UNDERSTANDING THE IN-SITU SOLAR WIND PROPERTIES WITH MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Arturo Avila-Lares, Brandyn Mahan, Andrea Newlands

Introduction

A plane at the equator of the sun, called the heliospheric current sheet (HCS), serves as the boundary between the magnetic dipoles.[1] The sun's magnetic field flips its pole approximately every 11 years[2] which, when combined with solar wind, causes the HCS to become tilted and wavy.[1] The standard deviation of the HCS's latitude (SD) and integrated slope (SL), called the HCS indexes, are two novel parameters that may aid in gauging the HCS's complexity and tracking the solar cycle.[3] A highly complex HCS indicates the Sun is near the solar maximum and that its poles will soon flip. During this time, solar winds are intensified.

Solar wind is a flow of particles made of mostly protons and electrons that are thrown from the sun's atmosphere, which is too hot for the sun's gravity to contain. It then travels along the sun's magnetic field lines before being ejected away from the Sun.[4] Solar wind is traveling at about one million miles per hour when it reaches Earth and can have immense impacts on different parts of life on Earth; affecting things like satellite operations, power grids, and telecommunication systems. Solar wind can also pose a threat to astronauts traveling in space.[5] It is thus important to understand both solar wind and the Sun's magnetic field in order to predict times of high solar wind activity as well as mitigate its effects.

We aim to investigate solar winds and HCS indexes in order to enhance our understanding of different categories of solar winds as well as forecast future HCS indexes. We first used hourly solar wind data (i.e. proton density, temperature, and speed) obtained from the Advanced Composition Explorer (ACE) Mission which launched in 1997 to create a base K-Means model with 4 clusters. We then enhanced this model by using a PCA to produce two principal components. While these clusters are close together, they are visually distinct. DBSCAN was unable to identify significant clusters within the solar wind data.

We had hoped to use the principal components derived from the ACE data to aid in the forecast of the HCS indexes, however the drastic reduction in the amount of HCS index data that could be used for forecasting caused by the ACE mission's start in 1997 as well as the removal of nuance from the ACE data caused by the aggregation of hourly into monthly averages created poor-performing models. For this reason, monthly sunspot numbers and lagged columns were used as the predictors for HCS indexes. After cross-validation across various models, Ridge Regression emerged as the best-performing model for both selected indices, SD_70 and SL_70_log10. Subsequently, we fine-tuned the hyperparameters to optimize performance, achieving a mean cross-validated R^2 values of 0.78 for SD_70 and 0.59 for SL_70_log10 on the test set.

Datasets

Advanced Composition Explorer Mission

The Advanced Composition Explorer (ACE) Mission is a satellite deployed by NASA to collect and analyze particles from solar, interplanetary, and interstellar origins[6]. While the ACE Mission collects various types of particle information, our project focus was on Solar Wind Electron, Proton, and Alpha Monitor (SWEPAM) measurements provided by ACE. The level 2 dataset from the SWEPAM is organized into hourly averages of solar wind parameters[7]. Data features for ion measurements include density, velocity coordinates, temperature, and speed. Missing data for a given time period is logged as -9999.9. Overall this dataset amounts to 204,768 rows. Data was downloaded in space-delimited file format.

Table 1: SWEPAM Level 2 Dataset	
Feature	**Description**
fp_year	fractional year.
fp_doy	fractional day-of-year.
ACEepoch	seconds since Jan 1 00
proton_density	Proton Density (cm^3).
proton_temp	Radial Component of proton temperature (deg. Kelvin).
He4toprotons	Ratio of alphas/protons
proton_speed	Proton Speed (km/s)
x,y,z_dot_GSE	X,Y,Z component of proton velocity in GSE coordinates (km/s)
r,t,n_dot_RTN	R,T,N component of proton velocity in RTN coordinates (km/s)
x,y,z_dot_GSM	X,Y,Z component of proton velocity in GSM coordinates (km/s)
Electron_temp	Electron Temperature (deg. Kelvin).
pos_gse_x,y,z	Components of spacecraft position in GSE (km).
pos_gsm_x,y,z	Components of spacecraft position in GSM (km).
A value of -9999.9 indicates bad or missing data.	

Sunspot Number

The sunspot number dataset is a collection of sunspot counts dating back to 1749. The US National Oceanic and Atmospheric Administration (NOAA) and the Solar Influences Data Analysis Center in Belgium maintain the official sunspot number dataset[8]. Sunspot numbers are calculated based on the initial count of sunspot groups along with the number of individual sunspots that can be observed. The sunspot number is the sum of individual sunspots plus the number of sunspot groups multiplied by ten. Sunspot groups are multiplied by ten due to the groups on average having ten individual spots within[9]. Our project used the monthly mean sunspot number dataset. Data was downloaded in TXT file format.

Relative Sunspot Number Calculation

$$R = K(10g + s)$$

Table 2: Sunspot Number	
Feature	Description
Year	Year of sunspot observation
Month	Month of year in integer value ranging 01-12
Median fractional date	Date in fraction of year for the middle of the corresponding month.
Mean Sunspot Number	Monthly mean total sunspot number
STD	Standard deviation of the sunspot numbers
Observations	Number of observations used to compute the monthly mean sunspot number

HCS Indexes

The HCS indexes dataset was developed by Dr. Zhao at the University of Michigan - Ann Arbor Department of Climate and Space Sciences and Engineering. The dataset focuses on two parameters, SD and SL, which are used to track the HCS structure over time. SD is the latitudinal deviation of the HCS from the Sun's equator. SL is the waviness of the HCS in reference to the Sun's equator. The dataset contains features for fractional year, the two indexes SD_70 and SL 70, and ten other features derived from SD_70 and SL_70 which were not of interest for this study. The dataset consists of 596 entries recorded from 1976 to 2022. Each observation corresponds to a single rotation of the sun which takes 27.2753 days, called a Carrington rotation.

Table 3: HCS Index

Feature	Description
fyear_CS	Fractional year value for data observation
SD_70	Integral of the wave's slope at a 70-degree latitude, quantifying the deviation of the HCS from the equator
SL_70	Measurement of the current plane's tilt at 70 degrees latitude, indicating the HCS's waviness.

Feature Engineering

A principal component analysis was performed on the ACE Mission data in order to reduce the dimensions in a way that would be easier to visualize in a 2D format as well as aid the clustering models. Two principal components were produced.

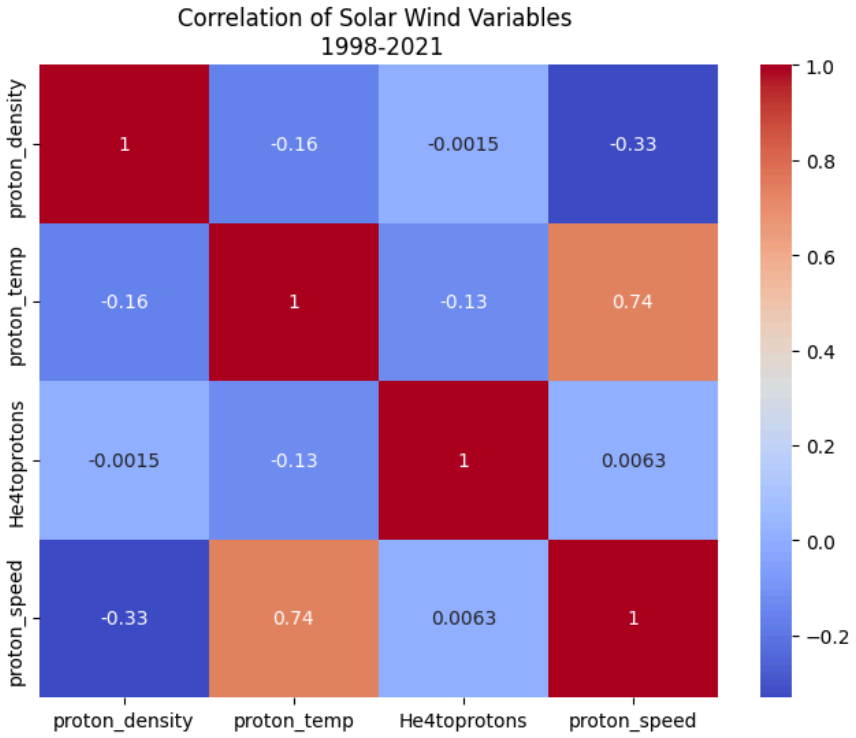
To ensure that SL_70 and SD_70 of the HCS indexes were on a similar scale, we calculated the logarithm base 10 of SL was calculated, which is referred to as "SL_70_log10" in the notebooks. Lag columns of the outcome variables were also calculated as potential features. The date field was used as an index to ensure that the lag was calculated based on the date and not on the position. Based on feature importances and evaluation metrics, it was determined that the lag features for one month before added the greatest value for the forecast model.

EDA

ACE Mission

The original dataset contains 204,768 between 1998 and 2021, however there are 121,969 after dropping rows with bad or missing data. The figure below shows a correlation matrix of the solar wind features.

