heritage

MDPI

Article

Modeling Chemical Tests and Fiber Identification of Paper Materials Using Principal Component Analysis and Specular Reflection FTIR Data

Julie H. Wertz [1], Arthur A. McClelland [2,*], Debora D. Mayer [3] and Penley Knipe [1]

[1] Straus Center for Conservation and Technical Studies, Harvard Art Museums, Cambridge, MA 02138, USA; juliehwertz@gmail.com (J.H.W.); penley_knipe@harvard.edu (P.K.)
[2] Center for Nanoscale Systems, Harvard University, Cambridge, MA 02138, USA
[3] Weissman Preservation Center, Harvard University, Cambridge, MA 02138, USA; debora_mayer@harvard.edu
* Correspondence: amcclelland@cns.fas.harvard.edu

Abstract: Paper materials and works of art on paper such as drawings, watercolors, prints, books, and manuscripts represent a large portion of museum, archive, and library collections. However, paper materials are infrequently the subject of technical studies due to inherent limitations in their analysis such as the fragility of the paper substrate, a lack of suitable sampling opportunities, and the presence of mixed, but chemically similar cellulosic materials. The application of principal component analysis (PCA) modeling to specular reflection FTIR data has the potential to provide a non-invasive means of analysis for major and minor components in paper materials. Using known study collection objects, PCA models distinguishing paper sizing materials and fiber types based on specular reflection FTIR data were successfully demonstrated thus providing a plausible alternative method for the identification of paper materials in collection objects without the need for destructive testing or sampling of the object.

Keywords: specular reflection FTIR; principal component analysis PCA; paper fiber identification; paper sizing identification

Citation: Wertz, J.H.; McClelland, A.A.; Mayer, D.D.; Knipe, P. Modeling Chemical Tests and Fiber Identification of Paper Materials Using Principal Component Analysis and Specular Reflection FTIR Data. *Heritage* **2022**, *5*, 1960–1973. https://doi.org/10.3390/heritage5030102

Academic Editor: Massimo Lazzari

Received: 1 July 2022
Accepted: 26 July 2022
Published: 1 August 2022

1. Introduction

Paper materials and works of art on paper represent a large portion of museum, archive, and library collections, including drawings, watercolors, prints, books, and manuscripts. Despite their prevalence, paper materials are infrequently the subject of technical studies due to inherent limitations in their analysis, such as the fragility of the paper substrate, a lack of suitable sampling opportunities, and the presence of mixed, but chemically similar cellulosic materials.

There are many potential questions a technical study of paper and works on paper can answer. It can be useful to understand the composition (furnish) of a paper substrate in terms of the raw fiber material (cotton, wood, linen, etc.), chemical treatments (gelatin, starch, or alum sizing), fillers (such as clay), and lignin content (an organic compound present in certain fibers). This information may inform conservation treatments, stability predictions, and provenance research. To date, most of these questions have been answered through invasive or micro-destructive analysis [1], including micro sampling for fiber identification with polarized light microscopy or chemical spot testing, such as the Biuret test, which uses aqueous copper (II) sulfate in a basic aqueous solution to detect proteins, indicating the presence of gelatin. The nature of these tests limits which objects can be studied.

A non-destructive analysis is always preferential for conservation research of collection materials [2–4]. Fourier-transform infrared (FTIR) spectroscopy is a particularly suitable analytical chemistry technique to detect and identify the organic materials present in paper

and works of art on paper [5–11]. Since the resulting vibrational spectra are cumulative representations of all materials present, it can be difficult to discern minor peaks among the bulk cellulose signal, as well as subtle variations within the cellulose signal. These small differences, however, are significant for material identification and classification; however, the human eye alone cannot make reliable distinctions in the patterns of the spectra.

With FTIR, identifications are usually made by interpreting individual peaks or comparing the spectrum to reference spectral libraries of known compounds. Computers are used to carry out this pattern matching task; however, the correlation algorithms used in pattern matching heavily weight the identification matches to the largest peaks in the spectra. For paper materials, the FTIR spectra are usually dominated by large cellulosic peaks. Minor chemical components with small peak contributions to the overall spectra, such as the sizing material used in the paper, end up being ignored by the spectral library matching algorithms. This means that with straight spectral library matching, all papers often end up being identified as the same.

This manuscript explores the application of principal component analysis (PCA), a method of data processing and analysis to determine correlations between measurements. The data can be processed in a way that ignores the commonalities between the spectra and highlights the systematic differences. Specifically, here, for instance, the cellulosic spectral contribution can be ignored and the signal from the materials used for paper sizing can be highlighted and explored.

The sample set for this research was mainly the Paper and Mediums Study Collection from Legacy Press compiled by Cathleen Baker (http://www.thelegacypress.com/study-collection.html, accessed on 1 September 2021). It consists of American and European book and writing papers from the 18th to the 21st centuries. It has previously been subjected to fiber analysis and chemical tests for gelatin, starch, and lignin following standard methods in the field of paper conservation. The known sample set was supplemented with alum sized papers from Debora Mayer's personal collection and pure fiber pulp stock samples from Walter Rantanen.

The analysis reported in this manuscript will demonstrate that the PCA model of the specular reflection FTIR spectra can provide the same information as the chemical tests- if not even better, providing a plausible alternative method for identification that does not require destructive testing and can be applied to a wider range of objects.

1.1. Spectroscopy Data Collection Methods

The FTIR spectra for this manuscript were collected using a non-contact specular reflectance method which avoids any risk of damage to the sample, making the technique suitable for fragile objects, and yet it provides rich molecular information about the object, as previously demonstrated by McClelland et al. in the analysis of coatings on historic salted paper prints [12].

Raman spectroscopy is the other optical vibrational spectroscopy analytical chemistry technique often used in the analysis of museum and library collections. It is also completely non-contact and provides similar molecular information to FTIR. However, Raman spectroscopy can suffer spectral interference from fluorescence in the object under investigation and it is, in general, not suitable in the analysis of papers as cellulosic materials tend to exhibit high amounts of fluorescence. Additionally, Raman spectroscopy uses highly focused laser light and is therefore not ideal for possibly photosensitive materials.

Attenuated total reflection instruments (ATR-FTIR) have been used extensively in cultural heritage analytical research; however, the ATR-FTIR requires intimate contact between the ATR crystal and the material, causing small dents in paper supports. In the end, different ATR-FTIR approaches all involve either sampling the object or the potential risk of permanently distorting soft materials such as paper. ATR-FTIR does have the distinct advantage over specular reflection FTIR of decades of reference spectra already existing in spectral reference libraries. For instance, a comprehensive library of reference spectra of ATR-FTIR and Raman spectra for cultural heritage materials has been built through

the Infrared and Raman User Group (IRUG) [13,14]. The PCA techniques presented here could certainly be applied to ATR-FTIR or Raman spectral data sets. For this manuscript, the question was about how much information could be extracted specifically from the specular reflection FTIR spectra.

