Case4

Matt Chinchilla, Drew Larsen, Rikel Djoko

10/18/2020

# Business Understanding

Influenza, commonly referred to as the flu, is a worldwide respiratory infectious disease that easily spreads from one person to another. In the North America region each year more than 2,000,000 people are hospitalized with flu, and about 36,000 die. Flu spreads mainly from person-to-person through coughs or sneezes, creating aerosols containing the virus from infectious individuals.

In this study, we used 20 years worth of data from 2000 to 2020 pulled from the CDC website(<https://www.cdc.gov/flu/>) to run and determine the appropriate autoregressive integrated moving average (ARIMA) model for influenza transmission in North America. Forecasting of seasonal infectious diseases, such as influenza, can help in public health planning and outbreak response.

We assume that the data on the CDC website is correct and represents flu cases in the United States as a whole. The results and conclusion of this analysis should only be applied to flu cases in the United States aggregate. In order to use ARIMA, we must make the assumption that the correlation structure of how current flu cases in the US relate to flu cases in the past does not change as we move forward in time.

# Data extraction and Evaluation

* Source: World Health Organization FluNet <https://apps.who.int/flumart/Default?ReportNo=7>
* Country: United States
* Time Period: 2000-2020(Week38)

Below is information about the structure of the data collected from the W.H.O. fluNet website. Twenty year’s worth of data was collected from 2000 to week 38 of 2020. There are 22 columns, most of which are continuous variables, and six categorical columns. Almost eighteen percent of the data consists of missing values.

**Overview Stats:**

rows: 1081

columns: 22

discrete\_columns: 6

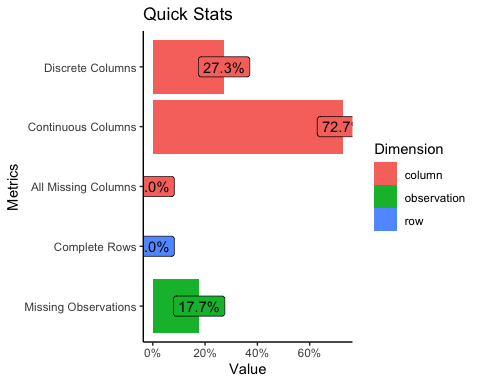
continuous\_columns: 16

all\_missing\_columns: 0

total\_missing\_values: 4213

complete\_rows: 0

total\_observations: 23782



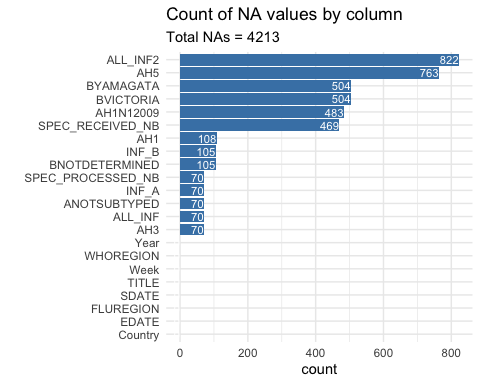
### Data columns and classes

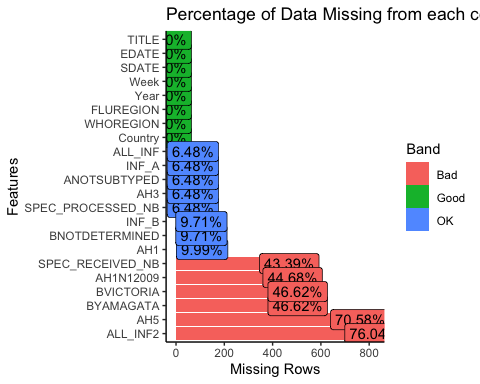
The response column we are interested in predicting is ALL\_INF, which is a weekly aggregate of positive cases for all strains of the flu. Since we are only focused on the United states there is redundant data in several columns including “Country”,“WHOREGION”,and “FLUREGION”.

