Air Quality Index as Covariate in Predicting Deaths Due to Chronic Lower Respiratory Diseases per Month

Brianna Mao PSTAT 174

Contents

Summary

Poor air quality is linked to poorer health, but it may be possible to know how many months of past air quality index data are significant in observing the percentage of deaths due to chronic lower respiratory diseases in a month. The monthly death counts and daily air quality measures in Kern County and Los Angeles County in California are studied to find out which combination of months of air quality data produces the most accurate model for the monthly death amounts. The study found that the current month's air quality and the current and past one month's air quality are the most significant in affecting the percentage of deaths in a month. However, the models do not necessarily lend to good or accurate forecasts of future deaths given predictions of air quality.

Introduction

The effects of pollution on health are often-studied, with the general consensus that increasing amounts of pollution leads to detrimental health. The World Health Organization (2018) estimates that 7 million people die each year due to breathing in polluted air, which can cause diseases such as "stroke, heart disease, lung cancer, chronic obstructive pulmonary diseases and respiratory infections".

It is intuitive that worse air qualities would have more adverse effects on the human body, but air quality varies from day to day. Is the effect of the air quality in a single month significant in affecting the number of people who die that month from respiratory-related diseases? Or perhaps the accumulated past couple of months of air quality has more significance, and if so, how many months? It may even be possible, though unlikely, that air quality does not help predict the number of deaths that would occur in a month more than the number of people who died in previous months. Being able to forecast how many deaths may occur based on the air quality could help healthcare providers prepare to care for patients or issue warnings to patients and prevent their health from declining.

Analyzing the Air Quality Index (AQI) and the number of deaths categorized under chronic lower respiratory diseases (CLRD) within two California counties over the span of fifteen years may give insight to these questions. Kern County has twelve times the rate of people who die of chronic respiratory diseases than that of the entirety of California (Perez 2018), making it a sensible choice to study. Los Angeles County boasts a large population and provides a good source of robust data for analysis.

Deaths per County

The California Department of Public Health, Center for Health Statistics and Informatics, Vital Statistics Branch curates a set of datasets called "Death Profiles by County". The data in each dataset span at least four years and contains the number of deaths that occur in each California county per month based on data entered on death certificates. The data can be filtered by cause of death, including death due to CLRD. Due to the data being collected based on death certificates, cases that need to be further investigated will lead to delays in certification, but this database is still a good starting point for many research questions. This data can help spot any trends or lack thereof in the number of people who die due to various causes over the years, from as early as 1970 to 2021.

Air Quality Index

The United States Environmental Protection Agency (U.S. EPA) collects air quality measurements from outdoor monitors in the United States, Puerto Rico, and U.S. Virgin Islands. The data is recorded hourly, daily, and annually. The daily AQI measurements by county in the United States can be downloaded as files for each year from the EPA's Air Data website.

Air quality data is often used in studies to track if areas are doing well in decreasing air pollution, or the effects of poor air quality on people's health. For example, one study of $PM_{2.5}$ and O_3 exposure, two

chemicals that make up part of the calculation of the general Air Quality Index,in California during 2012 concluded that "[n]onlocal emissions could lead to more deaths than local emissions" (Wang *et al.*, 2019). Air quality data thus allows for a range of exploration possibilites that can directly relate to people's lives.

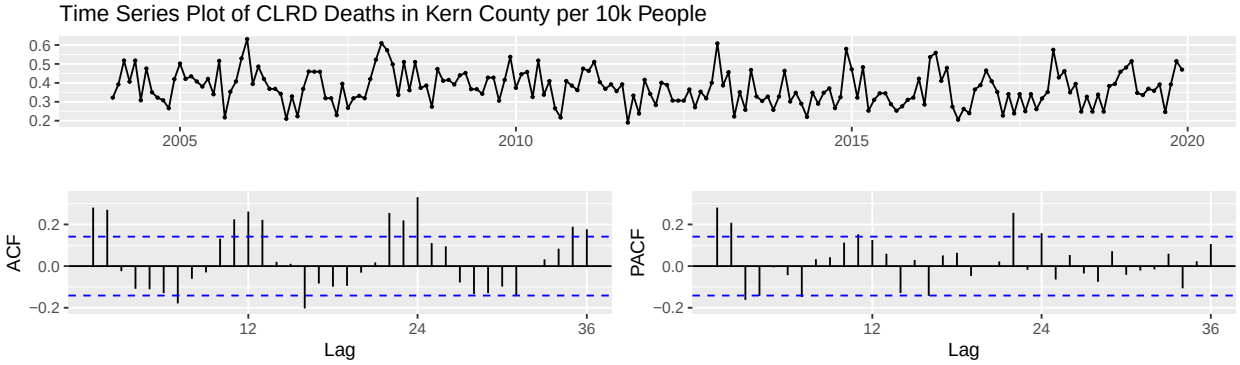
Data Analysis

To perform forecasting with models, often the time series is assumed to be stationary, having a constant mean, constant variance, and no autocorrelation dependent on time. Data in the real world oftentimes contain trends and seasonality, which must be removed before models can be fit on the residuals.

Kern County Deaths

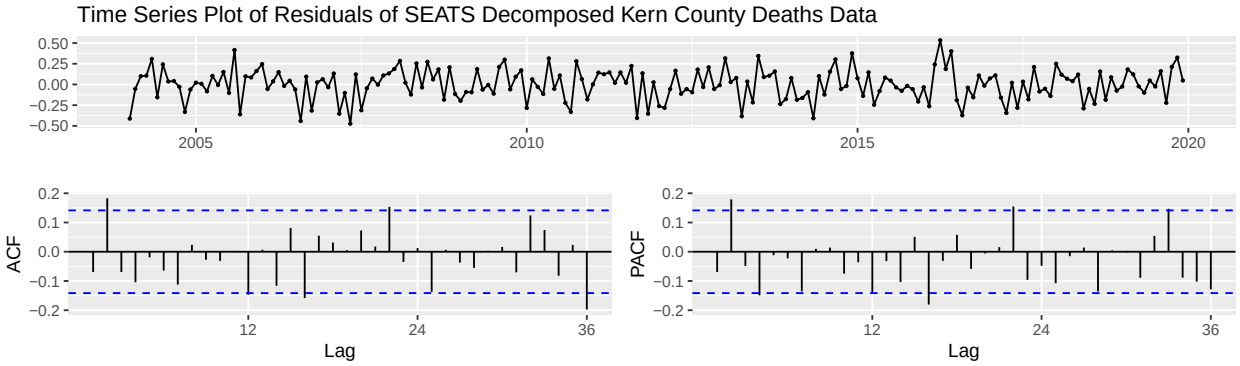
Monthly death counts in Kern County from 2004 to 2019 are obtained from the "Death Profiles by County" dataset. The number of deaths are measured monthly by counts and are converted to percentages per populations of 10,000 by dividing the count by the estimated population of Kern County in the previous year. The population estimates are derived from the United States Census Bureau estimates, the most accurate source for population estimates in the United States.

The `kern_deaths` dataset contains 192 observations of the percentage of the population who died of CLRD per 10,000 people and has no missing values. From the plot of the time series, it appears that the variance is non-constant. there is a slight downwards trend in the mean, and there may also be some seasonality in the time series.



The variance can be normalized by applying a log-transformation on the data. Since the time series is measured monthly, the Seasonal Extraction in ARIMA Time Series method, or SEATS, can be used to remove any trend and seasonality. A plot of the original data, seasonally adjusted data, and trend of the data is available in the Appendix (**Figure A**).

The plot of the residuals of the decomposed data shows that the residuals have a constant mean of about 0 and a more constant variance than before. From the autocorrelation plot of the residuals, there is significant autocorrelation at lags 2, 16, and 36. The partial autocorrelation plot of the residuals show significance at lags 2, 16, and 22.



To check if the residuals are stationary, the Augmented Dickey-Fuller is used to see if a time series has a unit root. The test assumes the model follows the form

$$Y_t = \mu + \beta t + \alpha Y_{t-1} + \sum_{j=1}^{p} \phi_j \Delta Y_{t-j} + \epsilon_t$$

and tests the hypotheses

$$H_0 : \alpha = 1 \text{ with } p = 0 \text{ (the series is non-stationary)} \quad \text{vs} \quad H_a : \text{the series is stationary}$$

If the p-value is smaller than 0.05, the null hypothesis is rejected in favor of the alternate hypothesis.

The ADF test returns a p-value of 0.01, thus the null hypothesis is rejected and the residuals of the transformed de-seasonalized data are stationary.

The residuals have 192 observations and no missing values. The residuals have a sample mean of

$$\hat{\mu} = \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t = \frac{1}{192}\sum_{t=1}^{192} Y_t = 0,$$

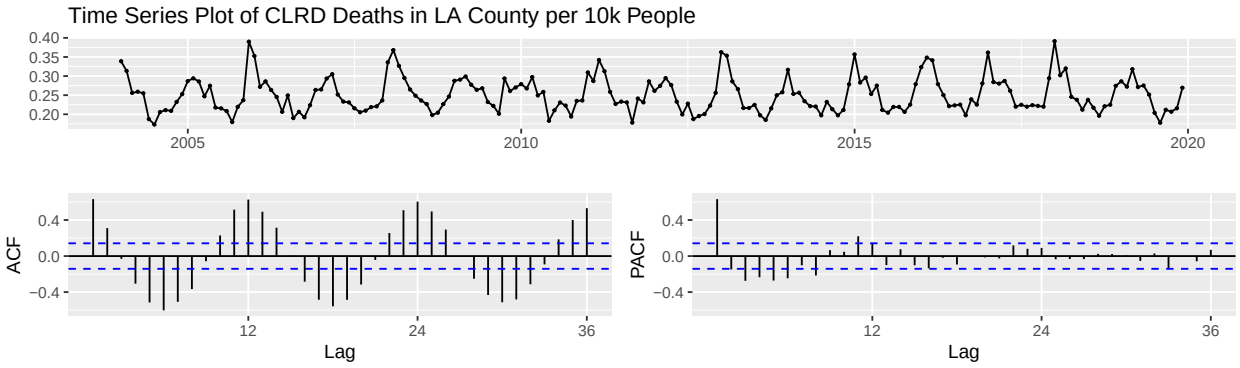
and sample variance of

$$\text{Var}(\bar{Y}) = \frac{1}{T}\sigma^2 \left[1 + 2\sum_{j=1}^{T-1}\left(1 - \frac{|j|}{T}\right)\rho(j)\right] = \frac{1}{192}(0.035)\left[1 + 2\sum_{j=1}^{191}\left(1 - \frac{|j|}{192}\right)\rho(j)\right] = 0.000009.$$

Los Angeles Deaths

Monthly death counts in Los Angeles County from 2004 to 2019 are obtained from the "Death Profiles by County" dataset. The number of deaths are measured monthly by counts and are converted to percentages per populations of 10,000 by dividing the count by the estimated population of Los Angeles County in the previous year. The population estimates are derived from the United States Census Bureau estimates, the most accurate source for population estimates in the United States.