Although infrared peak positions are based on the energy of the molecular bonds and do not change significantly between data collection modes, the peak shape and intensity will be strongly affected based on the method used. As such, an ATR-FTIR reference spectrum will not be directly comparable to an external reflectance spectrum. Ideally, a separate spectral reference library would be available where the reference and sample materials were scanned using the same type of instrument. A more thorough discussion of FTIR analysis and applications in cultural heritage can be found in McClelland et al. [12].

1.2. Principal Component Analysis (PCA)

PCA can be applied to many kinds of data, including spectra, and enables the visualization of relationships that may not be obvious to the eye alone, eliminating bias imposed by human interpretation. In recent years, PCA has been applied in cultural heritage research to FTIR spectra of degrading plastics [15], Raman spectra of drawing media in drawings by the artist Odilon Redon [16], various properties of 19th and 20th century Chinese papers including pH and tensile strength [17], NMR spectra of 13th-15th century Italian paper [18], microspectrofluorimetry of lake pigments [19], and FTIR spectra of archaeological Aztec resins [20]. This list, while not exhaustive, demonstrates the value of PCA applied to different types of data sets in cultural heritage as non-invasive analysis and material characterization techniques continue to improve.

A PCA model is generated by organizing the data in a matrix, effectively a table, where each sample becomes a row and each column a measurement. For FTIR spectra, each row is an individual spectrum, and each column is the spectral response at a particular wavenumber. Linear algebra algorithms are then applied to the data matrix to find the eigenvalues and eigenvectors. The eigenvectors are referred to as the principal components (PCs). The eigenvalues are referred to as the loadings. In practice, a computer takes care of all the calculations with relatively little user input needed other than correctly organizing the data into a table of samples versus measurements.

The PCA model will depend on the samples that were included in the matrix. In general, the more data available when building the model, the more robust the model will be. It is usually best to start with a collection of known and well-characterized samples and then try to apply the model built with the known samples to unknown objects.

The PCs can be thought of as a new basis set that is a better description for the original data set. A familiar example of using a different basis set to better describe the real-world is the use of longitude and latitude to describe positions on the surface of the earth, rather than Cartesian (x, y, z) coordinates from the center of the earth. The PCs are a new coordinate system that is made up of a linear combination of the different weightings of each of the measurements in the original data matrix that best describe the variance in the data. The different weightings are called the loadings. The PCs are numbered based on the amount of variance they represent, so PC 1 will represent more variance than PC 2 and so on. Variance is the statistical measure of how much spread the data set has along a specific PC. The total number of PCs will vary based on the complexity of the data set, and all PCs may not be significant. It is certainly possible to overfit the data and model the noise in the data set.

The PCA models also generate a Q residual value, which represents how well that sample fits to the model, and a Hotelling T-squared value that shows how far from the center of the model a particular sample lies. Looking for samples with high Q residuals and T-squared values is usually an easy way to spot outliers in the data set.

Prior to building a model, it is usually necessary to preprocess the sample data. Preprocessing treatment of raw data can have a significant effect on the strength of the model, since PCA models work best with linear relationships. Preprocessing the data describes any process applied to the data prior to modeling, including normalization, variable centering,

mathematical operations, and scaling. This treatment removes extraneous variation, for example, signal artifacts from the instrument, and linearizes relationships between variables. Spectral data can usually be improved with baseline corrections and mean centering the data. In FTIR, there is often a baseline offset from light that is scattered out of the optical collection path. This scattering usually has a weak wavelength dependence leading to a sloping baseline on the data. On paper media, the roughness of the paper contributes to the baseline offset as light is scattered out of the optical collection path of the microscope. Since the baseline shape does not contain chemical information about the object, it can be fitted to a polynomial and subtracted from the spectra. Mean centering is a data processing technique where the average (mean) of the entire data set is subtracted from each individual spectrum in the data set. This has the effect of suppressing what is the same between all the spectra and highlighting what is different.

2. Materials and Methods

2.1. Specular Reflectance FTIR Spectroscopy

This research employed a Bruker LUMOS FTIR microscope with a liquid nitrogen cooled mercury cadmium telluride (MCT) detector from 4000 to 600 cm^{-1} in specular reflectance mode at the Harvard Center for Nanoscale Systems (CNS). The technique is fully non-contact, and the system allows for multiple points in the analysis area to be selected and scanned. The microscope and camera allowed the close examination of the analysis area and documentation of the visual appearance. Thirty points were collected from each sample (fifteen each from two areas of the paper) to ensure statistical representation within the data model. Each point of data was collected from the default area size of 125 × 126 microns and was the average of 16 scans at 4 cm^{-1} spectral resolution. The background was taken against the built-in gold reference mirror on the instrument sample stage. Data collection was carried out using the accompanying Bruker OPUS software. Spectra were baseline corrected in Opus using the rubber band method.

2.2. Sample Material and Classification

Areas for spectral analysis were selected based on an absence of stains, foxing, print, ruling, or other visual irregularities. Spectra were collected from known papers from a range of sources. The Legacy Press Paper and Mediums Study Collection was the primary source of reference material. The collection, released in 2016 and compiled by Cathleen Baker, contains 42 examples of book paper and 21 examples of writing paper made from the 18th–21st centuries in Europe and North America. The papers were subject to chemical tests by Baker for protein (Biuret test), starch (iodine-potassium iodide), and lignin (phloroglucinol). The fiber content was assessed by Integrated Paper Services, Inc. The results of these investigations were included with the collection which enabled data correlation with the FTIR spectra. The decision to use the Legacy Press Collection of papers was made because it is a curated collection of papers that many art conservation labs have and can be referenced and studied by other researchers. Using known collections allows a comparison of test results and adds to the shared understanding of these papers, which are studied in art conservation training programs. This project took advantage of the testing already conducted for sizing, lignin, and fiber composition which is recorded with the collection. No additional testing was performed on these papers, such as filler content.

The alum sized papers were artists papers from the 1970s with 100% rag fiber content (cotton) from Debora Mayer's collection, and included Lenox 100 (Rising Co., Housatonic, MA, USA), Fabriano 5 watercolor paper (Fabriano, Italy) and BFK Rives printmaking paper (Rives, France).

The first step in the project was to take fiber samples of high cellulose content (fibers with no or extremely low lignin content) and determine if they could be readily separated using this model. The fiber groups first tested were pure pulp stock reference materials including cotton linter fibers, cotton textile fibers, rayon fibers, softwood bleached (BL)

kraft fibers, softwood bleached (BL) sulfite fibers, and softwood high alpha pulp fibers. The samples were provided by Walter Rantanen, SGS-IPS Testing.

The samples were assigned classes, the term used for a group of known samples that are the same chemically. Initial modeling was carried out with the data from the 21 writing paper samples because there was more variability in the test results than in the results for the book paper samples. A positive result for gelatin (G), starch (S), or lignin (L) was used for the model's classes. For example, the SL class all tested positive for starch and lignin but not gelatin. The N class had a negative result for all three chemical tests.

