MDPI

Article

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

Rana Gahwagy[1], Lauren Meyer[1], Grace Hartley[1]

[1] Smith College Department of Statistical and Data Sciences Northampon, MA 01063;

* Correspondence:

Version May 3, 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 the `pls` package [1] 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

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 normally 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 [2], so interpretation is needed to make them relevant between researchers. To interpret these absorbance values, we 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.* [3], 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.

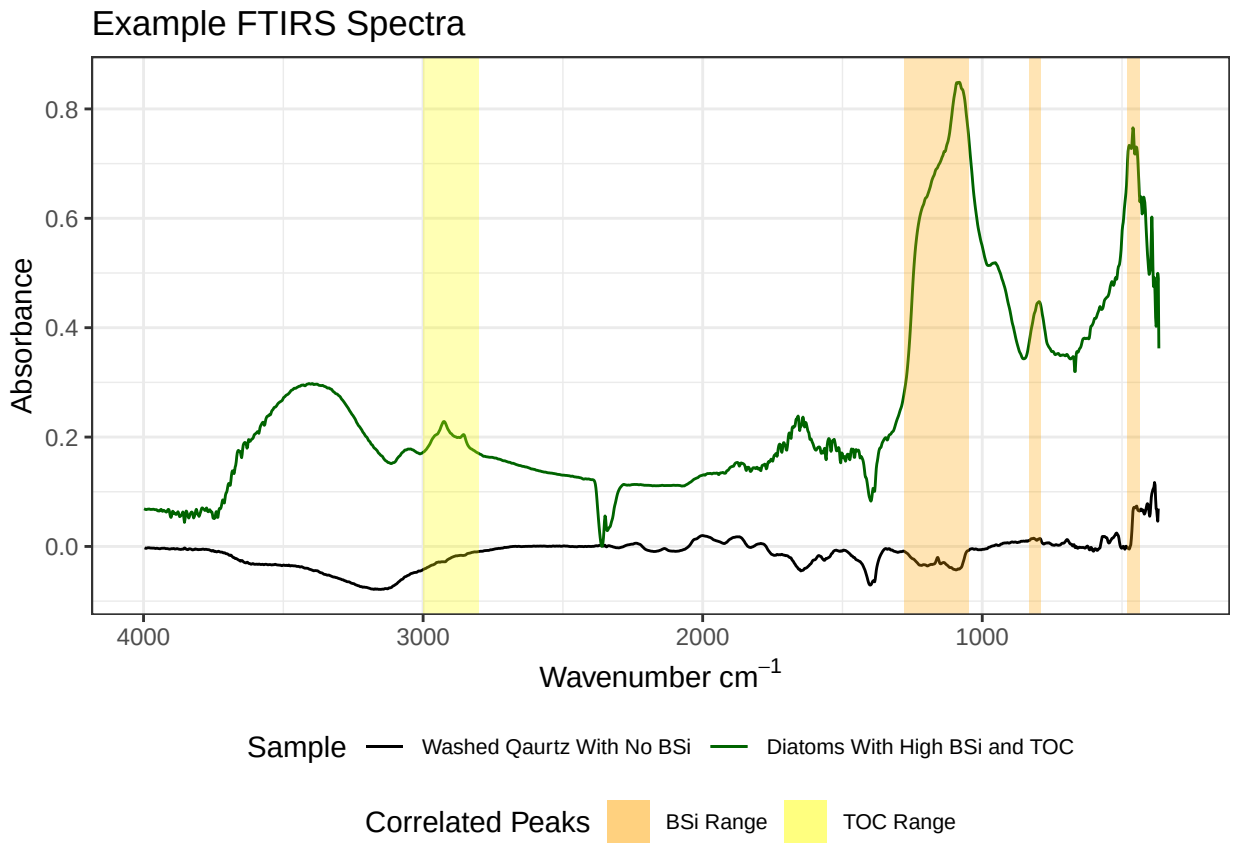

Figure 1. Comparison between absorbance values across the spectrum of wavelength for two calibration samples, one of pure diatoms, which have high biogenic silica (BSi) and total organic carbon (TOC) content, and one of washed quartz which has no BSi or TOC. Highlighted areas are ranges that correlate 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

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 collected from Alaska by Daryl Kaufman (Northern Arizona University), and samples of only diatoms and only washed quartz are used for calibration. 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 (measured in percentage by weight), calculated by traditional wet chemistry digestion methods.

