Forecasting Deaths Due to Influenza / Pneumonia

By John Russell

Abstract

Looking at a dataset of weekly deaths due to Influenza and Pneumonia from 2009 to 2018, time series analysis was performed in order forecast the deaths over the next 104 weeks (2 years). After the 2018 season, which was classified as high severity, the number of deaths due to influenza or pneumonia are supposed to decrease back towards the mean, which is about 7.2 percent.

Introduction

Both influenza and pneumonia are major concerns in the health field. Pneumonia is the world's leading cause of death for children under 5 years. It is also the leading cause of hospital admissions for adults in the US other than women giving birth (1). With that many hospitalizations, it is important that the treatment is effective and people are able to recover.

When looking at the given dataset, which is the collection of weekly deaths due to Influenza (Flu) and Pneumonia from 2009-2018, there are a number of possible questions that could be asked. Which region has the highest percentage of death due to the flu and pneumonia? Does age play a significant role in dying due to the infections? And, given how high the percentage of deaths due to the two infections in 2018 was, is there a possibility of an upwards trend?

Looking closer at the latter, this question is of significance because an upwards trend could indicate that our current methods used to combat both the flu and pneumonia are losing effectiveness, the strain of infection is changing faster than we can adapt, or other potentially dangerous causes. The hypothesis reached after an initial examination of the 10 years in which data was collected, is that the deaths due to influenza and pneumonia are not trending upwards and should head back towards the mean as time increases. Historically, there have been peaks in the deaths due to the infections, but overall they have been fairly constant. In order to test the hypothesis, time series analysis is used. Forecasting the next two years should give sufficient evidence towards reaching a conclusion.

Method

The data collected looks at the influenza and pneumonia deaths recorded by the Center of Disease Control's (CDC) Morbidity and Mortality Weekly Report (MMWR). The original dataset has been broken into three parts. The first, which is the one used for this analysis, has the first 479 observations of weekly national data with all ages from 2009-2018. More specifically, week 40 of 2009 to week 49 of 2018. The second has 1,438 observations categorized by age. The third has 4,791 observations categorized by region.

There are a total of 11 variables: geoid, which measures geographical data (either national or regional); Region, which measures the region of the United States that observation is from; State, which measure the state in the region; season, which measures the infection "season", each a year in length; MMWR Year/Week, which measures the week and year in which that data was collected by the MMWR; Deaths

from influenza, which measures the total deaths caused primarily by the flu; Deaths from pneumonia, which measures the total deaths caused primarily by pneumonia; Deaths from pneumonia and influenza, which measures the total deaths caused by either the flu or pneumonia; All Deaths, which measures the total number of deaths for that given week, regardless of cause; and Percent of deaths due to pneumonia or influenza, which measures the percentage of deaths caused by either the flu or pneumonia.

Of these 11 variables, geoid, Region, state, and age are not used in the first dataset, so they are removed. The response variable that will give the most information towards reaching a valid conclusion is the percentage of deaths due to influenza or pneumonia, so all other death-related variables are redundant. They are also removed. Finally, as both season and MMWR Year/Week are both time-related variables, only MMWR Year/Week is used due to being more precise. The final dataset (first 6 observations) is shown in Table 1.

Table 1: Dataset

MMWR.Year.Week	Pecent.of.deaths.due.to.pneumonia.or.influenza
200940	7.82771697
200941	8.346070222
200942	8.598343554
200943	8.737508615
200944	8.77311492
200945	8.864183963

Traditional time series analysis was performed (2). An ARIMA model was fitted using the `auto.arima` function in R. Forecasting was also done in R, using the `forecast` function. Two models were considered: one removing the seasonality component and the other leaving the model unadjusted for seasonality. The AIC (Akaike Information Criteria) was used to select the best fit. The best model was the regression with ARIMA(2,1,4) errors. The reason for this is due to standard ARIMA models not being able to handle non-integers for time values. Since the seasonal period of the data is actually 52.18 (365.25 / 7 to account for leap years), the results will not be optimal (3).

