PART 1

INTRODUCTION

For this assignment, the analysis of the U. S. seasonally adjusted personal consumption expenditures (PCE) with several forecasting models is going to be performed to evaluate and compare their effectiveness. The dataset "PCE.csv" contains time series data that shows the personal consumption expenditures over a certain period. This data has been seasonally adjusted to exclude those seasonal variations emanating from holidays, weather, or any other recurring seasonal patterns. These modifications are considered a necessary step to remove the noise and make it easier to identify the underlying patterns and cycles in the data. The dataset contains 779 rows and 2 variables which contain details related to PCE.

Variable	Data Type	Description
DATE	Character (dd-mm-yyyy)	The date when the expenditure was recorded
PCE	Numeric	The amount of personal consumption expenditure in U.S. dollars.

TABLE 1 DATA DICTIONARY OF PCE

DATA UNDERSTANDING AND PRE-PROCESSING

In the process of analysing the PCE (Personal Consumption Expenditures) dataset follow the following steps

Time series object: Import all the important libraries and load the data from 'pce.csv' then the 'DATE' column was converted to the Date data type "%d/%m/%Y". Time-series object was generated from the 'PCE' data with frequency=12 (monthly) and start=c(1959, 1), end =c(2023,11). This step was crucial for aligning the data in chronological order.

Missing values: Checking missing values using "is.na()" function. There are 43 missing values. Given the complexities potentially embedded within the PCE data, the na_kalman() method was considered appropriate for imputing missing data.

After the imputation process, two statistical tests were conducted to compare the mean and variance of the original and imputed datasets.

Two Sample Test	Value	p-value	Original vs imputed ratio
t-test (mean)	t = 0.84327	0.3992	1.03934
F-test (variance)	F = 1.0174	0.8123	1.0173

TABLE 2 TEST RESULTS

The chosen imputation method for handling missing values in the PCE data effectively maintained both statistical properties mean and variance of the original series.

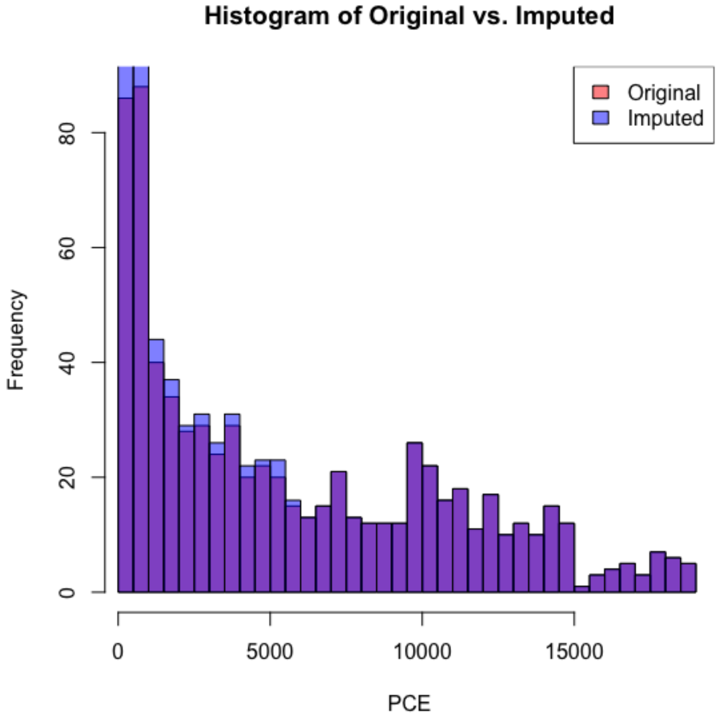


FIGURE 1 HISTOGRAM. ORIGINAL VS IMPUTED

Decomposition of time series: Use the decompose() function to decompose PCE imputed data.

Additive decomposition is suitable for the PCE data as it typically exhibits relatively stable seasonal patterns and linear trends, making it easier to separately analyse and understand the underlying trend, seasonality, and irregular components.

$$Y_t = T_t + S_t + E_t$$

Where:

Y_t, T_t, S_t, E_t observed value, trend, seasonality, and error at time t for PCE.

