Group 12 Final Report

Trina Shores[*], Group Leader Steven Macapagal[†], Editor/Analyst
Journey Martinez[‡], Computation/Analyst Yuan Yao[§], Editor/Analyst
Heather Nagy[¶], Analyst Kenneth Porter[‖], Editor

2022-08-09

Contents

[*] katrina.shores@tamu.edu - Graduate Certificate in Statistics (Distance)
[†] steven.macapagal@tamu.edu - Masters of Science in Statistics (Distance)
[‡] journeymartinez89@tamu.edu - Masters of Science in Statistics (Distance)
[§] bonedragona@tamu.edu - Masters of Science in Biology (Distance)
[¶] hnagy@tamu.edu - Masters of Science in Statistics (Distance)
[‖] kporte@tamu.edu - Masters of Science in Statistics (Distance)

Abstract

Food borne disease affects a significant amount of Americans yearly. Data on food borne illness from January 1998 through December 2015 was made stationary through log transformation and differencing, and then a model was selected based on ACF and PACF behavior, as well as residual diagnostics. The chosen model was an ARIMA$(1,1,1) \times (1,0,1)_{12}$, out of candidates such as ARIMA, GARCH, other SARIMA, and Prophet models. Forecast performance was used to validate model choice, and in all informative cases showed a decreasing seasonal trend in predicted instances of illness for the 48 months following December 2015.

Background and Research Goals

The CDC estimates that each year roughly 1 in 6 Americans (or 48 million people) gets sick, $128,000$ are hospitalized, and $3,000$ die of food borne diseases. Our data set provides data on food borne disease outbreaks reported to the CDC from 1998-2015. Data fields include year, state, location, reported food vehicle and contaminated ingredient, etiology, status, total illnesses, hospitalizations, and fatalities.

Our goal was primarily to describe the trends and variability in our illness data. After finding a valid model and analyzing the relationship between illnesses and hospitalizations, we wanted to check a large subset of the data (illnesses from Salmonella) to see behaviors after removing some variability from other illness sources.

The results of that analysis would serve as inspiration to form additional valid models for the sake of comparison, using AIC, BIC, and forecasting RMSE. Only after careful consideration of these could we decide on the best model for our data.

Stationarity

Figure 1 shows total illnesses per month from January 1998 through December 2015 (blue), with mean (purple). There appeared to be a downward trend in total illnesses per month, and a decrease in variability after 2010. There was a slight seasonal pattern; the highest illness counts seemed to be between February and May while the lowest were between July and November.

In order to make this time series stationary, we had to address two issues. The first was heteroskedasticity; the variability in the first 12 years was much larger than in the last 5. The second was trend; the mean of the data tended downward with time.

To remedy these, we first took a log transformation of our illness data to stabilize the variance. Afterward, we looked at both differencing and detrending; differencing the data cut down the autocorrelation more than detrending did (0.25 vs. -0.41 at lag 1). So, our final transformations were taking the log, then differencing the total illnesses; shown in plots 2 and 3 of Figure 1.

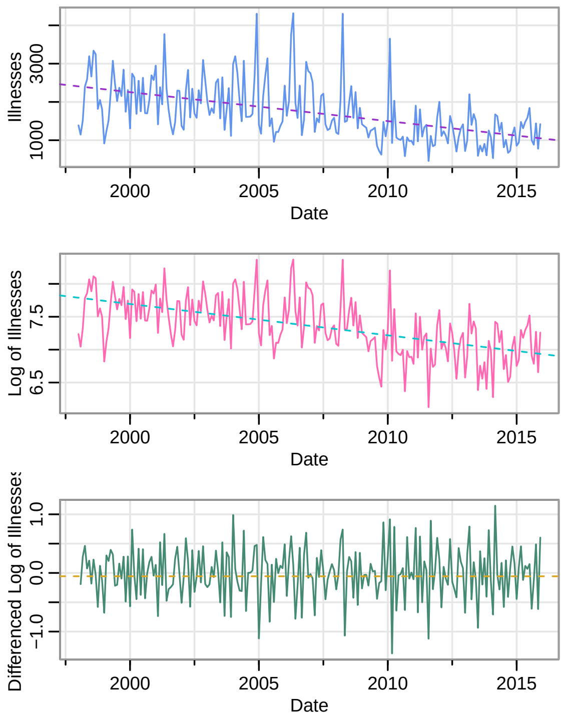


Figure 1: Transforming to Stationarity

Model Selection

Once we identified that this transformation achieved stationarity, we used the correlogram (ACF) and partial autocorrelogram (PACF) to select an initial form of the model. Notice how the ACF dropped off immediately after lag 1, while the PACF trailed off; the first few terms were significant, gradually decreasing until consecutively nonsignificant after lag 4. This behavior suggests an ARIMA$(0, 1, 1)$ model, since only one lag in the ACF was significant while the PACF trailed off with multiple significant lags.

