Time Series Analysis on Gross Domestic Product

Mit Desai (202103013)* and Sishir Mishra (202312003)†
Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India
SC475, Time Series Analysis

In this project, we first performed the Exploratory Data Analysis, which includes the visualization of the given data of GDP, feature selection using a correlation matrix between different features of the given data, and standardization and normalization of the data for the scaling of the selected feature. We have also transformed the data whenever and wherever we found the data to be fluctuating at a large level. After this, we found the Autocovariance function and autocorrelation function to verify whether data is stationary or not; then we did data splitting of 80% Training data and 20% Testing data and tried to fit different Time Series models for non-stationary data converted to stationary form on which the models were fitted. Then we find different errors in different models as they will be overfitting using which we choose the model with the least error. Now, we will be applying that model beyond the test data time and find out the futuristic data time series. Our main observation are that the GDP, CPI and PPI data is non-stationary and how to forecast their values is what we show in our project.

I. INTRODUCTION

This part of the report discusses the description of time series data that we considered for the analysis of Gross Domestic Product (GDP). The data is of a total 62 years since 1960 and includes quarterly data of each year till 2022. The data consists of originally five columns which are as under:

- **Gross Domestic Product:** GDP is a measure of sum total of the prices of all commodities and services produced in the region or country during particular period, for example, a year or a quarter.

  It offers a complete picture of a countrie´s economy performance and mostly is taken into account to appraise the condition and growth of a economy.

  GDP can be calculated using three approaches: the former (accumulation of the value added at each stage of production), the latter (aggregation of consumption, investment, government spending and net exports), and the last (combination of all the incomes earned by individuals and firms known as the income approach).

- **Producer Price Index:**PPI compares the change in what domestic sales receive in terms of their output value on a recurring basis.

  It keeps at the check the shifts in prices made at the original producers before the goods are sold to consumers.

  PPI in the PPI market is a commonly used a signal of the level of inflation pressures within the economy, which may later on affect the consumer prices.

- **Consumer Price Index:**CPI calculates the average level of cost increase in services and goods purchased by urban clients for the time period selected.

  It is an indicator of inflation that comes into effect and is used to monitor the changes in the prices of household purchases.

  CPI tracks a host of goods and services, such as housing, transportation, food, and medical care; and these are usually segmented into distinct classes which make it possible to map inflation rates more accurately.

- **M1 Money Supply:** M1 is a sum of level of notes and coins plus checking deposits and other assets, which can be easily converted into the money.

  It is a measure of the most liquid assets that make up the money supply and serve as a gauge of the funds availability for transactions in the economy. Changes in M1 can reflect shifts in consumer spending patterns, monetary policy actions by central banks, and overall economic conditions.

- **Debt-to-GDP Ratio (DDNSA):**The debt-to-GDP ratio is a measure of a country's total debt (government, corporate, and household) relative to its GDP. It provides insight into a country's ability to service its debt obligations relative to the size of its economy. A high debt-to-GDP ratio may indicate that a country is heavily indebted and could face challenges in repaying its debt, while a low ratio suggests a healthier fiscal position.

These economic indicators play crucial roles in understanding and monitoring the performance, stability, and trends within an economy. Policymakers, investors, businesses, and consumers rely on these indicators to make informed decisions and assess the overall economic health of a country.

*Electronic address: 202103013@daiict.ac.in
†Electronic address: 202312003@daiict.ac.in

