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

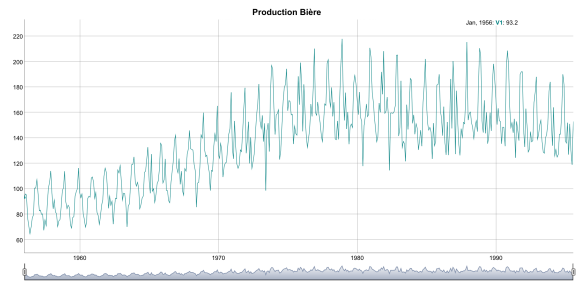
L'objectif de cet étude est de prédire la production de bière en Australie pour les trois prochaines années à partir de séries chronologiques. La production de bière est une variable endogène car elle est dépendante de d'autres facteurs comme le climat. Nous nous basons sur des données mensuelles, de janvier 1956 à août 1995. Comme il est difficile de visuellement discerner le modèle multiplicatif de l'additif nous analysons à la fois le modèle initial et le modèle transformée. Notre approche de vise en premier lieu à transformer les données, à éliminer la périodicité et la tendance avant de rendre le processus stationnaire. Puis des tests statistiques sont effectués sur les résidus afin de cibler les modèles les plus robustes. Finalement, nous utilisons les modèles choisis afin de prédire la production de bière sur les trois prochaines années. A COMPLETER/ MODIFIER...

Méthologie

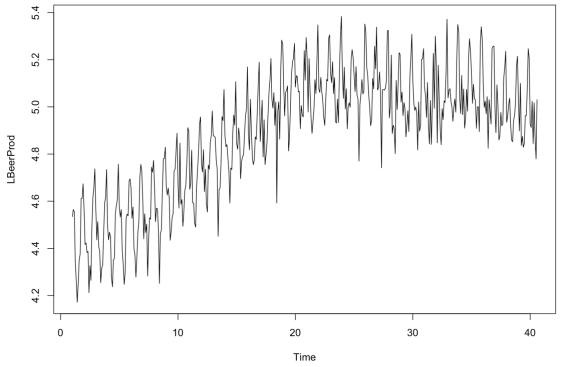
.1 Analyse initiale

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). Nous débutons notre approche en faisant une analyse visuelle des séries chronologiques.

À première vue nous voyons que nos séries sont non-stationnaire avec une tendance et une périodicité annuelle. Cette dernière est d'ailleurs confirmé sur les sorties ACF et PACF des séries initiales. Il est dur à dire si la tendance est quadratique ou linéaire. Nous nous baserons sur des tests supplémentaires pour être sûr. De plus, nous remarquons un accroissement subtile de la saisonnalité, nous laissant un doute quant à la nature multiplicative ou additive du modèle. Nous avons ainsi décidé de comparer les résultats obtenus des séries initiales, avec les résultats du logarithme des séries.



(a) Série Initiale



(b) Série avec une transformation logarithme

FIGURE 1 – Comparaison série initiale et avec transformation logarithmique

.2 Élimination de la tendance et de la saisonnalité –MEILLEUR TITRE ?

À partir de ces deux modèles, nous commencons 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 faire une prochaine différenciation, $(I - B^{12})$, soit D=1. Nous prenons ainsi les cas de différenciation 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 sorties ACF, PACF et à la fois grâce aux tests ADF et KPSS. Nous établirons nos prochains modèles à partir de Y_t (cas non-transformé) et YL_t (cas transformé)
Soit $(X_t)_{t\in\mathbb{Z}}$ le processus initial

$$Y_t = (I - B) * (1 - B^{12}) * X_t$$
$$YL_t = (I - B) * (1 - B^{12}) * log(X_t)$$

.3 Création de modèles

Série Initial, Y_t)

Nous utilisons 3 approches pour modéliser notre processus Y_t.

La première est une approche standard, qui permet de déterminer les p, q et P, Q qui minimiseraient le BIC. En effet, grâce à la fonction auto.arima de R nous obtenons en premier lieu SARIMA (0,1,3) X (0,1,2)[12]. Cela veut dire que notre modèle est de la forme suivante :

$$\theta....ECRIRELEMODELE$$

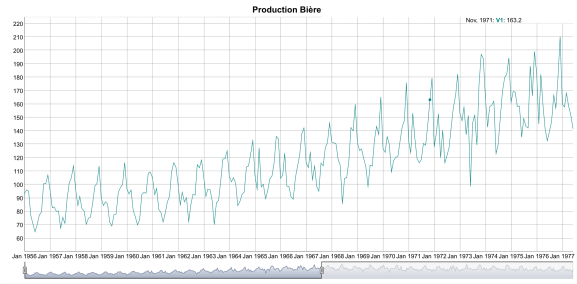
Puis nous tentons de développer un autre modèle potentiel. Comme notre série semble avoir une tendance quadratique, de part sa forme concave, nous considérons une nouvelle approche qui pourrait traiter la possibilité d'endogénéité de la variable production de bière. Ainsi nous supposons que la variable pourrait être dépendante de d'autres facteurs tel que le climat. Nous créons alors le modèle SARIMAX qui est de la forme suivante :

$$\theta....ECRIRELEMODELE$$

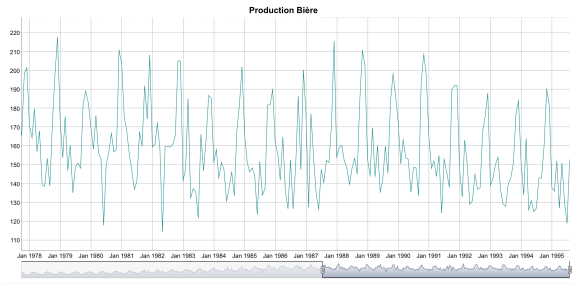
La dernière approche utilisée dans cet étude est la recherche du point de rupture.

Tout comme pour la deuxième approche nous choisissons cette méthode à cause de la tendance de la série. Nous voyons en effet, un point de rupture possible autour de 1970. Le modèle choisit par auto.arima est :

$$\theta....ECRIRELEMODELE$$



(a) Partie à gauche de la rupture



(b) Partie à droite de la rupture

FIGURE 2 – Analyse point de rupture

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**
1	SARIMA $(4,1,4)(0,1,1)_{12}$	α - à compléter
2	SARIMAX $(4,0,4)(0,1,1)_{12}$	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. 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).

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.

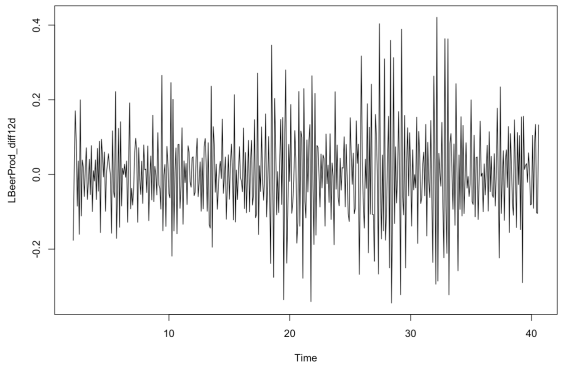


FIGURE 3 – Séries logarithmique avec des différenciations, d=1, D=1

.4 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 RESULTAT CONCLUANT

Modèle Initial, YL_t

METTRE PHOTOS CHECK UP RES DU RESULTAT CONCLUANT

.5 Prédictions

.6 Conclusion