Forecasting Influenza Trends in Chicago: A Comprehensive Data Analysis Approach

Latha Sai Mounika Kona

George Mason University

AIT 580- Analytics: Big Data to Information

Dr.Alla G. Webb

May 8, 2023

Abstract

This research aims to understand the hidden trends in the influenza records of Chicago over the last decade. Forecasting influenza is crucial because it allows healthcare workers and doctors to allocate resources. Forecasting provides an opportunity for hospital management to prepare for the upcoming surge in patients. In this research, I developed the Autoregressive Moving Average Model for forecasting the upcoming ICU admissions because of the influenza infectious disease in Chicago. The data was collected from October 14, 2015 to February 26, 2023. This research also focuses on data analysis to understand how testing of patients for the potential risk of flu has changed over time which helps health officials to promote vaccination campaigns etc. Moreover, this research emphasizes the different types of Influenza variants over the years which helps identify changes in the prevalence of each variant and their effect on the public. The results show that positive influenza cases were high during the months December to February and the influenza subtype A is most predominately spread over the years with high-risk levels. The forecast also suggests that there will be an increase in ICU hospital admissions in the future. Therefore, it is important that hospital management or public health officials must take all the precautionary measures like increasing the number of vaccines based on the subtype that is more common over the years and educating the people about the consequences of the virus, the importance of taking the flu vaccines to avoid the risk.

Introduction

Influenza is a contagious respiratory illness that has outcomes ranging from mild to serious and sometimes causes fatal ICU admissions. The infection spreads when an infected person coughs or sneezes. Several influenza variants include A, B, H1N1, and H2N3. Moreover, this virus can mutate and form subgroups within the above-mentioned groups. In a highly dense city such as

Chicago, the spread of any contagious disease is highly likely and can spread at a faster pace. The dataset is obtained from the Chicago data portal (Influenza Surveillance Weekly | City of Chicago | Data Portal, n.d.), it is important to implement Influenza surveillance in a city such as Chicago because it can shed light on how the risk of influenza has changed over time. Influenza cases are studied in a wide range of environments and divided into specific cases based on the type and condition of the patient. The current H1N1 influenza virus is not considered as a novel influenza virus for surveillance purposes. Suspected novel variants have symptoms of severe respiratory illness of unknown etiology related to international travel (Influenza - HAN - Chicago Health Alert Network, n.d.). Influenza A or B including H3N2 and H1N1 can be tested positive on PCR tests which are reported not immediately but within 24 hours. Because Chicago has one of the busiest international airports, with many people visiting Chicago and traveling via Chicago there are higher chances of new variants spreading across the suburbs and other cities. This can cause an increase in the number of new Influenza cases. For this purpose, this study used the influenza data of Chicago to investigate and forecast the upcoming surge in new cumulative ICU cases of Chicago. This paper is set to investigate the following research questions:

- Monthly Time series of total influenza-positive cases and total samples tested over the years.
- Risk levels over the years and the subtype with high-risk level.
- Comparison of the proportion of positive cases of different variants (A, B, H1N1, H3N2) of influenza over the years.
- The number of hospitalizations in the ICU due to influenza over the years.
- Forecasting cumulative ICU admissions.

Literature Review

Influenza is a contagious disease that spreads from infected individuals to others via coughing, sneezing and affects 9.3 % of children every year (CDC 2019). It commonly occurs in temperate regions with higher activity during the winter seasons (Siddhivinayak et al., 2016). Influenza A and B are the most prevalent types of variants of influenza that spread at a fast pace.

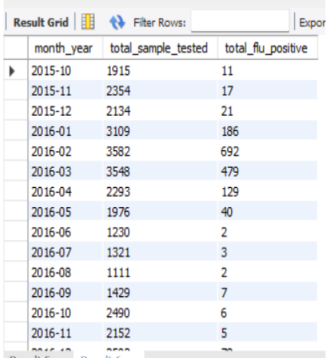
Influenza further includes subgroups H1N1 and H3N2 (Petrova & Russell, 2018). The role of the pandemic surveillance system cannot be ignored, in 2009 Illinois was affected by the influenza pandemic during the spring and fall season. The Illinois Department of public health initiated a surveillance system to track influenza spread by collecting data from hospitals, research units etc. This allowed Illinois to prepare well for the upcoming surge and understand influenza (Soyemi et al., 2014). The weekly influenza surveillance contains information on the week's start, and end date, hospital ICU admissions, Total Flu tested, total positive, etc. (Pandemic Influenza A (H1N1) in Chicago, 2009). In 2022, in a study, time series techniques were used to determine the severity of influenza positive cases in Bangladesh. In this case study, the annual seasonal pattern was identified (Berry et al., 2022). Similarly, in a study (Chen et al., 2020), the Seasonal Autoregressive Integrated Moving Average model was developed on influenza data from Shenyang China. And forecasted the upcoming new cases. The developed model includes both seasonal and Arima terms. The normalized BIC for the SARIMA(0,1,0)(0,1,2) with seasonal term m = 12 was used for urban areas and (1,1,1)(1,1,0) m= 12 was used for rural area dataset.

