Master 1 MIASHS

Option :

Méthodes quantitatives et modélisations pour l'entreprise

Méthode de Prévion

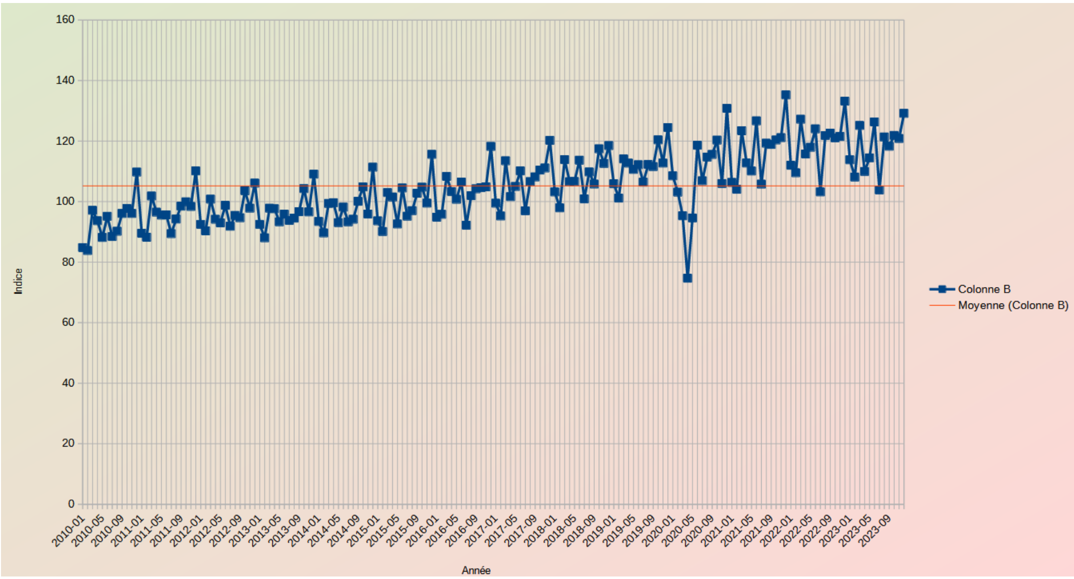
Indice de volume des ventes - Ensemble du Commerce en France

Travail réalisé par :
HADAD Ahmed Ali
Avril 2024

INTRODUCTION

L'analyse des données est cruciale pour comprendre les fluctuations et les tendances dans différents secteurs, afin de modéliser un modèle nous permettant de faire des prédictions dans le futur. Dans cette étude, nous examinerons la série chronologique de l'**Indice de volume des ventes dans l'ensemble du Commerce en France**, accessible en cliquant sur le lien suivant : INSEE - Indice de volume des ventes

a) **Indice de volume des ventes** : cet indice mesure les variations du volume des ventes dans le secteur du commerce. Il est utilisé pour évaluer les tendances générales des ventes de biens ou de services au fil du temps.
b) **Ensemble du Commerce** : cela indique que l'indice englobe l'ensemble du secteur du commerce, ce qui inclut généralement les activités de vente de détail, de vente en gros, ainsi que d'autres formes de commerce.



Une observation clé de notre analyse réside dans la légère tendance à la croissance de la série. Cette observation reflète une progression positive dans le secteur commercial au fil du temps.
Cependant, un événement s'est produit en **2020**, où nous observons une baisse marquée de l'indice. Cette anomalie dans la série chronologique peut être attribuée à l'effet direct de la pandémie **COVID-19**, perturbant les schémas commerciaux habituels et affectant de manière significative les performances économiques de la France.

Pour élaborer un modèle prédictif fiable, nous parcourons trois méthodes et comparons leurs puissances de prédiction :

→ Méthodes de décomposition

→ Le lissage : Holt, Holt-Winters additive et Holt-Winters multiplicative

→ Méthode de Box-Jenkins.

Information : Le travail a été mené sous le **logiciel R** pour la méthode de **Box-Jenkins** et sous **Excel** pour les deux **autres méthodes**.

Table des matières

1

Méthode de décomposition

Les méthodes de décomposition visent à désaisonnaliser la séries chronologique et à en extraire une tendance globale, ce qui est crucial pour effectuer des prévisions précises sur l'évolution future. Ainsi, l'objectif dans cette partie est l'élaboration d'une série sans variations saisonnières.
Pour ce faire, nous avons généralement le choix entre trois principaux modèles de décomposition :

a. Modèle additif,

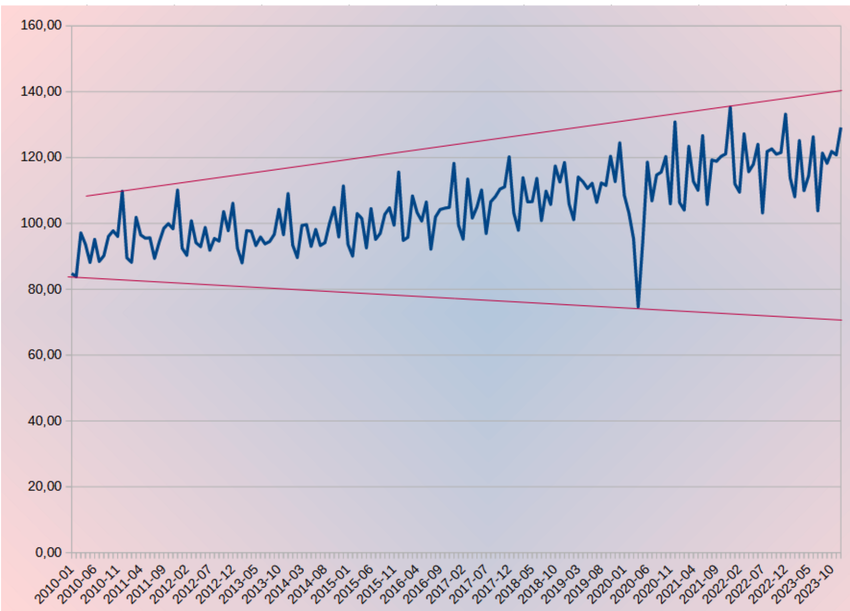
b. Modèle multiplicatif,

c. Modèle mixte,

Dans un premier temps, nous allons identifier le modèle approprié à cette série chronologique.

1.1 Choix de la méthode

La **méthode graphique** consiste à tracer deux courbes joignant les maxima (minima) distants des périodes. Si ces deux courbes sont à peu près parallèles, on choisit un modèle additif, dans le cas contraire nous adaptons le modèle multiplicatif :



Ainsi les deux droites ne sont pas parallèles alors un modèle multiplicatif est privilégié.

1.2 Méthode de decomposition : Modèle multiplicatif

Considérons une série $(X_t, t = 1, \cdots, T)$. Un modèle multiplicatif est un modèle de décompositon de la série chronologique qui s'écrit sous la forme :

$$X_t = T_t \times I_t \times \epsilon_t$$

→ T_t est une tendance déterministe

→ I_t est une composante saisonnière déterministe

→ ϵ_t est la composante résiduelle aléatoire.

