CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Seasonal Adjustment and Forecasting by Using X-12-ARIMA Method with Application on Monthly Retail Sales in U.S.

AUTHOR: Zhaoying Zheng

SUPERVISOR: Dr. Mohamedou Haye

DATE: April 6th, 2020

Abstract

Retail sales, which defined as the purchases of finished goods and services by consumers and businesses, usually measured once a month. Similar to Gross Domestic Product (GDP) and unemployment rate, retail sales is an important indicator of a country's overall economic health and direction. It reflects how much demand exists for consumer goods and thus projects the pulse of the economy. Hence, time series analysis for monthly retail trade data has become an important part for the government agencies in all over the world.

By analyzing the retail sales data, one can not only show the tendency and long-term movements of consumer behaviors in economy, but also allow us to understand the major factors that affect monthly retail sales. In addition, if we further forecast the future monthly retail sales, it will be benefit for the government agencies as well as the local retailers to understand whether any unusual occurrences have had major effects on the retail sales, for example, the 2008 Financial Crisis and the recent COVID-19 diseases outbreak. Therefore, the government agencies and local retailers can plan inventories and to develop the strategies to prevent the problems aroused from the economic issues, such as economic recession.

However, monthly retail sales vary strongly throughout the year. Throughout a year, the size of population, the levels of unemployment, and other measures of consumption market fluctuations due to seasonal events including changes in seasons, major holidays, and school schedules. For example, there are greater consumptions leading up to major holidays such as Christmas. In addition, trading days, moving holidays and leap years also have significant impacts on the monthly retail sales. Since these seasonal and calendar events follow some regular patterns each year, their influences on cyclical trends can be eliminated by seasonally adjusting the data from month to month. These seasonal adjustments will make us easier to observe and analyze the underlying trend, long-term movements and unusual occurrences in the time series.

One of the most popular and state-of-art methods for decomposing monthly data is X-12-ARIMA, which was originally developed by the U.S. Bureau of the Census. It is now widely used by the government agencies around the world. In this project, the US monthly retail sales data from US Census Bureau will be analyzed and forecasted by using this method. The mean average percentage error (MAPE) between testing and training datasets will be calculated as a measurement for the accuracy rate during the forecasting process.

Acknowledgment

For this Honours Project, I would like to show my deepest gratitude to my supervisor, Dr. Mohamedou Haye, a patient, kind, respectable and resourceful professor, who has provided me with valuable guidance on doing this Honours Project, as well as on my predictive modeling project in my Co-op internship during summer 2019. I would also like to express my gratitude to Prof. Ahmed Almaskut, who has reviewed my Honours Project carefully and provided valuable suggestions to improve my Honours Project. I also greatly appreciate to all professors who have taught me in the classes in the past four years in Carleton University, who help me to develop the fundamental and essential academic competence to complete this Honours Project.

Table of Notations

Notations	Description
Y_t	Original time Series
y_t	Logarithm transformation of the original time series, log (Y_t) or log $(\frac{Y_t}{d_t})$
C_t	Trend (or trend cycle)
I_t	Irregular fluctuations
S_t	Seasonal component
A_t	Seasonally adjusted time series
D_t	Trading day effects
E_t	Moving holiday effects
$(SI)_t$	SI ratios, a combination of Seasonal component and Irregular fluctuations

Table of Contents

1 Introduction 9

1.1 Project Structure 9

1.2 Introduction of X-12-ARIMA 9

2 Seasonal Adjustment with X-11 11

2.1 Time Series Components 11

2.2 Decompositions for a Seasonal Time Series 12

2.3 Moving Averages 13

2.3.1 Symmetric Trend Moving Average 13

2.3.2 Symmetric Seasonal Moving Average 16

2.4 Procedures of X-11 Seasonal Adjustment 17

Stage A: Initial Estimates 17

Stage B: Revised Estimates: Seasonal Factors and Seasonal Adjustment 19

Stage C: Final Estimates: Final Henderson Trend and Final Irregular 21

3 RegARIMA Model in X-12-ARIMA 23

3.1 Introduction 23

3.2 Built-in Regressors 25

3.3 Point Forecasting 27

4 Monthly Retail Sales in US: Data Analysis and Forecasting 30

4.1 Data Preprocessing 30

4.2 Plotting Time Series 31

4.3 Model Selection by Using Package ‘x12’ in R 32

4.3.1 Testing for Trading Day Effects 32

4.3.2 Testing for Easter Effects 34

4.3.3 Testing for Different ARIMA Parameters 35

4.4 Plots Analysis of Seasonal Adjustment 37

4.5 Forecasting and Validation 40

4.6 Comparison of Sales Prediction from Pure Seasonal ARIMA Model and X-12-ARIMA Method .. 41

5 Conclusions 44

5.1 Summary and Findings 44

5.2 Future Improvements 45

6 Bibliography 46

7 Appendices 47

7.1 Appendix A: AIC Test 47

7.2 Appendix C: R Code 48

7.3 Appendix C: Partial R Output 55

7.3.1 Partial Summary of Model T1 and E1 55

7.3.2 Partial Summary of Model T2 and A1 56

7.3.3 Partial Summary of Model E2 57

7.3.4 Partial Summary of Model E3 58

7.3.5 Partial Summary of Model E4 59

7.3.6 Partial Summary of Model E5 60

7.3.7 Partial Summary of Model A2 61

7.3.8 Partial Summary of Model A3 62

7.3.9 Partial Summary of Model A4 63

7.3.10 Diagnostics for the Pure Seasonal ARIMA Model 64

List of Tables

Table 1: Some Common Weight Functions $h_j^{(2H+1)}$ for the Henderson Moving Averages………………16

Table 2: Built-in Regressors for Monthly Series in X-12-ARIMA……………………………………......26

Table 3: Preview of Training Dataset of Monthly Retail Sales in U.S……………………………………30

Table 4: Time Series of Training Dataset of Monthly Retail Sales in U.S………………………………..30

Table 5: Day Composition for June 2005, 2006 and 2007………………………………………………..33

Table 6: AIC Comparison in Sample for Models with Different Regressors to Test the Significance of Trading Day Effects based on X-12-ARIMA Seasonal Adjustment Method……………………………..33

Table 7: AIC Comparison in Sample for Models with Different Regressors to Test the Significance of Easter Effects based on X-12-ARIMA Seasonal Adjustment Method……………………………………35

Table 8: AIC Comparison in Sample for Models with Different seasonal ARIMA parameters based on X-12-ARIMA Seasonal Adjustment Method………………………………………………………..36

Table 9: Observed Values and Predicted Values for Monthly Retail Sales (in Millions) in 2013 by Using Model A3 in X-12-ARIMA Method………………………………………………………………………40

Table 10: Observed Values and Predicted Values for Monthly Retail Sales (in Millions) in 2013 by Using Pure Seasonal ARIMA Model, ARIMA$(1,1,1) \times (1,1,1)_{12}$……………………………………………...42

List of Figures

Figure 1: Flow Diagram for Seasonal Adjustment with X-12 ARIMA…………………………………...10

Figure 2: Flow Diagram of the X-11 Seasonal Adjustment in X-12-ARIMA Method…………………...22

Figure 3: Plot of Original Time Series of Monthly Retail Sales (in Millions) from January 2004 to December 2012...31

Figure 4: Partial Summary of Model T2 in R…………………………………………………………..34

Figure 5: Autocorrelations of the Residuals of Model A3……………………………………………...36

Figure 6: Plots of Final Trend Estimate, Final Seasonal Component Estimate and Final Irregular Fluctuations Estimate of the Original Time Series by Using Model A3………………………………….37

Figure 7: Plots of Final Seasonally Adjusted Data by Using Model A3…………………………………38

Figure 8: Plot of Seasonal Factors and SI Ratios by Month by Using plotSeasFac() function…………..39

Figure 9: Plot of Original Series, Seasonally Adjusted Series, Trend, and Outliers by Using Model A3..39

Figure 10: Time Series with Point Forecasts and 95% Prediction Interval by Using Model A3…………40

Figure 11: ACF and PACF of the first differenced and then seasonally differenced time series, $(1 - B)(1 - B^{12})Y_t$……………………………………………………………………………….............42

1 Introduction

1.1 Project Structure

This Honours Project focuses on developing concepts of the X-12-ARIMA method, as well as applying the X-12-ARIMA method on project data analysis and forecasting. Since the original output from X-12-ARIMA method from U.S. Bureau of the Census is somehow static and it is not easy for users to extract the required information for further processing, we will use the R package x12 for project data analysis and forecasting. It allows summarizing, modifying and storing the outputs from X-12-ARIMA within a well-defined class-oriented implementation (Kowarik, Meraner, Templ and Schopfhauser, 2014).

In the next section 1.2, a briefly introduction of the X-12-ARIMA method will be provided. Chapters 2 and 3 will focus on the concepts, methodologies and diagnostic tests for the X-12-ARIMA method. In chapter 4, we will focus on the project data analysis by using x12 package in R, which we apply X-12-ARIMA method to decompose the time series and analyze the underlying trend of monthly retail sales data in U.S. from 2004 to 2012. It will then follow by models building and model selection in R. After that, the future monthly sales in 2013 will be forecasted, which will be compared to the actual monthly sales in 2013. Finally, the summaries, findings and future improvements for the project will be stated in chapter 5. The R code and major outputs for the project data analysis, models selection and forecasting will be shown in Appendices.

1.2 Introduction of X-12-ARIMA

The basic assumption of seasonal adjustment with X-12-ARIMA seasonal adjustment method of the US Census Bureau is the possibility to decompose a quarterly or monthly series (Y_t) into several components, namely the seasonal component (S_t), the trend (or trend cycle) component (C_t), and a residual (irregular) component (I_t), by using a combination of moving average filters in a seasonal adjustment method called X-11.

For preadjusting the series, an algorithm based on RegARIMA (linear regression model with ARIMA time series error) is used by X-12-ARIMA. RegARIMA models are used for forecasting and backcasting prior to seasonal adjustments, and to deal with outliers, moving holiday effect (E_t) such as Easter Day, trading day effect (D_t) (e.g., there are four or five weekends for different months in different years), leap year effect (the years where an extra day is added to the end of February), and user-defined regressors.

Summarized by Findley, Monsell, Bell, Otto and Chen (1998), the figure below provides an overview of how X-12-ARIMA method works for seasonal adjustment:

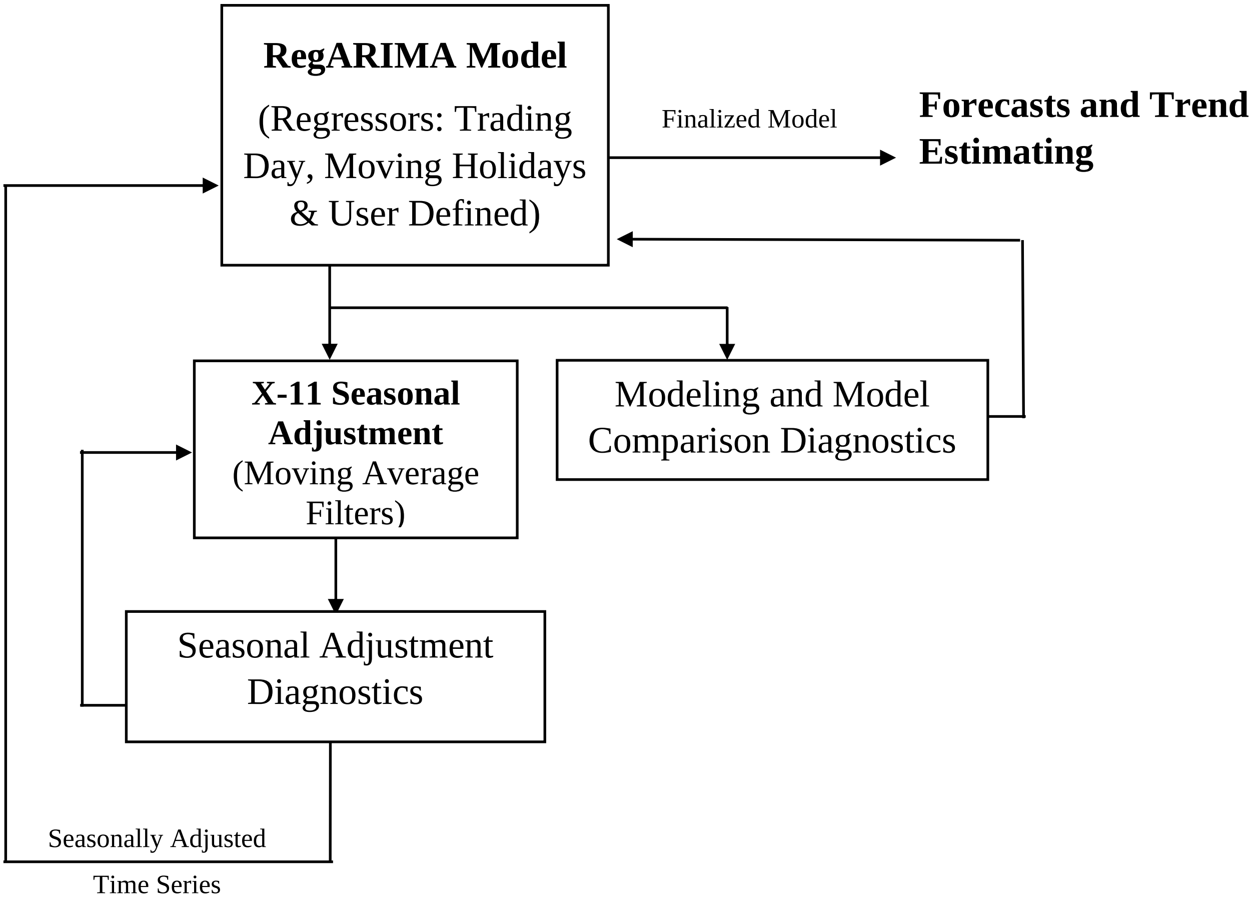

RegARIMA Model
(Regressors: Trading Day, Moving Holidays & User Defined)
Finalized Model
Forecasts and Trend Estimating
X-11 Seasonal Adjustment
(Moving Average Filters)
Modeling and Model Comparison Diagnostics
Seasonal Adjustment Diagnostics
Seasonally Adjusted
Time Series

Figure 1: Flow Diagram for Seasonal Adjustment and Forecasting with X-12-ARIMA

From the flow diagram above, we clearly see that a RegARIMA model is initially built for the time series to preadjust various effects such as outliers, moving holiday, trading day and leap year effects. Then it carries out the actual seasonal adjustment, which decomposes the pre-adjusted series, i.e. the output from the RegARIMA step, into three elements: trend, seasonal component, and irregular fluctuations. Finally, the final step of the program is to test the effectiveness of both modeling and seasonal adjustment by using a set of diagnostics tests.

As the two major components in X-12-ARIMA method are X-11 seasonal adjustment and RegARIMA models, the next chapters 2 and 3 will provide detailed concepts, methodologies and diagnostic tests on these two components, respectively.

2 Seasonal Adjustment with X-11

A time series is formed by the data that are collected over time. There are lots of data published monthly, or quarterly by the government agencies are time series, including Consumer Price Index (CPI), Gross Domestic Product (GDP), unemployment rate, and the one that we focus on in this project, the retail sales. Usually, those who analyzing the time series typically would like to establish the underlying trends, cyclical deviations from trend, long-term movements and unusual occurrences from the time series. However, the analysis on the original time series would not be straightforward, because of the short-term effects associated with the time of the year, which generally called seasonal component. It will be discussed under the section 2.1.

Seasonal adjustment is hence be introduced. The purpose of seasonal adjustment is to remove the seasonal component of a time series that exhibits a seasonal pattern. It is normal to report seasonally adjusted data for CPI, GDP, unemployment rate and retail sales to reveal the underlying trends and behaviours in economy. To understand the X-11 seasonal adjustment method, first, we need to understand the components in a time series, which will be introduced in section 2.1. Section 2.2 states the two different types of decomposition model in X-12-ARIMA method, along with the methodology of identifying which decomposition method should we choose for the time series. In section 2.3, we will briefly explain the three types of moving averages that being used during the X-11 seasonal adjustment process. Finally, we will provide the complete process of X-11 seasonal adjustment for the two different decomposition model in section 2.4.