Materials and Methods

The dataset is obtained from the Chicago Data portal (Influenza Surveillance Weekly | City of Chicago | Data Portal, n.d.) which is publicly available. It contains features week start, weekend, flu risk level, hospital flu ICU weekly, and lab flu tested. Total positive etc. These features represent, the start of the week, the weekend for the data collection, the level of the flu risk during that specific week, weekly hospital admissions, the number of flu tests conducted in the week, the number of flu tests that are positive in that week, total fraction of tests that are positive. The dataset contains 388 data points, where each row represents the set of observations collected on a specific week. In the initial step of the data analysis, a total number of missing values were identified to answer the forecast research question. I used Python and libraries such as pandas, NumPy matplotlib, and pmdarima. The cumulative weekly data was created using the cumsum() function and created a new feature "icu_cumulative_weekly" and the frequency of the data was set to weekly before using it for forecasting. The trend analysis was performed using the decomposition library from statsmodel.tsa.seasonal_decompose. Once the trend analysis was performed, the auto_arima model was used to optimize the ARIMA model parameters such as p,d,q (Selva Prabhakaran, 2019). The other research questions regarding total influenza positive and total samples tested, comparison of the variant proportion over the time and number of ICU hospitalizations were addressed using R programming language. Similarly, the monthly time series data was aggregated/created using SQL.

Results and Discussion

- Monthly Time series of total influenza-positive cases and total samples tested over the years.



month_year	total_sample_tested	total_flu_positive
2015-10	1915	11
2015-11	2354	17
2015-12	2134	21
2016-01	3109	186
2016-02	3582	692
2016-03	3548	479
2016-04	2293	129
2016-05	1976	40
2016-06	1230	2
2016-07	1321	3
2016-08	1111	2
2016-09	1429	7
2016-10	2490	6
2016-11	2152	5

In the table above the total sample tested, total flu positive cases were obtained based on the month for all the years.

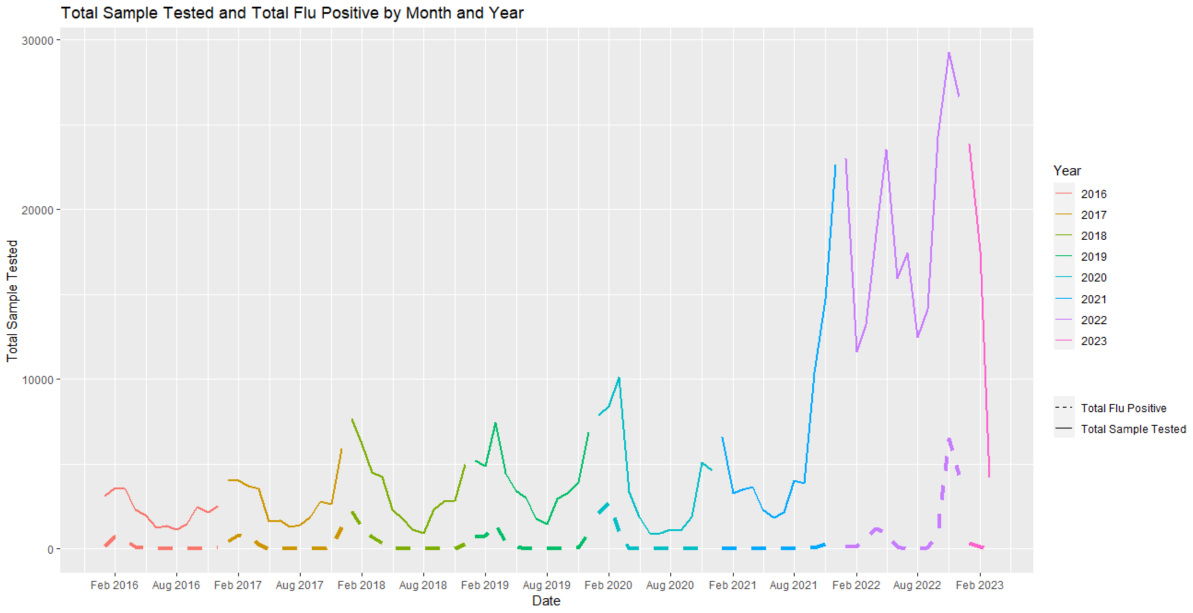
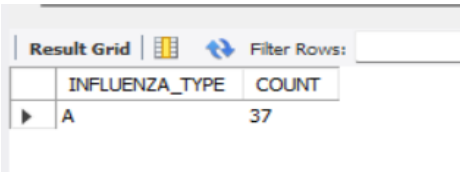


