Data Analysis 01 report
ADVANCED METHODS III 140.753

Abstract

For this assignment, we had to obtain the last 10 years of data for Ozone and PM 2.5 from the EPA website and explore the following points. (1) What are the trends? (2) Do the trends differ between Ozone and PM 2.5? (3) Are there any differences between the East and West of the United States? (4) Can we detect any change in regulation? In this document I describe the pre-processing and analysis I carried out.

1 Pre-processing

Once I downloaded all the data, I wanted to explore the impact of all the variables by looking at the full data set. Via simple summary analysis, I determined to keep the columns *State.Code, County.Code, Site.ID, POC, Unit, Method, Date, Start.Time, Sample.Value* as I determined that they had the most usefulness potential[1]. The size of the tables posed a problem for other pre-processing steps, so I used the *IRanges* package to compress the information from 18Gb to 2Gb in RAM[2]. Using *IRanges* I cleaned the information by keeping only *State, County, Site.ID, Date, Sample.Value* while using other tables from the AQS data site to obtain the state abbreviations and county names[3]. Furthermore, I adjusted the units to by keeping them in parts per million and micrograms per cubic meter for Ozone and PM 2.5 respectively. A few entries in liters per minute were discarded because they are incompatible with micrograms per cubic meter. Finally, I created two major summaries by calculating the mean, the standard error, first and third quartiles by grouping the information by the unique combination of State (abbr), County (name), Month (year and month, set day to 15), and Site ID. The simpler version did not consider the Site ID in creating the unique combinations. Doing so efficiently was much more challenging than what I expected[4] because the *tapply()* function was rather slow with objects from the *IRanges* package.

2 Exploratory data analysis

For this section, I used the summarized data that did not take into account the Site ID. While there are 96822 total entries once merging Ozone and PM 2.5, the ones without any missing information reduce to 40401. This an pose some problem as we can see in Figure 1 where we can clearly notice the missing information, which can be more frequent in some states such as *AK:Alaska* and *NE:Nebraska*.

Figure 1 provides a major indication as to when regulation might have been implemented during the period under analysis. That is because there are several states, such as WA and VT that had high values in the early years (closer to purple) which decrease during 2007 and 2008. I believe that this reflects the regulations the EPA undertook as summarized in `http://www.epa.gov/pm/actions.html`.

In Figure 1, one state that pops out is *AR:Arkansas* because of the high peak in PM 2.5 in 2011. Figure 5 (supplementary material) shows the trend in more detail.

Figure 2 is the most important one in this report. This figure shows the mean of the PM 2.5 and Ozone levels per month by averaging the information from the counties. I used natural splines

[1] More at data/pre-proc.R
[2] More at data/pre-small.R
[3] More at data/process-pre.R and data/get-ref.R
[4] More at data/summarize-clean.R

