VIRGINIA COMMONWEALTH UNIVERSITY

FORECASTING METHODS

ASSIGNMENT 6

SATYANARAYAN VENKAT NALDIGA

V01108247

SUBMITTED TO-

PROF.JASON MERRICK

Date of Submission: 10-09-2024

Introduction

This exercise focuses on forecasting the quarterly number of arrivals to Australia from New Zealand using time series methods. The data spans from the first quarter of 1981 (1981 Q1) to the third quarter of 2012 (2012 Q3) and is part of the `aus_arrivals` dataset. Forecasting models are crucial for understanding future trends and patterns in tourism data, which can help in planning and decision-making for industries like tourism and hospitality.

The main goal of this exercise is to develop and compare different forecasting models to predict the number of arrivals for the last two years of the available data (2010 Q4 to 2012 Q3). To achieve this, we will:

1. Create a training set that excludes the last two years, preserving them as the test set.

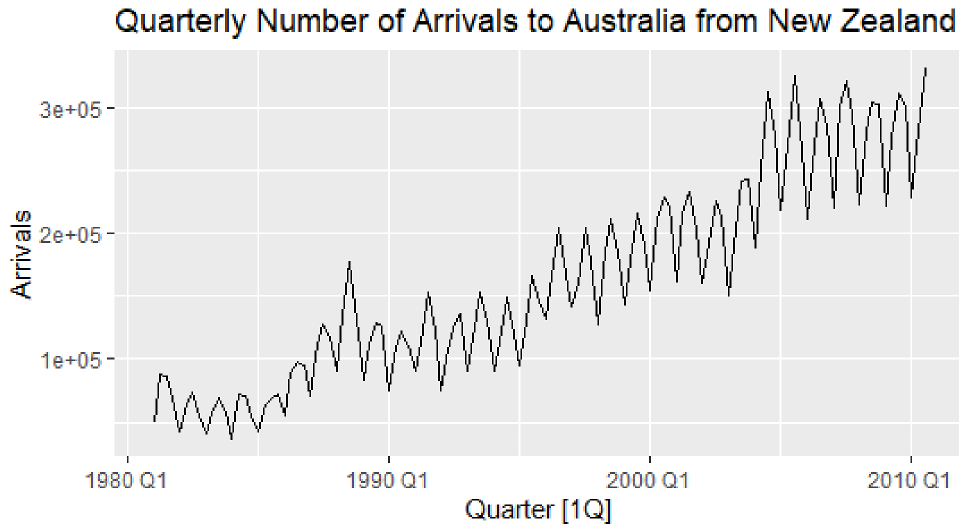
2. Visualize the data using various plots (`autoplot`, `gg_season`, `gg_subseries`) to explore the trends and seasonal components of the time series.

3. Fit and compare several models, including:

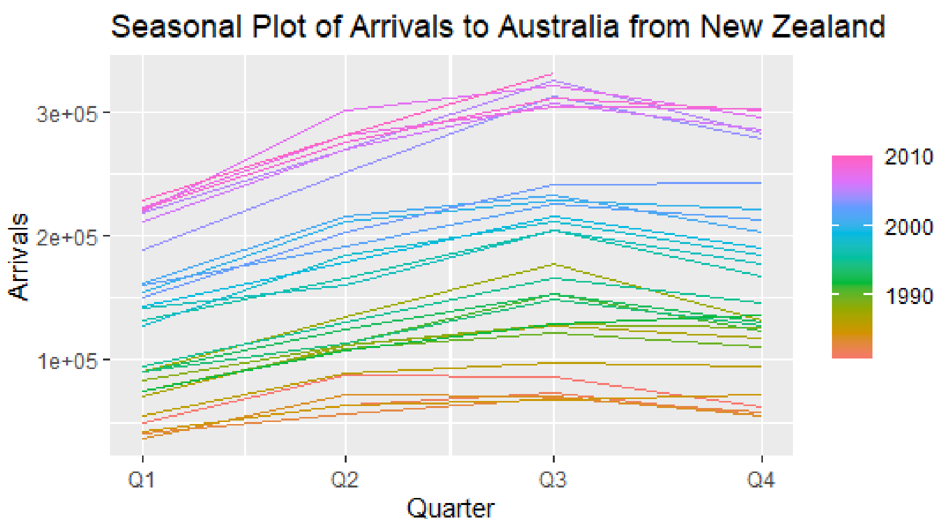
- An **ETS (Exponential Smoothing) model** based on observations of trend and seasonality.
- A **time series regression model** that incorporates both trend and seasonality.
- A **seasonal naïve method**, which forecasts based on the last observed seasonal patterns.