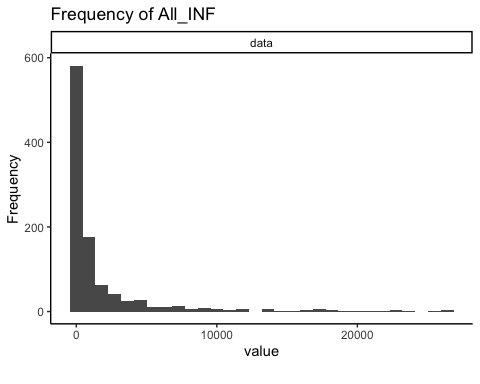
## 'data.frame': 1081 obs. of 22 variables:  
## $ Country : chr "United States of America" ...  
## $ WHOREGION : chr "Region of the Americas of WHO" ...  
## $ FLUREGION : chr "North America" ...  
## $ Year : int 2000 2000 ...  
## $ Week : int 1 2 ...  
## $ SDATE : chr "1/3/00" ...  
## $ EDATE : chr "1/9/00" ...  
## $ SPEC\_RECEIVED\_NB : int NA NA ...  
## $ SPEC\_PROCESSED\_NB: int 7450 5959 ...  
## $ AH1 : int 0 2 ...  
## $ AH1N12009 : int NA NA ...  
## $ AH3 : int 457 302 ...  
## $ AH5 : int NA NA ...  
## $ ANOTSUBTYPED : int 1153 713 ...  
## $ INF\_A : int 1610 1017 ...  
## $ BYAMAGATA : int NA NA ...  
## $ BVICTORIA : int NA NA ...  
## $ BNOTDETERMINED : int 2 3 ...  
## $ INF\_B : int 2 3 ...  
## $ ALL\_INF : int 1612 1020 ...  
## $ ALL\_INF2 : int NA NA ...  
## $ TITLE : chr "Widespread Outbreak" ...

### Missing Data

Below are charts to help visualize how much data is missing from each column and the frequency of data within ALL\_INF. The first chart shows the count of missing values in each column, the second chart depicts the percentage of missing data from each column. ALL\_INF is the primary column we are focused on predicting and we can see that there are 70 missing values in that column which only represents 6.48% of the overall data in the column. The chart Frequency ALL\_INF shows the frequency of values in the ALL\_INF column as expected the data is right skewed with the majority of values being less than 10,000.







# Modeling Building

In order to predict flu cases in the coming weeks, we have decided to build and compare multiple ARIMA(p,d,q) models. ARIMA is different than a normal deterministic model, such as a linear regression model, in that the data at previous lags is used to calculate the current data point. The ARIMA model is broken down into 3 parts: the “AR(p)” part, the “I” part, which is where we get our d, and the “MA(q)” part.

The “AR(p)” part, or Auto-Regressive model, is a function of the values at previous lags of the data set. The p signifies how far back we include lags in our model. If we have a p of 7, we look 7 lags back in order to calculate the current point.

The “MA(q)” part, or Moving Average model, is a function of the average of the previous q values in the variable. The average of previous lags can tell us something about the value of the variable at the current time.

The I, or integrated, part of the ARIMA model signifies how many times we “difference” the data. Differencing occurs when you take the value of the variable at a time and subtract the value of the variable at the time directly before it. The amount of times that you difference is the “d” in the ARIMA(p,d,q) model.

We also include seasonality in this data. Seasonality is similar to differencing, but instead of differencing the previous time step, we difference s time steps previously. If s = 4, then we subtract using the value of the variable 4 time steps prior.

We difference the data in order to make it “stationary.” Stationarity states that the mean, variance, and the dependence on previous lags do not change over time. AR, MA, and ARMA (a combination of AR and MA models) require that the data is stationary. Often, we check if the data is stationary and if not, we difference the data until it is. We then take the differenced data and fit an ARMA model to it.

These models are useful when we have temporally correlated data, as we do now. Having correlation in the data makes it difficult to create models using traditional statistics, as an assumption for these models typically is that the variable’s values are independent of itself, which is not the case with flu data over time. If you have a high number of cases yesterday, you are likely to have a high number of cases today. The ARIMA(p,d,q) model allows us to use the correlations present in the data in order to aid in our predictions.

We have pulled data from the CDC website relating to flu cases. There are many different types of flus, so we have decided to use the ALL\_INF variable as our data.

As globalization has increased rapidly in the past few years, it became clear to us that the trends in the data before 2015 may not be the same as the trends in the data after 2015, so we have decided to cut our data set to the past 5 years. There are no NA values present in this data set.

