Laboratory 3 Forecasting with Regression Including Smoothing Techniques such as Moving Averages and Exponential Smoothing

Purpose and Alignments Written Report

ANA 535 Forecasting

Name KOHEI NISHITANI Student ID# 1122867

Date May 2nd , 2025

Table of Contents

Introduction

Lab3 shifts its focus from diagnosing seasonality and stationery to producing usable forecast for Amtrak dataset. Building the foundation of periodicity insights in Lab2, this Lab3 exercise will enable researcher to familiarize time-series regression, moving-average smoothing, ETS and Holt-winters smoothing. Those methodologies translate observed patterns into predictable patterns. The workflow at this lab begins with exploratory data analysis. The target variable passenger-miles are inspected to see if it has normal distribution or any other skewness, then it go through log-transformation because of observed skewness. After this preliminary inspection, a tslm() regression spanning 1991-2024 establishes a first fitted-versus-actual benchmark. Then showing how model fitting can be changed with different training windows —1991-1997, 1991-2004, 1991-2016, and 1991-2020—to illustrate how structural breaks, most apparent the 2020 COVID-19 collapse, Then demonstrate other methodologies like ETS and Holt-Winter's model.

Background

The previous lab exercise established all necessary data cleaning and exploratory analysis to ensure reliable forecasting results. We cleaned the Amtrak passenger-miles data, sorting, filtering and then analyzed it using frequency-domain methods including Fast Fourier Transform and periodograms and STL decomposition to reveal both strong annual patterns. Consistent patterns were successfully identified and eliminated in periodic data during Lab 2, but that exercise did not proceed to generate meaningful forecasted results. The absence of regression analysis and smoothing techniques as well as accuracy evaluation means decision-making professionals still need quantitative projections together with uncertainty understanding for future demand.

This Lab 3 closes that gap. The exercise applies time-series regression and moving-average filters alongside Holt-Winters, ETS, exponential-smoothing models to generate forecasts from the seasonally adjusted dataset produced in Lab 2. The statistical evaluation comparing multiple prediction periods before and after the COVID pandemic with multiple different training period and the use of basic forecasting methods in Lab 3 supplies crucial information for Amtrak managers to make schedule changes based on changing travel patterns.

Data

Data from Amtrak source enables the exploration of time-series analysis periodicity finding and adjustment through this exercise. The Amtrak dataset covers a time span from January 1991 to June 2024 holding about four hundred monthly records. The records contain a month field along with Ridership, PassengerMiles and RidersReported fields. For modelling purposes, the target variable is PassengerMiles. Preliminary checks in Lab 2 confirmed strong annual seasonality and a long-run upward trend.

Methods and Procedures

A histogram shows strong left-skew at initial EDA(figure 1). Applying a log transform change the distribution to see if it made distribution more Normal curved distribution, satisfying the normality assumption used in later regression and smoothing models. However, this log transformation keeps the left-skewness after the transformation, so it doesn’t solve the skewness. Instead of log transformation, square transform changes the distribution more normally curved distribution.

Given square transformation, Using the transformed series, three polynomial time-trend regressions were fitted with ggplot2::geom_smooth(), cubic fit shows better fitting than other type of regressions(figure 2).

