Chemometric approaches for document dating: Handling paper variability

Carolina S.Silva[a], Maria Fernanda Pimentel[b], José Manuel Amigo[b,c], Carmen García Ruiz[d], Fernando Ortega Ojeda[d].

[a]Department of Fundamental Chemistry, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade Universitária, Recife, Brazil.

[b]Department of Chemical Engineering, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade Universitária, Recife, Brazil.

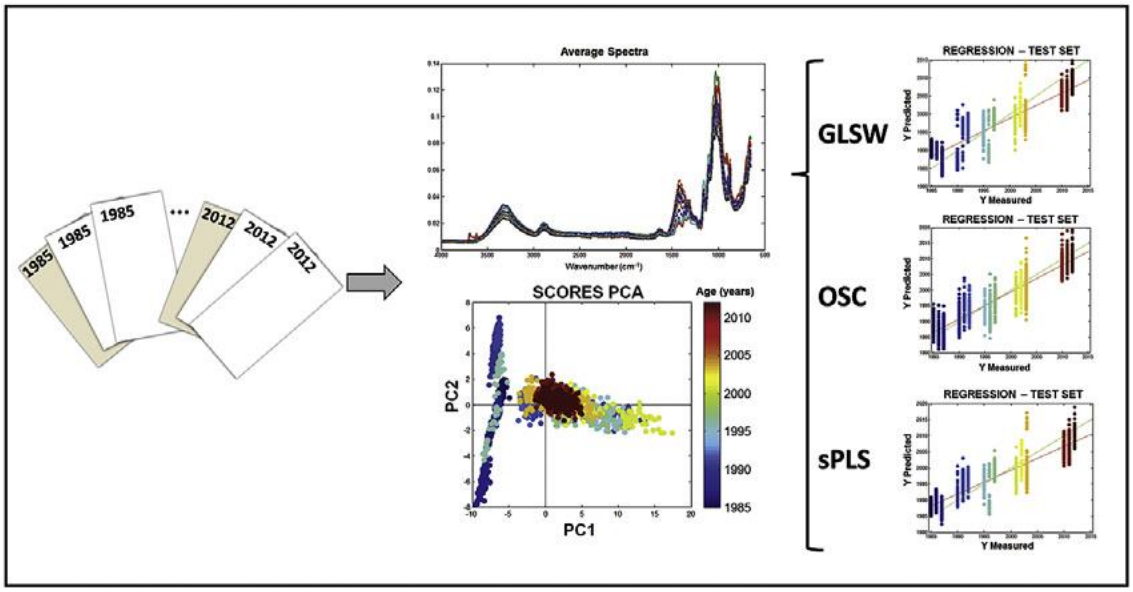
[c]Department of Food Science, University of Copenhagen, Rolighedsvej 30, Frederiksberg C, Copenhagen, Denmark.

[d]Department of Analytical Chemistry, Physical Chemistry and Chemical Engineering, University Institute of Research in Police Sciences (IUICP), University of Alcalá, Ctra. Madrid-Barcelona Km. 33.6, 28871, Alcalá de Henares, Madrid, Spain.

Abstract: A non-destructive methodology based on Fourier Transformed Infrared Spectroscopy (FTIR) is proposed in this research to estimate the age of documents of different ages. Due the variability in the samples caused by their different chemical compositions, chemometric approaches were proposed to build one unique regression model able to determine the age of the paper regardless of its composition. PLS models were built employing Generalized Least Squares Weighting (GLSW) and Orthogonal Least Squares (OLS) filters to reduce the variability of samples from the same year. Afterwards, sparse PLS, which is an extension of the PLS model including a variable selection step, was applied to compare its performance with the preprocessing filters. All techniques proposed were compared to the initial PLS models, showing the potential of the chemometric approaches applied to FTIR data to estimate the age of unknown documents.

Keywords: Documentoscopy; Chemometrics; Paper Dating; Forensic.

Grafical abstract:



1. Introduction

Document dating is still a major challenge in forensic document examination field [1]. Not only the variety of inks and papers, but also the mechanisms of degradation are some of the issues that make the study of the aging process a very complex topic. Although a number of research groups have studied the ink aging process, document dating focused on the paper aging process is still open for the development of new methodologies [2].

Paper samples are very complex mixtures. Although, the major compound in paper is cellulose, inorganic fillers are added during the paper manufacturing to provide proper characteristics such as whiteness, brightness and texture. Among the common inorganic compounds found in paper composition, calcium carbonate (CaCO3) and kaolinite (Si2Al2O5(OH)4) are the most common. Cellulose is a linear polymer mostly linked by β-1,4-glycosidic bonds. Due to its ability to aggregate and form highly structural entities its Degree of Polymerization (DP) may be defined. This is a parameter that measures uniformity, depending both on the vegetable fiber from which it originated and also changes over the time [3].

Overall, paper behavior over time can be very difficult to predict. During the paper aging process, several structural changes occur, including degradation of carbohydrates, cellulose, degradation by biological agents and oxidations. Evaluating those processes, including the variability in paper composition and storage conditions, certainly poses a challenge to the field of document analysis [3]. Cellulose being one of the major compounds in paper manufacturing means that there is a good deal of research in the literature employing different methods to evaluate cellulose degradation over time. Those studies are extremely informative and important for document date estimation.

Schedl and coworkers proposed a methodology employing mass spectrometry to quantify 2,5-dihydroxyacetophenone (DHAP) chromophore in cellulose samples and documents both artificially and naturally aged [4]. The authors used factorial design to evaluate how different conditions influenced the paper composition, using the concentration of DHAP as criteria to monitor paper aging under different temperatures, humidity, and iron ions presence. Preliminary works were developed in order to evaluate the kinetics of the reactions occurring in paper over time [5,6]. Those studies established a relationship between artificial and natural aging processes. They reported different behavior in the degree of cellulose polymerization when factors such as temperature and acidity were modified. Other groups employed different analytical techniques in order to monitor cellulose DP as well as other inorganic compounds present in paper [[7], [8], [9]].

Due to the added value of using non-destructive analysis to maintain document integrity, vibrational spectroscopy, such as infrared, has become seen as an appropriate analytical approach [10]. Ali and coworkers [11] evaluated the potential of Near and Middle Infrared (NIR and MIR, respectively) spectroscopy for dating artificially aged papers. Nine different papers were analyzed, in which spectral comparison and ratio between characteristic bands were identified and evaluated to monitor the changes in cellulose crystallinity over time. The authors also stated that NIR spectroscopy was able to differentiate paper from different sources in their “as-received” state.

Hajji and coworkers employed MIR spectroscopy, X-ray diffraction, and energy dispersive X-ray fluorescence to evaluate artificially aged documents. They compared the results with restored documents from different centuries, which had been stored under extreme conditions. Although the research was not in a forensic context, the analysis of the documents' spectral features under different conditions allowed the authors to identify changes in the paper composition and the storage conditions [7].

Trafela and coworkers [12] employed MIR spectroscopy and Partial Least Squares Regression in order to date and quantify the degree of polymerization of the cellulose, pH, ash, and lignin content in paper samples dated from 1850 to 2007. The authors commented on the difference in paper composition of documents pre-1850 and post-1850, making it necessary to build a different model for each period. For the age estimation models, the authors achieved standard error of prediction values of 8.6 years for documents pre and post 1850.

Some of the works reported changes in the degree of the polymerization of cellulose, and several others mentioned degradation processes that could be identified related to the age of the paper [3,13]. In some cases, the decrease in the magnitude of signal at 1425 cm^{-1}, 1370 cm^{-1}, and 900 cm^{-1} in the MIR spectral region were identified and related to the cellulose crystallinity [7]. The decrease of intensity in the absorption bands at 1010 cm^{-1} and 1420 cm^{-1}, and the 1086/1096 cm^{-1} ratio were also associated with the paper degradation [14]. In fact, the majority of works dealing with paper degradation are focused in specific bands rather the whole spectrum. Although important information can be achieved, the complex composition of paper may have been disregarded due the selection of only few (usually two) specific bands.

