MDPI

Article

Calibration, Validation, and Radiation: Predicting Biogenic Silica and Organic Carbon Percentages in Lake Sediment Core Samples

Rana Gahwagy[1,2,‡,*], Lauren Meyer[2,†,‡], Grace Hartley[2,†]

[1] Smith College Department of Statistical and Data Sciences Northampon, MA 01063; rgahwagy@smith.edu, lmeyer@smith.edu, ghartley@smith.edu

* Correspondence:

Version April 15, 2022 submitted to Water

Abstract: In many settings, biogenic silica (BSi) and total organic carbon (TOC) are widely used as proxies for temperature and/or environmental variations that are helpful in paleoclimate and paleoenvironmental reconstructions. Often, the methodology for analyzing these parameters in sediments can be expensive and time consuming (particularly for BSi). However, Fourier Transform Infrared (FTIR) Spectroscopy offers an efficient alternative where many samples can be run with minimal amount of sediment and time. This technique is advantageous in that it requires small volumes of sediment (~0.01g), minimal sample preparation (mixing sample with potassium bromide powder), and instrumental analysis times are relatively rapid (a few minutes per sample). FTIR Spectroscopy quantifies BSi and TOC using infrared radiation (IR) absorbance units—as opposed to percentages of BSi or TOC—which are difficult to compare across different studies and localities. Therefore, there is a need for a systematic way to convert the results from the FTIR Spectrometer into percentages. In this research project, we address this need by building a universal calibration model using partial least squares (PLS) regression that converts BSi absorbance to percentages. We developed this model using a PLS package in R and based our model on samples from Arctic lakes in Greenland and Alaska. Our preliminary model uses a k-fold cross-validation method and utilizes three components. Ongoing work intends to improve on the model's prediction accuracy, expand our calibration model to include TOC percentages, and incorporate more BSi samples from other locations. We aim for the model to be universal and integrated into a Shiny app, where paleoclimatologists can use it on samples from various localities and compare their results. This model will prove a valuable tool in paleoclimate reconstruction by facilitating FTIR Spectroscopy on lake sediments.

1. Introduction

Studying the content of biogenic silica and other organic compounds present in lake sediment cores can be a powerful tool in gaining insight into our reconstruction of past climates. The wet chemistry processes used to determine these proportions can be time consuming and costly. As a result, data on the amount of biogenic silica present in different samples is limited in quantity and resolution. Fourier-transform infrared (FTIR) spectroscopy is a promising technique to reduce the time and money needed to determine the proportion of these compounds in lake sediments. FTIR spectroscopy involves measuring the absorbance of the infrared radiation at different wavelengths. These absorbance values are arbitrary and unitless Kamat and Schatz [1], so interpretation is needed to make them relevant between researchers. To interpret these absorbance values, we seek to use a Partial Least Squares Regression (PLSR) model to predict the percentage of the biogenic silica in each sample. This technique has been pioneered successfully by Vogel *et al.* [2], but is not accessible to those

who might wish to use FTIR spectroscopy for their samples and lack the statistical background to implement a PLSR model.

The goal of this project is to create an interface where a user can input their FTIR spectroscopy data into a PLSR model to calculate the approximate biogenic silica and total organic carbon percentages. Our work primarily includes improving the accuracy of the existing model (https://github.com/people-r-strange/PLSmodel) and preparing for universal input. This includes selection of the most applicable diagnostic plots to determine the accuracy of the model. Because this work will be accessible to the public, our model and corresponding interface satisfy an existing need for efficient FTIR spectroscopy interpretation. Further, greater accessibility and use of this technology may eliminate the need for expensive wet chemical processes.

