Mémoire

Série Temporelle



Mikael BOZON, William LAURENT, Pierre MARJOLLET

Contents

[Mise en place du Framework 2](#_Toc509599112)

[Présentation du sujet et de notre objectif 2](#_Toc509599113)

[R, algorithmique orientée statistique 2](#_Toc509599114)

[Partie 1 : Production totale brute d’électricité 3](#_Toc509599115)

[1. Présentation de la série 3](#_Toc509599116)

[2. Premières approches 3](#_Toc509599117)

[3. Modélisation SARIMA 5](#_Toc509599118)

[Partie 2 : Importation totale d’électricité 10](#_Toc509599119)

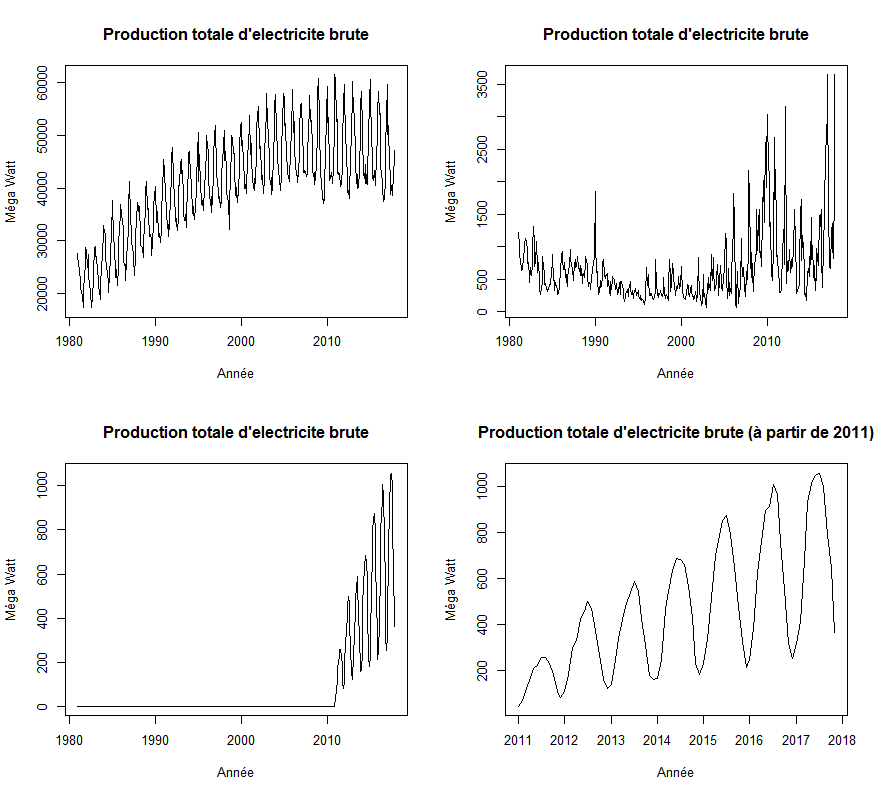
[Partie 3 : Production d’électricité photovoltaïque brute 11](#_Toc509599120)

# Mise en place du Framework

## Présentation du sujet et de notre objectif

Nous disposons d’une base de données contenant vingt-huit séries de données relatives à la production et à la consommation d’électricité en France. Ces données ont été calculées de manière mensuelle sur une durée de 36 ans (1981-2017)

Nous nous intéressons particulièrement aux séries de données suivantes :

* **prod.total** - Production totale brute d'électricité
* **import** - Total des importations physiques d'électricité
* **photo** - Production brute d'électricité photovoltaïque (à partir de 2011)

Ci-contre une photo de ces séries :

**L’objectif de ce travail est de fournir un modèle de prédiction pour les 12 prochains mois de chacune des séries de données.**

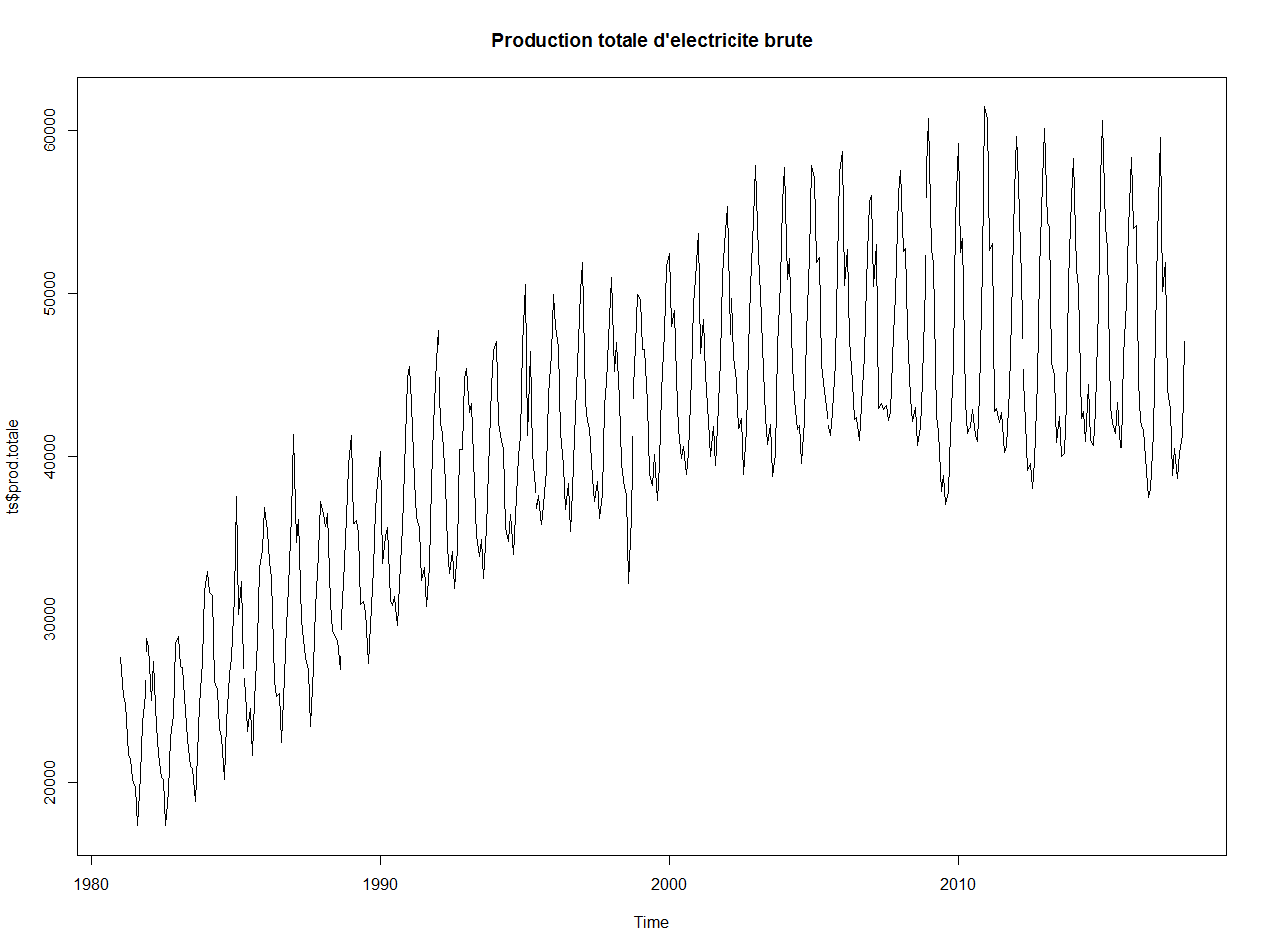
## R, algorithmique orientée statistique

Nos algorithmes seront interprétés par le logiciel open-source **R**. Nous en utiliserons les librairies suivantes :

* **magrittr** pour l’aspect code pure : magrittr permets de simplifier l’écriture des codes
* **stats** pour l’analyse des séries entières, stats fait parties des packages initialement pourvus dans R, développé par la « R Core Team »et agrémentés de contributions
* **corrplot** pour pouvoir développer des graphiques de corrélations […] On s’en sers même pas au final

# Partie 1 : Production totale brute d’électricité

### Présentation de la série



Cette première représentation appelle plusieurs commentaires.

