Analyse de la consommation d'électricité en France

Mémoire M2 IFMA Séries temporelles

Céline Nguyen-Tu, Jimmy Ly, Dimitrij Muller, Aras Chaigne

Janvier 2025

Résumé

La consommation d'énergie, en particulier d'électricité, est un indicateur crucial de l'activité économique et sociale d'un pays. En France, c'est une source d'énergie qui joue un rôle fondamental dans la vie quotidienne des citoyens et dans le fonctionnement des industries. Comprendre les tendances, les variations saisonnières et les anomalies potentielles dans la consommation d'électricité est essentiel pour une planification énergétique efficace, la gestion des ressources et la mise en œuvre de politiques énergétiques durables. Ce projet vise à analyser les séries temporelles de la consommation d'électricité en France. Les objectifs spécifiques incluent la compréhension des tendances de consommation des Français sur les dernières années, une modélisation ajustée de cette dernière en vue de la prévision de la consommation future.

Table des matières

1 Présentation des données

Les données que nous avons choisi d'analyser sont les données de consommation d'électricité en France métropolitaine, hors Corse. Notre jeu de données couvre la période de 2012 à 2024 et provient du site data.gouv.fr.

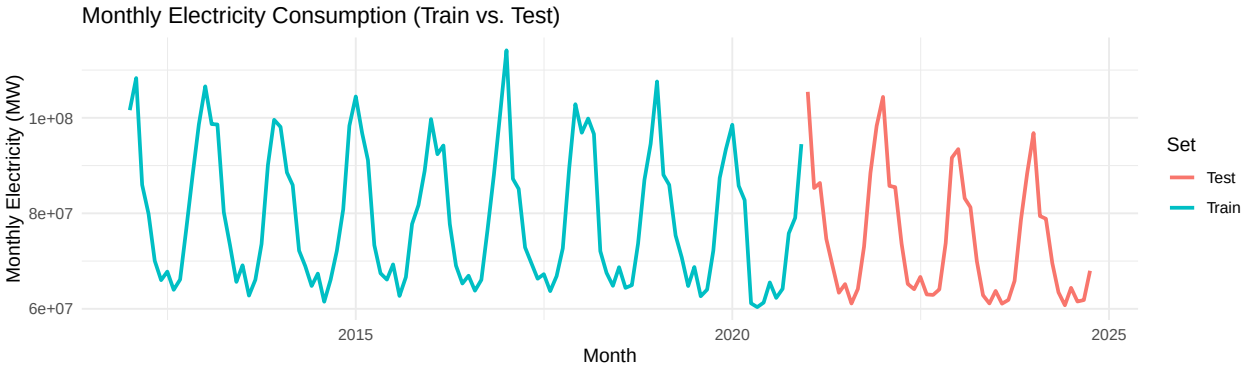
Le jeu de données comprend plusieurs variables que nous considérerons pour analyser les tendances de consommation. Les variables essentielles pour notre analyse sont :

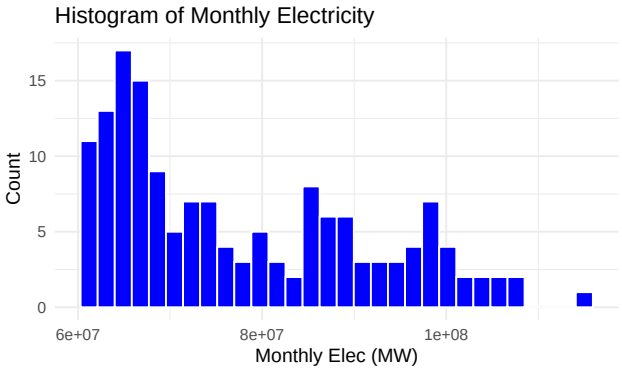
- Horodatage : Indique la date et l'heure de chaque enregistrement de la consommation, avec une granularité de 30 minutes.
- Consommation d'électricité (MW) : Représente la consommation brute d'électricité mesurée en mégawatts (MW) pour chaque intervalle de 30 minutes.
- Statut des données : Indique si les données sont définitives ou intermédiaires. Les années 2012 à 2020 sont au statut définitif, signifiant qu'elles ont été vérifiées et validées par des équipes terrain. Les données de 2021 à 2024 sont au statut intermédiaire, donc susceptibles de modifications ultérieures.