Nous allons désaisonnaliser la série, série sans variations saisonnières (**CVS**) :

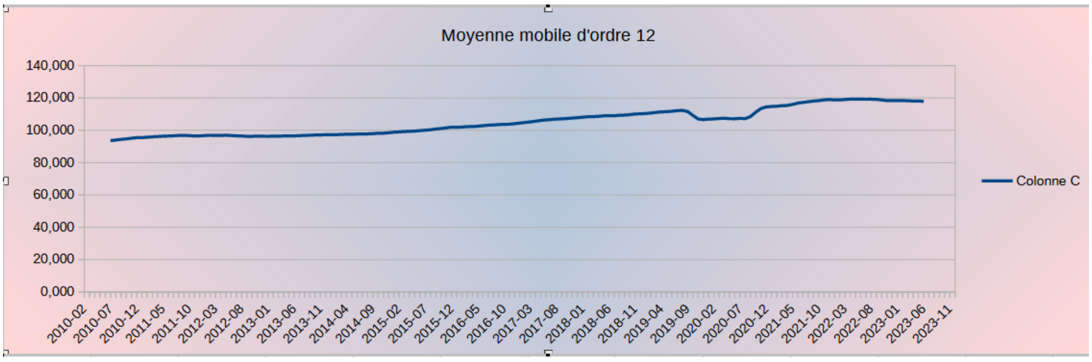
$$Y_t = \frac{X_t}{I_t}$$

Pour se faire nous pouvons commencé par déterminer la tendance T_t.

1.2.1 Calcul de la tendance par filtrage par moyenne mobile

Pour des données mensuelles, on considère la moyenne d'ordre 12 :

$$T_t = \frac{1}{12} \times \left(\frac{X_{t-6}}{2} + X_{t-5} + \cdots + X_{t-1} + X_t + X_{t+1} + \cdots + X_{t+5} + \frac{X_{t+6}}{2} \right)$$



Cette moyenne mobile a un effet lissant sur la série, atténuant ainsi le pic d'irrégularité qui était apparu précédemment. De plus, elle permet d'éliminer les légères variations saisonnières de la série. En conséquence, nous observons une tendance croissante globale se dégageant davantage.

1.2.2 Calcul des coefficients saisonniers

Pour estimer les variations saisonnières, on va enlever la tendance T_t. Pour un modèle multiplicatif, la série sans tandance est donnée par :

$$\text{série sans tandance : } = \frac{X_t}{T_t}$$

Ensuite, on met cette série sans tendance sur le tableau de **Buys-Ballot** qui va servir à calculer les coefficients saisonniers I_t :

→ On calcule la moyenne des données sans tendance pour chaque mois.

→ On corrige les coefficients obtenus pour qu'ils aient une moyenne nulle.

→ On obtient les coefficients saisonniers I_t en soustrayant aux moyennes mensuelles.

Tableau de Buys-Ballot pour les données sans tendance : Xt/Tt

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	moyenne pa	I
Janvier		0,937	0,956	0,960	0,960	0,947	0,931	0,952	0,955	0,959	1,015	0,927	0,943	0,961	0,954	0,954
février		0,922	0,932	0,914	0,921	0,910	0,939	0,908	0,903	0,914	0,963	0,906	0,922	0,913	0,921	0,920
mars		1,061	1,041	1,015	1,020	1,038	1,059	1,079	1,050	1,027	0,888	1,071	1,068	1,058	1,037	1,036
avril		1,004	0,973	1,012	1,021	1,022	1,010	0,962	0,981	1,012	0,695	0,977	0,970	0,931	0,967	0,967
mai		0,991	0,959	0,967	0,953	0,931	0,983	0,990	0,978	0,992	0,883	0,949	0,989	0,969	0,964	0,964
juin		0,991	1,021	0,993	1,005	1,047	1,035	1,035	1,043	1,003	1,107	1,084	1,040	1,071	1,037	1,036
juillet	0,945	0,925	0,951	0,970	0,954	0,951	0,894	0,909	0,925	0,949	0,996	0,902	0,866		0,934	0,933
août	0,960	0,973	0,989	0,976	0,963	0,967	0,987	0,997	1,005	0,999	1,069	1,013	1,022		0,994	0,994
septembre	1,018	1,017	0,983	0,997	1,022	1,019	1,007	1,010	0,967	0,999	1,065	1,007	1,030		1,011	1,011
octobre	1,032	1,033	1,076	1,074	1,068	1,036	1,009	1,029	1,071	1,102	1,081	1,017	1,019		1,050	1,050
novembre	1,010	1,019	1,014	0,994	0,976	0,980	1,011	1,033	1,023	1,053	0,934	1,019	1,027		1,007	1,007
décembre	1,150	1,141	1,101	1,121	1,131	1,134	1,136	1,115	1,075	1,167	1,143	1,136	1,125		1,129	1,129
											moyenne(moyenne par mois)				1,0003	

Coefficient saisonniers : I

janvier	février	mars	avril	mai	juin	juillet	août	eptemb	ctobr	ovembr	écembre
0,954	0,920	1,036	0,967	0,964	1,036	0,933	0,994	1,011	1,050	1,007	1,129

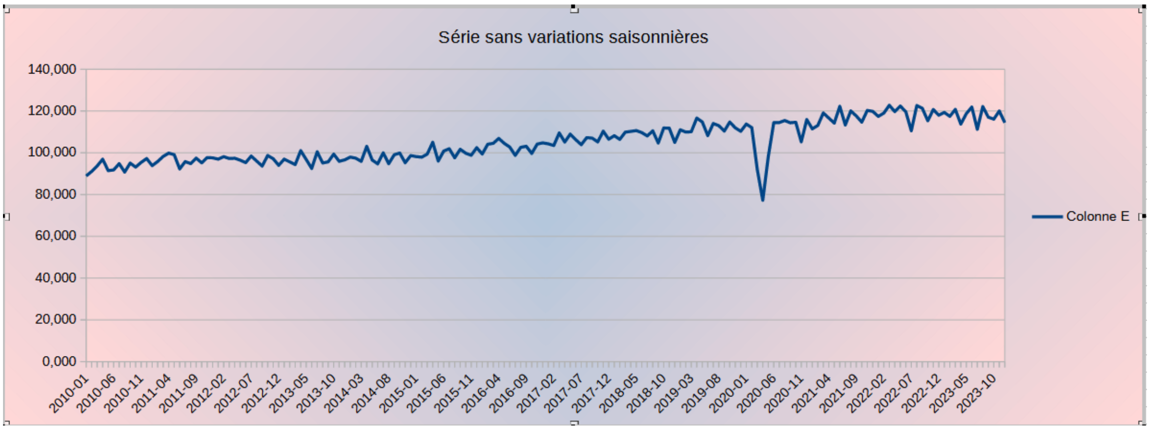
1.2.3 Série sans variations saisonnières : CVS

