Predicting Chickenpox Cases in Hungary

Oluwatomi Hassan

2023-11-01

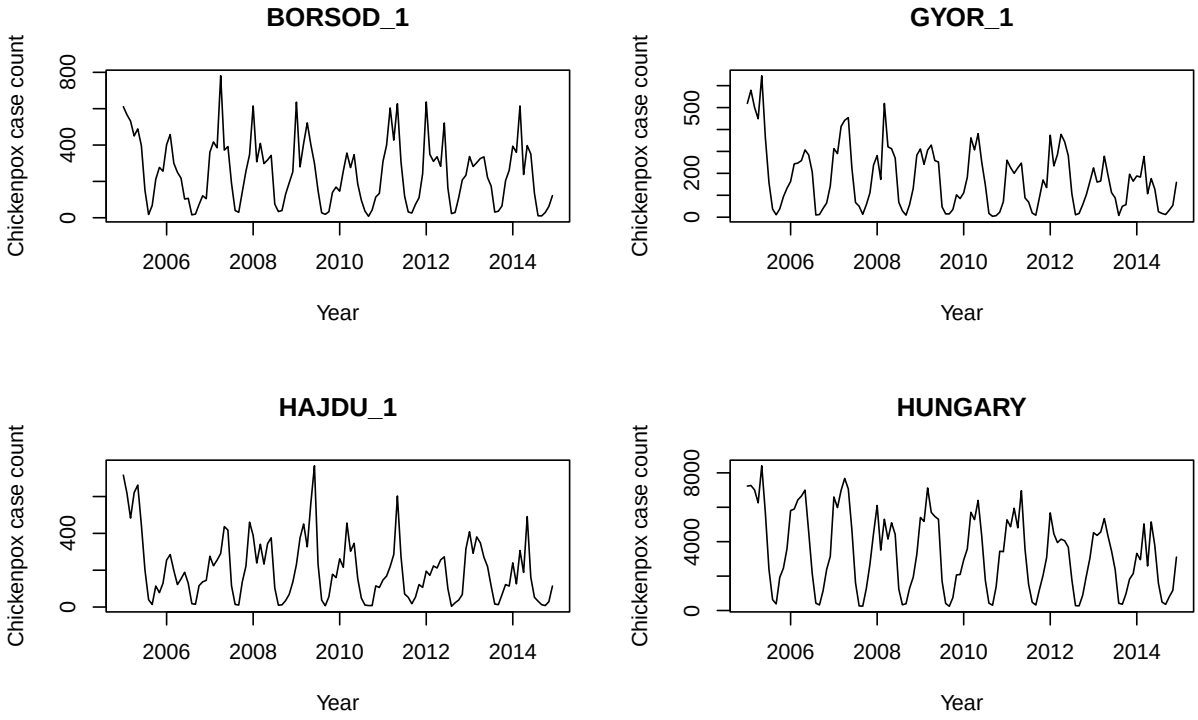
Introduction

This report contains a comprehensive time series analysis of chickenpox cases in Hungary between 2005 and 2015. The data set we will be using contains reported chickenpox cases for twenty geographic regions of Hungary (19 counties and the capital city of Budapest), and our analysis aims to achieve two primary objectives: county level case count prediction and nation level case count prediction. The first objective will be achieved by developing time series models that can effectively predict case levels for three different counties in Hungary (Győr-Moson-Sopron County, Hajdú-Bihar County, and Borsod-Abaúj-Zemplén County) and the second objective will be achieved by developing a time series model that can effectively predict case levels for the nation of Hungary as a whole. The models are trained using the first seven years of the data and tested using the last two years (2013,2014). The resulting models in this study can aid to predict highly contagious disease cases similar to chickenpox on a local or regional scale to improve public health response.

