Analyzing Foodborne Disease Outbreaks Over Time

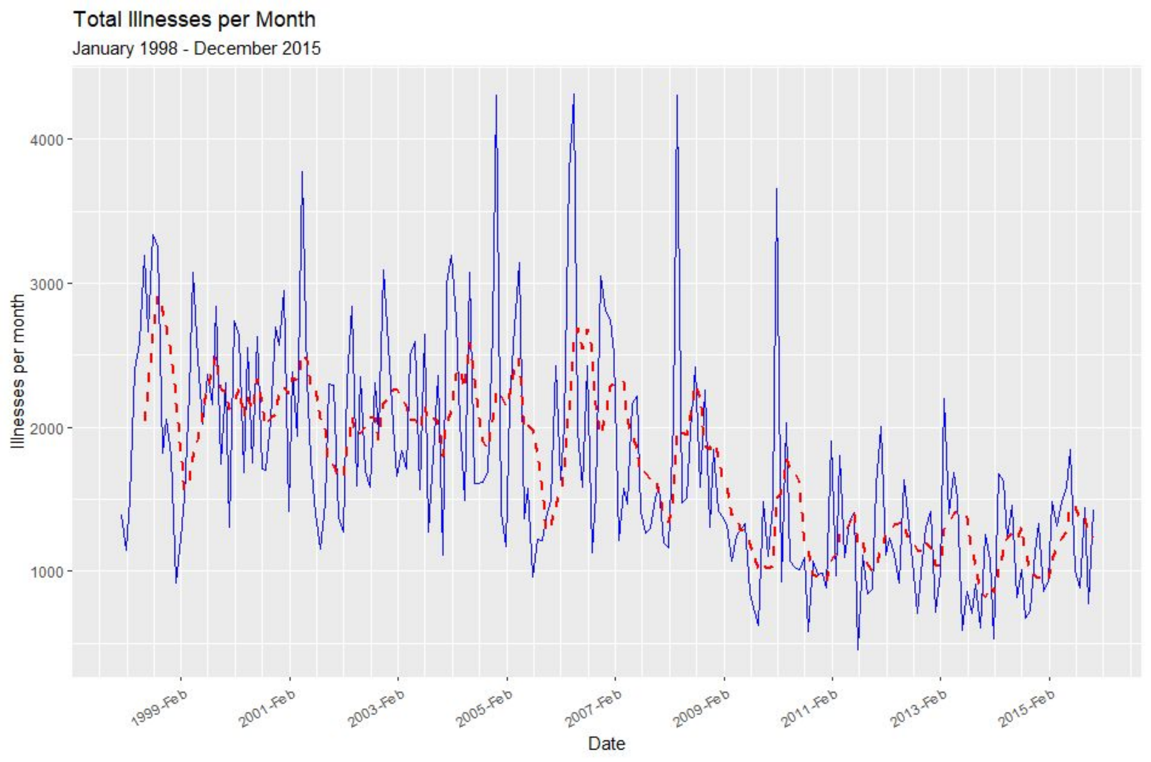
A look into illnesses and more

Group 12

- Trina Shores (group leader)
- Steven Macapagal
- Journey Martinez
- Yuan Yao
- Heather Nagy
- Kenneth Porter

Summary of prior findings

A reminder of what our data look like



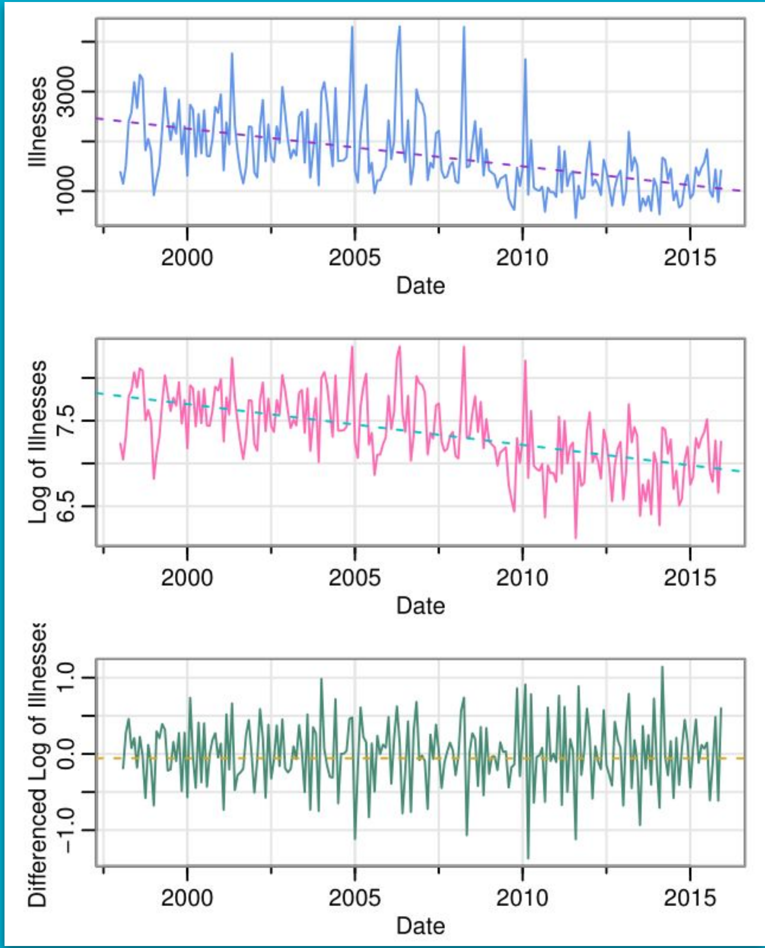
Solid blue line represents the original time series

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

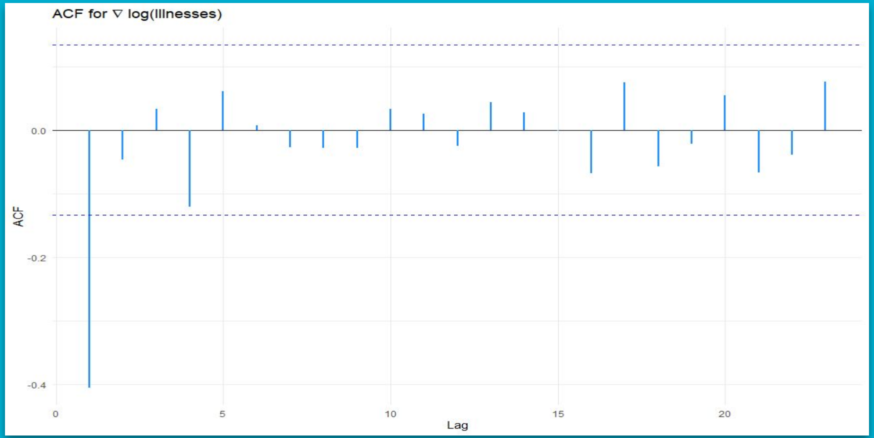
Stationarity

Variability is much greater from 1998 to 2010 and decreases from 2011 to 2015. Illnesses also seem to be trending downward and might be seasonal.

Log transforming and differencing achieved a more constant variance and got rid of the trend, giving us stationary data to work with.

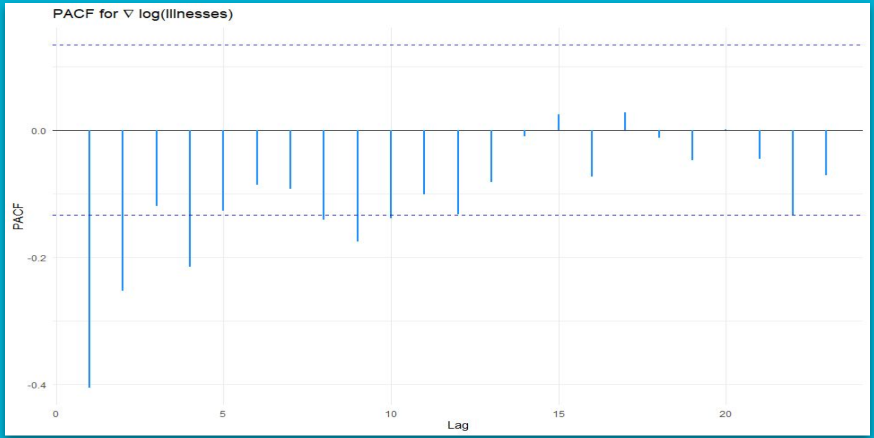


ACF and PACF



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

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



Parameter Estimation of MA(1)

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

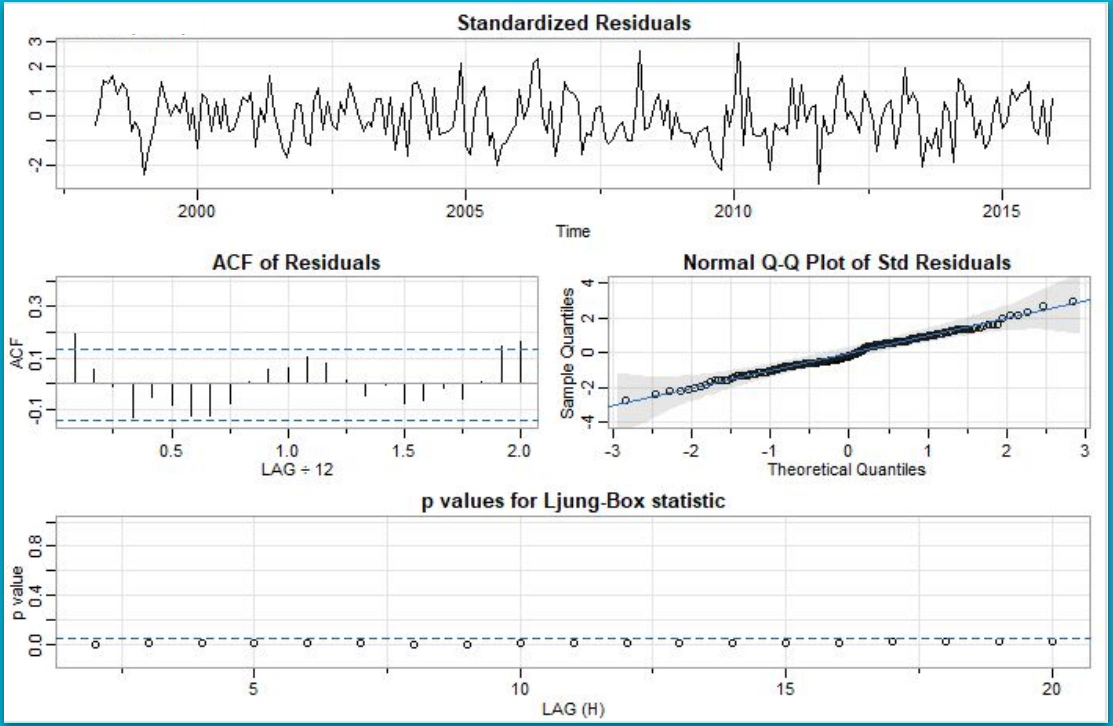
x_t = difference in log(Illnesses) for time t

