Published in final edited form as:

Anal Chim Acta. 2010 November 29; 681(1-2): 63–70. doi:10.1016/j.aca.2010.09.022.

Wavelet Analysis Used for Spectral Background Removal in the Determination of Glucose from Near-Infrared Single-Beam Spectra

Boyong Wan[†] and Gary W. Small^{*}

Department of Chemistry & Optical Science and Technology Center University of Iowa Iowa City, IA 52242

Abstract

Wavelet analysis is developed as a preprocessing tool for use in removing background information from near-infrared (near-IR) single-beam spectra before the construction of multivariate calibration models. Three data sets collected with three different near-IR spectrometers are investigated that involve the determination of physiological levels of glucose (1-30 mM) in a simulated biological matrix containing alanine, ascorbate, lactate, triacetin, and urea in phosphate buffer. A factorial design is employed to optimize the specific wavelet function used and the level of decomposition applied, in addition to the spectral range and number of latent variables associated with a partial least-squares calibration model. The prediction performance of the computed models is studied with separate data acquired after the collection of the calibration spectra. This evaluation includes one data set collected over a period of more than six months. Preprocessing with wavelet analysis is also compared to the calculation of second-derivative spectra. Over the three data sets evaluated, wavelet analysis is observed to produce better-performing calibration models, with improvements in concentration predictions on the order of 30% being realized relative to models based on either second-derivative spectra or spectra preprocessed with simple additive and multiplicative scaling correction. This methodology allows the construction of stable calibrations directly with single-beam spectra, thereby eliminating the need for the collection of a separate background or reference spectrum.

Keywords

near-infrared; wavelet analysis; multivariate calibration; glucose; partial least-squares

1. Introduction

Near-infrared (near-IR) spectroscopy has been studied for its potential use in the determination of clinically relevant analytes in biological matrixes,[1-4] as well as in direct noninvasive measurements in tissue.[5-7] In combination with multivariate calibration techniques such as partial least-squares (PLS) regression, near-IR spectroscopy can provide

© 2010 Elsevier B.V. All rights reserved.

^{*}Corresponding author. Tel: 1-319-335-3214; fax: 1-319-353-1115. gary-small@uiowa.edu..

[†]Present address. Bristol-Myers Squibb Co., 1 Squibb Dr., New Brunswick, NJ 08903.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

quick quantitative determinations of species such as blood glucose with minimal sample preparation.

One of the main impediments to the practical use of this approach is the instability of multivariate calibration models over time. This instability of calibration can be due to changes in the instrumental response, i.e., instrumental drift, or sample-based variation. The existence of such varying background information will not only worsen the precision of concentration predictions, but also complicate the multivariate models, which can lead to models with many terms that may be over-fit to the calibration data.

Removal of background information from spectra is typically performed by taking the ratio to a suitable reference or blank spectrum, followed by additional preprocessing to remove any residual background effects (e.g., baseline variation). Collection of a closely matching reference spectrum is complicated in aqueous-based samples, however, by the extreme temperature sensitivity of the water absorbance. Even slight temperature mismatches between aqueous blank and sample spectra produce baseline artifacts when the ratio is taken. Furthermore, if the spectral measurement is made directly in tissue, there is no convenient way to obtain a good matrix-matched blank.

To overcome these complications, our laboratory has explored the development of near-infrared calibrations directly with the single-beam sample spectra. This approach requires no collection of a reference background. Through the use of suitable preprocessing strategies, good calibrations have been developed,[8-10] although maintaining calibration stability over a long timeframe remains a challenge.

In the work reported here, wavelet analysis is explored as a way to remove background signatures from single-beam spectra and thereby allow stable calibration models to be formed. Wavelet analysis has found applications such as peak resolution, denoising, data compression, and baseline correction in analytical chemistry.[11] In near-IR spectroscopy, wavelet analysis has been used in conjunction with multivariate calibration for a number of applications,[12-16] although not for use in building calibrations with single-beam spectra.

In this research, wavelet analysis is applied to near-IR single-beam spectra from three different instruments to suppress non-constant background information. Following this preprocessing step, PLS calibration models are developed. These models are evaluated for their predictive ability when applied to data collected shortly after the calibration, as well as data collected over a time span of approximately six months. The performance of the wavelet preprocessing step is compared to second-derivative preprocessing, another common background suppression technique.

