AIR WATCH

PREDICTIVE
FINAL PROJECT

Group Members

Akanksha Arora
Pranjal Totala
Suranjana Chowdhury
Thao Dinh

TABLE OF CONTENT

OBJECTIVE 3

MEMBER CONTRIBUTION 3

1 TECHNICAL WORK 4

1.1 AIR QUALITY DATA 4
1.1.1 DATA COLLECTION: 4
1.1.2 ASSUMPTIONS: 4
1.1.3 DATA PREPARATION: 5
1.2 DATA PRE-PROCESSING 5
1.2.1 TIME SERIES DECOMPOSITION 5
1.2.2 STATIONARY TEST 5
1.2.3 ACF AND PACF PLOTS 6
1.2.4 TEST-TRAIN SPLIT 7
1.3 MODELLING 7
1.3.1 SIMPLE SMOOTHENING 7
1.3.2 HOLT WINTERS' ADDITIVE METHOD 7
1.3.3 ETS 7
1.3.4 ARIMA 7

2 EVALUATION 8

2.1 METHODOLOGY 8
2.2 TOP 5 CITIES 8
2.3 BOTTOM 5 CITIES 9

3 CONCLUSION 11

3.1 ALERT SYSTEM 11
3.2 ALERT: TOP 5 CITIES 11
3.3 ALERT: BOTTOM 5 CITIES 13

4 APPENDIX 15

4.1 BAKERSFIELD, CA: STATIONARY TS 15
4.2 FRESNO, CA: STATIONARY TS 16
4.3 VISALIA-PORTERVILLE, CA: STATIONARY TS 16
4.4 SAN FRANCISCO-OAKLAND-HAYWARD, CA 17
4.5 LOS ANGELES-LONG BEACH-ANAHEIM, CA 18
4.6 CHEYENNE, WY 19
4.7 WILMINGTON, NC 19
4.8 URBAN HONOLULU, HI 20
4.9 KAHULUI-WAILUKU-LAHAINA, HI 21
4.10 BANGOR, ME 22

Objective

Problem Statement: To predict the air quality index (PM2.5) of the top 5 most polluted cities and top 5 least polluted cities in USA and create an air quality alert when the PM2.5 prediction goes over safety range.

Air pollution is a major concern in many cities and predicting PM2.5 levels is crucial for protecting public health and making informed policy decisions. Predicting PM2.5 (particulate matter with a diameter of 2.5 micrometres or less) in air is important not for just protecting public health and preserving the environment, but also making informed policy decisions related to air quality management. The alert helps notify about the conditions that could be harmful people sensitive to air pollutions – elders, children, people with lung or heart disease. Additionally, it would help either government or involved parties have a strategic plan to prevent air pollution and improve the air quality.

Member Contribution

1. **Data Collection:** Each member downloaded data for 2-3 cities.
   - Akanksha Arora: Top 3 **Most** Polluted
   - Suranjana Chowdhury: Bottom 2 **Most** Polluted
   - Thao Dinh: Top 3 **Least** Polluted
   - Pranjal Totala: Bottom 2 **Least** Polluted
2. **Data Pre-processing:** One person appended all the datasets, one person performed data sanity checks; one person performed aggregation and last person reviewed everything
3. **Stationary Check & Model Building:** Cities were divided as follows, and each person performed stationary check, built models, and forecasted for future values for their share of cities -
   - Akanksha: Bakersfield, Fresno
   - Suranjana: Visalia, San Jose, Los Angeles
   - Thao Dinh: Cheyenne, Wilmington
   - Pranjal: Urban Honolulu, Kahului, Bangor
4. **Model Conclusion, Report Building and PowerPoint creation:** We create a shared document so that everybody can contribute equally to PowerPoint and report creation

1 Technical Work

1.1 Air Quality Data

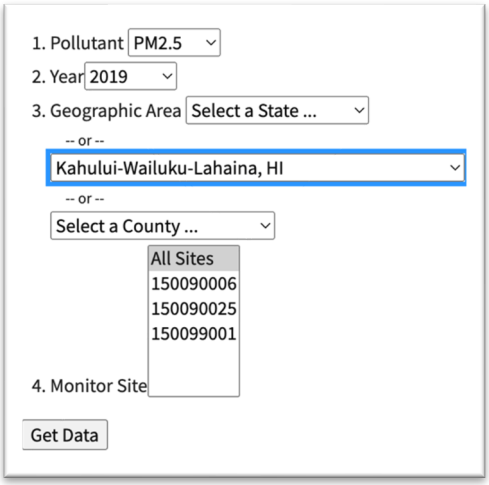
1.1.1 Data Collection:

We collected the data for top 5 most polluted and least polluted cities for 10 years from 2013 to 2022.
**Source used to identify top 5 most and top 5 least polluted cities -> American Lung Association*

Top 5 most polluted cities -
1. Bakersfield, California
2. Fresno-Madera-Hanford, California
3. Visalia, California
4. San Jose-San Francisco-Oakland, California
5. Los Angeles-Long Beach, California

Top 5 least polluted cities –
1. Cheyenne, WY
2. Wilmington, NC
3. Urban Honolulu, HI
4. Kahului-Wailuku-Lahaina, HI
5. Bangor, ME

The datasets for 10 cities for years 2013-2022 were first downloaded from US EPA website (US Environment Protection Agency). These datasets were collated for all cities and for all years to create one unified dataset for analysis. The website gave us PM2.5 value collected by different sites in these 10 cities. Each site collected data either daily, alternate or may be weekly.



1.1.2 Assumptions:

- The air quality index (PM2.5) does not usually change much over time (stationary)
- There are multiple sites across a city which report PM2.5 concentration. Daily PM 2.5 is assumed to be average of same day across multiple devices/sites
- Some dates are missing for some cities. To treat missing PM 2.5 values, moving average method is used. We have considered the last 2 and next 2 PM2.5 to calculate for the missing one.

1.1.3 Data Preparation:

Each site collected data either daily, alternate or may be weekly. We had some missing PM2.5 data for few dates and had some missing dates in our data.

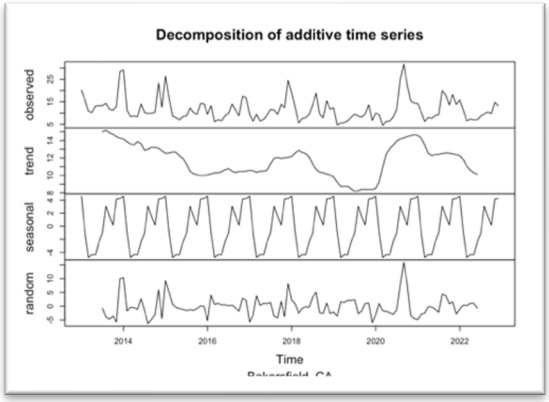
1. Missing PM2.5 data – We took the average PM2.5 value for a particular day from all sites. This was done to also make sure the data correctly represents the entire city and not just a part of the city.
2. Missing Dates – We first added the dates that were missing for our data and then imputed the values using the average of previous 2 and after 2 PM2.5 values.

Post this, we collated the data for all cities together in one dataset. Due to some model limitations in R, we summarized data at monthly level.

1.2 Data Pre-Processing

1.2.1 Time Series Decomposition

Visualizing the timeseries data helps to observe various trend & seasonality in the data. We plotted the time series for all 10 cities and looked at their decomposed graph to see trend and seasonality. From the graphs we could see some seasonality and trend in the time series for few cities.



1.2.2 Stationary Test

Looking at a few graphs we can easily see trends and seasonality in some cities. A stationary time series is the one in which mean, variance, autocovariance and autocorrelation remain constant with time. It is important for a time series to be stationary as sometimes non-stationary series lead to biased estimates and unreliable predictions. There are various tests that can be performed to check if a time series is stationary or not-

1. ADF Test:
   Null Hypothesis (HO) - time series is nonstationary
   If p-value < 0.05 - Reject Null Hypothesis and time series would be stationary
2. KPSS Test:
   Null Hypothesis (HO) - time series is trend stationary
   If p-value < 0.05 - Reject Null Hypothesis and time series is non trend stationary

We used ADF and KPSS test to take decision on if the given time series is stationary or not.