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

```
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 Q tests are all less than 0.05
    - Residuals are correlated!



Parameter estimation of ARIMA(1,1,1)

- MLE estimation (unconditional least squares) with the sarima function in the astsa library
- Using p = 1, d= 1 (using 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
  - $\hat{\phi} = 0.2193, \hat{\theta} = -0.9351, \hat{\sigma}_w^2 = 0.1241$

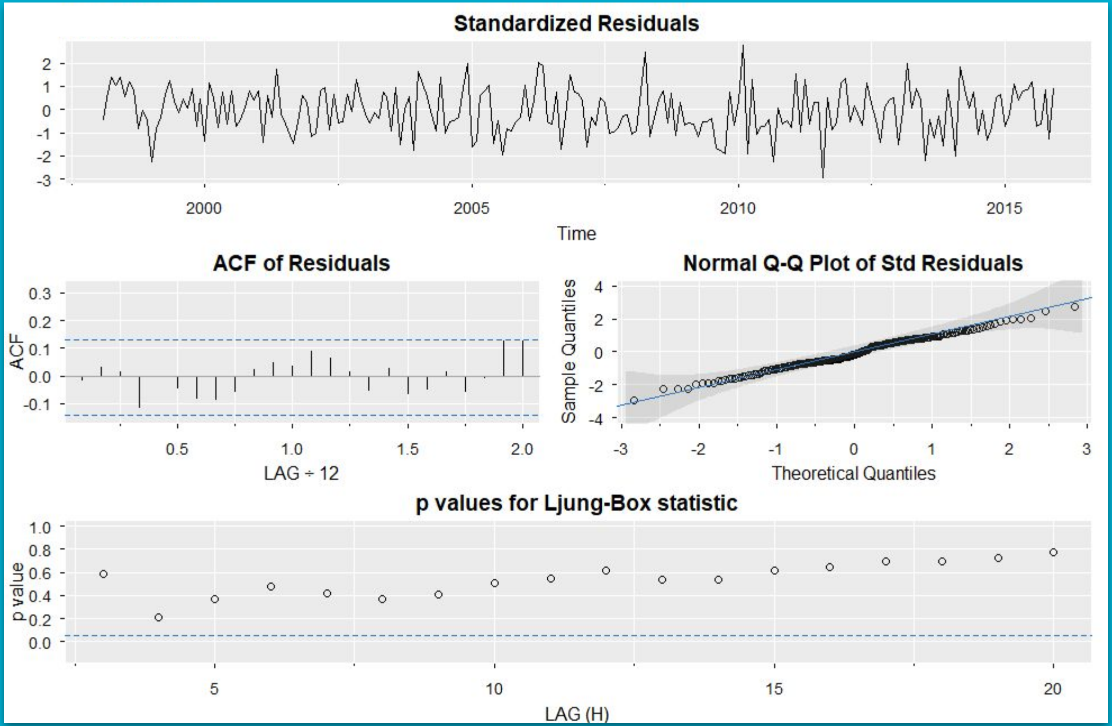
x_t = difference in log(Illnesses) for time t

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

```
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 ARIMA(1,1,1)

- Scattered, Normally distributed residuals
- ACF shows no departure from model assumptions
- p-values for Q tests all greater than 0.05
  - Residuals are uncorrelated

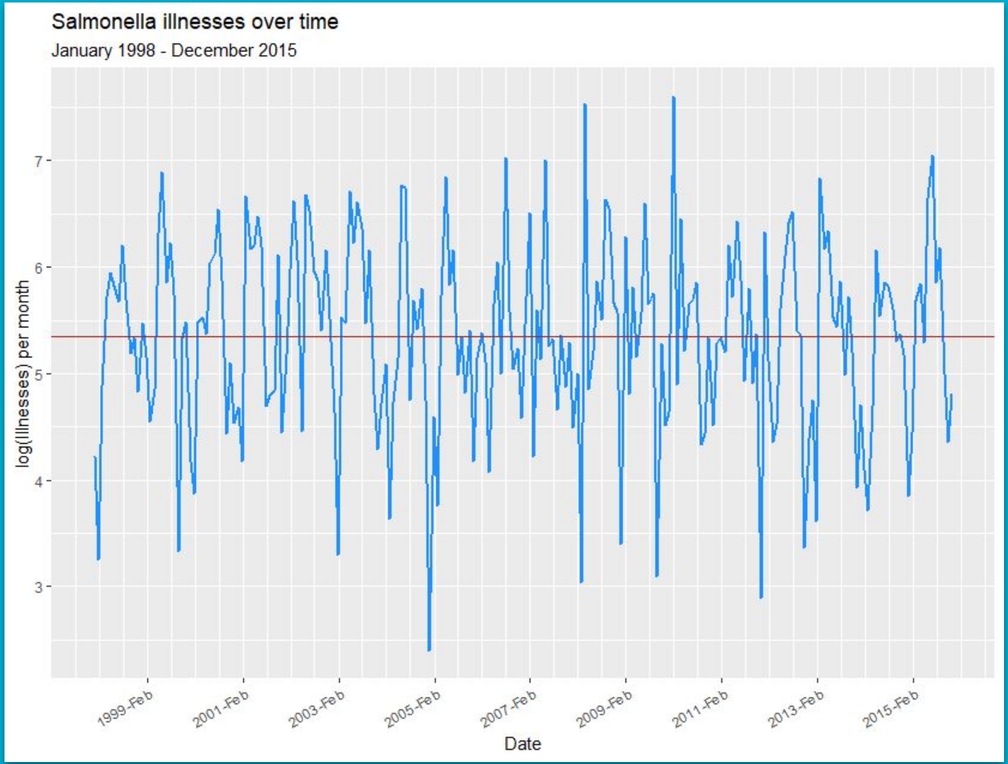


Our plan

1. Compare subsets of the data to see if there is an underlying pattern to foodborne disease outbreaks.
2. Compare models of foodborne illness to models of hospitalizations related to foodborne illness.
3. Compare different models to our previously established ARIMA(1, 1, 1) model on fit characteristics and prediction.

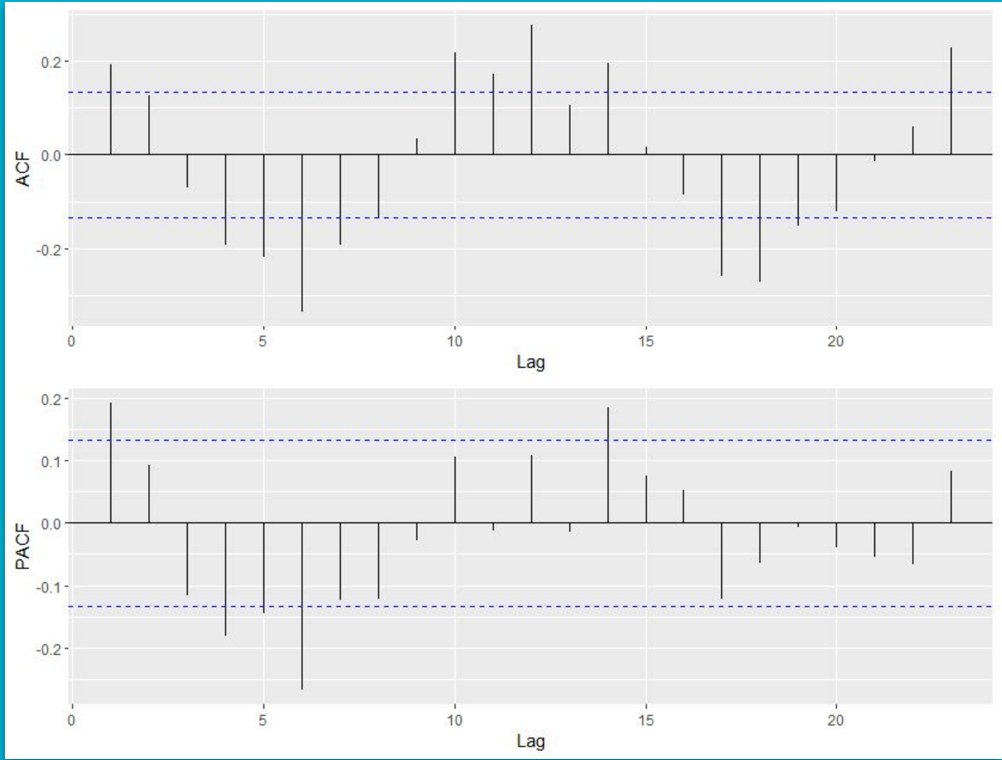
Analyzing source of outbreaks

Analysis of salmonella cases



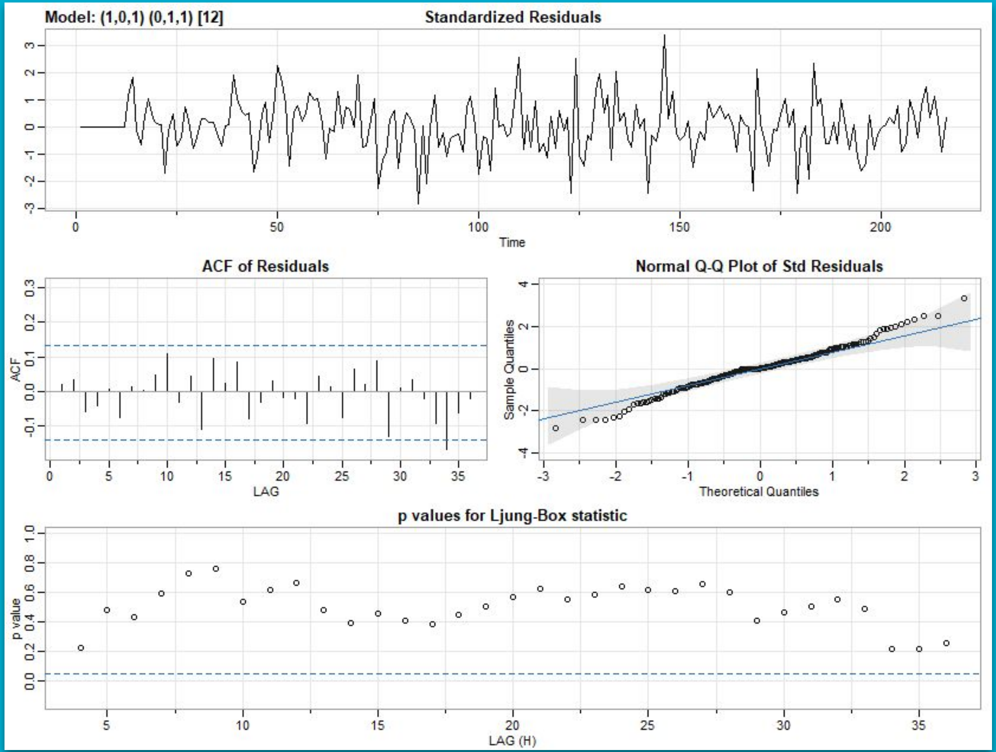
- Plot of log(illnesses) for salmonella cases appears to be stationary
- Seems to have a seasonal component

Analysis of salmonella cases



- Seasonal differencing needed: ACF is large and significant for every 12th lag, seems to tail off slowly
- ACF and PACF have significant 1st lags

Analysis of salmonella cases



SARIMA(1, 0, 1) x $(0, 1, 1)_{12}$

- Residuals appear to be normally distributed white noise
- P-values for Ljung-Box statistic appear to be nonsignificant