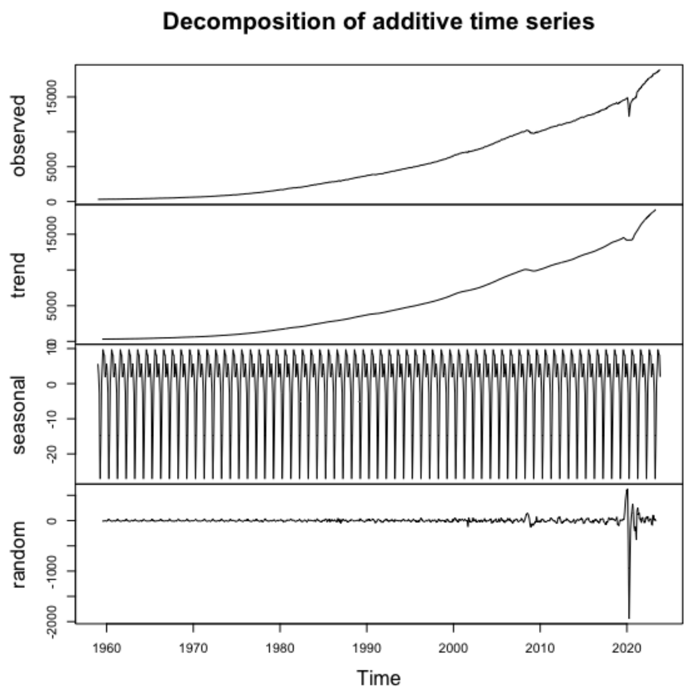


FIGURE 2 ADDITIVE DECOMPOSITION

TREND

The PCE time series exhibits a strong upward trend over the years, indicating consistent growth in personal consumption expenditures.

The linear regression model reveals a significant upward trend in PCE, a negligible p-value, and a high R-squared value of 0.9119, demonstrating a strong linear fit of PCE with the time index.

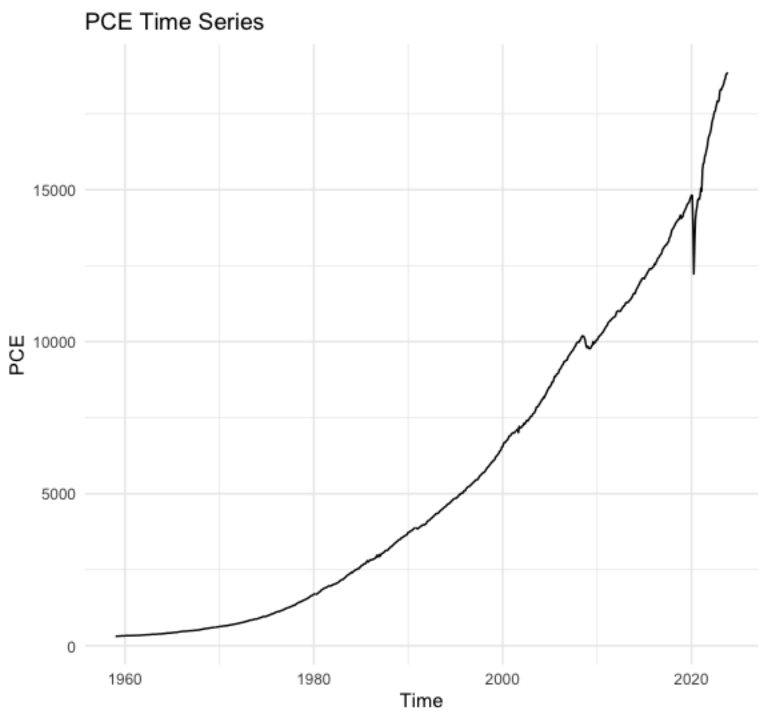


FIGURE 3 TREND OF PCE

Seasonality

The seasonal plot of PCE shows similar patterns each year, with noticeable drops in the year 2020, as the data is seasonally adjusted, these variations are likely smoothed out to reveal trends and cycles unrelated to seasonality better.

