Air Quality Time Series Report

Adam Sampson, Jay Bektasevic, Brenden McGale, Andrew Brill, Taylor Maxson

February 20, 2018

Introduction

This report is an analysis of data related to air quality. The exploratory question this report investigates is: How closely does weather data correlate with air quality data in Louisville?

Datasets

Weather Data

Hourly weather data was retrieved for Bowman Field Weather Station using the NCDC / NOAA simplified weather data download form at: https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505&countryabbv=&georegionabbv=&resolution=40.

col_name	description
DATE_TIME	Date in GMT
DIR	Wind Direction in compass degrees
SPD	Wind Speed in MPH
TEMP	Temperature in degrees F
PCP01	1-hour liquid precipitation in inches

Air Quality Data

Particulate matter ($PM_{2.5}$) is a broad term used for an airborne mixture of solid particles and liquid droplets. These particles are <= 2.5 microns in diameter and, although the composition is region-specific, are largely made up of sulphate, nitrate, carbon particles and soil. Fine particles are produced from all types of combustion, including motor vehicles, power plants, residential wood burning, forest fires, agricultural burning, and some industrial processes.

Along with ground-level ozone, fine particulate matter is one of the two major components of smog.

From a human perspective, excess fine particulates pose hazards for people with asthma, cardiovascular or lung disease, as well as children and the elderly. These health effects have been associated with both short term (daily) and long term (>year) exposure. However, even if you are healthy, you may feel temporary symptoms if you are exposed to high levels of particle pollution. Numerous scientific studies connect particle pollution exposure to a variety of health issues, including:

- irritation of the eyes, nose and throat
- coughing, chest tightness and shortness of breath
- reduced lung function
- irregular heartbeat
- asthma attacks
- heart attacks
- premature death in people with heart or lung disease

Ecologically, fine particulate matter can damage vegetation and can lead to soil erosion.

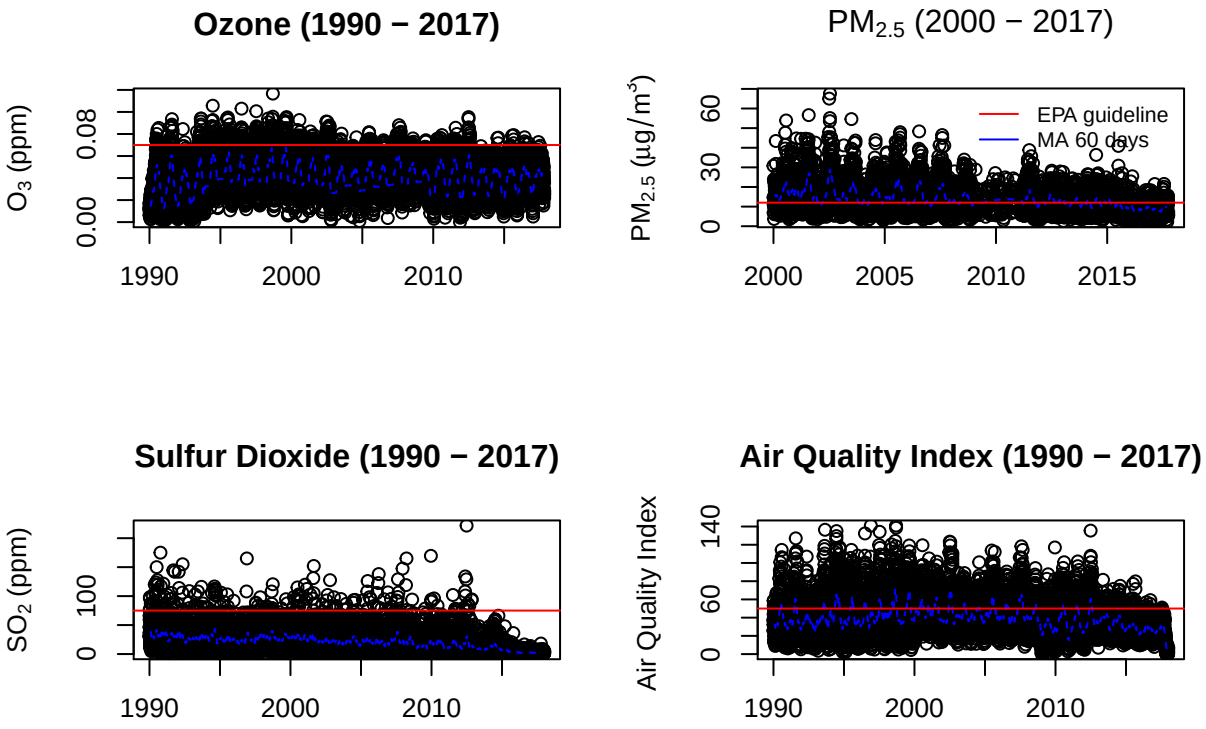
The EPA has set a National Ambient Air Quality Standards for six principal pollutants found here.

Air Quality Index Table

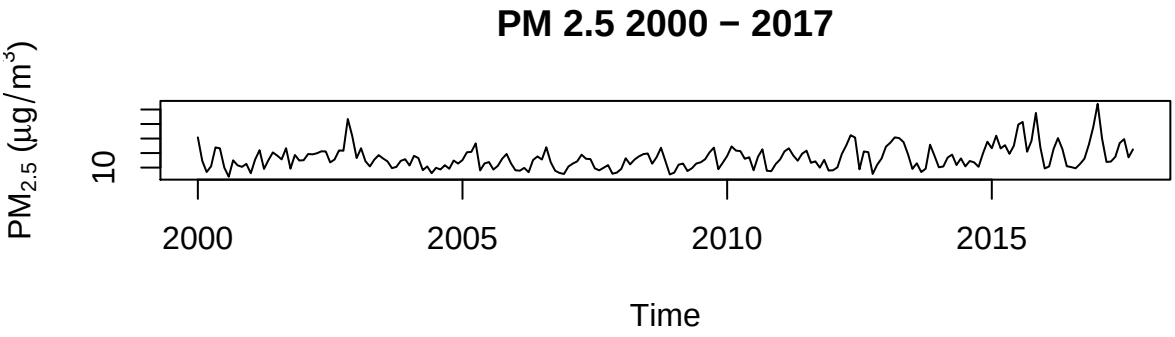
AQ Index	Health Concern	Meaning
0 to 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk.
51 to 100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
101 to 150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
151 to 200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
201 to 300	Very Unhealthy	Health alert: everyone may experience more serious health effects.
301 to 500	Hazardous	Health warnings of emergency conditions. The entire population is more likely to be affected.