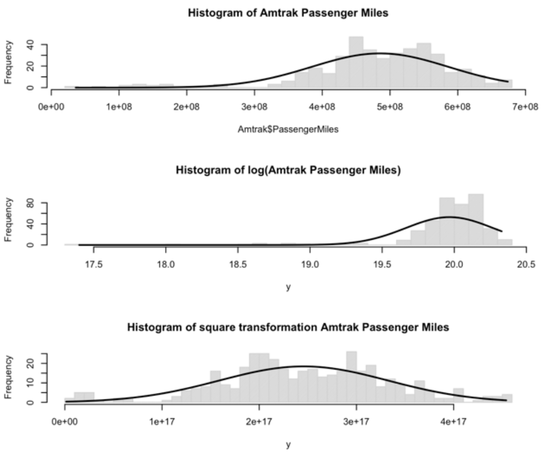

Figure. 1

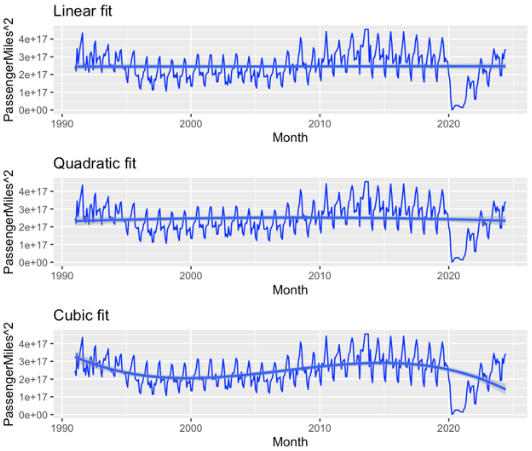

Figure. 2

After EDA four model variants were applied in a single model() call: (1) linear + trend, (2) linear + season(), (3) quadratic + season, and (4) cubic + season(figure 3). For all period, the accuracy statistics ranked the cubic-season specification best (i.e., lowest MAPE). However, this cubic fitting doesn't capture the drop well(figure 4), so re-estimated on four historical windows—1991-1997, 1991-2004, 1991-2016, and 1991-2020—to capture the best MAPE with best fitting(figure 5-12). Across four historical windows, cubic fitting has least MAPE among other fittings, as anticipated cubic fitting at 1991-2020 has largest MAPE than other periods.