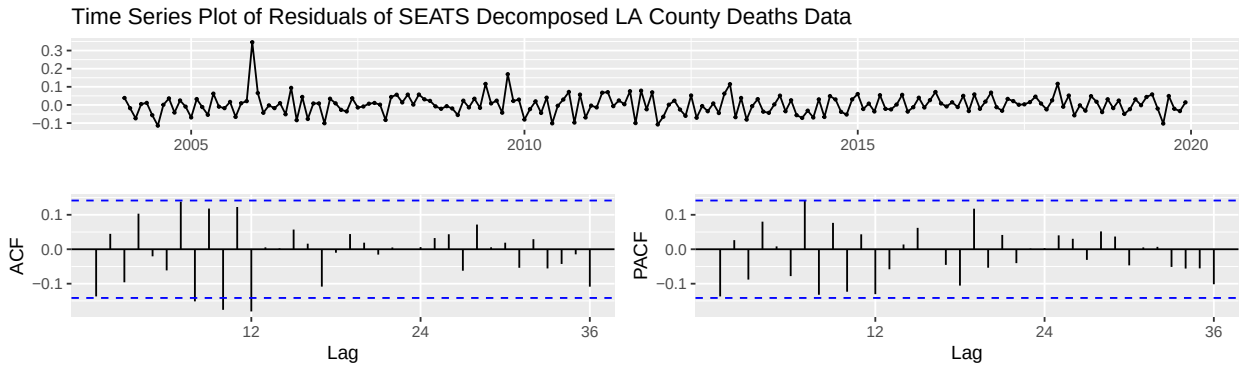
The `la_deaths` dataset contains 192 observations of the percentage of the population who died of CLRD per 10,000 people and has no missing values. From the plot of the time series, it appears that like the `kern_deaths` dataset, the variance is non-constant and the time series may also have seasonality.



The variance can be normalized by applying a log-transformation on the data. Since the time series is measured monthly, the SEATS method can be used to remove any trend and seasonality. A plot of the original data, seasonally adjusted data, and trend of the data is available in the Appendix (**Figure B**).

The plot of the residuals of the decomposed data shows that the data has a constant mean of about 0 and a temporary shock in December 2005, but otherwise constant variance. The autocorrelation plot of the

residuals shows that there is significant correlation at lags 10 and 12. The partial autocorrelation plot of the residuals shows no significant lags at all.



Once again, to check if the residuals are stationary, the Augmented Dickey-Fuller is used with the same model assumptions and hypotheses as before. The p-value from the test is 0.01 is smaller than 0.05, thus the residuals of the transformed de-seasonalized time series are stationary.

The residuals 192 observations and no missing values. The residuals have a sample mean of

$$\hat{\mu} = \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t = \frac{1}{192}\sum_{t=1}^{192} Y_t = 0.0018,$$

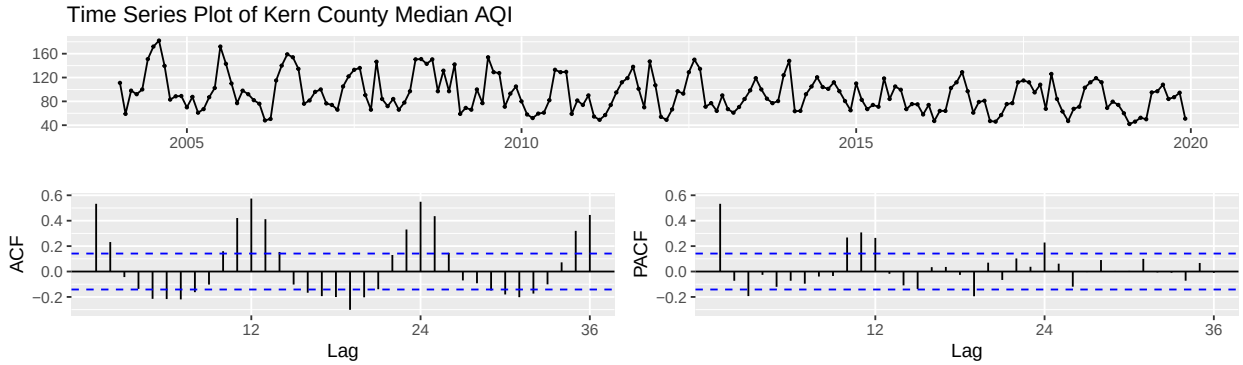
and sample variance of

$$\text{Var}(\bar{Y}) = \frac{1}{T}\sigma^2\left[1 + 2\sum_{j=1}^{T-1}\left(1 - \frac{|j|}{T}\right)\rho(j)\right] = \frac{1}{192}(0.0029)\left[1 + 2\sum_{j=1}^{191}\left(1 - \frac{|j|}{192}\right)\rho(j)\right] = 0.000002.$$

Kern County AQI

Median AQI in Kern County from 2004 to 2019 are obtained from the datasets created by the US EPA. The AQI are measured daily and are aggregated into monthly medians in order to have the same time periods as the county deaths data.

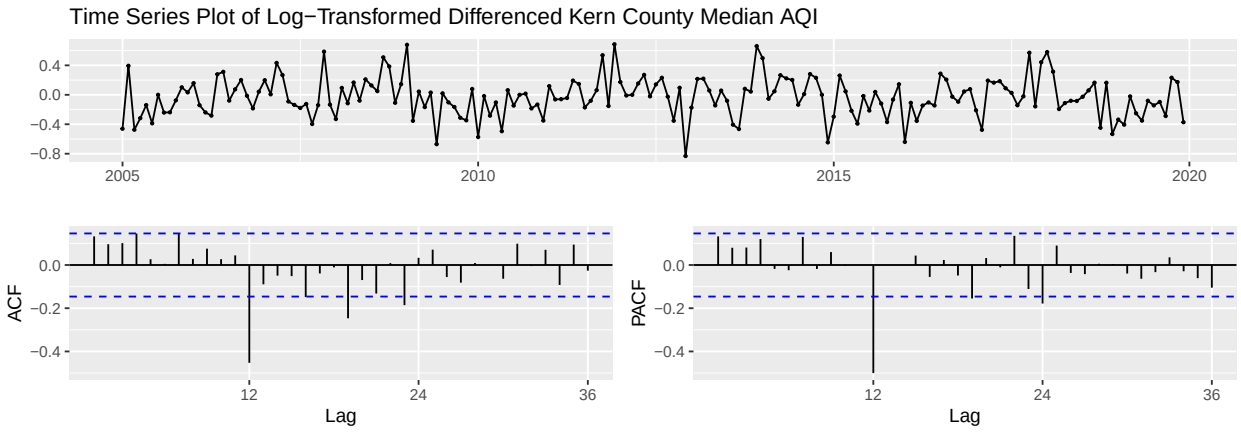
The `kern_aqi` dataset contains 192 observations of the median air quality index in Kern County and contains no missing values. From the plot of the time series, it appears that the variance is non-constant and there is a negative linear trend.



Since the data is also measured monthly, the SEATS decomposition can be used to remove trend and seasonality. However, after performing the SEATS decomposition on the Los Angeles County deaths data, the returned residuals resemble white noise. If the residuals are white noise, the data has no significant

structure and thus would not be a particularly beneficial covariate to have when regressing the monthly deaths data.

The data can be transformed in other ways. A logarithm is applied to normalize the difference, and a difference at lag set to 12 is used to remove the seasonality. The plot of the transformed data shows that the data has a constant mean of about 0 and a more constant variance than before. The autocorrelation plot of the residuals shows that there is significant correlation at lags 12 and 19. The partial autocorrelation plot of the residuals shows significance at lags 12 and 24 as well.



The Augmented Dickey-Fuller is applied to check stationarity of the residuals using the same assumptions and hypothesis as before. The test returns a p-value 0.01 which is less than 0.05, thus the null hypothesis is rejected and the transformed data is considered stationary.

The log-transformed differenced data contains 180 observations and no missing values. The data has a sample mean of

$$\hat{\mu} = \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t = \frac{1}{180}\sum_{t=1}^{180} Y_t = -0.03,$$

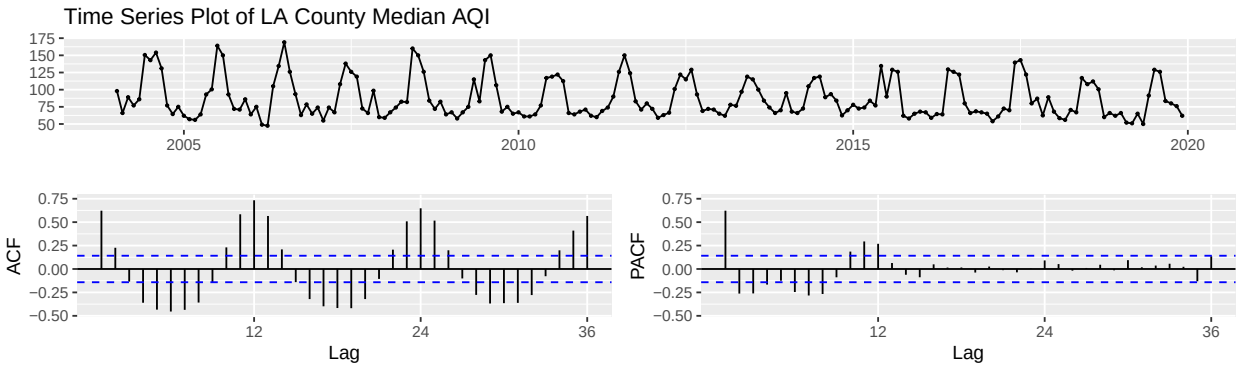
and sample variance of

$$\text{Var}(\bar{Y}) = \frac{1}{T}\sigma^2\left[1 + 2\sum_{j=1}^{T-1}\left(1 - \frac{|j|}{T}\right)\rho(j)\right] = \frac{1}{180}(0.07)\left[1 + 2\sum_{j=1}^{179}\left(1 - \frac{|j|}{180}\right)\rho(j)\right] = 0.00007.$$

Los Angeles County AQI

Median AQI in Los Angeles County from 2004 to 2019 are obtained from the datasets created by the US EPA. The AQI are measured daily and are aggregated into monthly medians in order to have the same time periods as the county deaths data.