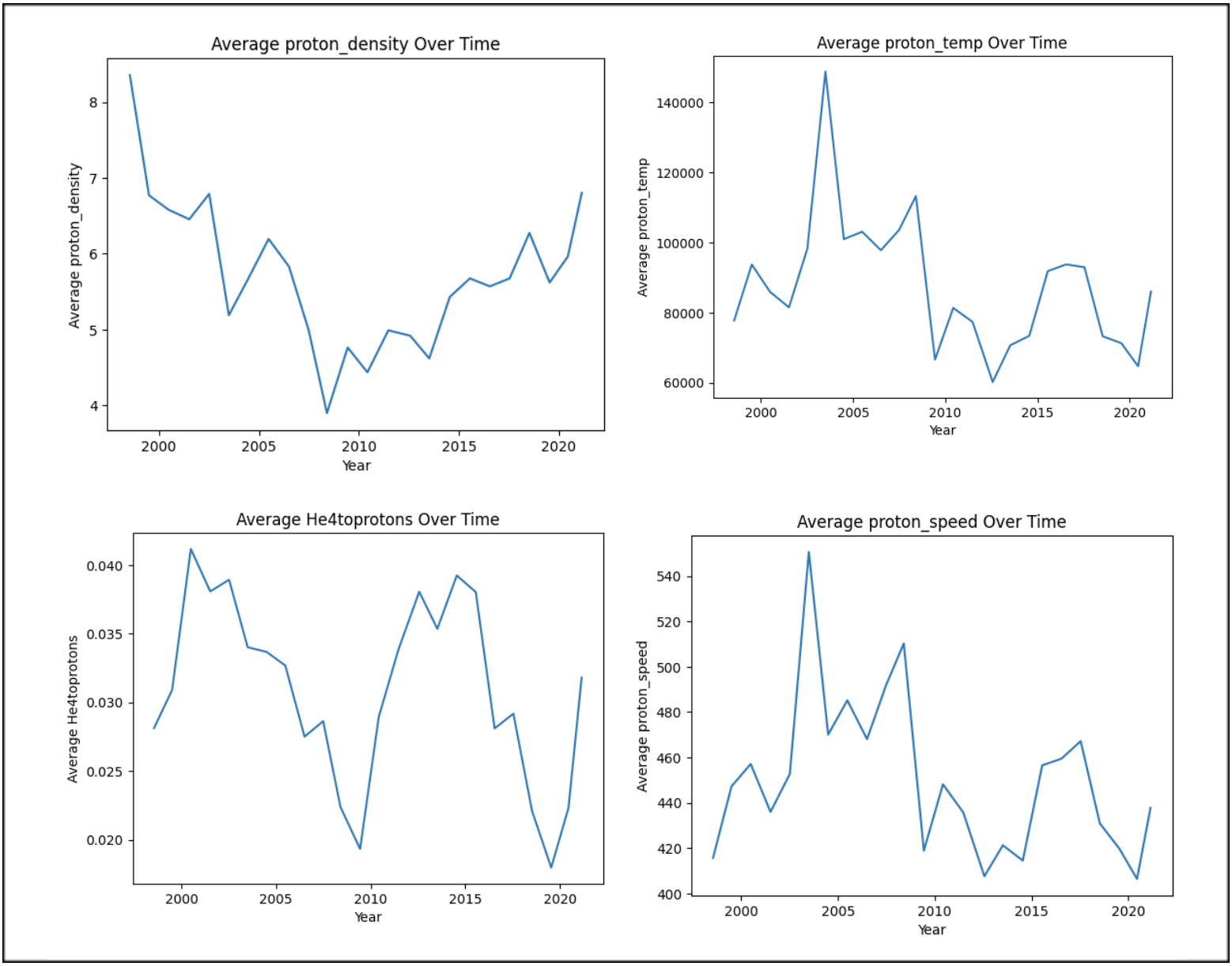
Figure 1



Correlation matrix of solar wind metrics

In addition, looking at the values over time show a slight resemblance to the Sun's solar cycle as the Sun's magnetic field influences the intensity of the solar winds.

Figure 2

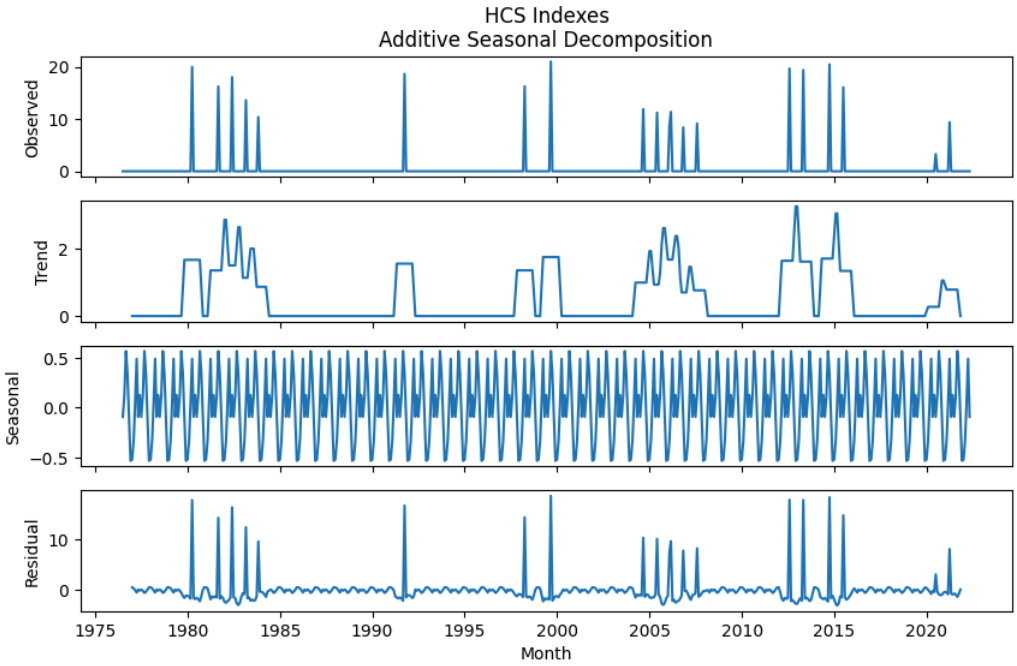


Solar wind metrics over time

HCS Indexes

Between 1976 and 2022, there is data missing for seventeen Carrington rotations, which we imputed with a zero to perform a seasonal decomposition analysis using an additive model. The graph below, which was nearly identical for both outcome variables, illustrates the pronounced seasonal pattern, with a trend component that shows occasional significant deviations from the norm. The residual component mostly lies around the zero line with some sporadic spikes, which might represent random or irregular effects not explained by the seasonality or trend.

Figure 3



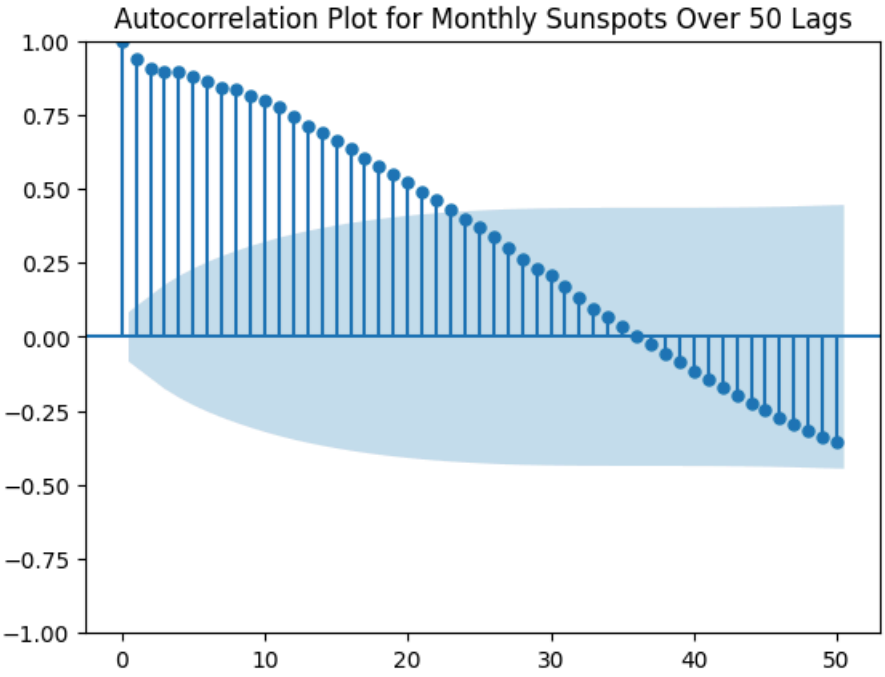
HCS Index SD_70 Additive Seasonal Decomposition

Sunspots

The dataset consists of 3,292 monthly observations raging from 1749 to 2023. By smoothing out the observations using a moving average, we were able to see more defined solar cycles (see figure in the appendix). Using the smoothed out data, we attempted to identify peaks in the data but weren't fully successful (see figure in the appendix). Thus, we opted for using an average of 11 years per cycle, which worked better yet the cycles didn't align exactly (see figure in appendix).

To further explore the cycles, an autocorrelation plot was created. The plot below indicates a strong, cyclical pattern in the monthly sunspot data. It also illustrates that the data is positively correlated at shorter lags, with the correlation diminishing and sometimes becoming negative as the lags increase, consistent with solar cycle behavior.

Figure 4



Autocorrelation plot of monthly sunspot numbers recorded from January 1749 to December 2023, revealing strong cyclical patterns.

To investigate the relationship between the HCS indexes and sunspot activity, the outcome variables and the monthly sunspot numbers were plotted against each other.

