# Portland Public Transport 1960 – 1969

## 1. Introduction

With a successful and long history on public transport, TriMet – a Portland-based transport public company – builds one of America's most admired transit systems. The region is nationally recognized for transportation innovation and results. Portland's admirable success in developing such system is thanks to the willingness to try new implementations and to adapt when circumstances change.

The purpose of this study is to identify the relationships between the volume of population who use the Portland public transport system and time (in months) using an appropriate model – autoregressive integrated moving average (ARIMA). By doing this, these relationships can be represented by equations with numeric estimated parameters, so as to make sense of the problem in a quantitative way. The study can potentially apply into historical, geographical studies, civil engineering or other areas where decision making in public transport or urban development is necessary.

The method used in this study to build an appropriate model includes assessing stationarity of the data, test-fitting models (ARIMA and Seasonal Means), model diagnostics and forecasting analysis, in order to study the pattern of commuter in Portland at the time.

## 2. Data Analysis

### 2.1. Data Description

This study aims to study the available past data of the Portland Public Transport system, from January 1960 to January 1969 (Appendix A). This data set is collected from Kaggle, an online community where data scientists and machine learners find and share data sets. The data set contains monthly counts of riders for the Portland public transportation system from January 1960 through June 1969. It contains two variables: one represents the months and the other represents the corresponding total number of riders

using the Portland public transportation system. There is a total of 114 observations, in which 109 is used to create the statistical model and the remaining used to assess this model's forecasting function. Figure 1 illustrates the data which shows an overall increasing trend and possibly seasonal, since there seems to be an overall decrease from January to August and then increase again. The mean and variances of this dataset are both not constant over time.

The public transport in Portland is impressively efficient although it has been around for decades. The system is minimally invasive with nature and does not involve as much underground construction as other efficient public systems. Portland's public transport system is an example of a successfully developed system that has been around for years (with their street cars since as early as 1888), which also overcame the challenges posed by rivers and its hilly topography. Therefore, this study can be used in the development of transport systems for countries or places with similar geographical, demographical and economical properties with Portland during the years studied.

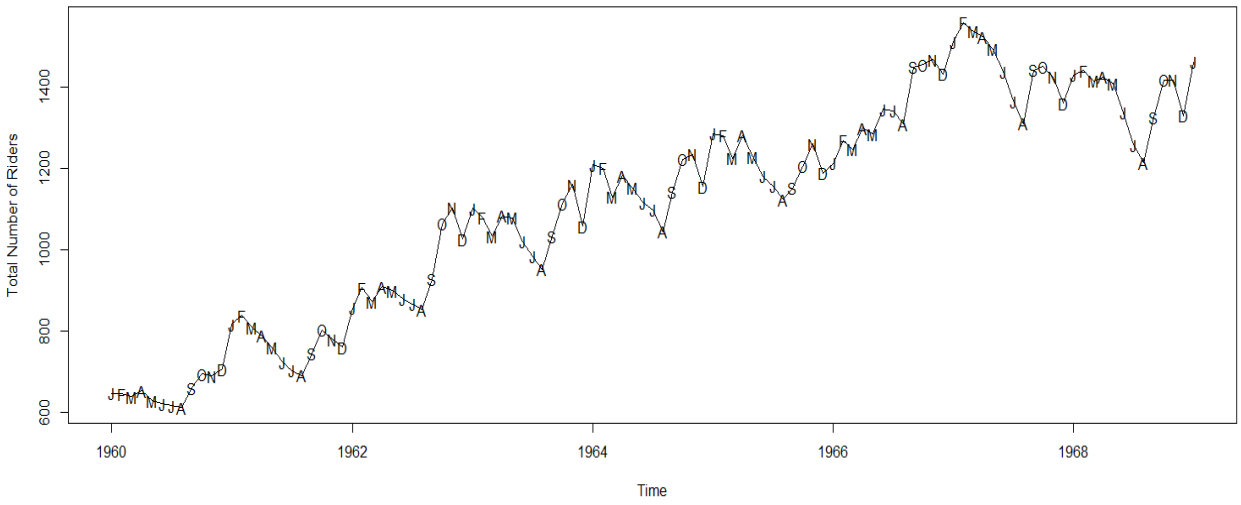


**Figure 1. Monthly total number of riders from January 1960 through June 1969**

### 2.2. Selected Model

For this data set, the ARIMA model is used to fit and forecast, specifically ARIMA (0, 1, 2). This is because ARIMA models are used in cases where data show evidence of non-stationarity, and differencing is involved to eliminate the non-stationarity.

