Time Series Analysis of Influenza-like-Illness Data

Introduction

The prevalence of Influenza-like-illnesses (ILI) exhibits an inexplicably seasonal trend, driven by cyclical spikes. In real life, professionals also encourage vaccinations during the "flu season", times when the number of ILI cases are high. Interestingly, although humans have been dealing with the flu since 1918, and still suffer from approximately 36,000 deaths per year in the United States due to flu-like-illnesses [1], we do not have a very concrete understanding of the mechanisms behind the virus' seasonality. This report will characterize features of seasonal ILI fluctuations; we will also observe correlations between ILI case prevalence with other climatic, geographical, and social features, in order to understand potential factors that drive the disease' seasonality.

Results and Discussion

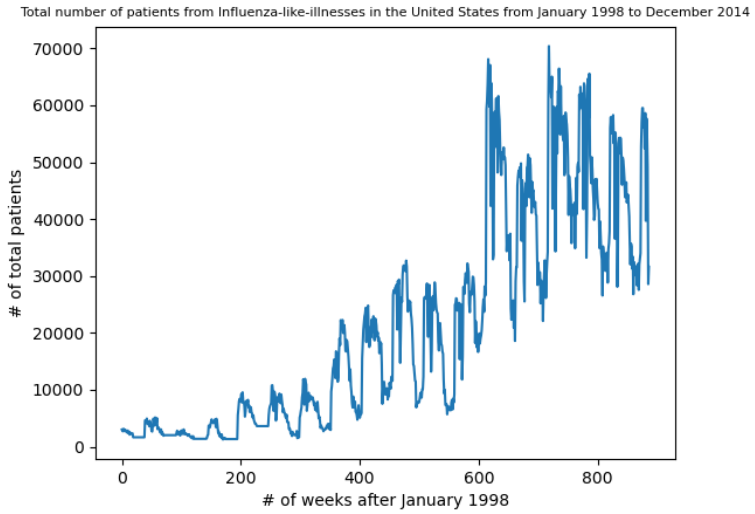

Figure 1: Total number of patients from Influenza-like-illnesses in the US between 1998 and 2014.

In Figure 1, we plot the total number of ILI patients in the US as a time series over an interval of 16 years [2]. Immediately from observing the raw data, we can see oscillations that seem to maintain the same frequency.

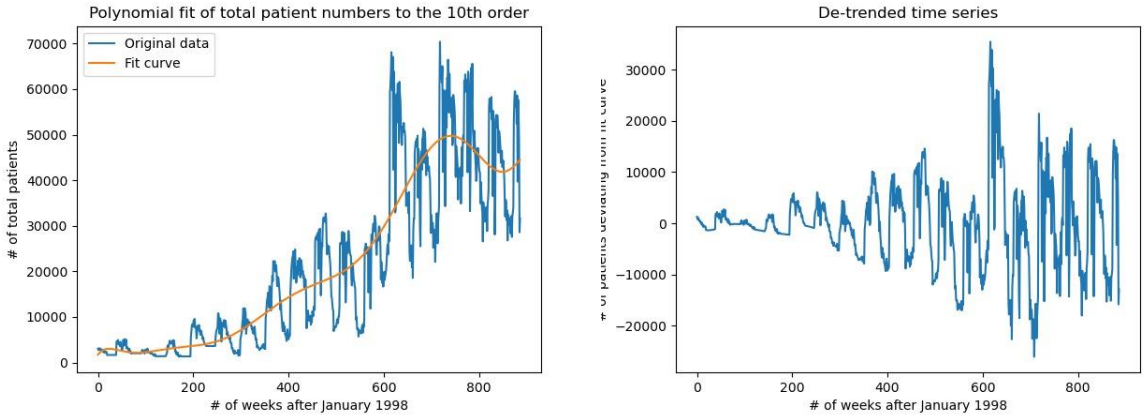

Figure 2: a) 10th order polynomial fit of the total number of ILI patients. b) De-trended total number of ILI patients.

Before conducting any Fourier analysis on the time series, we must clean up the data by de-trending it. The general, long-term trend of case numbers follows a non-linear trajectory due to the complex variety of factors that influence disease prevalence. Consequently, we need to use an appropriate high order polynomial fit curve to capture the long-term trend without overfitting into the annual fluctuation; in this case, the 10th order is appropriate. The de-trended data (Figure 2b) illustrates the annual fluctuations as periodic oscillations.