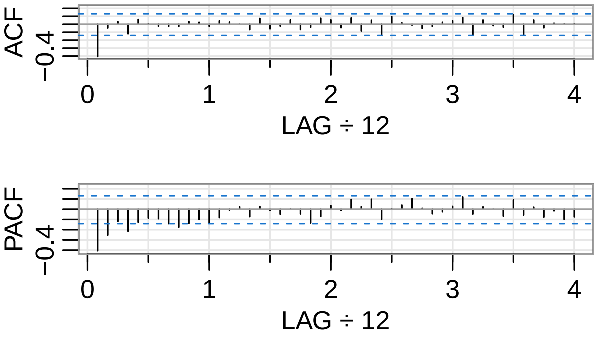


Figure 2: Stationary Illness Data

From there, we fit the ARIMA$(0, 1, 1)$ model to our stationary data using the sarima function from the astsa package. Our conditional sums of squares converged to -1.010483 and unconditional sums of squares (MLE) converged to -1.017639; the MA(1) term was significant by a p-value of 0 resulting from a test statistic of -31.3782 following a t distribution with 214 degrees of freedom. The AIC for the model was 0.8212041, and BIC was 0.8525589. The results of the residual analyses for the model showed that he standardized residuals had a decent scatter and followed a Normal distribution well, evidenced by the Normal Q-Q plot. However, one thing we noticed was that the Ljung-Box statistic plot showed all significant p-values for the Q-tests. This meant we rejected that the residuals were uncorrelated; meaning the residuals were not white noise. In order to remove the autocorrelation in the residuals and get white noise, we decided to add an AR(1) term, and refitted the model to be an ARIMA$(1, 1, 1)$.

Overfitting is an issue that impacts forecasting accuracy, so caution was taken when adding the parameter. After fitting, our conditional sums of squares converged to -1.014162 and unconditional sums of squares (MLE) converged to -1.039521; both the AR(1) and MA(1) terms were significant by p-values of 0.0023 and 0, resulting from test statistics of 3.0912 and -44.6319, each following a t distribution with 214 degrees of freedom. The AIC (0.7867413) and BIC (0.8337735) were smaller for this model compared to those from the previous model. Considering the BIC was still lower despite having more parameters, we felt confident it was best to add the AR(1) term. Using the coefficient estimates, our new model was expressed as follows.

$$\nabla \hat{x}_t = 0.2193_{(0.071)} \nabla x_{t-1} + \hat{\omega}_{t(0.1241)} - 0.9351_{(0.021)} \omega_{t-1}$$

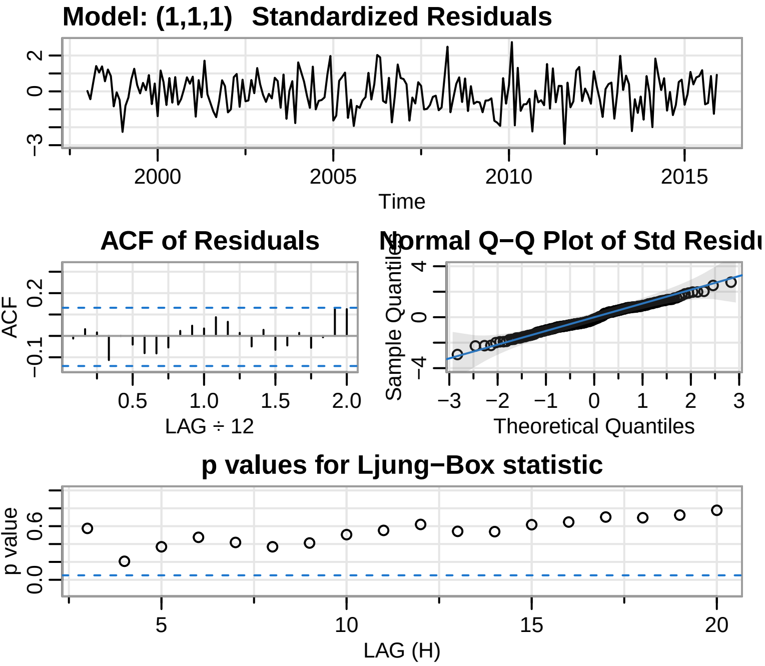


Figure 3: Residual Analyses of ARIMA(1,1,1)

Figure 3 displays the results of residual analyses. The ACF for the residuals showed an autocorrelation of zero for all lags and the residual plot itself showed no underlying pattern. The Ljung-Box Q-statistic looked at the accumulation of autocorrelation instead of the individual autocorrelations seen in the ACF. For this model, the p-values exceeded our significance level of $\alpha = 0.05$; we did not reject the null hypothesis that the residuals were white noise.

The relationship between illnesses and hospitalizations was studied in their cross-correlation (CCF). We plotted the CCF of log illnesses and log hospitalizations in the first plot of Figure 5. There were some significant and systematic cross-correlations, but the magnitude seemed to be somewhat small (less than -0.2). We then differenced the data to examine the effect of illnesses on the growth rate of hospitalizations. This is shown in the second plot of Figure 4, where there didn't appear to be any significant cross-correlations, except at lag 1. Again, the cross-correlation was about -0.2, meaning that an above average increase in log illnesses tended to be followed by a below average decrease in log hospitalizations about 1 month later (i.e. more people were being admitted to the hospital than expected in the period following the illnesses).

