FACULTÉ DES SCIENCES
UNIVERSITÉ D'ANGERS

Université d'Angers

Master Mathématiques et Applications

M2 DATA SCIENCE

Année académique 2021-2022

Rapport de Séries Chronologiques

Production mensuelle de bière en Australie entre 1956 et 1995

Etudiants :
Badreau Marie
Belnou Alexandre

Enseignants :
Okome Obiang Eunice
Proia Frédéric

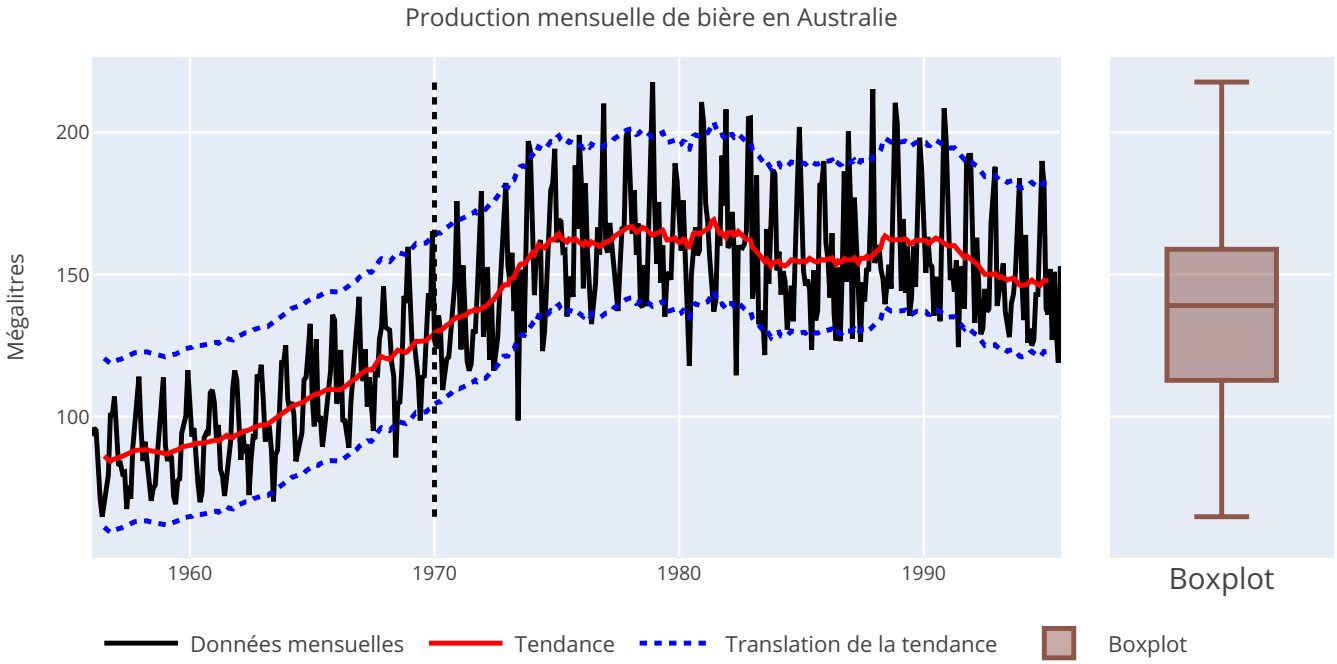
Sommaire

Contents

1. Introduction

1.1. Présentation des données

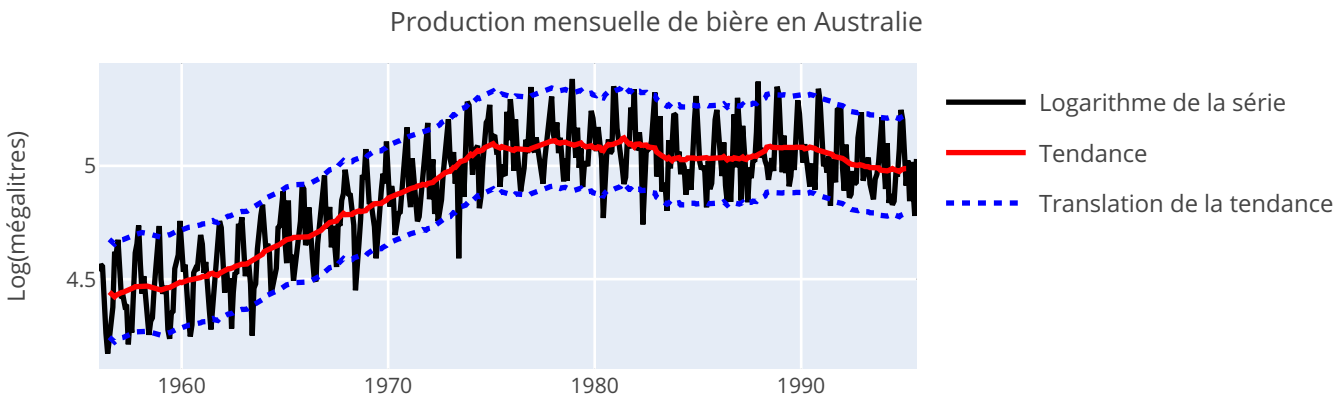
Les données présentées dans notre rapport concernent la production mensuelle de bière (en mégalitres) en Australie entre janvier 1956 et août 1995 compris, soit $n = 476$ observations. La série provient d'une base de données en libre accès sur le site https://www.kaggle.com et n'a pas de valeur manquante.



La représentation graphique de la série nous montre des données d'allure périodique, à tendance croissante jusqu'en 1975 environ, puis la tendance se stabilise, oscillant entre 114.6 et 217.8 mégalitres. La valeur minimale de la série est atteinte en juin 1956 avec une production de 64.8 mégalitres de bière, contre une valeur maximale de 217.8 mégalitres produits en décembre 1978. On note par ailleurs une moyenne globale de la série autour des 136 mégalitres.

On peut supposer de manière raisonnable que la période de cette série correspond à une année soit 12 mois (ou observations). Nous reviendrons plus en détail sur le motif saisonnier dans la partie suivante.

