Forecasting Climate-related Search Terms in Google Web Search

Repository: https://github.com/cynthiazyii/ChenjiaCynthiaWill_TSA_project

Zhenghao Lin, Chenjia Liu, Cynthia Zhou

2024-04-23

Introduction, Motivations, Relevance, Objectives

Climate change has become one of the most pressing issues facing humanity in the 21st century. With the escalating impacts of global warming, extreme weather events, and environmental degradation becoming more visible, public awareness and concern have surged. In this digital age, the internet has become a primary avenue for education and engagement on environmental issues. As the world's predominant search engine, Google offers invaluable insights into global interest in climate topics, making it a valuable resource for public concern evaluation.

This report utilizes Google Trends data to perform a comprehensive time series analysis of the search volume for "Climate" to identify underlying patterns, trends, and anomalies in public interest over time. Using advanced statistical tools and methods, this analysis will: 1. Forecast Future Trends: Apply various models to the original dataset to predict future changes in public interest concerning climate issues. 2. Assess the Impact of Outliers: Compare forecasts from the original data with those from a modified dataset that excludes outliers, evaluating how these anomalies affect overall predictions. 3. Analyze Spatial Variations: Conduct a spatial analysis of relative search volumes to map regional differences in interest regarding climate-related topics across the globe.

Through these analyses, the report seeks to provide a comprehensive overview of the public's engagement with climate issues, offering insights that could inform policymakers, educators, and environmental organizations in their efforts to foster greater awareness and action.

Data Information

Our data set is sourced from Google Trends. As the original data is interest index which only represents the popularity of total search term in certain time periods or regions, we use a Chrome extension called Glimpse-Google Trend Supercharged to convert interest index into absolute values. We captured the number of searches for climate-related search terms globally from 2004 to the present.

Table 1: Data Information

Data	Information
Source:	Google Trend
Search Term:	Climate
Time Range:	2004 to present
Spatial Range:	Global

Methodology

The models that are being used by this analysis include TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) model, ARIMA + FOURIER terms, STL+ETS (Seasonal and Trend decomposition using Loess + Error, Trend, Seasonal) model, Neural Network, and SSES (State Space Exponential Smoothing) model. A short description of the models are given below:

1 TBATS

TBATS was designed to forecast time series with multiple seasonal periods. For example, daily data may have a weekly pattern as well as an annual pattern. Or hourly data can have three seasonal periods: a daily pattern, a weekly pattern, and an annual pattern. In TBATS, a Box-Cox transformation is applied to the original time series, and then this is modelled as a linear combination of an exponentially smoothed trend, a seasonal component and an ARMA component. The seasonal components are modelled by trigonometric functions via Fourier series. TBATS conducts some hyper-parameter tuning (e.g. which of these components to keep and which to discard) using AIC.

2 ARIMA + FOURIER terms

The ARIMA model with Fourier terms for seasonal adjustment is a sophisticated approach for forecasting time series data, particularly when the data exhibits complex seasonality that cannot be adequately modeled by simple seasonal ARIMA (SARIMA) components alone.

3 STL+ETS

STL is a versatile and robust method for decomposing time series. STL is an acronym for "Seasonal and Trend decomposition using Loess", while loess is a method for estimating nonlinear relationships. The ETS models are a family of time series models with an underlying state space model consisting of a level component, a trend component (T), a seasonal component (S), and an error term (E).

4 Neural Network

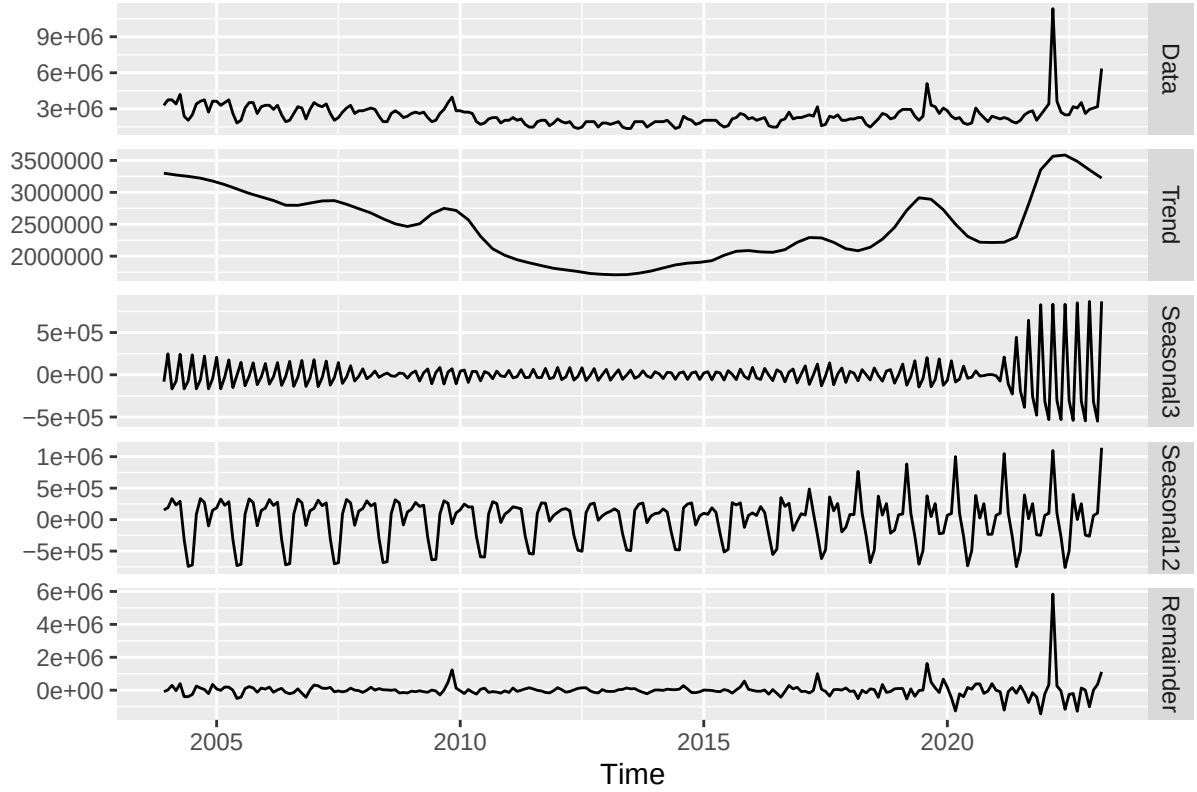
Neural networks or simulated neural networks are a subset of machine learning which is inspired by the human brain. They mimic how biological neurons communicate with one another to come up with a decision. A neural network consists of an input layer, a hidden layer, and an output layer. The first layer receives raw input, it is processed by multiple hidden layers, and the last layer produces the result.

