Exploratory Data Analysis of PM2.5 Emissions in USA

Luis Talavera

July 14th 2022

Introduction

Fine particulate matter (PM2.5) is an ambient air pollutant for which there is strong evidence that it is harmful to human health. In the United States, the Environmental Protection Agency (EPA) is tasked with setting national ambient air quality standards for fine PM and for tracking the emissions of this pollutant into the atmosphere. Approximatly every 3 years, the EPA releases its database on emissions of PM2.5. This database is known as the National Emissions Inventory (NEI). More information about the NEI can be found at the EPA National Emissions Inventory web site.

For each year and for each type of PM source, the NEI records how many tons of PM2.5 were emitted from that source over the course of the entire year. The data that we will use for this assignment are for 1999, 2002, 2005, and 2008.

The overall goal of this data analysis is to explore the National Emissions Inventory database and see what it say about fine particulate matter pollution in the United states over the 10-year period 1999–2008.

Import libraries and get data

First, we import the libraries needed and load the datasets to study.

```
library(dplyr)
library(ggplot2)
```

```
data.dir <- "./data"
if(!file.exists(data.dir)) {
  dir.create(data.dir)
}
zip.url <- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
zip.file <- paste(data.dir, "dataset.zip", sep="/")
if(!file.exists(zip.file))
{
  download.file(zip.url, destfile = zip.file)
}
unzip(zip.file, exdir = data.dir)
NEI <- readRDS(paste(data.dir, "summarySCC_PM25.rds", sep="/"))
SCC <- readRDS(paste(data.dir, "Source_Classification_Code.rds", sep="/"))
```

The data

The datasets we will analyze are:

PM2.5 Emissions Data (**summarySCC_PM25.rds**): This file contains a data frame with all of the PM2.5 emissions data for 1999, 2002, 2005, and 2008. For each year, the table contains number of tons of PM2.5 emitted from a specific type of source for the entire year.

Rows and columns

```
dim(NEI)
```

```
## [1] 6497651       6
```

First few rows.

```
head(NEI)
```

```
##     fips      SCC Pollutant Emissions  type year
## 4  09001 10100401  PM25-PRI    15.714 POINT 1999
## 8  09001 10100404  PM25-PRI   234.178 POINT 1999
## 12 09001 10100501  PM25-PRI     0.128 POINT 1999
## 16 09001 10200401  PM25-PRI     2.036 POINT 1999
## 20 09001 10200504  PM25-PRI     0.388 POINT 1999
## 24 09001 10200602  PM25-PRI     1.490 POINT 1999
```

Last few rows.

```
tail(NEI)
```

```
##           fips        SCC Pollutant   Emissions     type year
## 75051171 56011 2282020005  PM25-PRI 0.028598300 NON-ROAD 2008
## 75051181 53009 2265003020  PM25-PRI 0.003152410 NON-ROAD 2008
## 75051191 41057 2260002006  PM25-PRI 0.046869500 NON-ROAD 2008
## 75051201 38015 2270006005  PM25-PRI 1.012890000 NON-ROAD 2008
## 75051211 46105 2265004075  PM25-PRI 0.000486488 NON-ROAD 2008
## 75051221 53005 2270004076  PM25-PRI 0.001622670 NON-ROAD 2008
```

- **fips**: A five-digit number (represented as a string) indicating the U.S. county
- **SCC**: The name of the source as indicated by a digit string (see source code classification table)
- **Pollutant**: A string indicating the pollutant
- **Emissions**: Amount of PM2.5 emitted, in tons
- **type**: The type of source (point, non-point, on-road, or non-road)
- **year**: The year of emissions recorded

Source Classification Code Table (**Source_Classification_Code.rds**): This table provides a mapping from the SCC digit strings in the Emissions table to the actual name of the PM2.5 source. The sources are categorized in a few different ways from more general to more specific and you may choose to explore whatever categories you think are most useful. For example, source "10100101" is known as "Ext Comb /Electric Gen /Anthracite Coal /Pulverized Coal".

