Analyse de l'incidence de la varicelle

Marine C. Cambon

Table des matières

Préparation des données

Les données de l'incidence de la varicelle 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 1991 et se termine avec une semaine récente. L'URL est :

```
data_url = "https://www.sentiweb.fr/datasets/incidence-PAY-7.csv"
```

Pour nous protéger contre une éventuelle disparition ou modification du serveur du Réseau Sentinelles, nous faisons une copie locale de ce jeux de données que nous préservons avec notre analyse. Il est inutile et même risquée de télécharger les données à chaque exécution, car dans le cas d'une panne nous pourrions remplacer nos données par un fichier défectueux. Pour cette raison, nous téléchargeons les données seulement si la copie locale n'existe pas.

```
data_file = "varicelle.csv"
if (!file.exists(data_file)) {
    download.file(data_url, data_file, method="auto")
}
```

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)

Nom de colonne	Libellé de colonne
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

Nous chargeons maintenant le jeu de donné téléchargé sur notre machine.

```
data = read.csv(data_file, 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         7  8639    6010  11268     13          9        17         FR
## 2 202011         7 10209    7575  12843     16         12        20         FR
## 3 202010         7  9011    6691  11331     14         10        18         FR
## 4 202009         7 13631   10544  16718     21         16        26         FR
## 5 202008         7 10424    7708  13140     16         12        20         FR
## 6 202007         7  8959    6574  11344     14         10        18         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
## 1524 199102         7 16277   11046  21508     29         20        38
## 1525 199101         7 15565   10271  20859     27         18        36
## 1526 199052         7 19375   13295  25455     34         23        45
## 1527 199051         7 19080   13807  24353     34         25        43
## 1528 199050         7 11079    6660  15498     20         12        28
## 1529 199049         7  1143       0   2610      2          0         5
##      geo_insee geo_name
## 1524        FR   France
## 1525        FR   France
## 1526        FR   France
## 1527        FR   France
```

```
## 1528        FR   France
## 1529        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,]
```

```
##  [1] week       indicator  inc        inc_low    inc_up     inc100
##  [7] inc100_low inc100_up  geo_insee  geo_name
## <0 rows> (or 0-length row.names)
```

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 :

