Geostatistics

A Partial Least Squares Calibration Model for Predicting Biogenic Silica

Vivienne Maxwell

5/12/2021

Contents

I. Purpose

The aim of this project is to create a partial least squares (PLS) calibration model to facilitate paleoclimatologists in converting their raw absorbance spectral data into percentages of Total Organic Carbon (TOC) and Biogenic Silica (BSi) in lake core samples. The project herein described focuses mainly on developing a Partial Least Squares Regression model and cross-validating lake core samples. Future work encompasses (1) writing a PLS package and accompanying vignette for RStudio and (2) creating an interactive dashboard where paleoclimatologists can input their raw spectral data, adjust for certain parameters, and collect predicted percentage values.

General Topic/ Phenomenon of Study

In paleoclimatology, lake cores are used as high quality archives to study past climates. High values of biogenic silica (BSi) and total organic carbon (TOC) often indicate that the environment in which the cores were collected experienced high temperatures. This information is helpful when reconstructing past climates in arctic settings. Currently, paleoclimatologists have several ways of measuring BSi and TOC. The most popular method involves a wet chemical digestion of biogenic silica and while it yields important high resolution data, is a relatively expensive and time-consuming approach (Hurd, 1972; DeMaster, 1981, 1991; Eggiman et al., 1980; Mortlock and Froelich, 1989; Müller and Schneider, 1993; Landén et al., 1996). However, an alternative method to the wet chemistry method is the Fourier Transform Infrared Spectroscopy (FTIRS), which was first applied to lake core sediments by Vogel and Rosén (Vogel et al., 2008; Rosén et al., 2010; Rosén et al., 2011). FTIRS requires a small amount of sediment, produces fast sample analyses, and collects information on TOC and BSi (Rosén et al., 2011, Liu et al., 2013).

Why is this Relevant?

Paleoclimatologists run their lake core samples through the FTIRS to collect information on TOC and BSi. The offloaded OPUS Atmosphere Corrected data contains information on wavelength and absorbance. In this way, all of the current paleoclimatology literature refers to results in terms of absorbance, which is a relative value. This means that the absorbance spectra data is useful when comparing among the same sample but it can be difficult to generalize trends using absorbance spectra across different samples. Our goal with this project is to convert the absorbance values to percentages so as to provide paleoclimatologists with a more universal and distinct way of analyzing and understanding their data. While there has been great demand for this sort of calibration, some geoscientists lack the statistical background and expertise to create such a model.

Partial Least Squares Regression

Partial least squares regression is used to analyze the effects of one or more continuous explanatory variables on one or more than one continuous response variable. PLS is often used when working with and accounting for highly correlated explanatory variables. In our case, we are using absorbance and wavenumber values which are both highly correlated to one another, as we expect each absorbance value to relate to one another. We are using these multicollinear absorbance values to predict percentages of biogenic silica (BSi).

The general formula for a partial least squares regression model is:

Y = XB + E

Y = response matrix (n x m)

X = predictor matrix (n x p)

B = Coefficient matrix (p x m)

E = noise term (n x m)

Where n is number of observations, m is number of response variables and p is number of predictor variables.

In our model:

Y = percentages of BSi (39 x 1)

X = absorbance 39 samples (39 x 3698)

B = weights used to create various linear combinations (3698 x 1)

E = Error term (39 x 1)

Where n = 39, m = 1 and p = 3698

Essentially PLS determines the best way to combine X and Y, and then runs the regression to predict Y.

Why PLS and not some other regression model?

