Time Series on Lyme Data

Jeremy Ling and Alexis Rivera

Spring 2018

1 Literature Review

Lyme's disease is a tick-borne disease caused by Lyme borrelia. In North America, Lyme borrelia's, alson kown as Lyme disease, main cause to human disease is transferred by a microorganism known as *borrelia burgforferi* whereas in Europe, the microorganism that causes the disease is *borrelia afzelii*, *borrelia garinii*, *borrelia burgdorferi*, *borrelia spielmanii*, and *borrelia bavariensis.* For this research, we will focus on *borrelia burgforferi* and its causes of Lyme borrelia in the United States. *borrelia burgforferi* developed an absence in biosynthetic pathways which lead to a full dependency on its environment for nutritional value. The microorganism exists and transmitted to human flesh by ticks known as *Ixodes scapularis* and *Ixodes pacificus. Ixodes scapularis* population exists in the West Coast, as *I scapularis* is found in both the upper Midwest and Northeast. Such ticks is only fed once during each stage of its four stage life cycle – egg, larva, nymph, and adult which at all times can contain *borrelia burgforferi.* Such ticks life expediency is dependent on its environment and climate which can range from two to six years. The transmission of *borrelia burgdorferi* by such ticks through injection takes more than 36 hours and known to occur in peri-urban and rural areas, which is why such ticks exists in the Northeast of the US. It is also known that the risk of Lyme borreliosis decreases if most animals present in the environment are not act as reservoirs, like the species of deer and cattle, for the microorganism (Baker (2017)). Therefore, a good indication of the presence of *Ixodes scapularis* with *borrelia burgdorferi* if there a population of either deer or cattle. Despite people identifying areas with higher risks of such ticks to be present, it takes up to weeks for people to notice any developing symptoms.

People do not often identifying any early symptoms of Lyme's disease until several weeks from the bite of a tick with *borrelia burgforferi.* Early symptoms identified from Center for Disease Control (CDC) includes fever, chills, headache, fatigue, muscle/joint aches, swollen lymph nodes, and Erthymea Migrans[1]. There is no human vaccine for the Lyme's disease despite how impactful its been in the United States. Even if there was a vaccine for Lyme's disease, people still need to be aware of the exposure of tick base illnesses. Therefore, the best alternative in preventing Lyme's disease is reducing people's exposure of such ticks (Keesing and Ostfeld (2018)). Such attempts were made, but remain unsuccessful. In 2015, CDC reports that 95% of confirmed Lyme disease cases came from 14 states as mentioned in Table 1. CDC also notes that the location of cases reported reflect the where the patient lived during the time of the diagnosis of the disease (CDC (2017)).

2 Data

The time series data we are interested in forecasting is Lyme disease report data. For multivariate analysis, we include Google Trends data for the "Lyme disease" search topic. Initial plots of both time series are shown in Figure 1, where we can already see both time series' are shaped similarly.

2.1 Lyme Disease Reports

National Lyme disease reports are easily available, albeit in an annual format. In hopes of finding monthly data, we searched the 14 states where 95% of the Center of Disease Control's reported Lyme disease cases were reported from, finding that Wisconsin and Rhode Island yielded monthly data. With Wisconsin holding more data spanning from 2009 to 2016 with a total of 96 periods, we chose to use their

[1] A circular rash from the result of a tick bite

monthly reports for Lyme disease (of Health Statistics (2018)).

2.2 Google Trends Data

Google Trends normalizes their data in such a way so that searches can be specified to record weekly, monthly, and annually. Regardless of time frame, Google Trends normalizes their data such that the highest number of searches in that time period holds the value ”100”, with other values relative to that period. For example, a time period with exactly three quarters the amount of searches as that of the most popular time period would hold a value of ”75”. Unfortunately, Google Trends data begins only in 2004, making annual data unappealing for time series analysis.

In our case, when we mention Lyme disease, we mean a search *topic* instead of a search *term*. Data on search terms scan each word in the search phrase and return all terms in the query, while data on search topics returns terms that share the same concept. This means the search topic ”Lyme disease” would include searches for ”tick-born illnesses”. Searches have been limited to the state of Wisconsin to match our Lyme disease reports data.

3 Initial Plots

3.1 Seasonality

In our paper, we explore the log transformed models of both datasets. While there were no signs of trends, our initial plots showed signs of seasonality.

Figure 2’s seasonal plots revealed that both datasets looked similar, with Lyme disease peaking around June and July with a trough set in February. Meanwhile, Google Trend seasonal plots for the Lyme disease search term shows a broader peak stretching between May and July. We believe this is due to Wisconsin residents preemptively searching up Lyme disease prevention methods before the expected spike in reported cases occurs. Adult *I scapularis* are active in the environment during the months between October and May and the females transmit *borrelia burgforferi* to humans which develops into Lyme's disease. It takes up several of weeks for humans to finally notice symptoms as mentioned and it demonstrates on Figure 2. As suspected, people start to notice their symptoms and present them to clinicians in the spring and summer time as suggested in the literature review (Seifter et al. (2010)). Our polar and subseries plots in Figure 3 and Figure 4 reflect the observations made in our seasonal plots.

3.2 ACF and PACF

Our ACF plot has a distinct oscillating pattern, which is expected from our data having a seasonal pattern. It also takes many periods before the lags enter the Bartlett bands, forming a cone shape. Our PACF holds a similar shape, dropping past the Bartlett bands much faster in comparison. Both correlograms are shown in Figure 5.

3.3 Stationarity

We ran Augmented Dickey-Fuller tests on both Lyme disease and Google Trends data to test for a unit root. In both cases, our p-value was low enough to reject the null hypothesis that our data was non-stationary where $\alpha = .05$ and conclude that our data was already stationary. Additionally, we used the `ndiffs` command to check how many times our time series data would need to be differenced to become stationary and found that for both sets, the number of differences to reach stationarity would be zero, confirming our ADF test's results.

From these results, we can also conclude that during multivariate analysis, we can assume there is no cointegration between the two time series datasets. Because of this, we skip vector error correction models and proceed with variance autoregressive models.

4 Simple Methods

When modeling, we partitioned our data such that the 2009 to 2014 window was training data, leaving data from 2015 and 2016 as test data. Our first models included four simple methods, mean, naive, seasonal-naive, and random walk, all of which are shown in Figure 6. As we can see from the forecasts plotted below, the mean, naive, and random walk fits are largely horizontal in nature with wide confidence intervals and do a poor job capturing the seasonality in our data.

Meanwhile, the seasonal-naive fit model seems to do a decent job, with tighter 95% confidence interval bands that still contain the test data. When observing its residuals in Figure 7, they are normally distributed, albeit slightly skewed to the left. In addition, the residuals vs. time plot reveals that while they are centered around zero, they are not evenly spread around zero. Finally, the correlogram for our seasonal-naive residuals have some lags jutting out of our Bartlett bands.

5 Seasonal Linear Model

After finding that the seasonal-naive fit was a good start, we tried running a seasonal linear model for comparison, illustrated in Figure 8. While the model does a poor job fitting troughs, it decently fits over peaks and the transitions between the two. The two-year forecast still manages to capture the actual values within its 95% confidence interval. That being said, the residuals look worse than those of the seasonal-naive model, shown in Figure 9.

