Analyzing Foodborne Disease Outbreaks Over Time

A look into illnesses and more

Group 12

Trina Shores (group leader)

Steven Macapagal

Journey Martinez

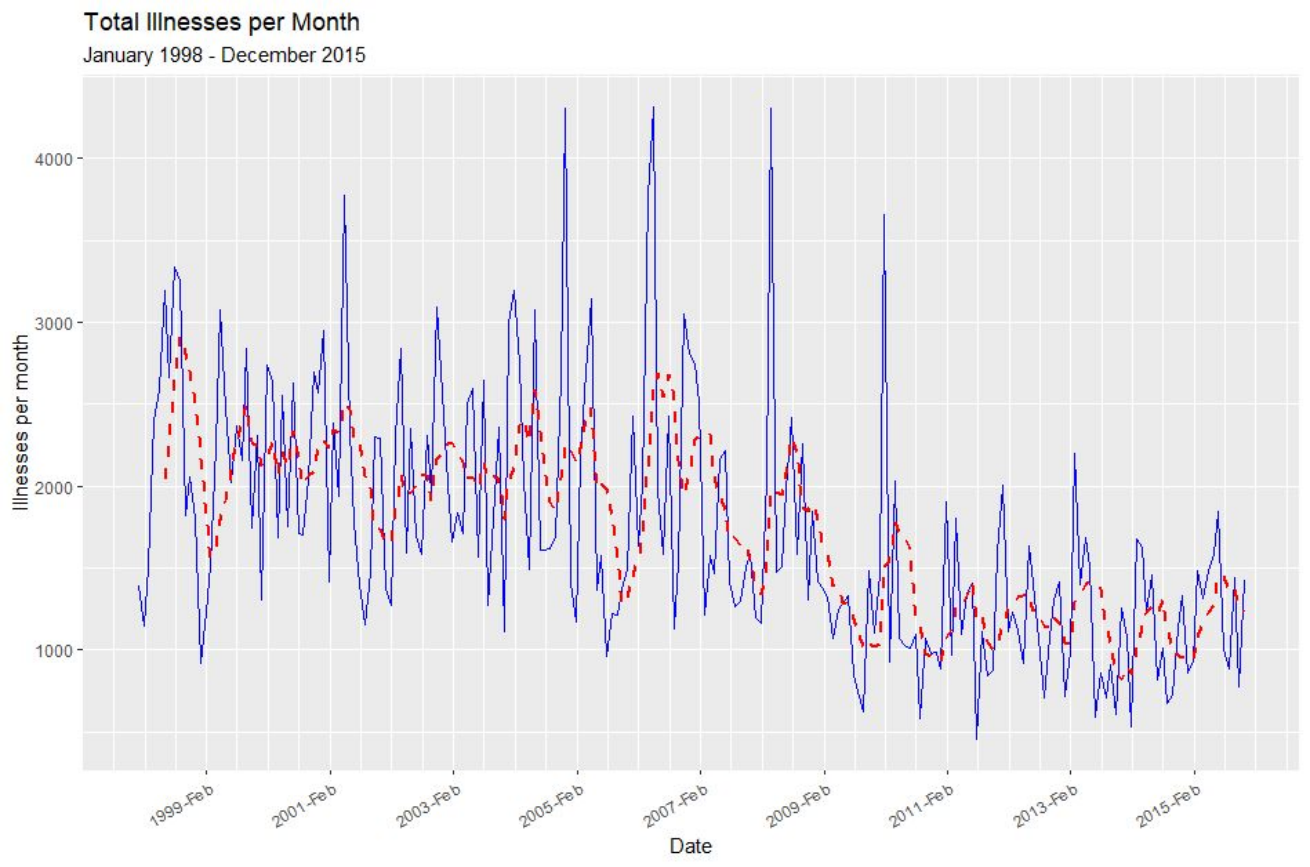
Yuan Yao

Heather Nagy

Kenneth Porter

Data on foodborne illnesses

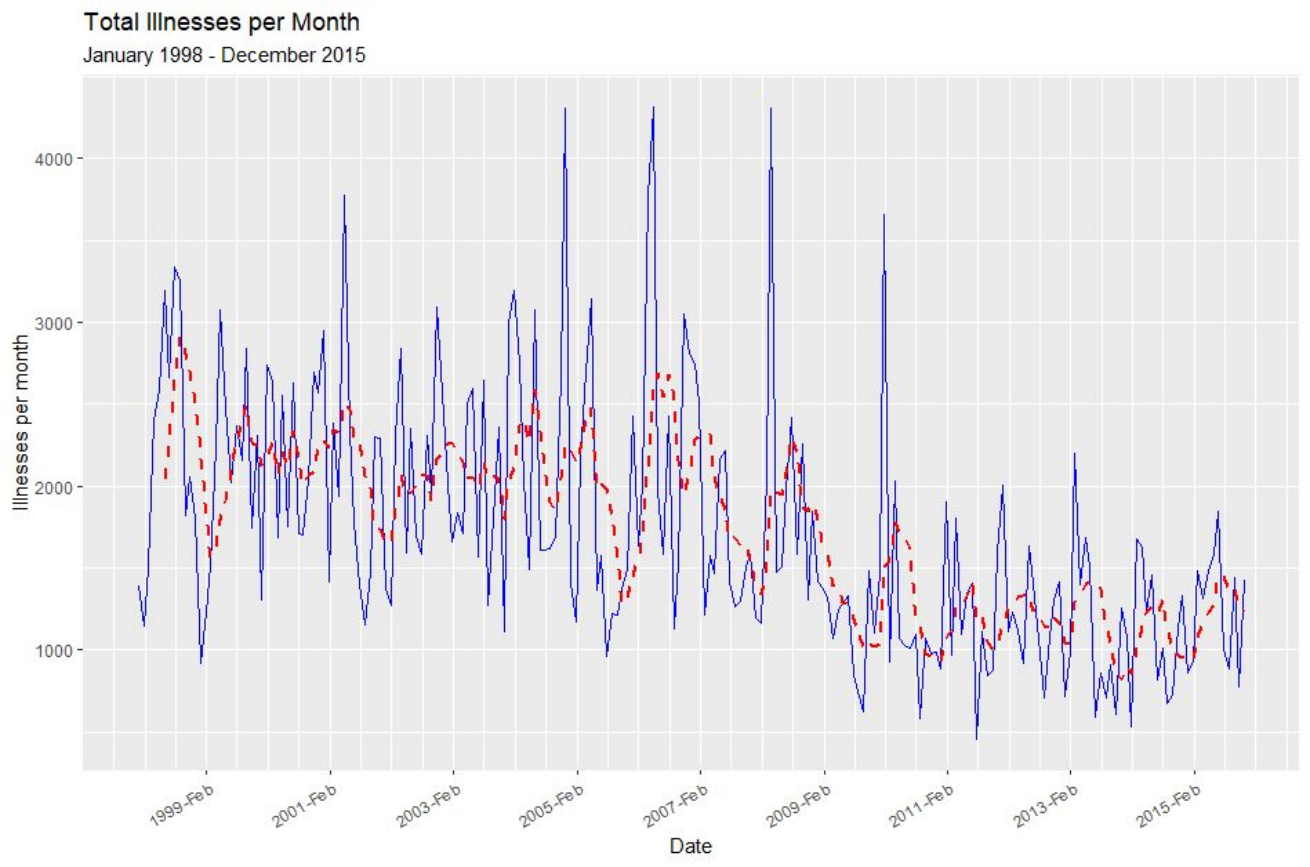
A reminder of what our data look like:



Solid blue line represents the original time series

Dotted red line represents a filtered time series over a 6-month period

Total illnesses per month appear to decline over time and are less variable over time.



Variability is much greater from 1998 to 2008 and decreases from 2009 to 2015. Illnesses also seem to be cyclical.