The model we are trying to fit has several explanatory variables that are highly correlated with one another and we are trying to run a regression model on a data set that has more covariates than observations. In this way, we need something more complex than a simple linear or multiple regression model. PLS separates the variability into two parts: (1) outer relations and (2) inner relations. The outer relations are the relationships between the predictor and response blocks individually, and can be shown by decomposing the predictor and response blocks. This is essentially what Partial Component Analysis does; it focuses on explaining the variance in X. However, we are interested in the inner relations that describe the relationships between the predictor and response blocks, basically how do the predictor and response variables relate? In order to determine this, the predictor variable is decomposed into T and the response variable is decomposed into U. Regression is then run on T and U. In this way, because we are interested in understanding how the predictor variable—the multiple absorbance values—relates to the predicted response variance—percentages of BSi—we need to perform a partial least squares regression.

How does it work? Linear Algebra!

The main objective of PLS is to decompose the predictor and response variables while taking into account how the predictor variable affects the response variables. This is performed by doing matrix algebra and using an algorithm called NIPALS or SIMPLS. This model uses Sijmen de Jong's SIMPLS algorithm. In the iterative SIMPLS algorithm, the predictor data (X) is projected onto a weight vector (R) and is used to calculate the scores (T). This relationship is depicted by the equation T = XR. We can derive R from the NIPALS algorithm using the formula $R = W(P'W)^{-}1$, where W is the weight vector and P is the loading. When we were creating loading plots, which illustrate which parts of the spectrum are being weighted more in our model, we had to decide whether to plot a set of weights R or loadings P. Because we are using PLS as a method for identifying a subspace within which to restrict and stabilize the regression vector, we need only to look at R. In this way, our PLS regression model is described as:

$\mathbf{Y = Xb + e}$, where b = $R(T'T)^{-}1T'y$

In this way, the regression vector is a linear combination of the weights R. B tells us which portions of X are most important for predicting Y, allowing the loading plots to illsutrate which portions of X are being weighted more in our model.

It has been interesting to develop and understand this PLS model as I simultaneously take Linear Algebra. I find myself drawing connections between this project and that course and it is rewarding to see linear algebra happen in "Real Life".

Some Focused Questions We Aim To Answer

- **Specific Spectrum vs. All Spectrum:** Is it beneficial to run the calibration model over a specific portion of the spectrum versus the entire spectrum? If so, which portion of the spectrum should we use?

As illustrated in the next section, we determined that it is more beneficial to run the model on a specific portion of the spectrum.

- **Recommended number of samples:** Can we pinpoint the recommended number of samples required to run the calibration model?
- **Universal Preprocessing:** Can all of the data be preprocessed in the same way, or do we need to provide our user with various options for preprocessing the data?

- **Transferable Results:** How transferable are these results from one lake site to another? Do we observe a difference between cold and warm climates? What about at different localities within a cold climate?
- **Marine cores:** How is this all transferable to the marine environment?

II. Data

The data set we are working with is from arctic lakes, where the relationship between diatoms and biogenic silica is more closely related. 12 "NAN-" lake core samples were collected from Nanerersarpik Lake in Southeast Greenland, 3 "FISK-" samples were collected from Fiskebol Lake in northern Norway, 8 "LSA-" samples were collected from the Lower Sermilik Lake in Southeast Greenland and two controls—clean beach sand and washed quartz—were included as well.

The NAN samples included several duplicates so we worked with a total of 39 OPUS "Atmosphere Corrected" Data point table files (dpt). These dpt files are the offloaded data from the FTIRS and include wavenumber and corresponding absorbance values. These 39 dpt files were the basis of the calibration model.

Data Validation

Cross-validation (CV) is a way of assessing how generalizable the results of a statistical analysis are to an independent data set. In our case, we used cross validation to assess how accurately our model would predict percentages of BSi. In a predictive model, two data sets are generally given: (1) data set of known data on which the training is run (training data set) and (2) data set of unknown data against which the model is tested (testing set). The objective of cross-validation is to test the model's ability to accurately predict new data that was not given and to highlight problems with the model such as overfitting or selection bias.

k-fold Cross-Validation

We based our cross-validation off of previous work by Rosén et al (2011). In their work, they used 10-fold cross validation, where 90% of the data was used for training and 10% of the data was used for testing. In our case, if we were to run a 10-fold, we would only have ~4 samples in each subset, which we felt was not sufficient for training and testing purposes. In this way, we chose a 5-fold cross-validation, where our original sample (n=39) was randomly partitioned into five subsamples. In this way, we had ~8 samples in each subset and reasoned this was an ideal number of samples for each subset. Four of the subsamples were used as the training set and one of the subsamples was used as the testing set. The cross-validation process was then repeated five times which each of the five subsamples used exactly once as the testing set. In this way, all of our observations were used for both training and testing, which is helpful as our sample size was relatively small.

