Stefan Maciolek

5/7/2020

16:960:565

Time Series Analysis of Cardiovascular and Neoplasm Deaths in the US

Introduction:

While catastrophic events such as wars, terror attacks, pandemics, and natural disasters receive the lion's share of media attention, the simple fact remains that the most Americans will die from one of two causes: heart disease and cancer. In 2018, of the recorded 723.6 deaths per 100,000 thousand people in the United States, 163.6 of them were caused by heart disease, and 149.1 of them by cancer (Xu et al). This means heart disease and cancer caused 22.6% and 20.6%, of all deaths in the US that year, respectively. While the speculation on the factors influencing the rates of death caused by these two scourges is beyond the scope of this paper, constructing time series models and making predictions nonetheless has much utility in the context of public health. Resources and personnel can be allocated to meet the expected numbers of those affected, and identification of times when the number of deaths is relatively low can inform future studies to identify factors causing the reduced number.

Methods:

Data were obtained from the CDC's WONDER database on the total numbers of deaths in the United States, in the years 1999 to 2018. The data were filtered to include deaths from cardiovascular system issues - representing heart disease, and deaths from neoplasms - representing cancer. The data were collected by year and by month for each. The total population of the United States was available for the yearly but not the monthly data, so it was added to the monthly data, but reflects the population of the US with a yearly, not a monthly resolution.

In order to remove the effects of the US's rapid population growth, the *rates* of death were analyzed, as opposed to the total number of deaths. Rates were expressed as percentages of the total population dying from each cause per unit time. Because the population data were only collected by year, the monthly percentages could be considered to be only approximations of the total death rates, since intra-year births and deaths were not factored in. However, because the size of the United States' population is very large compared to the annual numbers of births and deaths, this approximation is considered to be reasonably accurate for our purposes.

In order to control for general upwards and downwards trends in the data over time, first differences and seasonal differences were taken to see if they stabilized the data. Because both time series appear to have equal variance over time, log and power transformations were not considered. Then auto-correlation functions (ACFs) and partial autocorrelation functions (PACFs) were taken of each time series to inform which starting model was used, and a periodogram was used to confirm periodicity. Starting models included regular differencing and/or seasonal differencing if shown to be warranted from the initial analysis. The model started as a moving average of degree 1, and the degree was incremented by one until successive parameters were no longer significant. Then the seasonal moving average component was added, incrementing the degree by one each time, until there were no

further significant parameters. Finally, the same was done with autoregressive components. The final model in each case was evaluated with a residual plot, a residual autocorrelation plot, as well as QQ plots and histograms of the residuals. A Shapiro-Wilk test was not considered due to the large sample size of the data. Due to the complexity of the data, it was possible to produce very complicated models that had significant moving average and autoregressive components at high lags. However, in order to reduce the possibility of overfitting, favor was given to simpler models. This resulted in a compromise whereby models would be complex enough to fully model the data, but models with very high MA and AR orders were not considered.

Finally, using the fitted models, predictions were made for the percentage of deaths occurring from both causes for an additional five years. Note that data were only available through year end 2018, and so the data for 2019 and 2020 are predictions. The predictions do not take into account the possibility of medical advancements that would reduce the death rate from both causes, and so if such advancements were made in the prediction window, the rates would no longer be accurate. The discussion of such advancements is beyond the scope of this paper.

Results and Discussion:

Preliminary:

Figure 1 in Appendix 1 seemingly shows a worrying trend – total deaths in the United States due to cardiovascular issues dropped, bottomed out, and started increasing again, while neoplasm deaths steady increased over the years. However, if we normalize the deaths by the total population and instead look at the rate of the total population dying per year, the data tell a different story in Figure 2. The uptick in cardiovascular issues is still present, though much less pronounced, and the proportion of deaths due to neoplasm issues relatively constant. If we instead look at the proportion of the population dying per month, a new trend emerges, - one of seasonality. Figure 3 suggests that there is a winter peak in the proportion of people dying from cardiovascular issues, while there is a much lesser dip in the deaths caused by neoplasm issues in the spring. However, the latter effect seems to be most pronounced in the month of February for each year – perhaps the shorter duration of this month means fewer total deaths accumulate, and indeed, the dip seems less pronounced every four years, when February is slightly longer due to leap years. Nonetheless, there does seem to be some seasonality at play in the rates of neoplasm deaths, but with much less overall variation than cardiovascular deaths.

Cardiovascular Deaths:

The rate of cardiovascular deaths, as shown in Figure 4 in Appendix 2, seems to be both trending downward and highly seasonal. In order to account for this, the first regular difference was taken (Figure 5), as well as the first season difference (Figure 6), which resulted in a much more random and stochastic time series. This informed an ARIMA (0,1,1) X $(0,1,1)_{12}$ starting model. However, the autocorrelation function of the double-differenced series (Figure 7) shows significant values at the first three lags, and the significant value at lag 36 suggests more seasonal moving average terms were needed, as well. The partial autocorrelation Function (Figure 8) decayed much less rapidly, suggesting that the moving average component of this series dominates, as opposed the autoregressive component. A periodogram (Figure 9) of the double differenced series shows one dominant frequency, further lending support to the notion that these deaths are highly seasonal, with a yearly period.

One order was added to each component until the highest order coefficients were no longer significant, and then the process was repeated with the other components. Ultimately, this iterative method yielded several significant models, with the simplest being chosen. The final model was an ARIMA (1,1,3) X $(0,1,3)_{12}$. A residual plot, shown in Figure 10 of Appendix 3 appears to be mostly random, and Figure 11 shows that the residuals are not autocorrelated. Figure 12, which is a QQ plot of the residuals, and Figure 13, which is a histogram are not perfectly normal – they show a lighter-tailed distribution of residuals towards the mean than would be predicted under a perfectly normal approach. However, because the assumption of residual independence is not violated and there seems to be no skewness, this model is taken to be sufficient.

Neoplasm Deaths:

Figure 14 in Appendix 4 shows that neoplasm deaths are seasonal and do have a trend, but neither is as pronounced as with the cardiovascular deaths. In this case, just a seasonal difference was taken and shown to work well (Figure 15), which informed an ARIMA (0,0,1) X $(0,1,1)_{12}$ starting model. The autocorrelation function in Figure 16 this time has significant values only at lags 1 and 12, which suggests a simpler model will be enough in this case, and the partial autocorrelation function, shown in Figure 17, does not lend any credence to needing a more complex model. A periodogram of the seasonally differenced series, shown in Figure 18, does not show the presence of a single dominant frequency.

