Descriptive analysis: potential predictors for hurricane forecasting

File Description

This file contains a series of plots that are part of the exploratory analysis conducted for the paper "Simple yet Effective: A Comparative Study of Statistical Models for Yearly Hurricane Forecasting".

1. **Correlation Plots**
   - This section contains plots illustrating the correlation between various indexes and hurricane counts.
2. **Exploratory Plots**
   - This section includes plots related to specific variables and approaches used in the exploratory analysis.
3. **Cluster Analysis**
   - This section presents the results of a cluster analysis performed on the data.
4. **Principal Component Analysis**
   - This section includes the principal component analysis, highlighting the main components influencing hurricane counts.

The purpose of this file is to explicitly show some of the considerations developed during the variable selection process.

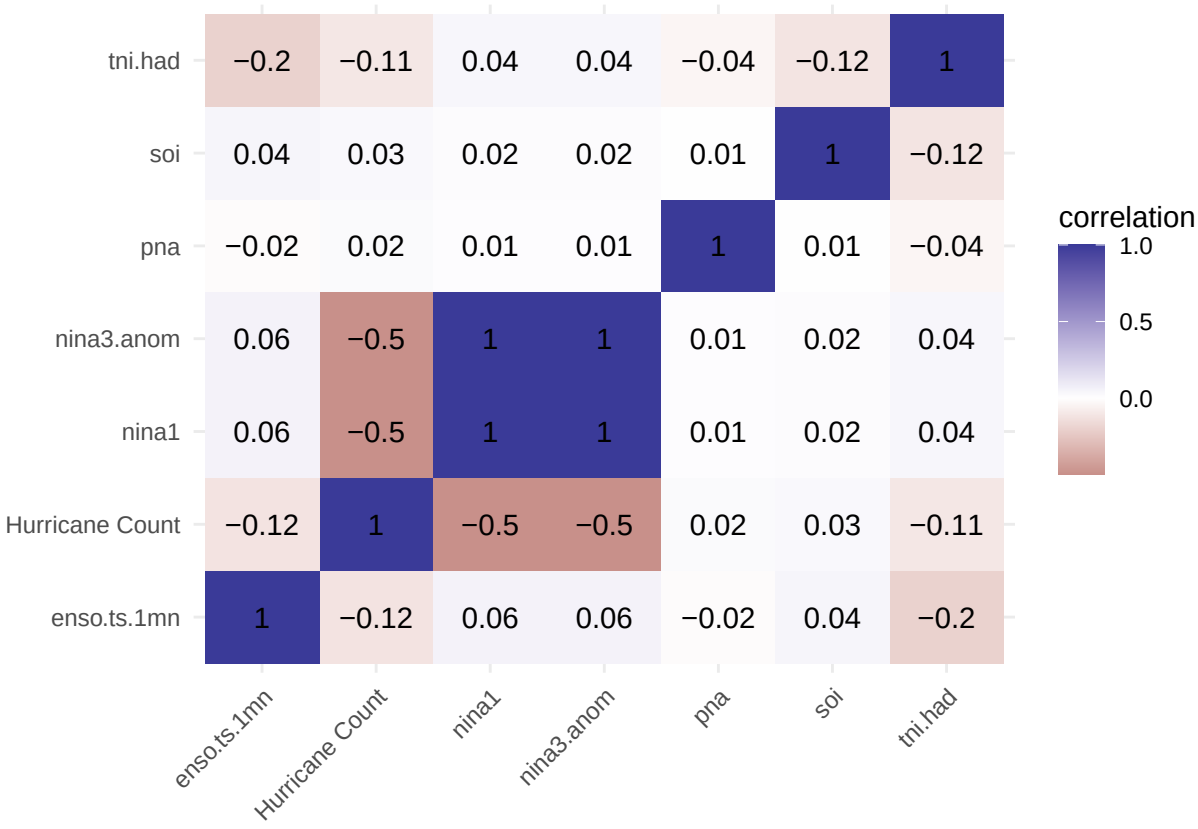
Data Collection

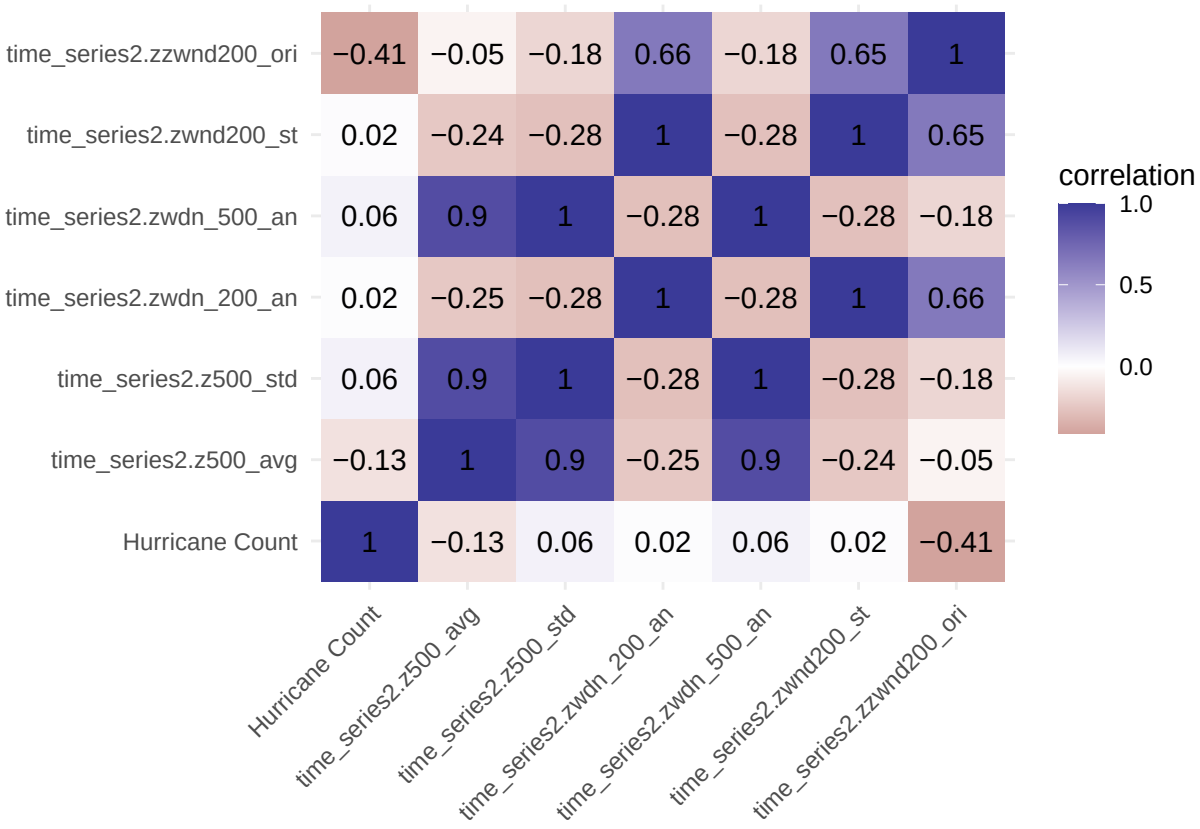
The dataset used in this experiment consisted of time-series data from 1979 to 2020, including: - **Hurricane Counts**: Number of hurricanes per year. - **200mb Zonal Wind Anomalies**: Deviation in the wind component at the 200mb pressure level over the central equatorial Pacific. - **Nino3.4 Anomalies**: Sea surface temperature anomalies in the Nino3.4 region, which is crucial for monitoring El Niño and La Niña events.

