CS575 - Applied Time Series Analysis

Mini Project

Komal Pandya (2021CS17)

Department of Computer Science and Engineering
Indian Institute of Technology Patna

April 2021

Table of Contents

DATASET USED FOR THE PROJECT -
https://www.statsmodels.org/devel/datasets/ generated/sunspots.html

Objectives

Stationarity tests for the dataset

- Visual Tests(from graphs)
- Check for similarity in Mean and Variance
- Using ADF and KPSS tests

Modelling time series

- Using ARIMA
- Using Seasonal ARIMA
- Using deep learning approaches(DNN, RNN, LSTM)

Objectives

Clustering

- Using KMeans
- Using Agglomerative Clustering

Outlier Detection

- Using Interquartile range method
- Using Standard deviation method
- Using Z-Score method

Methodology

Stationarity Tests

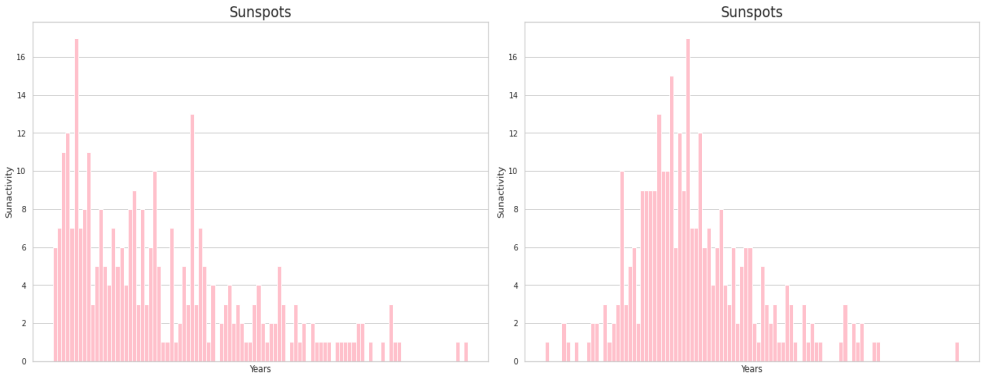
- Visual Tests(from graphs) - `If the data curve is showing any visible trend or cycle, it's non stationary, otherwise it's stationary.`
- Check for similarity in Mean and Variance - `Comparing mean and variance at different time values.If they varies heavily, data is non stationary.`
- Using ADF and KPSS tests - `For both tests, possibility for having unit roots is checked and statistics and critical values are compared.`

Results for Stationarity Tests

- The data was found to be non-stationary using above mentioned tests.
- It was then made stationary using differncing.
- After differncing each test was performed again and the new differenced series was found to be stationary.
- Slides below contain the result comparison for both the series i.e original vs differnced.

Test 1:Visualization

Figure: Original vs Differenced series



As shown in the figure, original data is non stationary while the differned data shows a bell shaped curve(gaussian distribution), hence stationary.

Test 2:Mean

The data series is partitioned into 3 equal parts, for calculating means of different partitions to compare.

Partition 1 of Original —- 45.269903

Partition 3 of Original —- 62.150485

Partition 1 of Differenced —- 0.392157

Partition 3 of Differenced —- -0.262745

There's a huge difference in means of different parts of the original series. However, after differencing, means are almost similar and closer to zero, indicating stationarity.

Test 3: ADF and KPSS Tests

Tests	Original	Differenced
ADF	Non-Stationary	Stationary
KPSS	Non-Stationary	Stationary

Table: Observations of the tests

The original tests showed the results as non stationary. After differencing, the series became stationary.

Methodology

Modeling Time Series

- ARIMA - It takes into account the previous values of the time series .From acf and pacf curves, the parameters(p,d,q) for AR, MA, ARIMA could be identified.Parameters with least MSE were identified using GridSearchCV.
- Seasonal ARIMA - For capturing seasonal patterns, SARIMA was used with the parameters having least AIC value.
- Deep Learning Approaches - Deep Neural network, Recurrent neural network and long short term memory network were used for modeling the data using deep learning.

Results for Time Series Modeling

- After finding the parameters(p,d,q) for AR(p), MA(q) and ARIMA(p,d,q) using GridSearchCV, the values obtained with least MSE are as follows - p - 9, d - 2, q - 3.
- For seasonal ARIMA, the parameters (P,D,Q,s) were found to be (2,2,2,2) respectively.(least AIC)
- The data was split into train and test sets with 2:1 ratio.

Results for Time Series Modeling

The models and their respective MSE are as follows -

- AR(9) - 293.047
- MA(3) - 490.999
- ARMA(9,3) - 324.704
- ARIMA(9,2,3) - 300.941
- SARIMA(2,2,2,2) - 248.863
- DNN - 226.502
- RNN - 0.0144
- LSTM - 0.0121

From the above table, certainly LSTM outperforms other models that were implemented.

