

# W271 Section 3 Lab 4

Kiersten Henderson, Zhaoning Yu, Daghan Altas

12/09/2017

```
knitr::opts_chunk$set(cache=TRUE)

library(easypackages)
packages("knitr","xts","forecast","ggfortify","ggplot2", "dplyr","plotly", "Hmisc",
         "tseries","stats","fpp", "forcats")

opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
rm(list=ls())
```

## I. Introduction

Our task was to analyze a time series entitled “Lab4-series2”. Because the time series is “unidentified” we did not know the domain of the data-generating process. Domain-specific knowledge often guides modeling choices.

Nonetheless, we were able to estimate a valid model (based on model diagnostics) with high forecast accuracy (measured by mean absolute percent error (MAPE) using out-of-sample fit) when compared to a naive forecast.

Of the many models we estimated, we favor an  $ARIMA(1, 1, 1)(0, 1, 1)_{12}$ .

$$x_t = \phi x_{t-1} + \theta w_{t-1} + \Theta w_{t-12} + w_t$$

and

$$\hat{x}_t = 0.93_{(0.038)} \hat{x}_{t-1} - 0.80_{(0.056)} \hat{w}_{t-1} - 0.89_{(0.052)} \hat{w}_{t-12} + \hat{w}_t$$

**expand this below** Interpret the model in words: The value for  $x$  at any given time is dependent on a combination of the error from the previous time point (one lag previous), the value of the previous time point, and the error in the measurement from one year ago.

## II. Data Loading and Cleaning

We began our analysis by loading the data, inspecting its structure, and checking for missing values.

```
# setwd('/Users/daghanaltas/Hacking/Berkeley/W271/Labs/w271_lab4')
df <- read.csv("./Lab4-series2.csv")
str(df)

## 'data.frame':    311 obs. of  2 variables:
## $ X: int  1 2 3 4 5 6 7 8 9 10 ...
## $ x: num  5.54 5.55 5.17 4.88 4.85 ...

cbind(head(df), tail(df))
```

```
##      X      x      X      x
## 1 1 5.544 306 5.240
## 2 2 5.555 307 5.546
## 3 3 5.172 308 5.078
## 4 4 4.878 309 4.907
## 5 5 4.851 310 4.599
## 6 6 4.686 311 4.681
```

```
sum(is.na(df)) # check if there is any NA
```

```
## [1] 0
```

There are no missing values and the first column of the dataframe is the index column, which we chose to discard. We proceeded by converting the data to a (xts) based time series due to the ease of subsetting xts-based objects.

```
ms <- ts(df$x, start = c(1990, 1), frequency = 12)
```

```
ms.training <- window(ms, 1990, c(2014, 12))
```

```
ms.test <- window(ms, 2015)
```

```
head(ms.training, 12)
```

```
##      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov
## 1990 5.544 5.555 5.172 4.878 4.851 4.686 5.127 4.968 5.215 5.249 5.504
##      Dec
## 1990 5.664
```

```
tail(ms.training, 12)
```

```
##      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov
## 2014 7.293 7.404 7.011 6.048 6.208 6.201 6.567 6.194 5.837 5.698 5.668
##      Dec
## 2014 5.498
```

