ENSAE Paris

Projet de Séries Temporelles Linéaires

Analyse d'un indice de production industrielle (base 100) : Production de vin (avec raisin)

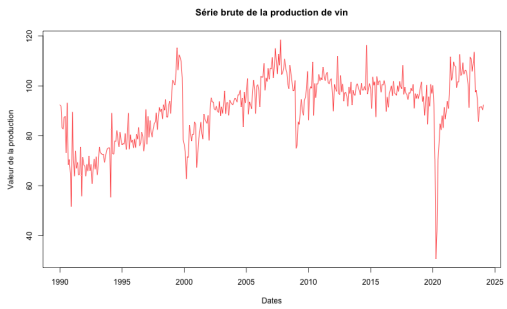
Eva-Andrée TIOMO et Tom ROSSA

21 Mai 2024

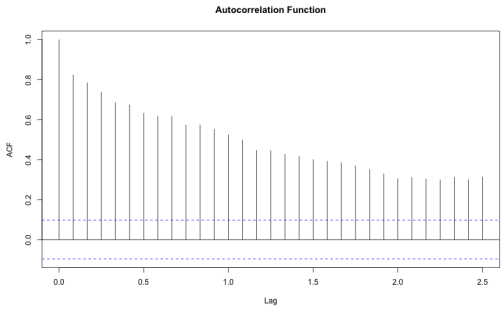
1 Présentation des données

1.1 Description de la série choisie

La série temporelle que nous allons étudier tout au long de ce projet est un des indices bruts de la production industrielle en France construit par l'INSEE. Nous avons choisi une série préalablement corrigée de ses variations saisonnières et des jours ouvrés (CVS-CJO) qui représente la production mensuelle de vin (avec raisin), en France Métropolitaine entre janvier 1990 et février 2024. Cette série a été révisée à deux reprises par les équipes de l'INSEE et nous avons choisi la révision la plus récente. Cet indice est exprimé en base 100 à partir de la donnée de la production de vin sur l'année 2021.



(a) Représentation graphique de la série brute



(b) Auto-corrélogramme de la série brute

En observant une représentation graphique de la série, on remarque qu'il ne se dégage pas de tendance déterministe nette sur la période et qu'elle ne semble pas avoir de saisonnalité. On observe plusieurs pics de forte croissance ou décroissance de la série qui peuvent être expliqués par des évènements météorologiques exceptionnels dont dépend fortement la production de vin. La stationnarité de la série ne semble pas très claire et bien que le test de Dickey-Fuller augmenté ne permette pas de rejetter l'hypothèse de non-stationnarité à des niveaux significatifs, nous pouvons en douter. D'autant que le test de KPSS indique qu'il est possible de rejetter à tous les niveaux l'hypothèse de stationnarité, et que l'autocorrélogramme indique l'autocorrélation d'ordre 1 de la série est assez proche de 1.

1.2 Transformation de la série

Ainsi, nous doutons de la stationnarité de notre série et soupçonnons qu'elle soit intégrée d'ordre 1. Nous avons donc procédé à une différenciation de la série pour travailler sur la série des $(\Delta \mathbf{X}_i)$. En observant la série différenciée, l'hypothèse de stationnarité semble bien plus plausible. La série semble centrée et varie autour de l'axe des abscisses sans qu'aucune tendance ne s'en dégage (Cf Annexes : Figure 7).

A titre indicatif, nous procédons à une régression linéaire de la série différenciée sur l'indice $\mathbf{t}$ et l'on voit très nettement que le coefficient associé est très proche de 0. Indice supplémentaire d'une

absence de tendance déterministe. Ainsi, nous procédons à un test de Dickey-Fuller augmenté (ADF) avec constante et sans tendance. Nous vérifions que les résidus du modèle de régression induit par ce test ne soient pas autocorrelés à l'aide d'un test de Ljung-Box, sans quoi nous devrons ajouter des retards au test ADF. Il se trouve que pour un test ADF à 0, 1 ou 2 retards, l'absence d'auto-corrélations des résidus jusqu'à l'ordre est rejetée au moins une fois par la série de tests de Ljung-Box. Nous ajoutons donc 3 retards au test ADF et voici la série de p-valeurs des tests de Ljung-Box que nous obtenons :

