Aerobiologia (2021) 37:15–28
https://doi.org/10.1007/s10453-020-09669-1

ORIGINAL PAPER

Testing the Raman parameters of pollen spectra in automatic identification

S. G. Pereira · A. Guedes · I. Abreu · H. Ribeiro

Received: 4 June 2020 / Accepted: 20 October 2020 / Published online: 28 October 2020

Abstract Pollen identification and quantification are used in many fields of application and research has been conducted to attain accurate automatic pollen recognition aiming to reduce the laborious work and subjectivity in human identification. The aim of our study was to evaluate the capacity of Raman parameters of pollen spectra, calculated for only 7 common band intervals in a limited spectral range, to be used as future technique in pollen automatic identification. There were analyzed 15 different pollen species considered to induce allergic reactions. Raman spectra were acquired at an excitation wavelength of 785 nm in a spectral region from 1000 to 1800 cm^{-1}, preprocessed and deconvoluted to determine the Raman parameters: wavenumber, full width at half maximum of the band and integrated intensity. Seven common band intervals of all Raman spectra, in the fingerprint areas 1000–1010, 1300–1460 and 1500–1700 cm^{-1}, were chosen for the classification of the pollen species using SVM (support vector machine). Our results showed that the classification accuracy of all pollen species was 100% in the training step, while in the testing step 14 out of the 15 pollen species were correctly assigned (93.3%), including the discrimination between 5 Poaceae species and between *Betula pendula* and *Corylus avellana*. It was also observed that all Raman parameters are important in the classification as well as all wavenumber areas considered. So, our study indicates that the Raman parameters of pollen spectra can be a promising methodology for automatic pollen recognition.

Keywords Pollen classification · Raman spectra · Spectroscopy · Support vector machine

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10453-020-09669-1) contains supplementary material, which is available to authorized users.

S. G. Pereira · A. Guedes · H. Ribeiro (✉)
Department of Geosciences, Environment and Spatial Plannings, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal
e-mail: helena.ribeiro@fc.up.pt

A. Guedes · I. Abreu · H. Ribeiro
Earth Sciences Institute (ICT), Pole of the Faculty of Sciences, University of Porto, Porto, Portugal

I. Abreu
Department of Biology of the Faculty of Sciences, University of Porto, Porto, Portugal

1 Introduction

Pollen analysis has been used in many fields of application such as environmental monitoring (Ribeiro et al. 2015), agriculture (Cunha et al. 2016), paleobotany (Seddon et al. 2019; Schopf et al. 2016), forensic science (Orijemie and Israel 2019; Pereira et al. 2020) and medicine (Lo et al. 2019; Medek et al.

2019). Pollen is one of the most common triggers of season allergic reactions, in some individuals when inhaled causes symptoms due to the proteins that it carries and the numbers of individuals suffering from allergies has grown exponentially in the last years (Sedghy et al. 2017),

Traditionally, pollen identification and quantification are performed manually by light microscope, a process that is time consuming and requires a trained observer to perform it objectively. Moreover, some important pollen types inducing allergies, like *Phleum* or *Dactylis* (García-Mozo 2017) cannot be identify to the genus level. The identification task would benefit with a faster and more resolved identification of pollen species and this is an area where research has been carried out along the years (Rittenour et al. 2012; Sharma-Ghimire et al. 2019).

Image-based applications have been used for pollen identification and other biological particles for a few years (France et al. 2000; Ranzato et al. 2007). It is based on microscopic image analysis through image processing detection techniques, and the introduction of texture characterization in the identification has led to improvements in the classification performance of the distinct pollen types (Marcos et al. 2015). More recently, the implementations of real-time automatic pollen recognition systems based on image processing techniques (Oteros et al. 2015) and digital holographic images (Sauvageat et al. 2020), have showed good results in the online identification of a number of pollen taxa.

Besides pollen morphological features that provide a good taxonomic distinction at the family, genus and even in some cases at the species level, pollen grains also present several differences concerning molecular features and chemical composition that allow identification (Depciuch et al. 2018; Zimmermann 2018). Recently, new methods in pollen identification and quantification have been developed, foreseeing automatic pollen identification (Šantl-Temkiv et al. 2020).

DNA-based techniques have been used for pollen identification and quantification in order to substitute the traditional methodology. The air samples analyze are collected by standard methods, and the DNA extraction occurs afterward following optimized methodologies (Rojo et al. 2019; Bell et al. 2016) specially because the pollen DNA extraction is challenging and some problems involving pollen abundance quantification may need other resolution (Baksay et al. 2020). Some studies have showed that this method could provide accurate qualitative discrimination among grass species (Brennan et al. 2019; Kraaijeveld et al. 2015) that until now were not possible to distinguish with image processing techniques.

Several spectrophotometric techniques have been tested and applied aiming automatic pollen identification. Fourier transformation infrared spectroscopy (FTIR Muthreich et al. 2020; Xu et al. 2018; Zimmermann and Kohler 2014), ultraviolet light induced fluorescence (UV-LIF) (Ruske et al. 2018; Forde et al. 2019), fluorescent spectroscopy (Mularczyk-Oliwa et al. 2012; Zhang et al. 2019) and Raman spectroscopy (Wang et al. 2015). At first, the different approaches were used only to discriminate bioaerosol, and in some cases even pollen, from other materials, biological or not, present in the air. In particular, the UV-LIF technique has evolved to a more elaborated system to distinguish between pollen families and genus. Fluorescence-based equipment is being used in the discrimination of materials in the air, nonbiological and biological compounds are easily distinguished due to intrinsic characteristics; however, bioaerosols like pollen and fungal spores are proving more challenge (Forde et al. 2019). Bağcıoğlu et al. (2015) tested 7 different FTIR and Raman spectroscopy methodologies to the same pollen samples and conclude that Raman microspectroscopy measurements, which are focused on the corpus region of pollen grains, achieved one of the best taxonomic-based differentiation of pollen.

The detection system and the collection of the data are just a part of an automatic pollen identification protocol, the data analysis/classification (Okwuashi and Ndehedehe 2020) is presently one important subject. Development in data science has given a valuable input into pollen classification based on pollen spectroscopic features. Some studies are using machine learning techniques for classification as supervised learning, where the whole data set is divided into training and testing set, as it happen in SVM (support vector machine), NN (neural networks) or *k*-nearest neighbors, while others choose unsupervised learning algorithms, where the data are analyzed as 1 group as, e.g., hierarchical cluster analysis or *k* means (Swanson and Huffman 2020).

The use of Raman spectroscopy in the pursuit of automatic pollen identification is not a new research field (Mondol et al. 2019; Wang et al. 2015; Ivleva et al. 2005; Schulte et al. 2008) but recent developments in term of classification algorithms, high-throughput screening (Mondol et al. 2019) and possible identification of airborne pollen (Doughty and Hill 2020; Guedes et al. 2014) can allow the increase in single pollen's spectra resolution and therefore better discrimination of pollen samples.