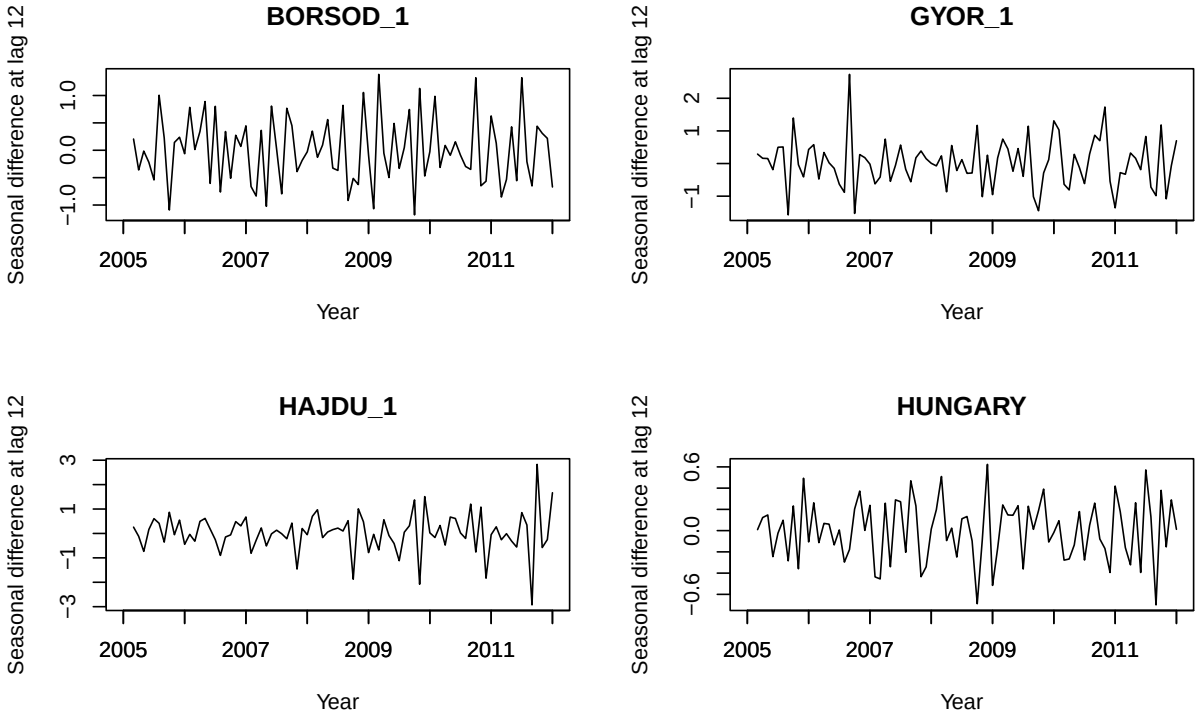
Methods

To explore the data, time plots for each of the three counties of interest and Hungry is plotted. Since the variation in the data decreases over time, a log transformation is performed to stabilize the variance. After the transformation, there is evidence of seasonality and trends in the plot. The first difference and seasonal difference at lag 12 was used to remove additional seasonality and trend. Once the plot was stationary, the data is modeled with Holt-Winter Smoothing method and ARIMA models. The ACF/PACF plots is used to determine the parameters of the ARIMA model. Each ARIMA model is evaluated with AIC and RMSE is used to compare the ARIMA and Holt-Winters Smoothing method(HW) models.

