Time Series Analysis of Change of Unemployment Rate in USA

Sidar YÜK
Middle East Technical University
Ankara, Turkey
e229105@metu.edu.tr

***Abstract*—This paper delves into predicting the Unemployment Rate- Of black or African Americans in the USA using various forecasting models, including ARIMA, ETS, TBATS, NNETAR, and PROPHET. The study predominantly occurs within the R-Studio environment, involving thorough data preprocessing steps such as anomaly detection and addressing stationary/nonstationary conditions. The optimal models are determined and fitted to produce forecasts following data cleaning and statistical procedures. The performance of these models is then compared on both training and test datasets. The research encompasses a comprehensive approach, including outlier analysis, data cleaning, and unit root checking to enhance the reliability of the results. Ultimately, the study aims to evaluate and compare the effectiveness of different forecasting models in predicting Unemployment Rate- Black and African American changes..**

Keywords—Forecasting, Unemployment, USA, ETS

I. INTRODUCTION

Unemployment is when an individual actively seeks employment but cannot secure a job. It is a crucial gauge of economic well-being, with the unemployment rate being the prevalent metric. Computed by dividing the number of unemployed individuals by the total labor force, this rate reflects the proportion of people facing joblessness. The significance of unemployment lies in its role as an economic indicator, shedding light on the capacity of workers to secure productive employment and contribute to the economy's overall output. More unemployed individuals correlate with reduced overall economic output(Hayes& Anderson,2023).

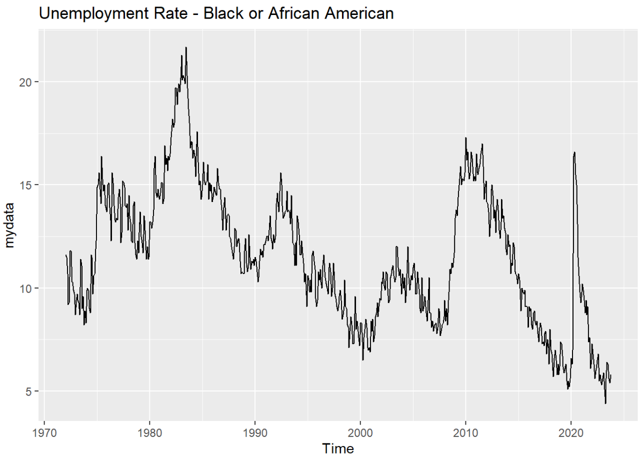
The examination will delve into the repercussions of economic fluctuations on unemployment rates, encompassing both periods of recession and expansion. This entails identifying and examining cyclic trends to gain deeper insights into how the job market reacts to overarching economic shifts. Additionally, the dataset under scrutiny might encompass irregularities or aberrations that offer valuable insights into extraordinary occurrences influencing unemployment rates. These outliers could be ascribed to economic shocks, policy alterations, the COVID-19 pandemic, or other external elements, providing valuable information about distinct incidents within the labor market.

This time series analysis aims to uncover critical insights into the monthly unemployment rate of Black or African Americans spanning from 1972 to 2023

In this study, aside from performing exploratory data analysis, a range of forecasting models, namely ARIMA, ETS, TBATS, Neural Network, and Prophet, were utilized to anticipate variations in the unemployment rate. The effectiveness of these models was assessed by scrutinizing metrics such as root mean squared error (RMSE) and mean absolute error (MAE). R-Studio was pivotal throughout the analysis, serving as the primary tool for data series exploration and unemployment rate forecasting. The subsequent stage involved a comparative evaluation of the model's predictive capabilities, relying on RMSE and MAE metrics.