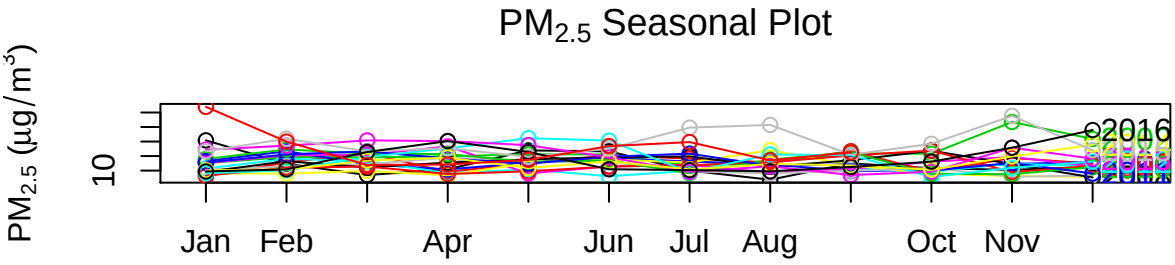
Data was aquired from the EPA website there are several monitoring stations in Jefferson County.

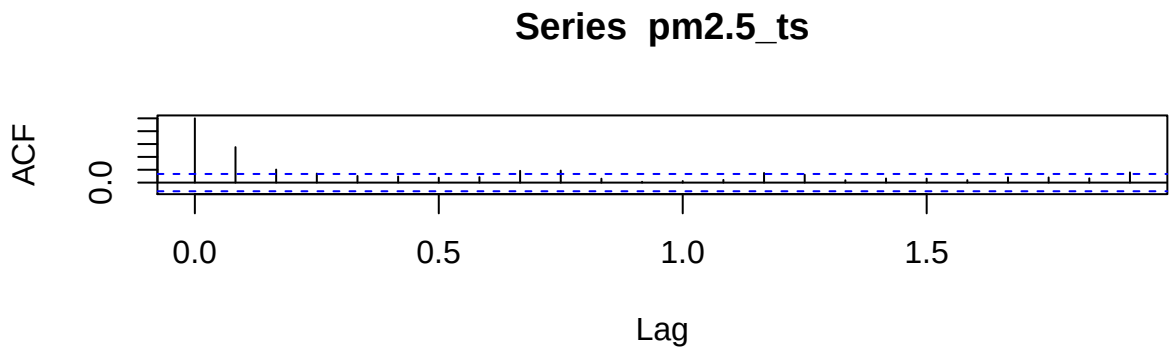
Exploratory Analysis of Air Data

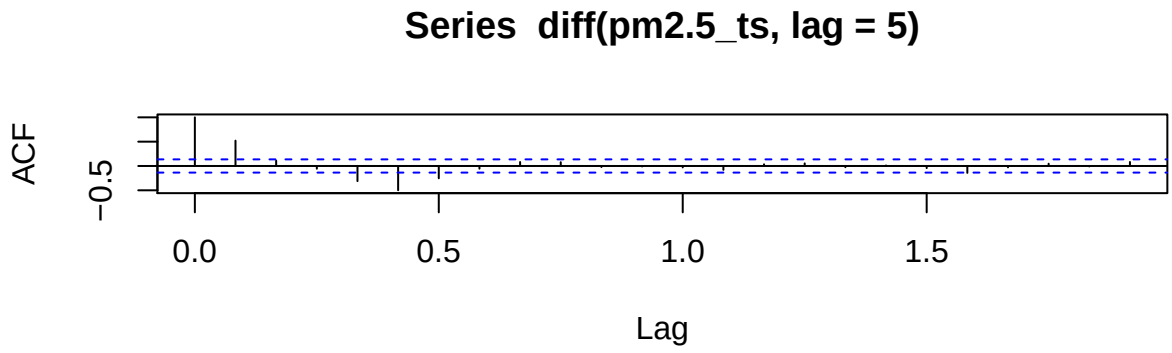


Seasonal Variation in Air Quality

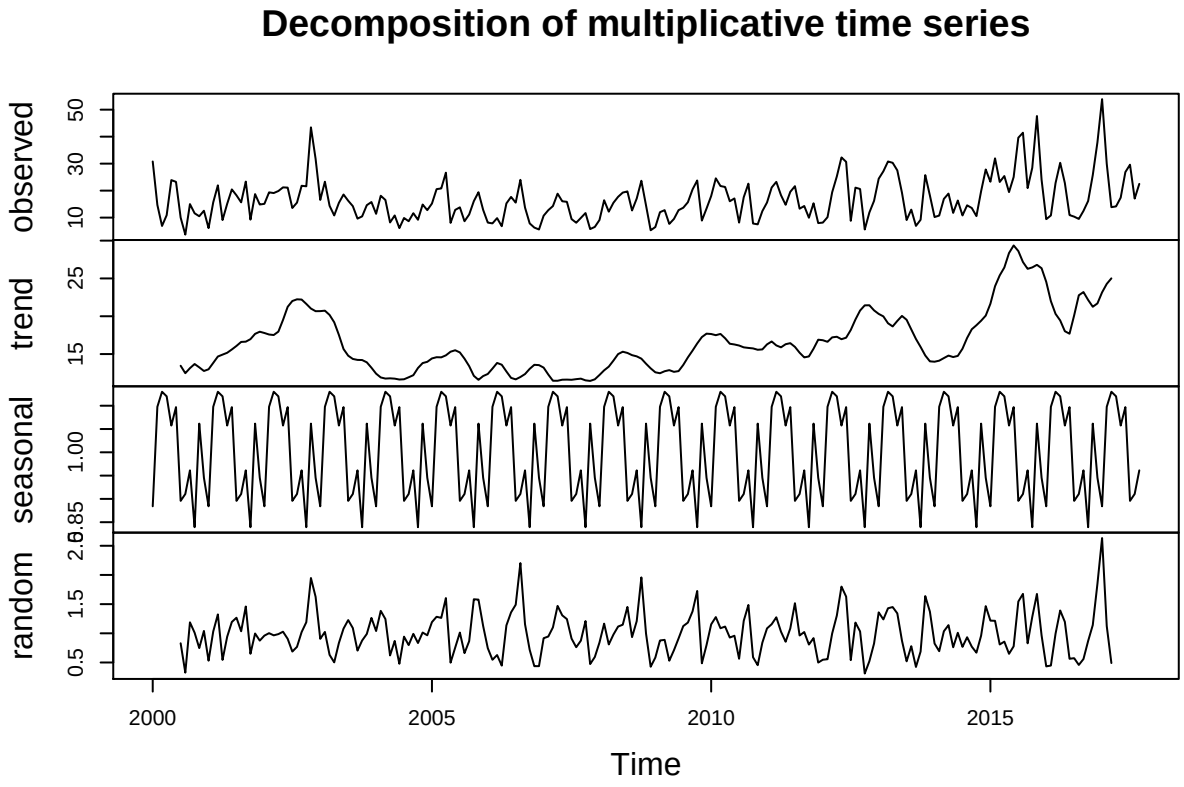









```
##
##  Augmented Dickey-Fuller Test
##
## data:  pm2.5_ts
## Dickey-Fuller = -5.3557, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

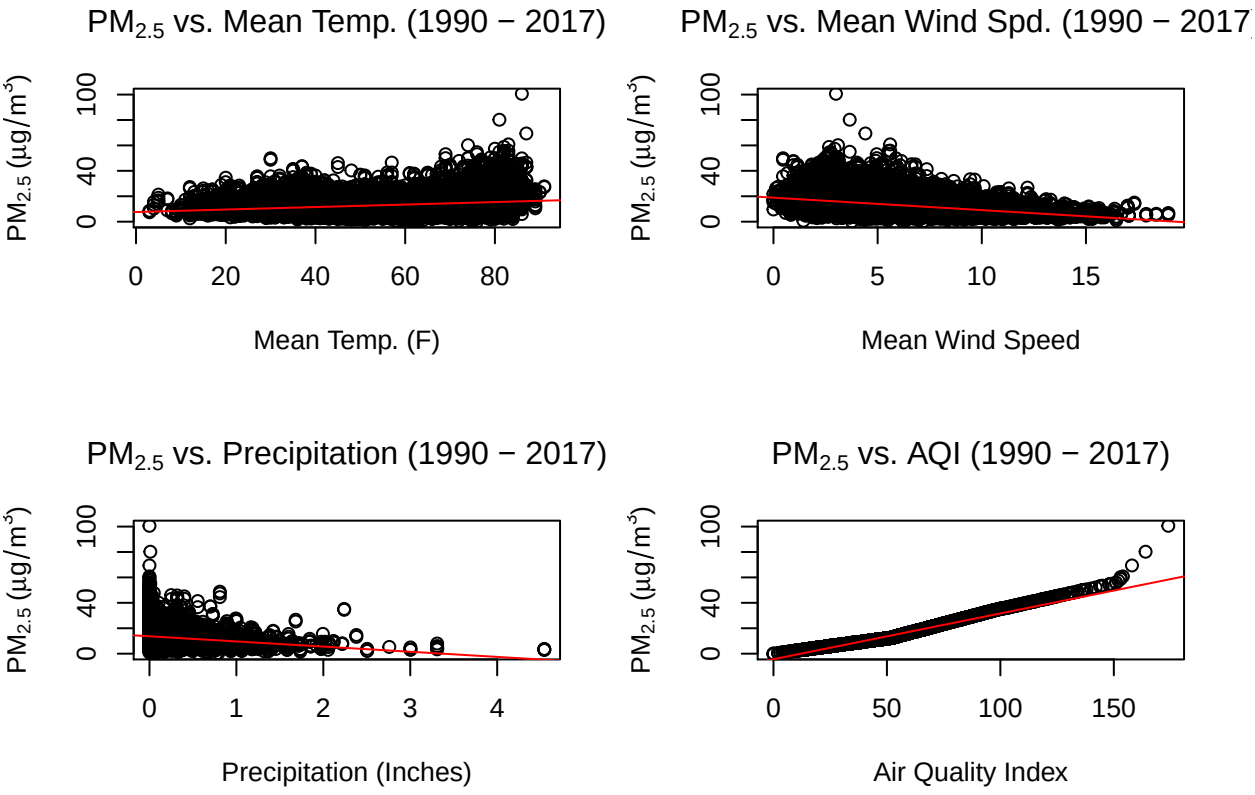


For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.

Correlations with other pollutants

In order to examine the cause of pollution we will need to look at correlation between the $PM_{2.5}$ and other weather variables such as temperature, wind, and precipitation.

