



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 gegalitre), 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 mouvements 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 . 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 : $X_t = m_t * S_t * Z_t$.

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 Différentes différentiations –MEILLEUR TITRE ?

Nous travaillons donc maintenant en parallèle sur la série initiale et la série passée au logarithme.

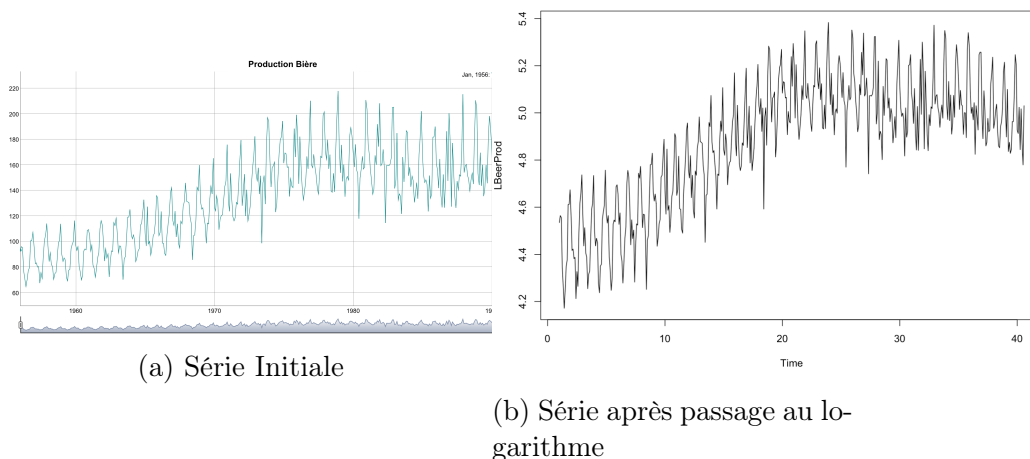


FIGURE 1 – Comparaison série initiale et avec transformation logarithmique

Différentiation locale ($d = 1$)

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 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 différenciations, et notamment des différenciations saisonnières, $(I - B^{12})$, soit $D=1$. Nous prenons ainsi les cas de différenciations suivants :

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

AJOUTER GRAPHIQUE : ACF, PACF - Petit $d = 1$

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 .

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 : $X_t(AMODIFIERPOUR TILDE)SARIMA(0, 1, 3)(0, 1, 2)_{12}$

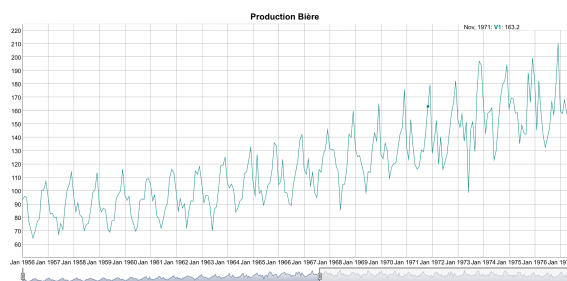
$\theta....$ ECRIRELEMODELE 

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 croît, ça aurait même tendance à décroître. 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`.

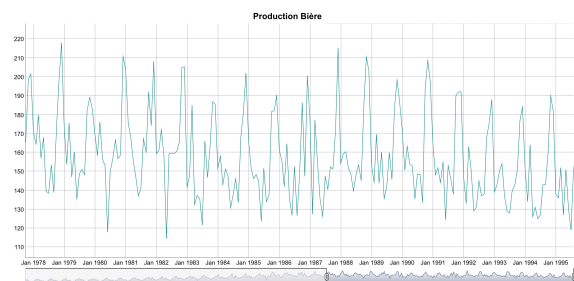
$\theta....$ ECRIRELEMODELE

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 : $p = 0, d = 0, q = 0, P = 1, D = 1, Q = 0$ donc $X2_t SARIMA(0, 0, 0)x(1, 1, 1)[12]$ 




(a) Partie à gauche de la rupture, Série X1



(b) Partie à droite de la rupture, Série X2

FIGURE 2 – Analyse point de rupture

$\theta....$ ECRIRELEMODELE

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 **qui minimise** le BIC, **il** ne semble pas considérer la qualité des résidus. En effet, le test de Box Jenkins nous montre que les résidus ne **sont pas bruits blancs** et qu'au final les modèles ne sont pas adaptés à faire les prévisions. Afin de palier 

à ce problème nous cherchons alors à augmenter manuellement les valeurs de p, q ainsi P et Q ce qui nous donnerait plus de flexibilité dans le modèle mais qui augmenterait le risque de 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 (ELEVER – 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 nous retrouvons avec les modèles suivants :

TABLE 1 – Modèles choisis

	Modèles	Equations
A	SARIMA (4, 1, 4)(0, 1, 1) ₁₂	α - à compléter
B	SARIMAX (4, 0, 4)(0, 1, 1) ₁₂	β - à compléter
C	SARIMA avec rupture (4, 1, 3)(0, 1, 1) ₁₂	γ - à compléter

Série avec transformation logarithmique, Y_{L_t})

Nous procédons de la même façon avec le modèle logarithmique. Tout comme pour la série initiale, en faisant une différenciation, $(I - B)$ et en faisant une différenciation saisonnière $(I - B^{12})$, nous obtenons une modèle stationnaire. Le test ADF et KPSS confirme ce résultat tout comme les sorties ACF et PACF (qui nous n'affiche plus de pics cycliques).