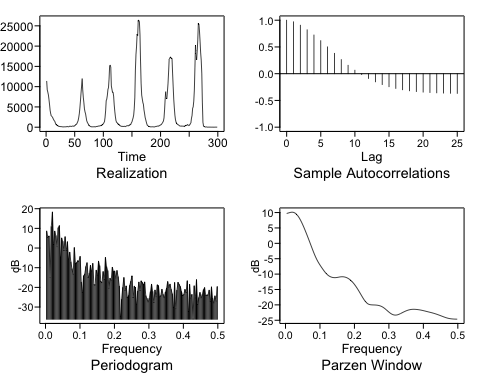
## The number of NAs in the time series is 0

Here we use the Ljung-Box test to decide if there is temporal dependence present in this data. A p-value less than 0.05 indicates that there is temporal dependence present, and we can move on with modeling our data using ARIMA. Had the p-value been greater than 0.05, there would not be enough evidence that there is temporal dependence in the data, and we may be fitting a model on white noise. Using an ARIMA model on white noise would be useless as there is no dependence on previous lags in this data, which is what ARIMA leverages to make predictions.

## Obs 0.9704079 0.9062268 0.820728 0.7233893 0.6177477 0.5014781 0.3808711 0.263803 0.157425 0.06513914 -0.0160519 -0.08913658 -0.1522069 -0.2033115 -0.243746 -0.2755208 -0.3005393 -0.3196223 -0.3339341 -0.3444659 -0.3522778 -0.3582031 -0.362771 -0.3661942

## $test  
## [1] "Ljung-Box test"  
##   
## $K  
## [1] 24  
##   
## $chi.square  
## [1] 1537.237  
##   
## $df  
## [1] 24  
##   
## $pval  
## [1] 0

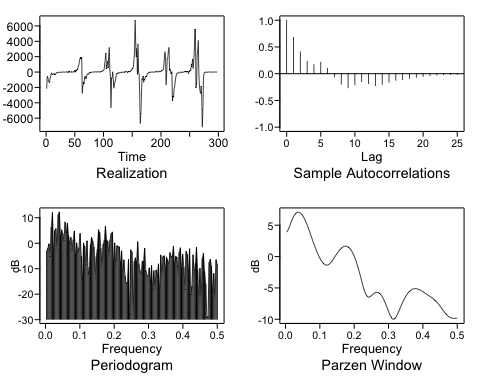
You can see the plot of the data set in the top left, the sample autocorrelations in the top right, the Periodogram in the bottom left and the Parzen window in the bottom right. The Parzen window is a smoothed version of the Periodogram, so we will focus on the Parzen window instead. Note the strong seasonal trend in the data. Every 52 weeks or so, flu cases peak. Since the flu is a seasonal virus, we expect to see that behavior in the data. The autocorrelation plot shows what lags the current time step depends on and how strongly it depends on the previous lags. Looking at the autocorrelation plot, we see a slow dampening of the autocorrelations. This indicates that we may need to difference the data, or that there may be a moving average component present in the data. Lastly, we look at the Parzen window. This plots the frequencies present in the data. The Parzen window has a peak around 0.0, which can indicate strong wandering behavior or a possible seasonal trend with a low frequency.



We use the Dickey-Fuller test below, which tests if the data is stationary. The data passes the Dickey-Fuller test, but this is a test with a high false positive rate. We made the decision to make multiple models and compare them, rather than relying only on the Dickey-Fuller test. These results support our first model, which assumes that the model doesn’t need to be differenced. The plots above suggest that there is some seasonality in the data and there might be some differencing necessary.

##   
## Augmented Dickey-Fuller Test  
##   
## data: fluts  
## Dickey-Fuller = -5.1263, Lag order = 6, p-value = 0.01  
## alternative hypothesis: stationary

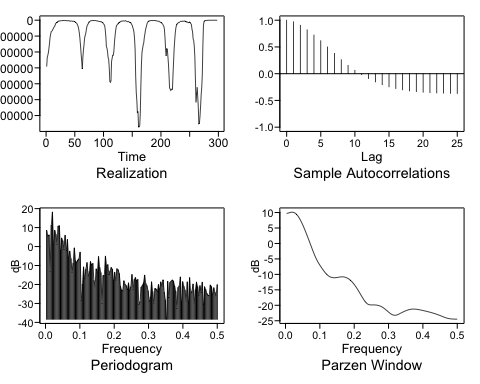
Here is the plot, the sample autocorrelations, the Periodogram and Parzen window of the first differenced data. Note that the seasonality is still present in the data, but there is less of a damping trend in the sample autocorrelations. Differencing definitely removes some of the wandering that the sample autocorrelations and Parzen window detected in the data.



