MOSAIC: Metadata Optimization and Statistical Anomaly Detection using Unsupervised Clustering for Geospatial Temperature Data

Interns: Neel Kansara, Darryl Tang, Jordan Tran

Mentors: Andrew Clark, Dr. Kevin Czajkowski, Brianna Lind, Dr. Rusanne Low, Ian Maywar, Sara Mierzwiak, Pramila Paudyal

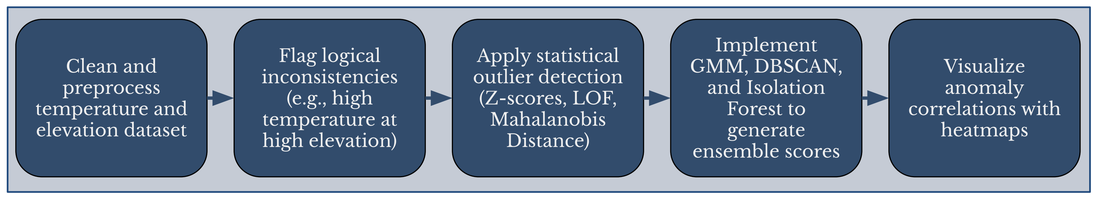
Introduction

The GLOBE dataset is a collection of hundreds of thousands of environmental data points collected by citizens around the world. The data spans 30 years and measures a wide range of different metrics, such as air temperature, soil moisture, and salinity. With the immense amount of data stored by the program, many students and researchers rely on the dataset for their own research studies. As this dataset relies on citizen scientists to collect and input information, it is imperative that this data is not only reliably accurate, but also easily accessible for future researchers to study global trends. However, as the current dataset consists of raw information recorded over three decades by a multitude of researchers, it consists of many data points with inconsistent information. Due to a multitude of reasons, including but not limited to human error, device error, and improper data collection, many points do not show accurate data. Including these data points in any research analysis could lead to inaccurate conclusions, which can be harmful for many reasons. Using various models and statistical tests, our research aims to provide researchers with accurate, interpretable flags and additional metadata to aid researchers in effectively using the GLOBE dataset.

Methodology

While the surface temperature dataset contains 67 columns of information, further columns are added to improve potential models, as well as providing information that wasn't previously given. Information was found through reverse geocoding inputted latitude/longitude, which provided country, continent, and biome data. Time stamped data was extracted to provide the month and year columns. Metadata added: Countries, Country Code, Continent, Year, Month, Biome, Season.

After the addition of metadata, a step-by-step process was used to traverse and flag the dataset through clustering and statistical tests. A system was developed outlined by this workflow:



Results

Figure 1: Geographical Mapping of Accuracy Scores Given by Clustering Models

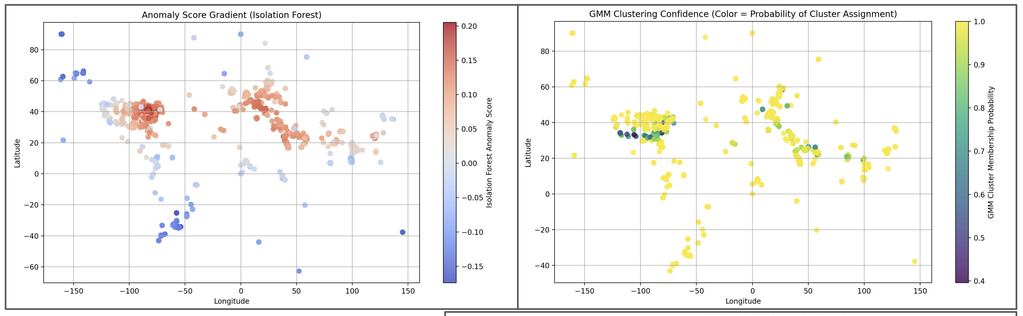


Figure 1A shows the anomaly score based on Isolation forest. Geographically distinct points are most likely to be flagged as anomalies by this model. Figure 1B shows Gaussian Mixture Modeling. Points with unusually high inverse correlation between temperature, elevation, and location are flagged by this model. Figure 1C shows scoring based on DBSCAN, GMM, and Isolation Forest. Using this ensemble scoring system solution between multiple clustering methods provides higher accuracy flagging than individual models.

