

# **MACT 4232**

# **Analysis of Time Series Data Project**

---

May 29th, 2024

**Sama Amr**  
900211296

**Mona Mahmoud**  
900212749



## Table of Contents

Table of Contents.....	1
Dataset Description.....	3
Purpose of Analysis.....	3
Data Analysis.....	4
Model Identification.....	5
Behavior of the process.....	8
Forecasting.....	8
Conclusion.....	9

## Dataset Description

This dataset, sourced from the U.S. Census Bureau, is hosted by the Federal Reserve Economic Database (FRED), which maintains and updates its information based on incoming data. It records the population data in the US. For national population data from 1900 to 1949, the figures exclude residents of Alaska and Hawaii. From 1940 to 1979, the data includes the resident population plus Armed Forces overseas. For all other years, only the resident population is covered.

We obtained the dataset from Kaggle via this link

<https://www.kaggle.com/datasets/census/population-time-series-data?select=POPH.csv>

The dataset contains 4 columns and 99 observations. The four variables are `realtime_start`, `value`, `date`, and `realtime_end`. The `value` variable records the population at the beginning of each year. The `date` variable shows the date of the observation and is recorded once at the beginning of every year from 1900 to 1999.

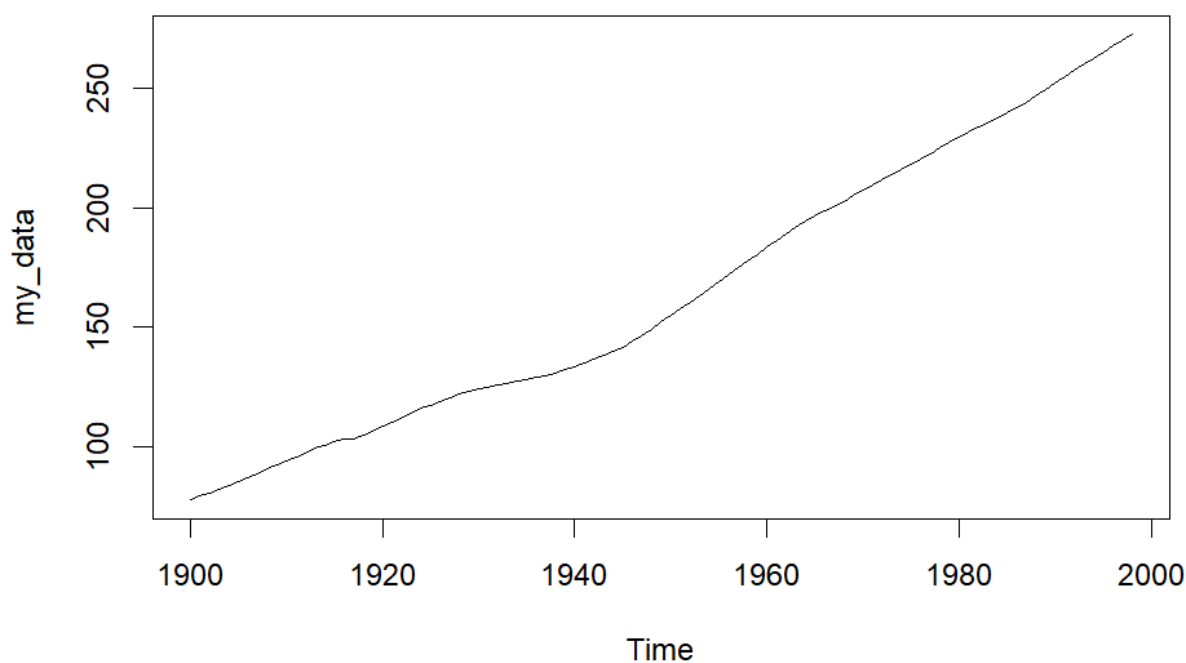
## Purpose of Analysis

The purpose of our analysis is to observe the population trend over the years in the US and forecast the population for the next 5 years. Forecasting the U.S. population offers significant benefits across various sectors, enhancing strategic planning and resource allocation. For government agencies, accurate population projections inform policy development and implementation in healthcare, education, infrastructure, and social services, ensuring resources are effectively distributed and future needs anticipated. Economically, population forecasts are crucial for labor market analysis, consumer demand predictions, housing market trends, and long-term growth strategies, guiding businesses in market analysis and strategic planning. Urban and regional planners rely on these forecasts to design sustainable environments, including housing, transportation, and public utilities, while healthcare planners use them to ensure

