Published in Talanta 82 (2010) 1784-1790

Use of genetic algorithms with multivariate regression for determination of gelatine in historic papers based on FT-IR and NIR spectral data

L. Cséfalvayová[1], M. Pelikan[2], I. Kralj Cigić[3], J. Kolar[4], M. Strlič[1]

[1]Centre for Sustainable Heritage, University College London, United Kingdom

[2]University of Missouri at St. Louis, Missouri Estimation of Distribution Algorithms Lab, USA

[3]University of Ljubljana, Faculty of Chemistry and Chemical Technology, Slovenia

[4]Morana RTD d.o.o., Ivančna Gorica, Slovenia

*correspondence: m.strlic@ucl.ac.uk

Abstract

Quantitative non-destructive analysis of individual constituents of historic rag paper is crucial for its effective preservation. In this work, we examine the potentials of mid- and near-infrared spectroscopy, however, in order to fully utilise the selectivity inherent to spectroscopic multivariate measurements, genetic algorithms were used to select spectral data derived from information-rich FT-IR or UV-VIS-NIR measurements to build multivariate calibration models based on partial least squares regression, relating spectra to gelatine content in paper. A selective but laborious chromatographic method for the quantification of hydroxyproline (HYP) has been developed to provide the reference data on gelatine content. We used 9-fluorenylmethyl chloroformate (FMOC) to derivatise HYP, which was subsequently determined using reverse-phase liquid chromatographic separation and fluorimetric detection. In this process, the sample is consumed, which is why the method can only be used as a reference method. The sampling flexibility afforded by small-size field-portable spectroscopic instrumentation combined

with chemometric data analysis, represents an attractive addition to existing analytical techniques for cultural heritage materials.

Keywords: gelatine, genetic algorithm, near-infrared spectroscopy, historic paper

1. Introduction

For centuries, paper has been the material of choice to record and communicate ideas and achievements. Not least due to its durability, this diverse material still remains the most universally used and understood information medium. With the aim to preserve the accumulated knowledge, characterisation and preservation of historic paper are areas of lively research.

In Europe, paper has been produced since the 13th century and until approximately the first half of the 19th century the technology remained largely unchanged. Paper sheets were produced from suspensions of fibres from old rags, and in order to make the paper surface less absorptive, sizing with gelatine was used [1]. Around ca. 1850, rosin sizing was introduced, and with it the era of acidic paper. Gelatine-sized papers were usually, but not exclusively, hand-made, of better quality fibres and of approximately neutral pH. Thus, they are more stable than rosin-sized papers, which are acidic. Due to its potentially beneficial role in paper degradation quantitative determination of gelatine is of high interest [2].

Gelatine is a denatured unfolded collagen, which is obtained by controlled hydrolysis of the fibrous insoluble collagen [3]. The amino acid composition of gelatine can be inferred from that of collagen. The secondary amino acid hydroxyproline (HYP), which is almost exclusively found in animal-derived collageneous proteins, is indispensable for the stabilization of the collagen superhelix because it increases the stability of the molecule by hydrogen bonding [4]. Due to its high gelatine specificity, the detection of HYP has long been used to estimate the amount, or presence, of gelatine in samples of

animal origin [5-7]. Gelatines used in papermaking varied in terms of their physical properties and amino acid content depending on the origin of protein, method of manufacture, and purity [8-10]. The main structural difference between fish and mammalian gelatines is the amino acid contents (proline and HYP), which is slightly higher in mammalian gelatines. However, gelatine determination methods based on HYP content are frequently in use and have even been standardised (TAPPI T-504 om-89 [11]), which is why our reference method was based on the same analytical principle, despite its drawbacks.

The standard colourimetric procedure is based on determination of HYP present in the extract of paper, from which the gelatine content is calculated. This procedure consists of oxidation of HYP with hydrogen peroxide followed by reaction with *p*-dimethylaminobenzaldehyde to form a magenta-coloured adduct. However, colour development is not reliable, which was attributed to difficulties with removal of the remaining peroxide [12, 13].

The detection and quantification of amino acids is routinely performed using high-performance liquid chromatography (HPLC) in combination with a variety of reagents and pre- or post-column derivatisation techniques [14-16]. The derivatising compound 9-fluorenylmethyl chloroformate (FMOC) rapidly reacts with primary and secondary amino acids and forms stable derivatives [17]. FMOC contains both chromophores and fluorophores, and can be determined using emission or absorption techniques [18].

The ageing behaviour of gelatine used in paper sizing was studied by Dupont who characterized aqueous extracts of paper by size-exclusion chromatography [2]. It was concluded that gelatine undergoes hydrolysis upon degradation resulting in an increase in the lower-molar-mass fractions. Stephens et al. recently quantified gelatine content in rag papers using seven amino acids [19]. However, since oxidation and hydrolysis occur during natural ageing of gelatine, no quantitative

method can provide an entirely accurate result: (i) if based on size exclusion chromatography, the results will depend on the hydrodynamic volume of the hydrolysed and oxidised macromolecules i.e. conformation of the degraded gelatine in the solvent; (ii) if based on amino acid analysis, the results will depend on their resistance against oxidative damage. Amino acids in collagen are prone to autoxidation, with proline, histidine, phenylalanine, lysine, and arginine being the least stable [20]. While HYP degrades, it is also formed from proline by way of oxidation of proline, and its content in gelatine degraded using various oxidative agents changed least, along with glycine, among all collagen amino acids. This justifies the use of HYP as a key amino acid for quantification of gelatine in historic paper. Recently, non-destructive approaches based on FT-(N)IR have been introduced to historic material characterisation [21, 22] finally allowing for in-situ measurements.