1. Bakersfield, CA: Stationary

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.0924, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.14389, Truncation lag parameter = 4, p-value = 0.1
```

2. Wilmington, NC: Non-Stationary

```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -2.8944, Lag order = 4, p-value = 0.2051
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.8084, Truncation lag parameter = 4, p-value = 0.01
```

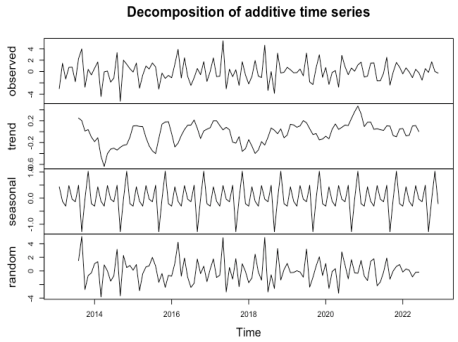
All other cities ADF and KPSS test result screenshot are in Appendix.

We ran ADF, KPSS test for all 10 cities. We found that 6 out of 10 cities were stationary and 4 were non-stationary.

- PM2.5 from Urban Honolulu, Kahului-Wailuku-Lahaina and Bangor are non-stationary with ADF p-value < 0.05 but KPSS p-value is < 0.05.
- PM2.5 from Wilmington is non-stationary with ADF p-value > 0.05 and KPSS p-value > 0.05

To convert these cities into stationary series we do detrend by differencing. Detrend by differencing helps to remove linear or non-linear trend from the time series by taking the first difference i.e., subtracting the value of each observation from its previous observation of the time series.

Below are the decomposed graphs for the Wilmington city after differencing them and converting them into stationary Time Series.



Wilmington

1.2.3 ACF and PACF Plots

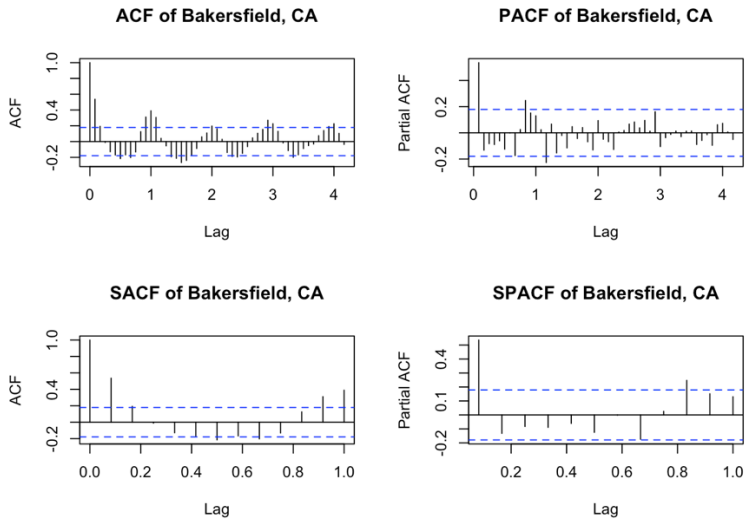
ACF and PACF plots helps to analyse the correlation structure of the time series.

ACF Plot: The ACF function calculates the similarity (correlation) of the values of a time series with its own lagged values. If the data has trend, ACF plot of small lags tends to be large and positive. When the data is seasonal, correlations will be large for seasonal lags than for other lags.

PACF Plot: The Partial ACF function measures the strength of the relationship between 2 variables, while controlling the effect of another variable. It helps identify the presence of a trend and/or cyclic behaviour in the time series.

Seasonal ACF Plot: The SACF is a variation of the ACF that is used when the time series exhibits seasonality. The SACF plot can help identify the presence of seasonal patterns in the time series.

Seasonal PACF Plot: The SPACF is a variation of the PACF that is used when the time series exhibits seasonality. The SPACF plot can help identify the presence of seasonal patterns in the time series, while controlling for the effects of the intervening lags.



Bakersfield, CA

Above is the ACF, PACF, SACF and SPACF plots for Bakersfield, CA. We see AR showing p=1 and MA showing P=1 with seasonal s=12.

For other cities and then their ACF-PACF plots, please refer to Appendix.

1.2.4 Test-Train Split

We split our dataset into train and test. We data for years 2013-2021 as train and year 2022 as test for each city.

1.3 Modelling

1.3.1 Simple Smoothening

Simple Exponential Smoothening takes weighted average of past observations to forecast future values. The weights decrease exponentially as the observations gets older.

1.3.2 Holt Winters' Additive Method

Holt Winters' Additive method uses level, trend, and seasonality to forecast the future values. Level means average value of time series.

1.3.3 ETS

ETS stands for Error, Trend and Seasonality. ETS uses error, trend, and seasonality to forecast the future values. Error helps to capture random fluctuations in data.

1.3.4 ARIMA

Autoregressive Integrated Moving Average models considers current observation and its past observations. The AR component captures the correlation between the current observation and its past observations, the differencing component removes the trend from the data, and the MA component captures the correlation between the current observation and the error terms from past observations

2 Evaluation