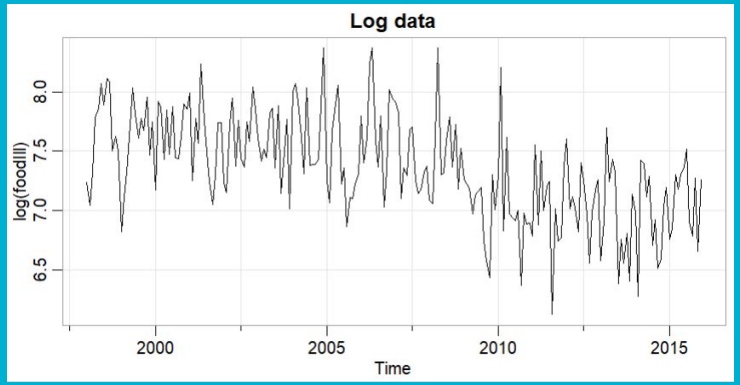
Our revised goals

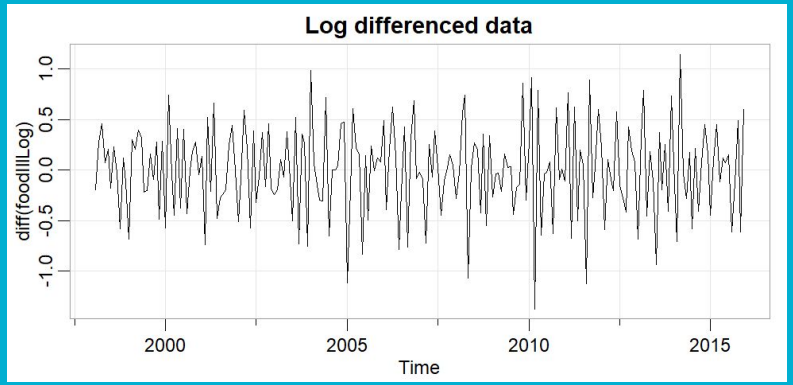
- How is the rate of foodborne illness changing over time?
  - Directional trends
  - Changes in variability
  - Cyclical trends
- Does type of illness or establishment affect the rate of foodborne illness over time?

Selecting and estimating a model

Addressing Nonstationarity in Data

- Heteroskedasticity
  - Because the variability changes throughout the plot, we will log transform the data. Log transformations bring extreme values closer to the rest of the data, removing some variability.

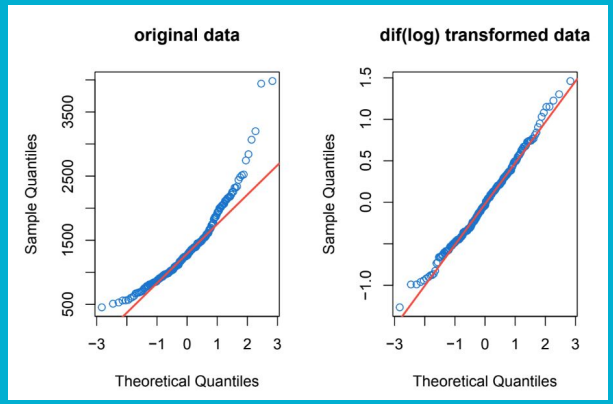


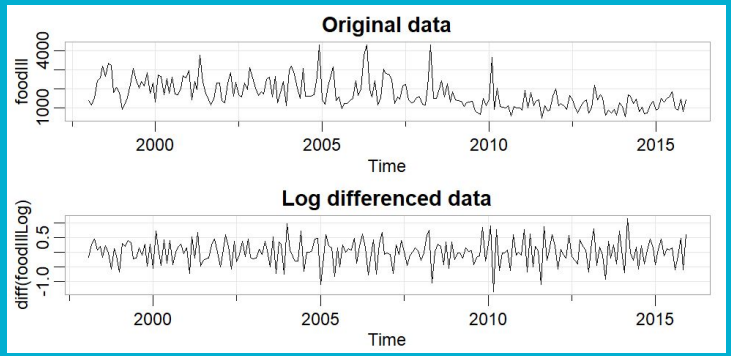


- Trend
  - Because there is a trend, we will then difference the data.