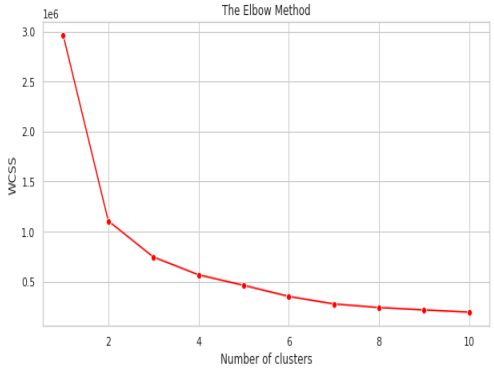
Methodology

Clustering

- KMeans- Clusters are identified using the elbow method and variance is minimized among data points to form clusters.
- Agglomerative Clustering - This is a bottom-up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Results for Clustering

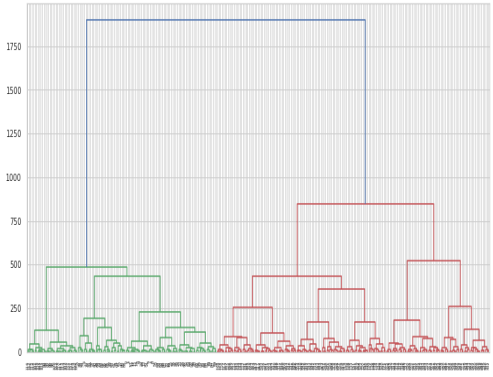
Figure: Elbow Method for KMeans



For KMeans, 3 clusters were selected.

Results for Clustering

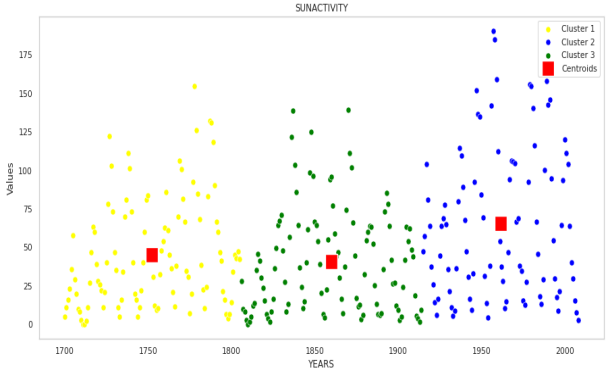
Figure: Dendrogram for Agglomerative Clustering



For Agglomerative clustering, 3 clusters were selected.

Results for Clustering

Figure: Viewing Clusters



Methodology

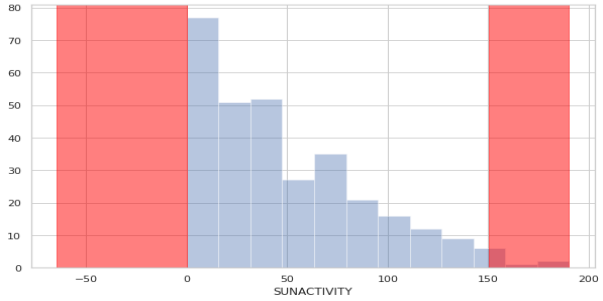
Outlier Detection

- InterQuartile Range method - Interquartile range(IQR) is the difference between the third quartile and first quartile.(IQR = Q3-Q1). Outliers then can be defined as the observations that are below (Q1 1.5x IQR) or above (Q3 + 1.5x IQR).
- Std Deviation method- After measuring the upper bound and lower bound values of standard deviation, points not falling in range, can be considered as outliers.
- Z-Score method - .if the Z-score value is greater than or less than 3 or -3 respectively(usually), that data point will be identified as outliers.

Results for Outlier Detection

The IQR is 53.8 The lower bound value is -64.69999999999999 The upper bound value is 150.5

Figure: IQR Method

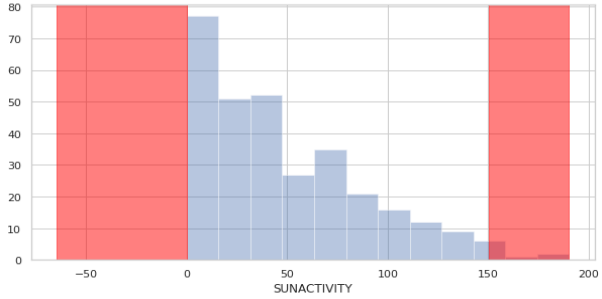


Outliers : 8

Results for Outlier Detection

The lower bound value is -71.60568131066171 The upper bound value is 171.10988843040286

Figure: Std Deviation Method

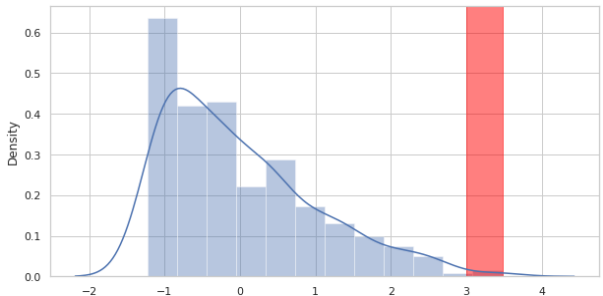


Outliers : 2

Results for Outlier Detection

Threshold for Z-Score : 3

Figure: ZScore Method



Outliers : 2

Conclusion

- The SUNSPOTS dataset from statsmodels was initially found to be non stationary using various tests.
- For models like AR and MA, the data needs to be stationary. Hence it was made stationary using differencing.
- Various statistical and deep learning models were then applied to the data and mean squared error for each of them was reported.
- The scatter plot of the original time series showed some possibility for clusters. Hence, clustering implementation was done using algorithms like KMEans and Agglomerative clustering.
- As the data was having a high variance, there could be some outliers too. Approaches like IQR Method, Z-score method, etc were utilized to find and report the number of outliers.