Prédiction de la Production de Bière en Australie

Par :

Philippine RENAUDIN
Elvina EURY

Master 2 Mathématiques et Applications
Parcours Data Science

Sous la direction de M. Frédéric Proia

2020-2021

Introduction

Dans le cadre de cette étude nous disposons de données de production mensuelle de bière en Australie (en megalitre), de janvier 1956 à août 1995. L'objectif est de prédire la production de bière pour les trois années suivantes, c'est-à-dire du mois de septembre 1995 au mois de septembre 1998, en modélisant les données à l'aide de séries chronologiques. Dans un premier temps, nous analysons les données afin de déterminer une présence ou absence de tendance et de périodicité. Ensuite nous mettons en place plusieurs modèles qui nous semblent pertinents, et ce en se basant sur différentes caractéristiques des données. Enfin, nous comparons les performances de nos modèles sur un critère prédictif et ne gardons que les deux meilleurs. A partir de ces deux modèles, nous prédisons donc les valeurs pour les 36 mois suivants.

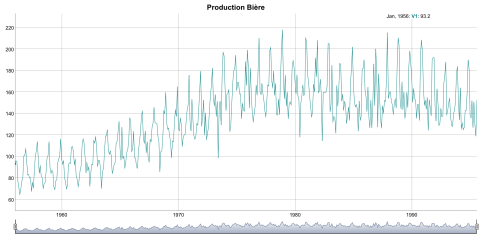
I Analyse des données et premières visualisations

I.1 Visualisation des données

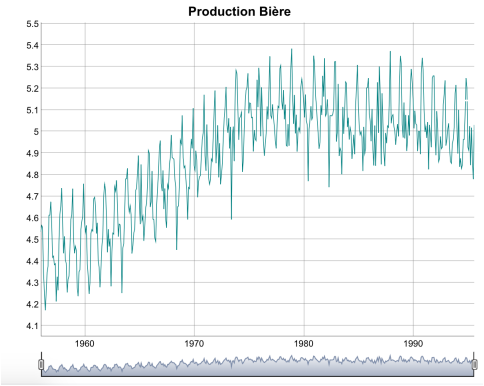
La modélisation par séries chronologiques peut s'écrire comme suit :
$X_t = m_t + S_t + Z_t$ où X_t est la valeur de la série à l'instant t.

m_t est la tendance et représente l'évolution de la série sur le long terme. S_t est la saisonnalité et représente l'évolution à court terme de la série. La saisonnalité traduit l'éventuelle périodicité des données. Z_t est la partie non déterministe du processus et est donc qualifiée de fluctuations. C'est le reste de l'information non expliquée par la tendance ou la saisonnalité. L'écriture de X_t ci dessus suppose que le modèle est additif. Or, il est tout à fait possible qu'il y ait des phénomènes d'amplification, auquel cas un modèle additif n'est pas adapté. On lui préférera un modèle multiplicatif de la forme : $X_t = m_t * S_t * Z_t$.

La Figure 1 représente la série (à gauche) et le logarithme de la série (à droite).
Une première observation visuelle de la série met en évidence une tendance (linéaire ou quadratique) ainsi qu'une périodicité (annuelle). De plus, l'amplitude des dernières périodes semble bien plus grande que celles des premières, ce qui nous pousse à également considérer le logarithme de la série dans la suite de notre étude.



(a) Série Initiale



(b) Série après passage au logarithme

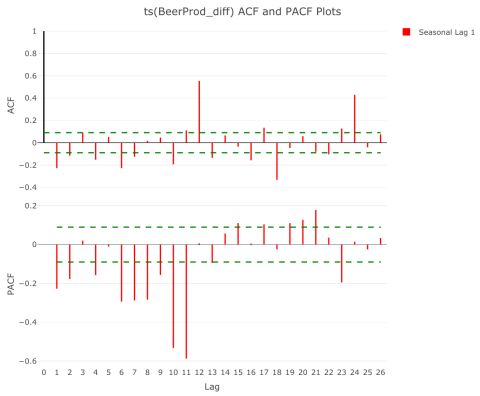
FIGURE 1 – Série initiale et série après transformation logarithmique