Raman spectroscopy is a nondestructive technique, that doesn't require sample preparation, which comes as an advantage to other techniques suggested for pollen identification and yet it is possible to analyze aqueous or air samples with minimal interference (Weiss et al. 2019; Guedes et al. 2014).

Raman spectroscopy evolved along the years, in the beginning was used to identify pollen of known samples (controls) to separate them and in the testing of different wavelengths to ascertain the best suited for pollen (Ivleva et al. 2005). Also, the bands in a Raman spectrum are characteristic and may be assigned to specific chemical compounds which makes it possible to discriminate them. The assignment of the pollen spectrum bands and the correlation of distinct pollen taxa that they seem to generate another important use of Raman spectroscopy to pollen identification and characterization (Schulte et al. 2008). Pollen grains have a characteristic of high fluorescence spectrum, and that has been a working issue that researchers using Raman spectroscopy must address. To enhance the information extracted and reduce noise, a variety of spectrum preprocessing techniques have been used as a baseline correction, normalization and smoothing (Fukuhara et al. 2019).

Additionally, Raman parameters obtained after deconvolution of the spectrum such as the wavenumber and other parameters as the intensity, the integrated intensity and the FWHM (full width at half maximum of the band) remarks to chemical compounds of the pollen wall and can be characteristic for a specific taxa.

Therefore, in this work we aim to evaluate the capacity of the Raman parameters of pollen spectra to be used as future technique in pollen automatic identification by simplifying the data acquisition and reducing the volume of information to analyze. We tested the use of parameters of only 7 common band intervals for all pollen species tested and used support vector machine with a data science software for the classification.

2 Material and methods

2.1 Pollen collection

The pollen samples analyzed by Raman microspectroscopy were collected, during the flowering season in 2018, in the Porto city, from gardens of the Faculty of Sciences of the University of Porto campus and in public parks. 15 different pollen species were analyzed from trees, shrubs and weeds (Table 1) consider to induced allergic reactions (Galán et al. 2017). Three plants per each species were sampled, and flowers/catkins were randomly collected from all quadrants of the plants, in different branches, until a small plastic box was filled. After separation of the anthers from the other plant structures, the anthers were dried at 25 °C during 24 h, after that time, shivered through different grades of sieves to separate the pollen from the rest of the plant materials. Pure pollen was then collected. The samples were stored at − 20 °C until analysis (Ribeiro et al. 2017).

2.2 Raman spectra acquisition and processing

Before the analyses, pollen samples were taken from the storage and left 10 min at room temperature.

Table 1 Pollen analyzed in the study, divided in type of plant and in pollen family

Plant type	Pollen family	Pollen species
	Aceraceae	*Acer negundo*
	Asteraceae	*Artemisia vulgaris*
	Betulaceae	*Alnus glutinosa*
		Betula pendula
		Corylus avellana
Trees and shrubs	Cupressaceae	*Cupressus lusitanica*
	Fagaceae	*Quercus robur*
	Oleaceae	*Fraxinus floribunda*
	Platanaceae	*Platanus x acerifolia*
	Salicaceae	*Salix atrocinerea*
Grasses	Poaceae	*Anthoxanthum adoratum*
		Dactylis glomerata
		Holcus lanatus
		Lagurus ovatus
		Lolium perenne

Raman spectra were acquired by an XploRA™ Raman microscope (Horiba Scientific, France) that combines optical microscopy with a Raman spectroscopy using a laser radiation which allows a “one shot” analysis. A 100 × objective lens was used to focus the laser beam on the sample and also to collect the Raman scattered radiation in backscattering geometry. The Raman signal was detected on a highly sensitive cooled charge-coupled device (CCD) detector.

Prior to each measurement, the Raman spectrum wavenumber was calibrated using a Si reference standard (520.6 ± 0.1 cm^{-1}). Pollen samples were placed on a glass slide, and for each species 3 spectra from 3 different pollen grains were collected at an excitation wavelength of 785 nm from a diode laser at a power of 25 mW with a range of diffraction gratings with 1200 lines mm-1 and slit of 300 µm. Extended scans were performed, with 5 scans of 50 s each measured on each pollen grain, in a spectral region from 1000 to 1800 cm^{-1} with approximately 1 cm^{-1} resolution.

Raman spectra were preprocessed involving an automatic polynomial baseline correction to attenuate the fluorescence influence followed by a denoise procedure using the Savitsky–Golay algorithm to increase spectra quality. The spectra were then normalized to a constant area, where the area under the curve is set to 100 (a.u.).

Afterward, each spectrum was deconvoluted using a mixed Gaussian–Lorentzian curve-fitting procedure to determine the precise Raman parameters: wavenumber (W), full width at half maximum of the band (FWHM) and integrated intensity (A). To reduce the influence of the natural variability of the intensity of the spectrum a new parameter was calculated, R_area (pondered area), the ratio between the integrated intensity with the total integrated intensity of the deconvolution curve. For the fit of the spectral sets, 18 bands were used, which correspond to the aggregate of principal bands present in the distinct pollen spectra.

The software LabSpec 6 (Horiba Scientific, France) was used for spectra acquisition and deconvolution.

2.3 Data analysis and pollen classification

A matrix with all the Raman parameters obtained for each of the 18 bands considered in the deconvolution process was created. Only the seven common band intervals to all Raman spectra, chosen by visual inspection were used in the classification of the pollen species. The common band intervals were the ones in the fingerprint areas: 1000–1010, 1300–1460 and 1500–1700 cm^{-1} and therefore the data matrix presented a total of 21 feature classifiers.

For the pollen classification analysis, it was used the open-source Orange 3.24.1. software package, with tools for data visualization and analysis, data mining and machine learning (Demsar et al. 2013).

The potential of Raman parameters to accurately classify the distinct pollen species was evaluated applying a supervised learning algorithm—SVM (support vector machine). SVM is based on the concept of finding a design function that best separates the analyzed features in different groups. A hyperplane represents that separation and the best hyperplane is the one that maximizes the distance between features and therefore gives the best classification or regression. This represents a linear classifier, but its usual to find nonlinear distribution for the data, and in that case, kernel functions are used (set of mathematical functions that allow for a nonlinear decision surface to be transformed into a linear higher dimensional space), the objective remains but the hyperplane adjusts differently to the data.