The Dickey-Fuller test says that the differenced data is stationary, so we can feel confident fitting an ARMA model to the differenced data.

##   
## Augmented Dickey-Fuller Test  
##   
## data: second  
## Dickey-Fuller = -6.3425, Lag order = 6, p-value = 0.01  
## alternative hypothesis: stationary

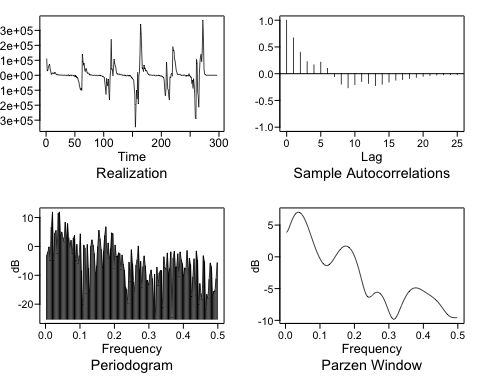
Here is the plot, the sample autocorrelations, the Periodogram and Parzen window of the seasonally differenced data with s=52. We chose 52 since we expect the flu trend to repeat every year and we have weekly data. Note that the plot looks like it flipped upside down and note that the wandering behavior is still present in the autocorrelations and the Parzen window.



The Dickey-Fuller test suggests that we have stationary data after seasonally differencing, so we can go ahead and fit an ARMA model to the remaining data.

##   
## Augmented Dickey-Fuller Test  
##   
## data: seasonal  
## Dickey-Fuller = -5.1597, Lag order = 6, p-value = 0.01  
## alternative hypothesis: stationary

Here is the plot, the sample autocorrelations, the Periodogram and Parzen window of the differenced and seasonally differenced s=52 data. Note that it looks like the plot of the differenced data is flipped, and that the autocorrelations and Parzen window look similar to the single differenced data.



The Dickey-Fuller test suggests that we have stationary data after seasonally and regularly differencing, so we can go ahead and fit an ARMA model to the remaining data.

##   
## Augmented Dickey-Fuller Test  
##   
## data: seasonalanddiff  
## Dickey-Fuller = -6.3172, Lag order = 6, p-value = 0.01  
## alternative hypothesis: stationary

Here, we use AIC to determine the best p and q for our non-differenced model. AIC suggests that we use p = 7 and q = 0 after testing all combinations of p between 0 and 10 and q between 0 and 5.

## Five Smallest Values of aic

## p q aic  
## 43 7 0 13.76056  
## 44 7 1 13.76366  
## 49 8 0 13.76421  
## 38 6 1 13.76768  
## 55 9 0 13.76959

Here, we estimate the parameters of the AR(7) model.

##   
## Coefficients of Original polynomial:   
## 1.6743 -0.7657 -0.0093 0.0134 0.3781 -0.4893 0.1423   
##   
## Factor Roots Abs Recip System Freq   
## 1-1.8293B+0.8642B^2 1.0584+-0.1924i 0.9296 0.0286  
## 1-0.6567B+0.6414B^2 0.5119+-1.1388i 0.8009 0.1828  
## 1+1.2372B+0.6031B^2 -1.0257+-0.7784i 0.7766 0.3967  
## 1-0.4256B 2.3498 0.4256 0.0000  
##   
##

The Ljung-Box test suggests that we modelled the temporal dependence out of the data with an AR(7) model, so we can move forward with this model.

## Obs -0.009515332 0.006248785 0.01510153 0.02191927 -0.006659231 -0.01263407 0.004101544 -0.009478087 -0.1202596 -0.06087474 0.1275253 -0.005015893 -0.08822134 0.04670769 0.02983544 -0.009385244 -0.04091451 0.01542245 -0.005579032 -0.01205638 -0.02146238 -0.01499615 -0.03314868 -0.01951292

## $test  
## [1] "Ljung-Box test"  
##   
## $K  
## [1] 24  
##   
## $chi.square  
## [1] 15.90351  
##   
## $df  
## [1] 24  
##   
## $pval  
## [1] 0.8915271

Here, we use AIC to determine the best p and q for our single differenced model. AIC suggests that we use p = 6 and q = 1 after testing all combinations of p between 0 and 10 and q between 0 and 5.