I.2 Etude de la stationnarité

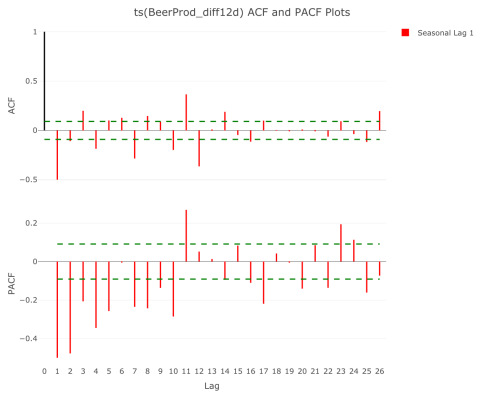
Nous travaillons donc maintenant en parallèle sur la série initiale et la série passée au logarithme. La première étape pour déterminer un modèle est d'étudier la stationnarité de la série. Les tests ADF et KPSS, ainsi que l'observation visuelle des deux séries considérées montrent qu'elles ne sont clairement pas stationnaires. Nous différencions alors plusieurs fois les séries jusqu'à obtention de la stationnarité.

Nous commençons par faire une différenciation locale sur les 2 séries (soit d=1). Les sorties ACF et PACF affichent des pics à 12, 24, et plus, nous poussant à considérer d'autres différenciations, et notamment des différenciations saisonnières (soit D=1).
Nous prenons ainsi les cas de différenciations suivants :

1. d=0, D=1
2. d=1, D=1



(a) Différenciation locale, $d = 1$



(b) Différenciation locale et saisonnière, $d = 1, D = 1$

FIGURE 2 – ACF et PACF après différenciations

Dans le premier cas, bien que la PACF ne comporte plus de pics saisonniers, l'ACF en comporte toujours et de plus la série est non-stationnaire d'après les tests ADF et KPSS. En revanche dans le deuxième cas (d=1,D=1), nous avons supprimé la périodicité et la série est stationnaire (tests ADF et KPSS).
Dans la suite on note $(X_t)_{t\in\mathbb{Z}}$ le processus initial, et $(XL_t)_{t\in\mathbb{Z}}$ le série transformée.

II Création de modèles

II.1 Série Initial, X_t

Nous nous focalisons en premier lieu sur la série non transformée X_t, puisque les approches mentionnées ci-dessous seront identiques pour la série XL_t. Nous utilisons 3 approches pour modéliser notre processus X_t :

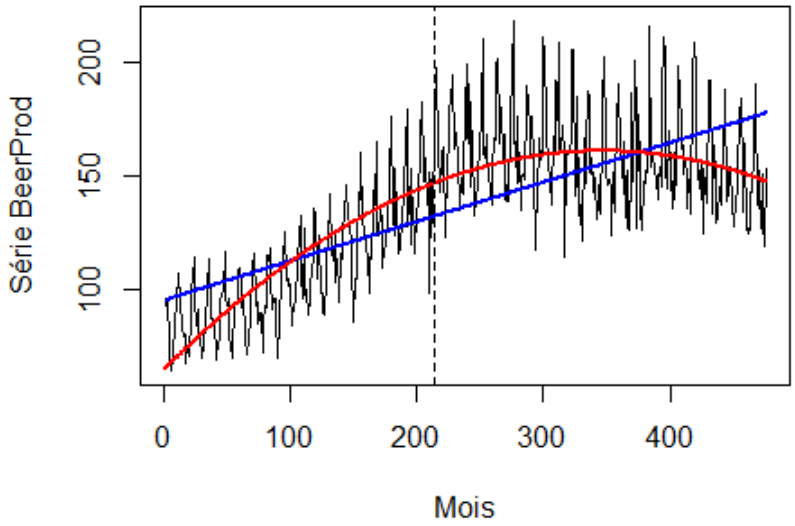


FIGURE 3 – Aperçu des différentes approches

a) Première approche

Dans un premier temps, nous considérons que la tendance observée sur les séries est linéaire. Nous cherchons donc grâce à la fonction auto.arima du package tseries de Rstudio les ordres p, q, P, Q optimaux, c'est à dire ceux permettant la minimisation du BIC (on impose d=1 et D=1 car ce sont ces ordres de différenciation qui éliminent la périodicité et donnent une série stationnaire).
Nous obtenons donc un premier modèle : $X_t \sim SARIMA(0,1,3)(0,1,2)_{12}$

