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 gigalitre), 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

Les séries chronologiques ont 3 composantes principales : La tendance (T), qui décrit les mouvement à long terme, la périodicité (S), qui est cyclique, et les fluctuations (ϵ).

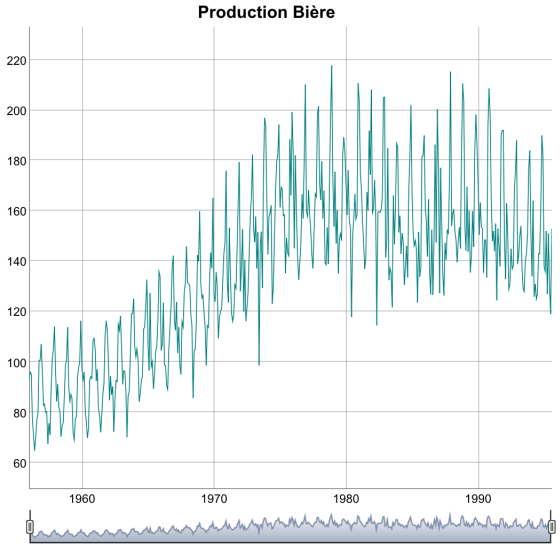
La modélisation par séries chronologiques peut s'écrire comme suit : $X_t = m_t + S_t + Z_t$ où X_t est la valeur de la série à l'instant t. mt est la tendance et représente l'évolution de la série sur le long terme. S_t est la saisonnalité et représente l'évolution à court terme de la série. La saisonnalité traduit l'éventuelle périodicité des données. Z_t est la partie non déterministe du processus et est donc qualifié 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 : $X_t = m_t * S_t * Z_t$.

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.

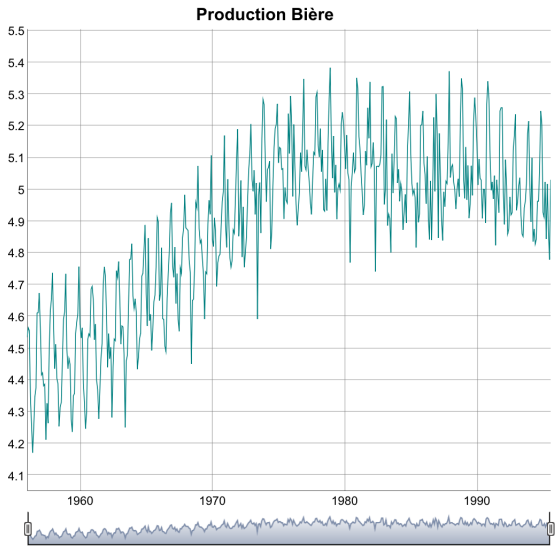
I.2 Éliminer la stationnarité

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 d'étude. Les tests ADF et KPSS, ainsi que l'observation visuelle de nos deux séries



(a) Série Initiale



(b) Série après passage au logarithme

FIGURE 1 – Comparaison série initiale et avec transformation logarithmique

d'études montrent qu'elles ne sont clairement pas stationnaires. Nous différencions alors plusieurs fois nos séries et étudions les acf et pacf des nouvelles séries obtenues, ainsi que leur stationnarité. À partir de ces deux modèles, nous commençons par faire une différenciation $(I - B)$, soit d=1. Les sorties ACF et PACF nous affiche des pics à 12, 24, et plus, nous poussant à considérer d'autres différentiations, et notamment des différentiations saisonnières, $(I - B^{12})$, soit D=1. Nous prenons ainsi les cas de différentiations suivants :

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

Pour le premier cas, la série est non-stationnaire, tandis que lorsque nous prenons d=1 et D=1, nous obtenons une série stationnaire, un résultat qui est vu à la fois via les 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.