Via the method described in the previous section, and ARIMA (1,0,1) X $(0,1,1)_{12}$ was found to be sufficient, with diagnostics shown in Appendix 5. The residual plot, in Figure 19 appears to be random, and an autocorrelation plot shown in Figure 20 shows the residuals are mostly independent. While the autocorrelation of the first lag of the series is significant, it is noted that much more complex models were fit which still did not resolve this, with no improvement to other diagnostic measures, and so this model was ultimately chosen. A QQ plot (Figure 21) and a histogram (Figure 22) show that the residuals appear to be very normally distributed.

Predictions:

Predictions for the proportion of the population dying of both causes were carried out for 60 months, and are plotted in Appendix 6, Figure 23, along with 95% confidence intervals for each. It is clear that cardiovascular deaths will continue to be seasonal, with the rate being higher in the winter months. There also appears to be a slight upward trend, reversing the early 2000's downward trend, though the widening confidence intervals make this difficult to state conclusively. Neoplasm deaths, on the other hand, are predicted to be relatively constant. They will be seasonal, but neatly contained within a narrow band.

Conclusion:

The validity of the predictions presented here is contingent on no major medical advancements becoming available in the next five years for these issues. There is predictive evidence of a slight increase of the proportion of deaths due to cardiovascular disease in the coming years, and from a public health perspective, that should be addressed due to the sheer magnitude of the number of people it will affect. Given that these deaths are also very seasonal, various factors that lower the rate of death in the summer could be leveraged to likewise lower them in the winter (speculation on these

factors is, as previously stated, beyond the scope of this paper). The neoplasm death rate, on the other hand is not projected to increase in the coming years. While every death is of course a tragedy, its lower overall rate, and the fact it is not going to increase, make neoplasm issues somewhat less urgent of an issue that cardiovascular disease in the context of public health.

Refences:

Xu JQ, Murphy SL, Kochanek KD, Arias E. Mortality in the United States, 2018. NCHS Data Brief, no 355. Hyattsville, MD: National Center for Health Statistics. 2020.

Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2018 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2018, as "compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html on Apr 30, 2020 7:37:30 AM

Cryer, J. D. and Chan, K.-S. (2008). Time Series Analysis with Applications in R (2ndEd). Springer.

Appendix:

Appendix 1: Preliminary Analysis:

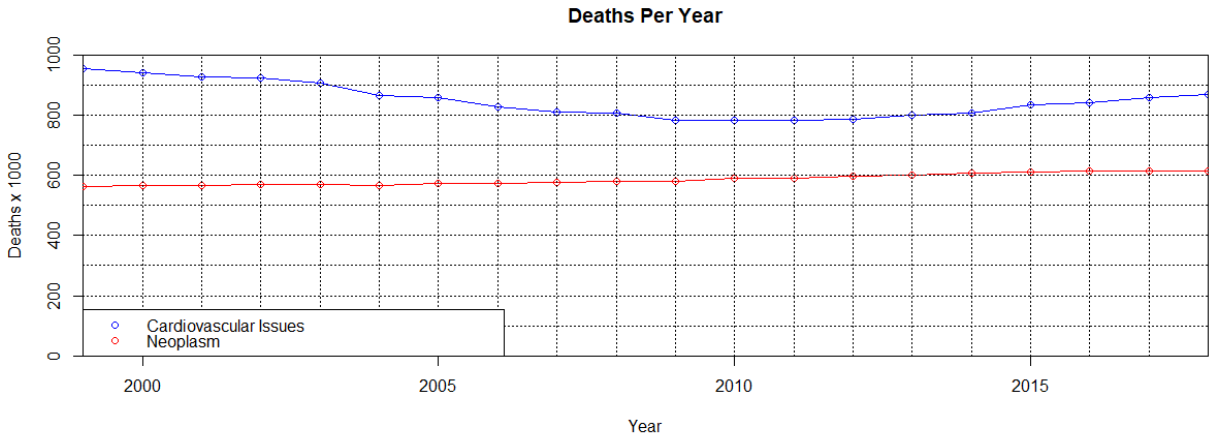

Figure 1: Total Number of Deaths per Year in the United States from Cardiovascular and Neoplasm issues

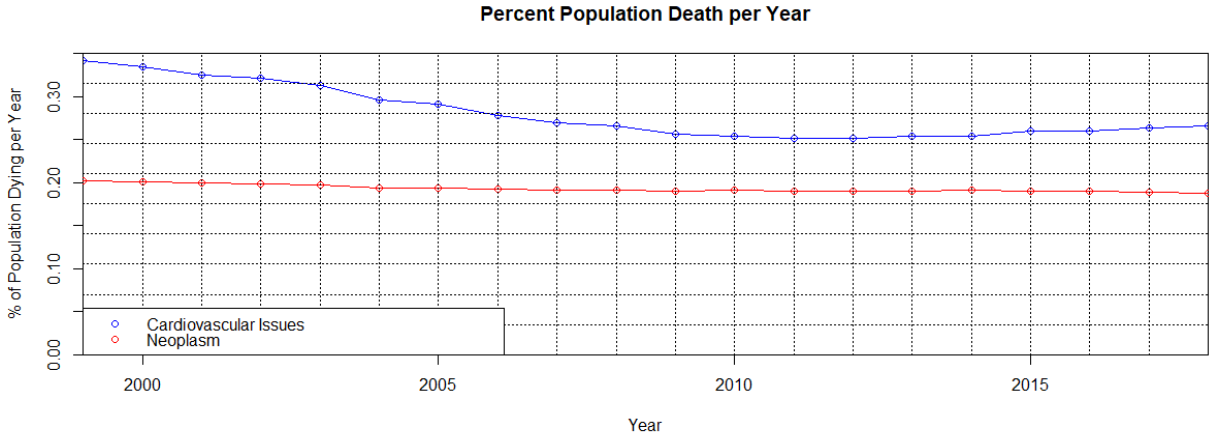

Figure 2: Percentage of the Population Dying per Year from Cardiovascular and Neoplasm Issues

