M1 BIBS Orsay
Analyse spectrale et Séries Chro.
2010-2011

Devoir, correction

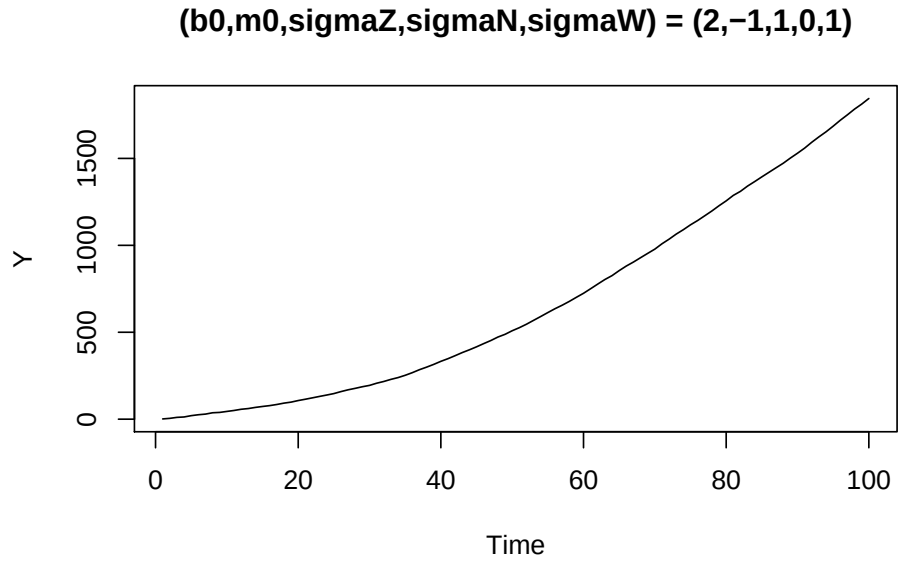
Une liste d'erreurs fréquentes est détaillée à la fin.

Exercice 1

1. On pose $\overrightarrow{X_t} = (B_t, m_t)^T$ pour $t \geqslant 0$. Alors, on observe que pour tout $t \geqslant 0$:
$$\overrightarrow{X_{t+1}} = F\overrightarrow{X_t} + \overrightarrow{V_t} \ ,$$
et :
$$Y_t = G\overrightarrow{X_t} + W_t \ .$$
Avec :
$$F = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad G = \begin{pmatrix} 0 & 1 \end{pmatrix} \ ,$$
et où :
$$\overrightarrow{V_t} = \begin{pmatrix} Z_t \\ N_t \end{pmatrix} \ .$$
est un bruit blanc de matrice de covariance $Q = \begin{pmatrix} \sigma_Z^2 & 0 \\ 0 & \sigma_N^2 \end{pmatrix}$ puisque Z et N sont indépendants. Par ailleurs, W est bien un bruit blanc, de matrice de covariance, $R = \sigma_W^2$. Pour en conclure qu'on a bien un modèle d'état, il faut vérifier que $(\overrightarrow{X_0}, (\overrightarrow{V_t})_{t\geqslant 0}, (\overrightarrow{W_t})_{t\geqslant 0})$ est une collection de variables décorrélées. Or cela découle du fait que $\overrightarrow{X_0} = (B_0, m_0)^T$ est déterministe et que Z, N et W sont indépendants.
2. On utilise le modèle d'état pour faire la simulation. Remarquons que m peut être interprétée comme la tendance de Y, puisque Y est égal à m plus un bruit blanc. Par ailleurs, lorsque $\sigma_N = 0$, $m_{t+1} - m_t = B_t$ donc on peut interpréter B comme la pente de la tendance m. Cette pente suit donc une marche au hasard, b_0 désigne la pente à l'origine (qui doit diriger la direction de m au début) et m_0 la valeur de m à l'origine (qui doit peu influencer l'allure de Y). Simulons Y avec $\sigma_N = 0$ tout d'abord :

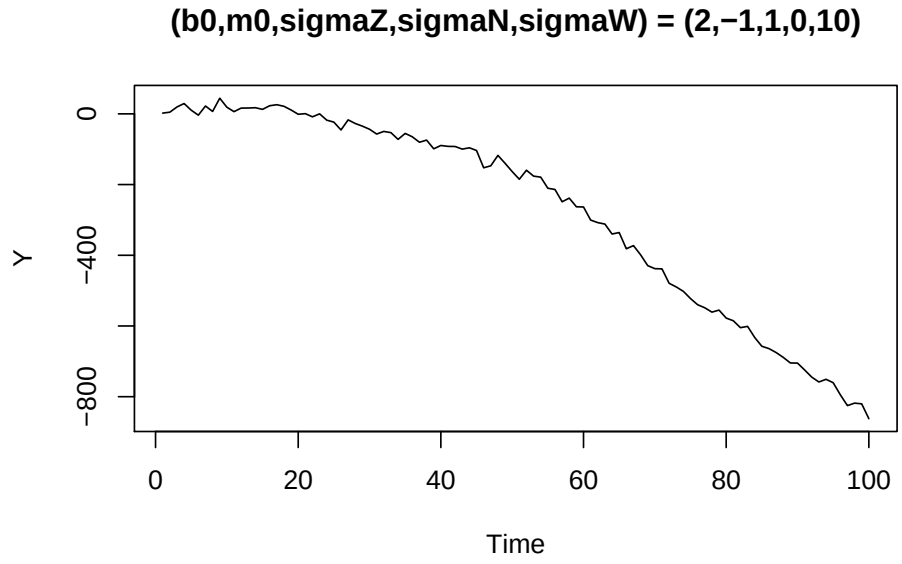
```
> n <- 100
> b0 <- 2
> m0 <- -1
> F <- matrix(c(1, 0, 1, 1), ncol = 2, byrow = TRUE)
> G <- matrix(c(0, 1), ncol = 2)
> sigmaZ <- 1
> sigmaN <- 0
> sigmaW <- 1
> etatexo1 <- function(b0, m0, n, sigmaZ, sigmaN, sigmaW) {
+     X <- matrix(c(b0, m0), ncol = 1)
+     Y <- rep(0, n)
+     for (t in 1:n) {
+         X <- F %*% X + matrix(c(rnorm(1, sd = sigmaZ), rnorm(1,
+             sd = sigmaN)), ncol = 1)
+         Y[t] <- G %*% X + rnorm(1, sd = sigmaW)
+     }
+     Y <- ts(Y)
+ }
```

```
> Y <- etatexo1(b0, m0, n, sigmaZ, sigmaN, sigmaW)
> plot(Y, main = paste("(b0,m0,sigmaZ,sigmaN,sigmaW) = (", paste(b0,
+     m0, sigmaZ, sigmaN, sigmaW, sep = ","), ")", sep = ""))
```



