Exploratory Data Analysis - Particles Emission

Mathieu Besancon

Tuesday, September 09, 2014

Introduction

Image source here.

This report aims at giving a quick overview of the particle emissions in the USA, based on the exploratory analysis of data provided by the the Environmental Protection Agency (EPA) and made publicly available. We will use data from the years 1999, 2002, 2005, and 2008.

```
grid.raster(img)
```

The purpose is to put in a proper report the graphs produced for the second project of *Exploratory Data Analysis*, an online course offered by the John Hopkins University on the Coursera plateform. You can have an overview of the course here.

Pre-processing

We begin by loading the data :

```
part_data=readRDS("summarySCC_PM25.rds")
SCC <- readRDS("Source_Classification_Code.rds")
```

We store in *part_data* the particle emissions data and in SCC the Source Classification Code of each emission type. For some of the plots, we will need the package *ggplot2*. We will also be using the function *multiplot* for this package, available here.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.0.3
```

The function multiplot is hidden to keep the document short. You can visite the website to see the details, the code remained untouched.

We can quickly look at the data to see on what exactly we are working : What are the attributes of the particle emissions data :

```
names(part_data)
```

```
## [1] "fips"      "SCC"       "Pollutant" "Emissions" "type"      "year"
```

How big is our data set :

```
dim(part_data)
```

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

What contains the first line :

```
part_data[1,]
```

```
##    fips      SCC Pollutant Emissions  type year
## 4 09001 10100401  PM25-PRI     15.71 POINT 1999
```

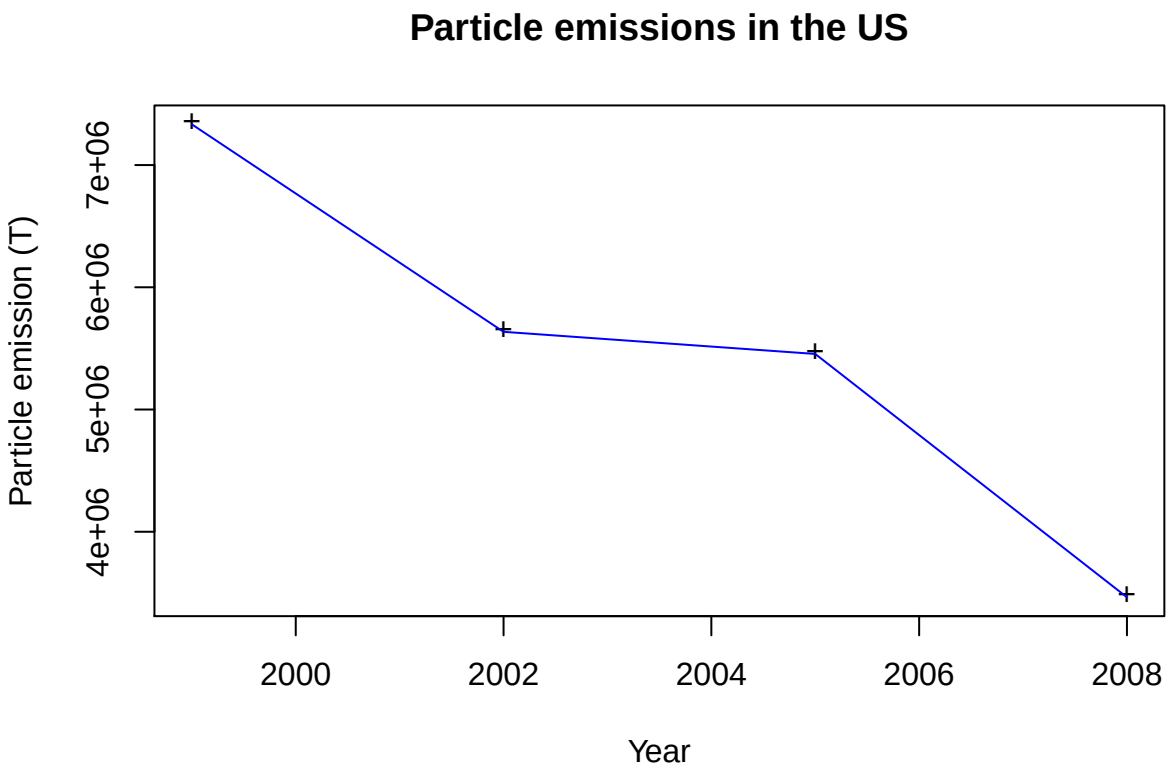
Plot I : Overall particle emissions in the US across the years

First of all, we aggregate the emissions data per year.