II. DATA DESCRIPTION AND PREPROCESSING

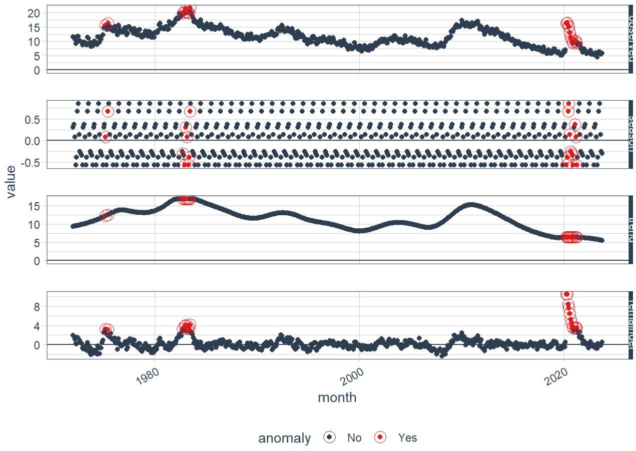
The data set is taken from https://fred.stlouisfed.org/, the Federal Reserve Bank of St. Louis web page, one of the 12 regional reserve banks in the United States. The data set contains 622 observations recorded annually from 1972 to 2023. Data includes information about the unemployment rate in the USA. This data is only based on one demographic variable, which is black and African American people. The data were obtained from a monthly survey of about 60,000 households.



Graph 1 Time Series Plot of Data Set

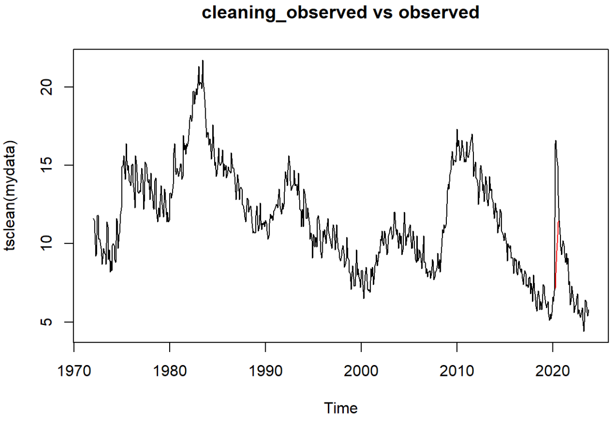
The time series shows there are both increasing and decreasing trends. There is an increasing trend from the 1970s to the end of 1980. A decreasing trend has been noticeable since this date. When we look at 2020, the graph suddenly peaked, possibly due to outliers. The variance problem can be seen clearly. The plot seems nonstationary.

After checking, the time series plot procedure begins with STL decomposition, which allows us to identify and analyze anomalies in the data, ultimately confirming the presence of abnormalities in the time series.



Graph 2Anomaly Detection Plot

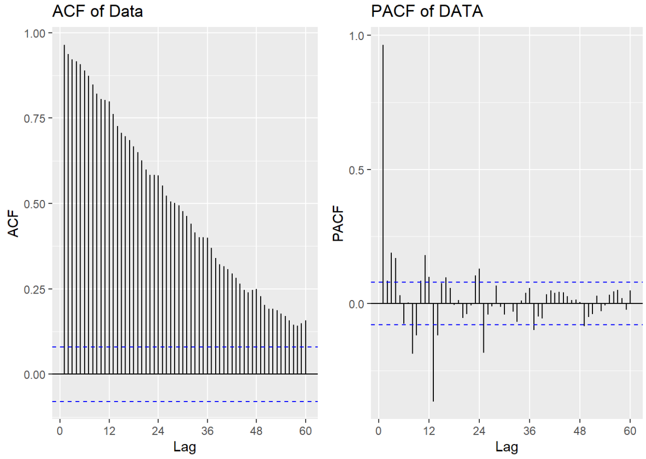
Anomaly points appear in 3 different periods around 1975, 1983, and 2021, respectively. Anomaly points mainly to the period between 2020 and 2021 when the unemployment rate increased incredibly. The reason for this is the COVID-19 pandemic.



Graph 3 time series plot after cleaning dataset and original dataset

The tsclean() function in R-Studio eliminates missing values and outliers by employing linear interpolation. Following the application of tsclean, the subsequent time series plot demonstrates the successful eradication of anomalies through the substitution with interpolated values.

The data set was divided into train and test sets using the window() function, a generic function that extracts the subset of the object x observed between the times, and a generic function that extracts the subset of the object x observed between the times start and end. The last 12 observations are kept as a test set because data includes annual observations.