Le bruit d'observation étant faible ($\sigma_W = 1$, à comparer aux valeurs prises par Y), on observe quelque chose de proche de m, qui est assez lisse ici. Augmentons le bruit d'observation :

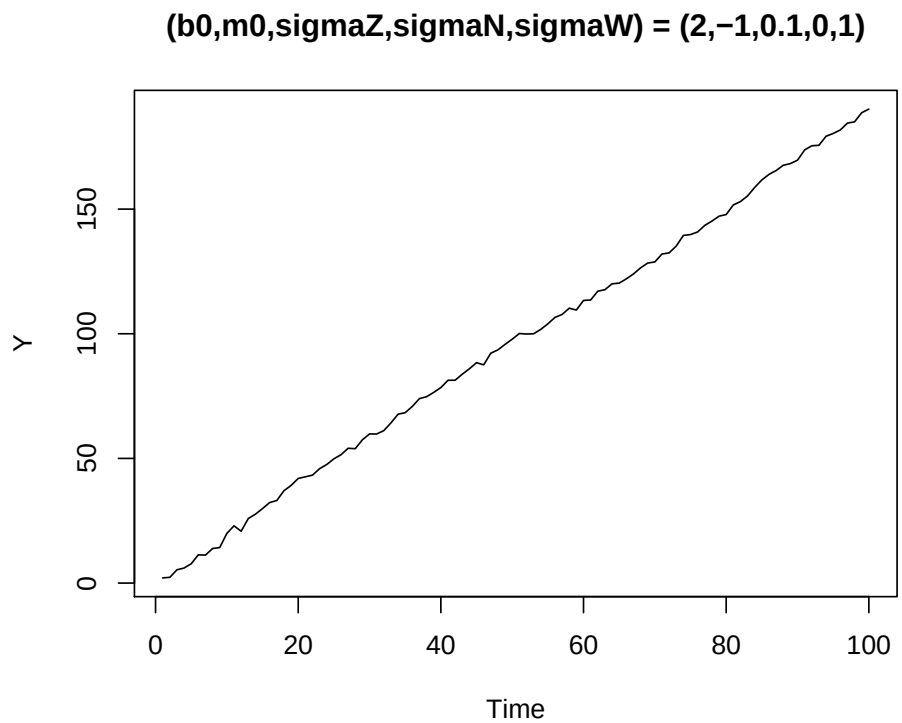
```
> sigmaW <- 10
> Y <- etatexo1(b0, m0, n, sigmaZ, sigmaN, sigmaW)
> plot(Y, main = paste("(b0,m0,sigmaZ,sigmaN,sigmaW) = (", paste(b0,
+     m0, sigmaZ, sigmaN, sigmaW, sep = ","), ")", sep = ""))
```



Cel ane change pas le type de tendance, mais ajoute un bruit additif. Diminuons au contraire le bruit de la marche au hasard.

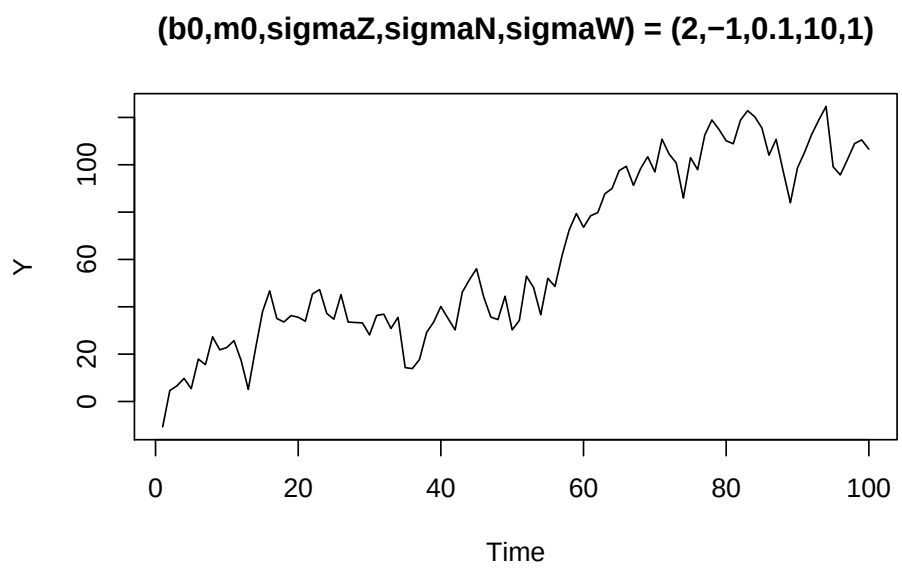
```
> sigmaZ <- 0.1
```

```
> sigmaW <- 1
> Y <- etatexo1(b0, m0, n, sigmaZ, sigmaN, sigmaW)
> plot(Y, main = paste("(b0,m0,sigmaZ,sigmaN,sigmaW) = (", paste(b0,
+     m0, sigmaZ, sigmaN, sigmaW, sep = ","), ")", sep = ""))
```



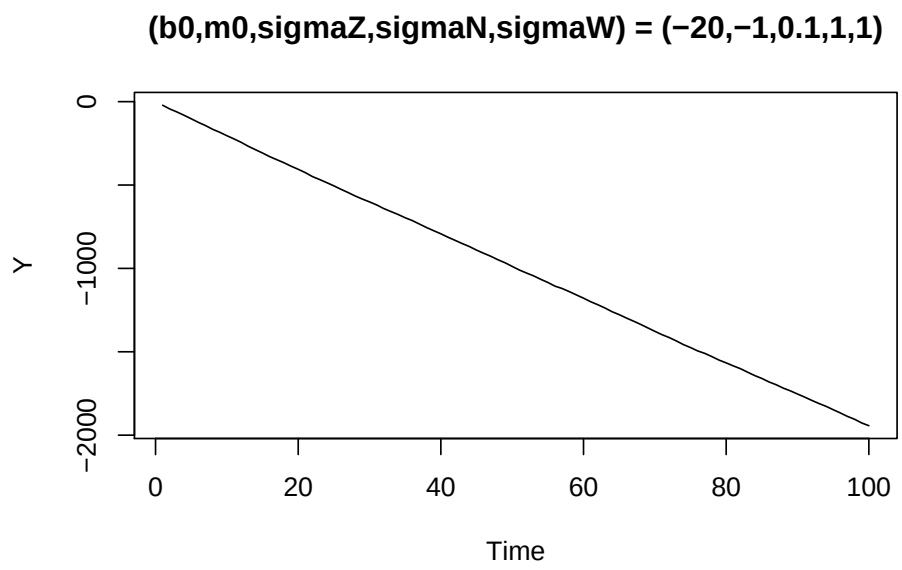
La tendance devient vraiment presque linéaire, on gardera la valeur de σ_Z dans la suite. . Maintenant, augmentons le bruit N. Ceci doit perturber la tendance m_t et donc casser le caractère lisse de la tendance. Exemple avec $\sigma_N = 10$:

```
> sigmaN <- 10
> sigmaW <- 1
> Y <- etatexo1(b0, m0, n, sigmaZ, sigmaN, sigmaW)
> plot(Y, main = paste("(b0,m0,sigmaZ,sigmaN,sigmaW) = (", paste(b0,
+     m0, sigmaZ, sigmaN, sigmaW, sep = ","), ")", sep = ""))
```



Enfin, examinons l'influence de b_0 sur un exemple.

```
> b0 <- -20
> sigmaN <- 1
> sigmaW <- 1
> Y <- etatexo1(b0, m0, n, sigmaZ, sigmaN, sigmaW)
> plot(Y, main = paste("(b0,m0,sigmaZ,sigmaN,sigmaW) = (", paste(b0,
+     m0, sigmaZ, sigmaN, sigmaW, sep = ","), ")", sep = ""))
```



Il devient difficile à la marche aléatoire B de changer de signe avant l'instant 100 lorsque $|b_0/\sigma_Z|$ est loin de 0.

