Tutoriel de séries temporelles linéaires

Thomas Dujardin, Bastian Dupoirieux

Avril - Mai 2022

Partie I : Les données

1. Que représente la série choisie ? (secteur, périmètre, traitements éventuels, transformation logarithmique, etc.)

La série choisie représente l'indice CVS-CJO (Corrigé des Variables Saisonnières - Corrigé des Jours Ouvrés) en base 100 en 2015 (donc de moyenne 100 en 2015) de la production aéronautique et spatiale entre janvier 1990 et janvier 2020.

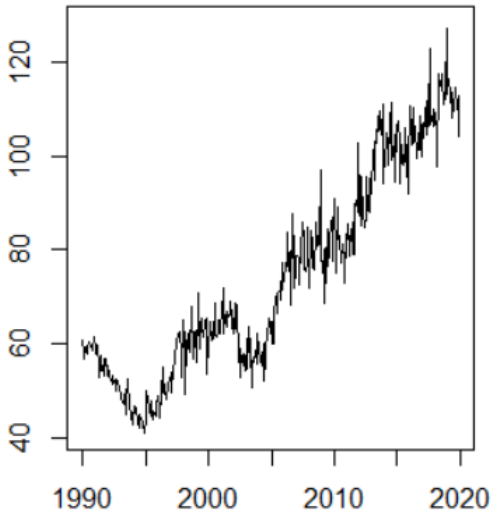

Figure 1 – Série initiale

2. Transformer si besoin la série pour la rendre stationnaire (différentiation, suppression de la tendance déterministe, etc.). Justifier soigneusement vos choix.

La figure 1 ci-dessus montre une tendance. Une régression linéaire de l'indice sur la date montre que les coefficients correspondant à une tendance (date et constante) sont significatifs (p-value de la nullité jointe inférieure à 0.05). Nous allons donc effectuer un test de la racine unitaire avec constante et tendance. Il faut d'abord vérifier la non-autocorrélation des résidus de la régression linéaire ci-dessus. Un premier ADF sans lags avec les 24 derniers résidus (donc sur deux ans) est rejeté à tous les ordres. Il faut alors inclure au minimum 7 lags afin de supprimer l'autocorrélation des résidus.

F-statistic: 2166 on 1 and 358 DF, p-value: < 2.2e-16

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.640647	0.858404	46.18	<2e-16
seq_int	0.191824	0.004121	46.54	<2e-16

TABLE 1 – Régression linéaire de l'indice sur la date

Ce test ADF à 7 lags donne une p-value pour la tendance de **0.1595** > 0.05 : on ne peut pas rejeter la non-stationnarité au seuil de 5 (pourcent). Il est donc nécessaire de différencier au moins une fois la série.

Testons la présence d'une tendance dans la série différenciée :

F-statistic: 0.06784 on 1 and 357 DF, p-value: 0.7947

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0170349	0.6078433	-0.028	0.978
seq_int[-1]	0.0007591	0.0029144	0.260	0.795

TABLE 2 – Régression linéaire de l'indice différencié sur la date

Nous déterminons ensuite qu'il faut 6 lags pour supprimer l'autocorrélation des résidus de cette seconde régression linéaire, et qu'un test ADF à 6 lags donne une p-value de moins de **0.01**, donc un rejet de l'hypothèse de racine unitaire à tous les niveaux, et donc la stationnarité de la série différenciée.

3. Représenter graphiquement la série choisie avant et après transformation.

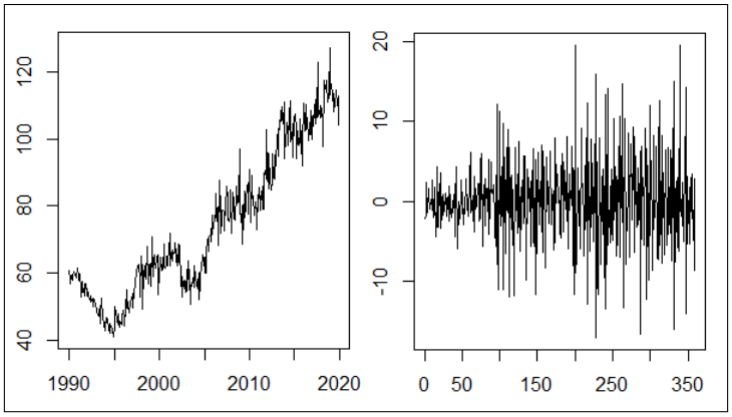

FIGURE 2 – Série brute (gauche) et série différenciée (droite)

Partie II : Modèles ARMA

4. Choisir, en le justifiant, un modèle ARMA(p,q) pour votre série corrigée X_t. Estimer les paramètres du modèle et vérifier sa validité.

On détermine les fonctions d'autocorrélations et d'autocorrélations partielles :

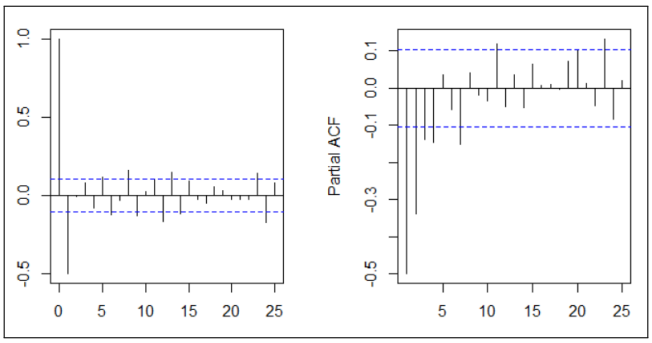

FIGURE 3 – ACF, PACF de la série différenciée

On observe des valeurs significatives jusqu'à des ordres élevés, mais on ne retiendra que les pics $p_{max} = 11$ (d'après la PACF) et $q_{max} = 14$ (d'après l'ACF).
On détermine alors les AICs et BICs des $ARIMA(p, 1, q)$ où $p \leqslant p_{max}$ et $q \leqslant q_{max}$. On obtient que $ARIMA(0, 1, 1)$ (donc MA(1)) et $ARIMA(2, 1, 2)$ minimisent respectivement BIC et AIC.