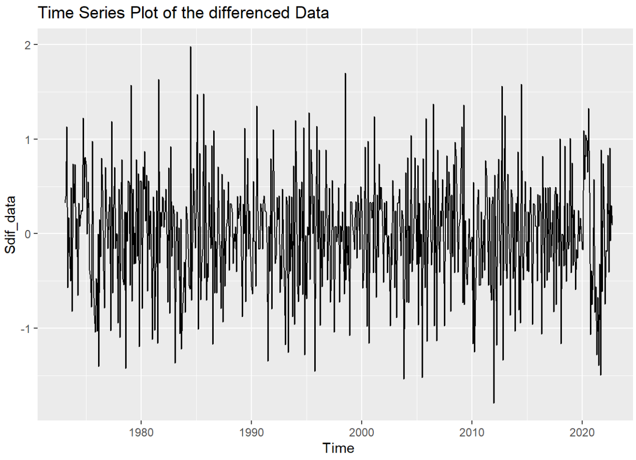
The ACF and PACF plots derived from the training set indicate a gradual decrease, suggesting a slow decay pattern. Such behavior is indicative of nonstationary processes. In light of this observation, caution is advised when interpreting the PACF plot, and making any definitive comments on it is discouraged by authorized individuals.

In model-based statistical analysis, certain assumptions must be met by the analyzed data. When dealing with time series, achieving covariance-stationarity is often a prerequisite for initiating the modeling process. Consequently, seeking a variance-stabilizing transformation that brings the data closer to satisfying this assumption is rational. Many statistical techniques assume a normal distribution of data and constant variance over time, referred to as covariance-stationarity in time series. However, real-life data often deviates from these crucial assumptions. By applying various transformation techniques, it is possible to align data that violates these assumptions with the required conditions..

This study made an effort to address non-constant variance by initially determining the lambda value. A Box-Cox transformation was employed as the resulting value was not equal to 1. The BoxCox function within the R-Studio was utilized to modify the data appropriately to execute the transformation.

Upon attaining stationarity in variance, further tests were conducted to ensure the mean and overall stationarity of the annual dataset used in this study. Both the KPSS and ADF tests were employed for analyzing non-stationarity. In addition to these, stationarity conditions were examined using the PP, HEGY, and CANOVA-HENSEN tests. The combination of these tests was utilized to assess and validate the stationary nature of the dataset comprehensively. The initial stage of the KPSS test reveals non-stationarity in the data with a significance level of $p<0.05$, indicating the presence of a unit root. Subsequently, the second stage of the test reaffirms the data's nonstationary nature, emphasizing that the trend is stochastic, also with a significance level of $p<0.05$. In summary, the first level highlights the existence of a unit root, while the second level underscores the stochastic nature of the trend. In addition, the ADF test indicates a nonstationary system with a stochastic trend, the pp test indicates the result is not stationary ($p>0.05$), and the HEGY test indicates a regular unit root problem. There may be a seasonal unit root problem. On the contrary, the Canova-Hensen test shows that the process is purely deterministic and stationary. Lastly, the ndiffs() and nsdiffs() functions in R studio within the forecast package give the number of regular and seasonal differences to be taken to achieve the stationary; both function's numbers are 1

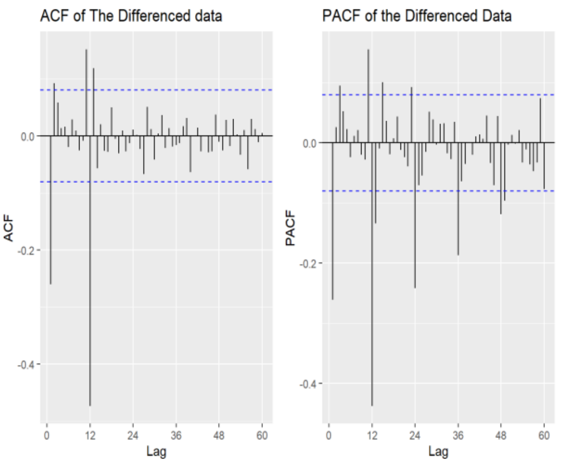
Except for the Canova-Hensen test, all other tests show that the process is a nonstationary and stochastic trend. Additionally, nonstationary and stochastic processes can be easily seen in the acf and pacf plots. In addition, the ndiffis and nsdiffs functions clarify that differences should be taken. The regular unit problem was overcome after taking the first regular order difference ($p>0.05$ for the KPSS test and $p<0.05$ ADF). Then, after seasonal differences were removed, there was no unit problem. ($p<0.05$ for the HEGY test)



Graph 4 Time Series Plot of differenced data set

III. MODEL SUGGESTION

Once a stationary time series has been acquired, the next step involves utilizing Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to guide the selection of an appropriate model. The ACF plot assists in identifying the suitable Moving Average (MA) order, while the PACF plot provides insights into determining the appropriate AutoRegressive (AR) order. Examining and combining these plots enables one to make informed decisions about the optimal model for the given data.