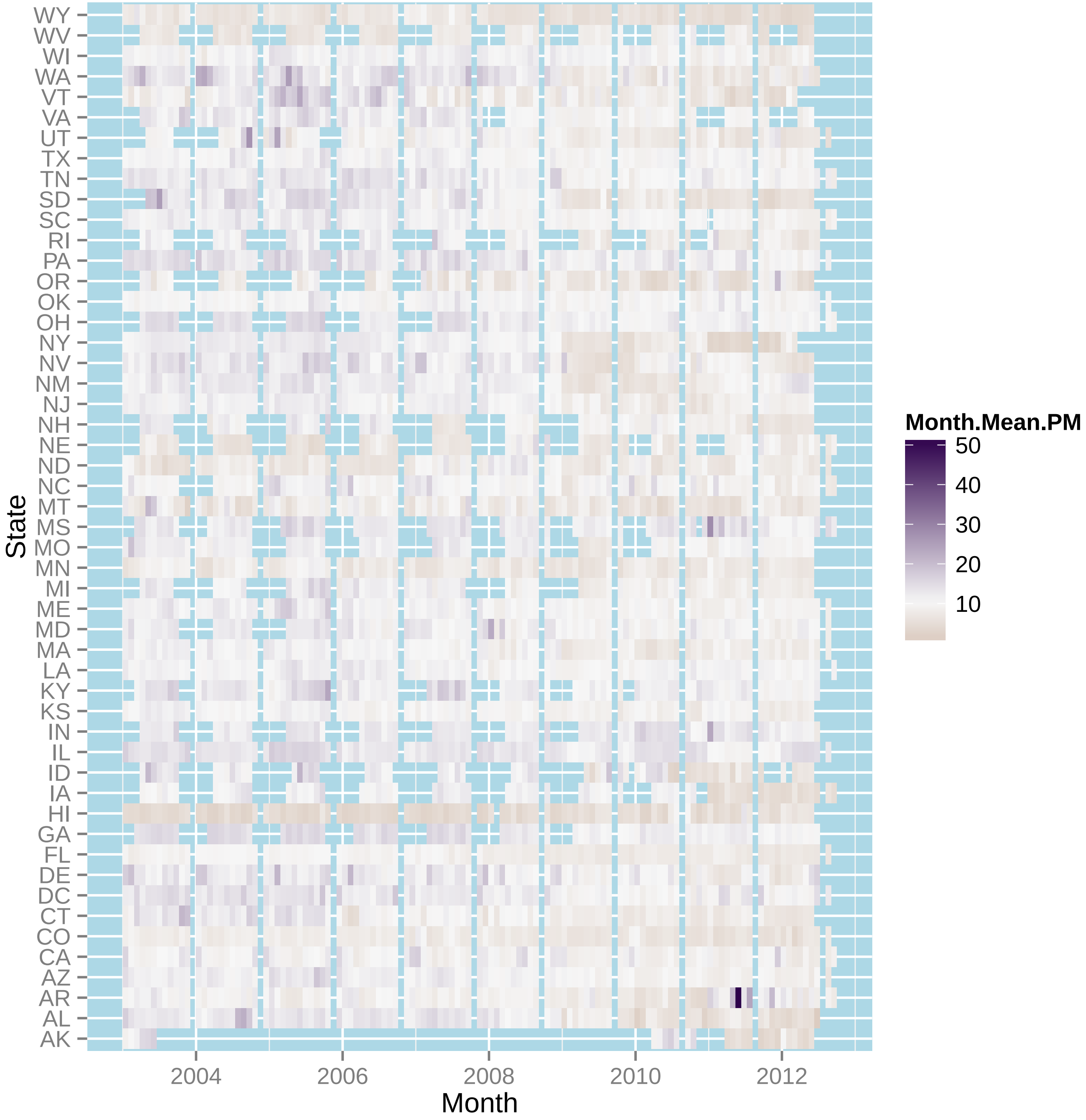

Figure 1: Montly mean PM 2.5 from the counties separated by state. Background is in light blue to clearly notice the missing observations. There are some months without any data in all states. Uses only complete cases.