The writing paper samples, results, and classes are given in Table 1. Manufacture dates and sample letters are from the notes accompanying the collection. The number of samples in each class is: N (1), L (1), S (3), SL (1), GS (7), GL (6), GSL (2). There were unfortunately no samples that tested positive for gelatin only. Figure 1 shows an average spectrum for each class prior to preprocessing. The most obvious difference in spectra is the proteinaceous amide I and II peaks from 1550 to 1650 cm^{-1} for the GL, GS, and GSL classes. In the -OH stretching region from 3000–3400 cm^{-1}, the slope of the GS spectrum is shallower than that of the other classes. The remaining variance across the spectra is less distinct and difficult to characterize by peak position and shape alone, hence the application of PCA. With data science classification techniques, such as PCA, it is not necessary to try to parse and assign each peak in a spectrum. The shape of the spectra can be viewed as a wholistic signature of a paper's specific chemical composition. Spectral processing techniques such as the Kubelka-Munk or Kramers-Kronig transforms are not necessary if the PCA model is built with data that is collected in the same method as the data from the unknown objects.

Table 1. Chemical test results and model classification for Paper and Mediums Study Collection writing paper.

		Chemical Test Results			
Writing Paper	**Publication Date**	**Gelatin +/− (G)**	**Starch +/− (S)**	**Lignin +/− (L)**	**Model Class ***
A	1790–1816	+	−	+	GL
B	1980s	+	+	−	GS
C	ca. 1819	−	−	+	L
D	1950s?	+	+	−	GS
E	1952?	−	+	−	S
F	1995	−	−	−	N
G	1995	−	+	−	S
H	2004	−	+	−	S
I	2001	−	+	−	S
J	ca. 1829	+	−	+	GL
K	1760?	+	−	+	GL
L	1900?	+	+	−	GS
M	early 19th c.?	+	−	+	GL
N	unknown; 18th c.?	+	−	+	GL
O	unknown; 19th c.?	+	−	+	GL
P	1900?	+	+	−	GS
Q	ca. 1835	+	+	−	GS
R	early 1900s?	+	+	−	GS
S	1830s?	+	+	+	GSL
T	ca. 1877	+	+	−	GS
U	mid-20th c.?	+	+	+	GSL

* G (gelatin), S (starch), and L (lignin) indicate classification based on a positive test result for that material; N indicates negative for all tests.

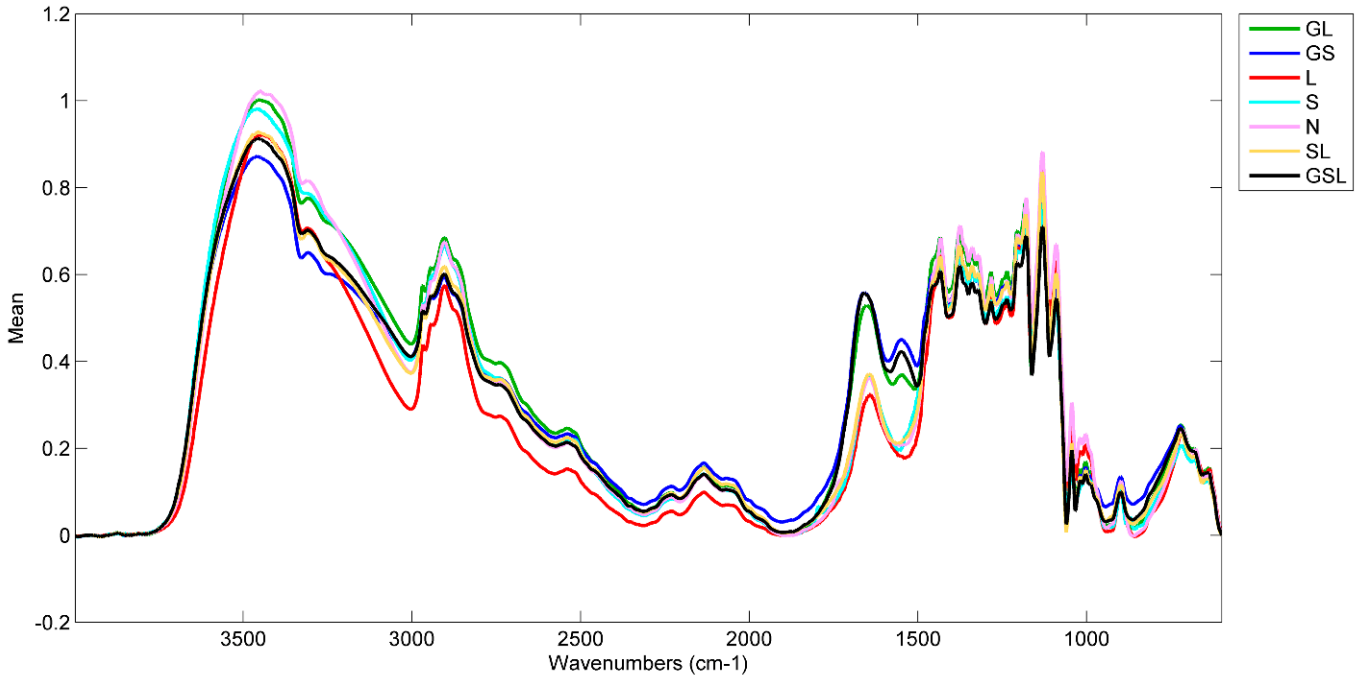

Figure 1. Average spectra of writing paper data by class before preprocessing. The similarity between the different spectra makes it hard for the human eye to distinguish the differences. Correlation matching functions used in spectral library software will also struggle to find the differences. Classification techniques, such as PCA, make it not necessary to try to parse and assign each individual peak in the spectrum. The shape of the spectrum can be viewed wholistically as a signature for a paper's particular chemical composition.

Of the 10 positive listings for lignin, only one was a moderate test result (Paper C). The other 9 were noted as slight or very slight test results.

2.3. Modeling

Modeling was carried out using the Solo software from Eigenvector Research Incorporated. The spectra were imported into the software and class assignments entered, first based solely on the results of the sizing in the paper. A second model was made including the known information about the sizing and the fiber types. A range of preprocessing conditions were tested to optimize the predictive capability of the model. The order of preprocessing steps does affect the outcome of the resultant model.

For spectral data, sources of variation that can obscure information of interest include measurement noise, baseline variation, and environmental CO_2 to name a few. These signals are often referred to as clutter and they make the modeling less robust. A strong model will show clustering within an assigned class, while individual classes are reasonably separated within the model space. Proper preprocessing can help minimize the clutter and tighten the clustering of the classes by just keeping the variance of interest in the data set and ignoring the variance that is not of interest.

For these spectra, the baselines were first corrected in OPUS using the rubber band method. Next, in Solo, a multiplicative signal correction (MSC) was applied as a weighted normalization treatment to remove magnitude variability. The MSC algorithm also performs a baseline removal, but better results were obtained using the OPUS baseline removal tool prior to importing the data into Solo. The weighted normalization treatment determines the weighting scale factor by regressing a measured spectrum against a reference spectrum calculated from the mean of the data, then correcting the spectrum using the slope and intercept of the fit. MSC can be applied to the mean or the median of the data. For this project the MSC (mean) algorithm was used.