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}. However, below 500 cm^{-1} and above 4000

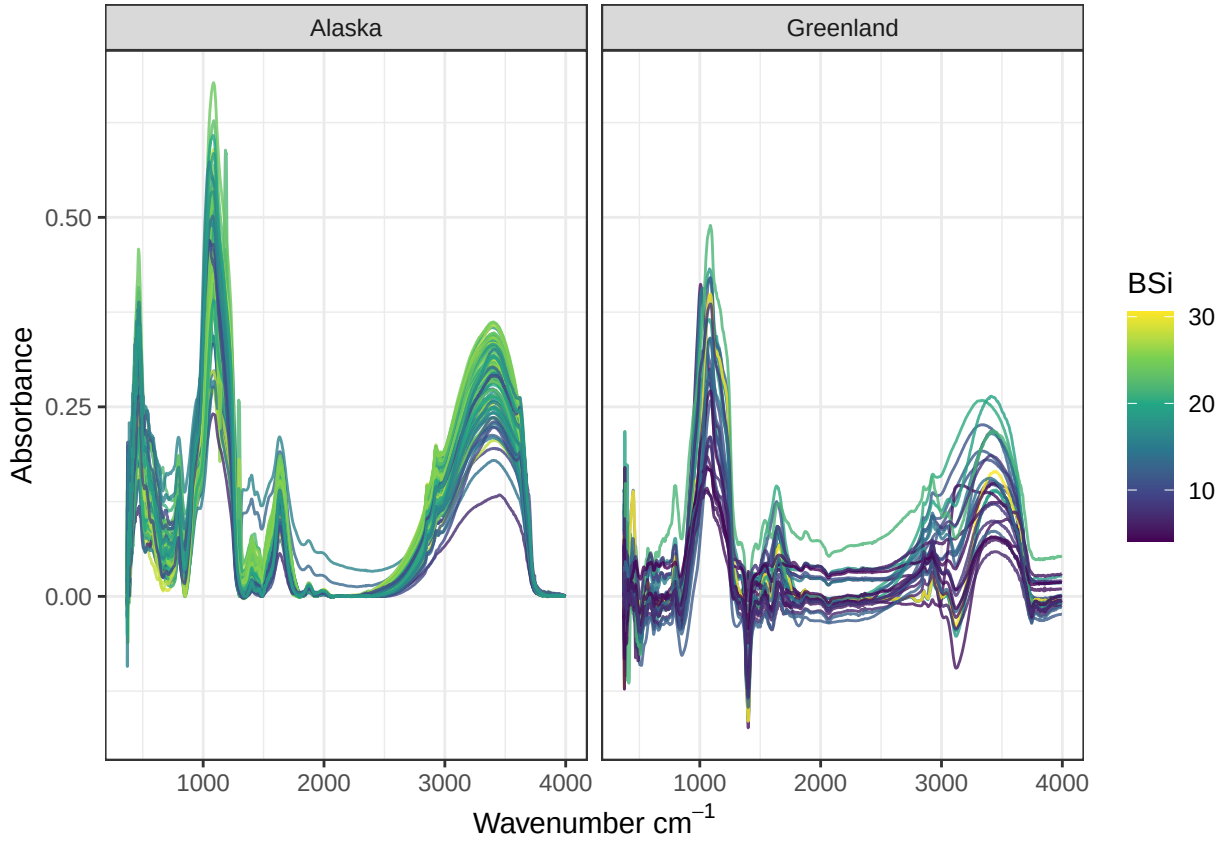

Figure 2. Each sample's spectrum is plotted here by region and color coded with actual biogenic silica (BSi) where dark blue colors are samples with low BSi and yellow denotes high BSi. Note that samples with higher BSi tend to have higher absorbance values.