The parameters estimated for this model are: $\boldsymbol{\theta_1} = 0.0216$ and $\boldsymbol{\theta_2} = 0.2394$

### 2.3. Selection Strategies

#### 2.3.1. Seasonal Means Model

Due to the potential seasonality of this dataset, the data is initially fitted using the Seasonal Means Model below (Table 1). Although this model has a high adjusted R-squared value of about 94% and low p-value for all variables, indicating a potential good fit, its standard error is very high. Normal Q-Q Plot shows a skewed plot at both tails which also do not fit the normal line well and a standard residual histogram that suggests this data do not meet the assumptions of normality (Figure 2), which is also confirmed by the p-value less than 0.05 in the Shapiro-Wilk Test (Table 2). Since the mean of this data set is not constant, the fit of the model, as illustrated by Figure 3, also fits adequately.

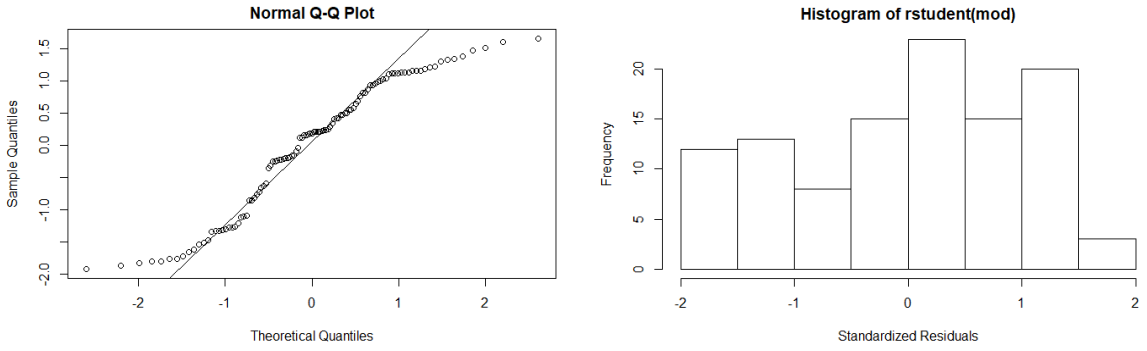


**Figure 2 – Q-Q Plot and Histogram of Standard Residuals for Seasonal Means Model**

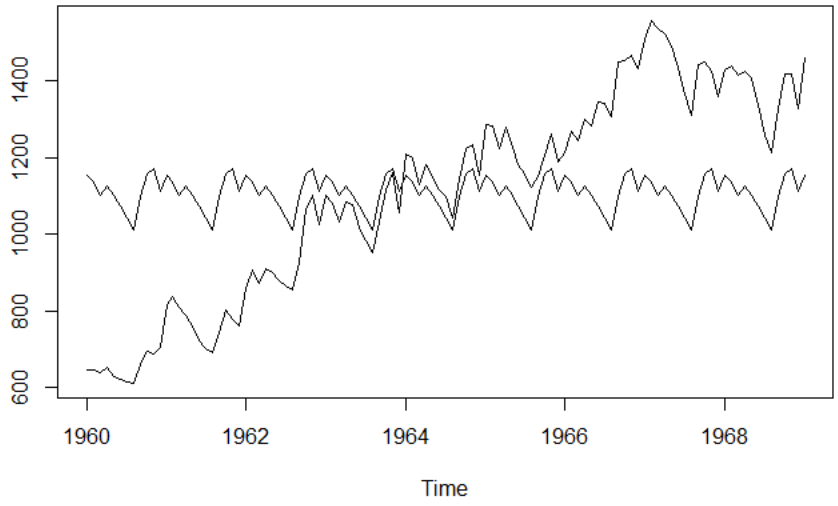


**Figure 3 – Fitted Data versus Modeled Data**

```
Call:
lm(formula = data ~ month. - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-505.10 -217.11   48.44  246.44  435.33

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
month.January     1153.10      88.85   12.98   <2e-16 ***
month.February    1135.56      93.65   12.12   <2e-16 ***
month.March       1100.67      93.65   11.75   <2e-16 ***
month.April       1127.11      93.65   12.04   <2e-16 ***
month.May         1103.56      93.65   11.78   <2e-16 ***
month.June        1073.56      93.65   11.46   <2e-16 ***
month.July        1043.33      93.65   11.14   <2e-16 ***
month.August      1012.22      93.65   10.81   <2e-16 ***
month.September   1097.33      93.65   11.72   <2e-16 ***
month.October     1158.22      93.65   12.37   <2e-16 ***
month.November    1170.56      93.65   12.50   <2e-16 ***
month.December    1112.78      93.65   11.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 281 on 97 degrees of freedom
Multiple R-squared:  0.9459,   Adjusted R-squared:  0.9393
F-statistic: 141.4 on 12 and 97 DF,  p-value: < 2.2e-16
```

**Table 1 – Seasonal Means Model**

```
Shapiro-Wilk normality test

data:  rstudent(mod)
W = 0.94163, p-value = 0.0001233
```

**Table 2 – Shapiro-Wilk Test for Seasonal Means Model**

### 2.3.2. ARIMA Model

A commonly used model to study time series and make forecasts in ARIMA. The first step in fitting an appropriate model is to assess the stationarity of the data and take differences if necessary, since the data needs to be stationary to be used to fit an ARIMA model. A time series is defined as stationary when its mean is constant over time and its covariance only depend on lag size and not lag time. To test whether this data is stationary, the Augmented Dickey-Fuller Test is used (Table 3). The results show a p-value of 0.2314 > 0.05 which leads us to not reject the null hypothesis that this time series is non-stationary.

```
Augmented Dickey-Fuller Test

data:  data
Dickey-Fuller = -2.8322, Lag order = 4, p-value = 0.2314
alternative hypothesis: stationary
```

**Table 3 – Augmented Dickey-Fuller Test 1**

To solve the problem of stationarity, a common method is to take the first difference by subtracting the previous observation from the current time series. By taking the difference, the time series now seems to have constant means and variances (Figure 4). Stationarity is also supported by the Augmented Dickey-Fuller Test applied to this time series, with p-value less than 0.05 (Table 4).

