Visualizing and Predicting Influenza Cases in New York State

A Master's Project
Presented to the
BIOMEDICAL AND HEALTH INFORMATICS
PROGRAM

In Partial Fulfillment
of the Requirements for the
MASTER OF SCIENCE DEGREE

State University of New York
at Oswego

By Daniel Truong

December 5th, 2019

Abstract

Influenza is a contagious and potentially lethal viral disease that affects a good amount of the U.S. population every year. Preventative treatment through periodic vaccination is instrumental in mitigating the deadly effect of the disease. Other factors that could affect the efficacy of treatment include climate, geolocation, and population density. As such, for researchers to get a better grasp on the effect that the aforementioned variables play, data must be collected periodically to make assertions. This project involves visualizing the occurrences of influenza cases in New York State from 2009 to the present. Also, a couple of prediction models were tested and evaluated on data from 2018 onwards. The result was that the prediction models detected the general trend of when the cases occurred, but not the exact magnitude. A project like this is designed to be adapted for other types of diseases or different municipalities.

Contents

Introduction

Influenza is a viral infection that predominantly affects the respiratory system. Colloquially referred to as “the flu”, symptoms of influenza include fever above 100 degrees Fahrenheit, aching muscles, headaches, nasal congestion, and persistent coughs. In most extreme cases, influenza can lead to more serious ailments such as pneumonia, heart problems, or even death.

There are four strains of the Influenza virus: types A, B, C, and D. In a study done by Martinez et al. (2019), types A & B are the most common infections to afflict the human body. Type C is said to be a more lethal strain, but less common compared to A & B. Type D is a newer strain, but rarely found to affect the human body. The type A strain is the most virulent; spread vectors usually include avian species (namely aquatic birds and poultry) and human beings. From a study by Su et al. (2017), some of history’s most devastating influenza outbreaks such as the H1N1 (swine flu) and H5N1 (bird flu) strains originated from the type A strain.

One of the ways that the effects of the influenza virus can be mitigated is with periodic vaccinations (usually done as a yearly “flu shot”). The influenza vaccine is a neutralized version of the influenza virus that gets introduced into the immune system of a human being. The immune system would become provoked by the introduction of the foreign virus and generate a biological response to protect the human body. From there, if the human body were to come into contact with a strain of the influenza virus, the immune system would be trained to be able to combat the virus (and protect the human body as such). The human body must be subjected to periodic vaccines due to the ever-changing behavior of the influenza strain in such a short time.

Environmental variables can also affect the spread of the influenza virus. Amongst more populous locations, the virus can easily spread from person to person. Places like schools, office buildings, buses, and other public congregations are often common threat vectors for influenza to spread. Extra care must be given to geolocational factors if a treatment plan were to be drafted to address an influenza epidemic.

Data Visualization

A study by Mirel and Görg (2014) was done to see if the visualization of disease mechanisms helped with the “sense-making” of complex disease topics amongst the general population. What they found (albeit amongst a small sample-size population) was that the understanding of such complex mechanisms came about from the supplementing of visual representations with background information that bolstered whatever point the researchers were aiming for. In short, for researchers (investigating disease mechanisms, epidemiology, or other higher-order topics) to make their point amongst the general population, not only was it important to empirically connect their evidence to their research, but it was also imperative to visually show that connection as well.

In medicine, epidemiology is the study of factors and variables that affect population health on a locational and temporal basis. For the most part, infectious diseases are studied for their effect on population health and spread vectors. Influenza would fall into this category. To get an adequate idea of how dangerous the influenza virus can spread and find ways to combat its spread, visualizing the geographical trends of influenza infections would certainly help researchers in gaining rapid insight into its spreading trends. A study done by Sakai et al. (2004) in Japan combined geographical evaluation with kriging analysis to not only track down the general infection cases originating in western Japan, but to also analyze seasonal trends to better target patients for preventative treatment. Another study, done by Wangara et al. (2019) in Kenya, found variances in the locational occurrences of leprosy and demonstrated a need to lock down public health policies in these areas. Having that visual association of diseases with geolocation will help researchers reach conclusions about their occurrences more coherently and rapidly.