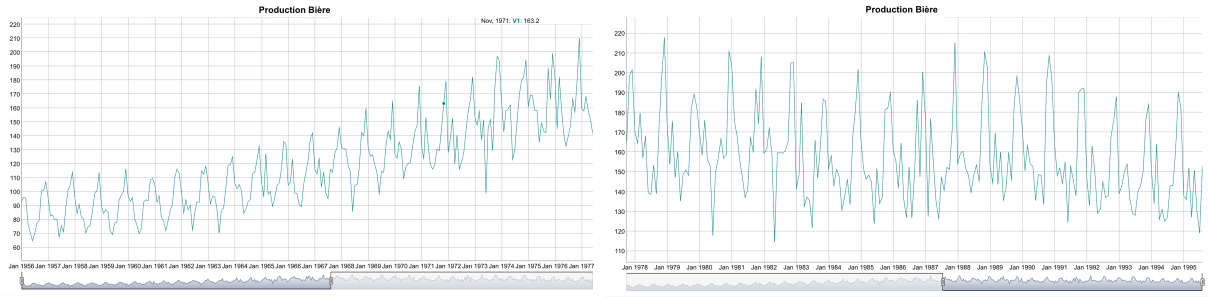
b) Deuxième approche

Dans un deuxième temps, nous essayons d'affiner ce modèle en ne considérant non plus que la tendance est linéaire mais quadratique. En effet, le début de la série est bien linéaire mais à partir d'un moment, la série semble arrêter de croître pour stagner ensuite, voire décroître.
Pour prendre en compte cette tendance quadratique, nous décidons donc d'effectuer une régression quadratique de X_t par rapport au temps et de modéliser les résidus de cette régression par un SARIMA, toujours à l'aide de la fonction auto.arima.
Nous obtenons donc pour les résidus : $res_t \sim SARIMA(2,0,2)(0,1,2)_{12}$

c) Troisième approche

Enfin, au lieu de considérer que la tendance est quadratique, nous considérons qu'il y a peut-être une rupture de la tendance à un certain moment et essayons donc de repérer cet éventuel point de rupture pour diviser la série en deux et ne travailler que sur la dernière partie de la série. En effet, un évènement à pu modifier la tendance et donc prendre en compte la première partie des données risquerait de biaiser les prédictions.

Nous déterminons par étude de l'erreur MSE que le point de rupture se trouve à l'abscisse 213 (c'est à dire en Septembre 1973). Une fois la série séparée en deux, nous ne considérons donc plus que la partie droite et modélisons cette série par un SARIMA grâce à la fonction auto.arima.
Nous obtenons donc un troisième modèle : $X2_t \sim SARIMA(0,0,0)(1,1,1)_{12}$



(a) Partie à gauche de la rupture, Série X_1 (b) Partie à droite de la rupture, Série X_2

FIGURE 4 – Séparation de la série au point de rupture

Avant d'aller plus loin dans la comparaison des modèles, il est important de valider la qualité des résidus . Or, nous remarquons que bien que la commande auto.arima ait sélectionné les modèles minimisant le BIC, elle ne semble pas considérer la qualité des résidus.
En effet, le test de Box-Jenkins nous montre qu'aucun des résidus des différents modèles ne peuvent être considérés blancs et les ACF et PACF présentent encore beaucoup de

pics significatifs. Il reste donc encore beaucoup d'information à aller chercher.

Afin de palier à ce problème, nous cherchons alors à modifier manuellement les valeurs de p, q ainsi que P et Q au risque d'avoir des estimateurs non significatifs. Une fois ces valeurs augmentées ou diminuées, nous regardons la significativité de chaque estimateur. Si un estimateur n'est pas significatif , nous l'enlevons du modèle et comparons les valeurs du BIC pour les modèles avec et sans. Nous gardons enfin le modèle qui minimise le BIC.