1. La série présente un trend et une saisonnalité, qui semblent tous les 2 déterministres.
2. La série n’est clairement pas stationnaire.
3. L’amplitude des oscillations semble être légèrement croissante, mais rien ne nous permet d’en déduire clairement que nous devons mettre en place un modèle multiplicatif.

### Premières approches

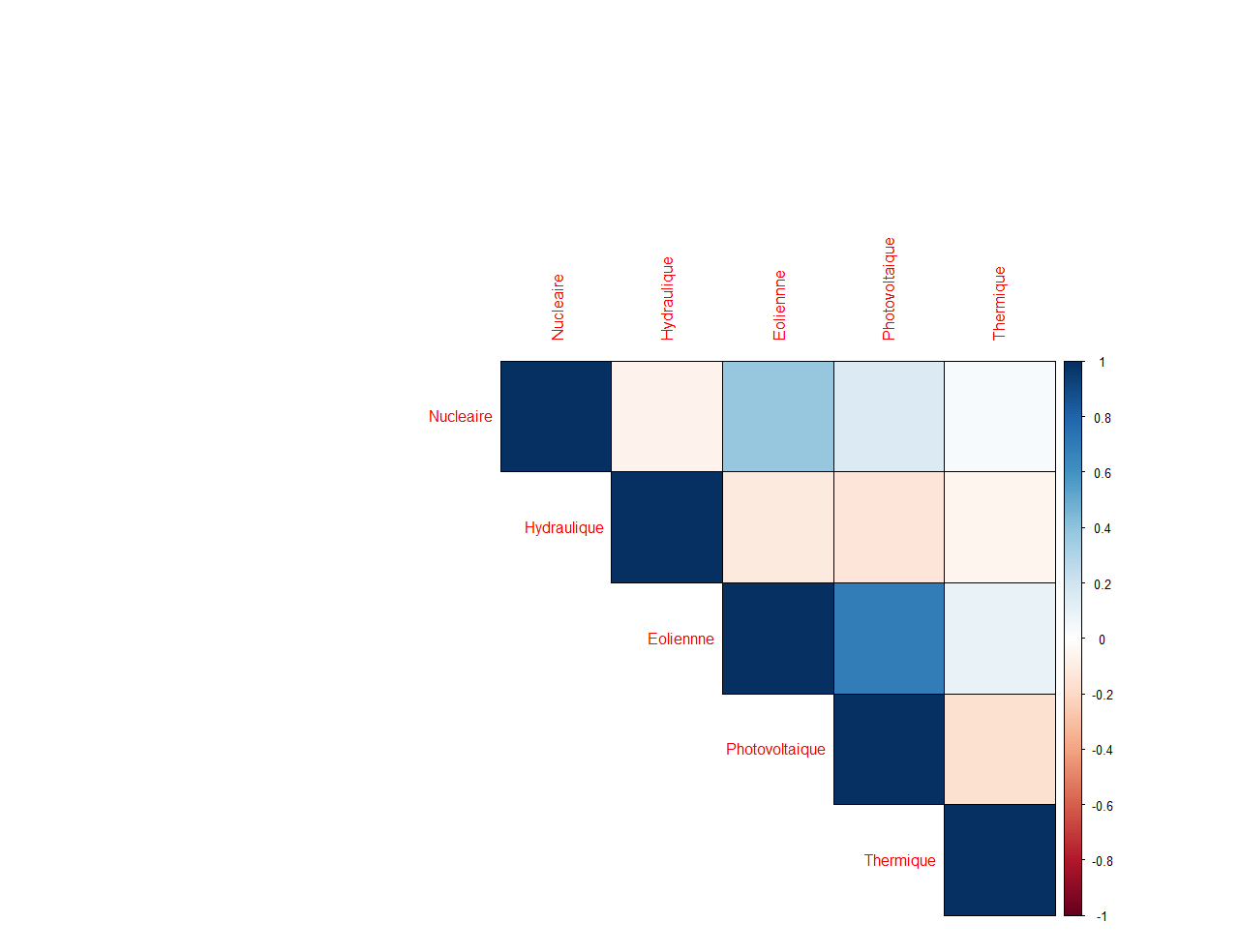
Avant de se lancer dans la construction d’un modèle sur la série globale, nous pouvons essayer d’étudier la construction de cette série à partir des autres séries à disposition.

Nous commençons par remarquer cette série est la somme des séries :

* Production brute nucléaire
* Production brute hydraulique
* Production brute éolienne
* Production photovoltaïque
* Production brute thermique

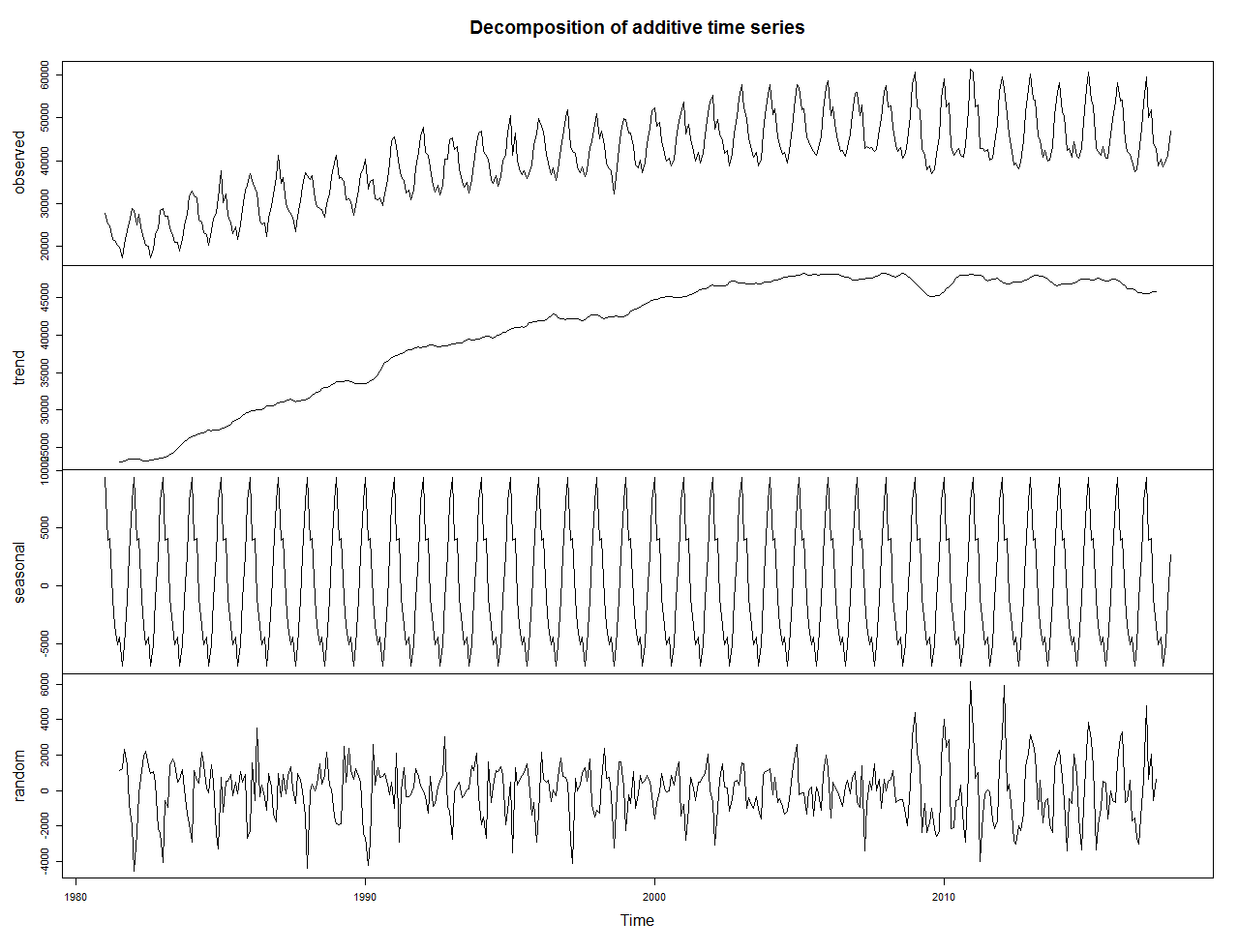
Nous allons d’abord étudier les corrélations entre ces séries. Si d’aventure il s’avérait qu’elles soient indépendantes, nous pourrons explorer la piste qui consiste à les modéliser toutes séparément et prendre la somme.