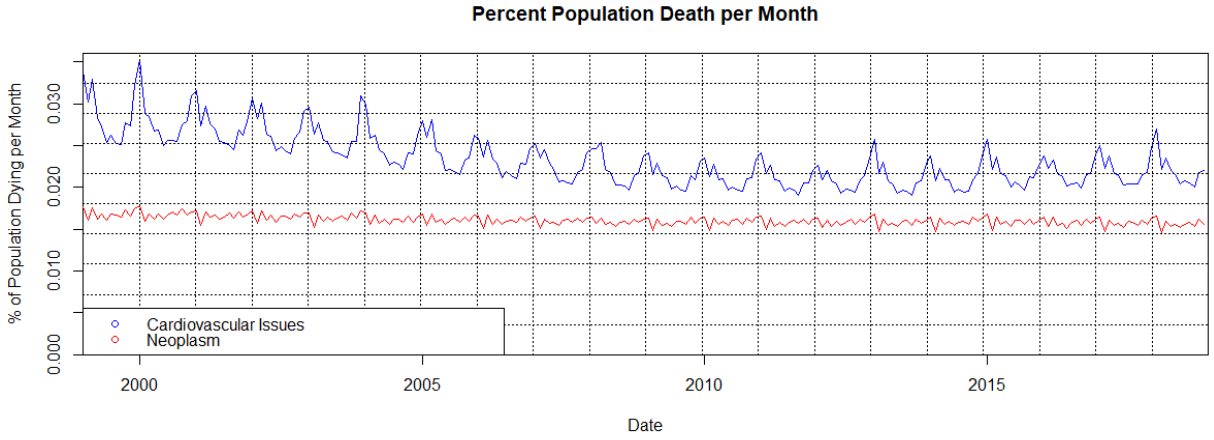

Figure 3: Percentage of the Population Dying per Month from Cardiovascular and Neoplasm Issues

Appendix 2: Preprocessing - Cardiovascular Deaths:

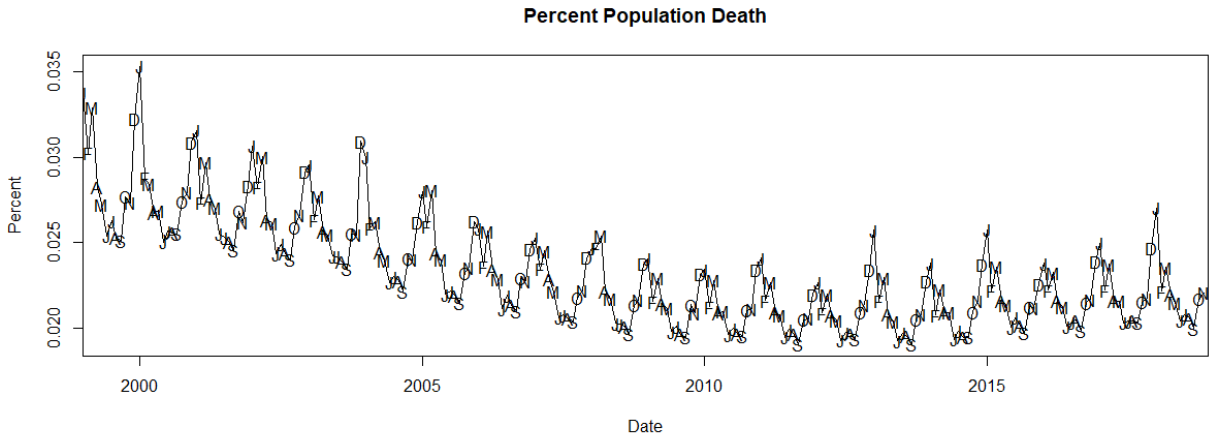

Figure 4: Percent of Monthly Population Deaths due to Cardiovascular Issues

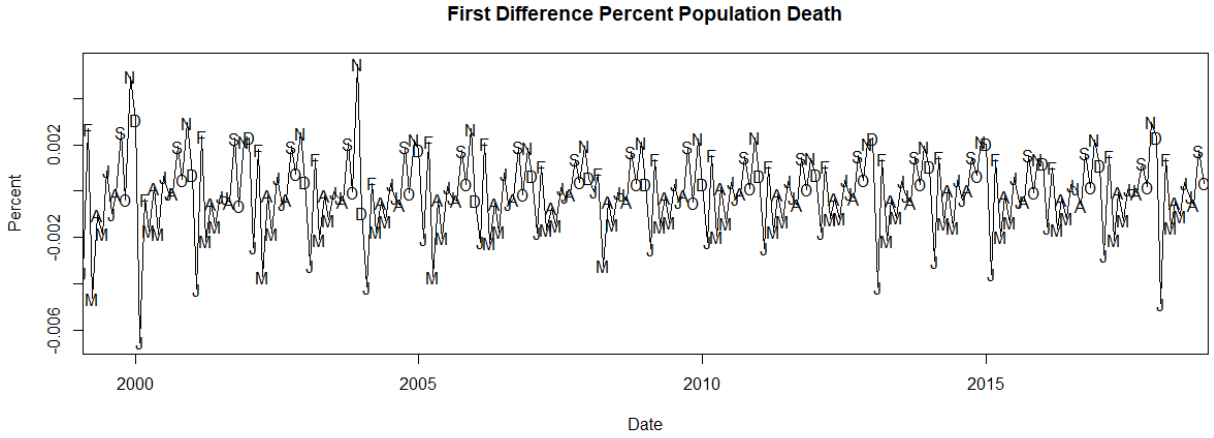

Figure 5: First Difference of Monthly Population Deaths due to Cardiovascular Issues

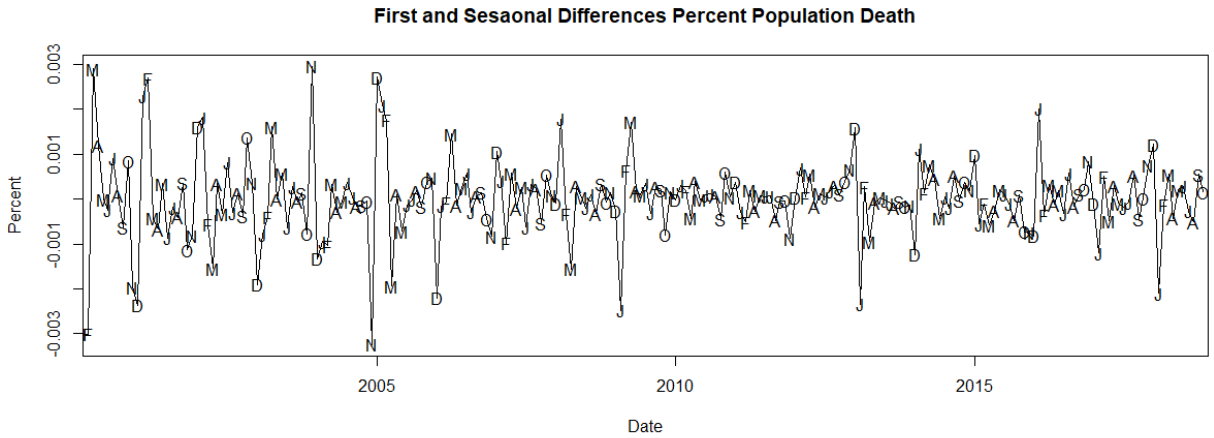

Figure 6: First Difference and First Seasonal Difference of Deaths due to Cardiovascular Issues