Project Methodology

To help with visualizing and predicting influenza trends, a web-based application will be developed for this purpose. The objective of the app is as follows:

- Display the magnitude of influenza cases in NYS per specific periods.
- Generate a prediction model for influenza cases from 2018 to the Present.
- Use open-source technology to allow for adaptation by other municipalities.

When developing this project, the main actors that we're concerned with developing for common users and other developers. The project defines users as professionals who will benefit from the data visualizations and predictions by viewing the generated info (i.e. NYS general population, primary care physicians, pharmacies, and other researchers primarily focused on influenza cases in NYS). Other developers are defined as users who will take the existing app and adapt it for other locations or diseases (i.e. epidemiologists, health care agencies, universities, etc.). Developing the app as a web-based program will allow access for a greater range of clients, as well as ease of use.

Data Source

NYS' Department of Health operates a public health data repository through its Statewide Planning and Research Cooperative System (SPARCS). Researchers can get access to public use data such as (but not limited to) birth rates, locations of medical facilities, nursing home reviews, and chronic illnesses. The data that this project will be concerned with is the lab-confirmed cases of influenza (separated by county and type as far back as 2009).

When downloading the data from its source, there are a total of 59,868 records amongst 9 variables:

- **Season**: the yearly period for which the flu data was recorded

- **Region**: the geographical description of where the data was recorded (i.e. Central New York, Capital District, Western New York, Metro, etc.)
- **County**: the County of the recorded incidents
- **CDC.Week**: designation of the end-of-week (for the year) of recording
- **Week.Ending.Date**: the date that denotes the CDC end-of-week recording
- **Disease**: whether the disease is classified as Influenza Type A, B or Unspecified
- **Count**: the amount of lab-confirmed cases of Influenza for the specified date
- **County.Centroid**: geographic coordinates for the centroid of the county
- **FIPS**: unique geographic code assigned to each county and state

Per the description of the dataset from NYS, the number of cases reported is somewhat under-represented due to the following factors:

1. The cases are only represented if an ill patient seeks medical care.
2. A medical specimen must be collected for testing.
3. Test results are reported to the Electronic Clinical Lab Reporting System (ECLRS).
4. Test results come back positive.

As such, it can be easy to see why a breakdown of processes could contribute to an underreporting of influenza cases. In any case, the data in its present form must be cleaned up and munged further if it's to be of use for the project.

Prediction Algorithm

The data for the influenza cases will occur over a set period. Because we do not have any other confounding factors in the data set (that could affect flu cases), we will need to frame the prediction as a univariate analysis of time series data. This analysis will be done on a per-county basis. Two prediction algorithms will be evaluated for this purpose: ARIMA and Holt-Winters Exponential Smoothing. Because the data exhibits some seasonality (repetition over a set period), both models will be utilized for the prediction component of the project.

An ARIMA (Autoregressive integrated moving average) model is a type of regression that analyzes univariate time series data that shows some evidence of non-stationarity. An example of such a study is one by Koppelova and Jindrova (2019) where periodic cell phone signals were collected to determine trends in adoption rate by the general population. Another study by Wang and Liu (2017) analyzes the passenger flow of inner-city expressways to predict future trends and make urban planning decisions from those trends.

Contrasting ARIMA in this project is Triple Exponential Smoothing, also known as the Holt-Winters model. Holt-Winter Exponential Smoothing is primarily concerned with forecasting future values – making predicted observations on out-of-sample observations. Holt-Winters works well if the data exhibit seasonality. One example of this model in action was a study done by Brutlag (2000) at WebTV to see if aberrant behavior in network traffic could be automatically detected without needing constant human supervision. It was found that while implementing such a model could be robust, the work operations of implementing a software solution for this model were not quite optimal at the time of writing. Additionally, a study by Puah et al. (2016) used the Holt-Winters method to forecast seasonal rainfall patterns in Malaysia. This study was used to gauge the effects of climate change on natural disaster occurrences.

