Cellulose (2020) 27:5323–5335
https://doi.org/10.1007/s10570-019-02892-1

ORIGINAL RESEARCH

Development of a chemometric methodology based on FTIR spectra for paper dating

Jingjing Xia · Yue Huang · Jixiong Zhang · Xiayu Du · Hong Yan · Qianqian Li · Yang Li · Yanmei Xiong · Shungeng Min

Received: 13 April 2019 / Accepted: 1 December 2019 / Published online: 16 April 2020

Abstract The dating of documents is one of the pending issues to be resolved in many fields. Although there are a variety of analytical methodologies focused on the inks of documents to estimate the dating of documents, the analysis of documents itself attracted little attention. A non-destructive method based on Fourier transformed infrared (FTIR) was proposed in this research to estimate the dating of documents. First, FTIR spectra of journals dated from 1940 to 1980, naturally aged and conserved in the China Agriculture University library were collected. Second, the least squares support vector machines (LS-SVM) was applied to distinguish documents of different dating, and the accuracy was 99.26%. In order to find interested wavenumber that influence the dating process of documents, sparse partial least squares (sPLS) was applied to select informative variables. The average of selected variables was 483 after 100 runs, and the selected variables were focused on the absorption peaks of inorganic components and cellulose. Splicing sPLS with LS-SVM (sPLS–LS-SVM) built model to see the effective of selected variables. Average accuracy of sPLS–LS-SVM model was 99.34%, even the best result could reach 100.00% after 100 circle times. The present work indicates that the possibility of FTIR combined with chemometrics can estimate the dating of documents accurately. Additionally, the wavenumber which influence the dating of documents are mostly focused on cellulose and inorganic components.

Keywords Dating · Documents · FTIR · sPLS · LS-SVM

J. Xia · J. Zhang · X. Du · H. Yan · Y. Li · Y. Xiong (✉) · S. Min (✉)
College of Science, China Agricultural University, Beijing 100193, People's Republic of China
e-mail: xiongym@cau.edu.cn

S. Min
e-mail: minsg@263.net

Y. Huang
College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100193, People's Republic of China

Q. Li
School of Marine Science, China University of Geosciences, Beijing 100086, People's Republic of China

Introduction

Regarding the forensic standpoint, the forensic analysis documents present a real interest, as cases of crimes like falsification, questioned signatures, threatening letters or even terrorist attacks may leave them as evidence (Ortiz-Herrero et al. 2018; Calcerrada and Garcia-Ruiz 2015). As is known, the amount of carbon 14 is a standard for determining the dating of documents and material. However, half-life of carbon 14 is 5730 years (Ward et al. 2018), only materials

from 1000 to 5000 years ago can determine the dating through carbon 14. Although it is helpful for archeology, there are still many restrictions on its use for forensic dating. In this study, we aimed to find a method which can provide the same function as carbon 14 but determine the dating of the recent documents within 5 years.

There are two directions to research the dating of documents. One is ink dating which is a well-developed technique to classify the dating of documents. The time of document under specific ink could be detected by several techniques such as HPLC (Andrasko 2001; Liu et al. 2006), GC (Brazeau and Gaudreau 2007), MS (Weyermann et al. 2007), FTIR (Wang et al. 2001) and UV–vis (Senior et al. 2012; Xu et al. 2006), etc. Although the year of document and ink are almost the same, sometimes deviations still exist between both sides. Thus, date tracing by documents itself is very meaningful, and can address the problem of time source.

The other direction is the study of paper itself. Studies on paper degradation have been reported (Calcerrada and Garcia-Ruiz 2015). However, the process of document aging is rather complicated. Main ingredients in documents are not only cellulose, but also include many inorganic fillers to keep proper characteristics (Silva et al. 2018). Therefore, the analysis of documents dating has been studied in a limited range. Studies on cellulose degradation refer to the characterization of cellulose (Missori et al. 2006; Souguir et al. 2017), degradation products (Dupont et al. 2012) and document date. Zięba-Palus et al. (2017) distinguished the documents of different degrees of dating by infrared and Raman spectroscopy and obtained the degradation mechanism of documents by 2D maps. In the study of Ortiz-Herrero and Blanco, py-GC/MS was used for estimating the document age through two different approaches based on paper analysis: (1) a direct method using a PLS of the pyrolytic profiles of synthetic samples correlated with their aging time in chamber, and (2) an indirect method was used to identify the characteristic paper components during different time periods. And they concluded that 5 h in chamber under the experimented conditions were equivalent to one natural year under police custody conditions (Ortiz-Herrero et al. 2018). Laser induced fluorescence spectroscopy was employed to monitor documents fibers in function of their natural dating time (Martínez et al. 2017). FTIR combined with chemometrics was used to predict the dating of document, the root mean squared errors (RMSE) were around 4 years in model (Silva et al. 2018). After building a sPLS model, average spectrum focused on kaolinite, calcium, and carbonate (Silva et al. 2018).

In this study, we proposed a preliminary research to estimate the dating of documents by FTIR. sPLS–LS-SVM was used to acquire stable variables for modeling after 100 repeating calculations. Selected variables were investigated to reveal how they contributed to the dating of documents.

Materials and methods

ATR-FTIR spectroscopy

Both background and samples were measured by attenuated total reflection flourier transformed infrared (ATR-FTIR) spectrometer, in the range of 4000–650 cm^{-1} at a resolution of 2 cm^{-1} and each spectrum was average of 16 scans. The samples were pressed against the diamond crystal of the ATR device until a torque knob ensured that the pressure applied was the same for all measurements (Gomez-de Anda et al. 2012). The crystal was cleaned by anhydrous ethanol between successive measurements to avoid minute contamination. The cleaned crystal was checked by running a background spectrum. Spectrum recording was conducted by a Nicolet iS5 ATR-FTIR spectrometer (Thermo Scientific Co, Ltd., U.S.). Spectral data were collected by OMNIC 9.7.43 software.

Samples

According to the RMSE of 4 years found by Silva et al. (2018), the book samples in this study were collected with an interval of 5 years from 1940 to 1980. All the samples were found and measured in the basement of China Agricultural University Library. Because the paper of documents was very thin, seven sheets of paper were stacked as a sample. Triplicate spectra of different positions of each sample (top, middle and bottom) were collected, generating nine spectra every seven sheets. Different parts of each book (front, middle and back) were sampled. Namely, a total of 27 spectra were collected from each book.

From 1940 to 1980, five books were chosen for each 5-year interval. At last, a total number of 1215 spectra were obtained for later analysis. All the samples were divided into 810 training sets and 405 test sets after systematic sampling.

Algorithms and modeling

Principal component analysis

Principal component analysis (PCA) is a multivariate analysis method, which provides an unsupervised interpretation without prior assumptions on identifying of different samples (Liu et al. 2018). PCA aims to reduce the dimensionality of the original data space by using a smaller and more efficient abstract space of latent variables, where the data can be displayed and the information of the original space is essentially kept (Mees et al. 2018).

Sparse partial least squares classification analysis

The sPLS based on PLS was developed by Chung and Keles (2010). PLS is a well-known dimension reduction method, and sPLS employs a "sparsity" solution to achieve variable selection and dimension reduction simultaneously. In the process of sPLS, only non-zero variables that contribute to explaining the variability present in the response variable can be selected. Meanwhile, sPLS reduces the noise contained in the irrelevant variables and can be implemented to select variables (Abdel-Rahman et al. 2014).

