REPLACE WITH COVER PAGE

SMU®

Table of Contents

Introduction

Every year, influenza viruses pose remarkable impacts on socio-economy, such as costs of medical care, loss of productivity, and deaths. Since economic considerations are essential for influenza control, decision makers often need to examine following questions for health interventions:

- How much will an influenza season cost a states/provinces or country?
- Which states/provinces or counties bear high costs?
- Where to distribute vaccines to achieve the maximum returns?

Influenza, commonly known as "the flu", is an infectious disease caused by an influenza virus. Three of the four types of influenza viruses affect humans: Type A, Type B, and Type C. Type D has not been known to infect humans, but is believed to have the potential to do so. Usually, the virus is spread through the air from coughs or sneezes. This is believed to occur mostly over relatively short distances.

Influenza spreads around the world in yearly outbreaks, resulting in about three to five million cases of severe illness. In the northern and southern parts of the world, outbreaks occur mainly in the winter, while around the equator, outbreaks may occur at any time of the year.

Larger outbreaks known as pandemics are less frequent. In the 20th century, three influenza pandemics occurred: Spanish influenza in 1918 (17–100 million deaths), Asian influenza in 1957 (two million deaths), and Hong Kong influenza in 1968 (one million deaths). The World Health Organization declared an outbreak of a new type of influenza A/H1N1 to be a pandemic in June 2009.

To date, only a small number of studies have estimated the economic impacts of influenza. The Office of Technology Assessment reported that the influenza accounts for $1 ~ 3 billion per year in medical costs. Meltzer, et al. argued that the annual economic burden of pandemic influenza could range from $71.3 ~ 166.5 billion. The latest estimation by Molinari et al. indicated that the short-term costs and long-term burden of seasonal influenza can be amounted to $26.8 ~ $87.1 billion a year. These studies have established systematic methods to analyze influenza economics and offered valuable guidance for interventions.

While most are interested in the medical and socio-economy costs of the flu, the Dacy Pharmacy Supply Chain is taking a different approach. Given that a new distribution center will be coming on-line in India, the organization needs predictable data to determine the the “just in time, just enough” level of electrolyte drinks that should be available to consumers. Given the spike of use in the drink from the US, the assumption is that all countries experience the same. Shipping liquids to India is an expensive proposition. The special instance of customs declaration and weight restrictions require the supply chain to be optimal to save inventory sitting in an expensive warehouse.

Method

TBD. (More to come using the ARIMA and ASE forecast graphs?)

”depending on criteria. Endpoint ASE performs the best at 1, 3, and 5 years of data used, getting worse the more data is included. This is likely due to the model’s inability to adequately predict the large outbreaks approximately every 5 years. Rolling ASE performs the best at 4 years, while adding more years of data is likely due to a large number of NA’s or sparse data.”

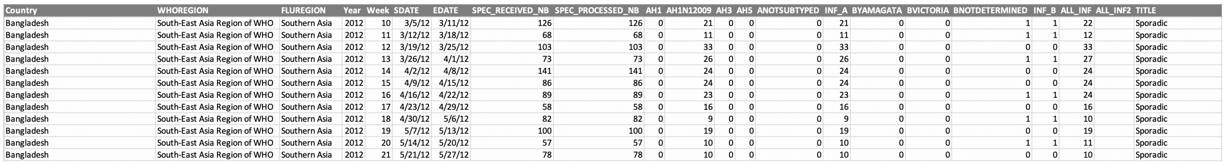
Data

Data Source

There are myriad datasources available with flu data and they can all be fairly specific. Given the nature of the flu strains and the seasonal aspect of the data, it is important to select a consistent and proven source. For this question of interest, the data selected will be provided by the World Health Organization (WHO). The data is categorized by region. The entire Southeast Asia region will be chosen given the expansion plans and distribution center reach The time frame select will be for 1995-2020.

The dataset can be organized by various categories such as Country (each country within the WHORegion (Southeast Asia)) and FluRegion which is sub-regions from WHHORegion. There are 11 countries within 3 sub-regions in the dataset.