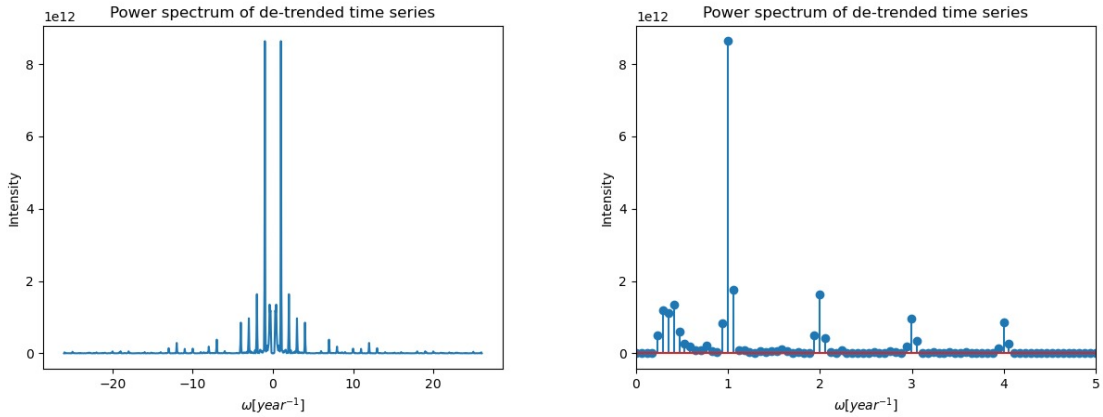

Figure 3: a) Full power spectrum of de-trended time series for total number of ILI patients. b) Power spectrum zoomed in at $\omega = 0 \sim 5$ [year^{-1}]

Results of the Fourier analysis on the de-trended time series show a very large peak at $\omega = 1$ [year^{-1}] of the power spectrum, suggesting that the time series oscillates at a frequency of 1 year^{-1}. This is consistent with intuition, that the seasonality of the flu is indeed annual, which is now identified by our analysis.

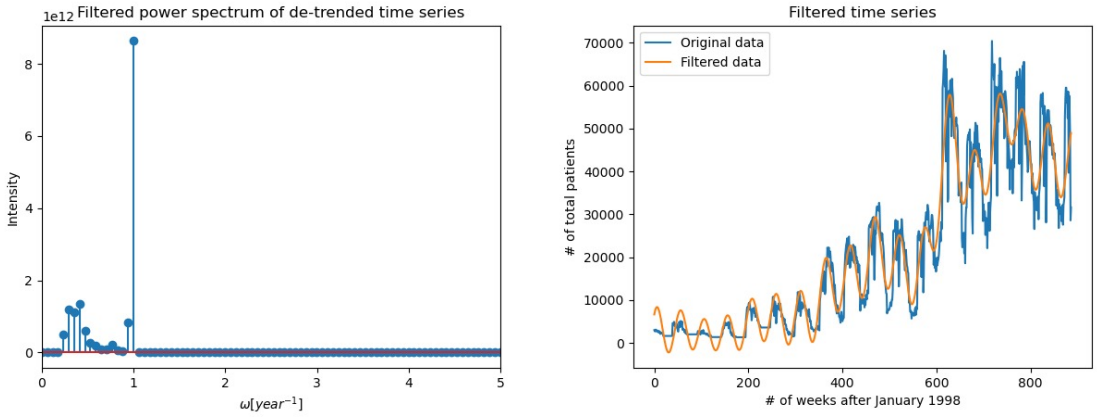

Figure 4: a) Power spectrum of the time series after filtering out all oscillations with ω > 1 [year^{-1}]. b) Filtered time-space data with recovered trend.

We inferred that the annual oscillations are characterized by the frequency ω = 1 [year^{-1}]. Thus, filtering the frequency-space spectrum by converting any value above ω = 1 [year^{-1}] to zero effectively removes high-frequency, low-intensity oscillations that exist as "noise" in the time-space. These high-frequency oscillations are not important for identifying seasonal oscillations, which reside in much lower frequencies. Figure 4b depicts the filtered time series after reverse Fourier transformation; it models the annual oscillations of long-term number of ILI patients.

