

Quantitative prediction of material properties using reflectance spectroscopy: A multivariate chemometrics-based approach

Daniel A. Shiley

ASD, Inc., 2555 55th St, Suite 100, Boulder, Colorado, USA

Abstract

Spectral reflectance-based multivariate models that utilize large numbers of spectral points to predict sample composition have been widely accepted for quantitative and qualitative analysis within agricultural, pharmaceutical and industrial markets. However, multivariate models have not been widely used for remote sensing applications or for quantitative modeling of minerals in rocks. In order to develop a multivariate calibration model we need to understand the error of the reference measurements so that realistic expectations of model performance can be established. Lab error and its effect on the calibration model will be discussed. Calibration and validation sets must contain the diversity, both spectral and compositional, that would be expected to be encountered in new samples. Additional considerations for the development of multivariate calibration models include sample presentation, creation of the calibration and validation sets, model statistics, calibration monitoring and performing updates to the calibration model.

Introduction

Many remote sensing applications and mineral identification programs have typically used univariate peak ratios or peak recognition software to relate spectral reflectance characteristics to some compositional feature or features present in the sample. However, because of the very complex and overlapping spectral features this approach usually cannot provide quantitative predictions for unknown samples. In the visible-near infrared and short wave infrared regions spectra have overlapping waveforms that limit the effectiveness for simple ratios and peak recognition routines. Quantitative

models for prediction of constituent concentrations may be developed using reflectance or absorbance formats, but it is frequently advantageous to convert reflectance spectra to absorbance, $\text{Log}(1/R)$ format when developing calibration models. This conversion can help to produce a more linear model than with reflectance spectra.

Through the use of multivariate modeling techniques such as Partial Least Squares (PLS), we can create quantitative calibration models that can simultaneously consider all wavelengths in the prediction. These multivariate models can be used to predict analyte concentration of unknown samples directly from the sample spectra. Figure 1 contains absorbance spectra from samples with varying amounts of muscovite, these spectra have been offset for clarity. As can be seen with these spectra, other minerals contained in the samples cause the overall spectral shape to differ significantly between the samples therefore, a simple ratio to peak height might not be very effective for determination of concentration.

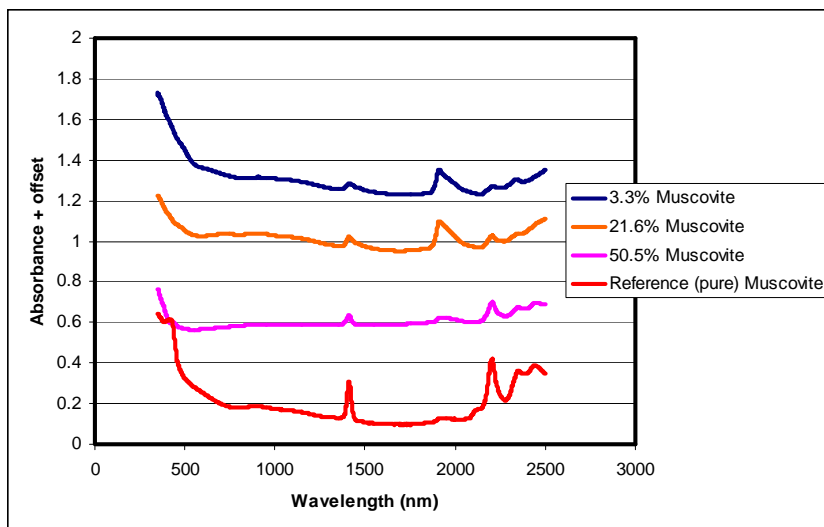


Figure 1, Absorbance spectra from mineral samples containing varying amounts of muscovite.

When developing near-infrared calibrations many sources of error can cause problems with the modeling process. These include instrumental issues, such as noise and spectral drift, errors due to sample inhomogeneity, errors due to incorrect handling of

spectral or data outliers and most importantly errors in the reference methods.