Original Time Series

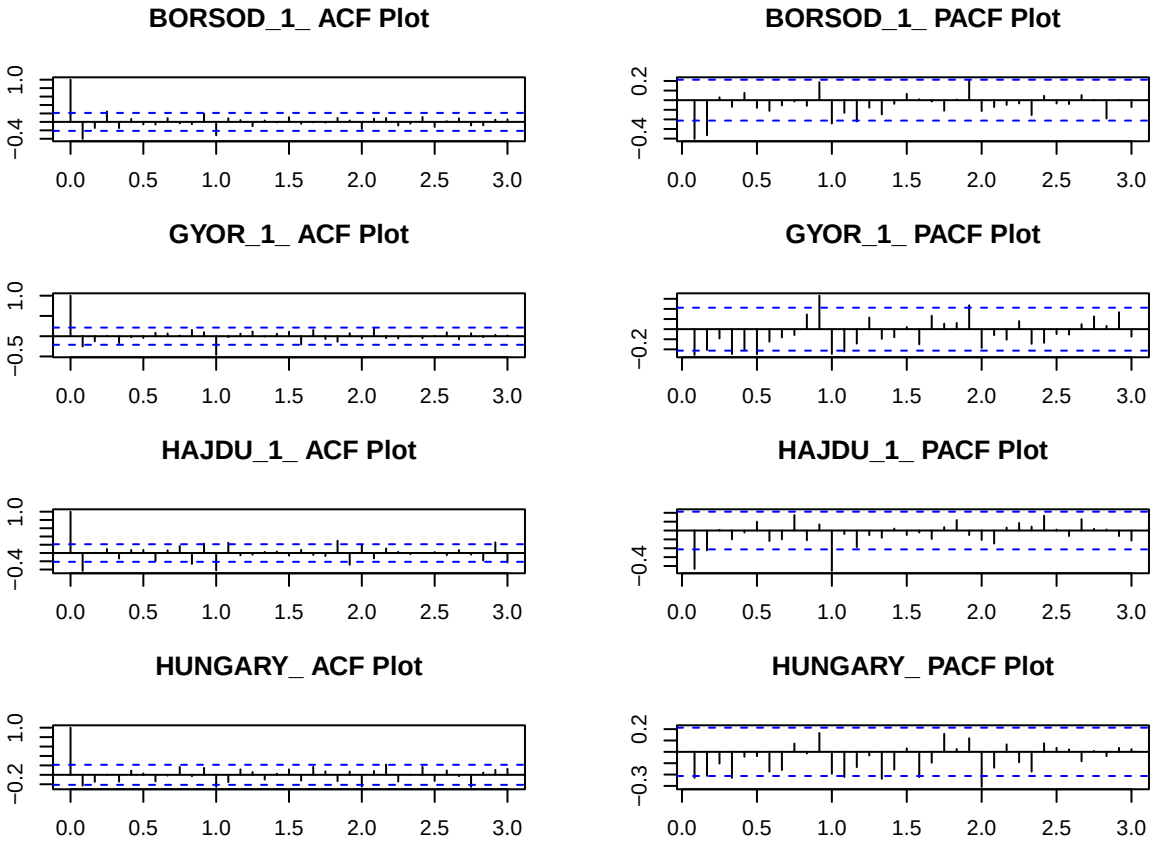


1st and Seasonal Difference of Log-Transformed Time Series.



Results

ARIMA Models



Borsod-Abaúj-Zemplén County For the seasonal component, ACF is non-zero at lag 1, suggesting $Q = 1$ and PACF is non-zero at lag 2, suggesting $P = 2$. For the non-seasonal component, the ACF seems non-zero at lag 1, suggesting $q = 1$ AND PACF is non-zero at lag 1 or 2, suggesting $p = 1$ or $p = 2$.

The ARIMA model ARIMA(1,0,1) x ARIMA(2,1,1) 12 produced the lowest AIC.

Győr-Moson-Sopron County For the seasonal component, the ACF appears to be non-zero at seasonal lag 1, suggesting $Q = 1$ and the PACF appears to be slightly non-zero at seasonal lag 1 suggesting $P = 1$ or $P = 0$.

For the non-seasonal component, the ACF appears to cut off after lag 1 suggesting $q = 1$ or $q = 0$ and the PACF suggests p is somewhere between 1 and 6.

The ARIMA model ARIMA(1,1,1) x ARIMA(0,1,1) 12 produced the lowest AIC.

Hajdú-Bihar County For the seasonal component, the ACF appears to be non-zero at seasonal lag 1, suggesting $Q = 1$ and the PACF appears to be non-zero at seasonal lag 1 suggesting $P = 1$.

For the non-seasonal component, the ACF appears to cut off after lag 1 suggesting $q = 1$ and the PACF appears to be non-zero at lag 1 suggesting $p = 1$.

The ARIMA model ARIMA(1,0,1) x ARIMA(0,1,1) 12 produced the lowest AIC.

Nation of Hungary For the seasonal part, the ACF appears to be non-zero at seasonal lag 1 and seasonal lag 2 suggesting $Q = 1$ or $Q = 2$ and the PACF appears to be zero at seasonal lag 1 but non zero at seasonal lag 2 suggesting $P = 0$ or $P = 2$.