## Five Smallest Values of aic

## p q aic  
## 38 6 1 13.81826  
## 49 8 0 13.82817  
## 55 9 0 13.82984  
## 50 8 1 13.83052  
## 43 7 0 13.83177

Here, we estimate the parameters of the ARIMA(6,1,1) model.

##   
## Coefficients of Original polynomial:   
## 1.3413 -0.4976 -0.0379 -0.0163 0.3524 -0.3256   
##   
## Factor Roots Abs Recip System Freq   
## 1-1.7623B+0.8323B^2 1.0587+-0.2840i 0.9123 0.0417  
## 1-0.7366B+0.7037B^2 0.5233+-1.0711i 0.8389 0.1777  
## 1+1.1575B+0.5560B^2 -1.0410+-0.8456i 0.7456 0.3914  
##   
##

The Ljung-Box test suggests that we modelled the temporal dependence out of the data with an ARIMA(6,1,1) model, so we can move forward with this model.

## Obs 0.0004102403 -0.002388576 0.003971036 0.01153195 -0.01053103 -0.05415727 0.02234946 0.0168021 -0.09214613 -0.03623838 0.1224524 -0.0117118 -0.09838454 0.03215974 0.01915885 -0.03522965 -0.07913208 -0.01660961 -0.02422907 -0.02849168 -0.03838718 -0.03272334 -0.0442563 -0.02854876

## $test  
## [1] "Ljung-Box test"  
##   
## $K  
## [1] 24  
##   
## $chi.square  
## [1] 17.09307  
##   
## $df  
## [1] 24  
##   
## $pval  
## [1] 0.8446657

Here, we use AIC to determine the best p and q for our seasonally differenced model. AIC suggests that we use p = 7 and q = 0 after testing all combinations of p between 0 and 10 and q between 0 and 5.

##   
## Five Smallest Values of aic

## p q aic  
## 43 7 0 21.66671  
## 44 7 1 21.66963  
## 49 8 0 21.67018  
## 54 8 5 21.67049  
## 38 6 1 21.67316

Here, we estimate the parameters of the ARIMA(7,0,0) model with s = 52.

##   
## Coefficients of Original polynomial:   
## 1.6551 -0.7348 -0.0223 0.0129 0.3777 -0.4822 0.1361   
##   
## Factor Roots Abs Recip System Freq   
## 1-1.8309B+0.8657B^2 1.0575+-0.1921i 0.9304 0.0286  
## 1-0.6559B+0.6382B^2 0.5138+-1.1414i 0.7989 0.1827  
## 1+1.2397B+0.6040B^2 -1.0261+-0.7762i 0.7772 0.3969  
## 1-0.4079B 2.4513 0.4079 0.0000  
##   
##

The Ljung-Box test suggests that we modelled the temporal dependence out of the data with an ARIMA(7,0,0) with s = 52 model, so we can move forward with this model.

## Obs -0.009477356 0.007117896 0.01523798 0.02167102 -0.006084615 -0.01171475 0.002099987 -0.01067088 -0.1207629 -0.06103191 0.1276156 -0.005315792 -0.08827839 0.04653268 0.02965834 -0.008841763 -0.04047308 0.01531056 -0.006094186 -0.01147281 -0.01999726 -0.01460777 -0.03328357 -0.017317

## $test  
## [1] "Ljung-Box test"  
##   
## $K  
## [1] 24  
##   
## $chi.square  
## [1] 15.82993  
##   
## $df  
## [1] 24  
##   
## $pval  
## [1] 0.8941165

Here, we use AIC to determine the best p and q for our differenced and seasonally differenced model. AIC suggests that we use p = 6 and q = 1 after testing all combinations of p between 0 and 10 and q between 0 and 5.

##   
## Error in aic calculation at 10 2   
## Five Smallest Values of aic

## p q aic  
## 38 6 1 21.72425  
## 49 8 0 21.73460  
## 55 9 0 21.73593  
## 43 7 0 21.73842  
## 45 7 2 21.73912

Here, we estimate the parameters of the ARIMA(6,1,1) model with s = 52.

