

MASTER I :

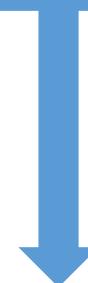
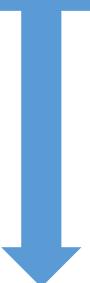
STATISTIQUE ET INFORMATIQUE DÉCISIONNELLE (SID)



PROJET DE SERIES TEMPORELLES:

MODELISATION DE LA PRODUCTION ELECTRIQUE

MENSUELLE DE 1985 A 2018



EXPOSANT :

ABDOU LAYE NDAO

MALICK FAYE

PROFESSEUR :

Dr Aba DIOP

ANNÉE SCOLAIRE 2022 / 2023

Table des matières

Introduction	1
I. Descriptions.....	2
1. Description des données	2
2. Description de la série.....	2
II. Décomposition de la série et Ajustement de la tendance en utilisant un modèle linéaire simple et multiple	3
1. Décomposition	3
2. Ajustement de la tendance	4
III. Etude de la saisonnalité et analyse des résultats.....	6
IV. Modélisation de la série par la méthode de Box & Jenkins.....	6
1. Tests de stationnarité.....	6
2. Choix du modèle et estimation	7
V. Prévision par la méthode ARIMA de la production électrique	10
1. Prévision pour une année	10
2. Prévision pour trois ans.....	11
VI. Prévision par la méthode de lissage exponentiel	11
1. Prévision pour une année	12
2. Prévision pour trois ans.....	12
3. Comparaison graphique de la méthode de lissage exponentiel et celle d'ARIMA	13
CONCLUSION	15

Introduction

La production électrique, est le processus de génération d'énergie électrique à partir de différentes sources d'énergie primaire. Elle est essentielle pour alimenter les foyers, les industries, les transports et les infrastructures de communication. Elle joue un rôle crucial dans le développement économique et le confort de vie des personnes. Ainsi une étude des séries temporelles dans la production d'énergie est essentielle pour analyser et prévoir la demande et d'optimiser la production. Dans ce contexte, cette étude vise à modéliser les données de la production électrique mensuelle d'un pays entre 1985 et 2018.

Les objectifs attendus dans cette étude sont nombreux. Parmi lesquelles, nous pouvons citer tout d'abord, l'étude de la tendance de la série en utilisant un modèle linéaire simple et multiple. Ensuite de faire une prédition de la production électrique pour les années à venir en utilisant les processus ARIMA et la méthode de lissage exponentiel et enfin de fournir des résultats fiables qui pourrons aider les décideurs à prendre des décisions éclairées sur la production d'électrique.

I. Descriptions

1. Description des données

La base de données de l'étude nommer **Base_production_Electrique** est composée de deux variables numériques et de 397 observations. La première variable **date**, nous renseigne sur les dates de production électrique en mois et la variable **production** nous indique la quantité électrique produit chaque mois entre 1985 et 2018.

2. Description de la série

La base de départ est maintenant considérée comme une série temporelle grâce à une conversion faite. La fréquence de la série est égale à 12, vue que l'année est la référence et si on compte le nombre d'unité dans l'année en termes de mois est égale à 12. Cette série a comme point de départ le 1^{ière} Janvier 1985 et se termine au 1^{ière} Janvier 2018. La représentation graphique de la série nous montre une tendance croissante de celle-ci.

