Anal Bioanal Chem (2014) 406:4715–4724
DOI 10.1007/s00216-014-7856-y

RESEARCH PAPER

Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least squares regression

Andrew Todd Weakley · Arthur L. Miller ·
Peter R. Griffiths · Sean J. Bayman

Received: 14 February 2014 / Revised: 15 April 2014 / Accepted: 24 April 2014 / Published online: 16 May 2014

Abstract The feasibility of measuring airborne crystalline silica (α-quartz) in noncoal mine dusts using a direct-on-filter method of analysis is demonstrated. Respirable α-quartz was quantified by applying a partial least squares (PLS) regression to the infrared transmission spectra of mine-dust samples deposited on porous polymeric filters. This direct-on-filter method deviates from the current regulatory determination of respirable α-quartz by refraining from ashing the sampling filter and redepositing the analyte prior to quantification using either infrared spectrometry for coal mines or x-ray diffraction (XRD) from noncoal mines. Since XRD is not field portable, this study evaluated the efficacy of Fourier transform infrared spectrometry for silica determination in noncoal mine dusts. PLS regressions were performed using select regions of the spectra from nonashed samples with important wavenumbers selected using a novel modification to the Monte Carlo unimportant variable elimination procedure. Wavenumber selection helped to improve PLS prediction, reduce the number of required PLS factors, and identify additional silica bands distinct from those currently used in regulatory enforcement. PLS regression appeared robust against the influence of residual filter and extraneous mineral absorptions while outperforming ordinary least squares calibration. These results support the quantification of respirable silica in noncoal mines using field-portable infrared spectrometers.

Keywords Partial least squares · Monte Carlo unimportant variable elimination · Silica measurement · FT-IR · Mine dust

Introduction

The US Mine Safety and Health Administration (MSHA) mandates the monitoring and quantification of occupational exposure to airborne respirable silica in US mines [1]. The most abundant polymorph of crystalline silica, α-quartz, is internationally recognized as a carcinogen [2–4]. Failure to quantify and mitigate worker exposure to respirable quartz leads to a decrease in lung function [5] and may result in respiratory diseases such as silicosis [6–9].

Since no field-portable method exists to monitor silica exposure, a current research goal of the National Institute for Occupational Safety and Health (NIOSH) Office of Mine Safety and Health Research (OMSHR) involves evaluating the efficacy of a direct-on-filter method for potentially measuring silica in the field. The two analytical techniques presently used to quantify silica in respirable samples collected in mines are mid-infrared (IR) spectrometry and X-ray diffraction (XRD), which are used for coal mines and noncoal mines, respectively [10–12]. The ubiquity, speed, and diminishing cost of instrumentation makes IR spectroscopy the preferred analytical technique for the purpose of field-portable measurements. However, the success of on-site assessment using portable FT-IR spectrometers is contingent upon overcoming critical implementation barriers associated with end-of-shift silica assessment [13, 14].

Field-portable IR methods will require procedures that anticipate filter-substrate mishandling and limit the need for

Electronic supplementary material The online version of this article (doi:10.1007/s00216-014-7856-y) contains supplementary material, which is available to authorized users.

A. T. Weakley (✉)
Department of Chemical and Materials Engineering, University of Idaho, PO Box 441021, Moscow, ID 83844-1021, USA
e-mail: aweakley@vandals.uidaho.edu

A. L. Miller · S. J. Bayman
National Institute for Occupational Safety and Health, 315 E. Montgomery Ave., Spokane, WA 99207, USA

P. R. Griffiths
Griffiths Consulting LLC, Ogden, UT 84403, USA

specialized training in quantitative spectroscopic analysis [12, 15]. Regulatory agencies currently utilize multiple preparation steps prior to acquiring an IR spectrum of airborne dust. These include pre- and post-weighing filters, removal of the filter along with any organic contaminants by plasma ashing, and finally redeposition of the treated dust on a clean polymeric filter prior to measurement. Sample preparation and spectral post-processing both afford distinct stages to introduce variability into the prediction [16, 17]. Technical aptitude in these domains still dictates the accuracy of quartz determination by experts.

Acquiring spectra from samples collected directly on the filter substrate would facilitate the rapid determination of α-quartz by circumventing the ashing and redeposition steps. This approach has been tested in the recent past [15, 18, 19]. A consequential improvement in prediction accuracy is expected as opportunities for sample loss or mishandling are removed from the protocol [13, 17]. Unfortunately, this approach leaves the filter in place along with any trace organic and mineral interferents within the IR sampling volume. Kaolinite clay is the most serious and common interference when using the MSHA-regulatory IR method for quartz prediction [1, 20], since its presence inflates the estimated mass of quartz [21, 22]. Although kaolinite-correction methods are commonly used, they can add to the inaccuracy of the determination of α-quartz, since there are several types of kaolin with slightly different IR spectra [23]. Thus the application of chemometric techniques, such as partial least squares regression [24–26], may produce calibrations resistant to confounder and substrate interference when these effects are appropriately modeled or at least suppressed by carefully selecting only stable predictors (channel-wavenumbers) quintessential to determining a target analyte.

Partial least squares (PLS) regression demonstrates remarkable selectivity and prediction accuracy when applied to complex multicomponent spectra [25–27]. A major appeal of PLS regression in spectroscopy relates to its handling of predictor collinearity and explicit modeling of the data structure (i.e., latent physicochemical effects). In other words, PLS regression assumes that a few latent variables, articulated mathematically as linear combinations of predictors (absorbances at each wavenumber), underscore the relationship between absorbance changes and analyte concentration. By isolating a common subspace between the concentration y-vector and the matrix of IR absorbances, PLS regression inherently minimizes the influence of irrelevant species on calibration while maximizing the linear relationship(s) between analyte concentration and target absorbance.

Using a PLS approach, improvement in the quantification of airborne silica is expected [28, 29]. Given a host of known background interferences and probable unknown day-to-day variation in geological interferences, an integrated approach to PLS calibration will better suit the direct-on-filter quantification of airborne silica. Furthermore, this feasibility study assesses the role of sample subset partitioning on PLS calibration, the gains in precision, and aid to latent variable interpretation imparted via a novel approach to Monte Carlo unimportant variable elimination (MCUVE) [30], and the ultimate basis for selecting a viable PLS model. Candidate PLS models generated from variable elimination are compared to an ordinary least squares (OLS) regression derived from the MSHA P-7 method.