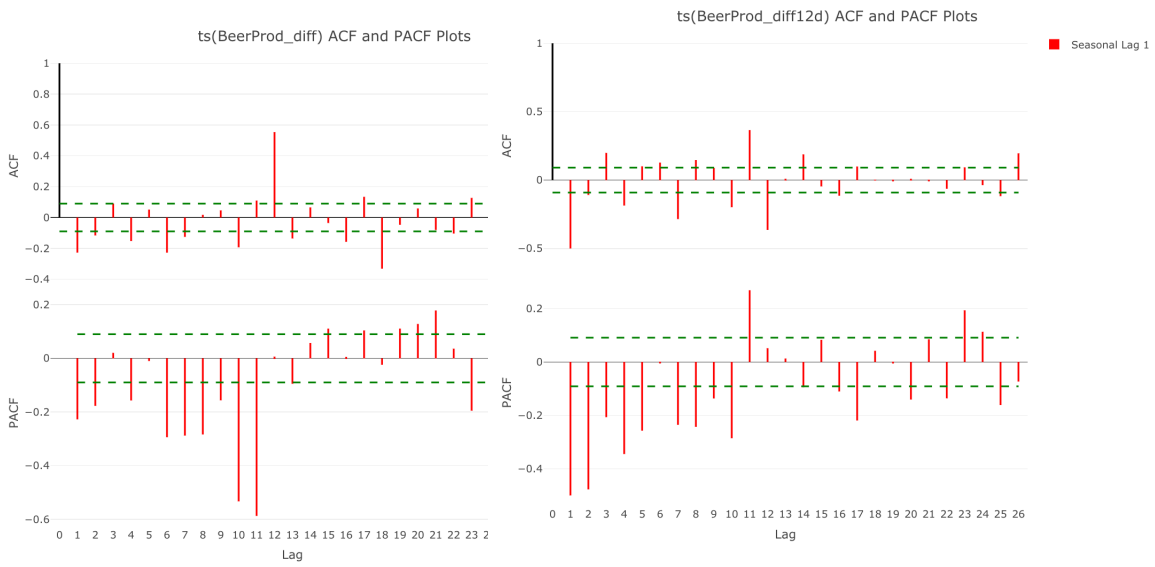
II Création de modèles

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 :

a) Dans un premier temps, nous considérons que la tendance observée sur la série X_t 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 ici) Nous obtenons donc un premier modèle avec p=0, q=3, P=0, Q=2. On a donc : $X_t \sim SARIMA(0,1,3)(0,1,2)_{12}$



(a) Une différentiation locale, $d = 1$

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

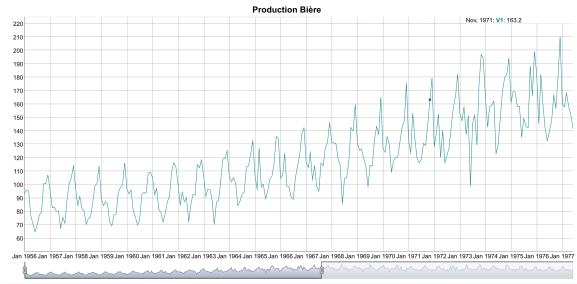
FIGURE 2 – ACF et PACF après une différentiation locale, $d = 1$

b) 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, on ne peut plus dire que ça croit, ça aurait même tendance à décroitre. Pour prendre en compte cette tendance quadratique, nous décidons donc d'effectuer une régression quadratique de X_t et de modéliser les résidus de cette régression par un SARIMA, toujours à l'aide de la fonction auto.arima.

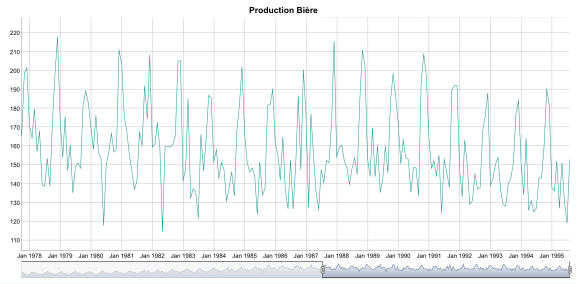
c) Enfin, au lieu de considérer que la tendance est quadratique, nous considérons qu'il y a peut-être une rupture de la tendance à un certain moment et essayons donc de repérer cet éventuel point de rupture pour diviser la série en deux et ne travailler que sur la deuxième partie de la série. En effet, 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 ce trouve à l'abscisse 213. 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 les ordres suivant : $p = 0, d = 0, q = 0, P = 1, D = 1, Q = 1$ donc $X2_t \sim SARIMA(0,0,0)(1,1,1)_{12}$

Avant d'aller plus loin dans la comparaison des modèles, il est important de valider la qualité des résidus des modèles créés. Or, nous remarquons que bien qu'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 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 augmenter manuellement les valeurs de p, q ainsi P et Q au risque



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



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

FIGURE 3 – Analyse point de rupture

d'avoir des paramètres non significatifs. Ainsi, pour chaque modèle créé nous effectuons un test statistique de student sur chaque coefficient. Dans chaque cas où les coefficients les plus (élevé ?– A VALIDER QUEL TERME METTRE) sont non significatifs, une variante du modèle, ne possédant aucune variable significative est créée. Le BIC de cette dernière est comparé à celui du modèle possédant les paramètres non significatifs. Nous obtenons avec les modèles suivants :

METTRE PHOTOS CHECK UP RES du modèle avec meilleur résultat pour log et non-log

TABLE 1 – Modèles retenus

	Modèles	**Equations**
1	SARIMA $(4,1,4)(0,1,1)_{12}$	α - à compléter
2	SARIMA $(4,0,4)(0,1,1)_{12}$	β - à compléter
3	SARIMA avec rupture $(4,1,3)(0,1,1)_{12}$	γ - à compléter

Série avec transformation logarithmique, YL_t)

Nous procédons de la même façon avec le modèle logarithmique.