3. Il suffit de voir que :

$$\nabla Y_t = \nabla m_t = B_{t-1} \quad \text{et } \nabla B_{t-1} = Z_{t-2} \ ,$$

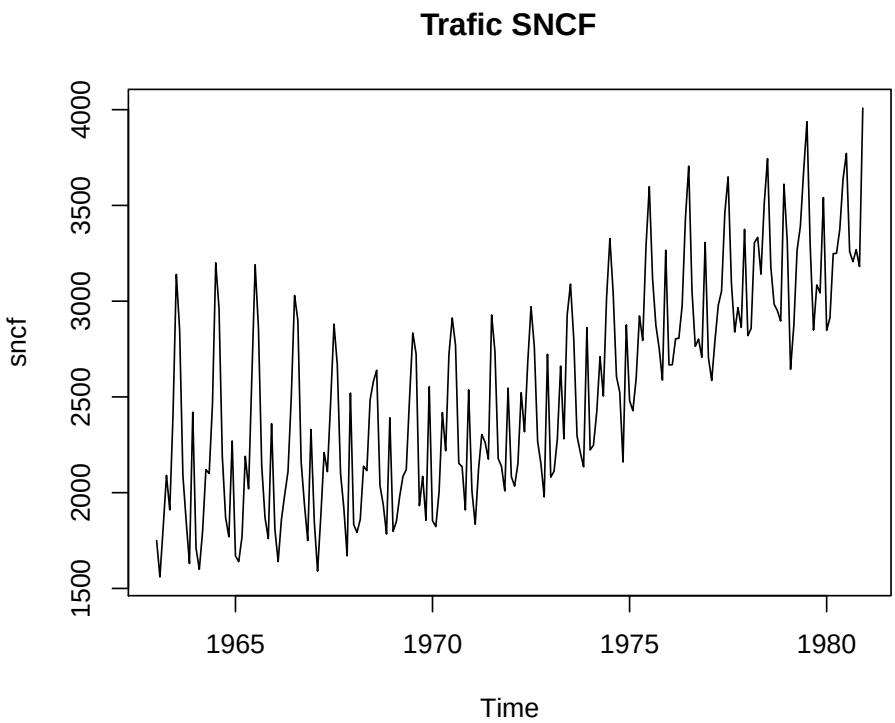
Donc :

$$\nabla^2 Y_t = Z_{t-2} \ .$$

Or $(Z_{t-2})_{t\in\mathbb{Z}}$ est un bruit blanc, ainsi Y est un $ARIMA(0,2,0)$.

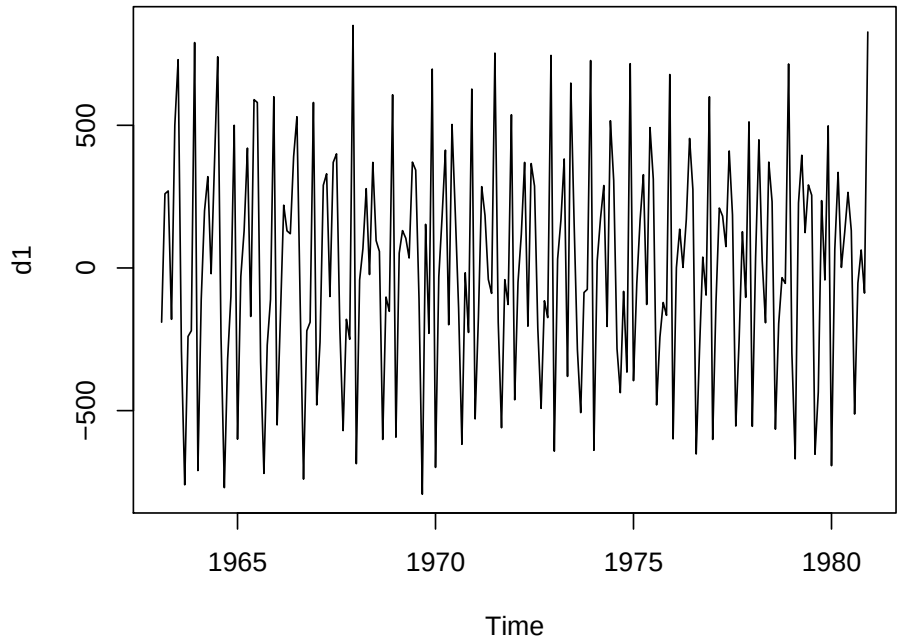
Exercice 2 Commençons par charger et tracer la série.

```
> load("sncf.rda")
> plot(sncf, main = "Trafic SNCF")
```