On calcule et représente graphiquement la matrice des corrélations entre ces séries :



Une rapide analyse nous mène rapidement à la conclusion qu’il sera fastidieux de décomposer la série globale selon les sous-séries dont elle est la somme. Si nous persistions dans ce sens, il nous faudrait estimer les corrélations pour pouvoir agréger les résultats.

Nous allons donc étudier la série comme telle.



Sur cette décomposition additive avec la fonction decompose(), on observe un trend qui n’est pas linéaire, la partie aléatoire n’est pas stationnaire. Seule la partie saisonnière semble correcte. Nous n’irons pas plus loin avec cette décomposition.

### Modélisation SARIMA

Partant du fait que la série n’est pas stationnaire et présente une saisonnalité, nous allons commençons par calibrer un modèle SARIMA.

Nous suivrons le processus général suivant :

Maintenant appliquons ce process à notre série :

Différenciation saisonnalité

**OUI**

Test stationnarité : KPSS

**NON**

Différenciation 1er ordre

**OUI**

Test stationnarité : KPSS

**NON**

Tests validation du modèle :

* Normalité des résidus : Kolmogrov – Smirnov, Shapiro-Wilk
* Blancheur des résidus : Ljung-Box

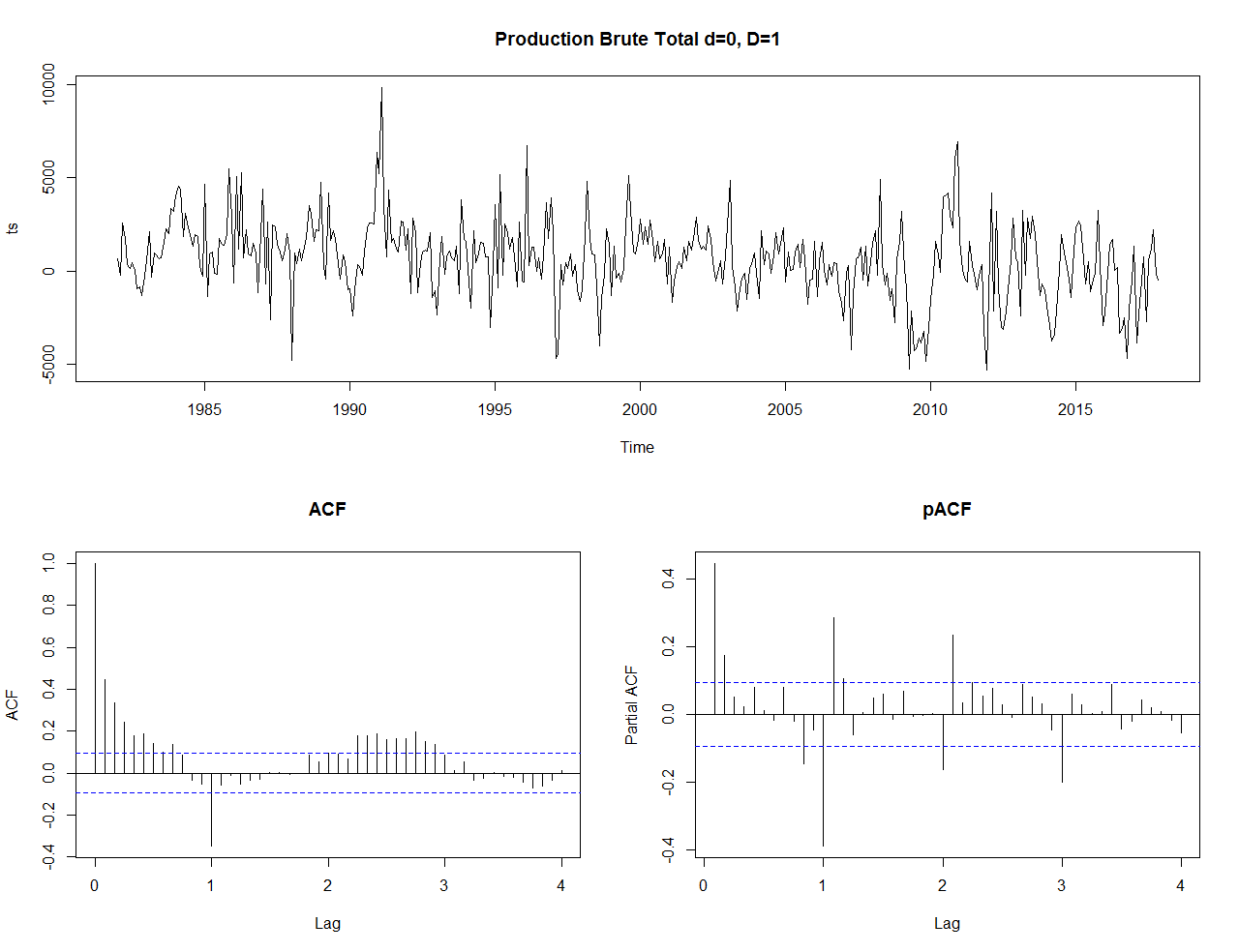
Analyse :

* ACF (pour p,q)
* pACF (pour P, Q)

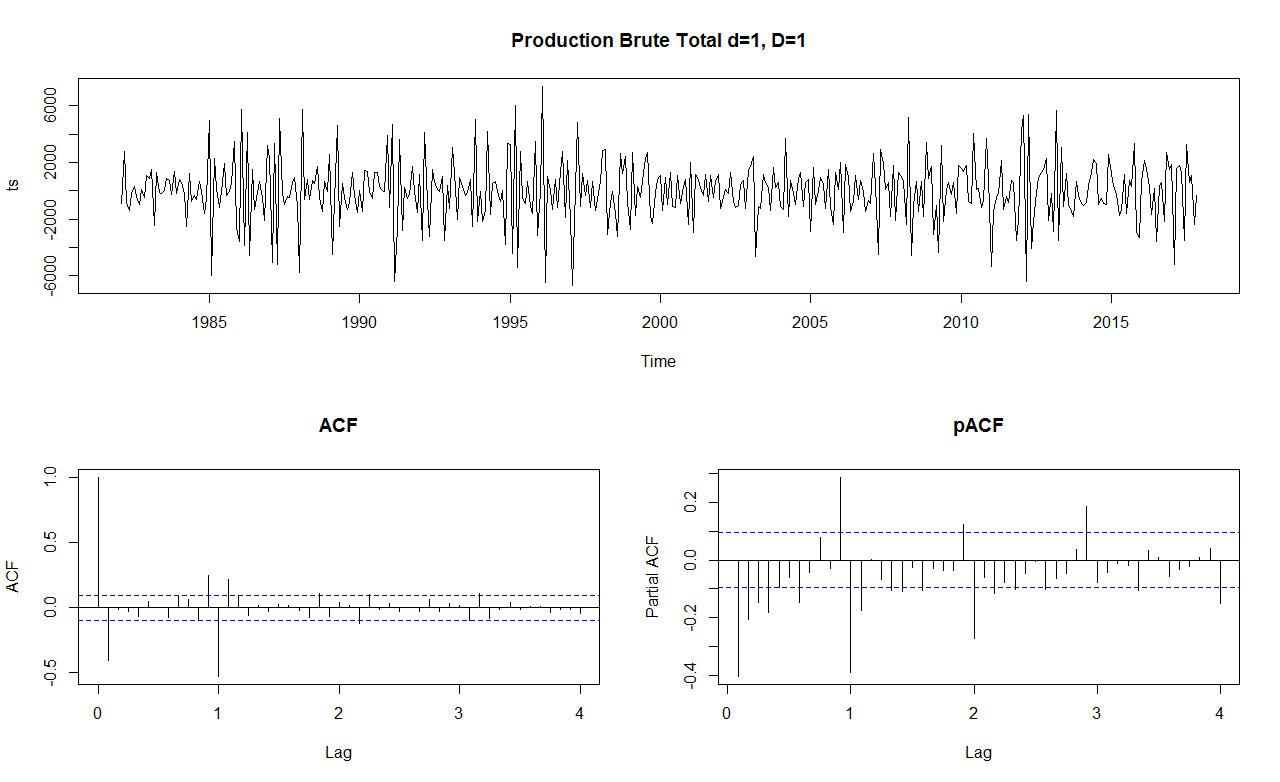
Fitting

Test significativité

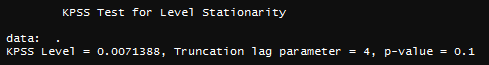
Prévisions

Rien qu’en se basant sur l’analyse de l’acf, on conclut que les résidus ne sont pas stationnaires. Le test KPSS nous le confirme par la suite. Nous allons donc différencier au premier ordre cette série.