Our data matrix is composed of 45 spectra × 21 classifier features and in order to minimize fitting problems at the classification algorithm, we randomly divided the data into 2 sets, one training group with 66.7% of the spectra (2 per pollen species, 30 total spectra) to estimate the best classification model and a separate testing group with 33.3% of the study cases (1 per pollen species, 15 total spectra). A radial basis function (RBF) kernel was selected and the tuning of its γ and c parameters was performed by testing several combinations until the best train-test classification was met (c = 1.6 and γ = 0.05). RBF kernel is a commonly used general kernel functions in SVM classification and is defined as $K_{RBF}(x, x') = \exp[-\gamma\|x - x'\|^2]$. The γ parameter allows to define how far is the influence reach of a single training example, while the c parameter (common to all SVM kernels) will act as a trade-off between a correct classification of training examples against maximization of the decision function’s margin, the smaller the value of c the larger margin will be accepted at the cost of training accuracy (Sammut and

Webb 2011). The precision (ratio of correctly classified objects to all object that should truly be correctly classified) and classification accuracy (ratio of correct classification to total classifications made) attained was analyzed, and a confusion matrix (two-dimensional table, where one dimension corresponds to the true class of an object and the other to the class that the classifier assigns) was used to summarize the performance of the classification algorithm (Sammut and Webb 2011).

3 Results and discussion

3.1 Spectra analysis

Raman spectra give information about the pollen chemical characterization, containing specific signals of macronutrients such as lipids, proteins, carbohydrates, water and even some pigments (Zimmermann 2010; Schulte et al. 2008; Zimmermann and Kohler 2014; Bağcıoğlu et al. 2015; Pummer et al. 2013; Kenđel and Zimmermann 2020; Weglinska et al. 2020). As a result, the spectra are quite complex and variable between different genera and even species, which can also be noticed in our results.

The Raman spectra obtained for the 15 different pollen species show distinct 18 bands, characteristic of each pollen species, and were selected due to being the ones that improve the deconvolution fitting line, in the functionality region between 1000 and 1800 cm^{-1} (Fig. 1), but only 7 band intervals were common to all studied species, distributed in three fingerprint regions (Fig. 2). In fact, the average Raman spectra present some differences between the studied species, being possible to distinguish particularities between the spectra of tree and grass species.

These 3 fingerprint regions are defined by the following 7 common band intervals at about [1000–1010 cm^{-1}], [1305–1335 cm^{-1}], [1340–1375 cm^{-1}], [1440–1460 cm^{-1}], [1525–1600 cm^{-1}], [1600–1615 cm^{-1}] and [1650–1665 cm^{-1}].

In the 1500–1700 cm^{-1} fingerprint area, the bands assigned to nucleic acids (adenine and guanine) by Diehn et al. (2020) were found mostly at ≈ 1580–1590 cm^{-1} for trees and at approximately 1565 cm^{-1} in the grass species. The exception is *Anthoxanthum adoratum* with a peak ≈ 1530 cm^{-1}, assigned to carotenoids (Diehn et al. 2020), this is the band with more heterogeneity of peak values.

The tree spectra also present a well-defined band in the interval [1600–1615 cm^{-1}] and 1 or 2 less intense bands (at 1525–1600 and 1650–1665 cm^{-1}), one before and other after one higher intensity peak in the region [1600–1615 cm^{-1}], most of times showed as shoulders more or less defined, while in grasses, 2, 3 or 4 medium–low intensity bands are observed in the same region. The band with the higher intensity in all tree spectra, is around 1608 cm^{-1} and has been assigned to mitochondrial activity but also to the ferulic acid and coumaric acid building blocks in sporopollenin (Diehn et al. 2020). This band is present

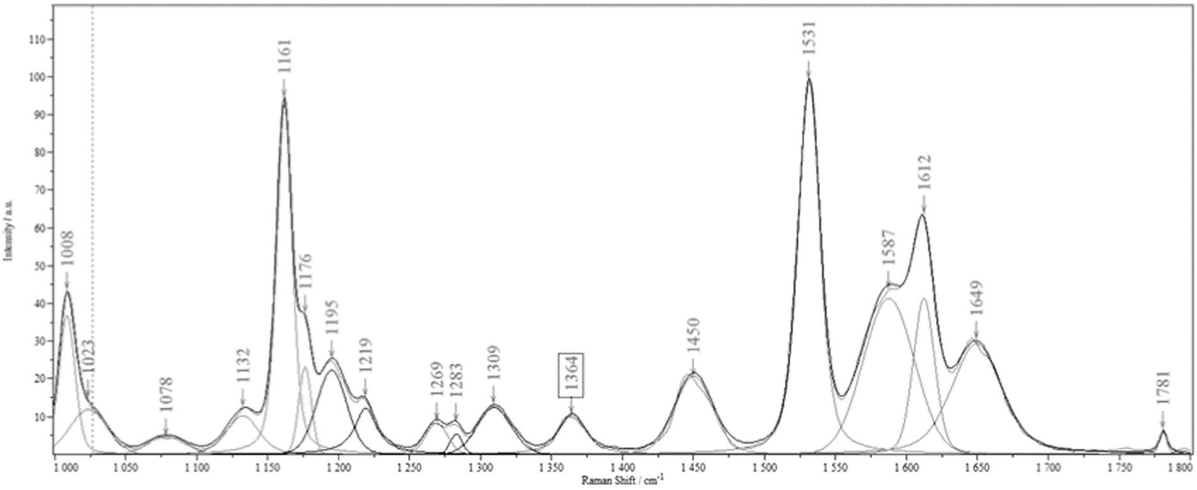
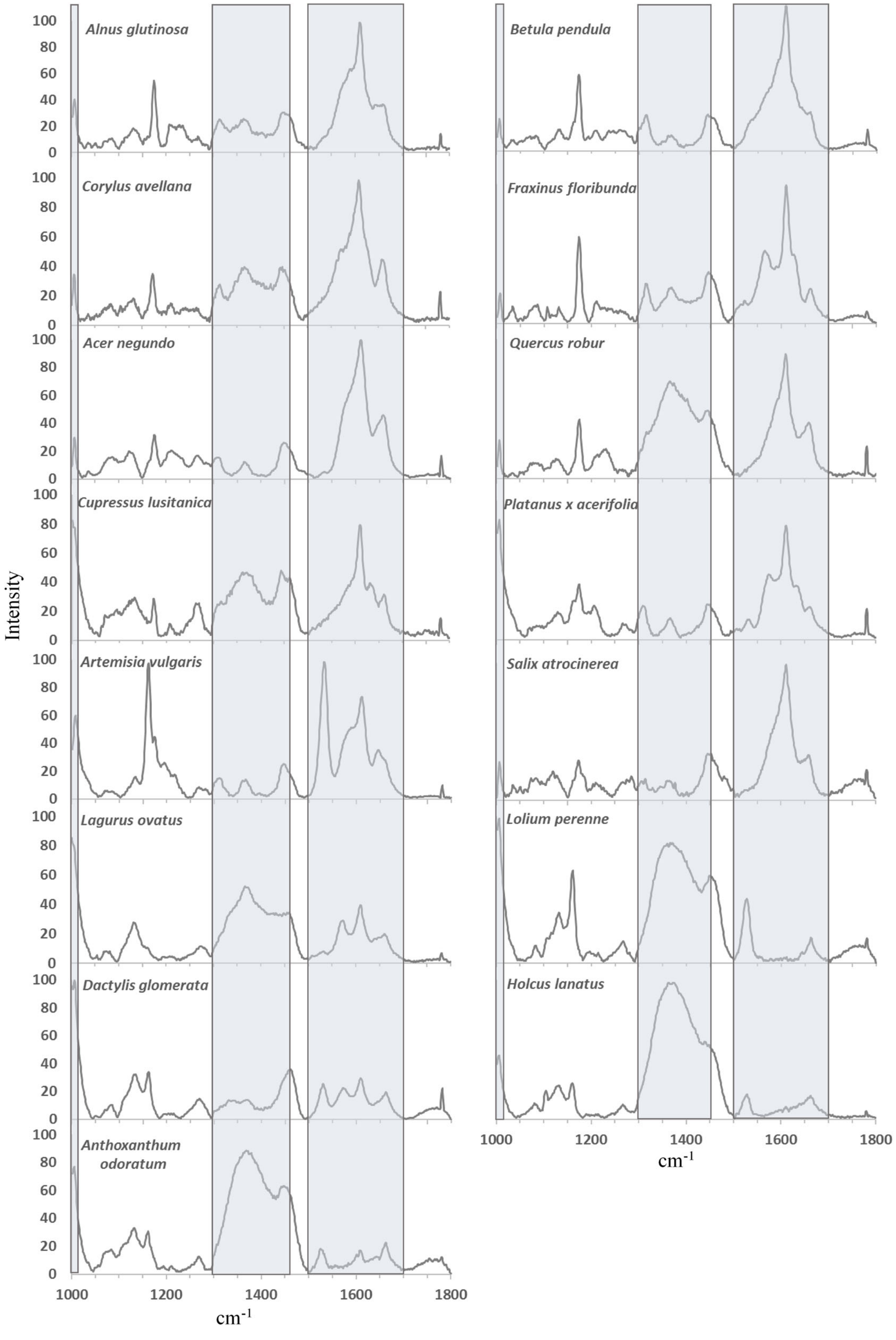