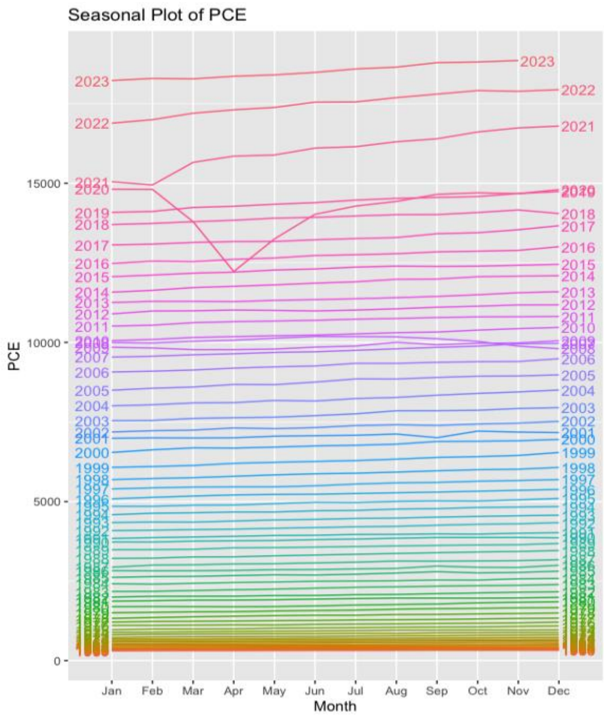


FIGURE 4 SEASONALITY OF PCE

Volatility

The graph shows a dramatic increase in the rolling standard deviation of PCE around 2020, indicating a significant rise in volatility, likely due to economic disruptions caused by the COVID-19 pandemic.

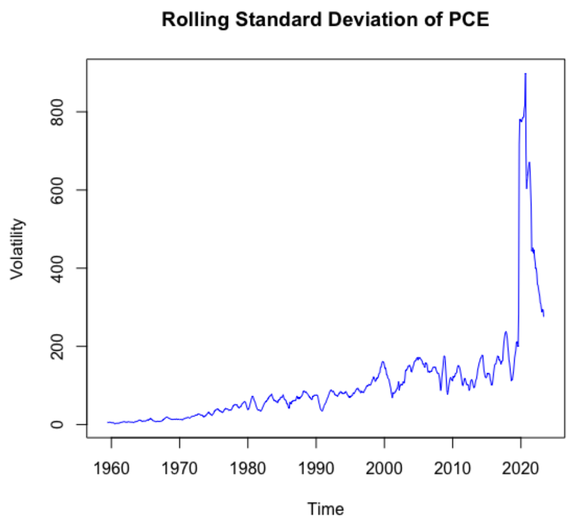


FIGURE 5 VOLATILITY OF PCE

STATIONARITY ANALYSIS

1. **Augmented Dickey-Fuller (ADF) Test:**
   - Null Hypothesis (H0): The PCE data has a unit root, suggesting it is non-stationary.
   - Alternative Hypothesis (H1): The PCE data does not have a unit root, indicating it is stationary.
2. **KPSS Test (Kwiatkowski-Phillips-Schmidt-Shin Test):**
   - Null Hypothesis (H0): The PCE data is stationary around a deterministic trend.
   - Alternative Hypothesis (H1): The PCE data is not stationary around a deterministic trend.

Note:

- p-value less than 0.05, for stationary in the ADF test.
- p-value greater than 0.05, for stationary in the KPSS test.

Imputed PCE Data				
Test	Value	lag	P value	Result
ADF	1.783	9	0.99	Not stationary
KPSS	10.449	6	0.01	Not stationary

FIGURE 6 TEST RESULTS FOR IMPUTED DATA

1ST Differenced PCE Data				
Test	Value	lag	P value	Result
ADF	-7.1521	9	0.01	Stationary
KPSS	2.782	6	0.01	Not Stationary

FIGURE 7 TEST RESULTS FOR 1ST DIFFERENCED DATA

2nd Differenced PCE Data				
Test	Value	lag	P value	Result
ADF	-15.044	9	0.01	Stationary
KPSS	0.00499	6	0.1	Stationary

FIGURE 8 TEST RESULTS FOR 2ND DIFFERENCED DATA

DATA TRANSFORMATION

The goal of applying the Box-Cox transformation to the PCE data would be to reduce skewness and variance in the series. We get optimal **lambda = 0.02773.**

