Time Series Analysis on Philadelphia Precipitation

Cory Boyko

Harrisburg University of Science and Technology

Pavlo Buryi, Ph.D.

ANLY 565 Time Series and Forecasting

Abstract

This study investigates if precipitation from Philadelphia can be modeled using standard time series models. The data was split into the following time frames: daily, weekly, monthly, seasonal, and yearly time series. Holt Winters, AR, linear regression, and ARMA models were used for the stationary series along with a SARIMA model to check. Overall, none of the models are too accurate as the best models are most likely outside the scope of this course.

Keywords: Meteorology, Time Series, Stationary Models, R programming

Time Series Analysis on Philadelphia Precipitation

Historically, weather, specifically precipitation, has been one of the hardest physical phenomena to predict, but what does that actually mean? Can we predict when it will rain based on past data or does it have too much randomness applied to it? This research studied those exact questions. Weather is a part of everyone’s daily lives whether you live in a climate where it rain can be expected most days, or a climate that experiences strong storm systems, or even a temperate climate with little precipitation. However, how much can be explained through time series? Precipitation was chosen for this research, because it can generally have the most effect on a person. If it doesn’t rain then it generally does not affect anyone. However, if it doesn’t rain for a long time then droughts can cause serious issues. When it does rain, it can affect many industries such as outside sporting events, and travel. One aspect to note about this is that precipitation includes snow as well as rain. Precipitation also appears to be somewhat cyclic. For example, in the average climate, it cannot rain every day, but it also has to rain some days. People also usually believe that certain time periods, whether its months or seasons, experience the same general amount of precipitation every year. It is also interesting to see if the series are stationary or if the variance and mean change due to climate change.

# Motivation and Hypothesis

To study this, fifty years of precipitation data for Philadelphia, Pennsylvania from the National Oceanic and Atmospheric Administration. Philadelphia was chosen, because Pennsylvania experiences the ‘typical’ four seasons of the year. They experience a snowy winter, a rainy spring, a summer that is prone to storms, and a fall that could experience rain or snow. However, it is not an extremely rainy climate such as Seattle or much of the United Kingdom. The weather in Philadelphia was possibly more ideal since there is a preconceived notion that the amount of precipitation changes with time on a yearly basis. However, there are many different time frames to study for weather with each serving its own uses.

The first question that was studied was whether a model could be created to predict the amount of precipitation and what type of model is best. Multiple types of models were compared that will be discussed in the methodology section. The second question that was studied, was if the time frames that were used to make the predictions were important. Those include daily values, weekly values, monthly values, seasonal values, and yearly values. The hypothesis of this study is that the larger time frames like seasonal and yearly precipitation models will be the best, possibly due to the fact that the volatility of weather averages out over time. Daily or weekly predictions will be difficult, because of the short-term fluctuations with weather. AR models will be the best models for the series.

# Data and Methodology

This dataset has an observation for each day from January 1, 1970 to December 31, 2019 with the amount of precipitation recorded at the Philadelphia International Airport. The airport was used, because it had the most complete set of data for the time period desired. This dataset was used as the base to create other aggregates that will be further explained.

As previously mentioned, one aspect of this study will be to look at different time frames. The first time frame that was investigated was daily. No aggregation was needed for this as the data came in a daily format. The second time frame that was investigated was weekly. This was done by creating a flag for what number week in the entire dataset it was. R was then used to replicate the week number seven times. Functions within the tidyverse packages were used to aggregate on the week number. The third time frame that was investigated was monthly. The year and month values were split in excel and then R was used to paste together the month and year variables. A new set was created by aggregating on the new month-year variable. The fourth time frame that was investigated was the season. This created four seasons, but not necessarily in the usual quarterly format. Winter was defined as December, January, and February. Spring was defined as March, April, and May. Summer was defined as June, July, and August. Fall was defined as September, October, and November. A new flag was created by concatenating on the season and year variable and finally aggregated. The fifth and final time frame investigated was yearly. The data was simply aggregated on each year. There are pros and cons to using different lengths of time. The shorter times like daily and weekly are more important, because the time is more immediate and would have a greater application. Longer times like monthly and seasonally are important still because people like to know what to expect, especially when the precipitation is in the form of snow and ice. However, the yearly time series is less important, because the model will tell you how much precipitation to expect, but no other information on when it will be.

The second part was the analysis of the data. The first step to analyzing the data was to view the plots of the series, the autocorrelation plots, and the partial autocorrelation plots. After that, the series were checked for stationarity. Assumptions could be made based on the plots, but stationarity was still checked by the Augmented Dickey-Fuller test. After that, multiple different stationary models were applied such as Holt Winters, AR(p), linear regression, and ARMA(p, q). SARIMA was tested for the yearly series as well. To evaluate the models, predictions were made on four lags ahead and were evaluated by using mean absolute percentage error (MAPE). Final model evaluation was based on MAPE as well as the significance of the coefficients. The residuals were investigated as well.

# Results

## Time Series, ACF, and PACF Plots

The first step in the analysis was to begin by inspecting the time series plots, as well as the plots for autocorrelation and partial autocorrelation. The first series was the daily series. Based off of the plot, the series appears to be stationary. One aspect of this plot to note is that many values are at 0. This makes sense since it rains less days than it doesn’t. In turn, this also makes it difficult to predict and difficult to use MAPE since many actual values will be 0 making the denominator of the formula 0. Because of this, the data is too random and will not be modeled. The ACF plot appears to have the first three lags to be significant with the first value being positive and the next two being negative. However, all three have a very small magnitude and outside of direction, are probably not practically significant. The partial autocorrelation plot also has the first two lags being significant, but most likely not practically significant.

The next series that was investigated was the weekly series. This series appears to be stationary as well. It appears that there are still weeks with no precipitation, but at a much less frequency than the daily series. One interesting aspect of this, is that the weekly series is small enough where you can still see the spikes in precipitation in certain weeks. The ACF plot shows that it doesn’t have many significant lags and those that are might be due to random error. Nonetheless, models were still attempted. The PACF plot similarly has few, random significant lags. This may be due to random error as well.

The third series to be investigated was the monthly series. This series appears to be stationary as well. This is the first series where it appears seasonal fluctuations can be seen. If seasonal effects were to exist then the series would not be stationary. This was checked and will be discussed in the next part of the paper. The ACF plot shows that the first lag is significant. Though it is relatively small, it is the highest seen between the other series so far and could be practically useful. The PACF plot also shows the first lag being significant as well as the nineteenth lag.

The fourth series that was investigated was the seasonal series. This one appears to stationary as well, thought the end of the series looks questionable. The ACF plot for the seasons has no significant lags meaning it could be similar to white noise. Modeling was still be used. The PACF plot has one lag that is significant, but still appears to be very low.

The fifth and final series that was investigated was the yearly series. Like the series before, this one appears to be stationary as well, but the right tail might be experiencing more variation. Both the ACF and PACF plots do not have any significant lags which shows the yearly series will be difficult to model as well.

## Augmented Dickey-Fuller Tests

Based off of the plots, all of the series appeared to be stationary. This will now be tested by using the Augmented Dickey-Fuller Tests. The daily series has a Dickey-Fuller value of -24.307 which produces a p-value less than 0.01. The null hypothesis cannot be rejected and the series is taken to be stationary. The weekly series has a Dickey-Fuller value of -12.881 which also produces a p-value less than 0.01. The weekly series is stationary as well. The monthly series produced a Dickey-Fuller value of -7.7675and a p-value of 0.01. The monthly series is stationary. The seasonal series had a Dickey-Fuller value of -6.6464 and again a p-value of 0.01. The seasonal series is stationary. The yearly series had a Dickey-Fuller value of -2.7625 which had a p-value of 0.2684. Therefore, the null hypothesis is rejected and the series is not stationary. However, based on the ACF and PACF plots show that is similar to white noise and is most likely still stationary and stationary models will still be applied as well as a non-stationary model out of an abundance of caution.

## Holt Winters

The first type of modeling that was used was Holt Winters. The model for the weekly series gives an alpha of 0.013, signifying that there is a lot of smoothing. This sounds reasonable due to the variation in weekly precipitation. Based off of the prediction of four lags ahead, the MAPE is 105.2883%. The model is overestimating the amount of precipitation that will be received, which could be seen as a positive. In the weeks where it does rain, the model is actually quite close. The ACF plot of the residuals shows that some lags are still significant, so overall this is not a great model.

The next series is the monthly series. Applying Holt Winters to the monthly series gives an alpha of 0.0655, again a lot of smoothing, but slightly less than the weekly series. The MAPE of this model was 152.414%. This model also overestimates the amount of precipitation, but at no point is that close to the actual precipitations. The ACF plot looks similar to the plot without the model. Similar to the weekly model, this is also not a good model.

The third series is the seasonal series. The alpha of the seasonal Holt Winters model is 0.0409, another model with a lot of smoothing. The MAPE of this model was 31.849% indicating that this was a relatively good model. The predictions were generally in the middle of the actuals, showing that this gives unbiased estimates. The ACF plot shows no significant lags overall, adding to evidence that this is a good model.

The fourth series is the yearly series. The alpha of this series is 0.1197, showing that this series needs to least smoothing out of the series tested. The MAPE of this model was 17.598% and showed unbiased estimates. The ACF plot of this model also shows no significant lags, so overall this is a good model as well.

## Autoregressive Models

The next set of modeling that was done was AR(p) models. Based off of the original ACF plots, AR looks like they could provide the best models. The best model for the weekly series is AR(6). However, only the 4th and 6th coefficients are statistically significant at the 95% confidence level with both being positive. MAPE for this model is 79.7363%, an improvement of about 25%. The ACF plot only shows one significant lag, which could be to random error. Overall, this is an ok model that still has a high MAPE and insignificant coefficients. It is an improvement over the Holt Winters model.

The best model for the monthly series is AR(1). This single coefficient is positive and statistically significant at the 95% level. The MAPE for this model at four lags is 113.6695%, again an increase, but overall not great because of such a high error rate. From the plot it also looks like it is overestimating the amount of precipitation. The ACF plot has two significant lags, but the rest appears to be noise. This is a better model than Holt Winters, but there is room for improvement.

The best model for the seasonal series is surprisingly AR(0), meaning that this is most likely just white noise. The MAPE for this model is 33.316% and is worse than the Holt Winters model, which makes sense since there are no actual AR terms. The ACF plot shows that there are still 3 significant lags. Overall, the AR model is not a good model for the seasonal series.

The best model for the yearly series is also AR(0). MAPE for this model is 16.387%, which is an improvement from the Holt Winters model, but since it is AR(0) that is possibly due to random error. The ACF plot looks practically unchanged as well. Overall, we can confirm that the for the longer time lengths, autoregressive only models will not provide a good fit for the series.

## Linear Regression

The next set of models are all done with linear regression. The independent variables for each are the seasonal variable as well as time. For the weekly series, neither time nor the number of the week in the year was statistically significant. With none of the independents being significant, the MAPE was 82.842%, which was worse than the AR series. There are also significant lags in the ACF plot showing that the model is overall not a very good fit.

For the regression model for the monthly series, neither time nor any of the months again are statistically significant. The MAPE for this model was 122.766% which was also worse than the AR model. The ACF plot shows that there are still some significant lags after the model. Linear regression is not a good fit for the seasonal model either.

For the seasonal series, none of the independent variables had statistically significant coefficients. However, this model did provide the lowest MAPE with a value of 30.16681%. The ACF plot shows no significant lags, showing that it is white noise left over. Overall, this is a good, not great model, because of the low MAPE and ACF plot, but the independents still aren’t statistically significant.

Since the yearly series does not have any ‘seasons’ in the typical sense, the only independent used was the time variable. Time had a coefficient of 0.02 and was statistically significant. The MAPE for this model was 16.543%, slightly higher than the previous models. However, this is most likely the best model due to the low MAPE and the significant independent variable. The ACF of the residuals shows no significant lags adding to the evidence that this is a good model.

## Autoregressive Moving Average

The last set of stationary models is the ARMA(p, q) models. These series were done with the maximum p and q being 6. The best order (based on AIC) for the weekly series is AR(6, 5). This model gives a MAPE of 82.937% which is greater than the AR and linear regression models. The ACF plot shows that only white noise is left over, but because the MAPE is greater than the previous models, it is not a good model.

The best order of the monthly series was ARMA(5, 4). The MAPE for this model was 129.563%, again worse than the AR and linear regression models. Based off of the prediction plots, it looks like the predictions were going in the opposite direction of the actual values. The ACF plot shows one significant lag that may be due to random error. Overall, this is not a good model based off of MAPE and the predicted values.

