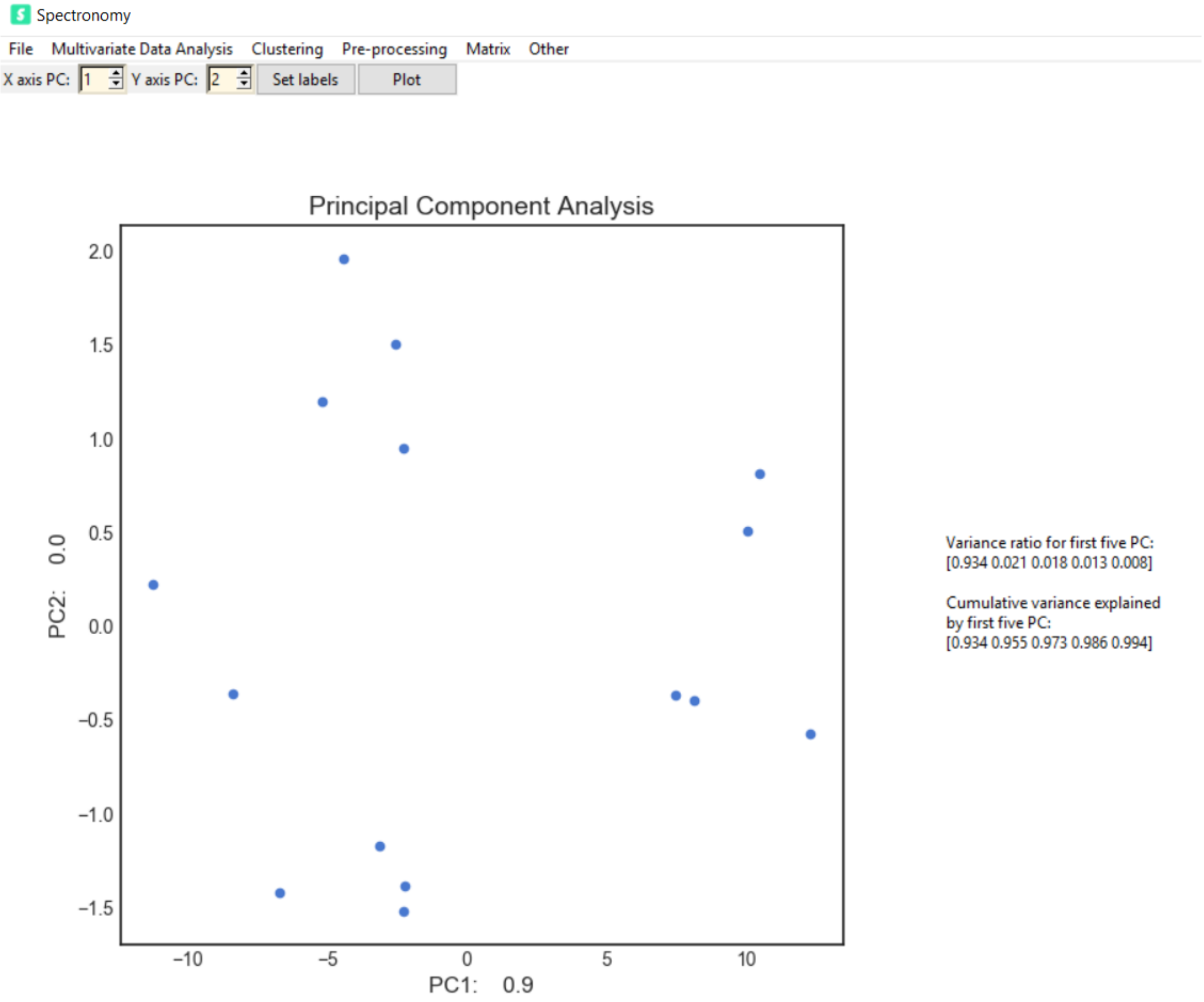


Fig. 43. PCA procedures implemented in the Spectronomy software.

Principal component analysis (PCA) is a multivariate technique that operates in an unsupervised manner and is used to analyze the inherent structure of the data. PCA is a widely used tool for dimensionality reduction, model building and data exploration with a huge potential for Raman and FTIR spectroscopy (Gautam et al. 2015). Considering that only a few principal components are necessary to represent a majority of a total variance in the data set, PCA is a great way to reduce the dimensionality of data (Reisner et al. 2011). By analyzing multiple Raman or FTIR spectra by PCA, each spectrum can be represented in terms of the principal component variables as a small set of values, called scores. The graphical user interface implemented in the Spectronomy system has a feature for a flexible operation on the selected principal components to generate a score plot adapted for the purposes defined by a user. In addition, two types of a score plot can be generated: (i) two dimensional (Fig. 44) and (ii) three-dimensional (Fig. 45).

© 2017 Tous droits réservés.

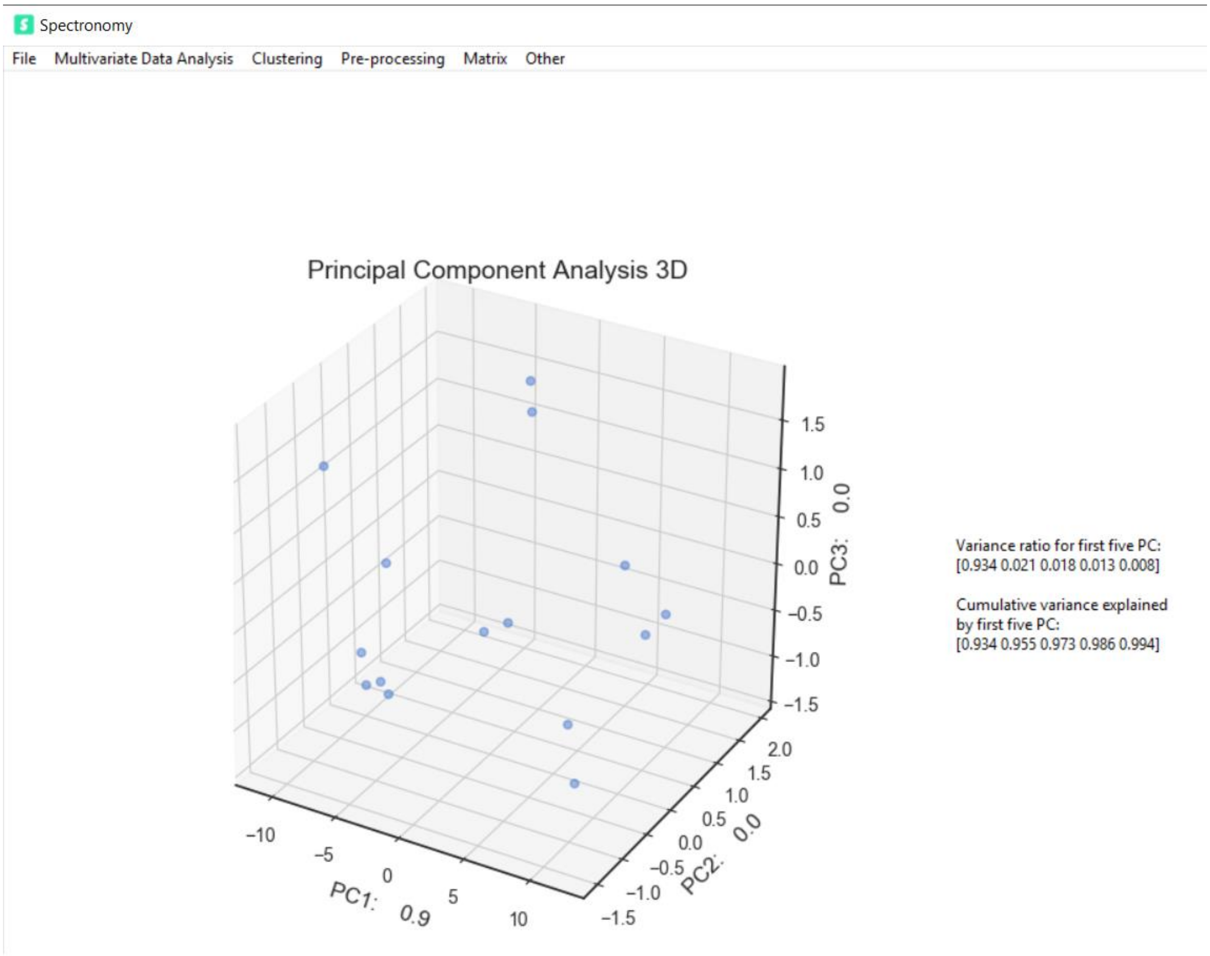


Variance ratio for first five PC:
[0.934 0.021 0.018 0.013 0.008]

Cumulative variance explained
by first five PC:
[0.934 0.955 0.973 0.986 0.994]

Fig. 44. Two-dimensional projection of the scores generated by PCA.

© 2017 Tous droits réservés.



Variance ratio for first five PC:
[0.934 0.021 0.018 0.013 0.008]

Cumulative variance explained by first five PC:
[0.934 0.955 0.973 0.986 0.994]

Fig. 45. Three-dimensional projection of the scores generated by PCA.

Kernel principal component analysis (kPCA) (Fig. 46) is a nonlinear form of PCA, which better exploits a complicated spatial structure of high-dimensional features. If the original data exist with a complex nonlinear relationship, kPCA is more suitable for the feature extraction (Shao et al. 2014).

© 2017 Tous droits réservés.

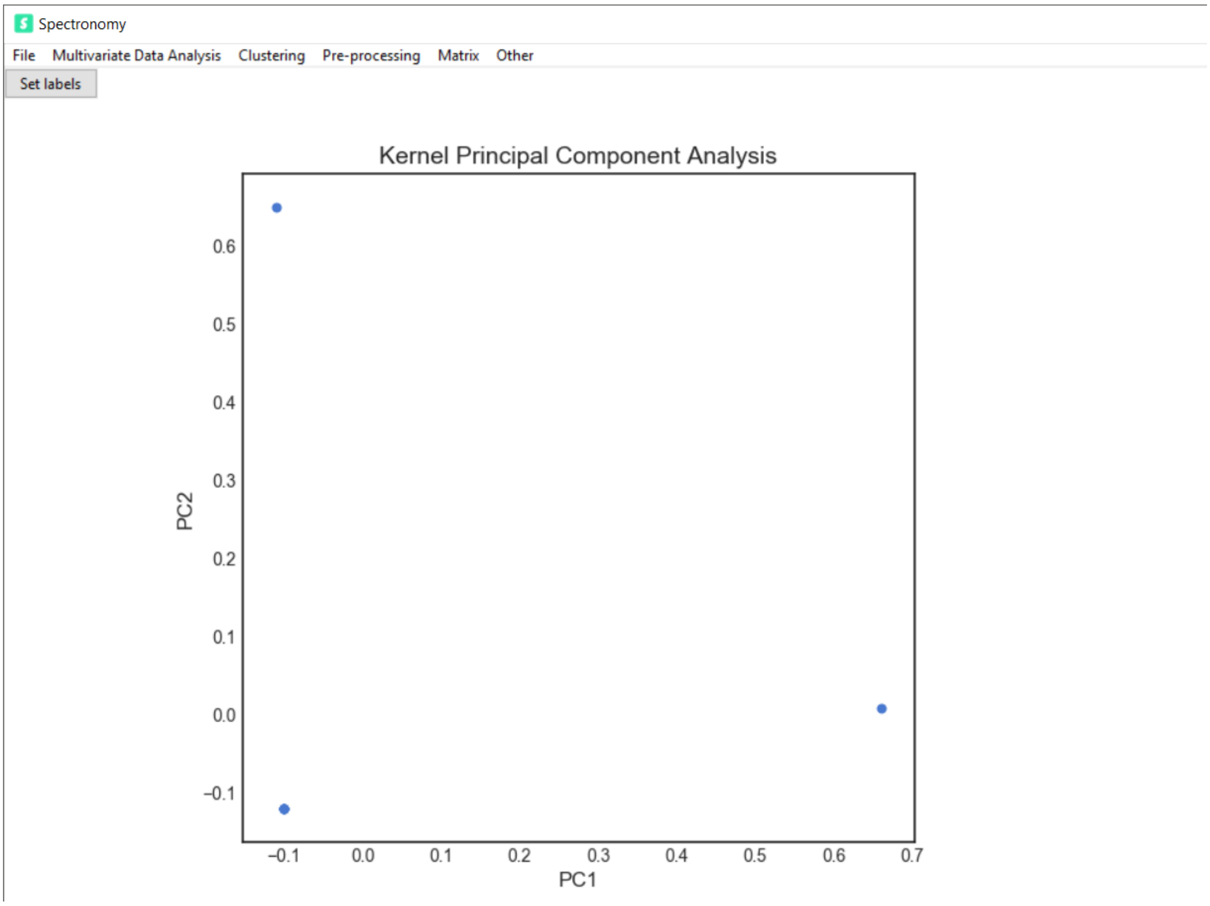
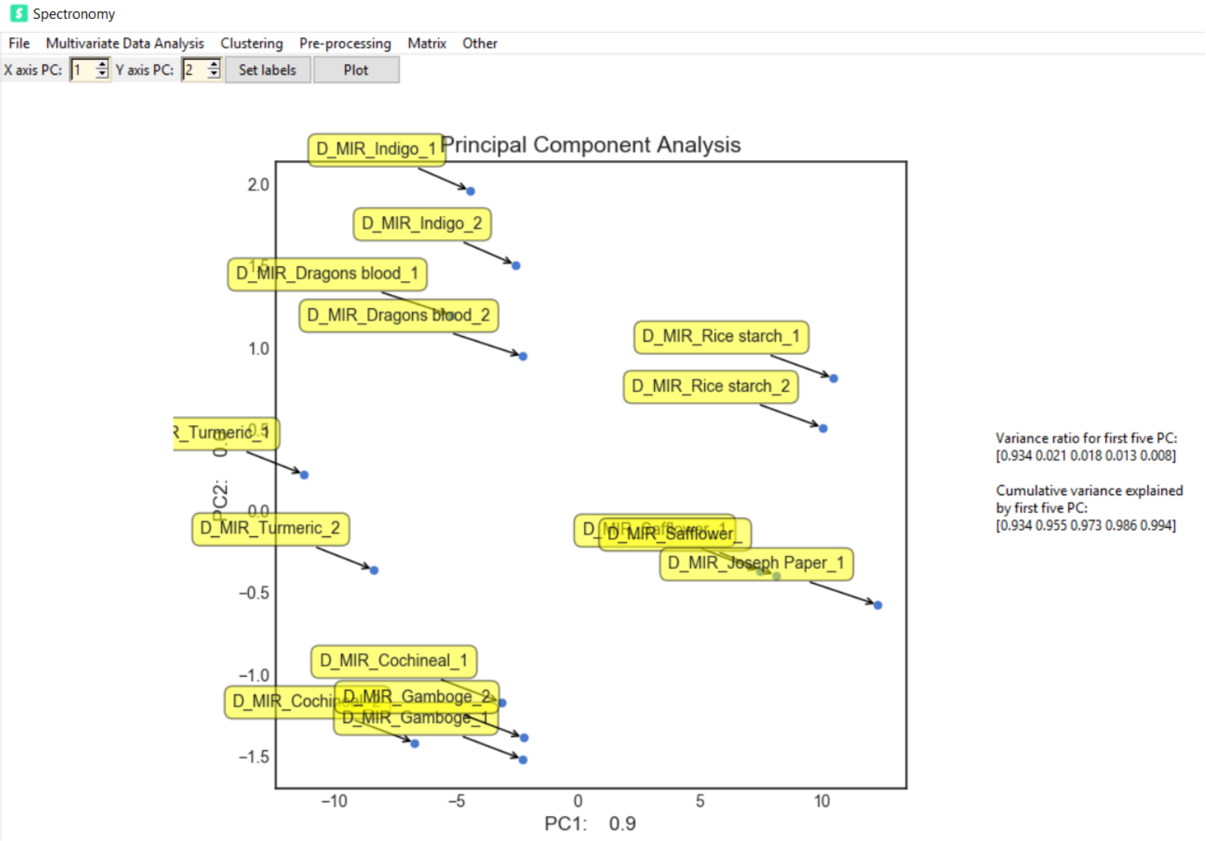


Fig. 46. Two-dimensional projection of the scores generated by kPCA.

Both techniques are accessible in the Spectronomy system to further score labeling function (Fig. 47).

© 2017 Tous droits réservés.



Variance ratio for first five PC:
[0.934 0.021 0.018 0.013 0.008]

Cumulative variance explained
by first five PC:
[0.934 0.955 0.973 0.986 0.994]

Fig. 47. Two-dimensional projection of the scores generated by PCA with score labels.

In addition, a possibility to display loading plots of the specified principal components (Fig. 48) and to export a PCA-reduced matrix is also available (Fig. 49).

© 2017 Tous droits réservés.

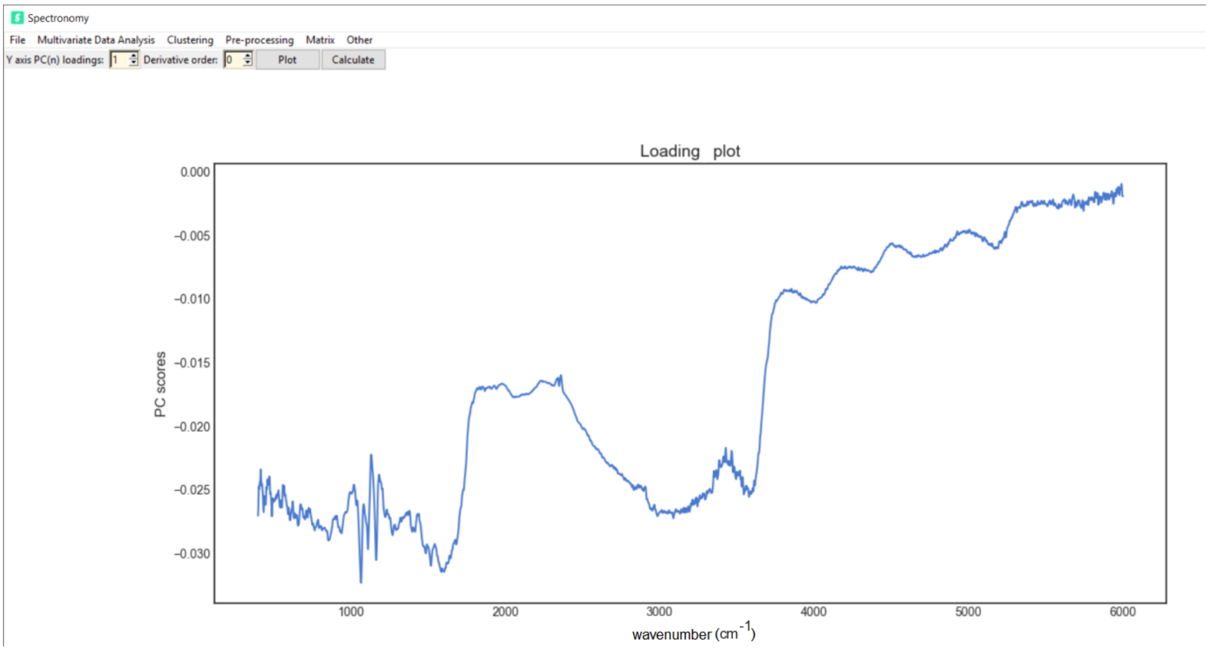


Fig. 48. Loading plot of the first PC.

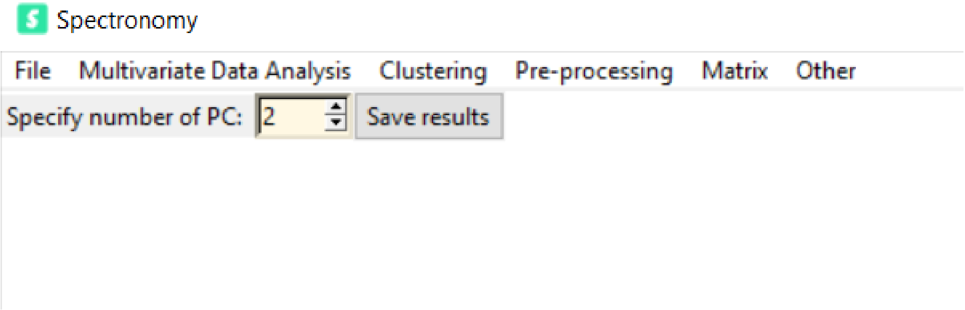
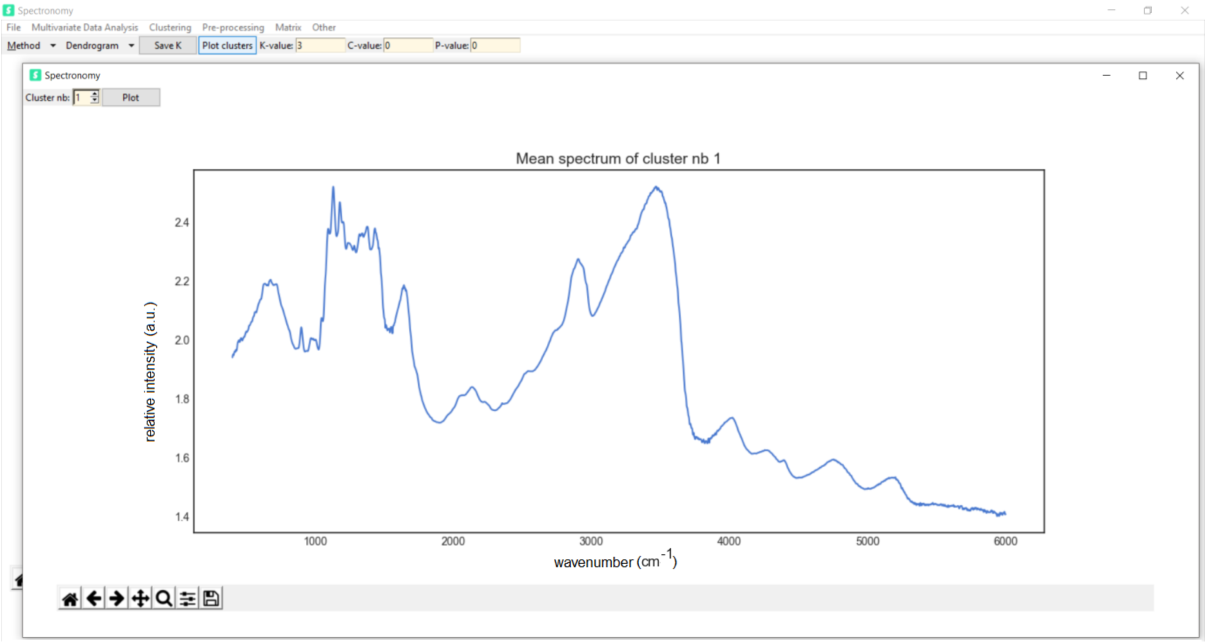


Fig. 49. Export function of the reduced PCA matrix included in the Spectronomy software.

Another novelty is a possibility to plot an average spectrum, which is associated with a specific cluster (Fig. 50).

© 2017 Tous droits réservés.



K-value: 3

C-value: 0

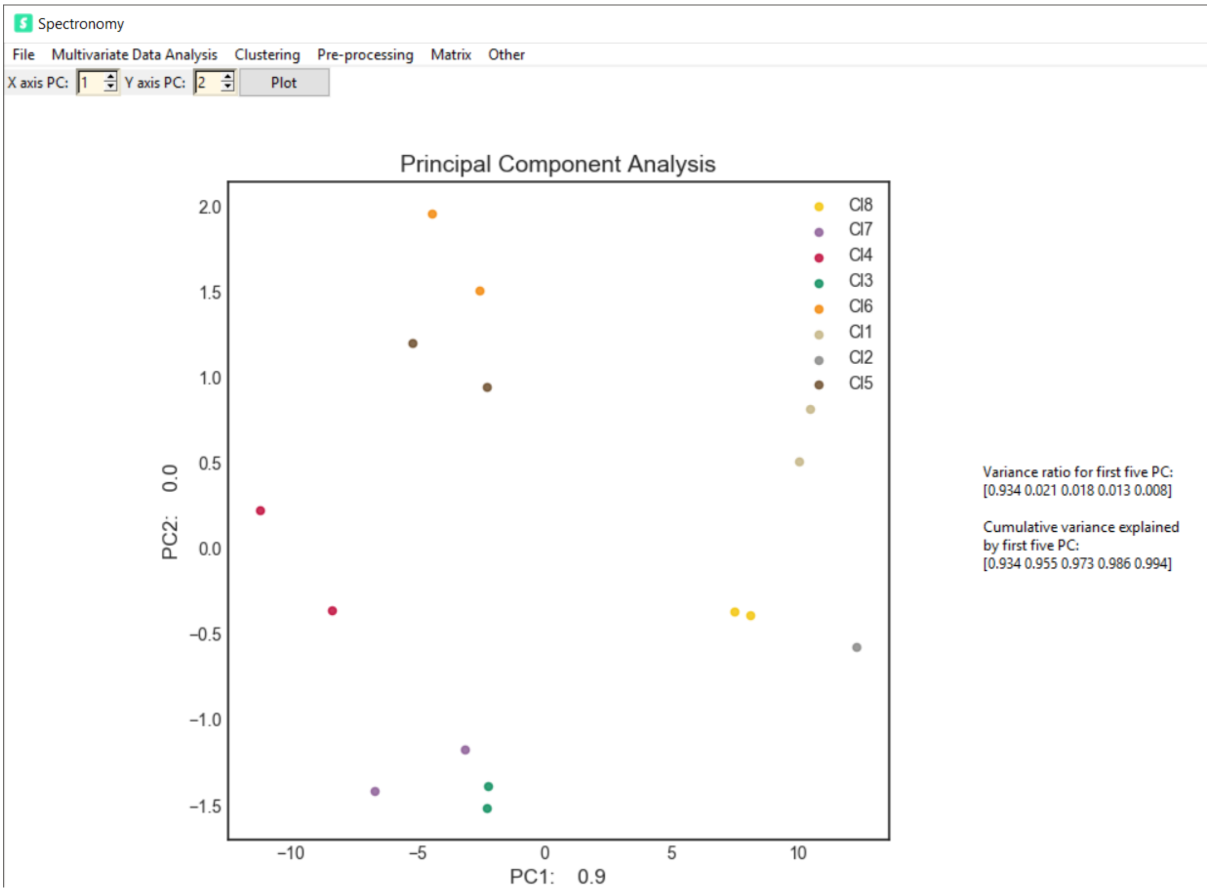
P-value: 0

Cluster nb: 1

Fig. 50. Mean spectrum of one cluster function accessible in the HCA clustering in the Spectronomy software.

The clustering results in the form of a comma-separated file can be exported and saved on the hard drive and then used to assign different cluster colors to the PCA scatter plot (Fig. 51). This function is especially useful in a combination of cluster analysis (CA) with principal component analysis (PCA) to obtain and characterize any relationships between these techniques

© 2017 Tous droits réservés.



Variance ratio for first five PC:
[0.934 0.021 0.018 0.013 0.008]

Cumulative variance explained by first five PC:
[0.934 0.955 0.973 0.986 0.994]

Fig. 51. PCA score plot with cluster labeling feature.

4.3. Evaluation of the system

The software performance and effectiveness were evaluated using two different data sets from Raman and FTIR spectroscopy. The computer used for the evaluation was equipped with a 2.0 GHz Intel Core i7-4750HQ processor and 8 GB of DDR3-RAM memory. Since the purpose of this paper is to introduce the system rather than analyze a particular data set, the specifics of the specimens are not considered relevant.

4.3.1 Evaluation of Raman data set

The fine powders of $CaCO_3$, $NaNO_3$ and Na_2SO_4 in the weight ratio 1:1:1 were ground in a mortar in order to obtain micron-sized particles of a diameter range from 1 to 10 micron. The particles were transferred to an Eppendorf tube with 5 ml of methanol.

© 2017 Tous droits réservés.

The Eppendorf tube was spinned in the vortex. Then, after 30 seconds, 10 µl volume was transferred on a silver foil substrate. The sample was left for 12 hours in a laminar flow chamber to dry off. Thus, the Raman data set consisted of 978 spectra was collected by the inVia Raman microscope (Renishaw, Wotton-under-Edge, UK) equipped with 785 nm Near-Infrared Diode Laser lines as the excitation source with a spectral resolution of ~ 1.3 cm^{-1} and a 100× (N.A. 0.9) Olympus objective, WiRE 4 ™ acquisition software, and thermoelectrically cooled CCD (1024 × 256 pixels) detector. The laser power was set at 100%. The wavenumber scale was calibrated using Si as a standard. The time of acquisition was set for 3 seconds with a 1 acquisition approach. The data matrix with dimensions of 1015 × 978 (wavenumbers × number of spectra) was constructed.

The evaluation of the first data set was dedicated to emphasize the software capabilities in the processing of a large number of Raman spectra. The system was first used to carry out the initial pre-processing of the raw spectra (Fig. 52).

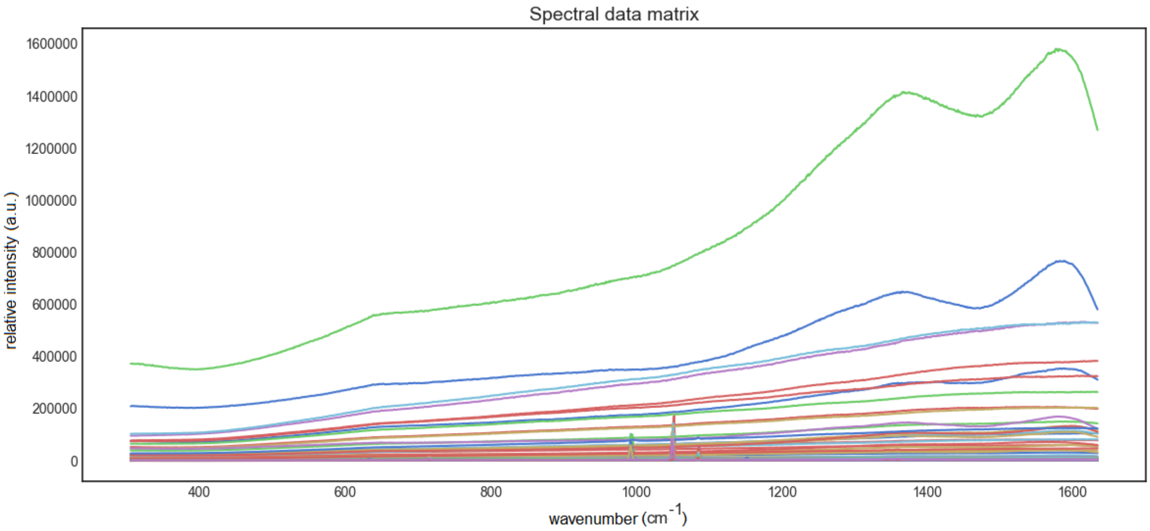


Fig. 52. Raw Raman spectra of $CaCO_3$, $NaNO_3$ and Na_2SO_4.

The loading of the first data matrix (Raman spectra) took 9.24 seconds. In turn, the spectra intensity scaling took 7.10 seconds, where the automatic baseline correction with the asymmetric least squares option took 45.39 seconds (Fig. 28).

© 2017 Tous droits réservés.

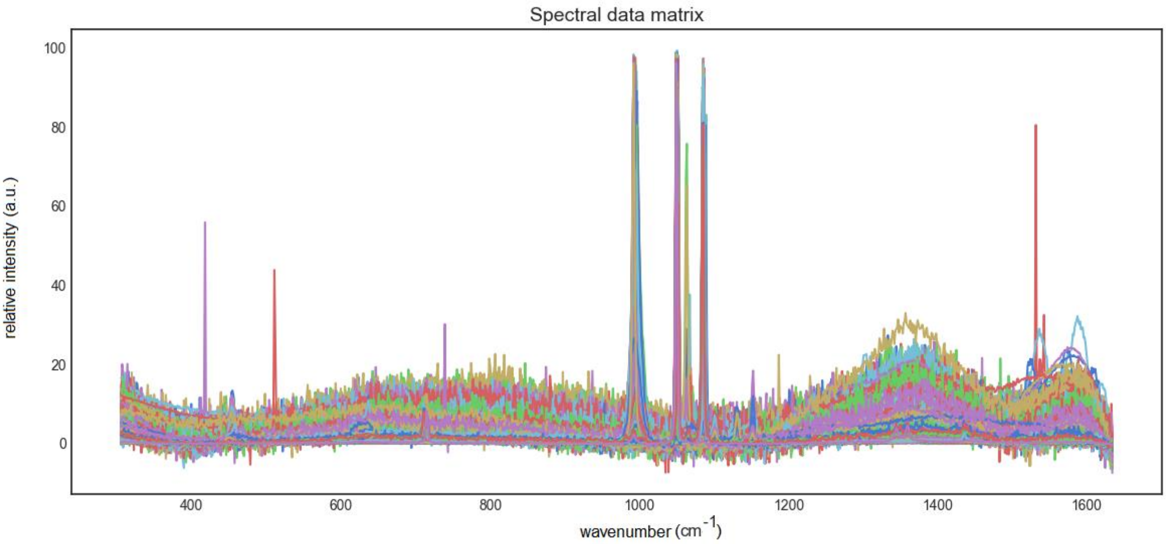


Fig. 53. Raman spectra after pre-processing of intensity autoscale and AsLS baseline correction (lambda = 10^7 and p=0.01).

Afterwards, the spectra were used to generate a PCA model and then projected in the form of the PCA score plot (Fig. 54).

© 2017 Tous droits réservés.

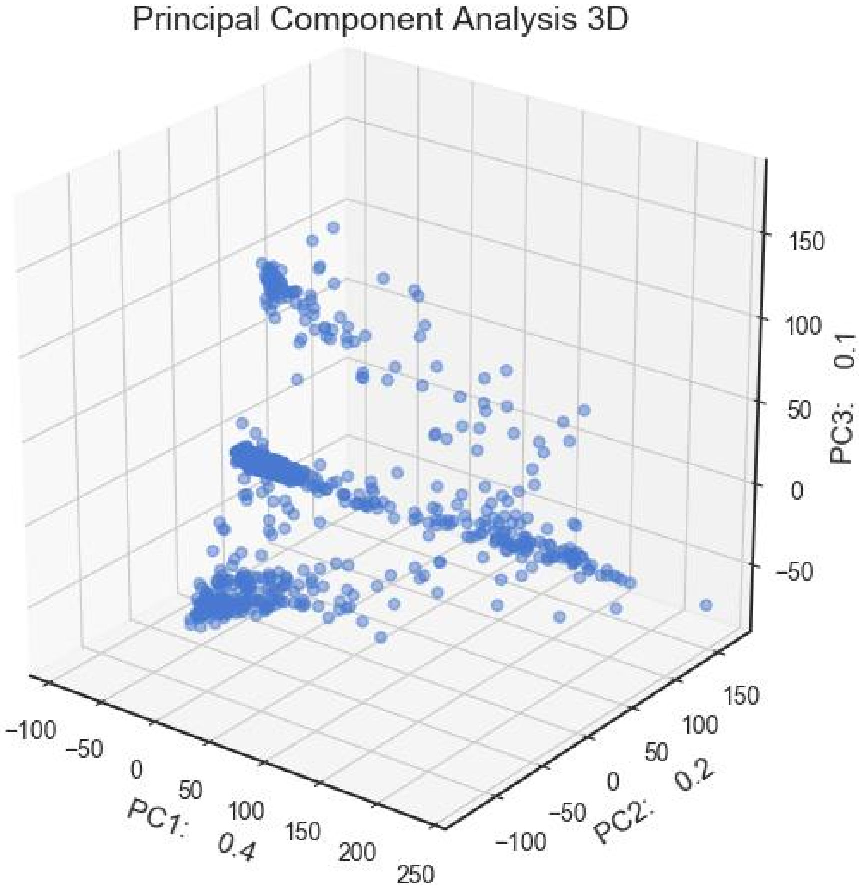


Fig. 54. 3-D PCA scatterplot of the Raman spectra data matrix.

This process took only 0.52 s. The most remarkable result of the PCA is a pattern formation on a scatter plot. In the graph presented in Fig. 54, four groups of the scores are well recognizable, which provide a number of the major clusters in the sample. The number of four clusters was used in the clustering methods based on this result.

In the next step, all the included clustering techniques were tested. Construction of the Ward's HCA model and the dendrogram projection took 5.56 seconds (Fig. 55).

© 2017 Tous droits réservés.

Then, the projection of a truncated mode of the dendrogram was faster and took 3.28 seconds.

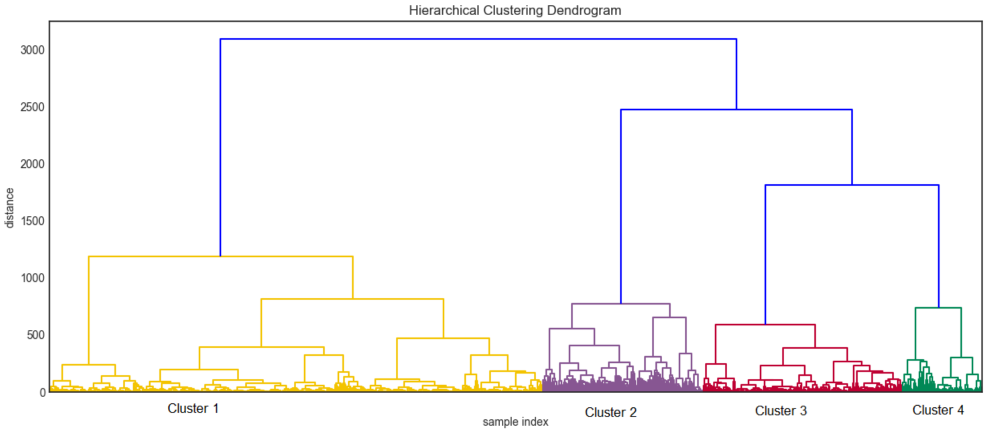
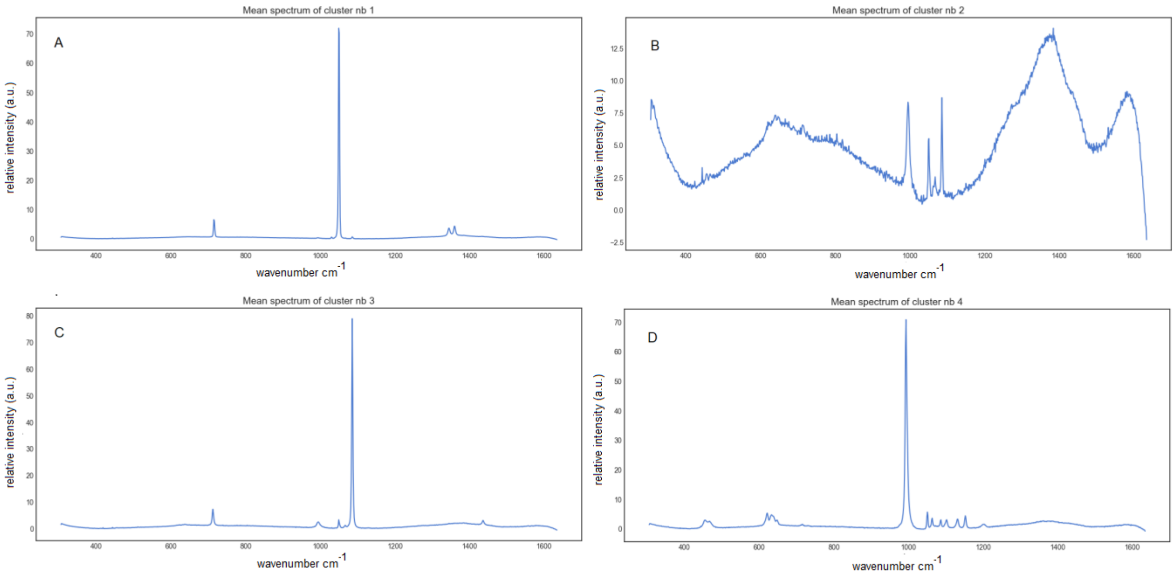


Fig. 55. Ward's HCA dendrogram of Raman spectra with specified 4 clusters.

Owing to the HCA results, the cluster 1 contains 516 spectra, which represents 52% of the sample. In turn, cluster 2 contains 168 spectra, which represents 17% of the sample. Finally, cluster 3 and cluster 4 represent 208 and 84 spectra respectively. Cluster 3 represent 22% of the sample population, where cluster 4 represents 9% of the population. For the specification of each cluster composition, the mean spectra of the clusters were generated which took only 1.42 second each (Fig. 56).



© 2017 Tous droits réservés.

Fig. 56. Mean spectra of 4 clusters generated by Ward's HCA algorithm. A – mean spectrum of cluster 1; B – mean spectrum of cluster 2; C – means spectrum of cluster 3; D – mean spectrum of cluster 4.

By identification of the mean cluster spectra, the cluster composition was given. The cluster 1 represents $NaNO_3$ particles, the cluster 2 represents mixed, heterogeneous particles, which were developed by grinding the powders, and the cluster 3 represents $CaCO_3$ particles, where the cluster 4 represents Na_2SO_4 particles.

In the case of the k-means clustering, the implemented system of a data validation was used. Due to the small distances between clusters, the Voronoi diagram was automatically replaced by the PCA score plot with points colored according to the cluster order (Fig. 57). This action may occur when the points in the Voronoi graph are closely accumulated. This type of clustering was relatively fast and took only 4.21 s. However, due to a decision making of an appropriate graph application the process lasted 3.42 minutes.

© 2017 Tous droits réservés.

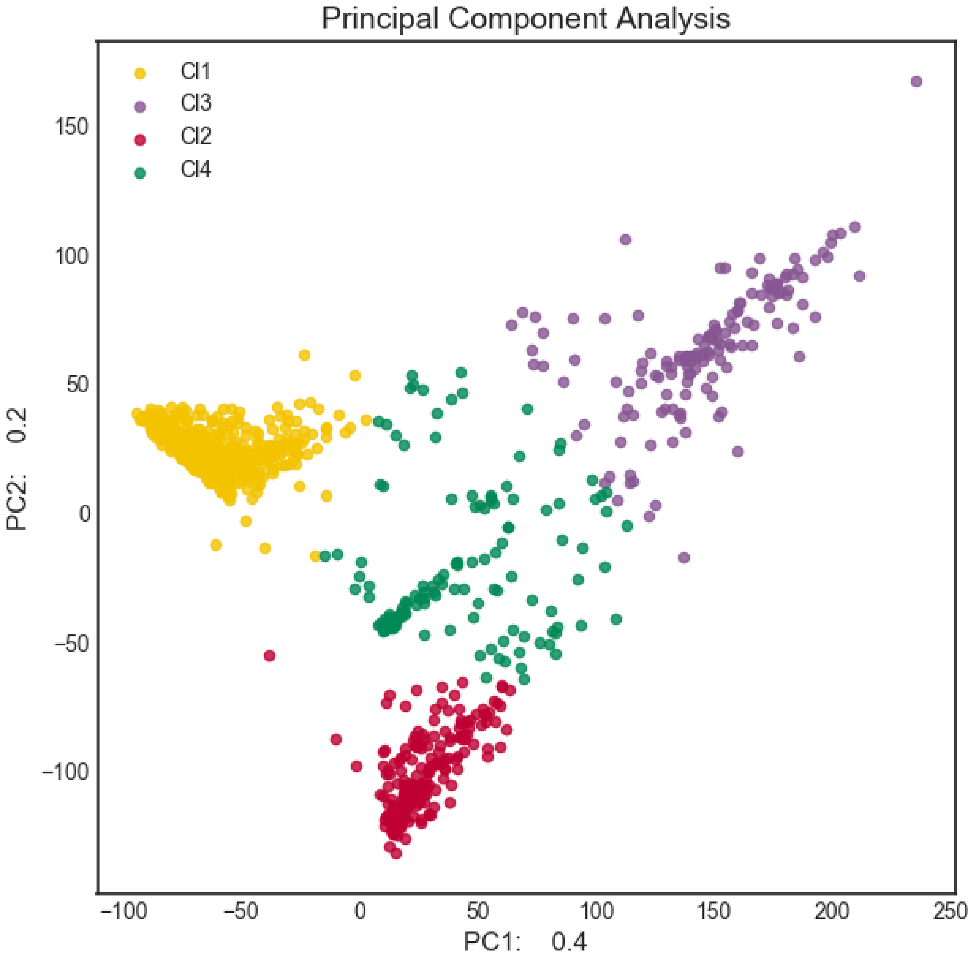


Fig. 57. 2-D PCA scatterplot with four labeled clusters from k-means clustering –Cl1: $NaNO_3$; Cl2: $CaCO_3$; Cl3: Na_2SO_4 and Cl4: mixture.

The clustering results generated by the k-means algorithm are complementary to the Ward's HCA results. The projection of the labeled scores by the k-means cluster vector shows the mixed cluster Cl4 which is situated in the middle of the other clusters at the scatter plot.

The final clustering task was done by fuzzy-C-means clustering. This task covers two steps: (i) calculation and visualization of the FPC values (Fig. 58); (ii) implementation of fuzzy-C-means clustering (Fig. 59). The first step took 8.54 seconds, where the second one was faster and took 2.35 seconds.

© 2017 Tous droits réservés.

