UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Multivariate approach-based system for the automated interpretation of spectra: application to pigments identification through Raman spectroscopy in art analysis

Juan José Gonzalez-Vidal

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Teoria del Senyal
i Comunicacions

Multivariate approach-based system for the automated interpretation of spectra: application to pigments identification through Raman spectroscopy in art analysis

Ph.D. Thesis

Juan José González-Vidal

Supervisors:
Rosanna Pérez-Pueyo and María José Soneira

Optical Communications Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, May 2017

Abstract

The application of spectroscopic techniques is crucial for art historians and conservators who require knowledge of materials used in works of art (pigments, dyes, binders, additives, ...) in particular instances. In this sense, the knowledge of pigments which were in use on the ancient artists' palettes is fundamental to preserve the art works. In addition, this knowledge is important to determine correct conservation approaches, to study degradation processes or authenticity-related issues. For instance, the proper interpretation of molecular signatures from a vibrational spectroscopy gives valuable information about the materials used by the artists. In this regard, the spectral identification is one of the essential interpretations to be performed, which is generally carried out by visual comparison between the unknown spectra with an appropriate database of reference spectra. This identification approach while being simple and intuitive may turn out a complex task which usually requires an experienced analyst and inevitably introduces an element of subjectivity linked to the intervention of the investigator. Besides, these analyses can be limited due to interferences from other phenomena like noises or admixtures. This task is further complicated when the spectra are to be interpreted by a software system. Hence, the noise impact must be reduced to have an effective identification and a robust strategy for processing multi-component spectra needs to be implemented. Clearly, a fully-automated data processing system for a reliable spectral interpretation is of practical interest.

Several automated methodologies were designed, developed and analysed in this Ph.D. Thesis for the purposes of art works analysis through Raman spectroscopy. In this sense, the usage of mathematical morphology together with p-spline fitting demonstrated to be a consistent combination in the application of data enhancement Raman spectra from artistic pigments. Besides, a generalised identification methodology to identify single- and multi- component spectra was developed. This identification method relies on automated spectral matching based on Principal Component Analysis (PCA) and Independent Components Analysis (ICA), being computationally efficient and conceptually simple. Moreover, a supervised classification methodology to automatically distinguish between Raman spectra showing small differences was developed. According to predefined reference training sets, the classification method is

able to classify unknown Raman spectra relying on PCA and Multiple Discriminant Analysis (MDA). Both the identification and classification methodologies successfully work using a single spectral observation for the unknown Raman spectra, with no user intervention or previous knowledge of the analysed sample.

The designed, developed and analysed automated methodologies for noise filtering and identification and classification of artistic pigments are integrated in a global system for the automated data interpretation of spectra from art works analysis implemented in this Ph.D. Thesis, namely ***PigmentsLab***. This software platform together with the integrated methodologies can play a good auxiliary role in the analysts' endpoint interpretation, providing insight from the raw spectral measurements into pigments. The system implementation provides an easy-to-use software platform and straightforward to update when new spectral data become available. The robust, reliable and consistent results obtained on Raman spectra demonstrated the competitiveness of the implemented data processing solutions. The system has great potential as an accurate and practical method for the automated interpretation of Raman spectra for not only pigment analysis, but essentially for any material group.

Resum

L'aplicació de tècniques espectroscòpiques és crucial per als historiadors i conservadors d'art que requereixen el coneixement dels materials utilitzats en obres d'art (pigments, colorants, aglutinants, additius, ...) en casos particulars. En aquest sentit, el coneixement de l'ús dels diferents pigments en les paletes dels artistes és fonamental per preservar les obres d'art. A més, aquest coneixement és important per determinar les estratègies de conservació correctes, per estudiar els processos de degradació o problemes relacionats amb l'autenticitat de les obres d'art. Per exemple, la interpretació adequada de les signatures moleculars d'una espectroscòpia vibracional proporciona informació valuosa sobre els materials utilitzats pels artistes. En aquest sentit, la identificació espectral és una de les interpretacions essencials a realitzar, que generalment es porta a terme mitjançant la comparació visual entre els espectres desconeguts amb una base de dades adequada dels espectres de referència. Aquesta estratègia d'identificació, tot i ser senzilla i intuïtiva, pot resultar una tasca complexa que requereix generalment d'un analista experimentat i inevitablement introdueix un element de subjectivitat vinculat a la intervenció de l'investigador. A més, aquestes anàlisis poden veure's limitades a causa d'interferències d'altres fenòmens com ara soroll o barreges de pigments. Aquesta tasca es complica encara més quan els espectres han de ser interpretats per una computadora. Per tant, l'impacte del soroll ha de ser reduït per tenir una identificació eficaç, i una estratègia robusta per al processament d'espectres de components múltiples ha de ser implementada. El desenvolupament d'un sistema de processament de dades totalment automatitzat per a una interpretació espectral fiable és d'evident interès pràctic.

Diverses metodologies automatitzades han estat dissenyades, desenvolupades i analitzades en aquesta tesi doctoral, focalitzades en l'anàlisi d'art mitjançant l'espectroscòpia Raman. En aquest sentit, l'ús de morfologia matemàtica juntament amb l'ajustament basat en *p-splines* va demostrar ser una combinació consistent en l'aplicació de millora de la qualitat d'espectres Raman de pigments artístics. A més, s'ha desenvolupat una metodologia d'identificació generalitzada per identificar els espectres Raman composats tant d'un sol pigment com de múltiples pigments. Aquest mètode d'identificació es basa en la cerca de coincidència espectral automatitzada basa-

da en l'anàlisi per components principals (PCA) i l'anàlisi per components independents (ICA), sent un mètode computacionalment eficient i conceptualment simple. D'altra banda, s'ha desenvolupat una metodologia de classificació supervisada per distingir automàticament entre espectres Raman que mostren petites diferències entre ells. A partir dels conjunts de referència predefinits de dades d'entrenament, el mètode de classificació és capaç de classificar els espectres Raman desconeguts mitjançant PCA i l'anàlisi discriminant múltiple (MDA). Tant la metodologia d'identificació com la de classificació funcionen correctament utilitzant només una sola observació espectral per als espectres Raman desconeguts, sense intervenció de l'usuari ni el coneixement previ de la mostra analitzada.

Les metodologies automatitzades dissenyades, desenvolupades i analitzades per al filtrat de soroll i la identificació i la classificació de pigments artístics estan integrades en un sistema global per a la interpretació automatitzada de dades a partir d'espectres mesurats en obres d'art què ha estat implementat en aquesta tesi doctoral, anomenat ***PigmentsLab***. Aquesta plataforma *software* juntament amb les metodologies integrades pot tenir un bon paper auxiliar en la interpretació de punt final dels analistes, proporcionant coneixement i valor a partir de les mesures espectrals en brut de pigments artístics. La implementació del sistema proporciona una plataforma fàcil d'utilitzar i també d'actualitzar quan es disposa de noves dades espectrals. Els resultats obtinguts en els espectres Raman analitzats, sent robustos, fiables i coherents, demostren la competitivitat de les solucions de tractament i processat de dades implementades. El sistema té un gran potencial com a mètode precís i pràctic per a la interpretació automàtica dels espectres Raman no només per a l'anàlisi de pigments artístics, sinó essencialment per a qualsevol grup de materials.

Acknowledgements

Quiero dar las gracias en primer lugar a mis directoras de tesis, Rosanna y María José, por transmitirme vuestra motivación e ilusión en la investigación, por vuestra confianza y la ayuda prestada para hacer realidad la implementación de este trabajo.

Al grupo de espectroscopía Raman de la UPC, por proporcionar los recursos necesarios para el desarrollo de esta tesis, a Conchi y Antonio, y en especial a Sergio.

I am very grateful to Wim Fremout and Steven Saverwyns from the Royal Institute for Cultural Heritage, Nadim C. Scherrer from the Bern University of Applied Sciences and Marta Anghelone from the Academy of Fine Arts Vienna for kindly providing a really valuable dataset of Raman spectra used in this Ph.D. Thesis.

Gràcies a l'equip de Gaia@DPCB, Nora, Javi, Jordi, i Marcial. Per tot el que he après amb vosaltres què de ben segur veureu reflectit en aquesta tesi.

A Ester, Cristian, Jus y Peter, porque con esta tesis ahora *tengo una zeta*! Al Miquel i al Xavi: això bé es mereix una bona celebració, *misters*! Gràcies també als que m'heu demostrat, com sempre, que us tinc al meu costat malgrat la distància, Laura, Membri i Lorenzo, i als que compartiu amb mi més que una afició una addicció per l'esport, Lluís, Pilar, Alex, Iker, José i la resta de la *secta*!

Vull dedicar un gràcies molt especial a les *malaltes*, Elena i Lourdes. És un autèntic plaer compartir el *bon-rotllisme* terapèutic de cada tarda amb unes autèntiques *cracks* com vosaltres. Per fer-me costat, per la Concòrdia i per la nostra sincronització mental què fins i tot ens espanta. Per tot el què vindrà i sé segur que viurem plegats, perqué sou la meva família que es tria. Gaudim del viatge, espartanes. Arooo!

Gràcies Ferran, per la teva paciència tots els cops que t'he fet desquadrar els *timings* (i els cops que queden!). Com aquesta tesi, espero que el procés de cerca d'espais i temps en comú ens porti a bon port. O millor dit, al Pirineu, on s'està tan a gust...

Per últim, vull donar les gràcies als meus germans, i molt especialment als meus pares, per tot el vostre suport, els ànims i consells i per creure sempre amb mi.

Acronyms and Abbreviations

BD Bhattarcharyya Distance
CI Continuous Integration
DBSCAN Density-Based Spatial Clustering of Applications with Noise
DMSO DiMethyl SulfOxide
ED Euclidean Distance
EM Expectation-Maximisation
FFT Fast Fourier Transform
FTIR Fourier Transform InfraRed Spectroscopy
FWHM Full Width at Half Maximum
GC Gas Chromatography
GCO Grup de Comunicacions Òptiques
GUIs Graphical User Interfaces
IB Identification Block
ICA Independent Components Analysis
IR InfraRed Spectroscopy
JDBC Java Database Connectivity
JMD Jeffries-Matusita Distance
LDA Linear Discriminant Analysis
LIBS Laser-Induced Breakdown Spectroscopy
MD Mahalanobis Distance
MDA Multiple Discriminant Analysis
MF Matching Factor
MSB Mixtures Separation Block
MVC Model-View-Controller
MySQL Open-source relational database management system based on SQL
OS Operating System
PC Principal Component
PCA Principal Component Analysis
PHP Hypertext Preprocessor
PLS Partial Least Squares
RMSE Root Mean Squared Error

SC Squared Cosine
SDF Spectral Data Format
SERS Surface-Enhanced Raman Spectroscopy
SNR Signal-to-Noise Ratio
SPC SPectroscopiC format
SQL Structured Query Language
SVN Apache SubVersioN
TSC Teoria del Senyal i Comunicacions
UML Unified Modelling Language
UPC Universitat Politècnica de Catalunya
UV-vis UltraViolet Visible Spectroscopy
WORA Write Once, Run Anywhere
XML eXtensible Markup Language
XRD X-Ray Diffraction
XRF X-Ray Fluorescence

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Thesis motivation

This Ph.D. Thesis belongs to the research developed by the Optical Communications Group -Grup de Comunicacions Òptiques (GCO)- of the Signal Theory and Communications departament -Teoria del Senyal i Comunicacions (TSC)- of the Technical University of Catalonia -Universitat Politècnica de Catalunya (UPC)- on the application of Raman spectroscopy in art analysis. Raman spectroscopy is an analytical technique, which provides qualitative and quantitative information regarding the molecular composition of an analysed material. The representation of the signature of a material obtained by Raman spectroscopy is called Raman spectrum. The Raman spectrum is composed of bands, whose positions are unique for each material and allow its identification. The main features of this technique are the following: the sample under test does not require any special preparation, allowing its analysis *in-situ*; it allows to obtain objective results in real time; it is a non-destructive technique; it allows a point to point analysis, that is, a sample mapping may be performed which allows to delimit accurately the analysed areas. Consequently, Raman spectroscopy provides a huge flexibility and versatility, and it is perfectly adapted to the demands of the analysis of the pigmentation of art works.

In this regard, the identification of a pigment from its Raman spectrum is generally carried out by comparison between an unknown Raman spectrum with an appropriate set of well-known Raman spectra of reference pigments, looking for which spectrum of the reference spectral library is the most similar to the unknown spectrum. Therefore, the developing of a rigorous and documented database of spectra from pigments acquired through any spectroscopic technique may be a top priority as it is a key element of the art materials identification.

On the other hand, a Raman spectrum can be divided into two parts: useful signal and noise. The useful signal is the part of the Raman spectrum that contains the

desired information (Raman bands) which allows the identification of the analysed material, while the noise is the part that does not correspond to Raman scattering, that is, the part that does not depend on the molecular structure of the material, such as random fluctuations of intensity for instance, which represents the largest source of uncertainty in the analysis of Raman signal. Thus, it is required to enhance the Raman information by filtering the noises out in order to avoid their effects on the spectral comparison and, therefore, on the identification. Furthermore, the comparison between spectra can be affected by the measurement conditions. Thus, a homogenization of the spectra by means of a spectral normalization is required, and some techniques, metrics and mathematical operators must be chosen to quantify the similarity between Raman spectra.

Frequently, the comparison between the unknown spectrum and the reference spectra is carried out by visual inspection by the analyst. This is an intuitive and simple method, but it is also slow and may be imprecise, especially in instances which spectra show a lot of bands and close together or if the analysed samples come from pigment mixtures, which usually are not in the reference spectral libraries. In addition, this way of working may introduce a component of subjectivity depending on the analyst's experience. This is why automating the identification process has become a hot topic of research nowadays. On another matter, artists often use combinations of pigments and mixtures in order to get different hues when making their art works. These combinations can mask the spectroscopic fingerprint of the analysed pigments and therefore mislead the processes used for the analysis and characterization of the palette used in a work of art. Thus, a particular analytical problem can be arisen when identifying unknown spectra that come from mixtures of pigments.

In conclusion, the general aim of this Ph.D. Thesis is to fully automate the data interpretation process devoted to the analysis of artistic pigments through Raman spectroscopy. The data interpretation process can be generally seen as a five-step process: acquisition, preparation, analysis, reporting and acting. Hence, the main objective of this Ph.D. Thesis is the automation of the data interpretation process applied to pigments analysis from the raw Raman spectra to the decision-making process in a systematic and objective way. This implies the development of several algorithms (noise filtering, matching-based identification, spectral classification, etc). For that purpose, the pigments interpretation process requires the design, development, implementation and analysis of a useful and supporting tool in order to retrieve a quick and automatic identification of the pigments used in works of art as part of their objective analysis. In addition, although the work developed in this Ph.D. Thesis is focused on the identification of pigments, the designed methodology is intended to be transparent to the spectroscopy and application, i.e. allowing the recognition and identification of any material group from its spectrum.

1.2 Thesis objectives

When analysing art materials through Raman spectroscopy, the measurement may be affected by several practical problems: shot noise, fluorescence's baseline, Raman bands distortion and shifting, ... These problems may hinder the comparison, whether automated or no, between the analysed sample spectrum and the database of reference Raman spectra, which eventually may make the identification impossible. In addition, a new problem can arise when analysing samples which were created with the mixture of several pigments. In this context, a spectroscopist may not achieve the identification of the unknown mixture due to the complexity of its spectral expression.

All these issues are way more complicated when working with spectral identification algorithms. For instance, a level of shot noise which is not significant when the recognition is visually carried out by an analyst may be critical for an automatic recognizer. In the same way, dealing with mixtures an identification system may become lost providing no match, and therefore no identification.

As a result, this Ph.D. is aimed to provide a useful tool in the identification of Raman spectra to analysts of art works independently of their experience. The main objectives to be achieved in this research project can be listed as follows:

A. The design, development and implementation of an automatic noise filtering methodology for enhancing Raman spectra from pigments, intended to obtain an improved Signal-to-Noise Ratio (SNR), making the Raman spectra easier to interpret

B. The design, development and implementation of an automatic identification methodology of Raman spectra from pigments, whether mixtures or not. That is, without prior knowledge of the composition of the analysed sample, analysing its performance in practical instances affected by differences between relative intensities, binding its impact and extending this analysis for the case of pigment mixtures, proposing a robust identification method. This implies the analysis of methods based on multivariate analysis allowing to extract the required information in order to achieve a proper pigments identification by means of their Raman spectrum

C. The design, development and implementation of an automatic classification methodology of Raman spectra from pigments, in order to perform a discrimination between the pigments found in natural and synthetic forms or in different crystalline structures, allowing the differentiation of several features (such as stability and hue) and also being used as chronological markers

D. The development of an extended, documented, detailed and robust database of Raman spectra of reference pigments. This implies the development of a visualisation system that will provide tools for exploring the spectral database allowing to handle spectra, as well as to show the pigments information and their spectroscopic details

E. The research study, design, development, implementation and analysis of a global automatic identification system of Raman spectra of pigments, stating its proper operation both in theoretical and experimental cases. This system may allow the pigments identification without prior knowledge of the composition of the analysed sample

F. The identification tool, although focused on the application to the analysis of art works, is aimed to be transparent to the spectroscopy and application, allowing the identification of any material based on its spectrum

1.3 Thesis outline

The main contents of this Ph.D. Thesis are structured in seven chapters, which develop the main objectives described in the previous section.

- **Chapter 1: Introduction.** In the current chapter, the justification, the objectives and the contents of this Ph.D. Thesis are introduced

- **Chapter 2: Literature review.** This chapter presents a review of different analytical techniques aimed at performing art analysis, and focusing on Raman spectroscopy, reviewing state-of-the-art data processing techniques in Raman spectroscopy applied to pigments analysis

- **Chapter 3: Acquisition and enhancement of Raman spectra from pigments.** This chapter describes the setup of the experimental measurement system and introduces the main noise sources present in Raman spectroscopy are introduced. Additionally, a novel automated noise filtering methodology aimed at helping in the interpretation of the Raman spectra from pigments is presented in this chapter in order to enhance the Raman information

- **Chapter 4: Automated analysis of Raman spectra from pigments.** This chapter describes the methodologies developed for the automated interpretation of Raman spectra from pigments. The developed methodologies were used to identify and classify Raman spectra from pigments which are commonly present in artist's paints in experimental environments, providing reliable and consistent results

- **Chapter 5: Global system of automated interpretation of spectra in art analysis.** This chapter describes the global software platform developed for the automated interpretation of spectra from pigments, which integrates the automated methodologies described in Chapter 3 and Chapter 4
- **Chapter 6: Raman characterisation of polymorphic forms of copper phthalocyanine blue under solvents and cleaning agents.** This chapter describes the Raman characterisation of different crystalline structures of one of the most widespread artists' blue pigment and the effect of applying solvents and cleaning agents on paint layers based on this pigment
- **Chapter 7: Conclusions and future work.** The main achievements are summarised and several promising directions for future work are exposed in this chapter

Finally, the main contributions associated to this research are listed in a **Publications** chapter, and also a **Bibliography** chapter is included picking up all the consulted sources, and several annexes as well containing additional information such as experimental results.

Chapter 2

Literature review

2.1 Chapter overview

This chapter introduces the importance of performing art analysis as well as presenting a review of different analytical techniques aimed at performing this kind of analysis. Among these techniques Raman spectroscopy stands out as an ideal technique suitable for the analysis of artistic pigments. Finally, a review of data processing in Raman spectroscopy applied to pigments analysis is presented.

2.2 Art analysis

Cultural heritage allows us to better understand previous generations and the history of where we come from. It is widely accepted that cultural heritage should be preserved to ensure long-term access and availability for future generations[1–3]. In this regard, essential information for tasks such as cataloguing and restoration may be obtained through the pigment analysis of an art work.

From prehistoric times humans have left their mark on their environment in the form of painted images, whether in the form of simple hand-prints, works of fine art or spray-can graffiti[4]. It seems that people have an underlying conscious or subconscious urge to mark their passing. It may be that primitive man made marks by scratching trees or rocks with stones as a way of marking a track, indicating a source of food or water or even marking territory. At some stage, however, it was discovered that some materials (called pigments) could be used to colour a surface, and the practice of painting was born and persists to this day.

In general, the art of painting was developed by different channels depending on the culture and the kind of civilization, and controlled by the availability of raw materials[5]. In fact, the selection of raw materials is directly influenced by climatic conditions, which involves geographical and chronological dependency, as part of the development of new

materials, keys to the evolution of artistic movements. Hence, a thorough knowledge of the pigments present in an art work is absolutely essential to gain insight into the materials composition and deterioration mechanisms in order to apply optimum restoration and conservation methodologies[6–9].

As might be expected, prehistoric painters used the pigments available in the vicinity of their homes. These were the so-called earth pigments, soot from burning animal fat and charcoal from the fire. The colours were yellow ochre, red ochre, and black. Water was the binding agent and enabled the pigment to be sprayed from the mouth or painted onto the surface using the fingers as brushes. Fig. 2.1 shows bison painted on a cave wall in Altamira, Spain. This painting is more than 30.000 years old.

Figure 2.1: Paintings of a bison in a cave wall in Altamira, Spain

The palette of these early people was limited to those materials readily to hand and requiring only the most basic technology for their preparation. Large parts of the spectrum of colours, notably blues and greens, were not available to them, yet they produced strikingly vivid images through skilful use of what they had.

The Egyptians began serious colour manufacture from about 4.000 BC. They introduced washing of pigments to increase their strength and purity. They also introduced new materials, the most famous of which was Egyptian blue - first produced around 3.000 BC. This is a very stable pigment and still appears as if fresh on wall paintings produced at that time.

Meanwhile, the Greeks' contribution to painting was the manufacture of white lead pigment which is still regarded as the whitest of the white pigments. It was the only white used in European easel paintings until the 19th century when its poisonous lead

content restricted its manufacture and sale as an artist's pigment.

The Romans made use of the pigments developed by the Egyptians and Greeks. One of the most important colours introduced by the Romans was Tyrian purple. It is mentioned in texts from 1.600 BC and was obtained from the hypobranchial gland of the molluscs murex trunculus and purpura haemastoma which were found in the Mediterranean Sea near Tyre.

Furthermore, the mediaeval palette and paintings were characterized by the use of clear, well-defined, bright colours. In addition to azurite, which had been used as a blue since the time of the ancient Egyptians, by far the most important blue in the middle ages was ultramarine. It was made by grinding the semi-precious mineral lapis lazuli, a rock containing the mineral lazurite, and was used in Afghanistan in the sixth century AC. During the renaissance, the colour blue was associated with purity and ultramarine was used to striking effect in paintings of the Virgin Mary. The combination of the price of the semi-precious stones and the cost of the process meant that ultramarine was more expensive than gold.

Alternatively, the development of the science of chemistry during the Industrial Revolution was partly driven by the textile dyeing industry, and led to the development of many new pigments. The first chemically synthesized pigment was made in Germany in 1704 by Diesbach. When using a batch contaminated with animal oil, he accidentally made a purple and then a blue pigment instead of the red he was trying to make. The blue became known as Prussian blue. However, ultramarine remained the most important blue pigment. Its cost was so high that in 1828 Jean-Baptiste Guimet manufactured a synthetic pigment, the so-called French ultramarine, chemically identical to Lapis Lazuli but with a cost of about a tenth of the current price for the cheapest Lapis[10]. In the early 19th century, many blue pigments (as the different forms of cobalt blue) were added to the existing varieties of blue[11] (see Fig. 2.2).

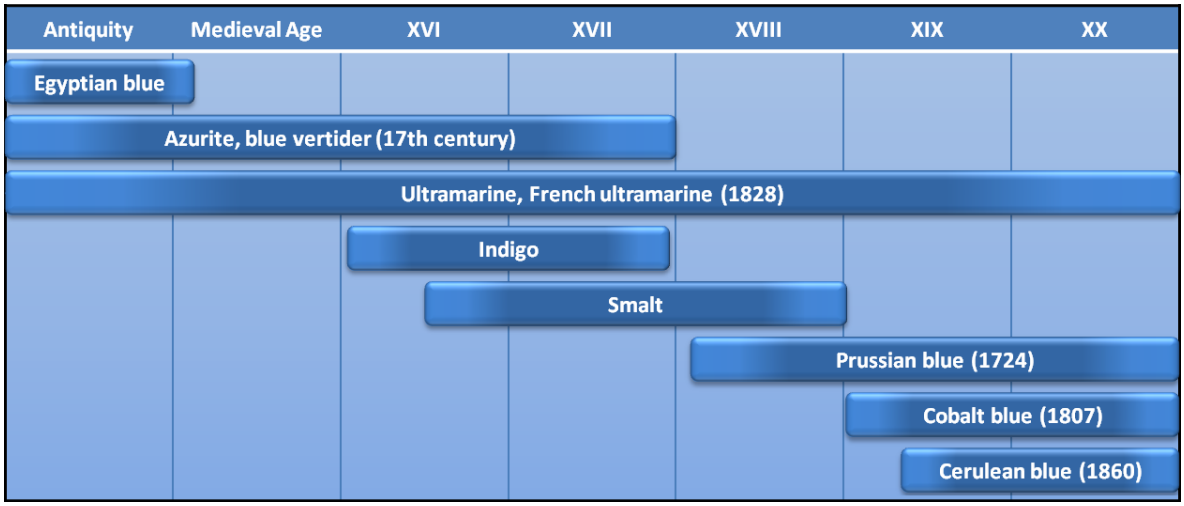


Figure 2.2: Usage timeline of main historical blue pigments

An art work containing the light sensitive pigment indigo would require strict conditions of exhibition, as faint light levels and controlled exposure times. Consequently, an objective and detailed study of the pigmentation of an art work is important to reveal information relating to the cataloguing, restoration and conservation of the work of art[12].

2.2.1 Analytical techniques

Since the first reported analytical studies and technical examinations of art and archaeological objects conducted in the late 18th century, analytical techniques and methods applied to the study of art works have exponentially increased[13]. An analytical approach to art and archaeological objects commenced around 1780 as a result of the progressive practical application of the ideas of art historians like Johann Wincklemann (1717-1768) on art and technical history[14]. According to these historians, knowledge of art objects should be based on the examination of the art work itself. From this time, analytical techniques and methods applied to the study of art works have constantly grown, particularly in the mid-late 20th century and, thus, nowadays there is a wide variety of scientific methodologies in the service of art and heritage conservation.

Traditionally, a number of naturally occurring substances have been used in art works given their ability. In addition to these natural products, synthetic substances were introduced into art a few years after the invention of the first synthetic polymer, cellulose nitrate, in 1846 by Schoenbein[15]. In the 20th century, the use of synthetic polymers has become widespread, and nowadays they are found not only as materials forming the art object but also as adhesives, varnishes or fillers of missing parts used in restoration works. Hence, the identification of the materials composing an art work is the most basic conservation science task[16]. Typically, this task is accomplished through non-destructive analytical techniques or, if needed, destructive techniques based on small samples analysis. The information yielded about the art work's material composition provides crucial information in the development of preventive conservation measures -such as lighting and humidity controls- and the selection of appropriate restoration treatments. Besides, it can also lead to discoveries about the art works' origin. Hence, analytical methods can provide insight into the time period of an art object, its authenticity, the authorship, and previous restoration treatments. In addition to leading to new interpretations, this information impacts the selection of treatments by conservators and conservation scientists.

The application of analytical techniques for the task of art materials analysis has been extensively improved in an attempt to enhance the sensitivity, repeatability and accuracy of the analytical results. Spectroscopic techniques, such as UltraViolet Visible Spectroscopy (UV-vis)[17], Fourier Transform InfraRed Spectroscopy (FTIR)[18] and

Raman spectroscopy[19], have been coupled with light microscopes for these purposes[20]. Synchrotron radiation FTIR microspectroscopy has been successfully applied to the analysis of art works[21]. Mass spectrometry has also been increasingly used as a detector system coupled with a chromatographic device[22]. Chromatographic methods have also improved in recent years[23–25]: Paper and thin layer chromatographic techniques have been progressively replaced with Gas Chromatography (GC), pyrolysis-GC, and high performance liquid chromatography. Although scarcely used owing to the limited availability of instrumentation, microbeam analytical techniques have also been incorporated into the list of instrumental techniques for art conservation purposes[26].

Nonetheless, it is important to note that the application of analytical techniques is usually difficult given the restrictions imposed on the number and size of the samples because of the unique and inimitable character of an art work. Non-invasive and non-destructive techniques are clearly favourable as they do not require the removal of art work samples. In this sense, Raman spectroscopy is now arguably the first-pass technique of choice for conservators and art historians who require knowledge of materials used in works of art (pigments, dyes, binders, additives, ...) in particular instances, due, among others, to its properties of non-destructivity, specificity and capability for in situ examination of art works[27,28]. Molecular signatures from Raman spectroscopy give much and valuable information about the materials used by the artists when making their art works[29,30]. In this sense, Raman spectroscopy is extended in analytical laboratories that work on art works because of their versatility for obtaining analytical information from both inorganic and organic materials. In addition to that, this techniques require a minimum preparation of sample enabling a high specificity as the bands in the Raman spectrum provide specific molecular fingerprintings. Additionally, an effective spatial resolution at the micron level is achieved, there is lower interference from inorganic substrates, and sample smoothness and transparency are not critical.

2.2.2 Raman spectroscopy

Raman Spectroscopy is a photonic technique that provides a signal scattered by a material under analysis when a beam of monochromatic light makes contact to it. The frequency of almost all the scattered energy matches the incident radiation -elastic scattering known as Rayleigh scattering- which, although intense, does not provide information on the composition of the analysed sample. However, there is a small part of the scattered light that is characteristic of the analysed material itself, resulting in a non-elastic scattering since there is an exchange of energy between the photon and the molecule during the collision. This scattering -known as Raman scattering- presents certain discrete frequencies located above and below the incident frequency, $v_i \pm v_r$, where these frequencies $\pm v_r$ are specific for each material, as are linked to its molecular

structure and its chemical bonds[31].

For this type of inelastic scattering two cases exist. If the radiation is scattered at a frequency lower than that of the incident light is known as Raman Stokes radiation. If, however, the radiation is scattered at a higher frequency than the incident one is called Raman Anti-Stokes. The signal scattered by the analysed material is graphically represented as a plot -the so-called Raman spectrum- of the light scattering (or intensity, in arbitrary units [a.u.]) versus Raman shift (in $[cm^{-1}]$) with respect to the frequency of the incident light (see Fig. 2.3). According to the Maxwell-Boltzmann energy distribution law, since most molecules are in the lowest energy state, it is much more likely the Stokes scattering to occur. Therefore, the Stokes scattered intensity is always higher than that of the Anti-Stokes scattering. Because of this difference, usually only the Stokes effect is measured, placing it in the positive x-axis.

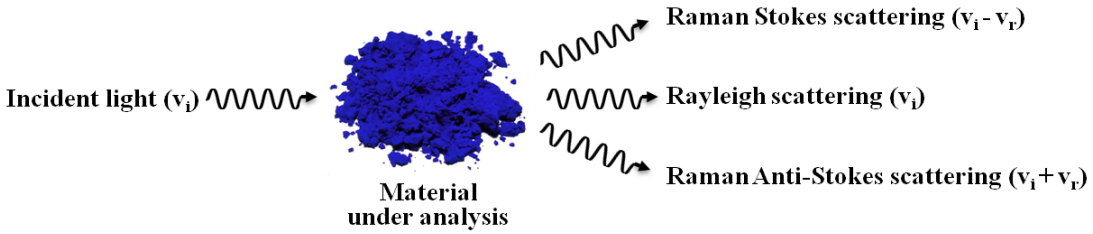


Figure 2.3: Light dispersion scheme when a monochromatic light makes contact to a material under analysis

A Raman spectrum is like a "fingerprint" of the analysed molecule: two Raman spectra coincident in number and position of their bands do not exist for two different molecules. Indeed, molecules composed of the same elements but in different proportions have different Raman spectra. Furthermore, molecules with exact chemical composition but different crystalline phase have different spectra. Therefore, the Raman spectrum is unique for each material. Moreover, it can be obtained from almost any chemical compound. This representation allows to view spectral bands (called Raman bands) centred at the Raman frequencies characteristic of each material (see Fig. 2.4). In a Raman spectrum the information is mainly on the position of each of the bands that identify unequivocally the material under analysis.

The identification of the analysed material through Raman spectroscopy is generally carried out by comparison between an unknown Raman spectrum with an appropriate set of reference Raman spectra. Once a Raman spectrum is measured, the main task of an analyst is to find out the material which matches the unknown one. Traditionally, this identification is based on visual comparison: when the spectrum of the analysed sample is obtained, it is compared with those well-known spectra, -called reference Raman spectra or simply patterns- looking for which Raman spectrum from those of the reference spectral library is the most similar to the unknown Raman spectrum (see 2.5). One limitation of the objective analysis of art works is the availability

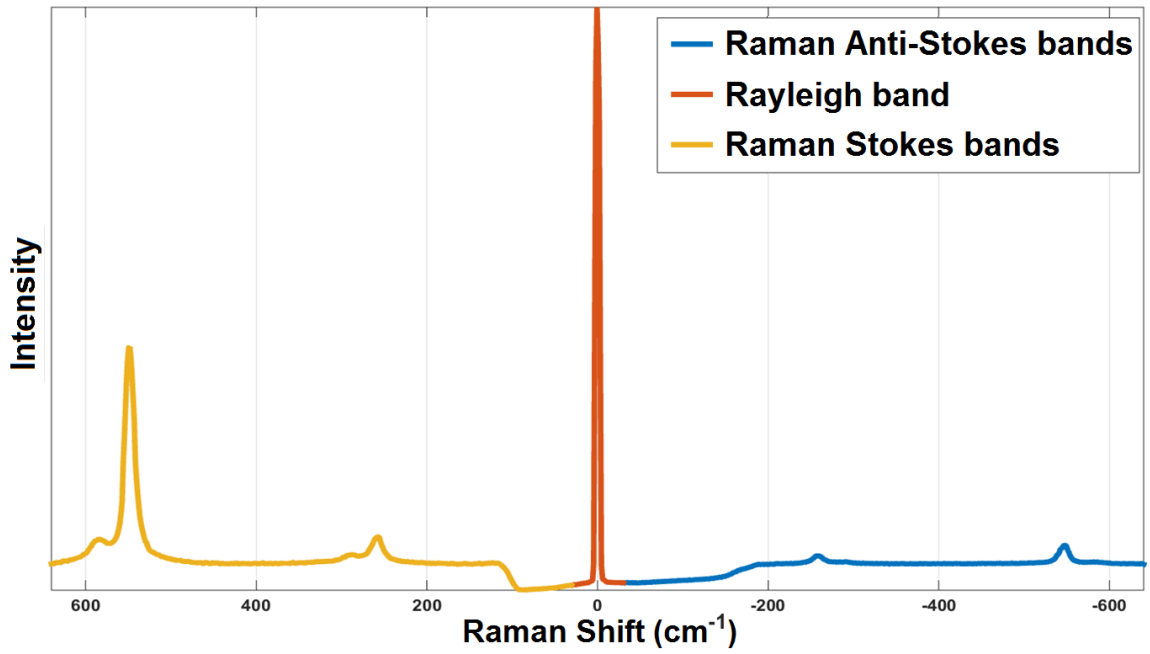


Figure 2.4: Raman spectrum of an ultramarine blue pigment showing Raman bands corresponding to the Raman Anti-Stokes (blue), Rayleigh (red) and Raman Stokes (yellow) dispersions

of high-quality reference spectra with useful historical, artistic and scientific information. The developing of a rigorous and documented database of pigments may be a top priority as it is a key element of the art materials identification[32].

The use of Raman spectroscopy as a technique of analysis in real-world applications is now a well documented reality both in fundamental theoretical aspects and in instrumentation and applications[33–35]. In particular, Raman spectroscopy has the invaluable ability to investigate precious art objects in a completely non-contact and non-destructive way. The identification of art materials is important in the research of art works and has been the subject of many research articles and monographic studies[36,37]. As a result, Raman spectroscopy is one of the techniques that best fits the analysis of art materials and pigmentation, being a natural application of this technique in this topic because of its non-destructive molecular examination allows the art materials identification. Thus, this way of working provides useful and objective information for the cataloguing, restoration, conservation of art works, helping in the preservation of the cultural heritage.

2.3 Data processing in Raman spectroscopy applied to art analysis

The knowledge of pigments which were in use on the palettes of the ancient artists is fundamental to preserve the works of art. The signature of a pigment obtained

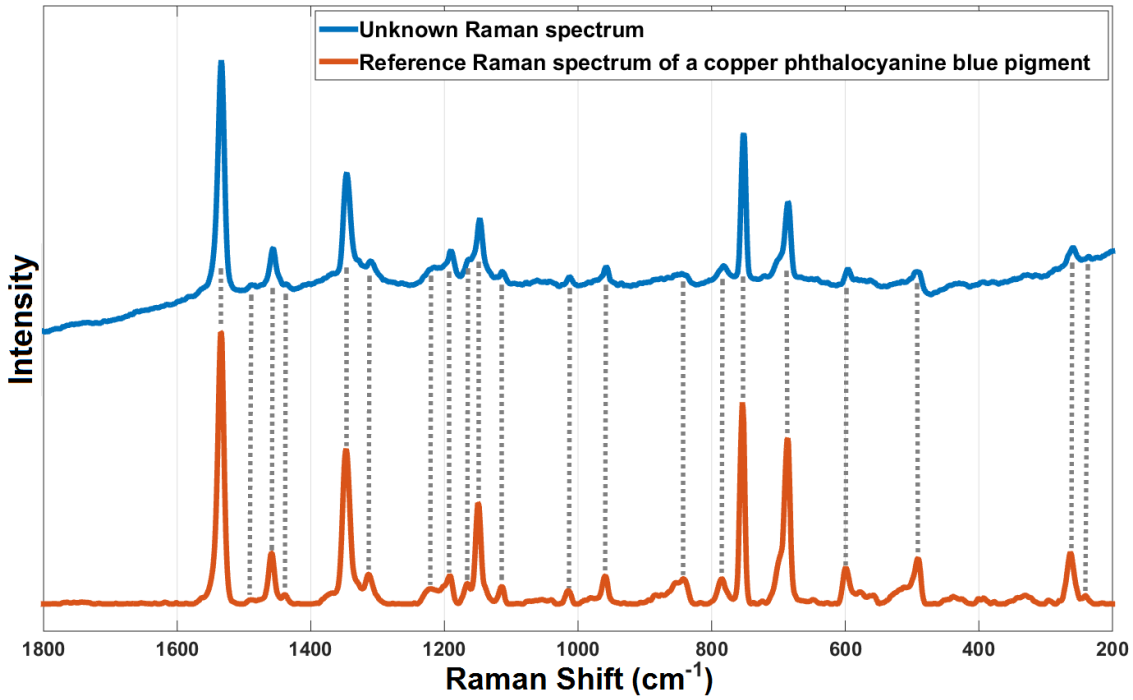


Figure 2.5: Identification of an experimental Raman spectrum (blue) through comparison with the reference Raman spectrum of a copper phthalocyanine blue pigment (red). Dashed lines highlight the main coincident Raman bands between these two Raman spectra

by Raman spectroscopy (see Fig. 2.6) is unique and allows the identification of the pigment through its molecular spectrum. Hence, Raman spectroscopy offers many advantages, providing qualitative and quantitative information regarding the molecular composition of the pigmentation of an art work without producing any damage to the analysed object.

As in many other fields of the knowledge, most of the recent developments of the Raman analysis are related to the use of the increasing power of the computers in the interpretation of the data. The development of new methods and data treatment of the spectral information is a field of continuous research in the field of cultural heritage. The study of modern pigments is often complex due to the presence of a large number of Raman bands, typical of the used organic molecules.

The identification of modern organic pigments by Raman spectroscopy is hampered by the large amount of different synthetic materials that exist. Therefore, an extended database of reference spectra is needed. Besides this spectral library, there is need for an accurate and fast-searching algorithm for selecting the reference spectrum that best corresponds with the unknown spectrum. Different algorithms have been developed based on different multivariate analysis approaches[38–42]: chemometrics, pattern recognition, supervised and unsupervised machine learning, automated detection, among many others.

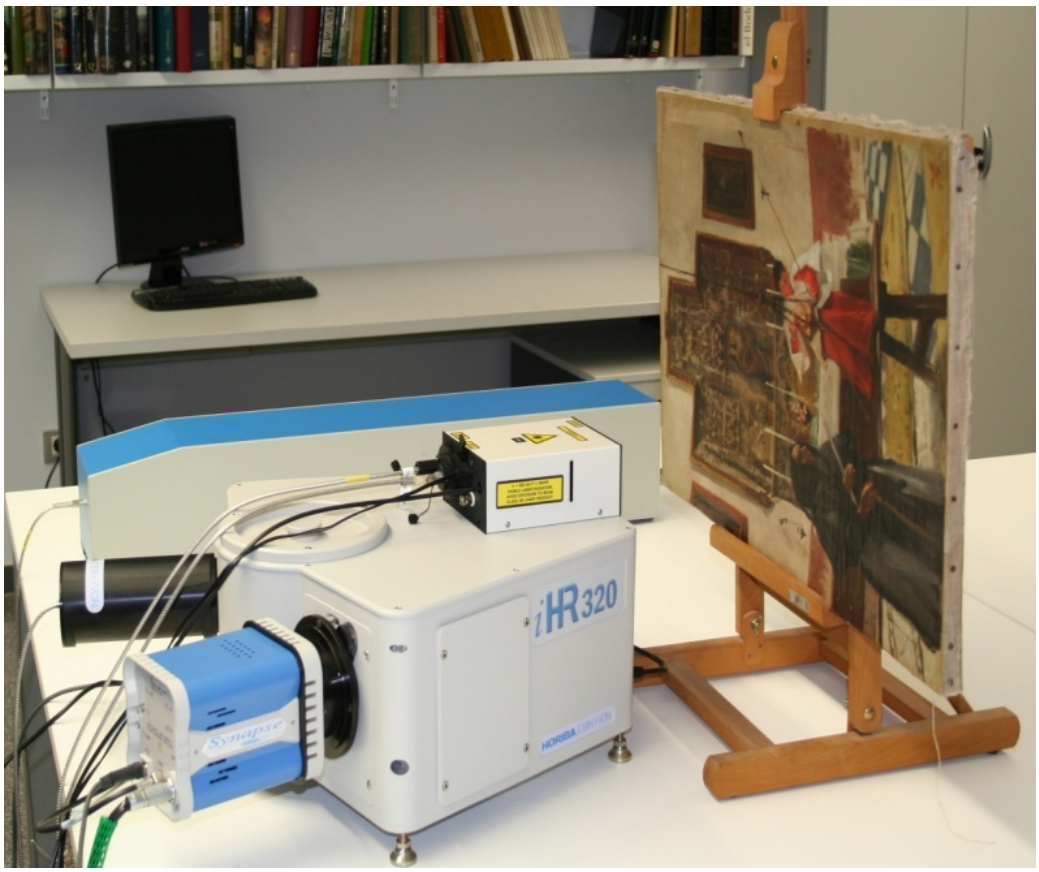


Figure 2.6: Measurement of a Raman spectrum from a work of art

2.3.1 Data enhancement

Pollutants and other environmental factors as well as interferences from the binding media and to ageing may have a direct impact on the quality of the Raman signal[43–46], which contributes to the difficulties in identifying pigments by Raman spectroscopy in the form of noise. An appropriate signal treatment expands the capabilities of the technique to non-invasively identify and quantify the chemical composition of paint layers in art works.

Generally, experimental Raman spectra are noisy, which complicates the analysis making it difficult for the analyst to locate spectral features of interest. Additionally, if the spectrum is to be interpreted by a computer, for example to search for the spectrum in a reference library, noise in the experimental spectrum may often lead to poor search results. Clearly, reducing the noise from a spectrum is of considerably practical interest. There are different noise reduction approaches[47–49]: by smoothing, based on fuzzy logic, or by merging continuous wavelet transform. In addition, the assignment of the vibrational modes is often compromised by the presence in the spectrum of an intense fluorescence background that covers the measured spectra. Several techniques have been employed to minimize the presence of this fluorescence in order to resolve and

analyse Raman spectra[50–52]: wavelength shifting, time gating, frequency-domain filtering, first- and second-order derivatives, simple curve fitting of the broadband variation with a high-order polynomial, and mathematical morphology.

2.3.2 Feature extraction

Numerous studies have been devoted to the topic of feature extraction. It is well-known that an appropriate chemometric preprocessing of Raman datasets is of great importance for further identification or classification procedures[53–56] and has to be considered carefully to accomplish good results. In this sense, dimensionality reduction techniques are usually applied as feature extraction in Raman spectroscopy with the aim of extracting spectral characteristics or markers of special interest for a proper data interpretation, therefore making dimensionality reduction to become an essential step. Hence, there is a need to reduce the dimension of the dataset without losing essential information as the computation time decreases in subsequent data processing steps (classification training, spectral library searching, ...). As reported in[57], a strategic approach based on dimensionality reduction may improve exponentially the accuracy of an automated identification or classification methodology - identification rates may be improved up to a 15%. Consequently, the extraction of spectral markers of interest through data reduction tools should be considered in every chemometric analysis in order to optimise the outcome of an identification or classification methodologies.

2.3.3 Automated analysis

The identification of paint materials by visually interpreting spectra can be a complex and time-consuming task, and may introduce subjectivity linked to the analyst's experience. This is why there is a strong move underway in the discipline towards computer-assisted analysis of art works. In this sense, different algorithms have been employed to aid in the recognition of spectra in an automatic and an objective way. These algorithms are based on multivariate analysis, using techniques such as principal components analysis, artificial neural network, linear discriminant analysis, support vector machine, case-based reasoning or fuzzy logic[58–62]. The goal of these algorithms is to find the reference spectrum of the database that matches the unknown spectrum. The main benefit of an automatic identification system base on a recognition algorithm lies in offering a useful supporting tool to the analyst in the decision-making process.

For spectral identification purposes, the process of comparison may be carried out in two different ways: comparing the bands or the spectrum as a whole. On one hand, if the first strategy is selected, it is needed to automatically localize the bands of the unknown spectrum[61,63]. Nevertheless, this localization may turn out to be a complex task, and once the bands are localized it is needed to identify which pigment

they correspond to. On the other hand, if the second strategy is selected, the band's localization is not needed since the comparison is developed by comparing point by point of the unknown spectrum and the patterns. In this case, the reference spectral library is composed by the whole spectral expression of the pattern pigments.

Nevertheless, real samples are always complex mixtures of original and degradations compounds that require new approaches to be implemented in the daily practice of Raman spectroscopy. In these complex environments, the ability to identify the constituent pigments within a mixture is extremely important. Raman spectroscopy is often used for this, and has been shown to be a rapid, non-contact detection method, which offers the ability for a single detector to identify a variety of substances.

Different automated techniques have been reported, mainly based on multivariate analysis, which has developed into a highly valuable instrument for the analysis of Raman spectra[58,64–66]. Some techniques require a set of several spectra measured from the same spot in order to be able to identify mixed pigments and demonstrated to be very effective when the number of mixed components is initially known[67]. Another methods are based on a Pearson correlation application to compare sets of multi-wavelength resonance-Raman or by using a chemometric technique such as Independent Components Analysis (ICA)[68–73] which recovers a set of independent signals from a set of measured signals. Furthermore, some commercial software packages[74,75] attend to address the automated identification of single- and multi-component spectra but they require user input at some stage of the identification process. Usually, these methods try to provide a list of single candidates and the user then needs to choose how many or which of those candidates are present in the mixture. Consequently, these techniques remain dependent on the user's experience.

Clearly, automated analysis based on the entire spectral range for multivariate analyses improve the pigments identification through Raman spectroscopy. In particular, recent studies[76–79] conclude that a full-spectrum matching algorithms exhibit excellent performance in identification and classification tasks. This class of algorithms supports both vector and trajectory input formats, exploiting all available spectral information. By combining these insights, optimal spectrum matching performance can be achieved using careful preprocessing i.e. data enhancement, and a vector similarity metric for the automated identification and classification of Raman spectra from pigments.

2.4 Chapter summary

In this chapter, the need of performing analysis of art works in order to preserve the cultural heritage was discussed, and, in particular, several analytical techniques aimed at performing this kind of analysis were reviewed. From these techniques, Raman spectroscopy stands out as it allows the unequivocal identification of the analysed

material in a non-destructive fashion requiring no sample preparation, ideal for the analysis of artistic pigments. Finally, the state of the art of data processing techniques in Raman spectroscopy applied to pigments analysis was reviewed, focusing on data enhancement methodologies, feature extraction techniques and automated analysis of Raman spectra from art works.

Chapter 3

Acquisition and enhancement of Raman spectra from pigments

3.1 Chapter overview

This chapter describes how the main experimental Raman spectra from pigments used in this research were acquired. Specifically, Sect. 3.2 provide details regarding the setup of the experimental measurement system. Additionally, the main noise sources present in Raman spectroscopy are introduced. In order to enhance the Raman information Sect. 3.3 presents a novel automated noise filtering methodology aimed at helping in the interpretation of the Raman spectra from pigments.

3.2 Data acquisition

Raman spectroscopy provides a huge flexibility and versatility, and it is perfectly adapted to the demands of the analysis of the pigmentation of art works. Nowadays, the latest advances in technology have enabled the development of equipments devoted to Raman spectroscopy to the point of allowing the portability of the instrumentation. A portable state-of-the-art Raman spectroscopy system is described below, which constitutes primarily the laboratory of the Raman spectroscopy research group of the UPC.

3.2.1 Measurement system: Experimental set-up

The experimental Raman spectra acquired for the purposes of this research were recorded using the portable Raman equipment iHR320 (HORIBA Jobin-Yvon) with a lens of 4.5x focus. The optical source employed for spectral acquisition was a He-Ne laser (632.8 nm) which provided approximately 17 mW. The light from the laser was guided by an optical fiber to the optical head and directed to the sample. The scattered light was collected and filtered by the corresponding edge filter built into the optical

head. It was then guided by the optical fiber to the monochromator and detected by a thermoelectrically cooled charge-coupled device (CCD). A schematic diagram of the measurement system is represented in Fig. 3.1. Acquisition times were around 30 seconds with 5 accumulations (150 seconds) on inorganic pigments and 300 seconds with 5 accumulations (1500 seconds) for the organic pigments in order to achieve the best trade-off between signal to noise ratio in the spectra from the sample and measurement time.

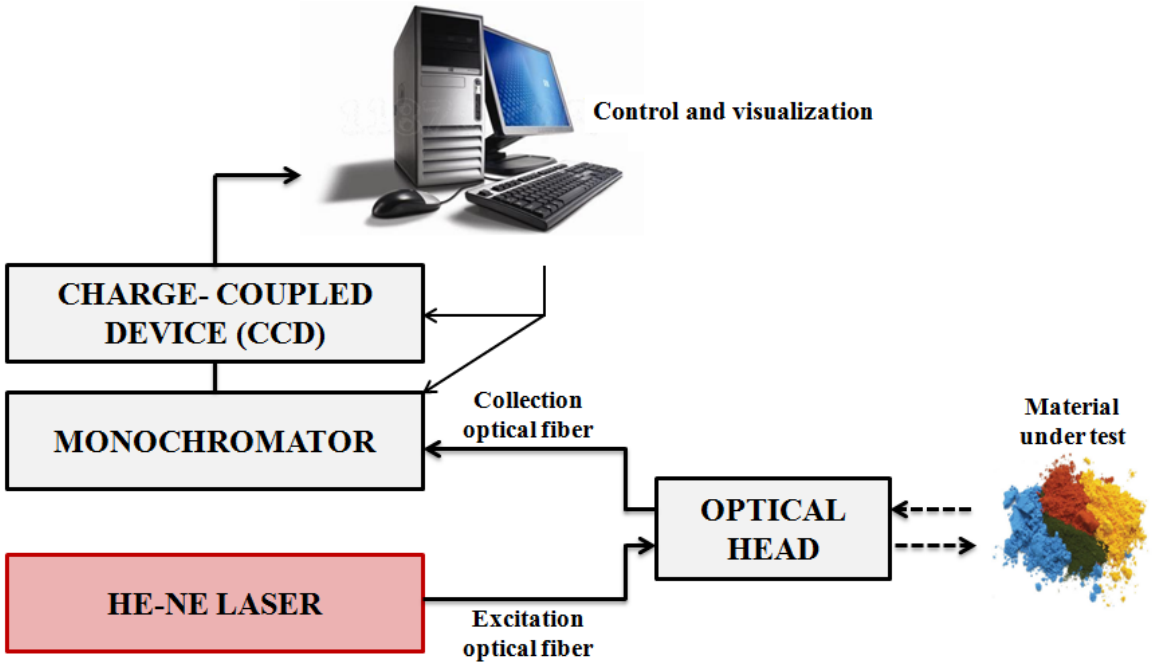


Figure 3.1: Schematic diagram of the measurement system based on the portable Raman equipment iHR320

3.2.2 Noise sources in Raman spectroscopy

A Raman spectrum may provide valuable information about the analysed sample but the quality of this information may be compromised due to the presence of interfering or unwanted signals. Broadly speaking a Raman spectrum can be divided into two parts: the useful signal and the noise[31]. In our case, the useful signal is the Raman information, which can be seen as a fingerprint signal in the form of a specific combination of peaks -the Raman bands- by which the analysed material can be unequivocally identified. In contrast, the noise is the part of the Raman spectrum that comes from undesired sources, which thus can adversely affect the interpretation of the analysed sample.

The most commonly found sources of noise in Raman spectroscopy are shot noise and fluorescence's baseline (see Fig. 3.2): the shot noise is an unavoidable noise source caused by the statistical nature of light, which may compromise the analysis of a

Raman spectrum; the fluorescence's baseline is sample-inherent usually of higher amplitude than the Raman information that can thus mask the Raman bands. Therefore, the noise impact should be reduced -i.e. filtered- before performing further analyses (whether through visual inspection or automated methodologies) in order to accomplish a proper interpretation of a Raman spectrum.

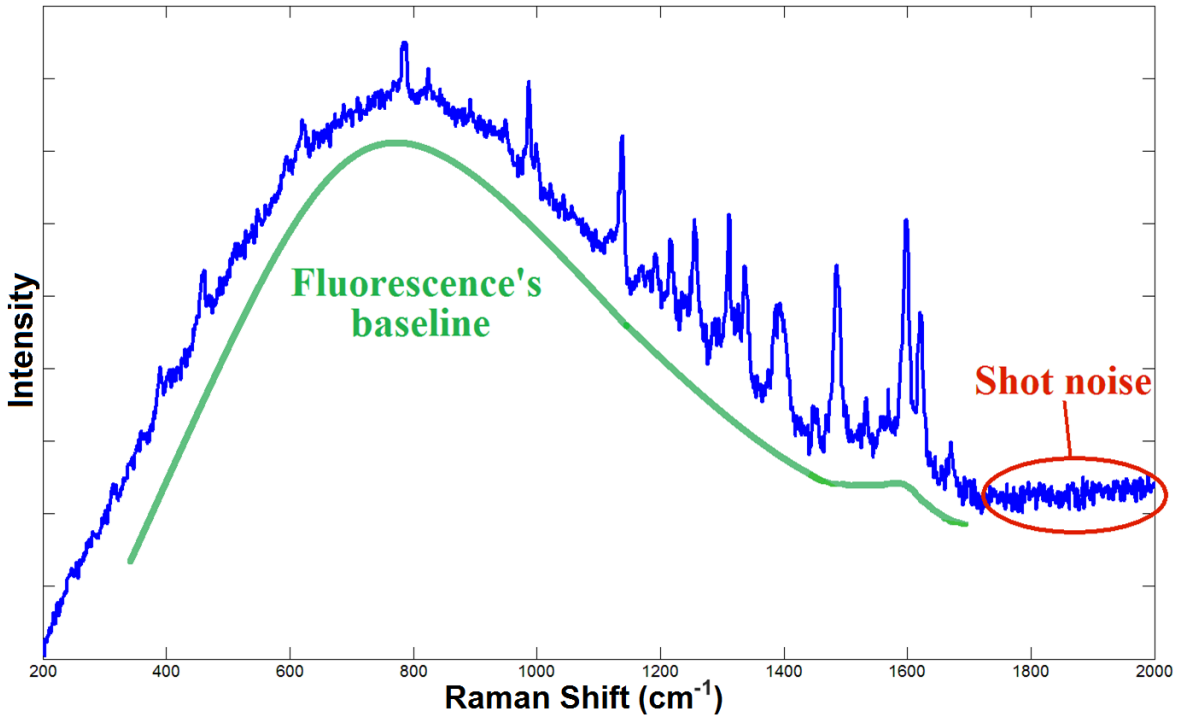


Figure 3.2: Noise sources most commonly found in Raman spectroscopy

3.3 Data enhancement: Automated noise filtering methodology

There is no single strategy for noise filtering in Raman spectroscopy. Several methods have been proposed to enhance the Raman information[80–96]. The most frequently used methods are software procedures, which do not require to upgrade the existing instrumentation. Such procedures are generally dedicated to filter one kind of noises separately, i.e. or shot noise or fluorescence's baseline. For instance, to reduce the shot noise the simplest procedure is the median filter, whilst to remove the fluorescence's baseline the simplest and widest used method is the polynomial fitting. The basic version of such methods involves user intervention in order to select appropriate key parameters, and this selection process is usually time consuming. For instance, choosing which Raman shifts belong to noise sources in non-Raman characteristic band regions or which ones belong to Raman characteristic band regions is a critical point, which may introduce subjectivity depending on the analysts' experience. Thus, several

methods have been developed in the last decade in order to avoid any user intervention. Generally, such methods are based on iterative solutions. Though these methods may provide successful results, they treat one kind of noise at a time and due to the high nonlinearity and complexity of a Raman spectrum they may not well smooth it or reject its fluorescence's baseline. As an alternative, a fully-automated noise filtering approach was developed in this research which pursues a twofold objective: the shot noise reduction and the fluorescence's baseline removal.