Figure 5

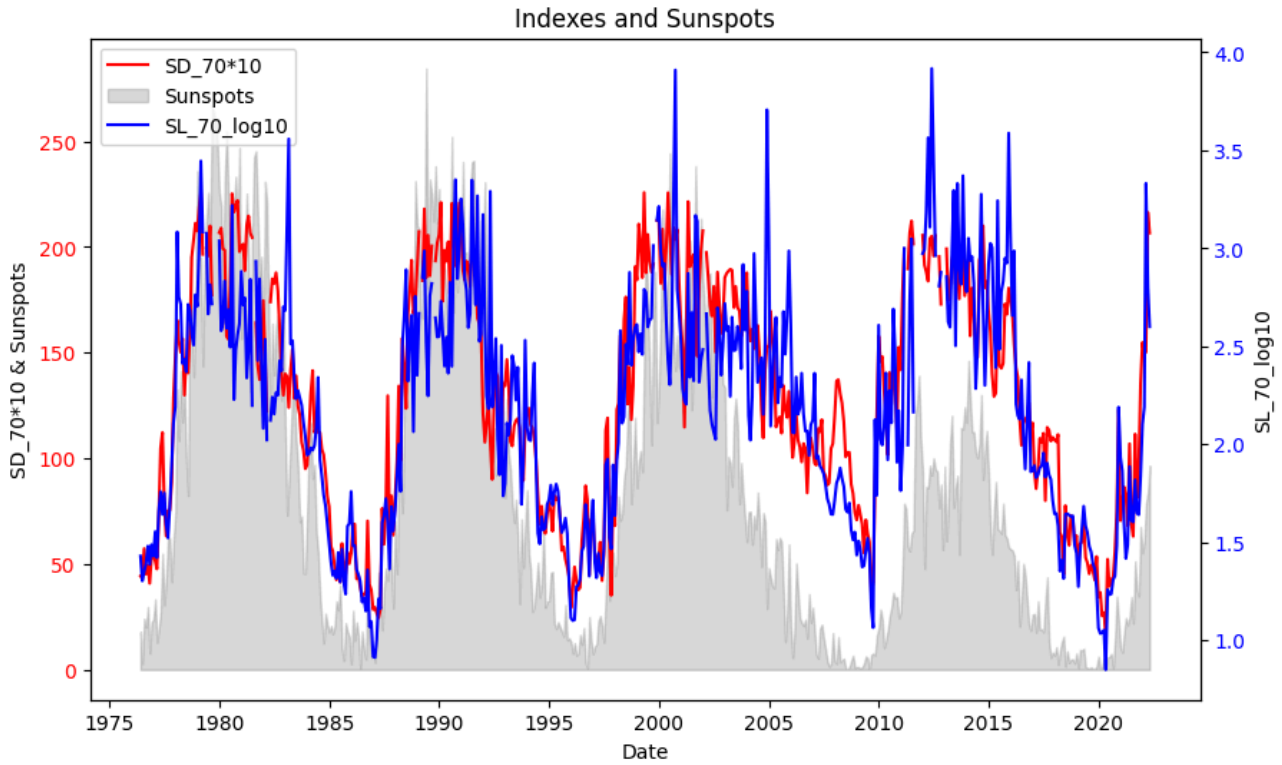


Chart depicting how the HCS indexes reflect patterns in monthly sunspot numbers, suggesting their potential use as effective predictors.

Clustering

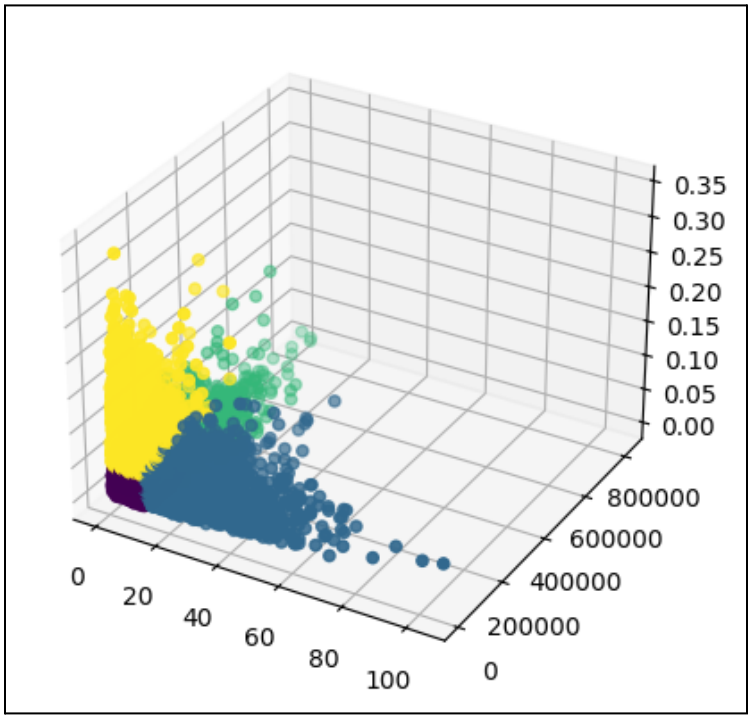
K-Means

ACE Mission data was imported with a focus on four features: proton_density, proton_temp, He4toprotons, and proton_speed. The data was scaled using Standard Scaler. Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI) were calculated for K-means models with cluster numbers ranging from 2 to 10. CHI and DBI were plotted at each cluster number to find the peak CHI and minimum DBI (see figure in the appendix). CHI was highest at 2 clusters while DBI was minimum at 4.

To resolve this difference, the Elbow Method was employed. By plotting the within-cluster sum of squares (WCSS) against k number of clusters, we were able to find the significant change in slope across k (the elbow) at 4 clusters[10] (see figure in the appendix). This determined that our K-means model should have 4 clusters.

The K-means model was set to 4 clusters and the resulting cluster labels were joined to the ACE dataset. Pair combinations of the 4 ACE mission features were iteratively plotted to see if any two variables showed the most distinct clusters (see figure in the appendix). Clusters were very close to each other with no clear decision boundary between the 4 clusters. Figure 4 below shows the clusters plotted in a 3-dimensional plane. The 3-dimensional plot shows how clusters are still close together without a clear decision boundary.

Figure 6

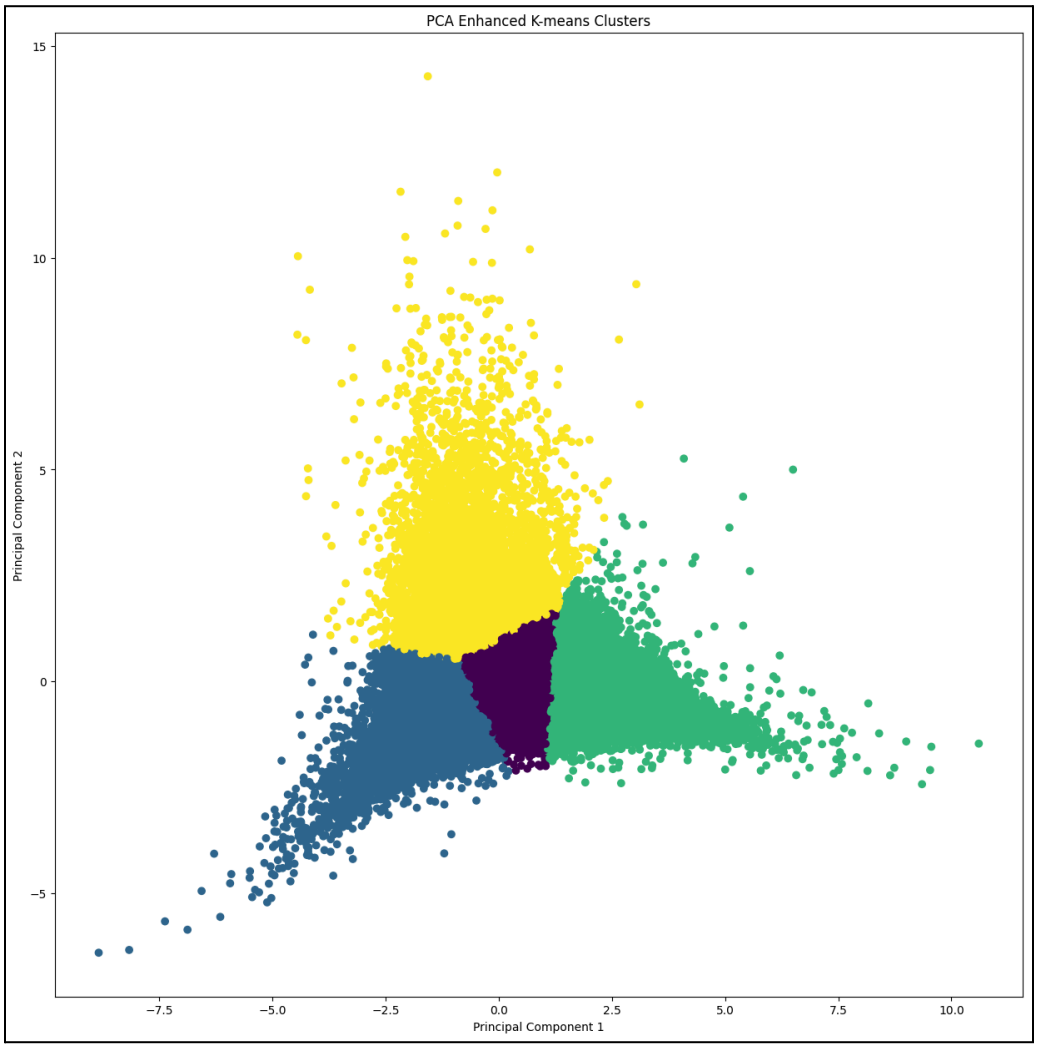


Proton Density, Proton Temperature, He4toProtons plotted in x, y, z respectively

Feature importance was inconclusive through K-means using the features provided by the ACE Mission data. Principal Component Analysis (PCA) was the next option for feature exploration. To determine the appropriate amount of components we plotted the cumulative explained variance against the number of principal components. The explained variance ratio calculated by PCA shows the amount of variance on the dataset by a particular component. By taking the cumulative sum of this ratio at each additional principal component we can determine the number of components that will give us the target variance desired. For our case, we looked for the number of components getting us as close to 80% as possible. Please refer to the appendix figure for results. Cumulative variance ratio determined 2 principal components are appropriate for the data.