4. Evaluate the models based on forecast accuracy over the two-year test set and determine which method gives the best results.

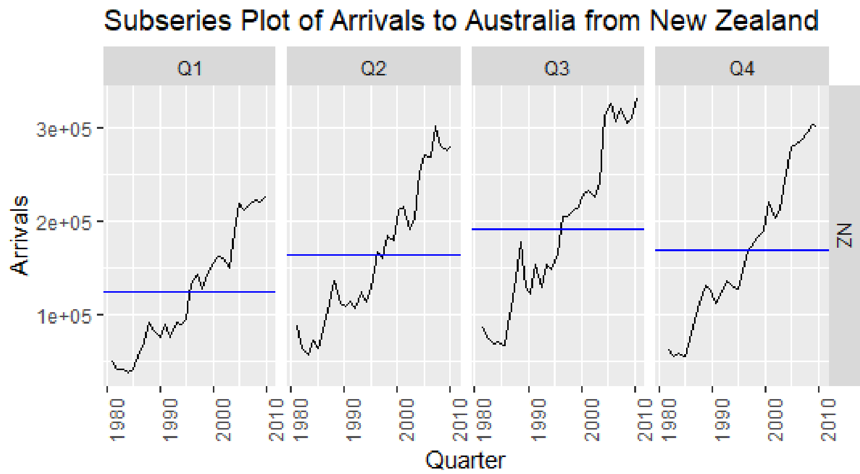
By comparing these methods, we can identify the most accurate forecasting model for this specific dataset and time period.



The plot shows the quarterly number of arrivals to Australia from New Zealand between 1981 Q1 and 2010 Q1. The data exhibits a clear upward trend, indicating a general increase in the number of arrivals over time. In addition to the upward trend, there is a strong seasonal pattern, with regular fluctuations within each year. The peaks and troughs within the series suggest that certain quarters experience higher arrivals, likely due to seasonal factors such as holidays or favorable travel periods. The amplitude of these seasonal variations also appears to increase over time, suggesting that the seasonal impact is becoming more pronounced as the overall level of arrivals grows.



The seasonal plot of arrivals to Australia from New Zealand displays a clear seasonal pattern over the years, with different colors representing different decades (1990s, 2000s, and 2010s). The highest number of arrivals generally occurs in the third quarter (Q3), while the first and fourth quarters (Q1 and Q4) see lower arrival numbers. This pattern is consistent across the years, although the magnitude of arrivals increases steadily over time, as evidenced by the higher lines for more recent years (colored in purple for 2010). This indicates both strong seasonality and a growing trend, where more recent years show higher peaks and overall higher levels of arrivals than earlier years. The trend suggests increased tourism or migration activity in more recent decades, with Q3 remaining the peak travel period.