Now we can plot the variables and see if they corrlate with each other.



As shown above, there is a slight correlation between $PM_{2.5}$ and mean temperature. It shows a slight increase in $PM_{2.5}$ as the mean temperature increases. Another interesting relationship can be observed by plotting $PM_{2.5}$ vs. Mean Wind Speed –we can observe that the $PM_{2.5}$ levels actually drop as the Mean wind speed increases. This is counter-intuitive because one would think that higher winds would cause $PM_{2.5}$ levels to spike. This phenomenon can actually be explained, since, we are in the valley and all the pollution gets trapped in the valley, higher wind speeds are attributed to weather changes such as weather fronts that push out air from the valley - hence the lover readings.

Increase in precipitation shows a similar relationship as with the wind speed. This phenomenon also makes sense as an increase in precipitation allows fine particulates to bind to rain droplets and fall to the ground.

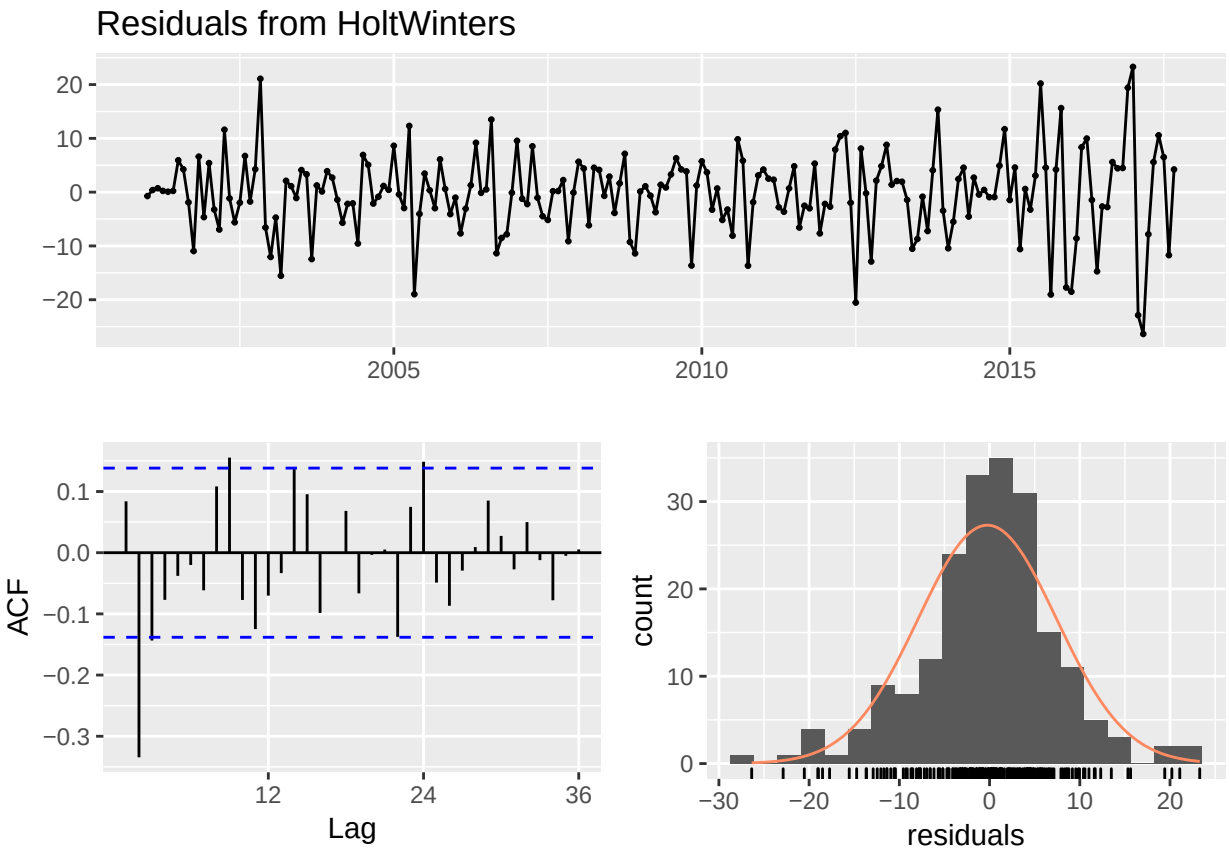
The obvious relationship is an increase in an Air Quality Index also proportionately increases $PM_{2.5}$ levels. Note – $PM_{2.5}$ levels are used in calculating AQI.

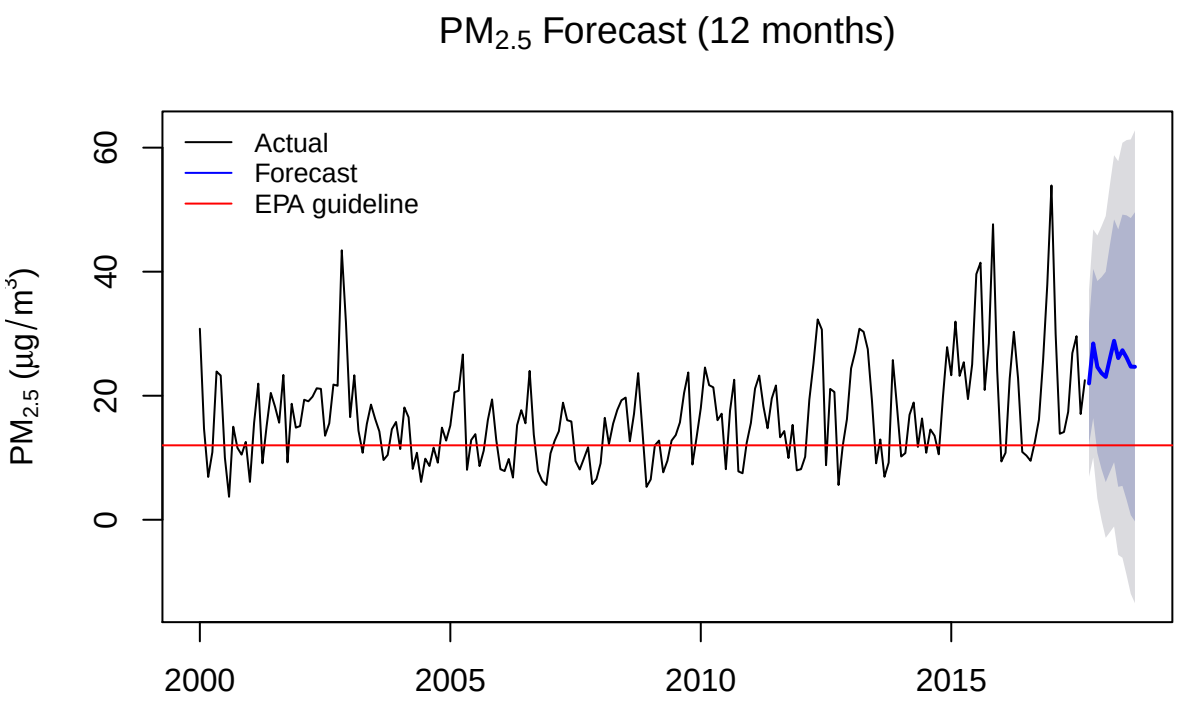
$PM_{2.5}$ Forecast

Now that we know that the time series is a stationary we can go ahead and forecast $PM_{2.5}$ levels 12 months in the future. I will use Holt-Winters model as it does a great job of picking up seasonality and the trend.

