Projet de Séries Temporelles

Rapport final

“Indice de la production industrielle pour l’extraction de pétrole brut en France”

NDZONDE FONKOU Yvan Landry et TATOU DEKOU Thibault

[Revised 23 mai 2024]

1 Les données

1.1 Que représente la série choisie ?

Ce projet examine l'indice CVS-CJO de la production industrielle pour l'extraction de pétrole brut entre Janvier 1990 et décembre 2023 en France. Cet indice mesure la production industrielle spécifique à l'extraction de pétrole brut en France. Il est corrigé des variations saisonnières et des jours ouvrés, ce qui signifie que les variations qui se répètent chaque année à la même période (comme les fluctuations saisonnières) sont ajustées, ainsi que les variations liées au nombre de jours ouvrés dans chaque mois. En outre, cet indice est calculé en base 100 de 2021, ce qui permet de mesurer ses variations par rapport à cette année de référence. Cette série donne ainsi un aperçu de l'activité industrielle dans le secteur de l'extraction de pétrole brut au fil du temps. Il s'agit d'une série mensuelle. Dans la suite la série originale sera notée $(X_t)_{t\in T}$. Sa représentation graphique est donnée par la figure 1 ci-après :

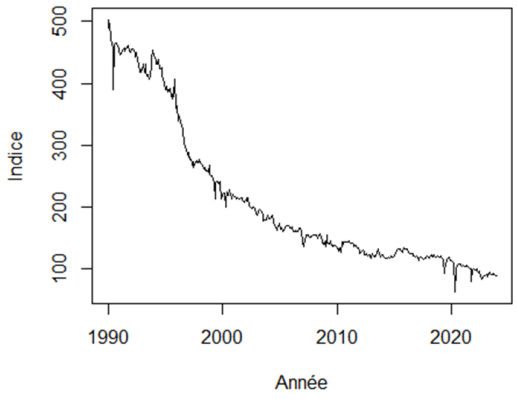

FIGURE 1 – Evolution de l'indice de 1990 à 2023

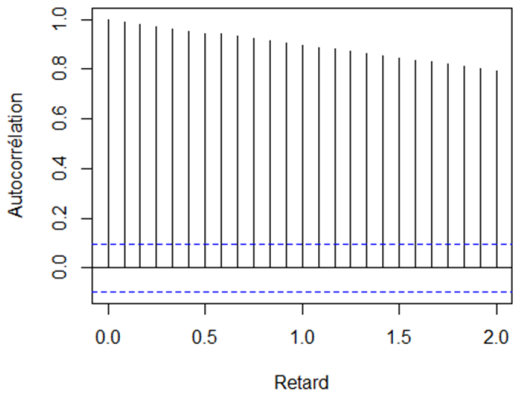

FIGURE 2 – Corrélogramme de la série

1.2 Transformation de la série

L'analyse de la série et de son corrélogramme suggère une possible non-stationnarité de la série. En effet, on observe une forte décroissance de l'indice au cours la période étudiée. Pour se prononcer définitivement sur la stationnarité de la série il est nécessaire de réaliser des tests de racines unitaires. Avant de procéder à ceux-ci, il convient de vérifier s'il y a une constante et/ou une tendance linéaire non nulle. La représentation graphique de la série montre que la tendance n'est probablement pas linéaire, mais que si on devait en choisir une, elle serait négative. Ce que confirme la régression de la série sur

ses dates, comme illustré sur la figure 13 en annexe. La significativité du coefficient de la tendance indique qu'il faudra se mettre dans le cas des tests de racine unitaire avec constante et éventuellement tendance non nulle. Par conséquent, le test augmenté de Dickey-Fuller (ADF) et le test KPSS adéquats sont réalisés pour vérifier la non-stationnarité de la série. Le test ADF nous permet de tester l'existence d'une racine unitaire en présence d'une tendance, il postule de fait la non-stationnarité de la série comme hypothèse nulle. Le test KPSS, quant à lui, postule stationnarité d'une série comme hypothèse nulle. Il sied de rappeler que le test ADF n'est valide que lorsque les résidus ne sont pas autocorrélés. Le test de Ljung-Box est effectué pour vérifier cela. Le nombre de retards requis pour le test ADF est augmenté tant que les autocorrélations ne sont pas nulles (voir annexe : 7). Les résultats de ces tests sont présentés ci-dessous.