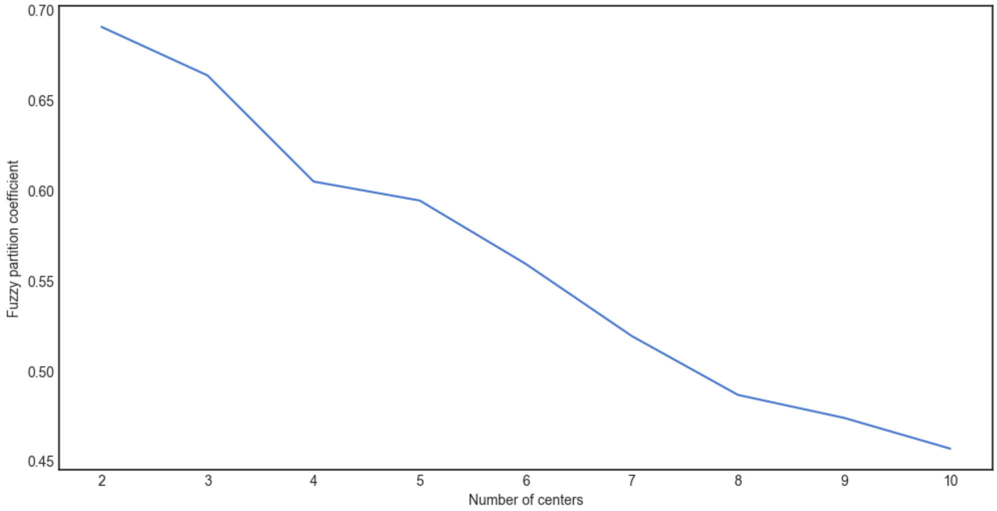
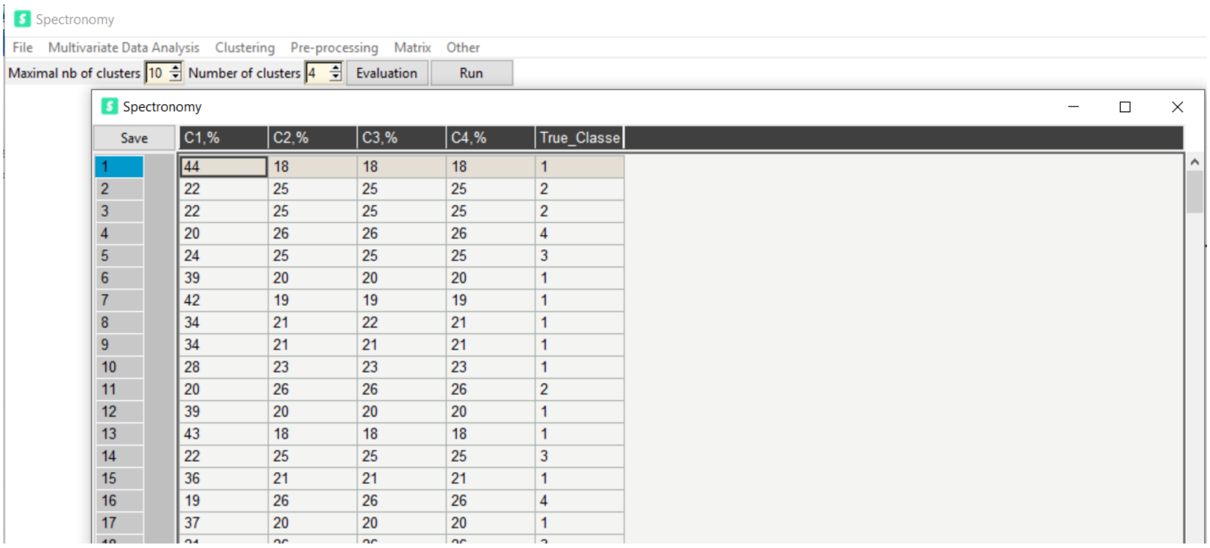


Fig. 58. Fuzzy partitioning coefficient calculated for Raman spectra data set.

The FPC projection does not provide clear information about a number of clusters for fuzzy-C-means clustering. Due to the data compatibility, the number of four clusters was used for the calculations (Fig. 59).



Maximal nb of clusters 10 Number of clusters 4

	C1,%	C2,%	C3,%	C4,%	True_Classe
1	44	18	18	18	1
2	22	25	25	25	2
3	22	25	25	25	2
4	20	26	26	26	4
5	24	25	25	25	3
6	39	20	20	20	1
7	42	19	19	19	1
8	34	21	22	21	1
9	34	21	21	21	1
10	28	23	23	23	1
11	20	26	26	26	2
12	39	20	20	20	1
13	43	18	18	18	1
14	22	25	25	25	3
15	36	21	21	21	1
16	19	26	26	26	4
17	37	20	20	20	1

Fig. 59. Print-screen of tabulated results generated by fuzzy-C-means clustering algorithm with cluster membership (%) for each spectrum in the data set.

For presented clustering results, correlation coefficients were calculated with the equation presented below (29):

© 2017 Tous droits réservés.

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \qquad (29)$$

Where: *Correl* – correlation coefficient, *X* – first numeric input (number, vector, etc.), *Y* – second numeric input (number, vector, etc.), $\bar{X}$– average of *X*, $\bar{Y}$– average of *Y*.

The numeric input in this case is a clustering vector obtained by each technique. The calculated correlation coefficient results are presented in Table 2. Fuzzy-C-means clustering has the smallest correlation coefficient (0.54) to both HCA and k-means techniques. This value is predictable because of the differences in the applied algorithms.

Table 2. Correlation coefficient in the different clustering algorithms.

Method/Correlation coefficient	K-means	HCA	Fuzzy-C-means
K-means	-	1	0.54
HCA	1	-	0.54
Fuzzy-C-means	0.54	0.54	-

4.3.2. Evaluation on FTIR data set

The Fourier transform near-infrared (FT-NIR) spectral measurements were performed using a Perkin-Elmer (Waltham, MA) analytical system consisted of a Spectrum One FT-NIR spectrometer coupled to a Spectrum Spotlight 400 NIR microscope. An optical fiber was coupled with the system of a sample analysis. The described data set represents 14 spectra of 6 different organic pigments and 4 spectra of a binding medium (rice starch). The spectral range was set for 400-1100 cm^{-1} (Fig. 60). The data set dimension was 876 × 18 (wavenumbers × number of spectra).

© 2017 Tous droits réservés.

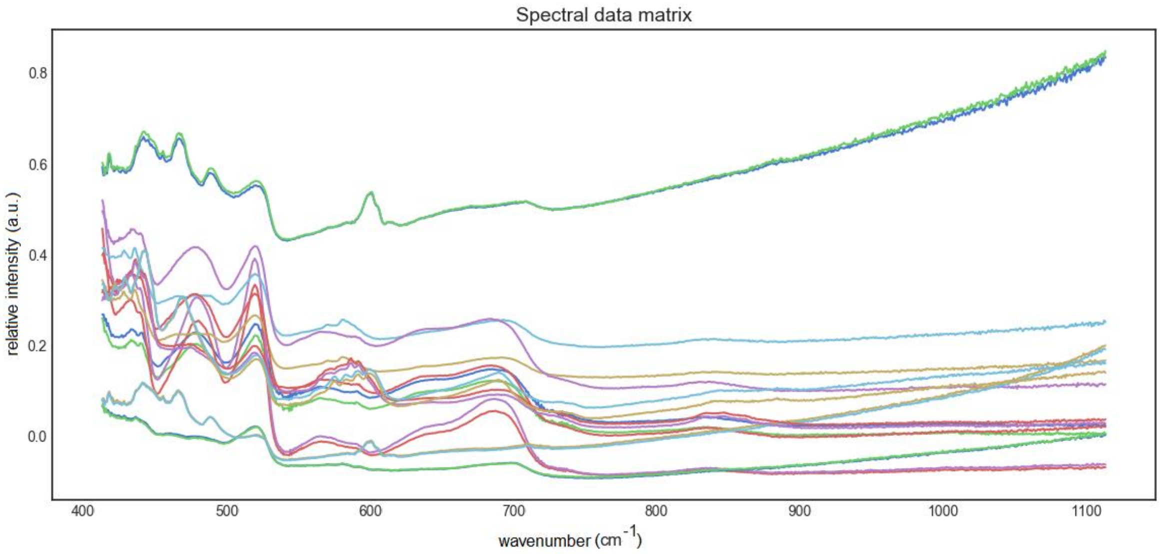


Fig. 60. Near-IR spectra of pigments and binder medium (rice starch).

Next, a series of tasks including the spectral pre-processing, principal component analysis and cluster analysis was applied for pattern recognition in the data set. The first step was an application of the 3rd polynomial EMSC for the pre-processing (Fig. 61).

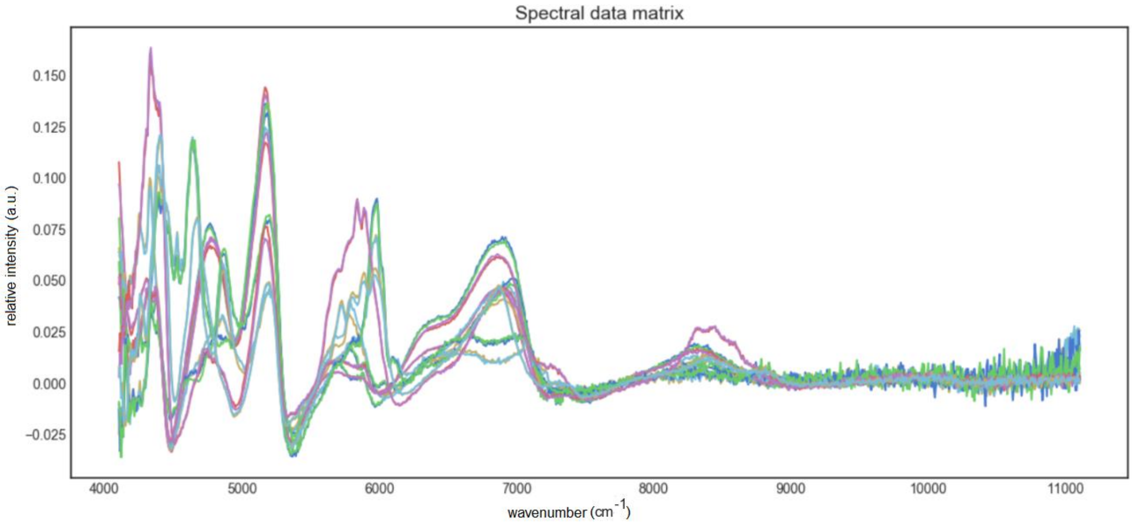


Fig. 61. Near-IR spectra after pre-processing step of 3rd EMSC and AsLS baseline correction (lambda = 10^5 and p = 0.1).

© 2017 Tous droits réservés.

The second step was an application of the PCA for the exploratory data analysis (Fig. 62).

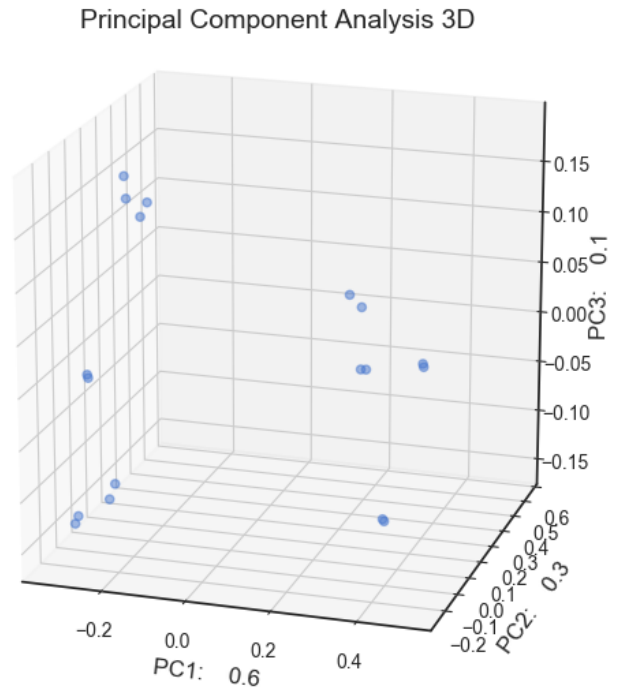


Fig. 62. Three-dimensional scatter plot of Near-IR spectra after PCA.

In the next step, all the three available clustering techniques were used to perform pattern recognition. The samples for the purpose of this evaluation were unmarked in the testing data set and then confronted with the labels determined during the spectra acquisition. The optimal number of clusters was calculated by the fuzzy coefficient and found 7 clusters (Fig. 63).

© 2017 Tous droits réservés.

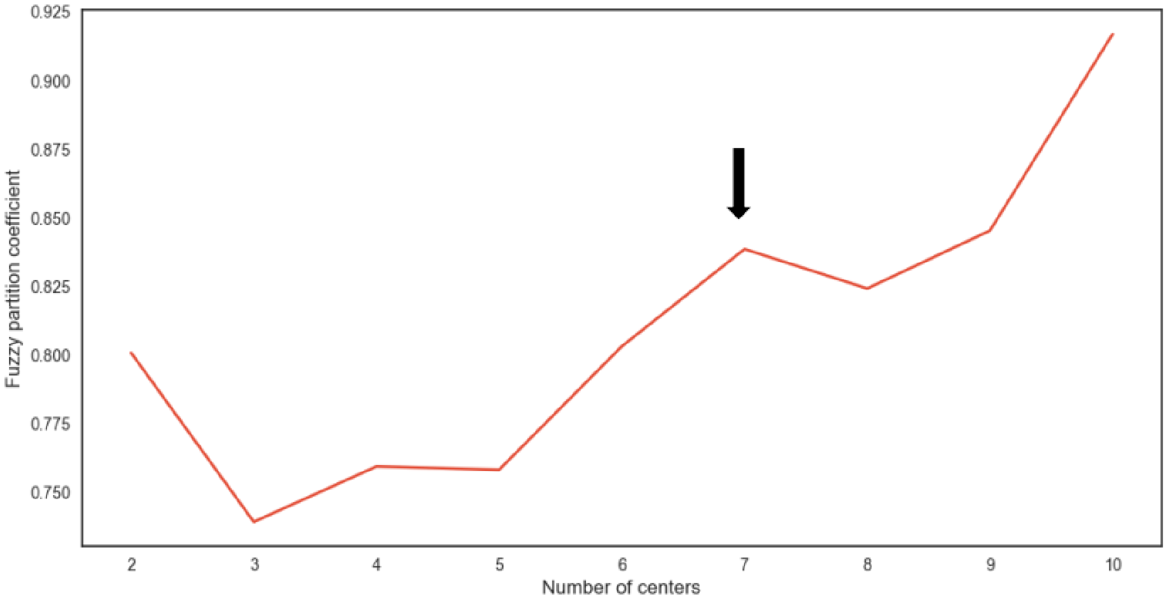
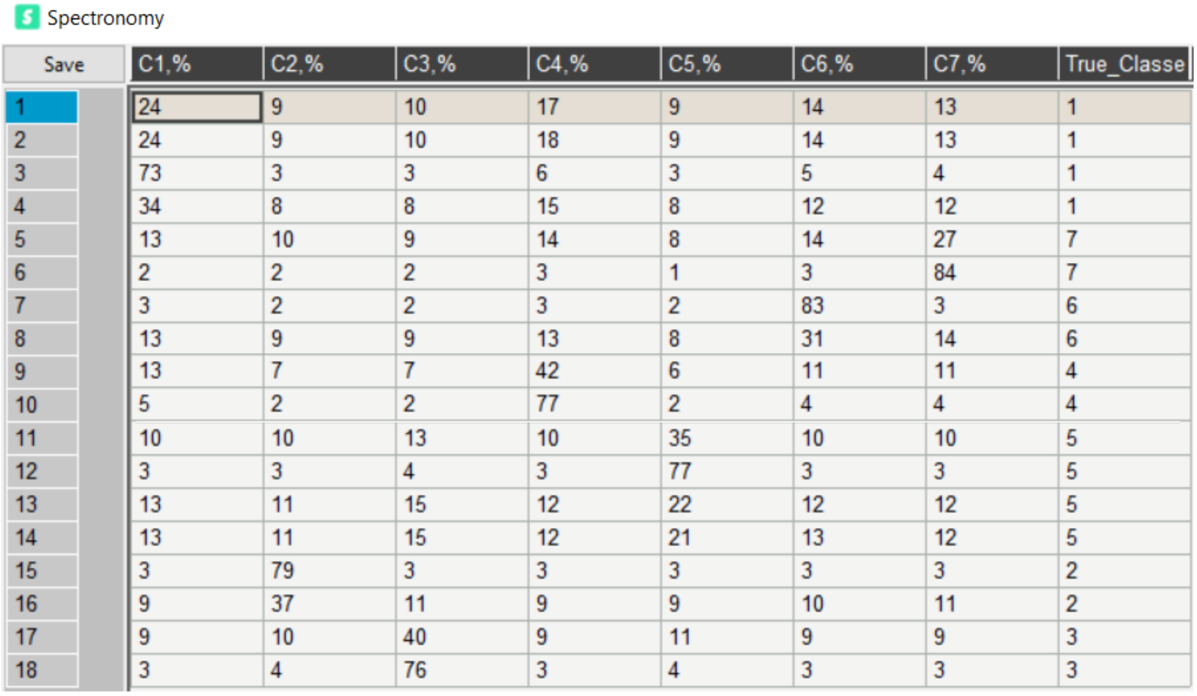


Fig. 63. Fuzzy partitioning coefficient calculated for NIR spectra data set. The black arrow shows the optimal number of clusters.

The FPC is defined in the range from 0 to 1, with 1 being the best. It is a metric, which demonstrates how accurately the data is described by a certain model. When the FPC is maximized, the data is described in the best way. In fact, the best practice is to find a significant knee in the FPC plot and then test fuzzy c—mean clustering, based on a designated value of clusters. The cluster analysis: fuzzy-C-means (Fig. 64); k-means (Fig. 65) and HCA (Fig. 66) pointed out the separated 7 clusters.

© 2017 Tous droits réservés.



	C1,%	C2,%	C3,%	C4,%	C5,%	C6,%	C7,%	True_Classe
1	24	9	10	17	9	14	13	1
2	24	9	10	18	9	14	13	1
3	73	3	3	6	3	5	4	1
4	34	8	8	15	8	12	12	1
5	13	10	9	14	8	14	27	7
6	2	2	2	3	1	3	84	7
7	3	2	2	3	2	83	3	6
8	13	9	9	13	8	31	14	6
9	13	7	7	42	6	11	11	4
10	5	2	2	77	2	4	4	4
11	10	10	13	10	35	10	10	5
12	3	3	4	3	77	3	3	5
13	13	11	15	12	22	12	12	5
14	13	11	15	12	21	13	12	5
15	3	79	3	3	3	3	3	2
16	9	37	11	9	9	10	11	2
17	9	10	40	9	11	9	9	3
18	3	4	76	3	4	3	3	3

Fig. 64. Print-screen of tabulated results generated by fuzzy-C-means clustering algorithm with cluster membership (%) for each spectrum in the data set.

© 2017 Tous droits réservés.

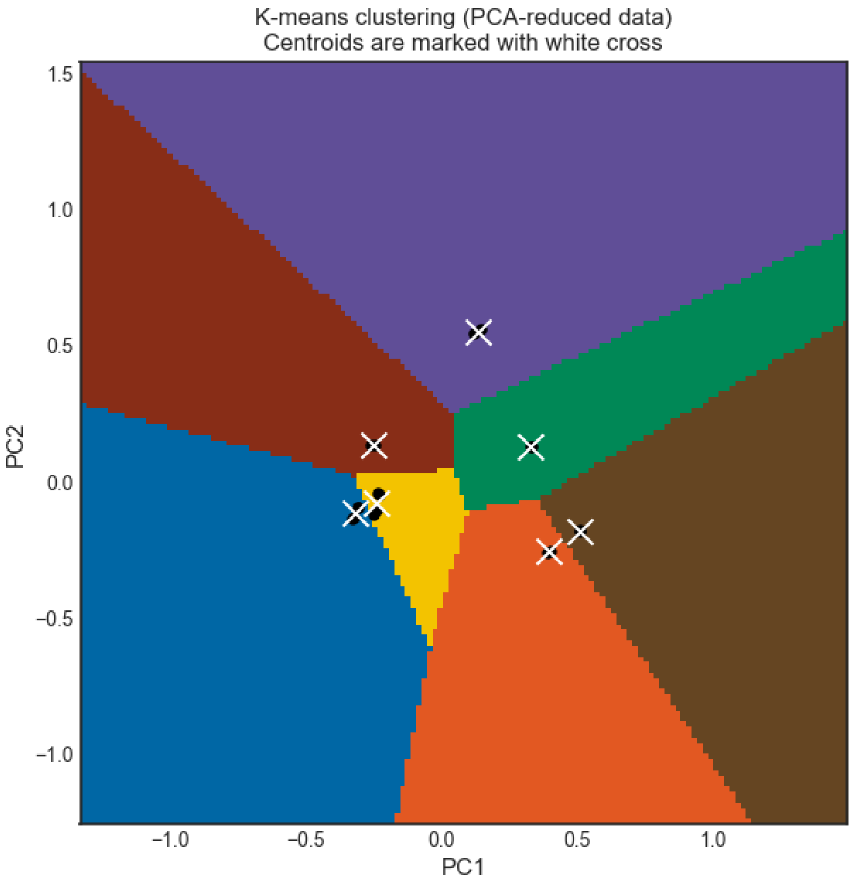


Fig. 65. Voronoi diagram generated by application of k-means clustering (7 clusters specified a priori) algorithm on PCA-reduced data with marked centroids.

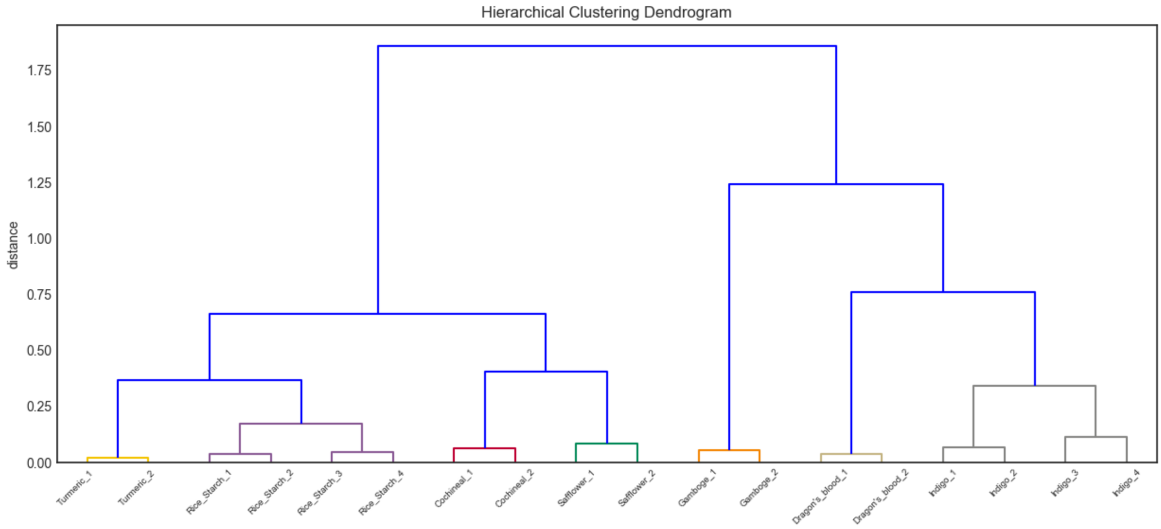


Fig. 66. Ward's HCA dendrogram of NIR-IR spectra of pigments and binding medium.

© 2017 Tous droits réservés.

The combination of both PCA and HCA techniques included in the separated feature is presented in Fig. 67.

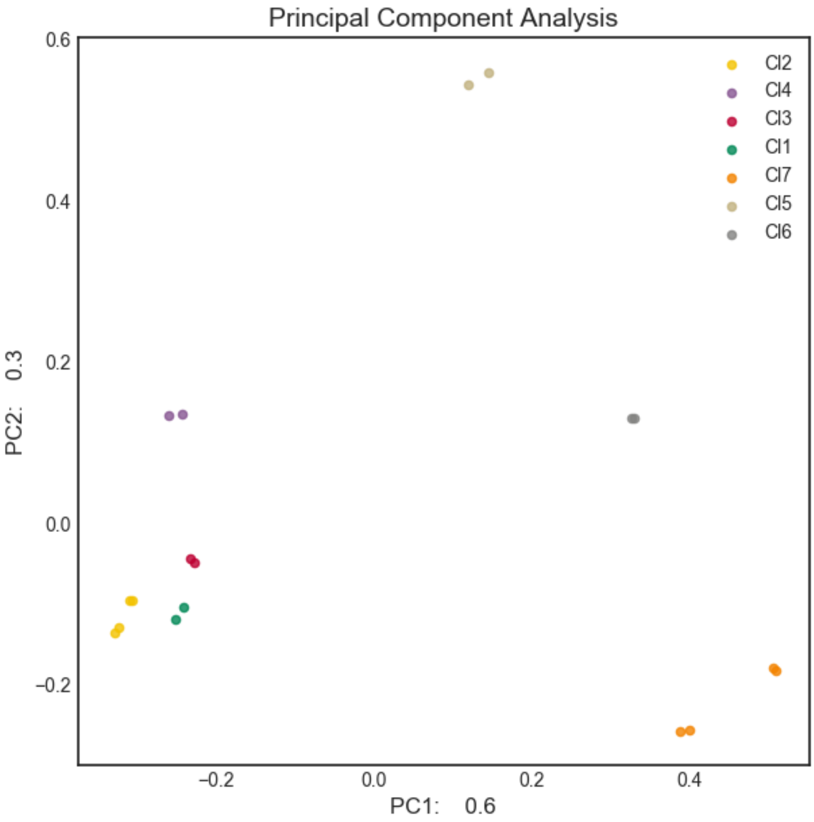


Fig. 67. PCA score plot results with cluster designed by HCA.

The accuracy of all techniques implemented in the software was 100%. The results show a great potential of the Spectronomy system for a fast and accurate pattern recognition applied to a spectral data matrix. However, it should be emphasized, that a quality of collected spectra and specification of appropriate pre-processing techniques is crucial for a success of such proceeding. Nevertheless, the Spectronomy system is flexible and adjusted for a trial and error method of solving problems, which significantly improves the effortlessness of a multivariate analysis for a user.

The system processing algorithms (intensity scaling, background subtraction,

© 2017 Tous droits réservés.

normalization, principal component analysis and cluster analysis) were previously used by our research group to help classify airborne aerosol particles and pigments.

4.4. Conclusions

The data analysis of multiple Raman and FTIR spectra is a complex process with no standard implementation of commonly executed methods. Therefore, the open-source, freely available system that handles the pre-processing, principal component analysis and cluster analysis dedicated to spectral data was developed. The Spectronomy software includes many easy-to access, powerful algorithms coupled with a well-developed graphical user interface. Several implementations for Raman and IR spectral processing were performed. The optimal number of cluster calculation, average cluster spectrum projection, fuzzy-C-means clustering with the FPC parameter specification is a matter of concern. Furthermore, the system combines two programming languages, which significantly improve a potential application scope, owing to a huge number of available packages. The evaluation of the system showed fast and accurate analytical capabilities for Raman and IR spectra. However, it should be noted that the Spectronomy system is not a complex software and it needs a constant development. There are numerous areas that could be improved, such as a cosmic spike filter, spectral feature extraction, spectral database management, spectra classification and a calculation performance for big data analysis. Nevertheless, we are confident that making this program available to users will result in a faster and better-designed development. The ultimate purpose of this software development is to provide a powerful tool for any application of Raman and FTIR spectroscopy.

The attributes included in the Spectronomy makes this software prominent. First, it is based only on the free, open-source and up-to-date packages, which significantly improves interoperability and consistency of the software with new operating systems. Second, the source code of the software is also made available. It gives a

© 2017 Tous droits réservés.

possibility to modify the system or to equip it with new features. Third, Spectronomy is based on a new release of Python 3 connected with the R programming language environment, which capabilities allow the implementation of almost any practicable feature. It should be emphasized, that a number of free, official packages for Python 3 and R programming language is estimated for 116 387 and above 10 000, respectively. The fourth benefit of the system can be found in the diversity of features that are included in the software, such as scale-based normalization, automated baseline correction, extended multiplicative signal correction (EMSC), standard normal variate correction (SNV), Savitzky-Golay filtering, principal component analysis, k-means clustering, fuzzy c-means clustering, hierarchical cluster analysis (HCA) and optimal number of cluster calculation algorithms.

Finally, in spite of the powerful techniques included in the Spectronomy software, it is still relatively easy to use. All features are accessible through a self-explanatory graphical user interface, which is supposed to facilitate an access to chemometric tools for researchers. However, it should be emphasized that Spectronomy is relatively small and designed for specific purposes and gives way to software such as SASIR (http://www.chimiometrie.fr/saisir_webpage.html). SASIR is a complete and universal package of command-line functions written for MATLAB, SCILAB and OCTAVE which can be used for research and routine works in chemometrics, which exceeds the capabilities and functions of the Spectronomy system. The Spectronomy software was created for relatively small datasets of special purpose in the field of Raman and FTIR spectroscopy, which at the moment does not translate into a development of a comprehensive chemometrics routine. Contrary to windowed environments such as Spectronomy, it has a great advantage to allow batch procedures. It also makes it possible to mix data of any origin (chemical and physical data, spectroscopic data, numeric images). In summary, the Spectronomy system is a development project that gives way to complete projects like SASIR.

© 2017 Tous droits réservés.

4.5. Future works

Several new algorithms are being developed for spectra pre-processing and analysis. Our intention is to add an ability to load files with other extensions – directly exported from the integral spectrometer software. The program will be supplemented with various classification methods, such as neural network classifier, support vector machine classifier, Naïve Bayes methods and stochastic gradient descent. An implementation of regression analysis and relational database framework is also planned. Moreover, our ambition is to adapt the program for a hyperspectral image analysis and an application of multivariate statistics for several matrices in a single run.

© 2017 Tous droits réservés.

CHAPTER 5: APPLICATION OF THE SPECTRONOMY ANALYTICAL SYSTEM

© 2017 Tous droits réservés.

In this chapter, we present the Spectronomy system deployment, conducted for a multivariate data processing of experimental, spectral data sets. The focus of the application is based on the noteworthy paradigms from the field of industrial and biogenic aerosol particle analysis, likewise non-destructive microanalysis of cultural heritage materials. The initial part of this chapter demonstrates the results of the coalmine particle analysis by SEM/EDS and RMS with correspondence to the bulk analysis techniques including ATR-FTIR and XRD. Subsequently, the characterization of pollen grains (an important constituent of the biological aerosol) by RMS with specification of plant species and potential contamination was described. Finally, the results from the Near-FTIR and Mid-FTIR analysis of organic pigments have been presented. The methods for spectral data analysis elaborated in this chapter are subsumed in the Spectronomy system and can be easily re-established by a reader.

5.1. SEM/EDS and RMS analysis of mining environment aerosol particles

5.1.1. Context

Global coal production is still is important to the economy of many countries. According to the World Coal Association (http://www.worldcoal.org), over 7 billion tonnes (Gt) of hard coal is currently produced worldwide, and around 800 million tonnes of brown coal/lignite. The world leaders of hard coal productions are China and USA; according to the data published by International Energy Association IEA statistics (http://www.iea.org), Poland is within the top ten coal producers in the world, with its average annual production of ca 140 Mt. Most of the global production of coal is dedicated to the local state consumption; only around 15% is destined for the international coal market (http://www.worldcoal.org). The contribution of the coal industry to ambient air pollution is a well-known issue, although it is mainly discussed with respect to coal-fired power plants due to emission of gaseous pollutants (such as SO_2, NO_x, CO_2, hydrocarbons) as well as

© 2017 Tous droits réservés.

particulate ones, e.g. coal dust and fly ash. There are also plentiful studies on the role of coal mining in increasing air pollution caused by particulate matter emission (Ghose & Majee 2007; Palmer et al. 2010; Aneja et al. 2012; Cvetković et al. 2013; Pandey et al. 2014; Roy et al. 2015; Kurth et al. 2015), but they are mainly related to opencast or surface mining, and to a lesser extent to underground mining. Elevated concentrations of suspended particulate matter (PM) caused by coal mining activity can pose an additional threat to human health due to its unusual chemical composition. Coal-derived suspended particulate matter is often rich in trace elements (Pandey et al. 2014; Finkelman 1999; Silva et al. 2009; Li et al. 2012; Smoliński et al. 2014) which may have adverse effect on human health, therefore monitoring of air quality and dust chemical composition in the coal mining sites or in their vicinity is required.

It is noteworthy that publications about an impact of underground coalmines on ambient air quality are significantly less frequent than these about open-pit (surface) mines. Underground mining's influence on the surrounding environment might appear less threatening with respect to air pollution, although it is necessary to recognize environmental cumulative effects, such as land subsidence, destruction of water resources, soil erosion, waste rock dumping etc. (Meng et al. 2009). Coal dust exposure among underground miners have already been investigating by a number of groups (Landen et al. 2011; Esch & Hendryx 2011; Hosgood et al. 2012) but mainly with respect to the risk assessment and epidemiological studies and not to the coal dust composition and morphology. Landen et al. (Landen et al. 2011) concluded that indeed there was an increased risk of mortality associated with cumulative exposure to coal dust, but also with coal rank, probably due to differences in the composition of coal mine particulates.

Coal dust comprises nano- and micrometre sized particles resulting from various industrial processes, such as drilling and blasting, transport and transfer of coal. Particles sampled in the non-diesel coal mines showed bimodal size distributions –

© 2017 Tous droits réservés.

one maximum around 17 to 20 µm and the other of about 5 to 8 µm (Burkhart et al. 1987). Carbon-rich particles in the coal dust samples were classified as of diesel-origin based on their diameter – diesel exhaust aerosol is mostly submicrometer in size, and coal dust aerosol is mostly greater than 1 µm in size (Cantrell & Rubow 1991). It was later confirmed by Birch and Noll (Birch & Noll 2004) by selective collection of coal dust particles according to their size – the size fraction below 1 µm was much more abundant in diesel coal mine than in non-diesel mines. Measurements of elemental carbon (EC) is often used as a surrogate to evaluate a content of diesel particulate matter (DPM) in underground mines since diesel engines are the only source of submicrometer EC in underground mines (Noll et al. 2007). The contribution of EC particles to the coal dust is not negligible, especially in the diesel-operated coalmines, which makes the composition of coal dust very complex.

Dust control and monitoring of miners' exposure is a routine practice in the mine shafts (Ren et al. 2011). According to Yan-qiang et al. (Yan-qiang et al. 2011), the dust concentration in the caving face can reach even 3000 mg/m^3, which is an alarming value. Therefore, research on coal dust – due to extreme risk of human exposure – needs to be carried out. The main directions of the coal dust research, as recognized by Yan-qiang et al. (Yan-qiang et al. 2011) are: (1) dust characterization, (2) the law regulation on coal mine, (3) dust explosion, (4) technology and method of prevention and control of dust. The need for research on coalmine dust is on top of that list. Coal deposits in Poland are exploited mainly in the underground mines. Most of them are located in the southern or southwestern parts of Poland (Silesia region). In this study, we focused on investigation the chemical composition of coal dust particles from the underground coalmine in this area. The particles were analysed individually by means of two microanalytical techniques: SEM/EDX and Raman microspectrometry (MRS). This approach, called single particle analysis (SPA), requires a large number of particles being measured, to ensure statistically reliable

© 2017 Tous droits réservés.

results. Collected data were supplemented by statistical data treatment by means of Hierarchical Cluster Analysis (HCA). Bulk analysis by means of ATR-FTIR and XRD was made to complement microanalysis. The main objective was to determine the elemental and molecular composition of coal dust particles collected at 900 m depth with special attention to the carbonaceous particles of coal origin, their evolution and mixing state. Indeed, individual particle analysis may provide typical tracers related to the coalmine particles that could be used for air quality assessment in mining environment.

5.1.2. Sampling description

Sampling of the coal dust was conducted in the underground coalmine. Sampling sites were distributed near an outtake shaft with a diameter φ 7.5 m and 996 m depth, in the coal seam. According to Philpott (Philpott 2002), the coal from this seam shows the following parameters: the ash content 10.02% to 38.47% (the average 21.71%) and the sulphur content varies from 0.82% to 2.16% (the average 1.27%). The seam shows a changeable morphology because of the coalbed thickness and mullock interlayers. Variability of the ash content is strongly linked with the non-coal rock interlayers, since their presence causes a decrease of the coal calorific value and increase of after-burning residue. Sulphur content varies also quite substantially, but it is not related to the presence of worthless material (mullock); it is likely to originate from sulphides such as pyrite being present in the coal exploited in these coal mine (Sawlowicz et al. 2005).

Dust samples were collected from gravitational deposition along the main gallery, beginning with a spot near the shaft exit (Sample 1) and moving gradually closer to the longwall (Sample 4).

© 2017 Tous droits réservés.

5.1.3. Single particle analysis

For single particle analysis by SEM/EDX, the collected dust samples were suspended in hexane, shaken with vortex and deposited on a silver foil. Each sample was then measured with the SEM/EDS system (Tescan Vega 3 SB) in high vacuum with an acceleration voltage of 10 kV. X-ray spectra from individual particles were collected with an energy-dispersive silicon drift detector (SDD) (Oxford Instruments) in the automatic mode with a help of the INCA software (Oxford Instruments). About 300 particles were analyzed per sample, based on the back scattered electron (BSE) image analysis, for 20 s of acquisition time and at magnification of 140x. Such measurement conditions are suitable for the determination of low-Z elements (starting from Z=6, carbon). Fitting of the spectra was made in Quantitative X-ray Analysis System (QXAS) by linear fitting with fitting model and an elemental library designed for each sample.

The semi-quantitative elemental composition of each particle was calculated with an iterative approximation method based on Monte Carlo simulations with the home-made software (Ro et al. 1999; Szaloki et al. 2000). The semi-quantification procedure provides the results within 10% accuracy between the calculated and nominal elemental concentrations (Ro et al. 2001). The large data matrix containing the elemental composition of each of 300 particles from 4 collected dust samples was subjected to the exploratory and multivariate statistical data analysis.

Single particle analysis was also performed by Raman microspectrometry (RMS). The Labram HR800 spectrometer (Horiba) was used in the experiments. Raman backscattering was excited with 473 nm wavelength laser beam, 5 s acquisition time, 3 accumulations and 600 gr/mm grating (with the center at 1255 cm^{-1}). The beam was focused on the sample surface through an optical objective (×50 Olympus objective with N.A 0.75). The diameter of the laser spot on the sample surface was ~1 µm^{2} for the fully focused laser beam. A total of 150 particles were measured for

© 2017 Tous droits réservés.

each sample. The identification of molecular compounds was performed by comparing measured and reference spectra. Curve fitting of the first-order spectral region characteristic of carbonaceous species (1100– 1700 cm^{-1}) was performed with the software program LabSpec 5 (Horiba Raman software). The fitting of extracted spectra was made by Gauss-Lorentz function after a linear baseline correction following the procedure described by Sadezky et al. (Sadezky et al. 2005). The goodness-of-fit was indicated by the error value <5% between the calculated fit curve and the observed.

5.1.4. Bulk analysis

Bulk analysis of the coal dust was performed by means of attenuated total reflection Fourier- transform infrared spectroscopy (ATR-FTIR) using the pellets made of collected samples. The system used for measurements was the Thermo Nicolet FTIR/FT-Raman Spectrometer (model 670) equipped with Ever-Glo mid-IR source, KBr beam splitter and DTGS-KBr detector. The analysis was controlled by OMNIC spectroscopy software. Samples were measured via ATR on the diamond crystal. The 240 scans of each sample were acquired at 4 cm^{-1} spectral resolution. After acquisition the baseline correction and normalization of the spectra was applied.

The mineral phase qualitative composition in the collected samples was determined by XRD using a Panalytical EMPYREAN X-ray diffractometer with CuKα radiation (λ=1.54128 Å), Ni Kβ filter and PIXcel3D detector. Data were collected in the 2θ angle range from 10 to 90^{o} with a step size 0.0130^{o} and generator settings 35 mA, 40 kV. Phase identification was made by comparison with the ICDD Pdf-4 database.

5.1.5. Statistical analysis

Single particle data from SEM/EDX analysis were imported into The Spectronomy system. Particles were analyzed through queries on particle composition and

© 2017 Tous droits réservés.

clustering using the Ward's hierarchical clustering algorithm. The optimal number of clusters (ONC) was set by calculation of Dindex. Hierarchical cluster analysis HCA (Ward's algorithm) was performed with cut-off labeling set by calculated ONC. HCA was also performed on the Raman spectra data set for each sample. The data pre-treatment covering baseline correction (AsLS) and spectral normalization was made. The final number of clusters was determined gradually by comparing the mean spectra of the clusters via dissimilarity axis.

5.1.6. Hierarchical Cluster Analysis from SEM/EDS

Analysis of dust particles to determine the concentration of light elements, such as carbon, nitrogen and oxygen, is significant to study the chemical behavior of the atmospheric aerosol structure and properties (Osán et al. 2000). It has been confirmed that many environmental particles contain low-Z elements in the form of nitrates, sulfates, oxides, or mixtures including a carbon matrix (De Hoog et al. 2005; Worobiec et al. 2006). The reasoning behind single particle analysis supplementing bulk analysis is the new type of information we can draw from the matrix of a large number of particles characterized individually by their X-ray spectra. With suitable data treatment it is possible to find correlations, groups (clusters) of particles with similar properties (E A Stefaniak et al. 2009).

The semi-quantified spectra of individual particles within each sample were subjected to HCA. Cluster distribution, with a corresponding number of clusters, is presented in the Table 1. The cluster types show high similarity among the samples, although some distinction could be drawn. The most abundant cluster is that composed of a Ca-C-O compound (most likely: calcium carbonate). It is the main component of sample 1, 2 and 3, and it is often present as mixed with silicates (Si-O rich) and aluminosilicates (Si-Al-O rich). Carbon is present practically in every cluster, which proves high mixing ratios with other components, which is quite natural for

© 2017 Tous droits réservés.

coalmine dust. In addition, there are clusters gathering carbonaceous particles, characterized by a high content of carbon and oxygen. These particles are – most likely – pure coal dust, sometimes with accessory minerals, such as silicates and aluminosilicates. The distinction between carbonaceous particles collected in the coal mine, including coal particles and diesel particulate matter (DPM), has already been discussed in the literature (Cantrell & Rubow 1991; Birch & Noll 2004; Noll et al. 2007) but only based on particle's diameter. More conclusive results were derived from the Raman investigation of airborne carbonaceous particles (Escribano et al. 2001; Sze et al. 2001) but it was related only to soot particles of various origins. It will be discussed in detail further in the text. The sulphur content found in the particles agrees with the low sulphur amount given by Philpott (Philpott 2002) for the coal exploited from the seam 388, where sample 1 was collected close to the elevator (shaft exit), where the air influx might influence the composition of dust particles. Ca-C-O rich particles are encountered for all clusters. The 80% of the total particle population in sample 2 belongs with the Ca-C-O–rich cluster (number 3 and 4). Moreover, the low-populated cluster (number 1) containing iron (38% mass fraction) and traces of chromium (3.6%) agglomerated with carbon and oxygen (probably of coal origin). Sample 3 shows a bit stronger variety (minerals typical for the crust composition) and sample 4 is the most homogeneous with the presence of carbon in all clusters. It is practically pure carbon-rich coal dust, with some accessory minerals such as Ca-S rich particles (very likely calcium sulphate), silicates and aluminosilicates

© 2017 Tous droits réservés.

Table 3. Clusters and their abundances in the four samples of the coal dust.

Sample 1			Sample 2		
Cluster	Composition	Abundance [%]	Cluster	Composition	Abundance [%]
1	O(46%) Ca(29%) C(12%) Si(4.6%) Al(2.9%)	15	1	Fe(39%) O(28%) C(24%) Cr(3.6%)	2
2	O(61%) Ca(18%) C(14%) Si(4.5%)	22	2	C(41%) O(39%) Ca(7.5%) Mg(3.5%) Si(2.8%)	18
3	O(56%) S(18%) Ca(11%) C(9.5%)	11	3	C(57%) O(19%) Ca(15%) Mg(2.9%) Si(2.4%)	10
4	O(62%) C(25%) Ca(2.7%) Si(2.5%) Mg(1.9%)	22	4	O(37%) Ca(33%) C(24%) Mg(2.7%)	70
5	C(44%) O(29%) Ca(7.8%) Si(3.8%) Fe(3.6%) Al(2.1%)	19			
6	Ca(56%) O(22%) C(15%) Fe(1.5%)	12			
Sample 3			Sample 4		
Cluster	Composition	Abundance [%]	Cluster	Composition	Abundance [%]
1	O(52%) Si(30%) C(7.8%) Al(3.2%) Fe(2.2%)	12	1	C(65%) O(14%) S(11%) Si(2.3%) Mg(2.3%)	37
2	O(62%) C(20%) Al(4.3%) Si(3.9%) S(3.6%)	13	2	C(66%) O(13%) Al(2.7%) Si(2.5%) Mg(2.4%)	34
3	O(44%) Ca(35%) C(12%) Si(3.9%) Al(2.7%)	7	3	C(57%) O(24%) S(7.1%) Si(4.1%) Al(2.5%)	29
4	Fe(40%) O(18%) C(16%) Cr(12%) Al(5%) Ni(4%)	5			
5	O(40%) C(37%) Ba(4.7%) Ca(3.6%) Si(2.9%) Al(2.9%)	24			
6	C(64%) S(10%) O(5.8%) Ca(4.8%) Mg(3.3%) Al(2.2%)	38			

High content of Ca-rich particles (most likely in the form of $CaCO_3$) in samples 1 and 2 is a result of technical operations in the mineshaft in order to minimize the negative influence of the acidified environment. The walls in the seams are coated with calcium carbonate suspension. In samples 1 and 3 we can observe Ca-S-O structures, which are related to the presence of calcium sulphate – oxidized sulphur from coal (most likely from pyrite) reacts with calcium carbonate forming calcium sulphate. Iron is also observed in both samples 2 and 3 – it is associated with carbon and oxygen, but also traces of other metals such as chromium and nickel. The origin of iron in coal dust particles is probably due to a presence of pyrite, that is very common in coal and is an important contributor to Black Lung Disease (Huang et al.

© 2017 Tous droits réservés.

2005). The presence of pyrite in the coal (as well as in the waste dump rock) exploited in LWB has been confirmed (Sawlowicz et al. 2005).

5.1.7. Raman microanalysis

RMS has been shown to be a powerful and versatile technique for determining molecular composition of single particles and describing chemical heterogeneity (Sobanska et al. 2006; Sobanska et al. 2014). RMS spectra were processed for background correction and compared to the reference data. Fig. 4 contains the Raman spectra for the most abundant compounds recognized by RMS. Carbon, calcite and gypsum were typical for all the four examined samples, but their contribution to the coal dust molecular composition was different, which is consistent with the SEM/EDS results. As expected, the molecular species recognized by RMS were $CaCO_3$ (calcite, based on the Raman band 1080 cm^{-1}) and $CaSO_4·2H_20$ (gypsum, 1001 cm^{-1}or 1007 cm^{-1}). Supplementary compounds associated with EDS results such as hematite (Fe_20_3), magnetite (Fe_30_4), quartz (SiO_2) were characterized by RMS in the collected samples. It is noteworthy that pyrite was not identified in the samples by RMS. The Fe-S association was not detected in the sample excluding the presence of pyrite in coal dust particles. This is a result of particle oxidation in air leading to transformation of pyrite to iron oxides and sulphate ions found in the collected particles. Weathering of pyrites in coal mines is a well-known process; in the presence of oxygen and humidity, pyrite is oxidized according to the equation (Singer & Stumm 1970):

$$2FeS_2 + 7O_2 + 2H_2O = 2Fe^{2+} + 4SO_4^{2-} +4H^+$$

This reaction is followed by further oxidation of Fe(II) to Fe(III) which in consequence accelerates further oxidation of pyrite with Fe(III) ions acting also as oxidants (Singer & Stumm 1970). Acidity generated in this reaction is very dangerous, especially in

© 2017 Tous droits réservés.

the coalmine waste landfills, full of coal residue and accompanying pyrite, which is the cause for Acid Mine Drainage. Oxidation of pyrite takes place also in the underground mine shafts, therefore a need for applying calcium carbonate as neutralizing agent. As a result, calcium sulphate (gypsum) is formed; this is confirmed by SEM/EDS and RMS in the coal dust particles.