While Mid-IR encompasses the fundamental vibrational bands of molecules, NIR spectra are generated from overtone and combination bands of key fundamental absorptions, resulting in a high degree of coupling. Due to the co-linearity between spectral or other variables, simple univariate calibration techniques cannot be used for quantitative analyses of NIR spectra. Model-building techniques such as Principal Component Regression (PCR) or Partial Least Squares regression (PLS) based on the use of latent variables have a built-in capability to extract relevant information from a complete spectrum although there is much evidence that employing selected spectral variables can improve the accuracy of predictions [23]. Such selection of variables can be considered as an optimization problem. Methods specifically aimed at variable selection for multivariate calibration have been developed, such as Interactive Variable Selection, Uninformative Variable Elimination, Iterative Predictor Weighting PLS, Interval PLS, significance tests of model parameters, simulated annealing and Genetic Algorithms [24]. Genetic algorithms (GAs) are search methods based on principles of natural selection and genetics [25-27]. Several reviews of GAs and their application to chemistry-related problems have been published

[28-30].

In this contribution, we present optimisation and implementation of a GA consisting of steps including initialization of candidate solutions, evaluation, selection, recombination, crossover and mutation, and replacement, using operators inspired by natural evolution and genetics. In a GA, the collection of variables whose values are to be optimized is termed a chromosome, and the individual variables are called genes. A chromosome represents a candidate solution to the optimization problem. A set of chromosomes is termed a population. The initial population of candidate solutions is usually generated randomly across the search space. Once the population is initialized, the fitness value of each of the individual chromosomes in the population is evaluated. The chromosomes with the best fitness are selected to generate a new set of child chromosomes through the methods of crossover and mutation in order to introduce diversity into the child chromosomes while preserving the information carried by the parents. The new population formed replaces the original, and the chromosomes with the best fitness values in the new population are again selected to reproduce through recombination and mutation. This procedure is an iterative evolutionary process in search of the chromosome with the highest fitness value. The algorithm terminates after a fixed number of generations or when chromosome with a user-specified level of fitness is found.

GAs applied to PLS have been shown to be very efficient optimization procedures [31, 32]. However, GAs have significant drawbacks as they present a tremendous configuration challenge because of the numerous parameters required to run a GA that affect the outcome.

We also compare Fourier Transform IR and dispersive UV-VIS-NIR spectroscopy in relation to determination of gelatine content in naturally aged historic papers. The reference procedure was developed and GA was employed for selection of spectral variables for construction of multivariate calibration models.

2. Experimental

2.1. Samples

We assembled a set of 79 papers dating from the 17th and the 19th century. The samples for analysis were taken from areas not printed, inked or discoloured in order not to compromise the analyses.

2.2. Reagents

The amino acids 4-hydroxyproline (HYP), sarcosine (≥99%, Fluka, Buchs, Switzerland), 9-fluorenylmethylchloroformate (FMOC, Sigma-Aldrich), HCl (37%, reagent grade, Scharlau Chemie, Spain), Na_2HPO_4 (>99%, Ultra – for molecular biology, Fluka, Buchs, Switzerland), KH_2PO_4 (p. a., Kemika, Zagreb, Croatia), methanol (gradient, 99.99%, HPLC grade, Scharlau Chemie S. A., Spain), acetonitrile (supergradient, 99.99%, HPLC grade, Scharlau Chemie S. A., Spain), gelatine (obtained from craft workshop) were used. All aqueous solutions were prepared from deionized water purified by a Milli-Q system (Millipore, Molsheim, France). The borate buffer solution was prepared by adding 305.0 mg of H_3BO_3 (p.a., Kemika, Zagreb, Croatia), 327.7 mg of KCl (p.a., Merck, Darmstadt, Germany) and 200.0 mg of NaOH (synthesis grade, Scharlau Chemie S. A., Spain) to 50 ml of deionized water. The pH of the borate buffer solution was adjusted to 9.9 by adding 4% HCl. The FMOC reagent was prepared by dissolving 25.0 mg of FMOC in 5 ml of acetonitrile. A series of aqueous standard solutions of sarcosine and HYP were prepared. The phosphate mobile phase was prepared by mixing the two stock solutions of 40 mM Na_2HPO_4 and 40 mM KH_2PO_4 to reach pH 7.8.

2.3. Chromatographic determination

The procedure consisted of gelatine extraction from paper, gelatine hydrolysis to obtain free amino

acids, and their subsequent derivatization and high performance liquid chromatographic determination. An Agilent 1100 Series HPLC system (Paolo Alto, USA) was used, equipped with a quartenary pump, degasser, autosampler, column thermostat set at 40 ˚C and a diode-array detector. Using the paper sample size of approximately 5 mg, gelatine was extracted with 1 ml 0.1 M HCl at 100 ^{o}C for 1 h. An aliquot of extract (0.9 ml) was hydrolyzed with 0.9 ml 12 M HCl, and maintained at 100 ^{o}C for 18 h. The excess HCl was removed by evaporation of the contents to dryness in an oil bath (110 ^{o}C). The hydrolyzate residue was re-dissolved in 0.5 ml of the phosphate buffer eluent and filtered before injection. Derivatization of amino acids was performed by reacting with 9-fluorenylmethyl chloroformate (FMOC). Experimental conditions were applied according to the existing protocol [33]. The amount of gelatine was calculated assuming the fraction by weight of HYP in a typical hide glue (0.126), the moderately pure gelatine similar to that used for paper sizing. The gelatine content is expressed as weight percentage (%(m/m)) on dry paper basis.