2.1 Methodology

To evaluate a time series forecasting model, AIC is one of the measures. AIC is used to evaluate the goodness of fit and complexity of a model. The lower the AIC score, the better the model is.

Another measure that we considered to evaluate our model is RMSE. RMSE measure the average magnitude of error between the actual and the predicted values. Again, lower the value, better is the model.

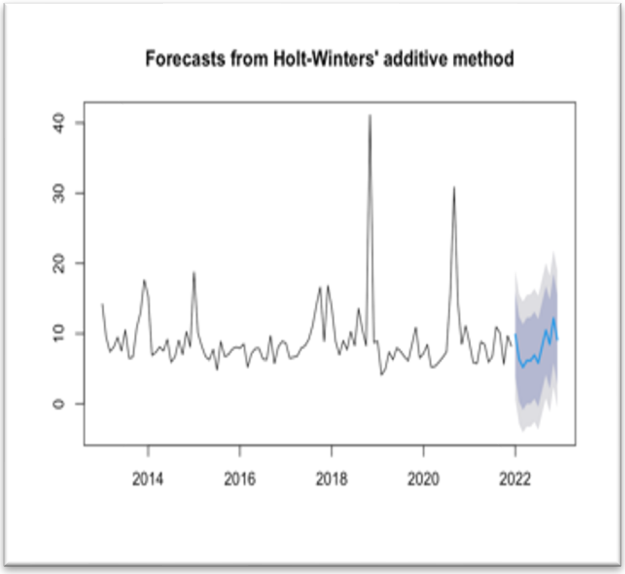
2.2 Top 5 Cities

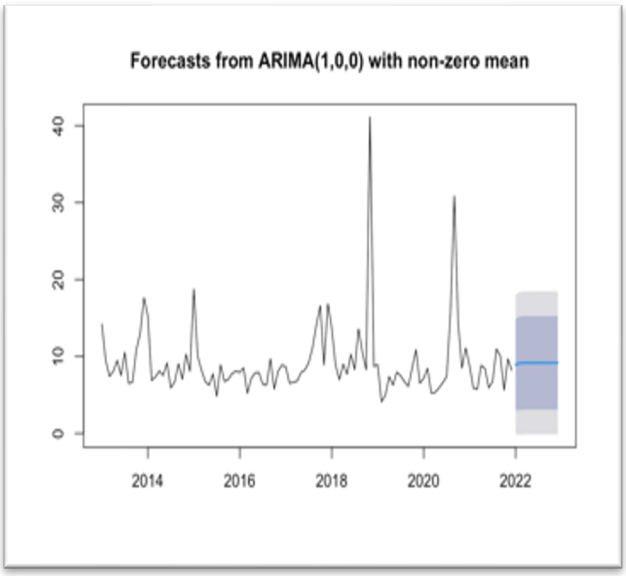
Below table looks at the RMSE and the AIC value for top 5 cities for the 4 models discussed in the previous section.

	Simple Exponential	Holt–Winters'	ETS Forecasting	ARIMA	Conclusion
Bakersfield, CA	AIC: 864.8442	AIC: 847.5029	AIC: 792.6041	AIC: 635.96	ARIMA
	RMSE: 5.1297	RMSE: 4.1584	RMSE: 4.3245	RMSE: 4.4055	
Fresno, CA	AIC: 929.6218	AIC: 914.9621	AIC: 823.7585	AIC: 700.52	ARIMA
	RMSE: 6.9237	RMSE: 5.6829	RMSE: 6.1566	RMSE: 5.8578	
Visalia–Porterville, CA	AIC: 950.8843	AIC: 955.795	AIC: 889.3072	AIC: 719.11	ARIMA
	RMSE: 7.6400	RMSE: 6.8654	RMSE: 6.9968	RMSE: 6.4268	
San Francisco–Oakland–Hayward, CA	AIC: 846.0338	AIC: 856.9399	AIC: 775.7197	AIC: 641.05	Holt–Winters'
	RMSE: 4.7019	RMSE: 4.3441	RMSE: 4.5102	RMSE: 4.5761	
Los Angeles–Long Beach–Anaheim, CA	AIC: 730.1852	AIC: 733.7015	AIC: 725.6362	AIC: 508.52	ARIMA
	RMSE: 2.7501	RMSE: 2.4554	RMSE: 2.7513	RMSE: 2.4477	

For cities Bakersfield, Fresno, Visalia & Los Angeles ARIMA models gives the lowest AIC and RMSE score. Thus, we go ahead with ARIMA.

But in case of San Francisco, when we looked at the forecasted value graph for ARIMA, we see a straight line that is not able to capture trend or seasonality. The forecasted values are almost same to that of constant. In this case we choose the model with the next best RMSE score that captures trend and seasonality both which is Holt Winters'.





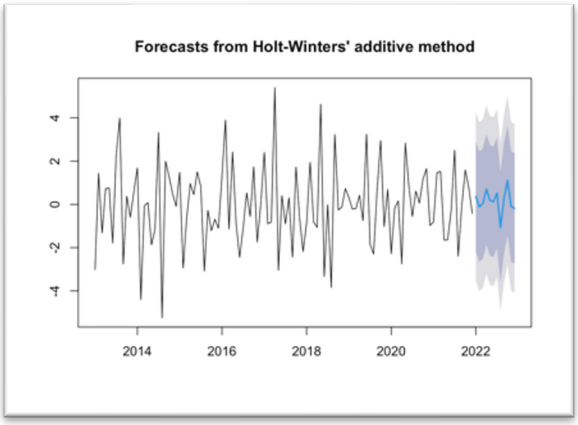
2.3 Bottom 5 Cities

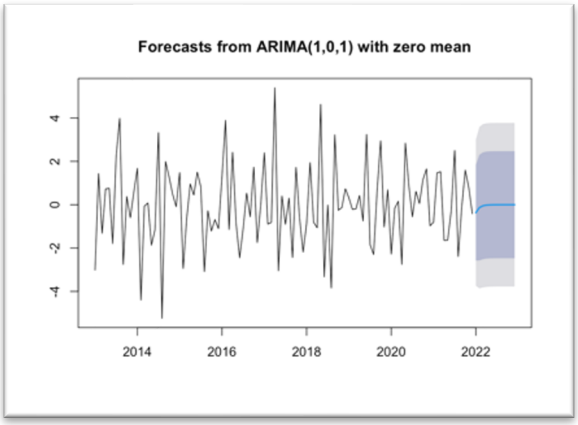
Below table looks at the RMSE and the AIC value for bottom 5 cities for the 4 models discussed in the previous section.

	Simple Exponential	Holt-Winters'	ETS Forecasting	ARIMA	Conclusion
Cheyenne, WY	AIC: 691.4068	AIC: 670.2716	AIC: 599.6235	AIC: 459.43	ARIMA
	RMSE: 2.2981	RMSE: 1.8305	RMSE: 1.9026	RMSE: 1.9329	
Wilmington, NC	AIC: 651.4828	AIC: 670.9731	AIC: 651.4828	AIC: 428.42	Holt-Winters'
	RMSE: 1.9103	RMSE: 1.8365	RMSE: 1.9103	RMSE: 1.7056	
Urban Honolulu, HI	AIC: 503.0748	AIC: 496.9938	AIC: 492.2524	AIC: 285.27	ARIMA
	RMSE: 0.9609	RMSE: 0.8207	RMSE: 0.8179	RMSE: 0.8665	
Kahului-Wailuku-Lahaina, HI	AIC: 517.5657	AIC: 524.7874	AIC: 517.5657	AIC: 289.71	ARIMA
	RMSE: 1.0276	RMSE: 0.9334	RMSE: 1.0276	RMSE: 0.8832	
Bangor, ME	AIC: 649.4932	AIC: 632.6641	AIC: 627.8886	AIC: 399.05	ARIMA
	RMSE: 1.8928	RMSE: 1.5380	RMSE: 1.5325	RMSE: 1.4261	

For cities Cheyenne, Honolulu, Kahului & Bangor ARIMA models gives the lowest AIC and RMSE score. Thus, we go ahead with ARIMA.

But in case of Wilmington, when we looked at the forecasted value graph for ARIMA, we see a straight line that is not able to capture trend or seasonality. The forecasted values are almost same to that of constant. In this case we choose the model with the next best RMSE score that captures trend and seasonality both which is Holt Winters'.


Forecasts from Holt-Winters' additive method
4
2
0
-2
-4
2014
2016
2018
2020
2022


Forecasts from ARIMA(1,0,1) with zero mean
4
2
0
-2
-4
2014
2016
2018
2020
2022

3 Conclusion

3.1 Alert System

After selecting the best model for each city, we plan to predict the PM2.5 levels for all 10 cities for the year 2023. For this purpose, we take the whole data for each 2013-2022 and run the best performing model. We predict the values for the year 2023 using the same model.

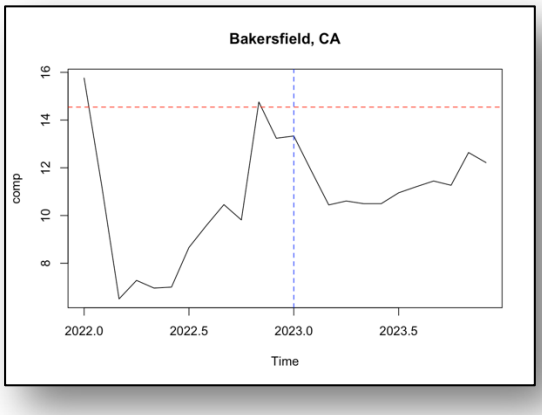
