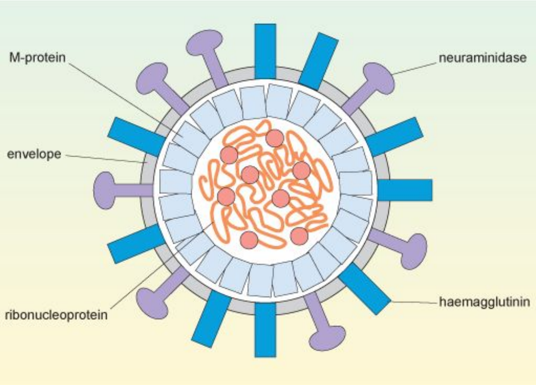
Project 2: USA Influenza Cases Time Series

Diyana Tial, ID: 16334584 (Preparing Slides,Computing,Editing,Revising)
Annalisa Berg, ID: 16339285 (Preparing slides,Computing,Editing,Revising)

Introduction



- Historically, seasonal influenza vaccines have been an integral component of U.S. public health since the late 1970s. Since implemented, the vaccines were designed to provide trivalent immunity, against three strains of influenza. Influenza vaccines typically contain viral components against influenza strains A (H1), A (H3), and strains of variation B. (Centers for Disease Control and Prevention [CDC], 2024)

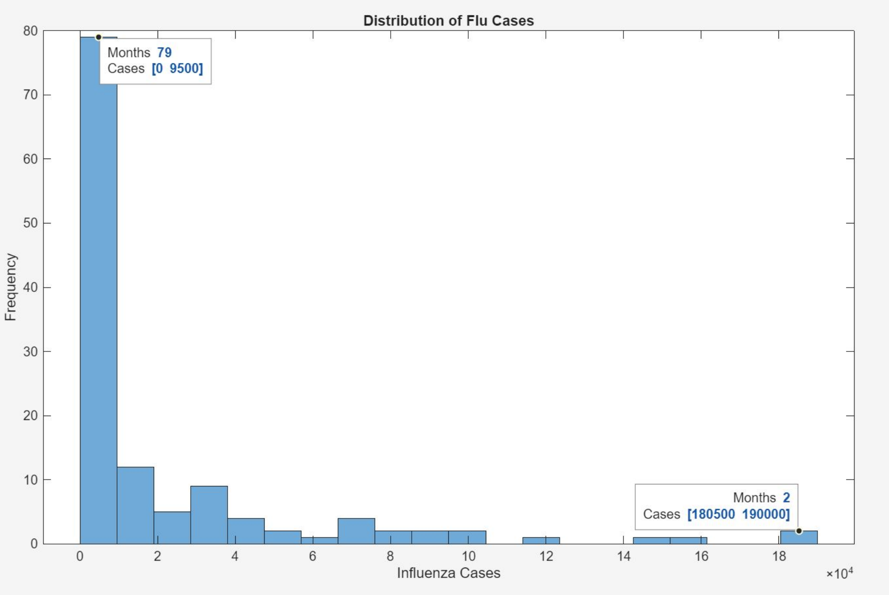
- Influenza vaccines are of this nature to provide protection against the evolving nature of viruses.

- During the 2013-2014 season, quadrivalent influenza vaccines were implemented, but were later replaced by the traditional trivalent vaccines in the 2024-2025 flu season.

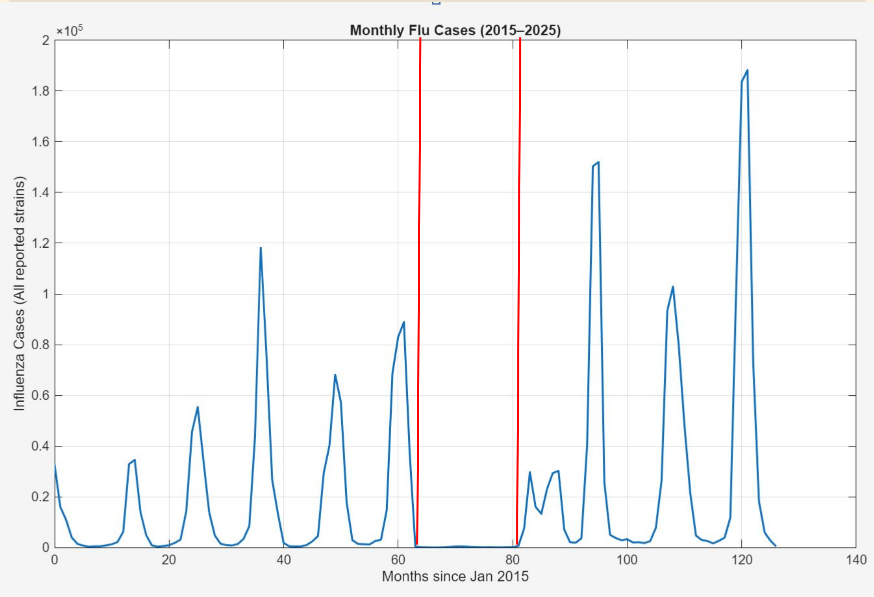
- Flu variations are denoted by their physical characteristics. Influenza A strains are characterized by the presence of hemagglutinin and neuraminidase proteins on their surface. Influenza B viruses belong to lineages, and are slow-evolving. Other influenza strains, C and D, are not commonly reported. (Centers for Disease Control and Prevention [CDC], 2024)

- **The aim of this project**, is to identify reported cases of Influenza, under the lens of the 2015-2025 time period. Taking into account a switch of vaccine methods that were implemented in the middle of the data set, we expect to see results that demonstrate a change in vaccine type.

Histogram



Plot



In the figure, red bars are used to indicate months during the COVID-19 pandemic, where flu case numbers are relatively low compared to other years.

Mean and Variance calculations

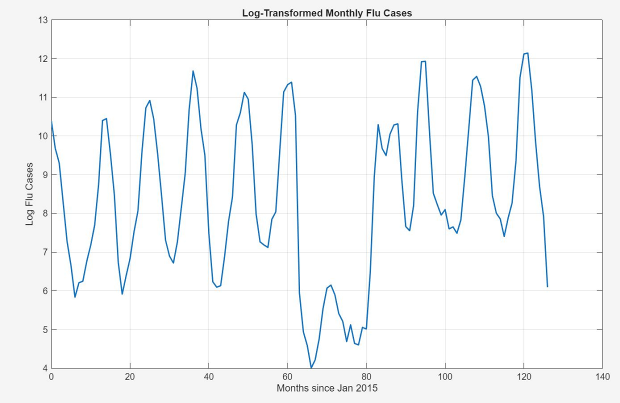
Overall mean of flu cases: 21314.98
Overall variance of flu cases:
1378039355.37
Rolling mean (first window): 8259.67
Rolling mean (last window): 67441.43
Rolling variance (first window):
157017478.24
Rolling variance (last window):
7167613909.62

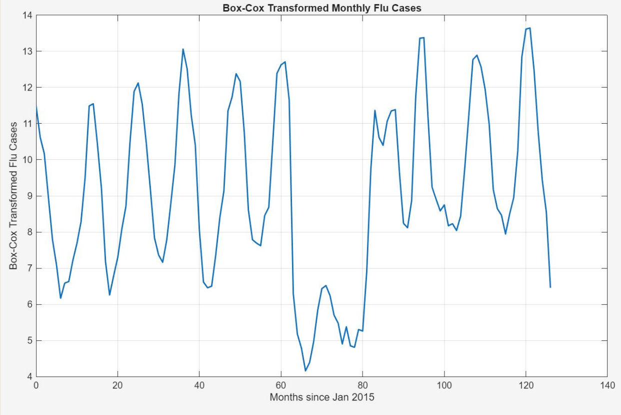
Rolling average and rolling variance were used to observe changes. In our data, the 12-month rolling mean rose from ~8.3k to ~67.4k (≈8×), while the rolling variance jumped from ~1.6×10^8 to ~7.2×10^9 (≈46×). This strong mean–variance indicates that there is a need for Log transformation of the data.