Analysis

A time series analysis method was used since the data was collected in weekly intervals. Choosing a good model and forecasting the data accurately was possible with the amount of observations given. All analysis was done using R software. The original data has the variable MMWR Year/Week as an integer. The first step taken was to convert that integer to a time variable. This was done using the `as.Date` function. The data is `checked` for stationarity. Figure 1 shows the graph of the original data.

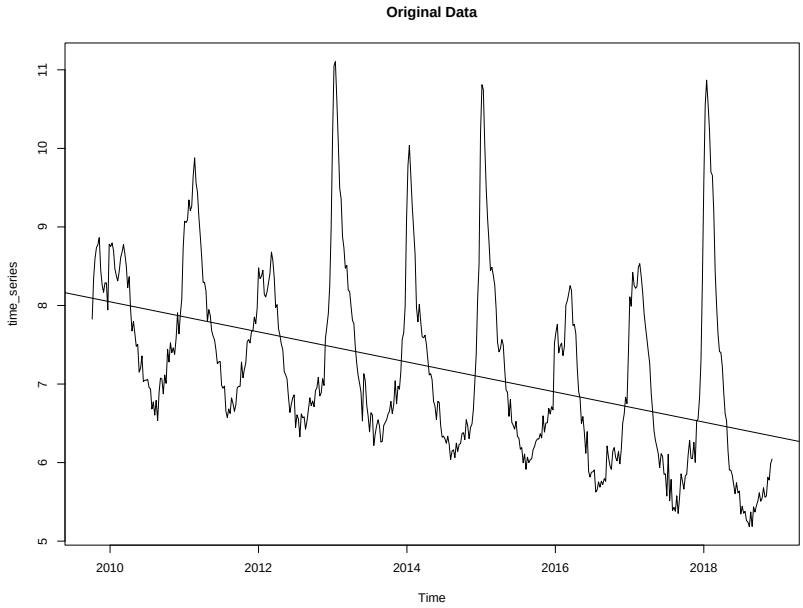


Figure 1: Original data

The graph shows that while the mean value seems to be decreasing, the difference between the mean and peak values is growing. Since the mean is not constant over time, differencing in required. Figure 2 shows the graph of the differenced data.

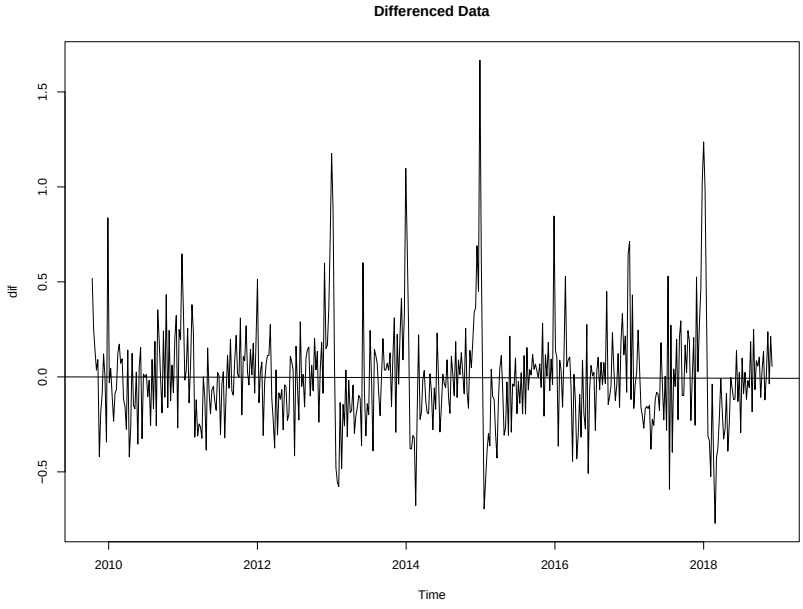


Figure 2: Differenced Data

Now that the mean is constant, the next issue is the potential seasonality effect. Figure 3 is the graph of the data after removing the lag, which in this case is 52, for the number of weeks in a given year.