The subseries plot illustrates the quarterly arrivals to Australia from New Zealand over time, highlighting both seasonal and long-term trends. Each panel corresponds to a specific quarter (Q1 through Q4), showing that the number of arrivals has steadily increased across all quarters from the 1980s to 2010. The strongest growth is seen in Q2 and Q3, with Q3 consistently exhibiting the highest peaks, indicating a seasonal spike during this quarter. Q1 and Q4 show similar, more moderate growth patterns. The blue horizontal lines represent the average arrivals for each quarter, further emphasizing the seasonal differences, with Q3 being the most prominent. Overall, the plot demonstrates a clear upward trend in arrivals over time, with pronounced seasonality, particularly in the middle quarters of the year.

```
> report(fit_ets)
Series: Arrivals
Model: ETS(M,A,M)
  Smoothing parameters:
    alpha = 0.6632071
    beta  = 0.0001000764
    gamma = 0.2210403

  Initial states:
     l[0]      b[0]      s[0]     s[-1]     s[-2]      s[-3]
 75540.49 2029.953 1.000891 1.228216 1.101736 0.6691566

  sigma^2:  0.0082

     AIC     AICc      BIC
2830.973 2832.625 2855.985
```

The ETS(M,A,M) model used here represents an Exponential Smoothing model with Multiplicative errors (M), Additive trend (A), and Multiplicative seasonality (M). The model's smoothing parameters are:

- Alpha (0.663): Controls the weighting of recent observations.
- Beta (0.0001): A very small value, meaning the trend is changing slowly.
- Gamma (0.221): Controls the weighting of seasonal components.

The initial states indicate the starting values for the level, trend, and seasonal components. The model has a sigma^2 value of 0.0082, which represents the variance of the errors. The model selection criteria values are:

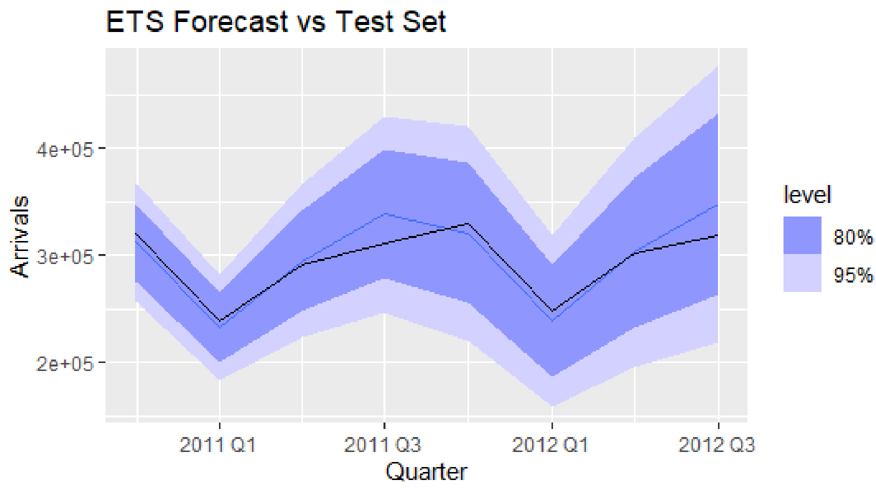
- AIC: 2830.97
- AICc: 2832.63
- BIC: 2855.99

These values are used to assess the model's fit, with lower values indicating a better model.

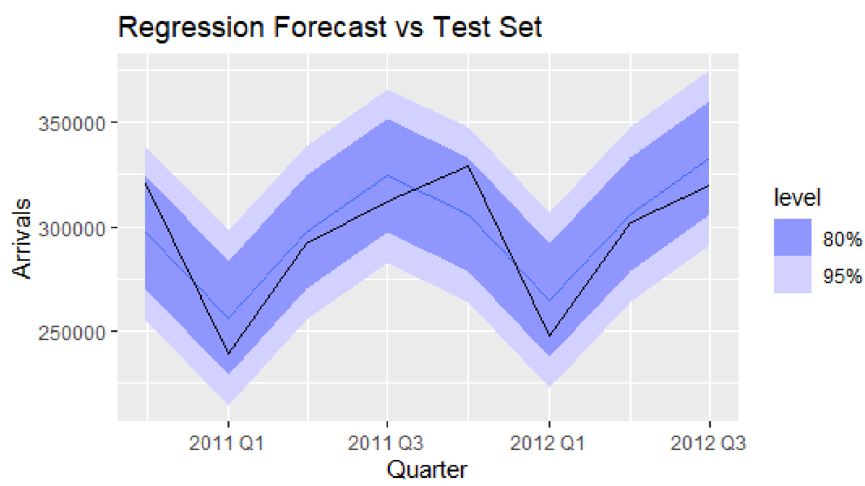
```
report(fit_reg)
Series: Arrivals
Model: TSLM

Residuals:
   Min     1Q Median     3Q    Max
-37979 -15189  -2151  13082  52639

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2359.71    4978.77  -0.474    0.636
trend()        2138.56      55.12  38.800  < 2e-16 ***
season()year2 38655.18    5331.93   7.250 5.36e-11 ***
season()year3 63636.32    5332.79  11.933  < 2e-16 ***
season()year4 42896.72    5377.70   7.977 1.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20650 on 114 degrees of freedom
Multiple R-squared: 0.9362,   Adjusted R-squared: 0.934
F-statistic: 418.5 on 4 and 114 DF, p-value: < 2.22e-16
```