2. Data

Our model is trained on spectroscopy data gathered from the analysis of lake bed samples. 26 samples are from Greenland collected by Greg de Wet, 103 are from Alaska collected by Daryl Kaufman at Northern Arizona University, and samples of only diatoms and only washed quartz. Each pre-processed dataset has two variables: wavenumbers tested, and corresponding absorbance values measuring the sample's absorbance of that specific wavenumber of light (Figure 1). Each sample also has a single associated measurement of biogenic silica, calculated by traditional wet chemistry digestion methods.

```
## Warning: Removed 1814 row(s) containing missing values (geom_path).
```

The Greenland samples were tested at 3,697 distinct wavenumbers, ranging from 368 cm^{-1} to 7497 cm^{-1}, while the Alaska samples were tested at 1,882 distinct wavenumbers ranging from 368 cm^{-1} to 3996 cm^{-1}. The relationship between absorbance and wavenumber is smooth for all samples between the wavenumbers of 500 cm^{-1} and 4000 cm^{-1}, though below 500 cm^{-1} and above 4000 cm^{-1} the line shows increased noise, with the noise being the worst at the highest wavenumbers. This is evident in figure 2, and even more notably in the higher wavenumbers of the wet quartz data in figure 3.

The Greenland samples and the Alaskan samples were analyzed at different times with slightly different spectroscopy settings, so the wavenumbers measured do not match exactly. Even though the difference is small, the regression model requires the wavenumber labels to be consistent as they are the explanatory variables in the model. In addition, the Alaskan samples were analyzed on a much smaller range of wavenumbers than the Greenland samples, though the resolution was about the same. Therefore, the two sets of data have a mismatch in the number of columns, which we resolve using linear interpolation and elaborate below.

3. Methods

We use functions from the `pls` (https://CRAN.R-project.org/package=pls) R package to create and evaluate our PLSR models. The original model was created using this same package. In order to run the PLSR model, we need all datasets to include the same amount of wavenumber columns with identical labels. This is because each wavenumber is treated as an independent variable in the model, and prediction is not possible if each sample has slightly different independent variables.

In order to solve this issue, we will linearly interpolate the absorbance spectrum from each Greenland sample to match the wavenumbers of the Alaskan samples, rounded to the nearest integer for ease of human interpretation. This is because the range of wavenumbers measured in the Alaskan samples is smaller, therefore interpolating Alaskan samples onto the Greenland spectra would not be possible. Since the absorbance spectra from both samples are high resolution and the differences between the original and the interpolated absorbance values are small, we are comfortable that this does not meaningfully alter the data.