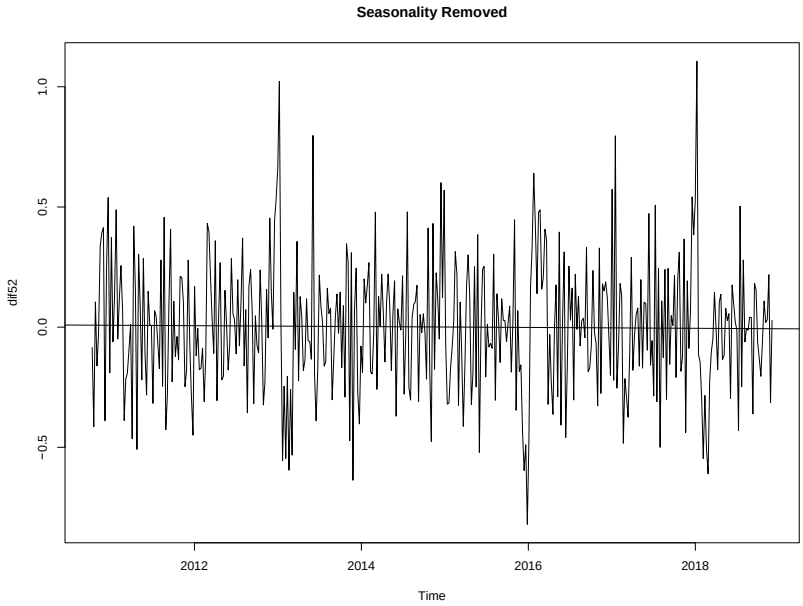


Figure 3: Graph of data adjusted for seasonality

Although the seasonal differenced model looks slightly better in terms of non-discerning patterns, the differenced model is also in the range of acceptable outcomes. Therefore, both models were used in model fitting.

The `auto.arima` function chooses the most optimal model for the data. This is the method that was used in picking the model. The two models chosen were the result of including seasonal models in the search and constricting the search to non-seasonal models. The models were a regression model with ARIMA(2,1,4) errors and a regression model with ARIMA(0,1,2)(1,0,0)[52] errors. The AIC of the two models were -53.19 and -41.75, respectively. Based solely on the AIC, the regression model with ARIMA(2,1,4) errors should be picked, but the residuals were checked on both models first to check assumptions. The residuals should look like white noise.

Figure 4 shows the residual plots for the regression model with ARIMA(2,1,4) errors. Figure 5 shows the residual plots for the regression model with ARIMA(0,1,2)(1,0,0)[52] errors.

