S3.03 Description et prévision de données temporelles.

Analyse de séries chronologiques

Analyse de l'indice de prix de production de l'industrie française pour le marché français – CPF 35.11 et 35.14 – Électricité tarif de détail pour les ménages

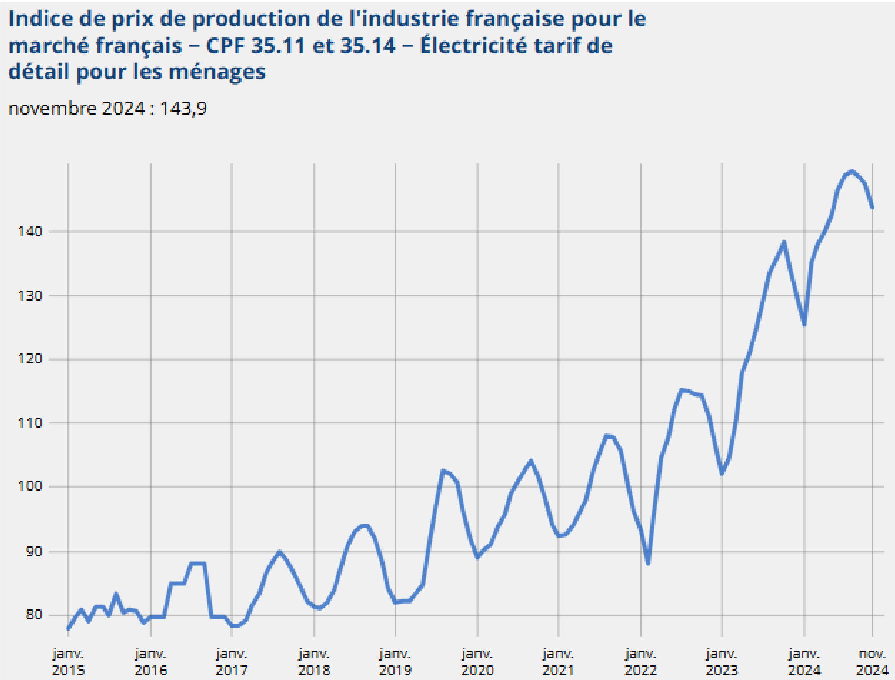
17/01/2025

Groupe : OIKOS

Diallo Thierno
ARANGO CATTY Imany

1. Partie 1 : Analyse de l'évolution de l'indice de production de l’électricité.

A. Choix du dataset



Le graphique de la série chronologique nous présente l'indice des prix de production de l'industrie française pour le marché français (électricité, tarif de détail pour les ménages) de janvier 2015 à novembre 2024.

Nous constatons que ce graphique présente une tendance à la hausse (le prix augmente), car les valeur ne cesse de monter dans le global, même si elle baisse a des moment. Aussi nous pouvons dire que cette série est basée sur un modèle multiplicatif, car la variabilité des données augmente au fur et à mesure que le temps augmente.

Nous pouvons également détecter une composante saisonnière car :

- **Fluctuations régulières** (12 mois) : On peut observer des pics et des creux réguliers, notamment au cours des années. Ces variations correspondent probablement à des changements saisonniers dans la demande d'électricité, par exemple, une augmentation durant les mois d'hiver en raison du chauffage.
- **Lien avec les cycles climatiques** : Les prix de l'électricité sont souvent influencés par des facteurs climatiques saisonniers (hiver/été), qui affectent la consommation énergétique des ménages.
- **Modèle récurrent** : Bien que la tendance générale de l'indice soit à la hausse sur la période, les fluctuations cycliques récurrentes suggèrent une composante saisonnière.