For every city if the PM2.5 level goes above the 95%ile line we send an alert to the city. It is important to note here that PM2.5 levels can vary widely depending on factors such as location, season, weather patterns. Thus, to monitor PM2.5 levels regularly we use the 95th percentile line instead of one standard line reference for all cities as a tool to take the decision.

The PM2.5 concentration ranges and corresponding AQI values for each category are as follows:

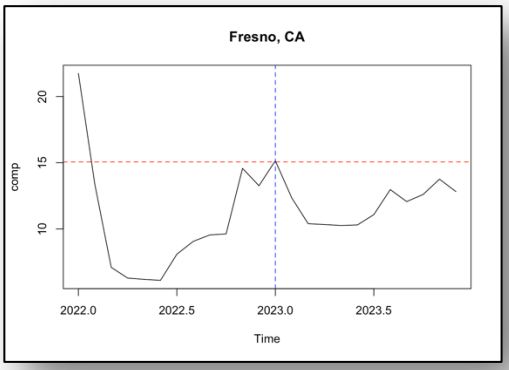
Good: 0-12.0 µg/m3
Moderate: 12.1-35.4 µg/m3
Unhealthy for sensitive groups: 35.5-55.4 µg/m3
Unhealthy: 55.5-150.4 µg/m3
Very unhealthy: 150.5-250.4 µg/m3
Hazardous: 250.5-500 µg/m3

3.2 Alert: Top 5 Cities

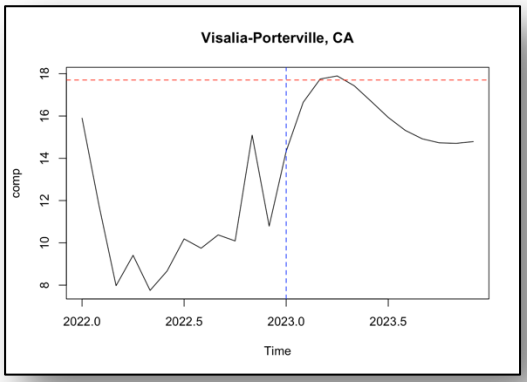
1. Bakersfield, CA: The graph below shows the 2023 predictions (after blue line) for the city. The red line shows the 95%ile line. If the values cross this threshold, then we send out the alert. In this case, we see that the PM2.5 levels first decrease and then increase slightly but never cross the threshold line. Thus, alert should **not** go out for this city. The AQI is always between the Good and Moderate range.



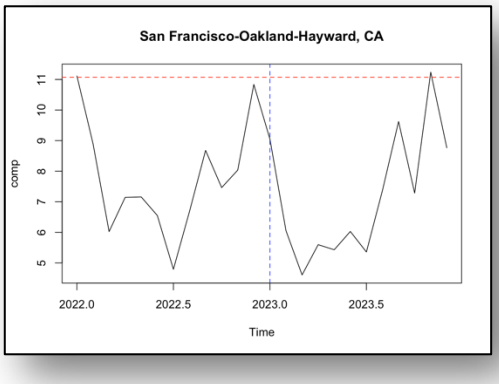
2. Fresno, CA: We see that the PM2.5 levels first decrease and then increase slightly in the later half of the year but never cross the threshold line. Thus, alert should **not** go out for this city. The AQI is always between the Good and Moderate range.



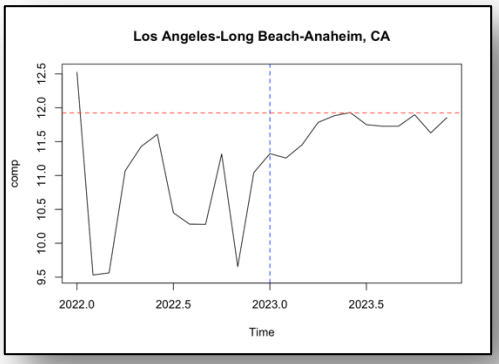
3. Visalia-Porterville, CA: We see that the PM2.5 levels first increases and then slightly decreases. We see that for March month PM2.5 level crosses the threshold range and thus, **alert should go out** for this city. The AQI is in Moderate range.



4. San Francisco-Oakland-Hayward, CA: We see sudden increase in the PM2.5 level for the month of November which crosses the threshold line. Although the AQI is in good range, values cross our threshold of 95%ile, thus **alert should go out**.

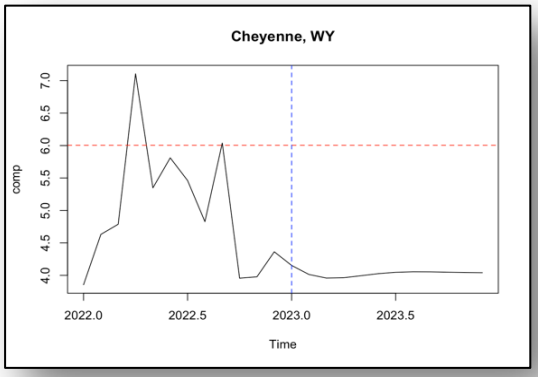


5. Los Angeles-Long Beach-Anaheim, CA: The PM2.5 values increase steadily for the city. It is at the danger level but has not yet crossed it. We do **not** send out an alert, but we should monitor the city with a close eye. The AQI value is in the Good Range.

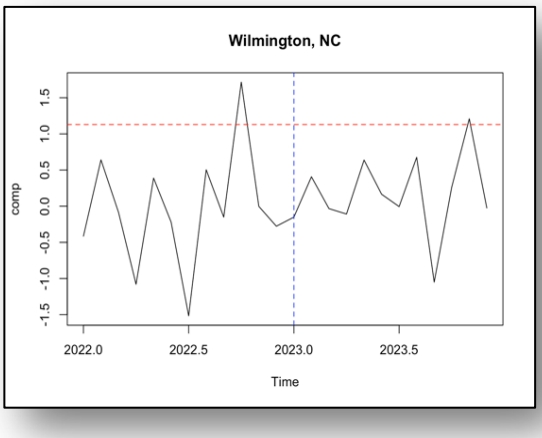


3.3 Alert: Bottom 5 Cities

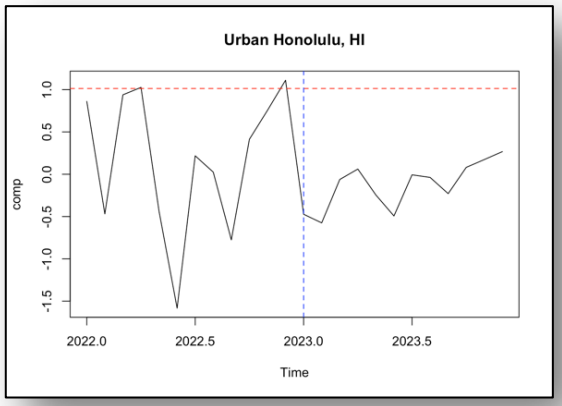
1. Cheyenne, WY: We see that the PM2.5 levels decrease and never cross the threshold line. Thus, alert should **not** go out for this city. The AQI is always Good.



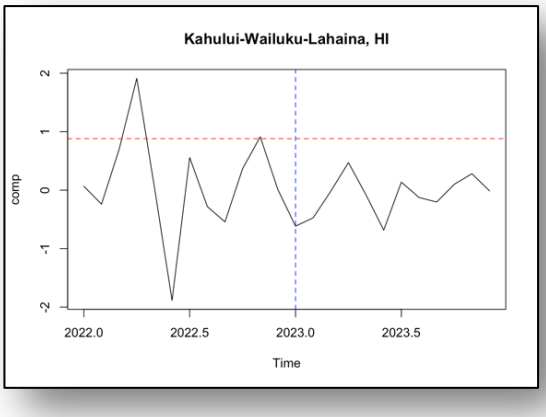
2. Wilmington, NC: We see that the PM2.5 levels crosses the threshold line for November 2023. Thus, **alert should go out** for this city. The AQI is always Good.



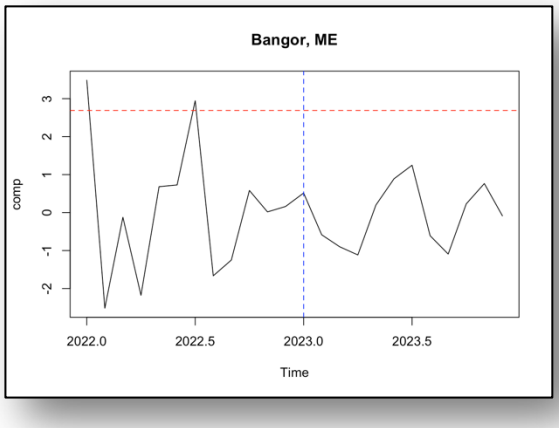
3. Urban Honolulu, HI: We see that the PM2.5 never crosses the threshold range and thus, alert should **not** go out for this city. The AQI is in good range.



4. Kahului-Wailuku-Lahaina, HI: We see that the PM2.5 never crosses the threshold range and thus, alert should **not** go out for this city. The AQI is in good range.

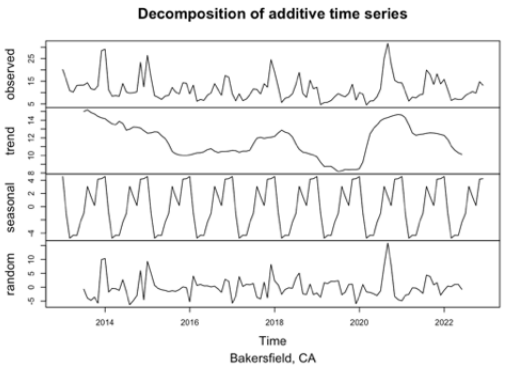


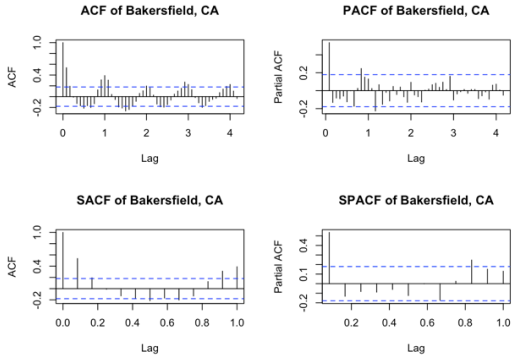
5. Bangor, ME: We see that the PM2.5 never crosses the threshold range and thus, alert should **not** go out for this city. The AQI is in good range.



4 Appendix

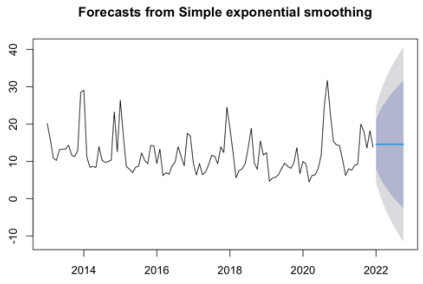
4.1 Bakersfield, CA: Stationary TS

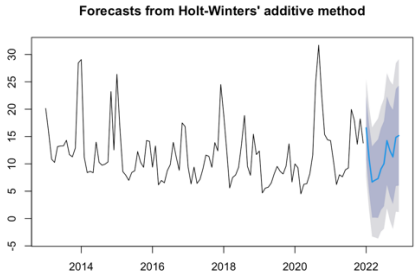


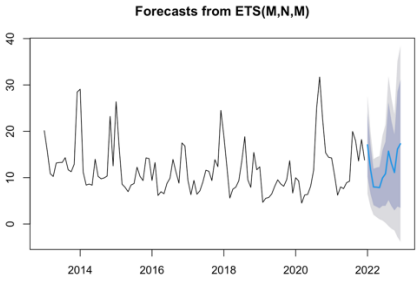


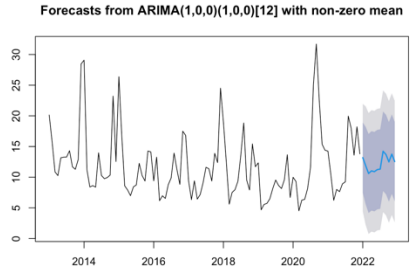

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.0924, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.14389, Truncation lag parameter = 4, p-value = 0.1
```