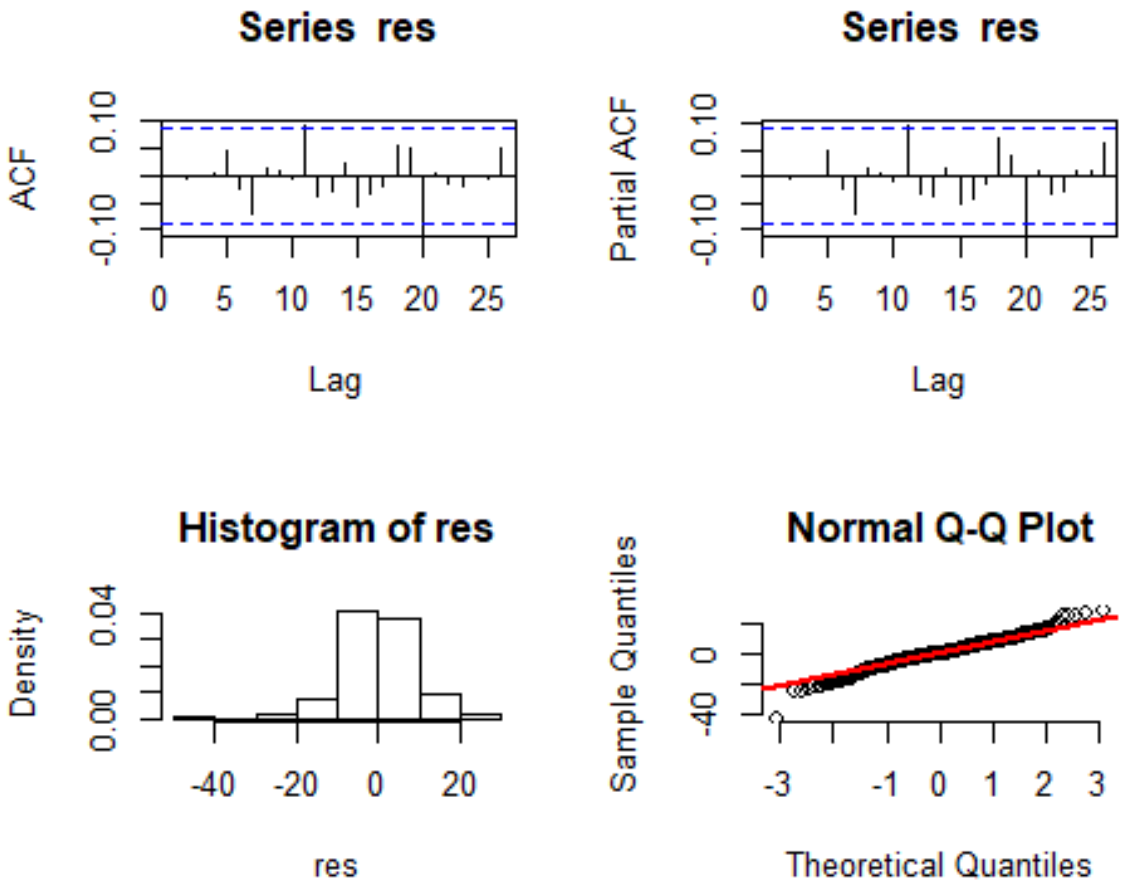


FIGURE 5 – Checkup des résidus obtenus après modifications des ordres

Nous obtenons donc finalement les modèles 'optimaux' suivants :

TABLE 1 – Modèles retenus-Série initiale

	Modèles
Mod1	SARIMA $(4,1,4)(0,1,1)_{12}$
Mod2	SARIMA $(4,0,4)(0,1,1)_{12}$ (sur les résidus)
Mod3	SARIMA $(4,1,3)(0,1,1)_{12}$ (rupture)

II.2 Série avec transformation logarithmique, XL_t

Nous procédons de la même façon avec la série transformée. Nous obtenons par les 3 mêmes approches et après ajustement des paramètres les modèles suivants :

TABLE 2 – Modèles retenus-Série transformée

	Modèles
ModL1	SARIMA $(3,1,4)(0,1,1)_{12}$
ModL2	SARIMA $(4,0,4)(0,1,1)_{12}$ (sur les résidus)
ModL3	SARIMA $(4,0,4)(0,1,1)_{12}$ (rupture)

III Comparaison des modèles et prédictions

III.1 Meilleurs modèles

Pour chaque série, nous comparons les 3 modèles obtenus sur la base d'un critère prédictif. Nous prédisons la dernière période de la série à l'aide des modèles tronqués et calculons l'erreur MSE pour chaque modèle (carré des écarts). Nous gardons le modèle qui minimise la MSE.
Les modèles retenus (Tables 3 et 4) sont donc les suivants :

Série X_t : Modèle 2 : SARIMA$(4,0,4)(0,1,1)_{12}$

TABLE 3 – MSE des modèles basés sur la série initiale

	Modèles	**MSE**
1	SARIMA $(4,1,4)(0,1,1)_{12}$	69.82345
2	SARIMA $(4,0,4)(0,1,1)_{12}$ (sur les résidus)	**66.83503**
3	SARIMA $(4,1,3)(0,1,1)_{12}$ (rupture)	77.87975