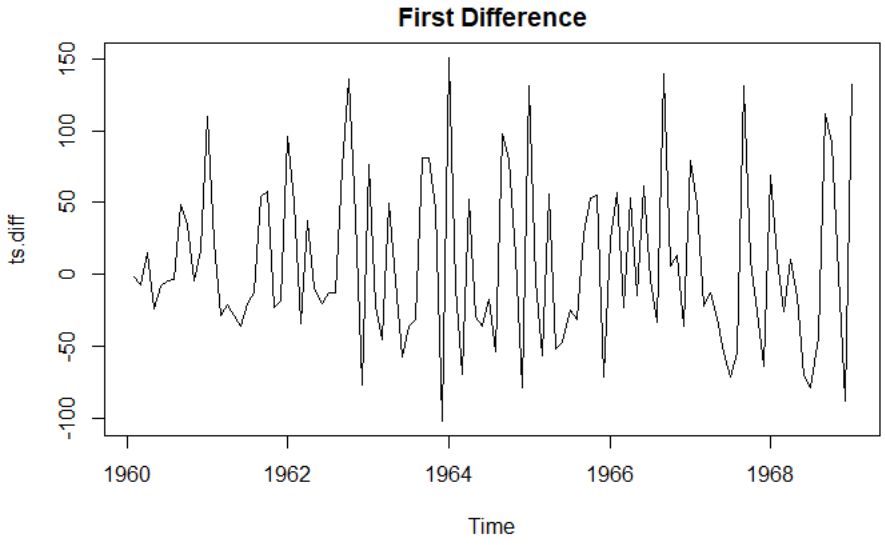


**Figure 4 – First Difference**

```
Augmented Dickey-Fuller Test

data:  ts.diff
Dickey-Fuller = -5.1091, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

**Table 4 – Augmented Dickey-Fuller Test 2**

Since the data is now stationary, we can proceed to fit the appropriate ARIMA (p, d, q) model by assessing the sample autocorrelation function (ACF), partial autocorrelation function (PACF) and extended autocorrelation function (EACF) (Figure 5). The original data's ACF and PACF plots suggest an AR (1) model since ACF dies of rather than cut off and PACF drops to zero after the first lag. Seasonality can also be a problem in this model because PACF is significant at lag 12 and 24. After taking the difference, both ACF and PACF show significantly high values at lag 12. ACF definitely displays a seasonal pattern because the value is consistently high at lag 12, 24, 36… EACF of the first difference suggests ARIMA (0, 1, 2) or (1, 1, 2).

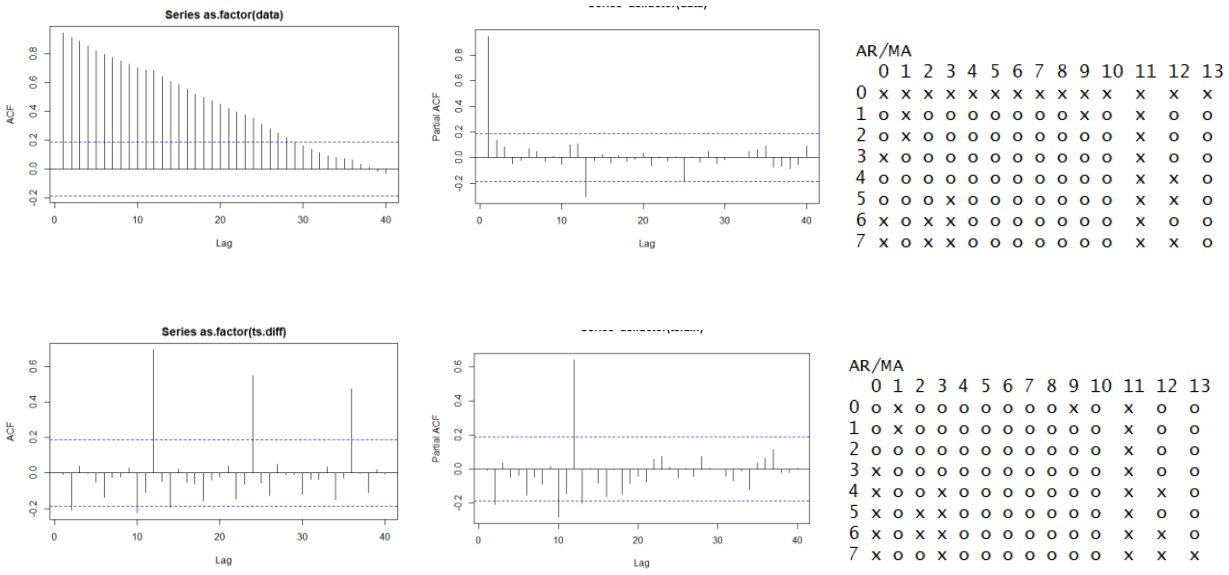


**Figure 5 – ACF, PACF, EACF Plots for Original Data (top) and First Difference (Left to Right)**

Fitting both ARIMA (0, 1, 2) and (1, 1, 2), we see that ARIMA (0, 1, 2) has better estimation error, and it is also advisable to use a lower level ARIMA model (Table 5).