RMS spectra revealed a high level of particles composed of amorphous carbon with its two typical D and G Raman bands. It is noteworthy, that silicon, aluminum and magnesium were also detected by SEM/EDS – they are likely to be components of silicate or aluminosilicates but their spectra were not detected by RMS. Indeed, silicate and aluminosilicates are detected by RMS with difficulties due to their low Raman cross section and the intense fluorescence signal that can be generated by clay minerals.

A closer look to the Raman spectra (Fig. 68) let us observe the species that are not associated with the EDS results. The Raman band at 145 cm^{-1} proves the presence of trace amounts of TiO_2 polymorph, anatase (Eg mode), which is a common crust mineral. The other bands typical for anatase (B1g/395 cm^{-1} and A1g/515 cm^{-1}) are missing in the spectra, but their intensity ratio is very low (A1g/Eg=0.16 and B1g/Eg=0.16) (Yan et al. 2013).

Taking into account the species correspondent to their Raman spectra in the coal dust particles, they were divided into clusters using HCA procedure. The results are presented in Table 4.

© 2017 Tous droits réservés.

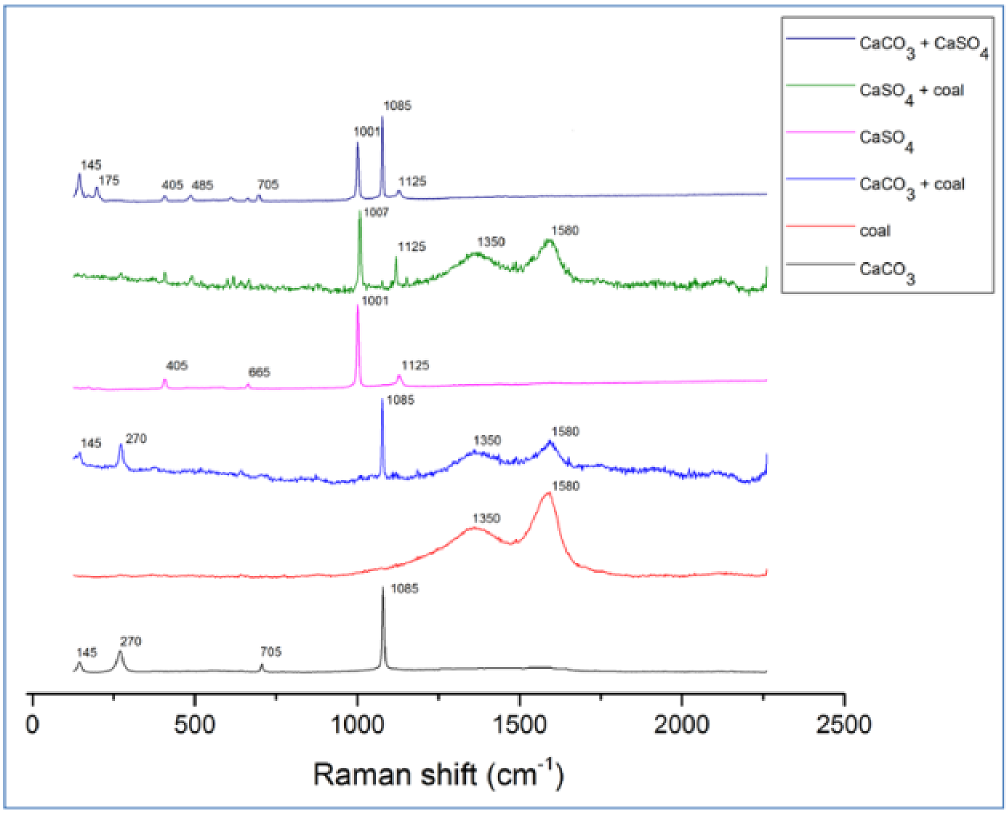
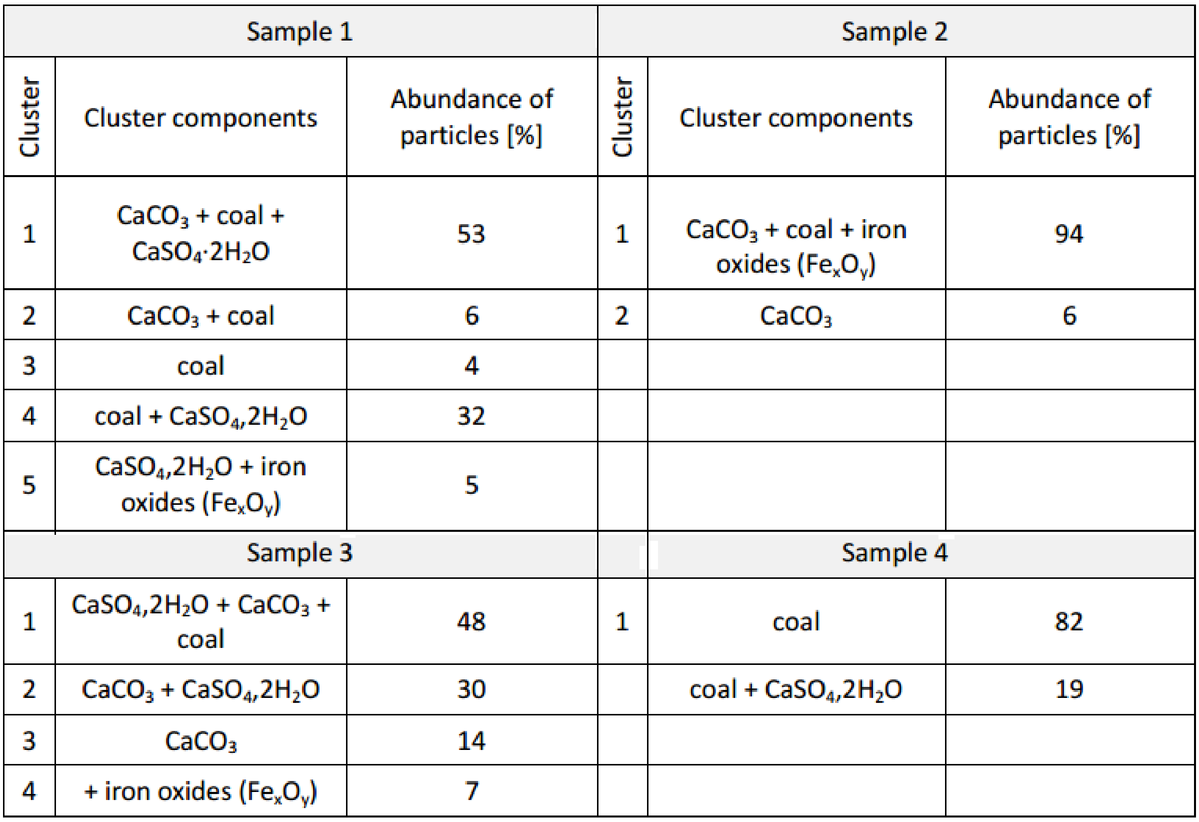


Fig. 68. Raman spectra of the most representative compounds detected in the coal dust samples. The numbers indicate the Raman shift position for the typical features in the Raman spectrum.

© 2017 Tous droits réservés.

Table 4. Clusters and their abundances in the four samples of the coal dust-based on the Raman spectra.



Sample 1			Sample 2		
Cluster	Cluster components	Abundance of particles [%]	Cluster	Cluster components	Abundance of particles [%]
1	$CaCO_3$ + coal + $CaSO_4 \cdot 2H_2O$	53	1	$CaCO_3$ + coal + iron oxides (Fe_xO_y)	94
2	$CaCO_3$ + coal	6	2	$CaCO_3$	6
3	coal	4			
4	coal + $CaSO_{4,}2H_2O$	32			
5	$CaSO_{4,}2H_2O$ + iron oxides (Fe_xO_y)	5			
Sample 3			Sample 4		
1	$CaSO_{4,}2H_2O$ + $CaCO_3$ + coal	48	1	coal	82
2	$CaCO_3$ + $CaSO_{4,}2H_2O$	30		coal + $CaSO_{4,}2H_2O$	19
3	$CaCO_3$	14			
4	+ iron oxides (Fe_xO_y)	7			

Cluster composition derived from RMS is consistent with the SEM/EDX results, even considering the limitations of both techniques. Calcium carbonate (calcite) and calcium sulphate (gypsum) are the main components encountered in samples 1, 2 and 3 and they are mixed with coal in each of the four samples for 91 %, 94 % and 49 % of particles, respectively. Sample 4 is mainly composed of coal-rich particles, but among them, it was possible to distinguish coal particles with sulphates (19%) and without (82%). SEM/EDX results showed also three clusters in the matrix of element abundances (Table 3). The major components of the element-rich clusters are carbon, oxygen (all three), sulphur (cluster 1 and 3) and trace elements from accessory clay minerals (Si, Al, and Mg). It should be pointed out that clustering based on RMS spectra revealed the structures with the strongest Raman scattering (carbon's D and G bands, stretching vibration of sulphate and carbonate ions) while

© 2017 Tous droits réservés.

clustering based on elemental composition divided the particles based on element's abundances. Both of them evidenced the chemical mixing for all particles i.e. one particle containing several compounds. Considering that, SPA is more specific for dust characterization than bulk measurements.

5.1.8. Carbonaceous particles

Coal dust is significantly enriched with carbonaceous particles, both pure (cluster "C-O rich" in Table 2 or "coal" in Table 4) and associated with abundant minerals such as calcite, gypsum, aluminosilicates, or even iron compounds as main species. Carbon-rich particles appeared in all four samples but their abundance is exceptionally high in the sample 4, which is collected at the closest spot to the longwall. Each of the carbonaceous particles exhibits the Raman spectrum with characteristic features in the region 1100-1700 cm^{-1}. In the first order Raman spectrum of carbonaceous materials, the two bands: D (disordered) at around 1350 cm^{-1} and G (graphitic) at around 1580 cm^{-1} are well represented (Fig. 69). These two wide and overlapping bands could be deconvoluted into five bands: D1, D2, D3, D4 and G, each of them representing a different vibration mode (Sadezky et al. 2005; Catelani et al. 2014). According to the model presented by Sadezky et al. (Sadezky et al. 2005), G band corresponds to an ideal graphitic lattice vibration, D1 (denoted as "D" in older publications; main peak at ~1350 cm^{-1}) and D2 (known also as D'; appears as a shoulder on the G band from the higher frequencies' side at around 1620 cm^{-1}) are related to disordered graphitic structure and defects in the graphitic layers, D3 (also known as A band, at ~1500 cm^{-1}) – presence of amorphous carbon and organic species, D4 (at ~1200 cm^{-1}) presence of impurities, inorganics or poly-enes (Sadezky et al. 2005; Catelani et al. 2014). The example of the D/G band deconvolution applied to the carbonaceous particles investigated in this work is shown in Fig. 69.

© 2017 Tous droits réservés.

In order to characterize soot, carbon blacks, graphite and graphite-based materials, coals and other carbonaceous materials, different parameters calculated from the D-G bands have been used so far. In the literature published before the paper of Sadezky's (Sadezky et al. 2005), the authors used only D and G band intensity ratio, without deconvolution. Bacsa et al. (Beyssac et al. 2003) characterized high-temperature treated carbon soot and glassy carbon (reference) with intensity ratio I_D/I_G. Jawhari (Jawhari et al. 1995) used that the dependence between the integrated intensity ratio I_D/I_G and its inversely proportional relationship to the microcrystalline planar size L_a to compare carbon blacks from different producers and microcrystalline graphite as a reference. Escribano (Escribano et al. 2001) and Sze (Sze et al. 2001) investigated carbonaceous aerosols and carbon-containing particles through the integrated intensity ratio D/G and D'/G. In 2003 Beyssac et al. (Beyssac et al. 2003) proposed three parameters for estimation of carbonaceous material degree of organization: D1 and G band positions (wavenumber at the peak center), R1 and R2 ratio, defined as D1/G intensity ratio – R1, and D1/(G+D1+D2) area ratio – R2 and FWHM (Full Width Half Maximum) of G and D1 bands. Sadezky (Sadezky et al. 2005) applied FWHM (full-width-half-maximum), band position and intensity (peak area) ratios of the deconvoluted bands, such as D1/G, D2/G, D3/G and D4/G. Antunes et al. (Antunes et al. 2006) performed a comparative study of first- and second-order Raman spectra of multi-walled carbon nanotubes (MWCNT) at visible and infrared laser excitation. As comparative materials the authors used carbon fibers, powdered graphite and highly ordered pyrolytic graphite (HOPG); the parameters applied for comparison were: relative band intensity ratios: D (nowadays: D1) to G and D' to D (with the new symbols: D2/D1) and D, D' and G band positions. They observed that D band downshifts with increasing wavelength and its relative intensity increases. D' band positions is also slightly affected (small downshift) while G band remains practically at the same wavenumber, despite the laser wavelength. Moreover, I_D/I_G also increases with the wavelength, so does $I_{D'}/I_G$

© 2017 Tous droits réservés.

but to a lesser extent.

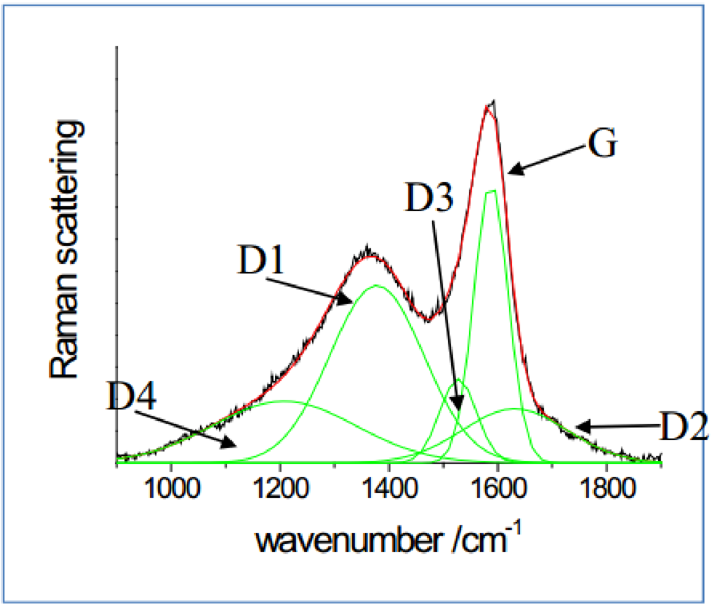


Fig. 69. Typical example for the deconvolution of a first order Raman spectrum of a carbonaceous particle from the investigated coal dust.

To evaluate changes in the chemical structure and reactivity of soot, Ivleva et al. (Ivleva et al. 2007) used D1 FWHM and I_{D3}/I_G to characterize airborne soot and other carbonaceous particles of atmospheric aerosol samples collected in a rural region. On the other hand, to follow structural changes in spark discharge soot (GfG) and light-duty diesel vehicle (LDV) soot upon oxidation, the same group (Markus Knauer et al. 2009) applied the following parameters: D1 FWHM, D3 FWHM and I_{D3}/I_G ratio. In 2008 Rusciano et al. (Rusciano et al. 2008) published their research on SERS study of nano-sized organic carbon particles produced in combustion processes. Owing to the Raman signal enhancement (*ca.* five orders of magnitude), they characterized nano-carbonaceous particles with respect to Di (general symbol for D1, D2, D3 and D4) to G band area ratio. GfG and heavy duty engine (EURO IV) soot were investigated again by Knauer, Ivleva and others (Ivleva et al. 2007; M. Knauer et al.

© 2017 Tous droits réservés.

2009) using – as before – D1 FWHM but also the area ratio of D3/(G+D2+D3), also called as R3 parameter. Soot reactivity was investigated by Schmid, Ivleva and others (Schmid et al. 2011) by means of RMS engaging three different laser wavelengths: 785 nm, 633 nm and 532 nm. They confirmed the relationship described in earlier publications about the decrease of the peak position (frequency, wavenumber) and intensity of D1 band, with G band being fairly resistant (independent on a laser wavelength). Schuster et al. (Schuster et al. 2011) proved (with NEXAFS and XPS) that structural disorder on the surface of diesel soot (Euro IV and Euro VI) is accompanied by a higher amount of oxygen functional groups. Recently, Catelani et al. (Catelani et al. 2014) presented the results of airborne soot characterization via a set of histograms with D1 FWHM and G FWHM. The authors emphasized that among many studies on soot and carbonaceous particles there is a lack of direct comparison of parameters based on D and G bands due to a use of different laser source, which always affects the position and shape of the diffuse D band. Therefore, it was "not possible to make a direct comparison on the absolute FWHM values, but only on the relative ones".

Finally, first order Raman spectrum was also used to characterize high-rank coals, Pre- Cambrian and Carboniferous coal samples, and carbonized anthracites (Marques et al. 2009; Kwiecinska et al. 2010; Rodrigues et al. 2011). The authors used D1 FWHM vs. D1 position, G FWHM vs. G position, I_{D1}/I_G, and even "G/all" area ratio. Hu et al. (Hu et al. 2015) investigated structural changes along the thermal annealing pathway of nanoporous carbon (NPC) by the following parameters: D, D', A (either known as D1, D2, D3) and G band positions, G band FWHM, I_A/I_G – which does not diverge from the parameters used before by other researchers. However, the position of TPA band (related to transpolyacetylene (TPA)-like structures, also known as D4) around 1150 cm^{-1} was also discussed.

According to the references listed above, the D1-FWHM and D1/G parameters are commonly used to characterize carbonaceous structure since they are strictly

© 2017 Tous droits réservés.

correlated with the degree of crystallinity of the material, and are most sensitive to modifications. These two parameters were calculated from the Raman spectra for all particles in the four samples. The frequency patterns presented in Fig. 6 and Fig. 7 give some clear differences related to the typology of the carbonaceous material. The FWHM values of D1 peaks show a distinct unimodal distribution pattern for the sample 4 with a main narrow peak between 130 cm^{-1} and 160 cm^{-1}. The frequency histogram plots for the other three samples exhibit a large distribution without any clear evidence of a bimodal distribution. The main D1-FWHM modes are about 160 cm^{-1} for samples 1 and 2, and 200 cm^{-1} for sample 3. The distribution for sample 1 shows a significant fraction of particles with D1 FWHM values lower than 150 cm^{-1} for sample 1 and 3 whereas modes above 200 cm^{-1} prevailed in sample 2. The previous studies reported that the D1-FWHM band might be correlated with the degree of crystallinity of the material, the FWHM value increases in amorphous and microcrystalline carbon, due to the more disordered structure. Frequency histogram plots reported in Fig. 70 clearly suggest higher crystalline degree in sample 4 when less ordered graphitic structure is found for samples 1 to 3. Actually, the disordered structure may also occur at the microscopic scale when impurities are embedded within the graphitic structure. This probably results in variation for D1-FWHM that was observed for samples 1 and 3, which contain many mineral impurities.

Fig. 71 presents the distribution of the D1/G band intensity ratio for carbonaceous particles in all the four coal dust samples. The distribution is unimodal for sample 2 and 3, with the highest abundance of the particles with I_{D1}/I_G value around 0.8. Sample 1 exhibits a large distribution of I_{D1}/I_G values, from 0.6 to 0.9. For sample 4 the maximum is shifted down to the lower values.

© 2017 Tous droits réservés.

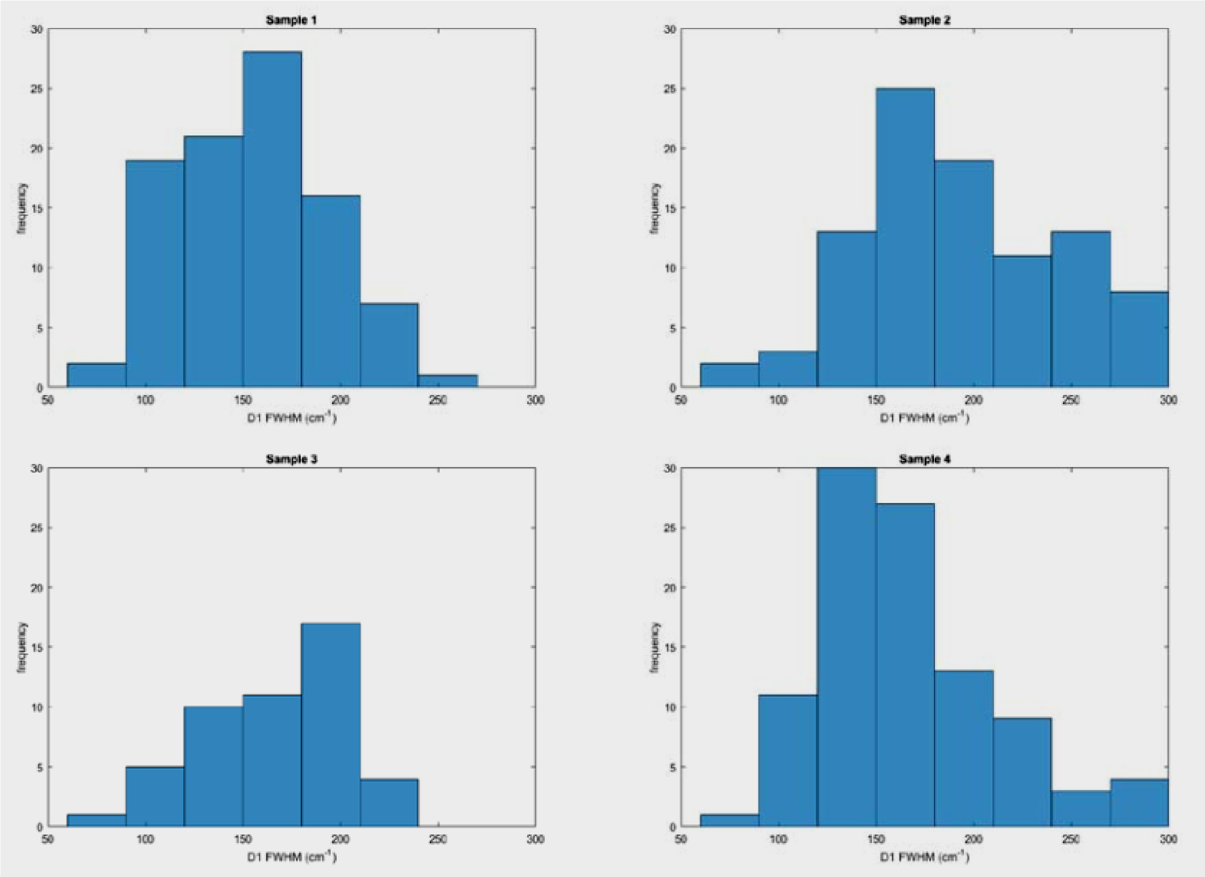


Fig. 70. D1 FWHM patterns for carbonaceous particles in all the four coal dust samples.

© 2017 Tous droits réservés.

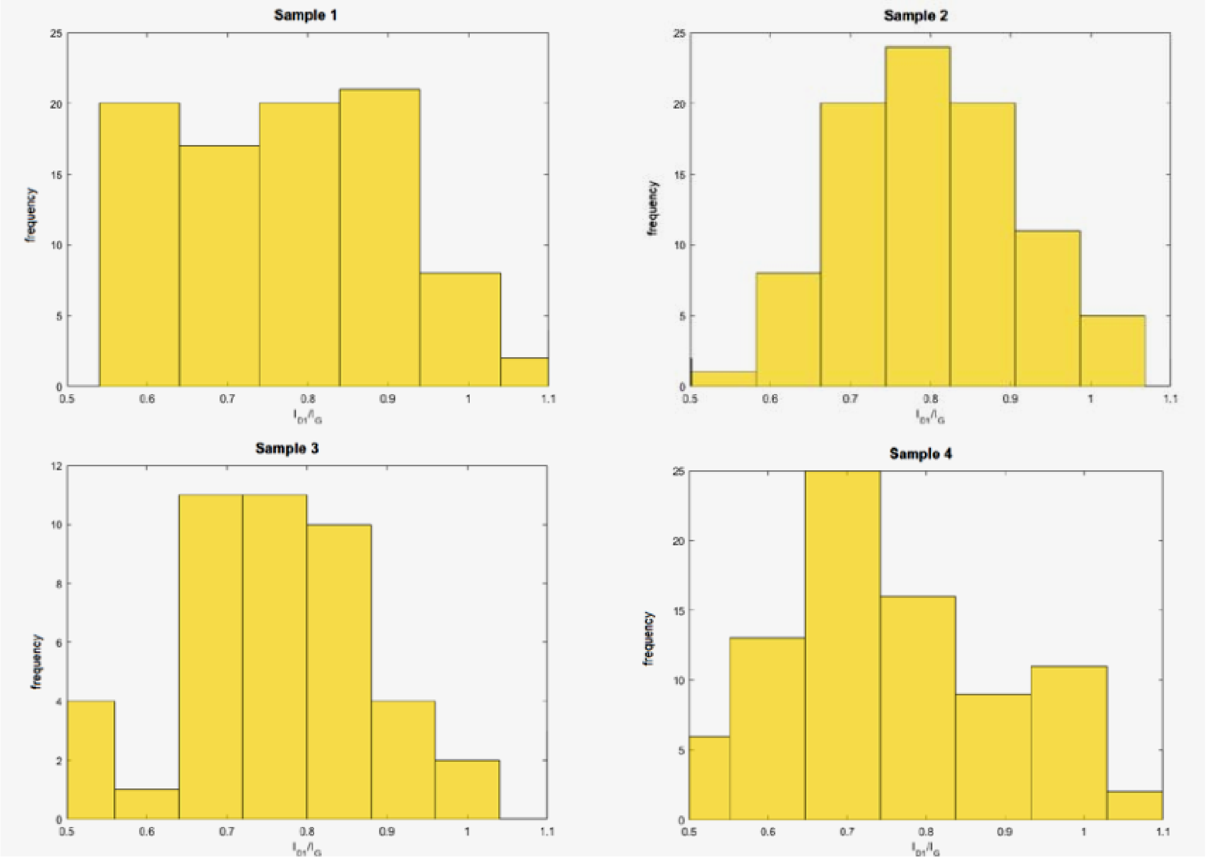


Fig. 71. Distribution of the D1/G area ratio for carbonaceous particles in all the four coal dust samples.

Since the D4 band is linked to the amount of impurities, such as ions, metals, and polyenes in the BC (Sadezky et al. 2005), the parameter D4/G area ratio might be reasonably assumed as a distinctive marker. However, in coal dust samples it seems that no significant results can be extracted from D4/G frequency patterns. The mineral impurities modifying the graphitic structure mainly influence the degree of crystallinity of the material. Finally, internally mixing state is found in most of coalmine dust particles and was highlighted using the D1-FWHM and D1/G frequency patterns. The attempt of identifying possible tracer of coalmine dust particles from our data suffers of the lack of comparison with other coalmine samples from different source. The frequency histogram plots were successfully used for

© 2017 Tous droits réservés.

tracing atmospheric carbonaceous particle aging (Catelani et al. 2014) and could be adapted for source identification of coal-mine particles.

5.1.9. FTIR and XRD

Fig. 72 shows ATR-FTIR spectra after baseline correction made in OMNIC software. The mixture of carbonate-rich particles with silicate minerals, sulphates and water was observed. The significant IR absorption bands referring to the carbonate group ($CO3^{2-}$) were observed in spectra at ~1405 (ν3, asymmetric stretching), 871 (ν2, out-of-plane bending), 1085 (ν1, symmetric stretching) and 712 cm^{-1} (ν4, in-plane bending) which probably corresponds to $CaCO_3$-rich particles. Such particles belong to the first three samples, where the presence of all bands is confirmed. However, in sample 3 due to the presence of strong C-H bending band (organic compound) at 1455 cm^{-1}, the ν_3 mode of carbonate ion (with the strongest intensity) is barely noticeable. The IR absorption bands for sulphate group (SO_4^{2-}) were observed at ~1115 (ν_3, asymmetric stretching) and 613-675 cm^{-1} (ν_4, asymmetric bending) for sample 1 and sample 3. All spectra contain bands, which indicate the presence of silicate minerals (e.g. aluminosilicates, quartz). The bands at 779 and 797 cm^{-1} are common for Si-O-Si stretching in quartz, and band at ~750 cm^{-1} is corresponding to AlO_4 group. Furthermore, the Al-OH bending, which is characteristic for clay minerals, is observed at ~910 cm^{-1}. The other bands of silicate minerals are related to Si-O vibrations and they were observed at ~1005, 1034, 1085 cm^{-1}. The presence of C=C and C-H stretching was observed at 1620 and 2860-2920 cm^{-1} respectively. It is noteworthy that the presence of organic material on graphitic material was evidenced only using ATR-FTIR. This associated organic species may results either from the aging process of coal particles or from aggregation with soot particles emitted by the machine used for coal extraction. The low-intense water bands are

© 2017 Tous droits réservés.

located within the 3300-3700 cm^{-1} for all samples likely due either to the low water adsorption on hygroscopic species or to structural water (e.g. gypsum).

The spectrum of sample 4 is significantly different from the others. The major species of sample 4 are silicate minerals, which are supported by their IR peaks at 750, 833, 910 and 1003 cm^{-1}, and low-intense OH stretching mode (3695 and 3691 cm^{-1}).

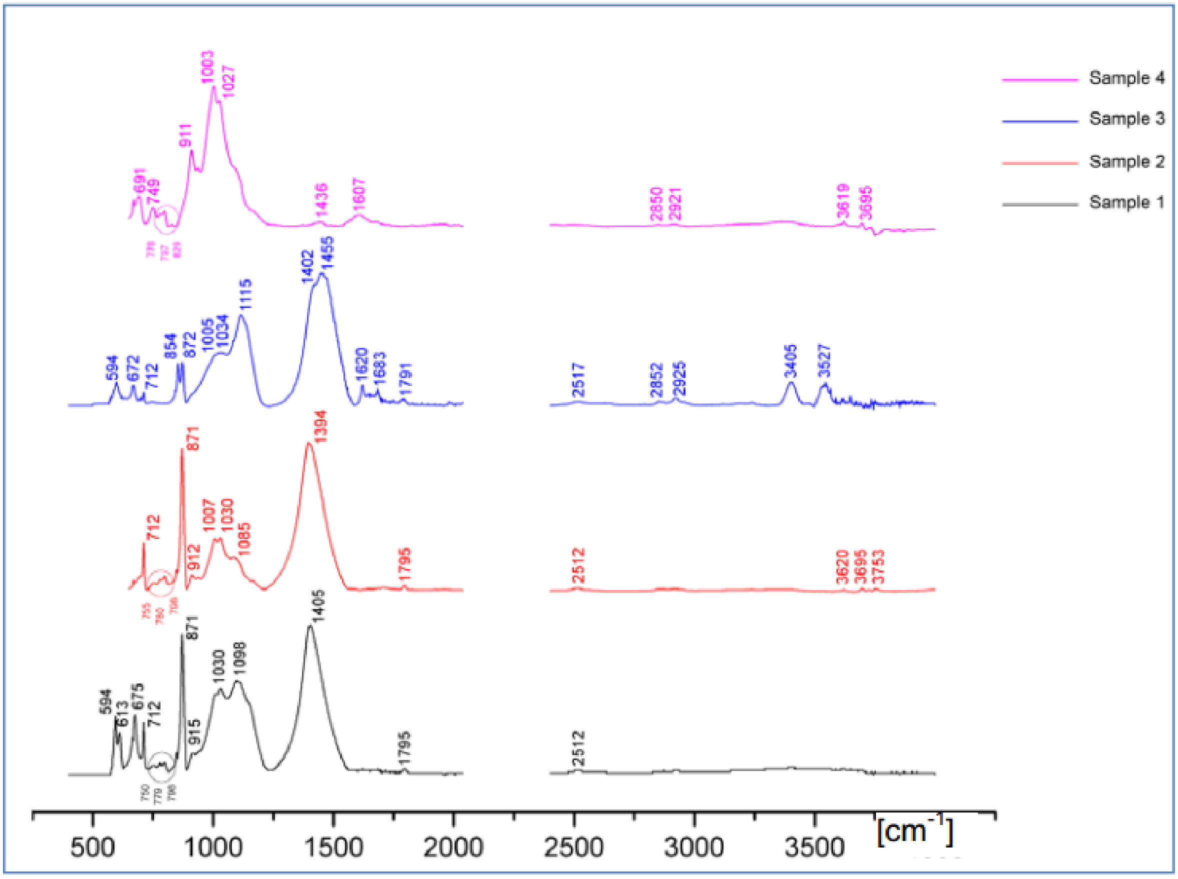


Fig. 72. ATR-FTIR spectra of the coal dust samples.

The results of XRD measurements are generally in compliance with the ones presented above. The software search through the XRD pdf-4 database revealed the presence of the following phases: sample 1: calcite $CaCO_3$, gypsum $CaSO_4{\cdot}2H_2O$ and quartz SiO_2; sample 2: calcite CaCO3, lo_w- magnesium calcite $(Mg_{0.06}\ Ca_{0.94})CO_3$ and

© 2017 Tous droits réservés.

quartz SiO_2; sample 3: calcite $CaCO_3$, gypsum $CaSO_4{\cdot}2H_2O$, and quartz SiO_2; sample 4: quartz SiO_2, aluminosilicate-type $Al_2Si_2O_5(OH)_4$, gypsum $CaSO_4{\cdot}2H_2O$. The XRD spectrum of sample 4 exhibits also a broad band in the 2θ range 20-25^o which can be attributed to d_{002} of coal particles, that are present as a majority in this sample (see Table 3 and Table 4).

5.1.10. Conclusions

The coal dust samples, collected in an underground coal mine (ca 900 m underground) were examined by means of four spectroscopic techniques in combination; two of them suitable for microanalysis such as SEM/EDS and RMS and the other two used for bulk analysis – ATR-FTIR and XRD. Four samples were collected in the main shaft at different distance from the elevator pitch (the exit). All four samples contained large quantity of carbonaceous particles, together with calcium carbonate, calcium sulphate, silicates, aluminosilicates and iron oxides. Carbonaceous particles were found to be both associated with minerals as internal or external mixture. Particles composed of carbon, oxygen and – in the minority – sulphur were attributed as "pure", i.e. coal-originated carbonaceous material. They were characterized by the D1 and G bands in the Raman spectra, which revealed their similarity with respect to the D1/G band intensity ratio and D1-FWHM bands for moderately ordered graphitic structure. These values are characteristic for the coal type explored in this underground mine, therefore it can be used as a tracer for carbonaceous particles in further investigation of ambient aerosols. The distance from the exit was significant with respect to the sample composition and the mixing state – the abundance of coal-originated particles increased with the distance from the exit. Both microanalytical and bulk techniques appeared compatible and complementary, especially for the species detected by only one technique (titanium dioxide) or two (clay minerals).

© 2017 Tous droits réservés.

5.2 Raman analysis of biogenic particles – pollen grains

Primary biological aerosol particles (PBAP) are emitted from vegetation and by other living organisms. PBAP include pollen grains, fungal spores, bacteria, viruses, cell fragments, and protozoans (Després et al. 2012) and they are ubiquitous in the atmosphere (Gregory 1961, Womack et al. 2010). The main research interest involves PBAP effects on humans, animals and agriculture (Waggoner 1983, Burge 1990), the environmental processes they contribute to, e.g. ice and liquid water cloud droplet activation (Després et al. 2012, Morris et al. 2013), and atmospheric chemistry (Deguillaume et al. 2008, Vaitilingom et al. 2013).

Pollen causes allergic symptoms in humans (Schappi et al. 1999, Barnes et al. 2000, Simon-Nobbe et al. 2008), in particular during spring, when pollen concentrations are typically the highest. Recently, there has been a growing interest in studying the impact of bioaerosols on cloud formation and precipitation (Möhler et al. 2007, Pöschl et al. 2010, DeMott et al. 2011, Morris et al. 2011). Pollen has been introduced to global climate models as sources of primary particles (Heald and Spracklen 2009, Hoose et al. 2010, Spracklen et al. 2010, Sesartic et al. 2013). Fungal spores, pollen grains, and their fragments have been shown to nucleate ice at relatively high temperatures in the laboratory, suggesting that these particle classes may contribute to atmospheric cloud formation and evolution if lofted in sufficient numbers (Diehl et al. 2001, Pummer et al. 2012, Haga et al. 2013), and *in situ* measurements at the ground level and in clouds at high altitude have corroborated this possibility (Prenni et al. 2009, DeLeonRodriguez et al. 2013, Huffman et al. 2013, Tobo et al. 2013). There are indications that biological particles could be important for the cloud water cycle, especially in a boreal forest region (Morris et al. 2013, Sesartic et al. 2013). It should be noted, that atmospheric pollutants may have a direct effects on pollen grains: (a) modifications of their biological and reproduction functions: decrease in viability and germination, (b) alteration of a physicochemical characteristics of a pollen surface, (c) change in the allergenic potential, and (d) adjuvant effect

© 2017 Tous droits réservés.

increasing their potential health hazards. Moreover, several works indicate an impact of pollen contaminants on honeybees (genus *Apis*) (McArt et al. 2017; Di Pasquale et al. 2016; de Oliveira et al. 2016). It is generally assumed that bees are exposed to pesticides during crop pollination, yet surprisingly little is known regarding how a focal crop pollen collection is related to a pesticide exposure, how a landscape context influences a crop pollen collection, and whether a magnitude of pesticide risk to bees is at levels warranting concern (McArt et al. 2017).

Pollen grains present several differences at its chemical composition level, for instance in a pollen wall (Duque et al. 2013), that can allow its identification. Therefore, during the last years, a chemical and spectroscopic examination of pollen has become increasingly important and improvements in the use of these techniques have resulted in a reduction of a sample consumption. Raman instrumentation has provided additional impulse for the adoption of Raman spectroscopic techniques in pollen identification, characterization and classification *in situ* without prior preparation (purification, extraction and contrast medium). Additionally, this technique offers high flexibility and good chemical and structural specificity, high spatial resolution, short acquisition times for analysis and can be used in a non-destructive and minimally invasive manner on pollen. The chemical–structural characterization of several pollen grains by Raman spectroscopy has been carried by several authors (Ivleva et al. 2005; Laucks et al. 2000; Schulte et al. 2008), and the works include the vibrational assignments of signals frequently found in Raman spectra of pollen specimens. In addition a comprehensive library of Raman spectra of pollen, which can be regarded as a precursor of a larger pollen database was introduced (Guedes et al. 2014).

In this subsection, in a brief form, we present a potential of the Spectronomy system for differentiation of pollen pellets Raman spectra and a potential impact of such a proceeding on a plant species identification and isolation of contaminated samples.

© 2017 Tous droits réservés.

5.2.1. Sample description and analysis

The pollen pellets were collected directly from the beehives. For the purpose of this subsection, two types of pollen pellets were analysed: (i) contaminated by pesticides (i.e. imidacloprid) and (ii) non-contaminated with pesticides. The pollen pellets without any preparation were analysed by means of the Labram confocal Raman microspectrometer (Horiba, Jobin-Yvon) equipped with a 100×, 0.9 numerical aperture Olympus objective. Raman scattering was excited with the 785 nm wavelength laser. The spot diameter of the laser beam at the sample was 1 µm. The applied system uses a high precision piezo translator and feedback signal to automatically track and adjust the laser focus on the sample - ensuring perfect focus for each measurement. The spectral range was 850 – 1800 cm^{-1}, with spectral resolution ~ 0.4 cm^{-1}. The spectra were acquired for several spots in the pollen pellets. The final data matrices contain: (i) 35 spectra of a pollen pellet with traces of pesticides and (ii) 42 spectra of a pollen pellet contaminated by pesticides. A single Raman spectrum of a pollen pellet spot contains 2476 variables. The pre-processing of Raman spectra was due to the application of AsLS background correction ($\lambda = 10^{-6}$ and p=0.001) and intensity scaling.

5.2.2. Results

The acquired spectra were gathered in the matrices, where each spectrum was located in the single row of the spreadsheet and wavenumbers were located in columns. Such a data structure was required for the Spectronomy system (see more in Chapter 4). The collected spectra are presented in Fig. 73.

© 2017 Tous droits réservés.

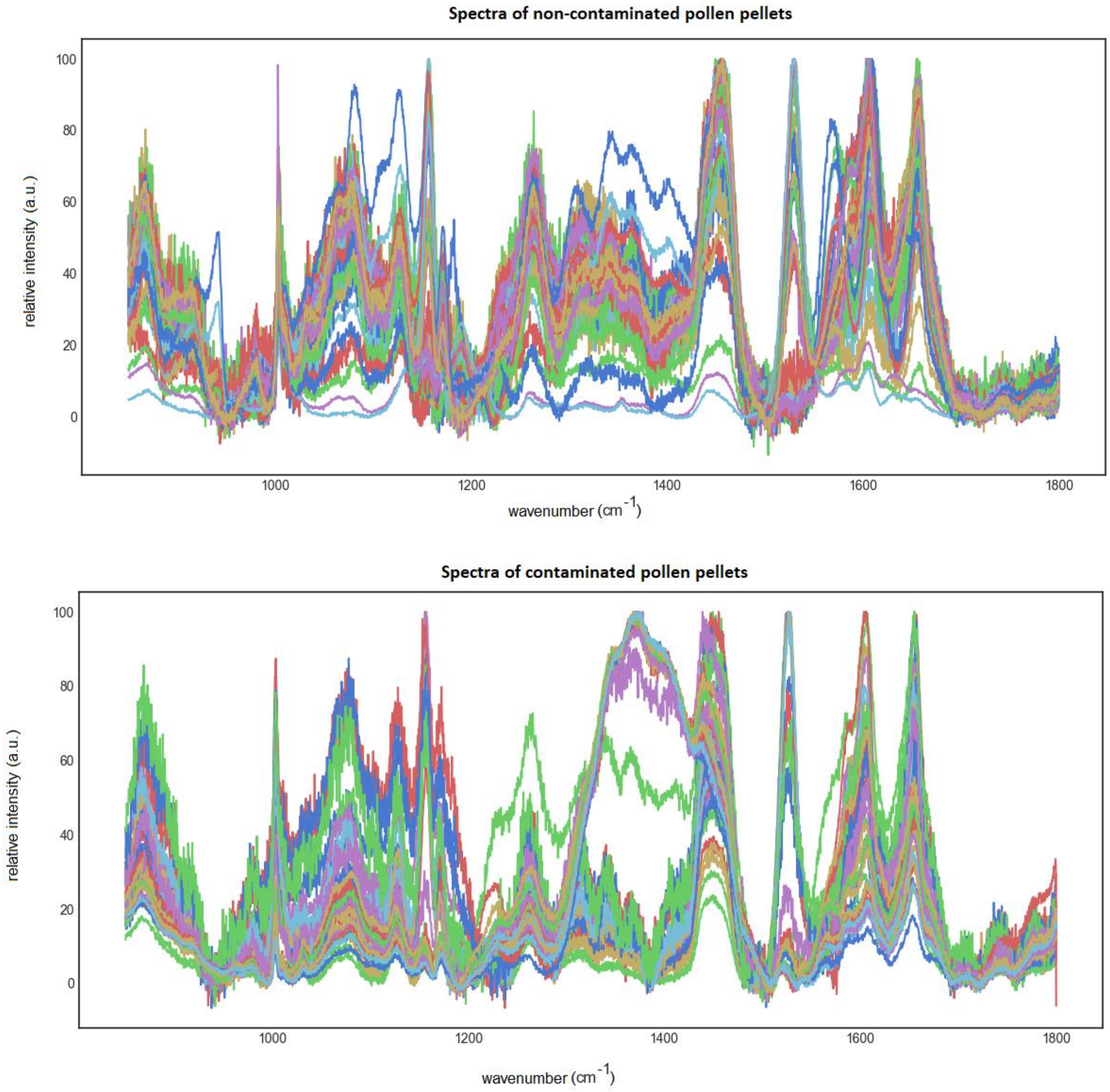


Fig. 73. The Raman spectra of contaminated and non-contaminated pollen pellets.

By visual inspection of the graphin Fig. 73, it is difficult to see the significant difference in some spectral range, which would give unmistakable information about the possibility of grouping the spectra. Thus, for in order to specify an actual number of components in the data set, Ward’s HCA was used.

© 2017 Tous droits réservés.

Non-contaminated pollen pellets

The generated Ward's HCA dendrogram is presented in Fig. 74. The *cut-off* point was designated by visual inspection of the dendrogram and was set for 3 clusters.

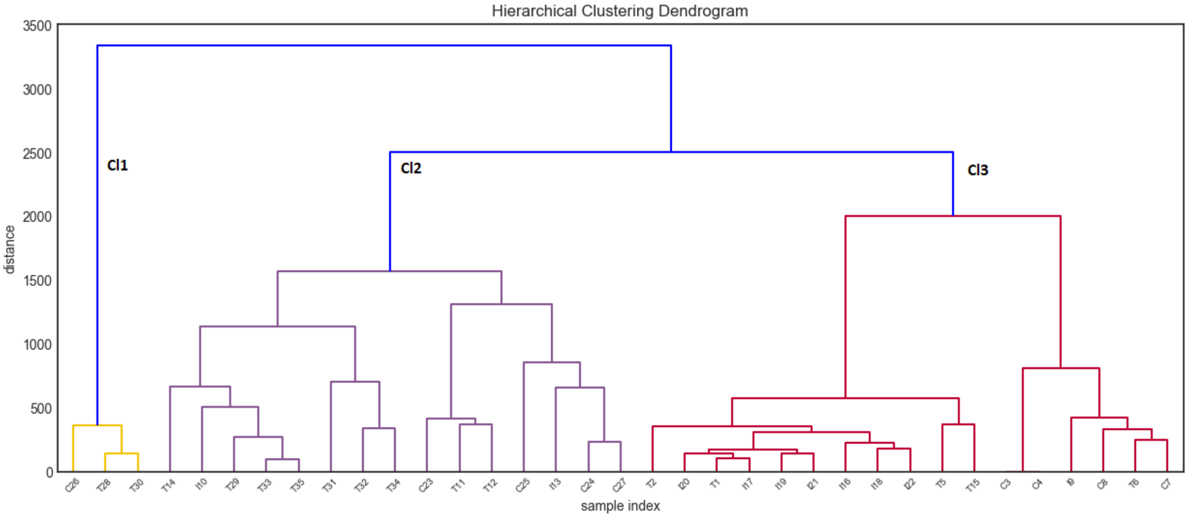


Fig. 74. Ward's dendrogram of non-contaminated pollen pellet spectra.

