Midterm Report

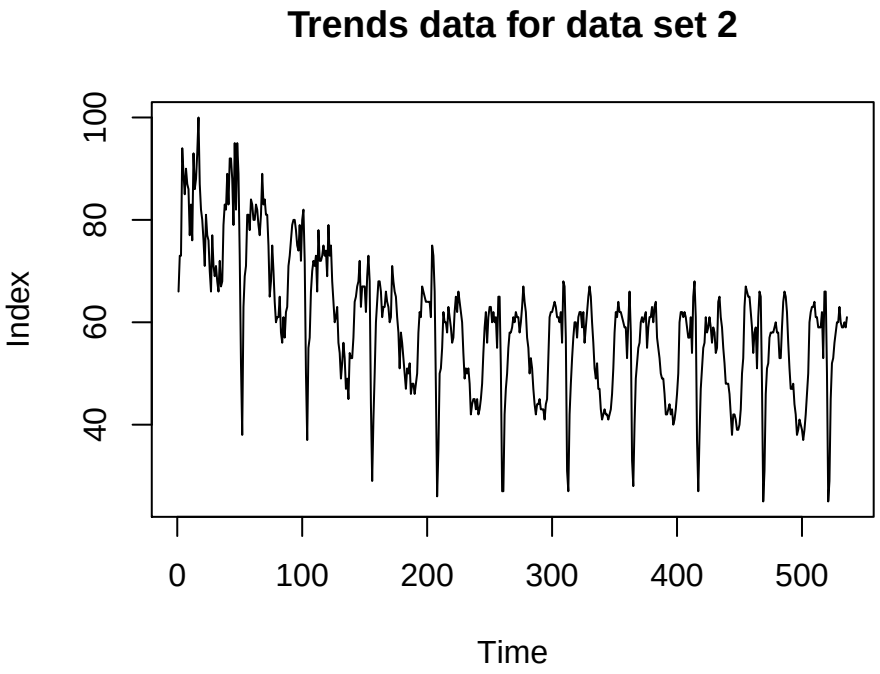
Lianghui Li(25180845)

April 17, 2016

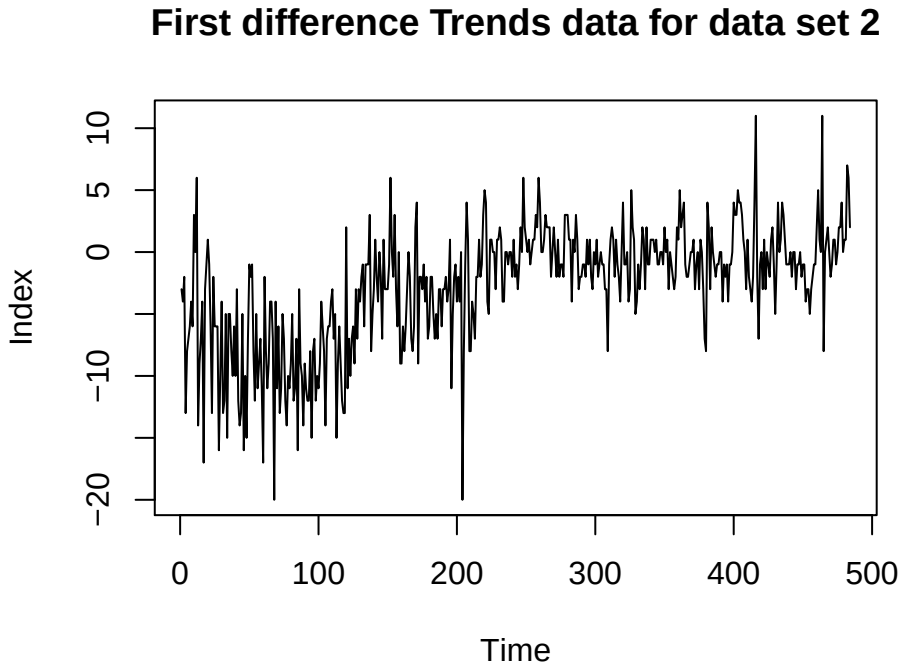
This report is on the analysis of dataset 2.

