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), 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

3 Conclusions

[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