PCA was applied to the scaled ACE dataset at 2 components and then fed into our K-means model. CHI and DBI both improved significantly from previous evaluations indicating stronger clusters were generated. The new cluster labels were joined with the ACE dataset and plotted with PC1 and PC2. Figure 5 shows the PCA clusters.

Figure 7



K-Means clustering using two principal components

PCA enhanced K-means provides more distinct clusters, however clusters are still close together. The ACE Mission data along with PC1 and PC2 were tested with the HCS index forecast model. Please refer to the forecasting section for results.

DBSCAN

K-means clustering displayed clusters close together without a determining feature of significance from the ACE data. DBSCAN does not require a parameter number of clusters to be set for training, but does require a value for epsilon. Epsilon is the radial distance between points in a cluster and determines how DBSCAN clusters points neighboring one another. Since the 4 features provided by ACE vary in scale, normalization with Standard Scaler was applied. A rule-of-thumb assumption (num_features + 1) for the minimum points was used, setting minimum points to 5.

Epsilon values ranging from 0.1 to 1.0 in increments of 0.1 were iterated in various DBSCAN models using a sample from the scaled ACE dataset. A sample size of 15% of original data was used for iteration to save on computation. Silhouette score, number of clusters, and number of datapoints labeled as noise was collected at each epsilon value (see table in the appendix). An epsilon value of 0.8 provided the best silhouette score at 5 clusters and only 160 (out of 18K) points labeled noise.

The cluster distribution was plotted to evaluate the effectiveness of DBSCAN. It was found that over 99% of all data points were clustered together with 4 other clusters containing 7 or less points (see figure in the appendix).

To verify if a more optimal epsilon value should be utilized, nearest neighbor distance was plotted across the full ACE dataset[11]. When evaluating the nearest neighbor distance, an approximate epsilon value of 0.4 was found to be optimal (see figure in the appendix). DBSCAN was run at epsilon set to 0.4 and the cluster distribution was once again plotted. Over 99% of all data points were clustered in one cluster while 31 other clusters contained 16 or less data points in each (see figure in the appendix).

DBSCAN was unable to highlight features significant for pattern recognition or finding clusters able to classify types of solar wind from the ACE mission data.

Forecasting

Data Partitioning

We created two models, one for each of the outcome variables (SD_70 & SL_70_log_10), where the monthly sunspot numbers, the date, and the lag columns of the outcome variables were used as features. We partitioned the data sequentially to avoid data leakage. Additionally, to ensure that our models did not inadvertently use data from the training set, we omitted the first row of the test set because it included lagged values from the previous month.

Model Creation

In order to decide on the “best model” for each of the outcome variables, we used the scikit-learn package to compare multiple algorithms using, and evaluate them using different accuracy metrics. Initially, we obtained base models with suspiciously high R^2 values. To address this, we implemented cross-validation and the averaged of the evaluation metrics for the different folds to re-assess performance, which provided a more robust evaluation (Tables 4 & 5).

That being said, for both outcome variables, Linear Regression and Ridge Regression emerged as the best performing algorithms with R^2 scores of 0.78 for SD_70 and 0.59 for SL_70_log10 (Tables 4 & 5). We thought ensemble models would perform better but, given the data size and the limited number of features available, these algorithms generally don’t perform as well.

Table 4: Base model evaluation cross-validation metrics for SD_70					
Model	**Mean CV R^2**	**Std Dev CV R^2**	**Test STD MAE**	**Test STD MSE**	**Test STD RMSE**
LinearRegression	0.784798	0.151703	1.489568	3.725764	1.930224
Ridge	0.784651	0.150915	1.495619	3.735733	1.932804
ExtraTreesRegressor	0.753322	0.167286	1.903254	5.732117	2.394184
GradientBoostingRegressor	0.741474	0.178344	2.329877	7.628636	2.761998
RandomForestRegressor	0.726895	0.197591	1.743157	4.952106	2.225333
Lasso	0.722385	0.184107	1.717816	4.847422	2.201686
DecisionTreeRegressor	0.452831	0.259352	2.619378	10.603354	3.256279

Table 5: Base model evaluation cross-validation metrics for SL_70_log 10					
Model	**Mean CV R^2**	**Std Dev CV R^2**	**Test STD MAE**	**Test STD MSE**	**Test STD RMSE**
Ridge	0.594941	0.287875	0.217282	0.082393	0.287041
LinearRegression	0.594879	0.287852	0.216890	0.082218	0.286736
ExtraTreesRegressor	0.513920	0.353744	0.220380	0.085086	0.291695
RandomForestRegressor	0.507842	0.319461	0.268256	0.109820	0.331391
GradientBoostingRegressor	0.461919	0.327613	0.300369	0.129636	0.360051
DecisionTreeRegressor	0.084178	0.790332	0.340981	0.210452	0.458750
Lasso	-0.524541	0.679005	0.598394	0.455507	0.674913

Hyperparameter Tuning & Cross-Validation

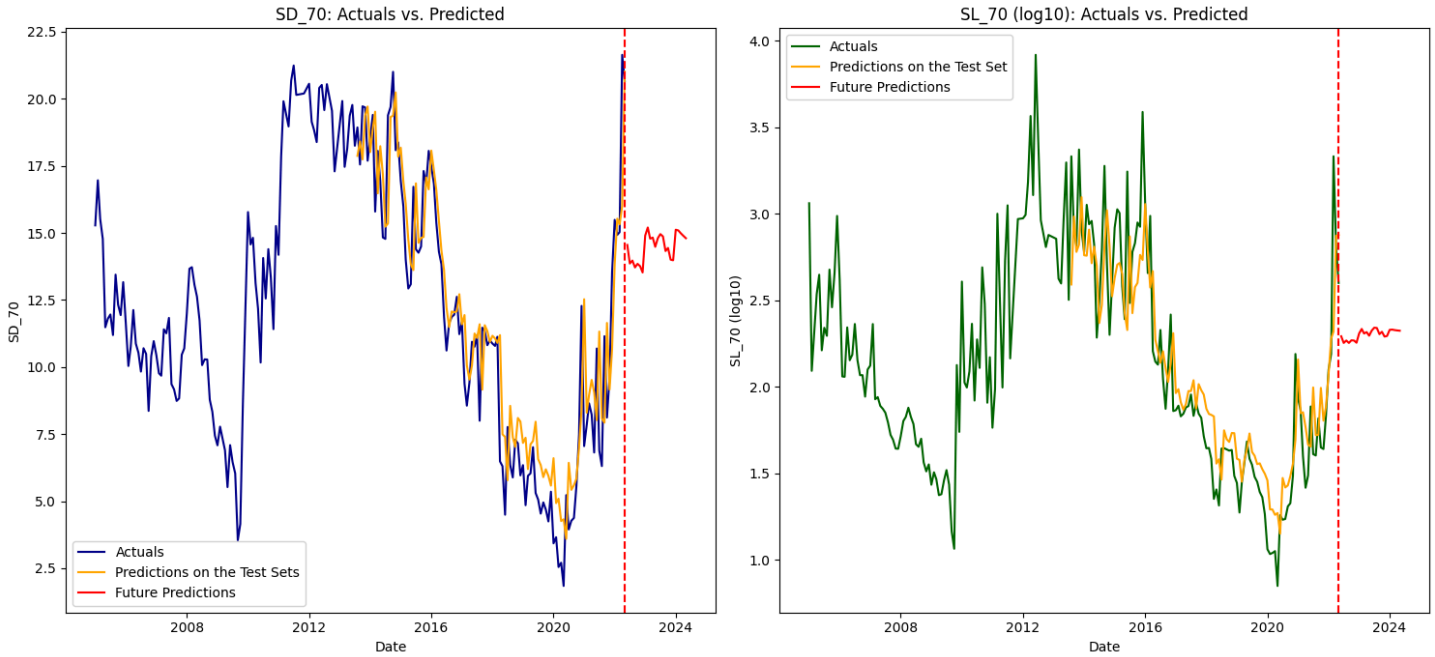
Given that Ridge Regression is particularly suited for hyperparameter tuning, we ran a grid search to identify the optimal parameters for each model. Following this, the best-tuned models were used to conduct cross-validation. Even though the results were slightly better than the base models (Table 6), we noticed inconsistencies in performance across various folds; some displayed high accuracy while others were notably lower. These variations underscore a critical need for robustness in the dataset, which is essential for building more dependable models.