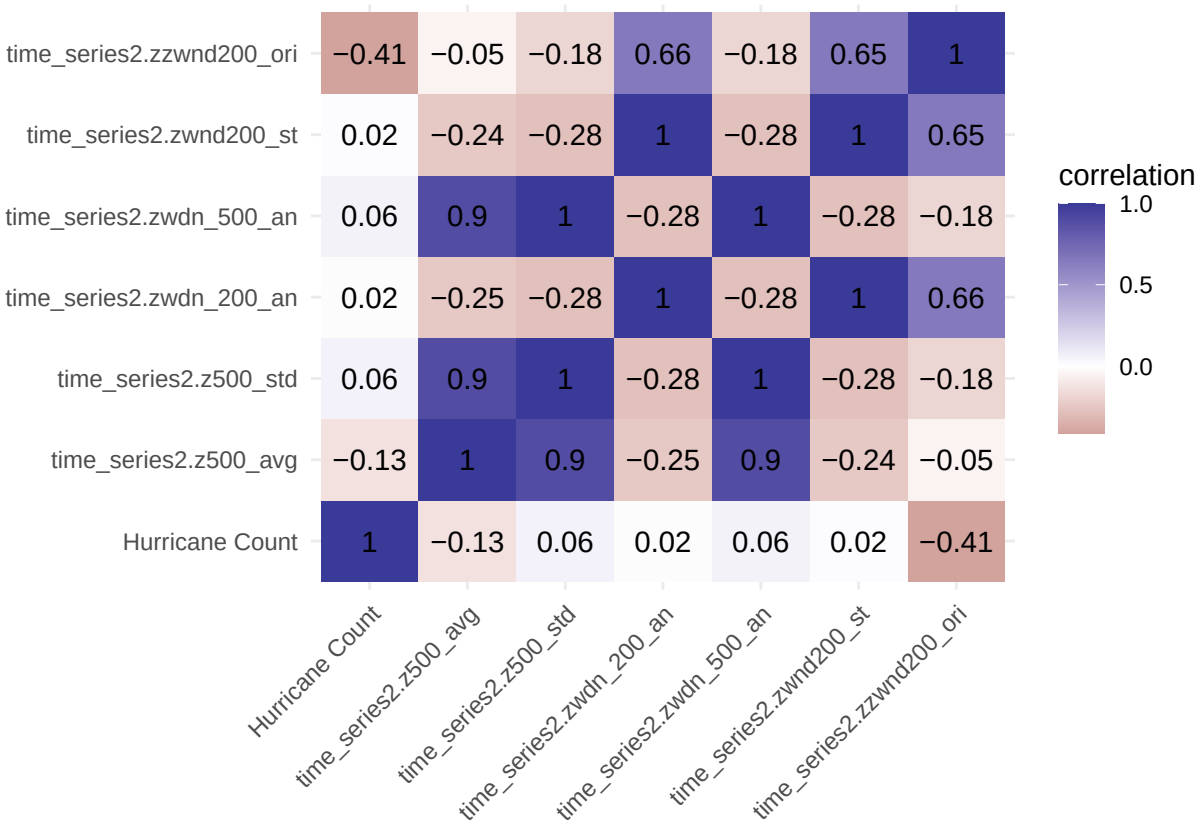
All data was preprocessed to remove missing values and standardized to ensure uniformity in scale.