Figure 1: Raw Data from the WHO



Country	WHOREGION	FLUREGION	Year	Week	SDATE	EDATE	SPEC_RECEIVED_NB	SPEC_PROCESSED_NB	AH1	AH1N12009	AH3	AH5	ANOTSUBTYPED	INF_A	BYAMAGATA	BVICTORIA	BNOTDETERMINED	INF_B	ALL_INF	ALL_INF2	TITLE
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	10	3/5/12	3/11/12	126	126	0	21	0	0	0	21	0	0	1	1	22		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	11	3/12/12	3/18/12	68	68	0	11	0	0	0	11	0	0	1	1	12		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	12	3/19/12	3/25/12	103	103	0	33	0	0	0	33	0	0	0	0	33		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	13	3/26/12	4/1/12	73	73	0	26	0	0	0	26	0	0	1	1	27		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	14	4/2/12	4/8/12	141	141	0	24	0	0	0	24	0	0	0	0	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	15	4/9/12	4/15/12	86	86	0	24	0	0	0	24	0	0	0	0	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	16	4/16/12	4/22/12	89	89	0	23	0	0	0	23	0	0	1	1	24		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	17	4/23/12	4/29/12	58	58	0	16	0	0	0	16	0	0	0	0	16		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	18	4/30/12	5/6/12	82	82	0	9	0	0	0	9	0	0	1	1	10		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	19	5/7/12	5/13/12	100	100	0	19	0	0	0	19	0	0	0	0	19		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	20	5/14/12	5/20/12	57	57	0	10	0	0	0	10	0	0	1	1	11		Sporadic
Bangladesh	South-East Asia Region of WHO	Southern Asia	2012	21	5/21/12	5/27/12	78	78	0	10	0	0	0	10	0	0	0	0	10		Sporadic

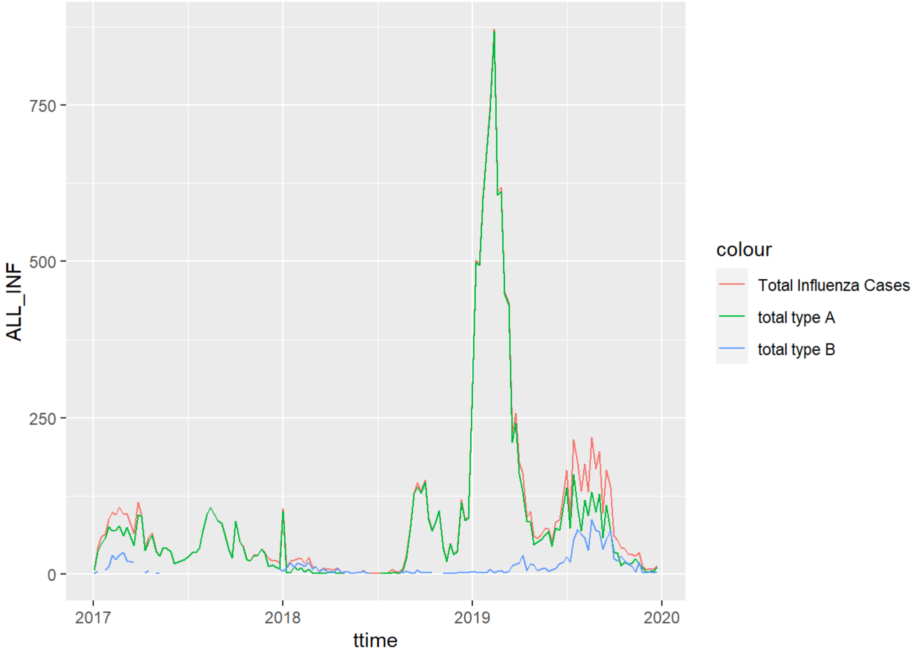
The World Health Organization (WHO) collects this data in weekly intervals over the course of 24.5 years. The information contains number of virus strands and individual cases. Personal demographics are left out for various reasons, but not necessarily require for the purposes of predicting supply needs. As the process matures, it might be valuable to foster a dataset that includes ages for special inventory for children and senior citizens. For unknown reasons, the years 1995-2010 are absent of the data required to predict the rate of flu in the future, therefore, the dataset was narrowed down to analyze the years 2017-2019 for the most common flu strands, A and B.

The data selected is not only important to predict cases by using the fields describing if someone had flu or not, but also the date/time it took place. Time series is a series of data points indexed (or listed or graphed) in time order and can adjust for seasonality of the data.

Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

All of these are important pieces of data separately, but most important is the time in which the measurement occurred since that can relatively affect the prediction.

