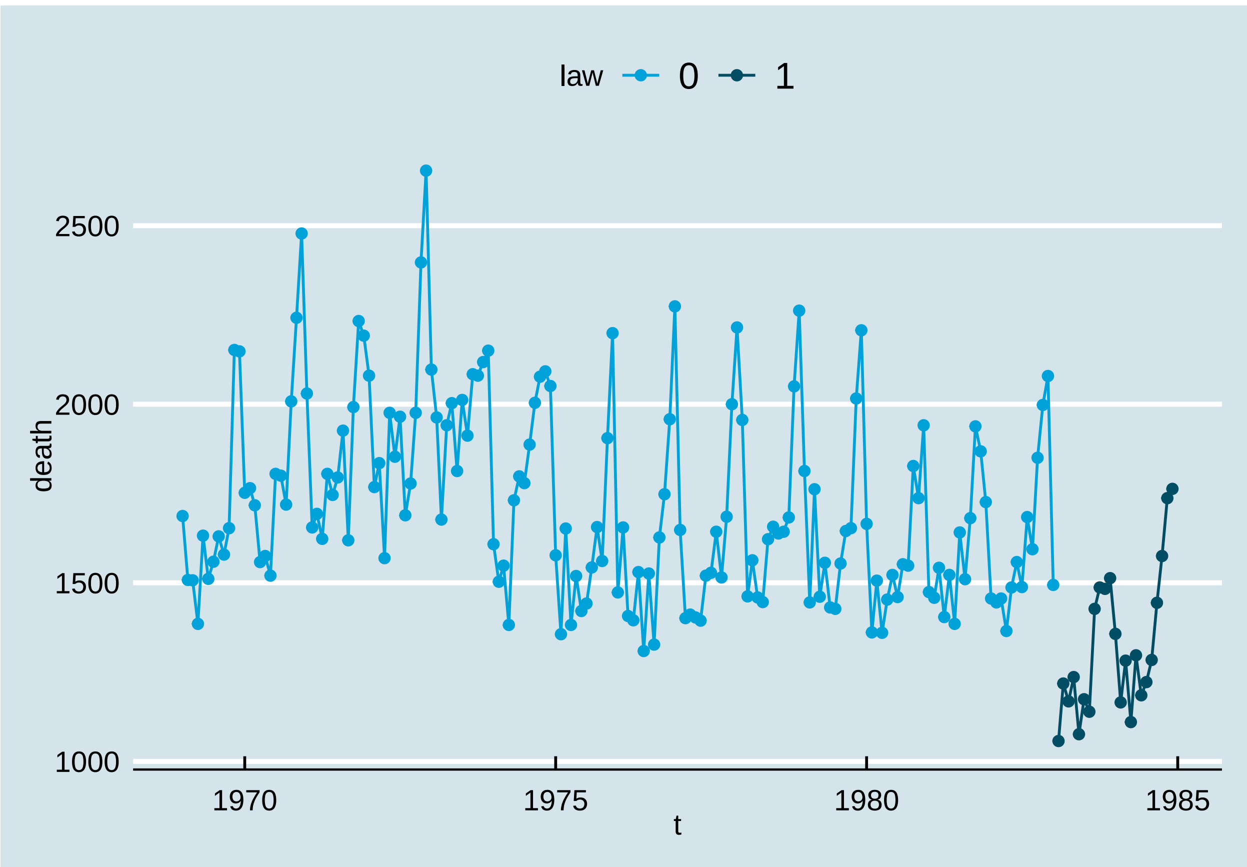
TS Modèles v.1

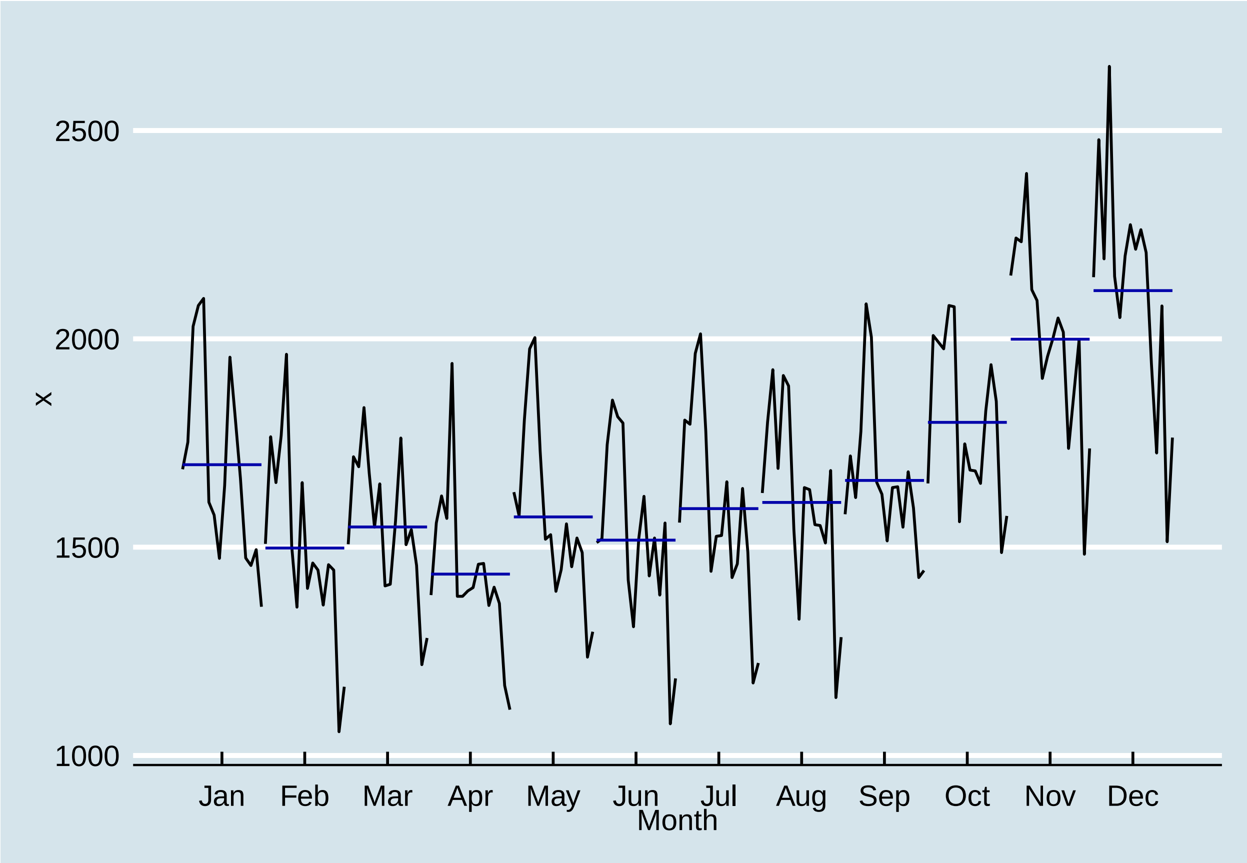
Loïc

Contexte

```
uk <- read_table("../data.txt", col_types = "if")
period <- seq(as.Date('1969-01-01'), as.Date('1984-12-31'), by = "month")
uk_ts <- ts(uk$death, start = c(1969,1), frequency = 12)
uk_df <- bind_cols(uk, t = period)
```



Ici, on cherche à démontrer que la périodicité est annuelle.
On peut remarquer sur chaque "petite série temporelle mensuelle" la tendance qui décroit.



Modèle 1: Régression linéaire

On cherche un modèle paramétrique de la forme:

$$\text{death} = m_t + s_t + \epsilon_t$$

1. Régression sur la tendance

J'ai laissé le modèle "optimal" (*en bleu foncé*) sur chaque plot.
Les modèles qu'on ajuste s'appelle "fit" (*en bleu clair*).

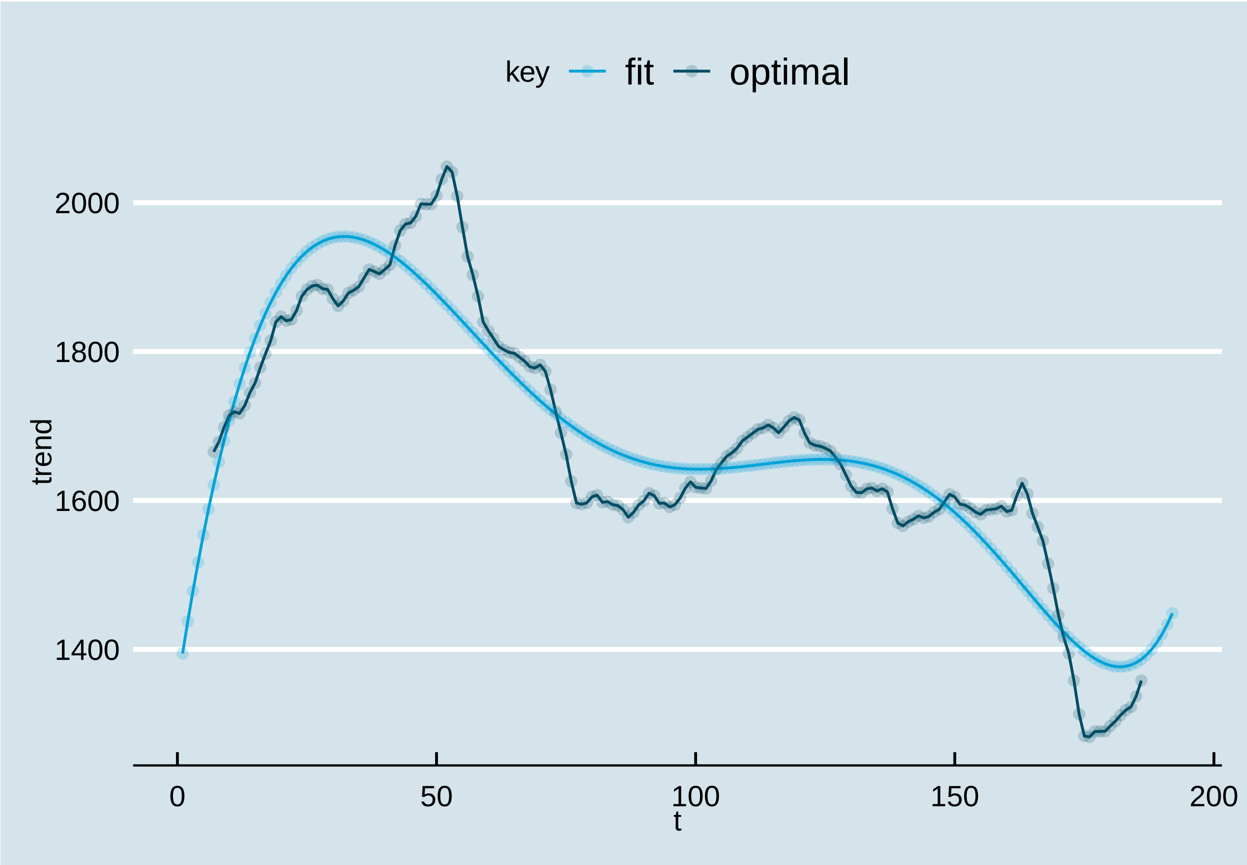
```
mod_fit <- lm(death ~ t + t2 + t3 + t4 + t8, data = df_fit)
```

J'ai essayé plusieurs modèles, j'ai essayé step aussi mais j'ai du réduire encore pour arriver à

$$m_t = 1349 + 46.5t - 1.17t^2 + \frac{t^3}{100} - 3.432 * 10^{-5}t^4 + 2.613 * 10^{-15}t^8$$

On peut même enlever t^4 et t^8 puisque leur coefficient sont proches de 0.

```
##
## Call:
## lm(formula = death ~ t + t2 + t3 + t4 + t8, data = df_fit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -391.81 -185.02  -44.44  118.54  763.27
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.349e+03  1.021e+02  13.215  < 2e-16 ***
## t            4.650e+01  9.236e+00   5.035 1.12e-06 ***
## t2          -1.171e+00  2.484e-01  -4.715 4.73e-06 ***
## t3           1.076e-02  2.526e-03   4.260 3.25e-05 ***
## t4          -3.432e-05  8.787e-06  -3.906 0.000131 ***
## t8           2.613e-15  8.831e-16   2.958 0.003495 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 241.5 on 186 degrees of freedom
## Multiple R-squared:  0.3228, Adjusted R-squared:  0.3046
## F-statistic: 17.73 on 5 and 186 DF,  p-value: 2.367e-14
```