Rows and columns.

```
dim(SCC)
```

```
## [1] 11717    15
```

First few rows and columns.

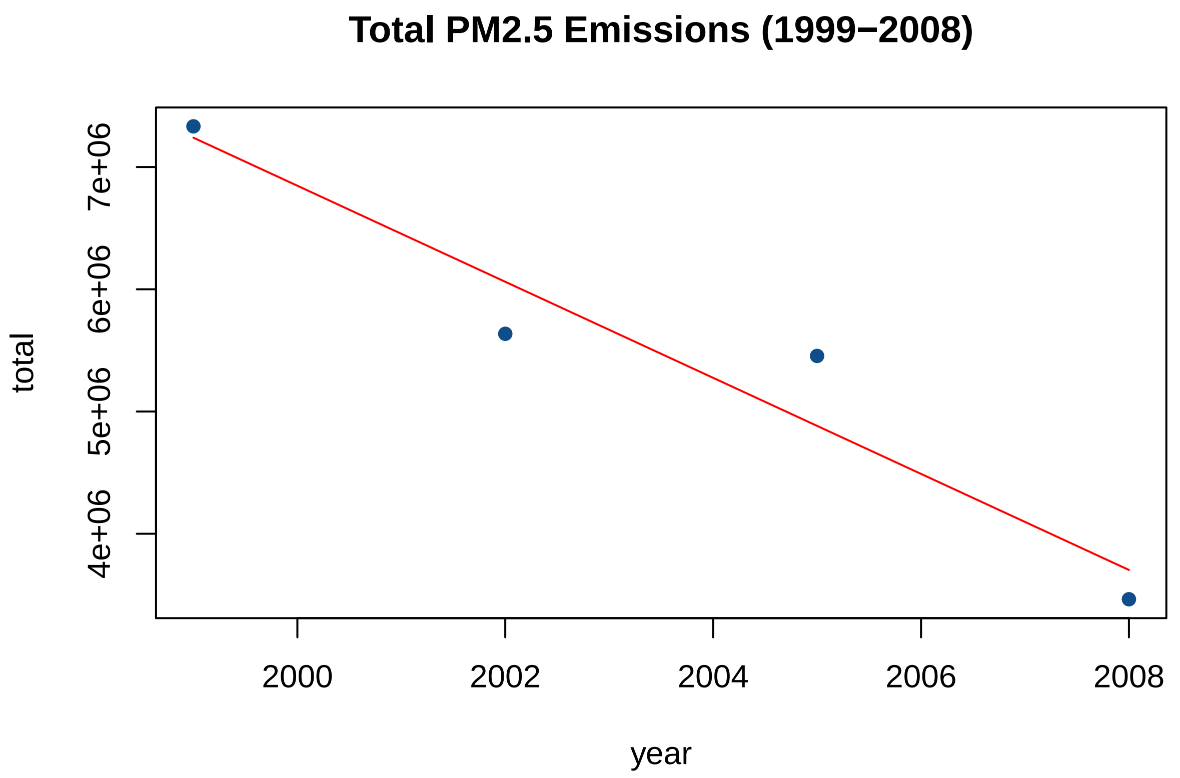
```
head(SCC[,1:5])
```

```
##        SCC Data.Category
## 1 10100101         Point
## 2 10100102         Point
## 3 10100201         Point
## 4 10100202         Point
## 5 10100203         Point
## 6 10100204         Point
##                                                                    Short.Name
## 1                   Ext Comb /Electric Gen /Anthracite Coal /Pulverized Coal
## 2 Ext Comb /Electric Gen /Anthracite Coal /Traveling Grate (Overfeed) Stoker
## 3       Ext Comb /Electric Gen /Bituminous Coal /Pulverized Coal: Wet Bottom
## 4       Ext Comb /Electric Gen /Bituminous Coal /Pulverized Coal: Dry Bottom
## 5                   Ext Comb /Electric Gen /Bituminous Coal /Cyclone Furnace
## 6                   Ext Comb /Electric Gen /Bituminous Coal /Spreader Stoker
##                                 EI.Sector Option.Group
## 1 Fuel Comb - Electric Generation - Coal
## 2 Fuel Comb - Electric Generation - Coal
## 3 Fuel Comb - Electric Generation - Coal
## 4 Fuel Comb - Electric Generation - Coal
## 5 Fuel Comb - Electric Generation - Coal
## 6 Fuel Comb - Electric Generation - Coal
```