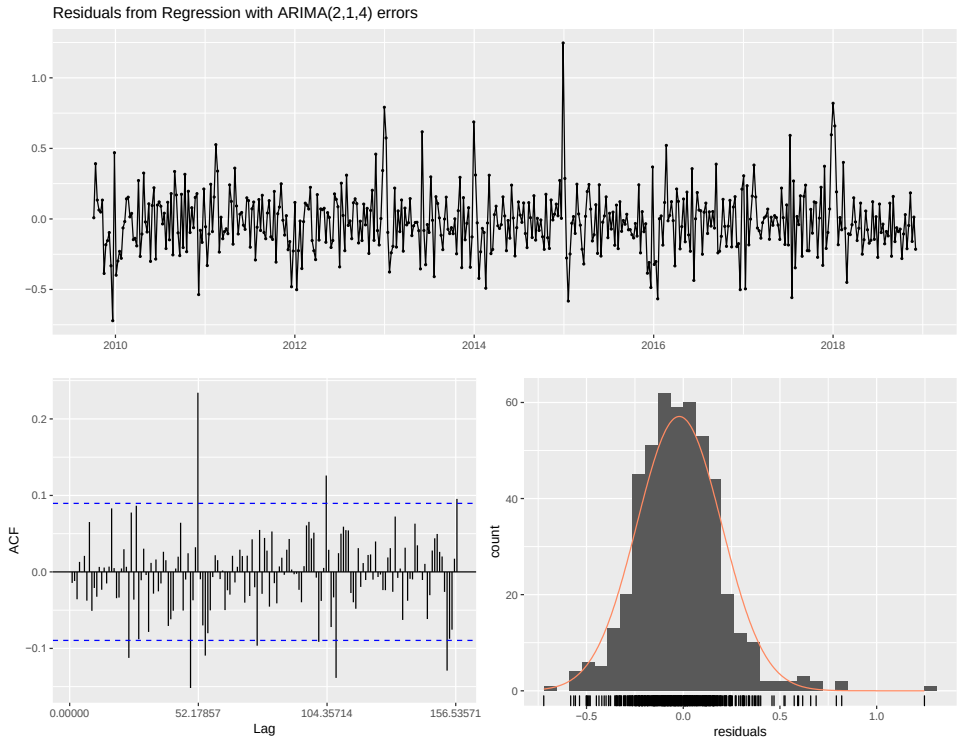


Figure 4: Residuals from Regression with ARIMA(2,1,4) errors

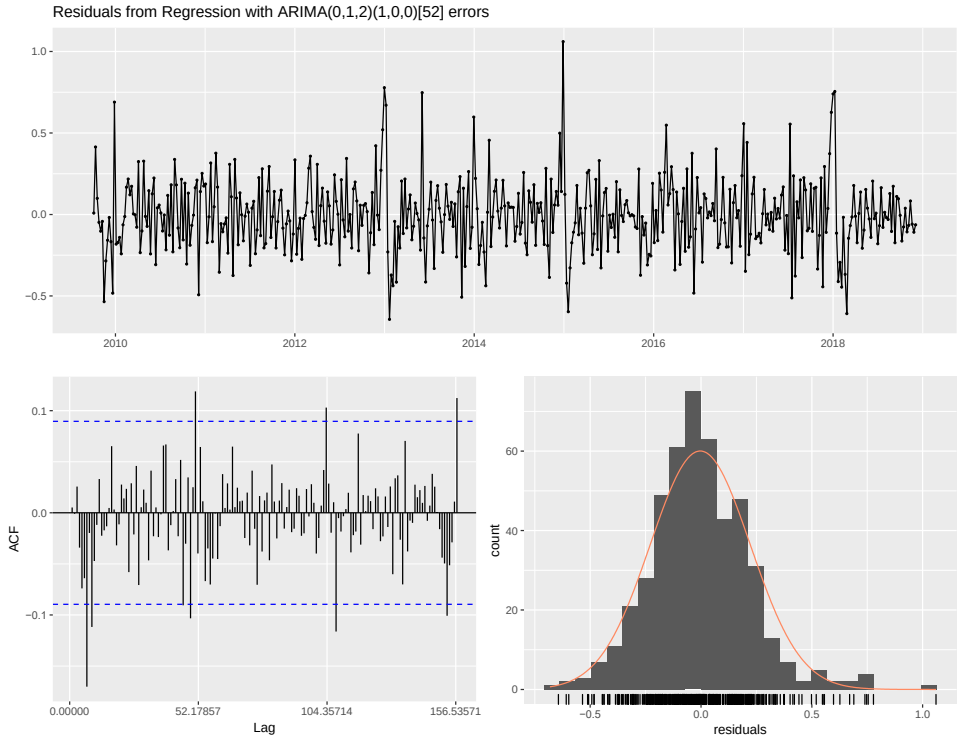


Figure 5: Residuals from Regression with ARIMA(0,1,2)(1,0,0) [52] errors

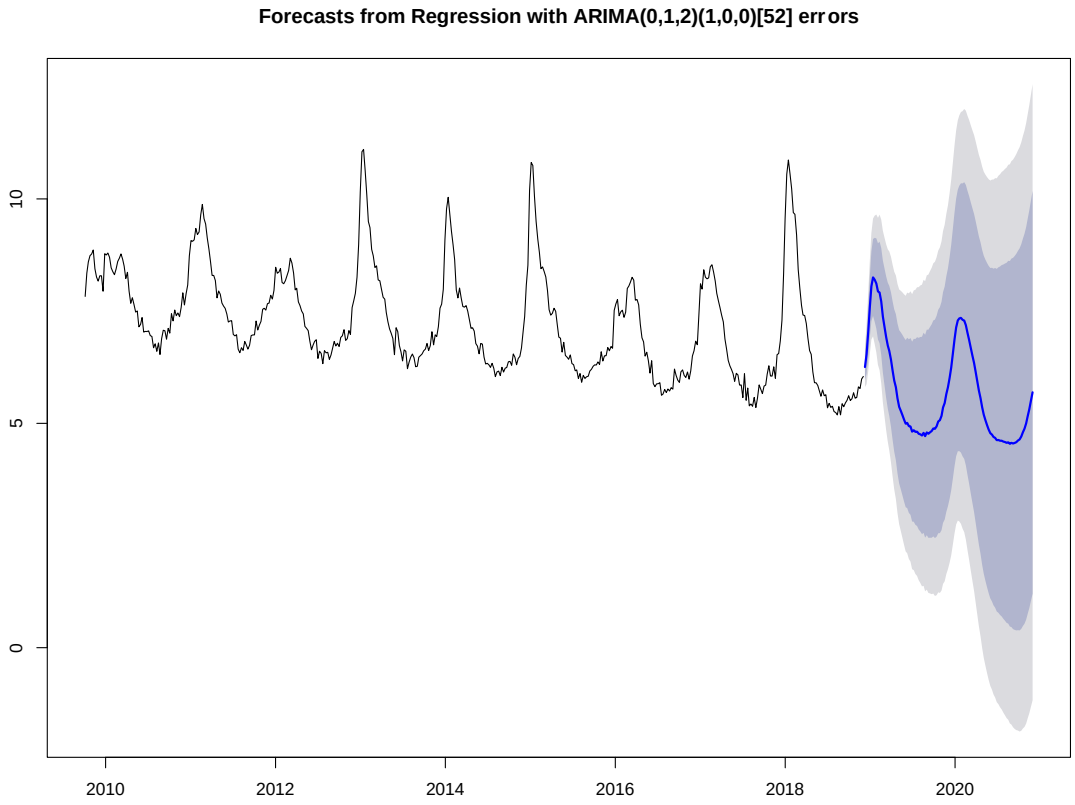
Both models have errors that appear to be white noise. However, it is worth noting that the Ljung-Box statistic for the regression model with ARIMA(2,1,4) errors is very high. The p-value is

significantly lower than the chosen alpha level of 0.05, while the p-value for the regression model with ARIMA(0,1,2)(1,0,0) [52] errors is 0.3. Thus, instead of choosing the model with the lower AIC, as the AIC values were not too different, the final model is a regression model with ARIMA(0,1,2)(1,0,0)[52] errors. Table 2 shows the variables in the final model.

Variable	Coefficient	SE	Ljung-Box Statistic	P-Value
MA1	0.1122	0.0464		
MA2	0.1542	0.0459		
SAR1	0.3616	0.0464		
S1-52	0.9635	0.2261		
C1-52	-0.7705	0.2270		
S2-52	-0.3334	0.1126		
C2-52	-0.1205	0.1120		
			94.569	0.3179

The MA1 and MA2 variables correspond to the MA(2) process of the error terms. The SAR2 corresponds to the seasonality effect. The rest of the variables are needed to fully capture the seasonality of the model.

After the model was chosen, it was used to forecast the next 104 weeks (2 more seasons) to see if the maximum percentage of deaths continues to increase. The `forecast` function was used to plot the graph. Figure 6 shows the results.



The forecast plot predicts the deaths to decrease back towards the mean. The confidence interval shows that there is a chance that the percentage of deaths continues to rise.

Conclusion

From this analysis, it appears that the past season, where the flu and pneumonia were classified as “high severity’, were more of an outlier than a trend. During every year, deaths caused by the flu and pneumonia are at their highest during winter. The data coincides with this; the weeks with the highest percentage of deaths are, on average, between the 40th week of the year and the 15th week of the following year. This is between October and January. During these periods of time, the percentage of deaths are at the maximum for that year, and the past year saw the percentage reach about 10%. From forecasting, the next two years see an expected maximum that is closer to the average, which is 7.2%.

