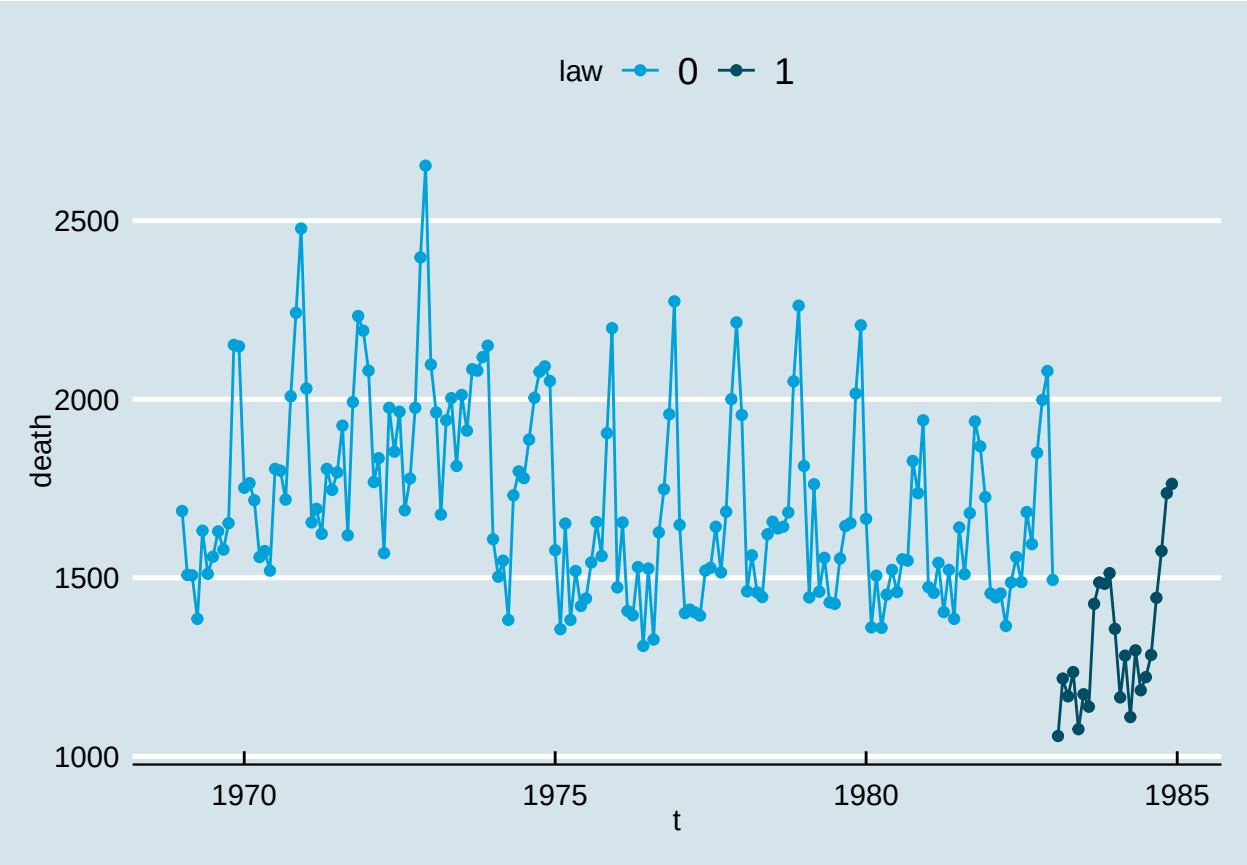
TS Modèles v.1

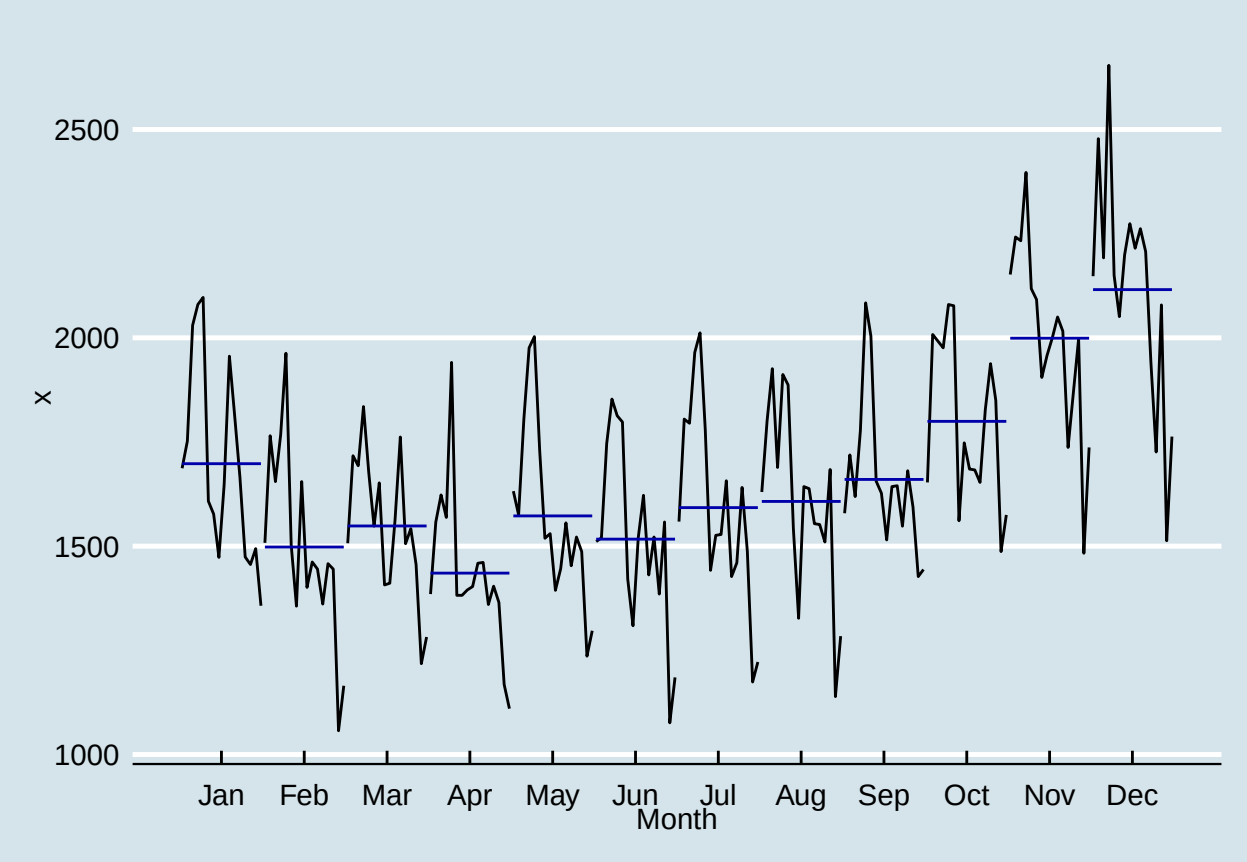
Loïc

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.



COUCOU DYLAN !!! ON COMMENCE ICI POUR MOI DU COUP COMME TU AS UNE MEILLEURE ANALYSE DE DONNEES !

Modélisation

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

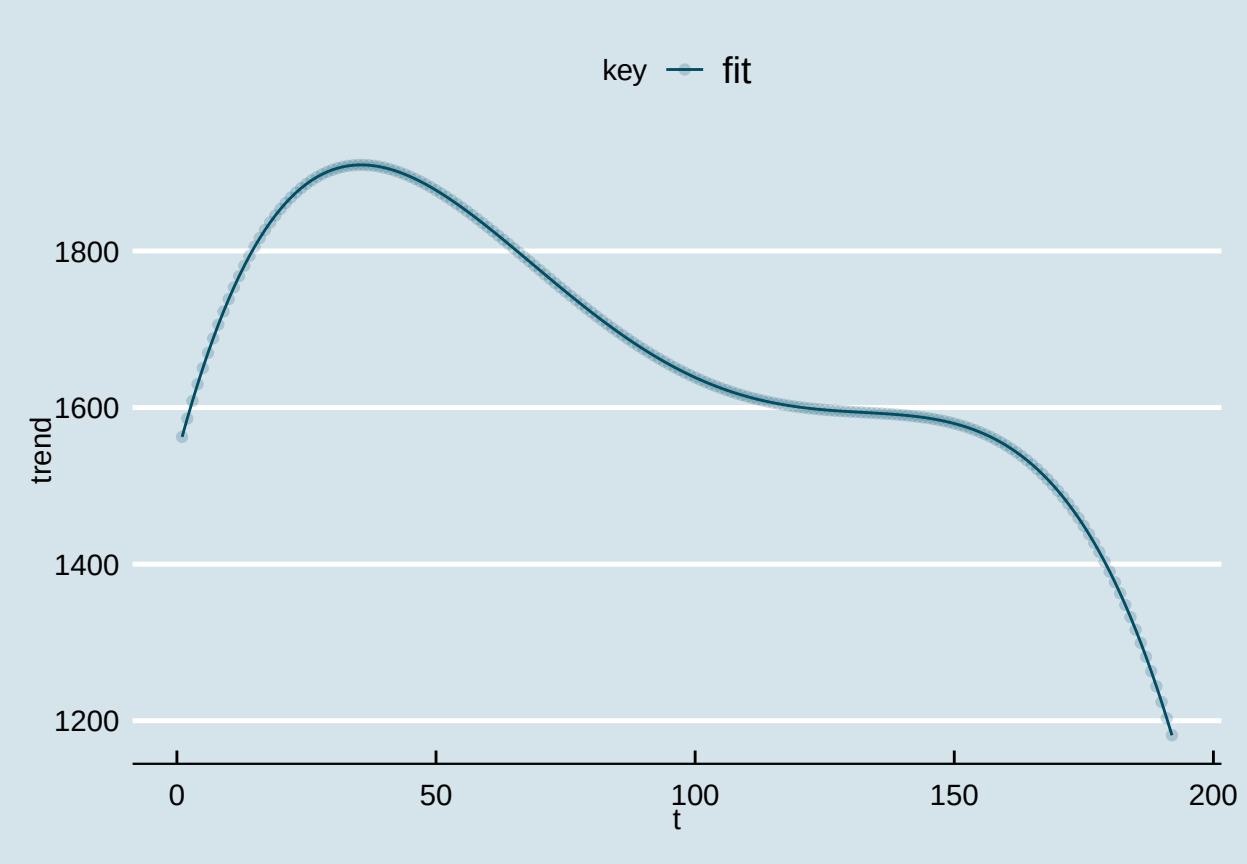
J'ai laissé le modèle "optimal" (*en bleu foncé*) sur chaque plot.
Les modèles qu'on ajuste s'appelle "fit" (*en bleu clair*).

```r
mod_fit <- lm(death ~ ., data = df_fit)
bestlm <- step(mod_fit, trace = F, direction = "both")
```

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

$$\hat{m}(t) = 1538 + 25.07t - 0.54t^2 + \frac{249t^3}{62500} - 9.906 * 10^{-6}t^4$$