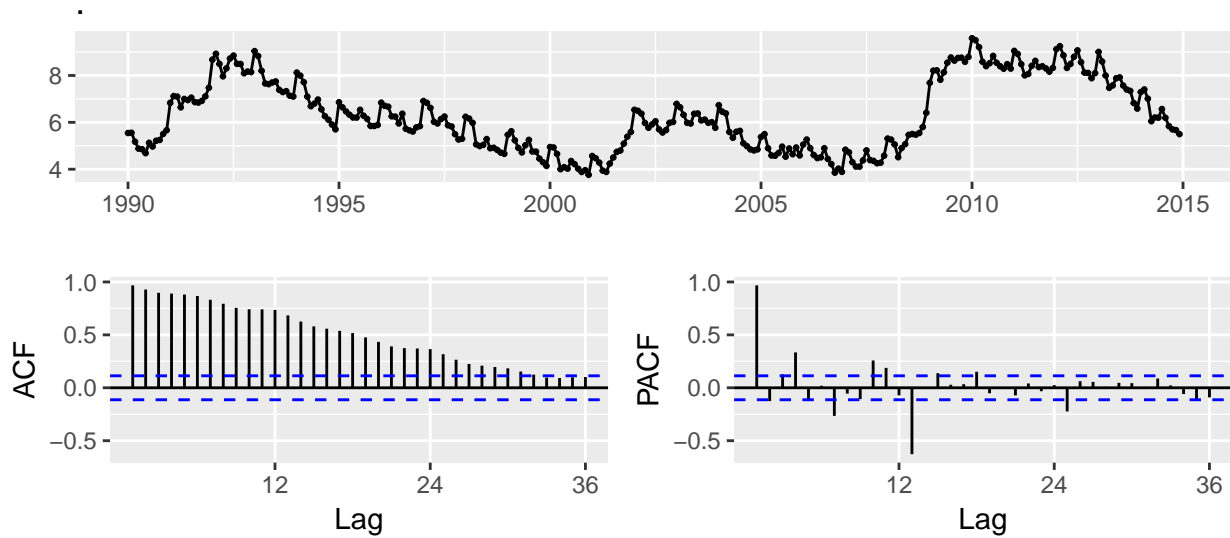
```
ms.test
```

```
##      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov
## 2015 6.449 6.425 5.929 5.536 5.472 5.240 5.546 5.078 4.907 4.599 4.681
```

### III. Exploratory Time Series Data Analysis

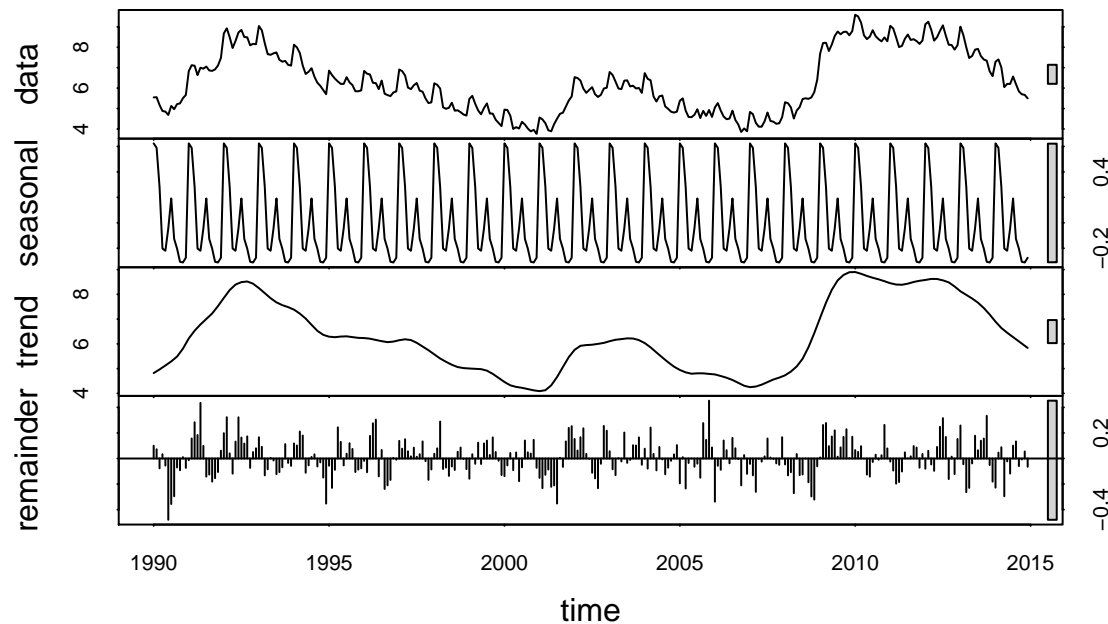
We initially plotted the untransformed time series with its ACF and PACF.

```
# there is an issue with X axis when plotting xts objects,
# converting to ts for plotting
ms.training %>% ggtsdisplay
```



We made use of STL decomposition to decompose the series into seasonal and trend components.

```
fit.stl <- stl(ms.training, t.window = 15, s.window = "periodic",
  robust = TRUE)
plot(fit.stl)
```



We observed that the untransformed series has both a trend and a seasonal component. Because the series is not stationary in the mean this immediately suggested the need for differencing.