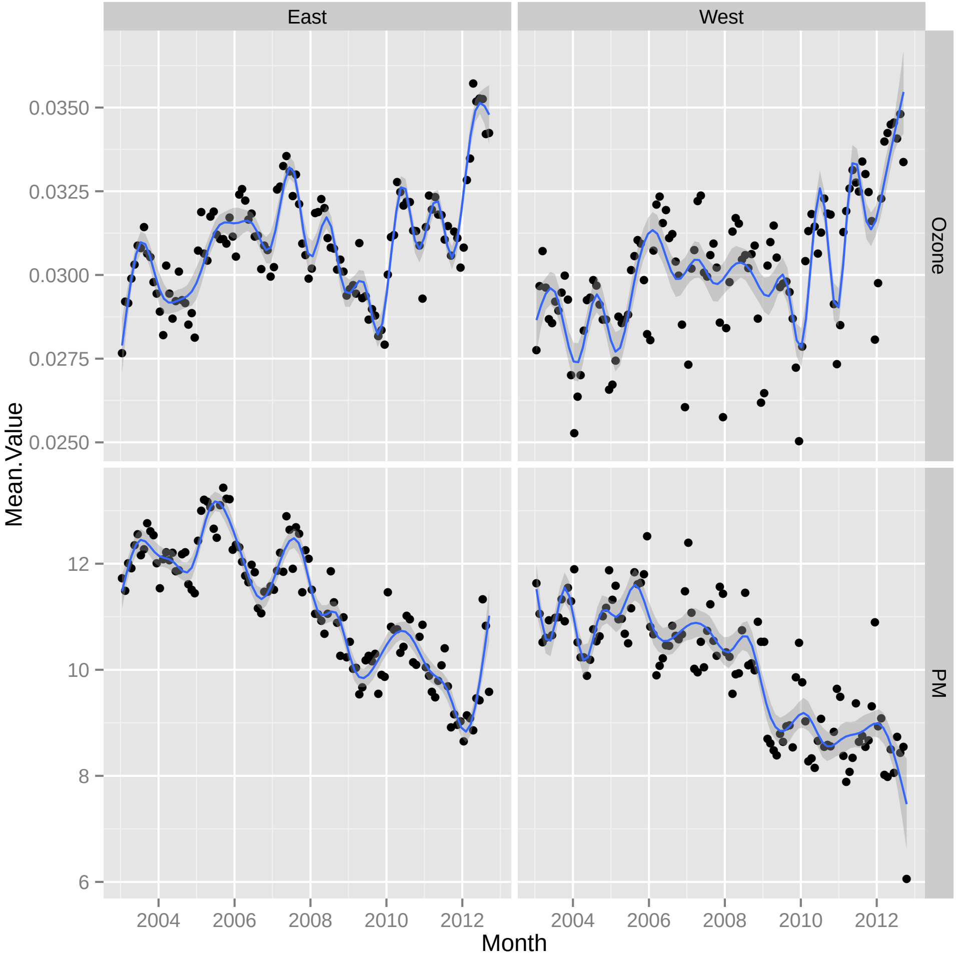

Figure 2: Mean PM 2.5 and Ozone levels per month from the counties. Smooth line is a from a natural spline with 20 degrees of freedom. Grey bands mark the 95% confidence bands. Uses full data as it does not pair up PM 2.5 and Ozone observations.

with 20 degrees of freedom to smooth the trends. We have data for 10 years and there seem to be clear seasonal trends around winter and summer. Thus, 20 degrees of freedom seems appropriate.

Figure 2 clearly shows that the PM 2.5 has been decreasing since 2008. It still maintains a seasonal trend with higher values during the summer and lower ones during the winter. Notably, the winter of 2012 has been rather low. While Ozone shares overall the seasonal trend with PM 2.5, it has been increasing since 2010. The fact that both 2011 and 2012 presented higher Ozone values than 2010 is alarming.

We can also use Figure 2 to compare the East versus the West states of the Unites States. Overall, the East states have higher PM 2.5 and Ozone values than the West states. This is expected in a sense because the majority of the densely populated areas of the US are in the East coast.

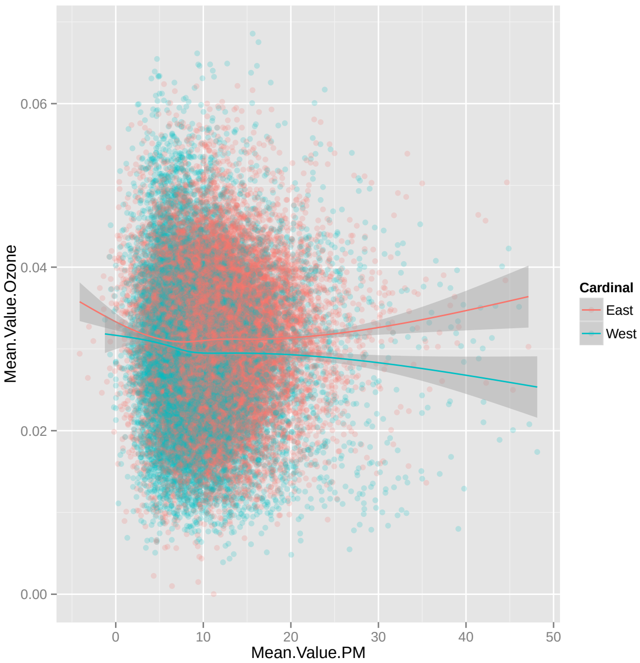

Figure 3: Mean PM 2.5 vs mean Ozone levels by county. Alpha set to 1/5 so 5 points are needed for alpha to reach 1. Smooth lines are from a natural spline with 5 degrees of freedom. Uses complete cases.