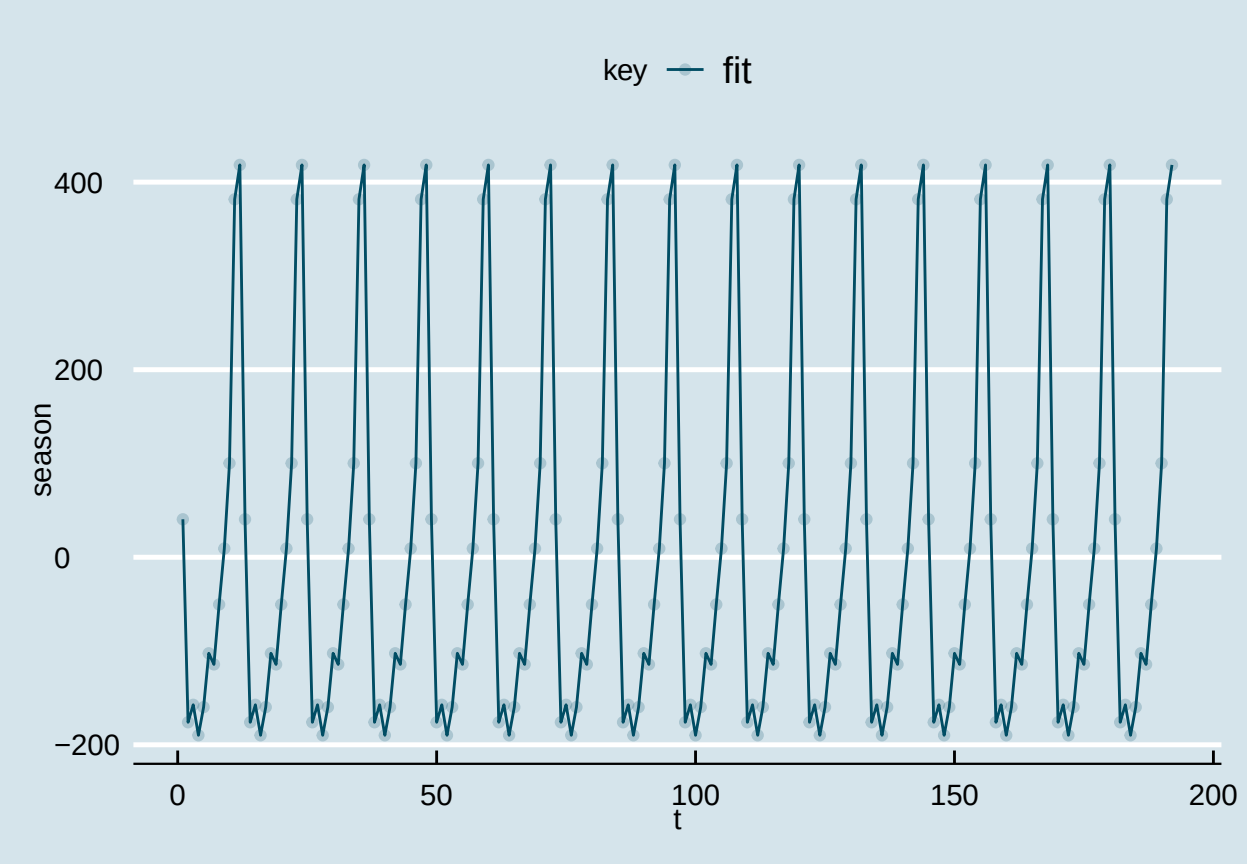
```
##
## Call:
## lm(formula = death ~ t + t2 + t3 + t4 + sin1 + sin2 + sin3 +
##     sin4 + sin5 + cos1 + cos2 + cos3 + cos4, data = df_fit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -312.11  -81.49   -3.82   86.74  348.30
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.538e+03  5.394e+01  28.523  < 2e-16 ***
## t            2.507e+01  3.846e+00   6.518 7.08e-10 ***
## t2          -5.396e-01  8.068e-02  -6.688 2.82e-10 ***
## t3           3.984e-03  6.272e-04   6.352 1.72e-09 ***
## t4          -9.906e-06  1.612e-06  -6.145 5.10e-09 ***
## sin1        -1.204e+02  1.480e+01  -8.137 6.75e-14 ***
## sin2        -6.241e+01  1.476e+01  -4.229 3.75e-05 ***
## sin3        -3.669e+01  1.475e+01  -2.487  0.01379 *
## sin4        -2.294e+01  1.475e+01  -1.555  0.12161
## sin5         2.252e+01  1.475e+01   1.527  0.12845
## cos1         2.014e+02  1.475e+01  13.656  < 2e-16 ***
## cos2         1.164e+02  1.475e+01   7.895 2.88e-13 ***
## cos3         5.938e+01  1.475e+01   4.026 8.38e-05 ***
## cos4         4.197e+01  1.475e+01   2.846  0.00495 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144.5 on 178 degrees of freedom
## Multiple R-squared:  0.7681, Adjusted R-squared:  0.7512
## F-statistic: 45.35 on 13 and 178 DF,  p-value: < 2.2e-16
```



2. Régression sur la saisonalité

D'après notre modèle de régression, on estime la saisonalité comme suit:

$$\hat{s}_t = \begin{cases} 201\, cos(\frac{\pi t}{6}) - 120\, sin(\frac{\pi t}{6}) + \\ 116\, cos(\frac{\pi t}{3}) - 62.41\, sin(\frac{\pi t}{3}) + \\ 59.38\, cos(\frac{\pi t}{2}) - 36.69\, sin(\frac{\pi t}{2}) + \\ 41.97\, cos(\frac{\pi 2t}{3}) - 22.94\, sin(\frac{\pi 2t}{3}) \end{cases}$$