The best order of the seasonal series was ARMA(0,0). Because of this, the model is the same as the AR model which had a MAPE of 33.316%. Both the prediction plots and ACF plots are the same as the AR model. Similar to the AR model, this is not a good fit.

The best order of the yearly series was ARMA(2,2), which is surprising since the best AR model was AR(0). However, with a MAPE of 19.64%, this model preforms the worst of all of the models. The ACF plot shows that only white noise is left over. Overall, this is not a good fit and ARMA is not a good fit for the larger time periods.

## Seasonal Autoregressive Integrated Moving Average

From the Augmented Dickey Fuller test, the yearly time series was nonstationary, despite other aspects showing that it’s probably stationary. Since it might be nonstationary, a seasonal autoregressive integrated moving average (SARIMA) model was fit to the data. All combinations of order (2, 2,2, 2, 2, 2) were checked and the best order was (0, 0, 2, 2,1,1). The MAPE of the model was 22.61% which was the worst of any model tried so far. The ACF plot also looks no different than the original ACF plot. Based on this, it is fair to assume that the stationary models do a better job at modeling the yearly time series than the nonstationary models do.

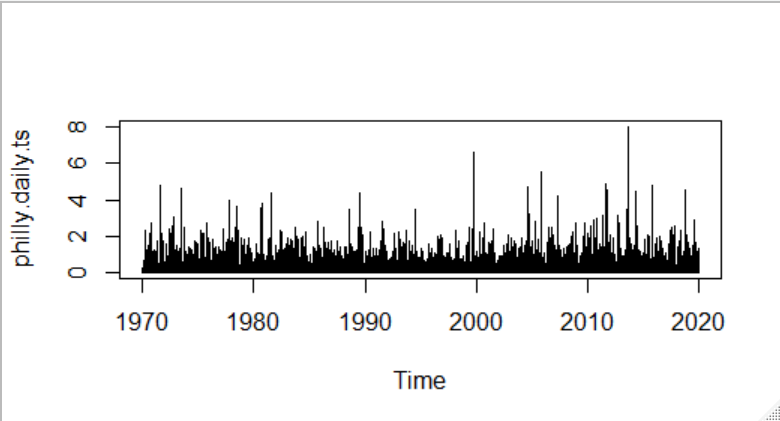
# Conclusion

In conclusion, it is difficult to get a good model to predict precipitation using general time series models. Nonetheless, multiple models were tried and some fit better than others. The best model for the weekly time series was the AR series which was AR(6). It had only two of the six coefficients being statistically significant, but two were significant and the AR model had the best MAPE. The best model for the monthly series was the AR model which was AR(1). The single coefficient was significant, but MAPE was still over 100% showing that the model was not very good. The best model for the seasonal data was the linear regression model. The model itself did not have any significant coefficients but had the best MAPE. Compared to other models, the MAPE was not that much lower than the other models and may be considered insignificant. The best model for the yearly time series was also the linear regression series. The time variable was statistically significant. The model itself had a slightly higher MAPE than the other models, but the other models did not have significant coefficients, so this is practically the best model.

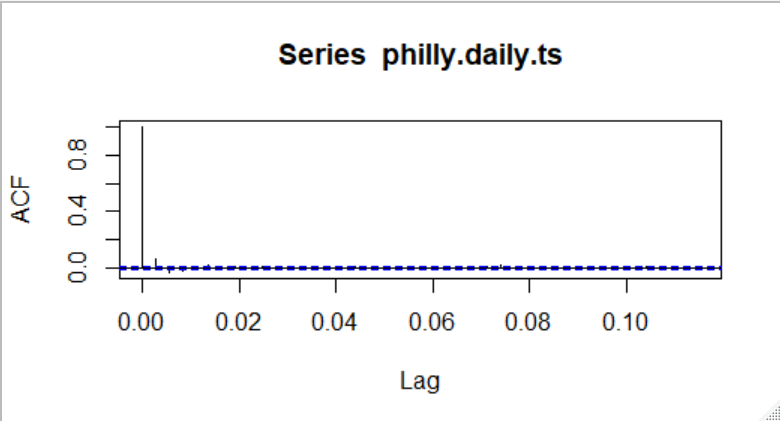
Like all studies, there were some limitations here as well. First, though there was a large sample size of 50 years, more data could be obtained to create the best possible model. Secondly, there are many more models that could have been used that are beyond the scope of the current knowledge. There are most likely specific models just for metrological series. Thirdly, because this research was only precipitation in Philadelphia, the only assumptions can be said are about the precipitation in Philadelphia. If future research is done on cities of similar climates, then the results may be extended, but as of now it can only apply to Philadelphia.

Future research on this topic can be looking at models for different cities in different climates. If cities in the same climate as Philadelphia are studied, the results of this studied can be confirmed or there can be better evidence for other models. If different climates are studied, more information about how the models used in this study can be applied. As more is learned in the field of time series, different models can also be applied. One other area of future research can be using other weather phenomena such as temperature and humidity to see how they relate to precipitation.

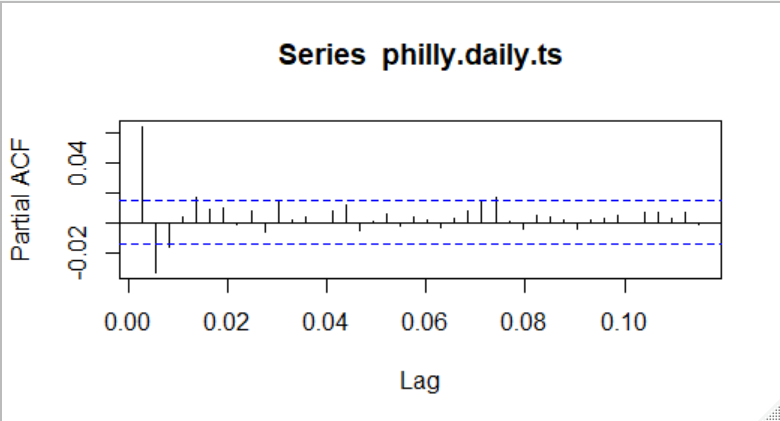
Appendix



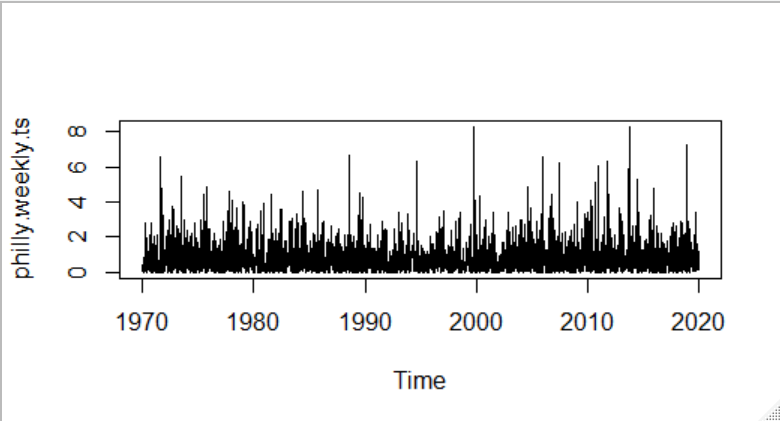
**Figure 1.** Daily Time Series Plot



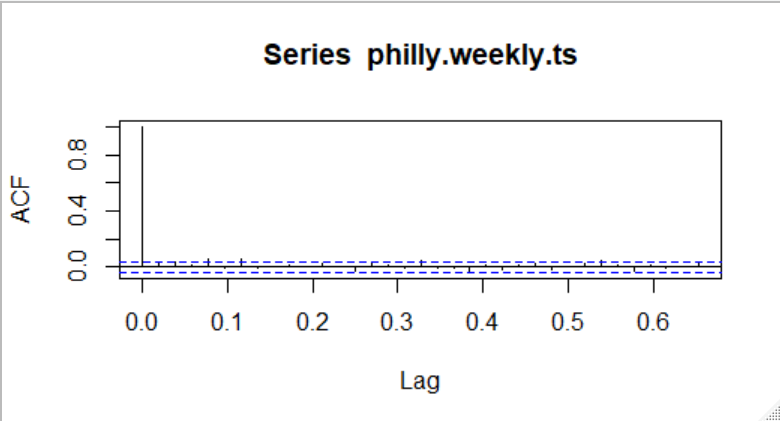
**Figure 2.** Daily Time Series ACF Plot



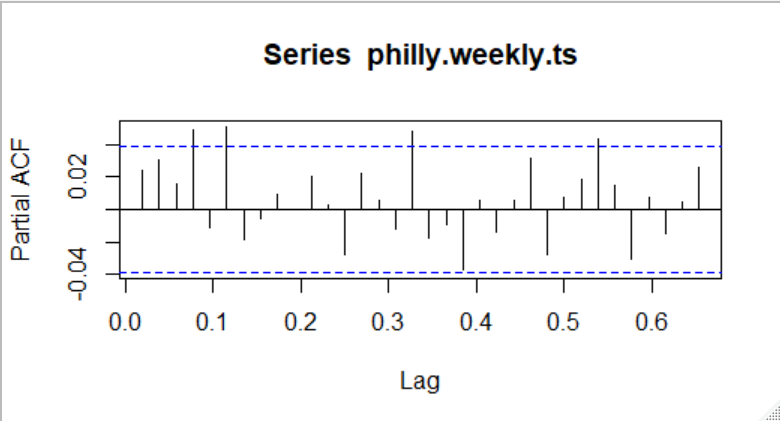
**Figure 3.** Daily Time Series PACF Plot



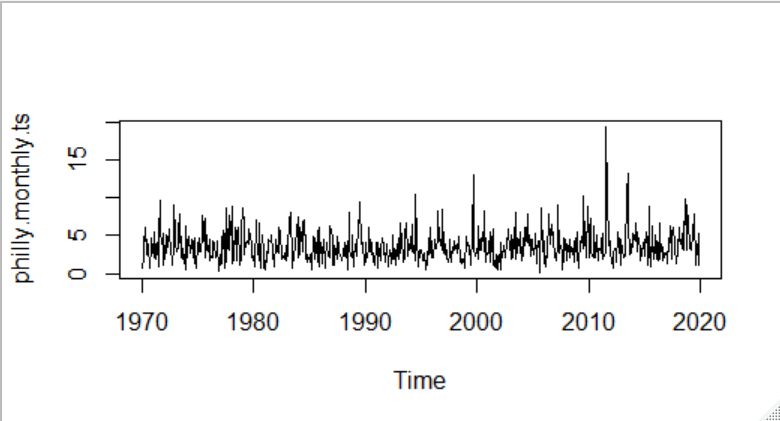
**Figure 4.** Weekly Time Series Plot



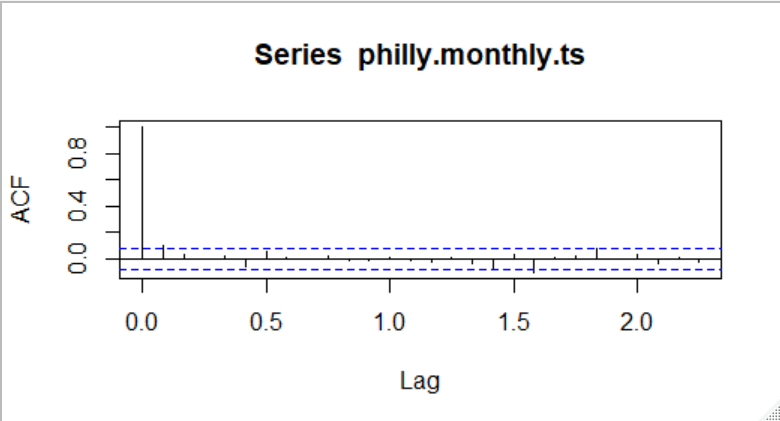
**Figure 5.** Weekly Time Series ACF Plot



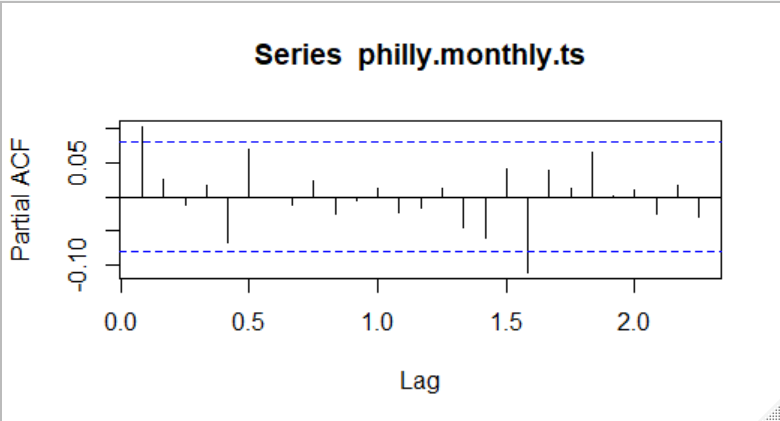
**Figure 6.** Weekly Time Series PACF Plot