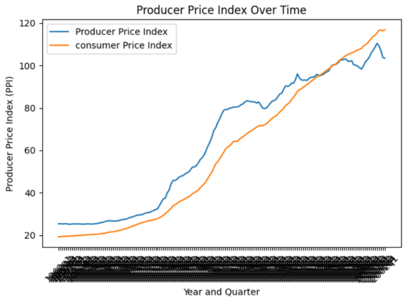


FIG. 1: PPI and CPI graph

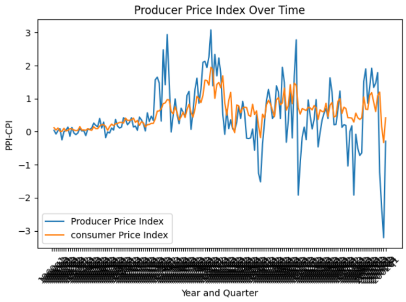


FIG. 2: dppi and dcpi plot

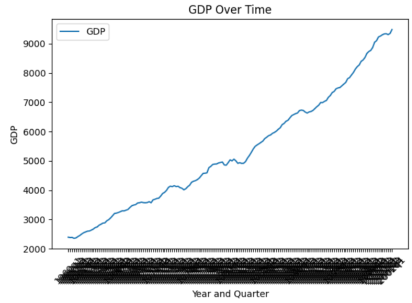


FIG. 3: GDP plot

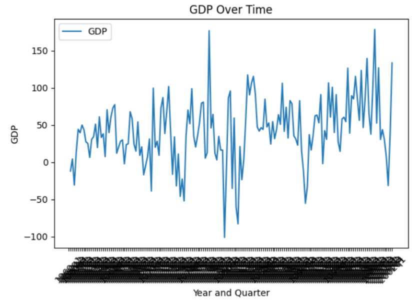


FIG. 4: GDP difference Vs Time plot

II. DATA VISUALIZATION

We plotted various types of line plots to see what relation is being established between various features of time series data of GDP that is found. There were no missing values found in the data. The data was discovered from the Kaggle website and was verified with the data found on the website of Reserve Bank of India. The first step in the approach towards visualizing the data was to describe the data in terms of a number of rows(count) in the data, mean, median, mode, min, 25 percentile, 50 percentile, 75 percentile, and max of the columns in the data of GDP. We also found the standard deviation of the columns. This helps us find out the outliers in the given data. We did not ignore any data point as only three data points breached the upper bound of 1.5x the max and 1.5x the min data. The line plots between different features of the GDP data is as under:

The first figure shows the plot of the Producer Price Index and Consumer Price Index over all the quarters of the year from 1960 to 2022. The total count is 169 time stamps. We observe that both PPI and CPI increase over time and after the year 2007 Q4, the Producer Price Index crosses the Consumer Price Index. This shows that the Consumer demand is more than what the producers are supplying after 2007 Q4. Hence, the inflation rate has increased thereafter(Law of supply and demand). The second figure shows the $PPI_t - PPIt - 1$ and $CPI_t - CPI_{t-1}$ with respect to time. The difference is taken to see the fluctuations in both the time series. We can see that the fluctuations in dPPI and dCPI increase with time. Thus, from the graph, it is evident that it is non-stationary as the variance (fluctuations) increases concerning time. For a time series to be stationary, Variance and mean must be constant and not changing concerning time. The third figure shows a plot of GDP concerning time. From the graph, it is clear that the GDP of the country has increased linearly over time. It is observed that the GDP has seen a fall over 2 consecutive years only for 4 times in 60 years. The fourth figure shows $dGDP = GDP_t - GDPt - 1$ plot concerning time. This plot shows fluctuations in GDP with concerning time. We see four major fluctuations upwards and downwards, simultaneously. Similar observation was noted in the GDP vs time graph as well. The fifth graph is similar to the fourth one as we have just taken the logarithm of values of GDP to descale the plot due to high original values of GDP in the data. This figure is