In this respect, a new and simple procedure was designed and implemented to reduce the shot noise and to remove the fluorescence's baseline simultaneously with a single strategy, which is independent of the Raman spectrum to be filtered. The underlying principle of this novel approach is based on the different "shapes" shown by the shot noise and the fluorescence's baseline in a Raman spectrum: the shot noise may be seen as an intensity fluctuation (rapid variation), whilst the shape of the fluorescence background is shown as a soft drifting baseline (slow variation). In this regard, the method uses mathematical morphology operations, which simplify and preserve the main features of the shapes, jointly with cubic penalized spline fitting for smoothing and baseline-removal of Raman spectra in a unified way. No parameter tweaking is needed and therefore no user intervention is required. The method was developed as an application-specific algorithm which improves the signal-to-noise ratio tackling at the same time both shot noise reduction and baseline rejection, preserving the shapes, positions and intensity ratios of the Raman bands.

The methodology presented here describes the core principles of the proposed approach for noise filtering. Finally, the results are discussed and evaluated for real-case examples.

3.3.1 Noise filtering methodology

A noise filter methodology is proposed which broadly speaking is based on a curve fitting technique intended to obtain an improved SNR, making the Raman spectra easier to interpret.

The use of piecewise polynomials to model regression functions and perform curve fitting has a long history[97–103]. In smoothing, the location of the points, or knots, in which the polynomial pieces are joined are arbitrary, which permits a very large class of possible fits. A widely used fit is based on splines[104–107], which are piecewise-defined by polynomial functions. Penalized splines (or p-splines)[108] are a very popular spline fitting approach, which has the following properties: efficient computation, flexibility, and ease of setup[109]. P-splines are regression splines fit by least-squares with a

roughness penalty which avoids overfitting[105]. According to[110], the optimal degree of this piecewise polynomial regression is 3, which generates the so-called cubic p-splines. The smoothness of the estimate varies as a function of the smoothing parameter, λ: the larger the smoothing parameter, the more the fit minimizes towards a polynomial fit, which in turn allows the estimate to deal with data gaps[108]. In our research, the λ-parameter was selected to be small enough so as to keep the estimates smooth and its value was fixed to 0.7. This constant value provides a good compromise between smoothness and polynomial fit regardless of the input spectra, whether simulated or experimental.

The choice of knots has been a subject of much research[111,112]. Equidistant knots can be used, but this allows only limited control over the fit. Instead, a smart knots selection is preferred so in the case noise filtering of Raman spectra the presence of noise is optimally reduced whilst the shape and positions of the Raman bands remain unaltered. This may be achieved through a strategic selection of knots according to the shape of the input data. To do so, the usage of mathematical morphology operations is proposed.

Mathematical morphology is a nonlinear technique based on classical set theory[113,114]. It finds application in many different research fields as it only involves the definition of sets of data taking advantage of the properties of those sets[115,116]. In particular, it is predominantly useful in fields in which the shape is the most important feature. Morphological operations transform the original function into another function looking for geometric structures (i.e. shapes) using the structuring element whose shape is chosen according to the *morphology* of the function and the special structures to be extracted. Choosing a suitable structuring element, we can use the information extracted from morphological operations to generate the knots sequences to be used in the cubic p-splines fitting to denoise a Raman spectrum. There are two basics operations in mathematical morphology, called *erosion* and *dilation* and the combination of such operations provides two more operators, namely closing and opening. The morphological closing of a function f by a structuring element Y, $\phi_Y(f)$, is described mathematically as

$$\phi_Y(f) = \epsilon_Y[\delta_Y(f)] \tag{3.1}$$

being $\epsilon_Y(f)(x) = \min_{s \in Y} f(x+s)$ the erosion and $\delta_Y(f)(x) = \max_{s \in Y} f(x+s)$ the dilation of the function. The closing smooths the function nonlinearly, removing holes and connecting nearby items thus taking always values that are higher or equal than those of the input function. Hence, the closing by a short structuring element may provide a rough estimation the shape of the Raman bands. In this case, this short structuring element, Y_m, is defined so that the closing can take into account any Raman band, and therefore fixed to three data points. The resulting closing is modified to further reduce

the shot noise influence as $\phi'_{Y_m}(f) = \phi_Y(f) \notin \epsilon_Y(f)$.

On the other hand, the morphological opening of a function f by a structuring element Y,

$$\gamma_Y(f) = \delta_Y[\epsilon_Y(f)] \tag{3.2}$$

smooths the input function too but differently since it removes the positive peaks, taking always values that are lower or equal than those of the input function.

Hence, the opening by the optimal structuring element, Y_{opt}, may provide a rough estimation of the fluorescence's baseline. The main morphological operations are graphically represented in Fig. 3.3.

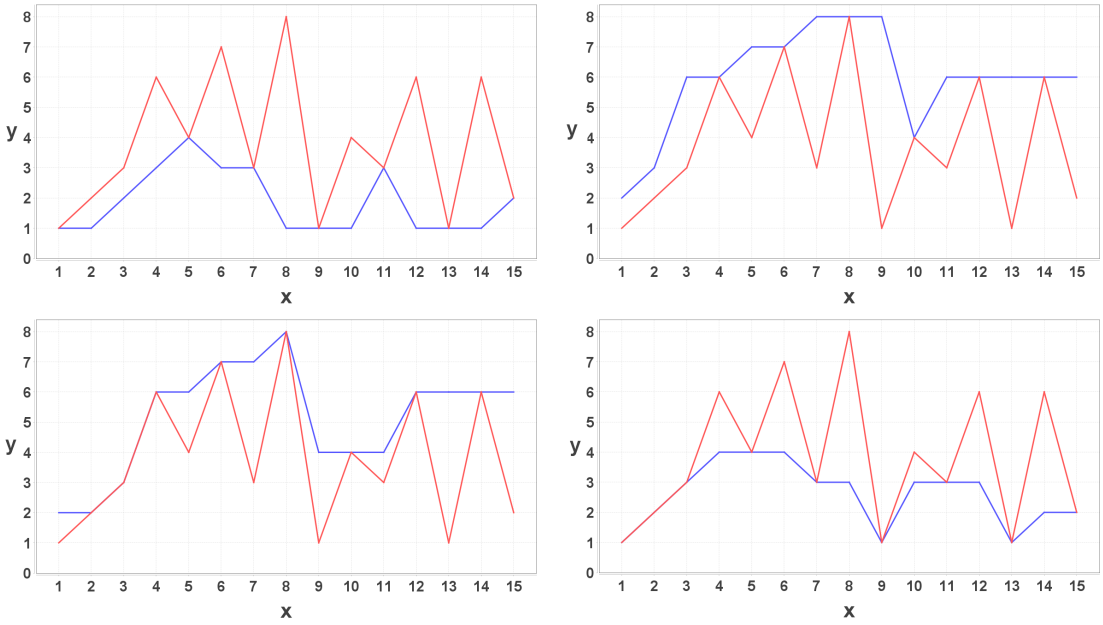


Figure 3.3: Representation of the main mathematical morphology operations (blue) on an input sequence (red) with a structuring element of 3 data points: erosion (top left), dilation (top right), closing (bottom left) and opening (bottom right)

The optimal structuring element is selected following a lookup procedure[51]:

- As starting point, the opening of the input spectra by the minimum structuring element is computed
- Iteratively, the opening by an incremented length of the structuring element is computed for each iteration
- The Root Mean Squared Error (RMSE) between the resulting opening and the opening of the previous iteration is computed
- The optimal structuring element is obtained when the RMSE gets stabilized, i.e. the same opening is obtained in 3 consecutive iterations

- The resulting opening is further modified in order to reduce any flaw in the peak regions as

$$\gamma'_{Y_{opt}}(f) = min(\gamma_{Y_{opt}}(f), \frac{\epsilon_{Y_{opt}}(f) + \delta_{Y_{opt}}(f)}{2}) \tag{3.3}$$

The methodological scheme of the noise filtering follows the flowchart shown in Fig. 3.4. Being f a noisy Raman spectrum, a knots sequence, K1, is obtained from the intersection of f with the modified closing by the minimum structuring element, $\phi'_{Y_{min}}(f)$. A cubic p-spline fit of f through K1 is performed which provides an intermediate function, g. Then, the optimal structuring element that follows the morphology of the baseline in g is achieved. Next, a new knots sequence, K2, is obtained from the intersection of g with the modified opening by the optimal structuring element, $\gamma'_{Y_{opt}}(g)$. A cubic p-spline fit of g through K2 provides an estimation of the fluorescence's baseline, h. Finally, the denoised spectrum, d, is obtained by subtracting the fluorescence's baseline estimation from the intermediate function. The morphology-based cubic p-spline fitting methodology for enhancing Raman spectra is graphically represented in Fig. 3.5, where it was applied to an experimental Raman spectrum from a sample of a PY1 pigment powder.

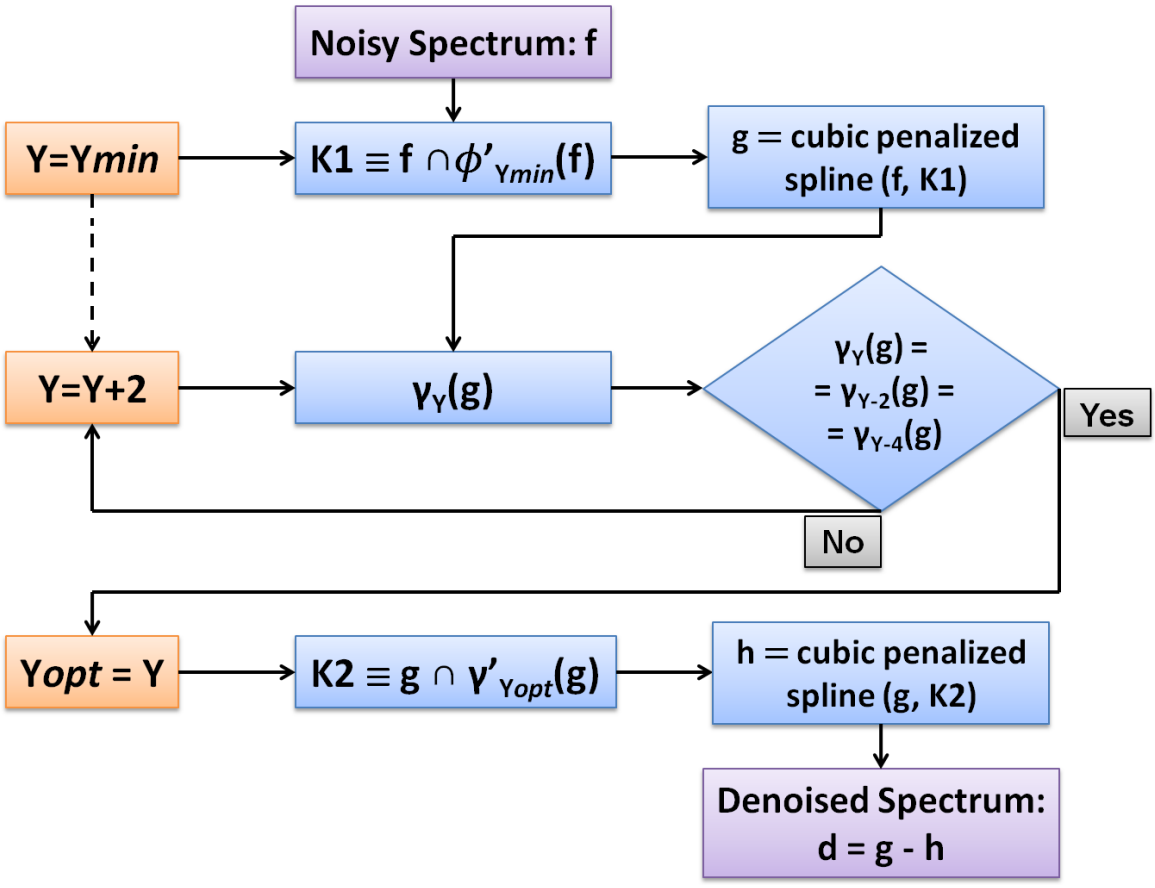


Figure 3.4: Noise filtering scheme, aimed to reduce the shot noise and to remove the fluorescence's baseline

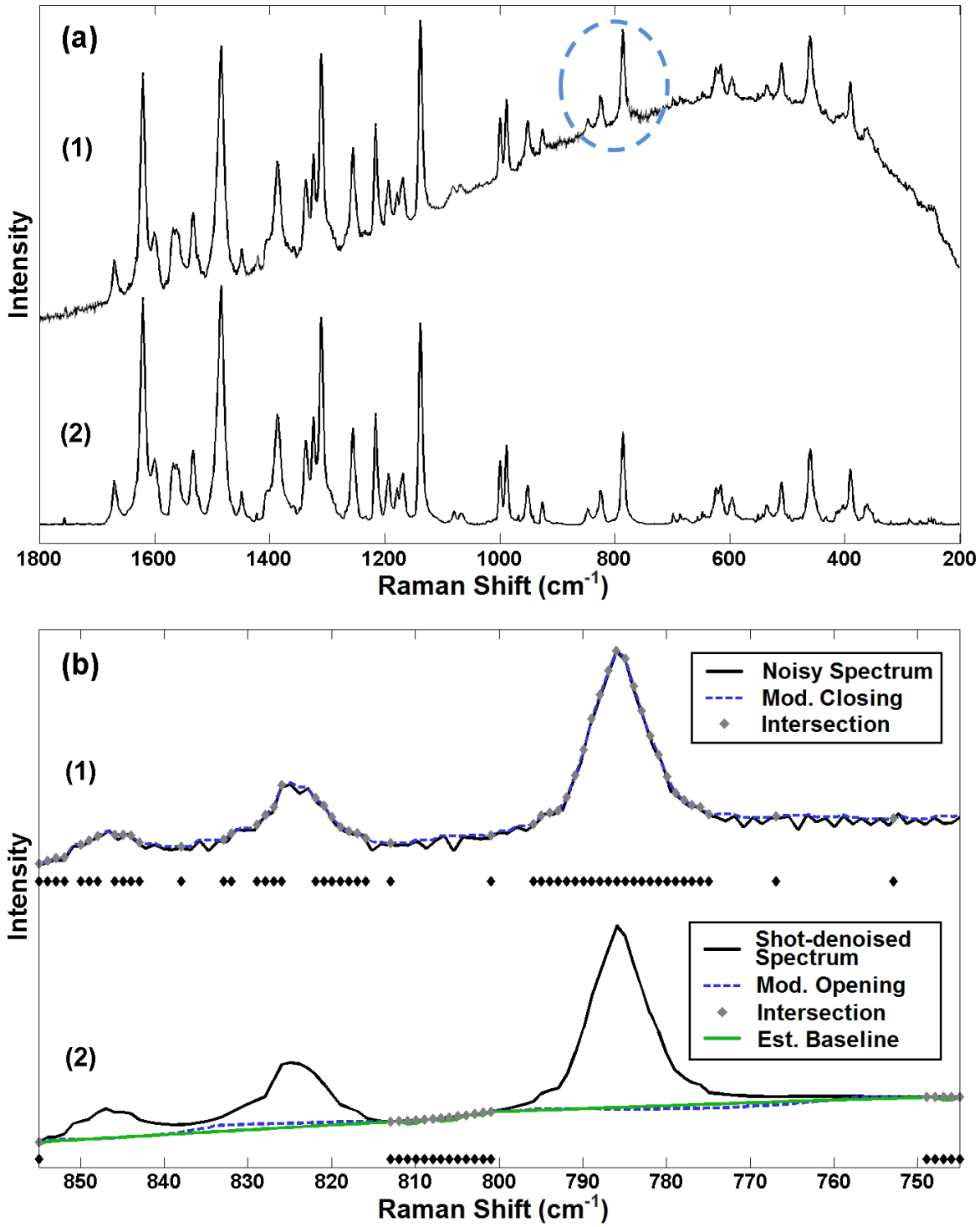


Figure 3.5: a) Graphical example of the developed noise filtering method applied to a measured Raman spectrum of sample of a PY1 pigment powder, b) zoom for Raman shifts from $740cm^{-1}$ to $860cm^{-1}$: 1) shot noise reduction by fitting a cubic penalized spline through the modified closing by the minimum structuring element, 2) baseline removal by fitting a cubic penalized spline through the modified opening by the optimal structuring element. The knot sequences are represented as black diamonds for both cases

Appendix A provides a performance analysis of the developed data enhancement methodology. Specifically, a benchmark study using simulated Raman spectra was presented providing a performance evaluation and comparison of the noise filtering algorithm developed in this research and conventional denoising algorithms in common use. The results show that the presented denoising approach outperformed all other algorithms that were tested in both shot noise and baseline tests. Additionally, several tests were performed using experimental Raman spectra providing reliable and suitable results as well.

3.4 Chapter summary

In this chapter, an overview of the data acquisition through Raman spectroscopy was presented, focusing on the experimental measurement system of the UPC laboratory and its main application: the analysis of artistic pigments. In this kind of analysis, external agents such as pollutants or binding media among others may increase the noise impact, thus degrading the quality of the Raman measurements. Consequently, a fully-automated denoising methodology was developed, which enhances the Raman information helping in the interpretation of Raman spectra.

The developed noise filtering approach uses the same novel and simple scheme for both shot noise reduction and fluorescence's baseline rejection. Specifically, the approach is based on p-spline fitting, a piecewise polynomial curve fitting technique generally used for data smoothing. A key point of the developed denoising methodology is how to obtain the location of the points (the so-called knots) in which the polynomial pieces are joined. A strategic selection of knots according to the shape of the input Raman spectra was discussed. Concretely, the usage of mathematical morphology operations for knots selection was presented. Hence, mathematical morphology operations are able to retrieve the morphology of the Raman information, which preserves the shapes, positions and intensity ratios of the Raman bands.

The usage of mathematical morphology together with p-spline fitting demonstrated to be a consistent combination in the application of data enhancement in Raman spectroscopy applied to pigments analysis. Several tests were performed providing successful results despite of requiring no user intervention: The method reduces the interferences coming from noise sources whilst enhancing the Raman information in an automatic fashion. Consequently, the developed noise filtering methodology has great potential as an accurate fully-automated practical method to help in the interpretation of Raman spectra, not only for artistic pigment analysis, but essentially for any material group as well.

Chapter 4

Automated analysis of Raman spectra from pigments

4.1 Chapter overview

This chapter describes the methodologies developed for the automated interpretation of Raman spectra from pigments. This answers the increasing motivation to automate the processes involved in the identification and classification of Raman spectra in paint instances against the manual analysis of Raman spectra which can be a subjective and time-consuming task. The developed methodologies presented in 4.3 and 4.4 were used to identify and classify Raman spectra from pigments which are commonly present in artist's paints in experimental environments, providing reliable and consistent results. Therefore, they can play a good auxiliary role in the analysts' endpoint identification and classification automated systems.

4.2 Spectral pre-processing and comparison

As explained in Sect. 2.2.2, the signature of a pigment obtained by Raman spectroscopy is unique and allows the identification and classification of the analysed pigment through its molecular spectrum. The identification and classification processes are generally carried out by spectral comparison between an unknown Raman spectrum with an appropriate set of reference Raman spectra previously stored in a reference database[117–121]. When applying an automated comparison-based methodology for recognising or classifying Raman spectra from pigments, it is necessary to rely on mathematical tools such as distance metrics, which allow the spectral comparison in an automated fashion. In this respect, it is crucial to make some spectral pre-treatment to properly address this comparison.

Spectral pre-processing

A uniform data format smooths the progress of comparisons between unknowns and patterns. Furthermore, as noise is inherent to the acquisition of a Raman spectrum, a denoising should be performed to enhance the Raman information as much as possible even assuming that it was collected under optimal conditions. Consequently, a three-step pre-processing sequence must be followed in order to ensure the success of the automated processes of identification and classification. Shot noise reduction and baseline correction is the first pre-processing step. In this regard, the methodology for shot noise reduction and baseline rejection described in 3.3 was used. The second pre-processing step is interpolation, so that all spectra are stored in a compatible way. The interpolation ensures that all spectra have a common set of Raman shifts, which is crucial when spectra collected with different measurement systems are used. Finally, the last pre-processing step is intensities normalization, which reduces the impact of measurement conditions so that the outcome of the data processing is independent of the acquisition instrument.

The intensities normalization used in this research was the *min-max* normalization, where the minimum intensity is scaled to 0 and the maximum to 1, meaning that a normalized spectrum x' maintains the relative ratio between its Raman bands from the input spectrum x by applying the following equation:

$$x'[i] = \frac{x[i] - min(x)}{max(x) - min(x)} \quad \forall i \tag{4.1}$$

where $x[i]$ is the Raman intensity at the Raman shift i.

These pre-processing steps ensure that all the spectra are baseline-corrected and fulfil a set of homogeneity conditions with respect to data format. That is, all spectra cover the same spectral range, are baseline subtracted, and their intensities are normalized. As a result, the spectral comparisons involved in the identification and classification of unknown Raman spectra may be properly carried out.

Spectral comparison

In order to perform the spectral comparison between an unknown Raman spectrum and the reference Raman spectra several distance metrics were used. An overview of these metrics is detailed hereafter.

Euclidean distance The Euclidean Distance (ED) quantifies the degree of similarity between spectra and is defined as:

$$ED(u,v) = \sqrt{\sum_{i=1}^{K}(u(i)-v(i))^2} \tag{4.2}$$

being the Euclidean distance between two vectors u and v. The lower the distance, the more similar the spectra.

Mahalanobis distance The Mahalanobis Distance (MD) provides a measure of similarity between a vector (v) and a given distribution (D_i):

$$MD(v,D_i) = \sqrt{(v-\mu_i)'\Sigma_i^{-1}(v-\mu_i)} \tag{4.3}$$

where μ_i is the arithmetic center (i.e. the centroid) of the i-th distribution.

Bhattarcharyya distance The Bhattarcharyya Distance (BD) measures how similar two distributions are and is defined as:

$$BD(i,j) = \frac{(\mu_i-\mu_j)\Sigma(\mu_i-\mu_j)'}{8} + \frac{ln|\Sigma|}{2\sqrt{|\Sigma_i||\Sigma_j|}} \tag{4.4}$$

where μ_k and Σ_k are, respectively, the centroid and the dispersion matrix (the autocovariance matrix) of a k-th distribution and $\Sigma = \frac{\Sigma_i+\Sigma_j}{2}$.

Jeffries-Matusita distance The Jeffries-Matusita Distance (JMD)[122,123] indicates how separated two distributions are and is defined as:

$$JMD(i,j) = 2(1-e^{BD(i,j)}) \tag{4.5}$$

Specifically, JMD ranges from 0 to 2.

4.3 Automated pigment recognition through Raman spectroscopy

Traditionally, the recognition of pigments through Raman spectroscopy has been carried out through visual comparison between the Raman spectra measured on art works with an appropriate set of reference Raman spectra (also known as *patterns*). Nevertheless, this simple identification approach may turn out to be a difficult and a tedious task in instances showing a large number of Raman bands located close together as

in some kinds of organic pigments, and therefore it relies heavily on the experience of the analyst. The pigment identification may be further complicated when dealing with spectra from pigment mixtures, attending to the fact that usually the spectrum of an unknown mixture is not available in a library of the reference spectra. Giving this issues, a robust automatic identification strategy is clearly of practical interest for determining the pigments used in paint from a spectrum that shows their spectroscopic fingerprint.

4.3.1 Reference database compilation

The reference spectral library plays a key role in the automated identification process up to the point that an unknown spectrum cannot be identified if the corresponding reference spectrum is not in the library. To optimize this situation, it is recommended that the set of reference spectra -also called patterns- should be suitable regarding the particular application and the specific analysis to be performed by the user. In fact, different strategies may be considered to compile a library when analysing the pigmentation of a work of art. For instance, if the art work is suspected to be created by certain artist the library may include pigments used by that artist, or if the art work is supposed to fit in a given artistic movement it may include pigments used in that period (see Fig. 4.1, which includes the usage periods of the main historical pigments used in this research).

These strategies may be carried out under artistic and historical documentation, main reason for the developing of a high-quality documented database of spectra from art materials.

In this research, the reference spectral library was composed by a total of 288 Raman spectra. In particular, fifty-one spectra were measured from inorganic pigments that have been used in paints for centuries. These spectra were obtained by measuring directly pure pigments, one high-quality spectrum (in terms of Raman effect vs. fluorescence and signal to noise ratio) for each pigment. The pigment powders were supplied by different manufacturers (Sennelier, Kremer and Mongay). The rest were taken from the database of synthetic organic pigments used in modern and contemporary paintings published in[124]. Inorganic pigments are here designed by their historic name and organic pigments following the *Colour Index* reference database[a]. Further details on the reference database contents are described in Appendix B.

[a]The *Colour Index* name is established and published by the *American Association of Textile Chemists and Colorists* and *The Society of Dyers and Colourists*. The colour index name is a generic category and does not refer to a specific pigment. While it enables the artist to form a general idea of opacity, transparency and lightfastness, for a pigment in a certain colour space, it does not provide definitive information. Many grades of pigment are available from a number of manufactures with a very wide range of physical attributes

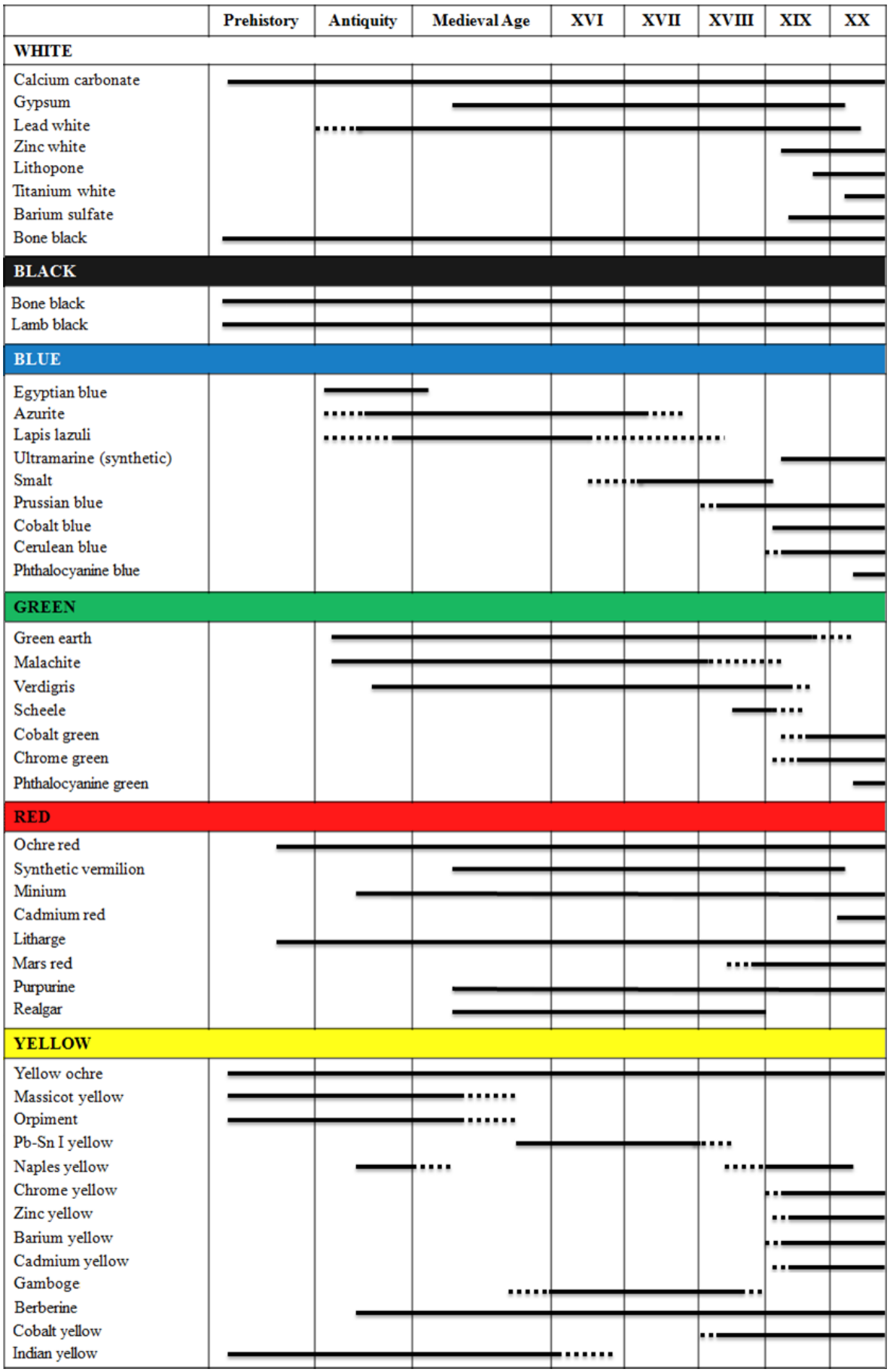


Figure 4.1: Usage periods of the main inorganic pigments used in this research. Solid lines: periods exactly known; discontinuous lines: periods of appearance and disappearance

4.3.2 Single-component identification

From a mathematical point of view, a given spectrum can be seen as vector of between 1000 and 2000 components typically. As a result, programmatically, the identification may become a time-consuming process. Hence, with the aim of speeding up the processing time and saving computing resources the usage of a data reduction tool was proposed as reported in 2.3.2. For the purpose of reducing the dimensionality of Raman datasets most often four standard techniques are used[125], namely Principal Component Analysis (PCA)[126], ICA[69], Linear Discriminant Analysis (LDA)[127] and Partial Least Squares (PLS)[128]. The performance of these techniques were tested on simulated datasets. The results of this comparative analysis are outlined in Sect. B.1 of Appendix B. Specifically, the best performing dimensionality reduction with no data loss providing the lowest processing time was obtain when applying PCA.

PCA is a multivariate method of the chemometric family, which is a chemical discipline that is based on mathematical and statistical methods to design and to select measure procedures and optimal experiments as well as to obtain the maximum information of the analysis[129,130]. It is extended as a data reduction tool mainly. Given a set of data, this technique provides a new space, known as the PC space, of same dimension as the original space, but its dimension can be reduced once some dimensionality reduction criterion is applied. In this reduced space, the original data are represented in such a way that the information is highlighted[39,40]. The main motivation to apply a dimensional reduction tool is to obtain a given set of data defined by N variables as a set of lower dimension, K, with $K << N$, but with equivalent information content. In the case of Raman spectra of artistic pigments processing, the initial considered set consists on P Raman spectra of the reference pigments (with $P << N$).

For this research, the dimensionality reduction criterion of considering 100% of variance of the P reference spectra in the original space was used, which provides a reduced space of $K = P - 1$ dimensions. In this way, a new space is generated through the PCA where spectra can be seen as points, and the measured unknown spectra can be compared and identified after projecting them onto this PCs space. Fig. 4.2 shows a two-dimension projection of the PCA transformation applied to the reference Raman spectra together with the corresponding biplot, which -in the case of Raman spectra- is an exploratory graph on both wavelength and scores of the PCA transformation matrix are displayed as a two-variable scatterplot. Fig. 4.3 shows the reference Raman spectra projected onto the PCs space (the so-called *scores*) as a function of PC together with the cumulative variance of the PCA projection also as a function of PC.

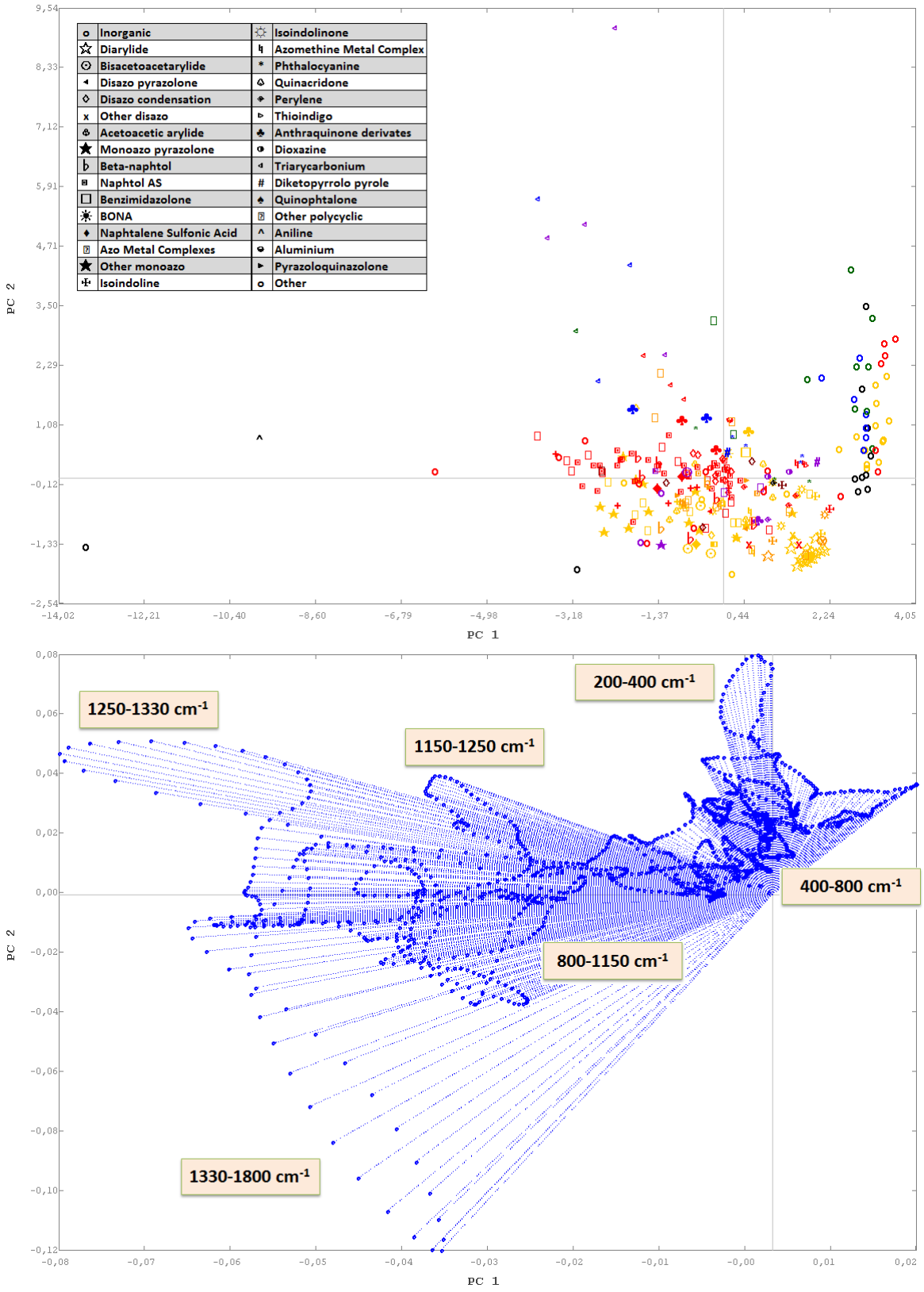


Figure 4.2: PC1-PC2 projection (top) and biplot (bottom) of the reference Raman spectra - item styles stand for chemical classes (see Sect. B.2 of Appendix B), item colour by *Colour Index*

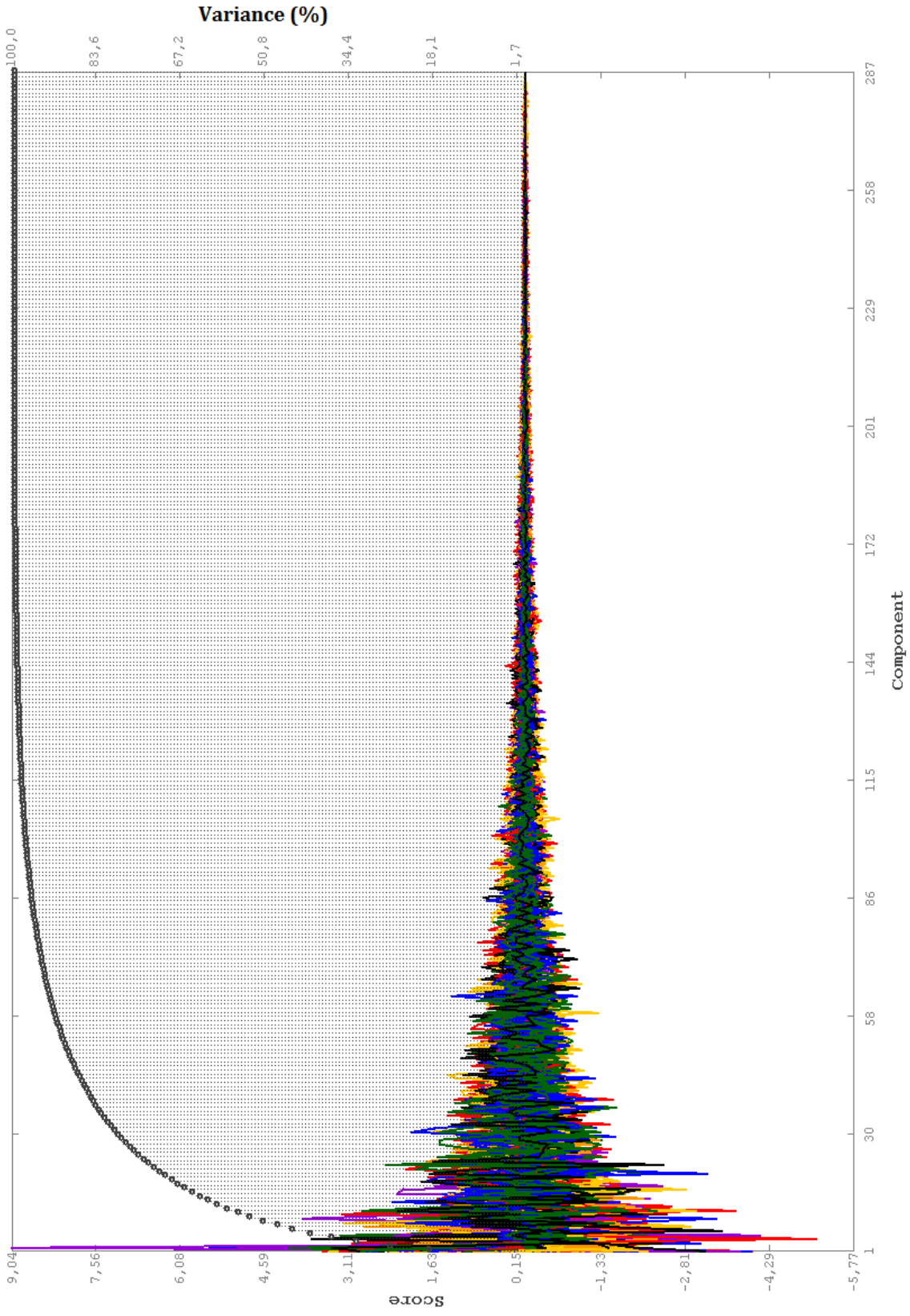


Figure 4.3: Scores of the reference Raman spectra and cumulative variance of PCA projection as a function of PC. Each colour represents a different reference Raman spectrum

Let x'_{unk} be an unknown spectrum after being preprocessed. The proposed identification methodology works by projecting its standard expression onto the PCs space, that is:

$$s_{unk} = \frac{x'_{unk} - m_{lib}}{\sigma_{lib}} C \tag{4.6}$$

where C is the transformation matrix obtained when applying the PCA to the reference spectra, and m_{lib} and σ_{lib} are the mean and the standard deviation values of each wavelength of the reference spectral library, respectively. The proposed identification methodology is developed based on the PCs space, and the Identification Block (IB) shown in Fig. 4.4 is based on mathematical operators and some identification criteria. The definition of the IB is described hereafter.

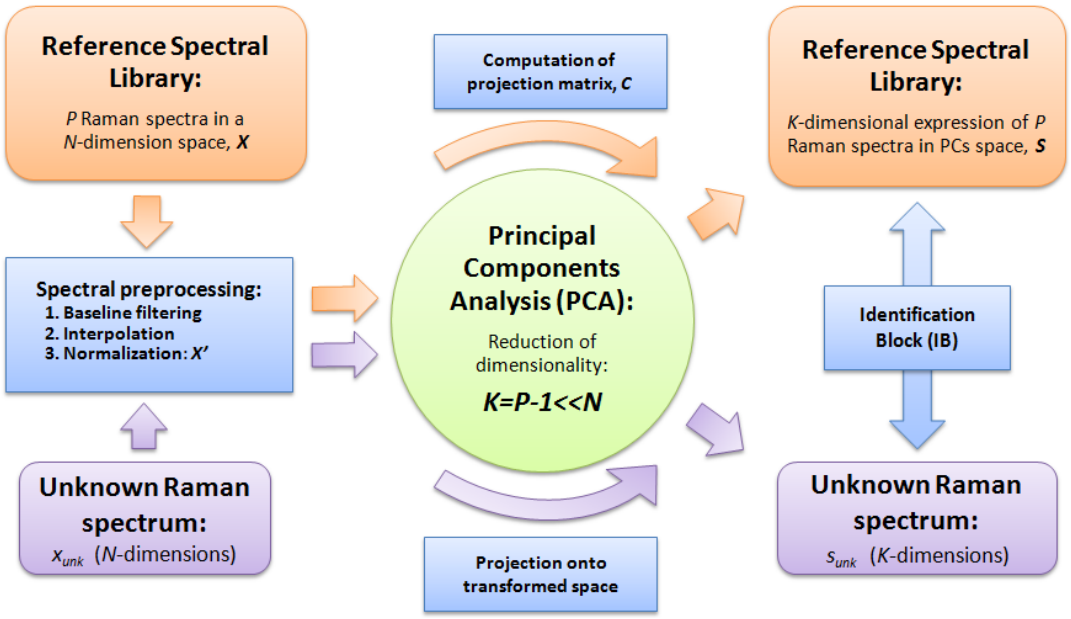


Figure 4.4: Spectra preprocessing and data reduction process: The expression of all spectra in a homogeneous and reduced format facilitates the comparison between the reference spectra and an unknown spectrum

Identification Block definition

The first stage of the identification methodology developed in this research is the Identification Block (IB). In order to extract candidate spectra (from the reference spectral library) to identify an unknown spectrum, the ED between the unknown spectrum and the patterns were computed. Besides, an additional metric was used, the so-called Squared Cosine (SC), which quantifies the quality of the representation of a spectrum projected onto the PCs space generated by the reference spectral library. It is defined as ratio of the norm of its expression in the PCs space, s_{unk}, to the norm of the standard

expression of the spectrum, x'_{unk}, that is,

$$SC(unk) = \frac{\|s_{unk}\|^2}{\|\frac{x'_{unk} - m_{lib}}{\sigma_{lib}}\|^2} \tag{4.7}$$

As a cosine, it fulfils that $0 \leq SC(unk) \leq 1$, and $SC(unk) = 0$ shows a low quality of the spectral representation in the PCs space, while $SC(unk) = 1$ indicates an optimal representation.

Two parameters regarding to the reference spectral library were defined:

- min_lib: The minimum distance of the reference spectral library
- min_k: The minimum distance between the k-th pattern and the rest spectra of the reference spectral library

These parameters are used to define the identification criteria. Specifically, the criteria to determine the candidate spectra that may identify the unknown spectrum are the following:

1. If the distance between the unknown spectrum and the k-th pattern is lower than the minimum distance of the reference spectral library, then the k-th pattern is candidate to be the pigment corresponding to the analysed sample:

   If $ED(unk, k - thpattern) < min\_lib$ then the k-th pattern is candidate

   The underlying idea is that two spectra are different if the distance between them is higher than min_lib. Therefore, this first criterion provides the candidate spectra whose distance to the unknown spectrum is lower than min_lib. Nevertheless, this criterion may turn out to be restrictive (for instance, when some spectra of the reference spectral library are very similar, which may imply a low min_lib), and whether it is not accomplished a new criterion is defined:

2. If the squared cosine of the unknown spectrum is higher than the ratio given by the minimum distance of the library and the minimum distance between the k-th pattern and the rest of the library, and if the distance between the unknown spectrum and the k-th pattern is lower than the minimum distance between the k-th pattern and the rest of the library, the k-th pattern is candidate:

   If $SC(unk) > \frac{min\_lib}{min_k}$ and if $ED(unk, k - thpattern) < min_k$ then the k-th pattern is candidate

Bearing in mind the underlying idea, the squared cosine is checked in order to guarantee a minimum quality of PC representation of the unknown spectrum. As

min_k is always higher or equal to min_lib this second criterion is less restrictive than the previous one. If the unknown spectrum is well represented this new criterion may provide more candidates to identify it. A graphical interpretation of these criteria is represented in Fig. 4.5.

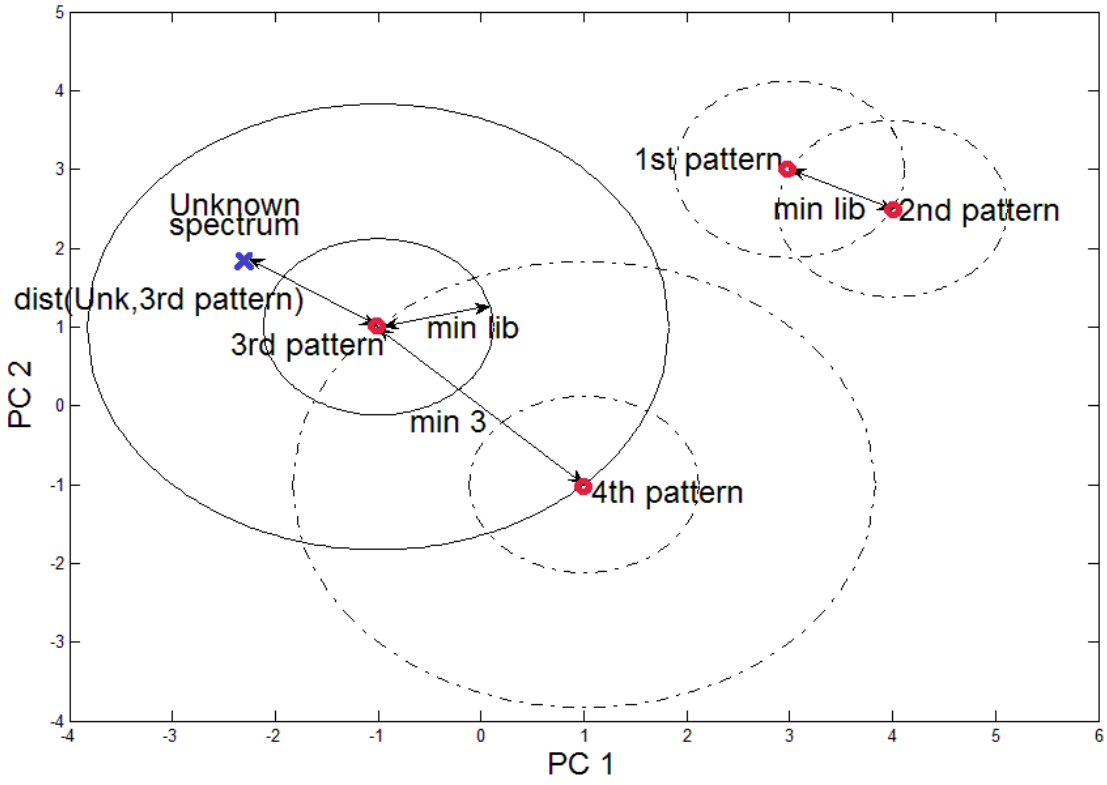


Figure 4.5: Graphical interpretation of the identification criteria in a two-dimensional space. min_lib stands for the minimum distance between the spectra of the database, and min_3 stands for the minimum distance between the 3-rd pattern and the rest of the patterns. In the presented example, if the unknown spectrum is well-represented, it may be identified as the 3-rd pattern, since the distance between the unknown spectrum and the 3-rd pattern is lower than min_3.

The identification criteria may be fulfilled by different patterns, so a parameter was defined to interpret the result of the identification, te so-called Matching Factor (MF):

$$MF_k = 1 - \frac{ED(unk, k-thpattern)}{max\{min_{cands}\}} \tag{4.8}$$

which is computed for each candidate k, and where $max\{min_{cands}\}$ is the maximum ED of the minimum ED between the candidates and the rest of the reference spectral library. This factor provides information regarding to the similarity between the unknown spectrum and the candidates: the more similar the unknown and the candidate, the higher the Matching Factor, and the other way around, the less similar, the lower the Matching Factor (as for instance in cases where other components like binding agents are present which may mask some of the pigments' bands). In this sense, the

quality of the measured spectra (in terms of signal to noise ratio) is a key factor in the identification process: the higher quality the unknown spectra have, the more similar the unknown spectrum with respect to the corresponding reference may be. The diagram shown in Fig. 4.6 summarizes the scheme of the identification methodology.

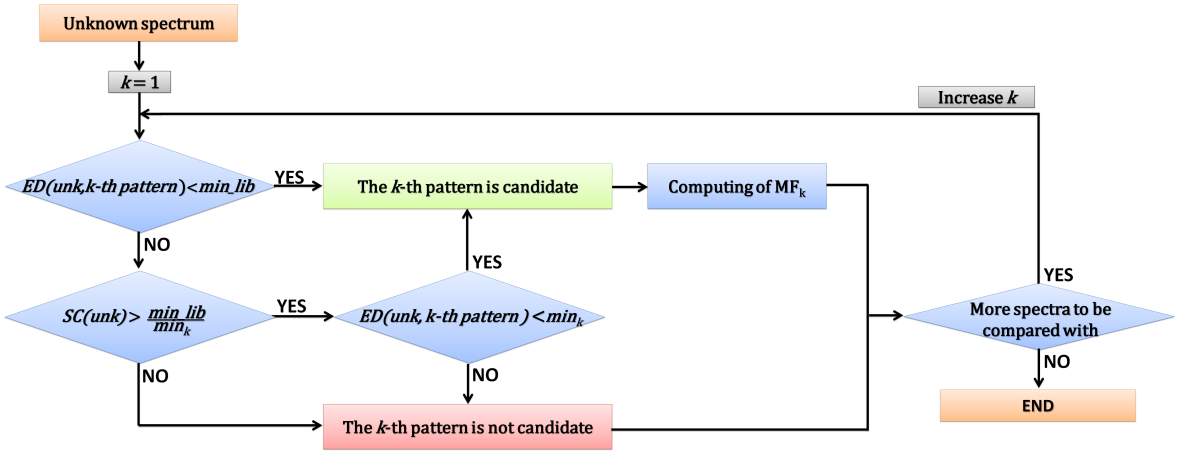


Figure 4.6: Overview of the identification scheme

4.3.3 Multi-component identification

A particular analytical problem can arise from the comparison between an unknown spectrum that comes from a mixture of pigments and the reference spectra of individual pigments since it is difficult to obtain the appropriate reference spectrum corresponding to the unknown mixture[131]. Additionally, in a practical situation it is not known initially whether the analysed spectrum originates from a single pigment or a mixture of several compounds. E.g. a green appearance may be achieved with either green pigments or a mixture of blue and yellow particles. Given these uncertainties that are particularly relevant to mobile systems with low spatial resolution or mixtures of synthetic organic pigments and dyes, the development of a mixtures identification strategy is interesting for identifying the components in Raman spectra from pigment mixtures.

In this sense, a preliminary analysis was proposed using a mathematical spectrum which represents a hypothetical mixture considering that a Raman spectrum of a pigment mixture contains separately the spectroscopic signature of each pigment. However, this preliminary proposal is only valid for the more commonly found mixtures, i.e. the binary mixtures. When applied to admixtures of more than 2 components the usage of mathematical spectra may fail so another approach was developed. The final multi-component identification methodology is a blind method. Specifically, it identifies single and multi-component spectra without user input or judgement using a single spectral observation, i.e. the unknown spectrum. There are thus no parameters to be

tweaked. Furthermore, it provides a Matching Factor on the resulting identification, aimed at becoming a useful support tool for the analyst in the decision-making process.

Binary mixtures identification

In practice, the pigments may have been used in mixtures (or in admixtures) with other pigments to produce special effects or tonal qualities. Binary mixtures are the more frequently mixtures used in paintings; for instance, it is usual the mixture of a yellow pigment with a blue pigment to produce green colours. In this situation, the pigment identification is actually done on Raman spectra of pigment mixtures. To decrease complexity whilst also speeding up the identification process, a system to automatically identify Raman spectra of binary mixtures of pigments was developed. The system is able to identify the two different pigments in the mixture from spectroscopic signature obtained by Raman spectroscopy. The technique has been proved with mixtures showing its robustness against some of the critical factors that could affect the application of Raman spectroscopy for pigment identification as distortion and wavenumber shifts in key Raman bands due to different measurement environmental conditions.

Let x'_i and x'_j be the i-th and j-th patterns which are candidates and x_m the original expression of their mixture, $x_m = x'_i + x'_j$. The PC-expression of the mixture's spectrum after some mathematics is

$$s_{mixture} = k \cdot s_i + k \cdot s_j + (2k-1)\frac{m_{lib}}{\sigma_{lib}}C \tag{4.9}$$

being s_i and s_j the PC-expression of the corresponding candidates and k the constant $\frac{1}{max(x_m)}$. The "spectrum" of a mixture built in this way (adding normalized spectra directly one by one), assumes that all bands from the individual pigments are present maintaining their relative intensities. In a practical situation this is rarely the case and may become a weakness of the proposed methodology. However, this strategy solves the incorporation to the reference spectral library of all possible mixtures. Dealing with binary mixtures in a reference spectral library of m spectra the binomial coefficient of m and 2 may be built. For instance, taking $m = 20$, 190 mixtures may be obtained. Moreover, taking into account all the possible configurations of a mixture in terms of proportions of the corresponding individual pigments, a large number of mixtures may be built. Thus, the possibility of building all the binary mixtures for a given reference spectral library has been completely ruled out. In addition, programmatically, this tactic would lead to a computing resources consuming code. A new strategy is developed instead, which avoids the manufacturing of all these mixtures and their subsequent measurement. The mixtures are only created when one of the following criteria is fulfilled:

1. If there are candidates with non-negligible Matching Factors of similar order, that

is, higher than 10% and with a difference lower than 30%, then these candidates may compound the mixture, i.e. if $MF_i, MF_j > 10\%$ and $|MF_i - MF_j| < 30\%$ then the mixture is created with the candidates i and j.

2. If there are not candidates or all candidates have a Matching Factor lower than 60%, the system automatically sorts the distances between the unknown spectrum and the rest of the library and get the two patterns that have the lowest distances. Then these two patterns may create the mixture.

It should be noted that the values taken in the mixture-building criteria can be modified to make the identification process more or less relaxed depending on the user requirements. The values proposed in this research were established after the analysis of the algorithm performance in a simulation stage, which best suited for proper operation. Once a mixture is built, if appropriate, it is seen as a new pattern. Then, the identification criteria are applied with the unknown spectrum and the created mixture, allowing the identification of spectra of binary mixtures. That is, the proposed algorithm confirms, scientifically, the presence of various pigments in an analysed sample through mathematical mixtures. The diagram shown in Fig. 4.8 summarizes the implemented identification methodology with binary mixtures handling. To show the performance of the implemented algorithm, the developed methodology was applied to Raman spectra from different handmade samples (see Fig. 4.7). These handmade samples were constructed by mixing artistic pigments up two by two with random proportions and were enhanced according to the methodology described in Sect. 3.3.