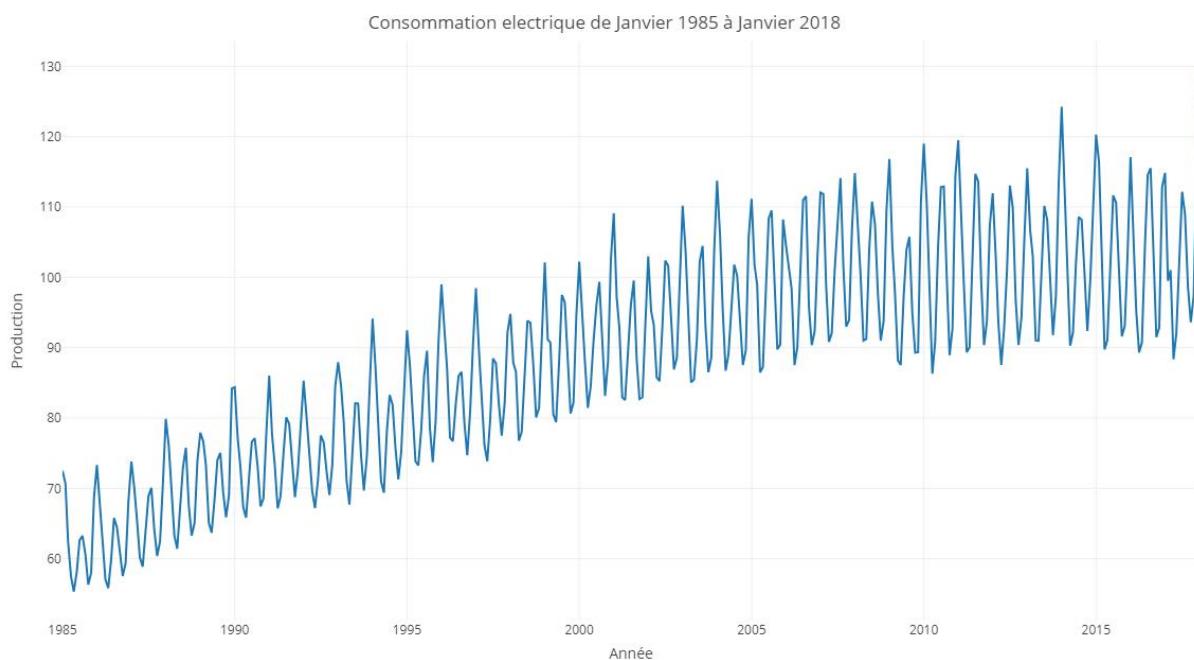


Figure 1 : Consommation électrique de 1985 à 2018

II. Décomposition de la série et Ajustement de la tendance en utilisant un modèle linéaire simple et multiple

1. Décomposition

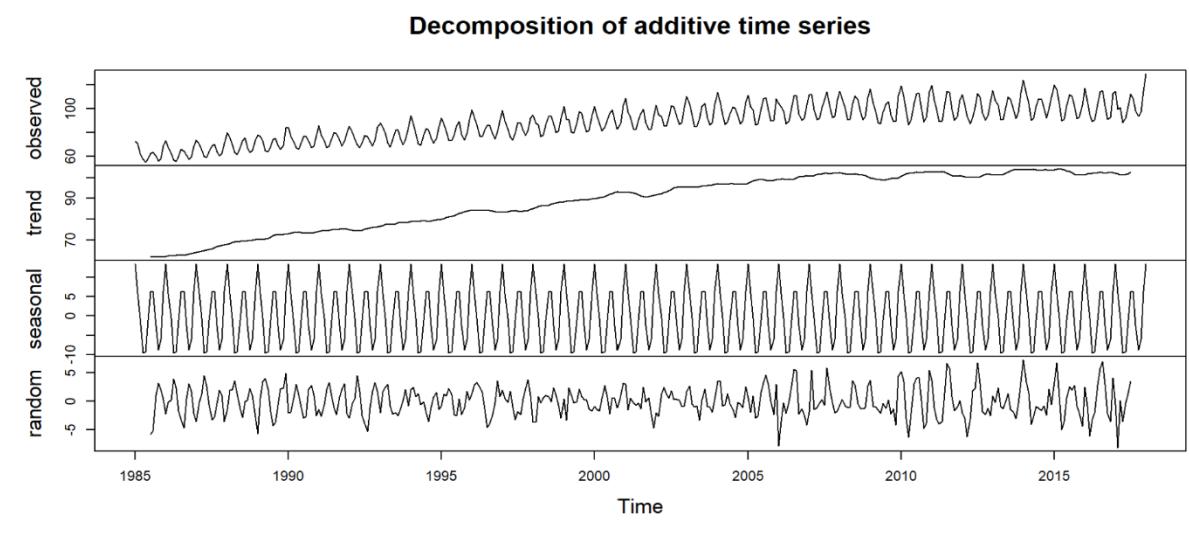


Figure 2 : Décomposition additive de la série

La figure ci-dessus représente la décomposition additive de la série. La première partie de la figure (observed) nous montre la représentation graphique de la série normale à décomposer. Cette décomposition nous donne les trois parties qui composent la série à savoir : la tendance (trend), la saisonnalité (seasonal) et l'effet aléatoire (random).

➤ La tendance :

Il s'agit de la composante de la série qui représente l'évolution à long terme des données. Cette tendance est représentée de manière lissée pour faciliter l'identification des mouvements à long terme dans la série temporelle. On peut voir que la tendance (trend) progresse de façon croissante tout au long de la série.

➤ La saisonnalité :

Il s'agit des variations régulières dans les données qui se produisent à des intervalles réguliers. Cette composante peut être utilisée pour identifier les variations saisonnières de la production électrique mensuelle, telles que les pics de production pendant les mois de Janvier vers Juillet ou les baisses de production pendant les mois de Septembre vers Novembre. On constate que la saisonnalité évolue selon des amplitudes égales.

➤ L'effet aléatoire

La dernière partie de la figure nous montre la composante aléatoire de la série temporelle. Il s'agit des variations imprévisibles dans les données qui ne peuvent pas être expliquées par la tendance ou la saisonnalité.

En somme, cette décomposition additive nous permet de mieux comprendre les différentes composantes qui composent la série de production électrique mensuelle, ce qui peut être utile pour la modélisation et la prédition des valeurs futures de la série temporelle.

2. Ajustement de la tendance

➤ Modèle linéaire simple

```
t=time(produc)
y=as.vector(produc)
x=as.vector(t)
x1<-x[7:391]
modelSimple=lm(y~x)
summary(modelSimple)

call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.6168 -6.3153 -0.4795  5.5767 21.5226 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.595e+03 8.957e+01 -28.98   <2e-16 ***
x             1.341e+00 4.475e-02   29.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.515 on 395 degrees of freedom
Multiple R-squared:  0.6946,    Adjusted R-squared:  0.6938 
F-statistic: 898.2 on 1 and 395 DF,  p-value: < 2.2e-16
```

Le modèle montre que la variable explicative **x** est significative car sa p-value < 0, 05, ce qui montre que la relation entre la tendance et le temps est statistiquement significative. Le coefficient de régression de la variable **x** est de 1,341, ce qui suggère que pour chaque unité de temps, la tendance augmente de 1,341 unités. On peut dire que le modèle est bon car il a un coefficient de détermination R^2 de 0,6946, ce qui explique que 69,46% de la variation de la tendance peut être expliquée par le temps dans le modèle de régression linéaire. Cela veut dire que le modèle a un bon pouvoir d'ajustement sur les données de production électrique et que le temps est une variable importante pour expliquer les variations de la tendance. Ainsi, il est important de noter qu'il existe d'autres types de régressions à savoir la régression multiple qui prend en compte d'autres facteurs importants qui pourra influencer la production électrique que le modèle linéaire simple ne prend pas en compte.

