R-PROGRAMMING RESEARCH PAPER

Submitted to Prof. BaijuP

Submitted by Mahak Maheshwari
(242PGF011)

1.TITLE OF THE PAPER

Uban Population Growth Prediction in India using ARMA models

2.INTRODUCTION

Urban population trends play a crucial role in shaping economic development, resource allocation, and environmental sustainability. Over the years, urbanization has been a key driver of global progress, but recent shifts in urban population dynamics pose new challenges. Understanding these trends is essential for policymakers, urban planners, and researchers to address the socio-economic and environmental implications of urban population changes.

This research paper aims to analyse and forecast urban population growth using historical data from 1986 to 2019. An ARIMA(0,1,1) model with drift has been employed to project population trends for the decade 2020–2029. The analysis highlights a declining trend in urban population, which could have significant policy and socio-economic implications.

3.LITERATURE REVIEW

(Ali, March 2020) defines urban population as a form of a "social transformation from traditional rural societies to modern, industrial and urban communities". The paper studies the trends of urbanization in India as well as the growth of cities, metropolitan cities and distribution of urban population in states (from 1901 to 2011). From the data computed, it is seen that the total population increased from 238.3 million to 1210.1 million in 2011 where as the urban population increased from 11 per cent in 1901 to 31 per cent in 2011. It is more than tenfold increase in the urban population growth of India. From 1901 to 1951, the urbanization growth has been slow, but it picked a sharp pace after 1951. From 1951-1961 the growth became slow, with an exponential growth of only 2.34. Then from 1961-1981 the growth increased, where as from 1981-2001 the growth again declined. But from 2001-2011, the growth rate increased due to globalization as it attracted more and more employment towards the urban areas.

(Satterthwaite, 2014) explores the concept of urbanization and urban population growth. It was found that the urban population growth has been showing a decreasing growth in Sub-Saharan Africa since the 1970s. Despite the slow growth rates, the number of people migrating to the urban cities is rising. The research paper also discusses regional growth. The high-level income country Europe has largely completed its demographic transition, shows a low population growth where as the Sub-Saharan Africa, which is a low-income country, is still experiencing high population growth. While these transitions provide a framework to understand urbanization, they don't capture the individual choices and struggles that urbanization shapes communities.

(Aswale, 2007) finds out that since urbanization and per capita income are positively correlated, there is no to less correlation between urbanization reduction of population below poverty line. There are many factors responsible for the reason such as neglecting of slums and so on. Three things are expected out of a economic development- rise in per capita income, reduction in the rate and magnitude and reduction of population below poverty line. It has been seen that India is far behind the high-income countries with respect to the level of urbanization.

(Rahman, 2012) studies the growth pattern to forecast urban population in SAARC countries. It has sourced data from UNPD from 1950 to 2000 in five-year intervals. The authors have tried to fit two models- exponential model and ARMA. It was found that the ARMA model was far superiorly fitting and explaining the predictions with respect to the exponential model. The authors have use ARMA

model to forecast the time trend behaviour of the urban population as a percentage of total population up to the year 2025. One of the findings of the research showed that urbanization in Bangladesh was faster than any other SAARC countries. ARMA models for all the seven countries of SAARC resulted in smaller and more significant accepted mean soot square than the fitted exponential growth models.

(BHAGAT, 2011) mentions that after the 2011 census urbanization had increased way faster than expected. There was a decreasing growth in the 1980s and 1990s, but after 1990s the growth was rapid where the urban population absolute growth was more than the rural population growth. The increase in urban population growth was from 2001-2011. Urbanization only occurs when the urban population growth rate is greater than the rural population growth rate. Hence, the urban population growth differential is crucial. The net rural-urban classification and net rural-to-urban migration were mainly responsible for higher urban-rural growth differential leading to speedy urbanization from 2001-2011.

4.DATA AND METHADOLOGY

4.1 Data Collection

The data for urban population growth rate has been collected from **World Bank**. Time period for analysis is from **1986 to 2019**. To forecast unemployment rates, we have used the R software.

TABLE 1: Urban Population Growth Rate (1986-2019)

Year	urban_population_growth	Year	urban_population_growth	Year	urban_population_growth
1986	3.203949848	1997	2.702997788	2008	2.552873022
1987	3.17436058	1998	2.66886959	2009	2.512308892
1988	3.146482897	1999	2.628230166	2010	2.49274629
1989	3.130152469	2000	2.598675553	2011	2.474032431
1990	3.092929503	2001	2.711573919	2012	2.470338202
1991	3.002043136	2002	2.937708255	2013	2.457264391
1992	2.87029913	2003	2.878886128	2014	2.423844627
1993	2.841324866	2004	2.82462859	2015	2.394052977
1994	2.811885018	2005	2.746251941	2016	2.413097263
1995	2.776692245	2006	2.660297138	2017	2.407475136
1996	2.737602928	2007	2.599633138	2018	2.359169975
				2019	2.315801985

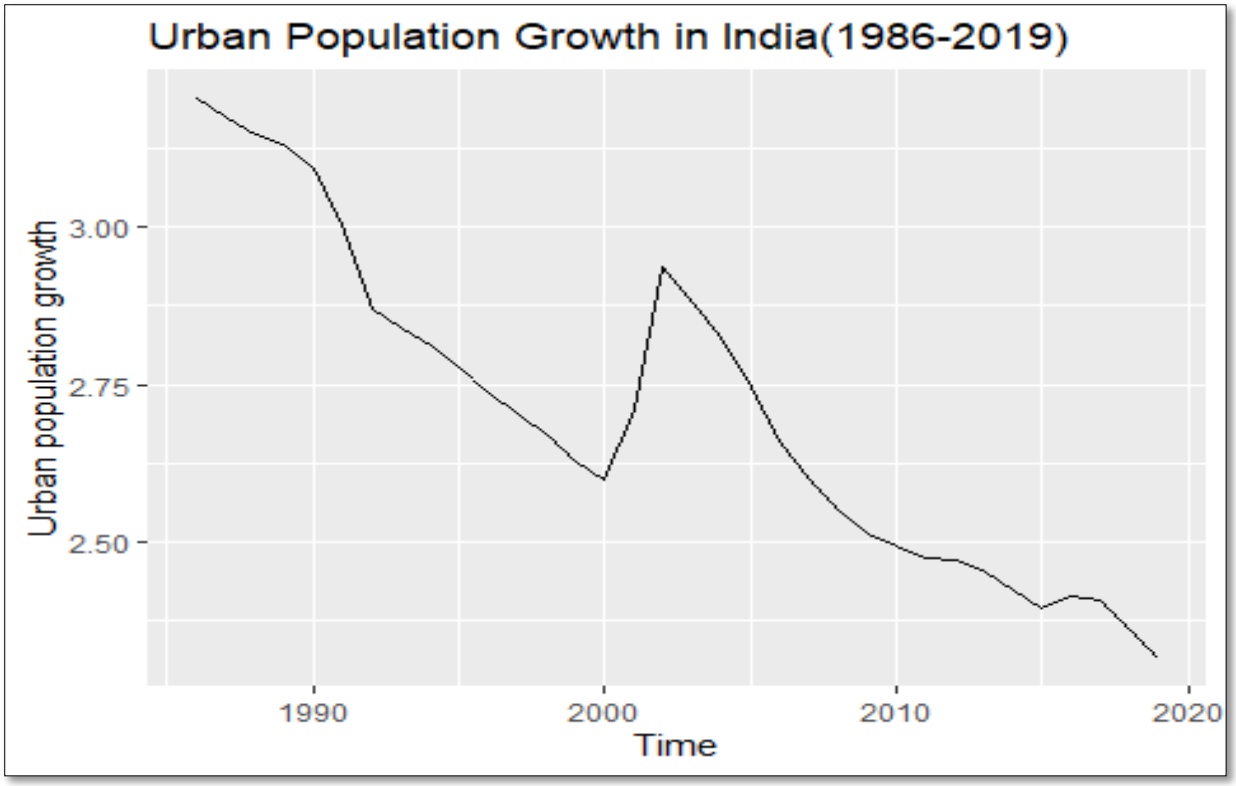


Figure 1. Annual urban population growth rate (%) in India (1986-2019)

From Figure 1, it can be seen that the urban population growth rate was showing a decreasing trend till 2000 and a sudden sharp rise after 2000. It is due to the significant economic reforms in 1991, when India moved towards liberalization, privatization and globalization. The effects of such policies were not immediate, and took its own time to impact urban growth (which can be seen by the sharp rise after 2000). Overall, the urban population growth shows a declining trend after 2003.