Experimental

Quartz sampling

Filter samples of noncoal mine dusts were obtained from field surveys in three active mines in Idaho, New York, and Ohio. Samples were acquired from mines with sedimentary and igneous rock formations. Seventeen samples were acquired from a hard rock silver mine on two separate occasions, 21 samples from a granite mine, and 8 samples from a limestone mine. A total of 46 samples were available for analysis.

Airborne dust was collected using sampling trains standardized for noncoal mines [11]. The sampling pump flow rate was set at 1.7 L min^{-1} and large particulate matter was removed using Dorr-Oliver style cyclones, which separate the respirable fraction. Respirable particles were deposited onto preweighed 37-mm polyvinyl chloride (PVC) filters with 5-μm pore size (SKC Corp., Inc.).[1] Filters were mounted within three-piece plastic cassettes. After collection, all filter samples were postweighed to determine the mass of loaded dust. For each group of filter samples collected during one mine visit, three unused filters from the same lot were set aside as controls.

FT-IR instrumentation and acquisition parameters

Spectra were collected in transmission mode at 4 cm^{-1} resolution by averaging 40 scans using the Bruker Optics model Alpha FT-IR spectrometer. Interferograms were processed using Blackman-Harris 3-term apodization prior to Fourier transformation. Spectra were saved from 399.5 to 3,998.5 cm^{-1} so that each spectrum contained 2,542 spectral channels, i.e., discrete wavenumbers, since the data spacing

[1] Since particles larger than 4 μm were removed by the Dorr-Oliver cyclone and the pore size is specified as 5 μm, it may be asked why all the particles did not pass through the pores. In practice, the PVC from which the filters were fabricated is porous and the pathways through it are sufficiently tortuous that small particles contact and adhere to the structural features via diffusion, interception, and impaction.

for the Bruker Alpha spectrometer operating at a nominal resolution of 4 cm^{-1} is 1.417 cm^{-1}. Each individual spectrum was ratioed against the spectrum of a blank filter to remove the filter background. Multiplicative scatter correction (MSC) was applied to remove scattering effects from each individual spectrum as well [31]. Figure 1 shows the resulting average spectra from 2,200 to 400 cm^{-1} for samples from the silver mine (see Electronic Supplementary Material Figs. S1 and S2 for the granite and limestone mine examples, respectively).

To compensate for the nonuniform spatial distribution of dust particles, each filter was carefully mounted using a stainless steel holder to ensure that the 6-mm diameter IR beam always interrogated the center of the filter. Miller et al. showed that the estimation of silica using only this center portion of the filter adequately captured data representative of the silica deposited onto the PVC filter, when samples were collected using any one of three common sampler fixtures (filter cassettes) and a Dorr-Oliver-style respirable cyclone [14]. After analysis, all samples were sent to an independent laboratory for XRD analysis, which is the primary analytical method used for regulatory measurement of silica in noncoal mines.

X-ray diffraction

Quantitative XRD analysis (NIOSH 7500 method) was performed independently (RJ Lee Group Inc, Monroeville, PA) to generate the empirical α-quartz standards for PLS and manual FT-IR calibration [11]. The filter and any organic matter in the dust was eliminated using a low-temperature radiofrequency plasma asher. The appropriate procedures were then followed to wash, suspend, and disperse intact particulate matter in solution. Thin film deposition of the suspended solid onto a qualified 25-mm silver membrane filter substrate followed. This was then placed in an XRD sample holder. Working standards of α-quartz (NIST SRM 2950) were prepared in triplicate and used to generate a calibration curve from the referenced-normalized, α-quartz using the primary diffraction peak at 26.66° (2θ). Pertinent diffraction peaks collected from respirable α-quartz were also normalized and the concentration of silica (μg SiO_2/filter) estimated using this calibration curve. The PANalytical Cubix Pro X-ray diffractometer equipped with rotating anode was used to gather all necessary diffractograms.

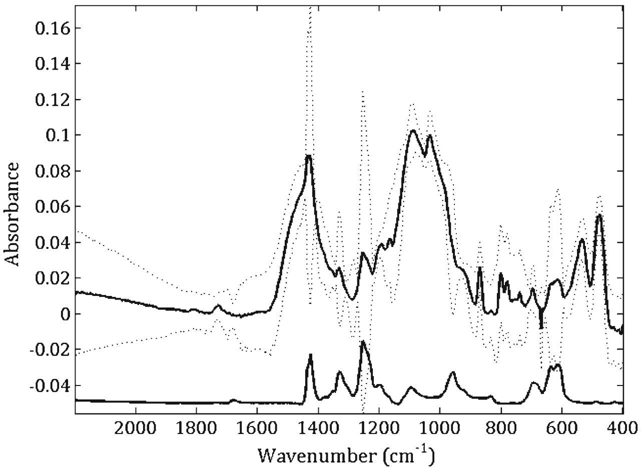

Fig. 1 Average spectrum calculated from 17 silver mine dust samples (top, solid) with accompanying ± one standard deviation on each absorbance (*dashed*). A baseline-corrected [62] and scaled (to 5 % original absorbance) spectrum of PVC is plotted with −0.05 absorbance offset. The peaks of the α-quartz doublet are clearly visible at 780 and 800 cm^{-1}

Spectral preprocessing for PLS regression

Although the single-beam spectra were ratioed against that of a blank filter, the spectra often included residual PVC bands because of differences in filter thicknesses. Low-frequency baseline perturbations due to substrate and particle scattering were also observed. Given the minimal baseline complexity, a formal baseline correction algorithm was not used. Rather, a smoothed first-derivative spectrum was produced using a Savitzky-Golay 21-point, second-order polynomial filter. This operation simultaneously removed the predominant baseline and (because of the smoothing effected by the filter) increased the signal-to-noise ratio (SNR) of each spectrum with a minimal sacrifice in resolution. Filtering removed the first 10 and the last 10 channels from each IR spectrum leaving 2,522 points in the first-derivative spectra.

Sample set partitioning based on joint *x-y* distances (SPXY) [32] was used to select calibration samples that optimally spanned the calibration range (0–281 μg-SiO_2/filter). From the original 46 samples, 2 were first removed as outliers, using principal component analysis (PCA) [27, 33], and SPXY was applied to select the best 29 (N_c) samples for training, and the remaining 15 spectra (N_p) were assigned for validation. Because 42 out of the 44 samples contained less than 200 μg-SiO_2/filter, SPXY selected all SiO_2 compositions in the higher concentration range. To ensure comparable coverage in the validation samples, one sample from the granite mine data (251 μg-SiO_2/filter) was exchanged with a low concentration SiO_2 sample from the calibration set.