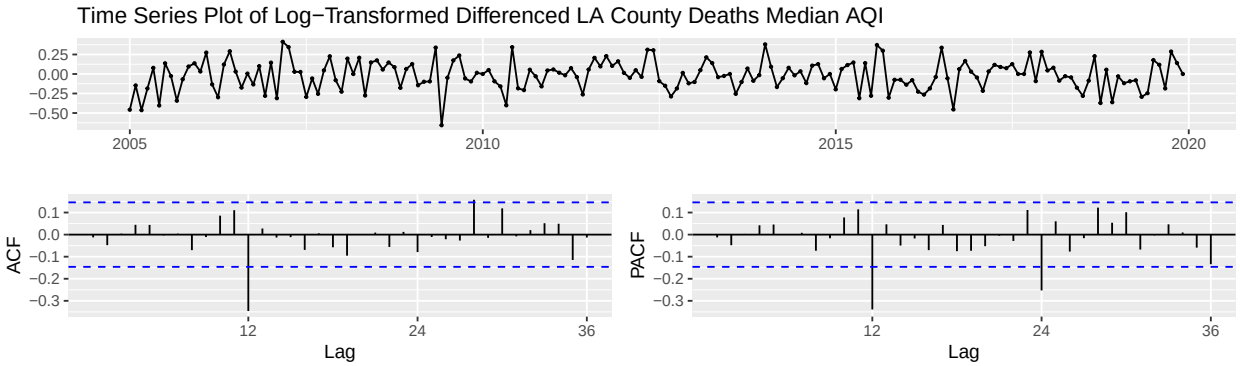
The `la_aqi` dataset contains 192 observations of the median air quality index in Los Angeles County and contains no missing values. From the plot of the time series, it appears that the variance is non-constant and there is a negative linear trend.



Since the data is also measured monthly, the SEATS decomposition can be used to remove trend and seasonality. However, after performing the SEATS decomposition on the Los Angeles County deaths data, the residuals resemble white noise, just like with the Kern County AQI data. Decomposing the data in this way would not create a beneficial covariate dataset.

Alternate methods can be used to transform the data. The variance can be normalized by applying a log-transformation on the dataset. Since the time series is measured monthly, differencing with the lag set to 12 might remove any seasonal trend.

The plot of the transformed data shows that the data has a constant mean of about 0 and a more normalized variance than before. The autocorrelation plot of the residuals shows that there is significant correlation at lag 12. The partial autocorrelation plot of the residuals shows significant correlation at lags 12 and 24.



The Augmented Dickey-Fuller is applied to check stationarity of the residuals using the same assumptions and hypotheses as before. The returned p-value of 0.01 is less than 0.05, thus the null hypothesis is rejected and the transformed data is stationary.

The log-transformed differenced data contains 180 observations and no missing values. The data has a sample mean of

$$\hat{\mu} = \bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t = \frac{1}{180}\sum_{t=1}^{180} Y_t = -0.02,$$

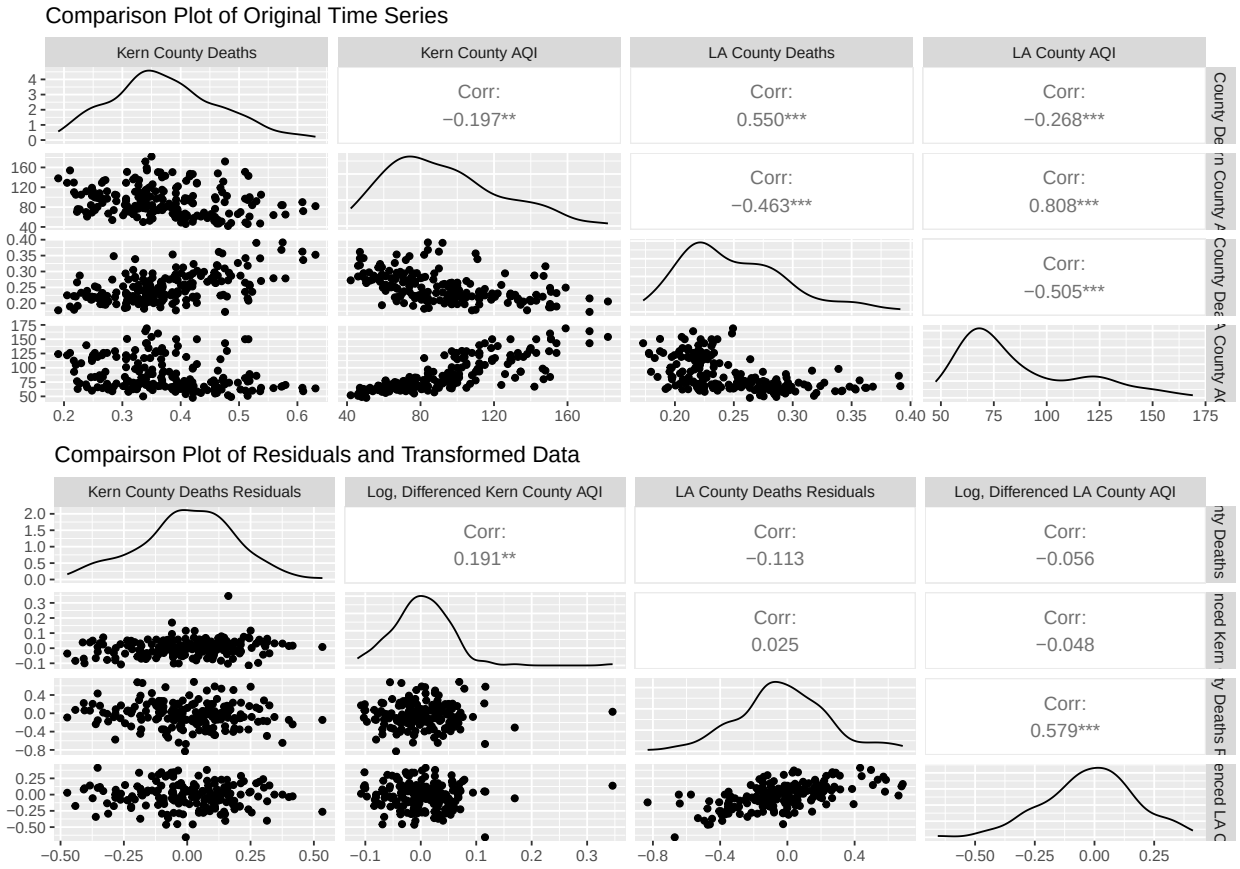
and sample variance of

$$\text{Var}(\bar{Y}) = \frac{1}{T}\sigma^2\left[1 + 2\sum_{j=1}^{T-1}\left(1 - \frac{|j|}{T}\right)\rho(j)\right] = \frac{1}{180}(0.04)\left[1 + 2\sum_{j=1}^{179}\left(1 - \frac{|j|}{180}\right)\rho(j)\right] = 0.00003.$$

Relationships Between Data

Assuming that air quality data can be used to estimate the number of deaths per month attributed to CLRD, correlations between the time series are expected. Correlation plots between the original datasets and the

transformed datasets are shown below.



Models

Time series data can be fit into many different types of models, but not all of the models are the best in predicting future data. Good models for time series data separate the response into different components and a white noise component, where the white noise is stationary and uncorrelated with constant variance. Models with lower AIC and lower BIC values perform better than models with higher AIC and higher BIC, thus the former model would be chosen in favor over the latter.

auto.arima Models on Deaths Data

The correlation between the deaths and AQI data for each of the counties is not very high, so it is possible that the number of deaths each month depends more on just the past deaths data.

The residuals of the SEATS decomposed `kern_deaths` and `la_deaths` time series are stationary, meaning they do not need to be differenced further to achieve a constant mean, so an ARMA or SARMA model may be suitable. An ARMA or SARMA model is built based on past observations of data, current and past observations of white noise, and sometimes a periodic component if the data is measured monthly, like it is here. The `auto.arima` model can fit the best ARIMA model as chosen by the computer.

The `auto.arima` command selects SARMA models for both the `kern_deaths` and `la_deaths` residuals. The SARMA$(p,q)\times(P,Q)_s$ model takes on the form

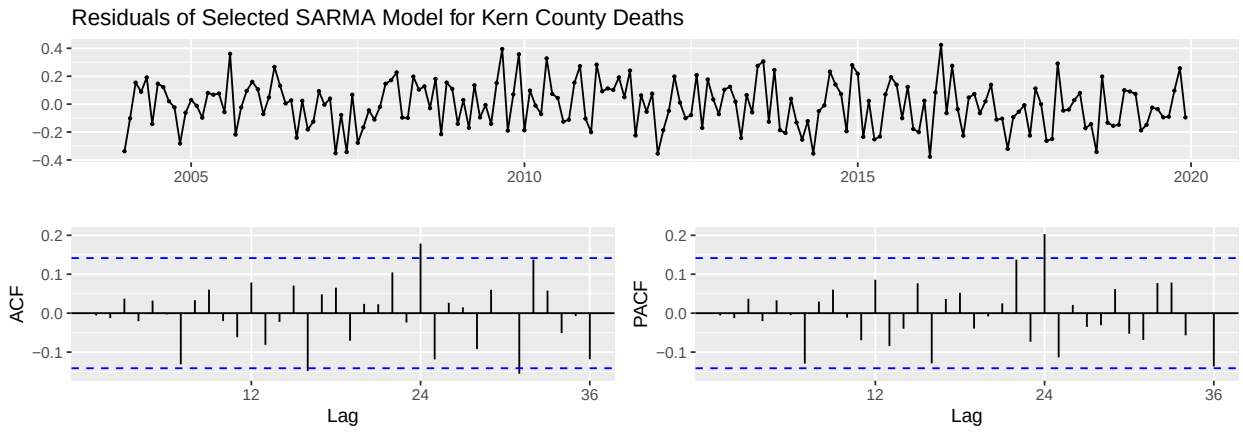
$$\begin{aligned}\phi(B)\Phi(B)_{P,s}Y_t &= \theta(B)\Theta(B)_{Q,s}\epsilon_t\\ &\text{where}\\ Y_t\cdot B^n &= Y_t - Y_{t-1} - Y_{t-2} - \cdots - Y_{t-n}\\ \epsilon_t\cdot B^n &= \epsilon_t - \epsilon_{t-1} - \epsilon_{t-2} - \cdots - \epsilon_{t-n}\\ \phi(B) &= 1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\\ \Phi(B)_{P,s} &= 1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{Ps}\\ \theta(B) &= 1+\theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\\ \Theta(B)_{Q,s} &= 1+\Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}\end{aligned}$$

where s is the length of the periodic component in the time series, Y_t is the observed number of deaths in month t, and ϵ_t is white noise observed in month t.

The chosen model for the residuals of the `kern_deaths` data is a SARMA$(1,3)\times(0,2)_{12}$ model:

$$(1-0.7214B)Y_t = (1-0.8815B+0.3550B^2-0.3391B^3)(1-0.3923B^{12}-0.3906B^{24})\epsilon_t$$

with an AIC of -113.8 and a BIC of -91. The residuals from this model are shown below.



