PROJET DE SERIES TEMPORELLES LINEAIRES

Rois-Céti DIMBAMBA L.C. et Amadou DIOUF (2AD ENSAE Paris)

14/05/2021

Contents

1 Données

1.1 Série étudiée

Dans ce projet, nous nous intéressons à l'indice de la production industrielle (base 100 en 2015) du secteur de l'exploitation de gravières et sablières ainsi que de l'extraction d'argiles et de kaolin. Cet indice est produit mensuellement par l'INSEE à partir des enquêtes mensuelles de branche réalisées par la Direction des statistiques d'entreprises (DES) et le Service de la Statistique et de la Prospective du ministère de l'agriculture. La série x_t ici étudiée couvre la période allant de janvier 1990 à janvier 2021, soit 373 observations. Il s'agit d'une série agrégée corrigée des variations saisonnières et des jours ouvrés (CVS-CJO) et ne présente par conséquent pas de saisonnalités.

L'observation de la figure ci-dessus nous fait soupçonner une présence d'hétéroscédasticité. Nous effectuons une transformation logarithmique dont le but est de lisser les observations pour à la fin se retrouver avec des données plus homoscédastiques. La série obtenue suite à cette transformation est noté $y_t = log(x_t)$. Cette dernière, toute comme la série initiale, semble présenter de tendance constante. Elle est donc centrée pour finalement obtenir la série z_t.

1.2 Stationnarité

Pour l'étude de la stationnarité de la série z_t, Nous utilisons le test de racine unitaire de Dickey-Fuller augmentée (ADF) dans le cas avec constante nulle et sans tendance. Ce test a pour hypothèse nulle la présence d'une racine unitaire synonyme de non-stationnarité. Plusieurs retards ont successivement été ajoutés jusqu'à ce que les coefficients associés ne soient plus significatifs. Le retard maximal trouvé est égal à 2. Le test donne une probabilité critique de $0,01$, ce qui permet de rejeter au seuil de 5% l'hypothèse nulle. La série z_t est donc considérée comme stationnaire.