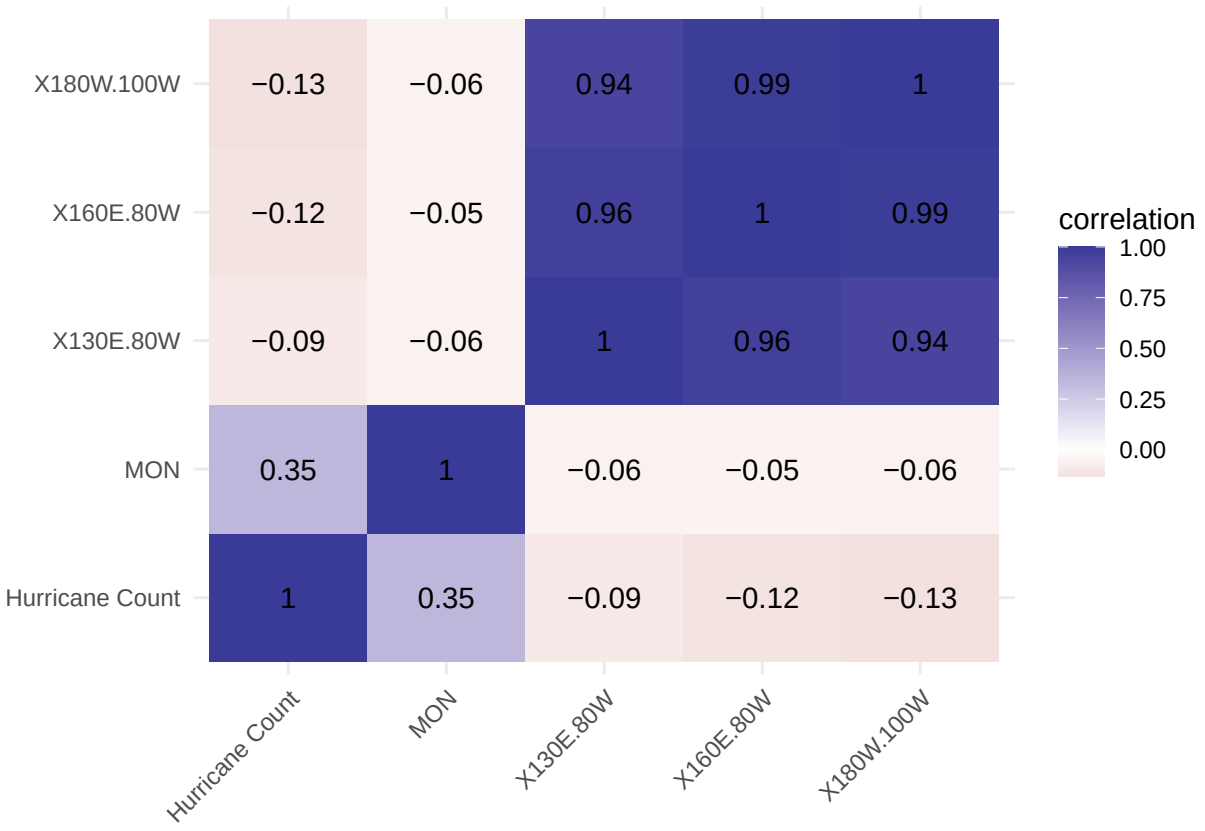
CORRELATION PLOT

1. What indexes are linearly correlated with the Hurricanes counts from 1979?


tni.had
soi
pna
nina3.anom
nina1
Hurricane Count
enso.ts.1mn
−0.2 −0.11 0.04 0.04 −0.04 −0.12 1
0.04 0.03 0.02 0.02 0.01 1 −0.12
−0.02 0.02 0.01 0.01 1 0.01 −0.04
0.06 −0.5 1 1 0.01 0.02 0.04
0.06 −0.5 1 1 0.01 0.02 0.04
−0.12 1 −0.5 −0.5 0.02 0.03 −0.11
1 −0.12 0.06 0.06 −0.02 0.04 −0.2
enso.ts.1mn
Hurricane Count
nina1
nina3.anom
pna
soi
tni.had
correlation
1.0
0.5
0.0


time_series2.zzwnd200_ori
time_series2.zwnd200_st
time_series2.zwdn_500_an
time_series2.zwdn_200_an
time_series2.z500_std
time_series2.z500_avg
Hurricane Count
−0.41 −0.05 −0.18 0.66 −0.18 0.65 1
0.02 −0.24 −0.28 1 −0.28 1 0.65
0.06 0.9 1 −0.28 1 −0.28 −0.18
0.02 −0.25 −0.28 1 −0.28 1 0.66
0.06 0.9 1 −0.28 1 −0.28 −0.18
−0.13 1 0.9 −0.25 0.9 −0.24 −0.05
1 −0.13 0.06 0.02 0.06 0.02 −0.41
Hurricane Count
time_series2.z500_avg
time_series2.z500_std
time_series2.zwdn_200_an
time_series2.zwdn_500_an
time_series2.zwnd200_st
time_series2.zzwnd200_ori
correlation
1.0
0.5
0.0


time_series2.zzwnd200_ori
time_series2.zwnd200_st
time_series2.zwdn_500_an
time_series2.zwdn_200_an
time_series2.z500_std
time_series2.z500_avg
Hurricane Count
−0.41 −0.05 −0.18 0.66 −0.18 0.65 1
0.02 −0.24 −0.28 1 −0.28 1 0.65
0.06 0.9 1 −0.28 1 −0.28 −0.18
0.02 −0.25 −0.28 1 −0.28 1 0.66
0.06 0.9 1 −0.28 1 −0.28 −0.18
−0.13 1 0.9 −0.25 0.9 −0.24 −0.05
1 −0.13 0.06 0.02 0.06 0.02 −0.41
Hurricane Count
time_series2.z500_avg
time_series2.z500_std
time_series2.zwdn_200_an
time_series2.zwdn_500_an
time_series2.zwnd200_st
time_series2.zzwnd200_ori
correlation
1.0
0.5
0.0



The correlation plot chunk contains 4 subfigures containing the linear correlations 21 variables downloded through the "https://psl.noaa.gov/enso/dashboard.html". A detailed explanation of each variable can be also find at the Noa website.

First CP - The first correlation plot shows a moderate negative correlation with two indexes "nina1" and "nina3.anom". The first index, beside the misleading label, represent the "Nino 4 Mean using ersstv5 from CPC". The second index instead represents the " Nino Anom 3.4 Index using ersstv5 from CPC". More info about both indexes are available "https://psl.noaa.gov/data/climateindices/list/. Yet preliminary, The"Nino indexes" are ameasure used to monitor sea surface temperature (SST) anomalies (mean) in the central equatorial Pacific region, specifically between 5°N-5°S latitude and 170°W-120°W longitude. The data for this index is derived from the NOAA's Extended Reconstruction Sea Surface Temperature version 5 (ERSSTv5), which uses a combination of buoy, satellite, and ship-based measurements to provide a comprehensive view of SSTs. The indexes are crucial for understanding and predicting El Niño and La Niña events.

Second CP - The second correlation plot shows a moderate linear correlation with ""time_series2.z500_std", refers to the standardized anomalies of the 500 hPa geopotential height (Z500). This index is used in meteorology and climate studies to understand variations in the atmospheric pressure at the 500 hPa level, which is approximately at the mid-troposphere, around 5.5 kilometers (18,000 feet) above sea level.