2. Régression sur la saisonalité

On part du même principe que les TP avec $cos(\frac{\pi kt}{6})$ et $sin(\frac{\pi kt}{6})$
Dans le code, quand c'est `cos k` $= cos(\frac{\pi kt}{6})$, pareil pour `sin k`

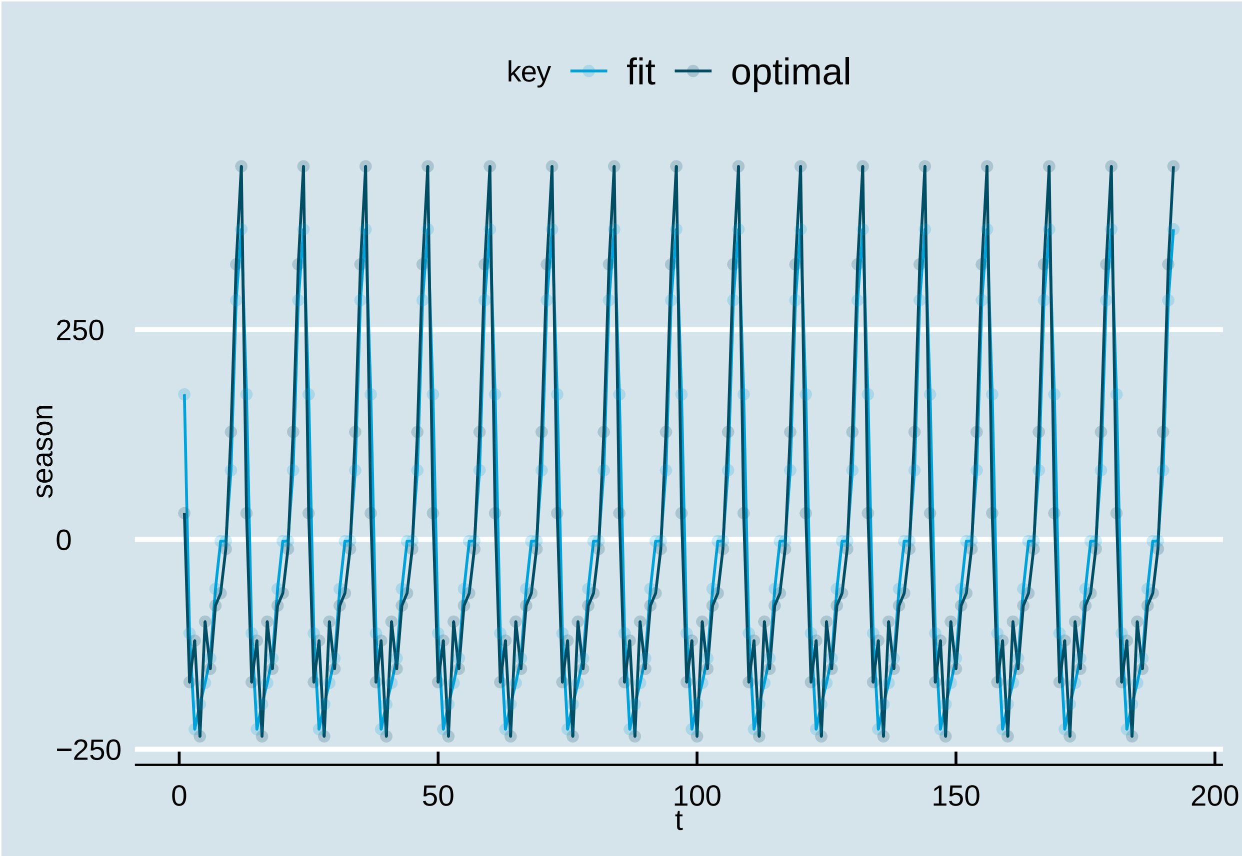
```
mod_fit_season <- lm(season ~ sin1 + cos1 + cos2 + cos3, data = df_fit)
```

Donc après plusieurs essais qui vont dans tous les sens, on arrive à un modèle qui fit bien:

$$s_t = 199\,cos(\frac{\pi t}{6}) + 114\,cos(\frac{\pi t}{3}) + 57\,cos(\frac{\pi t}{2}) - 112\,sin(\frac{\pi t}{6})$$

On remarque d'ailleurs que ça fit nettement bien avec l'optimal.

```
##
## Call:
## lm(formula = season ~ sin1 + cos1 + cos2 + cos3, data = df_fit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141.53  -43.24  -11.07   52.51  105.53
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.273e-14  4.968e+00    0.00        1
## sin1        -1.121e+02  7.026e+00  -15.96  < 2e-16 ***
## cos1         1.985e+02  7.026e+00   28.25  < 2e-16 ***
## cos2         1.140e+02  7.026e+00   16.22  < 2e-16 ***
## cos3         5.698e+01  7.026e+00    8.11 6.51e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.85 on 187 degrees of freedom
## Multiple R-squared:  0.8808, Adjusted R-squared:  0.8782
## F-statistic: 345.4 on 4 and 187 DF,  p-value: < 2.2e-16
```

