Projet de Séries Temporelles :

modélisation ARIMA d'une série temporelle

Victor DUC

14 mai 2023

Contents

1 Les données

Considérons l' indice CVS-CJO de la production industrielle (base 100 en 2015) - Fabrication de produits en caoutchouc et en plastique (NAF rév. 2, niveau division, poste 22) (Identifiant 010537483). Cet IPI, calculé par l'Insee à partir des enquêtes mensuelles de branche réalisées auprès d'un échantillon d'entreprises, permet de suivre l'évolution mensuelle de fabrication de produits en caoutchouc et en plastique en France métropolitaine. Il admet pour année de référence 2015 ce qui signifie qu'il a pour moyenne 100 en 2015. Cette série est corrigée des variations saisonnières (CVS) et des effets de calendrier (CJO).

```r
plot(xm, ylim=c(56,124), xaxt="n") #xaxt supprime les labels et graduations (abcisses)
axis(side=1, ylim=c(56,124), at=seq(0,376,12)) #cree nouveaux labels et graduations (abcisses)
```

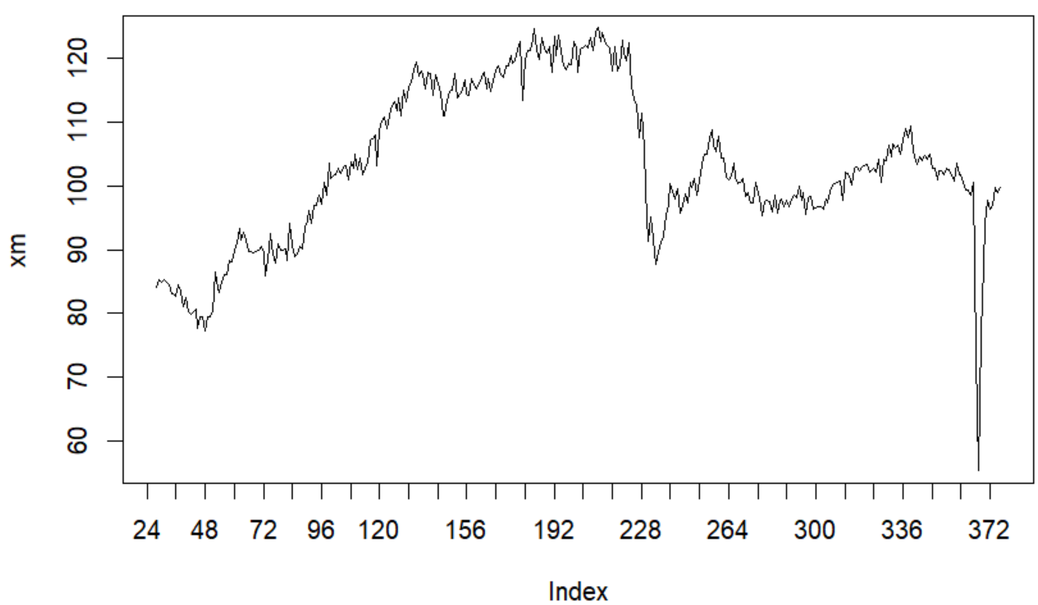

Figure 1: Chronogramme

Le chronogramme nous laisse présumer que la série considérée n'est pas stationnaire. En effet la série semble avoir une tendance non linéaire, voire non déterministe.

```r
acf(xm)
```