Model training proceeded using two wavenumber ranges for PLS regression as follows: the wavenumber range containing the α-quartz doublet (751–857 cm^{-1}) and the spectral range between 415 and 2,201 cm^{-1} (referred to subsequently as the half spectrum). It can be assumed a priori that most channels in a full first-derivative mid-IR spectrum (415–3,986 cm^{-1}) will not be useful to PLS regression. Furthermore, all of the fundamental lattice vibrations of silica are active below ~1,400 cm^{-1}. Confining the chemometric analysis to absorbance within the 415–2,201 cm^{-1} range accomplished three things. First, it left enough redundant/useless absorbance available to challenge our variable selection algorithm (discussed in detail below). Secondly, it accelerated the required runtime for the variable selection algorithm. And finally, this spectral range facilitated an exploration of additional silica vibrations outside the fairly narrow α-quartz doublet.

Model training and testing spectra, using the masses acquired by XRD as the primary calibration values (*Y*), were mean-centered prior to regression to stabilize the PLS1 algorithm. Preprocessing, PLS modeling, and feature selection were performed in MATLAB® (2007b, The Mathworks, Natick, MA) using either the open source libPLS (v. 1.6, Changsha Nice City, China) package or custom software.

Latent variable and feature selection in PLS regression

A major practical task of PLS regression is to estimate the number of latent variables best representing the relationship between the spectra and analyte composition. In this study, Monte Carlo cross-validation (MCCV) employed a 60–40 calibration-validation split with a calibration set resampling number set to 1,000 [34]. A minimized root mean squared error of cross-validation (RMSECV) identified the optimal number of latent variables to use in all PLS models tested.

There is a growing interest in modeling the latent structure while including only the best (and fewest) spectral features [30, 35, 36]. Wavenumber selection routines often apply an objective metric and/or heuristic device to eliminate those variables least important to the modeling problem. Data reduction often removes noise and excessive redundancy. This generally improves performance and, more importantly, assists latent variable interpretation.

In this study, a modified version of Cai, Li, and Shao's MCUVE is implemented in a novel backward elimination manner [30]. In this BMCUVE routine, the latent variables are estimated at each pass of the algorithm (using MCCV) [34, 37], followed by the selection of the best subset of wavenumbers needed to predict α-quartz. The process of latent variable estimation and wavenumber elimination proceeds until a termination criterion is reached (or the number of variables available for regression is exhausted). Similar to other feature-selection methods, only calibration data are used in model reduction. Prediction testing and cross-validation inform the choice of the final PLS model.

A single pass of the MCUVE algorithm used in this study proceeds as follows: 60 % of the calibration data is randomly sampled to develop a temporary training set, a PLS regression is performed using this training data, and a regression coefficient for each predictor variable is calculated. This entire process is repeated for 1,000 trials resulting in 1,000 regression coefficients for *each* predictor variable. Note that the number of latent variables remained fixed over the 1,000 trials where only the composition of the training samples was varied by random sampling. Additionally, Cai and colleagues recommend only 100 trials for an individual pass of MCUVE [30]. This was deemed inadequate for this particular data due to the small training set.

After resampling, each variable's relative importance is assessed using a statistic known as a reliability index (RI) [38]. Formally,

$$RI_j = \frac{mean\left(\beta_{ji}\right)}{std\left(\beta_{ji}\right)}\ i = 1, 2, \ldots, N_R\ \ j = 1, 2, \ldots, p \tag{1}$$

where *mean*(β_{ji}) is the average regression coefficient for the 1,000 trials (=N_R) on the *j*th wavenumber, *std*(β_{ji}) is the *j*th wavenumber's standard deviation, and *p* are the total quantity of wavenumbers available to BMCUVE. Due to some sampling-distribution asymmetry, a robust form of the RI criterion was used where the *median*(β_{ji}) and interquartile range, *IQR*(β_{ji}), was substituted for each regression coefficient's mean and standard deviation, respectively.

Typically, variables with an *RI* value above some estimated noise level are retained (i.e., deemed important) while those below this cutoff value are discarded [38]. Given a reasonable estimate of the cutoff value, wavenumber elimination ceases after discarding unimportant variables once. This approach is adequate at removing redundant wavenumbers from spectra containing only additive noise and minimal artifacts. For this study, we assumed that mid-IR transmission spectra were nonideal on both accounts, i.e., spectra were rife with absorbance redundancies, contaminated with scattering artifacts and interferences, and contained spectroscopic biases contingent upon varying geological factors.

As opposed to operating a single pass of MCUVE, the quantity of variables available for modeling was successively reduced by 10 % until only a small number of wavenumbers remained for PLS regression. For example, if approximately half the spectral channels (e.g., 415–2,201 cm^{-1} range=1,262 channels) were evaluated by BMCUVE, the first pass of the algorithm would remove the 126 least important variables according to their low *RI* values. The remaining 1,136 channels are then evaluated in the second pass of the algorithm resulting in the subsequent removal of another 114 least important variables, and so on. Starting with a given number of predictors (p_i) and fixed percentage of wavenumber removed at each pass (q), it can be easily shown that the number of passes of the MCUVE algorithm required to reach one remaining wavenumber is expressed as follows:

$$M = {}^{-\log_{10}(p_i)}\Big/\log_{10}\left(1-\frac{q}{100}\right). \tag{2}$$

Therefore, Eq. (2) estimates that the half-spectrum elimination will take 68 passes to reach a single remaining wavenumber (after rounding *M* up to the nearest integer). Note, rounding creates a situation in which a handful of these passes are redundant. To counter a repetitive evaluation, auxiliary code is executed to reduce the number of wavelengths by one for a repetitive pass. This influences the accuracy of Eq. (2).

This entire backward elimination routine was repeated five times in order to identify a class of viable models that showed good performance according to the root mean square error of calibration (RMSEC), RMSECV, root mean square error of prediction (RMSEP), and the cross-validated coefficient of determination (Q^2). The results of PLS regression with and without wavenumber selection were compared for the two aforementioned wavenumber ranges: 751–857 cm^{-1} and 415–2,201 cm^{-1}. To investigate whether the mass of silica could be predicted without the α-quartz doublet, the 415–2,201 cm^{-1} range was evaluated using BMCUVE with the 751–857 cm^{-1} range removed prior to analysis.