2.1 Time Series Components

According to the findings from National Statistics (2007), generally, time series data can be split into three major components: trend (or trend cycle) (C_t), seasonality (S_t) and irregular fluctuations (I_t), each representing the impact of certain types of real world events on data, where:

- Trend (or trend cycle) (C_t): represents the direction and underlying behaviours of the time series. It is a reflection of the medium-long term movement in the series and it is typically due to the influences such as population growth, general economic development and business cycle;

- Irregular fluctuations (I_t): represents the residuals of the time series after the identification of the trend and seasonal components, which are due to the unpredictable factors, such as sampling error, non-sampling error and anti-season weather. It also contains the outliers, which are the extreme values in the time series data. The outliers usually have identifiable causes, such as war, natural disasters and strikes. Although outliers are part of the irregular fluctuations (I_t), they will be identified and replaced during the seasonal adjustment process, see section 2.4 below.

- Seasonal component (S_t): comprises seasonal effects and calendar effects, where

    i. Seasonal effects: the cyclic patterns which are caused by the changes associated with the natural conditions, (e.g. seasons and weather patterns), the administrative measures (e.g. the start and end dates of the school year), and the social and cultural behaviour (e.g. Christmas);

    ii. Calendar effects: the impacts that relate to factors which do not necessarily occur in the same month each year, including

        a. Trading day effects (D_t): the months having different numbers of each day of the week from year to year. For example, there are four or five weekends for different months from year to year, and leap year effect for February;

        b. Moving holidays (E_t): the holidays that occur in different months from year to year (e.g. Easter Day).

2.2 Decompositions for a Seasonal Time Series

In X-12-ARIMA, the two most commonly used decomposition models of a time series Y_t at time t are additive model and multiplicative model, which are defined as below:

- Additive model: the observed time series (Y_t) is considered to be the sum of three independent components C_t, S_t and I_t, that is

$$Y_t = C_t + S_t + I_t \tag{2.2.1}$$

When the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls, the additive model is appropriate to be used.

- Multiplicative model: the observed time series (Y_t) is expressed as the product of three independent components C_t, S_t and I_t, that is

$$Y_t = C_t \times S_t \times I_t \tag{2.2.2}$$

When the amplitude of both the seasonal and irregular variations increase as the level of the trend rises, the multiplicative model is appropriate to be used.

In addition, in X-12-ARIMA (compared to the previous version X-11-ARIMA), a pseudo-additive model is added:

$$Y_t = C_t \times (S_t + I_t - 1) = C_t(S_t - 1) + C_t I_t \tag{2.2.3}$$

The pseudo-additive model is introduced because the multiplicative model cannot be used when the original time series Y_t has very small or zero values in the same month or months each year. In these cases, a pseudo-additive model combining of both additive and multiplicative models is used. However, since our retail sales data do not contain any very small or zero values, we will not focus on the analysis of pseudo-additive model in this project. For more information of using the pseudo-additive model for seasonal adjustment, see Findley, Monsell, Bell, Otto, and Chen (1998).

2.3 Moving Averages

Moving averages, which are the trend filters of the X-11 seasonal adjustment method, are used to average a shifting time span of data in order to produce a smoothed estimate of a time series. By applying symmetric trend moving averages to the original data, the trend or trend cycle will be smoother than the original data and captures the main movement of time series without all of the minor fluctuations.

In X-12 ARIMA, there are three moving averages used in different stages: symmetric centred moving average, Henderson moving average and symmetric seasonal moving average, where the first two belong to the trend moving average. These two types of moving averages will be discussed separately below.

2.3.1 Symmetric Trend Moving Average

i. Centered Moving Average

The idea of centered moving average is replacing the original data by an average of its current value and its neighbors in the past and future for the same length. Hence, the centered moving averages are symmetric. To derive the definition of centered moving average, we first introduce the simple moving average. According to Hyndman (2009), a simple moving average of order m can be written as

$$\hat{C}_t = \frac{1}{m}\sum_{j=-k}^{k} Y_{t+j} \tag{2.3.1}$$

where $m = 2k + 1$. That is, the estimate of the trend at time t is obtained by averaging values of the time series within k periods of t.

However, observing that the simple moving averages are symmetric only when the order m is an odd value, i.e. $m = 2k + 1$. If m is even, it would no longer be symmetric. By making an even-order moving average symmetric, Hyndman (2009) suggest that we can apply a moving average to a moving average. For example, if we initially take a moving average of order 12, and then apply another moving average of order 2 to the results, we have

$$\begin{aligned}\hat{C}_t &= \frac{1}{2} \times \left(\frac{Y_{t-6} + Y_{t-5} + \cdots + Y_t + Y_{t+1} + \cdots + Y_{t+5}}{12} + \frac{Y_{t-5} + Y_{t-4} + \cdots + Y_t + Y_{t+1} + \cdots + Y_{t+6}}{12}\right) \\ &= \frac{1}{24}Y_{t-6} + \frac{1}{12}Y_{t-5} + \cdots + \frac{1}{12}Y_t + \cdots + \frac{1}{12}Y_{t+5} + \frac{1}{24}Y_{t+6} \qquad (2.3.2)\end{aligned}$$

Now the weighted average of observations is symmetric. This filter is called centered (2×12) moving average. In fact, this filter is usually applied to monthly data, where each month of the year is given equal weight as the first and last term apply to the same month in consecutive years. As a result, the seasonal variation will be averaged out and the resulting values of $\hat{C}_t$ will have less seasonal variation remaining. The centered (2×12) moving average will be used in step 2 of Stage A in the procedures of X-11 seasonal adjustment (see section 2.4 below) for the initial trend estimate.

ii. Henderson Moving Average

We first introduce symmetric weighted moving average. Stated by Shumway and Stoffer (2006), in preference to simple moving averages, a symmetric weighted moving average is defined as

$$\hat{C}_t = \sum_{j=-k}^{k} a_j Y_{t+j} \tag{2.3.3}$$

where the weights a_j are symmetric and sum to one, i.e. $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$. Compared to the simple moving averages, the advantage of weighted moving averages is that the resulting trend estimate is much smoother.

Henderson moving averages are filters that derived by Robert Hendersn in 1916. Different from the centered moving averages and weighted moving average, the Henderson moving average filters are applied on A_t (seasonal adjusted time series), in Stage B and Stage C of the procedures of X-11 seasonal adjustment (see section 2.4 below). Using the concepts of symmetric weighted moving average, Henderson moving average is defined as:

$$\hat{C}_t = \sum_{j=-H}^{H} h_j^{(2H+1)} A_{t+j} \tag{2.3.4}$$

where $h_j^{(2H+1)}$are the weights for the Henderson moving averages, which are called the Henderson coefficients. Observed by Gray and Thomson (1996), the Henderson coefficients are given by

$$h_j^{(2H+1)} = q_j(H)(a + bj^2), \quad -H \leq j \leq H, \tag{2.3.5}$$

where $q_j(H)$ is defined as

$$q_j(H) = \{(H+1)^2 - j^2\}\{(H+2)^2 - j^2\}\{(H+3)^2 - j^2\},$$

and a and b are determined by

$$a \sum_{j=-H}^{H} q_j(H) + b \sum_{j=-H}^{H} q_j(H) j^2 = 1$$

and

$$a \sum_{j=-H}^{H} q_j(H) j^2 + b \sum_{j=-H}^{H} q_j(H) j^4 = 0$$

Equation (2.3.4) also shows that the Henderson moving averages are symmetric, i.e. $h_j^{(2H+1)} = -h_j^{(2H+1)}$. Note that they can be < 0.

To determine the optimum value of H, first let Δ denote the differencing operator, so that $\Delta A_{t+j} = A_{t+j} - A_{t+j-1}$. Then the smoothness measure is

$$E\left(\Delta^3 \sum_{j=-H}^{H} h_j A_{t+j}\right)^2 \tag{2.3.6}$$

Since the Henderson filter is the minimizer of the smoothness measure defined in equation (2.3.6) (Gray and Thomson, 1996), the optimum value of H will be the one that minimize the smoothness measure.

In X-12-ARIMA, the automatic selection procedure is the same as the other moving average filters, but the user can alternatively specify any odd-number length $2H + 1$ (Findley, Monsell, Bell, Otto and Chen, 1998). In fact, according to Hyndman (2009), we obtain the values of some common weight functions $h_j^{(2H+1)}$:

Filter Length	$h_0^{(2H+1)}$	$h_1^{(2H+1)}$	$h_2^{(2H+1)}$	$h_3^{(2H+1)}$	$h_4^{(2H+1)}$	$h_5^{(2H+1)}$	$h_6^{(2H+1)}$
5 Term (H=1)	0.558	0.295	-0.073				
7 Term (H=3)	0.412	0.294	0.059	-0.059			
9 Term (H=4)	0.330	0.267	0.119	-0.010	-0.041		
13 Term (H=6)	0.240	0.214	0.147	-0.066	0.000	-0.028	-0.019

Note: $h_j^{(2H+1)} = -h_j^{(2H+1)}$

Table 1: Some Common Weight Functions $h_j^{(2H+1)}$ for the Henderson Moving Averages

2.3.2 Symmetric Seasonal Moving Average

The goal of applying symmetric seasonal moving averages is to obtain the estimates of the seasonal component from SI ratios (detrended time series). The symmetric seasonal moving averages used in X-11 seasonal adjustment are the simple 3-term moving averages, of simple averages of odd length, $m = 2k + 1$ of SI ratios from the same calendar month each year as month t,

$$S_t^{3\times(2k+1)} = \frac{1}{3}\,(S_{t-12}^{2k+1} + S_t^{2k+1} + S_{t+12}^{2k+1}) \tag{2.3.7}$$

with

$$S_t^{2k+1} = \frac{1}{2k+1}\sum_{j=-k}^{k} SI_{t+12j} \tag{2.3.8}$$

$S_t^{3\times(2k+1)}$ is referred to as the $3 \times (2k+1)$ seasonal moving average, and it is symmetric. See section 2.3.1.i above as to how this is defined. In X-11 seasonal adjustment, the 3×3 seasonal moving average is used at step 3 in Stage A and the 3×5 seasonal moving average is used at step 3 in Stage B.

2.4 Procedures of X-11 Seasonal Adjustment

Summarized the findings by Findley, Monsell, Bell, Otto and Chen (1998), X-11 seasonal adjustment can be split into three stages in order: Stage A: Initial Estimates, Stage B: Revised Estimates, and Stage C: Final Estimates, which will be explained in the following sections, respectively.

Stage A: Initial Estimates

Frist of all, the original time series Y_t is decomposed into three components: Trend (C_t), seasonal component (S_t) and Irregular fluctuations I_t, using additive model, multiplicative model or pseudo-additive model (not discussed in this project). The equations for additive model and multiplicative model are given by the followings, respectively:

- For additive model,

$$Y_t = C_t + S_t + I_t$$

- For multiplicative model,

$$Y_t = C_t \times S_t \times I_t$$

Secondly, we obtain the initial estimate of the trend $C_t^{(1)}$ by applying a centered (2×12) moving average on the original time series Y_t:

$$C_t^{(1)} = \frac{1}{2} \times \left(\frac{Y_{t-6} + Y_{t-5} + \cdots + Y_t + Y_{t+1} + \cdots + Y_{t+5}}{12} + \frac{Y_{t-5} + Y_{t-4} + \cdots + Y_t + Y_{t+1} + \cdots + Y_{t+6}}{12}\right)$$

$$= \frac{1}{24} Y_{t-6} + \frac{1}{12} Y_{t-5} + \cdots + \frac{1}{12} Y_t + \cdots + \frac{1}{12} Y_{t+5} + \frac{1}{24} Y_{t+6}$$

Thirdly, we estimate the seasonal-irregular ratio $(SI)_t^{(1)}$ (detrended Time Series) by removing the initial estimate of the trend $C_t^{(1)}$ (given by Eq. 2.4.4) from the original time series Y_t:

- For additive model,

$$(SI)_t^{(1)} = Y_t - C_t^{(1)}$$

- For multiplicative model,

$$(SI)_t^{(1)} = Y_t / C_t^{(1)}$$

Then, we preliminarily estimate the seasonal component $\hat{S}_t^{(1)}$ (seasonal moving average), where the estimate of initial preliminary seasonal factors can be derived by applying a weighted 5-term (3×3) moving average to the $(SI)_t^{(1)}$ series for each month separately:

$$\hat{S}_t^{(1)} = \tfrac{1}{9} (SI)_{t-24}^{(1)} + \tfrac{2}{9} (SI)_{t-12}^{(1)} + \tfrac{3}{9} (SI)_t^{(1)} + \tfrac{2}{9} (SI)_{t+12}^{(1)} + \tfrac{1}{9} (SI)_{t+24}^{(1)}$$

After that, the initial estimate of seasonal factors is derived by normalizing the initial preliminary seasonal factors.

- For additive model,

$$S_t^{(1)} = \hat{S}_t^{(1)} - \left(\frac{1}{24} \hat{S}_{t-6}^{(1)} + \frac{1}{12} \hat{S}_{t-5}^{(1)} + \cdots + \frac{1}{12} \hat{S}_t^{(1)} + \cdots + \frac{1}{12} \hat{S}_{t+5}^{(1)} + \frac{1}{24} \hat{S}_{t+6}^{(1)}\right)$$

- For multiplicative model,

$$S_t^{(1)} = \hat{S}_t^{(1)} / \left(\frac{1}{24} \hat{S}_{t-6}^{(1)} + \frac{1}{12} \hat{S}_{t-5}^{(1)} + \cdots + \frac{1}{12} \hat{S}_t^{(1)} + \cdots + \frac{1}{12} \hat{S}_{t+5}^{(1)} + \frac{1}{24} \hat{S}_{t+6}^{(1)}\right)$$

Note that this step is done to ensure that the annual average of the seasonal factors is close to one.

At the same time, the outliers are identified, removed and replaced from $(SI)_t^{(1)}$ by eliminating $S_t^{(1)}$ from to obtain irregular component series $I_t^{(1)}$, calculating moving standard deviation via $I_t^{(1)}$, and calculating weights by using the function of moving standard deviation. Finally, according to the extreme value and the other values that closest to it during the same period, modifying the extreme value and letting it replace the origin one in origin series Y_t.

The final step in stage A is to obtain the preliminary seasonally adjusted series $A_t^{(1)}$ by eliminating the initial estimate of seasonal factors $S_t^{(1)}$ by:

- For additive model,

$$A_t^{(1)} = Y_t - S_t^{(1)}$$

- For multiplicative model,

$$A_t^{(1)} = Y_t / S_t^{(1)}$$

Stage B: Revised Estimates: Seasonal Factors and Seasonal Adjustment

The first step in Stage B is to obtain the revised estimate of the trend-cycle component by using a (2H+1)-term Henderson moving average:

$$C_t^{(2)} = \sum_{j=-H}^{H} h_j^{(2H+1)} A_{t+j}^{(1)}$$

Where $A_t^{(1)}$ comes from the initial estimates in Stage A, the (2H+1)-term Henderson coefficients are $h_j^{(2H+1)}$ defined in section 2.3.1.ii, and H is determined by data of Henderson moving average.

