Exploratory Data Analysis

Konstantin Burkin

2022-04-10

Fine particulate matter pollution

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). 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.

Dataframe

- The first file contains information of the location, year, source, and emission quantity. The columns of the dataframe are:
  - 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
- The second file contains information of a particular emission source for the corresponding SCC code.

```
# download data frame
url <- "https://d396qusza40orc.cloudfront.net/exdata%2Fdata%2FNEI_data.zip"
dir.create("data")
```

```
## Warning in dir.create("data"): 'data' already exists
```

```
download.file(url = url, destfile = "./data/file.zip")
unzip(zipfile = "./data/file.zip", exdir = "./data")

# read the data
NEI <- readRDS("./data/summarySCC_PM25.rds")
SCC <- readRDS("./data/Source_Classification_Code.rds")

# upload libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

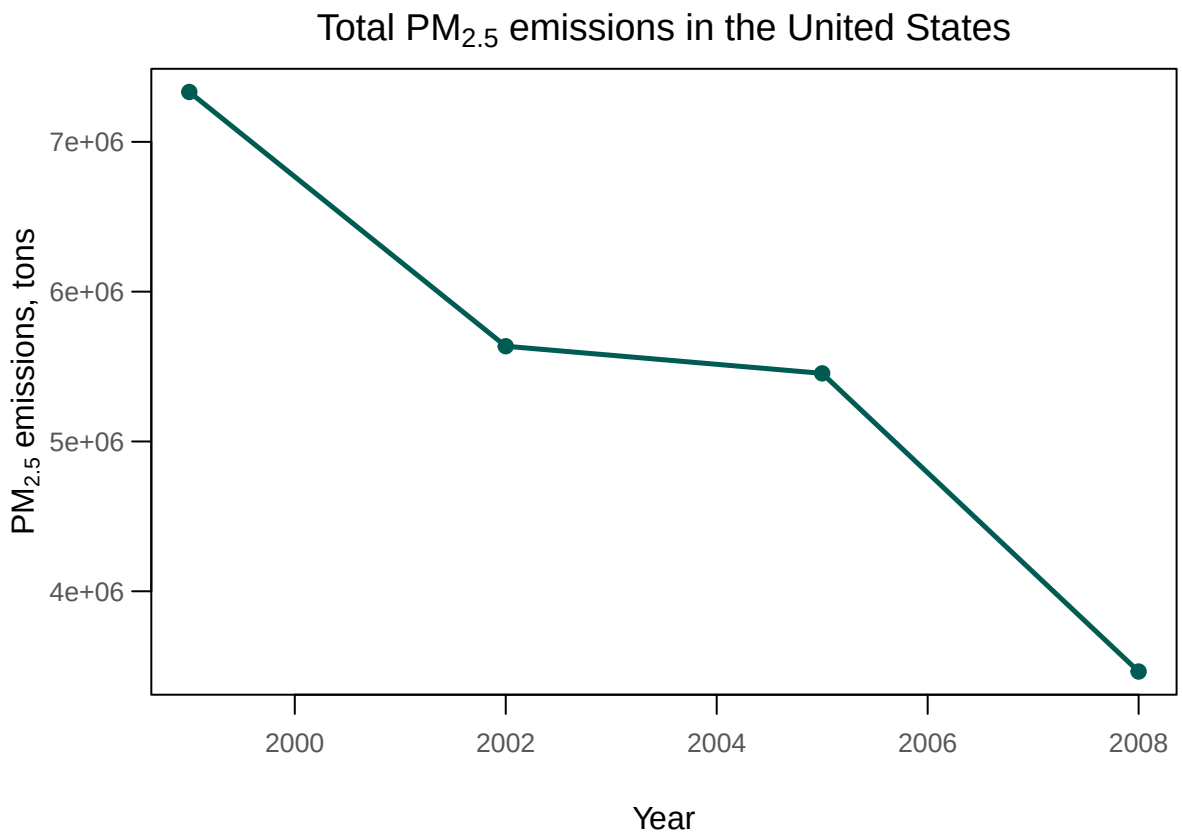
```
str(NEI)
```