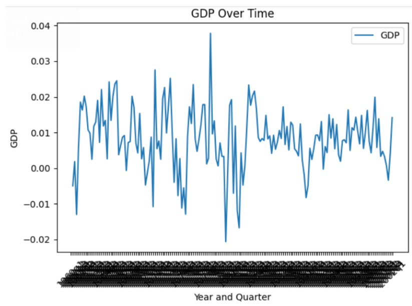


FIG. 5: Difference Natural Logarithmic GDP Vs Time plot

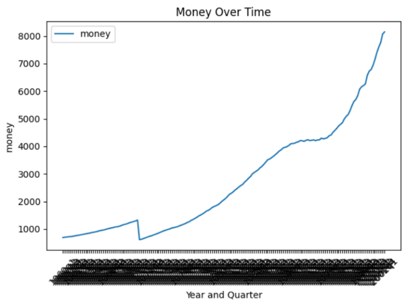


FIG. 6: Money Flow over Time

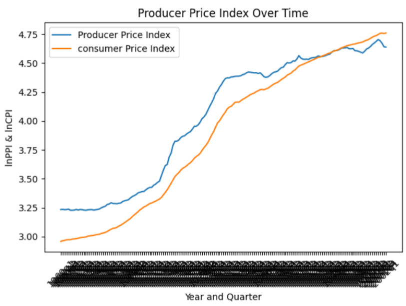


FIG. 7: Natural Logarithmic Difference of PPI and CPI over time

the graph of Money flow in the market with time. As the money flow increases the GDP increases and vice versa. This is the direct relation of money flow with GDP. Natural logarithmic function applied to PPI and CPI data of GDP are portrayed in the seventh figure which in turn shows the descaled value of PPI and CPI hence easy for interpretation. It is similar to PPI and CPI data with respect to time.

III. STATIONARITY OF THE DATA

In this section, we will be doing a stationarity test of all the columns concerning to time. The reason to check stationarity is that we need to forecast the future value of GDP. Stationarity is the property in which a time series is said to be stationary if the mean and variance of the time series are constant and not dependent on the time. The value of the autocorrelation function may depend upon the lag of the time series. For all the time series given in the data, we will be verifying whether they are stationary or not. For the stationarity test, we will use ADF test. A statistical technique called the Augmented Dickey-Fuller (ADF) test is used to assess if a time series is stable. A time series is said to be stationary if its statistical characteristics, such its variance and mean, remain constant over time. This is a summary of how the ADF test operates:

Null Hypothesis (H0): According to the ADF test, the time series is non-stationary since it has a unit root. Stated differently, H0 claims that the series is non-stationary.

Alternate Hypothesis(H1): The time series is stationary, according to the alternative hypothesis (H1). The series is assumed to be stationary if the alternative hypothesis is accepted rather than the null hypothesis.

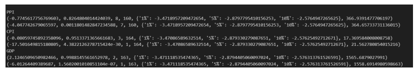
ADF Statistic:ADF Statistics Based on the existence of a unit root in the autoregressive model of the time series, the ADF test calculates a test statistic. If the test statistic is greater than or equal to crucial values from a certain distribution, the null hypothesis should be rejected.

Critical Values: The sample size, significance level (e.g., 1%, 5%, or 10%), and inclusion of a constant and/or trend in the model all influence the critical values. Critical values for various circumstances, such as those with or without a constant and/or trend, are provided by the ADF test.

Decision Rule: The time series is said stationary and the null hypothesis is rejected if the computed ADF statistic is smaller than the critical value. On the other hand, if the calculated ADF statistic exceeds the critical value, the series is non-stationary and the null hypothesis cannot be ruled out.

Interpretation: The time series is probably stationary if the null hypothesis is rejected. That being said, if the null hypothesis is accepted, it suggests that the series is non-stationary and may need further differencing or might be some other transformations necessary to make it necessary.