The function of mean cluster spectra projection was used to visualize the main differences of the clusters and potential momentous areas of spectra for clustering algorithm (Fig. 75).

© 2017 Tous droits réservés.

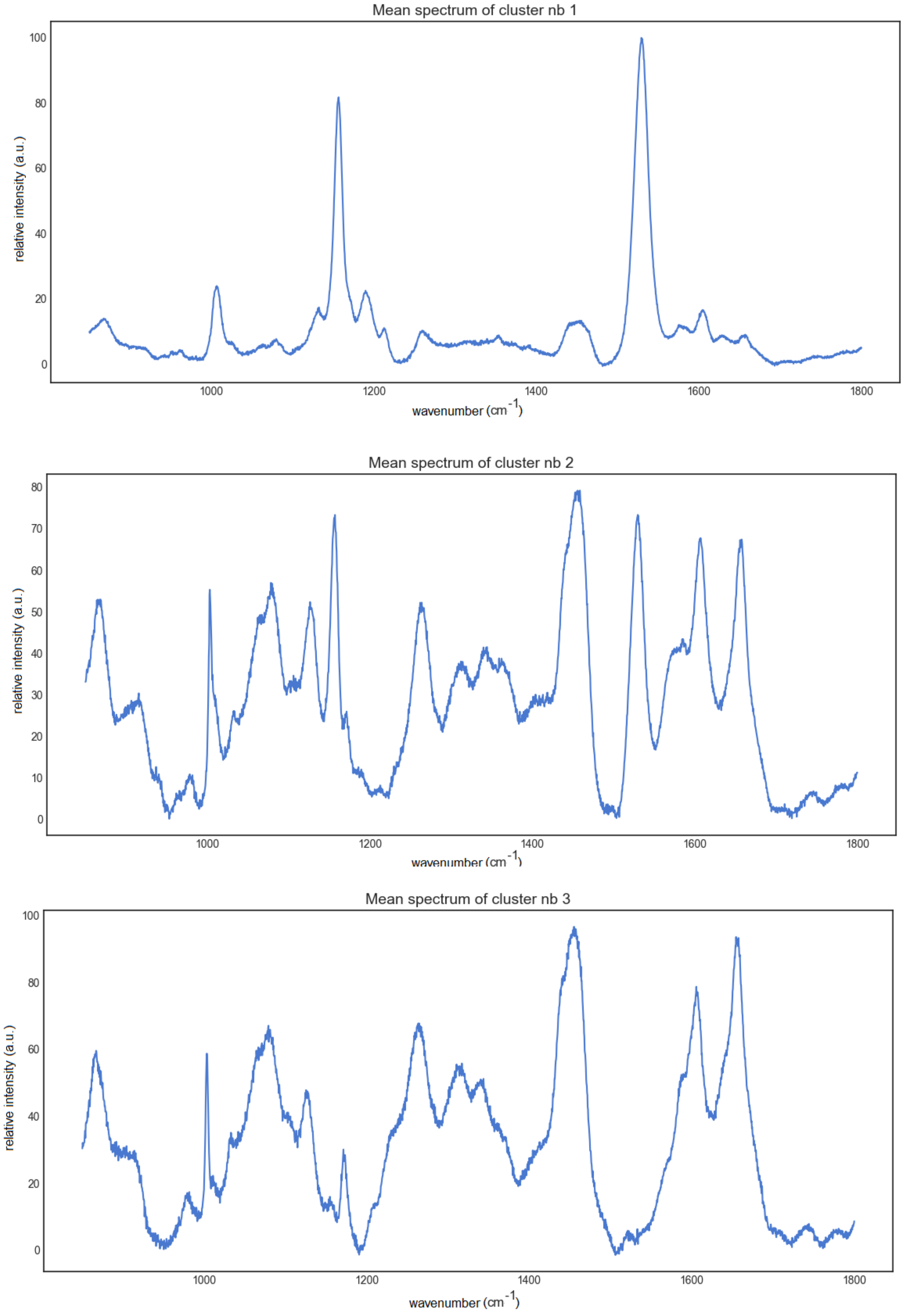


Fig. 75. Mean Raman spectra of Cl1, Cl2 and Cl3 from Ward's HCA performed on the non-contaminated pollen pellets matrix.

© 2017 Tous droits réservés.

In the mean spectrum of cluster 1 the two main peaks with the highest intensity value compared to the other peaks can be distinguished, i.e. ~1530 cm^{-1} and ~1156 cm^{-1}. These peaks also appear in the other mean spectra (Cl2 and Cl3). However, in the Cl3 the intensity value of these peaks is significantly lower than for the other clusters (Cl1 and Cl2). In turn, the higher horizontal level of the peaks intensity can be observed in the Cl2 and Cl3 – probably this is the main factor for the differentiation of the Cl1, which was placed in the separated node. The differentiation of the mean Cl2 and Cl3 spectra is due to the following peaks: ~1186 cm^{-1}, ~1339 cm^{-1} and ~1596 cm^{-1}. The remaining peaks correspond to: carbohydrates (420-440 cm^{-1}), S-S bond in amino acids (540 cm^{-1}), C-O-C glycosidic bond (540 cm^{-1}, 819 cm^{-1}, 1079 cm^{-1}), phosphate (780 cm-1), C-O-P-O-C in RNA (860 cm^{-1}) and C=O bond (stretching) (1650 cm^{-1}). However, the main differentiation of the spectra is observed for the peaks ~1156 cm^{-1} and ~1530 cm^{-1}, which can be assigned to pollen carotenoids (Schulte et al. 2009). Schulte et al. (2009) provides the first evidence of interspecies differences in pollen carotenoid content, structure, and/or assembly between plant species (Schulte et al. 2009). Therefore, it can be assumed that specified clusters are listed for different plant species. Unfortunately, the lack of the complex database of Raman spectra limits the possibility of plant species designation in the presented sample. Nonetheless, this lack of information in this area should be classified as another goal of this work, which should be achieved in the future.

Contaminated pollen pellets

The Ward's HCA dendrogram was generated for the contaminated pollen pellets matrix (Fig. 76). The *cut-off* point was designated by a visual inspection of the dendrogram and was set for 2 clusters.

© 2017 Tous droits réservés.

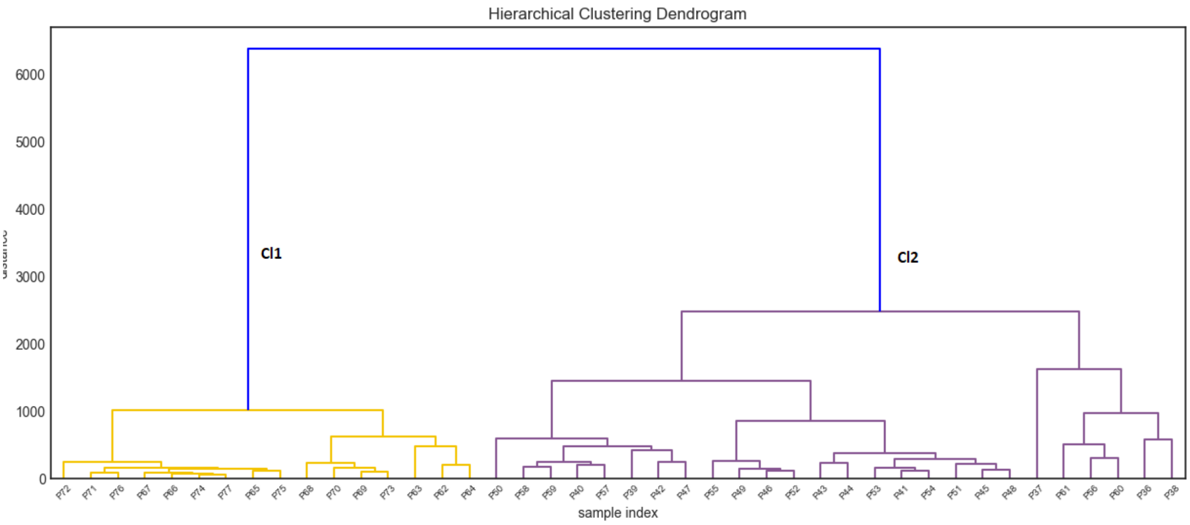


Fig. 76. Ward's dendrogram of contaminated pollen pellet spectra.

As in the case of the previous matrix, the function of mean cluster spectra projection was used to visualize the main differences between the clusters and potential momentous areas of the spectra for the clustering algorithm (Fig. 77).

© 2017 Tous droits réservés.

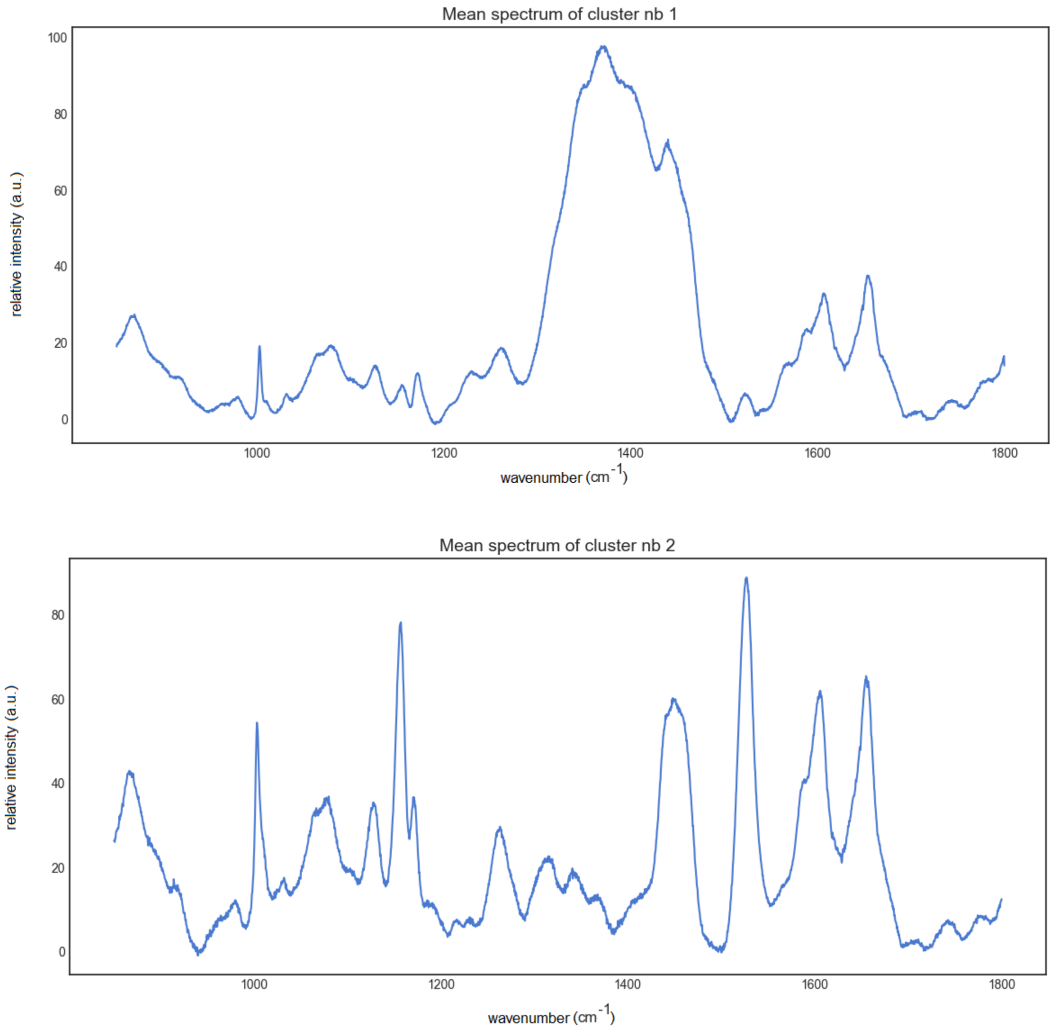


Fig. 77. Mean Raman spectra of Cl1 and Cl2 from Ward's HCA performed on contaminated pollen pellet spectra matrix.

A significant difference between the mean spectrum of Cl1 and the mean spectrum of Cl2 can be observed. In the first example a broad band with its centre ~1340 cm^{-1} (range 1250 – 1500 cm^{-1}) is observed. This band can be characterized by the highest intensity in the Cl1 spectrum. The bands were assigned by the other authors to: ~ 1366 cm$^{−1}$ corresponded to the bending of C–H and O–H bonds in the honey sample (Paradkar & Irudayaraj 2002); ~1460 cm$^{−1}$ was found the signal associated to a combination of the vibration of COO^{-} group of the bending vibration of CH2 group (Kizil et al. 2002; Nickless et al. 2014). This region was attributed to the presence of flavanols and organic acids. Therefore, the spectrum of Cl1 was assigned to the

© 2017 Tous droits réservés.

honey. It should be noted, that pollen pellets collected by honeybees are complex mixtures of biochemical compounds including proteins, saccharides and lipids from the pollens and compounds produced by honeybees themselves (e.g. honey). In turn, the spectrum of Cl2 is similar to the spectra exported from non-contaminated pollen pellets matrix (Fig. 77) and it was assigned to the pollen spectrum.

Differentiation between contaminated and non-contaminated pollen pellets

The matrices of non-contaminated and contaminated pollen pellets were concatenated into the single matrix. This matrix contains 77 Raman spectra collected from the two samples. The spectral pre-processing used before the cluster analysis and multidimensional data analysis was exactly the same as in the case of the previous matrices. The clustering algorithm was Ward's HCA, where the multidimensional data analysis was performed by PCA with a cluster labelling function included in the Spectronomy system. The generated dendrogram is presented in the Fig. 78.

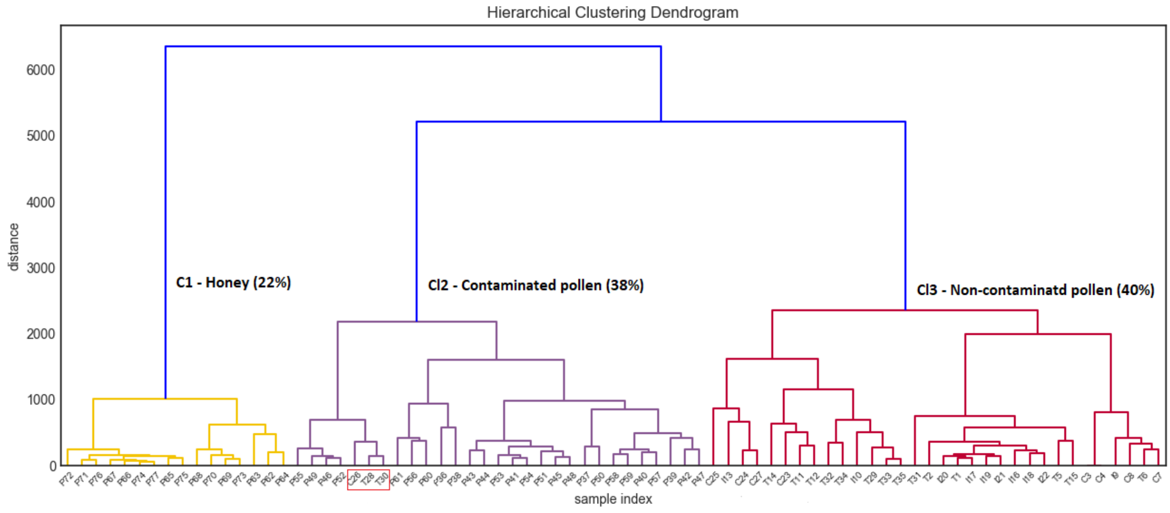


Fig. 78. Ward's dendrogram of contaminated and non-contaminated pollen pellets spectra (single matrix). The red rectangle indicates 3 wrongly clustered spectra of non-contaminated pollen pellets.

© 2017 Tous droits réservés.

The *cut-off* point was designated by visual inspection of the dendrogram and was set for 3 clusters. PCA was performed with cluster labels exported by Ward's HCA (Fig. 79).

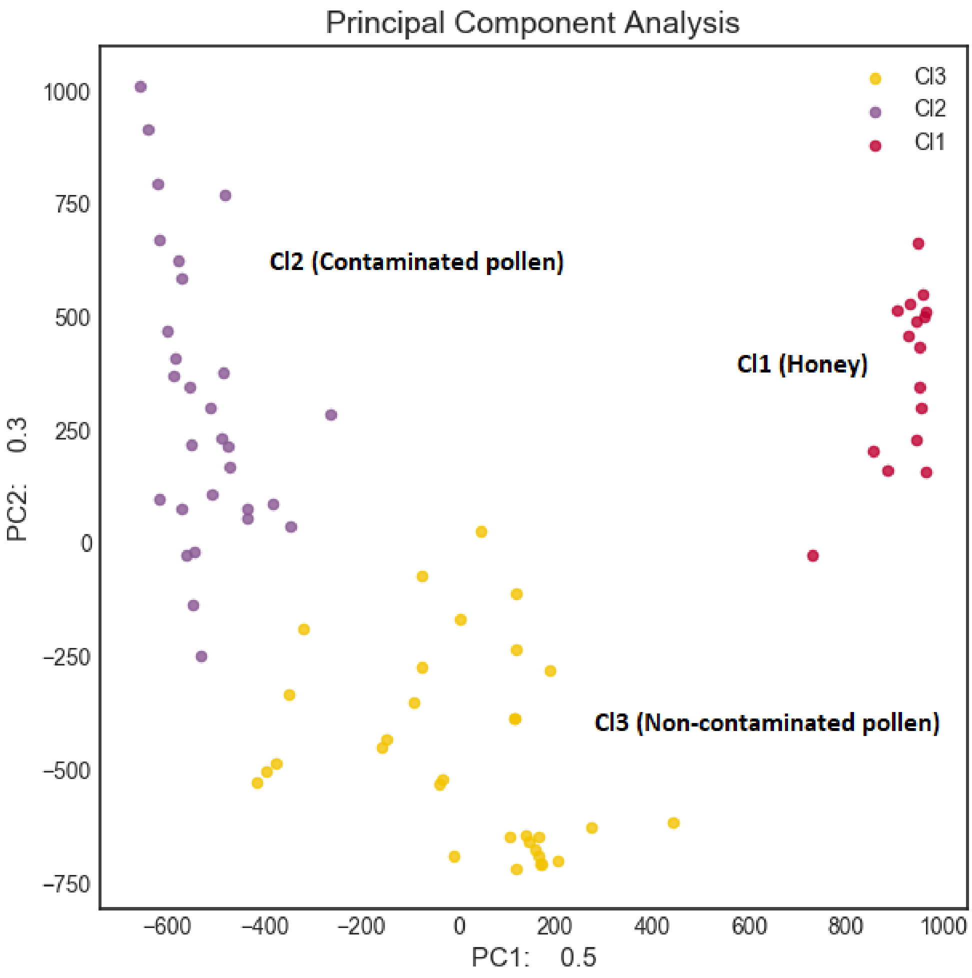


Fig. 79. 2D PCA scatter plot of pollen pellets spectral matrix.

The cumulative variance explained by 2 principal components is ~80% for the spectral pollen pellets matrix. In the 2D PCA scatterplot (Fig. 79), 3 groups can be specified is shown in the dendrogram (Fig. 78). 3 components can be distinguished in the pollen pellets matrix. Each specified cluster reflects the proper matrix structure separated within individual clusters. Cl2 corresponds to the contaminated pollen and constitute 38% of spectra population, where Cl1 contains honey spectra (22%) and

© 2017 Tous droits réservés.

Cl3 non-contaminated pollen spectra (40%). Only three spectra were wrongly clustered by the Ward's HCA algorithm (marked by the red rectangle in Fig. 78).

5.2.3. Conclusions

The preliminary results obtained in this study showed the feasibility of the proposed methodology. We demonstrated that it was possible to differentiate the Raman pollen pellets spectra, which can be assigned to the grouping of the distinct plan species. Thus, the pollen origin in the beehives could be specified. However, this work needs a significant development of the presented methodology, mostly in the preparation of the pollen Raman spectra. In addition, the distinction between contaminated and non-contaminated pollen grains, as well as honey spectra was obtained. The results show a great potential of Raman microspectroscopy coupled with chemometric methods in rapid separation of pollen grains based on their origin and contamination content.

5.3. FTIR analysis of colour organic pigments

For the identification of the pure pigments or colorants - already separated from a binding medium - Fourier Transform Infrared (FTIR) spectroscopy has been established as a effective analytical technique, as such spectra are characterized by very sharp and characteristic multitude of absorptions in the fingerprint region (Lewandowski et al. 2015). Identification of pigments by FTIR spectroscopy is usually processed by comparing the spectrum of an unknown sample with the spectra present in a database, which has to be as accurate as possible. Interpretation of pigment spectra in terms of chemical structural units is rarely processed, since organic molecules normally produce many overlaid vibrations, which can hardly be assigned to distinct molecular structures. Therefore, the spectra of unknown pigment

© 2017 Tous droits réservés.

samples are difficult to be identified if they do not match any of the library spectra. In that case, even a class assignment can be performed only by an expert analysis. Identification of such unknowns might be facilitated by considering characteristic absorption bands; however, this task is rather time-consuming due to a high number of variables and pigments. In this field, a multivariate data analysis has a significant potential, which can improve the process of pigments' recognition. Multivariate classification procedures can be categorized as supervised pattern recognition methods, whereas a variety of different approaches is known. In comparison, unsupervised pattern recognition methods refer mainly to cluster analysis methods, which were described previously. Relating to FTIR spectroscopy, the spectra consist often - depending on a range and resolution - of more than 3000 data points (absorption data per wavenumber). Thus, the specification of momentous variables is crucial for correct specification of a pigment in an unknown sample. It is even more complex to specify an unknown pigment type in mixtures. In addition, a huge limitation may be caused by a low contribution of a signal from the pigment in the FTIR spectrum, due to a background signal (from a deposition substrate or binding medium. It should be noted, that in the non-destructive methods of analysis there is no possibility to extract the pigment from the substrate. Thus, the main idea of the application of FTIR for such a purpose is to classify the signals from the samples based on the chemical composition of the pigments and then identify the average spectrum of the group. The momentous step of such a proceeding is an application of appropriate pre-processing methods before clustering. In that way, a clustering model of a high compliance can be build.

Oriental ink painting, called Sumi-e, is one of the most appealing painting styles that have attracted artists around the world. As a target of presenting approach, the differentiation of the pigments in the Sumi-e paintings is requested. Below we present the procedure for separation of the FTIR spectra from the joint data matrix based on the features included in the Spectronomy analytical system.

© 2017 Tous droits réservés.

5.3.1. Samples description

A set of 6 natural powder pigments: turmeric, dragon's blood, indigo, safflower, cochineal, gamboge (Kremer Pigmente) and one binding medium (rice starch) were prepared. As a deposition substrate, 4 different papers were tested i.e. Whatman, K14, K78 and M20. For the purpose of the current methodology, the specification of each paper is not necessary. Pigments were mixed with water and then deposited with a wooden spatula on the paper. Finally, the paper with deposited pigments was left to dry for at least 12 hours.

5.3.2. FTIR measurements

Three different measurements were conducted on blank papers, pure pigments and pigments deposited on papers by Near FTIR, mid-FTIR and micro FTIR spectroscopy.

In order to collect the spectra in the near infrared range, a portable ARCoptix FT-NIR Rocket equipped with a photodiode InGaAs detector with a working range of 11000 – 4000 cm^{-1} (900 – 2500 nm) was used. The instrument is equipped with an HL2000 halogen lamp (Ocean Optic, 20 watts). The spectra were collected with an optical fibre bundle (Y shaped) which is constituted of 7 optical fibres (fibre core size 400 µm - six illumination fibres around a collecting one). The measurement spot is of approximately 3 mm wide (diameter). The probe was positioned perpendicularly to the surface owing to a clamp standing at the end of an articulating arm, at a working distance ranging between 3 and 5 mm, in order to record the specular reflection component of the reflected light. The spectra were obtained by averaging 30 scans with an acquisition time about 20 seconds and at an 8 cm^{-1} spectral resolution. The instrument was calibrated using a white Spectralon® standard. The software used during the measurements was ARCspectro Rocket (ARCoptix S.A.).

© 2017 Tous droits réservés.

Micro-FTIR spectra were recorded in the mid infrared range by use of a Spotlight 400 Perkin-Elmer microspectrometer equipped with liquid nitrogen cooled mercury cadmium telluride (HgCdTe) detector. The spectra were collected over 200 scans, at a resolution of 4 cm^{-1} using the spectrum from a golden mirror plate for background acquisition. Spectra were collected in the range 500 – 6000 cm^{-1} and expressed as function of pseudo-absorbance (log(1/R)). The software used during the measurements was SpectrumIMAGE (Perkin-Elmer).

Mid-FTIR spectra were recorded by use of an ALPHA FTIR Spectrometer (Bruker) equipped with a DTGS detector and an external reflection module. The spectra were collected over 128 scans, at a resolution of 4 cm^{-1} using the spectrum from a golden mirror plate for background acquisition. Spectra were collected in the range 6000 – 400 cm^{-1} and expressed as function of pseudo-absorbance (log(1/R)). The software used during the measurements was OPUS 7 (Bruker Optik GmbH).

The two instruments were used in the Mid-IR range to evaluate the difference of spectral data processing for FTIR spectra collected with different spot size (micro/macro differentiation).

For optimization of the clustering model and to show its capabilities and limitations, five different types of data sets were collected: (i) near-IR spectra of papers, (ii) near-IR spectra of separated pigments and separated binding, (iii) near-IR spectra of deposited pigments with binding medium on individual paper, (iv) mid-IR spectra (macro FTIR mode (~ 5mm spot size)) of deposited pigments with binding medium on individual paper, (v) mid-IR spectra (µFTIR mode (~0.01 mm^2 spot area)) of deposited pigments with the binding medium on the unique paper.

The applied pigments were supplied from a single provider. The binding medium is corresponding to rice starch in all the examples. Due to the content beyond the assumptions of the present work, pigments have not been mixed together, as well as they not have been mixed with the binding medium.

© 2017 Tous droits réservés.

5.3.3. Results

Distinction between different papers (Near-IR)

The acquisition of 2 spectra of the M20, K14 and K78 was made, beside the Whatman paper where 4 spectra were collected. Spectra were obtained by averaging 30 scans with an acquisition time about 20 seconds and at an 8 cm^{-1} spectral resolution. The instrument was calibrated using a white Spectralon® standard. The several types of pre-processing procedures and their configuration were tested for optimization of the papers spectra separation in the clustering algorithms. The results presenting below are corresponding to the optimal procedure for the pre-processing of spectra, which includes the spectra scaling (intensity scaling) and application of the 3rd polynomial EMSC method for normalization (Fig. 80).

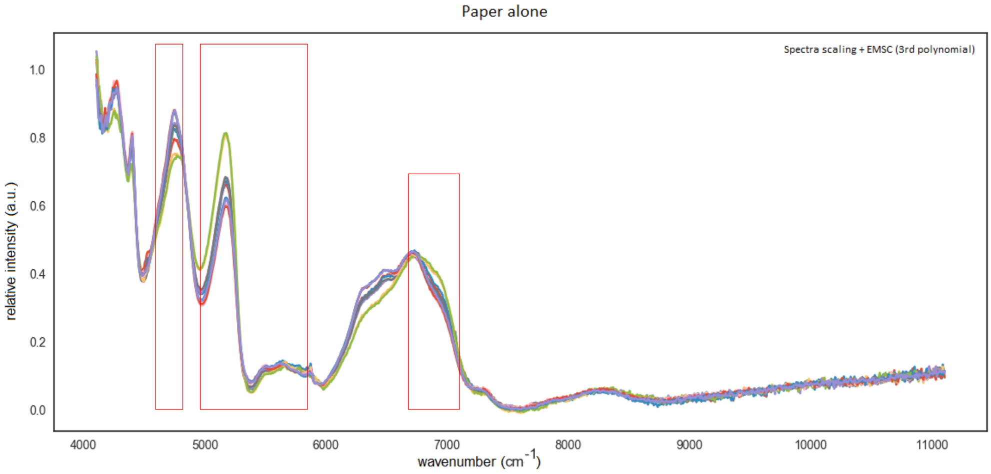
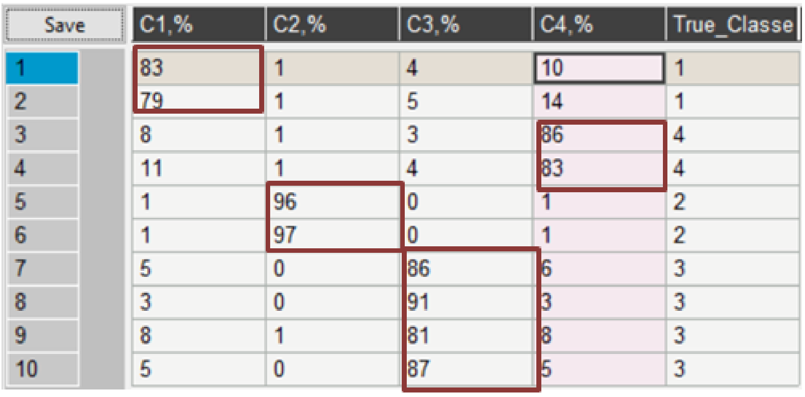


Fig. 80. Near-IR spectra of different papers after pre-processing (spectral scaling and 3rd polynomial of EMSC. (a.u. - arbitrary unit). The red rectangles represent potential momentous spectra areas (designated manually) for clustering.

In Fig. 80, we can visually distinguish the possible momentous areas within several ranges, i.e. ~4250 cm^{-1}, between 4500-5200 cm^{-1}, ~ 6500 cm^{-1} and 7200 cm$^{- 1}$ (see red marks in Fig. 80).

© 2017 Tous droits réservés.

The two different types of clustering (fuzzy c-means and Ward's HCA) as well as unsupervised multivariate data analysis (PCA) were applied as complementary techniques.



Save	C1,%	C2,%	C3,%	C4,%	True_Classe
1	83	1	4	10	1
2	79	1	5	14	1
3	8	1	3	86	4
4	11	1	4	83	4
5	1	96	0	1	2
6	1	97	0	1	2
7	5	0	86	6	3
8	3	0	91	3	3
9	8	1	81	8	3
10	5	0	87	5	3

Fig. 81. Fuzzy-c-means clustering table of near-IR spectra after pre-processing of different types of paper – in the left column the number indicates the number of spectrum and the first row C1 - C4 % indicates the cluster contribution.

In fuzzy c-means clustering (Fig. 81) each cluster is associated with a membership function that expresses the degree to which individual data points (spectra) belong to the cluster. 4 different clusters (C1 to C4) were separated with a significant cluster affiliation of more than 80% for each cluster. The results demonstrate that the separation between the spectra related to the different paper types is substantial, meaning that C1 is clearly associated with paper 1, C2 with paper 2, C3 with paper 3 and C4 with paper 4. However, it should be noted, that the Spectronomy software employs the random initialization seeding in fuzzy-c-means clustering which may change the results. Based on this assumption, another multivariate data analysis procedure in the form of PCA was used (Fig. 82). The results from fuzzy c-means clustering were exported in the form of clustering vector and then used in combination with PCA with a cluster labels function included in the Spectronomy

© 2017 Tous droits réservés.

system. This step was important to demonstrate the accuracy of fuzzy-c-means clustering even for the random initialization seeding.

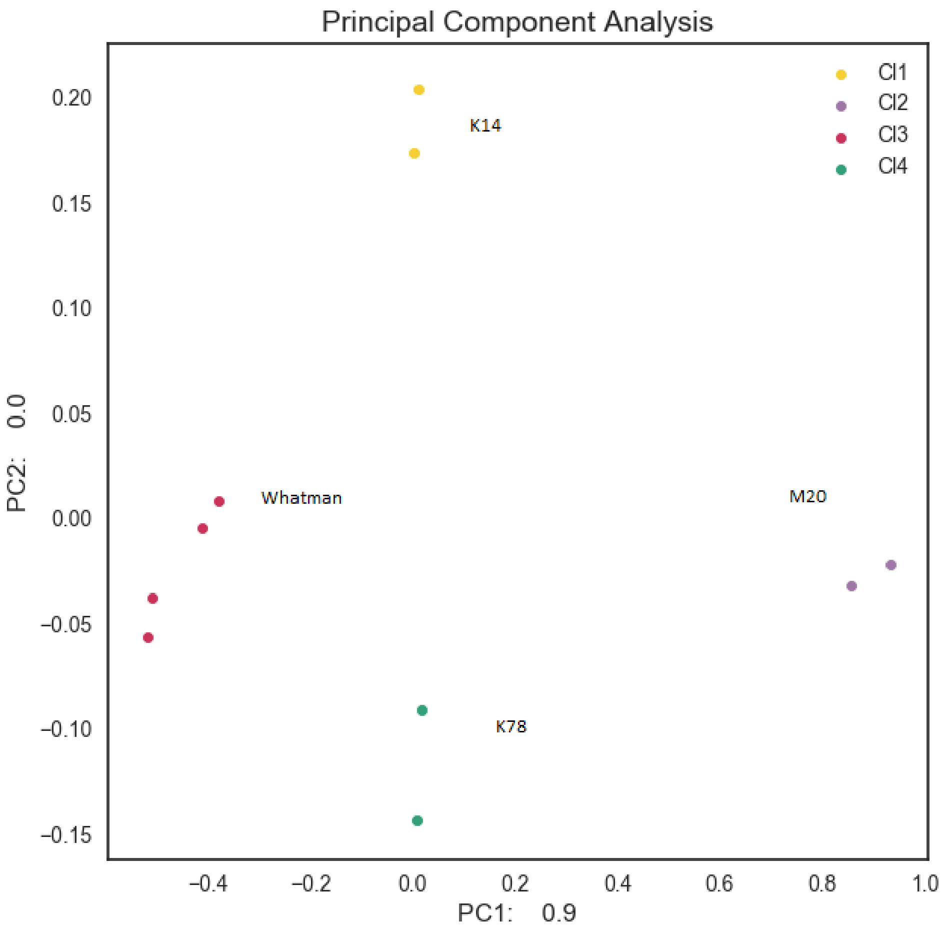


Fig. 82. PCA scatter plot of different papers near-IR spectra after pre-processing with labelled clusters from fuzzy c-means clustering. The papers are associated to the clusters as follows: Cl1 (K14), Cl2 (M2), Cl3 (Whatman), Cl4 (K78).

Total variance explained by PC1 is around 90%. The scores on the scatter plot are well separated (Fig. 83). The separated scores on the scatter plot correspond to the data pattern formation which is consistent with the fuzzy-means clustering results

© 2017 Tous droits réservés.

and thus demonstrates the effectiveness of spectra differentiating. By using the scores from two first PC (A matrix) Ward's HCA was also performed (Liu et al. 2003).

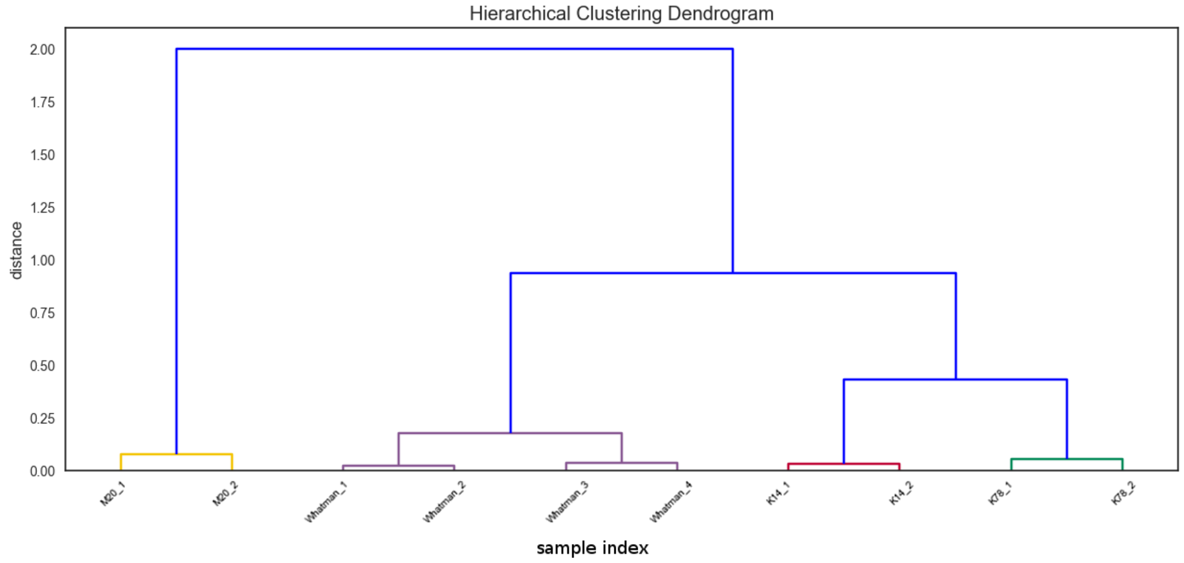


Fig. 83. Ward's HCA dendrogram of near-IR spectra of different papers after pre-processing.

A dendrogram based on hierarchical clustering analysis of the Near-IR spectral data was constructed (Fig. 83), which separated the paper spectra into 4 groups. However, it should be noted, that a more general clustering (into 3 clusters) could also be made, based on the Euclidean distance method of cut-off dendrogram point specification. The papers labelled as K14 and K78 are from the same supplier, despite the fact that their texture and composition differ from each other. Thus, by applying Ward's HCA, the structure and hierarchy of the data in the matrix can be visualized, which in this particular case was inaccessible by both fuzzy c-means clustering and PCA.

The presented results are consistent with the fuzzy c-means clustering with cluster numbers (labels) set automatically during clustering procedure, which may differ. PCA enabled a display of the dissimilarity among papers without prior knowledge. Therefore, PCA and CA (fuzzy c-means and HCA) from the Near-IR spectral data of

© 2017 Tous droits réservés.

papers could be used for rapid discrimination of a different paper type. In addition, Ward's HCA can provide information about more general similarities among the papers, such as the same supplier.

Despite the low number of spectra used in the processing, the separation of different papers is apparent. However, the same model rebuilt in the Spectronomy system may not be fitted ideally for larger matrices, especially with various types of papers, which significantly differ from the presented data. In such a case, a proper evaluation on the studied subject with modification of data analysis procedures (e.g. pre-processing) is needed. However, Near-IR spectroscopy seems to be an appropriate analytical method for a paper comparison.

Distinction between different pigments and binder (Near-IR)

The acquisition of 2 spectra of safflower (carthamin pigment), cochineal (carmine pigment), turmeric (curcumine pigment), gamboge pigment, dragon's blood pigment was made, beside the indigo pigment and rice starch where 4 spectra were collected for each component. Spectra were obtained by averaging 30 scans with an acquisition time about 20 seconds and at an 8 cm^{-1} spectral resolution. The several types of pre-processing procedures and their configuration were tested for optimization of the pigments and binding medium spectra separation in the clustering algorithms. The results presented below (Fig. 84) are corresponding to the optimal procedure for the pre-processing of spectra which include the application of the 3rd polynomial EMSC method for normalization and Asymmetric Least Squares Baseline Correction (AsLS) with lambda = 10^5 and p-value = 0.1 which constitute the optimal values for spectra separation. The optimal values were specified by iterative calculation of the clustering results with different baseline correction parameters starting from the default values parameters in presenting procedure (Baek et al. 2015; Eilers & Boelens 2005) which are: lambda = 10^7 and p-value = 0.01.

© 2017 Tous droits réservés.

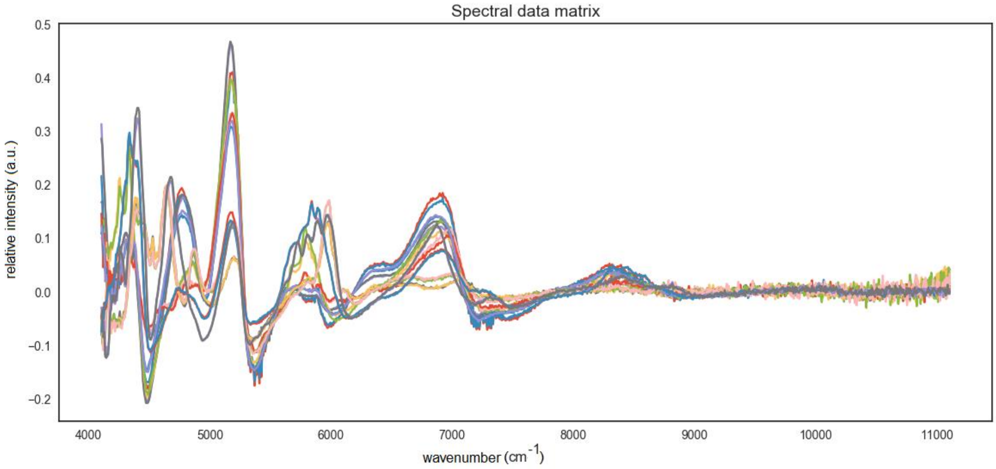


Fig. 84. Near-IR spectra of different pigments and binding medium (rice starch) after pre-processing 3rd polynomial of EMSC and AsLS (lambda = 10^5; p-value = 0.1).

The negative values of the relative intensity were created after 3rd polynomial of EMSC and AsLS baseline correction. These values have no legitimate physical counterpart. However, for the clustering purpose, the presented pre-processing procedure shows the best separation of the spectra. Moreover, after generating clustering vector, the raw Near-IR spectra or spectra after a distinct pre-processing step (intensity scaling only) can be sorted because of clustering results, which may help in identification of unknown samples.

The complementary results of spectra separation in clustering algorithms have been achieved by the 3rd polynomial EMSC method for normalization and Savitzky-Golay 1st derivative (Fig. 85).

© 2017 Tous droits réservés.

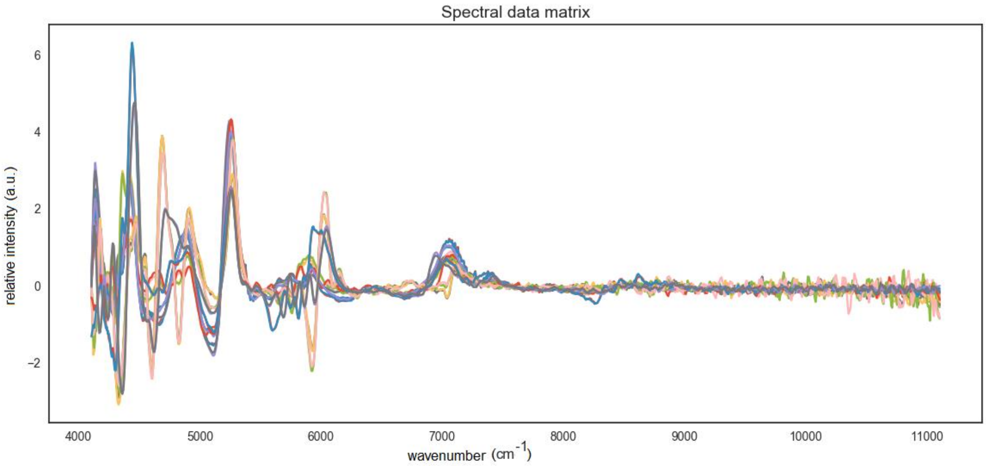


Fig. 85. Near-IR spectra of different pigments and binding medium (rice starch) after pre-processing of 3rd polynomial of EMSC and Savitzky-Golay 1st derivative.

As in the case of the 3rd polynomial of EMSC and AsLS baseline correction pre-processing, the 3rd polynomial of EMSC and S-G 1st derivative pre-processing produced some negative values of the relative spectra intensity. This condition is due to the spectra transformation after the pre-processing procedure. Comparing the spectra after AsLS baseline correction (Fig. 84) with those after EMSC with S-G procedure (Fig. 85) the variety of different spectral ranges can be observed. In the first case, the most noticeable ranges are ~4200 – 4800 cm^{-1}; ~5200 cm-1; 5400 – 5800 cm^{-1} and 6200-6900 cm^{-1}. In turn, in the second case the narrow spectral ranges may be specified, however, in a more complex way: ~4000 – 4300 cm^{-1}; ~4400 cm^{-1}; 4600-4800 cm^{-1}; 4900-5100 cm^{-1}; ~5200 cm^{-1}, 5600-6000 cm^{-1}; ~6100 cm^{-1} and ~7000 cm^{-1}. Due to the simpler way of major spectral ranges determination, the first type of pre-processing is preferred. Following in this subsection, the results of the data processing are presented based on the spectral data matrix after 3rd polynomial of EMSC and AsLS baseline correction. The specification of the number of components in the data set was reached by calculation of the fuzzy partition coefficient, which indicates the correct number of clusters (Fig. 86). This parameter

© 2017 Tous droits réservés.

was calculated for the spectral data set after 3rd polynomial of EMSC and AsLS baseline correction pre-processing.