Three parameters were supplied to the MCUVE function (libPLS v. 1.6, Changsha Nice City, China) prior to execution: the size of a temporary training set, the resampling number, and the number of latent variables. As described above, 60 % of the available calibration samples were randomly sampled, the resampling number was set to 1,000, and the latent variables estimated using MCCV.

Manual calibration

The manual calibration was similar to that used in previous work [13], which entailed using OPUS spectral analysis software (Bruker Optics) to perform peak integrations on IR spectra from the dust samples. We note here that band integration is susceptible to errors, especially for spectra with curved baselines. In a manner similar to that used in an earlier study [13], the approach included silica quantification via manual integration of the α-quartz doublet in the range 767–816 cm^{-1}. A correction protocol for kaolinite interference involved subtracting the potential contribution of kaolinite to the doublet using a peak-ratio technique previously published along with the kaolinite band at 915 cm^{-1} [39]. The area of the corrected doublet bands was used as a single predictor in an ordinary least squares (OLS) regression and regressed against the mass of XRD estimated silica. Although correction for kaolinite absorption was employed, the level of kaolinite in these samples was generally negligible as indicated by the absence of a strong peak at 915 cm^{-1}.

To foster simple comparisons, the samples were partitioned according to the prescription of the SPXY algorithm for the PLS regression in the region of the α-quartz doublet. A calibration curve was developed using 2/3 of the available samples for model training and the other 1/3 for prediction testing. OLS calibration, prediction, and model diagnostics were also performed using MATLAB packages.

Results

After 42 wavenumbers were removed using BMCUVE, PLS regression results were obtained using 34 wavenumbers from the quartz doublet region (Fig. 2). Performance statistics for select regression models are presented in Table 1. It may be noted that the calibration error was often higher than the prediction error for the 11 models shown in Table 1. This indicates that the training data often best spanned the calibration range while the validation set understated the model performance. This was probably caused by the manner in which the SPXY algorithm was used in this study and will be discussed in detail below.

Returning to the 34 predictor PLS model, visual inspection of Fig. 2 confirms the strong Q^2 statistic—that is, a strong linear relationship between the predictors and response variable. Remarkably, only one latent variable (LV) was required to predict quartz in the presence of the PVC substrate when the narrow wavenumber range was tested.

Table 1 shows that BMCUVE successfully suppresses redundant and/or artifact-related absorbance by removing less important wavenumbers; ultimately, this improves the performance of PLS regression relative to the cases where a wider range is used. Indeed, when all 76 wavenumber were employed for the doublet case, 2 latent variables were selected to achieve the minimized RMSECV. The removal of 42 wavenumbers led to the elimination of an additional latent variable showing an improved Q^2, RMSECV, and RMSEP. Only one true “effect,” which is captured by a corresponding PLS component in the 2, 8, and 34 wavenumber models,

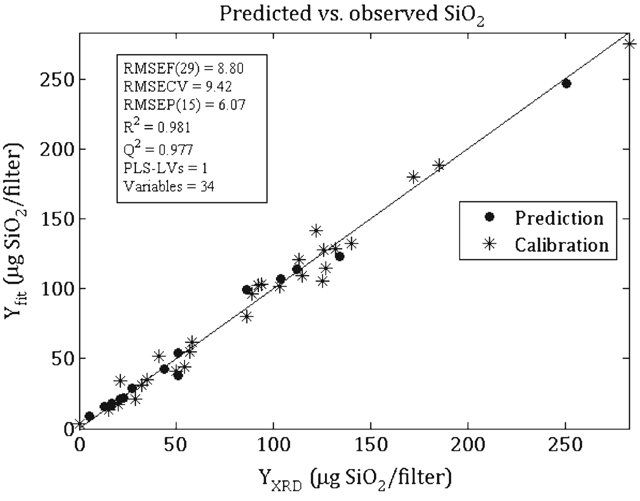

Fig. 2 Predicted versus observed α-quartz composition for the 34-variable PLS model

Table 1 Summary of select PLS regressions (lines 1–10) and manual OLS regression (line 11). The 20 wavenumbers chosen for model #7 were not the same as model #10. In the former case, all 20 variables resided within the α-quartz doublet. The asterisk (*) on the 415-2,201 range (Model's 8-10) indicate that 58 wavenumbers from the α-quartz doublet region were excluded from the regression and BMCUVE procedures

Model	Range (cm^{-1})	Variables (#)	LVs (#)	R^2	Q^2	RMSEC (μg SiO_2)	RMSECV (μg SiO_2)	RMSEP (μg SiO_2)
1	751–857	76	2	0.980	0.972	9.34	10.44	6.41
2		34	1	0.981	0.977	8.80	9.42	6.07
3		8	1	0.980	0.971	9.24	10.62	6.03
4		2	1	0.975	0.969	10.22	11.05	6.18
5	415–2,201	1,262	3	0.972	0.925	11.38	17.35	9.97
6		139	2	0.981	0.972	9.29	10.64	9.69
7		20	1	0.981	0.977	9.09	9.66	6.19
8	415–2,201*	1,204	3	0.972	0.918	11.78	18.04	10.44
9		223	2	0.981	0.972	9.24	10.71	9.28
10		20	2	0.985	0.980	8.15	8.95	9.42
11	767–816	1	–	0.935	0.922	15.62	17.02	11.27

accurately explains the relationship between the IR spectra (X) and the mass of silica (Y). The near coincidence of the X-loadings (p_1) and PLS loadings weights (w_1) (when plotted against wavenumber, not shown) [26] simplifies the interpretation of the latent "effect," i.e., the latent variable is simply mapping the spectral variation due to changes in the mass loading of silica on each filter sample.

A spectrum from the granite mine data illustrates the wavenumber elimination path for regressions involving the α-quartz doublet (Fig. 3). The absorbance and first-derivative spectrum are plotted with the selected wavenumbers clearly identified. The selection of wavenumbers from both components of the doublet demonstrates that the most successful PLS regressions required attributes from both the α-quartz lattice vibration at 780 cm^{-1} and the transverse optical (TO) Si-O-Si symmetric stretch at 800 cm^{-1} [22, 40–42].