Next, generalized least squares weighting (GLSW) was applied to the data. GLSW is a decluttering treatment where the data are weighted by the inverse square root of the clutter covariance. The weighting factor can be adjusted, with lower thresholds going further into the features from minor components. A GLSW threshold of 0.05 was found to give the tightest clustering of the classes with this data set. The GLSW algorithm shrinks the clutter dimensions without fully removing them, so any actual variance caught in the clutter calculation will still have some representation in the model. It does not account for magnitude difference, so it should be applied after a normalization process such as MSC.

"Clutter" is variance in the data that is not relevant to the question at hand. For example, in this case, the fiber type information would be "clutter", since the question was about whether there was gelatin, lignin, or starch in the object. The same FTIR data could be arranged into classes based on the fiber type and then the gelatin, lignin, and starch signals would be the "clutter".

Finally, the data were mean centered. Mean centering is a treatment where the mean of the data set is subtracted from each column of the matrix. This allows the model to capture variance around the mean of the data, emphasizing differences while eliminating repetitive information from the model. A similar treatment, class centering, where the mean of each class is subtracted from that particular class, can also be applied. The suitability of class centering depends on whether there is additional variance within a class, effectively a "sub-class". This is the case with the paper samples since there are spectral contributions from the fiber content that are not accounted for by the classifications for the chemical tests. As such, mean centering is the best choice for this data set and generates a better model than class centering [21].

In summary, the full preprocessing treatment for the data was rubber band baseline correction in OPUS and in Solo MSC (mean), GLSW (clutter source x-block classes, threshold 0.05), and mean centering preprocessing.

3. Results and Discussion

3.1. Lignin Determination

Of the writing papers in the Legacy Press collection, aside from paper C, all the positive occurrences of lignin were listed as slight or very slight. In most instances this would be from the unbleached portion of flax/hemp fibers used, which sometimes show up as dark flecks or fiber bundles in paper. The PCA models had trouble distinguishing the reported slight variation in lignin, and this may be due to the heterogeneity of the fiber distribution in the papers and the sampling area (125 $\times$ 126 microns) used for the FTIR capture. While 125 microns is a common analysis area for FTIR microscopy, it is a small area for collecting a representative sample of heterogenous fiber composition in paper. It is likely that the sample area for FTIR capture was too small to detect the slight variation in lignin content. Further investigation of the lignin content of the papers is left for future work with a different reference set of papers and different sample area parameters.

3.2. Paper Sizing Material Determination

A PCA model was built to explore if the paper sizing material could be determined from the specular reflection FTIR data of a variety of historic writing papers. Paper sizing are materials added to the paper to enhance certain desired properties of the paper. Some common sizing materials are gelatin, starch, and alum. The FTIR signal from the sizing is usually very small compared to the signal from the cellulose fibers in the paper. The gelatin and starch sized papers were from the Legacy Press Collection. The alum sized papers were from Debora Mayer's personal collection. Figure 2 shows the PCA model of the reference papers using the information about the sizing material. By and large, the different groups separate out nicely. The slight overlap between the gelatin size papers and the gelatin and starch sized papers is probably partly due to unaccounted for variance from the different fiber types in the papers and differences in the amount of the gelatin content between the papers.

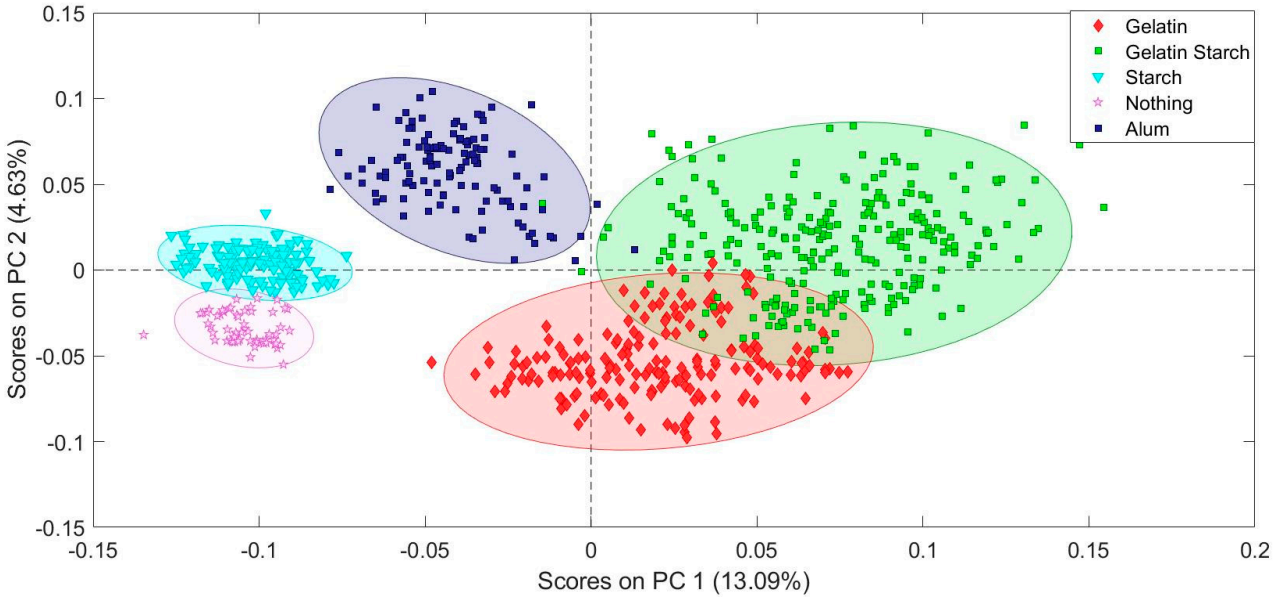

Figure 2. PCA model built on specular reflection FTIR data and known paper sizing material information for classification.

To test the model against a museum collection object, the paper that the photographs are mounted on in Alexander Gardner's *Photographic Sketch Book of War* (object number 2013.6.1), published in 1866 were used (Figure 3). The selection of this object was mainly due to the fact that specular reflection FTIR data had already been collected from it by the authors for another research project. Measurements from the blank paper area on the right edge of pages 68, 78, 96, and 97 are plotted in Figure 4. It unfortunately does not fall very neatly into any of the reference categories. This shows some of the limitations of the model based solely on the sizing material information.

3.3. Paper Fiber Determination

The papers in the Legacy Press collection do have different fiber types and fiber mixtures and this is a variable in the signal that was not taken into account by the model built on just the chemical tests. This led to the question of whether the specular reflection FTIR data had enough information to distinguish the different types of cellulosic material.