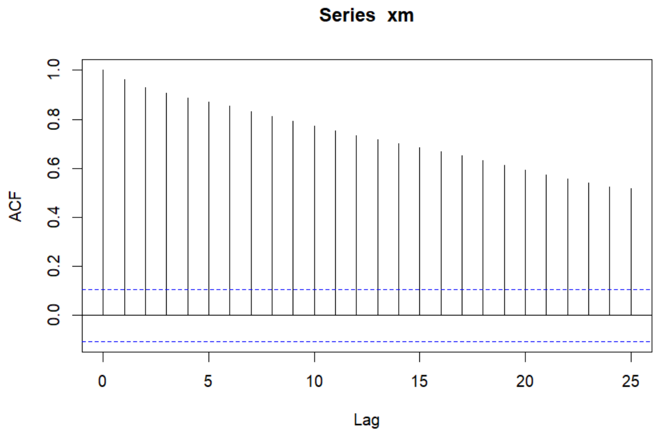

Figure 2: Fonction d'autocorrélation

Les valeurs d'autocorrélation sont relativement grandes *i.e.* proches de 1 donc le corrélogramme semble confirmer notre conjecture de non stationnarité faible. Avant de procéder aux tests de racine unitaire, il convient de vérifier s'il y a une constante et/ou une tendance linéaire non nulle. La représentation graphique de `xm` a montré que la tendance n'est probablement pas linéaire, mais si on devait en choisir une elle serait positive. Régressons `xm` sur ses `dates` pour le vérifier.

```r
dates <- as.yearmon(seq(from=1992, to=2021, by=1/12))
summary(lm(xm ~ dates))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -382.07888  153.44135  -2.490   0.0132 *
dates          0.24183    0.07647   3.162   0.0017 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le coefficient `dates` 0.24183 associé à la tendance linéaire est bien positif, et peut-être significatif (on ne peut pas vraiment le confirmer car le test n'est pas valide en présence de résidus possiblement autocorrélés). Il faudra donc se mettre dans le cas des tests de racine unitaire avec constante et éventuellement tendance non nulles. Le test de Dickey-Fuller augmenté (ADF) dans le cas avec constante et tendance consiste en la régression suivante, pour une variable X donnée

$$\Delta X_t = c + bt + \beta X_{t-1} + \sum_{\ell=\mathbb{1}_{k>0}}^{k} \phi_\ell \Delta X_{t-\ell} + \varepsilon_t$$

où $\beta + 1$ est l'autocorrélation à l'ordre 1 de X et k le nombre de retards nécessaires à considérer pour rendre les résidus non autocorrélés. L'hypothèse nulle de racine unitaire $H_0 : \beta = 0$ est testée par la statistique de test $\hat{\beta}/\hat{\sigma}(\hat{\beta})$ qui suit une loi de Dickey-Fuller dépendant du nombre d'observation et du cas du test dans lequel on se place.

```r
require(fUnitRoots) #tests de racine unitaire plus modulables
adf <- adfTest(xm, lag=0, type="ct") #test ADF dans le cas avec constante et tendance
```

Avant d'interpréter le test, vérifions que les résidus du modèle de régression sont bien non autocorrélés, sans quoi le test ne serait pas valide. Comme la série est mensuelle, testons l'autocorrélation des résidus jusqu'à l'ordre 24 (deux ans), sans oublier de corriger les degrés de liberté du nombre de régresseurs.

```r
Qtests(adf@test$lm$residuals, 24, fitdf = length(adf@test$lm$coefficients))
```

```
     lag        pval
[1,]   1          NA  [7,]   7 0.022876699 [13,]  13 0.275733144 [19,]  19 0.694105717
[2,]   2          NA  [8,]   8 0.044741545 [14,]  14 0.351442840 [20,]  20 0.741925819
[3,]   3          NA  [9,]   9 0.075621288 [15,]  15 0.432599912 [21,]  21 0.690873009
[4,]   4 0.007822992 [10,]  10 0.097378039 [16,]  16 0.507589442 [22,]  22 0.734772005
[5,]   5 0.028198964 [11,]  11 0.145196531 [17,]  17 0.563915826 [23,]  23 0.677581512
[6,]   6 0.012584107 [12,]  12 0.205496370 [18,]  18 0.632214360 [24,]  24 0.390702098
```

L'absence d'autocorrélation des résidus est rejetée au moins une fois (`Q(4)` à `Q(8)`), le test ADF avec aucun retard n'est donc pas valide. Ajoutons des retards de ΔX_t jusqu'à ce que les résidus ne soient plus autocorrélés.

```r
series <- xm; kmax <- 24; adftype="ct"
adf <- adfTest_valid(xm,24,adftype="ct")
```

```
ADF with 0 lags: residuals OK? nope
ADF with 1 lags: residuals OK? nope
ADF with 2 lags: residuals OK? nope
ADF with 3 lags: residuals OK? nope
ADF with 4 lags: residuals OK? nope
ADF with 5 lags: residuals OK? OK
```

Il a fallu considérer $k = 5$ retards au test ADF pour supprimer l'autocorrélation des résidus.

```r
adf #affichage des resultats du test valide maintenu
```

```
Test Results:
  PARAMETER:
    Lag Order: 5
  STATISTIC:
    Dickey-Fuller: -1.88
  P VALUE:
    0.6274
```

La racine unitaire n'est pas rejetée à un seuil de 95% pour la série en niveau, la série est donc au moins I(1). Différencions la série une fois.