5 SSES

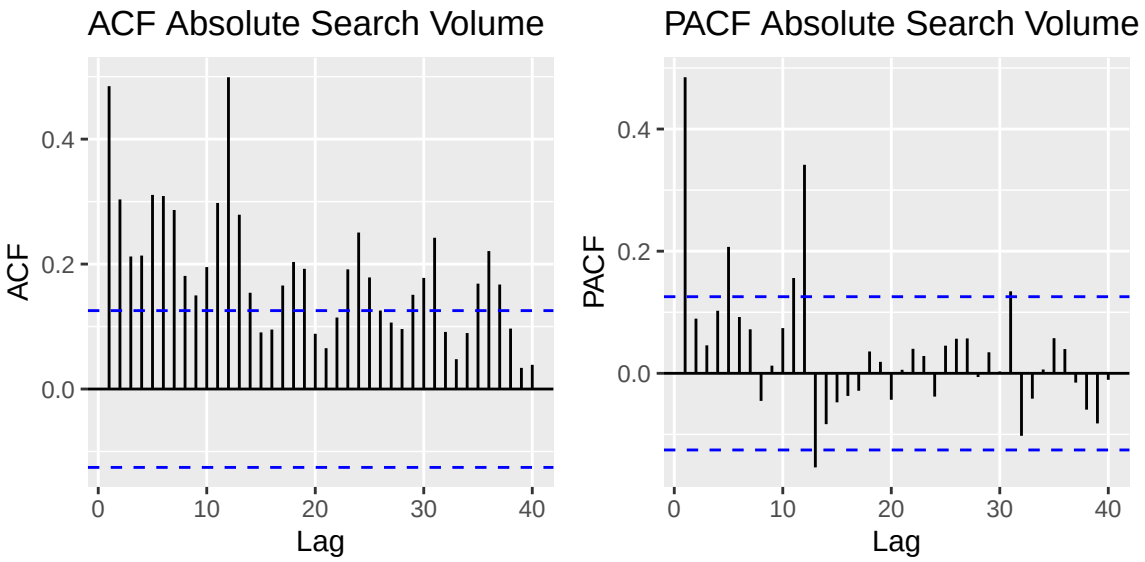
The exponential smoothing state space model is a statistical model that generates prediction intervals and point forecasts. It is a stochastic data generating process that can produce an entire forecast distribution. The exponential smoothing methods generate point forecasts, and the statistical models generate the same point forecasts, but can also generate prediction intervals.

Analysis

Forcast of original data



First, the data set was read and the Time is converted to "date". The variable of interest (absolute search volume) was renamed and converted to a time series object. A training set and test set was then obtained by filtering data. The last 12 observation was selected as holdouts. The time series data set was plotted to check for trends and seasonality. The seasonal period is set using 3 (quarterly) and 12 (manually).



The ACF plot displays the correlation between the time series data and its lagged values over different time lags. The plot shows several spikes that are outside the blue dashed confidence interval lines, especially at the initial lags. This suggests that there is a significant autocorrelation at those lags. The PACF plot, on the other hand, shows the partial correlation between the time series and its lagged values, controlling for the values at shorter lags. The significant spike at the first lag indicates a strong correlation. Since the spike at lag 1 is followed by a cutoff, it might suggest an AR(1) component.The slow decay in the ACF plot could indicate that the data is non-stationary or has seasonal patterns.