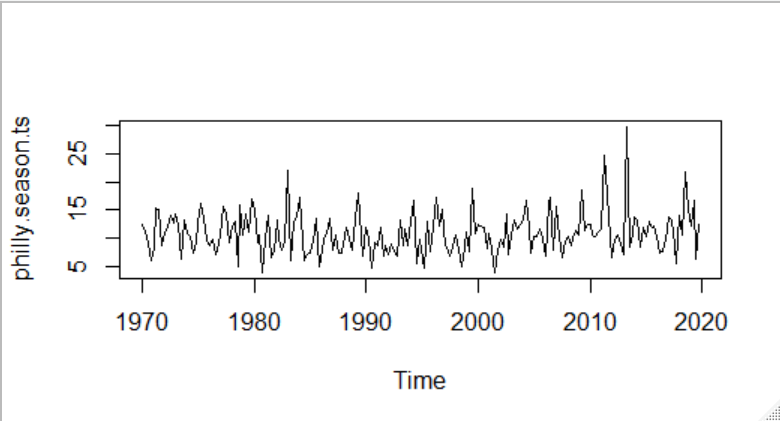
**Figure 7.** Monthly Time Series Plot



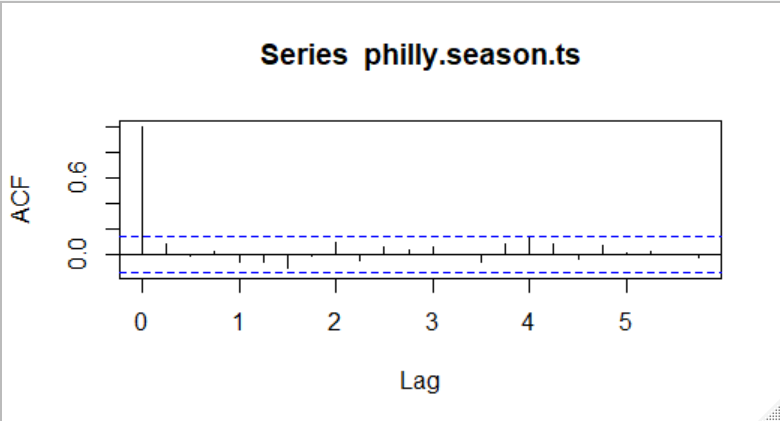
**Figure 8.** Monthly Time Series ACF Plot



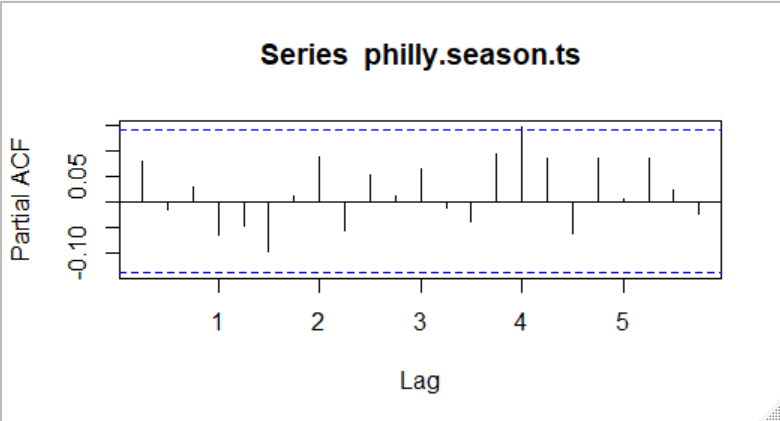
**Figure 9.** Monthly Time Series PACF Plot



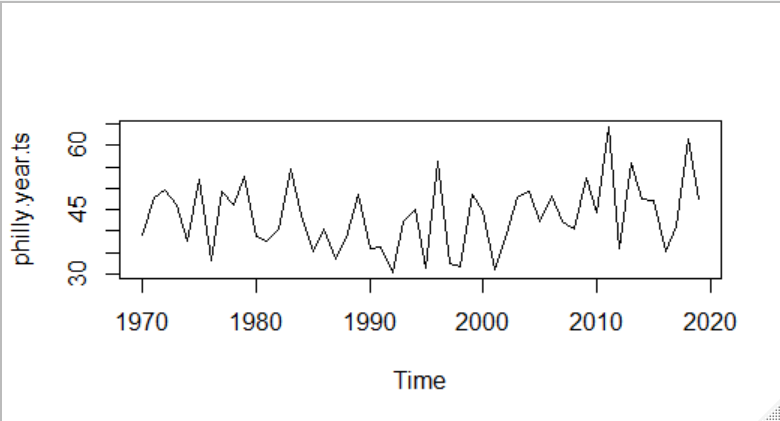
**Figure 10.** Seasonal Time Series Plot



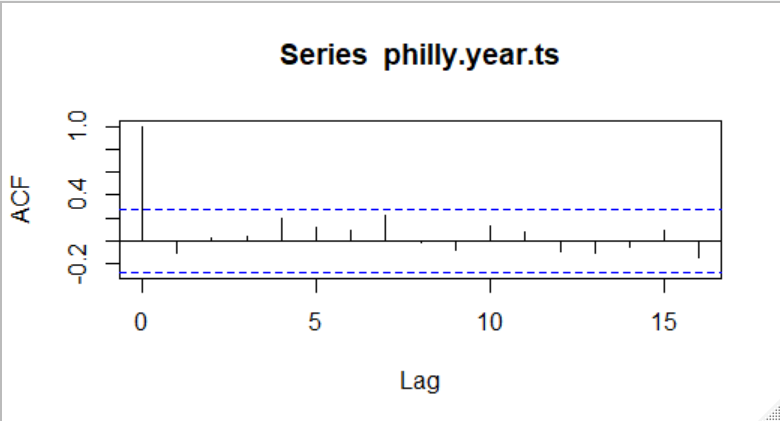
**Figure 11.** Seasonal Time Series ACF Plot



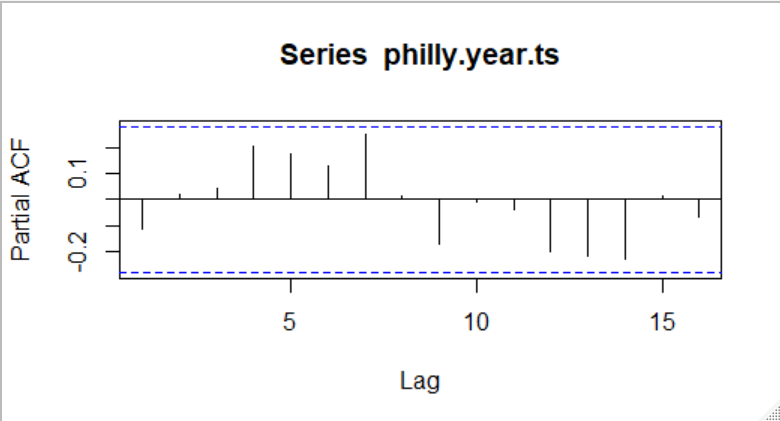
**Figure 12.** Seasonal Time Series PACF Plot



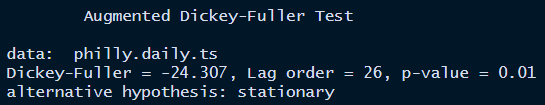
**Figure 13.** Yearly Time Series Plot



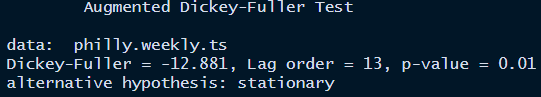
**Figure 14.** Yearly Time Series ACF Plot



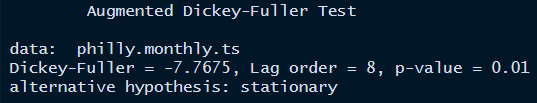
**Figure 15.** Yearly Time Series PACF Plot



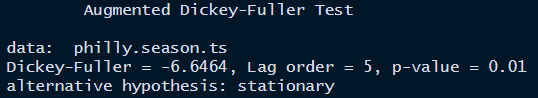
**Figure 16.** Daily Time Series ADF Test



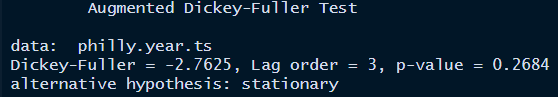
**Figure 17.** Weekly Time Series ADF Test



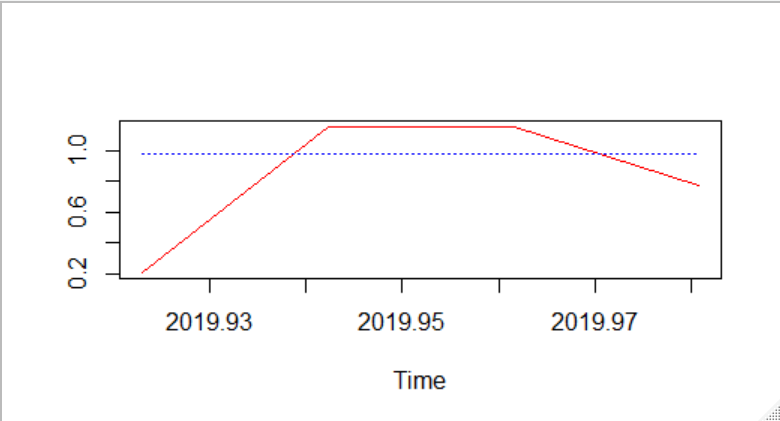
**Figure 18.** Monthly ADF Test



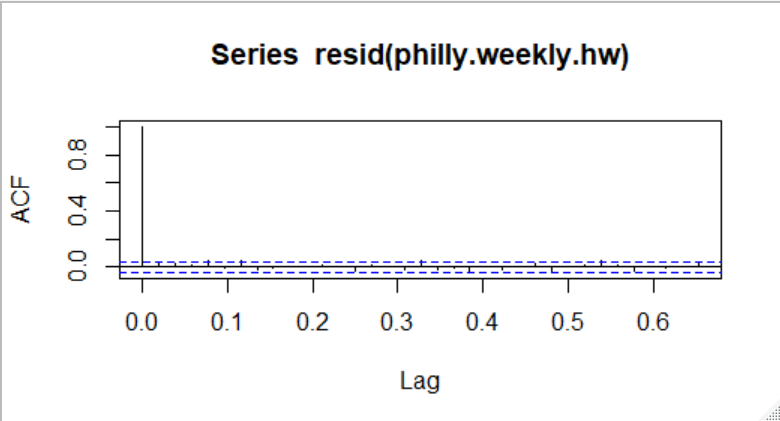
**Figure 19.** Seasonal ADF Test



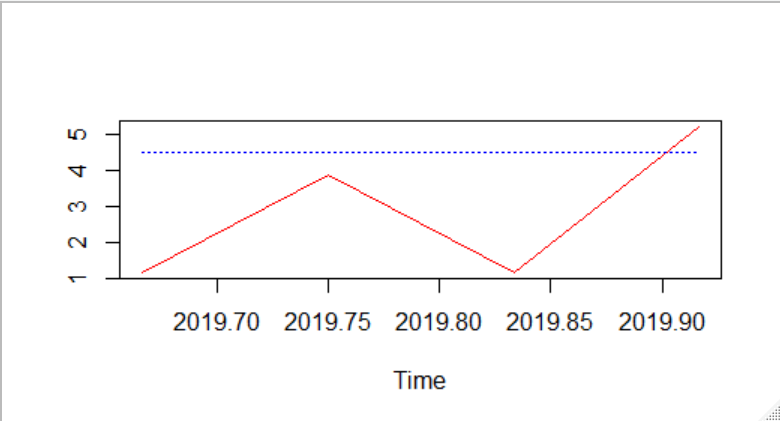
**Figure 20.** Yearly ADF Test



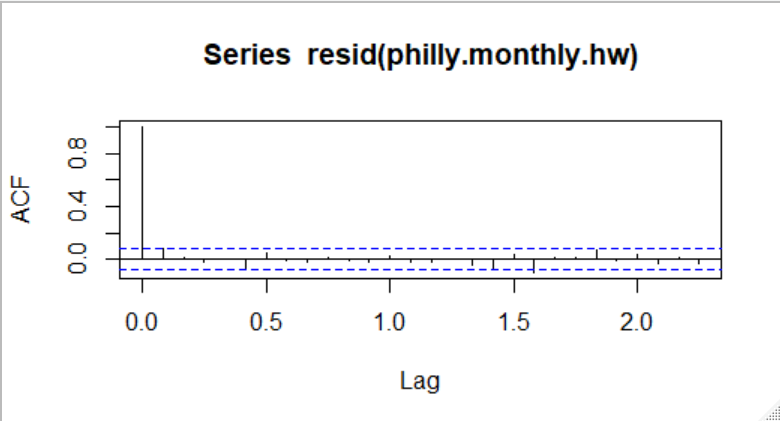
**Figure 21.** Weekly Series HW Prediction Plot