Pour obtenir la série corrigée des variations saisonnières (**CVS**), avec un modèle multiplicatif, on calcule :

$$\mathbf{CVS} = \frac{X_t}{I_t}$$

Tableau de Buys-Ballot pour la série CVS: Xt/It

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Janvier	88,897	93,846	96,939	96,949	97,945	98,165	99,434	104,278	108,178	111,019	113,756	111,481	117,415	119,323
février	91,086	95,845	98,149	95,650	97,399	97,899	104,081	103,484	106,407	109,905	112,035	113,078	118,989	117,457
mars	93,703	98,258	97,254	94,369	95,894	99,358	104,520	109,490	109,875	110,078	91,967	119,071	122,738	120,750
avril	96,908	99,836	97,394	101,015	103,064	105,019	106,881	105,143	110,244	116,596	77,262	116,627	119,710	113,751
mai	91,456	99,103	96,447	96,800	96,519	96,063	104,529	108,928	110,619	114,780	98,138	114,219	122,353	118,733
juin	91,795	92,268	95,288	92,490	94,709	100,827	102,747	106,269	109,676	108,238	114,452	122,220	119,673	121,873
juillet	94,772	95,779	98,414	100,471	99,946	101,950	98,768	103,878	108,078	114,002	114,495	113,317	110,553	111,249
août	90,772	94,838	96,026	95,130	94,768	97,616	102,618	107,268	110,499	112,995	115,440	120,050	122,616	122,123
septembre	95,051	97,446	93,656	95,665	99,029	101,671	103,126	107,004	104,679	110,389	114,366	117,632	121,332	117,038
octobre	93,137	95,195	98,692	99,349	99,883	99,797	99,692	105,199	111,879	114,708	114,623	114,680	115,328	116,100
novembre	95,409	97,703	97,147	95,925	95,240	98,795	104,148	110,325	111,815	111,944	105,241	120,286	120,694	120,008
décembre	97,237	97,565	94,003	96,617	98,663	102,420	104,741	106,513	104,980	110,279	115,896	119,821	117,987	114,416



Ainsi, les mouvements saisonniers se sont quasiment dégradés.

2

Le lissage

Les méthodes de lissage, telles que la méthode de **Holt** et **Holt-Winters**, sont des méthodes très puissantes utilisées pour analyser et prévoir les séries chronologiques.

2.1 Méthode de Holt

Soit la série chronologique $(X_t, t = 1, \cdots, T)$.

La méthode de **HOLT** permet de construire des prévisions linéaires de la forme :

$$\hat{X}_t = S_t + h \times T_t \qquad \text{où } h = 1, 2, \cdots$$

→ S_t est la composante de niveau

→ T_t est la composante de tendance

On détermine les deux composantes par :

$$\begin{cases} S_t = \alpha \times X_t + (1-\alpha) \times (S_{t-1} + T_{t-1}) \\ T_t = \gamma \times (S_t - S_{t-1}) + (1-\gamma) \times T_{t-1} \end{cases}$$

Elles dépendent de deux paramètres α et γ compris entre 0 et 1.

initialisation :

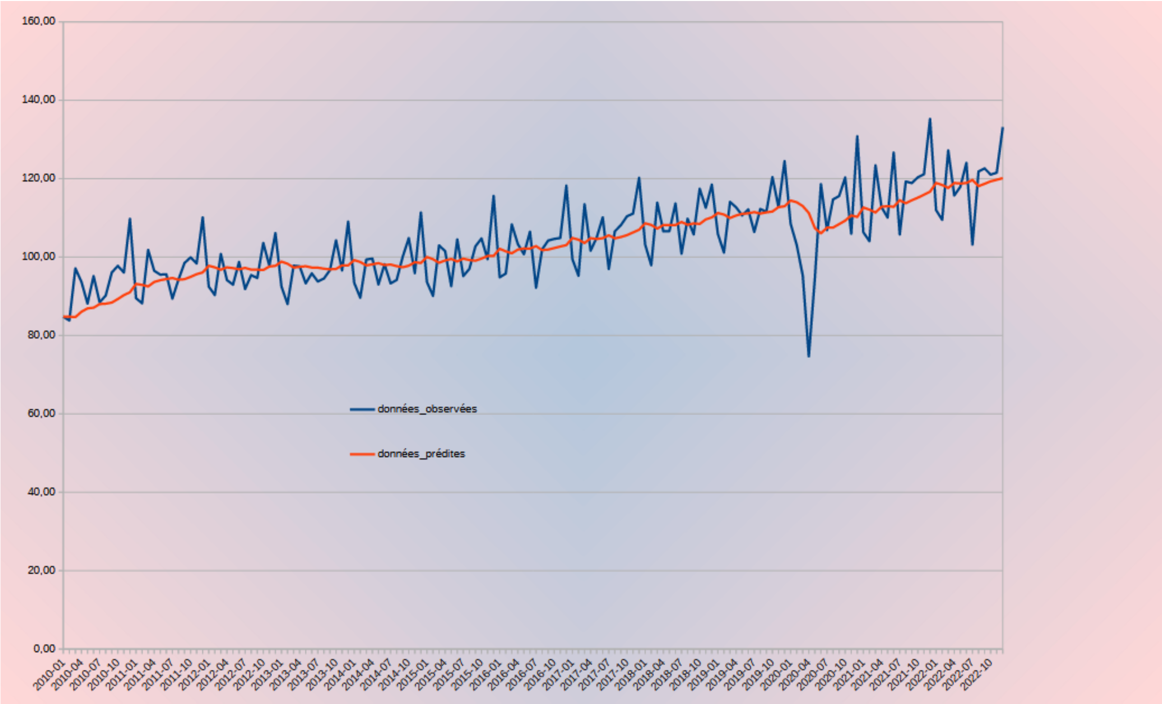
L'initialisation des relations est faite en choisissant :

$$\begin{cases} S_0 = X_1 \Longrightarrow S_1 = S_0 \\ T_0 = 0 \end{cases}$$

En appliqaunt la méthode, les valeurs trouvées par le solveur exel sont :

$$\begin{cases} \alpha = 0.108 \\ \gamma = 0.017 \\ \text{MAPE} = 5.59\% \end{cases}$$

Graphique représentant les valeurs prédites et les valeurs observées :



2.2 Méthode de Holt-Winters additive

Partant de la même idée de **Holt**, l'idée est d'introduire des coefficients saisonniers. La prévision est de la forme :

$$\hat{X}_t = S_t + h \times T_t + I_{t+h-s} \qquad \text{où } h = 1, 2, \cdots$$

I_t est un coefficient saisonnier, **s** est la saison, ici $s = 12$: On prend une saison égale une année. Les trois composantes sont calculées par :

$$\begin{cases} S_t = \alpha \times (X_t - I_{t-s}) + (1-\alpha) \times (S_{t-1} + T_{t-1}) \\ T_t = \gamma \times (S_t - S_{t-1}) + (1-\gamma) \times T_{t-1} \\ I_t = \delta \times (X_t - S_t) + (1-\delta) \times I_{t-s} \end{cases}$$

Elles dépendent de trois paramètres α, γ et δ compris entre 0 et 1.

Initialisation :

L'initialisation des relations est faite en prennant :

→ Les 12 valeurs initiales $t = -11, -10, \cdots, 0$

→ Les 12 valeurs initiales pour S : moyenne des 12 premières observations (celles de l'année 2010).