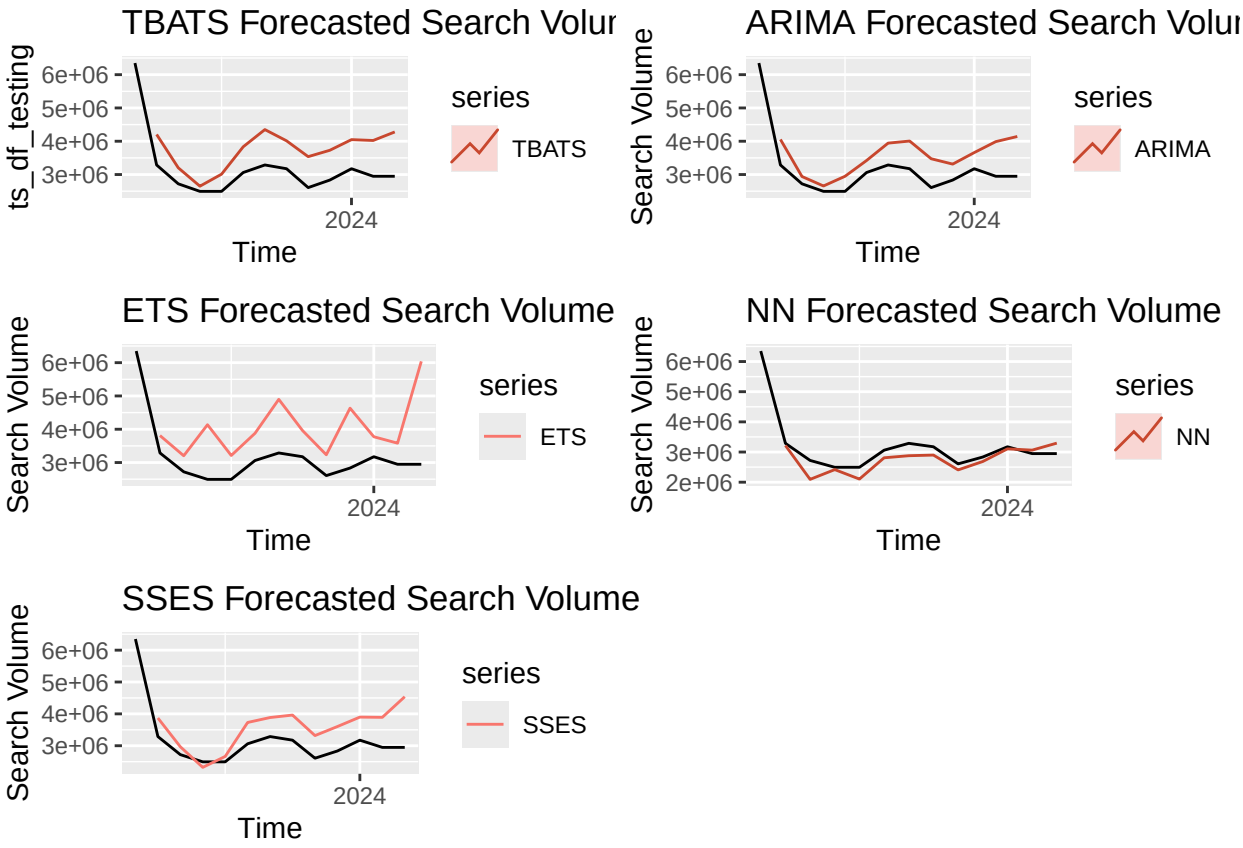
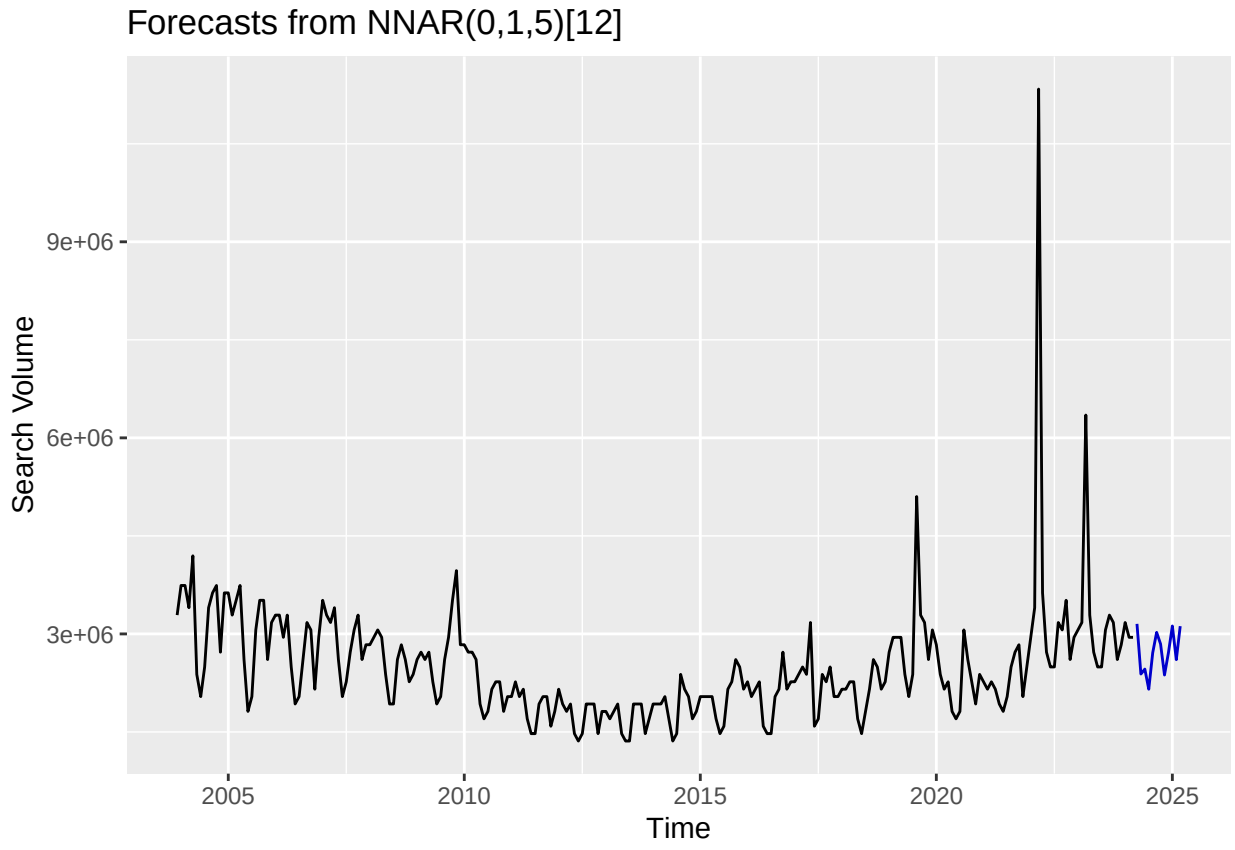