To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model. The Ljung-Box test tests the hypotheses

$$H_0 : \text{the data is i.i.d} \quad \text{vs} \quad H_a : \text{the data is not i.i.d}$$

by calculating

$$Q = T \sum_{j=1}^{k} \hat{\rho}_Y^2(j)$$

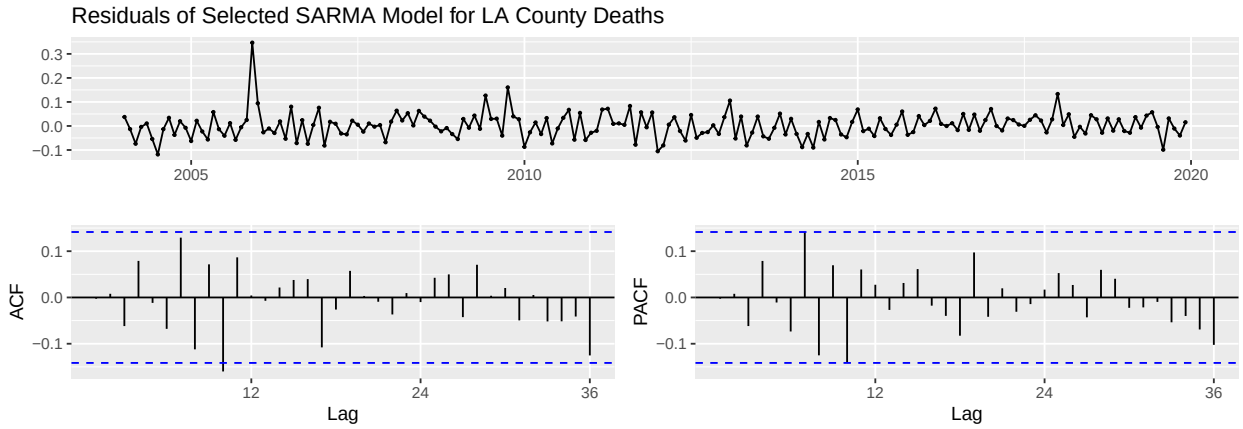
which can be approximated by a χ_k^2 distribution under the null hypothesis. If the p-value returned by the test is greater than 0.05, the null hypothesis is not rejected and the data appears to be independent identically distributed observations, which would indicate the data is also white noise. Otherwise, the data would not be white noise.

The test returns a p-value of 0.9348 which is greater than 0.05, thus the residuals of the fitted SARMA model are white noise.

The chosen model for the residuals of the `la_deaths` data is a SARMA$(0,1) \times (0,1)_{12}$ model:

$$Y_t = (1 - 0.1176B)(1 - 0.1946B^{12})\epsilon_t$$

with an AIC of -579.69 and a BIC of -570.05. The residuals from this model are shown below.



To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.97 which is greater than 0.05, thus the residuals of the fitted SARMA model are white noise.

Manual ARMA Models on Deaths Data

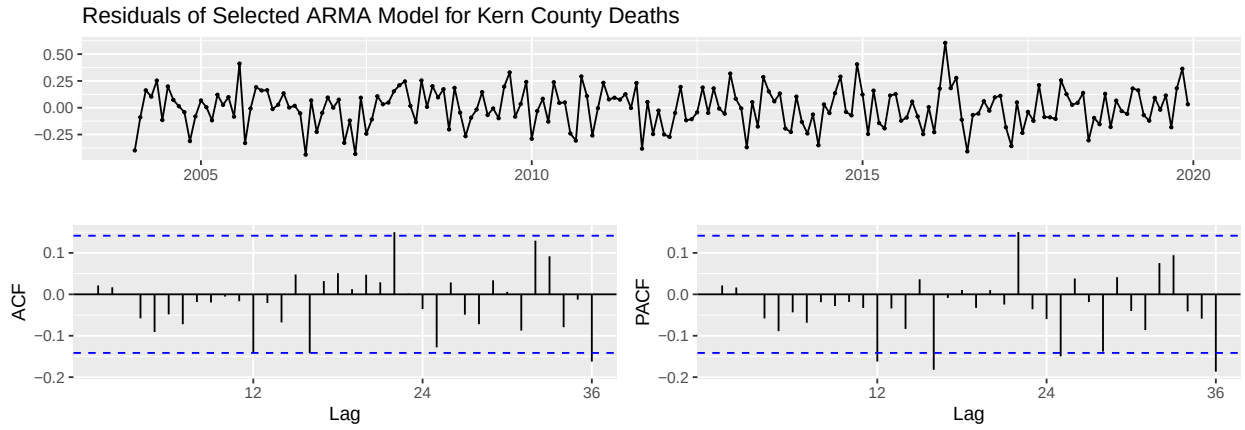
Although the computer supposedly selected the best ARIMA model, perhaps other models may do better. Looking at the ACF and PACF plots of the residuals of the `kern_deaths` data, there is a significant lag at time 2 for both. This may correspond to an ARMA(2, 2) model:

$$Y_t - \mu_Y = \phi_1(Y_{t-1} - \mu_Y) + \phi_2(Y_{t-2} - \mu_Y) + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2}.$$

The manually selected model has the parameters

$$Y_t = -0.7018(Y_{t-1}) - 0.5793(Y_{t-2}) + \epsilon_t + 0.6388\epsilon_{t-1} + 0.7383\epsilon_{t-2}$$

with an AIC of -99.4 and a BIC of -79.85. The residuals from this model are shown below.



To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.7651 which is greater than 0.05, thus the residuals of the fitted ARMA model are white noise.

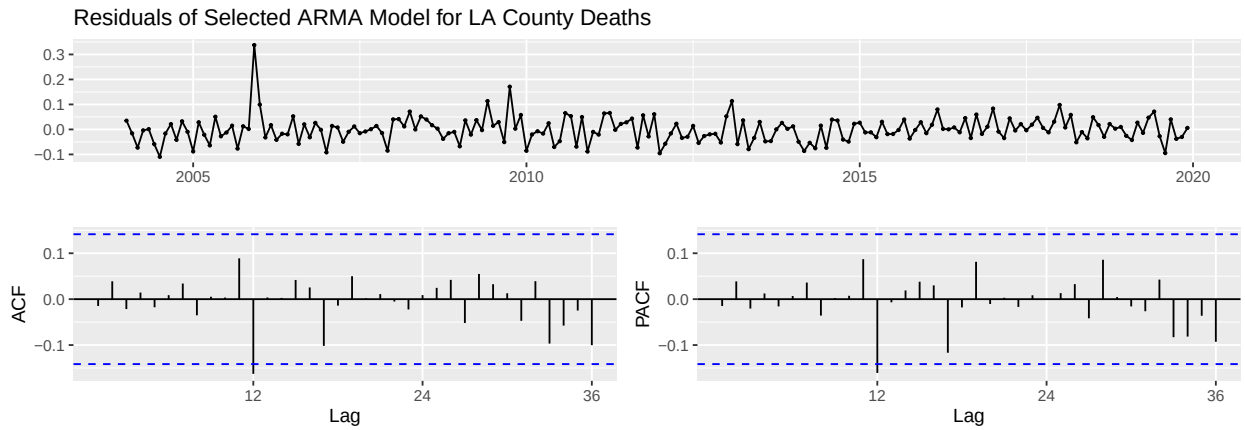
Looking at the ACF and PACF plots of the residuals of the `la_deaths` data, there is a significant lag at time 10 in the ACF and no significant lag in the PACF. This may correspond to an ARMA(0, 10) model:

$$Y_t - \mu_Y = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_{10}\epsilon_{t-10}$$

The manually selected model has the parameters:

$$\begin{aligned} Y_t - 0.002 = \epsilon_t &- 0.0592\epsilon_{t-1} - 0.0775\epsilon_{t-2} - 0.0324\epsilon_{t-3} + 0.0823\epsilon_{t-4} + 0.0091\epsilon_{t-5} \\ &- 0.0702\epsilon_{t-6} + 0.0755\epsilon_{t-7} - 0.0929\epsilon_{t-8} + 0.1088\epsilon_{t-9} - 0.1947\epsilon_{t-10} \end{aligned}$$

with an AIC of -569.96 and a BIC of -530.87. The residuals from this model are shown below.



To see if the ϵ_t component is white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.8357 which is greater than 0.05, thus the residuals of the fitted ARMA model are white noise.

The manually selected ARMA models do not have lower AIC and BIC than the SARMA models for both datasets, so the ARMA models do not perform better than the SARMA model in terms of AIC and BIC. This may indicate that the residuals of the data after removing seasonality and trend still contain a seasonal component. The SARMA models will be used to make forecasts without covariate data.

Regression Model

Assuming that AQI would have an effect on the number of chronic lower respiratory disease-related deaths, a regression model can be fit with both sets of data. The regression model assumes the form

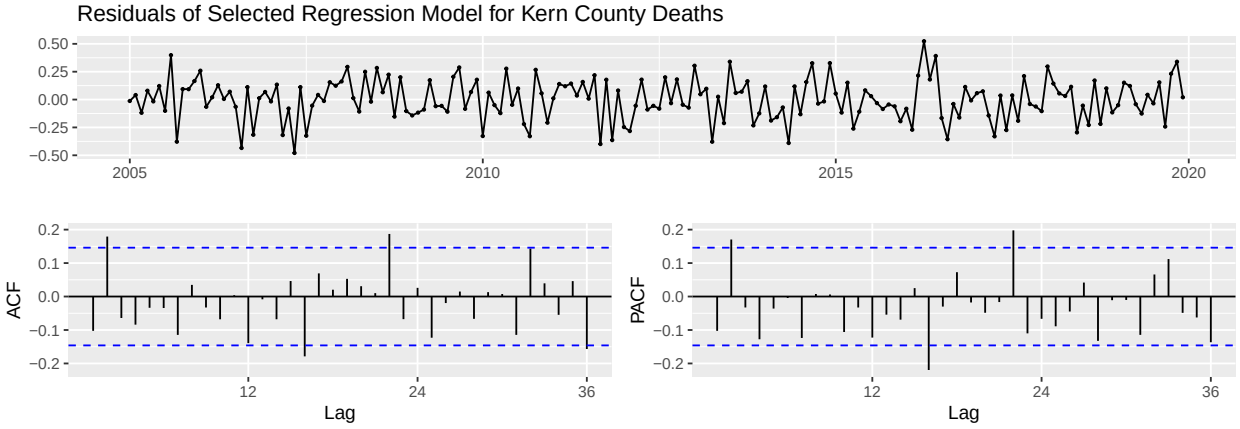
$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_q X_{t-q} + \epsilon_t$$

where Y_t and X_t are stationary random variables and ϵ_t is white noise. Y_t would be the residuals of the number of deaths per month and X_t would be AQI in month t with transformation performed to achieve stationarity.