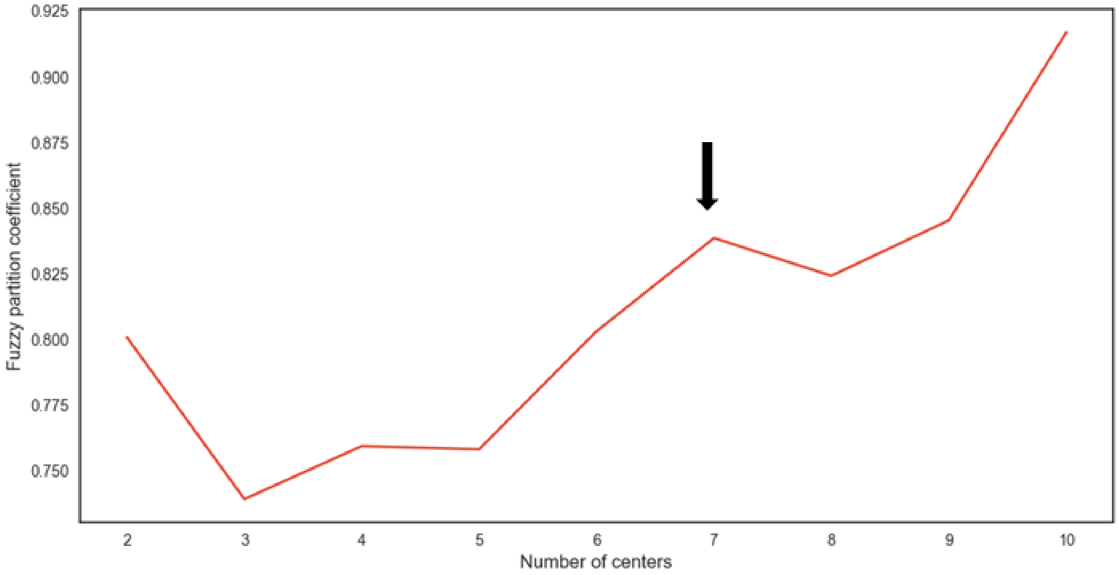
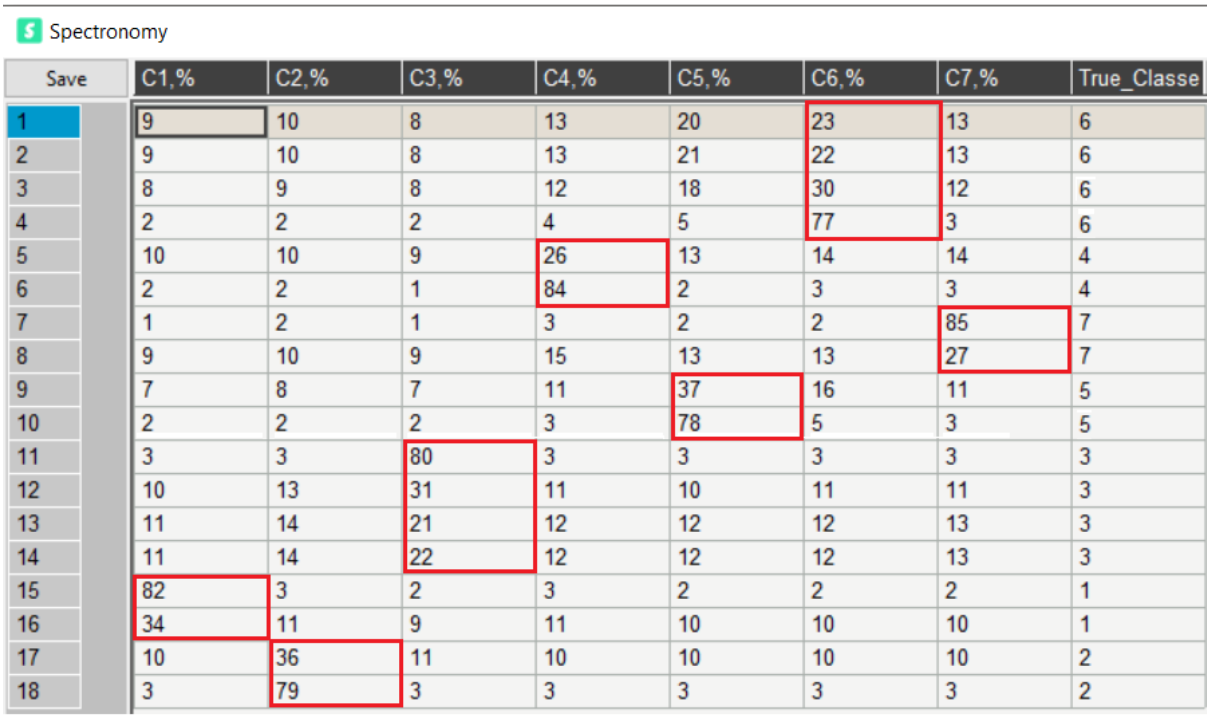


Fig. 86. Fuzzy partitioning coefficient calculated from the data set with pigments and binding medium (black arrow indicates the optimal cluster number value).

The fuzzy partitioning coefficient indicates the correct number of components in the data set corresponding to 6 pigments (safflower, gamboge, Dragon’s blood, turmeric, indigo, cochineal) and 1 binding medium (rice starch). In the results from fuzzy c-means clustering, each from 7 clusters was associated with a membership function of the spectra (Fig. 87).

© 2017 Tous droits réservés.



	C1,%	C2,%	C3,%	C4,%	C5,%	C6,%	C7,%	True_Classe
1	9	10	8	13	20	23	13	6
2	9	10	8	13	21	22	13	6
3	8	9	8	12	18	30	12	6
4	2	2	2	4	5	77	3	6
5	10	10	9	26	13	14	14	4
6	2	2	1	84	2	3	3	4
7	1	2	1	3	2	2	85	7
8	9	10	9	15	13	13	27	7
9	7	8	7	11	37	16	11	5
10	2	2	2	3	78	5	3	5
11	3	3	80	3	3	3	3	3
12	10	13	31	11	10	11	11	3
13	11	14	21	12	12	12	13	3
14	11	14	22	12	12	12	13	3
15	82	3	2	3	2	2	2	1
16	34	11	9	11	10	10	10	1
17	10	36	11	10	10	10	10	2
18	3	79	3	3	3	3	3	2

Fig. 87. Fuzzy c-means clustering table of Near-IR spectra after pre-processing of different types of pigments and binding medium.

7 different clusters of spectra were separated with a diverse cluster affiliation of spectra (in the range of 21-85%) (Fig. 87). Complementary multivariate data processing algorithms were used: (i) Ward's HCA; (ii) k-means clustering and (iii) principal component analysis with score labelling based on k-means clustering results. The corresponding parameters for specification of intrinsic component number, beside the fuzzy partitioning coefficient are Dindex and Hubert's index. Both parameters were calculated for both the Ward's HCA and k-means clustering algorithms. The Dindex value for Ward's HCA was set for 4 clusters where Hubert's index determines 6 clusters as an optimal value (Fig. 88). In the case of k-means clustering,the Dindex value is corresponding to the value of 7 clusters, as in the case of the fuzzy partitioning coefficient (Fig. 89), anyhow the Hubert's index indicates the 5 clusters as the optimal number.

© 2017 Tous droits réservés.

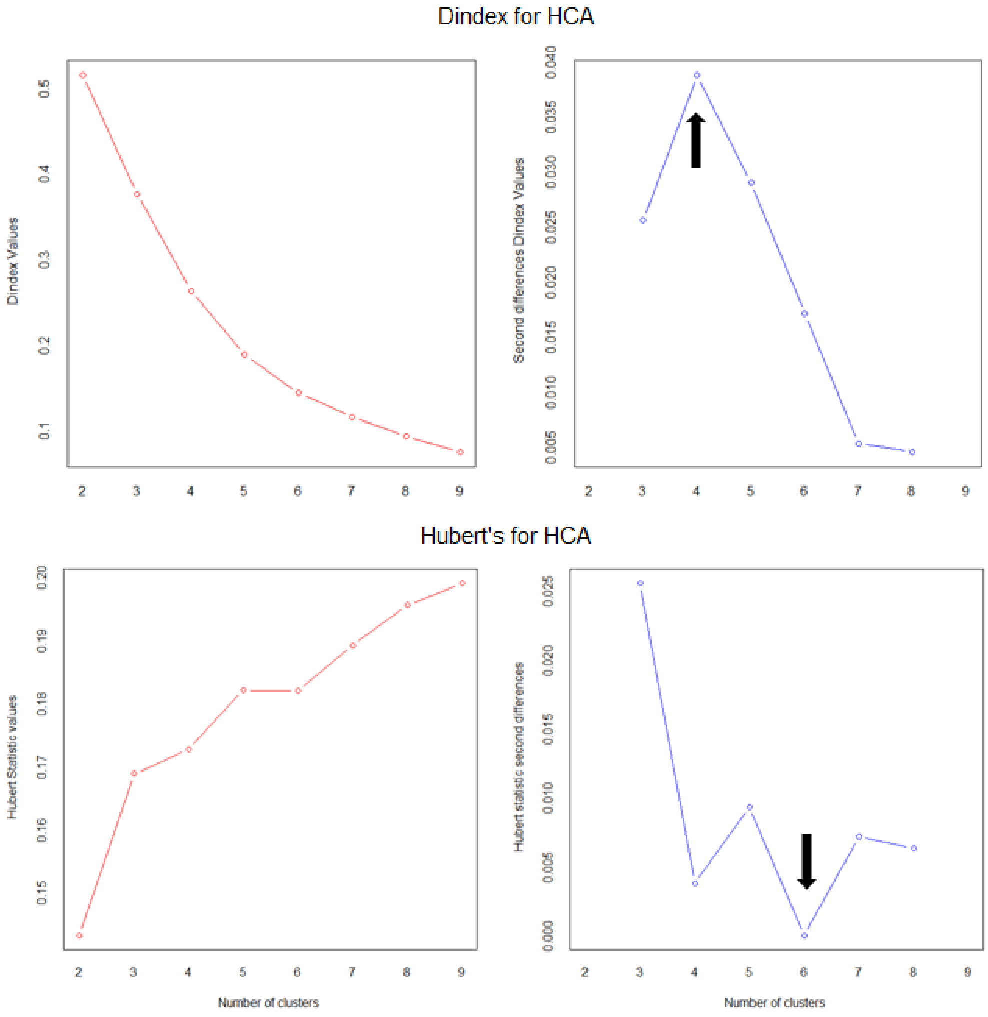


Fig. 88. Dindex and Hubert's index values for the HCA algorithm calculated from the data set with pigments and binding medium data set after pre-processing (black arrow indicates the optimal cluster number value).

© 2017 Tous droits réservés.

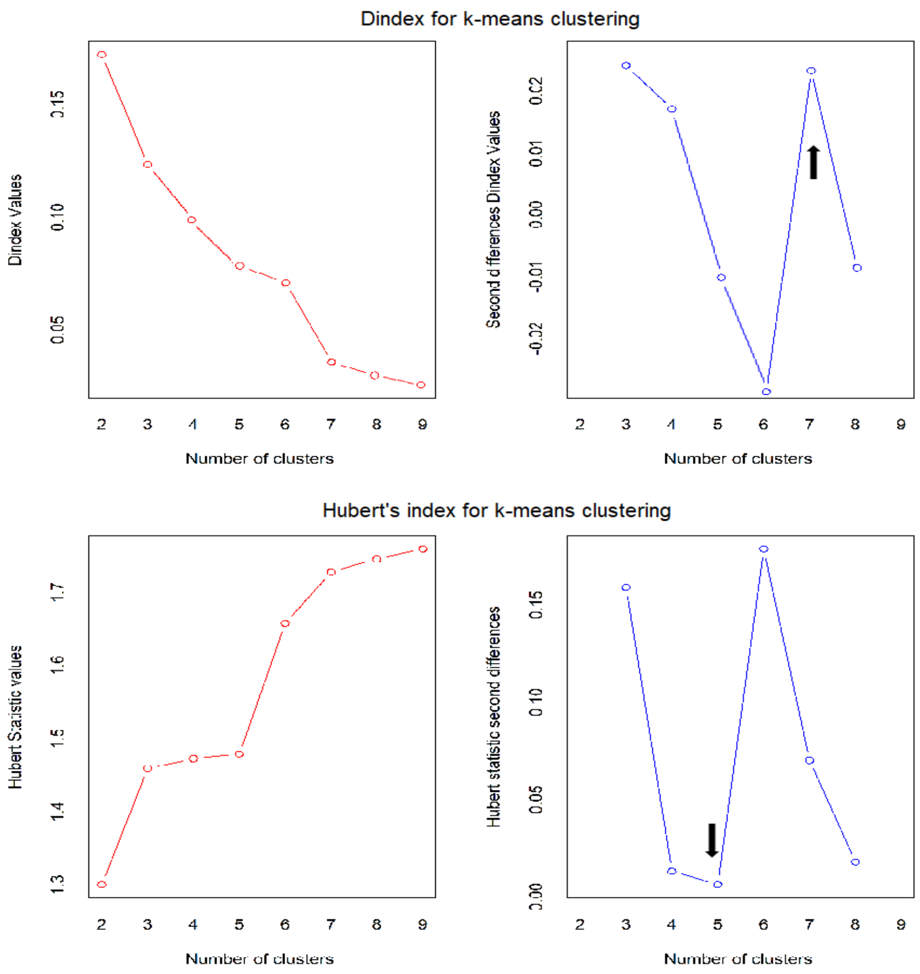


Fig. 89. Plot of Dindex and Hubert's index values for k-means clustering algorithm calculated from the data set with pigments and binding medium data set after pre-processing (black arrow indicates the optimal cluster number value).

The differences in the calculated values of the optimal number of clusters can be observed. The fuzzy partitioning coefficient and the Dindex of k-means clustering was set for 7 clusters, which corresponds to the actual data structure. However, the other calculated parameters (Dindex and Hubert's index for Ward's HCA; Hubert's index for k-means clustering) are different from the actual data structure, which in the presenting results are known *a priori*. This discrepancy is due to the difference in the clustering algorithms (Charrad et al. 2014) for which the grouping of data may

© 2017 Tous droits réservés.

differ. In addition, the specification of optimal clustering number may be problematic for these data sets where a number of variables is much higher than a number of samples (spectra) (Everitt et al. 2011). A dendrogram based on the hierarchical clustering analysis of the Near-IR spectral after pre-processing was constructed (Fig. 90). The spectra of pigments and binding medium were separated into 6 groups. The optimal number of clusters in this case was set based on the Dindex, which value was 4 clusters and it was marked on the dendrogram (Fig. 90). In turn, the Hubert's index value was calculated for 6 clusters. By a visual inspection of the dendrogram (Fig. 90) the application of the Dindex does not produce the desire results (underestimation of the data) in the case of HCA. However, it should be noted, that a more general clustering (into 2 clusters) could also be made, based on the Euclidean distance method of cut-off dendrogram point specification.

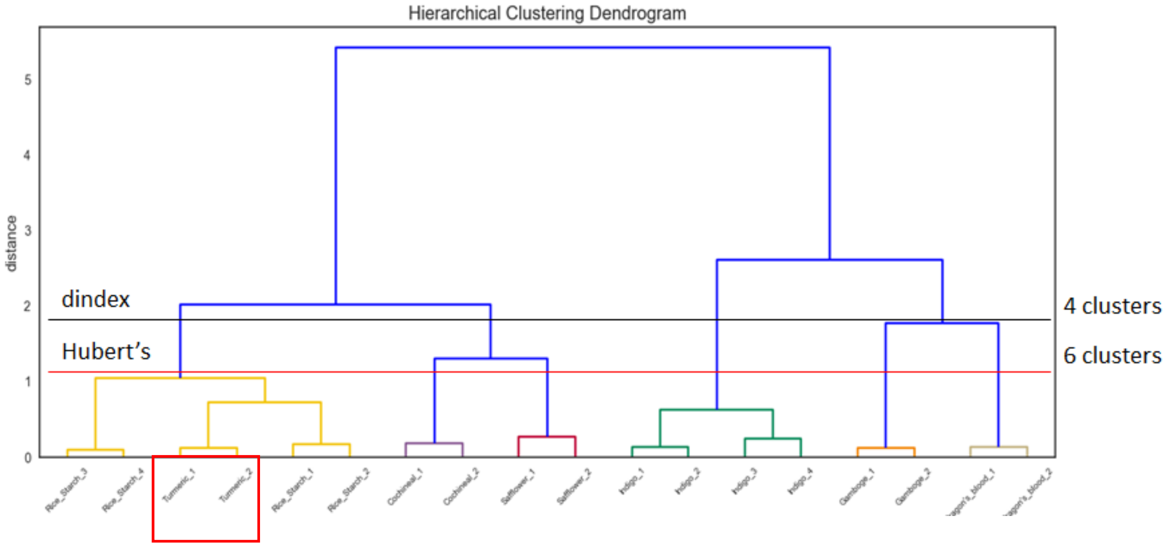


Fig. 90. Ward's HCA dendrogram of near-IR spectra of different pigments and rice starch (binding medium) after pre-processing with indicated Dindex and Hubert's index values. The red frame marked the wrongly clustered turmeric pigment, which was grouped with the binding medium (rice starch).

© 2017 Tous droits réservés.

The k-means clustering was performed for 7 clusters (Fig. 91). This procedure was performed due to the calculated Dindex for k-means clustering, as well as the fuzzy partitioning coefficient, which were set for 7 clusters.

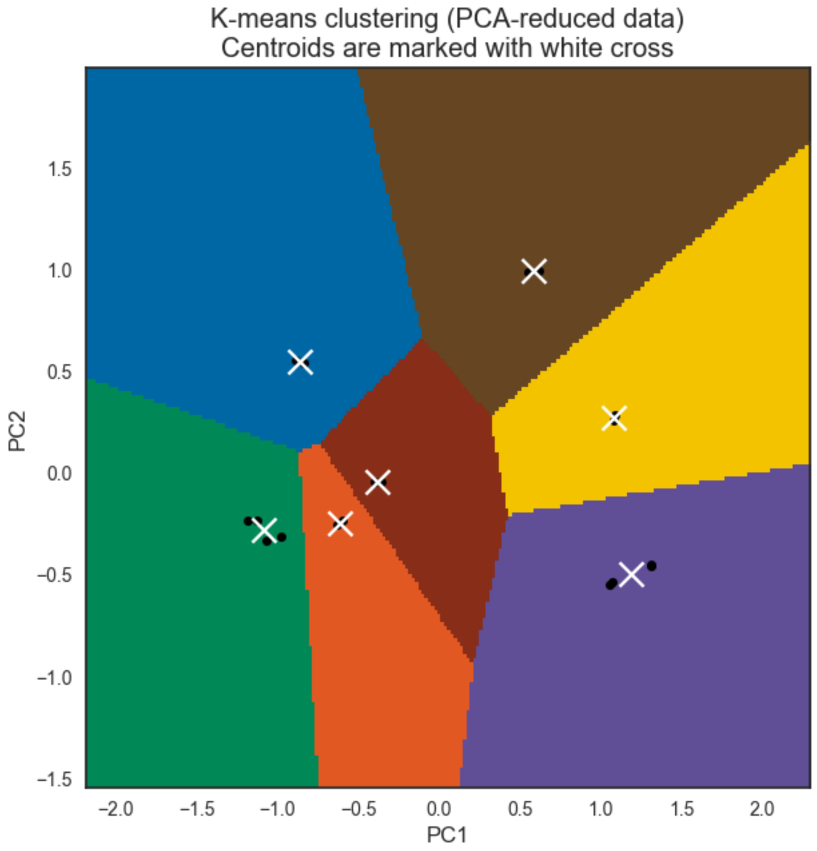


Fig. 91. Voronoi diagram of Near-IR spectra of pigments and binding medium projected on the PCA scatter plot plane.

In the Voronoi diagram (Fig. 91) the scores of the spectra are well separated and do not overlap with each other. However, green, orange and brown clusters are closely located on the diagram. PCA with scores labelling from k-means clustering was used to demonstrate the location of the matrix components in the scatter plot (Fig. 92).

© 2017 Tous droits réservés.

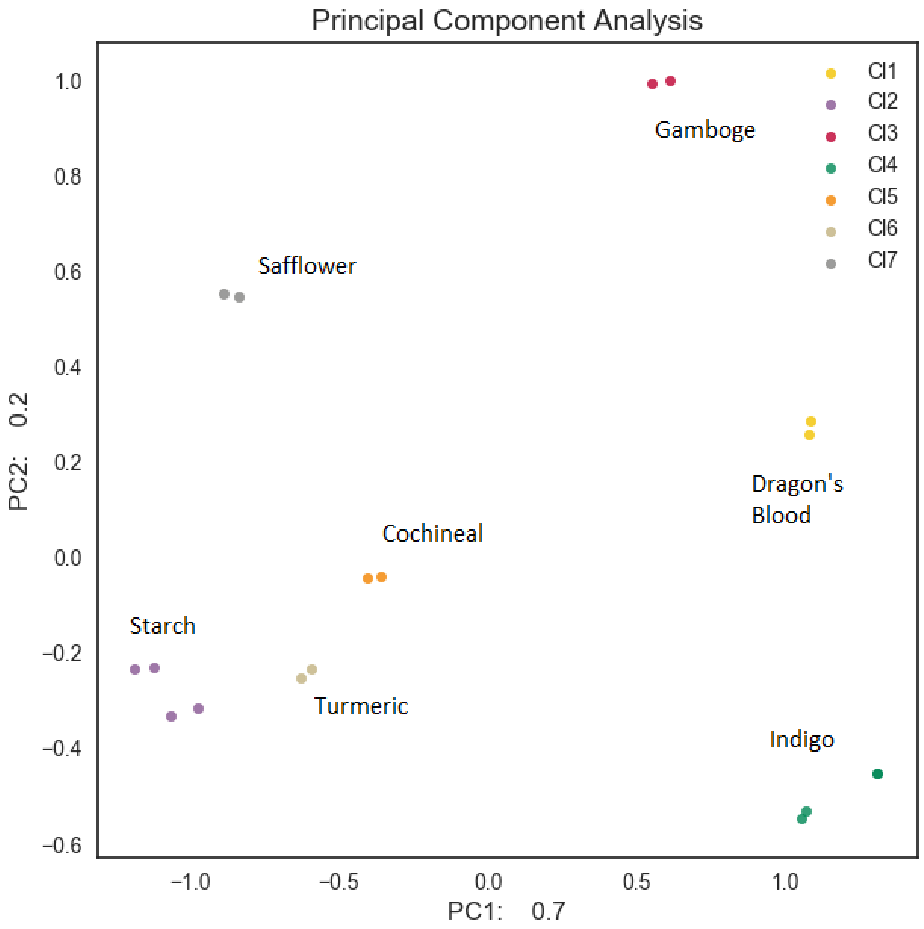


Fig. 92. PCA scatter plot of different pigments and rice starch (binding medium) Near-IR spectra after pre-processing. Scores are labelled based on the k-means clustering.

Total variance explained by two principal components is around 90%. After robust pre-processing the PCA scores corresponding to the rice starch, turmeric, and cochineal are located close to each other. Notwithstanding, separation of the pigments and rice starch is possible.

Unfortunately, the results from complementary clustering algorithms (fuzzy c-means and Ward's HCA) are not identical. In fact, fuzzy c-means clustering –due to the

© 2017 Tous droits réservés.

random initialization seeding used in the Spectronomy system – is not fully reproducible as Ward's HCA (Stetco et al. 2015). Here the grouping of spectra will be exactly the same for the same data matrix regardless the number of runs of Ward's HCA, where in the case of fuzzy c-means the spectra separation may slightly differ. Therefore, the pigments separation cannot be based only on fuzzy c-means clustering included in the Spectronomy system. The HCA dendrogram shows that only turmeric pigment is wrongly clustered (Fig. 89). Another problem is the specification of an actual number of components in the data set, which was previously discussed and presented (Fig. 87, Fig. 88 and Fig. 89). The results of the cluster analysis are summarized in Table 5.

Table 5. Clusters designated by different clustering techniques and components linked with these groups.

Component	Fuzzy clustering	K-means	HCA
Rice_Starch	7	2	1
Safflower	4	7	3
Cochineal	3	5	2
Turmeric	5	6	1
Indigo	2	4	4
Gamboge	1	3	5
Dragon's_blood	6	1	6

Fuzzy-c-means clustering and k-means clustering, produce the same partitioning, based on the numbers of groups calculated by the fuzzy partitioning coefficient

© 2017 Tous droits réservés.

(Fig. 86) and Dindex (Fig. 88), respectively. As mentioned previously, for Ward's HCA only turmeric pigment was wrongly clustered. Moreover, the specification of the optimal *cut-off* point of the dendrogram may be problematic The Dindex and Hubert's parameters are not corresponding to the actual number of components. In addition, the *cut-off* point set by a visual inspection (Euclidean distance method) leads to an underestimation of the data. To solve this problem, the detailed visual inspection of the 3D PCA scatterplot was made (Fig. 93).

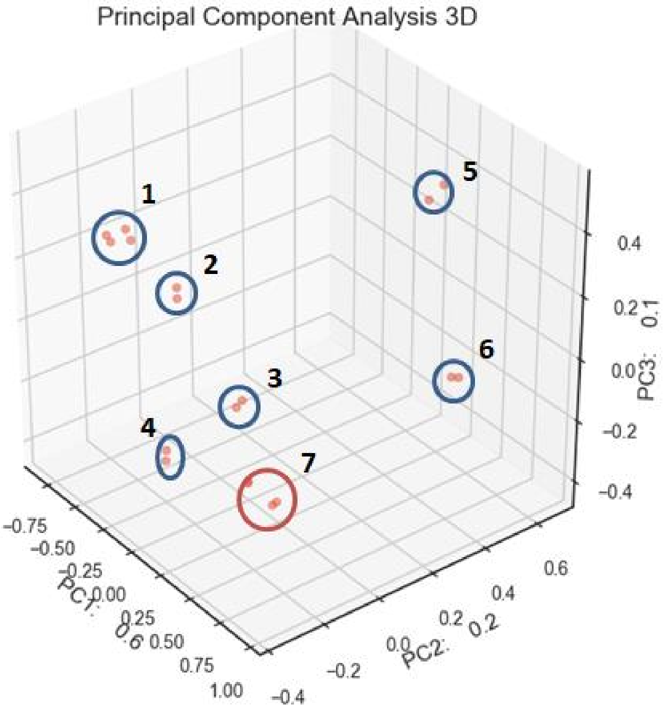


Fig. 93. 3D PCA scatter plot with marked groups of spectra.

The cumulative variance explained by 3 principal components is ~90%. At the 3D PCA scatterplot (Fig. 93), the 7 groups can be specified, what may play an important role in a process of specifying a number of components. In comparison to the 2D scatter plot, (Fig. 92) the pattern formation is more visible. For specification of the actual number of components in the data set, the two types of results were generated: internal parameters calculation (fuzzy partitioning coefficient, Dindex

© 2017 Tous droits réservés.

value, Hubert's index value) and 3D PCA scatter plot. The results of Dindex calculated for k-means clustering, fuzzy partitioning coefficient and 3D PCA scatter plot examination indicate the correct number of components in the data set. However, other internal parameters may mislead the specification of the optimal number of clusters. Thus, the application of many complementary techniques in this case is justified. Near-IR spectroscopy seems to be an appropriate technique for pigments separation, but specification of a number of components in the data set may be difficult, mostly for mixed pigments.

Homogeneity of the pigment (Near-IR)

The acquisition of 10 spectra of the dragon's blood pigment deposited on the paper was made in two contrasting areas – the homogeneous area, which by visual inspection was precisely covered by the pigment (5 spectra) and heterogeneous area with unequally deposited pigment on the paper (5 spectra). Spectra were obtained by averaging 30 scans with an acquisition time about 20 seconds and at an 8 cm^{-1} spectral resolution. The picture of the sample with the marked spot of analysis is presented below (Fig. 94).

© 2017 Tous droits réservés.

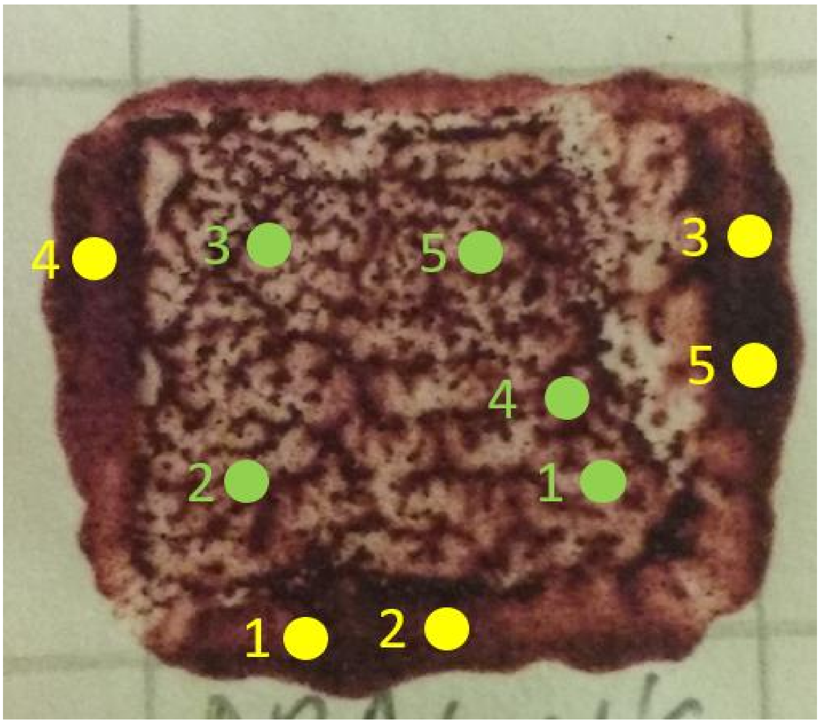


Fig. 94. The deposited dragon's blood pigment with marked spots of analysis (heterogeneous area – green and homogeneous area – yellow).

The several types of pre-processing procedures and their configuration were tested for optimization of the dragon's blood spectra separation from the homogeneous and heterogeneous areas in the clustering algorithms. The results presented below (Fig. 95) are corresponding to the optimal procedure for the pre-processing of spectra which includes the 3rd polynomial EMSC method for normalization and Asymmetric Least Squares Baseline Correction (AsLS) with lambda = 10^5 and p-value = 0.1.

© 2017 Tous droits réservés.

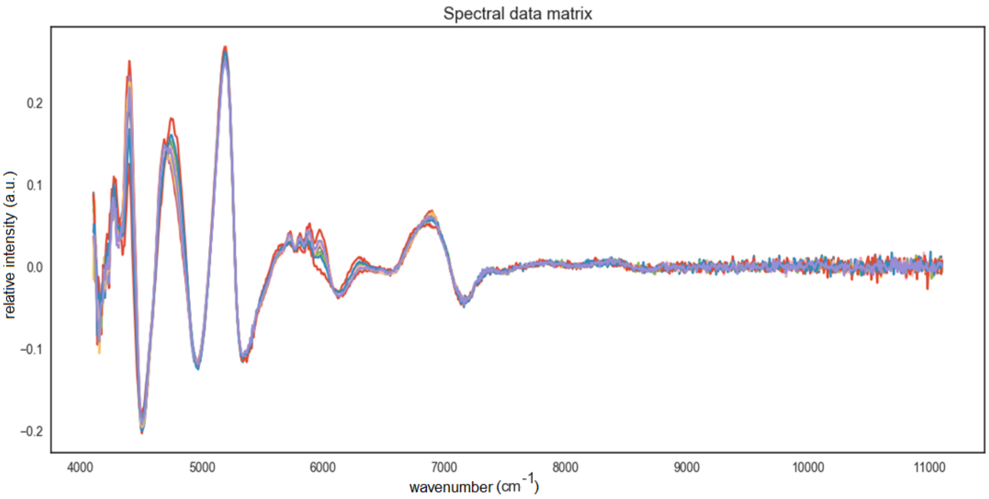


Fig. 95. Near-IR spectra of dragon's blood pigment spectra from homogeneous and heterogeneous areas after pre-processing 3rd polynomial of EMSC and AsLS (lambda = 10^5; p-value = 0.1).

In Fig. 95, the possible momentous spectral areas are hardly distinguished. However, presumably the important variation can be observed in the range ~4700 cm^{-1} and between 5700-6000 cm^{-1}.

The fuzzy partitioning coefficient (FPC) indicates a favorable separation of the data based on two clusters (Fig. 96).

© 2017 Tous droits réservés.

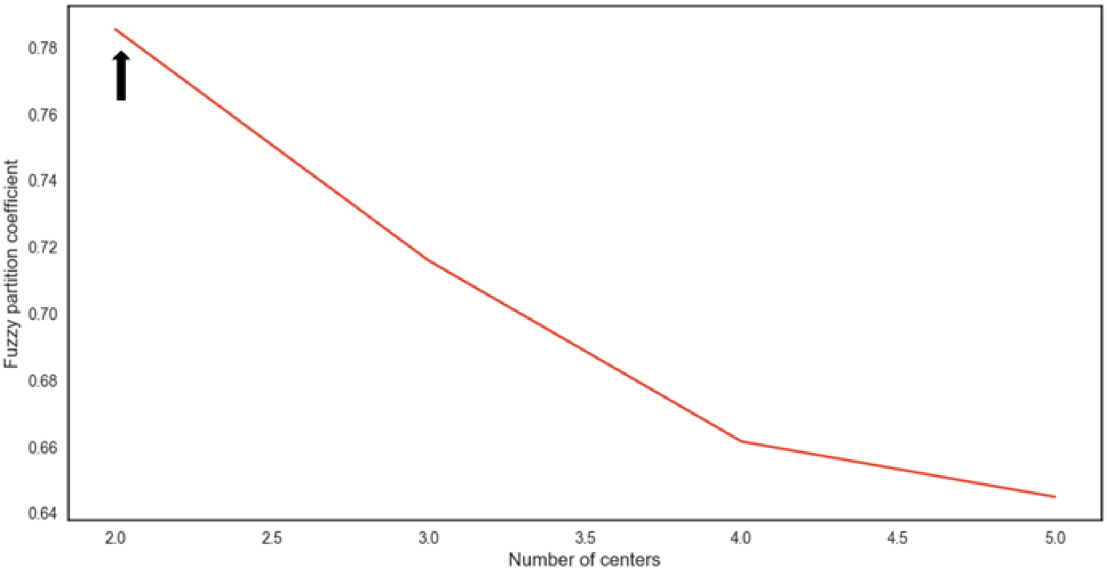
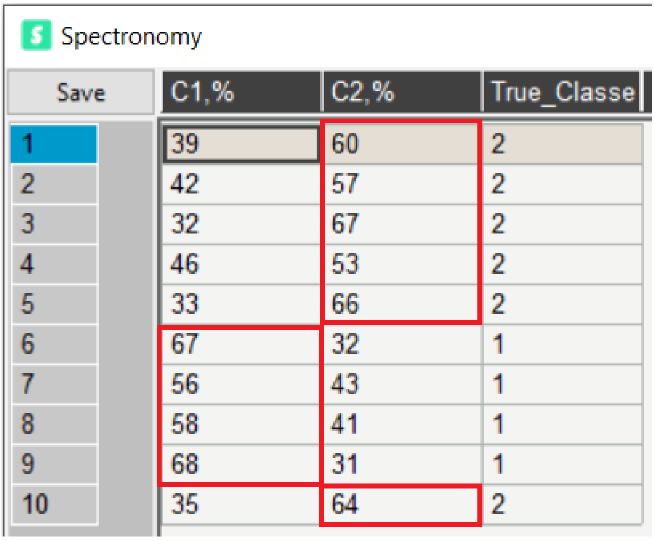


Fig. 96. Fuzzy partitioning coefficient calculated from the data set with spectra of dragon’s blood pigment deposited on paper, collected from homogeneous and heterogeneous areas.

Based on the FPC results, fuzzy c-means clustering was performed for 2 clusters (Fig. 97).

© 2017 Tous droits réservés.



	C1,%	C2,%	True_Classe
1	39	60	2
2	42	57	2
3	32	67	2
4	46	53	2
5	33	66	2
6	67	32	1
7	56	43	1
8	58	41	1
9	68	31	1
10	35	64	2

Fig. 97. Fuzzy c-means clustering table of Near-IR spectra after pre-processing of dragon's blood pigment deposited on paper, collected from homogeneous and heterogeneous areas.

2 different clusters were separated with the cluster affiliation of the spectra between 53-68% (Fig. 97). Complementary multivariate data processing algorithms were used: PCA with score labelling based on the fuzzy c-means results and Ward's HCA. PCA with scores labelling from fuzzy c-means clustering was used to demonstrate the location of the matrix components in the scatter plot (Fig. 98).

© 2017 Tous droits réservés.

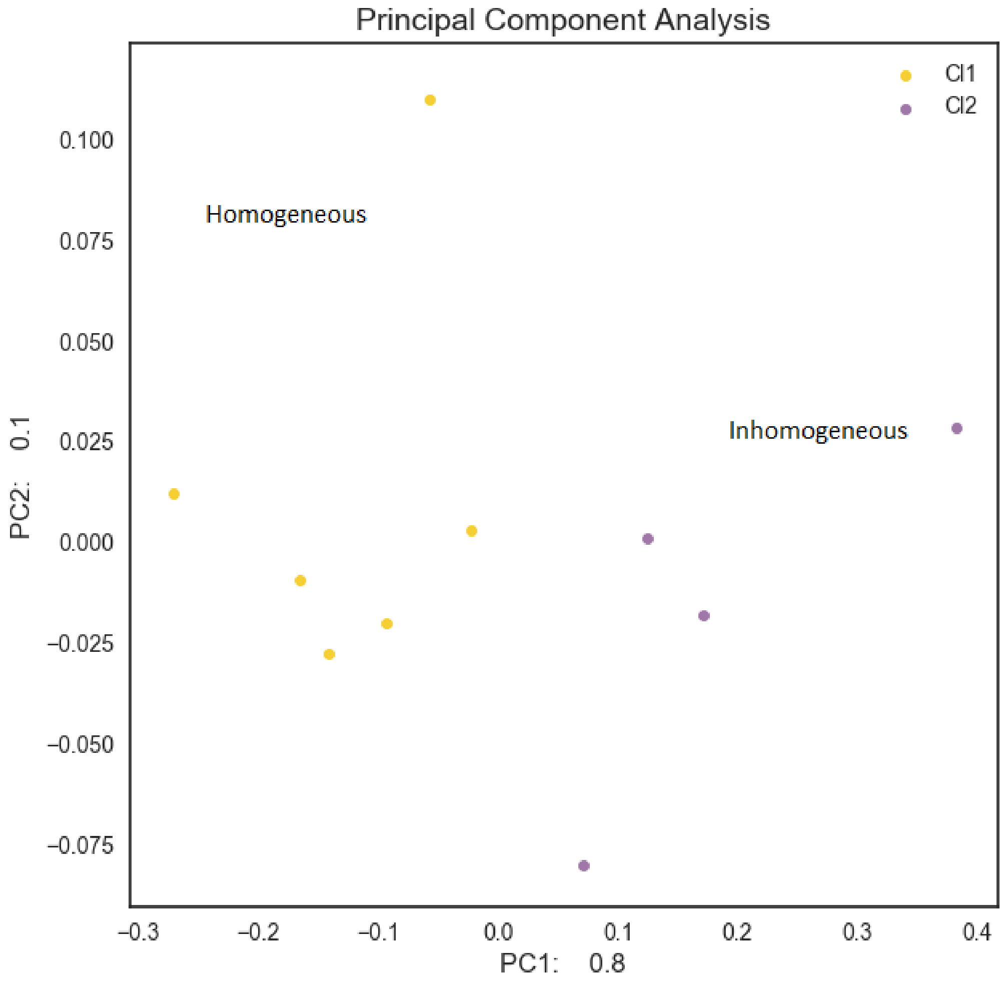


Fig. 98. PCA scatter plot of dragon's blood spectra collected from homogeneous and heterogeneous areas. Scores are labelled based on fuzzy c-means clustering.

Total variance explained by two principal components is ~ 90%. After robust pre-processing the PCA scores corresponding to two separated groups of the data can be distinguished. These groups, as previously assumed, are corresponding to the homogeneous and heterogeneous areas of the deposited pigment.

In turn, Ward's HCA indicates favourable separation of the data based on two clusters (complementary to PCA and fuzzy c-means clustering) (Fig. 99).

© 2017 Tous droits réservés.

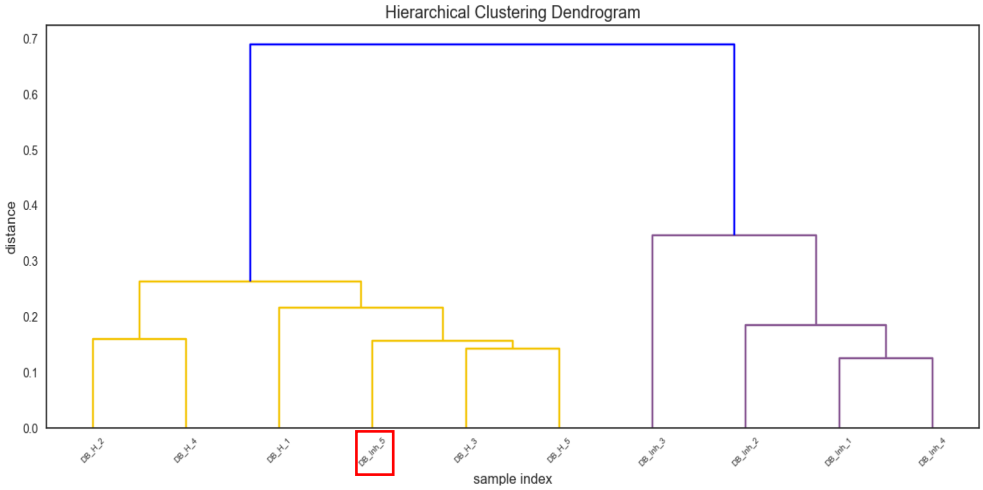
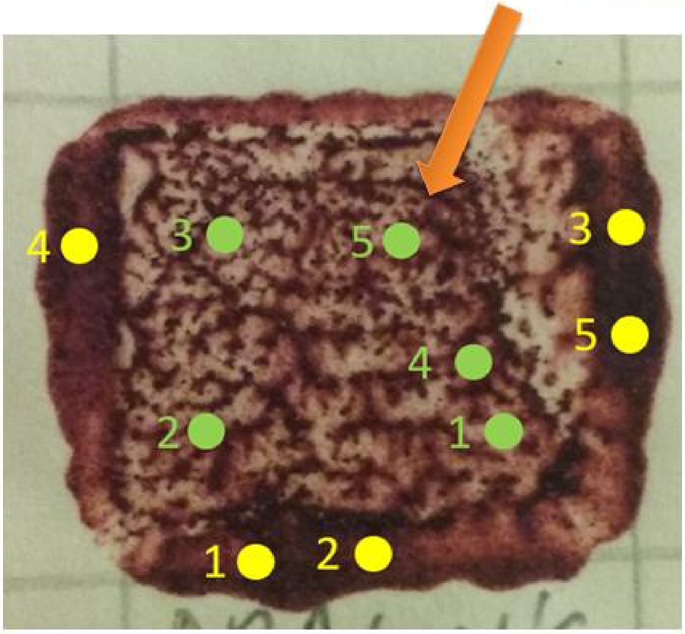


Fig. 99. Ward’s HCA dendrogram of Near-IR spectra of dragon’s blood pigments collected from homogeneous and heterogeneous areas. The red frame marked the wrongly clustered inhomogeneous spot.

The results from the presented techniques (PCA, fuzzy c-means and Ward’s HCA) show the separation of the spectra into two groups. However, one spot which was designated as inhomogeneous is wrongly clustered (Fig. 100).

Wrongly clustered inhomogeneous spot



© 2017 Tous droits réservés.

Fig. 100. Inhomogeneous spot incorrectly clustered to homogeneous areas' group by fuzzy c-means and Ward's HCA techniques.

This state of affairs is probably due to the degree of pigment coverage. The assumption of this process is outlined below (Fig. 101).

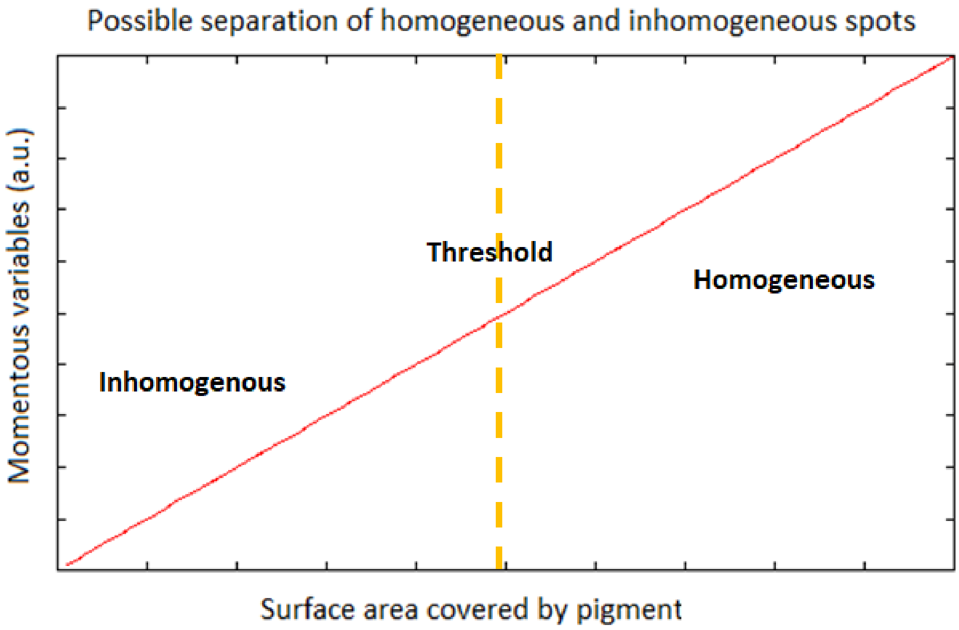


Fig. 101. The estimated threshold occurrence based on pigment coverage on paper.

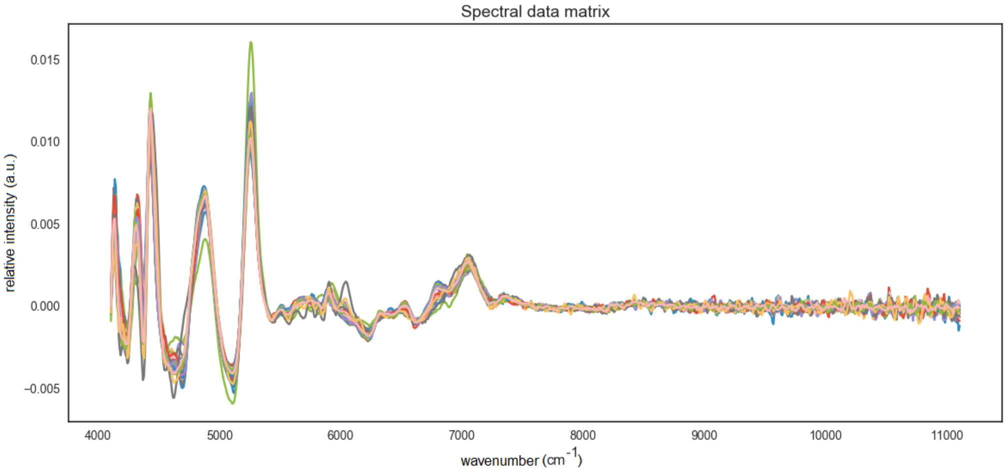
The description of the spots collected from the samples was important to determine the effectiveness of separation of homogeneous and heterogeneous areas. However, the bias in the specification of area labels by an investigator cannot be excluded. Based on fuzzy c-means clustering, PCA and HCA, the differentiation of two main groups was possible. One inhomogeneous spot was wrongly clustered in relation to the labels designated by the investigator. These results indicate that there is a requirement to compromise suggestive observations (by the investigator) and objective results (clustering results).

