

Analyse de l'incidence du syndrome grippal

Konrad Hinsén

Table des matières

Préparation des données	1
L'incidence annuelle	6

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"
path = "./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/

Nom de colonne	Libellé de colonne
geo_name	Libellé de la zone géographique (ce libellé peut être modifié sans préavis)

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

Téléchargement

Pour éviter les problèmes de changement d'URL ou de format on télécharge le fichier de données en local. Pour ça on commence par vérifier si le fichier a déjà été téléchargé, si il ne l'a pas été on le télécharge avec `read.csv` et on le stocke en local avec `write.csv2`, cette dernière fonction permet de stocker un tableau de données au format csv à l'emplacement indiqué.

```
if (file.exists(path)==T){
  data = read.csv(path, skip=1, header = T, sep = ",")
}else {
  data = read.csv(data_url, skip=1, header = T, sep = ",")
  write.csv2(data,path)
  print("Le fichier a bien été importé")
}
```

Regardons ce que nous avons obtenu :

```
head(data)
```

```
##      week indicator    inc inc_low inc_up inc100 inc100_low inc100_up geo_insee
## 1 202147           3 37236  31834 42638     56         48      64      FR
## 2 202146           3 30063  25306 34820     46         39      53      FR
## 3 202145           3 20364  16564 24164     31         25      37      FR
## 4 202144           3 18999  15042 22956     29         23      35      FR
## 5 202143           3 27040  21935 32145     41         33      49      FR
## 6 202142           3 28343  23382 33304     43         35      51      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
```

```
## 1930 198449      3 101073   81684 120462   184      149      219
## 1931 198448      3  78620   60634  96606   143      110      176
## 1932 198447      3  72029   54274  89784   131       99      163
## 1933 198446      3  87330   67686 106974   159      123      195
## 1934 198445      3 135223  101414 169032   246      184      308
## 1935 198444      3  68422   20056 116788   125       37      213
##      geo_insee geo_name
## 1930      FR   France
## 1931      FR   France
## 1932      FR   France
## 1933      FR   France
## 1934      FR   France
## 1935      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
## 1699 198919      3  0      NA      NA      0      NA      NA      FR
##      geo_name
## 1699   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)
```

```
## Warning: le package 'parsedate' a été compilé avec la version R 4.1.2
```

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 :

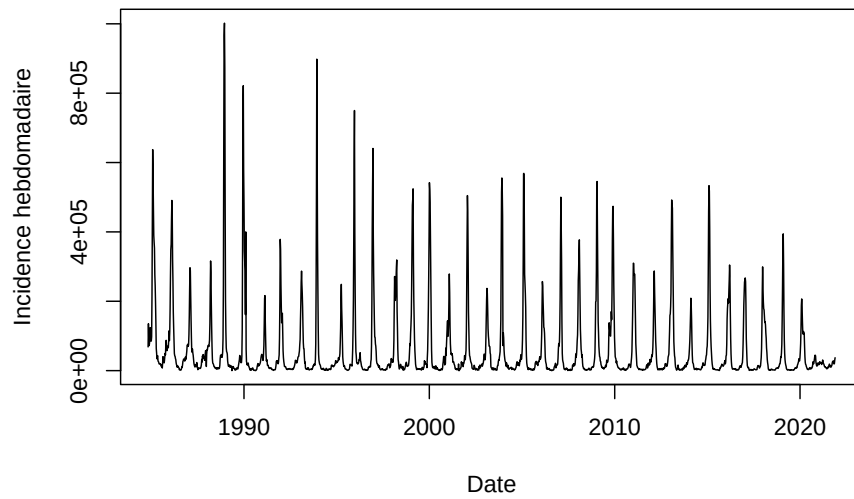
```
all(diff(data$date) == 7)
```

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

Inspection

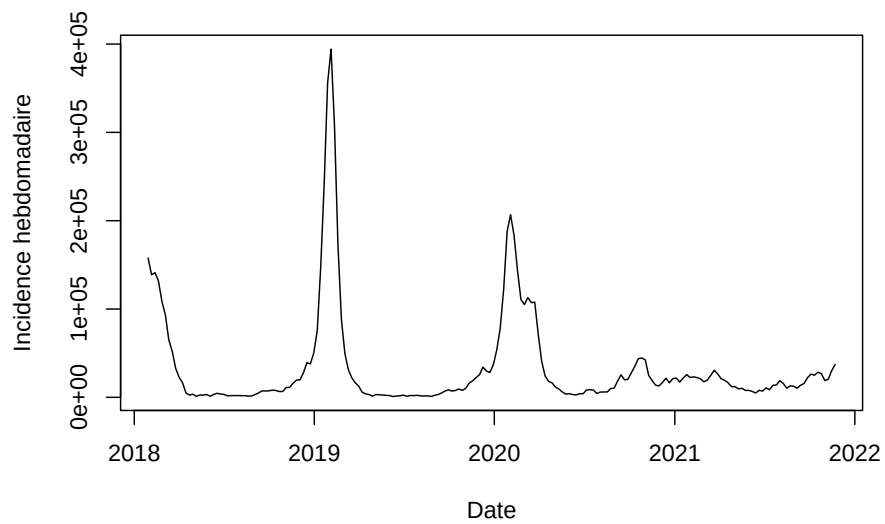
Regardons enfin à quoi ressemblent nos données!

```
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 jeu 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 appliquant la fonction `pic_annuel` à chaque année :

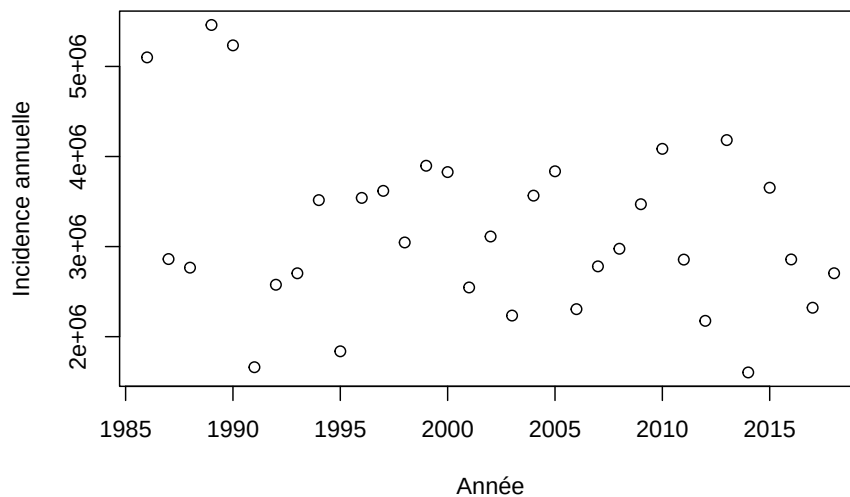
```
inc_annuelle = data.frame(annee = annees,  
                           incidence = sapply(annees, pic_annuel))  
head(inc_annuelle)
```

```
##   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")
```