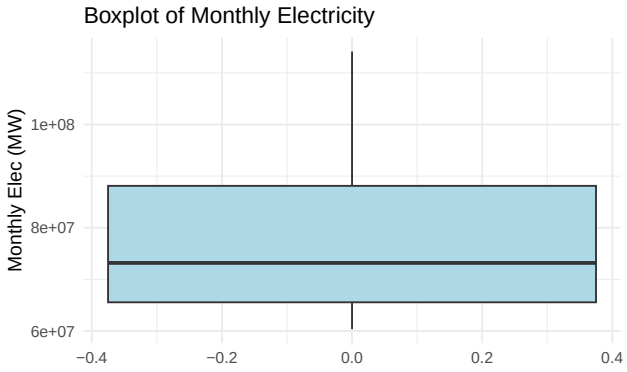
Pour la modélisation, nous avons divisé les données en ensembles d'entraînement et de test. Les données définitives (2012-2020) constituent l'ensemble d'entraînement, tandis que les données intermédiaires (2021-2024) forment l'ensemble de test. Cette séparation permet d'évaluer la capacité du modèle à généraliser sur

des données récentes des tendances de consommation. Nous tiendrons compte du fait que les données de 2021 à 2024, étant intermédiaires et susceptibles de modifications, peuvent contenir des anomalies. Cela pourrait affecter l'évaluation des performances du modèle, car les prédictions seraient comparées à des données potentiellement inexactes.

Nous avons agrégé les données à une échelle mensuelle pour lisser les variations quotidiennes et hebdomadaires. Cela nous permet de mieux identifier les tendances saisonnières et annuelles qu'avec une granularité plus courte. Cette approche réduit le bruit dans les données directement impactées par des variables exogènes comme la température, les weekends et les jours de vacances, et permet une analyse plus claire des schémas de consommation à long terme. Cette granularité est également plus visible sur un graphique à long terme.







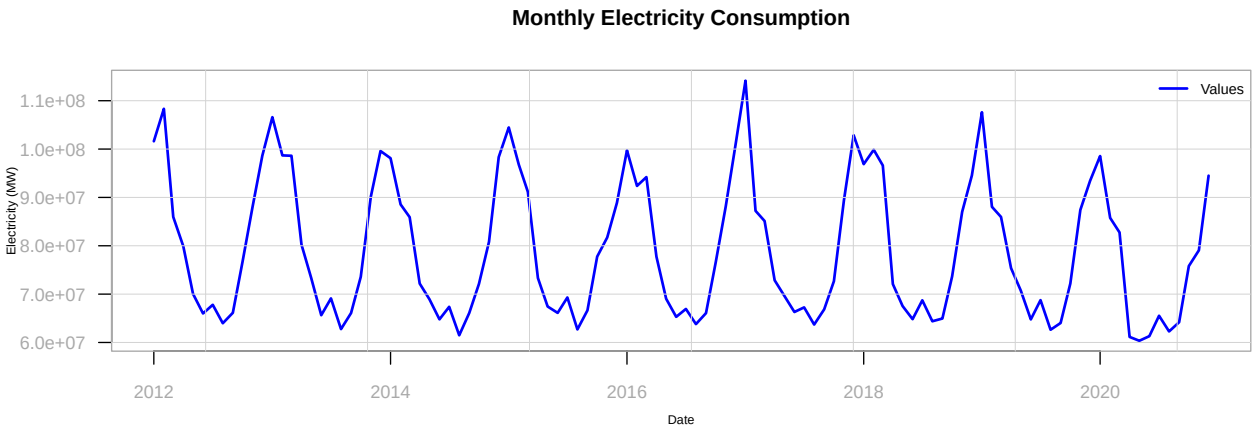
L'histogramme révèle que la consommation mensuelle se situe entre 6×10^7 MW et 8×10^7 MW. La distribution possède une queue à droite qui indique la présence de valeurs de consommation plus élevées mais de plus en plus rares. On retrouve des pics sur la queue de la distribution au-delà de 1×10^8 MW, probablement liés à des événements spécifiques qui ont nécessité plus d'électricité. Sans considérer ces événements rares, la consommation semble suivre une loi gamma.

Le boxplot montre que la médiane est légèrement supérieure à 7.5×10^7 MW. Les 50 % des valeurs centrales, représentés par la boîte, se situent entre 6.5×10^7 MW et 9×10^7 MW, donc la distribution est plutôt concentrée, l'écart-type est faible. Nous n'observons aucune valeur aberrante de consommation.