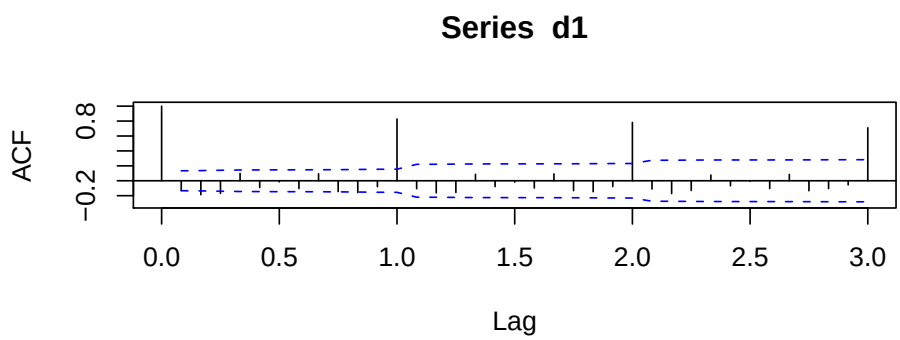


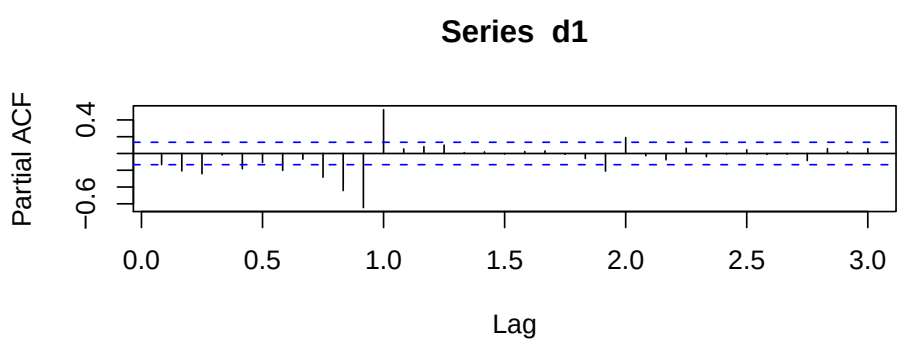
Il y a clairement (et très logiquement) un phénomène saisonnier de période 12, mais également une tendance. Notons X cette série chronologique, et examinons $(1-B^{12})X$, $(1-B)X$ et $(1-B)(1-B^{12})X$.

```
> d1 <- diff(sncf)
> d12 <- diff(sncf, lag = 12)
> d12.1 <- diff(d12)
> plot(d1)
```

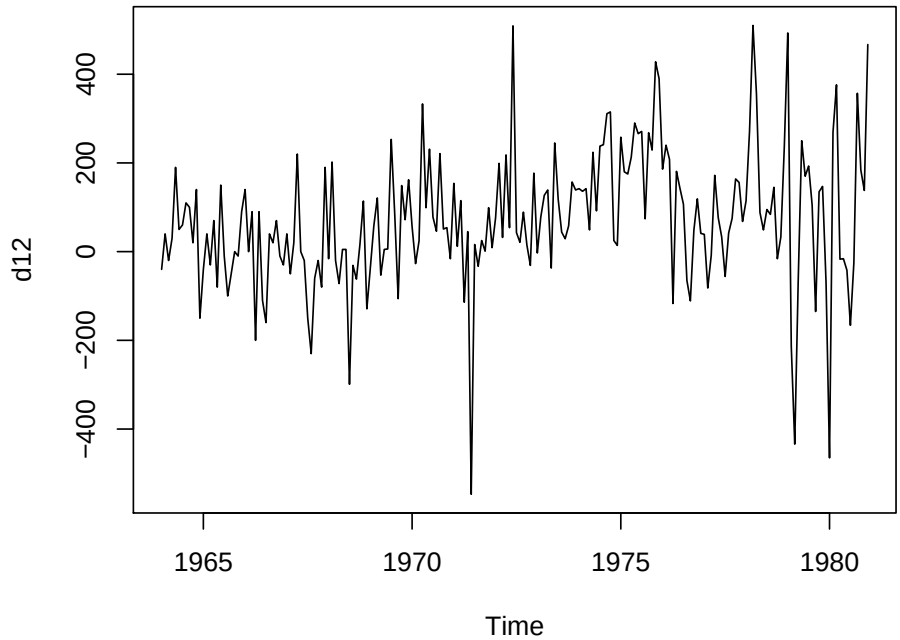



```
> par(mfrow = c(2, 1))
> acf(d1, ci.type = "ma", lag = 36)
> pacf(d1, lag = 36)
```

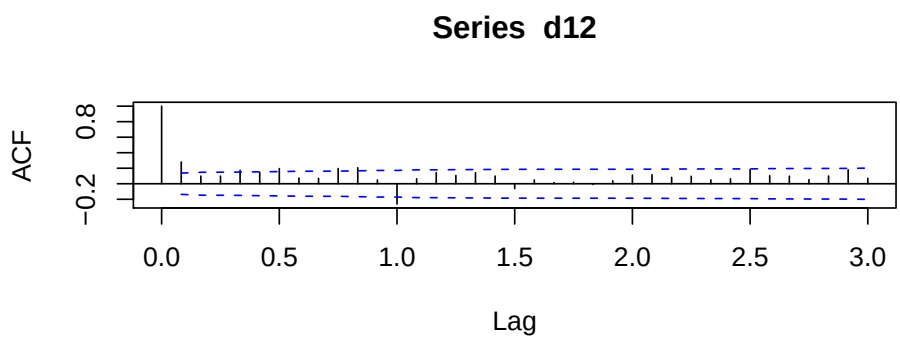


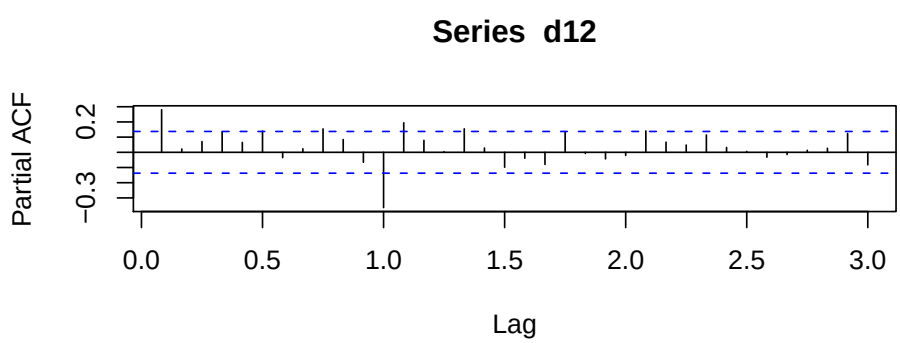



```
> plot(d12)
```

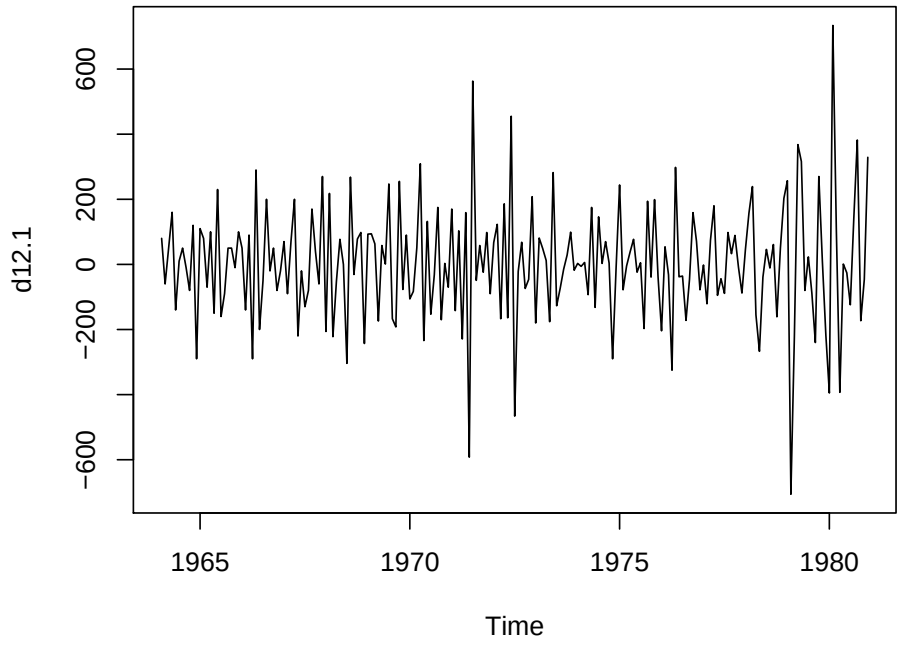



```
> par(mfrow = c(2, 1))
> acf(d12, ci.type = "ma", lag = 36)
> pacf(d12, lag = 36)
```