Data Cleaning and Preparation

To prepare the influenza data for modeling, the different influenza types were consolidated into one influenza check. Data from the 2009 - 2018 seasons were segregated out and used as training data for the prediction model. Each county was modeled separately from the other to better evaluate possible confounding factors afterward. For each year that the data was recorded, weeks 21 through 39 (roughly mid-May through late- September) were originally omitted from the source data set. Furthermore, data from the 2014 - 2015 season was missing for numerous counties. To properly train the forecasting models, dummy data were inserted to make up for the lack of observations. All dummy observations reflect zero occurrences of any influenza cases.

The programming language R (2018) was used for running the prediction models, in addition to cleaning the datasets with. R package **Forecast** (Hyndman and Khandakar, 2008) contains the function *auto.arima()* that will be used for training the forecasting algorithm. Meanwhile, the base R package has the *HoltWinters()* function that will be used for the exponential smoothing prediction. The data from the 2018-2019 and 2019-2020 season were used to test the efficacy of the prediction algorithms. In all, after cleaning the data, there were 29,078 observations in the training set and 3,720 observations in the test set.

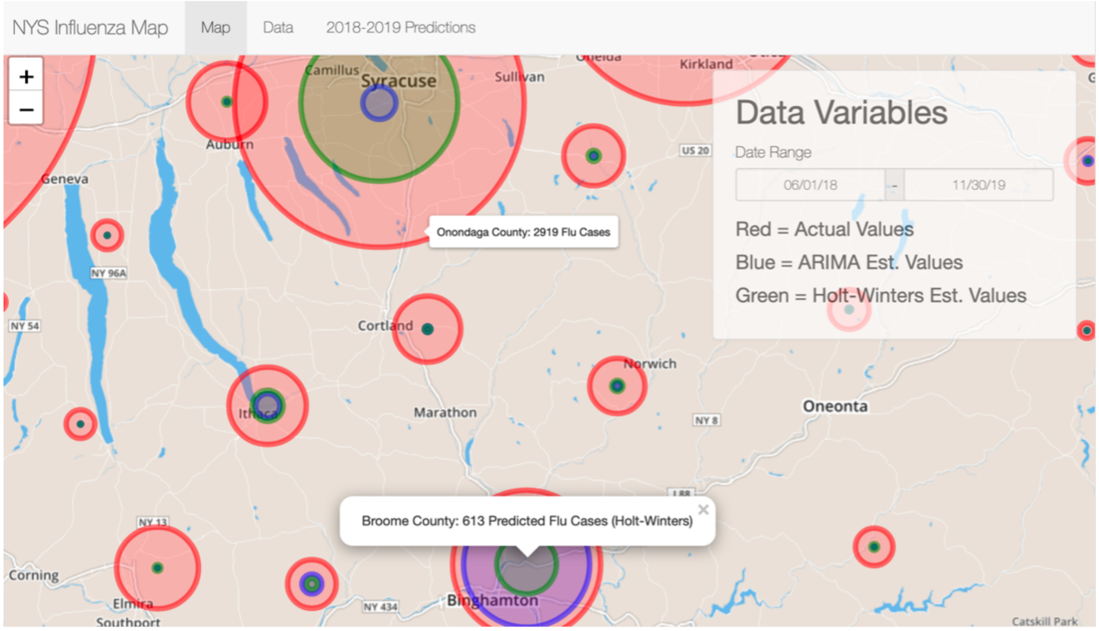
Software Implementation

To prepare the dataset for use with the interactive data map, R Studio was used to import the cleaned dataset from SPARCS. R Studio leverages the R programming language to allow researchers and developers to run statistics and transformations on datasets in a rapid manner.