ARIMA Approach

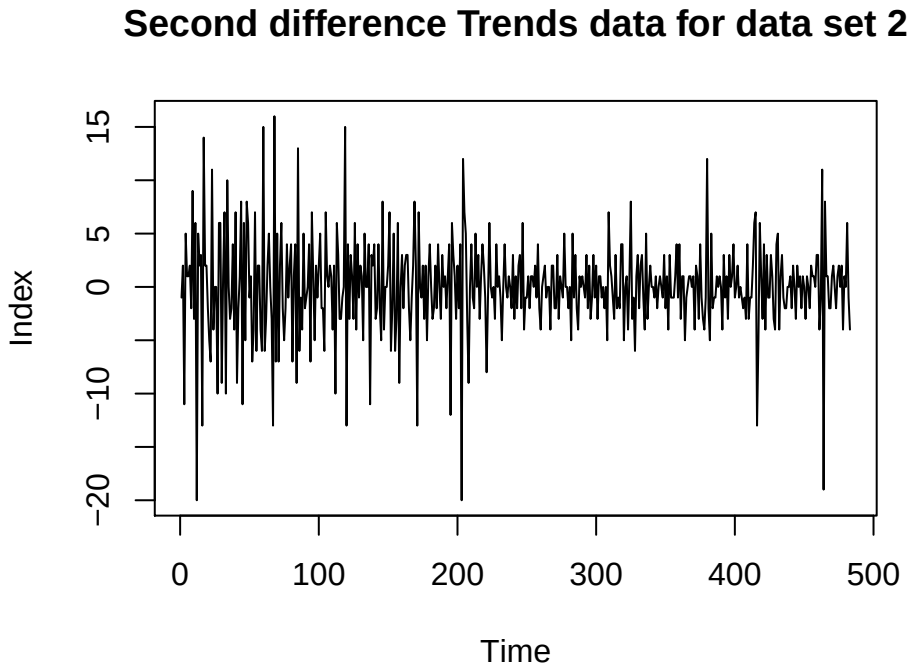
First, from looking at the original graph, we can see constant variance across different time periods, so I decide not to take a log of the trend or any other transformation to stabilize the variance.



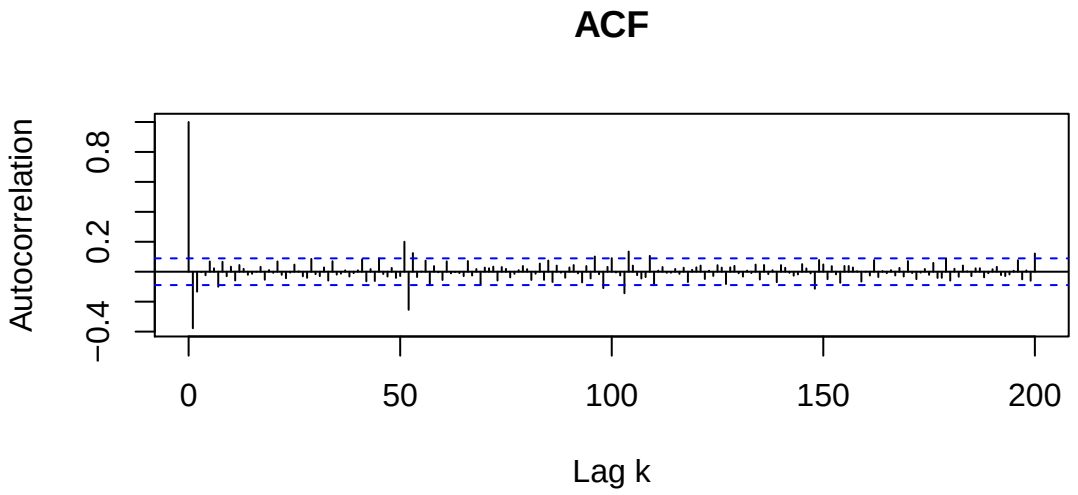
Second, I take a difference in order 52, since the data is the weekly data, and the plot has some obvious drop at each 52 periods, so I suspect the data has a seasonality factor of order 52.

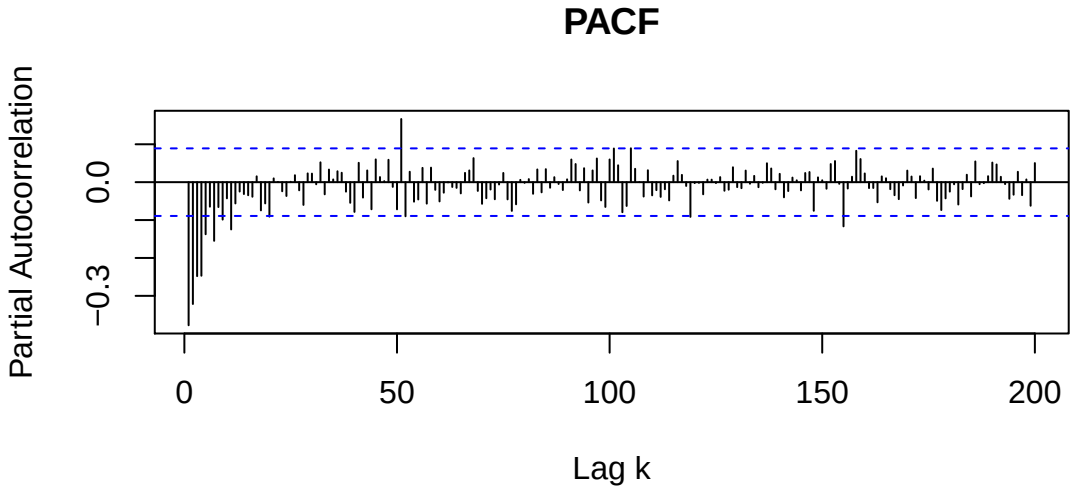


After dealing with seasonality, it's clear that there is still trend involved, since our analysis is about stationary process, we should take care of trend first by taking the difference in order 1.