systems can meet future demands, manage public health, and address challenges like aging populations. In education, forecasting helps plan for school capacities, staffing, and funding needs. Overall, population forecasting supports informed decision-making and strategic planning across multiple domains, fostering a well-prepared and adaptable society.

## Data Analysis

We applied appropriate scaling for our data where the population is described in millions and rounded to the nearest 2 decimal places. For example, according to the original dataset the population in 1901 was 77584000, we scaled it to 77.58 million for easier interpretation.



The plot of the population data vs time shows an increasing trend for the population across the years. The increase in population can be a result of advances in modern medicine and accelerating migration. It seems there are no irregular fluctuations in the data and no major

important events that occurred over that time period and had an effect on the process. It is a smooth increase, however it doesn't look stationary, so we performed the Augmented Dickey Fuller unit root test to make sure that it's non-stationary. The unit root test's output illustrated a  $p\_value$  of 0.154 for the  $z.lag.1$  which is greater than  $\alpha = 0.05$  and a  $p\_value$  of 0.0893 for the  $tt$  which demonstrates that there is a trend in our series; thus, confirming our hypothesis at the beginning. Therefore, we decided to take the first difference and test again for stationarity. This time the results of the unit root test show that we got rid of the stationarity and trend. While the `adf.test` command provided a high  $p\_value > 0.05$ . By this step we can then proceed with the first-difference model. We tried both tests at different lag points 1, 2 and 3.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.638152	0.289878	2.201	0.0302 *
$z.lag.1$	-0.006137	0.004271	-1.437	0.1540
$tt$	0.014871	0.008659	1.717	0.0893 .
$z.diff.lag$	0.815750	0.059343	13.746	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3032 on 93 degrees of freedom  
Multiple R-squared: 0.79, Adjusted R-squared: 0.7833  
F-statistic: 116.6 on 3 and 93 DF,  $p$ -value: < 2.2e-16