2 Analyse préliminaire

Dans cette section, nous allons étudier la structure et la stationnarité de la consommation d'électricité en France, une étape essentielle avant de passer à la modélisation. En effet, elle est importante car la plupart des modèles de séries temporelles présupposent que la série est stationnaire. Si une série est non stationnaire, elle doit être rendue stationnaire avant de pouvoir être modélisée efficacement. Cette première analyse permettra également de choisir une ou plusieurs approches adaptées à notre problème.

Commençons par utiliser des visualisations graphiques de la série afin de repérer visuellement les tendances ou cycles qui pourraient indiquer une non-stationnarité. Visualisons les données de l'échantillon d'entraînement :



À première vue, une saisonnalité se dégage de la courbe de consommation d'électricité. Cette observation était anticipée en raison du contexte et est expliquée par les variations des besoins énergétiques dûs aux températures et aux durées des journées dans les régions tempérées comme en France. Les variations saisonnières apparaissent à peu près à la même amplitude. La série oscille autour d'une moyenne non nulle représentant un signe de stationnarité et on n'observe pas de rupture structurelle apparente (changements soudains dans la tendance ou la variance).

On dispose de plusieurs outils et métriques pour vérifier la caractéristique de stationnarité. Les moyenne et variance glissantes ainsi que la fonction d'autocorrélation (ACF) sont des indicateurs de la stationnarité. En complément, on utilisera le test de Dickey-Fuller augmenté (ADF) et le test KPSS pour vérifier formellement le caractère stationnaire, bien qu'ils comportent certaines limites. Cette analyse permettra de déterminer s'il est nécessaire de transformer ou de stationnariser la série avant la modélisation mais également s'il est pertinent d'en extraire une partie tendancielle et saisonnière avant de modéliser la composante résiduelle stationnaire.

Test de Dickey-Fuller Augmenté

Le test de Dickey-Fuller Augmenté (ADF) est un test statistique utilisé pour vérifier la stationnarité d'une série temporelle. C'est une extension du test de Dickey-Fuller qui inclut des termes additionnels pour capturer des dynamiques plus complexes. La série est modélisée comme :

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \varepsilon_t$$

où $\Delta y_t = y_t - y_{t-1}$ est la différence de la série à l'instant t, y_{t-1} est la valeur de la série à l'instant $t-1$, α_i sont les coefficients du modèle autorégressif pour les lags passés et ε_t est le terme d'erreur ou résidu.
La statistique de test d'ADF est fondée sur l'estimation du coefficient β_1 dans ce modèle autorégressif. L'idée est de tester si $\beta_1 = 0$ en utilisant la statistique de la régression $t(\beta_1)$, qui est obtenue par la formule :

$$t(\beta_1) = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

où $\hat{\beta}_1$ est l'estimation du coefficient β_1 de la régression et $SE(\hat{\beta}_1)$ est l'erreur standard de $\hat{\beta}_1$.
- Hypothèse nulle H_0 : La série temporelle a une racine unitaire, c'est-à-dire qu'elle n'est pas stationnaire.
- Hypothèse alternative H_1 : La série temporelle est stationnaire.

Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin)

Le test KPSS est un test statistique utilisé pour vérifier la stationnarité d'une série temporelle. Contrairement au test de Dickey-Fuller, le test KPSS a comme hypothèse nulle que la série est stationnaire autour d'une moyenne ou d'une tendance déterministe. La série est modélisée comme :

$$y_t = \mu + \nu_t + \varepsilon_t$$

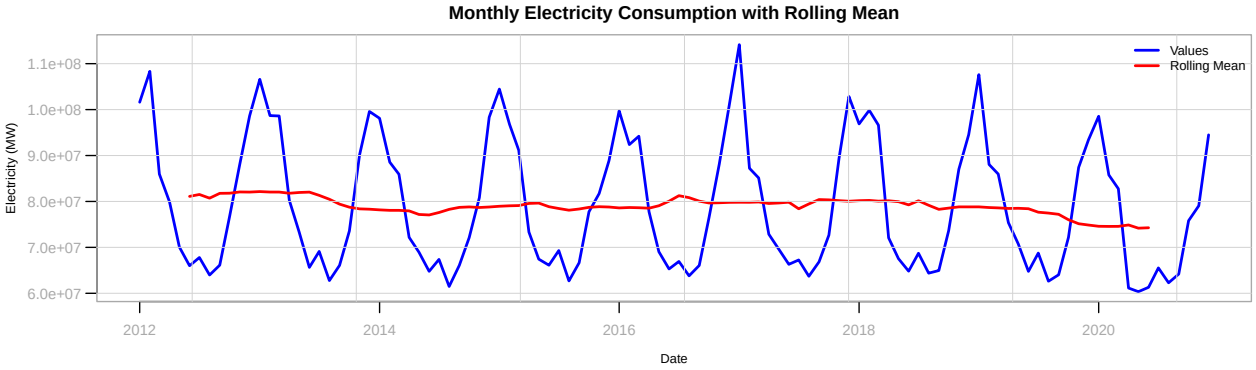
où μ est la constante de la tendance, ν_t est une tendance déterministe (ou une moyenne constante pour une série sans tendance) et ε_t est le terme d'erreur blanc (stationnaire et indépendant).