Understanding the sources of these errors and their net effect on models is of critical importance to the calibration developer (Mark and Workman, 2007).

The basic steps in creation of a multivariate model are as follows:

1. Obtain samples for calibration and validation sets.
2. Obtain reference data for these samples. This should include obtaining an estimate of laboratory error.
3. Collect spectra on the calibration and validation samples.
4. Create the multivariate model.
5. Validate the model using the validation set data.
6. Begin using the calibration as a probationary model.
7. Monitor model performance against the primary method on a frequent basis to assure continued model performance.
8. Update model as needed and revalidate the model.

Calibration and validation samples

Calibration samples should encompass the entire range of composition that would be expected from routine samples. The distribution of sample composition also is critically important. If most samples contain constituent values that are nearly the same, and only a few are toward the extremes then the model will likely not perform well and the predicted value for new unknown samples will be pulled toward the sample mean in the calibration set. However if the distribution of sample concentration is more even within the calibration set, then the predicted values will be more likely to predict properly. The constituents also should not be cross correlated, the value of each constituent should be independent of the other constituents contained in the sample as cross correlation will cause larger prediction errors. Due to problems with inter-relation, artificially created samples, where one or more samples are combined to achieve a better sample distribution are typically not much use for creation of natural product calibrations.

If blending of samples is done very sparingly, then this may be useful, but samples directly from the process (or nature) are preferred over those that have been created by blending.

One of the most frequently asked questions is how many samples will be required for a calibration? Essentially there is no set number, but a general guideline would be 50 – 75 samples for a feasibility study and 100 -150 for a good starting model. A robust model might require hundreds or even thousands of samples to represent the diversity and combinations of materials that could be encountered when using the model. To guard against over fitting the model, a reasonable guideline is that 10 or more samples should be used in the calibration set for every factor (principal component) in the model (Workman, J.J., Jr., 2008). Depending on the complexity and component concentration up to 15 or more factors (150 samples) might be required.

Selection of the validation set is critical. The samples selected for the validation set must encompass the same variation and sample characteristics as contained in the calibration set. The validation set should contain about 20% as many samples as the calibration set. An easy way to assure that the validation set is representative is to rank order the samples by their constituent value, then beginning with the first sample, mark every 5th sample for use in the validation set. Samples of certain types may also be added to the validation set if it is critical that the model performs well on that particular type. In no case should the selection of validation samples be based on which samples predict well on the calibration. Selection of samples must be made without knowledge of their predicted values. Also samples should not be removed from the validation set due to lack-of-fit from the prediction.

Reference Data

Reference data determination is of course, a critical task. The reference data should be from a method that is reproducible with laboratory error of these values as low as possible. The reference data used for creation of a calibration model can be either a

chemical test or the result from a physical test, such as viscosity. In either case, the spectra must be related to the reference assay or no calibration will be possible. The best way to determine reproducibility of a method is through the use of replicate analysis by the reference method. Standard Error of Laboratory (SEL) should be calculated using these replicates. It is also used to benchmark the reference method against another laboratory or method so that an estimate of accuracy can be produced. Problems frequently arise when creation of models are attempted on reference data that contains substantial analytic noise. It is always best to verify what the reference method error is rather than assuming that it is acceptable. The plots in figure 2 show the effect of 10 and 20% error on a data set. If the component percentage of the sample is low, then some error might be tolerated, but calibration error will skyrocket if the weight percentage of the component and/or the working range is high and the method error is high. If reference error originates due to error between replicates then performing replicate determination and averaging the results can help to lower the overall laboratory error. However, if the error was caused by a reference method with low precision and accuracy, then performing replicate analysis will be of less benefit.