Questions that have to be answered

1. Have total emissions from PM2.5 decreased in the United States from 1999 to 2008? Using the base plotting system, make a plot showing the total PM2.5 emission from all sources for each of the years 1999, 2002, 2005, and 2008.

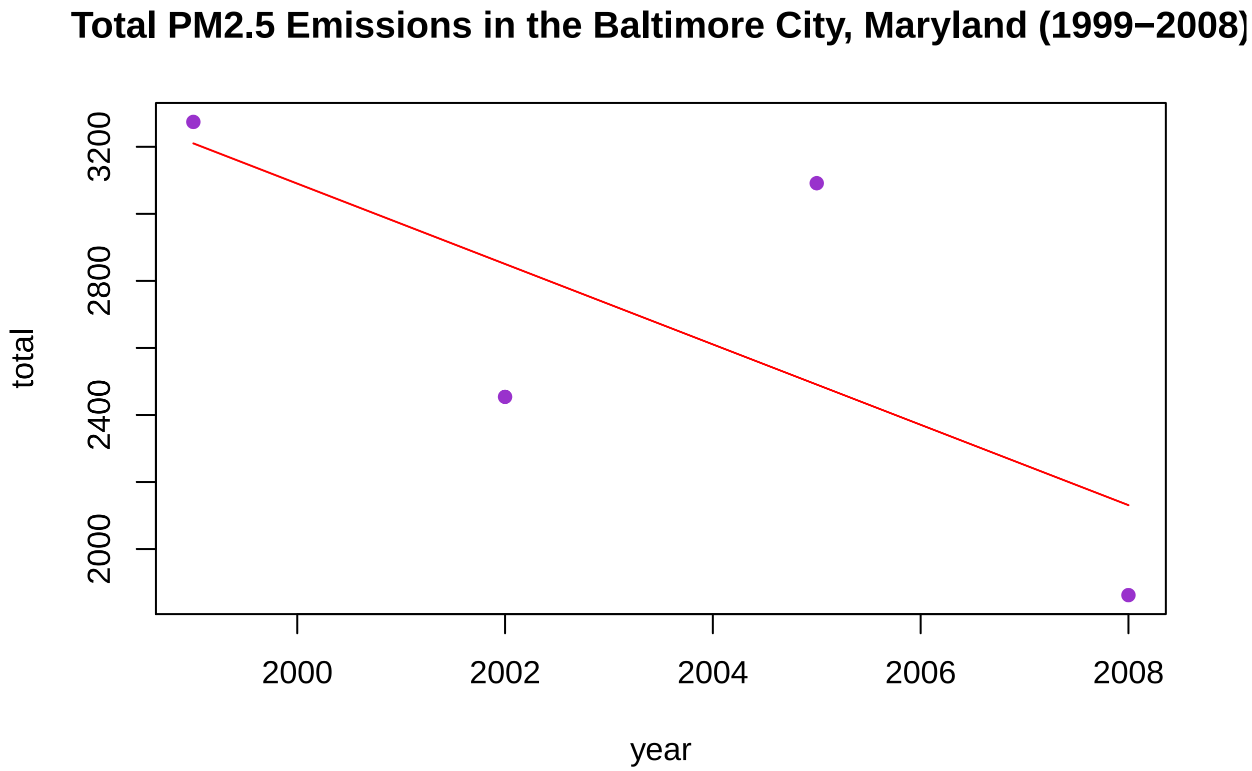
```
total.emissions.year <- NEI %>%
  group_by(year) %>%
  summarise(total = sum(Emissions))
plot(total.emissions.year,
     main="Total PM2.5 Emissions (1999-2008)",
     col = "dodgerblue4",
     pch = 16)
x <- total.emissions.year$year
y <- total.emissions.year$total
fit <- lm(y ~ x)
x0 <- seq(min(x), max(x), length = 10)
y0 <- predict.lm(fit, newdata = list(x = x0))
lines(x0, y0, col = 2)
```