© 2017 Tous droits réservés.

Clustering of pigment deposits on two different papers (Near-IR)

The acquisition of 2 Near-IR spectra of the safflower (carthamin pigment), cochineal (carmine pigment), turmeric (curcumine pigment), gamboge pigment, dragon's blood pigment, indigo pigment and rice starch was made. The pigments were deposited on the two papers (Whatman and K17) from different suppliers. In order to compare the variability of the substrate, the spectrum of each paper was also measured (2 spectra). The total number of collected spectra in the data matrix was 28 (14 spectra for each paper). The spectra were obtained by averaging 30 scans with an acquisition time about 20 seconds and at a 8 cm^{-1} spectral resolution.

The several types of pre-processing procedures and their configuration were tested for optimization of the pigments and binding medium spectra separation in the clustering algorithms. The results presented below are corresponding to the optimal procedure for the pre-processing which includes the application of the 3rd polynomial EMSC method for normalization and Asymmetric Least Squares Baseline Correction (AsLS) with lambda = 10^5 and p-value = 0.1 (Fig. 102).



© 2017 Tous droits réservés.

Fig. 102. Near-IR spectra of different pigments and binding medium (rice starch) deposited on two papers (Whatman and K17) after pre-processing 3rd polynomial of EMSC and AsLS (lambda = 10^5; p-value = 0.1).

By examination of the paper spectra after pre-processing (3rd EMSC and AsLS baseline correction) a small difference between spectra is observed (Fig. 103).

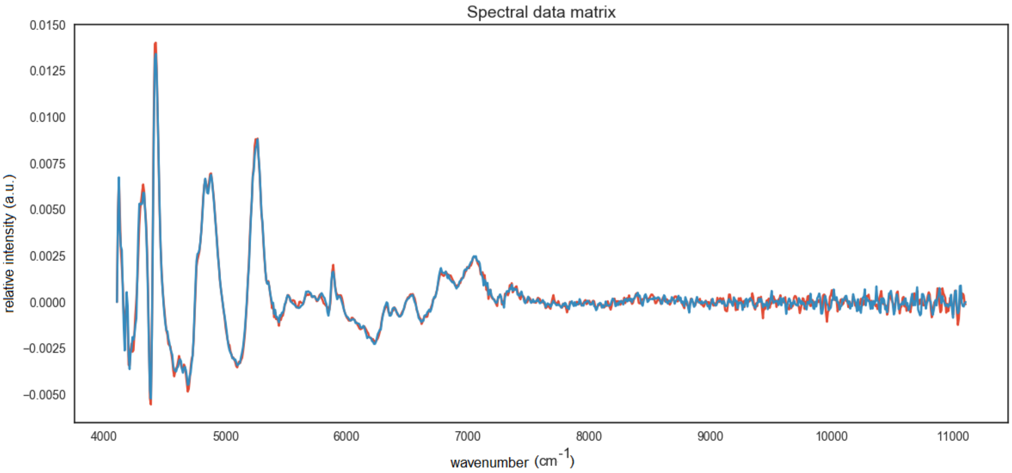


Fig. 103. Near-IR spectra after pre-processing 3rd polynomial of EMSC and AsLS (lambda = 10^5; p-value = 0.1) of two papers (Whatman and K17) used as substrates for pigment deposition.

The fuzzy partitioning coefficient indicates favorable separation of the Near-IR spectra from the pigments and binding medium (rice starch) deposited on two papers, into two clusters (Fig. 104).

© 2017 Tous droits réservés.

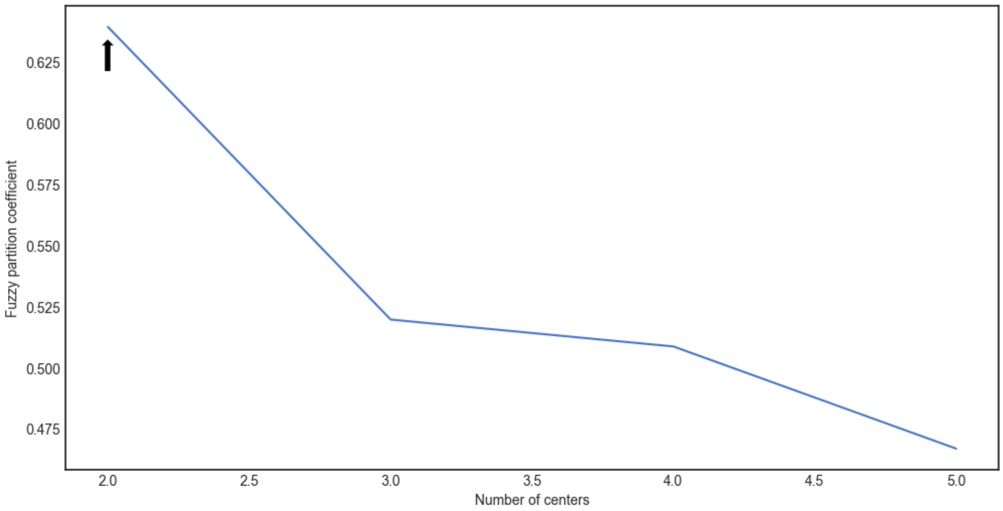
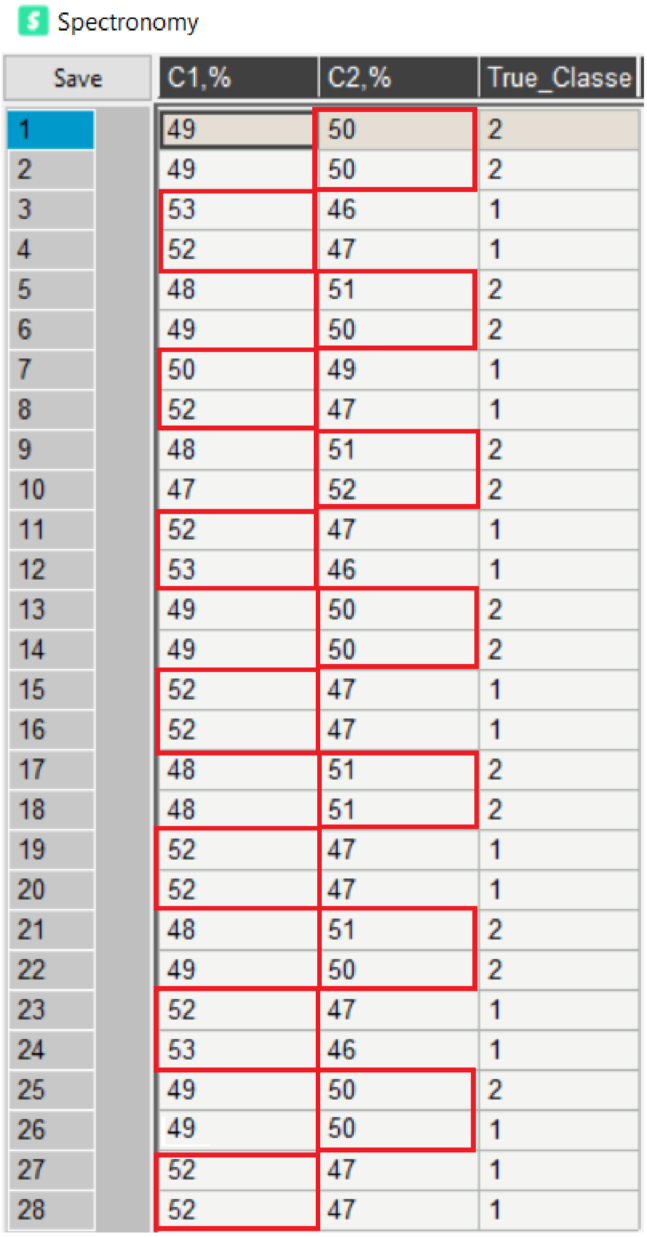


Fig. 104. Fuzzy partitioning coefficient calculated for the data set with spectra of 6 pigments and binding medium (rice starch) deposited on two different papers.

Based on the FPC results, fuzzy c-means clustering was performed for 2 clusters

© 2017 Tous droits réservés.



Spectronomy

Save	C1,%	C2,%	True_Classe
1	49	50	2
2	49	50	2
3	53	46	1
4	52	47	1
5	48	51	2
6	49	50	2
7	50	49	1
8	52	47	1
9	48	51	2
10	47	52	2
11	52	47	1
12	53	46	1
13	49	50	2
14	49	50	2
15	52	47	1
16	52	47	1
17	48	51	2
18	48	51	2
19	52	47	1
20	52	47	1
21	48	51	2
22	49	50	2
23	52	47	1
24	53	46	1
25	49	50	2
26	49	50	1
27	52	47	1
28	52	47	1

Fig. 105. Fuzzy c-means clustering table of Near-IR spectra after pre-processing of the 6 pigments and binding medium (rice starch) deposited on the two different papers.

2 different clusters of spectra were separated with the weak cluster affiliation of spectra between 50-53% (Fig. 105). Complementary multivariate data processing algorithms were used: PCA with score labelling based on fuzzy c-means results and Ward's HCA. PCA with scores labelling from fuzzy c-means clustering was used to demonstrate the location of the matrix components in the scatter plot (Fig. 85).

© 2017 Tous droits réservés.

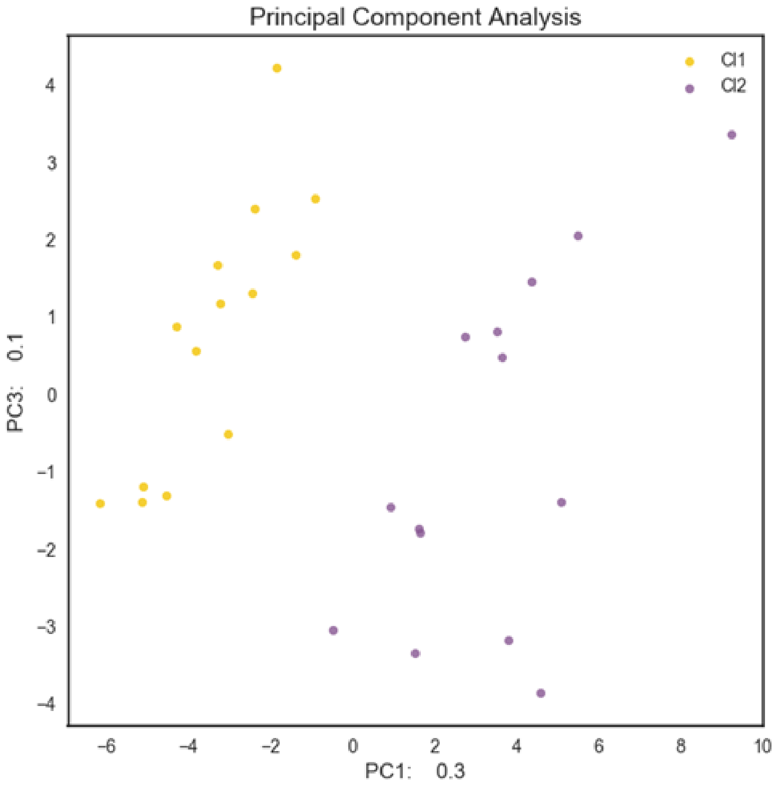


Fig. 105. PCA scatter plot of Near-IR spectra from 6 pigments and rice starch (binding medium) deposited on two different papers, after pre-processing.

Total variance explained by two principal components is around 40%. After the robust pre-processing, two separated groups from the data can be distinguished. These groups correspond to the pigments deposited on the different paper (yellow – Whatman and purple – K17). The same separation of the spectra can be observed on the generated dendrogram by Ward’s HCA (Fig. 106).

© 2017 Tous droits réservés.

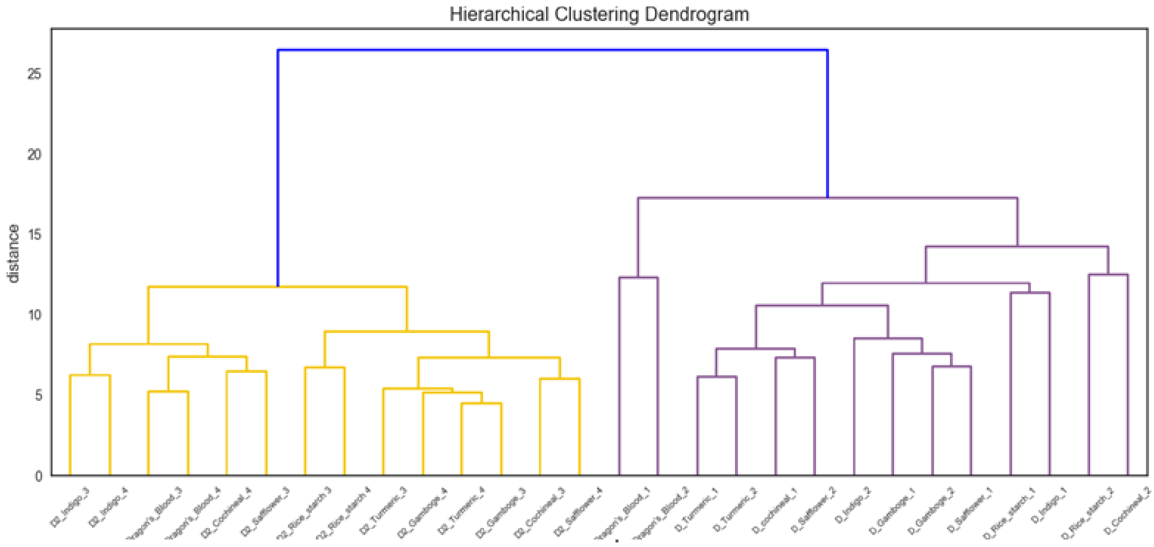


Fig. 106. Ward's HCA dendrogram of Near-IR spectra of 6 pigments and rice starch (binding medium) deposited on two different papers after pre-processing

Ward's HCA indicates favourable separation of the data based on two clusters (complementary to PCA and fuzzy c-means clustering). These clusters are corresponding to the pigments and binding medium deposited on the Whatman paper (cluster 1) and the K17 paper (cluster 2). However, the spectra of the matrix components (pigments and binding medium) in the lower branches of the dendrogram are not correctly separated (Fig. 107).

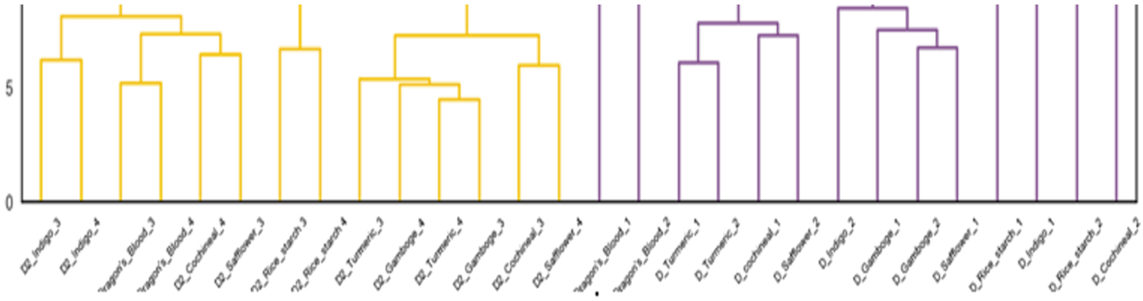


Fig. 107. Enlargement of the Ward's HCA dendrogram with mixed structure of pigments and binding medium.

In the graph presented above (Fig. 107), even pigments with contrasting colours, e.g. indigo and cochineal, are included in the same cluster. Moreover, the rice starch

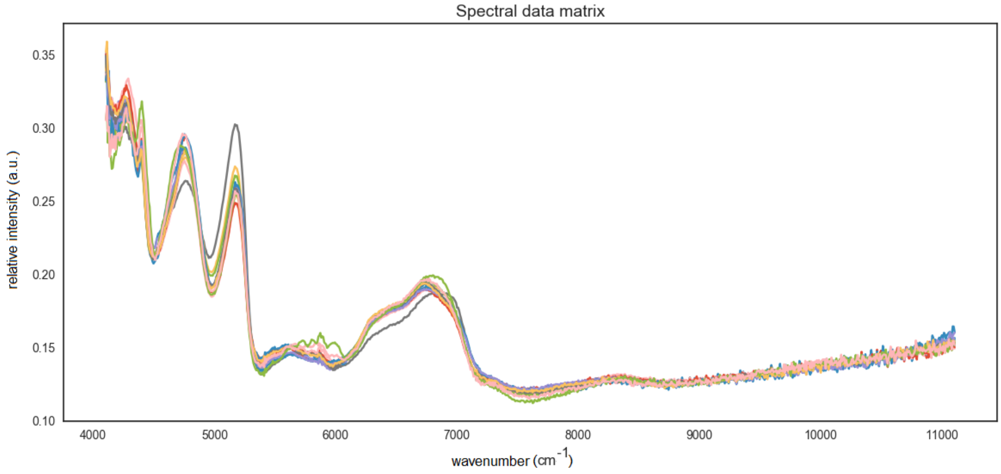
© 2017 Tous droits réservés.

(binding medium) was not classified in the disconnected group, with respect to the pigments.

This situation is also reflected in PCA and fuzzy c-means clustering, where exactly the same groups were created. In view of the above, the featured data set was divided into two smaller data sets with the spectra corresponding to the pigments and rice starch deposited on each paper independently.

Clustering of the pigment deposits on the Whatman paper (Near-IR)

The acquisition of 2 Near-IR spectra of the safflower (carthamin pigment), cochineal (carmine pigment), turmeric (curcumine pigment), gamboge pigment, dragon's blood pigment, indigo pigment and rise starch was made. The pigments were deposited on the Whatman paper. The total number of collected spectra in the data matrix was 14. The several types of pre-processing procedures and their configuration were tested for optimization of the pigments and binding medium spectra separation in the clustering algorithms. The results presented below are corresponding to the optimal procedure for the pre-processing of spectra, which includes only the 3rd polynomial EMSC method for normalization only (Fig. 108). The pre-processing step was reduced to highlight the differences between the spectra.



© 2017 Tous droits réservés.

Fig. 108. Near-IR spectra of different pigments and binding medium (rice starch) deposited on Whatman paper after pre-processing 3rd polynomial of EMSC.

The exploratory data analysis was performed by the 3D PCA (Fig. 109).

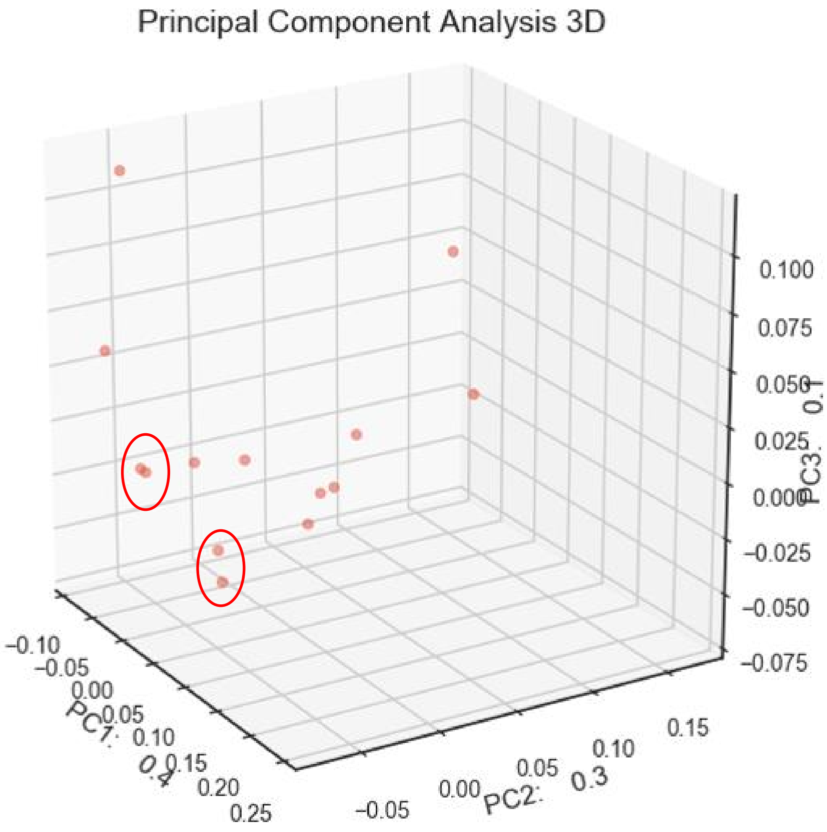


Fig. 109. 3D PCA scatter plot of the data set containing spectra of 6 pigments and binding medium (rice starch) deposited on Whatman paper.

Total cumulative variance explained by three first principal components is around 80%. In the scatter plot shown above (Fig. 109), no unquestionable group formation can be observed. The scores are distributed in a way that prevents a user from clearly

© 2017 Tous droits réservés.

identifying a number of groups (clusters) in the matrix. However, some of the scores transparently form pairs, what was marked by the red circles in the scatter plot (Fig. 109). Ward's HCA was performed as another multivariate data processing technique. The generated dendrogram is presented below (Fig. 110).

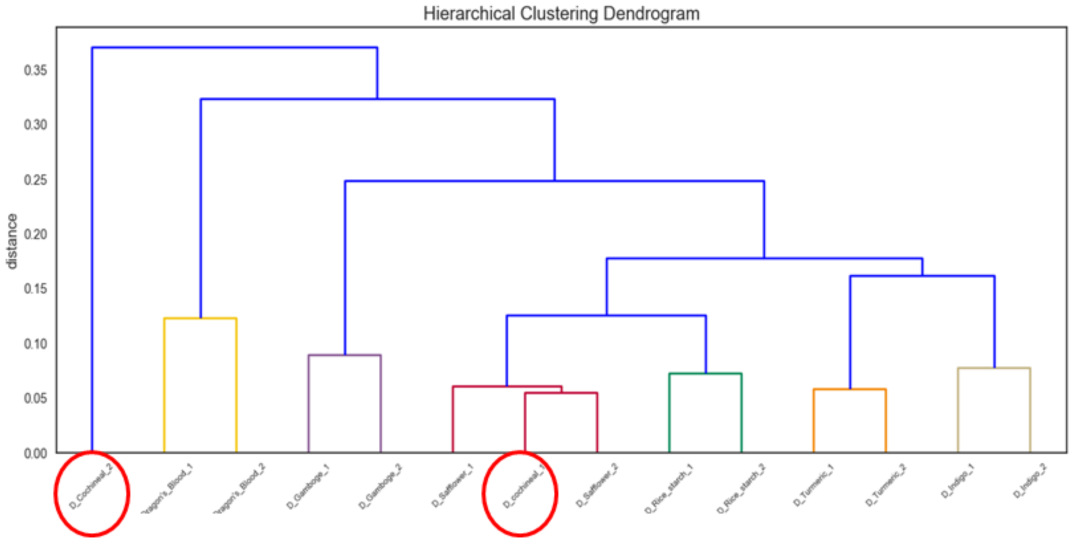


Fig. 110. Ward's HCA dendrogram of Near-IR spectra of different pigments and rice starch (binding medium) deposited on Whatman paper. The wrongly clustered components are marked by red circles.

In the generated dendrogram (Fig. 110) the weak differentiation of the spectra in the distinct dendrogram structure can be observed. The spectra of cochineal pigment are unpaired: (i) distributed in the separated node and (ii) in the safflower pigment cluster. The placement of the cochineal spectrum with the safflowers' makes a good sense due to the similar colours of these pigments. Unfortunately, the dendrogram structure does not create the coherent unity. The rice starch (binding medium) spectra occurred in the note with the safflower pigment and a single cochineal spectrum. The turmeric pigments were placed in the same node with contrasting indigo. The chemical structure of these pigments are also not complementary, which was also described elsewhere (Pozzi 2011).

© 2017 Tous droits réservés.

The calculated fuzzy partitioning coefficient shows the correct number of components in the data matrix (Fig. 111), where in turn the fuzzy c-means clustering algorithm (Fig. 112) produced an unclear form of the results, as in the case of PCA.

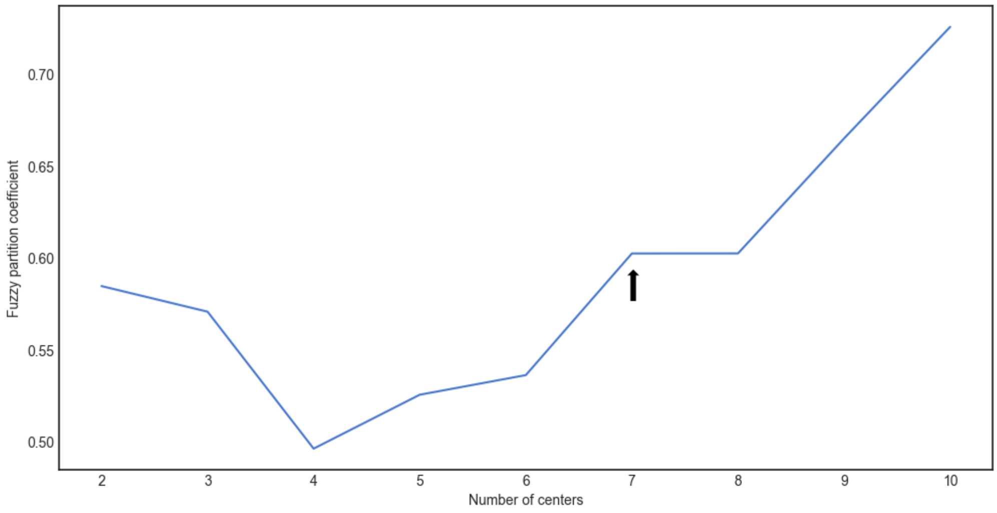
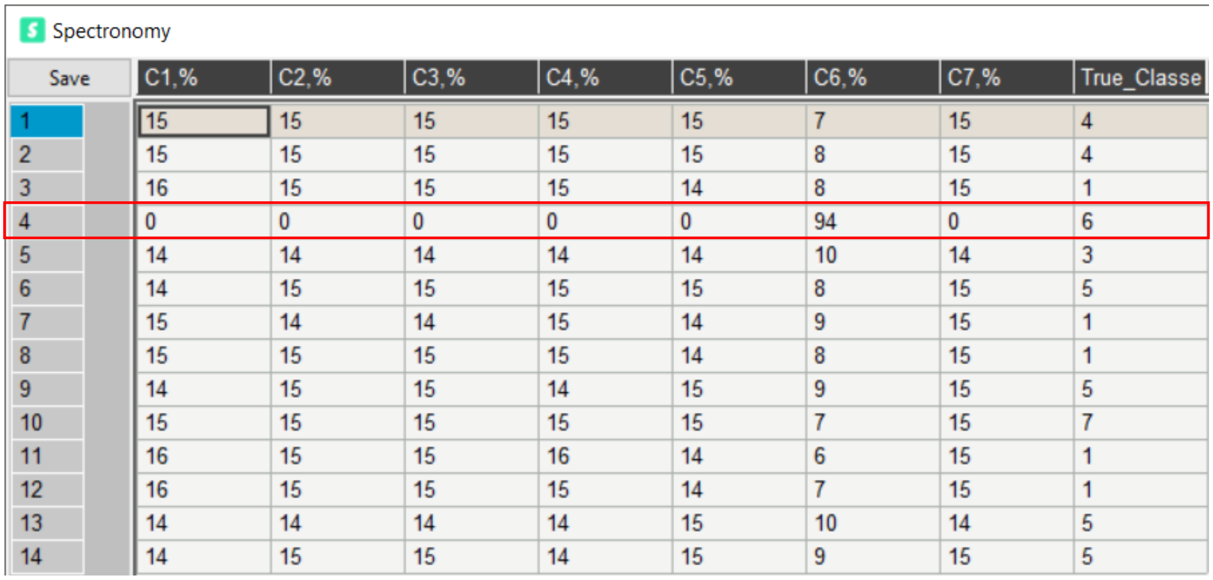


Fig. 111. Fuzzy partitioning coefficient calculated for the data set with spectra of 6 pigments and binding medium (rice starch) deposited on Whatman paper.

The black arrow in the FPC plot (Fig. 111) indicated the correct number of components in the data matrix (6 pigments and 1 binding medium). Nonetheless, the fuzzy c-means clustering results calculated for 7 clusters demonstrates very weak spectra separation (Fig. 112).

© 2017 Tous droits réservés.



Spectronomy

Save	C1,%	C2,%	C3,%	C4,%	C5,%	C6,%	C7,%	True_Classe
1	15	15	15	15	15	7	15	4
2	15	15	15	15	15	8	15	4
3	16	15	15	15	14	8	15	1
4	0	0	0	0	0	94	0	6
5	14	14	14	14	14	10	14	3
6	14	15	15	15	15	8	15	5
7	15	14	14	15	14	9	15	1
8	15	15	15	15	14	8	15	1
9	14	15	15	14	15	9	15	5
10	15	15	15	15	15	7	15	7
11	16	15	15	16	14	6	15	1
12	16	15	15	15	14	7	15	1
13	14	14	14	14	15	10	14	5
14	14	15	15	14	15	9	15	5

Fig. 112. Fuzzy c-means clustering table of Near-IR spectra after pre-processing of 6 pigments and binding medium (rice starch) deposited on Whatman paper.

The cluster membership was almost equally divided between the number of groups (~15%), beside one exception, where this value was calculated for 94%. This visible outlier is one of the cochineal spectrums and is corresponding to the results from Ward's HCA where this spectrum was placed in the separated node.

Clustering of the pigment deposits on the K17 paper (Near-IR)

The acquisition of 2 Near-IR spectra of the safflower (carthamin pigment), cochineal (carmine pigment), turmeric (curcumine pigment), gamboge pigment, dragon's blood pigment, indigo pigment and rise starch was made. The pigments were deposited on the K17 paper. The total number of collected spectra in the data matrix was 14. The pre-processing procedure was identical as in the case of the pigments deposited on the Whatman paper and included only the 3rd polynomial EMSC method for normalization only (Fig. 113).

© 2017 Tous droits réservés.

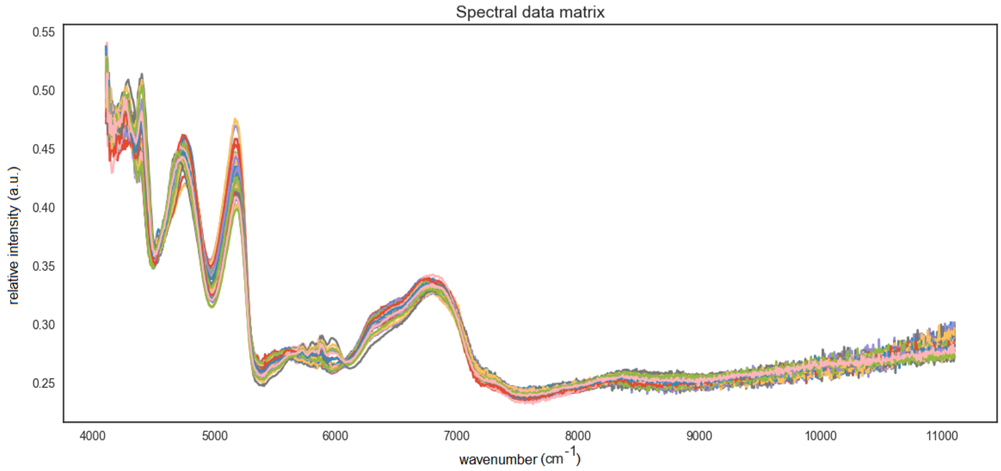


Fig. 113. Near-IR spectra of different pigments and binding medium (rice starch) deposited on K17 paper after pre-processing 3rd polynomial of EMSC.

The exploratory data analysis was performed by PCA (Fig. 114).

© 2017 Tous droits réservés.

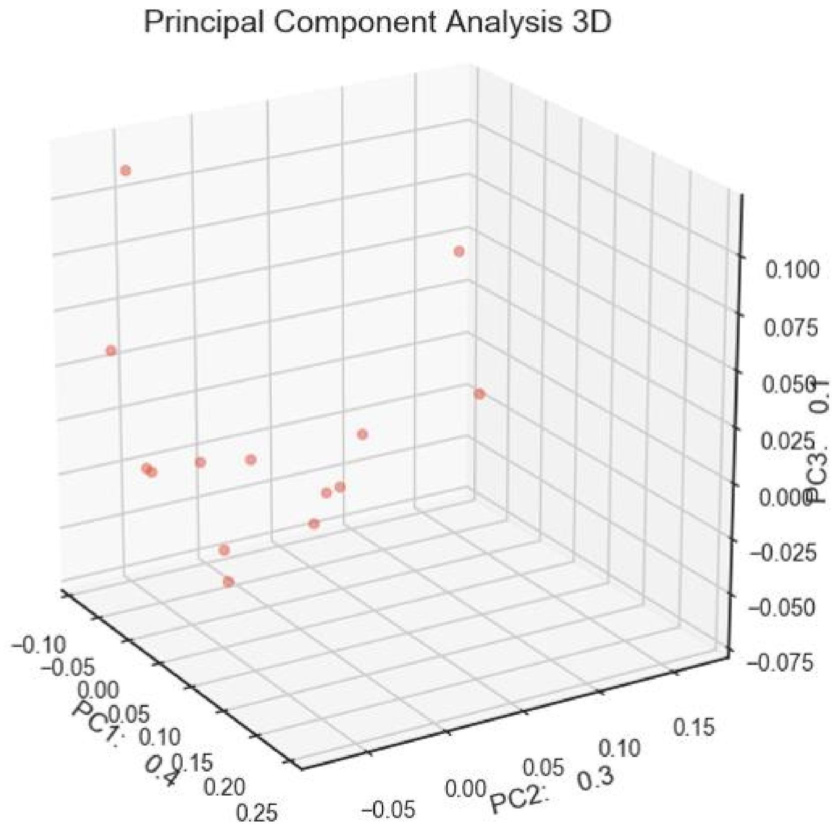


Fig. 114. 3D PCA scatter plot of the data set containing spectra of 6 pigments and binding medium (rice starch) deposited on K17 paper.

Total cumulative variance explained by three first principal components is around 80%. In the scatter plot shown above (Fig. 114), no unquestionable group formation can be observed as in the case of the pigments deposited on the Whatman paper. The scores are distributed in a way that prevents a user from clearly identifying the number of groups (clusters) in the matrix. However, some of the scores transparently form pairs, what was marked by the red circles on the scatter plot (Fig. 114). Ward's HCA was performed as another multivariate data processing technique. The generated dendrogram is presented below (Fig. 115).

© 2017 Tous droits réservés.

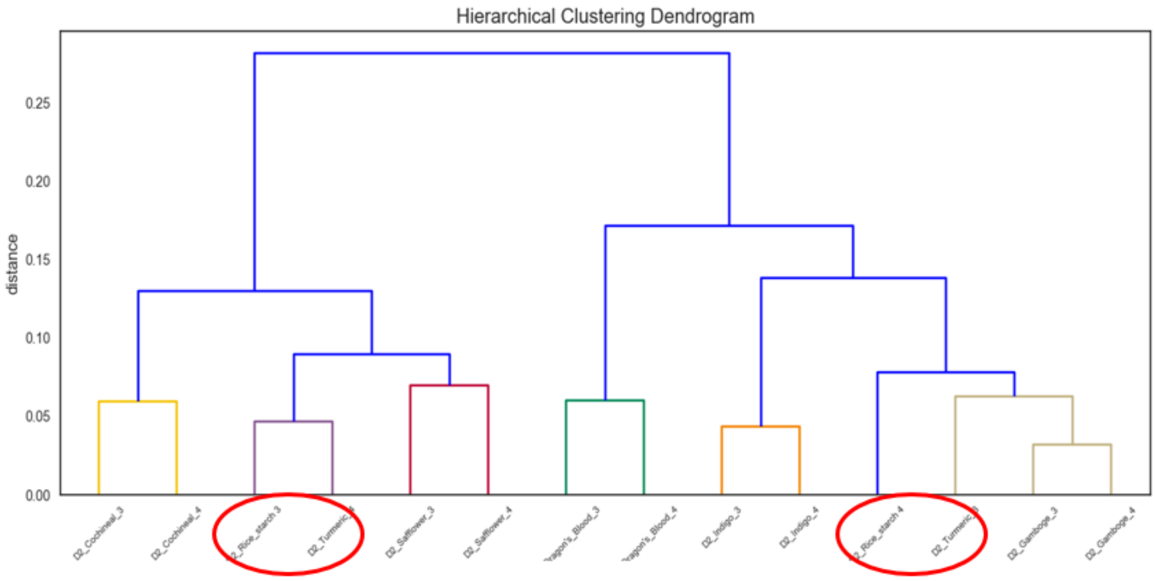


Fig. 115. Ward's HCA dendrogram of Near-IR spectra of different pigments and rice starch (binding medium) deposited on K17 paper. The wrongly clustered components are marked with the red circles.

In the generated dendrogram (Fig. 115), the differentiation of the spectra into two main groups can be observed. The spectra of the turmeric pigment are paired with the rice starch spectra and distributed in the separated nodes. The cochineal spectra are placed with the safflowers' spectra in the same node. However, the detailed node is a combination of the turmeric with rice starch pair and safflower pigment, which does not make a logical completeness. In the second main node, the dragon's blood pigment was placed in the same node with contrasting indigo and the poorly correlated gamboge pigment. The calculated fuzzy partitioning coefficient shows the incorrect number of components (5 components) in the data matrix (Fig. 116). The fuzzy c-means clustering algorithm (Fig. 117) also produced an unclear form of the results, as in the case of PCA.

© 2017 Tous droits réservés.

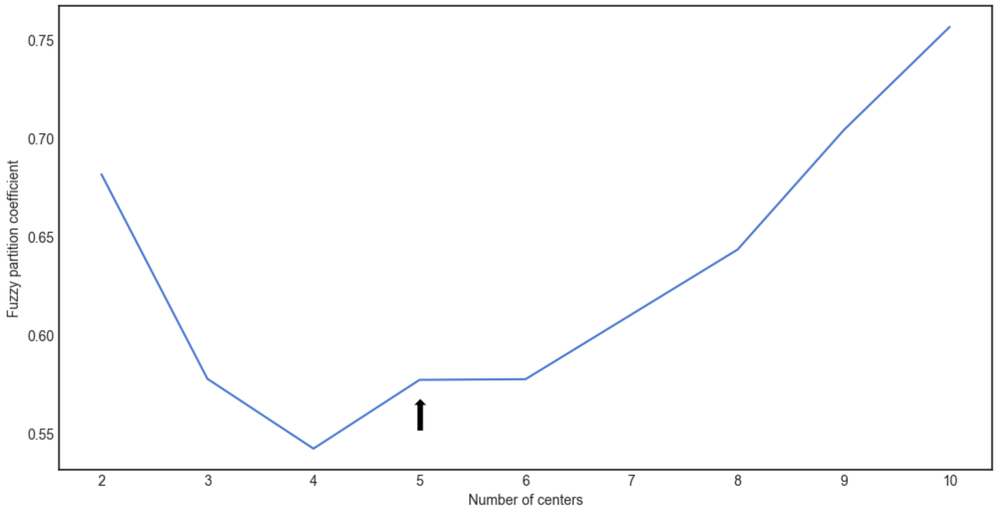
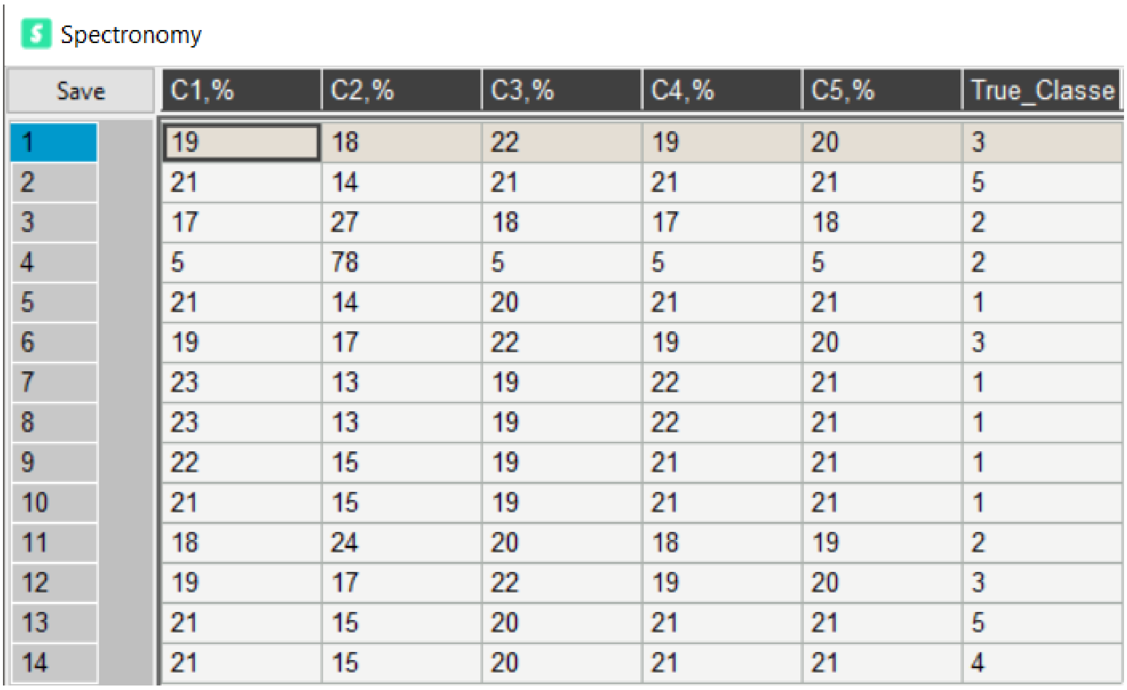


Fig. 116. Fuzzy partitioning coefficient calculated for spectra of 6 pigments and binding medium (rice starch) deposited on K17 paper.

The black arrow at the FPC plot (Fig. 116) indicated the incorrect number of the components in the data matrix (5 components). Not surprisingly, the fuzzy c-means clustering results calculated for 5 clusters demonstrate a weak spectra separation with no logical structure (Fig. 117).

© 2017 Tous droits réservés.



Save	C1,%	C2,%	C3,%	C4,%	C5,%	True_Classe
1	19	18	22	19	20	3
2	21	14	21	21	21	5
3	17	27	18	17	18	2
4	5	78	5	5	5	2
5	21	14	20	21	21	1
6	19	17	22	19	20	3
7	23	13	19	22	21	1
8	23	13	19	22	21	1
9	22	15	19	21	21	1
10	21	15	19	21	21	1
11	18	24	20	18	19	2
12	19	17	22	19	20	3
13	21	15	20	21	21	5
14	21	15	20	21	21	4

Fig. 117. Fuzzy c-means clustering table of Near-IR spectra after pre-processing of spectra from 6 pigments and binding medium (rice starch) deposited on K17 paper.

The weak cluster membership from the range of 20-24% can be observed, beside the one visible outlier of the 4th spectrum in the table, which has a membership of 78%. PCA and fuzzy c-means clustering were not useful in the pattern recognition of the deposited pigments on the K17 paper. Ward's HCA generates a weak separation of the pigments, which might be a problem while analysing the unknown samples.

In conclusion, based on these two examples, the Near-IR spectroscopy seems to be inappropriate for a separation of deposited pigments. Such a proceeding should be unambiguous and reproducible. In the case of real samples, a separation of pigments will be even more complicated due to their mixing and inhomogeneous paper structure.

Clustering of the pigment deposits on the paper (Mid-IR with spot diameter ~ 5mm)

The acquisition of 2 Mid-IR spectra of the deposited safflower (carthamin pigment), cochineal (carmine pigment), turmeric (curcumine pigment), gamboge pigment, dragon's blood pigment, indigo pigment and rise starch was made. The pigments

© 2017 Tous droits réservés.

were deposited on the absorbent Joseph paper. In addition, the single spectrum of the pure paper was recorded. The total number of collected spectra in the data matrix was 15 (12 spectra of pigments, 2 spectra of rice starch and 1 spectrum of Joseph paper). The total reflectivity was collected over 128 scans, at a resolution of 4 cm^{-1} using the spectrum from a golden mirror plate for a background acquisition. The spectra were collected in the range 6000 – 400 cm^{-1}.

The pre-processing procedure included only the spectra intensity scaling (Fig. 118).

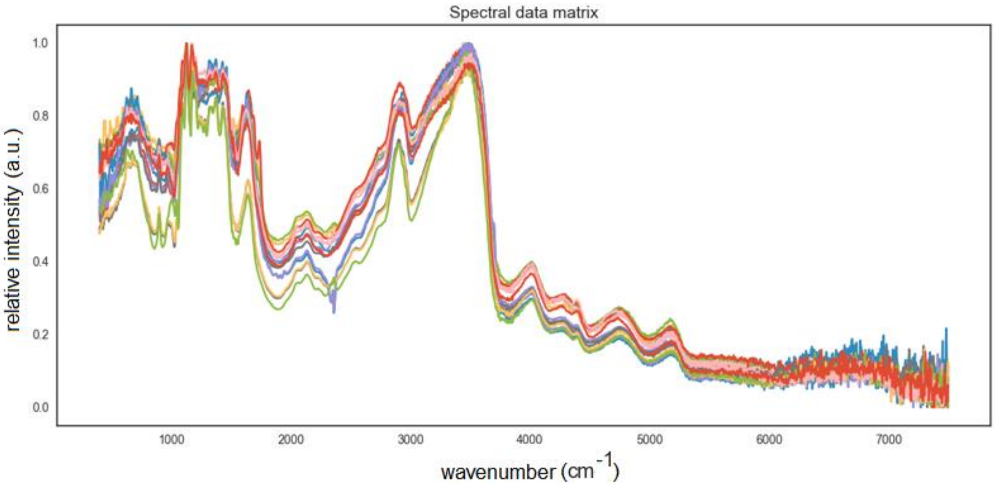


Fig. 118. Mid-IR spectra of different pigments and binding medium (rice starch) deposited on Joseph paper after intensity scaling.

By examination of the paper spectrum (Fig. 119) after the intensity scaling, the different variation of the spectral ranges in comparison to the pigments spectra is hardly observed (Fig. 118).

© 2017 Tous droits réservés.