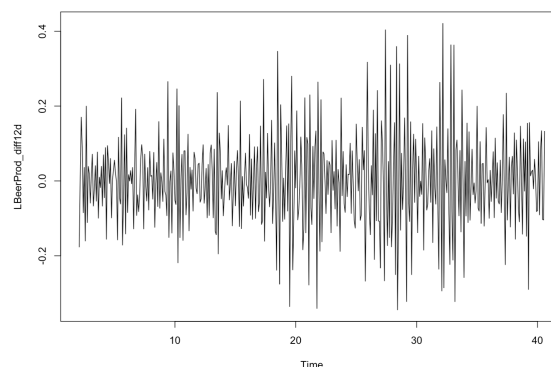


FIGURE 3 – Séries logarithmique avec des différenciations, $d=1$, $D=1$

Dans la Figure 2, nous remarquons que la variance n'est pas tout à fait constante. En effet, elle est plus élevée autours des périodes 20 et 30.

II.1 Analyses des résidus

Avant de faire des prédictions avec nos modèles nous cherchons à analyser les résidus. Pour cela nous allons faire à la fois une analyse visuelle et une analyse en se basant sur

les tests statistiques afin de voir si nos résidus sont bien des bruits blanc et s'ils sont gaussiens.



Modèle Initial, Y_t

METTRE PHOTOS CHECK UP RES du modèle avec meilleur résultat

Modèle logarithme, Y_{L_t}

METTRE PHOTOS CHECK UP RES du modèle avec meilleur résultat

II.2 Prédictions

Nous cherchons maintenant à analyser la performance prédictive des modèles sur 1 an ATTENTION PREDICTION 3 ANS demandé dans le mail ?. Pour le faire, une période, équivalente à 12 mois, est enlevée à partir d'août 1995. Les résultats ainsi prédits seront comparés avec les observations en utilisant le MSE, l'erreur moyenne au carré. Le modèle ayant le plus petit MSE sera le modèle sélectionné dans chacun des cas (avec et sans transformation logarithme).



Modèle Initial, Y_t

Figure 4 nous affiche à la fois les prédictions des 12 derniers mois et les vrais valeurs (en noir). Nous remarquons que les 3 modèles donnent visuellement des résultats assez proche des vrais observations. Nous confirmons cela à l'aide des MSE calculés (Table 2). Nous pouvoir établir que le modèle

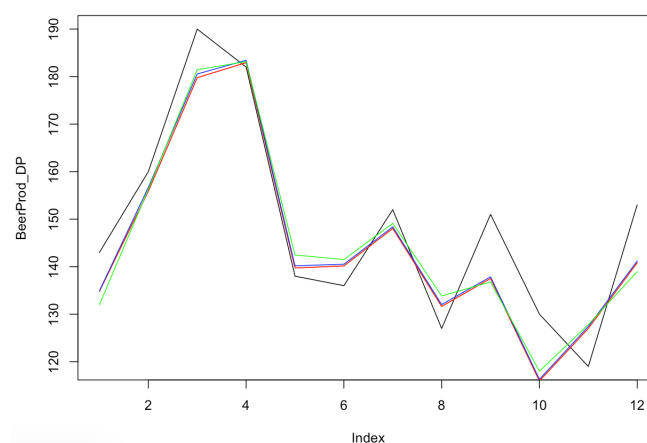


FIGURE 4 – Prédictions des 12 derniers mois et comparaison avec les vrai valeurs à partir de la série non-transformée

TABLE 2 – MSE

	Modèles	MSE
A	SARIMA (4, 1, 4)(0, 1, 1) ₁₂	69.82345
B	SARIMAX (4, 0, 4)(0, 1, 1) ₁₂	66.83503
C	SARIMA avec rupture (4, 1, 3)(0, 1, 1) ₁₂	77.87975



Modèle logarithme, YL_t

Les résultats obtenus à partir des logarithmes de la séries sont affichés dans la figure 5.

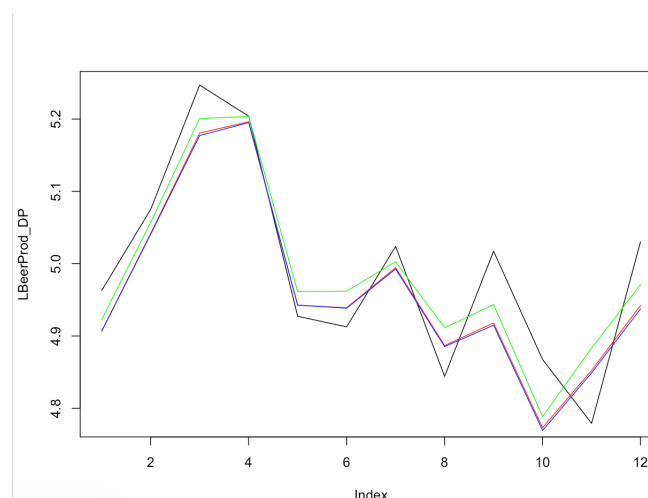


FIGURE 5 – Prédictions des 12 derniers mois et comparaison avec les vrai valeurs à partir de la série transformée

TABLE 3 – MSE

	Modèles	MSE
D	SARIMA A AJOUTER	0.003685328
E	SARIMAX A AJOUTER	0.003874816
F	SARIMA avec rupture A AJOUTER	0.003226944



II.3 Conclusion