21 Étude de cas sur l'analyse de données: Modifications de la pollution atmosphérique en particules fines aux États-Unis

Ce chapitre présente un exemple d'analyse de données portant sur l'évolution de la pollution atmosphérique aux États-Unis à l'aide des données de surveillance nationales disponibles gratuitement. Le but de ce chapitre est simplement de montrer comment les différents outils que nous avons abordés dans ce livre peuvent être utilisés pour lire, manipuler et résumer des données afin que vous puissiez développer des preuves statistiques pour des questions pertinentes du monde réel.

Regarder une vidéo de ce chapitre

21.1 Synopsis

Dans ce chapitre, nous décrivons les changements survenus dans la pollution de l'air extérieur par les particules fines (PM2.5) aux États-Unis entre 1999 et 2012. Notre hypothèse générale est que les PM2.5 en extérieur ont diminué en moyenne à travers les États-Unis en raison de exigences réglementaires découlant de la Clean Air Act. Pour étudier cette hypothèse, nous avons obtenu des données sur les P2,5 de l'Environmental Protection Agency des États-Unis, qui ont été collectées à partir de moniteurs situés aux États-Unis. Nous avons spécifiquement obtenu des données pour les années 1999 et 2012 (année complète la plus récente disponible). À partir de ces données, nous avons constaté qu'en moyenne aux États-Unis, les niveaux de PM2,5 ont diminué entre 1999 et 2012. Sur un moniteur individuel, nous avons constaté que les niveaux avaient diminué et que la variabilité des PM2,5 avait diminué. La plupart des États ont également connu une diminution de PM2,5,

21.2 Chargement et traitement des données brutes

Grâce au système de qualité de l'air de l' EPA, nous avons obtenu des données sur la pollution de l'air par les particules fines (PM2,5) qui sont surveillées aux États-Unis dans le cadre du réseau national de surveillance des particules. Nous avons obtenu les fichiers pour les années 1999 et 2012.

21.2.1 Lecture des données de 1999

Nous avons d'abord lu les données de 1999 à partir du fichier texte brut inclus dans l'archive zip. Les données sont un fichier délimité où les champs sont délimités avec le | caractère et les valeurs manquantes sont codées en tant que champs vides. Nous omettons certaines lignes commentées au début du fichier et initialement, nous ne lisons pas les données d'en-tête.

```
> pm0 <- read.table("pm25_data/RD_501_88101_1999-0.txt", comment.char = "#", header = FA
```

Après la lecture de 1999, nous vérifions les premières lignes (il y en a 117 421) dans cet ensemble de données.

```
> dim(pm0)
[1] 117421     28
> head(pm0[, 1:13])
  V1 V2 V3 V4 V5    V6 V7 V8  V9 V10      V11   V12    V13
1 RD  I  1 27  1 88101  1  7 105 120 19990103 00:00     NA
2 RD  I  1 27  1 88101  1  7 105 120 19990106 00:00     NA
3 RD  I  1 27  1 88101  1  7 105 120 19990109 00:00     NA
4 RD  I  1 27  1 88101  1  7 105 120 19990112 00:00  8.841
5 RD  I  1 27  1 88101  1  7 105 120 19990115 00:00 14.920
6 RD  I  1 27  1 88101  1  7 105 120 19990118 00:00  3.878
```

Nous attachons ensuite les en-têtes de colonne à l'ensemble de données et nous nous assurons qu'ils sont correctement formatés pour les images de données R.

```
> cnames <- readLines("pm25_data/RD_501_88101_1999-0.txt", 1)
> cnames <- strsplit(cnames, "|", fixed = TRUE)
> ## Ensure names are properly formatted
> names(pm0) <- make.names(cnames[[1]])
> head(pm0[, 1:13])
  X..RD Action.Code State.Code County.Code Site.ID Parameter POC
1    RD           I          1          27       1     88101   1
2    RD           I          1          27       1     88101   1
3    RD           I          1          27       1     88101   1
4    RD           I          1          27       1     88101   1
5    RD           I          1          27       1     88101   1
6    RD           I          1          27       1     88101   1
  Sample.Duration Unit Method     Date Start.Time Sample.Value
1               7  105    120 19990103      00:00           NA
2               7  105    120 19990106      00:00           NA
3               7  105    120 19990109      00:00           NA
4               7  105    120 19990112      00:00        8.841
5               7  105    120 19990115      00:00       14.920
6               7  105    120 19990118      00:00        3.878
```