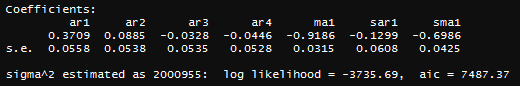
Ce résultat est bien plus satisfaisant. L’acff et le pACF présentent tous les 2 une croissance rapide vers 0. Cette série différenciée passe bien le test du KPSS



Nous retenons donc comme paramètres : d=1 et D= 12. D’après l’analyse, nous suspectons :

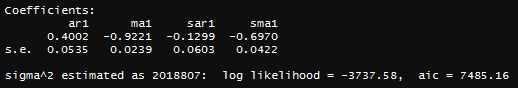
* Une composante AR(4) avec coefficient négatif, du fait des 4 premiers pics négatifs significatifs sur le pACF.
* Une composante SAR(2) avec les pics significatifs aux lags 1 et 2.
* Un MA(1) du fait du premier pic significatif sur l’ACF
* Un SMA(1) du fait du pic significatif u lag 1 sur l’ACF

Nous estimons alors un . Nous obtenons :

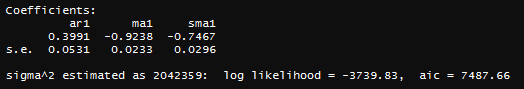


Les ar2/3/4 ne sont pas significatifs. Toutes les autres composantes sont significatives vec

Nous estimons donc un . Nous obtenons :

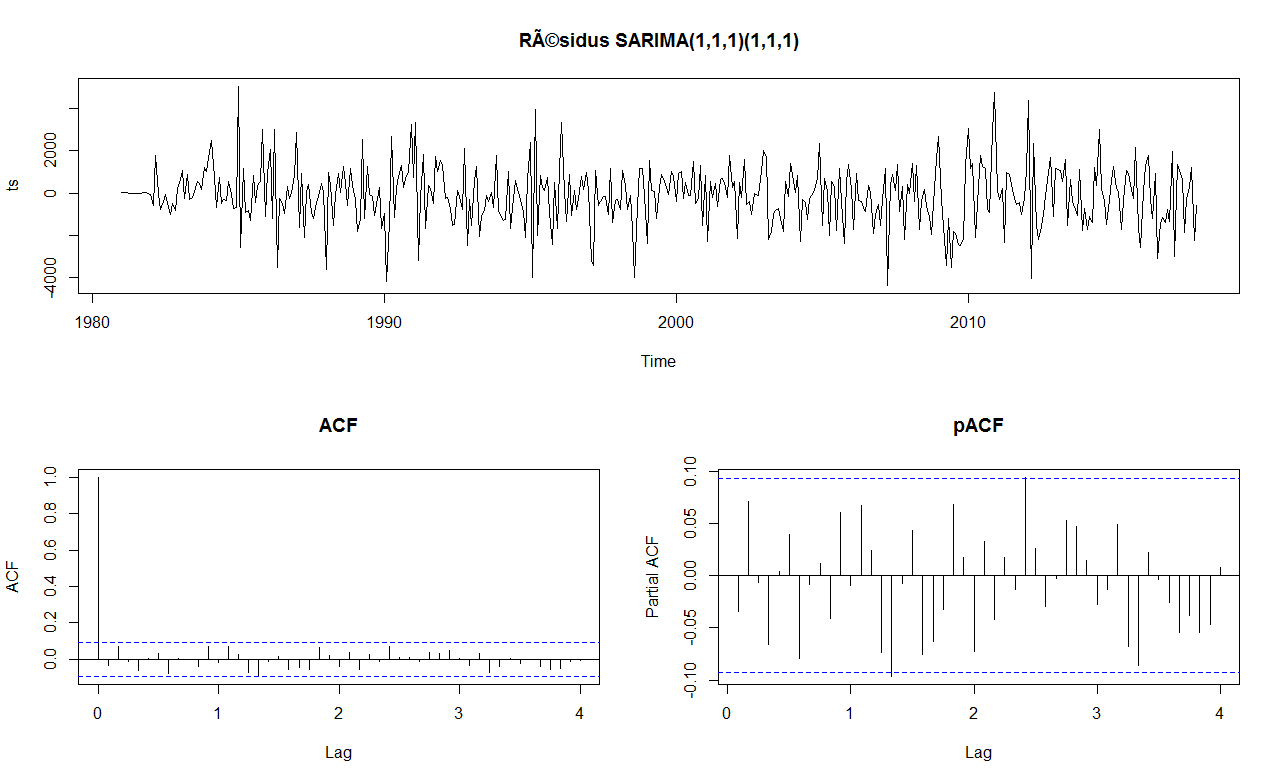


Tous les coefficients sont significatifs. Le SAR(1) est vraiment à la limite. On peut essayer de fit un



Mais l’AIC est plus grand que celui du . Nous retenons donc cette modélisation.

Nous affichons ses résidus et l’ACF et le pACF correspondant.

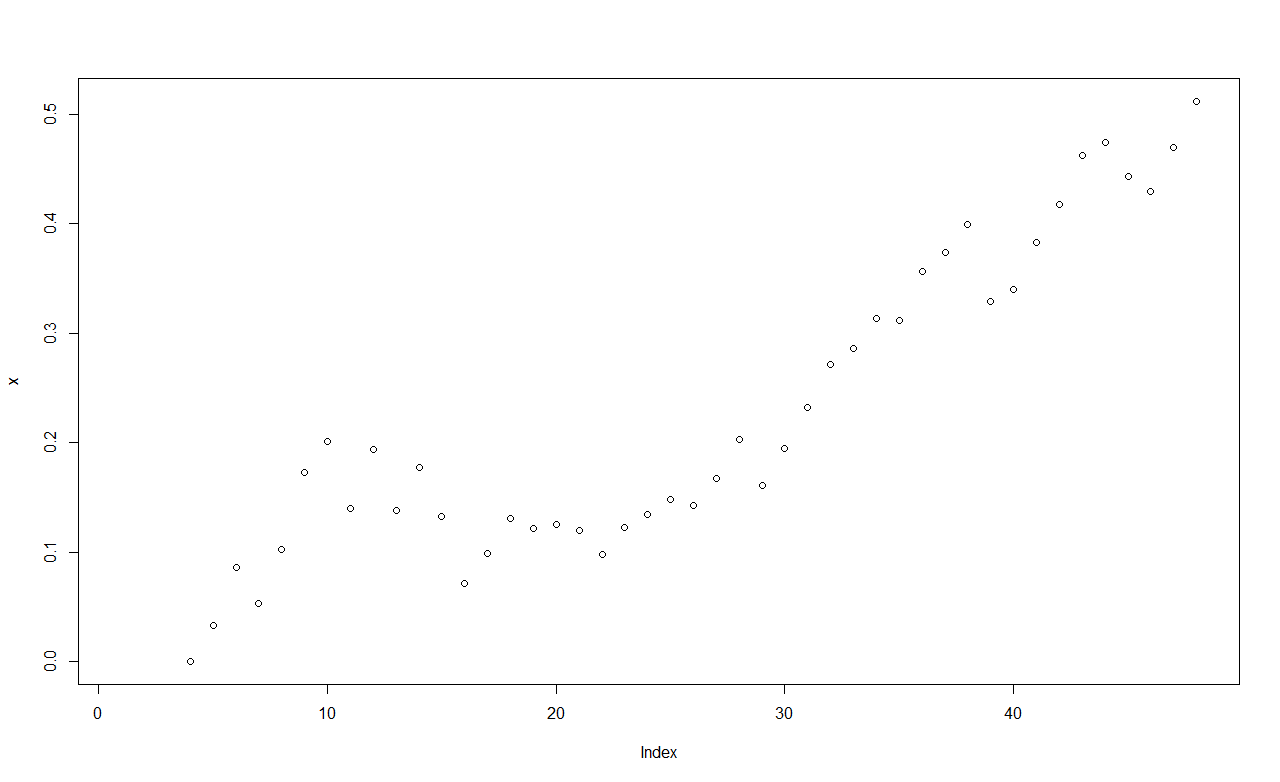


A première vue l’acf et le pACF sont compatibles avec l’hypothèse de blancheur des résidus.