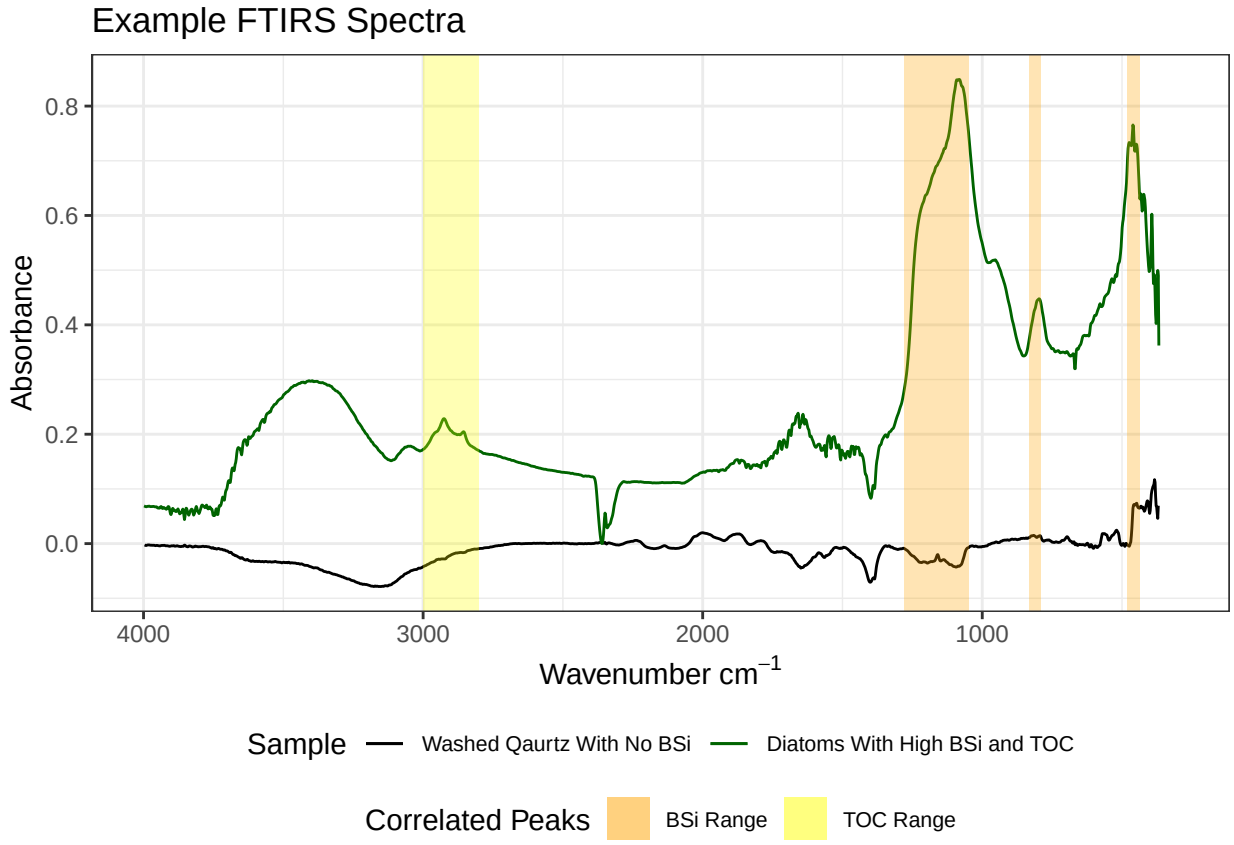

Figure 1. Comparison between absorbance values across the spectrum of wavelength for two calibration samples, one of pure diatoms, which have high biogenic silica (BSi) and total organic carbon (TOC) content, and one of washed quartz which has no BSi or TOC. Highlighted areas are ranges that correlate with BSi (in orange) and TOC (in yellow). The area under those peaks is what is usually reported in literature.