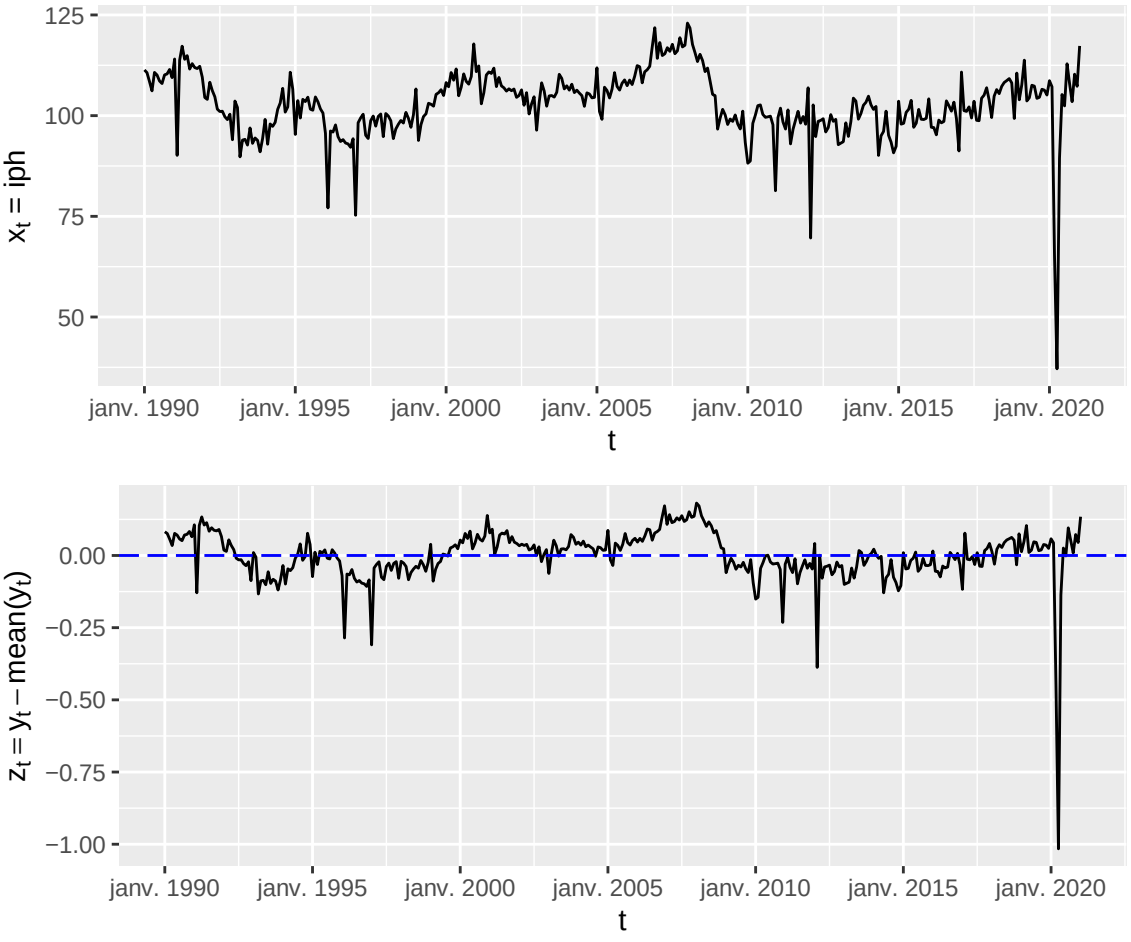

Figure 1: Représentation graphique de la série étudiée x_t et de la série transformée z_t

2 Modélisation ARMA : Méthodologie de Box et Jenkins

2.1 Identification des ordres maximaux pertinents

Dans cette étape, nous ferons usage de l'autocorrélogramme (ACF) et de celui partielle (PACF) pour déterminer les ordres maximaux q et p du modèle ARMA. En effet, il est prouvé que dans le cas d'un $MA(q)$ la fonction d'autocorrélation $\rho(h)$ est nulle lorsque $h > q$. Il s'agit donc de voir au niveau de l'ACF l'ordre à partir duquel les autocorrélations ne sont plus significativement différentes de 0 (à l'extérieur de la région de confiance). L'ordre q_{max} ainsi trouvé est de 18. Un résonnement similaire est observé pour ce qui est de l'ordre vu que la fonction d'autocorrélation partielle $r(h)$ d'un processus $AR(p)$ est également nulle lorsque $h > p$. L'ordre p_{max} est donc égal à 6.

