Analyse de l'incidence du syndrôme grippal

Marine C. Cambon

Table des matières

Préparation des données

Les données de l'incidence du syndrome grippal sont disponibles du site Web du Réseau Sentinelles. Nous les récupérons sous forme d'un fichier en format CSV dont chaque ligne correspond à une semaine de la période demandée. Nous téléchargeons toujours le jeu de données complet, qui commence en 1984 et se termine avec une semaine récente. L'URL est :

```
data_url = "http://www.sentiweb.fr/datasets/incidence-PAY-3.csv"
```

Voici l'explication des colonnes donnée sur le sur le site d'origine :

Nom de colonne	Libellé de colonne
`week`	Semaine calendaire (ISO 8601)
`indicator`	Code de l'indicateur de surveillance
`inc`	Estimation de l'incidence de consultations en nombre de cas
`inc_low`	Estimation de la borne inférieure de l'IC95% du nombre de cas de consultation
`inc_up`	Estimation de la borne supérieure de l'IC95% du nombre de cas de consultation
`inc100`	Estimation du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
`inc100_low`	Estimation de la borne inférieure de l'IC95% du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
`inc100_up`	Estimation de la borne supérieure de l'IC95% du taux d'incidence du nombre de cas de consultation (en cas pour 100,000 habitants)
`geo_insee`	Code de la zone géographique concernée (Code INSEE) http://www.insee.fr/fr/methodes/nomenclatures/cog/
`geo_name`	Libellé de la zone géographique (ce libellé peut être modifié sans préavis)

Chargement du jeu de données

Pour éviter les problèmes de lien cassés et pour gagner du temps de téléchargement, le jeu de données a été téléchargé en local, et est en suite chargé.

La première ligne du fichier CSV est un commentaire, que nous ignorons en précisant `skip=1`.

```
data = read.csv("incidence-PAY-3.csv", skip=1)
```

Regardons ce que nous avons obtenu :

```
head(data)
```

```
##     week indicator    inc inc_low inc_up inc100 inc100_low inc100_up geo_insee
## 1 202012         3  10125    7199  13051     15         11        19        FR
## 2 202011         3 102048   93969 110127    155        143       167        FR
## 3 202010         3 104977   96650 113304    159        146       172        FR
## 4 202009         3 110696  102066 119326    168        155       181        FR
## 5 202008         3 143753  133984 153522    218        203       233        FR
## 6 202007         3 183610  172812 194408    279        263       295        FR
##   geo_name
## 1   France
## 2   France
## 3   France
## 4   France
## 5   France
## 6   France
```

```
tail(data)
```

```
##        week indicator    inc inc_low inc_up inc100 inc100_low inc100_up
## 1842 198449         3 101073   81684 120462    184        149       219
## 1843 198448         3  78620   60634  96606    143        110       176
## 1844 198447         3  72029   54274  89784    131         99       163
## 1845 198446         3  87330   67686 106974    159        123       195
## 1846 198445         3 135223  101414 169032    246        184       308
## 1847 198444         3  68422   20056 116788    125         37       213
##      geo_insee geo_name
## 1842        FR   France
## 1843        FR   France
## 1844        FR   France
## 1845        FR   France
## 1846        FR   France
## 1847        FR   France
```

Y a-t-il des points manquants dans nos données?

```
na_records = apply(data, 1, function (x) any(is.na(x)))
data[na_records,]
```

```
##        week indicator inc inc_low inc_up inc100 inc100_low inc100_up geo_insee
## 1611 198919         3   0      NA     NA      0         NA        NA        FR
##      geo_name
```

```
## 1611   France
```

Les deux colonnes qui nous intéressent sont `week` et `inc`. Vérifions leurs classes :

```
class(data$week)
```

```
## [1] "integer"
```

```
class(data$inc)
```

```
## [1] "integer"
```

Ce sont des entiers, tout va bien !

Conversion des numéros de semaine

La gestion des dates est toujours un sujet délicat. Il y a un grand nombre de conventions différentes qu'il ne faut pas confondre. Notre jeux de données utilise un format que peu de logiciels savent traiter : les semaines en format ISO-8601. En `R`, il est géré par la bibliothèque parsedate :

```
library(parsedate)
```

Pour faciliter le traitement suivant, nous remplaçons ces semaines par les dates qui correspondent aux lundis. Voici une petite fonction qui fait la conversion pour une seule valeur :

```
convert_week = function(w) {
    ws = paste(w)
    iso = paste0(substring(ws, 1, 4), "-W", substring(ws, 5, 6))
    as.character(parse_iso_8601(iso))
}
```

Nous appliquons cette fonction à tous les points, créant une nouvelle colonne `date` dans notre jeu de données :

```
data$date = as.Date(convert_week(data$week))
```

Vérifions qu'elle est de classe `Date` :

```
class(data$date)
```

```
## [1] "Date"
```

Les points sont dans l'ordre chronologique inverse, il est donc utile de les trier :

```
data = data[order(data$date),]
```