```
Call:
arima(x = data, order = c(0, 1, 2), method = "ML")

Coefficients:
          ma1      ma2
      -0.0216  -0.2394
s.e.   0.0976   0.1007

sigma^2 estimated as 3151:  log likelihood = -588.29,  aic = 1180.59

Call:
arima(x = data, order = c(1, 1, 2), method = "ML")

Coefficients:
          ar1     ma1      ma2
      -0.1785  0.1487  -0.2387
s.e.   0.3732  0.3641   0.0970

sigma^2 estimated as 3145:  log likelihood = -588.2,  aic = 1182.39
```

**Table 5 – ARIMA Models**

Figure 6 shows a good fit of data using this model, suggested by the fitted values being very close to the observed values, indicated by the closeness of the 2 lines to each other.

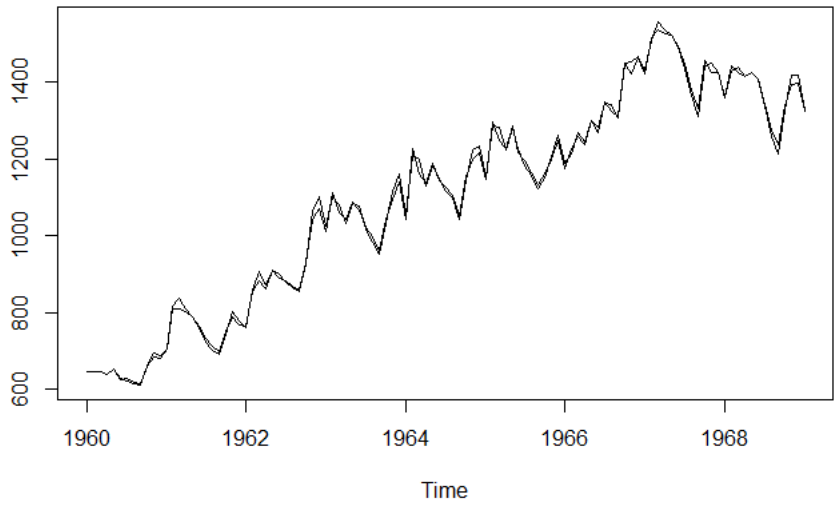


**Figure 6 – Fitted versus Observed Values**

Therefore, the model to be used for this study is ARIMA (0, 1, 2), with estimated parameters of the MA components $\boldsymbol{\theta_1} = 0.0216$ and $\boldsymbol{\theta_2} = 0.2394$.

The equation representing the model: $Y_t = Y_{t-1} - e_t - 0.0216e_{t-1} - 0.2394e_{t-2}$

### 2.4. Assessment of Performance

#### 2.4.1. Overfitting

To diagnose if a model is a good fit for a data set, we fit a more general model, of which the current model is a sub-model. In this case, we try to fit ARIMA (0, 1, 3) (Table 6). From the result, the estimate of the additional ma3 parameter is not significantly different from zero, and the standard error for the other two estimations increases. The estimates for parameter ma1 and ma2 also are not significantly different from the original model's estimates. As a result, we can conclude that ARIMA (0, 1, 2) is a sufficiently appropriate model for this time series**.**

```
Call:
arima(x = data, order = c(0, 1, 2), method = "ML")

Coefficients:
            ma1      ma2
      -0.0216  -0.2394
s.e.   0.0976   0.1007

sigma^2 estimated as 3151:  log likelihood = -588.29,  aic = 1180.59

Call:
arima(x = data, order = c(0, 1, 3), method = "ML")

Coefficients:
          ma1      ma2     ma3
      -0.0295  -0.2310  0.0527
s.e.   0.0976   0.1012  0.1120