Analysis of salmonella cases

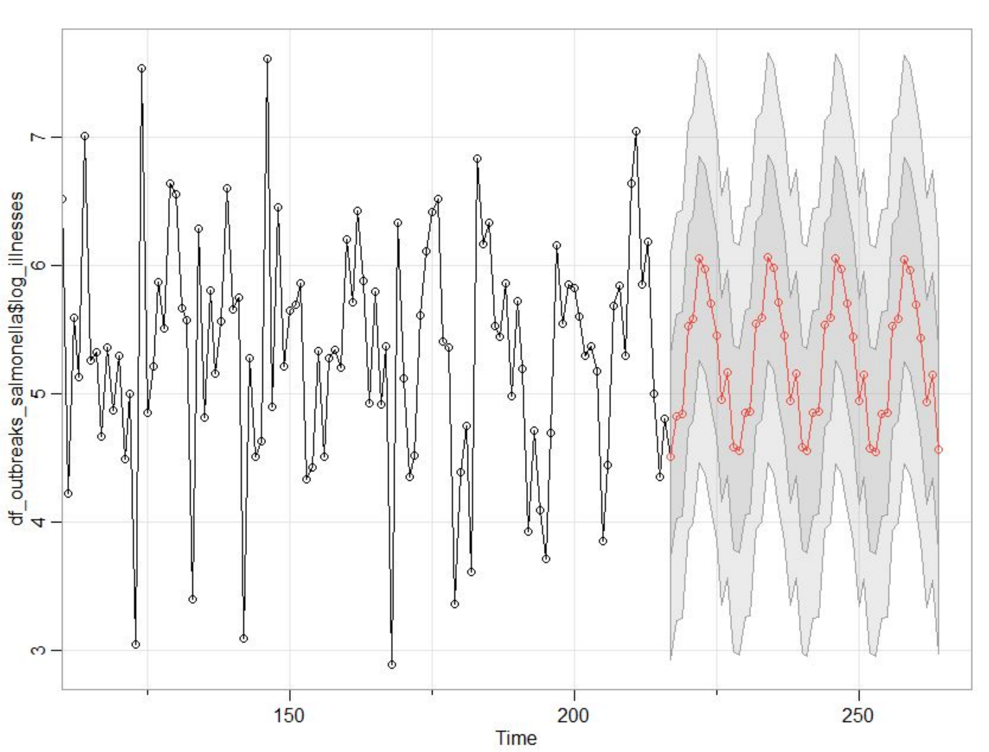
```
sigma^2 estimated as 0.5933:  log likelihood = -253.66,  aic = 517.33

$degrees_of_freedom
[1] 200

$ttable
         Estimate     SE t.value p.value
ar1        0.7605 0.1754  4.3361  0.0000
ma1       -0.8399 0.1487 -5.6491  0.0000
sma1      -0.9999 0.1044 -9.5770  0.0000
constant  -0.0006 0.0006 -1.0191  0.3094
```

SARIMA$(1, 0, 1) \times (0, 1, 1)_{12}$

Estimated coefficients:

$\varnothing_1 = 0.76$
$\theta_1 = -0.84$
$\Theta_1 = -1$
$\sigma^2 = 0.59$, df =200

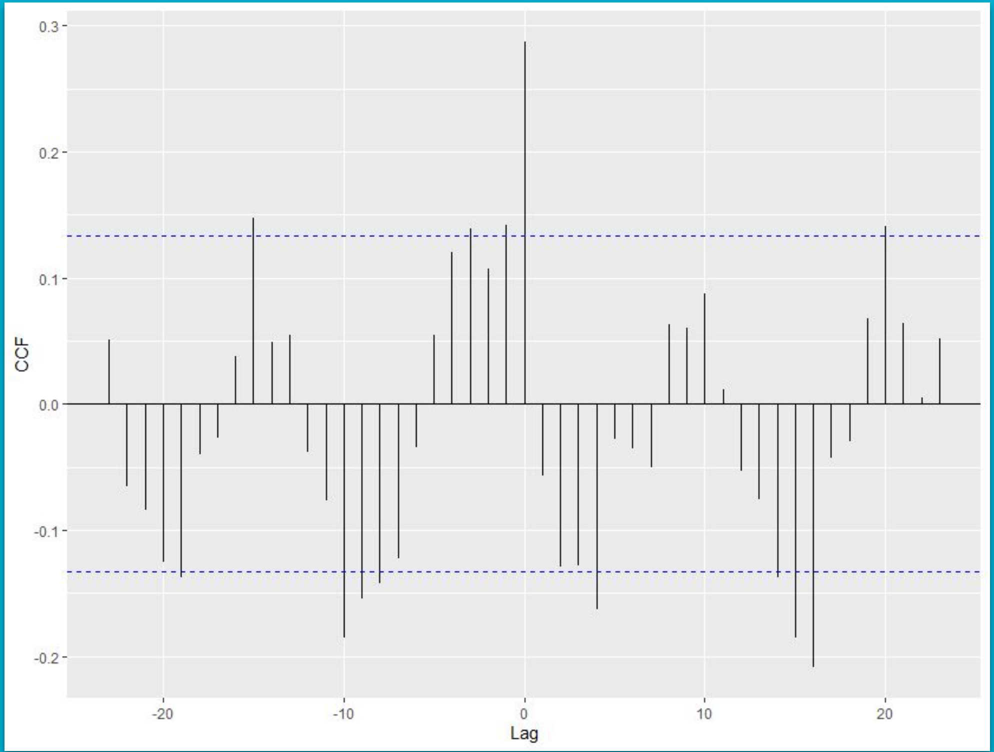
Analysis of salmonella cases



Seasonal forecasts of log(salmonella)

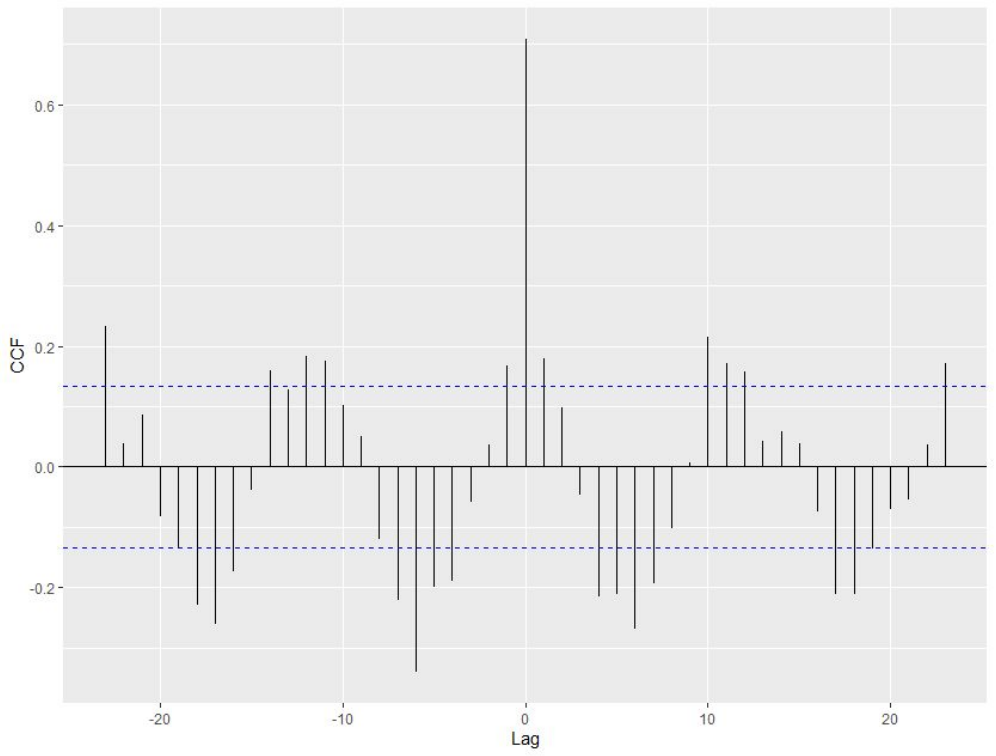
Relationships between illnesses and hospitalizations

CCF of overall illnesses and hospitalizations



- Strongest cross-correlation between illnesses and hospitalizations in the same period
- Appears to have some seasonality

Analysis of salmonella cases



CCF of log(illnesses) and log(hospitalizations)

- Seasonal patterns of illness still apparent
- Strongest cross-correlations in same period (hospitalizations follow a diagnosed illness closely)

Comparisons to baseline ARIMA(1, 1, 1) model

Steps

Modeling process:

1. Model formulation
2. Model estimation
3. Model diagnostics
4. Model selection

We will also compare their forecasts to the actual data released for 2016 through 2019.

Models:

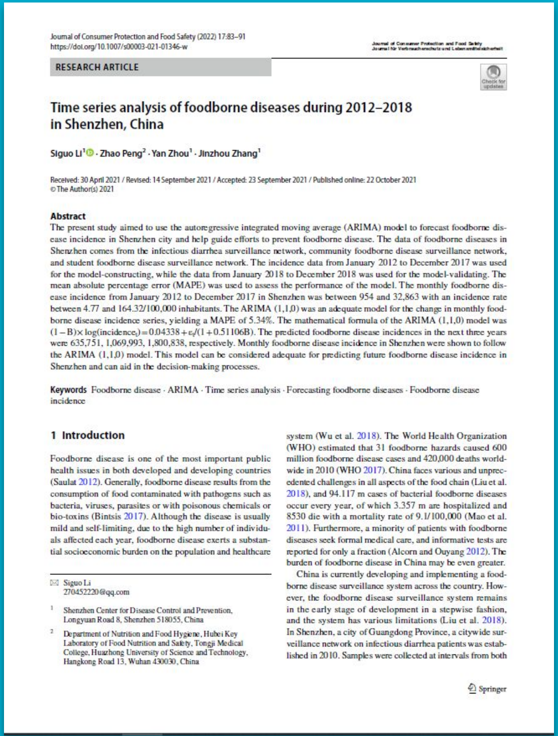
1. ARIMA(1, 1, 1) [baseline]
2. ARIMA(1, 1, 0) [Li et al. (2021)]
3. ARIMA(1, 1, 1) + GARCH(1, 0)
4. ARIMA(1, 1, 1) x $(1, 0, 1)_{12}$
5. ARIMA(1, 1, 1) x $(0, 1, 1)_{12}$
6. Prophet *(examined later)*

Comparing models suggested in prior literature

Li, et al. (2021) suggested an ARIMA(1, 1, 0) model to describe the incidence of foodborne illnesses over time.

- Performance and prediction compared to ARIMA(1, 1, 1) model we had previously found to fit our data

Li, S., Peng, Z., Zhou, Y., & Zhang, J. (2021). Time series analysis of foodborne diseases during 2012-2018 in Shenzhen, China. *Journal of Consumer Protection and Food Safety, 17*(2), 83-91.