Table 2: Forecast Accuracy for Daily Active Power

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
STL+ETS	-1110624.8	1339766.2	1110624.8	-38.44795	38.44795	-0.23679	4.11644
ARIMA+Fourier	-626317.7	697971.3	626317.7	-21.27047	21.27047	0.41068	1.94301
TBATS	-820800.9	873225.0	820800.9	-27.74279	27.74279	0.52540	2.50556
NN	171718.1	298203.2	248873.6	6.06823	8.68571	0.18905	0.88336
SSES	-634705.2	759560.7	663662.4	-21.21714	22.37811	0.49915	2.22684

The data was trained from December 2003 to March 2023 and tested from April 2023 to March 2024. Predictions were made for the next 12 months. Based on the residual scores, the neural network (NN) model outperformed others, achieving the lowest RMSE of 353,492.8. In contrast, the STL+ETS model registered the highest RMSE at 1,339,766.2. Consequently, the NN model was selected for predicting the absolute search volume. It's important to note that the residual scores were quite high. This high error rate is partly because our base numbers are substantial, typically in the millions. The NN's residual score, although in the hundreds of thousands, represents about 10 percent of the data, which is still considered a significant error. Moving forward, the next step involves removing outliers, which are causing huge spikes, and rerunning the candidate model.



Outlier Removal and Model Optimization

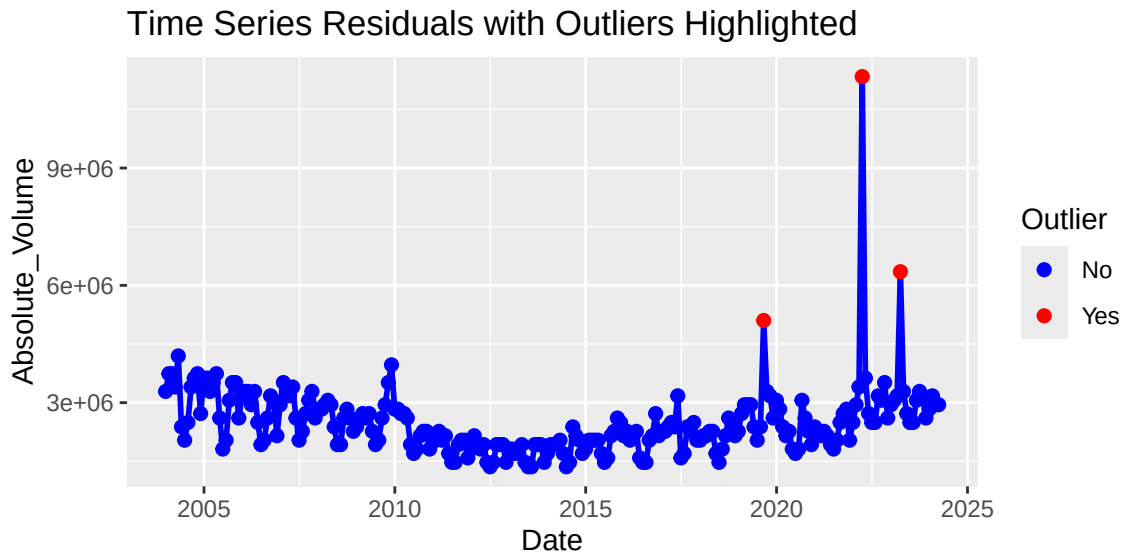
This section of the report builds upon the initial time series analysis by addressing outliers in the dataset and re-evaluating the forecasting models to identify the best performer in a refined dataset. The focus is to enhance model accuracy by minimizing the distortion effects caused by outliers.