##   
## Coefficients of Original polynomial:   
## 1.3406 -0.4864 -0.0459 -0.0154 0.3529 -0.3262   
##   
## Factor Roots Abs Recip System Freq   
## 1-1.7668B+0.8355B^2 1.0573+-0.2810i 0.9141 0.0413  
## 1-0.7353B+0.7002B^2 0.5251+-1.0736i 0.8368 0.1776  
## 1+1.1614B+0.5576B^2 -1.0414+-0.8419i 0.7467 0.3918  
##   
##

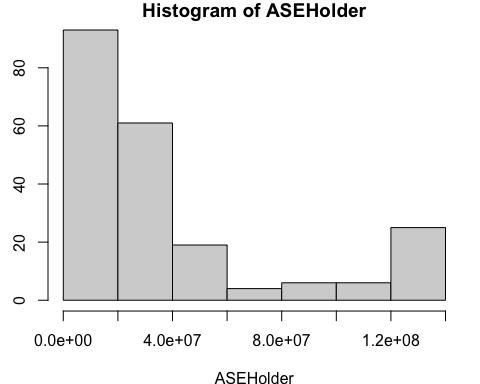
The Ljung-Box test suggests that we modelled the temporal dependence out of the data with an ARIMA(6,1,1) with s = 52 model, so we can move forward with this model.

## Obs 0.0004084121 -0.001409626 0.004690342 0.01163992 -0.01023082 -0.05290639 0.02027867 0.01528356 -0.09260387 -0.03584375 0.1239257 -0.01051667 -0.09715351 0.03324175 0.01993657 -0.03389561 -0.07777323 -0.01613628 -0.02451116 -0.02897782 -0.03902632 -0.03343543 -0.04532947 -0.02992604

## $test  
## [1] "Ljung-Box test"  
##   
## $K  
## [1] 24  
##   
## $chi.square  
## [1] 17.03383  
##   
## $df  
## [1] 24  
##   
## $pval  
## [1] 0.8472158

## Forecasting

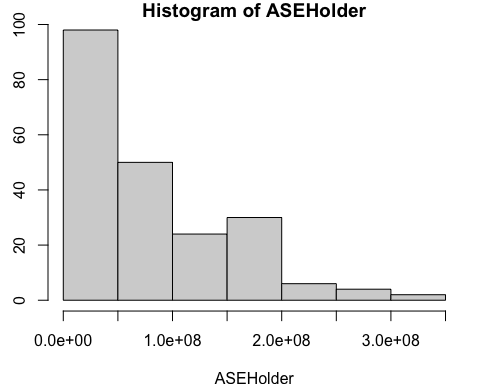
Here, we use a rolling window to determine the average squared error of our first model. Since there is temporal dependence in the data, we can’t just take a random subset of data and predict it using a model that was trained on a different subset of data. Instead, we must use a rolling window. In our case, we forecast 26 weeks ahead using the 60 previous data points and calculate the average squared error between our 26 predictions and the actual data. This is done many times by sliding the training and prediction window across the data, hence the “rolling window.” The rolling window ASE for the AR(7) model is 38,854,927.



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 985227 10087071 25114626 38854927 41154279 135932238

## [1] 38854927

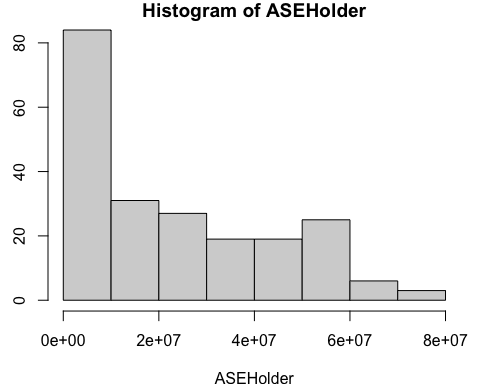
Again, we use a rolling window to calculate the average squared error. The ASE of the ARIMA(6,1,1) model is 79,473,892



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1267416 18498123 55851817 79473892 127360616 344827737

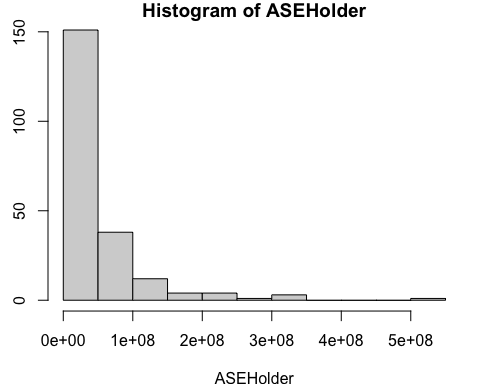
## [1] 79473892

Again, we use a rolling window to calculate the average squared error. The ASE of the ARIMA(7,0,0) with s = 52 model is 22,720,711.