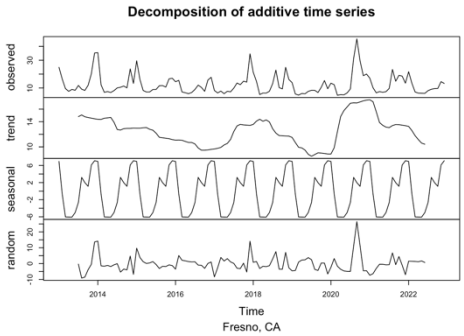


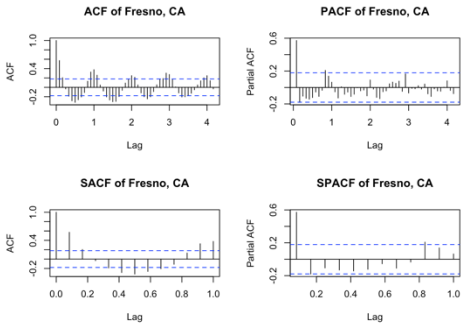






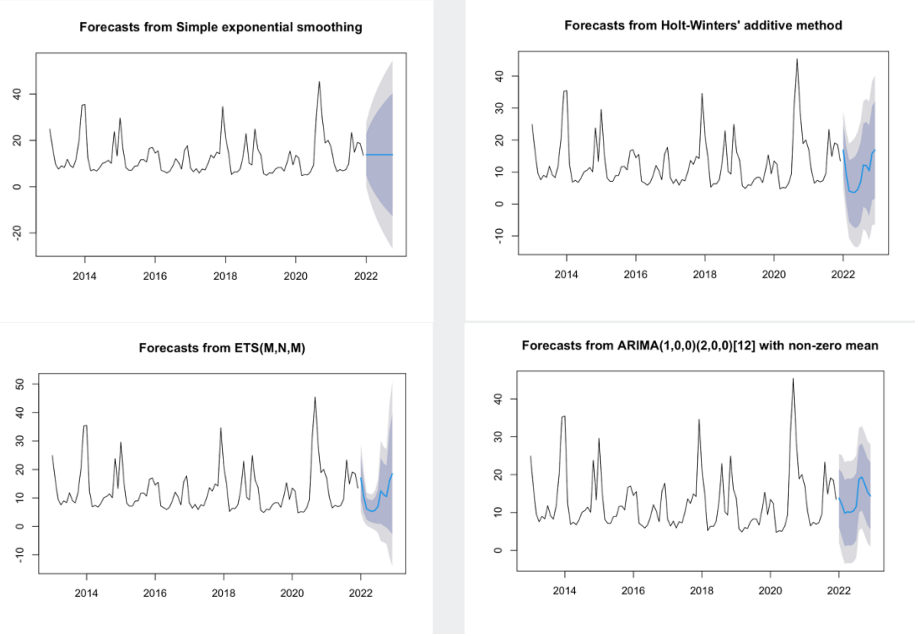
4.2 Fresno, CA: Stationary TS



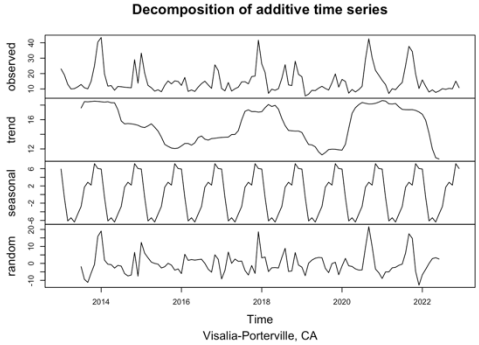


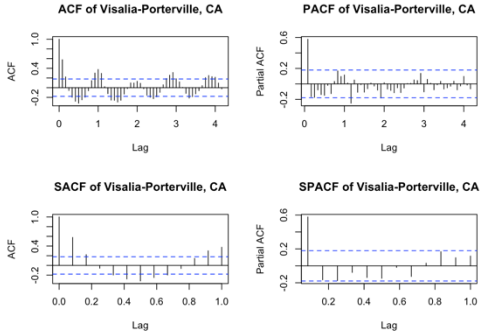

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.7383, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.043568, Truncation lag parameter = 4, p-value = 0.1
```



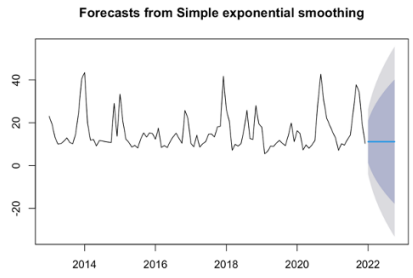
4.3 Visalia-Porterville, CA: Stationary TS

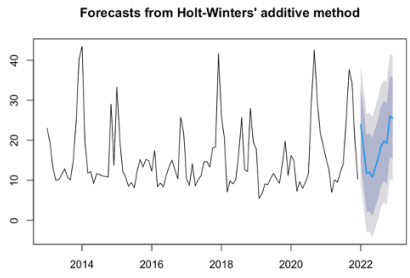


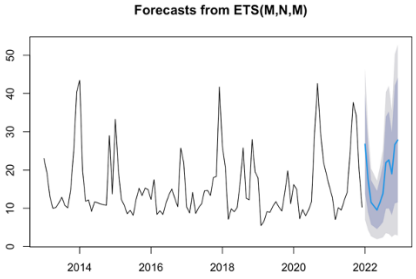


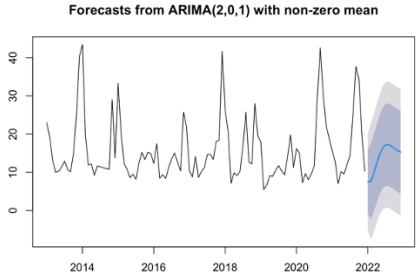