All period Best fitting model Cubic MAPE:17.3

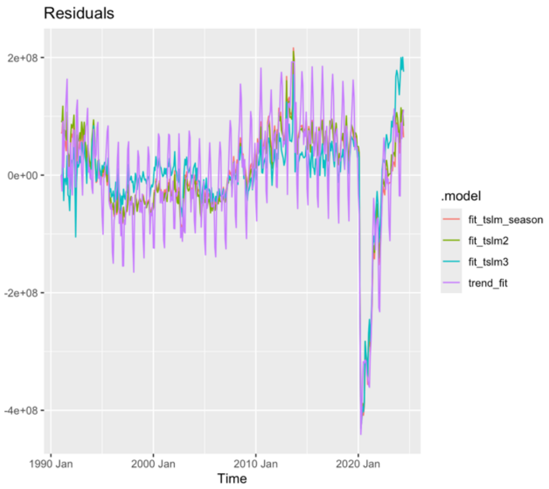

Figure. 3

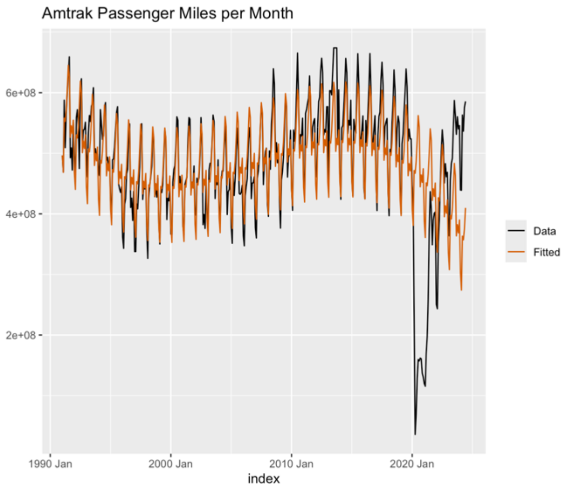

Figure. 4

1991-1997 Best fitting model Cubic: MAPE 3.56

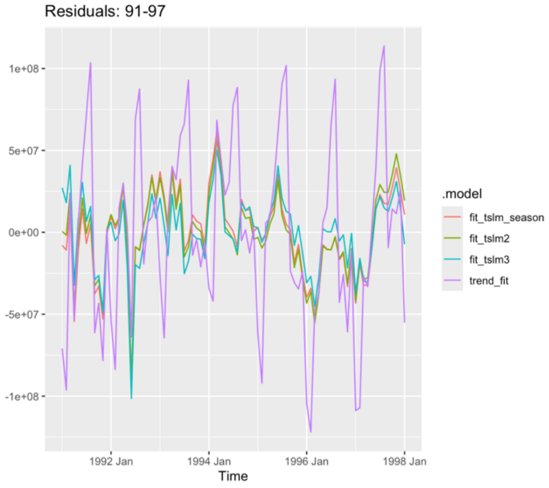

Figure. 5

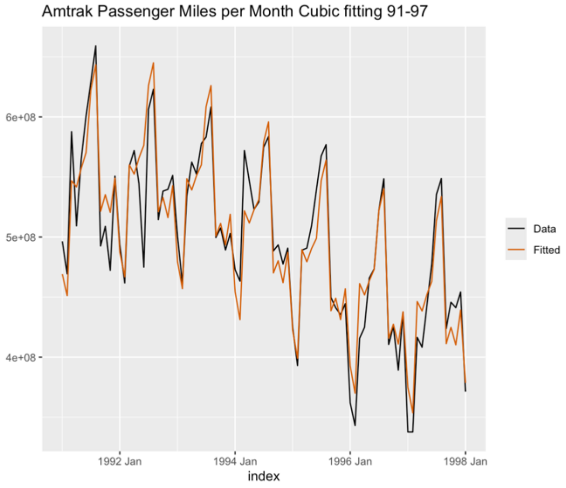

Figure. 6

1991-2004 Best fitting model Cubic: MAPE 4.15

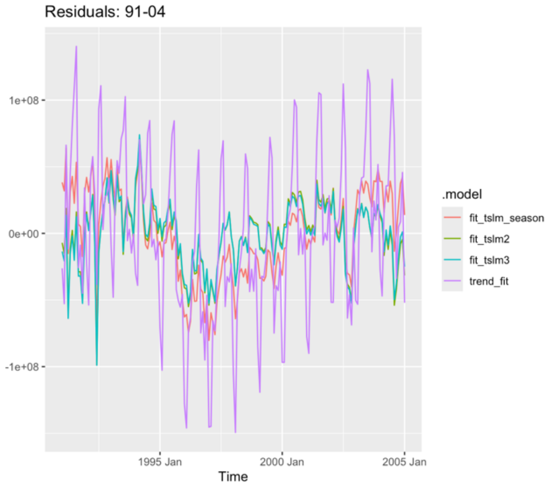

Figure. 7

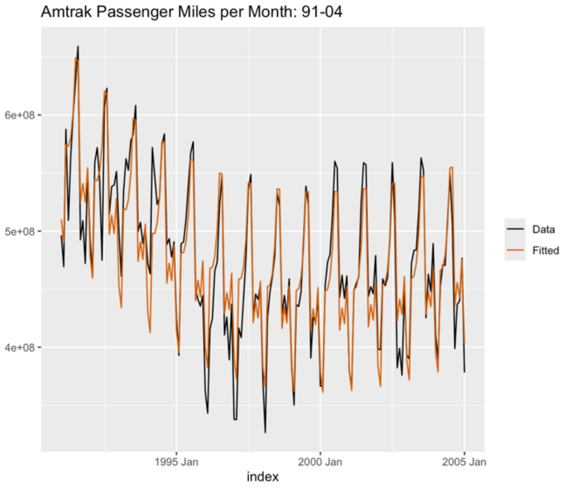

Figure. 8

1991-2016 Best fitting model Cubic: MAPE 4.43

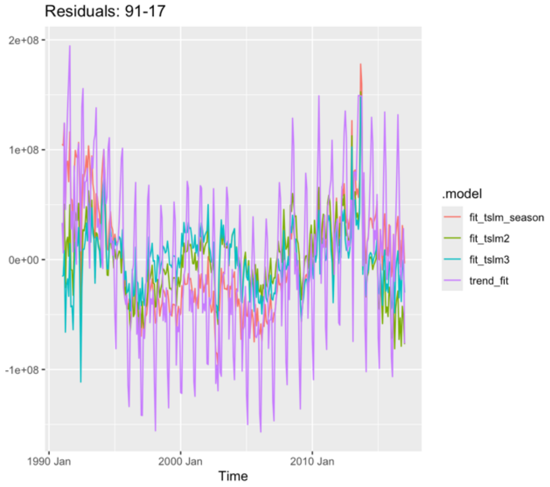

Figure. 9

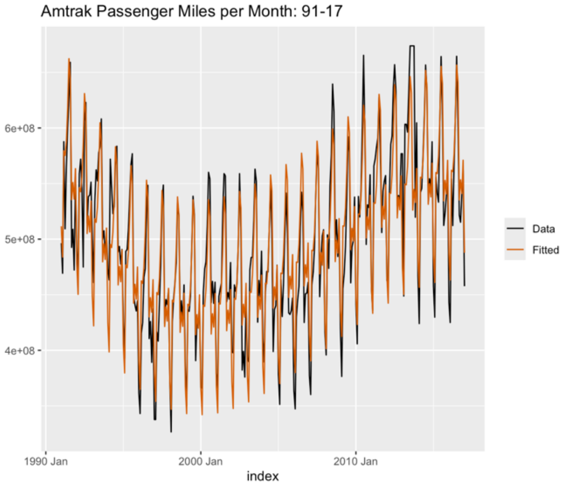

Figure. 10

1991-2020 Best fitting model Cubic: MAPE 14.0

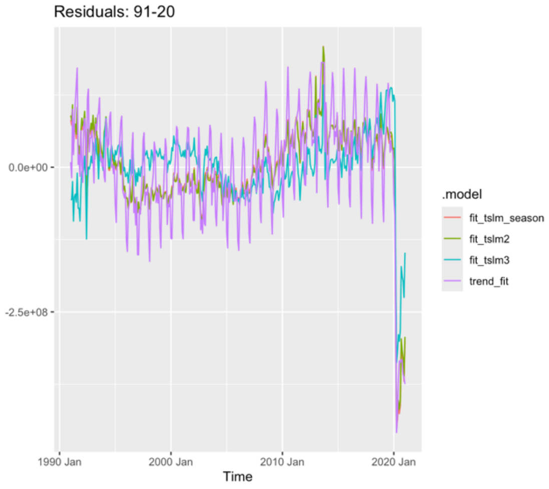

Figure. 11

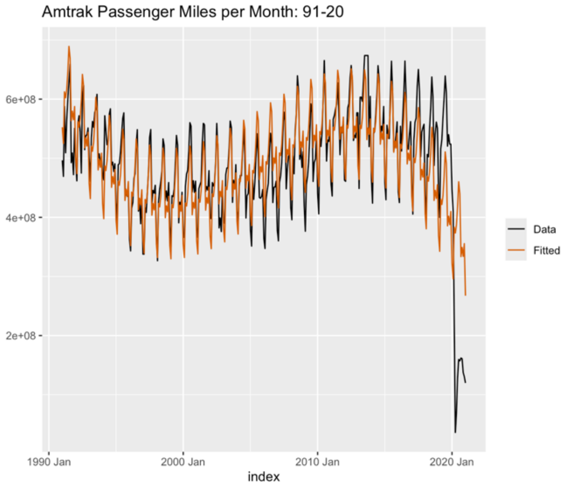

Figure. 12

Because of MAPE and the challenge of COVID sudden dropping, analysis focuses on 1991-2016 window instead of rest of it.(figure 13) Then gg_tsresiduals reveals residual of histogram, time-plot, and ACF of the residuals(figure 14). Overall, the residuals appear well-behaved for real-world data: the ACF drops off quickly, indicating little remaining autocorrelation. A spike at lag 12 hints at residual annual seasonality, but it is minor; a further seasonal difference does not seem warranted, so we proceed with the model as is. Then we decomposed the same window with decompose(), subtracted the seasonal and trend components(figure 15&16), and examined ACF/PACF after one and two seasonal differences(figure 17&18). The second difference offered no significant improvement(figure 19). As parallel exercise, using the natural-log series produced identical ACF behaviour, confirming that the original (variance-stabilised) scale was adequate(figure 20&21). We therefore retained the cubic-season specification without additional transformations for subsequent forecasting work.