```
## 'data.frame':    6497651 obs. of  6 variables:
##  $ fips     : chr  "09001" "09001" "09001" "09001" ...
##  $ SCC      : chr  "10100401" "10100404" "10100501" "10200401" ...
##  $ Pollutant: chr  "PM25-PRI" "PM25-PRI" "PM25-PRI" "PM25-PRI" ...
##  $ Emissions: num  15.714 234.178 0.128 2.036 0.388 ...
##  $ type     : chr  "POINT" "POINT" "POINT" "POINT" ...
##  $ year     : int  1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...
```

Including Plots

1. Have total emissions from $PM_{2.5}$ decreased in the United States from 1999 to 2008?

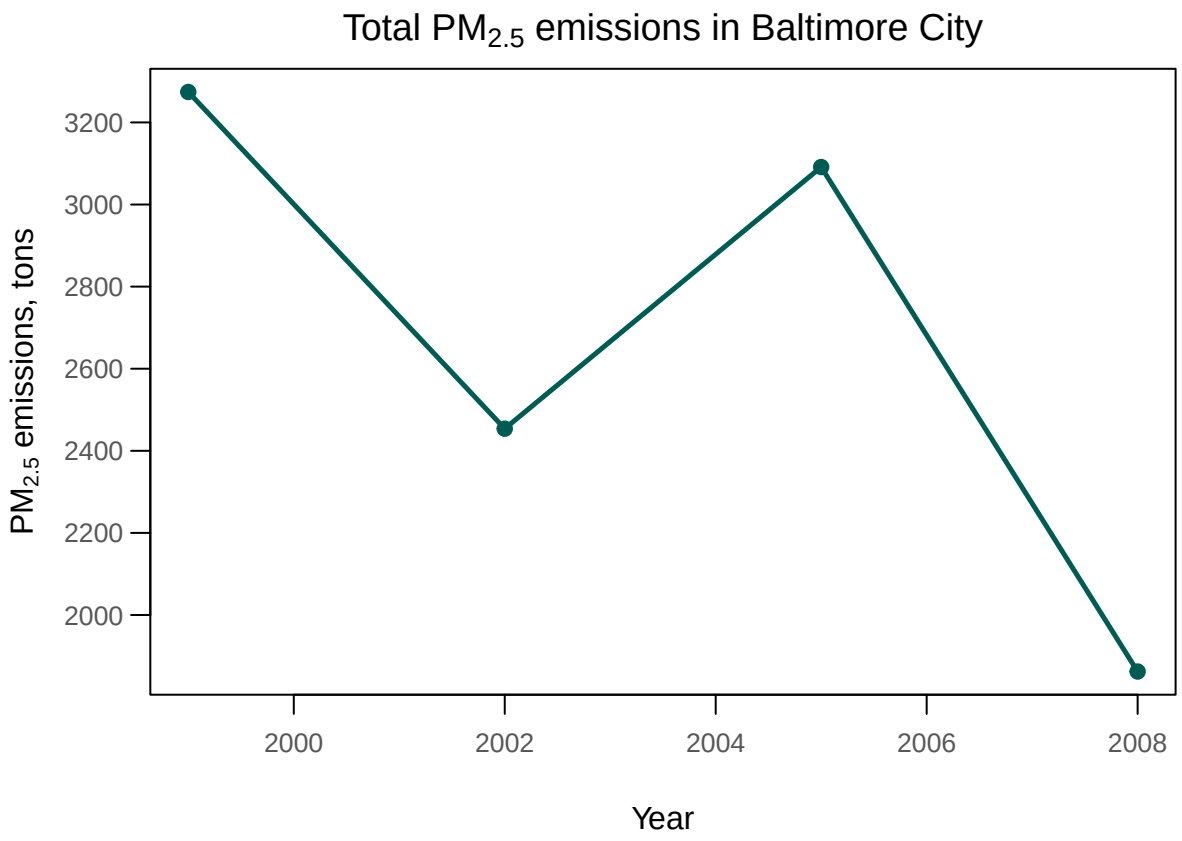
```
# extract data to visualize
# calculate sum of annual emissions across the US
total_emmisions <- tapply(NEI$Emissions, NEI$year, sum)