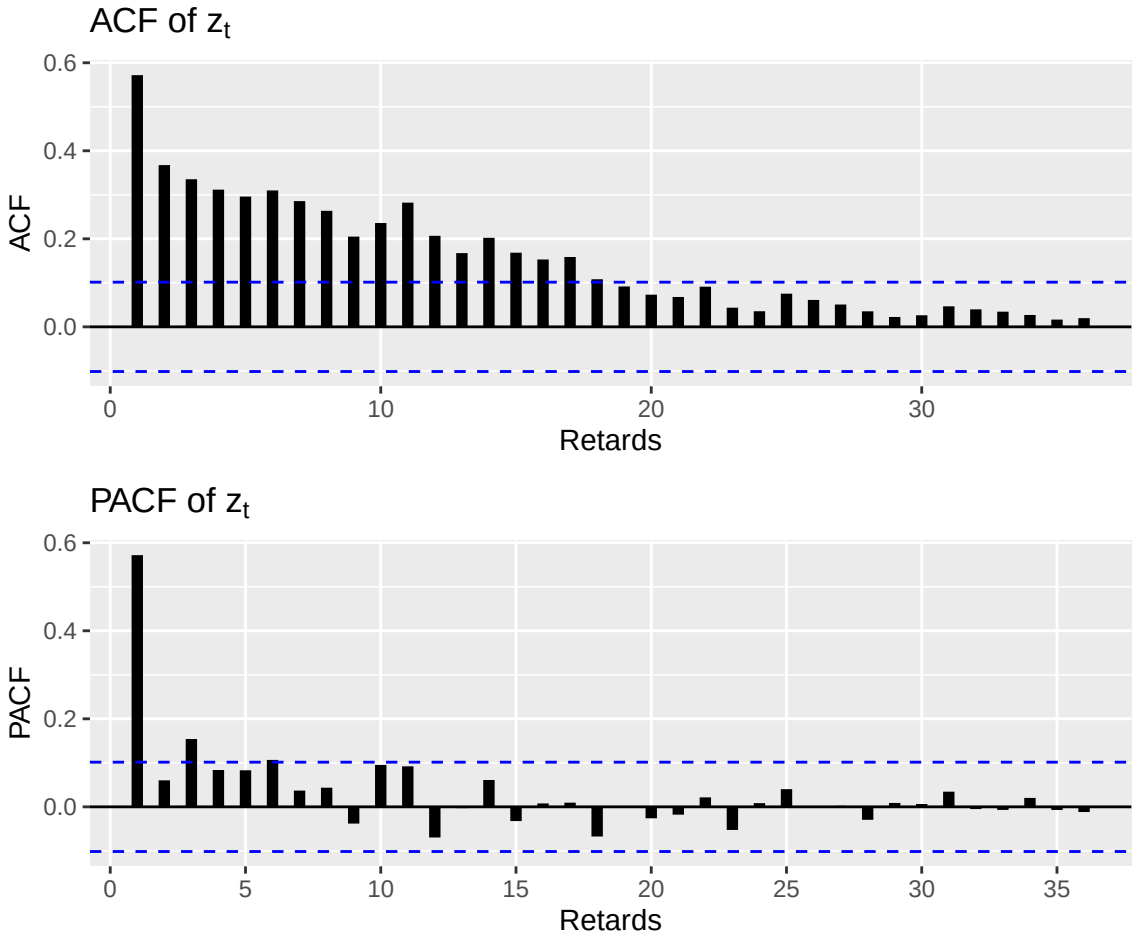

Figure 2: Représentation ACF et PACF de z_t

2.2 Sélection des modèles

Cette étape consiste à trouver lesquels parmi les 137 (19×7) modèles possibles satisfont les conditions de validité. Il s'agit de voir si, d'une part, le modèle est bien estimé en étudiant notamment la significativité des coefficients des ordres AR et MA les plus élevés. D'autres part, il est question de voir par le bias d'un test si les résidus issus de l'estimation des modèles sont autocorrélés. Dans le cas où ils le sont, le modèle ne sera pas retenu. C'est le test de Portemanteau/Ljung-box qui sera ici utilisé pour juger de cette autocorrélation. Cette sélection permet de se ramener à 15 modèles candidats qui vérifient les conditions précités (modèle bien estimé, non autocorrélation). Pour choisir parmi ces modèles lequel est le meilleur, on fait recours à des

mesure telles que le AIC et le BIC, le meilleur modèle (celui qu'on favorise le plus) étant celui dont la valeur est la plus faible. Ce dernier filtre met en évidence le modèle ARMA(3,4) et ARMA(1,2) qui minimisent respectivement le AIC et le BIC. On remarquera que le BIC favorise un modèle moins complexe, ce qui est dû à la pénalisation effectuée dans son calcul. C'est sur ce modèle que notre choix sera porté puisqu'étant plus parcimonieux. Le modèle retenu s'écrit alors :

$$z_t + 0,939 z_{t-1} = \epsilon_t - 0,449 \epsilon_{t-1} - 0,264 \epsilon_{t-2}$$

```
##  tests de nullité des coefficients :
##          ar1      ma1      ma2 intercept
## coef 0.93953 -0.44859 -0.26432   0.00315
## se   0.03016  0.06105  0.05540   0.01736
## pval 0.00000  0.00000  0.00000   0.85601
##
##  tests d'absence d'autocorrélation des résidus :
##      lag  pval lag  pval lag  pval lag  pval
## [1,]   1    NA   7 0.871  13 0.543  19 0.754
## [2,]   2    NA   8 0.883  14 0.509  20 0.792
## [3,]   3    NA   9 0.786  15 0.593  21 0.805
## [4,]   4 0.738  10 0.840  16 0.671  22 0.810
## [5,]   5 0.825  11 0.454  17 0.663  23 0.814
## [6,]   6 0.743  12 0.547  18 0.693  24 0.812
```