The reference materials of pure pulp stock were provided by Walter Rantanen of SGS-IPS testing, and include cotton linter fibers, cotton textile fibers, rayon fibers, softwood bleached (BL) kraft fibers, softwood bleached (BL) sulfite fibers, and softwood high alpha pulp fibers. Somewhat surprisingly, the different cellulose fiber types neatly split out into different groups (Figure 5). Even extremely similar fibers such as cotton linters (short fibers of the cotton boll) and cotton textile fibers (lint, the long cotton fibers commonly used in textiles) separated in PCA space.

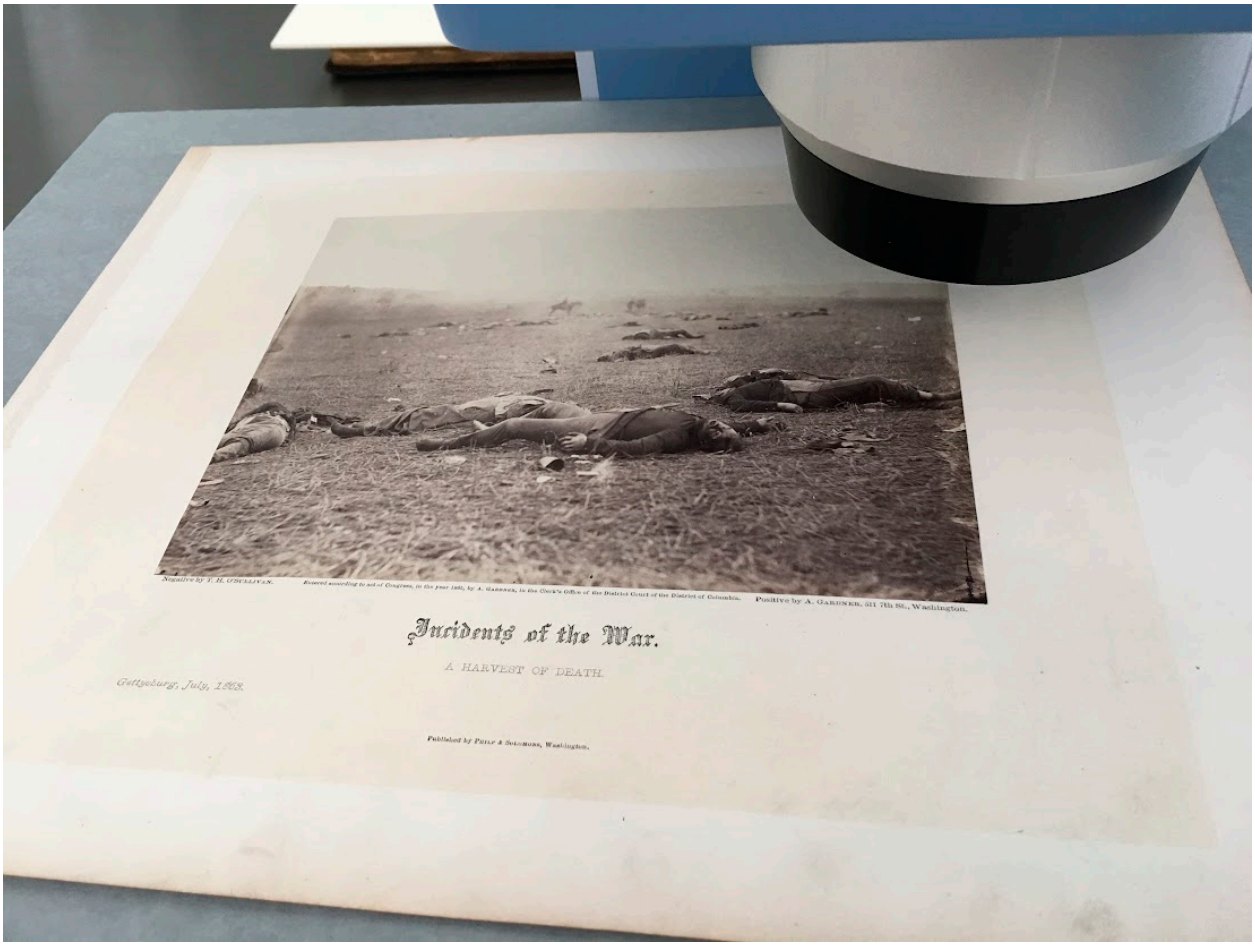

Figure 3. A loose print from Alexander Gardner's Photographic Sketch Book of War undergoing the specular reflection FTIR data collection for analysis of the paper. Alexander Gardner, *A Harvest of Death*, Albumen Silver photograph mounted overall to cardstock, Harvard Art Museums/Fogg Museum, Richard and Ronay Menschel Fund for the Acquisition of Photographs, 2008.22.

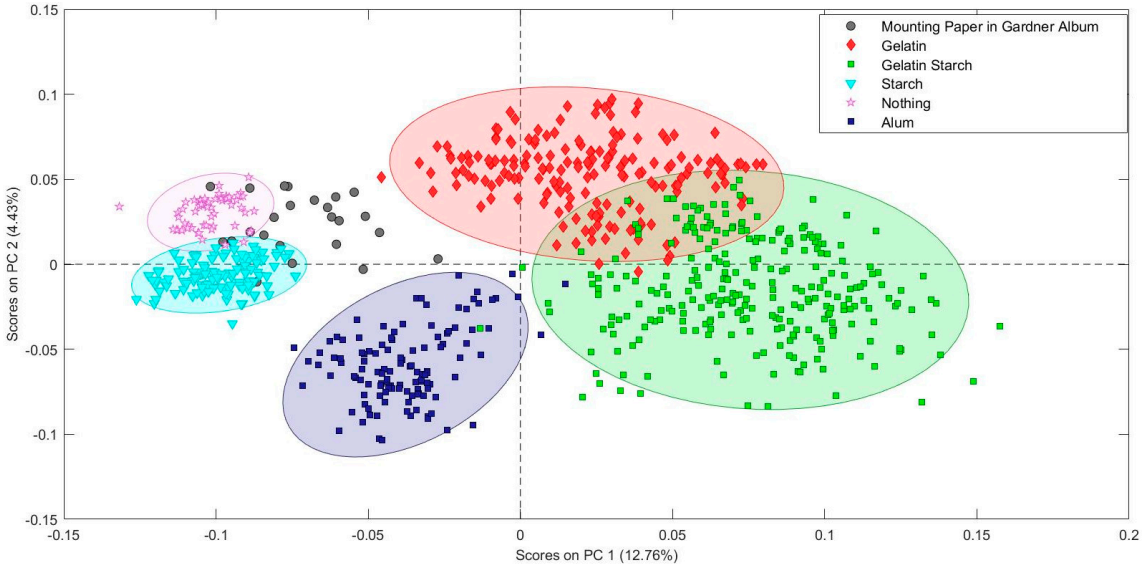

Figure 4. Plotting photograph mounting paper in Alexander Gardner's *Photographic Sketch Book of War* (Harvard Art Museum object number 2013.6.1), dark grey circles, plotted against the PCA model of paper sizing.