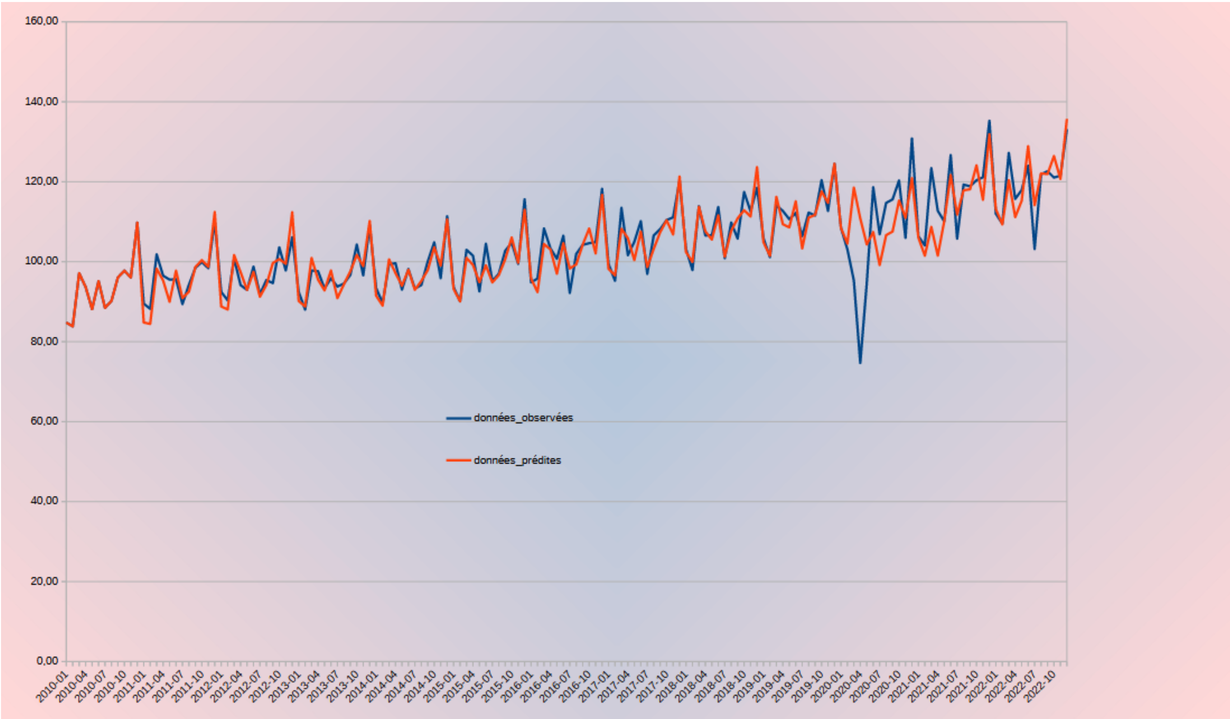
→ Les 12 valeurs initiales pour T, on prend 0

→ Les 12 valeurs initiales pour les coefficients saisonniers I : $X - S$ écart entre les 12 premières données (celles de 2010) et la valeur initiale pour S.

En appliqaunt la méthode, les valeurs trouvées par le solveur exel sont :

$$\begin{cases} \alpha = 0.126 \\ \gamma = 0.034 \\ \delta = 0.380 \\ \text{MAPE} = 2.75\% \end{cases}$$

Graphique représentant les valeurs prédites et les valeurs observées :



2.3 Méthode de Holt-Winters multiplicatif

La prevision est de la forme :

$$\hat{X}_t = (S_t + h \times T_t) \times I_{t+h-s} \qquad \text{où } h = 1, 2, \cdots$$

Les trois composantes sont calculées par :

$$\begin{cases} S_t = \alpha \times (\dfrac{X_t}{I_{t-s}}) + (1-\alpha) \times (S_{t-1} + T_{t-1}) \\ T_t = \gamma \times (S_t - S_{t-1}) + (1-\gamma) \times T_{t-1} \\ I_t = \delta \times (\dfrac{X_t}{S_t}) + (1-\delta) \times I_{t-s} \end{cases}$$

Initialisation :

L'initialisation des relations est faite en prennant :

→ Les 12 valeurs initiales $t = -11, -10, \cdots, 0$

→ Les 12 valeurs initiales pour S : moyenne des 12 premières observations (celles de l'année 2010).

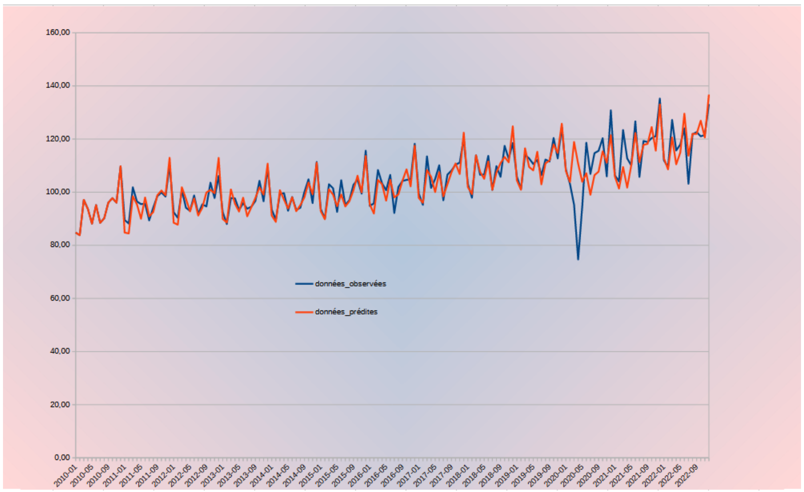
→ Les 12 valeurs initiales pour T, on prend 0

→ Les 12 valeurs initiales pour les coefficients saisonniers I : $\frac{X}{S}$ rapport entre les 12 premières données (celles de 2010) et la valeur initiale pour S.

En appliqaunt la méthode, les valeurs trouvées par le solveur exel sont :

$$\begin{cases} \alpha = 0.132 \\ \gamma = 0.029 \\ \delta = 0.354 \\ \text{MAPE} = 2.80\% \end{cases}$$

Graphique représentant les valeurs prédites et les valeurs observées :



2.4 Meilleure méthode

En comparant les **MAPE** (Mean Absolute Percentage Error) des differentes méthodes, le **MAPE** minimal est cel de la méthode de **Holt-Winters additive** de valeur $2,75\%$ contres $5,59\%$ et $2,80\%$ de la méthode de **Holt** et **Holt-Winters multiplicative** respectivement.