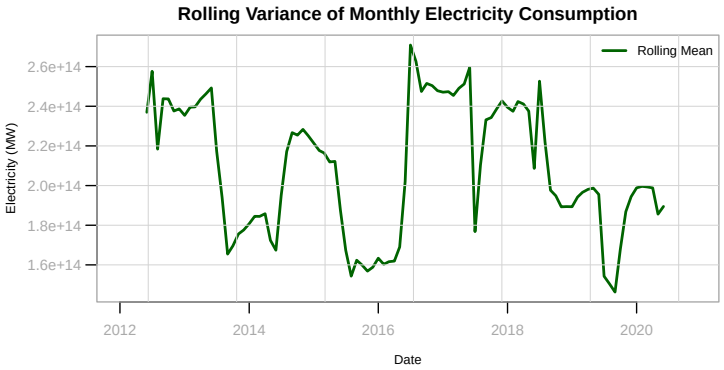
La statistique de test calculée est la suivante :

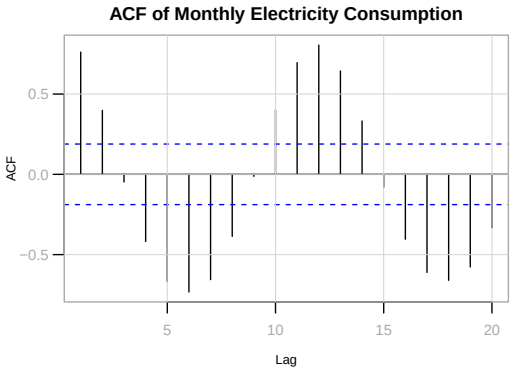
$$S = \frac{1}{n} \sum_{t=1}^{n} \hat{\varepsilon}_t^2$$

où $\hat{\varepsilon}_t$ est l'estimation des résidus du modèle de tendance.
- Hypothèse nulle H_0 : La série temporelle est stationnaire (soit autour d'une moyenne, soit autour d'une tendance déterministe).
- Hypothèse alternative H_1 : La série temporelle a une racine unitaire, c'est-à-dire qu'elle n'est pas stationnaire.

Analysons la stationnarité de la série temporelle de la consommation d'électricité :







ADF Test on Monthly Electricity Consumption p-value : 0.01
KPSS Test on Monthly Electricity Consumption p-value : 0.1

La moyenne glissante semble suivre une droite indiquant une très légère tendance linéaire. La forte composante saisonnière est confirmé par le graphe ACF montrant des autocorrélations significatives à des lags multiples de 12 mois. Puisque l'ACF ne tend pas rapidement vers zéro, on ne peut pas confirmer la stationnarité. La p-value du test ADF (p-value $= 0.01 < 0.005$) rejette l'hypothèse de non-stationnarité et la p-value du test KPSS (p-value $= 0.1 > 0.005$) ne rejette pas l'hypothèse de stationnarité. Cependant, ces tests sont influencés par une forte saisonnalité stable et prévisible, et peuvent ne pas détecter correctement la non-stationnarité dans ce cas. C'est pourquoi leur résultats sont contre-intuitifs et contraires à l'analyse de ACF. Néanmoins, cela nous donne l'indication que la série pourrait être stationnaire après ajustement de la saisonnalité, comme une différenciation saisonnière ou l'extraction de la composante saisonnière.

Nous pouvons alors envisager plusieurs approches de modélisation :
1. Il est possible d'appliquer une différenciation saisonnière pour tenter de rendre la série stationnaire et ensuite utiliser un modèle de type ARMA/ARIMA. Cependant, cela entraîne une perte d'informations sur la tendance et la saisonnalité, ce qui peut être utile pour l'analyse et l'interprétation du modèle. De plus, cette approche peut être inefficace si la saisonnalité est évolutive, ce que l'on suppose à priori au vu des facteurs exogènes pouvant influencer la consommation.