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 21
  STATISTIC:
    Dickey-Fuller: -1.7013
  P VALUE:
    0.7036
```

FIGURE 3 – Test ADF sur la série brute

```
        KPSS Test for Level Stationarity

data:  spread
KPSS Level = 5.8061, Truncation lag parameter = 5, p-value = 0.01
```

FIGURE 4 – Test KPSS sur la série brute

Le test ADF avec tendance, permet de ne pas rejeter, l'hypothèse nulle de présence de racine unitaire (p_value = 0,70 > 0,05) au seuil de signification de 5%. De même, le test KPSS rejette l'hypothèse nulle de stationnarité (p_value = 0,01 < 0,05) au niveau de signification de 5%. Partant des résultats du test ADF, on peut conclure que la série est au moins I(1). Pour stationnariser la série, le filtre en différences premières a été utilisé.

1.3 Représentation de la série différenciée

Comme le montre la Figure 5, la série différenciée semble être stationnaire. Pour confirmer cela, les tests KPSS, ADF et Phillips-Perron ont été réalisé comme précédemment. Au seuil de 5%, les tests ADF et Phillips-Perron rejettent l'hypothèse nulle d'existence d'une racine unitaire dans la série différenciée (Voir annexe : Figures 9 et 11), tandis que le test KPSS ne parvient pas à rejeter l'hypothèse nulle de stationnarité au niveau de signification de 5% (Voir annexe : Figure 10). Par conséquent, nous pouvons conclure que la série différenciée est stationnaire. La série considérée dans ce projet est donc I(1). La Figure 6 permet de comparer la série avant et après transformation.

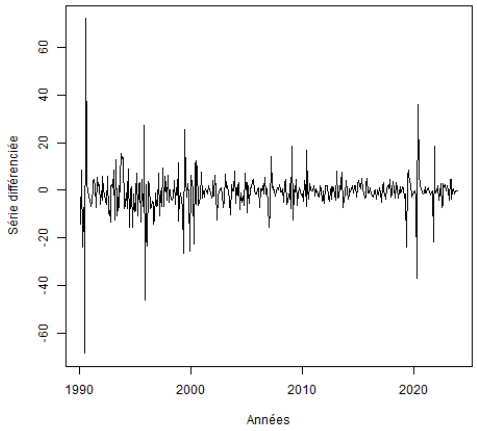

FIGURE 5 – Série différenciée à l'ordre 1

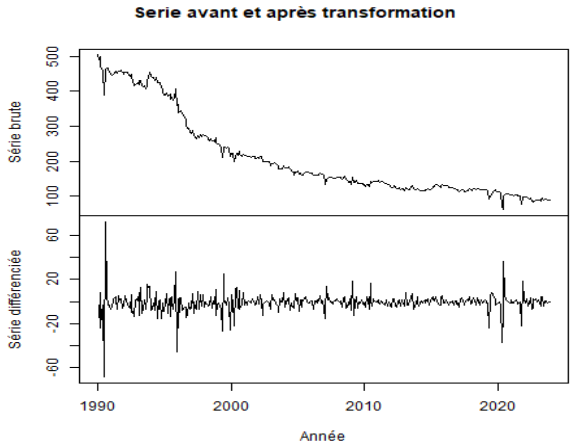

FIGURE 6 – Comparaison des séries

2 Modèle ARMA

2.1 Sélection du modèle, estimation des paramètres et validation

Cette section s'attèle à déterminer le modèle ARMA(p,q) qui correspond le mieux à la série différenciée que l'on sait stationnaire. Le corrélogramme et la fonction d'autocorrelation partielle représentées ci-dessous sont mis à contribution pour déterminer les ordres maximaux p et q du modèle.

L'analyse de la fonction d'autocorrélation nous montre que $q \leq 4$, tandis que la fonction d'autocorrelation partielle montre que $p \leq 4$ (voir Figure 14 en annexe). Aussi un choix délibéré est fait d'ignorer les pics significativement non nuls au delà de 4 retards. L'objectif étant d'obtenir des modèles parcimonieux. Les modèles ARMA(p,q) avec $0 \leq p, q \leq 4$ ont ensuite été