Due the complexity and variety of paper samples, multivariate analysis can contribute to the study by evaluating the complete spectrum of each sample from the dataset. The information contained in the entire spectrum is more precise (technical term) and can provide valuable knowledge about the samples. One major question that

arises with the application of multivariate techniques to such complex datasets is how to account for variability among the samples in the analysis. Most of methodologies proposed follow one of two approaches, either for naturally aged documents or artificially-aged documents. When naturally-aged documents are evaluated, the paper variety of document from the same year is not taken into consideration, which can lead to misinterpretation. Although differences between papers of different years are also related to the degradation process, paper composition due differences in raw material and the manufacturing process also show up in the model. On the other hand, for artificially-aged documents, most studies usually analyze the same sample as it has aged over time. This process, artificial aging, is out of the scope of the present work.

At the present time, the study of variations in sheet paper composition of documents from the same date has not been well explored, although this study forms the basis of future forensic applications. For non-destructive methods, which usually do not need sample preparation, variations in paper composition can be established by studying the spectral features, or by using multivariate techniques [15,16]. These are important since variations can be found both in the raw material used to produce the paper as well as among the inorganic compounds used for filling and coating the paper related to the purpose of use.

The goal of the present work is to propose a preliminary study to demonstrate the complexity of document dating problems and sample variability in a forensic context and employ different chemometric approaches to analyze these. To do this, Fourier Transformed Infrared (FTIR) spectroscopy was used for spectral acquisition and techniques such as preprocessing and variable selection were employed to deal with the high variability in sheet paper samples to estimate the age of the paper of a document.

2. Theory

PCA [17,18] is a popular exploratory analysis technique for describing the maximum variability of a dataset in a new space of reduced dimensionality. The **X** matrix containing the acquired data is decomposed into two other matrices **T** and $\mathbf{P}^T$, called scores and loadings matrices, respectively (Equation (1)):

(1) $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_x$

The dimensions for **X**, **T**, and $\mathbf{P}^T$ are (N x J), (N x A), and (A x J), respectively. A is the number of Principal Components needed to described the useful information in the data; N and J are the number of samples and variables, respectively; $\mathbf{E}_x$ is the residual matrix, with same dimensions as **X**. A PCA model aims to maximize the variance of the dataset and the spectral features that are related to the variance sources. As an exploratory technique, PCA is not able to estimate the age for an unknown sample. However, PLS can be employed for this purpose.

PLS is a well-known multivariate technique [19,20] that aims at building a mathematical model based on the covariance between the spectral information and the parameter of interest ***(y)***, in this particular case, document age. After building a model defined by an optimized number of Latent Variables (LV), a vector of regression coefficients is defined. This in turn, is used to estimate the $\mathbf{by}$, which is the predicted value for the parameter of an unknown sample with a given spectrum. In this case, the model will be built maximizing the covariance between **X** and **y** ruled by Equations (2), (3), (4), (5)), as described by Wold [20]. Initially, the X matrix is decomposed in **X**-scores (**T**) and X-loadings ($\mathbf{P}^T$), the product of which, when summed with X-residuals ($\mathbf{E}_x$) serves as a good predictor of the **X** matrix (Equation (1)). However, in PLS **T** must also be a good predictor of the parameter **y**, therefore Equation (3) is valid, in which **q** (A x 1) is the y-loading and $\mathbf{e}_y$ is the y-residual. **T** is a linear combination of the original variables of **X**, in which the coefficients are stored in a weight matrix **W** (J x A), that can be used to obtain the regression coefficients **(b)** of the PLS model as shown in Equation (4). Now, the unknown **y** value of a new sample can be estimated ($\mathbf{by}$) by its spectrum $\mathbf{x}_{unk}$, according Equation (5).

(2) $\mathbf{X= TP^T + E_x}$

(3) $\mathbf{y= Tq + e_y}$

(4) $\mathbf{b= W (P^TW)^{-1} q}$

(5) $\mathbf{by= x_{new}\, b}$

Several extensions of PLS models can also be found in the literature. Among them, the sparse PLS (sPLS) uses a penalty term to force an optimized number of regression coefficients to zero [21]. The main idea behind sPLS is to perform variable selection whilst reducing the noise generated by uninformative variables. The sPLS methods uses a Lasso approach to add a constraint to the sum of the squares criterium for regression purposes, as discussed by Refs. [22,23]. In this case, the regression vector is constrained to zeros to improve the performance of the regression model. Not

only the number of sparse Latent Variables (sLV) must be optimized like in PLS, but also the number of regression coefficients which are forced to zero must be optimized. The percentage of regression coefficients that are forced to zero is known as the sparsity level.

To assess the model performance and decide which is the most adequate, several figures of merit for the built models can be used as criteria. Examples of these parameters are the Root Mean Square Error (RMSE), the determination coefficient (R2) and bias for calibration, cross validation, and prediction sets.

Prior to building the model, spectral corrections are needed. The acquisition of spectral data provides relevant information about the presence (or absence) and the concentration of chemical compounds. The dataset would also include a large amount of information regarding physical phenomena, noise and/or errors, depending on such things as the technique employed, equipment, experimental conditions, and accessories. Irrelevant information can obscure the information related to the property of interest. Therefore, to avoid this problem, several mathematic tools known as preprocessing techniques can be used on the samples or variables.

Usually, preprocessing techniques such as normalization, baseline correction, smoothing and derivatives are extremely useful [24]. In some cases, a different kind of information related to interfering compounds needs to be attenuated; therefore, advanced preprocessing techniques are needed. Examples of such techniques are Orthogonal Signal Correction (OSC) [25,26] and Generalized Least Squares Weighting (GLSW) [[27], [28], [29]].

OSC is based in the concept that most variability present in a dataset have a small predictive value. Thus, it is possible to find and remove from the dataset the component that represent the maximum variability, which is orthogonal to the parameter of interest. On the other hand, GLSW aims to estimate a filtering matrix to down-weight clutter contribution, i.e., interfering related information. This information is estimated based on the differences between the samples that should be similar, and afterwards it is filtered from the original data. The filtering matrix G can be calculated using Equation (6):

(6) $\mathbf{G= VD^{-1} V^{t}}$

In which **V** is the eigenvector matrix, and **D** is the weighted version of singular values that can be calculated by Equation (7):

(7)

$$\boldsymbol{D} = \sqrt{\frac{\boldsymbol{S}^2}{\alpha} + 1_D}$$

In which **S** is the diagonal matrix of singular values, and the α-value is a scalar value settled to weigh the filtering matrix, which will depend upon the dataset. High α-values show less effect from the filter, while low α-values impose more filtering effect. For further information about both techniques, the reader is encouraged to perform additional reading [25,26,28,30].