2. Une autre approche consiste à décomposer la série pour modéliser séparément la tendance, la saisonnalité et les résidus stationnaires. Cette méthode peut offrir une meilleure compréhension de la série et un meilleur ajustement. Elle permet de capturer des structures temporelles plus sophistiquées et peut améliorer la précision des prévisions. De plus, extraire la tendance et la saisonnalité permet souvent d'appliquer des modèles plus simples à ajuster comme AR, MA ou ARMA.
3. Le modèle SARIMA est particulièrement adapté aux séries temporelles présentant des saisons. Il permet de modéliser simultanément la tendance, la saisonnalité et les résidus, tout en évitant la nécessité de stationnariser manuellement la série. Cette approche combine la capacité de capturer les composantes essentielles de la série temporelle tout en facilitant le processus de modélisation. Ce modèle est en revanche limité pour des saisonnalités évolutives.

Les deux dernières approches seront étudiées dans cette étude afin d'en déduire plusieurs modèles efficaces pour la prévision. Cette analyse permettra à la fois de mettre en pratique une large gamme de notions étudiées dans le cadre du cours et de choisir un modèle de prédiction le plus performant.

3 Modélisation par décomposition additive

3.1 Composantes tendance-saisonnalité-résidus

Cette étape nous permet d'identifier clairement les éléments déterministes (tendance et saisonnalité) et les éléments aléatoires (résidus) de la série. La tendance reflète la direction générale de la série, qui peut être croissante, décroissante ou plate. La saisonnalité capture les variations régulières et récurrentes dans la série, généralement liées à des facteurs externes comme la période de l'année, le mois ou la semaine. Enfin, les résidus représentent les fluctuations irrégulières de la série, qui ne peuvent pas être expliquées par la tendance ou la saisonnalité. La décomposition est essentielle pour mieux comprendre les dynamiques sous-jacentes de la série et pour permettre une modélisation plus adaptée et précise de ses composants.

Puisque l'amplitude des variations saisonnières reste constante sur toute la durée de la série, une décomposition additive est appropriée :

$$X_t = T_t + S_t + R_t$$

où X_t est la série temporelle de la consommation d'électricité, T_t est la composante de tendance, S_t est la composante de saisonnière qui peut être régulière (fixe) ou évolutive et R_t est la composante résiduelle ou l'erreur.

Les techniques pour estimer la tendance et la saisonnalité peuvent être semi-paramétriques (typiquement régression linéaire) ou non-paramétriques (moyenne mobiles simples ou pondérées, régressions locales, régressions par splines, lissages exponentiels, convolution par un noyau, décomposition dans une base d'ondelettes,...). Les résidus sont les valeurs restantes après avoir enlevé la tendance et la saisonnalité. Idéalement, ces résidus devraient être proches de 0 et ne montrer aucune structure discernable.

3.1.1 Modèle 1 - Méthode STL

La méthode STL (Seasonal-Trend decomposition using Loess) est une méthode de décomposition qui utilise la technique LOESS pour séparer une série chronologique en ses composantes de tendance, saisonnières et résiduelles. LOESS (Locally Estimated Scatterplot Smoothing) ou LOWESS (Locally Weighted Estimated Scatterplot Smoothing) est une méthode de régression non paramétrique utilisée pour lisser les données.

La régression localisée LOESS s'obtient en fixant une taille de fenêtre et on fait une régression linéaire ou polynomiale de X dans la fenêtre que l'on fait glisser. Cette méthode ne suppose pas connues les fonctions pouvant composer la série (comme dans le cas de la régression), mais juste une certaine régularité de la fonction. La fenêtre doit être suffisamment large pour capturer les tendances à long terme mais pas trop large pour ne pas lisser des cycles saisonniers.

Pour chaque point d'intérêt x_i de la série de donnée, les points de données voisins dans une fenêtre définie sont utilisés pour ajuster un modèle local. Soit (x_j, y_j) les points de données voisins de x_i dans une fenêtre

W. Le modèle de régression linéaire local est de la forme :

$$y_j \approx \beta_0(x_i) + \beta_1(x_i)(x_j - x_i)$$

où $\beta_0(x_i)$ et $\beta_1(x_i)$ sont les coefficients de la régression locale autour de x_i.