Table 6: Tuned Ridge Regression Cross Validation Metrics					
Outcome Variable	**Mean CV R^2**	**STD Dev CV R^2**	**Test MAE**	**Test MSE**	**Test RMSE**
SD_70	0.784802	0.151703	1.903254	5.732117	2.394184
SL_70_log_10	0.595062	0.287922	0.220380	0.085086	0.291695

Generating Future Forecasts

We generated predictions for both outcome variables over the next 24 months using the best-tuned models that were trained up to May 2022. To support these predictions, we utilized forecasted and historical features. Specifically, forecasted and historical sunspot numbers (after May 2022) were obtained from the Space Weather Prediction Center at the NOAA.[10] Additionally, we iteratively predicted the lag columns. The results for the predictions on the test set and the future predictions are depicted on Figure 8.

Figure 8



Side-by-side line charts displaying actual data, predictions on the test set, and forecasts for the next 24 months post-May 2022. Left chart: SD_70 series. Right chart: SL_70_log10 series.

Conclusion

We found four different categories of solar wind using our clustering model. Future investigations into what makes these clusters distinct from each other can help predict when those types of winds are mostly likely to occur as well as evaluate the risk they pose to Earth. In addition, future data could provide better features for forecasting HCS Indexes.

Ridge Regression was used to forecast HCS indexes using lagged columns and monthly sunspot numbers. The optimized models achieved mean cross-validated R^2 values of 0.78 for SD_70 and 0.59 for SL_70_log10 on the test set. As more HCS index data is collected, models may be updated to learn on a more robust data set. The addition of other features such as solar wind data should be explored in order to enhance forecast accuracy. We hope that this helps set the stage for more comprehensive and precise predictive models in the field in the future.

Ethical considerations

It's unclear how a stronger understanding of the solar wind cycles and HCS could be used for malpractice. In the context of predicting geomagnetic storms and how they affect satellite equipment and orbital activity, political or terrorist motives may be able to plan around the solar cycles to cause damage. While our work does not successfully map the solar cycle it could be used to build upon other research and lead to a greater understanding of the HCS.

Statement of Work

- Project Proposal: All
- Data Acquisition: Arturo
- Github Setup: Brandyn
- EDA: Andrea, Arturo
- ACE Models: Andrea, Brandyn
- HCS/SunSpot Models: Arturo
- Final Report: All
- Final Video: All

Our team worked together well and was efficient and constructive in our communication and feedback. We collaborated in ways that enhanced the workflow and results of our project.

References

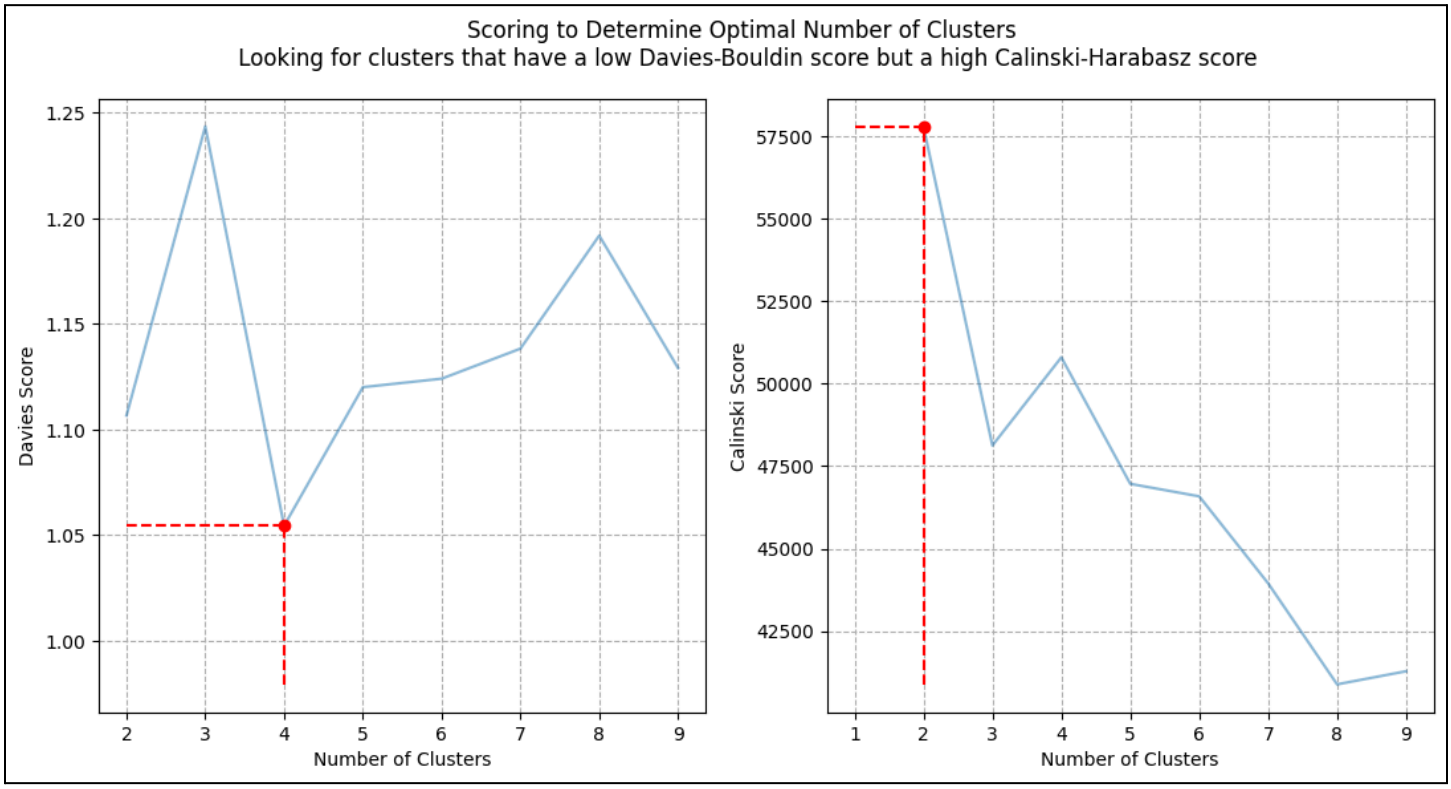
1. The heliospheric current sheet. (2001, August 01). *Journal of Geophysical Research*, *106*(A8), 15819-15831. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JA000120
2. *The Sun's Magnetic Field is about to Flip*. (2013, August 6). NASA. Retrieved April 14, 2024, from https://www.nasa.gov/science-research/heliophysics/the-suns-magnetic-field-is-about-to-flip/
3. Two novel parameters to evaluate the global complexity of the Sun's magnetic field and track the solar cycle. (2013, August). *The Astrophysical Journal*, *773*(2), 157. https://www.researchgate.net/publication/258259458_Two_novel_parameters_to_evaluate_the_global_complexity_of_the_Sun's_magnetic_field_and_track_the_solar_cycle
4. *Space Technology 5*. (n.d.). Space Technology 5. Retrieved April 15, 2024, from https://www.jpl.nasa.gov/nmp/st5/SCIENCE/solarwind.html
5. Lerner, L. (n.d.). *What is the solar wind?* UChicago News. Retrieved April 8, 2024, from https://news.uchicago.edu/explainer/what-is-solar-wind
6. NASA. (2023, October 12). *ACE*. ACE Mission. https://science.nasa.gov/mission/ace
7. *SWEPAM Level 2 Data Documentation*. ACE/SWEPAM Level 2 Data. (2007, November 7). https://izw1.caltech.edu/ACE/ASC/level2/swepam_l2desc.html
   National Oceanic and Atmospheric Administration. (2013, September 1). *Readme: Sunspot numbers - National Geophysical Data Center*. NOAA.gov. https://ngdc.noaa.gov/stp/space-weather/solar-data/solar-indices/sunspot-numbers/documentation/readme_sunspot-numbers.pdf
8. Solar Cycle 25 Prediction Panel. (n.d.). Predicted solar cycle: Sunspot number and radio flux values with expected ranges. S.I.D.C. Brussels International Sunspot Number; Penticton, B.C., Canada: 10.7 cm radio flux values. Retrieved April 14 from https://www.swpc.noaa.gov/products/solar-cycle-progression
9. *Monthly mean total sunspot number README*. SIDC. (n.d.) https://www.sidc.be/SILSO/infosnmtot
10. Sankalana, N. (2023, September 20). *K-means clustering: Choosing optimal K, process, and evaluation*

methods. Medium. https://medium.com/@nirmalsankalana/k-means-clustering-choosing-optimal-k-process-and-evaluation-methods-2c69377a7ee4