➤ Modèle linéaire multiple

```
x2=x^2
modelMultiple=lm(y~x+x2)
summary(modelMultiple)
```

```
Call:
lm(formula = y ~ x + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.5314 -6.4982 -0.4784  5.3515 25.4529 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.580e+05  1.950e+04 -8.103 6.79e-15 ***
x            1.566e+02  1.948e+01  8.039 1.07e-14 ***
x2          -3.880e-02  4.868e-03 -7.970 1.72e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.912 on 394 degrees of freedom
Multiple R-squared:  0.737,   Adjusted R-squared:  0.7356 
F-statistic: 552 on 2 and 394 DF,  p-value: < 2.2e-16
```

Le modèle montre que les variables explicatives **x** et **x2** sont significative car leur p-value < 0,05, ce qui montre que la relation entre la tendance et le temps est statistiquement significative. Le coefficient de régression de la variable **x** est de 1,566 et celui de **x2** est de -3,880, ce qui suggère que pour chaque unité de temps, la tendance augmente de 1,566 unités et diminue de 3,880 unités si le temps augmente au carré. On peut dire que le modèle est bon car il a un coefficient de détermination R² de 0,737, ce qui explique que 73,37% de la variation de la tendance peut être expliquée par le temps dans le modèle de régression linéaire multiple. Cela implique que le modèle a un bon pouvoir d'ajustement sur les données de production électrique et que les variables explicatives (le temps) sont très importantes pour expliquer les variations de la tendance.

➤ Comparaison des deux modèles linéaires

On note que le R²-ajusté du modèle linéaire multiple est de 0,7356 et que celui du modèle linéaire simple est de 0,6938. Ici, on constate que le R²-ajusté du modèle linéaire multiple est supérieur à celui du modèle linéaire simple. Donc on peut en déduire que le modèle linéaire multiple est le meilleur modèle pour ajuster les données de la production électrique.

III. Etude de la saisonnalité et analyse des résultats

```
[1] "Modele lineaire multiple avec la saisonnalite"
Call:
lm(formula = y ~ x + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.5314 -6.4982 -0.4784  5.3515 25.4529 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.580e+05 1.050e+04 -8.103 6.79e-15 ***
x           1.566e-02 1.948e+01  8.039 1.07e-14 ***
x2          -3.880e-02 4.868e-03 -7.970 1.72e-14 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.912 on 394 degrees of freedom
Multiple R-squared:  0.737, Adjusted R-squared:  0.7356 
F-statistic: 552 on 2 and 394 DF, p-value: < 2.2e-16

[1] "Modele lineaire multiple sans la saisonnalite"
Call:
lm(formula = tandancel ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8694 -1.1759  0.1793  1.0697  3.0143 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.692e+05 3.556e+03 -47.59 <2e-16 ***
x1          1.679e+02 3.553e+00  47.24 <2e-16 ***
x2          -4.160e-02 8.877e-04 -46.86 <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.336 on 382 degrees of freedom
Multiple R-squared:  0.9893, Adjusted R-squared:  0.9893 
F-statistic: 1.773e+04 on 2 and 382 DF, p-value: < 2.2e-16
```

Ici, on voit que dans le modèle linéaire multiple avec la saisonnalité que le R^2 (le coefficient de détermination) est égal à 0,737 et le coefficient de détermination du modèle linéaire multiple sans la saisonnalité est égal à 0,9893. De ce fait si on se base sur les coefficients de détermination, on peut conclure que le modèle sans la saisonnalité est meilleur. En plus si on se base sur les AIC, on voit que l'AIC du modèle avec la saisonnalité ($AIC=2773,911$) est supérieur à celui du modèle sans la saisonnalité ($AIC = 1320,781$). Or plus l'AIC est faible plus le modèle est meilleur. Donc on peut conclure dans les deux cas que c'est le modèle sans la saisonnalité qui est meilleur.

Modèle linéaire multiple avec la saisonnalité
AIC : 2773.911

Modèle linéaire multiple sans la saisonnalité
AIC : 1320.781

IV. Modélisation de la série par la méthode de Box & Jenkins

1. Tests de stationnarité

➤ Test augmenté de Dickey-Fuller

```
Augmented Dickey-Fuller Test
data: produc
Dickey-Fuller = -5.139, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Le test de Dickey-Fuller suppose au départ que le test est non stationnaire en H_0 . Après avoir effectué le test, on constate que le p-value est inférieure à 0,05. Donc l'hypothèse nulle est rejetée, d'où le test est stationnaire.

➤ TEST de Phillips-Perron