As we can see by the red line the total emissions from PM2.5 have decreased.

2. Have total emissions from PM2.5 decreased in the Baltimore City, Maryland (fips == "24510") from 1999 to 2008? Use the base plotting system to make a plot answering this question.

```
total.emissions.year <- NEI %>%
  filter(fips == "24510") %>%
  group_by(year) %>%
  summarise(total = sum(Emissions))
plot(total.emissions.year,
     main="Total PM2.5 Emissions in the Baltimore City, Maryland (1999-2008)",
     col = "darkorchid",
     pch = 16)
x <- total.emissions.year$year
y <- total.emissions.year$total
fit <- lm(y ~ x)
x0 <- seq(min(x), max(x), length = 10)
y0 <- predict.lm(fit, newdata = list(x = x0))
lines(x0, y0, col = 2)
```



As we can see by the red line the total emissions from PM2.5 in Baltimore city have decreased.

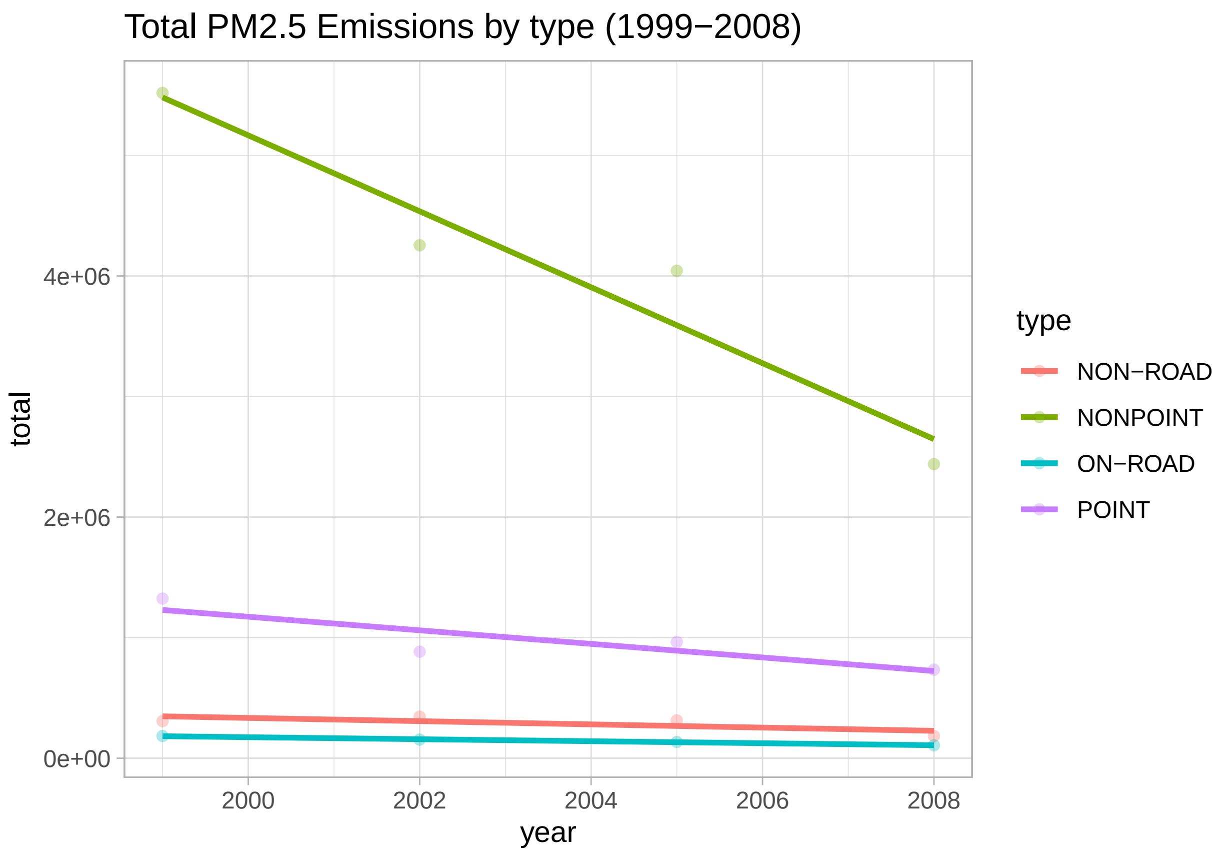