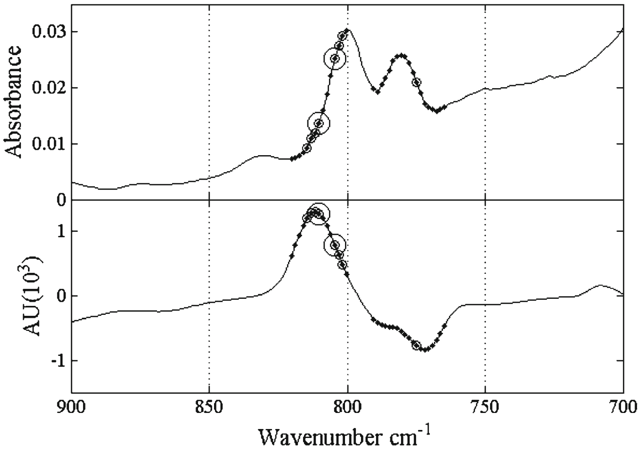

Fig. 3 Absorbance (*top*) and first-derivative (*bottom*) spectrum of the region containing the α-quartz doublet for the 20th granite mine sample. The original regression employed 76 wavenumbers from 751 to 857 cm^{-1}. BMCUVE selected the 34 (*dots*), 8 (*small circles*), and 2 (*large circles*)

Two latent variables were required in the 76-wavenumber regression. These were needed to accommodate contributions from scattering artifacts and residual PVC bands. Wavenumber elimination in first-derivative spectra began by eliminating wavenumbers with an absorbance of approximately zero followed by the weak PVC band at ~834 cm^{-1}. This order of elimination endorses the use of a reliability measure for wavenumber elimination. Overall, the most reliable variables had a large mean response and small resampling variance (IQR) which was equated to a large reliability index and thus a higher likelihood of retention. On the contrary, variations in PVC bands were random (large IQR) and uncorrelated with the silica concentration vector (small mean response), i.e., regression coefficients on PVC wavenumbers showed low reliability in quartz prediction leading to early rejection. Wavenumbers near a band's steepest points show the largest mean response, as would be expected for a first-derivative spectrum, as shown in Fig. 3; hence, this favored their retention in the PLS models as dictated by BMCUVE.

Three models were selected when the 415–2,201 cm^{-1} spectral range (1,262 channels) was screened using BMCUVE. The 1,262 wavenumber regression required 3 latent variables with an unacceptably high cross-validation and prediction error. Although model performance improved after only a single pass of BMCUVE, removing a total of 1,123 wavenumbers achieved substantially better regression performance. This was complemented by a reduction in the optimal number of PLS components. For fewer than 30 wavenumbers, only the spectral regions that included the α-quartz doublet were selected. This further highlights the value of the α-quartz region to silica prediction. The 139-wavenumber PLS regression was recorded and plotted to illustrate that other wavenumbers outside the doublet range were useful for quantitative analysis (Fig. 4). Although somewhat obscured by residual PVC bands, features that roughly represent known silica vibrations are evident.

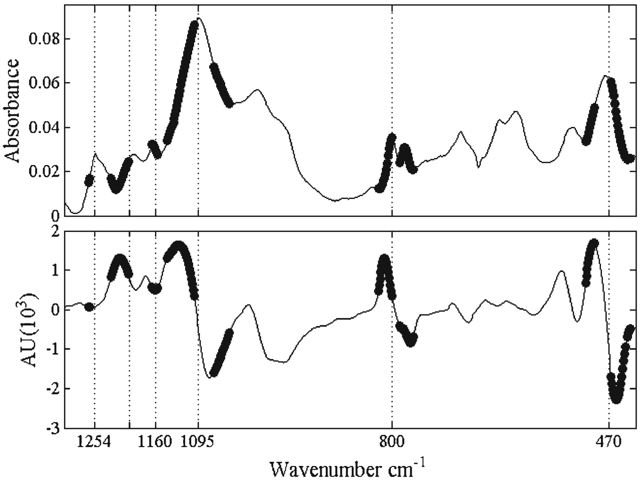

Fig. 4 Wavenumbers (*black dots*) retained upon removing 1,123 redundant predictors using BMCUVE for the 415 to 2,201 cm^{-1} range. *Vertical lines* mark the observed TO_1 mode (470 cm^{-1}), quartz doublet (780 and 800 cm^{-1}), TO_3 mode (~1,080 cm^{-1}), and tentative LO_4 (~1,160 cm^{-1}), TO_4 (1,200 cm^{-1}; unlabeled), and LO_3 mode (1,254 cm^{-1})

Adopting some shorthand from Innocenzi [41] and group symmetry nomenclature from Scott and Porto [22], bands in the spectrum shown in Fig. 4 are assigned as the Si-O-Si rocking vibration (TO_1, *E*) at 470 cm^{-1}, the antisymmetric A_2 mode at 780 cm^{-1}, the Si-O-Si symmetric stretch (TO_2, *E*) at 800 cm^{-1}, the Si-O-Si antisymmetric stretch at ~1,080 cm^{-1} (TO_3, *E*), and possibly the LO_4 (*E*) mode at ~1,160 cm^{-1} [22, 41–44]. Four wavenumbers selected between the 1,200 and 1,254 cm^{-1} band are possibly modeling effects from the TO_4 (*E*) mode and/or the LO_3 (*E*) modes, respectively. Extensive PVC interference together with the weak IR activity of the TO_4/LO_4 doublet in α-quartz greatly complicates determinations in this region [45].

Figure 4 illustrates that wavenumbers near the largest amplitude features in the derivative spectra were retained, with the exception of those near 1,160 cm^{-1}. All variables beyond 1,263 cm^{-1} were removed by BMCUVE thus nullifying the need to show the spectrum out to 2,201 cm^{-1}. Figure 4 illustrates that α-quartz transmission spectra exhibit a broad envelope spanning the region from 1,000 to 1,300 cm^{-1}. Assignment of specific bands is complicated by the polarization and orientation dependence of bands of α-quartz in this region [44, 46, 47]. In fact, large absorbance deviations (probably caused by scattering) near the tails of the 1,000–1,300 cm^{-1} envelope might support the hypothesis that this band is shifting and broadening as the sampling volume changes with α-quartz concentration. Therefore, for example, the assignment of the TO_3 mode is tentative for transmission IR spectra as this band is often unresolved from an A_2 mode unique to α-quartz [22]. Furthermore, our LO_3 assignment should not be confused with those observed in sol-gel and vitreous-SiO_2 thin films [41, 42, 48]. In those studies, the LO_3 mode often appears as an unmistakable broad shoulder between 1,200 and 1,260 cm^{-1} which shows an absorbance increase and band shift in proportion to incidence angle (Berreman effect) and thermal treatment [49, 50].