```
Phillips-Perron Unit Root Test
data: produc
Dickey-Fuller z(alpha) = -88.143, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary
```

Du même point de vue que le test de Dickey-Fuller, le test de Philips-Perron suppose que le test est non stationnaire à l'hypothèse nulle. Après le test, on constate que le p-value trouvé est inférieure à 0,05. Donc l'hypothèse nulle est rejetée, ce qui veut dire que le test est stationnaire.

2. Choix du modèle et estimation

➤ Partition de la base d'apprentissage et test

Notre base de départ comporte 397 observations et qui est répartie en base d'apprentissage et test. La base d'apprentissage représente les 85% ce qui équivaut à 337 observations et notre base de test représente les 15% qui reste (60 observations).

➤ Représentation du graphe de PACF

Series apprentissage

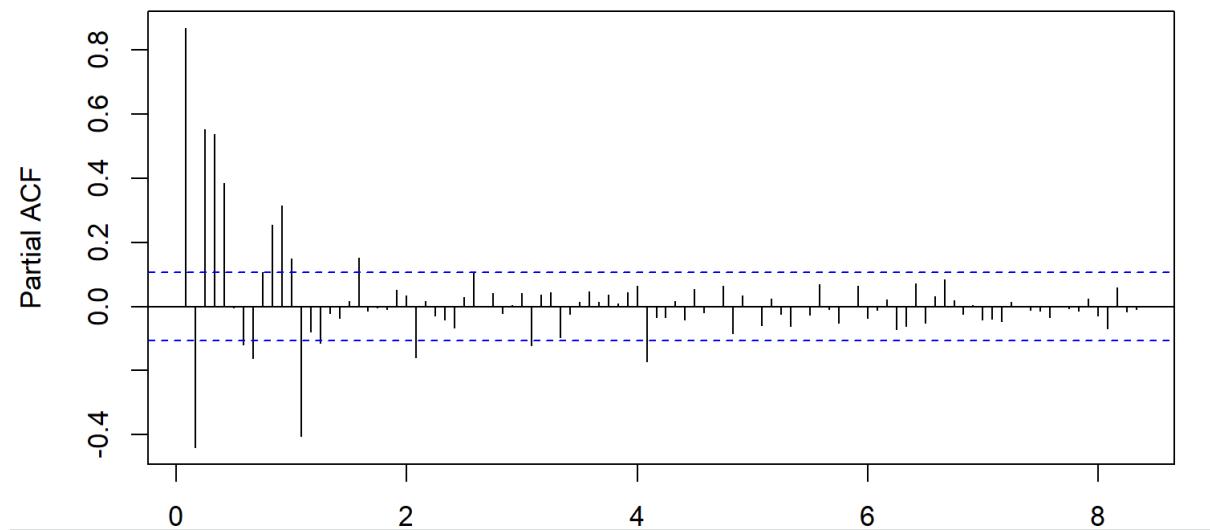


Figure 3 : PACF

En observant la courbe de la fonction d'autocorrélation partielle, on remarque qu'à un certain moment que les pics de la fonction restent à l'intérieur de l'intervalle de confiance, ce qui explique la nullité de l'autocorrélation partielle. La valeur du paramètre p est choisi entre les valeurs de la fonction dont les pics sont en dehors de l'intervalle de confiance.

➤ Représentation du graphe ACF

Series apprentissage

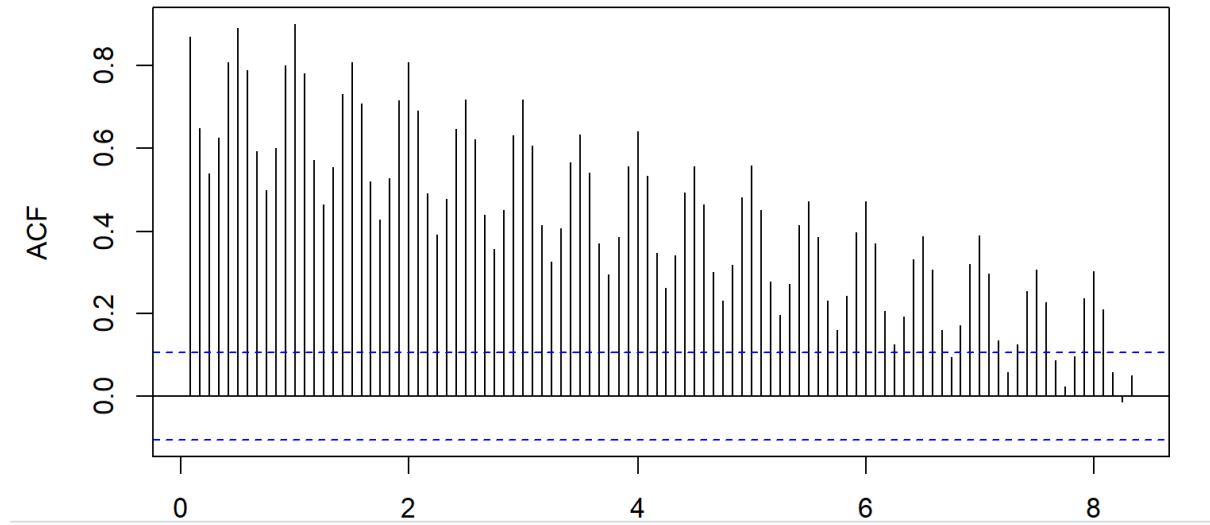


Figure 4 : ACF

En observant la courbe de la fonction d'autocorrélation simple, on peut remarquer que la décroissance n'est pas très rapide, ce qui explique que la série présente une certaine dépendance temporelle d'où la présence de la saisonnalité. Cela peut être confirmé en examinant les pics significatifs en dehors de l'intervalle de confiance, qui indiquent des corrélations non nulles entre les observations distantes dans le temps.

➤ Choix du modèle avec la fonction d'auto-arima

```

Series: apprentissage
ARIMA(2,1,1)(0,1,1)[12]

Coefficients:
          ar1      ar2      ma1      sma1
          0.5303 -0.0868 -0.9332 -0.7266
  s.e.   0.0602  0.0606  0.0277  0.0394