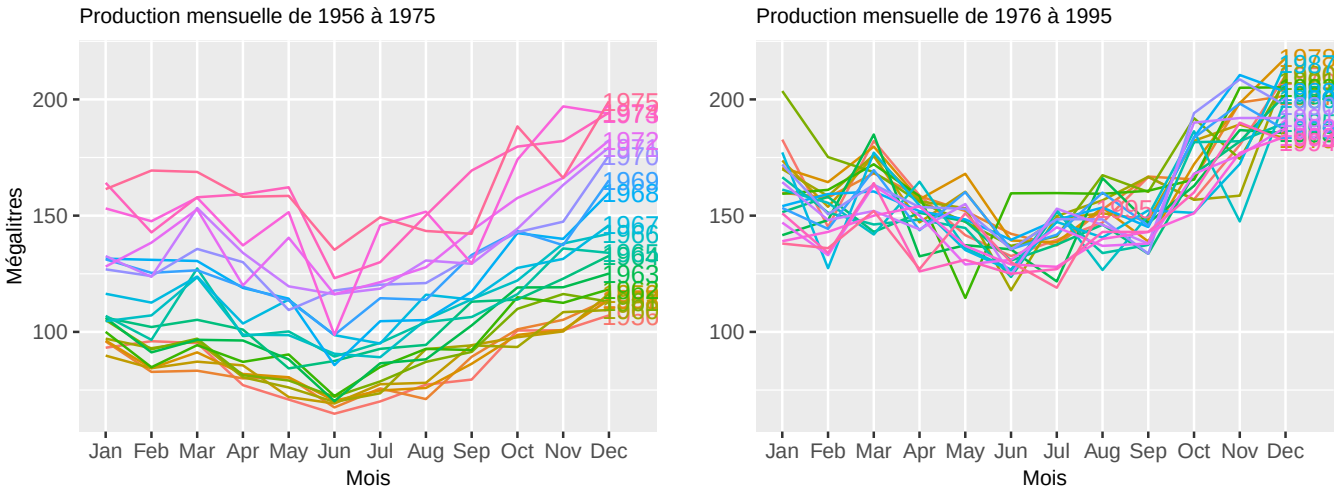
On remarque également une augmentation de la variance et de l'amplitude des motifs périodiques à partir des années 70 ; la courbe rouge représente l'estimation de la tendance de la série à partir de la fonction *decompose*, les courbes bleues correspondent à cette tendance à laquelle on a ajouté la valeur maximale (34.42) et minimale (-25.05) du motif périodique moyen estimé par cette même fonction. Ces différences d'amplitudes ne semblent pas être dues à des valeurs aberrantes, comme nous le montre le boxplot ci-dessus. Le graphique ci-dessous correspond à la représentation du logarithme des données associé aux différentes courbes présentées précédemment.



On voit que les amplitudes des motifs saisonniers sont plus homogènes sur toute la série dans ce deuxième graphique. Sauf mention contraire, nous travaillerons donc sur le logarithme des données dans la suite de ce rapport.

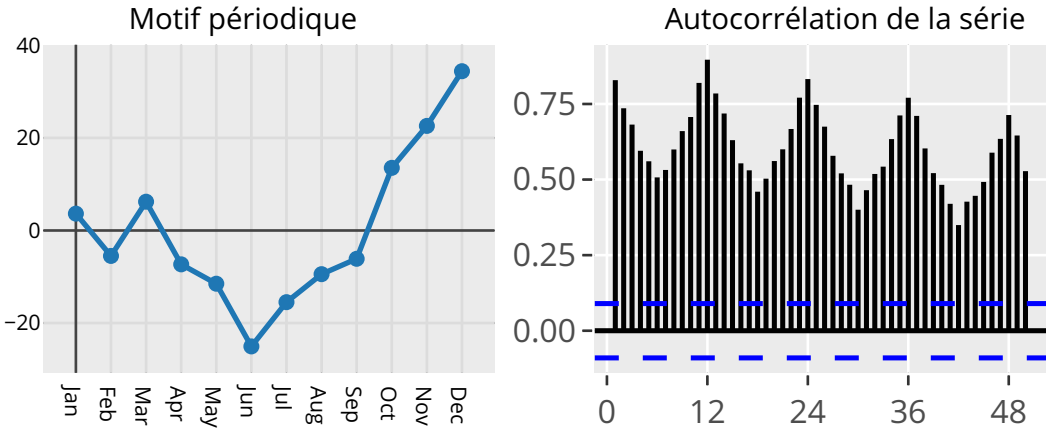
1.2. Étude du motif périodique

L'observation de données mensuelles nous laisse présager une périodicité annuelle. Nous pouvons étudier d'une part les données année par année :



Nous observons globalement une décroissance de la production sur la première moitié de l'année jusqu'en juin (excepté un pic de croissance au mois de mars) puis un forte croissance jusqu'à la fin de l'année. Ces observations se retrouvent dans le motif périodique estimé par la fonction *decompose* avec une périodicité de 12 mois.

D'autre part, l'étude de l'autocovariance de la série nous montre clairement une périodicité de fréquence 12.



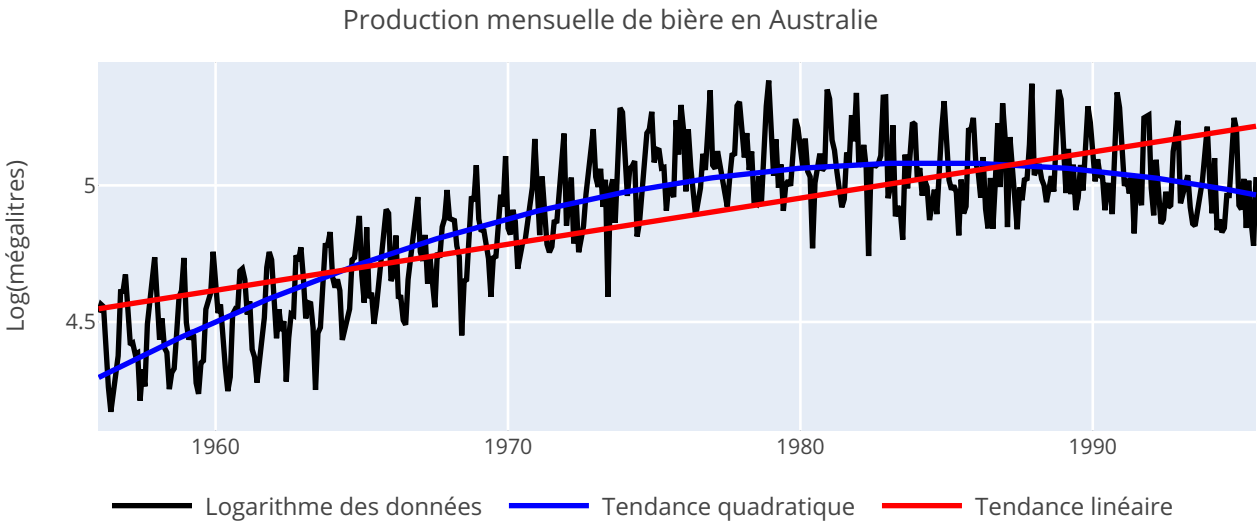
1.3. Tendance générale de la série

Revenons sur l'étude globale de la série, après passage au logarithme. La tendance générale des données peut être approximée par une régression quadratique (1) ou, plus simplement, linéaire (2) en fonction du temps. Les modèles précédents sont décrits par les formules suivantes :