analysés en utilisant la méthodologie de Box-Jenkins : estimation des modèles, validation des modèles (tests de non-autocorrélation des résidus (tests de Ljung-Box)), et sélection du modèle grâce aux critères d'information. Tous les tests sont effectués au seuil de signification de 5%. Les critères d'information AIC et BIC des différents modèles sont consignés dans les tableaux 1 et 2 en annexe.

Nous retenons parmi les modèles candidats ceux qui minimisent les critères d'informations. Selon le critère AIC, le meilleur modèle est ARMA(4,3) tandis que le critère BIC propose le modèle ARMA(0,1). Après examen, il s'avère que les deux modèles retenus sont valides et s'ajustent bien à nos données (les coefficients des ordres maximaux p et q sont significatifs pour les deux modèles et l'absence d'autocorrélation n'est jamais rejetée (Voir Figures 15 et 16 en annexe). Toutefois, les prévi-

sions des deux modèles étant proches et dans le soucis d'avoir le modèle plus parcimonieux posssible, le modèle minimisant le critère BIC est celui qui est retenu.

2.2 Expression du modèle ARIMA(p,d,q) pour la série brute

La série en différences premières étant stationnaire, il vient que d=1. De ce qui précède, on conclut que le modèle correspondant à la série brute initiale est **ARIMA(0,1,1)** dont l'équation est la suivante :

$$(1-L)X_t = (1+0.2595L)\epsilon_t$$

soit

$$X_t = X_{t-1} + \epsilon_t + 0.2595\epsilon_{t-1}$$

3 Prévision

Dans cette section, on désignera par T la longueur de la série. On suppose que $(\epsilon_t)_{t\geq 0}$ est un bruit blanc gaussien (avec $\epsilon_t \sim N(0, \sigma_\epsilon^2)$). On note également $\psi = -0.2595$ le coefficient de MA(1) dans le modèle ARMA(0,1) retenu pour la série différenciée.

3.1 Région de confiance de niveau α sur les valeurs futures (X_{T+1}, X_{T+2})

$$\begin{cases} X_{T+1} = X_T + \epsilon_{T+1} - \psi\epsilon_T \\ X_{T+2} = X_{T+1} + \epsilon_{T+2} - \psi\epsilon_{T+1} \end{cases} \implies \begin{cases} X_{T+1} = X_T + \epsilon_{T+1} - \psi\epsilon_T \\ X_{T+2} = X_T - \psi\epsilon_T + (1-\psi)\epsilon_{T+1} + \epsilon_{T+2} \end{cases}$$

$$\implies \begin{cases} \hat{X}_{T+1|T} = X_T - \psi\epsilon_T \\ \hat{X}_{T+2|T} = X_T - \psi\epsilon_T \end{cases}.$$

$$\implies \begin{cases} X_{T+1} - \hat{X}_{T+1|T} = \epsilon_{T+1} \\ X_{T+2} - \hat{X}_{T+2|T} = (1-\psi)\epsilon_{T+1} + \epsilon_{T+2} \end{cases}.$$

On suppose à présent ϵ_t est un bruit blanc gaussien de variance σ^2. Notons :