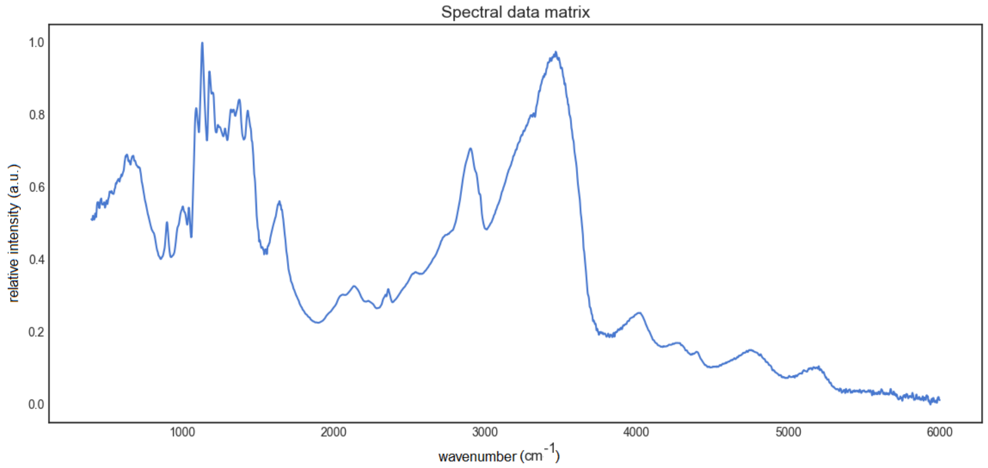


Fig. 119. Mid-IR spectrum after intensity scaling of the Joseph paper used as substrates for pigment deposition.

The fuzzy partitioning coefficient indicates a favourable separation of the Mid-IR spectra from the pigments, the binding medium (rice starch) and the substrate (Joseph paper), after intensity scaling (8 clusters) (Fig. 120).

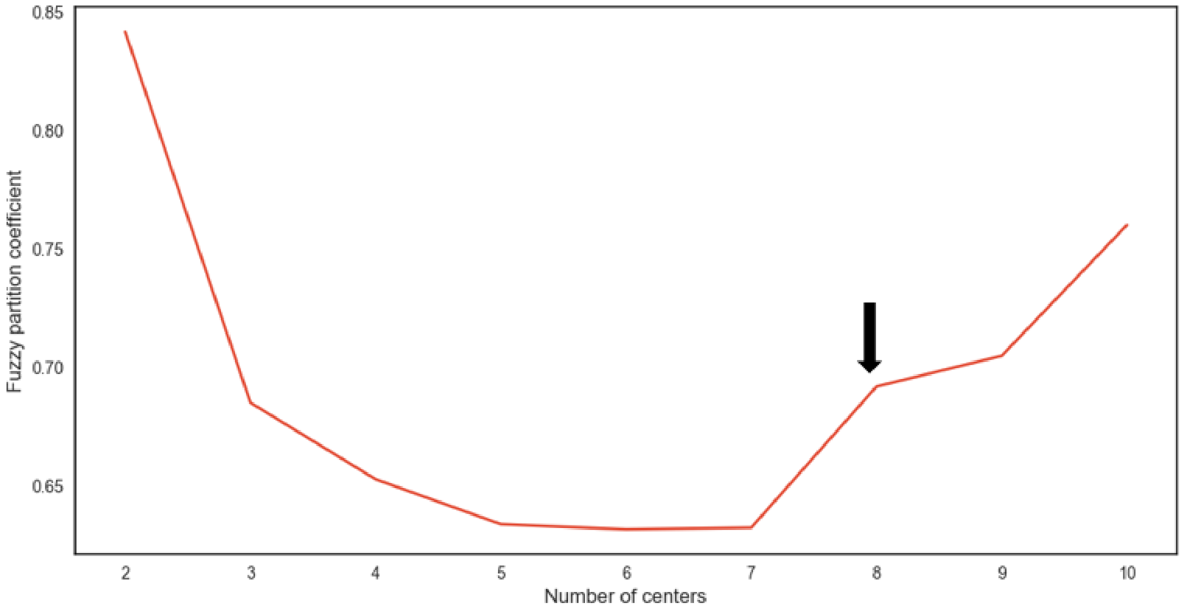
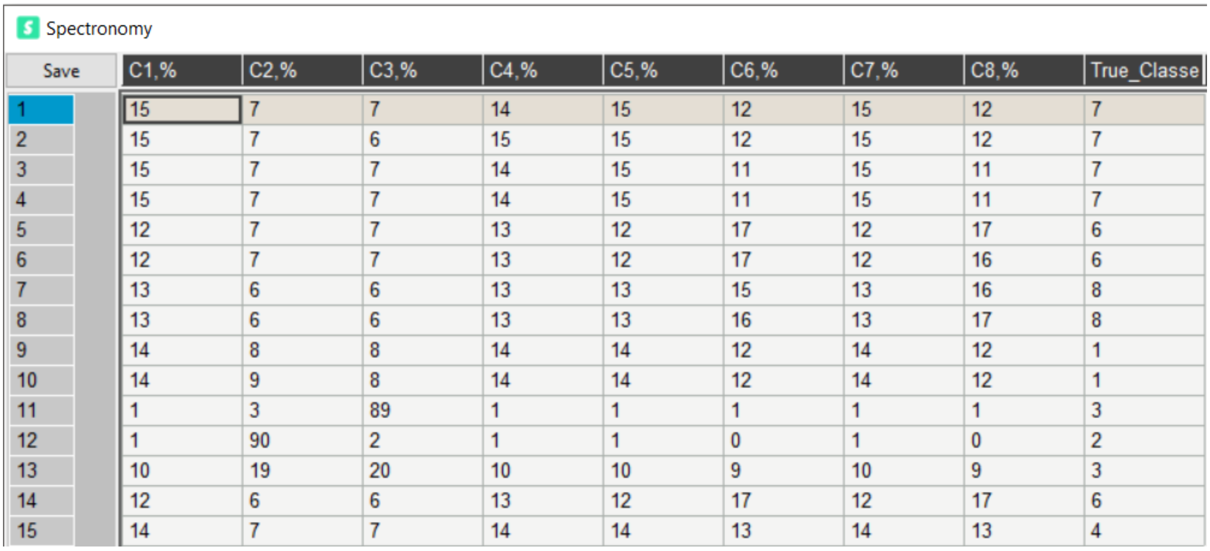


Fig. 120. Fuzzy partitioning coefficient calculated for the spectra of 6 pigments, binding medium (rice starch) and Joseph paper.

© 2017 Tous droits réservés.

Based on the FPC results, fuzzy c-means clustering was performed for 8 clusters (Fig. 121).



Save	C1,%	C2,%	C3,%	C4,%	C5,%	C6,%	C7,%	C8,%	True_Classe
1	15	7	7	14	15	12	15	12	7
2	15	7	6	15	15	12	15	12	7
3	15	7	7	14	15	11	15	11	7
4	15	7	7	14	15	11	15	11	7
5	12	7	7	13	12	17	12	17	6
6	12	7	7	13	12	17	12	16	6
7	13	6	6	13	13	15	13	16	8
8	13	6	6	13	13	16	13	17	8
9	14	8	8	14	14	12	14	12	1
10	14	9	8	14	14	12	14	12	1
11	1	3	89	1	1	1	1	1	3
12	1	90	2	1	1	0	1	0	2
13	10	19	20	10	10	9	10	9	3
14	12	6	6	13	12	17	12	17	6
15	14	7	7	14	14	13	14	13	4

Fig. 121. Fuzzy c-means clustering table of the Mid-IR spectra after intensity scaling of 6 pigments, binding medium (rice starch) and Joseph paper.

The spectra in the featured data matrix are not properly grouped. The 7th cluster contains safflower and cochineal pigments spectra, which are of the same colour range. In the 3rd cluster, the Joseph paper was clustered with one spectrum of rice starch. The dragon's blood pigment and turmeric are grouped in the 6th cluster, where the gamboge pigment and turmeric are located in the 8th cluster. Moreover, most of the spectra were separated with a weak cluster affiliation, between 14-17%, beside some exceptions for the spectra number 11 and 12 that are corresponding to rice starch. Complementary multivariate data processing algorithms were used: Ward's HCA and PCA with score labelling based on the Ward's HCA results. For the specification of the number of components in the data matrix, the Hubert's index was calculated and positioned in the dendrogram (Fig. 122).

© 2017 Tous droits réservés.

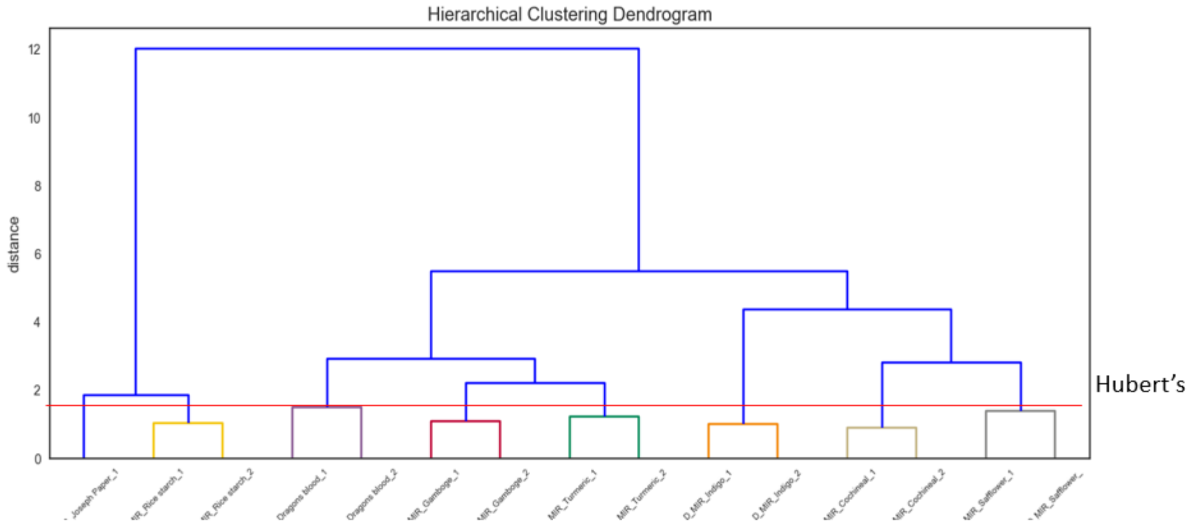


Fig. 122. Ward's HCA dendrogram of Mid-IR spectra of 6 pigments, rice starch (binding medium) and Joseph paper with specified threshold of optimal number of clusters calculated by via Hubert's index.

The Huber's index indicates the grouping of the matrix components into 8 clusters, which is complementary to the FPC (Fig. 121) and corresponds to the actual number of components. The Ward's HCA dendrogram demonstrates the logical separation of the spectra where the Joseph paper spectrum and the rice starch spectra are separated in the specific node of the dendrogram, disjointing these spectra from the pigments. The dragon's blood, turmeric and gamboge pigments are located in the same cluster (when a cut-off point is set for 4 clusters). These pigments are in the complementary range of colour, where gamboge and turmeric are more closely correlated and located in the descending node. In turn, to the right of the dendrogram the indigo pigments were disjoint from the cochineal and safflower pigments.

PCA with scores labelling from Ward's HCA clustering was used to demonstrate the location of the matrix components in the scatter plot (Fig. 123).

© 2017 Tous droits réservés.

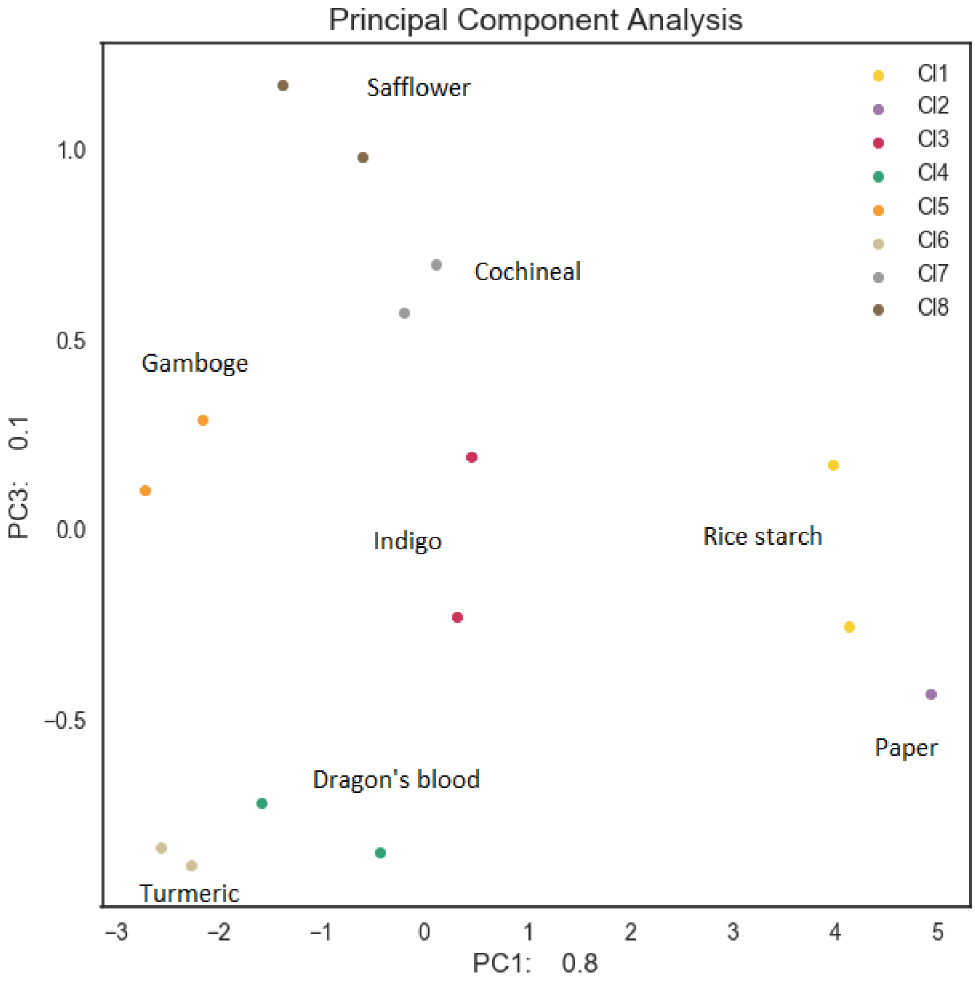


Fig. 123. PCA scatter plot of Mid-IR spectra from 6 pigments, rice starch (binding medium) and Joseph paper, after intensity scaling.

Total variance explained by two principal components is around 90%. After only intensity scaling, the separated pairs of pigments and rice starch can be distinguished. The disjoint spectra sets of pigments and rice starch with paper can be observed via first PC axis. This situation is also reflected in the Ward's HCA dendrogram. Fuzzy c-means clustering failed in the specification of the factual structure of the data matrix. However, the FPC calculation was indicative for the specification of the number of components in the data matrix. In view of the above, the featured data set was correctly grouped by two complementary techniques: PCA

© 2017 Tous droits réservés.

and Ward's HCA, where a number of components was specified by two parameters such Hubert's index and FPC.

The problem of clustering observations using a potentially large set of features may occur when the momentous variables need to be specified. The Witten and Tibshirani (Witten & Tibshirani 2010) proposed a novel framework for sparse clustering, in which one clusters the observations using an adaptively chosen subset of the features. In the case of the featured data set, the clustered features are corresponding to the momentous variables (wavenumbers) for spectra separation. The calculation was made by sparcl R package (Witten & Tibshirani 2010) which was coupled with the Spectronomy system. The sparse HCA clustering was used for the specification of the momentous variables during the pattern recognition. The p-value of feature weights was plotted on the spectra after intensity scaling (Fig. 124).

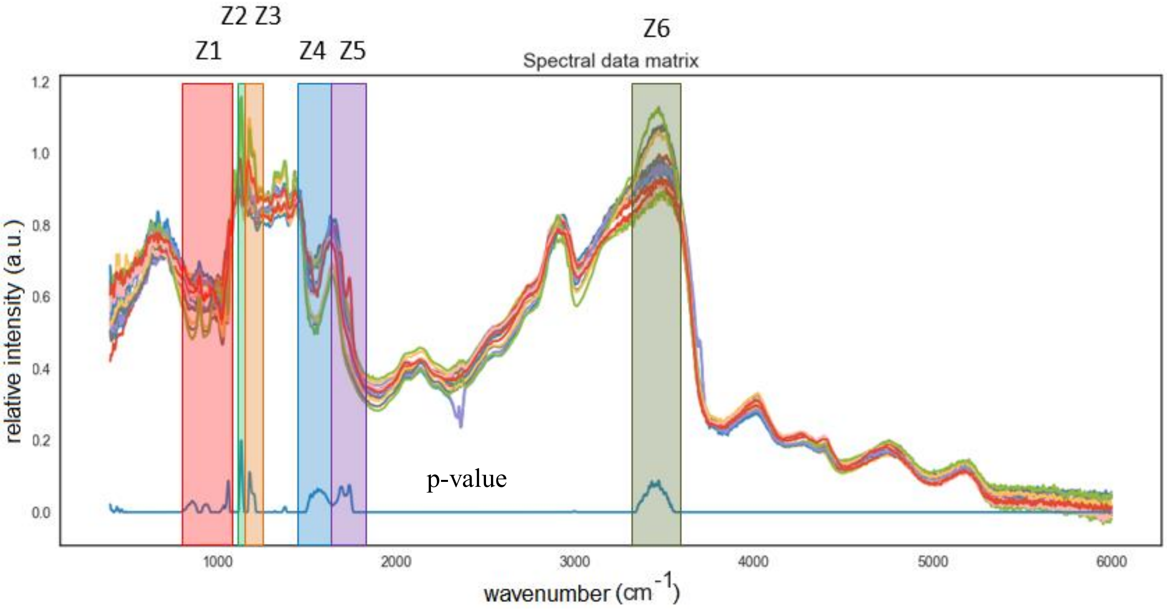


Fig. 124. Mid-IR spectra of deposited pigments and rice starch on Joseph paper. The p-value weights of features were plotted. 6 momentous spectral range were marked and labelled from Z1-Z6.

Mid-FTIR seems to be appropriate for distinguishing pigments deposited on one specific paper. The presented results of the automatic specification of the

© 2017 Tous droits réservés.

momentous variables may be significant. The „sparse" variables may be meaningful in terms of an identification of different pigments also with a support of a visual inspection of the spectra and spectral databases.

Clustering of the pigment deposits on the paper (µFTIR with spot area ~ 0.01 mm^2)

Reflection mid-FTIR spectra were also recorded *in situ* by use of an ALPHA micro FTIR spectrometer (Bruker) equipped with a DTGS detector and an external reflection module. The acquisition of the Mid-IR spectra of safflower (carthamin pigment) (3 spectra), cochineal (carmine pigment) (3 spectra), turmeric (curcumine pigment) (2 spectra), gamboge pigment (3 spectra), dragon's blood pigment (3 spectra), indigo pigment (3 spectra), rice starch (2 spectra) and a deposition substrate (Whatman paper) (2 spectra) was performed. The total number of collected spectra in the data matrix was 21. Total reflectivity was collected over 128 scans, at a resolution of 4 cm^{-1} using the spectrum from a golden mirror plate for background acquisition. The spectra were collected in the range 6000 – 400 cm^{-1}. The several types of pre-processing procedures and their configuration were tested for optimization of the spectra separation in the clustering algorithms. The results presenting below are corresponding to the optimal procedure for the pre-processing of spectra, which includes the spectra intensity scaling and the S-G derivatization (1st derivative of 2nd polynomial) (Fig. 125).

© 2017 Tous droits réservés.

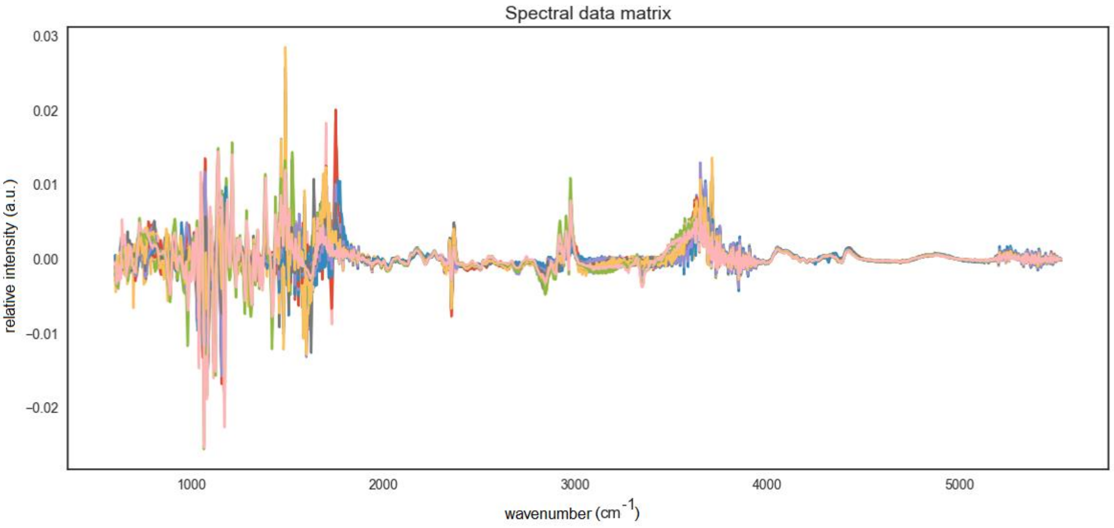


Fig. 125. Mid-IR spectra collected by µFTIR of 6 pigments and rice starch (binding medium) deposited on Whatman paper. Pre-processing step was conducted by spectra intensity scaling and S-G derivatization (1st derivative; 2nd polynomial).

Firstly, the exploratory data analysis was made by PCA (Fig. 126).

© 2017 Tous droits réservés.

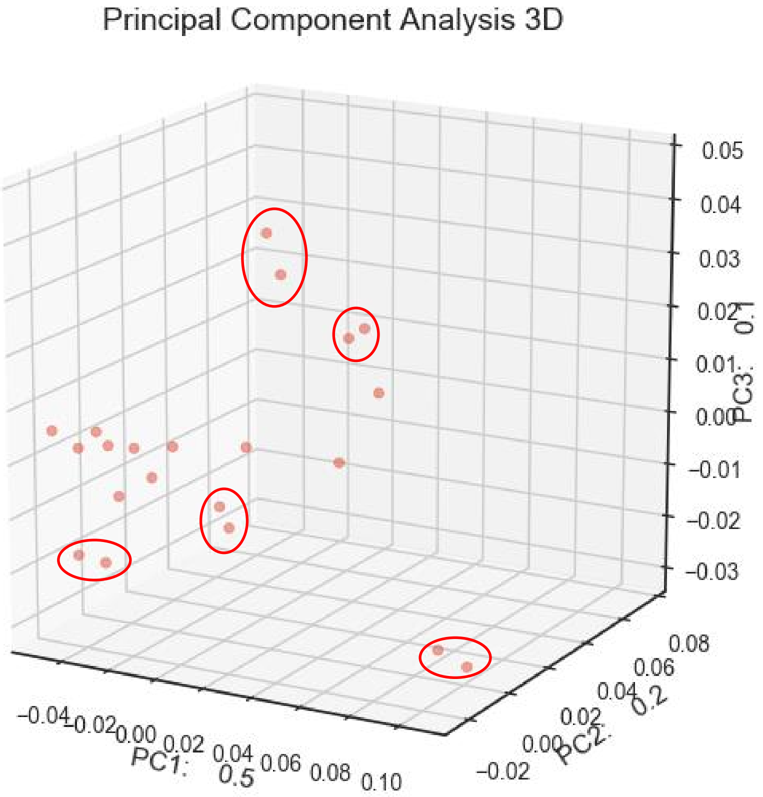


Fig. 126. 3D PCA scatter plot of Mid-IR spectra collected by μFTIR of 6 pigments, rice starch (binding medium) and Whatman paper (deposition substrate).

The cumulative variance explained by the 3 principal components is around 80%. The pattern formation in the scatter plot is hardly to distinguish. However, some of the scores are paired which was marked by red circles in the scatter plot (Fig. 126). The fuzzy partitioning coefficient indicates a favourable separation of the pigments, the binding medium (rice starch) and the substrate (Whatman paper) Mid-IR spectra after intensity scaling and S-G derivatization (8 clusters) (Fig. 127).

© 2017 Tous droits réservés.

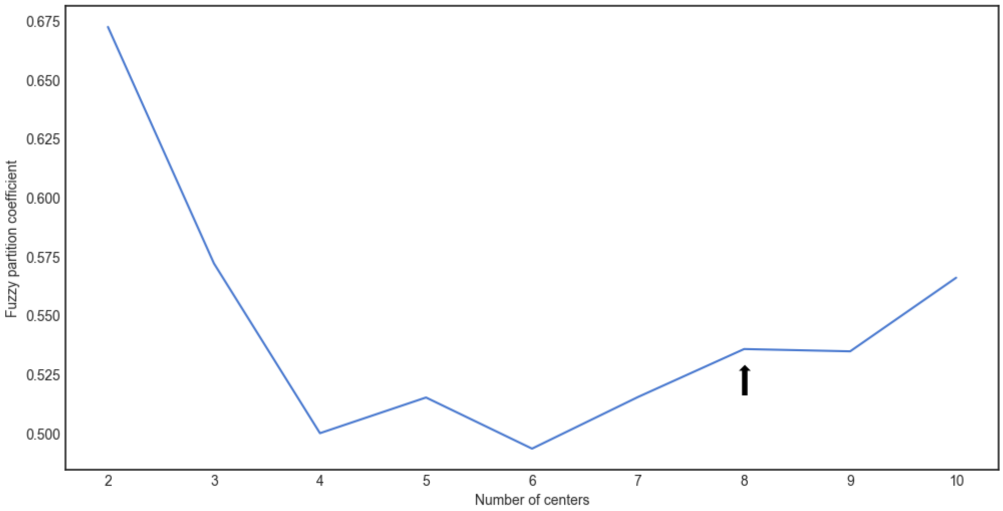
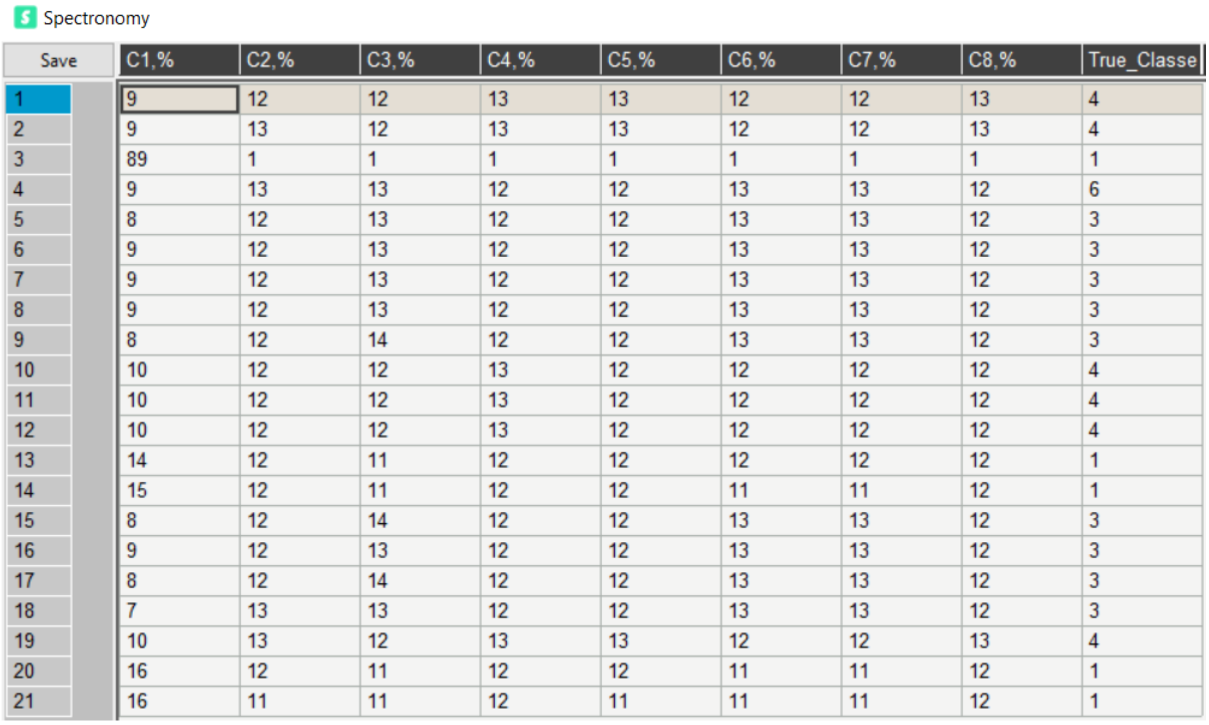


Fig. 127. Fuzzy partitioning coefficient calculated for the data set with spectra of 6 pigments, binding medium (rice starch) and Whatman paper.

Based on the FPC results, fuzzy c-means clustering was performed for 8 clusters (Fig. 128).



Save	C1,%	C2,%	C3,%	C4,%	C5,%	C6,%	C7,%	C8,%	True_Classe
1	9	12	12	13	13	12	12	13	4
2	9	13	12	13	13	12	12	13	4
3	89	1	1	1	1	1	1	1	1
4	9	13	13	12	12	13	13	12	6
5	8	12	13	12	12	13	13	12	3
6	9	12	13	12	12	13	13	12	3
7	9	12	13	12	12	13	13	12	3
8	9	12	13	12	12	13	13	12	3
9	8	12	14	12	12	13	13	12	3
10	10	12	12	13	12	12	12	12	4
11	10	12	12	13	12	12	12	12	4
12	10	12	12	13	12	12	12	12	4
13	14	12	11	12	12	12	12	12	1
14	15	12	11	12	12	11	11	12	1
15	8	12	14	12	12	13	13	12	3
16	9	12	13	12	12	13	13	12	3
17	8	12	14	12	12	13	13	12	3
18	7	13	13	12	12	13	13	12	3
19	10	13	12	13	13	12	12	13	4
20	16	12	11	12	12	11	11	12	1
21	16	11	11	12	11	11	11	12	1

Fig. 128. Fuzzy c-means clustering table of Mid-IR spectra from 6 pigments, binding medium (rice starch) and Whatman paper, after intensity scaling.

© 2017 Tous droits réservés.

The spectra in the featured data matrix are not properly grouped. The 3rd cluster contains dragon's blood, gamboge and safflower, which are not categorically in the same colour range. In the 1st cluster, there are both spectra of pigments, i.e. dragon's blood, as well as spectra of Whatman paper and rice starch. In the 4th cluster, all spectra of indigo pigment with a single spectrum of turmeric and cochineal are located. In the 6th cluster, only one single spectrum of dragon's blood is located. The fuzzy c-means clustering algorithm did not classify the spectra based on the cluster affiliation to the clusters number 5, 7 and 8. This fact is caused by significant similarities between the spectra which are unable to be properly separated by fuzzy c-means clustering (Stetco et al. 2015). Moreover, most of the spectra were separated with the weak cluster affiliation between 13-16%, beside one exception for spectrum number 3, which is corresponding to dragon's blood. Complementary multivariate data processing algorithms were used: Ward's HCA and PCA with score labelling based on the Ward's HCA results. For the specification of the number of components in the data matrix, the Hubert's index was calculated and positioned on the dendrogram (Fig. 129).

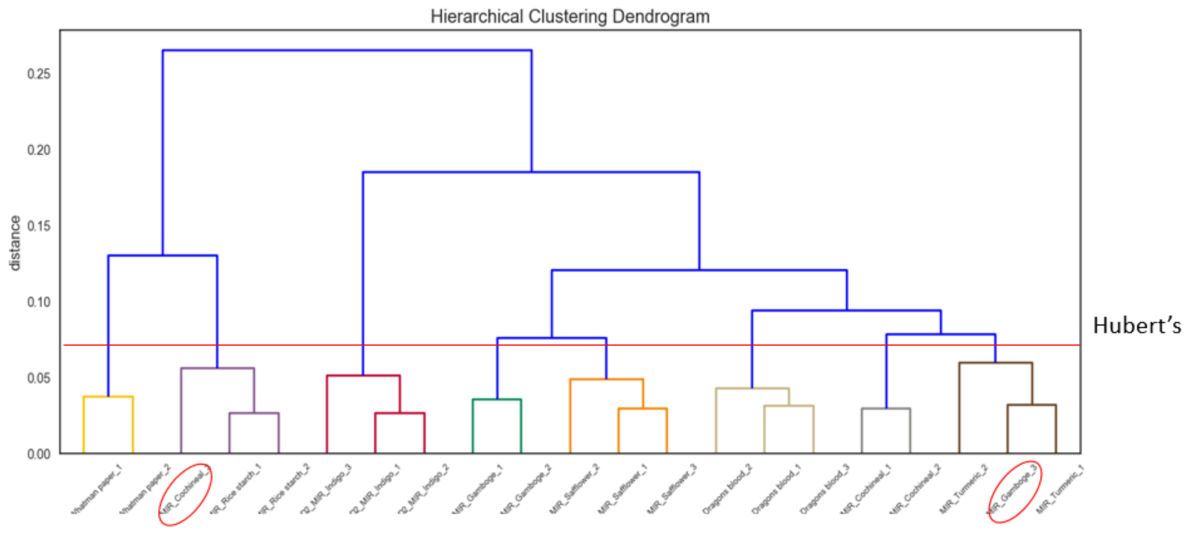


Fig. 129. Ward's HCA dendrogram of the Mid-IR spectra corresponding to 6 pigments, rice starch (binding medium) and Whatman paper, with specified threshold of the optimal cluster number of calculated via Hubert's index.

© 2017 Tous droits réservés.

The Huber's index indicates the grouping of the matrix components into 8 clusters, which is complementary to FPC (Fig. 127) and corresponds to the actual number of components. The Ward's HCA dendrogram demonstrates the logical separation of the spectra where the Whatman paper spectrum and the rice starch spectra are separated in the specific node of the dendrogram, disjointing these spectra from the pigments. Nonetheless, it should be highlighted that one by one spectrum of the gamboge pigment and cochineal pigment was wrongly clustered which was marked by the red circles. The first spectrum (gamboge) was grouped with the one from turmeric pigment, where the cochineal spectrum was located in the node with the rice starch spectra. The dragon's blood, turmeric and cochineal pigments are located in the same cluster (when the cut-off point is set for 5 clusters). These pigments are in the complementary range of colour (mostly dragon's blood and cochineal). In turn, to the left of the dendrogram, the indigo pigment was disjoint from the gamboge and safflower pigments.

PCA with scores labelling from Ward's HCA clustering was used to demonstrate the location of the matrix components in the scatter plot (Fig. 130).

© 2017 Tous droits réservés.

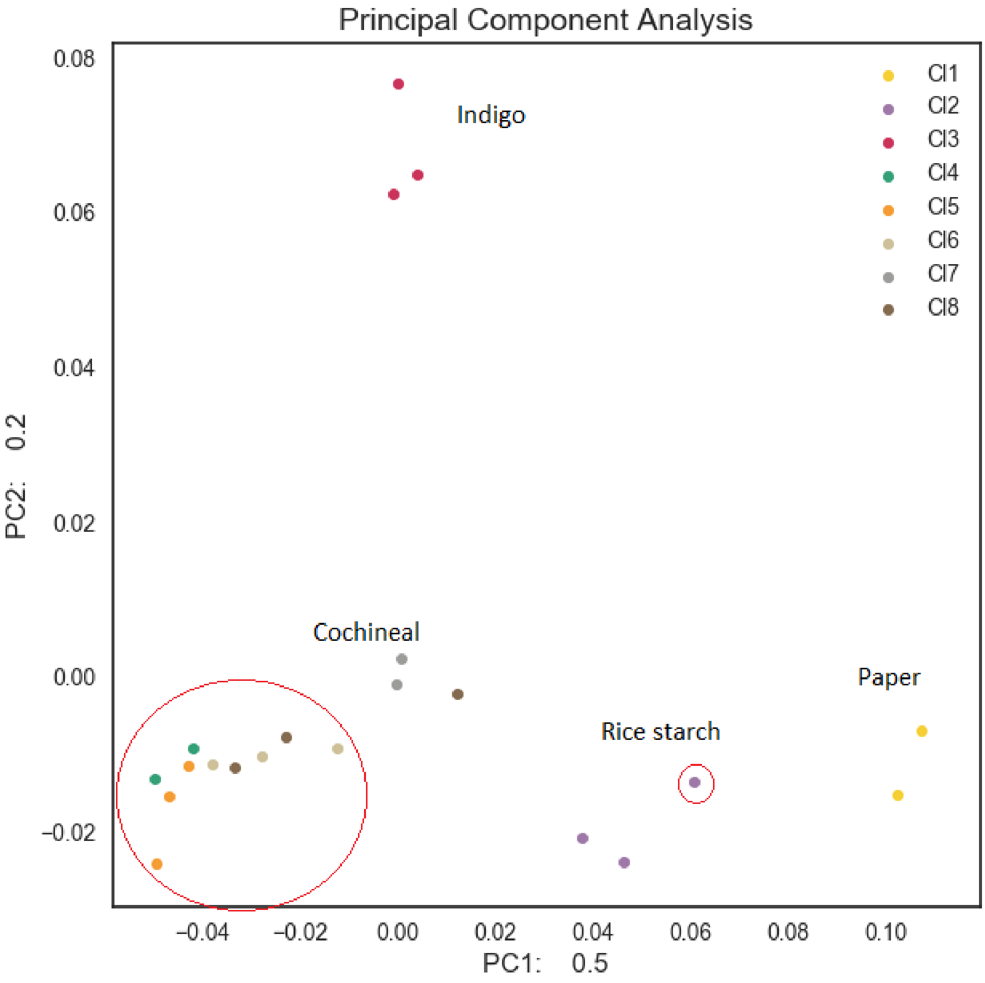


Fig. 130. PCA scatter plot of Mid-IR spectra from 6 pigments, rice starch (binding medium) and Whatman paper, after intensity scaling and S-G derivatization.

Total variance explained by two principal components is around 70%. After pre-processing, the grouping of spectra corrsponding to cochineal and indigo, rice starch and Whatman paper can be observed. However, the other pigment scores, i.e. dragon's blood, gamboge, turmeric and safflower are closely located in one area of the scatter plot (Fig. 130), what was marked by the red circle. It should be noted that the PCA with Ward's HCA labelling is not corresponding to the results presented in the dendrogram (Fig. 129). The degree of similarity of the spectra is close, which probably is reflected in the PCA scatter plot. In addition, the 3D PCA scatter plot did

© 2017 Tous droits réservés.

not visualize the group formation in more dimensions. Fuzzy c-means clustering algorithm also failed in the specification of the actual structure of the data matrix. However, the FPC calculation was indicative for the specification of the number of components in the data matrix. In view of the above, the featured data set was correctly grouped by only Ward's HCA, where the number of components was specified by two parameters such as Hubert's index and FPC.

Sparse HCA clustering was used for the specification of the momentous variables during the pattern recognition. The p-value of feature weights was plotted on the spectra after the pre-processing step (Fig. 131).

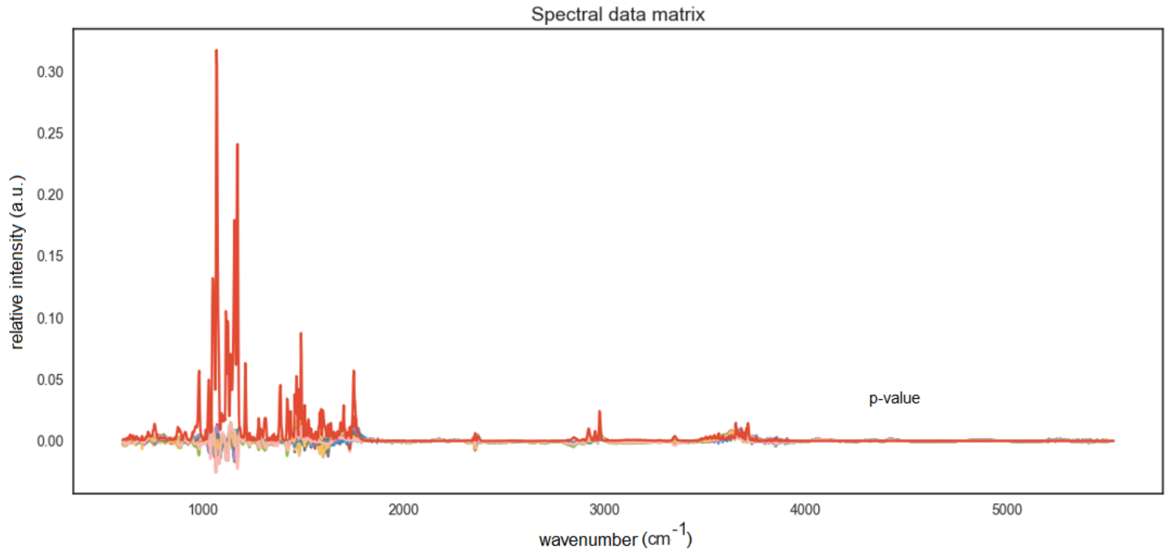


Fig. 131. Mid-IR spectra of deposited pigments and rice starch on Whatman paper. The p-value weights of features was plotted (red).

The significant variation of the p-value of the variables weight can be observed (Fig. 131). This variation is characterized by a large accumulation of variables of similar weight what in the case of sparse clustering is undesirable. This situation is probably due to the applied pre-processing method, which provides the best separation of the spectra in the Ward's HCA algorithm but limits the specification of the momentous variables for the featured data set.

© 2017 Tous droits réservés.

Mid-IR spectroscopy seems to be more appropriate for deposited pigments separation. For a spectrometer with a larger spot size (~ 5 mm), such proceeding is unambiguous. In the case of the analysis of the second data set by µFTIR (~0.01 mm^2), only two spectra were incorrectly clustered.

5.3.4. Conclusions

FTIR spectroscopy is a powerful analytical technique to study organic materials. However, in cultural heritage, since the sample under analysis is always a complicated matrix of several materials, a data analysis performed through peak-by-peak comparisons of the sample spectra with the standard ones is a tedious method that does not always provide good results (Sarmiento et al. 2011). To overcome this problem, a chemometric model based on PCA was developed to classify and identify pigments and binding media in artworks (Capobianco et al. 2017; Sessa et al. 2014; Carlesi et al. 2016). The independent problem is a classification of paper types based on their different properties.

In our work, we developed an innovative combination of cluster analysis methods combined with PCA to obtain the information about a number of pigments in the data matrix, as well as to separate the pigments' spectra based on their colour range and chemical structure. Near-IR spectroscopy was successfully applied to a differentiation of papers; moreover, a designation of distinctive papers with deposited pigments was also possible. In addition, a pigment coverage issue was taken under consideration with prospective results about a threshold presence. This might be a crucial aspect of clustering the spectra from real samples of Sumi-e art. The Mid-IR spectra of the deposited pigments were profitably grouped by the Ward's HCA method after the suitable pre-processing. It should be emphasized that an enormous contribution of the signal from the paper in the deposited pigments spectra was observed. This issue was solved owing to the optimized pre-processing and application of complementary multivariate data analysis techniques. Moreover,

© 2017 Tous droits réservés.

by application of the sparse clustering (sparse HCA), the designation of the momentous variables via the automatic mode was possible. The generated meaningful information about spectral ranges, which are important in the spectra classification, may have a crucial perspective in the identification of the spectra, which are covered by a signal from a substrate. We have obtained satisfactory results for pattern recognition based on the application of well-known pre-processing and multivariate data analysis techniques included in the Spectronomy system.

5.3.5. Future works

The development of the presented issue can take place through application of the classification models, such as a support vector machine classifier. In the future, a separation and classification of the mixed pigments will be taken under consideration. The blind source separation methods for a curve resolution, such as MCR-ALS will be used to specify pure components in the data matrix. The final step will be an application of the created model for real samples of the Sumi-e artwork for identification and classification of the pigments, binder and paper impurities. The presented methodology has a potential application for the analysis of aerosol particles, mostly where the differences of the spectra are inconsiderable.

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.

CONCLUSIONS AND FUTURE WORKS

© 2017 Tous droits réservés.

lilliad.univ-lille.fr

The objective of the present thesis was the development of multidimensional data analysis procedures dedicated to processing of Raman and FTIR spectra. We considered our main goals and presented several novel features of such a proceeding. Our results were adjusted to the major aspects of the Raman and FTIR analysis of the substrate-collected airborne aerosol particles. The potential of a single particle analysis by Raman microspectroscopy has been exploited by application of the originally designed analytical algorithm for an efficient description of chemical mixing of aerosol particles. The application of the algorithm to experimental data confirmed the potential in exceeding the limitations in trace component detection and quantitative analysis. Therefore, the new way of a sample description was presented. Due to this work, the new software that includes the described algorithm and several easy-to access, powerful chemometric techniques was provided. The developed data analysis system facilitates the reproducibility of the data processing applied to challenging aspects of pattern recognition in the scope of Raman and FTIR spectroscopy. The obtained results highlighted the potential of the presented system and aspects of its operation for data processing. Additionally, the important role of the suitable measurement conditions, such as particle collecting substrate, was evaluated. The work fulfills the set objectives and allows developing them in the future. A further development of the Spectronomy system and its application is planned. Several new algorithms are being developed for spectra pre-processing and analysis. Our intention is to add an ability to load files with other extensions – directly exported from an integral spectrometer software. The program will be supplemented with various classification methods, such as neural network classifier, support vector machine classifier, Naïve Bayes methods and stochastic gradient descent. An implementation of a regression analysis and relational database framework is also planned. Moreover, our hope is to adapt the program for a hyperspectral image analysis and an application of multivariate statistics for several matrices in a single run. The new features will be applied for the more complex matrices and big data.

© 2017 Tous droits réservés.

BIBLIOGRAPHY

Adachi, K. et al., 2014. Mixing state of regionally transported soot particles and the coating effect on their size and shape at a mountain site in Japan. *Journal of Geophysical Research*, 119(9.

Adams, F., 1994. Chemical characterization of atmospheric particles. *Topics in atmospheric*.

Afseth, N.K. & Kohler, A., 2012. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117.

Äijälä, M. et al., 2016. Chemometric analysis of aerosol mass spectra: exploratory methods to extract and classify anthropogenic aerosol chemotypes. *Atmospheric Chemistry and Physics Discussions*, (September).

Alessi, A. et al., 2013. Raman and IR investigation of silica nanoparticles structure. *Journal of Non-Crystalline Solids*, 362(1).

Aneja, V.P., Isherwood, A. & Morgan, P., 2012. Characterization of particulate matter (PM10) related to surface coal mining operations in Appalachia. *Atmospheric Environment*, 54.

Antunes, E.F. et al., 2006. Comparative study of first- and second-order Raman spectra of MWCNT at visible and infrared laser excitation. *Carbon*, 44(11).

Armstrong, J.T. & Buseck, P.R., 1985. A general characteristic fluorescence correction for the quantitative electron microbeam analysis of thick specimens, thin films and particles. *X-Ray Spectrometry*, 14(4).