La colonne qui nous intéresse est la `Sample.Value` colonne qui contient les mesures de PM2.5. Ici, nous extrayons cette colonne et en imprimons un bref résumé.

```
> x0 <- pm0$Sample.Value
> summary(x0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    7.20   11.50   13.74   17.90  157.10   13217
```

Les valeurs manquantes sont un problème courant avec les données environnementales et nous vérifions donc quelle est la proportion des observations manquantes (c.-à-d. Codées ainsi `NA`).

```
> mean(is.na(x0))  ## Are missing values important here?
[1] 0.1125608
```

Étant donné que la proportion de valeurs manquantes est relativement faible (0.1125608), nous avons choisi d'ignorer les valeurs manquantes pour le moment.

21.2.2 Lecture des données de 2012

Nous lisons ensuite dans les données de 2012 de la même manière que nous lisons les données de 1999 (les fichiers de données sont dans le même format).

```
> pm1 <- read.table("pm25_data/RD_501_88101_2012-0.txt", comment.char = "#",
+                   header = FALSE, sep = "|", na.strings = "", nrow = 1304290)
```

Nous définissons également les noms de colonne (ils sont identiques à ceux du jeu de données 1999) et extrayons la `Sample.Value` colonne de ce jeu de données.

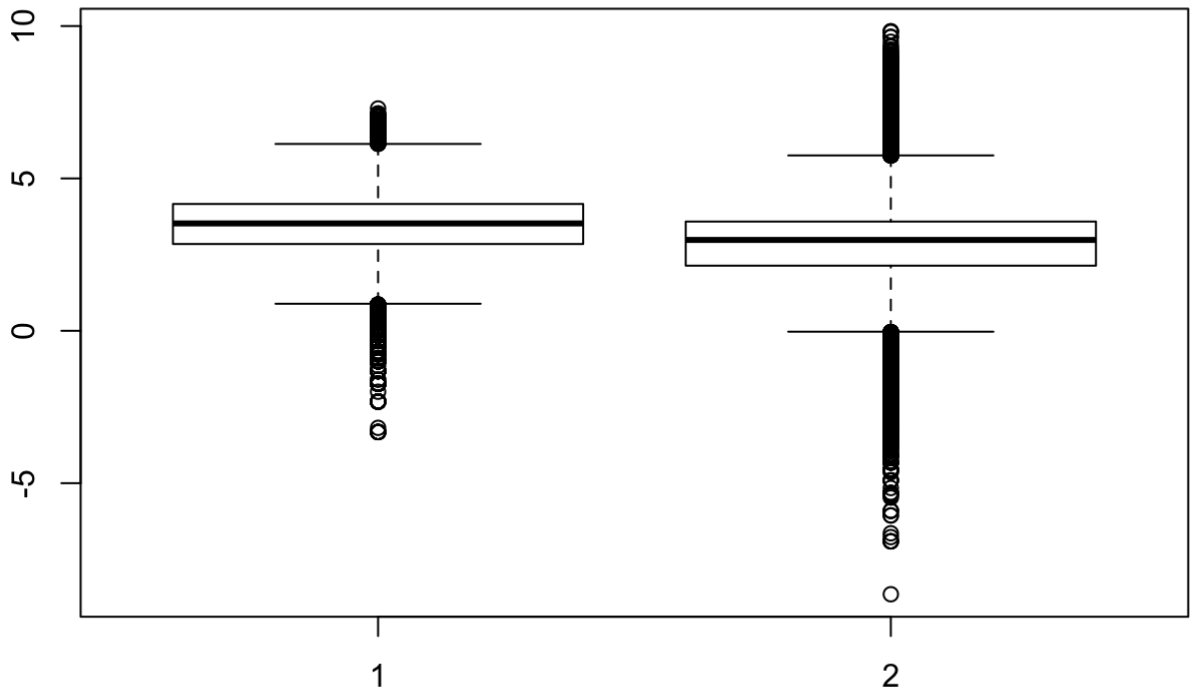
```
> names(pm1) <- make.names(cnames[[1]])
> x1 <- pm1$Sample.Value
```