sigma^2 = 4.969: log likelihood = -722.96
AIC=1455.92  AICc=1456.11  BIC=1474.83

```

Le meilleur modèle proposé par la commande `auto.arima` est ARIMA (2,1,1) (0,1,1) [12].

Cela signifie qu'il s'agit d'un modèle autorégressif intégré de moyenne mobile saisonnier d'ordre (2,1,1) avec une saisonnalité de période 12 (12 mois dans l'année), et un terme d'erreur saisonnier d'ordre 1.

➤ Validation du modèle avec les résidus

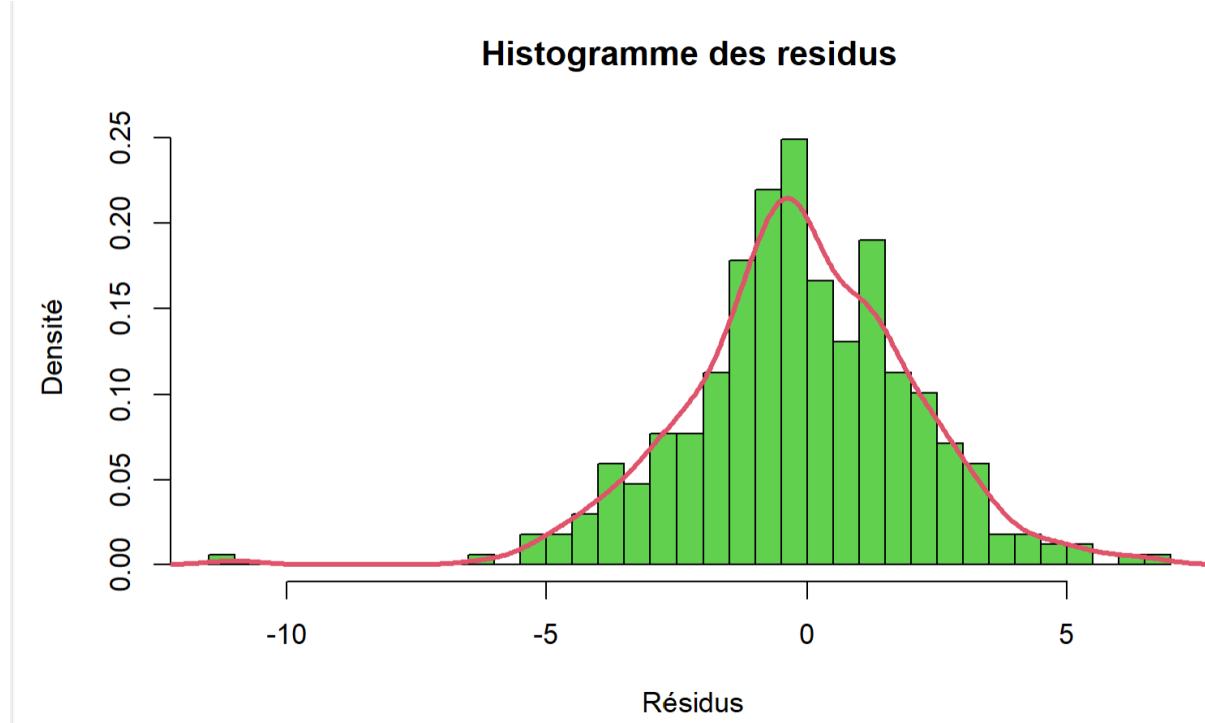


Figure 5 : Histogramme de résidus

Ici on constate que la distribution des classes est quasi symétrique. De plus, après avoir tracé la courbe de la densité des résidus on voit que la distribution suit une loi gaussienne. Donc, on peut dire que le modèle est bon.

➤ Validation du modèle par test

Summary prédition

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
89.05	93.45	100.87	102.08	110.14	117.07

Summary base test

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
88.35	93.02	102.40	102.94	110.26	129.40

On constate qu'il y'a pas une grande différence en comparant les minimums et maximums des valeurs prédites et ceux des vraies valeurs de la base de test. Le même constat a été fait sur les médianes et les moyennes des deux valeurs (prédites et base de test). Donc on peut dire qu'on a un modèle qui a un bon pouvoir prédictif.

V. Prévision par la méthode ARIMA de la production électrique

Dans cette partie, nous allons effectuer des prévisions pour une année et pour trois années à l'aide de la méthode ARIMA. Cette méthode nous permet de faire des prévisions basées sur des modèles statistiques qui prennent en compte les composantes de la série chronologique telles que la tendance, la saisonnalité et les fluctuations aléatoires.

1. Prévision pour une année

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
2018	114.31111	104.45857	92.09910	93.63140	104.31821	113.65996	112.58325	101.93541	93.85642	
2019	122.04284									
	Nov	Dec								
2018	97.12217	112.41629								
2019										

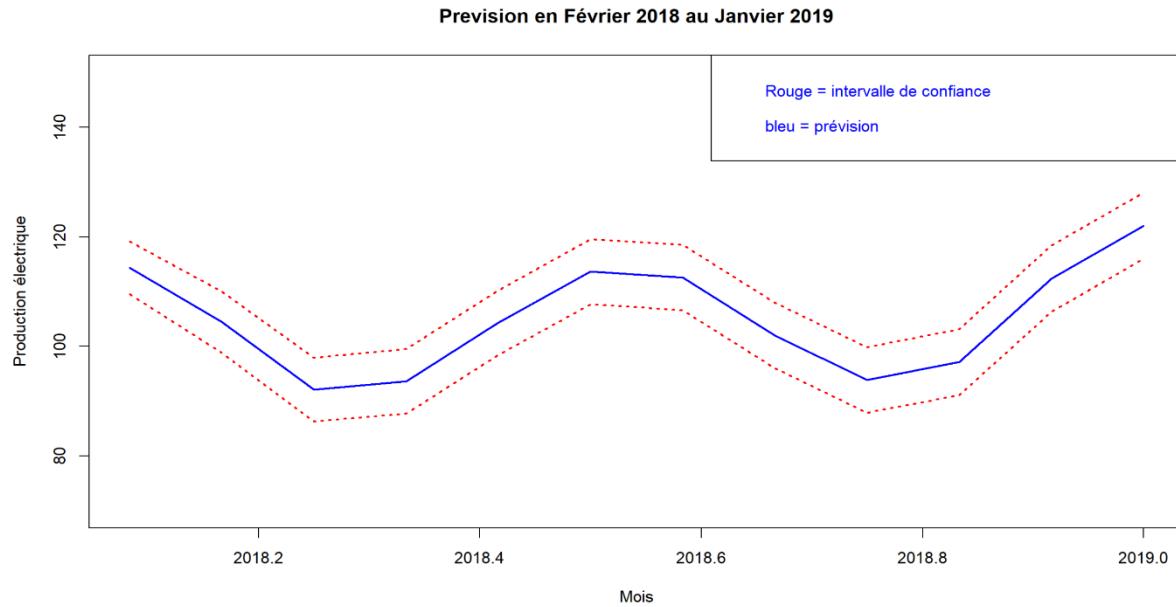


Figure 6 : Prédiction d'une année