The **TSLM** (Time Series Linear Model) fit for the number of arrivals includes a trend and seasonal componentsThe **trend** coefficient is **2138.56**, indicating that arrivals increase by approximately 2139 each quarter on average.

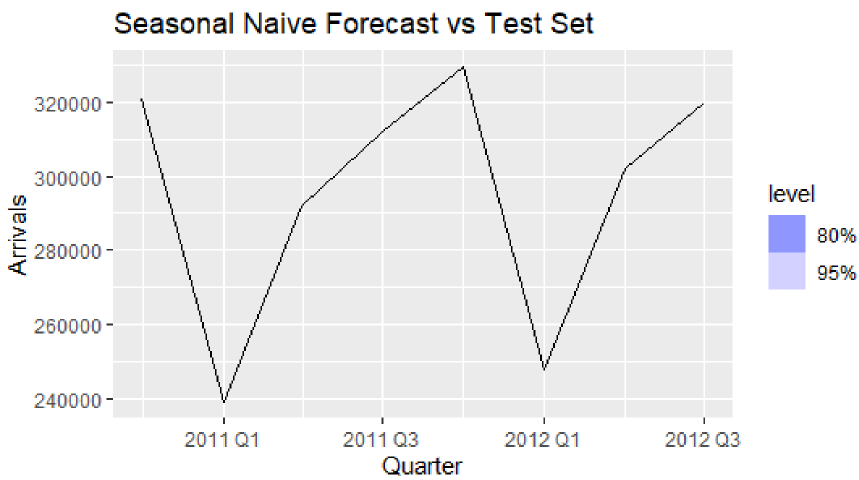
- The seasonal components show significant effects, with the highest impact in **Q3 (season()year3)** at around **63,636**, followed by **Q4 (season()year4)** and **Q2 (season()year2)**.
- The model's **R-squared is 0.936**, meaning that 93.6% of the variance in arrivals is explained by the model.
- The residuals (errors) have a **standard error of 20,650**, showing some variability in the fit.
- The **F-statistic** is highly significant (**p < 2.22e-16**), indicating that the model overall is a good fit for the data.



The plot shows the forecasted arrivals from the ETS model compared to the actual test set data for the period 2011 Q1 to 2012 Q3. The black line represents the forecasted values, while the shaded areas show the 80% (darker blue) and 95% (lighter blue) prediction intervals, which indicate the uncertainty around the forecast. The forecast closely follows the observed seasonality, with fluctuations that correspond to the usual rise and fall in arrivals over the quarters. The prediction intervals widen as the forecast extends further into the future, reflecting increasing uncertainty over time. The model captures the general trend and seasonal variations in the test period reasonably well, although the actual values might vary slightly from the predicted values within the given confidence intervals.