Les résidus passent le test KPSS pour leur stationnarité :

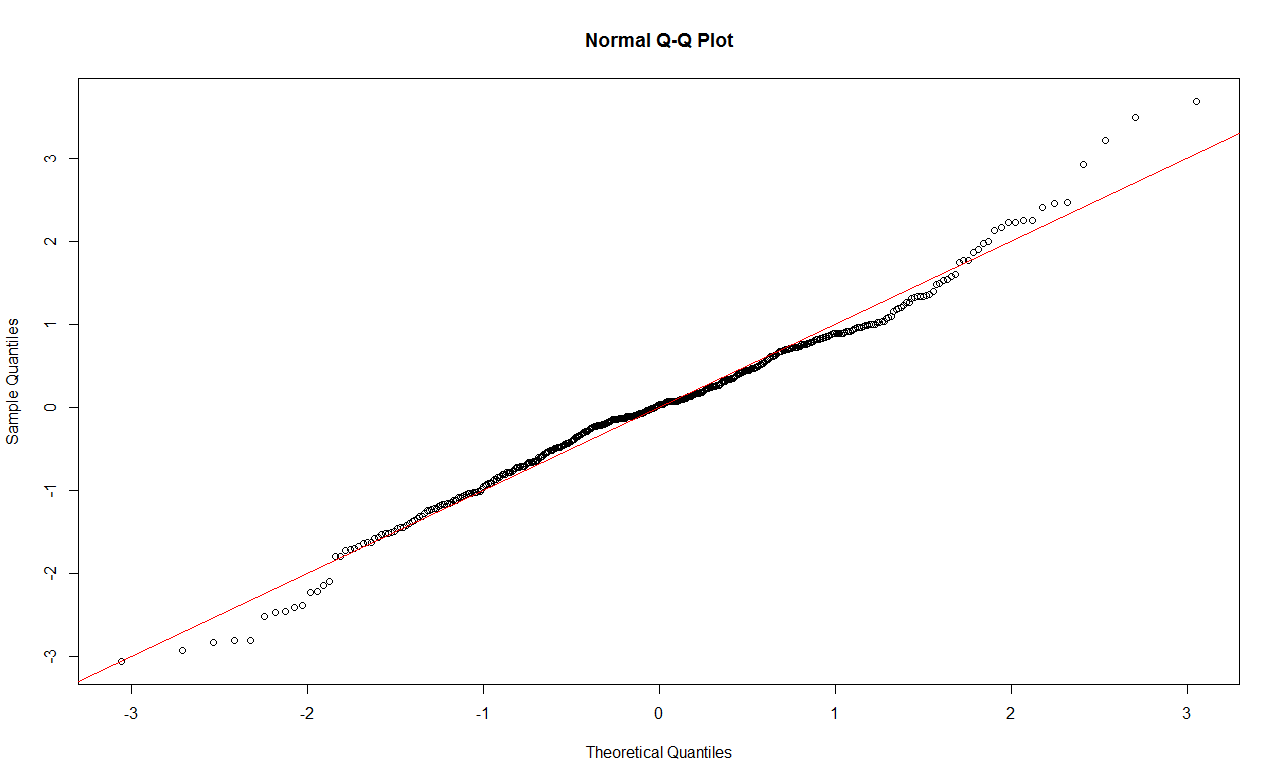


Nous testons la présence d’autocorrélations dans les résidus en affichant les p-values du test de Ljung-Box pour lequel nous avons retenu p+q+P+Q comme paramètre de degré de liberté.

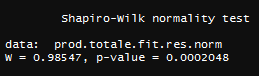
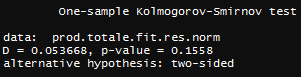


Le premier p-value est inférieur à 0.05, indiquant la présence d’autocorrélation. Les résidus sont donc un peu autocorrélés.

Enfin on teste la normalité des résidus. On trace que qqnorm et on réalise les test de Kolmogorov-Smirnov et de Shapiro-Wilk.

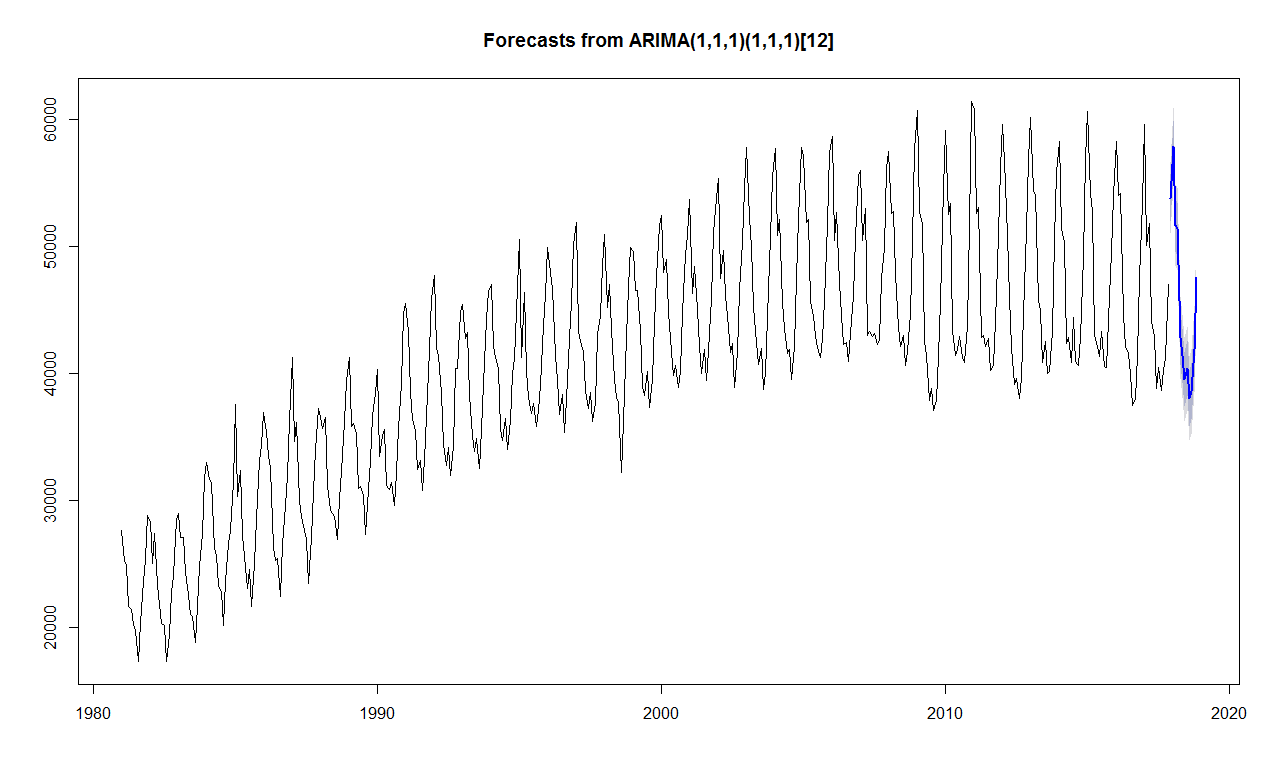


Les points semblent plutôt alignés sur la première bissectrice du plan, mais présentent des irrégularités sur les queues.