Standard error of the laboratory (SEL) reference method was calculated as

$$SEL = \sqrt{\frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\left(y_{ij} - \bar{y}_i\right)^2}{N(r-1)}},$$

where y_{ij} is the *j*th replicate of the *i*th sample, $\bar{y}_i$ is the mean value of all the replicates of the *i*th sample, N is the total number of samples, and r is the number of replicate analyses for each sample.

2.4. Spectroscopic analysis

A Perkin-Elmer Spectrum GX (Waltham, MA) equipped with a 76-mm Labsphere RSA-PE-200-ID (North Sutton, NH) integration sphere coated with Infragold, with a DTGS detector, was used to collect FT-IR reflectance spectra in the interval of 6500-500 cm^{-1}, at spectral resolution of 8 cm^{-1} resolution,

128 scans per sample with 4 layers of each paper sample.

Near-infrared reflectance spectra were recorded using a LabSpec 5000 spectrometer (Analytical Spectral Devices, USA) configured with three separate holographic diffraction gratings with three separate detectors with 512 element silicon photo-diode array for the spectral region 350-1000 nm, and two TE-cooled InGaAs for spectral regions 1000-1800 nm and 1800-2500 nm. The spectrometer was fitted with a Source Probe MugLite that features a tungsten quartz halogen light source with built-in DC current stabilizer circuitry, and a sapphire window. An FR Fiberoptic Jumper 1 m cable was required for its interfacing with the LabSpec instrument. UV-VIS-NIR spectra were measured over the range 350-2500 nm, using 50 scans. A Spectralon standard was used as a reflector for the background reference.

Each sample was analysed twice and spectra for the two runs were averaged. Thus the area of the sample scanned was increased and the sampling error reduced. All the measurements were performed under similar temperature and relative humidity conditions (23±2 ^{o}C; R.H. 25±5%).

2.5. Data analysis

The GA used in this study was inspired by the approach of Bernardo et al. [34]. Each candidate solution is represented by a subset of frequencies of size *k* where *k* is provided on input. The initial population of $n = 200$ subsets is generated at random according to the uniform distribution over all subsets of size *k*.

Ten-fold cross-validation is used to evaluate the quality of each subset in the initial population. The spectra are first divided into 10 approximately equally sized groups where each group corresponds to a range of values of gelatine content. The subset of frequencies is tested on each of these 10 groups of spectra as follows. First, a linear partial least-square model is built from the remaining 9 groups of

spectra. Then, the sum of square differences between the created model and the empirical data is computed for the group of spectra used for testing. In this manner, the model is always tested on different data than those used for learning and each group of spectra is used for testing. The average error from the 10 tests (using each of the 10 groups of spectra for testing) is then computed and the quality of the solution is estimated as the negative value of this error. The GA maximizes solution quality (minimizes average error).

Each GA iteration starts by creating a new population of subsets by applying crossover and mutation operators to the current population of subsets. First, the subsets in the current population are randomly divided into pairs of subsets and crossover is applied to each of the pairs of subsets with a given probability p_c (we used $p_c = 0.6$ in all experiments). The crossover first randomly reorders one solution. Then, it iterates through the elements in the two subsets and with probability 50% it exchanges elements in any position of the two subsets unless the exchange would make any of the two subsets invalid due to repeated occurrence of some elements. After applying the crossover, each subset undergoes mutation. Mutation iterates through the subset and modifies each element of the subset with probability p_m. Two types of modifications are used in this work: (1) Change to a randomly selected frequency according to the uniform distribution over all frequencies (with probability p_{mr}), and (2) change to a frequency selected at random from the entire population of subsets (with probability 1 - p_{mr}). For FT-IR spectra, we used $p_m = 0.01$ and $p_{mr} = 0.05$; for UV-VIS-NIR spectra, we used $p_m = 0.05$, and $p_{mr} = 0.3$. These settings were obtained through extensive empirical testing of the GA implementation, but comparable results were obtained for a wide range of parameter settings. Only mutations that lead to valid subsets (with no repeated elements) are allowed.

Once the new solutions have been created, they are evaluated using the 10-fold cross-validation and the best n solutions from the previous population and the population of new solutions are selected to form

the population in the next iteration. The procedure is repeated (starting with crossover and mutation) unless user-defined termination criteria are met. For example, the run may be terminated when a maximum number of iterations has been reached. The best solution obtained during the entire run is hreported. Other GA variants have been tested but the aforementioned implementation led to the best results, especially when the GA was allowed to run for a large number of iterations (1,000-10,000).

For a detailed description of partial least squares (PLS) regression refer to [35].