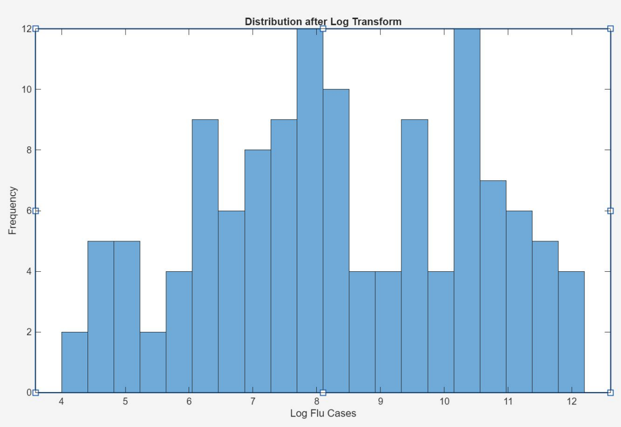
Yearly Mean & Variance through 2015-2025:

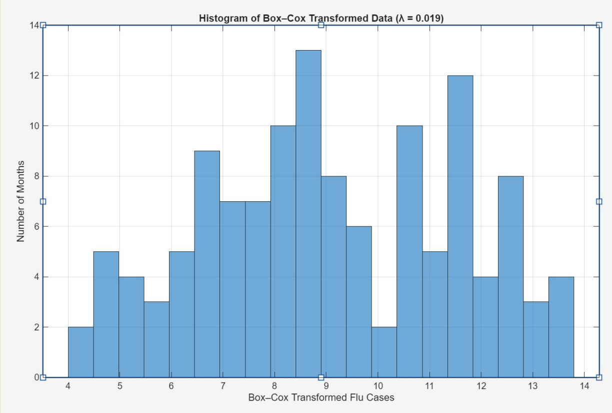
Year	MeanCases	Variance Cases
2015	5975.5	9.5275e+07
2016	9562.8	1.5107e+08
2017	17861	4.2968e+08
2018	22890	1.3837e+09
2019	23305	7.5668e+08
2020	17630	1.1344e+09
2021	3292.2	7.3553e+07
2022	39164	2.8888e+09
2023	14742	6.9316e+08
2024	31820	1.6033e+09
2025	67441	7.1676e+09

Variance Stabilization









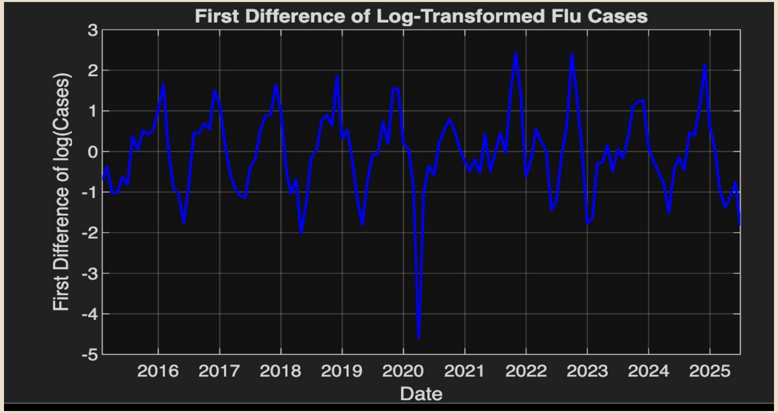
Both Box-Cox and Log transformations were tested on the data set.

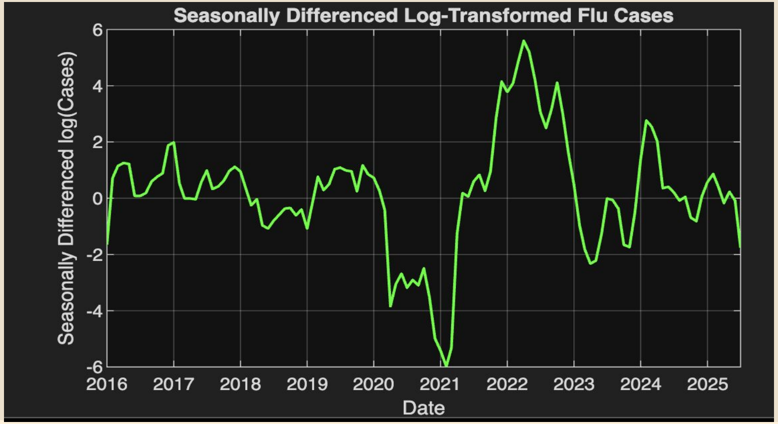
This revealed that The estimated Box–Cox λ value was close to 0, which corresponds to a log transformation.

We also can observe when comparing the Box-Cox transformation figures and Log transformation figures, that the data is visually similar.

Therefore, we proceed with log-transformed data for the rest of the analysis.

Detrending and Deseasonalizing



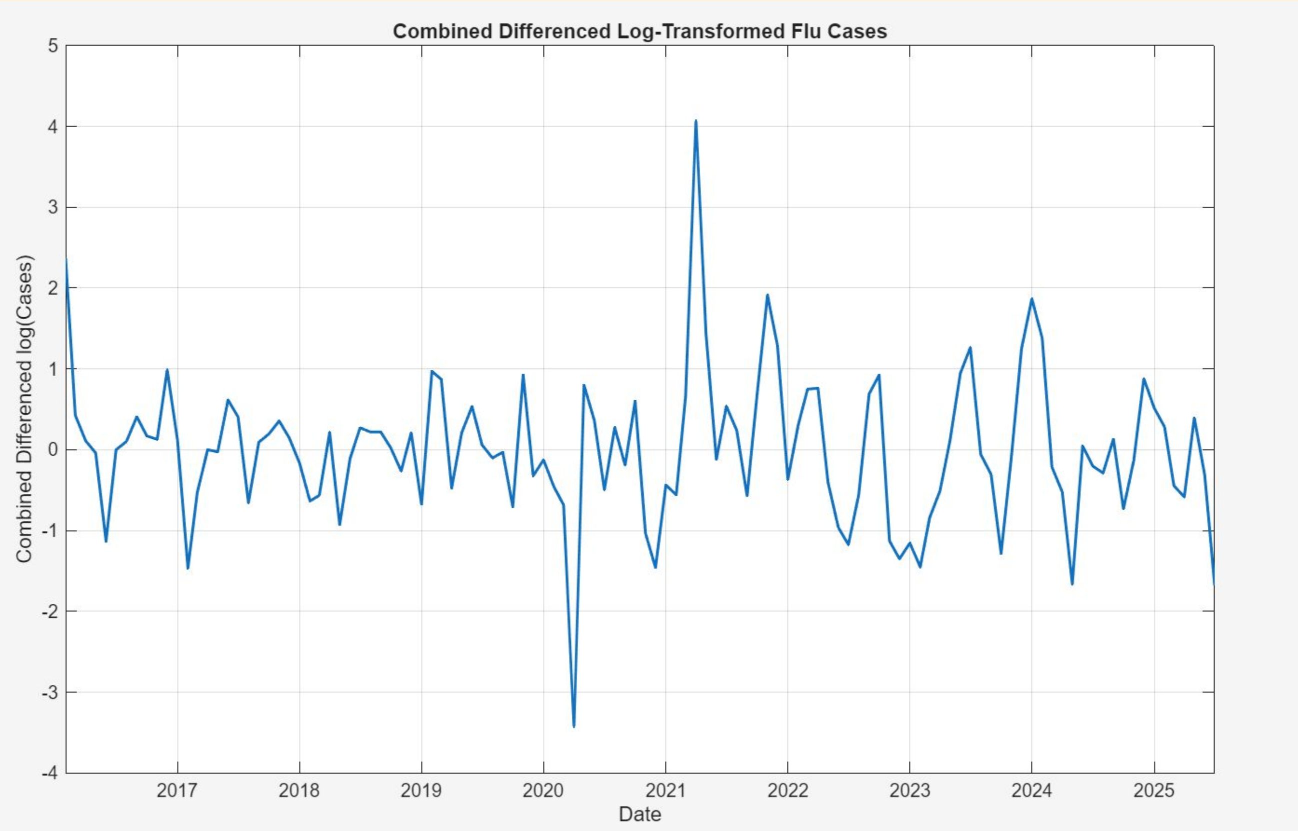


First differencing was taken of the data, shown in the first figure. First differencing removes trend so that the mean is stable.