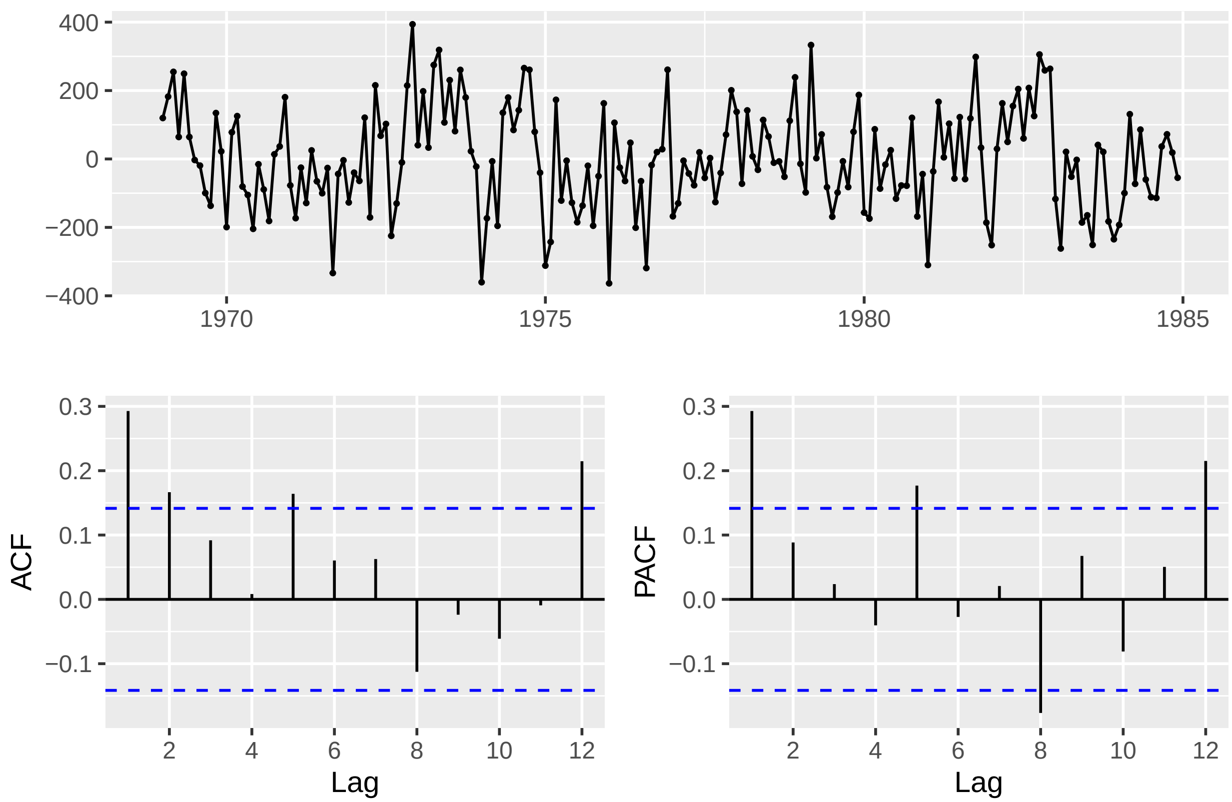

key
fit
optimal
season
250
0
−250
0
50
100
150
200
t

3. Résidus du modèle de régression

Comme c'est un modèle additif, il s'agit de la part non expliquée par la tendance et la saisonalité.

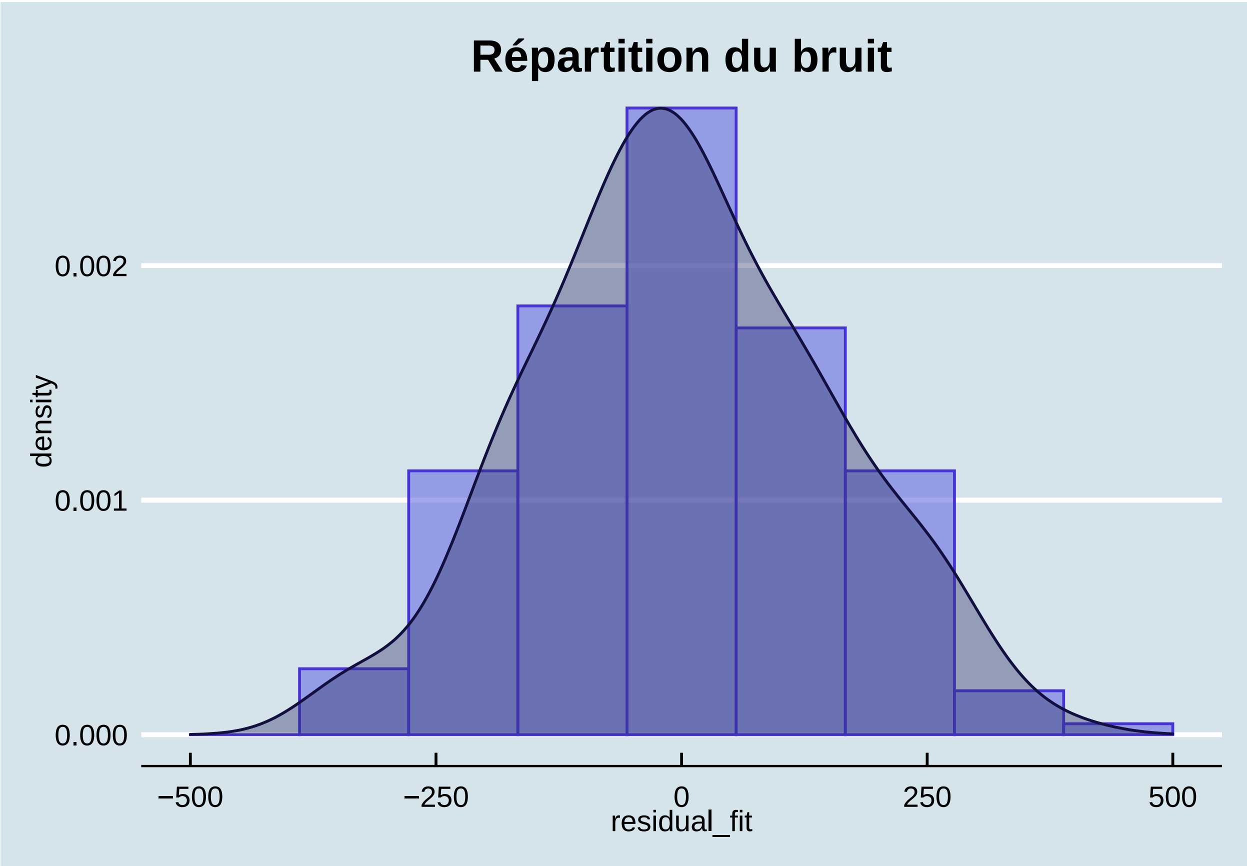
Ni l'ACF ni le PACF ne semblent décroitre du coup si on se base sur ce modèle ce sera ARMA ? ou ARIMA ?

```
residual_fit <- uk_ts - trend_fit - seasonal_fit
```



```r
shapiro.test(residual_fit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residual_fit
## W = 0.99425, p-value = 0.6684
```