The root mean squared error of prediction by cross-validation (RMSECV) was calculated

$$RMSECV = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}},$$

where y_i is the measured value and $\hat{y}_i$ is the leave-out prediction of gelatine content. The coefficient of determination (R^2) for modelled versus measured gelatine content, which indicates the percentage of the variance in the *y* (spectral data) that is accounted by the *x* (reference data), was also considered.

The genetic algorithm was implemented in C and for PLS modelling the software package PLSplus/IQ v. 8.0 (Thermo Scientific) was used.

3. Results and Discussion

3.1. HPLC Gelatine Assay

In order to obtain quantitative information on gelatine present in paper, based on FT-IR or NIR spectra, calibration must first be carried out using a reference method. In our study we modified the TAPPI standard method of determination of the content of HYP in paper as the basis for calculation of the amount of gelatine present [11]. In our approach, we used high-performance liquid chromatography instead of the colourimetric method in order to obtain more accurate data.

We added the amino acid sarcosine to the sample extract prior to hydrolysis in order to correct for

variations in sample preparation and chromatographic analysis and a possible loss of recovery. Using this internal standard method the reproducibility of quantitative determination of HYP in paper samples was improved as demonstrated by the average relative standard deviation (RSD) of 8.3% of the replicate determinations (as opposed to the average RSD 14.9% without the internal standard).

In order to obtain the value of error of the reference laboratory method, six replicate determinations for five samples with different gelatine contents were carried out. The established SEL was 0.2%, which is acceptable given the range of values from 0.2 to 9.7% in the sample set. Gelatine content was calculated according to TAPPI [11], in which HYP content of hide glue is given as 12.6 ±0.2%. While degradation of gelatine in paper and consequential changes in amino acid composition have not been fully elucidated yet, Dupont showed that upon heat/humid aging gelatine undergoes degradation [2] and hydrolysis and crosslinking were observed. The different sensitivities of amino acids in peptides to partial hydrolysis were suggested to be involved in preferential cleavage of polypeptide at specific points. However, as outlined in the Introduction, amino acids show different degrees of susceptibility towards oxidation, while HYP and glycin exhibit the highest stability [20]. To confirm this, we performed a study of the effect of heat/humid ageing on HYP content of gelatine. After subjecting a sample to accelerated degradation at 80 ^{o}C for 30 days, we analysed it for HYP content. The results indicated no statistically significant degradation-related loss of HYP : the contents in the original and degraded samples were essentially the same: 12.0% and 11.9%, respectively.

On the other hand, carbohydrates present in naturally aged papers may interfere with amino acid analysis due to condensation reactions between aldehyde functional groups in saccharides and amino acids during hydrolysis. This would lead to poor recoveries of amino acids and poor reproducibility. However, since the results of HYP determination showed good reproducibility, the results suggest that the interference by carbohydrates was insignificant. Further investigations on the amino acid

composition of gelatine from each historic paper were beyond the scope of this study.

3.2. FTIR analysis

Paper consists of cellulosic fibres, other organic components (gelatine, starch) and inorganic components (fillers, additives). Degraded naturally aged papers further contain a number of degradation products, which build up depending on the paper composition (especially its acidity), and on the environmental conditions (presence of oxygen, light, humidity). Despite the difficult interpretation of spectra, such changes should theoretically be detectable in the IR spectral region. To obtain the FT-IR spectra, we used an integrating reflectance accessory since the samples are not homogeneous in composition and thickness. Integrating spheres for the mid-infrared have become commonplace and have historically been used for quantitative measurements of reflectance and absorbance of samples and materials that exhibit a high degree of scattering [37]. However, such studies are of limited use due to physical limitations when analyzing whole documents or books. Additionally, the interpretation of mid-infrared spectra of complex mixtures may become extremely difficult. In Fig.1, we provide spectra of a historic rag paper with gelatine and of pure cellulose. The visible differences (highlighted) are minimal despite the different fibres, manufacturing technologies, presence of additives and sizing. The vibrations in the peptide bonds originate in bands known as amide A, B, and amide I, II, III. Bands related to proteins below 1400 cm^{-1} could be assigned to amide III, bands in the area 1700 cm^{-1}-1500 cm^{-1} to amide I and II and bands above 3000 cm^{-1} to amides A and B. The relatively broad complex envelope at 1700 cm^{-1}-1500 cm^{-1}, highlighted in Fig. 1, covers the amide I band, connected mainly with the C=O stretching vibration, and the amide II band connected with N-H in-plane bending vibration. The region between 4500 and 4000 cm^{-1} provides information about the nature of side chains in protein. A shoulder at 5000 cm^{-1} was reported as resulting from side

chain amide groups.

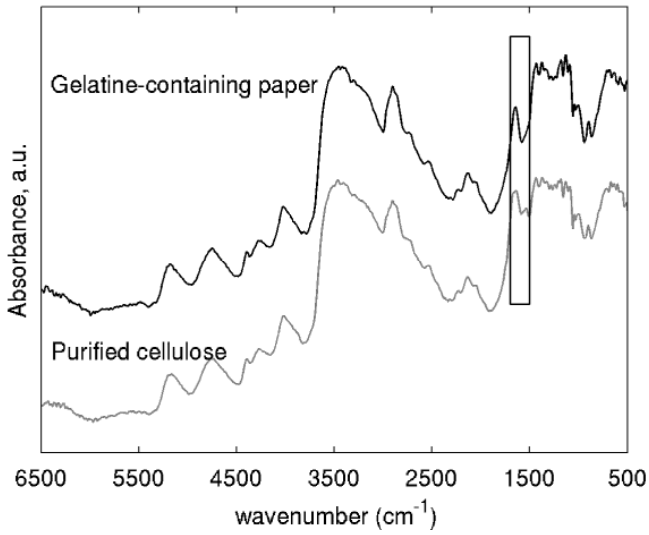

Fig.1 FT-IR spectra of gelatine-containing historic paper and of purified cellulose. The visually evident differences (highlighted) are minute. The spectra have been offset for clarity.