$$X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix} et, \hat{X} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix}$$

Alors on a :

$$(X - \hat{X}) = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ (1-\psi)\epsilon_{T+1} + \epsilon_{T+2} \end{pmatrix} \sim N(0, \Sigma)$$

avec,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & 1-\psi \\ 1-\psi & (1-\psi)^2 + 1 \end{pmatrix}$$

Il apparait alors que la matrice de variance-covariance Σ est inversible si et seulement si $\det(\Sigma) = \sigma^2 > 0$. En supposant que ce soit le cas, on sait que :

$$(X - \hat{X})^\top \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$$

de sorte que la région de confiance de niveau α s'écrit

$$\{X \in \mathbb{R}^2 \mid (X - \hat{X})^\top \Sigma^{-1} (X - \hat{X}) \leq q_{\chi^2_{1-\alpha}}(2)\},$$

où $q_{\chi^2_{1-\alpha}}(2)$ le quantile d'ordre $1-\alpha$ la distribution du $\chi^2(2)$.

3.2 Hypothèses formulées pour la détermination de la région de confiance

Dans l'optique d'obtenir la région de confiance précédente, certaines hypothèses ont été admises. Il s'agit notamment des hypothèses ci-après :

1) Le modèle est parfaitement identifié, i.e que l'on suppose que les paramètres estimés sont les vrais paramètres du modèle.
2) Les résidus sont distribués suivant une loi normale : $\epsilon_t \sim N(0, \sigma_\epsilon^2)$
3) On suppose σ_ϵ^2 connu, et tel que $\sigma_\epsilon^2 > 0$
4) On suppose enfin que $(\epsilon_t)_t$ est indépendamment et identiquement distribué.

En effet, en l'absence de telles hypothèses, la matrice de variance covariance est alors inconnue du fait non seulement de l'incertitude liée à l'estimation de la variance des résidus mais aussi du fait de l'incertitude liée à l'estimation des coefficients du modèle.

3.3 Représentation graphique de la région de confiance

La prévision en T+1 est la même en T+2 exactement comme le suggérait le modèle retenu. Ces prévisions sont visibles sur la figure 17 en annexe. Le couple (89.13917, 89.13917) représentant les

prévisions à ces dates futures, est le centre de l'ellipse de confiance. L'avantage qu'il y a à recourir aux régions de confiance, est qu'elles limitent le risque d'avoir de mauvaises prévisions, comparativement au scénario dans lequel les prévisions seraient réalisées les unes après les autres.

3.4 Question ouverte