Lag	**p-value**
1	0.9366109
2	0.9772932
3	0.9426064
4	0.9687178
5	0.7462071
6	0.3453311
7	0.3339071
8	0.322031
9	0.2897886
10	0.3158187

Table 1: p-valeurs test Ljung-Box

Coefficients	**Estimate**	**Std. Error**	**t value**	**P-valeurs**
(Intercept)	0.02496	0.34843	0.072	0.942928
z.lag.1	-1.90482	0.13986	-13.620	$< 2 \times 10^{-16}$ ***
z.diff.lag1	0.44696	0.11780	3.794	0.000171 ***
z.diff.lag2	0.25174	0.08691	2.897	0.003981 **
z.diff.lag3	0.14749	0.04933	2.990	0.002961 **

Table 2: Test ADF

Les tests de Ljung-Box ont des p-valeurs assez grandes qui montrent que l'on ne peut pas rejetter l'hypothèse d'absence d'auto-corrélation des résidus du test ADF. Ce test ADF qui est effectué sans tendance, avec constante et avec 3 retards est donc bel est bien valide. De plus, on voit qu'il rejette à tous les niveaux de significativité usuels l'hypothèse d'une racine unitaire de notre processus. On peut donc conclure que la stationnarité de notre série différenciée semble très plausible. On s'en convainct définitivement avec un test KPSS avec l'hypothèse nulle de stationnarité, que l'on ne peut rejetter à aucun niveau de significativité usuel (Cf Annexes : Sortie du test)

1.3 Comparaison avant et après différenciation

Ainsi, nous avons que l'hypothèse de stationnarité de notre série initiale paraissait assez peu plausible. Cela a été confirmé par le test de KPSS qui rejettait assez largement l'hypothèse. Néanmoins, après différenciation de la série, nous sommes parvenus à démontrer que l'hypothèse de stationnarité était là bien plus plausible et cela s'observe assez nettement sur le graphique conjoint ci-dessous :

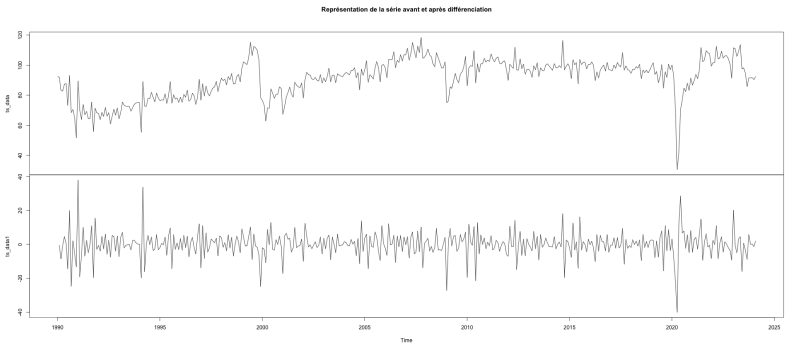
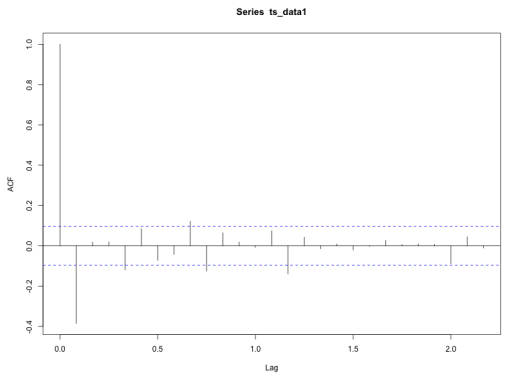