```
##
## Forecast method: HoltWinters
##
## Model Information:
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = pm2.5_ts)
```

```
##
## Smoothing parameters:
##  alpha: 0.7017798
##  beta : 0
##  gamma: 0.4709145
##
## Coefficients:
##            [,1]
## a   24.50950834
## b    0.23581002
## s1  -2.76010086
## s2   3.44442944
## s3  -0.58696649
## s4  -1.81290688
## s5  -2.65830107
## s6   0.07625436
## s7   2.69444983
## s8  -0.32851641
## s9   0.70069451
## s10 -0.72336208
## s11 -2.40994971
## s12 -2.67606548
##
## Error measures:
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -0.2494126 7.663517 5.600714 -13.04495 39.42491 0.6919845
##                    ACF1
## Training set 0.0839691
##
## Forecasts:
##          Point Forecast      Lo 80    Hi 80       Lo 95    Hi 95
## Oct 2017       21.98522 12.1447182 31.82572   6.9354725 37.03496
## Nov 2017       28.42556 16.4036454 40.44747  10.0396293 46.81149
## Dec 2017       24.62997 10.7657254 38.49422   3.4264366 45.83351
## Jan 2018       23.63984  8.1508686 39.12881  -0.0484984 47.32818
## Feb 2018       23.03026  6.0715064 39.98901  -2.9059141 48.96643
## Mar 2018       26.00062  7.6896920 44.31155  -2.0035292 54.00477
## Apr 2018       28.85463  9.2847244 48.42453  -1.0749569 58.78421
## May 2018       26.06747  5.3148315 46.82011  -5.6709528 57.80590
## Jun 2018       27.33249  5.4609809 49.20401  -6.1170982 60.78208
## Jul 2018       26.14425  3.2083795 49.08011  -8.9331351 61.22163
## Aug 2018       24.69347  0.7404952 48.64644 -11.9394431 61.32638
## Sep 2018       24.66316 -0.2654529 49.59178 -13.4618652 62.78819
```


Residuals from HoltWinters
20
10
0
−10
−20
2005
2010
2015
ACF
0.1
0.0
−0.1
−0.2
−0.3
12
24
36
Lag
count
30
20