Here, as per the above rule of ADF test, we can see that the value of ADF statistic in the **PPI**, **CPI**, and **GDP** exceeds the critical value, thus we reject the alternate hypothesis (H1). Hence, We conclude that all three of the **PPI**, **CPI**, and **GDP** time series are **non-stationary** (As shown in figure 8).



```
PPI
(-0.7745617756769603, 0.8264804014424039, 8, 160, {'1%': -3.4718957209472654, '5%': -2.8797795410156253, '10%': -2.5764947265625}, 366.9391477706197)
(-4.047742679065597, 0.0011801482847234588, 7, 160, {'1%': -3.4718957209472654, '5%': -2.8797795410156253, '10%': -2.5764947265625}, 364.6573373113 6015)
CPI
(-0.08059745892358096, 0.9513371365661683, 3, 164, {'1%': -3.47086589632514, '5%': -2.879330279087651, '10%': -2.57625492712671}, 17.369584008008758)
(-17.501649815180805, 4.3822126278715424e-30, 1, 164, {'1%': -3.47086589632514, '5%': -2.879330279087651, '10%': -2.57625492712671}, 21.562780854015216)
GDP
(2.1246509650982466, 0.9988145561652978, 2, 163, {'1%': -3.471118535474365, '5%': -2.8794405060097024, '10%': -2.576313761526591}, 1565.6879027991)
(-6.01264409389687, 1.5602001010851104e-07, 1, 163, {'1%': -3.471118535474365, '5%': -2.8794405060097024, '10%': -2.576313761526591}, 1558.6914980598663)
```

FIG. 8: ADF test result of PPI, CPI and GDP

We also tried another test of stationarity that we used to do in the course itself which is the **ACF** test. We produce the Autocorrelation function for each of the time series i.e. **PPI**, **CPI**, and **GDP**. The time lag taken in each of the autocorrelation functions of the three-time series is **5**. The ACF plot vs the time lag is given in figures 8, 9, and 10 of CPI, PPI, and GDP, respectively. Here, as the ACF plot of the CPI, PPI, and GDP is decaying very slowly, thus we again conclude using the ACF test that all three of the time series are non-stationary.

Now, we will try the ACF test on the **ddCPI**, **ddGDP** and **ddPPI** which is basically **ARIMA model** with differencing d=2 as we took differencing twice. We considered it as the ARIMA model because PPI, GDP, and CPI are non-stationary and so is any ARIMA model. The plot for the same is in figure 15, 16, 17, 18, 20 and 19.