Root Mean Square Error of Prediction (RMSEP)

The Root Mean Square Error of Prediction (RMSEP) is used to determine the optimal number of Partial Least Squares Regression components for prediction. RMSEP is essentially averaging the Root Mean Square Error for all of the testing folds. This gives us a sense of how accurate our model is in predicting BSi. A small RMSEP indicates a more accurate predictive power of the model. In our model, we created a cross-validated RMSEP curve to determine the number of components needed to predict BSi percentages. As observed in the plot below, three components yield the most accurate prediction in our model, while explaining the variance in X without overfitting the model.

When choosing the number of components, we are essentially defining the size of the subspace in which to run the regression. As observed in the table below, three components has the lowest cross-validated RMSEP,

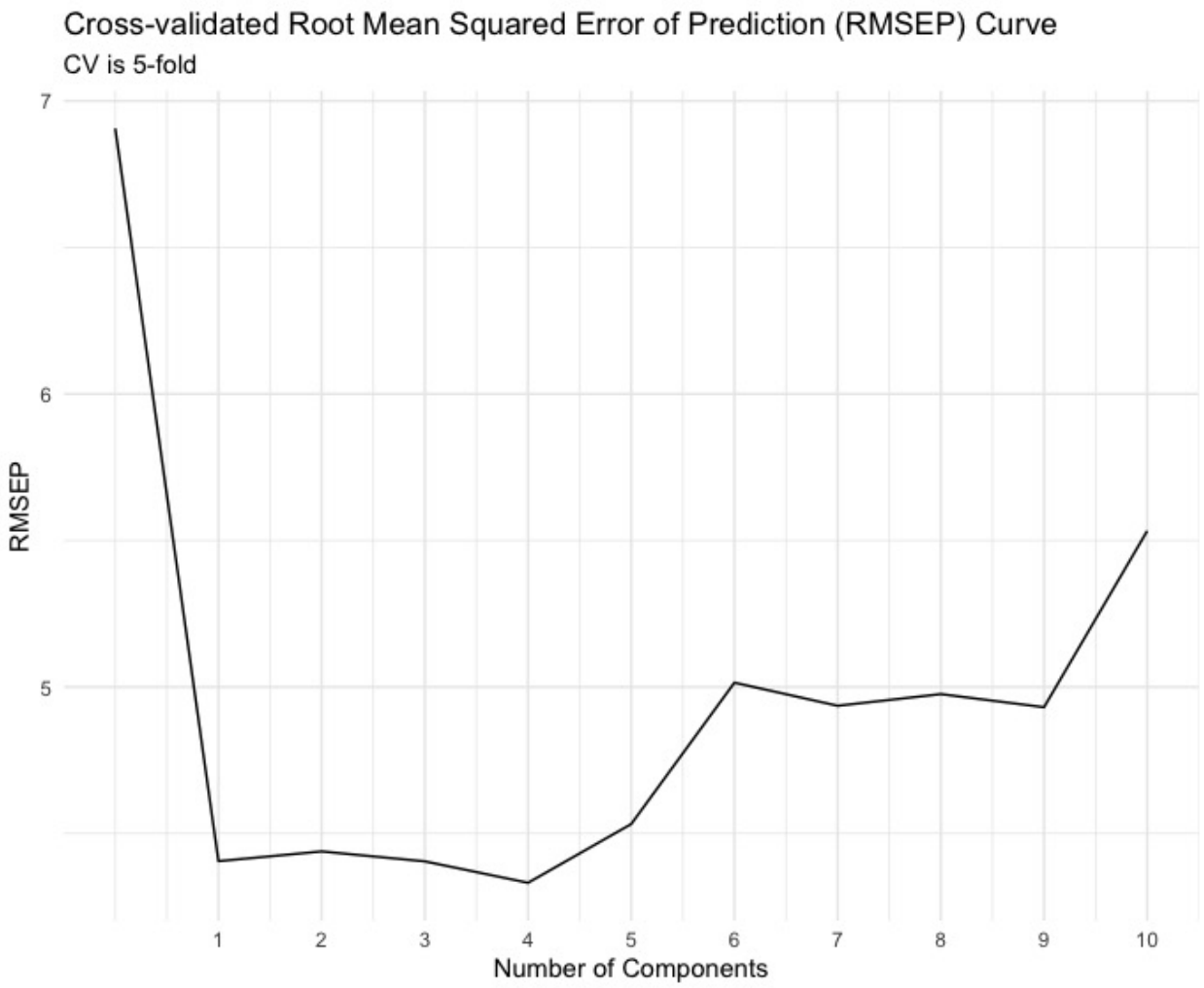


Figure 1: Cross-validated RMSEP Curve