We first considered PLS calibration model developed from full-spectra (750 variables) without any variable selection. The PLS model constructed required 10 latent variables corresponding to the optimum cross-validation error of 0.6% of gelatine content in the range of 0.2 to 9.7%. It is likely that the potential information on gelatine is related to the response variables in a complex manner, thus requiring a more complex model. This complexity can also be related to the variability of the samples. The plot of modelled versus measured gelatine content values is presented in the Fig. 2, which can be considered as an encouraging result.

It should be stressed that even if the PLS model built on the whole spectrum is able to provide fairly good predictions, the PLS latent variables calculated may also be affected by redundancies or the presence of irrelevant variables. Consequently, an additional improvement in gelatine determination using FT-IR spectroscopy may be gained by the selection of individual wavenumbers to be included in the calibration process, as opposed to using full-spectrum techniques. In order to do so, an

approach based on a genetic algorithm was used to select the wavenumbers which contribute most to the correlation. The pre-selection of wavenumbers is not based on a detailed interpretation of the spectral data. The application of a variable selection approach is likely to change the structure and/or the order of the latent variables of the PLS model.

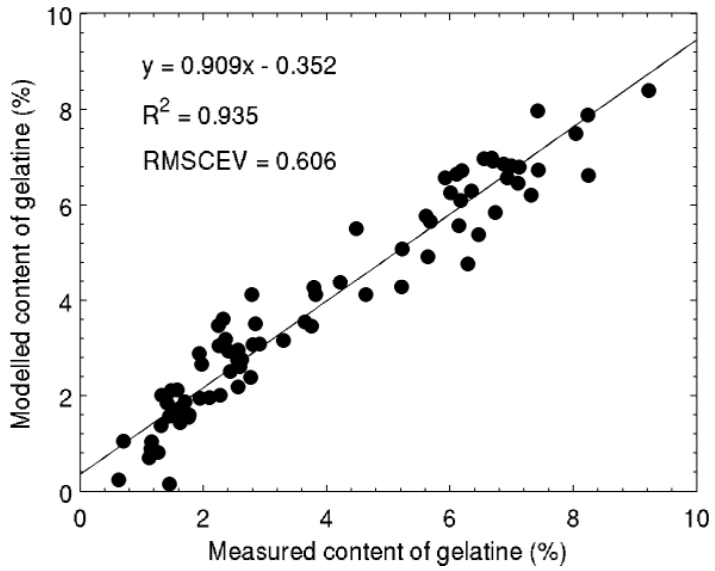

Fig.2 PLS model for estimation of gelatine content of historic paper, developed from full FT-IR spectra without variable selection.

A search was performed according to the procedure described in the experimental section and only 137 variables were retained as the optimal combination, which represents 18% of the full spectrum. An overview of the selected wavenumbers is given in Fig. 3. It can be noticed that the chosen variables are distributed throughout the whole spectral domain. Along with the spectral points relevant to the absorption of gelatine, some wavenumbers outside the analyte absorption bands could have been selected. These points may be of use by the model in performing baseline correction, or they may simply have been carried along through the recombination process. The number of latent variables employed would seem to depend on which wavenumbers have been selected, and the model obtained by the GA was based on three fewer latent variables. More importantly, the use of a variable selection

based on the genetic algorithm significantly improved the prediction results (corresponding to RMSECV = 0.2%) as compared to the model obtained on the complete spectrum.

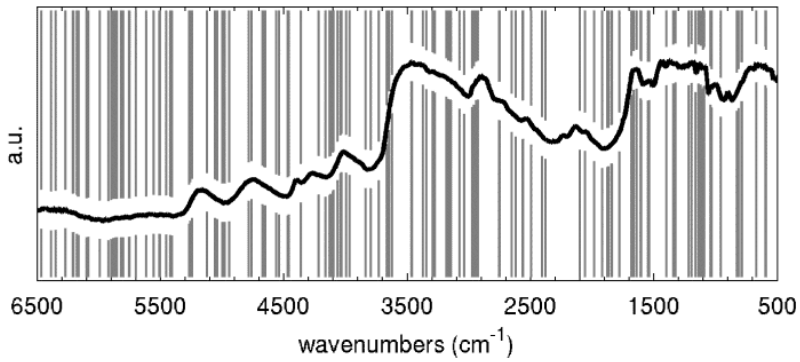

Fig.3 Overview of the spectral variables selected by GA. The FT-IR spectrum of the historic paper is overdrawn.

Evidently, the values modelled on the basis of FT-IR spectra using only the wavenumbers selected by GA correlate very well with the values determined by the reference chromatographic method (Fig. 4).