Multiple regression models are fit on the Kern County data, regressing the residuals of the number of deaths in a month with the residuals of the AQI in that month, the residuals of the AQI in that month and the previous month, and so on. The values of AIC and BIC are calculated for each of the models at the different values of lag (Appendix **Table 1**). The model with the lowest AIC and BIC is the ARDL(0,0) model with an AIC of -90.56477 and BIC of -80.985903:

$$Y_t = -0.001205 - 0.077873 X_t + \epsilon_t.$$

The residuals from this model are shown below.



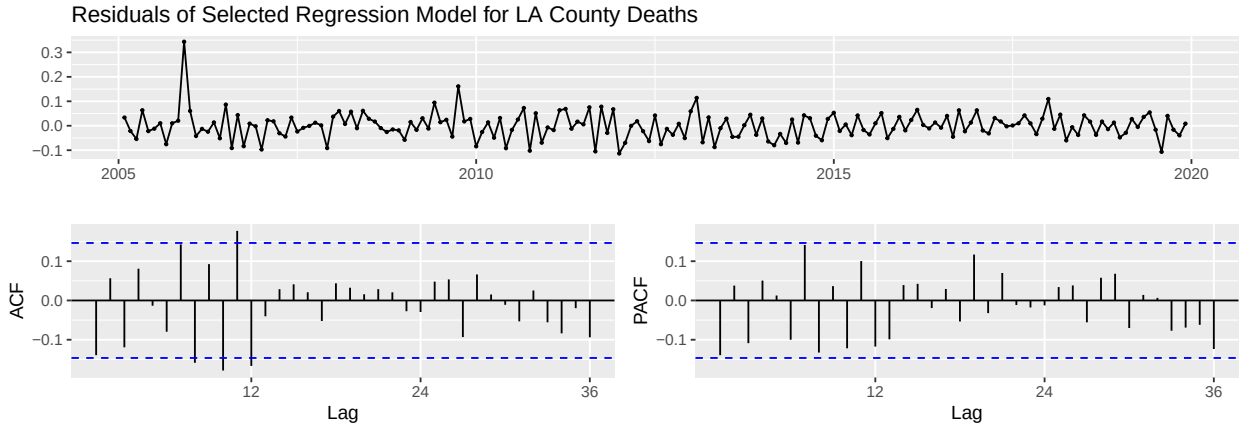
To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.1654 which is greater than 0.05, thus the residuals resemble white noise.

Multiple regression models are fit on the Los Angeles County data with the same procedure as on the Kern County data. The values of AIC and BIC are calculated for each of the models at the different values of lag (Appendix **Table 2**). The model with the lowest AIC is the ARDL(0, 12) model with an AIC of -538.2696 and a BIC of -491.4101, while the model with the lowest BIC is the ARDL(0, 0) model with an AIC of -532.6165 and a BIC of -523.0376. However, both of the residuals of the models fail the Ljung-Box test, indicating that the residuals do not resemble white noise and the model is not a good fit for the data.

The regression model that does not fail the Ljung-Box test and has the lowest AIC and BIC is the ARDL(0, 1) model with an AIC of -529.3588 and a BIC of -516.6093:

$$Y_t = 0.003419 - 0.019076 X_t + 0.016148 X_{t-2} + \epsilon_t$$

The residuals from this model are shown below.



To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.06, which is slightly larger than 0.05, so the null hypothesis is not rejected and the residuals resemble white noise.

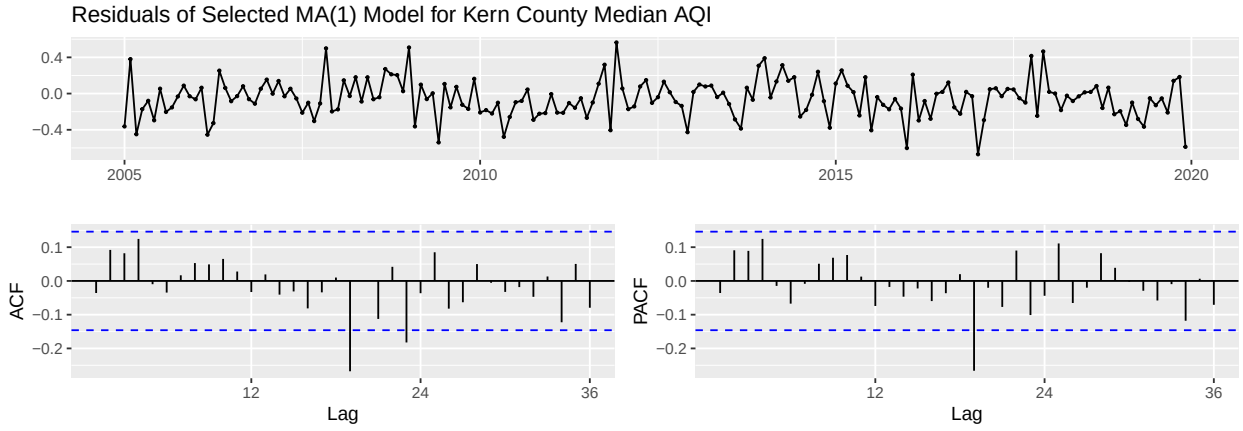
AQI Models

In order to perform forecasts using the models for the deaths per month regressed with the AQI data, forecasts must be made on the values of the AQI data as well. Forecasting the AQI is not the main objective of this study, so a reasonably accurate model will do. From earlier, the `auto.arima` does a good job of fitting a model with low AIC and BIC, so it can be used here again to fit models for the AQI data.

The `auto.arima` model on the residuals of the log-transformed differenced Kern County AQI data returns a SARMA$(0,0,1)\times(1,0,2)_{12}$ model with an AIC of -19.41 and BIC of -3.45:

$$(1+0.3766B^{12})Y_t = (1+0.1831B)(1-0.3766B^{12}-0.3703B^{24})\epsilon_t$$

The residuals of this model is shown below.

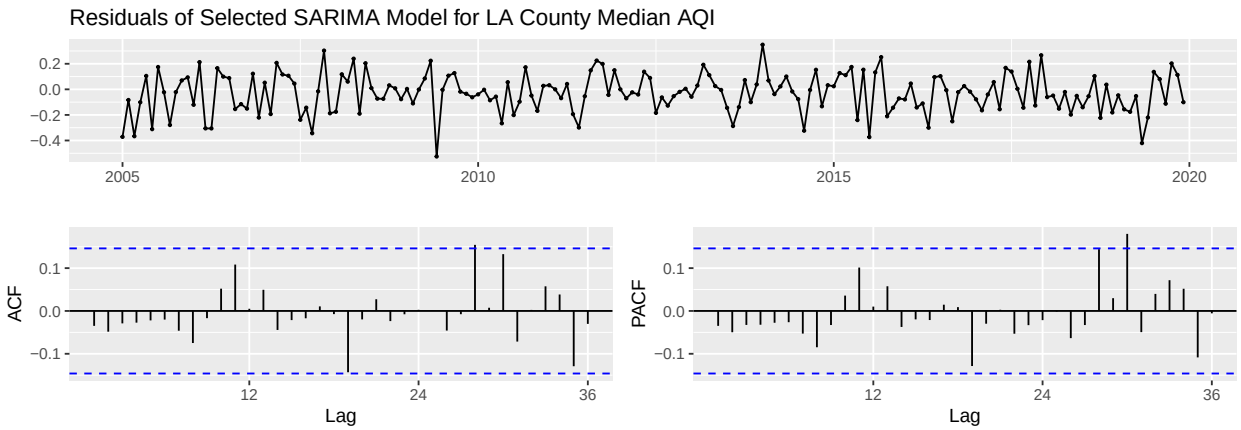


To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.6288, so the null hypothesis is not rejected and the residuals of the model are white noise.

The `auto.arima` model on the transformed and differenced Los Angeles County AQI data returns a SARMA$(4,0,1)\times(2,0,1)_{12}$ model with an AIC of -124.36 and BIC of -95.62:

$$(1-0.3925B+0.0321B^2-0.0327B^3-0.0567B^4)(1-0.1213B^{12}+0.0591B^{24})Y_t = (1-0.2989B)(1-0.8090B^{12})\epsilon_t$$

The residuals of this model is shown below.



To see if the ϵ_t component resembles white noise, the Ljung-Box test can be performed on the residuals of the model with the same test statistic and hypotheses as earlier. The test returns a p-value of 0.6379, greater than 0.05, indicating that the residuals of this model is white noise.

Forecasting

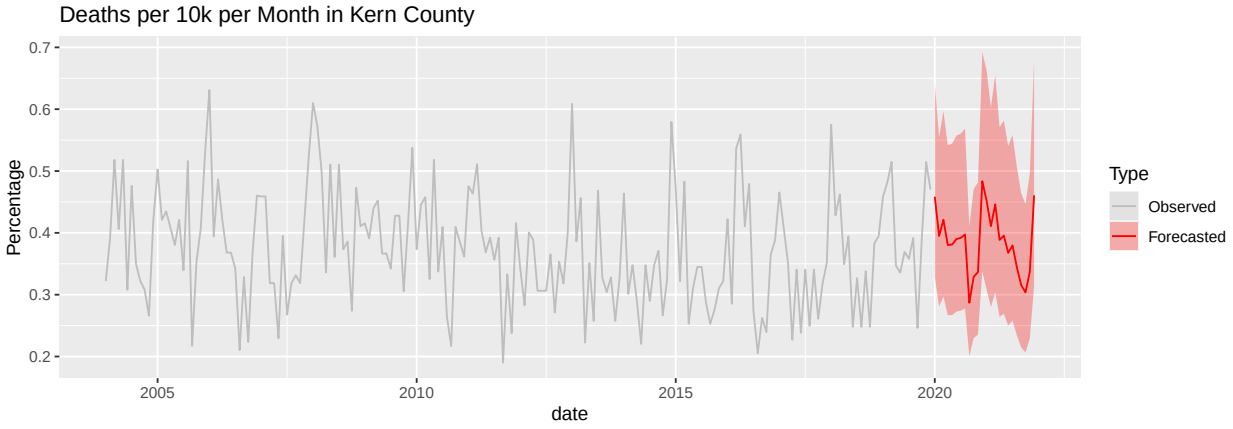
Forecasting Without Exogenous Variables

The `auto.arima` selected models perform the best in predicting the residuals of the transformed deaths per month data without exogenous variables. Using these models, forecasts for future months can be calculated under the model

