POPULATION PROJECTION ON CENSUS DATA OF ITALY FOR THE YEAR 2021

SOUMITA BANDYOPADHYAY

CONTENTS

LIST OF FIGURES:

LIST OF TABLES

Abstract:

This study attempts to develop a simple mathematical approach to predict the population of Italy for the year 2021, without attempting any justification of particular assumptions about fertility, mortality, and migration. In this present work, we focus on fitting an appropriate model using the past census years (1861-2011) data on the population to infer the future demographic trends of Italy. We use two methods: one, modeling using total population data, two modeling using male and female population data separately, and then combining it to get the total count. Population projection methods enable updating the population data with significant degrees of approximation to reality. The Linear, Exponential, Quadratic, Logistic (Pearl-Reed Logistic) Curves are used for assessing the status of a population. We find the Quadratic model to be the best fit in this case. Results have been neatly presented through graphs and they correspond to real-life census data.

Keywords: Italy, Census, Population projection, Forecast, Model fitting

1. INTRODUCTION

Italy, officially the Italian Republic, is a country consisting of a peninsula delimited by the Alps and several islands surrounding it. Italy is located in south-central Europe. A unitary parliamentary republic, with Rome as its capital and largest city, the country covers a total area of 301,340 km2 (116,350 sq. mi). With around 60 million inhabitants, Italy is the third-most populous member state of the European Union. At the beginning of 2020, Italy had 60,317,116 inhabitants. The resulting population density, at 202 inhabitants per square kilometer (520/sq. mi), is higher than that of most Western European countries. However, the distribution of the population is widely uneven.

A population census is the total process of collecting, compiling, evaluating, analyzing, and publishing or otherwise disseminating demographic, economic, and social data pertaining, at a specified time, to all persons in a country or a well-delimited part of a country. The first census of Italy was taken in 1861. Since then, a census has been taken every 10 years by the Italian National Institute of Statistics (IT ISTAT).

The main objective of this study is to develop a simple mathematical approach to predict the population of Italy for the year 2021 using past census years (1861-2011) data on population, without attempting any justification of particular assumptions about fertility, mortality, and migration. Mathematical curves are used when the growth rate of a population due to births, deaths, and migrations take place under normal situations and it is not subjected to any extraordinary changes like an epidemic, war, earthquake, or any natural disaster. The most popular mathematical curves i.e. Linear, Exponential, Quadratic, Logistic (Pearl-

Reed Logistic) Curves have been used for projecting the past population trend to get the value of population count for a future date. Population projection methods enable updating the population data with significant degrees of approximation to reality. In Section 2, the data has been described and the methodology discussed. Section 3 discusses the results and analysis.

2. DATA ANALYSIS

2.1. DATA DESCRIPTION

Here we have taken the population data of Italy over census years. Usually, Italy conducts a census every 10 years. The first census was carried out in 1861. Here we have accumulated the demographic data of Italy from 1861 to 2011 from the site: https://en.wikipedia.org/wiki/Demographics_of_Italy

We also have used the sex ratio of Italy as of 2020 as an approximate estimate of male and female population count for each of the years. The value used is 94.89 from the site: https://knoema.com/atlas/Italy/topics/Demographics/Population/Male-to-female-ratio

2.2. METHODOLOGY

2.2.1. MODEL SELECTION AND FITTING

From our data, we get the following trend diagram (fig 1). We find an increasing trend in the graph and it seems that the data follows an exponential trend. So, we keep choosing forecasting models from the lowest order(linear, quadratic, etc.) and fit the model to the data.

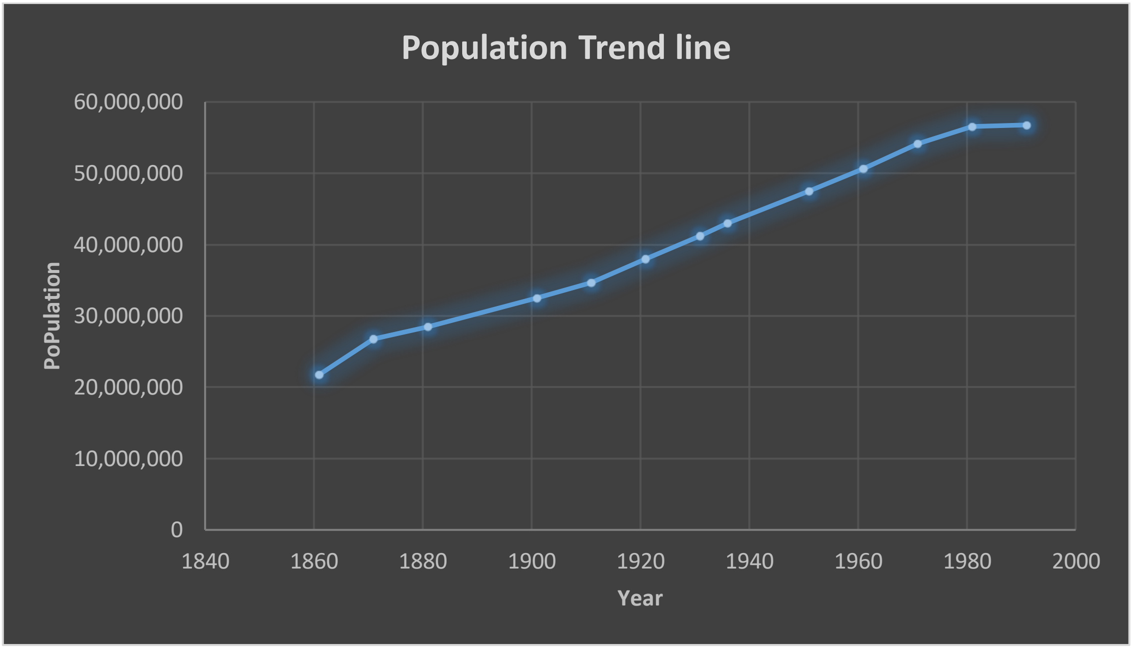

Figure 1: Population trend line obtained on complete census data from 1861-2011

By fitting, we mean estimating the unknown model parameters, usually by the method of least squares. We will also discuss methods for evaluating the quality of the model fit, and determining if any of the underlying assumptions have been violated. This will be useful in discriminating between different candidate models. Here we have selected the four models for further proceeding our research work. Some features of the four models are described below :

a) Linear Model:

The mathematical form of linear trend is

$$y_i = a_1 + b_1 x_i \tag{1}$$

Where y_i is the i^{th} observed value. Value of 'a' is merely the Y-intercept or the height of the line above the origin and 'b' represents the slope of the trend-line.

The estimates of 'a_1' and 'b_1' calculated using least square method from the normal equations:

$$\sum_{i=1}^{n} y_i = na_1 + b_1 \sum_{i=1}^{n} i \tag{2}$$

$$\sum_{i=1}^{n} iy_i = a_1 \sum_{i=1}^{n} i + b_1 \sum_{i=1}^{n} i^2 \tag{3}$$

The linear model assumes that population growth is growing at absolute equal increments per year, decade, or other units of time. It also assumes that growth will follow a similar pattern in future years. Since that is not a logical assumption for real-life population data, we try using the quadratic model.

b) Quadratic Model:

The mathematical form of quadratic trend is

$$y_i = a_2 + b_2 x + c_2 x_i^2 \tag{4}$$

Where y_i is the i^{th} observed value and 'a_2', 'b_2', and, 'c_2' are constants that can be determined by the method of least square.

The estimates of 'a_2', 'b_2', and 'c_2' calculated using least square method from the equations:

$$\sum_{i=1}^{n} y_i = na_2 + b_2 \sum_{i=1}^{n} i + c_2 \sum_{i=1}^{n} i^2 \tag{5}$$

$$\sum_{i=1}^{n} iy_i = a_2 \sum_{i=1}^{n} i + b_2 \sum_{i=1}^{n} i^2 + c_2 \sum_{i=1}^{n} i^3 \tag{6}$$

$$\sum_{i=1}^{n} i^2 y_i = a_2 \sum_{i=1}^{n} i^2 + b_2 \sum_{i=1}^{n} i^3 + c_2 \sum_{i=1}^{n} i^4 \tag{7}$$

By looking at the data plot (fig 1), we do not think that the quadratic model would give a good fit for the data and so we explore more models to find the appropriate one.

c) Exponential Model:

The mathematical form of exponential trend is

$$y_i = a_3 b_3^{x_i} \tag{8}$$

Where y_i is the i^{th} observed value and 'a_3', and 'b_3' are constants that can be determined by the method of least square.

The estimates of 'a_3', and 'b_3' calculated using least square equations:

$$\sum_{i=1}^{n} Y_i = nA + B \sum_{i=1}^{n} i \tag{9}$$

$$\sum_{i=1}^{n} iY_i = A \sum_{i=1}^{n} i + B \sum_{i=1}^{n} i^2 \tag{10}$$

Where $A = \log a_3$ and $B = \log b_3$, $Y_i = log y_i$

The exponential growth is not very realistic in the sense that it does not consider the environmental limits that have consequences on the population, according to this model; a population can grow with no limits.