La prédition donne les valeurs prévues pour chaque mois de Février 2018 à Janvier 2019. Les prévisions de production électrique pour les mois de Décembre à Mars et de Juin à Septembre 2018 sont plus élevées du fait que pendant ces périodes de l'année, on constate qu'il y a plus de fraicheur que dans les autres mois (comme les mois d'Avril, Mai, Octobre et Novembre). Donc on peut dire que la consommation d'électricité est plus importante dans les mois où il y a plus de fraicheur. Il est à noter que la prévision est à l'intérieur des intervalles de confiance. De plus, elle suit la même tendance que celle des intervalles de confiance.

2. Prévision pour trois ans

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2018	114.31111	104.45857	92.09910	93.63140	104.31821	113.65996	112.58325	
2019	122.04284	110.69451	103.24842	91.95762	93.91376	104.76087	114.16190	113.10687
2020	122.57875	111.23047	103.78440	92.49360	94.44975	105.29685	114.69789	113.64286
2021	123.11473							
	Sep	Oct	Nov	Dec				
2018	101.93541	93.85642	97.12217	112.41629				
2019	102.46692	94.39078	97.65757	112.95206				
2020	103.00290	94.92677	98.19356	113.48805				
2021								

Prévision en Février 2018 au Janvier 2021

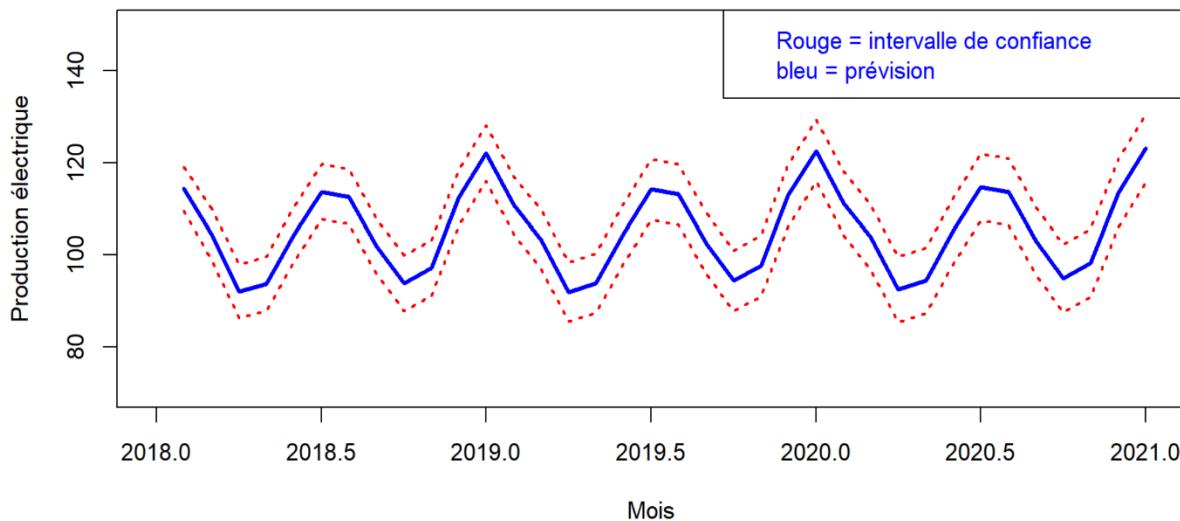


Figure 7 : Prédiction trois ans

On constate que dans les trois ans de prévision, la production électrique est plus importante dans les mois de Décembre à Mars et de Juin à Septembre. Tandis qu'il y'a une petite baisse de production entre les mois d'Avril, Mai, Octobre et de Novembre. Il est à noter que la prévision est à l'intérieur des intervalles de confiance. De plus, elle suit la même tendance que celle des intervalles de confiance. Cela donne une indication de l'habilitation de la prévision. Le fait que la prévision suive la même tendance que les intervalles de confiance indiquent également que le modèle est capable de capturer les variations de la série chronologique et de fournir une prévision qui suit la même tendance que les données observées. Cela renforce la confiance dans la qualité de la prévision obtenue à partir du modèle.

VI. Prévision par la méthode de lissage exponentiel

Dans cette partie, nous allons effectuer des prévisions pour une année et pour trois années à l'aide de la méthode de lissage exponentiel et comparer graphiquement ces résultats avec ceux