Journal of Consumer Protection and Food Safety (2022) 17:83–91
https://doi.org/10.1007/s00003-021-01346-w

Journal of Consumer Protection and Food Safety
Journal für Verbraucherschutz und Lebensmittelsicherheit

RESEARCH ARTICLE

Check for updates

Time series analysis of foodborne diseases during 2012–2018 in Shenzhen, China

Siguo Li[1] · Zhao Peng[2] · Yan Zhou[1] · Jinzhou Zhang[1]

Received: 30 April 2021 / Revised: 14 September 2021 / Accepted: 23 September 2021 / Published online: 22 October 2021

Abstract
The present study aimed to use the autoregressive integrated moving average (ARIMA) model to forecast foodborne disease incidence in Shenzhen city and help guide efforts to prevent foodborne disease. The data of foodborne diseases in Shenzhen comes from the infectious diarrhea surveillance network, community foodborne disease surveillance network, and student foodborne disease surveillance network. The incidence data from January 2012 to December 2017 was used for the model-constructing, while the data from January 2018 to December 2018 was used for the model-validating. The mean absolute percentage error (MAPE) was used to assess the performance of the model. The monthly foodborne disease incidence from January 2012 to December 2017 in Shenzhen was between 954 and 32,863 with an incidence rate between 4.77 and 164.32/100,000 inhabitants. The ARIMA (1,1,0) was an adequate model for the change in monthly foodborne disease incidence series, yielding a MAPE of 5.34%. The mathematical formula of the ARIMA (1,1,0) model was $(1-B)\times \log(\text{incidence}_t) = 0.04338 + \varepsilon_t/(1+0.51106B)$. The predicted foodborne disease incidences in the next three years were 635,751, 1,069,993, 1,800,838, respectively. Monthly foodborne disease incidence in Shenzhen were shown to follow the ARIMA (1,1,0) model. This model can be considered adequate for predicting future foodborne disease incidence in Shenzhen and can aid in the decision-making processes.

Keywords Foodborne disease · ARIMA · Time series analysis · Forecasting foodborne diseases · Foodborne disease incidence

1 Introduction

Foodborne disease is one of the most important public health issues in both developed and developing countries (Saulat 2012). Generally, foodborne disease results from the consumption of food contaminated with pathogens such as bacteria, viruses, parasites or with poisonous chemicals or bio-toxins (Bintsis 2017). Although the disease is usually mild and self-limiting, due to the high number of individuals affected each year, foodborne disease exerts a substantial socioeconomic burden on the population and healthcare system (Wu et al. 2018). The World Health Organization (WHO) estimated that 31 foodborne hazards caused 600 million foodborne disease cases and 420,000 deaths worldwide in 2010 (WHO 2017). China faces various and unprecedented challenges in all aspects of the food chain (Liu et al. 2018), and 94.117 m cases of bacterial foodborne diseases occur every year, of which 3.357 m are hospitalized and 8530 die with a mortality rate of 9.1/100,000 (Mao et al. 2011). Furthermore, a minority of patients with foodborne diseases seek formal medical care, and informative tests are reported for only a fraction (Alcorn and Ouyang 2012). The burden of foodborne disease in China may be even greater.

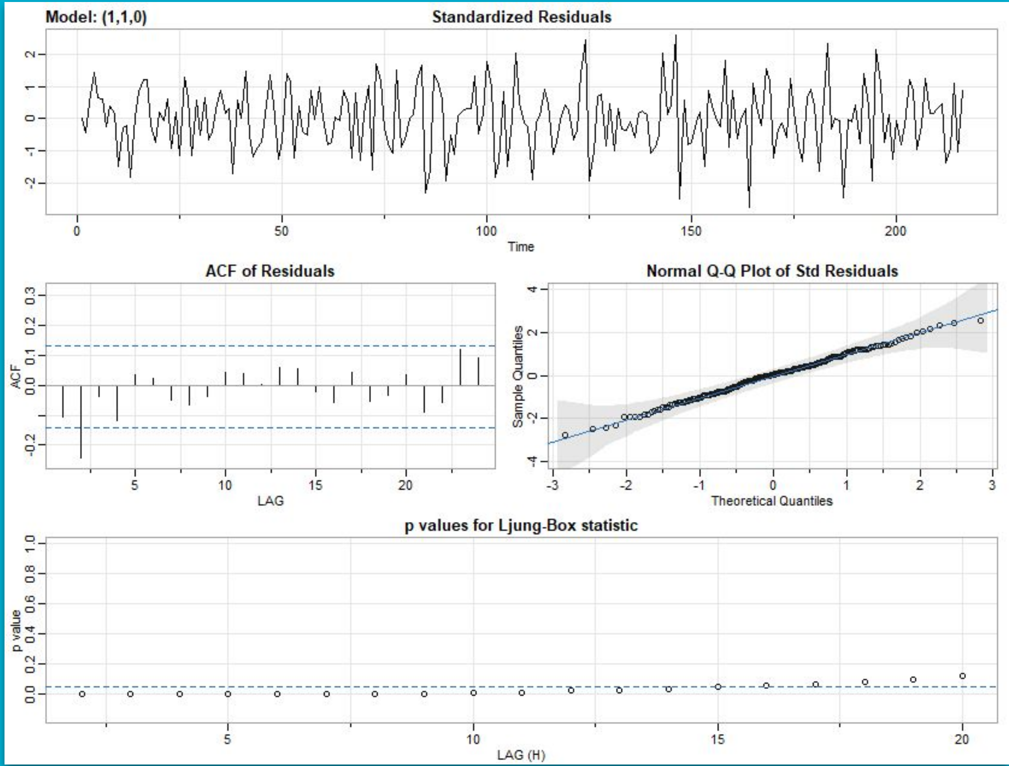
China is currently developing and implementing a foodborne disease surveillance system across the country. However, the foodborne disease surveillance system remains in the early stage of development in a stepwise fashion, and the system has various limitations (Liu et al. 2018). In Shenzhen, a city of Guangdong Province, a citywide surveillance network on infectious diarrhea patients was established in 2010. Samples were collected at intervals from both

✉ Siguo Li
270452220@qq.com

1 Shenzhen Center for Disease Control and Prevention, Longyuan Road 8, Shenzhen 518055, China

2 Department of Nutrition and Food Hygiene, Hubei Key Laboratory of Food Nutrition and Safety, Tongji Medical College, Huazhong University of Science and Technology, Hangkong Road 13, Wuhan 430030, China

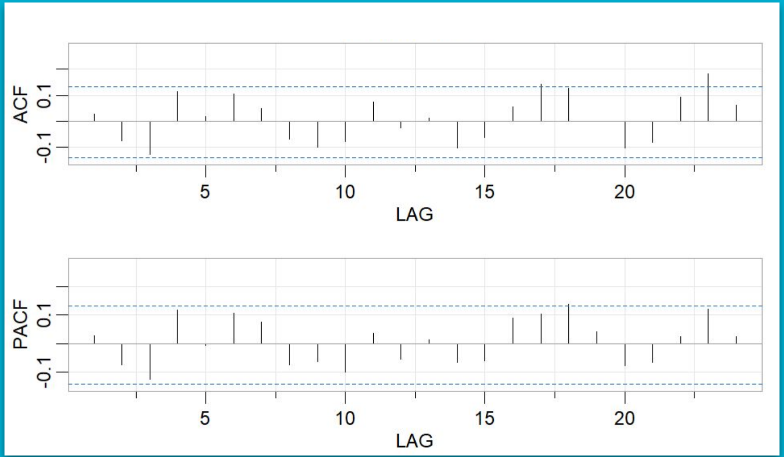
ARIMA(1, 1, 0) [Li et al., 2021]



- Diagnostics show residuals are not white noise, still have some autocorrelation
- Estimated coefficients: $\varnothing_1$ = -0.41 (0.06) σ^2 = 0.16, df = 214
- AIC = 1.02 BIC = 1.05

ARIMA(1,1,1) + GARCH(1,0)

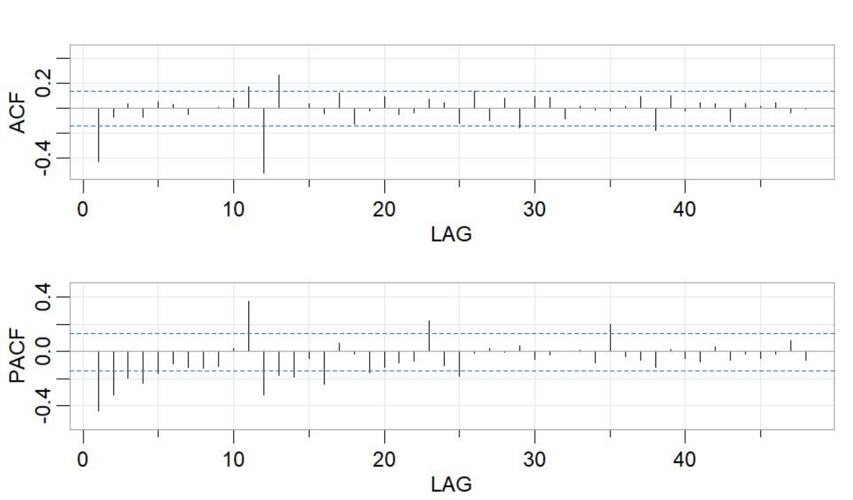
- Some small dependence left in ARIMA(1,1,1) squared residuals
    - Average of residuals = 0.012
- ARIMA(1,1,1) + GARCH(1,0) on transformed data does not have significant alpha term
- GARCH(1,0) not needed

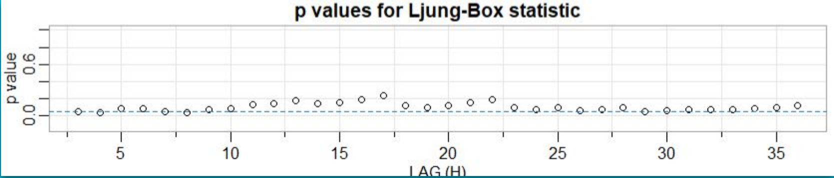



```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu       3.65873     0.46265    7.908 2.66e-15 ***
ar1      0.50357     0.06260    8.044 8.88e-16 ***
omega    0.14864     0.02054    7.237 4.57e-13 ***
alpha1   0.01944     0.08138    0.239   0.8112
shape   10.00000     3.93673    2.540   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log Likelihood:
 -97.76341    normalized:  -0.4526084
```

Seasonality

- There appears to be a seasonal component
  - ACF cuts off at lag 12, PACF tails off
  - Try adding SMA1 term
- ARIMA(1,1,1)x(0,1,1)$_{12}$ is possible
  - However, AR1 is not significant
- ARIMA(0,1,1)x(0,1,1) did not have white noise residuals
- ARIMA(1,1,1)x(1,1,1)$_{12}$
  - AR1 and SAR1 terms not significant





$ttable

	Estimate	SE	t.value	p.value
ar1	0.1371	0.0752	1.8244	0.0696
ma1	-0.9393	0.0290	-32.3892	0.0000
sma1	-0.9093	0.0857	-10.6136	0.0000

