Introduction à l'analyse des séries temporelles

TP1

```r
# Librairies utilisées
library(tidyverse)
library(lubridate)
library(Acf)
```

Cette première séance de travaux pratiques a pour but de se familiariser avec les commandes R permettant l'analyse des séries temporelles. On y abordera les commandes suivantes :

- Fonctions exploratrices et descriptives des séries temporelles :
  - `diff(series, lag = d)` permet de calculer les différences d'ordre d d'une série temporelle.
  - `decompose(series, type = c("additive", "multiplicative"))` permet de décomposer une série temporelle selon ses composantes tendancielle, saisonnière, et résiduelle : $x_t = T_t + S_t + \nu_t$. Attention : pour déterminer si T_t et S_t sont réellement présentes, il est impératif de comparer l'amplitude de ces composantes à l'amplitude de la composante résiduelle.
- Fonctions propres à la démarche de modélisation des séries temporelles par processus ARIMA :
  - **Identification** : `acf(series, lag.max = H)` et `pacf(series, lag.max = H)` permettent d'afficher les fonctions empiriques d'autocorrélation et d'autocorrélation partielle d'une série temporelle. Alternativement, on pourra préferer la fonction `Acf::Acf(series, lag.max = H)` qui effectue les deux à la fois.
  - **Estimation** : `arima(series, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S))` permet d'effectuer l'estimation d'un modèle SARIMA$(p, d, q)(P, D, Q)_S$ par une méthode de maximum de vraisemblance.
  - **Diagnostic** : `Box.test(series, lag = H, type = "Ljung-Box", fitdf = ddl)` permet d'effectuer un test d'hypothèse portant sur l'ensemble des pics d'autocorrélation jusqu'à l'ordre H.
  - **Prédiction** : `predict(model, n.ahead = m, se.fit = TRUE)` permet d'effectuer des prédictions d'un modèle SARIMA.
- Fonctions pour tester la normalité (au sens de sa loi) d'un échantillon :
  - `qqnorm(sample)` permet de tracer le graphique quantile-quantile associé à une loi normale.
  - Test de Kolmogorov-Smirnov : `ks.test(series, "pnorm")`.
  - Test de Shapiro-Wilk : `shapiro.test(series)`.

Partie 1 : identification des ordres de processus ARIMA

Planification de votre session de travail RStudio

Afin de rendre facile la manipulation des différents fichiers que vous allez manipuler au cours de ce TP, il est recommandé de créer un projet :

- Dans RStudio, créez un nouveau projet à l'emplacement de votre choix : `File > New Project...` (ou `Fichier > Nouveau Projet...`).
- Créez désormais un nouveau script R, ou un nouveau fichier RMarkdown, qui correspondra à votre fichier de travail.
- Téléchargez sur le site du cours le fichier `sim.RData`, puis déplacez ce fichier dans le répertoire de travail du projet.
- Dans votre script R, importez les données du fichier `sim.RData` : `load("sim.RData")`.

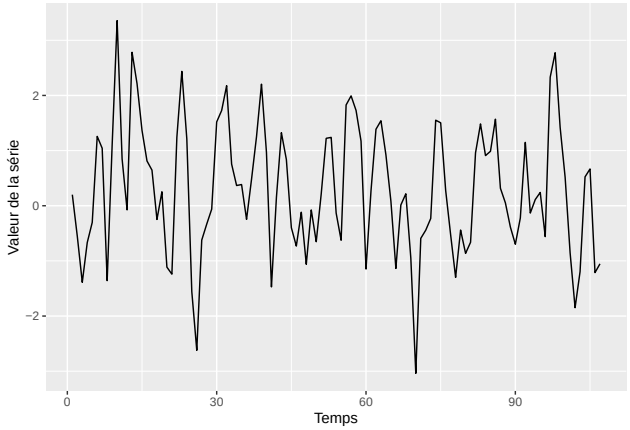
Consignes

Le fichier `sim.RData` contient 5 séries temporelles simulées, nommées `sim1` jusqu'à `sim5`. Pour chacune de ces séries, déterminez de quel processus SARIMA il s'agit, puis estimez les paramètres du modèle. Pour s'assurer que le modèle identifié est le bon, vérifiez que les résidus du modèle ne présentent plus d'autocorrélation.