Figure 4.7: Measurement of experimental Raman spectra from handmade samples used for assessing the performance of the implemented identification methodology

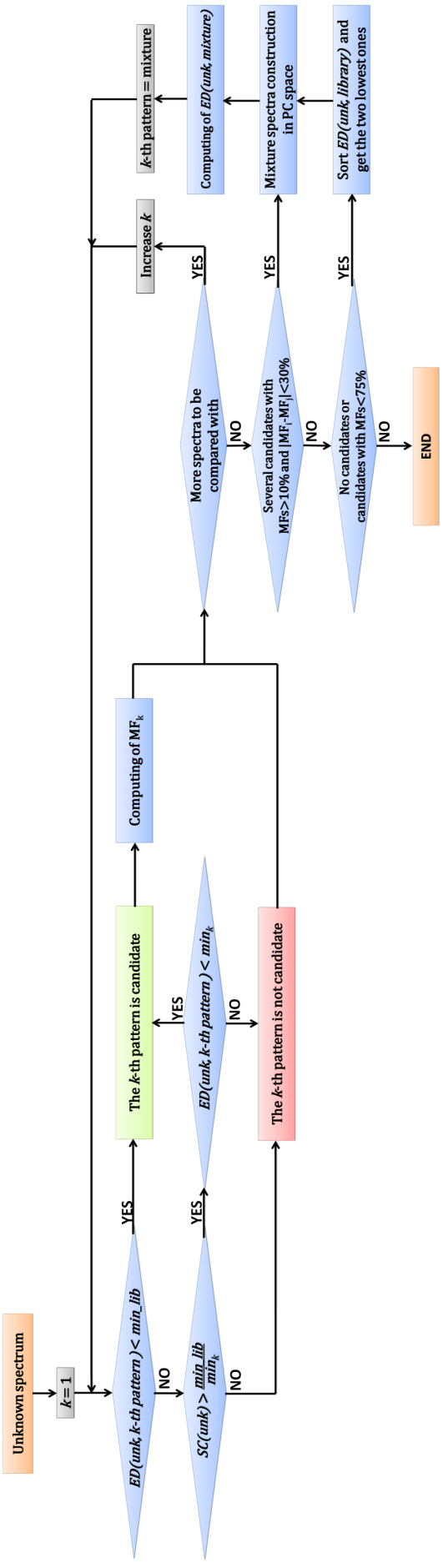


Figure 4.8: Overview of the identification scheme with binary mixtures handling

Sect. C.1.1 of Appendix C compiles the experimental results of applying the implemented identification methodology to the Raman spectra from handmade samples, which provided successful results for binary mixtures. The developed mixtures identification methodology was focused on binary mixtures since this kind of mixtures appears with relative frequency in art. However, in some cases mixtures of more than two components may appear as well. In the case of ternary mixtures for instance, the methodology does not identify the three components of the mixture since it is a construction limitation of the mixtures treatment algorithm.

In this sense, it was proposed to extend this algorithm to identify not only binary mixtures but ternary mixtures as well and then extrapolate it to identify mixtures of any number of components. Nevertheless, after testing in several cases this extrapolation, the mixture-building criteria failed to identify Raman spectra of mixtures of more than two components. To overcome this issue a new strategy was developed. This strategy is based on another chemometric technique, the so-called independent component analysis (ICA)[69], without changing the identification criteria but the mixture criteria, developing a more accurate methodology in terms of identification of mixtures, being a blind solution capable of identifying mixtures of more than two components. The generalisation of the identification methodology for single- and multi-component Raman spectra is described hereafter.

Generalised identification methodology for single- and multi- component Raman spectra: Mixtures Separation Block definition

The methodological scheme of the developed blind approach for identifying single- and multi- component Raman spectra is presented in Fig. 4.9. It follows a flowchart allowing iterative data processing and is built on a sequential two-step selection process with the initial Identification Block (IB) as described in Sect. 4.3.2 and later - applicable to multi-component spectra - the Mixtures Separation Block (MSB).

ICA is the core of the methodology block developed for mixtures handling, called the Mixtures Separation Block (MSB). Generally speaking, ICA is a technique that recovers a set of independent signals from a set of measured signals[70]. It is assumed that each measured signal is a linear combination of each of the independent signals, and that there are an equal number of measured signals and independent signals. Using vector-matrix notation, the ICA model is written as $x = As$, where x is the vector whose elements are the measured signals, s is the vector whose elements are the independent signals and A is the connecting matrix. All we observe is the vector x, and we must estimate both A and s using it. Then, after estimating the matrix A, we can compute its inverse, say W, and obtain the independent component simply by $s = Wx$.

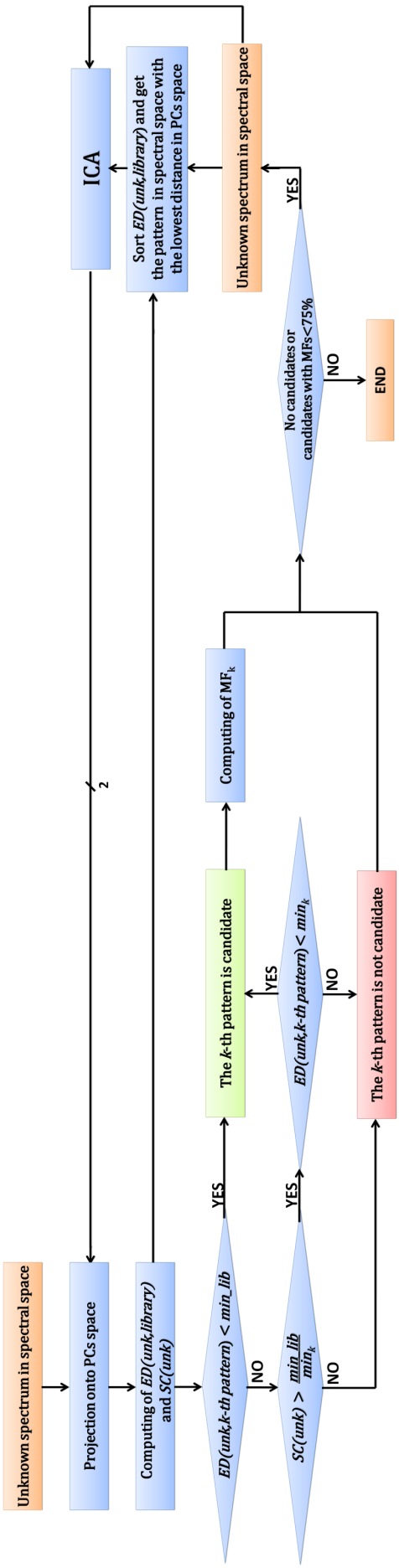


Figure 4.9: Overview of the implemented methodology based on independent component analysis for the automatic identification of single- and multi-component Raman spectra applied to pigments analysis

In general, in ICA-based applications at least n measured signals are needed to work with mixtures of n components. In the context of pigment identification in art works through Raman spectroscopy, this would imply to measure more than one Raman spectrum on the same spot of a work of art: two Raman spectra to identify binary mixtures, three Raman spectra to identify ternary mixtures, and so on. However, in a practical situation we do not know beforehand if the analysed spot is a mixture of pigments or not, and, in case it is a mixture, how many pigments were used. For instance, a spot corresponding to an orange hue on the art work may have been painted with an orange pigment or, a mixture of yellow, red and potentially white pigments. This would imply a drawback when working with ica, since the main idea of the identification methodology is that it must work with just one measured spectrum and without prior knowledge about the analysed sample.

Bearing in mind the above mentioned observation, the current research introduces an iterative solution based on the assumption that the spectral signature of the reference spectrum with the lowest ED is present in the unknown spectrum, and therefore they can be separated through ICA. Thus, the steps in the proposed solution are as follows:

1. The reference spectrum which has the lowest ED is selected
2. ICA is applied to this selected reference spectrum and the unknown spectrum. From this analysis two separated spectra are obtained
3. These two spectra are treated as two new unknown spectra and are delivered to the IB. If the two spectra are identified by the IB then the original unknown spectrum may come from a mixture of the reference spectra identified

These 3 steps can be applied iteratively, which eventually allows to identify the total number of reference spectra present in the unknown spectrum.

The presented MSB should be triggered only in case of dealing with spectra coming from mixtures. To do so, a "mixture criterion" was defined, which is applied to the outcome of the first stage of the proposed chained methodology, i.e. the IB. In this sense, and attending to the fact that the automated identification analysis done by the IB may be hindered and even avoided due to the presence of the spectroscopic signature of two or more components in the unknown spectrum, the "mixture criterion" was defined as follows:

If there are no candidates or all candidates have a MF lower than a certain value of MF (MF_{th}), then the unknown spectrum may come from a mixture.

An exhaustive study was conducted in order to determine the appropriate MF_{th} to trigger the MSB. This study was performed using simulated Raman spectra, and specifically fluorescence-free simulated spectra. The study was done by the following experiment:

1. A reference spectral library was simulated composed of one hundred different simulated spectra

2. One thousand unknown spectra were generated. Each of them was created by a linear addition of one, two or three reference spectra randomly picked up from the spectral library

3. For each of these unknown spectra the identification methodology was applied for several values of MF_{th} ranging from 0% to 100% in steps of 5%, which means that it was applied twenty-one times for each of the unknown spectra

4. As a quantitative measure of the quality of the "mixture criterion" the specificity and the sensitivity parameters were calculated for each MF_{th}. The specificity is defined as the percentage of unknown spectra recognized as mixture among the "true" mixed spectra. Similarly, the sensitivity is the percentage of unknown spectra recognized as single spectrum (i.e. not recognized as mixture) among the pure reference spectra. Furthermore, the success rate (defined as percentage of unknown spectra properly identified), was calculated for each MF_{th} from the identification results

The above four steps were repeated one hundred times, each time using a different simulated library and therefore different unknown spectra. Then, a statistical analysis was performed. The mean success rate as a function of MF_{th} is represented in Fig. 4.10. The figure reveals that the values of the MF_{th} ranging from 90% to 100% provide the highest mean success rate. The minimum MF_{th} providing the highest sensitivity and specificity was 90%.

From this study, we concluded that the appropriate MF_{th} is 90%. This value ensures that all the simulated mixture spectra triggered the MSB and also optimizes the processing time. Consequently, the MF_{th} of the "mixture criterion" in the presented identification methodology is henceforward fixed to 90%.

On the other hand, the system stops iterating when the "mixture criterion" is not fulfilled or when no new mixed components are identified. In addition, the EDs between the ICA outputs and their corresponding inputs are compared with the ED between the ICA inputs. This is done in order to filter out instances where the reference spectrum with the lowest ED is not present in the unknown mixture (as may be the case

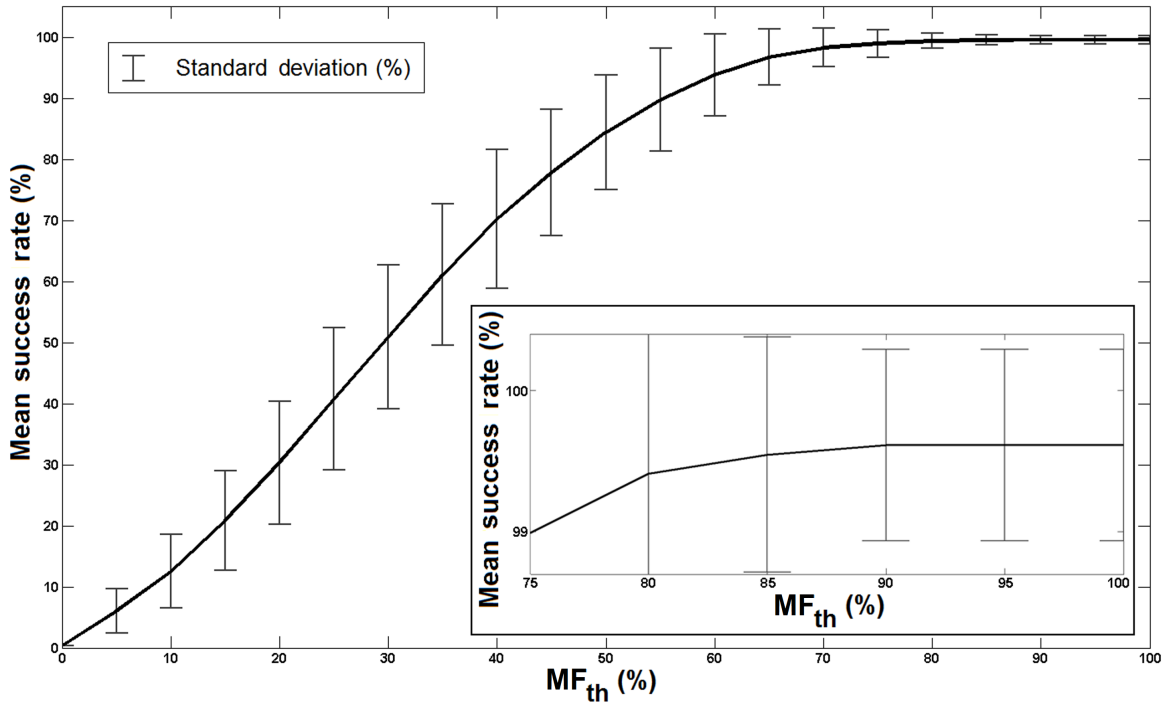


Figure 4.10: Mean success rate (and corresponding standard deviation as vertical bars) as a function of MF_{th}. For each MF_{th} (from 0% to 100% in steps of 5%), the mean success rate was computed from the percentage of unknown spectra successfully identified when applying the proposed method 100 times, each time identifying 1000 different unknown spectra. Inset figure shows a zoom for MF_{th} from 75% to 100%

when analysing an unknown spectrum whose corresponding reference spectrum is not available in the reference spectral library).

Finally, it must be pointed out that this strategy allows the identification of components used in a mixture with just one measured spectrum, as long as their respective spectroscopic signatures are in the unknown spectrum.

Results and discussion

In order to get a good overview of identification system response, two different scenarios were analysed. These two scenarios were based on simulated spectra and on experimental spectra. Sect. C.1.2 of Appendix C summarises the outcome of applying the generalised identification methodology in simulated environments, which provided successful identification results. The analysis on experimental environments are reported hereafter. In particular, the developed method was applied to experimental Raman spectra from handmade mixtures and from paintings. Specifically, the handmade samples were manufactured by mixing artist's pigments in random proportions. Three real-case examples are presented and discussed hereafter.

In a first example, the analysed spectrum was measured on a handmade mixture of the pigments PY1 and PG7, manufactured by Sennelier (see top of Fig. 4.11a). When the IB was applied to the spectrum one single candidate was found: the PY1 pigment with MF(PY1) = 45.7%. Since this MF is lower than the one established for the mixture criterion (90%), the MSB was triggered, and then the whole identification system provided two candidates: the PY1 with MF(PY1) = 77.8% and the PG7 with MF(PG7) = 71.1%. This result, which suggested that the unknown spectrum may correspond to a mixture of the PY1 and PG7 pigments, was consistent with the analysed mixture. In top of Fig. 4.11b the spectra can be examined.

The spectrum analysed in a second example was measured on a handmade sample made by mixing the pigments PY1, PR4 and PB15, manufactured by Sennelier (middle of Fig. 4.11a). The IB provided one candidate: the PY1 pigment with MF(PY1) = 20.9%. Thus, the MSB was triggered and three candidates were found: the PY1 pigment with MF(PY1) = 84.4%, the PB15 pigment with MF(PB15) = 85.4% and the PR4 pigment with MF(PR4) = 43.5%. This result suggested that the unknown spectrum may correspond to a mixture of the PY1, PB15 and PR4 pigments. Hence, the methodology was able to identify successfully the three pigments used in the handmade mixture, although it assigned different MFs to each candidate. Concretely, the PR4 pigment got a relatively low MF due to the fact that it has bands in common with the other candidates as can be seen in middle of Fig. 4.11b.

In a last example, the analysed spectrum (bottom of Fig. 4.11b) was measured directly on a spot with a pinkish hue of a painting representing an image of a *Saint Engratia* (bottom of Fig. 4.11a), linked to the Aragonese School (17th century). When the IB was applied to the spectrum two candidates were found, the vermilion pigment with MF(vermilion) = 30.9% and the white lead pigment with MF(white lead) = 18.4%. Thus, the MSB was triggered and three candidates were found: the white lead pigment with MF(white lead) = 84.8%, the vermilion pigment with MF(vermilion) = 82.7% and the barite pigment with MF(barite) = 74.6%. Consequently, the result suggested that the analysed sample may correspond to a mixture of the white lead, the vermilion and the barite pigments. Note that the fundamental band of the pigment identified with the highest MF is not the fundamental band in the mixed spectrum, which may show that the system has no difficulties in dealing with mixed spectra of different intensities. Furthermore, although some secondary peaks of the three pigments were lost due to noise, the MFs of the identification were relatively high (see bottom of Fig. 4.11c).

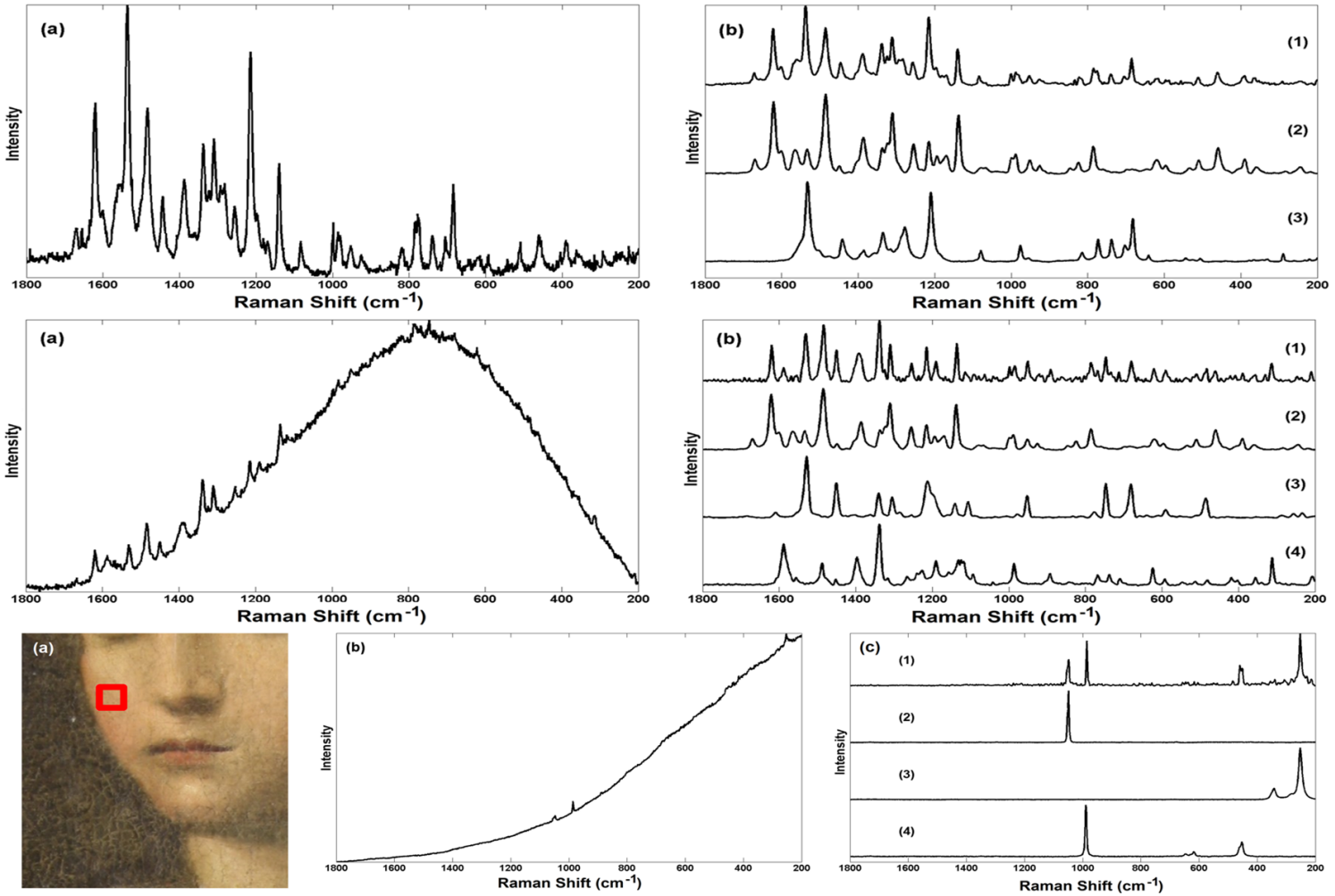


Figure 4.11: Top: a) Unknown Raman spectrum, b) Pre-processed unknown spectrum (1) together with the reference spectra of the pigments identified PY1 (2) and PG7 (3). Middle: a) Unknown Raman spectrum, b) Pre-processed unknown spectrum (1) together with the reference spectra of the pigments identified PY1 (2), PB15 (3) and PR4 (4). Bottom: a) Partial image of the analysed painting representing a *Saint Engratia* (17th century) -analysed spot marked with a red box-, b) Unknown Raman spectrum, c) Pre-processed unknown spectrum (1) together with the reference spectra of the pigments identified white lead (2), vermilion (3) and barite (4)

4.4 Automated pigment classification through Raman spectroscopy

As commented previously, it is well known that Raman spectroscopy is able to distinguish different molecular species based on the acquired Raman spectra. The discrimination between the pigments found in natural and synthetic forms[132–134] or in different crystalline structures[135–139] is an important topic in conservation science because the pigments may differ not only in their chemical and physical characteristics (such as stability, solubility and hue) but also appeared at different times on the paint market and thus they may be used as chronological markers. Certain pigments can be found in different crystalline structures as the copper-phthalocyanine blue pigment for instance, and the differences in their spectral data may go unnoticed. Indeed, these little differences in the spectral data may occasionally lead to a subjective interpretation or to the need of aggregating data from different analytical methods, making the identification a costly and time-consuming process. Automated distance-based identification algorithms as the IB algorithm described in the previous section may not be able to discriminate little differences as the distance metrics of the corresponding patterns may be too low that may invalidate the identification criteria implemented therein. Thus, the development of classification tools that can help the analyst in making decisions has become a trending topic[140–144].

Most papers that handle the classification issue are based on chemometrics where the identification features are manually retrieved from the spectra. As a result, a certain degree of subjectivity is still incorporated to infer the classification. Our premise, however, is that no user input should be required. This means that the process of assigning the class an unknown spectrum belongs should be fully automated. Hence, multivariate analysis techniques based on machine learning were explored in this research in order to design an analytical method to automatically classify artistic pigments from their Raman spectra in a transparent way regarding the classification topic: the material's provenance, the crystalline structure, or any other classification matter. Machine learning brings together computer science and statistics to quickly gain insights and make predictions from the input data. Hence, machine learning is used to find patterns in data and to build models that predict future outcomes based on historical data. Statistical classification is an example of a machine learning task, which is aimed at identifying to which of a set of categories a new observation may belong based on a set of reference data containing observations whose category membership is known beforehand. Generally speaking, two sets of data are defined in the context of machine learning: the training set and the testing set. Specifically:

- *Training set*: predefined set of reference data used to train the system. The

response values of the training dataset are known, i.e. each element of the dataset includes a label which identifies it with its own category.

- *Testing set*: set of unlabelled data used to validate the model.

Among the different types of machine learning algorithms, a crucial distinction is drawn between unsupervised and supervised learning:

- **Unsupervised machine learning**: These kinds of algorithms draw inferences from datasets consisting of input data without labelled responses in order to find patterns and relationships therein
- **Supervised machine learning**: These kinds of algorithms are trained on the training dataset using the corresponding labelled responses of each element in the dataset as a prior information. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset with unknown labelled responses

The following sections discuss the usage of unsupervised and supervised machine learning techniques for the discrimination of Raman spectra from pigments from different categories showing small differences among them.

4.4.1 Unsupervised classification methodology

The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find grouping in data. The clusters are modelled using a measure of similarity. In order to perform an objective discrimination of Raman spectra from artistic pigments, the use of clustering techniques was proposed. With this objective, different algorithms were analysed and evaluated depending on the configuration parameters of each technique using simulated spectra whose category was a priori known. In this sense, a simulated spectrum was generated by combining a variable number of Lorentzian-profile-based bands with random locations, amplitudes and Full Width at Half Maximum (FWHM), constrained such that it appeared qualitatively similar to real Raman spectra. Specifically, the Lorentzian function which implements the Raman bands in a simulated spectra is defined as:

$$f(x, x_0, A, B) = \frac{A}{1 + (\frac{(x-x_0)}{B/2})^2} \tag{4.10}$$

where x_0 is the band mean, A its amplitude, and B its bandwidth when the band amplitude has dropped by a half (FWHM). In particular, the techniques of k-means, Expectation-Maximisation (EM), hierarchical clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) were analysed.

k-means

The basic idea of the k-means[145] is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. To achieve this, the algorithm minimizes the within-cluster sum of squares (WCSS):

$$WCSS_i = \sum_{j=1}^{k} \|x_j - \mu_i\|^2 \tag{4.11}$$

where μ_i is the arithmetic centre of the i-th cluster (i.e. the cluster centroid). The algorithm starts by creating a set of k clusters randomly distributed. From this initial setup the algorithm reduces the distance between the members of the cluster and its centroid is updated at each iteration. The biggest drawback of the k-means algorithm is that the number of clusters must be set beforehand by the user. However in our application it is clear that this information will be known beforehand. The most important configuration options for the k-means algorithm are the number of clusters (namely k) and the initialisation method. In particular, the initial k centroids can be set using some criteria instead of randomly. The final result may depend on the initial position of the cluster centres as the solution may converge to a local optima. Therefore, the different initialization methods will be tested to select the method which best fits the requirements of this research. The tested initialization methods were: random (0), k-means++ (1), canopy (2) and farthest first (3).

Expectation-Maximisation

The Expectation-Maximization (EM)[146] is an optimization technique that assigns each element to a predefined cluster according to its probability of belonging to that particular group. For this, a Gaussian distribution function is used in order to adjust its parameters according to how the different elements are adapted to the distribution of each group. EM is two-step iterative process:

- Expectation: the first step of the process uses the values of the parameters whether initial or provided by the step maximization of the previous iteration in order to estimate the belonging probabilities of the elements to each of the models that characterize the different groups.

- Maximization: the second step of the process takes the belonging probabilities calculated in the expectation step, re-estimates the distribution parameters that maximize their likelihood.

The most important configuration options for the EM algorithm are the number of clusters (k) and the number of iterations expected to achieve convergence.

Hierarchical clustering

Hierarchical clustering[147] methods try to build a hierarchy of clusters. In the case of agglomerative algorithms, the process starts with each observation creating a cluster. From there, pairs of clusters are merged. This process is repeated until all the observations are grouped into a single cluster unless a stop threshold is configured. In this implementation, the stop threshold is the desired number of clusters to be produced. Variants of this method rely on how the distance between clusters is considered, the so called the linkage criteria. As clusters are extended objects, the distance between a pair of them can be computed from their centroids, their farthest elements, their closest elements, etc. The most important configuration options are the number of clusters (k) and the linkage criteria. Regarding the linkage criteria, the following configurations were tested: single (0), complete (1), average (2), mean (3), centroid (4), ward (5), adjusted complete (6), neighbour joining (7).

Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[148] is a density-based clustering algorithm. In DBSCAN, points are classified as core points and outliers. A given point becomes a core point if it has more than a predefined number of points in its surrounding area, in other words a certain number of points are found within a defined radius. These points are called *reachables*. If p is a core point, it forms a cluster with all points that are reachable from it. One of the features of this algorithm is that if points are isolated they might be considered as noise. In our application all points should be treated so a specific configuration will be used to avoid this behaviour. The most relevant configuration options for DBSCAN are:

- Threshold distance, ϵ: The distance within, for a given point, the others are considered *reachables*.
- Minimum points, N: The minimum number of points to consider within a defined radius to start grouping points together.

Comparative analysis of clustering techniques applied to Raman spectra

The clustering algorithms were evaluated under controlled test scenarios generated through simulations. In this way, the performance of the algorithms can be evaluated by comparing the *true* information (i.e. the correct category for each simulated spectra) provided by the simulation against the generated clustering results. It is desirable to quantify the success rate of the clustering in a single number. This can be used, not only to characterize the performance in a given situation, but to tune the parameters of a clustering algorithm in order to optimize the performance. From the simulations we

know the correct assignment of every simulated spectra to a unique category, i.e. the *true* clusters. In this analysis, the success rate has been defined as the ratio between the correct clustered spectra by the total number of spectra. In order to compute the correctly clustered spectra, a correspondence between true clusters and the resulting clusters from each method has to be found. This has been done by the maximum number of coincidences criteria using the training dataset as testing set.

Prior to the application of clustering, the dimensionality reduction of PCA was applied to the P reference spectra of the input dataset in order to reduce redundancies and speed up the processing time. In the recognition case, the main point to bear on mind when applying PCA is no data loss, achieved by a PCs space of $P-1$ dimensions. In this way, an unknown Raman spectrum projected onto the PCs space can be compared with the projected reference Raman spectra using all the available information. In the classification case though, it is important to highlight the differences between the input categories. As PCA provides a transformed space in which the axes are sorted according to the variability in the wavelength domain of the input dataset in a descending order, the inter-category differences are highlighted in the first PCs. In this sense, a lookup for the dimension of the PCs space providing the maximum success rate using the training set as testing set should be performed. This lookup is carried out through an evaluation process based on sweeping the PCs space dimension, applying the clustering analysis and then computing the success rate for each case. With this process, both the optimal PCs space dimension and the optimal configuration parameters for each clustering technique can be obtained.

The workflow of the clustering algorithm evaluation process is shown in Fig. 4.12. In particular, 100 different datasets were simulated. Each dataset contained $P = 60$ spectra divided in 3 different groups of 20 spectra each. For each dataset, PCA was applied sweeping the PCs space dimension from 2 to $P - 1 = 59$. Then, each of the analysed algorithms was applied and the success rate was computed for each case. This analysis allowed to obtain the optimal configuration parameters for each of the analysed algorithms (see Fig. 4.13). In general, the highest success rates were obtained with the lowest dimension of the PCs space. The k-means optimal configuration was k set to 3 (the number of *true* clusters) with random initialisation. For the EM case, the optimal k was also set to 3, iterations-independent. For the hierarchical case, the optimal configuration parameters were single linkage and k set to 3 as well. Finally, the optimal parameters for the DBSCAN case were a distance threshold set to 5 and minimum points set to 9. It was concluded that the algorithm that provides the best success rate and wall-clock performance using PCA as data reduction tool is k-means with k set to the number of clusters to be created, fixing the PCs space dimension to the lowest number of PCs providing the maximum success rate. Additional results are shown in Sect. C.2.1 of Appendix C.

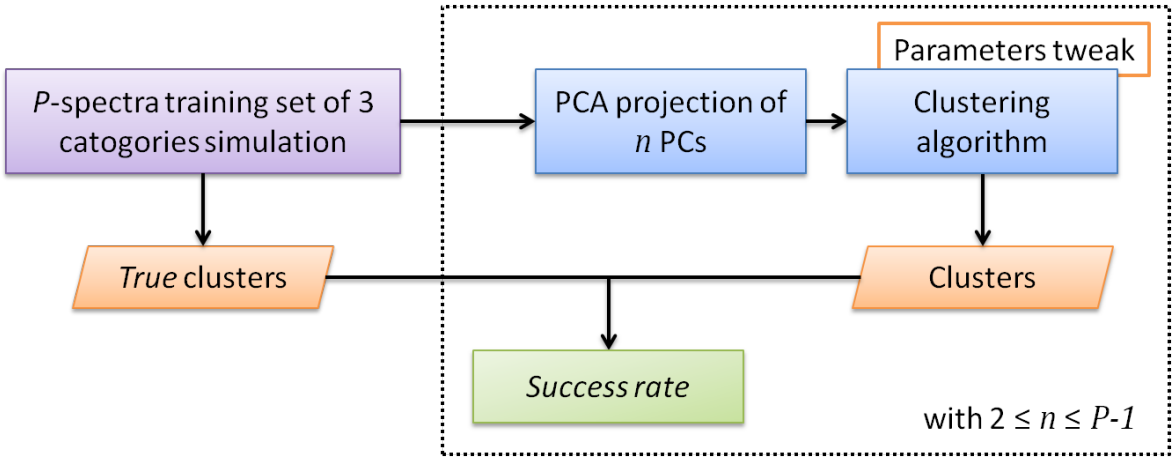


Figure 4.12: Schematic workflow of the clustering algorithm evaluation process through simulated datasets consisting of P spectra aimed at obtaining the optimal configuration parameters for each clustering technique using PCA as a data reduction tool sweeping the PCs space dimension from 2 to $P-1$

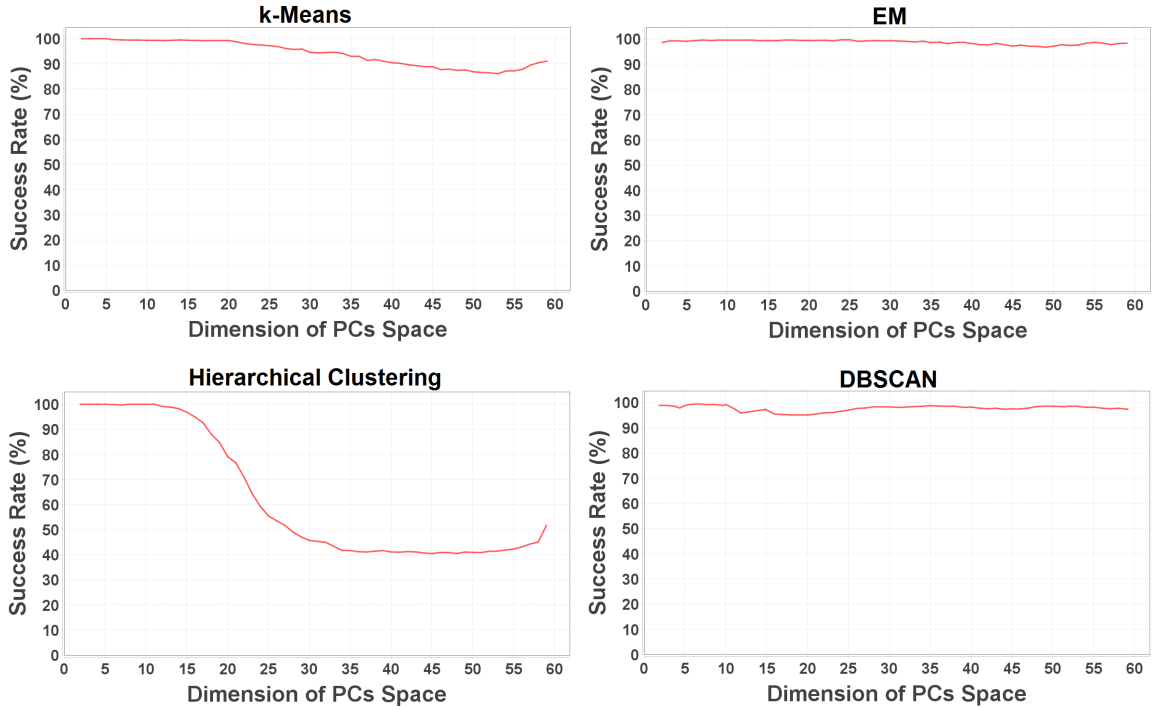


Figure 4.13: Success rate as a function of the PC space dimension using the the optimal configuration parameters: k-means (top left) with k set to 3 (the number of *true* clusters). EM (top right) with 5 iterations and k also fixed to 3. Hierarchical clustering (bottom left) with single linkage and k fixed to 3 as well. DBSCAN (bottom right) with a distance threshold set to 5 and minimum points fixed to 9

The methodological scheme of the unsupervised classification system is shown in Fig. 4.14. The classification of an unknown Raman spectrum is based on a two-step process. First, the system is trained with P reference Raman spectra looking for the optimal PCs space dimension, n_{opt}, i.e. the lowest dimension of the PCs space that provides the maximum success rate when applying k-means with random initialization and k set to the number of cluster to be created. From this process, the optimal PCA projection and the k-means centroids are retrieved. Finally, the unknown Raman spectrum is projected onto the PCs space using n_{opt} PCs and the nearest k-means centroid is computed for cluster assignment.

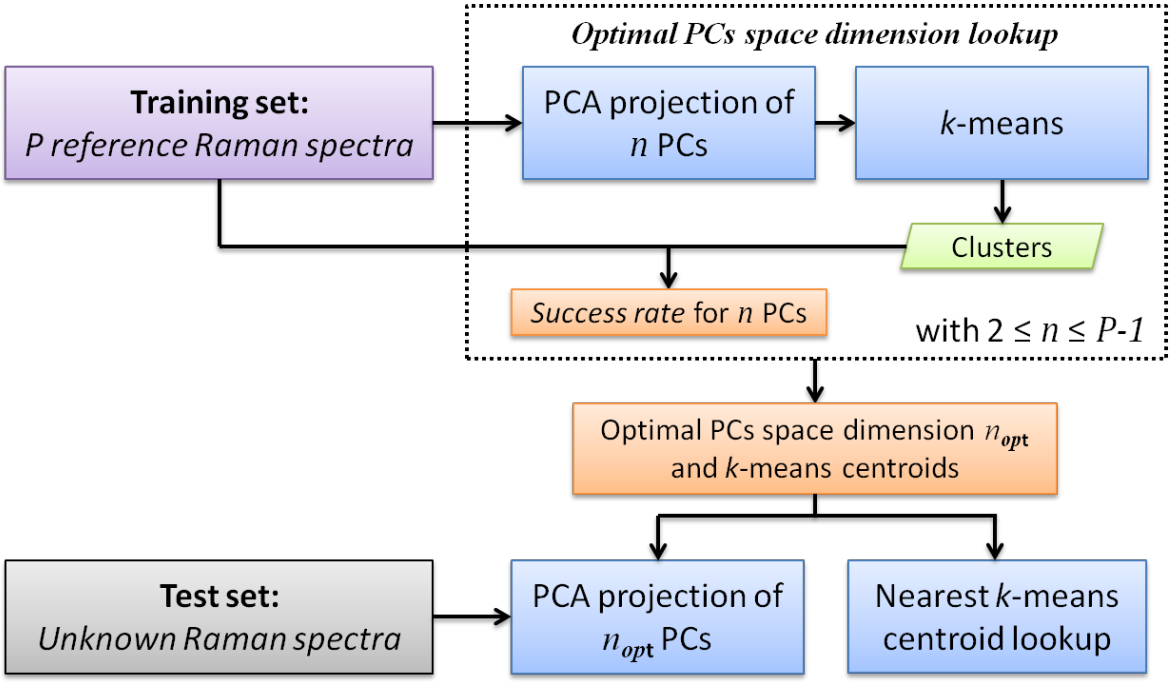


Figure 4.14: Overview of the unsupervised classification methodology

Section C.2.1 of Appendix C compiles the results of applying the described unsupervised classification methodology based on PCA and k-means with optimal parameters to experimental Raman spectra. In particular, the methodology provided successful results for Raman spectra measured with the same excitation wavelength but it could not perform a proper clustering of Raman spectra measured using different excitation wavelengths. This issue may represent a drawback for classifying Raman spectra measured using different instrument resolution, excitation wavelength or even laser power of the excitation source. Bearing on mind that a classification methodology should be a blind method, it should not depend on the measurement configuration of the input dataset. To overcome this limitation a supervised machine learning-based classification methodology is described in the following section.

4.4.2 Supervised classification methodology

A supervised classification methodology was developed in order to classify unknown Raman spectra according to predefined classes in a consistent way. To do so, it is necessary to rely on a specific classification strategy that allow the objective comparison between unclassified spectra and reference classes. This strategy is outlined hereafter.

The classification of artistic pigments fits the standard scheme of statistical classification[149], which is a supervised learning technique in the field of machine learning and statistics. It deals with the process of identifying to which of a set of classes an unclassified item belongs to, based on a training dataset containing references whose class membership is known beforehand. The standard classification scheme is built from two different stages: data acquisition and data processing. In the case of pigments analysis through Raman spectroscopy, the data acquisition stage is based on the Raman spectrometer. On the other hand, the data processing stage is composed of three different modules: feature extraction, classifier and decision-maker. First of all, the feature extraction is the process of defining a set of features, which most effectively represent the important information for classification. We selected PCA for this purpose as it is the technique that best fits the data dimensionality requirements for this research, as described in the previous section. Then, the classifier is the multivariate technique aimed at maximizing the inter-class distances whilst minimizing the intra-class differences from an appropriate set of class features. We selected Multiple Discriminant Analysis (MDA) for finding a combination of features that separates the user-defined classes, i.e. training dataset. From the set of extracted features by PCA, MDA provides a new space, the so-called classification space. Finally, the decision-maker is the procedure in which an unknown or unclassified element is projected onto the classification space and is assigned to one of the classes according to some metrics that will be discussed hereafter.

Next, we describe the characterization of the training dataset in the classification space and the procedure of class assignment for unknown spectra.

Characterization of the classification space

In the case of pigments analysis through Raman spectroscopy, the training dataset is composed of sets of reference Raman spectra, i.e. reference classes. These reference classes are decided by the user according to the classification purpose. The training dataset is represented by a matrix, S, which is divided in sub-matrices. Each sub-matrix S_i identifies a known class where each row is a spectrum of the $i-th$ reference class. The classification space is obtained by applying first PCA (feature extractor) to the training dataset and later MDA (classifier) over the PCA result. In this space, the training dataset for the $i-th$ class is now represented by a matrix C_i where each

row is a spectrum in the classification space. Each class is delineated by a region and is characterized by a centroid (the arithmetic center, μ_i) and a dispersion matrix (the auto-covariance matrix, Σ_i).

In order to perform an efficient classification, proper class separability in the classification space should be obtained. This class separability is checked by computing the JMD between all the classes. The classes in the classification space are totally distinguishable when JMD is equal to 2 while lower values indicate a worse separability. We use the class separability as a parameter to generate an adequate classification space, selecting a proper number of features. As we said previously, the PCs scores obtained in the feature extraction module are the distinctive features for each class. Then, we tune the number of PCs scores, successively until there is no improvement in the class separability. We considered a JMD value greater than 1.75 for achieving good class separability. In this way, starting from a number of PCs scores equals to the minimum number of spectra in $C_i \forall i$, the JMD is calculated in the tentative classification space obtained by performing PCA followed by MDA. The number of PCs scores used is increased by one until the desired JMD value is achieved or until the number of PCs scores is greater to the maximum number of spectra in $C_i \forall i$. If the class separability is achieved through the obtained number of PCs scores, it means that the user-defined reference classes allow the classification of unknown spectra. Otherwise, no class separability is achieved with the defined classes and must be re-defined. The organization chart of this procedure is outlined in Fig. 4.15.

Additionally, taking into account that outliers in a class can deform the class characterization, a basic statistical rule for outlier rejection is applied to each reference class in the classification space. This rule is defined as: if $x > \mu_i + 2\Sigma_i$ then x is rejected (being x a spectrum of the i-th class in the classification space). When a reference spectrum is rejected, the class parameters (centroid and dispersion matrix) are automatically recomputed for that class.

Class assignment Once the classification space is characterized, we defined a classification rule to assign an unknown spectrum to a reference class. For this purpose, we developed an autonomous matching technique based on distance metrics. Specifically, the ED and the MD were used.

First, we compute the ED between an unknown spectrum in the classification space (x) and the class centroid (μ_i). Second, we calculate the MD between an unknown spectrum in the classification space (x) and the class (C_i). These metrics express intuitive notions about the concept of distance. While the ED expresses how far apart an unknown spectrum and the centre of a class are, the MD takes into account the class dispersion and expresses how far apart an unknown spectrum and a class region

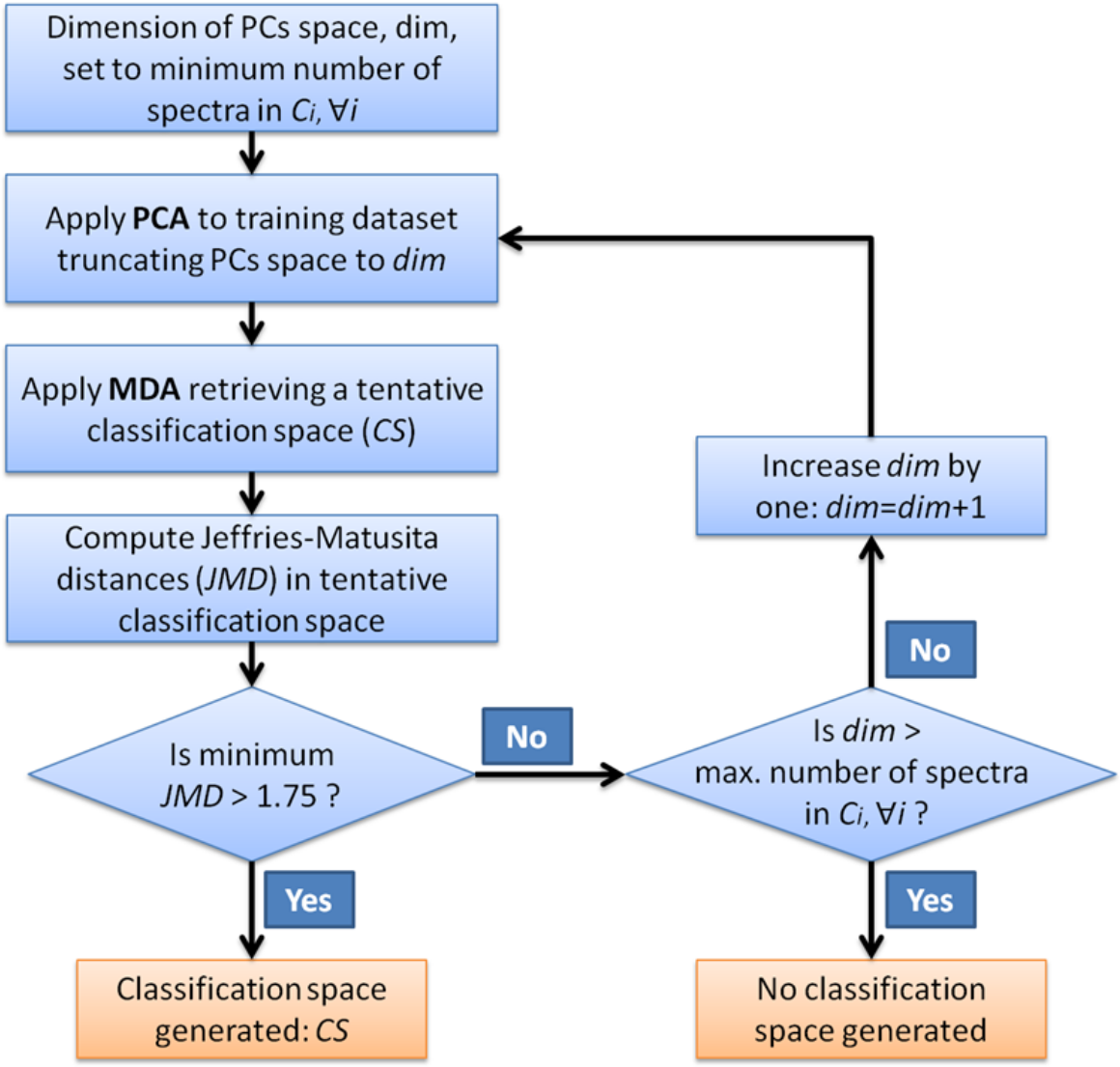


Figure 4.15: Classification space generation from a training dataset

are. Bearing in mind these meanings we define a classification distance combining the results of these two different distances, specifically:

$$CD_i(x) = ED(x, \mu_i)MD(x, C_i) \tag{4.12}$$

In order to classify an unknown spectrum we define a toolkit based on the above distance which allows to explore the matching of the unknown to a class. To do so, we firstly compute the classification distance (CD) between classes, the so-called *InterClassCD*, which provides a notion on how close the classes are. Note that there are as many values of *InterClassCD* as defined classes and the minimum value is due to the closest classes. Also, we calculate the so-called *IntraClassCD*, which provides an idea on how close a spectrum is to its own class. The farthest spectrum from a given class gives the maximum value of *IntraClassCD* for that class. Then, the classification of an unknown spectrum is performed by exploring the classification distance between

the unknown and each class. The assignment of the unknown spectrum to a reference class is performed by a matching function defined as:

$$MF_i(x) = \begin{cases} 1 & if\ CD_i(x) \leq maxIntraClassCD_i \\ 1 - \dfrac{CD_i(x)}{minInterClassCD} & if\ CD_i(x) \leq minInterClassCD \\ 0 & otherwise \end{cases}$$

where $CD_i(x)$ is the minimum value of CD between the unknown (x) and the reference classes, *minInterClassCD* is the minimum value of *InterClassCD* and the value $maxIntraClassCD_i$ the maximum value of *IntraClassCD* for the $i-th$ class. The matching function expressed in % is intended to help the analyst in the decision-making process. The methodological scheme of the classification system is shown summarized in Fig. 4.16. It illustrates the standard classification design together with the approach proposed in this research.

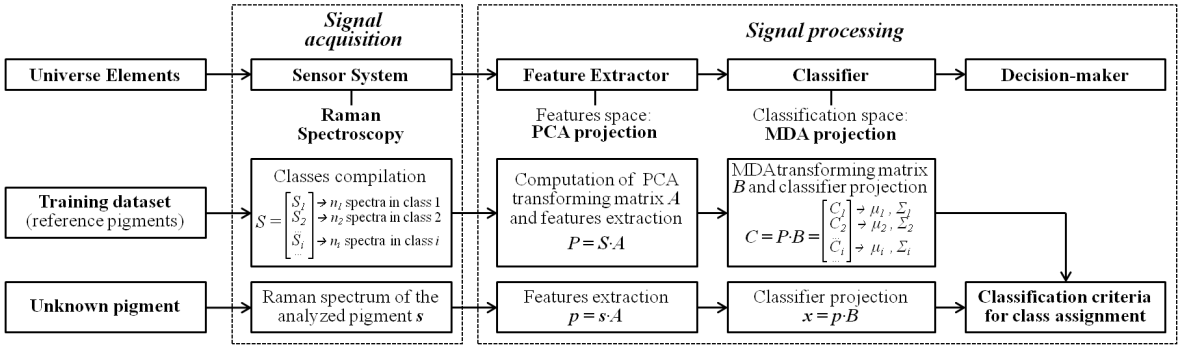


Figure 4.16: Overview of the supervised classification scheme

Results and discussion

Sect. C.2.2 of Appendix C compiles the verification and validation activities performed on the developed supervised classification methodology. Specifically, the performance of the methodology was diagnosed in an under-controlled environment using simulated data providing successful results. In order to show the performance of the implemented methodology in experimental environments the developed classification system was applied to unknown Raman spectra acquired from oil paintings[150] and art works. The spectra for the reference classes were acquired from reference pigment powders. The experimental spectra used in this research that were measured by the author were recorded using the portable Raman equipment as described in Sect. 3.2.

No assumptions regarding the input data are made by the classification system, which processes the data blindly through the presented automatic approach in a fully

transparent way. An experimental example is reported hereafter. In this case, we distinguish among ultramarine blue pigment in its natural form (as lapis lazuli) and in its synthetic form. Therefore, two reference classes were built. The natural form class was composed of six spectra acquired from Afghan, Siberian and Chilean lapis lazuli samples. The synthetic form class was composed of six spectra as well, which were acquired from several synthetic ultramarine blue pigment powders manufactured by Nubiola. The Feature Extraction module provided a 6-dimensional PCs space with an accumulative variance of 99.54% (see Fig. 4.17). The classification space is described by a straight line with two separated regions (one for each class) with a JMD equals to 2. The classification methodology was applied to twelve unknown spectra measured in our laboratory from different art works (see Fig. 4.18). Specifically, one of these unknown spectra was acquired from a Chilean art figure (see Fig. 4.19) whilst the remaining unknown spectra were measured from different oil paintings. Fig. 4.20 shows the projection of the unknown spectra onto the classification space. The classification results are reported in Table 4.1. The consistency of the results was assessed by inspection of the measured areas using a Leica MZ-12 stereomicroscope with a photomicrographic resolution of 600 magnifications. All the unknown spectra were successfully classified although with different value of the matching function, ranging from 47.32% to 100%. Specifically, the spectrum classified with the minimum matching value (painting 7 in Fig. 4.18) was deeply affected by undesired artifacts (unknown peaks) in the pre-processed spectrum.

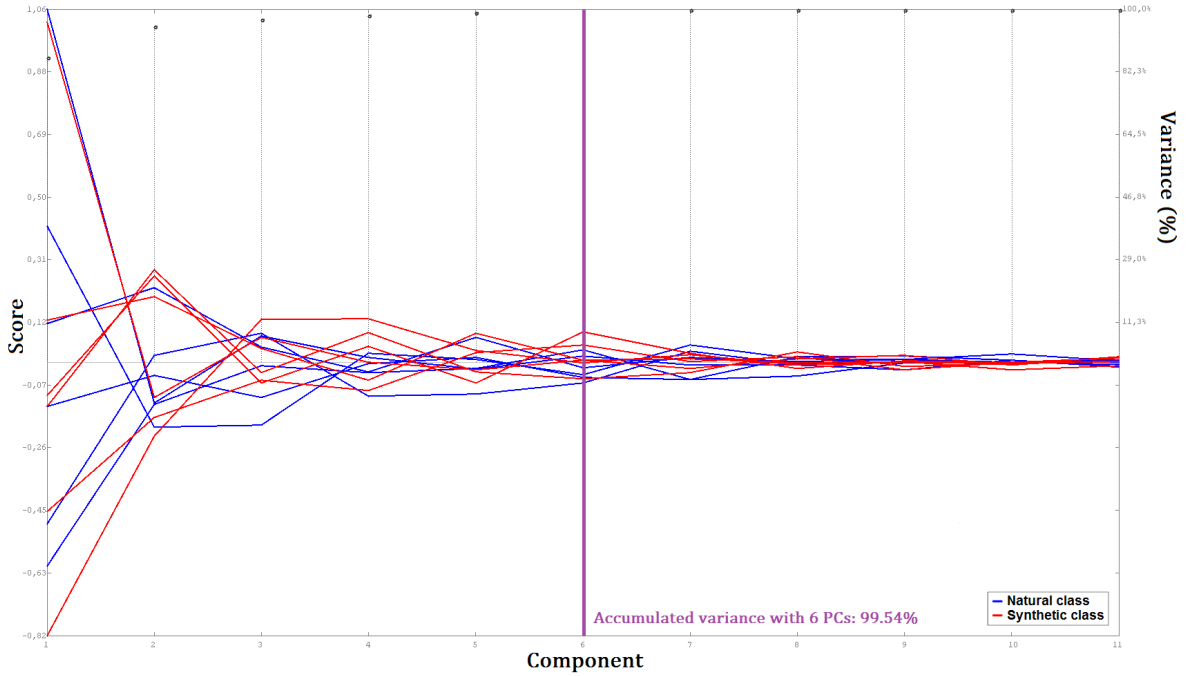


Figure 4.17: Scores of the training set and accumulated variance of PCA projection as a function of PC. The 6-dimensional PCs space accounts for an accumulative variance of 99.54%

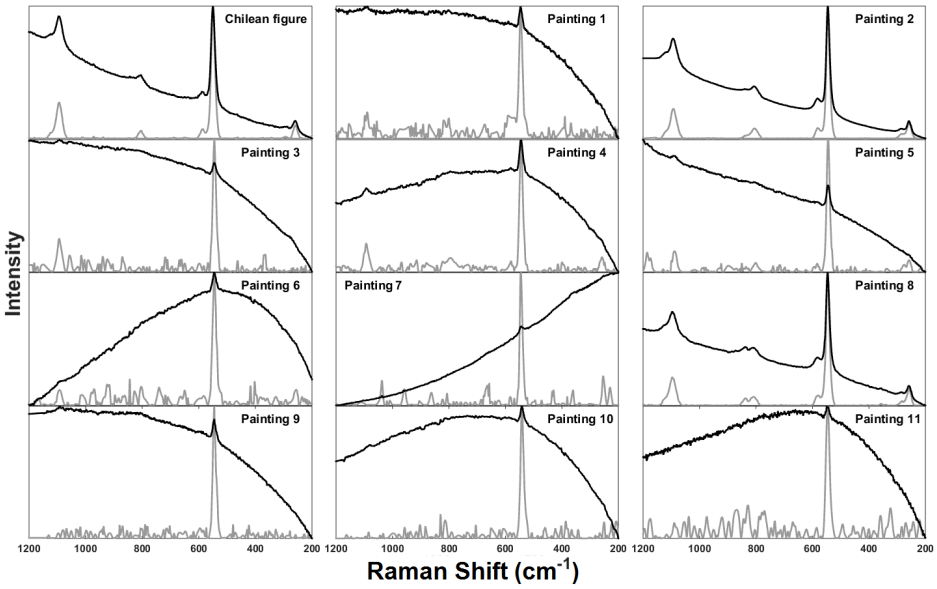


Figure 4.18: Experimental Raman spectra from ultramarine blue measured on a Chilean art figure and oil paintings: acquired spectra (black) and pre-processed spectra (gray)

Figure 4.19: Chilean art figure, expected to be manufactured from lapis lazuli (natural form of ultramarine blue pigment)

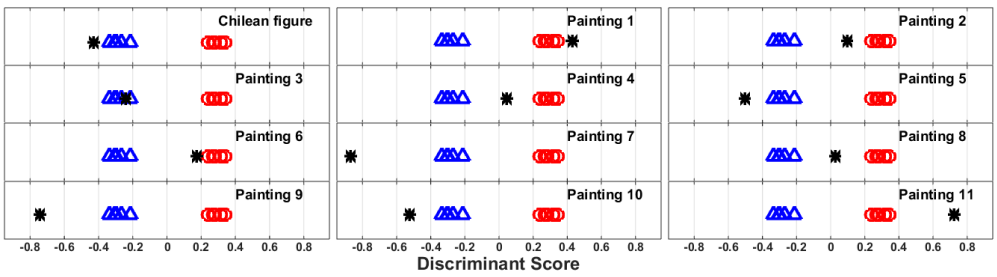


Figure 4.20: Projection of experimental Raman spectra from ultramarine blue onto the classification space: natural form class (blue triangles), synthetic form class (red circles) and unknowns (black asterisks)

Table 4.1: Classification of Raman spectra from ultramarine blue pigments

Artwork	**Expected Class**	**Assigned Class**	**MF (%)**
Chilean art figure	Natural	Natural	96.90
Painting 1	Synthetic	Synthetic	95.80
Painting 2	Synthetic	Synthetic	93.44
Painting 3	Natural	Natural	100.00
Painting 4	Synthetic	Synthetic	93.32
Painting 5	Natural	Natural	93.63
Painting 6	Synthetic	Synthetic	97.76
Painting 7	Natural	Natural	47.32
Painting 8	Synthetic	Synthetic	82.29
Painting 9	Natural	Natural	67.43
Painting 10	Natural	Natural	91.08
Painting 11	Synthetic	Synthetic	62.59

4.5 Chapter summary

A generalised methodology to automatically identify Raman spectra was presented. The method is able to identify single- and multi- component spectra from a single spectral observation, with no user input or previous knowledge of the analysed sample. The implemented algorithm is based on the automated matching of spectra using PCA and ICA, and it is computationally efficient and conceptually simple. However, it must be pointed out that the spectra need to be pre-processed to enhance the Raman information for the proposed method to work successfully.