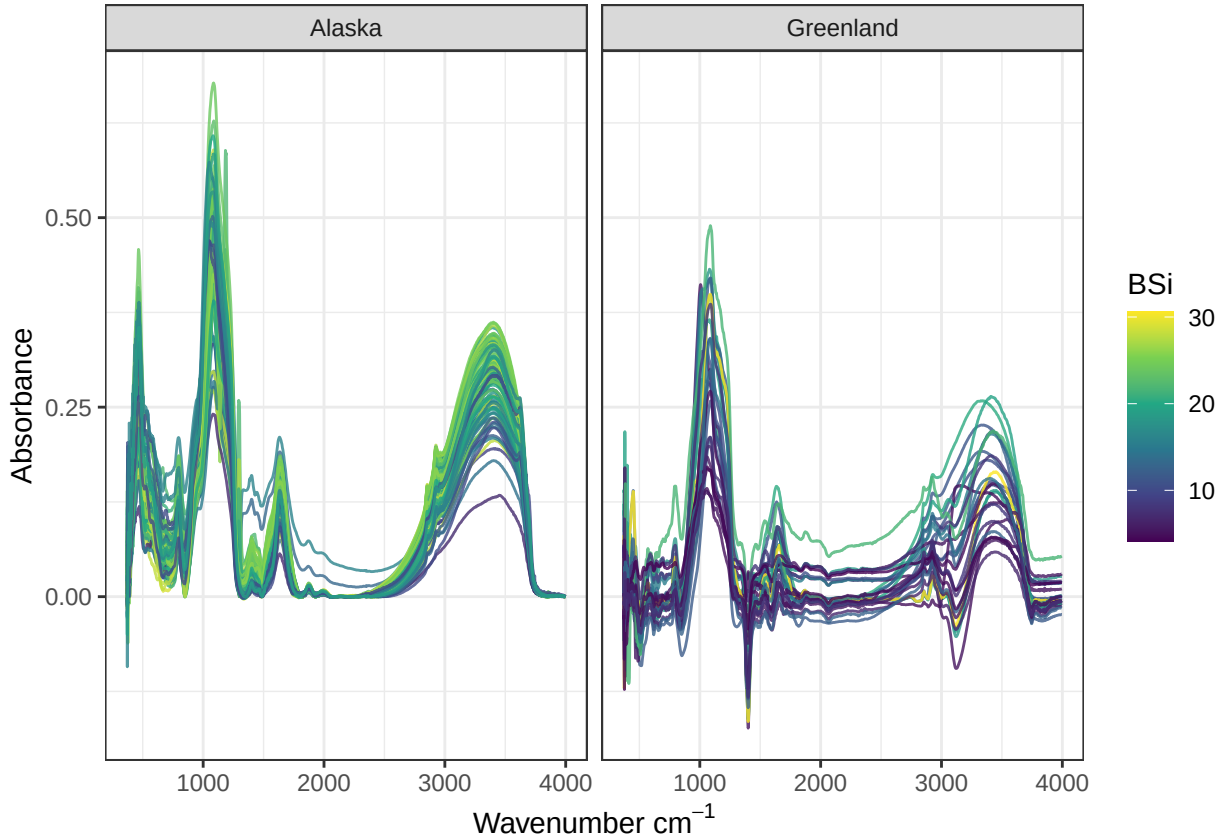

Figure 2. Each sample's spectrum is plotted here by region and color coded with actual biogenic silica (BSi) where dark blue colors are samples with low BSi and yellow denotes high BSi.

The original model was trained on only the Greenland samples (n = 28) to predict BSi content within the Greenland samples. To improve accuracy of the model, we integrate the Alaskan samples (n =103). In total, we built five different models to gauge the best performance with our data.

The first two models are region-specific; this includes the original Greenland model mentioned above, as well as a similar model that uses the Alaskan samples only. The next model uses the full spectrum of combined Greenland and Alaska samples. The last two models use sections of the combined Greenland and Alaskan spectra that are typically correlated with high BSi content. One of the two models includes the most reported peak between 1050-1280 cm^{-1}, and the other considers all associated peaks between 1050-1280 cm^{-1}, 790 - 830 cm^{-1}, and 435 - 480 cm^{-1} as highlighted in figure 1 Rosén *et al.* [3].

To compare the accuracy of these models, we look at the Mean Square Error (MSE) and Mean Absolute Deviation (MAD) values from each model. We employ assessment tools to determine the optimal number of components, notably Root Mean Squared Error of Prediction (RMSEP). These values are useful, because they are a simple, comparable metric. A good model will minimize the size of prediction errors, leading to a small RMSEP.

4. Results

Comparing the different models based on the metrics mentioned above, we find that the model that is trained on the full spectrum of the combined samples (Greenland + Alaskan) predicts BSi content the best. This is visible in table **??**, where we see that the combined full spectrum model has a Mean Squared Error of 6.84, which is considerably lower than even the next best model that has an MSE of 12.33.