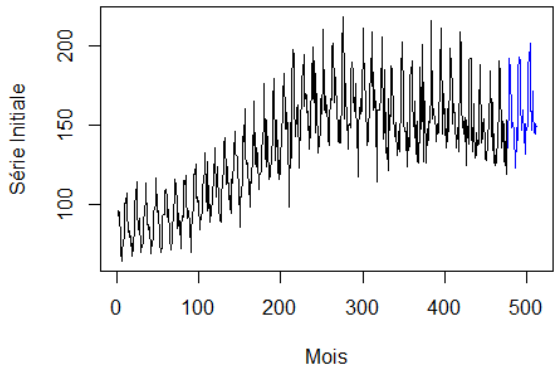
Série XL_t : Modèle L3 : SARIMA$(4,0,4)(0,1,1)_{12}$

TABLE 4 – MSE des modèles basés sur la série transformée

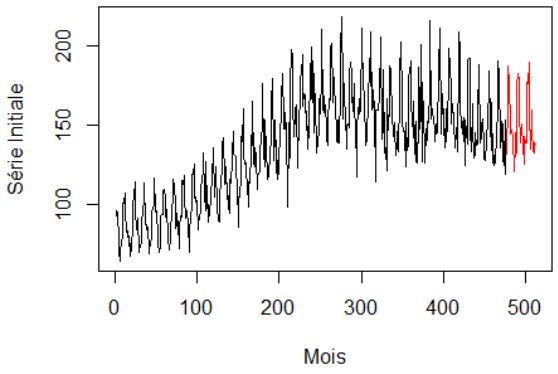
	Modèles	**MSE**
L1	SARIMA $(3,1,4)(0,1,1)_{12}$	0.003685328
L2	SARIMA $(4,0,4)(0,1,1)_{12}$ (sur les résidus)	0.003874816
L3	SARIMA $(4,0,4)(0,1,1)_{12}$ (rupture)	**0.003226944**

III.2 Prédictions

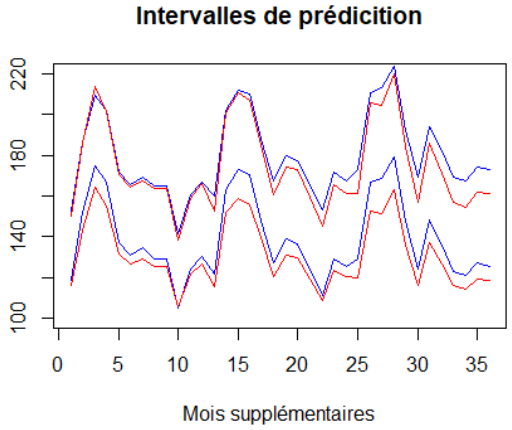
Une fois le meilleur modèle déterminé pour chaque séries, nous passons à l'étape de prédiction.
Les graphiques ci-dessous montrent les prédictions des deux modèles retenus, sur 3 ans. À gauche, la prédiction du modèle 2, à droite celle du modèle L3. Le dernier graphique représente les intervalles de prédiction de ces deux modèles.



(a) Prédiction avec le Modèle 2



(b) Prédiction avec le Modèle L3



(c) Intervalles de prédiction des deux modèles

FIGURE 6 – Prédictions sur 3 ans

Conclusion

Après une rapide étude visuelle de la série, nous avons pu déterminer trois différentes approches pour la modélisation de celle-ci. Dans chaque approche, nous étudions plus en détails les ACF et PACF des différentes différenciations effectuées. Nous étudions également la blancheur des résidus et la significativité des estimateurs des modèles renvoyés par la fonction auto.arima. Nous obtenons ainsi 6 modèles que nous comparons 3 à 3 sur un critère prédictif. Les erreurs MSE associées à chaque modèles nous permettent de déterminer un modèle optimal pour la série initiale et la série transformée. Nous prédisons ainsi avec ces deux modèles la production de bière en Australie pour les 3 années suivantes. Après une rapide étude visuelle des intervalles de prédiction nous serions tentées de dire que le modèle 2 est meilleur que le modèle L3 car ses intervalles sont plus 'petits'. Nous avons calculé l'écart moyen pour chaque intervalle et trouvons un écart plus faible pour le modèle 2 ce qui signifie que la prédiction est plus 'sûre' que celle du modèle L3. Nous choisissons donc de nous baser sur cet indicateur pour dire que le modèle 2 est le meilleur (parmi nos modèles) pour la modélisation de ce jeu de données.