Mixtures are handled through an iterative strategy based on ICA, which allows the components separation with high accuracy and no parameters to be configured. This strategy demonstrated to work successfully even when dealing with mixed spectra of different intensities. However, the separation of components with overlapping fundamental bands may cause some information loss significant for the identification of an overlapped minor component.

The system delivers fully automated identification, qualifying the result with a Matching Factor that is intended to help the judgement of the identification. Simulated spectra were used to assess the proper performance of the identification methodology. Moreover, several hand-made samples from mixed pigments were measured in order to

evaluate the proposed method and it was applied to real-case spectra from paintings as well. According to the consistency of the results, the system has great deliver an accurate and practical method for automated identification of Raman spectra, not only in pigment analysis, but essentially any material group.

Additionally, a methodology to automatically distinguish between Raman spectra showing small differences was presented. According to predefined reference classes, the method is able to classify unknown spectra from a single spectral observation, with no user input or previous knowledge of the analysed sample. The developed model is based on automated matching of unclassified spectra using PCA and MDA. The results showed that the method is suitable for art works analysis as it successfully classified the analysed Raman spectra in a consistent way. Moreover, the implemented method is an easy-to-use system and it is straightforward to update when new spectral data become available.

The implemented classification system has been applied to experimental Raman spectra, and the obtained results showed that it may play a good auxiliary role in the analysts' endpoint classification. Therefore, the system may become a useful tool to help in the decision-making process, in order to ease the management of pigment classification from Raman spectra whose reference classes are very similar.

Finally, it is worth noting that the methodologies make no assumptions with respect to the input data, applying a blind treatment of the Raman spectra and processing them in a transparent way regardless of the identification or classification purposes. Consequently, it is perfectly capable of dealing with spectra from different sources, i.e. recorded with different acquisition systems and measurement conditions. This fact may represent a significant advantage of the presented automated system in the applications of pigment identification and classification in art analysis through Raman spectroscopy, as it is independent of the measurement system and the configuration used for the acquisition of Raman spectra.

Chapter 5

Global system of automated interpretation of spectra in art analysis

5.1 Chapter overview

This chapter describes the global software platform developed for the automated interpretation of spectra from pigments, which integrates the automated methodologies described in Chapter 3 and Chapter 4. Specifically, Sect. 5.2 introduces the basics of data interpretation and describes the software system developed in this research, which is intended to provide insights from the raw spectra in order to help the spectroscopists and art analysts in the decision-making process. Besides, it provides an overview of the software platform development, including requirements specification, architectural pattern, data model definitions, among other software development tools. Finally, Sect. 5.4 shows an example of a use case devoted to the automatic interpretation of experimental spectra from an art work analysed through Raman mapping.

5.2 PigmentsLab: from raw spectra to insight into pigments

Data interpretation can be seen as a simple linear process, which includes five distinct steps that depend on each other: acquire, prepare, analyse, report and act (see Fig. 5.1). Indeed, data interpretation is an iterative process and findings from one step may require the previous step to be repeated with new information. Specifically:

- First of all, we need to obtain the source material before analysing it or acting on it. Thus, the first step in data interpretation is data acquisition. Acquire

includes all the processes devoted to retrieve data including finding, accessing, measuring and recording data.

- After getting the data, the next step is explore it. Exploring data is a part of the two-step data preparation process namely explore data and pre-process data through understanding the nature of the data (its quality and format) and pre-process the data, which includes cleaning or filtering data, modelling raw data into a more defined data model, packaging it using a specific data format, and integration of multiple data streams.

- The prepared data then may be ready for a subsequent analysis, which involves the selection of the analytical techniques to use, building a model of the data, and analysing the results. This step can take one or several iterations on its own or might require to go back to steps one and two to get more data or package data in a different way.

- The following step is reporting the insights gained from the analysis, which includes an evaluation of the analytical results presenting them in a visual way, creating reports that include an assessment of results with respect to success criteria, and making decisions for what actions should follow.

- The last step is acting, i.e. turning insights into action based on the purpose initially defined for instance to answer questions or for improving business processes.

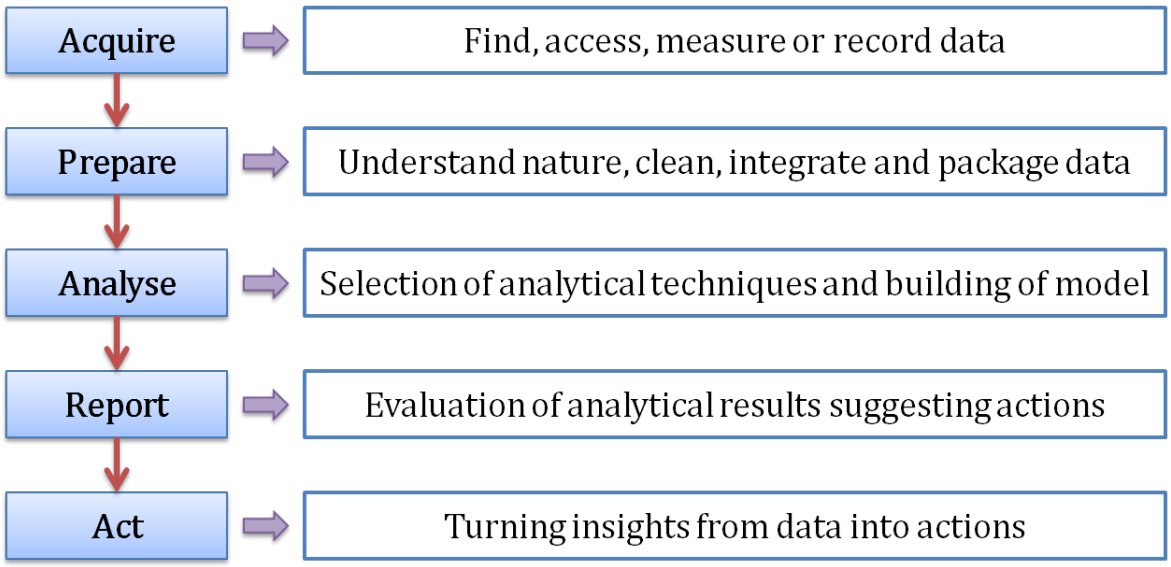


Figure 5.1: Schematic overview of the data interpretation process, which includes five steps: acquire, prepare, analyse, report and act

Nowadays, the multi-disciplinary scientific community devoted to the analysis and preservation of cultural heritage is immersed in the need of handling and interpreting an overwhelming amount of data. Indeed, the data volumes cultural heritage experts deal with is exponentially rising due to the emergence of the spectroscopic techniques. A clear example of the complex data handling processes encountered by scientists involved in the protection of cultural heritage roots in the multi-analytical approaches taken to the study of the constituent pigmentation in art works. This kind of studies deal with the characterization of pigments by several spectroscopies (as introduced in Sect. 2.2.1), mainly X-Ray Fluorescence (XRF)[151], X-Ray Diffraction (XRD)[152], Laser-Induced Breakdown Spectroscopy (LIBS)[153], InfraRed Spectroscopy (IR)[154], Raman spectroscopy and Surface-Enhanced Raman Spectroscopy (SERS)[155]. Thus, the spectral measurements of a single point in the art object under analysis provide a noteworthy volume of data that needs to be stored and properly processed and combined in order to be appropriately interpreted (see Fig. 5.2).

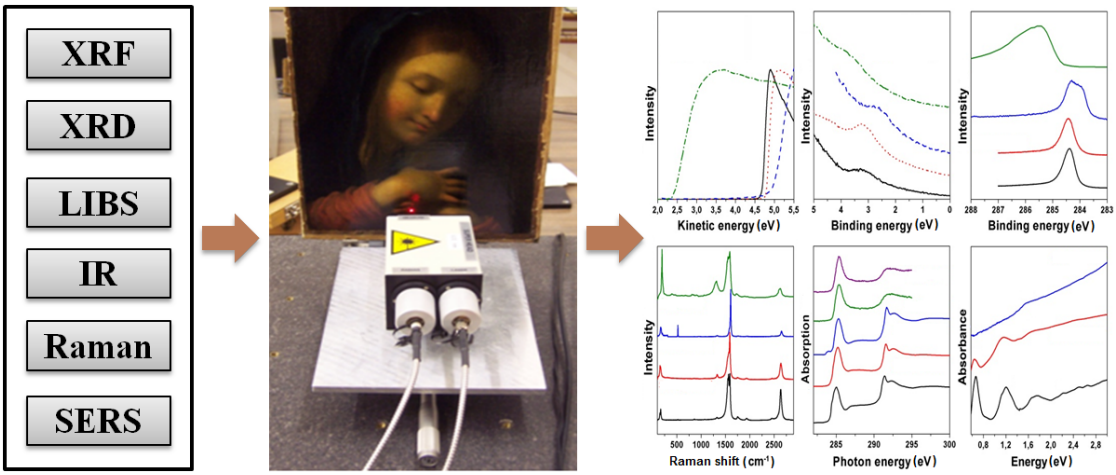


Figure 5.2: Example of data retrieval through multi-spectral characterization of the pigmentation in a given spot of an art work

As can be seen, the handling and treatment of a big volume of data is now relevant for the scientists working in the cultural heritage field, and therefore consuming sophisticated data software products and services can no longer be ignored by this scientific community. Actually, the means to extract insight from spectroscopic data are noticeably important as can be retrieved from the fact that the development of increasingly sophisticated techniques in the spectral data processing has already become a hot topic of research. In this sense, there is a wide variety of techniques that can be used to aggregate, manipulate, analyse, and visualise this sort of datasets. As presented in previous chapters, these techniques generally draw on disciplines such as statistics and computer science such as machine learning, data mining, pattern classification, cluster analysis, data fusion, signal processing, and pattern recognition among many others. In fact, cultural heritage researchers continue to develop new methodologies and im-

prove on existing ones, predominantly in response to the need to analyse innovative combinations of data. Presenting the spectral information in such a way that cultural heritage analysts can consume it effectively is a key challenge that needs to be met if analysing spectral data is to lead to concrete actions in the field of conservation and/or restoration of art works. For this reason, there is currently a remarkable amount of research and innovation in the field of visualisation, i.e. techniques and technologies used for creating images, diagrams, or animations to communicate, understand, and improve the results from data obtained through spectral analysis (see Fig. 5.3).

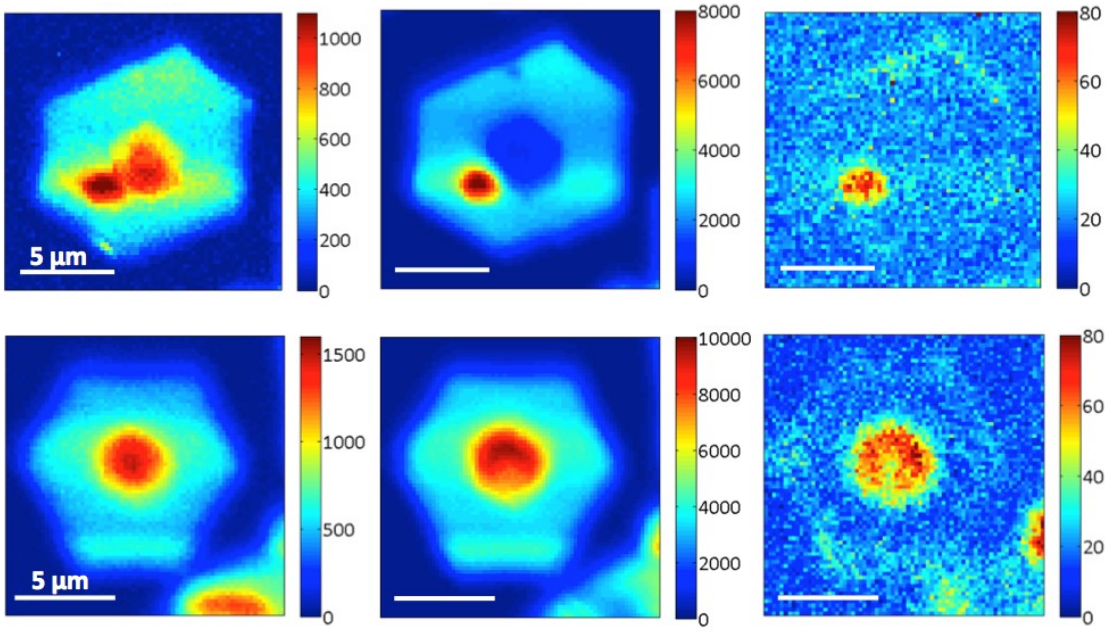


Figure 5.3: Example of measurements visualisation through spectral mapping

Spectral mapping has demonstrated to be a useful tool to gain insight into the composition of art works' multilayered structures by analysing tiny stratigraphic samples removed from the objects[156,157]. When sampling is limited or not possible at all, spectral imaging may still deliver valuable information on the composition and distribution of the pigments present in painted surfaces. In this way, the surface of an art work may be inspected to identify, for instance, discontinuities or alterations in the pigments application, which may give an indication of loss or change in material or overpaintings. Furthermore, restaurateurs often use different materials than the originals in order to eventually differentiate the original material from what was restored. Raman imaging or Raman mapping is a powerful technique which allows, among other things, the molecular structure detected through Raman spectroscopy to be visualised as false-colour images. Therefore, Raman identification mapping is a suitable method as it is aimed to generate detailed identification images based on the Raman spectra of the analysed sample. A complete spectrum is acquired at each and every pixel of

the image, and then interrogated to generate false-colour images based on the identification of the material composition and structure of the analysed area. Conclusively, an exhaustive and objective knowledge about the different materials used in art works is undoubtedly essential. Image data in the form of spectra dominate data storage volumes in spectral imaging. While a single file containing one spectral measurement can total a few kilobytes, a single spectral image can require thousands of megabytes or more to store. Additionally, in a multi-technique imaging approach, a combination of several spectral images coming from the different spectroscopies commented above may be needed, resulting in a large data volume of tens of gigabytes.

As commented previously, the identification of pigments used in cultural heritage is indispensable to determine correct conservation strategies, to study degradation processes and to answer authenticity-related questions. Once the spectroscopic measurements are taken from an art work, the main task of the analysts is to extract the information from the data sets in order to properly interpret the measurements. In this sense, one of the most crucial interpretations to be performed from the spectral data is based on their identification. The spectral identification is generally carried out by a visual comparison between the unknown spectra with an appropriate database of reference spectra. Since the 20th century, the introduction of synthetic organic pigments has enormously increased the number of available pigments. Therefore, a complete database of reference spectra is needed in order to ensure a correct identification. Although several reference spectra from pigments have been published, they are usually devoted to certain chemical classes. As new publications become available, the handling of spectral data by the analysts and spectroscopists is becoming more complex. Therefore, the need of a common platform holding an extensible reference database of spectra from pigments is of practical interest. The platform proposed herein attempts to create a reference database of the available pigments integrating at the same time the corresponding art history information and available spectra from the spectroscopies generally used in art works analysis.

Consequently, in this chapter we present the design, development and implementation of an integrated software platform aimed to the storage, processing, analysis and visualization of data coming from elemental and molecular spectroscopies applied to the study of artists' materials. The main aim of the system is to establish a reference framework of objective algorithmic solutions and high-performance visualisation technologies to aid in the interpretation of spectra. This global system is intended to provide insight from the raw spectral measurements to help the spectroscopists and art analysts in the decision-making process.

The implemented software platform, the so-called ***PigmentsLab***, is made up of

the following modules:

- **Database Explorer**: an extended, detailed and robust SQL-based database management system of reference pigments containing art historical information and integrating different spectral data from the most commonly used spectroscopies in the field of art analysis. This module provides several tools for exploring and querying the database as well as for showing the pigments information along with their spectral information.
- **Spectral Viewer**: an interactive application aimed at viewing spectroscopic measurements from art materials, implementing also elemental data analysis such as noise filtering or bands localisation to allow the analyst to quickly analyse the spectra in a visual way.
- **Virtual Spectroscopist**: an advanced application for making breakthroughs regarding the interpretation of spectra from art materials. The system provides the spectral characterization with no prior knowledge of the composition of the analysed sample.

A schematic overview of the ***PigmentsLab*** platform can be seen in Fig. 5.4.

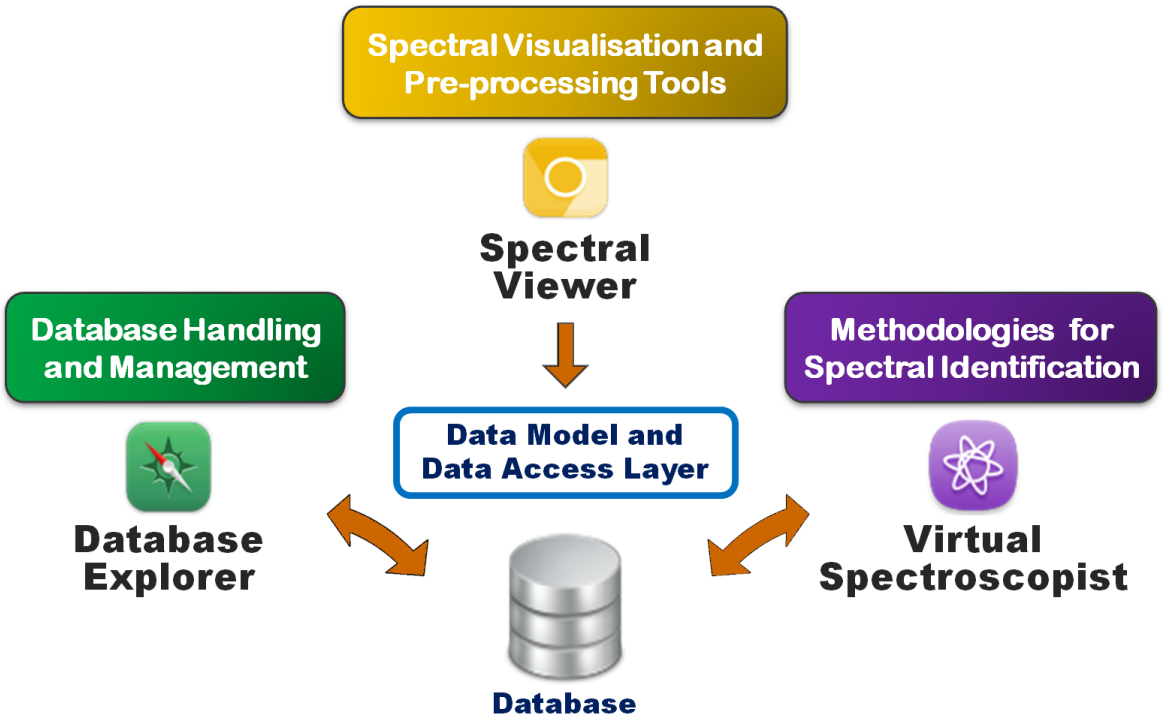


Figure 5.4: Schematic overview of ***PigmentsLab***: a three-module platform integrating art-historical and spectroscopic data from art materials as well as high-performance spectral visualisation and pre-processing technologies, database handling and management tools, and automated solutions to aid in the interpretation of spectra

5.2.1 Software platform development

An overall software development description of ***PigmentsLab*** is summarised hereafter, providing requirements specification, an architectural description, development tools among other software definitions such as:

- The architecture design, using information flowing characteristics, mapping them into the program structure
- The data design, describing structures that reside within the software, i.e. attributes and relationships between data objects
- The interface and procedural design, describing internal and external program interfaces, as well as the design of human interface.

Concretely, the development of the software platform follows the standard Waterfall model for software development (see Fig. 5.5), which is a sequential software development approach in which progress is seen as flowing steadily downwards (like a cascading waterfall) through the phases of analysis, design, implementation, testing and maintenance[158].

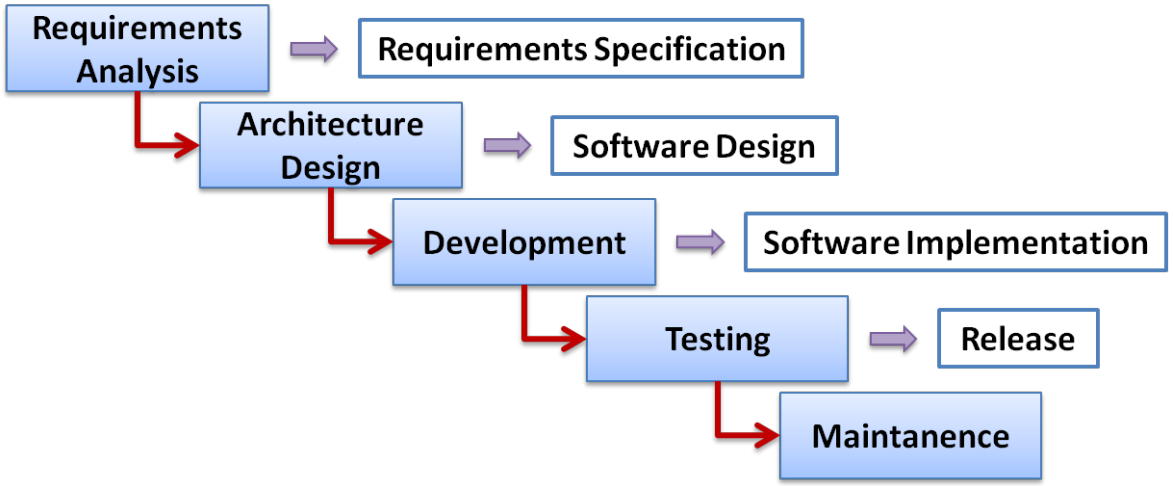


Figure 5.5: Waterfall model schematic diagram in which progress flows from the top to bottom

Software requirements specification A software system to be developed is described through a software requirements specification, which is a collection of requirements for a particular software product that performs certain functions in a specific environment and the criteria for determining whether those requirements are met[159]. One of the main software requirements is the programming language. The selected programming language for developing the ***PigmentsLab*** software platform is Java. Java is a general-purpose computer programming language that is concurrent, class-based,

object-oriented, and specifically designed to have as few implementation dependencies as possible. This programming language is intended to let application developers Write Once, Run Anywhere (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation.

According to International Standard ISO/IEC/IEEE 29148:2011 (Systems and software engineering - Life cycle processes - Requirements engineering), the benefits of documenting the software requirements include:

- It forces a rigorous assessment of requirements before design can begin and minimizes later redesign
- It provides a realistic basis for estimating product costs, risks, schedules and enhancements
- It provides an informed basis for deploying a product to new users or new operational environments

Appendix D (*Software Requirements Specification of PigmentsLab*) enlists the system requirements that are needed for the development of the software platform proposed in this chapter for the processing of spectroscopic data applied to the analysis of artists' materials.

Software architecture The structure of a computing system is known as *software architecture*, which comprise software elements, the externally visible properties of those elements, and the relationships among them[160]. Formally, the architecture is a reference frame in which competing interests may be presented, discussing requirements with users, and constraining the software implementation. In this sense, the architecture dictates organizational structure for development and maintenance activities. Also, it allows the achievement of a system's desired quality attributes such as performance, modifiability or usability. Consequently, the importance of the architecture for a project development such as user interfaces roots in the fact that it is a transferable and reusable abstraction of a computing system.

User interfaces are especially prone to change requests[161]. New functionalities may be added to an application, or existing ones may need to be extended upon user requests, or ported to a different platform. All these updates may imply code changes. Consequently, support for several user interface paradigms should be incorporated from the software design stage for developing a flexible software system. In this sense, changes to the user interface should be easy, and even possible at run-time: supporting different 'look and feel' standards or porting the user interface should not affect code in the core of the application. There exist different software architectures for the case of interactive applications with a flexible human-computer interface. Among them the

Model-View-Controller architectural pattern (Model-View-Controller (MVC)) stands out[162]. This software architecture divides an interactive application into three components: the model, the view and the controller, dividing the internal representations of information from the ways that information is presented to or accepted from the user:

- The **model** encapsulates core data and functional core of the application. It is independent of specific output representations or input behaviour. It encapsulates the appropriate data, and exports procedures that perform application-specific processing
- The **view** displays information to the user. A view obtains the data from the model. There can be multiple views of the model. Each view has an associated controller component
- The **controller** handles user input. In general, it receives input usually as events that encode mouse movement, activation of mouse buttons, or keyboard input. Events are translated to service requests for the model or the view. The user interacts with the system solely through controllers

Views and controllers together comprise the user interface. A change-propagation mechanism ensures consistency between the user interface and the model, maintaining a registry of the dependent components within the model. All views and also selected controllers register their need to be informed about changes. Changes to the state of the model trigger the change-propagation mechanism. The change-propagation mechanism is the only link between the model and the views and controllers. All the Graphical User Interfaces (GUIs) implemented as part of the ***PigmentsLab*** software platform are MVC-based. The general Java-based MVC scheme is represented in Fig. 5.6.

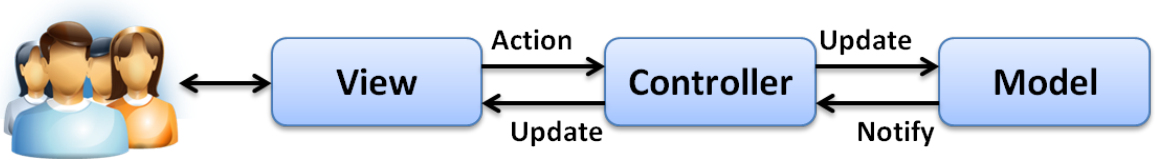


Figure 5.6: Java-based schematic diagram of the Model-View-Controller architectural pattern

Software development tools The platform implementation follows the standard Continuous Integration (CI) software development workflow[163]. Generally speaking, CI is a development practice that requires developers to integrate code into a shared repository with a relatively high frequency. Each *check-in* is then verified by an automated build, allowing to detect problems early. Indeed, by integrating regularly, errors

may be detected quickly, and therefore may be locate more easily. Concretely, the CI implementation for the development of ***PigmentsLab*** is represented in Fig. 5.7.

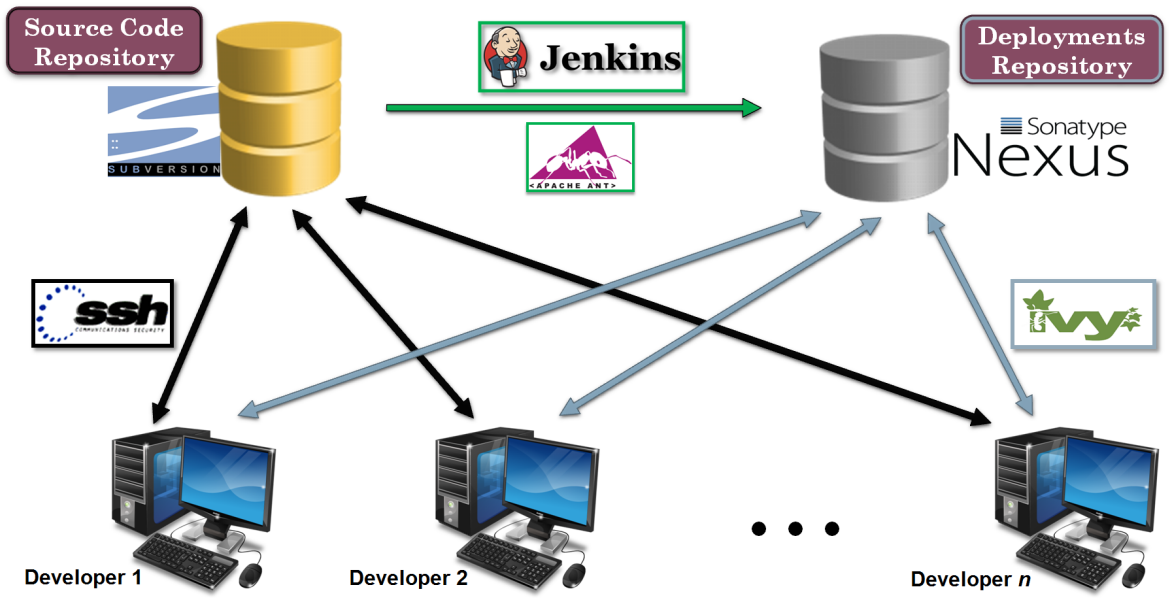


Figure 5.7: Overview of the ***PigmentsLab*** software platform development workflow, based on Continuous Integration (CI). The application of several standard software development tools is schematically shown, including mainly Apache SubVersioN (SVN) for revision control, Apache Ant for software build, Nexus Repository OSS for software deployments management, Apache Ivy for dependencies management, and Jenkins as integration tool

Specifically, the following standard software development tools are used:

- **Apache SubVersioN (SVN)**[164]: is a centralized software versioning and revision control system distributed as open source under the Apache License. Software developers use SVN to maintain current and historical versions of files such as source code, web pages, and documentation.

- **Apache Ant**[165]: is a software tool for automating software build processes, best suited to building Java projects.

- **Sonatype Nexus Repository**[166]: is a software deployment manager in charge of handling a storage repository location from which software packages may be retrieved and installed on a computer. It allows to organize, store, and distribute software components and supports all popular component formats and is always-on for continuous delivery and deployment.

- **Apache Ivy**[167]: is a dependency manager used primarily for Java projects. An Ivy eXtensible Markup Language (XML) file describes the software dependencies on other external modules and components. Then, dynamically, it downloads the project dependencies from the software repository, and stores them in a local

cache. This local cache of downloaded artifacts can also be updated with artifacts created by local projects.

- **Jenkins**[168]: is an open source automation server that helps to automate the non-human part of the whole CI development process. It is a server-based system running in a servlet container such as Apache Tomcat, supports version control tools such as SVN, can execute Apache Ant and allows automated repository deployment. Builds can be automatically triggered by various means, for instance by a commit in a version control system, ideal for CI software development.

Server specification The above explained software development tools require a computer device in order to provide the proper functionalities for developing the ***PigmentsLab*** software platform. Hence, these tools are integrated in a specific server for software development. The server is physically located in the UPC premises with restricted access and runs the software development tools described above. The server specifications are:

- Operating System (OS): SMP Debian 3.16.39 x86_64 GNU/Linux
- RAM: 2GB, HDD: 25GB, Processors: 1 Intel(R) QEMU Virtual CPU version 2.3.0 2.20GHz

To fulfil the requirements of the above commented software development tools, LAMP (Linux, Apache, Open-source relational database management system based on SQL (MySQL), Hypertext Preprocessor (PHP)) was adopted as a development standard in the server infrastructure. It uses *Linux* OS as the base layer, *Apache HTTP Server* as the web daemon sitting on top the OS, MySQL database to store all the information served by the web daemon, and PHP to drive and display the data.

Data model definition and data access layer With the aim of processing the spectra, a proper modelling of the raw data into a more specific model was defined, packaging the data by means of a specific data format. This abstract modelling allows to organize elements of data and standardizes how they relate to one another, based on the ability of a computer to fetch and store data at any place in its memory that can be itself stored in memory and manipulated by a software system. The implementation of a data model is generally represented by means of a Unified Modelling Language (UML) diagram and usually requires developing a set of procedures in order to create and manipulate instances of that model. This is generally known as data access layer. Within the ***PigmentsLab*** framework, the data model definition and data access layer are implemented following software development standards, providing routines for the spectroscopic measurements to be accessed, stored and processed in order to be appropriately analysed. In this sense, both the data model definition and data access

layer are compiled in a common library used throughout the ***PigmentsLab*** platform, namely *SpectralTools*. Fig. 5.8 shows a UML diagram, which provides a visualization of the data model design for the object-oriented-based definition of spectra. In particular, Table 5.1 shows the attributes description of the *RamanSpectrum* object as defined in *SpectralTools*. Additionally, Fig. 5.9 presents a UML diagram which schematically describes the data access layer implementation for the *RamanSpectrum* object handling implemented in *SpectralTools* - the rest of objects defined in the data model (*XrfSpectrum*, *XrdSpectrum*, *LibsSpectrum*, *IrSpectrum* and *SersSpectrum*) are accessed and handled similarly. Hence, database access is performed through Java Database Connectivity (JDBC), a standard application programming interface for Java. All objects defined in the data model may be stored to files in a file format specifically developed in this research, the so-called Spectral Data Format (SDF), which corresponds to the Java-objects serialisation. Standard formats used in spectroscopy such as SPectroscopiC format (SPC) are properly handled as well by the data access layer of *SpectralTools*.

Table 5.1: Attributes description of the *RamanSpectrum* object defined within the ***PigmentsLab*** data model

Attribute Name	**Description**
`id`	Unique identifier which identifies the pigment
`name`	The name which identifies the pigment
`type`	The type of the pigment: Inorganic or Organic
`typeCode`	The type code of the pigment: Inorganic types=1, organic types=2
`ciGroup`	The Colour Index group of the pigment: PB, PBk, PBr, PG, PO, PR, PV, PW, PY
`ciNumber`	The Colour Index number of the pigment
`ciSubNumber`	The Colour Index sub-number of the pigment
`chemicalName`	The chemical name of the pigment
`formulae`	The chemical formulae of the pigment
`description`	The description of the pigment
`artists`	The artists that used the pigment
`usage`	The usage of the pigment
`inUse`	Flag indicating whether this pigment is still in use or not
`xSamples`	Spectral domain, Raman shifts $[cm^{-1}]$ in the Raman case for instance
`ySamples`	Spectral intensities $[a.u.]$
`bands`	The list of spectral bands of the spectrum
`excitationSource`	The excitation source used for acquiring the spectrum [nm]
`sourcePower`	The source power used for acquiring the spectrum [mW]
`acquisitionTime`	The acquisition time used for acquiring the spectrum [s]
`accumulations`	The number of accumulations used for acquiring the spectrum

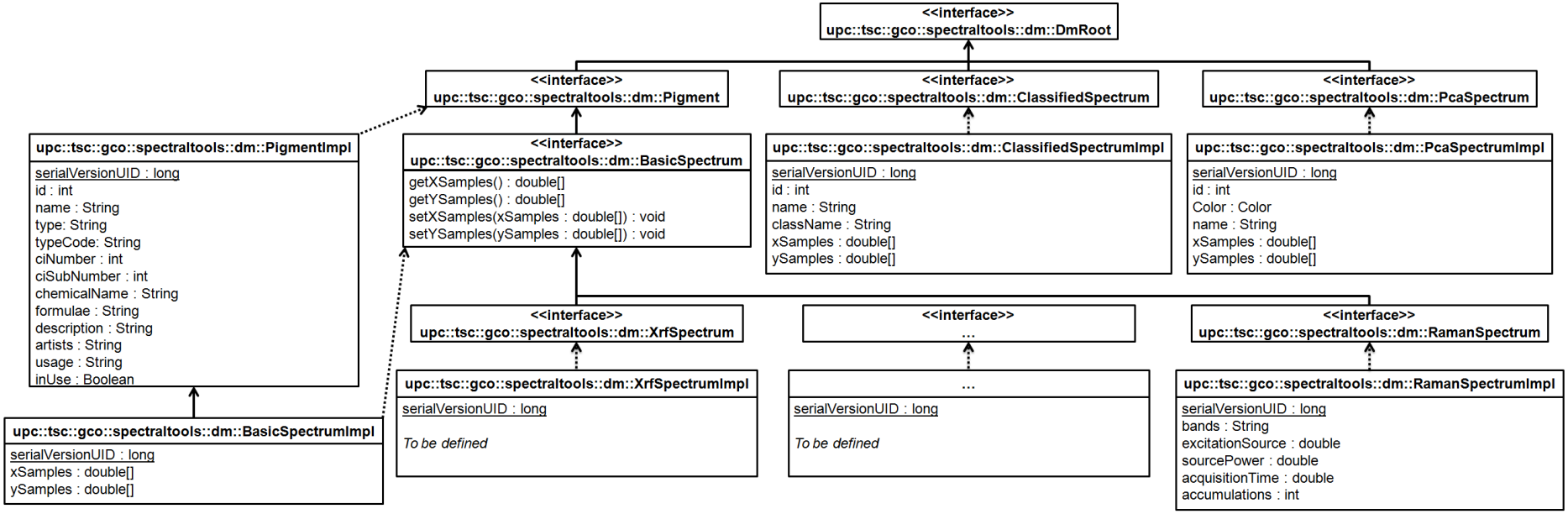


Figure 5.8: UML diagram depicting the main design of the data model implementation for the object-oriented-based definition of spectra

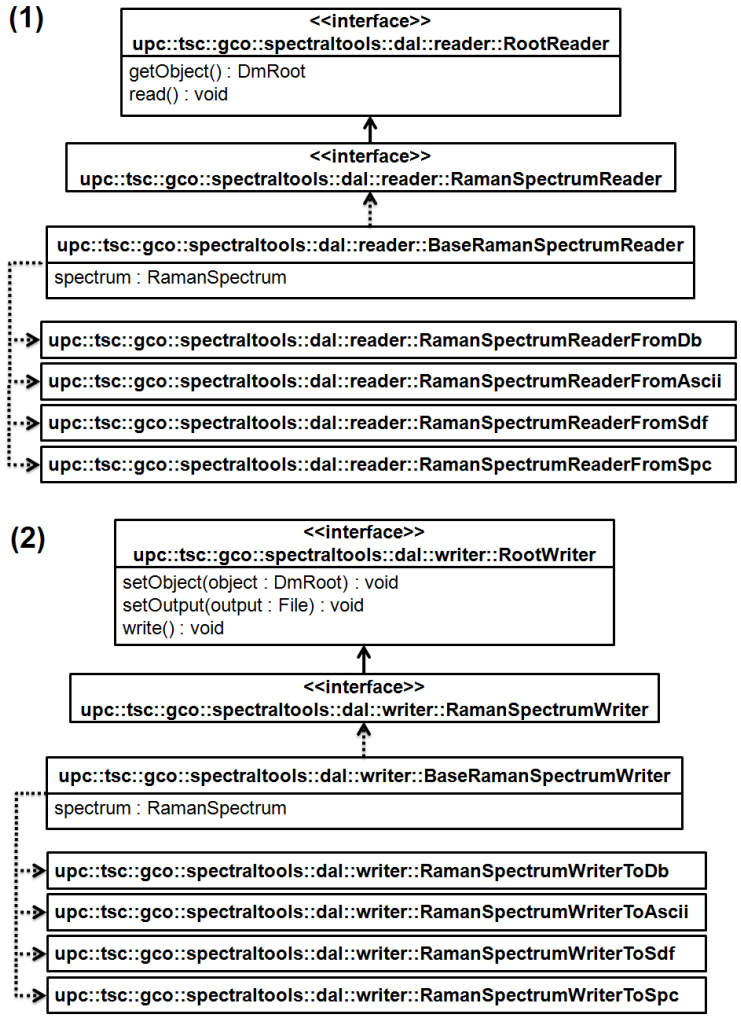


Figure 5.9: UML diagrams outlining the data access layer for the *RamanSpectrum* object handling within the ***PigmentsLab*** framework: objects devoted to data reading (1) and objects dedicated to data writing (2)

5.3 PigmentLabs: modules overview

This section provides an overview of the three main modules that conform the ***PigmentsLab*** software platform as schematically outlined in Fig. 5.4: *Database Explorer* (for reference spectral database browsing and management), *Spectral Viewer* (for spectral visualisation and pre-processing tools) and *Virtual Spectroscopist* (implementing automated methodologies to help in the spectral interpretation). In particular, a quick description of the GUIs implementing each module is presented hereafter, focusing on the main interactions between target users (spectroscopists or data analysts) and data processing (data handling and automated methodologies described previously).

5.3.1 Database handling

The ***PigmentsLab***-module devoted to database handling is called *Database Explorer*. This module implements a database management system developed to aid spectroscopists and art analysts to quickly retrieve information from the reference pigments present in the database, supporting data exploration, Structured Query Language (SQL)-based querying and database management. The main view of *Database Explorer* is presented in Fig. 5.10.

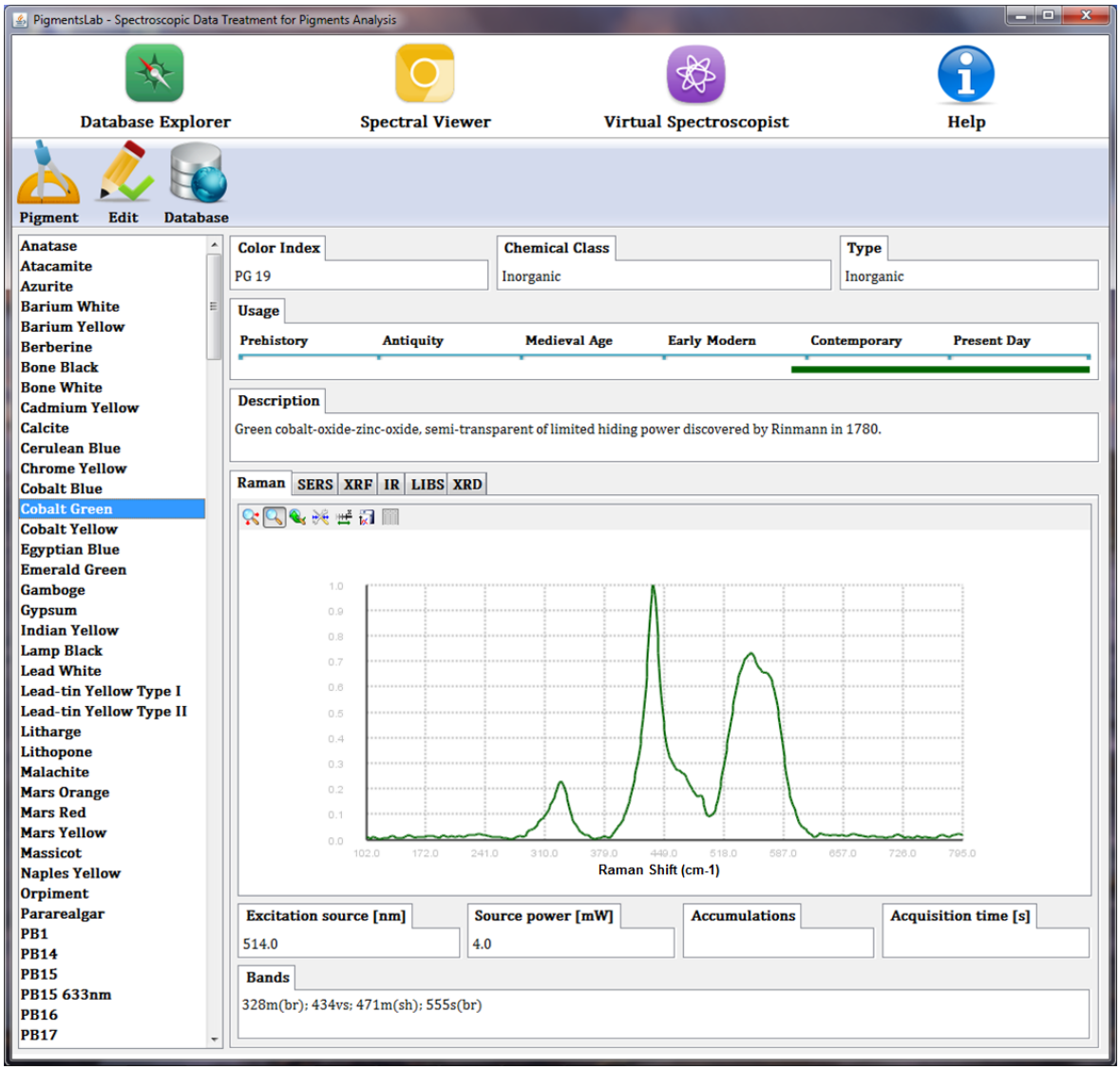


Figure 5.10: Main view of *Database Explorer*

In particular, it provides a `List Box`, which lists the reference pigments present in the database. When a reference pigment is selected from the `List Box`, the system accesses the database contents for the selected item through JDBC, maps its information into the *SpectralTools* data model and displays it to the user through the several `Text Boxes` implemented in the view, mainly:

- Art historic information:
  - *Colour Index*: group (PB, PBk, PBr, PG, PO, PR, PV, PW, PY) and index
  - Type (inorganic/organic)
  - Chemical class (according to description in Sect. B.2 of Appendix B)
  - General description
- Spectroscopic information. A `Tabbed Pane` allows to select the spectroscopic technique to be used: Raman, XRF, XRD, IR, LIBS or SERS. In the Raman case, for instance:
  - Plot of the spectrum in a `2D Canvas`
  - Excitation source [nm] and source power [mW]
  - Acquisition time [s] and accumulations

The information corresponding to the pigment usage is graphically displayed through a chronological panel which includes the main historical periods (see Fig. 5.11): prehistory, antiquity, medieval age, early modern, contemporary and present day.

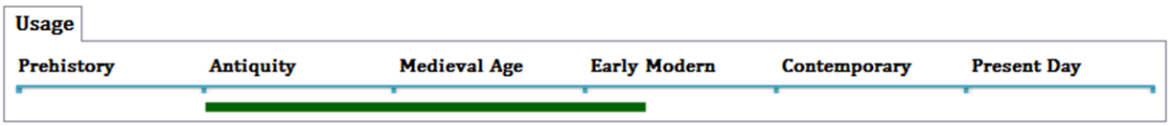


Figure 5.11: Chronological panel displaying the usage of malachite, one of the oldest known green pigments that occurs in Egyptian tomb paintings and in European paintings mainly in the 15th and 16th centuries

Finally, *Database Explorer* contains a menu with three options:

- **Pigment**: This menu allows to add, update and delete the information from a given reference pigment. These actions are performed through specifics MVCs. The *pigment update* view is shown in Fig. 5.12, showing an example of updating the database contents of malachite
- **Edit**: This menu provides pre-defined queries to filter pigments shown in the `List Box` by type, chemical class or *Colour Index*
- **Database**: This menu provides database management utilities such as database contents importing or exporting, useful for system maintenance and data backup activities

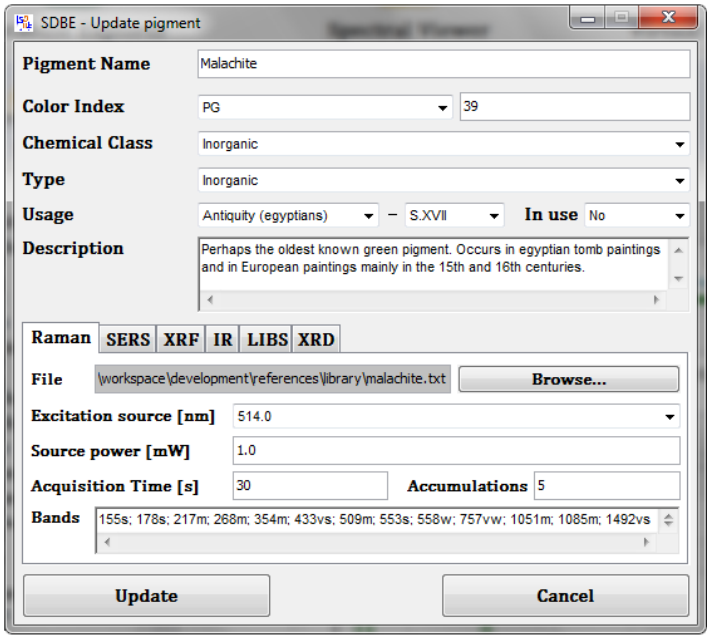


Figure 5.12: View devoted to updating the database contents of a selected reference pigment

5.3.2 Spectral visualisation and data enhancement tools

The ***PigmentsLab***-module devoted to spectral visualisation and data enhancement is called *Spectral Viewer*. This module provides an interactive application with the objective of visualising spectroscopic measurements obtained from art materials. It also provides access to different implementations of data pre-processing and enhancement techniques, such as noise filtering methodologies. The main view of *Spectral Viewer* is presented in Fig. 5.13. In particular, it provides a `2D Canvas` where the spectral plot is provided once a spectrum is loaded. Hence, the user may select different options from the *Spectral Viewer* menu, which implement the main functionalities for data visualisation and enhancement of the module:

- **Zoom**: This option allows to zoom in or zoom out the plot to spectral regions of interest
- **Crop**: This option allows to crop the spectral range under analysis
- **Pointer**: This option allows to add vertical markers or pointers in a selected point, providing information of both the spectral domain (Raman shift in the Raman case) and intensity
- **Intensities**: This option allows to modify the intensity of the spectra in the plot by a scalar, a factor or the min-max intensities normalisation (Eq. 4.1)
- **Filtering**: This option provides tools for performing the main data enhancement activities, i.e. noise filtering. Specifically, different shot noise filtering techniques

can be selected (Wiener filter[169], the median filter, the wavelet filter[170], the Fast Fourier Transform (FFT) filter, and the fuzzy filter[48]). Besides, the fluorescence's baseline rejection methods of polynomial filter or morphology filter[51] can be used. Finally, the automated methodology for noise filtering developed in this research and described in Chapter 3 can be applied as well from this menu

- **Bands**: This options allows to perform a bands localisation and modelling through different mathematical profiles (Lorentzian, Gaussian and Voigt)

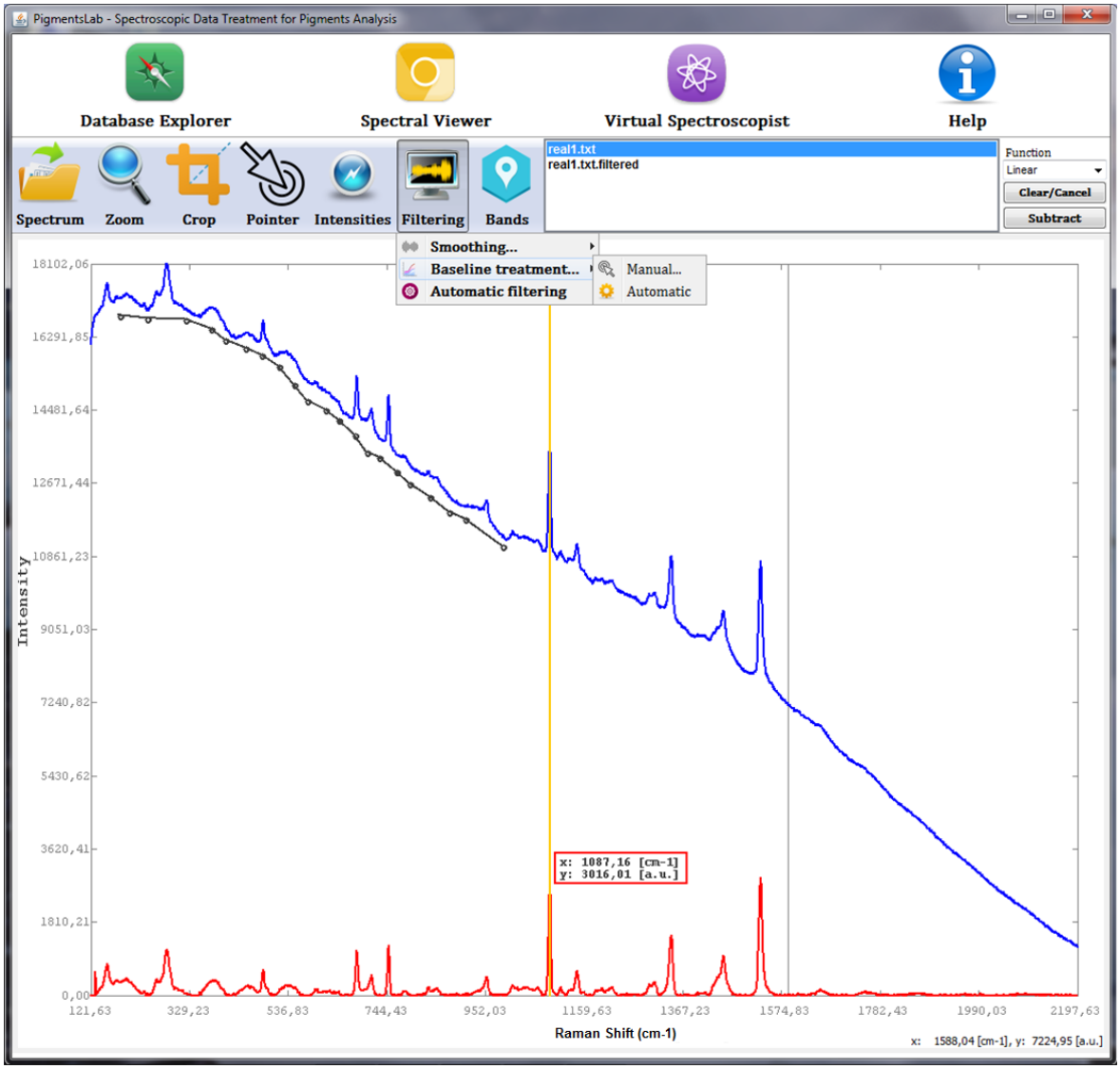


Figure 5.13: Main view of *Spectral Viewer*

5.3.3 Methodologies for automated interpretation of spectra

The ***PigmentsLab***-module devoted to the automated interpretation of spectra is called *Virtual Spectroscopist*. This module provides advanced solutions to help spectroscopists and art analysts on interpreting the data coming from spectroscopic techniques

applied to the analysis of art works. The main view of *Virtual Spectroscopist* is presented in Fig. 5.14. In particular, it is split in two views: the *references* view and the *unknowns* view. The former is devoted to the selection and visualisation of the reference spectral library with which the analysis will be performed. The later is focused on the data to be interpreted.