Now, we can begin our analysis by looking at the `ACF` and `PACF` from data after differencing.





First thing to notice is that there are some significant bunchings on the first two lag, around lag 52 and around lag 104 in our acf graph. However, from pacf, we are not as conclusive because there are decreasing trend of pacf in about first 10 lags that across blue bands as well as lag 52, which implies we need to fit more than 10 parameters in the AR order.

Being afraid of overfitting, I try to use the model under order 4 in each category, and basically concluded the performance from the acf graph, and I decide to fit `SARIMA` model instead of `ARIMA`, because using latter will involve using 52 parameters, which will easily cause a problem of overfitting that lower prediction power, since our data only has 536 periods.

Therefore, I start with a model with MA(2) (because the first two lags are over the blue bands in acf) multiply by MA(1) with seasonal 52(because the bunchings around lag 52), then overfit this model each time, and I conclude 6 models that seems `Reasonable` to me.

Model

$$ARMA(0,2) * (0,1)_{52}$$
$$ARMA(0,3) * (0,1)_{52}$$
$$ARMA(0,2) * (0,2)_{52}$$
$$ARMA(0,2) * (1,2)_{52}$$
$$ARMA(0,2) * (2,2)_{52}$$
$$ARMA(0,3) * (1,2)_{52}$$

Model Selection

AIC and BIC

From these 6 models, we compare their `AIC` and `BIC` separately

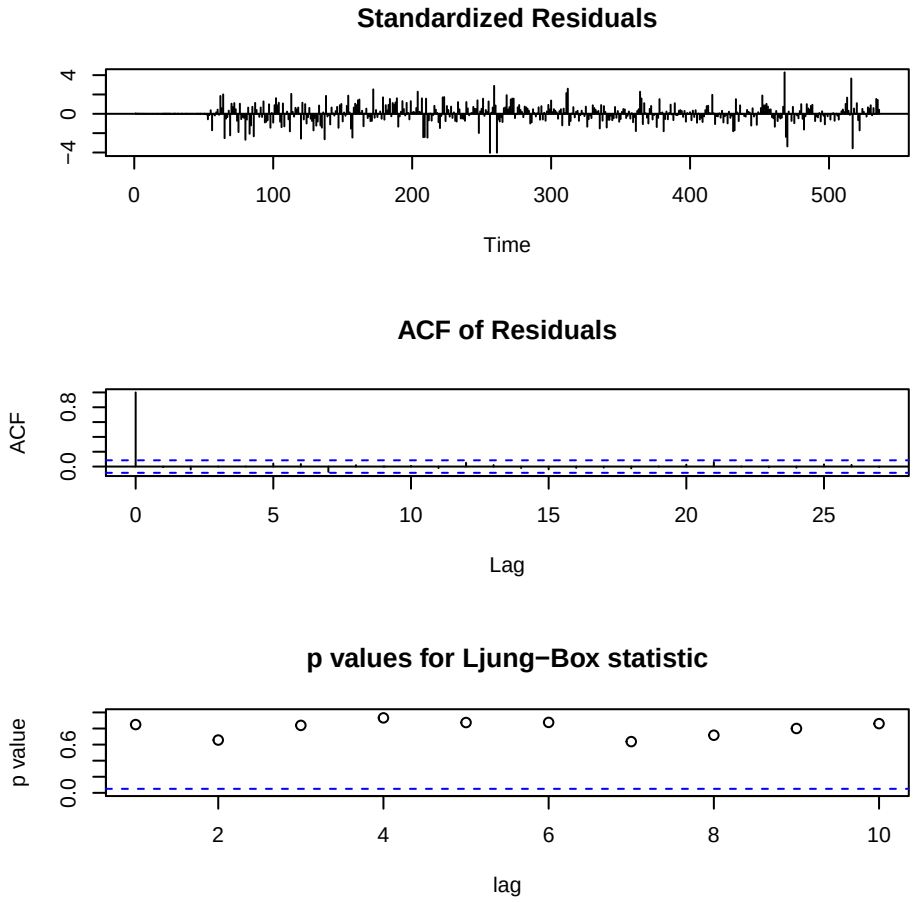
	1st	2nd	3rd	4th	5th	6th
AIC	2549	2552	2535	2527	2530	2530
BIC	2566	2572	2556	2552	2559	2559

AIC and **BIC** provides us a mean to assess the overall quality for the model, not only they will reward the goodness of fit from model(by likelihood function), but also penalize the increasing number of estimators, since increase the number of estimators always improve the goodness of fit. Overall, **AIC** and **BIC** are dealing with the information loss compared with the `True` model, so the lower those value are, the better model is (closer to the `True` model).