In the second figure, the data has been plotted after seasonal differencing. Seasonal differencing is used to remove seasonal trend.

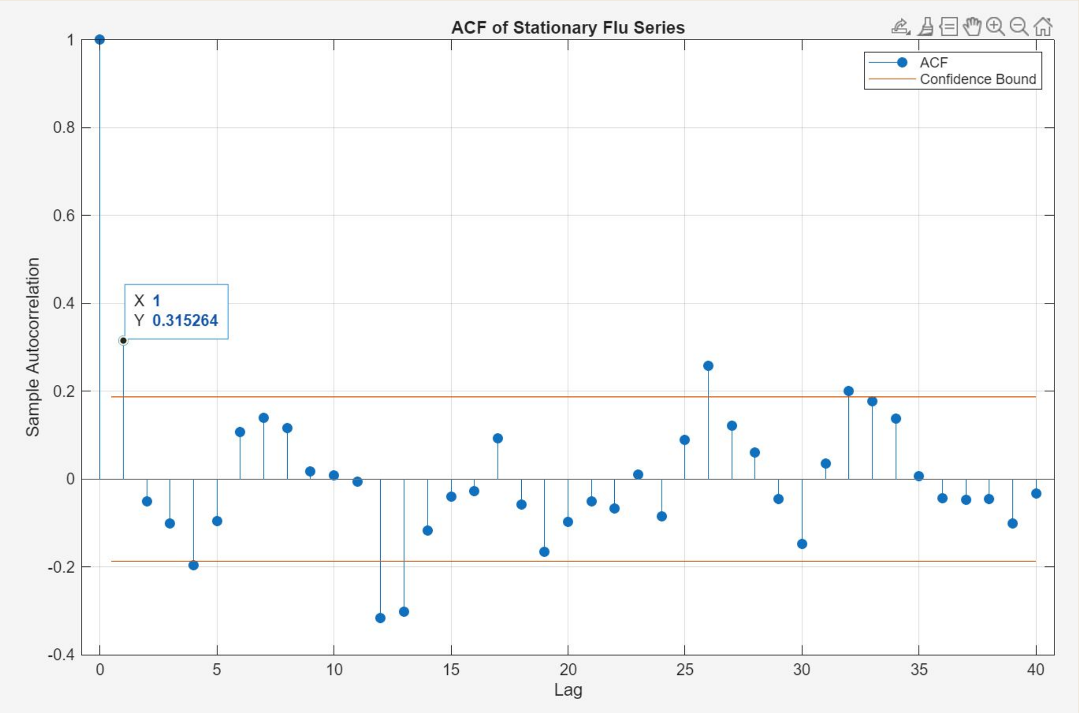
Both of these combined yield a more stationary data set compared to the raw or transformed data.

Combined Differenced Transformed Plot



This plot is a combined difference figure of the first differencing, and the seasonal differencing.

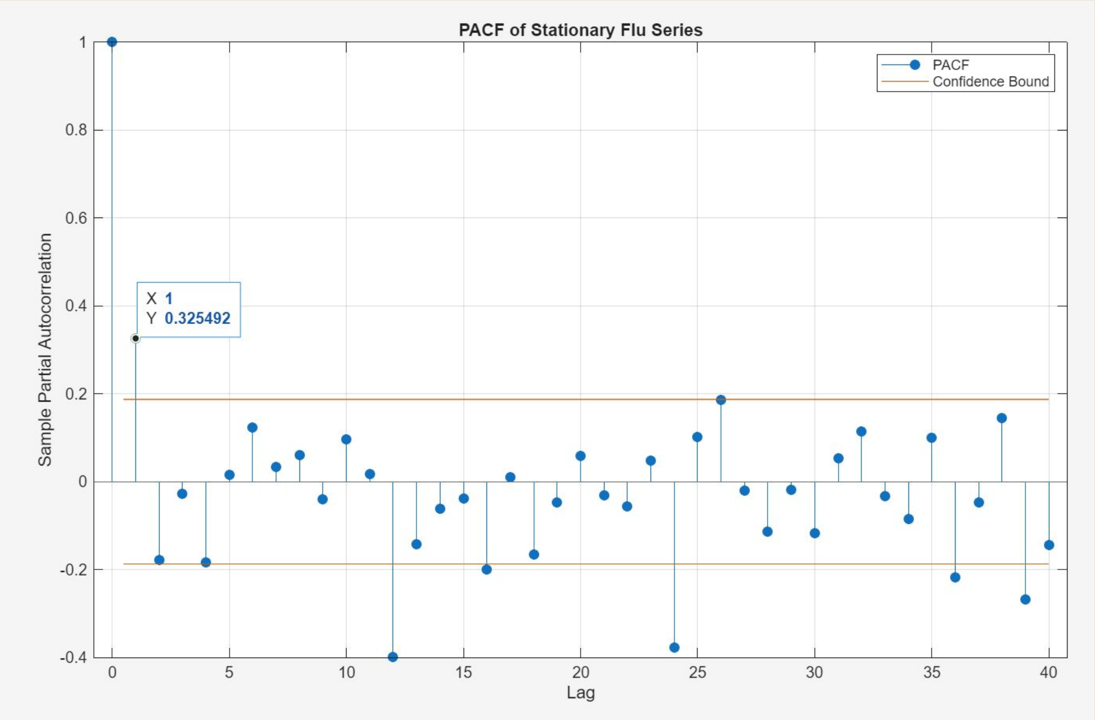
ACF Plot of Stationary Data



From the ACF plot, it is apparent that q=1.

Additionally, there is a notable significant second to last lag value at q=4.

PACF Plot of Stationary Data



From the PACF plot, we can determine a p=1 value for later necessary ARIMA/SARIMA modeling.

Also, there is a significant spike at p=4, as the second to last lag that extends past the confidence interval.

Selected SARIMA Models and Reasoning

Proposed SARIMA Models:

a. SARIMA(1,1,1)(0,1,1)[12]
b. SARIMA(1,1,1)(0,1,2)[12]
c. SARIMA(1,1,4)(0,1,1)[12]
d. SARIMA(1,1,4)(0,1,2)[12]
e. SARIMA(4,1,1)(0,1,2)[12]
f. SARIMA(4,1,1)(0,1,1)[12]

Justification of proposed SARIMA models: p is chosen by the last lag value of the PACF plot that is outside of the confidence interval, q is chosen by the last lag value of the ACF plot that is outside of the confidence interval (Kalyvas, 2023). This gave us p=1, q=1. The d value is chosen by the number of differencing on the data. Since we only removed seasonality through differencing, d=1.

For (P,D,Q)[S] values, [s]=12, from the PACF plot, because it is the largest value. Since X=12 results in Y=-0.31, the value is negative and therefore P=0, because P=0 when there is a negative ACF value at S. P+Q≤2 (Kalyvas,2023). So Q has two possible values, 1, and 2.

There is a second to last lag spike in the PACF plot at p=4. The corresponding ARIMA(4,1,1) model AIC value was tested and compared to other ARIMA values.