$$PCE = y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

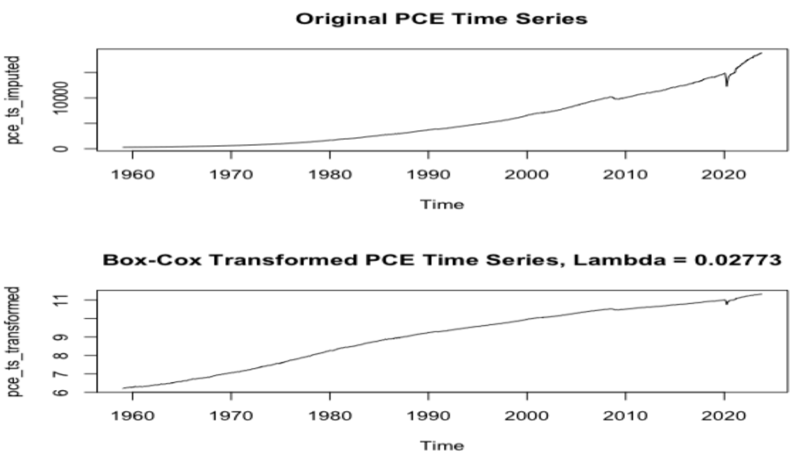


FIGURE 9 BOX-COX TRANFORMATION.

METRIC SELECTION

MAPE (Mean Absolute Percentage Error) is calculated as the average of the absolute differences between the forecasted and actual values, expressed as a percentage of the actual values. It provides errors as a percentage, making the relative forecasting error easy to understand.

$$\text{MAPE} = \left(\frac{100}{n}\right)\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

MASE (Mean Absolute Scaled Error) measures the accuracy of the forecast relative to a naïve baseline prediction, typically the one-period lag. It compares the model's error to a simple naïve model, indicating whether the advanced model offers any significant improvement.

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n} |Y_t - \hat{Y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n} |Y_t - Y_{t-1}|}$$

Where:

- Y_t , Y_{t-1} : PCE at time t and $t-1$.
- $\hat{Y}_t$: PCE forecast at time t.
- n : number of observations.

Note: NOT to consider **RMSE** metric because measures of the magnitude of the forecast error, giving a relatively high weight to large errors.

MODELLING

TRAIN TEST SPLIT

The PCE time series is divided into training and test data using a train-test split. The training data consists of 80% of the time series and ends in December 2010. The test data comprises approximately 20% of the time series and starts from January 2011.

Note: Fit all models on Box-Cox transformed PCE data with **lambda = 0.02773** as the modelling parameter.

DRIFT MODEL

The Drift Model would also be suitable as it would adapt to the long-term upward trend observed in the PCE data. The drift method projects the trend forward by calculating the average change between the first and last values and extending this trend into the future.

Forecast equation:

$$\hat{y}_{T+h} = y_{779} + h\left(\frac{y_{779} - y_1}{778}\right)$$

	MAPE	MASE
Test Set	18.5351	13.3155

FIGURE 10 DRIFT_ACCURACY

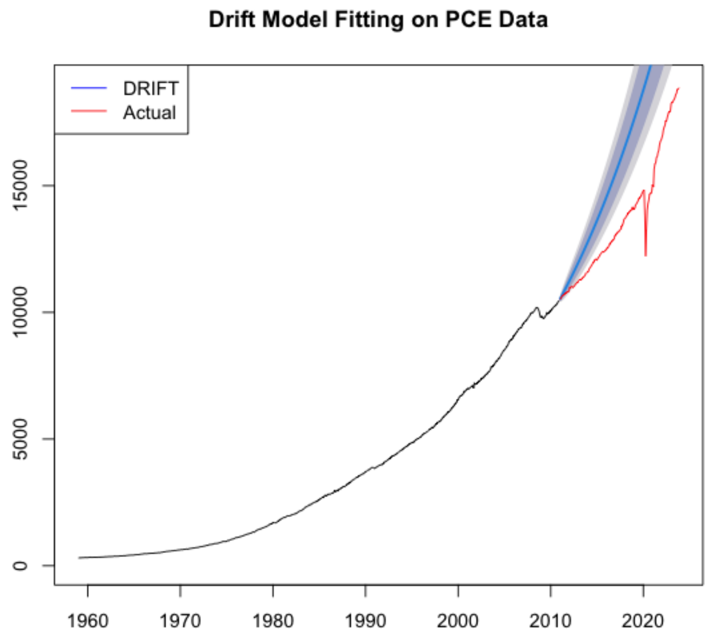


FIGURE 11 DRIFT_FORECAST_PLOT

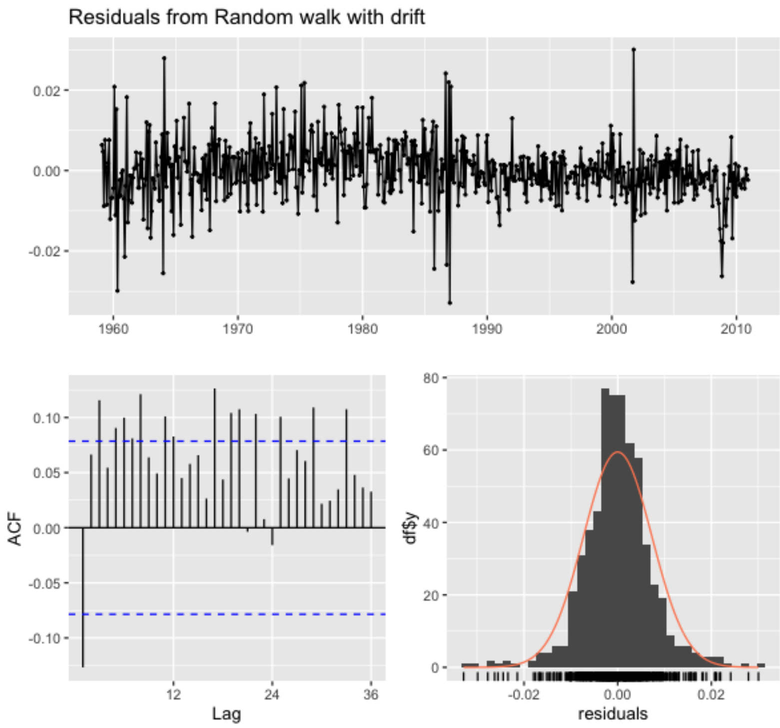


FIGURE 12 DRIFT_RESIDUAL _PLOT

The **Ljung-Box test** results for the residuals from the Random Walk with Drift model indicate significant autocorrelation (p-value < 1.051e-11), suggesting that the model does not adequately capture the underlying structure of the PCE data.

EXPONENTIAL MODEL

The **Holt linear method** is a time series forecasting technique that extends simple exponential smoothing to capture linear trends in the data. This method is especially suited for data exhibiting a trend but no seasonal fluctuations.

Forecast equation:

$$\hat{y}_{t+m} = \ell_t + mb_t$$

The optimal parameters for Holt's linear method suggest a significant smoothing coefficient **alpha = 0.765** for the level, indicating a high sensitivity to recent data changes. Additionally, a smaller smoothing coefficient **beta = 0.0334** for the trend implies a more gradual adjustment to changes in the trend. Holt's linear approach for forecasting involves utilising the starting **level l = 6.2003** and **trend b = 0.0053** to predict future values by iteratively adjusting the level and trend estimations using these smoothing parameters.