Outlier Detection and Interpolation

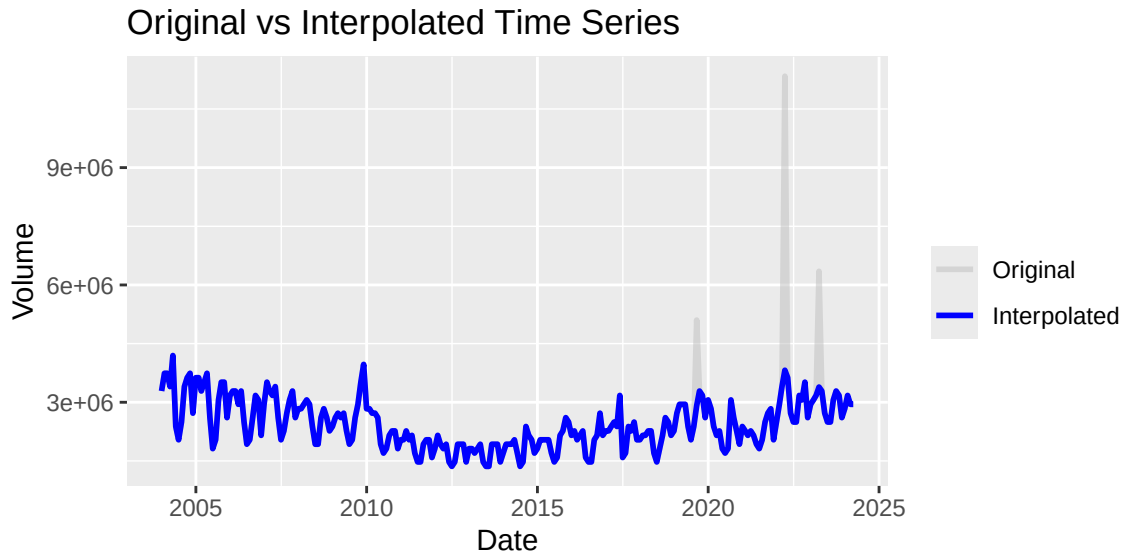
The process began by identifying outliers using the Interquartile Range (IQR) method, a robust technique that determines extreme values based on the distribution's quartiles. Outliers were defined as observations that fall below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR. These outliers were then replaced using spline interpolation, which provides a smooth estimate that is less likely to be influenced by extreme values, hence maintaining the integrity of the time series.

Methodology

The outliers in the Absolute_Volume data were first visually identified and marked in the dataset. Below is the graph of the time series plot with outliers identified through IQR method.

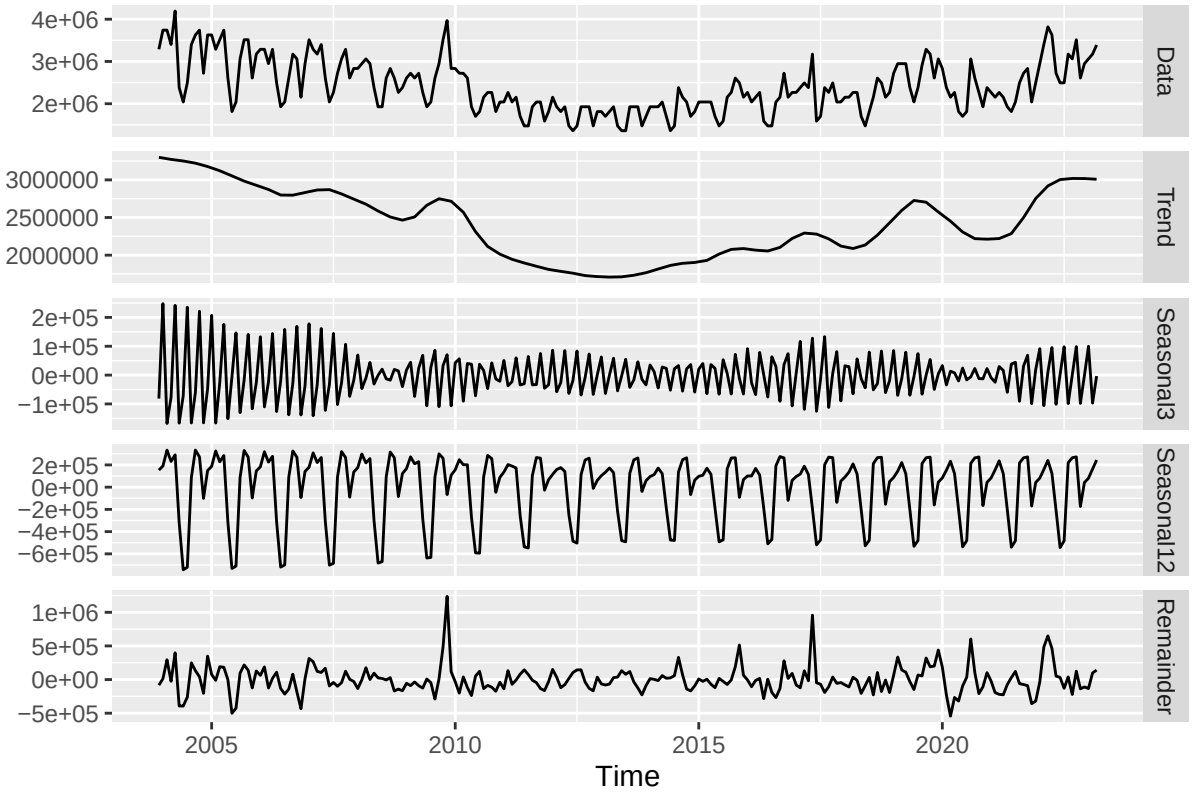


The na.spline function from the zoo package was then used to interpolate and fill in these outliers based on neighboring values, which helps in preserving the trend and seasonal patterns of the series. The following figure from the analysis illustrates the original data against the interpolated data, showing how the outliers have been integrated.