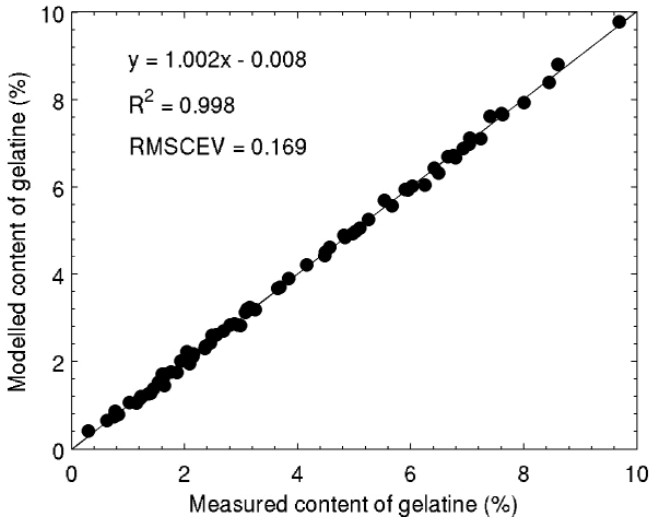

Fig. 4 Modelled vs. measured content of gelatine; the calculated model is based on the selected input wavenumbers by GA from FT-IR spectra and the partial least squares regression.

3.3. NIR analysis

While the FT-IR-based correlation is satisfactory, the technique itself is perhaps less suitable for heritage objects for practical reasons – for heritage applications, the instruments should ideally be small, portable and preferentially equipped with fibre optics to allow for maximum versatility. At present, NIR spectroscopy seems to fulfil these demands and we therefore chose to compare it with the more traditional FT-IR approach.

While an NIR spectrum could potentially be interpreted in a manner analogous to the interpretation of mid-IR spectra, much less is understood about the assignment of the observed bands due to overlapping of overtones of various orders and due to combinations of vibrations. For this reason, NIR spectroscopy has not been widely used in material characterisation yet. Vibrations involving bonds between the light atoms (e.g. CH, OH, NH) exhibit their first order overtones in the near-IR region. These are usually much stronger than the second, third, and fourth overtones. NIR absorption bands are typically 10-100 times weaker than their corresponding mid-IR absorption bands, which limits the sensitivity. The low NIR absorption coefficients are also an analytical advantage, since they allow the radiation to penetrate deeply through the inhomogeneous material and hence the direct analysis of samples in reflectance mode without further pre-treatments to be accomplished. A few studies exist on proteins with detailed band assignments [38].

The UV-VIS-NIR spectra for the gelatine containing historic paper along with that of purified cellulose are shown in Fig. 5. The spectrum of the historic paper containing gelatine is visually identical to the spectrum of pure cellulose, dominated by the first overtone O-H stretch centred at 1490 nm and the combination O-H centred around 1930 nm. The position of NIR bands can be more easily visualized in the processed spectra. The derivative spectrum of the gelatine containing paper reveals more features that can also be recognized in the region 2100 to 2400 nm. Proteins can be determined in this portion of

the NIR region by taking advantage of the associated functional groups, such as amides and various C-H functional groups.

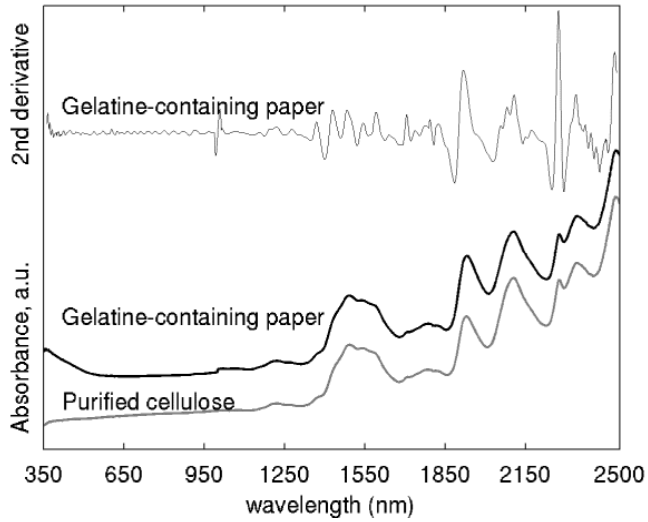

Fig. 5 UV-VIS-NIR spectra of purified cellulose, gelatine-containing paper and its second derivative spectrum using the Savitzky and Golay method with third order polynomial and 19 convolution points. In the process, we multiplied the derivative spectrum by -1. The spectra have been offset for clarity.

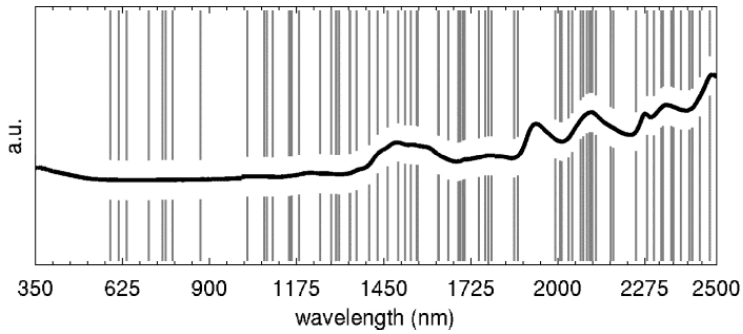

Fig. 6 Overview of the selected spectral variables by GA. The UV-VIS-NIR spectrum of the historic paper is overdrawn.