which means that a model with three components will have the most accurate predictive power. We also observe that three components explain ~85% of the variance.

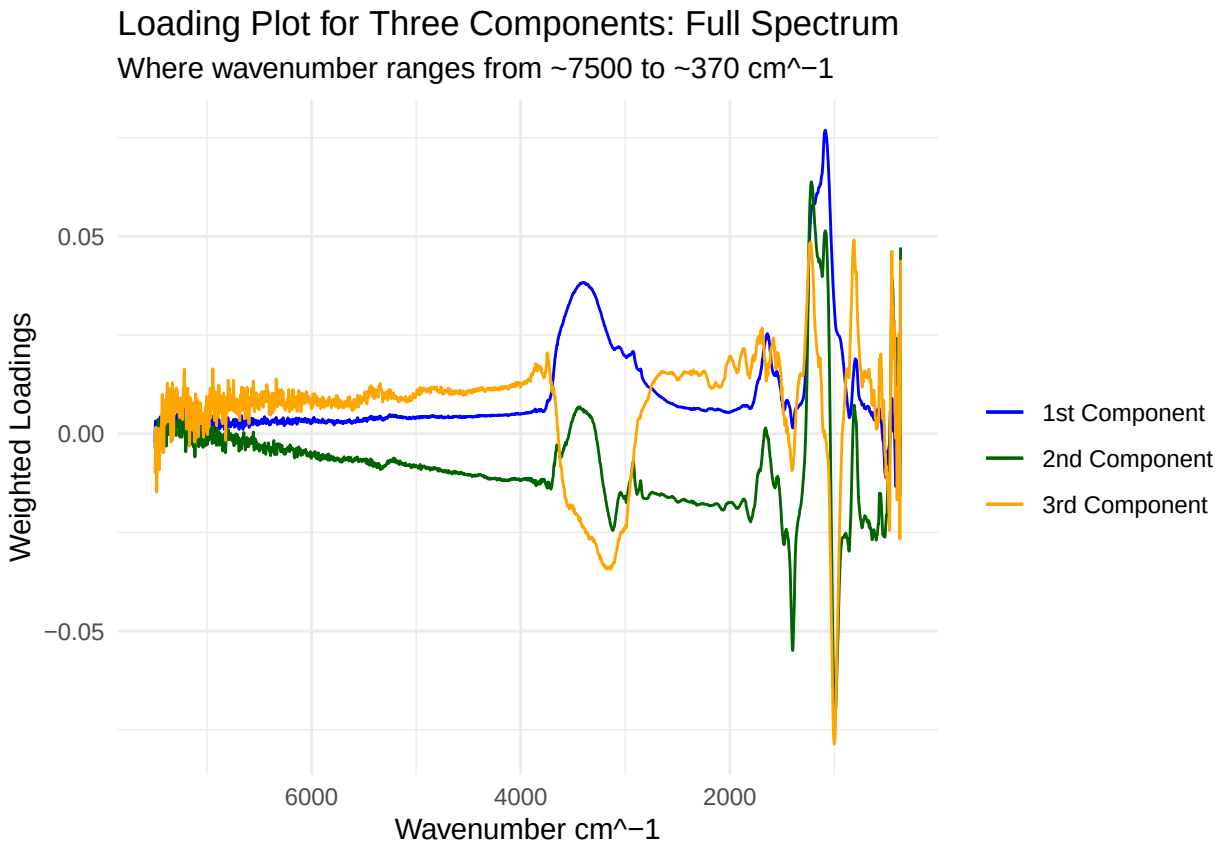
```
## Data:    X dimension: 39 3697
##  Y dimension: 39 1
## Fit method: kernelpls
## Number of components considered: 10
##
## VALIDATION: RMSEP
## Cross-validated using 5 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           6.906    4.550    4.697    4.540    4.669    4.832    5.015
## adjCV        6.906    4.532    4.565    4.398    4.530    4.659    4.719
##        7 comps  8 comps  9 comps  10 comps
## CV       5.108    5.503    5.805     5.815
## adjCV    4.780    5.058    5.303     5.278
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X             61.66    72.97    84.97    94.99    96.91    97.67    98.22
## BSiPercent    59.15    70.83    75.68    78.46    82.23    87.64    90.34
##             8 comps  9 comps  10 comps
## X             98.53       99     99.22
## BSiPercent    93.68       95     96.44
```

