Pyleoclim: A Python Package for the Analysis of Paleoclimate Data

Introduction

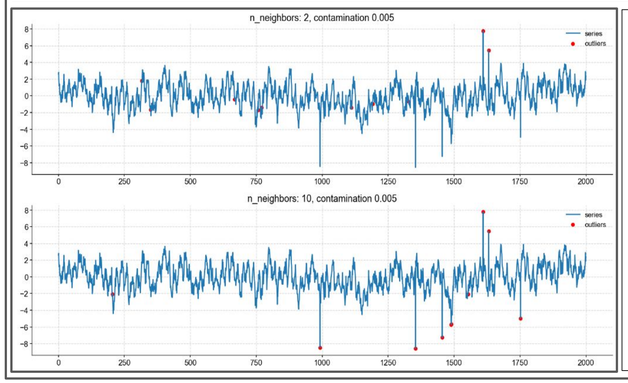
Pyleoclim is an object-oriented Python package for analyzing and visualizing time-series paleoclimate data, which offer unique challenges to the analyst, as they usually come in the form of timeseries with missing values and age uncertainties.
Our goal for this semester is to increase its functionalities such as anomaly detection, optimizing time-series analysis, and visualization styles.

Methods

- Understand the functionalities on generated data/examples from blog post
- Write functions in Pyleoclim that apply methods (ML for anomaly detection/visualization) to paleoclimate data
- Write CI tests
- Write Documentations
- Write a tutorial notebooks on how to use the method/functionality

Issue 1: KNN on anomaly detection (Sunny Lee)

This issue aims to explore how KNN works for timeseries data especially on paleoclimate data. Moreover, we also implement automate auto-tuning of the parameters using the silhouette method to heuristically determine all parameters

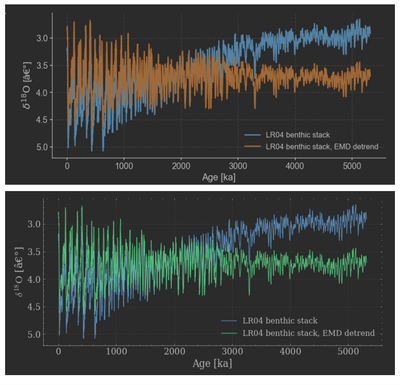


Approach

- For KNN on anomaly detection, it involves identifying data points that deviate significantly from the majority of the data
- KNN: Leveraging sklearn LocalOutlierFactor() function
- Silhouette method: A metric used to assess the quality of clustering, we're using it to evaluate the performance of models

Issue 2: SciencePlots Functionality (Ginny Barnes)

SciencePlots - API for developing standardized plot styles using MatPlotLib (plots for IEEE, science, nature articles)

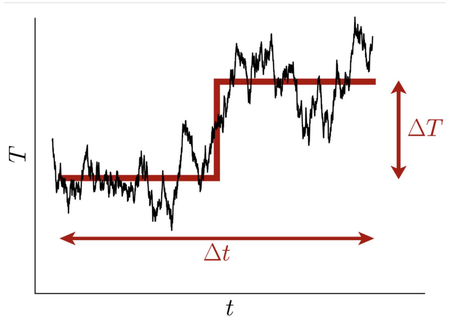


Approach

- Generate notebook with Pyleoclim/SciencePlot plot comparisons (IP)
- Add SciencePlots into Pyelo dependencies and environment (IP)
- Append new SciencePlot styles to Pyleo Plot.py file
- Create Unit Test to ensure smooth integration of Scienceplots
- Develop Scienceplots PyleoTutorial

Issue 3: Haar Fluctuation Analysis (Harman Pelia)

Current Pyleoclim Time Series analysis are effective yet costly due to our data's characteristics. Haar Analysis improves the efficiency, forecasting quality, and particularly useful for data compression and denoising applications. Implement a working Haar Fluctuation Analysis Method for our data based on previous research algorithms created for similar data.

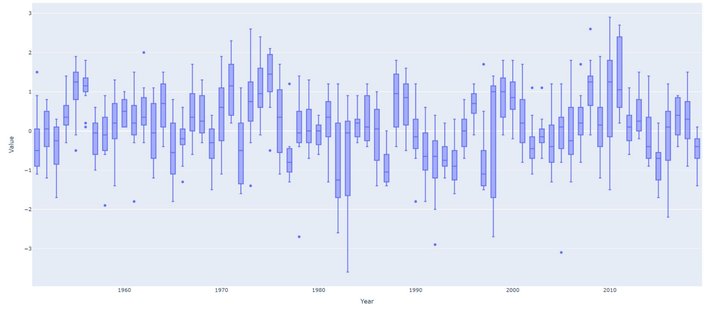


Approach

- Determine notation and good defaults values for our method.
- incorporate haar_v2 in Pyleoclim as a method of the Series class:
- Use this method to reproduce a findings similar to past research on some of Pyleoclim's pre-defined timeseries
- figure out a good unit test for this method
- implement it

Issue 4: Box plots for Time Series Data (Andrew Bae)

This issue contends with the use pandas and Plotly to represent change in paleoclimatic data over the course of decades through the use of box plots to measure change in Southern Oscillation Index value



Approach

- Use .astype function of pandas to group points from the same year into one dataset
- Generate boxplot using Plotly to juxtapose each year's data and measure annual changes in the Southern Oscillation Index value