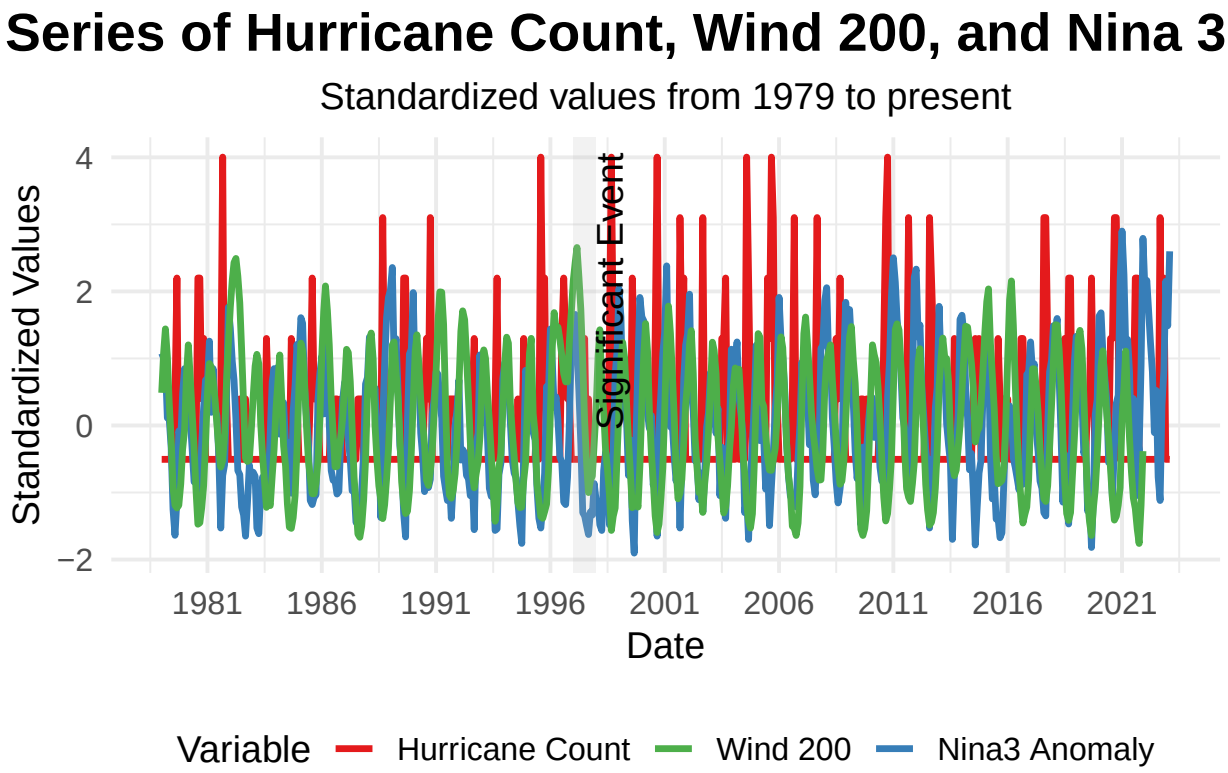
Third CP - The third plot shows a moderate low correlation with OLR index: Outgoing Longwave Radiation (OLR) area averages over the central equatorial Pacific (160°E-160°W).OLR and Convection: Low OLR values typically indicate high levels of convection and extensive cloud cover, as clouds and storms reflect and absorb much of the outgoing infrared radiation. Conversely, high OLR values suggest clear skies and less atmospheric convection. Hurricane Formation: Hurricanes form and intensify in regions of high atmospheric instability and significant convection. Thus, areas with persistently low OLR are conducive to the development of tropical cyclones, including hurricanes.

Fourth CP - The fourth plot does not shows anything significant as the label "MON" is referred to the

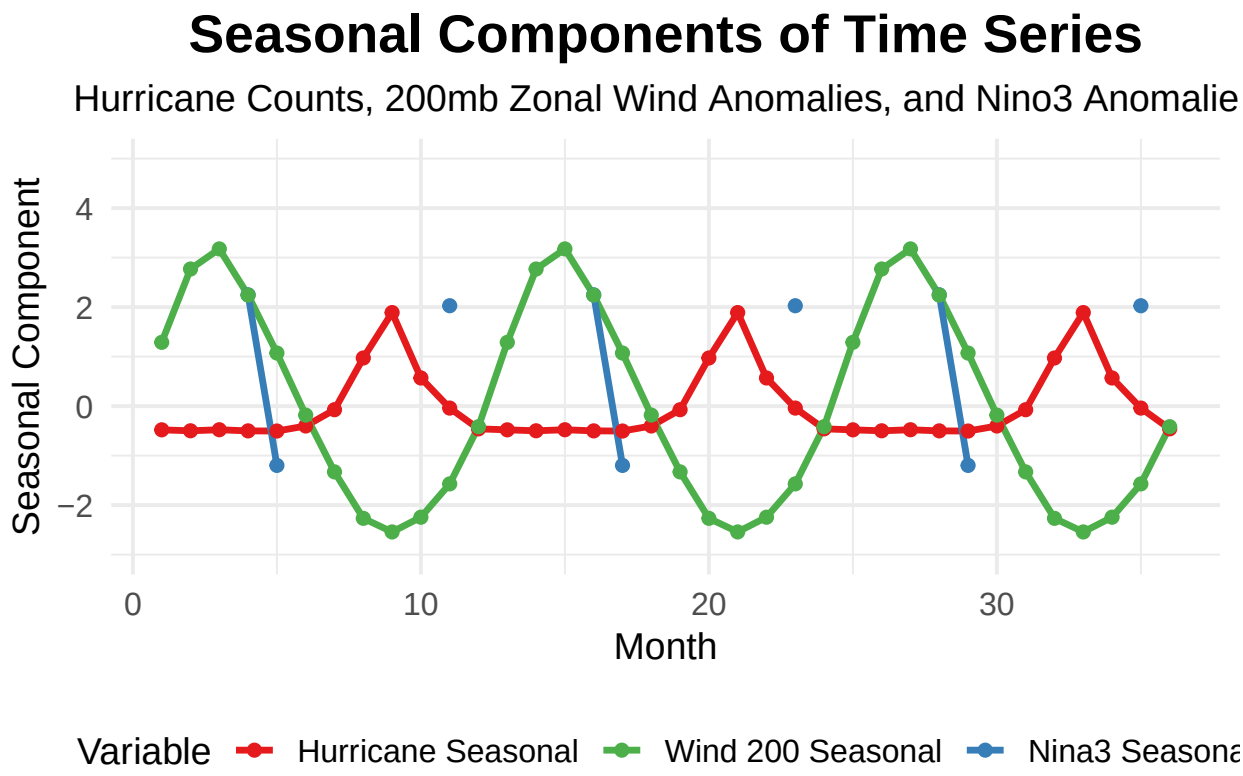
months occurrences. Therefore, a linear correlation with this time index indicates just and increasing pattern over time.