	MAPE	MASE
Test set	5.3471	4.1810

FIGURE 13 HOLT (A, N) _ACCURACY

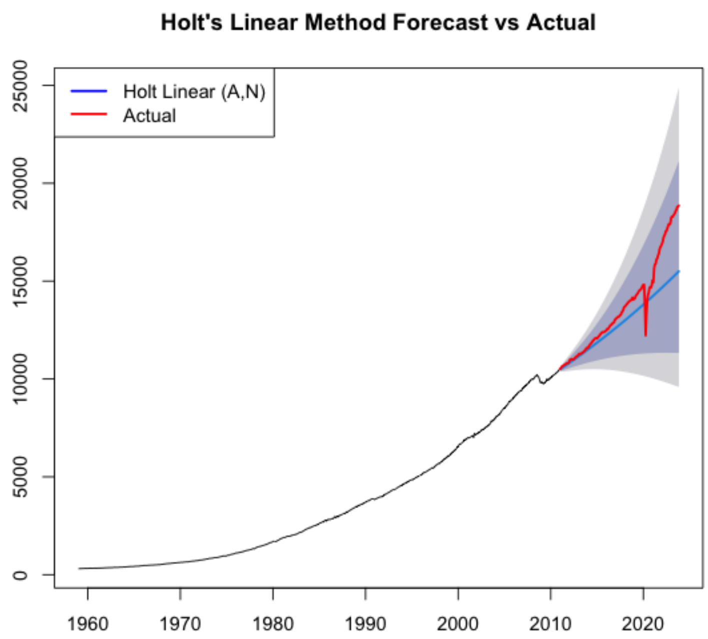


FIGURE 14 HOLT (A, N)_FORECAST

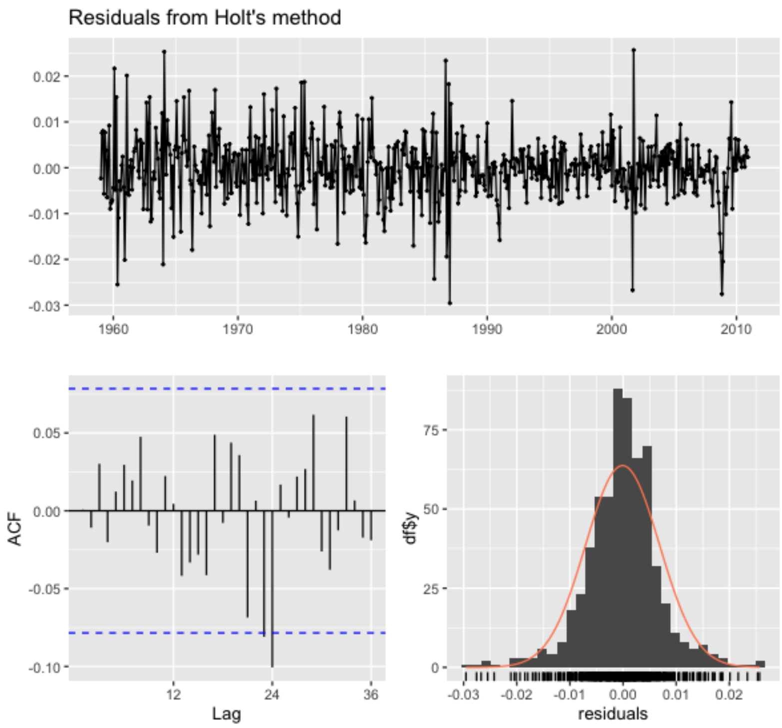


FIGURE 15 HOLT(A,N)_RESIDUAL_PLOT

The **Ljung-Box test** for the residuals from Holt's method shows a p-value of 0.4022, indicating significant autocorrelation in the residuals, suggesting that Holt's model may not adequately capture all the autocorrelative structures in the data.

ARIMA MODEL

The **ARIMA model** is crucial for analysing seasonally adjusted PCE data as it allows for detailed modelling of non-seasonal patterns and irregularities.

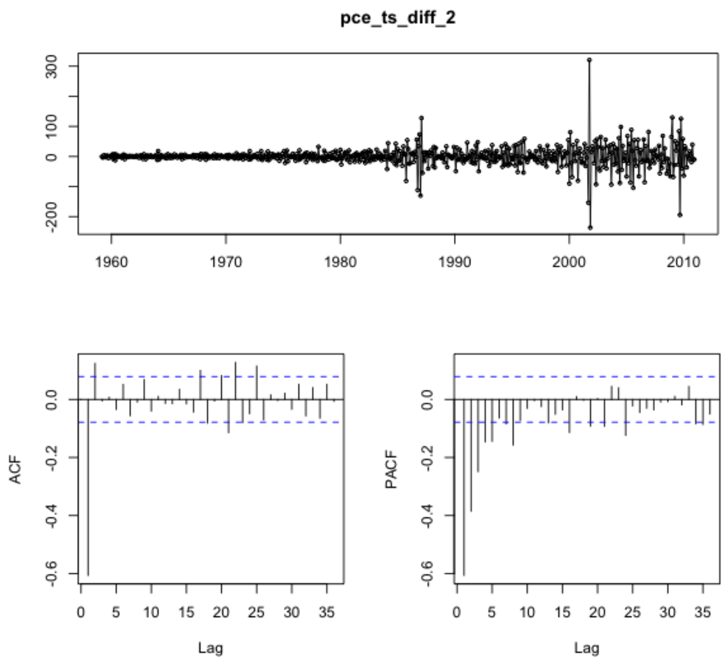


FIGURE 16 ACF AND PCF PLOT OF 2ND DIFFERENCED.

From the plot showing the ACF and PACF of the second-differenced PCE data, the PACF displaying significant spikes at lags 1, 2, and 3, combined with the ACF exhibiting significant correlations at lags 1 and 2, suggest an ARIMA model with parameters p=3, d=2, and q=2.