3. Materials and methods

To perform this study, reports from 15 different years (1985, 1986, 1987, 1990, 1991, 1992, 1995, 1996, 1997, 2001, 2002, 2003, 2010, 2011, 2012) were provided by the Spanish General Commissary of Scientific Police (Documentoscopy section, Spain). For each year, five reports were provided, each containing different number of sheets, but having an average of five sheets each. Different regions of each paper sheet (top, bottom, left, and right sides) were sampled to obtain two spectra (a duplicate) for every region; resulting in eight spectra per sheet, around, but not equal to, 3000 spectra in total. The samples were divided into Calibration and Prediction sets, with two Prediction sets built separately: (1) one whole report from each year, namely the Report Prediction set and (2) one sheet from each of the remaining reports, namely the Sheet Prediction set. The Report Prediction set was chosen in two ways: (1) dataset-PCA: for the individual report, a PCA was performed for each year and one complete report was chosen to include its variability within the total variability for that year. This is to guarantee that the variability of the prediction set was included in the model and no extrapolations were made; (2) dataset-RANDOM: a random choice among the individual reports was made, in order to challenge model performance and test its application to a forensic context. Thus, two datasets were employed, and after outlier removal, the number of spectra in the Calibration, Report Prediction and Sheet Prediction were 1883 (64%), 591 (20%) and 472 (16%), respectively for the Report Prediction set chosen by PCA and 1946 (66%), 600 (20%) and 400 (13%), for the random choice. Table 1 describes both datasets:

Table 1. Number of spectra in Calibration and prediction sets chosen for Dataset-PCA and Dataset-RANDOM.

Number of samples	Dataset-PCA	Dataset-RANDOM
Calibration	1883 (63%)	1946 (66%)
Report Prediction	591 (20%)	600 (20%)
Sheet Prediction	472 (16%)	400 (13%)
Total number of samples:	2946	

The spectra were acquired in the MIR region with a Nicolet iS10 spectrometer (ThermoFisher Scientific, MA, USA) using the ATR accessory Smart iTR diamond. The spectral range employed was 4000-650 cm^{-1}, with a resolution of 4 cm^{-1}, 0.482 cm^{-1} of increment, and 32 scans per spectrum. No sample pretreatment was needed since the extremities of the paper sheets could easily be inserted in the ATR accessory without damaging the document. The 2450-2235 cm^{-1} region, related to the CO_2 absorption, was removed to eliminate this variability from the dataset.

Afterwards, both datasets were preprocessed and the PLS models were built aiming to estimate the manufacturing year of the paper for each document. To assess the RMSE values for all models, the predicted y-values were rounded to its nearest integer. Different models were compared in order to identify differences among them, using the F-test to compare the RMSE values and the t-test to evaluate the bias of each model.

Different preprocessing techniques were evaluated to identify and minimize the differences between documents from the same year. To do this, OSC and GLSW filters were employed before the PLS modelling. Then, the results were then compared. The sparse method was also applied as a variable selection technique for comparison.

To build a sPLS models, two parameters must be optimized to define the best level of sparsity in the model: the number of sLV and the sparsity level. The models were built using 1 to 20 sLV with sparsity levels varying from 99 to 91%, i.e. including from 5 to 150 variables. To define the optimum sPLS models, two parameters were evaluated: the response surfaces to monitor the RMSEP, and R2 for the models. For all the models built, a cross validation was performed using the leave-one-document-out approach.

All the multivariate analyses were performed in Matlab using the PLS_Toolbox (Eigenvector Research Inc., USA). The sPLS algorithm was used as described in Ref. [31].

4. Results and discussion

4.1. Spectral features

After the data acquisition, the spectra were compressed to increase the spectral increment. This compression was performed using the average of intensity defined in a window of 4 points, reducing the spectral channel to ¼ of the original amount. This step was performed to minimize the noise (Fig. 1).

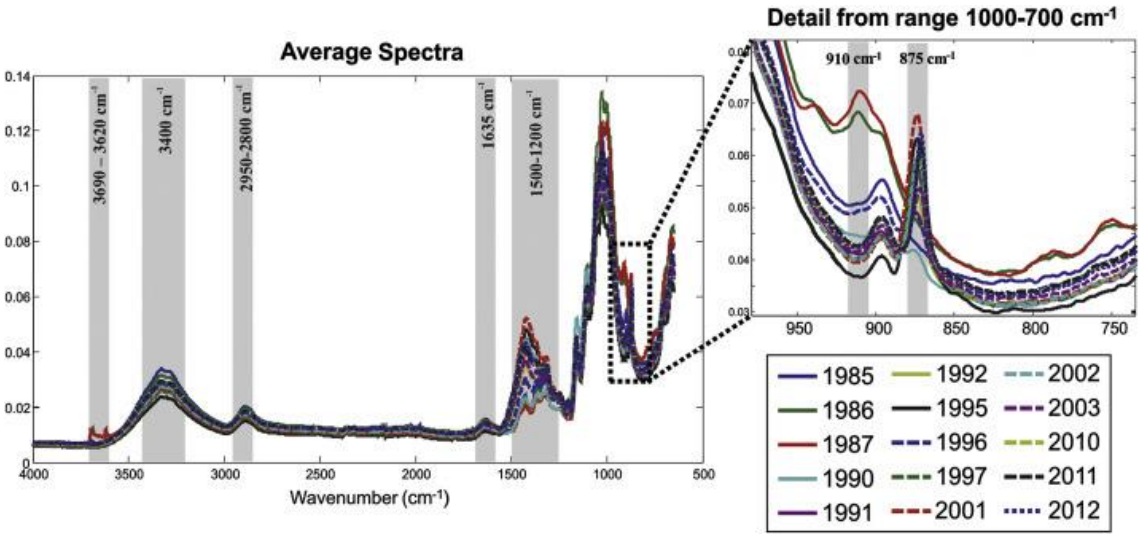


Fig. 1. Average compressed spectra of several documents from different years. The detailed 1000–700 cm−1 range emphasizes the calcium carbonate absorption region.

The MIR spectra revealed the spectral features associated with both calcium carbonate and kaolinite. In addition, different organic compounds present in the paper composition were expected to show up because of the variety of raw materials used in paper production. Characteristic bands for kaolinite can be seen at 3690 cm−1, 3620 cm−1, and 910 cm−1 in the 1986 and 1987 documents, and a small contribution in the paper manufactured in 1985. In fact, those documents are especially different from the others, since they were used for typewriters, and probably had a different composition. In addition, carbonate contribution can be noticed at 875 cm−1 in all

documents except the ones from 1985 to 1990 and 1996. The absorptions at 1410-1420 cm−1 could be attributed to calcium carbonate. However, this absorption band can be also overlapped with the cellulose bands that appear around 1420-1430 cm−1 [14,32,33], and are related to the cellulose crystallinity.

The inorganic and other organic compounds used in paper manufacturing have changed over the years and also among the types of paper produced by different companies. Nonetheless, cellulose is still the major compound, and its spectral contributions are present in the paper compositions regardless of brand or type. The absorptions at 1025 cm−1, 1160 cm−1, 1315 cm−1, and 2890 cm−1, related to different C-H, C-OH, C-CH2, C-O-C vibrations can also be found. Intramolecular vibrations from the H bond in OH-O can be seen at 3400 cm−1, while the absorbed water molecules provide an absorption band at 1635 cm−1. All these bands agreed with those reported in the literature for the most important compounds found in the paper spectrum (Table 2) [14,33].

Table 2. Important IR absorption bands present in paper compounds [14,33].

IR absorption bands (cm−1)	**Assignment**
712 and 870	Calcium carbonate
900–1200	CO and CC stretching
1500–1200	COH in plane bending; CCH and OCH def. stretching; HCH bending
1635	H2O absorbed
2800–2950	CH stretching in CH, CH2, CH3, symmetric
3400	OH-O intramolecular H bond
3690 and 3620	Kaolinite

4.2. Preprocessing and PCA

Different preprocessing techniques were evaluated in this study, and the results from their corresponding PLS models were used as criteria to choose the best technique. However, prior to building the PLS models, PCA models were carried out to investigate the spectral variability of the documents. Techniques such as SNV (Standard Normal Variate) normalization and Savitzky-Golay smoothing (2nd order polynomial and a 21-point window width) were applied to minimize the scattering and noise effects of the raw spectra.