2. Experimental

2.1. Instrumentation and data collection protocols

Near-infrared data from three different instruments were used in this work. The first instrument, termed instrument A, was a Digilab FTS-60A Fourier transform (FT) spectrometer (Varian, Inc., Randolph, MA). Instrument B was a Nicolet Nexus 670 FT spectrometer (Thermo Electron Corp., Madison, WI) and instrument C was a Thermo Nicolet 6700 FT spectrometer (Thermo Electron Corp.). Each instrument was equipped with a tungsten-halogen source, calcium fluoride beam splitter, and liquid nitrogen-cooled indium antimonide (InSb) detector. A K-band optical interference filter (Barr Associates, Westford, MA) was placed in the optical path to isolate the 5000-4000 cm^{-1} region.

For all instruments, samples were contained in the same demountable liquid transmission cell with a 20-mm diameter circular aperture (Wilmad-Labglass, Buena, NJ), sapphire windows (Meller Optics, Inc., Providence, RI), and a 1.5 mm path-length. Sample temperatures were controlled to the range of 36.9–37.1 °C by use of a refrigerated temperature bath (Model 9100, Fisher Scientific, Pittsburgh, PA) to circulate water through an integrated jacket in the sample cell. The temperature of the sample cell was monitored with a type-T thermocouple probe and a digital thermocouple meter (Omega Engineering, Inc., Stamford, CT). Due to the extreme temperature sensitivity of the background water absorbance, some temperature-induced variance was still present in the acquired spectral data.

In the data collection process, for each sample, three replicate interferograms were measured consecutively without removing the sample from the spectrometer. To minimize correlations between concentration and time during the measurement process, samples were run in a random order with respect to the component concentrations.

With spectrometers A and C, double-sided interferograms containing 8192 points and based on 256 co-added scans were sampled at every zero-crossing of the HeNe reference laser. With spectrometer B, double-sided interferograms containing 4096 points and based on 256 co-added scans were sampled at every two zero-crossings of the laser. All interferograms were Fourier processed with triangular apodization, Mertz phase correction, and one level of zero-filling to produce single beam spectra with a point spacing of about 1.9 cm⁻¹. The region of 5000–4000 cm⁻¹ was extracted from all spectra and used for subsequent calculations.

2.2. Reagents

Two sets of samples were used in this work. In both cases, the sample matrix consisted of α -D-glucose (analyte), sodium lactate, urea, sodium L-ascorbate, β -alanine, and triacetin in phosphate buffer. This sample matrix did not mimic a specific biological sample, but rather included components whose structural types are commonly found in such samples. As shown in Fig. 1, these six components have highly overlapped absorption peaks in the near-IR combination region of 5000–4000 cm⁻¹, thereby providing a good assessment regarding the ability of a calibration methodology to extract the glucose signature from single-beam spectra corresponding to a complex sample matrix.

The first sample set consisted of 80 samples. The solvent used was a pH 6.86, 0.025 M phosphate buffer prepared with distilled water using 0.025 M KH₂PO₄ and 0.025 M Na₂HPO₄. As a preservative, 5-fluoruracil (0.1% w/w) was added to the buffer. All samples were prepared by weighing appropriate amounts of α -D-glucose (ACS reagent, Aldrich Chemical Co., Inc., Milwaukee, WI), L-sodium lactate (Aldrich), urea (minimum 99.5%, Sigma Chemical Co., St. Louis, MO), sodium ascorbate (>99%, Aldrich), β -alanine (> 99%, Aldrich), and triacetin (approx. 99%, Sigma), and diluting with the phosphate buffer to 50 mL.

The concentration of each component for each sample was randomly assigned across the 80 samples. The concentration ranges for glucose, lactate, urea, ascorbate, alanine and triacetin were 0.68–34.27, 1.18–33.34, 1.20–30.23, 1.18–22.55, 1.75–38.32 and 0.46–27.95 mM, respectively, and correlation coefficients computed between the component concentrations ranged from -0.18 to 0.11, confirming that each species was independent. Avoiding correlations among the component concentrations in the calibration samples is important to ensure that the calibration model is based on the unique spectral signature of the analyte relative to the background matrix. This requirement holds for the samples used to compute

the calibration model but does not apply to samples to which the computed model would be subsequently applied to obtain predicted concentrations.

Each of the 80 samples was measured with spectrometers A and B as described above. Before collecting near-IR spectra, glucose concentrations were confirmed with a YSI 2300 STAT PLUS clinical analyzer (YSI, Inc., Yellow Springs, OH).