21.3 Résultats

21.3.1 Analyse américaine complète

Pour afficher les modifications globales des particules sur l'ensemble du réseau de surveillance, nous pouvons créer des boîtes à moustaches de toutes les valeurs de surveillance en 1999 et 2012. Nous prenons ici le journal des valeurs de particules pour ajuster le décalage des données.

```
> boxplot(log2(x0), log2(x1))
Warning in boxplot.default(log2(x0), log2(x1)): NaNs produced
Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
z$out[z$group == : Outlier (-Inf) in boxplot 1 is not drawn
Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
z$out[z$group == : Outlier (-Inf) in boxplot 2 is not drawn
```




```
> summary(x0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    7.20   11.50   13.74   17.90  157.10   13217
> summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -10.00    4.00    7.63    9.14   12.00  908.97   73133
```

Fait intéressant, `x1` il en ressort qu'il existe des valeurs négatives de PM, qui ne devraient généralement pas se produire. Nous pouvons enquêter un peu sur cela pour voir s'il y a quelque chose qui devrait nous inquiéter.

```
> negative <- x1 < 0
> mean(negative, na.rm = T)
[1] 0.0215034
```

Il y a une proportion relativement faible de valeurs négatives, ce qui est peut-être rassurant. Afin d'étudier cela un peu plus loin, nous pouvons extraire la date de chaque mesure du bloc de données d'origine. L'idée ici est que peut-être des valeurs négatives se produisent plus

souvent dans certaines parties de l'année que d'autres. Cependant, les données d'origine sont formatées sous forme de chaînes de caractères. Nous les convertissons donc au `Date` format de R pour faciliter la manipulation.

```
> dates <- pm1$Date
> dates <- as.Date(as.character(dates), "%Y%m%d")
```

Nous pouvons ensuite extraire le mois de chacune des dates avec des valeurs négatives et tenter d'identifier le moment où les valeurs négatives se produisent le plus souvent.

```
> missing.months <- month.name[as.POSIXlt(dates)$mon + 1]
> tab <- table(factor(missing.months, levels = month.name))
> round(100 * tab / sum(tab))
```

```
  January  February     March     April       May      June      July 
       15        13        15        13        14        13         8 
   August September   October  November  December 
        6         3         0         0         0 
```

Il ressort du tableau ci-dessus que la majeure partie des valeurs négatives se produit au cours des six premiers mois de l'année (janvier à juin). Cependant, au-delà de cette simple observation, il n'est pas clair pourquoi les valeurs négatives se produisent. Cela dit, étant donné la proportion relativement faible de valeurs négatives, nous les ignorerons pour le moment.

21.3.2 Modifications des niveaux de particules sur un moniteur individuel

Jusqu'ici, nous avons examiné l'évolution des niveaux de particules en moyenne dans l'ensemble du pays. L'un des problèmes soulevés par l'analyse précédente est que le réseau de surveillance aurait pu être modifié entre 1999 et 2012. Donc, si pour une raison quelconque, en 2012, il y avait plus de contrôleurs concentrés dans des régions plus propres du pays qu'en 1999, les niveaux de particules ont diminué alors qu'en réalité ils ne l'ont pas fait. Dans cette section, nous nous concentrerons sur un seul moniteur dans l'État de New York pour voir si les niveaux de particules *à ce moniteur ont* diminué de 1999 à 2012.