The **Shiny** package (Chang et al., 2019) for R was used to construct the user interface for the interactive data map. Shiny leverages datasets (loaded in R Studio) to create HTML5-based applications for clients to interact with. Shiny will allow us to display a map of NYS and plot circles to represent the magnitude of illnesses for each county. The map component leverages the **Leaflet** package (Cheng et al., 2018) to render the map. The completed app can then be uploaded online to ShinyApps.io or hosted on a dedicated web server.

Finished Product

The app has three components to it: a map view, a data table, and a section with predictions for the 2018-Present flu season. The map view (Figure 1) shows a map of NYS with small red dots over each county. The radius of the dots is proportional to the number of flu cases reported per season and week (both variables are user-controllable via a dropdown box towards the side). The larger the dot, the more flu cases were reported for that county. If a user selects a date within the 2018-2019 season range, they will also see blue and green colored dots of varying sizes. The blue dots represent the predicted amount of flu cases from the ARIMA model and the green dots represent values predicted from the Holt-Winters model.



Onondaga County: 2919 Flu Cases

Data Variables

Date Range

06/01/18 - 11/30/19

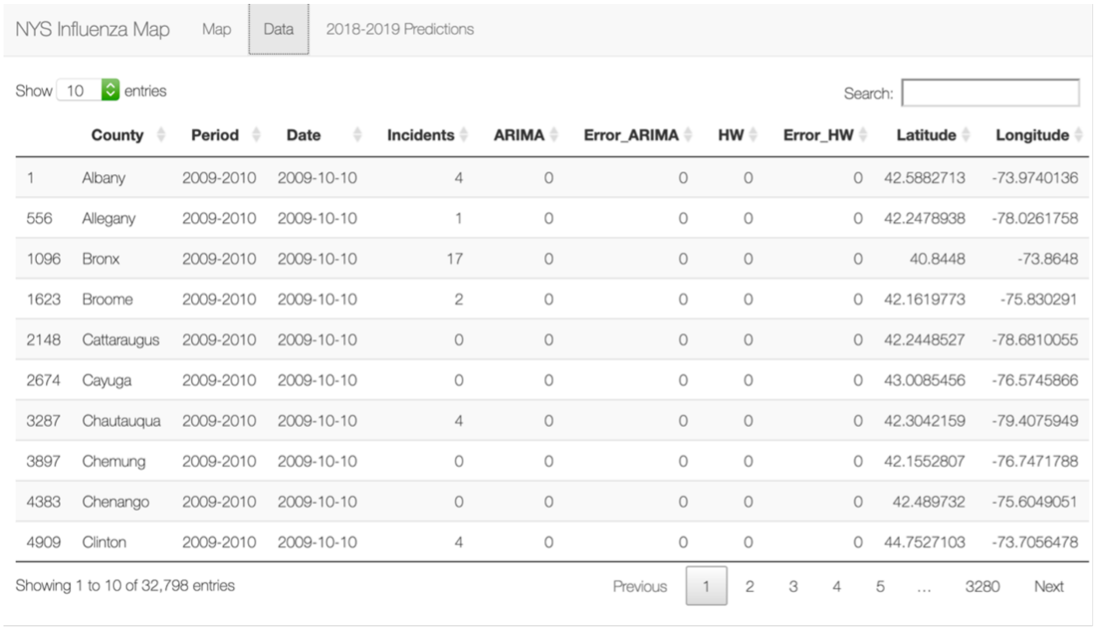
Red = Actual Values

Blue = ARIMA Est. Values

Green = Holt-Winters Est. Values

Broome County: 613 Predicted Flu Cases (Holt-Winters) ×

Figure 1: Map View showing flu cases (actual and predicted) in the Southern Tier.