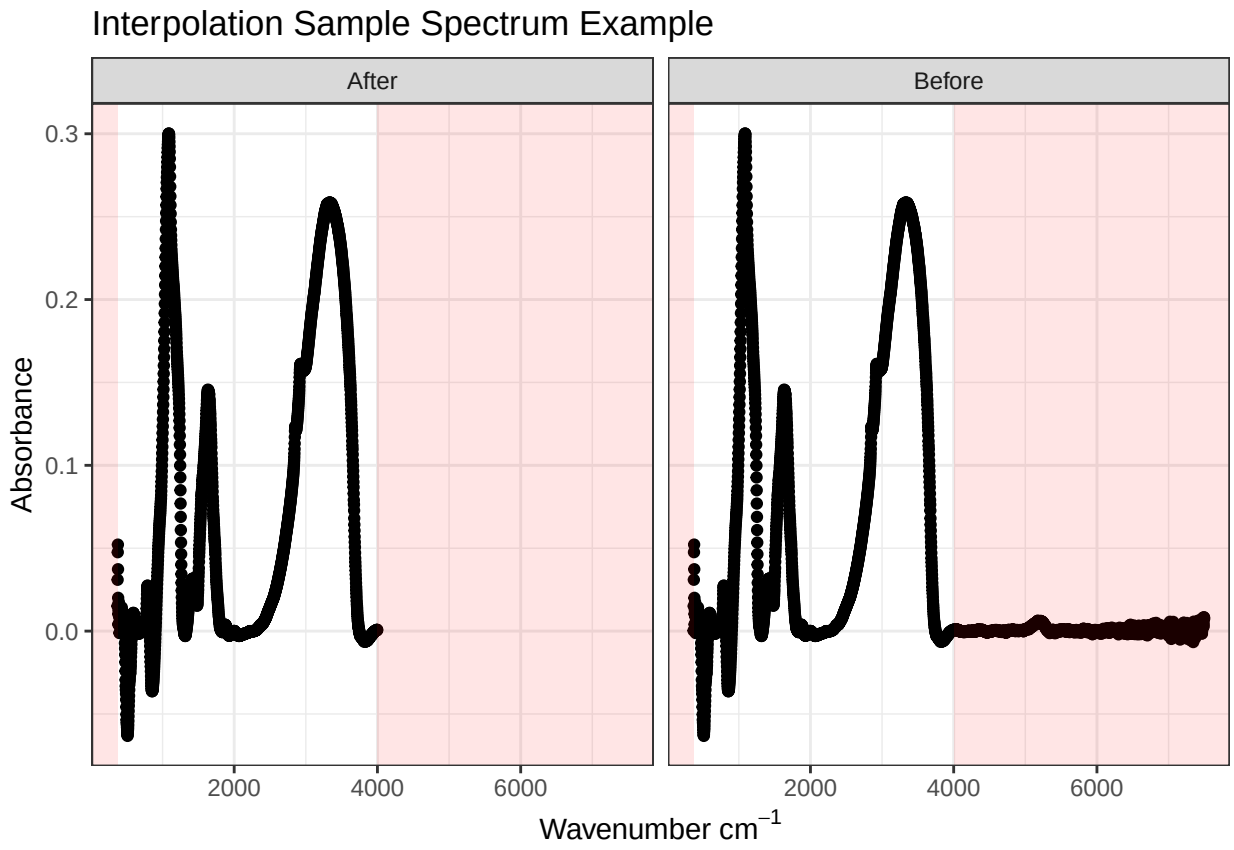

Figure 3. A closer look at the Absorbance spectrum of a random sample from Greenland shows that the raw data (Before) has a lot of noise in wavenumbers above 4000 cm^{-1}, which is consistent throughout all samples. After truncating the spectrum and interpolating the Greenland data (After) there were negligible changes.

Table 1. Overview of the five different models tested against each data set. The Mean Square Error (MSE) and Mean Absolute Deviation (MAD) were calculated to compare and the model that uses both the Alaska and Greenland full spectrum shows optimal results.

Data Trained On	Data Tested On	MSE	MAD
Greenland	Alaska	61.78174700958	6.97577898211303
Greenland	Combined	47.0020588160737	5.64642471647423
Alaska	Greenland	307.56239102917	14.9424681723402
Alaska	Combined	71.3391970518575	4.7130242367885
Combined	Combined	6.84194528643838	2.02673866084178
Limited - One Segment	Combined	12.3313102356506	2.84803813321514
Limited - Multi-Segment	Combined	13.2485397696818	2.78010069524131

Once we have the most appropriate model, we must choose how many principal components to include within the model. Principal components refer to latent variables that the algorithm predicts within the model to use for prediction of the desired outcome variable (pls package citation). Including too few principal components will not provide enough predictive power, but too many risks an overly complex or overfitted model. We look at a plot (figure 4) of RMSEP values for each potential number of components (n = 10), and find that using 10 components creates a model with the lowest RMSEP. Despite the high number of components, our model does not appear to overfit our data because it accurately generalizes to new data that was not included in the training data. Overall, our final model can be represented as: $BSi = x_1 + x_2 + ... + x_{1882} + b$ where the x_n variables are the individual wavenumbers and b is the intercept.

wordcountaddin::text_stats()

Abbreviations

The following abbreviations are used in this manuscript:

FTIR	Fourier-Transform Infrared
PLSR	Partial Least Squares Regression

References

1. Kamat, P.; Schatz, G.C. How to make your next paper scientifically effective, 2013.
2. Vogel, H.; Rosén, P.; Wagner, B.; Melles, M.; Persson, P. Fourier transform infrared spectroscopy, a new cost-effective tool for quantitative analysis of biogeochemical properties in long sediment records. *Journal of Paleolimnology* **2008**, *40*, 689–702. https://link-springer-com.libproxy.smith.edu/article/10.1007/s10933-008-9193-7, doi:10.1007/s10933-008-9193-7.
3. Rosén, P.; Vogel, H.; Cunningham, L.; Hahn, A.; Hausmann, S.; Pienitz, R.; Zolitschka, B.; Wagner, B.; Persson, P. Universally applicable model for the quantitative determination of lake sediment composition using Fourier transform infrared spectroscopy. *Environmental science & technology* **2011**, *45*, 8858–8865.