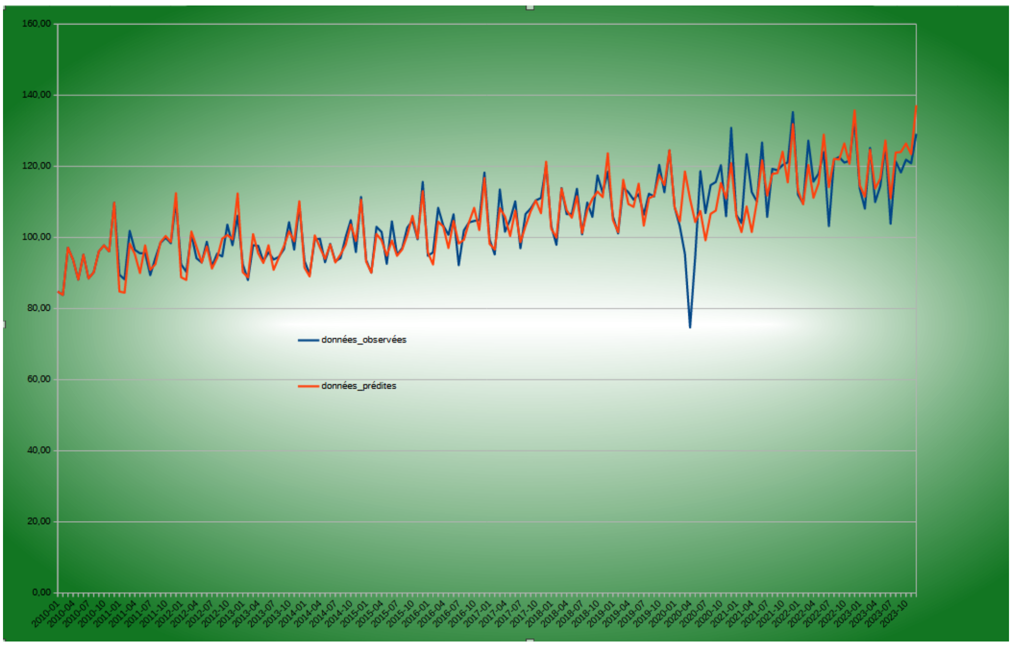
2.5 Prévision

Nous appliquons la meilleure méthode de lissage obtenue, à savoir **Holt-Winters additive**, pour prévoir les douze dernières données, qui correspondent aux données de l'année **2023**. Ensuite, nous calculerons le **MAPE**.

Performance de la prédiction :

En comparant les valeurs prédites avec les données observées de **2023**, nous obtenons un **MAPE** de **3,02%**, ce qui indique une marge d'erreur acceptable.

Graphique final sur l'ensemble de l'échantillon comportant les observations et les prévisions :

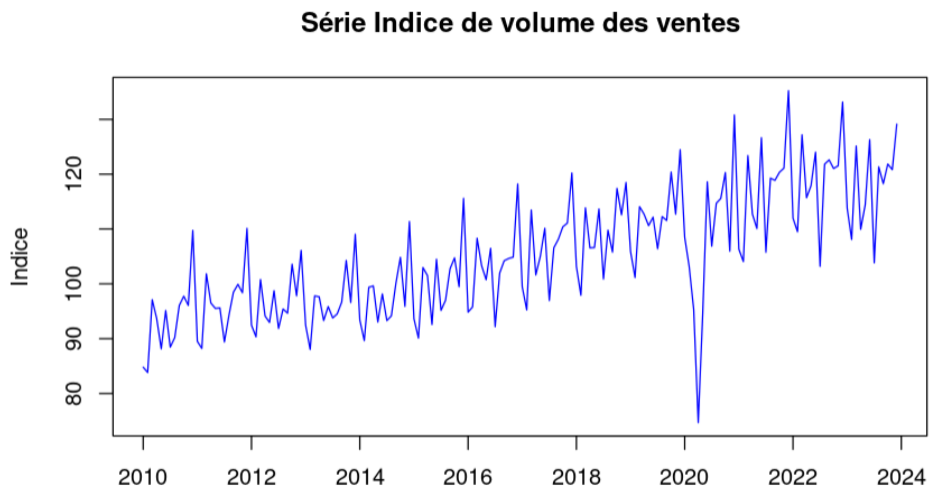


3
Méthode de Box-Jenkins

La méthode de **Box-Jenkins** est utilisée pour trouver un modèle **ARMA** capable de prédire la suite de la série chronologique. Bien qu'elle soit considérée comme l'une des méthodes de prévision les plus robustes, elle nécessite plusieurs étapes pour mettre en place le bon modèle :

3.1 Analyse des données

Cette étape repose sur l'examen graphique des données ou le **test de Dickey-Fuller** afin de vérifier si la série est stationnaire. Si ce n'est pas le cas, il est nécessaire de la rendre stationnaire en utilisant par exemple la différenciation.



La série, telle qu'elle se présente, révèle une tendance croissante qui suggère qu'elle est non stationnaire. Afin de confirmer cette hypothèse, nous allons effectuer un **test de Dickey-Fuller** pour vérifier la stationnarité de la série. Deux hypothèses s'affrontent :

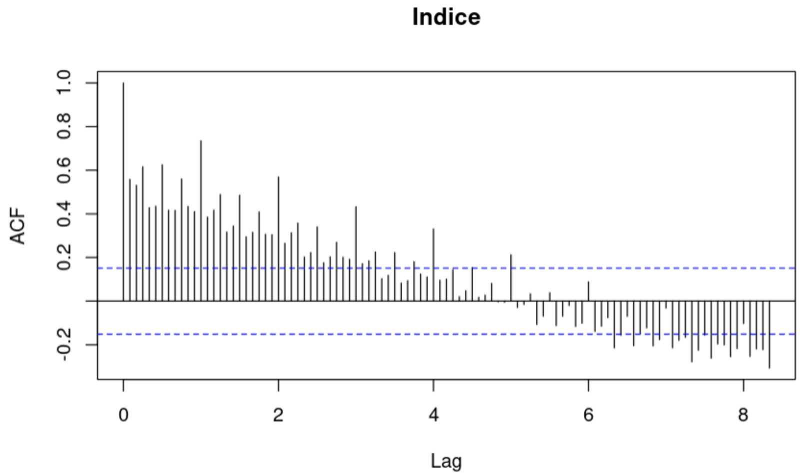
$$\begin{cases} H_0 : & \text{la série est non-stationnaire.} \\ H_1 : & \text{la série est stationnaire.} \end{cases}$$