Donc ça suit une loi normale, et on obtient:

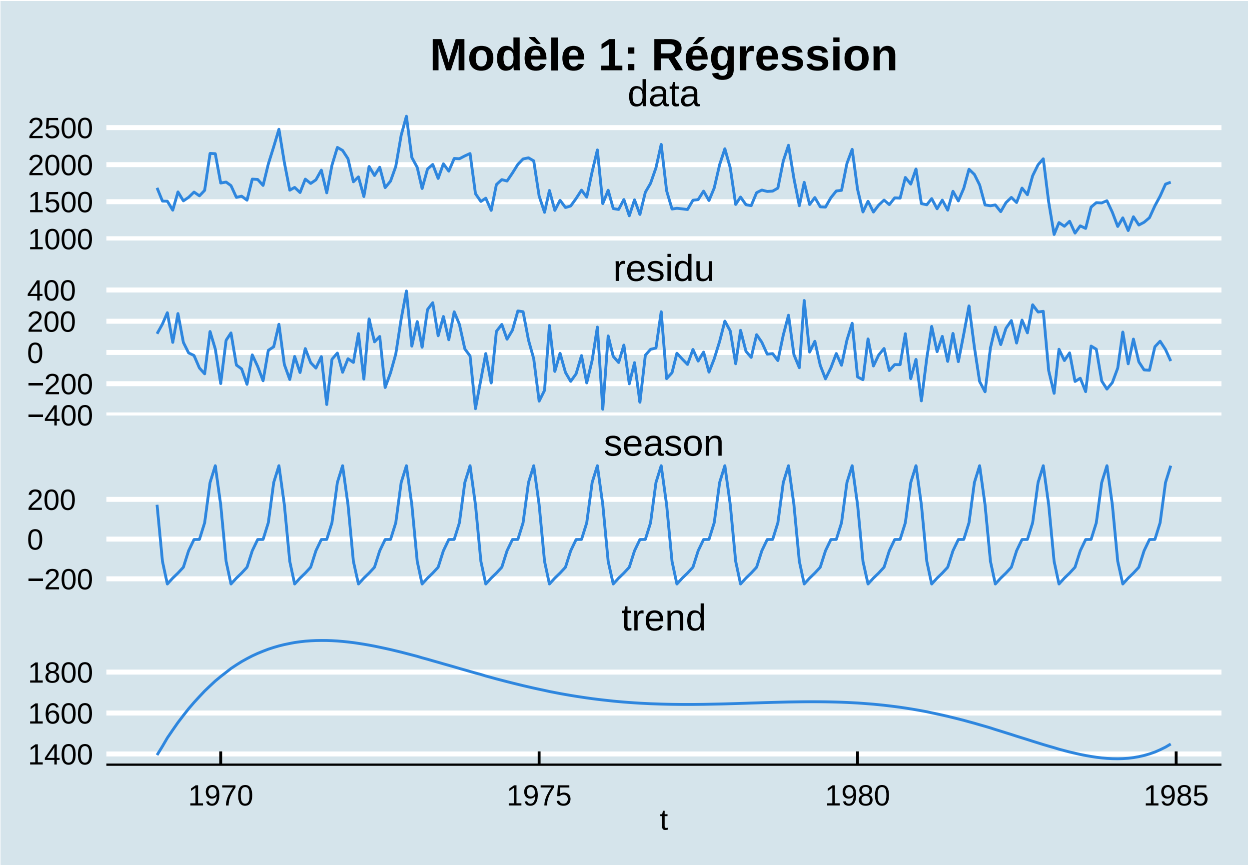
$$\epsilon_t \sim \mathcal{N}(\mu = 0, \sigma^2 = 22519)$$

4. Décomposition

Donc on a un modèle paramétrique de la forme

$$\text{death} = m_t + s_t + \epsilon_t$$

$$\text{avec} \begin{cases} m_t = 1349 + 46.5t - 1.17t^2 + \frac{t^3}{100} - 3.432 * 10^{-5}t^4 + 2.613 * 10^{-15}t^8 \\ s_t = 199\, cos(\frac{\pi t}{6}) + 114\, cos(\frac{\pi t}{3}) + 57\, cos(\frac{\pi t}{2}) - 112\, sin(\frac{\pi t}{6}) \\ \epsilon_t \sim \mathcal{N}(0, 22\,519) \end{cases}$$