On va maintenant tester les autocorrélations des résidus de ces deux modèles, et tester leur ajustement :

	ar1	ar2	ma1	ma2		ma1
	-1.0811	-0.1591	0.3543	-0.5954		-0.6869
s.e.	0.0752	0.0737	0.0607	0.0596	s.e.	0.0345
	ARIMA(2,1,2)					MA(1)

FIGURE 4 – Coefficients des modèles ARIMA(2,1,2) et MA(1)

Tous les coefficients ci-dessus sont significatifs, puisque la valeur absolue du rapport entre le coefficient estimé et l'erreur standard est toujours supérieure à 1.96. Les deux modèles sont donc ajustés. L'absence d'autocorrélation des résidus est rejetée pour quasiment tous les lags

lag	pval		
1	NA	12	0.03692930
2	NA	13	0.03555529
3	NA	14	0.05398024
4	NA	15	0.04862661
5	0.60901014	16	0.07069630
6	0.21885490	17	0.09920586
7	0.26689916	18	0.05345918
8	0.02558872	19	0.03762855
9	0.02740894	20	0.05349947
10	0.03746075	21	0.06768676
11	0.04615625	22	0.09104888
		23	0.09316970
		24	0.02000212

FIGURE 5 – Autocorrélation des résidus du modèle ARIMA(2,1,2)

pour le MA(1), mais seulement pour quelques-uns pour l'ARIMA(2,1,2). On sélectionne donc ce modèle faute de mieux, même s'il n'est pas parfaitement validé.

5. Exprimer le modèle ARIMA(p,d,q) pour la série choisie.

Nous avons sélectionné un $ARIMA(2, 1, 2)$, qui peut s'écrire :$X_t = -1.0811X_{t-1} - 0.159X_{t-2} + \varepsilon_t + 0.3543\varepsilon_{t-1} - 0.5954\epsilon_{t-2}$

Partie III : Prévision

6. Ecrire l'équation vérifiée par la région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2})

En supposant les résidus gaussiens de même écart-type σ ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$), et en notant $q_{1-\alpha}$ le quantile d'ordre $1-\alpha$ d'une $\chi^2(2)$ (2 étant le nombre de dimensions de la région de confiance), on a : $R_\alpha = \{x \in \mathbb{R}^2 | x^t \sigma x \leqslant q_{1-\alpha}\}$

7. Préciser les hypothèses utilisées pour obtenir cette région.

Nous avons besoin de deux hypothèses :
— Les coefficients estimés en Q4 et Q5 sont les coefficients théoriques de l'ARIMA ;
— Les (ε_t) sont i.i.d., centrées et gaussiennes d'écart-type σ^2 positif connu.

8. Représenter graphiquement cette région pour $\alpha = 95$ %. Commenter.

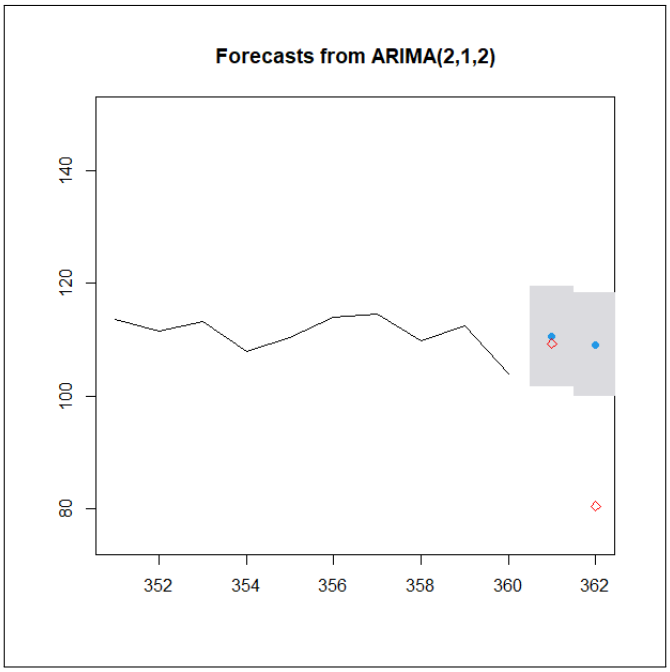

FIGURE 6 – Prédictions pour les mois de février et mars 2020. Les points bleus représentent les valeurs prédites, la zone grise représente la région de confiance et les losanges représentent les valeurs réelles.

On représente ici les valeurs (X_{T+1}, X_{T+2}) (T=janvier 2020) avec leur région de confiance à 95 %. La valeur prédite pour février 2020 est dans la région et est très proche de la valeur réelle, mais la valeur prédite pour mars 2020 n'est pas inclue dans la région de confiance. Cela s'explique par la forte baisse d'activité dans ce secteur conséquemment à la pandémie de COVID-19.

9. 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}. Sous quelle(s) condition(s) cette information permet-elle d'améliorer la prévision de X_{T+1} ? Comment la (les) testeriez-vous ?

Y_{T+1} est utile dans la prédiction instantanée de X_{T+1} ssi Y cause X instantanément au sens de Granger, soit : $\hat{X}_{T+1}|\{Y_u, X_u, u < T\} \cup \{Y_{T+1}\} \neq \hat{X}_{T+1}|\{Y_u, X_u, u < T\}$.
Afin de tester cette condition, (chapitre 5).