Initial analysis of the samples did not show significant variance between the regions of each sheet (top, bottom, left and right). Although there were differences between sheets among sheets of paper from the same report, the differences were not significant when compared to the variability among the documents. The results from the PCA models showed that the two initial principal components (PCs) explained 77% and 11% of data variability, respectively.

Fig. 2a showed two clusters in the scores scatter plot; and their corresponding loadings plot are shown in Fig. 2b. The positive values of the scores in PC1 are related to the absorption at 875 cm^{-1} and 1415 cm^{-1} (Fig. 2b). As previously mentioned, those contributions can be associated to the presence of calcium carbonate and its absorption at 1415 cm^{-1}. The other cluster showing negative contribution in PC1 is related to the absorptions at 910 cm^{-1}, 1000 cm^{-1}, 3620 cm^{-1}, and 3690 cm^{-1}, which are related to kaolinite [14,32,33].

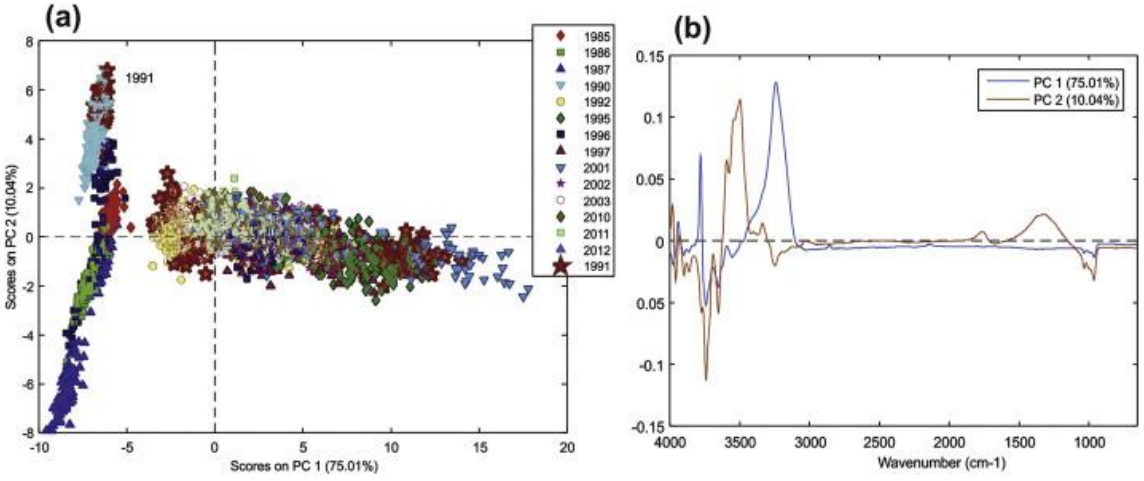


Fig. 2. Principal Component Analysis for all documents: (a) score and (b) loading plots for PC1 and PC2. Dataset preprocessed with SNV, smoothing filter and mean center.

In PC2, the positive contributions would relate to the absorption bands at 1000 cm^{-1}, 1055 cm^{-1}, and 1160 cm^{-1}, which are associated to the cellulose vibrations. The negative contributions in PC2 are associated to the absorptions at 1106 cm^{-1} and 913 cm^{-1}, which can be related to cellulose and kaolinite, respectively.

It is important to notice that the maximum variability among the datasets is related to inorganic fillers, i.e., the documents that had been coated with either calcium carbonate or kaolinite. Use of a coating compound depends on the type and brand of the sheet paper sample analyzed, which can vary significantly within documents from the same year. Indeed, as expected, the documents from 1991 showed a variability as high as the total (red stars in Fig. 2a), considering the two initial PCs of the dataset preprocessed with SNV, smoothing filter and mean center. This variability is not related to document age, and therefore, the changes related to the cellulose along the aging process must be evaluated separately from these interfering compounds. For this reason, one possibility is to employ different methods (OSC and GLSW filters) to suppress the interfering contributions from the data and access the information related to the aging process.

As previous described, two different datasets were chosen to evaluate model performance regarding Report Prediction set. Dataset-PCA was built by choosing one whole report according to PCA results for each year. No extrapolation regarding an unexpected variability of prediction set was found, thus all the variability in the dataset was taken into account. From a forensic point of view, however, it was important to evaluate model robustness for predicting samples with an unexpected variability. The dataset-RANDOM was built. Some of the documents chosen to compose the Report Prediction set showed a high variability when compared to the other documents from the same year. Fig. 3 shows the results of a PCA model for documents from 2003, from which Document 5 was chosen. Document 5 differed from the others regarding the absorption at 1410 and 870 cm^{-1}, which is found, in the literature as being the region associated to calcium carbonate content.

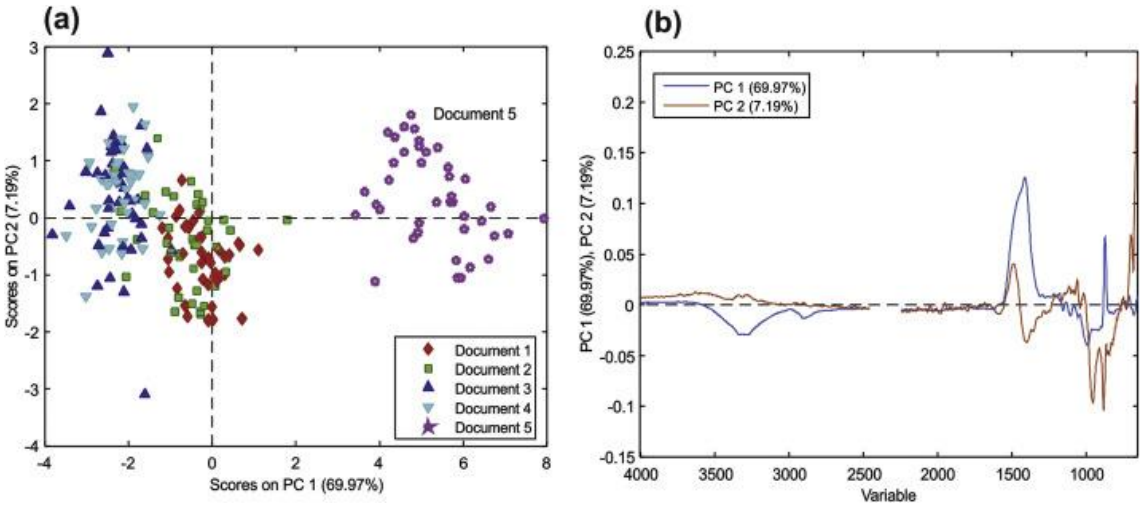


Fig. 3. Principal Component Analysis for 2003 documents: (a) score and (b) loading plots for PC1 and PC2. Dataset preprocessed with SNV, smoothing filter and mean center.

4.3. PLS models, preprocessing and variable selection

To predict document age, the models were built and subsequently tested using the calibration and prediction sets, respectively. Table 3 shows the results for the four models built: (1) model 1: PLS model with data preprocessed with SNV, smoothing and mean-centering; (2) model 2: PLS model with SNV, smoothing, OSC and mean-centering; (3) model 3: PLS model with SNV, smoothing GLSW and mean-centering; (4) model 4: sPLS model with SNV, smoothing and mean-centering.

Table 3. Results from PLS models using different preprocessing techniques Model 1 (PLS model with SNV, smoothing and mean-centering); model 2 (PLS model with SNV, smoothing, OSC and mean-centering); model 3 (PLS model with SNV, smoothing GLSW and mean-centering); model 4 (sPLS model with SNV, smoothing and mean-centering).