C'est l'occasion pour faire une vérification : nos dates doivent être séparées d'exactement sept jours :

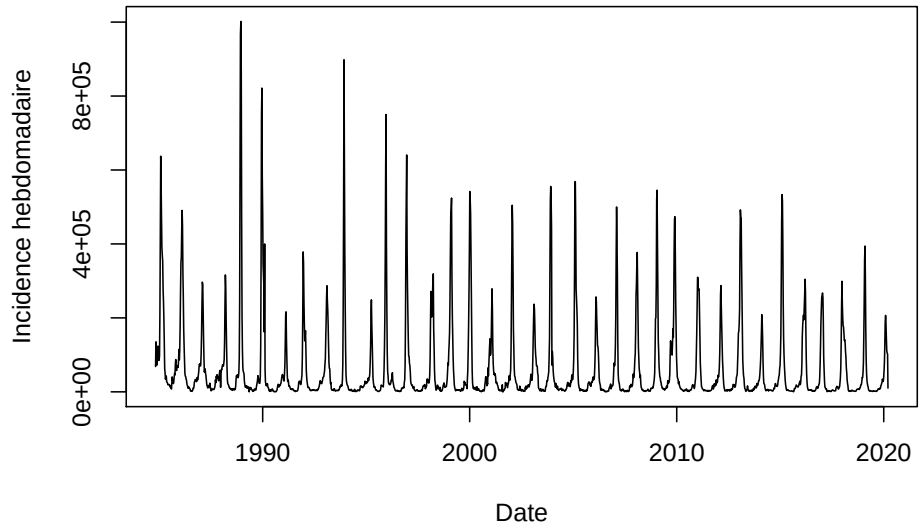
```
all(diff(data$date) == 7)
```

```
## [1] TRUE
```

Inspection

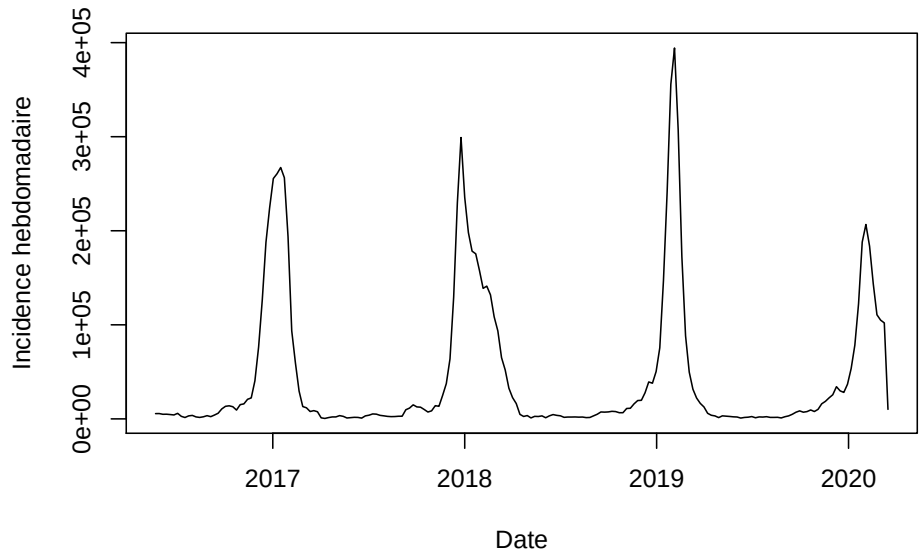
Regardons enfin à quoi ressemblent nos données!

```
plot(data$date, data$inc, type="l", xlab="Date", ylab="Incidence hebdomadaire")
```



Un zoom sur les dernières années montre mieux la localisation des pics en hiver. Le creux des incidences se trouve en été.

```
with(tail(data, 200), plot(date, inc, type="l", xlab="Date", ylab="Incidence hebdomadaire"
```



L'incidence annuelle

Calcul

Étant donné que le pic de l'épidémie se situe en hiver, à cheval entre deux années civiles, nous définissons la période de référence entre deux minima de l'incidence, du 1er août de l'année N au 1er août de l'année N + 1. Nous mettons l'année N + 1 comme étiquette sur cette année décalée, car le pic de l'épidémie est toujours au début de l'année N + 1. Comme l'incidence de syndrome grippal est très faible en été, cette modification ne risque pas de fausser nos conclusions. L'argument `na.rm=True` dans la sommation précise qu'il faut supprimer les points manquants. Ce choix est raisonnable car il n'y a qu'un seul point manquant, dont l'impact ne peut pas être très fort.

```
pic_annuel = function(annee) {
      debut = paste0(annee-1,"-08-01")
      fin = paste0(annee,"-08-01")
      semaines = data$date > debut & data$date <= fin
      sum(data$inc[semaines], na.rm=TRUE)
      }
```