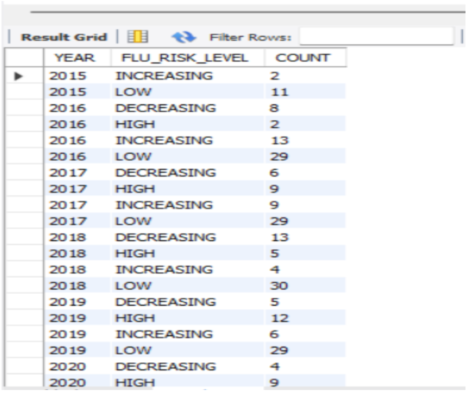
Figure 1 Samples tested vs total flu positive results per month and year.

From figure 1, it is evident that there is a tremendous increase in the number of samples tested between Aug 2021 and Aug 2022. Similarly, during this period there are a lot of tests that have positive results. Moreover, the sample testing has a cyclical pattern where there are more tests conducted during the winter months than the summer months.

- Risk levels over the years and the subtype with high-risk level.



INFLUENZA_TYPE	COUNT
A	37



YEAR	FLU_RISK_LEVEL	COUNT
2015	INCREASING	2
2015	LOW	11
2016	DECREASING	8
2016	HIGH	2
2016	INCREASING	13
2016	LOW	29
2017	DECREASING	6
2017	HIGH	9
2017	INCREASING	9
2017	LOW	29
2018	DECREASING	13
2018	HIGH	5
2018	INCREASING	4
2018	LOW	30
2019	DECREASING	5
2019	HIGH	12
2019	INCREASING	6
2019	LOW	29
2020	DECREASING	4
2020	HIGH	9

From the above tables, we can see that influenza subtype A has a high-risk level with a count of 37. In the second table count of each risk level was obtained for each year, 2015 year has an increasing count of 2, a low of count 11, 2020 has a decrease count of 4, and a high with a count of 9, etc.

- Comparison of the proportion of positive cases of different variants (A, B, H1N1, H3N2) of influenza over the years.