The second sample set consisted of 89 samples composed of separately designed groups of 64 calibration samples and 25 prediction samples. These samples were measured with spectrometer C. Reagents for the six solutes were the same as described above. The solvent was a 0.1 M phosphate buffer at pH 7.4 prepared by titrating NaH_2PO_4 (ACS reagent, Fisher Scientific) with 50% w/w NaOH (Fisher Scientific). Sodium benzoate (Fisher Scientific, 0.5% w/w) was added to the buffer as a preservative.

The calibration and prediction sets were designed separately such that each uniformly spanned the concentration range of 1-30 mM for glucose, lactate, alanine, urea, triacetin and 1-20 mM for ascorbate. Sample concentrations were designed by a combination of uniform design[17] and subset selection.[18] In the calibration data set, the correlation coefficients between the concentrations of different components ranged from -0.1 to 0.13, while for the prediction data, correlation coefficients between different components concentrations was in the range from -0.27 to 0.21.

2.3. Data collection and splitting

The prepared 80 samples in the first sample set were measured with instruments A and B in the same order. For each instrument, data were collected during several different data collection sessions which spanned approximately one week. Thus, two data sets were collected in total, which were termed data sets A and B, respectively, depending on which instrument was used for measurement. Data collected with spectrometer A were acquired over eight days during seven sessions. With spectrometer B, data were collected over seven days during six individual data collection sessions.

To evaluate the influence of possible time-dependent instrumental drift on the prediction performance of the calibration model, both data sets were partitioned into calibration and prediction sets according to the time order of the data collection. For each data set, the data from the last two days were used as the prediction set, while data collected before the last two days were used in developing calibration models. This data separation strategy provided some information regarding the time-related stability of the developed calibration models.

For the second sample set, a long-term prediction study was implemented to produce data set C. In this study, the total data collection with instrument C lasted 205 days. The 64 calibration samples were measured during the first four days, and the 25 prediction samples were measured repeatedly afterwards on a weekly or bi-weekly basis from day 9 to day 205. A total of 14 sets of prediction spectra were collected in this manner. The prediction samples were kept frozen between measurements, and the YSI 2300 STAT PLUS clinical analyzer was used periodically to confirm the integrity of the glucose concentrations. The same data collection protocols described previously were used.

2.4. Data analysis and implementation

All spectra were transferred to a Silicon Graphics Origin 200 R12000 server (Silicon Graphics, Inc., Mountain View, CA). This computer employed the Irix operating system (Version 6.5, Silicon Graphics, Inc.). Subsequent calculations were performed with a Dell Dimension 4550 computer (Dell Computer, Austin, TX) operating under Microsoft Windows XP (Microsoft Inc., Redmond, WA). Wavelet analysis was implemented with the

Wavelet Toolbox (version 3.0) of Matlab (version 7.0, The MathWorks, Inc., Natick, MA). The PLS_Toolbox 3.0 (Eigenvector Research, Manson, WA) was used for the second-derivative filtering, as well as the development and implementation of PLS calibration models.

3. Results and discussion

3.1. Overview of data preprocessing methodology

To decrease intensity differences among the single-beam spectra, multiplicative signal correction[19] (MSC) was applied to all spectra before any further analysis. The MSC method can remove multiplicative scaling artifacts present in single-beam spectra due to experimental variation.[9] In this calculation, the mean calibration spectrum was used as a standard, and each spectrum was regressed separately against the standard with a two-parameter linear least-squares fit. The regression was performed over 4000 to 5000 cm^{-1} . The additive and multiplicative regression coefficients (i.e., y-intercept and slope) were then used to correct the offset and scale of the spectrum. Fig. 2 plots the single-beam spectra from data set A before (A) and after (B) scaling with MSC. Following this preprocessing step, wavelet analysis or second-derivative filtering were applied and PLS calibration models were developed.

The discrete wavelet transform (DWT)[20] was used to decompose the single-beam spectra into their underlying components on the basis of projections onto a set of wavelet basis functions. These basis functions are derived from a mother wavelet that is shifted and dilated such that projections onto the individual functions provide information about both the frequency content of the signal and the occurrence of the frequency components along the x -axis of the data. By contrast to Fourier analysis, which models data through projections onto infinitely long sine and cosine functions, wavelet analysis provides a somewhat more flexible modeling approach. This flexibility derives from the large selection of mother wavelets to use and the fact that the irregular shape and finite duration of the wavelet basis functions can provide a better fit to the features that comprise a non-harmonic signal such as a single-beam spectrum.