Therefore, it seems that based on the information given by **AIC** and **BIC**, we should select model 4 in a comparative sense(the less the better); since **AIC** and **BIC** can not provide absolute quality of our models, it's possible that none of them fit the data well, and only the 4th fit slightly `less worse`; so it's better to have model diagnostic for all of our models.

Model diagnostic

In this case, because of the length restriction, only provide the plots for `4th` model.



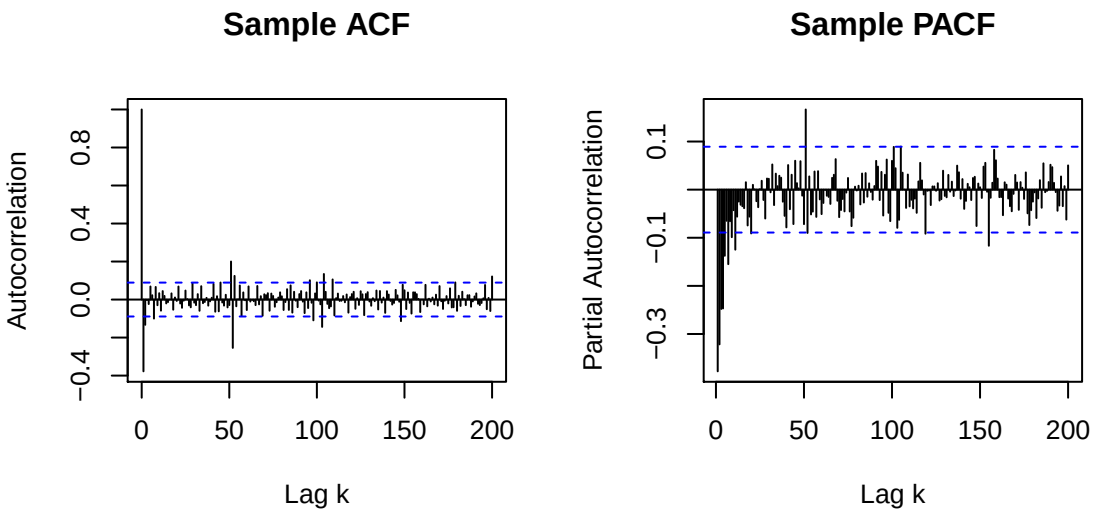
It's easy to see that `4th` model works really decent in this case, with no acf across the blue bands and the p-value Ljung-Box is quite high, meaning that the null hypothesis of the data coming from this model fails to reject, actually, this diagnostic works well with the other five models in this case.

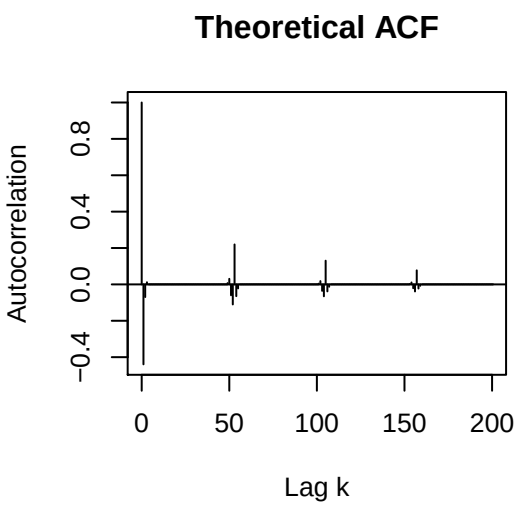
Theoretical and Sample Comparison

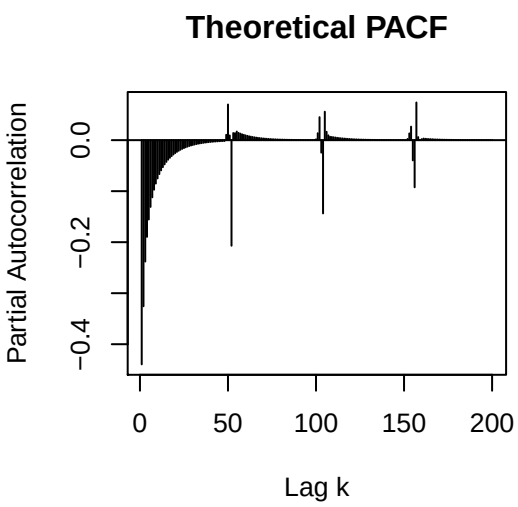
First, we need to get the estimate from our `4th` model

```
##        ma1        ma2       sar1       sma1       sma2
## -0.7076936 -0.1811358  0.5934393 -0.8592202  0.3515217
```

Then we plug the estimate into `ARMAacf` to see the theoretical plot for ACF and PACF







From the comparison, we could see similar significant pattern (bunchings in lag 52, 104, 156)in both ACF and PACF, however, even though, taken into the account of noise from the sample, there are still some unignorable discrepancies in the detail of bunching, such as the sign of bunching in PACF, warning us that the data may not be fully concluded in the `4th` model or some other existed `ARIMA` is better in this case.