sigma^2 estimated as 3144:  log likelihood = -588.18,  aic = 1182.37
```

**Table 6 – Overfitting Results**

#### 2.4.2. Residual Analysis

We assume that if we prescribe the appropriate model, the residuals are normally distributed with zero mean and a constant standard deviation. Therefore, to test if our model meets this assumption, we need to look at the Normal Q-Q Plots and the histogram of residuals for this model (Figure 7) and the Shapiro-Wilk Normality Test again (Table 7). Normal Q-Q Plot shows a prominent curve at the middle of the normal line, and the observations are skewed at the tail. The histogram of the residuals is also imbalanced to the left of the plot. Shapiro-Wilk Test gives a p-value much less than 0.05. Thus, we can conclude that

the normality assumption is not met for this dataset. However, similar to many other real-life data, the underlying distribution of residuals can be difficult to obtain, and thus we might have to proceed our study without an assumed distribution to find the best results possible.

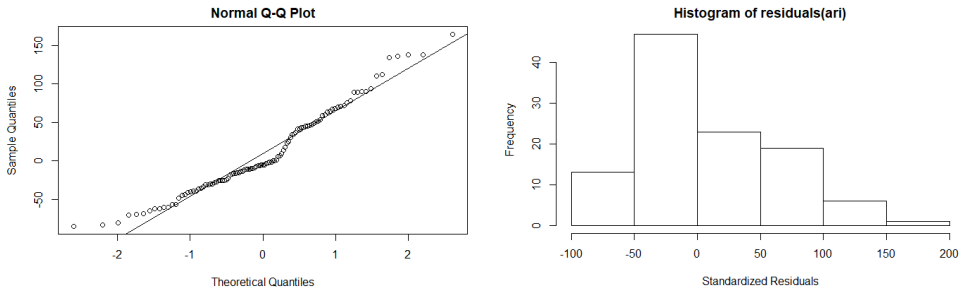


**Figure 7 – Normal Q-Q Plot and Histogram of Standard Residuals for ARIMA (0, 1, 2)**