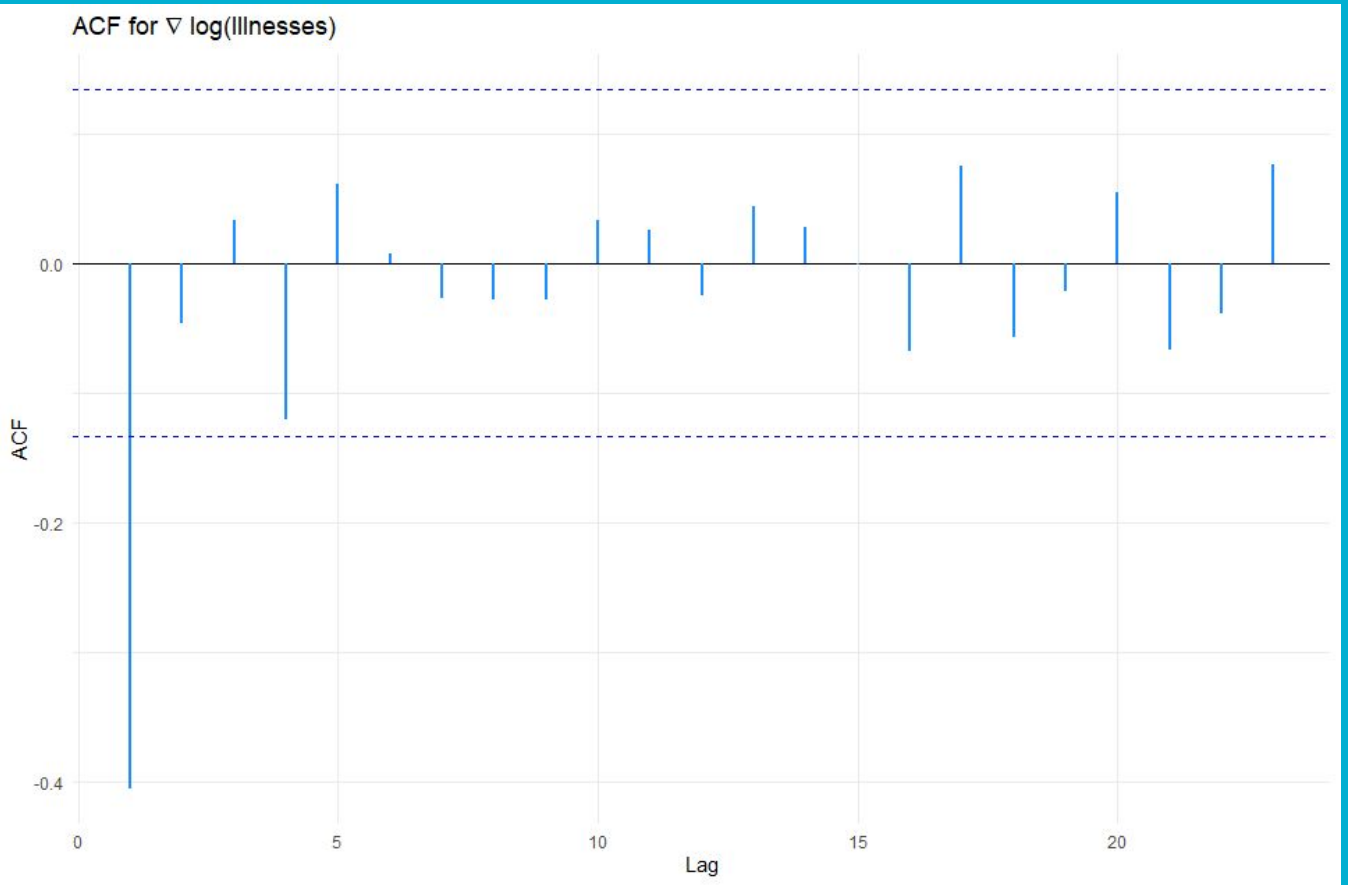
Transforming Data to Stationarity

- Log transformed to Normality
- First order differencing
  - Q-Q plot with both transformations shows linearity
- Augmented Dickey-Fuller test
  - p-value = 0.01 < 0.05
  - Transformed series is stationary
- KPSS test
  - p-value = 0.1, failed to reject Null Hypothesis
  - Transformed data is trend stationary