Similar to Stage A, the second step in Stage B is to obtain the revised estimate of the seasonal-irregular ratio $(SI)_t^{(2)}$by removing the revised estimate of trend component $C_t^{(2)}$ from the original time series Y_t:

- For additive model,

$$(SI)_t^{(2)} = \mathrm{Y_t} - C_t^{(2)}$$

- For multiplicative model,

$$(SI)_t^{(2)} = Y_t / C_t^{(2)}$$

Then, we can get an estimate of revised preliminary seasonal factors $\hat{S}_t^{(2)}$ by applying a 3×5 centered moving average to the revised $(SI)_t^{(2)}$ ratios:

$$\hat{S}_t^{(2)} = \frac{1}{15}(SI)_{t-36}^{(2)} + \frac{2}{15}(SI)_{t-24}^{(2)} + \frac{3}{15}(SI)_{t-12}^{(2)} + \frac{3}{15}(SI)_t^{(2)} + \frac{3}{15}(SI)_{t+12}^{(2)} + \frac{2}{15}(SI)_{t+24}^{(2)} + \frac{1}{15}(SI)_{t+36}^{(2)}$$

Consequently, the final revised estimate of seasonal factors $S_t^{(2)}$ can be derived by normalizing the revised preliminary seasonal factors $\hat{S}_t^{(2)}$:

- For additive model,

$$S_t^{(2)} = \hat{S}_t^{(2)} - \left(\frac{1}{24}\hat{S}_{t-6}^{(2)} + \frac{1}{12}\hat{S}_{t-5}^{(2)} + \cdots + \frac{1}{12}\hat{S}_t^{(2)} + \cdots + \frac{1}{12}\hat{S}_{t+5}^{(2)} + \frac{1}{24}\hat{S}_{t+6}^{(2)}\right)$$

- For multiplicative model,

$$S_t^{(2)} = \hat{S}_t^{(2)} / \left(\frac{1}{24}\hat{S}_{t-6}^{(2)} + \frac{1}{12}\hat{S}_{t-5}^{(2)} + \cdots + \frac{1}{12}\hat{S}_t^{(2)} + \cdots + \frac{1}{12}\hat{S}_{t+5}^{(2)} + \frac{1}{24}\hat{S}_{t+6}^{(2)}\right)$$

Finally, we obtain the revised estimate of seasonally adjusted series $A_t^{(2)}$ by eliminating the the final revised estimate of seasonal factors $S_t^{(2)}$ from the original time series Y_t:

- For additive model,

$$A_t^{(2)} = Y_t - S_t^{(2)}$$

- For multiplicative model,

$$A_t^{(2)} = Y_t / S_t^{(2)}$$

Stage C: Final Estimates: Final Henderson Trend and Final Irregular

Similar to the first step in Stage B, we first apply a (2H+1)-term Henderson moving average on the revised estimate of seasonally adjusted series $A_t^{(2)}$ to obtain the final estimate of the trend-cycle component $C_t^{(3)}$, which is defined by:

$$C_t^{(3)} = \sum_{j=-H}^{H} h_j^{(2H+1)} A_{t+j}^{(2)}$$

where $A_t^{(2)}$ comes from revised estimates in Stage B. H is determined by data of Henderson moving average.

Finally, we get the final estimate of the irregular component $I_t^{(3)}$ by removing the final estimate of the trend component $C_t^{(3)}$ from the revised estimate of seasonally adjusted series $A_t^{(2)}$:

- For additive model,

$$I_t^{(3)} = A_t^{(2)} - C_t^{(3)}$$

- For multiplicative model,

$$I_t^{(3)} = A_t^{(2)}/C_t^{(3)}$$

Thus, the seasonal adjustment of time series is completed. In conclusion, the final estimated decompositions are:

- For additive model,

$$Y_t = C_t^{(3)} + S_t^{(2)} + I_t^{(3)}$$

- For multiplicative model,

$$Y_t = C_t^{(3)} \times S_t^{(2)} \times I_t^{(3)}$$

We can summarize the procedures of X-11 seasonal adjustment by using the following flow chart:

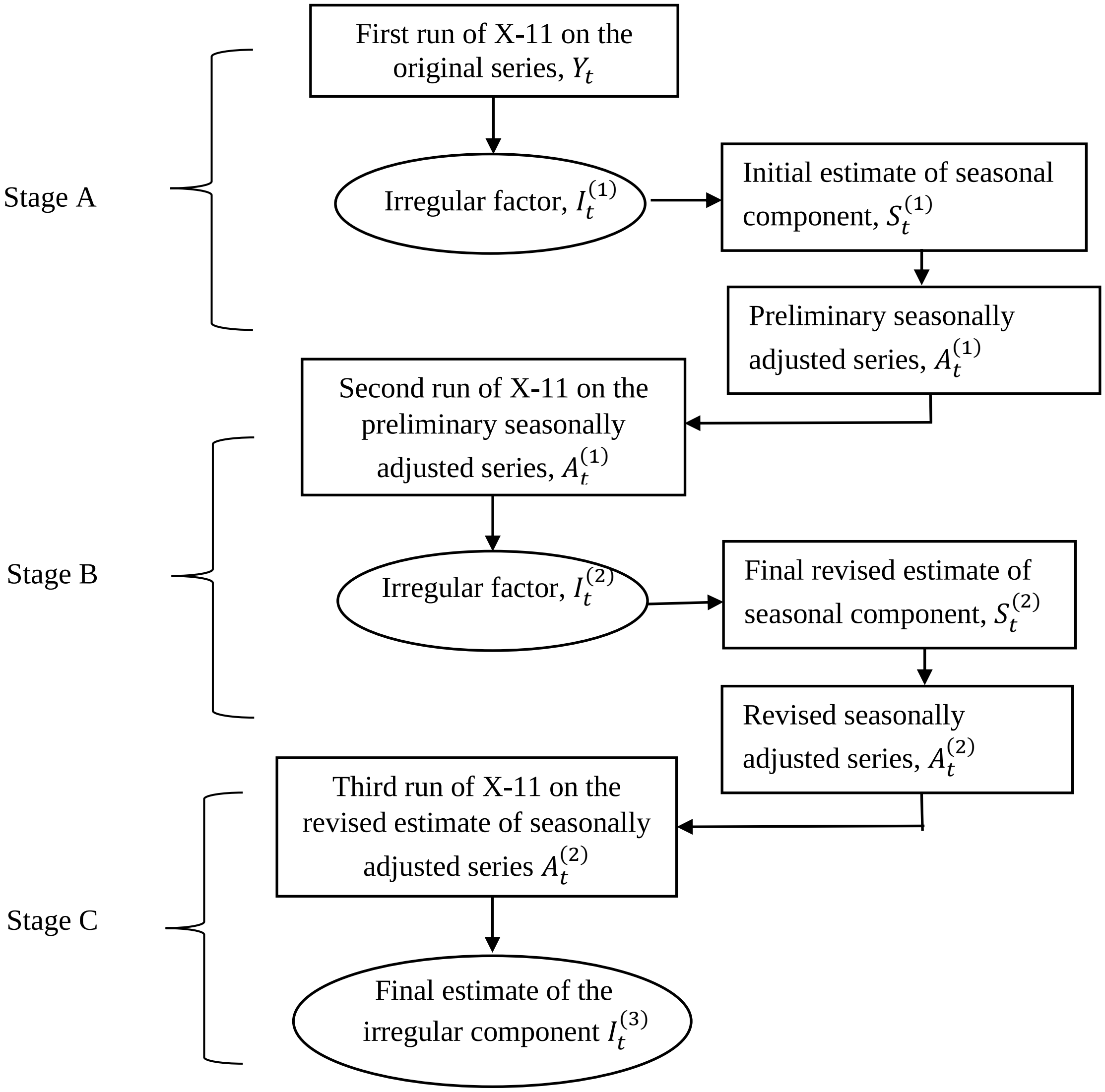

Stage A
First run of X-11 on the original series, Y_t
Irregular factor, $I_t^{(1)}$
Initial estimate of seasonal component, $S_t^{(1)}$
Preliminary seasonally adjusted series, $A_t^{(1)}$
Stage B
Second run of X-11 on the preliminary seasonally adjusted series, $A_t^{(1)}$
Irregular factor, $I_t^{(2)}$
Final revised estimate of seasonal component, $S_t^{(2)}$
Revised seasonally adjusted series, $A_t^{(2)}$
Stage C
Third run of X-11 on the revised estimate of seasonally adjusted series $A_t^{(2)}$
Final estimate of the irregular component $I_t^{(3)}$

Figure 2: Flow Diagram of the X-11 Seasonal Adjustment in X-12-ARIMA Method

3 RegARIMA Model in X-12-ARIMA

3.1 Introduction

The RegARIMA model that is used in X-12-ARIMA method, also called regression model with ARIMA error, is an extension of the autoregressive integrated moving average (ARIMA) class of models. For RegARIMA, the time series is modeled with linear regression whereas the error term follows a seasonal ARIMA model. A general RegARIMA model is defined as:

$$Y_t = \beta^T X_t + z_t = \sum_i \beta_i x_{it} + z_t \qquad (3.1.1)$$

where

- Y_t is the original time series;
- x_{it} are the regressors for trading day, leap-year effect, moving holidays, additive outliers, temporary changes, level shifts and other user-defined effects, which observed concurrently with $\mathrm{Y_t}$;
- β_i are unknown regression coefficients;
- $z_t = \mathrm{Y_t} - \sum_i \beta_i x_{it}$ is a seasonal ARIMA process.

Usually we need to transform the nonlinear observed time series $\mathrm{Y_t}$ into logarithms of $\mathrm{Y_t}$, so that it will be adequately fit by a RegARIMA model. More generally,

$$y_t = \log\left(\frac{Y_t}{d_t}\right) = \sum_i \beta_i x_{it} + z_t \qquad (3.1.2)$$

where d_t is the sequence of combined trading day and Easter holiday effect factors obtained from a regression model of the irregular component of Y_t (obtained from a preliminary run in X-11 seasonal adjustment) or other modification for known external effects (e.g. economic factors).

Note that applying the standard regression methodology to time series data, which is that standard regression model assumes that the regression errors z_t in (3.1.1) and (3.1.2) are uncorrelated over time. However, for time series data, the errors in (3.1.1) will usually be autocorrelated. Thus, it is not appropriate to use ARIMA model on the error term without differencing (and hence, ARMA model). Hence, the errors in (3.1.1) will often require differencing if the observations from a time series show seasonal effects or trend that depend on the time, since this means that the errors in (3.1.1) are non-stationary. We then introduce the seasonal ARIMA model for z_t.

According to Shumway and Stoffer (2006), a general multiplicative seasonal ARIMA model, or SARIMA model, denoted by ARIMA$(p, d, q) \times (P, D, Q)_s$, is given by:

$$\phi(B)\Phi_{\mathrm{P}}(B^s)(1-B)^d\,(1-B^s)^D z_t = \theta(B)\Theta_{\mathrm{Q}}(B^s)a_t \qquad (3.1.3)$$

where

- B is the back shift operator ($Bz_t = z_{t-1}$);
- s is the seasonal period (s = 12 for monthly data);
- $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$ is the nonseasonal autoregressive (AR) operator;
- $\Phi_{\mathrm{P}}(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_{\mathrm{P}} B^{Ps}$ is the seasonal autoregressive (AR) operator;
- $\theta(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$ is the nonseasonal moving average (MA) operator;
- $\Theta_{\mathrm{Q}}(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_{\mathrm{Q}} B^{Qs}$ is the seasonal moving average (MA) operator;
- $(1-B)^d$ represents nonseasonal differencing of order d;
- $(1-B^s)^D$ implies seasonal differencing of order D;
- a_t are i.i.d white noise with mean zero and variance σ_a^2.

Substituting (3.1.2) into (3.1.3), we obtain the general RegARIMA model that allowed by the X-12-ARIMA method:

$$\phi(B)\Phi_{\mathrm{P}}(B^s)(1-B)^d\,(1-B^s)^D(y_t - \sum_i \beta_i x_{it}) = \theta(B)\Theta_{\mathrm{Q}}(B^s)a_t \qquad (3.1.4)$$

The above equation implies that firstly the regression effects are subtracted from $\mathrm{Y_t}$ to get the zero mean series z_t. Suppose $\omega_t = (1-B)^d\,(1-B^s)^D(\mathrm{y_t} - \sum_i \beta_i x_{it})$. Since a_t is assumed to be a sequence of independent variables with mean 0 and constant variance σ_a^2, it follows from these constrain that ω_t is a covariance stationary time series that satisfies the difference equation $\phi(B)\Phi_{\mathrm{P}}(B^s)\omega_t = \theta(B)\Theta_{\mathrm{Q}}(B^s)a_t$ (U.S. Bureau of the Census, 1999).

Hence, another way to write the RegARIMA model in (3.1.4) is

$$(1-B)^d\,(1-B^s)^D y_t = \sum_i \beta_i\{(1-B)^d\,(1-B^s)^D x_{it}\} + \omega_t \qquad (3.1.5)$$

where ω_t follows the stationary ARMA model i.e. ARMA$(p, q) \times (P, Q)$. Equation (3.1.4) implies that both the regression variables x_{it} and the time series $\mathrm{Y_t}$ are differenced by the ARIMA model differencing operator $(1 - B)^d\ (1 - B^s)^D$.

RegARIMA models are used for forecasting and backcasting, so that we are able to use symmetric filters at the beginning and at the end of the series (see the application of symmetric moving average filters in chapter 2). In addition, RegARIMA model can deal with calendar effects (e.g. trading day, leap year, and Easter Day) and outliers through regressors.

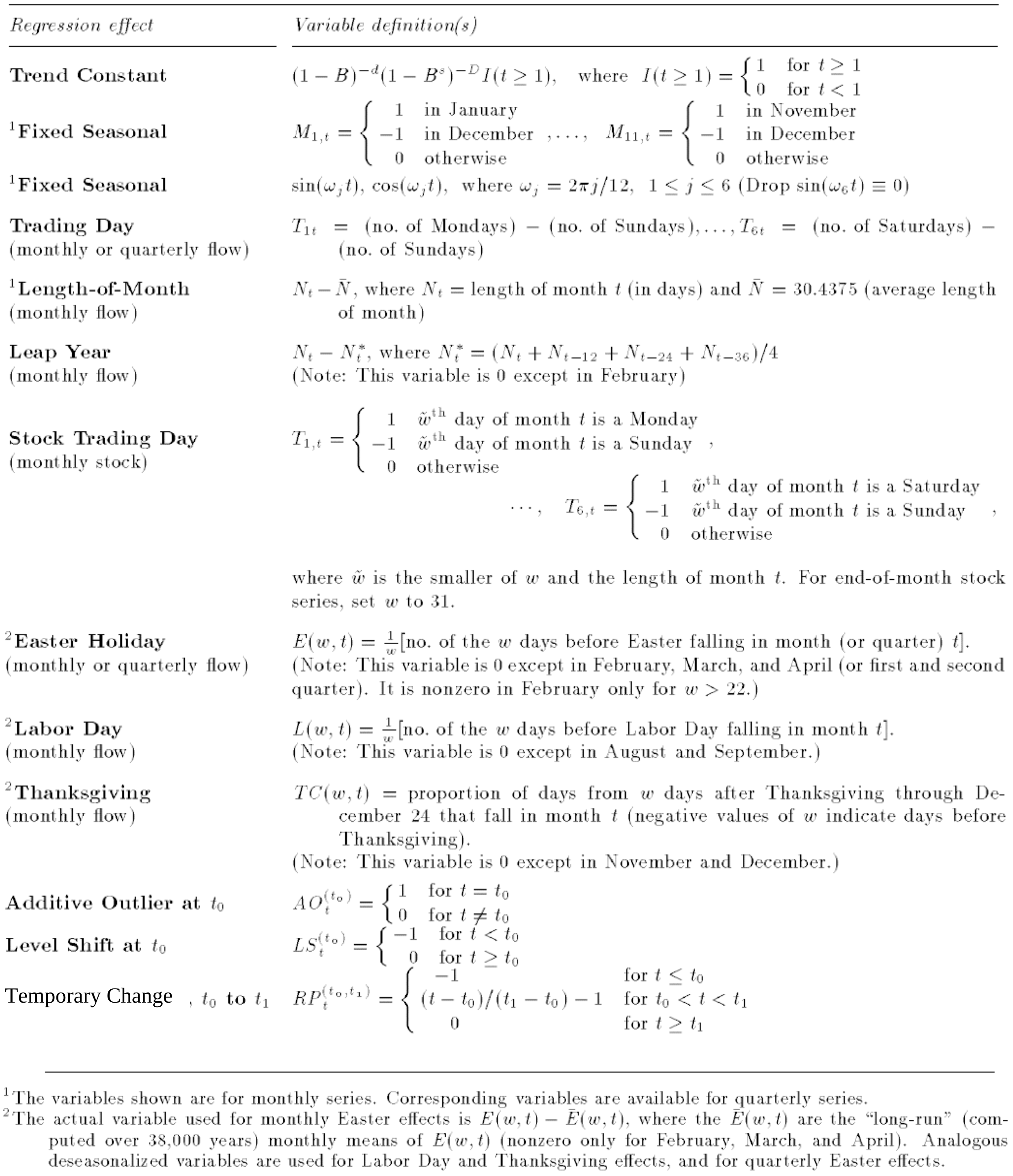
There are three types of outliers that can be identified by the model automatically: additive outliers, level shift outliers and temporary change outliers. Summarized by Chan (1995), these three types of outliers are defined as follows:

- Additive outlier (AO): an additive outlier appears as a surprisingly large or small value occurring for a single observation. Subsequent observations are unaffected by an additive outlier.
- Level shift (LS): for a level shift, all observations appearing after the outlier move to a new level. In contrast to additive outliers, a level shift outlier affects many observations and has a permanent effect.
- Temporary change (TC): temporary change outliers are similar to level shift outliers, but the effect of the outlier diminishes exponentially over the subsequent observations. Eventually, the series returns to its normal level.