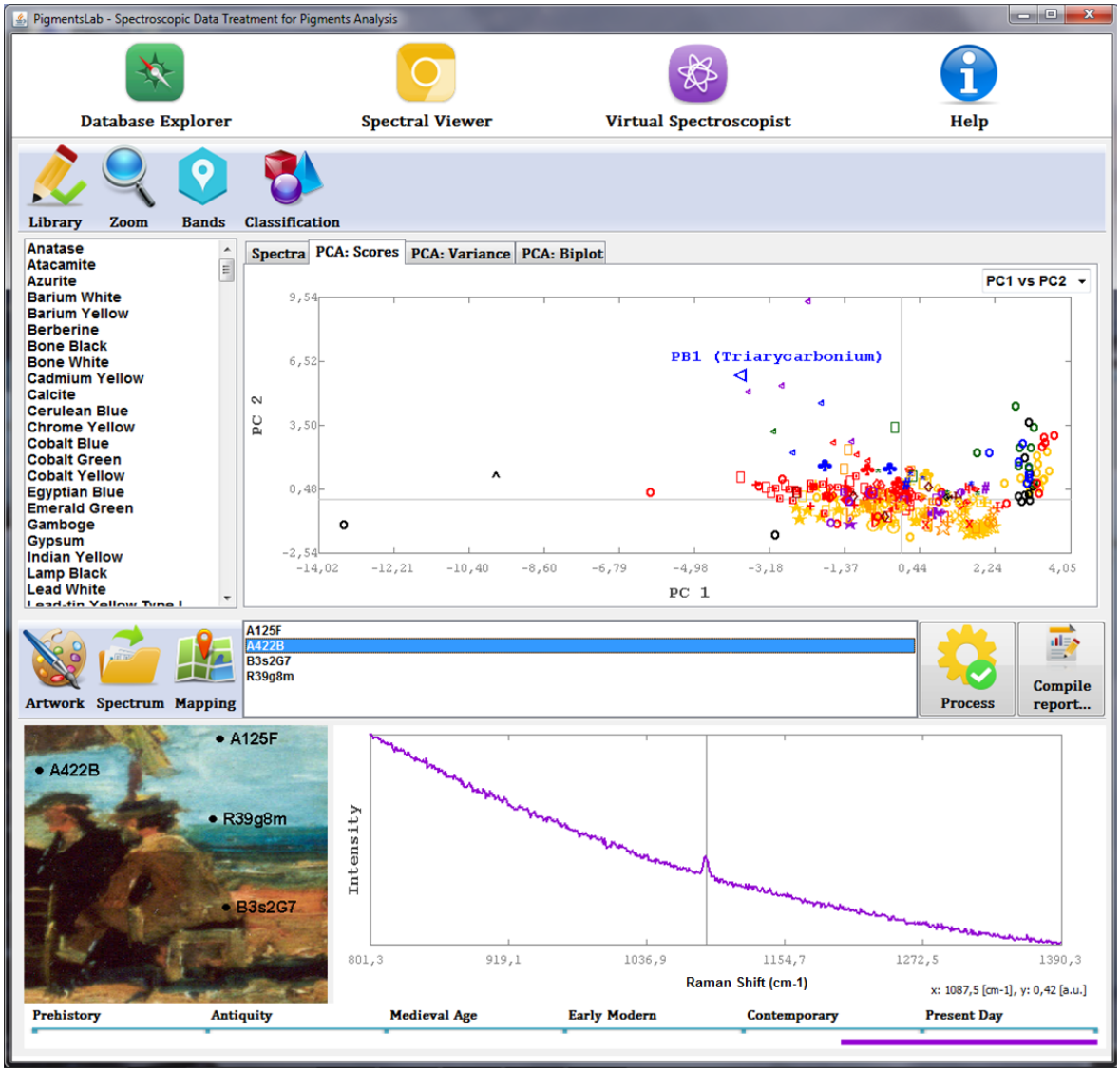


Figure 5.14: Main view of *Virtual Spectroscopist*

The *references* view provides a `List Box`, which similarly to the *Database Explorer* lists the reference pigments present in the database. When a reference pigment is selected from the `List Box`, the system accesses the database contents for the selected item through JDBC, maps its information into the *SpectralTools* data model and plots the corresponding spectrum in a `2D Canvas`. A `Tabbed Pane` allows to select the reference spectral library visualisation mode: spectral representation or PCA projection, which includes 2D-PC component plots, scores and variance as a function of component and 2D biplots. This view contains a menu with the following options:

- **Zoom**: To zoom in or zoom out the plot to spectral regions of interest
- **Bands**: To perform a bands localisation and modelling through different mathematical profiles (Lorentzian, Gaussian and Voigt)
- **Classification**: To customise the reference training sets in the database used for classification. This action is performed through a specific MVC. The *customise classification* view is shown in Fig. 5.15, showing an example of customisation of a training set of phthalocyanine blue

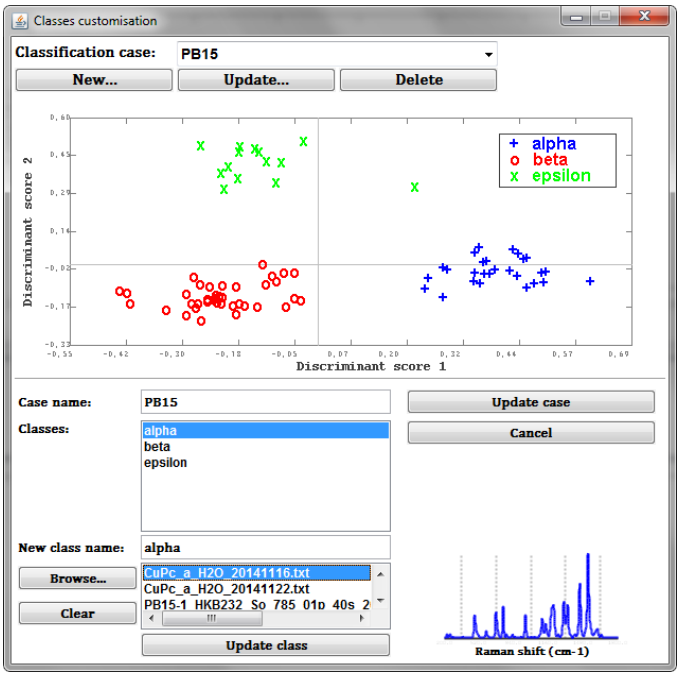


Figure 5.15: View devoted to classification customisation of training sets

The *unknowns* view provides two `2D Canvases`. The use of the first one is optional, and allows the visualisation of the art work under analysis to keep track of the spectral measurement positions on the analysed artwork's surface. The second one is focused on plotting the Raman spectra to be interpreted, which can be loaded through the *Spectrum* button -for single selection, i.e. one spectrum at a time- or through the *Mapping* button -for a set of Raman spectra from a Raman mapping analysis-. The application also provides a chronological panel aimed at graphically displaying information corresponding to the candidate period of creation of the art work under analysis. When the *Process* button is pressed, the system triggers the data interpretation process (introduced in Fig. 5.1) focused on the analysis of Raman spectra from pigments as implemented in ***PigmentsLab***. The data interpretation methodology is graphically represented in Fig. 5.16. Besides, the chronological panel is dynamically updated according to the pigments identification results. When finished, a PDF report may be compiled with the analysis results through the *Report* button.

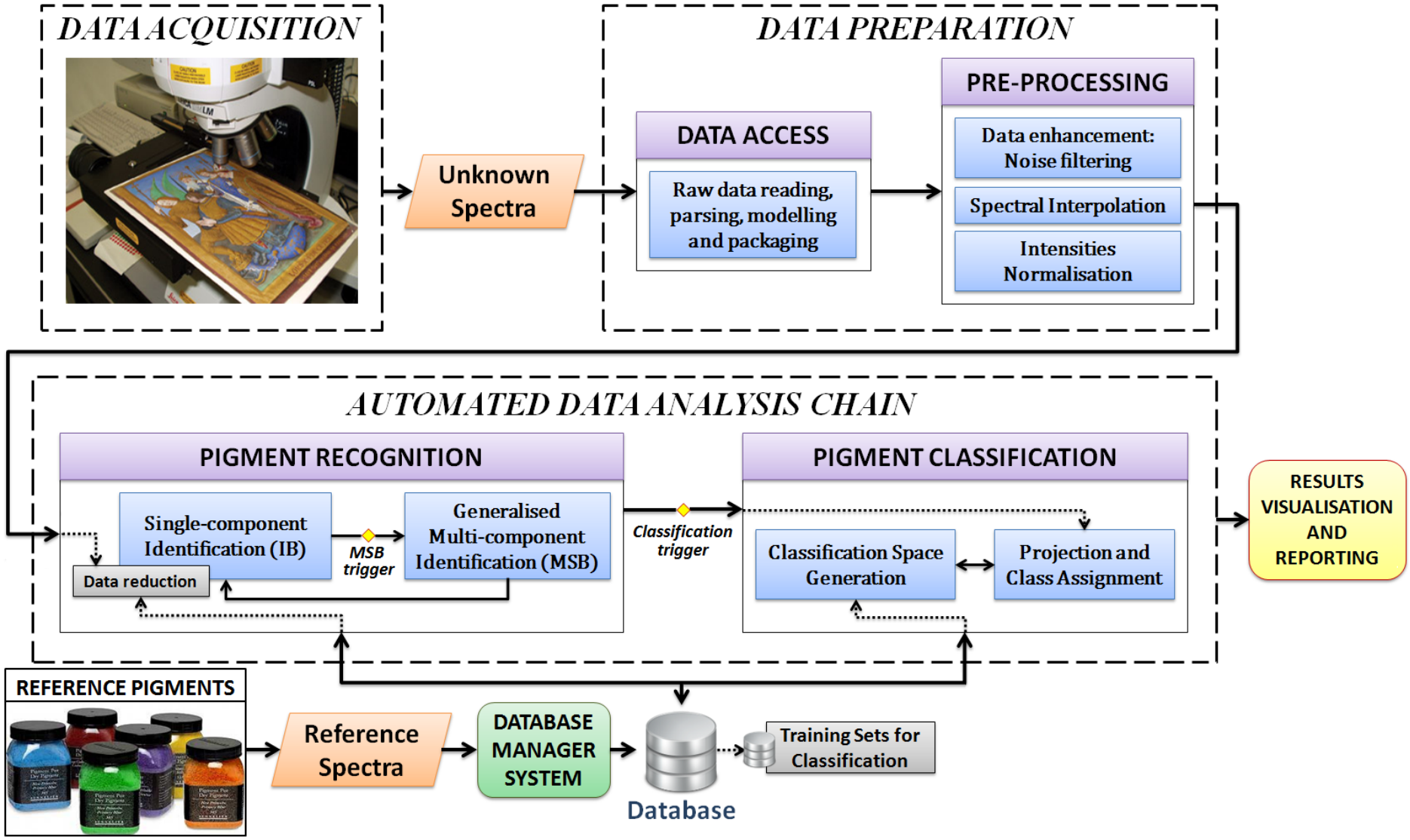


Figure 5.16: Schematic overview of the data interpretation process adapted to Raman analysis in pigments research as implemented in ***PigmentsLab***, which includes: data acquisition (through Raman spectroscopy), preparation (pre-processing), analysis (automated data analysis chain based on pigment recognition and classification), and results visualisation and reporting. The classification is triggered when an unknown spectrum is recognised as a pigment with a database entry of reference training sets

The previously described methodology implemented in ***PigmentsLab*** for the interpretation of spectra from pigments by means of the noise filtering approach described in Chapter 3 and the generalised recognition system together with the supervised classification method outlined in Chapter 4 allows to get insight into pigments in a fully-automated fashion, helping spectroscopists and art analysts such as art conservators in the decision-making process. Next, we provide a description of the data processing workflow implemented in *Virtual Spectroscopist* for the automated identification of Raman spectra from Raman mappings applied to art works analysis.

Raman mapping analysis Raman mapping is a powerful technique for generating detailed chemical images based on an area's Raman spectra. Specifically, a complete Raman spectrum is recorded at each and every pixel of the resulting image and interrogated to create false-colour images. In this sense, identifying pigments from Raman spectra in Raman mapping analysis yields images of pigments distribution. This kind of images is called *Raman identification mapping*. ***PigmentsLab*** allows the interpretation of the Raman spectra from a Raman mapping analysis (see Fig. 5.17): The acquired Raman spectra (u_{ij} with $1 \leq i \leq m$ and $1 \leq j \leq n$) are compiled in a matrix of unknown spectra, $\boldsymbol{U}$, and analysed through the data preparation and analysis chain based on the methodologies developed in this Ph.D. Thesis. Then, the data interpretation results are visualised through a Raman identification mapping where the colour of each pixel represents the pigment recognition and classification result for the corresponding unknown Raman spectrum u_{ij} in $\boldsymbol{U}$.

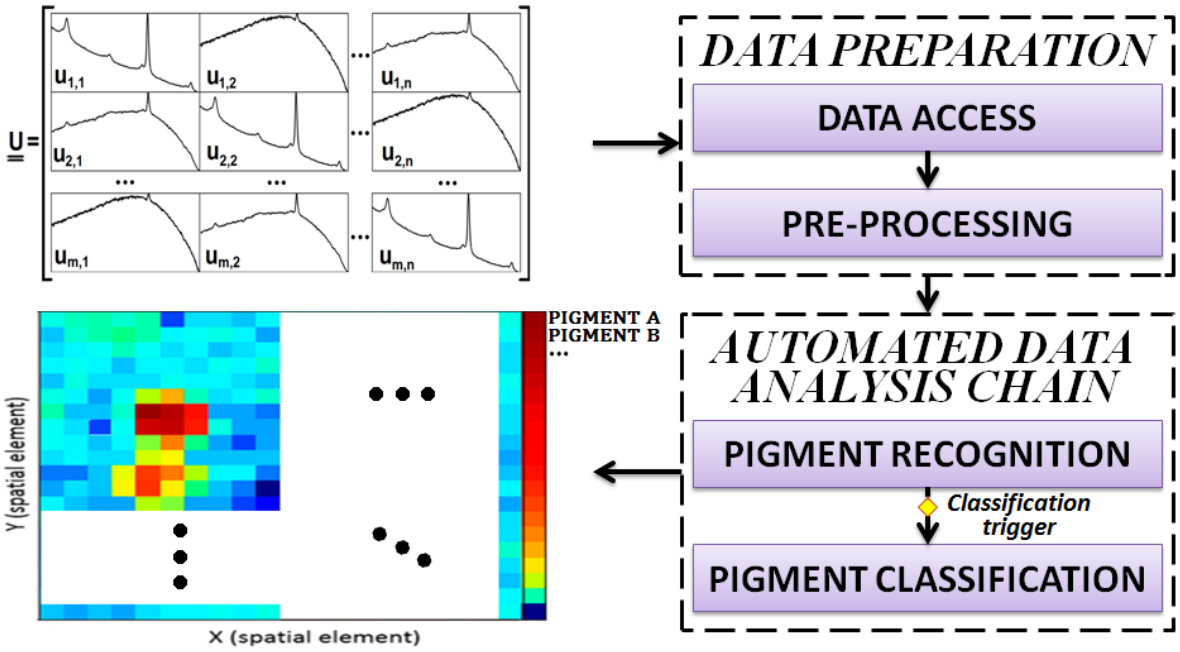


Figure 5.17: Schematic overview of the data interpretation process applied to Raman mapping analysis in pigments research as implemented in ***PigmentsLab***

5.4 Use case: Raman mapping interpretation

The analysis of the pigments composition of a painting was requested (see Fig. 5.18). The art work, signed by Cecilio Pla (València, Spain, 1860 - Madrid, Spain, 1934), was initially attributed to this artist, exponent of the Valencian modernist painting. To carry out this study a Raman mapping analysis was performed. Concretely, a specific area of the analysed art work was selected (the upper part of a sail), and the Raman mapping was recorded measuring Raman spectra in steps of 500μm of the selected area (see Fig. 5.19). Acquisition times were of 120 seconds with 3 accumulations for each Raman measurement.

Figure 5.18: Painting initially attributed to Cecilio Pla. The selected area under analysis (corresponding to the upper part of a sail) is marked with a red box

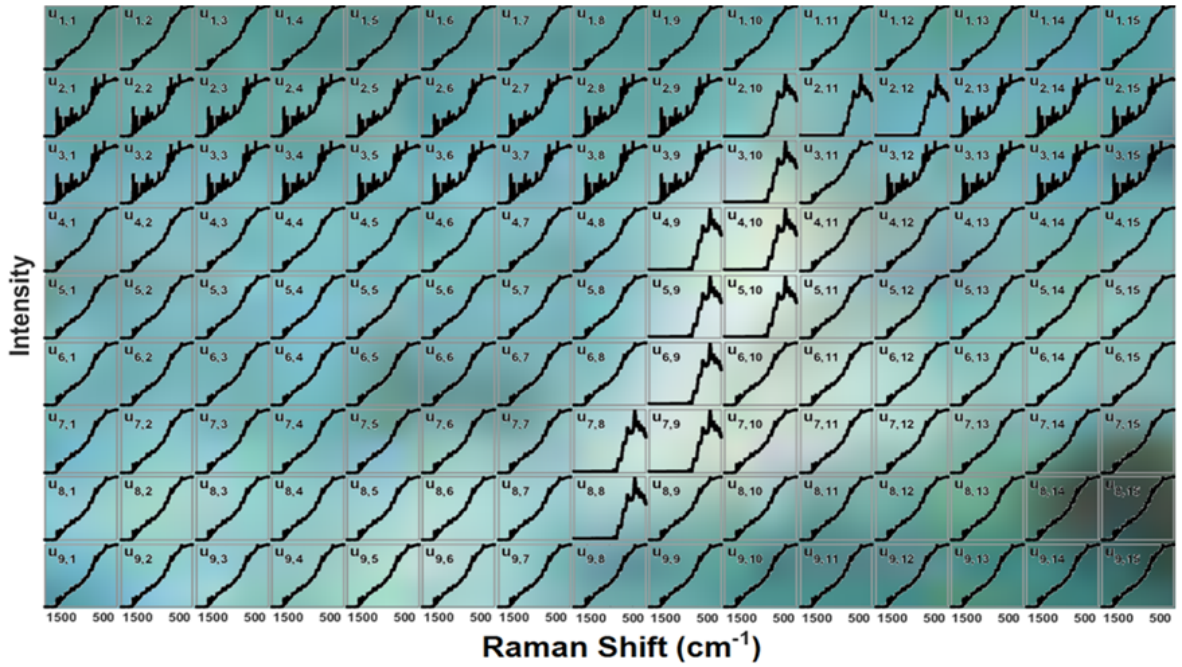


Figure 5.19: Experimental Raman spectra acquired through Raman mapping in the selected area under analysis of the art work - the upper part of a sail, shown in the background

The identification results obtained from the Raman spectra conforming the Raman mapping by means of the automated data analysis chain depicted in Fig. 5.16 and implemented in ***PigmentsLab*** are shown in Fig. 5.20(a) and (b). These results spotted out the usage of rutile (introduced in the European painting in 1945[5]), and α-copper phthalocyanine blue (first used in paintings around 1935[4]). These results together with the pigments analysis performed in other art work's surfaces were shown to be inconsistent with the date of death of the author (1934).

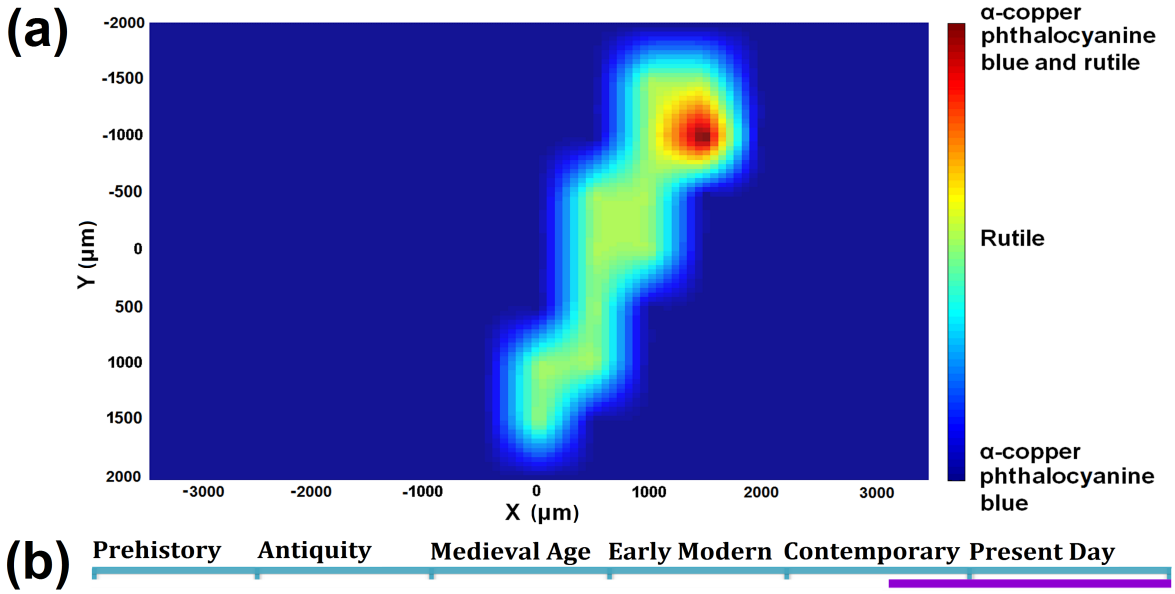


Figure 5.20: Resulting Raman identification mapping representing the identified pigments (b) and chronological information from this pigment analysis (c) provided by ***PigmentsLab***

5.5 Chapter summary

The design and implementation of an intuitive, reliable and user-friendly open-source cross-platform software platform has been presented. Specifically, the system integrates historical and spectroscopic data from art materials as well as advanced tools for the visualisation and pre-processing of spectra, database handling and management solutions, and high-performance methodologies to aid in the decision-making process for the interpretation of spectra. The platform lies in a niche market which calls for its development, implementation and release to the community.

The system implementation is based on a three-module scheme. Each module is focused on one of the three main tasks an art analyst usually performs on a daily basis: the *Database Explorer*, aimed at quickly retrieving information of reference materials as well as to handle main database management operations; the *Spectral Viewer*, provided as an interactive application to visualise spectroscopic measurements from

art materials, implementing sophisticated data treatment solutions such as noise filtering; and finally the *Virtual Spectroscopist*, including advanced solutions to help the analysts and spectroscopists on interpreting spectroscopic data from art works analysis.

The design of the system is based on a specific definition of data model and access layer, used in a common manner by the three main modules of the platform. In this sense, the proposed data model definition is certainly a key point in the dissemination of the platform to the scientific community in the field of cultural heritage, aimed at standardising the data format in a common fashion.

The global system of automated interpretation of spectra in art analysis here proposed will exceedingly benefit the scientific community devoted to the analysis and preservation of the cultural heritage, helping to make breakthroughs in processing and analysing spectroscopic data, and will be the forthcoming reference tool in the scientific exploitation and interpretation of spectroscopic data from art materials.

Chapter 6

Raman characterisation of polymorphic forms of copper phthalocyanine blue under solvents and cleaning agents

6.1 Chapter overview

This chapter presents an analysis of polymorphic forms of copper phthalocyanine blue. To do so, an overview of this pigment is provided in Sect. 6.2. A molecular characterisation aimed at discriminating the different polymorphic forms of the pigment was performed through Raman spectroscopy from dry pigments and under solvents and cleaning agents. The results of the spectral classification are provided in Sect. 6.3, which was carried out using the global system of automated interpretation of spectra described in Chapter 5.

6.2 Copper phthalocyanine blue: a brief overview

Copper phthalocyanine blue (CuPc) is the most important synthetic organic blue pigment from the 21st century artists' art works. This pigment is identified in the *Colour Index* as Pigment Blue 15 (PB15). Specifically, seven types of CuPc are included in this category[171]. Concretely, PB15:0 designs the unstabilised α-modification of CuPc, PB15:1 for the non-crystallising α-modification of CuPc, PB15:3 for the unstabilised β-modification of CuPc, PB15:4 for the non-flocculating β-modification of CuPc and PB15:6 for the unstabilised ϵ-modification of CuPc - PB15:2 (non-crystallising non-flocculating α-modification of CuPc) is in general not use as an artistic pigment, and PB15:5 (γ-modification of CuPc) is not produced by art manufacturers. The α-, β-,

and ϵ-modifications of CuPc are used as artistic pigments and are characterised by differences in stability, solubility and hue[172]. The α-modification of CuPc was introduced on the European market in 1935, the β-modification of CuPc was patented in 1953 and the ϵ-modification of CuPc was first used before the 1970s[173]. Ultramarine, Prussian blue, cobalt blue and cerulean blue remain in common use, but PB15 clearly appears as the most widespread artists' blue pigment. Thus, it is clear that identifying the polymorphic form of a CuPc may provide chronological details valuable to be used as marker to date and authenticate art works. In this sense, when a PB15 is detected in the original paint layers of an art work, the possibility of it to be created before 1935 may be discarded. Additionally, there is a special interest in the identification of CuPc blue pigment in the art conservation field[174]. Main conservation issues are related to the sensitivity of pigments to certain solvents. Hence, CuPc is insoluble in most of solvents but is partially soluble in aromatic solvents[175]. According to the literature, the application of these solvents on a painted surface containing α-modification of CuPc involves a risk of colour changes, resulting from crystallisation defects[176] consisting on a transformation from α-to-β-modification of CuPc.

Many studies agree that Raman spectroscopy is the most efficient technique to identify CuPc blue pigment in paints layers[143,177,178], even in case of complex mixtures irrespective to the CuPc concentration (see Fig. 6.1). In conclusion, the raw data acquired by means of the molecular spectroscopy of Raman spectroscopy provides insight into the pigments polymorphic form through the methodologies integrating the global system described in Chapter 5. The following section discusses the discrimination of experimental Raman spectra from CuPc using dry pigments and pigments under solvents and cleaning agents.

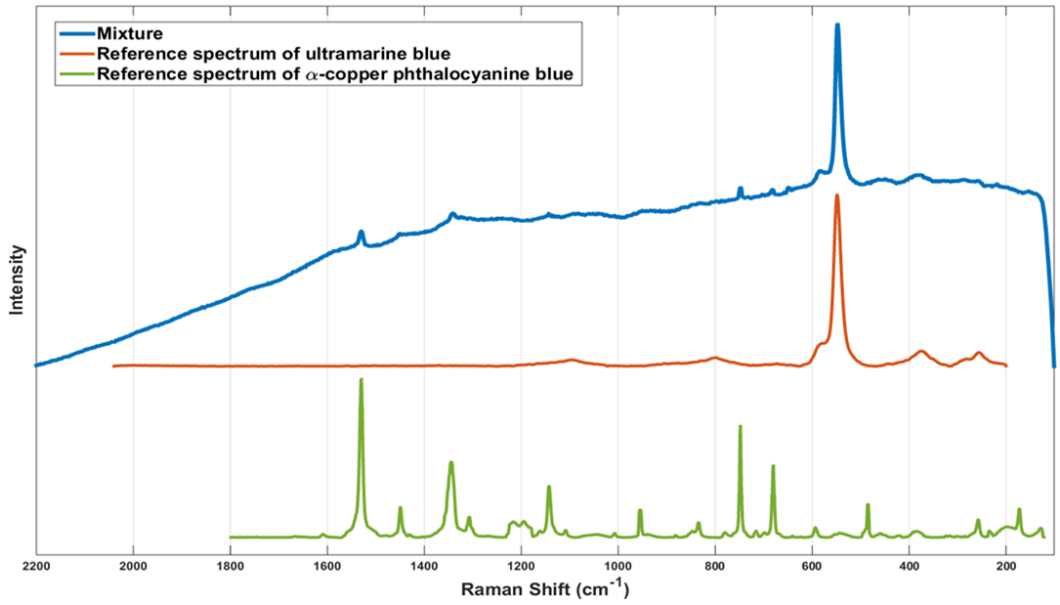


Figure 6.1: Experimental Raman spectra (blue) from a mixture of ultramarine blue and CuPc. Corresponding references are shown in red and green, respectively

6.3 Supervised classification of Raman spectra from copper phthalocyanine blue

The supervised classification methodology described in Sect. 4.4.2 of Chapter 4 was used to discriminate Raman spectra from the three main crystalline structures of copper phthalocyanine blue pigment, i.e. α-, β- and ϵ-modifications. Thus, three reference classes were built. Additionally to the reference spectra recorded by the author from pigment powders to generate the training dataset (see Fig. 6.2), a significant set of reference spectra were supplied by three different researchers and therefore recorded using different acquisition systems under different measurement conditions.

Figure 6.2: Pigment powders and hand-made samples from different crystalline structures of copper phthalocyanine blue

In this sense, it is well-known that there may be differences between Raman spectra recorded with different instruments, which may become a handicap for the purposes of pigment classification. Indeed, instrument resolution, excitation wavelength or even laser power of the excitation source can strongly influence the Raman bands[179] (see Fig. 6.3). Nevertheless, the classification methodology presented in Chapter 4 is not affected by these issues as illustrated by the results shown hereafter as long as the user-defined reference classes are properly defined and represented in the classification space: the implemented system automatically picked up the spectral markers for classification by means of the PCs scores regardless of the heterogeneous input data and discriminated the CuPc classes in the classification space.

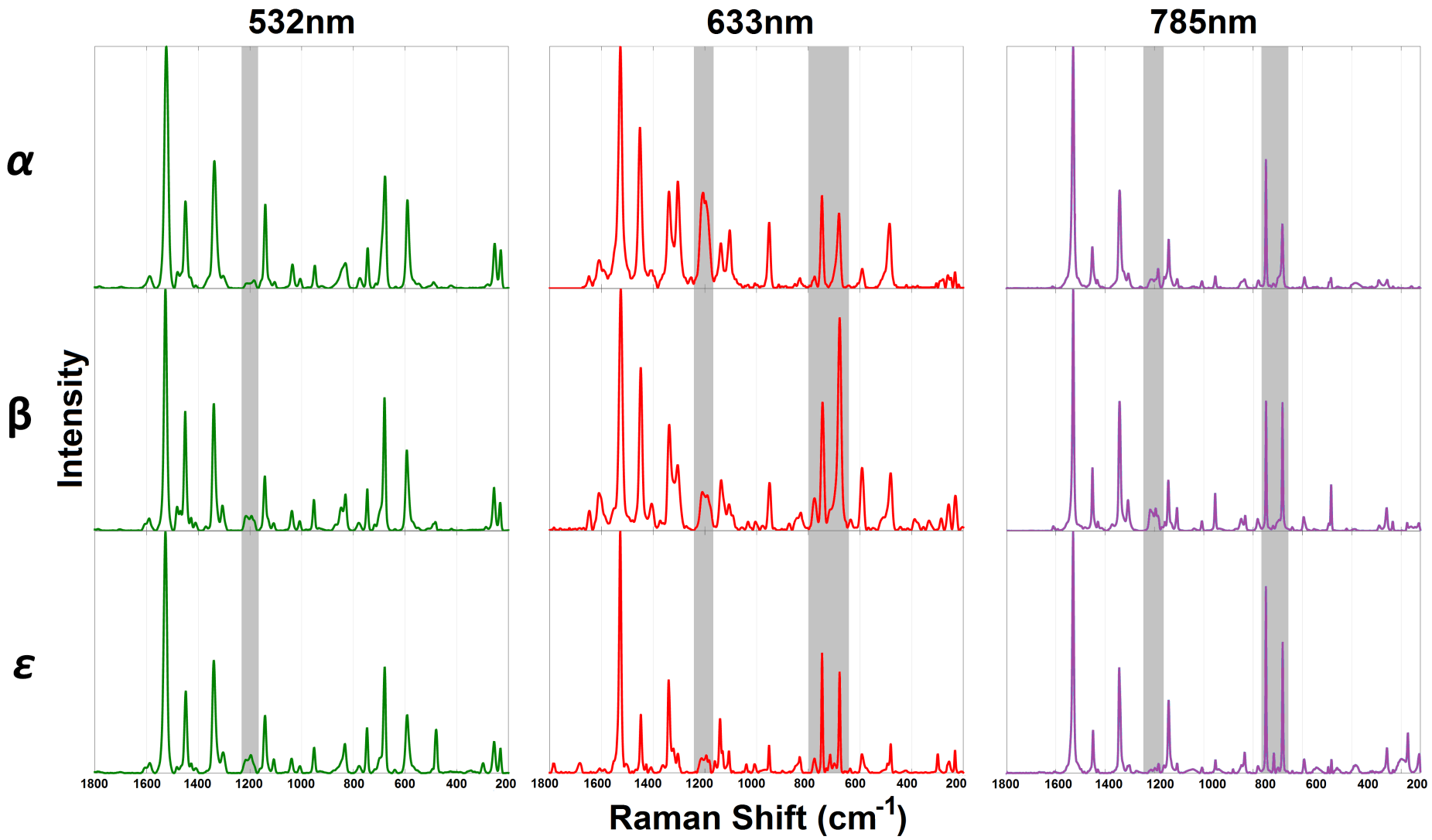


Figure 6.3: Reference Raman spectra from α-, β- and ϵ-modification of copper phthalocyanine blue showing spectral differences depending on the excitation wavelength used for acquisition - green (532nm), red (633nm) and infrared (785nm) excitation sources. Spectral markers generally used for visual discrimination between α-, β- and ϵ- classes for a given excitation wavelength are highlighted with a greyish shadow

Specifically, the α-modification class of copper phthalocyanine blue consisted of 27 Raman spectra: nine Raman spectra recorded using a 532nm excitation wavelength, ten Raman spectra recorded using a 633nm excitation wavelength, and eight Raman spectra recorded using a 785nm excitation wavelength. The β-modification class of copper phthalocyanine blue consisted of 38 Raman spectra: eleven Raman spectra were recorded using a 532 nm excitation wavelength, thirteen Raman spectra were recorded using a 633 nm excitation wavelength, and fourteen Raman spectra were recorded using a 785nm excitation wavelength. Finally, the ϵ-modification class of copper phthalocyanine blue consisted of 14 Raman spectra: ten Raman spectra were recorded using a 532nm excitation wavelength, one Raman spectrum was recorded using a 633nm excitation wavelength, and three Raman spectra were recorded using a 785nm excitation wavelength. The feature extraction module provided a 23-dimensional PCs space with an accumulative variance of 99.29% (see Fig. 6.4). The classification space is described by three class regions (one for each reference class) with a minimum JMD of 1.99 (see Fig. 6.5).

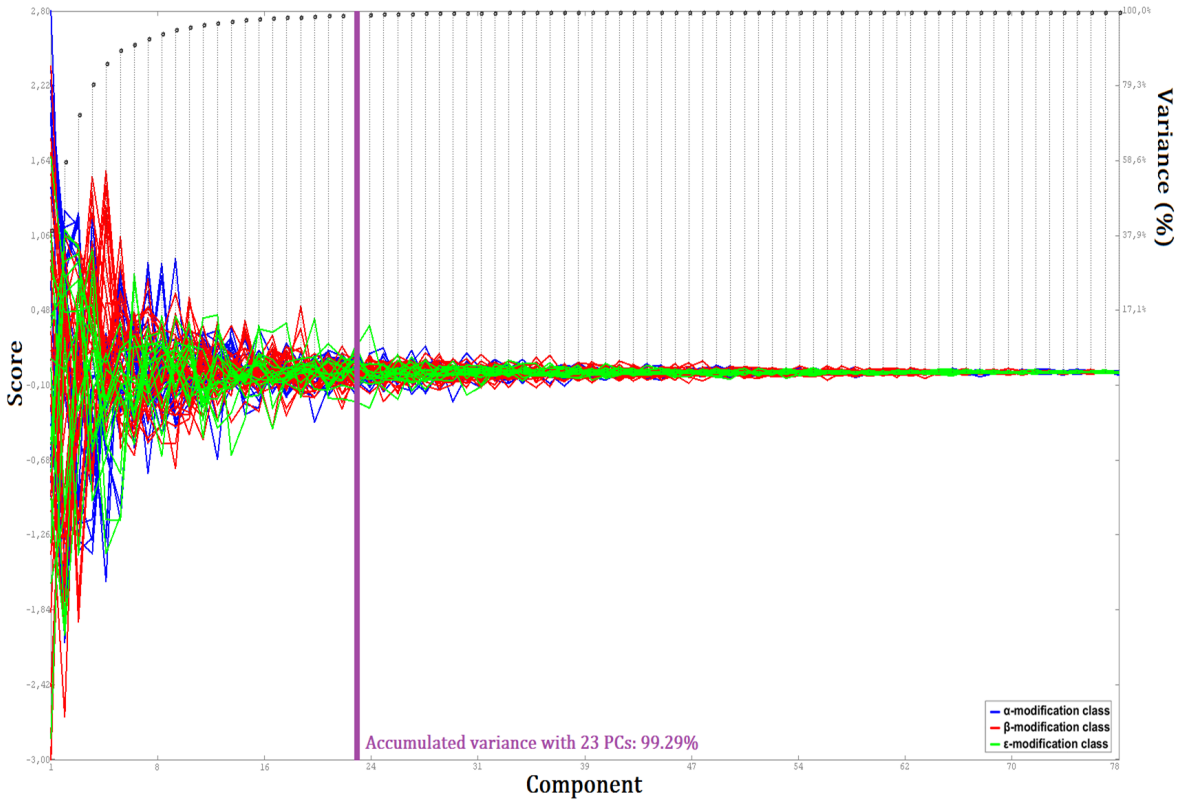


Figure 6.4: PCA scores of the reference training set build from Raman spectra from α-, β- and ϵ-modifications of copper phthalocyanine blue, together with the accumulated variance of PCA projection as a function of Principal Component. The 23-dimensional PCs space accounts for an accumulative variance of 99.29%

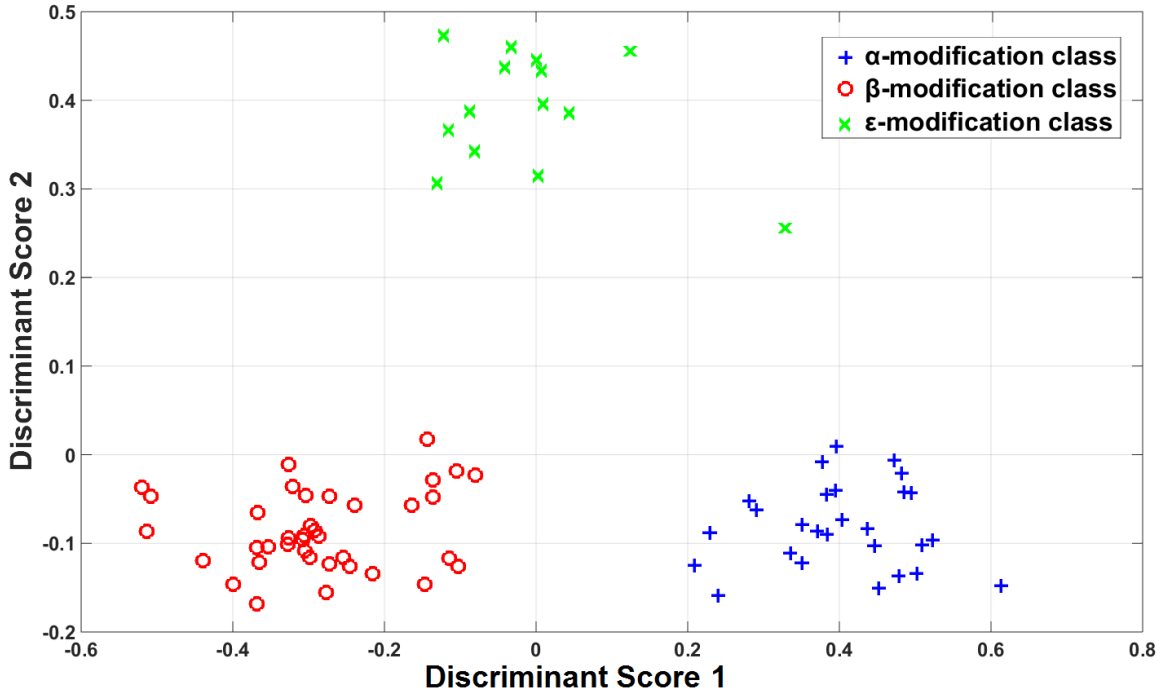


Figure 6.5: Classification space generated from the training dataset of reference Raman spectra from copper phthalocyanine blue pigment: α-modification class (+), β-modification class (o), ϵ-modification class (x)

6.3.1 Experimental results using dry pigments

Table 6.1 presents the classification results for Raman spectra kindly supplied by Marta Anghelone (Academy of Fine Arts Vienna) measured on hand-made samples built from copper phthalocyanine blue using dry pigments, i.e. without solvents or cleaning agents. These hand-made samples were prepared with PB15:1, PB15:3 and PB15:6 pigment powders manufactured by Kremer Pigments, which were mixed with several binding agents in different proportions and subjected to a UV aging process as reported in[150]. We applied the supervised classification methodology to a total of 36 Raman spectra (see Fig. 6.6) and we obtained a success rate of 100%, showing the consistency of the implemented supervised classification methodology presented in Chapter 4. Fig. 6.7 shows the projection of the unknown Raman spectra onto the classification space. The Raman spectra classified with lower matching values were affected by Raman band shifting, Raman band spreading and intensity inversions.

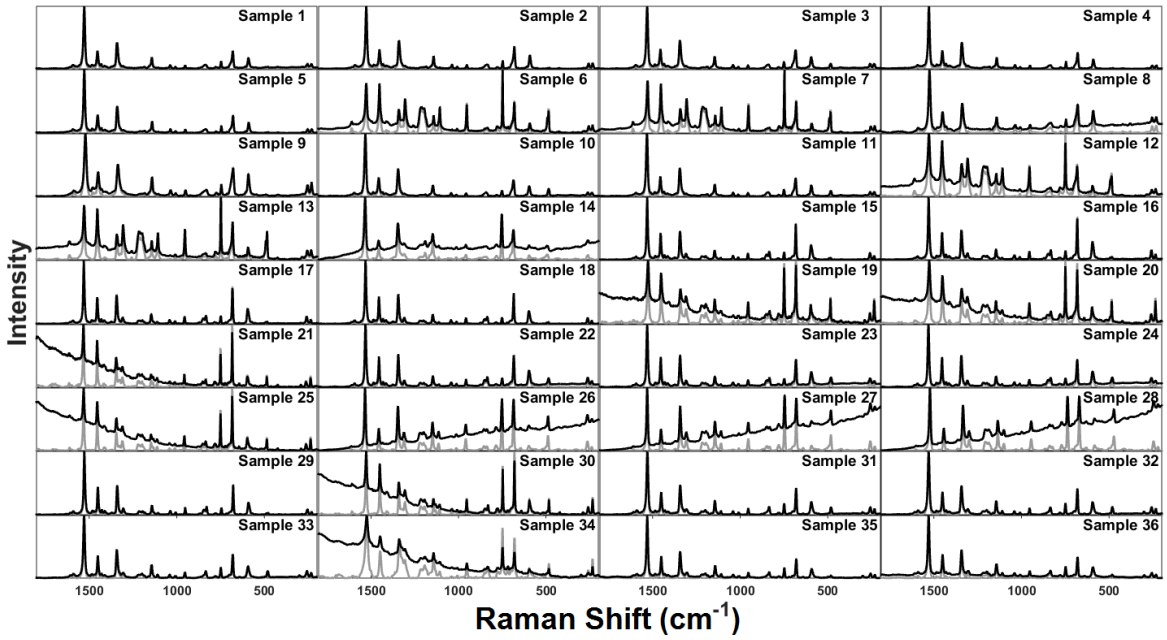


Figure 6.6: Experimental Raman spectra from copper phthalocyanine blue measured on hand-made samples: acquired Raman spectra (black) and pre-processed Raman spectra (gray)

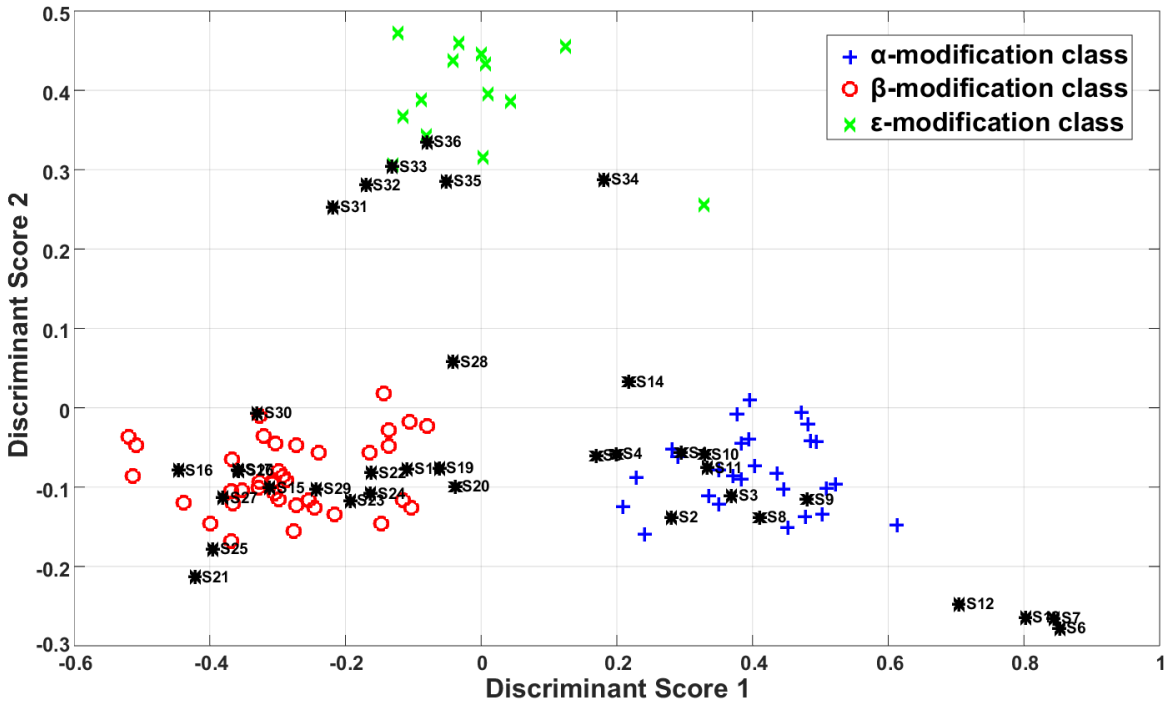


Figure 6.7: Projection of experimental Raman spectra from copper phthalocyanine blue onto the classification space: α-modification class (+), β-modification class (o), ϵ-modification class (x) and unknowns (black asterisks)

Table 6.1: Classification of Raman spectra from copper phthalocyanine blue pigments

	Sample	Expected Class	Assigned Class	MF (%)
S1	PB15:1, 532 nm, alkyd resin, 1	α-modification	α-modification	100.00
S2	PB15:1, 532 nm, alkyd resin, 2	α-modification	α-modification	100.00
S3	PB15:1, 532 nm, alkyd resin, 3	α-modification	α-modification	100.00
S4	PB15:1, 532 nm, linseed stand oil, 1	α-modification	α-modification	89.41
S5	PB15:1, 532 nm, linseed stand oil, 2	α-modification	α-modification	86.12
S6	PB15:1, 633 nm, linseed stand oil, 1	α-modification	α-modification	20.35
S7	PB15:1, 633 nm, linseed stand oil, 2	α-modification	α-modification	24.88
S8	PB15:1, 532 nm, acrylic binder, 1	α-modification	α-modification	100.00
S9	PB15:1, 532 nm, acrylic binder, 2	α-modification	α-modification	100.00
S10	PB15:1, 532 nm, acrylic binder, 3	α-modification	α-modification	100.00
S11	PB15:1, 532 nm, acrylic binder, 4	α-modification	α-modification	100.00
S12	PB15:1, 633 nm, acrylic binder, 1	α-modification	α-modification	57.99
S13	PB15:1, 633 nm, acrylic binder, 2	α-modification	α-modification	34.89
S14	PB15:1, 785 nm, acrylic binder	α-modification	α-modification	84.30
S15	PB15:3, 532 nm, alkyd resin, 1	β-modification	β-modification	100.00
S16	PB15:3, 532 nm, alkyd resin, 2	β-modification	β-modification	100.00
S17	PB15:3, 532 nm, alkyd resin, 3	β-modification	β-modification	100.00
S18	PB15:3, 532 nm, alkyd resin, 4	β-modification	β-modification	100.00
S19	PB15:3, 633 nm, alkyd resin, 1	β-modification	β-modification	100.00
S20	PB15:3, 633 nm, alkyd resin, 2	β-modification	β-modification	100.00
S21	PB15:3, 633 nm, alkyd resin, 3	β-modification	β-modification	85.12
S22	PB15:3, 532 nm, acrylic binder, 1	β-modification	β-modification	100.00
S23	PB15:3, 532 nm, acrylic binder, 2	β-modification	β-modification	100.00
S24	PB15:3, 532 nm, acrylic binder, 3	β-modification	β-modification	100.00
S25	PB15:3, 633 nm, acrylic binder	β-modification	β-modification	100.00
S26	PB15:3, 785 nm, acrylic binder, 1	β-modification	β-modification	100.00
S27	PB15:3, 785 nm, acrylic binder, 2	β-modification	β-modification	100.00
S28	PB15:3, 785 nm, acrylic binder, 3	β-modification	β-modification	71.83
S29	PB15:3, 532 nm, linseed stand oil	β-modification	β-modification	100.00
S30	PB15:3, 633 nm, linseed stand oil	β-modification	β-modification	100.00
S31	PB15:6, 532 nm, alkyd resin, 1	ϵ-modification	ϵ-modification	79.07
S32	PB15:6, 532 nm, alkyd resin, 2	ϵ-modification	ϵ-modification	87.80
S33	PB15:6, 532 nm, alkyd resin, 3	ϵ-modification	ϵ-modification	79.76
S34	PB15:6, 633 nm, alkyd resin	ϵ-modification	ϵ-modification	100.00
S35	PB15:6, 532 nm, acrylic binder, 1	ϵ-modification	ϵ-modification	100.00
S36	PB15:6, 532 nm, acrylic binder, 2	ϵ-modification	ϵ-modification	100.00

6.3.2 Experimental results using pigments under solvents and cleaning agents

In order to perform a Raman characterisation of polymorphic forms of the CuPc pigment under solvents and cleaning agents the following products commonly used by art conservators were selected: white spirit, dimethyl sulfoxide, formic acid, toluene and xylene. Specifically:

A. **White spirit:** Solvent extracted from petroleum, colourless or slightly yellowish, with odour of kerosene, very little soluble in water and with a boiling point between 140°C and 200°C. Used to dissolve oils, waxes, paraffins and resins.

B. **DiMethyl SulfOxide (DMSO):** Solvent widely used in the field of restoration because it is an optimal cleaning agent for various types of materials. It is somewhat toxic with respect to other products with similar characteristics. It dissolves most of the salts, many protein substances and the vegetable gums. It is therefore a solvent that, also considering easy availability, is used in numerous operations that are performed in the restoration, except for that in which the solvent action is too interested in some original materials of the paintings.

C. **Formic acid:** Formic acid is one of the most penetrating solvents. The most dangerous solvents for the original pictorial materials, evidently are those that being very penetrating, also present a strong and long retention. Therefore formic acid is in category one of the solvents which are the strippers; very penetrating and high and long retention, but also a powerful irritant to the skin and mucous membranes. It is effective when looking to remove layers based on proteins.

D. **Toluene:** It is an aromatic hydrocarbon, liquid insoluble in water with a characteristic smell to the thinner of paintings. It is used as solvent for paints, coatings, rubber, resins, thinner in nitrocellulose lacquers and in adhesives.

E. **Xylene:** It is a colourless liquid derived from flammable benzene and sweet odour. It is used as solvent in paints, rubber, leather and related industries. Xylenes are flammable, really irritant and poisonous.

Several experimental hand-made samples were prepared to analyse the effect of the selected solvents and cleaning agents on the α-, β- and ϵ-modifications of copper phthalocyanine blue. In particular, the two different α-modification of CuPc were used: the unstabilised α-modification of CuPc (PB15:0) manufactured by M. Graham (reference code 33-141), and the non-crystallising form (PB15:1) manufactured by Kremer Pigments (reference code 23050). In addition, the unstabilised β-modification of CuPc (PB15:3) and the unstabilised ϵ-modification of CuPc (PB15:6) were also studied, both

manufactured by Kremer Pigments as well (reference codes 23060 and 23070, respectively). In the first step of the samples preparation, a paint layer was painted with the mentioned polymorphic forms of CuPc. Then, the selected solvents and cleaning agents were applied to generate each sample, taking great care not to contaminate the samples, i.e. using a new and different solvent applicator for each sample and solvent product. Finally, the samples of pigments under solvents were subjected to a hot air-based drying process for a 15 minute period. A picture taken during the samples preparation showing the materials used to perform the Raman characterisation of CuPc under solvents and cleaning agents is presented in Fig. 6.8.

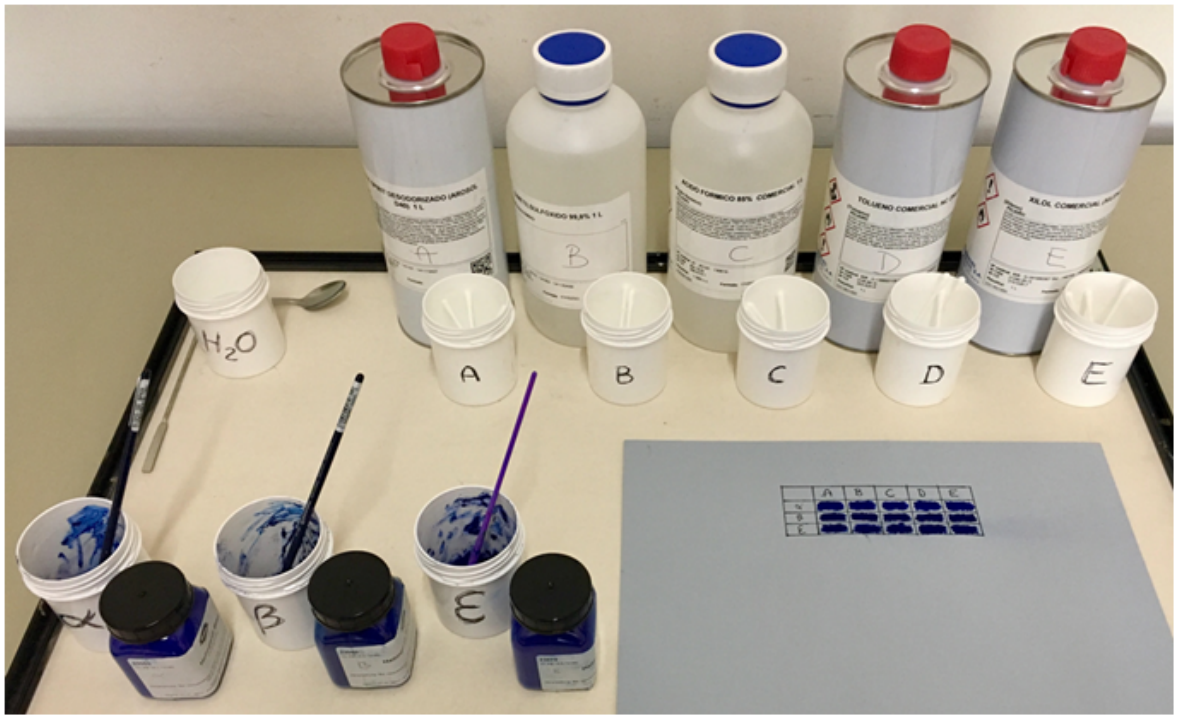


Figure 6.8: Experimental samples preparation using different polymorphic forms of copper phthalocyanine blue under solvents and cleaning agents

The resulting experimental samples built from the unstabilised α-modification of CuPc (PB15:0), the non-crystallising form (PB15:1), the unstabilised β-modification of CuPc (PB15:3) and the unstabilised ϵ-modification of CuPc (PB15:6) under solvents and cleaning agents -white spirit (A), dimethyl sulfoxide (B), formic acid (C), toluene (D) and xylene (E)- is shown in Fig. 6.9. Qualitatively, the effect of the solvents and cleaning agents on the samples was different: the PB15:1 and PB15:6 samples -more stable forms of CuPc- better withstood the solvents application than the PB15:0 and PB15:3 samples. DMSO and formic acid due to their penetration power very negatively affected all the samples, i.e. the pigment concentrations were severely decreased after the application of these solvents.

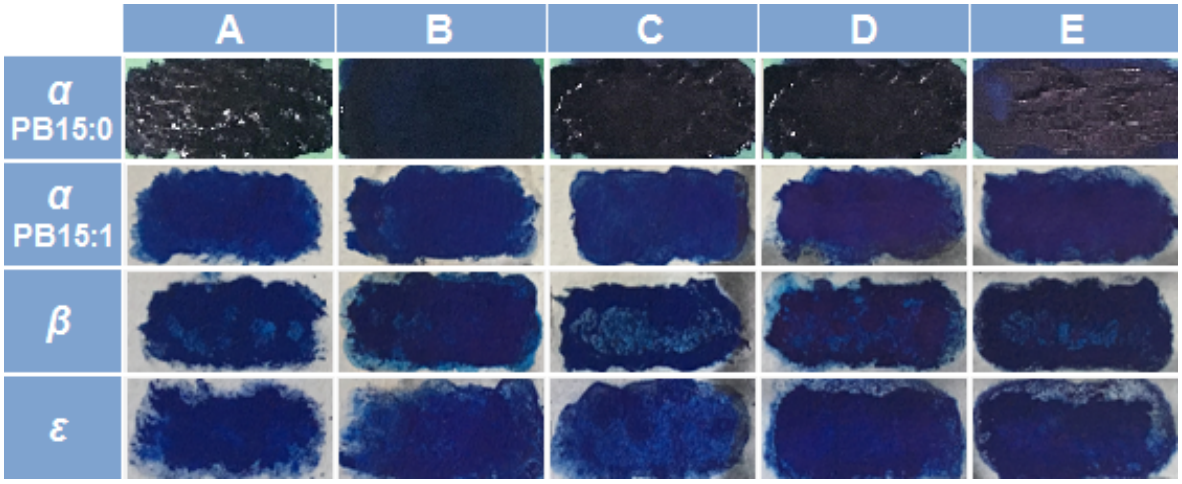


Figure 6.9: Experimental samples of α- (PB15:0 and PB15:1), β- and ϵ-modifications of copper phthalocyanine blue under solvents and cleaning agents: white spirit (A), dimethyl sulfoxide (B), formic acid (C), toluene (D) and xylene (E)

Raman spectra were acquire from these samples (see Raman spectra in Fig. 6.10). Acquisition times were of 100 seconds with 4 accumulations for each Raman measurement. Fig. 6.10 shows the projection of the Raman spectra onto the classification space. The resulting Raman identification mapping obtained through the automated data interpretation process implemented in ***PigmentsLab*** as described in Chapter 5 is shown in Fig. 6.12.