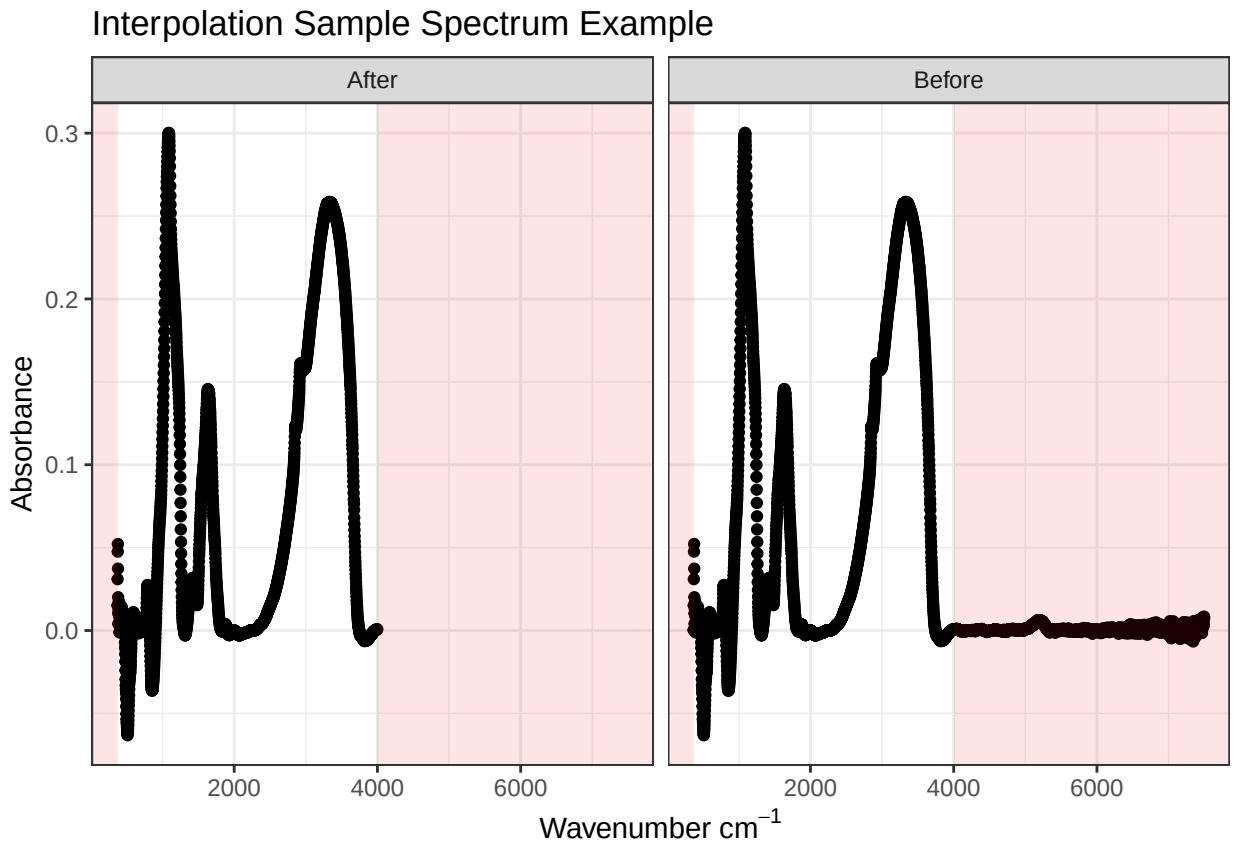

Figure 3. A closer look at the Absorbance spectrum of a random sample from Greenland shows that the raw data (Before) has a lot of noise in wavenumbers above 4000 cm^{-1}, which is consistent throughout all samples. After truncating the spectrum and interpolating the Greenland data (After) there were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

cm^{-1} the relationship disintegrates, with the signal 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 and 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, our 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 narrower range of wavenumbers than the Greenland samples, although the resolution was about the same. Therefore, the two sets of data have a mismatch in the number of explanatory variables, which we resolve using linear interpolation (see Section 3.1).