```r
dxm <- diff(xm,1)
```

Traçons le chronogramme de la série différenciée `dxm`.

```r
plot(dxm,type="l",xaxt="n")
axis(side=1, ylim=c(56,124), at=seq(0,376,12))
```

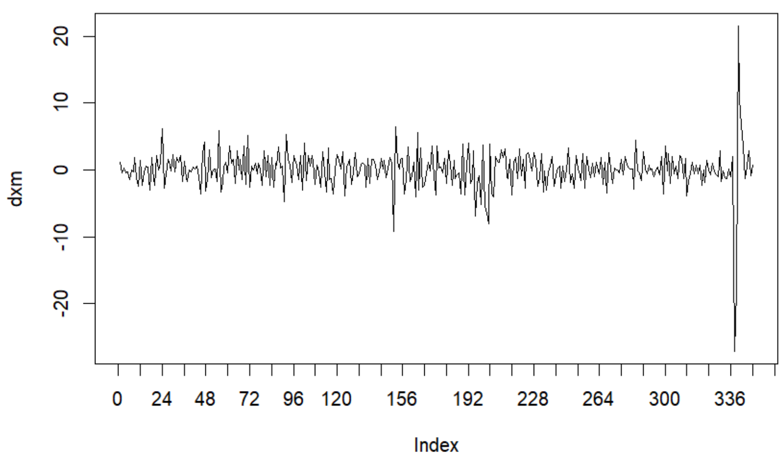

Figure 5: Chronogramme de la série différenciée 1 fois

Testons maintenant la racine unitaire pour la série différenciée `dxm`. La représentation graphique précédente semble montrer l'absence de constante et de tendance non nulle. Vérifions avec une régression.

```r
summary(lm(dxm ~ dates[-1])) #sans la premiere date car on a diffrencie la serie
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.95891   40.06521   0.673    0.501
dates[-1]   -0.01341    0.01997  -0.672    0.502
```

Il y a bien ni constante ni tendance significative. Effectuons donc le test ADF dans le cas sans constante ni tendance, en vérifiant l'absence d'autocorrélation des résidus.

```r
adf <- adfTest_valid(dxm,24,"nc")
```

```
ADF with 0 lags: residuals OK? nope
ADF with 1 lags: residuals OK? OK
```

Il a été nécessaire d'inclure un retard dans le test ADF.

```
adf
```

```
Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    Dickey-Fuller: -15.639
  P VALUE:
    0.01
```

Le test rejette la racine unitaire (`p-value` < 0.05), on dira donc que la série différenciée est "stationnaire" et retiendra un ordre $d^* = 1$. La série `xm` est donc I(1).

2 Modèle ARMA

Suivons la méthodologie de Box-Jenkins. Identifions les degrés du modèle. Étudions les fonctions d'autocorrélation et d'autocorrélation partielle de la série retenue.