For the non-seasonal part, the ACF appears to cut off after lag 1 suggesting $q = 1$ and the PACF is non-zero at lag 1 suggesting $p = 1$ or $p = 0$.

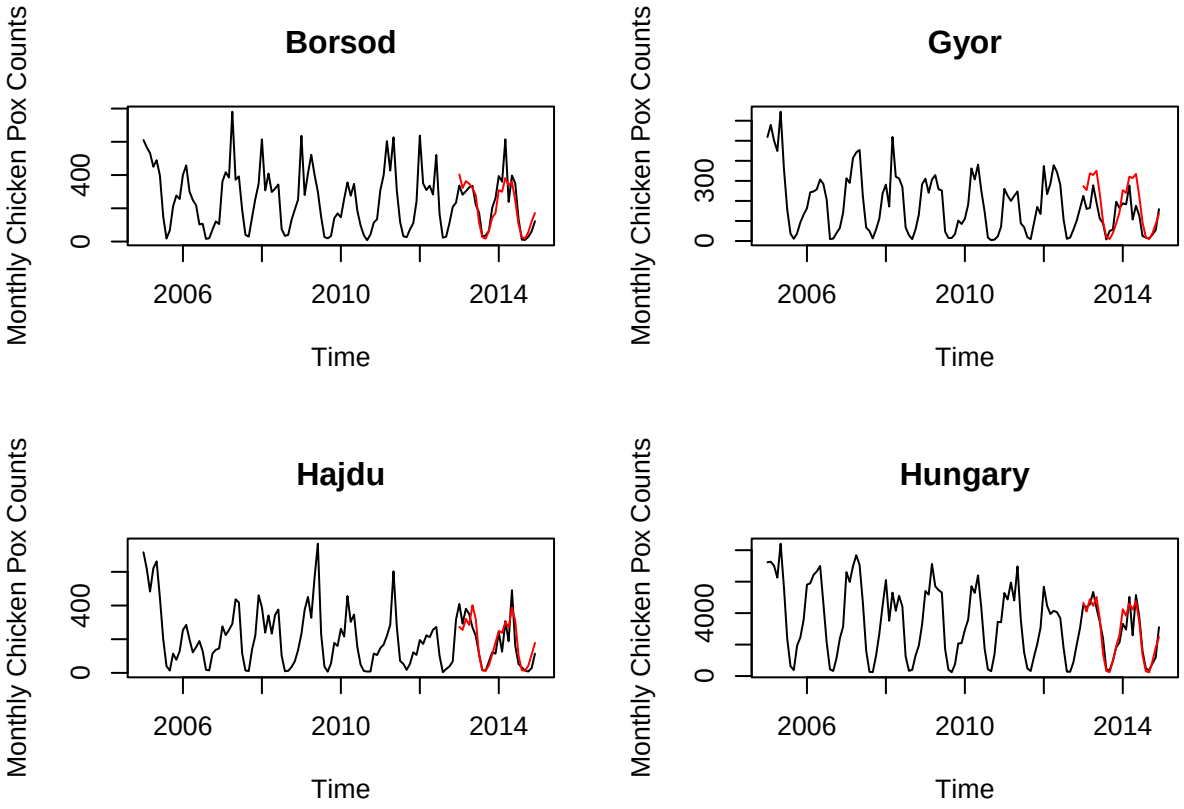
The ARIMA model ARIMA(1,1,1) x ARIMA(0,1,1) 12 produced the lowest AIC.

