Forecasting Unemployment Rate during the Pandemic

written by Mohit Verma
October 21, 2020


Forecasting

Forecasting, in simpler terms, is a process of predicting future values of a variable based on past data and other variables that are related to the variable being forecasted. For example, values of future demand for tickets for a particular airline company depend on past sales and the price of its tickets.

Time-series data is used for forecasting purposes. According to Wikipedia 'A time series is a series of data points indexed in time order. Most commonly, a time

series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data.' An example of time series data for monthly airline passengers is given below:

Figure 1
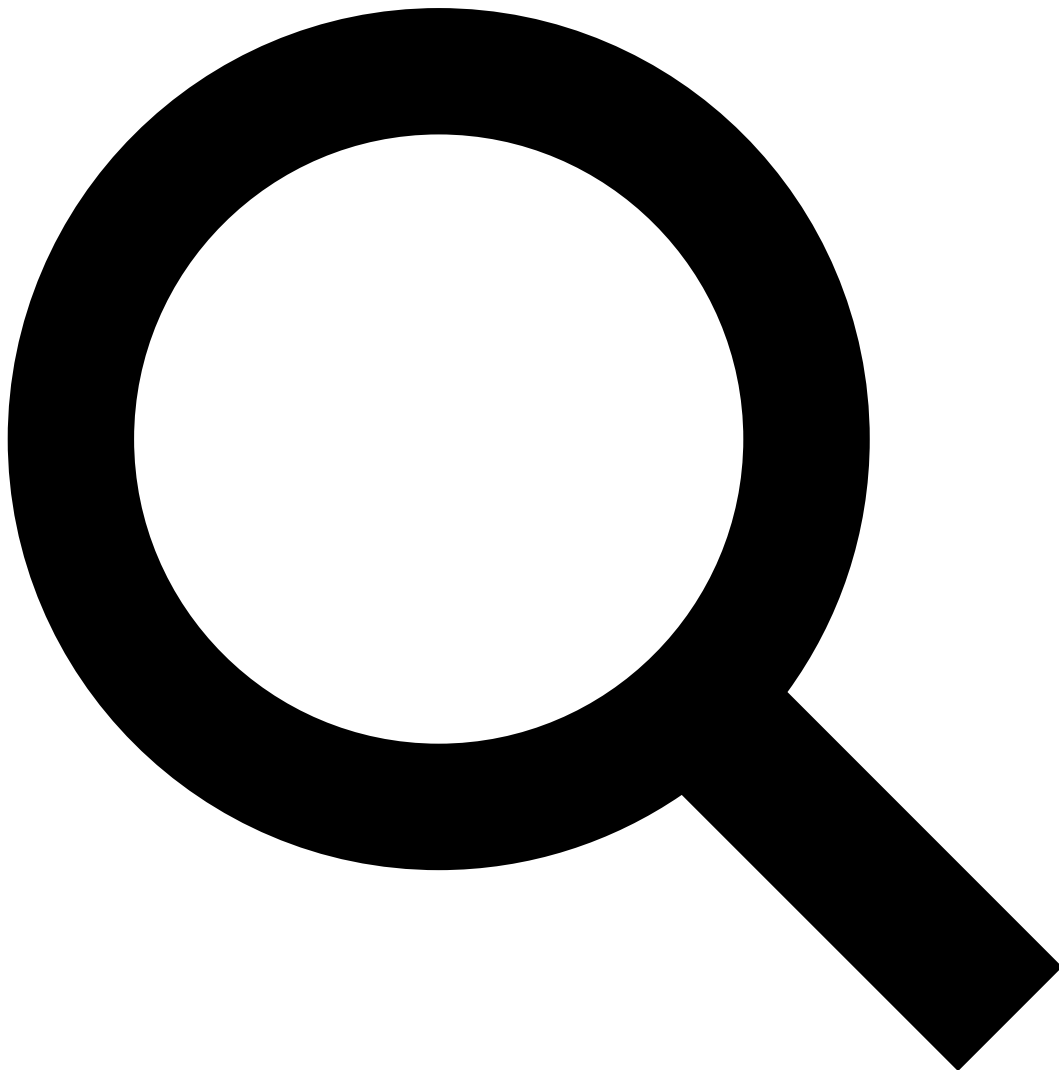
More technically, it is modelled through a stochastic process, _Y(t)_. In a time series data, we are interested in estimating values for _Y(t+h)_ using the information available at time _t._

Unemployment rate

Unemployment is the proportion of people in the labour force who are willing and able to work but are unable to find work. It is an indicator of the health of the economy because it provides a timely measure of the state of labour market and hence, overall economic activities. In wake of the impact of Covid-19 on economic activities throughout the world, unemployment rate analysis and forecasts have become paramount in assessing economic conditions.

In India, unemployment rates have been on the higher end in recent times. According to data released by Statistics Ministry, unemployment rate for FY18 was 6.1%, the highest in 45 years. It is no co-incidence that GDP rates have also been declining successively for the past few years. The shock that Covid-19 has given to the economy has only worsened our situation. The unemployment rate rose to 27.1% as a whopping 121.5 million were forced out of work.