4.2 Trends in the Urban Population Growth data for India (1986-2019)

TABLE 2: ACF Values for 9 period lags

Lag	ACF Values
[1]	1.00000000
[2]	0.87648100
[3]	0.73677262
[4]	0.60602838
[5]	0.47807966
[6]	0.35436071
[7]	0.25618031
[8]	0.19319957
[9]	0.13602458

[10]	0.07962406

These ACF values indicate that the urban population growth in India has a **strong short-term dependence** since years closer have a high correlation and it gradually decreases over longer periods.

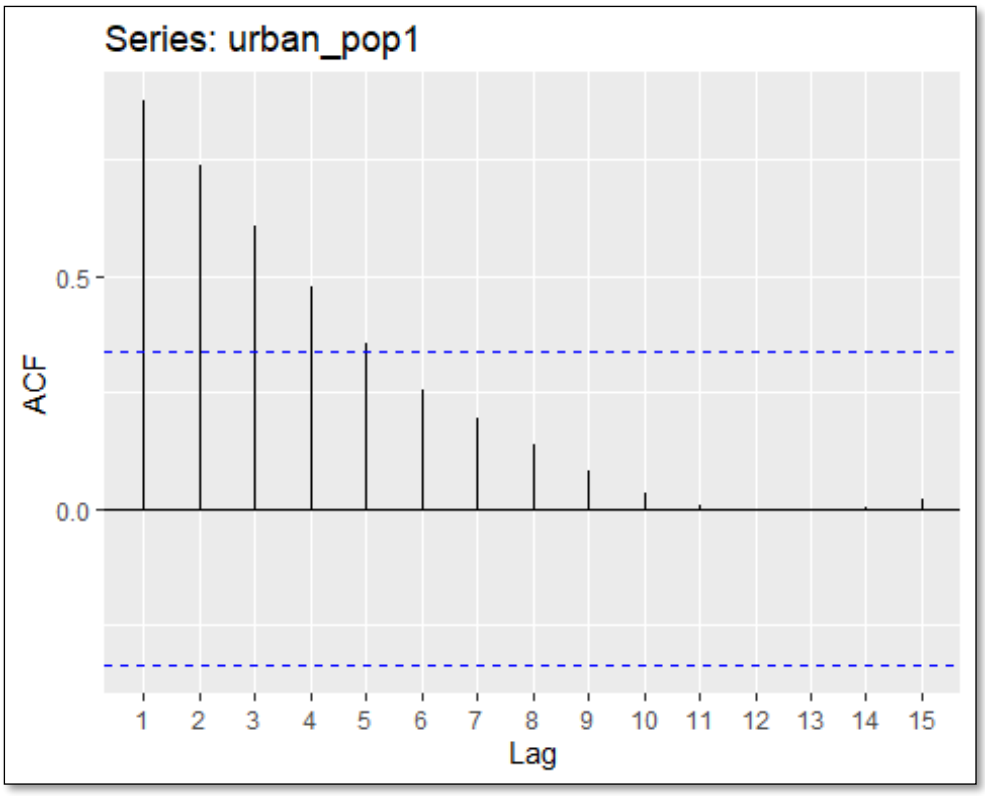


Figure 2. Correlogram

The gradual decline in the ACF plots or correlogram is associated with **non-stationary time series data**. Since the ACF values after one time period lag is not sharply falling to zero, it suggests that the urban population growth is influenced by its past values. The data likely shows a *trend*.

4.2.1 Finding the best forecasting method (when the data is non-stationary)

There are four main methods used to find the forecasts-Mean forecasting, *Naive forecasting, Seasonal naive forecasting and Drift*

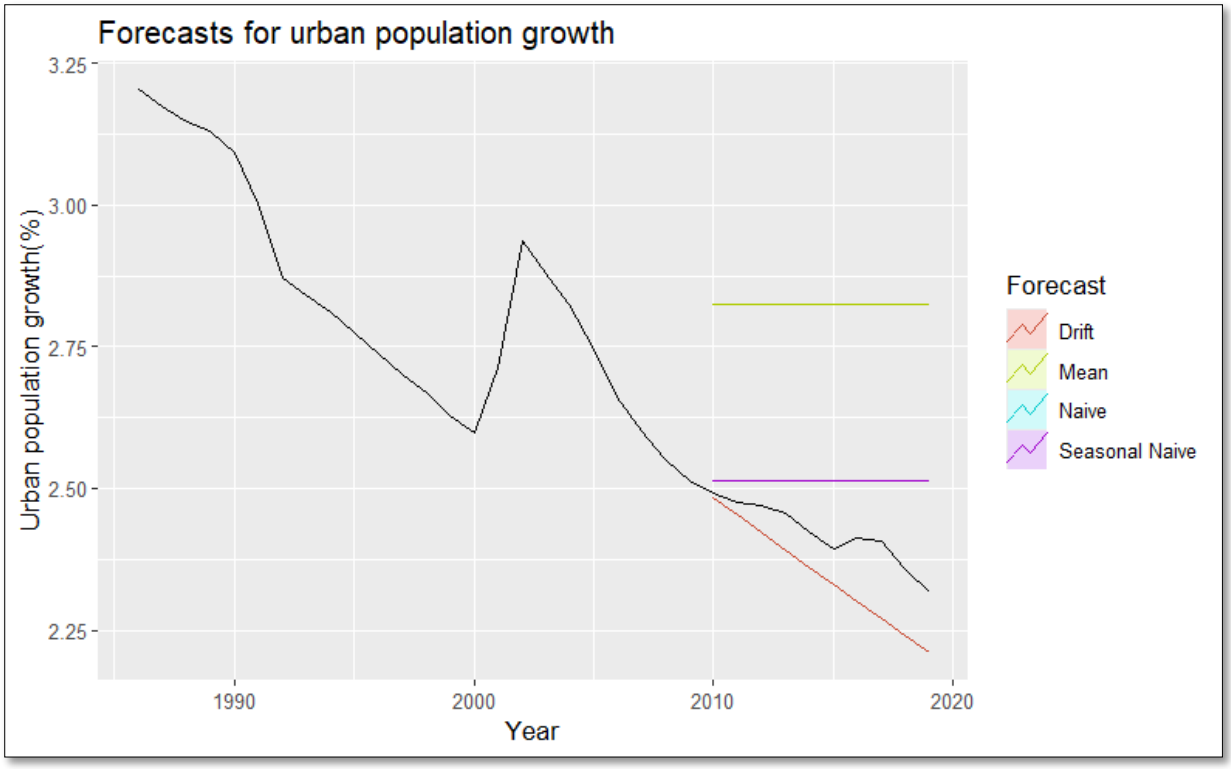


Figure 3. Forecasting Model with non-stationary urban population growth data

TABLE 3: RMSE VALUES FROM ALL FORECASTING METHODS

		ME	RMSE	MAE
MEAN FORECASTING	Training set	-9.253666e-17	0.2032379	0.1686414
	Test set	-4.046617e-01	0.4080550	0.4046617
NAIVE FORECASTING	Training set	-0.03007135	0.07506856	0.05955245
	Test set	-0.09152656	0.10552211	0.09152656
SEASONAL NAIVE FORECASTING	Training set	-0.03007135	0.07506856	0.05955245
	Test set	-0.09152656	0.10552211	0.09152656
DRIFT	Training set	2.123878e-16	0.06878229	0.03633368
	Test set	7.386584e-02	0.08387129	0.07386584

From the table above, we can say that the drift method's test set RMSE (0.0839) is lower than that of the mean (0.4080), naive (0.1055), and seasonal naive (0.1055) methods, confirming its robustness in predicting future values with minimal error.