ADDITIONAL EXPLOTATORY PLOT

In this section a more detailed look to some of the variables of interest is investigated.



Source: https://psl.noaa.gov/data/climateindices/

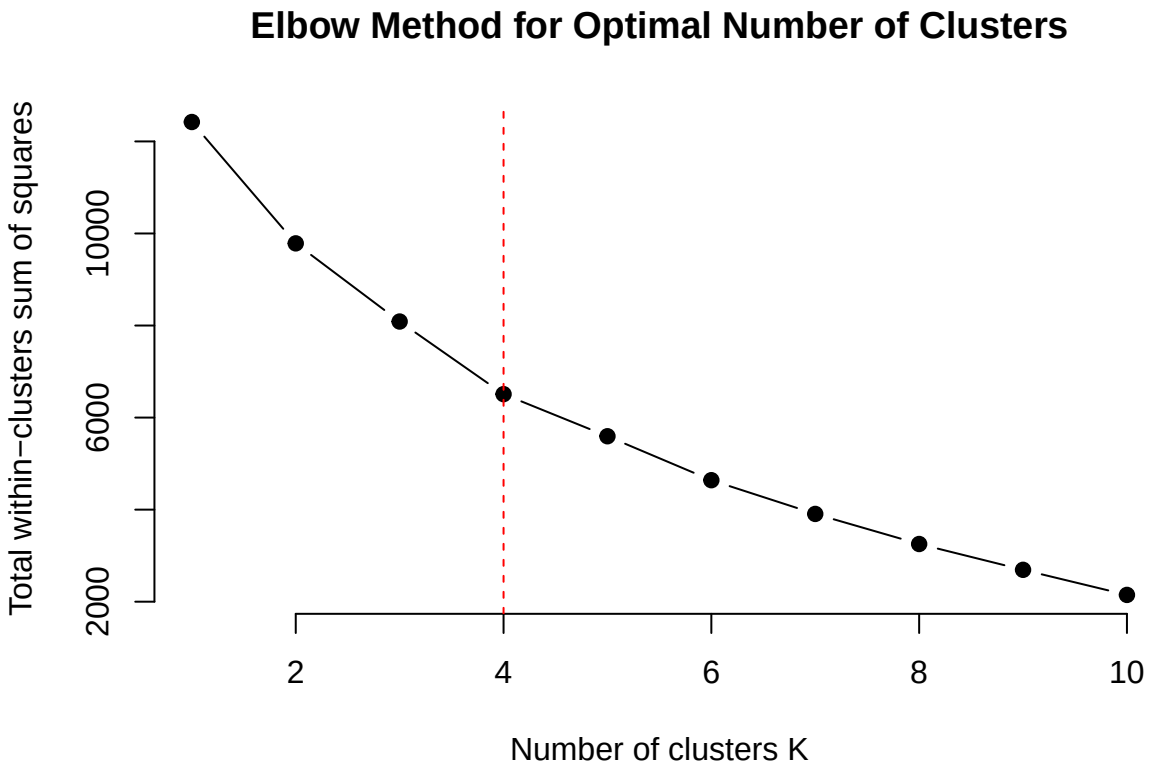


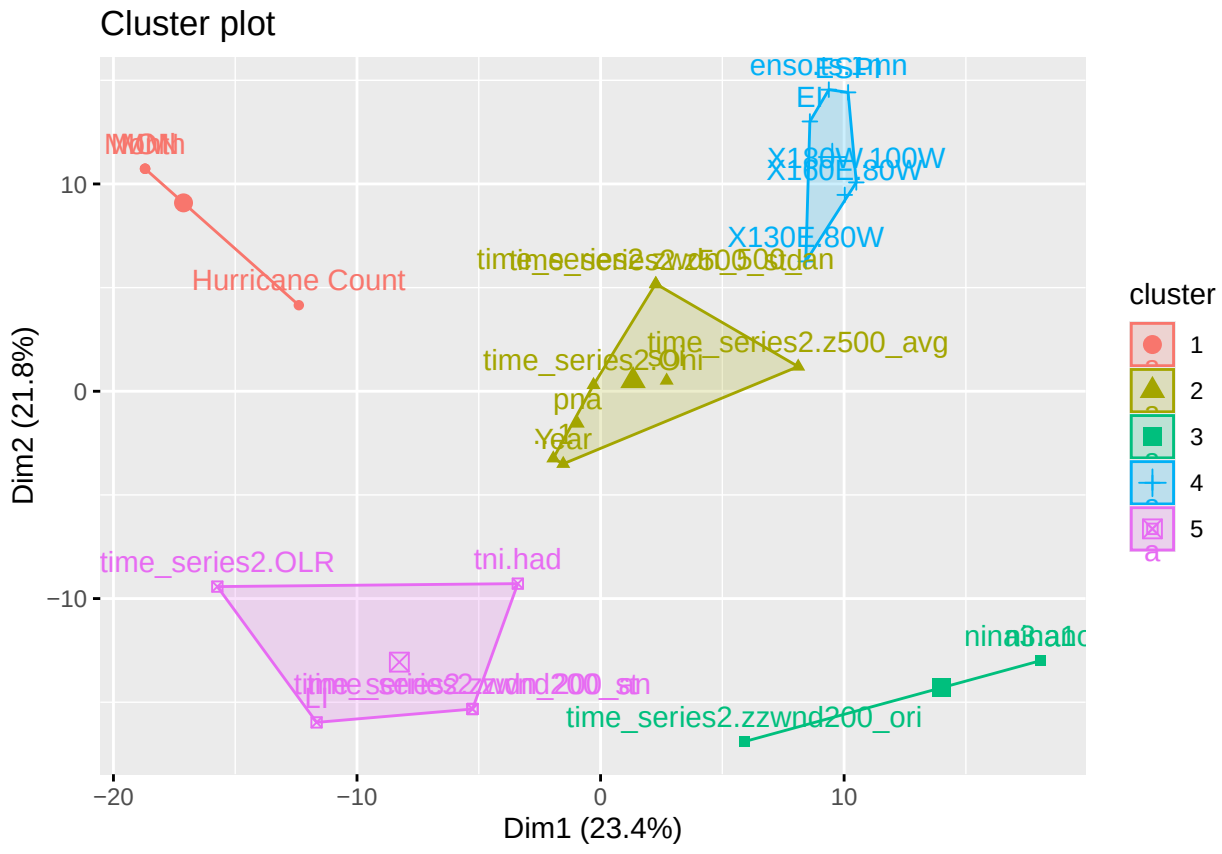
Source: https://psl.noaa.gov/data/climateindices/