Nous devons aussi faire attention aux premières et dernières années de notre jeux de données. Les données commencent en octobre 1984, ce qui ne permet pas de quantifier complètement le pic attribué à 1985. Nous l'enlevons donc de notre analyse. Par contre, pour une exécution en octobre 2018, les données se terminent après le 1er août 2018, ce qui nous permet d'inclure cette année.

```
annees = 1986:2018
```

Nous créons un nouveau jeu de données pour l'incidence annuelle, en applicant la fonction `pic_annuel` à chaque année :

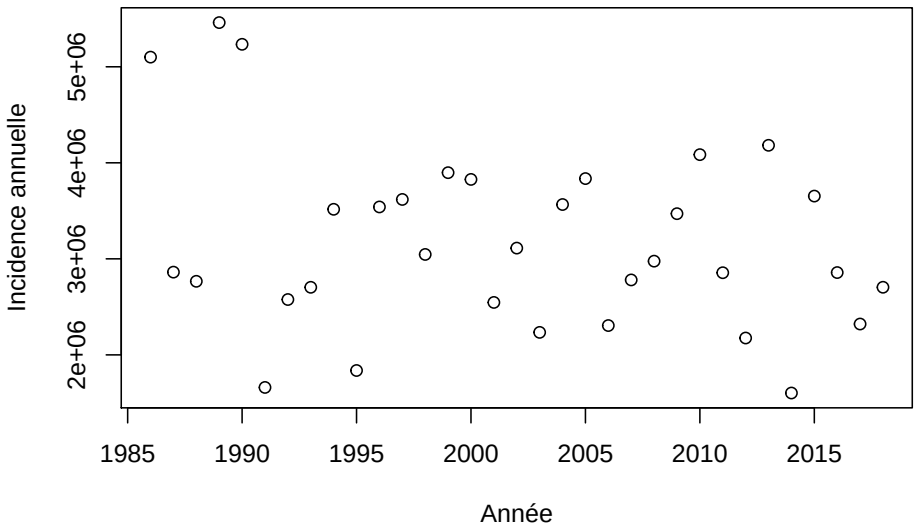
```
inc_annuelle = data.frame(annee = annees,
                          incidence = sapply(annees, pic_annuel))
head(inc_annuelle)
```

```
##   annee incidence
## 1  1986   5100540
## 2  1987   2861556
## 3  1988   2766142
## 4  1989   5460155
## 5  1990   5233987
## 6  1991   1660832
```

Inspection

Voici les incidences annuelles en graphique :

```
plot(inc_annuelle, type="p", xlab="Année", ylab="Incidence annuelle")
```



Identification des épidémies les plus fortes

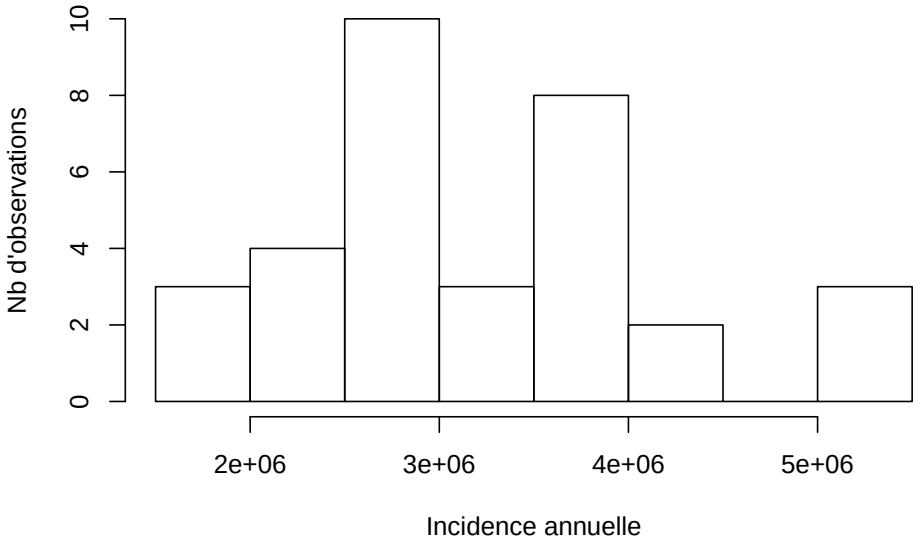
Une liste triée par ordre décroissant d'incidence annuelle permet de plus facilement repérer les valeurs les plus élevées :

```
head(inc_annuelle[order(-inc_annuelle$incidence),])
```

```
##    annee incidence
## 4   1989   5460155
## 5   1990   5233987
## 1   1986   5100540
## 28  2013   4182265
## 25  2010   4085126
## 14  1999   3897443
```

Enfin, un histogramme montre bien que les épidémies fortes, qui touchent environ 10% de la population française, sont assez rares : il y en eu trois au cours des 35 dernières années.

```
hist(inc_annuelle$incidence, breaks=10, xlab="Incidence annuelle", ylab="Nb d'observations
```


Nb d'observations
0
2
4
6
8
10
2e+06
3e+06
4e+06
5e+06
Incidence annuelle