Below are the forecast plots from the resulting ARIMA models that produced the best results:

```
integer(0)
```

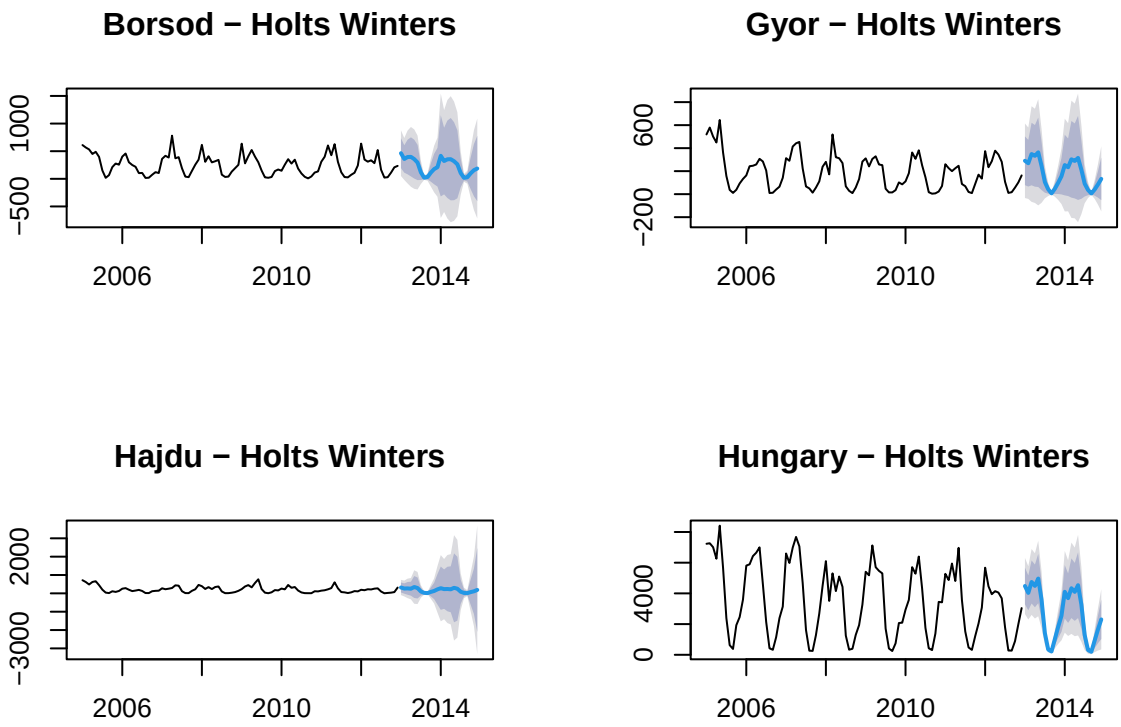
```
integer(0)
```

```
integer(0)
```




```
integer(0)
```

Holt Winters Smoothening Method



Evaluation of Models To compare the ARIMA models with the Holt Winters models, Root Mean Squared Error (RMSE) were performed and the results were as follows:

```
[1] "BORSOD ARIMA =  72.0977728010107"

[1] "BORSOD HW =  83.5800473510558"

[1] "GYOR ARIMA =  89.1256628721836"

[1] "GYOR HW =  88.1279188090823"

[1] "HAJDU ARIMA =  72.6881262222241"

[1] "HAJDU HW =  74.220103335884"

[1] "HUNGARY ARIMA =  587.676813349848"

[1] "HUNGARY HW =  576.85018506973"
```

While almost all of the time series exhibited a slight downward trend, only half of the models performed better when trend was included in the ARIMA model. Additionally, each county resulted in a different time series model: Borsod-Abaúj-Zemplén county was best predicted using an auto ARIMA model, Győr-Moson-Sopron was best predicted using Holt Winters Smoothing technique, and Hajdú-Bihar county was best predicted using a manual ARIMA model. The nation of Hungary was best predicted using Holt Winters Smoothing with multiplicative method technique.

Discussion

With this data set, the chickenpox case counts for three different counties in Hungary in addition to the chicken pox case count totals for the entire country with the goal of accomplishing two objectives: county level case count prediction and nation level case count prediction. The auto ARIMA model produced the lowest RMSE for Borsod-Abaúj-Zemplén County. The Holt Winters with multiplicative method model produced the lowest RMSE for Győr-Moson-Sopron County. The manual ARIMA model produced the lowest RMSE for Hajdú-Bihar County. The Holt Winters model produced the lowest RMSE for the nation of Hungary as a whole. The ARIMA models and the Holt Winters smoothing technique performed similarly. All of the models chosen produced models that were very close to the actual numbers that were observed for the years 2013 - 2014 effectively accomplishing the two objectives as defined.