10
0
−30
−20
−10
0
10
20
residuals



Even though the short-term forecast looks reasonable, this is a very crude model for a complex process. The residuals demonstrate that there is a lot of information that has not been captured with this model.

Techniques

Data Cleaning

Cleaning Weather Data

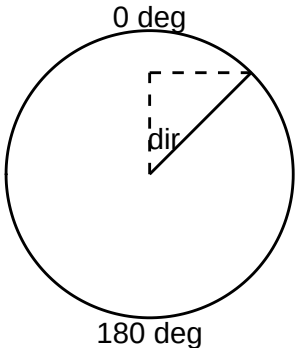
Variable	Cleanup
Date_TIME	Clean
DIR	Any * in field indicates NA
	990 - Variable - NA
	>360 - Uncertain - NA
SPD	Any * in field indicates NA
TEMP	Any * in field indicates NA
PCP01	Any * in field indicates NA

Converting Variables

Converting Wind Compass Direction to Inverval Data

Wind data is recieved in compass degrees. This is problematic when performing calculations because 0/360 degrees is between 350 degrees and 10 degrees, but the mathematical average of 350 degrees and 10 degrees is 180 which is exactly the opposite of the true average direction.

One way to make this work better is to convert from polar coordinates to cartesian coordinates.



$$sin(dir) = \frac{x}{1} \rightarrow x = sin(dir * \frac{180}{\Pi})$$
$$cos(dir) = \frac{y}{1} \rightarrow y = cos(dir * \frac{180}{\Pi})$$

An interesting bi-product of this transformation is that it is now possible to have a wind direction of {0,0} when there is no wind speed and lack of wind does not necessarily mean that wind direction has to be removed from calculations as NA.

Aggregating Data

Aggregating Weather Data to Daily

The Air Quality data is in daily format. Therefore it makes sense to aggregate the weather data to daily values as well.

Input	Output(s)
TEMP	min / mean / max
x.dir	min / mean / max
y.dir	min / mean / max
SPD	min / mean / max
PCP01	sum

Extracting addional information