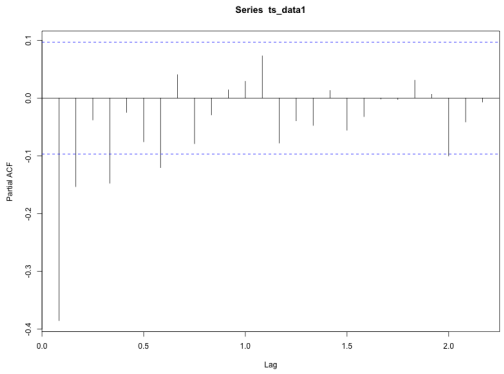
Figure 2: Représentation graphique de la série avant et après différenciation

2 Modèle ARMA

Dans cette partie, nous allons déterminer le modèle ARIMA qui permettent le mieux de modéliser notre série. Nous avons conclut dans la première partie que la série différenciée semblait stationnaire et nous allons de choisir un modèle ARMA pour cette série. Tout d'abord, il convient d'étudier les auto-corrélogrammes et auto-corrélogrammes partiels de la série :



(a) Auto-corrélogramme de la série différenciée



(b) Auto-corrélogramme partiel de la série différenciée

Figure 3: Deux graphiques côte à côte

Nous pouvons observer sur l'auto-corrélogramme de la série $(\Delta \mathbf{X}_i)$ qu'à partir du 14ème retard, les autocorrélations ne sont plus significativement différentes de 0. Cela laisse penser que le paramètre **q** du modèle ARMA serait donc inférieur à 14. De même, on observe sur auto-corrélogramme partiel de la série qu'à partir du 7ème lag, aucun coefficient n'est significativement supérieur à 0. Donc les ARIMA candidats sont donc ceux dont les coefficients **p** et **q** sont respectivement inférieurs à 7 et 14 et où $\mathbf{d} = 1$ puisque la série est intégrée d'ordre 1.

Dans un premier temps, nous allons régresser les modèles ARIMA candidats sur notre série et nous utiliserons les critères d'informations BIC et AIC pour faire une première sélection. Nous garderons les 2 modèles qui minimisent le BIC et les 2 qui minimisent le AIC parmi nos candidats. On rappelle que ces critères sont inversement proportionnels à la vraisemblance du modèle et pénalise les modèles les plus complexes.

Nous conservons les modèles ARIMA(5, 1, 3), ARIMA(6, 1, 3), ARIMA(1, 1, 2) et ARIMA(2, 1, 1) après cette première sélection. Il convient désormais de tester la significativité des coefficients AR et MA d'ordre le plus élevé à chaque fois. Pour le modèle ARIMA(5, 1, 3), on remarque que la p-valeurs associée au coefficient AR(5) et très élevée et ne permet pas de rejetter l'hypothèse de nullité de ce coefficient aux stades usuels (Cf Annexes : Table 4). Nous avons procédé de la même façon pour le coefficient AR(6) pour le modèle ARIMA(6, 1, 3) et avons montré que l'on ne pouvait pas rejetter l'hypothèse de nullité du coefficient AR(6) dont la p-valeur est élevée (Cf Annexes : Table 5).

A ce stade, il ne nous reste plus que 2 modèles dont les coefficients sont significatifs à tous les niveaux et dont les critères d'informations AIC et BIC sont les plus faibles. Il est nécessaire de tester l'absence d'autocorrélation des résidus. Nous procédons avec des tests de Ljung-Box pour des retards allant de 4, car la statistique de test converge vers une loi χ^2_{k-p-q}, à 50. Il se trouve que pour le modèle ARIMA(2, 1, 1), nous ne pouvons plus rejetter l'hypothèse nulle au-delà de 30 retards. Cela n'est pas le cas pour le modèle ARIMA(1, 1, 2) que nous avons donc choisi afin de modéliser au mieux notre série.