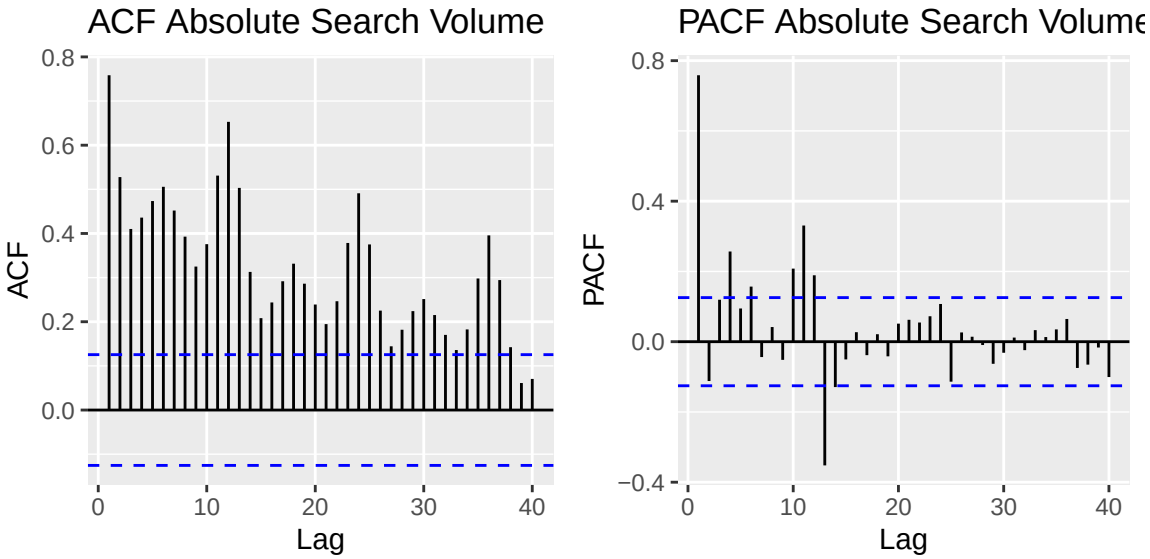


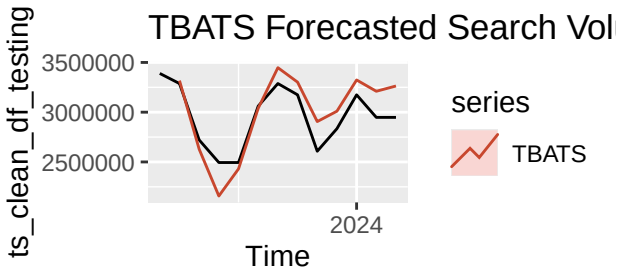
Model Re-evaluation

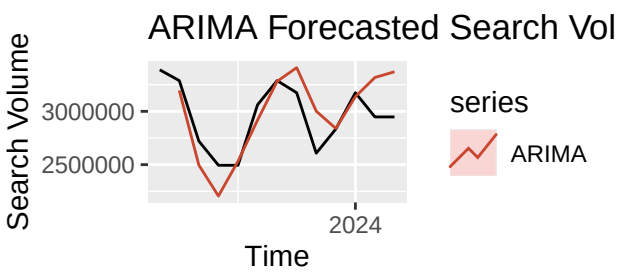
With the outliers addressed, the dataset was split into training and testing sets. The analysis method and forecasting models used in the previous section were re-applied to this cleaned data.

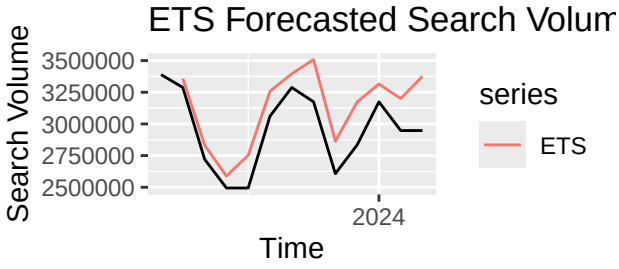


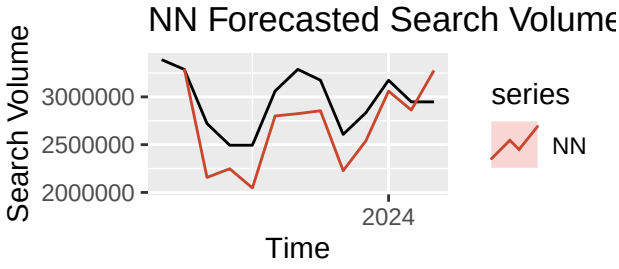
When looking at the ACF and PACF plots of the cleaned data frame, the ACF plot reveals strong positive autocorrelations at the initial lags and then gradually decline as the lags increase, indicating potential non-stationarity within the time series data. Furthermore, the ACF plot exhibits a potential seasonal pattern through its recurring spikes at specific lags. For the PACF plot, since it has a significant spike at the first lag and then cut off after it, the plot suggests at least an AR(1) model. This result align with the observation with the ACF and PACF plots of the original dataset.