Extracting Frequency Data from Precipitation

Dealing with NA values

Imputing NA values in Weather Data

Input	Impute Method
TEMP	na.approx
x.dir	0 if NA
y.dir	0 if NA
SPD	na.approx
PCP01	0 if NA

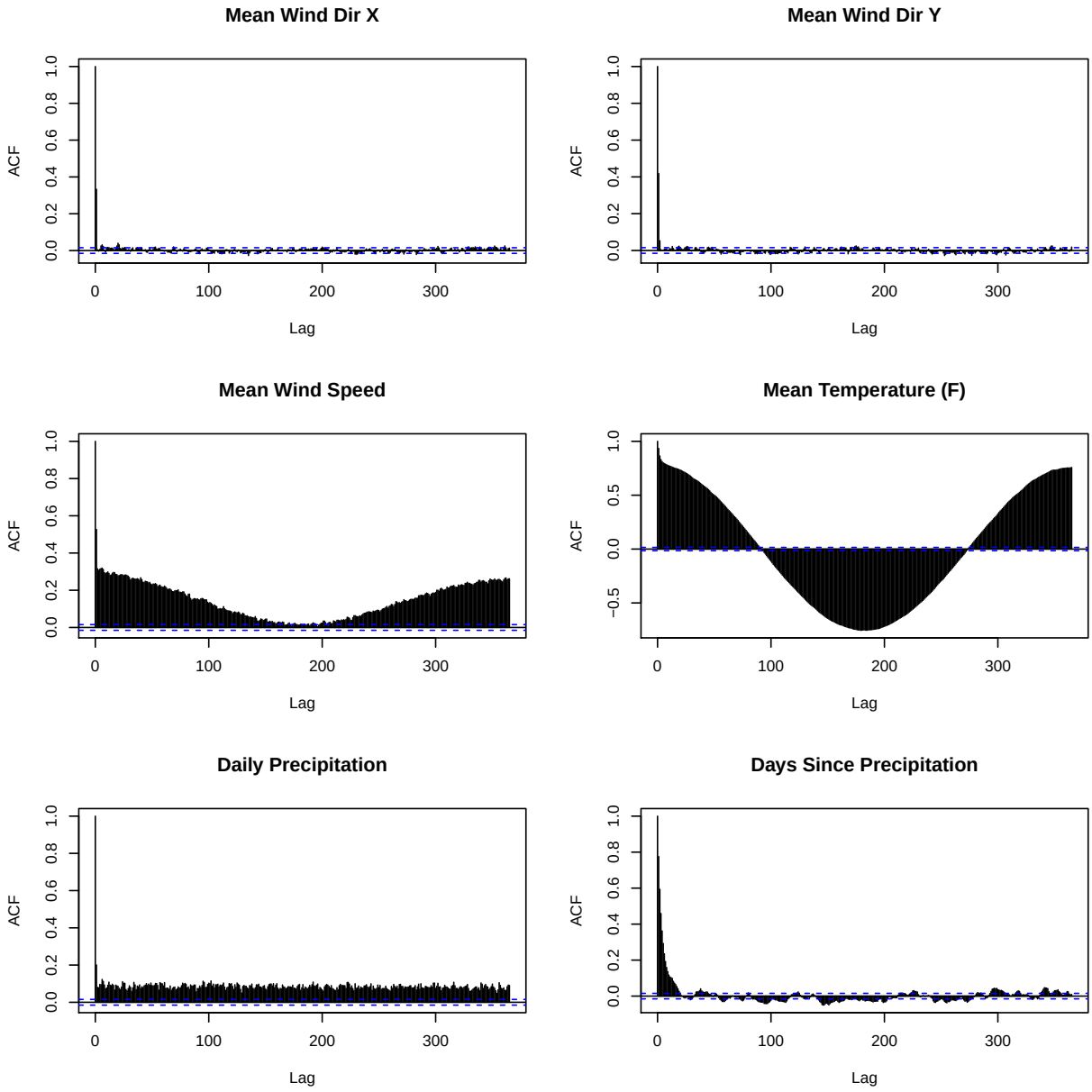
Checking for Seasonality and Cycles

The Autocorrelation of variables was checked using the `acf()` function.

Weather Data Seasonality

The charts below show that some of the weather data is seasonaly correlated over a year, and one of the sets of data is cyclically correlated to recent days.

- Wind Direction is not seasonally correlated. There is a minor correlation to the day before, but minimal correlation beyond 2 days.
- Wind speed is seasonally correlated on a yearly basis.
- Temperature is seasonally correlated on a yearly basis.
  - As expected
- Daily Precipitation is autocorrelated all over the place. For Louisville it appears that there is often rain followed by several days of no rain and then repeating.
  - This is siginficantly different from western states where rain is seasonally correlated to the time of year with wet winters and dry summers.
  - It may be possible to correlate amount of rain on a seasonal basis still
- Days since last rain is autocorrelated over several days and falling off the farther you get from 0 days. This re-enforces the concept that rain is correlated heavily to whether it rained recently or not.



Decomposition

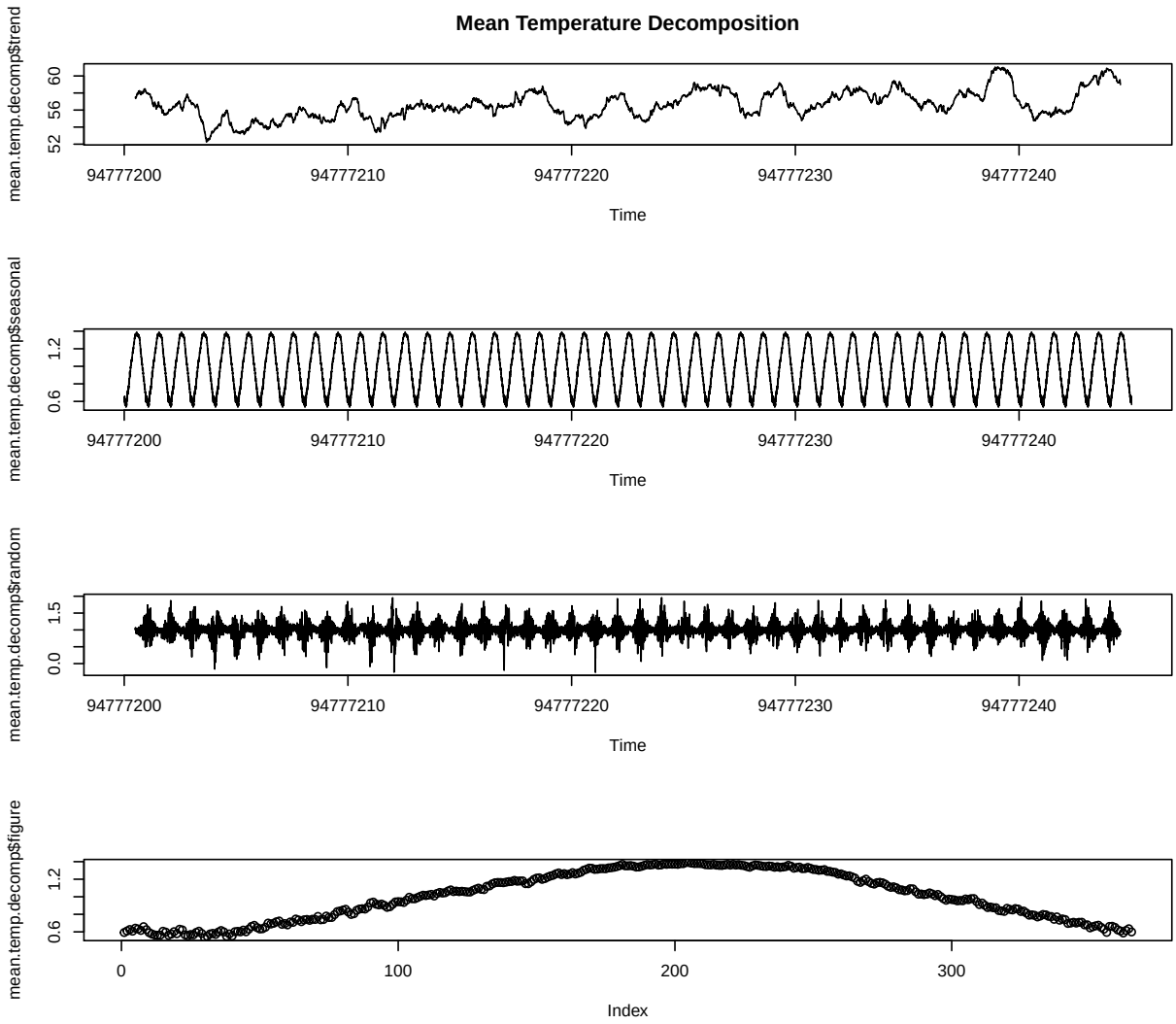
Weather Data Decomposition

For weather data it only makes sense to decompose Temperature, Wind Speed, and Rain Amount based on the results of the `acf()`.

Mean Daily Temperature