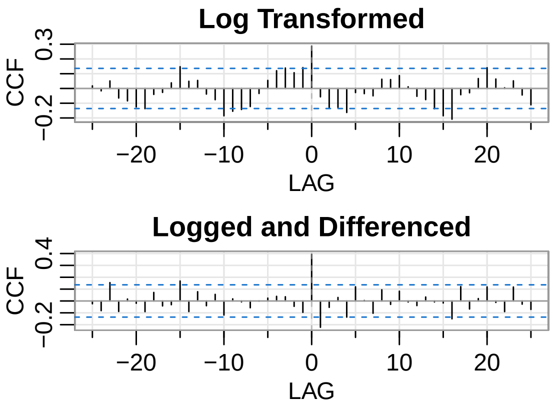


Figure 4: Cross-Correlation of Illnesses and Hospitalizations

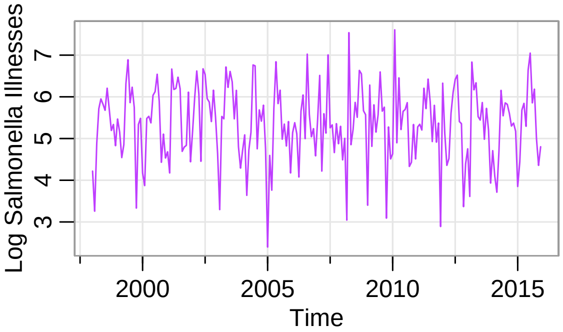


Figure 5: Log Salmonella Illnesses

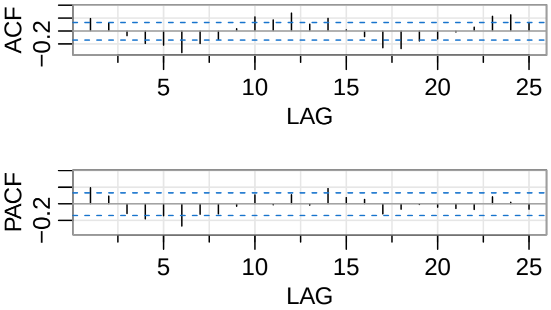


Figure 6: Log Salmonella Illnesses

Specifying Salmonella Source

One limitation of the prior analyses was the many different sources of disease, having different behaviors over time. We restricted our analyses to cases of illness from Salmonella to see if our chosen model would differ. Figure 5 shows the log transformed illnesses from Salmonella. The series appeared to be stationary, so we proceeded to look at the dependence structure. Notice the seasonal patterns every 12th lag. Because the seasonal ACF did not seem to decay over time, we differenced the series seasonally and added a seasonal MA(1) term.

Only the first ordinary lags of the ACF and PACF were significant, so we added both AR(1) and MA(1) terms. Therefore, our proposed model was an $\mathrm{ARIMA}(1,0,1)\times(0,1,1)_{12}$.

From there, we fit the $\mathrm{ARIMA}(1,0,1)\times(0,1,1)_{12}$ model to our stationary Salmonella illness data. Our conditional sums of squares converged to -0.192331 and unconditional sums of squares (MLE) converged to -0.175485; the MA(1), AR(1), and seasonal MA(1) terms were all significant by a p-values of 0 resulting from a test statistics of -5.6491, 4.3361, and -9.5770 respectively, following t distributions with 200 degrees of freedom. The AIC for the model was 2.535928, and BIC was 2.617254. This represented a considerably good model. Using the coefficient estimates given, our model was expressed as follows.

$$(1-0.7605_{(0.175)}B)\hat{x}_t = (1-0.8399_{(0.149)}B)(1-0.999_{(0.001)}B^{12})\hat{w}_t$$

Figure 7 displays the results of residual analyses for validating the model. The standardized residuals showed a good scatter, decently constant variance, and followed a Normal distribution well, evidenced by the Normal Q-Q plot. Both the ACF and Q tests agreed that there was no correlation remaining. Thus, the residuals were white noise and the model was valid. To further assess the strength of the model, we forecasted 48 months ahead. We saw that this seasonal forecast seemed to match the cyclic nature and variance seen in the past.

