Analyzing Google Flu Trends
Capstone Report

Mary Nelson
Faculty Mentors: Dr. Steven Chiou, Dr. Kang James
04/01/15
1. Introduction

The flu virus is prevalent all over the world. Different countries have different flu seasons when the flu virus is most common and it usually depends on climate zones. Outbreaks typically occur in fall or winter seasons when it is colder, and tropical climates typically have less pronounced flu seasons.

Information on flu counts and flu trends around the world is available online through Google Trends [1]. This information is found from analyzing and counting the number of Google searches from users looking up information on the flu. For example, if someone Googles “list of flu symptoms” or something related, this data would be collected. While not everyone who searches the flu on Google actually has the flu, this is a good indication to when flu seasons may be.

The information collected by Google regarding the number of flu related searches has been compared to information from US Centers for Disease Control and it turns out that they are very closely correlated which is shown in figure 1. Analysis of actual flu data from health agencies such as the CDC is very important, however most health agencies focus on a particular country or region and update slower, while Google Flu Trends provides data for a myriad of countries and is more frequently, providing a complement to existing systems from health agencies [1]. Additional information from sources such as Google could be very useful in predicting future epidemics and keep information on the flu as up to date as possible.

Figure 1:

There are two curves, blue is the Google Flu Trends, and orange is the US data http://www.google.org/flutrends/intl/en_us/about/how.html
2. Analytics

2.1 Preliminary Study

Google Trends started collecting flu related search data back in 2003. With data from the last 14 weeks in 2003 to the first 27 weeks in 2014, there were many flu season occurrences throughout the time period to use in my analysis. As you can see in figure 1, most of the flu outbreaks occur at the end of each year, or in the winter months. This pattern is consistent, except for in year 2009, which is apparent in figure 1. In 2009, there was a H1N1 pandemic that occurred in the months of April and May [2]. I considered this year an outlier and excluded it from my analysis and model formation. In addition, I performed log transformations to reduce the range of flu counts. In the original data set, flu counts range from about 1000 to 9000, which is larger than preferred for this analysis.

2.2 Simple Regression

I started my analysis by importing the weekly US flu search data from Google Flu Trends into R. The first thing I decided to do was to use regression analysis to find a model that fit the data accurately. I then needed to clean and organize the data by year so that each year could be graphed on the same axis as seen below in figure 2. Just by looking at it without testing any models, there seems to be a pattern of higher flu counts in the early and late weeks of each year compared to the weeks in the middle of the year like in figure 1. This makes sense because most flu seasons occur during the winter months.
After taking out data from 2009 and taking a log transformation to normalize the data, I was able to pick the best model of 15 that I tested. I tested a variety of different models from simple linear models, sin and cosine models. I came to the conclusion that the best model with the highest $R^2$ value was:

$$\log(\text{count}) = -1.280e-01 t + 2.310e-03 t^2 + 8.521$$

$\text{Std Errors: } t: 3.8e-03 \ t^2: 6.9e-05 \ \text{intercept:} 4.4e-02$

Where $t=$week

The dependent variable is flu counts during a particular week, and independent variables are a particular week as well as the square of that week. The standard errors listed support that the beta coefficients are not equal to zero. This model generates an $R^2$ value of 0.6905, and a very small p-value of less than 2.2e-16. With such a small p-value, I can conclude that this model is significant and a reasonable for the data, as seen in figure 2.

- Predicted week 28, 2014: 852 (Actual Google flu count: 1028)
- Predicted week 29, 2014: 856 (Actual Google flu count: 1008)
To reemphasize that 2009 is an outlier year, I plotted it against the other data and the model, as seen in figure 3. The data from 2009 is the blue line, and obviously doesn’t fit very well with the rest of the model and is outside of the 95% confidence interval. The yellow dots on the graph demonstrate how big the residual can be between the predicted values and the actual values for 2009.

**Figure 3:**

Log Transformation of Flu Counts by Week 2003-2014 (Without 2009)

---

**2.3 Time Series Analysis**

While using simple regression produces a reasonable model, after reading about time series analysis, I thought that an even better model could be found for this data set. This data seemed to be a good fit for time series analysis because the data has been consistently measured at equally spaced intervals, which is weekly from 2003 to 2014. Time series data is a sequence of successively measured data points made over a time period, whether it be days, weeks, months, etc. There are different types of time series data, but this flu data seems to be seasonal. Time Series lets us forecast future behavior based on how it has behaved in the past.