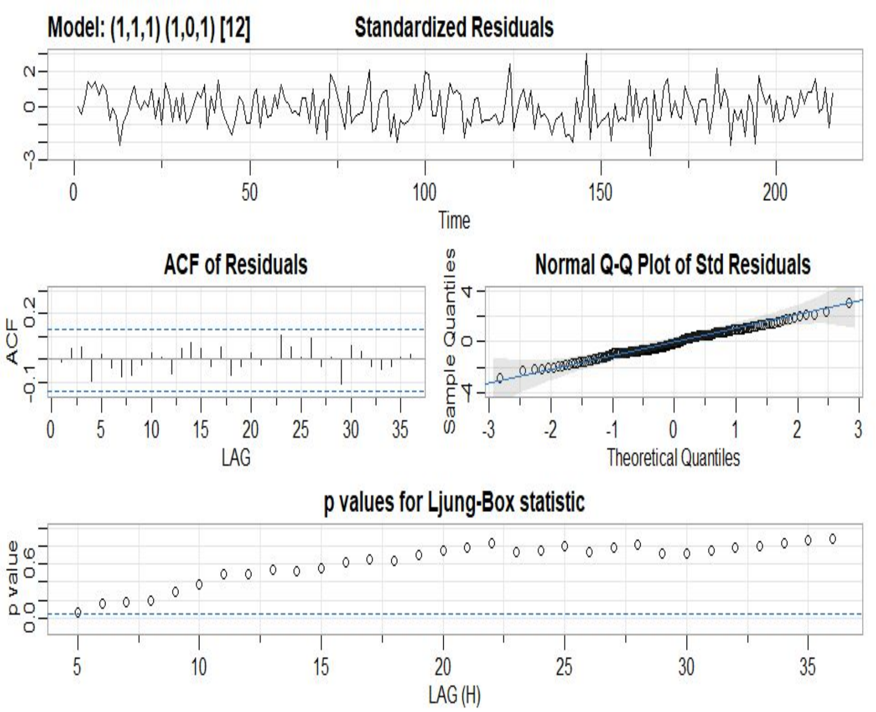
$ttable

	Estimate	SE	t.value	p.value
ar1	0.1275	0.0764	1.6691	0.0967
ma1	-0.9342	0.0303	-30.7925	0.0000
sar1	-0.0858	0.0917	-0.9352	0.3508
sma1	-0.8572	0.0925	-9.2709	0.0000

Selecting seasonal component

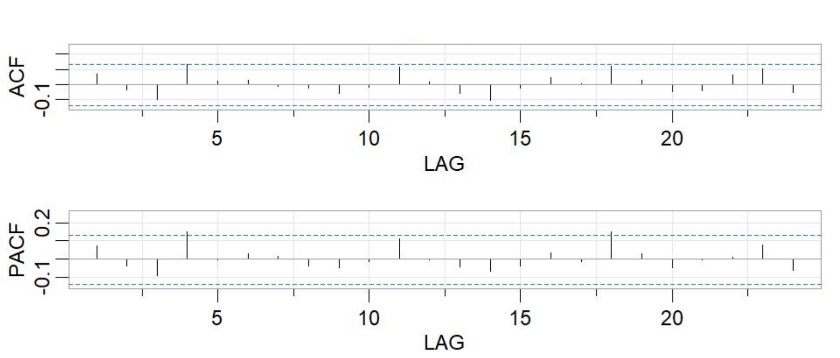
- ARIMA(1,1,1)x(1,0,1)$_{12}$ has significant estimates
- Slightly smaller AIC compared to ARIMA(1,1,1)
- Residuals appear to be white noise
- AIC = 0.769, BIC = 0.847
    - AIC =0.787 , BIC = 0.834 for baseline ARIMA

	Estimate	SE	t.value	p.value
ar1	0.1692	0.0729	2.3201	0.0213
ma1	-0.9403	0.0226	-41.5258	0.0000
sar1	0.9509	0.0912	10.4234	0.0000
sma1	-0.8796	0.1443	-6.0948	0.0000



ARIMA(1,1,1) x (1,0,1)$_{12}$

- Less dependence in squared residuals compared to ARIMA(1,1,1)
- Mean(residuals) = 0.006
  - Closer to zero than mean of ARIMA(1,1,1) residuals
- Less biased forecast



	Estimate	SE	t.value	p.value
ar1	0.1692	0.0729	2.3201	0.0213
ma1	-0.9403	0.0226	-41.5258	0.0000
sar1	0.9509	0.0912	10.4234	0.0000
sma1	-0.8796	0.1443	-6.0948	0.0000

$$x_t = (1+\phi)x_{t-1} - \phi x_{t-2} + \Phi x_{t-12} - \Phi(\phi+1)x_{t-13} + \Phi\phi x_{t-14} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}$$

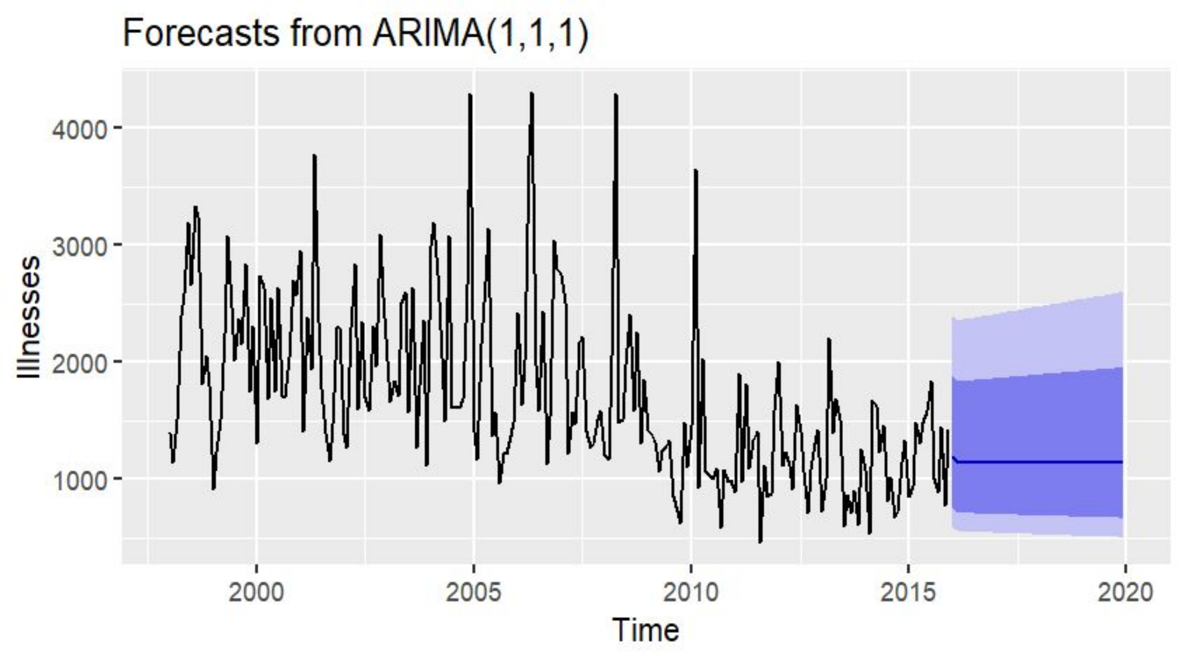
$$x_t = 1.1692x_{t-1} - 0.1692x_{t-2} + 0.9509x_{t-12} - 1.1118x_{t-13} + 0.1608x_{t-14} + w_t - 0.9403w_{t-1} - 0.8796w_{t-12} + 0.8271w_{t-13}$$

Models/Forecasts comparison

Model Comparisons

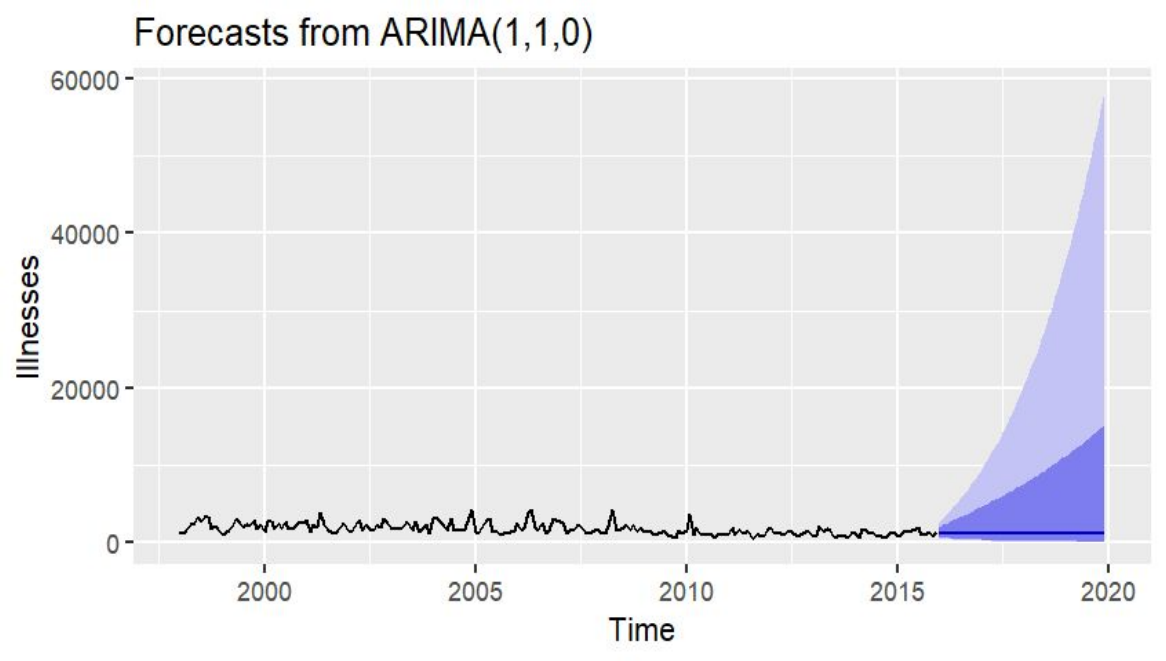
Model	AIC	BIC
ARIMA(1, 1, 1)	0.787	0.834
ARIMA(1, 1, 0)	1.020	1.051
$ARIMA(1,1,1)x(1,0,1)_{12}$	0.769	0.847
$ARIMA(1,1,1)x(0,1,1)_{12}$	0.881	0.946
ARIMA(1,1,1)-GARCH(1,0)	0.926	0.989

Forecast for ARIMA(1, 1, 1)