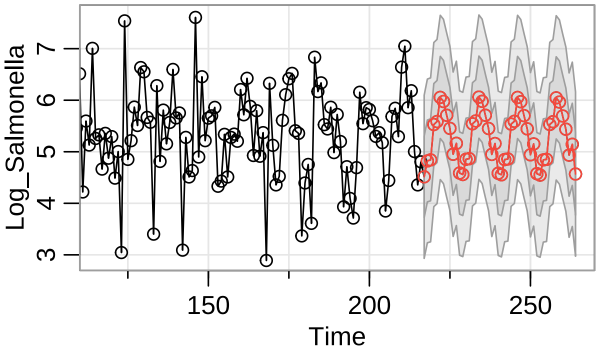


Figure 8: Forecast of Log Salmonella Illnesses

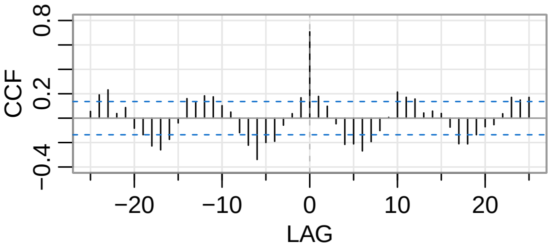


Figure 9: Log Transformed Salmonella Illnesses

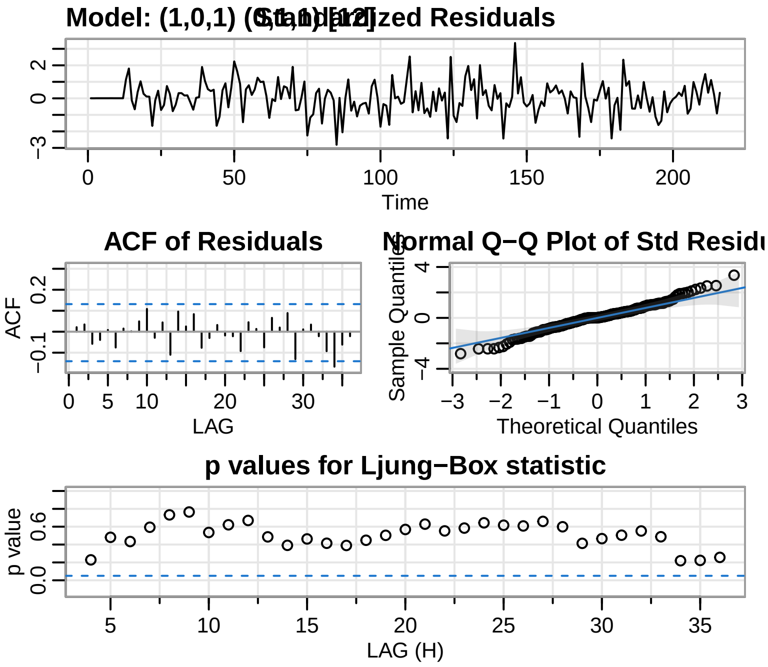


Figure 7: Residual Analyses of Log Salmonella Illnesses

Relationships Between Illnesses and Hospitalizations

When we looked at the salmonella cases only, the patterns we saw in the CCf plot from Figure 8 seemed to be much stronger. For lag 0, the most likely explanation for this pattern was that hospitals have a consistent reporting mechanism of food borne illnesses to the CDC, whereas cases diagnosed outside a hospital may not be reported consistently.

Comparisons to ARIMA(1,1,1)

Confirming our suspicions of seasonality in our data, we sought a seasonal model to compare to our ARIMA$(1, 1, 1)$. We also entertained models proposed in previous literature. Lastly, we fit both a GARCH and Prophet model, as they are known to handle the complex variance and seasonality that we also suspected in our data.

In prior literature written by Li, Peng, Zhou, & Zhang (2021), the authors proposed an ARIMA$(1, 1, 0)$ on incidence of food borne illness outbreaks. We tried fitting that model to our data to see if it would provide a better level of fit. However, the BIC was found to be 1.0763; higher than that of our chosen model. Also concerning was the several significant p-values from the Q tests indicating correlation in the residuals. Therefore, the model from Li et al. (2021) did not provide a better fit for our data.

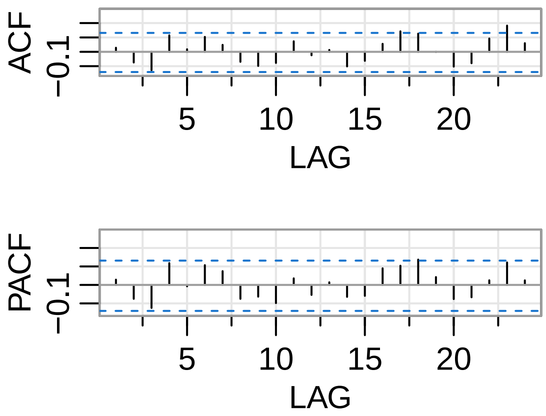


Figure 10: Squared Residuals from ARIMA(1,1,1)

Analyzing the squared residuals for ARIMA$(1, 1, 1)$ revealed a small bit of dependence (Figure 10). The mean of the residuals came out to 0.01191047, making them very slightly biased. Thus, we considered an ARIMA$(1, 1, 1)+$GARCH$(1, 0)$. But the α term lacked significance with a p-value of 0.8112. Using our original data (not log transformed), the α p-value was found to be 0.0698, which is borderline significant in the two-sided case. However, the AIC of this model jumped up to 15.87, much higher than the 0.926 for our transformed data. We then concluded that GARCH was not useful in our case after all.