Notre première tâche consiste à identifier un moniteur dans l'État de New York qui dispose de données en 1999 et 2012 (tous les moniteurs n'ont pas fonctionné au cours des deux périodes). Tout d'abord, nous sous-définissons les trames de données pour n'inclure que les données de New York (`State.Code == 36`) et uniquement les variables `County.Code` et le `Site.ID` (c'est-à-dire le numéro du moniteur).

```
> site0 <- unique(subset(pm0, State.Code == 36, c(County.Code, Site.ID)))
> site1 <- unique(subset(pm1, State.Code == 36, c(County.Code, Site.ID)))
```

Ensuite, nous créons une nouvelle variable qui combine le code du comté et l'ID de site en une seule chaîne.

```
> site0 <- paste(site0[,1], site0[,2], sep = ".")
> site1 <- paste(site1[,1], site1[,2], sep = ".")
> str(site0)
 chr [1:33] "1.5" "1.12" "5.73" "5.80" "5.83" "5.110" "13.11" ...
> str(site1)
 chr [1:18] "1.5" "1.12" "5.80" "5.133" "13.11" "29.5" "31.3" "47.122" ...
```

Enfin, nous voulons une intersection entre les sites présents en 1999 et 2012 afin de pouvoir choisir un moniteur contenant des données pour les deux périodes.

```
> both <- intersect(site0, site1)
> print(both)
 [1] "1.5"     "1.12"    "5.80"    "13.11"   "29.5"    "31.3"    "63.2008"
 [8] "67.1015" "85.55"   "101.3"
```

Ici (ci-dessus), nous pouvons voir que 10 moniteurs fonctionnaient dans les deux périodes. Cependant, plutôt que de choisir une au hasard, il vaut peut-être mieux en choisir une qui dispose d'une quantité de données raisonnable chaque année.

```
> ## Find how many observations available at each monitor
> pm0$county.site <- with(pm0, paste(County.Code, Site.ID, sep = "."))
> pm1$county.site <- with(pm1, paste(County.Code, Site.ID, sep = "."))
> cnt0 <- subset(pm0, State.Code == 36 & county.site %in% both)
> cnt1 <- subset(pm1, State.Code == 36 & county.site %in% both)
```

Maintenant que nous avons subdivisé les trames de données d'origine pour n'inclure que les données des moniteurs qui se chevauchent entre 1999 et 2012, nous pouvons fractionner les trames de données et compter le nombre d'observations sur chaque moniteur pour voir celles qui en ont le plus.

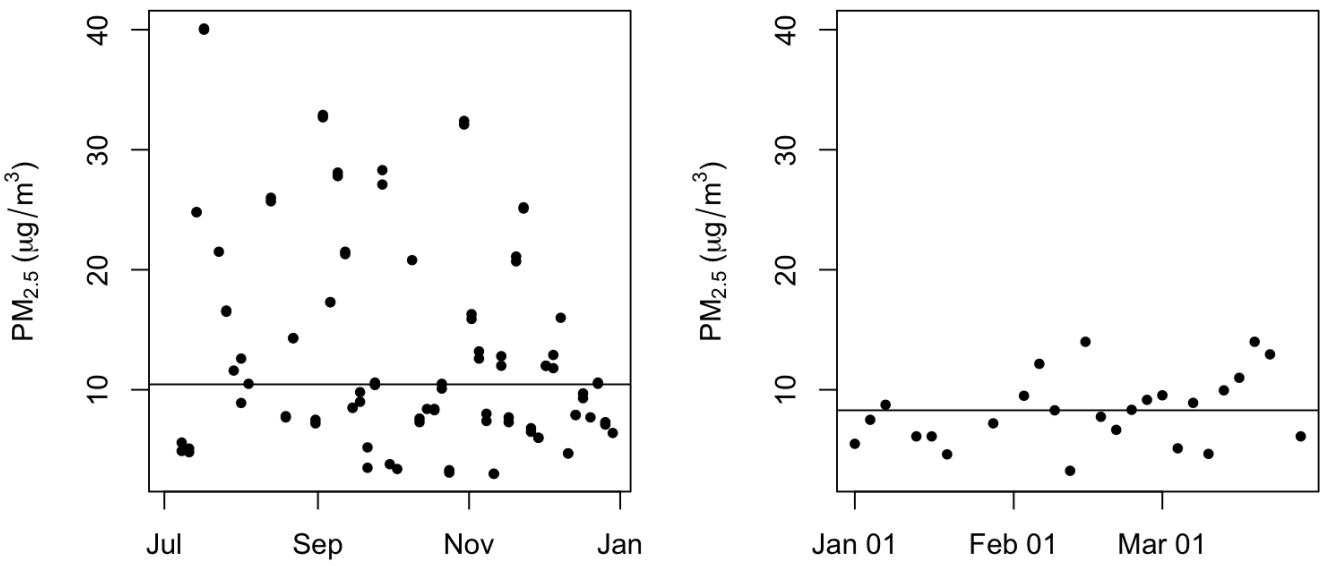
```
> ## 1999
> sapply(split(cnt0, cnt0$county.site), nrow)
   1.12     1.5   101.3   13.11    29.5    31.3    5.80 63.2008 67.1015
     61     122     152      61      61     183      61     122     122
  85.55
      7
> ## 2012
> sapply(split(cnt1, cnt1$county.site), nrow)
   1.12     1.5   101.3   13.11    29.5    31.3    5.80 63.2008 67.1015
     31      64      31      31      33      15      31      30      31
  85.55
     31
```