```
Shapiro-Wilk normality test

data:  residuals(ari)
W = 0.9587, p-value = 0.001897
```

**Table 7 – Shapiro-Wilk Normality Test for ARIMA (0, 1, 2)**

By observing the three-plot diagnostic plots from R's output (Figure 8) which includes plots of standardized residuals, sample ACF of residuals, and p-values for the Ljung-Box test statistic, we can draw additional conclusions about our data. The residuals of this model seem to stay close to zero, with a few of them above 2 and one observation at year 1964 above 3, suggesting a potential outlier in the data. The ACF plot shows a high autocorrelation at lag 12, suggesting a seasonal component, as expected from earlier. For the Ljung-box statistics plot, all of the p-values are high and way above the significant line at 5%. Thus, the residual terms are independent.

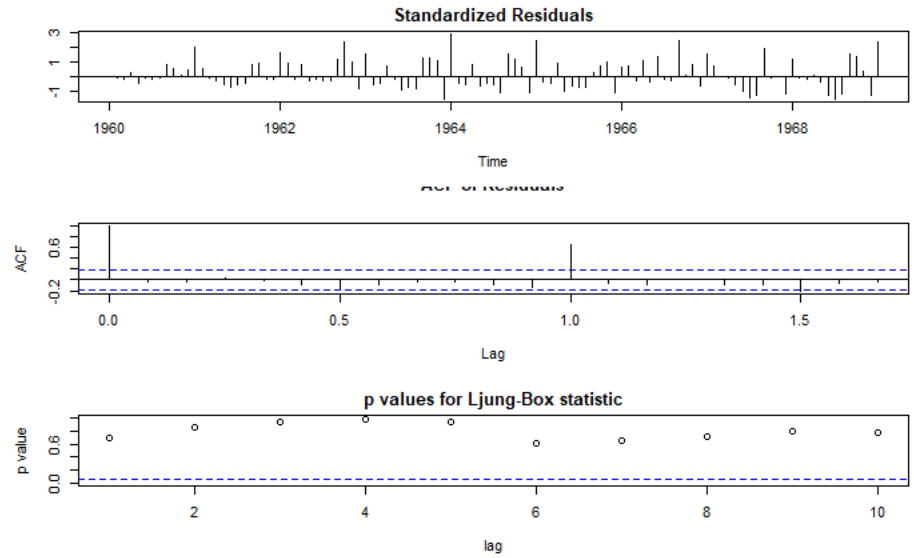


**Figure 8 – Tsdiag ouput**

### 2.5. Forecasting

Using the last five observations in this data for prediction, the below result is obtained (Table 8).

| | p1$lpi | p1$pred | p1$upi | forecast$Ridership |
|---|---|---|---|---|
| Feb 1969 | 1364.515 | 1474.531 | 1584.547 | 1425 |
| Mar 1969 | 1288.043 | 1441.957 | 1595.871 | 1419 |
| Apr 1969 | 1267.888 | 1441.957 | 1616.026 | 1432 |
| May 1969 | 1249.835 | 1441.957 | 1634.078 | 1394 |
| Jun 1969 | 1233.340 | 1441.957 | 1650.574 | 1327 |

**Table 8 – Forecasting Results**

The forecast is accurate for the first three months of the forecasting data but is far from the true number of ridership in May and June 1969 of the forecasting data. However, in May and June, the real observations see an unexpected drop in the number of riders, which can be outliers (as supported by Figure 1). Therefore, the model might not have taken into account these outliers. However, this model still predicts relatively accurately because all the predictions lie within 95% confidence intervals. Figure visualizes the prediction interval, which shows a reasonable range if we compare it to Figure 1.

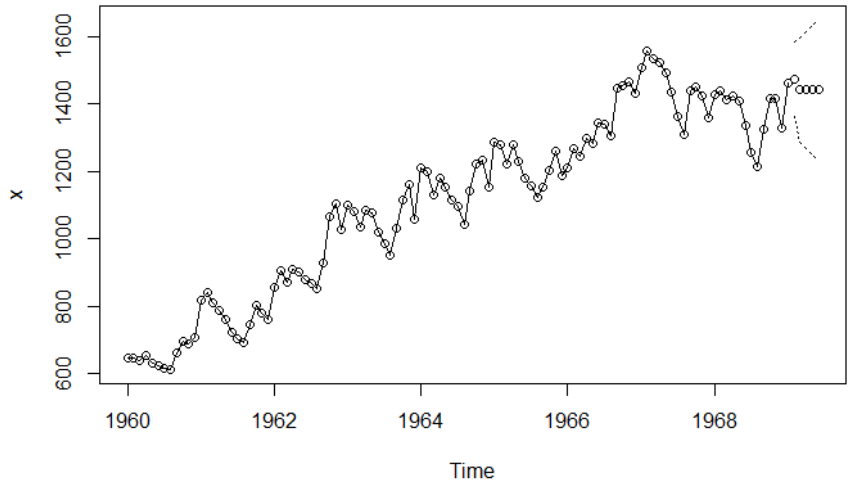