Chaque point x_j dans la fenêtre W est pondéré par une fonction de poids $w(x_j, x_i)$ (par exemple, la fonction tricube) qui dépend de la distance entre x_j et x_i. Puis, les coefficients $\beta_0(x_i)$ et $\beta_1(x_i)$ sont estimés en minimisant la somme pondérée des carrés des résidus :

$$\min_{\beta_0(x_i),\beta_1(x_i)} \sum_{j \in W} w(x_j, x_i) \left(y_j - \beta_0(x_i) - \beta_1(x_i)(x_j - x_i)\right)^2$$

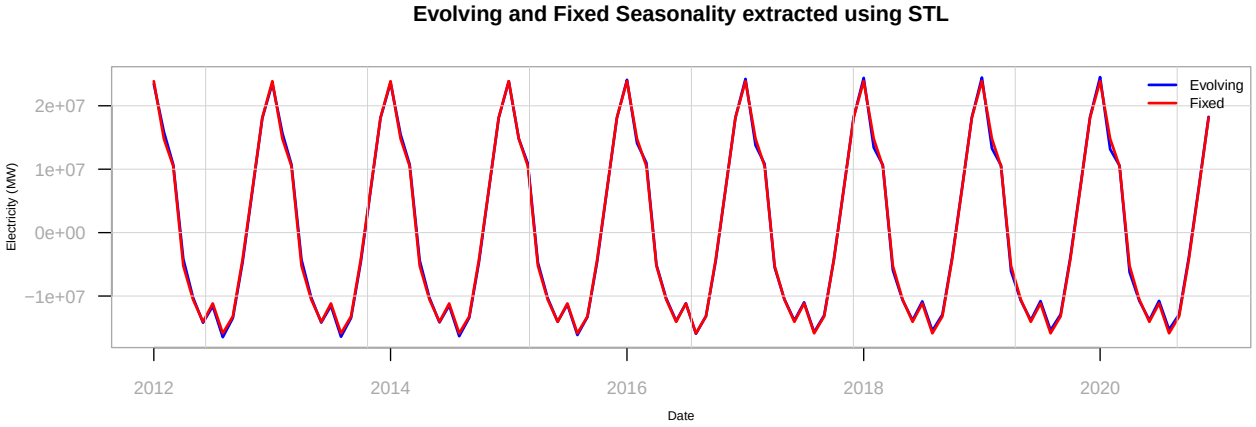
Les valeurs lissées $\hat{y}_i$ pour chaque x_i sont obtenues en évaluant le modèle local ajusté :

$$\hat{y}_i = \beta_0(x_i)$$

Ce processus permet de produire une courbe lisse qui s'adapte localement aux variations des données. Pour capturer la saisonalité, le lissage LOESS est appliqué sur les sous-séries saisonnières. Dans le cas de données mensuelles, ce sont les données correspondant au même mois pour chaque année. La méthode STL consiste à effectuer un lissage local de la tendance et de la saisonnalité, itératant ces procédures jusqu'à convergence. Cette technique permet de capturer les variations à court et à long terme des données, ce qui en fait une méthode robuste pour décomposer les séries temporelles présentant des modèles complexes et non linéaires.

Cependant, bien que cette méthode soit efficace pour identifier et comprendre les structures sous-jacentes, elle ne permet pas directement de faire des prévisions puisqu'elle repose sur des techniques de lissage non paramétriques. Pour effectuer des prévisions, il sera nécessaire d'adopter des modèles supplémentaires pour chaque composante.

La fonction `stl` de `stats` permet d'effectuer cette décomposition et propose également d'extraire une saisonnalité supposée régulière. Dans ce cas, le lissage est remplacé par une moyenne de groupe. On peut donc comparer les séries des saisonnalités obtenues dans les cas où l'on suppose la saisonnalité régulière ou évolutive. Les décompositions complètes calculées par des fonctions `R` sont disponibles en annexes.



On observe que la saisonnalité est légèrement évolutive au fil des années. En particulier, cette évolution semble linéaire et est plus visible au niveau des creux et des pics. Pour vérifier si cette évolution est statistiquement significative, on introduit un modèle de régression linéaire avec un terme d'interaction entre les mois et les années (on suppose que l'effet de l'année sur la saisonnalité peut varier en fonction du mois) :

$$\begin{aligned} S_t &= \beta_0 + \beta_1 \cdot CosMonth_t + \beta_2 \cdot Year_t + \beta_3 \cdot (CosMonth_t \times Year_t) + \varepsilon_t \\ CosMonth_t &= \cos(2\pi \cdot Month_t/12) \end{aligned}$$

où S_t est la composante saisonnière de la consommation d'électricité pour le mois t, $CosMonth_t$ est une variable continue qui indique le mois de l'année, $Year_t$ est une variable continue qui représente l'année (pour

tenir compte de l'évolution au fil du temps), $CosMonth_t \times Year_t$ est un terme d'interaction entre les mois et les années pour capturer les effets évolutifs de la saisonnalité et ε_t est l'erreur (résidus) à chaque point temporel.