4.2.2 Finding the best forecasting method (when the data is stationary)

Checking for Stationarity:

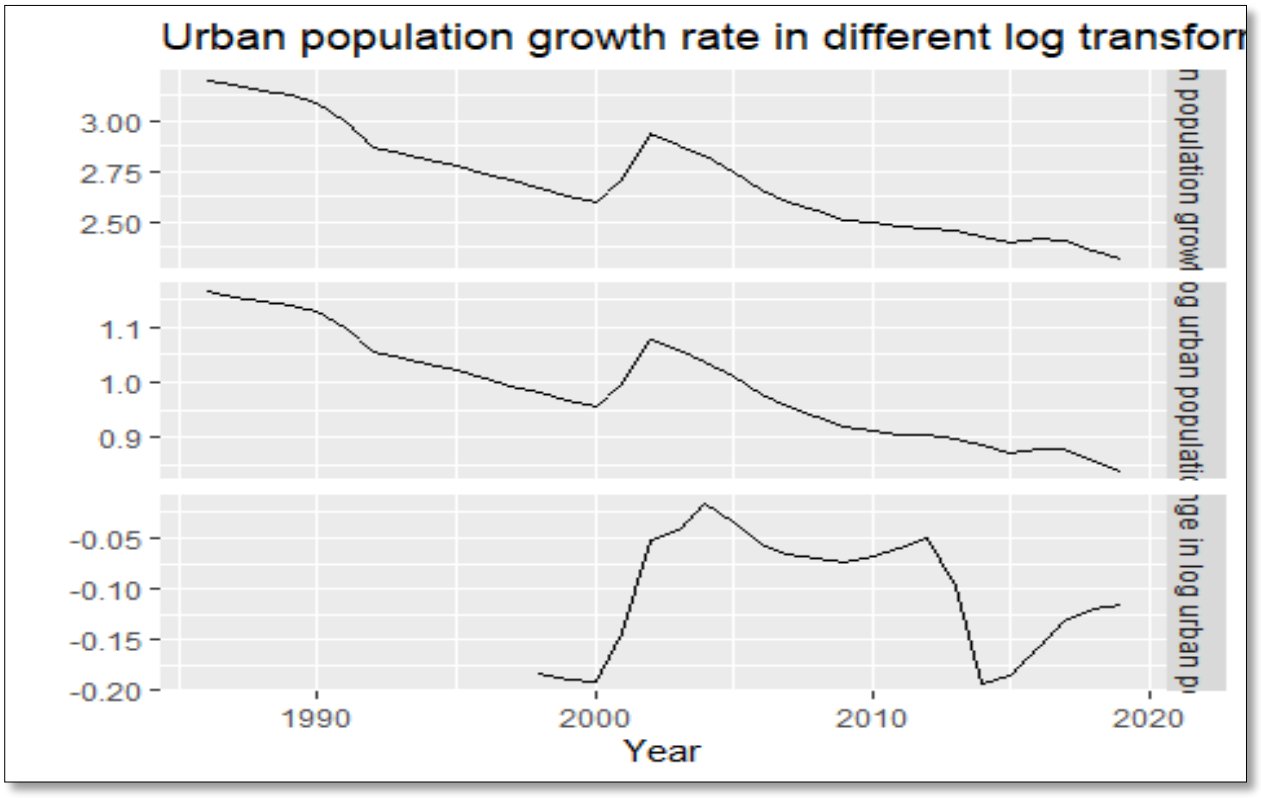


Figure 4. Urban population growth rate in different log transformation

KPSS Unit Root Test

Test is of type: mu with 3 lags.

Value of test-statistic is: 0.8428

Significance level	10pct	5pct	2.5pct	1pct
Critical values	0.347	0.463	0.574	0.739

Based on the KPSS test, where the test statistic is 0.8428 and is greater than all the critical values (at 10%, 5%, 2.5%, 1% levels), we reject the null hypothesis of stationarity, concluding that the data is non-stationary.

```
## ARIMA(0,1,1) with drift
##
## Coefficients:
##         ma1    drift
##       0.4891  -0.0270
## s.e.  0.1476   0.0134
```

```
##
## sigma^2 = 0.00291:  log likelihood = 50.43
## AIC=-94.85   AICc=-94.03   BIC=-90.36
##
```

TABLE 4: Training set error measures

	ME	RMSE	MAE	MPE	ACF1
Training set	8.70182e-05	0.05150742	0.0291149	0.006308642	0.01661087

The ARIMA(0,1,1) model with drift successfully captures the dynamics of the urban population time series. The model uses first-order differencing to make the data stationary, incorporates a moving average term (MA(1)) to account for short-term fluctuations, and includes a drift term (-0.0270) indicating a slight downward trend in the population. The model fit is strong, with low AIC (-94.85) and BIC (-90.36) values, and error metrics show good accuracy.

5.RESULTS AND DISCUSSION

Comparing the ARIMA(0,1,1) model with drift to the simple drift method used on the non-stationary data, we find that the ARIMA model provides a lower **RMSE of 0.0515**. It indicates improved accuracy over the drift-only method, which had a higher RMSE (0.0687). This suggests that, while the drift method performed relatively well for basic forecasting, the ARIMA(0,1,1) model with drift offers a more refined approach by incorporating both differencing and a moving average term to account for short-term variations, making it a more precise and reliable choice for forecasting the urban population. Therefore, the ARIMA(0,1,1) model with drift is the preferred method for accurate long-term forecasting.

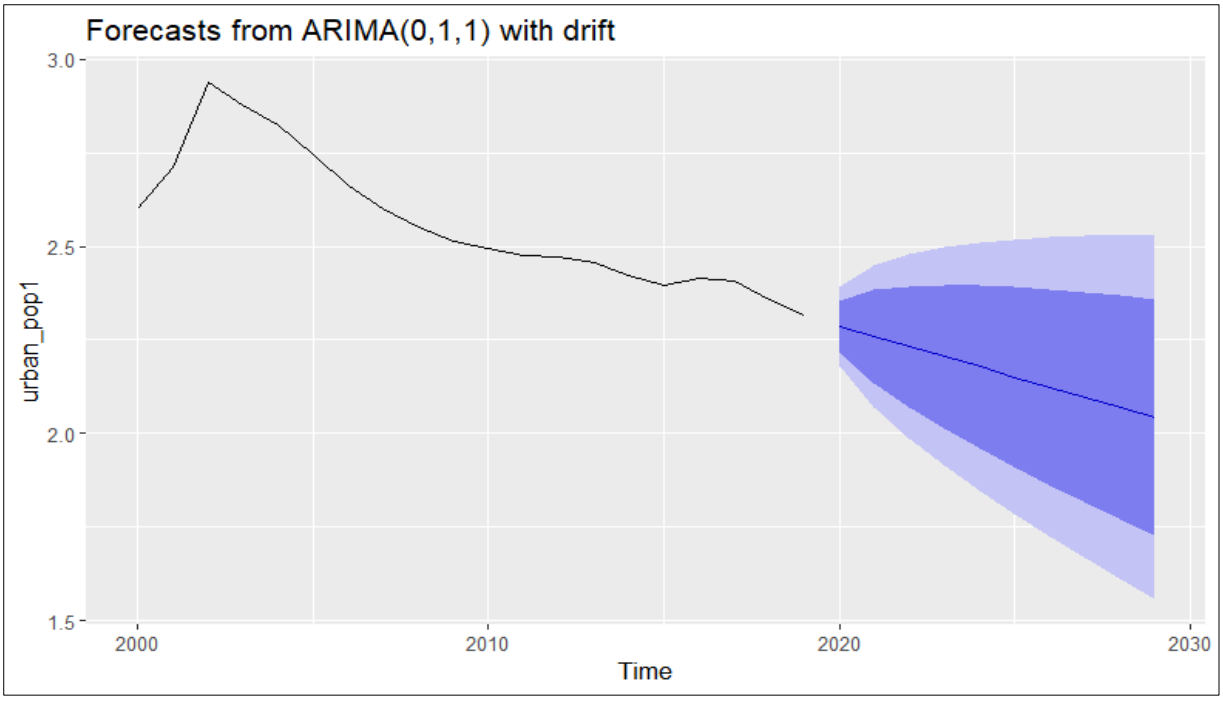


Figure 5. Forecasted urban population growth (from 2020-2029)

In the above figure 5, the black line represents the observed historical urban population growth data, where as the blue line represents the forecasted trend beyond 2019. The shaded regions are the confidence intervals. The forecasts suggest a slight declining trend in the urban population growth rate over the next decade. This could be due to factors such as declining birth rates, migration patterns, or other socio-economic factors.

6.POLICY IMPLICATIONS AND FUTURE RESEARCH

6.1 Policy Implications

- **Urban Attraction Strategies**: Governments should focus on improving job opportunities, urban infrastructure, and quality of life to attract and retain residents in urban areas.
- **Economic Adaptation:** A declining urban population growth may affect economic productivity requiring tailored economic policies.

6.2 Future Research

The main future research is to understand the reason for declining of urban population growth rate in India. That is to investigate demographic factors (e.g., birth rates, migration patterns), economic shifts, and quality-of-life issues contributing to the observed downward trend. Understanding whether rural-to-urban migration has slowed or if urban-to-rural migration is increasing can provide valuable insights into population shifts. Economic factors, such as changes in employment opportunities, rising living costs, and the affordability of urban housing, also require detailed analysis to determine how they influence individuals' decisions to stay in or leave urban environments. Furthermore, research should explore quality-of-life issues, such as access to education, healthcare, and public services, as well as environmental concerns like pollution and overcrowding, which may discourage urban living. By identifying these drivers, future studies can offer actionable insights to help policymakers develop strategies to address the causes of urban population decline and promote sustainable urban growth.