ADDITIONAL PLOT INTEPRETATION

The first plot shows a closer look to the yearly hurricane occurence with two indexes: the wind zonal anomalies and the Nino anomalies (label Nina3 Anolmaly). It shows that the yearly hurricane occurence tend to be higher when the two indexes are low. 200mb Zonal Wind Anomalies refer to deviations in the east-west wind component (zonal wind) from the average conditions at the 200 millibar (mb) pressure level, which is approximately 12 kilometers (7.5 miles) above sea level in the upper troposphere.El Niño Years: During El Niño years, enhanced westerly wind anomalies in the central Pacific can increase vertical wind shear in the Atlantic, suppressing hurricane activity there while potentially increasing it in the central and eastern Pacific. La Niña Years: Conversely, during La Niña years, stronger easterly winds at the 200mb level can reduce vertical wind shear in the Atlantic, creating more favorable conditions for hurricane development.

The second plot shows a comparison of the seasonal decomposition of 3 variables: hurricanes counts, Zonal Wind and Nina3 anomaly. Notably, the 3 Zonal Wind and Nina3 anomaly present an opposite seasonality compared to the hurricane count decomposition. As decomposition methods, a simple moving average has been used.


Elbow Method for Optimal Number of Clusters
Total within−clusters sum of squares
2000
6000
10000
2
4
6
8
10
Number of clusters K


Cluster plot
Hurricane Count
time_series2.OLR
tni.had
time_series2.z500_avg
time_series2.zzwnd200_ori
X130E.80W
pna
Year
cluster
1
2
3
4
5
Dim1 (23.4%)
Dim2 (21.8%)

CLUSTERING EXPERIMENT INTEPRETATION

The primary objective of this clustering experiment was to identify natural groupings within a dataset of meteorological variables, focusing on uncovering patterns that could provide insights into the relationship between different atmospheric phenomena, such as hurricane counts and wind anomalies.

Methodology

K-means Clustering was selected for this experiment due to its efficiency and simplicity in partitioning the dataset into distinct groups. The experiment followed these steps:

1. **Data Preprocessing**:
    - Standardized the data to have zero mean and unit variance.
    - Transposed the data matrix to treat each variable as an observation, clustering them based on their temporal patterns.
2. **Determining the Number of Clusters**:
    - A series of k-means clustering runs were performed with varying numbers of clusters (k = 2 to 10). The optimal number of clusters was determined using the Elbow method, which involved plotting the within-cluster sum of squares against the number of clusters and identifying the "elbow" point where the rate of decrease sharply slows and through different multidimentional plots.
3. **K-means Clustering**:
    - The k-means algorithm was initialized with 100 random starts to mitigate the effect of random initialization. The algorithm iteratively assigned each variable to the nearest cluster center and updated the cluster centers until convergence.
4. **Evaluation and Interpretation**:
    - The resulting clusters were analyzed to identify patterns and similarities within each group. Variables in the same cluster were interpreted as having similar temporal behaviors.

Results and Analysis

The optimal number of clusters was found to be five. Each cluster exhibited unique characteristics:

- **Cluster 1**: Included the Hurricanes counts and the time variables (either annual or montly). Indicating the no actuall variable seems to have similar temporal pattern to the hurricane counts besides the time.
- **Cluster 2**: Comprised variables like OLR, LI, zwdn_200_an, tni. These climate variables are interconnected because they all play roles in understanding and predicting atmospheric and oceanic processes, particularly in the tropics. OLR provides a measure of convective activity, while LI specifically focuses on precipitation anomalies during La Niña events. The 200mb zonal wind anomalies offer insights into upper-level atmospheric circulation, and TNI helps monitor shifts in ENSO phases.
- **Cluster 3**: Contained variables: enso.ts.1mn , ESPI, EI, X130E.80W, X180W.100W. It is clear why ESPI index and EI index are in the same cluster as EPSI is formed using the EI, and it is also understandable why they are in the same cluster of ENSO cycle as they are defined "ENSO precipitation indexes". While the X130E.80W, X180W.100W, represents the equatorial Upper 300m Temperature Average Anomaly in a specific area of the pacific and are known to be related to ENSO Phase. The equatorial upper ocean temperature anomalies are closely related to ENSO phases. During El Niño,

warm water accumulates in the eastern and central Pacific, raising temperatures above average. Conversely, during La Niña, cooler-than-average water dominates, especially in the central and eastern Pacific.

- **Cluster 4**: Represented variables zzwnd200_ori, nina3.anom,nina1. The variables are interconnected through their associations with atmospheric and oceanic processes, particularly those related to the El Niño-Southern Oscillation (ENSO) phenomenon.
- **Cluster 5**: Included the following variables variables: ENSO Influence: ONI, SOI, and related variables like z500_std and z500_avg are influenced by ENSO phases. During El Niño, ONI is positive, SOI is negative, and there can be significant changes in z500 patterns due to altered atmospheric circulation, including shifts in the jet stream.

Atmospheric Circulation Patterns: zwdn_500_an and z500 variables (std and avg) are closely related to the structure and dynamics of the mid-troposphere. Changes in these variables can indicate alterations in storm tracks, the position of the jet stream, and other weather-related phenomena.