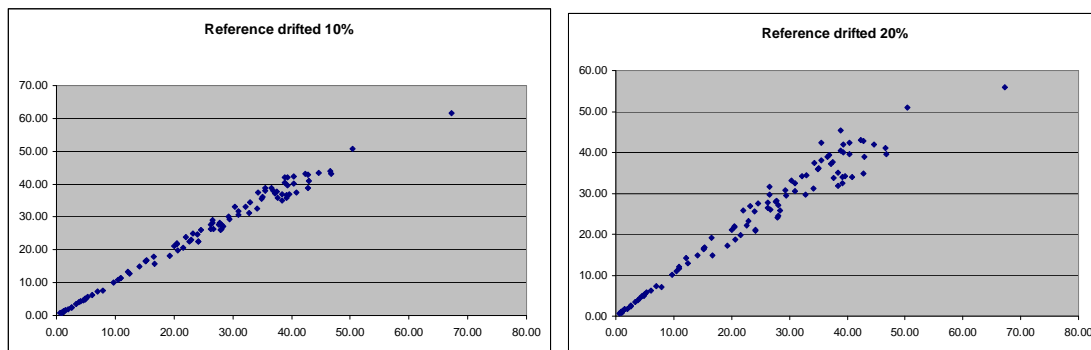


Figure 2, Reference data with 10 and 20% error

Once data has been generated for the calibration and validation samples, this data should be reviewed for cross correlation between the constituents. The easiest way to check for this is to plot constituent values for the same series of samples. The data plot

should resemble a shotgun pattern with no discernable trend. If a relationship is found to exist between the constituents, then additional samples should be found that do not conform to that trend.

Collection of Spectra

When collecting spectra for multivariate modeling it is important to control as many variables related to spectral acquisition as possible. System performance should be verified prior to collection of spectra. At a bare minimum the wavelength stability should be checked to assure that the system is functioning properly. If possible, the calibration samples should be ground or milled to the same particle size and homogenized prior to scan collection. If it is impractical to grind or mill to the same size, then a rotating sample presentation will likely provide the most reproducible spectrum. Sample illumination angles should be maintained throughout the spectral collection and while the calibration model that was based on these spectra are being used. If illumination angle changes, or particle size radically changes then the model will unlikely perform less accurately after these changes, unless the model has been stabilized for these effects. Temperature and moisture content are also important considerations. The calibration and validation samples should contain spectra collected on samples similar to the samples that will be analyzed with the developed model. Also if the form or particle size of the sample changes, then the model likely will not perform well until the calibration has been updated to include the new samples.

Creation of the Multivariate Calibration Model

Many chemometric programs are available to create the multivariate model. These programs combine the spectral data and reference data to produce a multivariate calibration model. GRAMS/AI with PLSPlus/IQ add-on, (Thermo, Woburn, MA), The

UNSCRAMBLER (CAMO Software, Inc, Woodbridge, NJ) and SIMCA P (San Jose, CA) are a few of the many programs commercially available. While the exact sequence differs from program to program, the basic steps remain the same.

1. Create a data file containing spectral and reference information.
2. Apply preprocessing to the spectra
3. Create experiment
4. Report the results of these experiments
5. Test the calibration using an independent test set of samples

Discussion of these tasks is beyond the scope of the paper and many books are available to explain these steps. The most important statistics that describe the calibration are standard Error of Cross-Validation (SECV) and Correlation coefficient – R-squared (RSQ). The number of factors (also called principal components and latent variables) used in the model are also important. As explained earlier there should be 10-20 samples in the calibration set for every one factor used in the model. Care should be taken when creating the calibration model to remove both spectral and concentration outliers that adversely affect the model. Samples may be removed from the model because their spectral characteristics are substantially different than the other samples in the model (spectral outliers) or because their reference data does not fit the relationship of the other samples (concentration outliers). In either case the goal is to remove as few samples as possible to produce the best calibration possible. Typically, if more than 10% of the samples are removed then the calibration developer might be trying to force a relationship where one does not exist. In that case when the independent test set is analyzed, the relationship of RSQ and standard error of prediction (SEP) of the test set would not be similar to the RSQ and SECV of the model. SECV is an estimate of SEP generated during the modeling process. Without the test set, the calibration developer would have little indication that the model would not work until it had been deployed (and failed). Samples removed as concentration outliers from the model should be rechecked by the primary assay if possible and should also have new spectra acquired for these samples if possible. If the reference data changes significantly following reanalysis, then the sample can be included back into the calibration set. If reanalysis

does not change the reference assay, then the calibration originator should investigate the cause for the lack of fit. For example, it might turn out that the sample had a very high or very low value for another constituent, which would indicate that additional samples of that type might be needed to help adequately describe the sample.