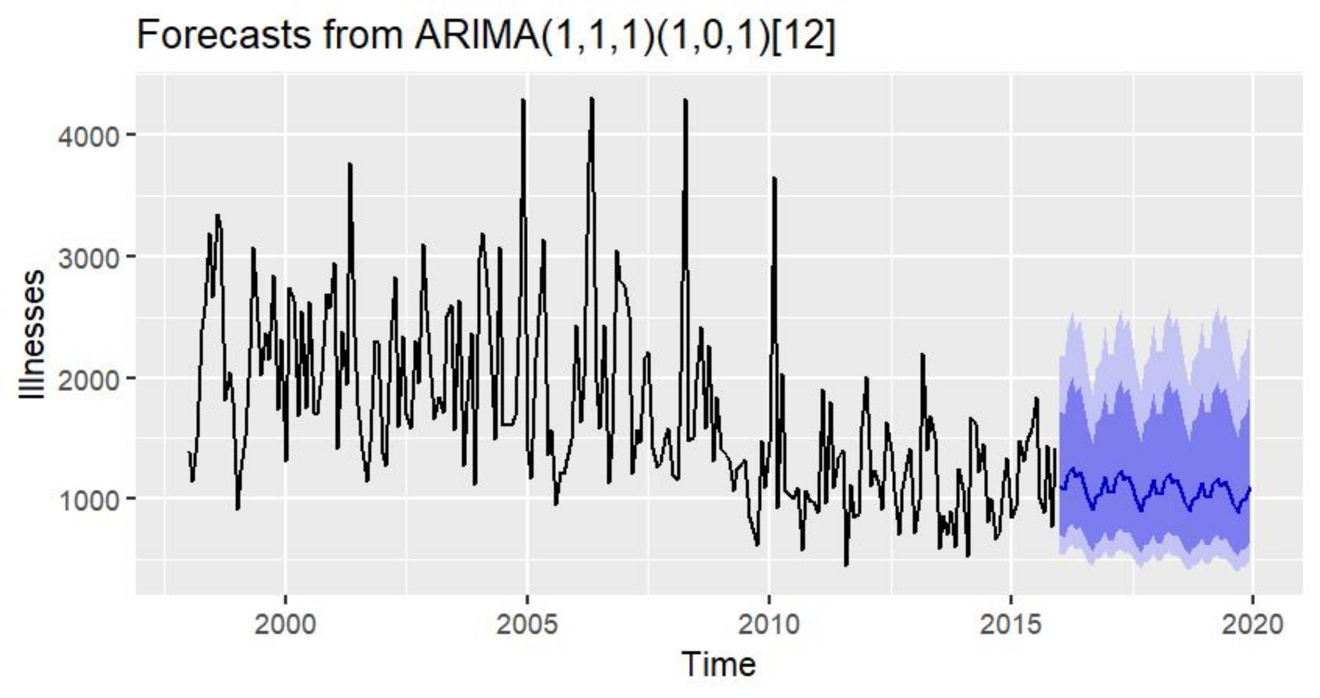
- The model mean's forecast appears to match mean of data
- But it doesn't appear to capture seasonal volatility well

Forecast for ARIMA(1, 1, 0)



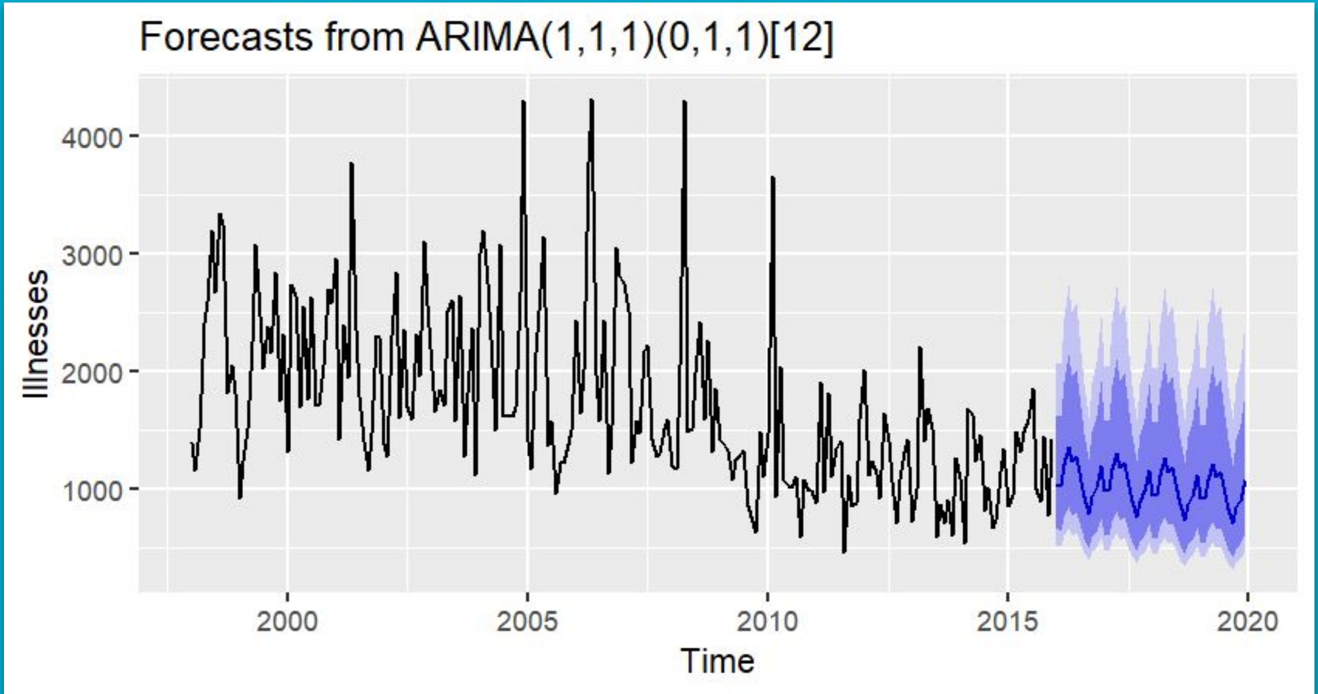
- Clearly not an ideal forecast due to the high upper CI bound
- However, the mean forecast appears to be accurate

Forecast for ARIMA(1,1,1)x(1,0,1)$_{12}$



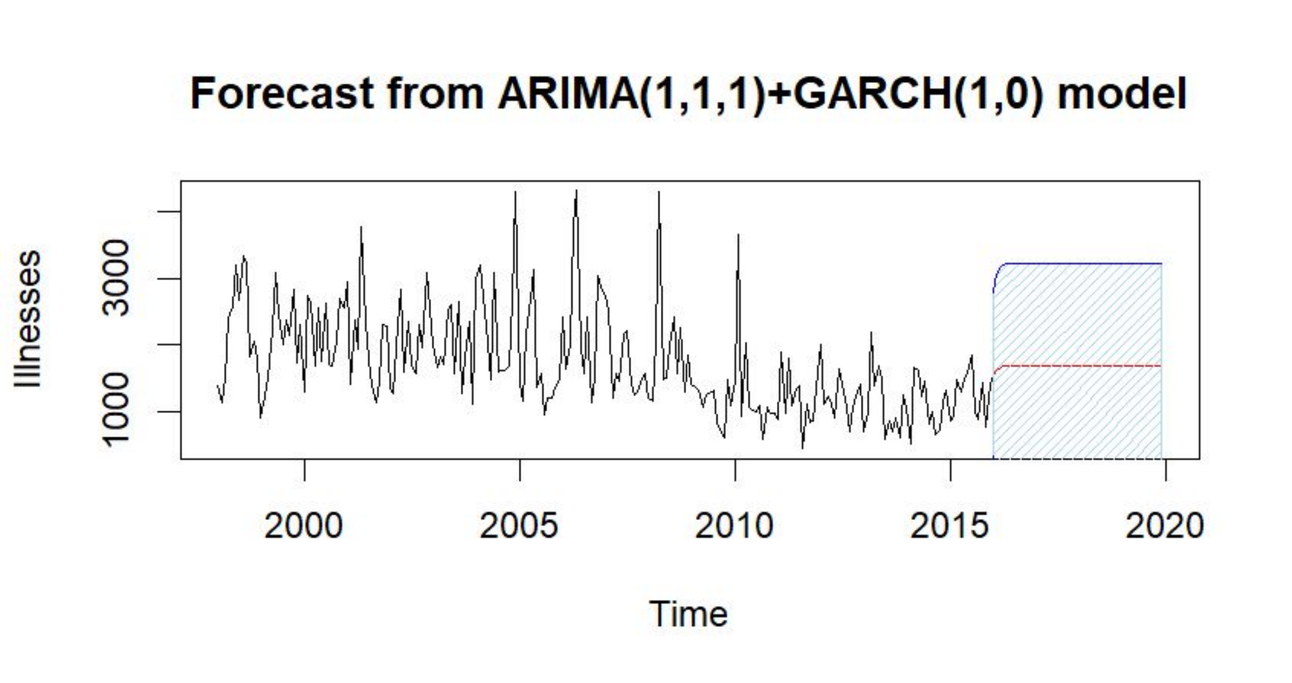
- Good performance for forecasting compared to actuals

Forecast for ARIMA(1,1,1)x(0,1,1)12



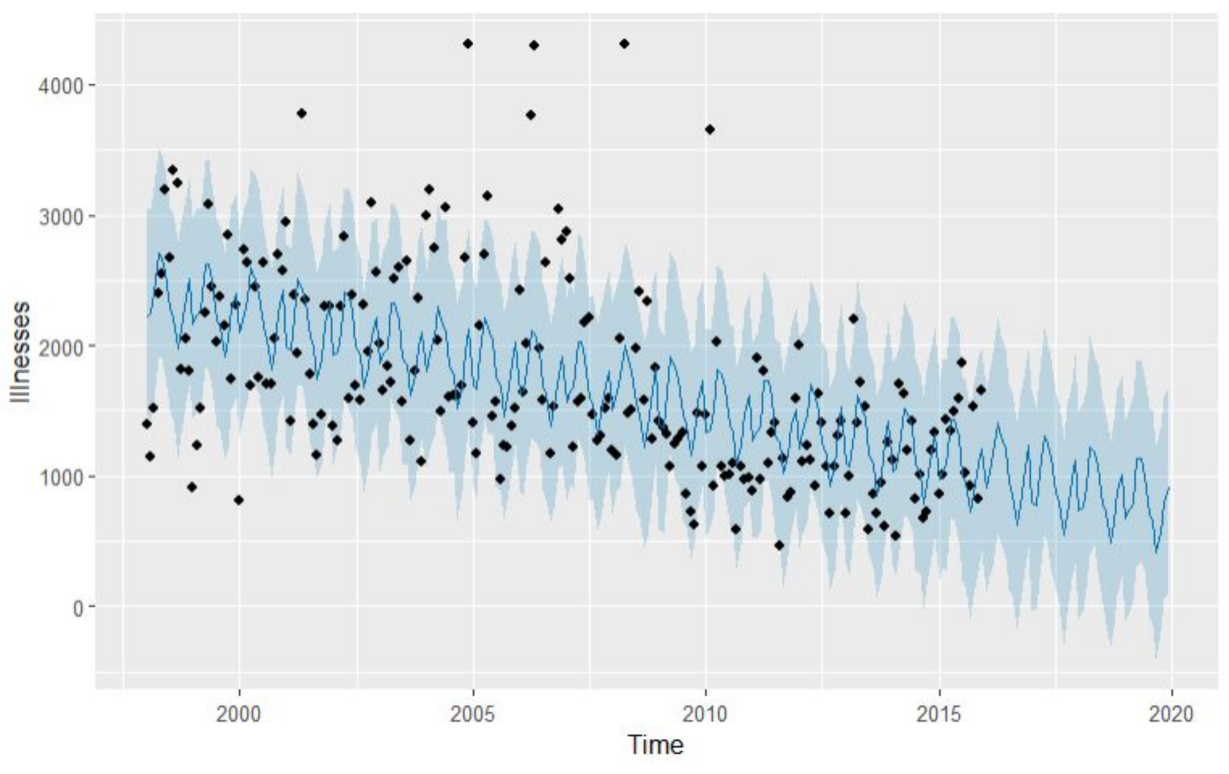
- This model appears to capture more of volatility
- This model did not have a significant AR1 term