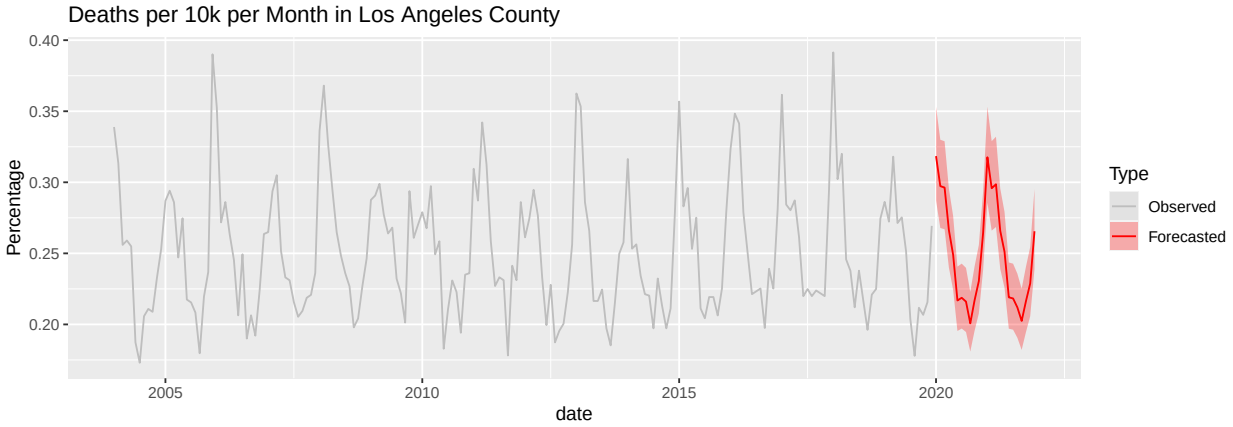
$$\hat{Y}_t = \exp\left\{\hat{T}_t + \hat{S}_t + \hat{R}_t\right\}$$

where $\hat{Y}_t$ is the predicted number of deaths at time t per 10,000 people, $\hat{T}_t$ is the predicted value of the trend component at time t, $\hat{S}_t$ is the predicted value of the seasonal component at time t, and $\hat{R}_t$ is the predicted value of the residuals at time t, estimated using the models selected.

The trend and seasonal component for the data can be estimated trivially using the past values of the trend component and seasonal component extracted with the SEATS method. Using the SARMA$(1,3) \times (0,2)_{12}$ model to predict the residuals for Kern County, the final forecasts for the next two years with 95% confidence intervals looks as follows.



The trend and seasonal component for the data can be estimated trivially once again for the Los Angeles data. Using the SARMA$(0,1) \times (0,1)_{12}$ model to predict the residuals for Los Angeles County, the final forecasts for the next two years with 95% confidence intervals looks as follows.



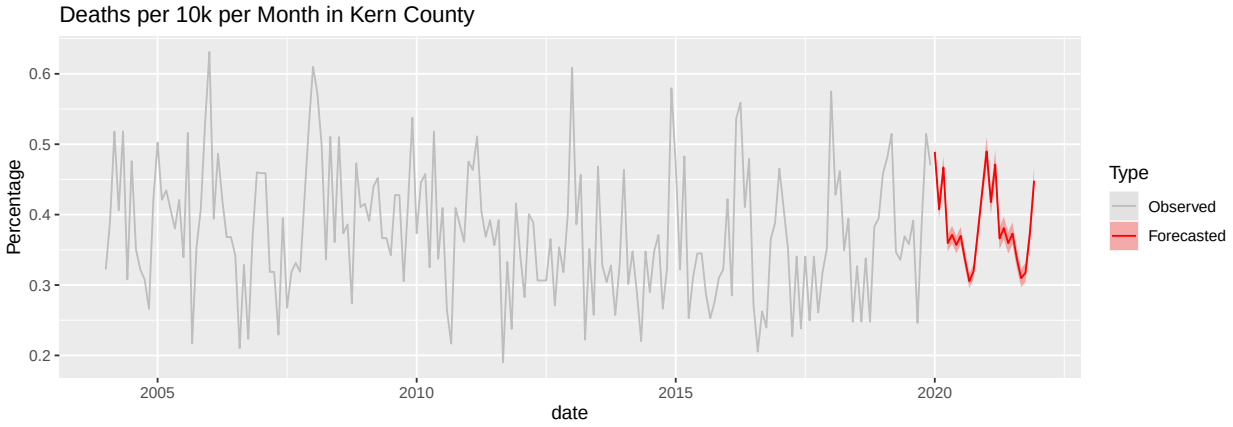
Forecasting with Exogenous Variables

Using the best regression models above, forecasts for the number of deaths per month can be found under the model

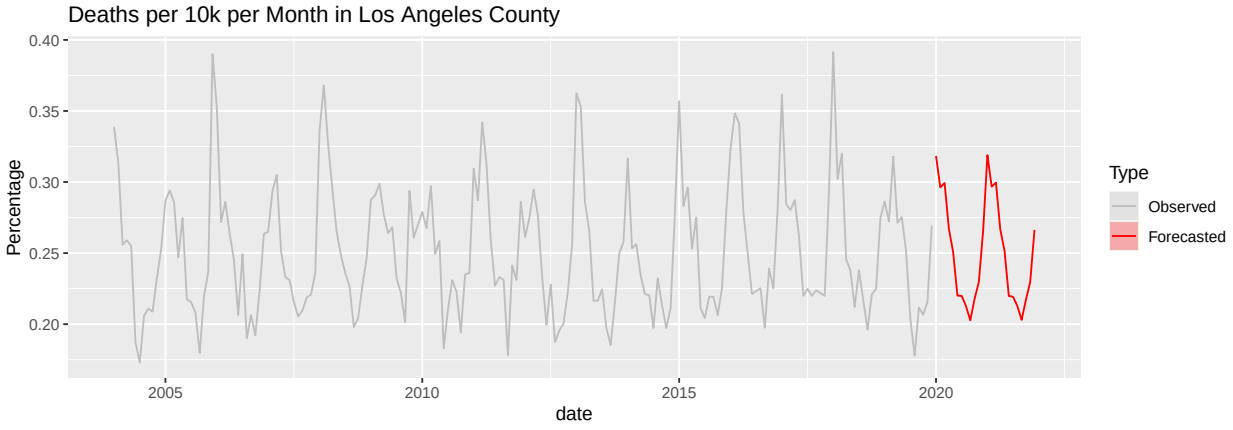
$$\hat{Y}_t = \exp\left\{\hat{T}_t + \hat{S}_t + f(r_t)\right\}$$

where $\hat{Y}_t$ is the predicted number of deaths at time t per 10,000 people, $\hat{T}_t$ is the predicted value of the trend component at time t, $\hat{S}_t$ is the predicted value of the seasonal component at time t, and $f(r_t)$ is the predicted value of the residuals at time t, a function of the log-transformed differenced AQI data.

Just as before, the trend and seasonal component for the data can be estimated trivially. Using the ARDL$(0, 0)$ model to predict the residuals for Kern County, the final forecasts for the next two years with 95% confidence intervals looks as follows.



The trend and seasonal component for the data can be estimated trivially again for the Los Angeles data. Using the ARDL$(0, 1)$ model to predict the residuals for Los Angeles County, the final forecasts for the next two years with 95% confidence intervals, though extremely tiny, looks as follows.

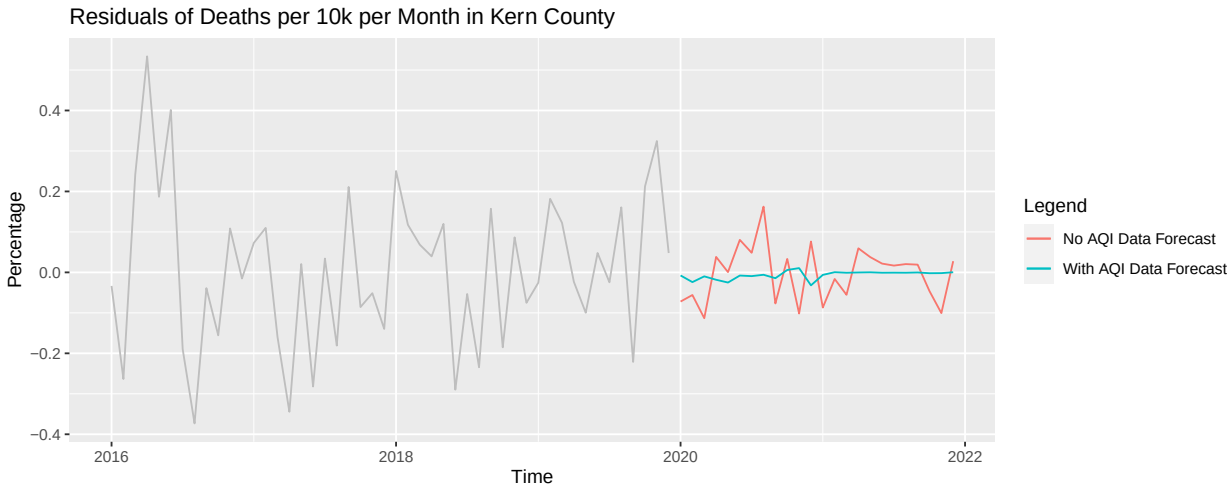


Discussion

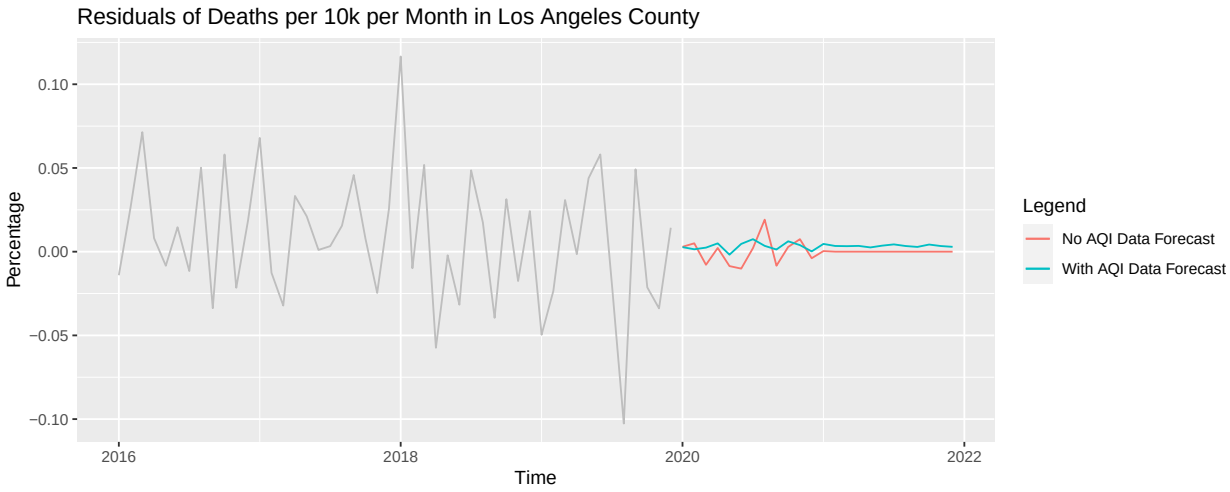
Both forecasting without AQI as an exogenous variable and forecasting with AQI as an exogenous variable create predictions that look uniformly periodic. There is no variation in the peaks as seen in the observed past data, most likely due to naive estimations of the trend component and seasonality. The 95% confidence intervals on the Kern County data are larger than the 95% confidence intervals on the Los Angeles County data, indicating that the models are more confident in estimating the percentage of people who die of chronic lower respiratory disease per 10,000 people in Los Angeles County than the models that estimate the same statistic in Kern County.