Least squares support vector machines

LS-SVM is based on standard support vector machine (SVM), which was first proposed by Suykens et al. (2002), and is a widely applicable and functional machine-learning technique for classification and regression (Yang et al. 2018). When using SVM or LS-SVM, there are three crucial problems that need to be solved, namely, the determination of the optimal input feature subset, proper kernel function, and the best kernel parameters. In this study, LS-SVM, radial basis function (RBF) kernel, the optimal combination of gam(c) and sig2 (r2) parameters were chosen when resulting in smaller root mean square error of cross validation (RMSECV) (Camps-Valls 2011).

Model estimation

To characterize prediction ability (efficiency) of classification model, the parameters of non-assigned samples, the error rate and accuracy were used.

Non-assigned samples are some samples which can't be recognized or classified in the model.

Error rate is the ratio of wrongly assigned samples and is calculated as Eq. 1:

$$\text{Error rate} = 1 - \frac{\sum_{g=1}^{G} \frac{n_{gg}}{n_g}}{G} \quad (1)$$

where G represents the sum of classes, ng is the total number of samples belonging to the g-th class and n_{gg} is the number of samples belonging to class g and correctly assigned to class g.

Accuracy is the ratio of correctly assigned samples, and is calculated as Eq. 2

$$\text{Accuracy} = \frac{\sum_{g=1}^{G} n_{gg}}{n} \quad (2)$$

where n is the total number of samples. Non-assigned samples are not considered for the accuracy calculation.

Results and discussion

Spectral features

As spectral resolution was set as 2 cm^{-1}, each spectrum would generate a large number of spectra points with a pretty slow scanning process. In order to keep the resolution, but compress the amount of data as well as improve the running speed, an average of five points was merged. Since seven sheets of paper were stacked as a sample, the thickness of samples was quite different. Normalization was used in order to eliminate difference in optical path (Fig. 1).

As known, the main component in paper documents is cellulose (Liu et al. 2018). Besides, the inorganic compounds are essential to keep documents white and smooth. Among these inorganic compounds found in documents, calcium carbonate ($CaCO_3$) and kaolinite ($Si_2Al_2O_5(OH)_4$) are the most common ingredients (Zięba-Palus et al. 2017; Silva et al. 2018).

Before statistical analysis, spectra of documents samples from 4000 to 650 cm^{-1} were collected as

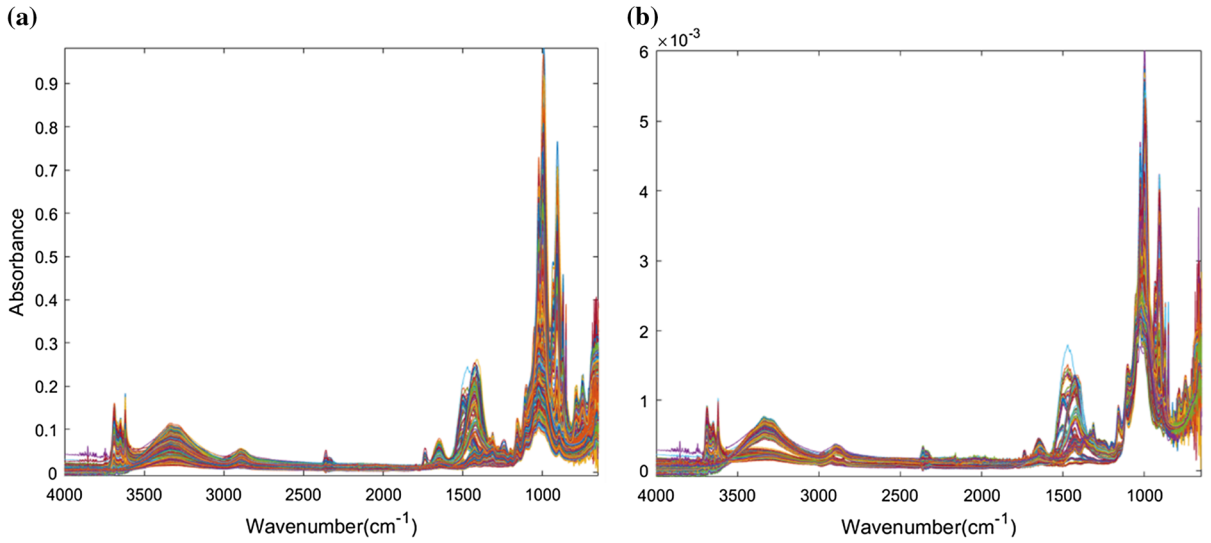


Fig. 1 Spectra with no preprocess (**a**); spectra after area normalization (**b**). The abscissa of both figures is wavenumber; the ordinate of the graph (**a**) is absorbance, the ordinate of the graph (**b**) is data which can be got after area normalization

shown in Fig. 2. There were two significant differences among spectra as black ellipses of Fig. 2. Thus, all spectra can be divided into two types based on the spectral differences. The representative samples in two types were chosen and assigned absorption peaks to chemical compounds (Fig. 3a, b).

The same parts in Fig. 3a, b are focused on the green part in Fig. 3a, HCH bending and wagging are observed in the 1200–1500 cm^{-1} range. Near the range of 1650 cm^{-1} are the absorption peaks of water. CH stretching in CH, CH2, CH3 are assigned in the 2800–2950 cm^{-1} range. OH–O intramolecular H

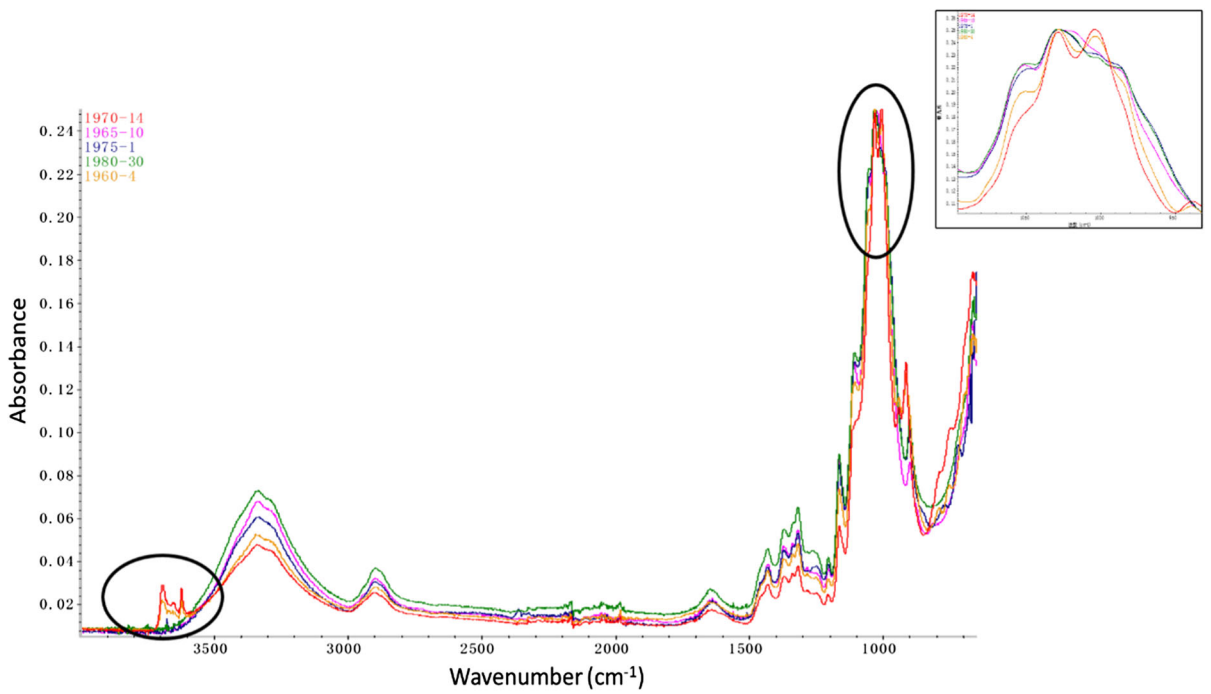


Fig. 2 Spectra of some samples. The black ellipses show the main differences in the absorption peaks