ARIMA(0,1,1) AIC = 322.46
ARIMA(1,1,1) AIC = 314.25
ARIMA(4,1,1) AIC = 285.41

There is also a similar spike in the ACF plot, q=4 was tested as well:

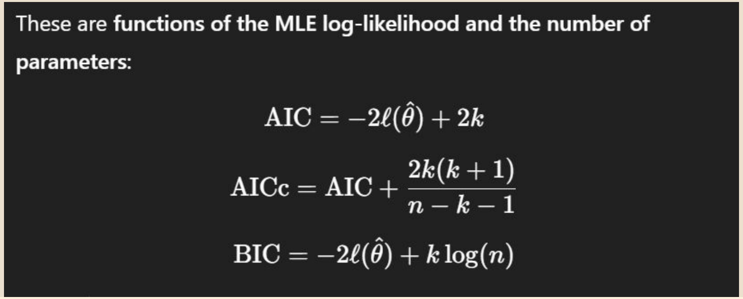
ARIMA(1,1,4) AIC= 290.42

MLE for Model Comparison

To check whether proposed AR and MR values are significant, we will use MLE, this is done by using the matlab tool “estimate” and “summarize”, these provide coefficient estimates, standard error, t-statistics, and p-values.

The following slide will show the overall result of this MLE, which provides us with AIC, AICc, and BIC values for all 6 proposed SARIMA models.

The resulting ℓ(θ^) value from this line of code:
[Est,EstParamCov,logL] = estimate(mdl, yT, 'Display','off'); is
-1.273622078154113e+02.



These are **functions of the MLE log-likelihood and the number of parameters:**

$$\mathrm{AIC} = -2\ell(\hat{\theta}) + 2k$$

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

$$\mathrm{BIC} = -2\ell(\hat{\theta}) + k\log(n)$$

AIC, AICc, BIC Comparison of proposed SARIMA Values

Model	K	AIC	AICc	BIC	Sigma^2
{'SARIMA(1,1,1)(0,1,1)[12]'}	4	281.09	281.42	292.47	0.50281
{'SARIMA(1,1,4)(0,1,2)[12]'}	8	273.24	274.46	295.99	0.44379
{'SARIMA(1,1,1)(0,1,2)[12]'}	5	274.67	275.17	288.89	0.47055
{'SARIMA(4,1,1)(0,1,2)[12]'}	8	265.33	266.55	288.08	0.41701
{'SARIMA(1,1,4)(0,1,1)[12]'}	7	277.24	278.18	297.15	0.46527
{'SARIMA(4,1,1)(0,1,1)[12]'}	7	268.72	269.67	288.63	0.4351

To determine the K value, the following equation was used: K=p+q+P+Q+1, where 1 accounts for the addition of the variance parameter.

However, compared to similar time series projects on Flu data, AR(4) or MA(4) is not typical and may result in an over-differentiated model, without significant improvement of the AIC value.

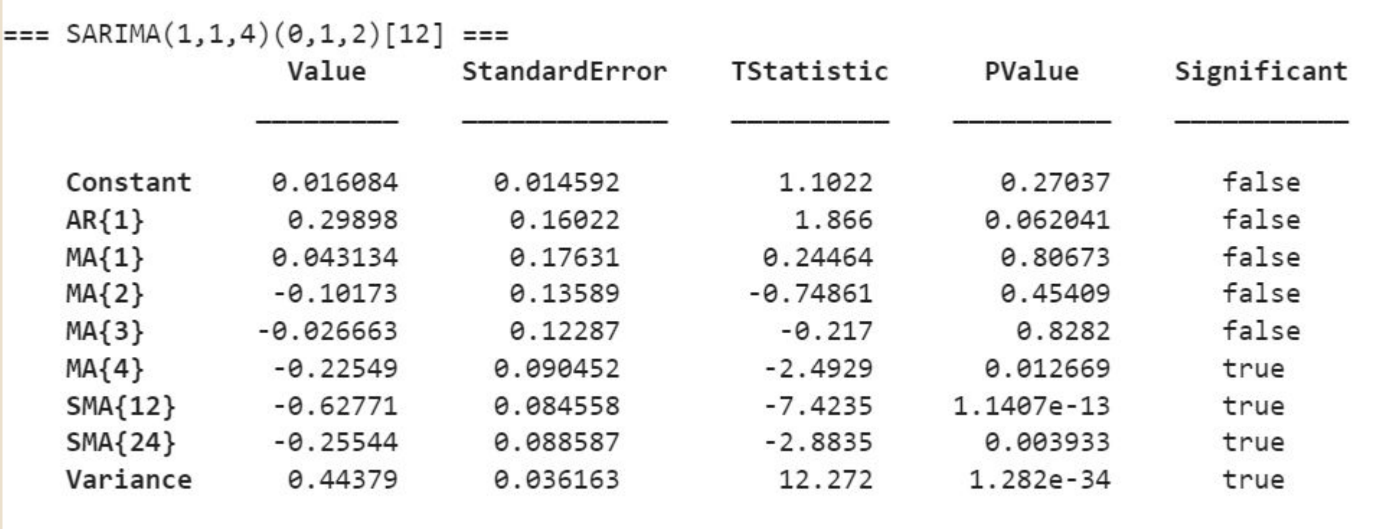
Using MLE for Parameter Significance

The final 3 most significant models were narrowed down by lowest AIC, AICc, and BIC values:

SARIMA(4,1,1)(0,1,2)[12]
SARIMA(4,1,1)(0,1,1)[12]
SARIMA(1,1,4)(0,1,2)[12]

Next, we proceeded with an MLE fit, for each individual SARIMA model. Tables were displayed with coefficient estimates, standard errors, t-statistics, and p-values for each parameter.

Using MLE for Parameter Significance

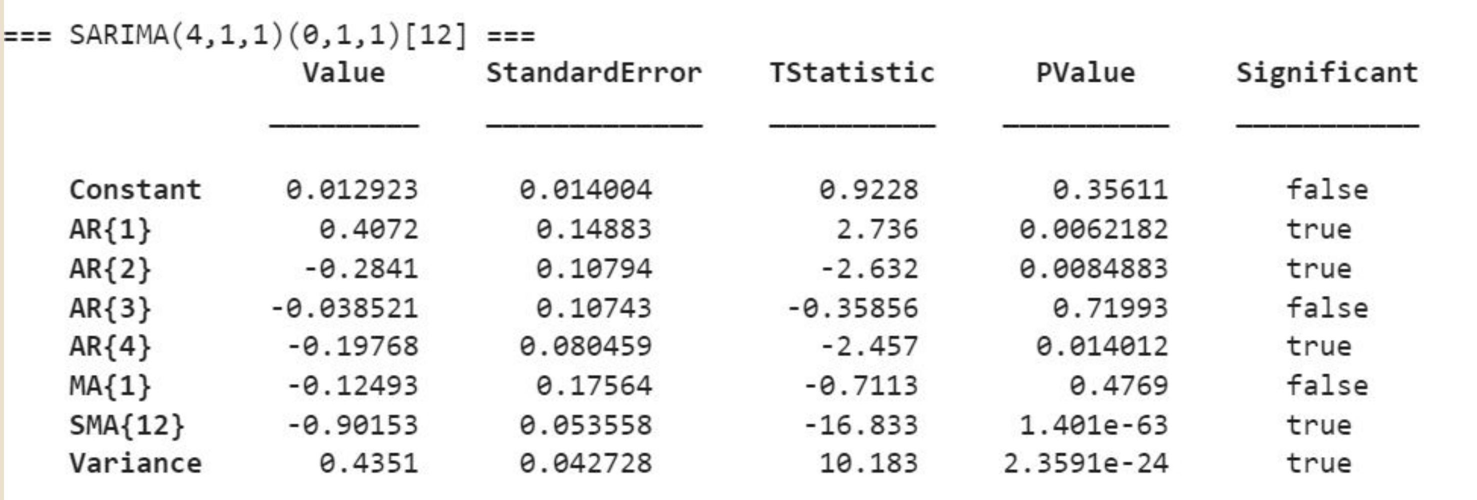


```
=== SARIMA(1,1,4)(0,1,2)[12] ===
                 Value      StandardError    TStatistic      PValue      Significant
               _________    _____________    __________    __________    ___________

    Constant    0.016084       0.014592         1.1022        0.27037        false
    AR{1}        0.29898        0.16022          1.866       0.062041        false
    MA{1}       0.043134        0.17631        0.24464        0.80673        false
    MA{2}       -0.10173        0.13589       -0.74861        0.45409        false
    MA{3}      -0.026663        0.12287         -0.217         0.8282        false
    MA{4}       -0.22549       0.090452        -2.4929       0.012669        true
    SMA{12}     -0.62771       0.084558        -7.4235     1.1407e-13        true
    SMA{24}     -0.25544       0.088587        -2.8835       0.003933        true
    Variance     0.44379       0.036163         12.272      1.282e-34        true
```

This MLE result indicates that MA(4) is the only significant MA value, and AR(1) is almost significant. This means that this model is driven by a single MA value, and can be simplified by removing other MA values.