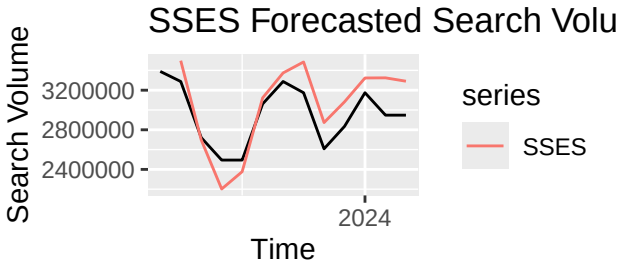












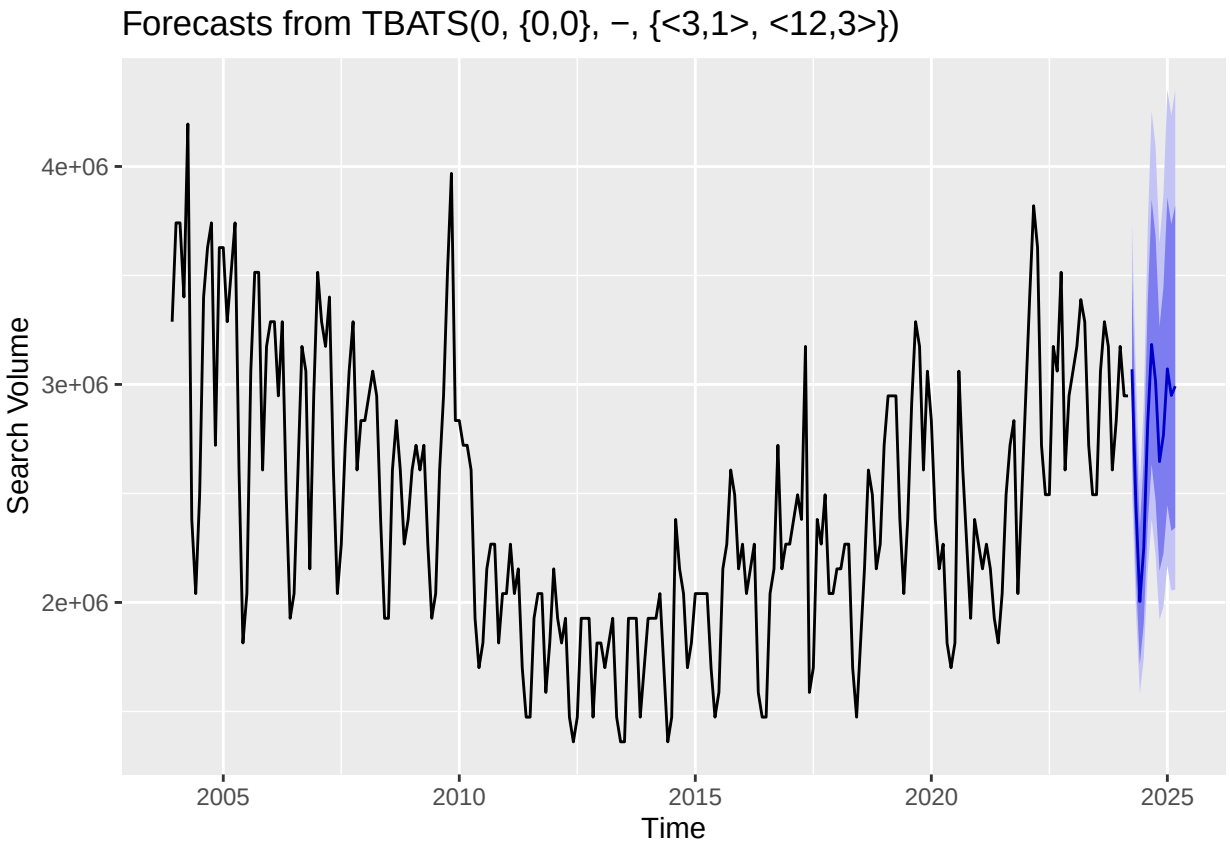
Results

Below is the model score table of cleaned dataset with outliers removed. Compared to the original data frame, the cleaning process significantly improved model performance. The TBATS model emerged as the best model with the lowest RMSE, indicating the highest forecasting accuracy among the evaluated models on the cleaned dataset. The improvement underscores the impact of outliers removal on enhancing model reliability and accuracy.