6 ARIMA Models

After playing around with a few ARMA models, the `auto.arima` command was used to generate R's best ARIMA model, producing an AR(2) MA(1) SAR(1) with non-zero mean model. The model itself does an excellent job fitting over the training data in Figure 10, especially fitting better in troughs in comparison with our seasonal linear model. However, forecasts underestimate the peaks in our test data. Even so, ARIMA model's residuals look the best in comparison with our previous ones. All lags on the correlogram lie within the Bartlett bands, the residuals are fairly normally distributed, and are more evenly distributed around 0 in comparison to previous models. These observations are shown in Figure 11. So far, we conclude that ARIMA modeling is the best candidate for forecasting Lyme disease reports.

7 Simple Linear Regression

When approaching multivariate analysis, we started by plotting a scatter plot between Lyme disease reports and Google Trends data. Finding a fairly strong positive linear trend, we decided to fit a simple linear regression over the data, as shown in Figure 12. After fitting the model over our scatter plot, we find that the model works fairly well, but still leaves a wide error range. Our summary statistics support our analysis, confirming that Google Trends data is statistically significant where $\alpha = .01$ and yielding an R^2 of .66. However, Figure 13's illustration of our linear regression residual analysis suggests univariate analysis does a better job forecasting than simple linear regression.

8 Vector Auto Regressions

When exploring VAR methods, we assume both time series' share a bidirectional relationship. On one hand, an increase in Lyme Disease reports can have a positive effect on Google Trends data because an increase in actual reports could lead to more searches online as the disease catches more attention. On the other, an increase in Google Trends data could have a positive influence on actual reports because individuals detecting early symptoms in themselves and those around them will search about it online before official reports are uploaded. When running our forecast, the 95% confidence intervals capture the actual values, as shown in Figure 14. Figure 15 shows residuals between both variables.

Figures 16 shows the impulse response function from the time series of Google trends of "Lyme's disease" on both time series. In the bottom graph of Figure 16, it demonstrates the shocks from Google trends on to actual reports decays and increases to zero until twenty four months. The top graph of Figure 16 demonstrates one shock from Reports will increase then slowly converges to zero until 24 months. Similarly in Figure 17, it demonstrates the impulse response function from actual reports of Lyme's disease on both time series has similar effect from Figure 16.

Figure 18 shows both FEVD plots for Google Trends data and actual reports. When observing our FEVD for Google Trends, we find that it has almost no explanatory power for actual reports. Conversely, we see the opposite effect for actual reports on Google Trends data. Because of this, we conclude that the relationship between Google Trends data and actual reports are not quite bidirectional, and that actual reports do a better job predicting Google Trends data than vice versa.

9 Conclusions

After finding that Google Trends data was unnecessary to forecast Lyme disease reports, we retired multivariate analysis in favor for univariate modeling. Among univariate models, the AR(2) MA(1) SAR(1) ARIMA model proved to be the best candidate, with forecasts that fit our test data fairly well, and the only model to produce residuals that resembled white noise. Figure 19 shows that when forecasting the next two years using this model, we find that it does a good job fitting over our dataset. The seasonal pattern is kept, but starts losing its effect in the second year. This is to be expected since we observe the same pattern in Figure 10, and that with only 96 observations to work with, we cannot expect to forecast too far into the future. In conclusion. actual reports does a better job in predicting the future. As explained by Baker in "Straight Talk About Chronic Lyme Disease", specialists often rely diagnosing patients based of prevalent symptoms like fatigue, chronic pain, irritable bowel symptoms, difficulty remembering periods of time, etc (Baker (2017)) People who believe they have Lyme's disease on the basis of these symptoms show negative results from blood tests and CDC criteria. Even if Google Trends data captures people with symptoms of Lyme's disease, it still remains up to a 2-tier blood test to fully in identify the individual with the disease.

10 References

References

Baker, P. J. (2017). Straight talk about chronic lyme disease.

CDC (2017). Data and statistics. ”https://www.cdc.gov/lyme/stats/index.html”.

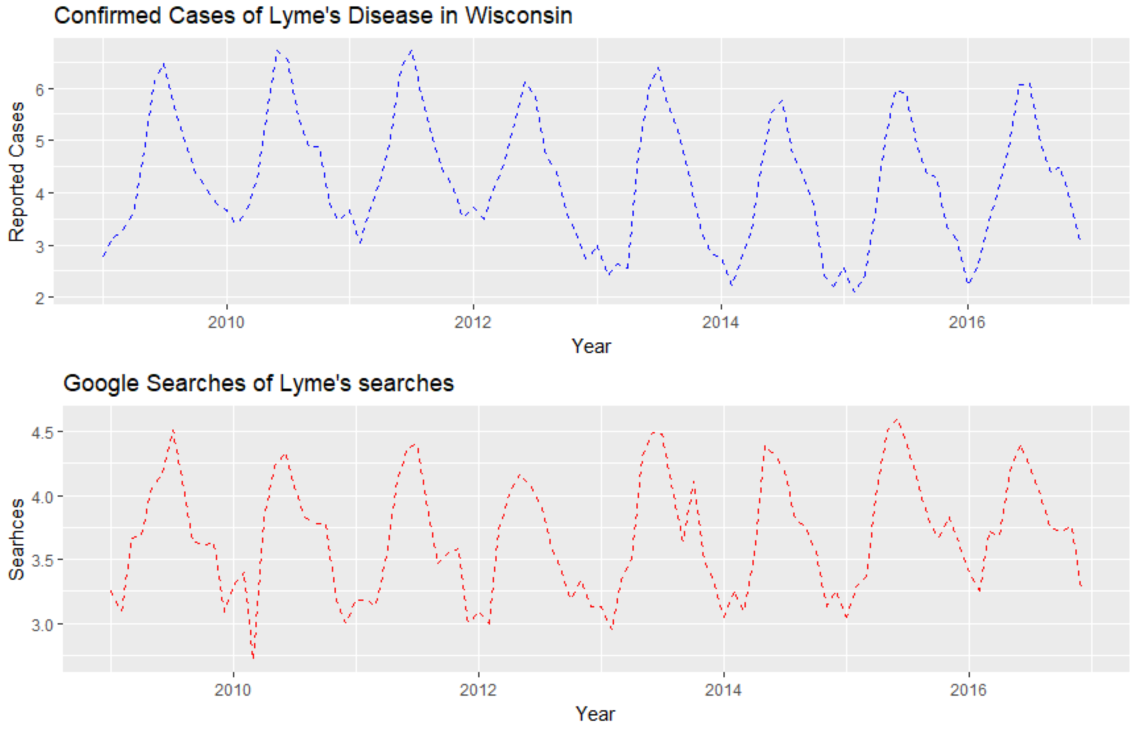
Keesing, F. and Ostfeld, R. S. (2018). The tick project: Testing environmental methods of preventing tick-borne diseases. *Trends in parasitology*.

of Health Statistics, W. D. (2018). Lyme disease - data and statistics. ”https://www.dhs.wisconsin.gov/tickborne/lyme/data.htm”.

Seifter, A., Schwarzwalder, A., Geis, K., and Aucott, J. (2010). The utility of “google trends” for epidemiological research: Lyme disease as an example. *Geospatial health*, 4(2):135–137.