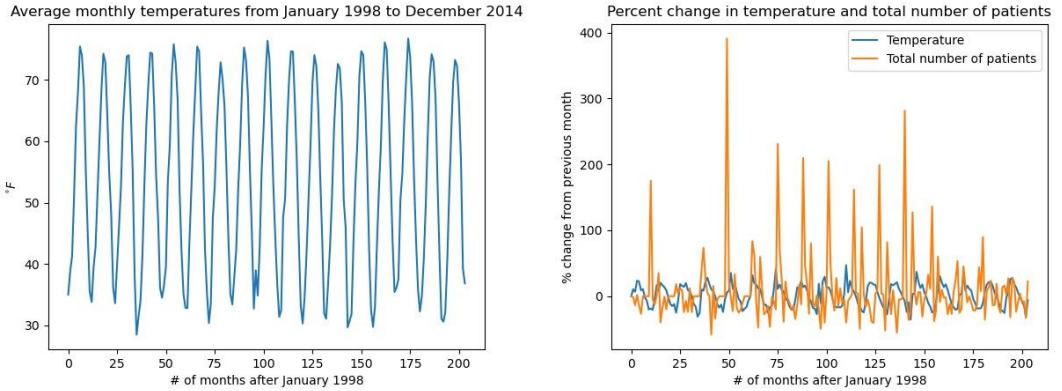

Figure 5: a) Average monthly temperatures in the United States from 1998 to 2014. b) Percent change in temperature and percent change in number of patients from their perspective, previous data points.

To investigate potential causes of the annual seasonality of ILI, we can obtain the similarity between patient data and temperature data. Figure 5a shows the average temperature in the US, which, from observing the raw data, exhibits roughly an annual frequency of 1 year^{-1}. In figure 5b, when the time series for percent change both in temperature and in case numbers are overlapped, we can see that peaks in temperature changes have roughly the same frequency as

peaks in the changes in patient numbers, with the temperature peaks offsetting to the right by approximately 6 months.

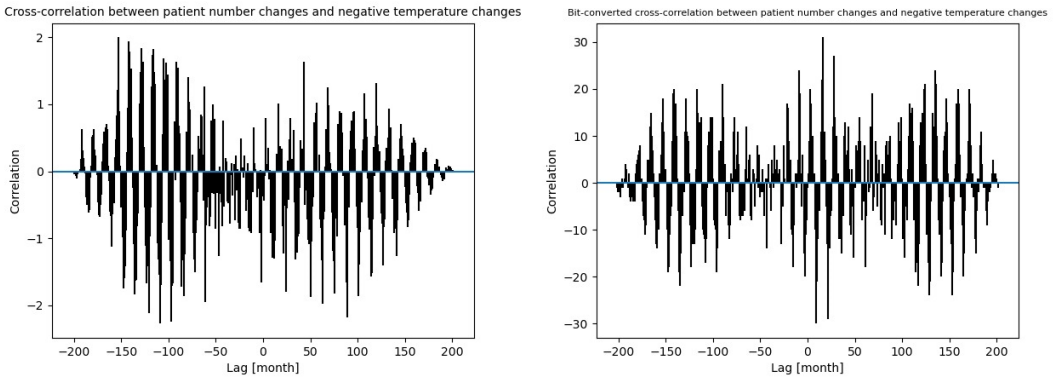

Figure 6: a) Cross-correlation between percent changes in ILI case numbers and negative percent changes in temperature. b) Bit-converted cross-correlation between percent changes in ILI case numbers and negative percent changes in temperature.

Cross-correlating the percent change time series in Figure 6a shows that negative average temperature changes correlate with ILI case number changes, which is exhibited by the peaks in the cross-correlation between the two time series. Importantly, there is a peak at lag = 0 [month], suggesting that the negative temperature changes directly correlate with case number changes, without shifting either time series. Bit-converting the time series before cross-correlating in Figure 6b shows an even larger magnitude of correlation, illustrating further that the changes happen in a relatively synchronous fashion. From this data, we can conclude that temperature is a potential cause of the seasonality of ILI cases.

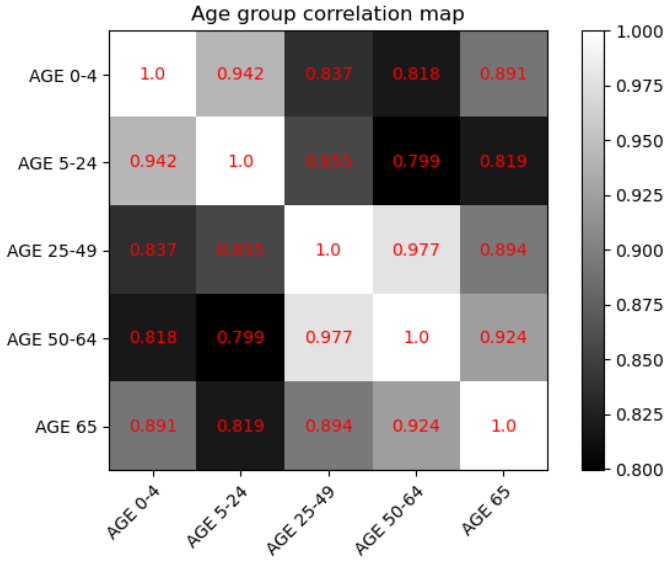

Figure 7: Matrix of normalized correlations between the case number time series of different age groups in the United States; x and y axes indicate the age groups correlated.