As implemented by Mallat,[21] the DWT algorithm represents the wavelet basis as a bank of finite impulse response high-pass and low-pass digital filters. The output of the filtering step produces two sets of wavelet coefficients. The high-pass filter outputs the high-frequency components of the signal that are termed the “details” (D_1). Similarly, the low-pass filter outputs the low-frequency components called the “approximations” (A_1). For an input signal with n sampled points, the Mallat algorithm incorporates downsampling such that D_1 and A_1 both have approximately $n/2$ points. A wavelet decomposition tree can be formed by applying the DWT to A_1 to produce A_2 and D_2 . This step can be repeated further, with the p^{th} decomposition producing A_p and D_p . Once the signal is decomposed, it can also be reconstructed by an inverse DWT. Modification of the wavelet coefficients during the reconstruction operation represents a way to obtain the original signal with certain features (e.g., noise or background components) suppressed.

In this work, for a p -level decomposition with a selected wavelet function, A_p was assumed to be a representation of the low-frequency background components. By zeroing A_p and then performing the reconstruction, these background components were removed and the resulting spectrum was used as the input to the PLS modeling procedure. The wavelet function and the value of p represent parameters to be optimized. Decomposition levels 4-9 were studied, and the Daubechies, symmlet and biorthogonal wavelet families were investigated. For the Daubechies and symmlet wavelets, orders 2-8 were studied. The biorthogonal family uses different wavelet functions for decomposition and reconstruction.

For decomposition and reconstruction, respectively, orders 2-6 (M) and 3-15 (N) were evaluated, with the requirement that $M+N$ is even.

Fig. 3A-C plots the results obtained when the Daubechies (6) wavelet was applied to the MSC-scaled single-beam spectra of the first three samples in data set A. Panels A, B, and C correspond to 5, 6, and 7 levels of decomposition, respectively. Three replicate spectra are plotted for each sample, but the replicates are so heavily overlapped that the individual spectra cannot be discerned. Inspection of the plots reveals an overall derivative-like character to the wavelet-processed spectra, with the prominent broad spectral background observed in Fig. 2 being suppressed to varying degrees. As the number of levels of decomposition increases, increasingly broader features are included in the reconstructed spectrum. The spectra of the three samples are distinct from each other, with the principal differences occurring in the region of 4500-4300 cm^{-1} . As observed in Figure 1, this region contains the most discriminating information for use in distinguishing the spectra of the six solutes.

For comparison to wavelet analysis, second-derivative filtering was also employed as a preprocessing procedure for the suppression of background information. This approach has been used in a variety of near-IR spectroscopy applications for removal of background artifacts.[22-25] The Savitzky-Golay least-squares approximation was used to estimate derivatives.[26,27] Adjustable parameters for this method are the polynomial order used for the derivative approximation and the spectral window size employed. Here, a second-order polynomial was used and the window size was varied over 5-41 points with an interval of 2.

Fig. 3D-F plots the results obtained when second-derivative processing was applied to the MSC-scaled single-beam spectra of the same three samples in data set A referenced above. Panels D, E, and F correspond to second-order polynomial derivative approximations that employed window sizes of 21, 31, and 41 points, respectively. The processed spectra are conceptually similar to those in Fig. 3A-C, with the broad background seen in Fig. 2 being largely suppressed. The discriminating information among the spectra of the three samples is again concentrated in the 4500-4300 cm^{-1} region. The changes across Fig. 3D-F are much more subtle than those observed across Fig. 3A-C, suggesting that more flexibility in the spectral decomposition is provided by the wavelet approach.

Inspection of Fig. 3 also suggests that there is no simple way to determine which wavelet or derivative parameters will be most effective in preprocessing single-beam spectra for the purpose of stabilizing the resulting calibration model. This suggests that a joint optimization of the preprocessing and modeling parameters must be undertaken.

3.2. Calibration optimization

After background removal, PLS calibration models were developed. For the PLS calculation, the spectral range used for calibration and the number of terms (factors or latent variables (LVs)) in the model were studied. To get the optimal combination of spectral range, number of LVs, and the wavelet or second-derivative parameters for the calibration model, a factorial experimental design was implemented. In this design, the number of LVs was varied from 6-20, and 20 spectral ranges in the window of 4700-4100 cm^{-1} were employed.

For each combination of the spectral range and wavelet or derivative parameters, a PLS model for glucose concentration was developed from the calibration data set. To determine the number of PLS factors for the model, leave-10%-out cross-validation with venetian blinds splitting was implemented. In defining the subsets of calibration samples to be withheld, the venetian blinds method takes the samples in a manner equally spaced across

the data set. For example, if the calibration set contained 60 samples, leaving 10% out would require 10 subsets of six samples. With venetian blinds sampling, the first subset would consist of samples, 1, 11, 21, 31, 41, and 51. The cross-validation method produces the root mean squared error of cross-validation (RMSECV) statistic that represents the pooled error in the predicted concentrations as each sample is withheld once.