Les étapes ci-dessous illustrent cette démarche pour la série `sim1` :

1. Tracez la série temporelle pour identifier s'il existe une tendance ou une saisonnalité (fonction `plot(series, type = "l")` ou utilisez les fonctions de la librairie `ggplot2`).

```
sim = tibble(x = seq_along(sim1), y = as.numeric(sim1))
ggplot(sim, aes(x = x, y = y)) +
    geom_line() +
    labs(x = "Temps", y = "Valeur de la série")
```

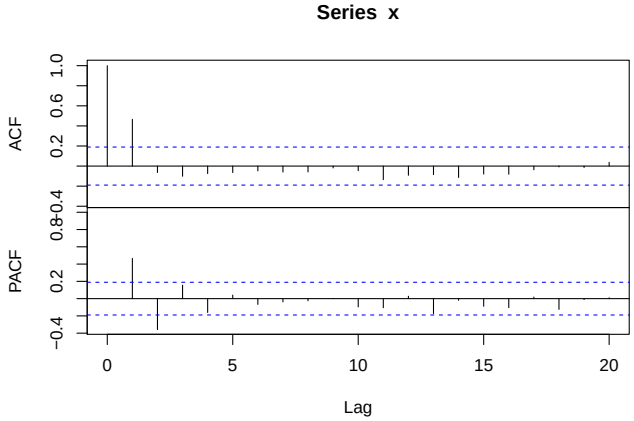


2. Le cas échéant, dégagez la tendance ou la saisonnalité en effectuant une différentiation de la série à l'ordre approprié.

```
# Ici, pas de tendance ni de saisonnalité identifiable à l'oeil nu.
```

3. À partir des ACF et PACF de la série, identifiez :
   a. s'il existe une structure à long terme et s'il s'agit d'une structure AR ou MA.
   b. le cas échéant, passez directement à l'étape 4.
   c. s'il existe une structure à court terme et s'il s'agit d'une structure AR ou MA.

```
Acf(sim1)
```




```
# La série ne présente pas de structure de long terme.
# L'analyse de la structure de court terme oriente vers un processus MA(1)
```

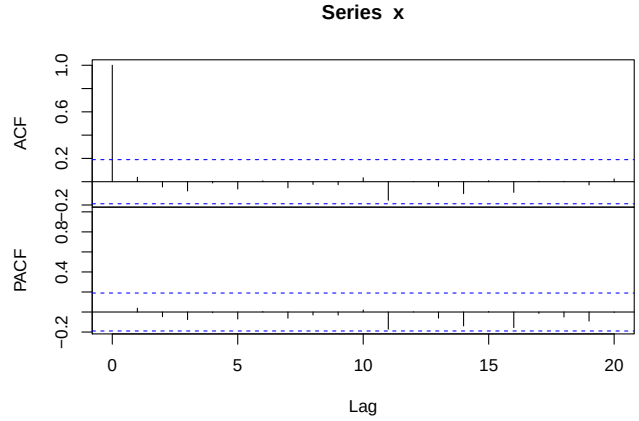
4. Effectuez ensuite l'estimation du modèle en spécifiant les ordres identifiés.
    a. L'espérance de la série est-elle différente de 0 ? Comment effectuer ce test d'hypothèse à partir des sorties du modèle ?

```
model = arima(sim1, order = c(0,0,1))
model
```

```
##
## Call:
## arima(x = sim1, order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##       0.6875     0.2866
## s.e.  0.0626     0.1559
##
## sigma^2 estimated as 0.9207:  log likelihood = -147.73,  aic = 301.45
```