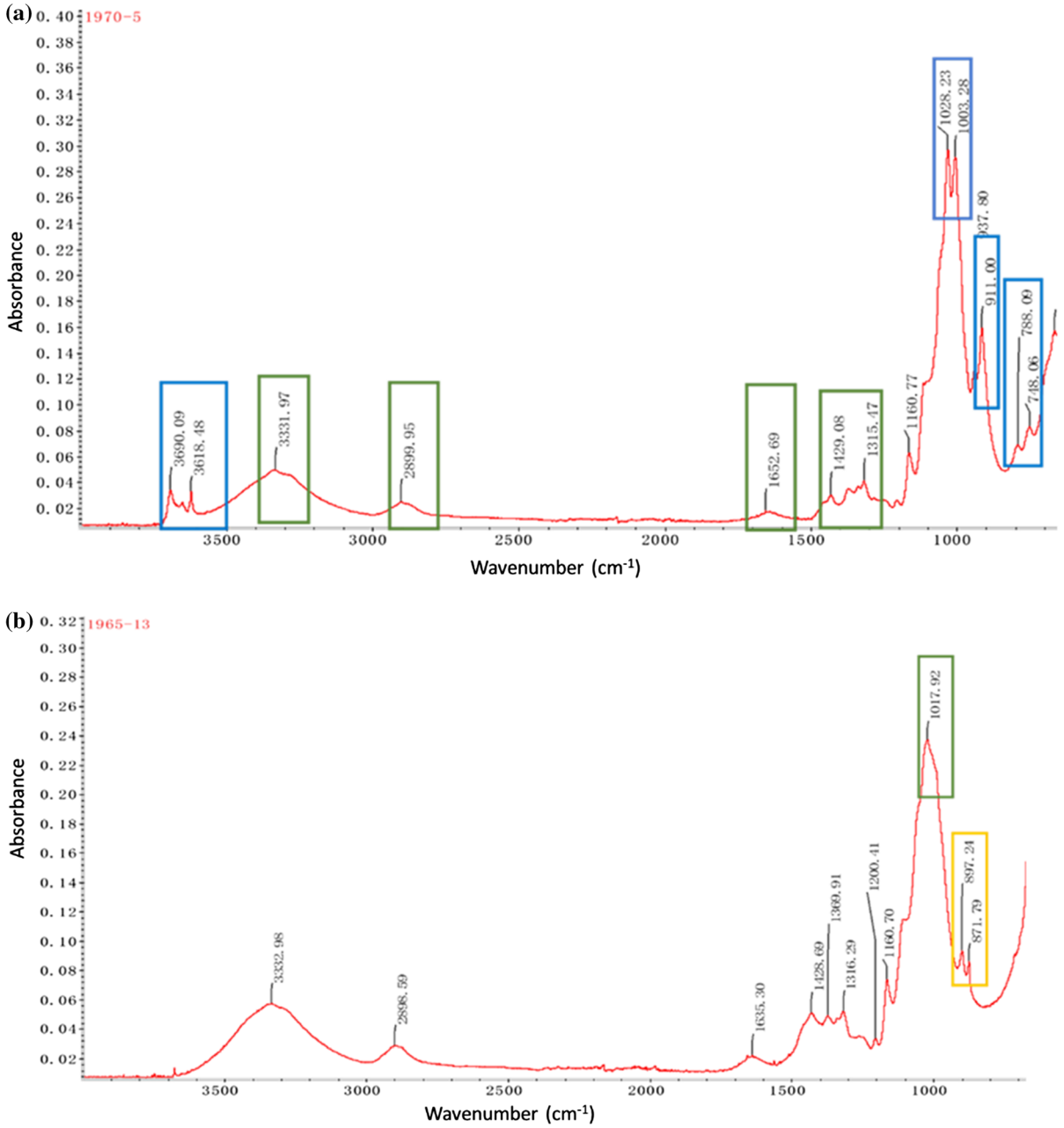


Fig. 3 The document spectrum with kaolinite (**a**); the document spectrum with carbonate (**b**). The absorption peaks enclosed by green rectangles belong to cellulose; the parts of blue are assigned to kaolinite; and the yellow rectangles represent carbonate

bond belongs to 3300–3500 cm^{-1} range. Based on some research, the absorption peaks enclosed by green rectangles belong to cellulose (Silva et al. 2018). The absorption peaks of 748, 788, 911, 1003, 1028, 3618 and 3690 cm^{-1} can be assigned to kaolinite (Silva et al. 2018) (The blue part of Fig. 3a). Carbonate contributions can be noticed not only at 875 cm^{-1} but also at 1410–1420 cm^{-1}. Because the range of 1410–1420 cm^{-1} overlaps with that at 1420–1430 cm^{-1} of cellulose (Martínez et al. 2017), no absorption peaks in 1410–1420 cm^{-1} range were observed, whereas peaks around 875 cm^{-1} were noted (see, the yellow part of Fig. 3b).

The analyze of principle components analysis

Principle components analysis (PCA) was used to classify the years of book. Figure 4a showed a two–dimensional diagram after PCA. The abscissa is PC1 which contains 82.00% of data information, and the ordinate is PC2 which contains 9.00% of data information. Figure 4b gave the three-dimensional diagram with PC3 which contains 5.00% of information. The samples from different years were grouped together. It is evident that it is impossible to distinguish all kinds of years only from Fig. 4.

The chemical information contained in each principal component (PC) is shown in Fig. 5. PC1 mainly includes the absorption bands of 907, 995, 1027, 1446, 3617 and 3691 cm^{-1}, which give mostly information