Using MLE for Parameter Significance



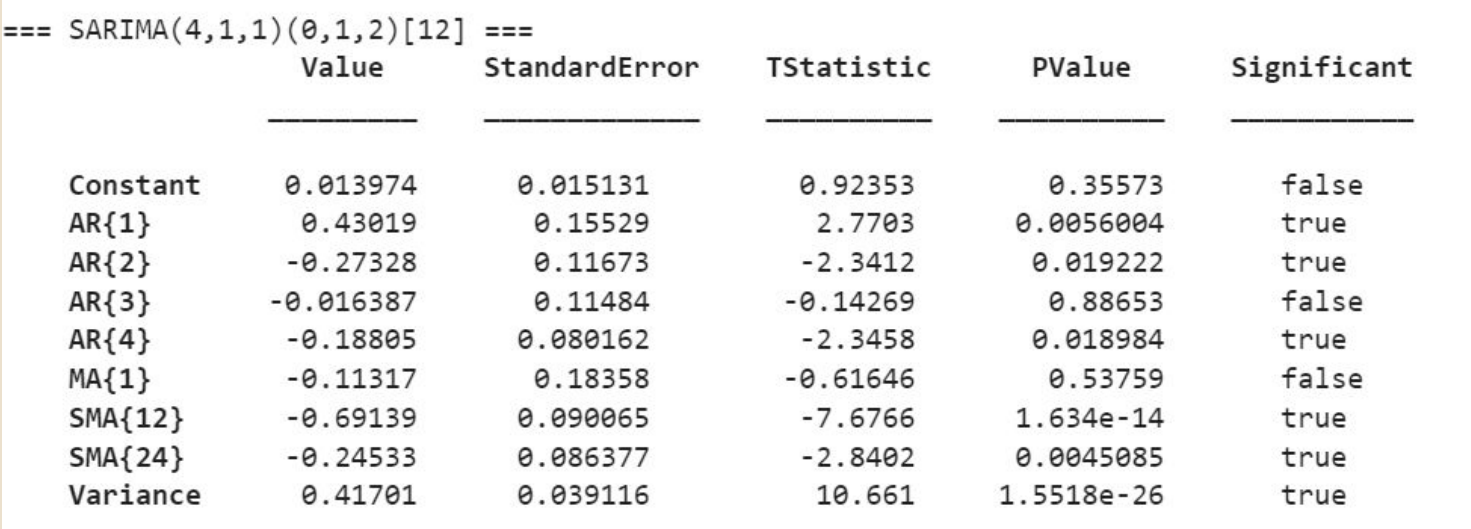
```
=== SARIMA(4,1,1)(0,1,1)[12] ===
                  Value      StandardError    TStatistic      PValue      Significant

               _________    _____________    __________    __________    ___________

    Constant    0.012923       0.014004         0.9228        0.35611        false
    AR{1}         0.4072        0.14883          2.736      0.0062182        true
    AR{2}        -0.2841        0.10794         -2.632      0.0084883        true
    AR{3}      -0.038521        0.10743       -0.35856        0.71993        false
    AR{4}       -0.19768       0.080459         -2.457       0.014012        true
    MA{1}       -0.12493        0.17564        -0.7113         0.4769        false
    SMA{12}     -0.90153       0.053558        -16.833      1.401e-63        true
    Variance      0.4351       0.042728         10.183     2.3591e-24        true
```

This MLE result indicates that AR(1), AR(2), AR(4) are the most significant parameters for this SARIMA model. This means that for this model, AR(3) and MA(1) can be set to 0.

Using MLE for Parameter Significance



```
=== SARIMA(4,1,1)(0,1,2)[12] ===
                   Value       StandardError     TStatistic       PValue       Significant
                ___________    _____________    ___________    ___________    ____________

    Constant      0.013974       0.015131         0.92353        0.35573         false
    AR{1}          0.43019        0.15529          2.7703       0.0056004        true
    AR{2}         -0.27328        0.11673         -2.3412        0.019222        true
    AR{3}        -0.016387        0.11484        -0.14269         0.88653        false
    AR{4}         -0.18805       0.080162         -2.3458        0.018984        true
    MA{1}         -0.11317        0.18358        -0.61646         0.53759        false
    SMA{12}       -0.69139       0.090065         -7.6766       1.634e-14        true
    SMA{24}       -0.24533       0.086377         -2.8402       0.0045085        true
    Variance       0.41701       0.039116          10.661      1.5518e-26        true
```

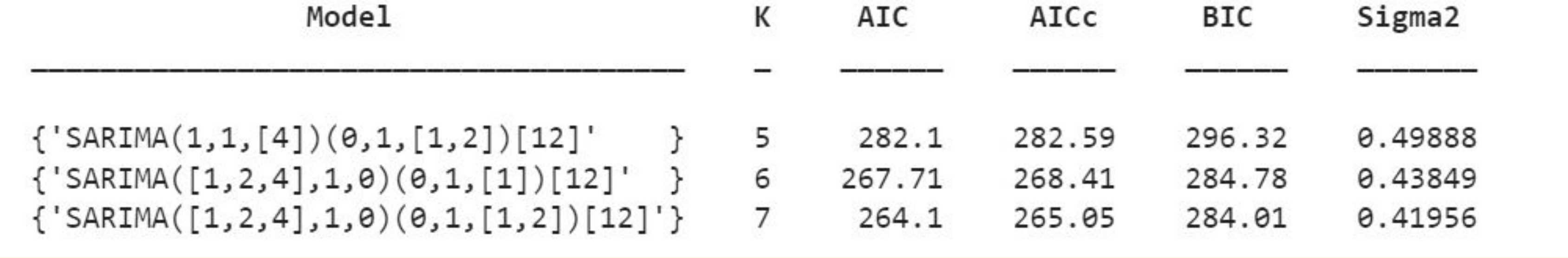
This MLE reveals that AR(1), AR(2), AR(4) are the most significant parameters for this SARIMA model. This means that for this model, AR(3) and MA(1) can also be set to 0.

AIC/AICc/BIC Values After Refinement

Proposed models with refinements:
SARIMA(1,1,4)(0,1,2)[12]
Refinement: Keep only MA(4) and seasonal terms, SMA(12, SMA(24).

SARIMA(4,1,1)(0,1,1)[12]
Refinement: AR(3) and MA(1) set to 0.