Si (Y_t) cause instantanément (X_t) au sens de Granger, alors Y_{T+1} nous apporte une information supplémentaire sur X_{T+1} s'il est possible d'avoir Y_{T+1} plus vite. Concrètement, cela signifie que Y_{T+1} fournit des informations supplémentaires pour prédire X_{T+1}. Par conséquent, l'erreur de prévision de X_{T+1} lorsque l'on considère les informations fournies par Y_{T+1} est inférieure à cette erreur de prévision lorsque Y_{T+1} n'est pas prise en compte. Typiquement, cela se traduit par :

$$\mathbb{E}[(X_{T+1} - \mathbb{E}(X_{T+1}|Y_{T+1}, X_T, \dots, Y_T, \dots))^2] \leq \mathbb{E}[(X_{T+1} - \mathbb{E}(X_{T+1}|X_T, \dots, Y_T, \dots))^2]$$

Avec la condition ci-après :

$$E[X_{T+1}|Y_{T+1}, X_T, ..., X_0, Y_T, ..., Y_0] \neq E[X_{T+1}|X_T, ..., X_0, Y_T, ..., Y_0].$$

Pour tester cela, nous considérons une série temporelle multivariée $Z_t = (X_t, Y_t)$ et nous procédons en deux étapes. Tout d'abord, nous estimons un modèle VAR sur la série Z_t. Si on désigne par ϵ_{1t} les résidus de l'équation en X_t (équation qui exprime X_t en fonction de ses valeurs passées et celles de Y_t) et par ϵ_{2t} ceux de l'équation en Y_t, l'étape suivante consiste alors à régresser ϵ_{1t} sur ϵ_{2t}. Si le coefficient de ϵ_{2t} n'est pas statistiquement significatif, on en déduit que les résidus sont non correlés et par conséquent que Y_t ne cause pas instantanément X_t et inversément. Dans le cas contraire on conclut que les deux séries se causent l'une l'autre instantanément au sens de Granger.

4 Annexe

	lag	pval
[1,]	1	NA
[2,]	2	NA
[3,]	3	NA
[4,]	4	6.939165e-09
[5,]	5	1.651100e-08
[6,]	6	3.144469e-08
[7,]	7	3.012061e-08
[8,]	8	9.451055e-08
[9,]	9	2.875316e-07
[10,]	10	8.036691e-07
[11,]	11	1.161869e-06
[12,]	12	2.334531e-06
[13,]	13	1.620985e-06
[14,]	14	3.452325e-06
[15,]	15	4.735570e-06
[16,]	16	9.810745e-06
[17,]	17	1.780275e-05
[18,]	18	2.611577e-05
[19,]	19	2.928222e-05
[20,]	20	3.862128e-05
[21,]	21	6.822014e-05
[22,]	22	1.106989e-04
[23,]	23	2.201752e-05
[24,]	24	5.805235e-06

FIGURE 7 – Test de Ljung-Box sur la série brute

	lag	pval
[1,]	1	NA
[2,]	2	NA
[3,]	3	NA
[4,]	4	NA
[5,]	5	NA
[6,]	6	NA
[7,]	7	NA
[8,]	8	NA
[9,]	9	NA
[10,]	10	NA
[11,]	11	NA
[12,]	12	NA
[13,]	13	NA
[14,]	14	NA
[15,]	15	NA
[16,]	16	NA
[17,]	17	NA
[18,]	18	NA
[19,]	19	NA
[20,]	20	NA
[21,]	21	NA
[22,]	22	NA
[23,]	23	NA
[24,]	24	NA

FIGURE 8 – Tests de Ljung Box sur la série différenciée

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 23
  STATISTIC:
    Dickey-Fuller: -2.4493
  P VALUE:
    0.0156
```