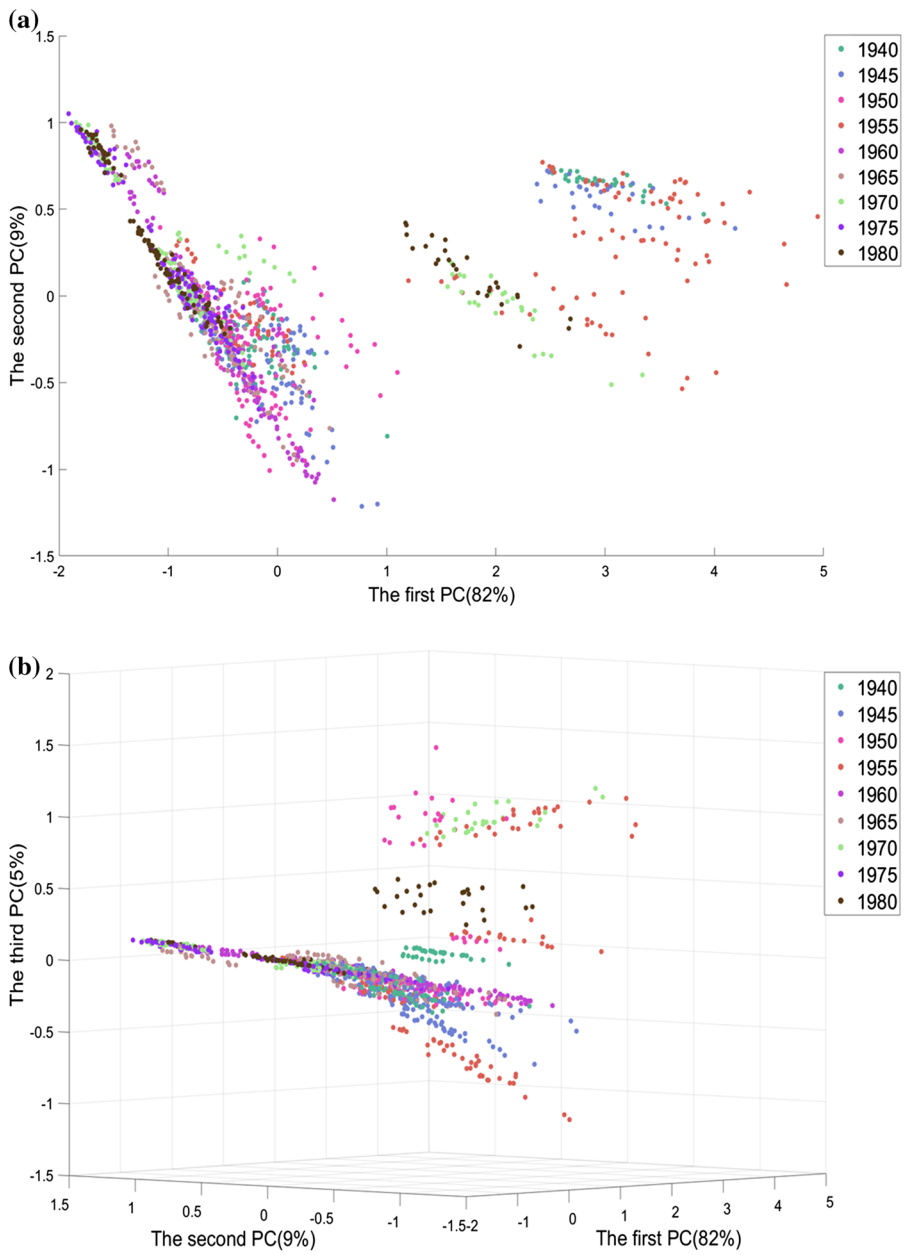


Fig. 4 Two-dimensional plots of PC scores (**a**); three-dimensional plots of scores (**b**)

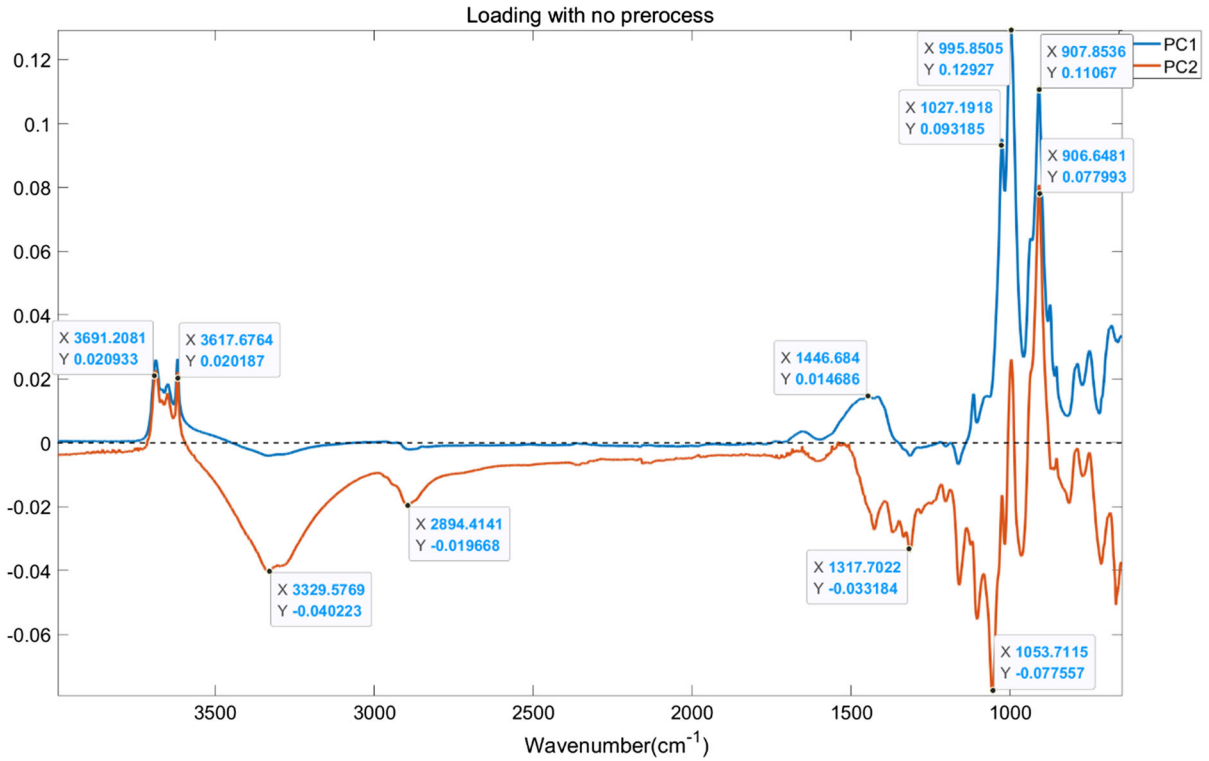


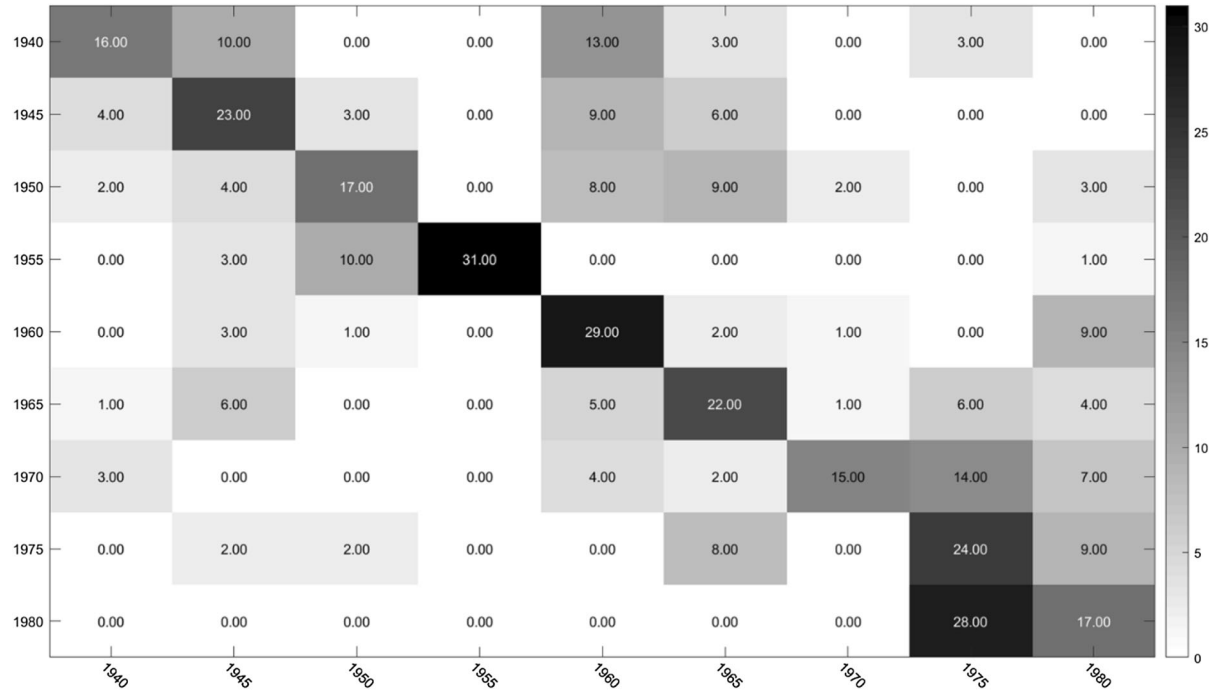
Fig. 5 Loadings with no preprocess. (PC1 gives the information of 82%, PC2 affords the information of 9%)