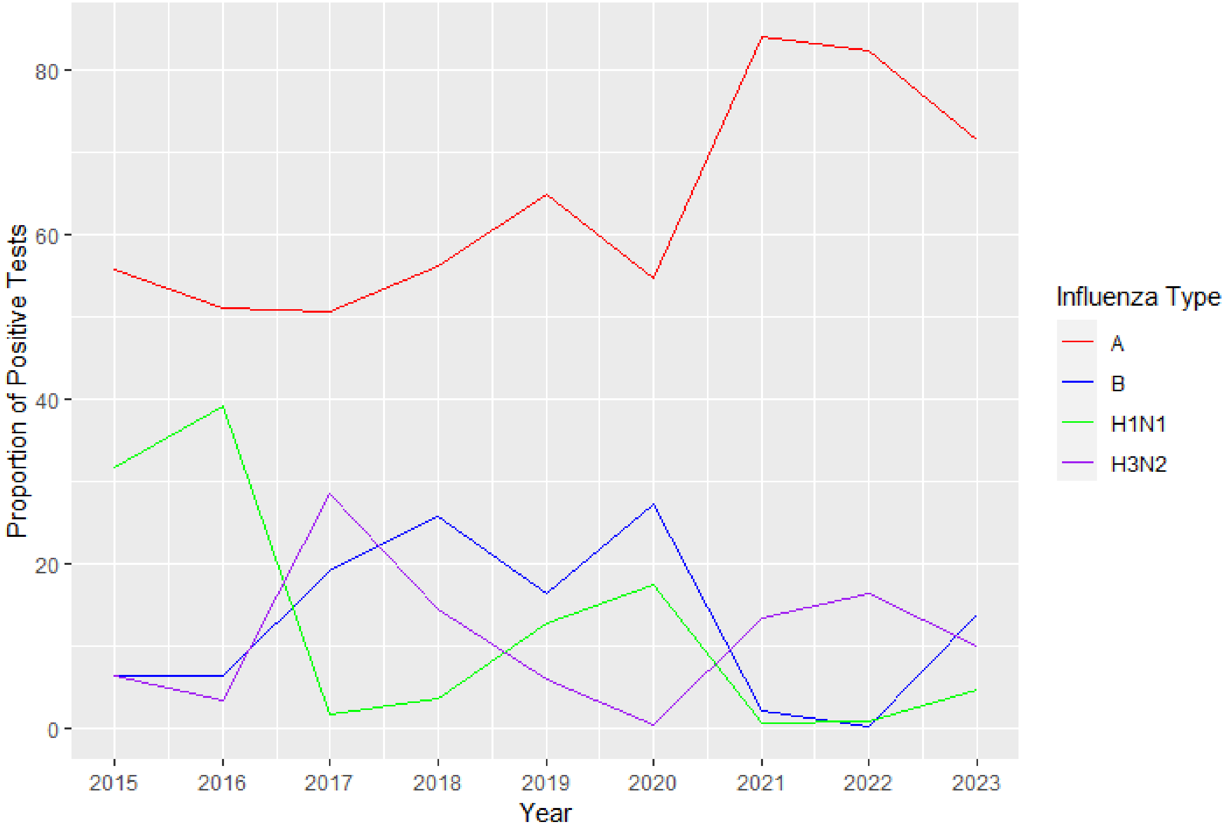


Figure 2 Trends of different types of Influenza from 2015-2023.

In the next step, the trend in the proportion of each type of influenza was plotted yearly. Over the period between 2015 and 2023, type A influenza-positive cases were prevalent in Chicago. The proportion of positive cases of type A influenza in the year 2015 is above 50%, the next highest proportion for 2015 was H1N1 type influenza with approximately 32%. From 2015 to 2023, the proportion of influenza type A increased tremendously. The proportion has increased to more than 80% in 2021 and gradually decreased to approximately 72% in 2023. Whereas the proportion of the type H3N3 influenza variant has gradually decreased from 2017 (approximately 30%) to 2020 and increased to 15% in 2022.

- The number of hospitalizations in the ICU due to influenza over the years.

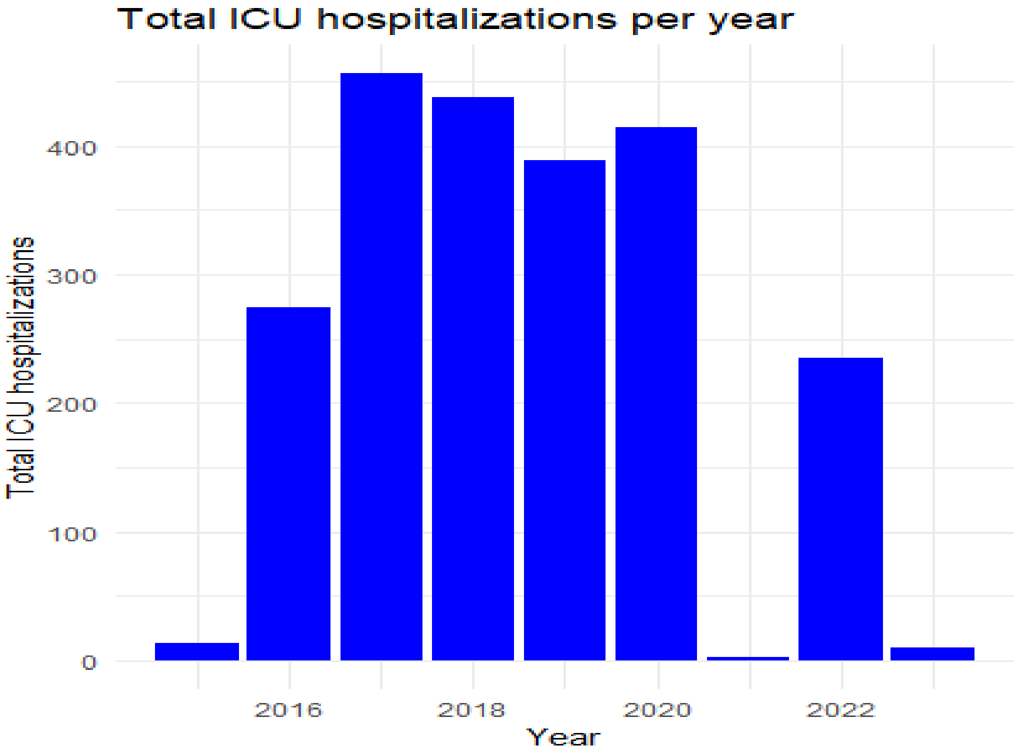


Figure 3 Number of individuals admitted into ICU due to severe Influenza.

As part of the exploratory data analysis, the number of ICU hospitalizations per year was plotted using R programming. From Figure 3 it is evident that in the year 2021, there are a smaller number of individuals who were admitted into ICU. Moreover, the mean ICU admissions value is 245 and the maximum ICU admissions is 456 which occurred in the year 2017.

- Forecasting cumulative ICU admissions.