As can be seen in the plot, there is seasonality, and we take one regular difference and one seasonal difference. Observing the ACF and PACF plots, it is evident that both graphs exhibit a cutoff after lag 3. It is crucial to highlight that the initial lags also hold significance in the ACF and PACF plots.

Suggested models: SARIMA(1,1,4)(4,1,1)$_{12}$, SARIMA(1,1,2)(4,1,1)$_{12}$, SARIMA(3,1,1)(4,1,1)$_{12}$, SARIMA(3,1,2)(4,1,1)$_{12}$

To initiate the modeling process, we begin constructing the highest-order model. Subsequently, we assess the significance of these models by examining the last estimated parameters for each component. If the ratio between these estimates and their corresponding standard errors (s.e) exceeds +2 or falls below -2, we can affirm the significance of the parameters and, consequently, the model. Following this, a determination is made regarding the significance of the last estimated parameters for each component. As the next step, we systematically reduce the order. After proposing several potential models that could suit the series, their significance, and performance are compared

Among the models SARIMA(1,1,2)(0,1,1)$_{12,}$ SARIMA(1,1,0)(0,1,1)$_{12,}$ SARIMA(1,1,0)(0,1,1)$_{12}$ are significant. In these three important models, BIC will be taken as the criterion. The best model is having smallest BIC. Of these three models, the model with the lowest BIC is SARIMA(1,1,0)(0,1,1)$_{12}$

Series: train2
ARIMA(1,1,0)(0,1,1)[12]
Coefficients:
ar1 sma1
-0.2528 -0.8422
s.e. 0.0396 0.0285
sigma^2 = 0.2222: log likelihood = -404.56
AIC=815.13 AICc=815.17 BIC=828.3

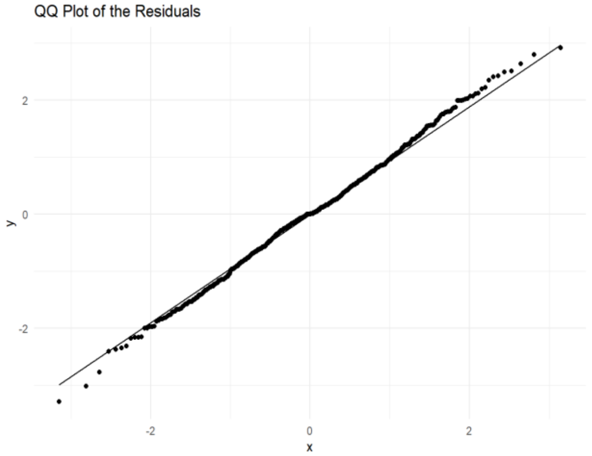
Table 1: summary of model

IV. MODELLING AND DIAGNOSTIC CHECKING

When developing and predicting with a time series model, it is crucial to assess its goodness of fit and validate the underlying assumptions. Achieving a model with an ideal fit is essential before generating ARIMA predictions. The time series analysis evaluation process resembles traditional regression analysis and relies on residual analysis.

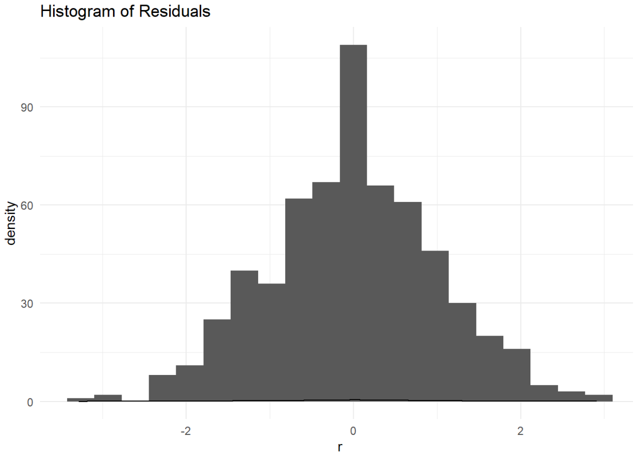
Formal tests like Shapiro-Wilk and Jarque-Bera and visual tools like Q-Q plots are employed to scrutinize the normality assumption. These methods help examine whether the residuals follow a normal distribution. Once the optimal model is identified, the diagnostic control phase commences, encompassing normality, serial correlation, and heteroscedasticity assessments.