on cellulose and inorganic components. In PC2, absorption bands of 906, 3617 and 3691 cm^{-1}, give the positive contribution and are related to carbonate and kaolinite as we mentioned. But, some absorption peaks of cellulose, such as 1053, 1317, 2894 and 3329 cm^{-1}, give negative contribution in PC2.

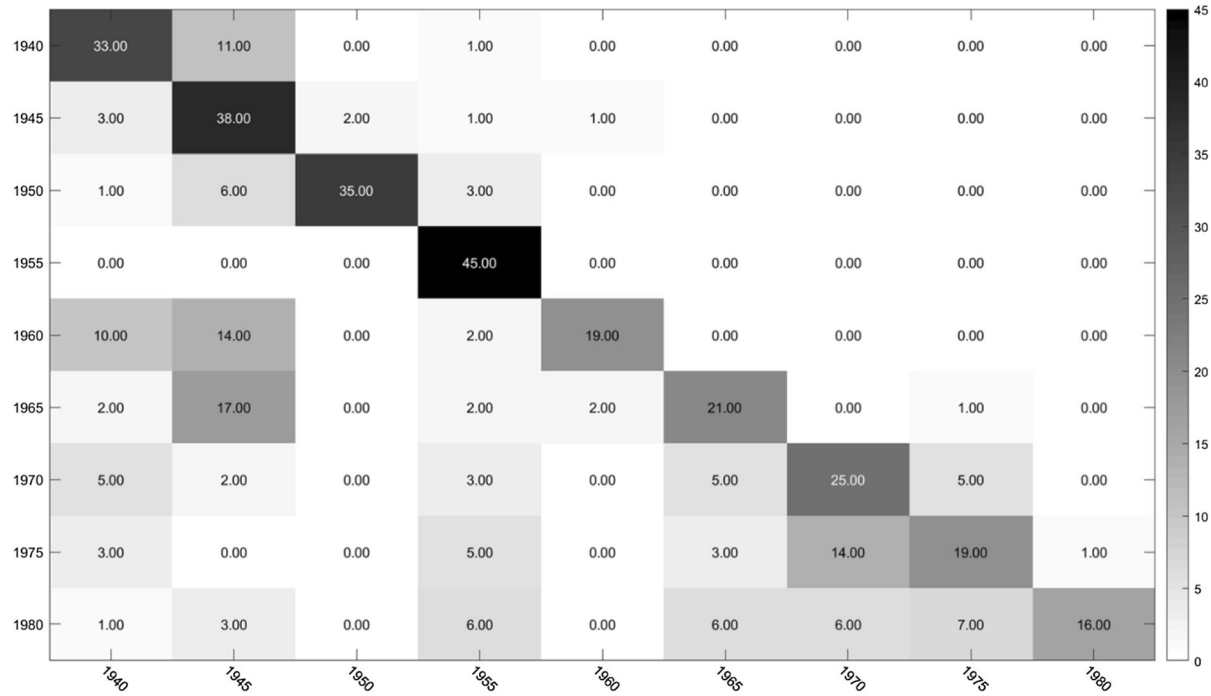
Modeling

Only PCA couldn't distinguish the year of the sample quite well, thus, some pattern recognitions were introduced to ascertain the accurate date of documents. Linear discriminant analysis (LDA), soft independent modeling of class analogy (SIMCA), and LS-SVM were used to establish models. The results were shown in Fig. 6. As shown in confusion matrix, the larger the number, the darker of the grid, wherein, only the correctly classified samples were on the diagonal while the misclassified samples fell in other areas. Figure 6a is the confusion matrix of LDA result, only nearly half of samples were correctly classified. The confusion matrix of SIMCA results (Fig. 6b), is better than LDA apparently. Thus, data were exhibiting non-linearity. When all variables from spectra were used to build LS-SVM model, only three samples were misclassified from confusion matrix of LS-SVM result (Fig. 6c). The classified accuracy of three algorithms was summarized in Fig. 6d. The accuracy of LDA was only 48.00%. Model built by SIMCA was with an accuracy of 62.00%. The best result was from LS-SVM, with the accuracy of 99.26%, the error rate of 0.74%.

The FTIR with LS-SVM is an effective way to identify the dating of documents. However, the specific spectral variables which contribute to the dating of documents were still unknown. Therefore, sPLS was used to extract the wavenumber variables that can decide the dating of documents. Especially, sPLS and LS-SVM were integrated together (sPLS–LS-SVM) for the excellent result of LS-SVM (99.26%). The process of sPLS–LS-SVM algorithm was operated as follows: sPLS was modeled with selected variables, and then LS-SVM was modelling on the selected variables. After sPLS, a total of 473 variables were selected as in Fig. 7a. The model was built by sPLS–LS-SVM with an accuracy of 99.01%,