In the next step, the Cumulative ICU weekly admissions were created and forecasted for a period of 30 weeks ahead using the Autoregressive integrated moving average model. The time series analysis was performed to analyze the trend and existing seasonality components in the data. From Figure 4, it is evident that the trend component represents the tendency of the dataset by capturing the underlying trend which is not affected by the cyclical fluctuations. From figure 4 it is evident that the trend of the cumulative ICU weekly admissions dataset has an exponential growth curve underneath the dataset from 2016 to mid-2020, after that the weekly ICU admissions seem to have reached a plateau.

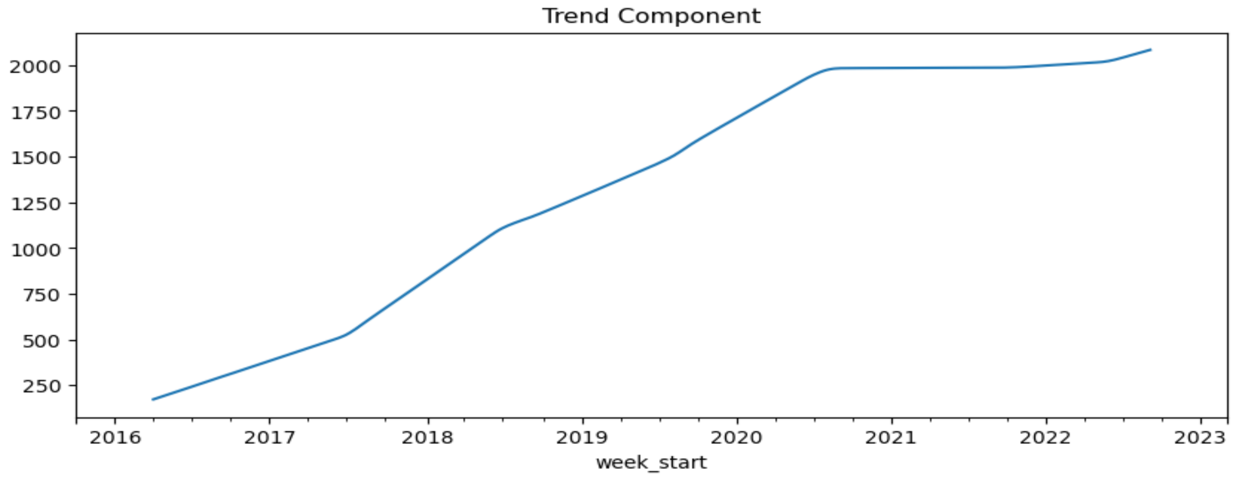


Figure 4 Trend component of the cumulative weekly ICU admissions in Chicago

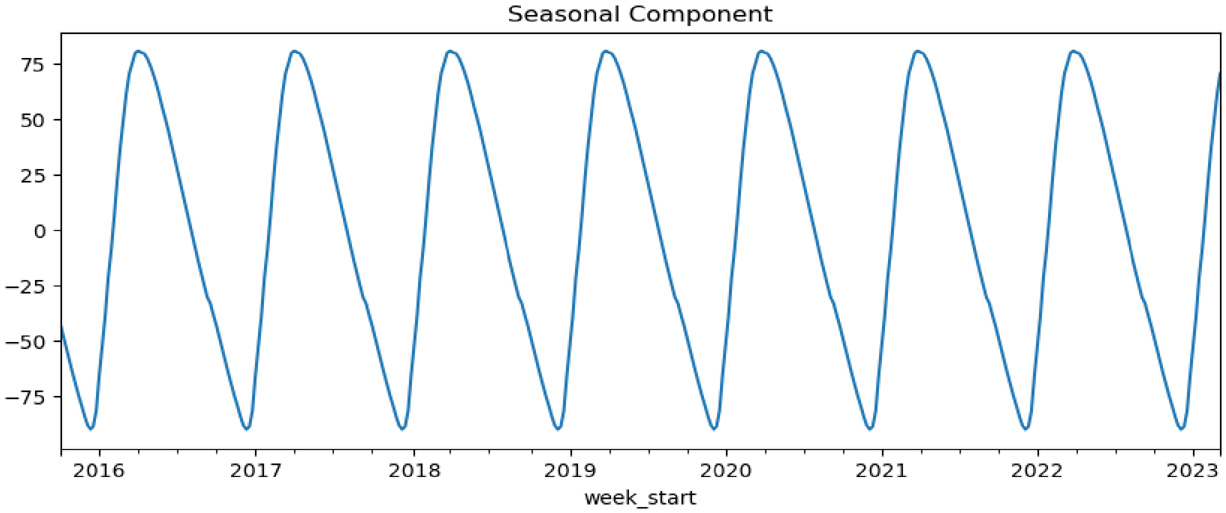


Figure 5. The seasonal component of the cumulative weekly ICU admissions in Chicago