To determine the seasonal components of our model, we revisited the correlograms of Figure 2, but couldn't visually determine the pattern. So we began with an SMA$(1)_{12}$ with seasonal differencing addition, yielding the unsatisfactory result of an insignificant AR(1) term, higher BIC, and more concerning higher AIC. This seasonal model, although technically valid, was not an improvement. We then realized that the lack of seasonal pattern in the correlograms was likely due to the model not requiring seasonal differencing. Hence, the ARIMA$(1, 1, 1) \times (1, 0, 1)_{12}$ was fit, mimicking our nonseasonal terms.

All terms (AR(1), MA(1), SAR(1), and SMA(1)) were significant by a p-values of 0.0213, 0, 0, and 0 resulting from test statistics of 2.3201, -41.5258, 10.4234, and -6.0948 respectively, following a t distribution with 211 degrees of freedom. Figure 11 displays the residual analyses; constant variance, following a Normal distribution, and lack of correlation indicated only white noise remained. Better yet, the AIC (0.7686196) was lower than the ARIMA$(1, 1, 1)$, while the BIC (0.8470065) was only slightly higher (despite the model having several additional parameters). The mean of the residuals were closer to 0, and the squared residuals were less correlated. Thus, the seasonal model was a clear improvement, shown below.

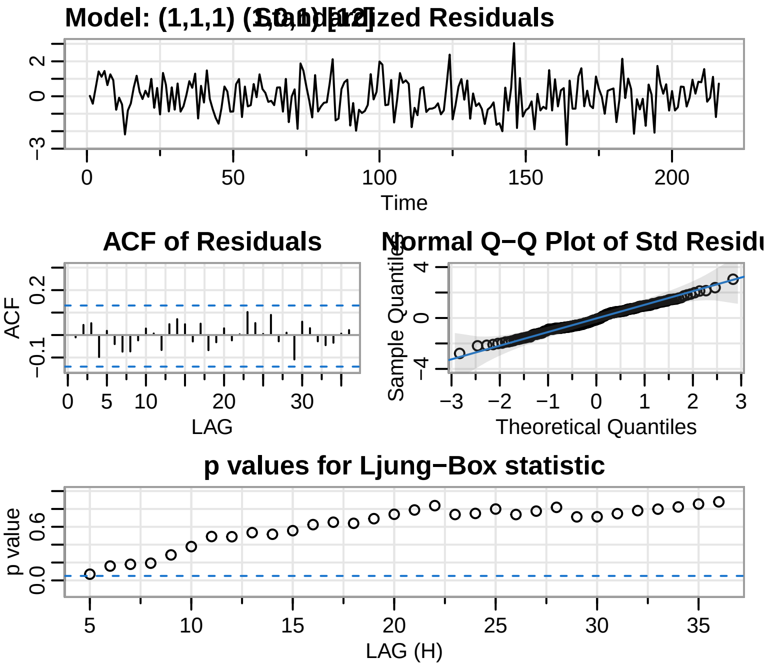


Figure 11: ARIMA(1,1,1)x(1,0,1)12

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

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

Forecasting

Because our original data set ended in December 2015, we were able to make forecasts from our models and compare their predictions to National Outbreak Reporting (NOR) data collected from 1998-2020 (although we did not include 2020, since the data appeared to be heavily influenced by the COVID-19 pandemic). Data coinciding from 1998-2015 was compared in Figure 12 to ensure the appropriateness of validating our models using the NOR data. Our data is shown in pink, the NOR data in blue. Few of the NOR data points showed from behind our data, meaning they mostly occurred in the same patterns and values, and similarly continued after our data ended. This was true for illness, hospitalization, and fatality counts, which gave us confidence to proceed.

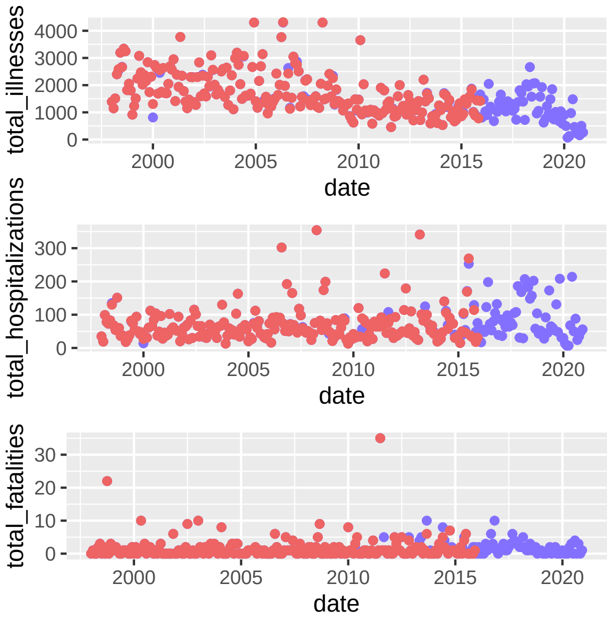
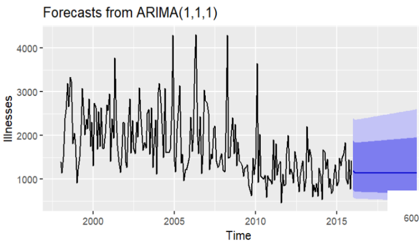
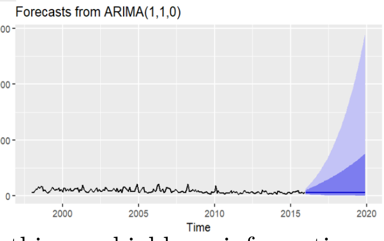
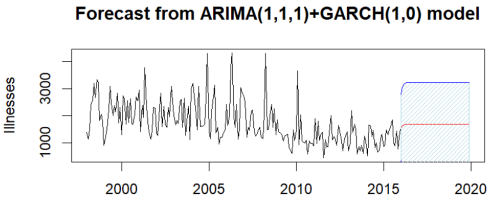