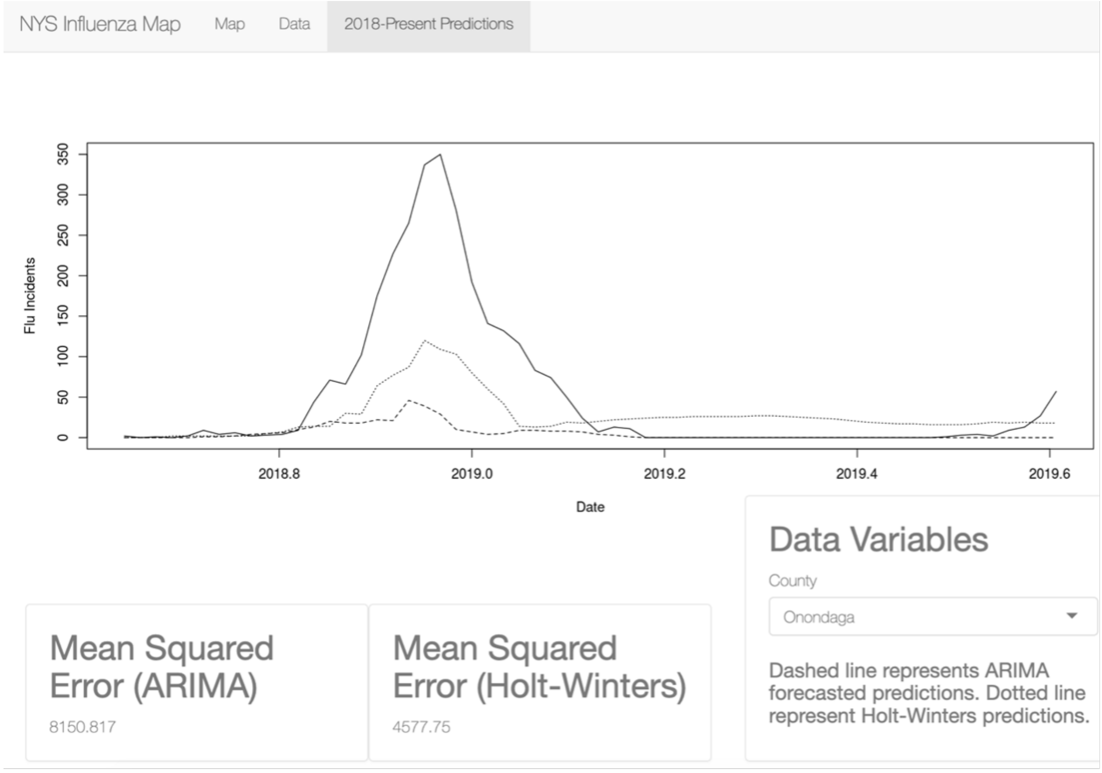


	County	Period	Date	Incidents	ARIMA	Error_ARIMA	HW	Error_HW	Latitude	Longitude
1	Albany	2009-2010	2009-10-10	4	0	0	0	0	42.5882713	-73.9740136
556	Allegany	2009-2010	2009-10-10	1	0	0	0	0	42.2478938	-78.0261758
1096	Bronx	2009-2010	2009-10-10	17	0	0	0	0	40.8448	-73.8648
1623	Broome	2009-2010	2009-10-10	2	0	0	0	0	42.1619773	-75.830291
2148	Cattaraugus	2009-2010	2009-10-10	0	0	0	0	0	42.2448527	-78.6810055
2674	Cayuga	2009-2010	2009-10-10	0	0	0	0	0	43.0085456	-76.5745866
3287	Chautauqua	2009-2010	2009-10-10	4	0	0	0	0	42.3042159	-79.4075949
3897	Chemung	2009-2010	2009-10-10	0	0	0	0	0	42.1552807	-76.7471788
4383	Chenango	2009-2010	2009-10-10	0	0	0	0	0	42.489732	-75.6049051
4909	Clinton	2009-2010	2009-10-10	4	0	0	0	0	44.7527103	-73.7056478

Showing 1 to 10 of 32,798 entries

Figure 2: Data Table

The data table (Figure 2) shows the cleaned-up data set for the user to explore. One can use the embedded search function to look for specific data on a county level, year, or season. Lastly, the prediction component (Figure 3) shows a time-series plot of actual observed flu cases vs. what the prediction algorithms have analyzed. The solid line represents the actual observations while the dashed and dotted lines represent the predicted observations (the ARIMA and Holt-Winters models respectively). The MSE (Mean squared error) for the forecasting model is also shown.