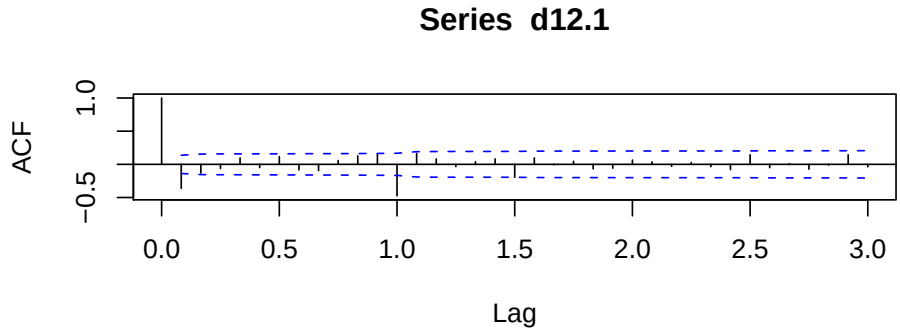


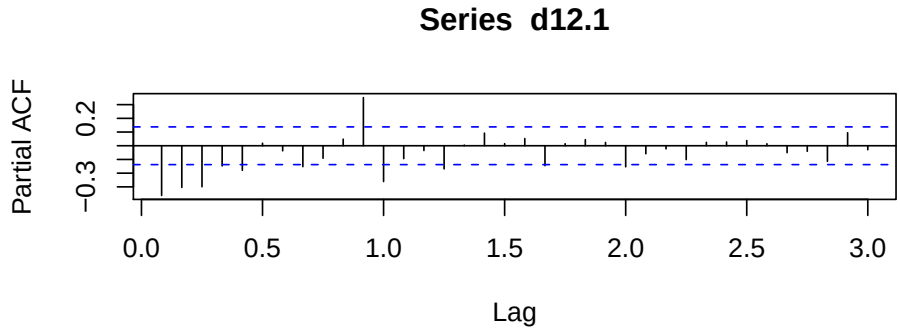



```
> plot(d12.1)
```




```
> par(mfrow = c(2, 1))
> acf(d12.1, ci.type = "ma", lag = 36)
> pacf(d12.1, lag = 36)
```

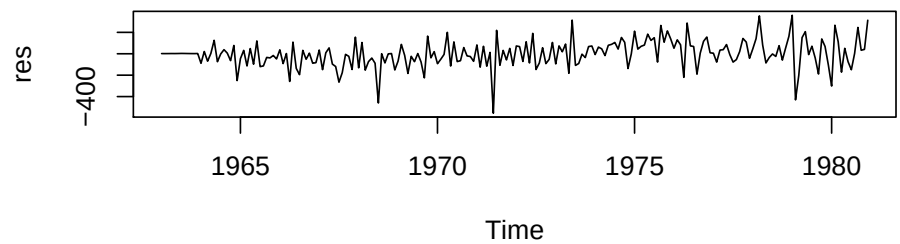


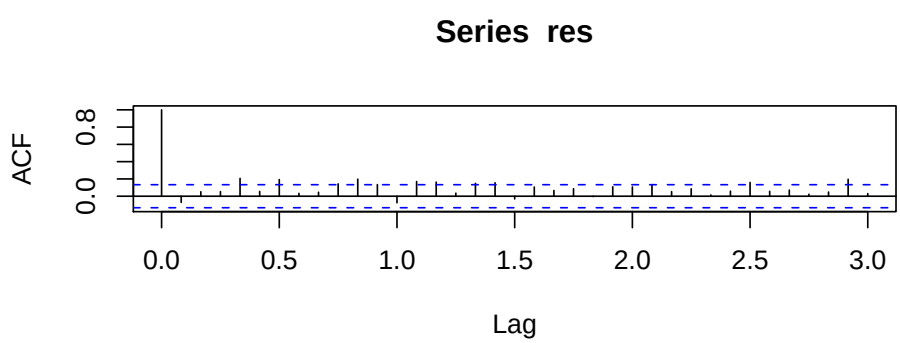


Il semble qu'il faille différentier au lag 1 , mais il n'est pas clair que la différentiation supplémentaire au lag 12 soit nécessaire. On va donc essayer plusieurs modèles.

– L'étude (examen de l'ACF et de la PACF) de la série $(1-B^{12})X$ suggère un $SARIMA(1,0,0)\times(1,1,0)_{12}$.

```
> n <- length(sncf)
> out <- arima(sncf, order = c(1, 0, 0), seasonal = list(order = c(0,
+     1, 1), period = 12), xreg = (1:n))
> res <- residuals(out)
> fitdf <- 2
> par(mfrow = c(2, 1))
> plot(res)
> acf(res, lag.max = 36)
> Box.test(res, lag = 36, type = "Ljung-Box", fitdf = fitdf)
        Box-Ljung test

data:  res
X-squared = 104.0567, df = 34, p-value = 4.992e-09
```





On rejette ce modèle au vu des trois éléments du diagnostic : les résidus sont laids (tendance non nulle), corrélation non nulle au lag 12 et test de Ljung-Box violemment rejeté.

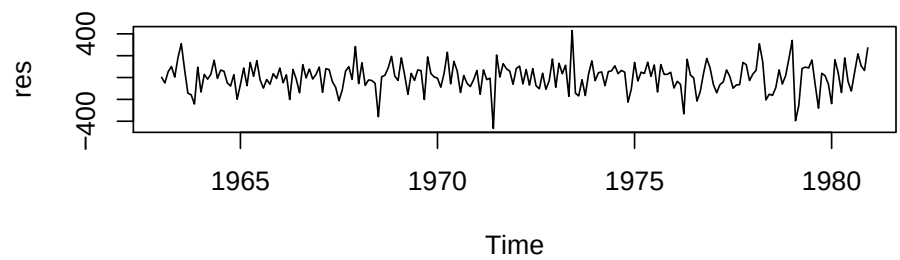
– L'étude de la série $(1-B)X$ suggère un $SARIMA(3,1,0)\times(2,0,0)_{12}$.