obtenus à l'aide de la méthode ARIMA. Le lissage exponentiel est une méthode de prévision largement utilisé pour les séries chronologiques. Elle consiste à prendre en compte la tendance et la saisonnalité de la série pour produire des prévisions.

1. Prévision pour une année

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2018	115.12458	109.15159	97.69496	100.10703	111.28663	120.85122	119.52513	
2019	128.16038							
	Sep	Oct	Nov	Dec				
2018	108.92212	100.62997	103.68895	118.96908				
2019								

Prévision en Février 2018 au Janvier 2021

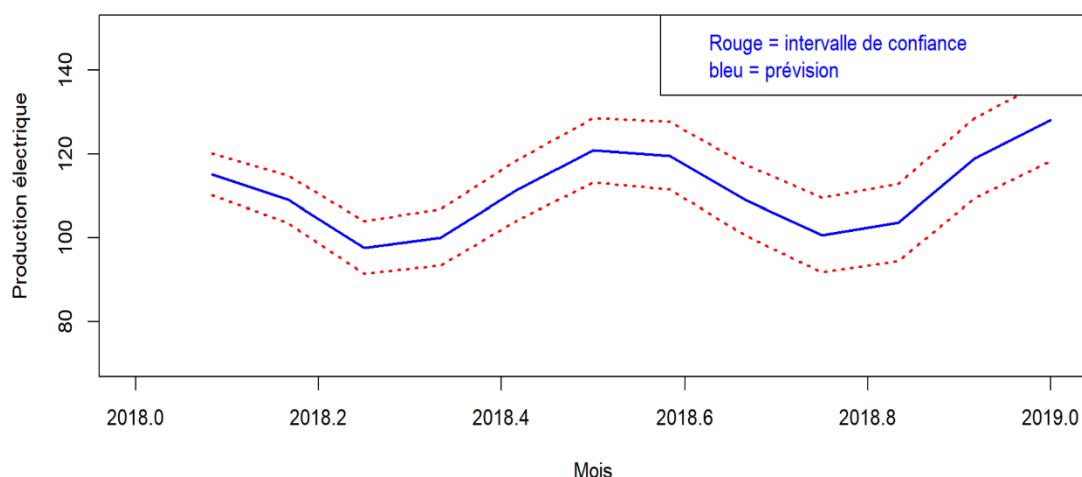


Figure 8 : Prévision pour une année par lissage exponentiel

Les valeurs de prévision sont données pour chaque mois de l'année 2018 avec la prévision du Janvier 2019. Les prévisions de production électrique pour les mois de Décembre à Mars et de Juin à Septembre 2018 sont supérieures à 107. Tandis que les valeurs prévues pour les mois de Avril, Mai, Octobre et Novembre 2018 sont inférieur à 107. Il est à noter que la prévision est à l'intérieur des intervalles de confiance. De plus, elle suit la même tendance que celle des intervalles de confiance.

2. Prévision pour trois ans

	Jan	Feb	Mar	Apr	May	Jun	Jul
2018	115.12458	109.15159	97.69496	100.10703	111.28663	120.85122	
2019	128.16038	115.95284	109.97986	98.52323	100.93530	112.11490	121.67948
2020	128.98864	116.78111	110.80812	99.35149	101.76356	112.94316	122.50775
2021	129.81691						
	Aug	Sep	Oct	Nov	Dec		
2018	119.52513	108.92212	100.62997	103.68895	118.96908		
2019	120.35340	109.75038	101.45824	104.51721	119.79734		
2020	121.18166	110.57865	102.28650	105.34548	120.62561		
2021							

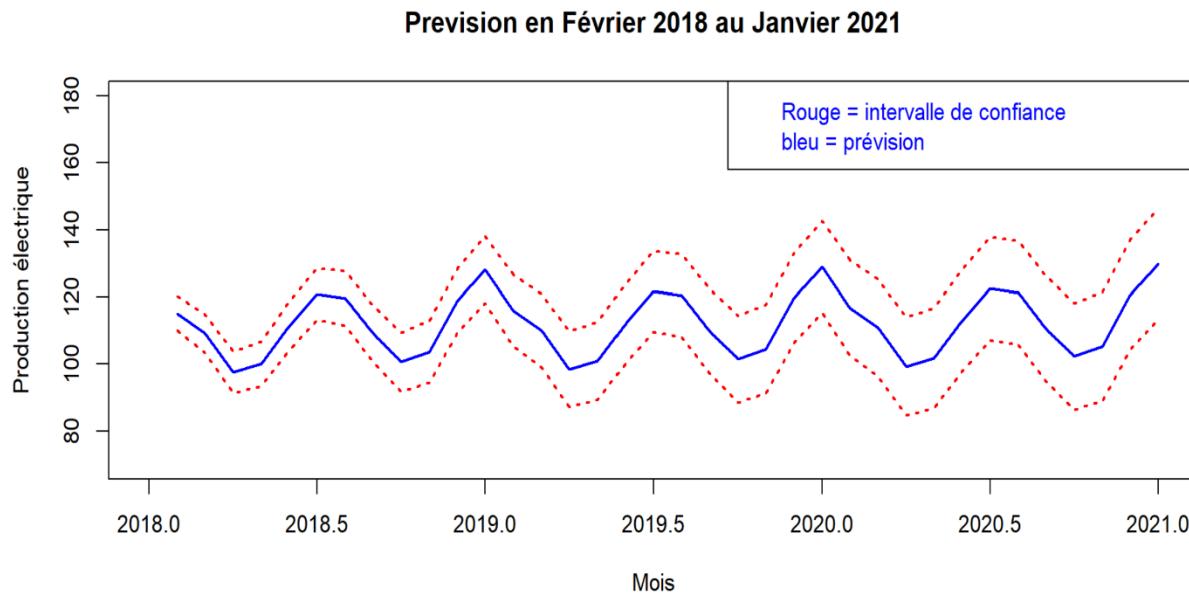


Figure 9 : Prévision pour trois ans par lissage exponentiel

On constate dans le graphe de la prévision qu'il y'a pas une grande différence entre l'année 2018 - 2019 et les deux années qui la succède. Alors, on peut dire que notre modèle a un bon pouvoir de prédiction sur les données de la production électrique. De plus, les intervalles de confiance permettent de quantifier l'incertitude associée aux prévisions.

3. Comparaison graphique de la méthode de lissage exponentiel et celle d'ARIMA

- Pour une année

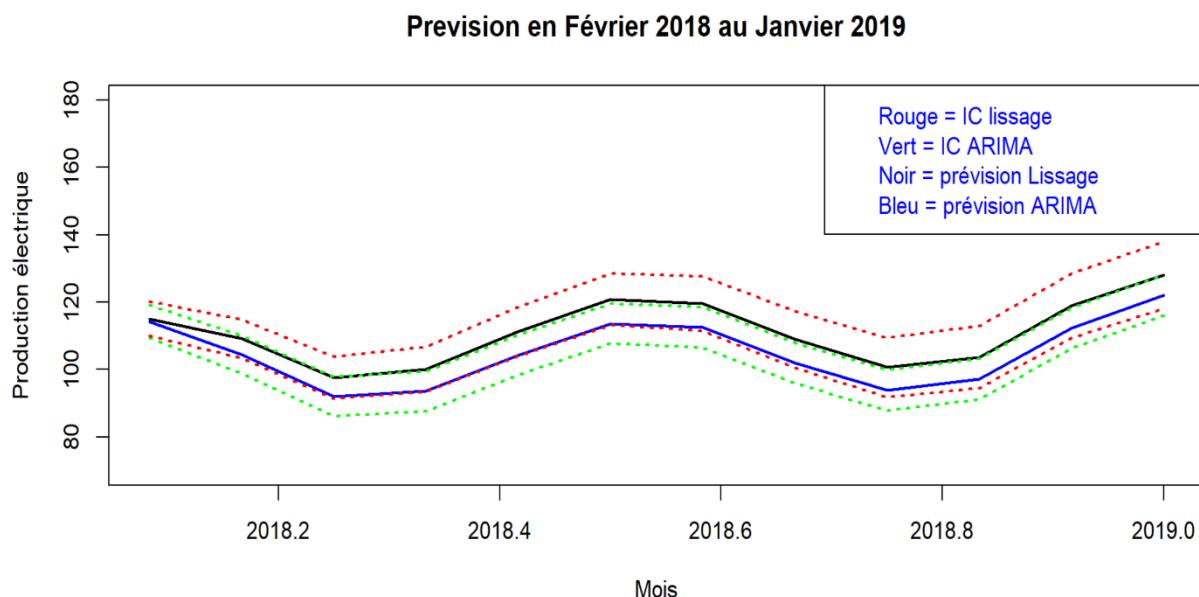


Figure 10 : Comparaison ARIMA et Lissage exponentiel pour un an