Forecast for ARIMA(1,1,1) + GARCH(1,0)



- Like the non seasonal ARIMA models but the lower CI limit is 0
- Similar pattern for forecasting using log data vs untransformed data

Forecast for Prophet



- The model forecast appears to do a good job of capturing the seasonality of the historical data
- The forecast also appears to have a slight negative trend

Forecasting Comparisons

Model	RMSE
ARIMA(1, 1, 1)	489.42
ARIMA(1, 1, 0)	474.94
$ARIMA(1,1,1)x(1,0,1)_{12}$	488.63
$ARIMA(1,1,1)x(0,1,1)_{12}$	487.91
ARIMA(1,1,1)-GARCH(1,0)	552.43
Prophet	638

Data from 2016 and beyond

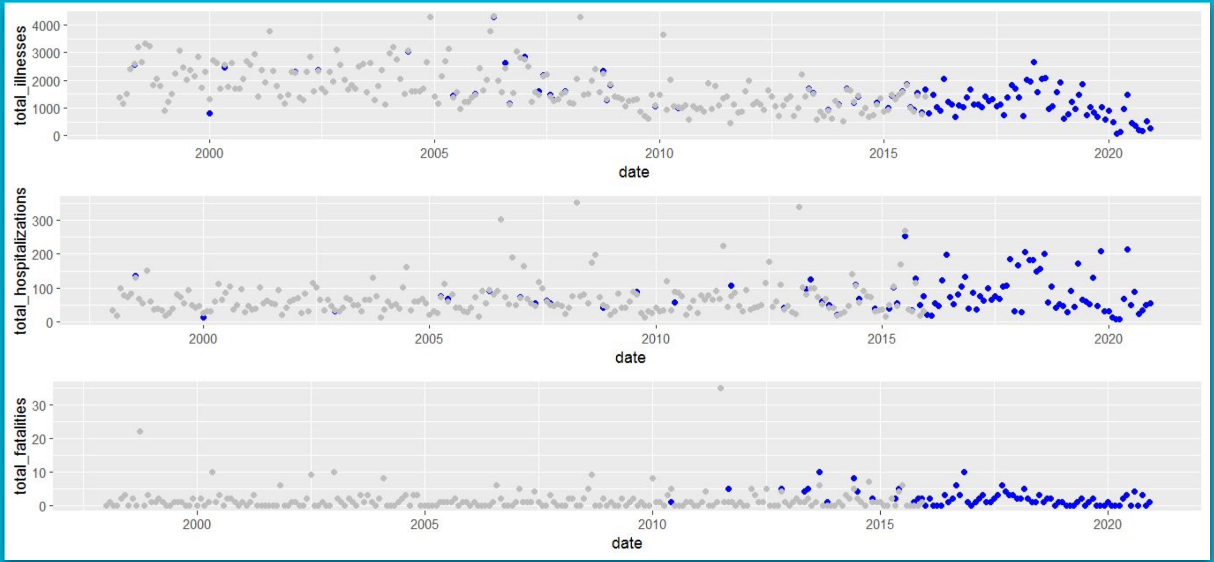
Data1: Foodborne illness data set (1998 - 2015):

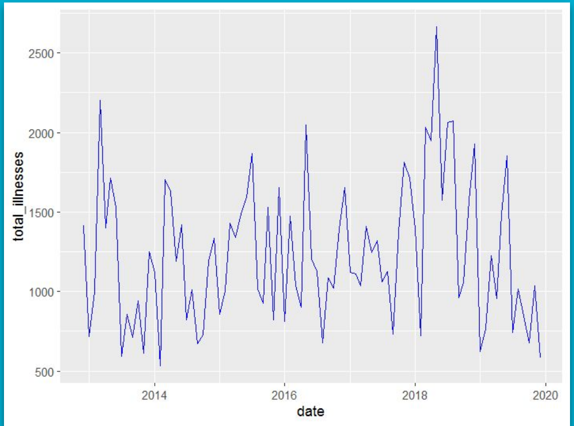
https://www.kaggle.com/datasets/cdc/foodborne-diseases

Data2: National Outbreak Reporting System(1998 - 2020)

https://wwwn.cdc.gov/norsdashboard

Validate 2016 and beyond data





2013 - 2020 illness from data 2
Choose 2016 - 2019(48 month)
For forecast comparing

Real data and forecasting of ARIMA(1,1,1)

Model output from data before 2016

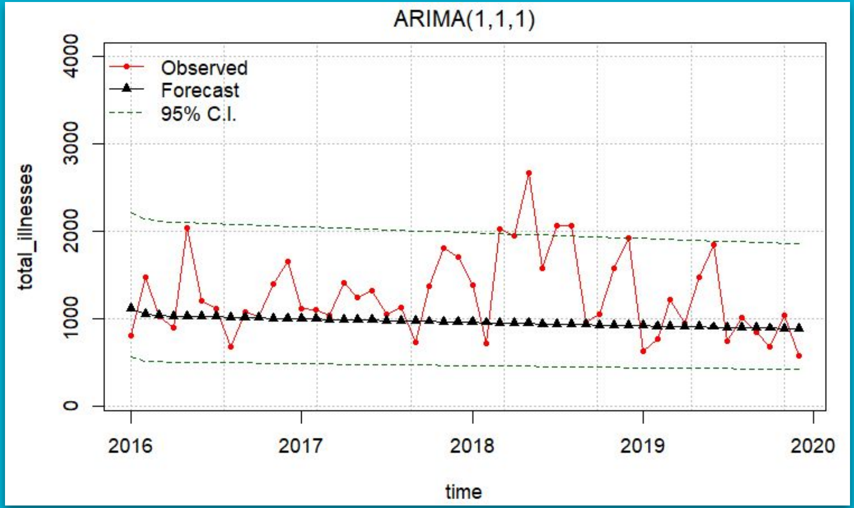
Data 1

```
$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
```

Data 2

```
$ttable
    Estimate     SE  t.value p.value
ar1   0.1998 0.0714   2.7971  0.0056
ma1  -0.9337 0.0216 -43.2306  0.0000
```

Forecast for data between 2016 - 2019



Real data and forecasting of ARIMA(1,1,0)

Model output from data before 2016

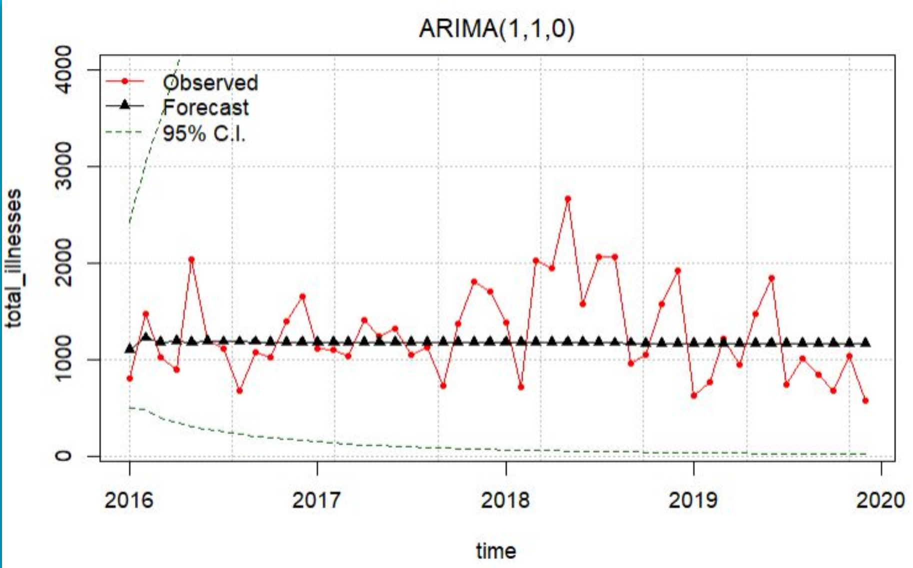
Data 1

```
$ttable
     Estimate      SE t.value p.value
ar1  -0.4078 0.0624  -6.535       0
```

Data 2

```
$ttable
     Estimate      SE t.value p.value
ar1  -0.4171 0.0622 -6.7047       0
```

Forecast for data between 2016 - 2019



Real data and forecasting of SARIMA(1,1,1)x(1,0,1)$_{12}$

Model output from data before 2016

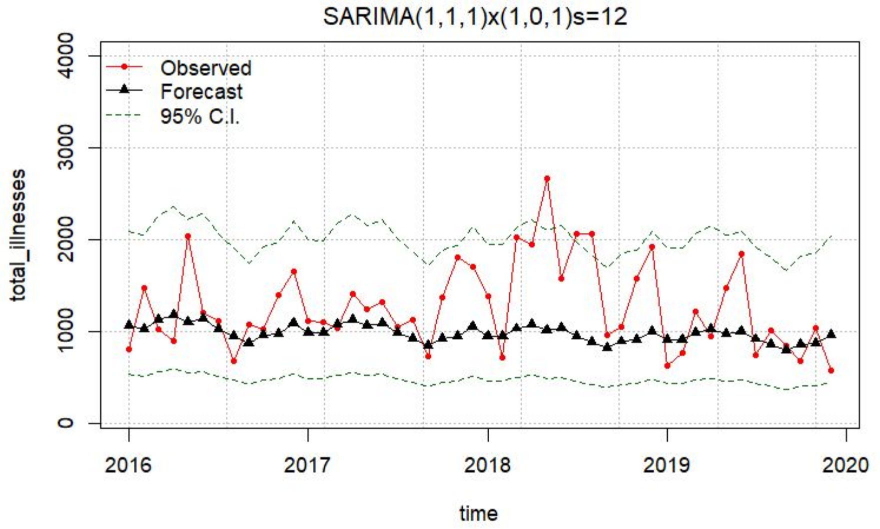
Forecast for data between 2016 - 2019

Data 1

```
$ttable
      Estimate     SE  t.value p.value
ar1     0.1692 0.0729   2.3201  0.0213
ma1    -0.9403 0.0226 -41.5258  0.0000
sar1    0.9509 0.0912  10.4234  0.0000
sma1   -0.8796 0.1443  -6.0948  0.0000
```

Data 2

```
$ttable
      Estimate     SE  t.value p.value
ar1     0.1501 0.0736   2.0395  0.0426
ma1    -0.9375 0.0237 -39.5128  0.0000
sar1    0.9474 0.0951   9.9615  0.0000
sma1   -0.8755 0.1472  -5.9484  0.0000
```



