Analysis of Speleothem Data Using SISALv3

Introduction

Speleothems, such as stalagmites and stalactites, are valuable archives of past climate conditions. They record various isotopic compositions that can be analyzed to infer historical climate data. This report details the methodology and findings of our analysis using the SISALv3 database, focusing on stable isotope compositions (δ18O and δ13C) in speleothem samples.

Methodology

Data Extraction

1. **Database Connection:** We connected to the SISALv3 MySQL database using SQLAlchemy. The relevant tables included `sample`, `d18O`, and `d13C`.
2. **SQL Query:** We constructed a query to extract `sample_id`, `depth_sample`, `d18O_measurement` (renamed as `d18O`), and `d13C_measurement` (renamed as `d13C`) from the respective tables, ensuring no NULL values were included.
3. **Data Loading:** The queried data was loaded into a Pandas DataFrame for analysis.

Data Processing

1. **Sorting and Splitting:** The data was sorted by `depth_sample` to maintain chronological order. The dataset was then split into training (80%) and testing (20%) sets.
2. **Time Series Analysis:**

- **ARIMA Model:** An ARIMA model (order 5,1,0) was fitted to the `d18O` values in the training set. The model was used to forecast `d18O` values for the testing set, and predictions were added to the DataFrame.
- **Performance Metrics:** Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated to evaluate the model's performance.

3. **PCA Analysis:**

- The `depth_sample`, `d18O`, and `d13C` columns were standardized.
- Principal Component Analysis (PCA) was performed to reduce dimensionality and identify principal components.

Visualization

1. **Time Series Plot:**

- Plotted the `d18O` values from the training, testing sets, and the model's predictions against `depth_sample`.
- Included a legend to distinguish between the training data, testing data, and predictions.

2. **PCA Plot:**

- Created a scatter plot of the first two principal components, color-coded by `depth_sample`.

Findings

Descriptive Statistics

- The dataset included 266,178 samples.
- Key statistics for `depth_sample`, `d18O`, and `d13C` were calculated and displayed.

ARIMA Model

- The model captured the overall trend in the training data but had limitations in accurately predicting the variability in the testing set.
- Performance metrics:
    - MAE: 2.28
    - MSE: 8.68
    - RMSE: 2.95

PCA Analysis

- The first two principal components explained a significant portion of the variance in the data.
- The scatter plot revealed patterns indicating changes in isotope composition with depth.

Conclusion

Our analysis provided insights into the isotopic composition of speleothems and their relationship with depth. The ARIMA model highlighted the challenges in time series prediction for speleothem data, while the PCA revealed underlying patterns. These findings contribute to our understanding of past climate conditions as recorded in speleothems.