3 Prévisions

3.1 Équation vérifiée par la région de confiance sur les valeurs futures

Nous nous intéressons désormais à de la prévision. Il est important de préciser que dans le tout le reste de notre papier, les résidus de la série sont supposés gaussiens. Ces prévisions sont essentielles en général dans les cas pratiques. Pour rappel, le modèle choisi pour la série (X_t) est un ARIMA (2,1,1) dont on dispose de T observations. On cherche à prédire X_{T+1} et X_{T+2}.
Comme (X_t) est un ARMA (2,1), on peut l'écrire de la manière suivante, $\forall \mathbf{t}$:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t - \psi_1 \epsilon_{t-1} \tag{1}$$

De plus, d'après le cours, la meilleures prédiction linéaire pour X_{T+1} est donnée par :

$$\hat{X}_{T+1|T} = \phi_1 X_T + \phi_2 X_{T-1} - \psi_1 \epsilon_T$$

Par ailleurs, la meilleure prédiction pour X_{T+2} est donnée par :

$$\hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \phi_2 X_T - \psi_1 \epsilon_{T+1|T}$$

Avec $\epsilon_{T+1|T} = 0$ car nous avons fait l'hypothèse qu'il s'agit d'un bruit blanc, ainsi, ϵ_{T+1} est décorrélé de ϵ_T. Ainsi, $E[\epsilon_T | X_{T_1}, X_{T-2}...] = 0$ puisque l'espace engendré par toutes les combinaisons linéaires des (X_T, X_{T-1}) est le même que celui engendré par les combinaisons linéaires des résidus passés.

Donc nous avons :

$$\hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \phi_2 X_T$$

Nous nous appuyons sur les deux équations suivantes pour le reste du projet :

$$\begin{cases} \hat{X}_{T+1|T} = \phi_1 X_T + \phi_2 X_{T-1} - \psi_1 \epsilon_T \\ \hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \phi_2 X_T \end{cases}$$

où:

- $\hat{X}_{T+1|T}$ est la prévision de X_{T+1} conditionnelle aux observations jusqu'à T
- $\hat{X}_{T+2|T}$ est la prévision de X_{T+2} conditionnelle aux observations jusqu'à T
- ϕ_1 et ϕ_2 sont les coefficients autoregressifs du modèle AR
- ψ_1 est le coefficient de la moyenne mobile du modèle MA
- X_t sont les valeurs observées de la série temporelle

On se concentre désormais sur les erreurs de prédiction $X_{T+1} - \hat{X}_{T+1|T}$ et $X_{T+2} - \hat{X}_{T+2|T}$ que nous cherchons à calculer.

On obtient :

$$\begin{cases} X_{T+1} - \hat{X}_{T+1|T} = \epsilon_{T+1} \\ X_{T+2} - \hat{X}_{T+2|T} = \epsilon_{T+2} + (\phi_1 + \psi_1)\epsilon_{T+1} \end{cases}$$

On note :

$$\hat{X} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix} \; et \; X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix} \tag{2}$$

On a : $X - \hat{X} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + (\phi_1 + \psi_1)\epsilon_{T+1} \end{pmatrix}$

Comme les résidus sont gaussiens, on sait que $X - \hat{X} \sim \mathcal{N}(0, \Sigma)$ avec Σ la matrice de variance-covariance :

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \phi_1 + \psi_1 \\ \phi_1 + \psi_1 & 1 + (\phi_1 + \psi_1)^2 \end{pmatrix}$$

On a $det(\Sigma) = \sigma^2 > 0$ (a priori). On en déduit que Σ est inversible.

Cette propriété sur la matrice de variance-covariance nous permet d'appliquer un théorème du cours:

$(X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$. Ce résultat nous permet de déduire un intervalle de confiance au niveau α:

$$\{X \in \mathbb{R}^2 | (X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \leq q_{1-\alpha}^{\chi^2(2)}\}$$

3.2 Hypothèses utilisées pour obtenir cette région de confiance

Pour obtenir cette région de confiance, plusieurs hypothèses sont posées. La série (X_t) doit être stationnaire avec une moyenne $\mathbb{E}[X_t] = m$ et une autocovariance γ déjà connues. On suppose également que la série (X_t) est un ARMA inversible et causal. Enfin, l'hypothèse sur les résidus est essentielle : il faut également que les résidus soient gaussiens.