Un certain nombre de moniteurs semblent convenir à la sortie, mais nous allons nous concentrer ici sur le comté 63 et le site ID 2008.

```
> both.county <- 63
> both.id <- 2008
>
> ## Choose county 63 and side ID 2008
> pm1sub <- subset(pm1, State.Code == 36 & County.Code == both.county & Site.ID == both.
> pm0sub <- subset(pm0, State.Code == 36 & County.Code == both.county & Site.ID == both.
```

Nous traçons maintenant les données de série chronologique des particules pour le moniteur au cours des deux années.

```
> dates1 <- as.Date(as.character(pm1sub$Date), "%Y%m%d")
> x1sub <- pm1sub$Sample.Value
> dates0 <- as.Date(as.character(pm0sub$Date), "%Y%m%d")
> x0sub <- pm0sub$Sample.Value
>
> ## Find global range
> rng <- range(x0sub, x1sub, na.rm = T)
> par(mfrow = c(1, 2), mar = c(4, 5, 2, 1))
> plot(dates0, x0sub, pch = 20, ylim = rng, xlab = "", ylab = expression(PM[2.5] * " ("
> abline(h = median(x0sub, na.rm = T))
> plot(dates1, x1sub, pch = 20, ylim = rng, xlab = "", ylab = expression(PM[2.5] * " ("
> abline(h = median(x1sub, na.rm = T))
```



Le graphique ci-dessus montre que les niveaux médians de particules (trait continu horizontal) ont légèrement diminué, passant de 10,45 en 1999 à 8,29 en 2012. Il est peut-être plus intéressant de noter que la variation (l'étendue) des valeurs de particules en 2012 est beaucoup plus grande. inférieure à celle de 1999. Cela suggère que non seulement les concentrations médianes de particules sont-elles plus basses en 2012, mais également qu'il y a moins de fortes pointes d'un jour à l'autre. Un problème avec les données ici est que les données de 1999 sont de juillet à décembre alors que les données de 2012 sont enregistrées de janvier à avril. Il aurait été préférable de disposer de données annuelles pour les deux années, car il pourrait y avoir une confusion saisonnière.