3. Of the four types of sources indicated by the type (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999–2008 for Baltimore City? Which have seen increases in emissions from 1999–2008? Use the ggplot2 plotting system to make a plot answer this question.

```
total.emissions.year <- NEI %>%
  group_by(year,type) %>%
  summarise(total = sum(Emissions))
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
g <- ggplot(total.emissions.year, aes(year, total))
g + geom_point(alpha=1/3, aes(color = type)) +
  geom_smooth(method="lm", se = FALSE, aes(color = type)) +
  theme_light() +
  labs(title = "Total PM2.5 Emissions by type (1999-2008)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

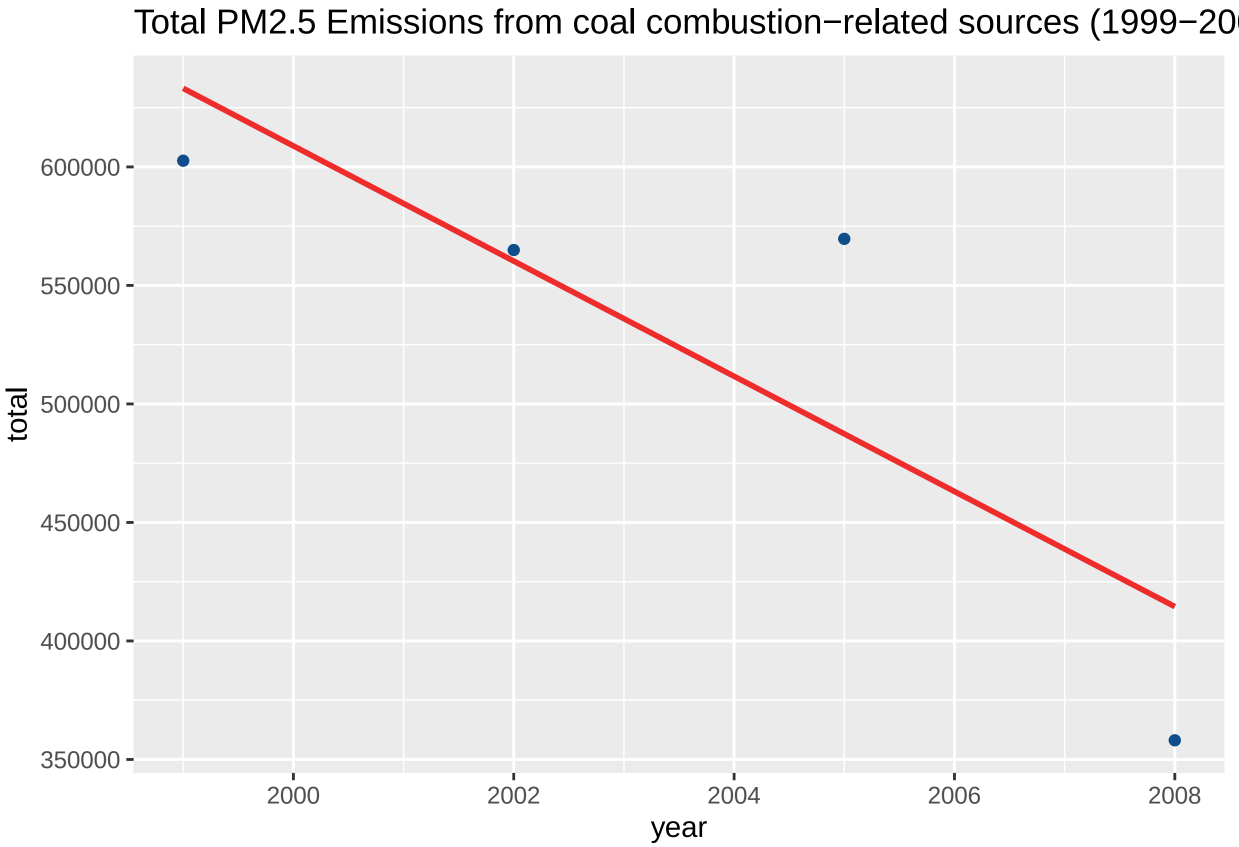


From the graph we can see that NON-POINT emissions have decreased significantly, POINT emissions have also decreased, but not as much as NON-POINT emissions, and NON-ROAD and ROAD emissions have hardly decreased at all.

4. Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?