The built-in regressors and auto-identified outliers in the X-12-ARIMA method will be presented in the sections 3.2. In section 3.3, we will briefly discuss the point forecasting using RegARIMA model in X-12-ARIMA.

3.2 Built-in Regressors

The set of built-in regressors for monthly series and their definitions is listed in Table 2 below. Note that the **Length-of-Month** and **Leap Year** regressors cannot be used with the **Trading Day** regressor, since the **Trading Day** regressor includes the effects of **Length-of-Month** and **Leap Year**. **Stock Trading Day** will be used only when the data collection day occurs on different days of the week. Applications using some of regressors are given in the project data analysis in chapter 4.



Regression effect	*Variable definition(s)*
Trend Constant	$(1-B)^{-d}(1-B^s)^{-D}I(t \geq 1)$, where $I(t \geq 1) = \begin{cases} 1 & \text{for } t \geq 1 \\ 0 & \text{for } t < 1 \end{cases}$
[1]**Fixed Seasonal**	$M_{1,t} = \begin{cases} 1 & \text{in January} \\ -1 & \text{in December} \\ 0 & \text{otherwise} \end{cases}, \ldots, \quad M_{11,t} = \begin{cases} 1 & \text{in November} \\ -1 & \text{in December} \\ 0 & \text{otherwise} \end{cases}$
[1]**Fixed Seasonal**	$\sin(\omega_j t)$, $\cos(\omega_j t)$, where $\omega_j = 2\pi j/12$, $1 \leq j \leq 6$ (Drop $\sin(\omega_6 t) \equiv 0$)
Trading Day (monthly or quarterly flow)	T_{1t} = (no. of Mondays) − (no. of Sundays), $\ldots, T_{6t}$ = (no. of Saturdays) − (no. of Sundays)
[1]**Length-of-Month** (monthly flow)	$N_t - \bar{N}$, where N_t = length of month t (in days) and $\bar{N} = 30.4375$ (average length of month)
Leap Year (monthly flow)	$N_t - N_t^*$, where $N_t^* = (N_t + N_{t-12} + N_{t-24} + N_{t-36})/4$ (Note: This variable is 0 except in February)
Stock Trading Day (monthly stock)	$T_{1,t} = \begin{cases} 1 & \tilde{w}^{\text{th}} \text{ day of month } t \text{ is a Monday} \\ -1 & \tilde{w}^{\text{th}} \text{ day of month } t \text{ is a Sunday} \\ 0 & \text{otherwise} \end{cases},$ $\ldots, \quad T_{6,t} = \begin{cases} 1 & \tilde{w}^{\text{th}} \text{ day of month } t \text{ is a Saturday} \\ -1 & \tilde{w}^{\text{th}} \text{ day of month } t \text{ is a Sunday} \\ 0 & \text{otherwise} \end{cases},$ where $\tilde{w}$ is the smaller of w and the length of month t. For end-of-month stock series, set w to 31.
[2]**Easter Holiday** (monthly or quarterly flow)	$E(w,t) = \frac{1}{w}$[no. of the w days before Easter falling in month (or quarter) t]. (Note: This variable is 0 except in February, March, and April (or first and second quarter). It is nonzero in February only for $w > 22$.)
[2]**Labor Day** (monthly flow)	$L(w,t) = \frac{1}{w}$[no. of the w days before Labor Day falling in month t]. (Note: This variable is 0 except in August and September.)
[2]**Thanksgiving** (monthly flow)	$TC(w,t)$ = proportion of days from w days after Thanksgiving through December 24 that fall in month t (negative values of w indicate days before Thanksgiving). (Note: This variable is 0 except in November and December.)
Additive Outlier at t_0	$AO_t^{(t_0)} = \begin{cases} 1 & \text{for } t = t_0 \\ 0 & \text{for } t \neq t_0 \end{cases}$
Level Shift at t_0	$LS_t^{(t_0)} = \begin{cases} -1 & \text{for } t < t_0 \\ 0 & \text{for } t \geq t_0 \end{cases}$
Temporary Change , t_0 **to** t_1	$RP_t^{(t_0,t_1)} = \begin{cases} -1 & \text{for } t \leq t_0 \\ (t-t_0)/(t_1-t_0)-1 & \text{for } t_0 < t < t_1 \\ 0 & \text{for } t \geq t_1 \end{cases}$

[1]The variables shown are for monthly series. Corresponding variables are available for quarterly series.

[2]The actual variable used for monthly Easter effects is $E(w,t) - \bar{E}(w,t)$, where the $\bar{E}(w,t)$ are the "long-run" (computed over 38,000 years) monthly means of $E(w,t)$ (nonzero only for February, March, and April). Analogous deseasonalized variables are used for Labor Day and Thanksgiving effects, and for quarterly Easter effects.

Table 2: Built-in Regressors for Monthly Series in X-12-ARIMA

Source: U.S. Bureau of the Census

URL: https://www.census.gov/ts/papers/jbes98.pdf

3.3 Point Forecasting

Point forecasts for future h months using RegARIMA model can be calculated using the following three steps:

1. Expand the RegARIMA equation so that y_t is on the left-hand side and all other terms are on the right.
2. Rewrite the equation by replacing t with $T + h$.
3. On the right-hand side of the equation, replace future observations with their forecasts, future errors with zero, and past errors with the corresponding residuals.

Beginning with $h = 1$, these steps are then repeated for $h = 2, 3, \dots$ until all forecasts have been calculated.

To better illustrate the process, here we will deploy an example. Suppose that for the logarithm of the original monthly time series y_t in (Eq. 3.1.1), the error term z_t follows ARIMA$(0,1,1) \times (0,1,1)_{12}$. From equation (3.1.5), we have the RegARIMA equation:

$$(1-B)^d\,(1-B^s)^D y_t = \sum_i \beta_i (1-B)^d\,(1-B^s)^D x_{it} + \omega_t$$

where ω_t follows ARMA$(p,q) \times (P,Q)$ is given by $\phi(B)\Phi_{\mathrm{P}}(B^s\,)\omega_t = \theta(B)\Theta_{\mathrm{Q}}(B^s)a_t$. Hence, now the RegARIMA model can be written as follows:

$$(1-B)(1-B^{12})y_t = \sum_i \beta_i\{(1-B)(1-B^{12})x_{it}\} + \omega_t$$

where ω_t follows ARMA$(0,1) \times (0,1)$is given by $\omega_t = \theta(B)\Theta(B^{12})a_t$. Then we expand the left hand side and right hand side to obtain

$$(1-B-B^{12}+B^{13})y_t = \sum_i \beta_i\{(1-B-B^{12}+B^{13})x_{it}\} + \omega_t$$

and applying the backshift operator ($By_t = y_{t-1}$) gives

$$y_t - y_{t-1} - y_{t-12} + y_{t-13} = \sum_i \beta_i\{x_{it} - x_{it-1} - x_{it-12} + x_{it-13}\} + \omega_t$$

Finally, moving all terms other than y_t to the right-hand side yields

$$y_t = y_{t-1} + y_{t-12} - y_{t-13} + \sum_i \beta_i \{x_{it} - x_{it-1} - x_{it-12} + x_{it-13}\} + \omega_t \tag{3.3.1}$$

where

$$\begin{aligned}\omega_t &= \theta(B)\Theta(B^{12})a_t \\ &= \left(1 + \hat{\theta}_1 B\right)\Theta(B^{12})a_t \qquad (\because \theta(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)) \\ &= \left(1 + \hat{\theta}_1 B\right)\left(1 + \widehat{\Theta}_1 B^{12}\right)a_t \qquad (\because \Theta_{\mathrm{Q}}(B^s) = 1 - \Theta_1 B^s - \cdots - \Theta_{\mathrm{Q}} B^{Qs}) \\ &= \left(1 + \hat{\theta}_1 B + \widehat{\Theta}_1 B^{12} + \widehat{\Theta}_1 \hat{\theta}_1 B^{13}\right)a_t \\ &= a_t + \hat{\theta}_1 a_{t-1} + \widehat{\Theta}_1 a_{t-12} + \widehat{\Theta}_1 \hat{\theta}_1 a_{t-13} \qquad (\because Ba_t = a_{t-1})\end{aligned}$$

This completes the first step. For the second step, we replace t with $T + 1$ in (3.3.1):

$$\begin{aligned}y_{T+1} &= y_{T+1-1} + y_{T+1-12} - y_{T+1-13} + \sum_i \beta_i \{x_{iT+1} - x_{iT+1-1} - x_{iT+1-12} + x_{iT+1-13}\} + \omega_{T+1} \\ &= y_T + y_{T-11} - y_{T-12} + \sum_i \beta_i \{x_{iT+1} - x_{iT} - x_{iT-11} + x_{iT-12}\} + \omega_{T+1}\end{aligned}$$

and

$$\omega_{T+1} = a_{T+1} + \hat{\theta}_1 a_T + \widehat{\Theta}_1 a_{T-11} + \widehat{\Theta}_1 \hat{\theta}_1 a_{T-12}.$$

Assuming we have observations up to time T, all values on the right-hand side are known except for a_{T+1}, which we replace with zero. For a_T, a_{T-11} and a_{T-12}, replacing with the last observed residuals e_T, e_{T-11}, and e_{T-12}, respectively returns

$$\widehat{\omega}_{T+1|T} = 0 + \hat{\theta}_1 e_T + \widehat{\Theta}_1 e_{T-11} + \widehat{\Theta}_1 \hat{\theta}_1 e_{T-12} \; = \; \hat{\theta}_1 e_T + \widehat{\Theta}_1 e_{T-11} + \widehat{\Theta}_1 \hat{\theta}_1 e_{T-12} \tag{3.3.2}$$

and hence

$$\hat{y}_{T+1|T} = y_T + y_{T-11} - y_{T-12} + \sum_i \beta_i \{x_{iT+1} - x_{iT} - x_{iT-11} + x_{iT-12}\} + \widehat{\omega}_{T+1|T}$$

where $\widehat{\omega}_{T+1|T}$ is defined in (3.3.2).

A forecast of y_{T+2} is obtained by replacing t with $T+2$ in (3.3.1). All values on the right hand side will be known at time T except y_{T+1} which we replace with $\hat{y}_{T+1|T}$, and a_{T+1} and a_{T+2}, both of which we replace with zero:

$$\hat{y}_{T+2|T} = \hat{y}_{T+1|T} + y_{T-10} - y_{T-11} + \sum_i \beta_i \{x_{iT+2} - x_{iT+1} - x_{iT-10} + x_{iT-11}\} + \widehat{\omega}_{T+2|T}$$

where

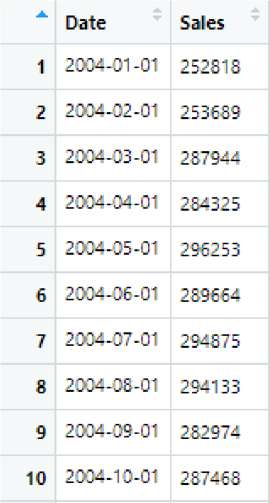
$$\widehat{\omega}_{T+2|T} = \widehat{\Theta}_1 e_{T-10} + \widehat{\Theta}_1 \hat{\theta}_1 e_{T-11}.$$

The process continues in this manner for all future time periods. In this way, any number of point forecasts can be obtained.

4 Monthly Retail Sales in US: Data Analysis and Forecasting

4.1 Data Preprocessing

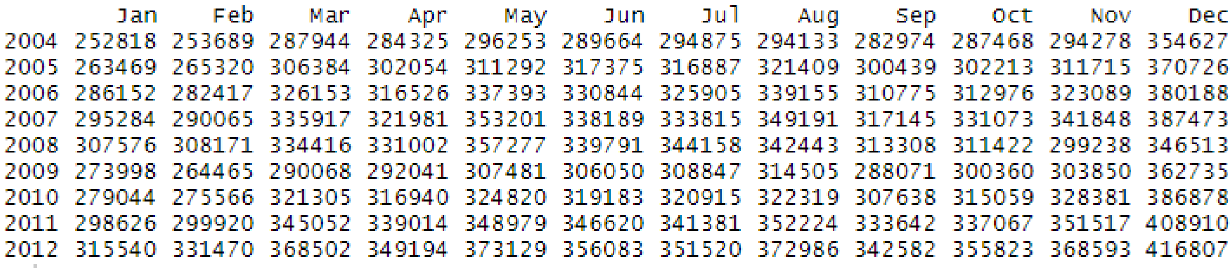
In this project, we investigate a non-confidential and non-seasonally adjusted retail sales dataset, which has been collected monthly by U.S. Bureau of the Census between calendar year 2010 to 2019. The cleaned dataset contains two columns, *Date* and *Sales*, with 108 observations on the monthly retail sales (in millions) in US for every month from 2004 to 2013. The dataset is split into a training dataset including 88.9% of the total monthly retail sales (96 observations between January 2004 and December 2012) and a testing dataset including 11.1% of the total monthly retail sales (12 observations between January 2013 and December 2013). A preview of the training dataset is shown below:



	Date	Sales
1	2004-01-01	252818
2	2004-02-01	253689
3	2004-03-01	287944
4	2004-04-01	284325
5	2004-05-01	296253
6	2004-06-01	289664
7	2004-07-01	294875
8	2004-08-01	294133
9	2004-09-01	282974
10	2004-10-01	287468

Table 3: Preview of Training Dataset of Monthly Retail Sales in U.S.

The monthly retail sales (Y_t) form a time series. Transforming the data frame above to time series format in R give us the table below:



```
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2004 252818 253689 287944 284325 296253 289664 294875 294133 282974 287468 294278 354627
2005 263469 265320 306384 302054 311292 317375 316887 321409 300439 302213 311715 370726
2006 286152 282417 326153 316526 337393 330844 325905 339155 310775 312976 323089 380188
2007 295284 290065 335917 321981 353201 338189 333815 349191 317145 331073 341848 387473
2008 307576 308171 334416 331002 357277 339791 344158 342443 313308 311422 299238 346513
2009 273998 264465 290068 292041 307481 306050 308847 314505 288071 300360 303850 362735
2010 279044 275566 321305 316940 324820 319183 320915 322319 307638 315059 328381 386878
2011 298626 299920 345052 339014 348979 346620 341381 352224 333642 337067 351517 408910
2012 315540 331470 368502 349194 373129 356083 351520 372986 342582 355823 368593 416807
```

Table 4: Time Series of Training Dataset of Monthly Retail Sales in U.S.

The Package ‘x12’ in R will be used to proceed with this project. The overall objectives for this project are two-fold: (1) to review, analyze and discuss historical patterns in monthly retail sales training dataset in US, which will be given in sections 4.2, 4.3, and 4.4. Moreover, (2) to test forecasting models and run short-term forecasting of monthly retail sales in US for future 12 months, and compared the predicted results to the testing dataset. The single variable as the monthly retail sales in US will be used to forecasting. This will be presented in section 4.5. Finally, the forecasting results will be compared between the models by using X-12-ARIMA method and pure seasonal ARIMA method in section 4.6. The R code for this project are in Appendix B.