lien du dataset : https://www.insee.fr/fr/statistiques/serie/010764284#Tableau

B. Établissement, analyse et prédiction de la CVS

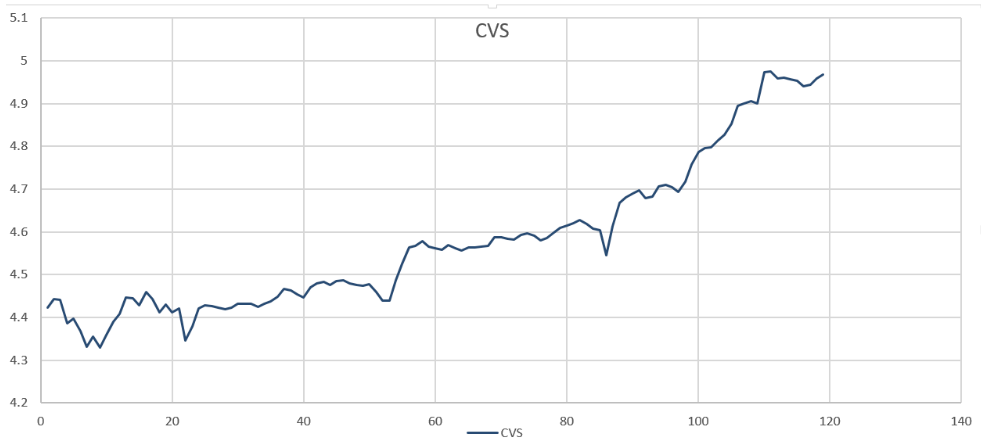
a. Établissement de la CVS et analyse

id	id²	Période	mois	x	ln(x)	mmc 12 : ln(x)	ln(x) - mmc12	s_hat (coef saisonier	s_hat - s_hat_bar
1	1	2015-01	01	77.9	4.35542595			-0.068354747	-0.067764166
2	4	2015-02	02	79.6	4.37701409			-0.067216668	-0.066626087
3	9	2015-03	03	80.8	4.39197697			-0.049783778	-0.049193197
4	16	2015-04	04	79	4.36944785			-0.017007378	-0.016416797
5	25	2015-05	05	81.2	4.39691525			-0.00144747	-0.000856889
6	36	2015-06	06	81.4	4.39937527			0.030572799	0.03116338
7	49	2015-07	07	80	4.38202663	4.386997201	-0.004970566	0.049835188	0.05042577
8	64	2015-08	08	83.3	4.42244855	4.388001331	0.034447218	0.066507784	0.067098365
9	81	2015-09	09	80.5	4.38825718	4.387482503	0.000774682	0.057689367	0.058279949
10	100	2015-10	10	80.8	4.39197697	4.389961504	0.002015462	0.03329181	0.033882391
11	121	2015-11	11	80.6	4.38949865	4.394917312	-0.005418663	-0.00010904	0.000481541
12	144	2015-12	12	78.9	4.36818123	4.398626146	-0.030444918	-0.041064842	-0.040474261
13	169	2016-01	01	79.7	4.37826959	4.404495159	-0.026225573	-0.068354747	-0.067764166
14	196	2016-02	02	79.7	4.37826959	4.410895338	-0.032625752	-0.067216668	-0.066626087
15	225	2016-03	03	79.7	4.37826959	4.416988644	-0.038719058	-0.049783778	-0.049193197
16	256	2016-04	04	85	4.44265126	4.420228722	0.022422534	-0.017007378	-0.016416797
17	289	2016-05	05	85	4.44265126	4.419294197	0.023357059	-0.00144747	-0.000856889
18	324	2016-06	06	85	4.44265126	4.419351161	0.023300096	0.030572799	0.03116338
19	361	2016-07	07	88.2	4.47960696	4.41908534	0.060521623	0.049835188	0.05042577
20	400	2016-08	08	88.1	4.47847253	4.417608508	0.060864025	0.066507784	0.067098365
21	441	2016-09	09	88.1	4.47847253	4.416660448	0.061812085	0.057689367	0.058279949
22	484	2016-10	10	79.8	4.3795235	4.414749887	-0.035226383	0.03329181	0.033882391
23	529	2016-11	11	79.8	4.3795235	4.412207192	-0.032683687	-0.00010904	0.000481541
24	576	2016-12	12	79.8	4.3795235	4.412238552	-0.032715048	-0.041064842	-0.040474261
25	625	2017-01	01	78.3	4.3605476	4.413206068	-0.052658465	-0.068354747	-0.067764166

id	CVS	ln(CVS)	prevision expo	prevision puissance	LES	LES 2	a LED	b LED	Prevision LED	composante de saisonalité
1	4.42319	1.486861	4.309450649	4.387061546	4.42319	4.42319				-0.067764166
2	4.44364	1.491474	4.313835885	4.387182189	4.42319	4.42319	0.02045	4.44364		-0.066626087
3	4.44117	1.490918	4.318225583	4.38738326	4.437505	4.42319	0.001591	4.45182	4.464090241	-0.049193197
4	4.385865	1.478387	4.322619749	4.38766476	4.440071	4.433211	0.000762	4.446931	4.453410765	-0.016416797
5	4.397772	1.481098	4.327018385	4.388026688	4.402126	4.438013	-0.00399	4.36624	4.447692899	-0.000856889
6	4.368212	1.474354	4.331421498	4.388469045	4.399078	4.412892	-0.00153	4.385265	4.362252892	0.03116338
7	4.331601	1.465937	4.335829091	4.38899183	4.377472	4.403223	-0.00286	4.351721	4.383729671	0.05042577
8	4.35535	1.471405	4.340241169	4.389595043	4.345362	4.385197	-0.00443	4.305527	4.348859919	0.067098365
9	4.329977	1.465562	4.344657737	4.390278685	4.352354	4.357313	-0.00055	4.347395	4.301101145	0.058279949
10	4.358095	1.472035	4.349078799	4.391042756	4.33669	4.353841	-0.00191	4.319539	4.346843934	0.033882391
11	4.389017	1.479105	4.35350436	4.391887255	4.351673	4.341836	0.001093	4.361511	4.317633268	0.000481541
12	4.408655	1.48357	4.357934424	4.392812183	4.377814	4.348722	0.003232	4.406906	4.362604031	-0.040474261
13	4.446034	1.492012	4.362368996	4.393817539	4.399403	4.369086	0.003369	4.42972	4.4101384	-0.067764166
14	4.444896	1.491756	4.366808081	4.394903323	4.432045	4.390308	0.004637	4.473781	4.433088222	-0.066626087
15	4.427463	1.487827	4.371251683	4.396069536	4.44104	4.419524	0.002391	4.462557	4.478418433	-0.049193197
16	4.459068	1.49494	4.375699807	4.397316178	4.431536	4.434585	-0.00034	4.428487	4.464947831	-0.016416797
17	4.443508	1.491444	4.380152457	4.398643248	4.450808	4.432451	0.00204	4.469166	4.428147981	-0.000856889
18	4.411488	1.484212	4.384609638	4.400050746	4.445698	4.445301	4.41E-05	4.446095	4.471205814	0.03116338
19	4.429181	1.488215	4.389071354	4.401538673	4.421751	4.445579	-0.00265	4.397923	4.44613943	0.05042577
20	4.411374	1.484186	4.393537611	4.403107029	4.426952	4.428899	-0.00022	4.425005	4.395275284	0.067098365
21	4.420193	1.486183	4.398008413	4.404755813	4.416048	4.427536	-0.00128	4.404559	4.424788472	0.058279949
22	4.345641	1.469173	4.402483764	4.406485025	4.418949	4.419494	-6.1E-05	4.418404	4.403282265	0.033882391
23	4.379042	1.47683	4.406963669	4.408294666	4.367634	4.419113	-0.00572	4.316154	4.418343399	0.000481541
24	4.419998	1.486139	4.411448132	4.410184736	4.375619	4.383077	-0.00083	4.368162	4.310434495	-0.040474261
25	4.428312	1.488018	4.415937159	4.412155233	4.406684	4.377857	0.003203	4.435512	4.367332971	-0.067764166

Tout d'abord, nous avons passé notre série au logarithme pour la transformer en une série additive.