```
coal.names <- grep('coal',SCC$Short.Name,value=TRUE,ignore.case=TRUE)
coal.scc <- SCC %>% filter(Short.Name %in% coal.names) %>% select(SCC)
NEI.coal <- filter(NEI, SCC %in% coal.scc$SCC)
total.emissions.year <- NEI.coal %>%
  group_by(year) %>%
  summarise(total = sum(Emissions))
g <- ggplot(total.emissions.year, aes(year, total))
g + geom_point(colour = "dodgerblue4") +
  geom_smooth(method="lm", se = FALSE, colour = "firebrick2") +
  theme(legend.position = "none") +
  labs(title = "Total PM2.5 Emissions from coal combustion-related sources (1999-2008)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

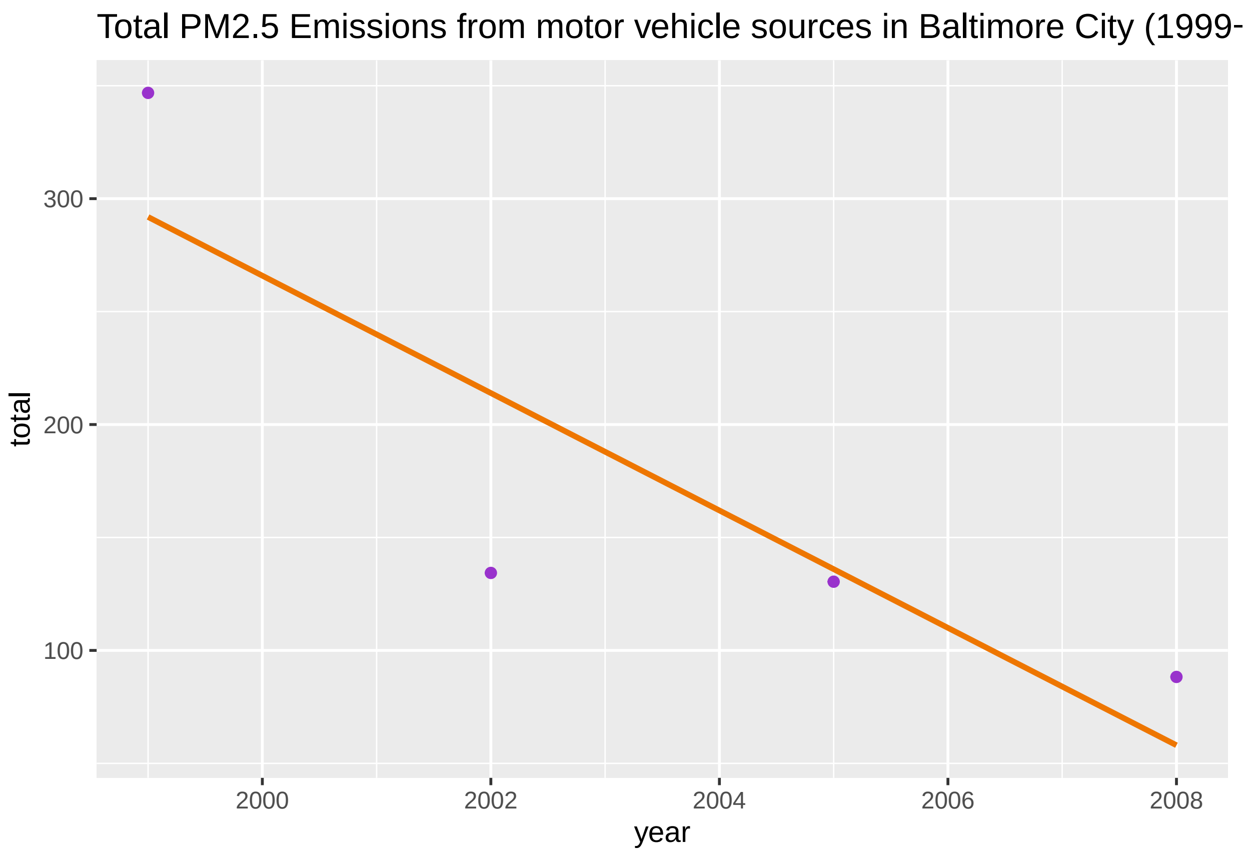


From the graphic, the emissions from coal combustion-related sources are decreasing.