#### Augmented Dickey-Fuller Test

data: my\_data  
Dickey-Fuller = -1.4371, Lag order = 1,  $p$ -value = 0.8096  
alternative hypothesis: stationary

```
[1] "Lag = 1"
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min       1Q   Median       3Q      Max
-1.42517 -0.14793 -0.02096  0.12771  1.18406

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.236388   0.103997   2.273  0.02535 *
z.lag.1      -0.172147   0.063229  -2.723  0.00775 **
tt           0.002363   0.001437   1.645  0.10348
z.diff.lag   -0.064844   0.103970  -0.624  0.53438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3075 on 92 degrees of freedom
Multiple R-squared:  0.0956, Adjusted R-squared:  0.06611
F-statistic: 3.242 on 3 and 92 DF,  p-value: 0.02563

Value of test-statistic is: -2.7226 2.5161 3.7071

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47

[1] "-----"

Augmented Dickey-Fuller Test

data: my_data_diff1
Dickey-Fuller = -2.7226, Lag order = 1, p-value = 0.2776
alternative hypothesis: stationary

[1] "Lag = 2"
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min       1Q   Median       3Q      Max
-1.42777 -0.16791 -0.03502  0.13340  1.17022

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.194093   0.106075   1.830  0.0706 .
z.lag.1      -0.140871   0.065310  -2.157  0.0337 *
tt           0.001992   0.001467   1.358  0.1778
z.diff.lag1  -0.107688   0.106247  -1.014  0.3135
z.diff.lag2  -0.186052   0.103488  -1.798  0.0756 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3053 on 90 degrees of freedom
Multiple R-squared:  0.1272, Adjusted R-squared:  0.08842
F-statistic: 3.279 on 4 and 90 DF,  p-value: 0.01474

Value of test-statistic is: -2.1569 1.6169 2.3262

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47

[1] "-----"

Augmented Dickey-Fuller Test

data: my_data_diff1
Dickey-Fuller = -2.1569, Lag order = 2, p-value = 0.5116
alternative hypothesis: stationary

[1] "Lag = 3"
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min       1Q   Median       3Q      Max
-1.4193 -0.1691 -0.0372  0.1412  1.1738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.192119   0.109073   1.761  0.0816 .
z.lag.1      -0.143470   0.067740  -2.118  0.0370 *
tt           0.002106   0.001518   1.388  0.1686
z.diff.lag1  -0.104321   0.112822  -0.925  0.3577
z.diff.lag2  -0.181743   0.107973  -1.683  0.0959 .
z.diff.lag3  0.009805   0.106497   0.092  0.9269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3085 on 88 degrees of freedom
Multiple R-squared:  0.1278, Adjusted R-squared:  0.07823
F-statistic: 2.579 on 5 and 88 DF,  p-value: 0.03174

Value of test-statistic is: -2.1179 1.5572 2.2445

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.04 -3.45 -3.15
phi2  6.50  4.88  4.16
phi3  8.73  6.49  5.47

[1] "-----"

Augmented Dickey-Fuller Test

data: my_data_diff1
Dickey-Fuller = -2.1179, Lag order = 3, p-value = 0.5277
alternative hypothesis: stationary
```

The ur command at all the 3 lags suggests that the series has already become stationary, yet the adf command suggests that the series is still not stationary. Therefore, we decided to proceed with the result from the adf command since it corresponds to our findings from the graph where the first difference is already enough to remove the trend from our data making it stationary.