Table 1: Qualité d'ajustement et Validité du modèle ARMA(1,2) choisi

lag	pval	lag	pval	lag	pval	lag	pval
1	NA	7	0.871	13	0.543	19	0.754
2	NA	8	0.883	14	0.509	20	0.792
3	NA	9	0.786	15	0.593	21	0.805
4	0.738	10	0.840	16	0.671	22	0.810
5	0.825	11	0.454	17	0.663	23	0.814
6	0.743	12	0.547	18	0.693	24	0.812

2.3 Tests complémentaires

Dans cette partie, le focus sera mis sur l'étude des propriétés de normalité et d'hétéroscedasticité des résidus obtenus avec le modèle ARMA(1,2). Pour ce qui est de l'hypothèse de normalité, elle intervient principalement dans la construction des intervalles de confiance des prévisions qui seront faites. Pour juger de la validité de cette hypothèse, nous utiliserons le test de Jarque-Bera. L'hypothèse nulle de ce test est la normalité de la série testée. Avec une probabilité critique de $2,2.10^{-26}$, cette hypothèse est rejetée au seuil de 1%. Les résidus ne sont donc pas normaux. Concernant l'hétéroscédasticité, le test de Breusch-Pagan dont l'hypothèse nulle est la l'homoscédasticité des résidus, est rejetée au seuil de 5% avec une valeur critique égale à $0,016$. Il n'y a donc pas d'homoscédasticité.

3 Prévisions

3.1 Région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2})

Notons T la longueur de la série Z_t. On va d'abord supposer que le modèle obtenu à la question précédente est bien celle qui régit le processus générateur des observations de ladite série, c'est-à-dire que :

$$Z_t \sim ARMA(1,2)$$

Dans ce cas, le modèle théorique (avec les vrais coefficients) s'écrit sous sa forme canonique comme suit :

$$(1-\phi_1 L)Z_t = (1-\theta_1 L)(1-\theta_2 L)\epsilon_t$$

soit de manière équivalente :

$$Z_t = \phi Z_{t-1} + \epsilon_t - (\theta_1+\theta_2)\epsilon_{t-1} + \theta_1\theta_2\epsilon_{t-2}$$

On obtient par la suite les prévisions linéaires de Z_{T+1} et de Z_{T+2} compte tenu de l'information qu'on a à la date T comme suit :

$$\begin{cases} {}_T Z_{T+1} = \mathbb{EL}[Z_{T+1}|Z_T, Z_{T-1}, ...] = \phi Z_T - (\theta_1+\theta_2)\epsilon_T + \theta_1\theta_2\epsilon_{T-1} \\ {}_T Z_{T+2} = \mathbb{EL}[Z_{T+2}|Z_T, Z_{T-1}, ...] = \phi \times_T Z_{T+1} + \theta_1\theta_2\epsilon_T = \phi^2 Z_T + (\theta_1\theta_2 - \phi(\theta_1+\theta_2))\epsilon_T + \phi\theta_1\theta_2\epsilon_{T-1} \end{cases}$$

Il en résulte :

$$\begin{cases} Z_{T+1} -_T Z_{T+1} = \epsilon_{T+1} \\ Z_{T+2} -_T Z_{T+2} = \phi(Z_{T+1} -_T Z_{T+1}) + \epsilon_{T+2} - (\theta_1+\theta_2)\epsilon_{T+1} = (\phi-\theta_1-\theta_2)\epsilon_{T+1} + \epsilon_{T+2} \end{cases}$$

En posant :

$$Z = \begin{pmatrix} Z_{T+1} \\ Z_{T+2} \end{pmatrix} \quad , \quad Z^* = \begin{pmatrix} {}_T Z_{T+1} \\ {}_T Z_{T+2} \end{pmatrix} \quad \text{et} \quad \epsilon = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} \end{pmatrix}$$