In logistic growth, population expansion decreases as resources become scarce, leveling off when the carrying capacity of the environment is reached, resulting in an S-shaped curve and hence proves to be a more realistic model of population growth than exponential growth. Hence, we consider the Pearl reed model.

d) Pearl- Reed Logistic Model:

The mathematical form logistic (Pearl – Reed Model) trend is

$$P_t = (L, r, \beta) = \frac{L}{1+e^{r(\beta-t)}} \tag{11}$$

Here the population size P is regarded as a function of L, r and β.Here the estimates of L, r and β can be obtained from the method of selective points.

Let us define,

$$d_1 = \frac{1}{P_0} - \frac{1}{P_1} = \frac{e^{r\beta}(1-e^{-rn})}{L} \tag{12}$$

$$d_2 = \frac{1}{P_n} - \frac{1}{P_{2n}} = \frac{e^{r(\beta-n)}(1-e^{-rn})}{L} \tag{13}$$

Now, the estimates are,

$$r = \frac{lnd_1 - lnd_2}{n} ; L = \frac{(d_1-d_2)}{{d_1}^2} ; \beta = \frac{\ln(\frac{L-P_0}{P_0})}{r} \tag{14}$$

<u>***2.2.2. MODEL VALIDATION***</u>

Model validation is the set of processes and activities intended to verify that models are performing as expected. Model validation can be based on two types of data: data that was used in the construction of the model and data that was not used in the construction. Validation based on the first type usually involves analyzing the goodness of fit of the model or analyzing whether the residuals seem to be random (i.e. residual diagnostics). Validation based on the second type usually involves analyzing whether the model's predictive performance deteriorates non-negligibly when applied to pertinent new data. The purpose of model validation is to check the accuracy and performance of the model basis on the past data for which we already have actuals.

Here we have adopted three accuracy measures to validate the model of population projection:

i) Mean Absolute Percentage Error:

The mean absolute percent error (MAPE) expresses accuracy as a percentage of the error.

$$\text{MAPE}=\frac{100\sum_{t=1}^{n}\left|\frac{(A_t-F_t)}{A_t}\right|}{n}$$

Here A_t is the actual value and F_t is the forecast value. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n.

ii) Mean Absolute deviation:

The mean absolute deviation (MAD) expresses accuracy in the same units as the data, which helps conceptualize the amount of error. This approach is useful when the size or size of a prediction variable is significant in evaluating the accuracy of a prediction. MAPE indicates how much error in predicting compared with the real value.

$$\text{MAD}=\frac{\sum_{t=1}^{n}|A_t-F_t|}{n}$$

Here A_t is the actual value and F_t is the forecast value. Here MAD is calculated over 'n' time points.

iii) Mean Square Deviation: The mean square deviation (MSD) measures the accuracy of the fitted time series values. Outliers have a greater effect on MSD than on MAD.

$$\text{MSD}=\frac{\sum_{t=1}^{n}(A_t-F_t)^2}{n}$$

Here A_t is the actual value and F_t is the forecast value. Here MSD is calculated over 'n' time points.

All three accuracy measures are based on one-period-ahead residuals. At each point in time, the model is used to predict the Y value for the next period in time. They are used to compare the fits of different time series models. Smaller values indicate a better fit. If a single model does not have the lowest values for all three accuracy measures, MAPE is usually the preferred measurement. In general, outliers have a greater effect on MSD than on MAD.

2.2.3. FORECASTING

After model validation, we can find out the appropriate model that will best fit for our data. Then we will use the model to generate the future forecasts of any time point.

3. ANALYSIS AND RESULTS:

3.1. SOFTWARES USED

To carry out the process of model selection and fitting, the software used are:

- Minitab 17
- MS-Excel 2016

3.2. RESULTS

The results obtained under each model are presented below:

- ***METHOD I:***

In this method, we will focus on population modeling of Italy and fitting an appropriate model using the past census years (1861-2011) data on population so that we can infer about the future demographic trends of Italy.

At first, we will consider the population of 13 census years from 1861-1991, and fit four different mathematical models for model validation using Minitab software. After that, we will compare the accuracy measures such as MAPE, MAD, and MSD of four models to find out the best fit for the entire population. Then we will fit the best-fitted model obtained by model validation over the 15 observations i.e. from 1861-2011 of the total population to predict the future population of Italy in 2021.

The model validation of the total population data of 13 census years has shown below-------

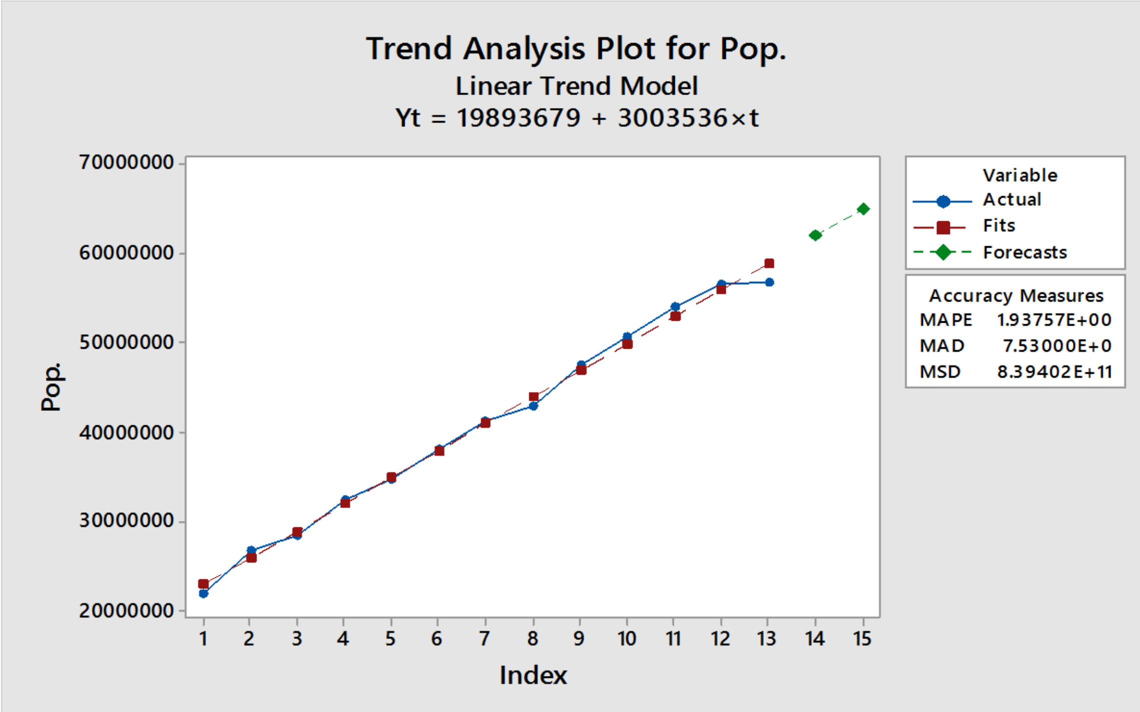

Figure 2: Linear trend model for population from 1861-1991 and estimates of 2001 and 2011

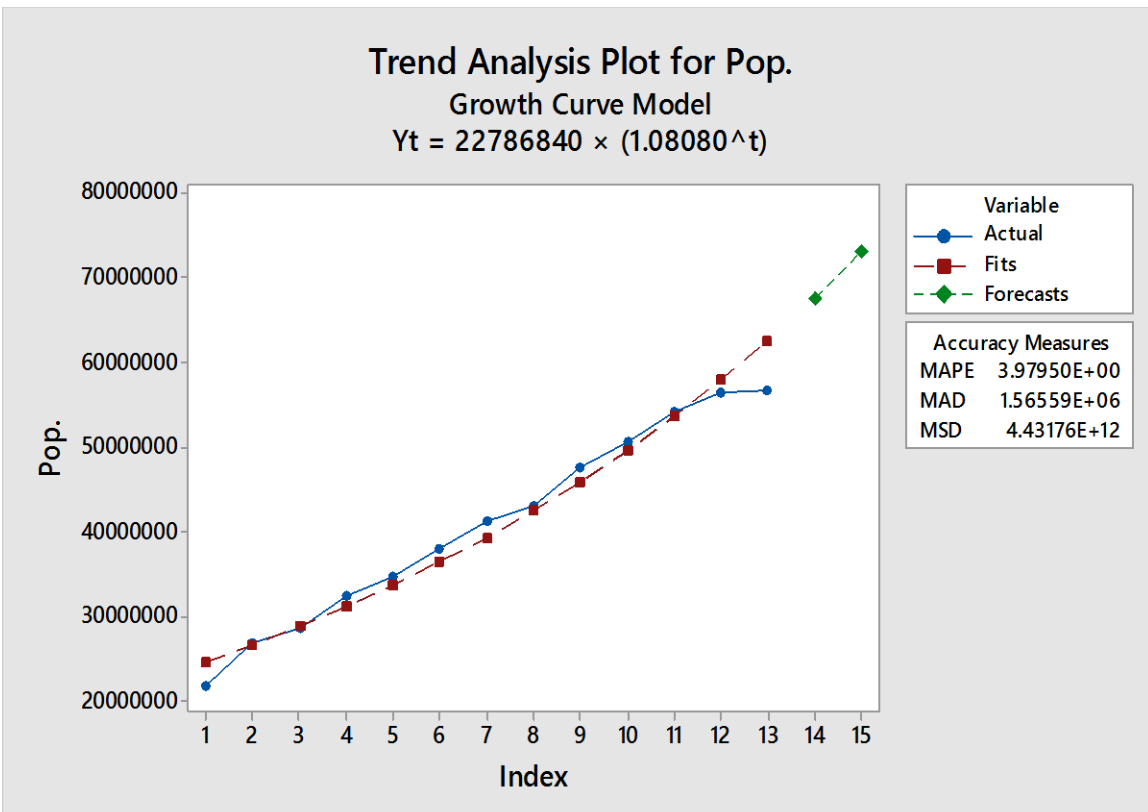

Figure 3: Exponential model for population from 1861-1991 and estimates of 2001 and 2011