The plots below show that mean daily temperature is seasonal over a yearly period. Interestingly, the random component is still seasonally affected. This is likely because the day-to-day variation changes between seasons. More advanced methods are needed to deal with this, however this decomposition should be sufficient for comparing to air quality data.

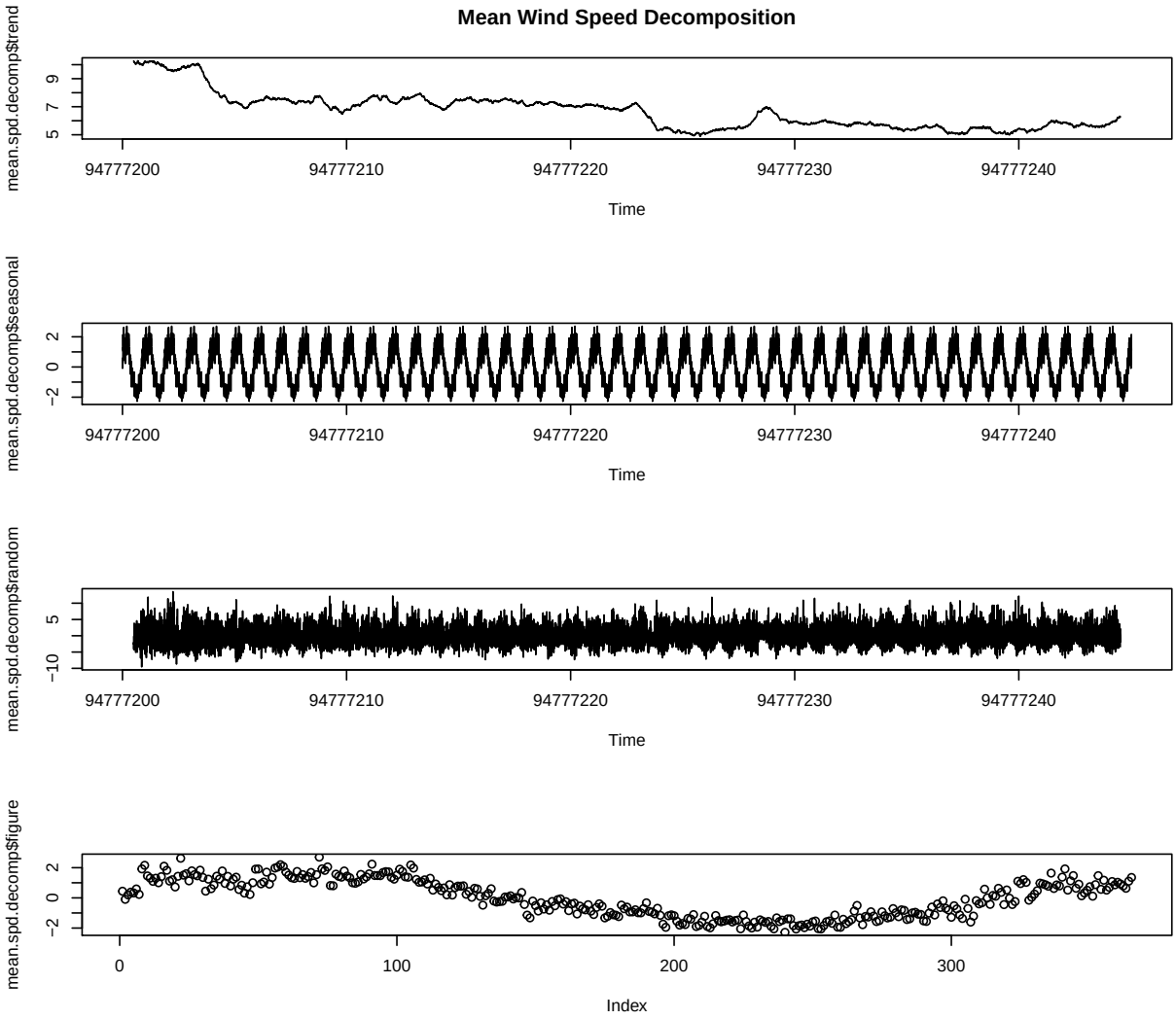
It is difficult to tell from this time-frame whether there is a real trend in temperature. More detailed analysis is required for that.



Mean Daily Wind Speed

The decomposition shows that there is likely yearly seasonality in the wind speed data. Once again, the random output shows likely seasonality as well. This indicates further work is required to finish the decomposition.

The trend retrieved from this decomposition is interesting. It shows that average daily wind speed is decreasing since 1973.

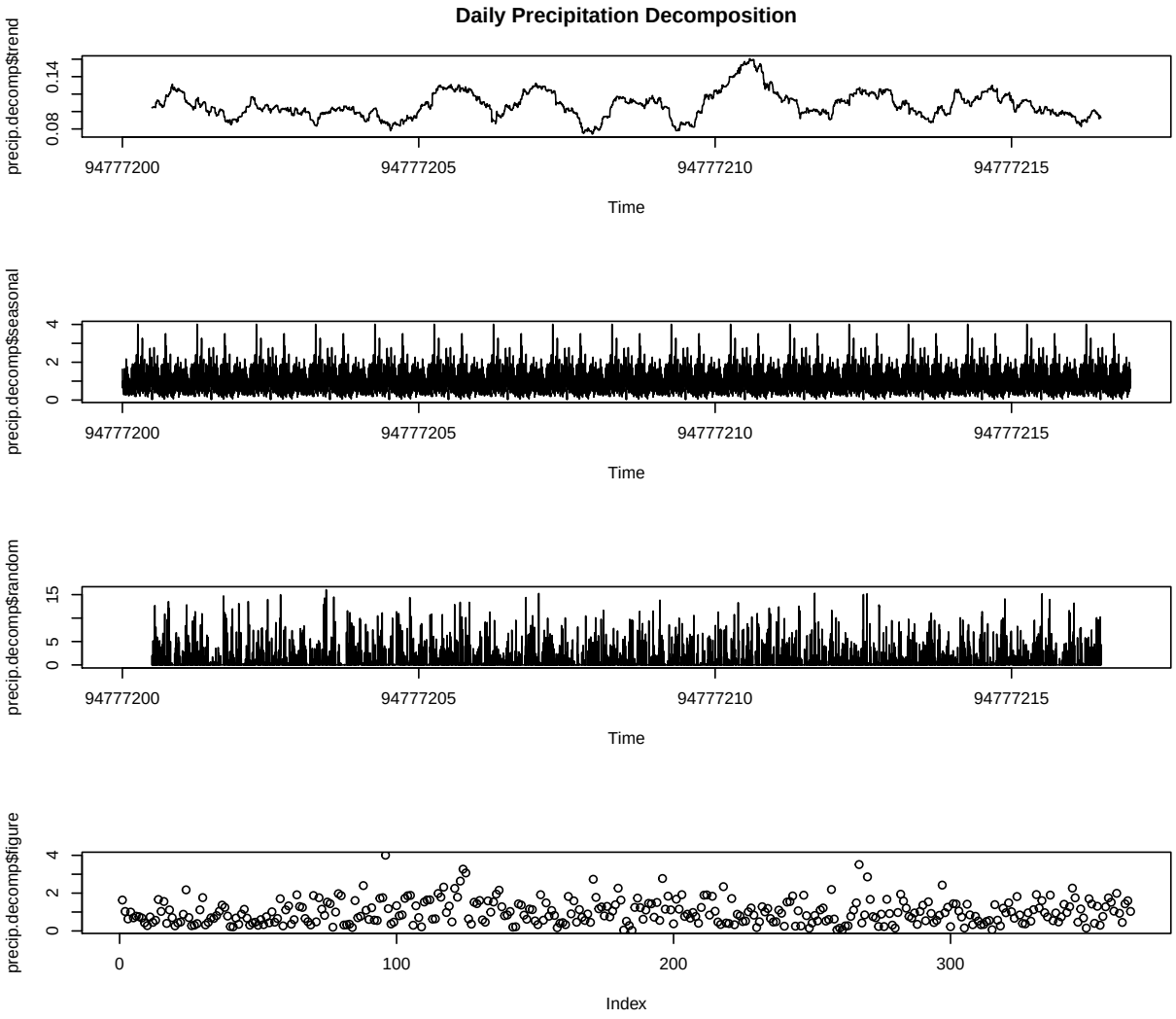


Daily Precipitation Decomposition

The decomposition of rainfall is fairly inconclusive. Visual review of the charts suggest that there may be higher daily rainfall possible during summer months (when there are thunderstorms and hurricanes), but the decomposition is not clean.

The data either needs to be transformed or a different method of decomposition needs to be used to validate whether the thunderstorm hypothesis is correct.

For the needs of this analysis it would be better to simply compare rainfall to air quality data with and without lag.



Conclusions

As windspeed increases, Particulate Matter Decreases.

As temperature increases, probability of high PM2.5 increases.

Heavey rainfall coincides with low PM counts.

Forecast model was not accurate.

External factors outside of weather were not considered (such as the effects of regulation or industrial changes).