Ce qui précède se réécrit :

$$Z - Z^* = A\epsilon$$

où

$$A = \begin{pmatrix} 1 & 0 \\ \phi-\theta_1-\theta_2 & 1 \end{pmatrix}$$

On va ensuite supposer que le terme d'erreur est un bruit blanc Gaussien. Ceci dit, on a :

$$\epsilon \sim \mathcal{N}_2(0, \sigma_\epsilon^2 I_2)$$

Il s'ensuit :

$$Z - Z^* \sim \mathcal{N}_2(0, \Sigma)$$

où

$$\Sigma = A(\sigma_\epsilon^2 I_2)A' = \sigma_\epsilon^2 AA'$$

c'est-à-dire

$$\Sigma = \sigma_\epsilon^2 \begin{pmatrix} 1 & 0 \\ \phi-\theta_1-\theta_2 & 1 \end{pmatrix}\begin{pmatrix} 1 & \phi-\theta_1-\theta_2 \\ 0 & 1 \end{pmatrix} = \sigma_\epsilon^2 \begin{pmatrix} 1 & \phi-\theta_1-\theta_2 \\ \phi-\theta_1-\theta_2 & 1+(\phi-\theta_1-\theta_2)^2 \end{pmatrix}$$

Puisque $det(\Sigma) = \sigma_\epsilon^4 > 0$, alors Σ est inversible et on en déduit :

$$(Z-Z^*)'\Sigma^{-1}(Z-Z^*) \sim \chi^2(2)$$

Dans le cadre d'un ARMA(1,2) comme spécifié ci-dessus, on peut montrer que :

$$\epsilon_t = \sum_{n=0}^{+\infty} c_n(Z_{t-n} - \phi Z_{t-n-1}) \qquad \text{où} \qquad c_n = \sum_{k=0}^{n} \theta_1^{n-k}\theta_2^k$$

L'observation de $(Z_t)_{t\leqslant 0}$ étant non disponible dans la base, on va émettre une hypothèse forte selon laquelle $Z_t = 0 \quad \forall t \leqslant 0$. Dans ce cas, à la date T, on sera en mesure (à coefficients AR et MA supposés connus) de calculer ϵ_t pour tout $t \leqslant T$ comme suit :

$$\epsilon_t = \begin{cases} 0 \quad \text{si} \quad t \leqslant 1 \\ \displaystyle\sum_{n=0}^{t-1} c_n(Z_{t-n} - \phi Z_{t-n-1}) \quad \text{où} \quad 1 < t \leqslant T \end{cases}$$

Ainsi, à la date T, compte tenu de l'information que l'on dispose, on est en mesure de calculer ϵ_t pour tout $t \leqslant T$, ${}_TZ_{T+1}$, ${}_TZ_{T+2}$ et $\hat{\sigma}_\epsilon^2 := \frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2$ (à coefficients AR et MA supposés connus), ce dernier étant un estimateur convergent de σ_ϵ^2 du fait que le bruit blanc est supposé Gaussien (et donc ergodique).
En définissant $\hat{\Sigma} = \hat{\sigma}_\epsilon^2 \begin{pmatrix} 1 & \phi-\theta_1-\theta_2 \\ \phi-\theta_1-\theta_2 & 1+(\phi-\theta_1-\theta_2)^2 \end{pmatrix}$ on a par le théorème de Slutsky :

$$(Z-Z^*)'\hat{\Sigma}^{-1}(Z-Z^*) \xrightarrow{Loi} \chi^2(2)$$

On peut donc déduire la région de confiance asymptotique (puisque T est assez grand) suivant pour $Z = \begin{pmatrix} Z_{T+1} \\ Z_{T+2} \end{pmatrix}$, de niveau $\alpha \in]0,1[$:

$$RC_\alpha^{asympt}(Z) = \left\{\underline{x} \in \mathbb{R}^2 : (\underline{x}-Z^*)'\hat{\Sigma}^{-1}(\underline{x}-Z^*) \leqslant q_\alpha^{Chi2}(2)\right\}$$

Il suffit par exemple de supposer que les coefficients du modèle ARMA estimés dans la partie 2. sont les vrais coefficients du modèle, et on détermine aisément cette région.
On en déduit aisément que pour $X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix}$, on a :

$$RC_\alpha^{asympt}(X) = exp\left(\begin{pmatrix} mean(y_t) \\ mean(y_t) \end{pmatrix} + RC_\alpha^{asympt}(Z)\right)$$