© 2022 by the authors. Submitted to *Water* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

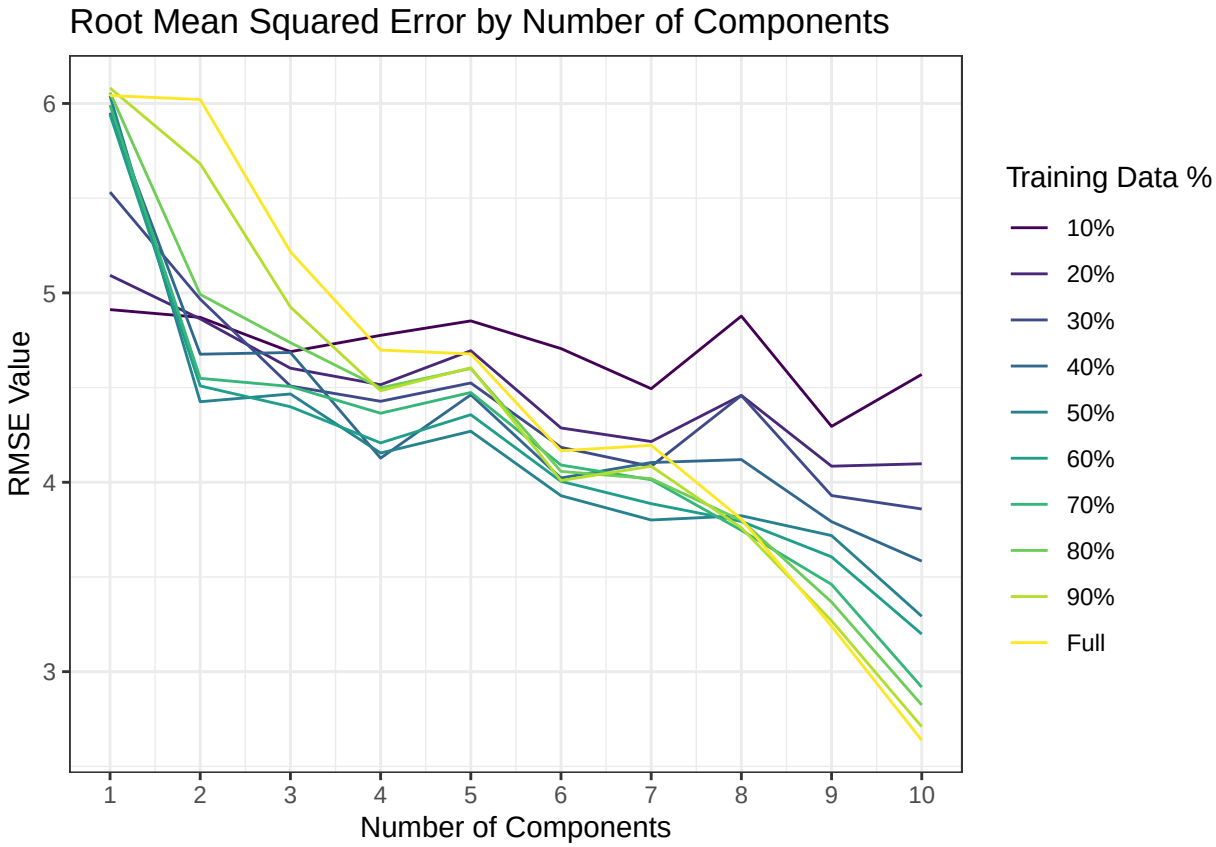

Figure 4. The Root Mean Square Error of Prediction (RMSEP) was calculated against new data for each component. Ten random samples were taken of the data for each percentage, and the model trained on that sample was then tested against the remaining data for each component, with the values shown above being the mean of the ten iterations for each combination, to limit the variation inherent in sampling. Based on this, we chose the model with 10 components as our final model.