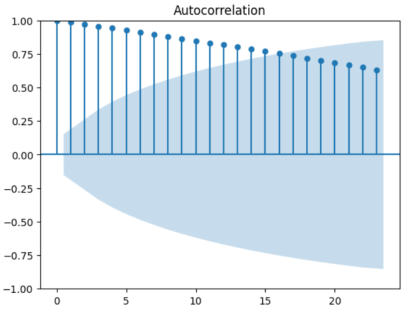


FIG. 9: ACF plot of PPI

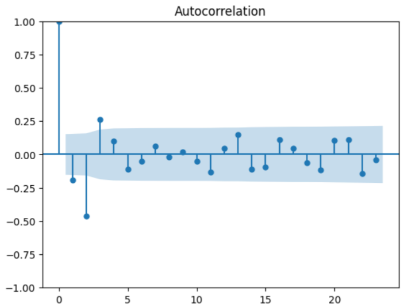


FIG. 12: ACF plot of ddPPI

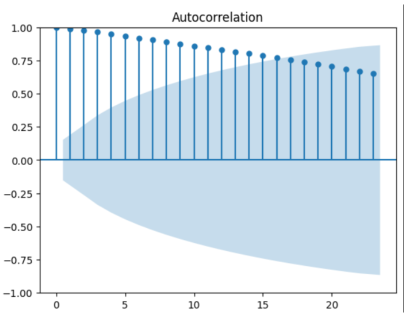


FIG. 10: ACF plot of CPI

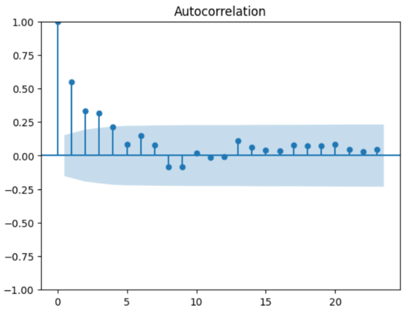


FIG. 13: ACF plot of ddCPI

IV. DECOMPOSITION OF TIME SERIES

There are 14 types of decomposition of the time series. But we mainly consider only 2 of them. One is **multiplicative time series** and **additive time series**. We denote the multiplicative time series as $x_t = T_t * S_t * \epsilon_t$. The additive time series is denoted by $x_t = T_t + S_t + \epsilon_t$. Here, T_t=Trend component of the time series (x_t), S_t=Seasonal component of the time series (x_t) and ϵ_t= Residual component of the time series (x_t). Now, here for CPI, PPI, and GDP we have the graphs of the trend, seasonality, and fluctuations decomposed from a single time series for both additive and multiplicative decompositions. We say that any time series is decomposed into additive if the seasonal component of the time series remains constant with respect to time. Whereas, it is said to be decomposed to multiplicative if **the seasonal component changes as per the level of the time series value at that time**. Here, as all the seasonal components of PPI, CPI, and GDP are increasing over time hence we can say that the given time series can be decomposed into **multiplication**. The plots for the same is in figure 9, 10, 11, 12, 13, and 14.

V. RESULTS

We then tried to fit the model of time series in the given data of GDP. So as we have proven that the time series of **PPI, CPI, and GDP** are all non-stationary. Now, for

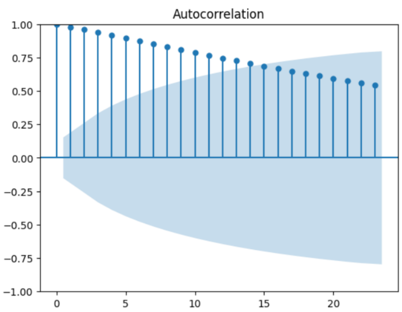


FIG. 11: ACF plot of GDP

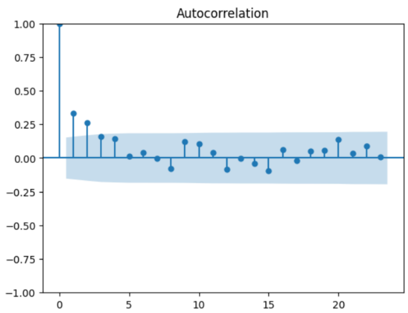


FIG. 14: ACF plot of ddGDP

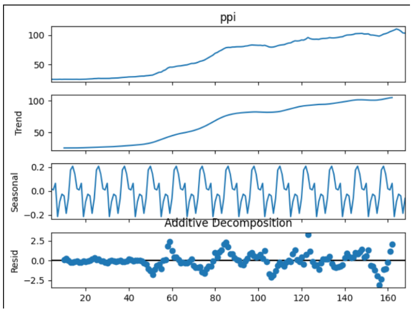


FIG. 15: Additive decomposition of PPI

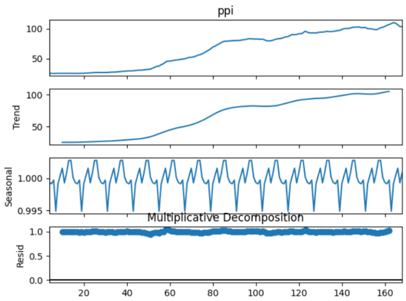


FIG. 16: Multiplicative decomposition of PPI

a non-stationary series, we fit **ARIMA(p,d,q)** model to the time series. Now, ARIMA is a non-stationary time series model which is a combination of **AR(p)** model which is of order p and, the MA(q) model which is of order q. **I** here stands for a number of times we do difference here, which will at the end become value of d. We will keep differencing until the final output time series becomes a stationary time-series model be it only AR(p) or MA(q) or else **ARMA(p,q)** models. Here, the model of CPI is taken as an example, as per the above graph figure number 13, the ACF plot decays rapidly of **ddCPI** that is when we take d=2 and differencing is applied **twice** to the given time series of CPI. Thus, we prove that dd-CPI is a stationary time series. Hence, only the value of stationary time series can be forecasted and it does not apply to the non-stationary time series. When we get the time series of CPI to ddCPI, the value of the order of the Autoregressive and Moving Average is **p=0** and q=3. The output of the same is given in figure 21.