```
agg_part<-aggregate(part_data$Emissions,list(part_data$year),FUN=sum)
```

The base plot system is used to create a simple plot of the data :

```
plot(agg_part$Group.1,agg_part$x,main="Particle emissions in the US",xlab="Year",ylab="Particle emissio
lines(agg_part$Group.1,agg_part$x,col="blue")
```



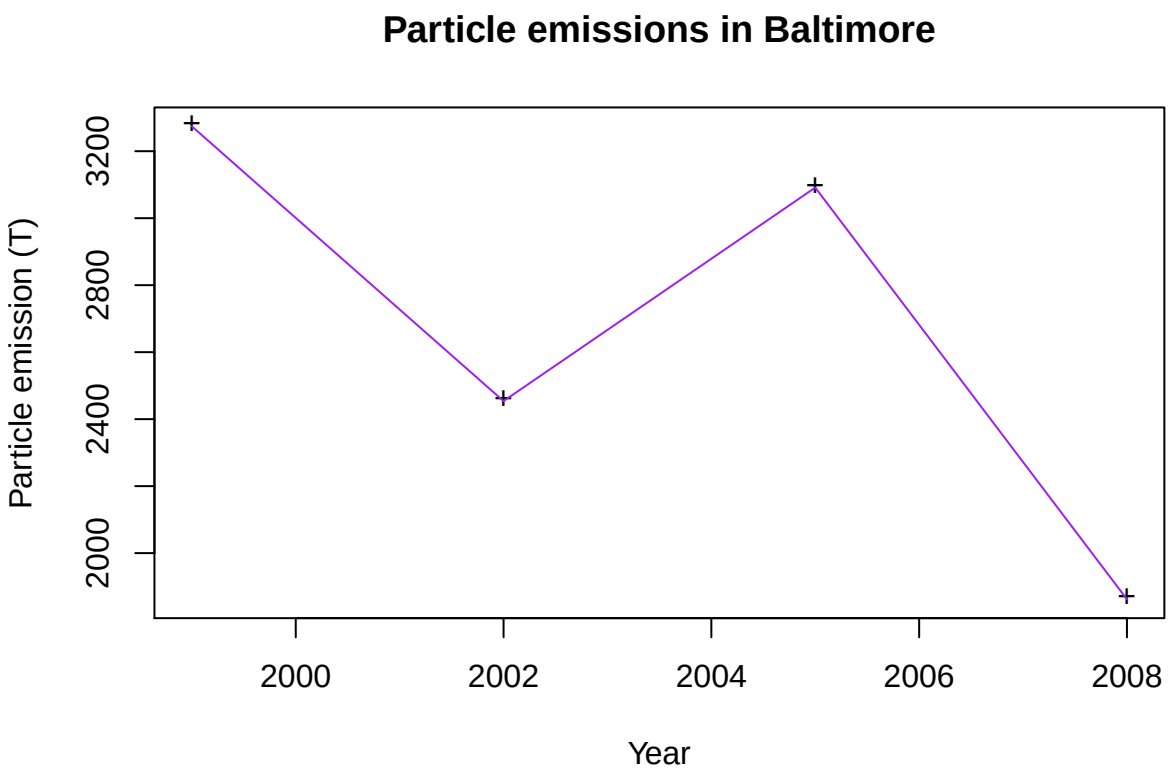
Plot II : Particle emissions per year in Baltimore

Baltimore is identified by the fips code "24510". We create a subset of the data only with the corresponding fips code, and aggregate the emissions per year.

```
balti_data<-subset(part_data,fips=="24510")
agg_balti<-aggregate(balti_data$Emissions,list(balti_data$year),FUN=sum)
```

We can then generate the plot.

```
plot(agg_balti$Group.1,agg_balti$x,main="Particle emissions in Baltimore",xlab="Year",ylab="Particle emi
lines(agg_balti$Group.1,agg_balti$x,col="purple")
```



Plot III : Emissions per type in Baltimore across years

Firstly, we each type from the Baltimore data we created for Plot 2.