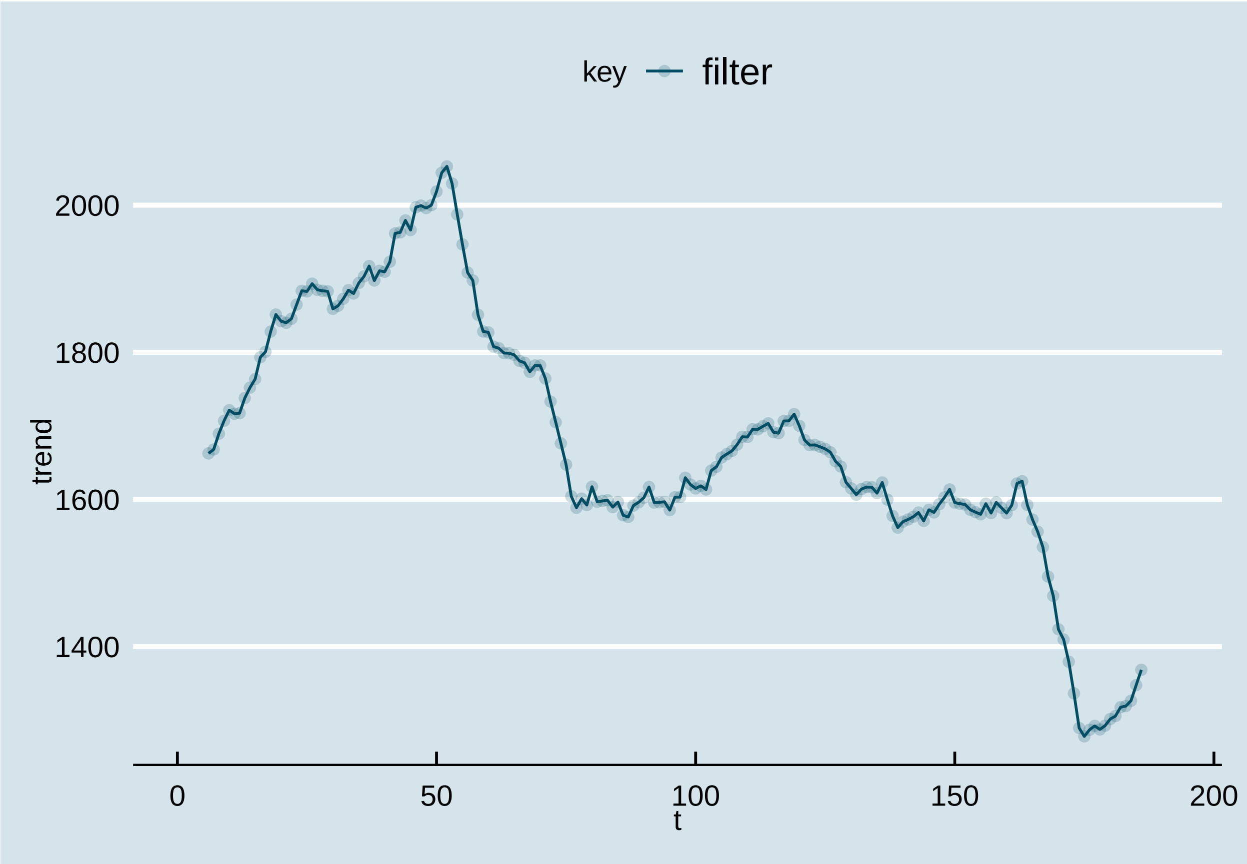
Modèle 2: Filtre linéaire

Alors le but ici c'est de ne pas passer par une régression linéaire mais plutôt par un filtre linéaire sur une moyenne mobile.
On n'aura pas forcément un modèle paramétrique dans ce cas, bien qu'on peut estimer les paramètres.

1. Filtre sur la tendance

Comme on a dit qu'il s'agit de périodes de 12 mois, on va donc appliquer un filtre linéaire à notre série temporelle avec une moyenne mobile sur 12 périodes.
J'ai pas mis la comparaison avec la tendance "optimale" puisque cette technique de filtre se superpose avec, du coup c'est l'optimale dans notre cas.

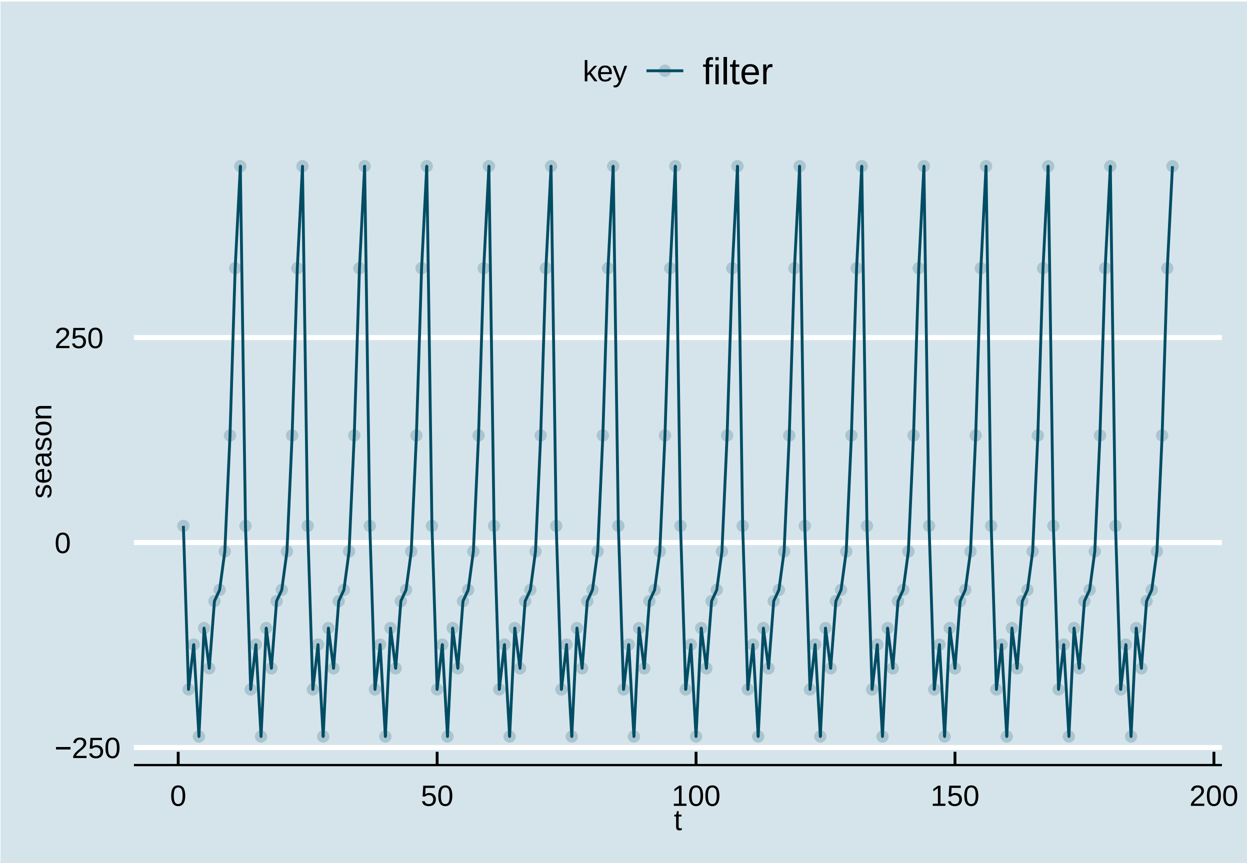
```
trend_filter <- stats::filter(
    uk_ts, rep(1/12, 12),
    method = "convolution",
    sides = 2
)
```