FIGURE 9 – Test ADF sur la série différenciée

```
        KPSS Test for Level Stationarity

data:  dspread
KPSS Level = 0.6895, Truncation lag parameter = 5, p-value = 0.0145
```

FIGURE 10 – Test KPSS test sur la série différenciée

```
        Phillips-Perron Unit Root Test

data:  dspread
Dickey-Fuller Z(alpha) = -453.07, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary
```

FIGURE 11 – Test PP test sur la série différenciée

```
        Phillips-Perron Unit Root Test

data:  dspread
Dickey-Fuller Z(alpha) = -453.07, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary
```

FIGURE 12 – Enter Caption

```
Call:
lm(formula = spread ~ dates)

Residuals:
   Min     1Q Median     3Q    Max
-80.83 -46.37 -13.73  47.77 111.60

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21939.3424   510.4501   42.98   <2e-16 ***
dates         -10.8279     0.2543  -42.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 13 – Régression linéaire de la série brute sur ses dates

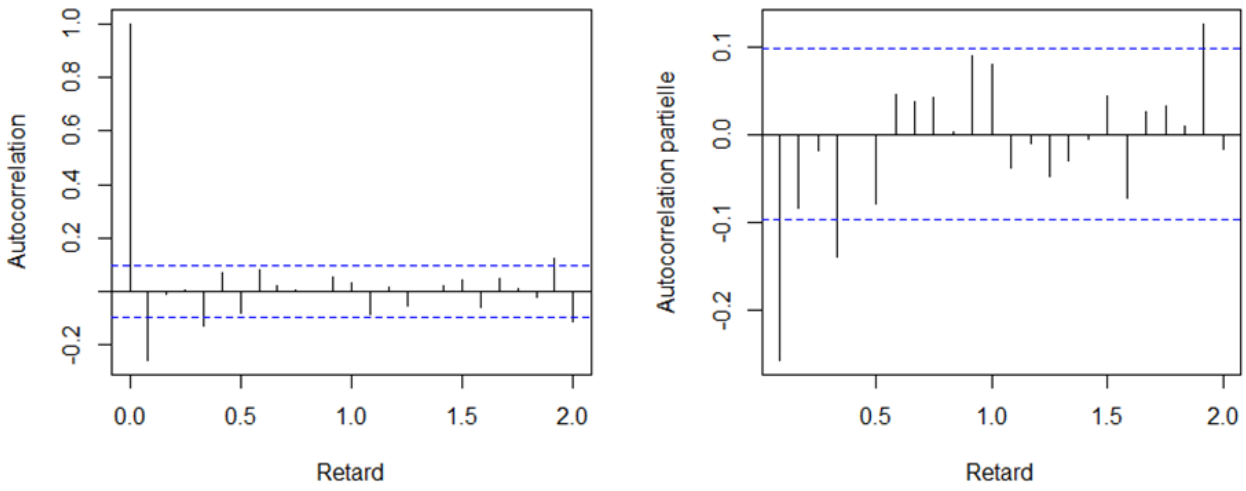

FIGURE 14 – Fonction d’autoccorelation et d’autocorrelation partielle de la série différenciée

p | q	0	1	2	3	4
0	2920	2900	2906	2912	2916
1	2902	2906	2910	2915	2918
2	2907	2908	2914	2914	2921
3	2913	2918	2914	2918	2922
4	2913	2917	2923	2921	2926

TABLE 1 – BIC des modèles ARMA(p,q)

p | q	0	1	2	3	4
0	2916	2892	2894	2896	2896
1	2894	2894	2894	2895	2894
2	2895	2892	2894	2890	2893
3	2896	2898	2890	2890	2890
4	2893	2893	2895	2889	2890

TABLE 2 – AIC des modèles ARMA(p,q)

```
> Qtests(arima413$residuals, 24, fitdf=length(arima413$coef))
      lag      pval
 [1,]   1        NA
 [2,]   2        NA
 [3,]   3        NA
 [4,]   4        NA
 [5,]   5        NA
 [6,]   6        NA
 [7,]   7        NA
 [8,]   8 0.1395566
 [9,]   9 0.1980554
[10,]  10 0.3531309
[11,]  11 0.3320915
[12,]  12 0.2898784
[13,]  13 0.3309938
[14,]  14 0.3906557
[15,]  15 0.4650020
[16,]  16 0.5635691
[17,]  17 0.6519179
[18,]  18 0.7342072
[19,]  19 0.5615386
[20,]  20 0.6403055
[21,]  21 0.7003879
[22,]  22 0.7032362
[23,]  23 0.1820269
[24,]  24 0.2117342
```

FIGURE 15 – Test de Ljung Box pour le modèle ARIMA(4,1,3)