11 Appendix

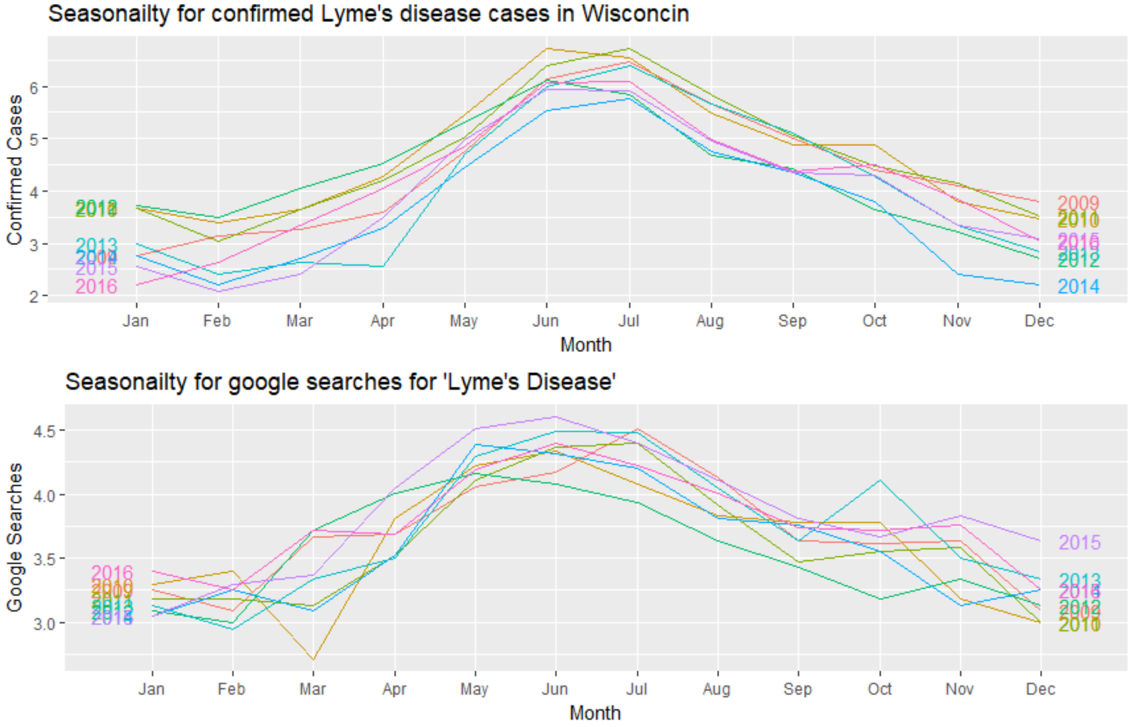
Figure 1:



Confirmed Cases of Lyme's Disease in Wisconsin

Google Searches of Lyme's searches

Figure 2:

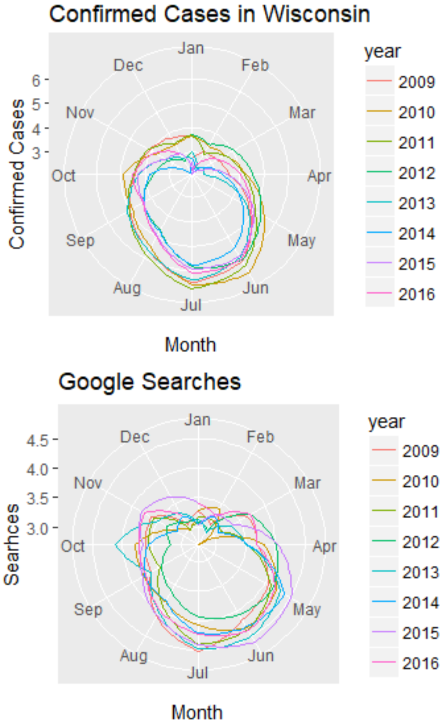

Seasonailty for confirmed Lyme's disease cases in Wisconcin

Seasonailty for google searches for 'Lyme's Disease'

Table 1: States from 95% of confirmed Lyme disease cases reported by CDC

Connecticut	New Jersey
Delaware	New York
Maine	Pennsylvania
Maryland	Rhode Island
Massachusetts	Vermont
Minnesota	Virginia
New Hampshire	Wisconsin

Figure 3:


Confirmed Cases in Wisconsin

Google Searches

Figure 4:

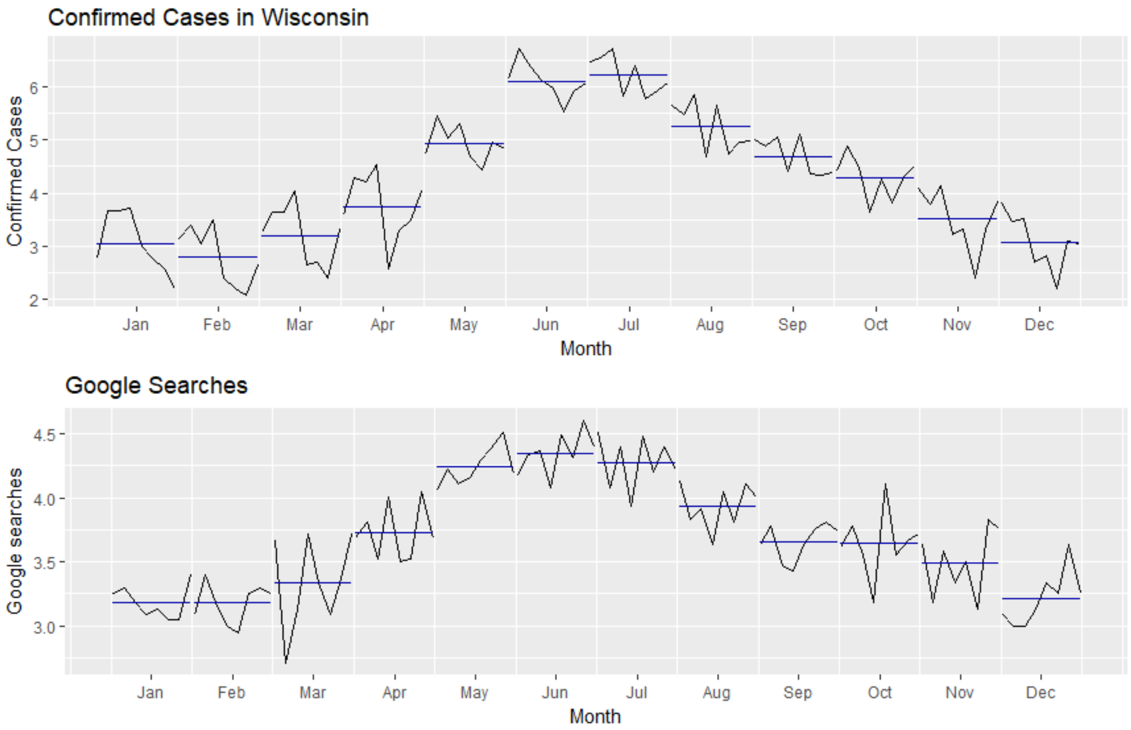


Figure 5:

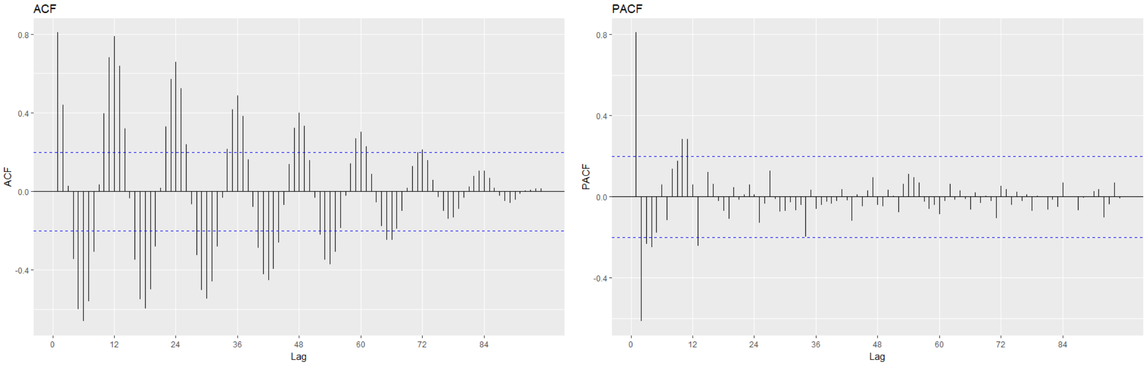
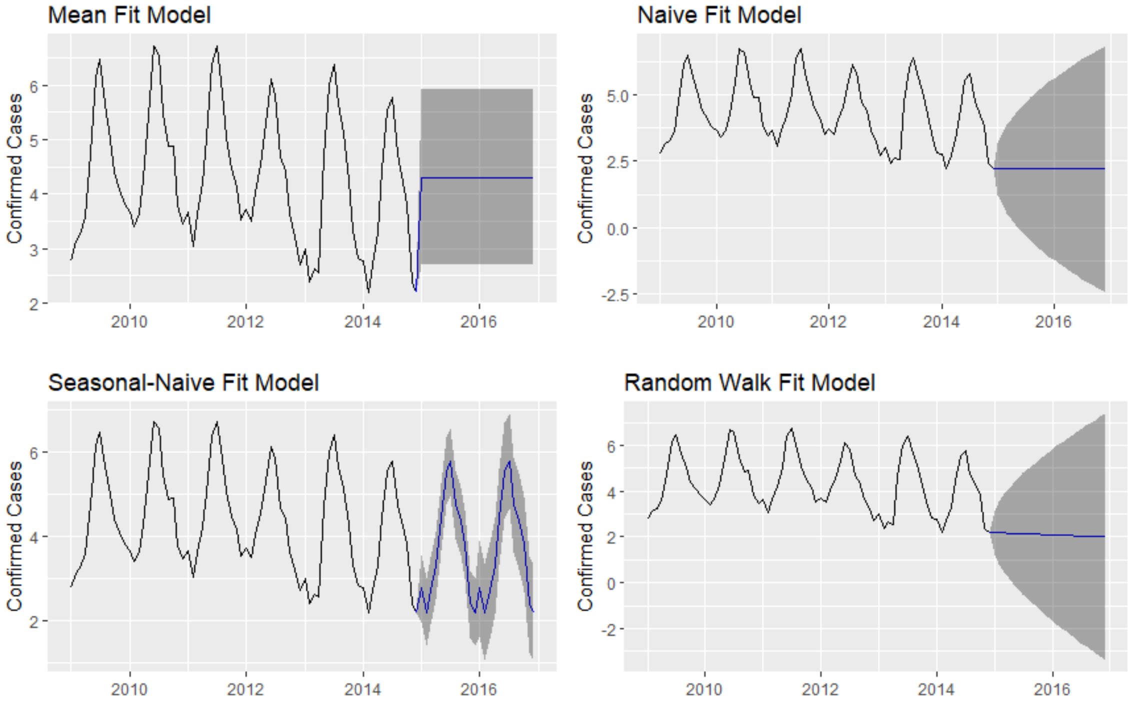


Figure 6:



Mean Fit Model

Naive Fit Model

Seasonal-Naive Fit Model

Random Walk Fit Model

Figure 7:

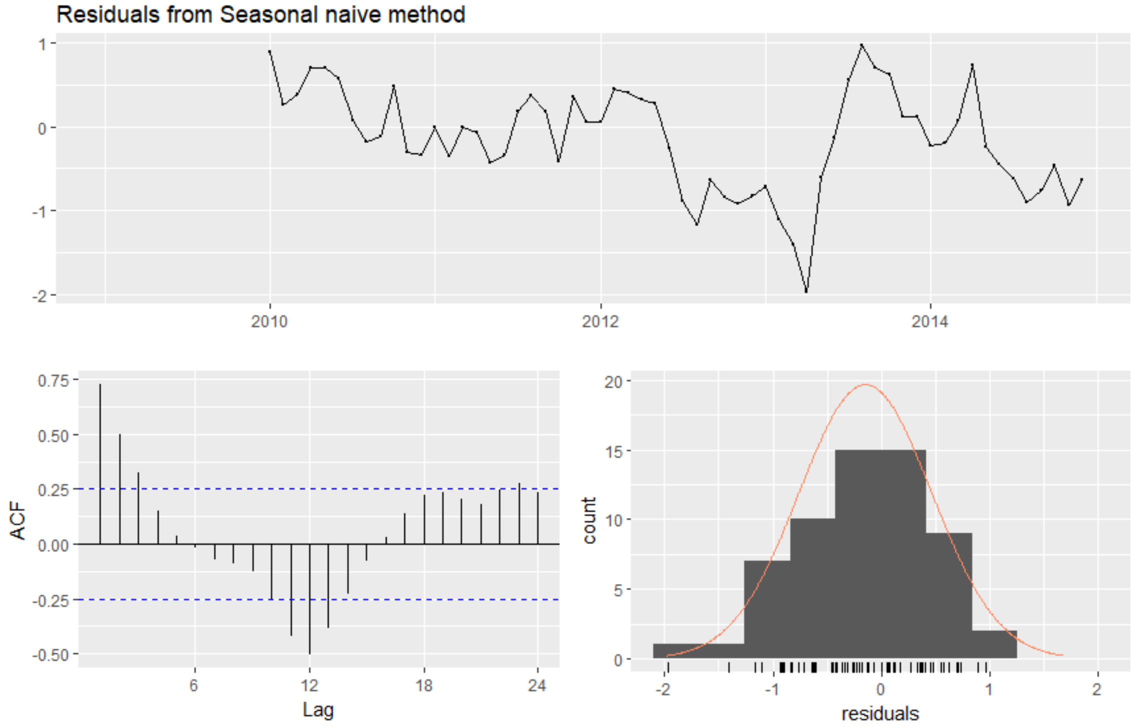


Figure 8:

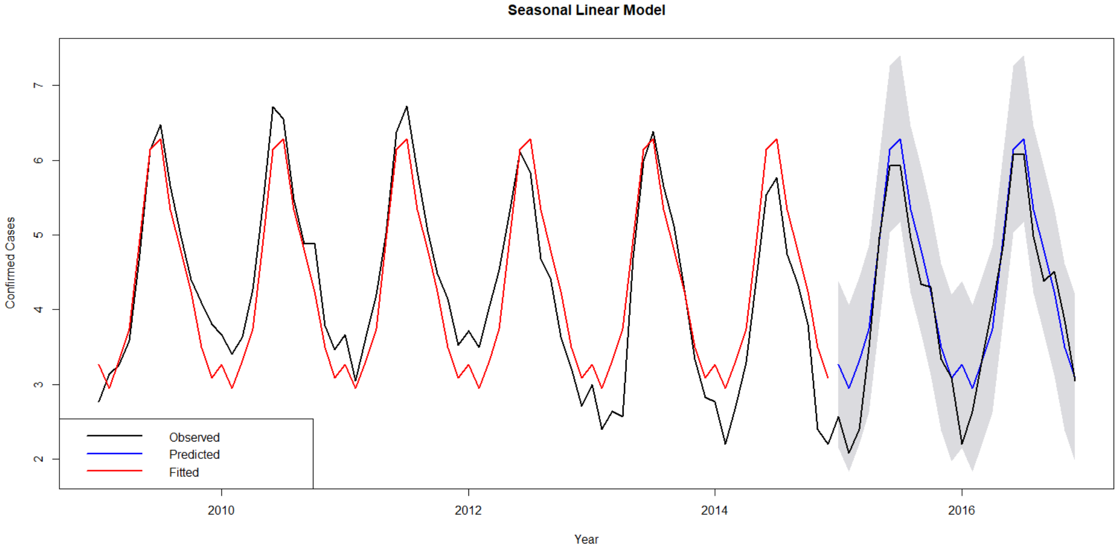


Figure 9:

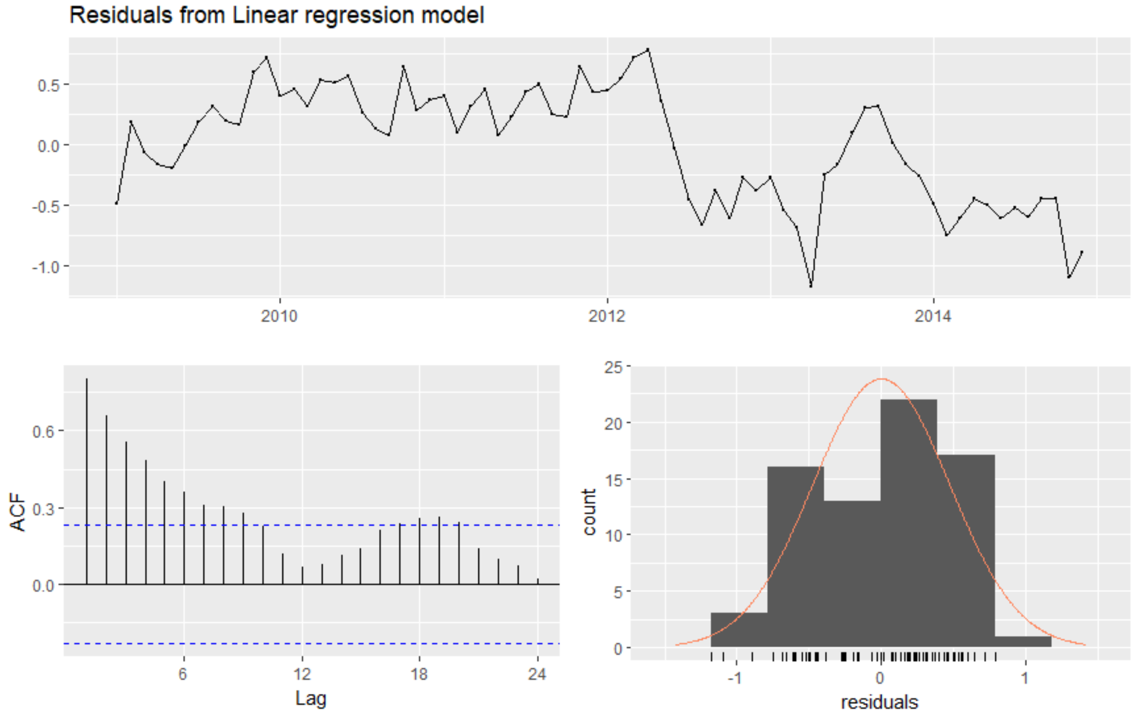

Residuals from Linear regression model
0.5
0.0
-0.5
-1.0
2010
2012
2014
ACF
0.6
0.3
0.0
6
12
18
24
Lag
count
25
20
15
10
5
0
-1
0
1
residuals

Figure 10:

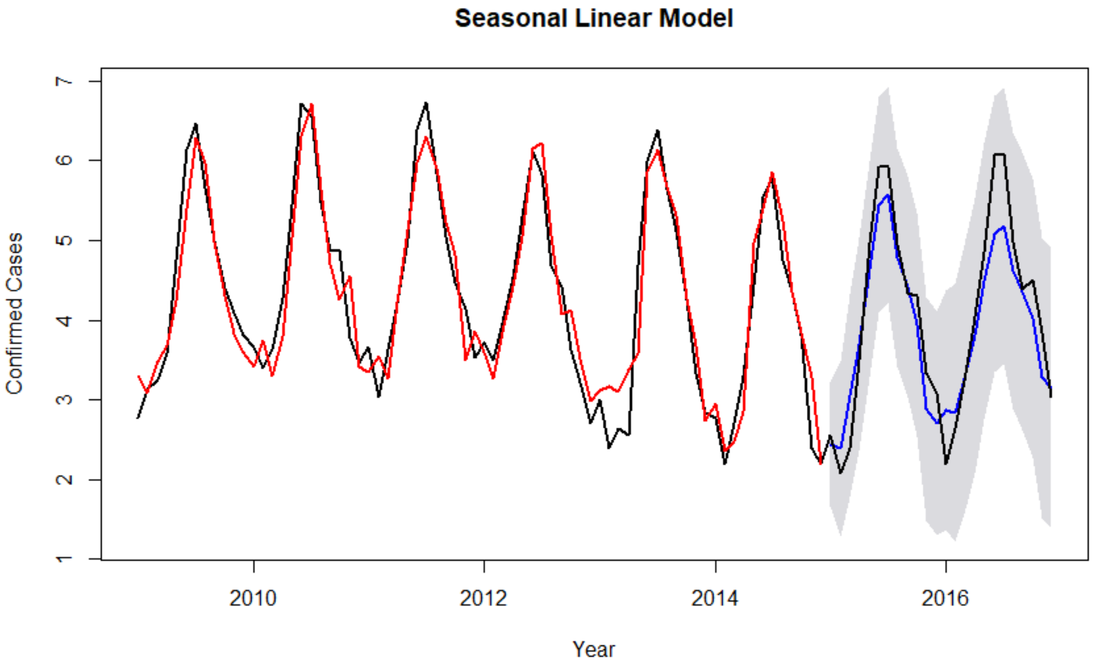

Seasonal Linear Model
Confirmed Cases
1
2
3
4
5
6
7
2010
2012
2014
2016
Year

Figure 11:

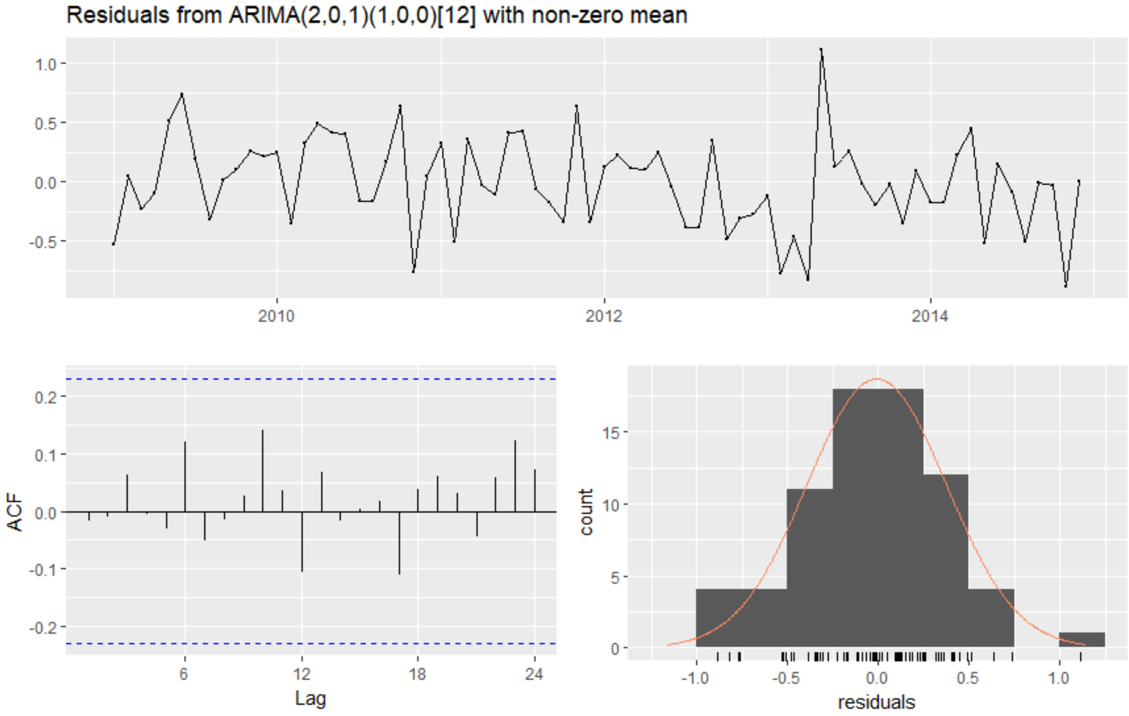
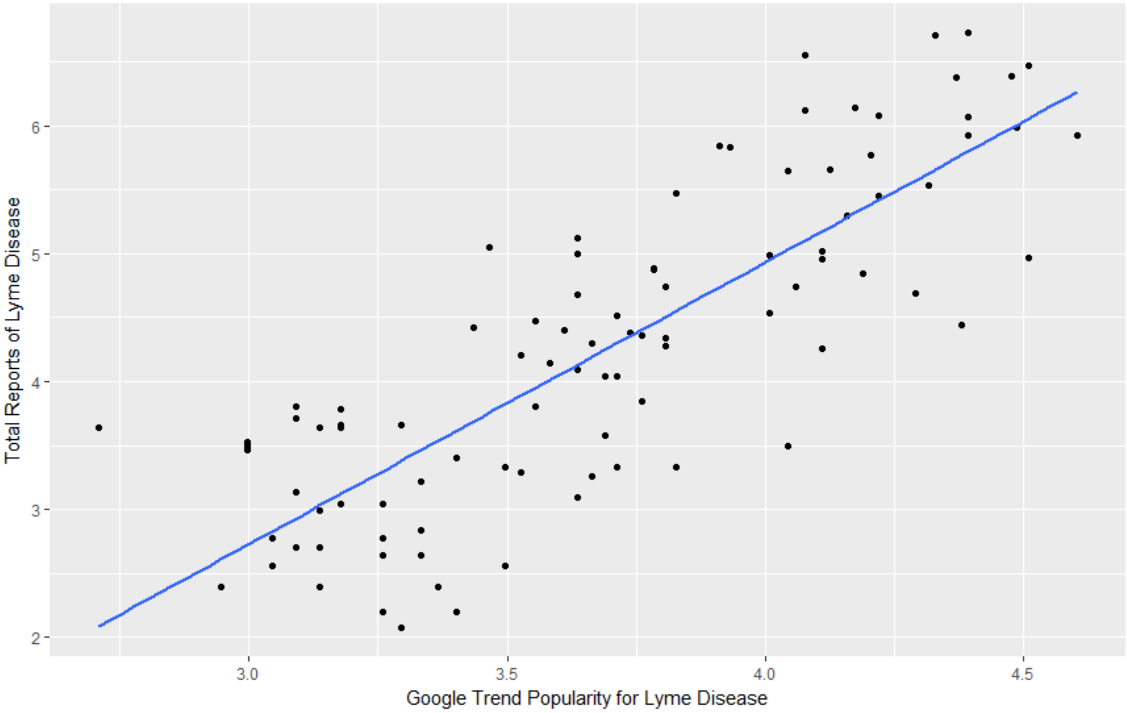

Residuals from ARIMA(2,0,1)(1,0,0)[12] with non-zero mean
1.0
0.5
0.0
-0.5
2010
2012
2014
ACF
0.2
0.1
0.0
-0.1
-0.2
6
12
18
24
Lag
count
15
10
5
0
-1.0
-0.5
0.0
0.5
1.0
residuals

Figure 12:



Total Reports of Lyme Disease
6
5
4
3
2
3.0
3.5
4.0
4.5
Google Trend Popularity for Lyme Disease

Figure 13:

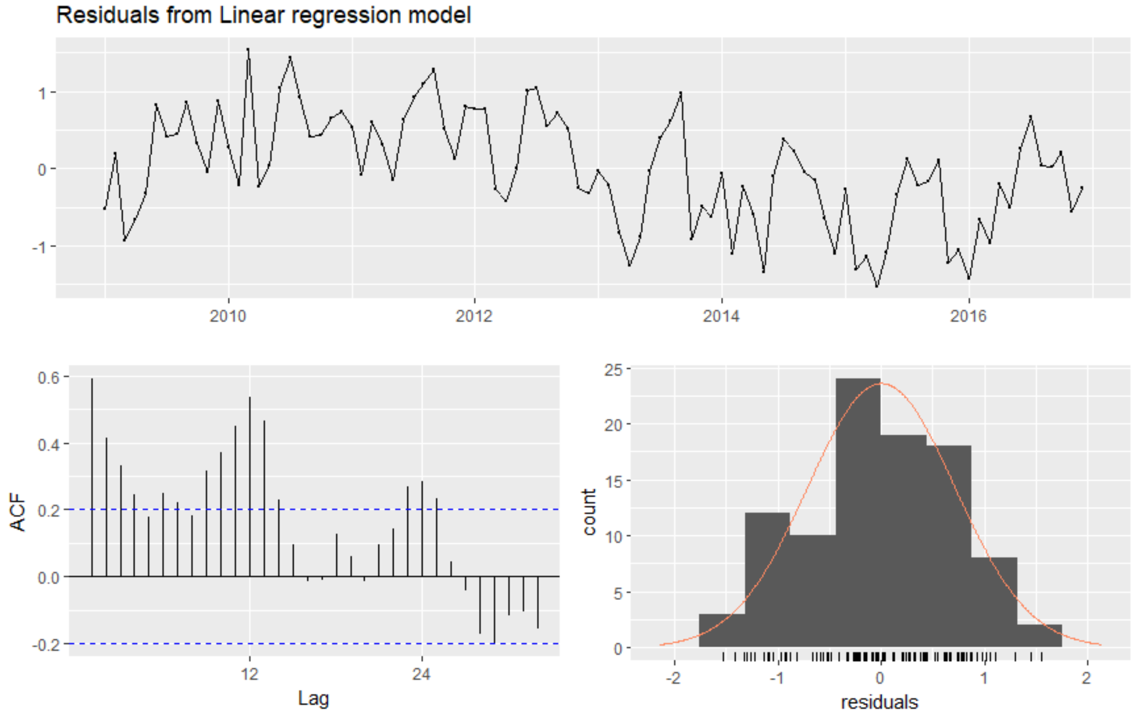

Residuals from Linear regression model
1
0
-1
2010
2012
2014
2016
ACF
0.6
0.4
0.2
0.0
-0.2
12
24
Lag
count
25
20
15
10
5
0
-2
-1
0
1
2
residuals