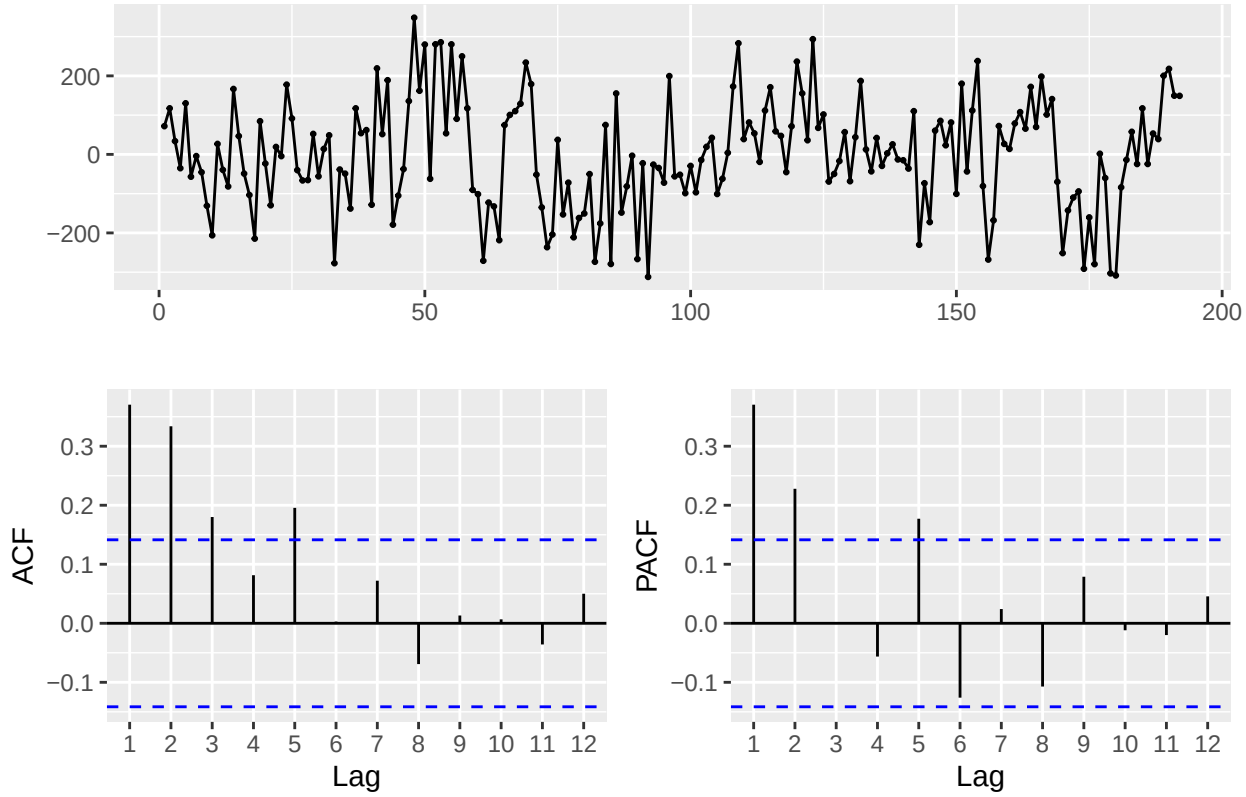

key
fit
400
200
0
−200
season
0
50
100
150
200
t

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.

L'ACF décroit rapidement.

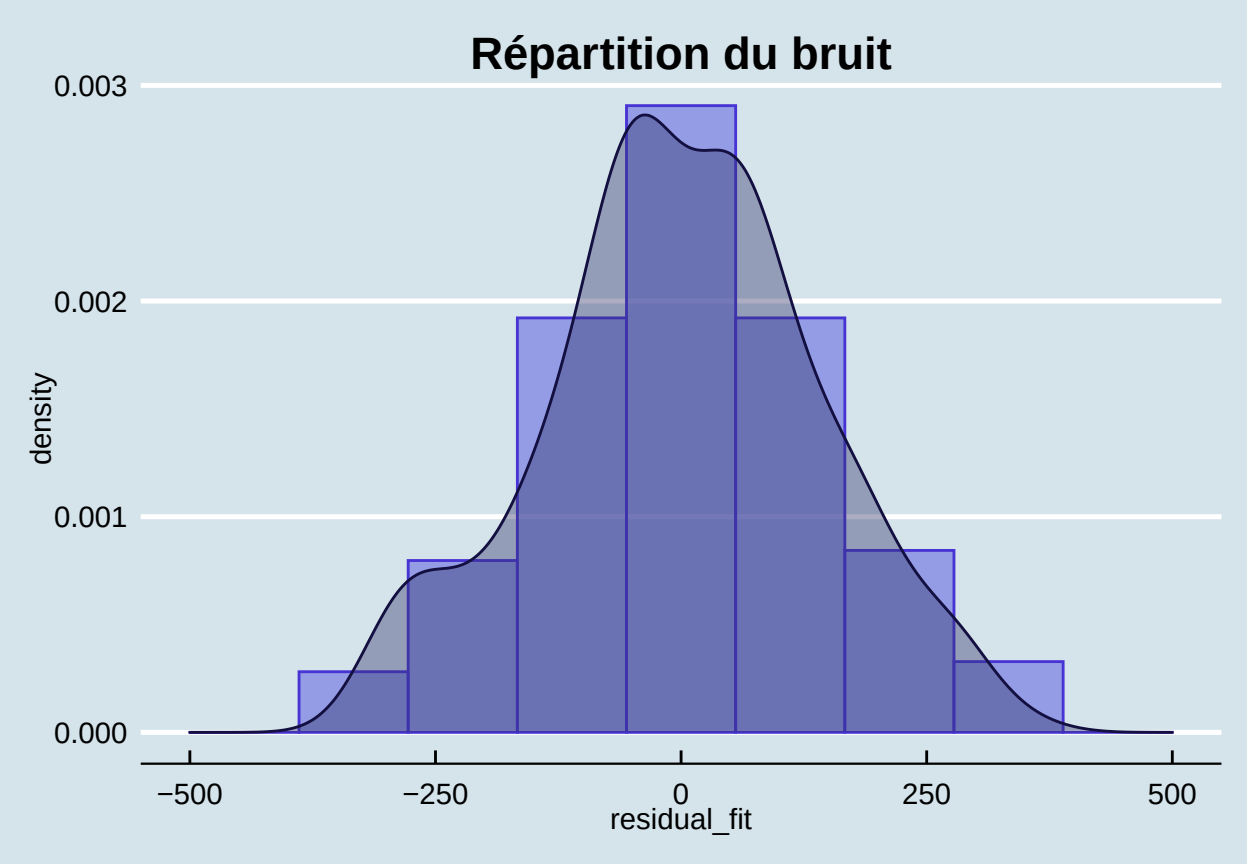
```
residual_fit <- bestlm$residuals
```