Forecast equation based on the above parameters:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-B)^2 Y_t=(1+\theta_1 B+\theta_2 B^2)\epsilon_t$$

	MAPE	MASE
Test set	5.3035	4.1489

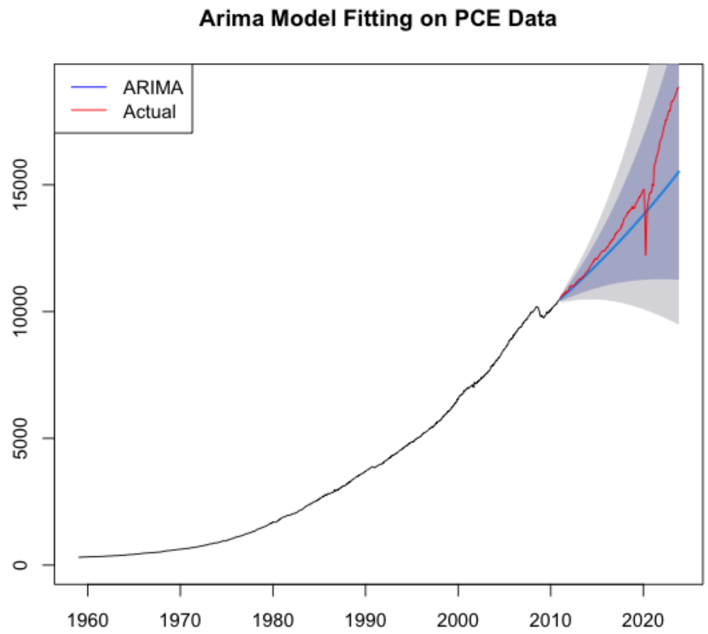


FIGURE 17 ARIMA_FORECAST

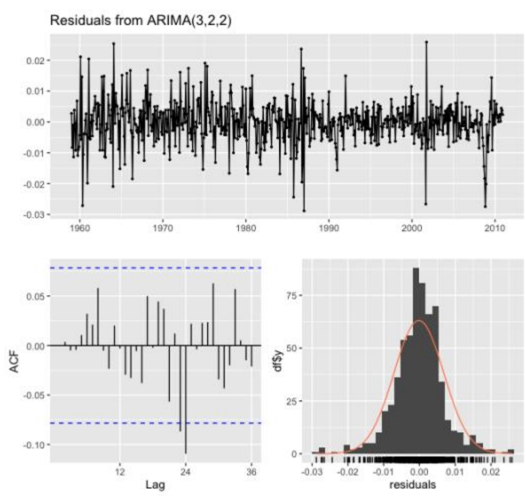
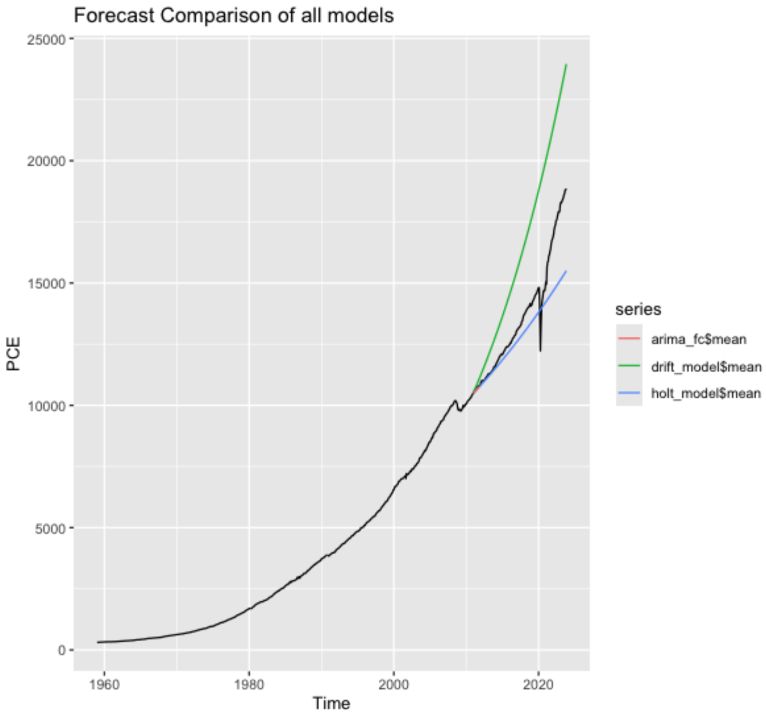


FIGURE 18 ARIMA_RESIDUAL

The **Ljung-Box test** for the residuals of the ARIMA(3,2,2) model indicates a p-value of 0.1731, suggesting that there is no significant autocorrelation in the residuals at the 5% level, supporting the adequacy of the model fit.