Mean Squared Error (ARIMA)

8150.817

Mean Squared Error (Holt-Winters)

4577.75

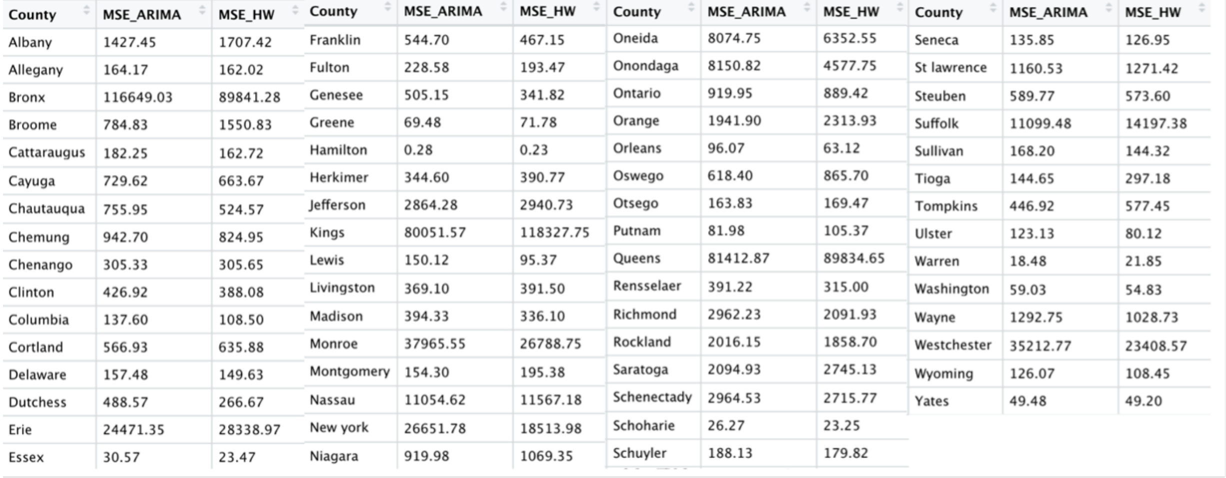
Data Variables

Dashed line represents ARIMA forecasted predictions. Dotted line represent Holt-Winters predictions.

Figure 3: App prediction component

Evaluation

When it comes to evaluating the accuracy of the prediction models in each county, we find mixed results. If we evaluate using MSE (see Figure 4 below), we find that some models work well for certain counties over one another.



County	MSE_ARIMA	MSE_HW
Albany	1427.45	1707.42
Allegany	164.17	162.02
Bronx	116649.03	89841.28
Broome	784.83	1550.83
Cattaraugus	182.25	162.72
Cayuga	729.62	663.67
Chautauqua	755.95	524.57
Chemung	942.70	824.95
Chenango	305.33	305.65
Clinton	426.92	388.08
Columbia	137.60	108.50
Cortland	566.93	635.88
Delaware	157.48	149.63
Dutchess	488.57	266.67
Erie	24471.35	28338.97
Essex	30.57	23.47
Franklin	544.70	467.15
Fulton	228.58	193.47
Genesee	505.15	341.82
Greene	69.48	71.78
Hamilton	0.28	0.23
Herkimer	344.60	390.77
Jefferson	2864.28	2940.73
Kings	80051.57	118327.75
Lewis	150.12	95.37
Livingston	369.10	391.50
Madison	394.33	336.10
Monroe	37965.55	26788.75
Montgomery	154.30	195.38
Nassau	11054.62	11567.18
New york	26651.78	18513.98
Niagara	919.98	1069.35
Oneida	8074.75	6352.55
Onondaga	8150.82	4577.75
Ontario	919.95	889.42
Orange	1941.90	2313.93
Orleans	96.07	63.12
Oswego	618.40	865.70
Otsego	163.83	169.47
Putnam	81.98	105.37
Queens	81412.87	89834.65
Rensselaer	391.22	315.00
Richmond	2962.23	2091.93
Rockland	2016.15	1858.70
Saratoga	2094.93	2745.13
Schenectady	2964.53	2715.77
Schoharie	26.27	23.25
Schuyler	188.13	179.82
Seneca	135.85	126.95
St lawrence	1160.53	1271.42
Steuben	589.77	573.60
Suffolk	11099.48	14197.38
Sullivan	168.20	144.32
Tioga	144.65	297.18
Tompkins	446.92	577.45
Ulster	123.13	80.12
Warren	18.48	21.85
Washington	59.03	54.83
Wayne	1292.75	1028.73
Westchester	35212.77	23408.57
Wyoming	126.07	108.45
Yates	49.48	49.20