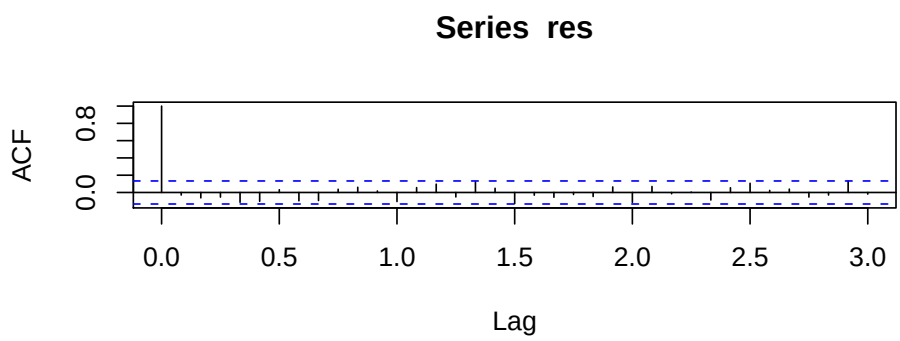
```
> out <- arima(sncf, order = c(3, 1, 0), seasonal = list(order = c(2,
+     0, 0), period = 12), xreg = (1:n))
> out
Call:
arima(x = sncf, order = c(3, 1, 0), seasonal = list(order = c(2, 0, 0), period = 12),
    xreg = (1:n))

Coefficients:
          ar1      ar2      ar3    sar1    sar2    (1:n)
```

```
      -0.5447  -0.3915  -0.3134  0.4238  0.5441  19.2518
s.e.   0.0658   0.0702   0.0648  0.0625  0.0638  48.8206

sigma^2 estimated as 16503:  log likelihood = -1365.54,  aic = 2745.08
```

Il semble qu'on puisse enlever le régresseur :

```
> out <- arima(sncf, order = c(3, 1, 0), seasonal = list(order = c(2,
+     0, 0), period = 12))
> res <- residuals(out)
> fitdf <- 5
> par(mfrow = c(2, 1))
> plot(res)
> acf(res, lag.max = 36)
> Box.test(res, lag = 36, type = "Ljung-Box", fitdf = fitdf)
        Box-Ljung test

data:  res
X-squared = 45.2885, df = 31, p-value = 0.04698
```





L'ACF semble correcte, mais on rejette ce modèle au vu du test de Ljung-Box. On pourra éventuellement y revenir pour essayer de l'améliorer si le troisième modèle ne marche pas.

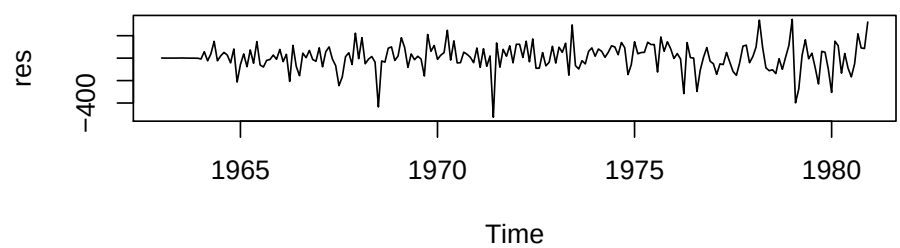
– L'étude de la série $(1-B)(1-B^{12})X$ suggère un $SARIMA(0,1,1)\times(0,1,1)_{12}$ ou $SARIMA(0,1,2)\times(0,1,1)_{12}$.

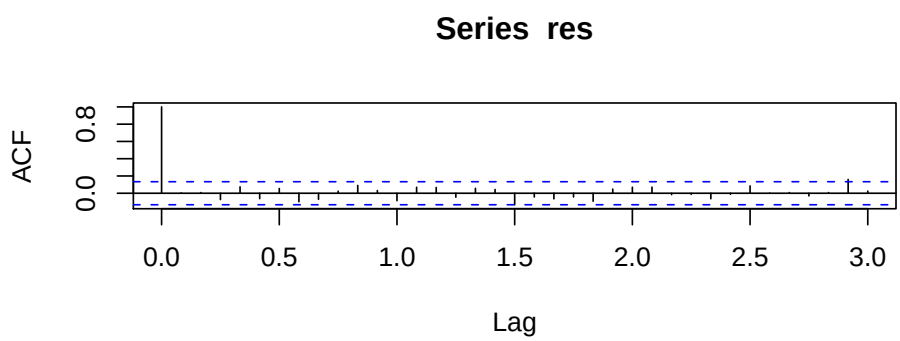
```
> out <- arima(sncf, order = c(0, 1, 2), seasonal = list(order = c(0,
+     1, 1), period = 12), xreg = (1:n)^2)
> out
Call:
arima(x = sncf, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 1), period = 12),
    xreg = (1:n)^2)
```

```
Coefficients:
          ma1      ma2     sma1  (1:n)^2
      -0.6742  -0.1942  -0.5577   0.0243
s.e.   0.0673   0.0660   0.0576   0.0240

sigma^2 estimated as 15363:  log likelihood = -1269.5,  aic = 2549
```

Il semble qu'on puisse enlever le régresseur (0 appartient à l'intervalle de confiance gaussien de niveau 0.95 centré autour du coefficient estimé).

```
> out <- arima(sncf, order = c(0, 1, 2), seasonal = list(order = c(0,
+     1, 1), period = 12))
> res <- residuals(out)
> fitdf <- 3
> par(mfrow = c(2, 1))
> plot(res)
> acf(res, lag.max = 36)
> Box.test(res, lag = 36, type = "Ljung-Box", fitdf = fitdf)

        Box-Ljung test

data:  res
X-squared = 36.5136, df = 33, p-value = 0.3087
```





Les résidus sont centrés, mais pas très jolis (on pourrait même se demander s'ils sont stationnaires). Néanmoins, l'ACF et le test de Ljung-Box sont bons, donc on conserve ce modèle. On a donc ajusté un $SARIMA(0,1,2)\times(0,1,1)_{12}$. On peut essayer de simplifier ce modèle en diminuant les valeurs de p et q.

```
> armaic <- function(x, M = 0, ...) {
+     AIC <- matrix(NA, M + 1, M + 1)
```

```
+     colnames(AIC) <- seq(0, M)
+     rownames(AIC) <- seq(0, M)
+     aic <- Inf
+     popt <- 0
+     qopt <- 0
+     for (p in 0:M) {
+         for (q in 0:(M - p)) {
+             outtemp <- arima(x, order = c(p, 0, q), optim.control = list(maxit = 600),
+                 ...)
+             if (aic > outtemp$aic) {
+                 out <- outtemp
+                 aic <- outtemp$aic
+                 popt <- p
+                 qopt <- q
+             }
+             AIC[p + 1, q + 1] <- outtemp$aic
+         }
+     }
+     res <- list(model = out, AIC = AIC, popt = popt, qopt = qopt)
+ }
> out.armaic <- armaic(d12.1, M = 2, seasonal = list(order = c(0,
+     0, 1), period = 12), include.mean = FALSE)
> out.armaic
```

```
$model

Call:
arima(x = x, order = c(p, 0, q), seasonal = ..1, include.mean = FALSE, optim.control = lis

Coefficients:
          ma1      ma2     sma1
      -0.6698  -0.1884  -0.5516
s.e.   0.0679   0.0654   0.0571

sigma^2 estimated as 15452:  log likelihood = -1269.98,  aic = 2547.95

$AIC
         0        1        2
0 2617.004 2553.469 2547.953
1 2592.637 2547.962       NA
2 2581.048       NA       NA

$popt
[1] 0

$qopt
[1] 2
```