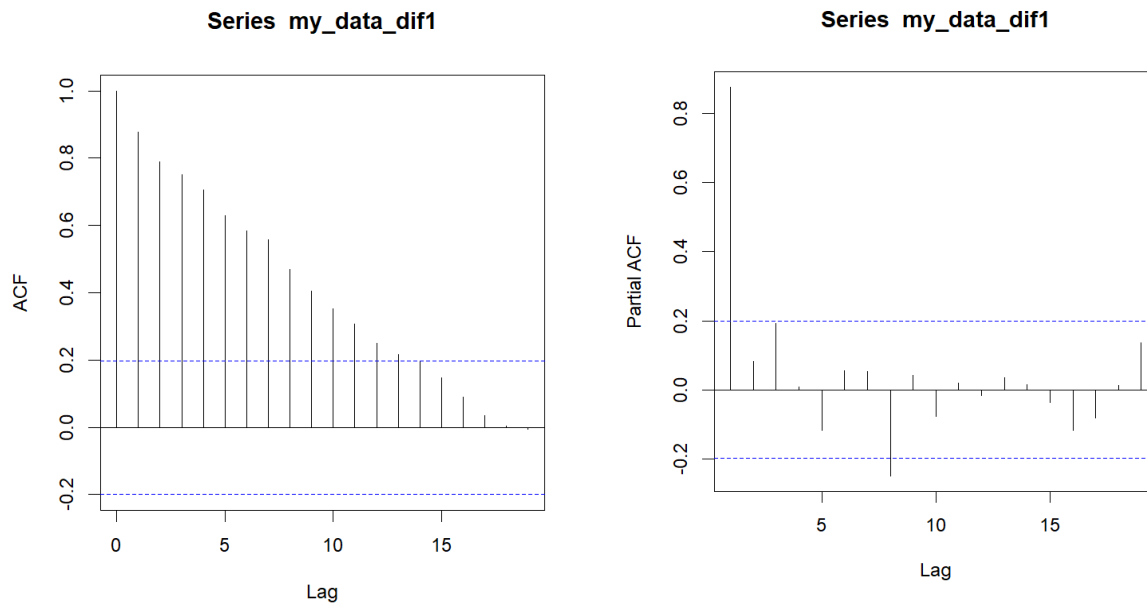
## Model Identification

After taking the first difference and testing for stationarity using the unit root tests, we verified that the data became stationary. When taking a glimpse of the data, and we can observe that it is



non-stationary since there is no clear pattern in time, we took the first difference for the data to then look as follows. That is simply because one can't proceed in the time series analysis to find the most suitable model without knowing for sure that the series is stationary, otherwise other steps are invalid. The difference made much of a difference where the data now has no pattern in time becoming stationary. Since the difference was used, this would identify an ARIMA model with  $d=1$ , yet we still need to identify the AR and MA parts. Next step was drawing the ACF and

PACF to identify if the other 2 parts exist or not and if they exist we identify their order. From the ACF plot we can see that it's decaying linearly. As for the PACF it cuts off after lag 1. Therefore, this would identify that we have an AR part of order 1, yet no AR part. Thus, reaching an ARIMA(1,1,0) where p,d,q are equal to 1,1,0, respectively.



The final model had a  $\sigma^2$  equal to 0.09 with an AIC= 59.04 and a negative log-likelihood with 2 estimated coefficients.

```
Call:
arima(x = my_data, order = c(1, 1, 0))
```

```
Coefficients:
      ar1
    0.9885
s.e.  0.0112
```

```
sigma^2 estimated as 0.09879:  log likelihood = -27.52,  aic = 59.04
```

$$X_t = 0.9885 * X_{t-1} + 0.09879$$

$$\phi = 0.9885$$

$$\varepsilon = 0.09879$$

## Model Comparison

We've also tried a model with MA of order 1 and the first difference just to make sure we were on the right track in terms of model selection.

```
call:
arima(x = my_data, order = c(0, 1, 1))
```

```
Coefficients:
          ma1
          0.9223
s.e.      0.0290
```

```
sigma^2 estimated as 1.312:  log likelihood = -153.32,  aic = 310.64
```