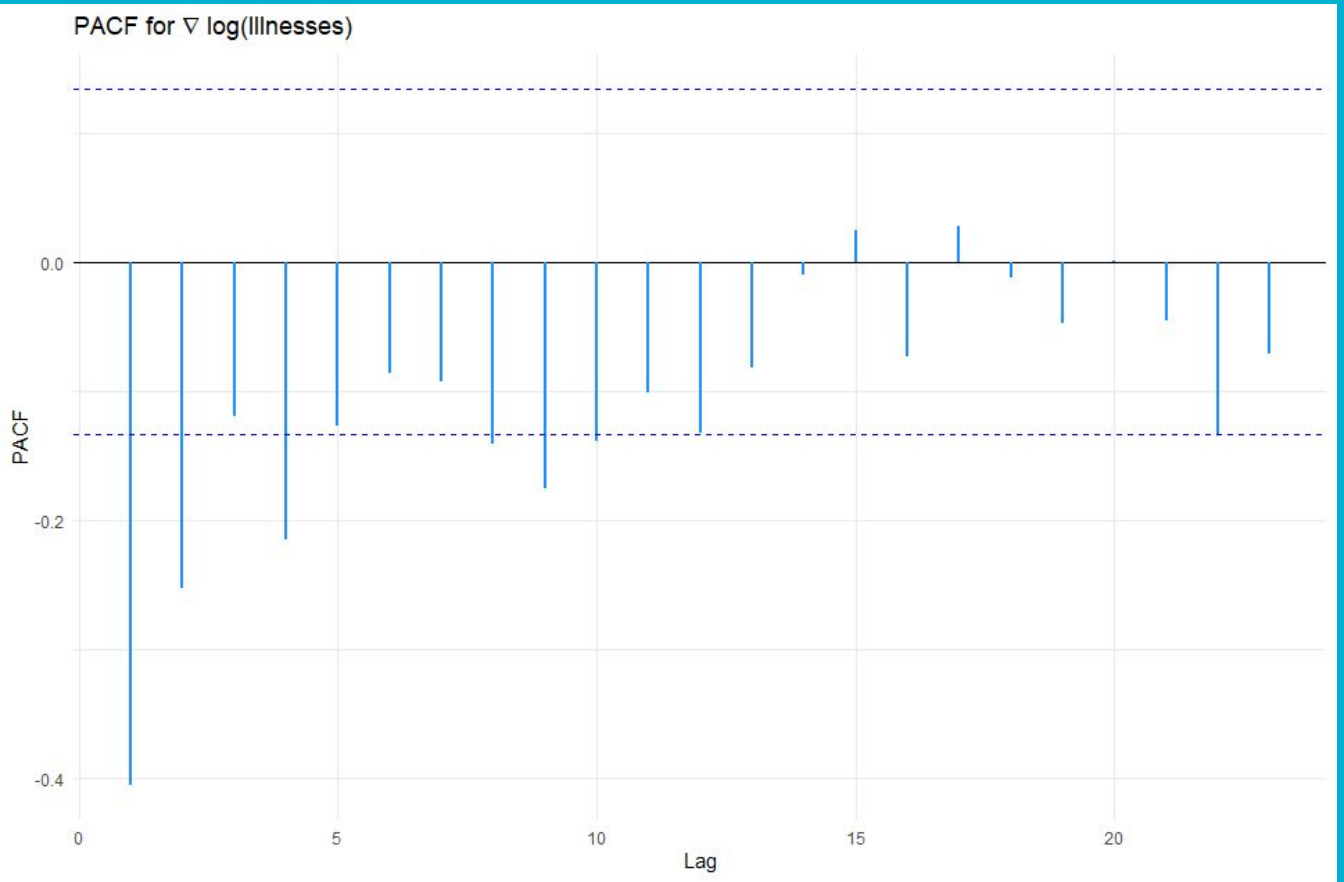


ACF



Only the first lag appears to be significant, while the ACF is not significantly different from zero afterwards.