$$Lmbp = \beta_0 + \beta_1 * temps + \beta_2 * temps_2 \quad (1)$$

$$Lmbp = \alpha_0 + \alpha_1 * temps \quad (2)$$

Lmbp représente ici le logarithme de nos données (*mbp* pour *Monthly Beer Production*), *temps* et $temps_2$ faisant respectivement référence au temps d'étude des données et au carré de ce vecteur temps.



On voit graphiquement que la régression linéaire n'est pas très adaptée à nos données tandis que le premier modèle (bleu) semble assez proche de la tendance générale. Les tests statistiques effectués dans le cadre d'une régression évaluent des hypothèses de significativité des différents coefficients :

- Significativité globale du modèle quadratique par la F-statistique :

$$\mathcal{H}_0 : "\beta_1 = \beta_2 = 0" \text{ vs } \mathcal{H}_1 : "\beta_1 \neq 0 \text{ ou } \beta_2 \neq 0"$$

- Significativité du cœfficient $j = 1, 2$:

$$\mathcal{H}_0 : "\beta_j = 0" \text{ vs } \mathcal{H}_1 : "\beta_j \neq 0"$$

Observons les résultats de ces tests sur nos différentes régressions :

Table 1: Coefficients de la régression quadratique (1)

	Estimate	Std. Error	t value	Pr(>|t|)
β_0	4.292196e+00	1.9465e-02	2.205085e+02	0e+00
β_1	4.603400e-03	1.8850e-04	2.442686e+01	0e+00
β_2	-6.700000e-06	4.0000e-07	-1.751659e+01	0e+00

Table 2: Test de Fisher et R^2 des régressions (1) et (2)

	statistic	p.value	r.squared	adj.r.squared
Reg. Quadratique	600.902	0	0.718	0.716
Reg. Linéaire	543.987	0	0.534	0.533

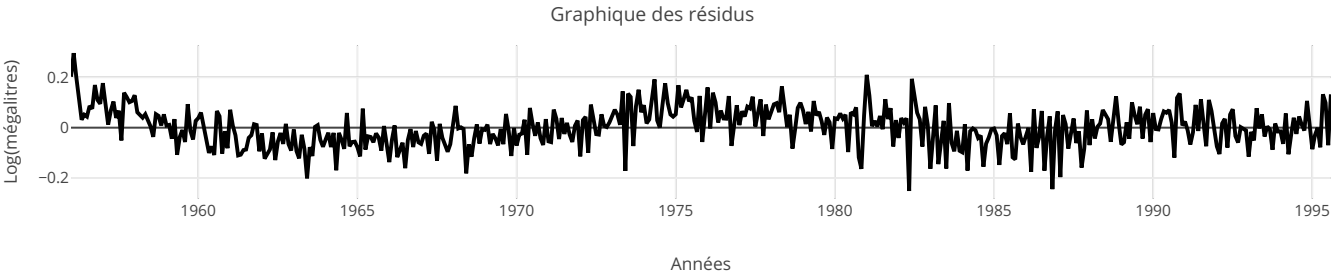
La régression quadratique est globalement significative, la p-value du test de Fisher étant largement inférieure à 5%. De même, les coefficients estimés par les deux régressions sont tous significatifs au seuil de 5% par rejet de l'hypothèse de nullité des coefficients. On note tout de même une estimation très proche de zéro pour les coefficients, mais rappelons que nous nous trouvons à l'échelle logarithmique, ce qui explique ces faibles valeurs. Le R^2 ajusté est nettement meilleur pour la régression quadratique que pour la régression linéaire, ce qui confirme l'impression visuelle de meilleure adéquation aux données.

L'analyse de variances sur les deux modèles nous donne une p-value égale à $2.5823044 \times 10^{-53}$ correspondant au test de Fisher des modèles emboîtés :

$$\mathcal{H}_0 : "\beta_2 = 0" \text{ vs } \mathcal{H}_1 : "\beta_2 \neq 0"$$

Nous pouvons donc rejeter fermement l'hypothèse nulle : le terme $temps_2$ est bien significatif. En conclusion, le modèle quadratique semble bien meilleur que la régression linaire simple.

Une première idée de modélisation serait donc d'associer le motif périodique à notre régression quadratique et envisager un modèle ARMA sur les résidus de cette régression, représentés ci-dessous :



Graphiquement, les résidus ne semblent clairement pas stationnaires : la moyenne n'est pas constante au cours du temps. De plus, la régression quadratique suggère une tendance inéluctable à la baisse, ce qui ne semble pas correspondre à la réalité : la production de bière en Australie ne tends pas vers un arrêt progressif de toute activité.

Nous pourrions envisager de différencier nos résidus pour étudier l'éventualité d'une modélisation par un SARIMA mais il semble plus raisonnable d'envisager que notre série présente une rupture dans les années 70, date à partir de laquelle la tendance globale de nos données semble osciller entre baisse et augmentation de la production annuelle. Nous allons donc chercher à estimer cette date de rupture afin d'ajuster de meilleurs modèles à nos données.