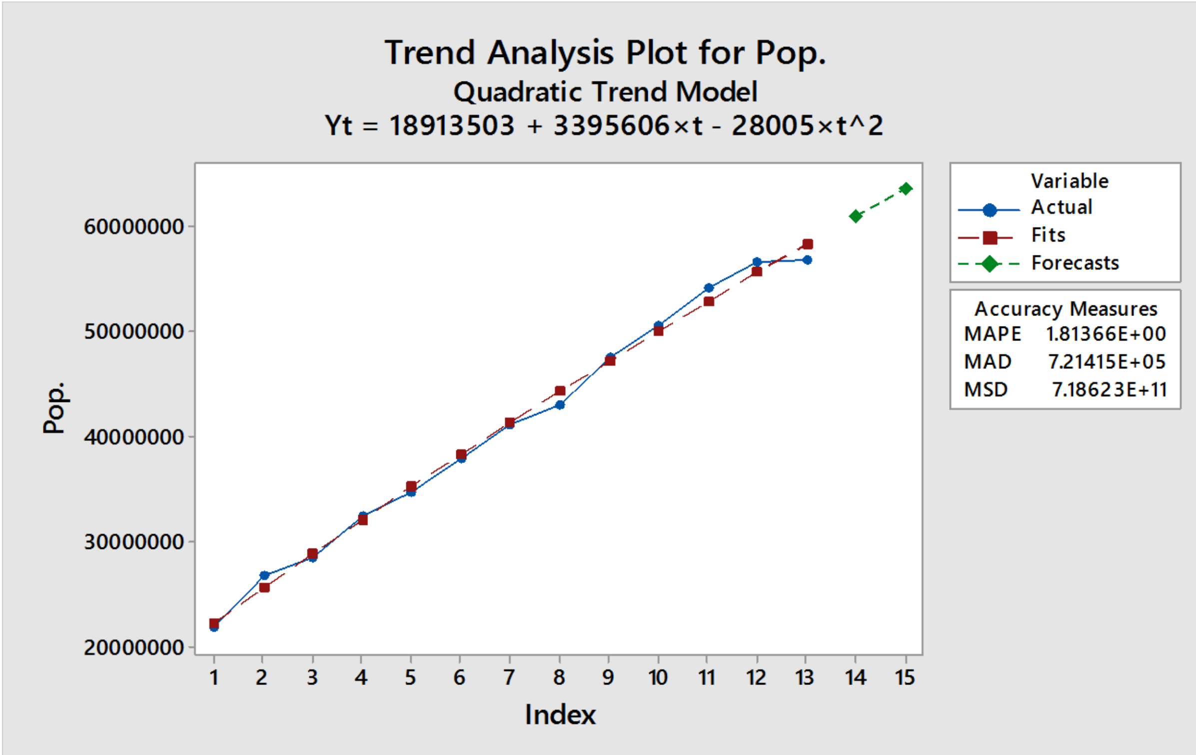

Figure 4: Quadratic trend model for population from 1861-1991 and estimates of 2001 and 2011

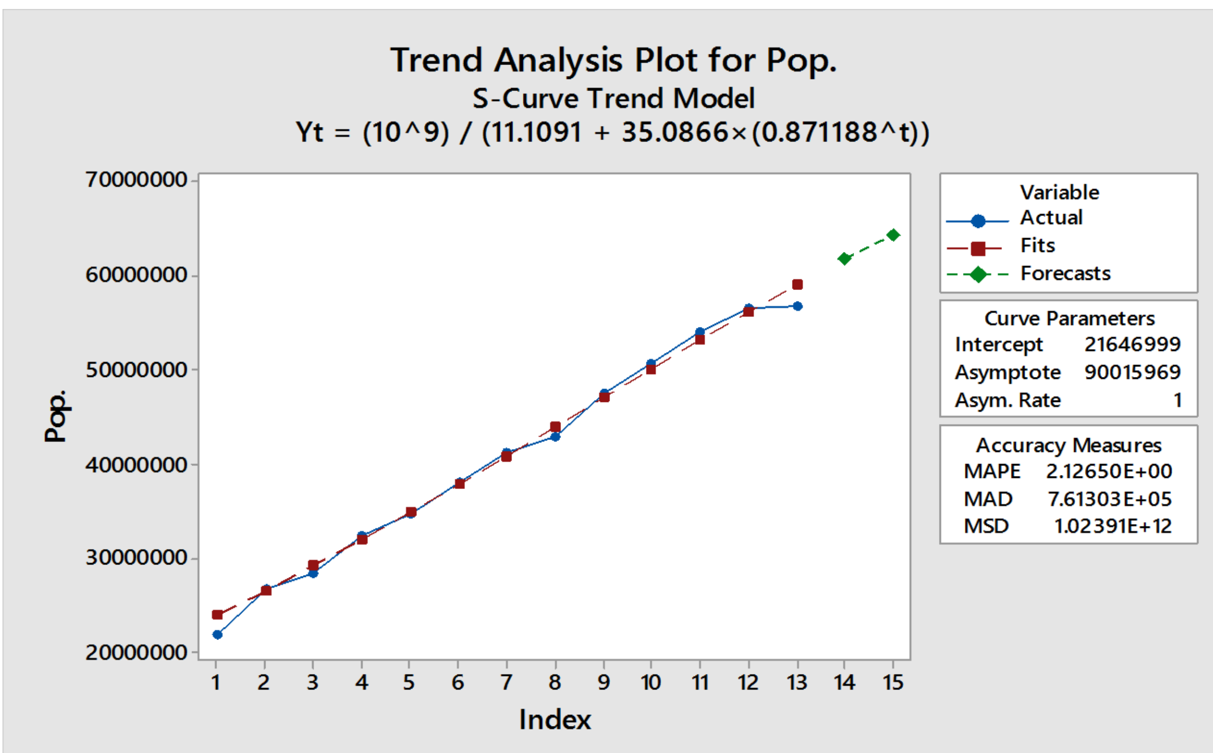

Figure 5: S-curve trend model for population from 1861-1991 and estimates of 2001 and 2011

Table 1: Accuracy measures on estimating the population data for years 2001 and 2011 by different models

ACCURACY MEASURES / MODEL NAMES	MAPE	MAD	MSD
LINEAR	1.93757	753000	839402000000
QUADRATIC	1.81366	721415	718623000000
EXPONENTIAL	3.9795	1565590	4431760000000
PEARL-REED LOGISTIC(S-CURVE)	2.1265	761303	1023910000000

From the table (table 1) it has been observed that the Quadratic Model has the least value for all the accuracy measures i.e. MAPE, MAD, and MSD. Then from model validation of the entire population, we find out that the Quadratic Model is the best-fitted (or suitable) model for our dataset.

- Now, we fit the Quadratic model over the 15 census years (1861-2011) data of total population to predict the value of total population for any future time points. To get a better projection, we use the data of recent two years i.e. 2001 and 2011 to capture the recent year's trend in the fitted model of the entire population.