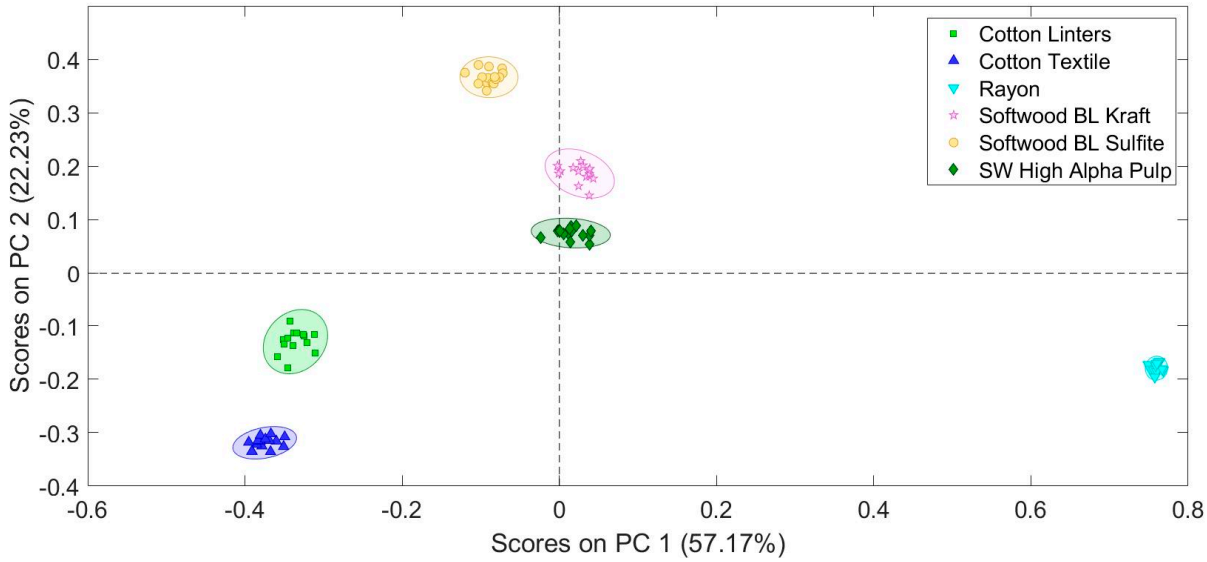

Figure 5. PCA of pure fibers samples provided by Walter Rantanen. The data was collected in specular reflection FTIR.

Encouraged by this result showing that different types of cellulose fibers can be clearly distinguished with specular reflection FTIR and principal component analysis, the Legacy Press writing paper data and alum sized papers were reclassified using the chemical tests and the paper fiber information (Figure 6). The potential lignin content was ignored again for reasons previously discussed. The starch wood paper (writing paper I) is far from the others in PCA space. Leaving pure wood fiber paper out of the model allows the data of the more similar papers to be more clearly plotted. This observation argues for developing a hierarchical model in the future. The hierarchical model can be thought of as a flow chart with different decision points. The first model would be used to determine the main fiber type family (wood, cotton, etc.) and the subsequent models would be used to determine the sizing materials and mixture of the fibers.

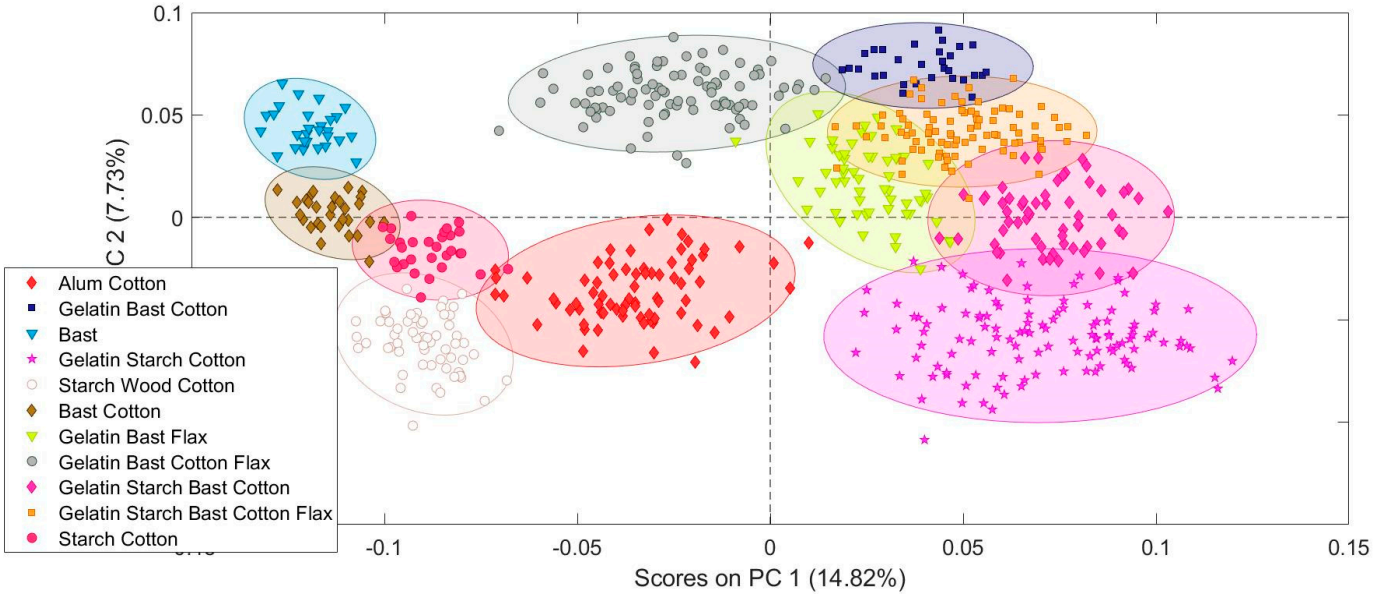

Figure 6. PCA model based on the sizing and fiber type information demonstrating the ability to distinguish mixed fiber papers.

Some of the classes are adjacent or slightly overlapping in PCA space, but there are surprisingly clear groups for the different mixtures that range from one to five components. In very broad strokes, PC1 picks up mostly on the variance due to gelatin content and PC2 picks up on variance due to bast content. While the confidence ellipses of the different fiber mixtures are slightly overlapping, multiple measurements from an object can give more confidence in the assignment of an unknown sample to a specific class. Spectra from the Gardner album do not align neatly with any one reference group, but appear to contain elements of gelatin, bast, cotton, starch, and flax (Figure 7). The fact that an almost unlimited number of specular reflection FTIR measurements can be taken from multiple positions on a collections object is a clear advantage. Other data science techniques such as a support vector machine discriminant analysis (SVMDA) may be a better choice for data that has such close boundaries, but exploration of that will be left for future work.