Dans ce modèle, $CosMonth_t$ représente la transformation trigonométrique appliquée à la variable $Month_t$. Elle permet de capturer les variations saisonnières sous forme de cycles réguliers sur 12 mois, ce qui est approprié car on suppose une relation continue cyclique entre les mois et la consommation d'électricité.

Table 1 – Table ANOVA

term	df	sumsq	meansq	statistic	p.value
cos_month	1	1.305768e+16	1.305768e+16	206.3272504	0.0000000
year	1	7.471500e+07	7.471500e+07	0.0000012	0.9991351
cos_month :year	1	9.071683e+09	9.071683e+09	0.0001433	0.9904704
Residuals	104	6.581773e+15	6.328628e+13	NA	NA

Avec l'analyse de la variance (ANOVA), on évalue l'effet global des facteurs et de leur interactions. On observe que le facteur $Year$ a une F-statistic nulle et n'est pas significatif (p-value = 0.9991) dans le modèle. De manière similaire, le terme d'interaction a une F-statistic très proche de zéro (F-value = 0.00014) et indique l'absence de significativité (p-value = 0.9904). Ainsi, la saisonnalité est principalement expliquée par le motif cyclique des mois et son évolution subtile d'une année à l'autre n'est pas suffisamment prononcée pour être statistiquement significative dans ce modèle.

Cela confirme que l'on peut supposer la saisonnalité régulière pour notre modèle par décomposition, mais également qu'il est pertinent envisager une approche par modélisation SARIMA. À présent, on ne conserve que la décomposition STL avec saisonalité régulière.

La composante tendancière étant approximativement linéaire, on peut la modéliser en ajustant un modèle de régression linéaire sur le temps, ce qui permettra des prévisions.

$$T_t = \beta_0 + \beta_1 \cdot Date_t + \varepsilon_t$$

Table 2 – Coefficients du modèle linéaire

term	estimate	std.error	statistic	p.value
(Intercept)	103936505.260	2416288.4863	43.01494	0
month_start	-1468.743	142.1796	-10.33019	0

Table 3 – Statistiques globales du modèle linéaire

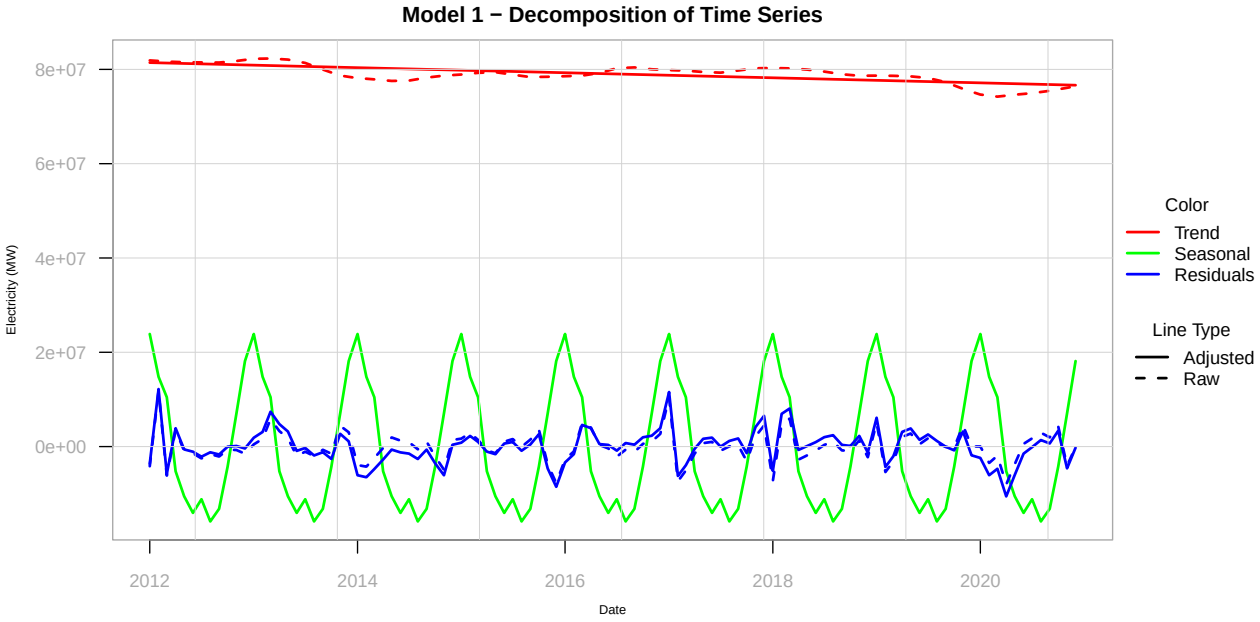
	R_squared	Adjusted_R_squared	F_statistic	p_value_F_statistic
value	0.5016758	0.4969746	106.7129	0