7.REFERENCES

Ali, E. (March 2020). Urbanisation in India: Causes, Growth, Trends, Patterns, Consequences & Remedial Measures. *ResearchGate*.

Aswale, S. (2007). Growing Trends Towards Urbanization and its impact on Economic Development. *Southern Economic Journal* .

BHAGAT, R. B. (2011). Emerging Pattern of Urbanisation in India. *Economic and Political Weekly*, Vol. 46, No. 34, pp. 10-12.

Rahman, M. A. (2012). Time Series Models for Growth of Urban Population in SAARC Countries . *International Scientific Press*, vol.2, 2012, 109-119.

Satterthwaite, G. M. (2014). Urbanisation concepts and trends . *International Institute for Environment and Development*.

8. R CODE

Urban Population Growth Forecasting using ARMA

1. Data Import and Transformation

```
file.exists("C:\\Users\\mahak\\Downloads\\urban_pop(1986-2019).xls")
```

```
## [1] TRUE
```

```
data<- readxl::read_excel("C:\\Users\\mahak\\OneDrive\\Documents\\R-Script\\urban_pop(1986-2019).xls")
data
```

```
## # A tibble: 34 × 2
##    Year urban_population_growth
##   <dbl>                   <dbl>
## 1  1986                    3.20
## 2  1987                    3.17
## 3  1988                    3.15
## 4  1989                    3.13
## 5  1990                    3.09
## 6  1991                    3.00
## 7  1992                    2.87
## 8  1993                    2.84
## 9  1994                    2.81
## 10  1995                   2.78
## # ℹ 24 more rows
```

```
urban_pop1<- ts(data[,2], start= 1986, end = 2019)
urban_pop1
```

```
## Time Series:
## Start = 1986
## End = 2019
## Frequency = 1
##      urban_population_growth
##  [1,]                3.203950
##  [2,]                3.174361
##  [3,]                3.146483
##  [4,]                3.130152
##  [5,]                3.092930
##  [6,]                3.002043
##  [7,]                2.870299
##  [8,]                2.841325
##  [9,]                2.811885
## [10,]                2.776692
## [11,]                2.737603
## [12,]                2.702998
## [13,]                2.668870
## [14,]                2.628230
## [15,]                2.598676
## [16,]                2.711574
## [17,]                2.937708
## [18,]                2.878886
## [19,]                2.824629
## [20,]                2.746252
## [21,]                2.660297
## [22,]                2.599633
## [23,]                2.552873
## [24,]                2.512309
## [25,]                2.492746
```

```
## [26,]          2.474032
## [27,]          2.470338
## [28,]          2.457264
## [29,]          2.423845
## [30,]          2.394053
## [31,]          2.413097
## [32,]          2.407475
## [33,]          2.359170
## [34,]          2.315802
```

2. Visualization

```
library(fpp2)
```

```
## Warning: package 'fpp2' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──── fpp2 2.5 ──
```

```
## ✓ ggplot2   3.5.1      ✓ fma       2.5
## ✓ forecast  8.23.0     ✓ expsmooth 2.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

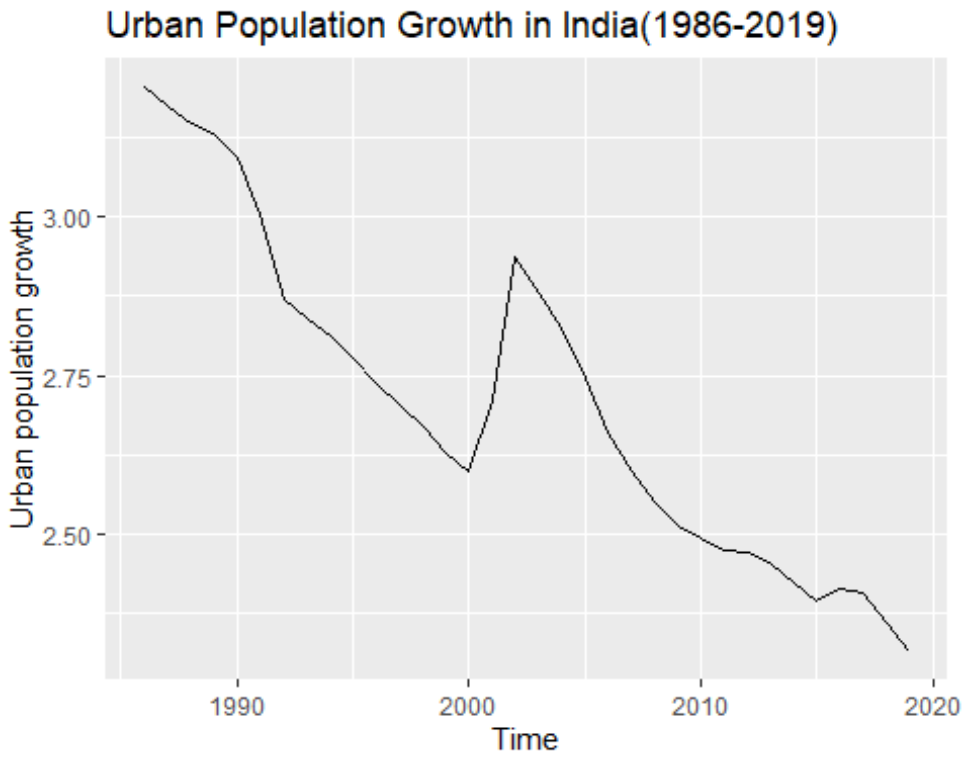
```
## Warning: package 'forecast' was built under R version 4.3.3
```

```
## Warning: package 'fma' was built under R version 4.3.3
```

```
## Warning: package 'expsmooth' was built under R version 4.3.3
```

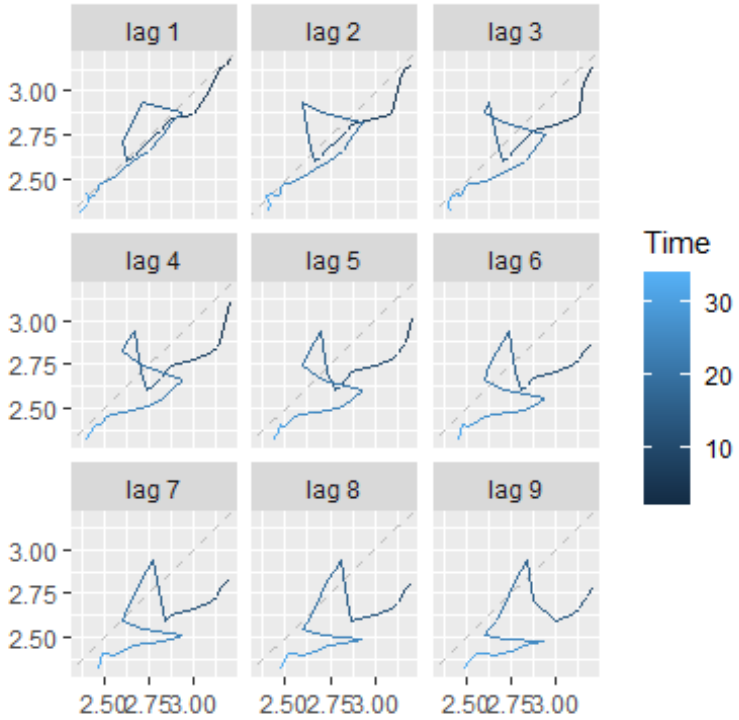
```
##
```

```
autoplot(urban_pop1)+xlab("Time")+ylab("Urban population growth")+ggtitle("Urban Population G
rowth in India(1986-2019)")
```



3. Checking for Time Trend and Seasonality Lagplot:

```
gglagplot(urban_pop1)
```



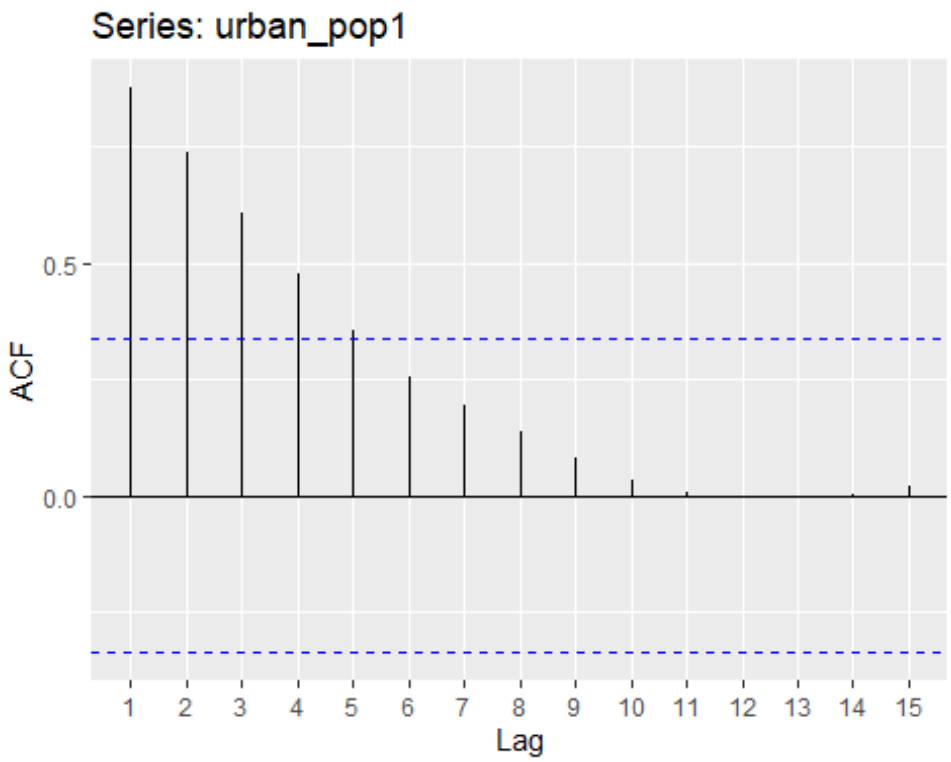
ACF values:

```
acf_values <- acf(urban_pop1, lag.max=9, plot=F)
acf_values$acf
```

```
## , , 1
##
##            [,1]
##  [1,] 1.00000000
##  [2,] 0.87648100
##  [3,] 0.73677262
##  [4,] 0.60602838
##  [5,] 0.47807966
##  [6,] 0.35436071
##  [7,] 0.25618031
##  [8,] 0.19319957
##  [9,] 0.13602458
## [10,] 0.07962406
```

Plotting the correlogram:

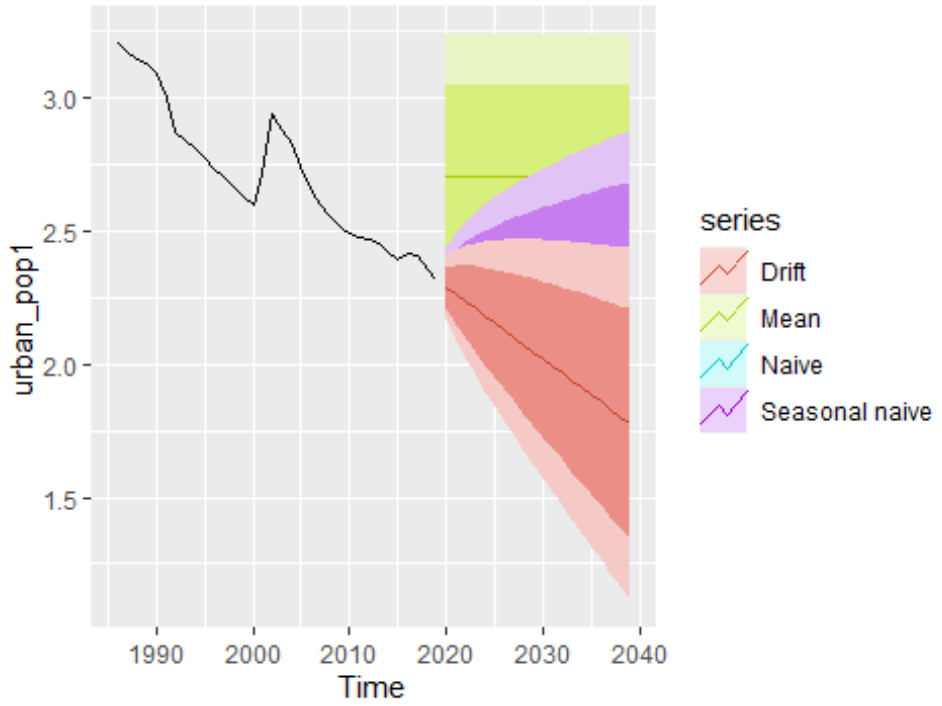
```
ggAcf(urban_pop1)
```



4. Model Estimation

not that: we are finding the best fit model using non-stationary data Plotting all the forecasting methods:

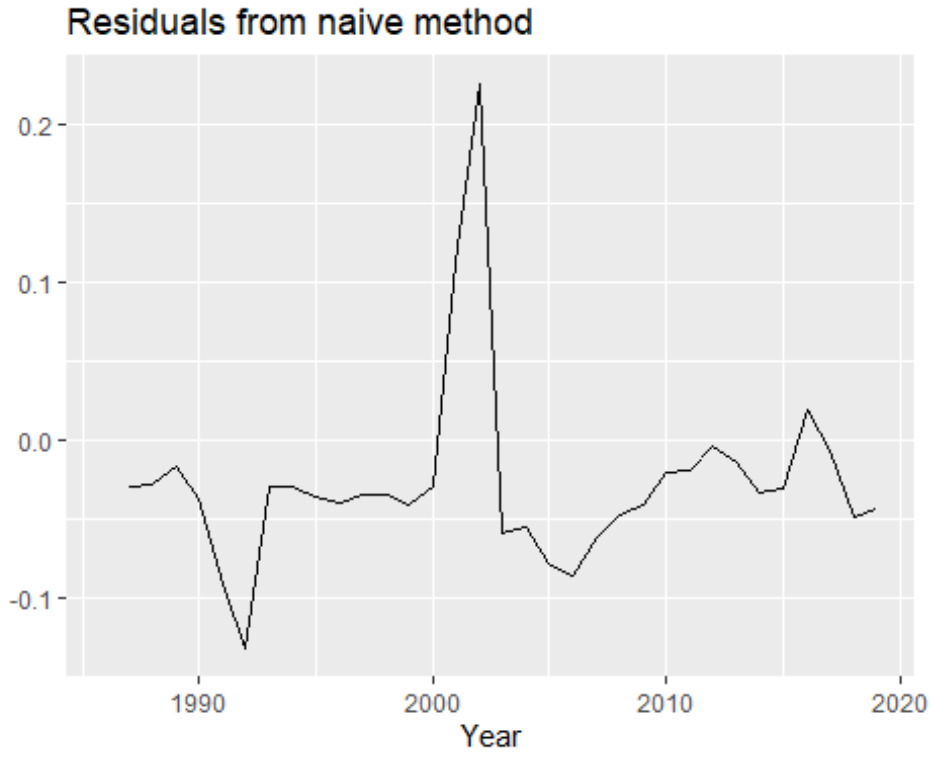
```
autoplot(urban_pop1)+
  autolayer(meanf(urban_pop1,h=20), series="Mean",PI=T)+
  autolayer(naive(urban_pop1,h=20), series="Naive",PI=T)+
  autolayer(snaive(urban_pop1,h=20), series="Seasonal naive",PI=T)+
  autolayer(rwf(urban_pop1,h=20,drift = T), series="Drift",PI=T)
```



Plotting the residual graph:

1.*Using naive method:*

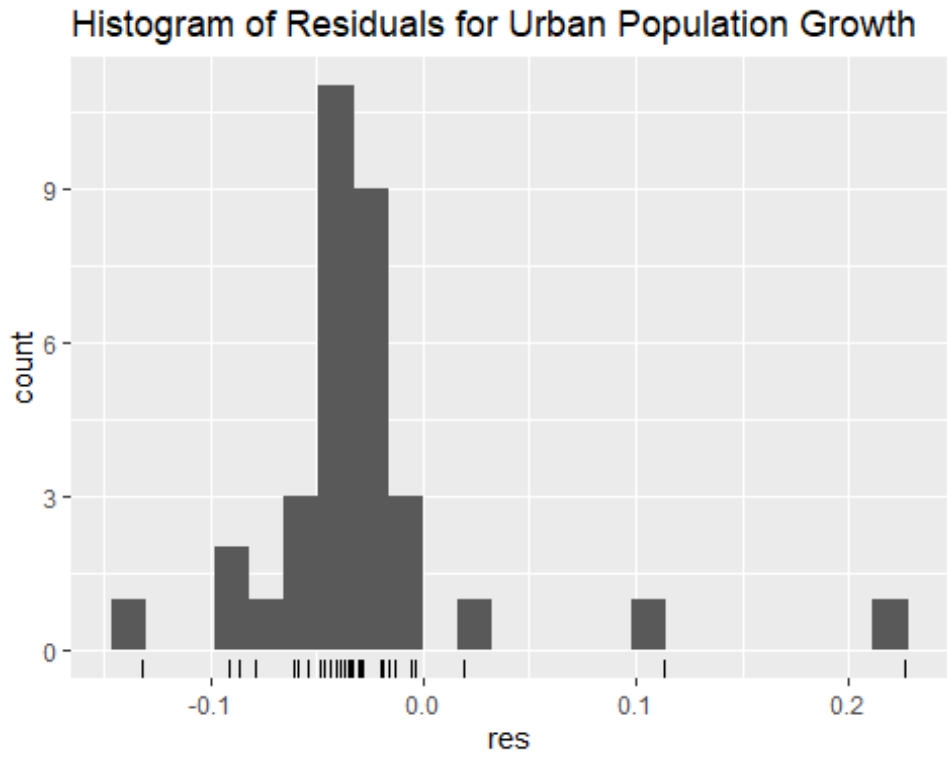
```
res<- residuals(naive(urban_pop1))
autoplot(res)+xlab("Year")+ylab("")+ggtitle("Residuals from naive method")
```



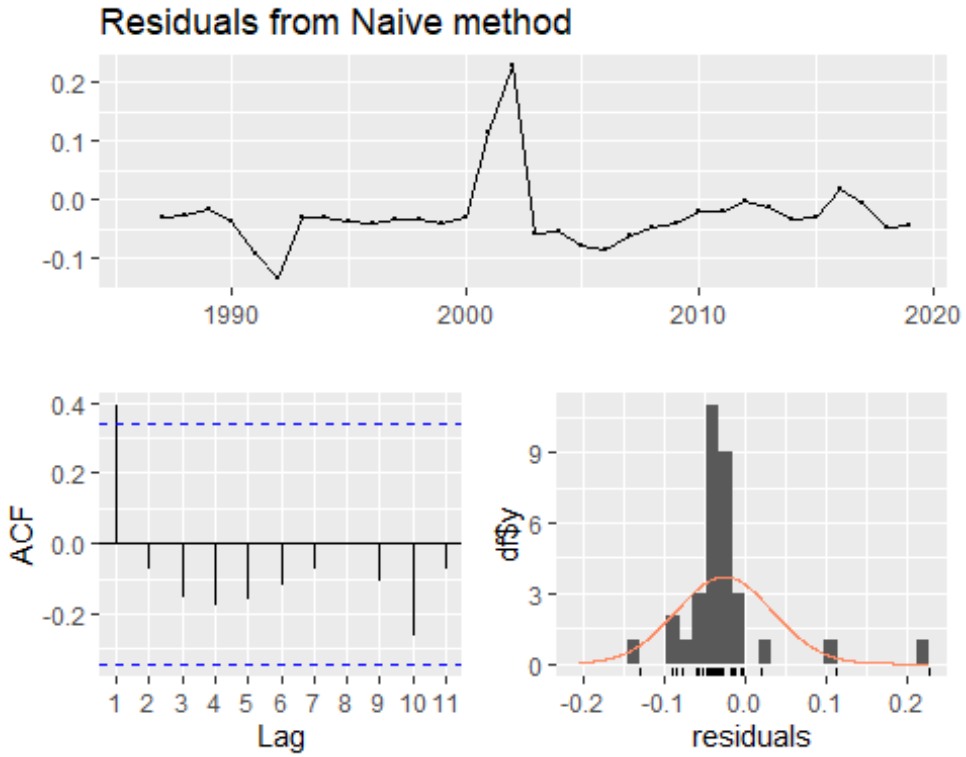
Plotting the histogram of residuals:

```
gghistogram(res)+ggtitle("Histogram of Residuals for Urban Population Growth")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



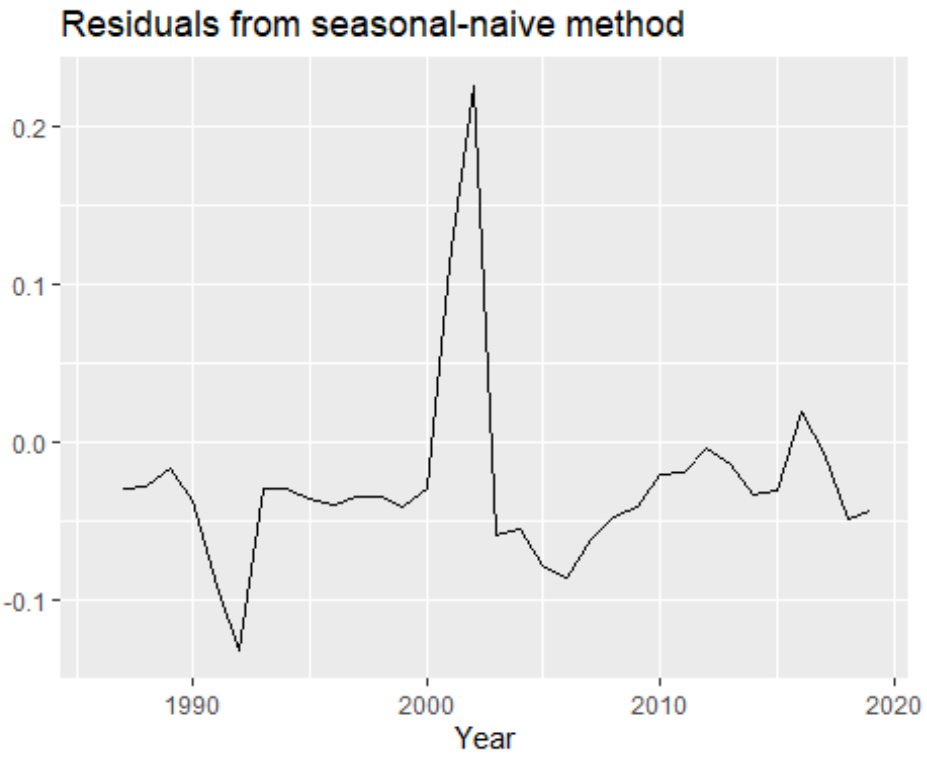

```
checkresiduals(naive(urban_pop1))
```



```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 9.6612, df = 7, p-value = 0.2086
##
## Model df: 0.   Total lags used: 7
```

2. *Using seasonal naive method:*

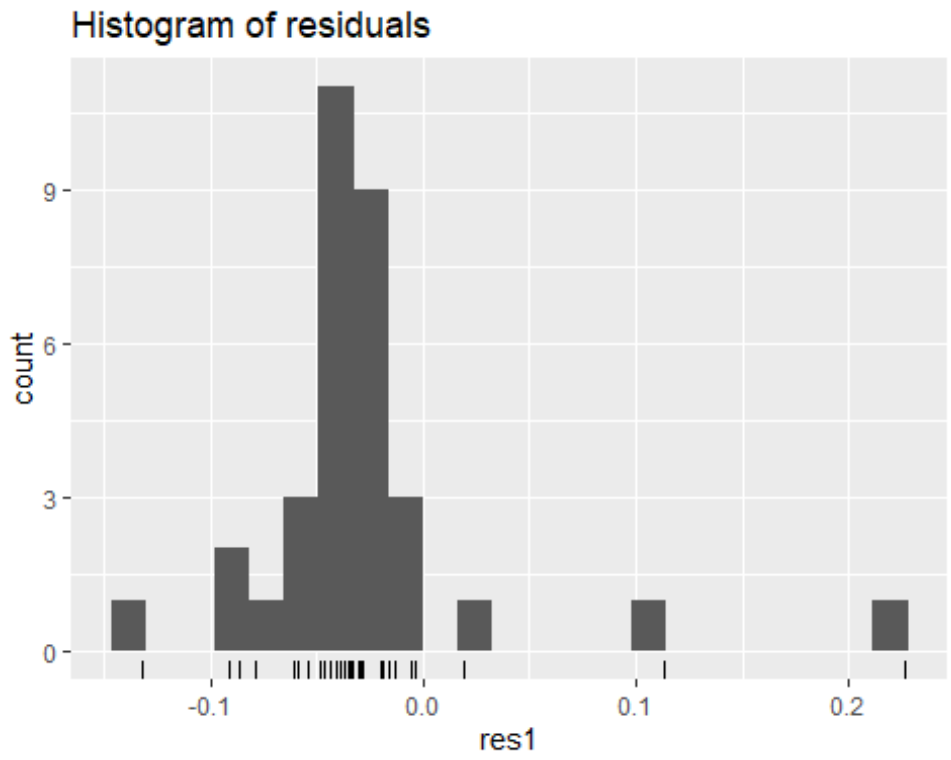
```
res1<- residuals(snaive(urban_pop1))
autoplot(res1)+xlab("Year")+ylab("")+ggtitle("Residuals from seasonal-naive method")
```



plotting the residual histogram:

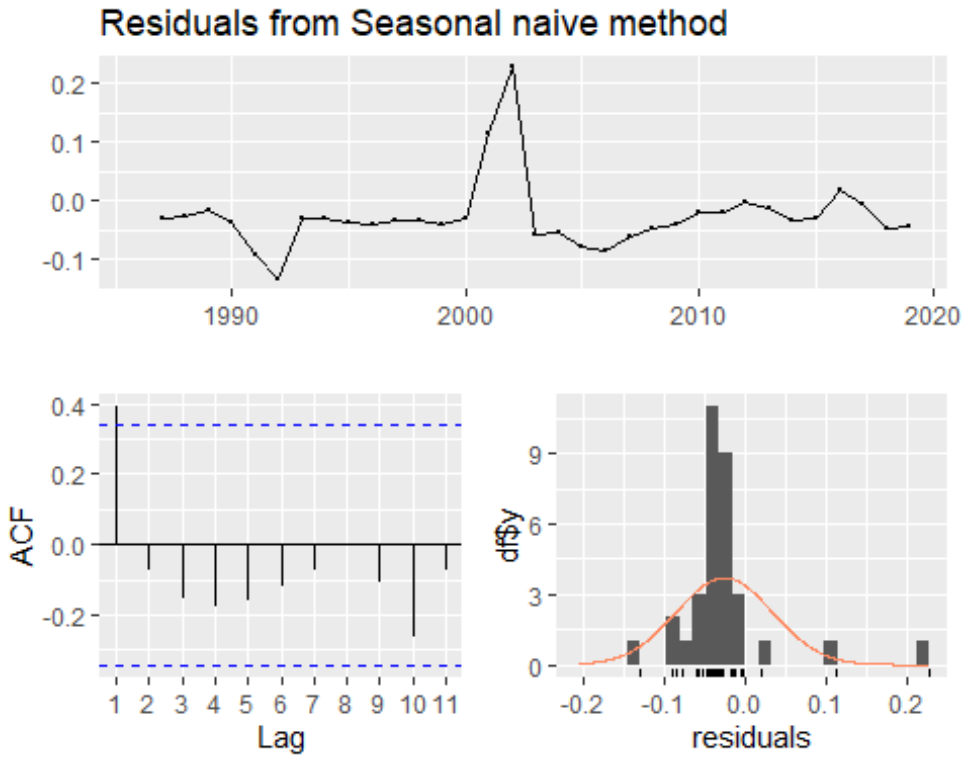
```
gghistogram(res1)+ggtitle("Histogram of residuals")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



all in one graph:

```
checkresiduals(snaive(urban_pop1))
```



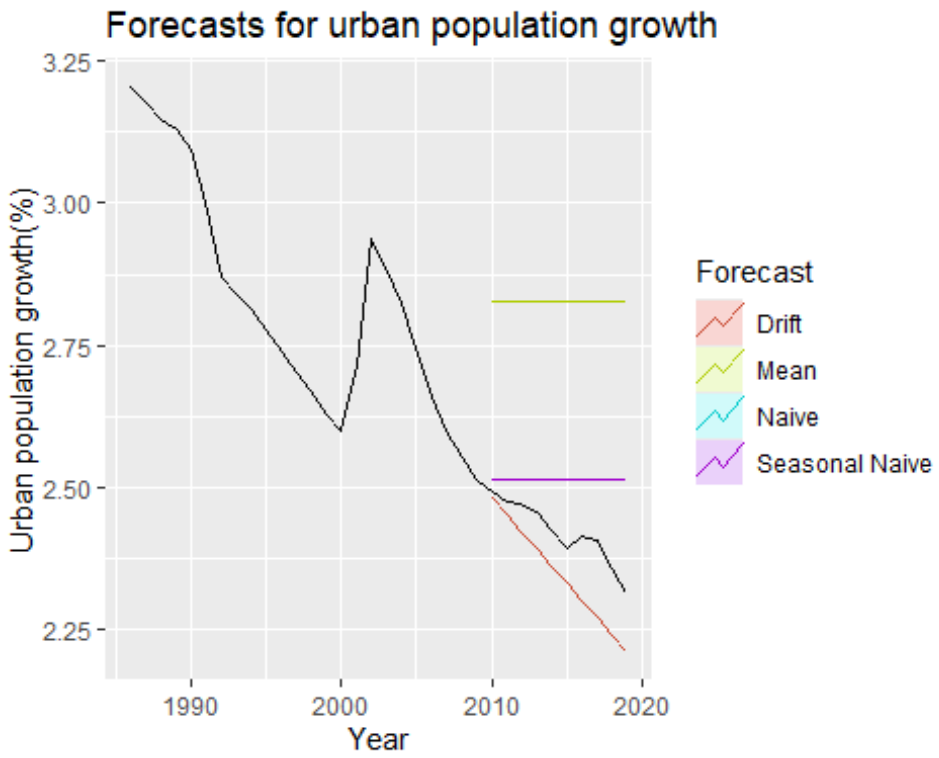

```
##
##  Ljung-Box test
##
## data:  Residuals from Seasonal naive method
```

```
## Q* = 9.6612, df = 7, p-value = 0.2086
##
## Model df: 0.   Total lags used: 7
```

6. Model Evaluation Evaluating forecasting accuracy:

```
urban_pop_training_data<- window(urban_pop1, end=2009) #training data
urbanpop1<-  meanf(urban_pop_training_data, h=10)
urbanpop2<- rwf(urban_pop_training_data, h=10)
urbanpop3<- snaive(urban_pop_training_data, h=10)
urbanpop4<- rwf(urban_pop_training_data, drift = T, h=10)

autoplot(urban_pop1)+ #plotting all the data
  autolayer(urbanpop1, series="Mean", PI= F)+
  autolayer(urbanpop2, series="Naive", PI= F)+
  autolayer(urbanpop3, series="Seasonal Naive", PI= F)+
   autolayer(urbanpop4, series="Drift", PI= F)+
  xlab("Year")+ylab("Urban population growth(%)")+
  ggtitle("Forecasts for urban population growth")+
  guides(colour= guide_legend(title="Forecast"))
```



Checking the accuracy: 1.mean forecasting method:

```
urban_pop_test_data<- window(urban_pop1, start=2010) #test data
accuracy(urbanpop1, urban_pop_test_data)
```

```
##                      ME      RMSE       MAE        MPE      MAPE     MASE
## Training set -9.253666e-17 0.2032379 0.1686414  -0.5061189  5.928916 2.831813
## Test set     -4.046617e-01 0.4080550 0.4046617 -16.7717346 16.771735 6.795047
##                   ACF1 Theil's U
## Training set 0.812913        NA
## Test set     0.576704  14.66482
```

2.Naive method

```
accuracy(urbanpop2, urban_pop_test_data)
```

```
##                     ME       RMSE        MAE       MPE     MAPE     MASE
## Training set -0.03007135 0.07506856 0.05955245 -1.092739 2.124149 1.000000
## Test set     -0.09152656 0.10552211 0.09152656 -3.830288 3.830288 1.536907
##                    ACF1 Theil's U
## Training set 0.3940311        NA
## Test set     0.5767040  3.956703
```

3.Seasonal naive method

```
accuracy(urbanpop3, urban_pop_test_data)
```

```
##                     ME       RMSE        MAE       MPE     MAPE     MASE
## Training set -0.03007135 0.07506856 0.05955245 -1.092739 2.124149 1.000000
## Test set     -0.09152656 0.10552211 0.09152656 -3.830288 3.830288 1.536907
##                    ACF1 Theil's U
## Training set 0.3940311        NA
## Test set     0.5767040  3.956703
```

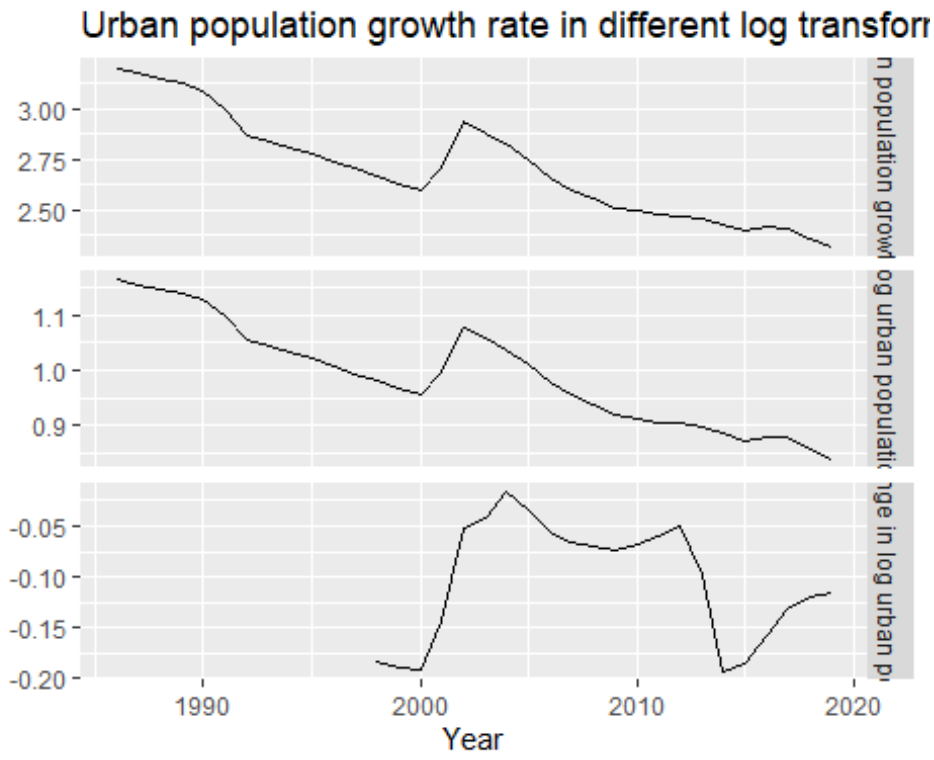
4.Drift method