2. Filtre sur la saisonalité

L'idée ici est de calculer la saisonalité sur la série sans la tendance.
Comme au tout début on a noté une périodicité annuelle, on déduira la saisonalité par la moyenne mensuelle.

```
seasonal_filter <- (uk_ts - trend_filter) %>%
matrix(12) %>% t() %>%
colMeans(na.rm = T) %>%
rep(length(uk_ts)/12)
```



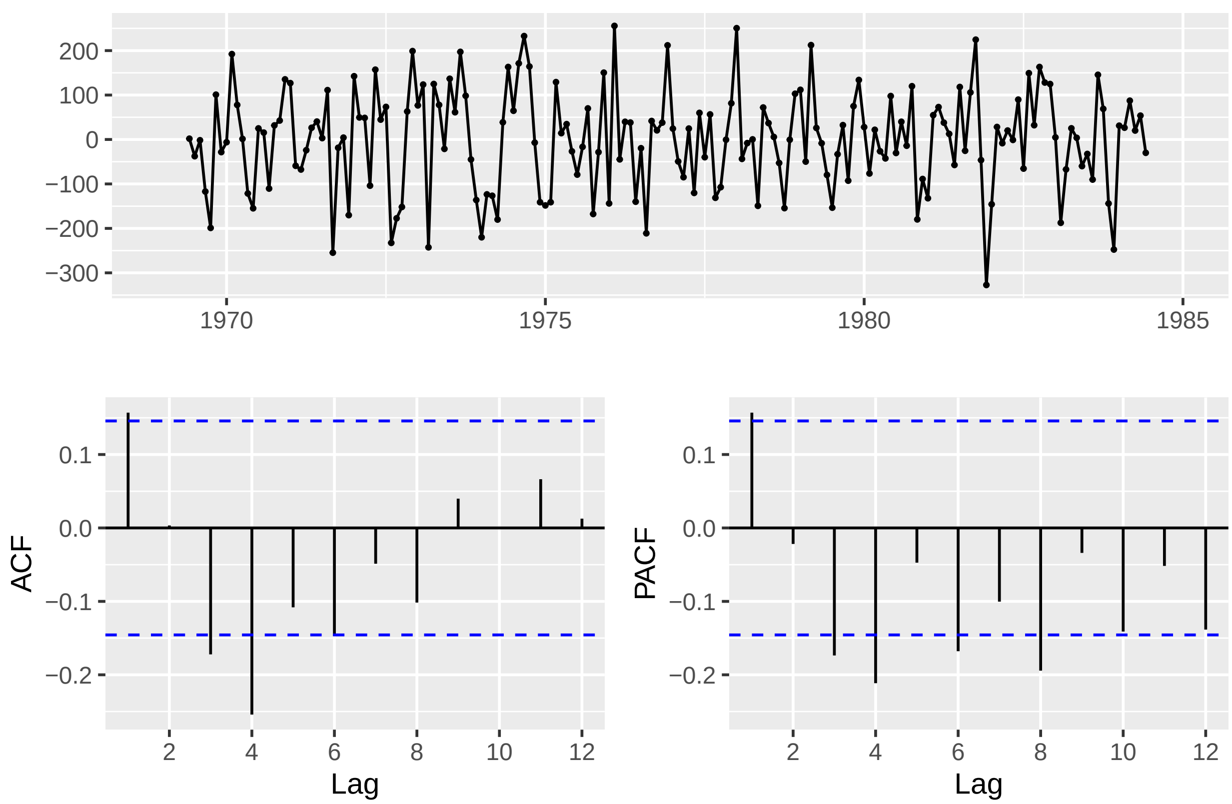
On arrive quasiment à la saisonalité du *modèle 1 avec la Régression linéaire*, l'amplitude est juste plus grande sur ce modèle au mois de décembre.

3. Filtre sur les résidus

Comme précédement, c'est un modèle additif donc let's go.

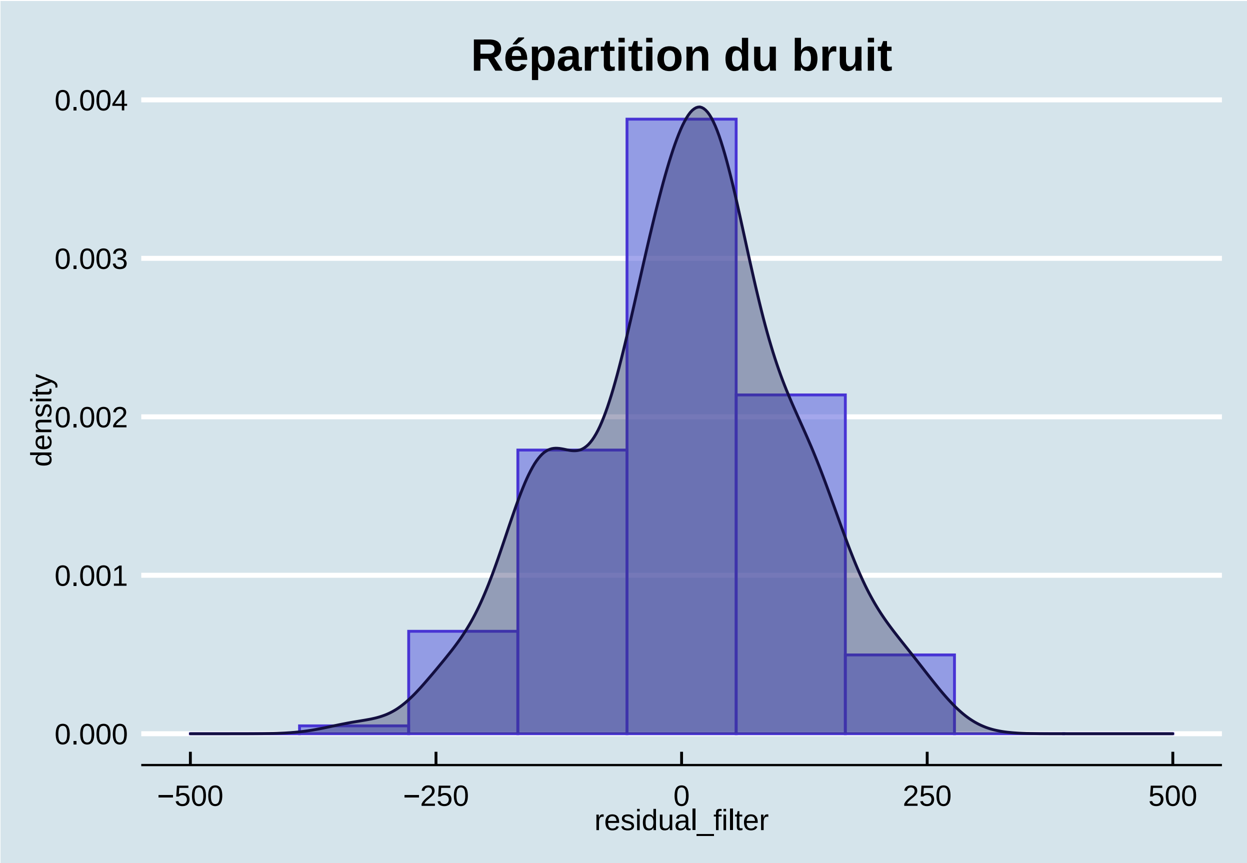
Là on a un ACF qui décroit, on part sur un MA avec un lag = 4.

```
residual_filter <- uk_ts - trend_filter - seasonal_filter
```



```r
shapiro.test(residual_filter)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residual_filter
## W = 0.9913, p-value = 0.3456
```