There are two models I looked at when analyzing this data, ARMA and ARIMA.
ARMA stands for autoregressive moving average model. The autoregressive part of this model represents a type of random process, or a time-varying process in nature, which our flu data is. This part also gives us the theory that individual values in time can be described by previous observations. The moving average part of this model means that the mean is continuously reevaluated as new data becomes available. It uses a certain period of time and it progresses by dropping the earliest value and adding the latest. We use two values in this model. P is the order of autoregressive, or the number of previous data points our model is dependent on. Q is the order of moving average, or the number of previous data points used to calculate our moving average.

The second type of model is ARIMA, or autoregressive integrated moving average model. This model is similar to the ARMA model except for we are adding the integrated part. This integrated part is used when trend filtering is required, or differentiation. If our process is non-stationary, or parameters of the cycle change over time, differentiation is needed until our process becomes stationary, and fluctuates around a constant value. The extra value in this process, d, is the number of times we differentiate our data to make it stationary. So differentiation is subtracting on value from the previous value.

In order to determine properties of a time series, we use correlograms. An ACF or autocorrelation function measures the correlation, or similarity, between observations at different distances apart. This helps us identify a preliminary model for time series. An ACF correlogram helps us see if our series is stationary or not, depending on if the ACF plot converges to zero given a confidence interval at a 5% significance level. The second plot we use in time series data analyses is PACF or partial autocorrelation function. This helps us identify the extent of the lag in an autoregressive function, or the order. Where this function converges to zero, given a confidence interval at a 5% significance level, is the number of lags we choose.

The first thing I looked at was the ACF plot (autocorrelation function) and PACF plot (partial autocorrelation function) to identify the parameters for ARIMA model. I saw that the ACF plot didn’t converge to zero, meaning that it is a non-stationary series, as
seen in figure 4. In order to fix this problem and make it stationary, differencing needed to be done. Differencing the data once wasn’t enough to make the data stationary, so I took it a second time. Then we would be modeling the “change in the changes” of the original data.

**Figure 4:**

![ACF Plot for Counts](image1.png) ![PACF Plot for Counts](image2.png)

The new ACF plot after differencing looks much different as seen in Figure 5. In converges to 0 rather quickly, after about 6 lags. This data has a lag of 6 meaning that the flu counts one week will depend on the flu counts from 6 weeks before.
Once a stationary process was obtained, I was able to fit an ARIMA model to the data. I tried three different ARIMA models with lag 6 and differencing of 2, but different moving averages of 0, 1, and 2. I found that ARIMA with lag 6, differencing of 2, and moving average of 0 was the most accurate because it had the lowest AIC value. With this model, we are able to forecast future flu counts.

A common goal in time series is to extrapolate past behavior into the future. Figure 6 shows this forecast for the next two weeks along with the 95% confidence intervals. The first forecasted week predicts flu counts to be about 954 (actual Google count: 1028) and the second week to be about 955 (actual Google count: 1008). I do not have permission to access CDC flu data, otherwise I would compare these numbers to actual flu counts.
3. Summary of Results

To summarize my results, the regression model forecasted flu counts that produced residuals of 176 and 152 while the time series model forecasted flu counts that produced residuals of 74 and 53. From this we can see that the time series model is the better model. These results make sense because when you think about it, flu outbreaks happen because it gets spread from people who already have it. Thus the flu counts of today or this week depend on people from previous weeks spreading the flu. Time series data is dependent on the past and so is this data.

4. Future Research

If I were to continue researching this topic, I would explore multiple linear regression with Google Flu Counts. An interesting twist on this subject would be to look into trends that may be related to flu seasons such as tissue and flu medicine purchasing activities, Doctor visits, weather patterns and temperatures, etc. I would need to obtain rights to these sources of data to explore this option. In my regression analysis course I learned skills to analyze multiple variables, which this type of research would require such skills. In future classes or research opportunities, this is something I would be interested in pursuing.
5. Educational Experience

Working on this project with Dr. Steven Chiou and Dr. Kang James enhanced my educational experience here at UMD in more ways than one. I learned how to plan and execute a long-term research project, which required excellent time management skills and perseverance. I learned how to use R software and know that skill will be very useful in future projects and in my future career. Also, in addition to strengthening my general regression analysis skills, I was able to become familiar with time-series analysis and apply it to my project in R. I believe creating models to predict outbreaks of not only the flu, but other viruses can be very beneficial in reducing the impact of the viruses on populations. Learning how to apply my skills to real life issues such as flu outbreaks is eye opening and rewarding. I am very thankful for all the guidance from Dr. Kang James and Dr. Steven Chiou and I enjoyed working with them on this project. The overall experience of this project was extremely educational and I am very thankful for the experience.

6. References and Sources to Aid in Project