```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.6302, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.043519, Truncation lag parameter = 4, p-value = 0.1
```

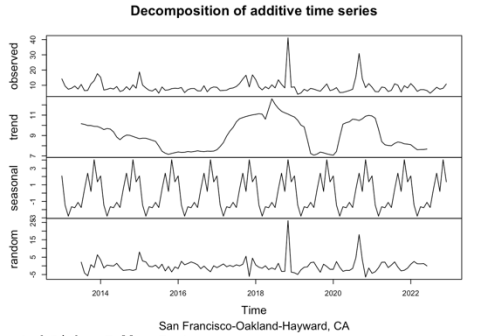


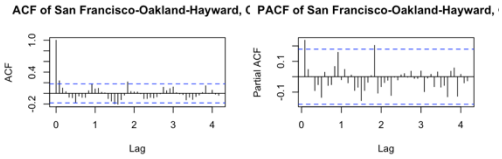


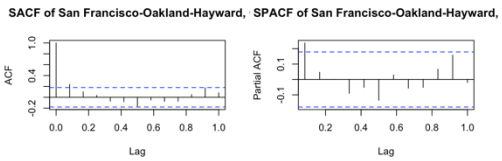




4.4 San Francisco-Oakland-Hayward, CA

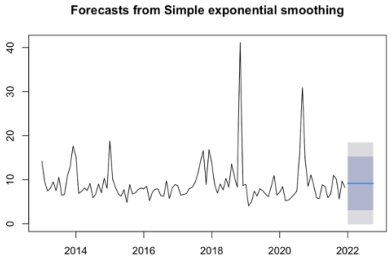


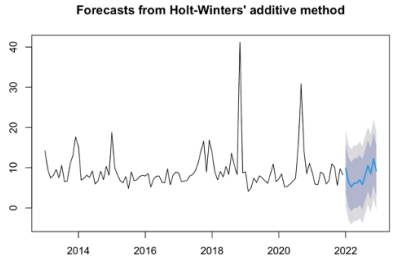


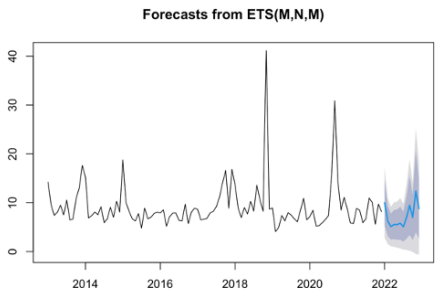


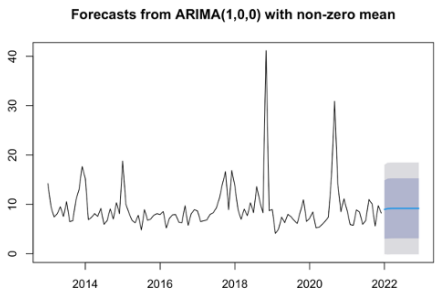

```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -4.8697, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.061811, Truncation lag parameter = 4, p-value = 0.1
```

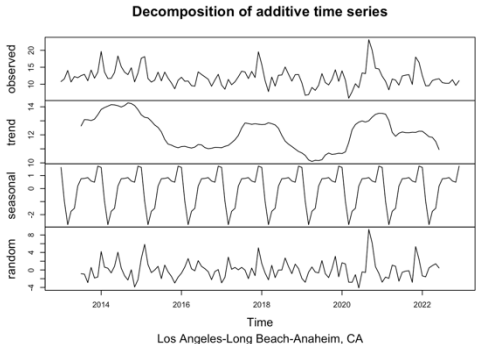


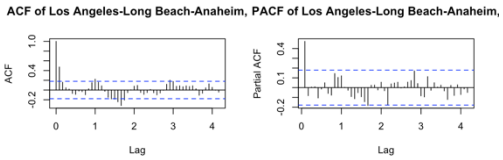


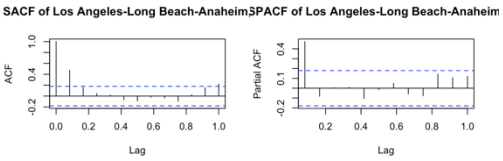




4.5 Los Angeles-Long Beach-Anaheim, CA

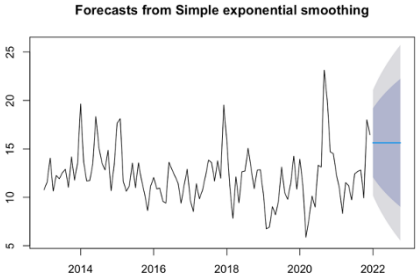


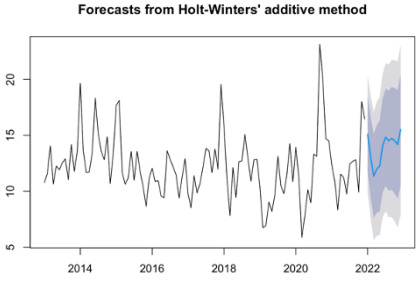


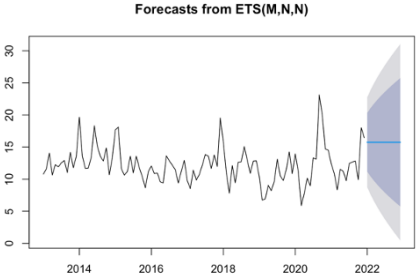


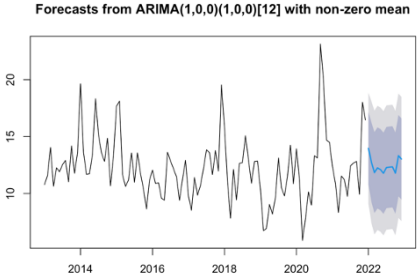

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -4.7625, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.20178, Truncation lag parameter = 4, p-value = 0.1
```