# make plot using base plotting system
par(las=1, oma=c(0,0,0,0), mar=c(4.2,4.8,2,1), cex.axis = 0.8, mgp=c(2.7,0.7,0))

plot(x = unique(NEI$year), y = total_emmisions,
     main = expression("Total PM"[2.5]*" emissions in the United States"),
     font.main = 2,
     col.axis = "#575757",
     ylab = expression("PM"[2.5]*" emissions, tons"),
     xlab = "Year",
     pch=19,
     col="#005C53",
     cex=1)
lines(unique(NEI$year), total_emmisions, col="#005C53", lwd=2.5)
```



The plot proves that total $PM_{2.5}$ decreased in the United States from 1999 to 2008. This can be the reason to the development of better emission filters for cars and factories.

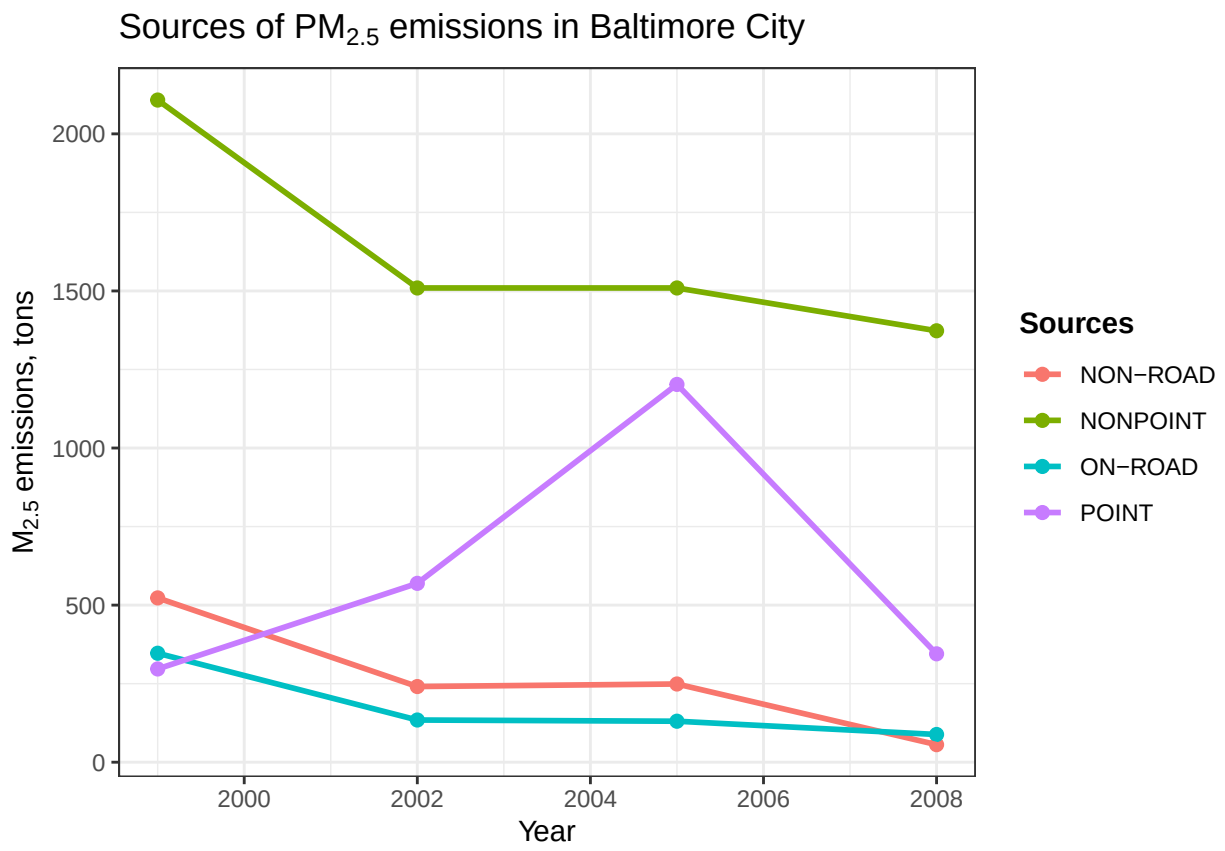
2. Have total emissions from $PM_{2.5}$ decreased in Baltimore City, Maryland from 1999 to 2008?