Figure 4: Table of MSE values

Visually interpreting the results, we find that the predicted results generally reflect the temporal trends for when influenza occurs, but not the exact magnitude of the cases. Figures 5 and 6 show this in effect.

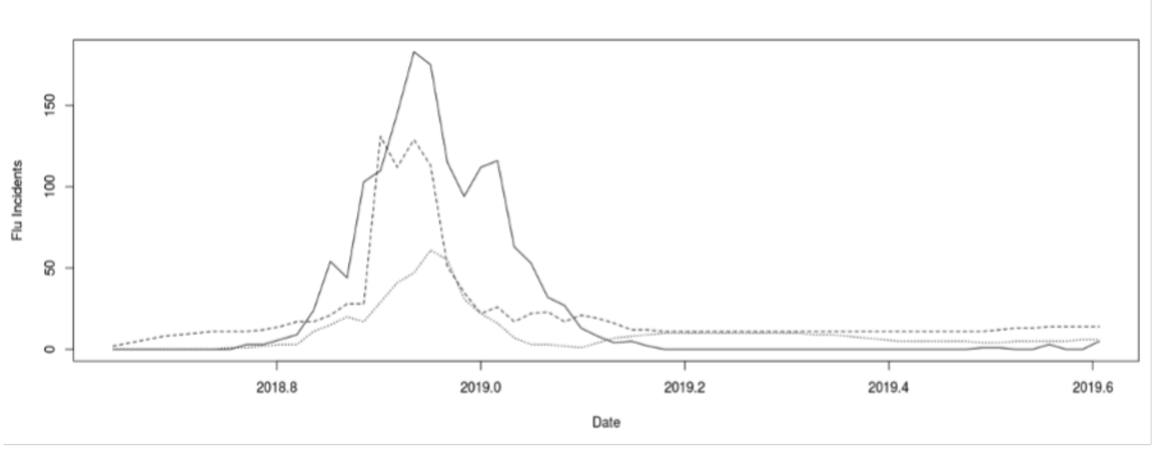


Figure 5: Predicted vs. Actual flu cases for 2018-Present for Broome County. Dashed lines represent the predicted cases.

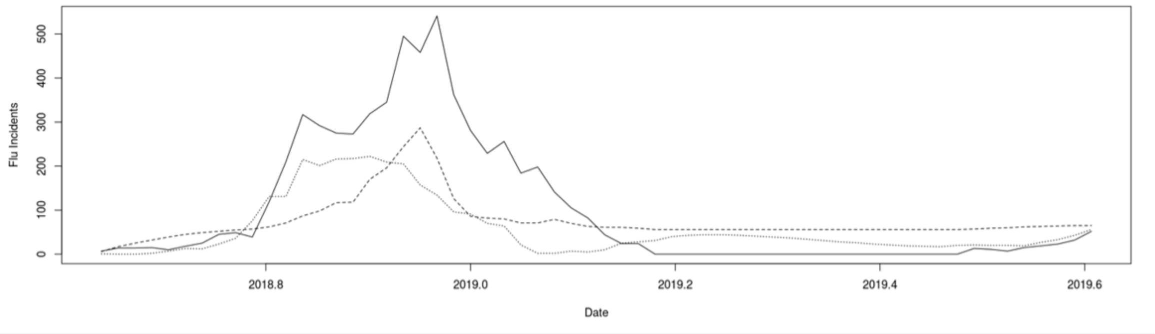


Figure 6: Predicted vs. Actual flu cases for 2018-Present for Nassau County.