5. How have emissions from motor vehicle sources changed from 1999–2008 in Baltimore City?

```r
mvehicle.names <- grep('vehicle',SCC$EI.Sector,value=TRUE,ignore.case=TRUE)
mvehicle.scc <- SCC %>% filter(EI.Sector %in% mvehicle.names) %>% select(SCC)
NEI.mvehicle <- filter(NEI, SCC %in% mvehicle.scc$SCC & fips == "24510")
total.emissions.year <- NEI.mvehicle %>%
  group_by(year) %>%
  summarise(total = sum(Emissions))
g <- ggplot(total.emissions.year, aes(year, total))
g + geom_point(colour = "darkorchid") +
  geom_smooth(method="lm", se = FALSE, colour = "darkorange2") +
  theme(legend.position = "none") +
  labs(title = "Total PM2.5 Emissions from motor vehicle sources in Baltimore City (1999-2008)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the graphic, the emissions from motor vehicle sources in Baltimore City are decreasing.

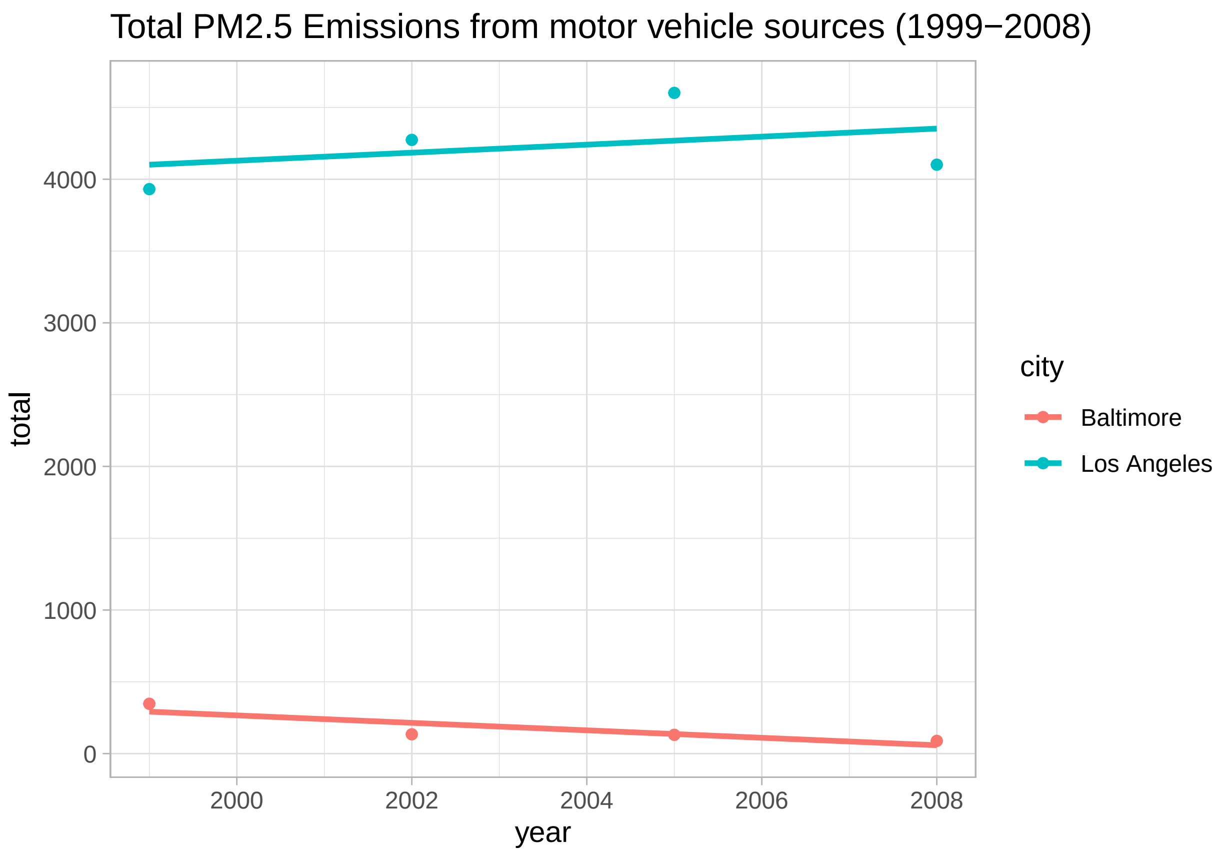
6. Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California (fips == "06037"). Which city has seen greater changes over time in motor vehicle emissions?

```
NEI.mvehicle <- filter(NEI, SCC %in% mvehicle.scc$SCC & (fips == "06037" | fips == "24510"))
cities <- list("06037" = "Los Angeles", "24510" = "Baltimore")
NEI.mvehicle <- mutate(NEI.mvehicle, city = as.character(cities[fips]))
total.emissions.year <- NEI.mvehicle %>%
  group_by(year,city) %>%
  summarise(total = sum(Emissions))
```

```
## ‘summarise()‘ has grouped output by ’year’. You can override using the ‘.groups‘ argument.
```

```
g <- ggplot(total.emissions.year, aes(year, total))
g + geom_point(aes(color = city)) +
  geom_smooth(method="lm", se = FALSE, aes(color = city)) +
  theme_light() +
  labs(title = "Total PM2.5 Emissions from motor vehicle sources (1999-2008)")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Emissions in Los Angeles are signifincantly higher than Emissions in Baltimore, also Baltimore Emissions levels are decreasing but on the other hand Los Angeles Emissions levels are increaseing

Conclusion

We can conclude, in general, total PM2.5 Emissions in the United States have decreased through the time including all emissions sources and cities. But it should be noted that emissions levels in Los Angeles are significantly high and have been increasing over the years, perhaps due to population growth, although there is no information that allows us to offer a justification for the levels and the increase.