Fig. 1 Example of a spectrum of *Artemisia vulgaris* with a total of 18 bands deconvolution


Intensity
Alnus glutinosa
Corylus avellana
Acer negundo
Cupressus lusitanica
Artemisia vulgaris
Lagurus ovatus
Dactylis glomerata
Anthoxanthum odoratum
Betula pendula
Fraxinus floribunda
Quercus robur
Platanus x acerifolia
Salix atrocinerea
Lolium perenne
Holcus lanatus
0
20
40
60
80
100
1000
1200
1400
1600
1800
cm-1

◀**Fig. 2** Average Raman spectra of the 15 pollen species analyzed and main fingerprint areas marked in gray (1000–1010, 1300–1460 and 1500–1700 cm^{-1})

in all pollen species analyzed, less in *Lolium perenne* where the higher intensity peak appears at ≈ 1600 cm^{-1}, and has been assigned to phenylalanine and tyrosine (Guedes et al. 2014) or to ring stretches of phenyl structures (Ivleva et al. 2005).

A common band observed in the interval [1650–1665 cm^{-1}] was the one at ≈ 1662 cm^{-1} that has been assigned to vibrations of proteins (Diehn et al. 2020; Schulte et al. 2008) and is present in all pollen species with the exception of *Artemisia vulgaris* pollen where a band is observed in the 1650 cm^{-1} position and may be assigned to Amide I system (C = O) (Guedes et al. 2014; Ivleva et al. 2005).

The fingerprint area between 1300–1460 cm^{-1} can be considered characteristic of grass species. In this region, compared with other species, a wide band with high intensity is observed in the interval [1340–1375 cm^{-1}], with most grass species presenting the peak at ≈ 1370 cm^{-1}, the exception is *Dactylis glomerata*. For tree species, this area is quite different, being observed a set of smaller bands around 1360 cm^{-1}, that can be assigned to nucleic acids (adenine and guanine) (Diehn et al. 2020). Also, in this fingerprint area (1300–1460 cm^{-1}) 2 peaks are found at ≈ 1450 cm^{-1} (in all species) and at ≈ 1313 cm^{-1} (in all trees and in *Dactylis glomerata* pollen) that correspond, respectively, to deformation made of C–H$_2$ groups of aliphatic carbon chains (Guedes et al. 2014) and to ferulic acid and coumaric acid building blocks in sporopollenin. In the other grass species, the peak is shifted to ≈ 1322 cm^{-1}, that is associated to carbohydrates (Diehn et al. 2020).

Finally, in a third fingerprint area (1000–1010 cm^{-1}) a band at ≈ 1006 cm^{-1} is characteristic of all pollen and can be assigned to carotenoids (Diehn et al. 2020; Schulte et al. 2008).

3.2 Classification analysis

The Raman spectra complexity and variability between distinct pollen types also enables its identification and classification, by applying data analysis, to taxonomic levels that are many times not possible to be discriminated by humans under light microscopy (Kraaijeveld et al. 2015; Mondol et al. 2019).

Our study assessed the possibility of the Raman parameters of the seven common bands to all pollen species to be sufficient for the classification process, a different approaches to what has been usually done in other studies that use information of the full or reduced spectral range. We must highlight that this was a preliminary proof-of-concept for the methodological approach using a smaller spectra data set.

The classification potential was evaluated in three combinations: all 15 tested species data sets, and by plant´s habitat only tree species and only grass species. The best classification possible of these data sets is achieved when used the Raman parameters of the wavenumber (W), full width at half maximum of the band (FWHM) and integrated intensity (A) (Table 2). In our study, the R_area (pondered area) parameter did not improve the classification.

The classification performance using all pollen species was very high, being perfect in the training step with a classification accuracy (CA) of 100% and a precision of 100%, while in the testing step 14 out of the 15 pollen species were correctly assigned (precision of 90% and CA of 93.3%). The exception was *Salix atrocinerea* pollen, which was misclassified as *Acer negundo*. It was possible to perform the distinction between pollen from *Betula pendula* and *Corylus avellana*, 2 taxa belonging to the same family presenting very similar morphologies, which can pose some classification challenges for the methods based on image processing (Sauvageat et al. 2020).

Comparing our classification accuracy with the ones obtained when the full or reduced spectral range is used by other authors in the pollen discrimination (Diehn et al. 2020; Ivleva et al. 2005; Zimmermann and Kohler 2014), it is possible to see that the Raman parameters can be a good alternative to pollen good classification, but it must be kept in mind that our spectra data set is small. As observed by Schulte et al. (2008), even though pollen taxa related to the genus and family level present chemical similarities, which are indicative of both phylogenetic relationship and mating behavior, in our study it was possible to discriminate between the distinct pollen species.