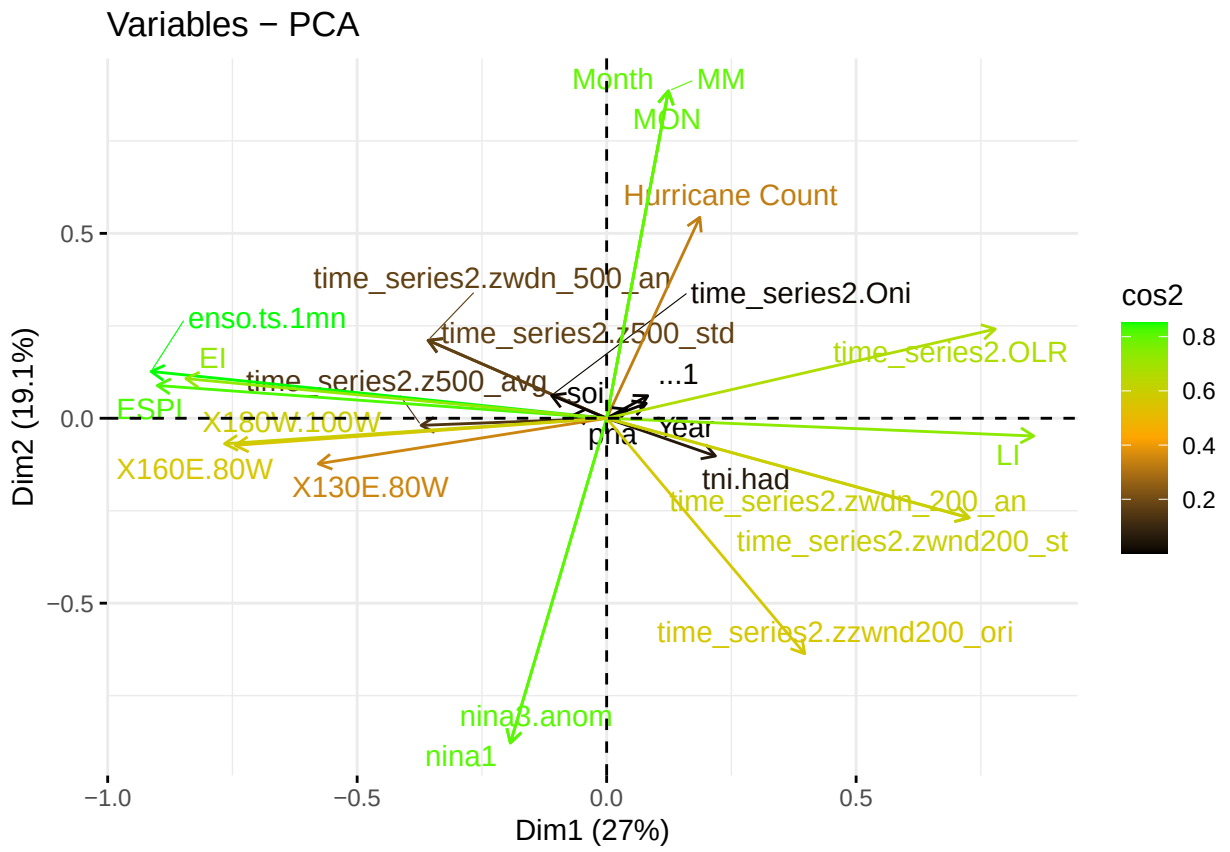
Teleconnection Patterns: PNA is a teleconnection pattern that describes the relationship between atmospheric pressure variations in the Pacific and North America. It interacts with ENSO, where certain phases of ENSO can enhance or suppress the PNA pattern. For example, El Niño conditions often correlate with a positive PNA phase.

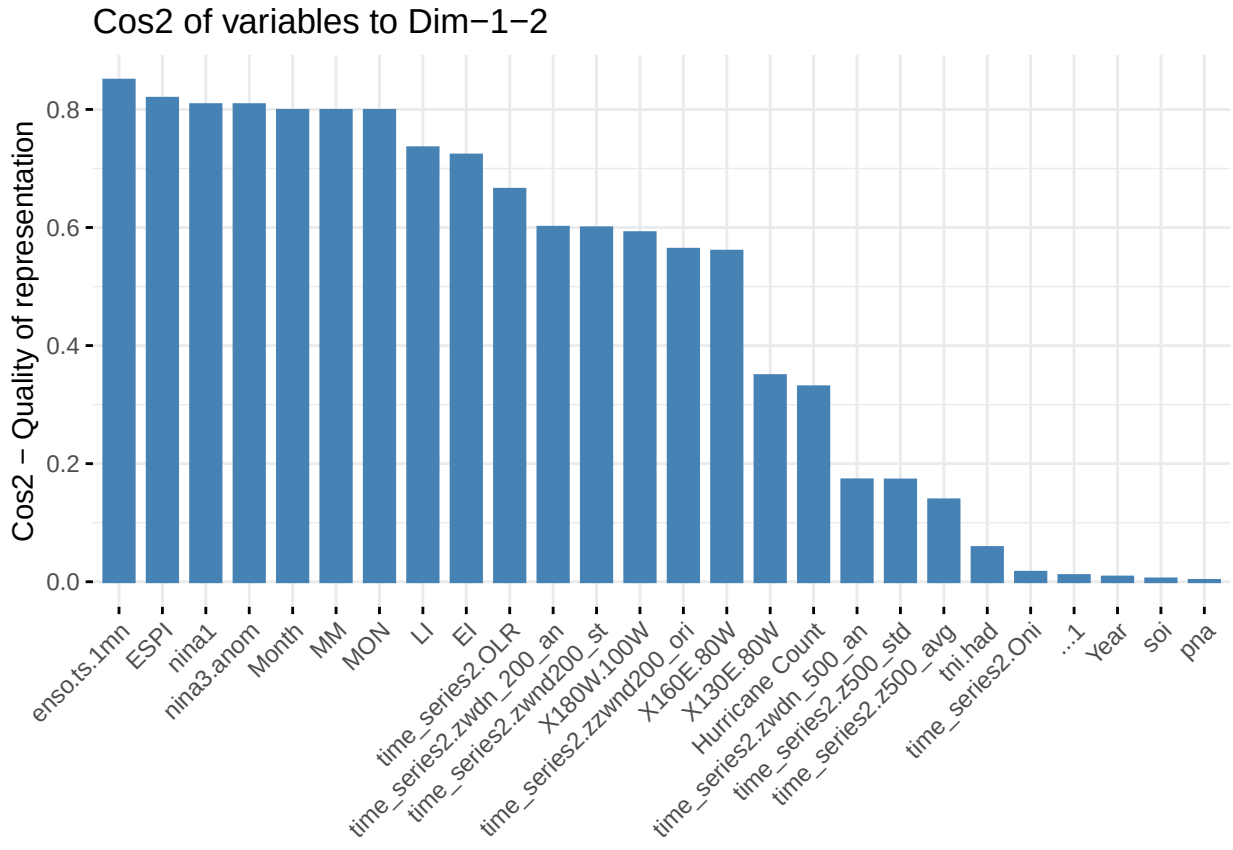
Feedback Mechanisms: The interplay between these variables can create feedback loops that amplify or mitigate weather patterns. For instance, a strong El Niño (high ONI, low SOI) can lead to significant changes in z500 patterns, influencing PNA and other teleconnection indices.

