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 14, 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, 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.* [1], 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.

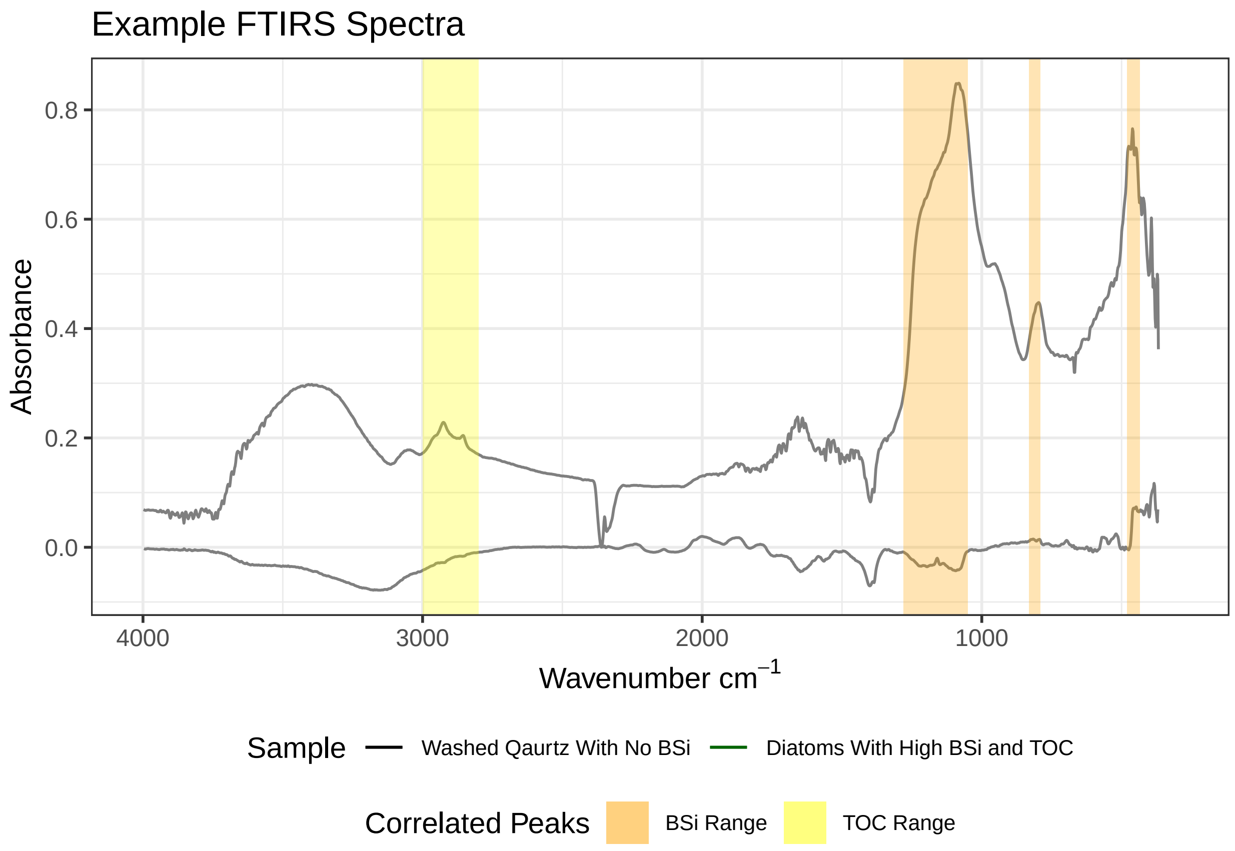

Figure 1. Comparison between absorbance values across the spectrum of wavelength for two calibration samples, one of pure diatoms, which have high biogenic silica (BSi) and total organic carbon (TOC) content, and one of washed quartz which has no BSi or TOC. Highlighted areas are ranges that correlate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

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

The model runs on spectroscopy data gathered from the analysis of lake bed samples. 26 samples are from Greenland, 100 are from Alaska, and two are specially made calibration samples. Each sample has two dimensions: a list of wavelengths tested, and a corresponding list of absorbance values measuring the sample's absorbance of that specific wavelength of light (figure 1). Each sample also has a single associated measurement of biogenic silica, calculated by traditional wet chemistry.