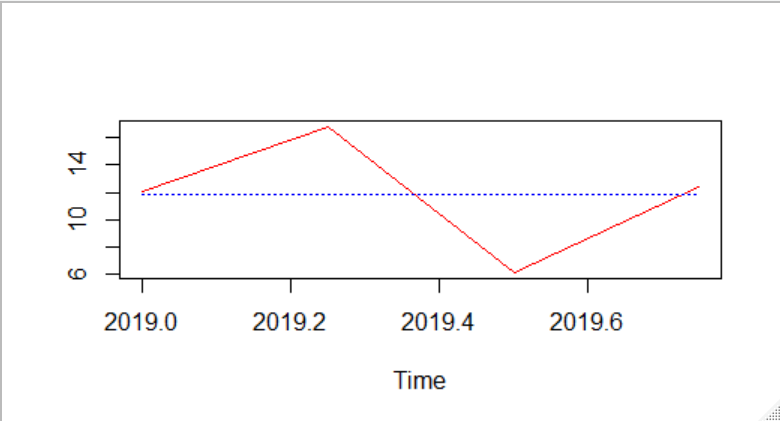
**Figure 22.** Weekly HW Residuals ACF Plot



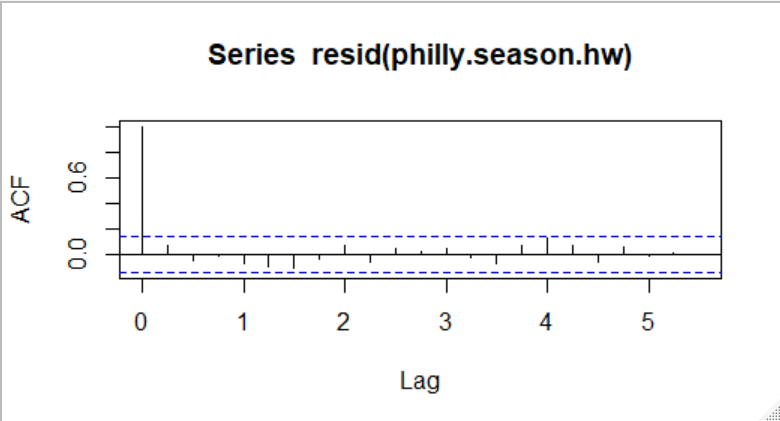
**Figure 23.** Monthly Series HW Predicted Plot



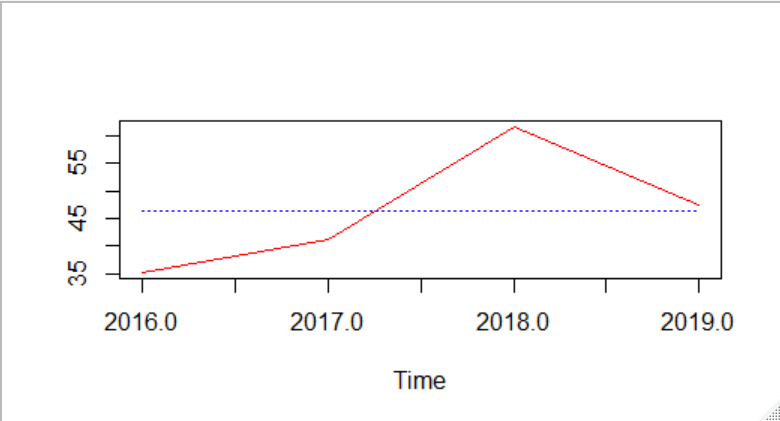
**Figure 24.** Monhtly HW Residuals ACF Plot



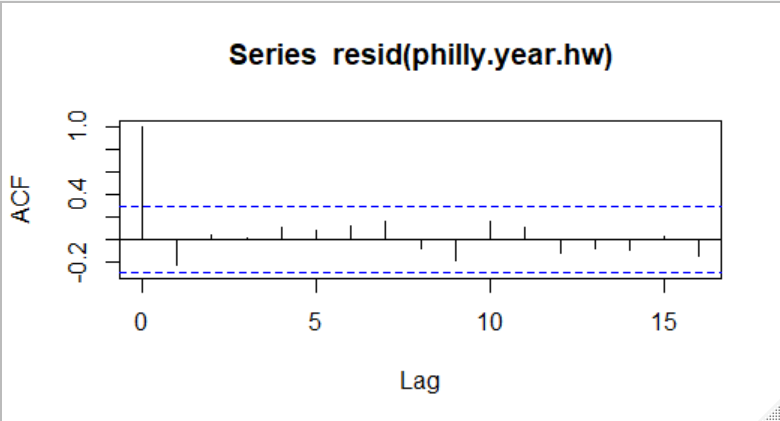
**Figure 25.** Seasonal Series HW Predicted Plot



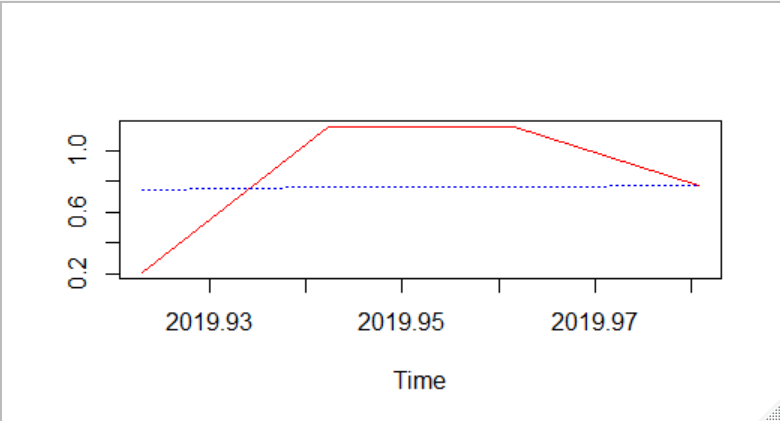
**Figure 26.** Seasonal HW Residuals ACF Plot



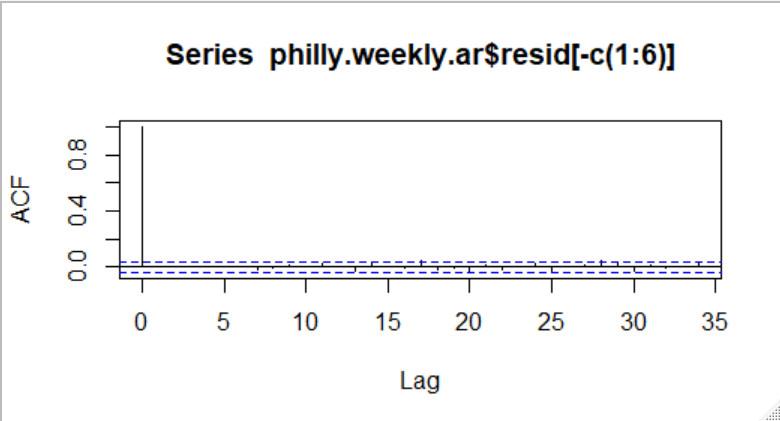
**Figure 27.** Yearly Series HW Predicted Plot



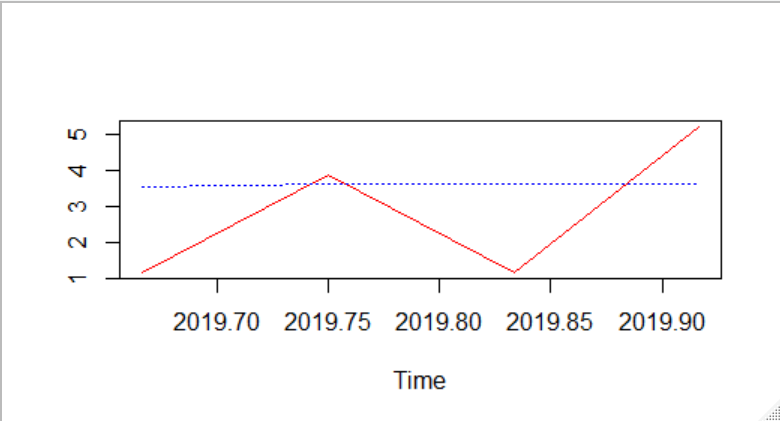
**Figure 28.** Yearly HW Residuals ACF Plot



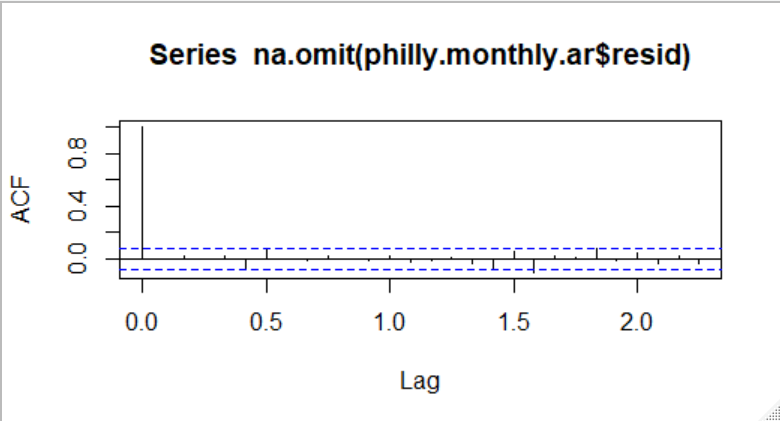
**Figure 29.** Weekly AR Predictions Plot



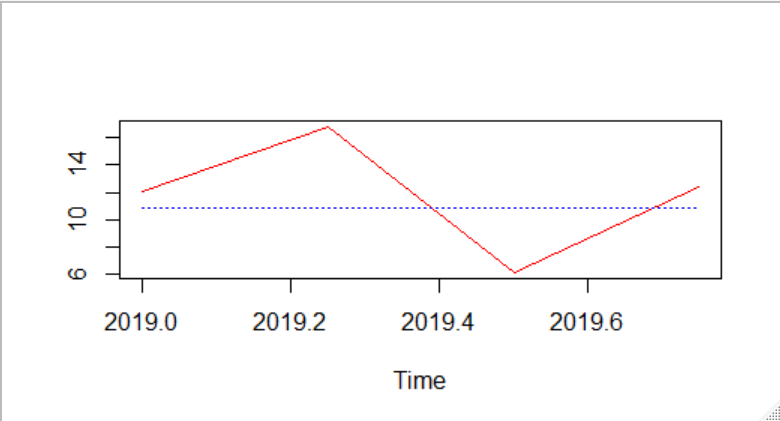
**Figure 30.** Weekly AR Residuals ACF Plot



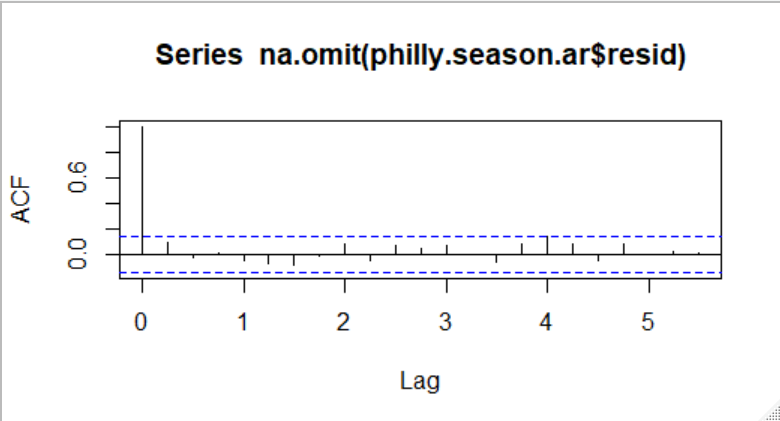
**Figure 31.** Monthly AR Prediction Plot