COMPARATIVE ANALYSIS



	MAPE	MASE
Drift	18.5351	13.3155
Holt Linear	5.3471	4.1810
Arima	5.3035	4.1489

FIGURE 19 COMPARATIVE ACCURACY

The ARIMA model, specifically ARIMA (3,2,2), demonstrated superior forecasting accuracy for the PCE data with the lowest MAPE of 5.3035 and MASE of 4.1489, indicating its effectiveness in capturing the dynamics of the series compared to Drift and Holt Linear models.

OCTOBER 2024 PREDICTION

Now we will predict the mean value of October 2024 using the **ARIMA model** with parameters **p=3, d=2, and q=2** on Box-Cox transformed PCE data with **lambda = 0.02773.** Also arima model was a good fit using the Ljung-Box test

Final forecasted PCE for October 2024: **19682.6142756194**

ROLLING FORECAST

One-step forecasting without re-estimation for PCE data involves using a pre-determined model to predict the next data point based on historical data, without updating the model's parameters as new actual data becomes available.

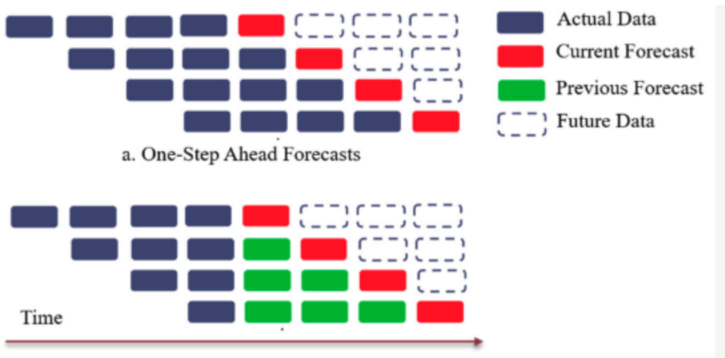


FIGURE 20 ROLLING FORECAST(ONE-STEP) (SAURABH SURADHANIWAR ET AL., 2021)

	MAPE	MAE
Drift	0.6643489	92.44065
Holt Linear	0.5038304	69.78976
Arima	0.5303565	73.31249

FIGURE 21 ROLLING ACCURACY

Holt linear is best for One-step forecasting without parameter re-estimation for PCE data.

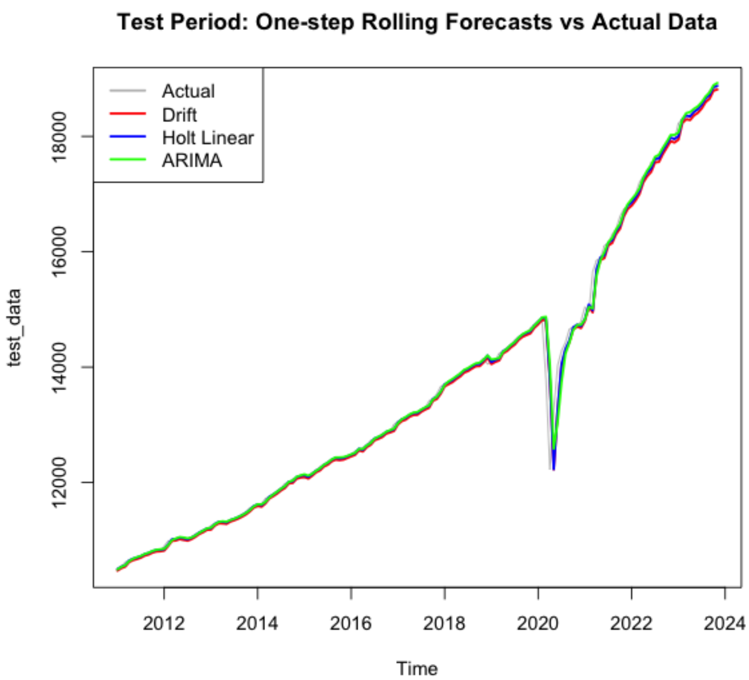


FIGURE 22 ROLING_FORCAST COMPARISON