3. Methods

We chose PLSR for this modeling task for several reasons. Traditional linear regression models require a specific set of conditions to be satisfied in the data, and once those conditions no longer hold the results of the model are no longer necessarily accurate. The main issue with our data in regards to traditional multiple regression is that our independent variables are not truly independent of each other, and are instead highly collinear. Absorbance measurements at adjacent wavenumbers will nearly always vary together. Correlated independent variables in multiple regression can cause overestimates of the model's predictive power, while the actual model fails to perform to these estimates.

The second issue with using multiple regression for this dataset is that the independent variables far outnumber the samples, causing multiple regression models to overfit to the existing data and lose

the ability to generalize. Given that the goal of this model is to predict new data, losing predictive power in this way is not ideal. PLSR solves both of these problems using a form of dimension reduction, where it predicts a few internal "latent variables," or components, using highly correlated variables from the original data, and then predicts the dependent variable based on those components, functionally aggregating all of the collinear variables into a single independent variable. This eliminates the covariance and collinearity and also reduces the actual number of independent variables to the point that multiple regression is appropriate [4].

We use functions from the `pls` R package to create and evaluate our PLSR models [1]. In order to run the PLSR model, we need all datasets to include the same number of wavenumber measurements 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 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 narrower, 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.

The "Maxwell 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 Maxwell model mentioned above, trained only on the Greenland samples, as well as the "Alaskan model" that uses the Alaskan samples only. The "Full 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. The "Limited Spectrum model" includes only the most reported peak between 1050-1280 cm^{-1}, and the "Segmented Spectrum" model considers all associated peaks between 1050-1280 cm^{-1}, 790 - 830 cm^{-1}, and 435 - 480 cm^{-1} as highlighted in figure 1 [5].

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 Full Spectrum model 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 percent BSi squared, which is considerably lower than even the next best model that has an MSE of 12.33.

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	61.78	6.98
Greenland	Combined	47.00	5.65
Alaska	Greenland	307.56	14.94
Alaska	Combined	71.34	4.71
Combined	Combined	6.84	2.03
Limited - One Segment	Combined	12.33	2.85