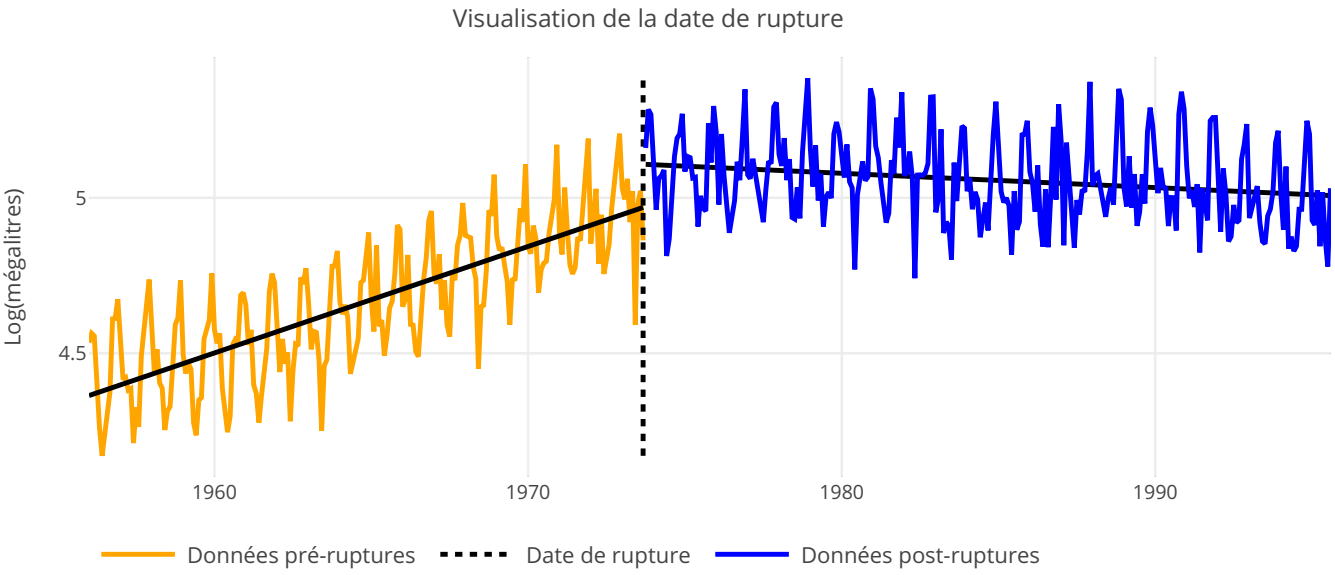
1.4. Date de rupture

Afin de déterminer la date de rupture, nous allons séparer notre série en deux parties : la première contiendra les observations d'indices compris entre 1 et R, avec R un entier à valeur dans $[1, n]$, et la seconde partie contiendra le reste des valeurs (soit celles comprises entre $R+1$ et n, on rappelle que $n = 476$). On estime ensuite un modèle de régression linéaire sur chacune des deux parties puis on calcule la MSE (*Mean Squared Error : Erreur quadratique moyenne*) induite par cette modélisation et dont voici la formule :

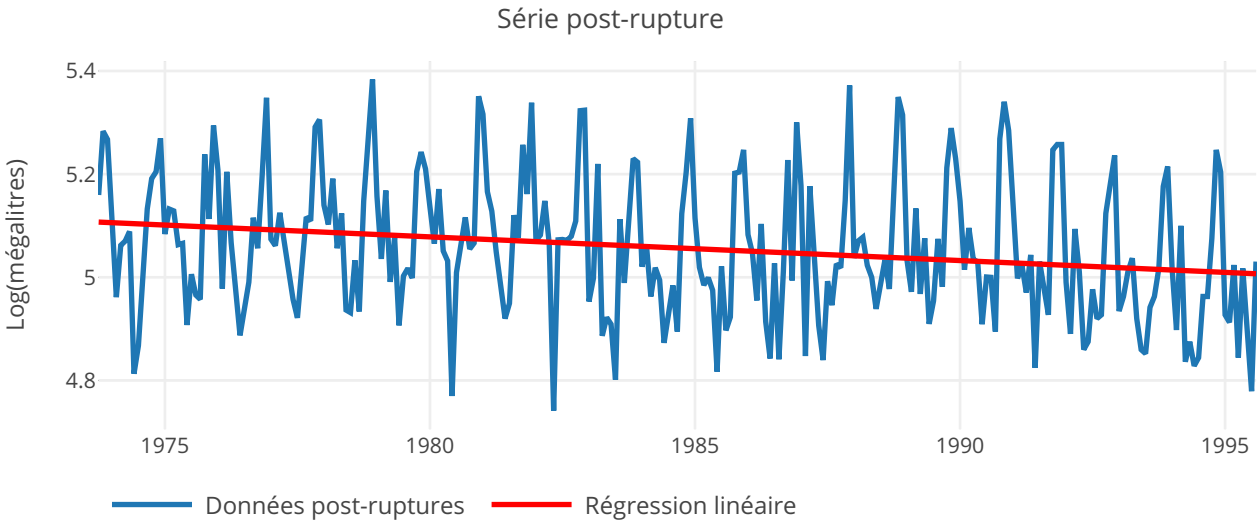
$$MSE = \frac{\sum_{i=1}^{R}[Lmbp_i - (\alpha_0 + \alpha_1 * temps_i)]^2 + \sum_{i=R+1}^{n}[Lmbp_i - (\beta_0 + \beta_1 * temps_i)]^2}{n}$$

où α et β sont respectivement les vecteurs des coefficients des régressions avant et après la date de rupture R.

En faisant varier la valeur de R, on obtient une MSE minimale lorsque $R = 213$, ce qui correspond au mois de septembre 1973. Dans la suite de ce rapport, nous utiliserons cette date comme date de rupture.



2. Étude de la partie post-rupture



2.1. Stationnarité

D'après le graphique, la série post-rupture présente une tendance décroissante et ne peut donc être considérée comme stationnaire. Nous pouvons effectuer les tests suivants pour confirmer cette intuition :

- Test de Dickey-Fuller Augmenté (ADF) :

$$\mathcal{H}_0 : \text{"La trajectoire est issue d'un processus non-stationnaire"} \text{ vs } \mathcal{H}_1 : \text{"}\bar{\mathcal{H}_0}\text{"}$$

- Test de Kwiatkwoski-Phillips-Schmidt-Shin (KPSS) :

$$\mathcal{H}_0 : \text{"La trajectoire est issue d'un processus stationnaire"} \text{ vs } \mathcal{H}_1 : \text{"}\bar{\mathcal{H}_0}\text{"}$$