Figure 14:

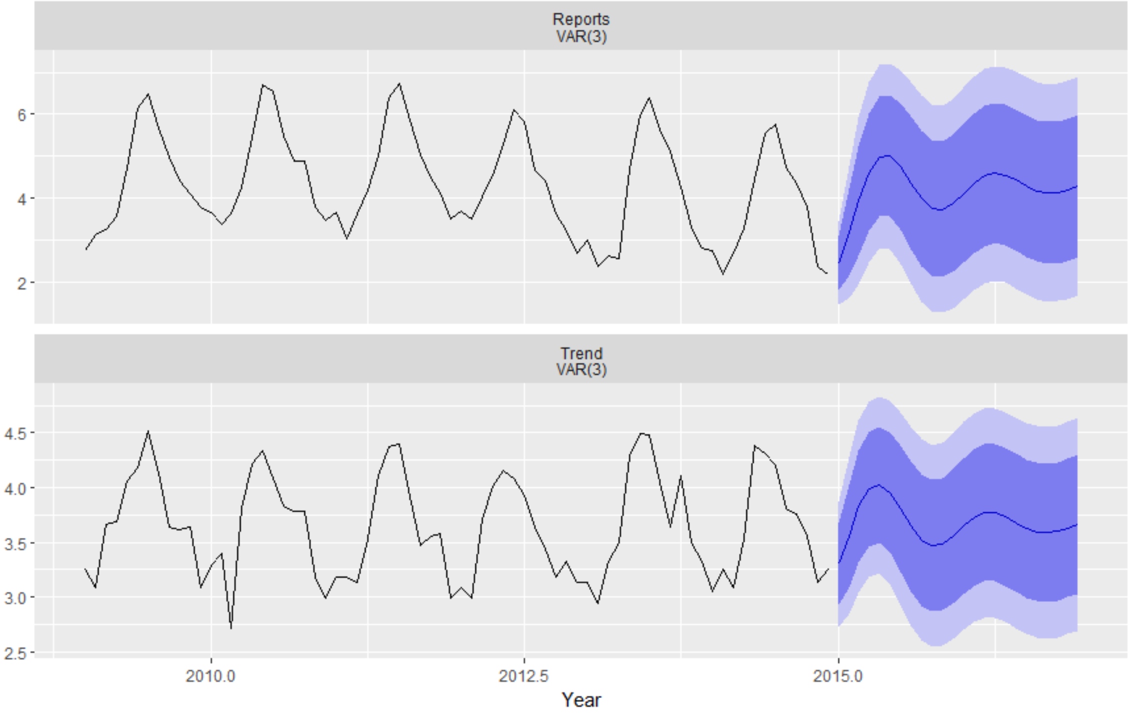

Reports
VAR(3)
6
4
2
Trend
VAR(3)
4.5
4.0
3.5
3.0
2.5
2010.0
2012.5
2015.0
Year

Figure 15:

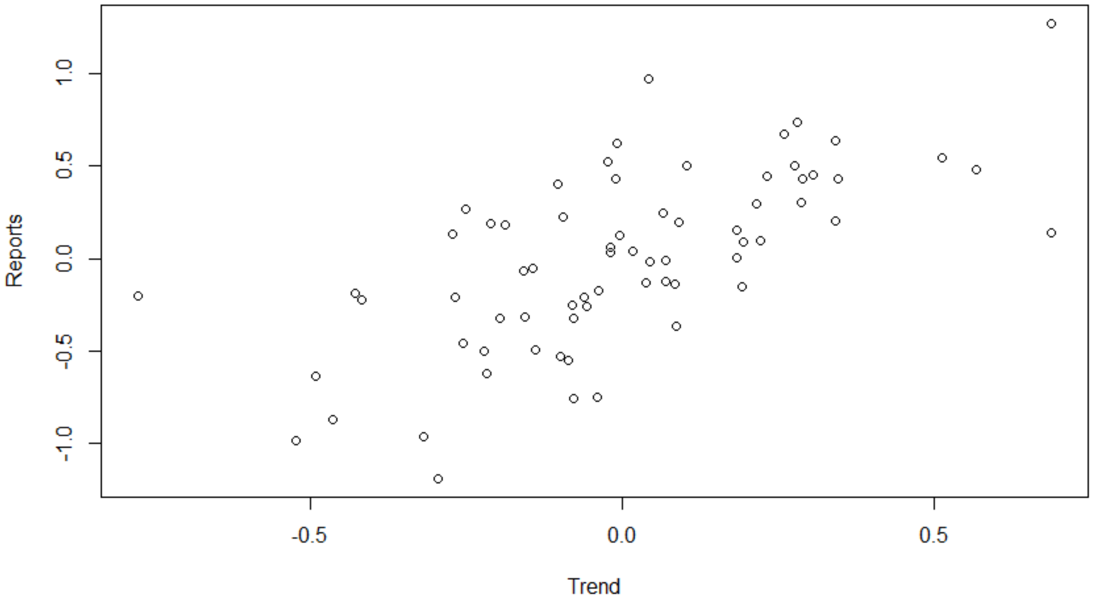


Figure 16:

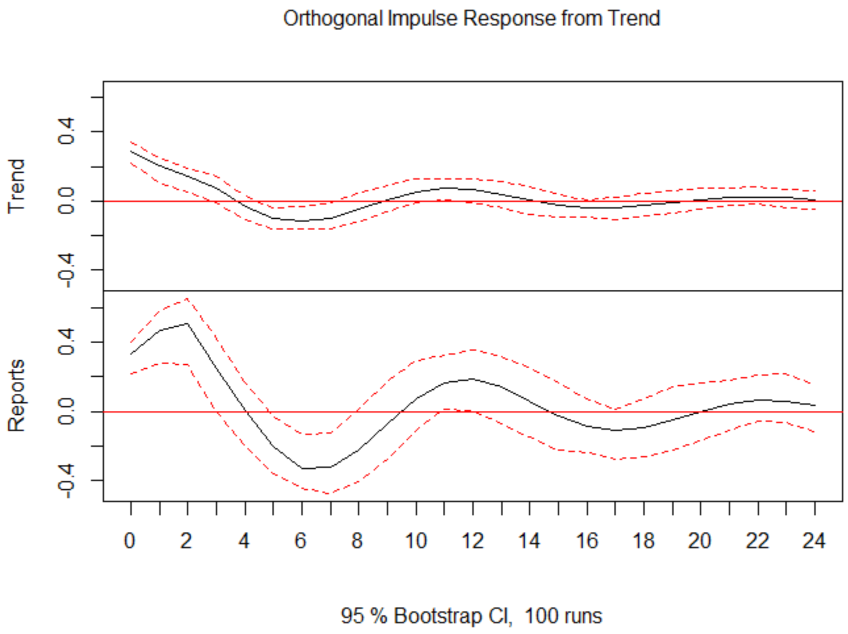

Orthogonal Impulse Response from Trend
Trend
Reports
0.4
0.0
-0.4
0.4
0.0
-0.4
0
2
4
6
8
10
12
14
16
18
20
22
24
95 % Bootstrap CI, 100 runs