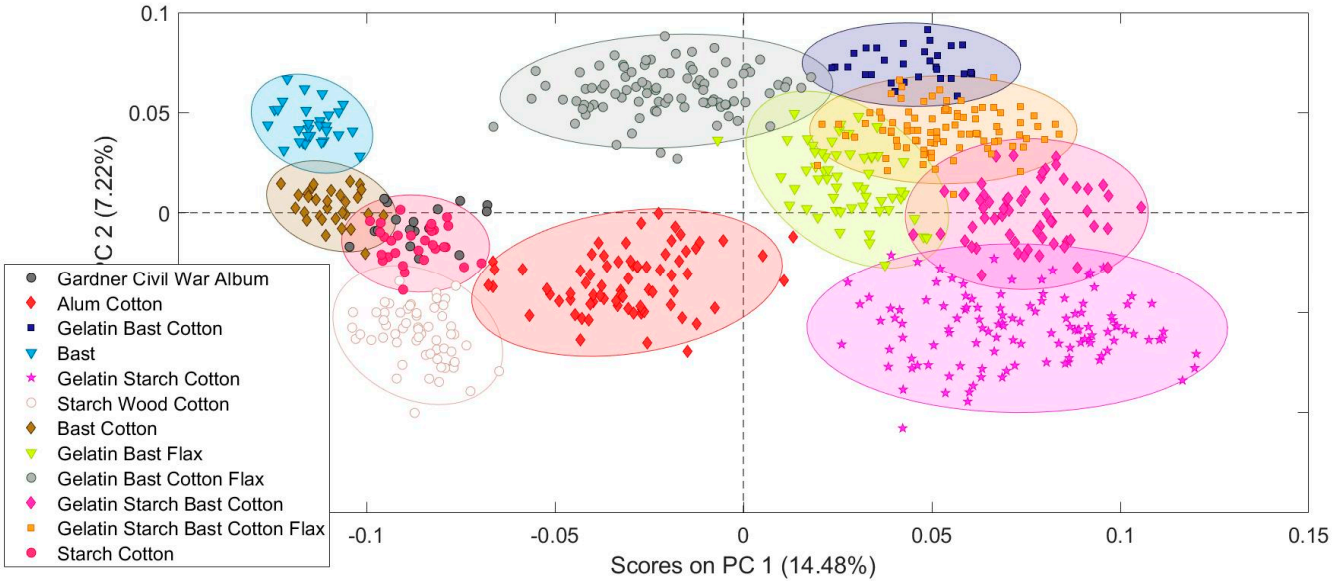

Figure 7. Plotting photograph mounting paper in Alexander Gardner's *Photographic Sketch Book of War* (Harvard Art Museum object number 2013.6.1), dark grey circles, against the model based on sizing material and fiber type. The mounting pages 68, 78, 96, and 97 now clearly fall in the cotton paper sized with starch classification area.

4. Conclusions

The application of PCA modeling to specular reflection FTIR spectral data has the potential to provide a non-invasive means of analysis for major and minor components in paper materials. Relative to the bulk cellulose signal, the presence of components such as gelatin, starch, and lignin are fairly minimal and may be difficult to detect from examining individual peaks alone. Traditional methods for detecting these materials involve wet chemical tests (Biuret, iodine, phloroglucinol) that are ultimately destructive and not applicable to museum and collection objects. Taking these two primary concerns into consideration, a PCA model of non-contact specular reflectance FTIR spectra was tested as an alternative method for the wet chemical tests.

The models demonstrated here, generated from the Legacy Press Paper and Mediums Study Collection, demonstrate the effectiveness of the method. Preprocessing is a significant part of making a robust PCA model, and the conditions are dependent on the type of data being modeled. In this instance, MSC, GLSW, and mean centering were found to be the most useful for minimizing similarity (bulk cellulose) and emphasizing variance within the specular reflectance FTIR spectra. Classifications for the calibration data were made based on the notes accompanying the collection, which came from chemical tests carried out by collection compiler Cathleen Baker and the fiber type analysis from Integrated Paper Services, Inc.

Pure fiber pulp stock samples were tested. The ability to distinguish different sources and processing of cellulosic material was demonstrated.

Including the fiber types and chemical test data produced a more precise classifications model for the paper study collection. The photograph mounting paper used by Alexander Gardner in Gardner's Photographic Sketch Book of War was compared to two different PCA models. The first model only included information about the paper sizing. The second model included sizing and fiber type information. The second model provides a much clearer identification of the paper Gardner used. The data from multiple pages lands neatly in the cotton/bast paper sized with starch area of the second PCA model, allowing a positive identification with reasonable confidence. The fiber composition result with the PCA model matches the results obtained by traditional fiber analysis performed by Debora Mayer using broken fragments from the edge of a page. The cotton fibers were identified based on the presence of extinction bands visible in cross-polarized light microscopy. Since the specular reflection FTIR measurement is completely non-sampling and non-contact, several points on multiple pages could be investigated.

Based on this demonstration, PCA modeling of non-invasive specular reflectance FTIR is a viable method for detecting the presence of minor chemical components within the bulk cellulose matrix of the paper without risking harm to the object. Differences in fiber types can also be distinguished. Since fiber identification also traditionally requires sampling and examination by optical microscopy, this is another place where PCA modeling of FTIR spectra has the potential to provide information for objects that would otherwise not be suitable for analysis involving sample removal. In the analysis of fiber composition of papers, use of PCA modeling with specular reflectance may better approximate industry standards on sampling paper to achieve representative results, all nondestructively, in contrast to traditional methods of removing samples for examination using optical microscopy. This is a practice changer for art conservators and other museum and library professionals in the analysis of collections.

As data science techniques become more widely accessible to the non-specialist, more information about objects can be determined. While the commercial Solo software by Eigenvector Research, Inc. was used here, similar data analysis could be carried out using free software packages such as python or R.

More complex models should be developed for general paper analysis. Since wood fibers land so far from cotton, bast, and flax fibers in PCA space, a hierarchal model should be used to first distinguish general fiber type, and then a more refined model to distinguish specific fiber mixture and sizing materials. The ability to build robust models depends on having relevant well-characterized reference materials to collect data from.

Future directions for this research include expanding the papers analyzed to papers of known fiber mixtures, including percentages along with processing chemistry and modern sizing agents, and broadening the scope of papers to study to include non-western papers such as Indian, Islamic World and East Asian papers to understand how the techniques outlined might be applied to papers from around the world. It is also hoped that once paper supports are better understood, the techniques might be used to study media that has been difficult to analyze because there is so little material present. This might include, for example, drawing inks on old master drawings. The use of principal component analysis and specular reflection FTIR may open new avenues for the study of artists' materials, allowing a deeper understanding of materials that were traditionally very difficult to analyze.