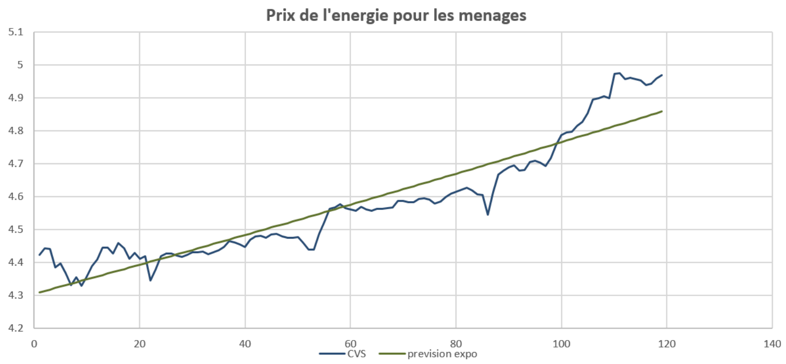
Afin d'établir la CVS, nous avons commencé par effectuer une moyenne mobile de fenêtre 12 pour d'effacer toute la saisonnalité et de récupérer la tendance (quand la fenêtre de la moyenne mobile est un multiple de la périodicité de, elle efface toutes la saisonnalité, dans notre cas c'était 12 mois). Puis nous avons retiré la tendance à nos données brute (Xt- mmc). Puis grâce à cette nouvelle série, nous avons calculé les coefficient saisonnier puis nous les avons centré, et finalement avons calculé la CVS.



- Analyse : Ce graphique nous présente la courbe de la série corrigée de ses variations saisonnières. Premièrement, on remarque que la courbe de notre CVS est croissante dans son ensemble, même si on observe quelques points décroissants et des variations irrégulières autour de la tendance (erreur). On constate également une accélération dans l'augmentation de la tendance(elle monte de plus en plus vite) vers la fin. L'allure semble ressembler à une fonction exponentielle ou une fonction polynomiale.

b. prédiction de la cvs par plusieurs modèles:

- **Modèle exponentiel :**



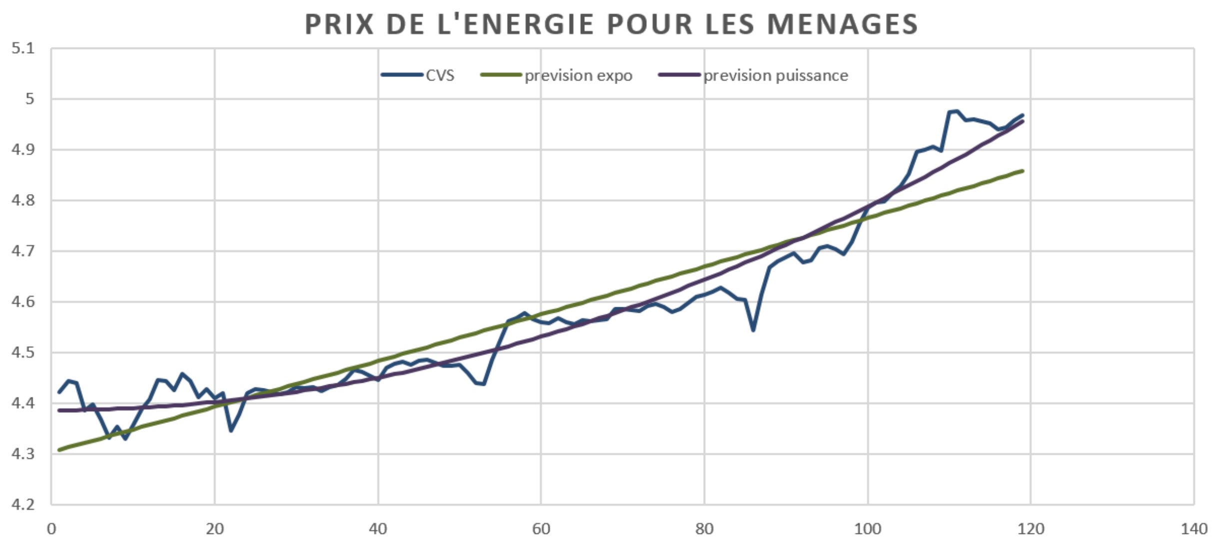
soit le modèle suivant :

$y = be^a \iff ln(y) = ln(be^{ax})$
$\iff ln(y) = ln(b) + ax * ln(e)$
$\iff ln(y) = ln(b) + ax$

posons $Y = ln(y)$ et $B = ln(b)$ donc $Y = ax + B$

Nous cherchons à modéliser la cvs par un modèle exponentiel . Le résultat obtenu est : **a = 0,001 et b = 4,305**. Nous constatons qu'il respecte assez bien la courbe de la CVS en suivant son allure mais nous observons qu'il y à pas mal d'écarts mais également, vers la fin, notre modélisation semble sous estimé la vitesse de croissance de la CVS. Après avoir calculé le coefficient de détermination(R^2) on trouve la valeur de 0.8750, ce qui est déjà pas mal. Le modèle exponentiel ne semble donc pas être le plus adapté dans notre situation. nous pensons pouvoir être encore plus précis en utilisant d'autres modèles (polynomiales) ou méthode lissage exponentielle.

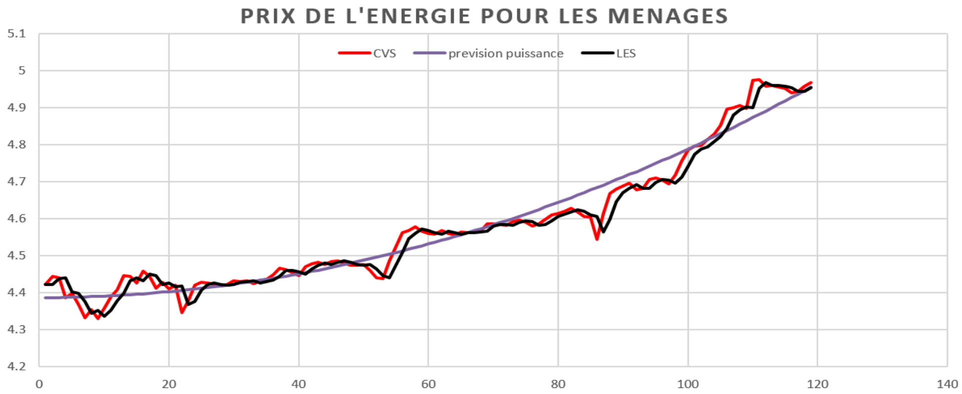
- **Modèle polynomial**
  Soit le modèle suivant : $y = ax^2+b \iff y = aX^2 + b$ avec $X = x^2$



Après avoir obtenu un résultat peu satisfaisant avec le modèle exponentiel, nous essayons avec un modèle polynomial de degré 2. Le résultat obtenu est : **a = 4.021*10e-5 et b = 4.387**. Avec ce modèle nous constatons que la courbe respecte bien l'allure de la CVS en prenant assez bien en compte les écarts. Ce modèle respecte également la vitesse de croissance de notre CVS sans la sous-estimé ou la surestimé.Nous avons calculé le coefficient de détermination pour ce modèle et nous avons trouvé 0.9540 , ce qui est très précis , encore plus que pour le modèle exponentiel. Dans notre cas de figure, nous pouvons d'hors et déjà affirmer que ce modèle polynomial est plus efficace que le modèle exponentiel.