3.2 Hypothèses utilisées

Récapitulons. Si l'on suppose les hypothèses suivantes :
`H1` : Le modèle spécifié est le vrai processus générateur des observations de Z_t
`H2` : Les erreurs ϵ_t suivent un bruit blanc Gaussien
`H3` : La série Z_t considérés est tronquée : $Z_t = 0 \quad \forall t \leqslant 0$
`H4` : Les coefficients estimés sont les vrais paramètres (paramètres théoriques) du modèle.
Alors, la région de confiance obtenu ci-dessus est valide.
La modélisation faite à la partie 2. rend compte des hypothèses `H1` (choix du modèle par qualité d'ajustement et validité) et `H2` (test de portemanteau, d'hétéroscédasticité et de normalité). Une fois celles-ci admises, les estimateurs des paramètres théoriques sont convergents. Donc T étant assez grand, l'hypothèse `H4` semble ne pas poser un grand soucis. L'hypothèse `H3` de troncature quand à elle, souvent admise est pratique, notamment quand on a un T assez grand, tel le cas ici, nous semble assez restrictif car cela reviendrait à supposer que l'indice de production industrielle du secteur d'activité d'exploitation de gravières et sablières, d'extraction d'argiles et de kaolin, était toujours le même avant janvier 1990, date à partir de laquelle on a des observations dans notre base de données.

3.3 Représentation graphique de cette région pour $\alpha = 95\%$

On représente graphiquement la région de confiance pour un niveau $\alpha = 95\%$. La limite supérieure (resp.inférieure) de la zone en bleu représente la frontière supérieure (resp.inférieure) de la région de confiance. Le trait en bleu foncé relit les deux prédictions ponctuelles pour les dates T+1 et T+2.

Table 2: Prévision de la série originale à T et T+1

x
109.0342
105.0048

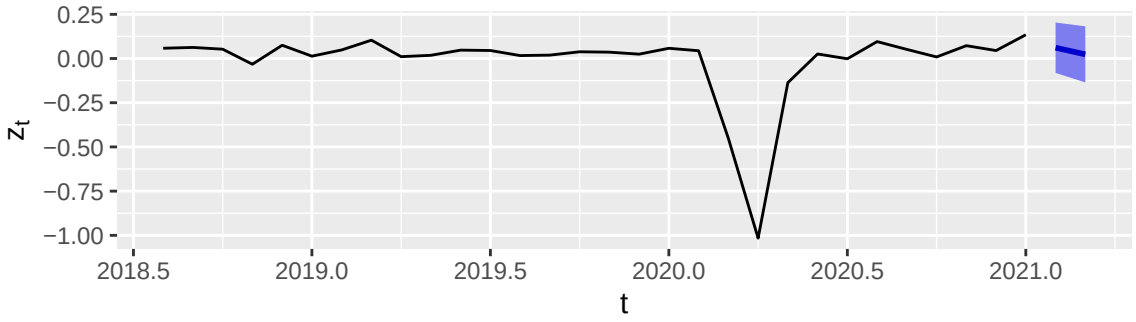

Figure 3: Prévisions en T et T+1 avec région de confiance à 95%

Commentaires : Comme on a vu dans la partie 2. que les résidus n'étaient pas normaux, la région de confiance, dont la détermination dépend fortement de cette hypothèse, n'est donc plus interprétable.

3.4 Question ouverte

Soit Y_t une série stationnaire disponible de t = 1 à T. On suppose que Y_{T+1} est disponible plus rapidement que X_{T+1}. Il est question ici de chercher à savoir sous quelles conditions cette information permet-elle d'améliorer la prévision de X_{T+1}.
Soit $\hat{X}_{T+1|\{X_t,t\leqslant T\}}$ (resp. $\hat{X}_{T+1|\{Y_t,X_t,t\leqslant T\}\cup\{Y_{T+1}\}}$) la prévision linéaire optimal à la date t (au sens L^2) de la variable X_t étant données les variables $\{X_t, t \leqslant T\}\}$ (resp. $\{Y_t, X_t, t \leqslant T\} \cup \{Y_{T+1}\}$). Si Y_t *cause instantanément* X_t, alors $\hat{X}_{T+1|\{X_t,t\leqslant T\}} \neq \hat{X}_{T+1|\{Y_t,X_t,t\leqslant T\}\cup\{Y_{T+1}\}}$ et donc Y_{T+1} est *utile* pour prédire X_{T+1} à la date T. Cette condition de *causalité instantanément* permet ainsi d'améliorer la prévision de X_{T+1} à la date T.