The optimal number of PLS factors was taken as the minimum number that produced a value of RMSECV that was not statistically different from the minimum RMSECV at a significance level of 5%. An *F*-test was used to make this determination. After that, models determined this way were sorted in ascending order of RMSECV. Models with the lowest RMSECV were selected for test with the prediction data.

3.3. Analysis of data Set A

3.3.1. Data assessment—Box plots were constructed to display the distributions of the maximum intensities in the single-beam spectra of the samples collected each day. These plots revealed median intensity variations in the spectra from day to day on the order of 2% of the intensity maximum. This reveals the existence of non-constant background variation in these single-beam spectra.

Noise analysis can provide information about the quality of the measured spectra. This was accomplished by taking the ratio of each spectrum to its replicates and converting the resulting transmittance spectra to absorbance units (AU).[28,29] A second-order polynomial least-square fitting procedure was then applied to the spectral range of 4500 – 4300 cm^{-1} to correct baseline variations caused by fluctuations in sample temperature. The 4500 – 4300 cm^{-1} spectral window was chosen because this range contains the glucose C-H combination band at 4400 cm^{-1} which is known to be an important spectral feature for quantitative analysis.[28] After the quadratic fitting, root-mean-square (RMS) deviations were then computed for all three possible ratios for each sample. For the 80 samples measured with instrument A, the estimated minimum, mean and maximum RMS noise estimates were 2.42, 4.44 and 16.40 μAU , respectively.

3.3.2. Calibration results without background removal—Before doing any background removal, calibration models were developed from the MSC-processed single-beam spectra directly. Twenty calibration models were developed corresponding to the 20 spectral ranges investigated. Top models were selected according to the sorted RMSECV values. The parameters for two representative top models and their corresponding calibration and prediction performance are listed in Table 1. These values should be considered baseline results for use in evaluating the background removal strategies. Fig. 4A shows a correlation plot between the predicted and prepared concentrations. Fig. 4B is the corresponding residual plot for the calibration and prediction data. In this plot, concentration residuals are plotted with respect to the time sequence of the data collection.

3.3.3. Background removal with wavelet Analysis—Calibration models were optimized as described previously, this time incorporating the wavelet processing parameters into the experimental design. Table 1 lists the performance characteristics for the top two models found. Fig. 5 displays the concentration correlation plot and the corresponding residual plot for the PLS model developed after background removal using the Daubechies (6) wavelet. We can see that, with optimization of the wavelet processing and spectral range parameters, the prediction performance of the calibration model was much improved. With the use of the Daubechies (6) wavelet and a decomposition level of 7, a standard error of calibration (SEC) of 0.56 mM and a standard error of prediction (SEP) of 0.70 mM were obtained with a 13-factor PLS model based on 4568–4141 cm^{-1} . This

represents a 27% improvement in SEP relative to the best result when only MSC preprocessing was used.

These results suggest that low-frequency background variation is a significant contributor to prediction error. Suppression of such non-constant low-frequency background variation with wavelet analysis can significantly improve the predictive performance of the calibration model.

3.3.4. Background removal with second-derivative filtering—Calibration models were optimized as described previously, incorporating the window size of the derivative calculation into the experimental design. Table 1 lists the top two models with second-derivative filtering and their corresponding calibration and prediction statistics. From the table, we can see that, compared with no background removal, second-derivative filtering does not offer any improvement in the predictive performance of the model. On the contrary, the calibration models become more complicated, which results in more PLS factors and a higher prediction error. This reflects the fact that, although second-derivative filtering may be able to remove some low-frequency information, it provides little freedom in specifying the values of cutoff frequencies and the degree of attenuation. This inflexibility may cause second-derivative filtering to remove an excessive amount of the analyte signal along with the background components. These results are consistent with the assessment afforded by Fig. 3 in which it was concluded that background removal by wavelet analysis has greater flexibility.

3.4. Analysis of data set B

Box plots were also constructed with the single-beam spectra in data set B. Variation in signal intensity over the six data collection sessions was similar to that found with data set A, thereby again suggesting the benefit of MSC preprocessing. The analysis of RMS noise levels in the 4500–4300 cm^{-1} range was also performed as before. For this data set, the minimum, mean and maximum RMS noise values were 2.70, 6.97 and 47.69 μAU , respectively.