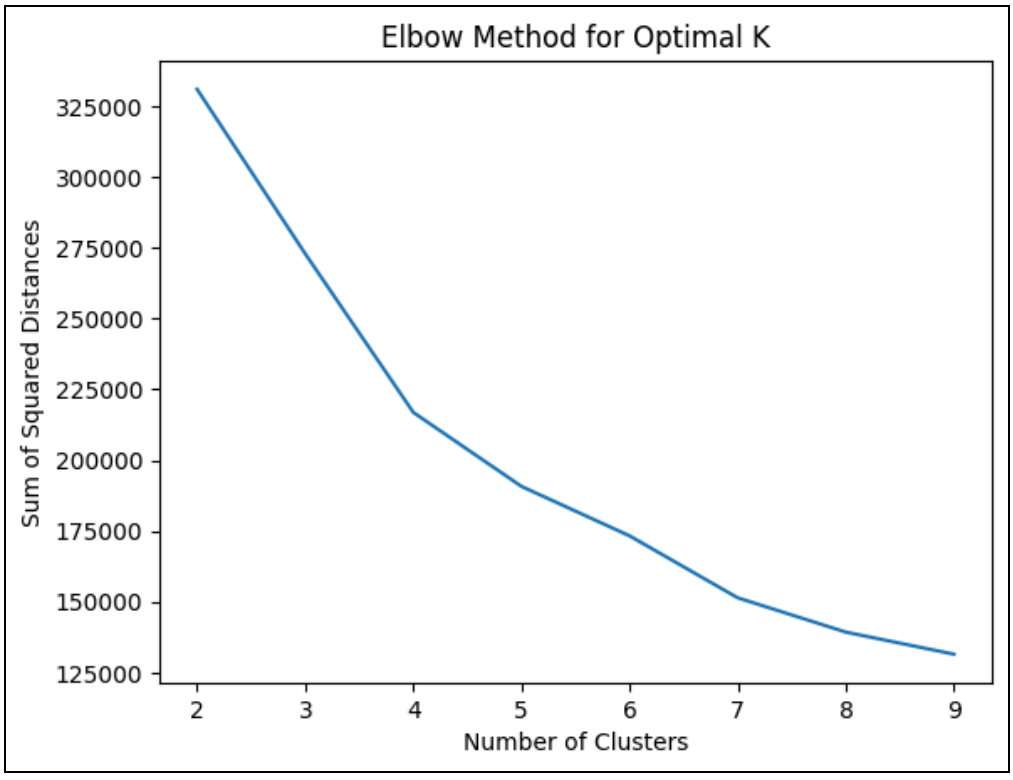
11. *How to determine Epsilon and MinPts parameters of DBSCAN clustering*. Amir Masoud Sefidian - Sefidian Academy. (2023, April 25). https://sefidian.com/2022/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/

Appendix

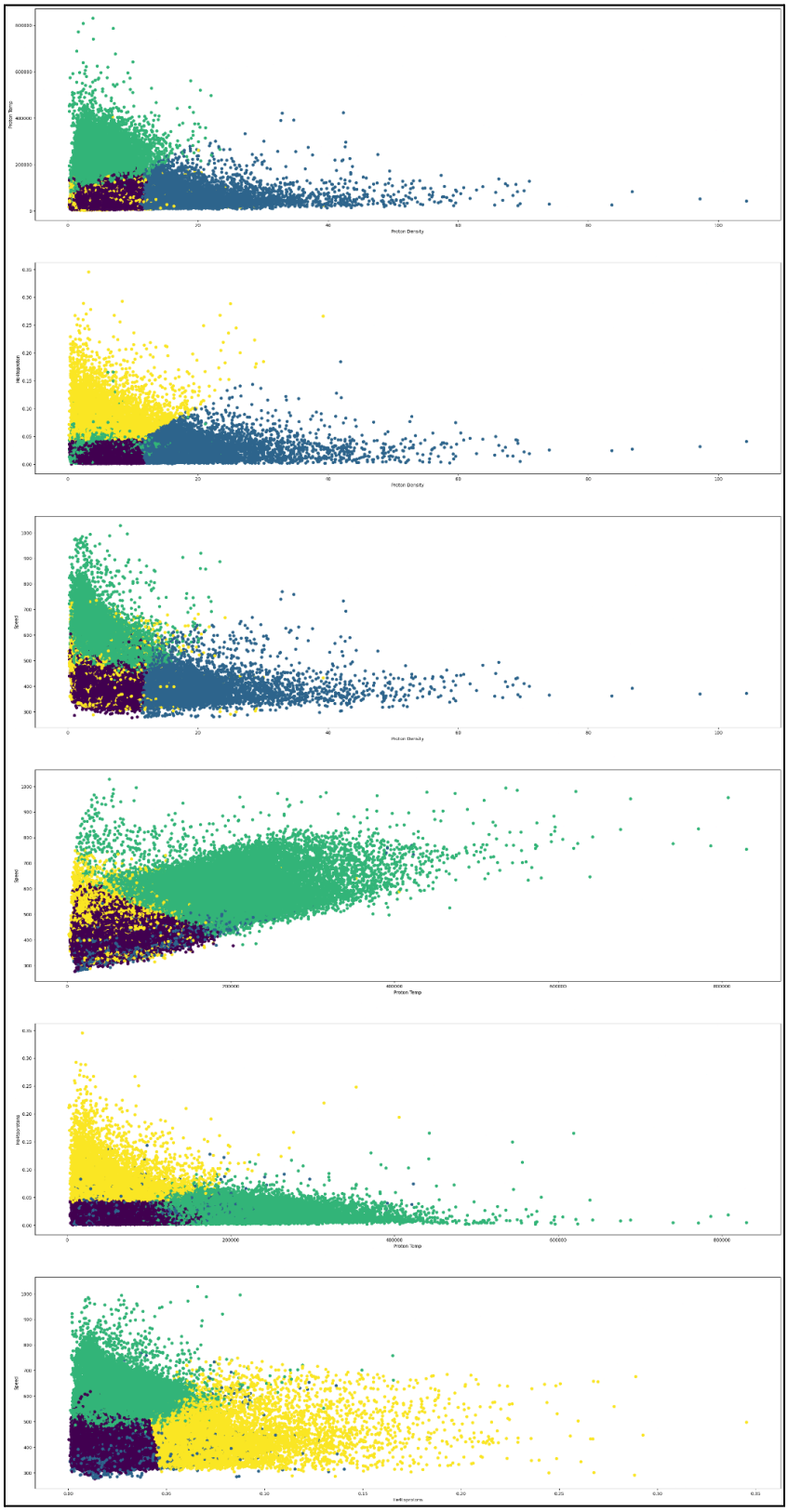
DBI/CHI vs Number of Clusters to Determine Optimal k