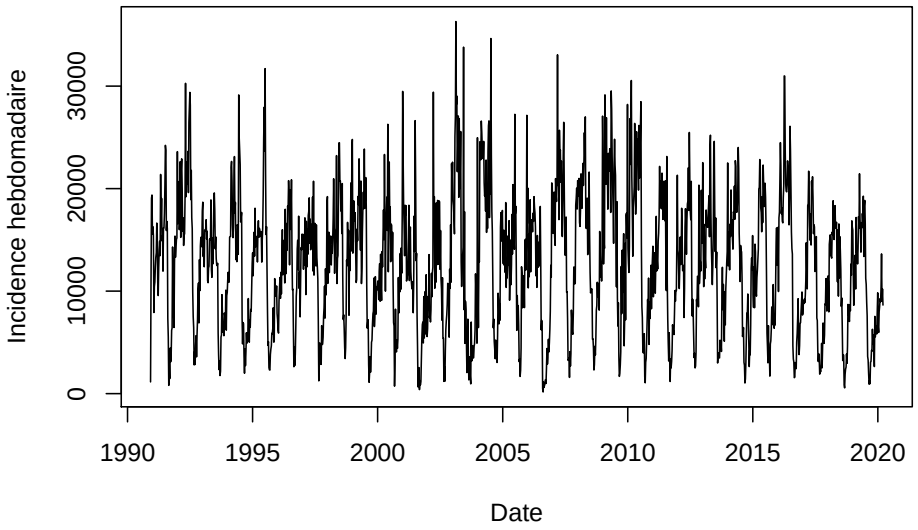
```
all(diff(data$date) == 7)
```

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

Inspection

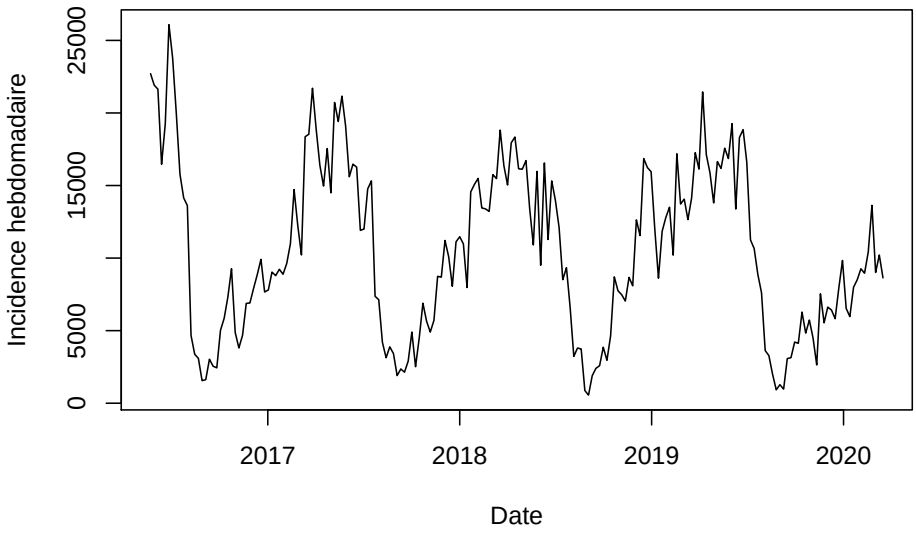
Regardons enfin à quoi ressemblent nos données!

```
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 septembre de l'année N au 1er septembre 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,"-09-01")
      fin = paste0(annee,"-09-01")
      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 = 1992:2019
```

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

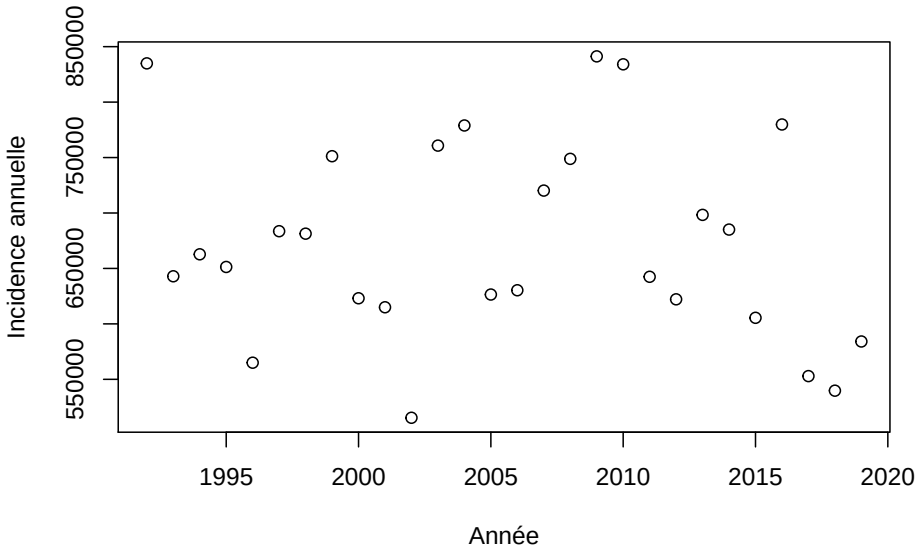
```
inc_annuelle = data.frame(annee = annees,
                          incidence = sapply(annees, pic_annuel))
head(inc_annuelle)
```

```
##   annee incidence
## 1  1992    834935
## 2  1993    642921
## 3  1994    662750
## 4  1995    651333
## 5  1996    564994
## 6  1997    683577
```

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
## 18  2009    841233
## 1   1992    834935
## 19  2010    834077
## 25  2016    779816
## 13  2004    778914
```

```
## 12  2003    760765
```

L'année à l'incidence la plus élevée est :

```
inc_annuelle[which.max(inc_annuelle$incidence),]
```

```
##    annee incidence
## 18  2009    841233
```

```
inc_annuelle[which.min(inc_annuelle$incidence),]
```

```
##    annee incidence
## 11  2002    515343
```

Celle à l'incidence la plus faible est :

```
inc_annuelle[which.min(inc_annuelle$incidence),]
```

```
##    annee incidence
## 11  2002    515343
```

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
```