The assessment of normality assumption typically starts with a visual examination using the QQ Plot, followed by formal tests such as Shapiro-Wilk and Jarque-Bera to analyze the observations depicted in the graph systematically



Graph 5 QQ Plot of The Standard Residuals

The plot shows that most of the model's residuals lie on the 45-degree straight line. This indicates that residuals are normally distributed.

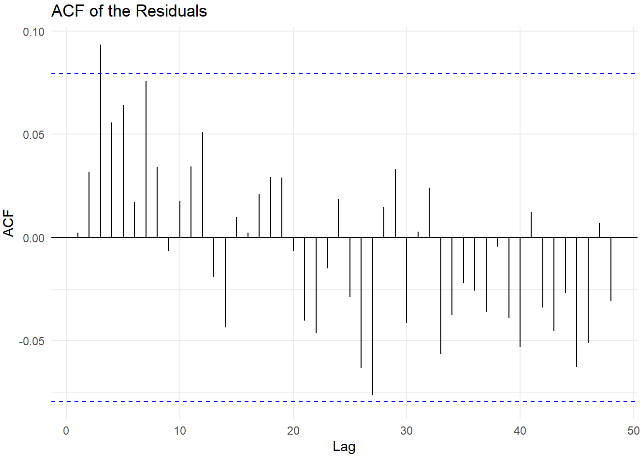


Graph 6 Histogram of the Standard Residuals

A Histogram of the residuals shows that they might have a symmetric distribution.

To be sure indicators of normality, Shapiro-Wilk and Jarque-Bera should be applied. Both Jargue Bera and Shapiro-wilk tests indicate residuals are normally distributed because p-values > 0.05

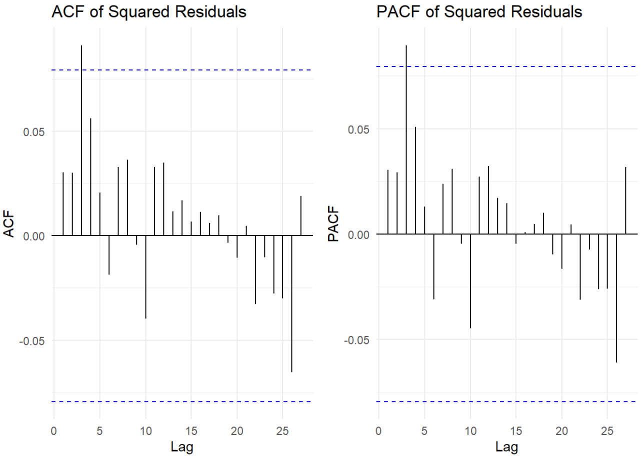
After that, The initial step in assessing serial correlation involves analyzing the Autocorrelation Function (ACF) plot of residuals. All spikes in the plot must align with the White Noise band to confirm the absence of correlation issues. Subsequently, the examination of serial correlation in residuals starts with a visual review of residual plots, resembling the method used for assessing normality. After this visual scrutiny, formal tests, such as the Breusch-Godfrey, Box-Ljung, and Box-Pierce tests, are employed to analyze further and quantify the serial correlation's existence. A thorough evaluation of serial correlation in the residuals is attained by integrating both visual examination and rigorous statistical tests.



Graph 7 ACF Plot of The Standard Residuals

ACF, almost all spikes are in the WN band, which shows no serial correlation in the residuals. To be sure, we have to apply the formal test mentioned previously, and all three tests show that residuals belonging to the best model are uncorrelated because all of them have p-values> 0.05

The final step of diagnostic checking is the heteroscedasticity of the residuals. To test this assumption, we can look at the ACF and PACF plots of squared residuals, and to be sure, we conduct formal tests, which are ARCH Engle's test and the Bresuch-Pagan test.



It is evident from the analysis that a majority of the squared residuals fall within the 95% White Noise Band, indicating homoscedasticity. This conclusion is further supported by formal tests, such as the Studentized Breusch-Pagan test and ARCH Engle's test, where non-significant results (p>0.05) affirm the homoscedastic nature of errors. Consequently, employing a GARCH-type model is unnecessary, given the consistent variance over time.