The same data processing and model optimization procedures were applied to data set B. The top models and corresponding calibration and prediction results are listed in Table 2 for the cases in which (1) only MSC was applied, (2) wavelet analysis was performed, and (3) second-derivative preprocessing was used. Compared with the best model obtained without background removal, the best calibration model developed after wavelet processing had fewer latent variables, and resulted in a 31% improvement in SEP. In the best case, an SEP of 0.24 mM was obtained with 10 PLS factors.

For second-derivative filtering, the top two optimized calibration models both had nine PLS factors. However, compared with the best model obtained without background removal, the optimal calibration models that included the derivative calculation did not offer an improved SEP. Again, this reflects the lack of flexibility in second-derivative filtering. Although the derivative calculation can remove the low-frequency background information, it also tends to remove an excessive part of the analyte information.

Consideration of the results presented in Tables 1 and 2 allows comparisons to be made between the calibration models computed with data sets A and B. Generally speaking, both before and after background removal with wavelet analysis, in terms of calibration and prediction results, the model developed with data set B provided better performance than the model developed with data set A. The better performance by data set B may be due to a higher level of instrumental stability of instrument B. As instrument A was significantly older, the instrumental response may have been more varying compared with instrument B.

Although background removal methods, such as wavelet analysis, can decrease such variation, they cannot be eliminated completely. Consequently, such variation will still have a large influence on the performance of the developed calibration model.

3.5. Analysis of data set C

To study the performance of the signal processing methods on long-term data, the same investigations were performed with data set C. To study the long-term performance of the calibration model, PLS models were developed with the calibration data collected in the first four days, and then applied to the data collected in each prediction week. In this way, the prediction performance of the calibration model could be evaluated over time as the spectrometer drifted.

Box plots in Fig. 6 show the distribution of maximum intensities in the single-beam spectra collected in each of the 14 prediction sets. Although the same set of samples was measured in each prediction set, spectral intensity variation was observed from week to week. This indicates existence of non-constant background variation. Analysis of RMS noise levels was also performed as before. The corresponding minimum, mean and maximum RMS values over the 4500-4300 cm^{-1} range were 2.85, 7.88 and 35.90 μAU , respectively.

As before, the calibration and prediction data were first preprocessed with MSC to remove additive and multiplicative scaling effects. After that, calibration models were developed directly or after further background suppression with wavelet analysis or second-derivative filtering. Model optimization and the selection of the top models were performed in the same way described in the sections above. The same levels of the parameters discussed previously were investigated in searching for the optimal models. Table 3 shows the parameters for the optimized calibration models for different processing methods.

After the optimization study, the calibration models were applied to the long-term prediction data week by week. Bar plots in Fig. 7 show the week-by-week prediction performance of the calibration models obtained with the different processing methods. As a general trend, the prediction error increases with time. For the influence of the background-removal procedures, it can be seen that for the first three weeks, there is little difference in the prediction performance of models developed with or without background removal. However, from prediction set 4 onward, we can see that the calibration model developed after wavelet analysis is more stable and less sensitive to instrumental drift, while models developed without background removal degraded more quickly with time. The model based on second-derivative filtering did not provide any improvement in long-term prediction performance. Again, this reflects the lack of ability to tune the second-derivative filtering calculation to optimize the tradeoff between the removal of the background and analyte information. If the average SEP is considered over the 14 prediction sets, the model based on wavelet analysis produced a 33% improvement over the model with only MSC scaling and a 40% improvement over the model based on second-derivative filtering.

Given the relatively large numbers of latent variables used with the PLS models reported here and the potential for these models to be overly fitted to the calibration data, SEP values were also computed for the 14 prediction sets with one and two fewer latent variables included in the models. For all three types of data preprocessing, average SEP values increased relative to those reported in Table 3. This result lends some confidence to the cross-validation and *F*-test procedures used to estimate the optimal number of latent variables.

4. Conclusions

The results presented in this paper demonstrate that the existence of non-constant instrumental variation may degrade the performance of calibration models quickly with time. To improve the robustness of the models, signal processing methods can be used to suppress the time-related variation of the instrumental response. By applying wavelet analysis to data collected from three different instruments, including a long-term study, models developed after background suppression were more stable with time and more robust to the effects of instrumental drift. Improvements in SEP values on the order of 30% were obtained with each of the three data sets tested.

On the basis of the work performed here, wavelet analysis appears to more efficient than second-derivative filtering in suppressing non-constant background variation. This is attributed to greater flexibility in the spectral decomposition step. With wavelet preprocessing, stable calibration models could be developed directly with single-beam spectra, thereby eliminating the requirement to collect a separate background spectrum.