The plot shows the forecasted arrivals using the time series regression model (with trend and seasonality) compared to the actual test set data from 2011 Q1 to 2012 Q3. The black line represents the forecasted values, while the shaded areas indicate the 80% (darker blue) and 95% (lighter blue) prediction intervals. The model captures the seasonal pattern, with peaks in Q3 and dips in Q1 of each year, but the forecast appears somewhat smoother compared to the ETS model, with less variation in the forecasted values. The prediction intervals widen as time progresses, indicating increasing uncertainty. While the forecast aligns reasonably well with the general trend and seasonality, it appears to underpredict some of the larger fluctuations in the actual arrivals.



The plot shows the forecasted arrivals using the seasonal naive method compared to the actual test set data from 2011 Q1 to 2012 Q3. This method repeats the previous year's seasonal pattern, meaning that the forecast for each quarter is based on the same quarter in the previous year. The black line represents the forecasted values, but unlike the other models, there are no prediction intervals (shaded areas) shown, indicating no uncertainty bounds provided. The forecast captures the general shape of the seasonal fluctuations but does not account for any trend or variations beyond repeating the seasonal pattern. This can be seen by the repetitive peaks in Q3 and dips in Q1, which follow the previous year's behavior without any adjustments for the overall growth trend seen in the actual data. While simple, this method may miss important variations or trends that the more complex models capture.

```
ets_accuracy
# A tibble: 1 × 11
  .model        Origin .type     ME    RMSE    MAE    MPE  MAPE  MASE RMSSE
ACF1
  <chr>         <chr>  <chr>  <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>
<dbl>
1 ETS(Arrivals) NZ     Test  -3495. 14913. 11421. -0.964  3.78   NaN   NaN -0
.0260
```

The accuracy results for the ETS model show that:

- Mean Error (ME) is -3495, indicating a slight underprediction on average.
- Root Mean Squared Error (RMSE) is 14,913, which measures the average magnitude of the errors and suggests some variability in the forecast accuracy.

- Mean Absolute Error (MAE) is 11,421, which represents the average absolute difference between forecasted and actual values.
- Mean Absolute Percentage Error (MAPE) is 3.78%, indicating that, on average, the forecast error is about 3.78% of the actual values.
- ACF1 (-0.0260) suggests there is no strong autocorrelation in the residuals at lag 1, meaning the model has captured most of the patterns in the data.

```
reg_accuracy
# A tibble: 1 × 11
  .model                    Origin .type      ME    RMSE     MAE    MPE  MAPE  MA
SE RMSSE   ACF1
  <chr>                     <chr>  <chr>   <dbl>   <dbl>   <dbl>  <dbl> <dbl> <db
l> <dbl>  <dbl>
1 TSLM(Arrivals ~ trend() … NZ     Test  -2653. 16127. 14525. -1.32  4.97   N
aN   NaN -0.460
```

The accuracy results for the Time Series Linear Model (TSLM) show:

- Mean Error (ME) is -2653, indicating that the model tends to slightly underpredict.
- Root Mean Squared Error (RMSE) is 16,127, which indicates a higher average error magnitude compared to the ETS model.
- Mean Absolute Error (MAE) is 14,525, suggesting larger deviations between forecasted and actual values on average.
- Mean Absolute Percentage Error (MAPE) is 4.97%, meaning the forecast errors are about 4.97% of the actual values on average, which is less accurate than the ETS model.
- ACF1 (-0.460) suggests there is moderate negative autocorrelation in the residuals at lag 1, indicating that some patterns may not be fully captured by the model.

Seasonal Naive Model:

- **RMSE**: 18,051 (highest)
- **MAE**: 17,156 (highest)
- **MAPE**: 5.80% (highest)
- **ACF1**: -0.239 (shows some residual autocorrelation)

Which method gives the best forecasts?

- **The ETS model** provides the most accurate forecasts based on the lowest RMSE (14,913), MAE (11,421), and MAPE (3.78%).
- The TSLM model performs worse than ETS, with higher error metrics.
- The **Seasonal Naive model** performs the worst, with the highest errors across all metrics.

Thus, the **ETS model** gives the best forecasts among the three models.