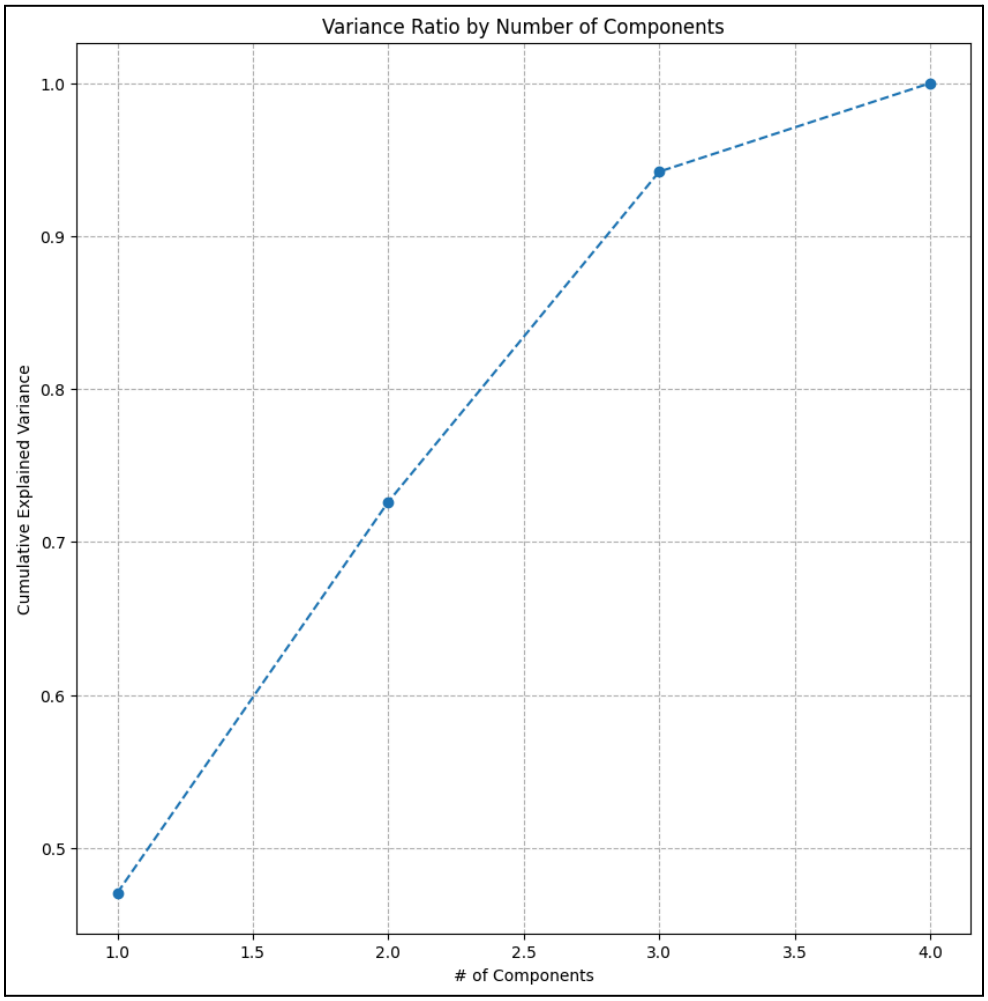
Elbow Method



Iterative Pair Combination Cluster Plots



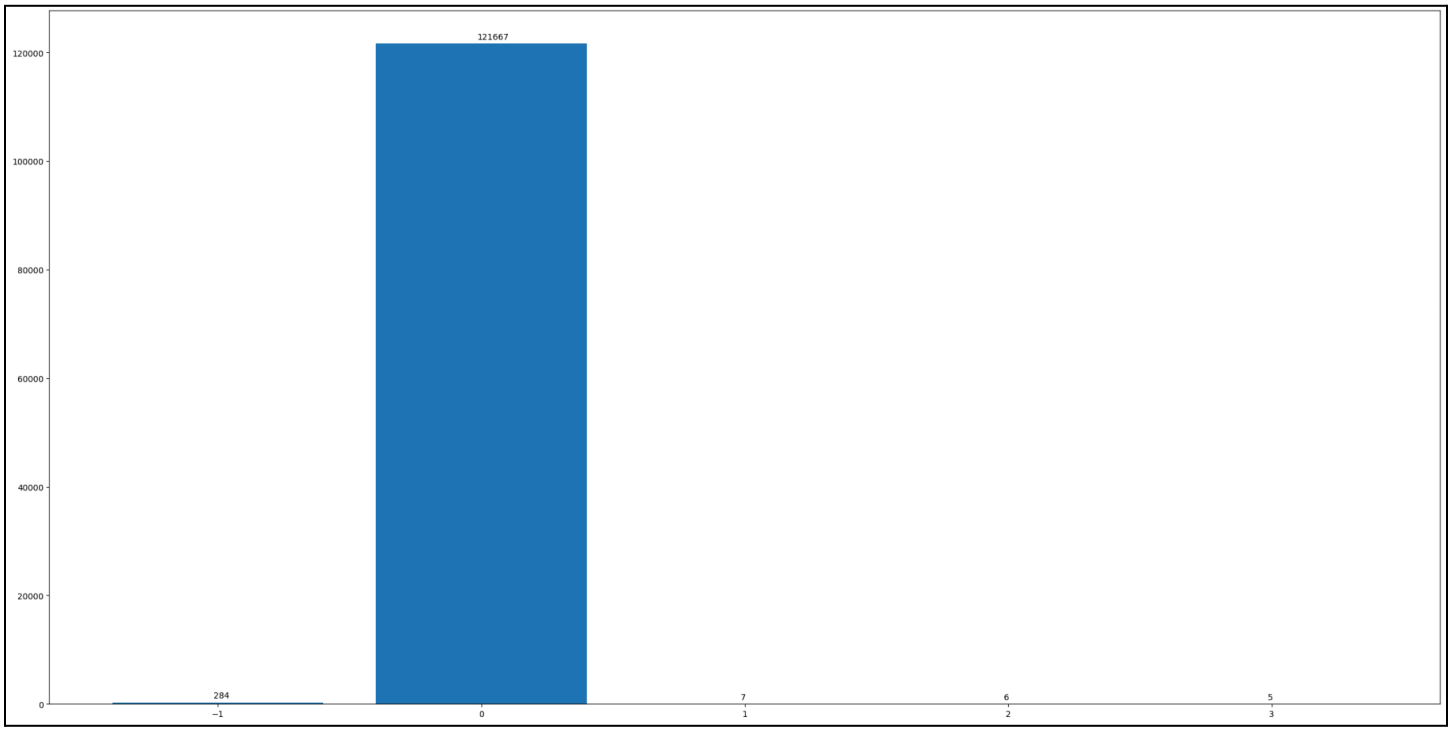
Cumulative Variance Ratio for Optimal Principal Components



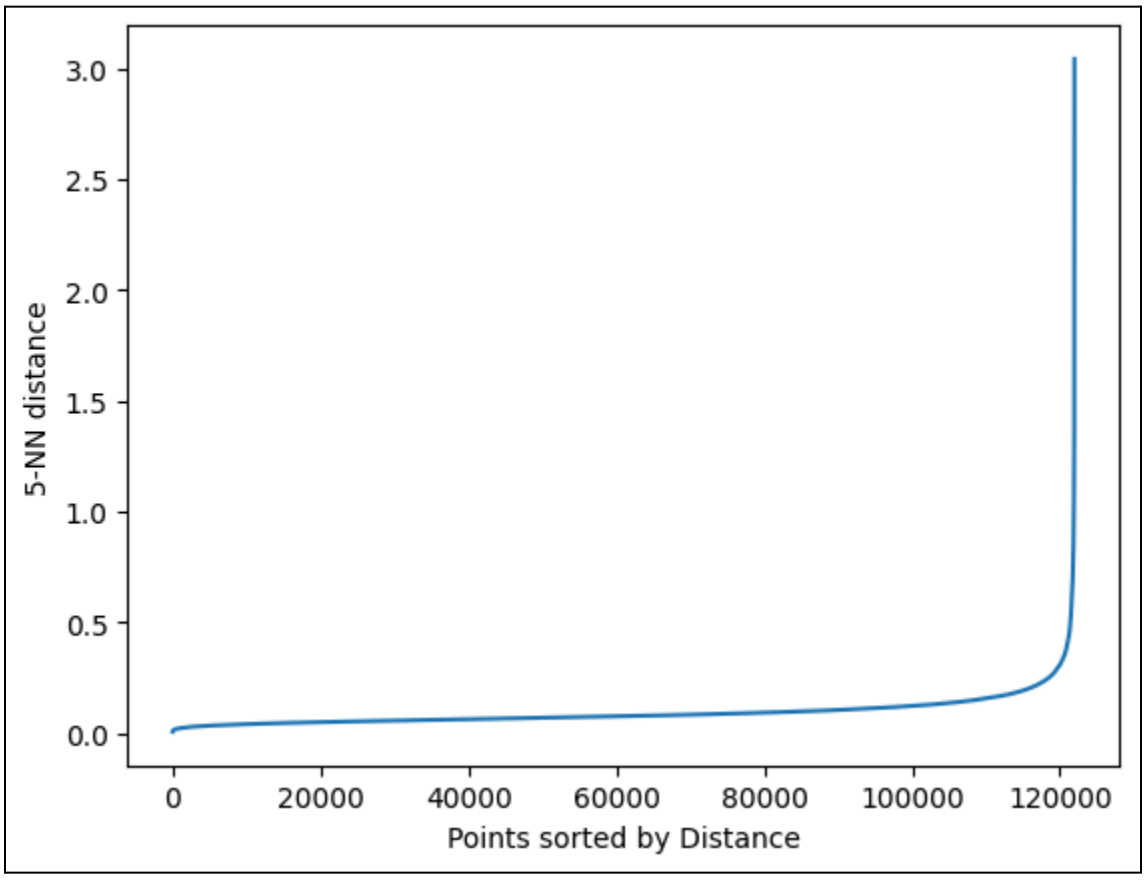
Epsilon at n Clusters

Epsilon at n Clusters				
epsilon	min_samples	Silhouette Score	Noise	#_clusters
0.1	5	-0.627849	17615	97
0.2	5	-0.586816	5934	89
0.3	5	-0.353779	2102	36
0.4	5	0.071716	1032	9
0.5	5	-0.001235	547	15
0.6	5	0.30411	335	8
0.7	5	0.296112	240	6
0.8	5	0.480025	160	5
0.9	5	0.420209	120	6
1	5	0.429544	75	3

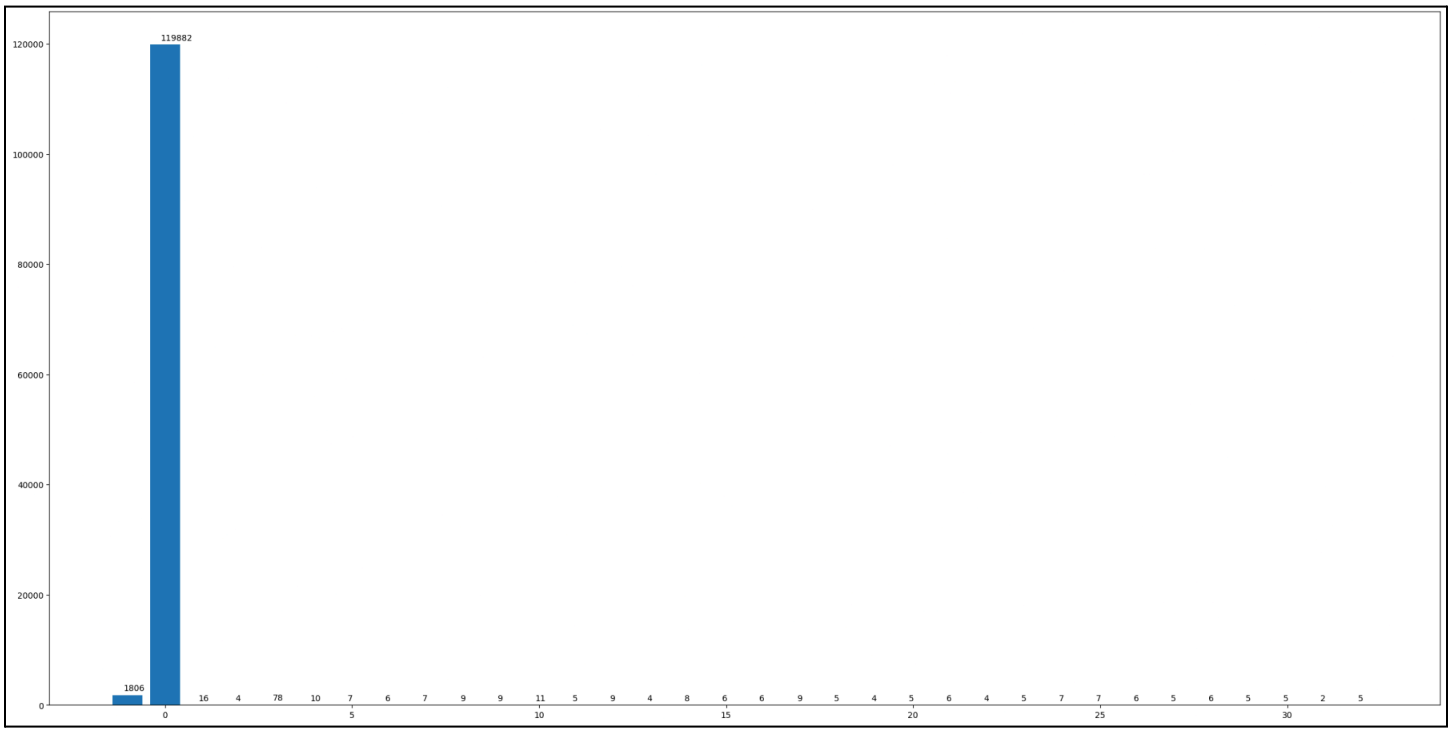
Data Point Distribution in Clusters at DBSCAN Epsilon 0.8