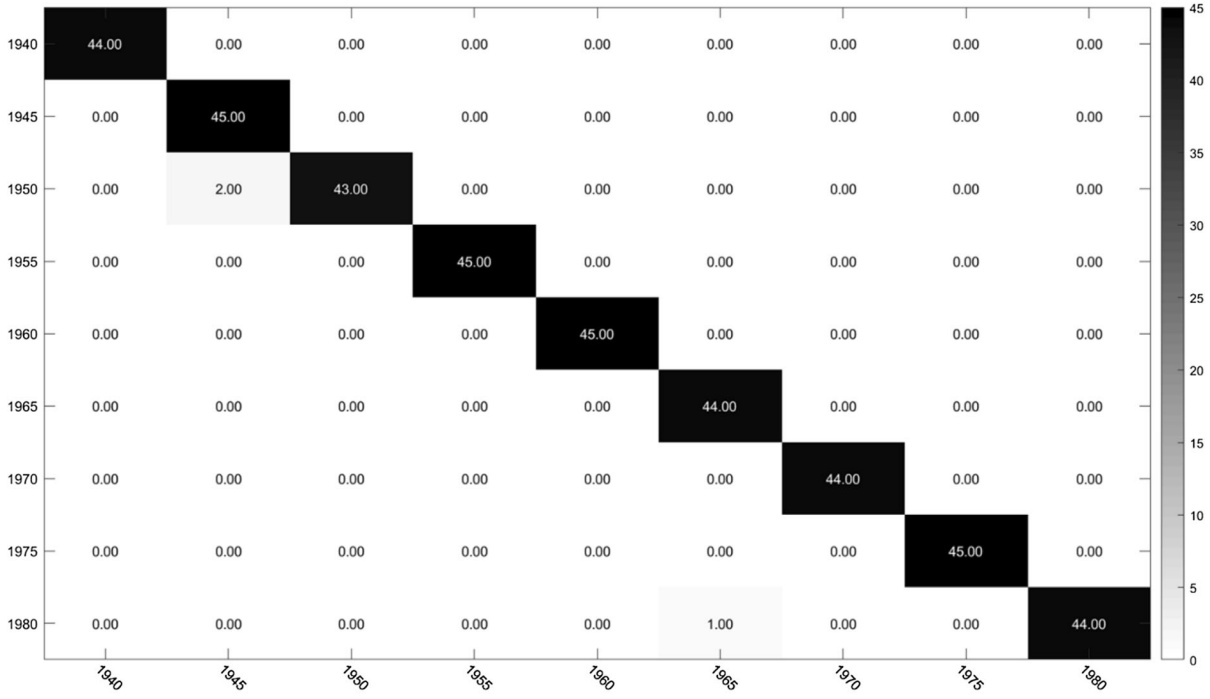


(a) The confusion matrix of LDA

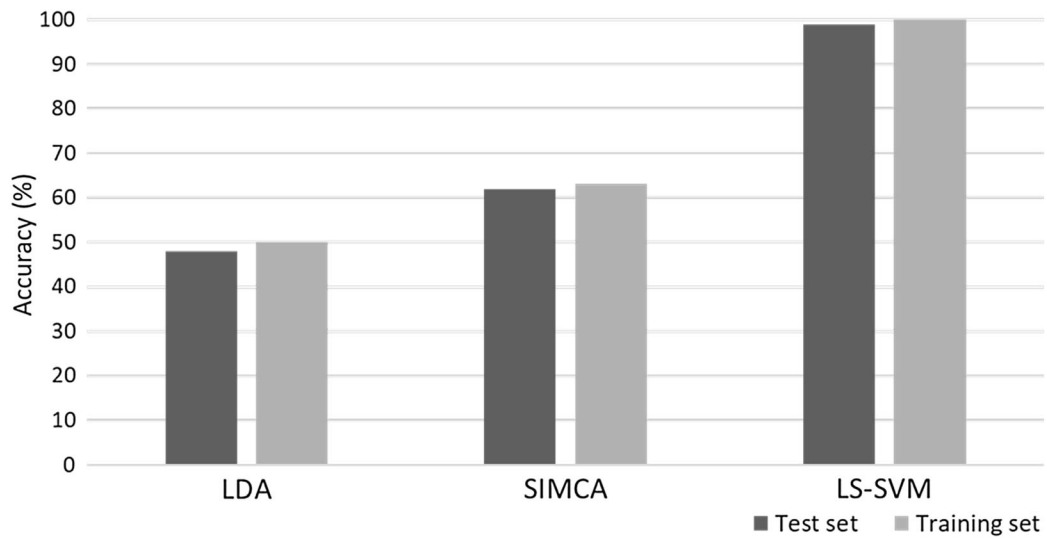


(b) The confusion matrix of SIMCA

Fig. 6 The confusion matrix of LDA result (**a**); the confusion matrix of SIMCA result (**b**); the confusion matrix of LS-SVM result (**c**); the classification results of three algorithms (**d**)