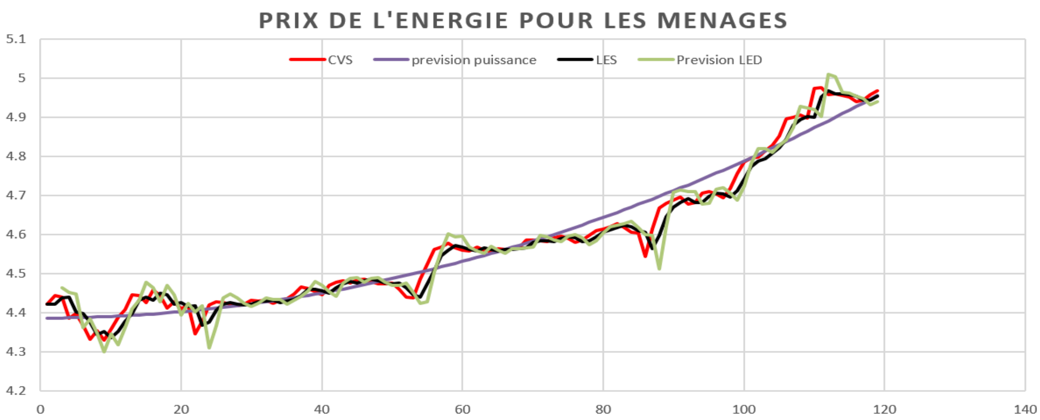
c. Méthode de lissage exponentiel

- Lissage Exponentiel Simple (LES) : β = 0.3



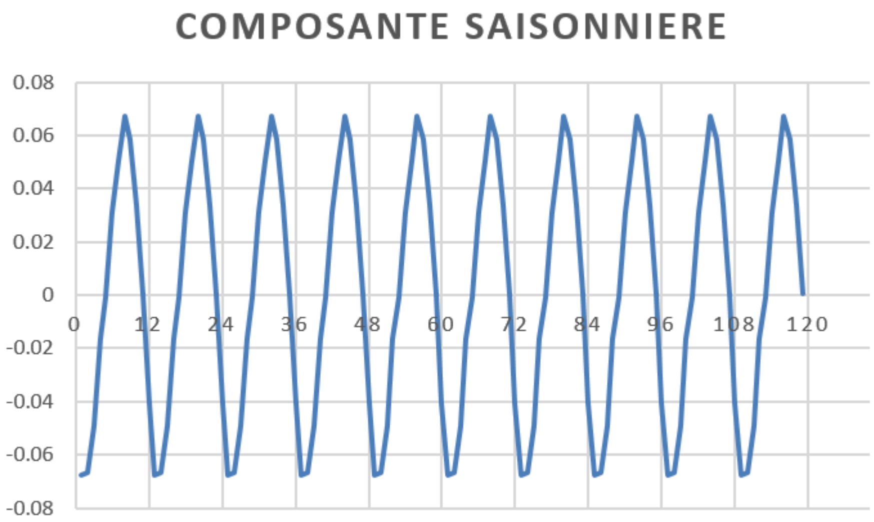
Pour effectuer notre lissage, nous avons choisi comme valeur de β = 0.3. Ce choix nous semble constituer un bon compromis entre réactivité et stabilité. Un coefficient de lissage trop faible aurait tendance à lisser excessivement les données et donc de sous-estimer la vitesse de croissance de notre CVS vers la fin, nous avons donc voulu donner plus d'importance aux valeurs les plus récentes. Avec la méthode de lissage exponentiel simple, on observe une courbe qui suit presque parfaitement la CVS, les écarts à noter sont très minimes. La courbe est beaucoup plus précise que celle du modèle polynomial. Cependant, la méthode par lissage exponentiel simple n'est pas forcément la plus adaptée pour notre cas de figure car elle ne prend pas en compte la tendance pour les futures prédictions, c'est pourquoi nous allons essayer un lissage exponentiel double.

- **LED**



Nous avons choisi une valeur élevée de β (β=0.9) car notre série temporelle présentée montre des fluctuations rapides ainsi qu'une tendance générale croissante et surtout une croissance rapide vers la fin. Ce choix nous permet de donner davantage de poids aux observations récentes, ce qui garantit que notre modèle réagit rapidement aux changements tout en capturant efficacement la tendance globale. Avec la méthode du lissage exponentiel double, on observe un bon respect de la courbe de la CVS, ce modèle prend bien en compte la tendance et cela nous facilite la tâche pour effectuer des prévisions. Malgré le fait que le lissage exponentiel simple semble être plus précis, la méthode du lissage exponentiel double reste plus adaptée, car nous souhaitons effectuer des prévisions à des horizons supérieurs à 1 et comme notre série comporte une tendance, nous pourrons la prendre en compte avec cette méthode.