Figure 2
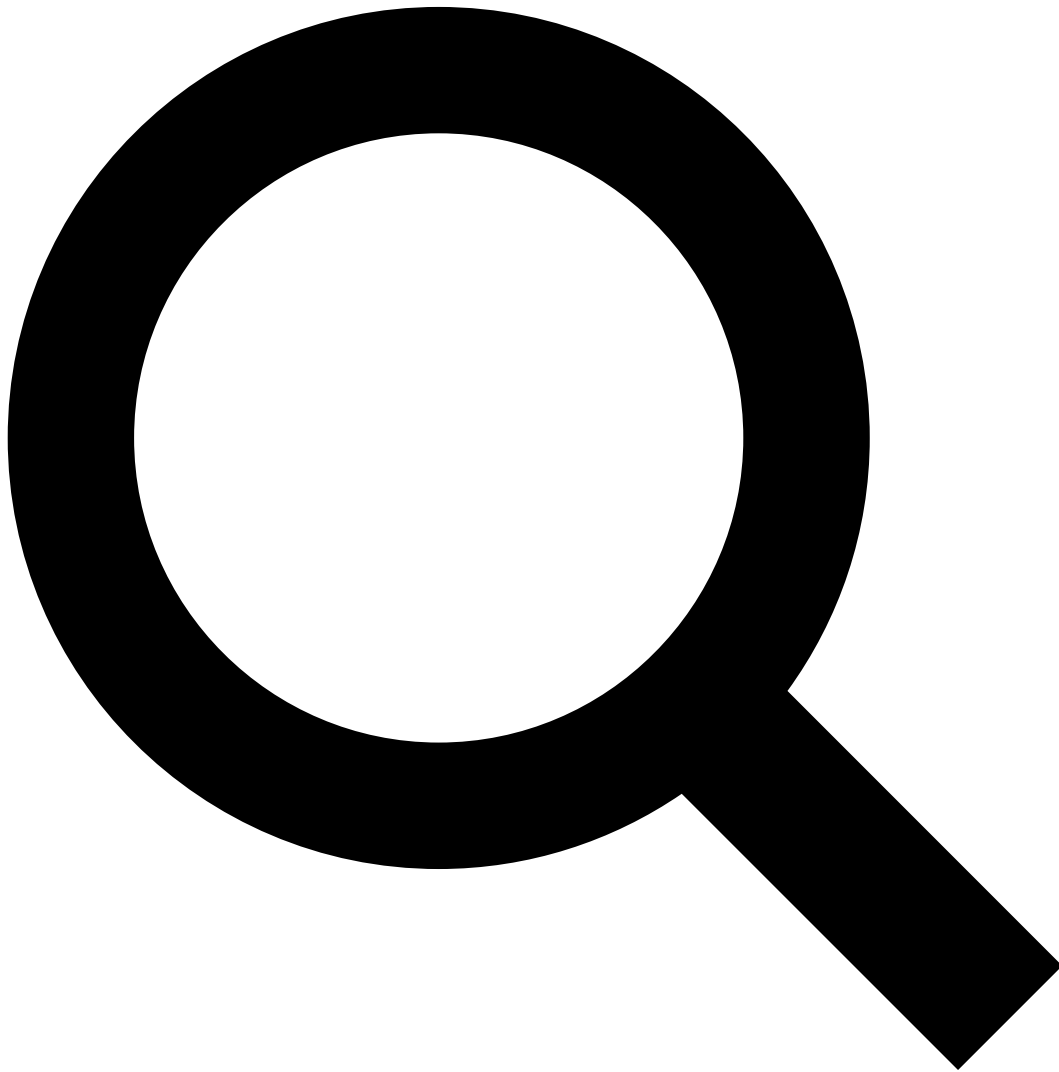
Source: CMIE

Methodology

The data used to forecast unemployment rates was sourced from CMIE website, which surveys over 43,000 households to generate monthly estimates since January 2016. The data has 56 monthly observations ranging from January 2016

to August 2020, data before 2016 was not available.

Four popular econometric forecasting models (ARIMA, Naïve, Exponential Smoothing, Holt's winter method) were used and the best performing model was chosen to forecast unemployment till December 2020.

The forecasting models were programmed in R. The relevant codes are available upon request with the author. The Dicky-Fuller test and the Chow test for structural breaks were conducted using STATA, results of which are presented further in the article.

Before beginning the analysis, I believe that the limitations of the analysis should be mentioned:

- The sample size of 56 observations is not sufficient for a thorough analysis, ideally the sample size should have been 2-3 times larger than the available data. Smaller sample sizes lead to skewed forecasting results which are prone to errors.
- The unemployment data from CMIE is an estimate and is a secondary source. In India, primary data is only collected once in 3-4 years, thus the forecasting results are only as good as the source of the data.
- This is a univariate analysis, an Okun's law based analysis of Unemployment rate as a function of GDP (output) and past trends would have been more suitable. However, since GDP data is only available quarterly and there are only 56 monthly observations available, it would have rendered the analysis insignificant with only 19 quarterly observations.
- Forecasting being based on past trends, is prone to errors. The negative shock provided by Covid-19 to the economies worldwide has made it all the more difficult to forecast. A Bloomberg study analysed over 3,200 forecasts by IMF since 1999 and found that over 93% of the forecasts underestimated or overestimated the results with a mean error of 2 percentage points.

Checking the stationarity of data

In order to model build a model, we need to make sure that the series is stationary. For intuitively checking the stationarity, I plotted the data over time as indicated in Figure 2 above. I also plotted the correlograms (autocorrelations versus time lags) as shown in Figure 8 and 9 in appendix. The plot of data over time indicate varying mean, variance and covariance. The ACF and PACF plot show that autocorrelations function are persistent indefinitely.

We perform the Augmented Dickey Fuller test at 2 lags. Result of the ADF test is shown in Table 1 below. The test statistic is insignificant at 5 per cent and the p-value is 0.1709, which is more than the accepted benchmark of 0.05. We fail to reject the null hypothesis of non-stationarity. We conclude that our series is non-stationary.

Dicky-Fuller test on raw data

Table 1

		—– Interpolated Dickey-Fuller —–		
	Test statistic	1% critical value	5% critical value	10%critical value
Z(t)	-2.303	-3.576	-2.928	-2.599