21.3.3 Modifications des niveaux de PM à l'échelle de l'État

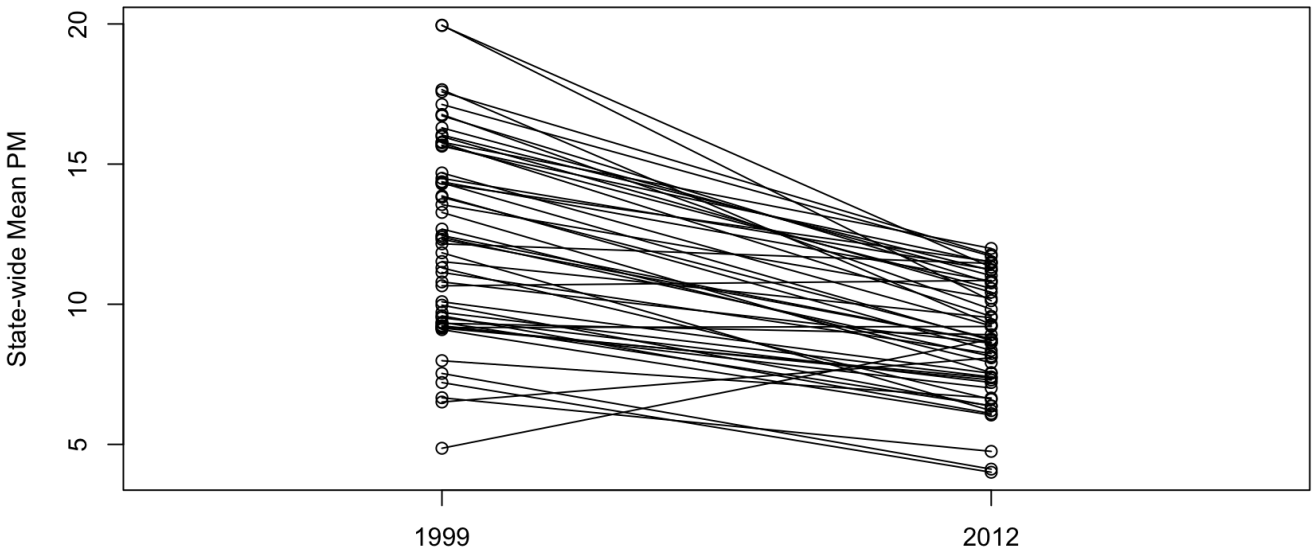
Bien que les normes de qualité de l'air ambiant soient définies au niveau fédéral aux États-Unis et touchent donc l'ensemble du pays, la réduction et la gestion effectives des particules sont laissées à la discrétion des États. Les États qui ne sont pas «en voie de réalisation» doivent élaborer un plan de réduction des particules afin qu'ils atteignent (éventuellement). Par conséquent, il pourrait être utile d'examiner les modifications de PM au niveau de l'état. Cette analyse se situe quelque part entre le pays tout entier et un moniteur individuel.

Ce que nous faisons ici est de calculer la moyenne des particules pour chaque état en 1999 et 2012.

```
> ## 1999
> mn0 <- with(pm0, tapply(Sample.Value, State.Code, mean, na.rm = TRUE))
> ## 2012
> mn1 <- with(pm1, tapply(Sample.Value, State.Code, mean, na.rm = TRUE))
>
> ## Make separate data frames for states / years
> d0 <- data.frame(state = names(mn0), mean = mn0)
> d1 <- data.frame(state = names(mn1), mean = mn1)
> mrg <- merge(d0, d1, by = "state")
> head(mrg)
  state    mean.x    mean.y
1     1 19.956391 10.126190
2    10 14.492895 11.236059
3    11 15.786507 11.991697
4    12 11.137139  8.239690
5    13 19.943240 11.321364
6    15  4.861821  8.749336
```

Faites maintenant un graphique qui montre les moyennes de 1999 à l'état dans une «colonne» et les moyennes de 2012 à l'état dans une autre colonne. Nous traçons ensuite une ligne reliant les moyennes pour chaque année dans le même état afin de mettre en évidence la tendance.

```
> par(mfrow = c(1, 1))
> rng <- range(mrg[,2], mrg[,3])
> with(mrg, plot(rep(1, 52), mrg[, 2], xlim = c(.5, 2.5), ylim = rng, xaxt = "n", xlab =
> with(mrg, points(rep(2, 52), mrg[, 3]))
> segments(rep(1, 52), mrg[, 2], rep(2, 52), mrg[, 3])
> axis(1, c(1, 2), c("1999", "2012"))
```



Le graphique ci-dessus montre que de nombreux États ont diminué les niveaux moyens de particules de 1999 à 2012 (bien que quelques États aient en fait augmenté leurs niveaux).