```
# extract data to visualize
# calculate sum of annual emissions in Baltimore City
total_emmisions <- tapply(NEI$Emissions[NEI$fips=="24510"], NEI$year[NEI$fips=="24510"], sum)

# make plot using base plotting system
par(las=1, oma=c(0,0,0,0), mar=c(4.2,4.8,2,1), cex.axis = 0.8, mgp=c(2.7,0.7,0))

plot(x = unique(NEI$year), y = total_emmisions,
     main = expression("Total PM"[2.5]*" emissions in Baltimore City"),
     font.main = 2,
     col.axis = "#575757",
     ylab = expression("PM"[2.5]*" emissions, tons"),
     xlab = "Year",
     pch=19,
     col="#005C53",
     cex=1
)
lines(unique(NEI$year), total_emmisions, col="#005C53", lwd=2.5)
```



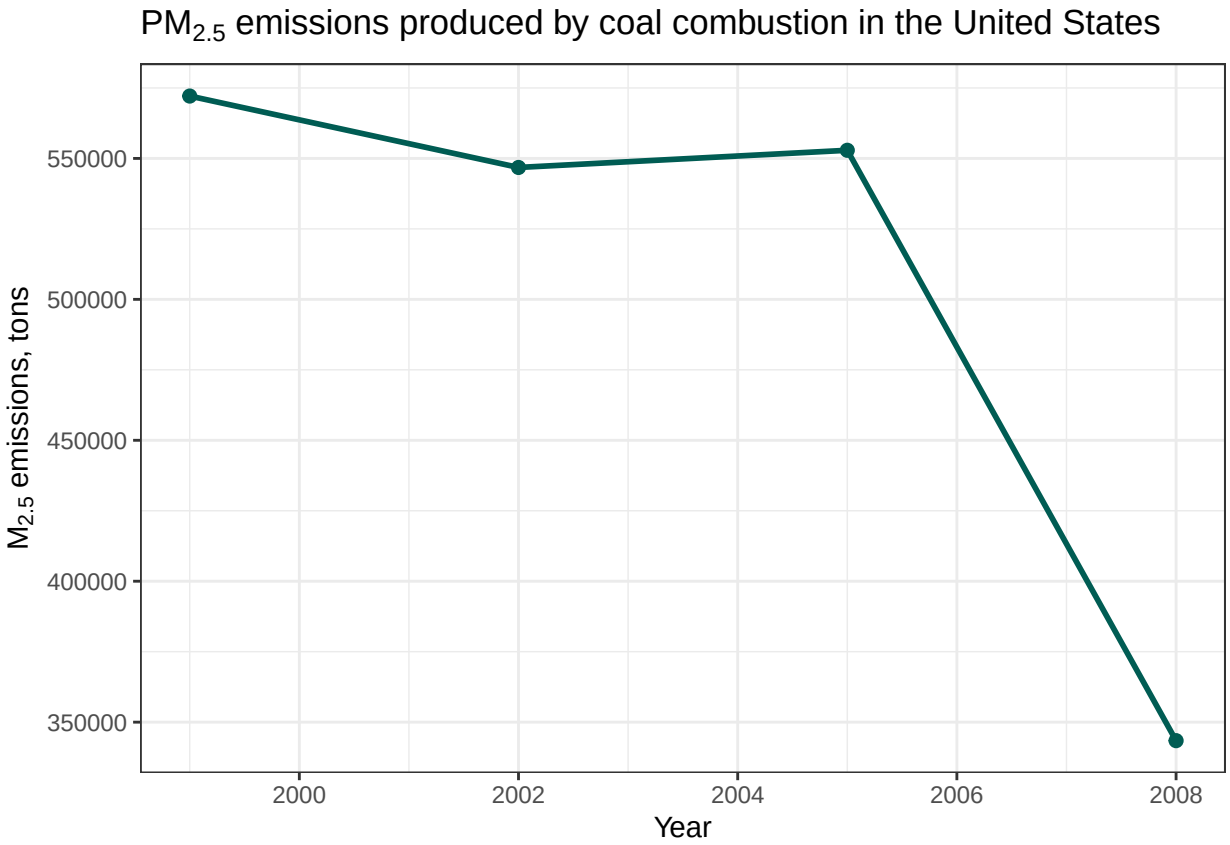
3. Which of these four sources (point, nonpoint, onroad, nonroad) have seen decreases in emissions from 1999–2008 for Baltimore City?