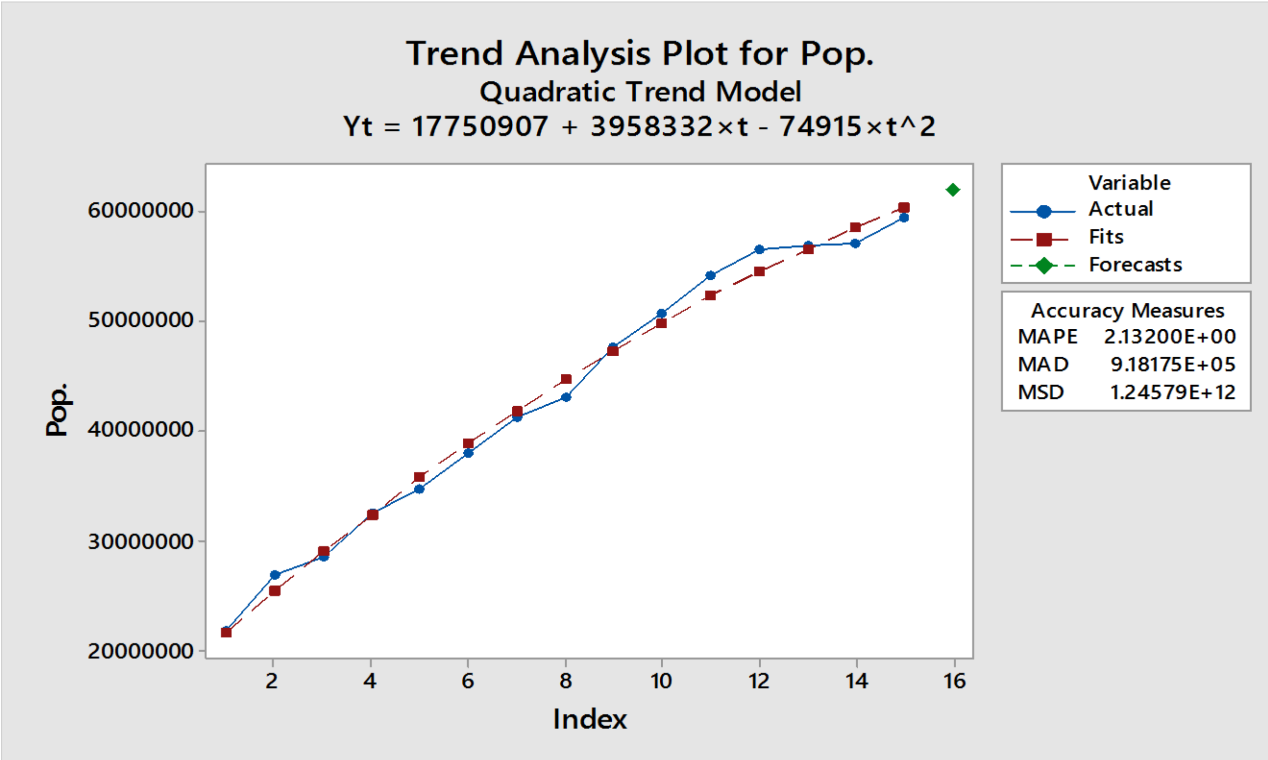

Figure 6: Quadratic trend model for population from 1861-2011 and estimate of 2021

- So, our estimated population census value for the year 2021 is 61,906,028.

• *METHOD II:*

Projection Of The Male population :

Now we project the Male Population of Italy separately. For that, we will take the data of census years of the male population by omitting the last 2 census years i.e. 2001 and 2011. Then we will validate four mathematical models to find out the best-fitted model for the male population by comparing three accuracy measures i.e. MAPE, MAD, and MSD. The model validation of the male population has been done below----

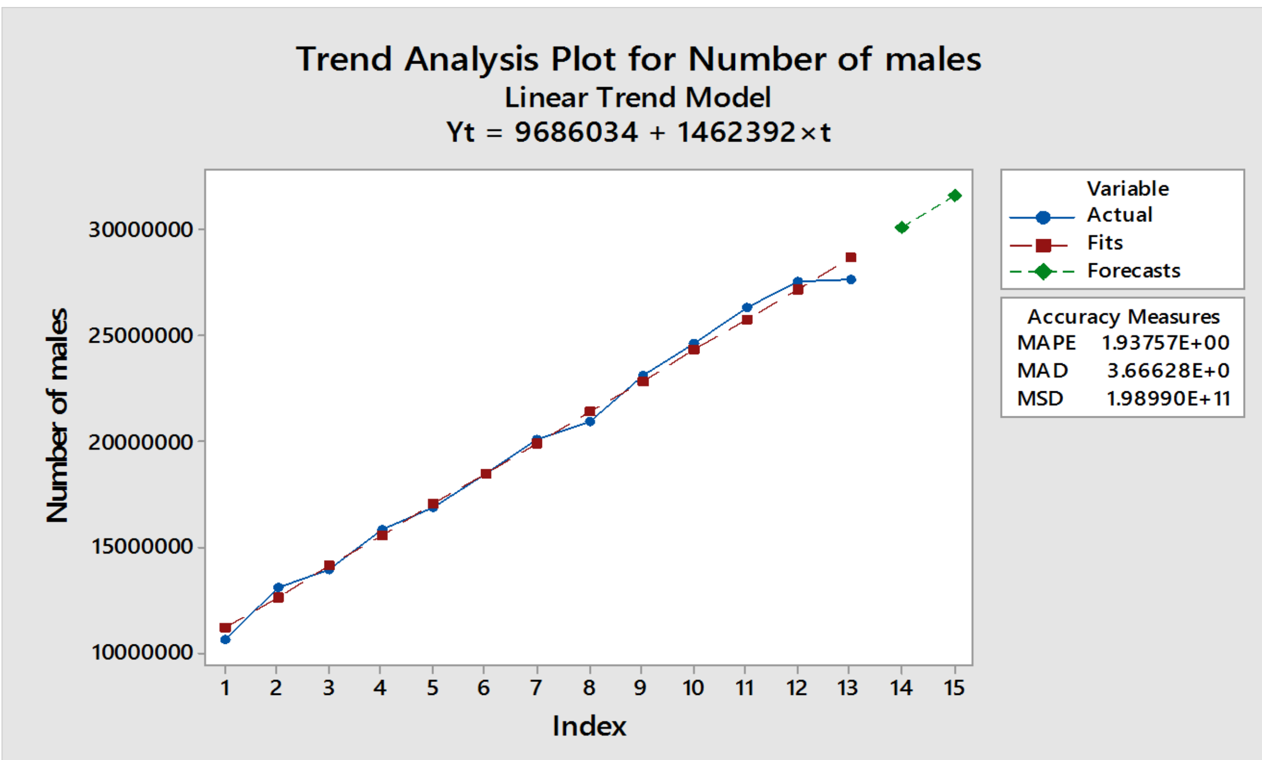

Figure 7: Linear trend model for male population from 1861-1991 and estimates of 2001 and 2011

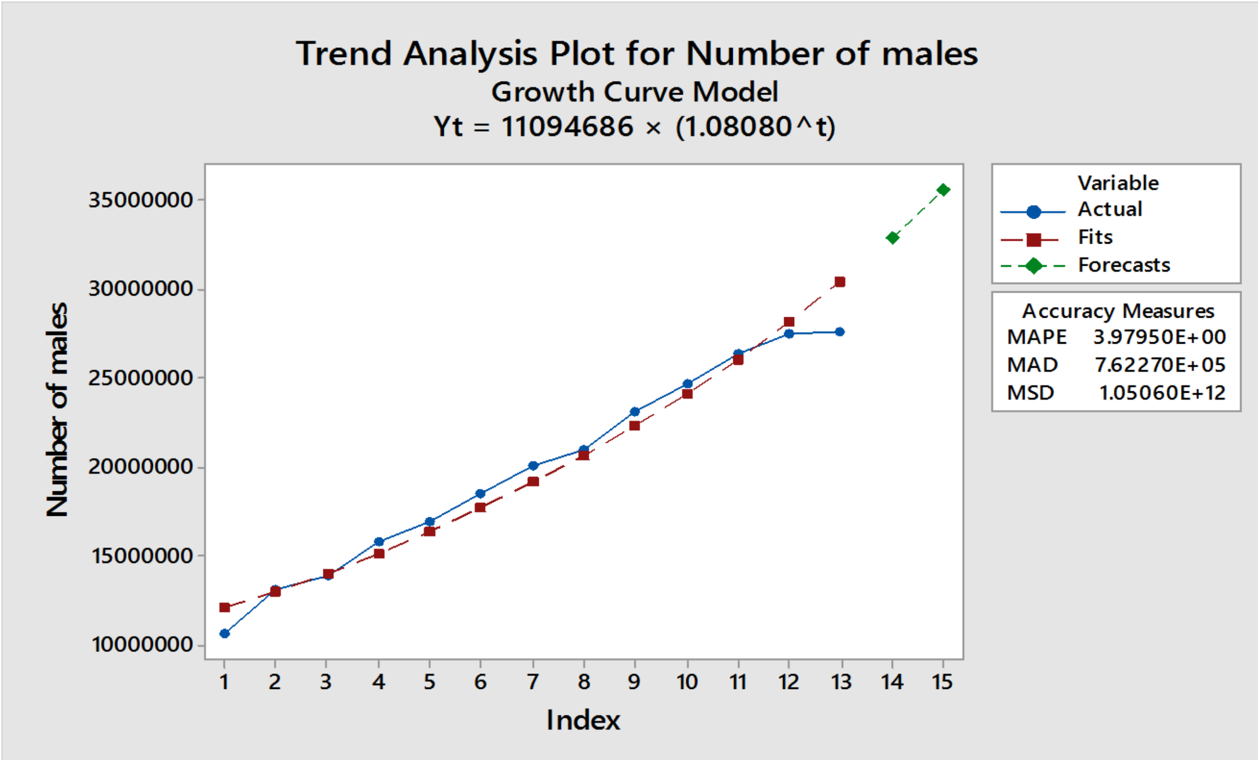

Figure 8: Exponential model for male population from 1861-1991 and estimates of 2001 and 2011