The time series appears to have time-constant variance and our intuition is supported by the White test.

```
# white test
```

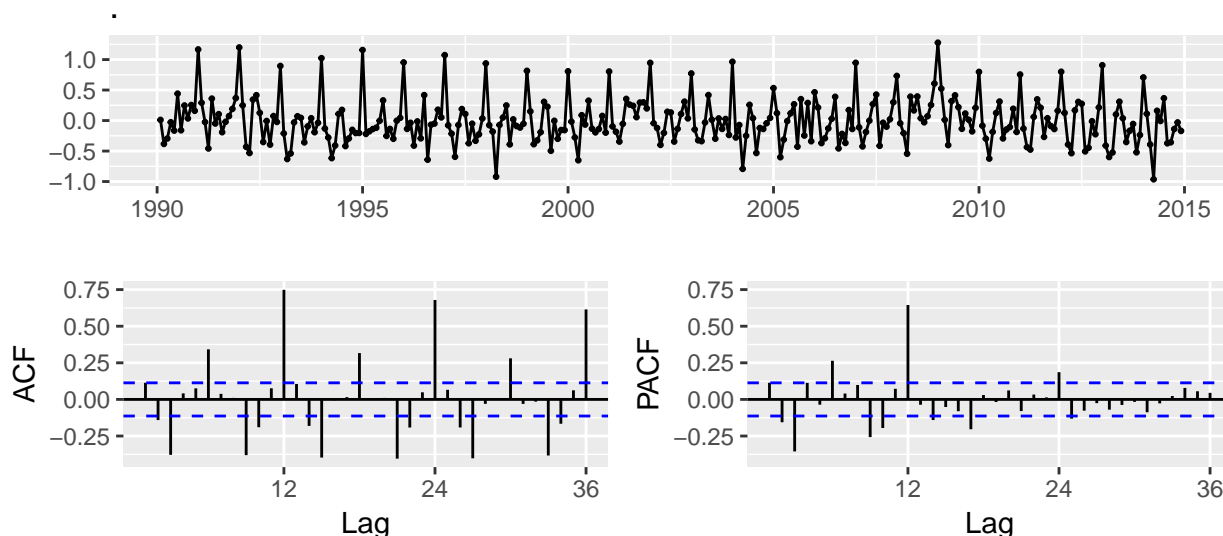
Thus, we do not see any need to stabilize the variance by transforming the series.

## Transformations

### Differencing for Trend

We identified a trend component in the untransformed time series, so we performed a first difference of the time series.

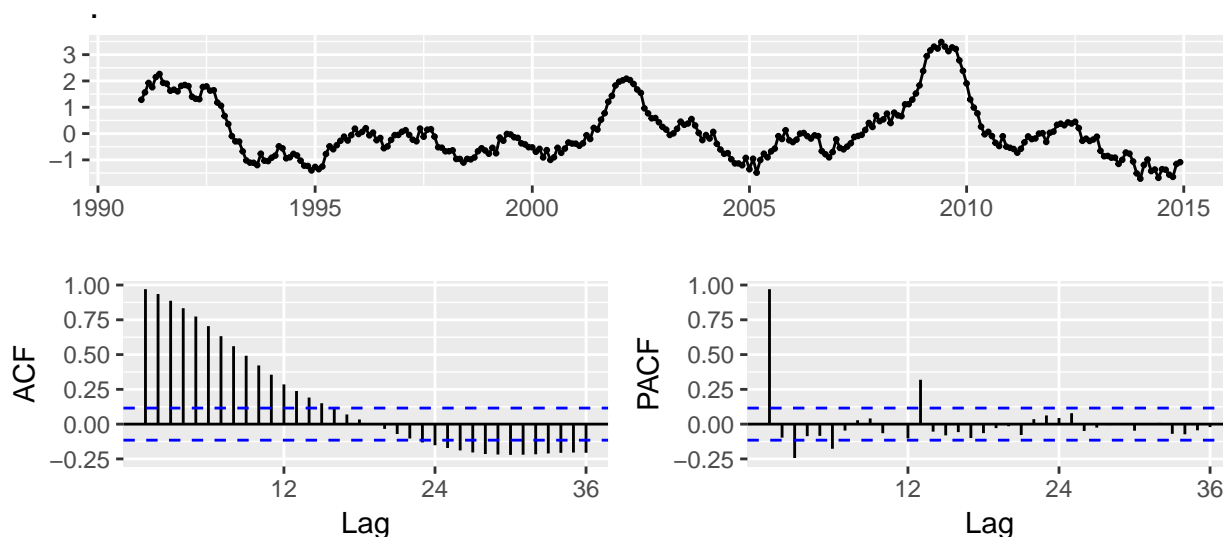
```
# First differencing only
ms.training.1d <- diff(ms.training, lag = 1)
ms.training.1d %>% ggtsdisplay
```



With only the first-differencing, we observed that the series appears somewhat stationary. However, when we examined the time series plot as well as the PACF chart, we observed a strong yearly seasonal component (at lag 12). This yearly seasonal component warrants attention and so we proceeded to explore the seasonality of our time series.