(c) The confusion matrix of LS-SVM

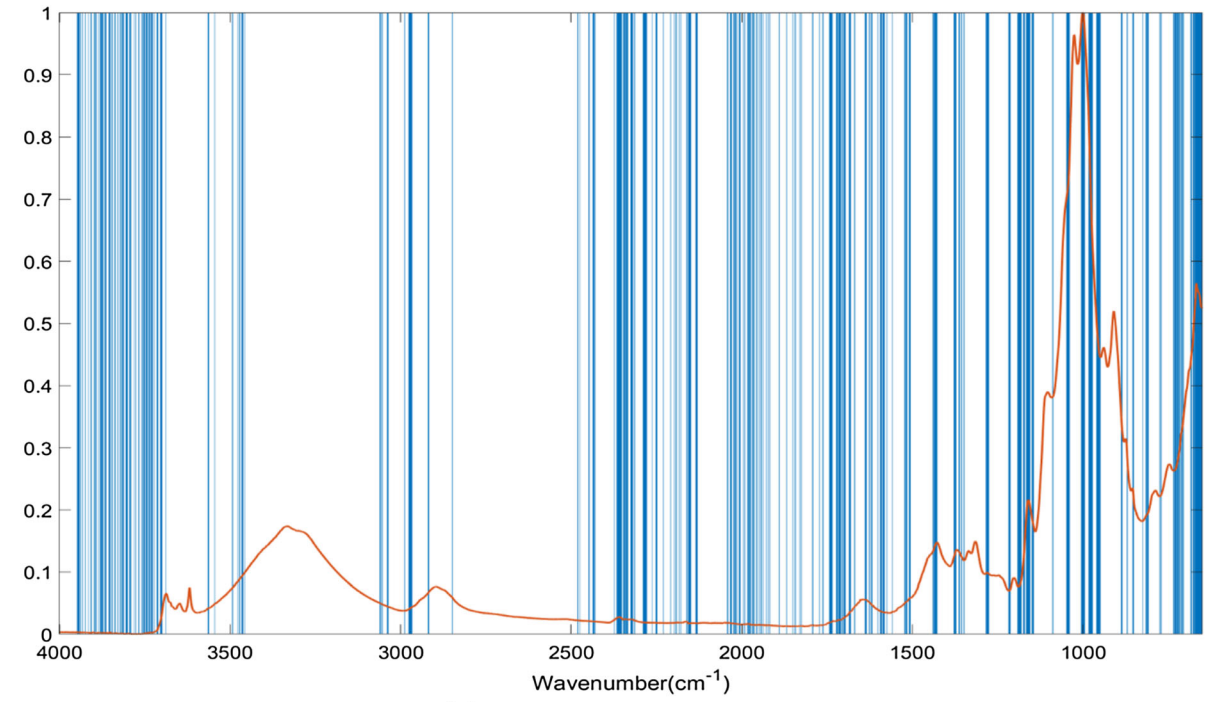


(d) Classification result of the three algorithms

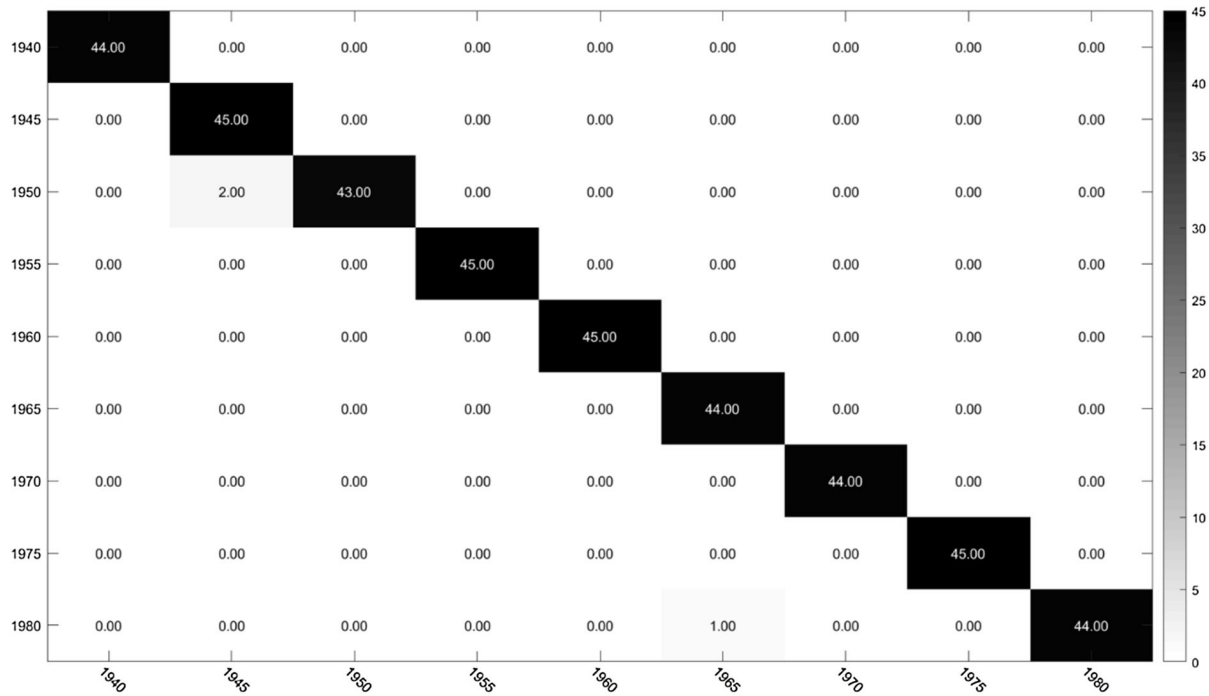
Fig. 6 continued

and the error rate of 0.99% as in Fig. 7b, only four samples were misclassified and six samples were not-assigned. The selected variables after sPLS give an outstanding effect to classify the dating of documents.

As known, only a pair of training set and test set cannot promise a convincing result, on the other hand, the selected variables would be different after every calculation. In order to acquire the stable variables, we renewed the training set and test set 100 times by Monte Carlo random sampling, and every time we executed sPLS–LS-SVM integrally. After 100 runs,



(a) The selected variables by sPLS



(b) The confusion result of LS-SVM result

Fig. 7 The selected variables of sPLS (**a**); the confusion matrix of sPLS–LS-SVM (**b**)

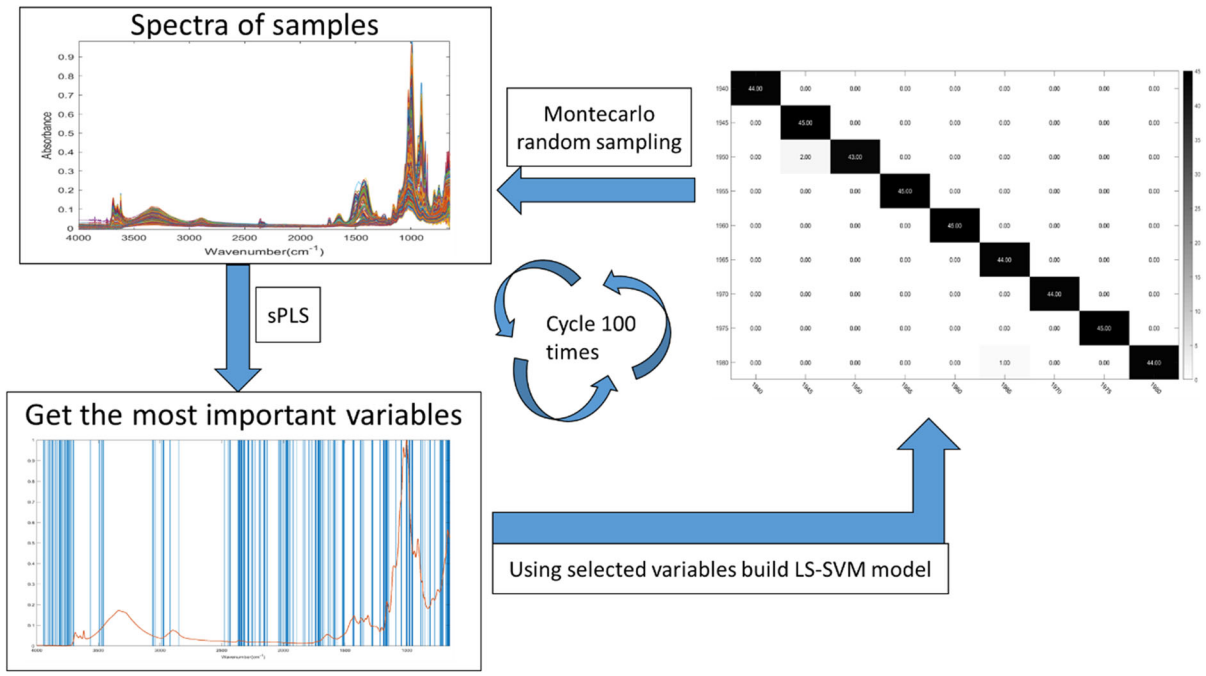


Fig. 8 The cycle process of sPLS–LS-SVM

the accuracy and group variables were obtained. The cycle process of sPLS–LS-SVM was shown in Fig. 8.

After 100 calculation runs, we got 100 accuracy. Among them, the minimum accuracy was 97.96%, the maximum accuracy was 100%; the average accuracy was 99.34% and the standard deviation was 0.43%. The model was stable because both accuracy and variance were acceptable.

The average number of 100 sets variables was around 483. A more important variable would match a higher frequency. The result of variables was shown in Fig. 9, with the wavenumber as abscissa and frequency as ordinate. The variables of 770–900, 940–1100, 1110–1820, 1800–2400, 2800–3100, and 3600–3900 cm^{-1} posed the high frequencies. Thus, these variables actually contributed more information in determining the date of documents. The variables of 770–900, 940–1100 and 3600–3900 cm^{-1} are the absorption peaks of kaolinite. The variables of 1100–1820 and 2800–3100 cm^{-1} are the absorption peaks of cellulose HCH (see Fig. 3). Both cellulose and inorganic components affect the dating of documents. However, the inorganic components are stable and hardly change over time, the reason for choosing inorganic components may be due to small differences between paper substrates. The variables are concentrated at 1500–2300 cm^{-1}, which are the absorption peaks of H_2O and CO_2. Because the interference of H_2O and CO_2 can't be avoided in the air, the influences were not investigated here.

Conclusion

Compared with the study of Silva et al., a different solving method was used. Silva et al. used RMSECV to evaluate the model, but in this study, accuracy was used. Both two methods gave as a result that FTIR combine with chemometrics can accurately distinguish the age of paper. In this study, the journal documents from 1940 to 1980 were chosen for analysis. Compared with other documents, the journal is closer to the age of the paper itself. And every journal has an average of two hundred sheets. Subsequently, the training set and test set were renewed 100 times by Monte Carlo random sampling. Therefore, the last selected variables are more statistically significant and more stable.