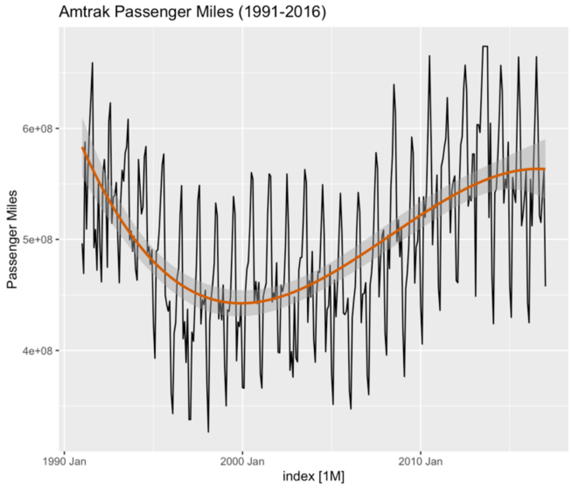

Figure. 13

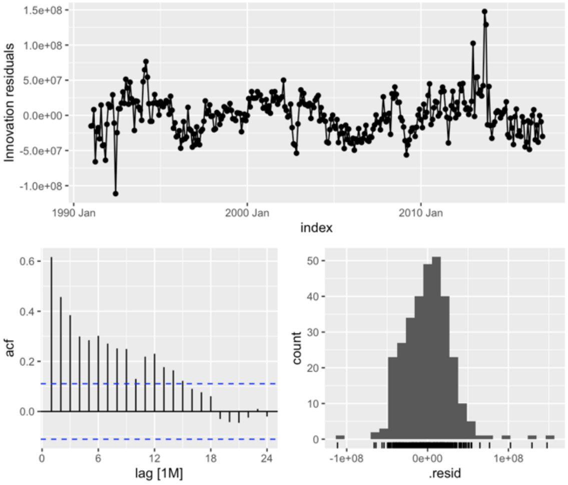

Figure. 14

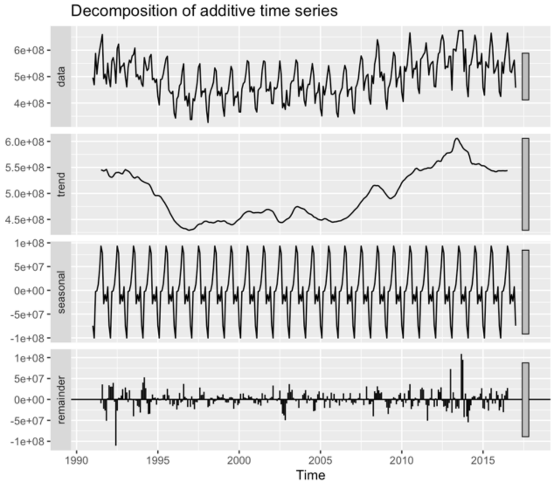

Figure. 15

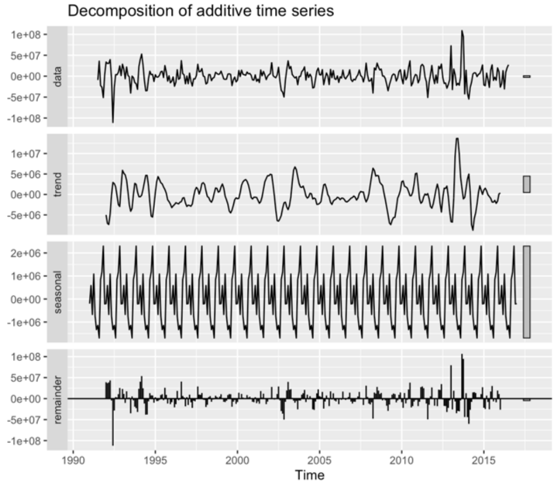

Figure. 16

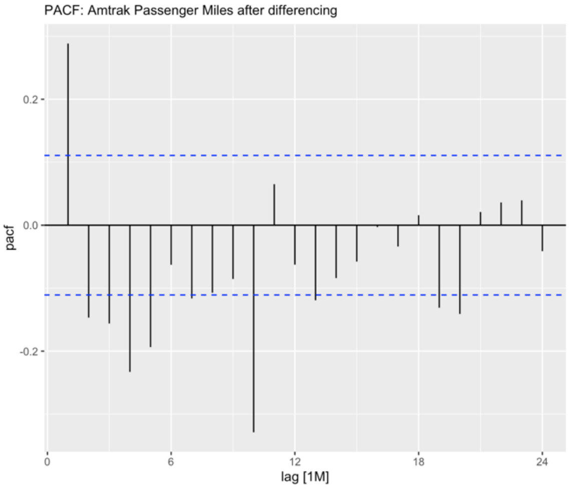

Figure. 17

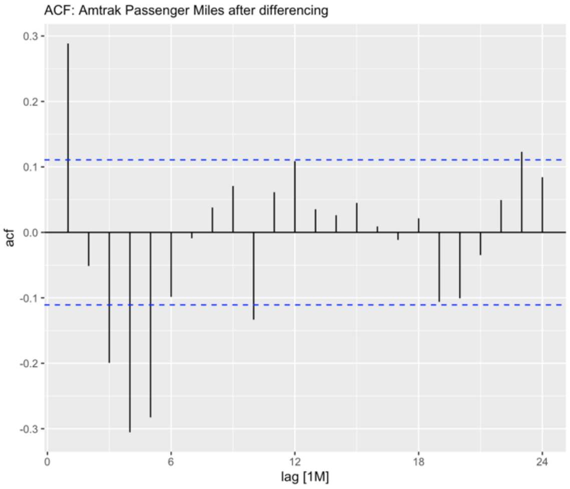

Figure. 18

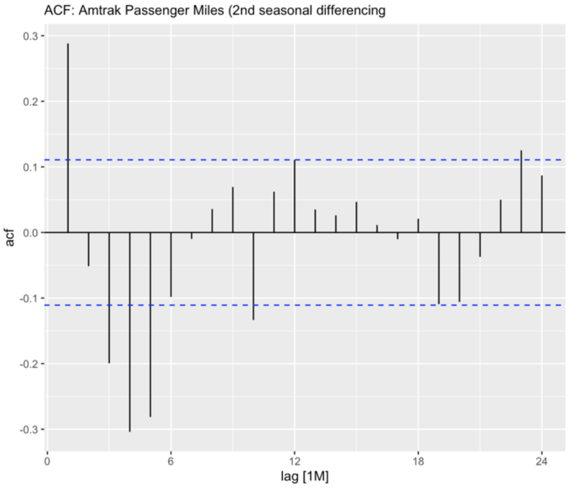

Figure 19

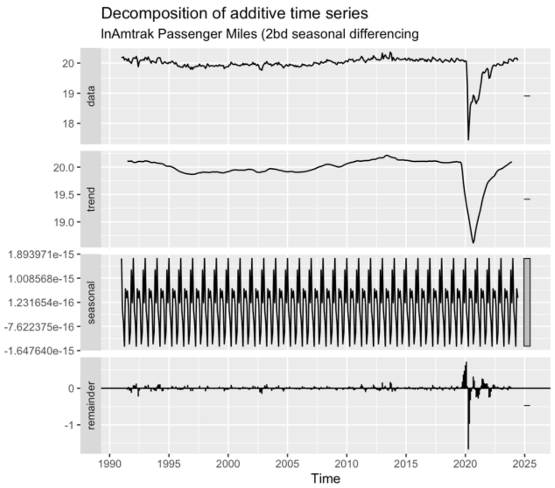

Figure. 20

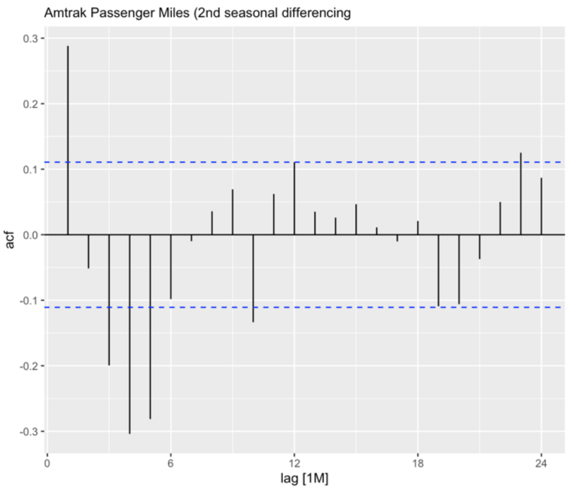

Figure. 21

Results

Because two rounds of differencing did not remove the annual signal enough, a short moving-average filter was introduced to remove the remaining seasonal noise and to provide a simple benchmark forecast. The analysis first focused on the deseasonalised, detrended series for 1991-2016, a span chosen to avoid pandemic distortions. A window length had to balance noise reduction against information loss: although the dominant cycle is yearly, a 12-month average would erase sub-annual variation, so a three-month window was selected(figure 22). Both centred and trailing versions were calculated, plotted alongside the original line(figure 23), and visually inspected to show how the centred filter tracks the peaks while the trailing filter lags turning points. To demonstrate other approach(Shumeli's book, Practicval Time Series Forecasting with R.), 1991-2004 was partitioned into a training and validation segments(figure 24&25). A 12-month trailing average was fitted to the training data and its final value was projected flat across the hold-out period; the resulting overlay, labelled “Training,” “Validation,” and “Future,” demonstrates how smoothing flattens volatility yet fails to anticipate sudden shifts because moving-average need to be de-seasonalized. This lab3 also introduce other methods like naïve forecast, ETS and Holt-Winter’s model. ETS fit automatically defaulted to an M