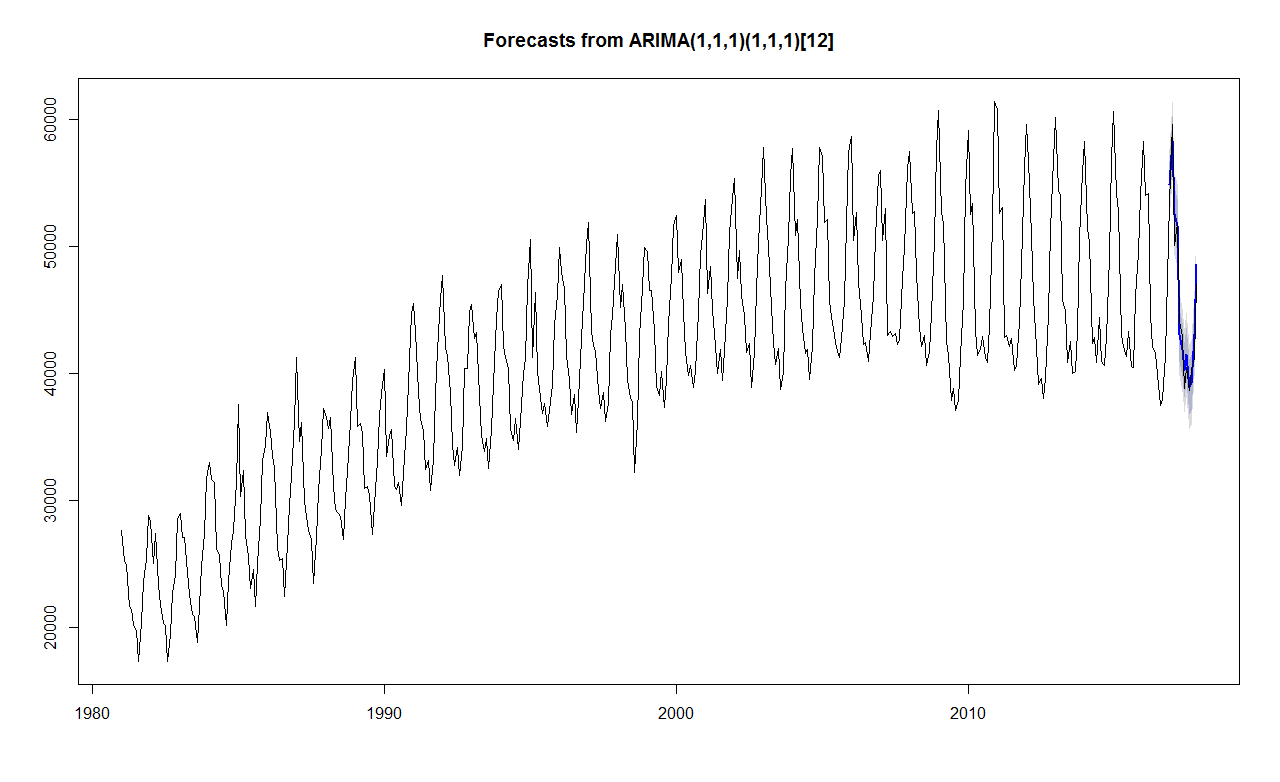


Les résidus passent le test de Kolmogorov-Smirnov, ne nous permettant pas de conclure en la défaveur de leur normalité, mais ne passent pas le test de Shapiro-Wilk.

Nous projetons ce modèle sur 12, et affichons les prévisions :



Nous réalisions également un back test qui se rélvèle être plutôt concluant pour notre modèle retenu :



# Partie 2 : Importation totale d’électricité

## Présentation de la série

La série est disponible (figure 1).

Les commentaires suivant sont fait à l’œil sur le graph :

* La série admettrait difficilement un trend, le cas échéant, on ne serait pas sur un trend linéaire
* La série admet difficilement une saisonnalité dans les valeurs hautes des piques. Ces oscillations ne sont pas constantes
* Le parttern de l’amplitude des oscillations n’est pas discernable aisément

Nous sommes dans un cas peu favorable à la modélisation : on ne saurait pas prolonger la série à la main contrairement aux deux autres que nous étudions ici.

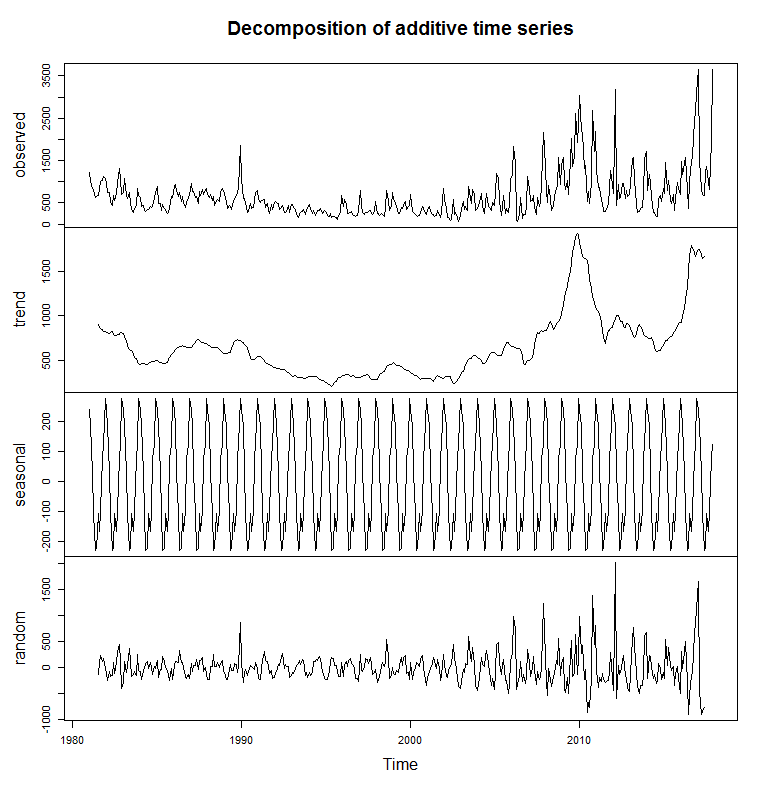
On se demande d’ailleurs s’il n’est pas plus favorable de considérer les données qu’à partir d’un certain seuil, nous avons préféré suivre le sujet et prendre toute la série [remarque à valid » avec Christian Robert]