Dataset-PCA

Training Set					Report Prediction					SHEET Prediction	
Model	LV R2pred	RMSEC bias	R2cal	bias	RMESCV	R2cv	biascv	RMSEP	R2pred	bias	RMSEP
1	4	4.4	0.86	−0.01	4.7	0.83	0.04	3.8	0.90	0.35	4.3
2	1	2.5	0.96	−0.00	4.5	0.85	0.02	4.0	0.89	0.32	3.7
3	2	4.2	0.87	0.01	4.6	0.86	0.01	3.6	0.91	0.22	4.2
4	5 (105*)	4.2	0.87	0.01	4.5	0.88	−0.06	4.0	0.88	0.15	4.5

R2pred	bias
0.86	0.05
0.90	0.24
0.87	0.07
0.85	0.00

Dataset-RANMDOM

Training Set					Report Prediction					SHEET Prediction	
Model	LV R2pred	RMSEC bias	R2cal	bias	RMESCV	R2cv	biascv	RMSEP	R2pred	bias	RMSEP
1	4	4.1	0.77	0.00	4.4	0.74	0.01	5.1	0.74	2.11	4.0
2	1	2.6	0.90	0.00	4.5	0.74	0.04	4.3	0.80	1.46	3.6
3	3	3.8	0.80	0.00	4.2	0.76	−0.00	5.0	0.75	1.95	3.7
4	5 (100*)	4.4	0.86	0.00	4.3	0.73	0.87	4.7	0.86	0.64	4.3

RP2pred	bias
0.78	0.44
0.82	0.22
0.82	0.34
0.87	0.97

4.3.1. PLS (Model 1)

PLS model was built using as preprocessing only Savitzky-Golay smoothing filter, SNV and mean centering as previously described. In general, it is possible to notice that regarding the criteria for choosing the prediction set, both models shows similar results with respect to error, except for the RMSEP value for the Report prediction set. RMSEP value for Report prediction set can change significantly with respect to the way the dataset has been chosen. This is due the fact that, with the dataset-RANDOM, the variability of some of the documents selected to compose the dataset is not included in the model, as previously discussed. This reflects an important error for the Report Prediction set (RMSEP = 5.1 years) and also for bias (see Table 3), which according a t-test was significant for the Report Prediction set in model 1 using the Dataset-RANDOM.

4.3.2. OSC filter (Model 2)

The OSC filter was employed using one component to attenuate differences between samples from the same year. The model built for both datasets provide similar results, except for the Report Prediction set, due the high variability in the prediction set for dataset-RANDOM. The bias value for dataset-RANDOM was high and significant, according the t-test results. On the other hand, the Sheet Prediction set was well predicted independently of how the dataset is built.

4.3.3. GLSW filter (Model 3)

As previously mentioned for GLSW, the α value needed to be adjusted to remove information from the interfering compounds without losing relevant variability among the data. However, since the variability of the documents from the same year had the same order of the variability in the whole dataset, the α value was kept as high as possible, otherwise to prevent noisy spectral profiles from the filter effect.

To evaluate the effect of α in the results, RMSEC and RMSECV from the PLS models were monitored whilst varying the α-values (Fig. 4) for both datasets. Fig. 4 shows the effect of the GLSW filter in the spectral profile, for the dataset-PCA, although dataset-RANDOM showed the same behavior. It is possible to visualize how the model error changed with the α-value. When α had a high value (α = 6 or higher), the effect of

the filter decreased and the interfering contribution was higher, leading to models that still needed to be improved, in fact a model similar to the PLS without any filtering. When α was equal to 1.66, RMSEC and RMSECV were not only similar but minimum, providing the best model for both datasets. However, as this value decreased, the effect of the filter produced noisier spectral profiles, in which the RMSEC and RMSECV showed a sudden increase when the α-value was set to 0.001. This means, as expected, that at some point the filter effect was so strong it started to remove part of the relevant information from the data, leading to poorer models.

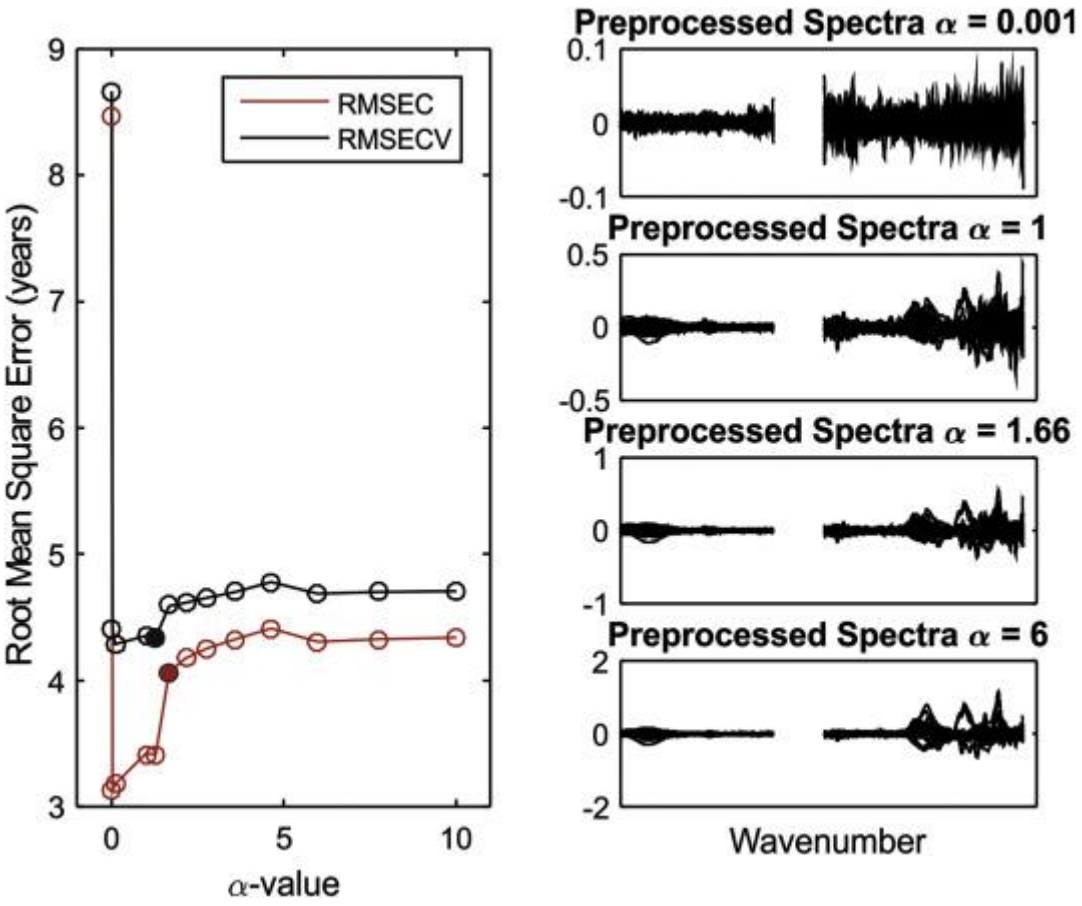


Fig. 4. (a) RMSECV and RMSEP values for various α-values when using the GLSW preprocessing; (b) spectral differences for several α-values. Results for Dataset-PCA.

In this work, the α-value chosen as the most suitable for the dataset was 1.66, for both dataset-PCA and dataset-RANDOM. Although the model built was not significantly improved, according the F-test for RMSE values, when compared to the PLS model without the filters, the contribution from the interfering compounds was attenuated without noticeable loss of the significant information.

4.3.4. sPLS model (Model 4)

Another approach performed to compare the prediction ability of the models was to employ variable selection methods. For this reason, sPLS models were built using the preprocessed spectra with SNV, smoothing, and mean centering, as previously described. To reach the best model, not only the number of LVs had to be optimized, but also the sparsity level, i.e. the percentage of regression coefficients that are forced to be zero.