Donc ça suit une loi normale, et on obtient:

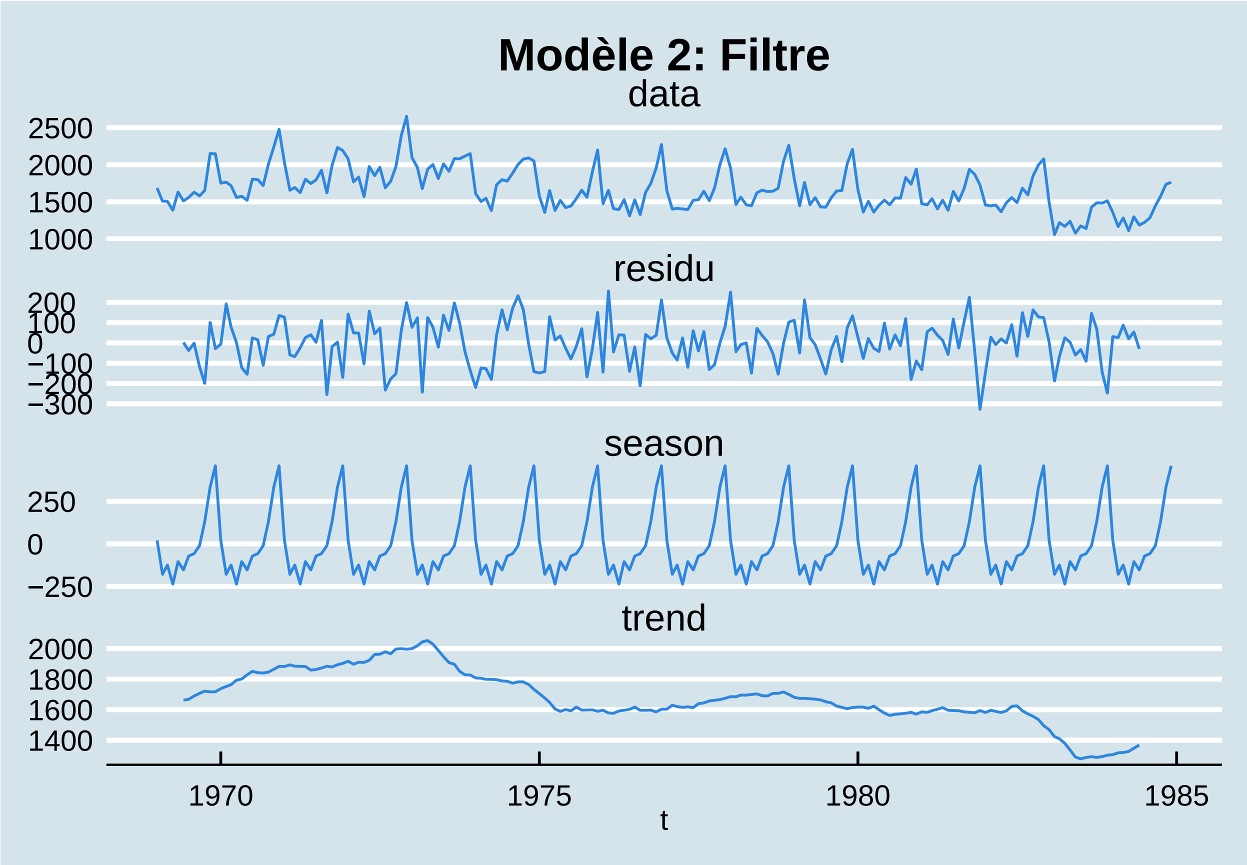
$$\epsilon_t \sim \mathcal{N}(\mu = 0, \sigma^2 = 12\ 446)$$

4. Décomposition

Donc on a un modèle dont on connaît certains paramètres de la forme

$$\text{death} = m_t + s_t + \epsilon_t$$

$$\text{avec} \begin{cases} s_t = \beta_0 + \beta_1\, cos(\frac{\pi t}{6}) + \beta_2\, cos(\frac{\pi t}{3}) + \beta_3\, cos(\frac{\pi t}{2}) - \beta_4\, sin(\frac{\pi t}{6}) + (\tau) \\ \epsilon_t \sim \mathcal{N}(0, 12\,446) \end{cases}$$



Modèle 3: Décomposition automatique

C'est ce qu'on a fait dans le pdf que j'ai envoyé (*explorer.pdf*), du coup je ne vais pas la refaire ici.
Il faut quand même noter que le modèle par filtre et la décomposition automatique donnent des résultats proches.

Remarque:
Les modèles 2 et 3 sont meilleurs que 1 puisque leurs résidus capturent moins de données. Ca voudrait donc dire qu'on a plus de contrôle sur le modèle et qu'il y a moins d'aléa.