```
> Qtests(arima011$residuals, 24, fitdf=1)
      lag       pval
 [1,]   1         NA
 [2,]   2 0.47042871
 [3,]   3 0.62672153
 [4,]   4 0.05024683
 [5,]   5 0.08753454
 [6,]   6 0.09517268
 [7,]   7 0.05089340
 [8,]   8 0.05914342
 [9,]   9 0.08557384
[10,]  10 0.11928344
[11,]  11 0.09827320
[12,]  12 0.12782311
[13,]  13 0.08146397
[14,]  14 0.11028460
[15,]  15 0.10570105
[16,]  16 0.14159920
[17,]  17 0.16559866
[18,]  18 0.18157387
[19,]  19 0.20623298
[20,]  20 0.21444070
[21,]  21 0.24957083
[22,]  22 0.29825883
[23,]  23 0.15416197
[24,]  24 0.09542079
```

FIGURE 16 – Test de Ljung Box pour le modèle ARIMA(0,1,1)

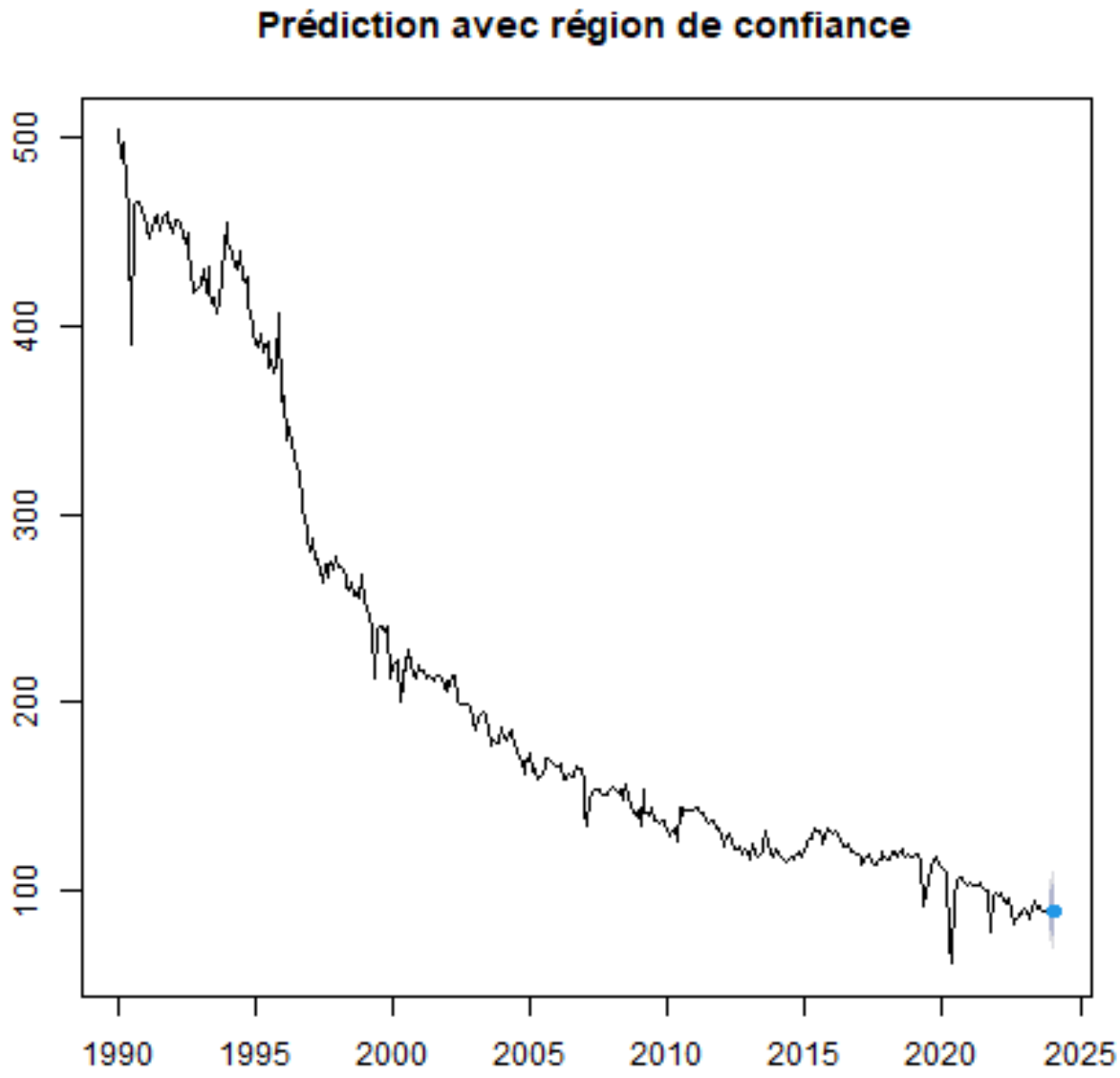

FIGURE 17 – Prévision