Fig. 5a shows the map of RMSECV values for the models built with 1–15 sLV (y-axes) and sparsity level varying from 98 to 91% (x-axes). High values of RMSECV are represented in red, while the best models are marked with a dark blue color. It is possible to notice a stability region after 3 sLV. Although there are no expressive changes according to the sparsity levels, there can be seen a minimum value for RMSECV with sparsity levels lower than 94% for models with 5 and 6 sLV. The models in the stability region have similar performance, however the model with 5 sLV and 94% of sparsity level was chosen as the simplest one. In this case, the RMSECV obtained was 4.5 years, R2 of 0.88 and bias of −0.06 years, considered not significant according to the t-test. Fig. 5b shows the included variables in the chosen model, while Fig. 5c shows the RMSEC and RMSECV according sLV with sparsity level of 94%.

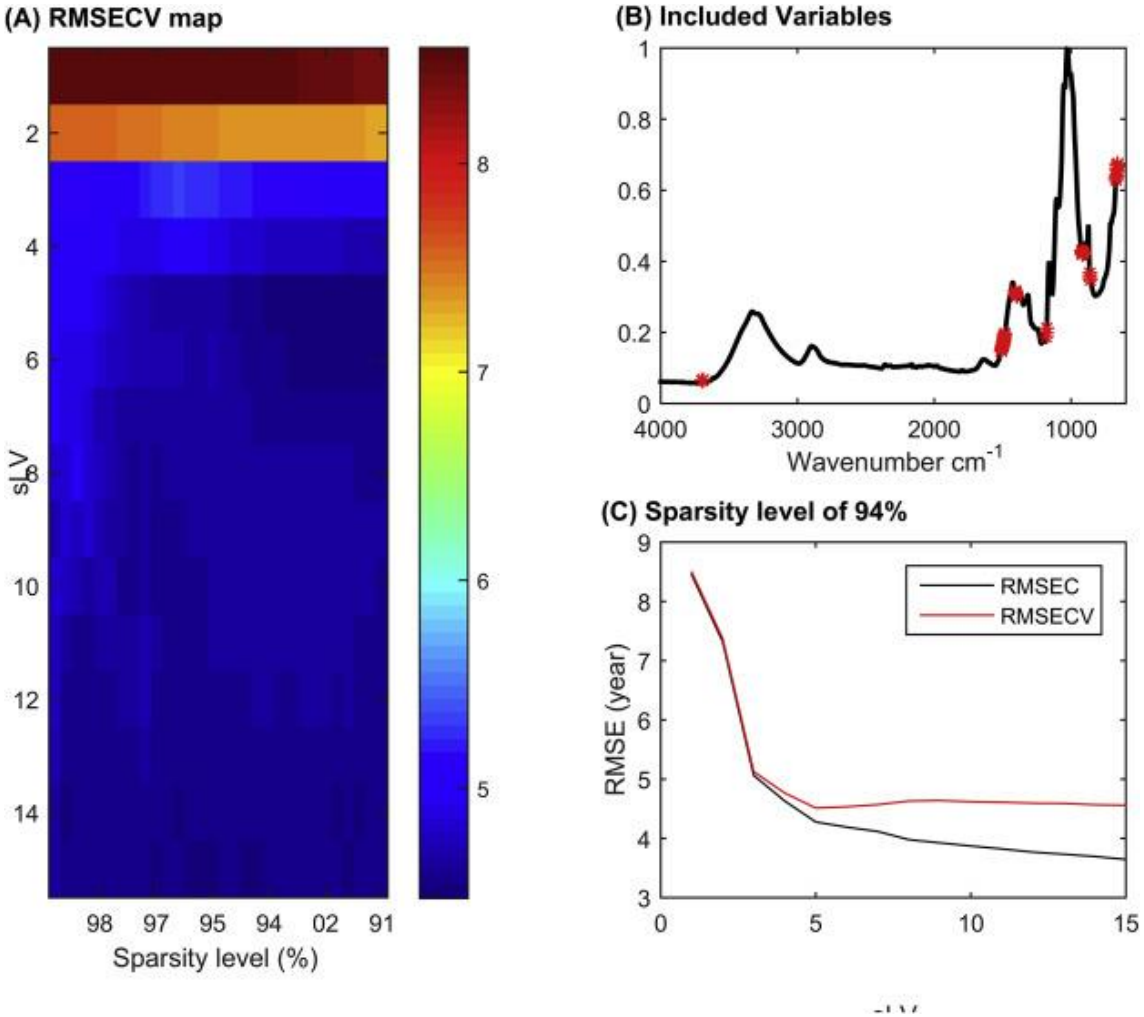


Fig. 5. (a) RMSECV map for the models built with sLV from 1 to 15 against the sparsity level; (b) the average spectra the variables included in the model; (c) RMSEC and RMSECV value according to sLV.

4.3.5. Comparing models

The different models were built and used to compare not only regarding the different preprocessing techniques mentioned above, but also how the datasets were built (Table 3). Regarding the table of results (Table 3), it is possible to notice the change in the number of LVs in the PLS models with and without the OSC and GLSW filters. This decrease of LV when applying the filters is due the fact that the filters removed part of the variance in X, which is not related to the age of paper, leading to a simplification of both datasets. It is important to emphasize that, at a 95% of confidence level, the bias of the Report Prediction set was considered significant for all models built with dataset-RANDOM, except for the sPLS model.

Without any filters, the PLS model built provided a RMSECV value of 4.7 years for dataset-PCA while, for the dataset-RANDOM the RMSECV was equal to 4.4 years. It is clear that the cross-validation step for dataset-PCA has taken into account the document papers with the highest number of variations for each year studied; but when a leave-one-document-out validation was performed, that variability was easily detected in the high value of RMSECV. As a consequence, prediction sets are more successful when compared with the dataset-RANDOM. When a filter was applied, however, that behavior was not observed. Regarding the errors, it is possible to notice certain stability when comparing the results of dataset-PCA and dataset-RANDOM for each model. Although the RMSE values shows stability, the bias for the Report Prediction set of dataset-RANDOM, on the other hand, reflects a variability that was not included in the models. As stated before, those values were considered significant at the 95% confidence level.

In a fair comparison between models, dataset-PCA is indeed the most appropriated to be considered. In a real-life application, it is mandatory to perform previous tests to check if the variations in an unknown specimen, in this case paper, are within the model. Fig. 6 shows the regression models for dataset-PCA. The scores of Variable Importance in Projection (VIP) in Fig. 3 show that the absorptions bands at $1483\,cm^{-1}$, $1415\,cm^{-1}$, $1026\,cm^{-1}$, $914\,cm^{-1}$, and $887\,cm^{-1}$ were, in general, the most important for the model building. The interfering compounds had also high contributions in the model building, except for model 2 (built with the OSC filter). In fact, model 2 showed the best results in a general point of view and is the only model that does not include variables related to the filling compounds, only cellulose-related absorption bands are important in this scenario.