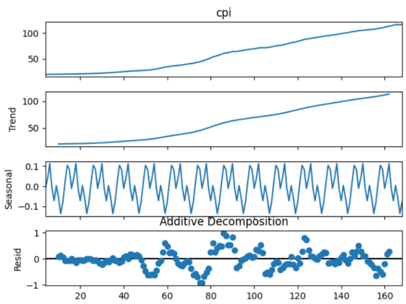


FIG. 17: Additive decomposition of CPI

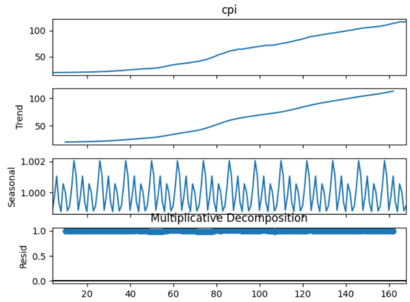


FIG. 18: Multiplicative decomposition of CPI

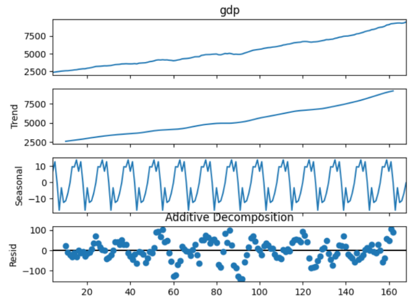


FIG. 19: Multiplicative decomposition of GDP

VI. CONCLUSIONS

From the given project, we conclude that the time series analysis of given data of GDP can be forecasted further but the time series of GDP was really tough to find of what order was it of in terms of ARIMA(p,d,q). So, for its prediction, we need to advance more towards other non-stationary models than the ARIMA models. We here conclude that, the problem statement would have got forecasted further if we have had a larger data size which is tough to find as we could have tried many other lags in the given time series and would have proven it stationary for any value of d less than 2. That would have made

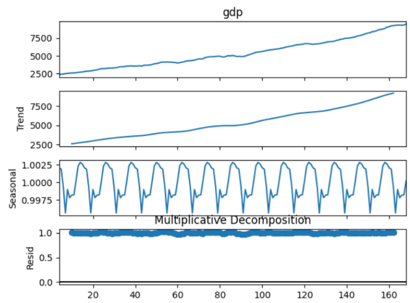


FIG. 20: Multiplicative decomposition of GDP

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  165
Model:               SARIMAX(0, 2, 3)   Log Likelihood                  -0.702
Date:                Fri, 10 May 2024   AIC                              9.404
Time:                        14:22:40   BIC                             21.779
Sample:                             0   HQIC                            14.428
                                - 165
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.2307      0.064     -3.629      0.000      -0.355      -0.106
ma.L2         -0.4860      0.062     -7.817      0.000      -0.608      -0.364
ma.L3          0.3872      0.065      5.973      0.000       0.260       0.514
sigma2         0.0587      0.005     12.633      0.000       0.050       0.068
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                42.56
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):              10.04   Skew:                            -0.64
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.15
```

FIG. 21: ARIMA model of order p=0, d=2 and q=3 with seasonality integrated to it

our work of forecasting easier. The other thing that we realized is that the forecasting would have been easier if though the data size was less but if there would have been multiple time series in the data and if a lot of parameters were under consideraton more than these five than we would have tried forming a certain kind of network through the correlation between each of them and what to consider and what not to, this would have made us learnt that how a model learns by itself about what to consider and what not to hence making the model self sufficient. The last but not the least thing that we conclude is that, the time series analysis had taught us really well about how to examine the given data, what patterns to find and how to forecast the future data. The major purpose was to consider studying the subject through the project and that we have tried our best to fulfil.

[1] Lecture Slides of SC475 taught by Dr. Mukesh Tiwari.

[2] Introduction to time-series and forecasting by Peter. J. Brockwell and Richard .A. Davis