Using the information of only seven common band intervals, we are able to reduce the volume of data necessary to classify the pollen species as well as the

Table 2 – Confusion matrix resulted from the SVM analysis on test step of the Raman parameters (wavenumber, full width at half maximum of the band and integrated intensity) of the 7 common wavenumbers from the Raman spectra of the pollen from 15 plant species

Classification %	*Acer negundo*	*Alnus glutinosa*	*Betula pendula*	*Corylus avellana*	*Cupressus lusitanica*	*Fraxinus floribunda*	*Platanus x acerifolia*	*Quercus robur*	*Salix atrocinerea*	*Artemisia vulgaris*	*A. adoratum*	*Dactylis glomerata*	*Holcus lanatus*	*Lagurus ovatus*	*Lolium perenne*
Acer negundo	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alnus glutinosa	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
Betula pendula	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
Corylus avellana	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
Cupressus lusitanica	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
Fraxinus floribunda	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
Platanus x acerifolia	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
Quercus robur	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
Salix atrocinerea	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Artemisia vulgaris	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
A. adoratum	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
Dactylis glomerata	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
Holcus lanatus	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
Lagurus ovatus	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
Lolium perenne	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

time of analysis and spectra acquisition due to the small spectral region studied.

Table 3 shows the confusion matrix for the tree pollen species classification corresponding to the test step.

The trees CA declined a little compared when all species were used, but the majority were accurately classified. The exceptions were *Salix atrocinerea* that remains misclassified as *Acer negundo* and now *Alnus glutinosa* is classified as *Quercus robur*. For this analysis we obtained a CA of 77.8% and a precision value of 66.7%. As it happens for the total species SVM analysis, in the train step, the CA and precision value were 100%.

It is interesting to observe that *Alnus* pollen was correctly discriminated when all studied species were considered. When we use only the Raman parameters of common band in tree pollen, we are distinguishing among more similar spectra. The Raman spectrum of *Quercus robur* pollen in the 1300–1460 cm^{-1} fingerprint areas has much more similarities to the grass spectra and therefore when grasses are included in the training step the *Q. robur* would be set further from the tree species. Additionally, for the same fingerprint area, *Salix atrocinerea* and *Acer negundo* spectra are very similar and for that these similarities can cause the CA decline.

When only grass species are tested, the classification renders the best performance with all species being correctly classified in both the training and testing steps (Table 4), with a CA and precision value of 100%.

Through high-throughput screening Raman spectroscopy (HTS-RS), Mondol et al. (2019) used the Raman spectra fingerprint region (758–1800 cm^{-1}) from pollen of 15 genera belonging to the Poaceae family and applied PCA-SVM for their classification. The predictions among Poaceae genera were high (around 79% accuracy and sensitivity of 80%), but the number of pollen grains/species analyzed was much higher compared with our study, which justify the lowest classification performance. In our study, we tested a small data set to ascertain the possibility of the Raman parameters to be sufficient for the classification process.

We tested also if the Raman peaks observed for Poaceae pollen species in the fingerprint area 1300–1460 cm^{-1}, with distinct spectral features among the tested species, could be enough for a correct classification among them. It was observed that

Table 3 Confusion matrix resulted from the SVM analysis on test step of the Raman parameters (wavenumber, full width at half maximum of the band and integrated intensity) of the 7 common wavenumbers from the Raman spectra of the pollen from tree plant species

Classification %	*Acer negundo*	*Alnus glutinosa*	*Betula pendula*	*Corylus avellana*	*Cupressus lusitanica*	*Fraxinus floribunda*	*Platanus x acerifolia*	*Quercus robur*	*Salix atrocinerea*
Acer negundo	100	0	0	0	0	0	0	0	0
Alnus glutinosa	0	0	0	0	0	0	0	100	0
Betula pendula	0	0	100	0	0	0	0	0	0
Corylus avellana	0	0	0	100	0	0	0	0	0
Cupressus lusitanica	0	0	0	0	100	0	0	0	0
Fraxinus floribunda	0	0	0	0	0	100	0	0	0
Platanus x acerifolia	0	0	0	0	0	0	100	0	0
Quercus robur	0	0	0	0	0	0	0	100	0
Salix atrocinerea	100	0	0	0	0	0	0	0	0

Table 4 Confusion matrix resulted from the SVM analysis of the Raman parameters (wavenumber, full width at half maximum of the band and integrated intensity) of the 7 common band intervals from the Raman spectra of the pollen from 5 grass species

Classification %	*A. adoratum*	*Dactylis glomerata*	*Holcus lanatus*	*Lagurus ovatus*	*Lolium perenne*
Anthoxanthum adoratum	100	0	0	0	0
Dactylis glomerata	0	100	0	0	0
Holcus lanatus	0	0	100	0	0
Lagurus ovatus	0	0	0	100	0
Lolium perenne	0	0	0	0	100

it was not sufficient for a good classification, in the train step the value of accuracy was 80% and the precision of 83.3% where *A. adoratum* was misclassified as *H. lanatus* and *H. lanatus* as *L. ovatus*.

Distinction among airborne Poaceae genera, or even species if possible, is important in terms of pollen-related allergy issues. Grass species are one of the most common and higher allergenic species and their wide distribution around the globe as well as number of species, causing several allergic reactions in susceptible individuals (García-Mozo 2017). However, not all grass species induce allergies, but a few genera like *Lolium* spp., *Dactylis* spp., *Anthoxanthum* spp., *Phleum* spp., among others, are the most allergic ones (Brennan et al. 2019; García-Mozo 2017). With an extensive flowering season, that lasts around 4–5 months between March–July and September, and with several annual peaks in airborne pollen concentration it makes months of suffering for grass pollen allergen suffers (Ribeiro and Abreu 2014). Presently, the grass airborne pollen season is not discriminated by the different genera or species, and the pollen season of the most allergenic ones may be common to other type of pollen season what may enhance the allergic individual reaction (García-Mozo 2017). So, it becomes clear the importance to identify the different airborne pollen contributors, and among the Poaceae it is a real challenge to exactly defined the traits of the flowering seasons, beginning and ending, for the different genera (Brennan et al. 2019). In fact, the morphological similarities among the airborne pollen from the different Poaceae genera makes almost impossible their distinction. Features such as number of apertures, shape and texture are quite similar, posing great analytical challenges to image processing algorithms (Ronneberger et al. 2002), although Poaceae pollen morphology is so typical that are easily distinguished among other airborne non-Poaceae pollen. With our study it was possible to distinguish between 5 different species of Poaceae by using the Raman parameters of only 7 common band intervals from the full pollen Raman spectrum.