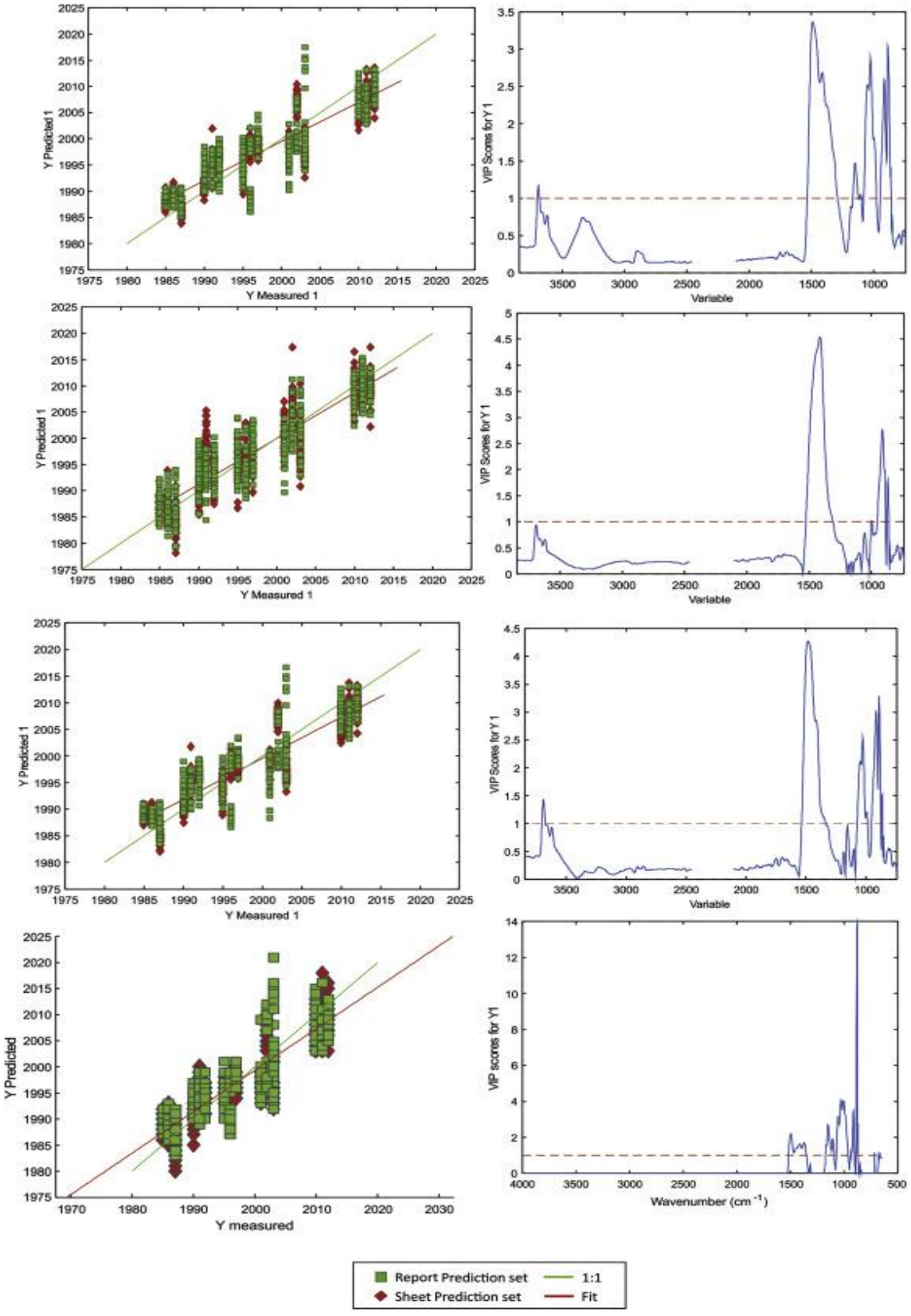


Fig. 6. Results for the regression models using (a-b) model 1: PLS model; (c-d) model 2: PLS with OSC (1 component); (e-f) model 3: PLS with GLSW (α = 1.660); (g-h) model 4: sPLS model.

The effect of the filter is noticeable in the regression plots and VIP scores of the PLS models (Fig. 6). The VIP scores plot shows the difference between the variable importance in each model. For the model built with the OSC filter, the variables responsible for the model building were related to the 1412 and 914 cm^{-1}, which according the literature [7], shows a modification during the aging process due the loss of cellulose crystallinity. Some other research reports have stated that these absorptions are related to filling compounds [14,33], making the regions ambiguous regarding determination of the paper's aging process. Although all models showed the mentioned variables as important, models 1, 3 and 4 showed the influence of other spectral regions, such as 3690 and 870 cm^{-1}, suggesting that the models were still under influence of the inorganic compounds.

Looking the results in Table 3 and the regression plots in Fig. 6, we can see that the OSC filter is the most successful in attenuating the variations between documents from paper manufactured in the same year. The reflection of the filter can be observed not only for the cross-validation step in dataset-PCA, but also for all values for dataset-RANDOM.

5. Conclusions

In the present study, a non-destructive methodology using infrared spectroscopy and chemometric techniques is proposed for dating the sheet paper used in documents from different years. A PCA model was built to identify the differences on the spectral features of the document pages, identifying an important influence from the inorganic fillers used in paper manufacturing. Compounds such as kaolinite and calcium carbonate were identified in the samples analyzed. They showed a high influence on the initial PLS models, which made them poor for any prediction purposes.

Three chemometric approaches were proposed to overcome variations in the sample set, which was built in different ways: (i) employing a variable selection method (sPLS); and exploring different preprocessing techniques such as (ii) OSC and (iii) GLSW. The preprocessing techniques, especially the OSC filter, had a very important role in the model building. The high variability among the documents from the same year was attenuated with the GLSW and OSC filters, providing adequate models for document dating. Although, acceptable values for RMSECV and RMSEP were obtained (around 4 years), the VIP scores for the models showed that the inorganic compounds were still

influencing the models, suggesting that other techniques need to be explored to improve the results.

The model employing OSC filter showed superior performance compared to the models previously built. In fact, it was also more robust since the different ways of building the dataset did not show a significative influence on model performance with respect to error. Bias value, however, was proven to be significative when the dataset was randomly built. The complexity of the model also decreased with the use of the OSC filter.

Dating documents by analyzing only the information about the paper they are written or printed on can be tricky due to the chemical differences caused by the paper composition. Different chemical characteristics can lead to differences in the aging process, and for forensic applications, this variability must be explored. The main objective of this research is to open a discussion about the advantages and drawbacks when implementing analytical techniques in forensic contexts, more specifically for document analysis, providing also chemometric approaches to deal with real problems faced by forensic experts. This study shows not only the potential of infrared spectroscopy, already discussed in the literature, but also how chemometric techniques can be useful for document dating, providing analytical methodologies that can be employed in scientific laboratories on a daily basis.

Acknowledgments

The authors would like to acknowledge the Spanish General Commissary of Scientific Police (Documentoscopy section, Spain) for providing the analyzed documents. Also the funding agencies INCTAA (Processes nº.: CNPq 573894/2008-6; FAPESP 2008/57808-1), NUQAAPE –FACEPE (APQ-0346-1.06/14), Núcleo de Estudos em Química Forense – NEQUIFOR (CAPES AUXPE 3509/2014, Edital PROFORENSE 2014), CNPq (PVE400264/2014-5), FACEPE and CAPES (PDSE scholarship process number BEX 7712/15-4), are acknowledged. The English version was revised by Sidney Pratt, Canadian, BA, MAT (The Johns Hopkins University), RSAdip (TEFL) (Cambridge University).

References

[1] M. Ezcurra, J.M.G. G ongora, I. Maguregui, R. Alonso, Analytical methods for dating modern writing instrument inks on paper, Forensic Sci. Int. 197 (2010) 1e20, https://doi.org/10.1016/j.forsciint.2009.11.013.

[2] M. Calcerrada, C. García-Ruiz, Analysis of questioned documents: a review, Anal. Chim. Acta 853 (2015) 143e166, https://doi.org/10.1016/j.aca.2014.10.057.

[3] M.C. Area, H. Cheradame, Paper aging and degradation: recent findings and research methods, BioResources 6 (2011) 5307e5337, https://doi.org/10.1016/j.polymdegradstab.2008.03.016.