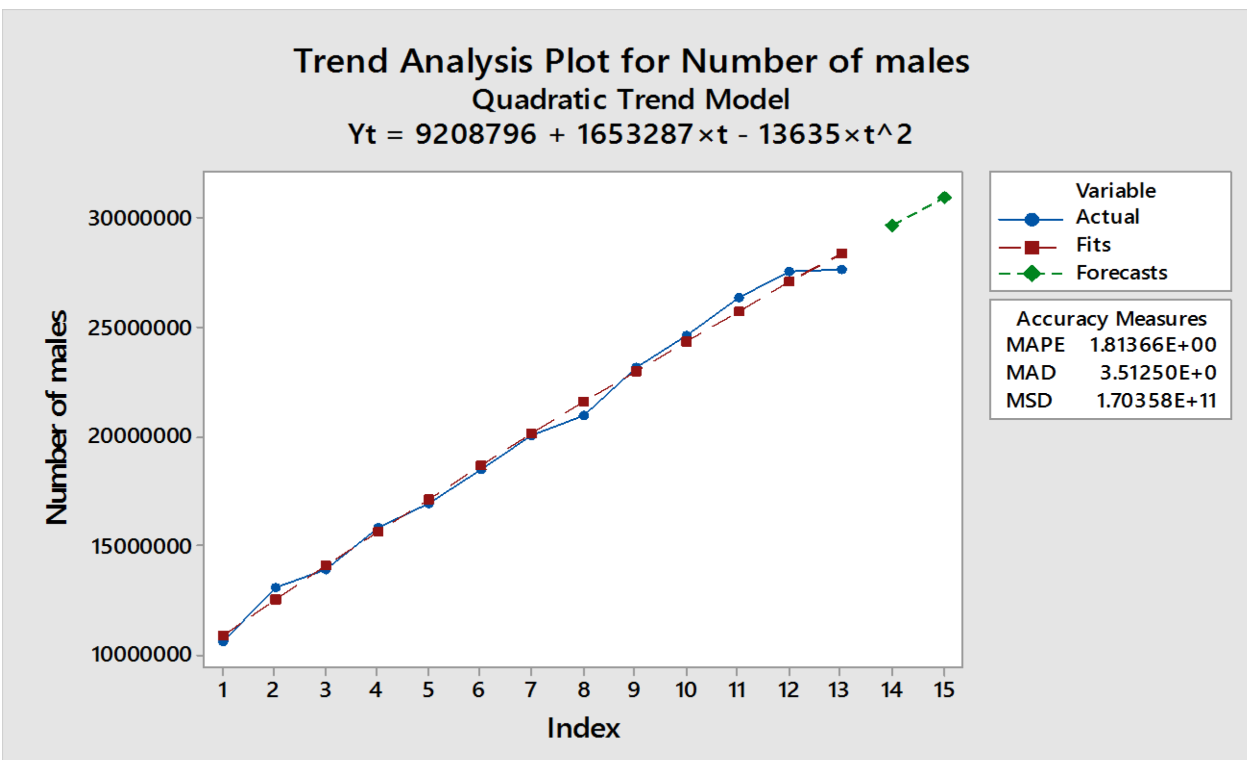

Figure 9: Quadratic trend model for male population from 1861-1991 and estimates of 2001 and 2011

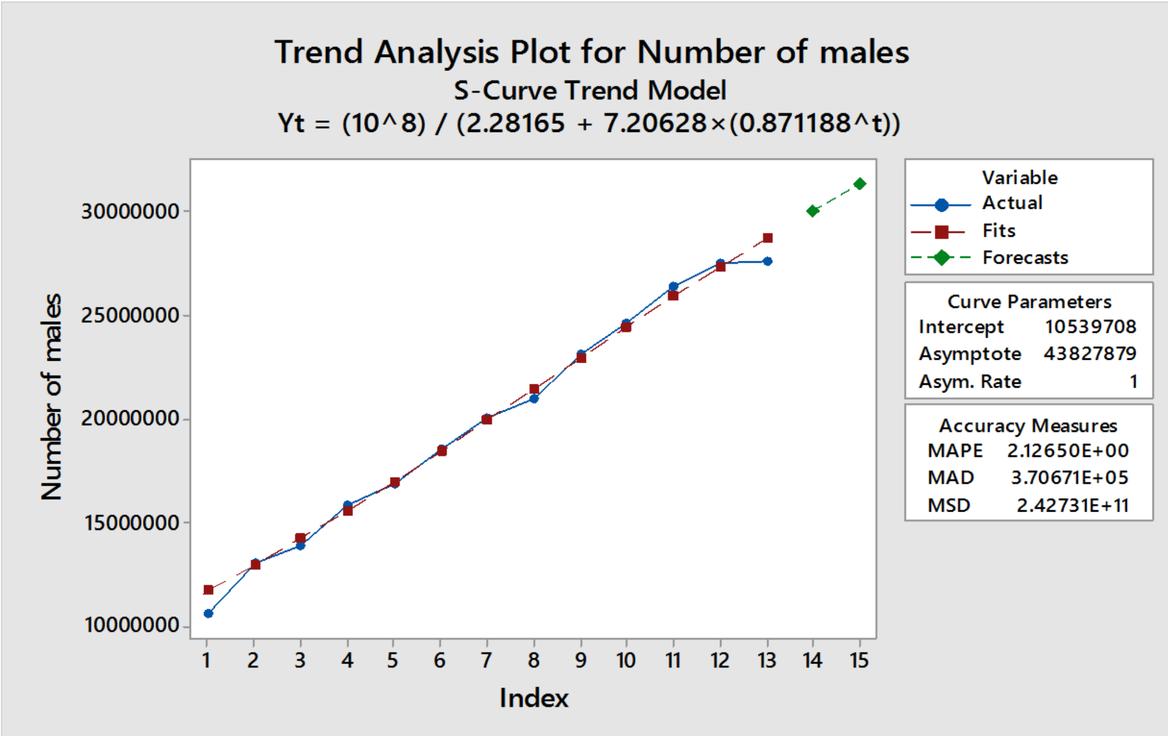

Figure 10: S-curve trend model for male population from 1861-1991 and estimates of 2001 and 2011

Table 2: Accuracy measures on estimating the male population data for years 2001 and 2011 by different models

ACCURACY MEASURES / MODEL NAMES	MAPE	MAD	MSD
LINEAR	1.93757	366628	198990000000
EXPONENTIAL	3.9795	762270	1050600000000
QUADRATIC	1.81366	351250	170358000000
S-CURVE	2.1265	370671	242731000000

From the table (table 2), it has been observed that the Quadratic Model has the least value for all the accuracy measures i.e. MAPE, MAD, and MSD. Then from model validation of the male population, we find out that the Quadratic Model is the best fitted (or suitable) model for our dataset.

- Now, we fit the Quadratic model over the 15 census years (1861-2011) data of the male population to predict the value of the male population for any future time points. To get a better projection, we use the data of recent two years i.e. 2001 and 2011 to capture the recent year's trend in the fitted model of the male population.

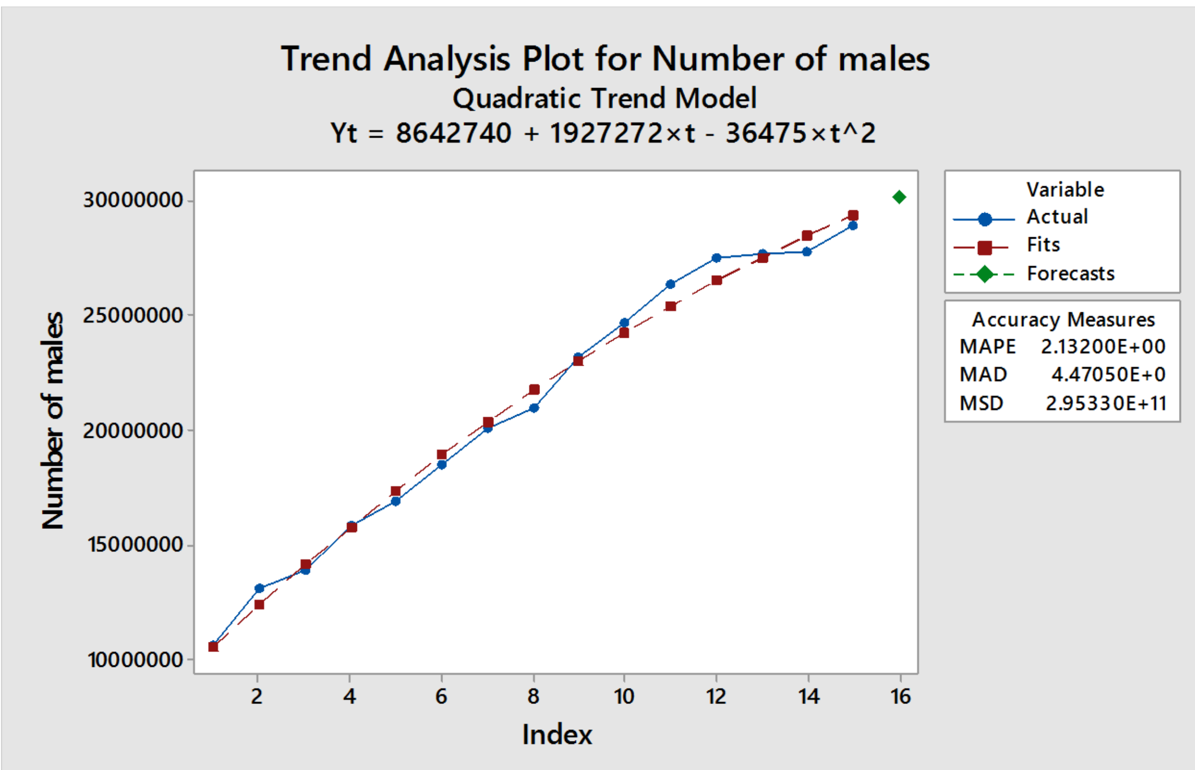

Figure 11: Quadratic trend model for male population from 1861-2011 and estimate of 2021

- So, our estimated male population census value for the year 2021 is 3,01,41,429.