Ault, A.P. et al., 2010. Characterization of the single particle mixing state of individual ship plume events measured at the port of Los Angeles. *Environmental Science and Technology*, 44(6).

Ault, A.P. et al., 2014. Heterogeneous reactivity of nitric acid with nascent sea spray aerosol: Large differences observed between and within individual particles. *Journal of Physical Chemistry Letters*, 5(15).

Ault, A.P. et al., 2013. Inside versus outside: Ion redistribution in nitric acid reacted sea spray aerosol particles as determined by single particle analysis. *Journal of the American Chemical Society*, 135(39).

Ault, A.P. et al., 2012. Single-particle SEM-EDX analysis of iron-containing coarse

© 2017 Tous droits réservés.

particulate matter in an urban environment: Sources and distribution of iron within Cleveland, Ohio. *Environmental Science and Technology*, 46(8).

Ault, A.P. et al., 2012. Single-particle SEM-EDX analysis of iron-containing coarse particulate matter in an urban environment: sources and distribution of iron within Cleveland, Ohio. *Environmental science & technology*, 46(8).

Ault, A.P. & Axson, J.L., 2017. Atmospheric Aerosol Chemistry: Spectroscopic and Microscopic Advances. *Analytical Chemistry*, 89(1).

Axson, J.L. et al., 2016. Lake Spray Aerosol: A Chemical Signature from Individual Ambient Particles. *Environmental Science and Technology*, 50(18).

Baccini, A., 2010. Statistique descriptive multidimensionnelle. *Publications de l'Institut de Mathématiques de Toulouse*.

Baek, S.-J. et al., 2015. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *The Analyst*, 140(1).

Banks, D. et al., 2002. Contaminant source characterization of the San Jose Mine, Oruro, Bolivia. *Geological Society, London, Special Publications*, 198(1).

Barnes, R.J., Dhanoa, M.S. & Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5).

Batonneau, Y. et al., 2006. Confocal microprobe Raman imaging of urban tropospheric aerosol particles. *Environmental Science and Technology*, 40(4).

Batonneau, Y. et al., 2003. Polarization effects of confocal Raman microspectrometry of crystal powders using interactive self-modeling analysis. *Journal of Physical Chemistry B*, 107(7).

Batonneau, Y. et al., 2001. Self-modeling mixture analysis of Raman microspectrometric investigations of dust emitted by lead and zinc smelters. In *Analytica Chimica Acta*.

Baustian, K.J. et al., 2012. Importance of aerosol composition, mixing state, and morphology for heterogeneous ice nucleation: A combined field and laboratory approach. *Journal of Geophysical Research Atmospheres*, 117(6).

Bergholt, M.S., Albro, M.B. & Stevens, M.M., 2017. Online quantitative monitoring of live cell engineered cartilage growth using diffuse fiber-optic Raman spectroscopy. *Biomaterials*, 140.

© 2017 Tous droits réservés.

Bernstein, S. et al., 2008. Application of CCSEM to heavy mineral deposits: Source of high-Ti ilmenite sand deposits of South Kerala beaches, SW India. *Journal of Geochemical Exploration*, 96(1).

Bersani, D. et al., 2016. Methodological evolutions of Raman spectroscopy in art and archaeology. *Anal. Methods*, 8(48).

Beyssac, O. et al., 2003. On the characterization of disordered and heterogeneous carbonaceous materials by Raman spectroscopy. In *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*.

Bi, Y. et al., 2016. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica Chimica Acta*, 909.

Birch, M.E. & Noll, J.D., 2004. Submicrometer elemental carbon as a selective measure of diesel particulate matter in coal mines. *Journal of environmental monitoring : JEM*, 6(10).

De Bock, L.A. et al., 2000. Single particle analysis of aerosols, observed in the marine boundary layer during the Monterey area ship tracks experiment (MAST), with respect to cloud droplet formation. *Journal of Atmospheric Chemistry*, 37(3).

Bondy, A.L. et al., 2017. Inland Sea Spray Aerosol Transport and Incomplete Chloride Depletion: Varying Degrees of Reactive Processing Observed during SOAS. *Environmental Science and Technology*, 51(17).

Bondy, A. L.; Kirpes, R. M.; Merzel, R. L.; Pratt, K. A.; Banaszak Holl, M. M.; Ault, A. P., 2017, Atomic Force Microscopy-Infrared Spectroscopy of Individual Atmospheric Aerosol Particles: Subdiffraction Limit Vibrational Spectroscopy and Morphological Analysis. *Analytical Chemistry, 89*, (17).

Bro, R. & Smilde, A.K., 2014. Principal component analysis. *Analytical methods*, 6.

Brunamonti, S. et al., 2015. Redistribution of black carbon in aerosol particles undergoing liquid-liquid phase separation. *Geophysical Research Letters*, 42(7).

Brusentsova, T.N. et al., 2010. Far infrared spectroscopy of carbonate minerals. *American Mineralogist*, 95(10).

Buajarern, J., Mitchem, L. & Reid, J.P., 2007. Characterizing the formation of organic layers on the surface of inorganic/aqueous aerosols by Raman spectroscopy. *Journal of Physical Chemistry A*, 111(46).

© 2017 Tous droits réservés.

Burkhart, J.E., McCawley, M.A. & Wheeler, R.W., 1987. Particle size distributions in underground coal mines. *American Industrial Hygiene Association journal*, 48(2).

Butler, H.J. et al., 2016. Using Raman spectroscopy to characterize biological materials. *Nature Protocols*, 11(4).

Bzdek, B.R., Pennington, M.R. & Johnston, M. V., 2012. Single particle chemical analysis of ambient ultrafine aerosol: A review. *Journal of Aerosol Science*, 52.

Cantrell, B.K. & Rubow, K.L., 1991. Development of personal diesel aerosol sampler design and performance criteria. *Mining Engineering*, 43(2).

Capobianco, G. et al., 2017. Chemometrics approach to FT-IR hyperspectral imaging analysis of degradation products in artwork cross-section. *Microchemical Journal*, 132.

Carlesi, S. et al., 2016. Multivariate analysis of combined reflectance FT-NIR and micro-Raman spectra on oil-paint models. *Microchemical Journal*, 124.

Carvalho-Oliveira, R. et al., 2015. Chemical composition modulates the adverse effects of particles on the mucociliary epithelium. *Clinics (São Paulo, Brazil)*, 70(10).

Castranova, V., 2000. From coal mine dust to quartz: Mechanisms of Pulmonary Pathogenicity. *Inhalation Toxicology*, 12.

Catelani, T., Pratesi, G. & Zoppi, M., 2014. Raman characterization of ambient airborne soot and associated mineral phases. *Aerosol Science and Technology*, 48(1).

Charrad, M. et al., 2014. NbClust : An R Package for Determining the. *Journal of Statistical Software*, 61(6).

Chen, H. et al., 2013. Chemical imaging analysis of environmental particles using the focused ion beam/scanning electron microscopy technique: microanalysis insights into atmospheric chemistry of fly ash. *The Analyst*, 138(2).

Chen, Y.C. & Thennadil, S.N., 2012. Insights into information contained in multiplicative scatter correction parameters and the potential for estimating particle size from these parameters. *Analytica Chimica Acta*, 746.

Cheng, W. et al., 2013. Surface chemical composition of size-fractionated urban walkway aerosols determined by x-ray photoelectron spectroscopy. *Aerosol Science and Technology*, 47(10).

© 2017 Tous droits réservés.

Christie, R.M., Robertson, S. & Taylor, S., 2007. Colour: Design & Creativity, 1(5).

Ciobanu, V.G. et al., 2009. Liquid - Liquid Phase Separation in Mixed Organic / Inorganic Aerosol Particles.

Claudio, T., Fuzzi, S. & Kokhanovsky, A., 2017. Primary and Secondary Sources of Atmospheric Aerosol.

Cochran, R.E. et al., 2017. Molecular Diversity of Sea Spray Aerosol Particles: Impact of Ocean Biology on Particle Composition and Hygroscopicity. *Chem*, 2(5).

Cprek, N. et al., 2007. Computer-controlled scanning electron microscopy (CCSEM) investigation of quartz in coal fly ash. *Fuel Processing Technology*, 88(11–12).

Craig, D., Mazilu, M. & Dholakia, K., 2015. Quantitative detection of pharmaceuticals using a combination of paper microfluidics and wavelength modulated Raman spectroscopy. *PLoS ONE*, 10(5).

Craig, R.L. et al., 2017. Spectroscopic Determination of Aerosol pH from Acid-Base Equilibria in Inorganic, Organic, and Mixed Systems. *Journal of Physical Chemistry A*, 121(30).

Craig, R. L.; Bondy, A. L.; Ault, A. P., 2015, Surface Enhanced Raman Spectroscopy Enables Observations of Previously Undetectable Secondary Organic Aerosol Components at the Individual Particle Level. *Analytical Chemistry, 87*, (15).

Crawford, I. et al., 2015. Evaluation of hierarchical agglomerative cluster analysis methods for discrimination of primary biological aerosol. *Atmospheric Measurement Techniques*, 8(11).

Cusack, M. et al., 2013. Source apportionment of fine PM and sub-micron particle number concentrations at a regional background site in the western Mediterranean: A 2.5 year study. *Atmospheric Chemistry and Physics*, 13(10).

Cvetković, Ž., Logar, M. & Rosić, A., 2013. Mineralogy and characterization of deposited particles of the aero sediments collected in the vicinity of power plants and the open pit coal mine: Kolubara (Serbia). *Environmental Science and Pollution Research*, 20(5).

Dallemagne, M.A., Huang, X.Y. & Eddingsaas, N.C., 2016. Variation in pH of Model Secondary Organic Aerosol during Liquid–Liquid Phase Separation. *The Journal of Physical Chemistry A*, 120(18).

Darchuk, L. et al., 2010. Argentinean prehistoric pigments' study by combined

© 2017 Tous droits réservés.

SEM/EDX and molecular spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 75(5).

Delhaye, M. & Dhamelincourt, P., 1975. Raman microprobe and microscope with laser excitation. *Journal of Raman Spectroscopy*, 3.

Doughty, D.C. & Hill, S.C., 2017. Automated aerosol Raman spectrometer for semi-continuous sampling of atmospheric aerosol. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 188.

Duda, R.O. & Hart, P.E., 1973. *Pattern Classification and Scene Analysis*.

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3).

Duque, L. et al., 2013. Elemental characterization of the airborne pollen surface using Electron Probe Microanalysis (EPMA). *Atmospheric Environment*, 75.

Dymińska, L., 2015. Imidazopyridines as a source of biological activity and their pharmacological potentials - Infrared and Raman spectroscopic evidence of their content in pharmaceuticals and plant materials. *Bioorganic and Medicinal Chemistry*, 23(18).

Ebben, C.J. et al., 2013. Size-resolved sea spray aerosol particles studied by vibrational sum frequency generation. *Journal of Physical Chemistry A*, 117(30).

Eilers, P.H.C. & Boelens, H.F.M., 2005. Baseline Correction with Asymmetric Least Squares Smoothing. *Life Sciences*.

Esch, L. & Hendryx, M., 2011. Chronic cardiovascular disease mortality in mountaintop mining areas of central Appalachian states. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*, 27(4).

Escribano, R. et al., 2001. Raman spectroscopy of carbon-containing particles. *Vibrational Spectroscopy*, 26(2).

European·Union, 2005. Thematic Strategy on Air Pollution. *Europa.eu*, 5.

Everitt, B.S. et al., 2011. *Cluster Analysis*.

Falgayrac, G. et al., 2006. Heterogeneous chemistry between PbSO4 and calcite microparticles using Raman microimaging. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 64(5).

© 2017 Tous droits réservés.

Falgayrac, G., Sobanska, S. & Brémard, C., 2013. Heterogeneous microchemistry between CdSO4 and CaCO3 particles under humidity and liquid water. *Journal of Hazardous Materials*, 248–249(1).

Falgayrac, G., Sobanska, S. & Brémard, C., 2014. Raman diagnostic of the reactivity between ZnSO4 and CaCO3 particles in humid air relevant to heterogeneous zinc chemistry in atmosphere. *Atmospheric Environment*, 85.

Ferraro, J.R., Nakamoto, K. & Brown, C.W., 2003. *Introductory Raman Spectroscopy*.

Finkelman, R.B., 1999. Trace elements in coal: environmental and health significance. *Biological trace element research*, 67(3).

Fitzgerald, E. et al., 2015. Comparison of the mixing state of long-range transported Asian and African mineral dust. *Atmospheric Environment*, 115.

Fletcher, R.A. et al., 2011. Microscopy and Microanalysis of Individual Collected Particles. In *Aerosol Measurement: Principles, Techniques, and Applications: Third Edition*.

Fontalvo-Gómez, M. et al., 2013. In-line near-infrared (NIR) and raman spectroscopy coupled with principal component analysis (PCA) for in situ evaluation of the transesterification reaction. *Applied Spectroscopy*, 67(10).

Forina, M. et al., 2008. Class-modeling techniques, classic and new, for old and new problems. *Chemometrics and Intelligent Laboratory Systems*, 93(2).

Fraund, M. et al., 2017. Elemental Mixing State of Aerosol Particles Collected in Central Amazonia during GoAmazon2014/15. *Atmosphere*, 8(9).

Fu, H. et al., 2012. Morphology, composition and mixing state of individual carbonaceous aerosol in urban Shanghai. *Atmospheric Chemistry and Physics*, 12(2).

Gabey, A.M. et al., 2011. The fluorescence properties of aerosol larger than 0.8 μ in urban and tropical rainforest locations. *Atmospheric Chemistry and Physics*, 11(11).

Gaffney, J.S., Marley, N.A. & Smith, K.J., 2015. Characterization of fine mode atmospheric aerosols by Raman microscopy and diffuse reflectance FTIR. *Journal of Physical Chemistry A*, 119(19).

Gautam, R. et al., 2015. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(1).

© 2017 Tous droits réservés.

Gauvin, R., Hovington, P. & Drouin, D., 1995. Quantification of Spherical Inclusions in the Scanning Electron-Microscope Using Monte-Carlo Simulations. *Scanning*, 17(4).

Geladi, P., MacDougall, D. & Martens, H., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3).

Genga, A. et al., 2012. SEM-EDS investigation on PM10 data collected in Central Italy: Principal Component Analysis and Hierarchical Cluster Analysis. *Chemistry Central Journal*, 6(Suppl 2).

Ghose, M.K. & Majee, S.R., 2007. Characteristics of hazardous airborne dust around an Indian surface coal mining area. *Environmental Monitoring and Assessment*, 130(1–3).

Goodwillie, T.G., 2003. Calculus III: Taylor series. *Geometry and Topology*, 7.

Gorry, P.A., 1990. General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method. *Analytical Chemistry*, 62(6).

Gosain, A. & Dahiya, S., 2016. Performance Analysis of Various Fuzzy Clustering Algorithms: A Review. In *Procedia Computer Science*.

Van Grieken, R. et al., 2000. Characterisation of Individual Aerosol Particles for Atmospheric and Cultural Heritage Studies. *Water, Air and Soil Pollution*, 4(123).

Griffiths, P.R. & De Haseth, J.A., 2007. *Fourier Transform Infrared Spectrometry*.

Guedes, A. et al., 2014. Pollen Raman spectra database: Application to the identification of airborne pollen. *Talanta*, 119.

Guilment, J.; Markel, S.; Windig, W., 1994, Appl. Spectrosc., 48, 320.

Guo, H., He, L. & Xing, B., 2017. Applications of surface-enhanced Raman spectroscopy in the analysis of nanoparticles in the environment. *Environmental Science: Nano*.

Guo, Q., Wu, W. & Massart, D.L., 1999. The robust normal variate transform for pattern recognition with near-infrared data. *Analytica Chimica Acta*, 382(1–2).

Hair, J.F. et al., 2010. Multivariate Data Analysis. *Vectors*.

Hamilton, J. C.; Gemperline P. J. , 1990, J. Chemom., 4,1-13

Hartigan, J.A., 1975. Clustering Algorithms. *Information Retrieval Data Structures and*

© 2017 Tous droits réservés.

Algorithms, 2.

Hartmann, D.L., Tank, A.M.G.K. & Rusticucci, M., 2013. IPCC Fifth Assessment Report, Climate Change 2013: The Physical Science Basis. *Ipcc*, AR5(January 2014).

Hartonen, K., Laitinen, T. & Riekkola, M.L., 2011. Current instrumentation for aerosol mass spectrometry. *TrAC - Trends in Analytical Chemistry*, 30(9).

Hinds, W.C., 1999. *Aerosol technology: Properties, Behavior, and Measurement of Airborne Particles.*

Hoffmann, T., Huang, R.J. & Kalberer, M., 2011. Atmospheric analytical chemistry. *Analytical Chemistry*, 83(12).

Hong, X. et al., 2010. Magnetic-field-assisted rapid ultrasensitive immunoassays using Fe3O4/ZnO/Au nanorices as Raman probes. *Biosensors and Bioelectronics*, 26(2).

De Hoog, J. et al., 2005. Thin-window electron probe X-ray microanalysis of individual atmospheric particles above the North Sea. *Atmospheric Environment*, 39(18).

Hoornaert, S., Moreton Godoi, R.H. & Van Grieken, R., 2004. Elemental and single particle aerosol characterisation at a background station in Kazakhstan. *Journal of Atmospheric Chemistry*, 48(3).

Hosgood, H.D. et al., 2012. Coal mining is associated with lung cancer risk in Xuanwei, China. *American Journal of Industrial Medicine*, 55(1).

Hovington, P. et al., 1997. CASINO: A new Monte Carlo code in C language for electron beam interaction---part I: Description of the program. *Scanning*, 19(1).

Hritz, A.D., Raymond, T.M. & Dutcher, D.D., 2016. A method for the direct measurement of surface tension of atmospherically relevant aerosol particles using atomic force microscopy. *Atmospheric Chemistry and Physics*, 16.

Hu, C. et al., 2015. Raman spectroscopy study of the transformation of the carbonaceous skeleton of a polymer-based nanoporous carbon along the thermal annealing pathway. *Carbon*, 85.

Huang, X. et al., 2005. Mapping and prediction of Coal Workers' Pneumoconiosis with bioavailable iron content in the bituminous coals. *Environmental Health Perspectives*, 113(8).

Huber, A.J. et al., 2007. Simultaneous IR material recognition and conductivity mapping by nanoscale near-field microscopy. *Advanced Materials*, 19(17).

© 2017 Tous droits réservés.

Hubert, L.J. & Levin, J.R., 1976. A general statistical framework for assesing categorical clustering in free recall. *Psychological Bulletin*, 83.

IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]

Ivleva, N.P. et al., 2013. Identification and characterization of individual airborne volcanic ash particles by Raman microspectroscopy. *Analytical and Bioanalytical Chemistry*, 405(28).

Ivleva, N.P. et al., 2007. Raman microspectroscopic analysis of size-resolved atmospheric aerosol particle samples collected with an ELPI: Soot, humic-like substances, and inorganic compounds. *Aerosol Science and Technology*, 41(7).

Ivleva, N.P., Niessner, R. & Panne, U., 2005. Characterization and discrimination of pollen by Raman microscopy. *Analytical and Bioanalytical Chemistry*, 381(1).

J . Osán , B . Alföldy , S . Kurunczi , S . Török , L . Bozó , J . Injuk, A.. W. and R.. V.G., 2001. Characterization of atmospheric aerosol particles over Lake Balaton , Hungary , using X-ray emission methods. *Idõjárás*, 105.

Jackson, J.E., 2005. *A User's Guide to Principal Components*.

Jacobsen, S.E., 2011. The Situation for Quinoa and Its Production in Southern Bolivia: From Economic Success to Environmental Disaster. *Journal of Agronomy and Crop Science*, 197(5).

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8).

Jain, A.K. & Dubes, R.C., 1988. Algorithms for Clustering Data. *Prentice Hall*, 355.

Jaumot, J., de Juan, A. & Tauler, R., 2015. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems*, 140.

Jawhari, T., Roid, A. & Casado, J., 1995. Raman spectroscopic characterization of some commercially available carbon black materials. *Carbon*, 33(11).

Jentzsch, P. V.; Kampe, B.; Ciobota, V.; Rosch, P.; Popp, J., 2013, Inorganic salts in atmospheric particulate matter: Raman spectroscopy as an analytical tool. *Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy, 115.*

© 2017 Tous droits réservés.

Jentzsch, P. V., Ciobotă, V., Kampe, B., Rösch, P., Popp, J., 2012, Origin of salt mixtures and mixed salts in atmospheric particulate matter. JRS 43,(4)

Vargas Jentzsch, P.; Bolanz, R. M.; Ciobotă, V.; Kampe, B.; Rösch, P.; Majzlan, J.; Popp, J., Raman spectroscopic study of calcium mixed salts of atmospheric importance. *Vibrational Spectroscopy* **2012,** *61*, (0), 206-213.

Jestel, N.L., 2010. Raman Spectroscopy. In *Process Analytical Technology*.

Jimoda, L. a., 2012. Effects of particulate matter on human health, the ecosystem, climate and materials: A review. *Facta universitatis-series*. 9(1).

Jonić, S., Sorzano, C.O.S. & Boisset, N., 2008. Comparison of single-particle analysis and electron tomography approaches: An overview. *Journal of Microscopy*, 232(3).

de Juan, A., Jaumot, J. & Tauler, R., 2014. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, 6(14).

De Juan, A. & Tauler, R., 2003. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta*, 500(1–2).

Jung, H.-J. et al., 2014. Combined use of quantitative ED-EPMA, Raman microspectrometry, and ATR-FTIR imaging techniques for the analysis of individual particles. *The Analyst*, 139(16).

KÃ¤mmer, E. et al., 2016. Single particle analysis of herpes simplex virus: comparing the dimensions of one and the same virions via atomic force and scanning electron microscopy. *Analytical and Bioanalytical Chemistry*, 408(15).

Kelly, S.T. et al., 2013. An environmental sample chamber for reliable scanning transmission x-ray microscopy measurements under water vapor. *Review of Scientific Instruments*, 84(7).

Khalilia, M.A. et al., 2014. Improvements to the relational fuzzy c-means clustering algorithm. *Pattern Recognition*, 47(12).

Kim, K.-H., Kabir, E. & Kabir, S., 2015. A review on the human health impact of airborne particulate matter. *Environment International*, 74.

Kizil, R., Irudayaraj, J. & Seetharaman, K., 2002. Characterization of irradiated starches

© 2017 Tous droits réservés.

by using FT-Raman and FTIR spectroscopy. *Journal of Agricultural and Food Chemistry*, 50(14).

Knauer, M. et al., 2009. Changes in structure and reactivity of soot during oxidation and gasification by oxygen, studied by micro-Raman spectroscopy and temperature programmed oxidation. *Aerosol Science and Technology*, 43(1).

Knauer, M. et al., 2009. Soot structure and reactivity analysis by Raman Microspectroscopy, Temperature-Programmed Oxidation, and High-Resolution Transmission Electron Microscopy. *Journal of Physical Chemistry A*, 113(50).

Krieger, U.K., Marcolli, C. & Reid, J.P., 2012. Exploring the complexity of aerosol particle properties and processes using single particle techniques. *Chemical Society Reviews*, 41(19).

Krzanowski, W.J. & Lai, Y.T., 1988. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics*, 44(1).

Kulkarni, P., Baron, P.A. & Willeke, K., 2011. *Aerosol Measurement: Principles, Techniques, and Applications: Third Edition*.

Kurth, L. et al., 2015. Atmospheric particulate matter in proximity to mountaintop coal mines: sources and potential environmental and human health impacts. *Environmental Geochemistry and Health*, 37(3).

Kwiecinska, B. et al., 2010. Raman spectroscopy of selected carbonaceous samples. *International Journal of Coal Geology*, 84(3–4).

Landen, D.D. et al., 2011. Coal dust exposure and mortality from ischemic heart disease among a cohort of U.S. coal miners. *American Journal of Industrial Medicine*, 54(10).

Laskin, A. et al., 2016. Progress in the Analysis of Complex Atmospheric Particles. *Annual Review of Analytical Chemistry*, 9(1).

Laskina, O. et al., 2015. Size Matters in the water uptake and hygroscopic growth of atmospherically relevant multicomponent aerosol particles. *Journal of Physical Chemistry A*, 119(19).

Laucks, M.L. et al., 2000. Physical and chemical (RAMAN) characterization of bioaerosols-pollen. *Journal of Aerosol Science*, 31(3).

Leona, M. et al., 2011. Nondestructive identification of natural and synthetic organic colorants in works of art by surface enhanced raman scattering. *Analytical*

© 2017 Tous droits réservés.

Chemistry, 83(11).

Lavine, B. & Workman, J., 2013. Chemometrics - review. *Analytical Chemistry*, 85(2).

Lewandowski, C.M., Co-investigator, N. & Lewandowski, C.M., 2015. *Infrared spectroscopy in Conservation Science*.

Li, J. et al., 2012. High quality of Jurassic Coals in the Southern and Eastern Junggar Coalfields, Xinjiang, NW China: Geochemical and mineralogical characteristics. *International Journal of Coal Geology*, 99.

Li, W. et al., 2016. A review of single aerosol particle studies in the atmosphere of East Asia: Morphology, mixing state, source, and heterogeneous reactions. *Journal of Cleaner Production*, 112(January 2013).

Li, W. et al., 2010. Size, composition, and mixing state of individual aerosol particles in a South China coastal city. *Journal of Environmental Sciences*, 22(4).

Li, W. & Shao, L., 2010a. Characterization of mineral particles in winter fog of Beijing analyzed by TEM and SEM. *Environmental Monitoring and Assessment*, 161(1–4).

Li, W. & Shao, L., 2010b. Mixing and water-soluble characteristics of particulate organic compounds in individual urban aerosol particles. *Journal of Geophysical Research Atmospheres*, 115(2).

Li, Y. et al., 2016. Method development and validation for pharmaceutical tablets analysis using transmission Raman spectroscopy. *International Journal of Pharmaceutics*, 498(1–2).

Li, Y. et al., 2015. Organelle specific imaging in live cells and immuno-labeling using resonance Raman probe. *Biomaterials*, 53.

Li, Y.-S. & Church, J.S., 2014. Raman spectroscopy in the analysis of food and pharmaceutical nanomaterials. *Journal of Food and Drug Analysis*, 22(1).

Liu, L. et al., 2003. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23).

Liu, Y. et al., 2008. Hygroscopic behavior of substrate-deposited particles studied by micro-FT-IR spectroscopy and complementary methods of particle analysis. *Analytical Chemistry*, 80(3).

Madden, H.H., 1978. Comments on the Savitzky-Golay Convolution Method for

© 2017 Tous droits réservés.

Least-Squares Fit Smoothing and Differentiation of Digital Data. *Analytical Chemistry*, 50(9).

Marques, M. et al., 2009. Correlation between optical, chemical and micro-structural parameters of high-rank coals and graphite. *International Journal of Coal Geology*, 77(3–4).

Martens, H., Nielsen, J.P. & Engelsen, S.B., 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, 75(3).

Maskey, S. & Ro, C.U., 2011. Quantitative energy-dispersive electron probe X-ray microanalysis for single-particle analysis and its application for characterizing atmospheric aerosol particles. *Pramana - Journal of Physics*, 76(2).

Massart, D.L. et al., 1997. *Handbook of Chemometrics and Qualimetrics*.

Matei, A. et al., 2005. Far-infrared spectra of amino acids. *Chemical Physics*, 316(1–3).

McArt, S.H. et al., 2017. High pesticide risk to honey bees despite low focal crop pollen collection during pollination of a mass blooming crop. *Scientific Reports*, 7.

McCreery, R.L., 2000. *Raman Spectroscopy for Chemical Analysis*.

McMurry, P.H., 2002. Chapter 17 A review of atmospheric aerosol measurements. *Developments in Environmental Science*, 1(C).

Mcneill, V.F., 2017. Atmospheric Aerosols : Clouds , Chemistry , and Climate.

Meng, L. et al., 2009. Environmental cumulative effects of coal underground mining. In *Procedia Earth and Planetary Science*.

Milligan, G.W. & Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2).

Moffet, R.C. et al., 2010. Automated chemical analysis of internally mixed aerosol particles using X-ray spectromicroscopy at the carbon K-edge. *Analytical Chemistry*, 82(19).

Moffet, R.C. et al., 2012. Iron speciation and mixing in single aerosol particles from the Asian continental outflow. *Journal of Geophysical Research Atmospheres*, 117(7).

Moffet, R.C. et al., 2016. Morphology and mixing of black carbon particles collected

© 2017 Tous droits réservés.

in central California during the CARES field study. *Atmospheric Chemistry and Physics*, 16(22).

Moricz, F., Walder, I.F. & Madai, F., 2009. Geochemical and Mineralogical Characterization of Waste Material from the Itos Sn-Ag Deposit, Bolivia. In *8th International Conference on Acid Rock Drainage*.

Navel, A. et al., 2015. Combining microscopy with spectroscopic and chemical methods for tracing the origin of atmospheric fallouts from mining sites. *Journal of Hazardous Materials*, 300.

Nickless, E.M. et al., 2014. Analytical FT-Raman spectroscopy to chemotype Leptospermum scoparium and generate predictive models for screening for dihydroxyacetone levels in floral nectar. *Journal of Raman Spectroscopy*, 45(10).

Noll, J.D. et al., 2007. Relationship between elemental carbon, total carbon, and diesel particulate matter in several underground metal/non-metal mines. *Environmental Science and Technology*, 41(3).

Nozière, B. et al., 2015. The Molecular Identification of Organic Compounds in the Atmosphere: State of the Art and Challenges. *Chemical Reviews*, 115(10).

O'Brien, R.E. et al., 2015. Chemical imaging of ambient aerosol particles: Observational constraints on mixing state parameterization. *Journal of Geophysical Research: Atmospheres*, 120(18).

O'Brien, R.E. et al., 2015. Liquid–Liquid Phase Separation in Aerosol Particles: Imaging at the Nanometer Scale. *Environmental Science & Technology*, 49(8).

Offroy, M. et al., 2015. Pushing back the limits of Raman imaging by coupling super-resolution and chemometrics for aerosols characterization. *Scientific Reports*, 5(1).

Ofner, J., Kamilli, K.A., et al., 2015. Chemometric analysis of multi-sensor hyperspectral images of coarse mode aerosol particles for the image-based investigation on aerosol particles. 17.

Ofner, J., Kamilli, K.A., et al., 2015. Chemometric Analysis of Multisensor Hyperspectral Images of Precipitated Atmospheric Particulate Matter. *Analytical Chemistry*, 87(18).

Ofner, J.; Deckert-Gaudig, T.; Kamilli, K. A.; Held, A.; Lohninger, H.; Deckert, V.; Lendl,

© 2017 Tous droits réservés.

B., 2016, Tip-Enhanced Raman Spectroscopy of Atmospherically Relevant Aerosol Nanoparticles. *Analytical Chemistr, 88*, (19), 9766-9772.

de Oliveira, R.C. et al., 2016. Bee pollen as a bioindicator of environmental pesticide contamination. *Chemosphere*, 163.

Osán, J. et al., 2000. Light Element Analysis of Individual Microparticles Using Thin-Window EPMA. *Microchimica Acta*, 132(2–4).

Otero, V. et al., 2014. Characterisation of metal carboxylates by Raman and infrared spectroscopy in works of art. *Journal of Raman Spectroscopy*, 45(11–12).

Ozaki, Y. & Šašić, S., 2007. Introduction to Raman Spectroscopy. *Pharmaceutical Applications of Raman Spectroscopy*.

Palmer, M.A. et al., 2010. Mountaintop Mining Consequences. *Science*, 327(5962).

Pandey, B., Agrawal, M. & Singh, S., 2014. Assessment of air pollution around coal mining area: Emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and principal component analysis. *Atmospheric Pollution Research*, 5(1).

Paradkar, M.M. & Irudayaraj, J., 2002. Discrimination and classification of beet and cane inverts in honey by FT-Raman spectroscopy. *Food Chemistry*, 76(2).

Di Pasquale, G. et al., 2016. Variations in the availability of pollen resources affect honey bee health. *PLoS ONE*, 11(9).

Pavilonis, B. et al., 2017. Characterization and risk of exposure to elements from artisanal gold mining operations in the Bolivian Andes. *Environmental Research*, 154.

Pedregosa, F. & Varoquaux, G., 2011. Scikit-learn: Machine learning in Python.

Pedrycz, W., 2005. Clustering and Fuzzy Clustering. *Knowledge-Based Clustering*.

Philpott, K.D., 2002. Evaluation of the Bogdanka Mine, Poland. *Polish Geological Institute Special Papers*, 7.

Potgieter-Vermaak, S.S. et al., 2005. Micro-structural characterization of black crust and laser cleaning of building stones by micro-Raman and SEM techniques. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 61(11–12).

Pozzi, F., 2011. Development of Innovative Analytical Procedures for the Identification

© 2017 Tous droits réservés.

of Organic Colorants of Interest in Art and Archaeology. *Thesis, Doctoral*.

Pósfai, M. & Buseck, P.R., 2010. Nature and Climate Effects of Individual Tropospheric Aerosol Particles. *Annual Review of Earth and Planetary Sciences*, 38(1).

Prather, K.A., Hatch, C.D. & Grassian, V.H., 2008. Analysis of Atmospheric Aerosols. *Annual Review of Analytical Chemistry*, 1(1).

Proctor, A. & Peter, M.A.S., 1980. Smoothing of Digital X-Ray Photoelectron Spectra by an Extended Sliding Least-Squares Approach. *Analytical Chemistry*, 52(14).

RAMAN, C. V. & KRISHNAN, K.S., 1928. A New Type of Secondary Radiation. *Nature*, 121(3048).

Randolph, T.W., 2006. Scale-based normalization of spectral data. *Cancer biomarkers : section A of Disease markers*, 2.

Reddy, D. & Jana, P.K., 2012. Initialization for K-means Clustering using Voronoi Diagram. *Procedia Technology*, 4.

Reisner, L.A., Cao, A. & Pandya, A.K., 2011. An integrated software system for processing, analyzing, and classifying Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 105(1).

Ren, T.X., Plush, B. & Aziz, N., 2011. Dust controls and monitoring practices on Australian longwalls. In *Procedia Engineering*.

Rinnan, Å., Berg, F. van den & Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, 28(10).

Ro, C.U. et al., 2003. A Monte Carlo program for quantitative electron-induced X-ray analysis of individual particles. *Analytical Chemistry*, 75(4).

Ro, C.U. et al., 2001. Chemical speciation of individual atmospheric particles using low-Z electron probe X-ray microanalysis: characterizing "Asian Dust" deposited with rainwater in Seoul, Korea. *Atmospheric Environment*, 35(29).

Ro, C.U., Kim, H.K. & Van Grieken, R., 2004. An Expert System for Chemical Speciation of Individual Particles Using Low-Z Particle Electron Probe X-ray Microanalysis Data. *Analytical Chemistry*, 76(5).

Ro, C.U., Osán, J. & Van Grieken, R., 1999. Determination of low-Z elements in individual environmental particles using windowless EPMA. *Analytical Chemistry*,

© 2017 Tous droits réservés.

71(8).

Robinson, N.H. et al., 2013. Cluster analysis of WIBS single-particle bioaerosol data. *Atmospheric Measurement Techniques*, 6(2).

Rodrigues, S. et al., 2011. Development of graphite-like particles from the high temperature treatment of carbonized anthracites. *International Journal of Coal Geology*, 85(2).

Rötting, T.S. et al., 2014. Environmental distribution and health impacts of As and Pb in crops and soils near Vinto smelter, Oruro, Bolivia. *International Journal of Environmental Science and Technology*, 11(4).

Roussel, S. et al., 2014. Process Analytical Technology for the Food Industry.

Roy, D., Singh, G. & Gosai, N., 2015. Identification of possible sources of atmospheric PM10 using particle size, SEM-EDS and XRD analysis, Jharia Coalfield Dhanbad, India. *Environmental Monitoring and Assessment*, 187(11).

Rusciano, G. et al., 2008. Surface-enhanced Raman scattering study of nano-sized organic carbon particles produced in combustion processes. *Carbon*, 46(2).

Ruspini, E.H., 1969. A new approach to clustering. *Information and Control*, 15(1).

Sadezky, A. et al., 2005. Raman microspectroscopy of soot and related carbonaceous materials: Spectral analysis and structural information. *Carbon*, 43(8).

Sarmiento, A. et al., 2011. Classification and identification of organic binding media in artworks by means of Fourier transform infrared spectroscopy and principal component analysis. *Analytical and Bioanalytical Chemistry*, 399(10).

Savitzky, A. & Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8).

Sawlowicz, Z., Łatkiewicz, A. & Stefaniak, E., 2005. Two-dimensional natural pyrite crystals and their formation. *Mineralogical Magazine*, 69(4).

Schill, G.P. & Tolbert, M.A., 2014. Heterogeneous ice nucleation on simulated sea-spray aerosol using Raman microscopy. *Journal of Physical Chemistry C*, 118(50).

Schmid, J. et al., 2011. Multiwavelength raman microspectroscopy for rapid prediction of soot oxidation reactivity. *Analytical Chemistry*, 83(4).

Schulte, F. et al., 2009. Characterization of pollen carotenoids with in situ and high-performance thin-layer chromatography supported resonant Raman

© 2017 Tous droits réservés.

spectroscopy. *Analytical Chemistry*, 81(20).

Schulte, F. et al., 2008. Chemical characterization and classification of pollen. *Analytical Chemistry*, 80(24).

Schuster, M.E. et al., 2011. Surface sensitive study to determine the reactivity of soot with the focus on the European emission standards IV and VI. *Journal of Physical Chemistry A*, 115(12).

Sessa, C., Bagán, H. & García, J.F., 2014. Influence of composition and roughness on the pigment mapping of paintings using mid-infrared fiberoptics reflectance spectroscopy (mid-IR FORS) and multivariate calibration. *Analytical and Bioanalytical Chemistry*, 406(26).

Shakya, K.M. et al., 2013. Similarities in STXM-NEXAFS spectra of atmospheric particles and secondary organic Aerosol generated from Glyoxal, α-Pinene, Isoprene, 1,2,4-Trimethylbenzene, and d-Limonene. *Aerosol Science and Technology*, 47(5).

Shao, R. et al., 2014. The fault feature extraction and classification of gear using principal component analysis and kernel principal component analysis based on the wavelet packet transform. *Measurement*, 54.

Silva, L.F.O. et al., 2009. Characterization of Santa Catarina (Brazil) coal with respect to human health and environmental concerns. *Environmental geochemistry and health*, 31(4).

da Silva Torres, E.A.F., Garbelotti, M.L. & Moita Neto, J.M., 2006. The application of hierarchical clusters analysis to the study of the composition of foods. *Food Chemistry*, 99(3).

Singer, P.C. & Stumm, W., 1970. Acidic Mine Drainage: The Rate-Determining Step. *Source: Science, New Series*, 167(3921).

Smith, D. S.; Kramer, J. R., 1999, EnViron. Int., 25, 295

Smith, R., Wright, K.L. & Ashton, L., 2016. Raman spectroscopy: an evolving technique for live cell studies. *The Analyst*, 141(12).

Smoliński, A. et al., 2014. Chemometric study of trace elements in hard coals of the upper Silesian Coal Basin, Poland. *Scientific World Journal*, 2014.

Sobanska, S. et al., 2006. Chemistry at level of individual aerosol particle using multivariate curve resolution of confocal Raman image. *Spectrochimica Acta -*

© 2017 Tous droits réservés.

Part A: Molecular and Biomolecular Spectroscopy, 64(5).

Sobanska, S. et al., 2012. Investigation of the chemical mixing state of individual asian dust particles by the combined use of electron probe X-ray microanalysis and raman microspectrometry. *Analytical Chemistry*, 84(7).

Sobanska, S. et al., 2000. Micro-characterisation of tropospheric aerosols from the Negev Desert, Israel. *Journal of Aerosol Science*, 31(SUPPL.1).

Sobanska, S. et al., 2014. Resolving the internal structure of individual atmospheric aerosol particle by the combination of Atomic Force Microscopy, ESEM-EDX, Raman and ToF-SIMS imaging. *Microchemical Journal*, 114.

Song, J. & Peng, P., 2009. Surface characterization of aerosol particles in Guangzhou, China: A study by XPS. *Aerosol Science and Technology*, 43(12).

Song, Y.C. et al., 2010. Chemical speciation of individual airborne particles by the combined use of quantitative energy-dispersive electron probe X-ray microanalysis and attenuated total reflection Fourier transform-infrared imaging techniques. *Analytical Chemistry*, 82(19).

de Souza Lins Borba, F., Saldanha Honorato, R. & de Juan, A., 2015. Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks. *Forensic Science International*, 249.

Stefaniak, E.A. et al., 2009. Combined SEM/EDX and micro-Raman spectroscopy analysis of uranium minerals from a former uranium mine. *Journal of Hazardous Materials*, 168(1).

Stefaniak, E.A. et al., 2009. Determination of chemical composition of individual airborne particles by SEM/EDX and micro-Raman spectrometry: A review. *Journal of Physics: Conference Series*, 162.

Stefaniak, E.A. et al., 2006. Molecular and elemental characterisation of mineral particles by means of parallel micro-Raman spectrometry and Scanning Electron Microscopy/Energy Dispersive X-ray Analysis. *Spectrochimica Acta - Part B Atomic Spectroscopy*, 61(7).

Steinier, J., Termonia, Y. & Deltour, J., 1972. Comments on Smoothing and Differentiation of Data by Simplified Least Square Procedure. *Analytical Chemistry*, 44(11).

Stetco, A., Zeng, X.J. & Keane, J., 2015. Fuzzy C-means++: Fuzzy C-means with

© 2017 Tous droits réservés.

effective seeding initialization. *Expert Systems with Applications*, 42(21).

Strola, S.A. et al., 2014. Single bacteria identification by Raman spectroscopy. *Journal of Biomedical Optics*, 19(11).

Sun, Y. et al., 2016. Aerosol characterization over the North China Plain: Haze life cycle and biomass burning impacts in summer. *Journal of Geophysical Research: Atmospheres*, 121(5).

Szaloki, I. et al., 2000. Quantitative characterization of individual aerosol particles by thin-window electron probe microanalysis combined with iterative simulation. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 55(7).

Sze, S.-K. et al., 2001. Raman spectroscopic characterization of carbonaceous aerosols. *Atmospheric Environment*, 35(3).

Tang, S.C.N. & Lo, I.M.C., 2013. Magnetic nanoparticles: Essential factors for sustainable environmental applications. *Water Research*, 47(8).

Tapia, J. et al., 2012. Geochemical background, baseline and origin of contaminants from sediments in the mining-impacted Altiplano and Eastern Cordillera of Oruro, Bolivia. *Geochemistry: Exploration, Environment, Analysis*, 12(1).

Tauler, R., Izquierdo-Ridorsa, A. & Casassas, E., 1993. Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution. *Chemometrics and Intelligent Laboratory Systems*, 18(3).

U.S Department Of Health And Human Services, 2003. Handbook for Dust Control in Mining. *Information Circular 9465*.

Wehbe, K. et al., 2013. The effect of optical substrates on micro-FTIR analysis of single mammalian cells. *Analytical and Bioanalytical Chemistry*, 405(4).

Wikipedia, 2016. Jablonski diagram. *Wikimedia Foundation*.

Windig, W. 1992, Chemom. Intell. Lab. Syst., 16,1.

Windig, W.; Guilment, 1991, J. Anal. Chem., 63, 1425.

Windig, W.; Heckler, C. E.; Agblevor, F. A.; Evans, R., 1992, J. Chemom. Intell. Lab. Syst. 14, 195

Windig, W.; Stephenson, D. A., 1992, Anal. Chem. 64, 2735.

Windig, W., 1994, Chemom. Intell. Lab. Syst. 23, 71.

© 2017 Tous droits réservés.

Windig, W.; Markel, S., 1993, J. Mol. Struct. 292, 161.

Windig W., 1997, Chemom. Intell. Lab. Syst., 36,3.

Windig, W. et al., 2002. Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach. *Analytical Chemistry*, 74(6).

Windig, W., 1997. Spectral data files for self-modeling curve resolution with examples using the Simplisma approach. *Chemometrics and Intelligent Laboratory Systems*, 36(1).

Windig, W. & Guilment, J., 1991. Interactive Self-Modeling Mixture Analysis. *Analytical Chemistry*, 63(14).

Witten, D.M. & Tibshirani, R., 2010. A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490).

Wold, S., Esbensen, K. & Geladi, P., 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3).

Worobiec, A. et al., 2007. Characterisation of concentrates of heavy mineral sands by micro-Raman spectrometry and CC-SEM/EDX with HCA. *Applied Geochemistry*, 22(9).

Worobiec, A. et al., 2006. Characterisation of individual atmospheric particles within the Royal Museum of the Wawel Castle in Cracow, Poland. *e-Preservation science*, 3.

Wu, J., Chen, J., et al., 2009. External validation measures for K-means clustering: A data distribution perspective. *Expert Systems with Applications*, 36(3 PART 2).

Wu, J., Xiong, H. & Chen, J., 2009. Adapting the right measures for K-means clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*.

Xu, L. et al., 2008. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta*, 616(2).

Yan-qiang, L. et al., 2011. New Progress on Coal Mine Dust in Recent Ten Years. *Procedia Engineering*, 26.

Yan, J. et al., 2013. Understanding the effect of surface/bulk defects on the photocatalytic activity of TiO2: anatase versus rutile. *Physical Chemistry Chemical*

© 2017 Tous droits réservés.

Physics, 15(26).

Yotova, G.I. et al., 2016. Urban air quality assessment using monitoring data of fractionized aerosol samples, chemometrics and meteorological conditions. *Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering*, 51(7).

Yu, D. et al., 2007. Computer-controlled scanning electron microscopy (CCSEM) investigation on the heterogeneous nature of mineral matter in six typical Chinese coals. In *Energy and Fuels*.

Ziegel, E.R., 2004. Statistics and Chemometrics for Analytical Chemistry. *Technometrics*, 46(4).

Zimmermann, B. et al., 2015. Characterizing aeroallergens by infrared spectroscopy of fungal spores and pollen. *PLoS ONE*, 10(4).

Zimmermann, F. et al., 2007. Environmental scanning electron microscopy (ESEM) as a new technique to determine the ice nucleation capability of individual atmospheric aerosol particles. *Atmospheric Environment*, 41(37).

Zong, S. et al., 2013. Surface enhanced Raman scattering traceable and glutathione responsive nanocarrier for the intracellular drug delivery. *Analytical Chemistry*, 85(4).

© 2017 Tous droits réservés.

© 2017 Tous droits réservés.