```
# extract data to visualize
# calculate sum of annual emissions in Baltimore City for each emission source
NEI_subset <- subset(x = NEI, subset = NEI$fips=="24510", select = c(Emissions, type, year))
total_emmisions <- aggregate(NEI_subset$Emissions, by = list(NEI_subset$type, NEI_subset$year), FUN = su

# make plot using ggplot2
ggplot(data = total_emmisions,
       aes(x = Group.2, y = x, color=Group.1) ) +
       geom_point(size=2) +
       geom_line(size=1) +
       labs(
       title = expression("Sources of PM"[2.5]*" emissions in Baltimore City"),
       x = "Year",
       y = expression("M"[2.5]*" emissions, tons")
       ) +
       theme_bw() +
       scale_colour_discrete(name = "Sources") +
       theme(legend.title = element_text(face = "bold"))
```



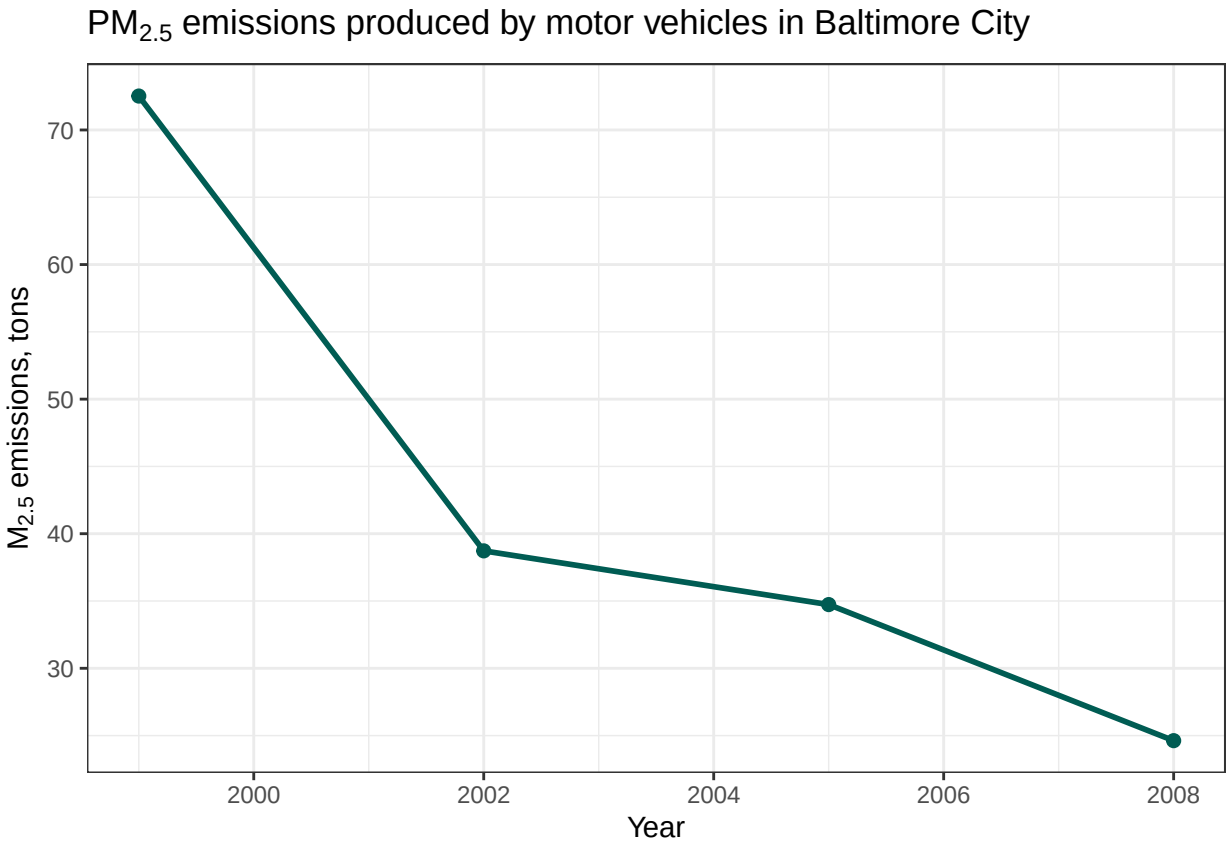
4. Across the United States, how have emissions from coal combustion-related sources changed from 1999–2008?