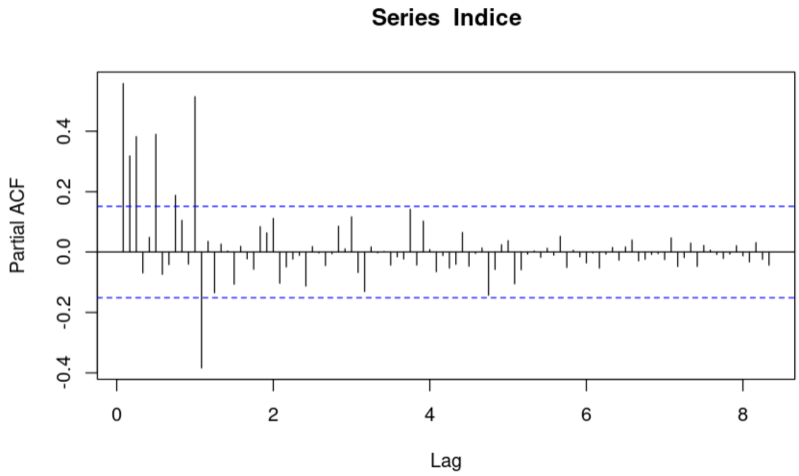
```
        Augmented Dickey-Fuller Test

data:  Indice
Dickey-Fuller = -4.4443, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Avec une **p-value** de 0.01, inférieure à un niveau de signification typique de 0.05, le test rejette l'hypothèse nulle. Ainsi, contrairement à ce que la tendance détectée pourrait nous faire croire quant à une série non stationnaire, la série est en réalité **stationnaire**.

3.2 Identification du modèle

Cette étape est basée sur l'analyse des autocorrélogrammes de la **série stationnaire** pour aboutir à un choix d'un modele **ARIMA(d, q, p)**.





Nous observons des fonctions d'autocorrélation (**ACF**) globalement décroissantes, avec des pics d'autocorrélation partielle (**PACF**) nuls à partir du 14ème pics. Ces deux critères, combinés au fait que nous avons confirmé que la série est stationnaire, nous permettent de conclure que le modèle adapté est un **modèle autorégressif d'ordre 13** (**AR(13)**).

3.3 Apprendre un modèle AR(13)

Nous avons réduit l'échantillon en enlevant les 12 dernières données, qui seront utilisées pour la validation du modèle.

L'idée est de diviser l'échantillon en deux : un **échantillon d'entraînement**, utilisé pour former le modèle et faire des prévisions, et un **échantillon de test** (les 12 dernières données), utilisé pour évaluer la qualité des prédictions du modèle construit.

```
Call:
arima(x = donnes_test, order = c(13, 0, 0))

Coefficients:
         ar1     ar2     ar3      ar4     ar5     ar6      ar7     ar8     ar9    ar10
      0.5120  0.0466  0.0470  -0.0793  0.0877  0.1176  -0.0557  0.0501  0.0532  0.0317
s.e.  0.0731  0.0606  0.0598   0.0593  0.0598  0.0581   0.0608  0.0600  0.0603  0.0599
         ar11    ar12     ar13  intercept
      -0.1293  0.7082  -0.4104   105.4364
s.e.   0.0605  0.0601   0.0750    10.2167

sigma^2 estimated as 24.71:  log likelihood = -477.34,  aic = 984.68
```

3.4 Validation du modèle

Le modèle choisi doit produire des résidus présentant un comportement de bruit blanc, tandis que ses coefficients doivent être significatifs.

3.4.1 Significativités des coefficients

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)    
ar1         0.511969   0.073099  7.0038 2.491e-12 ***
ar2         0.046584   0.060617  0.7685   0.44219    
ar3         0.047045   0.059779  0.7870   0.43129    
ar4        -0.079317   0.059342 -1.3366   0.18135    
ar5         0.087683   0.059847  1.4651   0.14289    
ar6         0.117645   0.058066  2.0261   0.04276 *  
ar7        -0.055725   0.060781 -0.9168   0.35924    
ar8         0.050100   0.060043  0.8344   0.40405    
ar9         0.053223   0.060331  0.8822   0.37768    
ar10        0.031695   0.059883  0.5293   0.59661    
ar11       -0.129342   0.060495 -2.1380   0.03251 *  
ar12        0.708185   0.060074 11.7886 < 2.2e-16 ***
ar13       -0.410446   0.074972 -5.4746 4.384e-08 ***
intercept 105.436369  10.216720 10.3200 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Il est à noter que seuls les coefficients **AR1**, **AR6**, **AR12**, **AR13** et l'**intercepte** sont significatifs. Par conséquent, nous allons concerver que les termes (retards) du modèle associés à ces coefficients.

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)    
ar1         0.535108   0.065714  8.1430 3.856e-16 ***
ar6         0.166879   0.049223  3.3903 0.0006982 ***
ar11       -0.033974   0.046968 -0.7233 0.4694755    
ar12        0.742797   0.054725 13.5732 < 2.2e-16 ***
ar13       -0.447974   0.072177 -6.2066 5.413e-10 ***
intercept 105.178402   7.063307 14.8908 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)    
ar1         0.526456   0.064894  8.1125 4.957e-16 ***
ar6         0.164035   0.048948  3.3512 0.0008046 ***
ar12        0.730509   0.052392 13.9432 < 2.2e-16 ***
ar13       -0.453244   0.071361 -6.3514 2.134e-10 ***
intercept 105.343736   7.735900 13.6175 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Alors le modèle **AR(13)** retenu est de la forme :

$$X_t = \varphi_1 X_{t-1} + \varphi_6 X_{t-6} + \varphi_1 X_{t-12} + \varphi_1 X_{t-13} + \epsilon_t$$

3.4.2 Les Résidus sont-ils des bruits blancs ?

Pour être considérés comme des bruits blancs, les résidus doivent avoir une moyenne nulle et ne pas être corrélés entre eux.

Les résidus sont-ils de moyenne nulle ?

On applique le **test de Student** de moyenne nulle :

$$\begin{cases} H_0 : & \mathbb{E}(\epsilon_t) = 0 \\ H_1 : & \mathbb{E}(\epsilon_t) \neq 0 \end{cases}$$

La statistique de test est de 1.684636, ce qui est inférieur à 1.96. Par conséquent, avec un niveau de signification typique de 0.05, nous ne rejetons pas l'hypothèse nulle (H_0), ce qui signifie que les résidus ont une moyenne différente nulle.

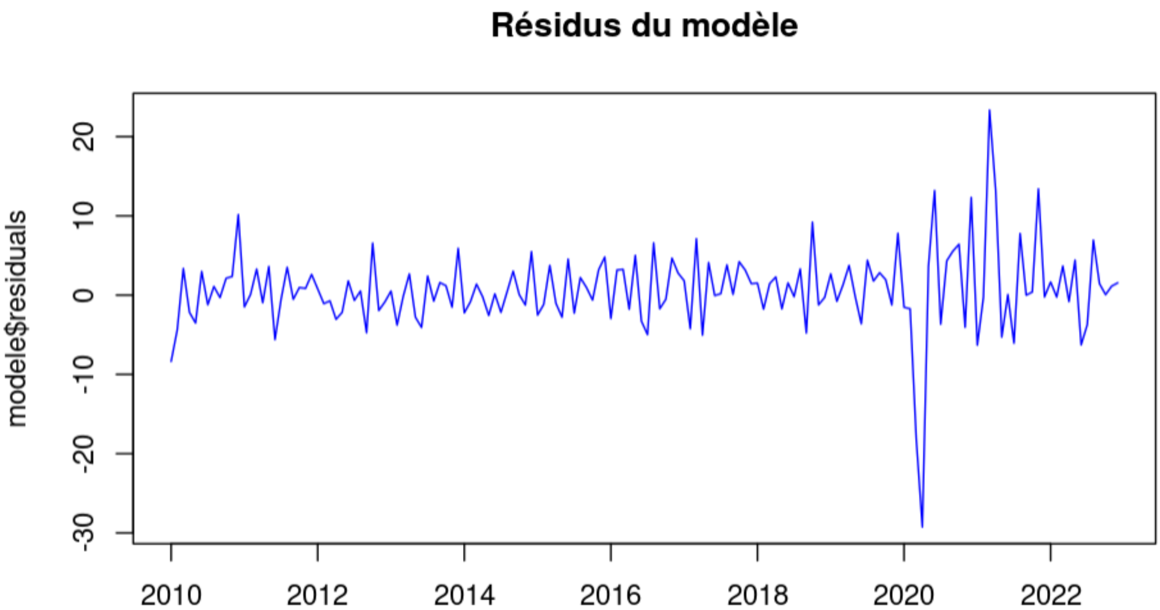
Les résidus sont-ils auto-corrélés ? test Portmanteau :

$$\begin{cases} H_0 : & \rho_1(\epsilon) = \rho_2(\epsilon) = \cdots = \rho_k(\epsilon) = 0 : \text{Les résidus sont indépendants et ne présentent pas d'autocorrélation.} \\ H_1 : & \exists \quad \rho_i(\epsilon) \quad \text{tel que} \quad \rho_i(\epsilon) \neq 0 : \text{Il y a de l'autocorrélation dans les résidus.} \end{cases}$$