```
par(mfrow=c(1,2))
pacf(dxm,24);acf(dxm,24) #on regarde jusqu'a deux ans de retard
```

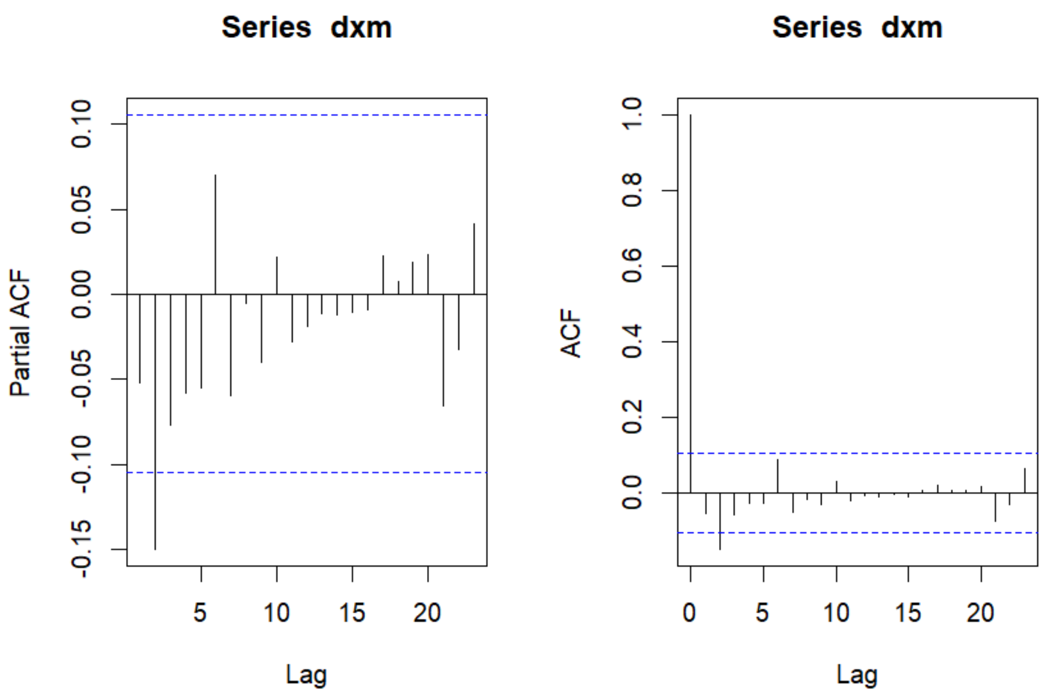

Figure 6: Fonctions d'autocorrélation et d'autocorrélation partielle de la série `dxm`

- La `pacf` est significative jusqu'à l'ordre 2 maximum ($\hat{r}(2) \neq 0$ et, pour tout $r > 2$, $\hat{r}(h) \approx 0$), on choisira donc $p^* = 2$.
- L'`acf` est significative jusqu'à l'ordre 0 au maximum ($\hat{q}(0) \neq 0$ et, pour tout $h > 0$, $\hat{q}(h) \approx 0$), on choisira donc $q^* = 0$.

```
pmax=2;qmax=0
```

Les modèles possibles sont tous les ARIMA(p,1,q) pour `xm` où $p \leqslant 2$ et $q \leqslant 0$: ARIMA(1,1,0) et ARIMA(2,1,0).

```
AICs
BICs
```

```
        q=0
p=1 1780.618
p=2 1774.767
```

```
        q=0
p=1 1788.322
p=2 1786.323
```

L'ARMA (2,0) minimise l'AIC et le BIC. Plaquons tout d'abord un ARMA(2,0).

```r
y <- dxm - mean(dxm) #
arima200 <- arima(y,c(2,0,0))
arima200
```

```
Call:
arima(x = y, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      -0.060  -0.1490    -0.0004
s.e.   0.053   0.0529     0.1363

sigma^2 estimated as 9.437:  log likelihood = -884.38,  aic = 1776.77
```

Le coefficient AR(2) est significatif *i.e.* statistiquement non nul (le rapport entre le coefficient estimé et son erreur standard est plus grand en valeur absolue que 1.96), l'ARIMA(2,0) est donc bien **ajusté**.

```r
Box.test(arima200$residuals, lag=24, type="Ljung-Box", fitdf=2)
```

```
        Box-Ljung test

data:  arima200$residuals
X-squared = 19.491, df = 22, p-value = 0.6148
```

Le test de Ljung-Box délivre une $p-$value > 0.05.

```r
Qtests(arima200$residuals, 24, 2)
```

```
     lag      pval
[1,]   1        NA  [7,]   7 0.1845271 [13,]  13 0.6534988 [19,]  19 0.9404627
[2,]   2        NA  [8,]   8 0.2747298 [14,]  14 0.7317735 [20,]  20 0.9588268
[3,]   3 0.1016988  [9,]   9 0.3251204 [15,]  15 0.7969433 [21,]  21 0.9332325
[4,]   4 0.1922187 [10,]  10 0.4029502 [16,]  16 0.8492606 [22,]  22 0.9278627
[5,]   5 0.2744723 [11,]  11 0.4787114 [17,]  17 0.8821926 [23,]  23 0.9138816
[6,]   6 0.1722162 [12,]  12 0.5736901 [18,]  18 0.9139062 [24,]  24 0.6148475
```

Les résidus des lags jusqu'à 24 sont non corrélés, se comportent comme un bruit blanc. Le modèle est bien **valide**. Vérifions que le modèle n'est pas simplifiable en considérant un ARMA(1,0).