Acknowledgments

This research was supported by the National Institutes of Health under grant DK60657.

References

1. Hazen KH, Arnold MA, Small GW. *Anal. Chim. Acta*. 1998; 371:255.
2. Ham FM, Cohen GM, Kostanic IN, Gooch BR. *Physiol. Meas.* 1996; 17:1. [PubMed: 8746372]
3. Amerov AK, Chen J, Small GW, Arnold MA. *Anal. Chem.* 2005; 77:4587. [PubMed: 16013877]
4. Ciurczak, EW.; Drennen, JK, III. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*. Marcel Dekker; New York: 2002.
5. Maruo K, Tsurugi M, Tamura M, Ozaki Y. *Appl. Spectrosc.* 2003; 57:1236. [PubMed: 14639751]
6. Arnold MA, Small GW. *Anal. Chem.* 2005; 77:5429. [PubMed: 16131049]
7. Li L-N, Li Q-B, Zhang G-J. *J. Infrared, Millimeter, Terahertz Waves.* 2009; 30:1191.
8. Lu G, Zhou X, Arnold MA, Small GA. *Appl. Spectrosc.* 1997; 51:1330.
9. Ding Q, Small GW, Arnold MA. *Applied Spectroscopy.* 1999; 53:402.
10. Kramer KE, Small GW. *Appl. Spectrosc.* 2009; 63:246. [PubMed: 19215656]
11. Ehrentreich F. *Anal. Bioanal. Chem.* 2002; 372:115. [PubMed: 11939178]
12. Foca G, Cocchi M, Li Vigni M, Caramanico R, Corbellini M, Ulrici A. *Chemom. Intell. Lab. Syst.* 2009; 99:91.
13. Liang L, Wang B, Guo Y, Ni H, Ren Y. *Vib. Spectrosc.* 2009; 49:274.
14. Liu Z, Cai W, Shao X. *Analyst (Cambridge, U. K.)*. 2009; 134:261.
15. Shao Y, He Y, Bao Y, Mao J. *Int. J. Food Prop.* 2009; 12:644.
16. Zhang Y, Cong Q, Xie Y, Yang J, Zhao B. *Spectrochim. Acta, Part A*. 2009; 71A:1408.
17. Fang, KT. *Uniform Design and Uniform Design Tables*. Science Press; Beijing: 1994.
18. Carpenter SE, Small GW. *Anal. Chim. Acta*. 1991; 249:305.
19. Isaksson T, Naes T. *Appl. Spectrosc.* 1988; 42:1273.
20. Jensen, A.; La Cour-Harbo, A. *Ripples in Mathematics: The Discrete Wavelet Transform*. Springer-Verlag; Heidelberg: 2001.
21. Mallat S. *IEEE Pattern Anal. and Machine Intell.* 1989; 11:674.
22. Myers DE, Anderson LD, Seifert RP, Ortner JP, Cooper CE, Beilman GJ, Mowlem JD. *J. Biomed. Opt.* 2005; 10:034017/1. [PubMed: 16229661]
23. Miyazawa M, Sonoyama M. *J. Near Infrared Spectrosc.* 1998; 6:A253.
24. Cooper CE, Elwell CE, Meek JH, Matcher SJ, Wyatt JS, Cope M, Delpy DT. *Pediatr. Res.* 1996; 39:32. [PubMed: 8825383]

25. Izutsu K-I, Fujimaki Y, Kuwabara A, Hiyama Y, Yomota C, Aoyagi N. J. Pharm. Sci. 2006; 95:781. [PubMed: 16498574]
26. Savitzky A, Golay MJE. Anal. Chem. 1964; 36:1627.
27. Steinier J, Termonia Y, Deltour J. Anal. Chem. 1972; 44:1906.
28. Hazen KH, Arnold MA, Small GW. Appl. Spectrosc. 1994; 48:477.
29. Chen J, Arnold MA, Small GW. Anal. Chem. 2004; 76:5405. [PubMed: 15362899]

Fig. 1.

Pure-component absorbance spectra of the six components obtained by taking the ratio of single-beam spectra of each constituent to a spectrum of phosphate buffer. The trace of glucose (solid line) corresponding to a concentration of 150 mM clearly shows the two C-H combination bands located at 4300 and 4400 cm^{-1} and the O-H combination band located at 4700 cm^{-1} . The other five components, lactate, urea, ascorbate, alanine and triacetin, were prepared at concentration levels of 400, 200, 60, 150 and 250 mM, respectively. Negative absorbance values arise due to differences in water concentration between the sample and reference spectra used in the absorbance calculation.