```
accuracy(urbanpop4, urban_pop_test_data)
```

```
##                      ME       RMSE        MAE         MPE     MAPE      MASE
## Training set 2.123878e-16 0.06878229 0.03633368 -0.01734643 1.286798 0.6101123
## Test set     7.386584e-02 0.08387129 0.07386584  3.07942295 3.079423 1.2403493
##                    ACF1 Theil's U
## Training set 0.3940311        NA
## Test set     0.6965269  3.134765
```

now, converting the non-stationary data into stationary data and then comparing the two models:
Checking for stationarity:

```
cbind("Urban population growth(%)"= urban_pop1, "Monthly log urban population growth"= log(urb
an_pop1), "Yearly change in log urban popn growth"= diff(log(urban_pop1),12)) %>%
  autoplot(facets=T)+xlab("Year")+ylab("")+ggtitle("Urban population growth rate in different log tr
ansformations:")
```



test:

```
library(urca)
```

```
## Warning: package 'urca' was built under R version 4.3.3
```

```
urban_pop1 %>% ur.kpss() %>% summary()
```

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: mu with 3 lags.
##
## Value of test-statistic is: 0.8428
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

since the data is non-stationary, we need to finf the best fitted model to convert it into a stationary series. which model to use?

```
summary((fit<- auto.arima(urban_pop1)))
```

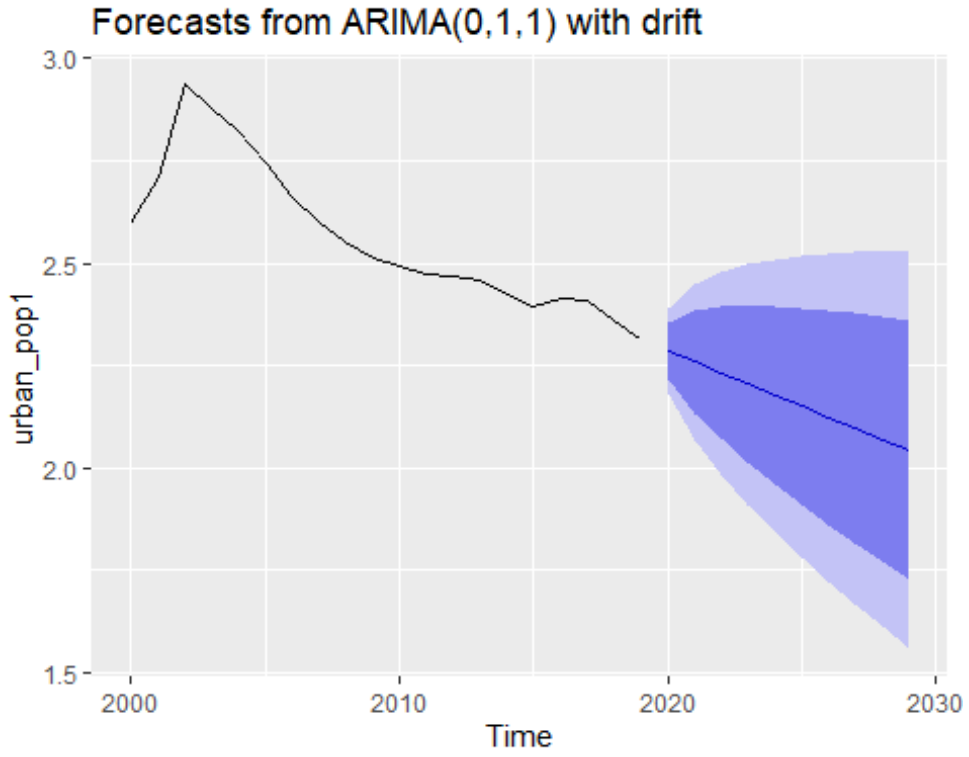
```
## Series: urban_pop1
## ARIMA(0,1,1) with drift
##
## Coefficients:
##          ma1    drift
##       0.4891  -0.0270
## s.e.  0.1476   0.0134
```

```
##
## sigma^2 = 0.00291:  log likelihood = 50.43
## AIC=-94.85   AICc=-94.03   BIC=-90.36
##
## Training set error measures:
##                       ME       RMSE       MAE         MPE     MAPE      MASE
## Training set 8.70182e-05 0.05150742 0.0291149 0.006308642 1.049959 0.5988847
##                    ACF1
## Training set 0.01661087
```

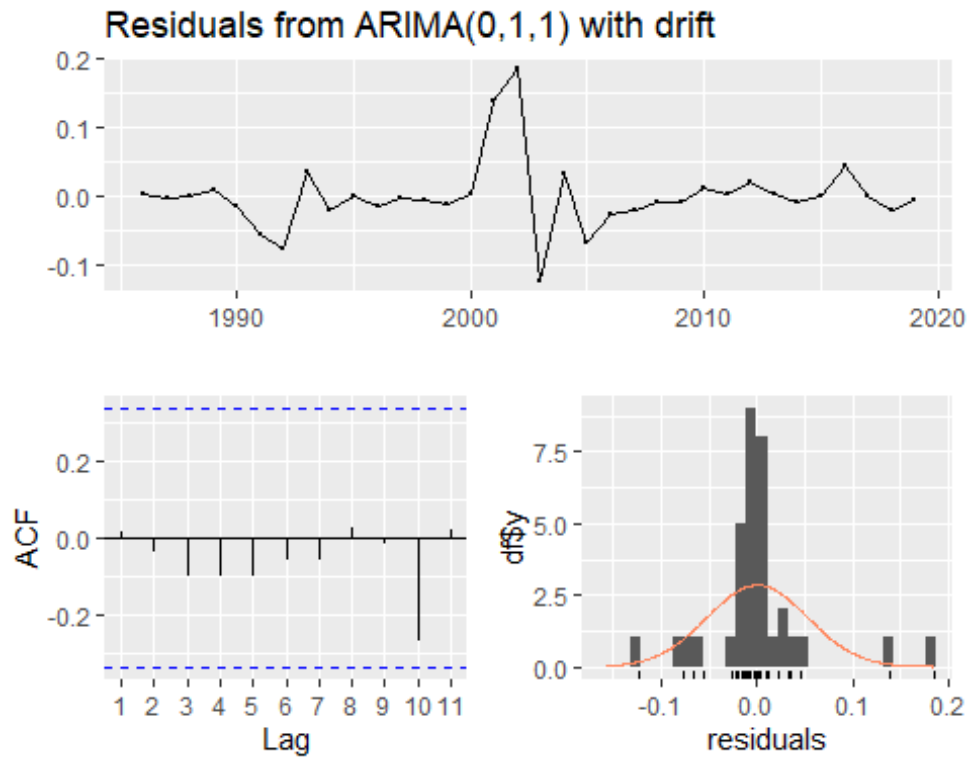
We use auto.arima() to get the best fitted model. Here the best fitted model is ARIMA(0,1,0) Note that: The RMSE value when the data is converted to a stationary series is less than the previous RMSE value where we directly forecasted the data.

now we will forecast the next 10 years: Forecasting:

```
fit%>% forecast(h=10)%>% autoplot(include=20)
```




```
checkresiduals(fit)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1) with drift
## Q* = 1.6111, df = 6, p-value = 0.9518
##
## Model df: 1.   Total lags used: 7
```