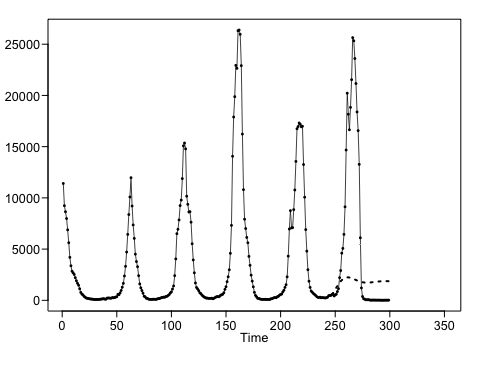
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 66698 4965516 19192360 22720711 38484593 74327599

## [1] 22720711

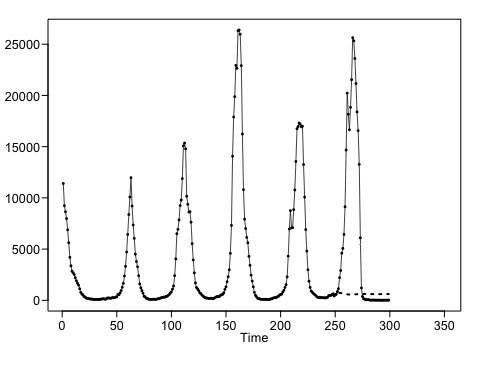
Again, we use a rolling window to calculate the average squared error. The ASE of the ARIMA(6,1,1) with s = 52 model is 47,934,334. 

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 65982 11192869 33736013 47934334 54896583 502193525

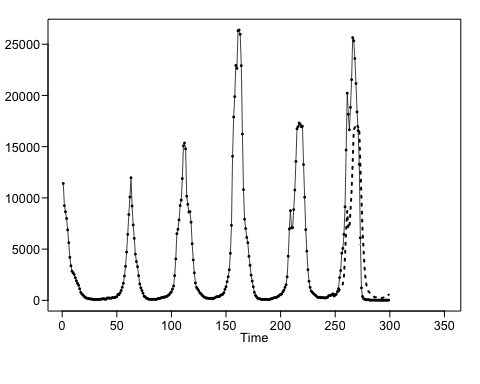
## [1] 47934334

Below are the predictions for 52 weeks of flu data using AR(7). The predictions are very flat and do not capture the flu season spike. 

## [1] 695.6282 900.2809 1082.3049 1237.5945 1394.9775 1549.3236 1711.9913  
## [8] 1854.6409 1968.3824 2059.5851 2129.5073 2185.9875 2224.0048 2241.3441  
## [15] 2240.8113 2225.0471 2199.5321 2165.8102 2125.0170 2079.2867 2030.6928  
## [22] 1982.2626 1935.7506 1892.2395 1852.6693 1817.7106 1788.2140 1764.5730  
## [29] 1746.7834 1734.6131 1727.6324 1725.4321 1727.4862 1733.1548 1741.7383  
## [36] 1752.5046 1764.7851 1777.9591 1791.4539 1804.7561 1817.4153 1829.0741  
## [43] 1839.4638 1848.3974 1855.7631 1861.5136 1865.6655 1868.2908 1869.5064  
## [50] 1869.4625 1868.3311 1866.2989

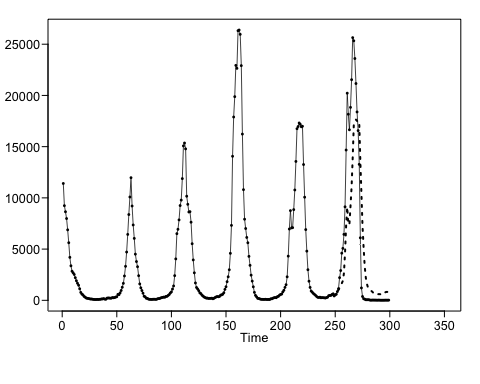
Below are the predictions for 52 weeks of flu data using ARIMA(6,1,1). The predictions are flat again and do not capture the flu season spike. 