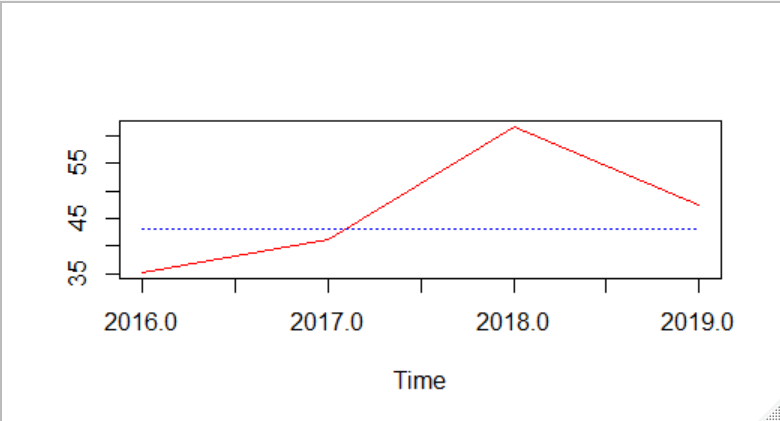
**Figure 32.** Monthly AR Residuals ACF Plot



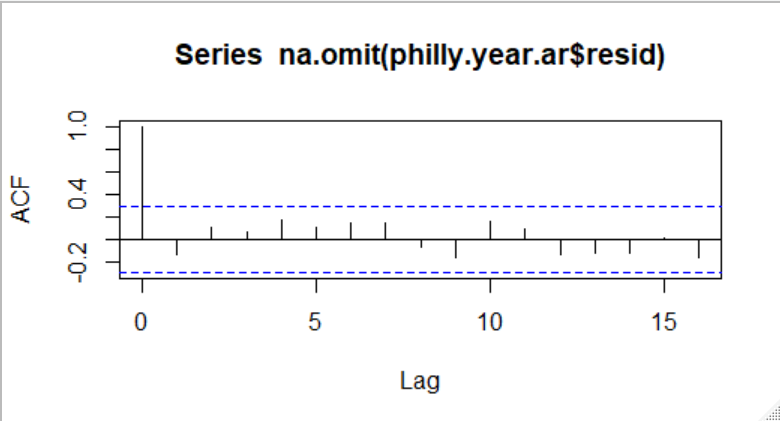
**Figure 33.** Seasonal AR Prediction Plot



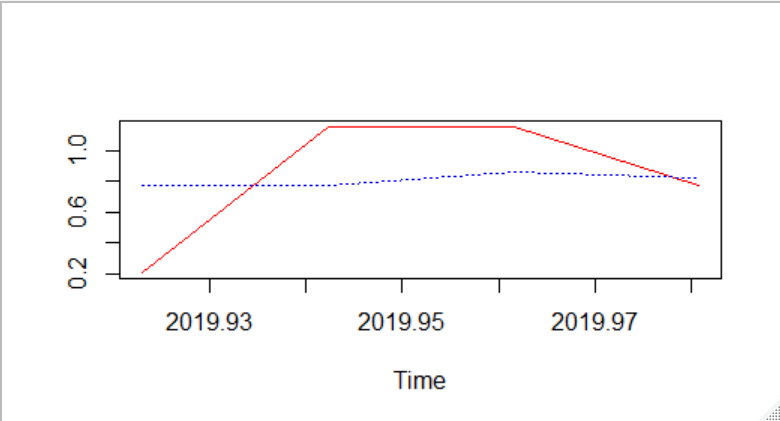
**Figure 34.** Seasonal AR Residuals ACF Plot



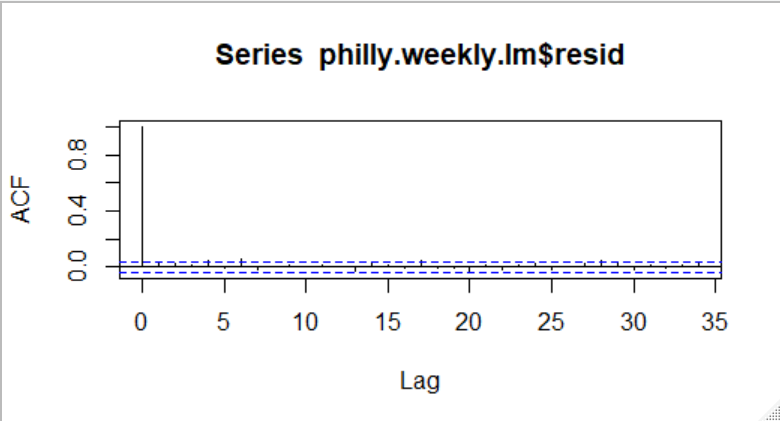
**Figure 35.** Yearly AR Prediction Plot



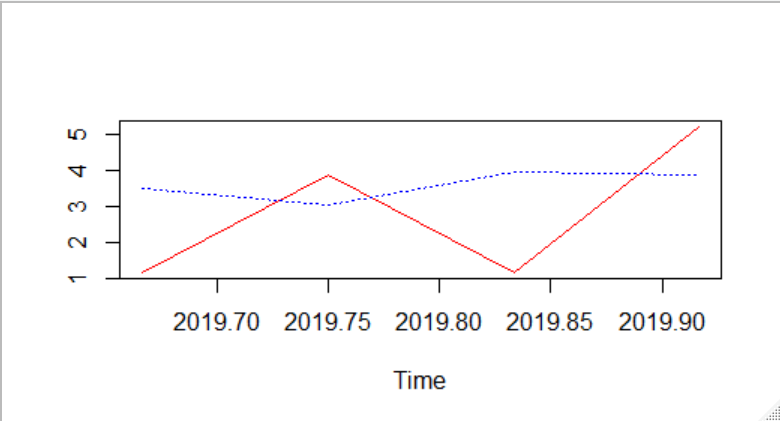
**Figure 36.** Yearly AR Residual ACF Plot



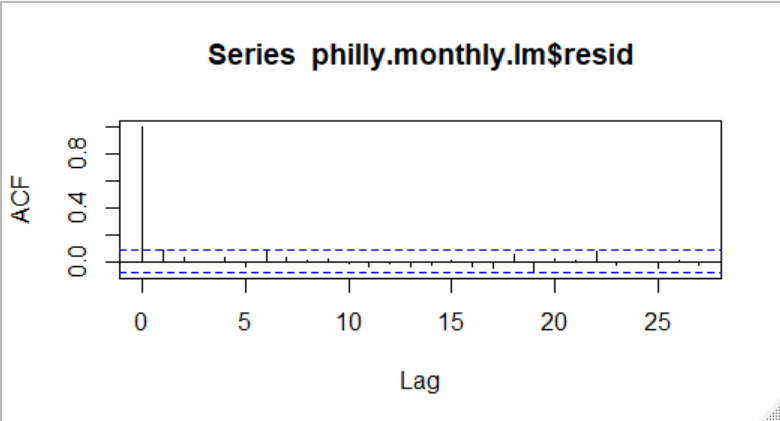
**Figure 37.** Weekly Linear Regression Prediction Plot



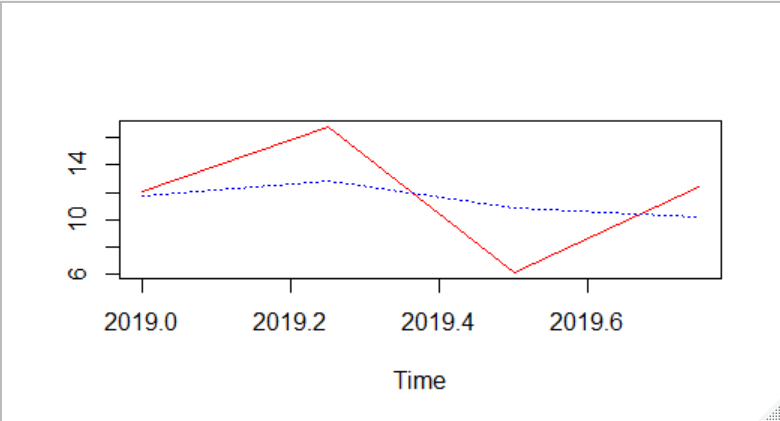
**Figure 38.** Weekly Linear Regression Residual ACF Plot



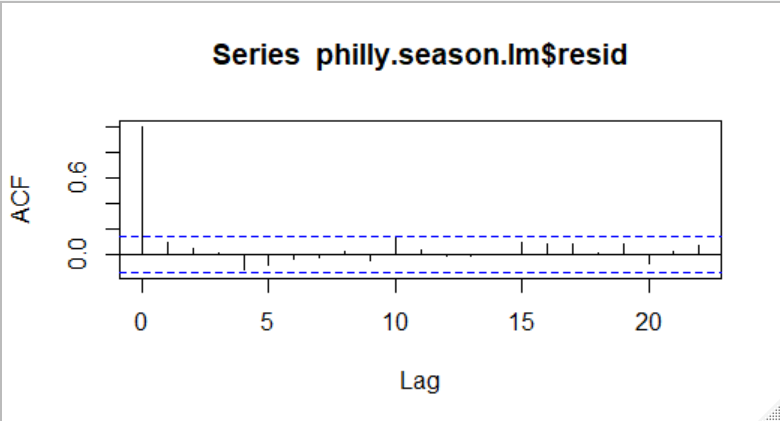
**Figure 39.** Monthly Linear Regression Prediction Plot



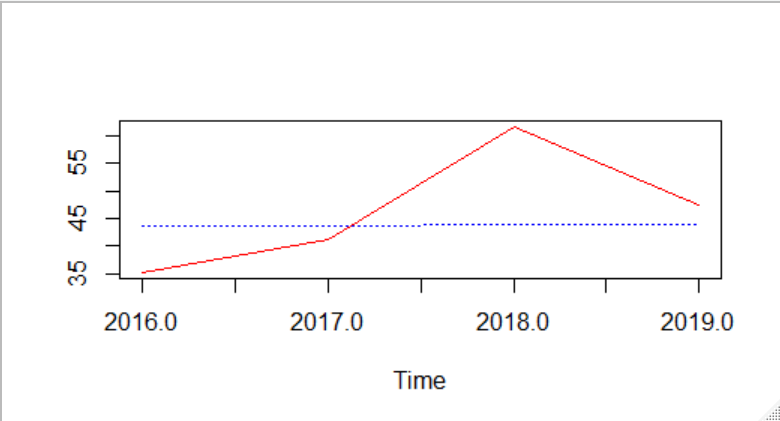
**Figure 40.** Monthly Linear Regression Residuals ACF Plot



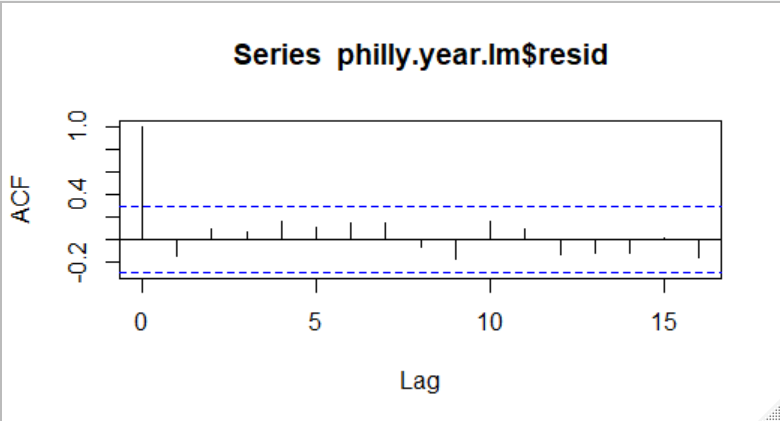
**Figure 41.** Seasonal Linear Regression Prediction Plot



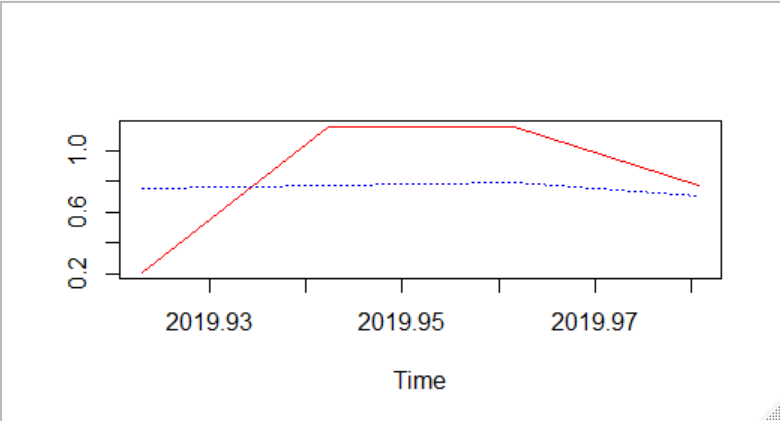
**Figure 42.** Seasonal Linear Regression Residual ACF Plot