C. Composante saisonnière.



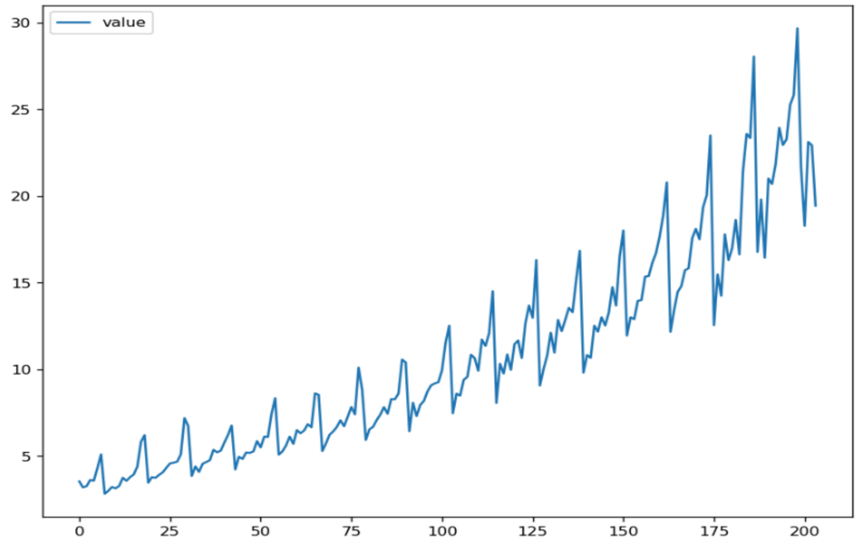
Le graphique présente la composante saisonnière de l'indice d'évolution du prix de production de l'énergie de détail pour les ménages, en fonction du temps (mois). Il met en évidence les variations périodiques de l'indice, qui suivent un schéma de 12 mois.

On remarque que les amplitudes saisonnières sont les mêmes chaque année. On remarque qu'à chaque début d'année les valeurs sont au plus bas et qu'environ tous les milieux d'années les valeurs sont au plus haut. On constate donc que les valeurs ont tendance à augmenter et atteindre un pic en été et descendre en hiver.

Cette saisonnalité peut être interprétée comme le reflet d'habitudes de consommation ou de production spécifiques à l'électricité, telles qu'une demande accrue durant les périodes hivernales (chauffage) et une diminution pendant les mois estivaux. Ces variations pourraient également être influencées par des facteurs externes comme la tarification réglementée.

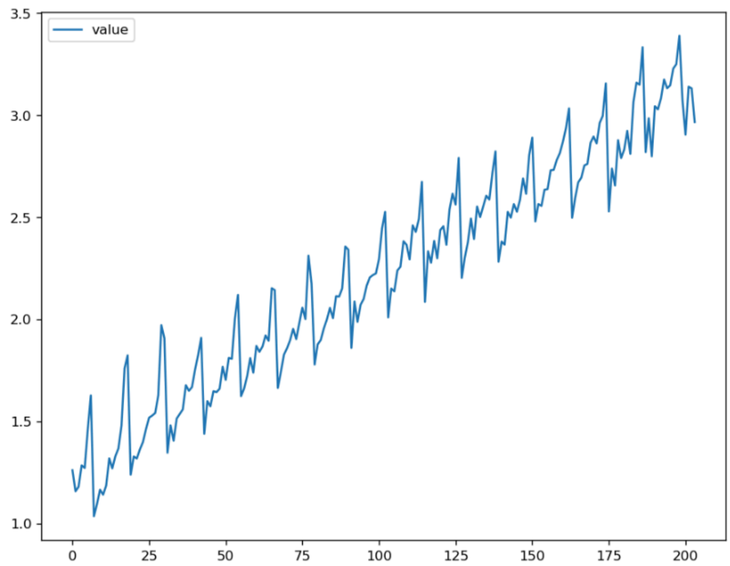
Partie 2 : Mise en place du modèle ARIMA

A. Conditions du modèle ARIMA



Ici nous avons le graphique qui présente l'évolution des valeurs de notre dataset dans le temps. Premièrement, nous observons sur notre graphique un schéma régulier tous les 12 mois ce qui peut suggérer une saisonnalité. On remarque également que c'est un modèle multiplicatif car la variabilité augmente avec le temps. On constate aussi qu'il y a une tendance croissante. On peut donc en conclure que notre série chronologique n'est pas stationnaire.
Nous souhaitons utiliser un modèle ARIMA est pour le faire notre série doit respecter les conditions suivantes: Être stationnaire, c'est-à-dire avoir une moyenne et une variance constante dans le temps.
Pour corriger la non-stationnarité de la variance, nous allons utiliser une transformation de type logarithmique afin de rendre la variance constante dans le temps.
Puis, pour éliminer les tendances, nous allons devoir remplacer la série d'origine par la série des différences adjacentes(différenciation).

B. transformation de la variance



Nous avons passé notre série au logarithme, nous constatons maintenant que notre variance semble être constante. Donc il nous reste plus qu'à tester la stationnarité de notre série pour plusieurs ordre de différenciation afin d'éliminer la tendance.

C. Test de la stationnarité(adfuller test) et niveau de différentiation (paramètre d)

Voici les résultats statistiques du test de stationnarité:

ADF Statistic: -0.988733
p-value: 0.757351

Nous constatons que la P-Valeur étant supérieur au seuil de confiance (0.05) donc nous ne rejetons pas l'hypothèse nulle selon laquelle la série est non-stationnaire. Donc notre série est non-stationnaire. Nous allons donc rechercher le niveau de différenciation nécessaire pour la rendre stationnaire.

Résultat différenciation d'ordre 1 :

différenciation (1) : ADF Statistic: -4.519432
différenciation (1) : p-value: 0.000181

Dans le cas de la première différenciation, on remarque que la P-Valeur est inférieur au seuil de confiance 0.05, dans ce cas nous pouvons rejeter l'hypothèse nulle. La série chronologique est donc stationnaire après une première différenciation. Cependant nous pouvons sans doute effectuer une autre différenciation afin de produire un série encore plus stable, en faisant attention de ne pas sur-différenciée.