Both models for the Kern County data on average forecast a maximum of around 4,900 deaths per 10,000 people in the December to January and a minimum of around 3,000 deaths per 10,000 people in August to September. Both models for the Los Angeles County data on average forecast a maximum of 3,200 deaths per 10,000 people in the winter and a minimum of 2,000 deaths per 10,000 people in September (Appendix **Table 4**).

The estimation of trend and seasonality for each county is the same for the forecasts performed with regression and the forecasts performed without regression as they are calculated with the same methods on the same data, so looking at the residuals of the model is enough to make comparisons. The values of the calculated residuals can be found in the Appendix (**Table 5**).



From the above plot of the residuals of both models on the Kern County data, the forecasted residuals for the two models do not have as much variation in range as the observed past residuals. The residuals made by including the AQI data as an exogenous variable are much flatter than the residuals made without exogenous variables. Based on the past residuals, where there exists many peaks and valleys, such a trend does not seem like a very good fit. Visually it appears that the forecasts made without AQI as an exogenous variable may be more reasonable to use.



Similarly for the Los Angeles County data, the forecasted residuals for both models do not have as much variation in range as the observed past residuals. The residuals made by including the AQI data as an exogenous variable are again much flatter than the residuals made without exogenous variables, although the residuals of the forecasts made without AQI data are not as varied in the Los Angeles Data compared to the Kern County data. However, by the second year the model without AQI data included starts to predict zeros for every value, while the model that includes AQI data still has slight variations in the residuals. Thus, though visually the former model forecasts residuals closer to the observations of residuals in the past in terms of behavior, this only holds true for a short forecast window. The latter model may be better for longer term forecasts.

The regression models chosen to forecast the amount of people who die of chronic lower respiratory disease per 10,000 people use lags of 0 and 1. This means only the AQI of the current month and possibly the current and past month are most significant in determining the number of people who die that month of respiratory disease. The present air quality is more likely impacts a person's health immediately to the point of possible death, rather than the effects of constant exposure to air pollutants building up over a longer stretch of time leading to death.

However, the forecasts made with the two types of models for each county create very similar forecasts. In the Kern County data, the forecasted values for each month have an average absolute difference of 0.02, while for the Los Angeles data this difference is 0.001. This might mean that air quality data is not a significant variable when analyzing deaths due to chronic lower respiratory disease. Given the multiple scientific studies performed by others on the effects of air quality on health, such a conclusion does not seem very likely.

Perhaps the lack of drastic difference in these two methods can be attributed to the transformations performed on the data. The SEATS decomposition was used on the county deaths data while simple differencing was performed on the AQI data; this possibly makes a difference when using AQI as a variable in the model. Or, perhaps the `auto.arima` selected model for the AQI data does not capture all the structure present in the data and thus makes less accurate forecasts of the AQI, which in turn leads to less accurate forecasts of the deaths per 10,000 people. The models with the lowest AIC and BIC may also be the best models when modeling the past data, but not the best data when forecasting future data. Additionally, all of the tests for stationarity and resemblance for white noise used the same two tests, the ADF test and the Ljung-Box test. There are other tests that test for the same things, such as the Box-Pierce test for independence. Using those tests may lead to different conclusions regarding the suitability of the dataset, what transformations to perform, or what models to fit than the one reached here.

Conclusion

Based on the study performed, the air quality in recent months are more significant in affecting the number of people who die in that month due to CLRD than a prolonged exposure. In Kern County, only the current

month is significant while in Los Angeles County, both the current month and past month are significant. However, the models do not appear to be particularly good at forecasting future values of deaths in the short term when compared to forecasts made without including AQI as an explanatory variable. Alternate transformations or model selection criteria may need to be performed to see if this holds true in all cases.

Appendix

Figure A

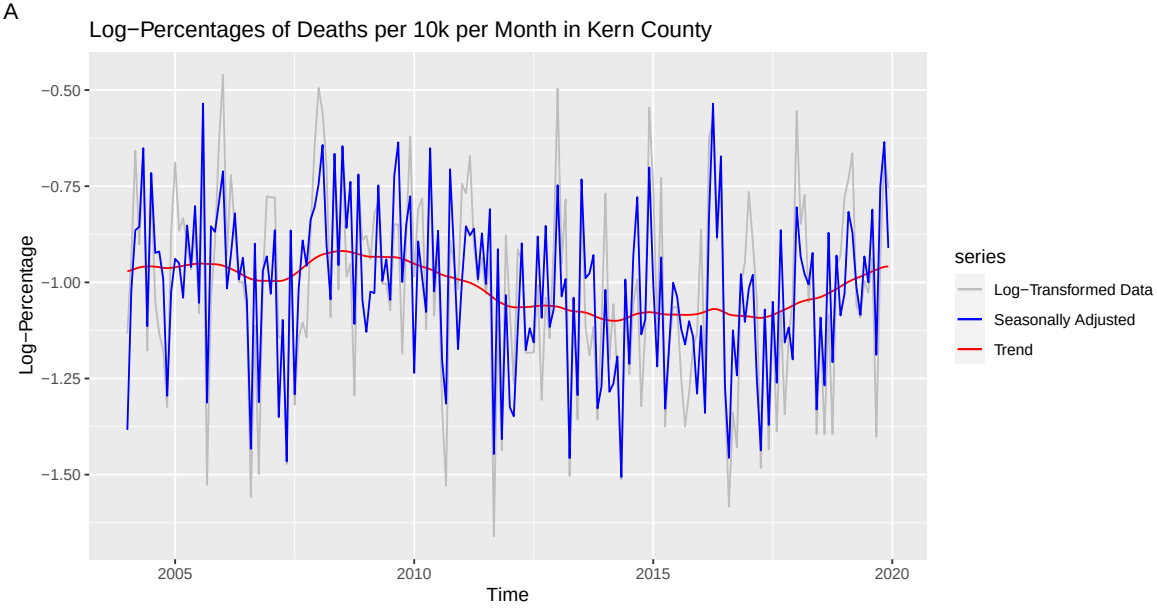


Figure B

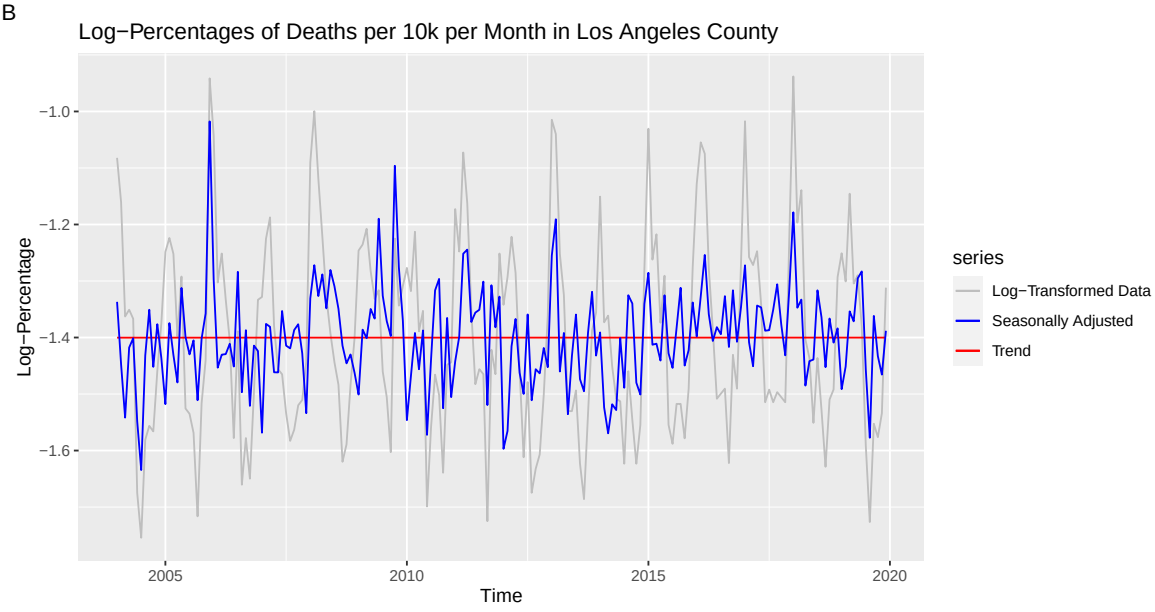


Table 1: Model Selection Criteria for Kern County Deaths Data Regressed with Lags of AQI Data

lag	AIC	BIC	White Noise Reisudals?
0	-90.56477	-80.985903	TRUE
1	-90.36787	-77.618328	TRUE
2	-90.37256	-74.463640	TRUE
3	-87.27151	-68.214616	TRUE
4	-84.18250	-61.989116	TRUE
5	-81.68900	-56.370711	TRUE
6	-78.59905	-50.167557	TRUE
7	-75.60810	-44.075179	TRUE
8	-75.61689	-40.994453	TRUE
9	-76.47339	-38.773427	TRUE
10	-73.21628	-32.450898	TRUE
11	-70.00597	-26.187389	TRUE
12	-68.68006	-21.820597	TRUE
13	-67.93147	-18.043571	TRUE
14	-67.80416	-14.900371	TRUE
15	-65.27487	-9.367850	TRUE
16	-61.88790	-2.990437	TRUE
17	-59.57481	2.300196	TRUE
18	-57.14973	7.689789	TRUE
19	-57.80189	9.989004	TRUE
20	-58.15135	12.577651	TRUE
21	-55.57422	18.079480	TRUE
22	-54.59119	21.973684	TRUE
23	-55.57878	23.883615	TRUE
24	-53.41553	28.930579	TRUE
25	-51.18124	34.034661	TRUE
26	-48.57927	39.492353	TRUE
27	-54.06634	36.846800	TRUE
28	-54.58538	39.154920	TRUE
29	-58.40242	38.150539	TRUE
30	-56.66350	42.687463	TRUE
31	-56.38268	45.751492	TRUE
32	-56.68156	48.220871	TRUE
33	-54.05221	53.603362	TRUE
34	-54.30862	56.084825	TRUE
35	-52.66774	60.448144	TRUE
36	-52.50198	63.320738	TRUE