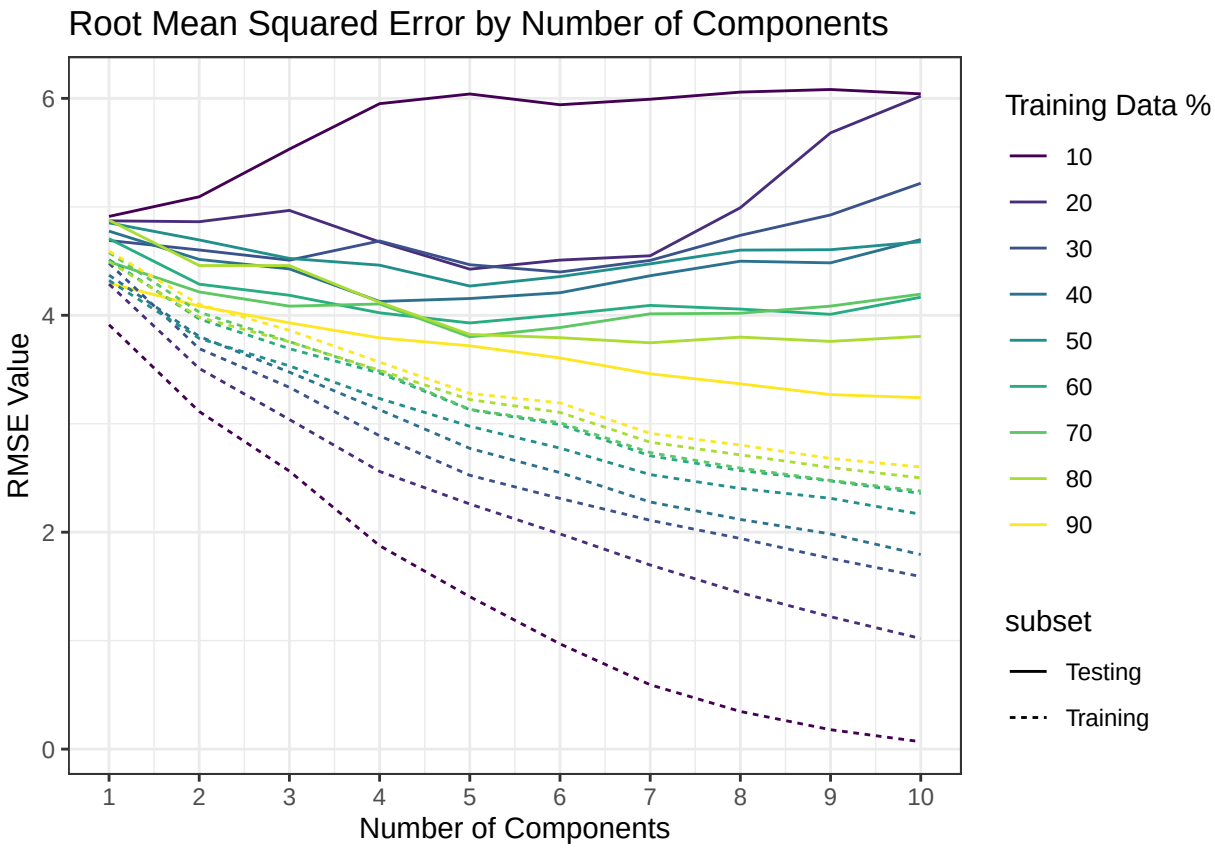

Figure 4. The Root Mean Square Error of Prediction (RMSEP) was calculated against new data for each component. Ten random samples were taken of the data for each percentage, and the model trained on that sample was then tested against the remaining data for each component, with the values shown above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

Data Trained On	Data Tested On	MSE	MAD
Limited - Multi-Segment	Combined	13.25	2.78

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 [1]. 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 5 components creates a model with the lowest RMSEP. The model does not appear to overfit our data because the cross-validated error remains small as components are added. Overall, our final model can be represented as:

$$BSi = \beta_0 + \beta_1 x_1 + x_2 + \cdots + \beta_{1882} x_{1882} + \varepsilon$$

where the x_n variables are the individual wavenumbers, β_0 is the intercept, and $\varepsilon = N(0, \sigma_\varepsilon)$ for some fixed σ_ε.

With our model and number of principal components in place, our model is ready to predict percent BSi from FTIR Spectroscopy Data. To facilitate this, we create an app **??** using R Shiny [**?**]. The convenient and accessible interface allows users to select a directory on their local computer that contains the FTIR Spectroscopy data for their samples. The predicted percent BSi in each sample,

estimated by our Full Spectrum model, is outputted to the screen, with an option to download these values.

5. Conclusion

5.0.1. Ethical Concerns

There are several ethical concerns that our project raises, regarding both the data we use and the model we have created. Because we did not collect the lake sediment core samples ourselves, we do not know the conditions that they were collected under. Specifically, we do not know what sort of relationship the data collectors have with the people who live where these samples were collected, nor if they had permission to take samples. We assume that best protocol was followed, but should investigate further to confirm that these samples were ethically collected. We also should wonder– are there environmental concerns to this practice? Does conducting the wet chemical digestion process leach harmful chemicals? Does removing the cores disrupt any natural life or processes? Are the carbon emissions of American geologists flying to Greenland to collect samples negligible? Potentially not. If our model brings accessibility to assessing biogenic silica composition in these samples, perhaps it will encourage researchers, who might not otherwise collect samples in the field, to do so. Since our work makes the work of paleoclimatologists more cost-effective, these concerns are potentially amplified.

If our work draws more paleoclimatologists to use FTIR Spectroscopy and subsequently, our PLSR model, we must consider the ethical implications of allowing our model to be used by scientists whose statistical background might be limited. Since our model provides an easy-to-use interface that requires little to know understanding of coding or statistical modeling, it would be easy for someone without the background to draw false conclusions. An uninformed person could feed their samples into our model, alter it slightly, and draw conclusions that best fit their interest, disregarding the appropriateness. For instance, our interface shows the model with up to 10 components (which demonstrates that the model is not a "final answer", but a *model* that holds room for error). A user could choose to look at the number of components that maximizes their results and provides the highest R^2 value, without really understanding what either of these values indicates.