**Figure 9 – Forecasting results of ARIMA (0,1,2)**

### 2.6. Final Recommendation

The ARIMA model, while takes care of non-stationarity of the data, does not take into consideration the seasonal feature of the data (as illustrated in the ACF plot in Figure 5). It is observable that the data has some seasonal trend, which can be expected because the number of people using public transport can be affected by weather and seasons. However, Seasonal Means Model is also not applicable in this case because the data's normality assumption is strongly violated and data mean and variances are not constant, the fit of this model is also poor. Therefore, a more appropriate model to be used in this case is a Seasonal ARIMA (SARIMA) model, that tackles both the non-stationarity and seasonal components. However, this model is beyond the purpose of the current project, thus the best model we can fit is the ARIMA (0, 1, 2) model.

## 3. Conclusion

In conclusion, ARIMA (0, 1, 2) is the best option for this study to forecast the monthly number of Portland's public transport riders based on the outcome of model-fitting and forecasting. However, keeping the consideration that this data is historically far back (1960-1969), it might not be suitable to use for Portland right now and might need modifications. However, the study can still be useful for similar problems elsewhere.

The use of this model is also limited since it does not cover the seasonality aspect of the data and does not meet the normality assumptions for its residuals. A better model for further study would be a seasonal ARIMA model.

**Appendix A: Portland's monthly total number of riders from January 1960 through June 1969**

| | Month | Ridership | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | 1960-01 | 648 | **39** | 1963-03 | 1034 | **78** | 1966-06 | 1345 |
| **2** | 1960-02 | 646 | **40** | 1963-04 | 1083 | **79** | 1966-07 | 1341 |
| **3** | 1960-03 | 639 | **41** | 1963-05 | 1078 | **80** | 1966-08 | 1308 |
| **4** | 1960-04 | 654 | **42** | 1963-06 | 1020 | **81** | 1966-09 | 1448 |
| **5** | 1960-05 | 630 | **43** | 1963-07 | 984 | **82** | 1966-10 | 1454 |
| **6** | 1960-06 | 622 | **44** | 1963-08 | 952 | **83** | 1966-11 | 1467 |
| **7** | 1960-07 | 617 | **45** | 1963-09 | 1033 | **84** | 1966-12 | 1431 |
| **8** | 1960-08 | 613 | **46** | 1963-10 | 1114 | **85** | 1967-01 | 1510 |
| **9** | 1960-09 | 661 | **47** | 1963-11 | 1160 | **86** | 1967-02 | 1558 |
| **10** | 1960-10 | 695 | **48** | 1963-12 | 1058 | **87** | 1967-03 | 1536 |