3 PM 2.5 versus Ozone

Using the data from all the counties, it would be great if we could find a direct relationship between PM 2.5 and Ozone levels. However, Figure 3 does not indicate to us that anything is going on. This affirmation is further supported by the poor results of a multiple linear regression as shown in Table 1. The adjusted R-squared for this regression is 0.0147 which is very weak. Also, the cardinality (East or West) has a larger coefficient than other variables!

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0E-02	6.2E-04	3.3E+01	4.7E-238
Mean.Value.PM	-1.5E-04	7.7E-05	-1.9E+00	5.4E-02
SE.PM	1.1E-04	3.0E-05	3.7E+00	2.2E-04
Q1.PM	1.9E-04	5.0E-05	3.9E+00	9.9E-05
Q3.PM	-1.7E-05	3.2E-05	-5.5E-01	5.8E-01
CardinalWest	-1.3E-03	8.6E-05	-1.6E+01	1.9E-54
Month	7.6E-07	4.2E-08	1.8E+01	5.6E-73

Table 1: Linear regression summary results.

4 Conclusions

PM 2.5 levels are decreasing, specially after regulation was implemented in 2008, while Ozone levels have been increasing lately. Also, the East states have higher levels than the West states. However, an early analysis at using the means by county data to relate PM 2.5 to Ozone directly failed.

A Supplementary figures

Figure 4 shows the densities for the monthly mean per county for PM 2.5 and Ozone. It uses the complete case data. We can observe how the Ozone data looks fairly bell-curved and symmetric while the PM 2.5 data is highly skewed. In addition, the East has higher values than the West for both PM 2.5 and Ozone levels.

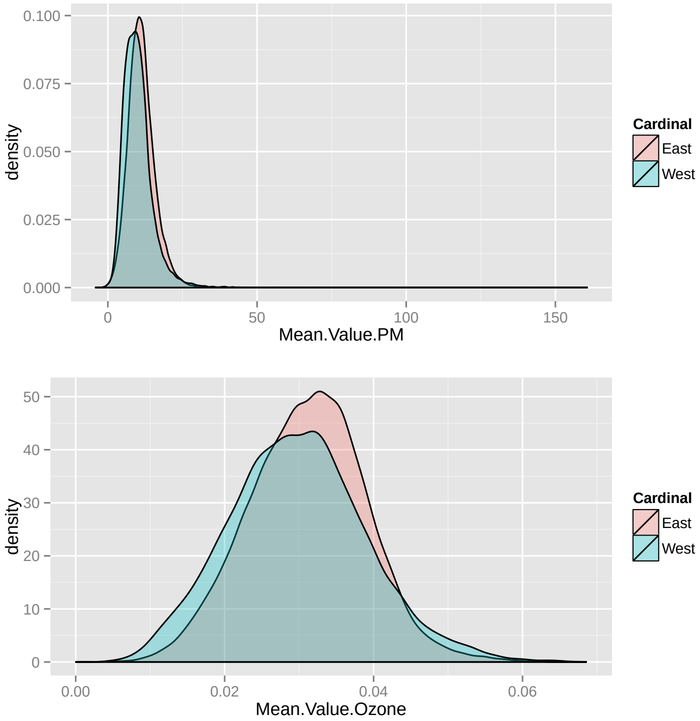

Figure 4: Densities of the mean monthly values per county for PM 2.5 and Ozone. Uses complete case data.

Figure 5 shows the trend for PM 2.5 for the counties in the state of Arkansas. The major outlier is from June 2011 from the county of Crittenden. I could not find a major event explaining this rise in PM 2.5 beyond seasonal wildfires. More at `http://alg.umbc.edu/usaq/archives/2011_06.html`.