Figure 2: GMM Clustering in PCA Feature Space with Low-Confidence Highlighted

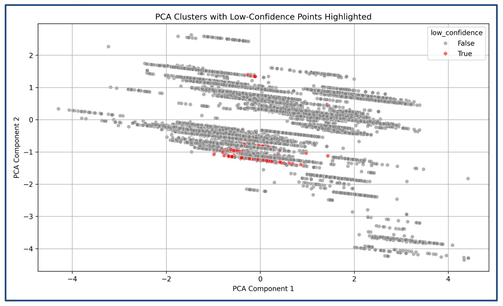


Figure 2 shows Gaussian Mixture Model (GMM) clustering results on the temperature data projected into two principal components. Each point represents a surface observation, color-coded by cluster confidence. Points in red indicate low-confidence assignments (probability < 0.6), suggesting ambiguous or anomalous data. These points may correspond to extreme environmental conditions, metadata errors, or sensor inconsistencies. The separation of clusters and clear low-confidence zones support the use of GMM for uncertainty-aware anomaly detection.

Figure 3: Correlation Heatmap Between Statistical and Logical Flags

Figure 3 shows correlations between the logical and statistical flags. While high elevation and hot temperature flags didn't show significant correlations with statistical models, they showed a high positive correlation (0.86) with flag_sum, the total flag count per data entry. This may suggest elevation and temperature outliers trigger other logical flags. The duplicated coord flag had a high inverse correlation (-0.72) with the local outlier factor flag. This might suggest a high duplication of points negatively affects statistical outlier testing.

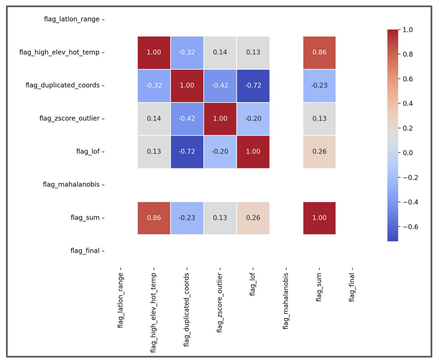


Figure 4: Metadata used to represent the difference of monthly surface temperatures in the US vs India

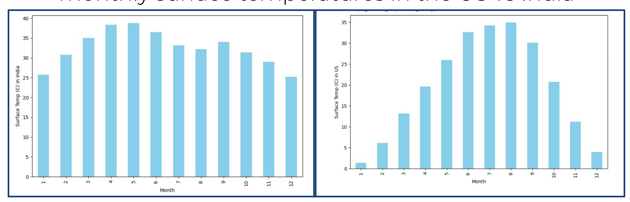


Figure 4 shows an example of temperature analysis on country and month, using the points flagged as accurate by our model. Month and country are two examples of the added metadata columns to the dataset, so these graphs are not able to be generated using the raw dataset.

Conclusion

This analysis produces a robust, interpretable flagging system for identifying anomalies in surface temperature data collected as a part of the GLOBE mission, through metadata augmentation and the use of unsupervised clustering algorithms and statistical tests. Figure 1 highlights the varied results of individual clustering models, justifying an ensemble scoring approach between multiple unsupervised clustering models. Figure 3 demonstrates the partial correlations among various anomaly flags, showing that the flagging system captures multiple dimensions of abnormality. The metadata augmentation and flagging approach added these columns to the GLOBE dataset:

- Metadata: Country, Country Code, Continent, Year, Month, Biome, Season
- Flags: latlon_range, high_elev_hot_temp, duplicated_coords, zscore_outlier, lof, mahalanobis, ensemble_score

This flagging system enhances the trustworthiness and usability of the data, providing future researchers with a scalable strategy to detect anomalies in Earth science datasets.

Acknowledgements

Our team would like to thank NASA, CSR, and the SEES program for enabling us to conduct research with industry professionals and experts. We would like to thank our project mentors for providing amazing guidance, as our research would not have been possible without them. Our team would also like to thank all the researchers who have contributed to the GLOBE dataset over the past 30 years, providing crucial information for earth science studies which can have large implications.