Therefore, it is not precise to draw the conclusion for our model selection, since we use all of the data to fit; if the model is good enough, it will still be better than other models in a sense that fewer residuals from different models.

It is necessary to use a rigorous procedure **Cross Validation** that determines how much closer to the real data that `4th` model perform in each segment of the data(test group), in which I define as [121, 224], [225, 328], [329, 432], [433,536], because our final prediction is the length of 104, I try to mimic this procedure by setting the length of each time frames equal 104. Comparing the real value and predictive value by calculating the MSE:

$$MSE = \sum (y - \hat{y})^2$$

MSE

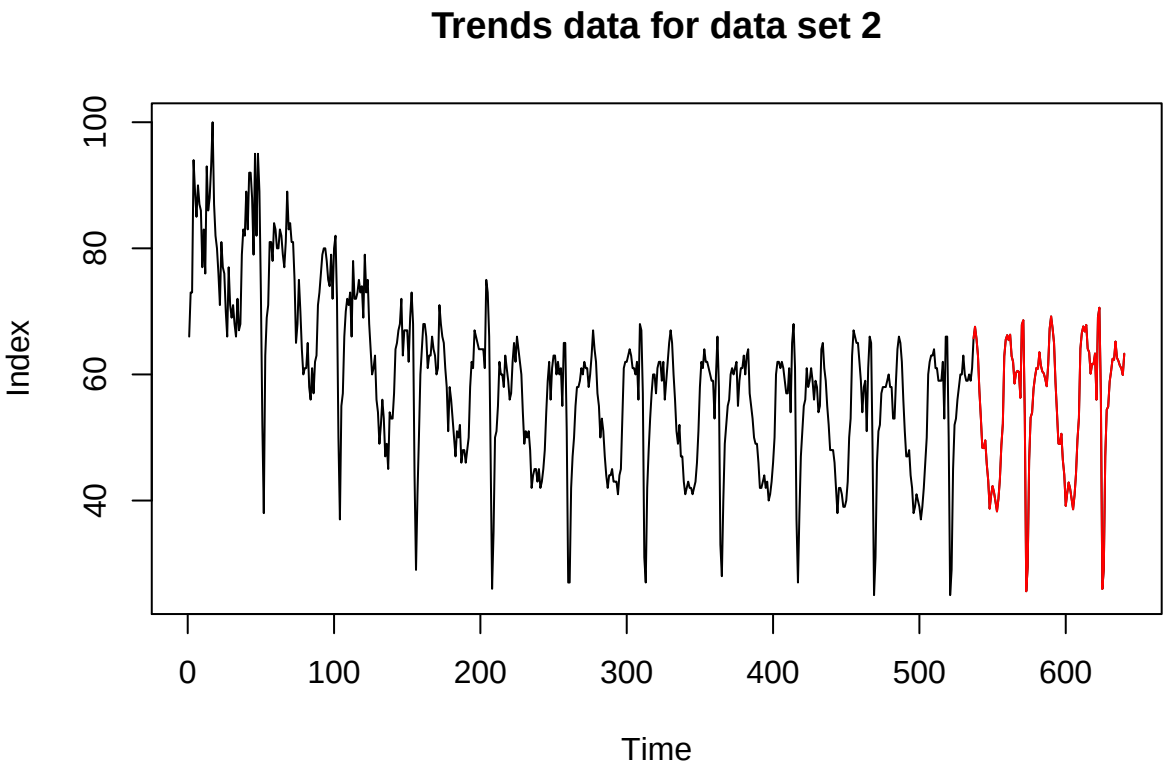
Model/Time	[121, 224]	[225, 328]	[329, 432]	[433,536]
1st	18.46	5.88	35.5	13.96
2nd	18.68	5.86	35.51	13.96
3rd	17.75	6.89	35.81	14.04
4th	16.92	7.63	34.86	14.1
5th	17	8.6	35.4	14.12
6th	16.98	7.37	34.71	14.1

We would expect the better usually has lower `MSE`, since its prediction is closer to the `Real` value. Unlike previous cases, using **Cross Validation** leads us to a variety of `good` model, especially when we focus on the least MSE yields in each time frames, such as `1st`, `3rd`, `4th`. since in this data set, we don't have significant outliers that cause concern so as to weight differently in each time frame, but in this case, we should treat them equally and take these three models into consideration when forecasting

Finally, I calculate my original data prediction based on the weighted average of prediction yields by these three models; since `4th` also performs better, I decide to weight it more, and lastly my prediction comes from these formula:

$$Orignial = \frac{3}{5} * 4th + \frac{1}{5} * 1st + \frac{1}{5} * 3rd$$

After getting the prediction, we can look at the prediction along with previous data to see if there are some others abnormal variation exists.