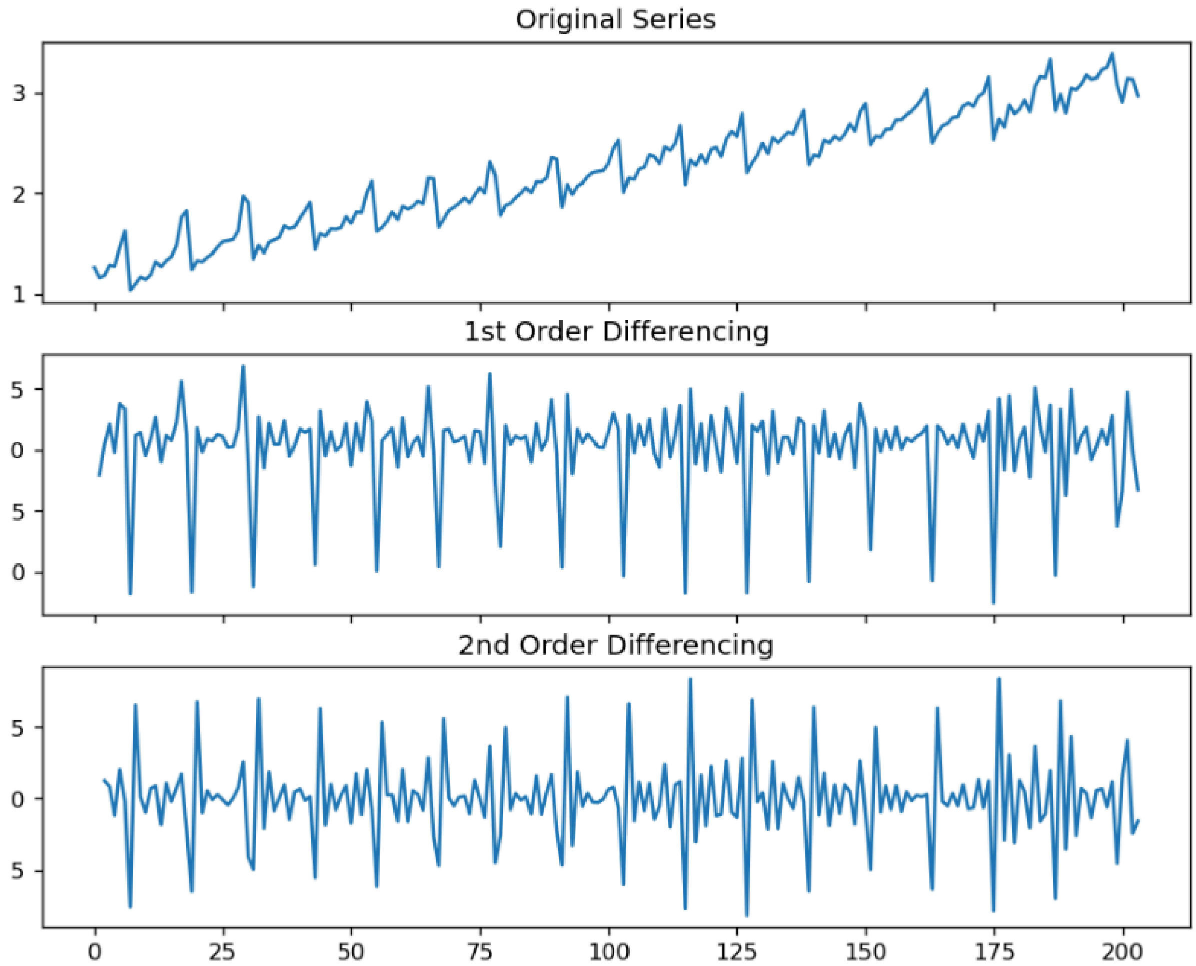
Résultat différenciation d'ordre 1 :

différenciation (2) : ADF Statistic: -10.037294
différenciation (2) : p-value: 0.000000

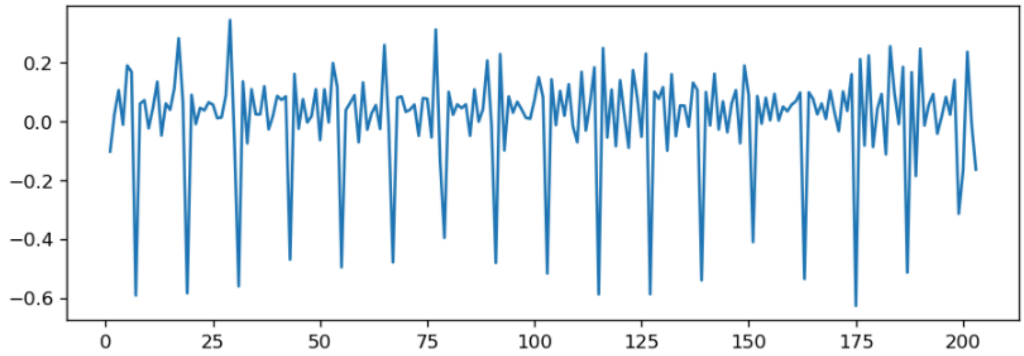
Dans le cas de la deuxième différenciation la P-valeur est toujours inférieur à 0.05 mais à un degré supérieur car le résultat est nul. Cela améliore la stationnarité de la série comparé à la 1er différenciation.

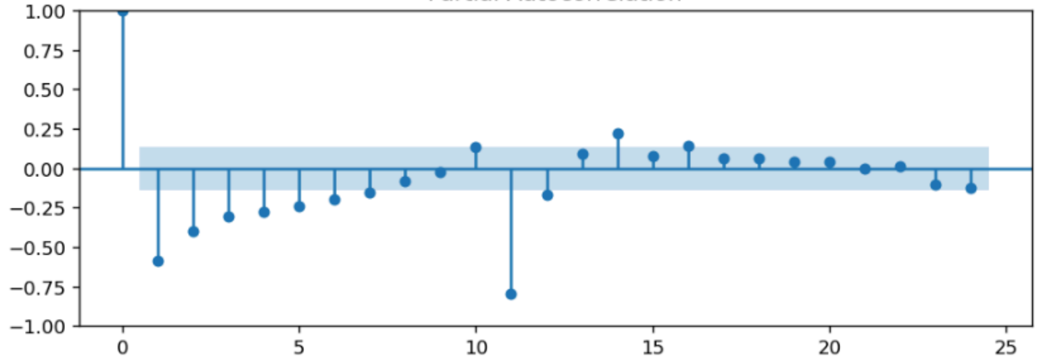
On peut voir ci-dessous les graphes correspondant à la série d'origine passée au logarithme et aux deux différenciations. Nous constatons que pour les serie différencier, toute la tendance a été supprimé, cependant pour l'ordre 2, les valeur de la série sont répartie de manière uniforme sur l'axe Y (entre -5 et 5)

Pour conclure au niveau des différenciations, si nous devions choisir un des niveaux de différenciation, nous aurions opté pour la différenciation de niveau 1 car elle est plus simple à élaborer et elle garantit déjà une très bonne fiabilité au niveau de la stationnarité de la série.