Figure 6 is similar to Figure 2. I didn't know how to make the X-axis comparable in this version. Note that Figure 6 uses the complete case data which is why the values are a bit different compared to Figure 2.

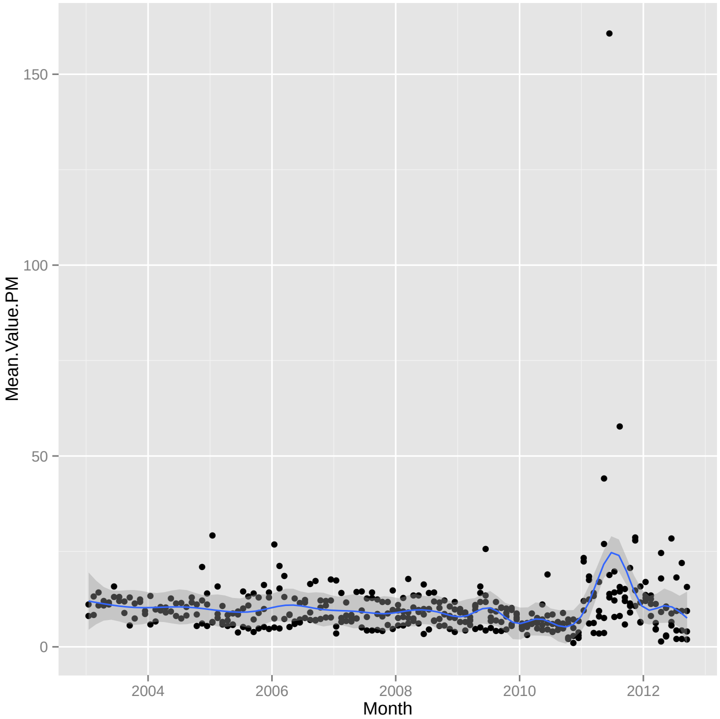

Figure 5: Mean PM 2.5 per month for Arkansas counties. Smooth line is from a natural spline with 20 degrees of freedom. Grey bands mark the 95% confidence bands. Uses complete cases.

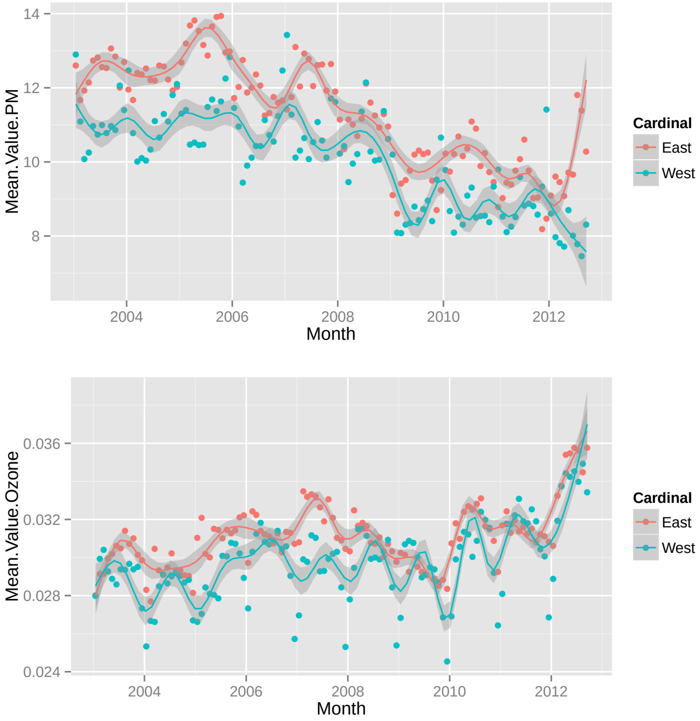

Figure 6: Mean PM 2.5 and Ozone levels per month from the counties. Smooth line is a from a natural spline with 20 degrees of freedom. Grey bands mark the 95% confidence bands. Uses complete cases.