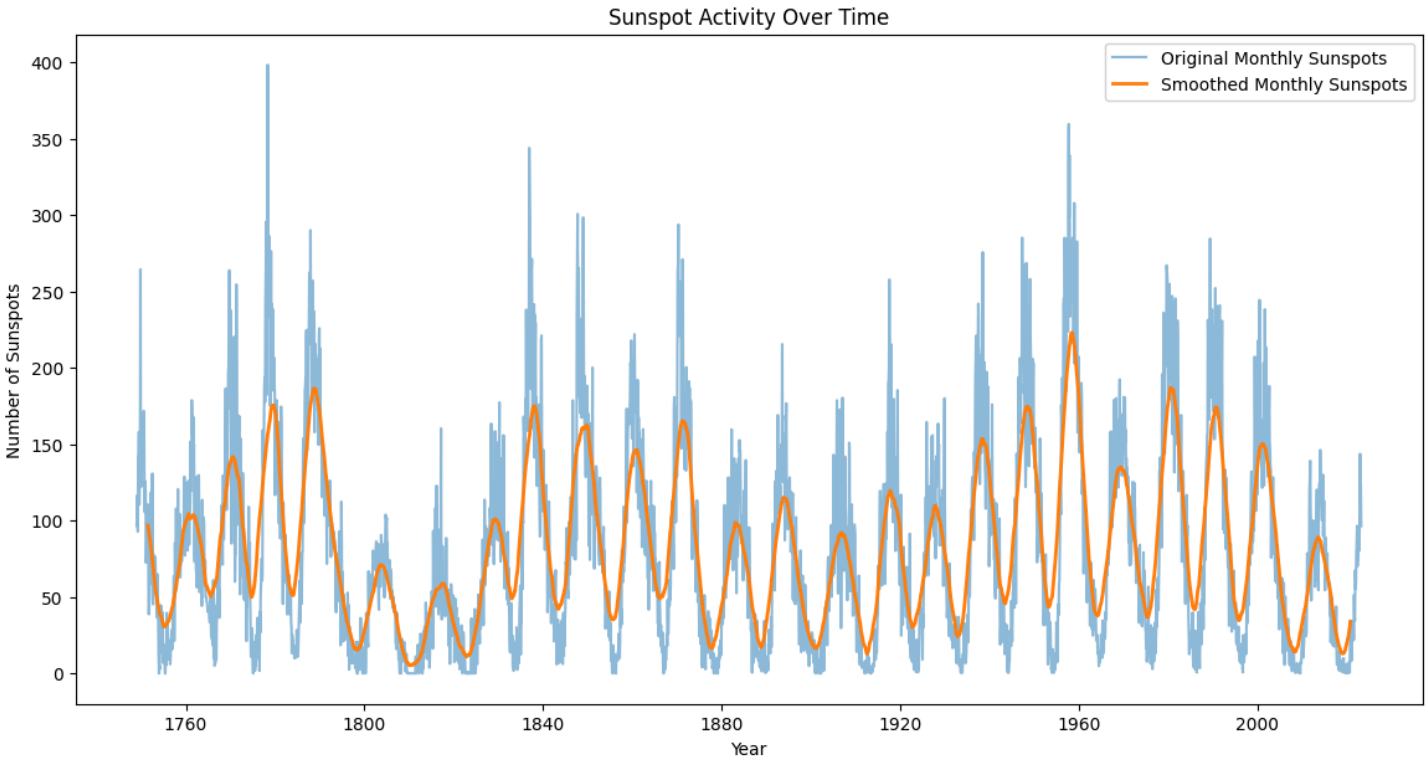
5 Nearest Neighbors Distance at Each Data Point



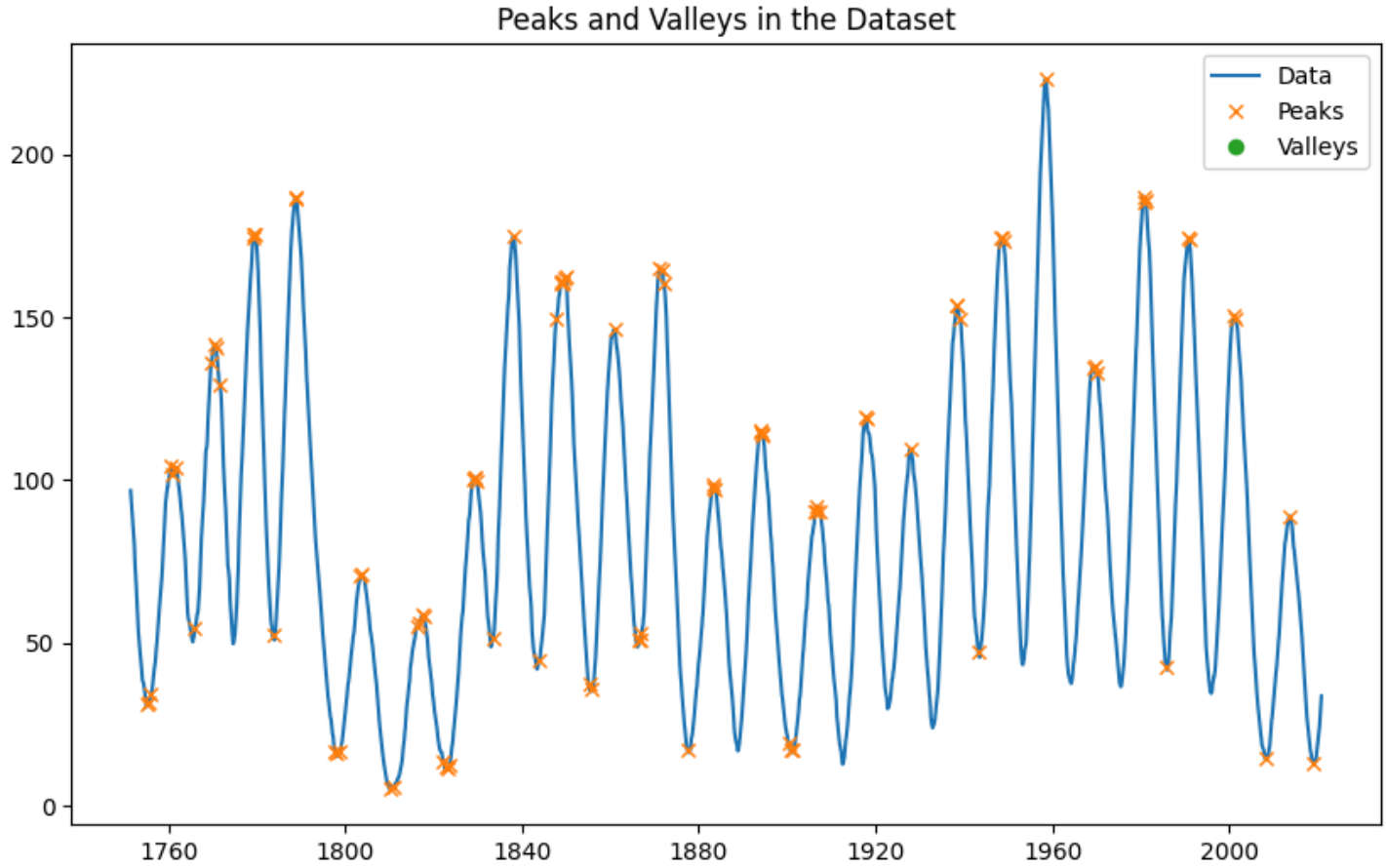
Data Point Distribution in Clusters at DBSCAN Epsilon 0.4



Smoothed Out Monthly Sunspot Numbers



Peaks and Valley of Smoothed Monthly Sunspot Numbers



Solar Cycles using 11-year average