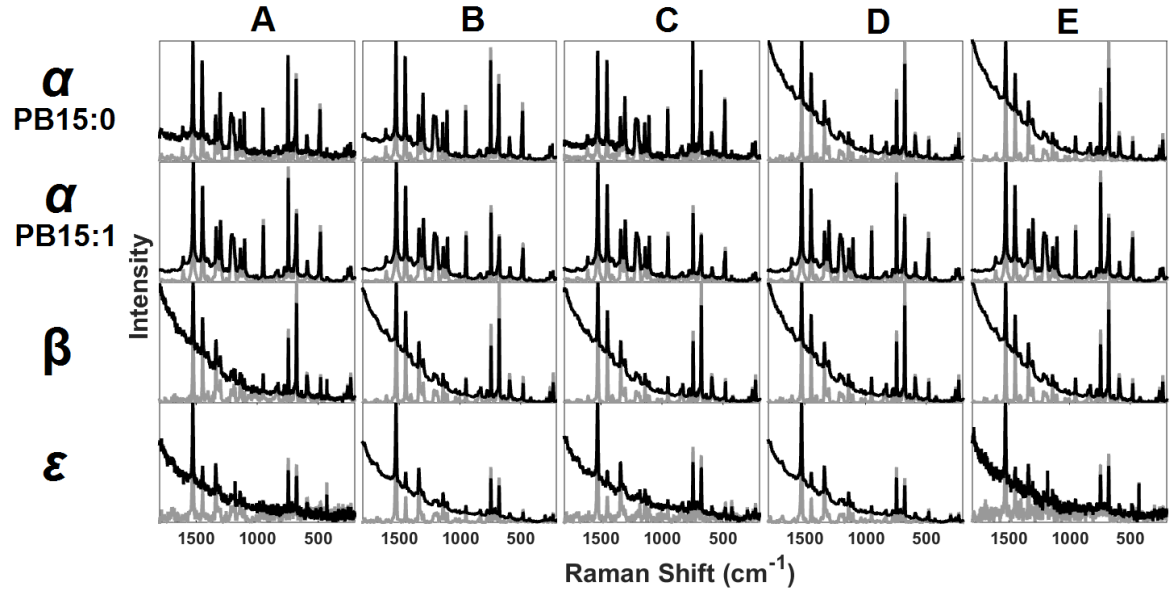


Figure 6.10: Experimental Raman spectra from α- (PB15:0 and PB15:1), β- and ϵ-modifications of copper phthalocyanine blue under solvents and cleaning agents: white spirit (A), dimethyl sulfoxide (B), formic acid (C), toluene (D) and xylene (E). Acquired Raman spectra (black) and pre-processed Raman spectra (gray)

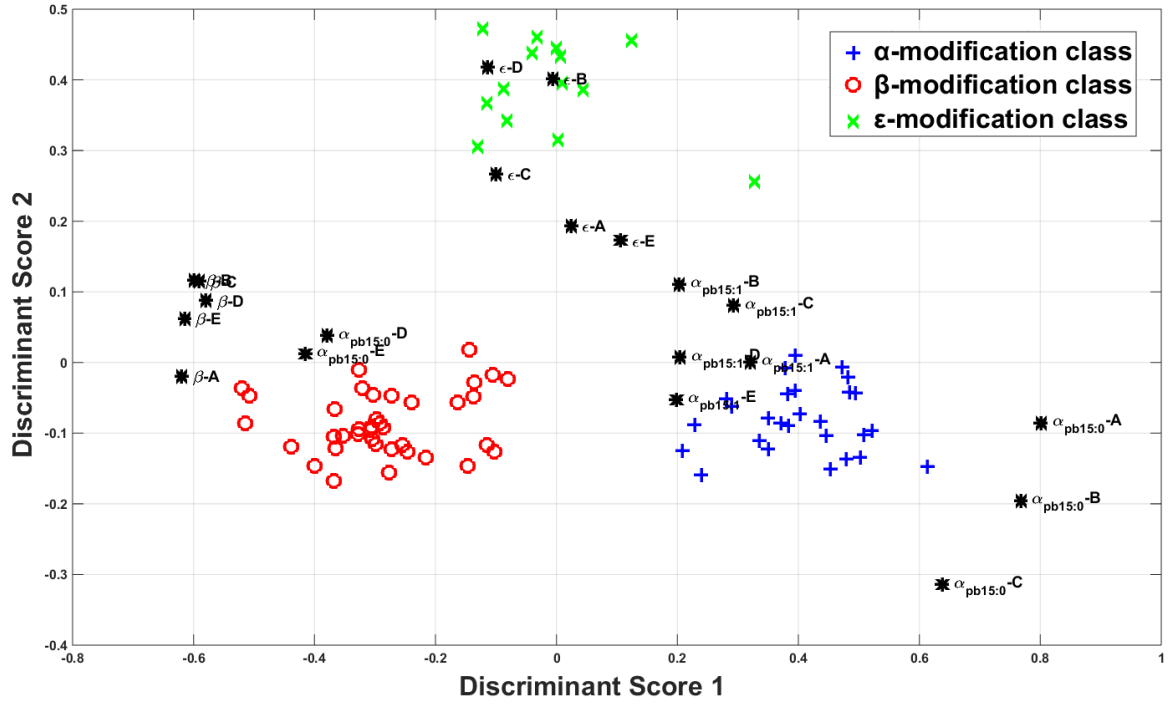


Figure 6.11: Projection onto the classification space of experimental Raman spectra from CuPc under solvents and cleaning agents -white spirit (A), dimethyl sulfoxide (B), formic acid (C), toluene (D) and xylene (E)-: α-modification class (+), β-modification class (o), ϵ-modification class (x) and unknowns (black asterisks)

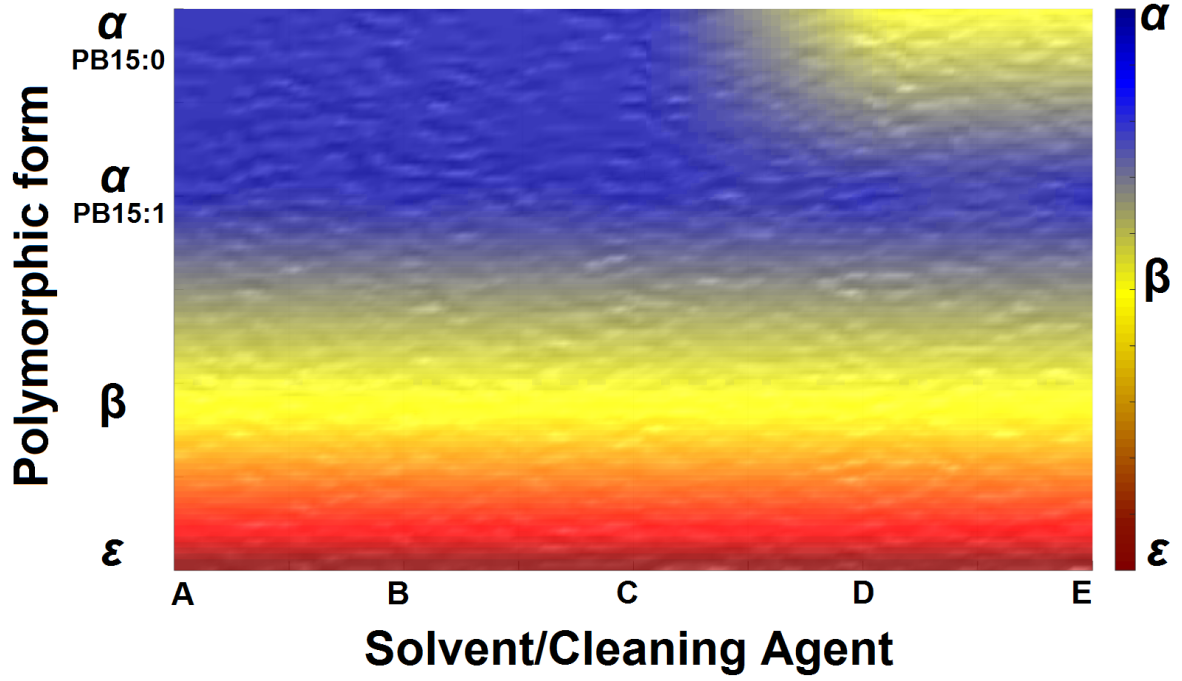


Figure 6.12: Resulting Raman identification mapping representing the identified and classified pigments obtained through ***PigmentsLab***

The colours of the lower rows (yellow and red) of the Raman identification mapping indicate that no polymorphic form transformation was produced after the application of the selected solvents and cleaning agents (white spirit (A), dimethyl sulfoxide (B), formic acid (C), toluene (D) and xylene (E)), i.e. the non-crystallising α-modification of CuPc (PB15:1), the unstabilised β-modification of CuPc (PB15:3) and the unstabilised ϵ-modification of CuPc (PB15:6) correspondingly remained in their original polymorphic form.

However, we can see a polymorphic form transformation in the two latest columns of the top row of the Raman identification mapping: a transformation from α- to β-modification of CuPc occurred when the solvents toluene (D) and xylene (E) were applied to the unstabilised α-modification of CuPc (PB15:0). Consequently, the application of these cleaning agents caused a change in the molecular crystallization of the PB15:0 pigment, which was detected in the raw data acquired through Raman spectroscopy and processed, analysed and visualised through ***PigmentsLab***.

6.4 Chapter summary

A Raman characterisation of polymorphic forms of copper phthalocyanine blue (CuPc) from dry pigments and under solvents and cleaning agents was presented. The main importance of performing this kind of analysis roots on the fact that the CuPc being the most widespread artists' blue pigment may provide chronological details valuable to be used as marker to date and authenticate art works. Indeed, the differences in the crystalline structures of CuPc (α-, β- and ϵ-modifications of CuPc) may go unnoticed, which may occasionally lead to a subjective interpretation.

Art conservators most frequently use solvents to remove highly discoloured or disfiguring varnishes. In addition to being a potential health hazard to the user, solvents can cause serious and irreversible damage to sensitive paint layers. In this sense, the CuPc is partially soluble in aromatic solvents, which may represent an issue in the art conservation field, since the application of these solvents on a painted surface containing α-modification of CuPc may produce crystallisation defects resulting from an α-to-β-modification of CuPc transformation.

The automated data interpretation process described in Chapter 5 was used to discriminate Raman spectra from the three main crystalline structures of copper phthalocyanine blue pigment, i.e. α-, β- and ϵ-modifications of CuPc. Thus, three reference classes were built to create the reference training set, not only using Raman spectra

measured by the author but also acquired in different research laboratories, specifically by Nadim C. Scherrer from the Bern University of Applied Sciences and Marta Anghelone from the Academy of Fine Arts Vienna. In this sense, the implemented system automatically obtained the spectral markers for classification through the PCs scores regardless of the heterogeneity of the input data used to train the system.

In particular, two cases were analysed. In the first case, the study of dry pigments was analysed. In this case, several experimental Raman spectra from hand-made samples were used where the class-membership was *a priori* known. The global system of automated interpretation of spectra implemented in this Ph.D. Thesis, ***PigmentsLab***, was used to process the Raman spectra from the hand-made samples. It successfully classified the analysed Raman spectra in an automated way, demonstrating the consistency of the implemented methodology.

In the second case, the effect of several pre-selected solvents and cleaning agents on the crystalline structures of CuPc was analysed. In this case, the application of ***PigmentsLab*** to a set of Raman spectra from a Raman mapping analysis performed on samples built from different polymorphic forms of copper phthalocyanine blue helped to visualise the effect of those solvents in the molecular crystallisation of CuPc. Indeed, a transformation from α- to β-modification of CuPc occurred when the solvents toluene and xylene were applied to the unstabilised α-modification of CuPc. The α- to β transformation resulted to be irreversible.

Conclusively, permanent damage can easily result from even the most cautious attempts to clean a painting although removing varnish layers is not always advisable. Should it become necessary to clean a painting, its pigments composition should be identified. If a CuPc is present then the usage of aromatic solvents such as toluene and xylene should be avoided, as retrieved from the results exposed in this chapter, where a change in the molecular crystallization of the unstabilised α-modification of CuPc was detected under the application of those solvents.

Chapter 7

Conclusions and future work

7.1 Summary of conclusions

This Ph.D. Thesis belongs to the challenging research area of data interpretation, specifically in the field of Raman spectroscopy applied to art works analysis. The data interpretation is generally outlined in a five-step process composed by acquisition, preparation, analysis, reporting and acting. The fully automation of the data interpretation process was developed in this Ph.D. Thesis to gain insight from raw spectra into pigments in a systematic and objective way. In this sense, the automation of the spectral interpretation process implied the development and analysis of several algorithms such as noise filtering, matching-based identification and spectral classification. For that purpose, the pigments interpretation process required the design and implementation of a useful supporting tool to retrieve an automatic identification of Raman spectra from artistic pigments.

Data acquisition on art works through Raman spectroscopy is based on the Raman scattering, which is produced when a monochromatic light beam makes contact to the analysed material and provides molecular information of the material under analysis. This non-destructive technique allows *in-situ* analysis obtaining objective results in real time. Consequently, Raman spectroscopy is a suitable technique that meets the demanding requirements of art works analysis. Pigment identification may deliver decisive information for the study of both historic and modern paint in the fields of art conservation and the forensic sciences: the signature of a pigment obtained by Raman spectroscopy is unique and allows the unambiguous identification of the analysed pigment through its molecular spectrum. After a brief description of the Raman spectroscopy equipment available in the Universitat Politècnica de Catalunya (UPC) laboratory and its main application -art analysis-, several issues were exposed. In this kind of analysis, coatings, pollutants or binding media among other external agents may degrade the quality of the Raman measurements by increasing the noise

impact. Consequently, a fully-automated noise filtering methodology was designed, developed, implemented and analysed, which enhances the Raman information helping in the interpretation of Raman spectra.

The novel denoising approach that was developed uses the same scheme for both shot noise reduction and fluorescence's baseline removal. Concretely, the developed noise filtering method is based on p-spline fitting, a piecewise polynomial curve fitting technique generally used for data smoothing. One of the key points of the developed denoising methodology is the retrieval of the location of the control points or knots in which the polynomial pieces are joined. Hence, a strategic selection of knots according to the shape of the input Raman spectra was developed by making use of mathematical morphology operations for knots sequence retrieval. In this sense, mathematical morphology operations retrieve the morphology, i.e. the shape, of the Raman information, and adequately preserves positions and intensity ratios of the Raman bands.

The developed fully-automated noise filtering methodology relies on mathematical morphology together with p-spline fitting and demonstrated to be a consistent approach for data enhancement in Raman spectroscopy applied to pigments analysis. The consistency of the denoising method was shown through several tests on both simulated and experimental cases which provided successful results. The method reduces the interferences coming from the main noise sources in Raman spectroscopy -shot noise and fluorescence's baseline- and enhances the Raman information in fully automatic way, i.e. requiring no user intervention.

Thanks to the data preparation performed through the developed noise filtering methodology, the proper data analysis stage may be carried out. To do so, a generalised methodology to automatically identify Raman spectra was designed, developed, implemented and analysed. This recognition method is able to identify single- and multi-component spectra from a single spectral observation. This recognition is performed with no user input or previous knowledge of the analysed sample. The implemented matching-based identification algorithm relies on Principal Component Analysis (PCA) and Independent Components Analysis (ICA) and is computationally efficient and conceptually simple.

The developed generalised identification methodology handles multi-component Raman spectra by means of an iterative strategy based on ICA, which allows to deblend the components in the mixture with high accuracy and no parameters to be configured. This deblending strategy demonstrated to work successfully even when dealing with mixed spectra in different concentrations, i.e. showing their spectral fingerprints with different intensities. Besides, it avoids the manufacturing of reference mixtures from all possible mixtures of pure pigments, with all the variability regarding to relative intens-

ities that this could involve and allows the application of the proposed identification criteria without adding complexity. That is, the generalised identification methodology speeds up the identification process saving computing resources and time.

The spectral recognition system delivers fully automated identification, qualifying the result with a Matching Factor (MF) that is intended to provide guidance in the identification process and should be taken as a value to help the user to make a final decision. The performance of the identification methodology was assessed through simulated spectra and evaluated through several hand-made samples from mixed pigments, and it was applied to real-case spectra from paintings providing consistent results. Therefore, the developed identification system become a practical method for the automated identification of Raman spectra, not only in pigment analysis, but essentially any material group.

The identification system was extended with an automatic classification methodology in order to distinguish between Raman spectra showing small differences among them. In a first attempt, the classification methodology relied on an unsupervised machine learning technique based on clustering analysis. Specifically, it was built from a combination of PCA for feature selection and k-means for data clustering. Nevertheless, the unsupervised-based classification method was not able to perform a proper clustering of Raman spectra measured using different excitation wavelengths. This issue represents a drawback for classifying Raman spectra measured in different Raman laboratories that most likely use different excitation sources: a classification methodology should be a blind method, it should not depend on the measurement configuration of the input dataset. To overcome this issue a supervised machine learning-based classification methodology was developed.

According to predefined reference classes, the developed supervised classification methodology is able to classify unknown spectra from a single spectral observation, with no user intervention or *a priori* knowledge of the analysed sample. The developed classifier relies on PCA and Multiple Discriminant Analysis (MDA) and demonstrated to be a suitable tool for art works analysis as it successfully classified the analysed Raman spectra in a consistent way. The implemented classification system was applied to experimental Raman spectra from pigments, and the obtained results showed that it may play a good auxiliary role in the analysts' endpoint classification.

The methodologies implemented in this Ph.D. Thesis were integrated in a global system for the automated interpretation of spectra from pigments analysis, namely ***PigmentsLab***. Hence, the design and implementation of an intuitive, reliable and user-friendly open-source cross-platform software platform was developed. Specifically, besides the above-commented methodologies, the system integrates historical and

spectroscopic data from art materials as well as several tools for spectral visualisation and database management, in order to aid in the decision-making process for the interpretation of spectra.

The system implementation is built upon a three-module scheme according to the three main tasks spectroscopists and art analysts usually perform on a daily basis: the *Database Explorer* -to retrieve information of reference materials as well as to handle main database management operations-, the *Spectral Viewer* -interactive application to visualise spectroscopic measurements from art materials-, and finally the *Virtual Spectroscopist* -including the developed automated methodologies in this Ph.D. Thesis to help on interpreting spectroscopic data from art works analysis. The developed system relies on a specific definition of data model and data access commonly used throughout the implemented platform, aimed at standardising the spectral data format.

The combination of the automated methodologies for recognition classification is the basis of the main automated data interpretation process implemented in ***PigmentsLab***. It is worth noting that, together with the developed noise filtering approach, these methodologies make no assumptions with respect to the input data, applying a blind treatment of the Raman spectra and processing them in a transparent way regardless of the interpretation purposes. Consequently, it is perfectly capable of dealing with spectra from different sources, i.e. recorded with different acquisition systems and measurement conditions. This fact may represent a significant advantage of the presented automated system in the applications of pigment identification and classification in art analysis through Raman spectroscopy, as it is independent of the measurement system and the configuration used for the acquisition of Raman spectra.

The global system of automated interpretation of spectra in art works analysis integrates the developed noise filtering and identification and classification methodologies. It is expected to exceedingly benefit the scientific community devoted to the analysis and preservation of the cultural heritage, helping to make breakthroughs in processing and analysing spectroscopic data as a reference tool in the scientific exploitation and interpretation of spectroscopic data from art materials.

Finally, a Raman characterisation of the most widespread artists' blue pigment -copper phthalocyanine blue (CuPc)- was performed using ***PigmentsLab***. Analysing the crystalline structures of CuPc (α-, β- and ϵ-modifications of CuPc) may provide chronological markers to date and authenticate art works, but also may provide guidance to art conservators in painting cleaning through solvents application. In this sense, the CuPc is partially soluble in aromatic solvents, and applying those solvents may produce crystallisation defects resulting from an α-to-β-modification of CuPc transformation.

Two different cases were analysed. In the first case, the study of dry pigments was analysed through experimental Raman spectra from hand-made samples where the class-membership was *a priori* known. ***PigmentsLab*** was used to process the Raman spectra from the hand-made samples and successfully classified them in an automated fashion, demonstrating the consistency of the implemented methodology. In the second case, the effect of several pre-selected solvents and cleaning agents on the crystalline structures of CuPc was analysed through a set of Raman spectra from a Raman mapping analysis. This analysis was performed on samples built upon different polymorphic forms of copper phthalocyanine blue under solvents and cleaning agents. The automated data interpretation process and visualisation routines implemented in ***PigmentsLab*** depicted an irreversible transformation from α- to β-modification of CuPc occurred when the solvents toluene and xylene were applied to the unstabilised α-modification of CuPc, confirming the transformation effect reported in the literature through Raman spectroscopy. Attending to the obtained results, the usage of aromatic solvents such as toluene and xylene should be avoided if CuPc is present in the paint layer.

Summarizing, the usage of the designed, developed and analysed automated methodologies integrated in the implemented global system for the automated data interpretation of spectra from art works analysis, ***PigmentsLab***, can play a good auxiliary role in the analysts' endpoint interpretation, providing insight from the raw spectral measurements into pigments. The implementation is an easy-to-use system and straightforward to update when new spectral data become available. The system has great potential as an accurate and practical method for the automated interpretation of Raman spectra for not only pigment analysis, but essentially for any material group.

7.2 Future work

From the research developed in this Ph.D. Thesis several promising topics for future work were raised. The following list compiles preliminary directions for these new research topics.

1. Compilation of a larger high-quality reference spectral database

The compilation of a larger high-quality reference spectral database is indispensable to proceed in all the topics for future work. In addition to acquiring more data, the larger high-quality reference spectral database to be compiled may include high-quality spectra coming from different spectroscopic techniques (mainly X-Ray Fluorescence (XRF), X-Ray Diffraction (XRD), Laser-Induced Breakdown Spectroscopy (LIBS),

InfraRed Spectroscopy (IR), Raman spectroscopy and Surface-Enhanced Raman Spectroscopy (SERS)).

Eventually, the database should include all the spectra from the reference pigments used in art, including all artistic movements from prehistory to nowadays. This implies that the database should be dynamically updated in order to include the spectra from new materials that may be developed. Hence, the database will provide increased value to the community devoted to the analysis of cultural heritage, gathering together all the information (art historic and spectroscopic) from artistic materials in a common reference framework.

As part of increasing the value provided by a larger high-quality reference spectral database, one may complete not only the labelling of already existing spectral data but the new spectra to be measured from reference pigments as well. In this sense, the following research lines will benefit from a larger high-quality reference spectral database:

- Including reference spectra from all the materials used in art works may increase the system robustness in the recognition stage, avoiding ambiguities in a systematic and objective way

- Extending the reference datasets used to train the system may allow the handling of more classification cases

An example of benefit from a larger database is based on extending the reference spectra for classification from the power and vibrant cadmium-based pigments, which were beloved of masters including Monet, Matisse, Cézanne and Dalí. These pigments face a ban recently raised in the European Union thanks to its potential toxicity. Therefore, the identification and classification of cadmium pigments will be a top priority in the near future. In this sense, the automated methodologies presented in this Ph.D. Thesis may serve as a helpful tool in the analysis of cadmium pigments. Hence, pigment powders of reference cadmium-based pigments were newly purchased, including cadmium yellow and cadmium red from different manufacturers (such as Winsor and Newton), which need to be measured and labelled.

2. Algorithmic improvements

Several improvements may be included in the automated data processing analysis chain in order to increase the robustness of the overall automatic recognition methodology. Specifically, two main improvements may be developed:

- **2.1 Application of clustering analysis to the reference spectral library:** Including a pre-processing stage in the reference spectral library projection based

on clustering analysis may be useful to better handle similar reference Raman spectra in the PCs space from pigments from the same chemical category for instance, easing the identification of unknown Raman spectra of such instances

- **2.2. Deblending process improvement in spectral recognition:** The ICA-based approach for spectral deblending in the generalised recognition methodology of Raman spectra may be improved through band modelling through different profiles such as Lorentzian, Gaussian or Voigt functions. This band modelling may simplify the handling of multicomponent Raman spectra as suggested by preliminary analysis performed using Lorentzian-based band modelling where the separation of components with overlapping fundamental bands did not generate information loss

3. Framework improvements

In general software systems, the data is usually picked and transferred to the place where processing happens and then the data is shown to the user. Downloading data from a given repository may consume much of an institution's network resources and severely impact the overall analysis time. Storing and analysing big volumes of data is becoming a challenge not just for independent researchers but for large-scale research centres as well. By greatly reducing both the download time and the storage size of the data, it is demonstrated that the "big data" paradigm of *moving computation to the data* can be of practical interest.

Consequently, the main platform infrastructure of the global system of automated interpretation of spectra from artistic pigments developed in this Ph.D. Thesis platform infrastructure may be updated to a remote server-based system. This way, the data may be located in a single repository node, therefore reducing the time devoted to data downloading and the disk usage of the analyst' computer. Besides, the data processing (mainly noise filtering and identification plus classification) may be performed on demand in the server node, reducing the CPU processing load of the end-user.

This new framework implies the study of two topics. First, the different architectures of data location need to be analysed in order to develop a fault tolerant system in terms of data redundancy. Finally, job scheduling systems should be deployed in the processing node in order to serve to the users appropriately in terms of preventing concurrency data access issues.

4. Extend Raman characterisation of solvent-unstable pigments

Paint layers in art works may acquire a wide variety of deposits or coatings in their lifetime, any of which might be considered to harm their aesthetic or historical integrity therefore justifying removal. Hence, cleaning of paintings by removing varnish layers are activities generally performed in the art conservation field. In this sense, the prime objective of art conservators is to reveal the original paint layer, whilst reducing the associated risks regarding the health of the art conservators. The application of specific solvents and cleaning agents may produce risk to the integrity of the original -even with the mildest cleaning agents there will always be some risk of damaging the art object.

In order to prevent permanent damage resulting from cleaning activities the identification of pigments in art works should be carried out. Hence, successful painting cleaning is based on identifying a cleaning agent which is able to remove the coating without affecting the underlying art materials. Good practice in painting cleaning relies on a risks evaluation and on selecting proper cleaning agents and strategies through careful documentation. As a result of the above-described issue, two new lines of research were raised:

1. Knowing which pigments are solvent-unstable is a first-pass analysis. Thus, retrieving a list of solvent-unstable pigments used in art needs to be performed. The existing literature and materials documentation may provide support to perform this activity. Once this list is compiled, a Raman characterisation (like the one described in this Ph.D. Thesis) should be performed in order to evaluate the risks of applying cleaning agents to provide feedback to art conservators in terms of suggestions on which solvents to apply depending on the original art materials present in the paint layer

2. Additional Raman analysis of polymorphic forms of copper phthalocyanine blue may be performed under further solvents and cleaning agents. In this sense, a total of 20 new solvents and cleaning agents generally used by art conservators were newly purchased. The effect of these solvents should be evaluated through a Raman characterisation, taking special care of any crystalline structure transformation

5. Add support to disaggregating methodologies

The target software platform may include data from different spectral sources. The combination of such data may provide great insights into pigments. Therefore, the global system of automated data interpretation developed in this Ph.D. Thesis may include support to disaggregating methodologies and data fusion methods. Hence,

increased value will be provided based on designing, developing and implementing automated interpretation approaches based on combining spectral data coming from different spectroscopic techniques to help the spectroscopists in the decision-making process. In this sense, signal processing techniques should be explored to implement data fusion and disaggregating methodologies in order to extract patterns from data including association rule learning for discovering interesting relationships, i.e. "association rules", among variables in large databases.

Appendices

Appendix A

Noise filtering methodology: Performance analysis

A.1 Analysis on simulated Raman spectra

The proposed noise filtering methodology was tested using simulated spectra. Shot noise was simulated from a zero-mean Gaussian distribution and variable variance. Also, different artificial profiles were simulated to mimic the fluorescence's baseline, which were selected in a heuristic way but similar in appearance to that of Raman spectra. In particular, four simulated profiles were used: polynomial, linear, sigmoidal and sinusoidal.

Comparisons were performed with the proposed filtering approach and several techniques in common use to filter the shot noise, such as the Wiener filter, the median filter, the wavelet filter, the Fast Fourier Transform (FFT) filter, and the fuzzy filter. For the Wiener filter the noise was estimated from the ideal spectra, and the response function was used with no smearing. The median filter was run several times with window sizes ranging from 3 to 11 data points and the window providing the lowest RMSE between the denoised and the ideal spectra was selected. The wavelet filter was performed by means of the standard wavelet soft thresholding with default parameters. The FFT filter was run several times with rectangular filters of different sizes, selecting the one providing the lowest RMSE.

Additionally, comparisons with respect to baseline rejection were carried out with the here proposed filtering methodology, the morphology-based filtering approach published in[51] and the conventional polynomial approach, being the last one the most popular method in Raman spectroscopy for subtracting the fluorescence's baseline. The conventional polynomial method was run several times selecting the polynomial degree that provided the lowest RMSE.

Unlike the proposed methodology, the general techniques in common use tested are focused on either shot noise filtering or baseline rejection. Therefore, to perform a proper comparison with respect to the filtering approach presented in the current paper, the previously commented shot noise filtering techniques were combined with the baseline filters above-mentioned. In particular, 100 noisy spectra were simulated and the RMSE between the ideal and the filtered spectra was computed to compare the results of the here proposed approach and each of the combinations of a shot noise filtering technique with a baseline filter.

The results, i.e. mean RMSE and standard deviation, are shown in Table A.1 - the best-degree polynomial filter is represented as PF and the morphology-based filter[51] is represented as MF. On average, at the noise levels tested the here presented method outperforms the combination of the other techniques. From the results we may also say that the proposed filter provides the least distortion of the Raman bands, which is very useful when the spectrum must be subsequently processed in order to identify the material or quantify its proportion in mixtures. Additional test results using simulated spectra can be found in the Fig. A.2 and Fig. A.1.

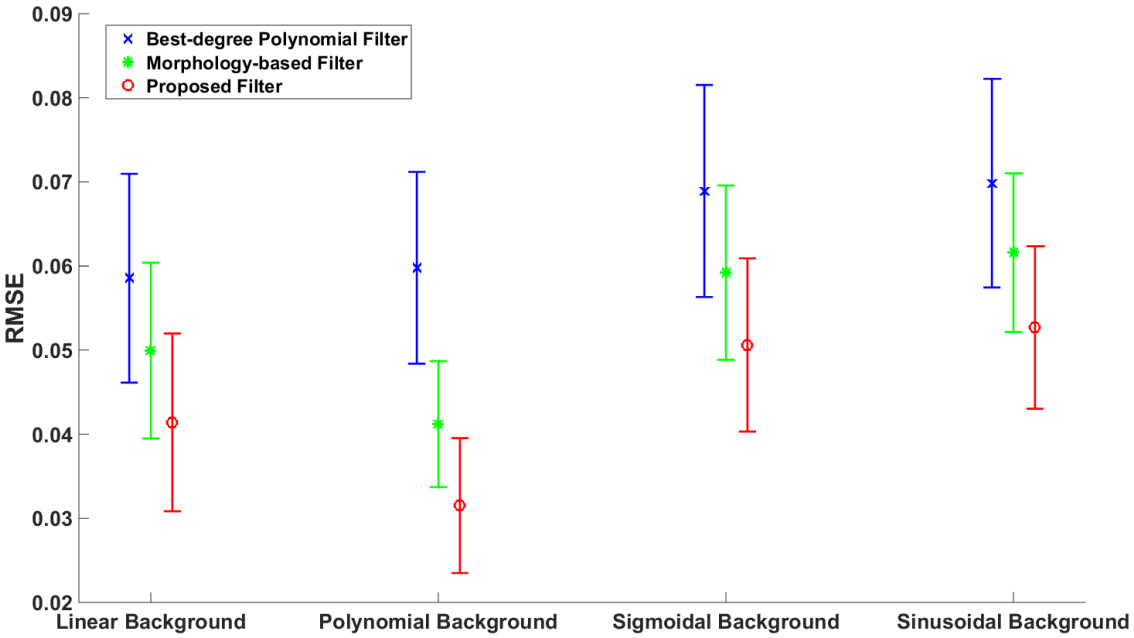


Figure A.1: Comparison of the performance on simulated spectra of the best-degree polynomial approach, the morphology filter, and the presented noise filtering method

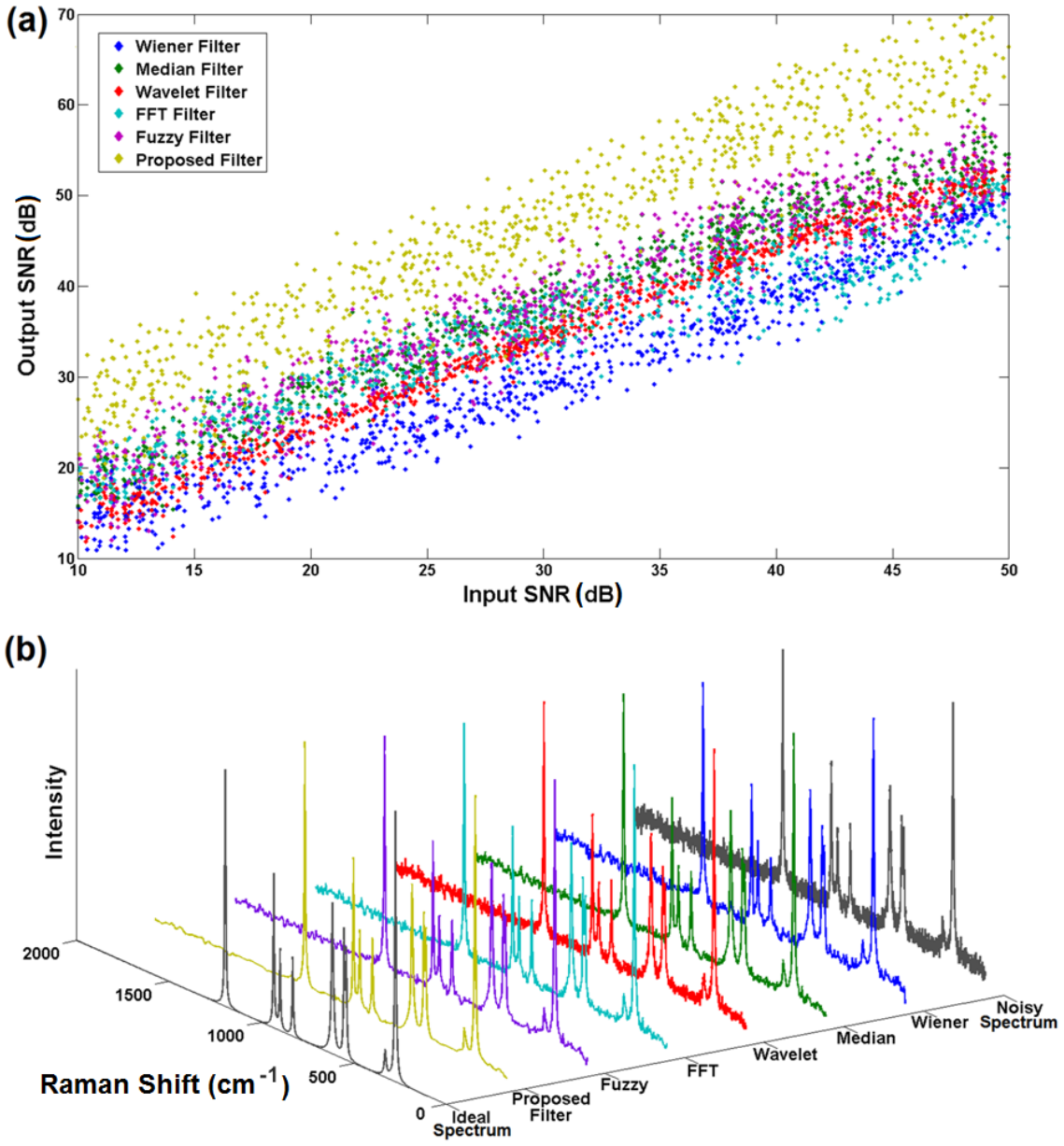


Figure A.2: a) Comparison of the performance of several shot noise filtering techniques on simulated spectra, b) shot noise filtering examples of a noisy simulated spectrum (SNR=10dB) being filtered by Wiener, median, wavelet, FFT, fuzzy and the proposed method

Table A.1: RMSE between ideal and filtered spectra using the proposed approach, and combinations of a baseline filter (conventional best-degree polynomial filter, PF, and morphology filter, MF) with a shot noise filter (Wiener, median, wavelet, FFT and fuzzy filters), using simulated spectra with different baseline profiles (linear, polynomial, sigmoidal and sinusoidal)

	Linear baseline	Polynomial baseline	Sigmoidal baseline	Sinusoidal baseline
Proposed filter	0.0319 ± 0.0097	0.0408 ± 0.0096	0.0230 ± 0.0070	0.0402 ± 0.0102
PF+Wiener filter	0.0560 ± 0.0169	0.0577 ± 0.0176	0.0543 ± 0.0356	0.0768 ± 0.0326
PF+median filter	0.0575 ± 0.0140	0.0703 ± 0.0216	0.0640 ± 0.0235	0.0806 ± 0.0310
PF+wavelet filter	0.0503 ± 0.0087	0.0640 ± 0.0189	0.0565 ± 0.0205	0.0736 ± 0.0288
PF+FFT filter	0.0465 ± 0.0122	0.0498 ± 0.0151	0.0470 ± 0.0137	0.0522 ± 0.0178
PF+fuzzy filter	0.0528 ± 0.0136	0.0655 ± 0.0211	0.0593 ± 0.0227	0.0757 ± 0.0302
MF+Wiener filter	0.0350 ± 0.0066	0.0451 ± 0.0067	0.0361 ± 0.0065	0.0448 ± 0.0066
MF+median filter	0.0486 ± 0.0124	0.0486 ± 0.0124	0.0497 ± 0.0124	0.0482 ± 0.0124
MF+wavelet filter	0.0416 ± 0.0072	0.0423 ± 0.0073	0.0423 ± 0.0070	0.0417 ± 0.0072
MF+FFT filter	0.0447 ± 0.0129	0.0448 ± 0.0138	0.0460 ± 0.0119	0.0456 ± 0.0118
MF+fuzzy filter	0.0449 ± 0.0118	0.0450 ± 0.0119	0.0461 ± 0.0120	0.0447 ± 0.0128

A.2 Analysis on experimental Raman spectra

To show the performance of the implemented methodology in realistic environments, we applied the developed method to Raman spectra from art works. In particular, some of the experimental Raman spectra used in this research were kindly provided by Nadim C. Scherrer from the Bern University of Applied Sciences. The experimental Raman spectra measured by the author used were acquired following the procedure described in Sect. 3.2..

Fig. A.3 presents some real-case examples of experimental Raman spectra measured from works of art, for which the proposed noise filtering technique was applied. These Raman spectra were acquired from different art works and therefore they show different shot noise realisations and different shapes of fluorescence's baseline. Specifically, the Raman spectrum before (in black) and after applying the proposed noise filtering methodology (in grey) are shown in all pictures. As it can be seen, in all the examples the Raman band shapes and positions were unchanged, and also their intensity ratios were maintained while reducing the shot noise and rejecting the baseline. Table A.2 shows a comparative on the experimental Raman spectra presented in Fig. A.3 carried out in the same way as for the simulated spectra. The performance of the different methods can be seen in Fig. A.4, A.5, A.6, A.7, A.8, A.9. The here proposed filtering approach provided the highest signal-to-noise ratio compared to the combination of conventional denoising techniques. As it can be seen, the Raman bands were visibly enhanced in the denoised spectrum.

Table A.2: SNRs of the denoised experimental Raman spectra using the proposed noise filtering approach, and combinations of a baseline filter (conventional best-degree polynomial filter, PF, and morphology filter, MF) with a shot noise filter (Wiener, median, wavelet, FFT and fuzzy filters)

	Spectrum a	Spectrum b	Spectrum c	Spectrum d	Spectrum e	Spectrum f
Proposed filter	28.1873	22.4613	32.3791	21.2516	31.8763	39.4224
PF+Wiener filter	16.6748	11.8768	8.9701	17.4773	27.3828	23.1225
PF+median filter	16.6137	19.0401	9.2666	19.1921	30.6203	27.9877
PF+wavelet filter	15.6451	16.5602	9.3726	15.9537	29.8357	27.1419
PF+FFT filter	11.5824	14.5842	11.7827	14.0882	12.3609	24.8403
PF+fuzzy filter	16.7221	21.7162	18.5150	19.7585	30.8861	29.5737
MF+Wiener filter	20.0519	9.6524	23.9501	8.5052	30.5533	24.6272
MF+median filter	21.2140	13.2364	26.8809	14.6687	30.7646	28.6767
MF+wavelet filter	21.0518	12.0132	26.8187	11.6223	30.6575	27.2624
MF+FFT filter	21.1334	14.2837	23.6079	12.2891	30.7159	28.1590
MF+fuzzy filter	22.7607	15.2989	27.5784	15.3349	31.2740	29.6025

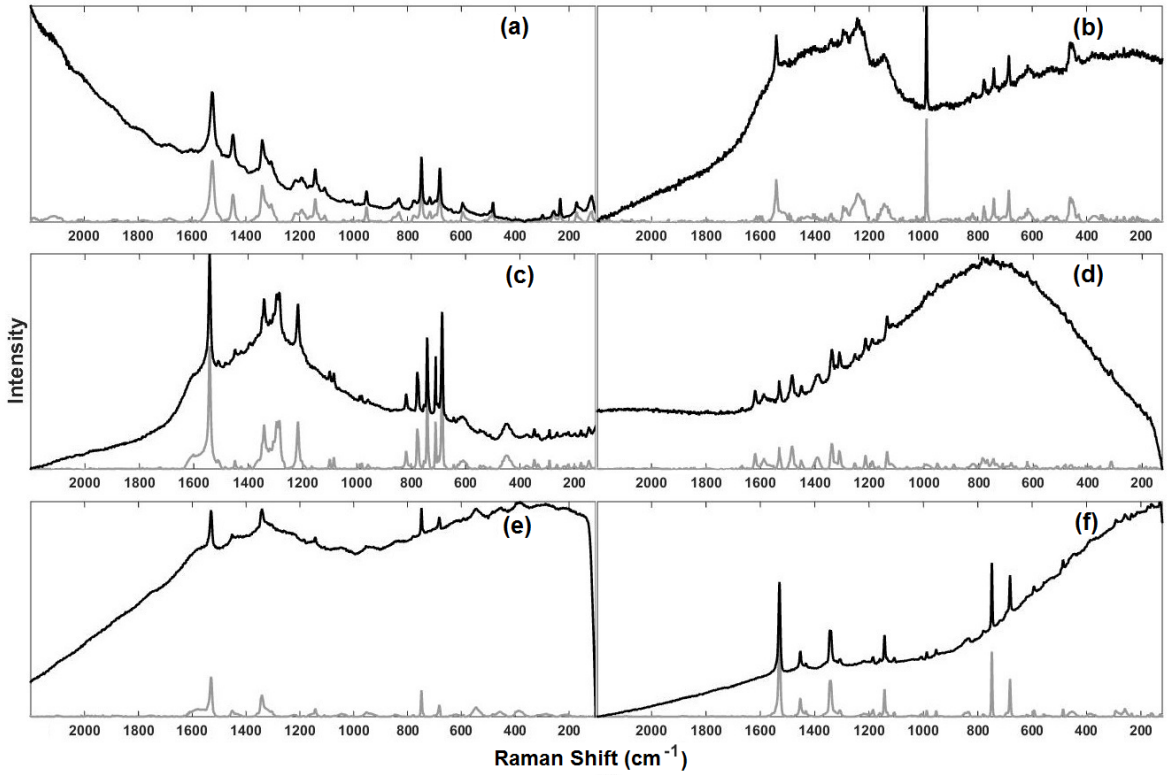


Figure A.3: Examples of experimental Raman spectra measured in art works, prior (in black) and subsequent (in grey) to apply the proposed noise filtering methodology: (a) copper-phthalocyanine blue, (b) mixture of calcite and a copper-phthalocyanine blue, (c) mixture of rutile and copper-phthalocyanine green, (d) mixture of a copper-phthalocyanine blue, carbon black and rutile, (e) mixture of a PY1, a PR4 and a copper-phthalocyanine blue, and (f) copper-phthalocyanine blue

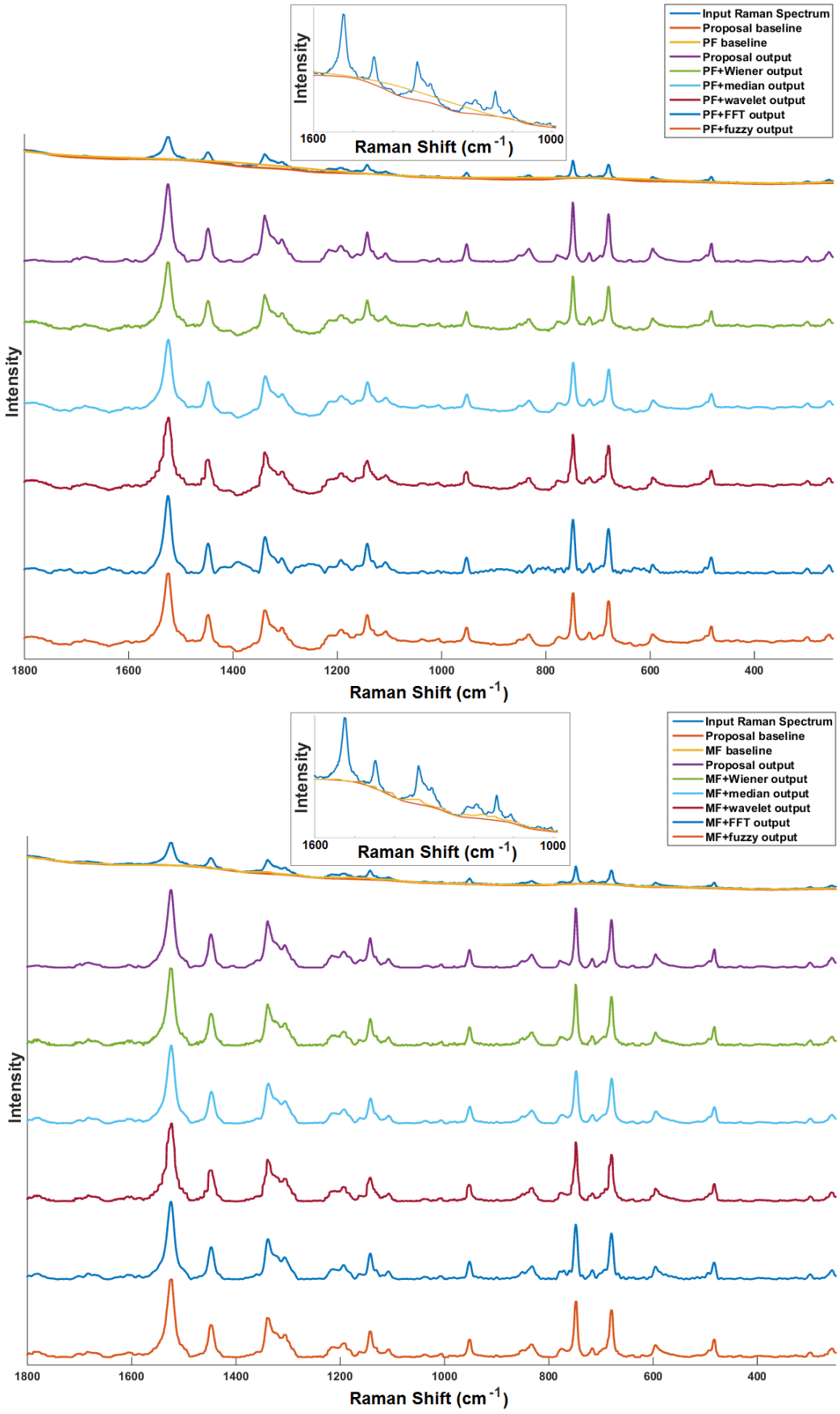


Figure A.4: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (a)

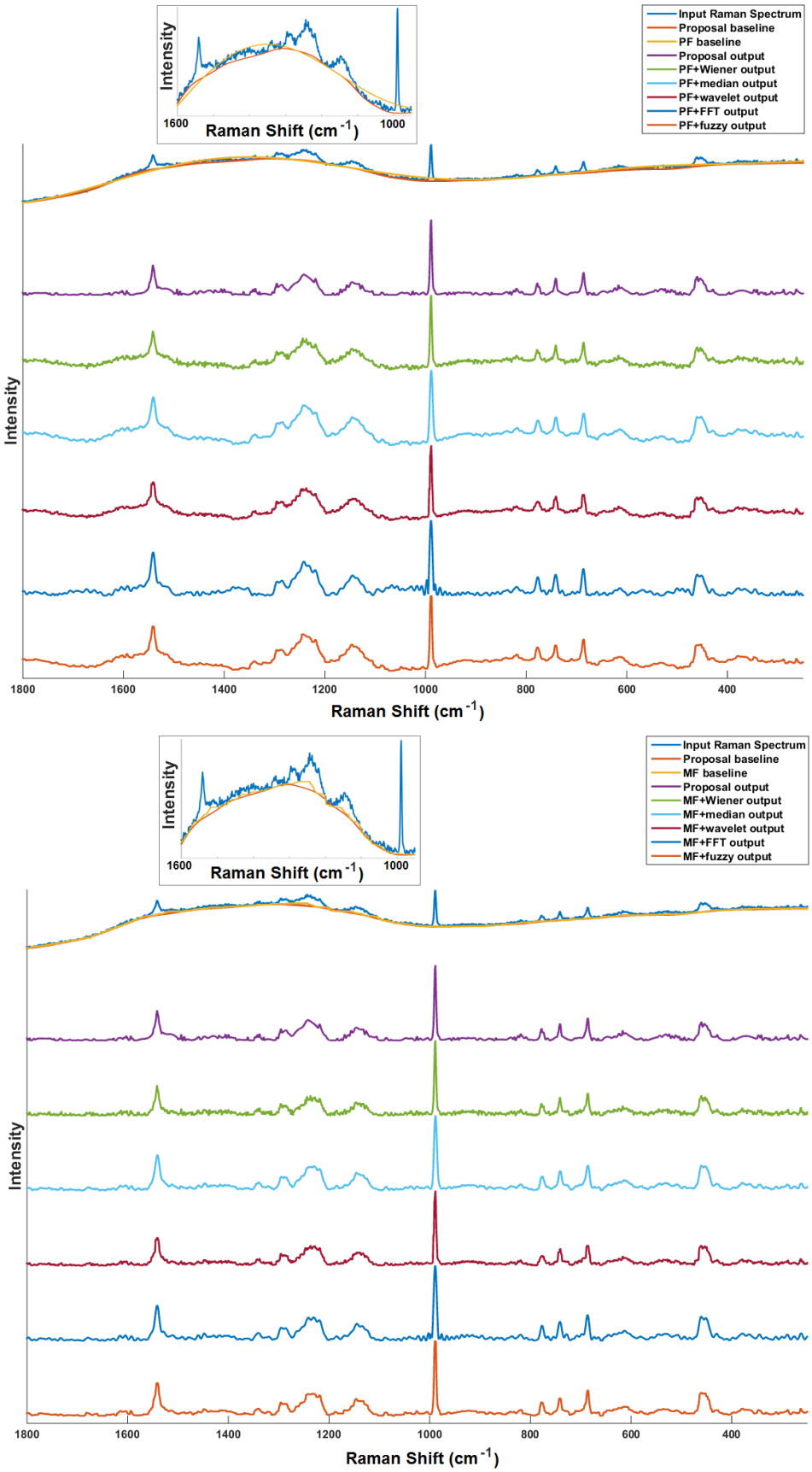


Figure A.5: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (b)

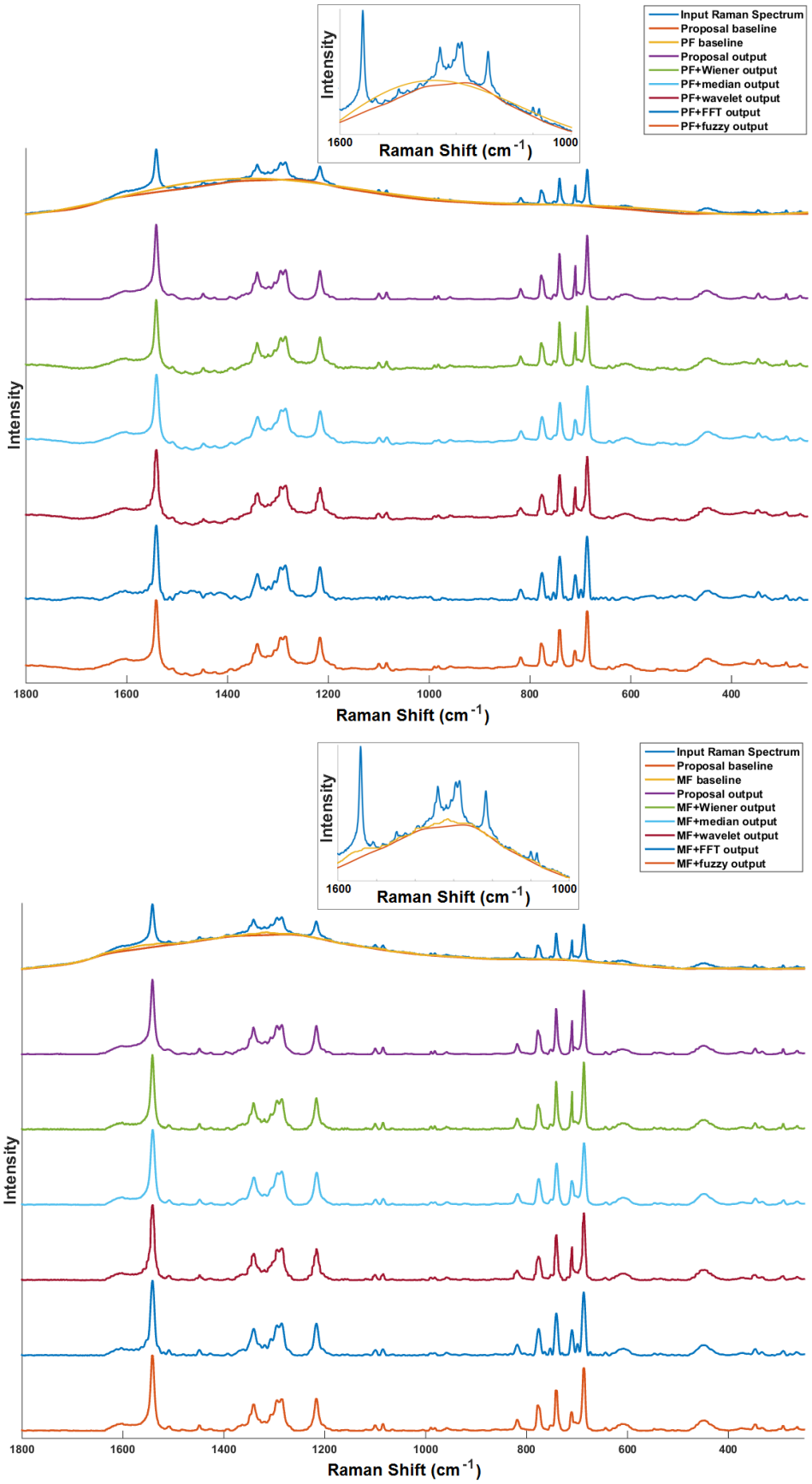


Figure A.6: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (c)