$$X_t = 0.9223 * \varepsilon_{t-1} + 1.312$$

$$\theta = 0.9223$$

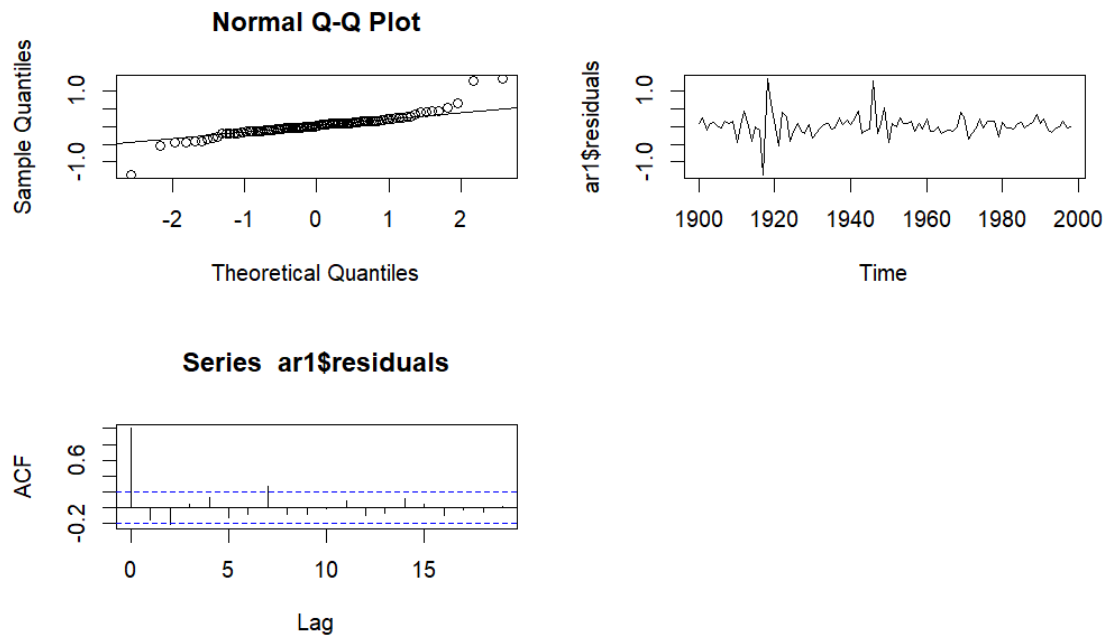
$$\varepsilon = 1.312$$

From the observed AIC, the AR model's AIC (59.04) was much less than the MA's AIC (310.64) indicating a better model. Therefore we proceeded with the AR model.

## Nice Assumptions

The following step was to check the residuals and validate the NICE assumptions. The 3 graphs below each demonstrate the NICE assumptions. For the Q-Q plot it shows that our data nearly follows a normal distribution which validates our first assumption. As for the ACF of the residuals, it shows a non-significant ACF with a stationary residual vs time plot which is what we are typically aiming for. For the residual vs time plot it suggests that the expectation is equal to zero with nearly a constant variance.





Finally, the Box-Pierce test was applied to see if the residuals were independent or not. The observed p-value which is more than 0.05 suggests that we fail to reject the  $H_0$  which means that error/ residuals are indeed independent.

#### Box-Pierce test

```
data: ar1$residuals
X-squared = 25.043, df = 19, p-value = 0.1591
```

## Behavior of the process

Since we're studying an ARIMA Model, we assured first that it's stationary by taking the first difference to ensure a better fit model. Thus, the next step was to observe the behavior of our data through plotting the ACF and PACF of the model which turned out to show an AR model of order 1. Moving on we examined the behavior of the residuals and validated the NICE assumptions, which show a no significant ACF. Lastly, we applied the box-pierce test which shows that the residuals are independent with an expectation of zero.

## Forecasting


The final step in our analysis was forecasting upcoming values to predict how they'd be. The following R output shows the forecast of the next 5 observations with their prediction intervals. Note that Point forecasts are in millions as previously mentioned.

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
1999	275.1019	274.6991	275.5047	274.4859	275.7180
2000	277.4861	276.5896	278.3827	276.1150	278.8573
2001	279.8429	278.3493	281.3364	277.5587	282.1271
2002	282.1725	279.9957	284.3493	278.8434	285.5017
2003	284.4753	281.5407	287.4100	279.9872	288.9635

As for the prediction interval it was measured using the following formula:

$\text{predicted value} \pm Z_{\alpha} \sqrt{v(\text{prediction error})}$  and to better illustrate the length of each of the interval produced we developed the following table, where it supports our hypothesis which is as

Lower 95%	Upper 95%	Difference
274.4859	275.7180	1.232053
276.1150	278.8573	2.742283
277.5587	282.1271	4.568409
278.8434	285.5017	6.658298
279.9872	288.9635	8.976362



The forecasted value is further away, the length of the interval is much bigger since we get more uncertain about our result. Getting to interpret our results relating them to real life, the forecasted points predict how much the population is estimated to increase in the following 5 years.

## Conclusion

To sum everything up, after analyzing our data, the most suitable model was found to be  $ARIMA(1,1,0)$  with both a difference and an AR part, yet no MA part. That was identified after plotting both the ACF and PACF graphs. The behavior of the model was then evaluated to validate the NICE assumptions related to the residuals which turned out to be all ok. Finally, we forecasted the population for the next 5 years to evaluate how the population is expected to grow over the following years.