Conclusion

The prediction models as-is tend to match well for seasonal trends in influenza data, but not for the volume of cases. This could be due to factors such as missing data affecting the models, population changes, socioeconomic demographics, or confounding factors (i.e. H1N1 influenza strain of 2009). If this app is to be adopted for widespread usage amongst researchers, additional prediction algorithms will need to be further explored; ideas to refine this component could entail gaining additional data from the U.S. Center for Disease Control or exploring the prediction as a function of a multivariate linear model. The user interface will also need additional work to be optimized on mobile interfaces; the spirit of this software package is to be running on a 7 or 10-inch mobile tablet for quick and easy evaluation amongst medical professionals. As it stands, the app is primarily optimized and tested on a desktop environment. If this app is to be utilized quickly and efficiently by fellow researchers, further testing on mobile environments will be necessary to ensure that the app meets ADA Standards for Accessible Design.

References

Brutlag, J. (2000). Aberrant behavior detection in time series for network monitoring.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). *Shiny: Web Application Framework for R*.

Cheng, J., Karambelkar, B., and Xie, Y. (2018). *Leaflet: Create Interactive Web Maps with the Javascript 'Leaflet' Library*.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 26(3):1–22.

Koppelova, J. and Jindrova, A. (2019). Application of exponential smoothing models and arima models in time series analysis from telco area. *Agris on-Line Papers in Economics and Informatics*, 11(3):73–84.

Martinez, A., Soldevila, N., Romero-Tamarit, A., Torner, N., Godoy, P., Rius, C., Jane, M., and Dominguez, A. (2019). Risk factors associated with severe outcomes in adult hospitalized patients according to influenza type and subtype. *PLOS One*, 14(1):1–15.

Mayo Clinic (2019). Influenza (flu). `https://www.mayoclinic.org/diseases-conditions/flu/symptoms-causes/syc-20351719`.

Mirel, B. and Görg, C. (2014). Scientists sense making when hypothesizing about disease mechanisms from expression data and their needs for visualization support. *BMC Bioinformatics*, 15(1):1–22.

New York State Department of Health - SPARCS (2009). Influenza laboratory - confirmed cases by county: Beginning 2009-10 season. `https://health.data.ny.gov/Health/Influenza-Laboratory-Confirmed-Cases-By-County-Beg/jr8b-6gh6`.

Puah, Y., Huang, Y., Chua, K., and Lee, T. (2016). River catchment rainfall series analysis using additive holt-winters method. *Journal of Earth System Science*, 125(2):269–283.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sakai, T., Suzuki, H., Sasaki, A., Saito, R., Tanabe, N., and Taniguchi, K. (2004). Geographic and temporal trends in influenzalike illness, japan, 1992-1999. *Emerging Infectious Diseases*, 10(10):1822–1826.

Su, S., Fu, X., Li, G., Kerlin, F., and Veit, M. (2017). Novel influenza d virus: Epidemiology, pathology, evolution and biological characteristics. *Virulence*, 8(8):1580–1591.

Wang, Y. and Liu, X. (2017). Seasonal passenger flow model of an inter-city expressway based on arima. *Advances in Transportation Studies*, 3:111–120.

Wangara, F., Kipruto, H., Ngesa, O., Kayima, J., Masini, E., Sitienei, J., and Ngari, F. (2019). The spatial epidemiology of leprosy in kenya: A retrospective study. *PLoS Neglected Tropical Diseases*, 13(4):1–11.