```
sub_point<-subset(balti_data,type=='POINT')
sub_non_point<-subset(balti_data,type=='NONPOINT')
sub_onroad<-subset(balti_data,type=='ON-ROAD')
sub_non_road<-subset(balti_data,type=='NON-ROAD')
```

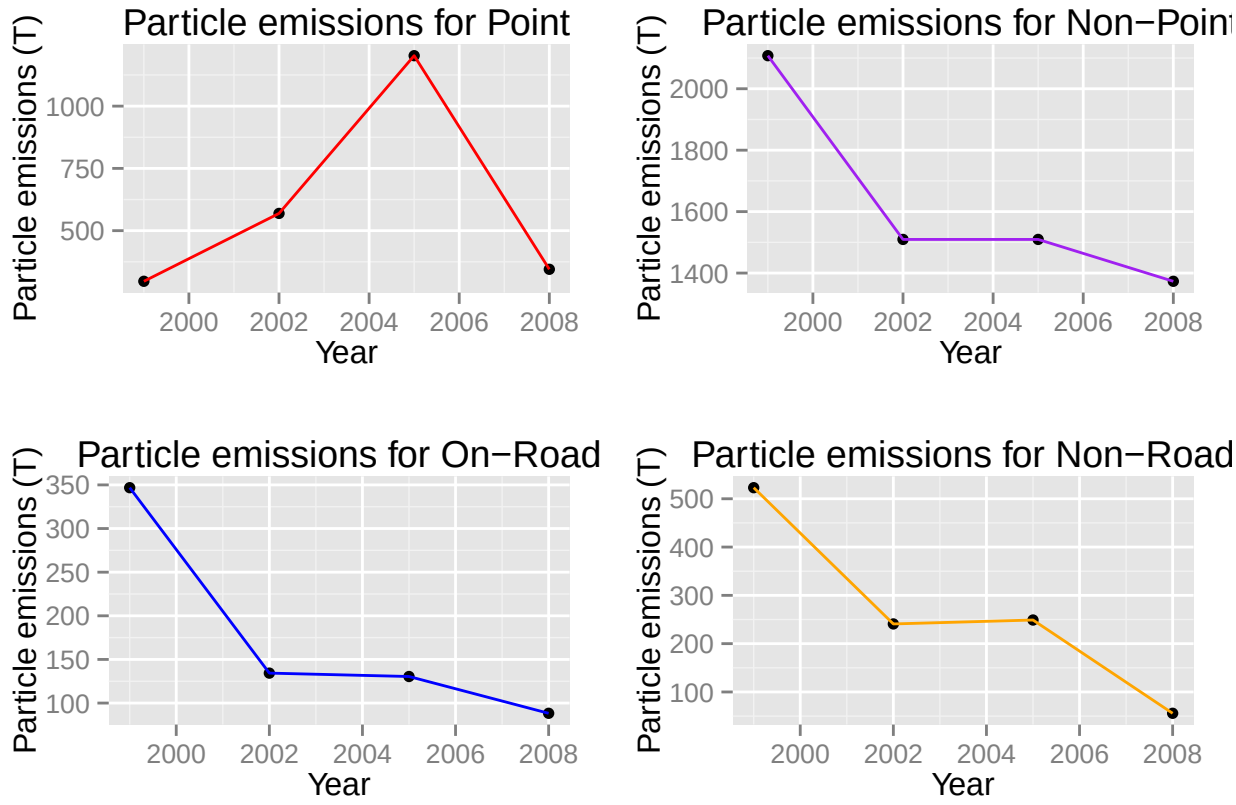
The data of each type are aggregated per year :

```
agg_point<-aggregate(sub_point$Emissions,list(sub_point$year),FUN=sum)
agg_non_point<-aggregate(sub_non_point$Emissions,list(sub_non_point$year),FUN=sum)
agg_onroad<-aggregate(sub_onroad$Emissions,list(sub_onroad$year),FUN=sum)
agg_non_road<-aggregate(sub_non_road$Emissions,list(sub_non_road$year),FUN=sum)
```

In order to display all the plot at once, we store them in objects, and call the multiplot function with these objects as arguments.

```
plot_point<-qplot(Group.1,x,main="Particle emissions for Point",data=agg_point,xlab="Year",ylab="Particl
plot_onroad<-qplot(Group.1,x,main="Particle emissions for On-Road",data=agg_onroad,xlab="Year",ylab="Pa
plot_non_point<-qplot(Group.1,x,main="Particle emissions for Non-Point",data=agg_non_point,xlab="Year",
plot_non_road<-qplot(Group.1,x,main="Particle emissions for Non-Road",data=agg_non_road,xlab="Year",ylab

multiplot(plot_point,plot_onroad,plot_non_point,plot_non_road,cols=2)
```



Plot IV : Emissions from coal combustion

In order to find the emissions codes related to coal combustion, we look for the emissions types names containing "Coal-fired". We extract the indexes of these emissions categories and create a subset of the classification data with those indexes.