SARIMA(4,1,1)(0,1,2)[12]
Refinement: AR(3), MA(1) set to 0.



```
                 Model                       K      AIC       AICc      BIC      Sigma2
 ______________________________________     _    ______    ______    ______    _______

{'SARIMA(1,1,[4])(0,1,[1,2])[12]'     }     5     282.1    282.59    296.32    0.49888
{'SARIMA([1,2,4],1,0)(0,1,[1])[12]'   }     6    267.71    268.41    284.78    0.43849
{'SARIMA([1,2,4],1,0)(0,1,[1,2])[12]'}      7     264.1    265.05    284.01    0.41956
```

Final Chosen Model

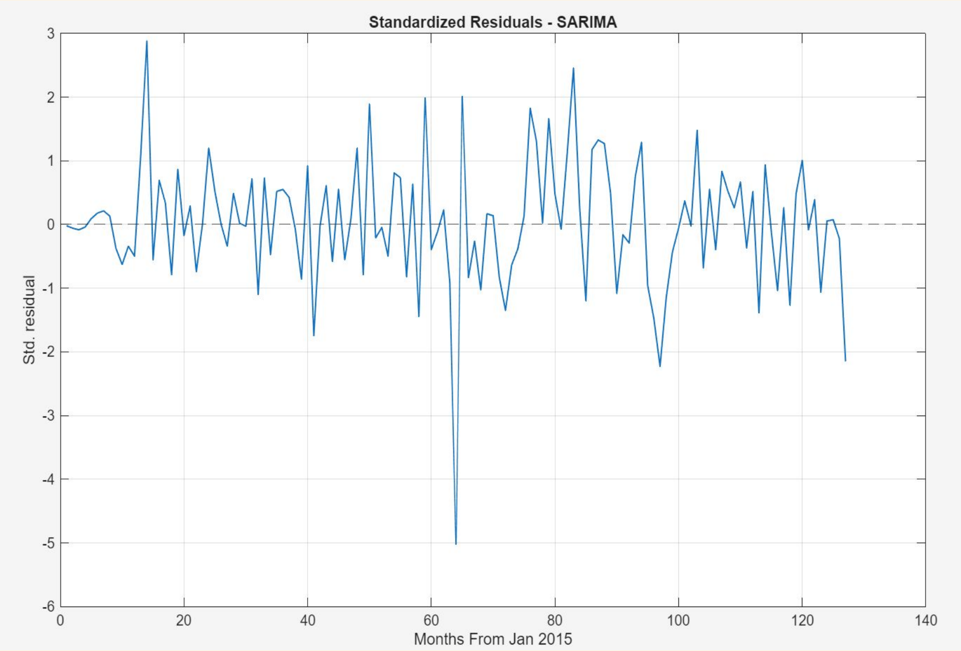
From the previous slide, using the AIC/AICc/BIC values, we can proceed with a model that has the best fit. The refined SARIMA([1,2,4],1,0)(0,1,[1,2])[12] model displays the best fit.

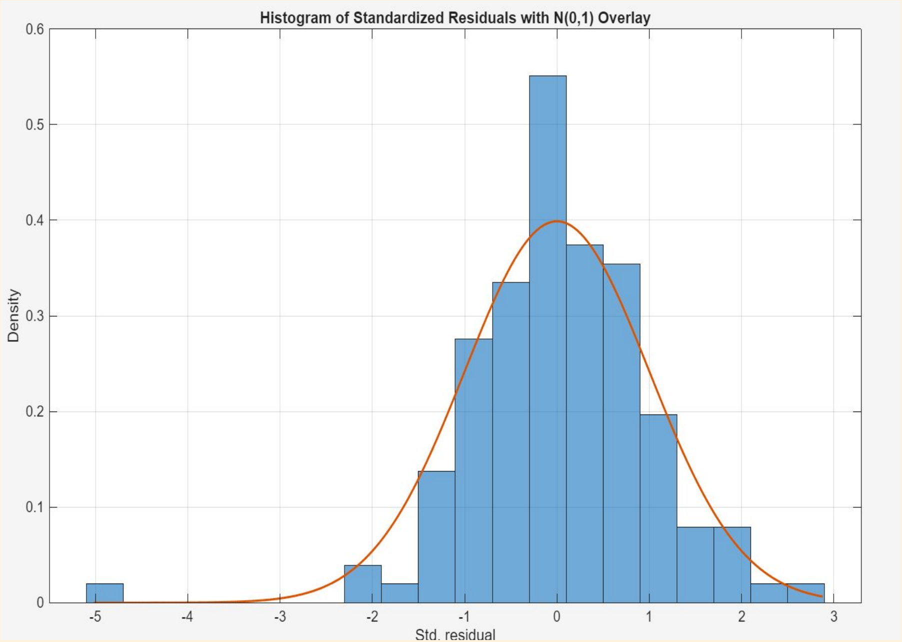
However, the best fit model with the least parameters, or, the most parsimonious model with the best fit is the SARIMA([1,2,4],1,0)(0,1,[1])[12] model. Because there is only a slight decrease in AIC values from this model to the previously mentioned model, the expense of an extra parameter may not always be worth the better fit.

Taking into account that our data is flu data, we choose the SARIMA([1,2,4],1,0)(0,1,[1,2])[12] model to proceed.