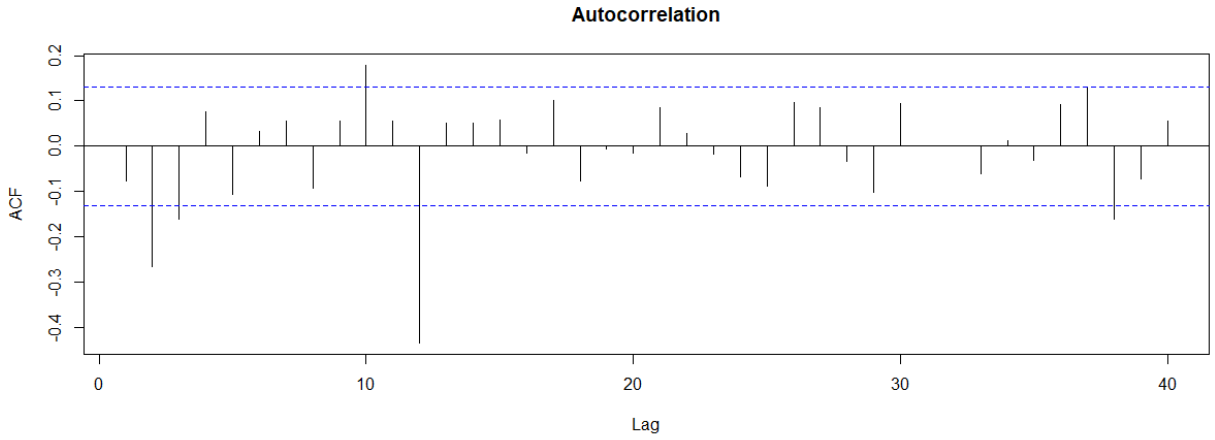

Figure 7: Autocorrelation Function of First and First Seasonal Differenced Deaths due to Cardiovascular Issues

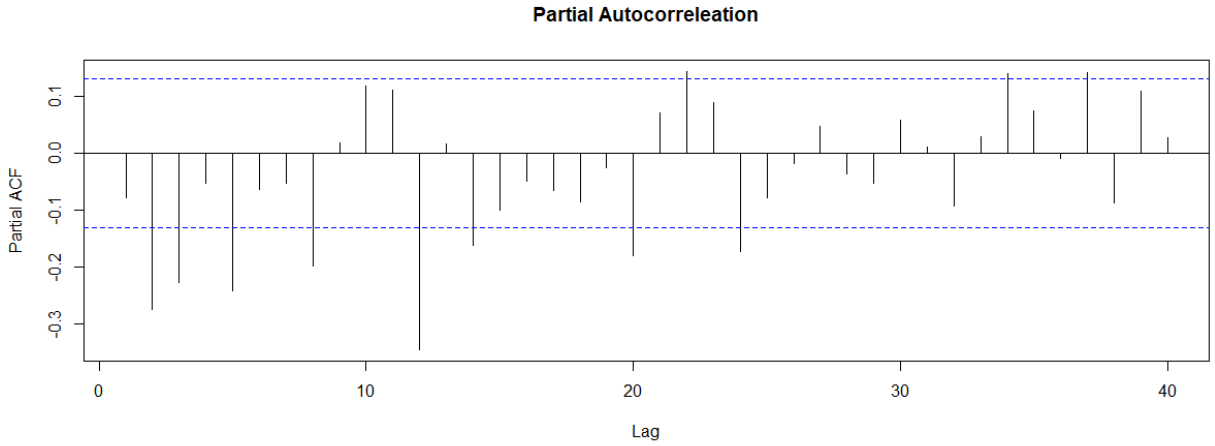

Figure 8: Partial Autocorrelation Function of First and First Seasonal Differenced Deaths due to Cardiovascular Issues

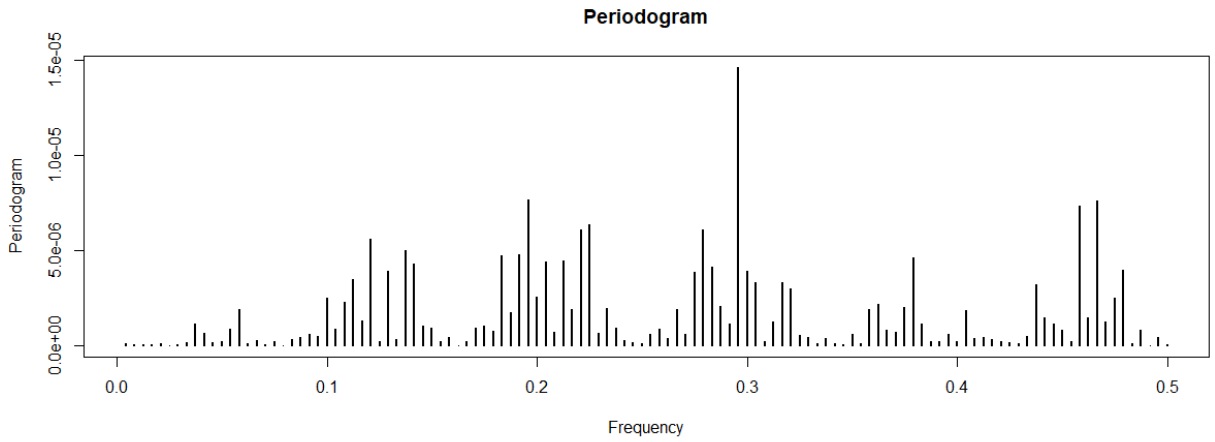

Figure 9: Periodogram of First and First Seasonal Differenced Deaths due to Cardiovascular Issues

Appendix 3: Model fit and Diagnostics- Cardiovascular Deaths:

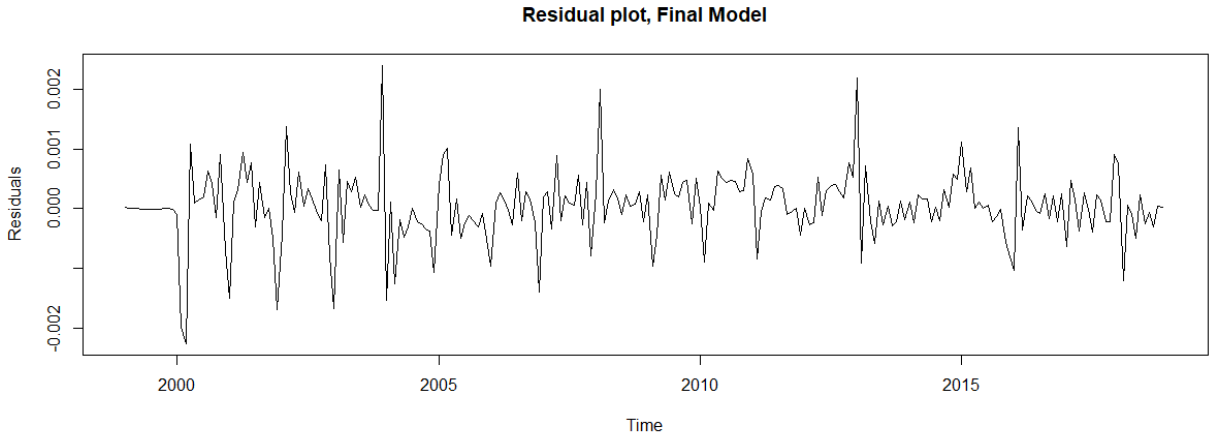

Figure 10: Residual Plot for Final Model of Monthly Deaths Due to Cardiovascular Issues

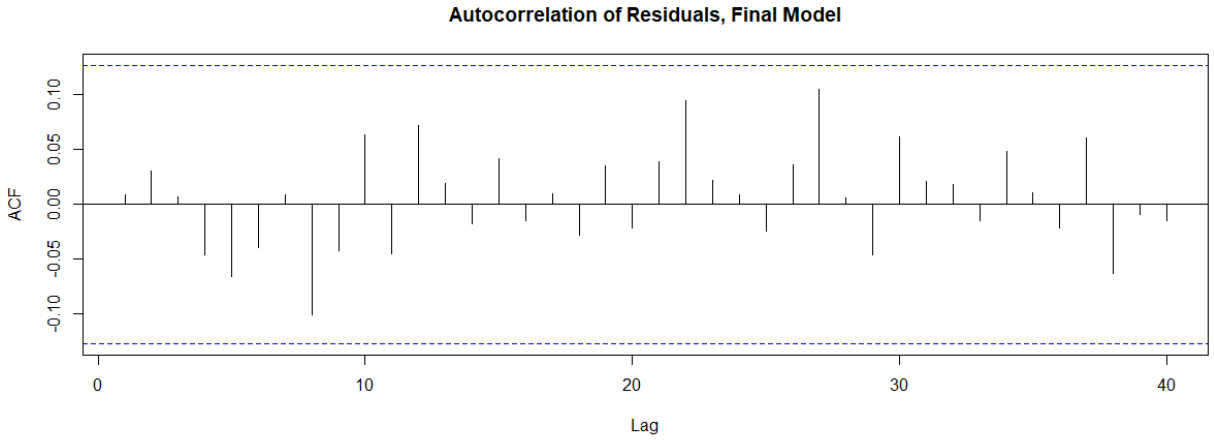

Figure 11: Autocorrelation Plot of Residuals for Final Model of Monthly Deaths Due to Cardiovascular Issues

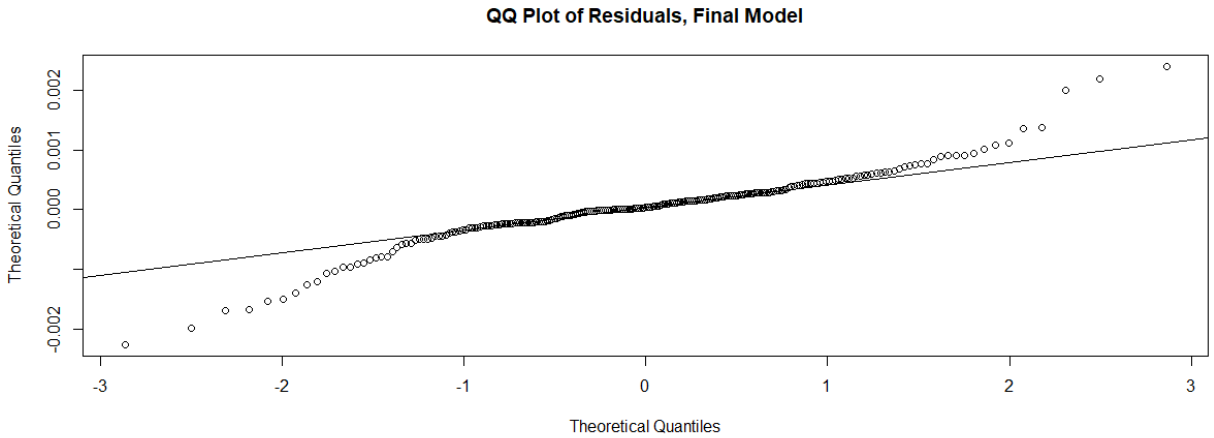

Figure 12: QQ plot of Residuals for Final Model of Monthly Deaths Due to Cardiovascular Issues

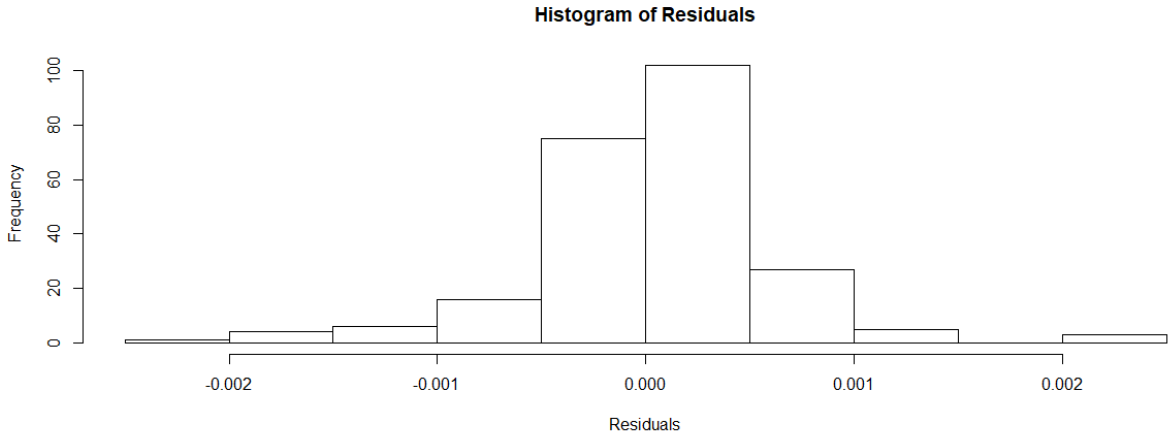

Figure 13: Histogram of Residuals for Final Model of Monthly Deaths Due to Cardiovascular Issues