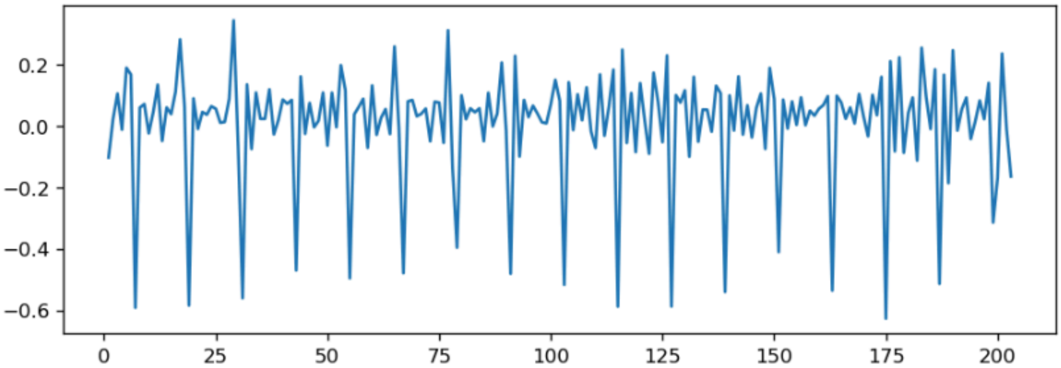
D. recherche de la valeur du paramètre p (de AR(p))

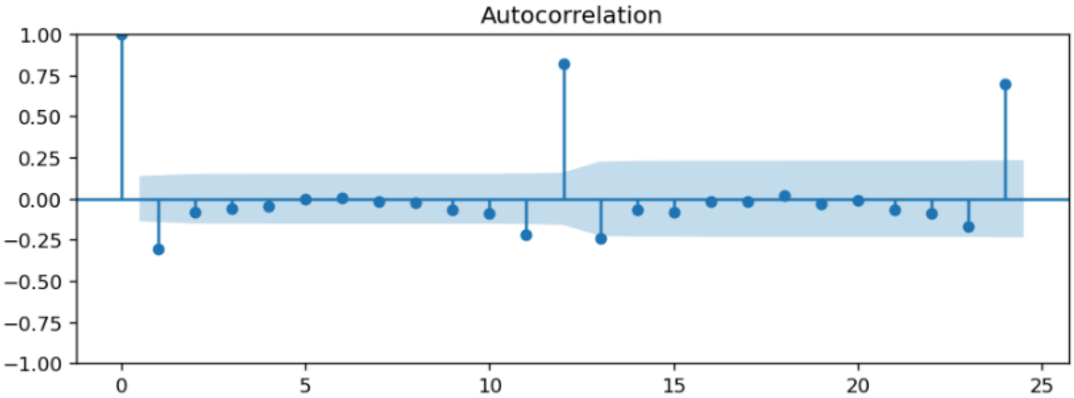




Ici nous avons le graphe du PACF (partial autocorrelation function)permettant de déterminer le nombre de termes autorégressifs pour une différenciation d'ordre 1. On observe une décroissance rapide des coefficients de corrélation partielle après le premier retard, et les points de l'autocorrélation partiel commencent à tendre vers 0 à partir 1, la valeur actuelle de la série est principalement liée à sa valeur précédente. Nous pouvons également noter qu'il y a un pic significatif à 12, cela est probablement dû au fait que nos données soit saisonnière (12 mois), mais les bar de l'autocorrélation reste généralement dans notre zone de confiance(proche de 0). Pour notre modèle ARIMA, nous choisissons donc **p = 1**. recherche de la valeur du paramètre q (de MA(q))

E. Recherche de la valeur du paramètre q (de MA(q))





Ici nous avons le graphe de ACF(autocorrelation function) permettant de déduire le nombre de moyennes mobiles pour un modèle ARIMA. Sur ce graphique on observe également une décroissance rapide des coefficients d'auto-corrélation après le premier retard, et les points de l'autocorrélation partiel commencent à tendre vers 0 à partir 1. Cela signifie que la valeur à l'instant t de la série est fortement influencée par la valeur de l'erreur à l'instant t-1 et très faiblement par les erreurs des autres instants. Cependant on constate également des pic au niveau de 12, 24 etc…, ces pic sont certainement dus au fait que notre série soit saisonnière, excepté ces pic qui sont significatifs, les autres valeurs sont bien dans notre zone de confiance(autour de 0).

F. Mise en place des modèle ARIMA(1,1,1) et ARIMA(1,1,2)

- **ARIMA(1,1,1)**

```
==============================================================================
Dep. Variable:                  value   No. Observations:                  204
Model:                 ARIMA(1, 1, 1)   Log Likelihood                  82.829
Date:                Sun, 26 Jan 2025   AIC                           -159.658
Time:                        18:20:22   BIC                           -149.719
Sample:                             0   HQIC                          -155.637
                                - 204
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3360      0.154      2.175      0.030       0.033       0.639
ma.L1         -0.8327      0.081    -10.292      0.000      -0.991      -0.674
sigma2         0.0258      0.002     10.332      0.000       0.021       0.031
===================================================================================
Ljung-Box (L1) (Q):                   0.54   Jarque-Bera (JB):                68.13
Prob(Q):                              0.46   Prob(JB):                         0.00
Heteroskedasticity (H):               0.90   Skew:                            -1.06
Prob(H) (two-sided):                  0.68   Kurtosis:                         4.88
===================================================================================
```

Nous avons obtenu le résumé statistique ci-dessus pour le modèle ARIMA(1,1,1). On constate au niveau des P-Valeur que les paramètres de notre sont toutes inférieures au seuil de confiance 0.05 ce qui signifie que les paramètres sont bien significatifs. On remarque aussi que le AIC est de -159.7, le BIC de -149.7 et le HQIC de -155.6. Ces valeurs sont considérablement faibles et cela peut indiquer que le modèle ARIMA(1,1,1) est performant. Nous les comparerons aux valeurs obtenues pour le modèle ARIMA(1,1,2).