Le test de Shapiro-Wilk ne rejette pas l'hypothèse de la normalité de nos résidus.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  residual_fit
## W = 0.99061, p-value = 0.2445
```



Donc $\hat{\epsilon}$ suit une loi normale, et on obtient:

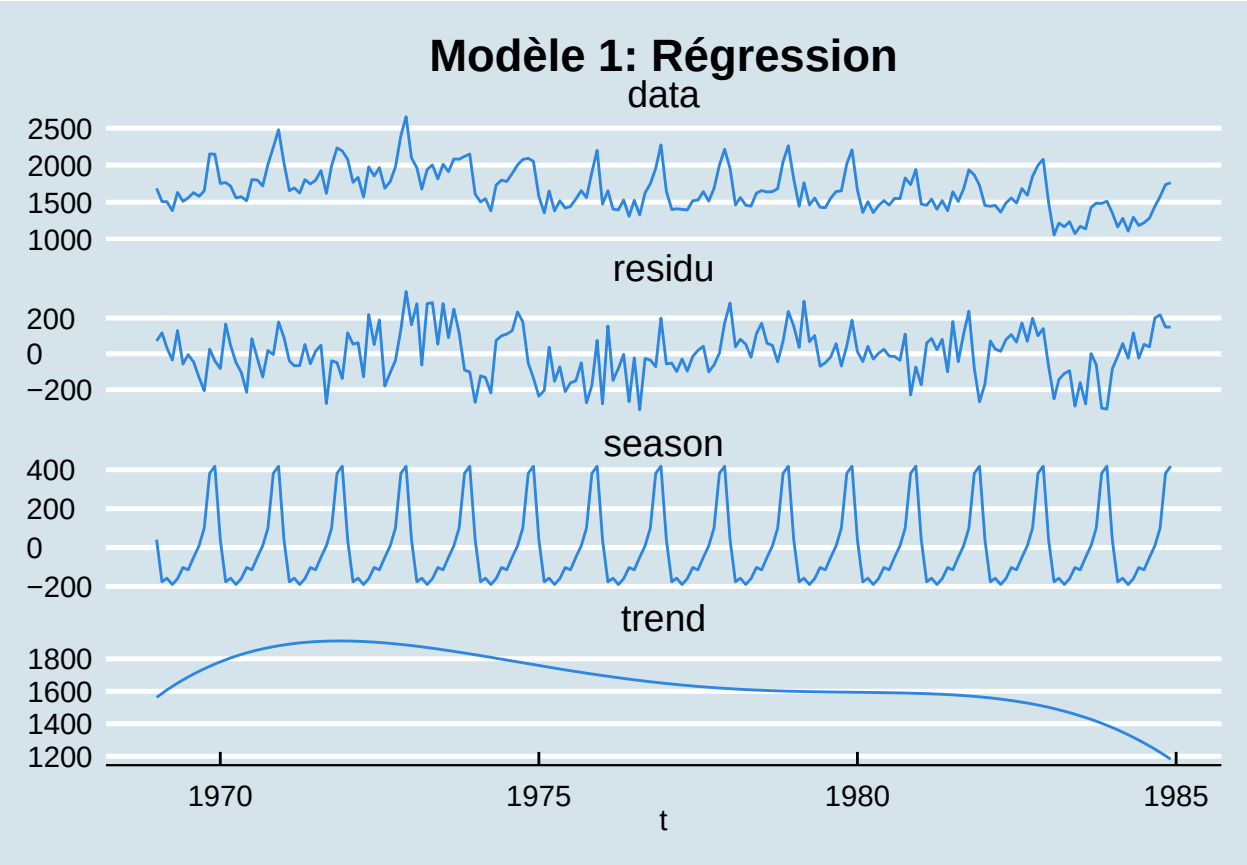
$$\hat{\epsilon}_t \sim \mathcal{N}(\mu = 0, \sigma^2 = 19\,451)$$

4. Décomposition

Nous obtenons ainsi un premier modèle, paramétrique, de la forme

$$\text{death} = \hat{m}(t) + \hat{s}_t + \hat{\epsilon}_t$$

$$\text{avec} \begin{cases} \hat{m}(t) = 1538 + 25.07t - 0.54t^2 + \frac{249t^3}{62500} - 9.906 * 10^{-6}t^4 \\ \hat{s}_t = ... \\ \epsilon_t \sim \mathcal{N}(0, 19\,451) \end{cases}$$