\*\* add adf test and interpretation

```
# Seasonal differencing only
ms.training.12d <- diff(ms.training, lag = 12)
ms.training.12d %>% ggtsdisplay
```



## Differencing for Yearly Seasonality

We observed that the seasonal differencing significantly smoothed the time series. Furthermore, the effect of seasonal differencing is also apparent in the ACF and PACF graphs.

*\*\*in what way?*

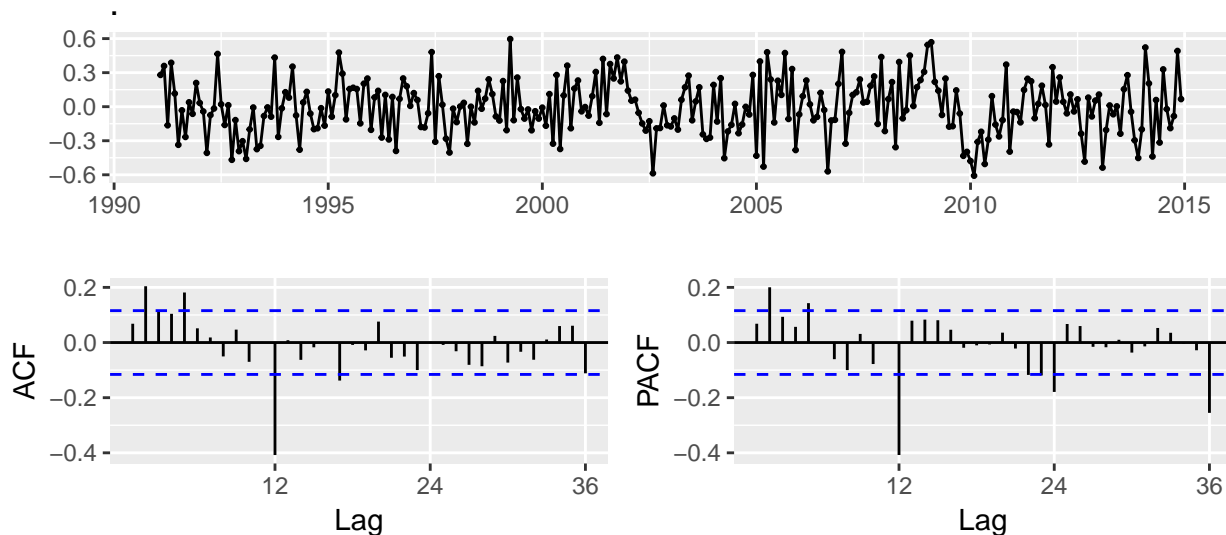
However, there is still obvious trend in the series and it is therefore, not stationary.

*\*\*adf and discussion*

We will combine the seasonal and non-seasonal components for our next exploratory graph

## Differencing for Trend and Yearly Seasonality

```
# Both the trend and seasonal differencing
ms.training.1d.12d <- diff(diff(ms.training, lag = 1), lag = 12)
ms.training.1d.12d %>% ggtsdisplay
```



```
pp.test.results <- cbind(round(rbind(pp.test(ms.training.1d)$statistic,
  pp.test(ms.training.12d)$statistic, pp.test(ms.training.1d.12d)$statistic),
  2), rbind("Can reject H0", "Can not reject H0", "Can reject H0"))
row.names(pp.test.results) <- c("1D", "12D", "1D.12D")
pp.test.results
```

```
##          Dickey-Fuller Z(alpha)
## 1D      "-199.64"              "Can reject H0"
## 12D     "-12.55"              "Can not reject H0"
## 1D.12D  "-327.85"            "Can reject H0"
```

We found that the combination of first-difference for trend plus differencing for lag-12 seasonality produces a time series that appears much more stationary than did differencing for either component alone. The augmented Dickey-Fuller test supported our intuition. Using the test, we were able to reject the null hypothesis that a unit root existed for the series ( $p = 0.01$ ) and thus proceed with the alternate hypothesis that the differenced series is weakly stationary.