- **ARIMA(1,1,2)**

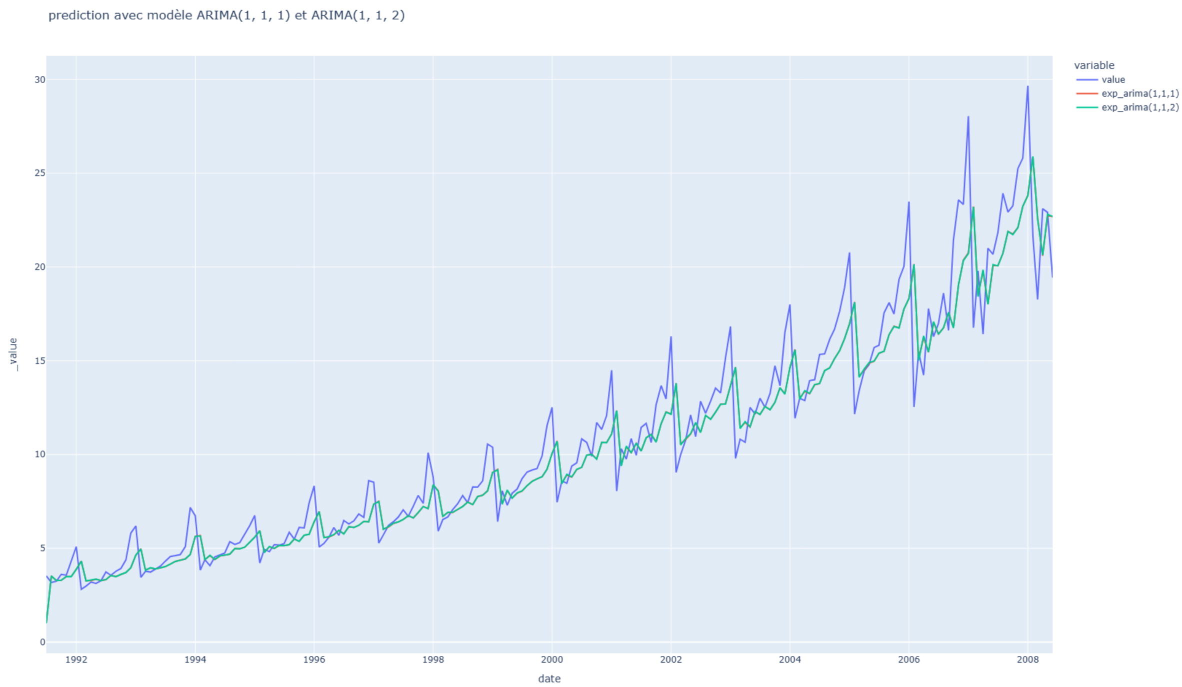
```
==============================================================================
Dep. Variable:                  value   No. Observations:                  204
Model:                 ARIMA(1, 1, 2)   Log Likelihood                  82.849
Date:                Sun, 26 Jan 2025   AIC                           -157.698
Time:                        18:26:50   BIC                           -144.445
Sample:                             0   HQIC                          -152.336
                                - 204
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2994      0.421      0.710      0.477      -0.527       1.125
ma.L1         -0.7924      0.432     -1.836      0.066      -1.639       0.054
ma.L2         -0.0314      0.301     -0.104      0.917      -0.621       0.558
sigma2         0.0258      0.003     10.307      0.000       0.021       0.031
===================================================================================
Ljung-Box (L1) (Q):                   0.65   Jarque-Bera (JB):                66.13
Prob(Q):                              0.42   Prob(JB):                         0.00
Heteroskedasticity (H):               0.91   Skew:                            -1.05
Prob(H) (two-sided):                  0.70   Kurtosis:                         4.85
===================================================================================
```

Ici nous avons le résumé statistique pour le modèle ARIMA(1,1,2). On constate au niveau des P valeurs, des valeurs très peu significatives notamment pour le AR et le MA2 car elles sont fortement supérieures à 0.05. Cela signifie que ces coefficient ne sont pas significatifs, et peuvent être supprimés. On remarque aussi que le AIC est de -157.7, le BIC de -144.4 et le HQIC de -152.3.

G. Comparaison

Le modèle ARIMA(1,1,2), comparé au modèle ARIMA(1,1,1) les P-valeurs laissent supposer que le modèle (1,1,2) est moins fiable que le modèle(1,1,1). De plus si l'on compare cette fois-ci les AIC, les BIC et les HQIC, les valeurs du modèle (1,1,2) sont toutes supérieures à celle du modèle (1,1,1), or on sait que le meilleur modèle est celui dont les valeurs sont les plus petites. On peut donc conclure que le modèle le plus pertinent à utiliser est le modèle ARIMA (1,1,1).

Pour avoir une meilleure visualisation de la situation nous allons effectuer un graphique sur lequel nous pourrons explicitement comparer les 2 modèles ainsi que les valeurs brutes.



De plus, sur le graphique nous constatons que les 2 modèles ARIMA (1,1,1,) et (1,1,2) se superposent quasi-parfaitement, et que tout deux ils prédisent relativement bien les valeurs réelles de notre dataset. Malgré tout, on peut remarquer que les prédictions sont généralement inférieures aux valeurs réelles surtout au mois de Janvier. Cela pourrait s'expliquer par le fait que le modèle ARIMA ne prend pas en compte la saisonnalité et donc ne laisse pas transparaître totalement les variations saisonnières(pour les variations saisonnières, nous devons utiliser le modèle SARIMA). Avec ce graphique seul on ne peut donc pas vraiment affirmer quel modèle entre le ARIMA (1,1,1) et le (1,1,2) est le meilleur, les deux semble parfaitement égaux (similaire).
Cependant, dans ce cas précis nous optons pour le modèle **ARIMA(1,1,1)** car il **présente le moins de complexité** dans le cadre de sa réalisation mais présente également de **meilleures statistiques (AIC,BIC,HQIC)** par rapport au modèle ARIMA(1,1,2).