Figure 17:

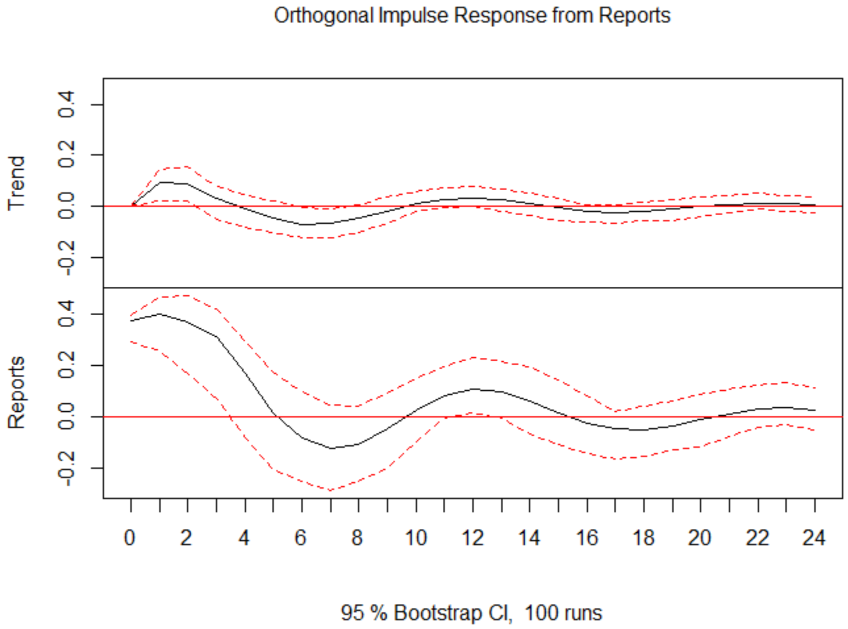

Orthogonal Impulse Response from Reports
Trend
Reports
0.4
0.2
0.0
-0.2
0.4
0.2
0.0
-0.2
0
2
4
6
8
10
12
14
16
18
20
22
24
95 % Bootstrap CI, 100 runs

Figure 18:

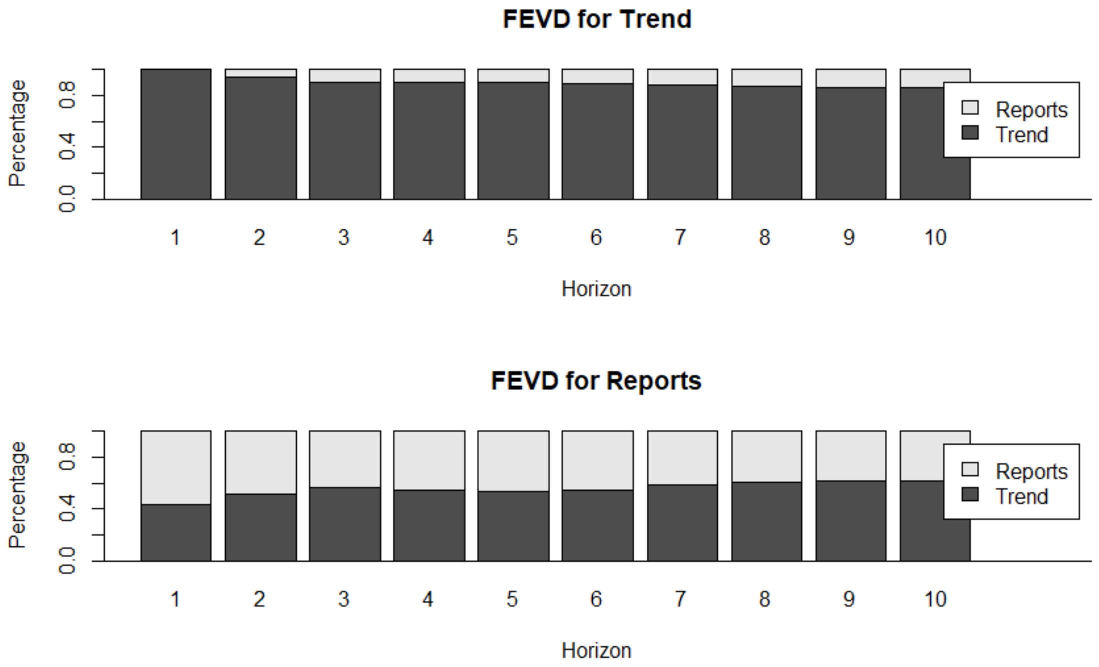
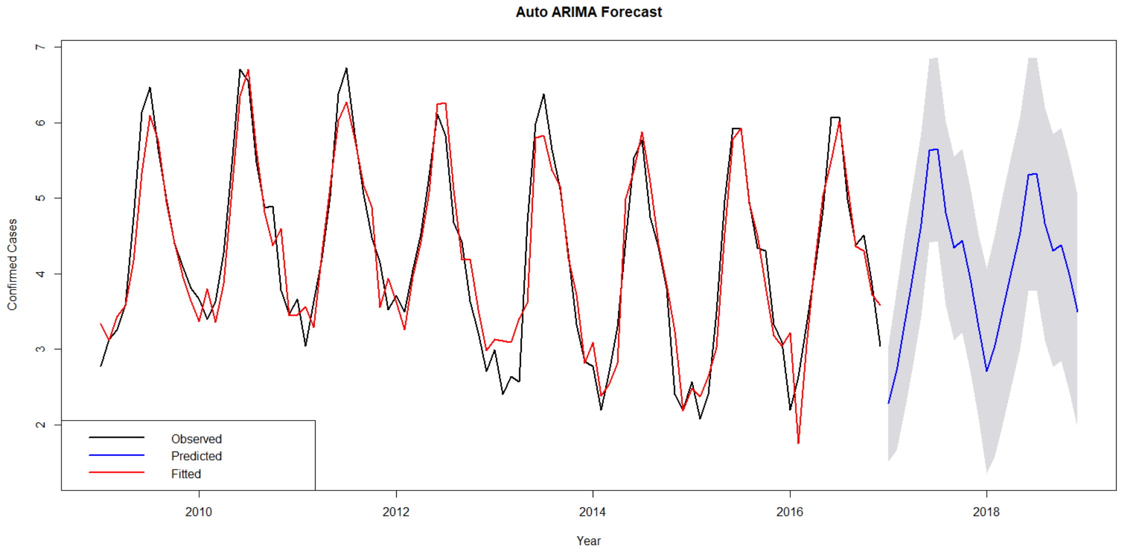


Figure 19:


Auto ARIMA Forecast
Confirmed Cases
Observed
Predicted
Fitted
Year