III. Loading Plots

Once we performed the appropriate cross-validation and selected the optimal number of components, we were able to create loading plots for each component. Loading plots are important as they highlight the parts of the spectrum weighted heavily in our model.

As illustrated in the plot below, we observe the loading plots to spike between 1820 and 1370cm^{-1}, 1300 and 850cm^{-1} and 850 and 700cm^{-1}. This is generally in line with what Rosen et al. (2011) report in their calibration model.

Future work involves rerunning the model on a subset of the spectrum. We aim to truncate any wavenumbers >3750cm^{-1} as that part of the spectrum was not weighted heavily. We would also like to focus on specific peaks, i.e. 1820 and 1370cm^{-1} or 1300 and 850cm^{-1}, to observe whether we detect any changes.



IV. Biogenic Silica Model Comparison

After running the calibration model, we compared the predicted BSi percentages against the actual wet chemical digestion percentages. The relationship is illustrated in the graph below. It is interesting to note that the model neither overfits nor underfits. Rather, we observe overfitting for some samples (i.e. LSA1-30A, LSA1-30B, NANB3A2-10.5 and NANDB-2) and underfitting for other samples (i.e. FSK-10, NANDB4). Most interesting, is how the model reacts to our two control samples SS and WQ. For both controls, the model predicts negative percentages of BSi. It is exciting to see that our model runs, but more work is needed to fine-tune the model.

Future Work

The Model

We would like to vary the preprocessing of the data using the OPUS software. The current samples were "Atmosphere Corrected" and we would like to experiment with other preprocessing settings to determine whether it would affect the model in any way. We would also like to run the model on subsets of the spectra. We would like to compare the model run on the entire spectrum with the model run on a specific portion of the spectrum to determine whether focusing on specific portions of the spectrum yields more accurate predictions.

We would also like to compare ranks of actual BSi percentages with the ranks of the predicted BSi to determine whether our model is accurately predicting the correct rank but simply getting the numbers wrong, or also inaccurately predicting the rank.