We were thus able to examine and interpret the ACF and PACF plots to identify potential *auto-regressive* and *moving-average* components for model estimation.

There are statistically significant ACF/PACF components at lag 2 and 5 without a clear pattern pointing in either MA or AR direction. ACF graph suggests an MA(2) model, whereas PACF graph suggests an AR(2) model. Both graphs hint at the possibility of a lag(5) component. In addition, when examining the PACF, we observe a strong component at lag 12. This suggests a seasonal  $MA(1)_{12}$  component.

**expand discussion to ACF**

## Summary of Exploratory Time Series Data Analysis

- summarize rationale for differencing decision
- Our exploratory analysis of the time series suggests an  $ARIMA(p, 1, q)(0, 1, 1)_{12}$  model
- Based on our interpretation of the ACF and PACF plots, our exploration to identify non-seasonal AR/MA order ( $p$  and  $q$ ) considered order up to 5.
- However, to avoid overfitting, the goal of our model estimation process was to identify appropriate models where the order of  $p \in (1, 2)$  and  $q \in (1, 2)$ .

## IV. Identifying the Dependence Orders for Estimated Models

Based on the EDA, we will search  $p/q$  values from 1 to 5, that minimize various **Information Criteria**. We will make our final choice based on the mean absolute percentage error (**MAPE**) criterion. During our search, we will hold  $d=1$ ,  $D=1$ ,  $P=1$ ,  $Q=1$  constant (based on our EDA).

```
results <- data.frame(p = 1:25, q = 1:25, AIC = 0, AICc = 0,
  BIC = 0)
for (p in 1:5) {
  for (q in 1:5) {
    m <- ms.training %>% Arima(order = c(p, 1, q), seasonal = list(order = c(0,
      1, 1), period = 12))
    index <- (p - 1) * 5 + q
    results[index, ] = c(p, q, m$aic, m$aicc, m$bic)
  }
}
rbind(results[which.min(results$AIC), ], results[which.min(results$AICc),
  ], results[which.min(results$BIC), ])
```

```
##    p q      AIC      AICc      BIC
## 6  2 1 -125.8566 -125.6430 -107.5591
## 61 2 1 -125.8566 -125.6430 -107.5591
## 1  1 1 -122.7133 -122.5715 -108.0754
```

## Candidate Models

We found that the grid search corroborated our exploratory analysis. We have looked for an optimal  $p, q$  combination within the  $p \in 1, 5$  and  $q \in 1, 5$ . Based on the information criteria optimization, we decided to focus on the following models:

- $ARIMA(1,1,1)(0,1,1)_{12}$  because it minimized the BIC.
- $ARIMA(2,1,1)(0,1,1)_{12}$  because it minimized the AIC and AICc.

## Model Parameter Estimation

```
fit111 <- ms.training %>% Arima(order = c(1, 1, 1), seasonal = list(order = c(0,
  1, 1), period = 12))
summary(fit111)
```

```
## Series: .
## ARIMA(1,1,1)(0,1,1)[12]
##
## Coefficients:
##          ar1          ma1          sma1
##          0.9311      -0.8047      -0.8909
## s.e.    0.0388      0.0560      0.0520
##
## sigma^2 estimated as 0.03519:  log likelihood=65.36
## AIC=-122.71  AICc=-122.57  BIC=-108.08
##
## Training set error measures:
##              ME          RMSE          MAE          MPE          MAPE
## Training set -0.009339336  0.1825311  0.1427566 -0.1059365  2.389875
##              MASE          ACF1
## Training set  0.1763078 -0.1050292
```

```
fit211 <- ms.training %>% Arima(order = c(2, 1, 1), seasonal = list(order = c(0,
  1, 1), period = 12))
summary(fit211)
```