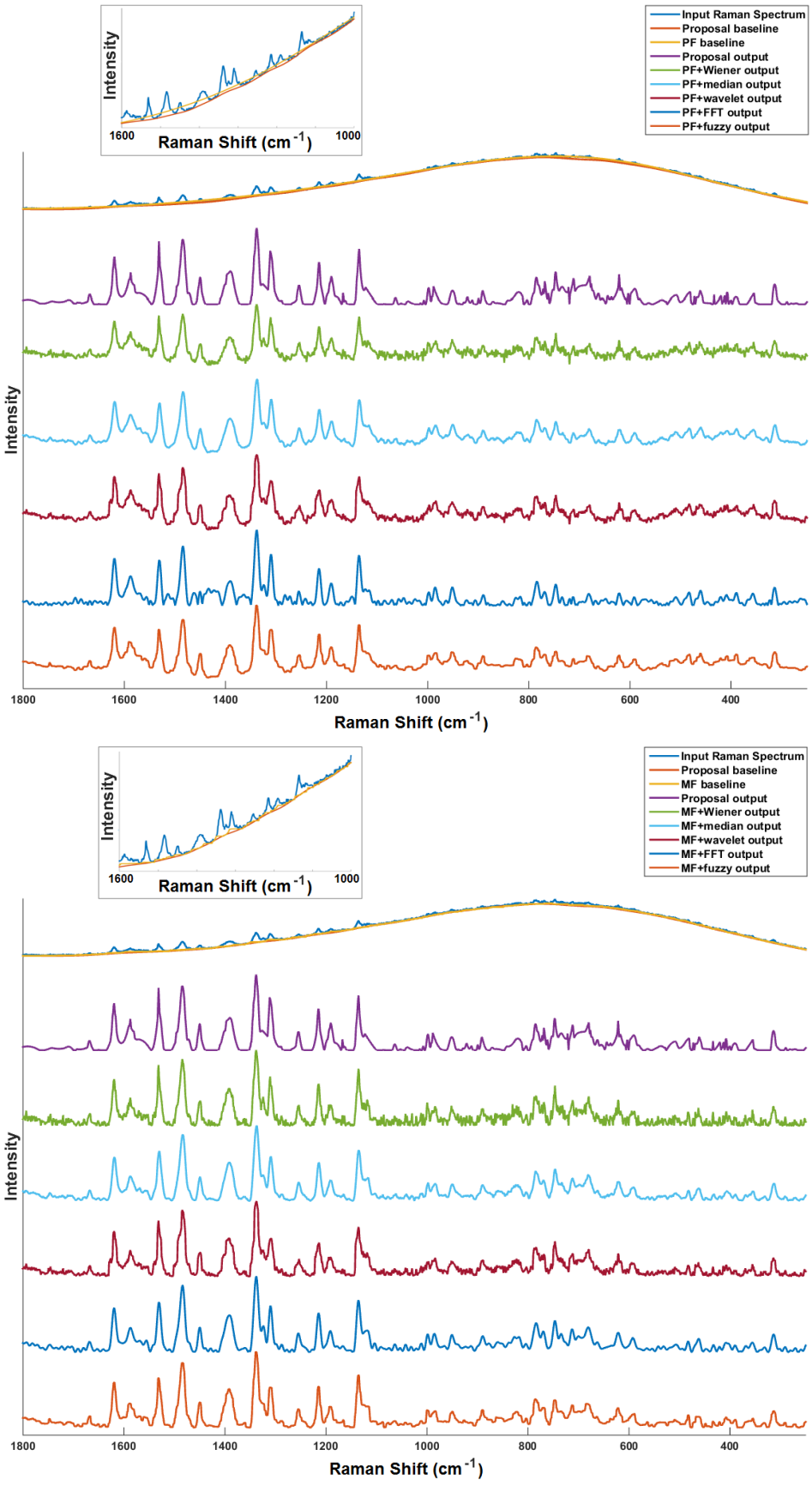


Figure A.7: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (d)

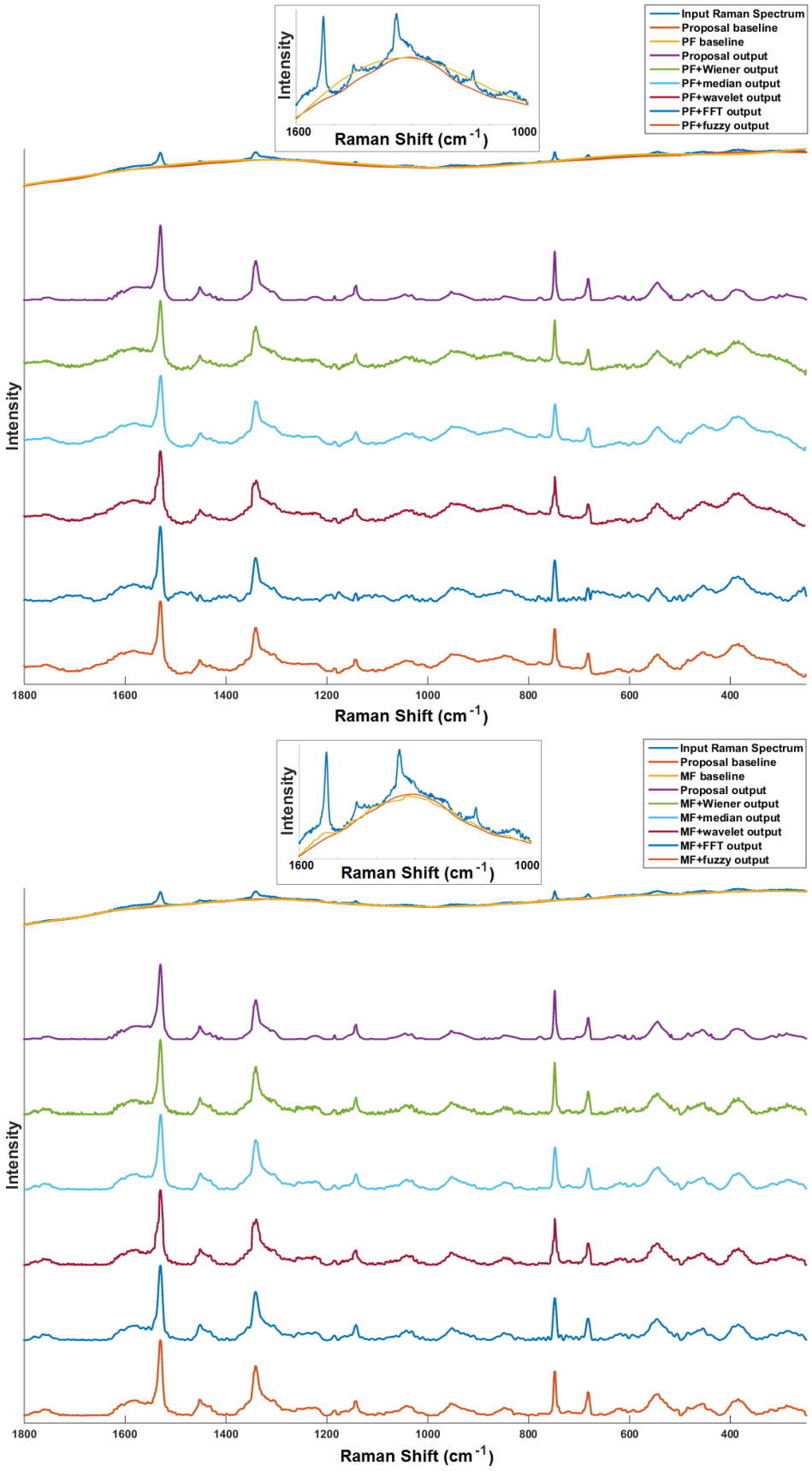


Figure A.8: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (e)

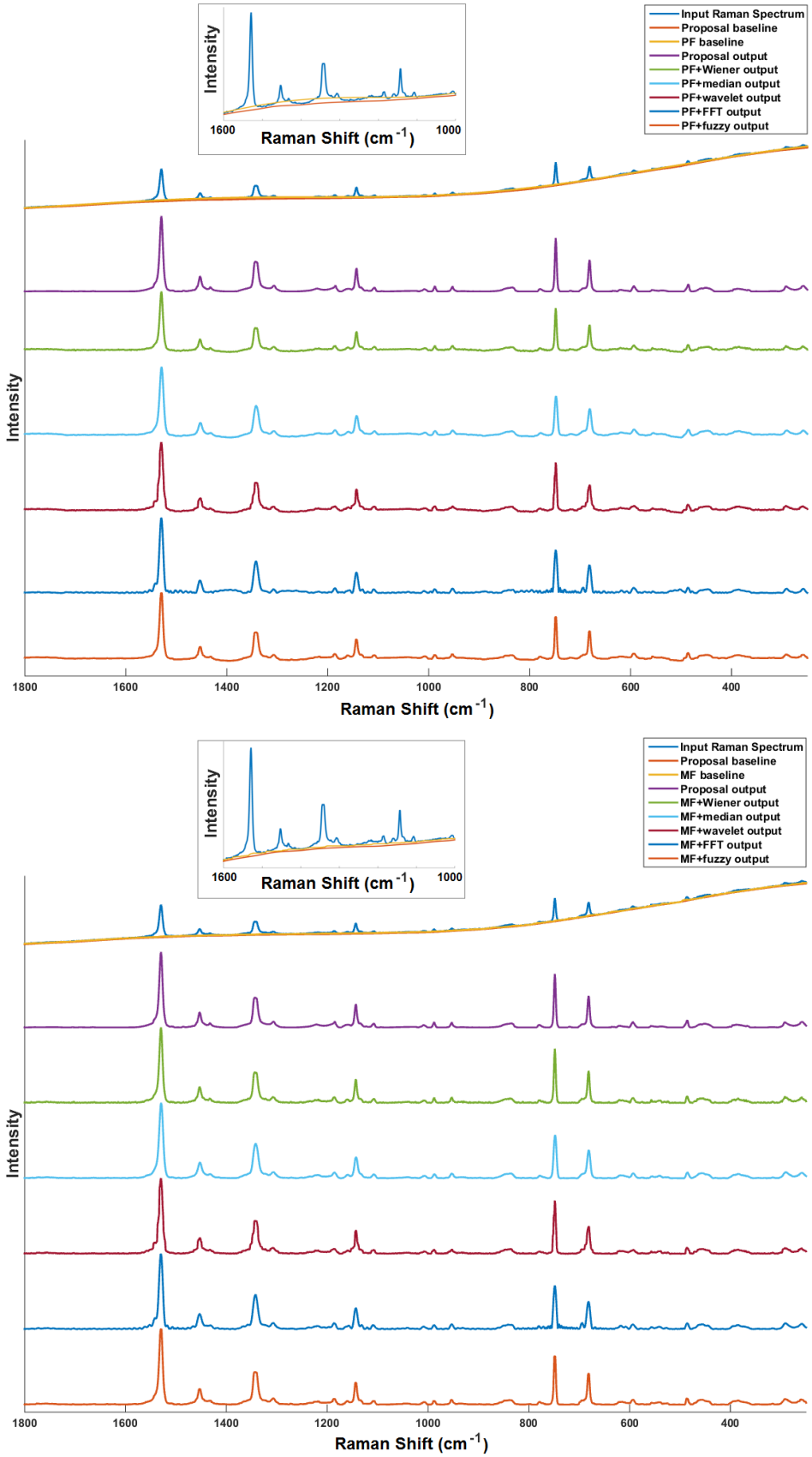


Figure A.9: Proposed filter versus conventional best-degree polynomial baseline filter (PF) -top- and morphology baseline filter (MF) -bottom- plus Wiener, median, wavelet, FFT and fuzzy filters applied to the experimental Raman spectrum (f)

These tables and figures provide a qualitative visual inspection of the performance of the noise filtering methodology presented in the current work. The denoising method reduced the influence of shot noise and removed the fluorescence's baseline without changing the shapes or positions of the Raman bands, maintaining their intensity ratios. The results show the effectiveness of the proposed denoising methodology as a fully-automated tool, that is, without requiring any user input, to help the analyst in the interpretation of Raman spectra.

Appendix B

Reference Spectral Library characterisation

B.1 Feature extraction: comparative analysis

A comparative study of techniques for dimensionality reduction is presented in this section. In particular, the most often four standard techniques were analysed: PCA, ICA, LDA and PLS.

PCA The basic idea of PCA is to decorrelate the input dataset[126]. So the starting point is the covariance matrix $\boldsymbol{K}$ defined by $\boldsymbol{K} = \boldsymbol{X}^t \cdot \boldsymbol{K}$. Because of construction c_{ij} is equal to c_{ji}, K is diagonalizable

$$\boldsymbol{D} = \boldsymbol{U}^t\boldsymbol{K}\boldsymbol{U} = \boldsymbol{U}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{U} = (\boldsymbol{X}\boldsymbol{U})^t(\boldsymbol{X}\boldsymbol{U}) = \boldsymbol{X}'^t\boldsymbol{X}'$$

So, if the data $\boldsymbol{X}$ is mapped by the orthogonal matrix $\boldsymbol{U}$, the resulting data matrix $\boldsymbol{X}$ is decorrelated; thus, $cor(\vec{x_i'}, \vec{x_j'}) = \mathbf{0}$ if $i = j$. However, no further information is used and therefore the technique works without supervision.

ICA The idea of ICA is very similar to the PCA with the difference that the goal is the achievement of 'non-gaussianity' of the dataset[69]. The assumption is that noise would result in a Gaussian behavior of the dataset. Information on the other hand shows a non-Gaussian behavior. Several methods and variants of the ICA exist - the so-called fastICA version is used here, where the neg-entropy estimates the 'non-gaussianity' of the dataset and a Newton fixed-point iteration is used. As well as PCA, ICA works without supervision.

LDA LDA[127] is generally used for classification, but it also can be used for dimensionality reduction. The method is supervised due to the fact that the class membership of the classes has to be known. The scatter matrices $\boldsymbol{F_i}$ are calculated by

$$\boldsymbol{F_i} = \sum_{\vec{x_j} \in \boldsymbol{X_i}} (\vec{x_j} - \vec{\mu_i})(\vec{x_j} - \vec{\mu_i})^t$$

where $\vec{x_j} \in \boldsymbol{X_i}$ are spectra from the i-th class and $\vec{\mu_i}$ the mean spectrum of $\boldsymbol{X_i}$. The scatter matrix is computed by

$$\boldsymbol{F_B} = \sum_{i=0}^{k} \boldsymbol{N_i} (\vec{\mu_i} - \vec{\mu})(\vec{\mu_i} - \vec{\mu})^t$$

where $\boldsymbol{N_i}$ is the size of the i-th class and $\vec{\mu}$ is the mean of the whole dataset. The inter-class variability has to be maximized over the variability of the whole dataset. The linear transformation is obtained through

$$\boldsymbol{W} = \text{argmin} \frac{det(\boldsymbol{W^t F_B W})}{det(\boldsymbol{W^t F W})}$$

PLS The basic idea of the supervised method of PLS[128] dimension reduction is the decomposition of the input matrix $\boldsymbol{X}$ and the output matrix $\boldsymbol{Y}$

$$\boldsymbol{X} = \boldsymbol{PU} + \boldsymbol{E} \text{ and } \boldsymbol{Y} = \boldsymbol{QV} + \boldsymbol{H}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the score matrices and $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the loading matrices. The matrices $\boldsymbol{E}$ and $\boldsymbol{H}$ are the error terms. The decomposition is constructed in such a way that the covariance of $\boldsymbol{P}$ and $\boldsymbol{Q}$ is maximized and that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices. The orthonormal matrix $\boldsymbol{V}$ is used for dimension reduction.

Comparative analysis The comparative analysis was performed using simulated spectra. Specifically, 100 datasets were simulated containing 3 different categories each dataset and 25 spectra per category. For each dataset, the four projections were applied (see an example in Fig. B.1). For each projection, the computation time was measured and the category separability was computed through the JMD (see Eq. 4.5). Table B.1 compiles the resulting time and JMD (mean and standard deviation) for each data reduction technique. From the results obtained, we may conclude that in the case of Raman datasets in a simulated environment the traditional linear dimensionality reduction technique of PCA outperforms the other data reduction techniques that were tested.

Table B.1: Mean time and JMD and standard deviation using simulated data and the most often used techniques in Raman spectroscopy

Technique	Time [ms]	JMD
PCA	77.76±8.87	2.00±0.00
PLS	81.13±32.08	1.98±0.07
LDA	1802.23±324.15	1.82±0.13
ICA	5402.23±1237.41	1.79±0.41

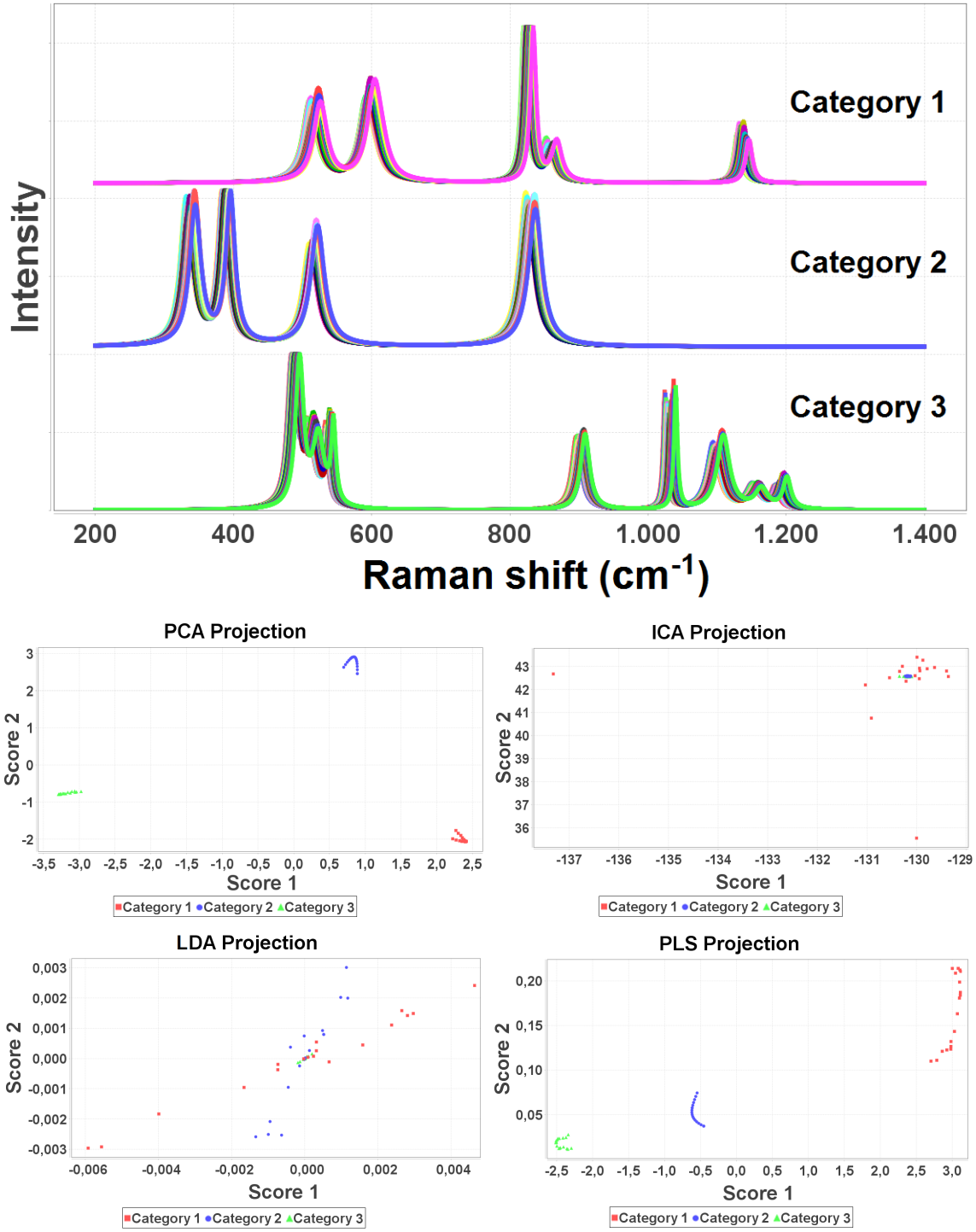


Figure B.1: 2D score space obtained through PCA (middle left), ICA (middle right), LDA (bottom left) and PLS (bottom right) applied to an input dataset of simulated Raman spectra of 3 categories (top)

B.2 Database characterisation

Identification of pigments in art works is indispensable to determine correct conservation treatments or to answer questions regarding authenticity issues. Since the early 20th century, the introduction of synthetic organic pigments has exponentially increased the number of artistic pigments. To ensure a correct identification an as complete as possible library of reference Raman spectra is needed. A subset of the spectral library used in this research was composed of the collection of Raman spectra of synthetic organic pigments published in[124].

Specifically, the reference spectral library is composed of over 250 synthetic organic pigments, represented by their colour index name, were collected from research institutes, pigment manufacturers, resellers and artist's paints makers. The exact source of the spectra can be found on `http://modern.kikirpa.be`, together with the instrumental parameters that were used to obtain the spectrum.

An overview of the synthetic organic pigments that were studied and their chemical classification according to[175] is given in Table B.2, following the research results published in[124]. Specifically, the Raman spectra of this collection were acquired using a 785-nm near-infra-red excitation source. Hence, the published Raman spectra of pigments allowed studying pigments with similar structure, which produce only slightly different spectra. This is due to the fact that pigments belonging to a given chemical class are usually characterized by some common Raman bands, a feature that may be used to discriminate different pigment classes.

Table B.2: Overview and classification of the synthetic organic pigments used in this research compiled from the data published in[124]

Pigment class	**Pigments** (*Colour Index*)
	Azo pigments - Disazo pigments
Diarylide yellow	PO16, PY12, PY13, PY14, PY17, PY55, PY63, PY81, PY83, PY87, PY113, PY126, PY127, PY152, PY170, PY172, PY174, PY176, PY188
Bisacetoacetarylide	PY16, PY155
Disazo pyrazolone	PR38, PR41, PO13, PO34
Disazo condensation	PR144, PR166, PR214, PR220, PR221, PR242, PR262, PY93, PY128, PBr23, PBr41, PBr42
Other	PR139

Pigment class	**Pigments** (*Colour Index*)
	Azo pigments - Monoazo pigments
Acetoacetic arylide pigments and lakes	PY1, PY1:1, PY2, PY3, PY4, PY6, PY61, PY62, PY65, PY73, PY74, PY75, PY97, PY111, PY116, PY168, PY169
Pyrazolone pigments and lakes	PY100, PY183, PY191
β-naphthol pigments and lakes	PR1, PR3, PR4, PR6, PR49, PR49:1, PR49:2, PR51, PR53, PR53:1, PO5, PO46
Naphthol AS pigments and lakes	PR2, PR5, PR7, PR8, PR9, PR12, PR14, PR17, PR18, PR21, PR22, PR23, PR31, PR32, PR112, PR146, PR147, PR150, PR170, PR184, PR187, PR188, PR210, PR213, PR223, PR237, PR238, PR239, PR243, PR245, PR247, PR253, PR256, PR258, PR266, PR268, PR269, PO24, PO38, PV44
Benzimidazolone	PR171, PR175, PR176, PR185, PR208, PO36, PO60, PO62, PO72, PY120, PY151, PY154, PY156, PY175, PY180, PY181, PY194, PY214, PV32, PBr25
BONA lakes	PR48:1, PR48:2, PR48:3, PR48:4, PR52:1, PR52:2, PR57:1, PR57:2, PR58:4, PR63:1, PR63:2
Naphthalene sulfonic acid lakes	PR54, PR60, PR60:1, PY104
Metal complexes	PY150, PG8, PG10
Other	PR211, PY213, PV51, PV52
Other azo pigments	PR276, PR277, PO74, PO79, PY205, PY206, PY209, PY209:1, PY210, PY212, PY219
	Heterocyclic (azo)methine pigments
Isoindoline/ isoindolinone	PR260, PO61, PO69, PO86, PY109, PY110, PY139, PY173, PY185, PBr38
Metal complexes	PR257, PR271, PO59, PO68, PY117, PY129, PY153
	Polycyclic pigments
Phthalocyanine	PG7, PG36, PB15, PB15:1, PB15:2, PB15:3, PB15:4, PB15:6, PB16, PB17
Quinacridone	PR122, PR202, PR206, PR207, PR209, PO47, PO48, PO49, PV19, PV42
Perylene/ perinone	PR123, PR149, PR178, PR179, PR190, PR194, PR224, PO43, PV29, PBk31

Pigment class	**Pigments** (*Colour Index*)
Thioindigo	PR88, PR181, PV36
Anthraquinone derivates	PR83, PR168, PR177, PR216, PO51, PY24, PY108, PY147, PB52, PB60, PV5, PV5:1
Dioxazine	PV23, PV37
Triarylcarbonium	PR81, PR81:1, PR81:2, PR81:3, PR81:4, PR81:5, PR169, PG1, PB1, PB14, PB62, PV1, PV2, PV2:2, PV3, PV3:1, PV27
Diketopyrrolo pyrole	PR254, PR255, PR264, PO71, PO73
Quinophthalone	PY138
Other	PR204
	Other classes
Aniline black	PBk1
Aluminium lakes	PR172, PR173, PB63
Pyrazoloquinazo-lone	PR251, PR252, PO67
Other	PO64, PR47, PR279, PR280, PR285, PV53, PBk21

Finally, in addition to the PC1-PC2 projection plot shown in Fig. 4.2 (Chapter 4), the PC2-PC3 and PC1-PC3 projection plots of the PCA transformation applied to the reference spectral library -including both inorganic pigments and the synthetic organic pigments described above- used in this research are compiled hereafter, together with the corresponding biplots. A brief explanation of the symbols (dot styles) used to represent the different chemical classes is shown in Table B.3.

Table B.3: Correspondence between symbols (dot styles) and pigment classes used in Fig. 4.2, B.2 and B.3, according to the classification described in Table B.2

o	Inorganic	☼	Isoindolinone
☆	Diarylide	ϟ	Azomethine Metal Complex
⊙	Bisacetoacetarylide	*	Phthalocyanine
◂	Disazo pyrazolone	♲	Quinacridone
◊	Disazo condensation	◈	Perylene
x	Other disazo	▹	Thioindigo
♣	Acetoacetic arylide	♣	Anthraquinone derivates
★	Monoazo pyrazolone	◐	Dioxazine
♭	Beta-naphtol	◂	Triarycarbonium
◘	Naphtol AS	#	Diketopyrrolo pyrole
□	Benzimidazolone	♠	Quinophtalone
☀	BONA	?	Other polycyclic
♦	Naphtalene Sulfonic Acid	^	Aniline
?	Azo Metal Complexes	◒	Aluminium
★	Other monoazo	▸	Pyrazoloquinazolone
✠	Isoindoline	o	Other

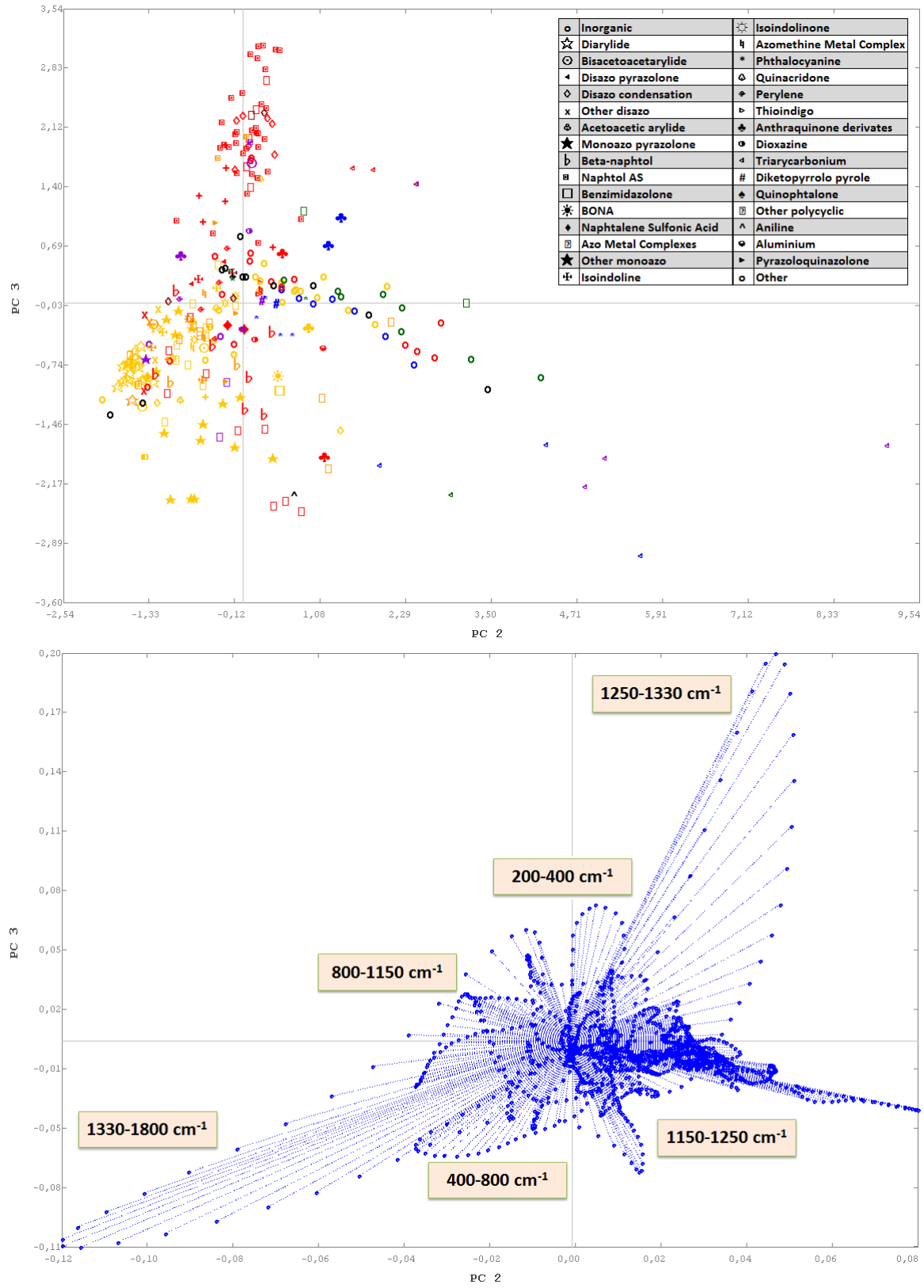


Figure B.2: PC2-PC3 projection (top) and biplot (bottom) of the reference Raman spectra - item styles stand for chemical classes, item colour by *Colour Index*

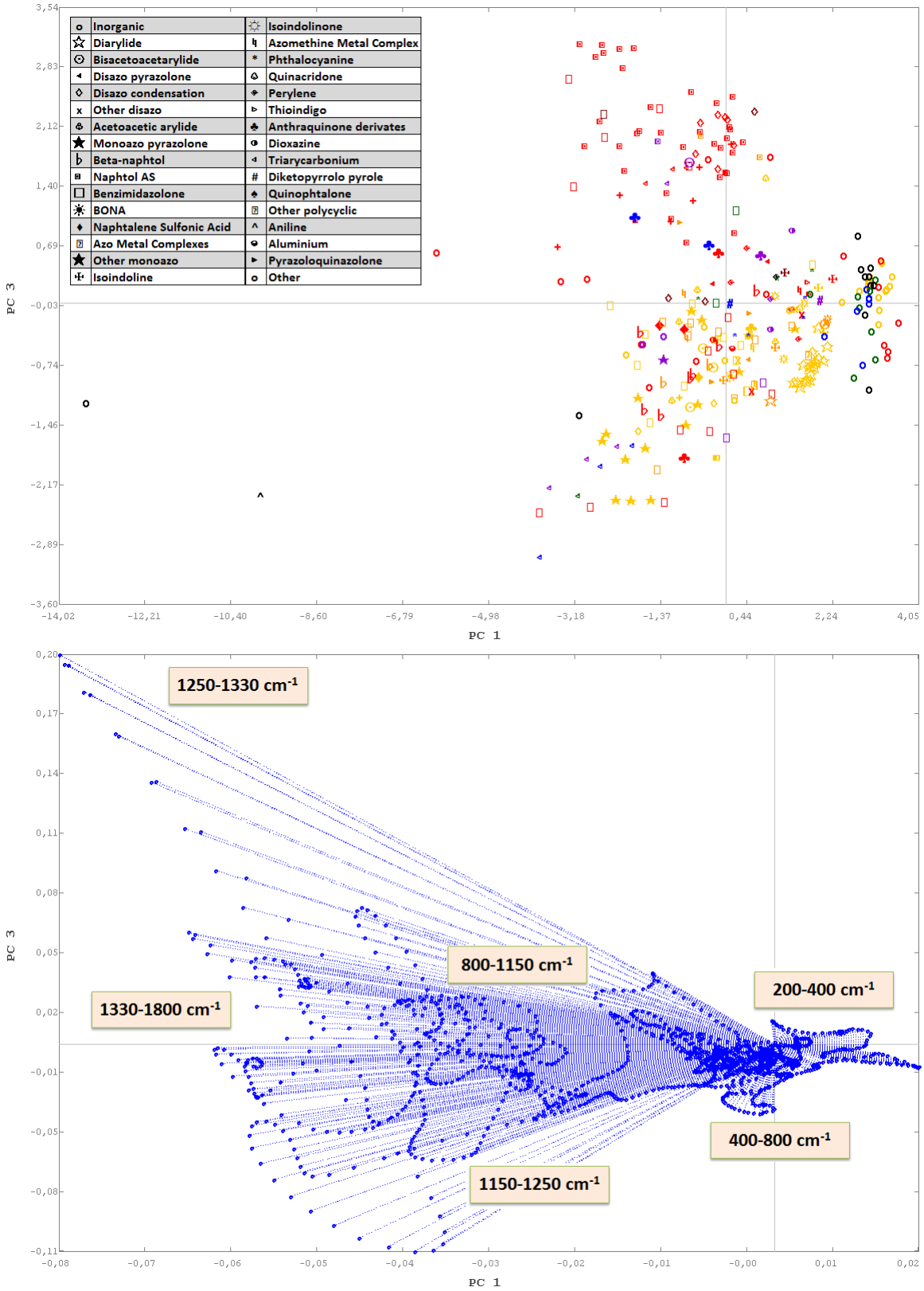


Figure B.3: PC1-PC3 projection (top) and biplot (bottom) of the reference Raman spectra - item styles stand for chemical classes, item colour by *Colour Index*

Appendix C

Automated analysis of Raman spectra: Performance analysis

C.1 Identification of Raman spectra from pigments

C.1.1 Binary mixtures handling-based identification algorithm

Experimental results The binary mixtures handling-based algorithm was applied to four experimental cases. In a first example, the analysed samples was a mixture of rutile and ultramarine blue. When the identification criteria were applied to the spectrum of the measured sample no separated candidates were found. Hence, the mixture-building criteria were applied creating a fictitious mixture with the two patterns which had the lowest ED to the studied spectrum: rutile and ultramarine blue. After applying the identification criteria the result was that the created mixture matched the unknown spectrum from the sample with a MF of MF(Mixture of rutile and ultramarine blue)=72.24%. This result led to conclude that the analysed sample corresponded to a mixture of the pigments rutile and ultramarine blue (Fig. C.1(1)).

The sample analysed in a second example was the mixture of the pigment PY1 and the pigment PR3. When the identification criteria were applied to the spectrum of the measured sample it was found a separated candidate, the pigment PR3 with a MF of MF(PR3)=42.39%. As this MF was lower than the value established to build mathematical spectra of mixtures (60%), the mixture-building criteria were applied creating a fictitious mixture with the two patterns which had the lowest ED to the analysed spectrum: the PY1 and the PR3 pigments. After applying the identification criteria the created mixture spectrum matched the unknown spectrum from the sample with a MF of MF(Mixture of PY1 and PR3)=81.69%. This result concludes that the sample may correspond to a PY1+PR3 mixture (Fig. C.1(2)).

In a third example, the pigment Sennelier 547 (which is not in the reference spectral library) was directly measured and analysed. The label of this pigment indicates that it is manufactured as a mixture of a PR4 pigment and a PY1 pigment. When applying the identification methodology, two separated candidates were found: the PR4 pigment and the PY1 pigment with a MF, respectively, of MF(PR4)=43.87% and MF(PY1)=39.65%, which could correspond to the mixture that is indicated on the label of the Sennelier 547 pigment. Attending to the mixture-building criteria, the system created the mixture from the reference spectra belonging to the PR4 pigment and the PY1 pigment. Then the identification criteria were applied over this fictitious mixture obtaining a MF of MF(Mixture of PR4 and PY1)=60.52%. This result led to conclude that the analysed sample corresponded to a mixture of PR4 and PY1, being a consistent result with the pigment label (Fig. C.1(3)).

In a last example, the analysed analysed was a mixture of PY1 and PB60. When the identification criteria were applied to the acquired Raman spectrum from the sample, two separated candidates were found: the PB60 pigment with a MF of MF(PB60)=19.78% and the PY1 pigment with a MF of MF(PY1)=12.88%. As the two candidates had non-negligible MFs of a same order the system created a mixture of the reference spectra corresponding to the pigment PY1 and the pigment PB60. After applying the identification criteria the spectrum of the created mixture matched the unknown spectrum from the sample with a MF of MF(Mixture of PB60 and PY1)=75.13%. This result led to conclude that the studied sample corresponded to a mixture of PY1 and PB60 (Fig. C.1(4)).

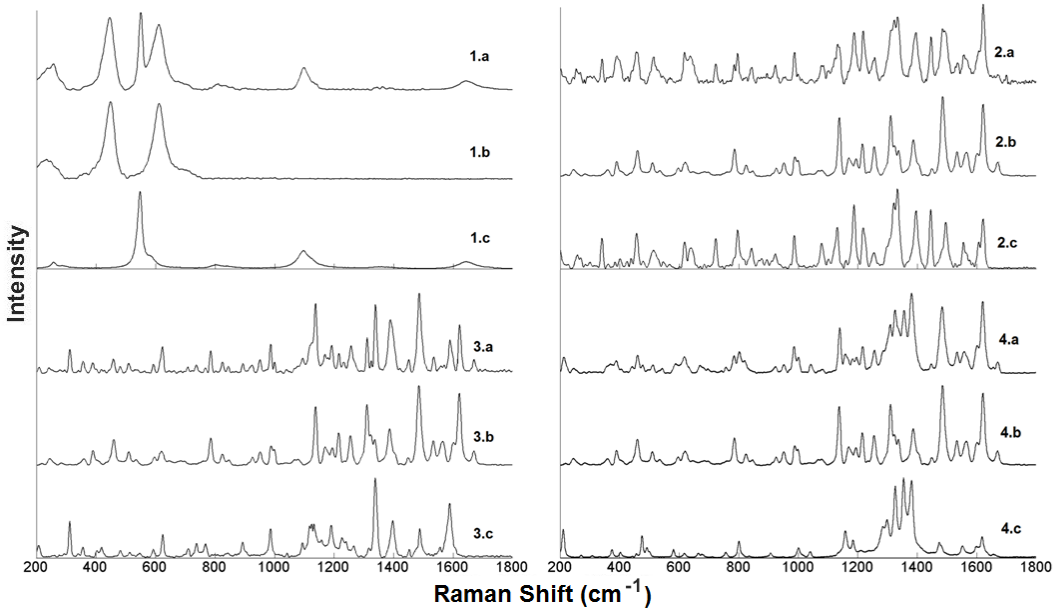


Figure C.1: Unknown (1.a) and reference spectra of rutile (1.b) and ultramarine blue (1.c). Unknown spectrum (2.a) and reference spectra of PY1 (2.b) and PR3 (2.c). Unknown spectrum (3.a) and reference spectra of PY1 (3.b) and PR4 (3.c). Unknown spectrum (4.a) and reference spectra of PY1 (4.b) and PB60 (4.c)

Theoretical analysis of a multicomponent case

The theoretical performance of the binary mixtures handling-based identification methodology when processing an unknown multicomponent spectrum is presented hereafter. Hence, a reference spectral library was simulated through 10 reference spectra ($P = 10$), Lorentzian-profile-based, in the range of $[200, ..., 1800]cm^{-1}$, which implies a dimension of 1600 variables ($N = 1600$) and therefore a K dimension PCs space of $K = P - 1 = 10 - 1 = 9 << N = 1600$.

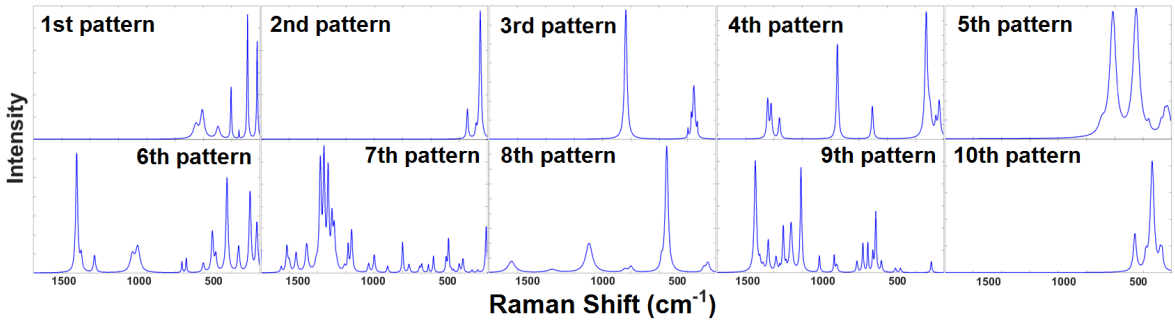


Figure C.2: Simulated reference spectral library

An unknown spectrum was generated by mixing the 2nd, 5th and 7th simulated spectra (see Fig. C.3).

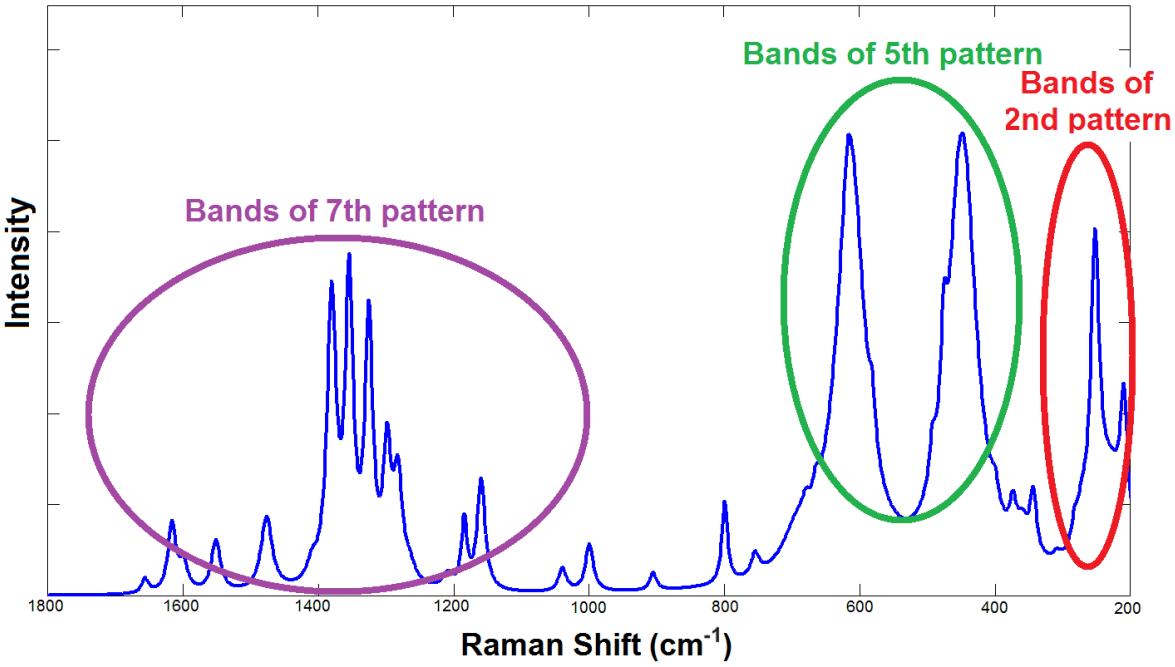


Figure C.3: Unknown spectrum simulating a ternary mixture generated by mixing the 2nd, 5th and 7th simulated spectra

When applying the binary mixtures handling-based identification methodology, two separated candidates were found: 5th and 7th reference spectra with a MF, respectively, of MF(5th pattern)=26.06% and MF(7th pattern)=8.49%. Attending to the mixture-building criteria, the system created the mixture from the reference spectra belonging to the 5th and 7th simulated patterns. Then, the identification criteria were applied over this fictitious mixture obtaining a MF of MF(Mixture of 5th and 7th)=83.12%. Hence, the methodology suggests that the unknown spectrum may correspond to the mixture of patterns 5 and 7. Attending to Fig. C.3 pattern 2's bands can be seen However, this pattern could not be recognised by the binary mixtures handling-based identification algorithm.

Additionally, the following list shows the reference spectra sorted in descending order according to the EDs to the unknown spectrum:

5th, **7th**, 6th, 1st, 10th, 9th, 4t,h **2nd**, 8th, 3rd

As shown, the first two patterns (5th and 7th patterns) correspond to two of the three components of the unknown spectrum, which were successfully recognised by the binary mixtures handling-based identification methodology. Nonetheless, the third component (2nd pattern) appears in the 8th position in the list demonstrating that the mixture-building criteria cannot be extrapolated to allow the identification of mixtures of more than two components. Conclusively, the developed binary mixtures handling-based identification methodology successfully identifies single-component spectra and also binary mixtures, focusing on this kind of mixtures as they may appear with relative frequency in art. However, in some cases, multicomponent spectra of more than two components may appear as well. In these multicomponent cases, the presented methodology may not be able to identify all the components in the mixture, since it is a construction limitation of the binary mixtures handling-based identification algorithm. Therefore, the generalised identification methodology for single- and multicomponent spectra from pigments was developed.

C.1.2 Analysis of generalised identification methodology in simulated environments

With the aim of showing the correct performance of the generalised identification methodology, it was tested in a simulation stage. Thereby, 100 different simulated reference spectral libraries were used. For each of these libraries 1000 unknown spectra were simulated and analysed by applying the presented methodology. On average, the identification of the unknown spectra was successful for 99.63% (with a standard deviation of 0.68%). Fig. C.4 (left) shows the histogram of MFs of the identification results (mostly ranging from 95% to 100%).

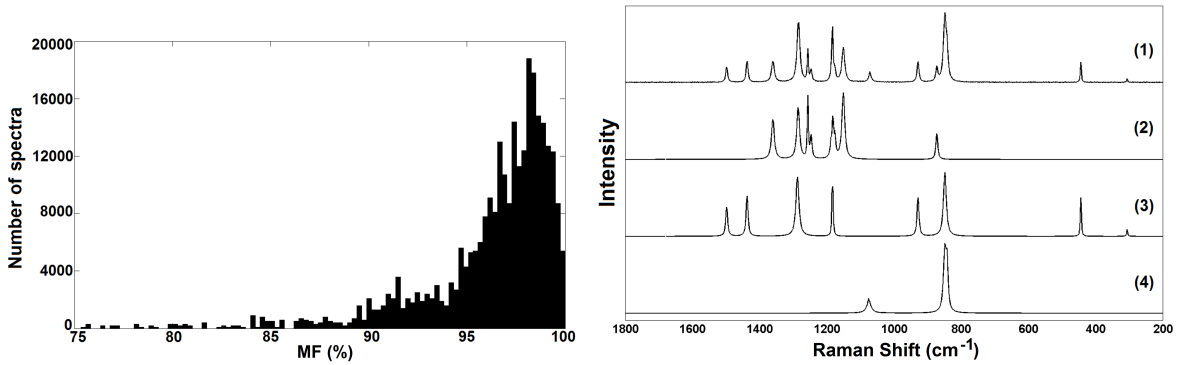


Figure C.4: Left: Histogram of Matching Factors (MFs) of the identification results using simulated Raman spectra. Right: Example of spectral identification from a simulated three-component mixture whose components have overlapping bands: (1) Unknown mixture, (2) and (3) components identified and (4) component not identified

In the remaining cases (0.37%) not all the individual components of the mixtures were identified, even though all components identified were correct. It was checked whether the fundamental band of these unidentified components overlapped with the fundamental band of an identified component. In these cases, the system interprets these overlapping bands as part of one of the individual components only (the most similar one to the unknown spectrum as seen by the IB) so that the others may miss some valuable Raman information when the MSB is applied and therefore may be identified with a lower MF or even may be left unidentified. An example of this issue is presented in Fig. C.4 (right). The system analysed the spectrum of mixture (1) of three components (2, 3 and 4) and identified the components (2) with a MF of 98.7% and (3) with a MF of 84.3%. Component (4) was not identified since its fundamental band overlaps the fundamental band of component (3) and therefore it was missed in the unknown spectrum identifying component (3). This issue may be a drawback of the proposed method, but the study of these instances by traditional methods is equally complex. Still, the casuistry of this situation tends to be relatively low, and yet, the system is able to identify the main components present in the mixture.

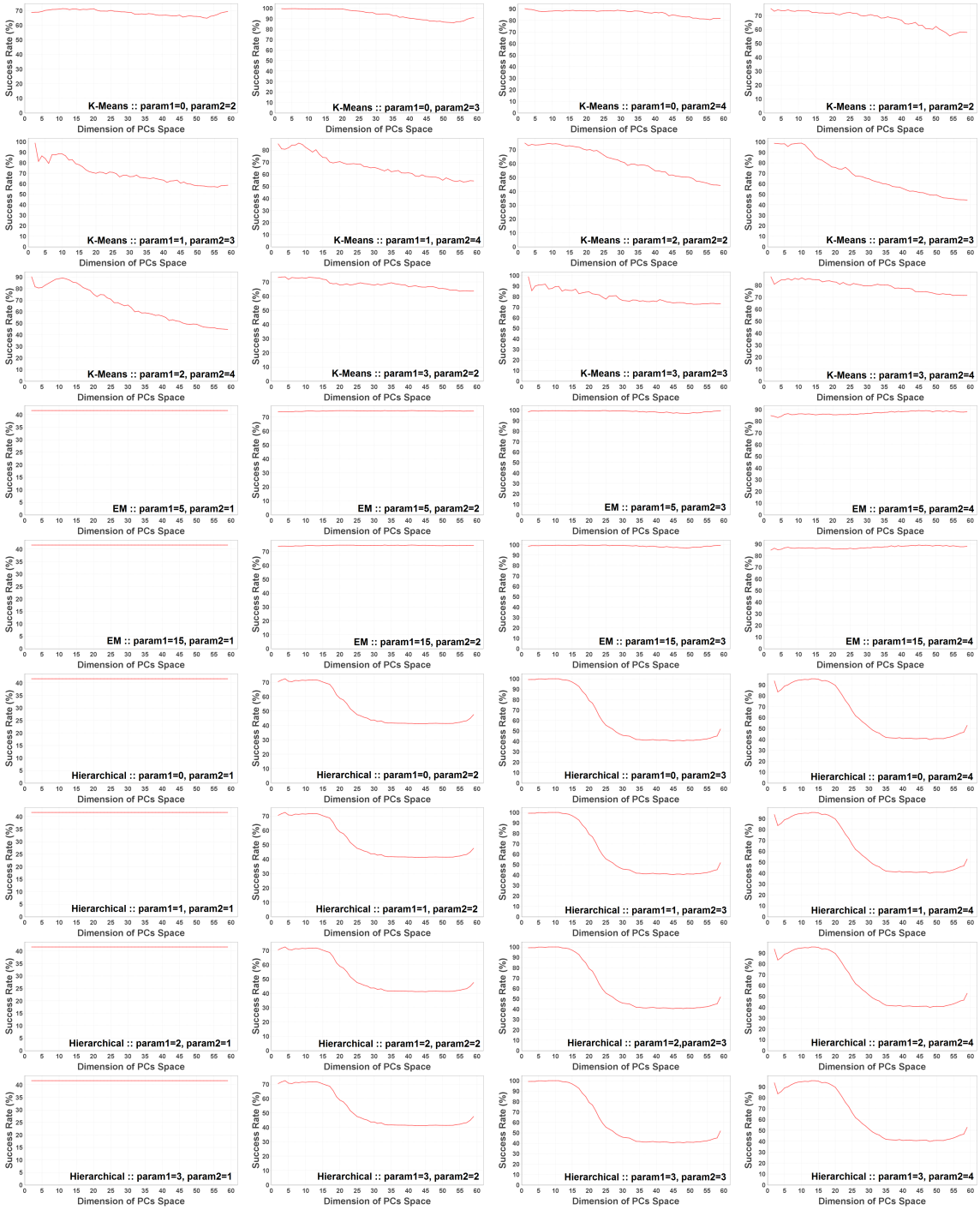
On the other hand, an additional validation of the identification system was carried out in order to study the MSB independently of the IB. To do so, every time the MSB was triggered in the above described simulation the residual between the pattern with the lowest ED and the corresponding component provided by the MSB was evaluated in terms of goodness-of-fit by means of several statistics: the R-square and the RMSE. In this sense, a good fit would have a RMSE closer to 0 and an R-square closer to 1. The following goodness-of-fit statistics (mean and standard deviations) were obtained: R-square $= 0.9959 \pm 0.0119$, and RMSE $= 0.0040 \pm 0.0045$. Taking the mean R-square for instance (99.59%), it indicates a good fit of the components separated by the MSB, quite close to the ideal case.

C.2 Classification of Raman spectra from pigments

C.2.1 Unsupervised classification of Raman spectra

Performance evaluation results using simulated datasets

Complementary results to those presented in Fig. 4.13 regarding the performance evaluation process using the selected clustering algorithms through simulated Raman spectra are shown in Fig. C.5. These figures allowed to determine the best-performing configuration parameters for each clustering technique.



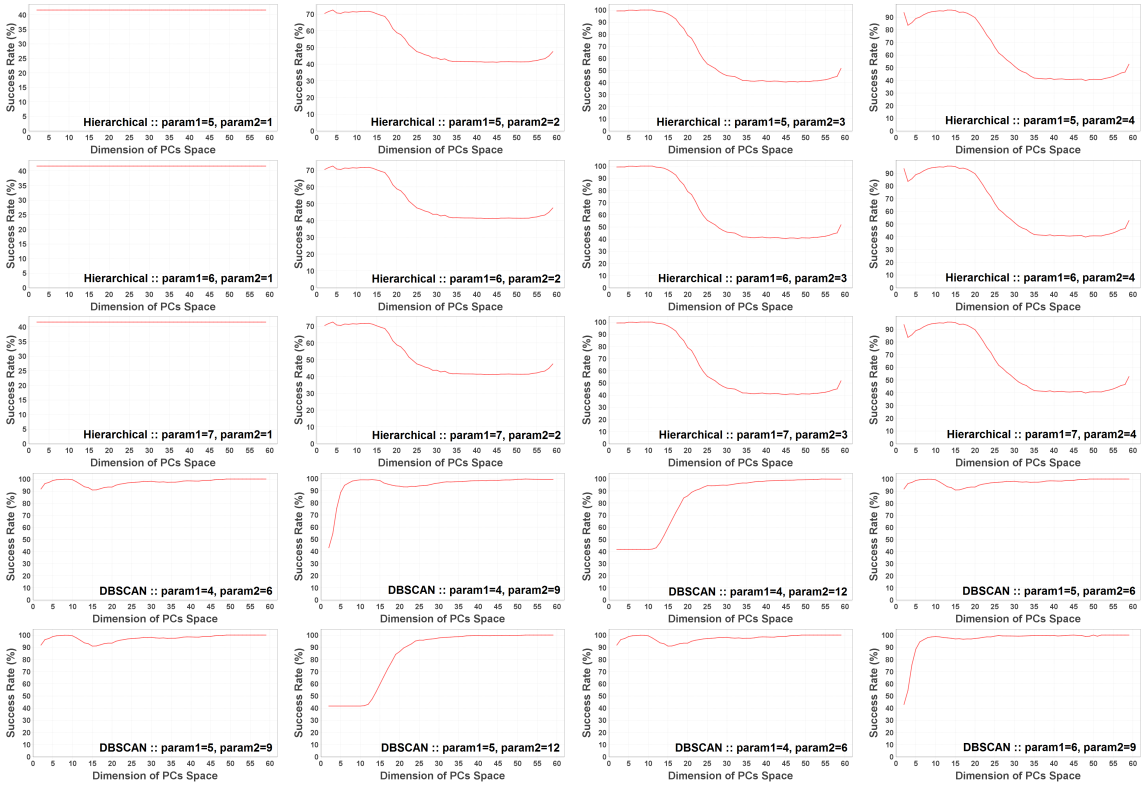


Figure C.5: Success rate as a function of the PCs space dimension of the analysed clustering algorithms

Unsupervised classification: Experimental cases

The unsupervised classification methodology based on PCA and k-means was applied to experimental Raman spectra. Specifically, we distinguish between three different crystalline structures of copper-phthalocyanine blue pigment: α-, β- and ϵ-modifications. In a first case, the Raman spectra were recorded using a single excitation wavelength of 785 nm. In a second case, the Raman spectra were recorded using multiple excitation wavelengths (532nm, 633nm and 785nm).

Single excitation wavelength case The input Raman spectra used as training dataset consisted on 12 Raman spectra (see Fig. C.6). In particular, the α-modification class consisted of 5 spectra, the β-modification class consisted also of 5 spectra, and the ϵ-modification class consisted of 2 spectra. All the spectra were recorded using a 785nm excitation wavelength. The PCA projection is represented in bottom of Fig. C.6 together with the k-means centroids. The methodology was validated through a cross-validation based on using all the spectra in the training set as a test set. This cross-validation provided a 100% of success rate. Additionally, an unknown Raman spectrum expected to be from an α-CuPc pigment was used a test instance and the methodology provided a successful result, i.e. the unknown Raman spectrum was clustered as a Raman spectrum from an α-CuPc pigment (see Fig. C.7).