By applying the partial least squares regression analysis to the UV-VIS-NIR spectra using only selected wavelengths, highly satisfactory predictions of gelatine content were obtained with a 0.2% RMSECV prediction error (Fig. 7). The model obtained by the GA was based on a significantly fewer spectral points (76 vs. 2150) and five fewer latent variables (4 vs. 9 latent variables) than that based on full

spectra (results not shown). Fig. 6 presents an overview of the subsets of variables selected in the GA procedure. The selected wavelengths occur near the absorption bands of proteins, but also where overlapping with another major component peak such as water is minimal. Here also, the final model produced may have selected variables that are outside of the absorption bands of the analyte and interferences.

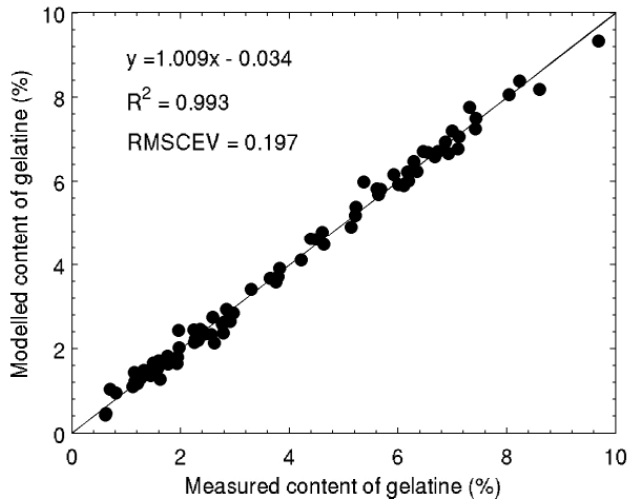

Fig. 7 Modelled versus measured content of gelatine; the calculated model is based on the selected input wavelengths by GA, from UV-VIS-NIR spectra, and the partial least squares regression.

The information on gelatine content in NIR spectral data is held in a range of wavelengths. Generally, in the region from 2100 to 2400 nm bands related to proteins located below 2200 nm are due to N-H stretching and bending modes, while C-H stretching and bending modes give rise to absorptions above 2200 nm [39]. Judging from the wavelengths selected by GA (Fig. 6), key bands associated with gelatine are from 1750 to 1840 nm and from 2000 nm to 2200 nm, representing the carbonyl stretch of the primary amide, combination band consisting of N-H bend second overtone; the C-H stretch/C=O stretch combination; and the C=O stretch/N-H in-plane bend/C-N stretch combination bands [40]. Other important wavelengths related to gelatine content include those at 1650 nm and 1685 nm

indicating the aromatic C-H stretch first overtone, which could theoretically represent aromatic containing amino acids such as HYP. The bands at 1500-1600 nm indicating N-H stretching vibrations are also useful for gelatine determination. Other intervals of selected wavelengths appear to be in the region from 2200 nm to 2450 nm, sensitive to the side chains of proteins, and the minor ones exist around 800 nm and 1100 nm. It is interesting to note that one of the major wavelength intervals falls within the first water peak spectral envelope 1400 to 1500 nm, while wavelengths of the second water peak 1900 to 2000 nm clearly should not be selected, which might relate to different structure of water associated with the fibres. This also indicates that the determination of proteins in paper using NIR will probably be influenced by water content in paper, which is the area of our future research.

4. Conclusions

Determination of gelatine in historic degraded papers is an interesting analytical problem, not least because the analyte is an ill-defined natural material and potentially degraded due to natural ageing. Because of this, any analytical method will be burdened with systematic errors, which are difficult to overcome because of the lacking reference. In this work, we presented:

- Development of a quantitative method based on chromatographic determination of HYP which has been shown to be among the most stable amino acids in collagen, and determined gelatine content in 79 historic paper samples. However, this method is not suitable for original documents, as the sample is consumed during analysis.
- Development of a new, non-destructive analytical method for determination of gelatine in historic papers based on the chemometric approach and on reference data obtained using the chromatographic method.
- Comparison of Fourier-transform IR and dispersive NIR spectroscopy. Despite the complexity

of spectral data of degraded paper samples, the implemented GA procedure for wavelength selection and partial least squares regression for quantitative determination of gelatine present in paper performed highly satisfactorily. The results are comparable, although the choice of technique will finally be influenced by instrument versatility, portability and cost.

Acknowledgements

The authors gratefully acknowledge financial support of the Slovenian Research Agency, Programme no. P1-0153. The financial support of the NSF under grant ECS-0547013, of AFOSR under grant FA9550-06-1-0096, and of UMSL ITS under HPCC is also gratefully acknowledged.