## Premières approches

### Une série difficile a expliqué en l’état

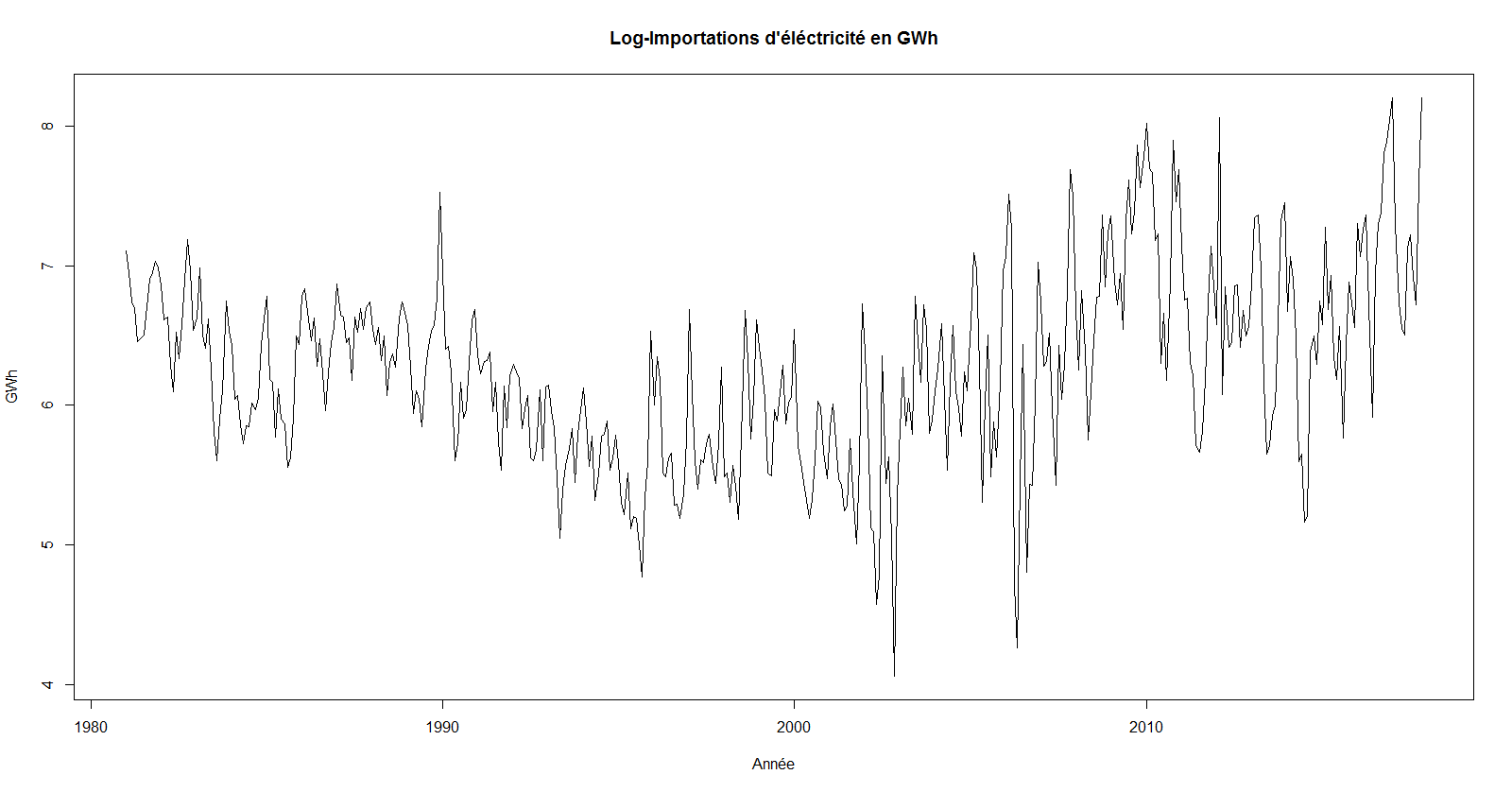
Nous avons cette série que l’on pourrait rattaché aux autres séries de ce dataset, cela nous obligerait à toutes les estimer et le fait qu’elles soient corrélé fait qu’on ne maitrisera probablement moins bien ce modèle que ceux que nous allons dresser plus tard.

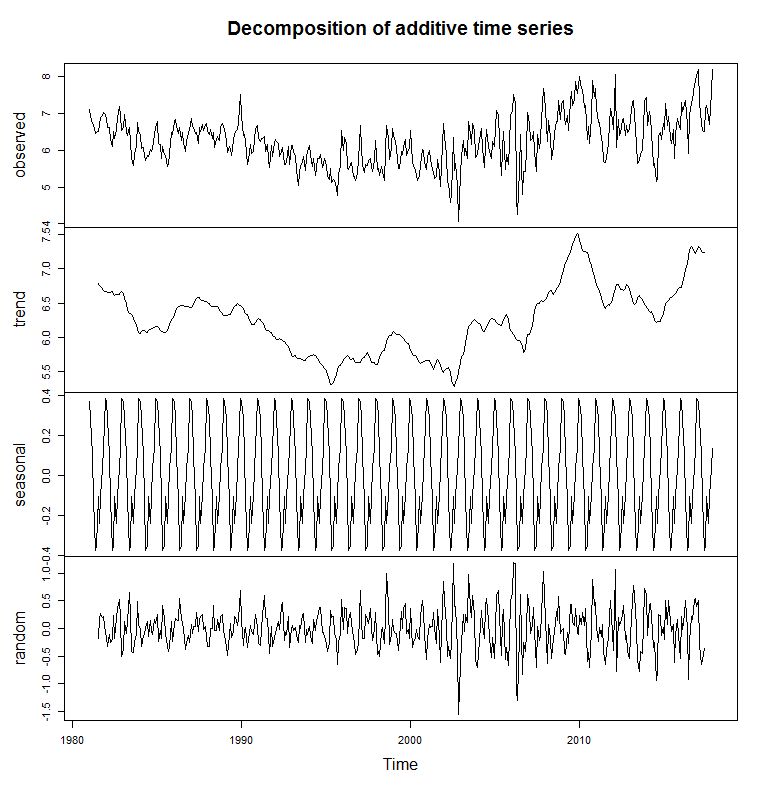
### La décomposition additive est mauvaise, passons au logarithme



Nous utilisons la fonction decompose(), en prenant soins de choisir une décomposition additive pour réfuter d’emblée cette hypothèse. Le bruit n’est pas stationnaire et le trend n’est manifestement pas linéaire. On réfute également que le modèle additif peut être probant sur une troncature de cette série. Nous passons au logarithme\_e de la série. Le logarithme est strictement croissant pour les valeurs strictement positives et s’inverse très bien. Il permet également de passer à un modèle multiplicatif et les intervalles de confiance que nous obtiendrons se propagent aisément

La série sur laquelle nous effectuerons la suite des analyses est la suivante :

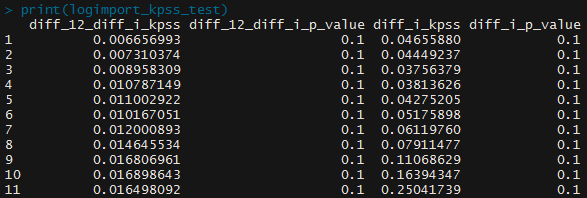




La décomposition est bien plus efficace, le trend n’est malheureusement pas linéaire, nous ne garderons pas de modèle car nous ne serons pas en projeter le trend. Le bruit n’étant pas blanc, nous décidons d’abandonner cette modélisation.

### Modélisation SARIMA

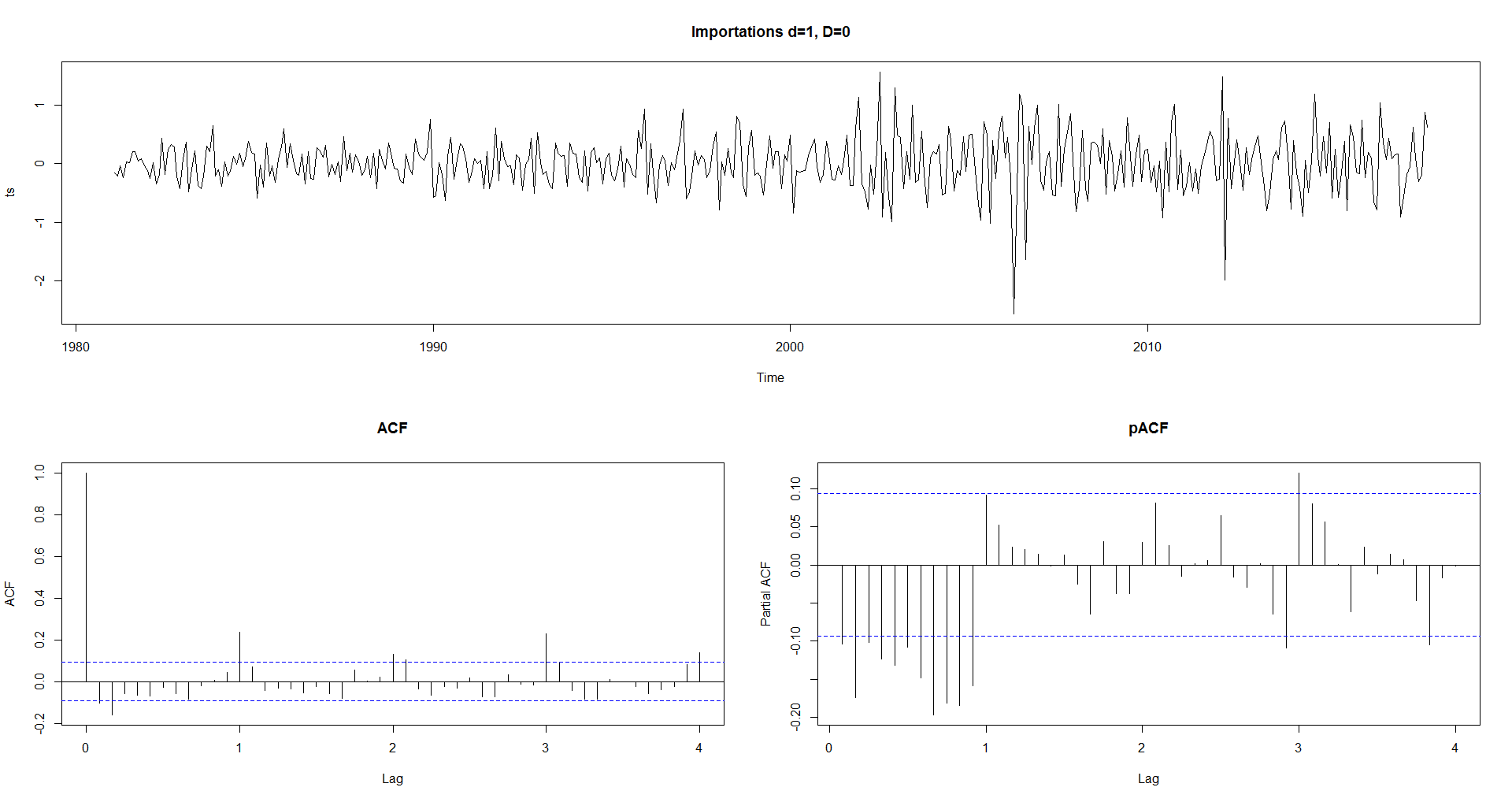
Le processus a été expliqué en annexe. Nous avons cherché à trouver de la stationnarité en différentiant la série à tous les ordres possible. Et nous en calculons le KPSS. Voici la matrice des résultats que nous avons obtenus :