Multivariate analysis was repeated using half the spectral range while intentionally excluding the α-quartz doublet to investigate whether absorbance residing entirely outside the doublet was capable of predicting silica (Table 1, rows 8–10). The prediction error minimum was achieved using 223 wavenumbers, some of which included features from known quartz modes at 516.9 and 695.5 cm^{-1}. Five variables corresponding to baseline (>1,430 cm^{-1}) were also included in this model, suggesting a slightly suboptimal result (see Electronic Supplementary Material Fig. S3). In fact, a 20-wavenumber model was tabulated because it showed the lowest RMSECV error and utilized absorption features exclusively from the TO_1 (469 cm^{-1}) mode and A_2-TO_3 envelope (~1,080 cm^{-1}). Given a larger calibration and validation sample set, it is anticipated that an optimal model resides somewhere between 20 and 223 predictors. Notably, the exclusion of the α-quartz doublet led to higher prediction error on average.

The results of the manual OLS regression are shown in Table 1 (#11; see Electronic Supplementary Material Fig. S4). Kaolinite interference was not visibly evident in the silver and granite mine spectra. Not surprisingly, substantial calcite content was observed in limestone mine spectra (Electronic Supplementary Material Fig. S2), which obscured the ability to clearly identify the presence of kaolinite using the 915 cm^{-1} band. However, the absence of easily recognizable OH stretching vibrations of kaolinite (>3,500 cm^{-1}) in those spectra suggest that kaolin was not present [51], therefore nullifying the need for confounder correction for these samples.

Table 1 indicates a substantial increase in training and cross-validation error for the OLS model. This was complemented by a decline in the R^2 and Q^2 statistics relative to the PLS results. Prediction testing yielded error statistics greater than all PLS models (although PLS model #8 showed comparable performance).

Discussion

Manual calibration and PLS regression

The RMSEC and RMSECV statistics from the OLS regression were either higher than or similar to the largest predictor PLS models (Table 1, #5 and #8). In other words, PLS regression predicted α-quartz comparably to the manual calibration when a broad range of wavenumbers was indiscriminately used, i.e., when wavenumbers were not screened using BMCUVE. A comparative advantage of the PLS method is revealed here, namely that the success of calibration is no longer contingent on the intervention of the practitioner, i.e. since manual calibrations may require some judgment and care when estimating the area of the quartz doublet, especially in the presence of a curved baseline. Many of the dilemmas common to manual calibration are traded for the simpler choice of a derivative filter in PLS regression.

Wavenumber selection used in tandem with PLS regression greatly improves prediction relative to OLS regression. With the exception of a single wavenumber model (not tabulated), PLS regression achieved a better prediction then the OLS regression for every model, across all performance measures for the doublet range (#1–4, columns 5–9). Additionally, PLS models developed using wavenumbers from 415 to 2,201 cm^{-1} (excluding the α-quartz doublet; #8–10) performed substantially better than the manual method, achieving the best calibration (RMSEC and RMSECV) overall. The best predictive models (RMSEP) clearly required the use of the doublet region (#1–4, #7). Ultimately, PLS approaches estimated silica dust more effectively and, when used with feature selection, acted in an exploratory capacity, i.e., BMCUVE identified hitherto unexploited silica vibrations relevant to the prediction of airborne α-quartz.

Methodological aspects of PLS regression and wavenumber selection

Figure 2 illustrates that training samples selected by SPXY filled the calibration space and maximized scatter about the ideal reference line. Simply using the remaining 15 samples for prediction testing failed to adequately account for dispersion within the modeling space thereby making RMSEP artificially small. In spite of this, SPXY had an unforeseen benefit of stabilizing backward elimination by improving the odds of selecting the same wavenumbers for repeated application of the full routine.

The results of BMCUVE are not precisely reproducible due to the resampling variance inherent to Monte Carlo estimators [52]. Rerunning the entire wavenumber selection routine led to the selection of slightly different channel-wavenumbers for fixed initial conditions. Wavenumbers that were selected often resided within the same IR vibrations from trial-to-trial. In other words, the same fundamental normal modes were deemed important to the PLS modeling problem even if the exact channels were not selected. Trial-to-trial resampling variability dictates the outcome of latent variable estimation in MCCV as well; namely, repeated runs of MCCV may result in the selection of a different number of latent variables, all other things being equal. Only an infinite amount of resampling completely eliminates this effect, although reasonable performance is often achieved for a moderately sized resampling number [30]. Future studies will investigate ways to limit trial-to-trial variance as well as experiment with objective model selection criterion (e.g., Akaike information criterion) [53].

The application of routines such as BMCUVE are typically justified in the PLS literature in terms of their ability to create more parsimonious predictive models [36, 54]. Certainly, fewer channel-variables aid in interpreting the behavior of a model's predictive performance (e.g., Fig. 4), improve precision (e.g., Table 1), and possibly minimize uncertainty due to the estimation of PLS parameters [55]. From a practical standpoint, parsimony for its own sake is a relatively unimportant element to a useful calibration, i.e., moderate redundancy is completely acceptable when the latent structure is well captured by the PLS components.

The 223-wavenumber regression (Table 1, row 8) aptly illustrates this point. Relative to the 20-predictor model, an insignificant change in precision is observed with spectral features spanning ~400–1,800 cm^{-1}. More importantly, two components are required to predict silica in both regressions. Although redundancy is clearly present in the larger model, the identical latent effects are captured sufficiently over the larger spectral range. The routine, on-site quantification of airborne silica might find this redundancy useful when identifying outlying samples. In fact, intentionally including wavenumbers in the regression that correspond to known interferences (e.g., kaolinite, ~915 cm^{-1}) may facilitate the determination of silica in coal dust samples using models developed in this study.

Infrared spectra and PLS regression

The presence of uncompensated PVC bands did not hamper quantitative analysis. The retention of absorption features near α-quartz vibrations after MCUVE confirms this hypothesis. For example, modeling error was surprisingly low for the best PLS regressions (e.g., Table 1, #10) when considering the close proximity of the A_2 and TO_3 modes of quartz to a PVC band (1,086 and 1,102 cm^{-1}; Fig. 1). Furthermore, wavenumbers from the quartz E mode at 695.5 cm^{-1} were discarded late in the MCUVE routine in spite of residual PVC background (C-Cl stretch at ~690 cm^{-1}) [56–58], obscuring any clear identification.