MacKinnon approximate p-value for Z(t) = 0.1709

Converting the non-stationary series into stationary

In order to transform the non-stationary series into stationary, we use differencing method (computing difference between consecutive observations). We plot the data over time, ACF and PACF again as shown in Figure 5 below and figure 10 and 11 in appendix, respectively. From the figures, we can intuitively say that the transformed series is stationary. Further, we used Augmented Dickey-Fuller tests to ascertain the stationary of our series. Table 2 shows the result of the ADF test. The test statistic is significant at 1,5 and 10 per cent levels and the p-value is less than 0.05. We reject the null hypothesis of non-stationarity of our series. The tests confirm that the series is stationary.
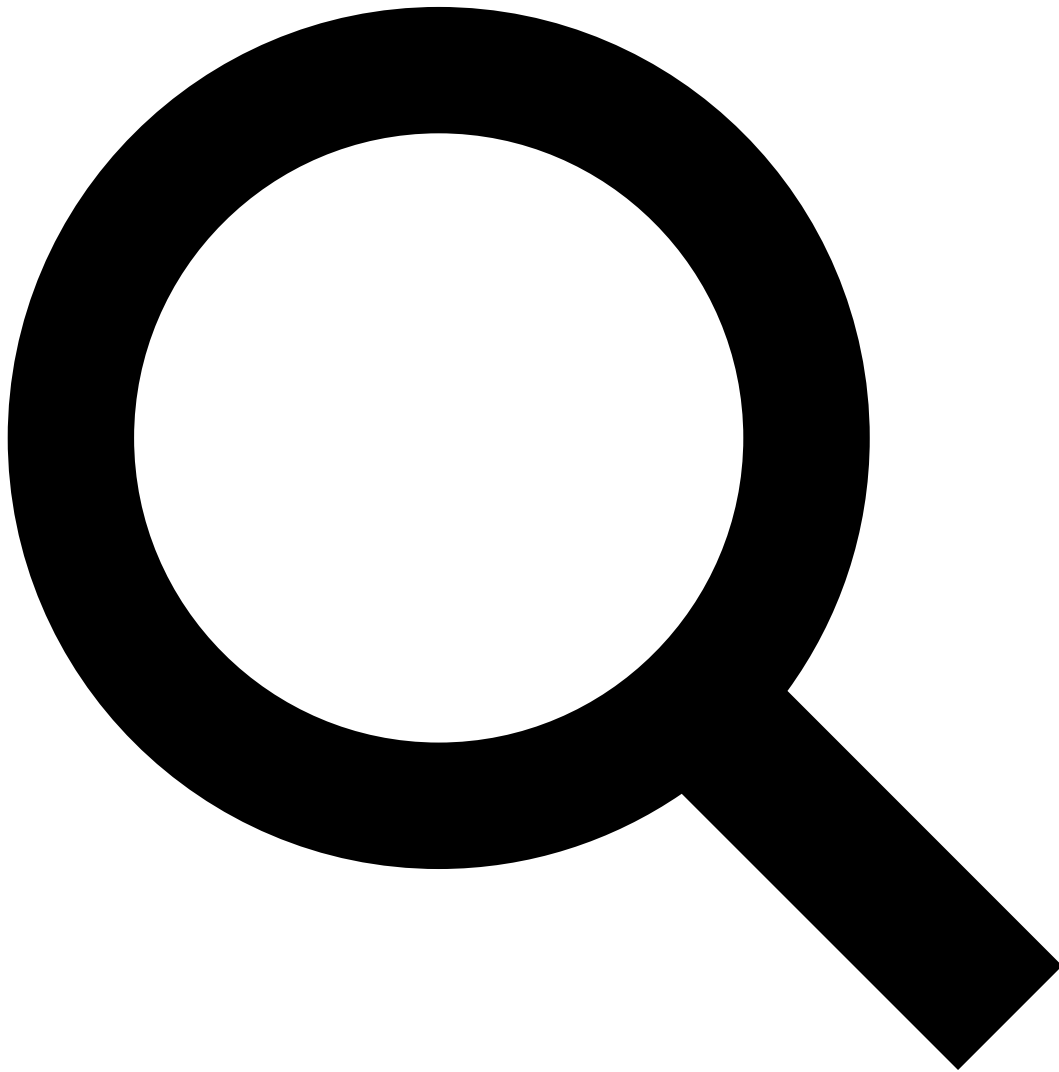
Dicky-Fuller test on first difference data

Table 2

	Test statistic	—– Interpolated Dickey-Fuller —–		
		1% critical value	5% critical value	10%critical value
Z(t)	-5.035	-3.576	-2.928	-2.599

MacKinnon approximate p-value for Z(t) = 0.0000

Figure 3

Naïve model

Naïve models are the simplest of forecasting models and provide a benchmark against which other more sophisticated models can be compared. Thus, a Naïve model serves as an ideal model to start any comparative analysis with. In a naive

model, the forecasted values are simply the values of the last observation. It is given by

$$\hat{y}_{t+h|t} = y_t.$$

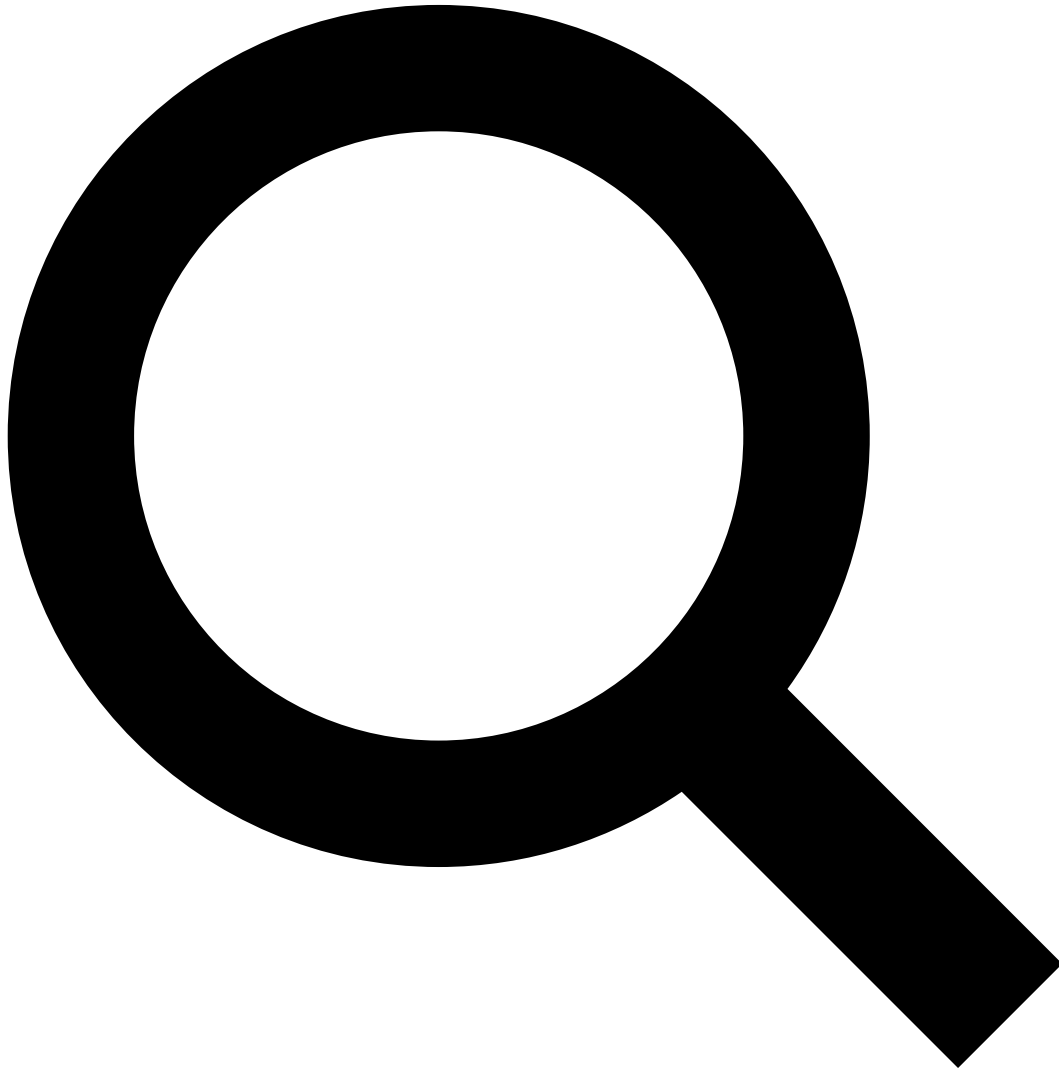Forecast results from Naïve method are presented below in figure 4 and table1.