5.0.2. Limitations

Our model faces several limitations that are mostly bounded by the small number of samples we have. The error rate and standard deviation for the model are still high and do not allow for very precise prediction. Since we only have a small sample size, we have yet to test and validate the final model extensively using new data. Most of the testing has been done using pre-existing data, but since collecting new training data with wet chemical digestion is costly, we rely on collaborations from outside sources to obtain new data. In addition to new data, we also need diverse data from different regions of the world in order to know if we can generalize our model and if it can be used universally. All of our current data is from the arctic region, but we expect different relationships in BSi occur in sample from different regions such as temperate, tropical or marine regions. For example, scientists analyzing Lake Malawi found BSi profiles in tropical regions are less interpertable than in Northern regions (Johnson *et al.* [6]). Currently, the model only accounts for samples from the Arctic region and could inaccurately predict BSi percentages for other regions. Our model is also contingent on the users of the model who are not familiar with PLSR being able to select the best number of the components and interpret the model, though this problem may be minimized by preselecting the best number of components for them.

5.0.3. Future Work

As we note above, our project was bounded by limitations that future work could eliminate. Future work can be done to improve our model itself and our Shiny App interface. First, future work

that incorporates samples from non-arctic locations could allow this to be a more universal model and explore whether it is best to account for regions within the model as an additional variable. We also would like for the model to predict percent Total Organic Carbon found in samples as it is a widely used measurement and detectable by FTIR Spectroscopy. Overall, future work should aim to incorporate more data (from arctic and non-arctic locations alike) and conduct a more robust investigation into the best number of components to be used in the PLSR model. These new findings should be reflected in the Shiny App interface. The interface should be able accept more file types than .csv alone, particularly .dpt and .txt files since . Ideally, we hope to see a powerful Shiny interface that allows a user to select which model they would like to predict with, provides summary diagnostic statistics of each model (such as R^2, RMSEP), visualizations to assess model performance, and background information that contextualizes prediction results.

5.0.4. Final Thoughts

The work presented in this paper has been a very substantial step towards the goal of developing a baseline universal calibration model to predict natural compounds in lake sediment cores. We have laid the layout and provided room for improvement, but there is much work to be done. Incorporation of more data will reduce the error margins of the model, which could render a useful tool for paleoclimatology research in past climate reconstruction by using biogenic silica as a proxy .

5.0.5. Acknowledgements

Our work would not be possible without several individuals and groups. We would like to thank the entire de Wet Lab at Smith College for their contributions and samples, as well as Daryl Kaufman at Northern Arizona University for providing us with the Alaskan sample data that was necessary to build our model. We hold much gratitude for Vivienne Maxwell and Professor Sara Stoudt for paving the way with their code that served as the framework of this project. Finally, we extend great thanks to Professor Greg de Wet and Professor Ben Baumer for their extensive support and guidance throughout the course of this project.

Abbreviations

The following abbreviations are used in this manuscript:

FTIR	Fourier-Transform Infrared
PLSR	Partial Least Squares Regression

References

1. Liland, K.H.; Mevik, B.H.; Wehrens, R. *pls: Partial Least Squares and Principal Component Regression*, 2021. R package version 2.8-0.
2. Kamat, P.; Schatz, G.C. How to make your next paper scientifically effective, 2013.
3. 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.
4. Helland, I.S. Partial Least Squares Regression and Statistical Models. *Scandinavian Journal of Statistics* **1990**, *17*, 97–114.
5. 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.
6. Johnson, T.C.; Brown, E.T.; Shi, J. Biogenic silica deposition in Lake Malawi, East Africa over the past 150,000 years. *Palaeogeography, Palaeoclimatology, Palaeoecology* **2011**, *303*, 103–109.