From the results, it was confirmed that FTIR spectroscopy combined with sPLS–LS-SVM can be an effective method to identify the dating of documents, with satisfied classification accuracy. By

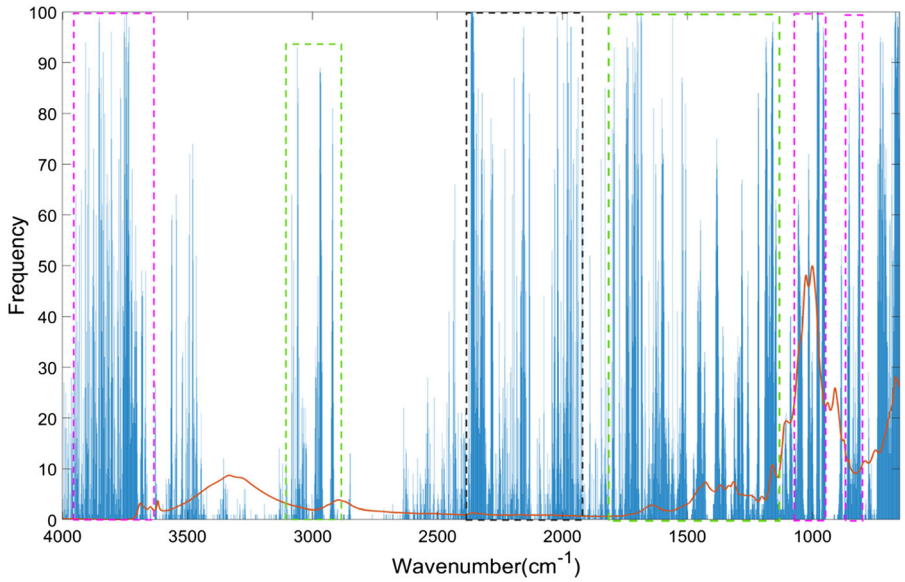


Fig. 9 The result of variable selection after 100 times

contrast, the accuracy of LDA, SIMCA and LS-SVM was 48.00%, 62.00% and 99.26%, respectively. The selected variables by sPLS were 483 after 100 times, and the selected variables were focused on the absorption peaks of inorganic components and cellulose. The average of sPLS–LS-SVM accuracy was 99.34% in 100 runs.

The accuracy (99.34%) is excellent when selecting samples in each 5-year interval, however, a more precise model should be developed with samples each year in further study. In practice, the dataset should be expanded with more years of documents.

Acknowledgments The authors would like to thank Yue Huang, Hong Yan and Qianqian Li for the critical review of the manuscript. The authors are also grateful to China Agricultural University Library for providing samples used in this work.

Compliance with ethical standards

Conflict of interest Authors declare that they have no conflict of interests.

References

Abdel-Rahman EM, Mutanga O, Odindi J, Adam E, Odindo A, Ismail R (2014) A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. Comput Electron Agric 106:11–19

Andrasko J (2001) HPLC analysis of ballpoint pen inks stored at different light conditions. J Forensic Sci 46:21–30

Brazeau L, Gaudreau M (2007) Ballpoint pen inks: the quantitative analysis of ink solvents on paper by solid-phase microextraction. J Forensic Sci 52:209–215

Calcerrada M, Garcia-Ruiz C (2015) Analysis of questioned documents: a review. Anal Chim Acta 853:143–166

Camps-Valls G (2011) Support vector machines in remote sensing: the tricks of the trade. In: SPIE remote sensing conference

Chung D, Keles S (2010) Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol. https://doi.org/10.2202/1544-6115.1492

Dupont AL, Seemann A, Lavedrine B (2012) Capillary electrophoresis with electrospray ionisation-mass spectrometry for the characterisation of degradation products in aged papers. Talanta 89:301–309

Gomez-de Anda F, Gallardo-Velazquez T, Osorio-Revilla G, Dorantes-Alvarez L, Calderon-Dominguez G, Nogueda-Torres B, de-la-Rosa-Arana JL (2012) 'Feasibility study for the detection of *Trichinella spiralis* in a murine model using mid-Fourier transform infrared spectroscopy (MID-FTIR) with attenuated total reflectance (ATR) and soft independent modelling of class analogies (SIMCA). Vet Parasitol 190:496–503

Liu YZ, Yu J, Xie MX, Liu Y, Han J, Jing TT (2006) Classification and dating of black gel pen ink by ion-pairing high-performance liquid chromatography. J Chromatogr A 1135:57–64

Liu YJ, Tran T, Postma G, Buydens LMC, Jansen J (2018) Estimating the number of components and detecting outliers using Angle Distribution of Loading Subspaces (ADLS) in PCA analysis. Anal Chim Acta 1020:17–29

Martínez JR, Nieto-Villena A, de la Cruz-Mendoza JA, Ortega-Zarzosa G, Guerrero AL (2017) Monitoring the natural aging degradation of paper by fluorescence. J. Cult Herit 26:22–27

Mees C, Souard F, Delporte C, Deconinck E, Stoffelen P, Stevigny C, Kauffmann JM, De Braekeleer K (2018) Identification of coffee leaves using FT-NIR spectroscopy and SIMCA. Talanta 177:4–11

Missori M, Mondelli C, De Spirito M, Castellano C, Bicchieri M, Schweins R, Arcovito G, Papi M, Castellano AC (2006) Modifications of the mesoscopic structure of cellulose in paper degradation. Phys Rev Lett 97:238001

Ortiz-Herrero L, Blanco ME, García-Ruiz C, Bartolomé L (2018) Direct and indirect approaches based on paper analysis by Py-GC/MS for estimating the age of documents. J Anal Appl Pyrolysis 131:9–16

Senior S, Hamed E, Masoud M, Shehata E (2012) Characterization and dating of blue ballpoint pen inks using principal component analysis of UV–Vis absorption spectra, IR spectroscopy, and HPTLC. J Forensic Sci 57:1087–1093

Silva CS, Pimentel MF, Amigo JM, Garcia-Ruiz C, Ortega-Ojeda F (2018) Chemometric approaches for document dating: handling paper variability. Anal Chim Acta 1031:28–37

Souguir Z, Dupont AL, De La Rie ER (2017) Formation of brown lines in paper: characterization of cellulose degradation at the wet-dry interface. Biomacromolecules 9:2546–2552

Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines, Chapter 3. World Scientific: 5 TohTuck Link, Singapore, pp 71–75

Wang J, Luo G, Sun S, Wang Z, Wang Y (2001) Systematic analysis of bulk blue ballpoint pen ink by FTIR spectrometry. J Forensic Sci 46:1093–1097

Ward JL, Snow MS, Olson JE, Ball D, Adamic ML (2018) Carbon-14 content in tree and soil samples at the Idaho National Laboratory nuclear site. Nucl Instrum Methods B 437:103–109

Weyermann C, Kirsch D, Costa Vera C, Spengler B (2007) A GC/MS study of the drying of ballpoint pen ink on paper. Forensic Sci Int 168:119–127

Xu Y, Wang J, Yao L (2006) Dating the writing age of black roller and gel inks by gas chromatography and UV–Vis spectrophotometer. Forensic Sci Int 162:140–143

Yang B, Shao Q, Pan L, Li W (2018) A study on regularized weighted least square support vector classifier. Pattern Recog Lett 108:48–55

Zięba-Palus J, Wesełucha-Birczyńska A, Trzcińska B, Kowalski R, Moskal P (2017) Analysis of degraded papers by infrared and Raman spectroscopy for forensic purposes. J Mol Struct 1140:154–162