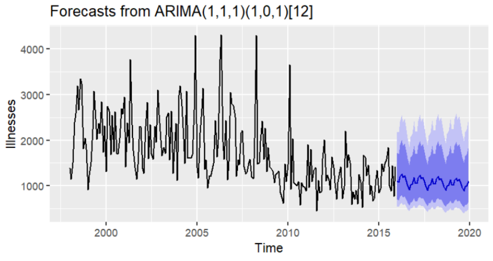
Figure 12: Comparing Original and New Data

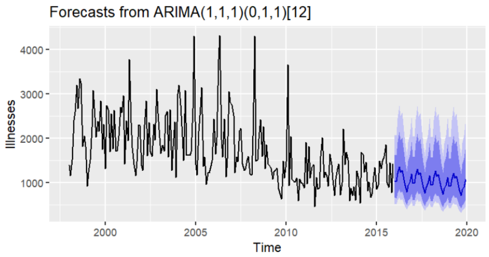
Forecasts from all models previously discussed are displayed below, as well as a Prophet model fit to our data. We chose to introduce the Prophet model in forecasting for its ability to detect complex seasonal patterns in data and use them to provide more accurate forecasts. First, ARIMA$(1,1,1)$ is displayed first. The forecasts borrowed the mean and spread of the preceding 6 years and made no attempt to predict cyclical behavior. Second, ARIMA$(1,1,0)$, is shown second. As before, the forecasts were flat near the mean of the data, accounting for no seasonality or trend. The confidence bands increased exponentially as time progressed, indicating diminishing confidence. Third, the ARIMA$(1,1,1)$ + GARCH$(1,0)$ forecasts. Note the flat trend at the overall mean with wide confidence bands stretching to 0; this was highly uninformative, confirming the poor performance of the model. Next, the ARIMA$(1,1,1) \times (0,1,1)_{12}$ was proposed. The forecasts captured the seasonal volatility well with neater, more informative confidence bands. However, the AR(1) term of this model was insignificant, making it a misleading representation. Our best model, the ARIMA$(1,1,1) \times (1,0,1)_{12}$, showed a similar forecast having seasonality and downward trend, but less volatility. The confidence bands were very similar; capturing the cyclic nature and only encompassing the spread of the preceding 6 years. Last considered was the Prophet model. It captures the seasonal component of the data with volatility comparable to the ARIMA$(1,1,1) \times (0,1,1)_{12}$. In contrast, it predicts a steeper downward trend than the previous models.