```
# extract data to visualize
# calculate sum of annual coal emissions in the US
# use dplyr %>% function to make pipeline
df <- NEI %>% subset(select = c(Emissions, SCC, year)) %>%
      merge(SCC[,c("SCC", "EI.Sector")], by = "SCC") %>%
      subset(select = c("Emissions", "year", "EI.Sector"))
df <- filter(.data = df, grepl(pattern = '[Cc]oal', x = df$EI.Sector))
df <- aggregate(df$Emissions, by = list(df$year), FUN = sum)

# make plot using ggplot2
ggplot(data = df, aes(x = Group.1, y = x)) +
      geom_point(size=2, col="#005C53") +
      geom_line(size=1,  col="#005C53") +
      labs(
        title = expression("PM"[2.5]*" emissions produced by coal combustion in the United States"),
      x = "Year",
      y = expression("M"[2.5]*" emissions, tons")
      ) +
  theme_bw()
```



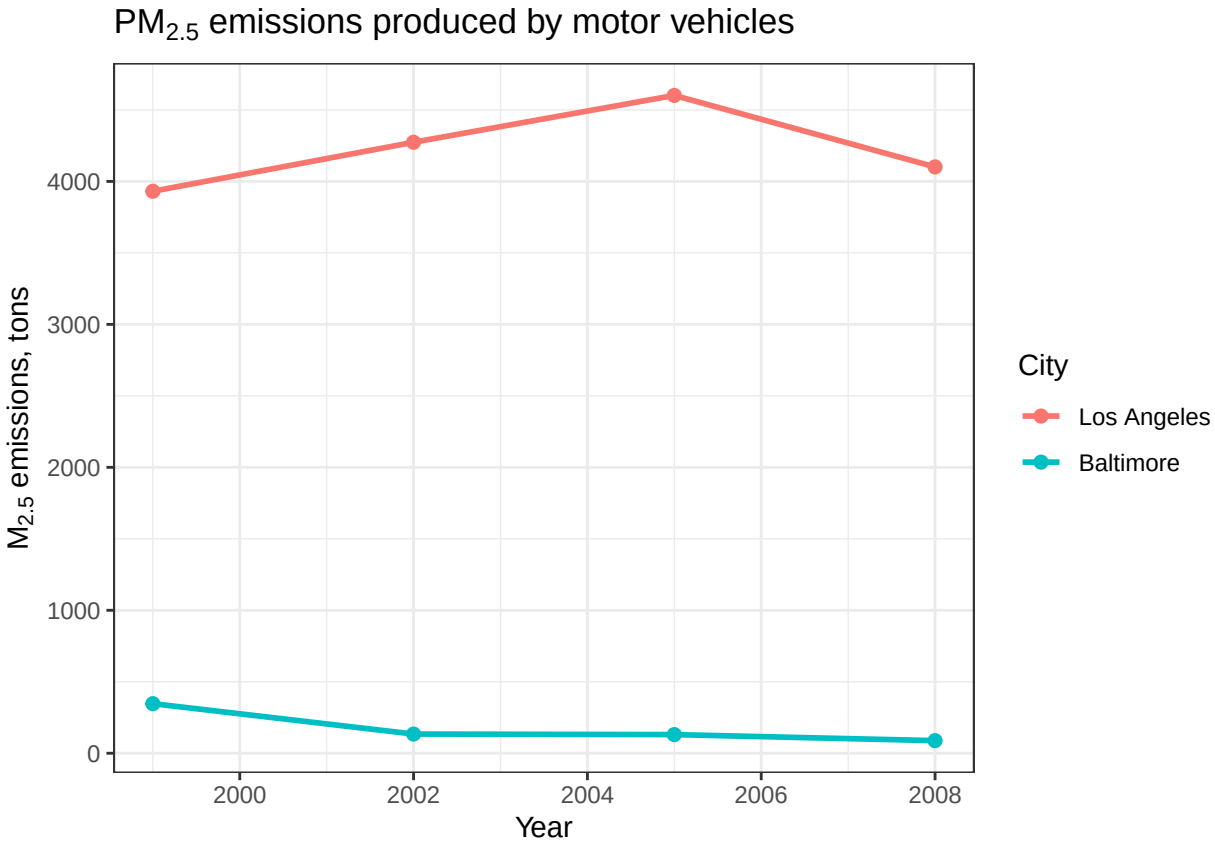
5. How have emissions from motor vehicles changed from 1999–2008 in Baltimore City?