In our study, we also tested the contribution of each Raman parameter or their different combinations in pollen classification (Fig. 3).

It was observed that the combination of two or more parameters gave better results than using only a single parameter. In Fig. 3 (left side), we can see the precision values obtained in the training and testing steps.

Considering each Raman parameter alone, the integrated intensity (A) is the one that less contributes to the classification (in training: 59% and in testing: 37%). The wavenumber (W) and full width at half maximum of the band (FWHM) showed equal precision in the testing step (63%) but the FWHM achieved best classification in the train step (90%).

When the parameters are combined in groups of 2, the W + FWHM gave better performance than the other combinations. However, only the combination of all parameters allows the best classification with 100% CA in the training step and 90% in the test one.

Finally, with the question in mind if all the 7 common wavenumbers are important to the discrimination of the pollen from the 15 studied species and therefore avoid overfitting of the classification algorithm, we tested removing the parameter's data of each wavenumber considered at a time without any other change in the remaining data. It was observed that all wavenumbers are important for the correct classification of the pollen species. CA in the training step and even less in the test one was very low when any wavenumber is removed (Fig. 3).

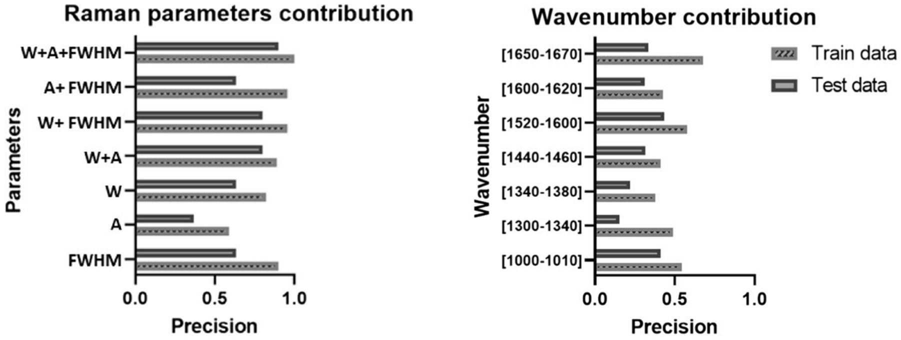


Fig. 3 Contribution of each Raman parameter (wavenumber (W), full width at half maximum of the band (FWHM) and integrated intensity (A)) or their combination (graphic on the right) and each of the 7 common band intervals (graphic on the left) in the pollen classification performance. Values of the precision obtained using SVM analysis in Orange software with the same configuration used in the classification of the species. Both train and test sets of data were analyzed

One interesting observation was when we test the removal of only the last band interval [1650–1670 cm^{-1}], the discrimination in the train step is negatively affected although in a smaller percentage (precision value of 94%) when compared with the removal of the other intervals.

The results shift from no misclassification, to 50% wrong classification for a few species but all trees (*Alnus glutinosa* pollen misclassify as *Betula pendula*, *Salix atrocinerea* as *Acer negundo*, *Corylus avellana* as *Fraxinus angustifolia*) being still possible to make distinction between the tree and the grass species. This behavior was not observed when all the other intervals were removed at a time, and grasses were misclassified as trees and vice versa. So, the band interval [1650–1670 cm^{-1}], seems to contribute to the classification between trees and grasses as well as among the different tree species. This band has been assigned to the vibrations of proteins (Diehn et al. 2020; Schulte et al. 2008) and the differences can be due to distinct molecular conformations arrising from distict protein content in each species.

All parameters are important in the classification, the wavenumber values are one of the most important though this parameter alone can be tricky. The calibration made in the equipment it is basically a calibration of the wavelength, and that if not taken as a routine can induce differences in this parameter. Zimmerman et al. (2014) described small shifts in the wavenumber position, even in pollen spectra of the same species in different geographical regions. In fact, in our study, in the test group the performance of this parameter alone is not good.

The proposed methodology in our study could be a promising approach for Raman-based automatic pollen classification, however, one drawback in the small data set used in the training and testing of the classification algorithm. It would be interesting to test in the future the efficiency of the 7 common band intervals in discriminating between the studied pollen types using high-throughput analysis methodology.

4 Conclusion

Our study focused on testing the possibility of using the band Raman parameters: wavenumber (W), full width at half maximum of the band (FWHM) and integrated intensity (A) instead of the all spectrum into pollen classification. All parameters are important in the classification, with the wavenumber and FWHM, contributing the most to the classification.

The results obtained proved to be possible, using the Raman parameters of 7 band intervals, common to all pollen types, to achieve a successful classification of different pollen species. Fourteen out of 15 pollen species were discriminated including some that are morphologically very difficult or even impossible to identified by the human eye, e.g., between 5 Poaceae species and between 2 species of Betulaceae, as *Betula pendula* and *Corylus avellana*.

It would be interesting to further test the proposed methodology using a larger number of species, including fresh pollen and more Poaceae species, as well as the minimal acquisition time to still achieve a precise classification.

Acknowledgements This work was funded by National Funds through FCT—Fundação para a Ciência e a Tecnologia for the project PTDC/ATP-EAM/0817/2014. ICT—Earth Sciences Institute is funded by National Funds through FCT projects UIDB/04683/2020 e UIDP/04683/2020. The authors thank Pereira J. for her help with Raman spectra acquisition.

References

Bağcıoğlu, M., Zimmermann, B., & Kohler, A. (2015). A multiscale vibrational spectroscopic approach for identification and biochemical characterization of pollen. *PLoS ONE, 10*(9), e0137899. https://doi.org/10.1371/journal.pone.0137899

Baksay, S., Pornon, A., Burrus, M., Mariette, J., Andalo, C., & Escaravage, N. (2020). Experimental quantification of pollen with DNA metabarcoding using ITS1 and trnL. *Scientific Reports, 10*(1), 4202. https://doi.org/10.1038/s41598-020-61198-6

Bell, K. L., de Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., et al. (2016). Pollen DNA barcoding: Current applications and future prospects. *Genome, 59*(9), 629–640. https://doi.org/10.1139/gen-2015-0200

Brennan, G. L., Potter, C., de Vere, N., Griffith, G. W., Skjoth, C. A., Osborne, N. J., et al. (2019). Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. *Nature Ecology and Evolution, 3*(5), 750–754. https://doi.org/10.1038/s41559-019-0849-7

Cunha, M., Ribeiro, H., & Abreu, I. (2016). Pollen-based predictive modelling of wine production: Application to an arid region. *European Journal of Agronomy, 73,* 42–54. https://doi.org/10.1016/j.eja.2015.10.008

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research, 14*(35), 2349–2353.

