ARTICLE IN PRESS

Analytica Chimica Acta xxx (2018) 1–10

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Chemometric approaches for document dating: Handling paper variability

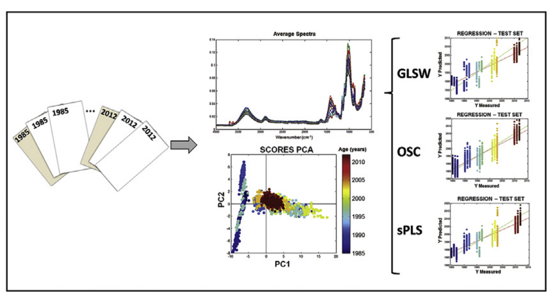
Carolina S. Silva [a, *], Maria Fernanda Pimentel [b], José Manuel Amigo [b, c], Carmen García-Ruiz [d], Fernando Ortega-Ojeda [d]

[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

HIGHLIGHTS

- Infrared Spectroscopy and PLS for document dating.
- Variable selection and preprocessing attenuated chemical differences between papers.
- OSC models showed better results regarding RMSEP value.
- Methodology developed was effective estimating the age of documents.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 January 2018
Received in revised form
7 June 2018
Accepted 11 June 2018
Available online xxx

Keywords:
Documentoscopy
Chemometrics
Paper
Dating
Forensic

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.

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].

* Corresponding author.
E-mail address: carolina.santossilva@ufpe.br (C.S. Silva).

https://doi.org/10.1016/j.aca.2018.06.031

Please cite this article in press as: C.S. Silva, et al., Chemometric approaches for document dating: Handling paper variability, Analytica Chimica Acta (2018), https://doi.org/10.1016/j.aca.2018.06.031

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 ($CaCO_3$) and kaolinite ($Si_2Al_2O_5(OH)_4$) 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 [7–9].

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 $\mathbf{P}^\mathbf{T}$, called scores and loadings matrices, respectively (Equation (1)):

$$X = TP^T + E_x \tag{1}$$

The dimensions for **X**, **T**, and $\mathbf{P}^\mathbf{T}$ 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

Please cite this article in press as: C.S. Silva, et al., Chemometric approaches for document dating: Handling paper variability, Analytica Chimica Acta (2018), https://doi.org/10.1016/j.aca.2018.06.031