PACF



The partial autocorrelogram appears to tail off over time. The first few lags are significant, and the magnitude of PACF decreases over time.

Model Selection Based on ACF and PACF

We choose MA(1) since only the ACF value for lag 1 was significant while the PACF had multiple significant values

Parameter Estimation of MA(1)

- MLE estimation (unconditional least squares) with the sarima function in the astsa library
- Using p = 0, d= 0 (using differenced log data) and q = 1 yielded best results
  - Values converged = > reasonable model
  - Conditional SS = -1.010483
  - Unconditional SS = -1.017639
  - MA(1) term significant
  - $\hat{\theta} = -.9103, \hat{\sigma}_w^2 = 0.1296$

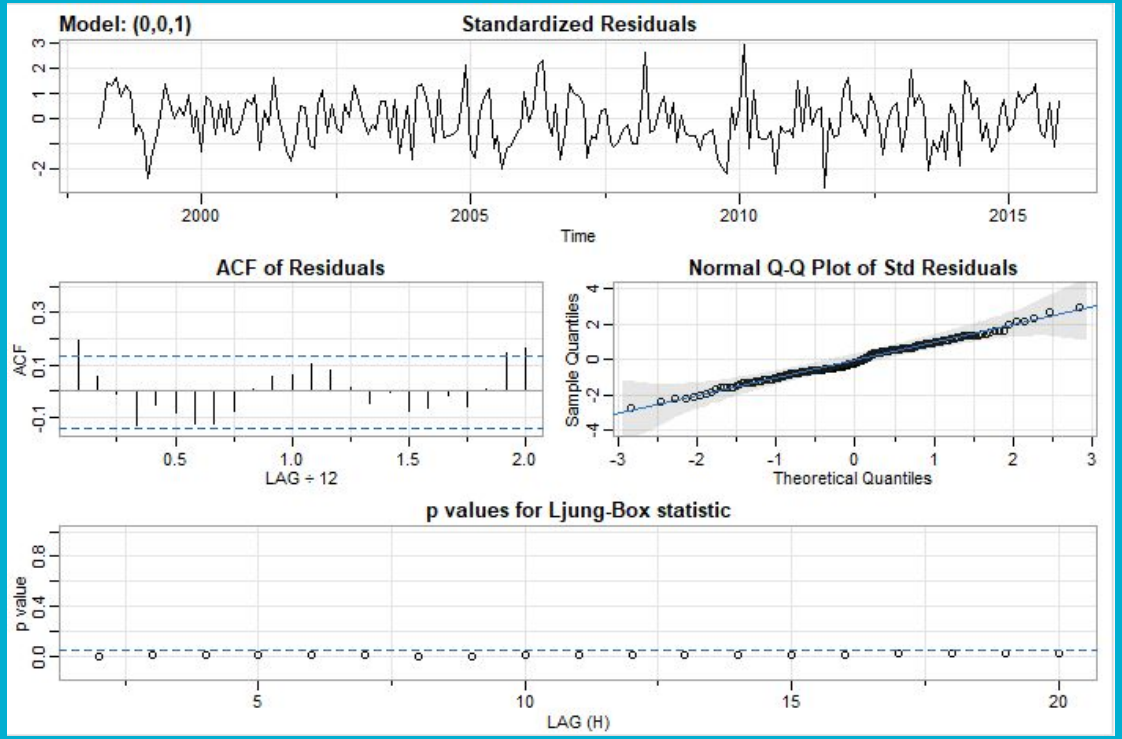
x_t = difference in log(Illnesses) for time t