Modèle 2: Régression linéaire sur le log

COUCOU DYLAN DU COUP JE DCRIS ICI

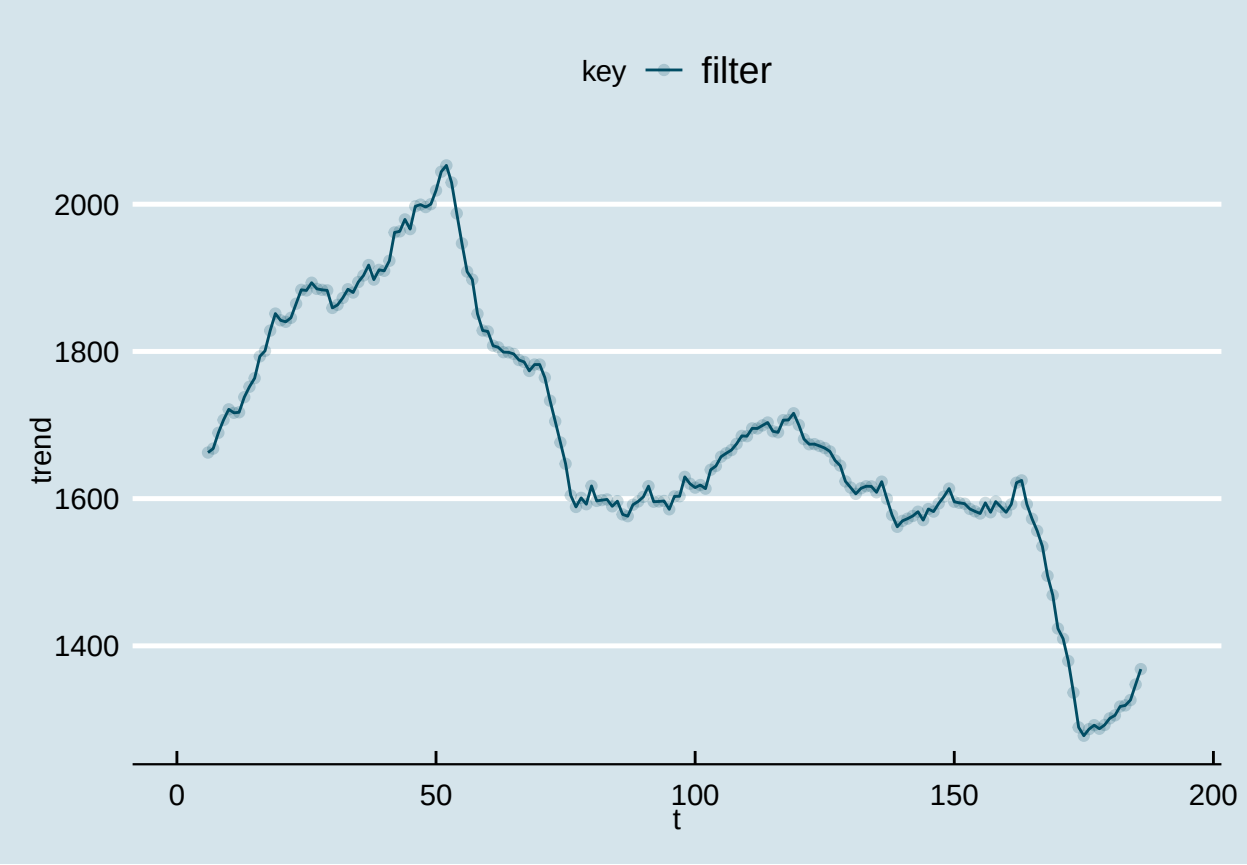
Modèle 3: Filtre linéaire

On suppose toujours death $= m_t + s_t + \epsilon_t$.
Au lieu de passer par une régression linéaire on applique plutôt un filtre linéaire sur une moyenne mobile (MA).
C'est une méthode locale non-paramétrique permettant de filtrer la tendance en éliminant la saisonalité.
On estimera ensuite cette dernière avec la moyenne arithmétique des mois et on en déduira les résidus.

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.

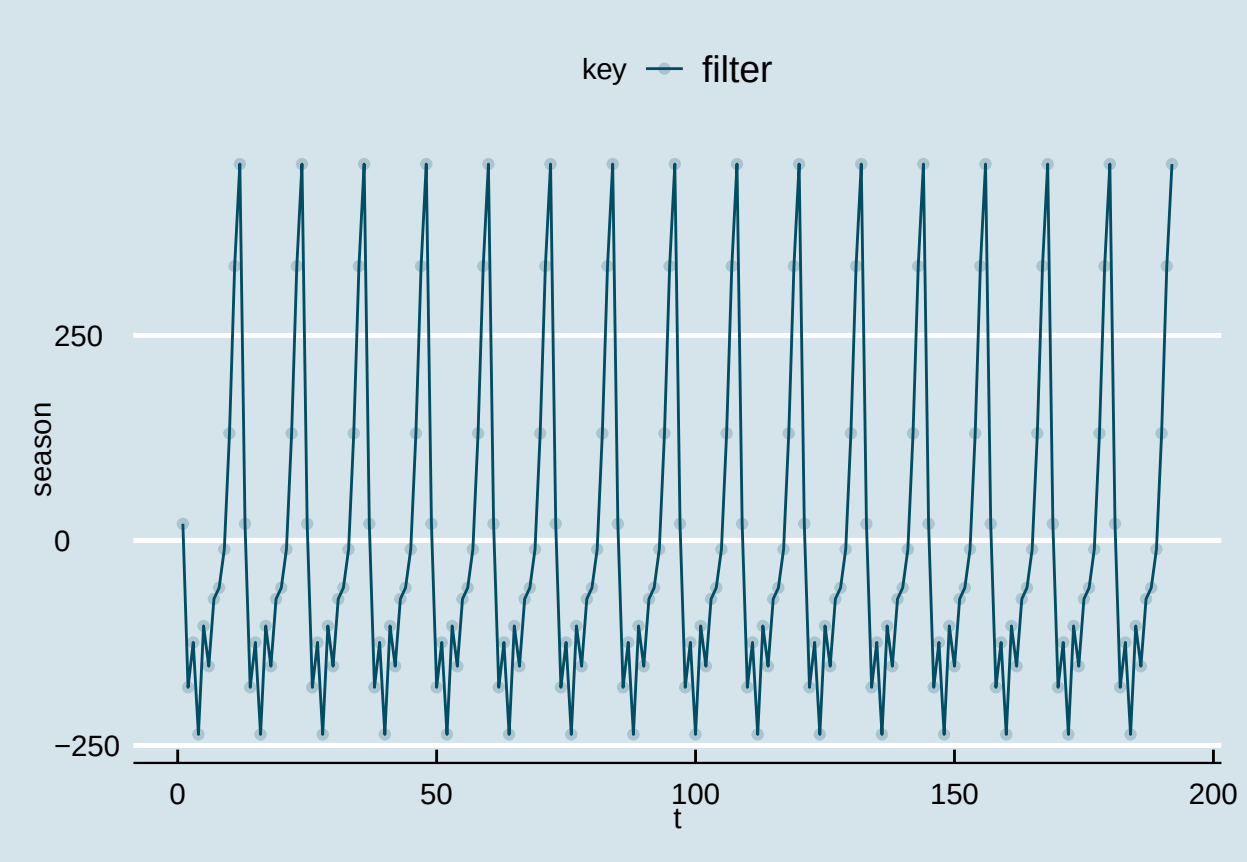
```
trend_filter <- stats::filter(
    uk_ts, rep(1/12, 12),