5. Récupérez les résidus du modèle (via le getter `$residuals`). Tracez l'ACF et la PACF des résidus pour vérifier qu'ils ne présentent plus d'autocorrélation.
    a. S'il existe une autocorrélation résiduelle, le modèle doit alors être ajusté à partir de ces nouvelles observations : repassez à l'étape 3.

```
residus = model$residuals
Acf(residus)
```




```
# Les résidus ne présentent pas d'autocorrélation, et se comportent comme un bruit blanc.
# Le modèle est donc bien ajusté.
```

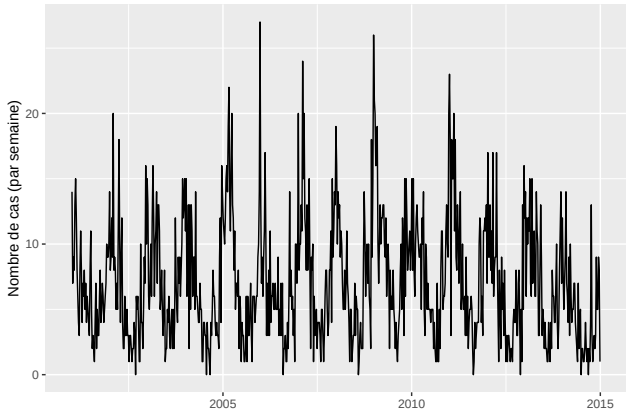
Partie 2 : manipulation d'une série issue de la surveillance épidémiologique

Téléchargez sur le site du cours le fichier `mp.RData` et importez-le dans votre session RStudio. Il contient un objet, `mp`, qui décrit l'évolution hebdomadaire du nombre de cas de méningites à pneumocoque, de l'année 2001 à l'année 2014, comme proxy des infections invasives à pneumocoque. On souhaite effectuer l'analyse temporelle de cette série, afin de pouvoir en effectuer des prédictions à court terme. Pour cela, nous allons estimer un modèle SARIMA.

0. Quel est le format de l'objet `mp` tel qu'il est stocké sur R ? Tracez la série `mp`. Des pistes sont données ci-dessous.

```
dates = mp %>% time() %>% as.numeric() %>% date_decimal() # vecteur dates
nombre = mp %>% as.numeric()                              # vecteur nombre de cas
meningites = tibble(dates = dates, nombre = nombre)       # data.frame combinant les deux
```

```r
ggplot(meningites, aes(x = dates, y = nombre)) +
    geom_line() +
    labs(x = "", y = "Nombre de cas (par semaine)")
```



1. Rappelez toutes les hypothèses rentrant en compte dans la modélisation par des processus SARIMA.
2. À l'instar de l'analyse des séries simulées en partie 1, modélisez la série `mp` par un processus SARIMA bien choisi.

La présence de très légers pics d'autocorrélation sur l'ACF des résidus peut laisser penser qu'il existe une structure d'autodépendance non prise en compte par le modèle, invalidant l'hypothèse que l'innovation de la série temporelle est un bruit blanc. Pour analyser plus finement l'ACF, nous proposons un test statistique, le test de Ljung-Box, qui s'intéresse à l'ensemble des pics d'autocorrélation jusqu'à un décalage H donné :

- Hypothèse nulle $\mathcal{H}_0$: $\quad \forall\, h = 1, \dots, H, \quad \rho(h) = 0,$
- Hypothèse alternative $\mathcal{H}_1$: $\exists\, h = 1, \dots, H, \quad \rho(h) \neq 0.$
- Statistique de test et sa loi sous H_0 :

$$Q = n(n+2) \sum_{h=1}^{H} \frac{\widehat{\rho}_h^2}{n-h} \underset{n\to\infty}{\sim} \chi^2_{ddl},$$

  où le nombre de degrés de liberté de la loi du χ^2_{ddl} est donné par $ddl = H - p - q - P - Q$, avec p, q, P et Q les ordres identifiés dans le modèle SARIMA$(p,d,q)(P,D,Q)_S$.
- Région de rejet univariée au niveau de risque α : $\mathcal{R} = \{Q > \chi^2_{ddl,1-\alpha}\}$.