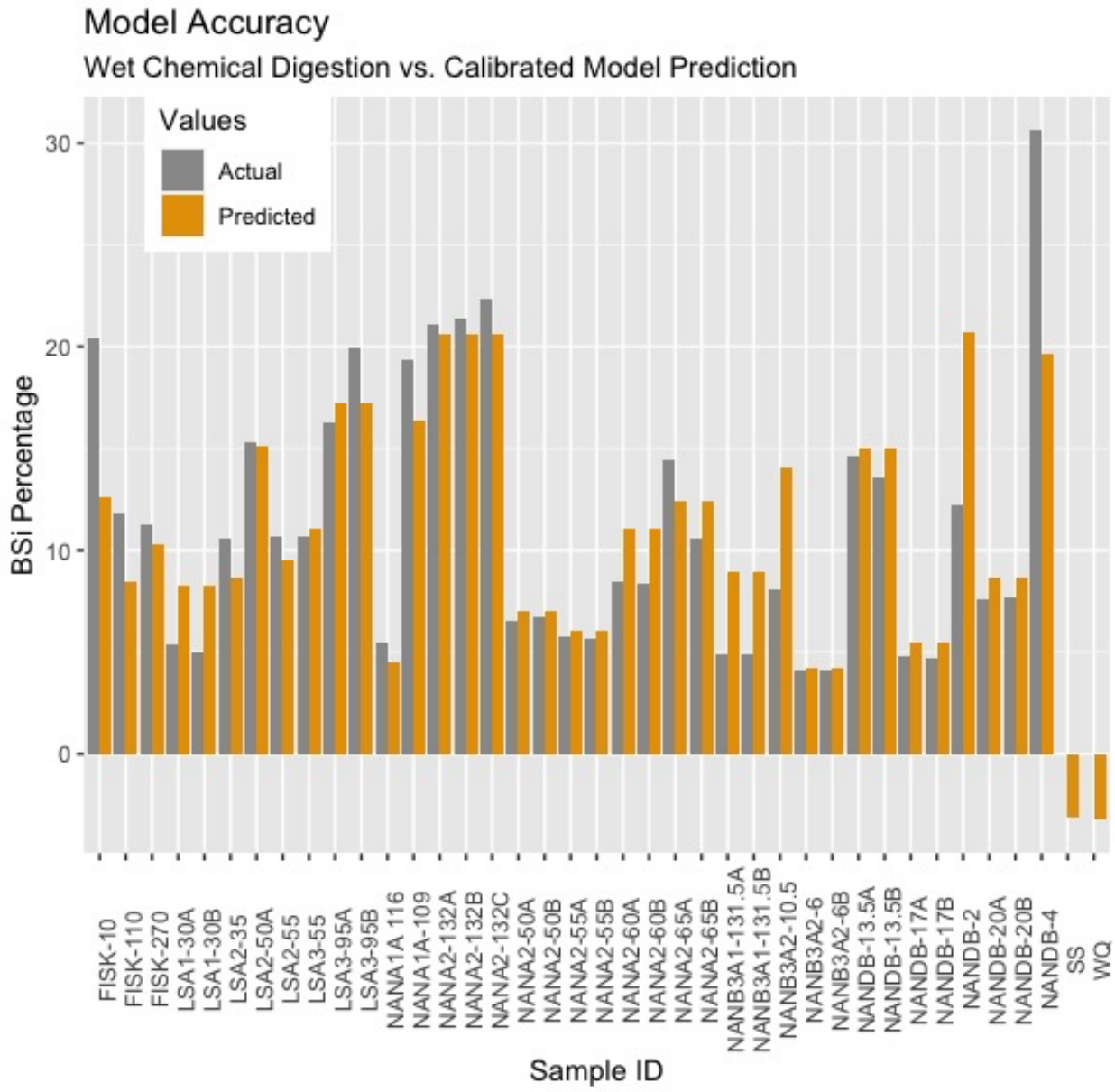


Figure 2: Actual vs Predicted BSi Percentage

Our samples are from various arctic settings in Greenland and Norway. We would like to map out the various samples and plot residual errors in order to better understand whether there are prediction trends related to locality.