Furthermore, we examine the correlation between the total case number time series of different age groups. Figure 7 shows a matrix of every age group correlated with every other age group, with their normalized correlations shown as the brightness and number of each square. From this matrix, we see that age groups correlate better with neighboring age groups than age groups with larger age gaps. Obviously, every age group compared to itself has a normalized correlation of 1. If one observes the squares beside any white, diagonal, 1.0 correlation squares, the color of the squares immediately beside them are much brighter than ones further away. In fact, there is a roughly decreasing brightness further away from the 1.0 correlation diagonal. This indicates that age groups play a factor in the trend of disease prevalence.

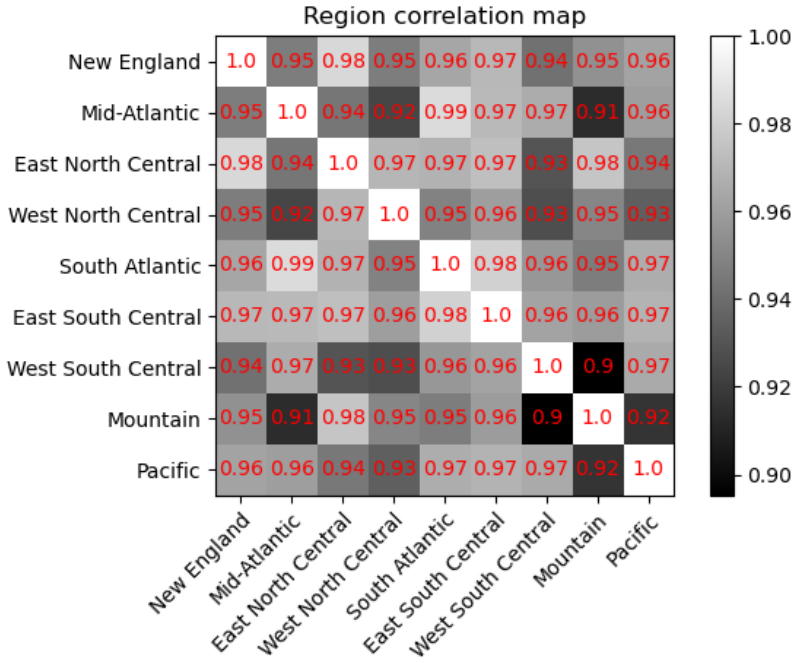

Figure 8: Matrix of normalized correlations between the case number time series of different regions in the United States; x and y axes indicate the regions correlated.

We then do the same analysis between different regions in the United States in Figure 8. This result is particularly interesting, since there is not a consistent trend of higher correlations between regions that are geographically closer or climatically similar. Instead, we see instances of low correlations between regions that have large differences in humidity. For instance, the Mountain region correlates the least with the Pacific region and the Mid-Atlantic region, regions where humidity differ the most. These results are consistent with literature [3], which suggests that humidity has an effect on the likelihood of infections.

Conclusion

Fourier analysis is a powerful tool that gives insights to periodic trends in data. This became especially effective in analyzing ILI, whose time series exhibits a very strong periodicity. In addition, cross-correlation data allowed us to compare similarities between two time series, which provides us with candidate factors that cause particular trends. An important

caveat to note, however, is that correlation does not necessitate causation. Although we have shown that temperature correlates nicely with the patient numbers, it is unclear whether temperature is the reason that the seasonality of disease prevalence exists.

References:

1. Krause, Edward, "Surveillance for Influenza-like-Illness in the United States from 1997-2014", Harvard Dataverse, V1 (2015)
2. Lofgren, Eric, et al. "Influenza seasonality: underlying causes and modeling theories." Journal of virology 81.11 (2007): 5429-5436.
3. Lowen, Anice C., et al. "Influenza virus transmission is dependent on relative humidity and temperature." PLoS Pathog 3.10 (2007): e151.