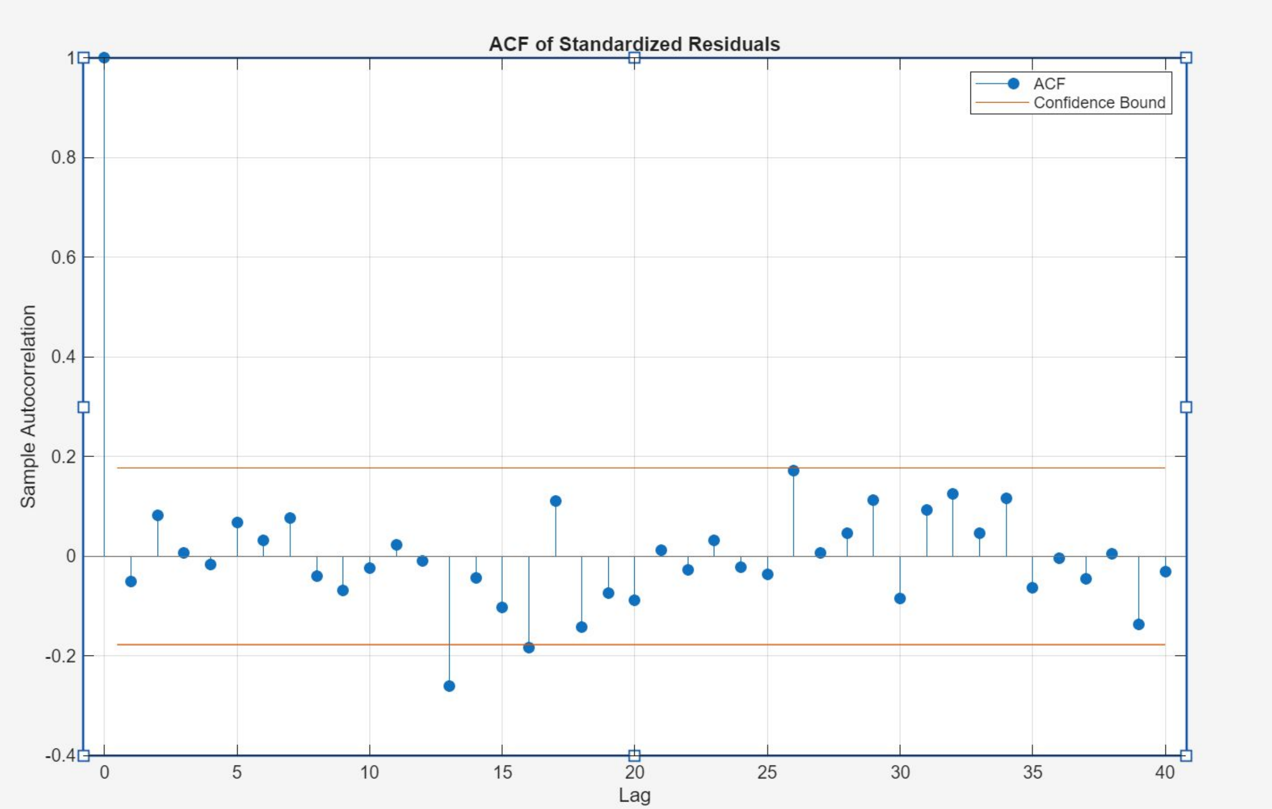
Plot and Histogram of Residuals

Residuals were calculated using the infer tool in matlab



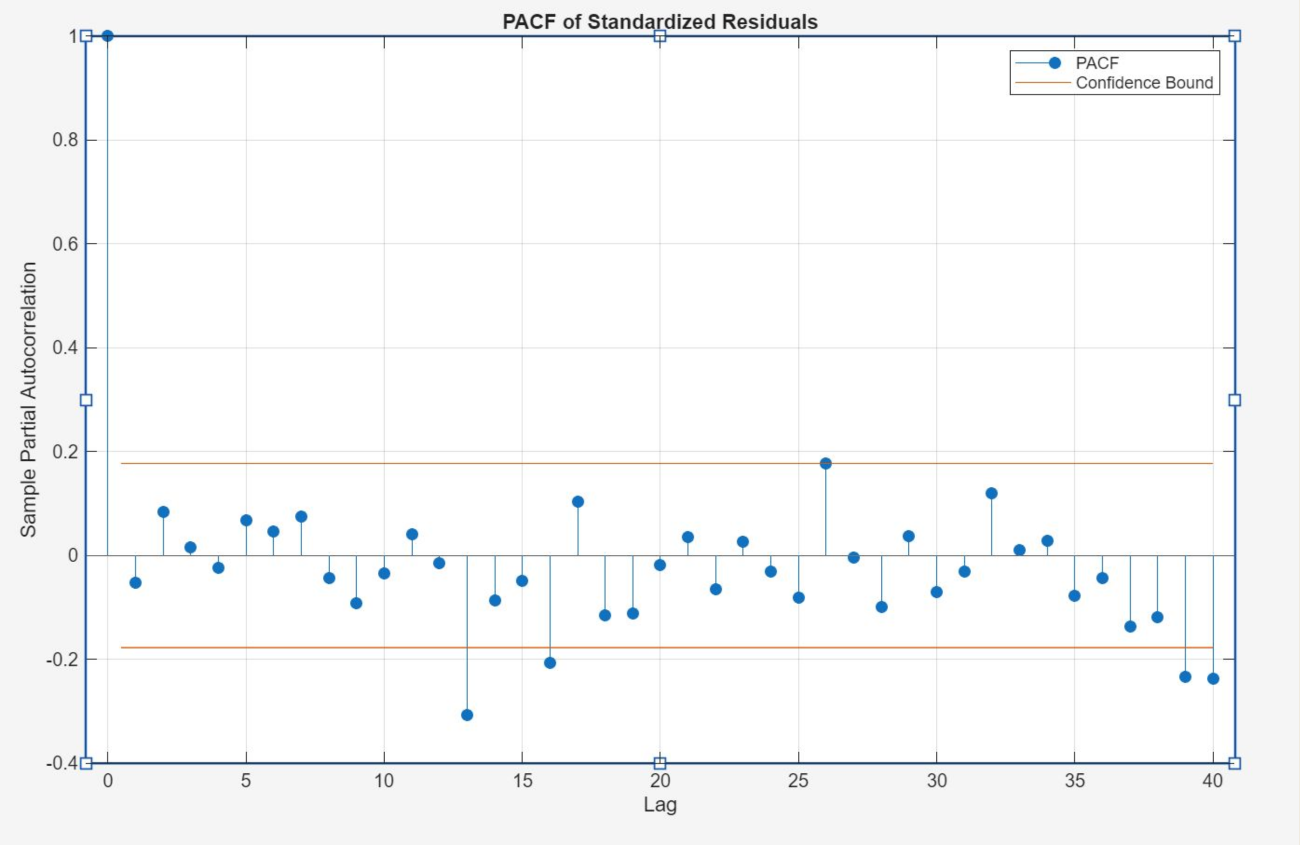


Standardized Residuals ACF



Residuals were standardized by dividing each innovation by its conditional standard deviation, yielding N(0,1) scaled residuals for valid normality and white noise diagnostics.

Standardized Residuals PACF



Randomness / Independence Tests

```
Jarque–Bera results:
   h=1 (1=reject),
   p=0.0010,
   JB=114.04
```

Interpretations:
These values indicate that residuals are not normal (heavy tails and/or skew).

```
McLeod-Li  results:
   (squared res):
   h12=0 p12=0.9947|
   h20=0 p20=0.9950|
   h24=0 p24=0.9984
```

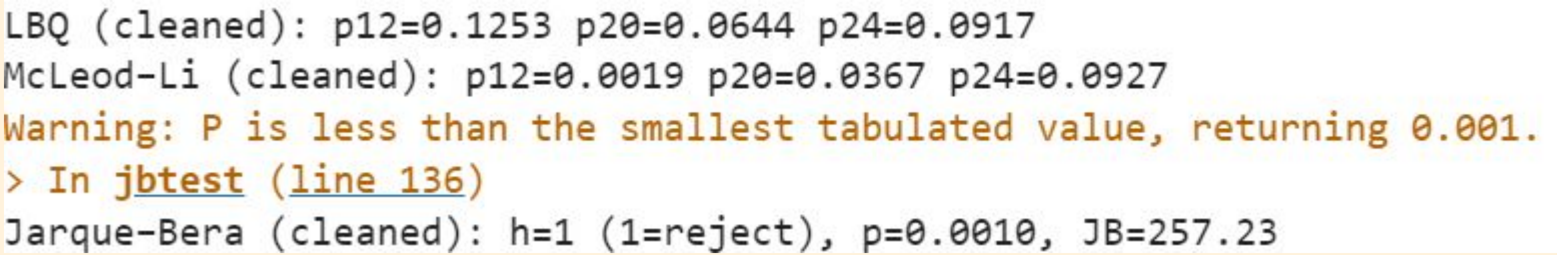
These results indicate that the residuals are non-Gaussian.

```
Ljung-Box results:
h12=0 p12=0.9856
h20=0 p20=0.1259
h24=0 p24=0.2718
```

Residuals do behave like white noise, as expected.

Refining Data Based on Interpretations

Because of the results in the randomness and independence tests, we realized it may benefit to throw outliers out of the data set, to provide better forecasting moving forward. In the plot of the raw data, it is apparent that there are significant outliers during the months of the COVID-19 pandemic. These outliers were removed, and randomness tests were re-conducted.

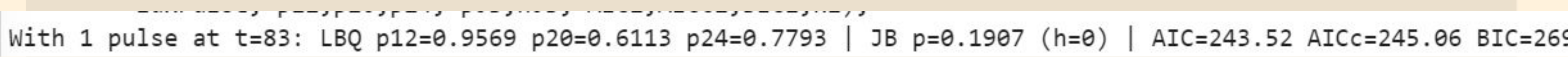


```
LBQ (cleaned): p12=0.1253 p20=0.0644 p24=0.0917
McLeod-Li (cleaned): p12=0.0019 p20=0.0367 p24=0.0927
Warning: P is less than the smallest tabulated value, returning 0.001.
> In jbtest (line 136)
Jarque-Bera (cleaned): h=1 (1=reject), p=0.0010, JB=257.23
```

However, the results indicate that the model is still behaving non-normally. So, we must use a different approach.