4.2 Plotting Time Series

Applied ggplot() function in R to plot the time series of monthly retail sales from January 2004 to December 2012 give us the following graph:

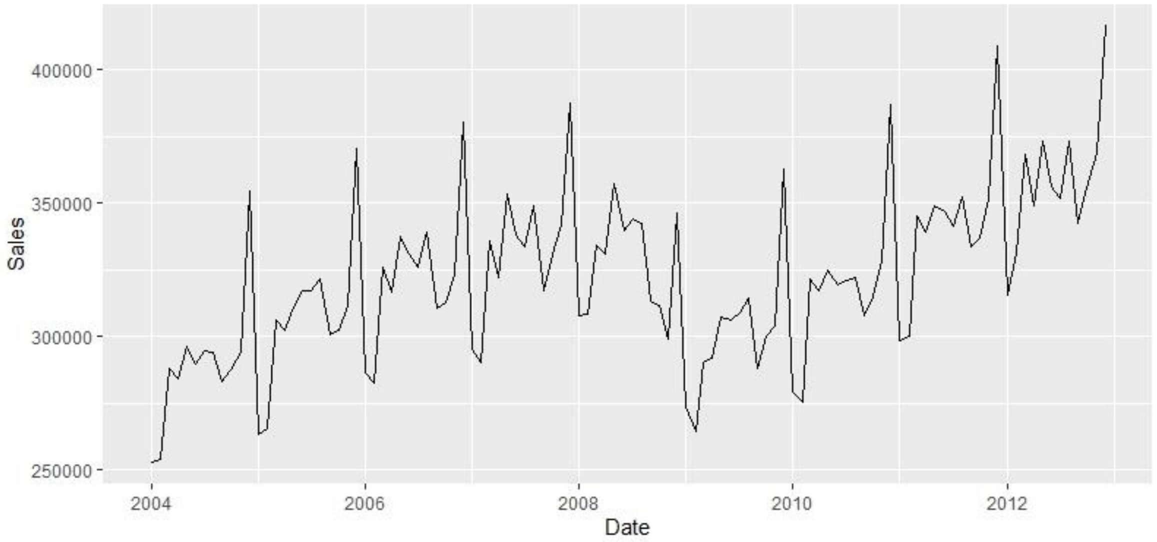


Figure 3: Plot of Original Time Series of Monthly Retail Sales (in Millions) from January 2004 to December 2012

We can see from this time series that there seems to be seasonal variations in the retail sales per month: there is a peak every December, and generally, the summer retail sales and March retail sales are higher than the other months. Also, from the plot we can see that there may be an increasing trend from 2004 to the end of 2008, and then there is a sudden decrease and a level shift at the end of 2008. After that, it seems that there is an increasing trend again. One of the major reasons causes this level decline at the end of 2008 may be related to the 2008 Financial Crisis in US, which started on October 2008. This crisis led to the Great Recession, where the purchasing power of people might drop significantly.

Since the time series appears to have some seasonality and trend, it is non-stationary. Hence, the time series will be required differencing. In addition, it seems that the sizes of both the seasonal and irregular variations roughly increase as the level of the time series rises, this may be due to the inflation, which associated to the consumer price index (CPI). Hence, a multiplicative model could be suggested for decomposing the time series (see section 2.2). Moreover, we may need to take the natural log on the original time series.

To better understand the regular seasonality of monthly retail sales year over year, the general trend of monthly retail sales from 2004 to 2012, as well as how the 2008 Financial Crisis affect the retail sales in US, we will apply the X-12-ARIMA method (Package 'x12' package in R) to decompose and analyze the time series data.

4.3 Model Selection by Using Package 'x12' in R

Recall the RegARIMA model used in X-12-ARIMA method in (3.1.2) that the regressors x_{it} are the regressors for trading day effects, moving holidays effects (majorly Easter effects), additive outliers, temporary change outliers, level shift outliers and other user-defined effects, which observed concurrently with the transformed time series y_t; whereas the error term z_t follows a seasonal ARIMA model. The program will automatically detect and select the model for the seasonal ARIMA error z_t.

For testing the significance of regressors, one can use the AIC statistics. The AIC statistics can be used to evaluate whether or not a particular regressor is preferred, compared to not having the particular regressor in the model. It quantifies the goodness of fit. When comparing two models, the one with the lower AIC is generally better (see Appendix A).

4.3.1 Testing for Trading Day Effects

Trading day effects arise since the number of occurrences of each day of the week, in a month, differs from year to year. For example, the number of occurrences for June is calculated across three years, from 2005 to 2007, shown as below Table 5:

	Number of Days in June		
Year	**2005**	**2006**	**2007**
Monday	4	4	4
Tuesday	4	4	4
Wednesday	5	4	4
Thursday	5	5	4
Friday	4	5	5
Saturday	4	4	5
Sunday	4	4	4

Table 5: Day Composition for June 2005, 2006 and 2007

These differences may cause regular effects in the monthly retail sales series. For example, the retail sales on Friday and Saturday are generally higher than on other days of the week. Thus, it is likely that the series has a slightly lower value in June 2005 than in June 2006 and June 2007. The trading day effects can be removed by adding the regressor **Trading Day** to the model. Hence, we would like to test the significance of trading day effects by comparing the AIC coefficients for the models with and without **Trading Day** regressor.

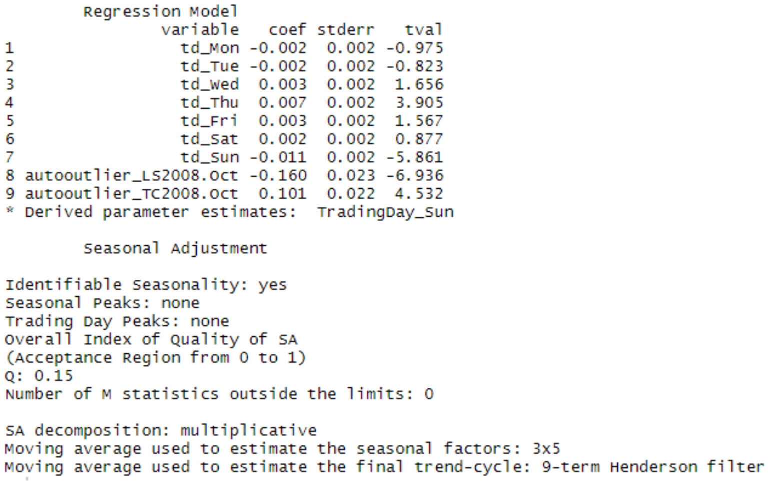
Since from section 4.2, it is obvious that there is a sudden decrease and level shift at the end of 2008, automatic outliers regressor will be first added to the model. Using the natural log transformation on the original time series and multiplicative decomposition, together with the automatic outliers detection regressor, the automatic seasonal ARIMA model detected by the program is $\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$, and the automatic outlier regressors added by the program are an additive outlier on October 2008 and a level shift outlier on October 2008. This shows as Model T1 in Table 6 below. The AIC coefficient in the summary of this model is 1949.0331.

	Model		
Model ID	**Regressor(s)**	**ARIMA Error**	**AIC**
T1	Automatic Outliers	$\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$	1952.5514
T2	Trading Day + Automatic Outliers	$\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$	1864.3378

Table 6: AIC Comparison in Sample for Models with Different Regressors to Test the Significance of Trading Day Effects based on X-12-ARIMA Seasonal Adjustment Method

Adding a trading day regressor give us Model T2 in Table 6, which has a lower AIC value 1864.3378. Hence, the better model to fit the time series is Model T2, i.e. the trading day effects are significant.

Alternatively, from the below summary of Model T2 (Figure 4), we see that the t-values for the trading days are significant for most of days in the week, i.e. the coefficients are significantly different from zero, except for Monday, Tuesday and Saturday. However, the p-value shows that the **Trading Day** regressor is significant as a combination for all seven days in a week. Hence, we conclude that the trading day effects are needed to be adjusted, i.e. we will add the regressor of trading day into our final model.



```
        Regression Model
                variable   coef stderr   tval
1                 td_Mon -0.002  0.002 -0.975
2                 td_Tue -0.002  0.002 -0.823
3                 td_Wed  0.003  0.002  1.656
4                 td_Thu  0.007  0.002  3.905
5                 td_Fri  0.003  0.002  1.567
6                 td_Sat  0.002  0.002  0.877
7                 td_Sun -0.011  0.002 -5.861
8 autooutlier_LS2008.Oct -0.160  0.023 -6.936
9 autooutlier_TC2008.Oct  0.101  0.022  4.532
* Derived parameter estimates:  TradingDay_Sun

        Seasonal Adjustment

Identifiable Seasonality: yes
Seasonal Peaks: none
Trading Day Peaks: none
Overall Index of Quality of SA
(Acceptance Region from 0 to 1)
Q: 0.15
Number of M statistics outside the limits: 0

SA decomposition: multiplicative
Moving average used to estimate the seasonal factors: 3x5
Moving average used to estimate the final trend-cycle: 9-term Henderson filter
```

Figure 4: Partial Summary of Model T2 in R

4.3.2 Testing for Easter Effects

The date of Easter Sunday is the first Sunday after the first full moon of the spring equinox, and based on the Georgian calendar, it can be anywhere between March 22nd and April 25th. As with seasonal effects, it is desirable to estimate and remove Easter effects from time series to help interpretation since there may be increases in sales due to Easter holiday in different months of the year.

The Easter effects can be removed by adding the regressor **Easter[*w*]** to the model, which indicates the change in the level of activity occurs *w* days prior to the Easter Sunday. For example, **Easter[1]** would mean that the holiday effect is estimating the change in the level of daily activity for the Saturday before Easter, whereas **Easter[8]** assumes the level of activity occurs or the whole week prior to Easter Sunday, i.e. all 8 days before.

If no Easter correction is made by the X-12-ARIMA method, the additional sales will be initially put in the SI ratios. The monthly SI ratios are then smoothed, which leaving the Easter effect in the irregular fluctuations. The final seasonal adjusted series will thus show peaks and troughs due to the effects of Easters.

Using the similar analysis as testing the significance of trading day effects, by adding, removing and changing some Easter effect parameters, we obtain the results in Table 7 below:

	Model		
Model ID	**Regressor(s)**	**ARIMA Error**	**AIC**
E1(= T1)	Automatic Outliers	ARIMA$(0,1,1) \times (0,1,1)_{12}$	1952.5514
E2	Automatic Outliers + Easter[1]	ARIMA$(0,1,1) \times (0,1,1)_{12}$	1954.036
E3	Automatic Outliers + Easter[8]	ARIMA$(0,1,1) \times (0,1,1)_{12}$	1954.1752
E4	Automatic Outliers + Easter[15]	ARIMA$(0,1,1) \times (0,1,1)_{12}$	1954.4974
E5	Automatic Outliers + Trading Day + Easter[1]	ARIMA$(0,1,1) \times (0,1,1)_{12}$	1866.3021

Table 7: AIC Comparison in Sample for Models with Different Regressors to Test the Significance of Easter Effects based on X-12-ARIMA Seasonal Adjustment Method

Where Model E1 is the same as Model T1 in section 4.3.1. The results above indicate that the Easter effects are not significant in this monthly retail sales time series, as adding an **Easter[w]** regressor to the model does not have significant impact on the coefficients of AIC. In addition, comparing the AIC of Model E5 to the AIC of Model T2 from the previous section, one suggests that with a **Trading Day** regressor in the model, adding an **Easter[w]** regressor reduces the performance of the model. Hence, we will not include the **Easter[w]** regressor in the final model.

From the analysis of sections 4.3.1 and 4.3.2, by far, the best model is Model T2 in section 4.3.1, which is combined with **Trading Day** regressor**, Automatic Outliers** regressor and ARIMA$(0,1,1) \times (0,1,1)_{12}$ errors. The AIC coefficient of this model is 1864.3378.

4.3.3 Testing for Different ARIMA Parameters

Although in most cases the automatic modelling options will provide satisfactory models (at least for the limited forecasting horizons used in seasonal adjustment) (National Statistics, 2007), we can change some seasonal ARIMA parameters in the model and compare to the automatic model to see if there is better model than the automatic one. Changing the automatic ARIMA model for the error term in Model T2 from $\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$ to $\text{ARIMA}(1,1,0) \times (1,1,0)_{12}$, $\text{ARIMA}(0,1,1) \times (1,1,1)_{12}$ and $\text{ARIMA}(1,1,1) \times (1,1,1)_{12}$ generating the following results in AIC coefficients:

	Model		
Model ID	**Regressor(s)**	**ARIMA Error**	**AIC**
A1 (= T2)	Automatic Outliers + Trading Day	$\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$	1864.3378
A2	Automatic Outliers + Trading Day	$\text{ARIMA}(1,1,0) \times (1,1,0)_{12}$	1902.1655
A3	Automatic Outliers + Trading Day	$\text{ARIMA}(0,1,1) \times (1,1,1)_{12}$	1850.5602
A4	Automatic Outliers + Trading Day	$\text{ARIMA}(1,1,1) \times (1,1,1)_{12}$	1851.9084

Table 8: AIC Comparison in Sample for Models with Different seasonal ARIMA parameters based on X-12-ARIMA Seasonal Adjustment Method

The results in Table 8 imply that Model A3 has the lowest AIC among the four models, where the error term follows $\text{ARIMA}(0,1,1) \times (1,1,1)_{12}$. This indicates that Model A3 is the best model among the models that we introduce above.

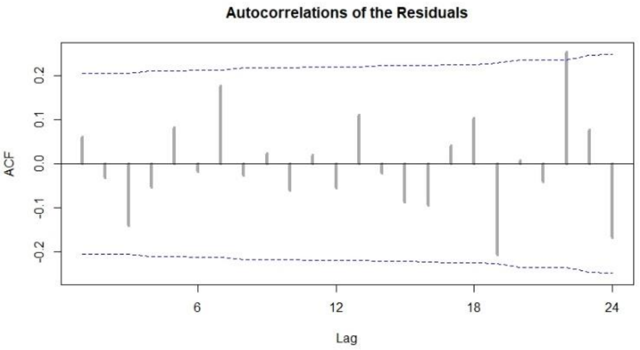


Figure 5: Autocorrelations of the Residuals of Model A3

Also, by plotting the autocorrelations of the residuals of Model A3 in Figure 5, we observe that he residuals are not significantly different from white noise. Hence, we will focus on applying Model A3 as the seasonal adjustment model in this project to analyze the time series as well as to predict the future monthly retail sales.

4.4 Plots Analysis of Seasonal Adjustment

Applying Model A3 in X-12-ARIMA method to decompose the original time series, we obtain the following equation as the final estimated decomposition (see section 2.4):

$$Y_t = C_t^{(3)} \times S_t^{(2)} \times I_t^{(3)}$$

where $C_t^{(3)}$ is the final trend estimate of the original time series, $S_t^{(2)}$ is the final seasonal component estimate of the original time series, and $I_t^{(3)}$ is the final irregular fluctuations estimate of the original time series. Plotting them respectively show the below results:

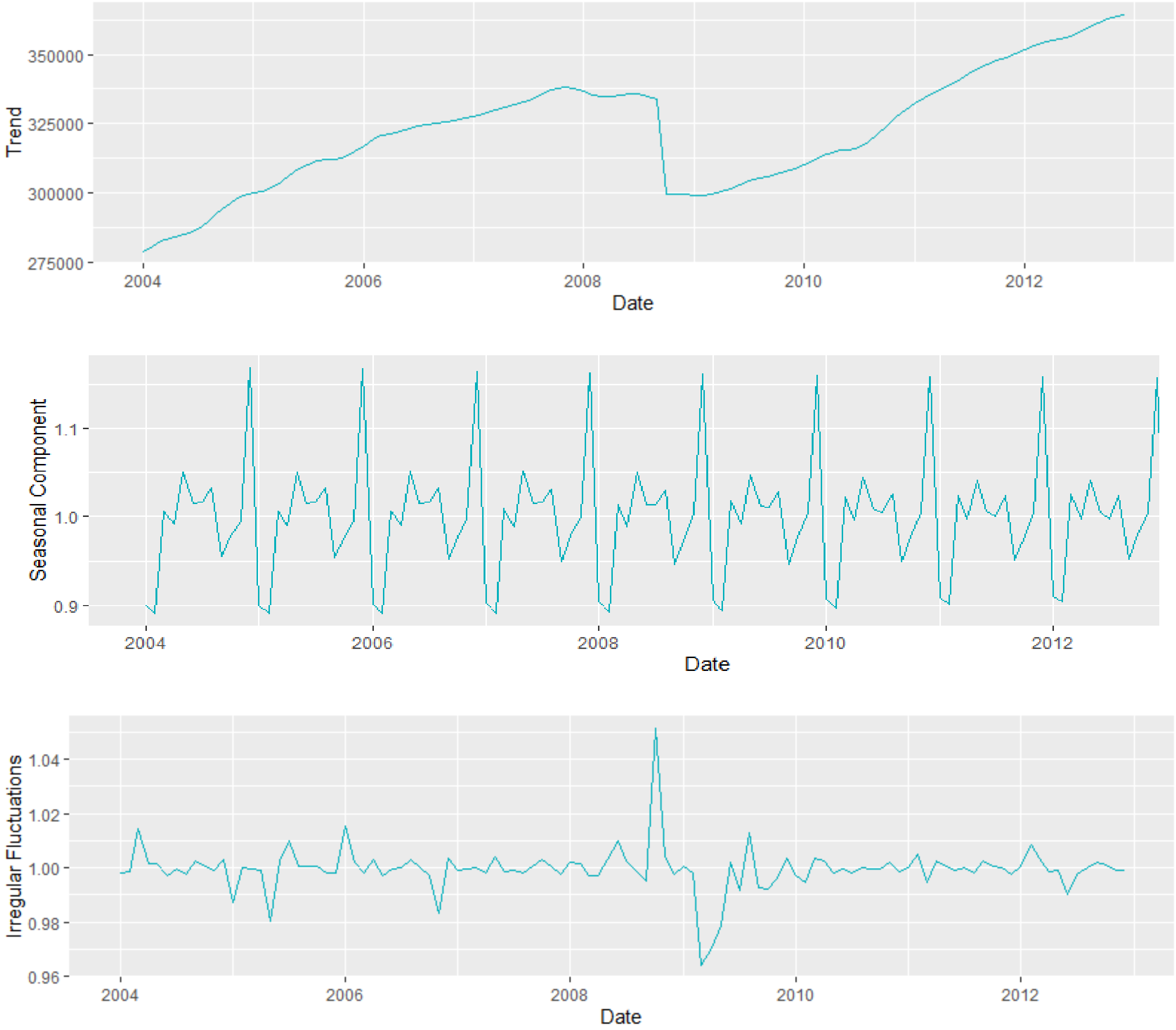


Figure 6: Plots of Final Trend Estimate, Final Seasonal Component Estimate and Final Irregular Fluctuations Estimate of the Original Time Series by Using Model A3

The final trend estimate $C_t^{(3)}$demonstrates the long-term changing trend of retail sales in U.S. with seasonal components and irregular fluctuations filtered from the original time series. From the plot of final trend estimate, we can see that the 2008 Financial Crisis causes the level of retail sales shifts down dramatically by around 11.54%, and which does not include the increasing inflation rate yet. It takes almost two and a half years for the retail sales in U.S. to increase back to its level before the 2008 Financial Crisis. Besides the level shift in October 2008, generally speaking, the retail sales in US tends to increase a little annually in long term.