Figure 4

Table 1

	Point forecast	Lo 80	High 80	Low 95	High 95

Sept	8.35	4.861900	11.83810	3.0154109	13.68459
Oct	8.35	3.417081	13.28292	0.8057517	15.89425
Nov	8.35	2.308433	14.39157	-0.8897794	17.58978
Dec	8.35	1.373799	15.32620	-2.3191783	19.01918

Box-Jenkins Approach

1. Identification of ARIMA (p, d, q) model

The data was split into training and testing dataset in 80:20 ratio. The training data was used for estimating the model, while the model was tested on the remaining 20 percent data. This is done in order to forecast the future values of the time series data.

p, d and q in (p, d, q) stand for number of lags, difference and moving average respectively.

The model best fitting the data was (0,1,3) as its Akaike Information Criterion (AIC) was the lowest amongst all the possible combinations of the order of the ARIMA model.

The residuals from Arima model were found to be normally distributed, with a mean of 0.09 and zero correlation. This causes a bias in the estimates. To solve the problem of bias, we will add 0.09 to all forecasts. The ACF and line graph of residuals is attached in the appendix.

After identification and estimation, several diagnostic tests were conducted to check if there were any uncaptured information in the model. Results of the diagnostics tests have been omitted from the article in interest of length.

2. Forecasting

The model that has been constructed was used to forecast unemployment rates for the next four months. The results are presented below in figure 5 and table 2.
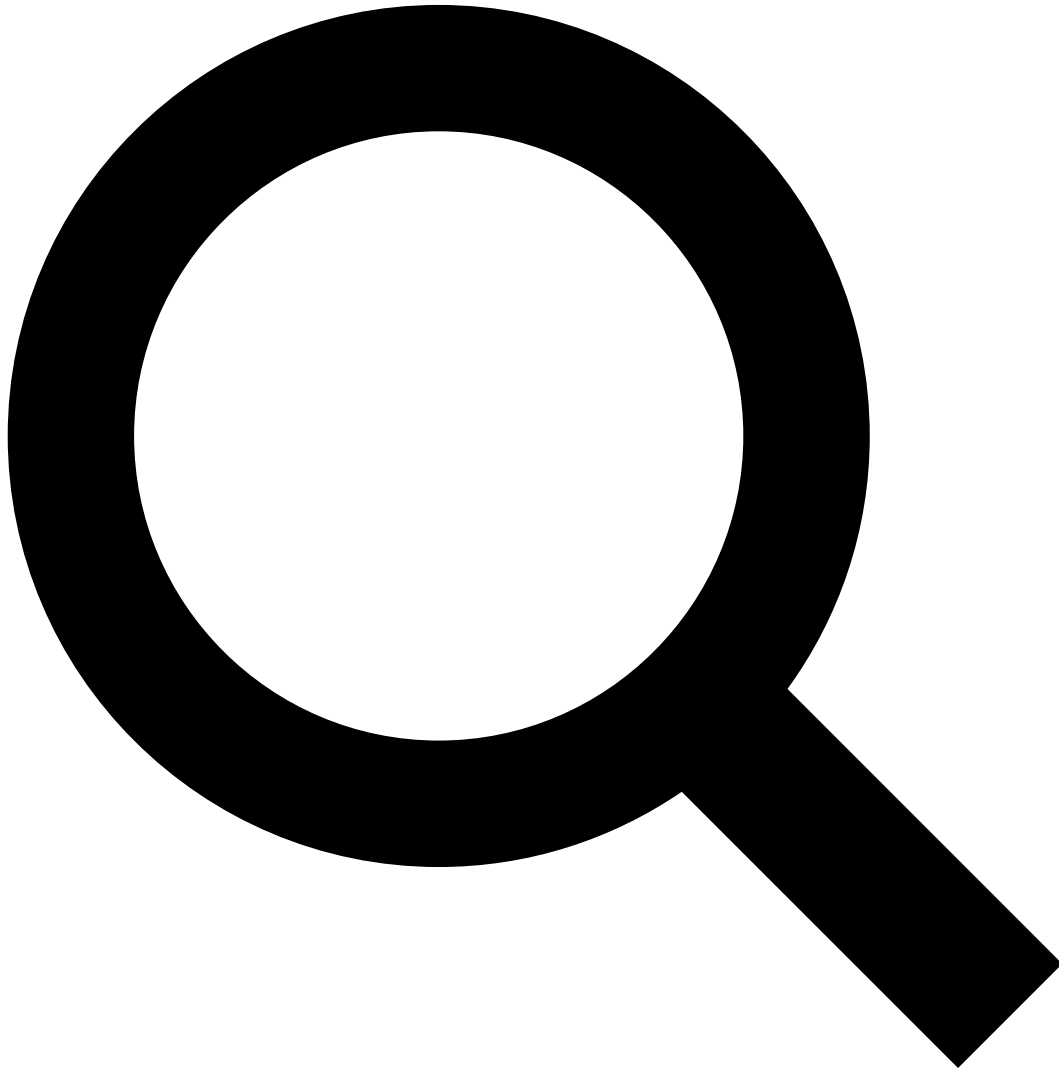
Figure 5

Table 2

	Point forecast	Lo 80	High 80	Low 95	High 95

Sept	9.04	5.978858	11.93987	4.401073	13.51765
Oct	9.77	5.183039	14.1951	2.797671	16.58054
Nov	10.3	5.364191	15.06267	2.797157	17.62971
Dec	10.3	5.280182	15.14668	2.668678	17.75819

Exponential Smoothing method

It is one of the most popular classic forecasting models. It gives more weight to recent values and works best for short term forecasts when there is no trend or seasonality in dataset. The model is given by:

$\hat{Y}_{(t+h|t)} = \square y_{(t)} + \square(1-\square)y_{(t-1)} + \square(1-\square)^2 y_{(t-2)} + \ldots$

with $0<\square<1$

As observed in the model, recent time periods have more weightage in the model and the weightage keeps decreasing exponentially as we go further back in time.