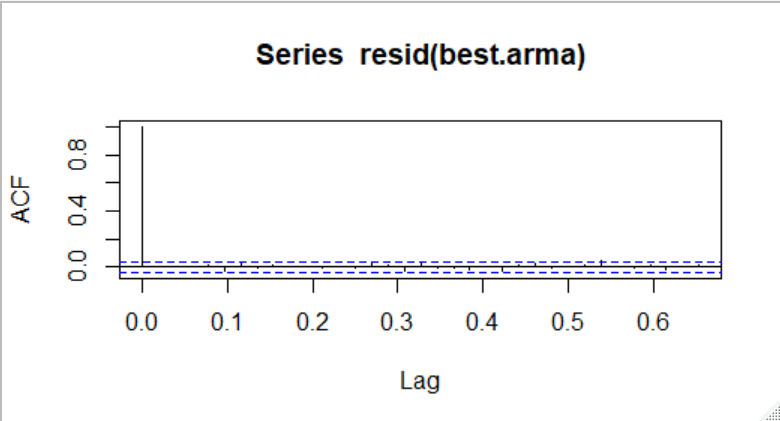
**Figure 43.** Yearly Linear Regression Prediction Plot



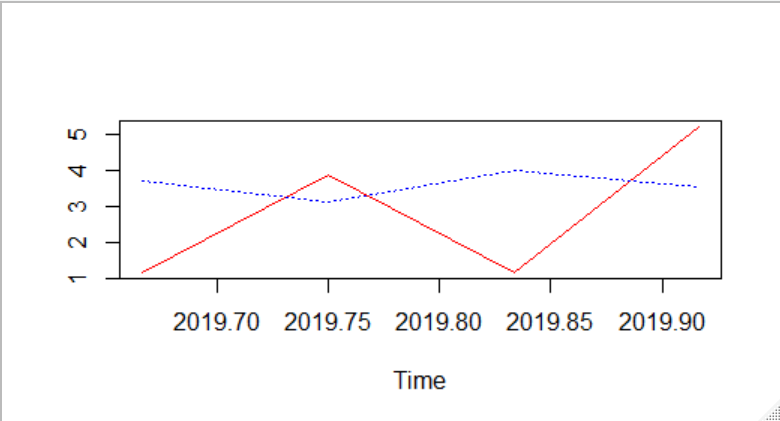
**Figure 44.** Yearly linear Regression Residual ACF Plot



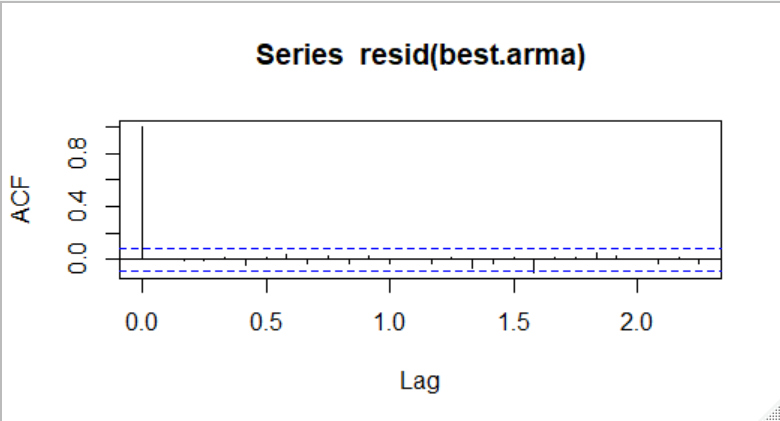
**Figure 45.** Weekly ARMA Prediction Plot



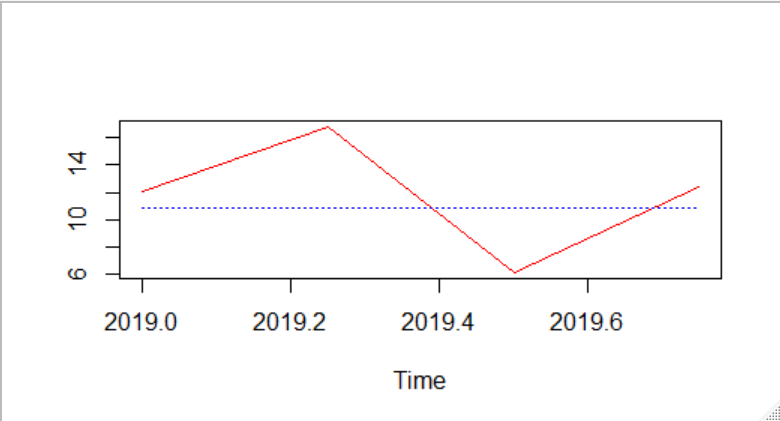
**Figure 46.** Weekly ARMA Residual ACF Plot



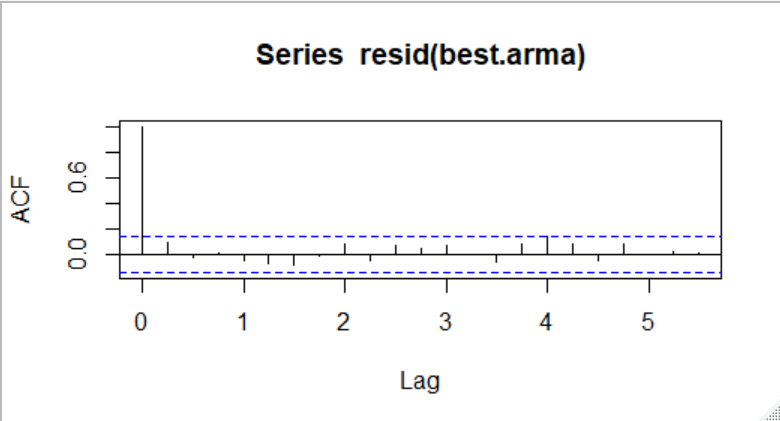
**Figure 47.** Monthly ARMA Prediction Plot



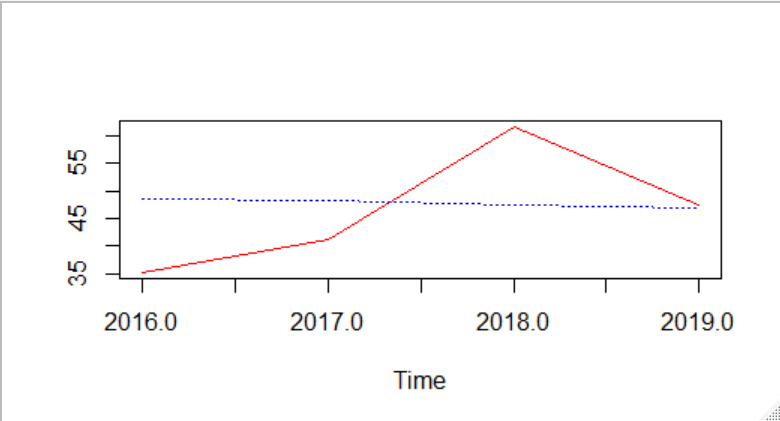
**Figure 48.** Monthly ARMA Residual ACF Plot



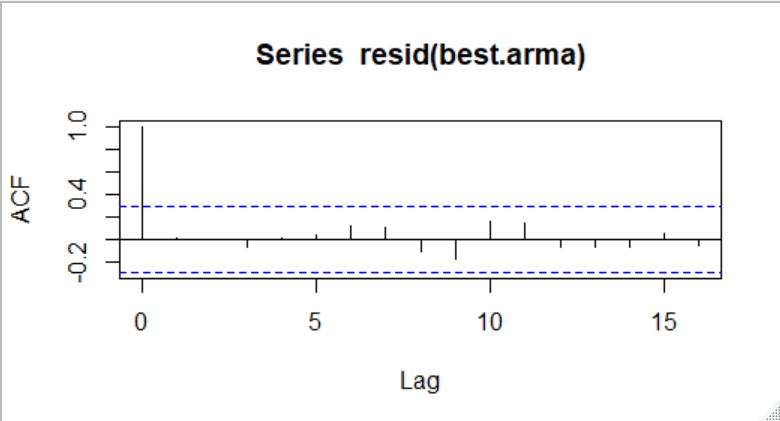
**Figure 49.** Seasonal ARMA Prediction Plot



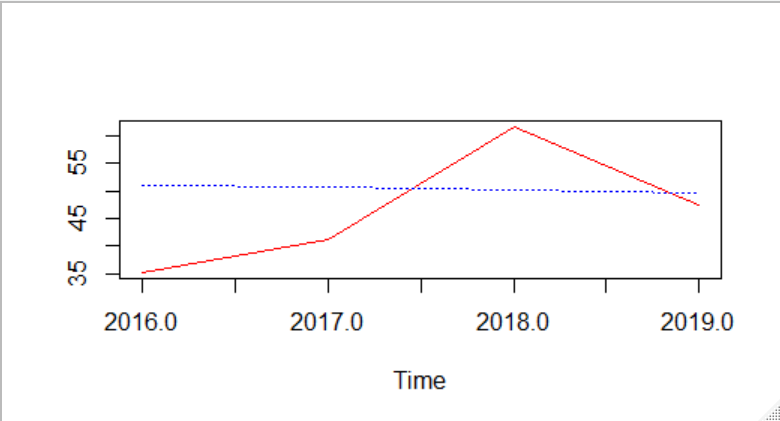
**Figure 50.** Seasonal ARMA Residual ACF Plot



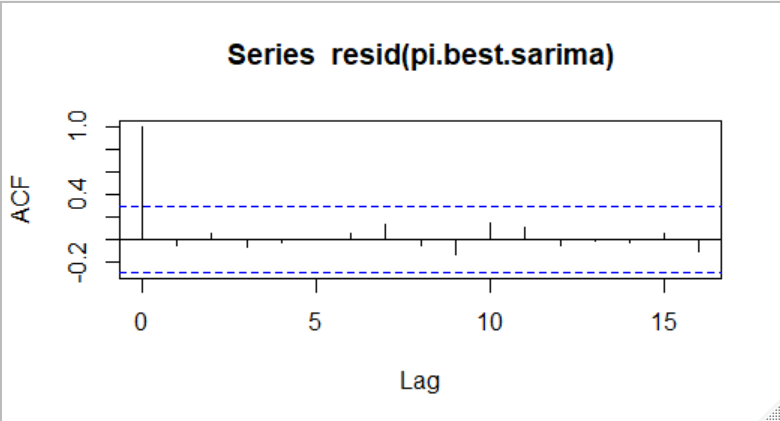
**Figure 51.** Yearly ARMA Prediction Plot