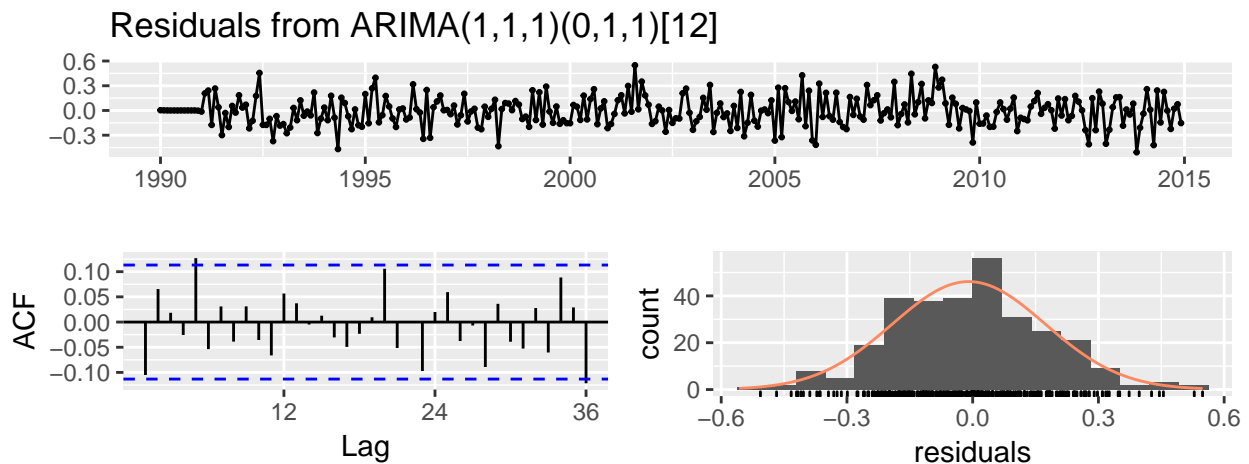
```
## Series: .
## ARIMA(2,1,1)(0,1,1)[12]
##
## Coefficients:
##          ar1          ar2          ma1          sma1
##          0.7289      0.1592      -0.7036      -0.8825
## s.e.    0.1030      0.0680      0.0900      0.0512
##
## sigma^2 estimated as 0.03476:  log likelihood=67.93
## AIC=-125.86  AICc=-125.64  BIC=-107.56
##
## Training set error measures:
##              ME          RMSE          MAE          MPE          MAPE
## Training set -0.008926836  0.1810804  0.1416233 -0.1017597  2.364688
##              MASE          ACF1
## Training set  0.1749081 -0.004085812
```

The model parameters we estimated for both our candidates are all statistically significant (criterion =  $2 \times \text{SE} < \text{Parameter}$ ).

## VI. Model Validation with Model Diagnostics

We performed and examined the model diagnostics for the two models we favored. `## ARIMA(1,1,1)(0,1,1)[12]`  
`**we need a hist of residuals here. instead of ggtsdisplay(res111), we could do #checkresiduals(fit111) - this also performs the Ljung-box test a`

```
res111 <- residuals(fit111)
# ggtsdisplay(res111)
checkresiduals(fit111)
```



```
##
## Ljung-Box test
##
## data: Residuals from ARIMA(1,1,1)(0,1,1)[12]
## Q* = 24.11, df = 21, p-value = 0.2878
##
## Model df: 3. Total lags used: 24
```

**\*\* ADD Verbiage saying we remove 211 for brevity\*\***

The interpretation of the model diagnostics for both of our candidate models is similar. There are very few significant lags for the model residuals in ACF plots, and the Box-Ljung tests cannot reject the null hypothesis that the residuals have remaining autocorrelations ( $p=0.29$ , and  $p=0.52$ , respectively). Thus, we conclude that the model residuals sufficiently resemble white noise. In addition, the residuals from both models follow a normal distribution.

These diagnostic tests provide support for accepting the model assumptions required for valid time series modeling and forecasting. We can therefore proceed to forecasting with confidence.

## VII. Forecasting

### Standard for Evaluating our Model

We wanted to have some kind of “standard” for model performance, in order to evaluate the performance of our model accuracy. Naive forecast is typically used for as a standard against which forecasts are compared (ref?). One hopes that thier model forecast is more accurate than simply using what has either just happened or happened at the same time period last year to forecast. In our case, we wanted to know if our favored model would outperform simply using the same values for last year as the current year’s forecast.

```
# insert naive seasonal forecast calculation of MAPE from
# Kiersten V1.
```

We proceeded by performing an 11-month ahead forecast of the series in 2015 using both models.

```
forecast111 <- fit111 %>% forecast(h = 11)
forecast211 <- fit211 %>% forecast(h = 11)
```