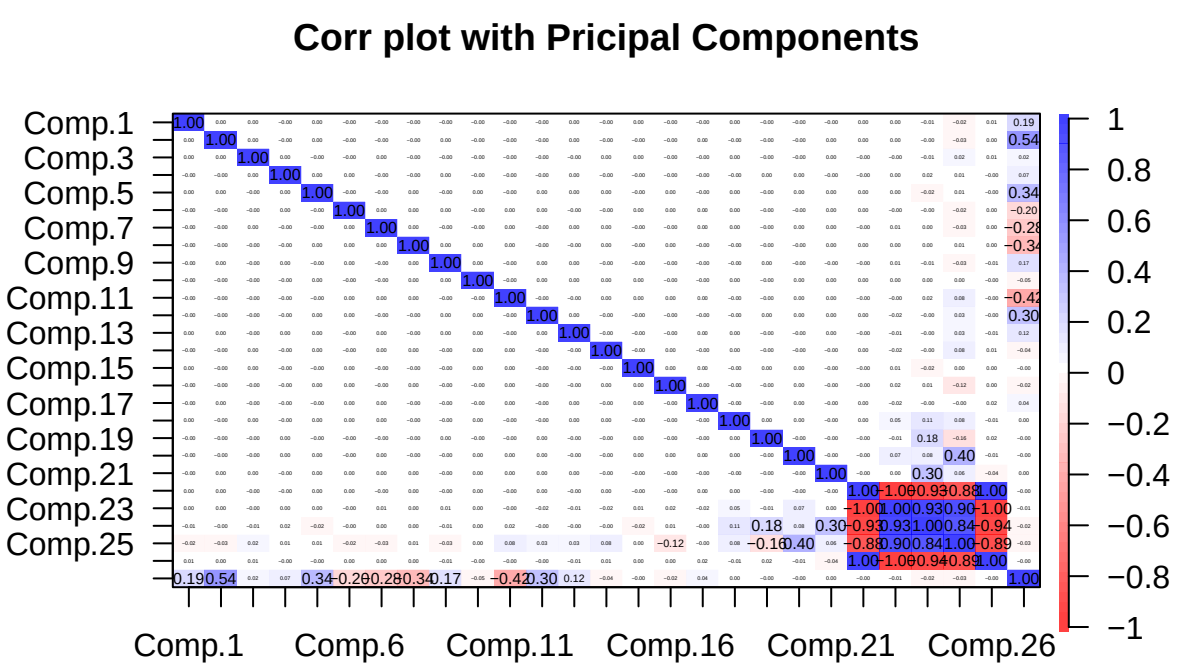
Conclusion

The clustering experiment successfully identified distinct groups within the dataset, revealing underlying patterns and relationships between different meteorological variables. These insights could inform further studies on the interactions between atmospheric conditions and hurricane occurrence, potentially aiding in improved predictive models for weather and climate phenomena.

PCA


Variables − PCA
Month
MM
MON
Hurricane Count
time_series2.zwdn_500_an
time_series2.Oni
enso.ts.1mn
time_series2.z500_std
EI
time_series2.OLR
time_series2.z500_avg
soi
...1
ESPI
X180W.100W
pna
year
LI
X160E.80W
X130E.80W
tni.had
time_series2.zwdn_200_an
time_series2.zwnd200_st
time_series2.zzwnd200_ori
nina3.anom
nina1
cos2
0.8
0.6
0.4
0.2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
Dim2 (19.1%)
Dim1 (27%)


Cos2 of variables to Dim−1−2
Cos2 − Quality of representation
0.8
0.6
0.4
0.2
0.0
enso.ts.1mn
ESPI
nina1
nina3.anom
Month
MM
MON
LI
EI
time_series2.OLR
time_series2.zwdn_200_an
time_series2.zwnd200_st
X180W.100W
time_series2.zzwnd200_ori
X160E.80W
X130E.80W
Hurricane Count
time_series2.zwdn_500_an
time_series2.z500_std
time_series2.z500_avg
tni.had
time_series2.Oni
...1
Year
soi
pna



PCA INTERPRETATION

During the model development phase, we considered using principal components as potential predictors of hurricane counts. Although this approach did not yield satisfactory results, the principal component analysis (PCA) provided valuable insights that confirmed previous findings.

The cos2 value for each variable indicates the proportion of the variable's variance captured by the principal components. The PCA analysis reaffirmed some earlier observations, such as the association of hurricane occurrence primarily with temporal variables (MM, Month, MON). There appears to be a negative relationship with the variables nina.anom and nina1, which aligns with seasonal decomposition findings of nina3.anom. Additionally, the ENSO cycles are associated with the ESPI and EI indices.

The color coding in the PCA biplot reveals that the variance explained by the first two principal components is around 50% for hurricane counts and exceeds 80% for ENSO-related variables, including ESPI, EI, nina1, and nina3.anom. These relationships are significant and underscore the importance of these variables in understanding hurricane occurrences.

The following plots illustrate the correlation between hurricane counts and the principal components, as well as the explained variance by these components.