| **11** | 1960-11 | 690 | **49** | 1964-01 | 1209 | **88** | 1967-04 | 1523 |
| **12** | 1960-12 | 707 | **50** | 1964-02 | 1200 | **89** | 1967-05 | 1492 |
| **13** | 1961-01 | 817 | **51** | 1964-03 | 1130 | **90** | 1967-06 | 1437 |
| **14** | 1961-02 | 839 | **52** | 1964-04 | 1182 | **91** | 1967-07 | 1365 |
| **15** | 1961-03 | 810 | **53** | 1964-05 | 1152 | **92** | 1967-08 | 1310 |
| **16** | 1961-04 | 789 | **54** | 1964-06 | 1116 | **93** | 1967-09 | 1441 |
| **17** | 1961-05 | 760 | **55** | 1964-07 | 1098 | **94** | 1967-10 | 1450 |
| **18** | 1961-06 | 724 | **56** | 1964-08 | 1044 | **95** | 1967-11 | 1424 |
| **19** | 1961-07 | 704 | **57** | 1964-09 | 1142 | **96** | 1967-12 | 1360 |
| **20** | 1961-08 | 691 | **58** | 1964-10 | 1222 | **97** | 1968-01 | 1429 |
| **21** | 1961-09 | 745 | **59** | 1964-11 | 1234 | **98** | 1968-02 | 1440 |
| **22** | 1961-10 | 803 | **60** | 1964-12 | 1155 | **99** | 1968-03 | 1414 |
| **23** | 1961-11 | 780 | **61** | 1965-01 | 1286 | **100** | 1968-04 | 1424 |
| **24** | 1961-12 | 761 | **62** | 1965-02 | 1281 | **101** | 1968-05 | 1408 |
| **25** | 1962-01 | 857 | **63** | 1965-03 | 1224 | **102** | 1968-06 | 1337 |
| **26** | 1962-02 | 907 | **64** | 1965-04 | 1280 | **103** | 1968-07 | 1258 |
| **27** | 1962-03 | 873 | **65** | 1965-05 | 1228 | **104** | 1968-08 | 1214 |
| **28** | 1962-04 | 910 | **66** | 1965-06 | 1181 | **105** | 1968-09 | 1326 |
| **29** | 1962-05 | 900 | **67** | 1965-07 | 1156 | **106** | 1968-10 | 1417 |
| **30** | 1962-06 | 880 | **68** | 1965-08 | 1124 | **107** | 1968-11 | 1417 |
| **31** | 1962-07 | 867 | **69** | 1965-09 | 1152 | **108** | 1968-12 | 1329 |
| **32** | 1962-08 | 854 | **70** | 1965-10 | 1205 | **109** | 1969-01 | 1461 |
| **33** | 1962-09 | 928 | **71** | 1965-11 | 1260 | **110** | 1969-02 | 1425 |
| **34** | 1962-10 | 1064 | **72** | 1965-12 | 1188 | **111** | 1969-03 | 1419 |
| **35** | 1962-11 | 1103 | **73** | 1966-01 | 1212 | **112** | 1969-04 | 1432 |
| **36** | 1962-12 | 1026 | **74** | 1966-02 | 1269 | **113** | 1969-05 | 1394 |
| **37** | 1963-01 | 1102 | **75** | 1966-03 | 1246 | **114** | 1969-06 | 1327 |
| **38** | 1963-02 | 1080 | **76** | 1966-04 | 1299 | | | |
| | | | **77** | 1966-05 | 1284 | | | |