Depciuch, J., Kasprzyk, I., Drzymała, E., & Parlinska-Wojtan, M. (2018). Identification of birch pollen species using FTIR spectroscopy. *Aerobiologia, 34*(4), 525–538. https://doi.org/10.1007/s10453-018-9528-4

Diehn, S., Zimmermann, B., Tafintseva, V., Seifert, S., Bağcıoğlu, M., Ohlson, M., et al. (2020). Combining chemical information from grass pollen in multimodal characterization. *Fronteirs in Plant Science*. https://doi.org/10.3389/fpls.2019.01788

Doughty, D. C., & Hill, S. C. (2020). Raman spectra of atmospheric particles measured in Maryland, USA over 22.5 h using an automated aerosol Raman spectrometer. *Journal of Quantitative Spectroscopy and Radiative Transfer, 244,* 106839. https://doi.org/10.1016/j.jqsrt.2020.106839

Forde, E., Gallagher, M., Walker, M., Foot, V., Attwood, A., Granger, G., et al. (2019). Intercomparison of multiple UV-LIF spectrometers using the aerosol challenge simulator. *Atmosphere, 10*(12), 797.

France, I., Duller, A. W. G., Duller, G. A. T., & Lamb, H. F. (2000). A new approach to automated pollen analysis. *Quaternary Science Reviews, 19*(6), 537–546. https://doi.org/10.1016/S0277-3791(99)00021-9

Fukuhara, M., Fujiwara, K., Maruyama, Y., & Itoh, H. (2019). Feature visualization of Raman spectrum analysis with deep convolutional neural network. *Analytica Chimica Acta, 1087,* 11–19. https://doi.org/10.1016/j.aca.2019.08.064

Galán, C., Dahl, A., Frenguelli, G., & Gehrig, R. (2017). Airborne pollen in Europe. In A. B. Singh (Ed.), *Allergy and allergen immunotherapy: New mechanisms and strategies* (pp. 127–162). Toronto, NJ: Apple Academic Press.

García-Mozo, H. (2017). Poaceae pollen as the leading aeroallergen worldwide: A review. *Allergy, 72*(12), 1849–1858. https://doi.org/10.1111/all.13210

Guedes, A., Ribeiro, H., Fernández-González, M., Aira, M. J., & Abreu, I. (2014). Pollen Raman spectra database: Application to the identification of airborne pollen. *Talanta, 119,* 473–478. https://doi.org/10.1016/j.talanta.2013.11.046

Ivleva, N. P., Niessner, R., & Panne, U. (2005). Characterization and discrimination of pollen by Raman microscopy. *Analytical and Bioanalytical Chemistry, 381*(1), 261–267. https://doi.org/10.1007/s00216-004-2942-1

Kenđel, A., & Zimmermann, B. (2020). Chemical analysis of pollen by FT-Raman and FTIR spectroscopies. *Fronteirs in Plant Science*. https://doi.org/10.3389/fpls.2020.00352

Kraaijeveld, K., de Weger, L. A., Ventayol García, M., Buermans, H., Frank, J., Hiemstra, P. S., et al. (2015). Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resourses, 15*(1), 8–16. https://doi.org/10.1111/1755-0998.12288

Lo, F., Bitz, C. M., Battisti, D. S., & Hess, J. J. (2019). Pollen calendars and maps of allergenic pollen in North America. *Aerobiologia, 35*(4), 613–633. https://doi.org/10.1007/s10453-019-09601-2

Marcos, J. V., Nava, R., Cristóbal, G., Redondo, R., Escalante-Ramírez, B., Bueno, G., et al. (2015). Automated pollen identification using microscopic imaging and texture analysis. *Micron, 68,* 36–46. https://doi.org/10.1016/j.micron.2014.09.002

Medek, D. E., Simunovic, M., Erbas, B., Katelaris, C. H., Lampugnani, E. R., Huete, A., et al. (2019). Enabling self-management of pollen allergies: A pre-season questionnaire evaluating the perceived benefit of providing local pollen information. *Aerobiologia, 35*(4), 777–782. https://doi.org/10.1007/s10453-019-09602-1

Mondol, A. S., Patel, M. D., Rüger, J., Stiebing, C., Kleiber, A., Henkel, T., et al. (2019). Application of high-throughput screening Raman spectroscopy (HTS-RS) for label-free identification and molecular characterization of pollen. *Sensors, 19*(20), 428. https://doi.org/10.3390/s19204428

Mularczyk-Oliwa, M., Bombalska, A., Kaliszewski, M., Włodarski, M., Kopczyński, K., Kwaśny, M., et al. (2012). Comparison of fluorescence spectroscopy and FTIR in differentiation of plant pollens. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 97,* 246–254. https://doi.org/10.1016/j.saa.2012.05.063

Muthreich, F., Zimmermann, B., Birks, H. J. B., Vila-Viçosa, C. M., & Seddon, A. W. R. (2020). Chemical variations in *Quercus* pollen as a tool for taxonomic identification: Implications for long-term ecological and biogeographical research. *Journal of Biogeography, 47*(6), 1298–1309. https://doi.org/10.1111/jbi.13817

Okwuashi, O., & Ndehedehe, C. E. (2020). Deep support vector machine for hyperspectral image classification. *Pattern Recognition, 103,* 107298. https://doi.org/10.1016/j.patcog.2020.107298

Orijemie, E. A., & Israel, I. (2019). Palynomorphs and travel history of vehicles in Nigeria. *Aerobiologia, 35*(3), 497–510. https://doi.org/10.1007/s10453-019-09577-z

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., et al. (2015). Automatic and online pollen monitoring. *International Archives of Allergy and Immunology, 167*(3), 158–166. https://doi.org/10.1159/000436968

Pereira, J. S. R., Ribeiro, H., & Abreu, I. (2020). Spatial and temporal environmental pollen analysis of footwear worn in the area of Barcelos, North-West Portugal, in a forensic context. *Aerobiologia, 36*(1), 89–94. https://doi.org/10.1007/s10453-019-09598-8

Pummer, B. G., Bauer, H., Bernardi, J., Chazallon, B., Facq, S., Lendl, B., et al. (2013). Chemistry and morphology of dried-up pollen suspension residues. *Journal of Raman Spectroscopy, 44*(12), 1654–1658. https://doi.org/10.1002/jrs.4395

Ranzato, M., Taylor, P. E., House, J. M., Flagan, R. C., LeCun, Y., & Perona, P. (2007). Automatic recognition of biological particles in microscopic images. *Pattern Recognition Letters, 28*(1), 31–39. https://doi.org/10.1016/j.patrec.2006.06.010

Ribeiro, H., & Abreu, I. (2014). A 10-year survey of allergenic airborne pollen in the city of Porto (Portugal). *Aerobiologia, 30*(3), 333–344. https://doi.org/10.1007/s10453-014-9331-9