The seasonal component provides information on the seasonal pattern and cycles in the dataset. In the Chicago time series dataset, the seasonality component is clearly visible with a recurrent yearly pattern. This annual pattern is evident from Figure 5, every year in the first and second quarters there are more Influenza positive cases than in the later quarters of the year. This pattern is consistent throughout the period of study which is from 2015 to 2023. Similarly, the residual component is also investigated for any undefined hidden patterns that are not captured by the seasonal and trend components of the time series data. From figure 6, it is evident that the residuals are randomly distributed from 2015 to 2023. But in some years, such as 2018, 2020, 2021, and 2022, the residuals are significantly different from zero and this means that some of the data is not being explained by the seasonal and trend components and some external factor is influencing the dataset pattern.

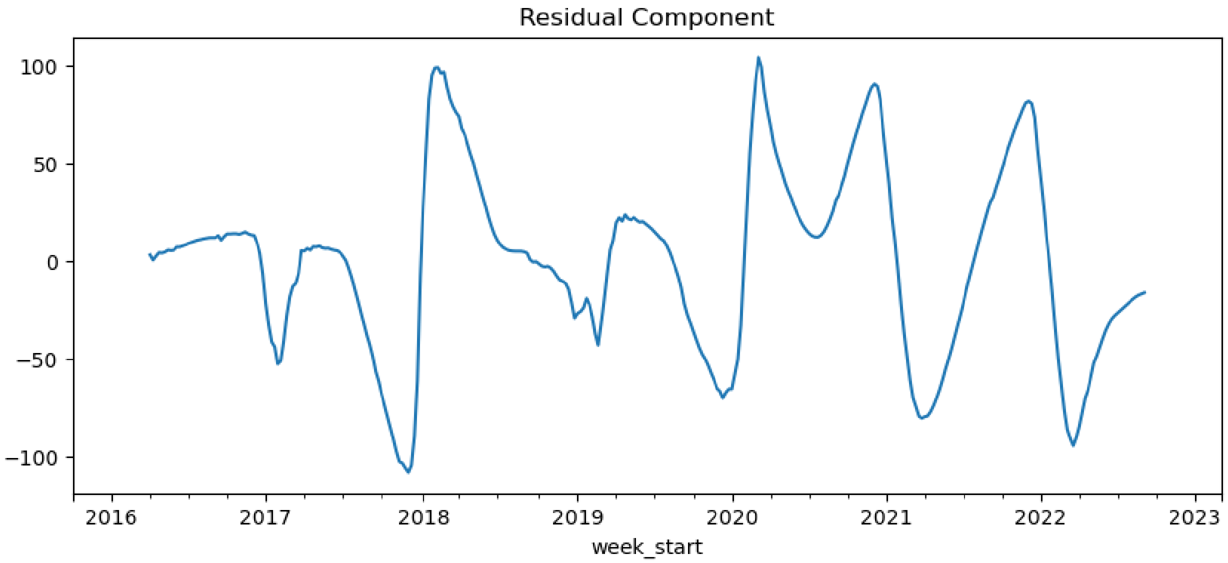


Figure 6. Residual component of the cumulative weekly ICU admissions in Chicago.