Having completed the ARIMA modeling phase, we focus on determining the most effective exponential smoothing model through the 'ets' function within the R forecasting package. The ideal exponential smoothing model for the given series is given below. This information is then assimilated to create a comprehensive forecasting strategy.

```
ETS(M,Ad,M)

 Call:
  ets(y = new_train, model = "MAM")

   Smoothing parameters:
     alpha = 0.6348
     beta  = 0.0528
     gamma = 0.0798
     phi   = 0.8478

   Initial states:
     l = 10.7079
     b = -0.4835
     s = 0.9267 0.9602 0.9577 0.9835 1.0057 1.0956
            1.0835 0.9436 0.9699 1.0113 1.039 1.0233

   sigma:  0.0559

#      AIC     AICc      BIC
 3354.266 3355.424 3433.709
```

Table 2: Summary of ETS MODEL

The notation "ETS(M, Ad, M)" indicates an exponential smoothing model characterized by Multiplicative error, Additive trend, and Multiplicative seasonality. These models are widely employed in time series forecasting to capture diverse patterns and trends in the data effectively. After fitting the model, the residuals of the ETS model are checked by the Shapiro-Wilk test, and it is seen that they do not follow normal distribution. (p<0.05). After applying the exponential smoothing model to the time series, a TBATS model was subsequently fitted, as outlined in Table 3

Table 3 : Summary of tbats model

```
   TBATS(0.866, {0,0}, 0.8, {<12,5>})

     Call: tbats(y = new_train)

     Parameters

       Lambda: 0.866193

      Alpha: 0.5830655

       Beta: 0.1742711

       Damping Parameter: 0.8

       Gamma-1 Values: -0.0003664937

       Gamma-2 Values: -0.006374737

     Seed States:
                   [,1]

       [1,]  8.08526788

       [2,]  0.06960364

       [3,] -0.18737934

       [4,]  0.36611197

       [5,] -0.09966732

       [6,]  0.07010455

       [7,]  0.09101711

       [8,]  0.11829016

       [9,]  0.12003440

      [10,]  0.22541976

      [11,] -0.04126089

      [12,]  0.03730990

      attr(,"lambda")

      [1] 0.8661932


      Sigma: 0.4253113

      AIC: 3298.739
```

Following the model fitting process, the residuals of the TBATS model were subjected to a Shapiro-Wilk test, revealing that they exhibit a normal distribution. (p>0.05). In the next step in our analysis, a Neural Network model was employed, incorporating past observations as input variables. The subsequent step involved fitting the Neural Network model. Specifically, the model was constructed utilizing historical observations as input variables. Detailed information about the model can be explored in Table 4.

```
 Series: new_train

  Model:  NNAR(25,1,13)[12]

  Call:   nnetar(y = new_train)

  Average of 20 networks, each of which is

  a 25-13-1 network with 352 weights

  options were - linear output units

  sigma^2 estimated as 0.05747
```

Table4:Summary of NNETAR Model

The model is NNAR(25,1,13)$_{12}$. It is a neural network with inputs Y_{t-1}, Y_{t-2}, Y_{t-24} and 13 neurons in the hidden layer.

when analyzing the residuals of NNETAR model, it was observed that do not follow normal distribution as indicated by Shapiro-Wilk tests ($p<0.05$).

Lastly, After fitting the Prophet model and providing its details in the appendix, forecast values were generated from each method using the forecast function. Subsequently, accuracy measurements were calculated for the training and test sets, and the results were summarized, respectively.

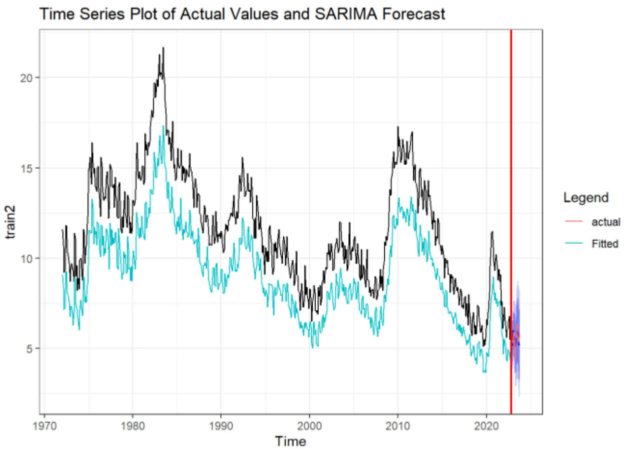
	SARIMA	ETS	PROBHET	TBATS	NNETAR
ME	2.536	-0.007	0.100	-0.007	0.001
RMSE	2.662	0.616	1.877	0.590	0.239
MAE	2.536	0.487	1.499	0.468	0.186
MPE	22.000	-0.221	-1.812	-0.194	-0.100
MAPE	22.000	4.352	13.502	4.261	1.773
ACF1	0.569	0.043	0.944	-0.022	-0.013