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

The Greenland samples were tested at 3,697 wavelengths, from 368 cm^{-1} to 7497 cm^{-1}, while the Alaska samples were tested at 1,882 wavelengths from 368 cm^{-1} to 3996 cm^{-1}. The relationship between absorbance and wavelength is smooth for all samples between the wavelengths of 500 cm^{-1} and 4000 cm^{-1}, though below 500 cm^{-1} and above 4000 cm^{-1} the line shows increased noise,

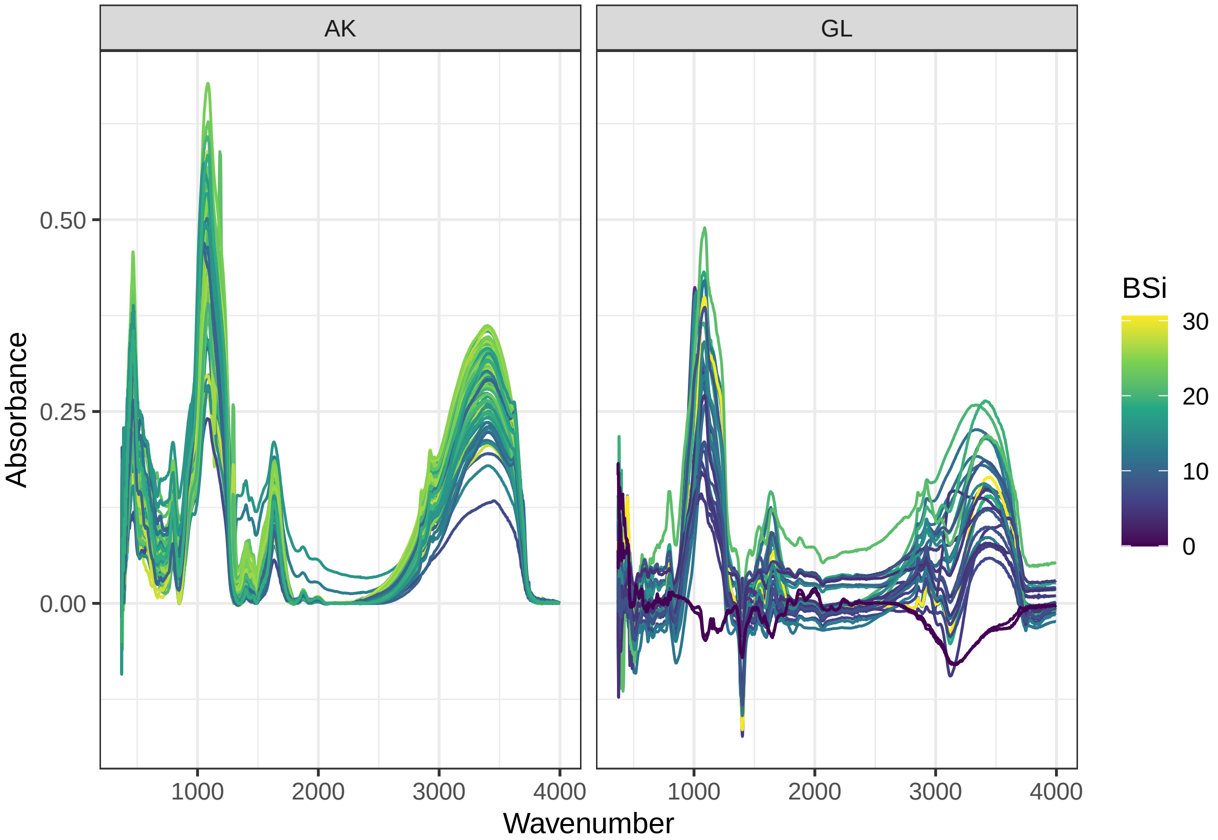

Figure 2. Each sample's spectrum is plotted here by region and color coded with actual biogenic silica (BSi) where dark blue colors are samples with low BSi and yellow denotes high BSi.