```
        Box-Ljung test

data:  modele$residuals
X-squared = 8.9466, df = 10, p-value = 0.5372
```

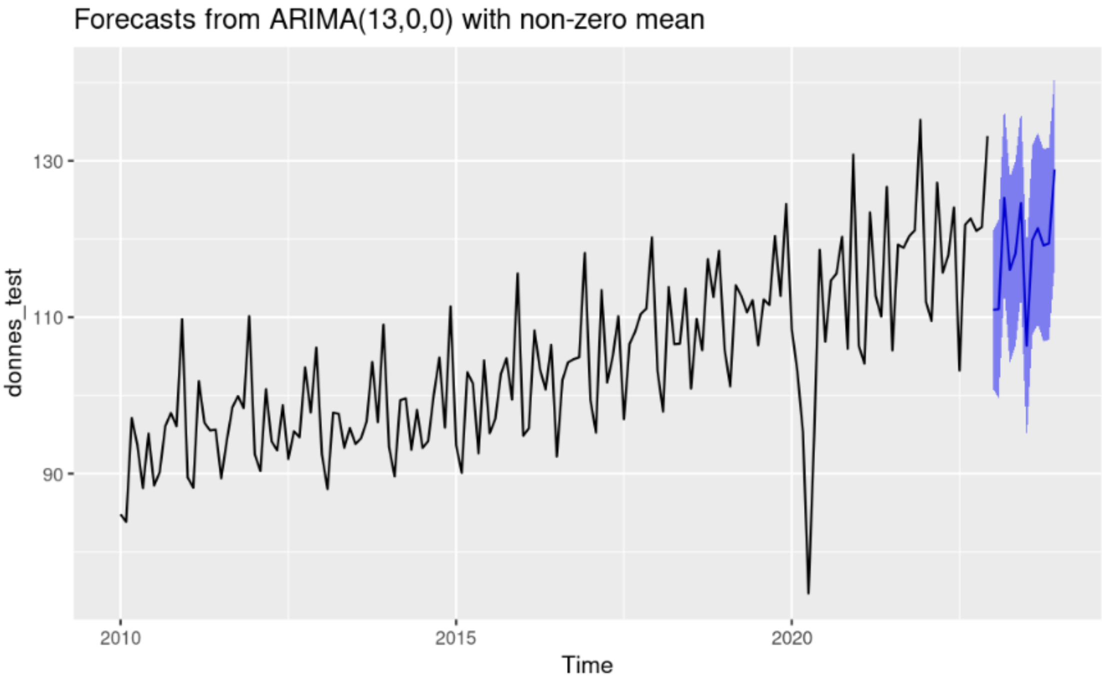
La **p-value** du **test de portmanteau** est de 0.537172 supérieure à 0.05, ce qui suggère qu'avec un niveau de significativité typique de 0.05, il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle d'absence d'autocorrélation des résidus.

Par conséquent, les résidus sont des bruits blancs.



3.5 Prévisions

Avec ce modèle obtenu, on va prédire les 12 futurs observations, représentées par la partie en couleur bleue.



3.6 Analyse de la capacité prédictive du modèle : MAPE

Le modèle a été construit à partir des données "*donnes_test*" : les données sans les 12 dernieres données de la base. Et puis on a prédit ces 12 données par le modèle.
On va mesurer la capacité prédictive du modèle en regardant le **MAPE** commis entre les données prédites et les données observées préalablment mits dans la table "*donnes_valid*".

Le **MAPE** du modele est obtenu par :

$$\text{MAPE} = \text{moyenne}\left(\left|\frac{\text{erreur}}{X : \text{données\_observées}}\right|\right)$$

$$\text{erreur} = \text{données\_observées} - \text{données\_prédites}$$

Pour les douze futurs données prédites, le modèle présente un **MAPE** de **2.1063838%**.

4 Conclusion

En comparant les deux méthodes de modélisation, le **lissage** et la méthode de **Box-Jenkins**, nous observons des performances différentes en termes de précision des prédictions.

Dans le cas du **lissage**, nous avons utilisé plusieurs variantes de la méthode de **Holt**, notamment la méthode de **Holt**, **Holt-Winters additive** et **Holt-Winters multiplicatif**. Les **MAPE** obtenus sont respectivement de **5.59%**, **2.75%**, et **2.80%**. Ces valeurs indiquent une précision variable des prédictions, avec une amélioration significative observée avec la méthode de **Holt-Winters additive**, qui a un **MAPE** optimal de **2.72%**.

En ce sens, on a appliqué ce meuilleur modèle retenu, **Holt-Winters additive**, pour prédire les 12 données futures (données de 2023) et on a eu un **MAPE** de **3.02%**. En revanche, la méthode de **Box-Jenkins** a produit un **MAPE** de **2.106838%** sur la prévision des données des 12 données futures, démontrant une précision supérieure par rapport au **Holt-Winters additive**.

Ainsi, sur la base de la comparaison des **MAPE**, la méthode de **Box-Jenkins** semble fournir des prédictions plus précises que les méthodes de **lissage** examinées dans cette analyse.

Voici le tableau illustrant les **données observées**, les **données prévues** par **Holt-Winters additive** et les **données prévues** par **Box-Jenkins**, avec leurs **MAPE** associés.

	Données observées	Prévision par Holt-Winters additive	Prévision par Box-Jenkins
2023-01	113,81	114,69	110,9446
2023-02	108,10	111,45	111,0278
2023-03	125,14	124,64	125,2479
2023-04	109,95	113,74	116,0636
2023-05	114,44	116,54	118,1383
2023-06	126,30	127,25	124,5825
2023-07	103,84	111,02	106,3526
2023-08	121,34	123,87	119,8268
2023-09	118,28	124,03	121,3816
2023-10	121,85	126,38	119,1813
2023-11	120,84	123,41	119,4373
2023-12	129,14	137,19	128,9102
MAPE		3,02 %	2.106838 %

5 Annexe

````
```{r}
library(readxl)

Indice <- read_excel("serie_010542960_05042024.xlsx",
                     col_types = c("text", "numeric"))
Indice = Indice[, -1]

Indice <-ts(data=Indice,start=c(2010,1), frequency=12)
```

# Faites un graphique de la série.

```{r}
plot(Indice)
```
# 1. Stationnarisation de la série

Rappélons que la prémiere étape de la méthode de  Box et Jenkins est la stationnarisation de la série ainsi que la correction des valeurs aberrantes.