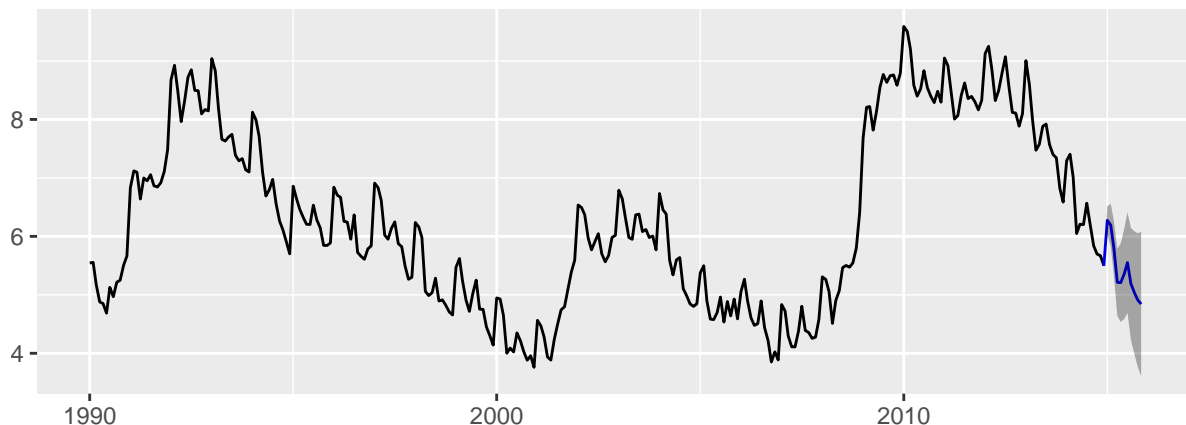
```
(results <- rbind(accuracy(forecast111, ms.test), accuracy(forecast211,
ms.test))[, 1:5])
```

##		ME	RMSE	MAE	MPE	MAPE
## Training set		-0.009339336	0.1825311	0.1427566	-0.1059365	2.389875
## Test set		0.026623361	0.2002727	0.1779535	0.1710571	3.319050
## Training set		-0.008926836	0.1810804	0.1416233	-0.1017597	2.364688
## Test set		-0.012986097	0.2163455	0.1986888	-0.6007220	3.770132

We compared our forecast fit to our reserved test data using the MAPE score. We note that the  $ARIMA(1,1,1)(0,1,1)_{12}$  has not only a lower score than the  $ARIMA(2,1,1)(0,1,1)_{12}$ , (3.2 versus 3.8), but also the lowest out of sample MAPE score for any of the models we estimated. We therefore performed forecasting with the  $ARIMA(1,1,1)(0,1,1)_{12}$  model.

**\*expand** In addition, our favored model forecast substantially outperformed the seasonal naive forecast in terms of MAPE. We are therefore confident that we have developed a valid and accurate forecast, which can be seen below.

```
forecast111 %>% autoplot()
```



The 11 month-ahead seasonal ARIMA model forecast follows the recent trend in the data due to the combination of seasonal and trend differencing. The confidence intervals are quite small for the first third of the forecast, however the confidence intervals rapidly increase during the latter two-thirds of the forecast. The point forecasts initially increase and then trend downward. The prediction interval also initially increases but allows for the data to either trend upwards or downwards during the latter two-thirds of the forecast period. We conclude that, as expected, our model is better at predicting the near-future.

Based on our forecast we expect the value of the dependent variable to continue to show seasonality over the next year and to mimic the downward trend present in the most recently observed values. We must however, be mindful of the possibility that the series will experience one of its unpredictable sudden upward shifts during the forecast period. If an upward shift occurred, it would lead to considerably different and higher point values than our forecast predicts.

**?Can you overlay the actual values on the forecast plot???**

## VIII. Conclusions

Based on our EDA we proceeded to...

We identified several appropriate models. Our criteria for choosing our favored model was... -good model diagnostics - the model residuals are statistically indistinguishable from white noise and are normally

distributed -good in sample fit as measured by low values of AIC, AICc, BIC -low out of sample forecast as measured by out of sample MAPE forecast

Our favored model is: ARIMA()

Interpret the model in words: The value for  $x$  at any given time is dependent on a combination of the error from the previous time point (one lag previous), the value of the previous time point, and the error in the measurement from one year ago.

Briefly summarize forecast predictions again...

Why we are happy with the model : the lowest out of sample accuracy that we identified for all the valid models we estimated. Only 3.2% error when compared to actual values realized in the time series. Compare this to xx% for the naive seasonal forecast, seems like an excellent forecast.