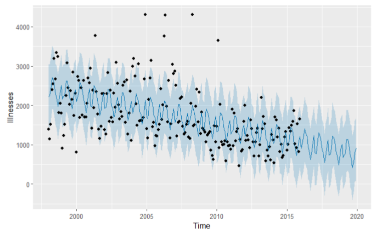


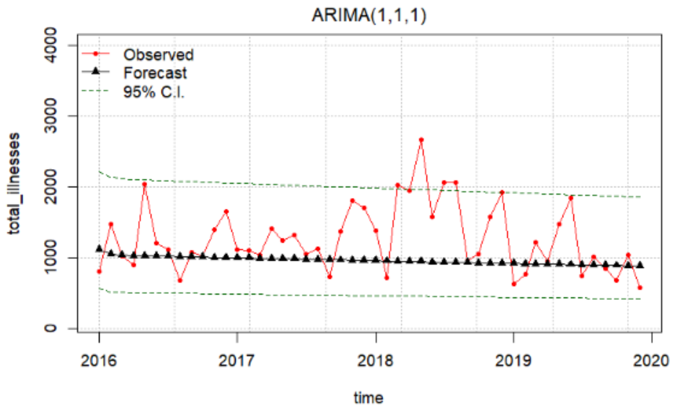


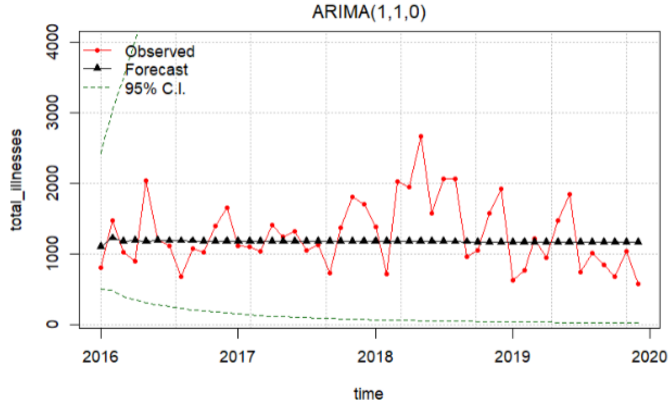


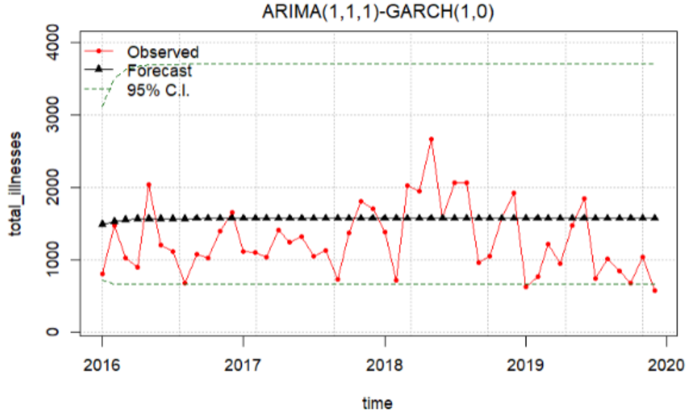


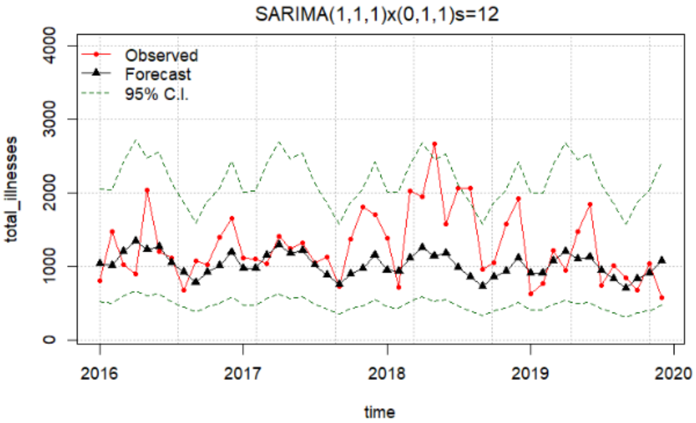




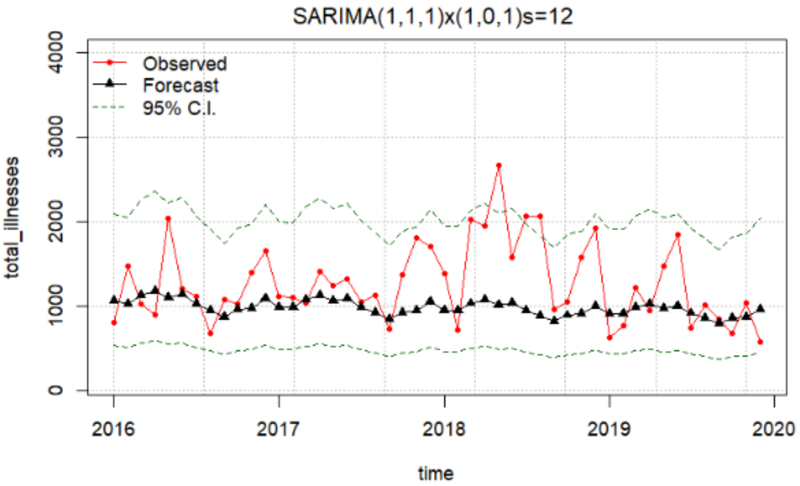


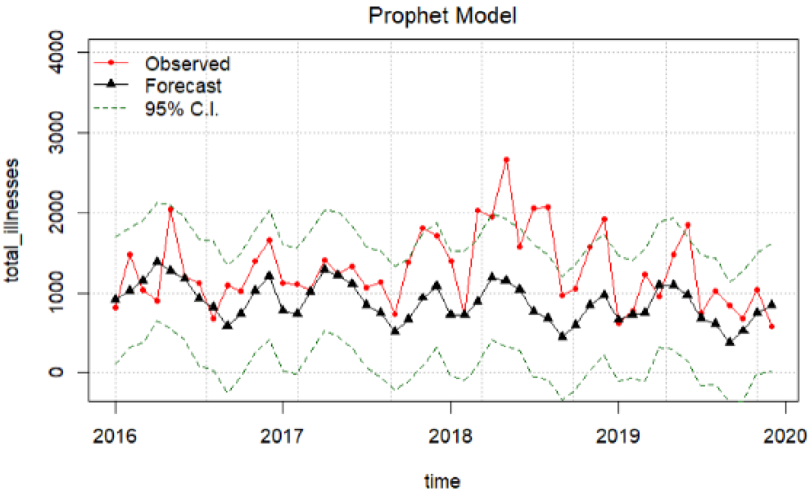






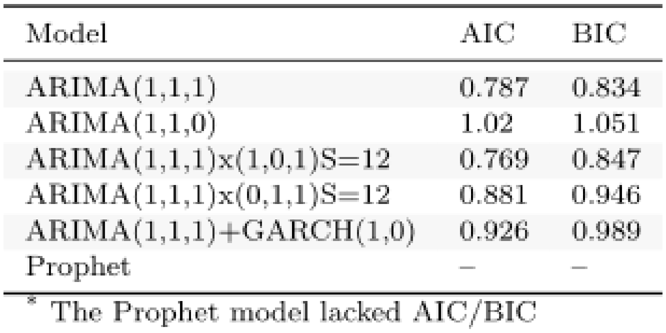
The following plots display the NOR data over the previous forecasts. Unsurprisingly, the models lacking seasonal components gave poor predictions. Most notable was the difference between the Prophet and seasonal ARIMA models; the Prophet model failed to predict spikes in 2018, having strict confidence bands and steeper downward trend. The ARIMA$(1,1,1)\times(0,1,1)_{12}$ captured the most NOR data points, with the ARIMA$(1,1,1)\times(1,0,1)_{12}$ close behind.





Conclusion

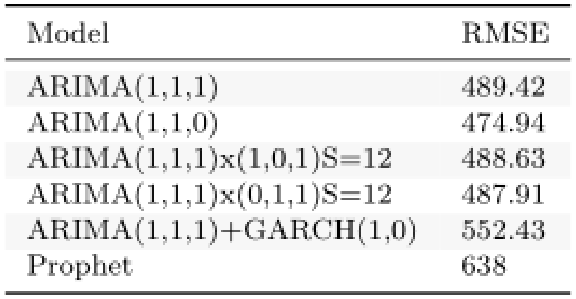
Table 1: Comparison of AIC and BIC



Model	AIC	BIC
ARIMA(1,1,1)	0.787	0.834
ARIMA(1,1,0)	1.02	1.051
ARIMA(1,1,1)x(1,0,1)S=12	0.769	0.847
ARIMA(1,1,1)x(0,1,1)S=12	0.881	0.946
ARIMA(1,1,1)+GARCH(1,0)	0.926	0.989
Prophet	–	–

[*] The Prophet model lacked AIC/BIC

Table 2: Comparison of RMSE for Forecasts



Model	RMSE
ARIMA(1,1,1)	489.42
ARIMA(1,1,0)	474.94
ARIMA(1,1,1)x(1,0,1)S=12	488.63
ARIMA(1,1,1)x(0,1,1)S=12	487.91
ARIMA(1,1,1)+GARCH(1,0)	552.43
Prophet	638

Table 1 displays the AIC and BIC from fitting each model to the original data. The lowest AIC occurred in the ARIMA$(1,1,1) \times (1,0,1)_{12}$, while the lowest BIC occurred in the ARIMA$(1,1,1)$. We note that BIC penalizes more heavily for the addition of parameters, to discourage over fitting. Considering that the seasonal model has several more parameters than the non-seasonal model, the small increase in BIC is telling of the seasonal model having a better fit.