Les p-values sont toutes supérieures à 0.1, le KPSS ne sait réfuter l’absence de stationnarité.

On va différencier tel que les des résidus avec des acf et pcf se rapprochent de bruits blancs, on jouera ensuite avec p, q, P et Q pour avoir la meilleure modélisation possible

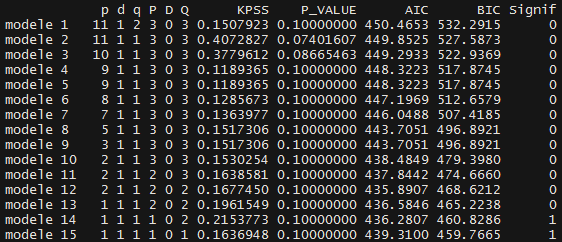
On trouve avec d = 1 et D = 0 les résidus suivants :



D’après cette analyse, nous émettons les remarques suivantes :

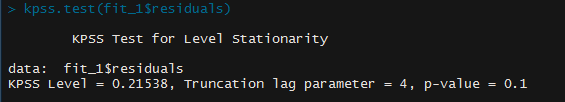
* Une composante en AR(11) avec les onze piques significatifs du pACF
* Une composante en SAR(3) avec le pique significatif en trois du pACF
* Un MA(2) du fait des deux petits pics de l’ACF
* Un SMA(4) du fait des quatre pics périodiques de l’ACF

Nous dressons un . Intuitivement, nous obtiendrons un modèle avec moins de paramètre, nous allons donc écrémer ce modèle petit à petit. Récapitulons les itérations faites et le modèle obtenu :



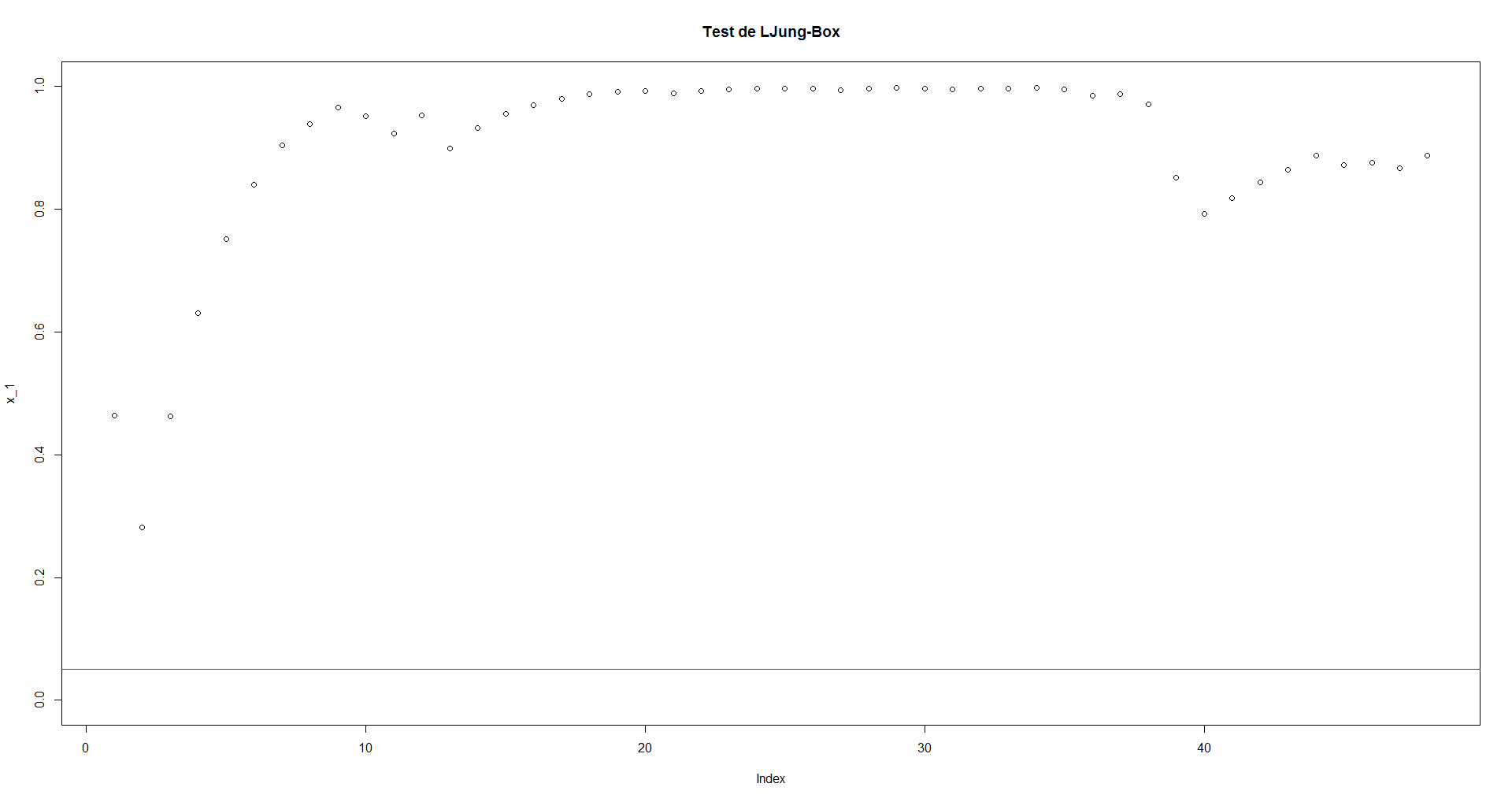
On se trouve avec deux bons modèles, le modèle 14 et le modèle 15. Nous avons optimisé sur critère du BIC. Modèle 14 se retrouve avec environ 1.1 points d’écart de BIC du modèle 15. Le modèle 15 par contre a 3 points d’AIC de moins. Il paraît plus raisonnable de prendre le modèle 14 car on se satisfait d’un écart de BIC qu’au dessus de 2 points en pratique.

On a le test du KPSS suivant



Rien ne nous permet d’affirmer que la série n’est pas stationnaire

Le test de LJung box est franchis haut la main, il n’y a plus d’autocorrelation dans nos résidus :



# Annexes

Annexe 1 : sélection pratique de nos SARIMA

Sélection d’un paramètre à optimiser : BIC, AIC, etc…

**NON**

**OUI**

Analyse :

* ACF (pour q,Q)
* pACF (pour p, p)

Premier fit

Test significativité

Si série saisonnière :

Différenciation saisonnalité : D = 1

**NON**

Test stationnarité : KPSS

Différenciation 1er ordre

Test stationnarité : KPSS

Tests validation du modèle :

* Normalité des résidus : Kolmogrov – Smirnov, Shapiro-Wilk
* Blancheur des résidus : Ljung-Box

Prévisions