Real data and forecasting of $SARIMA(1,1,1)x(0,1,1)_{12}$

Model output from data before 2016

Data 1

$ttable

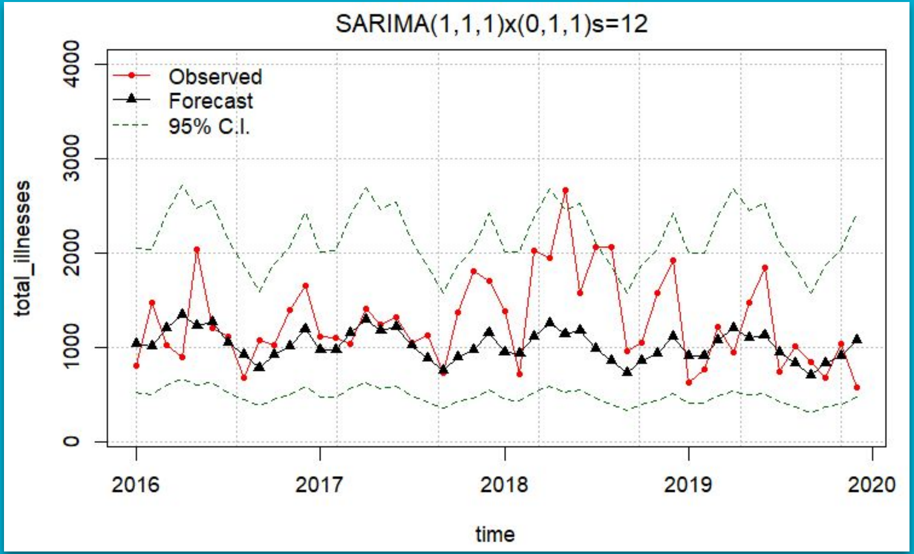
	Estimate	SE	t.value	p.value
ar1	0.1371	0.0752	1.8244	0.0696
ma1	-0.9393	0.0290	-32.3892	0.0000
sma1	-0.9093	0.0857	-10.6136	0.0000

Data 2

$ttable

	Estimate	SE	t.value	p.value
ar1	0.1164	0.0759	1.5346	0.1265
ma1	-0.9336	0.0300	-31.1513	0.0000
sma1	-0.9034	0.0820	-11.0127	0.0000

Forecast for data between 2016 - 2019



Real data and forecasting of ARIMA(1,1,1)-GARCH(1,0)

Model output from data before 2016

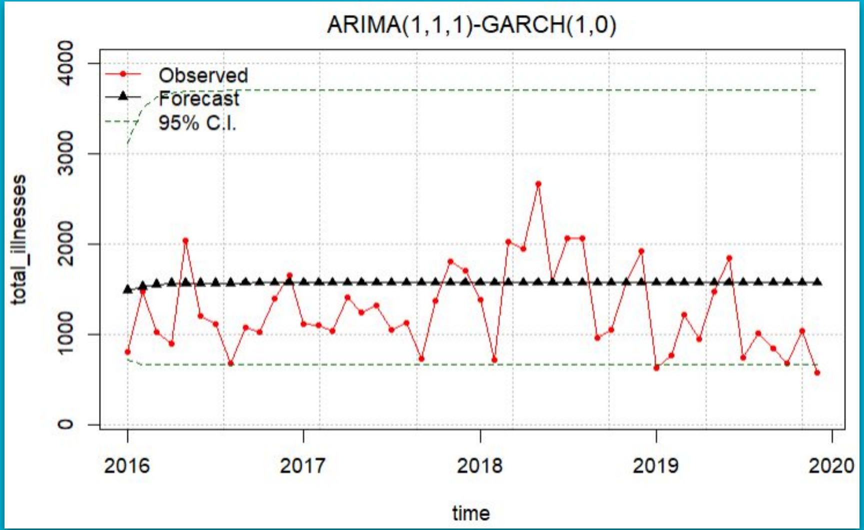
Data 1

```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu       3.65004     0.47009    7.765 8.22e-15 ***
ar1      0.50416     0.06355    7.934 2.22e-15 ***
omega    0.14052     0.01639    8.575  < 2e-16 ***
alpha1   0.01375     0.06644    0.207    0.836
```

Data 2

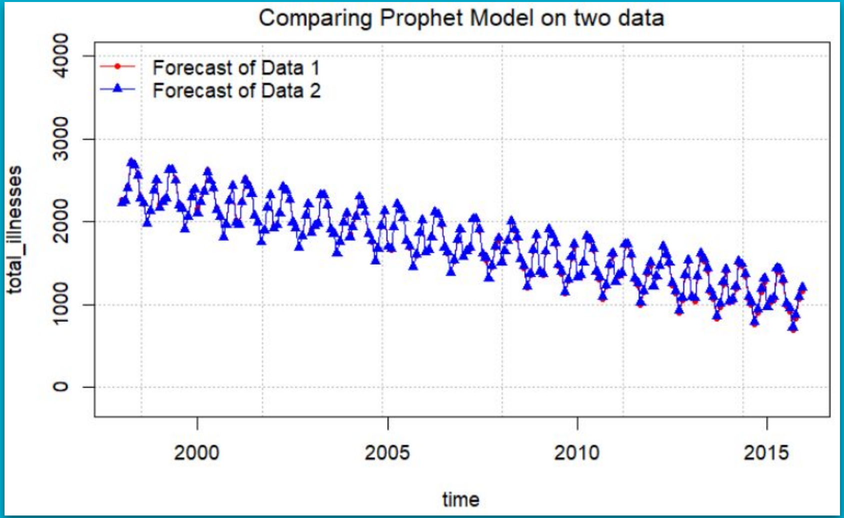
```
Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu       3.74032     0.48335    7.738 9.99e-15 ***
ar1      0.49186     0.06530    7.532 5.00e-14 ***
omega    0.14107     0.01671    8.444  < 2e-16 ***
alpha1   0.03958     0.07106    0.557    0.578
```

Forecast for data between 2016 - 2019

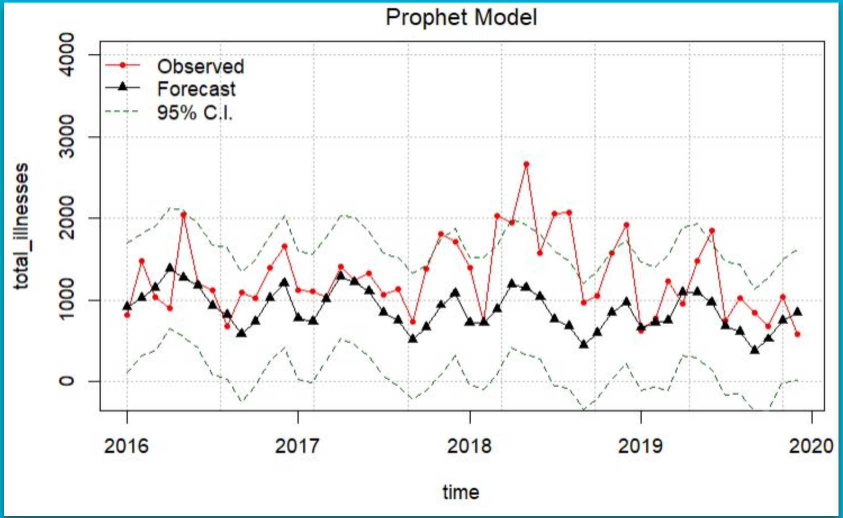


Real data and forecasting of Prophet Model

Fitting output from data before 2016



Forecast for data between 2016 - 2019



Conclusions

Key takeaways

1. *Model formulation and estimation.* We constructed six models (two ARIMA, two SARIMA, ARIMA-GARCH, Prophet) and compared their fit and forecasts.

2. *Model selection.* We chose the ARIMA(1, 1, 1) x $(1, 0, 1)_{12}$ model, based on these criteria:

- The seasonal ARIMA model has lower AIC and BIC values.
- The forecast captures the seasonality of foodborne disease outbreaks better than non-seasonal models.

3. *Validation.* The RMSE of the forecast is the lowest of all models that include seasonality.

Learnings

Key Concepts:

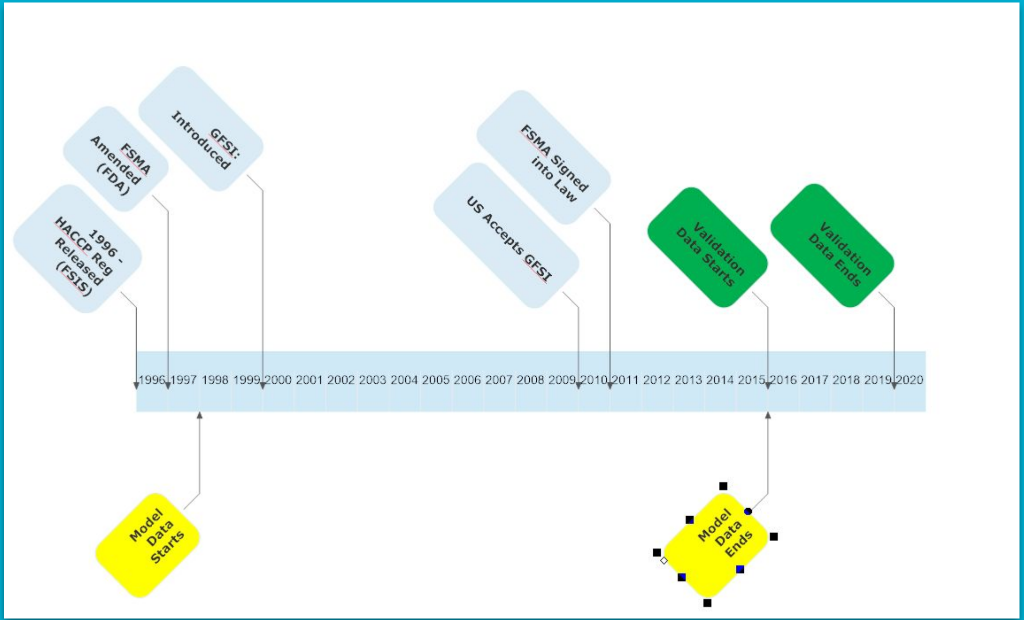
- Importance of the data
- Time changes everything
- Impact of the Data

Importance of Data

- Data Cleaning
  - Source of the information
  - Understanding the basics of the project
  - Consumer habits
    - 2022 - Sal - 572 (92 / 2) Backyard Poultry

Time Changes Everything

- Regulatory Changes
  - Time Delays in Implementation
- Improvements in Data Collection
- Improvements in Testing



"Salmonella contamination on broiler chickens (carcasses) decreased by 56 percent from 1995, before the HACCP final rule was announced, to 2000. The number of foodborne illness cases attributed to Salmonella on broilers was 190,000 lower in 2000 than in 1995." — Williams and Ebel study, 2012

Outbreaks

- 1999 - Hot Dogs - LM - 100
- 2006 - Spinach - EC 205
- 2006 - Taco Bell - EC 71
- 2009 -Peanut Butter - SAL - 714
- 2011 - Canteloupes - LM 147
- 2011 -Ground Turkey - SAL - 136
- 2013 - Chicken - Sal - 634
- 2015 - Chipotle - EC - 55
- 2015 - Mexi Cucumbers - SAL - 907

- 2016 - Flour - EC 63
- 2017 - Leafy Greens - EC 25
- 2018 - Romaine - EC 272
- 2019 - Ground Beef - EC 209
- 2019 - Flour - EC 167

Impacts of the Data

- Determines success of the industry
- Basis for Regulatory Goals and Changes
    - Healthy People Goals - CDC
    - USDA / FDA Strategy
- Regulatory Changes -
    - Sal Adulterant - Kiev / Cordon Blue
    - Flour Mills
    - Exclusion of Supply

Questions, comments, suggestions?

Thank you for your engagement and feedback!

Surprisingly good of Prophet Model