    method = "convolution",
    sides = 2
)
```



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 arithmétique mensuelle.

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



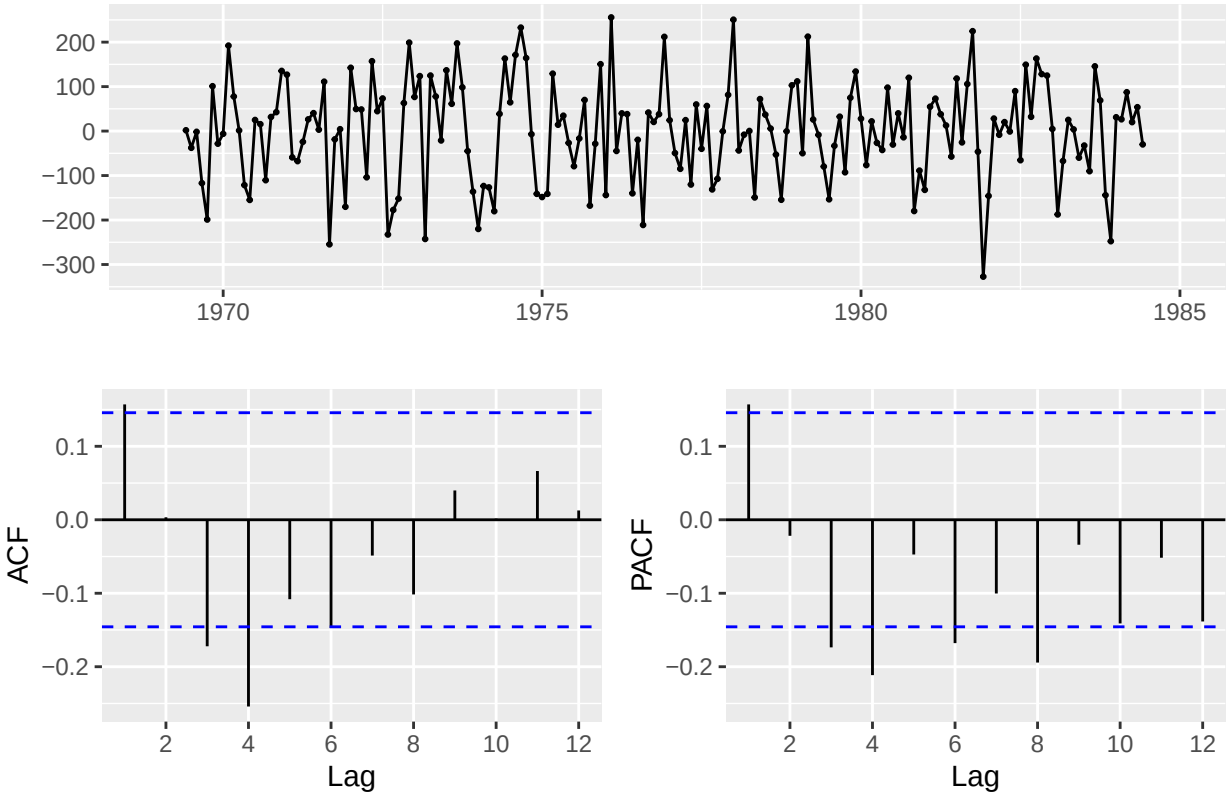
On arrive quasiment à la saisonalité du *modèle 1 avec la Régression linéaire*, mais avec une magnitude plus élevée.

3. Filtre sur les résidus

Nous avons notre modèle de tendance et de saisonalité. On en déduit alors les résidus: $\hat{\epsilon} = \text{death} - \hat{m}(t) - \hat{s}_t$

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

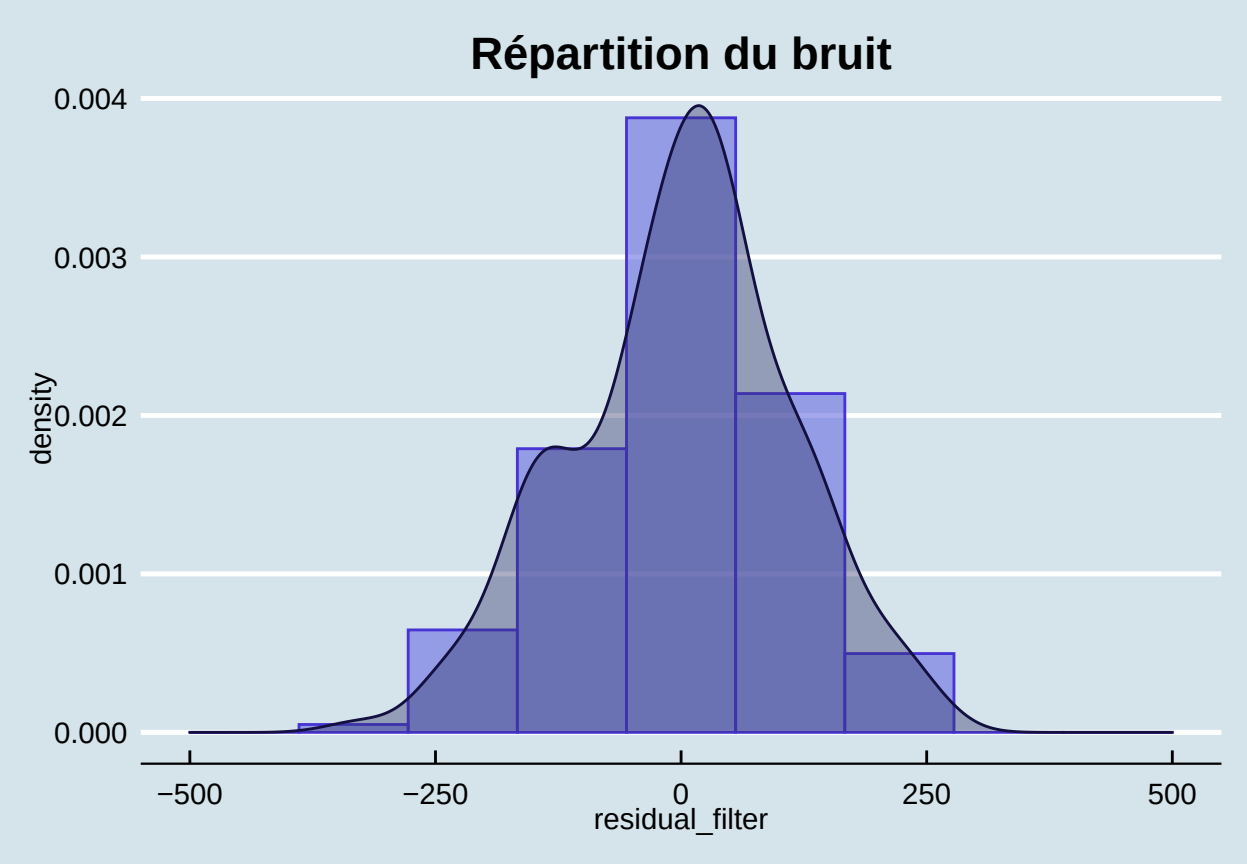
```
residual_filter <- uk_ts - trend_filter - seasonal_filter
```