On conserve donc le $SARIMA(0,1,2) \times (0,1,1)_{12}$, dont on rappelle les coefficients :

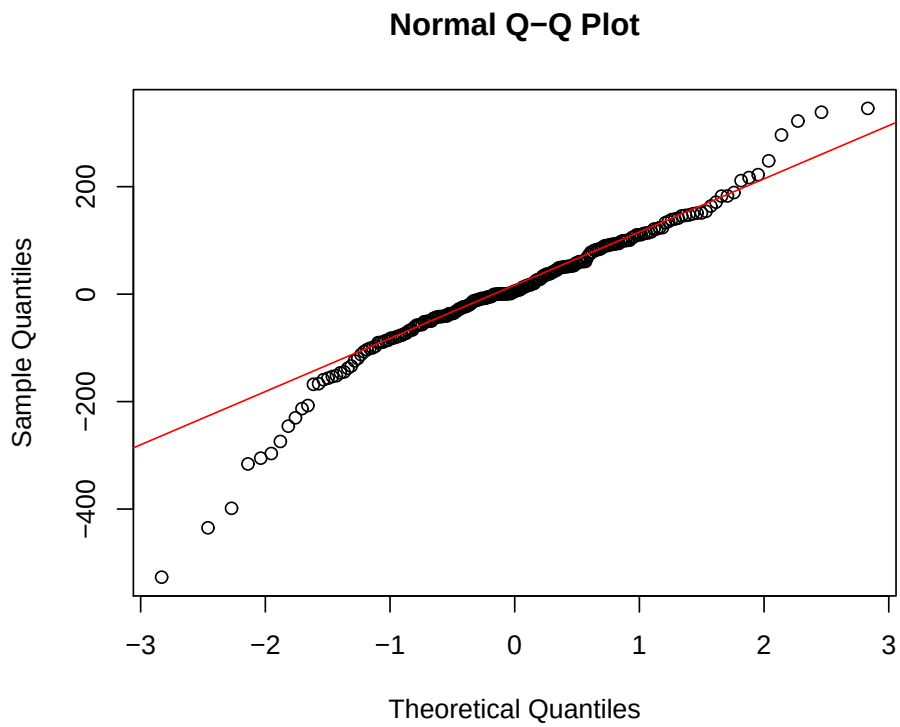
```
> out
```

```
Call:
arima(x = sncf, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 1), period = 12))
```

```
Coefficients:
          ma1      ma2     sma1
      -0.6698  -0.1884  -0.5516
s.e.   0.0679   0.0654   0.0571

sigma^2 estimated as 15452:  log likelihood = -1269.98,  aic = 2547.95
```

Pour finir, on prédit la dernière année à l'aide de la fonction `predict()`. Pour donner des bornes de confiance, on peut se fonder sur une hypothèse de gaussianité des résidus ... si elle est justifiée. Vérifions cela.

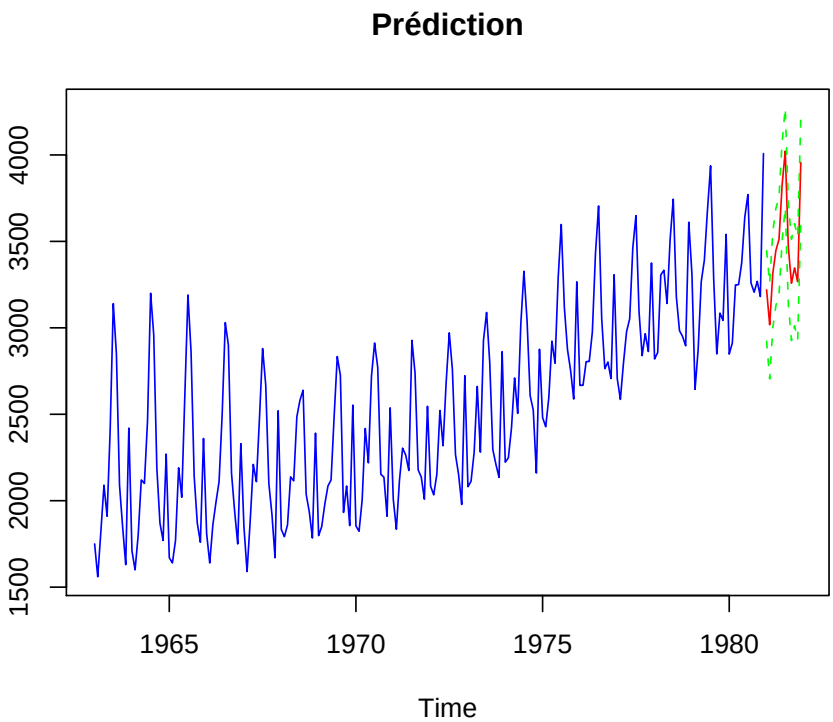
```
> qqnorm(res)
> qqline(res, col = "red")
```



Cela laisse franchement à désirer. On préfèrera donc estimer les quantiles d'ordre 0.025 et 0.975 par les quantiles empiriques. Pour cela, il faudrait effectuer une estimation pour chacun des 12 types d'erreurs de prédiction (i.e prédiction à 1 pas, prédiction à deux pas, etc.). Voici une façon rapide d'avoir un résultat grossier :

```
> q1 <- quantile(res/sd(res), 0.025)
> q1

     2.5%
-2.390264

> q2 <- quantile(res/sd(res), 0.975)
> q2

   97.5%
1.827022
```

```
> outpred <- predict(out, n.ahead = 12)
> pred <- outpred$pred
> predlow <- pred + q1 * outpred$se
> predup <- pred + q2 * outpred$se
> ts.plot(sncf, pred, predlow, predup, lty = c(1, 1, 2, 2), col = c("blue",
+     "red", "green", "green"), main = "Prédiction")
```



On note au passage que les quantiles sont assez différents des quantiles gaussiens.