All in all, it seems to be a good prediction graphically(prediction in red), and we finish the model fitting and forecast using **ARIMA**.

Parametric Approach

As a matter of fact, I use another approaches as well, which is parametric fitting by linear regression. From the original data, we could see there could be a quadratic trend, after fitting trend, we could analyze the residual by using similar step before(without differencing), in addition, since the data has seasonality, we could fit different types of sinusoid function that we have learned before.

- Fit the original data by quadratic function(LM) or sinusoid
- After fitting, deriving its residual
- Annlyze residual by ARIMA function
- Predict future residual by selected ARIMA function
- Predict future trend using quadratic function or sinusoid function
- Derive Oringal prediction by the sum of the prediction from trend and residual

Actually, it may seem more efficient by the regression method and these steps should lead us to the similar result, however, after using **Cross Validation**, the MSE in this case are much higher than the our previous model before, so I decide to discard this result because of its limitation.

Appendix

```
# loading data
data_all <- list(numeric(), numeric(), numeric(), numeric(), numeric())
for (i in 1:5) {
    ds_raw <- read.delim(file = paste0("q", i, "train.csv"))
    ds_raw <- ds_raw[1:536, 1]
    len <- length(ds_raw)
    ds <- rep(0, len)
    for (j in 1:len) {
        ds[j] <- as.numeric(unlist(strsplit(as.character(ds_raw[j]), ","))[2])
    }
    data_all[[i]] <- ds
}

ds1 <- data_all[[1]]
ds2 <- data_all[[2]]
ds3 <- data_all[[3]]
ds4 <- data_all[[4]]
ds5 <- data_all[[5]]

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

# dataset 1
plot(ds1, type = "l")
# take 52nd difference because of seasonlity
ds1.d <- diff(ds1, 52)
plot(ds1.d, type = "l")
# there is small linear trend
ds1.dd <- diff(ds1.d)
plot(ds1.dd, type = "l")
par(mfrow = c(2, 1))
acf(ds1.dd, lag.max = 100)
pacf(ds1.dd, lag.max = 100)

m11 <- arima(ds1, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m11)
# as we can see the value at lag 10 is approximatly 0.05, which is
# significant
m12 <- arima(ds1, order = c(0, 1, 4), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m12)

m13 <- arima(ds1, order = c(1, 1, 4), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m13)

m14 <- arima(ds1, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m14)
# the best so far in terms of low AIC
m15 <- arima(ds1, order = c(2, 1, 5), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m15)

len <- length(ds1)
computeCVmse <- function(order.totry, seasorder.totry, dataset) {
```

```
    MSE <- numeric()
    for (k in 4:1) {
        train.dt <- dataset[1:(len - 104 * k)]
        test.dt <- dataset[(len - 104 * k + 1):(len - 104 * (k - 1))]
        mod <- arima(train.dt, order = order.totry, seasonal = list(order = seasorder.totry,
            period = 52))
        fcast <- predict(mod, n.ahead = 104)
        MSE[k] <- mean((fcast$pred - test.dt)^2)
    }
    return(MSE)
}
MSE1 <- computeCVmse(c(0, 1, 3), c(0, 1, 1), ds1)
MSE2 <- computeCVmse(c(0, 1, 4), c(0, 1, 1), ds1)
MSE3 <- computeCVmse(c(1, 1, 4), c(0, 1, 1), ds1)
MSE4 <- computeCVmse(c(2, 1, 1), c(0, 1, 1), ds1)
MSE5 <- computeCVmse(c(2, 1, 5), c(0, 1, 1), ds1)

# afraid of overfitting, according to MSE and BIC, we use weighted average
# of 2rd and 3rd arima function(using the weight 2/5, 3/5 since model 3 is
# more accurate under MSE)
pred11 <- predict(m12, n.ahead = 104)$pred
pred12 <- predict(m13, n.ahead = 104)$pred
orig_pred1 <- 1/2 * pred11 + 1/2 * pred12
plot(1:(length(ds1) + length(orig_pred1)), c(ds1, orig_pred1), type = "l")
points((length(ds1) + 1):(length(ds1) + length(orig_pred1)), orig_pred1, col = "red",
    type = "l")