N A / M N M structure—no trend, only repeating seasonality—so its forecasts behave like a seasonal-naïve baseline(Figure 26&27). Holt-Winters was therefore introduced next because its additive and multiplicative forms allow both level and seasonal amplitude to evolve. At this exercise, Holt-Winter model(Additive and Multiplicative) is applied to 1991-2019 and 2020-2024 window, each models shows almost same MAPE, however shows almost identical MAPE values. Yet the training window makes a clear difference: Holt-Winters models calibrated on the full 1991-2019 history start from a higher level and recover quickly, whereas those trained only on 2020-2024 begin lower and lag the rebound(Figure 28). The contrast quantifies the COVID-era shift and confirms that multiplicative Holt-Winters—while marginally more accurate—remains highly sensitive to the chosen calibration period, highlighting the need to decide whether pre-pandemic behaviour is still relevant for forward planning.

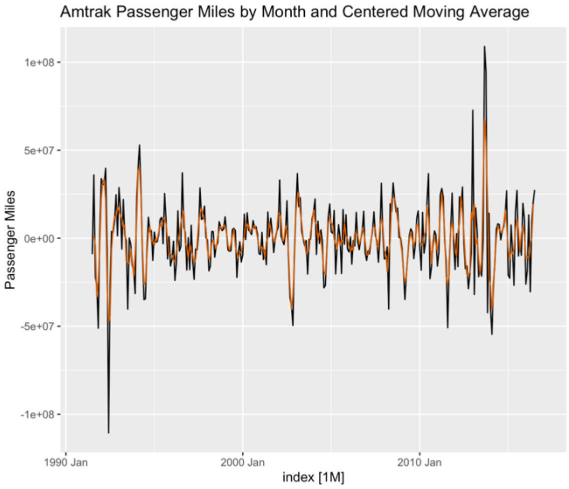

Figure. 22

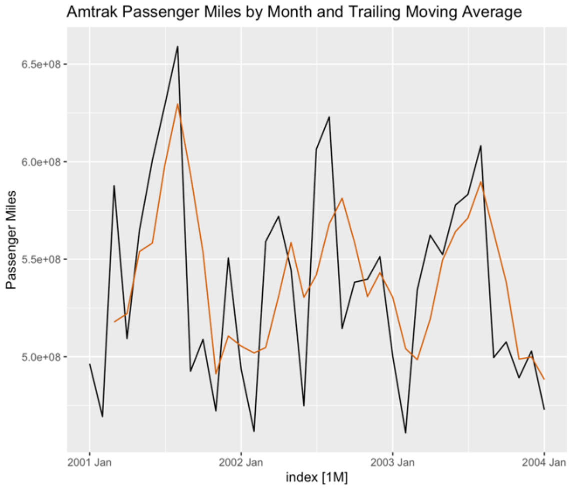

Figure. 23

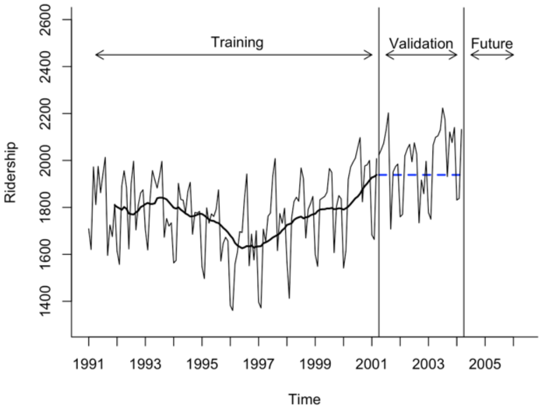