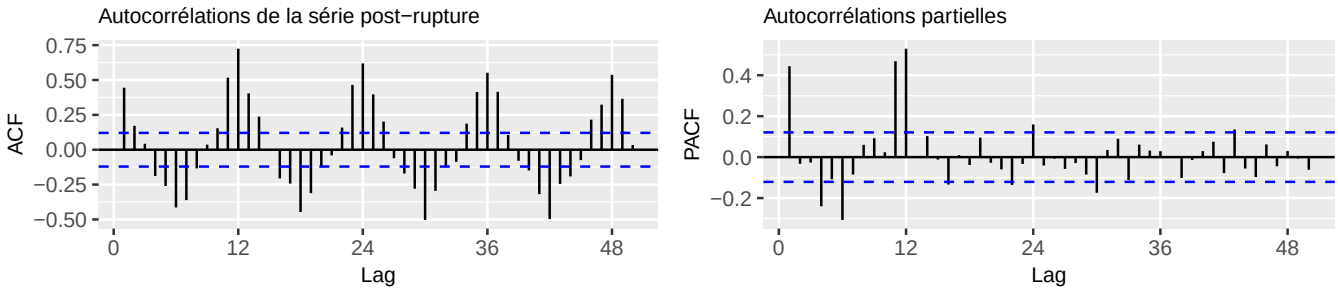
```
#> Warning in adf.test(Lmbp_postR): p-value smaller than printed p-value
#>
#>  Augmented Dickey-Fuller Test
#>
#> data:  Lmbp_postR
#> Dickey-Fuller = -11.385, Lag order = 6, p-value = 0.01
#> alternative hypothesis: stationary
#>
#>  KPSS Test for Level Stationarity
#>
#> data:  Lmbp_postR
#> KPSS Level = 0.68442, Truncation lag parameter = 5, p-value = 0.01496
```

Ainsi, d'après le test ADF nous pouvons rejeter l'hypothèse nulle de non-stationnarité au seuil de 5%. Or, d'après le test KPSS, nous pouvons également rejeter l'hypothèse nulle (stationnarité des données) au seuil de 5%, ce qui contredit le résultat précédent. A noter que la valeur de la p-value étant assez proche de 0.05, la fiabilité du résultat peut être remise en question.

En conclusion, nous pouvons émettre des doutes sur la stationnarité des données ci-dessus.

2.2. Analyse des autocorrélations

L'analyse des autocorrélations (ACF) de la série post-rupture nous montre une périodicité annuelle des données, comme relevée précédemment. Les valeurs estimées pour la fonction d'autocorrélations partielles (PACF) semblent rentrer dans le couloir de significativité à partir du douzième *lag*, ce qui est en accord avec la théorie des ARMA. Etant donné la périodicité relevée dans l'estimation de l'ACF, nous pouvons tenter de modéliser nos données par un modèle SARIMA, ce que nous exposerons dans la partie suivante.



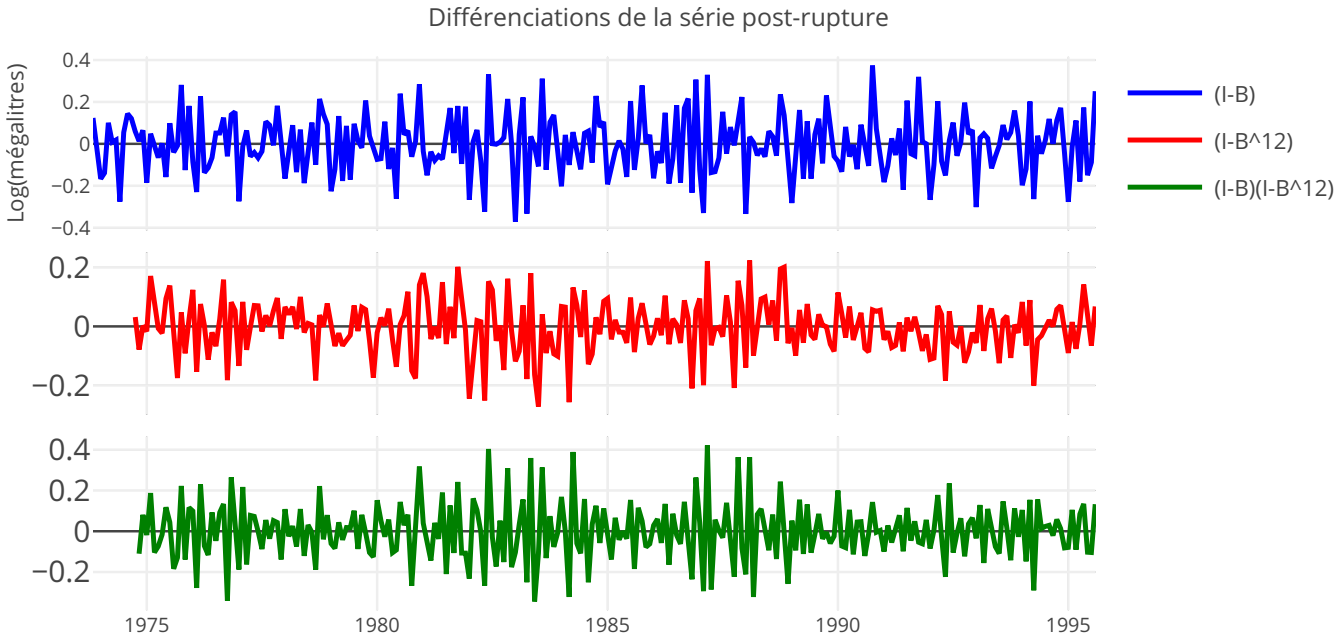
3. Définition des modèles

Nos analyses nous amènent donc à définir trois modèles SARIMA basés sur les différenciations suivantes :

- 1er modèle : $(I - B)$ afin de supprimer la tendance linéaire
- 2ème modèle : $(I - B^{12})$ afin de supprimer la tendance saisonnière
- 3ème modèle : $(I - B)(I - B^{12})$ afin de supprimer à la fois la tendance linéaire et la tendance saisonnière

3.1. Modélisation

Les différenciations de la série selon les opérateurs ci-dessus nous donnent les trois graphiques suivants :



Nous avons ainsi supprimé la tendance linéaire dans les trois modélisations qui semblent toutes trois stationnaires avec néanmoins un doute sur la stabilité de la variance dans la troisième série, voire la deuxième. Observons les résultats des tests ADF et KPSS :

Table 3: Résultats des tests de stationnarité

	(I-B)	$(I - B^{12})$	$(I - B)(I - B^{12})$
ADF	<0.01	<0.01	<0.01
KPSS	>0.1	>0.1	>0.1

Les tests ADF nous conduisent tous trois à rejeter l'hypothèse nulle de non-stationnarité des données, tandis que le test KPSS nous indique de ne pas rejeter l'hypothèse de stationnarité des données dans les trois cas de figure.