Projection Of The Female population :

Next we project the Female Population of Italy separately. For that, we will take the data of census years of the female population by omitting the last 2 census years i.e. 2001 and 2011. Then we will validate four mathematical models to find out the best-fitted model for the female population by comparing three accuracy measures i.e. MAPE, MAD, and MSD. The model validation of the female population has been done below----

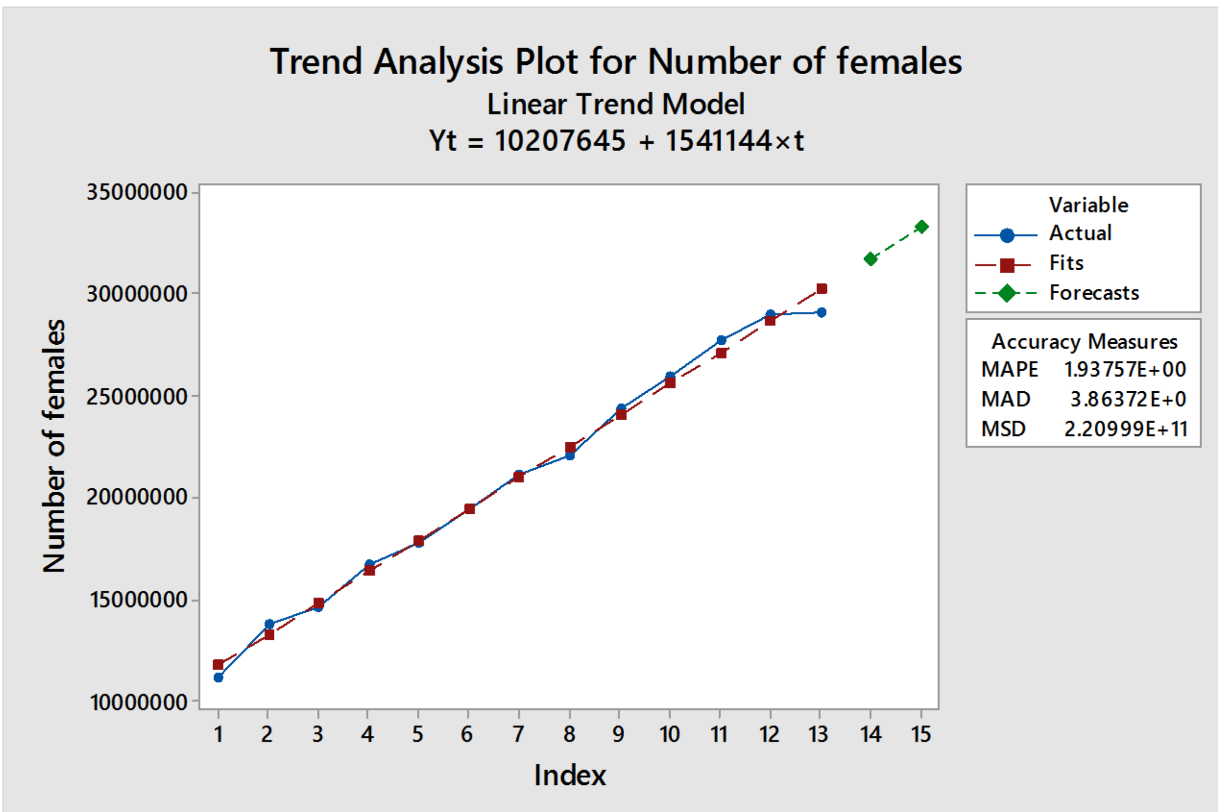

Figure 12: Linear trend model for female population from 1861-1991 and estimates of 2001 and 2011

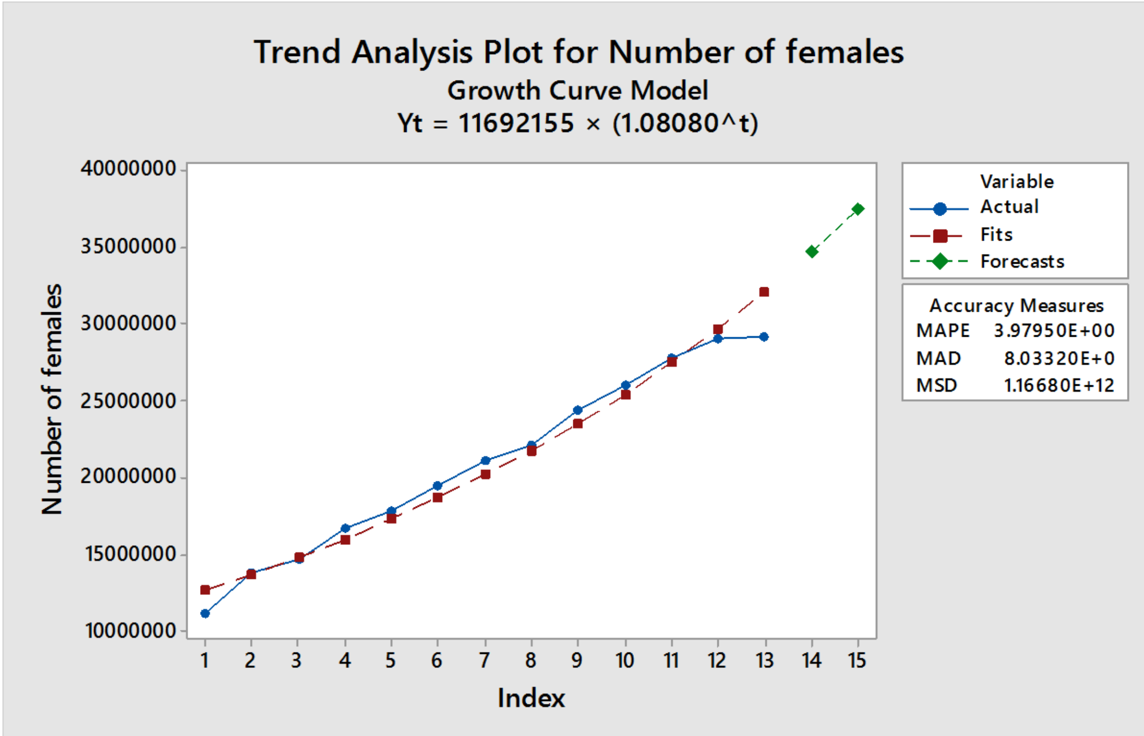

Figure 13: Exponential model for female population from 1861-1991 and estimates of 2001 and 2011