Figure 2: Initial Visual of the Data

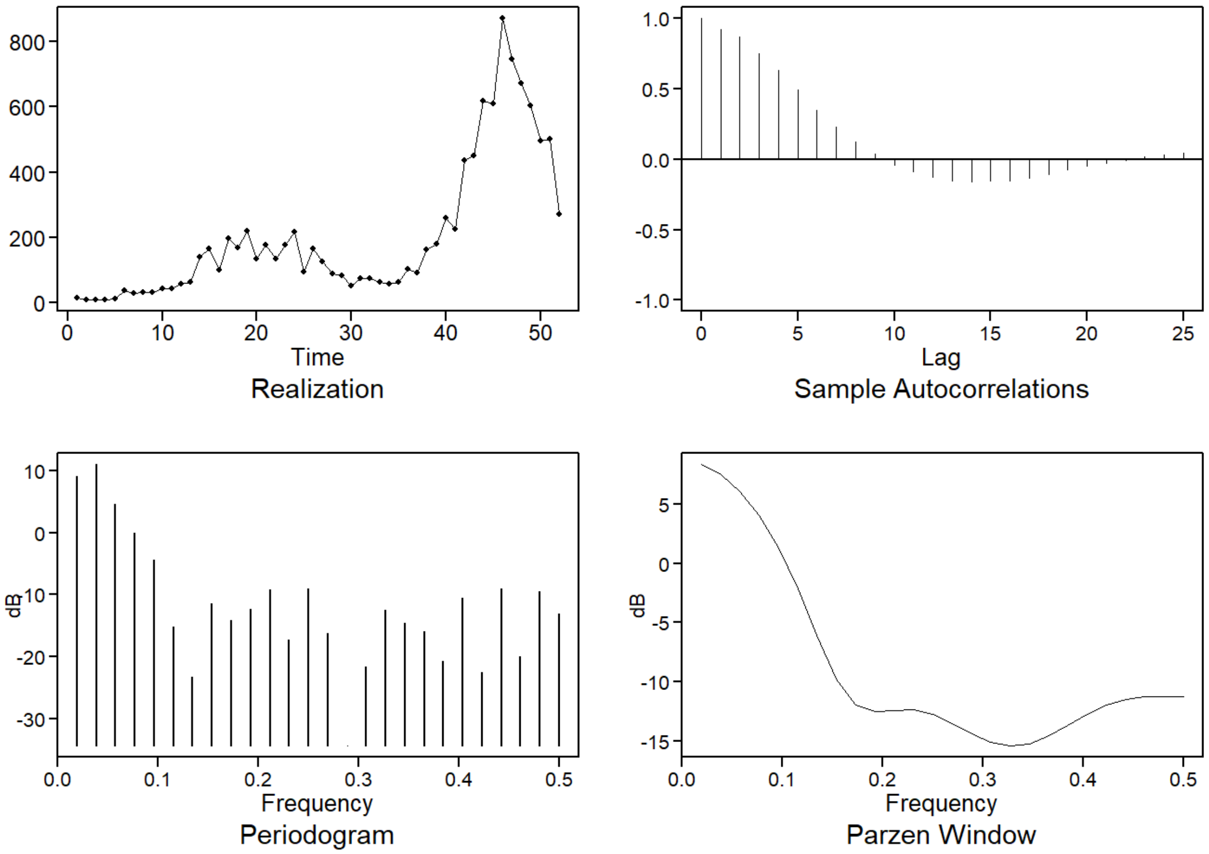


The clearest way to examine a regular time series manually is with a line chart such as the one above. The initial visualization of the data shows that the Type A and B Flu strains are the most common and that there was a huge spike in 2019 most likely attributed to the COVID-19 virus that was likely confused with Flu. This is why 2017-2019 will be the best data to use for a consistent time-series measurement.

Exploratory Data Analysis

Selecting the appropriate years to forecast is an important step in the data analysis. Various methods will be employed to help determine what the data says about the years.

Figure 3: Time Series Fit Graphs



Lag is essentially delay. Just as correlation shows how much two time series are similar, autocorrelation describes how similar the time series is with itself. Consider a discrete sequence of values, for lag 1. Compare the time series with a lagged time series. In other words, shift the time series by 1 before comparing it with itself. Proceed doing this for the entire length of time series by shifting it by 1 every time. An autocorrelation function is then born of the data which basically visualizes how much it correlates with itself.

For any time series, there will be a perfect correlation at lag/delay = 0, given the comparison of the same values with each other. As the time series shifts, the correlation values will begin decreasing. If time series comprises of completely random values, the only correlation will be at lag=0, and no correlation will exist anywhere else. In most of the time series datasets, this is not the case, as values tend to decrease over time, having some correlation at low lag values.

More to come through the PELT and BinSeg and MSE graphs?

Conclusion

TBD

Appendix

As with any technology; the system required to generate meaningful data, cleaning of the data, as well as the models that were implemented require various software packages. The raw “code” used to explore the methods detailed above can be found in the following libraries:

- GitHub Repository