Figure. 24

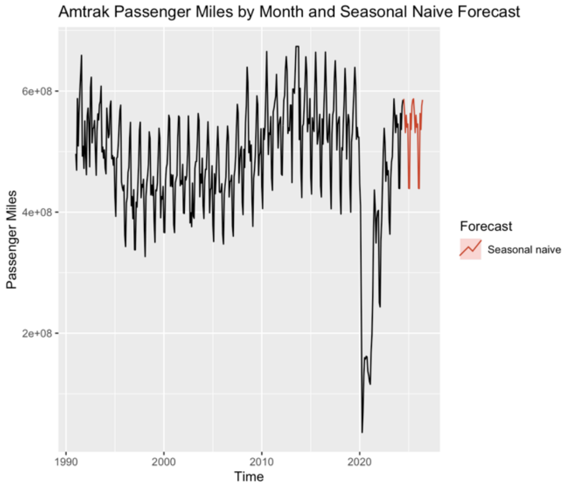

Figure. 25

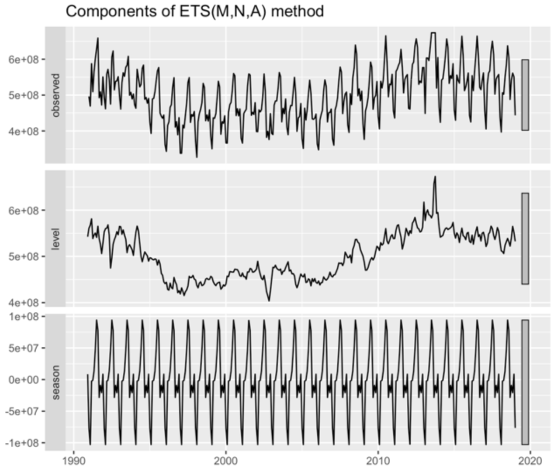

Figure. 26

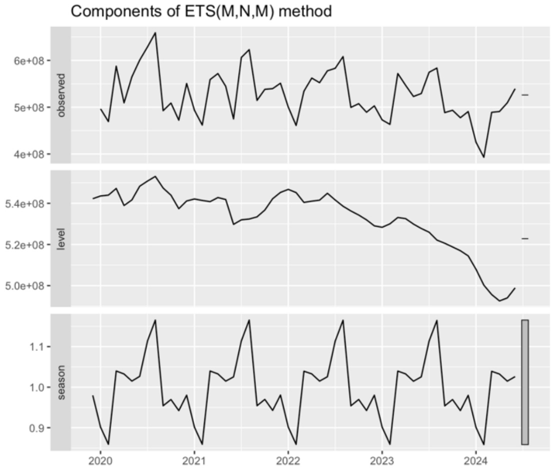

Figure. 27

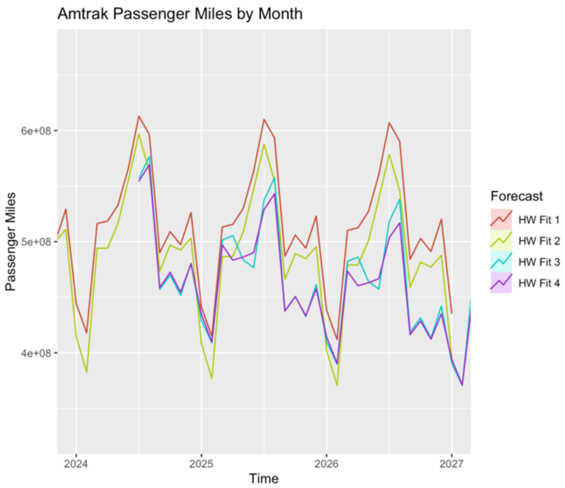

Figure 28

Conclusions

This lab3 demonstrates Three important aspects about time-series forecasting modeling. Sometime log transform fails to resolve issues; alternative power transforms along with visual checks must be used whenever distribution fits poorly. A successful forecasting process starts by removing every remainder of seasonality using twice-differencing and short moving-average filters. The length of data used for calibration influences Holt-Winters predictions because removing pre-2020 information caused the forecasts to remain lower and recover more slowly which shows that optimal historical data selection is vital for producing dependable forecasts.

References

Galit Shmueli. (2016) Practical Time Series Forecasting with R: A Hands-On Guide

Script

```
#Laboratory 3 Forecasting with regression including smoothing techniques
#such as moving averages and exponential smoothing.
```