```
# extract data to visualize
# calculate sum of annual vehicle emissions in the Baltimore City
df <- NEI %>% subset(subset=(fips=="24510"), select = c(Emissions, SCC, year)) %>%
  merge(SCC[,c("SCC", "Short.Name")], by = "SCC") %>%
  subset(select = c("Emissions", "year", "Short.Name"))
df <- filter(.data = df, grepl(pattern = "[Vv]ehicle", x = df$Short.Name))
df <- aggregate(df$Emissions, by = list(df$year), FUN = sum)

# make plot using ggplot2
ggplot(data = df, aes(x = Group.1, y = x)) +
       geom_point(size=2, col="#005C53") +
       geom_line(size=1,  col="#005C53") +
       labs(title = expression("PM"[2.5]*" emissions produced by motor vehicles in Baltimore City"),
       x = "Year",
       y = expression("M"[2.5]*" emissions, tons")) +
       theme_bw()
```



6. Is there any difference in emissions from motor vehicles between Baltimore City and Los Angeles?Which city has seen greater changes over time in motor vehicle emissions?

```
# extract data to visualize
# calculate sum of annual vehicle emissions in the Baltimore City and Los Angeles
df <- NEI %>% subset(subset = ((fips == "24510" | fips == "06037") & NEI$type == "ON-ROAD"),
                                select = c(Emissions, year, fips))
df1 <- aggregate(Emissions ~ year + fips, df, FUN = sum)


# make plot using ggplot2
ggplot(data = df1, aes(x = year, y = Emissions, color= fips)) +
       geom_point(size=2) +
       geom_line(size=1) +
       labs(
       title = expression("PM"[2.5]*" emissions produced by motor vehicles"),
       x = "Year",
       y = expression("M"[2.5]*" emissions, tons")) +
       scale_colour_discrete(name = "City", labels = c("Los Angeles", "Baltimore")
       ) +
       theme_bw()
```


PM2.5 emissions produced by motor vehicles
M2.5 emissions, tons
4000
3000
2000
1000
0
2000
2002
2004
2006
2008
Year
City
Los Angeles
Baltimore