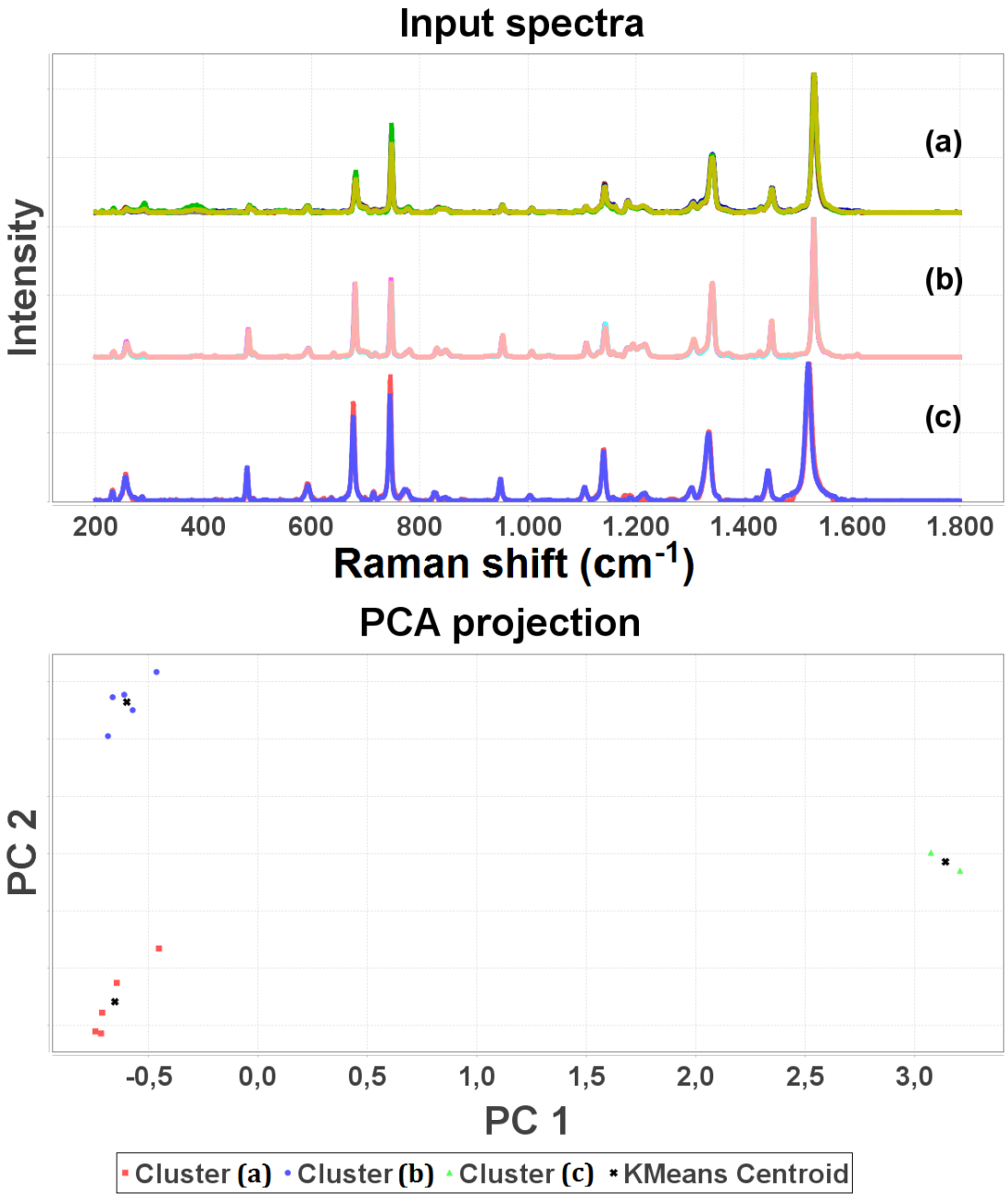


Figure C.6: Input Raman spectra used as training dataset consisting on 12 Raman spectra (a) 5 spectra from the α-modification, b) 5 spectra from the β-modification and c) 2 spectra from the ϵ-modification) measured with a 785nm excitation wavelength (top) together with the corresponding PCA projection (bottom)

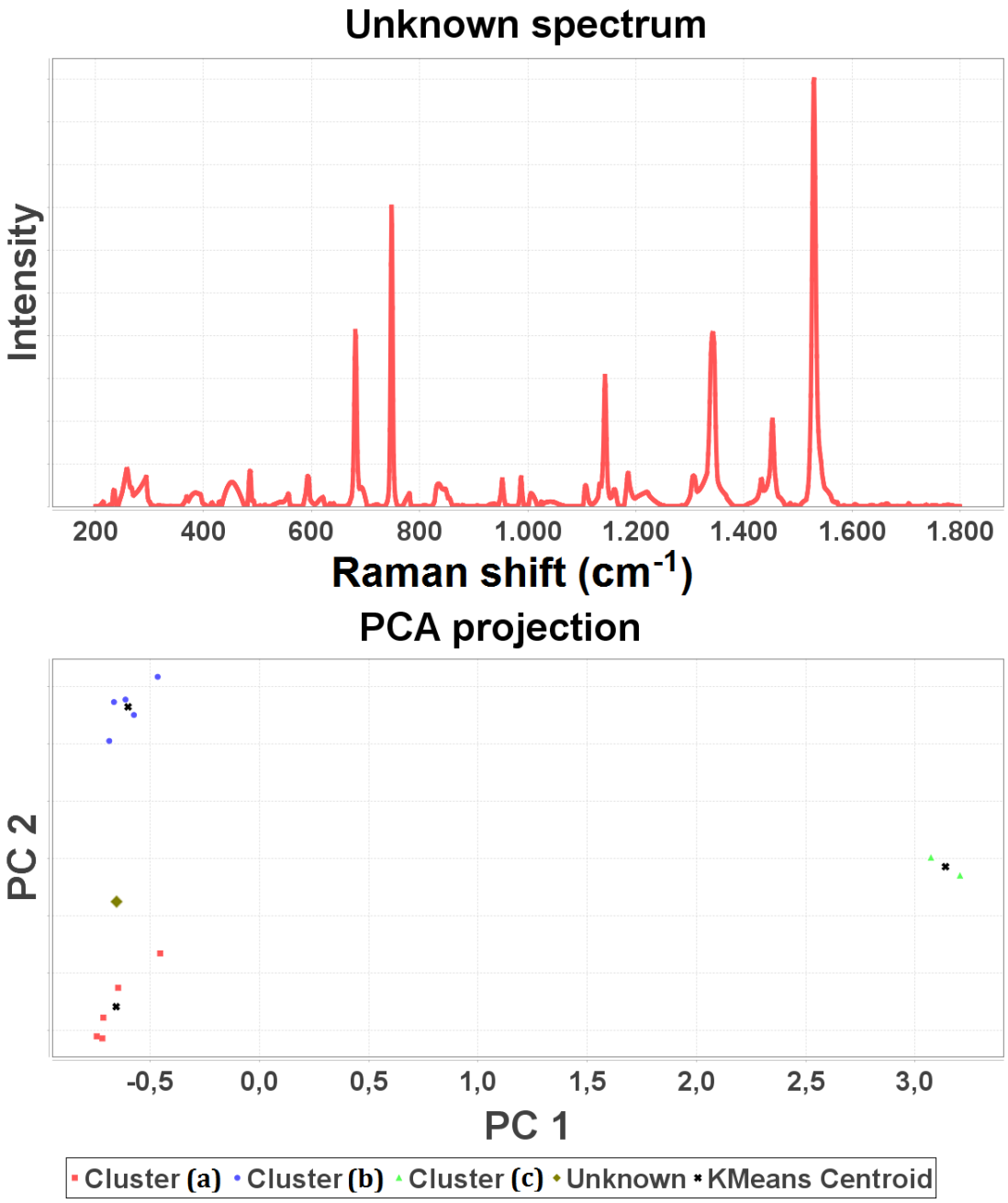


Figure C.7: Unknown Raman spectrum expected to be a Raman spectrum from a α-modification CuPc pigment (top) together with the corresponding PCA projection

Multiple excitation wavelength case In this case, the input Raman spectra used as training dataset consisted on 79 Raman spectra (see top of Fig. C.8). In particular, the α-modification class consisted of 27 spectra: nine spectra recorded using a 532nm excitation wavelength, ten spectra recorded using a 633nm excitation wavelength, and eight spectra recorded using a 785nm excitation wavelength. The β-modification class consisted of 38 spectra: eleven spectra were recorded using a 532 nm excitation wavelength, thirteen spectra recorded using a 633 nm excitation wavelength, and fourteen spectra recorded using a 785nm excitation wavelength. Finally, the ϵ-modification class consisted of 14 spectra: ten spectra were recorded using a 532nm excitation wavelength, one spectrum recorded using a 633nm excitation wavelength, and three spectra recorded using a 785nm excitation wavelength.

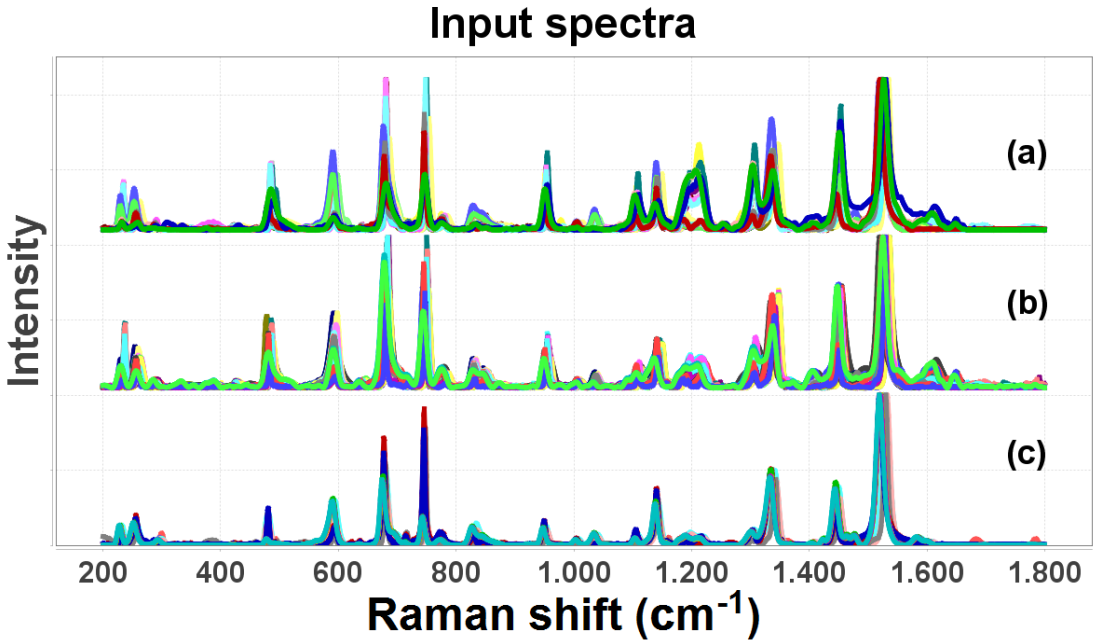


Figure C.8: Input Raman spectra used as training dataset consisting on 79 Raman spectra (a) 27 spectra from the α-modification, b) 38 spectra from the β-modification and c) 14 spectra from the ϵ-modification) measured with multiple excitation wavelengths

The maximum success rate (56.96%) was obtained with a 33 dimension PCs space when applying the cross-validation using all the training set as test set (see Fig. C.9). As seen in Fig. C.10, it is difficult to group the projected Raman spectra into separated clusters for the α-, β- and ϵ-modifications of the CuPc pigment. This non-separated distribution translated into a relatively low success rate. On the other hand, the unknown Raman spectra used as test instance as in the previous case was clustered using the 33 dimension PCs space as a Raman spectrum from a ϵ-CuPc pigment in this case (see Fig. C.10), whilst being expected to be from an α-CuPc pigment.

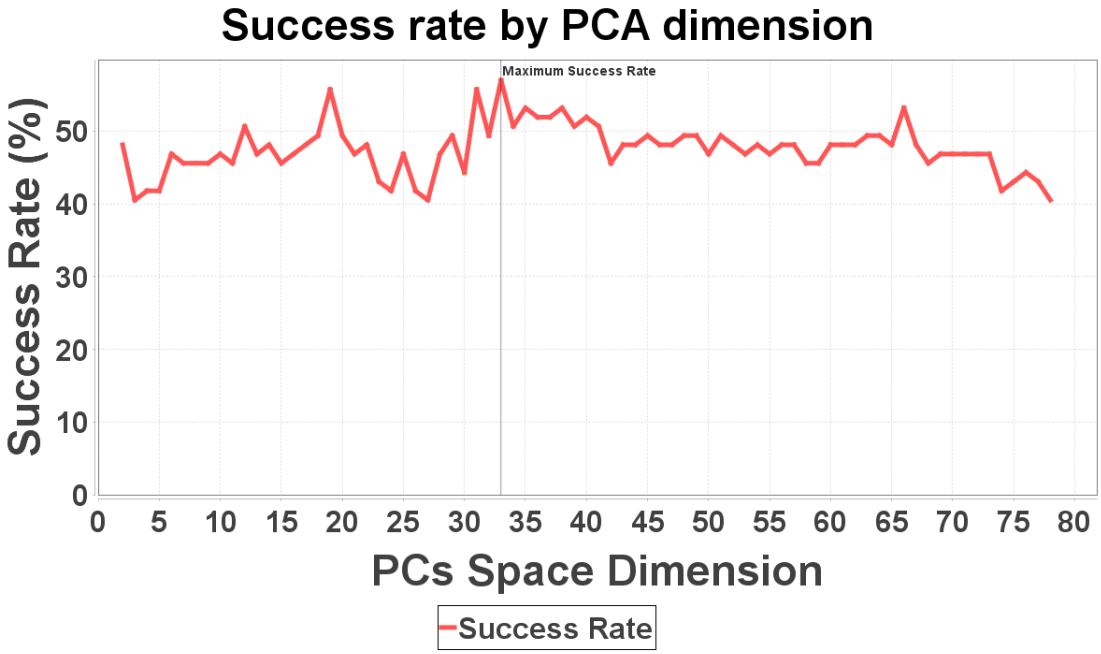


Figure C.9: Success rate as a function of PC dimension. Maximum success rate obtained with a PCA projection of 33 PCs

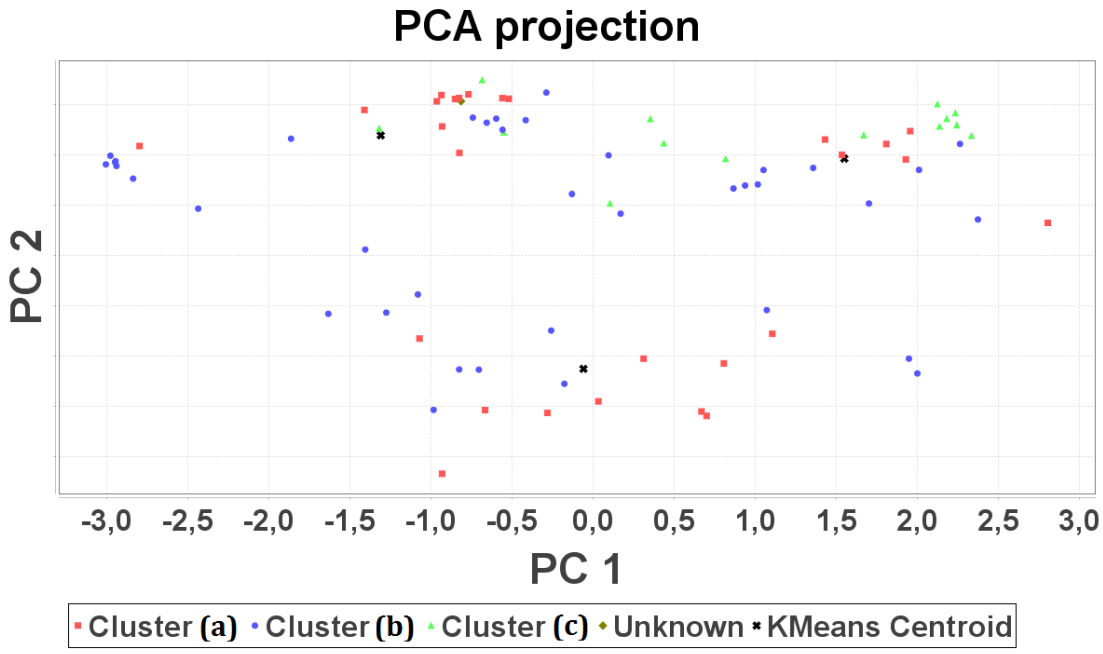


Figure C.10: PCA projection of input reference Raman spectra and projected unknown Raman spectrum together with the k-means centroids

C.2.2 Supervised classification of Raman spectra: Verification and validation

In order to diagnose the performance of the proposed classification methodology, it was analysed in a simulation stage. In particular, three different classes were created. To do so, three different spectra were generated, which simulated spectra measured from three different pigments. The only difference between these spectra was the amplitude of two selected bands between $650cm^{-1}$ and $800cm^{-1}$, as can be seen in Fig. C.11. Then, ten different simulated spectra were generated for each class, simulating different realizations for each of the three pigments. These different realizations were generated through random variations in band locations, amplitudes and bandwidths. Specifically, normal distribution functions were used giving random variations of $\pm 5cm^{-1}$ in band locations, $\pm 0.05a.u.$ in normalized intensities, and $\pm 2cm^{-1}$ in bandwidths (see inset figure in Fig. C.11).

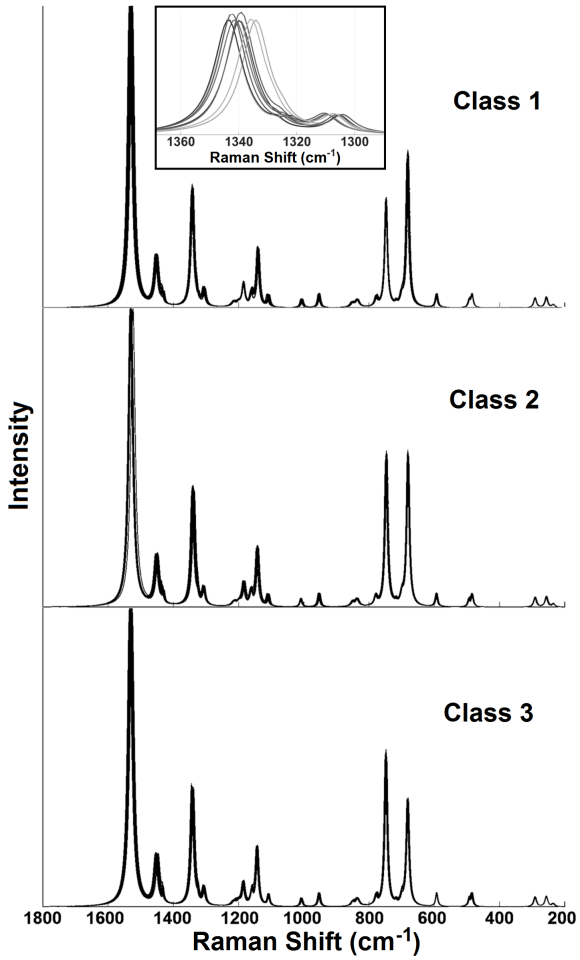


Figure C.11: Example of simulated spectra used for cross-validating the implemented classification methodology

With the simulated classes, we applied the Lachenbruch procedure (also called leave-one-out cross-validation)[180], which is a standard model validation technique for assessing the predictive performance of a methodology. It involves using one spectrum as the test set (for which we certainly know the corresponding class) and the remaining spectra as the training dataset. Specifically, based on this cross-validation procedure, we applied the following five-step sequence:

- Let the i-th spectrum form the test set (test spectrum)
- Get the classification space using the remaining spectra (29 spectra)
- Apply the classification criterion for class assignment on the test spectrum
- Check the classification outcome with respect to the expected class
- Repeat step 1 for i=1,...,n with n being the total number of spectra ($n = 30$)

We obtained a success rate of 100%, which shows a good predictive performance of the presented classification methodology in an under-controlled environment using simulated data.

Appendix D

Software Requirements Specification of PigmentsLab

The requirements listed in this appendix are applicable to the implementation, test and operation of ***PigmentsLab***, software platform designed and developed in this research. This appendix not only lists the several requirements (both scientific and technical) -in agreement to[181]-, but it also includes some system descriptions.

D.0.1 Definitions

The requirements set out in this appendix conform the following labelling scheme:

LABEL-X-SCOPE-xxx

where: *LABEL* is the Software Product Label, *X* is the requirement type, *SCOPE* is a three letter scope specification, and *xxx* is an identification number.

Each requirement is presented with its unique label and a number in the following form:

LABEL-X-SCOPE-xxx	Version	Priority	Status
Description			
Parent			

with:

LABEL-X-SCOPE-xxx
: The unique identifier of the requirement (see above)

Version
: Version number of the requirement

Priority
: Priority of the requirement

Status
: Status identifier

Parent
: Higher level requirement or requirements, comma separated list

D.1 General Description

D.1.1 Product Functions

PigmentsLab is in charge of:

- Storing and browsing existing items in the database, e.g. the reference spectra from pigments
- Adding and updating reference spectra from pigments
- Providing, for each existing item in the database, information regarding:
  - General description: pigment name, colour index, chemical class and usage
  - Spectroscopic information: spectral plot (including the following spectroscopic techniques: Raman, SERS, XRF, IF and LIBS), band positions if known, excitation source [nm], source power [mW], accumulations and acquisition time [s]
- Allowing spectral amplitude adjustment (scaling and factoring, and intensity normalisation)
- Providing band markers, modelling and localisers through different band profiles (Lorentzian, Gaussian or Voigt)
- Providing pre-processing techniques such as spectral enhancement, noise filtering, fluorescence's baseline rejection and shot noise reduction
- Zooming in and out to interesting regions in spectral plots
- Recognising unknown spectra (whether single- or multi-component) through automated matching-based spectral identification methodologies
- Classifying unknown spectra by means of machine learning-based methodologies trained through predefined categories of reference spectra from pigments

The system is expected to evolve over several releases, becoming, eventually, a useful tool to manage and browse a complete spectral library from pigments, to apply sophisticated visualisation and pre-processing tools, and to perform advanced interpretation techniques for identification and classification of spectra from pigments.

D.1.2 User Features

Reference spectral database management: The reference spectral database from artistic pigments shall be a powerful tool. That is, the spectroscopists may obtain information regarding different features for each element in the database. In this sense, the developing of a reference spectral database applied to artistic pigments is an open issue: when new artistic materials appear, new spectra shall be added to the existing reference spectra. Thus, spectroscopists need an robust interface to manage the spectra for whichever analytical purpose.

Visualisation and enhancement of spectra from pigments: The visualisation of spectroscopic measurements from art materials and the elemental data analysis such as noise filtering or bands localisation shall support the spectroscopists in the tasks of analysing spectra in a visual way.

Identification and classification of spectra from pigments: The automated interpretation of spectra from art materials through matching and machine learning-based methodologies for identification or classification of spectra from pigments shall provide an advanced application for making breakthroughs in art works analysis, providing a spectral characterization with no prior knowledge of the composition of the analysed sample.

D.1.3 General Constraints

SDBE-T-CO-001	1.0	MAN	Draft
The system's design, code, and maintenance documentation shall conform to the Development Standards.			
Parent: None			

SDBE-T-CO-002	1.0	MAN	Draft
The system shall use the current corporate standard MySQL database engine.			
Parent: None			

D.1.4 Assumptions and Dependencies

SDBE-T-DE-001	1.0	MAN	Draft
The operation of the system depends on changes being made in the system to browse existing items in the database, to visualise spectral measurements applying pre-processing tools, or to characterise unknown spectra through automated identification or classification methodologies.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

D.2 Specific Requirements

D.2.1 External Interface Requirements

None.

D.2.2 Functional Requirements

SDBE-T-STO-001	1.0	AUT	Issued
SDBE shall store spectra from pigments and their corresponding information in the database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-ADD-001	1.0	AUT	Issued
SDBE shall allow the user to add new spectra from pigments.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-ADD-002	1.0	AUT	Issued
When adding new spectra from pigments SDBE shall ask for chemical class.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*			

SDBE-T-ADD-003	1.0	AUT	Issued
When adding new spectra from pigments and chemical class being supplied SDBE shall store all the information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*, *SDBE-T-ADD-002*, *SDBE-T-STO-001*			

SDBE-T-ADD-005	1.0	AUT	Issued
When adding new spectra from pigments SDBE shall ask for colour index information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*			

SDBE-T-ADD-006	1.0	AUT	Issued
When adding new spectra from pigments and colour index information being supplied SDBE shall store all the information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*, *SDBE-T-ADD-005*, *SDBE-T-STO-001*			

SDBE-T-ADD-007	1.0	AUT	Issued
When adding new spectra from pigments SDBE shall ask for art-historical information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*			

SDBE-T-ADD-008	1.0	AUT	Issued
When adding new spectra from pigments and art-historical information being supplied SDBE shall store all the information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*, *SDBE-T-ADD-007*, *SDBE-T-STO-001*			

SDBE-T-ADD-009	1.0	AUT	Issued
When adding new spectra from pigments SDBE shall ask for the corresponding spectroscopic information (excitation source [nm], source power [mW], accumulations, acquisition time [s], and bands positions if known).			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*			

SDBE-T-ADD-010	1.0	AUT	Issued
When adding new spectra from pigments and the corresponding spectroscopic information (excitation source [nm], source power [mW], accumulations, acquisition time [s], and bands positions if known) being supplied SDBE shall store all the information.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-ADD-001*, *SDBE-T-ADD-007*, *SDBE-T-STO-001*			

SDBE-T-ADD-011	1.0	AUT	Issued
When adding new spectra from pigments, if key fields are left unfilled, the SDBE shall show a warning message and wait for the fields to be filled by the user			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-EXP-001	1.0	AUT	Issued
If no spectra from pigments in the database SDBE shall show a message of empty database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-002	1.0	AUT	Issued
SDBE shall allow the user to explore the existing database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-003	1.0	AUT	Issued
When an item is selected, SDBE shall provide the existing information, mainly general description (pigment name, colour index, chemical class and usage) and spectrosopic description (excitation source [nm], source power [mW], accumulations, acquisition time [s], and bands positions if known) of the corresponding item.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*			

SDBE-T-EXP-004	1.0	AUT	Issued
When an item is selected, SDBE shall provide a plot of the selected spectra from pigments according to the selected spectroscopic technique used for measurement (Raman, SERS, XRF, IF and LIBS).			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-005	1.0	AUT	Issued
When plotting a spectrum from a pigment, the spectral range shall be configurable by the user.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-006	1.0	AUT	Issued
When plotting a spectra from pigments, the samples shall be interpolated (linear interpolation) to assure homogeneity.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-007	1.0	AUT	Issued
When plotting a spectra from pigments, the intensity of the samples shall be normalized (min-max normalisation) to assure homogeneity.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-STO-001*			

SDBE-T-EXP-008	1.0	AUT	Issued
SDBE shall allow the user to update the information whether general or spectroscopic of an existing item in the database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*, *SDBE-T-EXP-001*			

SDBE-T-EXP-009	1.0	AUT	Issued
When updating an item and adding new chemical class SDBE shall store it.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*			

SDBE-T-EXP-010	1.0	AUT	Issued
When updating an item and adding new colour index information SDBE shall store it.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*			

SDBE-T-EXP-011	1.0	AUT	Issued
When updating an item and adding new art-historical information SDBE shall store it.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*			

SDBE-T-EXP-012	1.0	AUT	Issued
When updating an item and adding new spectrosopic information (excitation source [nm], source power [mW], accumulations, acquisition time [s], and bands positions if known) SDBE shall store them.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EXP-001*, *SDBE-T-STO-001*			

SDBE-T-EXP-013	1.0	AUT	Issued
If no spectra from pigments is selected, the action of update a spectra from pigments must be ignored by the SDBE showing a warning message.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-EID-001	1.0	AUT	Issued
SDBE shall allow the user to export the existing database to a backup file. This is mandatory for backing up and restoring. Also useful for data portability			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-EID-002	1.0	AUT	Issued
SDBE shall allow the user to import data to the database from a backup file. This action allows restoring the data. Also useful for portability			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-EID-003	1.0	AUT	Issued
When the exporting/importing actions are requested SDBE must assure losslessness of data.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*, *SDBE-T-EID-001*, *SDBE-T-EID-002*			

SDBE-T-EID-004	1.0	AUT	Issued
If the file to be read when the importing action is required, the SDBE must ignore the action, showing a warning message.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-EID-005	1.0	AUT	Issued
If the file to be exported already exists when the exporting action is required, the SDBE must ignore the action, showing a warning message.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-001	1.0	AUT	Issued
SDBE shall allow the user to delete an existing spectra from pigments.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-002	1.0	AUT	Issued
When deleting an existing spectra from pigments all the corresponding information shall be removed from the database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-003	1.0	AUT	Issued
When deleting an existing spectra from pigments the other spectra from pigments must remain in the database.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-004	1.0	AUT	Issued
When deleting an existing spectra from pigments the other spectra from pigments must update their identification field. The rest of fields must remain untouched.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-005	1.0	AUT	Issued
If no spectra from pigments is selected, the action of delete a spectra from pigments must be ignored by the SDBE showing a warning message.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SDBE-T-DEL-006	1.0	AUT	Issued
After deleting a spectra from pigments, the list of existing spectra from pigments shall be refreshed.			
Parent: *SDBE-T-CO-001*, *SDBE-T-CO-002*			

SV-T-VIEW-001	1.0	AUT	Issued
SpectralViewer shall read spectral files.			
Parent: None			

SV-T-VIEW-002	1.0	AUT	Issued
If the loading of a spectral file is cancelled by the user, the action of loading must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-003	1.0	AUT	Issued
When a spectral file is loaded, SpectralViewer show provide a graphical representation of the loaded spectrum.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-004	1.0	AUT	Issued
SpectralViewer shall allow spectral amplitude adjustment by a given scalar.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-005	1.0	AUT	Issued
SpectralViewer shall allow spectral amplitude adjustment by a given factor.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-006	1.0	AUT	Issued
SpectralViewer shall allow spectral amplitude adjustment by the min-max normalisation.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-007	1.0	AUT	Issued
If no spectra is loaded and spectral amplitude adjustment is selected, the action must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-008	1.0	AUT	Issued
SpectralViewer shall provide tools for band markers, and when a given coordinate in the domain axis is clicked, a band marker should be placed in the plot.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-009	1.0	AUT	Issued
If no spectra is loaded and band markers option is selected, the action must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-010	1.0	AUT	Issued
SpectralViewer shall provide tools for band modelling and localisation, automatically recognising the bands present in the loaded spectrum through different profiles to be chosen by the user: Lorentzian, Gaussian or Voigt.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-011	1.0	AUT	Issued
If no spectra is loaded and band modelling or localisation is selected, the action must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-012	1.0	AUT	Issued
SpectralViewer shall provide tools for pre-processing techniques such as spectral enhancement, noise filtering, fluorescence's baseline rejection and shot noise reduction.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-013	1.0	AUT	Issued
If no spectra is loaded and any pre-processing techniques option is selected, the action must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-014	1.0	AUT	Issued
SpectralViewer shall provide tools for zooming in and out to interesting regions in spectral plots.			
Parent: *SV-T-VIEW-001*			

SV-T-VIEW-015	1.0	AUT	Issued
If no spectra is any loaded and zooming in or out option is selected, the action must be ignored by the SpectralViewer showing a warning message.			
Parent: *SV-T-VIEW-001*			

VS-T-INT-001	1.0	AUT	Issued
VirtualSpectroscopist shall provide tools for recognising unknown spectra (whether single- or multi-component) through automated matching-based spectral identification methodologies.			
Parent: None			

VS-T-INT-002	1.0	AUT	Issued
VirtualSpectroscopist shall provide tools for classifying unknown spectra by means of machine learning-based methodologies trained through predefined categories of reference spectra from pigments.			
Parent: None			

VS-T-FW-001	1.0	AUT	Issued
VirtualSpectroscopist shall read spectral files from unknown spectra.			
Parent: None			

VS-T-FW-002	1.0	AUT	Issued
If the loading of a spectral file is cancelled by the user, the action of loading must be ignored by the VirtualSpectroscopist showing a warning message.			
Parent: *VS-T-FW-001*			

VS-T-FW-003	1.0	AUT	Issued
VirtualSpectroscopist shall load image files from pictures of the art work being analysed.			
Parent: None			

VS-T-FW-004	1.0	AUT	Issued
If the loading of a image file is cancelled by the user, the action of loading must be ignored by the VirtualSpectroscopist showing a warning message.			
Parent: *VS-T-FW-003*			

VS-T-FW-005	1.0	AUT	Issued
VirtualSpectroscopist shall provide tools for adding markers representing spectral measurements			
Parent: *VS-T-FW-003*			

VS-T-FW-006	1.0	AUT	Issued
VirtualSpectroscopist shall provide tools for generating a report with the outcome of the interpretation of spectra from pigments			
Parent: *VS-T-INT-001*, *VS-T-INT-002*			

D.2.3 Non-Functional Requirements

SDBE-T-NFUN-001	1.0	MAN	Issued
PigmentsLab shall be completely developed in Java language. This excludes scripts or utilities for launching or monitoring the platform, which may be done in other languages (bash or python preferably).			
Parent: None			

Publications

Peer-reviewed articles published in international journals

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *Automatic morphology-based cubic p-spline fitting methodology for smoothing and baseline-removal of Raman spectra*, 2017, Journal of Raman Spectroscopy, DOI: 10.1002/jrs.5130

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *Automatic classification system of Raman spectra applied to pigments analysis*, 2016, Journal of Raman Spectroscopy, 47(12), 1408

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Independent component analysis-based algorithm for automatic identification of Raman spectra applied to artistic pigments and pigment mixtures*, 2015, Applied Spectroscopy, 69(3), 314

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Automatic identification system of Raman spectra in binary mixtures of pigments*, 2012, Journal of Raman Spectroscopy, 43(11), 1707

Contributions to international conferences

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *Raman characterisation of copper phthalocyanine blue under solvents and cleaning agents*, International Congress on the Application of Raman Spectroscopy in Art and Archaeology, 2017, *submitted*

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *A software platform proposal for the automated interpretation of spectra in artworks analysis*, International Conference on Innotavion in Art Research and Technology, 2016, Book of Abstracts, p.78, ISBN: 978-94-6197-367-2

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, *Proposal of a classification system of Raman spectra applied to pigments analysis*, International Congress on the Application of Raman Spectroscopy in Art and Archaeology, 2015, Book of Abstracts, p.78, ISBN: 978-83-60043-27-1

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Fully automated fluorescence background removal and shot noise filtering in Raman spectroscopy applied to pigments analysis*, International Conference on Raman Spectroscopy Applied to Earth Sciences, 2014, Book of Abstracts, p.69

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Shot noise reduction through principal components analysis*, International Congress on the Application of Raman Spectroscopy in Art and Archaeology, 2013, Book of Abstracts, p.134, ISBN: 978-961-6902-38-0

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *ICA-based algorithm for pigment mixtures identification*, International Conference on Raman Spectroscopy Applied to Earth Sciences, 2012, Book of Abstracts, p.145

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Pattern recognition based on principal component analysis in Raman spectroscopy applied to pigments analysis*, International Congress on the use of Multivariate Analysis and Chemometrics in Cultural Heritage and Environment, 2012, Book of Abstracts, p.54, ISBN: 978-887-5473-32-7

J. J. González-Vidal, R. Pérez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Automatic identification system of Raman spectra of pigment mixtures*, International Congress on the Application of Raman Spectroscopy in Art and Archaeology, 2011, Book of Abstracts, p.37, ISBN: 978-88-97162-20-9

Bibliography

1. Edwards, H. G. M., Vandenabeele, P., Jehlicka, J. & Benoy, T. J. *Spectrochim. Acta and Part A.* **598,** 118 (2014).
2. Tournie, A., Prinsloo, L. C., Paris, C., Colomban, P. & Smith, B. *J. Raman Spectrosc.* **42,** 399 (2011).
3. Casadio, F., A. Bezúr, I. F., Muir, K., Trad, T. & Maccagnola, S. *J. Raman Spectrosc.* **43,** 1761 (2012).
4. Barnett, J. R., Miller, S. & Pearce, E. *Optics & Laser Technology* **38,** 445 (2006).
5. Mayer, R. *The artist's handbook of materials and techniques* (Tursen Hermann Blume, Oxford, 1999).
6. Edwards., H. G. *Historical Pigments: A Survey of Analytical Chemical Archaeometric Usage and Terminology for Forensic Art Analysis* (Encyclopedia of Analytical Chemistry, 2015).
7. Miguel, C., Claro, A., Gonalves, A. P., Muralha, V. S. F. & Melo, M. J. *J. Raman Spectrosc.* **40,** 1966 (2009).
8. Irazola, M. *et al. J. Raman Spectrosc.* **43,** 1676 (2012).
9. Domínguez-Vidal, A. & de la Torre-López, M. J. *J. Raman Spectrosc.* **45,** 1006 (2014).
10. Guimet, J. B. *Ann. Chim.* **46,** 431 (1831).
11. Duoma, M. *Pigments through the Ages* `http://www.webexhibits.org/pigments`. Accessed: 2017-04-13.
12. Cariati, F. *Raman Spectroscopy. Modern Analytical Methods in Art and Archaeology* (John Wiley, Oxford, 2000).
13. Domènech-Carbó, M. T. *Analytica Chimica Acta* **621,** 109 (2008).
14. Winckelmann, J. *History of the Art of Antiquity* (Getty Publications, Michigan, 2006).
15. Schoenbein, C. F. *Proceedings of the Natural Science Research Society* **1,** 26 (1846).

16. Whitmore, P. *Conservation Science Research: Activities and Needs and and Funding Opportunities* (National Science Foundation, Pittsburgh, 2005).

17. Cavaleri, T., Giovagnoli, A. & Nervo, M. *Procedia Chem.* **8,** 45 (2013).

18. Silva, C. E., Silva, L. P. & Edwards, H. G. M. *Anal. and Bioanal. Chem.* **386,** 2183 (2007).

19. Madariaga, J. M. *J. Raman Spectrosc.* **41,** 1389 (2010).

20. Bersani, D. *Anal. Methods* **8,** 8395 (2016).

21. Favero, P. A., Mass, J. & Delaney, J. K. *Heritage Science* **5,** 13 (2017).

22. Hutanu, D., Woods, A. G. & Darie, C. *Modern Chem. App.* **1,** 3 (2013).

23. Sutherland, K. *Stud. Conserv.* **45,** 4 (2000).

24. White, R. *Stud. Conserv.* **23,** 57 (1975).

25. Mills, J. S. & White, R. *The Organic Chemistry of the Museum* (Butterworths, London, 1987).

26. Tove, P. A., Sigurd, D. & Petersson, S. *Nucl. Instrum. Methods* **1,** 441 (1980).

27. Ropret, P., Miliani, C., Centeno, S. A., Tavzes, C. & Rosi, F. *J. Raman Spectrosc.* **41,** 1462 (2010).

28. Saverwyns, S. *J. Raman Spectrosc.* **41,** 152 (2010).

29. Edwards, H. G. M. *Spectrochim. Acta and Part A.* **80,** 14 (2011).

30. Kirmizi, B., Colomban, P. & Blanc, M. *J. Raman Spectrosc.* **41,** 1240 (2010).

31. Pelletier, M. J. *Analytical applications of Raman Spectroscopy* (Blackwell Science, Oxford, 1999).

32. Scherrer, N. C. & et al. *Spectrochim. Acta and Part A.* **73,** 505 (2009).

33. Long, D. A. *The Raman Effect. A Unified Treatment of the Theory of Raman Scattering by Molecules* (John Wiley, Oxford, 2002).

34. Turrell, G. & Corset, J. *Raman Microscopy Developments and Applications* (Academic Press and Harcourt Brace Company, Oxford, 1996).

35. McCreery, R. L. *Raman Spectroscopy for Chemical Analysis* (Wiley-interscience publication, Oxford, 2005).

36. Eastaugh, N., Walsh, V. & Chaplin, T. *Pigment Compendium: A Dictionary of Historical Pigments* (Elsevier Butterworth-Heinemann, Oxford, 2004).

37. Gettens, R. J. *Artist's Pigments: A Handbook of ther History and Characteristics* (A. Roy (ed.) and National Gallery of Art, Washington, 1997).

38. Smith, L. I. *A tutorial on Principal Components Analysis* (John Wiley, Washington, 2002).

39. Beebe, K. R., Pell, R. J. & Seasholtz, M. B. *Chemometrics: a practical guide* (John Wiley, New York, 1998).

40. Brereton, R. G. *Applied chemometrics for scientists* (John Wiley, New York, 2007).

41. Van der Heijden, F., Duin, R. P. W., de Ridder, D. & Tax, D. M. J. *Classification and Parameter Estimation and State Estimation: An Engineering Approach using Matlab* (John Wiley, New York, 2004).

42. Huber, J. *Annals of Statistics* **13(2),** 1985 (1985).

43. Centeno, S. A. *J. Raman Spectrosc.* **47,** 9 (2015).

44. Smith, G. D. & Clark, R. J. H. *J. Cult. Herit.* **3,** 101 (2002).

45. Gunn, M., Chottard, G., Rivière, E., Girerd, J. J. & Chottard, J. C. *Stud. Conservat.* **47,** 12 (2002).

46. Hernanz, A. *J. Raman Spectrosc.* **47,** 571 (2015).

47. Kandjani, A. & Griffin, M. J. *J. Raman Spectrosc.* **44,** 608 (2013).

48. Soneira, M. J., Pérez-Pueyo, R. & Ruiz-Moreno, S. *J. Raman Spectrosc.* **138,** 599 (2002).

49. Rowlands, C. J. & Elliott, S. R. *J. Raman Spectrosc.* **42,** 370 (2011).

50. Lieber, C. A. & A.Mahadevan-Jansen. *App. Spectrosc.* **57,** 1363 (2003).

51. Pérez-Pueyo, R., Soneira, M. J. & Ruiz-Moreno, S. *App. Spectrosc.* **64,** 595 (2010).

52. Beier, B. & Berger, A. *Analyst* **134,** 1198 (2009).

53. Bocklitz, T. *et al. J. Raman Spectrosc.* **40,** 1759 (2009).

54. Maerz, A. & Bocklitz, T. *Anal. Chem.* **83,** 8337 (2011).

55. Bocklitz, T., Walter, A., Hartmann, K. & Roesch, P. *Anal. Chem.* **47,** 704 (2011).

56. Friedman, J. H. *Data Min. Knowl. Discovery* **1,** 55 (1997).

57. Schumacher, W., Stoeckel, S., Roesch, P. & Popp, J. *J. Raman Spectrosc.* **45,** 930 (2014).

58. Castanys, M., Pérez-Pueyo, R., Soneira, M. J., Golobardes, E. & Fornells, A. *J. Raman Spectrosc.* **42,** 1553 (2011).

59. Schumacher *et al. J. Raman Spectrosc.* **42,** 383 (2011).

60. Pallipurath *et al. J. Raman Spectrosc.* **12,** 4291 (2013).

61. Ramos, P. M., Ferré, J., Andrikopoulos, K. S. & Ruisánchez, I. *App. Spectrosc.* **58,** 848 (2004).

62. Pérez-Pueyo, R., Soneira, M. J., Castanys, M. & Ruiz-Moreno, S. *Applied Spectrosc.* **63,** 947 (2009).

63. Vandenabeele, P. *Spectrochim. Acta and Part A* **80,** 27 (2011).

64. Ramos, P. M., Ruisánchez, I. & Andrikopoulos, K. S. *Talanta* **75,** 926 (2008).

65. Piantanida, G., Menart, E., Bicchieri, M. & Strlic, M. *J. Raman Spectrosc.* **44,** 1299 (2013).

66. Omar, J., Sarmiento, A., Olivares, M., Alonso, I. & Etxebarria, N. *J. Raman Spectrosc.* **43,** 1151 (2012).

67. Lunsford, R. *et al. J. Raman Spectrosc.* **43,** 1472 (2012).

68. Gobinet, C. *et al. IEEE EMBS* **1,** 6207 (2007).

69. Hyvarinen, A. & Oja, E. *Neural Networks* **13,** 411 (2000).

70. Vrabie, V. *et al. Biomed. Signal Process. Control* **2,** 40 (2007).

71. Wang, W. & Adah, T. *IEEE APSI* **1,** 109 (2005).

72. Wang, W., Adali, T., Li, H. & Emge, D. *IEEE MLSP* **1,** 259 (2005).

73. Bell, A. J. & Sejnowski, T. J. *Neural Comput.* **7,** 1129 (1995).

74. *GRAMS/AI Spectroscopy Software Suite and Adept Scientific.* `http://www.adeptscience.co.uk/products/lab/gramsai`. Accessed: 2017-04-13.

75. *KnowItAll Informatics System and Bio-Rad Laboratories* `http://www.bio-rad.com/en-es/product/raman-software`. Accessed: 2017-04-13.

76. Muro, C. K. & Lednev, I. K. *Anal. Bioanal. Chem.* **409,** 287 (2017).

77. Zhao, J., Frano, K. & Zhou, J. *Applied Spectrosc.* `https://doi.org/10.1177/0003702817694381` (2017).

78. Carey, C., Boucher, T., Mahadevan, S., Bartholomew, P. & Dyar, M. D. *J. Raman Spectrosc.* **46,** 894 (2015).

79. Lowry, S. R. *Automated Spectral Searching in Infrared and Raman and Near-Infrared Spectroscopy* (John Wiley, Oxford, 2006).

80. Mosier-Boss, P. A., Lieberman, S. H. & Newbery, R. *App. Spectrosc.* **49,** 630 (1995).

81. Xie, C. & Li, Y. *J. App. Physics* **93,** 2982 (2003).

82. Rowlands, C. J. & Elliott, S. R. *J. Raman Spectrosc.* **42,** 1761 (2011).

83. Schulze, H. G. & et al. *App. Spectrosc.* **66,** 757 (2012).

84. Li, Z. *et al. Analyst* **138,** 4483 (2013).

85. Osticioli, I., Zoppi, A. & Castellucci, W. *App. Spectrosc.* **61,** 839 (2007).

86. Matousek, P., Towrie, M. & Parker, A. W. *J. Raman Spectrosc.* **33,** 238 (2002).

87. Zhao, J., Carrabba, M. M. & Allen, F. S. *App. Spectrosc.* **53,** 834 (2002).

88. Rosi, F. *J. Raman Spectrosc.* **41,** 452 (2010).

89. Bertinetto, C. G. & Vuorinen, T. *App. Spectrosc.* **68,** 155 (2014).

90. Liland, K. H., Rukke, E., Fjaervol, E. & Isaksson, T. *Chemom. Intell. Lab. Syst.* **109,** 51 (2011).

91. Zhao, J. *App. Spectrosc.* **61,** 1225 (2007).

92. Schulze, H. G. *App. Spectrosc.* **65,** 75 (2011).

93. Carvajal, R. C. *App. Spectrosc.* **70,** 604 (2016).

94. Urbas, A. A. & Choquette, S. J. *App. Spectrosc.* **65,** 665 (2016).

95. Cao, A. *J. Raman Spectrosc.* **38,** 1199 (2007).

96. Galloway, C. M., Ru, E. C. L. & Etchegoin, P. G. *App. Spectrosc.* **63,** 1370 (2009).

97. Wand, M. P. *Comput. Statist.* **15,** 443 (2000).

98. Hudson, D. J. *J. Amer. Statist.* **61,** 1097 (1966).

99. Fuller, W. A. *Austr. J. Agric. Econ.* **13,** 35 (1969).

100. Studden, W. J. & Arman, D. J. V. *Ann. Math. Statist.* **40,** 1557 (1969).

101. Gallant, A. R. & Fuller, W. A. *J. Amer. Statist.* **68,** 144 (1973).

102. Wold, S. *Chemica Scripta* **1,** 97 (1971).

103. Smith, P. L. *J. Amer. Statist.* **33,** 57 (1979).

104. Boor, C. D. *A Practical Guide to Splines* (Springer Verlag, 2001).

105. Eilers, P. H. C. *Anal. Chem.* **75,** 3631 (2003).

106. De Rooi, J. J. & Eilers, P. H. C. *Chemom. Intell. Lab. Syst.* **117,** 56 (2012).

107. Yang, L. & Hong, Y. *Comput. Statist.* **108,** 70 (2016).

108. Eilers, P. H. C. & Marx, B. D. *Statistical Science* **11,** 89 (1996).

109. Wood, S. N. *Stat. Comput.* **1,** 1 (2016).

110. Knott, G. D. *Interpolating cubic splines* (Springer Science and Business Media, 2000).

111. Kooperberg, C. & Stone, C. J. *J. Comput. And Graph. Statist.* **4,** 301 (1992).

112. Friedman, J. H. & Silverman, B. W. *Technometrics* **1,** 3 (1989).

113. Matheron, G. *Random Sets and Integral Geometry* (Wiley, 1975).

114. Serra, J. *Image Analysis and Mathematical Morphology. Theoretical Advances* (Academic Press, New York, 1988).

115. Zhang, E., Wang, F., Li, Y. & Bai, X. *Bio-Med. Mater. Eng.* **24,** 50 (2014).

116. González-Castro, V., Debayle, J. & Pinoli, J. *Pattern Recognition Letters* **47,** 50 (2014).

117. Cardeira, A. M. *Appl. Spectrosc.* **67,** 1376 (2013).

118. Cesaratto, A. *Appl. Spectrosc.* **67,** 1234 (2013).

119. Maia, L. F. *J. Raman Spectrosc.* **44,** 560 (2013).

120. Gautier, G. *Appl. Spectrosc.* **63,** 597 (2009).

121. Ropret, P., Centeno, S. A., & Bukovec, P. *Spectrochim. Acta and Part A* **69,** 486 (2007).

122. Thomas, I. L., Ching, N. P., Benning, V. M. & D'Aguanno, J. A. *Int. J. Rem. Sens.* **8,** 331 (1987).

123. Thacker, N. A., Aherne, F. J. & Rockett, P. I. *Kybernetika* **34,** 363 (1997).

124. Fremout, W. & Saverwyns, S. *J. Raman Spectrosc.* **43,** 1536 (2012).

125. Van der Maaten, L. J. P., Postma, E. O. & van den Herik, H. J. *J. Mach. Learn. Res.* **10,** 66 (2009).

126. Pearson, K. *Phylos. Mag.* **2,** 559 (1901).

127. Fisher, R. A. *Ann. Eugenics* **7,** 179 (1936).

128. Boulesteix, A. L. *Stat. Appl. Genet. Mol. Biol.* **3,** 33 (2004).

129. Navas, N., Romero-Pastor, J., Manzano, E. & Cardell, C. *J. Raman Spectrosc.* **41,** 1486 (2010).

130. Nevin, A. *et al. Anal. Chem.* **79,** 6143 (2007).

131. Breitman, M., Ruiz-Moreno, S. & Gil, A. L. *Spectrochim. Acta and Part A.*

132. Bacci, M. *et al. Vib. Spectrosc.* **49,** 80 (2009).

133. Ali, E. M. A. & Edwards, H. G. M. *Spectrochim. Acta and Part A* **121,** 415 (2014).

134. Torres, A. R. D. *et al. J. Raman Spectrosc.* **45,** 1279 (2014).

135. Svarcova, S., Cermakova, Z., Hradilova, J., Bezdicka, P. & Hradil, D. *Spectrochim. Acta and Part A* **132,** 514 (2014).

136. Beaulieu-Houle, G., Gilson, D. F. R. & Butler, I. S. *Spectrochim. Acta and Part A* **117,** 61 (2014).

137. Lomax, S. Q., Lomax, J. F. & Luca-Westrate, A. D. *J. Raman Spectrosc.* **45,** 448 (2014).

138. Brostoff, L. B., Centeno, S. A., Ropret, P., Bythrow, P. & Pottier, F. *Anal. Chem.* **81,** 6096 (2009).

139. Defeyt, C. *et al. J. Raman Spectrosc.* **43,** 1772 (2012).

140. Hoehse, M., Paul, A., Gornushkin, I. & Panne, U. *Anal. Bioanal. Chem.* **402,** 1443 (2012).

141. Almeida, M. R. D., Correa, D. N., Rocha, W. F. C., Scafi, F. J. O. & Poppi, R. J. *Microchem. J.* **109,** 107 (2013).

142. Barone, G. *et al. J. Raman Spectrosc.* **49,** 898 (2015).

143. Defeyt, C., Pevenage, J. V., Moens, L., Strivay, D. & Vandenabeele, P. *Spectrochim. Acta and Part A* **115,** 636 (2013).

144. De Souza, F., Borba, L., Honorato, R. S. & de Juan, A. *Forensic Sci. Int.* **249,** 73 (2015).

145. MacQueen, J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1,** 281 (1967).

146. Dempster, A., Laird, N. & Rubin, D. *J. of the Royal Stat. Soc.* **39,** 1 (1977).

147. Rokach, L. & Oded, M. *Clustering methods. Data mining and knowledge discovery handbook* (Springer, Oxford, 2005).

148. Ester, M., Kriegel, H., Sander, J. & Xu, X. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* **1,** 226 (1996).

149. Duda, R. O. & Shortliffe, E. H. *Science* **220,** 4594 (1983).

150. Anghelone, M., Jembrih-Simuergera, D. & Schreiner, M. *Spectrochim. Acta and Part A* **149,** 419 (2015).

151. Beckhoff, B., Kanngieer, B., Langhoff, N., Wedell, R. & Wolff, H. *Handbook of Practical X-Ray Fluorescence Analysis* (Springer, 2006).

152. Rampazzi, L., Cariati, F., Tanda, G. & Colombini, M. P. *J. Cult. Herit.*

153. Miziolek, A. W., Palleschi, V. & Schechter, I. *Laser Induced Breakdown Spectroscopy* (2006).

154. Derrick, M. R., Stulik, D. & Landry, J. M. *Infrared Spectroscopy in Conservation Science and Scientific Tools for Conservation* (Getty Publications, 2000).

155. Xu, X. *et al. Advanced Functional Materials* **23,** 4332 (2013).

156. Halac, E. B., Reinoso, M., Luda, M. & Marte, F. *J. of Cultural Heritage* **13,** 469 (2012).

157. Zeng, G., Li, K., Yang, H. & Zhang, Y. *Vib. Spectrosc.* **68,** 38+ (2013).

158. Benington, H. C. *IEEE Annals of the History of Computing* **5,** 350 (1983).

159. Pressman, R. *Software Engineering: A Practitioner's Approach* (McGraw Hill, Boston, 2010).

160. Bass, L., Clements, P. & Kazman, R. *Software Architecture in Practice* (Addison-Wesley Longman Publishing, Boston, 2003).

161. Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P. & Stal, M. *Pattern-oriented Software Architecture and a System of Patterns* (John Wiley, Oxford, 1996).

162. Goldberg, A. *Smalltalk-80: The Interactive Programming Environment* (Addison-Wesley, 1983).

163. Booch, G. *Object Oriented Design: With Applications* (Benjamin Cummings, 1991).

164. Directory, F. S. *Subversion* `http://directory.fsf.org/wiki/Subversion`. Accessed: 2017-04-28. 2013.

165. S. Loughranh, E. H. *Ant in Action* (Manning Publications Company, 2007).

166. Sonatype. *Nexus IQ Server Release Notes* `http://books.sonatype.com/sonatype-clm-book/html/release-notes/index.html`. Accessed: 2017-04-28.

167. Foundation, A. S. *Apache Ivy Release Note and Roadmap* `http://grails.org/Roadmap`. Accessed: 2017-04-28. 2013.

168. Jenkins. *Jenkins Documentation* `https://jenkins.io/doc/`. Accessed: 2017-04-28.

169. Pratt, W. K. *IEEE Trans. on Comput.* **21,** 636 (1972).

170. Xu, Y., Weaver, J. B., Healy, D. M. & Lu, J. *IEEE Trans. on Image Proc.* **3,** 747 (1994).

171. *Colour Index Dyes and Pigments* (Society of Dyers et al., Bradforf and Yorkshire, 1971).

172. Bernstein, J. *Polymorphism in molecular crystals* (University press, Oxford, 2002).

173. Moser, F. H. & Arthur, L. T. *Phthalocyanine compounds* (Reinhold, New York, 1963).

174. Defeyt, C. & Strivay, D. *e-Preservation Science* **11,** 6 (2014).

175. Herbst, W. & Hunger, K. *Industrial Organic Pigments: Production and Properties and Applications* (John Wiley, 2004).

176. Smith, F. M. & Easton, J. D. *J. of the Oil and Colour Chemists' Assoc.* **49,** 614 (1966).

177. Shaibat, M. A., Casabianca, L. B., Siberio-Pérez, D., Matzger, A. J. & Ishii, Y. *J. Physical Chemistry B* **114,** 4400 (2010).

178. Basova, T. V. *J. Structural Chemistry* **41,** 770 (2000).

179. Scherrer, N. C. *Book of Abstracts: RAA2011 and ISBN: 978-88-97162-20-9 and Timeo Editore,* 203 (2011).

180. Zhang, Y. & Yang, Y. *J. Econometrics* **187,** 95 (2015).

181. IEEE-830-1984. *IEEE Guide to Software Requirements Specifications* (ISBN 0-7381-4418-5, 1984).