Appendix 4: Preprocessing – Neoplasm Deaths:

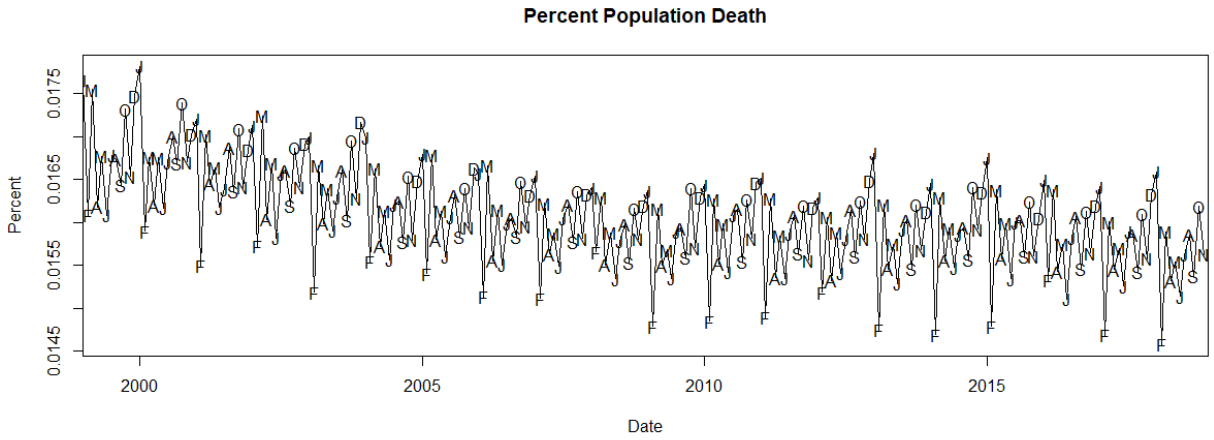

Figure 14: Monthly Percentage of Population Deaths due to Neoplasm Issues

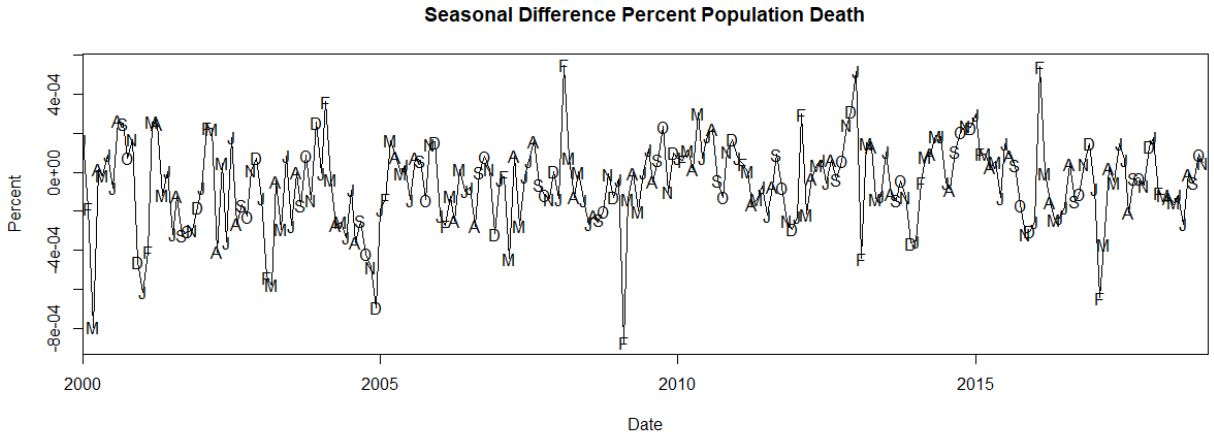

Figure 15: First Seasonal Difference of Population Deaths due to Neoplasm Issues

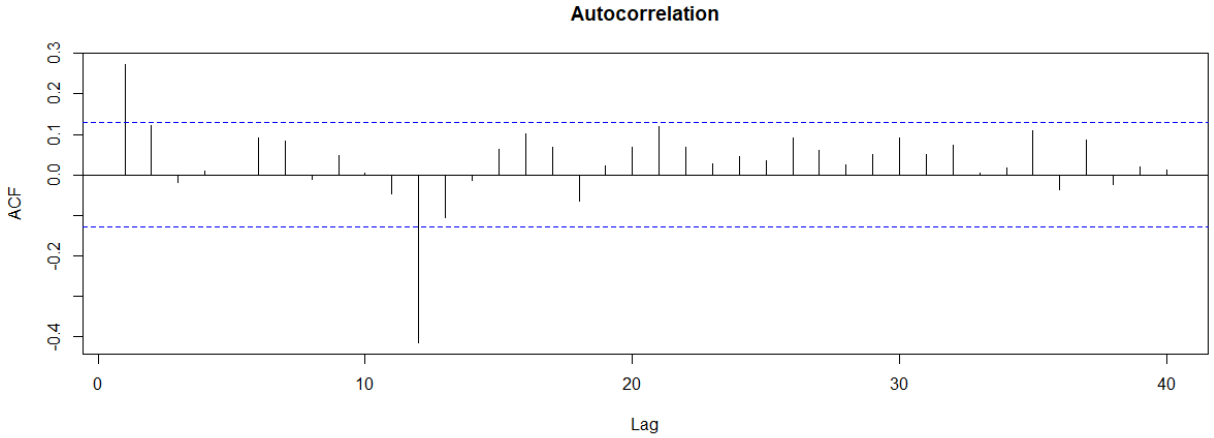

Figure 16: Autocorrelation Function of First Seasonal Differenced Deaths due to Neoplasm Issues