# Appliquons un tyest de stationnarité pour verifier:
test de stionnarité avec la focntion adf.test:
$$H0 : \text{non stationnaire c-à-d il y'a une racine unitaire du polynome caracteristique.}$$
````

$$H0 : \text{non stationnaire c-à-d il y'a une racine unitaire du polynome caracteristique.}$$

````
$$H1 : \text{stationnaire c-à-d il n y'a pas une racine unitaire du polynome caracteristique.}$$
````

$$H1 : \text{stationnaire c-à-d il n y'a pas une racine unitaire du polynome caracteristique.}$$

````
```{r}
# Load the tseries package
library(tseries)
```
````

````
  {r}

test = adf.test(Indice)

if (test$p.value < 0.05) {
        cat("Le test présente un p-value egale à", test$p.value, " inférieur à un niveau typique comme 0.05. Alors on rejette l'hypythèse nulle, ainsi la serie est stationnaire")
    }else{
       cat("Le test présente un p-value egale à", test$p.value, " superieur à un niveau typique comme 0.05. Alors on accepte l'hypythèse nulle, ainsi la serie est non-stationnaire")}


}
```

# 2. identification d'un modèle

Maintenant on cherche à mettre en place un modele de prévision. Pour cela, on commence à explorer les auto-corellations

# Autocorellations

```{r}
acf(Indice, lag.max = 100)
pacf(Indice, lag.max = 100)
```

On a des ACF globalement decroissante avec des PACF nuls à partir du 14ème pics. Alors ce deux critere, sachant qu'on a confirmé que la série est stationnaire, nous permet de dire que le modele adapté est AR(13).

# Réduire l'échantillon pour conserver les 12 dernières données pour la validation.

L'idée est d'avoir deux echantillon, echantillon d'entrainenement pour pouvoir faire des prevision et echantillon test (les 12 dernières données pour la validation) pour mesurer la qualité de la prediction du modele construit.

```{r}
nb_lignes = length(Indice)

# Extraire les 12 dernières données pour la validation
donnes_valid = tail(Indice, 12)

# Mettre à jour la série temporelle en enlevant les 12 dernières données
donnes_test = head(Indice, nb_lignes - 12)
```
````
````

````
# Construction du Modele à partir des données test:

Ici on apprend le moedele choisit aux donner entrainement pour ensuite faire des previsions dans le futur, ces previsions seront comparer avec les données réelles de validation pour mesurer la capacité predictive du modele.

```{r}
library(tseries)

modele = arima(donnes_test,order=c(13, 0, 0))
```

```{r}
library(lmtest)
```

```{r}
coeftest(modele)
```


# Suppressions de terme non significatives

Des termes qui ne contribuent pas significativement à la robustesse du modele, ne servent à rien et logiquement il complixifie le modele pour rien. Ces termes sont à supprimer pour garder un modele simple.

```{r}
# Ajuster le modèle ARIMA avec les retards significatifs
modele <- arima(donnes_test, order = c(13, 0, 0), fixed = c(NA, 0, 0, 0, 0, NA, 0, 0, 0, 0, NA, NA, NA,NA))

# Afficher le modèle ajusté
coeftest(modele)

```

```{r}
# Ajuster le modèle ARIMA avec les retards significatifs
modele <- arima(donnes_test, order = c(13, 0, 0), fixed = c(NA, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, NA, NA,NA))

# Afficher le modèle ajusté
coeftest(modele)

```
````

```
Apres avoir construit ce modele, on doit la valider en s'assurant que les residus sont des bruit blanc:

# 6. Validation du model:
# Les résidus sont-ils de moyenne nulle ?

On pose les hypotheses:
$$H0 : E[\epsilon] = 0  \quad  \text{contre}  \quad H1: E[\epsilon] \neq 0$$
```

$$H0 : E[\epsilon] = 0 \quad \text{contre} \quad H1: E[\epsilon] \neq 0$$

```

 sous H0 on a √Te/σ ~ N(0,1).
 On rejette H0 avec un risque $\epsilon= 5\%$ lorsque la statistique
 t =√Te/σ est supérieure à $1.96$ en valeur absolue.

# Les résidus sont-ils auto-corrélés ? test Portmanteau:

$$H0 : ρ_1(\epsilon) = ρ_2(\epsilon) = ⋯ = ρ_K(\epsilon) = 0: \text{Les résidus sont indépendants et ne présentent pas d'autocorrélation.}$$
```

$$H0 : \rho_1(\epsilon) = \rho_2(\epsilon) = \cdots = \rho_K(\epsilon) = 0 : \text{Les résidus sont indépendants et ne présentent pas d'autocorrélation.}$$

```
Vs

$$H1 : \text{il existe un} \quad ρ_i(\epsilon) \quad \text{tel que} \quad ρ_i(\epsilon)\neq 0: \text{Il y a de l'autocorrélation dans les résidus.}$$
```

$$H1 : \text{il existe un} \quad \rho_i(\epsilon) \quad \text{tel que} \quad \rho_i(\epsilon)\neq 0: \text{Il y a de l'autocorrélation dans les résidus.}$$

````
```{r}
# Calcule de la statistique de test
n <- length(donnes_test)
moyenne_innov <- mean(na.omit(modele$residuals))
Ec_typ <- sqrt(modele$sigma2)
stat_test <- sqrt(n) * (moyenne_innov / Ec_typ)

teste_portmanteau_Ljung <- Box.test(modele$residuals, lag = 10, type = "Ljung-Box")

if (stat_test < 1.96 & teste_portmanteau_Ljung$p.value > 0.05 ) {
  cat("La Statistique de test est de", stat_test, " inférieure à 1.96 donc les résidus ont une moyenne nulle.\n")
  cat("La p-value du test de portmanteau est de", teste_portmanteau_Ljung$p.value, " supérieure à 0.05 donc les résidus ne présentent pas d'autocorrélation.\n")

  cat("Les résidus sont des bruits blancs.\n")
} else {
  cat("Les résidus ne sont pas des bruits blancs.\n")
}
```

# Graphes des résidus

```{r}
plot(modele$residuals)
```

# 7. calcul des prévisions
Avec ce modele present, on va prédire les 12 futur observations qui sont été supprimé.

```{r}
library(forecast)
```

```{r}
predictions = forecast(modele, level = 0.95, h=12)
```

```{r}
modele %>%
   forecast(level = 0.95, h=12) %>%
  autoplot()
```

#  Analyse  de la capacité prédictive de ce modèle.

```{r}
prediction = predictions$mean

# Calcul de l'erreur absolue moyenne (MAE)

MAPE <- mean(abs((donnes_valid - prediction)/donnes_valid))*100

cat("Le modèle presente un MAPE de", MAPE,"%.")
```
Avec un MAPE de 2.10%, on peut se confier à notre modele: c'est bon modele.
````