**Figure 52.** Yearly ARMA Residual ACF Plot



**Figure 53.** Yearly SARIMA Prediction Plot



**Figure 54.** Yearly SARIMA Residuals ACF Plot

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 1. MAPE Across Different Models | | | | | |
| MAPE | HW | AR | LR | ARMA | SARIMA |
| Weekly | 105.2883 | 79.7363 | 82.8428 | 82.937 | X |
| Monthly | 152.414 | 113.6695 | 122.766 | 129.56% | X |
| Seasonal | 31.84925 | 33.31567 | 30.1668 | 33.3157 | X |
| Yearly | 17.598 | 16.38687 | 16.5433 | 19.64% | 22.61% |

|  |  |  |
| --- | --- | --- |
| Table 2. Holt Winters Values | | |
| HW | Alpha | Coefficient |
| Weekly | 0.013 | 0.974 |
| Monthly | 0.0655 | 4.522 |
| Seasonal | 0.0409 | 11.804 |
| Yearly | 0.1197 | 46.421 |

|  |  |  |
| --- | --- | --- |
| **Table 3.** Weekly AR(6) Coefficients | | |
| AR | 2.50% | 97.50% |
| 1 | -0.0165 | 0.062 |
| 2 | -0.0129 | 0.0657 |
| 3 | -0.0248 | 0.0536 |
| 4 | 0.0085 | 0.0869 |
| 5 | -0.0511 | 0.0275 |
| 6 | 0.0113 | 0.0898 |

|  |  |  |
| --- | --- | --- |
| **Table 4.** Monthly AR(1) Coefficient | | |
| AR | 2.5 | 97.5 |
| 1 | 0.022 | 0.185 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 5.** Weekly Linear Regression Summary | | | | |
| Variable | Estimate | Std. Error | t- value | p-value |
| Time | 0.001782 | 0.001297 | 1.373 | 0.17 |
| factor(Seas)1 | -2.831292 | 2.59107 | -1.093 | 0.275 |
| factor(Seas)2 | -2.829926 | 2.591095 | -1.092 | 0.275 |
| factor(Seas)3 | -2.74456 | 2.59112 | -1.059 | 0.29 |
| factor(Seas)4 | -2.775995 | 2.591144 | -1.071 | 0.284 |
| factor(Seas)5 | -2.641629 | 2.591169 | -1.019 | 0.308 |
| factor(Seas)6 | -2.826663 | 2.591194 | -1.091 | 0.275 |
| factor(Seas)7 | -2.733497 | 2.591219 | -1.055 | 0.292 |
| factor(Seas)8 | -2.892332 | 2.591244 | -1.116 | 0.264 |
| factor(Seas)9 | -2.842966 | 2.591269 | -1.097 | 0.273 |
| factor(Seas)10 | -2.8416 | 2.591294 | -1.097 | 0.273 |
| factor(Seas)11 | -2.924035 | 2.591319 | -1.128 | 0.259 |
| factor(Seas)12 | -2.810469 | 2.591344 | -1.085 | 0.278 |
| factor(Seas)13 | -2.709503 | 2.591369 | -1.046 | 0.296 |
| factor(Seas)14 | -2.814537 | 2.591394 | -1.086 | 0.278 |
| factor(Seas)15 | -2.670372 | 2.591419 | -1.03 | 0.303 |
| factor(Seas)16 | -2.715206 | 2.591443 | -1.048 | 0.295 |
| factor(Seas)17 | -2.69004 | 2.591468 | -1.038 | 0.299 |
| factor(Seas)18 | -2.629674 | 2.591493 | -1.015 | 0.31 |
| factor(Seas)19 | -2.702109 | 2.591518 | -1.043 | 0.297 |
| factor(Seas)20 | -2.743143 | 2.591543 | -1.058 | 0.29 |
| factor(Seas)21 | -2.763577 | 2.591568 | -1.066 | 0.286 |
| factor(Seas)22 | -2.761611 | 2.591593 | -1.066 | 0.287 |
| factor(Seas)23 | -2.681246 | 2.591618 | -1.035 | 0.301 |
| factor(Seas)24 | -2.60528 | 2.591643 | -1.005 | 0.315 |
| factor(Seas)25 | -2.831914 | 2.591668 | -1.093 | 0.275 |
| factor(Seas)26 | -2.696349 | 2.591693 | -1.04 | 0.298 |
| factor(Seas)27 | -2.671583 | 2.591717 | -1.031 | 0.303 |
| factor(Seas)28 | -2.800217 | 2.591742 | -1.08 | 0.28 |
| factor(Seas)29 | -2.616451 | 2.591767 | -1.01 | 0.313 |
| factor(Seas)30 | -2.591286 | 2.591792 | -1 | 0.318 |
| factor(Seas)31 | -2.54292 | 2.591817 | -0.981 | 0.327 |
| factor(Seas)32 | -2.728554 | 2.591842 | -1.053 | 0.293 |
| factor(Seas)33 | -2.563588 | 2.591867 | -0.989 | 0.323 |
| factor(Seas)34 | -2.511423 | 2.591892 | -0.969 | 0.333 |
| factor(Seas)35 | -2.604457 | 2.591917 | -1.005 | 0.315 |
| factor(Seas)36 | -2.728891 | 2.591942 | -1.053 | 0.293 |
| factor(Seas)37 | -2.701925 | 2.591967 | -1.042 | 0.297 |
| factor(Seas)38 | -2.53416 | 2.591992 | -0.978 | 0.328 |
| factor(Seas)39 | -2.853994 | 2.592016 | -1.101 | 0.271 |
| factor(Seas)40 | -2.742428 | 2.592041 | -1.058 | 0.29 |
| factor(Seas)41 | -2.613063 | 2.592066 | -1.008 | 0.314 |
| factor(Seas)42 | -2.461497 | 2.592091 | -0.95 | 0.342 |
| factor(Seas)43 | -2.534531 | 2.592116 | -0.978 | 0.328 |
| factor(Seas)44 | -2.990165 | 2.592141 | -1.154 | 0.249 |
| factor(Seas)45 | -2.585 | 2.592166 | -0.997 | 0.319 |
| factor(Seas)46 | -2.781434 | 2.592191 | -1.073 | 0.283 |
| factor(Seas)47 | -2.688468 | 2.592216 | -1.037 | 0.3 |
| factor(Seas)48 | -2.858902 | 2.592241 | -1.103 | 0.27 |
| factor(Seas)49 | -2.788727 | 2.591689 | -1.076 | 0.282 |
| factor(Seas)50 | -2.785088 | 2.591714 | -1.075 | 0.283 |
| factor(Seas)51 | -2.796755 | 2.591739 | -1.079 | 0.281 |
| factor(Seas)52 | -2.720667 | 2.591764 | -1.05 | 0.294 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 6.** Monthly Linear Regression Summary | | | | |
| Variable | Estimate | Std. Error | t- value | p-value |
| Time | 0.008379 | 0.005887 | 1.423 | 0.155 |
| factor(Seas)1 | -13.437668 | 11.744502 | -1.144 | 0.253 |
| factor(Seas)2 | -13.897966 | 11.744992 | -1.183 | 0.237 |
| factor(Seas)3 | -12.948864 | 11.745483 | -1.102 | 0.271 |
| factor(Seas)4 | -13.073162 | 11.745973 | -1.113 | 0.266 |
| factor(Seas)5 | -12.962461 | 11.746463 | -1.104 | 0.27 |
| factor(Seas)6 | -12.793359 | 11.746954 | -1.089 | 0.277 |
| factor(Seas)7 | -12.394057 | 11.747444 | -1.055 | 0.292 |
| factor(Seas)8 | -12.642555 | 11.747935 | -1.076 | 0.282 |
| factor(Seas)9 | -12.720211 | 11.745556 | -1.083 | 0.279 |
| factor(Seas)10 | -13.515195 | 11.746047 | -1.151 | 0.25 |
| factor(Seas)11 | -13.551812 | 11.746537 | -1.154 | 0.249 |
| factor(Seas)12 | -13.097408 | 11.747028 | -1.115 | 0.265 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 7.** Seasonal Linear Regression Summary | | | | |
| Variable | Estimate | Std. Error | t- value | p-value |
| Time | 0.02206 | 0.01828 | 1.207 | 0.229 |
| factor(Seas)1 | -32.84683 | 36.44791 | -0.901 | 0.369 |
| factor(Seas)2 | -31.76296 | 36.45248 | -0.871 | 0.385 |
| factor(Seas)3 | -33.64154 | 36.45705 | -0.923 | 0.357 |
| factor(Seas)4 | -34.31746 | 36.46162 | -0.941 | 0.348 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 8.** Yearly Linear Regression Summary | | | | |
| Variable | Estimate | Std. Error | t- value | p-value |
| Time | 0.021697 | 0.000575 | 37.74 | <2e-16 |

**R Code**

setwd("C:\\Users\\cboyk\\OneDrive\\Desktop\\HU Analytics\\565 Time Series\\Research Project")

# Libraries

library(readxl)

library(tidyverse)

# Reading in data

philly <- read\_excel("Philadelphia Precipitation 19702020.xlsx")

# Making a flag for week

week <- rep(1:2608,each = 7, 1)

week <- as.data.frame(week)

week2 <- rep(2609, each = 6, 1)

week2 <- as.data.frame(week2)

colnames(week2) <- "week"

week3 <- rbind(week, week2)

philly$week <- week3$week

# Making a Flag for Month

philly$month.year <- paste(philly$MONTH, philly$YEAR, sep = " ")

# Making a Flag for Season

philly$season.year <- paste(philly$SEASON, philly$YEAR, sep = " ")

# Making Week subset

philly\_week <- philly %>% group\_by(week) %>%

summarize(PRCP = sum(PRCP))

# Making Month subset

philly\_month <- philly %>% group\_by(month.year) %>%

summarize(PRCP = sum(PRCP))

# Making Seasonal Subset

philly\_season <- philly %>% group\_by(season.year) %>%

summarize(PRCP = sum(PRCP))

# Making Yearly subset

philly\_year <- philly %>% group\_by(YEAR) %>%

summarize(PRCP = sum(PRCP))

# Saving month and season series since they are sorted incorrectly

library(openxlsx)

write.xlsx(philly\_month, "Philly Month Series.xlsx")

write.xlsx(philly\_season, "Philly Season Series.xlsx")

# Bringing Month and Season series back in

philly\_month <- read\_excel("Philly Month Series.xlsx")

philly\_season <- read\_excel("Philly Season Series.xlsx")

# Making TS

philly.daily.ts <- ts(philly$PRCP, start = c(1970), end = c(2019+364/365), freq = 365)

philly.weekly.ts <- ts(philly\_week$PRCP, start = 1970, end = 2019 + 51/52, freq = 52)

philly.monthly.ts <- ts(philly\_month$PRCP, start = c(1970, 1), end = c(2019, 12), freq = 12)

philly.season.ts <- ts(philly\_season$PRCP, start = 1970, end = 2019.75,freq = 4)

philly.year.ts <- ts(philly\_year$PRCP, start = c(1970), end = c(2019), freq = 1)

# Making pre and post sets for each

philly.daily.ts.pre <- window(philly.daily.ts, end = 2019+354/365)

philly.daily.ts.post <- window(philly.daily.ts, start = 2019+355/365)

philly.weekly.ts.pre <- window(philly.weekly.ts, end = 2019+47/52)

philly.weekly.ts.post <- window(philly.weekly.ts, start = 2019+48/52)

philly.monthly.ts.pre <- window(philly.monthly.ts, end = c(2019, 8))

philly.monthly.ts.post <- window(philly.monthly.ts, start = c(2019, 9))

philly.season.ts.pre <- window(philly.season.ts, end = 2018.75)

philly.season.ts.post <- window(philly.season.ts, start = 2019)

philly.year.ts.pre <- window(philly.year.ts, end = 2015)

philly.year.ts.post <- window(philly.year.ts, start = 2016)

# Time series plots

ts.plot(philly.daily.ts)

ts.plot(philly.weekly.ts)

ts.plot(philly.monthly.ts)

ts.plot(philly.season.ts)

ts.plot(philly.year.ts)

# ACF plots

acf(philly.daily.ts)

acf(philly.weekly.ts)

acf(philly.monthly.ts)

acf(philly.season.ts)

acf(philly.year.ts)

# PACF plots

pacf(philly.daily.ts)

pacf(philly.weekly.ts)

pacf(philly.monthly.ts)

pacf(philly.season.ts)

pacf(philly.year.ts)

# Decomposition of each series

plot(decompose(philly.daily.ts))

plot(decompose(philly.weekly.ts))

plot(decompose(philly.monthly.ts))

plot(decompose(philly.season.ts))

# Holt Winters --------------------------------------------------------------------------------------

philly.daily.hw <- HoltWinters(philly.daily.ts.pre, beta =F, gamma = F)

philly.daily.hw$SSE

philly.daily.hw.predict <- predict(philly.daily.hw, n.ahead=10)

ts.plot(philly.daily.ts.post, philly.daily.hw.predict, lty = c(1,3), col=c("red","blue"))

# Cannot do MAPE because of values being 0 so divide by 0 error

philly.weekly.hw <- HoltWinters(philly.weekly.ts.pre, beta =F, gamma = F)

philly.weekly.hw$SSE

philly.weekly.hw.predict <- predict(philly.weekly.hw, n.ahead=4)

ts.plot(philly.weekly.ts.post, philly.weekly.hw.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.weekly.ts.post - philly.weekly.hw.predict))/(philly.weekly.ts.post))/

length(philly.weekly.ts.post)\*100 # 105.2883

acf(resid(philly.weekly.hw))

philly.monthly.hw <- HoltWinters(philly.monthly.ts.pre, beta =F, gamma = F)

philly.monthly.hw$SSE

philly.monthly.hw.predict <- predict(philly.monthly.hw, n.ahead=4)

ts.plot(philly.monthly.ts.post, philly.monthly.hw.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.monthly.ts.post - philly.monthly.hw.predict))/(philly.monthly.ts.post))/

length(philly.monthly.ts.post)\*100 # 152.414

acf(resid(philly.monthly.hw))

philly.season.hw <- HoltWinters(philly.season.ts.pre, beta =F, gamma = F)

philly.season.hw$SSE

philly.season.hw.predict <- predict(philly.season.hw, n.ahead=4)

ts.plot(philly.season.ts.post, philly.season.hw.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.season.ts.post - philly.season.hw.predict))/(philly.season.ts.post))/

length(philly.season.ts.post)\*100 # 31.84925

acf(resid(philly.season.hw))

philly.year.hw <- HoltWinters(philly.year.ts.pre, beta =F, gamma = F)

philly.year.hw$SSE

philly.year.hw.predict <- predict(philly.year.hw, n.ahead=4)

ts.plot(philly.year.ts.post, philly.year.hw.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.year.ts.post - philly.year.hw.predict))/(philly.year.ts.post))/

length(philly.year.ts.post)\*100 # 17.59806

acf(resid(philly.year.hw))