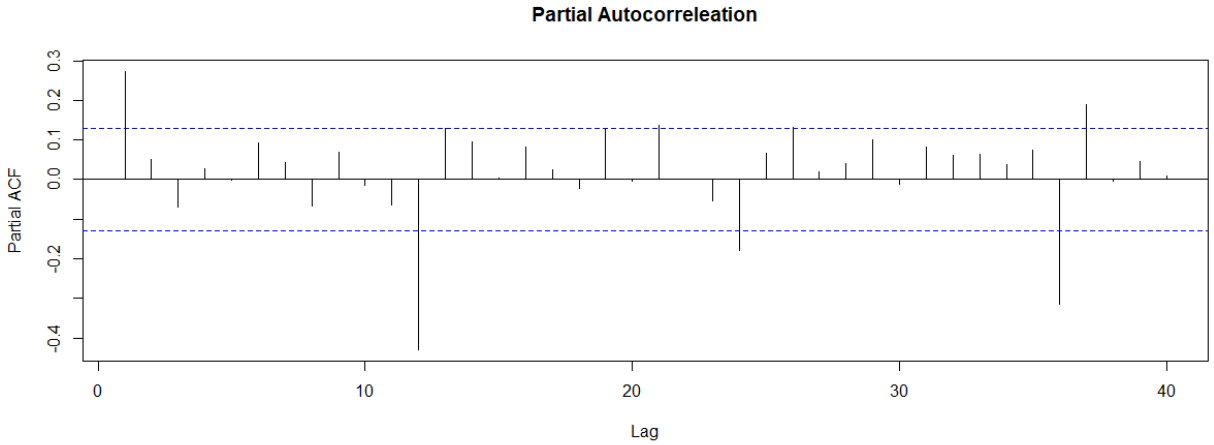

Figure 17: Partial Autocorrelation Function of First Seasonal Differenced Deaths due to Cardiovascular Issues

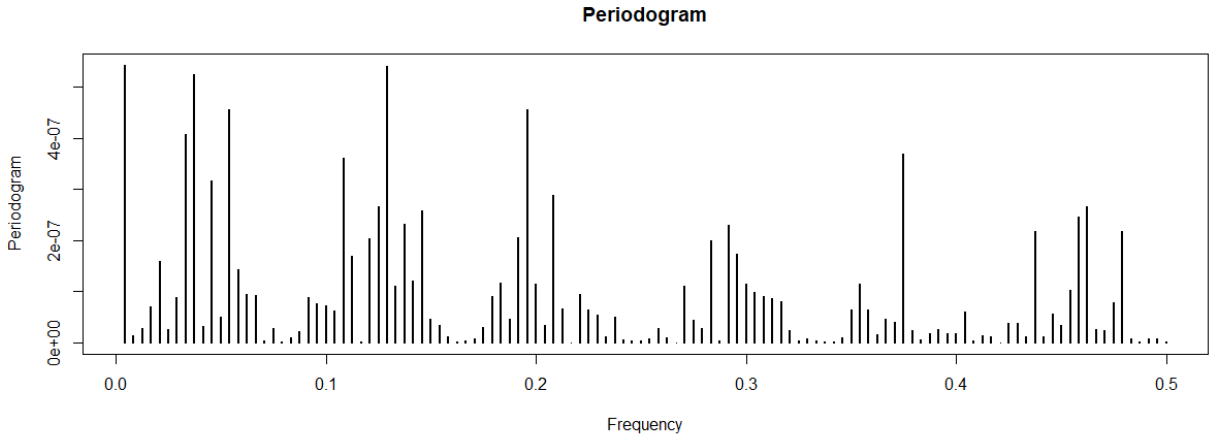

Figure 18: Periodogram of First Seasonal Differenced Deaths due to Cardiovascular Issues

Appendix 5: Model fit and Diagnostics- Neoplasm Deaths:

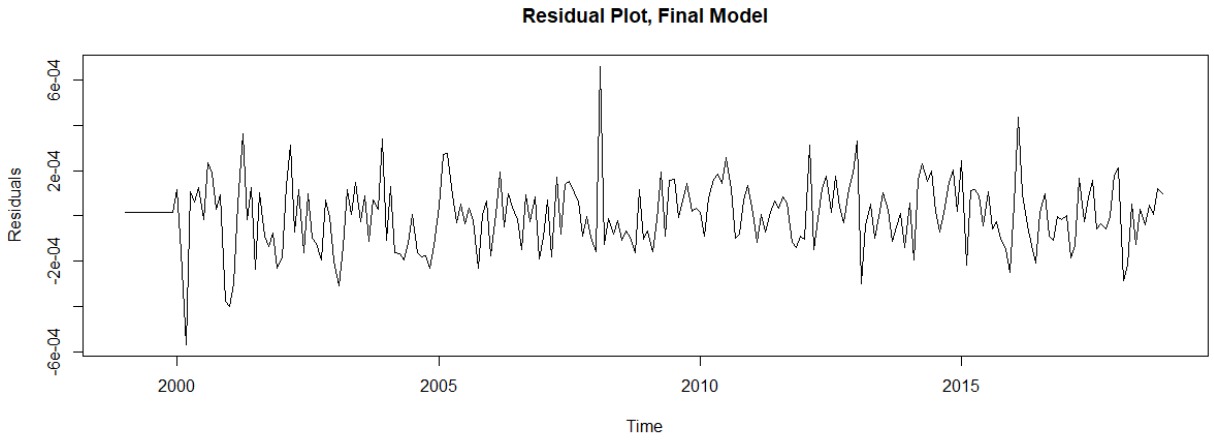

Figure 19: Residual Plot for Final Model of Monthly Deaths Due to Neoplasm Issues

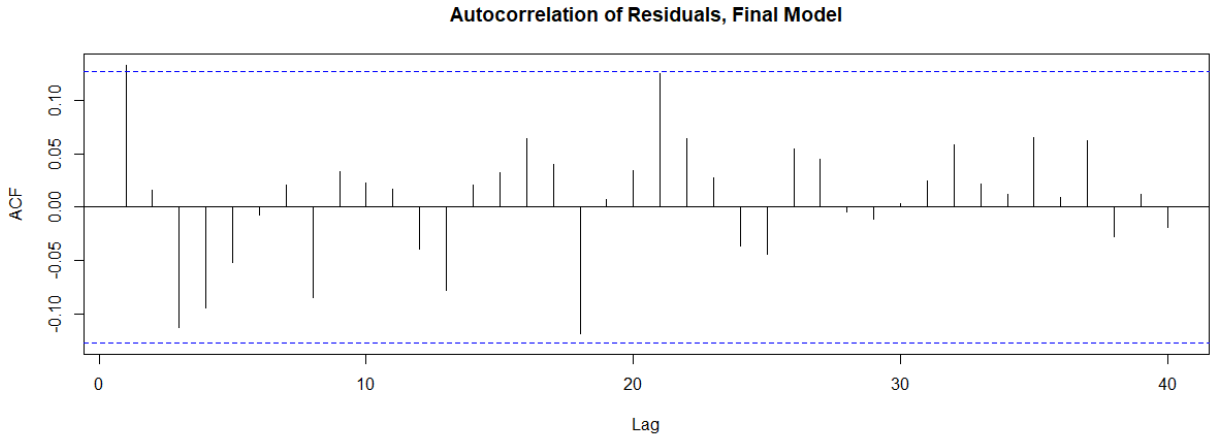

Figure 20: Autocorrelation Plot of Residuals for Final Model of Monthly Deaths Due to Neoplasm Issues