```
index_coal<-which(grepl('Coal-fired',SCC$Short.Name)==TRUE)
sub_coal<-SCC[index_coal,]
```

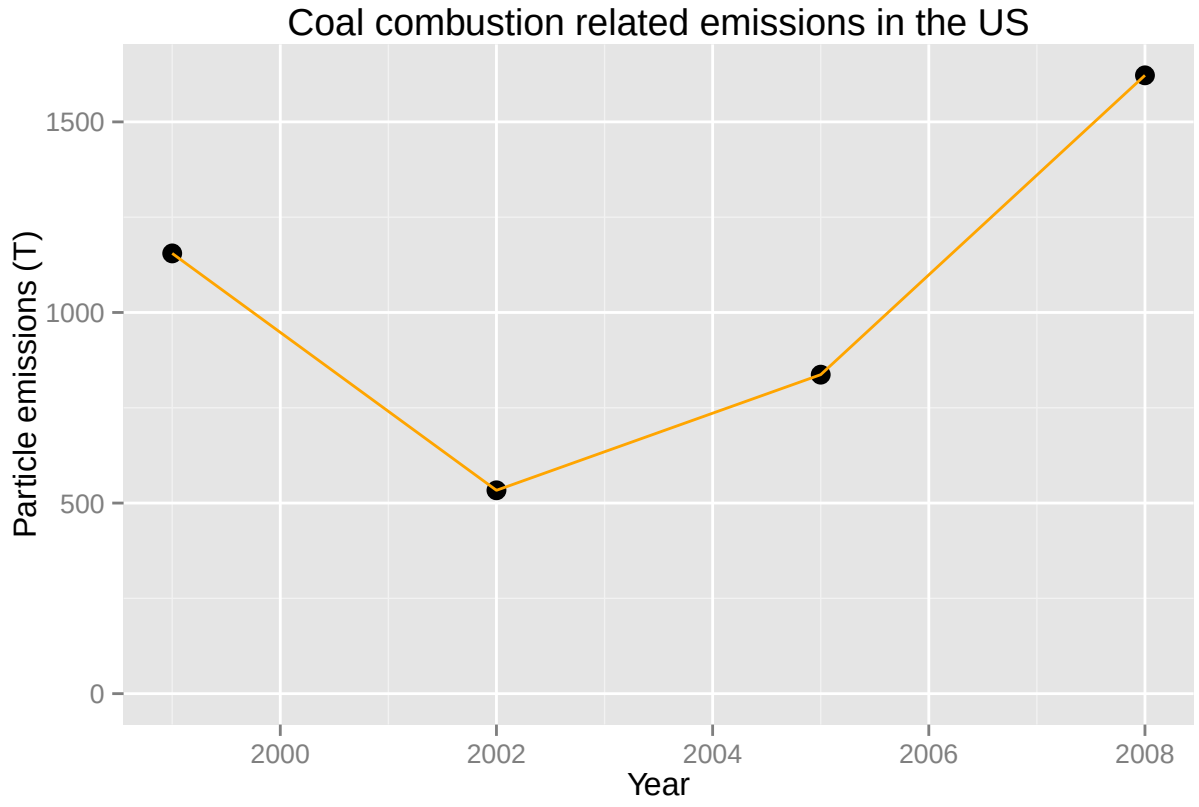
Then, a subset of the emissions data, *coal_data* can be generated based on their SCC code, which has to belong to the coal-fired categories.

```
coal_data<-part_data[which(part_data$SCC%in%sub_coal$SCC),]
```

The *coal_data* subset can then be aggregated per year as done before, and the plot can be generated.