Appendix

```
# Data Preprocessing
library(lubridate)
library(dplyr)
library(forecast)

k = 5 #BORSOD county - Ken
b = 8 #GYOR county  - Ben
t = 9 #HAJDU county - Tomi
bkt = 21 #HUNGARY - Ben Ken Tomi

#Read Chickenpox Case Data from CSV file
cpox <- read.csv('hungary_chickenpox.csv')
cpox$Date <- as.Date(cpox$Date, format = "%d/%m/%Y")
cpox$Month <- month(cpox$Date)
cpox$Year <- year(cpox$Date)

#Group data by monthly instead of weekly
cpox_new <- cpox %>%
  group_by(Year, Month) %>%
  select(-Date) %>%
  summarise(across(everything(), list(sum)))

#Create empty lists for county and nation data
county_list <- list()
county_list_log <- list()
ts_county <-list()
ts_county_log <-list()
ts_county_train <-list()
ts_county_test <-list()

#Store county and nation time series data into lists
for (i in 1:21) {
  if (i==21) {
    col_name <- 'HUNGARY'
    col_value <- rowSums(cpox_new[,3:ncol(cpox_new)])
    county_list[[col_name]] <- col_value
    col_value_log <- log(rowSums(cpox_new[,3:ncol(cpox_new)]))
    county_list_log[[col_name]] <- col_value_log
```

```
  } else {
    col_name <- colnames(cpox_new)[i+2]
    col_value <- cpox_new[,i+2]
    county_list[[col_name]] <- col_value
    col_value_log <- log(cpox_new[,i+2])
    county_list_log[[col_name]] <- col_value_log
  }

  ts_county[[col_name]] <- ts(county_list[[col_name]], start = c(2005, 1), frequency = 12)
  ts_county_log[[col_name]] <- ts(county_list_log[[col_name]], start = c(2005, 1), frequency = 12)
  ts_county_train[[col_name]] <- ts(ts_county_log[[col_name]][1:96], start = c(2005, 1), frequency = 12)
  ts_county_test[[col_name]] <- ts(ts_county_log[[col_name]][97:length(ts_county_log[[col_name]])], sta
}


par(mfrow = c(2, 2))

#Plot County and Nation Plots Before Log Transformation
for (i in c(k, b, t, bkt)) {
  plot(ts_county[[i]], xlab = "Year", ylab = "Chickenpox case count",
main = names(ts_county)[i])
}


#Plot 1st Difference of log(Chickenpox Case Count)
diff1 <- list()

for (i in 1:21) {
  col_name = names(ts_county_train)[i]
  diff1[[col_name]] <- c(NA, diff(ts_county_train[[i]]))
  diff1[[col_name]] <- ts(diff1[[col_name]], start = c(2005,1), deltat = 1/12)
}

par(mfrow = c(2, 2))

#Take seasonal difference at lag 12
diff12 <-list()

for (i in 1:21) {
  col_name = names(diff1)[i]
  diff12[[col_name]] <- c(NA, diff(diff1[[i]], lag = 12))
  diff12[[col_name]] <- ts(diff12[[col_name]], start = c(2005,1), deltat = 1/12)
  if (i %in% c(k, b, t, bkt)) {
    plot(diff12[[i]], xlab = "Year", ylab = "Seasonal difference at lag 12",
    main = names(diff1)[i])
    axis(1, at=seq(from=2005, to=2014, by=1), labels=seq(from=2005, to=2014, by=1))
  }
}


#Plot ACF and PACF for the differenced series

par(mfrow = c(4, 2), mar = c(3, 3, 3, 3))
```

```r
for (i in c(k, b, t, bkt)) {
  acf(diff12[[i]], lag.max = 36, na.action = na.pass,
  main = paste(names(ts_county_train)[i]," ACF Plot",sep = "_"))
  pacf(diff12[[i]], lag.max = 36, na.action = na.pass,
  main = paste(names(ts_county_train)[i]," PACF Plot",sep = "_"))
}


k_model <- auto.arima(ts_county_train[[k]])