Nous obtenons par les 3 mêmes approches et après ajustement des paramètres les modèles suivants :

TABLE 2 – Modèles retenus

	Modèles	**Equations**
L1	SARIMA $(3,1,4)(0,1,1)_{12}$	α - à compléter
L2	SARIMA xreg $(4,0,4)(0,1,1)_{12}$	β - à compléter
L3	SARIMA avec rupture $(4,0,4)(0,1,1)_{12}$	γ - à compléter

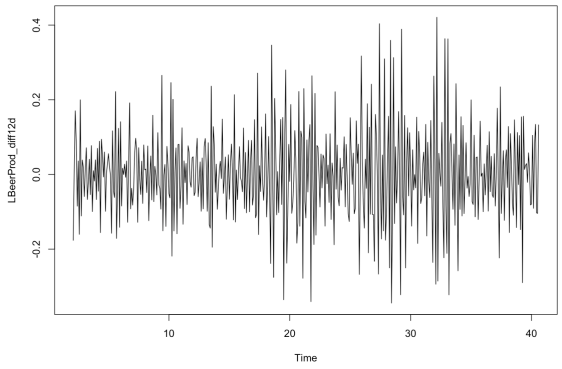


FIGURE 4 – Séries logarithmique avec des différenciations, d=1, D=1

II.1 Comparaison des modèles et prédictions

Meilleurs modèles

Pour chaque série, nous comparons les 3 modèles obtenus sur la base d'un critère prédictif. Nous prédisons à nouveau 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 retenues (Tables 2 et 3) sont donc les suivants :

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

TABLE 3 – MSE à partir de la série initiale

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

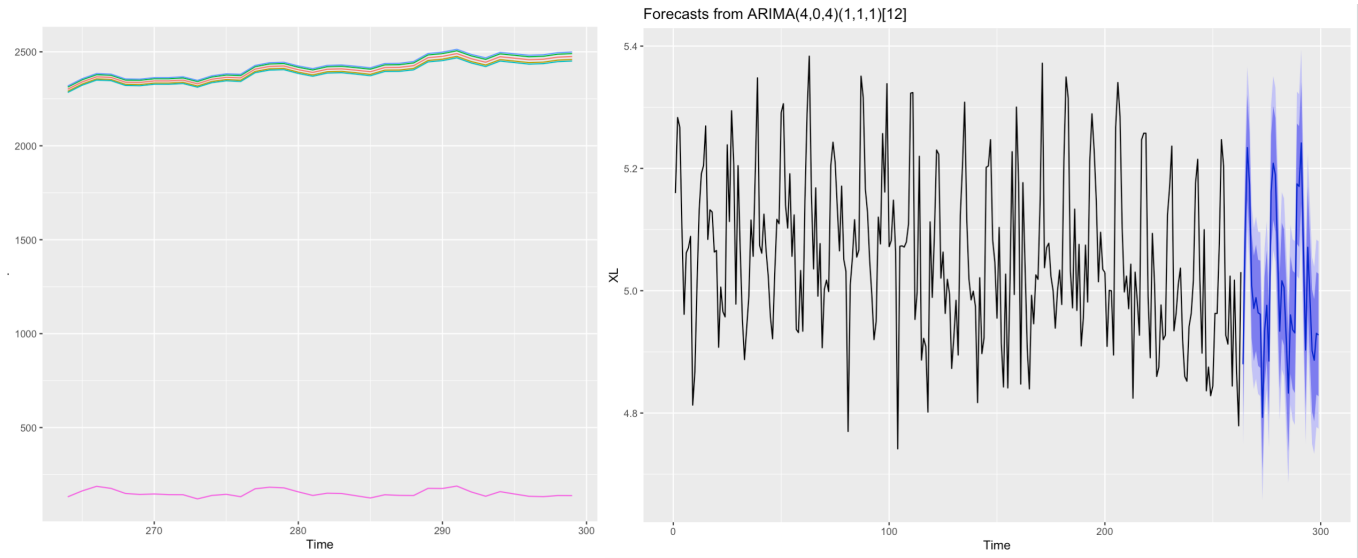
Série XL_t : Modèle L3 : SARIMA - A AJOUTER

TABLE 4 – MSE à partir de la série transformée

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

Les graphiques ci-dessous montrent les prédictions des deux modèles retenus sur 3 ans. À droite, nous voyons la prédiction du modèle L3, et à gauche celle du modèle 2 (en comparaison avec l'exponentiel du la prédiction de L3).

JE NARRIVE PAS À FAIRE AFFICHER LES GRAPHIQUE PREDICTION - VOICI CE QUE JE SORS :



(a) Modèle 2 avec intervalle de confiance

(b) Modèle L3

FIGURE 5 – Prédiction sur 3 ans

II.2 Conclusion

Le modèle retenu, soit celui avec le meilleur critère prédictif, est le modèle 2.

(a) Modèle 2 avec intervalle de confiance

(b) Modèle L3 avec intervalle de confiance

FIGURE 6 – Prédiction sur 3 ans