Ribeiro, H., Costa, C., Abreu, I., & Esteves da Silva, J. C. G. (2017). Effect of O3 and NO2 atmospheric pollutants on Platanus x acerifolia pollen: Immunochemical and spectroscopic analysis. *Science of the Total Environment, 599–600,* 291–297. https://doi.org/10.1016/j.scitotenv.2017.04.206

Ribeiro, H., Guimaraes, F., Duque, L., Noronha, F., & Abreu, I. (2015). Characterisation of particulate matter on airborne pollen grains. *Environmental Pollution, 206,* 7–16. https://doi.org/10.1016/j.envpol.2015.06.015

Rittenour, W. R., Hamilton, R. G., Beezhold, D. H., & Green, B. J. (2012). Immunologic, spectrophotometric and nucleic acid based methods for the detection and quantification of airborne pollen. *Journal of Immunological Methods, 383*(1–2), 47–53. https://doi.org/10.1016/j.jim.2012.01.012

Rojo, J., Núñez, A., Lara, B., Sánchez-Parra, B., Moreno, D. A., & Pérez-Badia, R. (2019). Comprehensive analysis of different adhesives in aerobiological sampling using optical microscopy and high-throughput DNA sequencing. *Journal of Environmental Management, 240,* 441–450. https://doi.org/10.1016/j.jenvman.2019.03.116

Ronneberger, O., Schultz, E., & Burkhardt, H. (2002). Automated pollen recognition using 3D volume images from fluorescence microscopy. *Aerobiologia, 18*(2), 107–115. https://doi.org/10.1023/A:1020623724584

Ruske, S., Topping, D. O., Foot, V. E., Morse, A. P., & Gallagher, M. W. (2018). Machine learning for improved data analysis of biological aerosol using the WIBS. *Atmospheric Measurements Techniques, 11*(11), 6203–6230. https://doi.org/10.5194/amt-11-6203-2018

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Boston, MA: Springer.

Šantl-Temkiv, T., Sikoparija, B., Maki, T., Carotenuto, F., Amato, P., Yao, M., et al. (2020). Bioaerosol field measurements: Challenges and perspectives in outdoor studies. *Aerosol Science and Technology, 54*(5), 520–546. https://doi.org/10.1080/02786826.2019.1676395

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., et al. (2020). Real-time pollen monitoring using digital holography. *Atmospheric Measurements Techniques, 13*(3), 1539–1550. https://doi.org/10.5194/amt-13-1539-2020

Schopf, J. W., Calça, C. P., Garcia, A. K., Kudryavtsev, A. B., Souza, P. A., Félix, C. M., et al. (2016). In situ confocal laser scanning microscopy and Raman spectroscopy of bisaccate pollen from the irati subgroup (Permian, Paraná Basin, Brazil): Comparison with acid-macerated specimens. *Review of Palaeobotany and Palynology, 233,* 169–175. https://doi.org/10.1016/j.revpalbo.2016.03.004

Schulte, F., Lingott, J., Panne, U., & Kneipp, J. (2008). Chemical characterization and classification of pollen. *Analytical Chemistry, 80*(24), 9551–9556. https://doi.org/10.1021/ac801791a

Seddon, A. W. R., Festi, D., Robson, T. M., & Zimmermann, B. (2019). Fossil pollen and spores as a tool for reconstructing ancient solar-ultraviolet irradiance received by plants: An assessment of prospects and challenges using proxy-system modelling. *Photochemical and Photobiological Sciences, 18*(2), 275–294. https://doi.org/10.1039/c8pp00490k

Sedghy, F., Sankian, M., Moghadam, M., Ghasemi, Z., Mahmoudi, M., & Varasteh, A. R. (2017). Impact of traffic-related air pollution on the expression of *Platanus orientalis* pollen allergens. *International Journal of Biometeorology, 61*(1), 1–9. https://doi.org/10.1007/s00484-016-1186-z

Sharma Ghimire, P., Tripathee, L., Chen, P., & Kang, S. (2019). Linking the conventional and emerging detection techniques for ambient bioaerosols: A review. *Reviews in Environmental Science and Bio/Technology, 18*(3), 495–523. https://doi.org/10.1007/s11157-019-09506-z

Swanson, B. E., & Huffman, J. A. (2020). Pollen clustering strategies using a newly developed single-particle fluorescence spectrometer. *Aerosol Science and Technology, 54*(4), 426–445. https://doi.org/10.1080/02786826.2019.1711357

Wang, C., Pan, Y.-L., Hill, S. C., & Redding, B. (2015). Photophoretic trapping-Raman spectroscopy for single pollens and fungal spores trapped in air. *Journal of Quantitative Spectroscopy and Radiative Transfer, 153,* 4–12. https://doi.org/10.1016/j.jqsrt.2014.11.004

Weglinska, M., Szostak, R., Kita, A., Nems, A., & Mazurek, S. (2020). Determination of nutritional parameters of bee pollen by Raman and infrared spectroscopy. *Talanta, 212,* 120790. https://doi.org/10.1016/j.talanta.2020.120790

Weiss, R., Palatinszky, M., Wagner, M., Niessner, R., Elsner, M., Seidel, M., et al. (2019). Surface-enhanced Raman spectroscopy of microorganisms: Limitations and applicability on the single-cell level. *Analyst, 144*(3), 943–953. https://doi.org/10.1039/C8AN02177E

Xu, X.-L., Zheng, Y.-Z., Chen, X.-C., Zhu, F.-L., & Miao, X.-Q. (2018). Identification of cattail pollen, pine pollen and bee pollen by fourier transform infrared spectroscopy and two-dimensional correlation infrared spectroscopy. *Journal of Molecular Structure, 1167,* 78–81. https://doi.org/10.1016/j.molstruc.2018.04.076

Zhang, M., Klimach, T., Ma, N., Könemann, T., Pöhlker, C., Wang, Z., et al. (2019). Size-resolved single-particle fluorescence spectrometer for real-time analysis of bioaerosols: Laboratory evaluation and atmospheric measurements. *Environmental Science and Technology, 53*(22), 13257–13264. https://doi.org/10.1021/acs.est.9b01862

Zimmermann, B. (2010). Characterization of pollen by vibrational spectroscopy. *Applied Spectroscopy, 64*(12), 1364–1373. https://doi.org/10.1366/000370210793561664

Zimmermann, B. (2018). Chemical characterization and identification of pinaceae pollen by infrared microspectroscopy. *Planta, 247*(1), 171–180. https://doi.org/10.1007/s00425-017-2774-9

Zimmermann, B., & Kohler, A. (2014). Infrared spectroscopy of pollen identifies plant species and genus as well as environmental conditions. *PLoS ONE, 9*(4), e95417. https://doi.org/10.1371/journal.pone.0095417