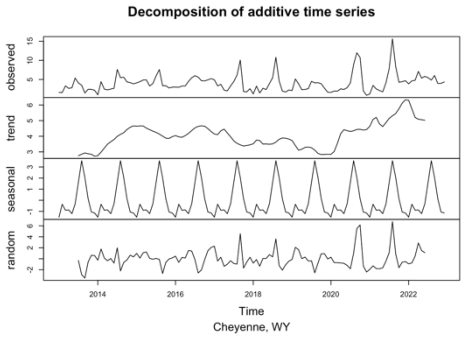


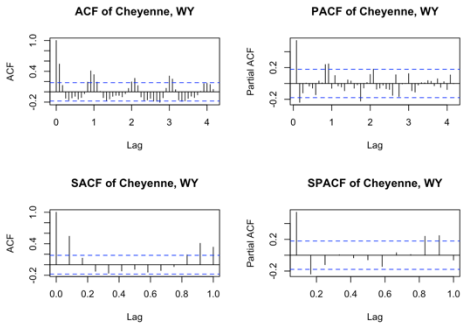






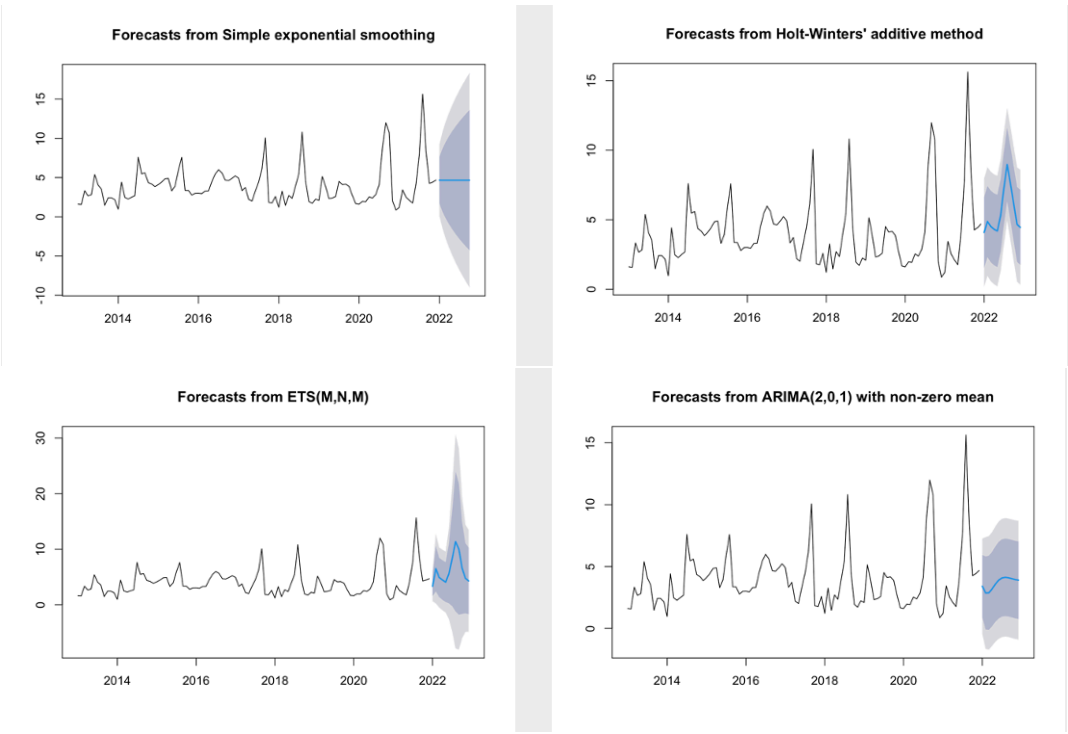
4.6 Cheyenne, WY



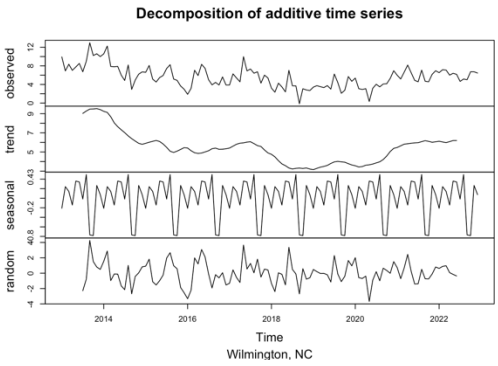


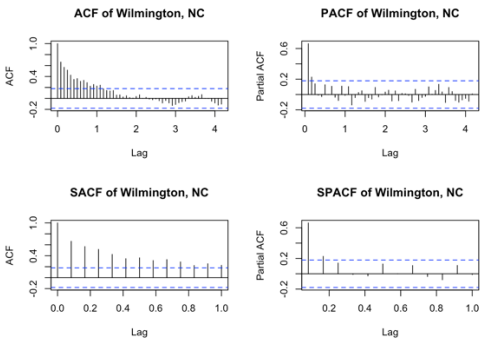

```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.2041, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.28013, Truncation lag parameter = 4, p-value = 0.1
```



4.7 Wilmington, NC

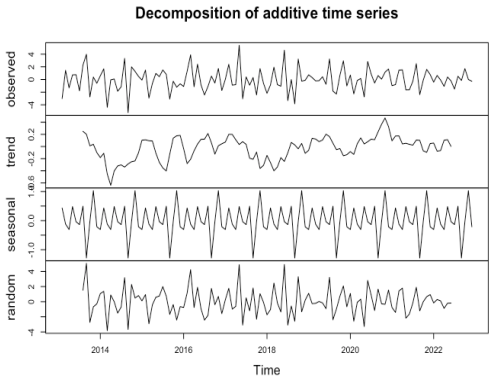


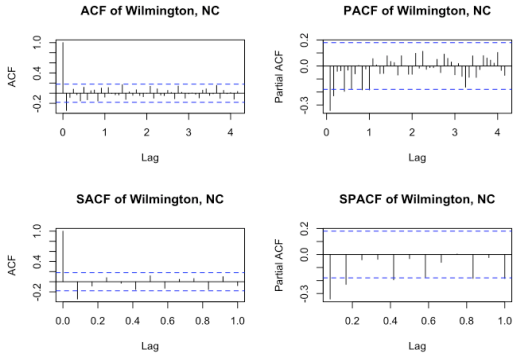


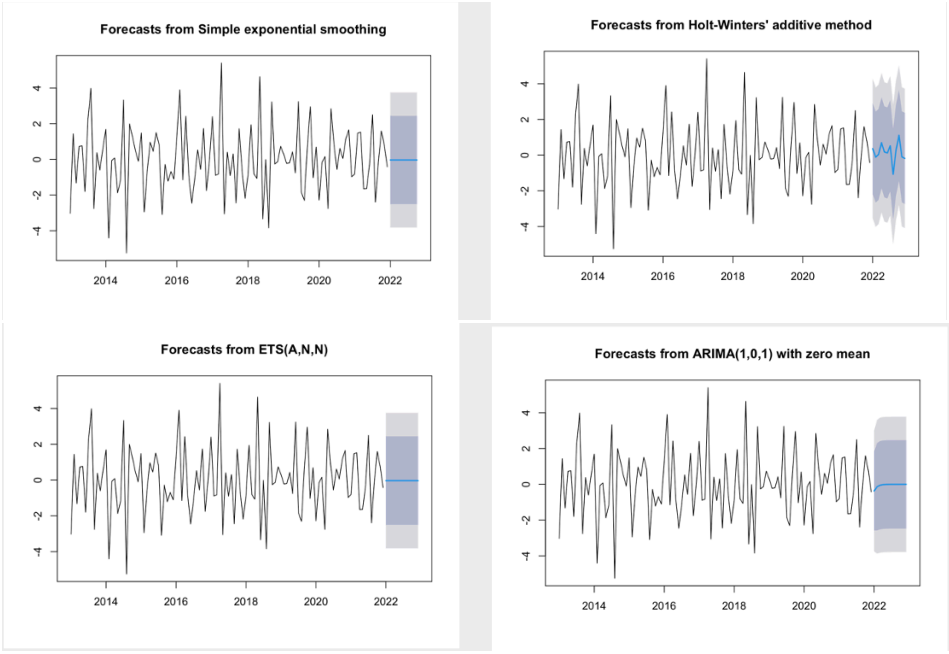
```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -2.8944, Lag order = 4, p-value = 0.2051
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.8084, Truncation lag parameter = 4, p-value = 0.01
```

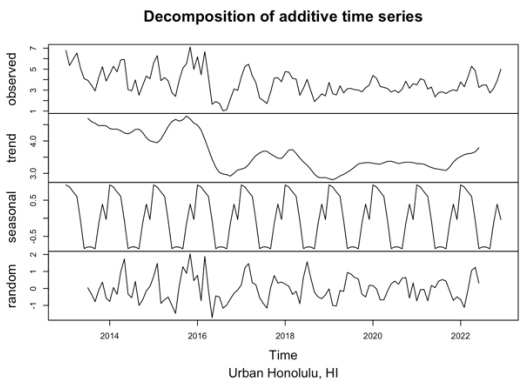
NON-STATIONARY: DIFF () METHOD-

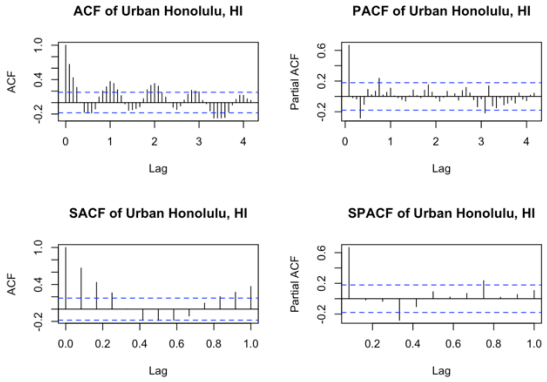






4.8 Urban Honolulu, HI

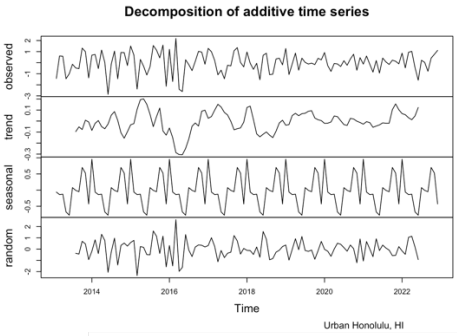


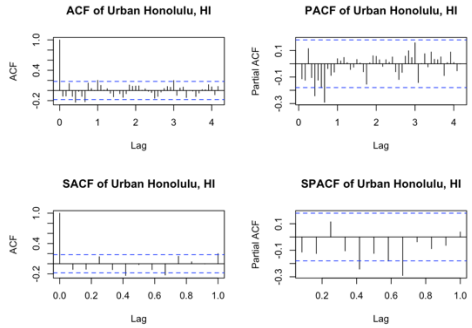


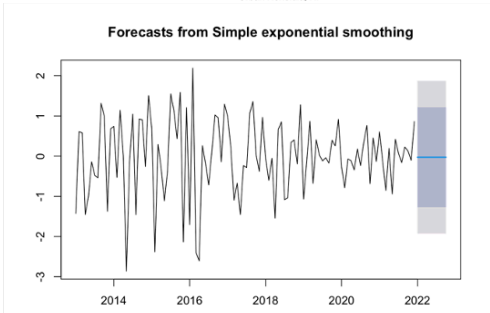
```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.98, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
```