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 wavelengths measured do not match exactly. The difference is small to the point of negligence, but the regression model requires the wavelength labels to be consistent. In addition, the Alaskan samples were analyzed on a much smaller range of wavelengths than the Greenland samples, though the resolution was about the same. To solve both problems at once, we linearly interpolate the absorbance curve from each Greenland sample to match the wavelengths of the Alaskan samples, as those do not vary. The absorbance curves from both samples were high enough resolution and the differences in wavelengths measured were small enough that we are comfortable that this did not meaningfully alter the data in any way, besides to make it all fit seamlessly in the same model. These minute differences in wavelengths is seen in figure 4, where the interpolated data (red) is mapped on top of the original data (black). Since the black points are obscured behind the red dots, we see that the differences in wavenumbers is negligible.

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

Data Trained On	Data Tested On	MSE	MAD
Combined	Combined	6.84194528643838	2.02673866084178
Limited - One Segment	Combined	12.3313102356506	2.84803813321514
Limited - Multi-Segment	Combined	13.2485397696818	2.78010069524131

wordcountaddin::text_stats()

Abbreviations

The following abbreviations are used in this manuscript:

FTIR	Fourier-Transform Infrared
PLSR	Partial Least Squares Regression

References

1. 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.

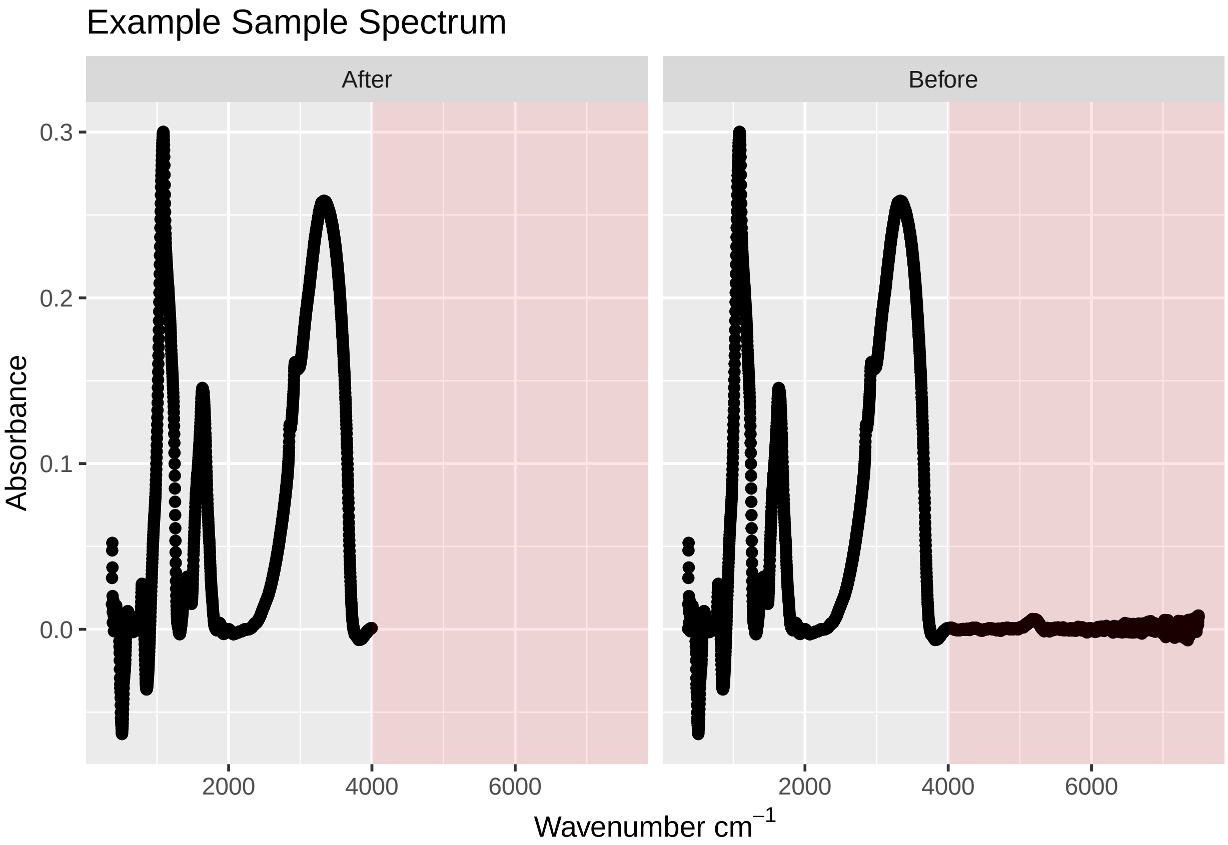

Figure 3. A closer look at the Absorbance spectrum of a random sample from Greenland shows that the raw data (Before) has a lot of noise in wavenumbers above 4000 cm^{-1}, which is consistent throughout all samples. After truncating the spectrum and interpolating the Greenland data (After) there were no observable changes.