The $\square$ *is the smoothing factor here whose value was chosen to be 0.9 since it had the lowest RMSE among all other values.*

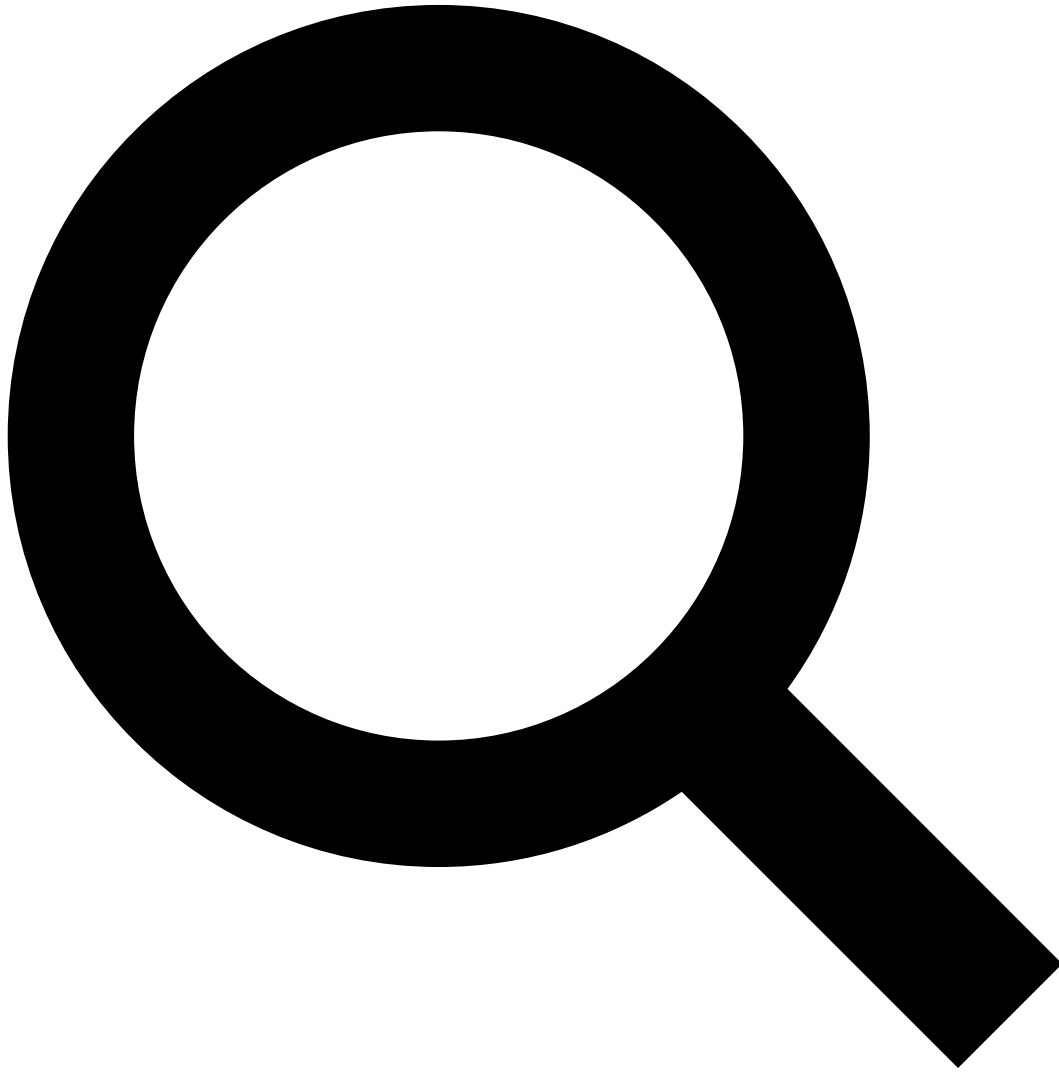The forecast results are presented below:

Figure 6

Table 3

	Point forecast	Lo 80	High 80	Low 95	High 95
Sept	8.30	4.739288	11.87260	2.8512134	13.76068

Oct	8.30	3.507498	13.10439	0.9673541	15.64454
Nov	8.30	2.532806	14.07908	-0.5233096	17.13520
Dec	8.30	1.700403	14.91149	-1.7963595	18.40825

Holt Winters' method

The simple exponential function cannot be used effectively for data with trends. Holt-Winters' exponential smoothing method is a better suited model for data with trends. This model contains a forecast equation and two smoothing equations. The linear model is given by:

$y_{t+h} = l_t + hb_t$

$l_t = \alpha y_t + (1-\alpha)l_{t-1}$

$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$

where, l_t is the level (smoothed value).

h is the number of steps ahead.

b_t is the weighted average of the trend.

Just like the simple exponential smoothing method, l_t shows that it is a weighted average of y_t

The α *is the smoothing factor here whose value was chosen to be 0.99 and the* β *value 0.0025 since they had the lowest RMSE among all other values.*
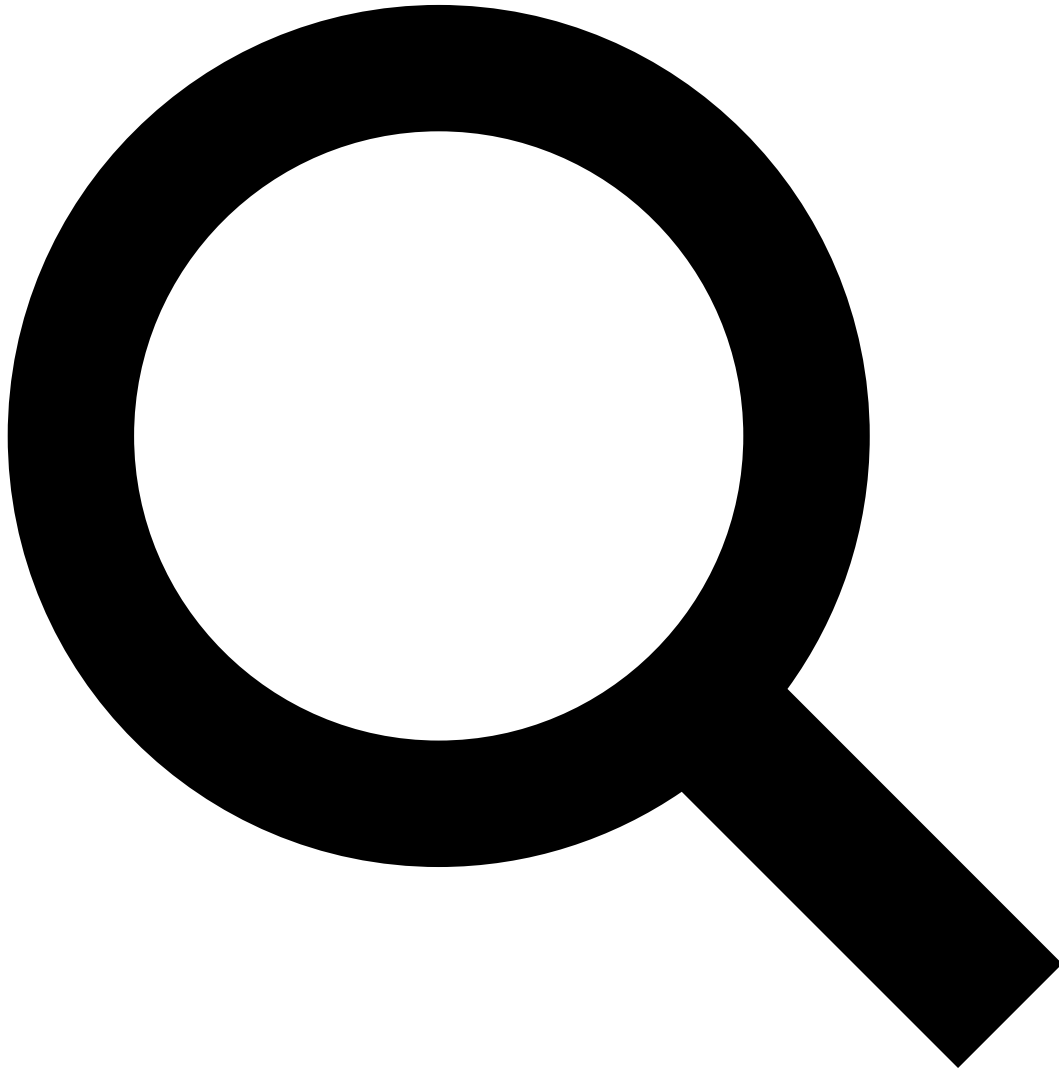
The forecast results are presented below:

Figure 7

Table 4

	Point forecast	Lo 80	High 80	Low 95	High 95

Sept	8.34	4.749288	11.9326	2.84121	13.84
Oct	8.33	3.24	13.4243	0.54541	16.11977
Nov	8.32	2.0800	14.5678	-1.2253	17.87316
Dec	8.31	1.0963	15.53419	-2.725103	19.35565

Evaluation

To compare the models the two parameters chosen are:

- Root mean square error (RMSE)
- Mean absolute error (MAE)

MAE is a measure of mean error in a set of observations/predictions. RMSE is the square root of the mean of squared differences between prediction and actual observation. RMSE is more useful when large errors are not desirable and MAE is useful otherwise.

RMSE and MAE statistics for all the models are presented below:

	Naive	ARIMA	Exp Smoothing	Holt Winters'
RMSE	2.72	2.24	2.73	2.7
MAE	1.05	1.034	1.06	1.05

From the table it is clear that ARIMA/Box Jenkins method has both the lowest RMSE and MAE among the models under consideration while Exponential smoothing method has the highest MAE and RMSE among all.

Therefore, the unemployment rate forecasts as per the Box Jenkins method for the next four months are:

Sept	9.04
Oct	9.77
Nov	10.3
Dec	10.3

The way ahead?

- The unemployment rate is expected to rise in the coming months. This is a bad sign for an economy that is already suffering.
- With GDP forecasts getting lower and lower for the current financial year, the govt needs to act quick to mitigate the potential damage.
- It is impossible to correctly ascertain the total impact of covid-19 on the economy and the range of the impact, but it is safe to say that we will be seeing the effects for a long time to come in some form or other.
- We might see more and more people slip into poverty, depression, increased domestic violence and with potentially long term impact on human development parameters like child mal-nutrition, enrolment rates etc among other things.

Some possible solutions

1. Expansionary monetary policy: It is a common tool of dealing with high unemployment rate in the short term. Under expansionary monetary policy, the central bank reduces the rate of interest on which it lends money to the banks, subsequently the banks lower their rates which leads to a higher amount of loans being taken by business owners. This extra capital helps businesses to hire more workers and expand production, which in turn reduces unemployment rate.
2. Expansionary fiscal policy: Under expansionary fiscal policy the government increases its spending, particularly in the infra-structure

sector. It spends more money to build dams, roads, bridges, highways etc. This increased spending leads to an increase in employment as these projects require labour.

3. <u>Expand the scope of NREGS to urban areas permanently and a higher minimum wage for all :</u> NREGS has proved to be really effective in alleviating poverty, improving quality of life and decreasing unemployment rate in rural areas. Given the unprecedented circumstances, the govt can consider expanding its scope to urban areas, so that it could provide employment to the millions of unemployed workers there. This increase in expenditure could also help the govt revive consumer demand, which is essential if we want to help the GDP get back on track.