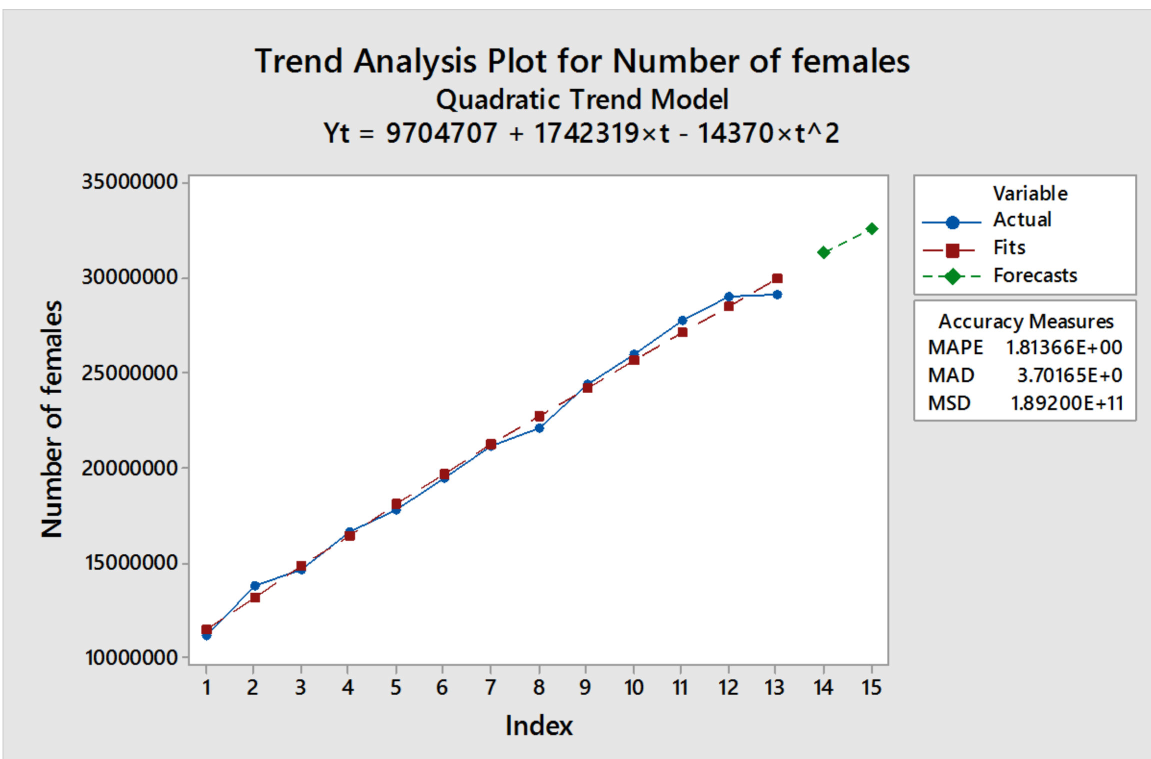

Figure 14: Quadratic trend model for female population from 1861-1991 and estimates of 2001 and 2011

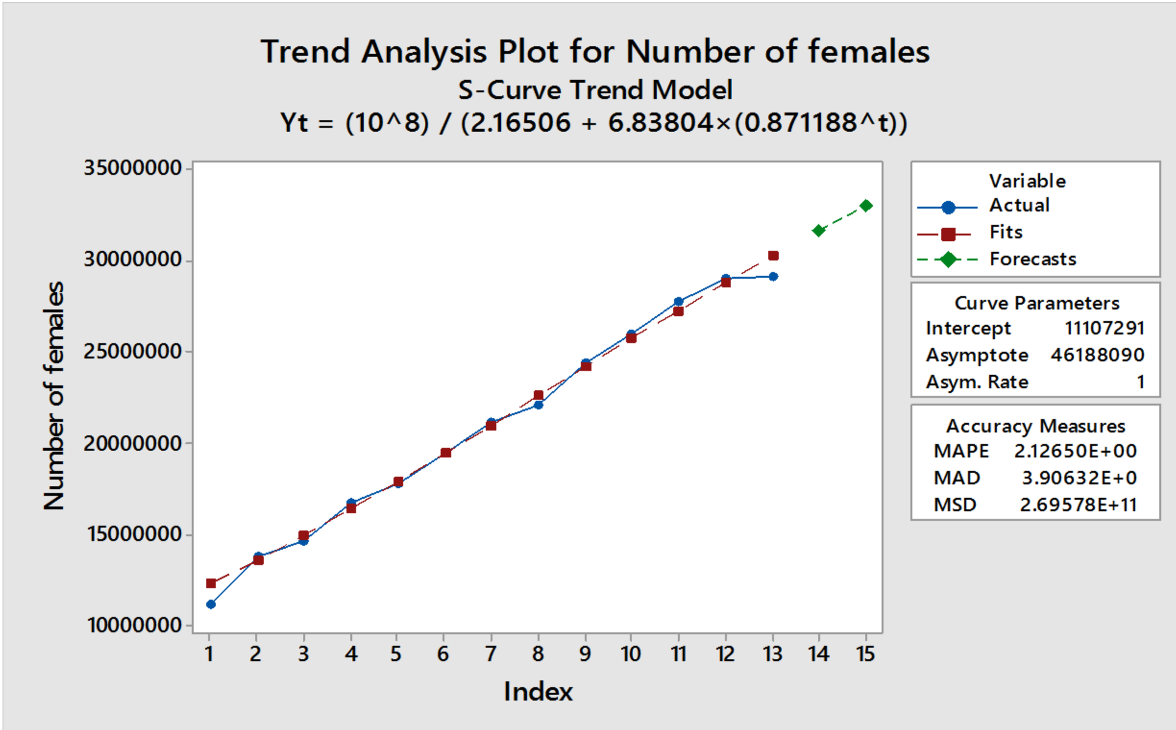

Figure 15: S-curve trend model for female population from 1861-1991 and estimates of 2001 and 2011

Table 3: Accuracy measures on estimating the female population data for years 2001 and 2011 by different models

ACCURACY MEASURES / MODEL NAMES	MAPE	MAD	MSD
LINEAR	1.93757	386372	220999000000
EXPONENTIAL	3.9795	803320	1166800000000
QUADRATIC	1.81366	370165	189200000000
S-CURVE	2.1265	390632	269578000000

From the table (table 3), it has been observed that the Quadratic Model has the least value for all the accuracy measures i.e. MAPE, MAD, and MSD. Then from model validation of the female population, we find out that the Quadratic Model is the best-fitted (or suitable) model for our dataset.

- Now, we fit the Quadratic model over the 15 census years (1861-2011) data of the female population to predict the value of the female population for any future time points. To get a better projection, we use the data of recent two years i.e. 2001 and 2011 to capture the recent year's trend in the fitted model of the female population.

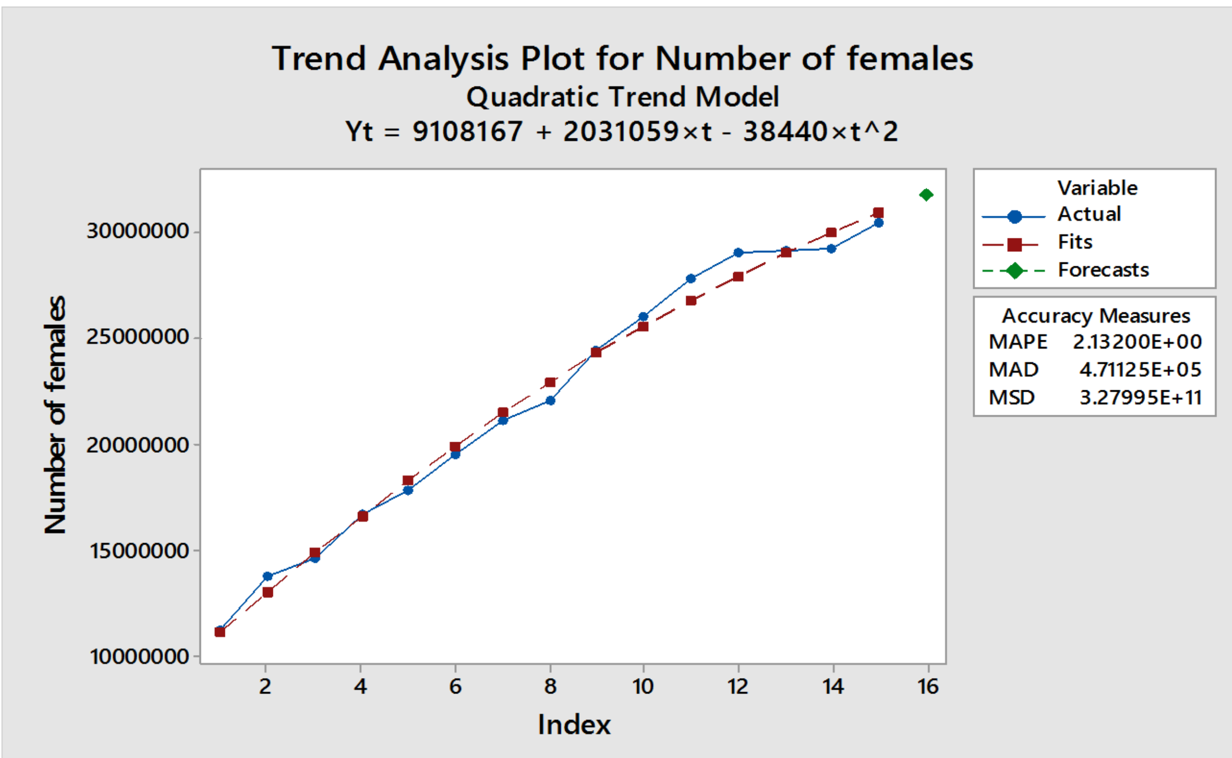

Figure 16: Quadratic trend model for female population from 1861-2011 and estimate of 2021

- Estimated female population census value for the year 2021 is 3,17,64,600

Using the predicted values of male and female population for the year 2021, we get the Total census value by adding up those values as :

Estimated population total= Male population +female population

=3,01,41,429+3,17,64,600=61,906,029

4.CONCLUSION

So we now plot our forecasted value of population for the year 2021 together with the past census values (1861-2011) and observe the following figure (fig 17)