New Suggested Refinement

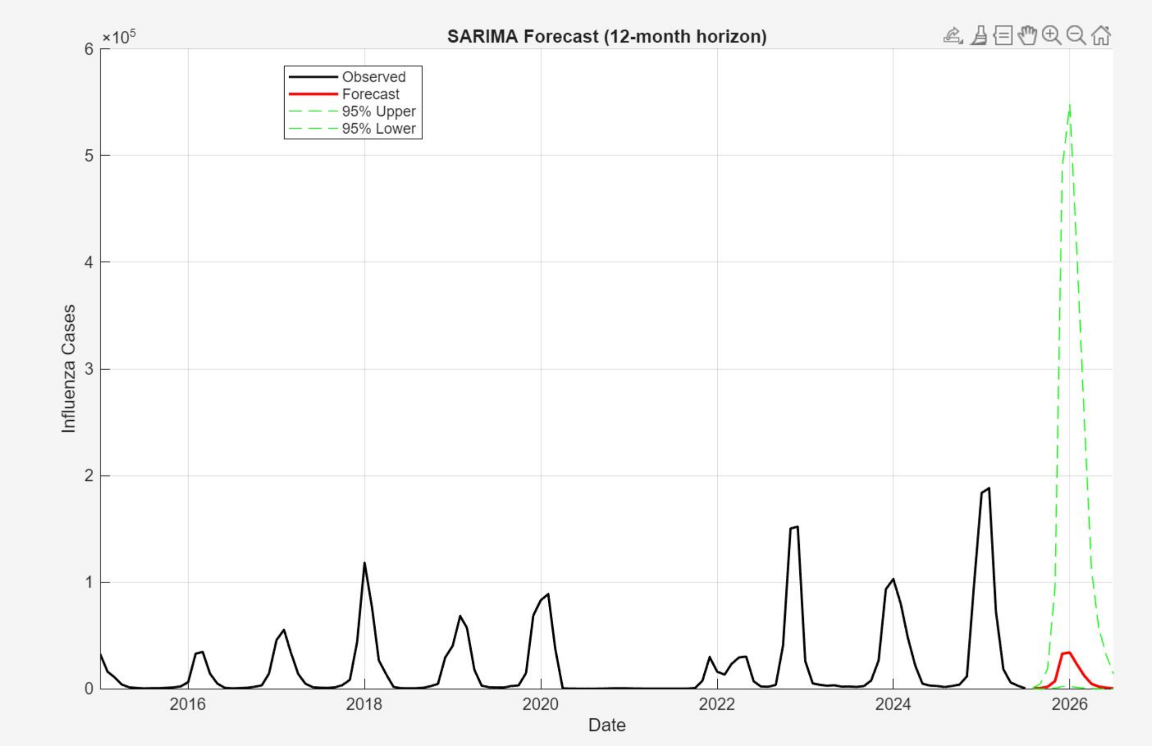
Because of the impact of the COVID-19 pandemic, we must find a method to refine the flu data without removing outliers, but instead, adjusting the averages of the data. Similar flu models have done, we can add a level-shift for the months of COVID, as well as an outlier pulse.



```
With 1 pulse at t=83: LBQ p12=0.9569 p20=0.6113 p24=0.7793 | JB p=0.1907 (h=0) | AIC=243.52 AICc=245.06 BIC=26
```

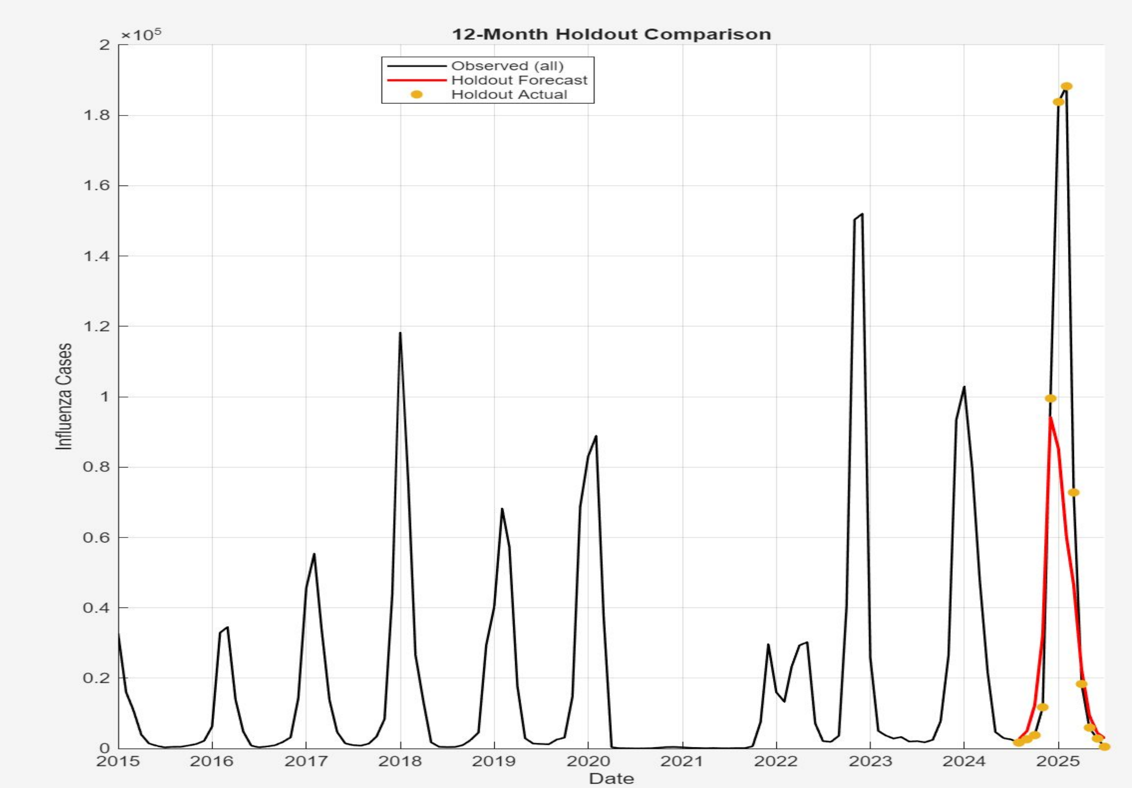
This resulted in better scores for randomness tests.
So, we proceed with the modified data and previously proposed SARIMA model.

Forecasting Using SARIMA



To generate a forecast, we use the chosen SARIMA model, and forecast out for the 2025-2026 period of flu cases.

12-Month holdout Forecast



In order to test the accuracy of our chosen SARIMA model, we can generate a forecast while omitting the last year of the data.

Then, overlaying this forecast with the holdout data, we can see that our model did not accurately predict the real influenza case data of the 2024-2025 period.

However, it accurately predicted the trend of cases, just not the magnitude of cases.

RMSE	MAE	MAPE
47736	25192	117.15

RMSE/MAE: Large absolute errors (~25k–47k cases per month), dominated by outbreak peaks.

MAPE: >100%, indicating forecasts underestimated severe outbreaks.

These results confirm the model captures timing/seasonality but **systematically underestimates peak magnitudes**.

Conclusion

The forecast errors were high mainly because *flu outbreaks are unpredictable*, not because the model was set up incorrectly.
Flu cases sometimes spike sharply when new strains appear or when vaccination rates are low, and these sudden peaks are very hard for SARIMA models to capture.

The model was able to recognize the seasonal pattern and it predicted when flu tends to rise each year but it consistently underestimated the magnitude of cases.
This explains why the error values, especially MAPE are large.

The COVID-19 pandemic also disrupted normal flu patterns in ways that are difficult to model, even after adding special adjustments.
In short, the model did well at predicting the *timing* of flu outbreaks, but not their *size*, which shows how uncertain epidemic forecasting can be.

Sources

Choosing a SARIMA model for time series:
https://www.linkedin.com/pulse/time-series-episode-1-how-select-correct-sarima-vasilis-kalyvas-jqcjf/

Shift-Step Justification:
https://pmc.ncbi.nlm.nih.gov/articles/PMC10461650/

Example of other studies that used SARIMA modeling for flu data:
https://pmc.ncbi.nlm.nih.gov/articles/PMC6926639/

Information on the Influenza Virus:
https://www.cdc.gov/flu/about/viruses-types.html