La fonction `Box.test` implémente ce test sous R. L'argument `fitdf` de la fonction doit être égal à $p+q+P+Q$.

3. Effectuez le test de Ljung-Box sur les résidus du modèle identifié en question 1, jusqu'à un décalage maximum approprié (en général deux fois la saisonnalité de la série). Qu'en concluez-vous ? Reprenez le modèle estimé en question 1 si nécessaire.

On s'intéresse maintenant à valider l'hypothèse de normalité des innovations de la série.

4. Récupérez les résidus du modèle et tracez leur évolution. Que remarquez-vous ? Que faut-il prendre en compte ?
5. Effectuez une analyse de la normalité des résidus. Pour cela, vous pouvez vous aider d'un QQ-plot, d'un PP-plot, de tests de normalité, etc. Qu'en concluez-vous ?
6. Ecrivez la forme simplifiée du modèle choisi, du type SARIMA$(p,d,q)(P,D,Q)_S$. Ecrivez également la forme détaillée du modèle, en explicitant la formule de récurrence suivie par la série.
7. Ecrivez l'intervalle de confiance pour le paramètre de long terme de la structure d'autocovariance du processus.

On s'intéresse désormais à effectuer des prédiction du modèle à court terme, afin de déterminer l'évolution probable du nombre de méningites à pneumocoque au début de l'année 2015.

8. Rappelez comment est prédite l'observation x_{n+m} en fonction des données observées $x_1, \ldots, x_n$.

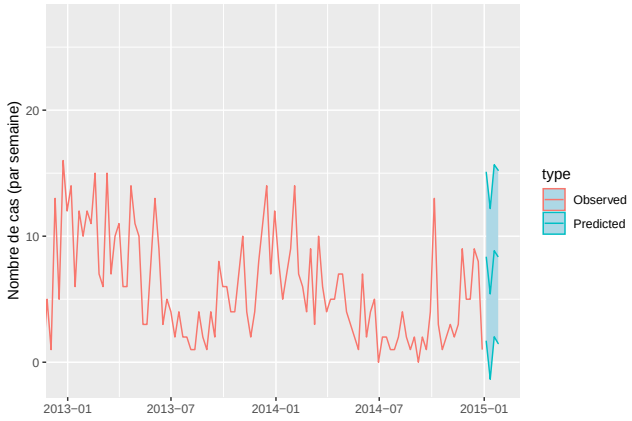
9. Effectuez et tracez les prévisions de la série sur un mois. Sur la figure, faites apparaitre l'intervalle de prédiction à 95%. Des pistes sont données ci-dessous.

```
predicted = predict(model, n.ahead = 4, se.fit = TRUE) # Prévisions à 4 semaines

# Création d'un data.frame qui contient les prévisions et les intervalles de
# prédiction des prévisions
predicted_df = tibble(dates = predicted$pred %>% time() %>% as.numeric() %>% date_decimal(),
                      nombre = predicted$pred %>% as.numeric(),
                      high = predicted$pred + 1.96 * predicted$se %>% as.numeric(),
                      low = predicted$pred - 1.96 * predicted$se %>% as.numeric(),
                      type = "Predicted")

# On combine le data.frame des prévisions avec celui des observations
meningites = meningites %>%
    mutate(type = "Observed") %>%
    bind_rows(predicted_df)

# Graphique des prévisions
ggplot(meningites, aes(x = dates, y = nombre, col = type, fill = type)) +
    geom_ribbon(aes(ymin = low, ymax = high), fill = "lightblue") +
    geom_line() +
    coord_cartesian(xlim = c(as_datetime("2013-01-01"), max(predicted_df$dates))) +
    labs(x = "", y = "Nombre de cas (par semaine)")
```



10. A partir des prédictions, déterminez la probabilité que le nombre de méningites à pneumocoque dépasse 10 la dernière semaine de janvier.

11. Répétez maintenant la question 9, en effectuant des prédictions sur trois ans. Que remarquez-vous ? Comment pouvez-vous l'expliquer ?