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

# dataset 2
t <- 1:536
plot(ds2, type = "l")
lg2 <- ds2
plot(lg2, type = "l")
ds2.d <- diff(lg2, 52)
plot(ds2.d, type = "l")
ds2.dd <- diff(ds2.d)
plot(ds2.dd, type = "l")
par(mfrow = c(2, 1))
acf(ds2.dd, lag.max = 200)
pacf(ds2.dd, lag.max = 200)

m21 <- arima(lg2, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m21)
m22 <- arima(lg2, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m22)
m23 <- arima(lg2, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 2), period = 52))
tsdiag(m23)
# best so far in terms of AIC
m24 <- arima(lg2, order = c(0, 1, 2), seasonal = list(order = c(1, 1, 2), period = 52))
tsdiag(m24)
m25 <- arima(lg2, order = c(0, 1, 2), seasonal = list(order = c(2, 1, 2), period = 52))
tsdiag(m25)
```

```r
m26 <- arima(lg2, order = c(0, 1, 3), seasonal = list(order = c(1, 1, 2), period = 52))
tsdiag(m26)

MSE21 <- computeCVmse(c(0, 1, 2), c(0, 1, 1))
MSE22 <- computeCVmse(c(0, 1, 3), c(0, 1, 1))
MSE23 <- computeCVmse(c(0, 1, 2), c(0, 1, 2))
MSE24 <- computeCVmse(c(0, 1, 2), c(1, 1, 2))
MSE25 <- computeCVmse(c(0, 1, 2), c(2, 1, 2))
MSE26 <- computeCVmse(c(0, 1, 3), c(1, 1, 2))

pred24 <- predict(m24, n.ahead = 104)$pred
pred21 <- predict(m21, n.ahead = 104)$pred
pred23 <- predict(m23, n.ahead = 104)$pred
orig_pred2 <- 3/5 * pred24 + 1/5 * pred21 + 1/5 * pred23
plot(1:(length(ds2) + length(orig_pred2)), c(ds2, orig_pred2), type = "l")
points((length(ds2) + 1):(length(ds2) + length(orig_pred2)), orig_pred2, col = "red",
    type = "l")

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

# dataset 3
plot(log(ds3), type = "l")
ds3.d <- diff(ds3, 52)
ds3.dd <- diff(ds3.d)
plot(ds3.d, type = "l")
par(mfrow = c(2, 1))
acf(ds3.dd, lag.max = 200)
pacf(ds3.dd, lag.max = 200)
m31 <- arima(ds3, order = c(1, 1, 2), seasonal = list(order = c(1, 1, 0), period = 52))
tsdiag(m31)
# the simplist?
m32 <- arima(ds3, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 0), period = 52))
tsdiag(m32)
# best in terms AIC
m33 <- arima(ds3, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 52))
tsdiag(m33)
m34 <- arima(ds3, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m34)
# best box (similar acf)
m35 <- arima(ds3, order = c(1, 1, 3), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m35)
m36 <- arima(ds3, order = c(2, 1, 2), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m36)
m37 <- arima(ds3, order = c(2, 1, 3), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m37)

m38 <- arima(ds3, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m38)
m39 <- arima(ds3, order = c(0, 1, 3), seasonal = list(order = c(1, 1, 1), period = 52))
tsdiag(m39)
m310 <- arima(ds3, order = c(2, 1, 3), seasonal = list(order = c(1, 1, 1), period = 52))
tsdiag(m310)
```

```r
# can't find
MSE31 <- computeCVmse(c(1, 1, 2), c(1, 1, 0))
MSE32 <- computeCVmse(c(1, 1, 1), c(1, 1, 0))
# can't find
MSE33 <- computeCVmse(c(1, 1, 1), c(1, 1, 1))
MSE34 <- computeCVmse(c(1, 1, 1), c(0, 1, 1))
MSE35 <- computeCVmse(c(1, 1, 3), c(0, 1, 1))
MSE36 <- computeCVmse(c(2, 1, 2), c(0, 1, 1))
MSE37 <- computeCVmse(c(2, 1, 3), c(0, 1, 1))
MSE38 <- computeCVmse(c(0, 1, 3), c(0, 1, 1))
MSE39 <- computeCVmse(c(0, 1, 3), c(1, 1, 1))
MSE310 <- computeCVmse(c(2, 1, 3), c(1, 1, 1))

pred32 <- predict(m32, n.ahead = 104)$pred
pred35 <- predict(m35, n.ahead = 104)$pred
pred37 <- predict(m37, n.ahead = 104)$pred
pred38 <- predict(m38, n.ahead = 104)$pred
pred39 <- predict(m39, n.ahead = 104)$pred
orig_pred3 <- 1/9 * pred32 + 2/9 * pred35 + 2/9 * pred37 + 3/9 * pred38 + 1/9 *
    pred39
plot(1:(length(ds3) + length(orig_pred3)), c(ds3, orig_pred3), type = "l")
points((length(ds3) + 1):(length(ds3) + length(orig_pred3)), orig_pred3, col = "red",
    type = "l")

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