Table 3: Forecast Accuracy for Daily Active Power with Cleaned Dataset

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
STL+ETS	-215676.44	241749.4	215676.4	-7.48832	7.48832	0.11827	0.73889
ARIMA+Fourier	-56154.16	240161.5	188096.7	-1.87295	6.64871	0.43404	0.69486
TBATS	-83613.70	199893.7	169354.0	-2.62990	5.98907	0.61850	0.60026
NN	236617.64	332128.9	292698.4	8.44210	10.34003	0.10969	1.00338
SSES	-134830.08	235477.6	206972.2	-4.29787	7.17737	0.60346	0.67696

Below is the forecast of the search volume of term "Climate" on google for next 12 months with TBATS forecasting model using cleaned dataset with outliers removded.



Spatial Characteristics

In the global distribution map of climate-related topic web search, African countries enjoy the most popularity as a fraction of total searches in their region. Related search terms account for a moderate proportion in North America, Australia and South Asia, while it was not that popular in South America, Europe and other Asia regions. On a country level, Ethiopia and Zimbabwe had the largest proportion of climate search terms in their total searches.

Population can be a possible reason for the difference. In African countries, where few people have access to the Internet, the proportion of climate search terms may be magnified. In addition, international organizations and non-governmental organizations such as the United Nations Environment Program are headquartered in Africa, leading to related behaviors. The extent to which different countries are affected by climate change is also important. Countries like the Philippines, which are at the forefront of climate

change, rank very high both in the interest index and in absolute values. People who live there are naturally more concerned about the climate.

Limiations

Our dataset inherently possesses several limitations that could skew our findings' accuracy. Initially, we utilized both the absolute and relative search volumes of the term "Climate". However, the ambiguity of this term means it is often searched in contexts unrelated to climate change, such as weather conditions or trends in various fields. This broad usage can lead to an overestimation of both absolute and relative search volumes. Additionally, while Google remains the most popular search engine globally, it is not the only one with significant market presence. For example, Baidu holds a substantial share of the market, particularly in China—the country with the second largest population. By not incorporating data from Baidu, particularly given its dominance in Chinese searches, our research may overlook a critical segment of global search trends, potentially resulting in biased outcomes.

Within the data obtained from Google, several issues still persist. Our analysis from first section highlighted the presence of outliers—abnormal spikes in the data which could distort our forecasts. To mitigate potential biases from these anomalies, we removed these outliers to try to achieve more reliable outcomes. Although this method generally improved model performance, it raised some new concerns, especially the risk of overfitting. By employing sophisticated techniques to detect and eliminate outliers, we unavoidably increased the likelihood that our models might become overly fitted to our current dataset. Such overfitting could undermine the accuracy of future predictions by making our models less capable of generalizing from new data. Moreover, our current dataset may not be extensive enough to conclusively differentiate between genuine seasonal effects and anomalies. Some of the spikes might not represent underlying seasonal trends but could instead be one-time events or products of irregular fluctuations. These non-recurring spikes could potentially skew the forecasts.

Another notable issue is the underrepresentation of certain countries in our study. The spatial analysis section excludes countries with low search volumes, many of which are also experiencing significant impacts from climate change. This omission could further contribute to an incomplete global perspective in our research.

Next Steps

In the future, if resources and time allowed, we will take following steps to further our investigation.

Firstly, we need to expand our data sources. To address the issue of over-reliance on Google and the consequent geographic and demographic biases, we should integrate data from other significant search engines, such as Baidu. By broadening our dataset to include multiple search engines, we can achieve a more balanced and globally representative analysis. This expansion will help us better understand regional differences in search behaviors and refine our interpretations of global interest in climate-related topics.

Secondly, improving data filtering techniques to better differentiate between searches related to climate change and those triggered by unrelated factors is essential. We hope that Google can develop more sophisticated algorithms that can contextually analyze search terms and distinguish between searches for climate change information and other uses of the word "climate." This might involve natural language processing (NLP) techniques or machine learning models trained.

Additionally, addressing the problem of data outliers and overfitting requires a revision of our current methodologies. We should implement more robust statistical methods to handle outliers without removing them entirely, thus preserving valuable data that might be relevant for understanding extreme but significant phenomena.

Lastly, we must strive to include underrepresented regions in our analysis. It's important to develop strategies to estimate search volumes in countries with lower internet penetration or those typically excluded from digital data analyses. This might involve proxy measures or collaborations with local data providers to capture a fuller picture of the global discourse around climate change.

Conclusion

This report analyzes the search volume trends of the term "Climate" using Google Trends data to gain insights into public interest over time. Our methodology involved a comprehensive time series analysis on both original dataset and the adjusted data frame with outliers removed, and an assessment comparison of search behaviors globally.

The results of our analysis indicate a steady increase in the future search volume for "Climate." This upward trend suggests a growing public awareness and interest in climate-related issues. By successfully removing outliers and refining the data, the predictive models, particularly the TBATS model, provided robust forecasts that reinforce our understanding of this increasing engagement. This increase in search volume can be interpreted as a positive sign that climate change is becoming a more significant concern for people worldwide, potentially signaling greater support for policies and initiatives aimed at combating climate change. As we move forward, it will be important to continue monitoring these trends and expand our data collection efforts to include a wider array of search engines and regions to further validate and refine our forecasts. By doing so, we can ensure that our models remain adaptable and accurate, providing valuable insights to policymakers, educators, and environmental organizations aiming to foster an informed and engaged public.