[4] A. Schedl, T. Zweckmair, F. Kikul, U. Henniges, T. Rosenau, A. Potthast, Aging of paper e ultra-fast quantification of 2,5-dihydroxyacetophenone, as a key chromophore in cellulosics, by reactive paper spray-mass spectrometry, Talanta 167 (2017) 672e680, https://doi.org/10.1016/j.talanta.2017.02.053.

[5] X. Zou, T. Uesaka, N. Gurnagul, Prediction of paper permanence by accelerated aging I. Kinetic analysis of the aging process, Cellulose 3 (1996) 243e267, https://doi.org/10.1007/BF02228805.

[6] X. Zou, T. Uesaka, N. Gurnagul, Prediction of paper permanence by accelerated aging II. Comparison of the predictions with natural aging results, Cellulose 3 (1996) 269e279, https://doi.org/10.1007/BF02228806.

[7] L. Hajji, A. Boukir, J. Assouik, S. Pessanha, J.L. Figueirinhas, M.L. Carvalho, Artificial aging paper to assess long-term effects of conservative treatment. Monitoring by infrared spectroscopy (ATR-FTIR), X-ray diffraction (XRD), and energy dispersive X-ray fluorescence (EDXRF), Microchem. J. 124 (2016) 646e656, https://doi.org/10.1016/j.microc.2015.10.015.

[8] F. Ka cík, D. Ka cíkov a, M. Jablonský, S. Katu s c ak, Cellulose degradation in newsprint paper ageing, Polym. Degrad. Stabil. 94 (2009) 1509e1514, https:// doi.org/10.1016/j.polymdegradstab.2009.04.033.

[9] J.R. Martínez, A. Nieto-Villena, J.A. Cruz-Mendoza, G. Ortega-Zarzosa, A.L. Guerrero, Monitoring the natural aging degradation of paper by fluorescence, J. Cult. Herit. (2017) 1e6, https://doi.org/10.1016/j.culher.2017.01.011.

[10] C.K. Muro, K.C. Doty, J. Bueno, L. Halamkova, I.K. Lednev, C.K. Luber, K.C. Doty,

J. Bueno, L. Hal amkov a, I.K. Lednev, Vibrational Spectroscopy : recent developments to revolutionize forensic science, Anal. Chem. 87 (2015) 306e327, https://doi.org/10.1021/ac504068a.

[11] M. Ali, A.M. Emsley, H. Herman, R.J. Heywood, Spectroscopic studies of the ageing of cellulose paper, Polymer 42 (2001) 2893e2900, https://doi.org/

10.1016/S0032-3861(00)00691-1.

[12] T. Trafela, M. Strlic, J. Kolar, D.A. Lichtblau, M. Anders, D.P. Mencigar, B. Pihlar, Nondestructive analysis and dating of historical paper based on IR spectroscopy and chemometric data evaluation, Anal. Chem. 79 (2007) 6319e6323, https://doi.org/10.1021/ac070392t.

[13] J.C. Williams, A review of paper quality and paper chemistry, Lybrary Trends. Paper Qual (1981) 203e224.

[14] J. Zieba-Palus, A. Weselucha-Birczynska, B. Trzicinska, R. Kowalski, P. Moskal, Analysis of degraded papers by infrared and Raman spectroscopy for forensic purposes, J. Mol. Struct. (2016) 1e9, https://doi.org/10.1016/j.molstruc.2016.12.012.

[15] F.M. Udris¸ tioiu, I.G. T anase, A. a. Bunaciu, H.Y. Aboul-Enein, Paper analysis: nondestructive and destructive analytical methods, Appl. Spectrosc. Rev. 47 (2012) 550e570, https://doi.org/10.1080/05704928.2012.682285.

[16] A. Kher, M.A.R.Y.M. Ulholland, B. Reedy, P. Maynard, Classification of document papers by infrared spectroscopy and multivariate statistical techniques, Appl. Spectrosc. 55 (2007) 1192e1198, https://doi.org/10.1366/ 0003702011953199.

[17] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37e52, https://doi.org/10.1016/0169-7439(87) 80084-9.

[18] R. Bro, A. Smilde, Principal component analysis, Anal. Methods (2014) 2812e2831, https://doi.org/10.1002/wics.101.

[19] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986), https://doi.org/10.1016/0003-2670(86)80028-9, 1e117.

[20] S. Wold, M. Sj€ostr€om, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109e130, https://doi.org/ 10.1016/S0169-7439(01)00155-1.

[21] H. Chun, S. Keles¸ , Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. Ser. B Stat. Methodol 72 (2010) 3e25, https://doi.org/10.1111/j.1467-9868.2009.00723.x.

[22] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, J. Chemom. 26 (2012) 42e51, https://doi.org/10.1002/cem.1418.

[23] M.A. Rasmussen, R. Bro, A tutorial on the Lasso approach to sparse modeling, Chemometr. Intell. Lab. Syst. 119 (2012) 21e31, https://doi.org/10.1016/j.chemolab.2012.10.003.

[24] Å. Rinnan, F.V.D. Berg, S.B. Engelsen, Review of the most common preprocessing

techniques for near-infrared spectra, TrAC Trends Anal. Chem. (Reference Ed.) 28 (2009) 1201e1222, https://doi.org/10.1016/j.trac.2009.07.007.

[25] T. Fearn, On orthogonal signal correction, Chemometr. Intell. Lab. Syst. 50 (2000) 47e52, https://doi.org/10.1016/S0169-7439(99)00045-3.

[26] S. Wold, H. Antti, F. Lindgren, J. €Ohman, Orthogonal signal correction of nearinfrared spectra, Chemometr. Intell. Lab. Syst. 44 (1998) 175e185, https://doi.org/10.1016/S0169-7439(98)00109-9.

[27] H. Martens, M. Høy, B.M.Wise, R. Bro, P.B. Brockhoff, Pre-whitening of data by covariance-weighted pre-processing, J. Chemom. 17 (2003) 153e165, https://doi.org/10.1002/cem.780.

[28] B.M. Wise, N.B. Gallagher, R. Bro, J. Shaver, W. Windig, R.S. Koch, Chemometrics

Tutorial for PLS _ Toolbox and Solo, Eigenvector Research, Inc.,Wenatchee, 2006.

[29] R. Fern andez-Varela, J.M. Andrade, S. Muniategui, D. Prada, Comparing the weathering patterns of six oils using 3-way generalized Procrustes rotation and matrix-augmentation principal components, Anal. Chim. Acta 683 (2010)84e91, https://doi.org/10.1016/j.aca.2010.10.020.

[30] N.B. Gallagher, Detection, classification, and quantification in hyperspectral images using classical least squares models, in: H.F. Grahn, P. Geladi (Eds.),Tech. Appl. Hyperspectral Image Anal, John Wiley & Sons Ltd, West Sussex, 2007, pp. 181e202.

[31] R. Calvini, A. Ulrici, J.M. Amigo, Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral

imaging, Chemometr. Intell. Lab. Syst. 146 (2015) 503e511, https://doi.org/10.1016/j.chemolab.2015.07.010.

[32] V. Causin, C. Marega, A. Marigo, R. Casamassima, G. Peluso, L. Ripani, Forensic differentiation of paper by X-ray diffraction and infrared spectroscopy, Forensic Sci. Int. 197 (2010) 70e74, https://doi.org/10.1016/j.forsciint.2009.12.056.

[33] F.M. Udris¸ tioiu, A. a. Bunaciu, H.Y. Aboul-Enein, I.G. T anase, Application of Micro-Raman and FT - IR spectroscopy in forensic analysis of questioned documents, GU J Sci. 25 (2012) 371e375, https://doi.org/10.1080/05704928.2012.673188.