Le choix des couples (d,D)=(1,0), (d,D)=(0,1) et (d,D)=(1,1) semble donc pertinent pour une modélisation SARIMA. Nous utiliserons la commande *auto.arima* du package *forecast* pour déterminer les nombres p, q, P et Q de paramètres ainsi que leurs estimations. Nous testerons la significativité des coefficients β_i par le test de Student:

$$\mathcal{H}_0 : "\beta_i = 0" \; vs \; \mathcal{H}_1 : "\beta_i \neq 0"$$

Le test de Student nous conduit à rejeter $\mathcal{H}_0$ au seuil de confiance α si la statistique $t = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta_i}}$ est strictement supérieure en valeur absolue au quantile d'ordre α de la loi normale centrée réduite.

3.1.1. Modèle 1

```
auto.arima(Lmbp_postR,d=1,D=0)
#> Series: Lmbp_postR
#> ARIMA(1,1,1)(2,0,0)[12]
#>
#> Coefficients:
#>           ar1      ma1   sar1    sar2
#>       -0.1122  -0.9839  0.593  0.2576
#> s.e.   0.0439   0.0280  0.020  0.0176
#>
#> sigma^2 estimated as 0.007053:  log likelihood=271.6
#> AIC=-533.2   AICc=-532.97   BIC=-515.36
```

La statistique pour la significativité des coefficients correspond donc à l'estimation du coefficient divisé par son écart-type, statistique qui ne doit pas être comprise entre -1.96 et 1.96 pour que le coefficient soit significatif au seuil de 5%. On voit qu'ici tous les coefficients correspondent à ce critère même si nous pouvons émettre un doute sur la significativité du paramètre $ar1$ au seuil de 1%. Nous pouvons donc tenter d'estimer notre premier modèle à l'aide d'un SARIMA(0,1,1)x(2,0,0)[12] comme ceci :

```
Mod1 = Arima(Lmbp_postR,order=c(0,1,1),seasonal = list(order=c(2,0,0),period=12))
```

```
#>                   ma1       sar1      sar2
#> M1_coeff  -0.98807830 0.57260143 0.2537649
#> M1_se      0.01250056 0.06052196 0.0616061
#> t_stat   -79.04272782 9.46105234 4.1191517
```

Les coefficients sont bien tous significatifs. Nous procédons de la même manière par la suite : recherche du meilleur modèle par la fonction *auto.arima* avec le couple de paramètres (d,D) correspondant puis réduction du modèle par élimination progressive des coefficients non significatifs pour les paramètres de plus haut rang.

3.1.2. Modèle 2

Le deuxième modèle sera donc un SARIMA(0,0,0)x(1,1,1)[12] avec une tendance linéaire :

```
Mod2 = Arima(Lmbp_postR,order=c(0,0,0),seasonal = list(order=c(1,1,1),period=12),include.drift=TRUE)
```

```
#>               sar1         sma1         drift
#> M2_coeff 0.2320122  -0.99999618 -0.0003336736
#> M2_se    0.0665181   0.07681995  0.0000967804
#> t_stat   3.4879558 -13.01740140 -3.4477388835
```

3.1.3. Modèle 3

Le modèle 3 sera basé sur un SARIMA(2,1,2)x(0,1,1)[12] :

```
Mod3 = Arima(Lmbp_postR,order=c(2,1,2),seasonal = list(order=c(0,1,1),period=12))
```

```
#>                 ar1         ar2        ma1        ma2         sma1
#> M3_coeff -0.7070092 -0.32858757 -0.4039668 -0.4231583  -0.86145976
#> M3_se     0.1426741  0.06349383  0.1425601  0.1314174   0.05322621
#> t_stat   -4.9554131 -5.17511042 -2.8336608 -3.2199564 -16.18487774
```