The plot of seasonal component estimate $S_t^{(2)}$ indicate the seasonal distribution characteristic of retail sales in U.S. and the regularity that seasonal components change with time. The seasonal components reach the highest peak in December, which may be due to the Christmas holiday each year, and they reach the lowest peak around January and February. In addition, the sub-peaks appear during the summer time each year. Therefore, it is obvious that the time series of retail sales in U.S. is affected by seasons.

After seasonal components and the trend are separated, the irregular fluctuations estimate $I_t^{(3)}$ including unpredictable fluctuations in the retail sales in U.S. Besides the outliers at the end of 2008 and at the beginning of 2019, there are little peaks in March in the most of years, which are due to the fact that we removed the Easter effects regressor from the model.

Finally, the final seasonally adjusted data $A_t^{(2)} = Y_t/S_t^{(2)} = C_t^{(3)} \times I_t^{(3)}$ (i.e. eliminating the seasonality from the original time series) shows as following, which just contains the trend and irregular fluctuations:

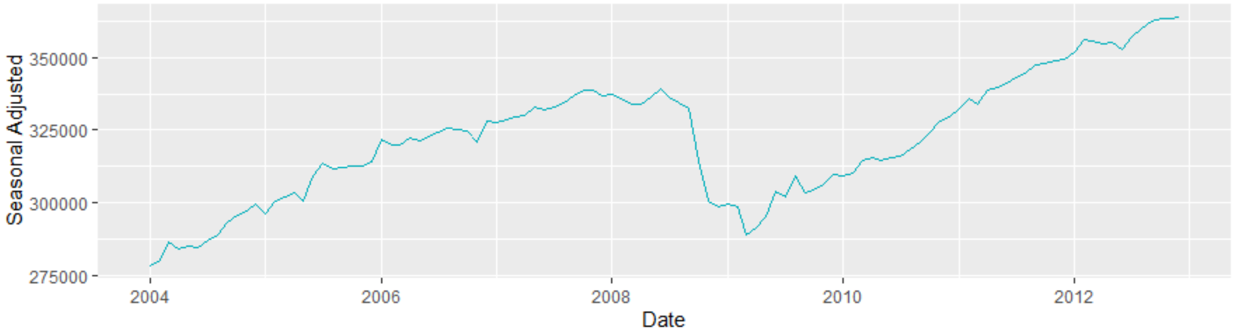


Figure 7: Plot of Final Seasonally Adjusted Data by Using Model A3

There are also some built-in plots options for the package x12 in R. For example, plotting the Seasonal factors and SI ratios by month using plotSeasFac() function gives us Figure 8:

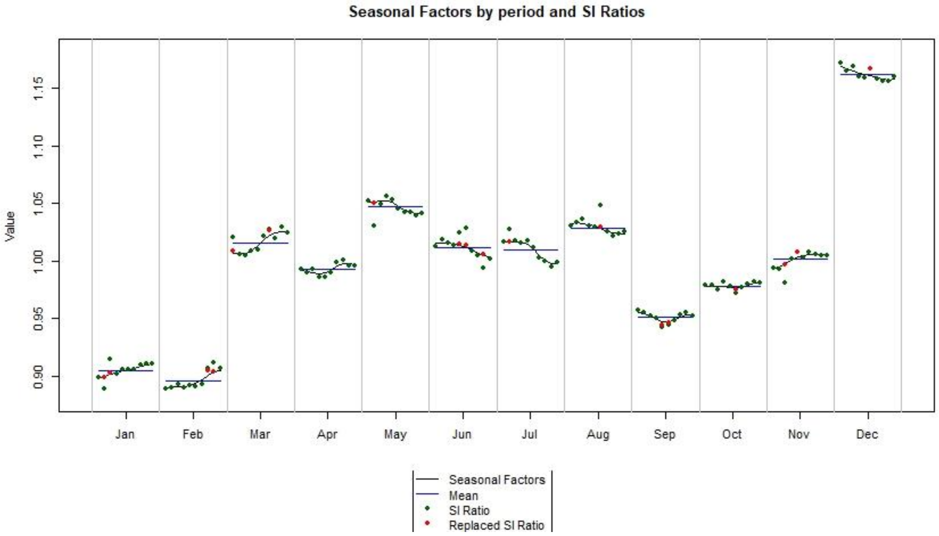


Figure 8: Plot of Seasonal Factors and SI Ratios by Month by Using plotSeasFac() function

Where the red dots are which the outliers are identified, removed and replaced (see Stage A in X11 seasonally adjustment procedures). The seasonal factors and SI ratios in Figure 8 indicate that generally speaking, the retail sales in U.S. are the highest in December each year. Furthermore, the retail sales during summer are higher than in other months, except in March, where there may be the effects from the retail sales boosting by Easter holiday (in the irregular fluctuations).

Figure 9 below shows the original series, seasonally adjusted series and trend together in one plot, as well as the automatically detected outliers, they are: an additive outlier on October 2008, a level shift on October 2008, and a temporary change on March 2009.

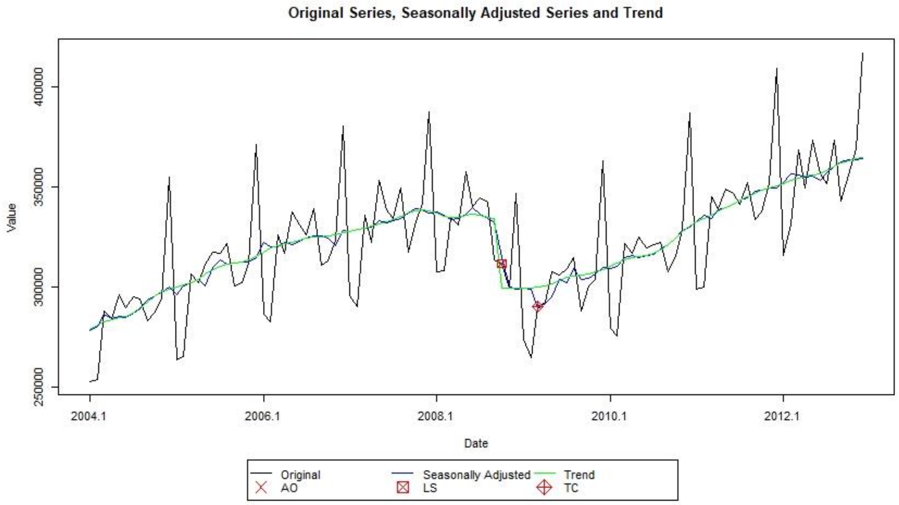


Figure 9: Plot of Original Series, Seasonally Adjusted Series, Trend, and Outliers by Using Model A3

4.5 Forecasting and Validation

Figure 10 presents the plot of forecasting by the Model A3 for January 2013 to December 2013. Here the predicted values are plotted as a blue line, and the 95% interval as a light blue shaded area. We can see that the time series of forecasts follows the regular seasonal pattern and the slightly increasing trend. However, the time series of forecasts is slightly smoother than the time series of the original data.

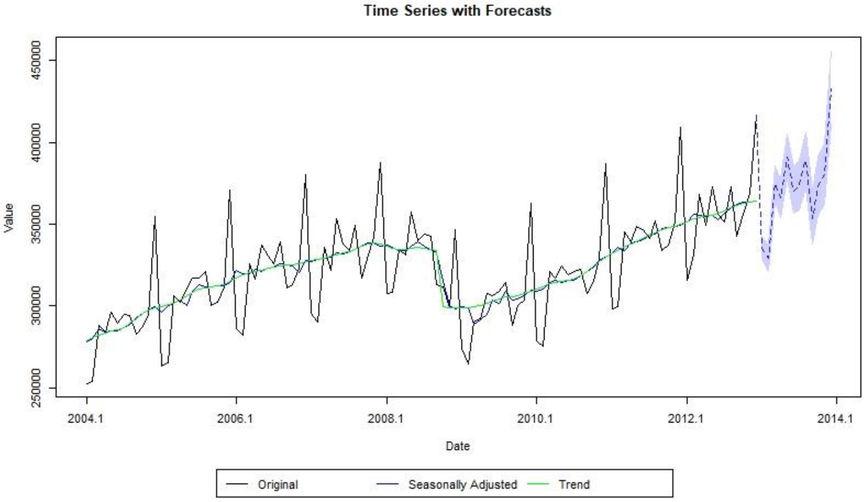
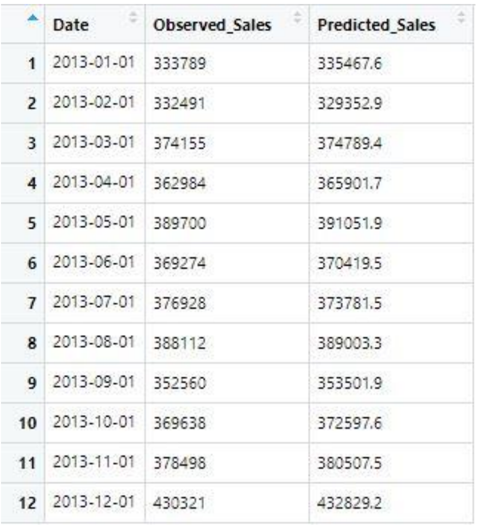


Figure 10: Time Series with Point Forecasts and 95% Prediction Interval by Using Model A3

Table 9 below summarizes the observed values (actual values) of retail sales in U.S. in 2013 and the predicted values forecasted by Model A3:



	Date	Observed_Sales	Predicted_Sales
1	2013-01-01	333789	335467.6
2	2013-02-01	332491	329352.9
3	2013-03-01	374155	374789.4
4	2013-04-01	362984	365901.7
5	2013-05-01	389700	391051.9
6	2013-06-01	369274	370419.5
7	2013-07-01	376928	373781.5
8	2013-08-01	388112	389003.3
9	2013-09-01	352560	353501.9
10	2013-10-01	369638	372597.6
11	2013-11-01	378498	380507.5
12	2013-12-01	430321	432829.2

Table 9: Observed Values and Predicted Values for Monthly Retail Sales (in Millions) in 2013 by Using Model A3 in X-12-ARIMA Method

As a measure of the accuracy of the forecasts, we can calculate the mean absolute percentage error (MAPE) for the forecast errors, which expresses the accuracy as a ratio defined by the formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where A_t are the actual values and F_t are the forecast values, $t = 1, \dots n$. Computing the MAPE by using R provides the result below:

$$MAPE_{A3} = 0.00526 = 0.526\%$$

which indicates that the observed values and the predicted values are relatively closed, as the value of MAPE is very low.

4.6 Comparison of Sales Prediction from Pure Seasonal ARIMA Model and X-12-ARIMA Method

Recall from sections 2.4 and 3.1, unlike X-12-ARIMA method, which applies a combination of X-11 method and differencing the error term at a lag that equal to the number of seasons (s) to seasonal adjust the time series, a traditional pure seasonal ARIMA model only uses differencing at a lag equal to the number of seasons (s) to remove the seasonal effects. Here, $s = 12$. In addition, modelling using X-12-ARIMA method could cover the calendar effects by adding built-in or user-defined regressors in the model, whereas pure seasonal ARIMA model does not have this function.

In the above section 4.5, we apply the X-12-ARIMA method to obtain the predicted values of retail sales for future twelve months. To illustrate the performance of the predictive model by using X-12-ARIMA method, as well as to compare the results to the pure seasonal ARIMA model, we will build a pure seasonal ARIMA model on the original time series of retail sales, Y_t, and apply the model to forecast the monthly retail sales from January 2013 to December 2013 in this section.

By analyzing the ACF and PACF of the first differenced and then seasonally differenced time series, i.e. $(1-B)(1-B^{12})Y_t$ (see Figure 11 below), we consider both ACF and PACF dies down. Hence, according to Shumway and Stoffer (2006), we would model a tentative ARMA(1,1) to the differenced series, i.e. pure seasonal ARIMA model $\text{ARIMA}(1,1,1) \times (1,1,1)_{12}$ for the original time series Y_t.

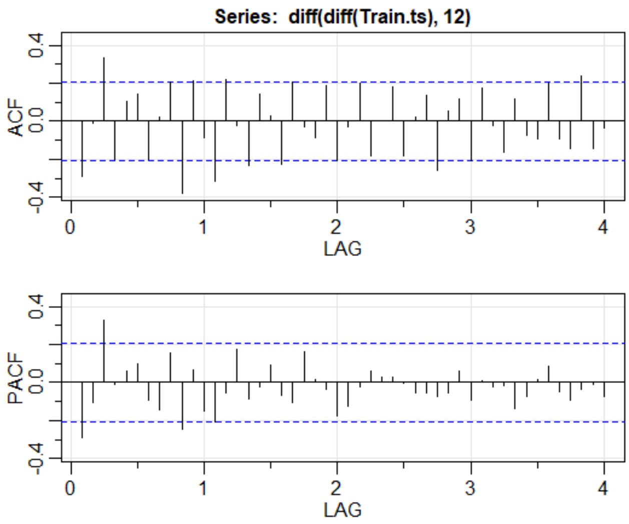
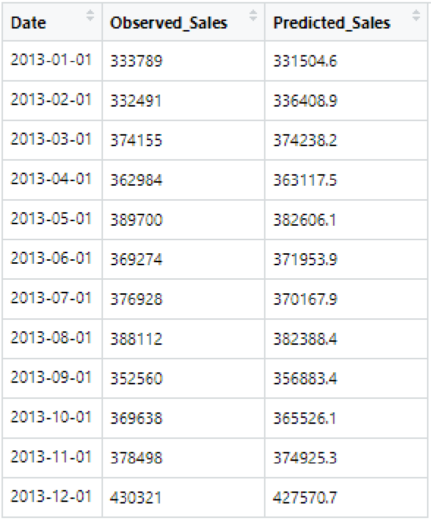


Figure 11: ACF and PACF of the first differenced and then seasonally differenced time series, $(\mathbf{1}-\boldsymbol{B})(\mathbf{1}-\boldsymbol{B}^{\mathbf{12}})\boldsymbol{Y}_{\boldsymbol{t}}$

Table 10 below summarizes the observed values (actual values) of retail sales in U.S. in 2013 and the predicted values forecasted by the pure seasonal ARIMA model ARIMA$(1,1,1) \times (1,1,1)_{12}$:



Date	Observed_Sales	Predicted_Sales
2013-01-01	333789	331504.6
2013-02-01	332491	336408.9
2013-03-01	374155	374238.2
2013-04-01	362984	363117.5
2013-05-01	389700	382606.1
2013-06-01	369274	371953.9
2013-07-01	376928	370167.9
2013-08-01	388112	382388.4
2013-09-01	352560	356883.4
2013-10-01	369638	365526.1
2013-11-01	378498	374925.3
2013-12-01	430321	427570.7

Table 10: Observed Values and Predicted Values for Monthly Retail Sales (in Millions) in 2013 by Using Pure Seasonal ARIMA Model, ARIMA$(\mathbf{1},\mathbf{1},\mathbf{1}) \times (\mathbf{1},\mathbf{1},\mathbf{1})_{\mathbf{12}}$

Similar to section 4.5, computing the mean absolute percentage error by using R provides:

$$MAPE_{SARIMA} = 0.009786 = 0.979\%$$

which is greater than $MAPE_{A3} = 0.526\%$, the mean absolute percentage error of Model A3 by using X-12-ARIMA method in section 4.5 above, i.e. $MAPE_{SARIMA} > MAPE_{A3}$. Hence, Model A3 by using X-12-ARIMA method is more accurate than the pure seasonal ARIMA model, regarding to predicting the retail sales for future twelve months.