b_model <- arima(ts_county_train[[b]], order = c(1, 1, 1),
                      seasonal = list(order = c(0, 1, 1), period = 12))

t_model <- arima(ts_county_train[[t]], order = c(1, 0, 1),
                      seasonal = list(order = c(0, 1, 1), period = 12))

bkt_model <- arima(ts_county_train[[bkt]], order = c(1, 1, 1),
                      seasonal = list(order = c(0, 1, 1), period = 12))


par(mfrow = c(2, 2))
# Borsod County
k_pred <- forecast(k_model, h = 24)

plot(exp(ts_county_log[[k]]), xlim = c(2005, 2015), main = 'Borsod', ylab = "Monthly Chicken Pox Counts
lines(exp(k_pred$mean), col = 'red')

# Gyor County
b_pred <- predict(b_model, n.ahead = 24)

plot(exp(ts_county_log[[b]]), xlim = c(2005, 2015), main = 'Gyor', ylab = 'Monthly Chicken Pox Counts')
lines(exp(b_pred$pred), col = 'red')

# Hajdu County
t_pred <- predict(t_model, n.ahead = 24)

plot(exp(ts_county_log[[t]]), xlim = c(2005, 2015), main = 'Hajdu', ylab = 'Monthly Chicken Pox Counts'
lines(exp(t_pred$pred), col = 'red')

# Hungary
bkt_pred <- predict(bkt_model, n.ahead = 24)

plot(exp(ts_county_log[[bkt]]), xlim = c(2005, 2015), main = 'Hungary', ylab = 'Monthly Chicken Pox Cou
lines(exp(bkt_pred$pred), col = 'red')


par(mfrow = c(2, 2))

hw_k <- hw(exp(ts_county_train[[k]]), seasonal = 'multiplicative', h = 24)
hw_b <- hw(exp(ts_county_train[[b]]), seasonal = 'multiplicative', h = 24)
hw_t <- hw(exp(ts_county_train[[t]]), seasonal = 'multiplicative', h = 24)
hw_bkt <- hw(exp(ts_county_train[[bkt]]), seasonal = 'multiplicative', h = 24)
```

```r
plot(hw_k, main = 'Borsod - Holts Winters')
plot(hw_b, main = 'Gyor - Holts Winters' )
plot(hw_t, main = 'Hajdu - Holts Winters')
plot(hw_bkt, main = 'Hungary - Holts Winters')


# RMSE Calculations

k_predictions <- data.frame(
  actual = exp(ts_county_test[[k]]),
  prediction_arima = exp(k_pred$mean),
  prediction_hw = hw_k$mean
)

b_predictions <- data.frame(
  actual = exp(ts_county_test[[b]]),
  prediction_arima = exp(b_pred$pred),
  prediction_hw = hw_b$mean
)

t_predictions <- data.frame(
  actual = exp(ts_county_test[[t]]),
  prediction_arima = exp(t_pred$pred),
  prediction_hw = hw_t$mean
)

bkt_predictions <- data.frame(
  actual = exp(ts_county_test[[bkt]]),
  prediction_arima = exp(bkt_pred$pred),
  prediction_hw = hw_bkt$mean
)

paste('BORSOD ARIMA = ' , sqrt(mean((k_predictions$actual - k_predictions$prediction_arima)^2)))
paste('BORSOD HW = ' , sqrt(mean((k_predictions$actual - k_predictions$prediction_hw)^2)))
paste('GYOR ARIMA = ' , sqrt(mean((b_predictions$actual - b_predictions$prediction_arima)^2)))
paste('GYOR HW = ' , sqrt(mean((b_predictions$actual - b_predictions$prediction_hw)^2)))
paste('HAJDU ARIMA = ' , sqrt(mean((t_predictions$actual - t_predictions$prediction_arima)^2)))
paste('HAJDU HW = ' , sqrt(mean((t_predictions$actual - t_predictions$prediction_hw)^2)))
paste('HUNGARY ARIMA = ' , sqrt(mean((bkt_predictions$actual - bkt_predictions$prediction_arima)^2)))
paste('HUNGARY HW = ' , sqrt(mean((bkt_predictions$actual - bkt_predictions$prediction_hw)^2)))
```