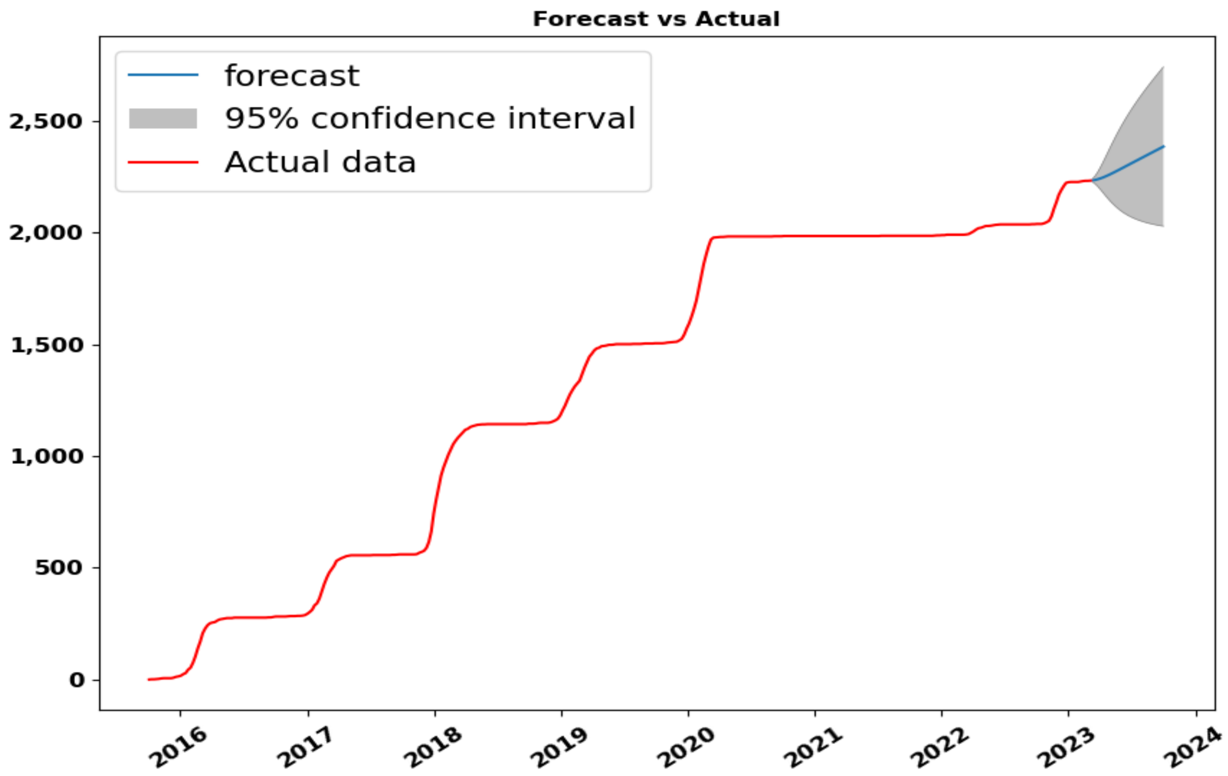
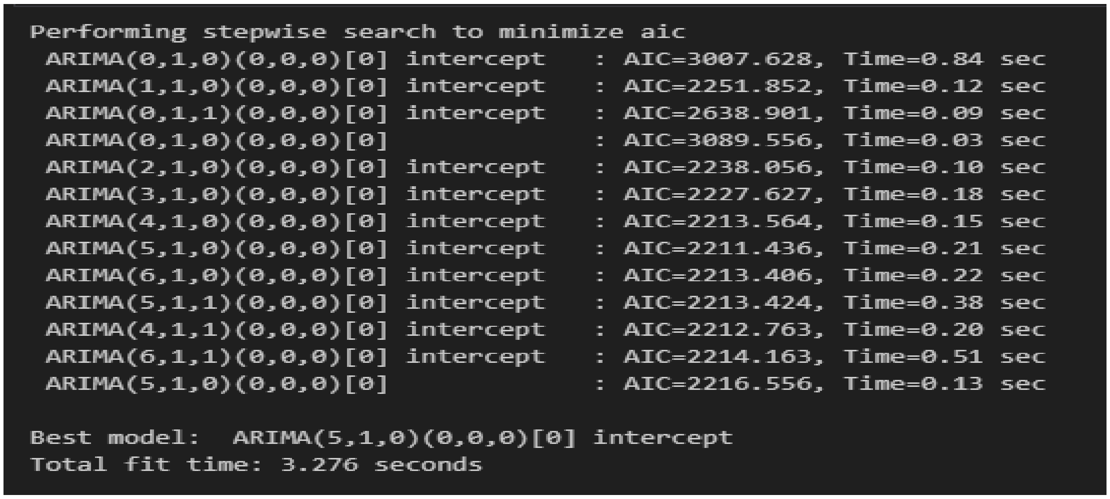