On ne rejette pas l'hypothèse de la normalité de nos résidus

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

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



Donc $\hat{\epsilon}$ suit une loi normale, et on obtient:

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

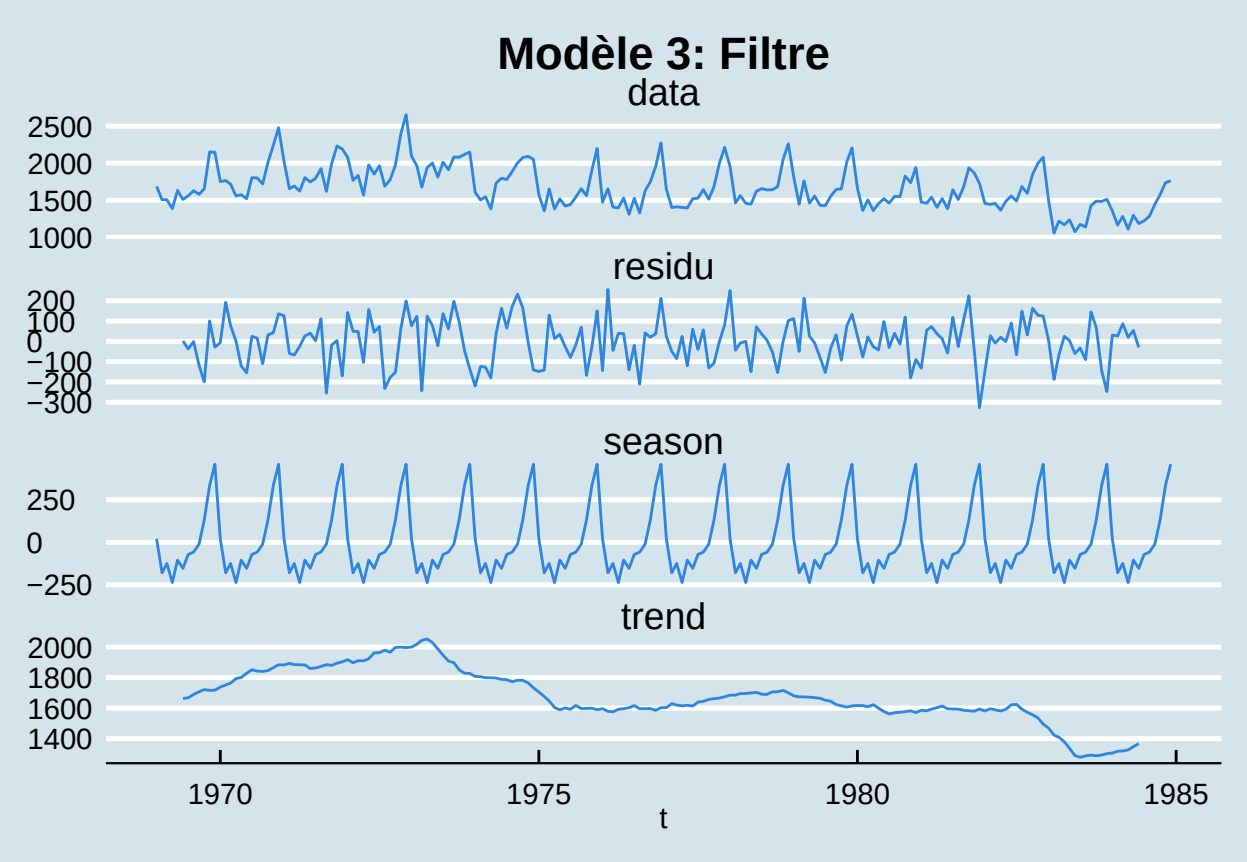
4. Décomposition

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

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

$$\text{avec} \begin{cases} \epsilon_t \sim \mathcal{N}(0, 12\,446) \end{cases}$$

Ce modèle détecte dans sa tendance le changement avec l'entrée en vigueur de la loi sur le port de ceinture.



Modèle 4: 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.