Fig. 2.

Unprocessed single-beam spectra from data set A (A) and spectra after application of the MSC transform (B). Intense water absorption bands on the edges of the spectrum effectively zero the light intensity outside of the 4900-4100 cm⁻¹ range. No analyte features are visible in the single-beam spectra at the mM concentration levels used.

Fig. 3.

Single-beam spectra corresponding to the first three samples in data set A after application of wavelet (A-C) or second-derivative (D-F) preprocessing. The spectra were scaled first by MSC. Three replicate spectra are shown for each sample, although the close similarity of the replicates does not allow them to be visually distinguished in the plots. Wavelet processing used the Daubechies (6) wavelet and 5 (A), 6, (B), or 7 (C) levels of decomposition. Second-derivatives were based on a second-order polynomial approximation and window sizes of 21 (D), 31 (E), or 41 (F) points. The range of 4700–4100 cm^{-1} is shown, corresponding to the spectral window explored in the construction of the calibration models.

Fig. 4.

(A) Correlation plot of predicted vs. prepared concentrations for single-beam spectra in data set A with only MSC preprocessing. Squares and triangles denote calibration and prediction points, respectively. The model used was based on 14 PLS factors computed from the range of 4568-4141 cm^{-1} . The values of SEC and SEP were 0.66 and 1.10 mM, respectively. (B) Plot of concentration residuals as a function of the sequence of data collection for the calibration model described in plot A. Symbols are the same as in plot A.

Fig. 5.

(A) Correlation plot of predicted vs. prepared concentrations for single-beam spectra in data set A with MSC preprocessing, followed by background removal with the Daubechies (6) wavelet and a decomposition level of 7. Squares and triangles denote calibration and prediction points, respectively. The model used was based on 13 PLS factors computed from the range of $4568\text{--}4141\text{ cm}^{-1}$. The values of SEC and SEP were 0.56 and 0.70 mM, respectively. (B) Plot of concentration residuals as a function of the sequence of data collection for the calibration model described in plot A. Symbols are the same as in plot A.

Fig. 6. Distributions of the maximum intensities in single-beam spectra collected in each prediction set with instrument C. The horizontal line in each box denotes the median intensity, while the box limits are drawn to the upper and lower quartiles of the distribution. The whiskers attached to each box specify the maximum and minimum values of the distribution.

Fig. 7. Prediction performance of the calibration models with data set C before and after background suppression with wavelet analysis and second-derivative filtering. Values of SEP are provided for each of the 14 prediction sets. White, gray, and black bars, respectively, correspond to scaling with MSC only, preprocessing with MSC and wavelet analysis, and preprocessing with MSC and second-derivative filtering. Although still subject to performance degradation with time, the calibration models that incorporated wavelet analysis clearly performed the best.

Table 1

Results for data set A.

Processing Method	Spectral Range (cm ⁻¹)	No. of LV	SEC (mM)	SEP (mM)
MSC only	4606-4181	13	0.63	0.99
	4568-4141	14	0.66	1.10
MSC + wavelet analysis (wavelet, ^a decomposition level)	Db (6), 7	13	0.56	0.70
	Sym (7), 8	14	0.56	0.72
MSC + second-derivative filtering (window size, polynomial order)	41, 2	15	0.56	0.99
	41, 2	17	0.57	1.09

^aWavelet families listed in the table were Daubechies (Db) and symmlet (Sym). The number in parentheses denotes the order of the specific wavelet functions studied.

Table 2

Results for data set B.

Processing Method	Spectral Range (cm ⁻¹)	No. of LV	SEC (mM)	SEP (mM)
MSC only	4606-4181 4587-4163	12 12	0.30 0.30	0.37 0.35
MSC + wavelet analysis (wavelet, ^a decomposition level)	Sym (7), 7	10	0.30	0.24
	Sym (5), 8	11	0.32	0.26
MSC + second- derivative filtering (window size, polynomial order)	31, 2	9	0.32	0.35
	31, 2	9	0.31	0.38

^aThe order of the symmlet wavelet used is given in parentheses.

Table 3

Results for data set C.

Processing Method	Spectral Range (cm ⁻¹)	No. of LV	SEC (mM)	Average SEP (mM)
MSC only	4613-4226	13	0.38	1.29
MSC + wavelet analysis (wavelet, ^a decomposition level)	Db (5), 8 4613-4226	12	0.39	0.86
MSC + second- derivative filtering (window size, polynomial order)	29, 2 4613-4226	11	0.38	1.43

^aThe order of the Daubechies wavelet used is given in parentheses.