# AR(p) ------------------------------------------------------------------------------------------

philly.weekly.ar <- ar(philly.weekly.ts.pre)

philly.weekly.ar$order

philly.weekly.ar.predict <- predict(philly.weekly.ar, n.ahead = 4)

ts.plot(philly.weekly.ts.post, philly.weekly.ar.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.weekly.ts.post - philly.weekly.ar.predict$pred))/(philly.weekly.ts.post))/

length(philly.weekly.ts.post)\*100 # 79.7363

philly.weekly.ar$ar[1] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[1,1]))

philly.weekly.ar$ar[2] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[2,2]))

philly.weekly.ar$ar[3] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[3,3]))

philly.weekly.ar$ar[4] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[4,4]))

philly.weekly.ar$ar[5] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[5,5]))

philly.weekly.ar$ar[6] + c(-2, 2) \* sqrt(as.numeric(philly.weekly.ar$asy.var.coef[6,6]))

acf(philly.weekly.ar$resid[-c(1:6)])

philly.monthly.ar <- ar(philly.monthly.ts.pre)

philly.monthly.ar$order

philly.monthly.ar.predict <- predict(philly.monthly.ar, n.ahead = 4)

ts.plot(philly.monthly.ts.post, philly.monthly.ar.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.monthly.ts.post - philly.monthly.ar.predict$pred))/(philly.monthly.ts.post))/

length(philly.monthly.ts.post)\*100 # 113.6695

philly.monthly.ar$ar[1] + c(-2, 2) \* sqrt(as.numeric(philly.monthly.ar$asy.var.coef[1,1]))

acf(na.omit(philly.monthly.ar$resid))

philly.season.ar <- ar(philly.season.ts.pre)

philly.season.ar$order

philly.season.ar.predict <- predict(philly.season.ar, n.ahead = 4)

ts.plot(philly.season.ts.post, philly.season.ar.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.season.ts.post - philly.season.ar.predict$pred))/(philly.season.ts.post))/

length(philly.season.ts.post)\*100 # 33.31567

acf(na.omit(philly.season.ar$resid))

philly.year.ar <- ar(philly.year.ts.pre)

philly.year.ar$order

philly.year.ar.predict <- predict(philly.year.ar, n.ahead = 4)

ts.plot(philly.year.ts.post, philly.year.ar.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.year.ts.post - philly.year.ar.predict$pred))/(philly.year.ts.post))/

length(philly.year.ts.post)\*100 # 16.38687

acf(na.omit(philly.year.ar$resid))

# Linear Regression ---------------------------------------------------------------------------------

Seas <- cycle(philly.weekly.ts.pre) # saving the seasons

Time <- time(philly.weekly.ts.pre) # saving the time

philly.weekly.lm <- lm(philly.weekly.ts.pre ~ 0 + Time + factor(Seas))

summary(philly.weekly.lm)

new.t <- seq(2019+48/52., len = 4, by = 1/52)

alpha <- coef(philly.weekly.lm)[1]

beta <- rep(coef(philly.weekly.lm)[2:52], 1)

new.dat <- data.frame(Time = new.t, Seas = rep(1:4, 1))

philly.weekly.lm.predict <- predict(philly.weekly.lm, new.dat)[1:4]

ts.plot(philly.weekly.ts.post, philly.weekly.lm.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.weekly.ts.post - philly.weekly.lm.predict))/(philly.weekly.ts.post))/

length(philly.weekly.ts.post)\*100 # 82.84284

acf(philly.weekly.lm$resid)

Seas <- cycle(philly.monthly.ts.pre) # saving the seasons

Time <- time(philly.monthly.ts.pre) # saving the time

philly.monthly.lm <- lm(philly.monthly.ts.pre ~ 0 + Time + factor(Seas))

summary(philly.monthly.lm)

new.t <- seq(2019+9/12, len = 4, by = 1/12)

alpha <- coef(philly.monthly.lm)[1]

beta <- rep(coef(philly.monthly.lm)[2:12], 1)

new.dat <- data.frame(Time = new.t, Seas = rep(1:4, 1))

philly.monthly.lm.predict <- predict(philly.monthly.lm, new.dat)[1:4]

ts.plot(philly.monthly.ts.post, philly.monthly.lm.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.monthly.ts.post - philly.monthly.lm.predict))/(philly.monthly.ts.post))/

length(philly.monthly.ts.post)\*100 # 96. 30368

acf(philly.monthly.lm$resid)

Seas <- cycle(philly.season.ts.pre) # saving the seasons

Time <- time(philly.season.ts.pre) # saving the time

philly.season.lm <- lm(philly.season.ts.pre ~ 0 + Time + factor(Seas))

summary(philly.season.lm)

new.t <- seq(2019, len = 4, by = 1/4)

alpha <- coef(philly.season.lm)[1]

beta <- rep(coef(philly.season.lm)[2:52], 1)

new.dat <- data.frame(Time = new.t, Seas = rep(1:4, 1))

philly.season.lm.predict <- predict(philly.season.lm, new.dat)[1:4]

ts.plot(philly.season.ts.post, philly.season.lm.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.season.ts.post - philly.season.lm.predict))/(philly.season.ts.post))/

length(philly.season.ts.post)\*100 # 30.16681

acf(philly.season.lm$resid)

Time <- time(philly.year.ts.pre) # saving the time

philly.year.lm <- lm(philly.year.ts.pre ~ 0 + Time)

summary(philly.year.lm)

new.t <- seq(2016, len = 4, by = 1)

alpha <- coef(philly.year.lm)[1]

new.dat <- data.frame(Time = new.t, Seas = rep(1:4, 1))

philly.year.lm.predict <- predict(philly.year.lm, new.dat)[1:4]

ts.plot(philly.year.ts.post, philly.year.lm.predict, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.year.ts.post - philly.year.lm.predict))/(philly.year.ts.post))/

length(philly.year.ts.post)\*100 # 16.54327

acf(philly.year.lm$resid)

# Harmonic -----------------------------------------------------------------------------------------

SIN <- COS <- matrix(nr = length(philly.monthly.ts.pre), nc = 6)

for (i in 1:6) {

SIN[, i] <- sin(2 \* pi \* i \* time(philly.monthly.ts.pre))

COS[, i] <- cos(2 \* pi \* i \* time(philly.monthly.ts.pre))

}

TIME <- (time(philly.monthly.ts.pre) - mean(time(philly.monthly.ts.pre)))/sd(time(philly.monthly.ts.pre))

mean(time(philly.monthly.ts.pre))

sd(time(philly.monthly.ts.pre))

philly.monthly.ts.pre.lm1 <- lm(log(philly.monthly.ts.pre) ~ TIME + I(TIME^2) + I(TIME^3) + I(TIME^4) + SIN[,1] + COS[,1] + SIN[,2] + COS[,2] +

SIN[,3] + COS[,3] + SIN[,4] + COS[,4] + SIN[,5] + COS[,5] + SIN[,6] + COS[,6])

coef(philly.monthly.ts.pre.lm1)/sqrt(diag(vcov(philly.monthly.ts.pre.lm1)))

# ARMA ---------------------------------------------------------------------------------------------

best.aic <- Inf

for (i in 0:6) for (j in 0:6) {

fit.aic <- AIC(arima(philly.weekly.ts.pre, order = c(i, 0,j) ) )

if (fit.aic < best.aic) {

best.order <- c(i, 0, j)

best.arma <- arima(philly.weekly.ts.pre, order = best.order)

best.aic <- fit.aic

}}

best.order

acf(resid(best.arma))

philly.weekly.arma.predict <- predict(best.arma, n.ahead = 4)

ts.plot(philly.weekly.ts.post, philly.weekly.arma.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.weekly.ts.post - philly.weekly.arma.predict$pred))/(philly.weekly.ts.post))/

length(philly.weekly.ts.post)\*100 # 82.936

best.aic <- Inf

for (i in 0:6) for (j in 0:6) {

fit.aic <- AIC(arima(philly.season.ts.pre, order = c(i, 0,j) ) )

if (fit.aic < best.aic) {

best.order <- c(i, 0, j)

best.arma <- arima(philly.season.ts.pre, order = best.order)

best.aic <- fit.aic

}}

best.order

acf(resid(best.arma))

philly.season.arma.predict <- predict(best.arma, n.ahead = 4)

ts.plot(philly.season.ts.post, philly.season.arma.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.season.ts.post - philly.season.arma.predict$pred))/(philly.season.ts.post))/

length(philly.season.ts.post)\*100 # 36.81892

best.aic <- Inf

for (i in 0:6) for (j in 0:6) {

fit.aic <- AIC(arima(philly.monthly.ts.pre, order = c(i, 0,j) ) )

if (fit.aic < best.aic) {

best.order <- c(i, 0, j)

best.arma <- arima(philly.monthly.ts.pre, order = best.order)

best.aic <- fit.aic

}}

best.order

acf(resid(best.arma))

philly.monthly.arma.predict <- predict(best.arma, n.ahead = 4)

ts.plot(philly.monthly.ts.post, philly.monthly.arma.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.monthly.ts.post - philly.monthly.arma.predict$pred))/(philly.monthly.ts.post))/

length(philly.monthly.ts.post)\*100 # 129.563

best.aic <- Inf

for (i in 0:6) for (j in 0:6) {

fit.aic <- AIC(arima(philly.year.ts.pre, order = c(i, 0,j) ) )

if (fit.aic < best.aic) {

best.order <- c(i, 0, j)

best.arma <- arima(philly.year.ts.pre, order = best.order)

best.aic <- fit.aic

}}

best.order

acf(resid(best.arma))

philly.year.arma.predict <- predict(best.arma, n.ahead = 4)

ts.plot(philly.year.ts.post, philly.year.arma.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.year.ts.post - philly.year.arma.predict$pred))/(philly.year.ts.post))/

length(philly.year.ts.post)\*100 # 19.63586

# ADF Test -----------------------------------------------------------------------------------------

library(tseries)

adf.test(philly.daily.ts)

adf.test(philly.weekly.ts)

adf.test(philly.monthly.ts)

adf.test(philly.season.ts)

adf.test(philly.year.ts)

adf.test(decompose(philly.year.ts)$random)

# SARIMA

get.best.sarima <- function(x.ts, maxord = c(1,1,1,1,1,1))

{

best.aic <- 1e8

n <- length(x.ts)

for (p in 0:maxord[1]) for(d in 0:maxord[2]) for(q in 0:maxord[3])

for (P in 0:maxord[4]) for(D in 0:maxord[5]) for(Q in 0:maxord[6])

{

fit <- arima(x.ts, order = c(p,d,q),

seas = list(order = c(P,D,Q),

frequency(x.ts)), method = "CSS", optim.control = list(maxit = 10000) )

fit.aic <- -2 \* fit$loglik + (log(n) + 1) \* length(fit$coef)

if (fit.aic < best.aic)

{best.aic <- fit.aic

best.fit <- fit

best.model <- c(p,d,q,P,D,Q)

}

}

list(best.aic, best.fit, best.model)

}

get.best.sarima(philly.year.ts.pre, maxord = c(2, 2, 2, 2, 2, 2))

pi.best.sarima <- arima(philly.year.ts.pre, order = c(0,0,2), seas=list(order=c(2,1,1),1))

philly.year.sarima.predict <- predict(pi.best.sarima, n.ahead = 4)

ts.plot(philly.year.ts.post, philly.year.sarima.predict$pred, lty = c(1,3), col=c("red","blue"))

sum((abs(philly.year.ts.post - philly.year.sarima.predict$pred))/(philly.year.ts.post))/

length(philly.year.ts.post)\*100 # 22.60865

acf(resid(pi.best.sarima))