Reliable regression coefficients were identified near the TO_4, LO_4, and LO_3 quartz modes for the 139-predictor PLS model. These modes were unearthed using BMCUVE over the half-spectral range, as illustrated in Fig. 4. Three factors complicate these band assignments, particularly for TO_4 and LO_4. They relate to interplay between PVC interference, resolution degradation due to derivative filtering, and the theoretical probability of observing TO_4–LO_4 splitting in α-quartz spectra. First, PVC has a CH_2-rocking and CH-wagging mode near the theoretical locations of the TO_4 and LO_3 modes, confounding the ability to determine visually which modes are displayed in Fig. 4 [56, 59]. Without visual verification, it is possible that BMCUVE accidentally retained redundant PVC background in the 139-wavenumber regression.

If PVC bands were inadvertently selected for the 139-predictor regression, this would indicate an early and major failure of BMCUVE. Because approximately 89 % of the redundant variables were eliminated prior to the 139-predictor regression, it is unlikely that the normal operating behavior of BMCUVE was compromised. As alluded to in the discussion of Fig. 3, wavenumber elimination appeared to behave hierarchically—wavenumbers comprising near-zero absorbance were first removed, followed by PVC background absorption bands, then by silica modes obscured by PVC, and so on. Considering this elimination order, the wavenumbers from ~1,200 to 1,230 cm^{-1} used in the PLS regression may measure a concentration-dependent variation in the TO_4 mode (although slightly perturbed by a constant PVC band contribution). Additionally, convolution of TO_4 with LO_3 features (1,252 cm^{-1}) may have resulted from the Savitzky-Golay derivative transformation. Therefore, PLS may have measured the combined changes in LO_3 and TO_4 absorbance superimposed on a constant PVC background (LO_3+TO_4+PVC constant).

To circumvent a detailed discussion of TO-LO splitting and its impact on IR activity therein, Fig. 5 shows a reference spectrum (1-cm^{-1} resolution) of Min-U-Sil 5 silica dust (US Silica, Berkeley Springs, WV) pressed into a KBr disk. The second-derivative spectrum reveals the presence of the suspected TO and LO modes at 1,081 and 1,166 cm^{-1} but the LO_3 mode was not resolvable in this spectrum. Calculated [44, 56, 60] and experimental spectra [61] further support these assignments. The TO_3 vibration appears quite broad in part due to the contribution from the A_2 mode at ~1,060 cm^{-1}, which is barely resolved even in the second-derivative spectrum measured at 1-cm^{-1} resolution. Broadening is attributed to the random orientations of the crystalline particles within the KBr substrate, yielding mixed absorption and reflection contributions to the band. Analogous or worse artifacts are anticipated for crystallites deposited onto PVC substrates for mine samples.

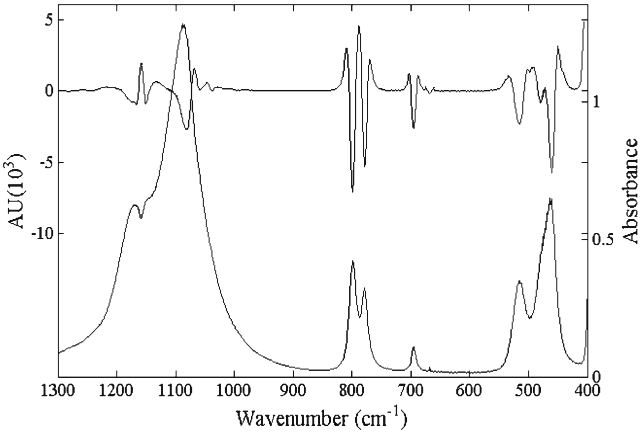

Fig. 5 Second-derivative (*left ordinate*) and absorbance spectra (*right ordinate*) for Min-U-Sil 5 analytical standard for silica

Conclusions

In US mines, respirable air samples are subjected to analysis for silica, and the nonexplicitly stated exposure limit for silica is an 8-h time-weighted average (TWA) concentration of 100 μg/m^3. Precise estimates of airborne silica, especially near this limit, are crucial for the eventual realization of end-of-shift exposure assessment. In order to prove the efficacy of a field-portable FT-IR method for the measurement of silica directly on filter samples, the interplay of preprocessing, wavenumber selection, and validation of PLS regression is paramount.

Careful preprocessing was shown to be critical to building acceptable models for the IR spectra of noncoal mine samples. PVC background correction and first-derivative transformations removed most scattering and filter interference. SPXY chose samples near the ends of the calibration range (0 and ~293 μg-SiO_2/filter). This ensured that high-concentration samples were well represented in the calibration samples. This was particularly important in this study where few dust samples were available beyond 150 μg-SiO_2/filter. An appropriate preprocessing routine, such as those using scatter correction and derivative transformation, should always precede FT-IR calibration with PLS regression, particularly for direct-on-filter α-quartz collection.

Even with some uncertainty in band assignments, wavenumbers known to constitute quartz vibrations were always incorporated into high-performing regressions by BMCUVE. Often, fewer than 50 wavenumbers realized the best predictive performance irrespective of the number of initial wavenumbers supplied to the selection routine. The BMCUVE algorithm led to the effective rejection of PVC spectral residues and other redundancies. Overall, IR absorption at the periphery of the quartz doublet led to an accurate prediction of the mass of α-quartz at the slight cost of precision and parsimony. Regardless, practical considerations may dictate that a less parsimonious model will better handle confounding absorbance, particularly for the positive interferant kaolinite.

Measuring the silica content of dust-laden filter samples using field-portable FT-IR spectroscopy appears viable using a PLS regression. In most instances, PLS regression clearly outperforms the ordinary least squares approach (MSHA P-7 analogous), which suggests potential improvement in the sensitivity and accuracy of a PLS-based analytical method for quantifying silica at end-of-shift. Notably, prediction improves when wavenumbers corresponding to known quartz vibrations are selected using BMCUVE. Given the success and minimal technical intervention of the user in the proposed

PLS modeling approach, it is recommended that models 1–4 (Table 1) be utilized for end-of-shift airborne α-quartz assessment in noncoal mines.