Indeed, from the diagnostics for the pure seasonal ARIMA model ARIMA$(1,1,1) \times (1,1,1)_{12}$ (see Appendix 7.3.10), we observe a significant amount of outliers in the series as exhibited in the plot of the standardized residuals and the tails of normal Q-Q plot, and some autocorrelations that still remains. All of these implies that the pure seasonal ARIMA model does not fit as well as Model A3 by using X-12-ARIMA method. One of the reasons may be coming from the lack of trading day effects covered by the pure seasonal ARIMA model.

5 Conclusions

5.1 Summary and Findings

This paper provides analysis and forecasting of monthly retail sales in U.S. based on X-12-ARIMA method. The best X-12-ARIMA model contains regressors for trading day effects and automatic outliers, whose errors follow seasonal ARIMA model ARIMA$(0,1,1) \times (1,1,1)_{12}$, since the Akaike Information Criterion (AIC) and the mean absolute percentage error (MAPE) of this model are lower than the other models.

Using the above-mentioned model to decompose the time series of retail sales in U.S., it shows that the time series of retail sales in U.S. follows the changing regularity of highly consistent seasonality. The financial crisis in 2008 and the resulting recession causes the level shifts down by around 11.54%, and it takes roughly two and a half year to increases back to its level before the recession. Besides that, the trend of retail sales in U.S. increases constantly in the normal periods.

The forecasts for the monthly retail sales in 2013 also suggest that the trend is continuously increasing in the normal periods. If these results can be generalized for future years under normal situation, then it suggests that both the U.S. government and the private retailers need to prepare for the increased amount of retail sales in U.S., and the government would need to implementing policies to control the inflation rate, and hence the pricing. For example, the government can control the money supply though the central bank to ensure that the inflation rate will be adjusted to an appropriate level.

Compare X-12-ARIMA method to the pure seasonal ARIMA method in section 4.6, we observe that the model by using X-12-ARIMA method is more accurate than the model by using pure seasonal ARIMA method. The reasons for this may including the significance of trading day effects in the time series as well as the different seasonal adjustment methods that the two models are using.

Furthermore, an advantage to the X-12-ARIMA method is that it is relatively easy to use even for people with limited statistical background, since this approach relies on a finite set of empirically developed moving averages. The user can either specify the particular moving averages used for the time series or simply let the program R choose them automatically according to the empirical criteria that presented in section 2.4.

In summary, X-12-ARIMA method is a reliable and efficient method for seasonal adjustment, trend estimation, outlier detection, and point forecasting, for the time series that has significant seasonal patterns and calendar effects, which could provide more accurate results than the pure seasonal ARIMA model.

5.2 Future Improvements

As the modelling by using X-12-ARIMA method can be compatible with user-defined regressors, one of the user-defined regressors that may affect monthly retail sales may be the monthly unemployment rates, since the unemployment rates are highly negatively correlated to the purchasing power as well as inflation rate, in an economic perspective. In addition, we can explore more user-defined regressors to train the model, in order to obtain more accurate forecasting results.

Furthermore, this analysis method may also help analyzing how the recent COVID-19 diseases outbreak impact on the retail industry in U.S., by using the data before and after the 2008 Financial Crisis as the base, and then adding other regressors as needed. If we can generate the appropriate forecasts by training this model by using X-12-ARIMA method, one may help the government and the private retailers to prepare for the changing in retail sales and the economy.

6 Bibliography

Aho, K., Derryberry & D., Peterson, T. (2014), "Model selection for ecologists: the worldviews of AIC and BIC", *Ecology*, **95** (3): 631–636, doi: 10.1890/13-1452.1

Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19 (6): 716–723, doi: 10.1109/TAC.1974.1100705

Chan, W. (1995). Understanding the effect of time series outliers on sample autocorrelations. Test 4, 179–186. URL: https://doi.org/10.1007/BF02563108

Findley, D., Monsell, B., Bell, W., Otto, M., & Chen, B. (1998). New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program. *Journal of Business & Economic Statistics, 16*(2), 127-152. doi:10.2307/1392565

Gray, A. & Thomson, P. (1996). “Design of Moving-Average Trend Filters Using Fidelity and Smoothness Criteria”, *Athens Conference in Applied Probability and Time Series, Vol.II:Time Series Analysis in Memory of E. J. Hannan*, New York: Springer-Verlag, pp.205-219

Hyndman, R.J. (2009) “Moving Averages”, *International Encyclopedia of Statistical Science*, 866-869, doi: 10.1007/978-3-642-04898-2_380

Hyndman, R.J., & Athanasopoulos, G. (2012), “6.1 Time Series Components”, *Forecasting: Principles and Practice*, OTexts: Melbourne, Australia. URL: OTexts.com/fpp

Kowarik, A., Meraner, A., Templ, M. & Schopfhauser, D. (2014). Seasonal Adjustment with the R Package x12 and x12GUI. *Journal of Statistical Software, 62(2)*. doi:10.18637/jss.v062.i02

Ma, W. (2016), “Time Series Analysis of Receipt of Fire Alarms Based on Seasonal Adjustment Method”, *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, doi: 10.1109/IHMSC.2016.208

National Statistics (2007). Guide to Seasonal Adjustment with X-12-ARIMA. *ONS Methodology and Statistical Development*. URL: https://ec.europa.eu/eurostat/cache/metadata/Annexes/lci_esqrs_uk_an1.pdf

Rossi, Richard J. (2018). Mathematical Statistics: An Introduction to Likelihood Based Inference. New York: John Wiley & Sons. p. 227. ISBN 978-1-118-77104-4

Shumway, R. H., & Stoffer, D. S. (2006). *Time series analysis and its applications: With R examples*. New York: Springer

U.S. Bureau of the Census (1999), X-12-ARIMA Reference Manual, U.S. Department of Commerce, Washington, DC

7 Appendices

7.1 Appendix A: AIC Test

The AIC, which is called Akaike Information Criterion, is an estimator that estimates the quality of each model, relative to each of the other models for a given set of data (Aho, Derryberry, Peterson, 2014). Thus, AIC is used for model selection. According to Akaike (1974), the AIC value of the model is defined as

$$\text{AIC} = -2\log(\hat{\mathcal{L}}) + 2k \tag{7.1.1}$$

where $\hat{\mathcal{L}}$ is the likelihood of the estimated model, k is the total number of parameters that are estimated in the model.

Note that the AIC uses a model's log-likelihood, $\log(\hat{\mathcal{L}})$ as a measure of fit. The likelihood function $\mathcal{L}$ measures the goodness of fit of a statistical model to the sample of data for given values of the unknown k parameters. According to Rossi (2018), the Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed model, the observed data can be explained the best. The natural log of the likelihood is used as a computational convenience. Hence, the model with the maximum likelihood $\arg\max\log(\hat{\mathcal{L}})$ is the one that fits the observed data the best. Also, the model with the higher value of $\log(\hat{\mathcal{L}})$ means it fits the observed data better.

Observed form equation (7.1.1), assume the number of parameters k is fixed, then as the value of $\log(\hat{\mathcal{L}})$ increases, the value of AIC decreases. Hence, the value of AIC is low for those models with high log-likelihoods, which also adds a penalty term for models, $2k$, with higher parameter complexity. Since the more parameters, the more likely the model overfits to the training data.

In conclusion, the smaller the value of AIC indicates the better model.

7.2 Appendix C: R Code

```
# Honours Project Code
# Created by Zhaoying Zheng

# ################ 0. Install packages ###################
install.packages('x12')
install.packages('dplyr')
install.packages('tidyr')
install.packages('readxl')
install.packages("ggplot2")
install.packages("Metrics")
install.packages('astsa')

library(x12)
library(dplyr)
library(tidyr)
library(readxl)
library(ggplot2)
library(Metrics)
library(astsa)
# ##############################



# ################# 1. Data Preprocessing #####################

# Load training dataset from Excel to RStudio
Train_data <- read_excel("C:/Users/zzhen/Desktop/Honours Project/Data/Monthly Retail Sales US -
Copy.xlsx", sheet="Training")
# Load testing dataset from Excel to RStudio
Test_data <- read_excel("C:/Users/zzhen/Desktop/Honours Project/Data/Monthly Retail Sales US -
Copy.xlsx", sheet="Testing")

# Convert tibble to dataframe
Train_df <- as.data.frame(Train_data)
Test_df <- as.data.frame(Test_data)
```

```
# Date conversion from POSIXct to Date format
Train_df$Date <- as.Date(Train_df$Date)
Test_df$Date <- as.Date(Test_df$Date)

# Transform dataframes to time series
Train.ts <- ts(Train_df$Sales, start=c(2004,1), end=c(2012,12), frequency=12)

# #############################



# ################## 2. Plotting Time Series ######################

# Plot the original time series of training dataset by using ggplot
ggplot(data = Train_df, aes(x = Date, y = Sales))+
  geom_line(color = "black", size = 0.7)

# #############################



# ################# 3. Model Selection by Using x12 ###################

# ------------ 3.1  check the significance of Trading Day effects ------------
# Automatic Seasonal ARIMA Model Detection
xb_trading <- new("x12Batch", list(Train.ts, Train.ts))
# Setting parameters: add trading day effect regressor
xb_trading <- setP(xb_trading,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
outlier.types="all"),1)
xb_trading <- setP(xb_trading,list(automdl=TRUE, estimate= TRUE,  regression.variables = c("td"),
outlier.types="all"),2)
xb_trading <- x12(xb_trading)
summary(xb_trading@x12List[[1]], fullSummary = TRUE)   # AIC = 1952.5514
summary(xb_trading@x12List[[2]], fullSummary = TRUE)   # AIC = 1864.3378



# ------------ 3.2 check the significance of Easter effects ------------
# Create new x12 batch object 'xb1' with 5 time series and change some Easter effect parameters
xb_easter <-new("x12Batch", list(Train.ts,Train.ts, Train.ts, Train.ts, Train.ts))
xb_easter <- setP(xb_easter,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
outlier.types="all"),1)
xb_easter <- setP(xb_easter,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
regression.variables = c("Easter[1]"), outlier.types="all"),2)
xb_easter <- setP(xb_easter,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
regression.variables = c("Easter[8]"), outlier.types="all"),3)
xb_easter <- setP(xb_easter,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
regression.variables = c("Easter[15]"), outlier.types="all"),4)
xb_easter <- setP(xb_easter,list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
regression.variables = c("td", "Unemploy_Rate"), outlier.types="all"),5)

xb_easter <- x12(xb_easter)
summary(xb_easter@x12List[[1]], fullSummary = TRUE)   # AIC = 1952.5514
summary(xb_easter@x12List[[2]], fullSummary = TRUE)   # AIC = 1954.036
summary(xb_easter@x12List[[3]], fullSummary = TRUE)   # AIC = 1954.1752
summary(xb_easter@x12List[[4]], fullSummary = TRUE)   # AIC = 1954.4974
summary(xb_easter@x12List[[5]], fullSummary = TRUE)   # AIC = 1866.3021



# ------------ 3.3 Check the AICs using different ARIMA parameters  ------------
# Create new x12 batch object with 4 time series and change some ARIMA parameters
xb_arima <-new("x12Batch", list(Train.ts, Train.ts, Train.ts, Train.ts))
xb_arima <- setP(xb_arima, list(automdl=FALSE))
xb_arima <- setP(xb_arima, list(arima.model=c(0,1,1),arima.smodel=c(0,1,1), estimate= TRUE,
regression.variables = c("td"), outlier.types="all"),1)    # Model 1 (automatic model)
xb_arima <- setP(xb_arima, list(arima.model=c(1,1,0),arima.smodel=c(1,1,0), estimate= TRUE,
regression.variables = c("td"), outlier.types="all"),2)    # Model 2
xb_arima <- setP(xb_arima, list(arima.model=c(0,1,1),arima.smodel=c(1,1,1), estimate= TRUE,
regression.variables = c("td"), outlier.types="all"),3)    # Model 3
xb_arima <- setP(xb_arima, list(arima.model=c(1,1,1),arima.smodel=c(1,1,1), estimate= TRUE,
regression.variables = c("td"), xreg = Unemploy_Rate.ts, regression.user = c("Unemploy_Rate.ts"),
outlier.types="all"),4)    # Model 4
xb_arima <- x12(xb_arima)
summary(xb_arima@x12List[[1]], fullSummary = TRUE)   # AIC = 1864.3378
summary(xb_arima@x12List[[2]], fullSummary = TRUE)   # AIC = 1902.1655
summary(xb_arima@x12List[[3]], fullSummary = TRUE)   # AIC = 1850.5602     # The lowest AIC
summary(xb_arima@x12List[[4]], fullSummary = TRUE)   # AIC = 1851.9084


# Model 3 in xb_arima has the lowest AIC 1850.5602 among the 4 different models
# Apply Model 3 to analyze the historical time series and forecast future values

model <- xb_arima@x12List[[3]]

# ###########################


# ######### 4. Plots Analysis of Seasonal Adjustment ##########

# ------------ 4.1 Manually Creating Plots ------------
# ------ For Trend, Seasonality, Irregular fluctuations and Seasonal Adjusted Series------

# Final Trend
trend <- model@x12Output@d12
trend_df <- data.frame(Date=as.Date(paste(1, zoo::as.yearmon(time(trend))),format = "%d %b %Y"),
Sales = as.matrix(trend))
trend_plot <- ggplot(data = trend_df, aes(x = Date, y = Sales))+
  geom_line(color = "#00AFBB", size = 0.7) + labs(y="Trend", x="Date")      # Plot the trend
trend_plot

# Final Seasonal Component
seasonal <- model@x12Output@d10
seasonal_df <- data.frame(Date=as.Date(paste(1, zoo::as.yearmon(time(seasonal))),format =
"%d %b %Y"), Sales = as.matrix(seasonal))
seasonal_plot <- ggplot(data = seasonal_df, aes(x = Date, y = Sales))+
  geom_line(color = "#00AFBB", size = 0.7) + labs(y="Seasonal Component", x="Date")      # Plot the
Seasonal Component
seasonal_plot

# Final Irregular flutuations
irregular <- model@x12Output@d13
```

```
irregular_df <- data.frame(Date=as.Date(paste(1, zoo::as.yearmon(time(irregular))),format =
"%d %b %Y"), Sales = as.matrix(irregular))
irregular_plot <- ggplot(data = irregular_df, aes(x = Date, y = Sales))+
  geom_line(color = "#00AFBB", size = 0.7) + labs(y="Irregular Fluctuations", x="Date")     # Plot the
Irregulars
irregular_plot

# Final Seasonally Adjusted Data
seasAujusted <- model@x12Output@d11
seasAujusted_df <- data.frame(Date=as.Date(paste(1, zoo::as.yearmon(time(seasAujusted))),format =
"%d %b %Y"), Sales = as.matrix(seasAujusted))
seasAujusted_plot <- ggplot(data = seasAujusted_df, aes(x = Date, y = Sales))+
  geom_line(color = "#00AFBB", size = 0.7) + labs(y="Seasonal Adjusted", x="Date")      # Plot the
Irregulars
seasAujusted_plot



# ------------ 4.2 Creating Built-in Plots ------------

# SI ratios and replacements plot
SeasFac_plot <- plotSeasFac(model)

# Plot everything including trend, seasonlly adjusted ts, outliers
All_plot <- plot(model,showAllout=TRUE,sa=TRUE,trend=TRUE, lwd_out=1,
pch_ao=4,pch_ls=7,pch_tc=9)

# Plot of autocorrelatoins of the residuals
acf_residual <- plotRsdAcf(model, which = "acf")

# #########################



################## 5. Point Forecasting for Future 12 Months #####################

# Forecasting values
forecast <- model@x12Output@forecast   # Estimate with CI
estimate <- model@x12Output@forecast@estimate
```

```
estimate_df <- data.frame(Date=as.Date(paste(1, zoo::as.yearmon(time(estimate))),format =
"%d %b %Y"), Predicted_Sales = as.matrix(estimate))

# Join the Observed values and predicted values in the same data frame
checkAccuracy_df <- left_join(Test_df, estimate_df, by = "Date", copy = FALSE)

# Check the mean absolute percentage error for the predicted values
mape(estimate_df$Predicted_Sales, Test_df$Observed_Sales)    # MAPE = 0.526%

# Plot the forecasting values and prediction intervals
forecast_plot <- plot(model,trend=TRUE, sa=TRUE,forecast = TRUE)

# ###########################



################## 6. Compare to Seasonal ARIMA Model #####################

# Fit a seasonal ARIMA model
# ACF of the original time series
acf(Train.ts,96)    # indicates seasonality, hence we should take the seasonal difference

# ACF and PACF of the first differenced and then seasonally differenced time series
acf2(diff(diff(Train.ts),12),48)
# Both ACF and PACF dies down
# Indicates an ARMA(1,1) model

# Build the seasonal ARIMA model
sarima(Train.ts,1,1,1,1,1,1,12)    # ARIMA(1,1,1)x(1,1,1)_12

# Predict values for future 12 months
estimate_sarima <- sarima.for(Train.ts,12,1,1,1,1,1,1,12)$pred
estimate_sarima_df <- data.frame(Date=as.Date(paste(1,
zoo::as.yearmon(time(estimate_sarima))),format = "%d %b %Y"), Predicted_Sales =
as.matrix(estimate_sarima))

# Join the Observed values and predicted values in the same data frame
```

```
checkAccuracy_sarima_df <- left_join(Test_df, estimate_sarima_df, by = "Date", copy = FALSE)

# Check the mean absolute percentage error for the predicted values
mape(estimate_sarima_df$Predicted_Sales, Test_df$Observed_Sales)     # MAPE = 0.979%

# ###########################
```