# dataset 4
lg4 <- log(ds4)
plot(ds4, type = "l")
plot(diff(ds4, 52), type = "l")
lg4.dd <- diff(diff(lg4, 52))
plot(ds4.dd, type = "l")
plot(diff(lg4.dd))
par(mfrow = c(2, 1))
acf(lg4.dd, lag.max = 200)
pacf(lg4.dd, lag.max = 200)
lg41 <- arima(lg4, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(lg41)
lg42 <- arima(lg4, order = c(3, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(lg42)
lg43 <- arima(lg4, order = c(3, 1, 1), seasonal = list(order = c(1, 1, 0), period = 52))
tsdiag(lg43)
lg44 <- arima(lg4, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 2), period = 52))
tsdiag(lg44)
lg45 <- arima(lg4, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 2), period = 52))
tsdiag(lg45)
lg46 <- arima(lg4, order = c(3, 1, 1), seasonal = list(order = c(2, 1, 2), period = 52))
tsdiag(lg46)
lg47 <- arima(lg4, order = c(3, 1, 1), seasonal = list(order = c(1, 1, 2), period = 52))
tsdiag(lg49)
lg48 <- arima(lg4, order = c(4, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
lg49 <- arima(lg4, order = c(4, 1, 1), seasonal = list(order = c(1, 1, 1), period = 52))
lg410 <- arima(lg4, order = c(4, 1, 2), seasonal = list(order = c(0, 1, 1),
```

```
    period = 52))
lg411 <- arima(lg4, order = c(3, 1, 1), seasonal = list(order = c(0, 1, 0),
    period = 52))
tsdiag(lg411)

MSEl42 <- computeCVmse(c(3, 1, 1), c(0, 1, 1))
MSEl41 <- computeCVmse(c(2, 1, 1), c(0, 1, 1))
# none
MSEl43 <- computeCVmse(c(3, 1, 1), c(1, 1, 0))
MSEl44 <- computeCVmse(c(1, 1, 1), c(0, 1, 2))
MSEl45 <- computeCVmse(c(2, 1, 1), c(0, 1, 2))
# none
MSEl46 <- computeCVmse(c(3, 1, 1), c(2, 1, 2))
# none
MSEl47 <- computeCVmse(c(3, 1, 1), c(1, 1, 1))
MSEl48 <- computeCVmse(c(4, 1, 1), c(0, 1, 1))
MSEl49 <- computeCVmse(c(4, 1, 1), c(1, 1, 1))
MSEl410 <- computeCVmse(c(4, 1, 2), c(0, 1, 1))
MSEl411 <- computeCVmse(c(3, 1, 1), c(0, 1, 0))
# possibily 511 011

pred42 <- predict(lg42, n.ahead = 104)$pred
pred41 <- predict(lg41, n.ahead = 104)$pred
pred44 <- predict(lg44, n.ahead = 104)$pred
pred48 <- predict(lg48, n.ahead = 104)$pred
pred410 <- predict(lg410, n.ahead = 104)$pred
orig_pred4 <- 1/5 * exp(pred42) + 1/5 * exp(pred48) + 3/5 * exp(pred410)
plot(1:(length(ds4) + length(orig_pred4)), c(ds4, orig_pred4), type = "l")
points((length(ds4) + 1):(length(ds4) + length(orig_pred4)), orig_pred4, col = "red",
    type = "l")

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

# dataset 5
plot(log(ds5), type = "o")
lg5 <- log(ds5)
plot(diff(diff(lg5, 52)), type = "o")
lg5.dd <- diff(diff(lg5, 52))
ds5.dd <- diff(diff(ds5, 52))
acf(lg5.dd, lag.max = 200)
pacf(lg5.dd, lag.max = 200)
acf(ds5.dd, lag.max = 200)
pacf(ds5.dd, lag.max = 200)
m51 <- arima(lg5, order = c(3, 1, 1), seasonal = list(order = c(1, 1, 1), period = 52))
tsdiag(m51)
m52 <- arima(lg5, order = c(3, 1, 1), seasonal = list(order = c(1, 1, 0), period = 52))
tsdiag(m52)
m53 <- arima(lg5, order = c(3, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m53)
m54 <- arima(lg5, order = c(4, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m54)
m56 <- arima(lg5, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m56)
```

```r
m57 <- arima(lg5, order = c(1, 1, 1), seasonal = list(order = c(0, 1, 1), period = 52))
tsdiag(m57)

MSE51 <- computeCVmse(c(3, 1, 1), c(1, 1, 1))
MSE52 <- computeCVmse(c(3, 1, 1), c(1, 1, 0))
MSE53 <- computeCVmse(c(3, 1, 1), c(0, 1, 1))
MSE54 <- computeCVmse(c(4, 1, 1), c(0, 1, 1))
MSE56 <- computeCVmse(c(2, 1, 1), c(0, 1, 1))
MSE57 <- computeCVmse(c(1, 1, 1), c(0, 1, 1))

pred53 <- predict(m53, n.ahead = 104)$pred
pred56 <- predict(m56, n.ahead = 104)$pred
orig_pred5 <- 1/2 * exp(pred53) + 1/2 * exp(pred56)
plot(1:(length(ds5) + length(orig_pred5)), c(ds5, orig_pred5), type = "l", col = 1)
points((length(ds5) + 1):(length(ds5) + length(orig_pred5)), orig_pred5, type = "l",
    col = 2)
```