Table 5: The Train Accuracy of models

	SARIMA	ETS	PROPHET	TBATS	NNETAR
ME	0.059	-0.186	-0.036	-0.455	-0.540
RMSE	0.411	0.429	0.468	0.598	0.884
MAE	0.370	0.317	0.353	0.461	0.619
MPE	0.552	-3.982	-1.215	-8.675	-10.040
MAPE	6.761	6.094	6.695	8.788	11.523
ACF1	0.160	0.169	0.245	0.052	0.474

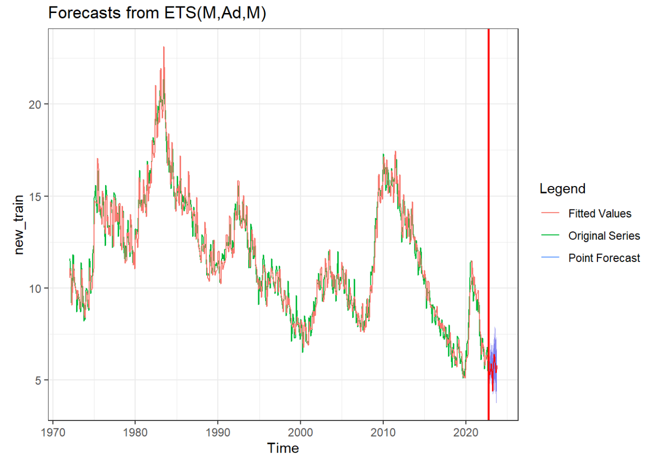
Table 6: The Forecasting Performance of Models

The predictive performance of the models is also clearly seen by examining the relevant graphs. Analyzing these visual representations can provide additional information and complement the quantitative measurements provided. It provides a more comprehensive understanding of each model's strengths and weaknesses in capturing key patterns and trends in the data. The red vertical line in the chart signifies the forecast origin, set for August 2022. The red line indicates the test set within the forecast zone, and the shaded blue area represents the 95% prediction interval. In this context, the blue line denotes the point forecast. Interestingly, the SARIMA model uniquely displays the original series in black, while other models adhere to conventional series representations. The first graph depicts results from the SARIMA model.



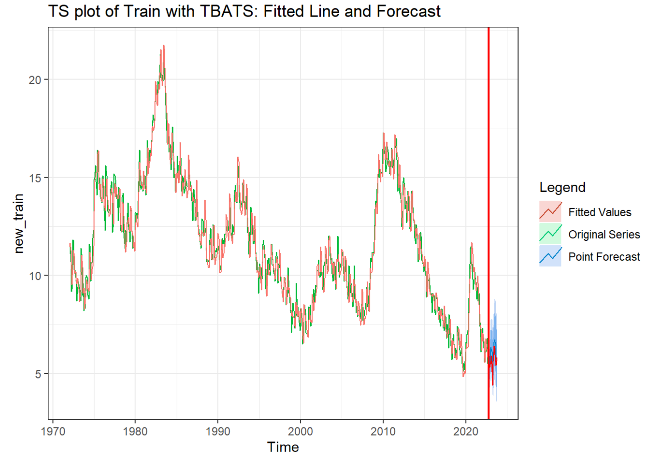
Graph 8 Forecast Plot Of Sarıma

The SARIMA model lacks alignment between the specified line and the actual series. Post-estimates, actual values, and point estimates are generally consistent, except for the decline and subsequent rise in the unemployment rate at actual values, which are not reflected in the point estimates. The difference between actual values and point estimates is relatively small at this rate of rise and fall. Graph number nine is constructed based on the ETS model:



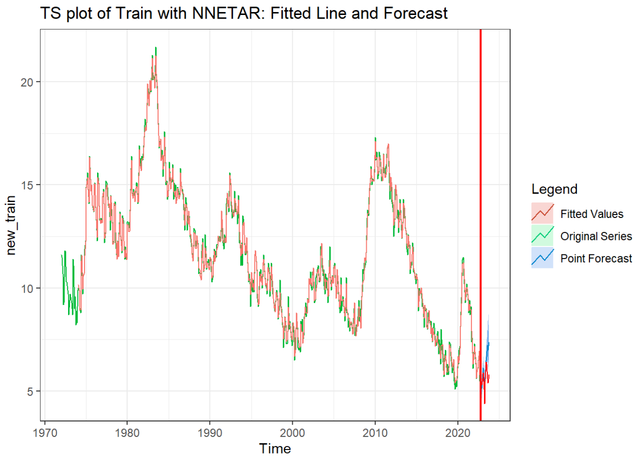
Graph 9 Forecast Plot of ETS

ETS seems to be offering the most precise alignment with the training data, encompassing a notable portion of the actual data points within the training set. Furthermore, the graphical representation substantiates the accuracy of the findings. Within the forecast zone, the actual values closely align with the point forecasts, indicating ETS's ability to capture minor fluctuations in the forecasted range. Noteworthy is the remarkably narrow 95% prediction interval compared to other models. These collective outcomes strongly affirm that ETS stands out as the superior choice among the models. Moving on to the graph ten it pertains to TBATS:



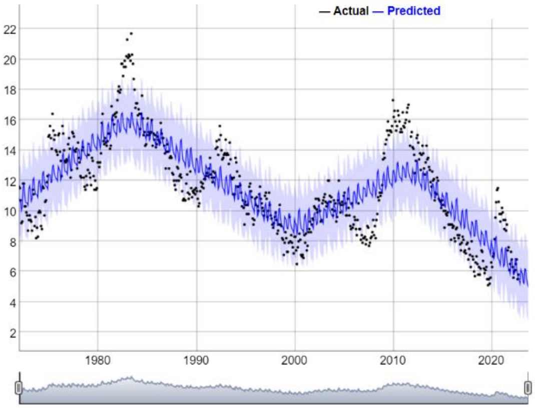
Graph 10 Forecast Plot of TBATS

In the TBATS model, the fitted value and the original series match each other well. In the forecast zone, the actual value and point forecast are close; the difference between them is slight. As a result, it is a good model for TBATS. Nevertheless, there are better options. The following plot showcases the lines fitted and predicted by the Neural Network model:



Graph 11Forecast Plot of NNETAR

There does not appear to be any problem with the compatibility of the fitted value and the original series. The forecast zone is incompatible with the point forecast and the actual value. The result we found in accuracy was confirmed once again with the graph. NNETAR is the worst among the models. 5. Lastly, the below graph is the product of series modeling using Prophet



Graph 12 Forecast Plot of Probhet

The shaded light blue region signifies the confidence intervals. While the predicted values on the plot may not align perfectly with the actual data points, it is worth noting that a substantial portion of the actual points lies within these confidence intervals.

V. DISCUSSION AND CONCLUSION

This research examines fluctuations in unemployment rates among Black-African individuals in the United States, employing a thorough time series analysis conducted in R-Studio. The methodology commenced with identifying anomalies, and the application of Box-Cox transformation helped stabilize the variance. Stationarity was assessed through KPSS and ADF tests, revealing a consistent unit root. Regular differencing was applied to address this issue. Subsequently, the HEGY test indicated a seasonal unit root problem, addressed by implementing seasonal differencing. A careful examination of ACF and PACF plots guided the selection of potential models, emphasizing parameter significance and BIC values.

Diagnostic checks were carried out on the residuals of the chosen model, confirming normality and ruling out serial correlation or heteroskedasticity, indicating constant variance over time. The final phase involved forecasting using R packages like ARIMA, ETS, TBATS, PROPHET, and NNETAR. ETS emerged as the most effective model, displaying superior performance in modeling the series and predicting future data points.

This project comprehensively explored time series analysis in R-Studio, covering descriptive data analysis, data preparation, relevant testing, and interpretation of results. Grounded in actual data, the study underscores the significance of comprehending and predicting changes in unemployment rates among Black-African individuals in the United States.

VI. REFERENCES

Hayes, A. (2023, August 9). *What is unemployment? Understanding causes, types, and measurement*. Investopedia. https://www.investopedia.com/terms/u/unemployment.asp