Table 2: Model Selection Criteria for Los Angeles County Deaths Data Regressed with Lags of AQI Data

lag	AIC	BIC	White Noise Reisudals?
0	-532.6165	-523.0376	FALSE
1	-529.3588	-516.6093	TRUE
2	-524.0041	-508.0952	TRUE
3	-518.7638	-499.7069	TRUE
4	-528.0738	-505.8804	TRUE
5	-523.2499	-497.9316	TRUE
6	-519.1780	-490.7465	TRUE
7	-515.0452	-483.5123	TRUE
8	-508.9850	-474.3625	TRUE
9	-506.4400	-468.7401	TRUE
10	-501.0088	-460.2434	TRUE
11	-502.3036	-458.4850	TRUE
12	-538.2696	-491.4101	FALSE
13	-534.9645	-485.0766	FALSE
14	-535.2782	-482.3744	FALSE
15	-528.9221	-473.0151	FALSE
16	-524.4787	-465.5812	FALSE
17	-519.3660	-457.4910	FALSE
18	-512.9731	-448.1336	FALSE
19	-510.9249	-443.1340	FALSE
20	-511.5495	-440.8205	TRUE
21	-509.2517	-435.5980	TRUE
22	-505.7463	-429.1814	TRUE
23	-500.5735	-421.1111	TRUE
24	-494.4965	-412.1504	TRUE
25	-498.4802	-413.2643	FALSE
26	-492.0372	-403.9656	FALSE
27	-487.6892	-396.7761	TRUE
28	-489.1606	-395.4203	FALSE
29	-483.7348	-387.1819	FALSE
30	-477.9072	-378.5562	TRUE
31	-476.1957	-374.0616	TRUE
32	-470.0508	-365.1484	TRUE
33	-463.4111	-355.7555	TRUE
34	-457.3615	-346.9680	TRUE
35	-453.0075	-339.8917	TRUE
36	-453.2682	-337.4455	FALSE

Table 3: Comparison of Forecasts from Regression Models

	Kern County				Los Angeles County			
Date	Regression Forecasted Deaths	AQI	Deaths Difference from Last Month	AQI Difference from Last Month	Regression Forecasted Deaths	AQI	Deaths Difference from Last Month	AQI Difference from Last Month
Jan 2020	0.489081	65.14189	NA	NA	0.318373	66	NA	NA
Feb 2020	0.407635	56.306695	-0.081445	-8.835194	0.296226	53.952259	-0.022147	-12.047741
Mar 2020	0.467002	51.395922	0.059367	-4.910773	0.299343	58.384002	0.003117	4.431743
Apr 2020	0.359539	65.322076	-0.107463	13.926154	0.267161	76.782596	-0.032182	18.398594
May 2020	0.371367	67.955272	0.011828	2.633196	0.250583	53.071797	-0.016578	-23.710799
Jun 2020	0.356975	103.422385	-0.014393	35.467113	0.220081	126.397371	-0.030502	73.325574
Jul 2020	0.369859	107.564656	0.012885	4.142271	0.219908	158.999925	-0.000173	32.602555
Aug 2020	0.335842	114.691358	-0.034017	7.126701	0.212691	121.642711	-0.007217	-37.357215
Sep 2020	0.305354	99.50631	-0.030488	-15.185048	0.202717	80.81051	-0.009973	-40.832201
Oct 2020	0.320004	79.380171	0.01465	-20.126139	0.21766	86.982214	0.014942	6.171705
Nov 2020	0.376463	81.278455	0.056459	1.898284	0.22972	70.501778	0.01206	-16.480437
Dec 2020	0.433909	75.78227	0.057446	-5.496185	0.265603	56.673927	0.035883	-13.827851
Jan 2021	0.489778	69.441626	0.055869	-6.340644	0.318983	72.534043	0.05338	15.860116
Feb 2021	0.417748	55.107478	-0.072031	-14.334147	0.296812	54.844588	-0.022171	-17.689455
Mar 2021	0.471195	51.198364	0.053447	-3.909115	0.299605	59.21512	0.002792	4.370532
Apr 2021	0.366074	64.493009	-0.105121	13.294646	0.266755	78.199806	-0.03285	18.984685
May 2021	0.380894	66.714846	0.01482	2.221837	0.251669	53.813515	-0.015086	-24.386291
Jun 2021	0.359467	102.966515	-0.021426	36.251669	0.219849	133.906432	-0.031819	80.092916
Jul 2021	0.373047	106.83208	0.01358	3.865565	0.219224	165.412831	-0.000626	31.5064
Aug 2021	0.337544	114.140598	-0.035503	7.308518	0.212671	119.795916	-0.006553	-45.616916
Sep 2021	0.309719	98.237192	-0.027826	-15.903407	0.203031	79.927924	-0.00964	-39.867991
Oct 2021	0.317462	80.237617	0.007744	-17.999574	0.217237	88.84151	0.014206	8.913586
Nov 2021	0.37183	81.954488	0.054368	1.716871	0.229592	68.684451	0.012355	-20.157059
Dec 2021	0.448199	74.278888	0.076368	-7.6756	0.266327	55.592516	0.036735	-13.091935

Table 4: Comparison of Forecasts from Forecasting Models

	Kern County			Los Angeles County		
Date	Without AQI Data	With AQI Data	Absolute Difference	Without AQI Data	With AQI Data	Absolute Difference
Jan 2020	0.458556	0.489081	0.030525	0.318436	0.318373	6.3e-05
Feb 2020	0.394843	0.407635	0.012792	0.297284	0.296226	0.001059
Mar 2020	0.421142	0.467002	0.04586	0.29629	0.299343	0.003053
Apr 2020	0.380408	0.359539	0.020869	0.266427	0.267161	0.000734
May 2020	0.381061	0.371367	0.009693	0.248879	0.250583	0.001704
Jun 2020	0.389873	0.356975	0.032898	0.216853	0.220081	0.003227
Jul 2020	0.391821	0.369859	0.021962	0.218736	0.219908	0.001172
Aug 2020	0.397387	0.335842	0.061545	0.216045	0.212691	0.003354
Sep 2020	0.286897	0.305354	0.018458	0.200766	0.202717	0.001952
Oct 2020	0.32885	0.320004	0.008846	0.216927	0.21766	0.000733
Nov 2020	0.336544	0.376463	0.039919	0.230536	0.22972	0.000817
Dec 2020	0.483537	0.433909	0.049628	0.264519	0.265603	0.001084
Jan 2021	0.45183	0.489778	0.037949	0.317622	0.318983	0.001361
Feb 2021	0.410813	0.417748	0.006934	0.295801	0.296812	0.001012
Mar 2021	0.446211	0.471195	0.024983	0.298618	0.299605	0.000987
Apr 2021	0.388575	0.366074	0.022502	0.265836	0.266755	0.000919
May 2021	0.395466	0.380894	0.014573	0.25103	0.251669	0.000639
Jun 2021	0.367663	0.359467	0.008196	0.21906	0.219849	0.000789
Jul 2021	0.37962	0.373047	0.006572	0.218272	0.219224	0.000951
Aug 2021	0.344814	0.337544	0.007269	0.211953	0.212671	0.000718
Sep 2021	0.315757	0.309719	0.006038	0.202455	0.203031	0.000575
Oct 2021	0.303713	0.317462	0.013749	0.216314	0.217237	0.000923
Nov 2021	0.336857	0.37183	0.034974	0.228821	0.229592	0.000771
Dec 2021	0.460609	0.448199	0.01241	0.265559	0.266327	0.000768
Average			0.022881			0.001224

Table 5: Comparison of Residuals from Forecasting Models

	Kern County			Los Angeles County		
Date	Without AQI Data	With AQI Data	Absolute Difference	Without AQI Data	With AQI Data	Absolute Difference
Jan 2020	-0.072053	-0.007608	0.064445	0.002914	0.002716	0.000198
Feb 2020	-0.055917	-0.024033	0.031884	0.005002	0.001435	0.003567
Mar 2020	-0.113205	-0.009842	0.103363	-0.007825	0.002424	0.01025
Apr 2020	0.0382	-0.018221	0.056421	0.002219	0.004972	0.002753
May 2020	0.000668	-0.025098	0.025767	-0.008604	-0.001782	0.006822
Jun 2020	0.080336	-0.00782	0.088156	-0.010125	0.004647	0.014772
Jul 2020	0.048428	-0.009255	0.057683	0.002122	0.007467	0.005344
Aug 2020	0.162384	-0.005886	0.16827	0.019122	0.003475	0.015647
Sep 2020	-0.076747	-0.014397	0.06235	-0.008381	0.001294	0.009675
Oct 2020	0.033202	0.005933	0.027269	0.00283	0.006203	0.003373
Nov 2020	-0.101559	0.010532	0.112091	0.007468	0.00392	0.003549
Dec 2020	0.076247	-0.032046	0.108293	-0.003924	0.000168	0.004091
Jan 2021	-0.08683	-0.006182	0.080648	0.000355	0.00463	0.004276
Feb 2021	-0.016267	0.000472	0.016739	0	0.003414	0.003414
Mar 2021	-0.055383	-0.000905	0.054478	0	0.003298	0.003298
Apr 2021	0.059443	-0.00021	0.059653	0	0.003449	0.003449
May 2021	0.037776	0.00023	0.037546	0	0.002542	0.002542
Jun 2021	0.021683	-0.000861	0.022544	0	0.003597	0.003597
Jul 2021	0.016792	-0.000672	0.017465	0	0.004349	0.004349
Aug 2021	0.020478	-0.00083	0.021308	0	0.003381	0.003381
Sep 2021	0.019103	-0.000205	0.019308	0	0.002838	0.002838
Oct 2021	-0.046317	-0.002041	0.044276	0	0.004259	0.004259
Nov 2021	-0.10063	-0.00185	0.098781	0	0.003365	0.003365
Dec 2021	0.027669	0.000356	0.027313	0	0.00289	0.00289
Average			0.058585			0.005071

References

Perez, N. (2018) *Despite decades of cleanup, respiratory disease deaths plague California county.* Available at: https://www.ehn.org/chronic-respiratory-disease-california-2621765230.html

Wang, T., Zhao, B., Liou, K.N., *et al.*. (2019) 'Mortality burdens in California due to air pollution attributable to local and nonlocal emissions', *Environment International*, 133 (Part B). Available at: https://doi.org/10.1016/j.envint.2019.105232

World Health Organization. (2018) *9 out of 10 people worldwide breathe polluted air, but more countries are taking action.* Available at: https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action