$x_t = \omega_t - .9103_{(0.02)}\omega_{t-1}$

```
Coefficients:
          ma1
      -0.9103
s.e.   0.0290

sigma^2 estimated as 0.1296:  log likelihood = -86.28,  aic = 176.56

$degrees_of_freedom
[1] 214

$ttable
    Estimate     SE  t.value p.value
ma1  -0.9103 0.029 -31.3782       0

$AIC
[1] 0.8212041
```

Estimation Output of MA(1)

- Scattered, Normally distributed residuals
- ACF shows no departure from model assumptions
- p-values for Ljung Box are all > 0.05
    - Residuals are correlated!



Addition of AR(1) to Model

Based on the results of the Ljung Box plot, we reject the null hypothesis that the residuals are uncorrelated (Q-tests are all significant). Because our residuals are correlated, we know they are not white noise.

To alleviate this, we add an AR(1) term to the model, giving us an ARMA(1,1).

Parameter estimation of ARMA(1,1)

- MLE estimation (unconditional least squares) with the sarima function in the astsa library
- Using p = 1, d= 0 (using differenced log data) and q = 1 yielded best results
  - Values converged = > reasonable model
  - Conditional SS = -1.014162
  - Unconditional SS = -1.039521
  - AR(1) and MA(1) terms significant
  - Smaller AIC and BIC than AR(1) and MA(1) models
  - $\hat{\phi} = 0.2193, \hat{\theta} = -0.9351, \hat{\sigma}_w^2 = 0.1241$

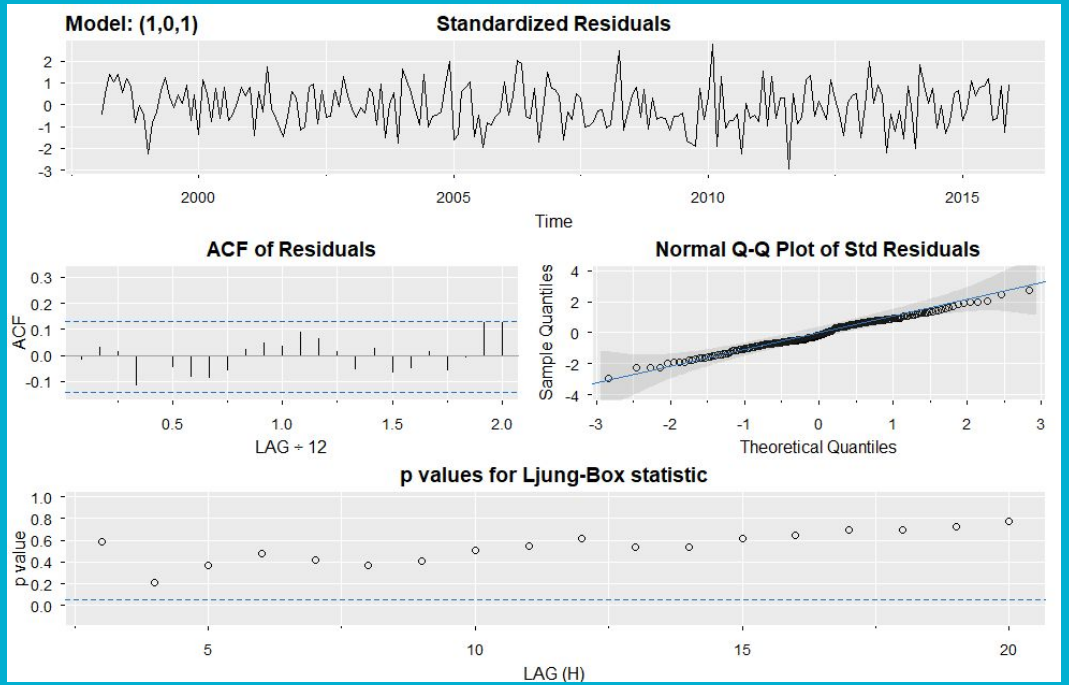
x_t = difference in log(Illnesses) for time t