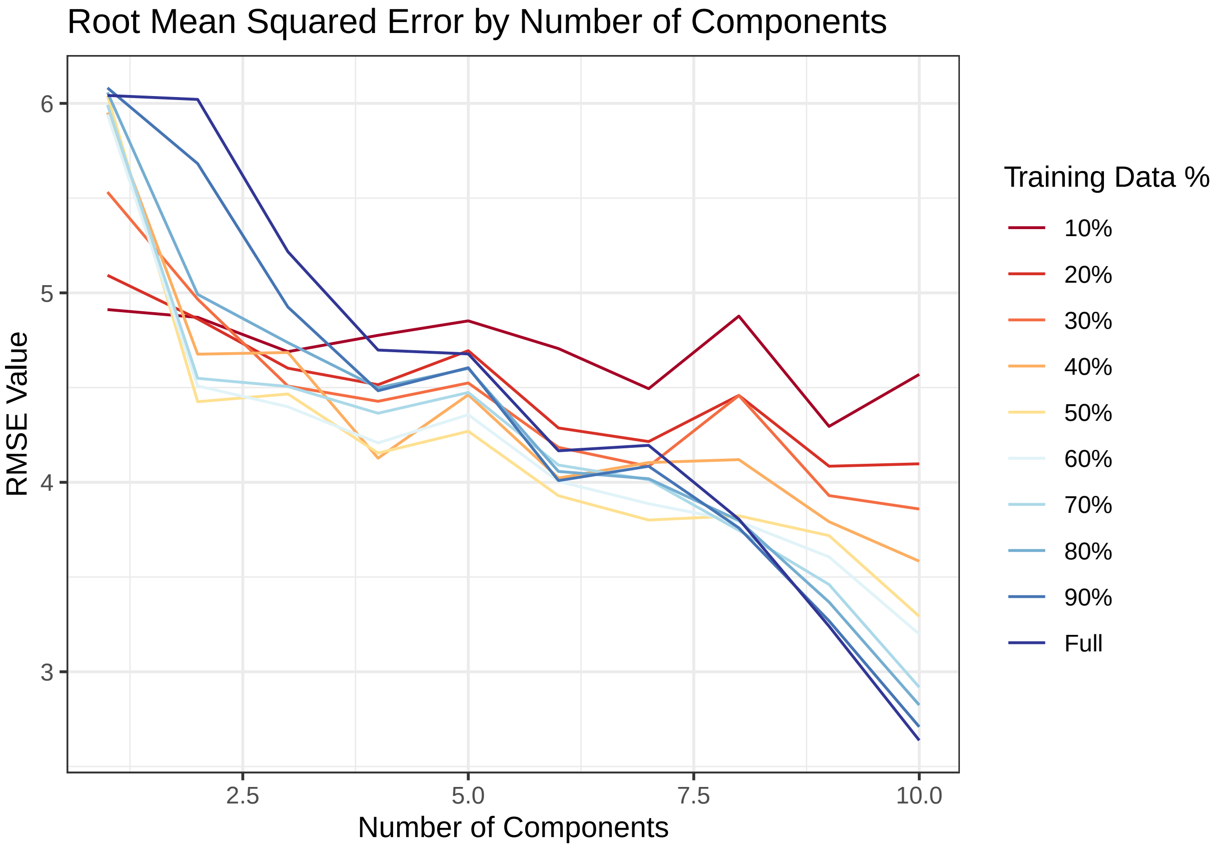

Figure 4. The Root Mean Square Error of Prediction (RMSEP) was calculated against new data for each component. Ten random samples were taken of the data for each percentage, and the model trained on that sample was then tested against the remaining data for each component, with the values shown above 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.