Nous avons fait le choix d'un modèle ARMA (2,1). Nous nous assurons qu'il est bien inversible et causal car on sait déjà que la série sur laquelle nous travaillons est stationnaire (grâce à un travail de différenciation [cf partie 2]).

Pour tester ces hypothèses, on s'assure que les polynomes associés aux modèles AR(2) et MA(1) n'admettent pas de racine de module inférieur à 1. Nous cherchons dans un premier temps à définir explicitement les polynômes ϕ et ψ qui sont les polynômes de notre ARMA(2,1). Pour rappel, on a ϕ_i (resp. ψ_i) les coefficients du polynôme ϕ (resp. ψ). Nous obtenons la valeur des coefficients :

Coefficients du modèle ARMA(2,1)

Coefficients	Valeurs
ϕ_1	0.4867946
ϕ_2	0.2590099
ψ_1	-0.9633876

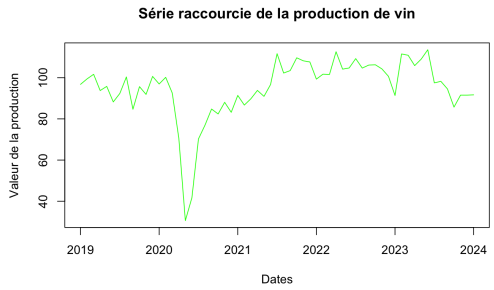
Figure 5: Coefficients des polynômes ϕ et ψ

Nous effectuons ensuite des tests qui nous permettent d'observer que les polynômes associés aux modèles AR(2) et MA(1) n'admettent pas de racine unitaire (ni de racine de module inférieur à 1). On conclut que le modèle ARMA développé est causal et inversible. Nous pouvons ainsi représenter graphiquement et justement l'intervalle de confiance que nous avons défini précédemment.

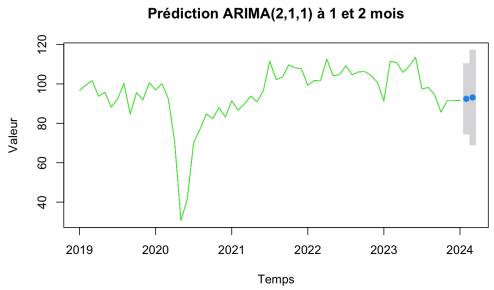
3.3 Représentation graphique

Dans cette partie, nous nous penchons sur les prévisions liées à notre série. Nous comparons notre série adaptée au modèle ARIMA (2,1,1) entre 2019 et début 2024 et une série comprenant des prédictions pour les deux premiers mois de janvier.

Nous avons donc tronqué la série initiale afin de pouvoir comparer les prédictions de notre modèle ARIMA pour Janvier et Février 2024, sachant uniquement les valeurs antérieures de la série. On observe que les valeurs prédites sont très proches des valeurs réellement observées de la production de vin sur ces deux dates. Concernant l'intervalle de confiance, on constate logiquement qu'il est plus large pour la prédiction de Février 2024. C'est tout à fait cohérent puisque la prédiction est fondée en partir sur la valeur prédite pour le mois de Janvier. L'erreur de prédiction vaut : $\epsilon_{T+2} + (\phi_1 + \psi_1)\epsilon_{T+1}$ dont la variance est supérieur à celle de l'erreur de prédiction pour la date antérieure.



(a) Production de vin entre 2019 et février 2024



(b) Prévisions Janvier et Fevrier 2024

Figure 6: Comparaison entre réalité et prévision

3.4 Ouverture et discussions

On suppose dans cette partie que (Y_t) est une série stationnaire disponible de t=1 à T. On suppose également que Y_{T+1} est disponible plus rapidement que X_{T+1} et on se demande dans quelles conditions cette information peut améliorer la prédiction de X_{T+1}.