Samples that are removed as spectral outliers should also have new spectra collected if possible, as the leading cause of spectral outliers is improper spectral collection. If the spectra for the sample are recollected, yet still is marked as a spectral outlier then additional samples of that type should be identified and added to the calibration library to better represent that type. Outliers can tell the calibration developer important information, so they should not simply be discarded without additional investigation.

The partial least squares (PLS1) modeling technique is used to produce a quantitative predictive model. This technique simplifies the spectral data by producing a series of descriptive vectors called factors. The factors are then regressed against the reference data to produce the model. SECV is a validation technique that produces an estimate of the calibration error on unknown samples of similar type. In this technique, each sample is automatically removed, a calibration developed with the remaining samples and the error calculated on the removed sample. Then that sample is returned and the next removed. Finally, the error for each sample while it was not part of the calibration is calculated. If the modeling has been performed correctly, SECV should approximate the error of a true independent test set as measured with SEP. If SECV and SEP values are found to differ significantly, this would be an indication that too many samples were removed during the modeling process or that the model has been over-fitted.

Probationary Model Use

Once a model has been tested using an independent set of samples typically it can be used in a probationary status. New samples then are analyzed with the model and comparisons are made in the predicted results to the primary reference assay. If the

calibration error on these probationary samples meets the expectations for accuracy then the model can be used for any purpose.

Ongoing Monitoring of Model Performance

An NIR calibration will need to be updated on a frequent basis or when the prediction error rises beyond the acceptable limits. Most models will need updates within the first year, then less frequently as more samples are added and more variability has been explained by the model. The performance of all multivariate models should be continually verified using comparisons to the primary reference assays. If the spectral residuals are found to be increasing in the new samples, then investigation should be made as to the cause. Spectral residuals are measures of spectral “sameness” produced by the model. A high spectral residual that indicates that the sample is not well represented by the calibration model. If the near infrared (NIR) system has been verified using performance standards and has been shown to be performing properly, then it may be likely that something has changed in the samples such that they are no longer well represented in the calibration, perhaps due to a new type of sample that was not included in the current model. Additional samples of that particular type would need to be added to the calibration to assure adequate and proper representation in the model. If the concentration residuals were found to be increasing in the monitor samples, this too might indicate that the sample was not represented adequately by the model, or this might indicate that there had been a change in the reference method, such that it no longer was producing data that was similar to that used in the model.

Summary

When properly constructed and maintained multivariate models can provide near instantaneous predictions base on the spectra collected by NIR spectrometers. These models can be quite robust and they provide an indication of the overall suitability for a

particular sample (spectral residuals). Verification of system performance during spectral collection and prior to use of the developed model will help to reduce the error of the calibration to the lowest possible amount. The error of all models must be continually monitored and all models should be frequently updated, especially shortly after the model has been initiated, as it is quite common to discover that all sources of variation within the samples were not adequately explained in the original model. When samples do not fit the model it is important to investigate the causes of these discrepancies and update the model when it is appropriate to do so. NIR-based models can provide a productivity boost to many laboratories or operations, but cannot replace all reference assay analyses because these methods are still needed to update and test the calibration models. Careful application of NIR calibrations can provide more robust quantitative models than univariate or peak identification routines and may be useful in mining and remote sensing applications that have not been widely served by these techniques.

References

Mark, H. and Workman, J.J., Jr, *Chemometrics in Spectroscopy*, 2007, Academic Press, pp. 421-425

Workman, J.J., Jr., NIR Spectroscopy Calibration Basics. *Handbook of Near-Infrared Analysis*, ed. D.A. Burns and E.W. Ciurczak, 2008, CRC Press, pp.123-149.