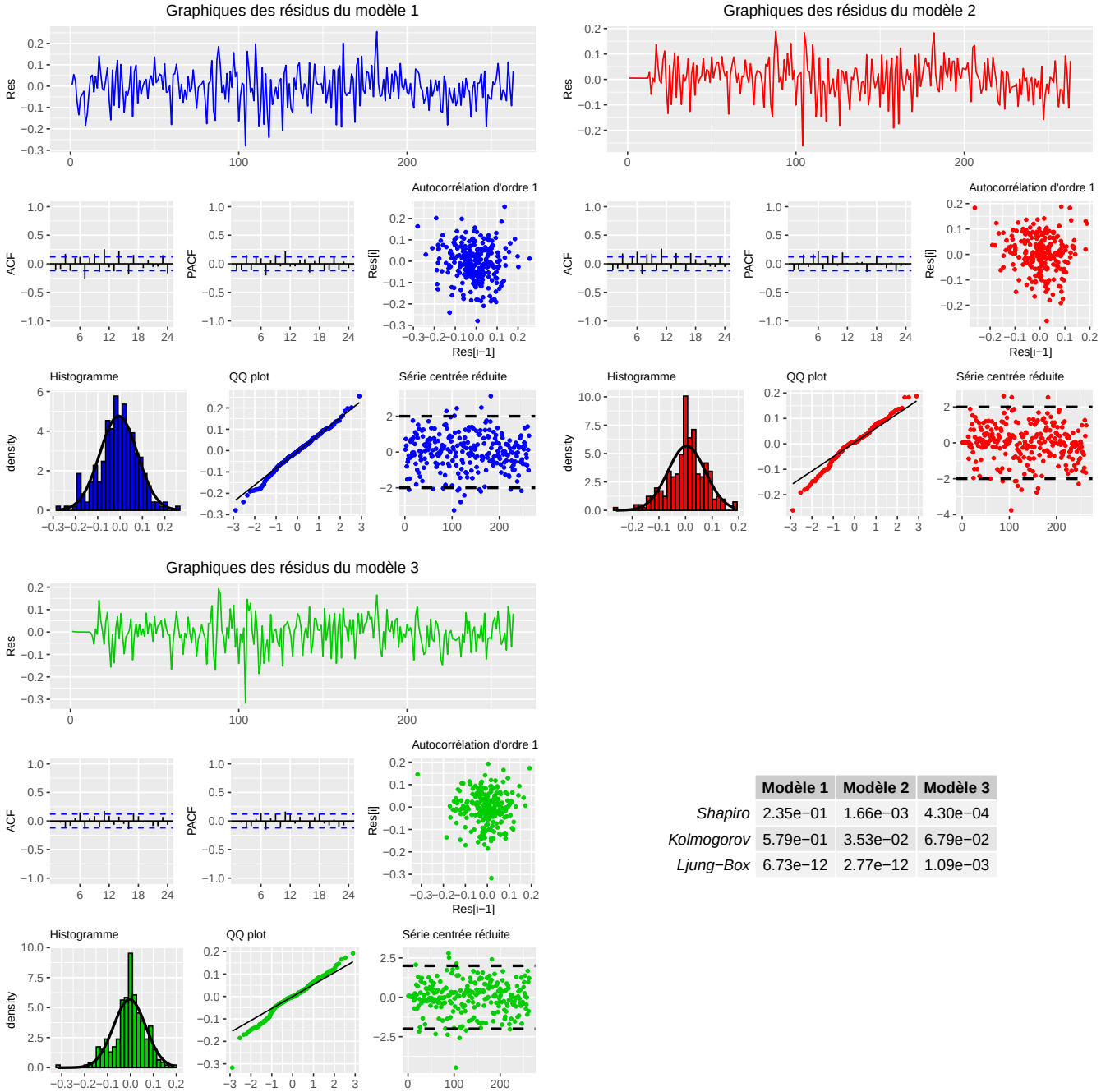
3.2. Etude des résidus

Dans cette partie, nous cherchons à étudier si les résidus de nos trois modèles peuvent être assimilés à un bruit blanc gaussien par des appréciations graphiques et différents tests statistiques : - Normalité des résidus : Tests de Shapiro et de Kolmogorov-Smirnov

$$\mathcal{H}_0 : \text{"Les données suivent une distribution normale"} \text{ vs } \mathcal{H}_1 : \text{"}\bar{\mathcal{H}_0}\text{"}$$

- Bruit blanc : Test de Ljung-Box

$$\mathcal{H}_0 : \text{"Les données ne présentent pas d'auto-corrélation d'ordre 1 à r"} \text{ vs } \mathcal{H}_1 : \text{"}\bar{\mathcal{H}_0}\text{"}$$



	Modèle 1	Modèle 2	Modèle 3
Shapiro	2.35e−01	1.66e−03	4.30e−04
Kolmogorov	5.79e−01	3.53e−02	6.79e−02
Ljung−Box	6.73e−12	2.77e−12	1.09e−03

Les tests de Shapiro et Kolmogorov s'accordent dans le non-rejet de l'hypothèse de normalité des résidus du premier modèle. En revanche, le test de Shapiro rejette fermement la normalité des résidus des modèles 2 et 3 tandis que les résultats du test de Kolmogorov, plus sensible aux grands échantillons, semble plus douteux dans ces deux cas. D'un point de vue graphique, les trois modèles admettent la majorité de leurs résidus centrés réduits entre -2 et 2. Cependant, les histogrammes et diagrammes quantile-quantiles des modèles 2 et 3 semblent moyennement

assimilable à une distribution normale tandis que les résidus du modèle 1 montrent une meilleure adéquation à une loi gaussienne.

Le test du Ljung-Box rejette fermement l'hypothèse de non auto-corrélation des résidus de l'ordre 1 à 12 pour les trois modèles. Graphiquement cependant, nous n'observons pas d'auto-corrélations entre les résidus que ce soit dans le nuage de points ou l'analyse des ACF et PACF.

En conclusion, le modèle 1 semble être le plus pertinent du point de vue de l'analyse des résidus puisque ce sont les seuls qui semblent suivre une loi normale. Cependant, nos trois modèles semblent cohérents vis-à-vis de l'absence de corrélations dans les résidus.

3.3. Erreur de prédiction

Nous cherchons ici à estimer l'erreur de prédiction de chacun de nos modèles en les appliquant à notre série tronquée de sa dernière période et en calculant l'erreur quadratique moyenne sur l'estimation de cette dernière période. La MSE a d'abord été calculée sur le logarithme de nos données puis, dans un deuxième temps, sur notre série d'origine. La transformation s'est effectuée de la manière suivante :

Soit $\hat{Lmbp_t}$ le prédicteur de $Lmbp_t$ correspondant au logarithme des données. On a alors, sous l'hypothèse de normalité des résidus (modèle 1) :

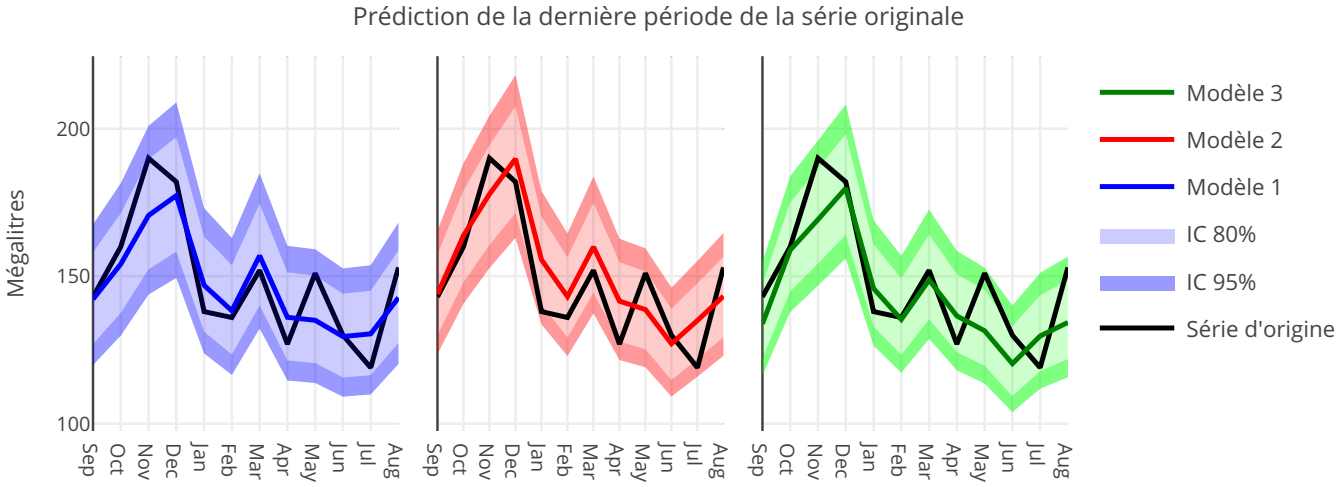
$$\hat{mbp}_{n+1} = e^{\hat{Lmbp}_{n+1} + \frac{\hat{\sigma}^2}{2}} \text{ où } mbp \text{ fait référence à la série originale et } \sigma^2 \text{ à l'écart-type de la série.}$$