Samples

We have more samples from Alaska, and would like to run the model on those samples to determine how the predictions between the current sample set and the new samples differ.

Prepare for Publication

As aforementioned, this type of predictive modeling is in demand and several paleoclimatologists would benefit from using this model. In this way, we will be developing a user-friendly PLS package in R specifically for paleoclimatologists, as well as an accompanying vignette to document our code. We will also create a Shiny App so as to increase accessibility. A manuscript will also be drafted in preparation for publication.

Acknowledgments

Thank you to Smith College and Posse for funding this research over the summer. Many thanks to Sara who provided helpful statistical and coding background and to Greg who provided the data and supervised this special studies.

References

"Cross-Validation (Statistics)." Wikipedia, Wikimedia Foundation, 4 May 2021, en.wikipedia.org/wiki/Cross-validation_%28statistics%29#Exhaustive_cross-validation.

DeMaster, D.J., 1981. The supply and accumulation of silica in the marine environment. Geochim. Cosmochim. Acta 45, 1715–1732.

DeMaster, D.J., 1991. Measuring biogenic silica in marine sediments and suspended matter. In: Hurd, D.C., Spenser, D.W. (Eds.), Marine Particles: Analysis and Characterization. American Geophysical Union, pp. 363–368.

"Different Kinds of PLS Weights, Loadings, and What to Look at?" Eigenvector, 12 Sept. 2009, eigenvector.com/different-kinds-of-pls-weights-loadings-and-what-to-look-at/.

Eggiman, D.W., Manheim, F.T., Betzer, P.R., 1980. Dissolution and analysis of amorphous silica in marine sediments. J. Sed. Petrol. 50, 215–225.

Hurd, D.C., 1972. Factors affecting solution rate of biogenic opal in seawater. Earth Planet. Sci. Lett. 15, 411–417.

Landen, A., Holby, O., Hall, P.J., 1996. Determination of biogenic silica in marine sediments—selection of pretreatment method and sample size. Vatten 52, 85–92.

Mevik, Bjørn-Helge, and Ron Wehrens. "Introduction to the PLS Package." University Center for Information Technology, University of Oslo and Biometris, Wageningen University & Research, 4 Aug. 2020.

Mortlock, R.A., Froelich, P.N., 1989. A simple method for the rapid determination of biogenic opal in pelagic marine sediments. Deep-Sea Res. 36, 1415–1426.

Müller, P.J., Schneider, R., 1993. An automated leaching method for the determination of opal in sediments and particulate matter. Deep-Sea Res. 40, 425–444.

Rosen, P., Vogel, H., Cunningham, L., Reuss, N., Conley, D., Persson, P., 2010. Fourier transform infrared spectroscopy, a new method for rapid determination of total organic and inorganic carbon and opal concentration in lake sediments. J. Paleolimnol. 43, 247–259.

Rosen, P., Vogel, H., Cunningham, L., Hahn, A., Hausmann, S., Pienitz, R., Zolitschka, B., Wagner, B., Perssos, P., 2011. Universally Applicable Model for the Quantitative Determination of Lake Sediment Composition Using Fourier Transform Infrared Spectroscopy. Environ. Sci. Technol. 45, 8858–8865.

"Partial Least Squares Regression Analysis." YouTube, YouTube, 3 June 2019, www.youtube.com/watch?v=2GoyL_SsrBk.

Vogel, H., Rosen, P., Wagner, B., Melles, M., Persson, P., 2008. Fourier transform infrared spectroscopy, a new cost-effective tool for quantitative analysis of biogeochemical properties in long sediment records. J. Paleolimnol. 40, 689–702.

U.S. National Library of Medicine, National Institutes of Health, openi.nlm.nih.gov/detailedresult?img=PMC4444114_pone.0