## [1] 621.7635 700.4756 734.2257 727.4804 728.2956 723.5923 724.5342 714.4800  
## [9] 687.3236 658.4254 631.6189 613.0982 599.2828 585.1589 572.8867 564.2427  
## [17] 561.7175 564.4892 569.5128 575.3830 581.6434 588.8089 596.7998 604.4875  
## [25] 610.8822 615.5093 618.5987 620.5552 621.4697 621.2804 619.9952 617.8810  
## [33] 615.3603 612.7682 610.2822 607.9763 605.9330 604.2763 603.1062 602.4441  
## [41] 602.2312 602.3773 602.8049 603.4518 604.2523 605.1261 605.9893 606.7744  
## [49] 607.4406 607.9680 608.3474 608.5754

Below are the predictions for 52 weeks of flu data using ARIMA(7,0,0) with s = 52. These predictions have the same seasonal trend as the previous year, and certainly look more accurate. This model predicts a spike during flu season and a low amount of cases during the offseason, which is much better than our previous two models. 

## [1] 557.3051 702.0819 717.9231 784.0950 828.7345 830.3661  
## [7] 1019.8662 1143.0202 1378.8221 1602.6988 2336.6202 4326.9308  
## [13] 6947.7345 8718.1281 7011.1216 7017.7644 8769.3399 10683.1511  
## [19] 13467.7862 16662.9085 16876.4806 17233.1577 17107.8858 16878.6852  
## [25] 16940.2449 13214.8981 10037.6163 6884.9812 4792.5597 2987.1457  
## [31] 1872.4906 1272.6063 922.4598 809.0246 691.3516 604.4921  
## [37] 511.5900 399.7922 317.2431 300.0989 323.5015 263.6065  
## [43] 279.5583 262.4840 254.4954 226.6809 268.1151 300.8550  
## [49] 445.9385 471.3870 469.2043 554.3818

Below are the predictions for 52 weeks of flu data using ARIMA(6,1,1) with s = 52. This is very similar to the ARIMA(7,0,0) with s = 52 model, but doesn’t get as low during the flat flu period.



## [1] 575.7059 744.3152 799.2017 897.0758 968.1218 1004.9840  
## [7] 1227.8061 1388.5993 1654.5786 1900.3458 2656.3740 4666.4603  
## [13] 7308.0617 9095.4173 7399.5603 7414.9187 9172.8785 11092.7352  
## [19] 13881.5599 17077.5407 17289.5084 17642.5059 17512.7423 17278.2076  
## [25] 17333.0695 13599.8829 10413.9220 7252.5588 5151.7667 3338.3594  
## [31] 2216.1932 1609.4033 1253.2920 1135.0771 1013.8429 924.5903  
## [37] 830.3597 718.2598 636.3942 620.8050 646.4662 589.3295  
## [43] 608.3646 594.5561 589.8817 565.2981 609.7509 645.1946  
## [49] 792.6012 819.9557 819.2467 905.4628

## Conclusion

|  |  |  |
| --- | --- | --- |
| Model | ASE | Seasonal Trend? |
| AR(7) | 38,854,927 | NO |
| ARIMA(6,1,1) | 79,473,892 | NO |
| ARIMA(7,0,0) with s=52 | 22,720,711 | YES |
| ARIMA(6,1,1) with s=52 | 47,934,334 | YES |

We propose using our best ARIMA model for Flu data: the ARIMA(7,0,0) with s=52. This model had the best rolling window ASE when we use 60 data points to predict 26 weeks in the future. This model also had seasonal trend that the AR(7) and ARIMA(6,1,1) models lacked. This trend is crucial to predicting long term flu trends and we expect to see a similar spike in subsequent years. This model will need to be retrained in future years to maintain relevance.

We would recommend that in the future we should test predicting flu data using other time series models such as VAR and LSTM. The LSTM model allows for encoding of more complex time series behavior which we believe will improve the long-term forecast. The VAR model allows other variables to be included in the model such as average weekly temperature, population density, and other relevant factors.