Dans les modèles 2 et 3, le prédicteur de la série d'origine est donné par :

$$\hat{mbp}_{n+1} = e^{\hat{Lmbp}_{n+1}} * \frac{1}{n} \sum_{k=1}^{n} e^{\epsilon_k} \text{ où } \epsilon_k \text{ correspond au } k^{ieme} \text{ résidu du modèle.}$$

Table 4: MSE avant et après transformation

	Modèle 1	Modèle 2	Modèle 3
MSE Lmbp	4.19e-03	5.24e-03	6.21e-03
MSE mbp	93.027	114.21	135.19



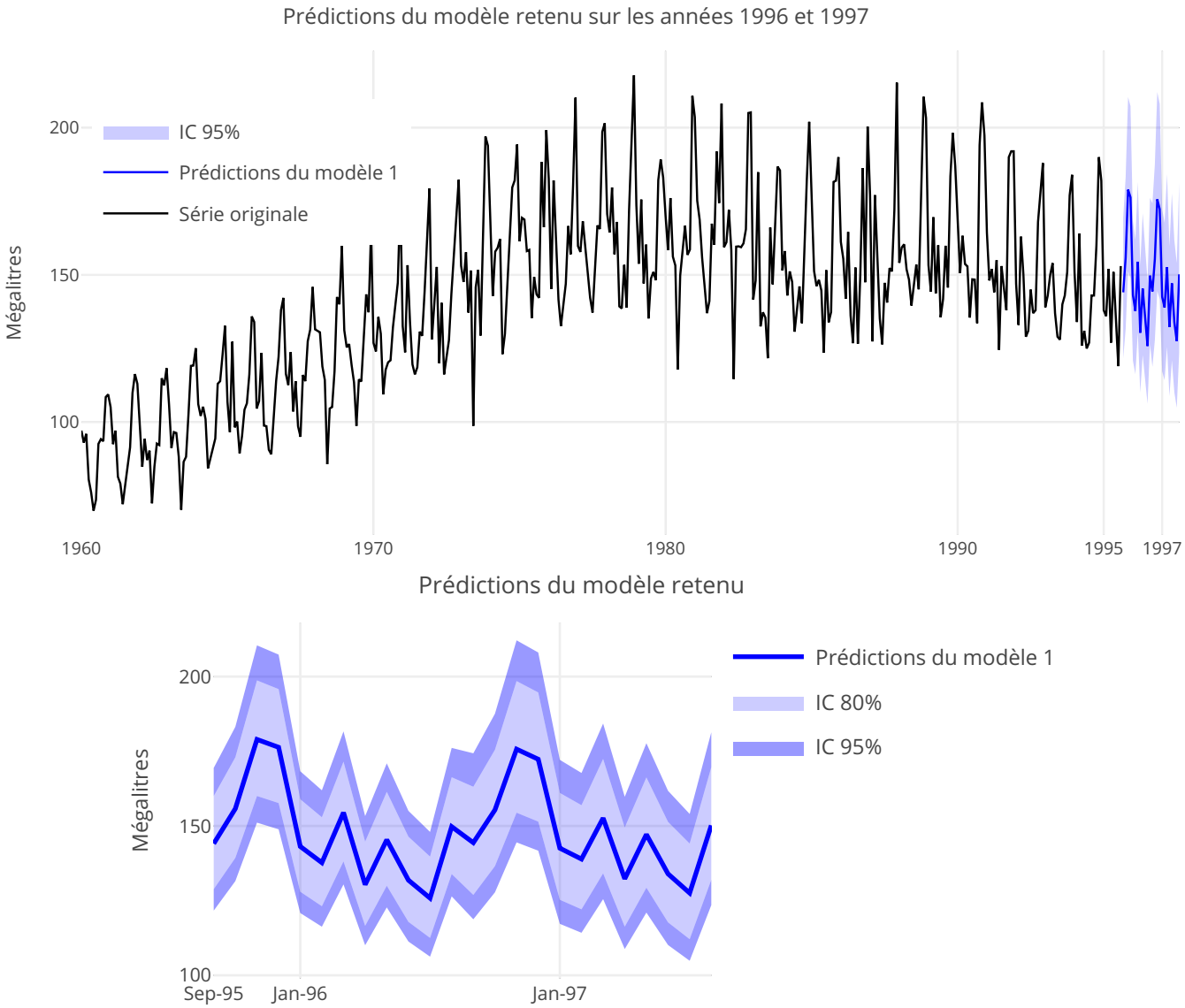
Ainsi, le premier modèle est celui qui minimise la MSE de la prédiction et dont l'ensemble des valeurs de la série se situe dans l'intervalle de confiance à 95% et même 80%. Etant donné l'ensemble de notre analyse, nous choisissons de retenir le premier modèle qui semble être le plus adapté à la prédiction de nos données. Dans notre dernière partie, nous tenterons de prédire la production mensuelle de bière pour les années 1996 et 1997.

4. Conclusion

Le modèle retenu est donc un SARIMA(0,1,1)x(2,0,0)[12] dont l'écriture est déterminée par l'équation suivante :

$$(I - B)(I - \alpha_1 B^{12} - \alpha_2 B^{24}) Lmbp_t = (I + \theta_1 B)\epsilon_t$$

où les α_i correspondent aux coefficients $sar1$ et $sar2$, et θ_1 est estimé par le coefficient $ma1$ de la fonction $Arima$.



L'estimation des deux années supplémentaires semble en accord avec le motif saisonnier de la série estimé par la fonction *decompose* au début de notre rapport. On retrouve une forte croissance à partir de septembre jusqu'à la fin de l'année, puis une décroissance jusqu'en juillet à l'exception du mois de mars où la production est en hausse. On remarque tout de même l'apparition d'un deuxième pic de production au mois de mai.

Pour améliorer ce modèle, nous pourrions envisager de prendre en considération de nouvelles variables comme la consommation mensuelle de bière en Australie, les exportations australiennes mensuelles de bière, ... L'ajout de ces informations permettrait peut-être une meilleure prédiction de la tendance générale de nos données.