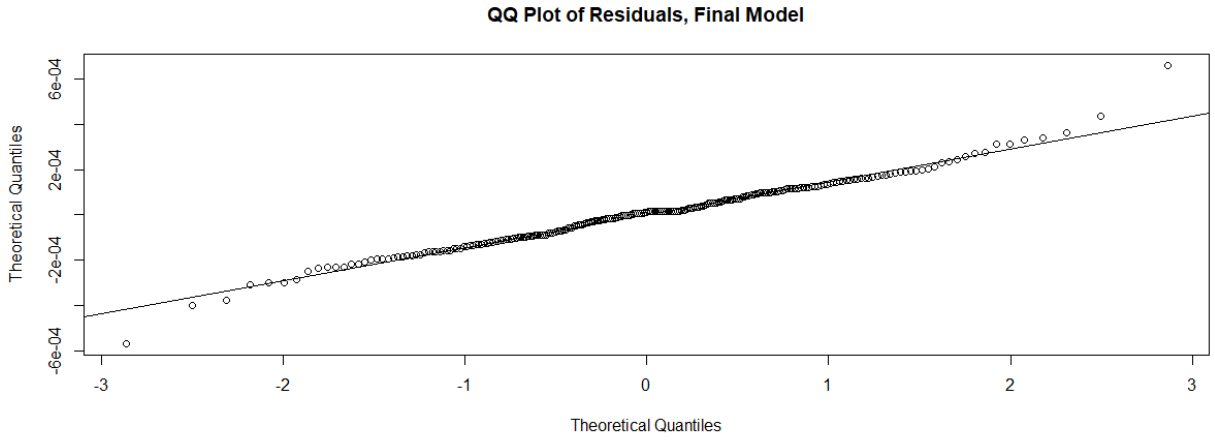

Figure 21: QQ Plot of Residuals for Final model of Monthly Deaths Due to Neoplasm Issues

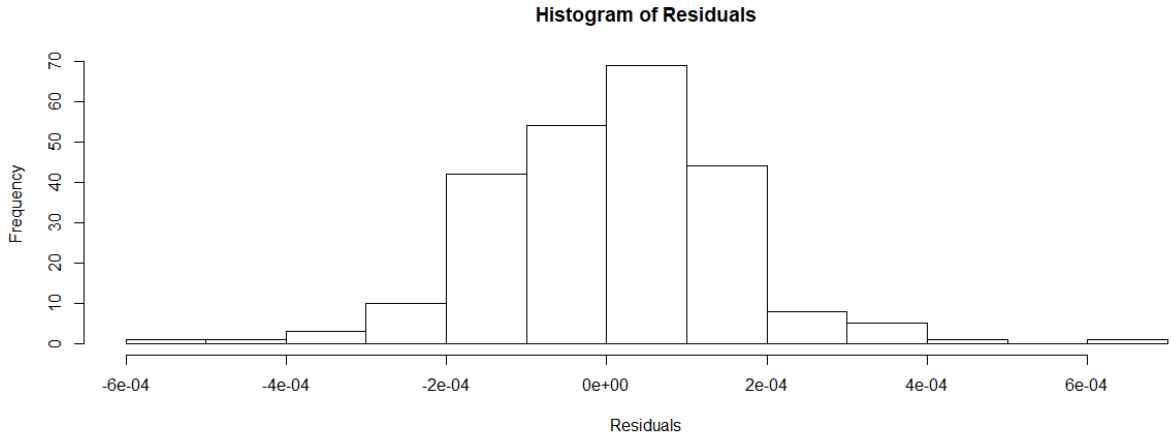

Figure 22: Histogram of Residuals for Final model of Monthly Deaths Due to Neoplasm Issues

Appendix 6: Final Predictions:

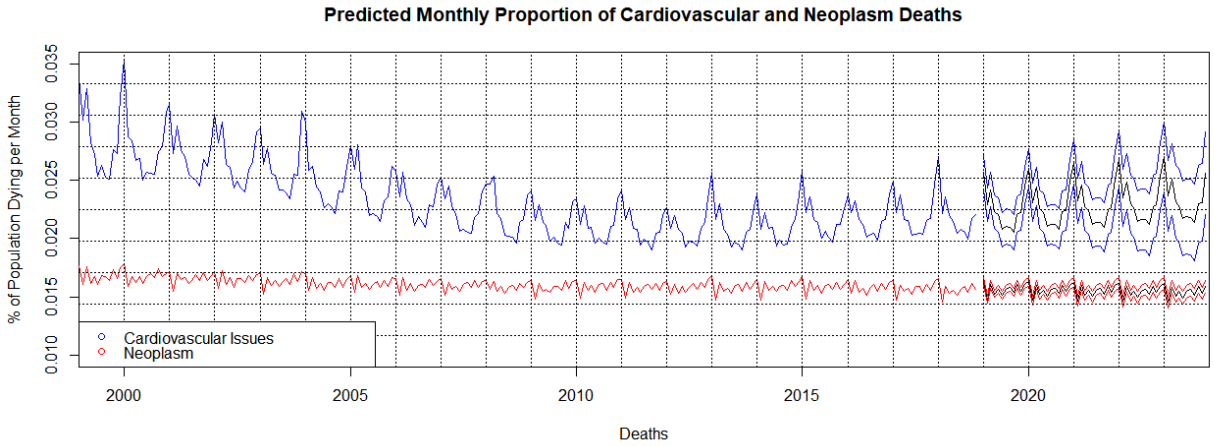

Figure 23: Plot of Monthly Deaths due to Cardiovascular and Neoplasm Issues, with 60 months predicted and 95% confidence intervals.