Figure 7. Forecast of cumulative weekly ICU admissions in Chicago

The weekly ICU cumulative admissions in Chicago were forecasted using ARIMA (5,1,0) with the lowest AIC = 2213 out of the combinations of p,d, and q with p_max = 10 and q_max = 10. Auto Arima of the pmdarima library was used to test the various possible combinations to select the best model, the AIC values were evaluated to select the best model. ARIMA(5,1,0), which has 5 autoregressive terms that take 5 past values into account to predict the future values. The 1 represents the 1 differencing of the time-series, the model uses the first difference values of the actual value to train and predict the future values. Where the MA term is 0 which refer to the past errors in the time series to predict the future. Thus the ARIMA(5,1,0) (Figure 8) uses 5 autoregressive terms, one difference and no moving average terms to model the time series dataset. This model uses the past 5 values to predict the future values, while considering the underlying trend and seasonality of the dataset. Using the optimized model, the forecast of the ICU cumulative admissions in Chicago was forecasted. The figure 6 shows the real values and the forecast values of 30 weeks into the future. The red line represents the real data, the blue line represents the forecast data and the shaded region represents the 95% confidence interval. Based on the best model, the forecast ICU admission values are gradually increasing with the forecasted value of 2500. The 95% confidence interval is reasonably wide with a lower limit of 2000 and upper limit of 2800.



```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=3007.628, Time=0.84 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=2251.852, Time=0.12 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=2638.901, Time=0.09 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=3089.556, Time=0.03 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=2238.056, Time=0.10 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=2227.627, Time=0.18 sec
 ARIMA(4,1,0)(0,0,0)[0] intercept   : AIC=2213.564, Time=0.15 sec
 ARIMA(5,1,0)(0,0,0)[0] intercept   : AIC=2211.436, Time=0.21 sec
 ARIMA(6,1,0)(0,0,0)[0] intercept   : AIC=2213.406, Time=0.22 sec
 ARIMA(5,1,1)(0,0,0)[0] intercept   : AIC=2213.424, Time=0.38 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=2212.763, Time=0.20 sec
 ARIMA(6,1,1)(0,0,0)[0] intercept   : AIC=2214.163, Time=0.51 sec
 ARIMA(5,1,0)(0,0,0)[0]             : AIC=2216.556, Time=0.13 sec

Best model:  ARIMA(5,1,0)(0,0,0)[0] intercept
Total fit time: 3.276 seconds
```

Figure 8. The best model selected by the auto_arima in a stepwise manner to select the best model.

Limitations & Future Research

There were a few limitations in the study that should be taken into consideration when interpreting the results. Firstly, the dataset used in this study only contains information about influenza surveillance weekly in the city of Chicago. Therefore, the conclusions drawn from this study may not be generalizable to other cities or states. Additionally, the analysis and forecasts presented in this study were only possible for the city of Chicago, as there were no records available for other cities or states in the dataset.

To overcome this limitation, further studies could use datasets that include information from multiple cities or states to gain more comprehensive understanding of the influenza surveillance weekly phenomenon.

Secondly, this study utilized ARIMA for forecasting, however, other deep learning models such as LSTM or Prophet could also be employed to further improve the forecasting accuracy. Further study could explore and compare the performance of different forecasting models to select the most appropriate one for influenza surveillance weekly data.

Conclusion

The influenza dataset of Chicago was investigated and used for exploratory data analysis to identify the total samples tested and total positive results for influenza from 2015 to 2023. The change in risk level over the years was reported. Similarly, the proportion of different variants A,B, H1N1 and H3N2 of influenza was studied. The best model ARIMA (5,1,0) was developed and the forecasts with 95% confidence intervals were reported. In this report the computational tools ;Python, R programming language and SQL were utilized to accomplish the task and derive meaningful insights.

References

Berry, I., Rahman, M., Flora, M. S., Shirin, T., Alamgir, A. S. M., Khan, M. H., ... & Fisman, D. N. (2022). Seasonality of influenza and coseasonality with avian influenza in Bangladesh, 2010–19: a retrospective, time-series analysis. *The Lancet Global health*, *10*(8), e1150-e1158.

CDC. (2019, September 13). *Key Facts About Influenza (Flu)*. Centers for Disease Control and Prevention; CDC. https://www.cdc.gov/flu/about/keyfacts.htm

Chen, Y., Leng, K., Lu, Y., Wen, L., Qi, Y., Gao, W., ... & Dong, J. (2020). Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018. *Epidemiology & Infection*, *148*.

Influenza - HAN - Chicago Health Alert Network. (n.d.). HAN. https://www.chicagohan.org/diseases-and-conditions/influenza

Influenza Surveillance Weekly | City of Chicago | Data Portal. (n.d.). Data.cityofchicago.org. Retrieved May 6, 2023, from https://data.cityofchicago.org/Health-Human-Services/Influenza-Surveillance-Weekly/6xmk-qk57

Pandemic Influenza A (H1N1) in Chicago. (2009). https://www.chicago.gov/content/dam/city/depts/cdph/infectious_disease/Communicable_Disease/IP_CDInfo_FEB2010_PandemicFlu.pdf

Petrova, V. N., & Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, *16*(1), 47-60.

Selva Prabhakaran. (2019, February 18). *ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+*. Machine Learning Plus. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Siddhivinayak, H., Newman, L. P., Paget, J., Azziz-Baumgartner, E., Fitzner, J., Niranjan, B., ... & Zhang, W. (2016). Influenza seasonality in the tropics and subtropics-when to vaccinate?. *PLoS ONE*, *11*(4).

Soyemi, K., Medina-Marino, A., Sinkowitz-Cochran, R., Schneider, A., Njai, R., McDonald, M., ... & Aiello, A. E. (2014). Disparities among 2009 pandemic influenza A (H1N1) hospital admissions: a mixed methods analysis–Illinois, April–December 2009. *PloS one*, *9*(4), e84380.