D'après les résultats, les prévisions par ARIMA semblent être plus proches des valeurs réelles que les prévisions par lissage. Cela peut être dû au fait que le modèle ARIMA prend en compte la tendance et la saisonnalité, ainsi que les erreurs résiduelles, pour exécuter ses prévisions. En revanche, le lissage exponentiel ne prend en compte que les tendances passées et ne tient pas compte des erreurs passées, ce qui peut conduire à des prévisions moins précises. Toutefois, il est important de noter que la performance des modèles dépend fortement des données utilisées pour l'entraînement et la validation, ainsi que de la qualité de la modélisation effectuée.

➤ Pour trois ans

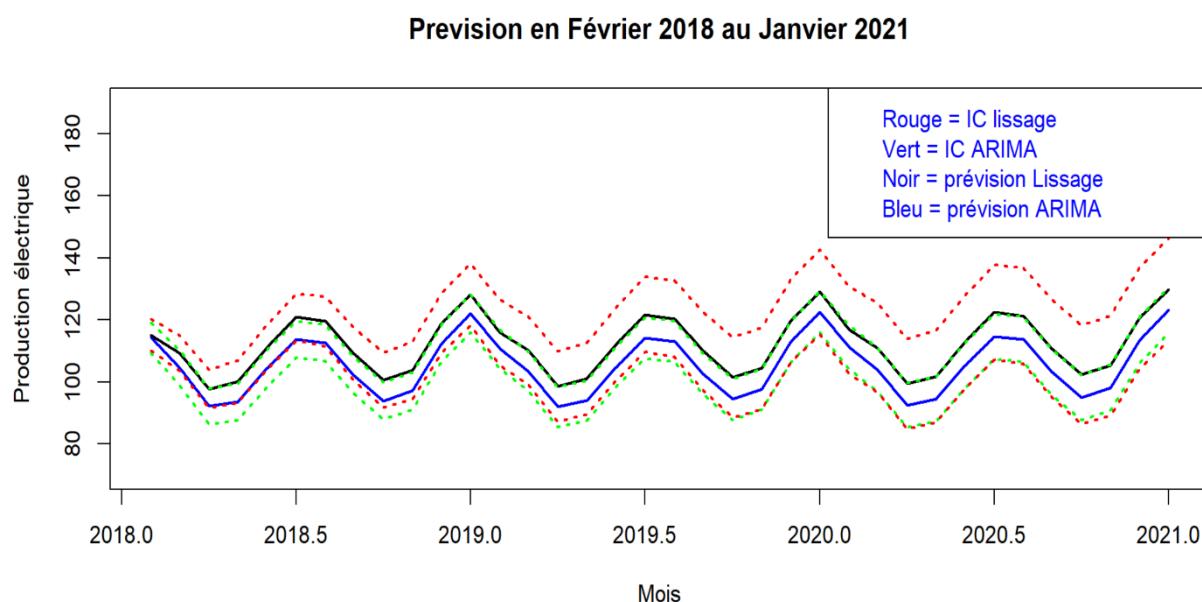


Figure 11 : Comparaison ARIMA et Lissage exponentiel pour trois ans

En examinant les résultats des prévisions pour une période de 3 ans, nous pouvons voir que les deux méthodes prévoient une tendance générale à la hausse de la série chronologique. Cependant, les prévisions par lissage exponentiel prévoient des valeurs plus élevées pour la plupart des mois. En effet, les prévisions par ARIMA semblent être plus stables et plus lisses, tandis que les prévisions par lissage sont plus volatiles, avec des variations plus marquées d'un mois à l'autre.

CONCLUSION

En conclusion, cette étude a pour but d'utiliser la méthode de lissage exponentiel et la méthode ARIMA pour modéliser et prévoir les données de la production électrique. Les résultats montrent que les deux méthodes sont capables de fournir des prévisions raisonnables pour la production électrique à court terme. Cependant, pour des prévisions à plus long terme, la méthode ARIMA s'est avérée plus performante. Il est important de noter que les modèles ne fournissent pas des prédictions exactes de l'avenir, mais plutôt des estimations basées sur les données historiques observées. Il est donc recommandé de continuer à surveiller la production électrique et de mettre à jour régulièrement les modèles pour s'adapter aux changements dans les tendances de production. Ces résultats peuvent aider les entreprises et les organisations à planifier leur production électrique et à prendre des décisions éclairées pour améliorer leur efficacité énergétique.