4. <u>A stimulus package aimed at putting money into the hands of the poor :</u>

The govt should also consider providing at least a one-time transfer of funds to people just like the US govt did. Such a transfer of putting money directly into the hands of the poor is the most effective way of reviving consumer demand in the economy and many economists around the world have been calling for such a plan to be implemented. There is no better way of increasing consumer expenditure other than putting money into the hands of cash-starved people.
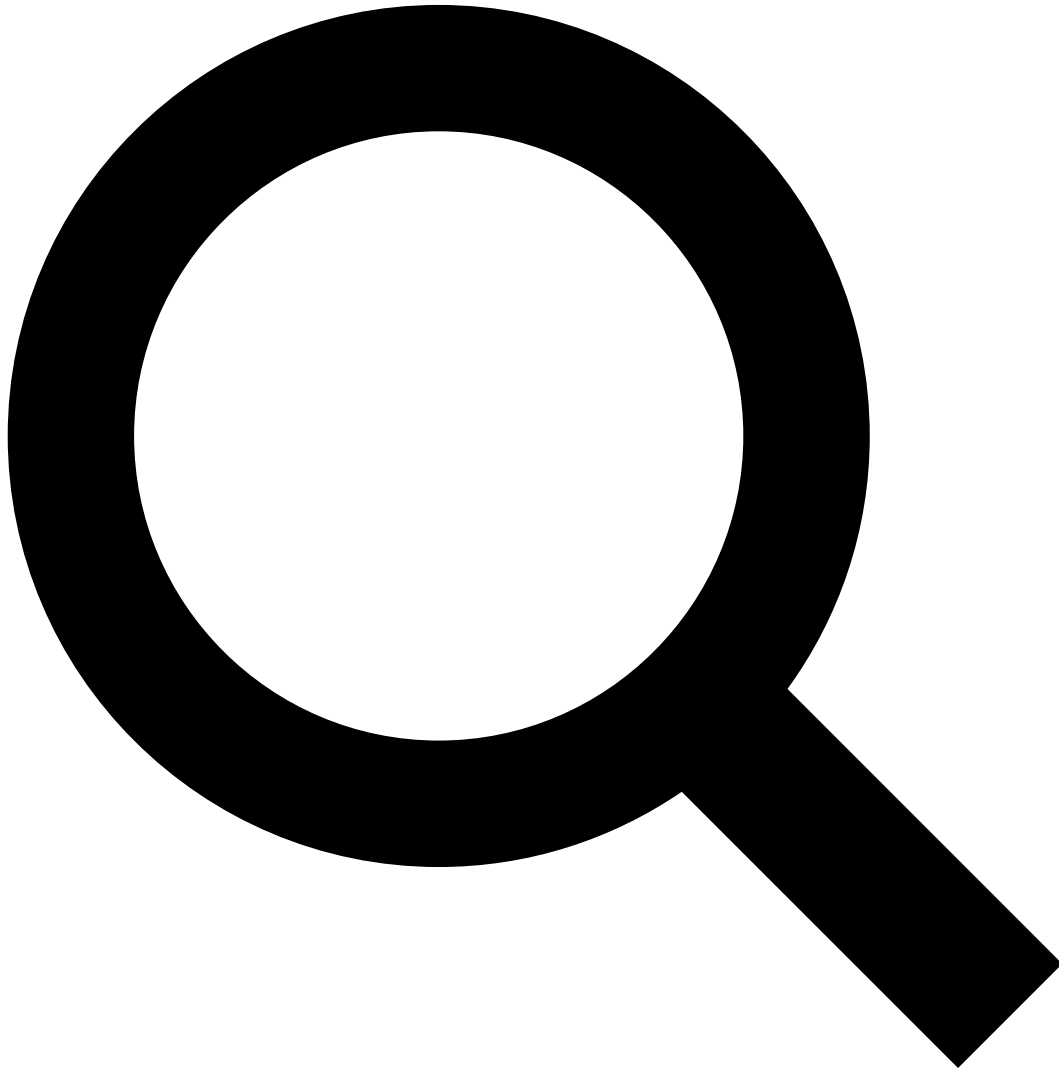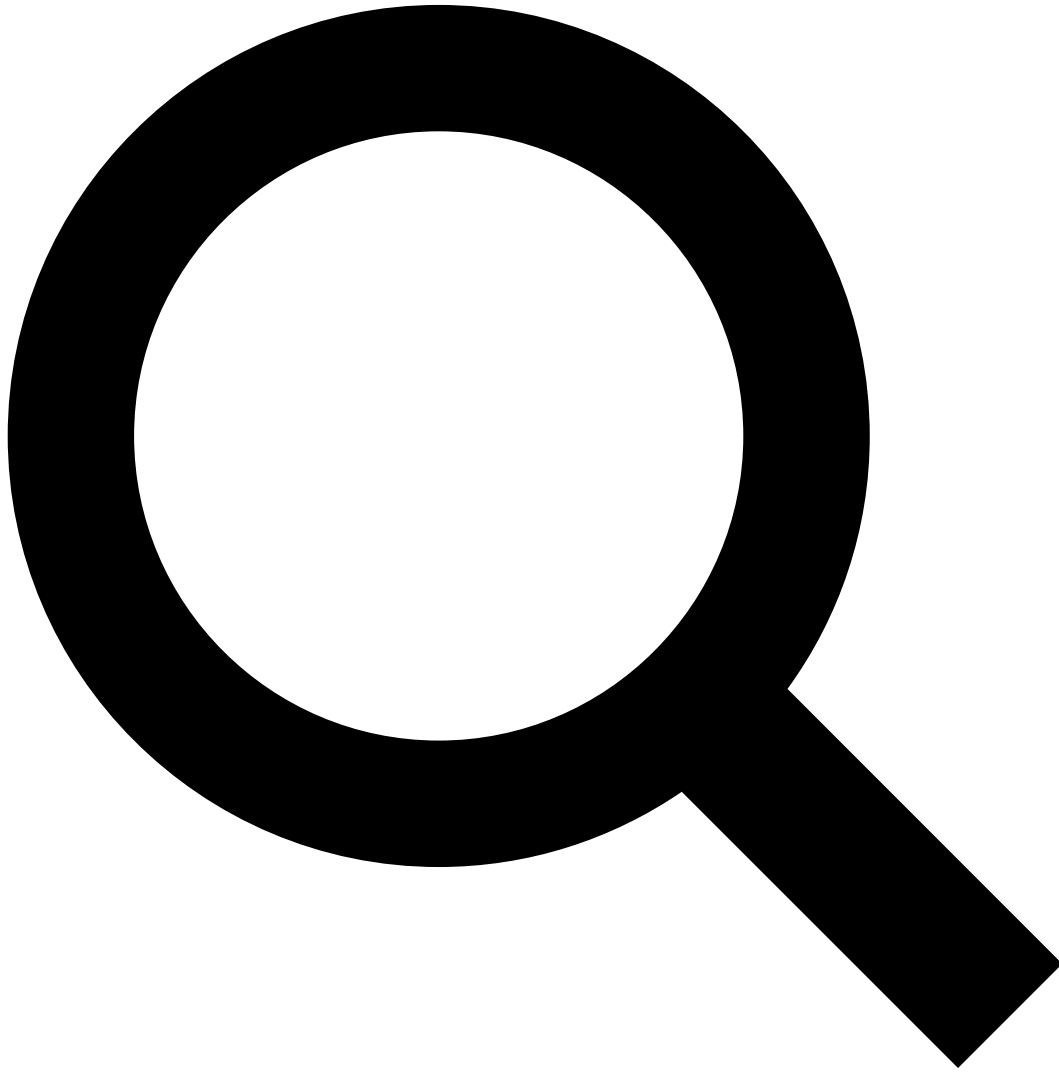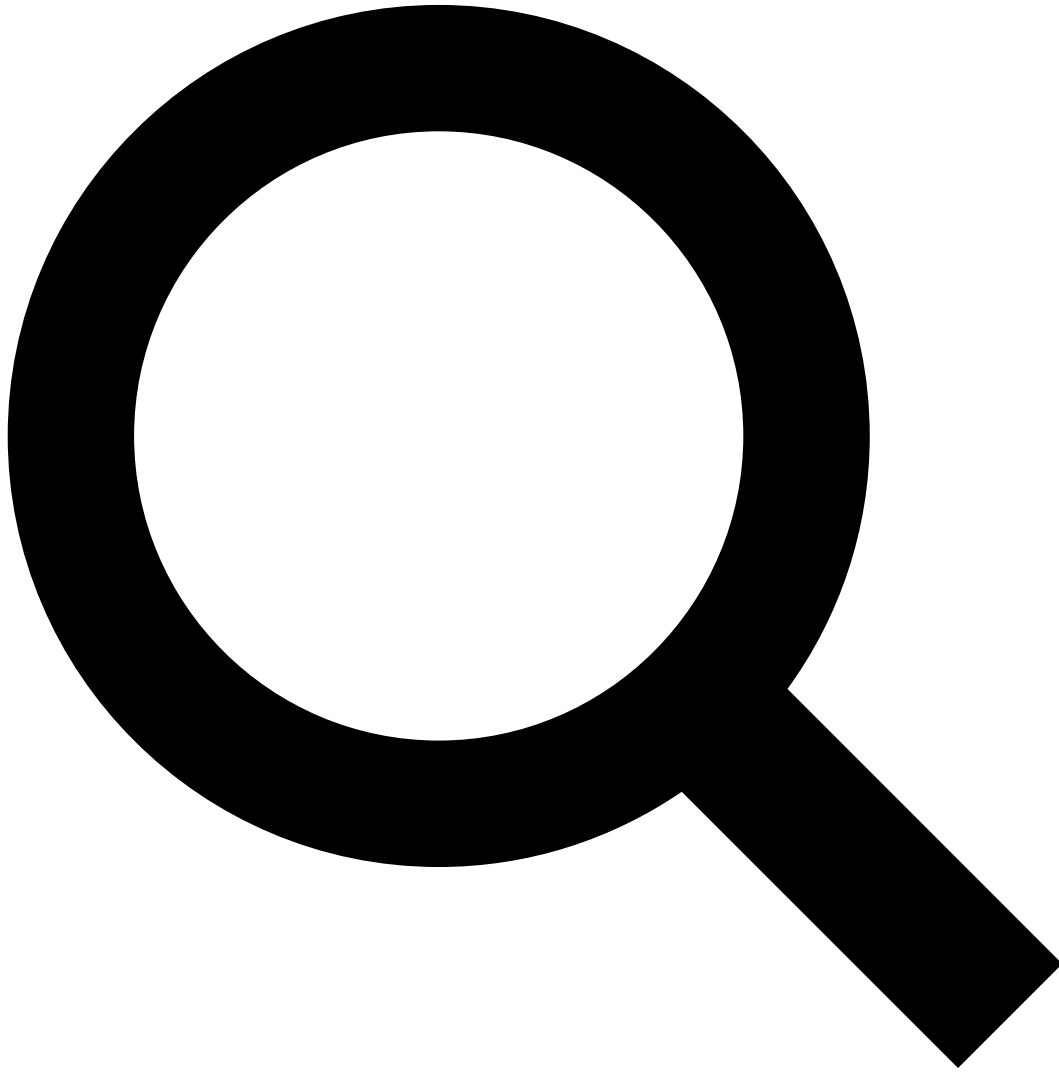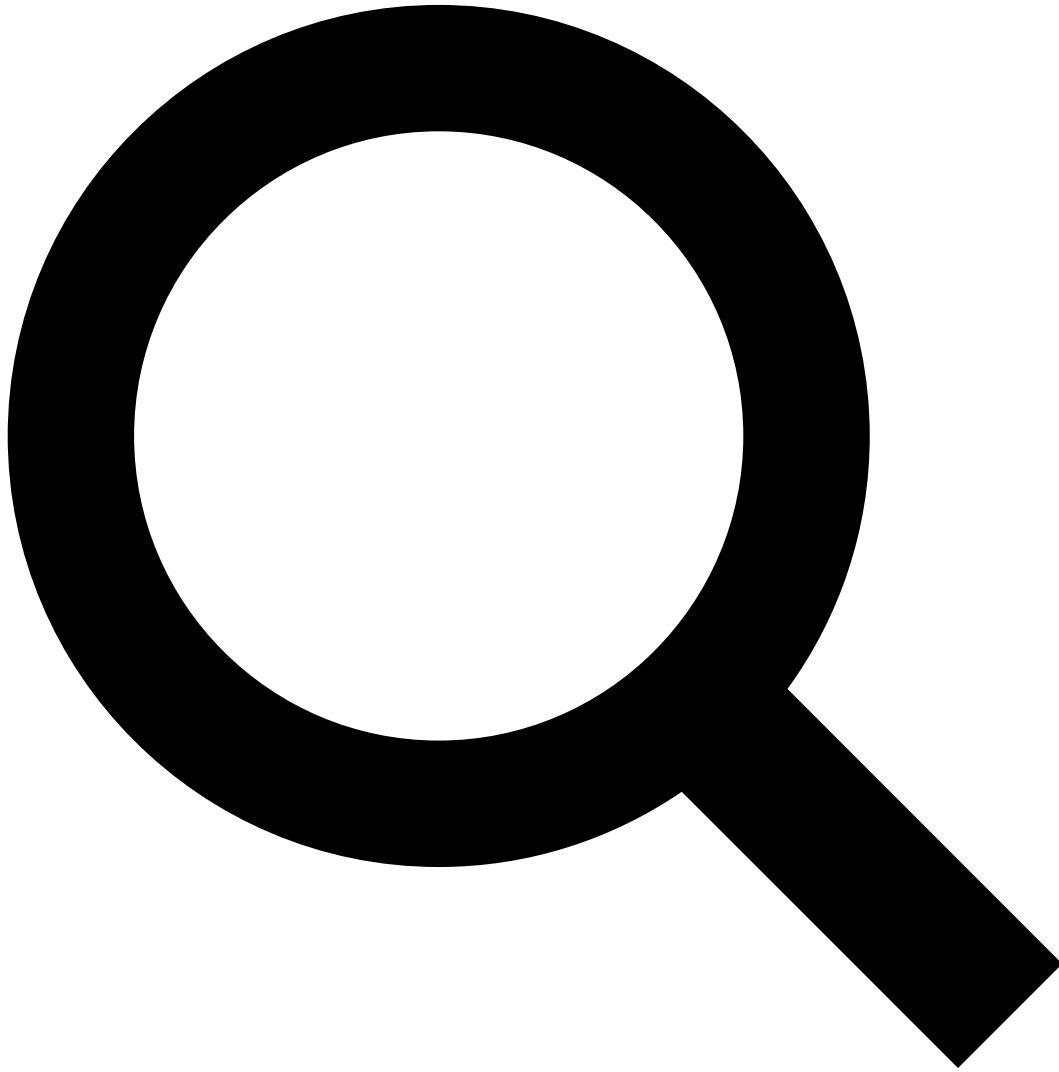
Appendix:

Figure 8

Figure 9

Figure 10

Figure 11

Figure 12

Figure 13