Table 2 shows the root mean squared error (RMSE) of the model forecasts. Although the ARIMA$(1,1,0)$ has the lowest RMSE, it was not a valid model for our data and the forecasts had exponentially widening confidence bands; we chose to ignore it. Consistent with our observations above, the ARIMA$(1,1,1) \times (0,1,1)_{12}$ had the lowest RMSE and the ARIMA$(1,1,1) \times (0,1,1)_{12}$ was a close second. However, the former was not a well fit model and possessed insignificant parameters, while the latter was an excellent fit overall.

In consideration of all of the above measures, we concluded that the ARIMA$(1,1,1) \times (0,1,1)_{12}$ out performed all other models in data fitting and prediction. It is thus recommended for modeling and predicting food borne illness counts from 1998-2019.

It is worth noting that the decrease in volatility after the year 2010 coincides with the acceptance of the Global Food Safety Initiative, as well as the signing of the Food Safety Modernization Act. Perhaps these measures worked to reduce outbreaks of food borne illness which had been contributing to the large spikes in illness counts preceding 2010, or perhaps they inspired more accurate testing and data collection which reduced the inclusion of false positives and cases of illness not actually contracted from food.

References

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.

"National Outbreak Reporting System." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), 3 Feb. 2022, wwwn.cdc.gov/norsdashboard.

"Foodborne Disease Outbreaks, 1998–2015." *Kaggle*, 2017, www.kaggle.com/datasets/cdc/foodborne-diseases.

"Estimates of Foodborne Illness in the United States." *Centers for Disease Control and Prevention*, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Foodborne, Waterborne, and Environmental Diseases (DFWED), 5 Nov. 2018, www.cdc.gov/foodborneburden/index.html.

Shumway, R. H., & Stoffer, D. S. (2019). *Time series: A data analysis approach using r.* CRC Press, Taylor & Francis Group.

Stoffer, D., & Poison, N. (2022). *Astsa: Applied statistical time series analysis.* https: //CRAN.R-project.org/package=astsa