Si Y_{T+1} est connu plus tôt que X_{T+1} d'une manière quelconque, cela peut potentiellement améliorer la prédiction de X_{T+1} si Y_{T+1} est corrélé avec X_{T+1} ou si elle fournit des informations supplémentaires sur le processus sous-jacent de X_t. C'est-à-dire que la prédiction de X_{T+1} sachant les données antérieures des deux séries seraient différente si l'on ajoute la donnée de la valeur de Y_{T+1}. Dans ce cas, on aura que le processus (Y_t) cause instantanément notre processus (X_t) au sens de Granger. Cela pourrait notamment s'expliquer si, (Y_t) peut être modéliser par un ARMA, et que son résidu ϵ_t^* est corrélé au résidu ϵ_t de notre série (X_t). Ajouter Y_{T+1} ajouterait de l'information à propos du résidus de X_T qui est décorrélé complétement des valeurs antérieures des séries et qui était jusqu'à lors la partie imprévisible du processus (X_t)

Pour tester cette hypothèse, nous pourrions effectuer une analyse de corrélation entre Y_t et X_t pour évaluer la force de leur relation. De plus, des méthodes de modélisation supplémentaires, telles que la régression ou les modèles de séries temporelles multivariées, pourraient être utilisées pour exploiter l'information supplémentaire fournie par Y_t dans la prédiction de X_{T+1}. Parmi ces méthodes, on peut notamment privilégier des modèles VAR ou SVAR en ce qu'ils permettent d'inclure Y_t comme une variable exogène dans le modèle de X_t si besoin.
Enfin, nous pourrions effectuer des tests statiques tels que les tests de de causalité de Granger pour vérifier si Y_t aide à prédire X_t en général.

4 Annexes

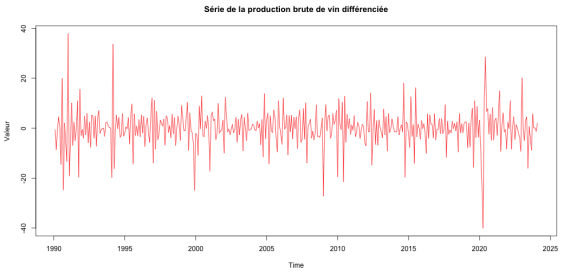


Figure 7: Représentation graphique de la série différenciée

Test	KPSS Unit Root Test			
Type	mu with 5 lags			
Test Statistic	0.0201			
Significance Level	**10%**	**5%**	**2.5%**	**1%**
Critical Value	0.347	0.463	0.574	0.739

Table 3: Résumé du test KPSS avec les valeurs critiques : test de stationnarité de la série différenciée

	ar1	**ar2**	**ar3**	**ar4**	**ar5**	**ma1**	**ma2**	**ma3**
Coef	-1.0898	0.0395	0.9085	0.2954	0.0478	0.6446	-0.5631	-0.9687
SE	0.0526	0.0819	0.0745	0.0791	0.0518	0.0228	0.0283	0.0234
p-value	0.0000	0.6292	0.0000	0.0002	0.3559	0.0000	0.0000	0.0000

Table 4: Résumé des termes AR et MA avec coefficients, erreurs standard et valeurs p pour ARIMA (5, 1, 3)

	ar1	ar2	ar3	ar4	ar5	ar6	ma1	ma2	ma3
Coef	-1.0890	0.0368	0.8922	0.2978	0.0749	0.0238	0.6429	-0.5659	-0.9705
SE	0.0524	0.0813	0.0811	0.0788	0.0780	0.0513	0.0231	0.0277	0.0240
p-value	0.0000	0.6510	0.0000	0.0002	0.3369	0.6434	0.0000	0.0000	0.0000

Table 5: Résumé des termes AR et MA avec coefficients, erreurs standard et valeurs p pour ARIMA (6, 1, 3)