One limitation of the analysis is the fact that age and region were not considered. There is a possibility that either one of those factors might have a significant impact on the percentage of deaths due to influenza and pneumonia. Another limitation is that influenza and pneumonia were grouped together. One of the two infections might see a different outcome when looked at alone, especially since the numbers are heavily skewed towards pneumonia. This is very significant, due to the fact that one of the risk factors of pneumonia is recently having a viral respiratory infection, like influenza (4).

A future study could involve the factors mentioned previously: age, region, and separating the flu and pneumonia. This would allow a more precise conclusion to be reached. This study would require multivariate time series applications to handle the extra variables.

References

1. “Top 20 Pneumonia Facts.” *American Thoracic Society,* American Thoracic Society, 2015, www.thoracic.org/.
2. Wei, William W. S. *Time Series Analysis: Univariate and Multivariate Methods*. Pearson Education, 2006.
3. Hyndman, Rob J. “Forecasting Weekly Data.” *Forecasting Weekly Data*, 4 Mar. 2014, robjhyndman.com/hyndsight/forecasting-weekly-data/.
4. “Pneumonia Symptoms, Causes, and Risk Factors.” *American Lung Association,* 15 Oct. 2018, www.lung.org/lung-health-and-diseases/lung-disease-lookup/pneumonia/symptoms-causes-and-risk.html.

Appendix

```
library(lubridate)
library(zoo)
library(forecast)
data <- read.csv("DataFileSp19.csv", header = T)
names(data)
dataset_1 <- data[1:479,]
yrwk <- dataset_1[, 6]
date1 <- as.Date(paste(yrwk, 1), "%Y%U %u")

time_series <- ts(dataset_1$Pecent.of.deaths.due.to.pneumonia.or.influenza, frequency =
          365.25/7, start = decimal_date(ymd(date1[1])))

#No Seasonality
best_fit1 <- list(aicc = Inf)
for(i in 1:25)
{
 fit1 <- auto.arima(time_series, xreg = fourier(time_series, K=i), seasonal = F)
 if(fit1$aicc < best_fit1$aicc)
   best_fit1 <- fit1
 else break;
 k_value1 <- max(i)
}

#Verify model
dif <- diff(time_series)
plot(dif, main = "ARIMA(2,1,4)")
acf_plot <- acf(dif, plot = F)
acf_plot$lag <- acf_plot$lag * 52
plot(acf_plot, col = "blue", xlab = "Lag (Weeks)")
pacf_plot <- acf(dif, lag.max = 104, type = "partial", plot = F)
pacf_plot$lag <- pacf_plot$lag * 52
plot(pacf_plot, col = "red", xlab = "Lag (Weeks)")

checkresiduals(best_fit1)


#Seasonality
best_fit2 <- list(aicc = Inf)
for(i in 1:25)
{
```

```
  fit2 <- auto.arima(time_series, xreg = fourier(time_series, K=i), seasonal = T)
  if(fit2$aicc < best_fit2$aicc)
    best_fit2 <- fit2
  else break;
  k_value2 <- max(i)
}

#Verify Model
dif52 <- diff(dif, lag = 52)
plot(dif52, main = "ARIMA(0,1,2)(1,0,0)[52]")
acf_plot_season <- acf(dif52, lag.max = 104, plot = F)
acf_plot_season$lag <- acf_plot_season$lag * 52
plot(acf_plot_season, col = "blue", xlab = "Lag (Weeks)")
pacf_plot_season <- acf(dif52, lag.max = 104, type = "partial", plot = F)
pacf_plot_season$lag <- pacf_plot_season$lag * 52
plot(pacf_plot_season, col = "red", xlab = "Lag (Weeks)")

checkresiduals(best_fit2)

#Forecasting
pred <- forecast(best_fit2, xreg = fourier(time_series, K = 2, h = 104))
plot(pred
```