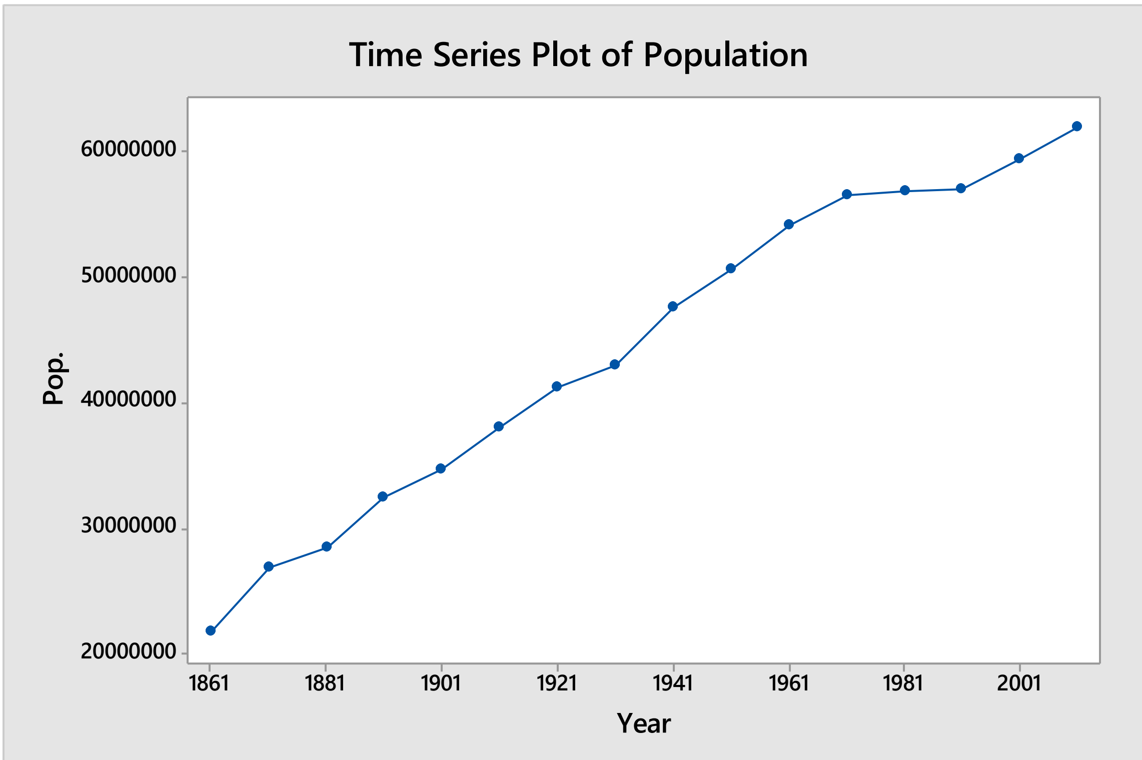

Figure 17: Plot of population data of Italy through the years

From the time series plot, it has been observed that there is a similar kind of sudden downfall of the population between 1871-1891, 1921-1941 and 1981-2001. After 1931, an increasing trend was observed in the population of Italy till 1971. From our study, we suspect that the population of Italy will increase in 2021, reaching an approximate value of 61,906,029.

5. FUTURE WORK

We find, from the source, https://en.wikipedia.org/wiki/Demographics_of_Italy that the estimated population size for 2021 is 59,257,566. Our estimated population size is 61,906,029, which has a percentage error of 4.469. So, in the future, we can work to find better models for estimating the future population trends so as to minimize the error.

6. BIBLIOGRAPHY

1. Fundamentals of Statistics-Vol II by A.M. Gun, M.K Gupta, B Dasgupta
2. Fundamentals of Applied Statistics by S.C. Gupta
3. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/moving-average/interpret-the-results/all-statistics-and-graphs/

7. ACKNOWLEDGEMENT

With utmost gratification, I present this project on “Population Projection of Italy for the year 2021”

Inspiration and motivation have played a key role in the success of any venture. I take this opportunity to acknowledge my indebtedness and my deep sense of gratitude to my guide, Dr. Abisa Sinha, whose valuable guidance, experience, and kind supervision throughout the course have shaped the present work. In addition, I want to extend my heartfelt thanks for her consultative help and constructive suggestion on the matter in this project.

I am greatly obliged to the professors of Bethune College, Kolkata for their help in the completion of the project. Last but not the least; my parents are also an inspiration for me. So with due regards, I express my gratitude towards them for their assistance and guidance.

Thank you

Soumita Bandyopadhyay

8. APPENDIX

- ***MINITAB OUTPUTS:***

1. For Total Population:

Trend Analysis for Population (Model type: Linear)

Data Pop.

Length 13

NMissing 0

Fitted Trend Equation

Y_t = 19893679 + 3003536×t

Accuracy Measures

MAPE 1.93757E+00

MAD 7.53000E+05

MSD 8.39402E+11

Forecasts

Period	Forecast
14	61943181
15	64946717

Trend Analysis for Pop.(Model type: Exponential growth)

Data Pop.

Length 13

NMissing 0

Fitted Trend Equation

Y_t = 22786840 × (1.08080^t)

Accuracy Measures

MAPE 3.97950E+00

MAD 1.56559E+06

MSD 4.43176E+12

Forecasts

Period Forecast

14 67628669

15 73093210

Trend Analysis for Pop. (Model type: Quadratic)

Data Pop.

Length 13

NMissing 0

Fitted Trend Equation

$Y_t = 18913503 + 3395606 \times t - 28005 \times t^2$

Accuracy Measures

MAPE 1.81366E+00

MAD 7.21415E+05

MSD 7.18623E+11

Forecasts

Period Forecast

14 60963005

15 63546466

Trend Analysis for Pop. (Model type: S-Curve)

Data Pop.

Length 13

NMissing 0

Fitted Trend Equation

$Y_t = (10^9) / (11.1091 + 35.0866 \times (0.871188^t))$

Accuracy Measures

MAPE 2.12650E+00

MAD 7.61303E+05

MSD 1.02391E+12

Forecasts

Period	Forecast
14	61732224
15	64336158

2. *For Male Population:*

Trend Analysis for Number of males (Model type:Linear)

Data Number of males
Length 13
NMissing 0

Fitted Trend Equation

$Yt = 9686034 + 1462392 \times t$

Accuracy Measures
MAPE 1.93757E+00
MAD 3.66628E+05
MSD 1.98990E+11

Forecasts

Period	Forecast
14	30159518
15	31621910

Trend Analysis for Number of males (Model type: Exponential growth)

Data Number of males
Length 13
NMissing 0

Fitted Trend Equation

Yt = 11094686 × (1.08080^t)

Accuracy Measures

MAPE 3.97950E+00
MAD 7.62270E+05
MSD 1.05060E+12

Forecasts

Period	Forecast
14	32927725
15	35588356

Trend Analysis for Number of males (Model type: Quadratic)

Data Number of males
Length 13
NMissing 0
Fitted Trend Equation

Yt = 9208796 + 1653287×t - 13635×t^2

Accuracy Measures

MAPE 1.81366E+00
MAD 3.51250E+05
MSD 1.70358E+11

Forecasts

Period	Forecast
14	29682280
15	30940141

Trend Analysis for Number of males (Model type: S-Curve)

Data Number of males
Length 13
NMissing 0

Fitted Trend Equation

Yt = (10^8) / (2.28165 + 7.20628×(0.871188^t))

Accuracy Measures

MAPE 2.12650E+00
MAD 3.70671E+05
MSD 2.42731E+11

Forecasts

Period Forecast
14 30056805
15 31324635

3. *For Female Population:*

Trend Analysis for Number of females(Model type: Linear)
Data Number of females
Length 13
NMissing 0

Fitted Trend Equation

Yt = 10207645 + 1541144×t

Accuracy Measures

MAPE 1.93757E+00
MAD 3.86372E+05
MSD 2.20999E+11

Forecasts

Period Forecast
14 31783663
15 33324808

Trend Analysis for Number of females (Model type: Exponential growth)

Data Number of females
Length 13
NMissing 0

Fitted Trend Equation

Yt = 11692155 × (1.08080^t)

Accuracy Measures

MAPE 3.97950E+00
MAD 8.03320E+05
MSD 1.16680E+12

Forecasts

Period	Forecast
14	34700943
15	37504854

Trend Analysis for Number of females (Model type: Quadratic)

Data Number of females
Length 13
NMissing 0

Fitted Trend Equation

Yt = 9704707 + 1742319×t - 14370×t^2

Accuracy Measures

MAPE 1.81366E+00
MAD 3.70165E+05
MSD 1.89200E+11

Forecasts

Period	Forecast
14	31280725
15	32606324

Trend Analysis for Number of females(Model type: S-Curve)

Data Number of females
Length 13
NMissing 0

Fitted Trend Equation

Yt = (10^8) / (2.16506 + 6.83804×(0.871188^t))

Accuracy Measures

MAPE 2.12650E+00
MAD 3.90632E+05
MSD 2.69578E+11

Forecasts

Period	Forecast
14	31675419
15	33011524