$$x_t = 0.2193_{(0.07)} x_{t-1} - .9351_{(0.02)} \omega_{t-1} + \omega_t$$

```
Coefficients:
         ar1      ma1
      0.2193  -0.9351
s.e.  0.0710   0.0210

sigma^2 estimated as 0.1241:  log likelihood = -81.57,  aic = 169.15

$degrees_of_freedom
[1] 213

$ttable
    Estimate     SE  t.value p.value
ar1   0.2193 0.071   3.0912  0.0023
ma1  -0.9351 0.021 -44.6319  0.0000

$AIC
[1] 0.7867413
```

Estimation Output of ARMA(1,1)

- Scattered, Normally distributed residuals
- ACF shows no departure from model assumptions
- p-values for Ljung Box > 0.05
    - Residuals are uncorrelated

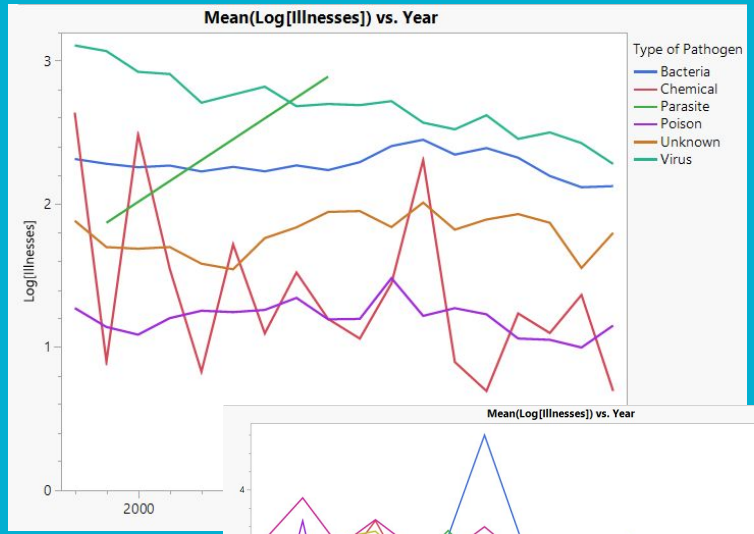


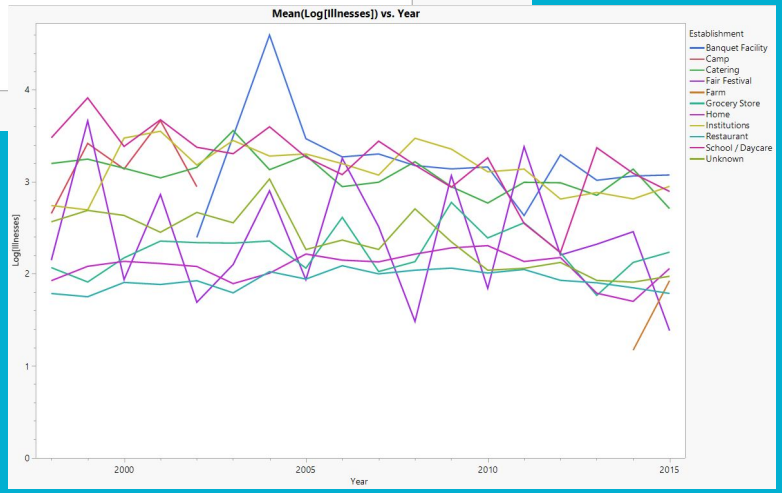
Future areas of research

- Trying to fit an ARIMA model (suggested in prior literature on foodborne illnesses)
- Analyze trends based on type of illness and establishment more closely
- Forecast data for years after 2015 and hopefully check against CDC data from that same period

Future Look Preview

- Statistical differences between Types
- Statistical Differences between Illness Source
- Interactions observed between type and illness Source





Questions, comments, suggestions?

Thank you for your engagement and feedback!