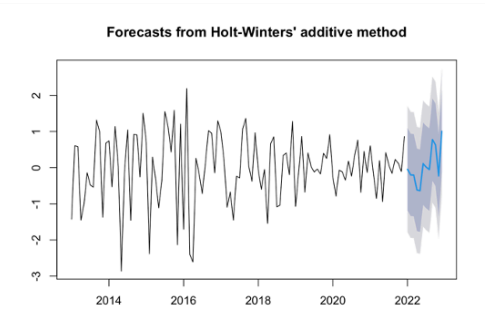
```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 0.65721, Truncation lag parameter = 4, p-value = 0.01744
```

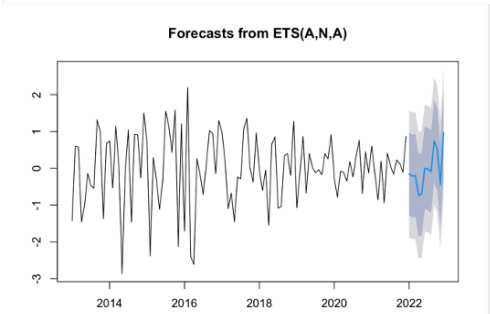
NON-STATIONARY: DIFF () METHOD-

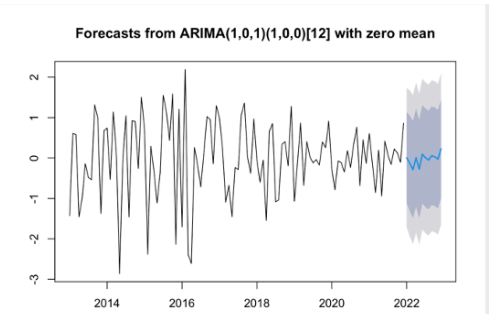




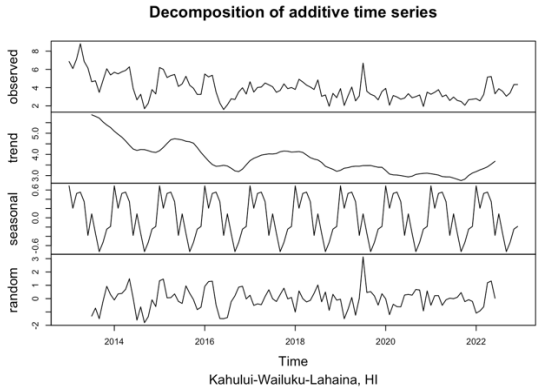


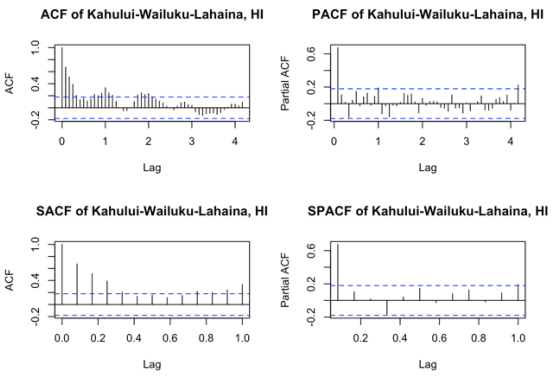






4.9 Kahului-Wailuku-Lahaina, HI

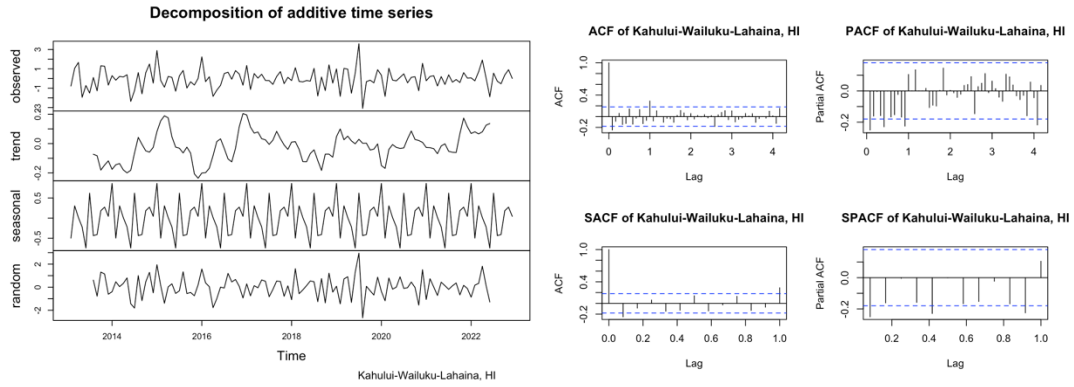


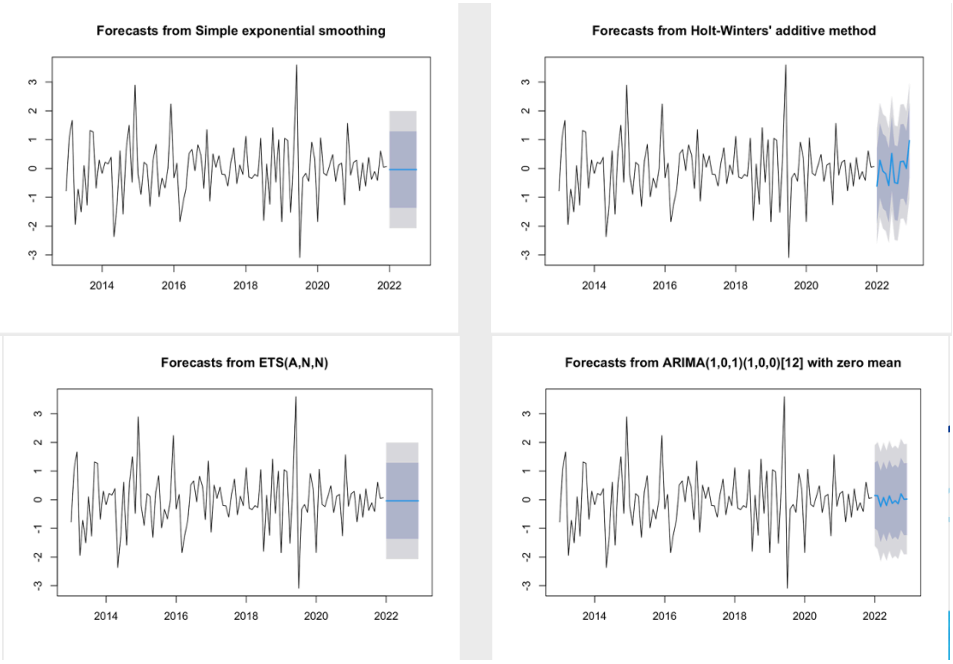


```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.1082, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

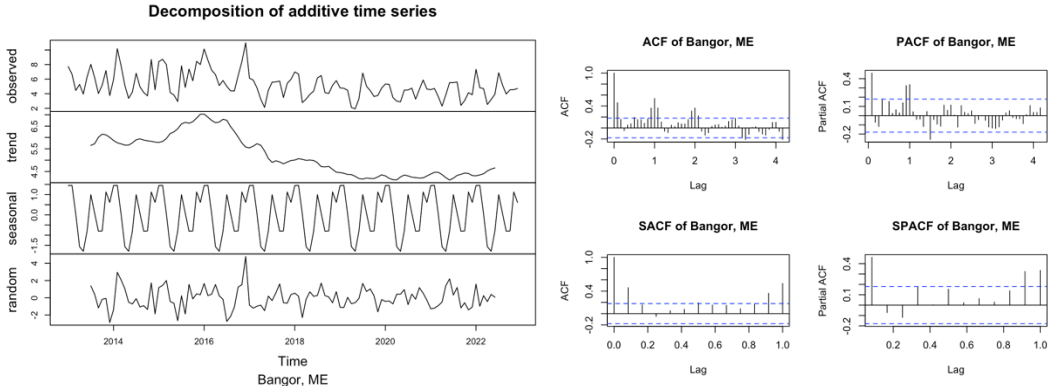
```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 1.1495, Truncation lag parameter = 4, p-value = 0.01
```

NON-STATIONARY: DIFF () METHOD-





4.10 Bangor, ME



```
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -5.4675, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
##  KPSS Test for Level Stationarity
##
## data:  ts_data
## KPSS Level = 1.1557, Truncation lag parameter = 4, p-value = 0.01
```

NON-STATIONARY: DIFF () METHOD-