```
## L'objet suivant est masqué _by_ .GlobalEnv:
##
##     SCC
```

```
p<-qplot(agg_coal[,1],agg_coal[,2],ylim=c(0,NA),main='Coal combustion related emissions in the US',
xlab="Year",ylab="Particle emissions (T)",size=5)
p<-p+geom_line(size=0.5,col="orange")
p<-p+theme(legend.position="none")
print(p)
```



Plot V : Motor vehicles emissions in Baltimore

We will reuse the *balti_data* subset but this time, we extract the emissions due to motor vehicles. Just as we did for coal, we extract the relevant classification codes from the SCC data, and use it to subset the *balti_data*. We name this subset *car_data1* :

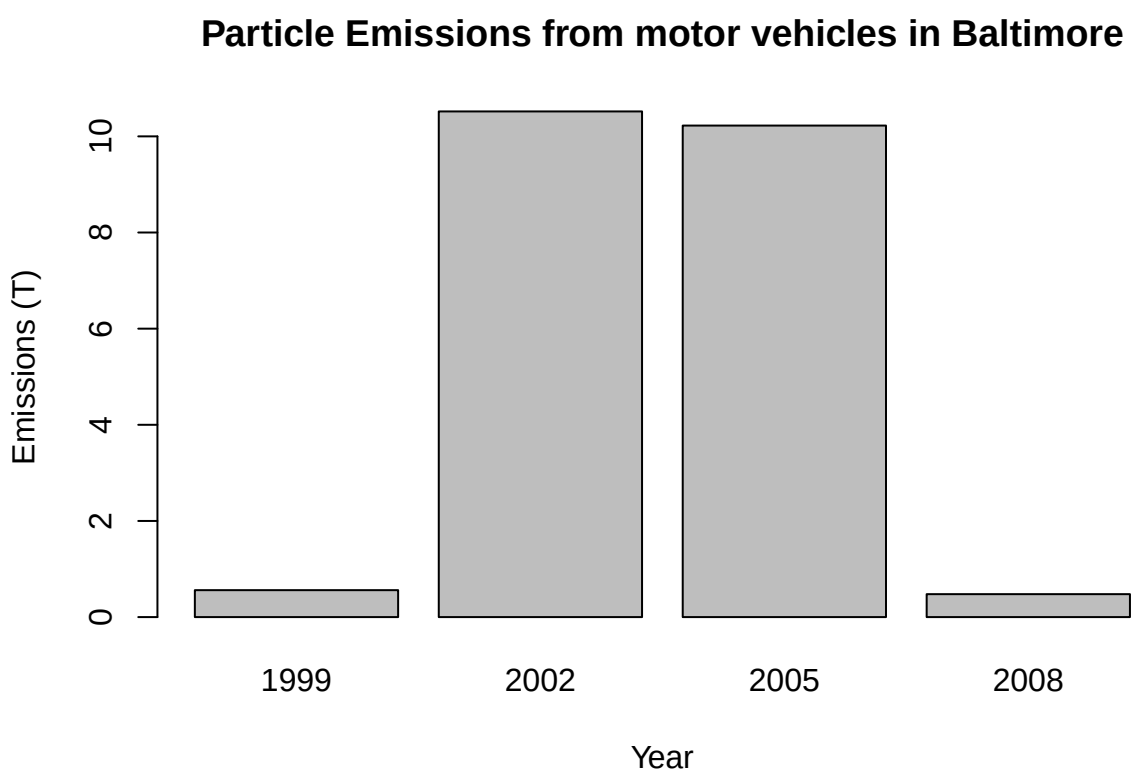
```
index_car<-which(grepl('Motor',SCC$Short.Name)==TRUE)
sub_car<-SCC[index_car,]

car_data1<-balti_data[which(balti_data$SCC%in%sub_car$SCC),]
```

As usual, we aggregate our data per year into *agg_car* and plot the result. We use this time the bar plot from the base plotting system for this purpose :

```
## L'objet suivant est masqué _by_ .GlobalEnv:
##
##     SCC
```

```
barplot(agg_car1[,2],names.arg=(agg_car1[,1]),xlab='Year',ylab='Emissions (T)',
main='Particle Emissions from motor vehicles in Baltimore')
```



Plot VI : Comparing motor vehicles emissions from Baltimore and Los Angeles County

We would like to discover which of these two cities had the greater change in particle emissions related to motor vehicles.

We extract the emissions data from the two cities, for which we know the fips code, respectively 24510 und 06037.

```
two_cities_data<-subset(part_data,(fips=="24510")|(fips=="06037"))
```

We create the subset *car_data* from the *two_cities_data* with only the relevant emissions records, and aggregate it in *agg_car*. We create a new attribute in this subset, **City**, which contains the string "Log Angeles" or "Baltimore", depending on the fips code.

```
car_data<-two_cities_data[which(two_cities_data$SCC%in%sub_car$SCC),]
agg_car<-aggregate(Emissions~year+fips,data=car_data,FUN=sum)
agg_car$City<-NA
for (i in 1:length(agg_car[,1])){
     if (agg_car[i,2]=="24510"){
           agg_car[i,4]<-"Baltimore"
     }else{agg_car[i,4]<-"Los Angeles"}
}
```

We can now plot the two variations of emissions per year on two different graphs using ggplot and the *facet_grid* function.

```
p<-ggplot(data=agg_car,aes(year,Emissions,colour=factor(City)))
p<-p+geom_point(size=4,shape=7)
p<-p+geom_line(size=0.8)
p<-p+facet_grid(fips~.,scales="free_y")
print(p)
```