References

[1] K. Garlick, AIC Book and Paper Group Annual 5 (1986) 94-107.
[2] A.-L. Dupont, J Chromatography A. 950 (2002) 113-124.
[3] P.I. Rose, Gelatin, in: H.F. Mark, N. Bikales, C.G. Overberger, G. Menges, J.I. Kroschwitz (Eds.), Encyclopedia of Polymer Science and Engineering, Wiley, New York, 1987, pp. 488-513.
[4] D. J. Prockop, Collagens, in: W.J. Lennarz, M.D. Lane (Eds.), Encyclopedia of Biological Chemistry, Academic Press, New York, 2004, pp. 482-487.
[5] J.E. Eastoe, Biochem J. 61 (1955) 589-600.
[6] R.E. Neuman, M.A. Logan, J Biol Chem. 186 (1950) 549-556.
[7] T.R. Keenan, Gelatin, in: R.E. Kirk, D.F. Othmer (Eds.), Encyclopedia of Chemical Technology, John Wiley & Sons Inc., New York, 1994, pp. 406-416.
[8] P.L. Privalov, in: C.B. Anfinsen, J.T. Edsall, F.M. Richards (Eds.), Advances in Protein Chemistry, Academic Press, New York, 1982, pp. 1-104.
[9] P.I. Rose, in: T.H. James (Ed.), The Theory of the Photographic Process, Mcmillan, New York, London, 1977, pp. 51-56.
[10] A.A. Leach, J.E. Eastoe, in: A.G. Ward, A. Courts (Eds.), The science and technology of gelatin, Academic Press, London, 1977, pp. 475-506.
[11] TAPPI. 1991a. Glue in paper (qualitative and quantitative determination), T504 om-89, in: TAPPI Test Methods, Atlanta, Ga.: Technical Association of the Pulp and Paper Industry, 1991.
[12] R.E. Neuman, M.A. Logan, J Biol Chem. 184 (1950) 299-306.
[13] H. Stegemann, Hoppe-Seylers Z Physiol Chem. 311 (1958) 41-45.
[14] F. Li, C.K. Lim, in: K. Blau, J.M. Halket (Eds.), Handbook of Derivatives for Chromatography, Wiley, Chichester, 1993, pp.157-174.
[15] G. Lunn, L.C. Hellwig, Handbook of Derivatization Reaction for HPLC, Wiley, New York,

1998.
[16] J.D. Russell, J.M. Dolphin, M.D. Koppang, Anal Chem. 79 (2007) 6615-6621.
[17] N. Seiler, Fluorescent derivatives, in: K. Blau, J.M. Halket (Eds.), Handbook of Derivatives for Chromatography, John Wiley and Sons; New York, 1993, pp. 175-213.
[18] S. Einarsson, B. Josefsson, S. Lagerkvist, J Chromatogr. 282 (1983) 609-618.
[19] C.H. Stephens, T. Barrett, P.M. Whitmore, J.A. Wade, J. Mazurek, M. Schilling, J. Am. Inst. Conserv. 47 (2008) 201-215.
[20] K. Uchida, Y. Kato, S. Kawakishi, J. Agric. Food Chem. 40 (1992) 9–12.
[21] T. Trafela, M. Strlič, J. Kolar, D.A. Lichtblau, M. Anders, D. Pucko Mencigar, B. Pihlar, Anal Chem. 79 (2007) 6319-6323.
[22] D.A. Lichtblau, M. Strlič, T. Trafela, J. Kolar, M. Anders, Appl Phys A. 92 (2008) 191-195.
[23] C.W. Brown, P.F. Lynch, R.J. Obremski, D.S. Lavery, Anal Chem. 54 (1982) 1472-1479.
[24] R. Leardi, in: R. Leardi (Ed.), Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks, Elsevier, Boston, 2003, pp. 169-198.
[25] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.
[26] A.S. Fraser, Austral J Biol Sci. 10 (1957) 492-499.
[27] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Massachusetts, 1989.
[28] C. B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99-145.
[29] R. Leardi, J. Chemometrics 15 (2001) 559-569.
[30] Z. Xiaobo, Z. Jiewen. M. J. W. Povey, M. Holmes, M. Hanpin, Anal. Chim. Acta 667 (2010) 14-32.
[31] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311-317.
[32] A. Durand, O. Devos, C. Ruckebusch, J. P. Huvenne, Anal. Chim. Acta 595 (2007) 72-79.
[33] J.W. Henderson, R.D. Ricker, B.A. Bidlingmeyer, C. Woodward, Agilent Technologies Pub. No. 5980-1193E, 2000.
[34] P. Bernado, E. Mylonas, M.V. Petoukhov, M. Blackledge, D.I. Svergun, J Am Chem Soc. 129 (2007) 5656-5664.
[35] H. Martens, T. Naes, Partial least squares regression, Mutivariate Calibration, Wiley, Chichester, 1989.
[36] M. Badadani, S.V. SureshBabu, K.T. Shetty, J Chromatogr B 847 (2007) 267-274.
[37] L.M. Hanssen, K.A. Snail, in: J.M. Chalmers, P.R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, Wiley, Chichester, 2002, pp. 1175-1192.
[38] K. Murayama, B. Czarnik-Matusewicz, Y. Wu, R. Tsenkova, Y. Ozaki, Appl Spectrosc. 54 (2000) 9 78-985.
[39] J. Wang, M.G Sowa, M.K. Ahmed, H.H. Mantsch, J Phys Chem. 98 (1994) 4748-4755.
[40] J.S. Shenk, J.J. Workman, M.O. Westerhaus, in: D.A. Burns, E.W. Ciurczak (Eds.), Handbook of Near-Infrared Analysis, Marcel Dekker, New York, 2001, pp. 419-474.