References

1. Mine Safety and Health Administratsion (2013) Infrared Determination of Quartz in Respirable Coal Mine Dust - Method No. MSHA P7. Pittsburgh Safety and Health Technology Center, Pittsburgh
2. World Health Organization (1997) Silica Volume 68. In: Monographs on the Evaluation of Carcinogenic Risks to Humans. Lyon, France
3. Steenland K, Mannetje A, Boffetta P, Stayner L, Attfield M, Chen J et al (2001) Cancer Causes Control 12:773–784
4. Weeks JL, Rose C (2006) Am J Ind Med 49:523–534
5. Hochgatterer K, Moshammer H, Haluza D (2013) Lung 191:257–263
6. Calvert GM, Rice FL, Boiano JM, Sheehy JW, Sanderson WT (2003) Occup Environ Med 60:122–129
7. Mannetje A, Steenland K, Attfield M, Boffetta P, Checkoway H, DeKlerk N et al (2002) Occup Environ Med 59:723–728
8. Mazurek JM, Attfield MD (2008) Am J Ind Med 51:568–578
9. Leung CC, Yu ITS, Chen W (2011) Lancet 379:2008–2018
10. National Institute of Occupational Safety and Health (2003) Silica, Crystalline by IR (KBr pellet)- Method 7602 In: NIOSH Manual of Analytical Methods (NMAM), 4th edn. Center for Disease Control and Prevention, Atlanta
11. National Institute of Occupational Safety and Health (2003) Silica, Crystalline, by XRD (filter redeposition)- Method 7500 In: NIOSH Manual of Analytical Methods (NMAM), 4th edn. Center for Disease Control and Prevention, Atlanta
12. Madsen FA, Rose MC, Cee R (1995) Appl Occup Environ Hyg 10: 991–1002
13. Miller AL, Drake PL, Murphy NC, Noll JD, Volkwein JC (2012) J Environ Monit 14:48–55
14. Miller AL, Drake PL, Murphy NC, Cauda EG, LeBouf RF, Markevicius G (2013) Aerosol Sci Technol 47:724–733
15. Kauffer E, Masson A, Moulut JC, Lecaque T, Protois JC (2005) Ann Occup Hyg 49:661–671
16. Eller PM, Feng HA, Song RS, Key-Schwartz RJ, Esche CA, Groff JH (1999) Am Ind Hyg Assoc J 60:533–539
17. Schwerha DJ, Orr CS, Chen BT, Soderholm SC (2002) Anal Chim Acta 457:257–264
18. Health and Safety Executive (2005) Crystalline silica in respirable airborne dusts Direct-on-filter analyses by infrared spectroscopy and X-ray diffraction In: Methods for the Determination of Hazerdous Substances. HSE Books, Sudbury
19. Chen CH, Tsaia PJ, Lai CY, Peng YL, Soo JC, Chen CY et al (2010) J Hazard Mater 176:389–394
20. Nayak P, Singh BK (2007) Bull Mater Sci 30:235–238
21. Painter PC, Coleman MM, Jenkins RG, Whang PW, Walker PL (1978) Fuel 57:337–344
22. Scott JF, Porto SPS (1967) Phys Rev 161:903–910
23. Lee T, Chisholm WP, Kashon M, Key-Schwartz RJ, Harper M (2013) J Occup Environ Hyg 10:425–434
24. Abdi H (2010) Wiley Interdiscip Rev Comput Stat 2:97–106
25. Næs T, Isaksson T, Fearn T, Davies T (2002) A User-friendly Guide to Multivariate Calibration and Classification. NIR Publications, Chichester
26. Wold S, Sjöström M, Eriksson L (2001) Chemom Intell Lab Syst 58:109–130
27. Kalivas JH, Gemperline PJ (2006) In: Gemperline PJ (ed) Practical Guide to Chemometrics, 2nd edn. CRC/Taylor & Francis, Boca Raton
28. Bye E (1992) Chemom Intell Lab Syst 14:413–417
29. Ritz M, Vaculikova L, Plevová E, Matýsek D, Mališ J (2012) Acta Geodyn Geomater 9:511–520
30. Cai W, Li Y, Shao X (2008) Chemom Intell Lab Syst 90:188–194
31. Isaksson T, Næs T (1988) Appl Spectrosc 42:1273–1284
32. Galvão RKH, Araujo MCU, José GE, Pontes MJC, Silva EC, Saldanha TCB (2005) Talanta 67:736–740
33. Abdi H, Williams LJ (2010) Wiley Interdiscip Rev Comput Stat 2: 433–459
34. Xu HS, Liang YZ (2001) Chemom Intell Lab Syst 56:1–11
35. Höskuldsson A (2001) Chemom Intell Lab Syst 55:23–38
36. Balabin RM, Smirnov SV (2011) Anal Chim Acta 692:63–72
37. Shao J (1993) J Am Stat Assoc 88:486–494
38. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C (1996) Anal Chem 68:3851–3858
39. Ainsworth S (2005) J ASTM Int 2:1–14
40. Saikia B, Parthasarathy G, Sarmah NC (2008) Bull Mater Sci 31: 775–779
41. Innocenzi P (2003) J Non-Cryst Solids 316:309–319
42. Osswald J, Fehr KT (2006) J Mater Sci 41:1335–1339
43. Hirata T (1999) Solid State Commun 111:421–426
44. Piro OE, Castellano EE, González SR (1988) Phys Rev B Condens Matter Mater Phys 38:8437–8443
45. Kirk CT (1988) Phys Rev B Condens Matter Mater Phys 38:1255–1273
46. Spitzer WG, Kleinman DA (1961) Phys Rev 121:1324–1335
47. Ocaña M, Fornes V, Garcia-Ramos JV, Serna CJ (1987) Phys Chem Miner 14:527–532
48. Almeida RM, Pantano CG (1990) J Appl Phys 68:4225–4232
49. Berreman DW (1963) Phys Rev 130:2193–2198
50. Gallardo J, Durán A, Di Martino D, Almeida RM (2002) J Non-Cryst Solids 298:219–225
51. Prost R, Dameme A, Huard E, Driard J, Leydecker JP (1989) Clays Clay Miner 37:464–468
52. Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman & Hall, Boca Raton
53. Burnham KP, Anderson DR (2004) Sociol Methods Res 33: 261–304
54. Mehmood T, Liland KH, Snipen L, Sæbø S (2012) Chemom Intell Lab Syst 118:62–69
55. Martens H, Høy M, Westad F, Folkenberg D, Martens M (2001) Chemom Intell Lab Syst 58:151–170
56. Krimm S (1968) Pure Appl Chem 16:369–388
57. Stromberg RR, Straus S, Achhammer BG (1958) J Res Natl Bur Stand 60:147–152
58. Tabb DL, Koenig JL (1975) Macromolecules 8:929–934
59. Ramesh S, Leen KH, Kumutha K, Arof AK (2007) Spectrochim Acta A 66:1237–1242
60. Sato RK, McMillan PF (1987) J Phys Chem 91:3494–3498
61. Francis S, Stephens WE, Richardson N (2009) Environ Health 8(S4):1–4
62. Weakley AT, Griffiths PR, Aston DE (2012) Appl Spectrosc 66:519–529