```r
estim <- arima(y,c(1,0,0)); arimafit(estim)
```

```
tests de nullite des coefficients :
        ar1 intercept
coef -0.052     0.000
se    0.053     0.158
pval  0.329     0.999
```

```
  tests d'absence d'autocorrelation des residus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.031  13 0.249  19 0.651
[2,]   2 0.004   8 0.050  14 0.316  20 0.707
[3,]   3 0.007   9 0.073  15 0.386  21 0.632
[4,]   4 0.017  10 0.101  16 0.458  22 0.668
[5,]   5 0.034  11 0.140  17 0.518  23 0.650
[6,]   6 0.022  12 0.191  18 0.587  24 0.369
```

Le modèle ARMA(1,0) n'est pas valide. Finalement

$$Y_t = (I - L)X_t \sim \text{ARMA}(2,0)$$

Vérifions la **causalité**, l'inversibilité du polynôme $\hat{\phi} := 1 - \hat{\phi}_1 X - \hat{\phi}_2 X^2$ *i.e.* que les racines sont de modules strictement supérieur à 1.

```r
Mod(polyroot(sort(arima200$coef[c('intercept','ar1', 'ar2')])))
```

```
[1]   2.531167 131.279155
```

Le processus $(\varepsilon_t)_t$ est le processus des **innovations** de $(X_t)_t$. On peut donc passer à l'ARIMA(2,1,0).

3 Prévision

Le vrai modèle théorique sous-jacent est un AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$$

et nous estimons

$$\hat{X}_t = \hat{\phi}_1 X_{t-1} + \hat{\phi}_2 X_{t-2} + \hat{\varepsilon}_t$$

Les hypothèses sont

H_1 : $(X_{1-d}, \ldots, X_0) = (X_{-1}, X_0)$ et $(Y_t)_{t \geqslant 1}$ sont décorrélés.

H_2 : Les paramètres $\hat{\phi}_1$ et $\hat{\phi}_2$ estiment bien le modèle.

H_3 : Les résidus $(\varepsilon_t)_t$ sont gaussiens (et indépendants).

H_4 : La matrice de variance-covariance est inversible.

Sous ces hypothèses

$$X_t = Y_t - \sum_{i=1}^{2} \binom{2}{i} (-1)^i X_{t-i} \sim \text{ARIMA}(2,1,0) + \mathbb{E}[X]$$

Les erreurs de prédictions e_{T+1} à horizon 1 et e_{T+2} à horizon 2 sont

$$e_{T+1} := X_{T+1} - \hat{X}_{T+1} \qquad\qquad e_{T+2} := X_{T+2} - \hat{X}_{T+2}$$

Par hypothèse H_2, la matrice de variance-covariance vérifie $\hat{\Sigma} = \Sigma$.

$$\Sigma = \begin{pmatrix} \text{Var}(\varepsilon_{T+1}) & \text{Cov}(\varepsilon_{T+1}, \varepsilon_{T+2}) \\ \text{Cov}(\varepsilon_{T+1}, \varepsilon_{T+2}) & \text{Var}(\varepsilon_{T+2}) \end{pmatrix}$$

Posons $e := \begin{pmatrix} e_{T+1} \\ e_{T+2} \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$ et $q_{1-\alpha}^{\chi^2(2)}$ le quantile d'ordre $1-\alpha$ de la loi du χ^2 à 2 degrés de liberté. La région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2}) est l'ellipse

$$\left\{ x \mid e^\top \Sigma e < q_{1-\alpha}^{\chi^2(2)} \right\}$$

Affichons les prévisions pour les deux prochains mois.

```
xmp
```

```
 arima200
100.64594
 73.39979
```

Question ouverte. Si l'on suppose que la série $(Y_t)_t$ cause instantanément la série $(X_t)_t$, alors l'ajout de l'information sur Y_{T+1} peut améliorer la prévision de X_{T+1}.

Pour tester cela, une approche possible serait d'utiliser une méthode de validation croisée. On pourrait diviser les données disponibles en deux ensembles : un ensemble d'apprentissage contenant les observations de X_1 à X_T, et un ensemble de validation contenant les observations de X_{T+1} et Y_{T+1}. On pourrait alors ajuster deux modèles ARIMA, un basé uniquement sur les observations de l'ensemble d'apprentissage pour prévoir X_{T+1}, et un autre qui utilise également l'observation de Y_{T+1}. On pourrait ensuite comparer les erreurs de prévision des deux modèles sur l'ensemble de validation pour déterminer si l'ajout de l'observation de Y_{T+1} a amélioré la prévision de X_{T+1}.

Forecasts from ARIMA(2,1,0)