Exercice 3

1. X est un $AR(1)$ causal car $|3/4| < 1$, donc sa fonction d'autocovariance γ_X vaut (cf. le cours) :

$$\forall h \in \mathbb{Z},\ \gamma_X(h) = \frac{\sigma^2}{1-\phi^2}\phi^{|h|}\ .$$

Par ailleurs, il est facile de voir que :

$$\forall t \geqslant 1, \quad Y_t = \sum_{i=1}^{t} X_i\ .$$

Donc :

$$\overrightarrow{Y} = A\overrightarrow{X}\ ,$$

où :

$$\overrightarrow{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \text{et } \overrightarrow{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}\ .$$

Soit Γ_Y la matrice de covariance de $\overrightarrow{Y}$ et Γ_X celle de $\overrightarrow{X}$. On a donc :

$$\Gamma_Y = A\Gamma_X A^T \ .$$

On effectue les calculs avec R :

```
> A <- matrix(0, 4, 4)
> A[lower.tri(A, diag = TRUE)] <- 1
> phi <- 3/4
> sigma2 <- 4
> gammaX <- toeplitz(phi^(0:3)) * sigma2/(1 - phi^2)
> gammaY <- A %*% gammaX %*% t(A)
> gammaY
```

```
          [,1] [,2]     [,3] [,4]
[1,]  9.142857   16 21.14286   25
[2,] 16.000000   32 44.00000   53
[3,] 21.142857   44 65.14286   81
[4,] 25.000000   53 81.00000  106
```

2. On applique la méthode du cours. En notant que Y est centré, on a :

$$P(Y_4|Y_1, \ldots, Y_3) = \alpha_1 Y_1 + \alpha_2 Y_2 + \alpha_3 Y_3 \ ,$$

où $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ est solution de :

$$\Gamma\alpha = \gamma \ ,$$

avec Γ la matrice de covariance de $(Y_1, \ldots, Y_3)$ et γ le vecteur :

$$\gamma = \begin{pmatrix} \mathbb{E}(Y_1 Y_4) \\ \mathbb{E}(Y_2 Y_4) \\ \mathbb{E}(Y_3 Y_4) \end{pmatrix} \ .$$

La variance de l'erreur de prédiction est alors :

$$v = \mathsf{Var}(Y_4) - \alpha^T \gamma \ .$$

On effectue les calculs avec R :

```
> gamma <- gammaY[1:3, 4]
> Gamma <- gammaY[1:3, 1:3]
> alpha <- solve(Gamma, gamma)
> v <- gammaY[4, 4] - t(alpha) %*% gamma
> alpha
```

```
[1]  2.256313e-15 -7.500000e-01  1.750000e+00
```

```
> v
```

```
     [,1]
[1,]    4
```

Une façon plus élégante de résoudre le problème est la suivante. L'espace vectoriel engendré par (X_1, X_2, X_3) est égal à celui engendré par (Y_1, Y_2, Y_3). En effet, Y_1, Y_2 et Y_3 sont comniaisons linéaires de X_1, X_2 et X_3 et réciproquement. Donc :

$$\begin{aligned}
P(Y_4|Y_1,\dots,Y_3) &= P(Y_4|X_1,\dots,X_3) \\
&= P(X_1+X_2+X_3+X_4|X_1,\dots,X_3) \\
&= X_1+X_2+X_3+P(X_4|X_1,\dots,X_3) \\
&= X_1+X_2+X_3+\phi X_3 \\
&= Y_1+(Y_2-Y_1)+(Y_3-Y_2)+\phi(Y_3-Y_2) \\
&= -\phi Y_2+Y_3(1+\phi)\ .
\end{aligned}$$

Puis, pour la variance de l'erreur de prédiction :

$$\begin{aligned}
v &= \mathbb{E}[(Y_4-P(Y_4|Y_1,\dots,Y_3))^2] \\
&= \mathbb{E}[(X_1+X_2+X_3+X_4-(\phi X_3+X_1+X_2+X_3))^2] \\
&= \mathbb{E}[(X_4-\phi X_3)^2] \\
&= \mathbb{E}[Z_4^2] \\
&= 4\ .
\end{aligned}$$

Erreurs fréquemment rencontrées

D'une manière générale :

1. Il faut rendre les graphiques importants.
2. Il faut commenter (sans faire de blabla) vos résultats.
3. Ces commentaires doivent être, de préférence, écrit en dehors du code, sur une feuille. L'avantage est que cela vous oblige à en faire une présentation synthétique, et à ne parler que de ce qui est important. Si la rédaction est bonne, je n'aurai même pas besoin d'aller voir le code.

Par ailleurs :

Exercice 1 Question 1. Le modèle d'état doit être le plus simple possible. Notamment, la dimension d'état doit être la plus petite possible (du moment que ça reste facile à faire). Ici, une dimension 2 suffit. Par ailleurs, il faut, pour suivre la définition du cours, que $(X_0, (V_t)_{t\geqslant 0}, (W_t)_{t\geqslant 0})$ forme une collection de variables aléatoires deux à deux décorrélées.

Question 3. Tout d'abord, c'étati une question théorique (il ne s'agissait pas d'ajuster un modèle $ARIMA$). Par ailleurs, sous prétexte que :

$$B_t = b_0 + \sum_{i=0}^{t-1} Z_t\ ,$$

il ne faut pas dire que "B_t est un $ARMA(0,t)$"!!! Tout d'abord, le processus B n'est pas stationnaire. Ensuite, on dit d'un processus que c'est un $ARMA$, pas d'une unique variable aléatoire (essayez donc de dire que B est un $ARMA(0,t)$ ...). Enfin, les paramètres d'un $ARMA$ sont des constantes (p,q), et ne peuvent varier en fonction du temps.

Exercice 2

1. Il faut commenter la série après l'avoir tracée : tendance ? phénomène saisonnier ? Ici, il faut parler de phénomène saisonnier plutôt que de saisonnalité (le motif n'étant pas assez semblable d'une année sur l'autre).
2. Revoir les règles de choix de (p, q) et (P, Q). Dans le cas présent (après avoir différentié une fois au lag 1 et une fois au lag 12), l'acf et la pacf nous indiquent $p = 5$, $q = 1$ ou 2, $P = 2$ et $Q = 1$. Comme $p > q$, cela nous incite à ajuster un $ARMA(0, 1)$ pour la partie non saisonnière, et comme $P > 1$, cela nous incite à ajuster un $ARMA(0, 1)$ pour la partie saisonnière. Soit, sur la série différentiée, un $SARIMA(0, 0, 1) \times (0, 0, 1)_{12}$. Beaucoup ont proposé un $SARIMA(5, 0, 1) \times (2, 0, 1)_{12}$, ce qui fait beaucoup de paramètres inutiles.
3. Personne n'a vérifié si on pouvait supposer nulle la moyenne de la série différentiée. Et pourtant, vous n'avez pas ajouté l'option `include.mean=FALSE` dans l'estimation du $SARIMA$ sur la série différentiée. Si jamais cette moyenne ne peut pas être supposée nulle, il faudra ajuster sur la série originale un $SARIMA$ avec `(1:n)^2` comme régresseur.

Exercice 3 Ne pas oublier de mentionner la causalité de X pour utiliser la formule de la fonction d'autocovariance d'un $AR(1)$ causal.