Author Contributions: Conceptualization, all authors.; methodology, J.H.W. and A.A.M.; software, J.H.W. and A.A.M.; formal analysis, J.H.W. and A.A.M.; resources, D.D.M. and P.K.; data curation, J.H.W. and A.A.M.; writing—original draft preparation, J.H.W.; writing—review and editing, A.A.M., D.D.M., P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was performed in part at the Harvard University Center for Nanoscale Systems (CNS); a member of the National Nanotechnology Coordinated Infrastructure Network (NNCI), which is supported by the National Science Foundation under NSF award no. ECCS-2025158. J.W. was supported by the Harvard Art Museums Beal Family Postgraduate Fellowship in Conservation Science. The APC was funded by MDPI.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would also like to acknowledge and thank Leonie Müller for fruitful discussions during this project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strlič, M.; Liu, Y.; Lichtblau, D.A.; de Bruin, G.; Knight, B.; Winther, T.; Kralj Cigić, I.; Brereton, R.G. Development and Mining of a Database of Historic European Paper Properties. *Cellulose* **2020**, *27*, 8287–8299. [CrossRef]
2. Brown, N.; Coppola, F.; Modelli, A.; Amicucci, F.; Lichtblau, D.; Strlič, M. Non-Destructive Collection Survey of the Historical Classense Library. Part I: Paper Characterisation. *Herit. Sci.* **2020**, *8*, 88. [CrossRef]
3. Liu, Y.; Fearn, T.; Strlič, M. Quantitative NIR Spectroscopy for Determination of Degree of Polymerisation of Historical Paper. *Chemom. Intell. Lab. Syst.* **2021**, *214*, 104337. [CrossRef]
4. Strlic, M.; Kolar, J.; Lichtblau, D.; Trafela, T.; Anders, M.; de Bruin, G.; Knight, B.; Martin, G.; Palm, J.; Selmani, N.; et al. Le Projet SurveNIR et La Caractérisation du Papier Ancien TT-The SurveNIR Project and the Characterization of Old Paper. *Support Tracé* **2009**. Available online: https://arsag.fr/resources/public/sts/summaries/support-trace-9.pdf (accessed on 15 June 2022).
5. Xia, J.; Zhang, J.; Zhao, Y.; Huang, Y.; Xiong, Y.; Min, S. Fourier Transform Infrared Spectroscopy and Chemometrics for the Discrimination of Paper Relic Types. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *219*, 8–14. [CrossRef] [PubMed]
6. la Russa, M.F.; Ruffolo, S.A.; Barone, G.; Crisci, G.M.; Mazzoleni, P.; Pezzino, A. The Use of FTIR and Micro-FTIR Spectroscopy: An Example of Application to Cultural Heritage. *Int. J. Spectrosc.* **2009**, *2009*, 893528. [CrossRef]
7. Yan, Y.; Wen, C.; Jin, M.; Duan, L.; Zhang, R.; Luo, C.; Xiao, J.; Ye, Z.; Gao, B.; Liu, P.; et al. FTIR Spectroscopy in Cultural Heritage Studies: Non-Destructive Analysis of Chinese Handmade Papers. *Chem. Res. Chin. Univ.* **2019**, *35*, 586–591. [CrossRef]
8. Trafela, T.; Strlič, M.; Kolar, J.; Lichtblau, D.A.; Anders, M.; Mencigar, D.P.; Pihlar, B. Nondestructive Analysis and Dating of Historical Paper Based on IR Spectroscopy and Chemometric Data Evaluation. *Anal. Chem.* **2007**, *79*, 6319–6323. [CrossRef] [PubMed]
9. Stanley, T. The Examination and Analysis of Dunhuang and Turfan Manuscript Materials at Princeton University Library's East Asian Library. *J. Am. Inst. Conserv.* **2017**, *56*, 194–210. [CrossRef]
10. Walker, J.; Hodgkins, R.; Berrie, B. On the Surface: Reflectance FTIR Spectroscopy in Cultural Heritage Research. *Microsc. Microanal.* **2021**, *27*, 2800–2804. [CrossRef]
11. Crocombe, R.A.; Leary, P.E.; Kammrath, B.W. (Eds.) *Portable Spectroscopy and Spectrometry*; Wiley: Hoboken, NJ, USA, 2021.
12. McClelland, A.; Bulat, E.; Bernier, B.; Murphy, E.L. Specular Reflection FTIR: A Non-Contact Method for Analyzing Coatings on Photographs and Other Cultural Materials. *J. Am. Inst. Conserv.* **2019**, *59*, 123–136. [CrossRef]
13. Lomax, S.Q.; Price, B.A.; Lins, A.; Davis, C.; Pretzel, B.; Picollo, M.; Richards, G.; Rice, S. The IRUG Raman Spectral Web Database: Objectives, Progress and Plans. *E-Preserv. Sci.* **2013**, *10*, 38–41.
14. Price, B.; Pretzel, B.; Carlson, J.; Ehrman, K.; Lins, P.A. Web-Based Exchange of Infrared and Raman Spectra: A New IRUG Initiative. In Proceedings of the Art 2002: 7th International Conference on Non-destructive Testing and Microanalysis for the Diagnostics and Conservation of the Cultural and Environmental Heritage, Antwerp, Belgium, 2–6 June 2002; Congress Centre Elzenveld: Antwerp, Belgium, 2002.
15. Mitchell, G.; France, F.; Nordon, A.; Tang, P.L.; Gibson, L.T. Assessment of Historical Polymers Using Attenuated Total Reflectance-Fourier Transform Infra-Red Spectroscopy with Principal Component Analysis. *Herit. Sci.* **2013**, *1*, 28. [CrossRef]
16. Daly, N.S.; Sullivan, M.; Lee, L.; Delaney, J.K.; Trentelman, K. Odilon Redon's Noir Drawings: Characterization of Materials and Methods Using Noninvasive Imaging and Spectroscopies. *Herit. Sci.* **2019**, *7*, 43. [CrossRef]
17. Brown, N.; Lichtblau, D.; Fearn, T.; Strlič, M. Characterisation of 19th and 20th Century Chinese Paper. *Herit. Sci.* **2017**, *5*, 47. [CrossRef]
18. Proietti, N.; Roselli, G.; Capitani, D.; Pettinari, C.; Pucciarelli, S.; Basileo, S.; Scognamiglio, F. Characterization of Handmade Papers (13th–15th Century) from Camerino and Fabriano (Marche, Italy). *J. Cult. Herit.* **2020**, *42*, 8–18. [CrossRef]
19. Nabais, P.; Melo, M.J.; Lopes, J.A.; Vitorino, T.; Neves, A.; Castro, R. Microspectrofluorimetry and Chemometrics for the Identification of Medieval Lake Pigments. *Herit. Sci.* **2018**, *6*, 13. [CrossRef]

20. Piña-Torres, C.; Lucero-Gómez, P.; Nieto, S.; Vázquez, A.; Bucio, L.; Belio, I.; Vega, R.; Mathe, C.; Vieillescazes, C. An Analytical Strategy Based on Fourier Transform Infrared Spectroscopy, Principal Component Analysis and Linear Discriminant Analysis to Suggest the Botanical Origin of Resins from Bursera. Application to Archaeological Aztec Samples. *J. Cult. Herit.* **2018**, *33*, 48–59. [CrossRef]
21. *Model Building: Preprocessing Methods*; Eigenvector Research, Inc.: Washington, DC, USA. Available online: https://wiki.eigenvector.com/ (accessed on 15 June 2022).