7.3 Appendix C: Partial R Output

7.3.1 Partial Summary of Model T1 and E1

```
> summary(xb_trading@x12List[[1]], fullSummary = TRUE)   # AIC = 1949.0331
--------------------------   Series_1   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(0,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 1
Automatically Identified Outliers: 1
Number of ts values that were almost identified as outliers: 0

        Regression Model
                variable   coef stderr   tval
1 autooutlier_LS2008.Nov -0.107  0.015 -6.978

        Likelihood Statistics
AIC:    1952.5514
AICC:   1952.9959
BIC:    1962.7669
HQ:     1956.6793
Log Likelihood:233.6246

Average absolute percentage error
        in out of sample forecasts
Last year:     2.1509
Last-1 year:   1.8427
Last-2 year:   2.1182
Last 3 years:  2.4918
```

7.3.2 Partial Summary of Model T2 and A1

```
> summary(xb_trading@x12List[[2]], fullSummary = TRUE)   # AIC = 1864.3378
--------------------------   Series_2   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0 1 1)(0 1 1) (Automatic Model Choice)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Trading Day + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 2
Automatically Identified Outliers: 2
Number of ts values that were almost identified as outliers: -

        Regression Model
                variable   coef stderr   tval
1                 td_Mon -0.002  0.002 -0.975
2                 td_Tue -0.002  0.002 -0.823
3                 td_Wed  0.003  0.002  1.656
4                 td_Thu  0.007  0.002  3.905
5                 td_Fri  0.003  0.002  1.567
6                 td_Sat  0.002  0.002  0.877
7                 td_Sun -0.011  0.002 -5.861
8 autooutlier_LS2008.Oct -0.160  0.023 -6.936
9 autooutlier_TC2008.Oct  0.101  0.022  4.532
* Derived parameter estimates:  TradingDay_Sun

        Likelihood Statistics
AIC:    1864.3378
AICC:   1867.5185
BIC:    1892.4305
HQ:     1875.6894
Log Likelihood:284.7314

Average absolute percentage error
        in out of sample forecasts
Last year:     1.2711
Last-1 year:   1.0927
Last-2 year:   1.4108
Last 3 years:  1.3097
```

7.3.3 Partial Summary of Model E2

```
> summary(xb_easter@x12List[[2]], fullSummary = TRUE)   # AIC = 1954.036
--------------------------   Series_2   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(0,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Easter[1] + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 1
Automatically Identified Outliers: 1
Number of ts values that were almost identified as outliers: 0

        Regression Model
                variable   coef stderr   tval
1 autooutlier_LS2008.Nov -0.107  0.015 -6.977
2    Easter[1]_Easter[1] -0.006  0.008 -0.719

        Likelihood Statistics
AIC:    1954.036
AICC:   1954.7101
BIC:    1966.8054
HQ:     1959.1958
Log Likelihood:233.8823

Average absolute percentage error
        in out of sample forecasts
Last year:     2.1461
Last-1 year:   1.8275
Last-2 year:   2.1131
Last 3 years:  2.4976
```

7.3.4 Partial Summary of Model E3

```
> summary(xb_easter@x12List[[3]], fullSummary = TRUE)   # AIC = 1954.1752
--------------------------   Series_3   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(0,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Easter[8] + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 1
Automatically Identified Outliers: 1
Number of ts values that were almost identified as outliers: 0

        Regression Model
                variable   coef stderr   tval
1 autooutlier_LS2008.Nov -0.107  0.015 -6.974
2    Easter[8]_Easter[8] -0.005  0.008 -0.614

        Likelihood Statistics
AIC:    1954.1752
AICC:   1954.8493
BIC:    1966.9445
HQ:     1959.3349
Log Likelihood: 233.8128

Average absolute percentage error
        in out of sample forecasts
Last year:     2.1473
Last-1 year:   1.8312
Last-2 year:   2.1139
Last 3 years:  2.4967
```

7.3.5 Partial Summary of Model E4

```
> summary(xb_easter@x12List[[4]], fullSummary = TRUE)   # AIC = 1954.4974
--------------------------   Series_4   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(0,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Easter[15] + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 1
Automatically Identified Outliers: 1
Number of ts values that were almost identified as outliers: 0

        Regression Model
                variable   coef stderr   tval
1 autooutlier_LS2008.Nov -0.107  0.015 -6.975
2  Easter[15]_Easter[15] -0.002  0.009 -0.233

        Likelihood Statistics
AIC:    1954.4974
AICC:   1955.1715
BIC:    1967.2667
HQ:     1959.6571
Log Likelihood:233.6516

Average absolute percentage error
        in out of sample forecasts
Last year:     2.1506
Last-1 year:   1.8447
Last-2 year:   2.1154
Last 3 years:  2.4919
```

7.3.6 Partial Summary of Model E5

```
> summary(xb_easter@x12List[[5]], fullSummary = TRUE)   # AIC = 1866.3021
--------------------------   Series_5   ------------------------------------
------------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(0,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Trading Day + Easter[1] + Automatically Identified Outliers


        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 2
Automatically Identified Outliers: 2
Number of ts values that were almost identified as outliers: 0

        Regression Model
                 variable   coef stderr   tval
1                  td_Mon -0.002  0.002 -0.977
2                  td_Tue -0.002  0.002 -0.837
3                  td_Wed  0.003  0.002  1.624
4                  td_Thu  0.007  0.002  3.793
5                  td_Fri  0.003  0.002  1.483
6                  td_Sat  0.002  0.002  0.868
7                  td_Sun -0.011  0.002 -5.708
8  autooutlier_LS2008.Oct -0.160  0.023 -6.926
9  autooutlier_TC2008.Oct  0.101  0.022  4.529
10    Easter[1]_Easter[1] -0.001  0.004 -0.190
* Derived parameter estimates:  TradingDay_Sun

        Likelihood Statistics
AIC:    1866.3021
AICC:   1870.107
BIC:    1896.9486
HQ:     1878.6856
Log Likelihood: 284.7493

Average absolute percentage error
        in out of sample forecasts
Last year:     1.2706
Last-1 year:   1.0924
Last-2 year:   1.4057
Last 3 years:  1.3136
```

7.3.7 Partial Summary of Model A2

```
> summary(xb_arima@x12List[[2]], fullSummary = TRUE)   # AIC = 1902.1655
--------------------------   Series_2   ------------------------------------
-----------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (1,1,0)(1,1,0)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Trading Day

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 0
Automatically Identified Outliers: 0
Number of ts values that were almost identified as outliers: 4

        Regression Model
                   variable   coef stderr   tval
1                    td_Mon -0.002  0.002 -1.043
2                    td_Tue  0.001  0.002  0.683
3                    td_Wed  0.000  0.002  0.200
4                    td_Thu  0.008  0.002  5.101
5                    td_Fri  0.001  0.002  0.840
6                    td_Sat  0.003  0.002  1.865
7                    td_Sun -0.012  0.002 -6.786
8  almostoutlier_AO2005.May -3.589 -2.298 -2.818
9  almostoutlier_LS2008.Oct -0.287 -3.709 -2.177
10 almostoutlier_AO2009.Aug  3.557  2.251  2.468
11 almostoutlier_TC2009.Sep -2.517 -3.150 -3.410
* Derived parameter estimates:  TradingDay_Sun

        Likelihood Statistics
AIC:    1902.1655
AICC:   1904.2832
BIC:    1925.1504
HQ:     1911.4531
Log Likelihood:263.8176

Average absolute percentage error
        in out of sample forecasts
Last year:     2.1996
Last-1 year:   0.9607
Last-2 year:   1.4724
Last 3 years:  4.1657
```

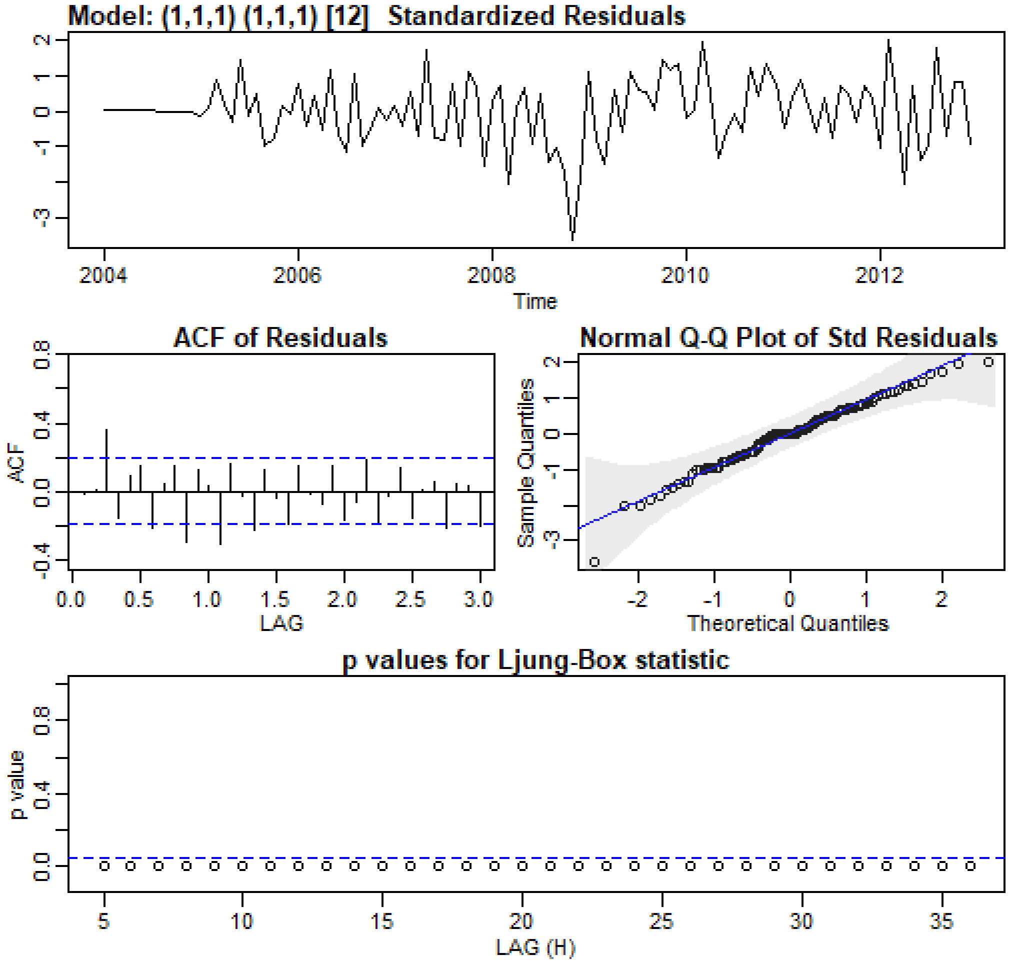
7.3.8 Partial Summary of Model A3

```
> summary(xb_arima@x12List[[3]], fullSummary = TRUE)   # AIC = 1850.5602
# The lowest AIC
--------------------------   Series_3   ------------------------------------
------------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (0,1,1)(1,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Trading Day + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 3
Automatically Identified Outliers: 3
Number of ts values that were almost identified as outliers: 0

        Regression Model
                 variable   coef stderr   tval
1                  td_Mon -0.003  0.002 -1.831
2                  td_Tue -0.001  0.002 -0.755
3                  td_Wed  0.003  0.002  1.967
4                  td_Thu  0.007  0.002  4.196
5                  td_Fri  0.002  0.002  1.542
6                  td_Sat  0.001  0.002  0.903
7                  td_Sun -0.009  0.002 -5.704
8  autooutlier_AO2008.Oct  0.050  0.010  5.069
9  autooutlier_LS2008.Oct -0.109  0.012 -9.405
10 autooutlier_TC2009.Mar -0.040  0.009 -4.505
* Derived parameter estimates:  TradingDay_Sun

        Likelihood Statistics
AIC:    1850.5602
AICC:   1855.054
BIC:    1883.7606
HQ:     1863.9757
Log Likelihood: 293.6202

Average absolute percentage error
        in out of sample forecasts
Last year:     1.3967
Last-1 year:   0.8808
Last-2 year:   1.6223
Last 3 years:  1.6869
```

7.3.9 Partial Summary of Model A4

```
> summary(xb_arima@x12List[[4]], fullSummary = TRUE)   # AIC = 1851.9084
--------------------------   Series_4   -------------------------------------
------------------------------------------------------------------------------
------

        Time Series

Frequency: 12
Span: 1st month,2004 to 12th month,2012

        Model Definition

ARIMA Model: (1,1,1)(1,1,1)
Model Span: 1st month,2004 to 12th month,2012
Transformation: Automatic selection : Log(y)
Regression Model: Trading Day + Automatically Identified Outliers

        Outlier Detection

Outlier Span: 1st month,2004 to 12th month,2012
Critical |t| for outliers:
aocrit1 aocrit2 lscrit1 lscrit2 tccrit1 tccrit2
"3.827"     "*" "3.827"     "*" "3.827"     "*"
Total Number of Outliers: 3
Automatically Identified Outliers: 3
Number of ts values that were almost identified as outliers: 0

        Regression Model
                 variable   coef stderr   tval
1                  td_Mon -0.003  0.002 -1.988
2                  td_Tue -0.001  0.002 -0.612
3                  td_Wed  0.003  0.002  1.863
4                  td_Thu  0.007  0.002  4.254
5                  td_Fri  0.002  0.002  1.497
6                  td_Sat  0.002  0.002  1.101
7                  td_Sun -0.009  0.002 -5.828
8  autooutlier_AO2008.Oct  0.052  0.010  5.290
9  autooutlier_LS2008.Oct -0.113  0.011 -9.996
10 autooutlier_TC2009.Mar -0.039  0.009 -4.347
* Derived parameter estimates:  TradingDay_Sun

        Likelihood Statistics
AIC:    1851.9084
AICC:   1857.1584
BIC:    1887.6627
HQ:     1866.3558
Log Likelihood:293.9461

Average absolute percentage error
        in out of sample forecasts
Last year:     1.5414
Last-1 year:   0.8438
Last-2 year:   2.0458
Last 3 years:  1.7345
```

7.3.10 Diagnostics for the Pure Seasonal ARIMA Model


Model: (1,1,1) (1,1,1) [12]   Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
LAG (H)