Geostatistics

Creating a Partial Least Squares Calibration Model for Paleoclimatologists

Vivienne Maxwell

5/12/2021

Contents

I. Purpose

The aim of this project is to create a 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 the 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 associated 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 wet chemistry and 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?

Once paleoclimatologists input their lake core samples into the FTIRS and offload the OPUS data files, the data contains information on wavelength and absorbance. In this way, all of the current paleoclimatology literature refers to results in terms of absorbance, which a hard value to understand. Our goal with this project is to convert those absorbance values to percentages so as to provide paleoclimatologists with a more meaningful way of analyzing and understanding their data.

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?
- **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

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 28 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 28 dpt files were the basis of the calibration model.

Preprocessing

We decided on a five-fold cross-validation as our sample size (n=28) was relatively small. In order to determine the number of components needed, we created a Cross-validated Root Mean Squared Error of Prediction (RMSEP) curve. As observed in the plot below, three components was enough.

III. Loading Plots

Once we determined the appropriate cross-validation and number of components, we were able to create loading plots for each component. Loading plots are important as they [. . .]

As illustrated in the plot below, [. . . .]

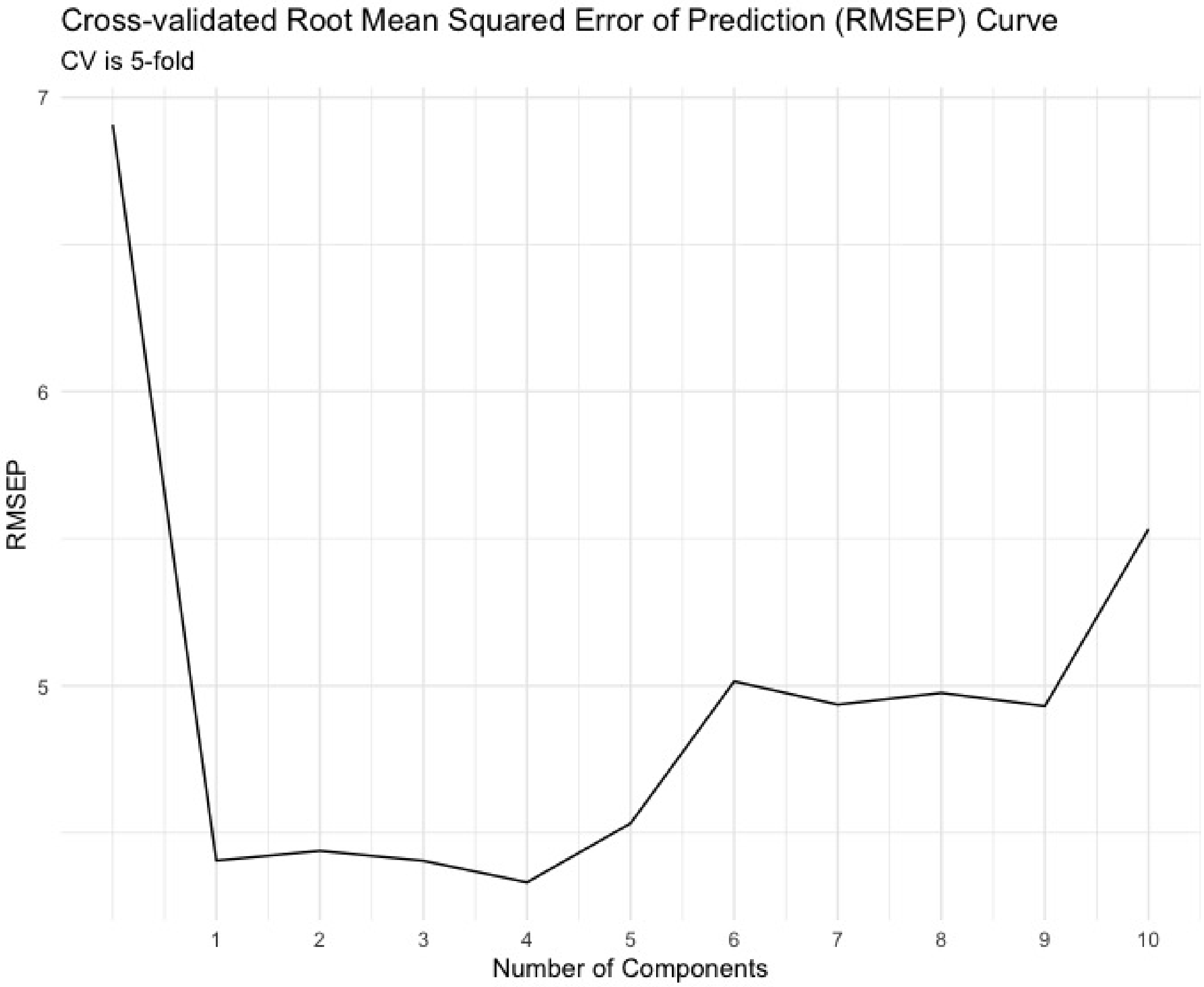

Figure 1: Cross-validated RMSEP Curve

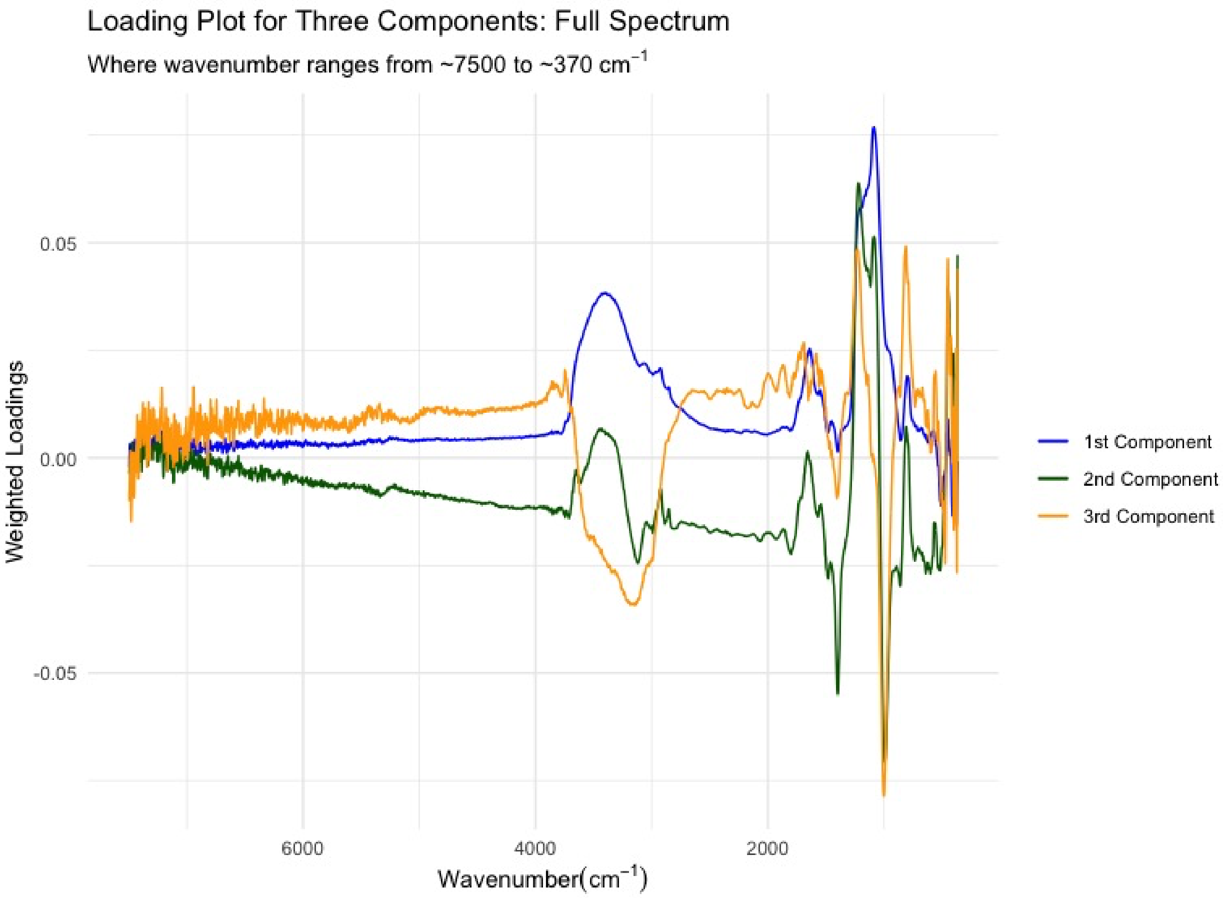

Figure 2: Full Spectrum Loading Plot for first three components