Les p-values associées aux coefficients sont inférieures au seuil de significativité de 5% (< 2e-16), ce qui indique que ces coefficients sont statistiquement significatifs. Cela signifie qu'il existe une relation robuste entre la variable indépendante (le temps) et la variable dépendante (la tendance). De plus, la F-statistic du modèle est élevée (106.7129), ce qui indique que le modèle global est significatif. Autrement dit, la régression linéaire capture bien la tendance dans les données. Dans ce contexte, n'avons pas besoin de vérifier l'hétéroscédasticité

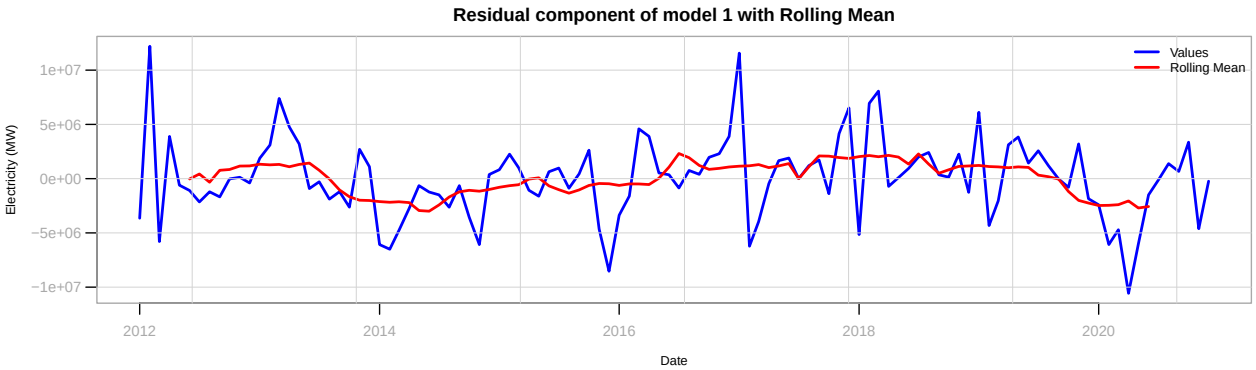
des résidus. En effet, le modèle reste valide pour capturer la composante de tendance car les résidus seront analysés et traités séparément dans la composante résiduelle du modèle global avec des modèles plus adaptés.

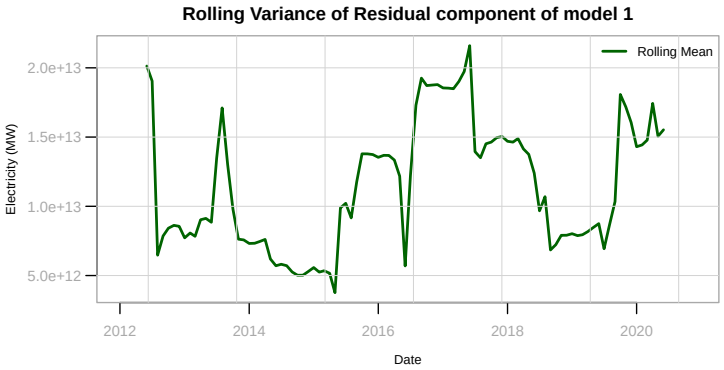
Soit T'_t la droite de régression modélisant la tendance, on peut réajuster la tendance et la composante résiduelle de sorte à récupérer les résidus de la régression linéaire dans la composante résiduelle mais en conservant la moyenne nulle de cette dernière. On obtient alors les composantes réajustées $\hat{T}_t$ et $\hat{R}_t$:

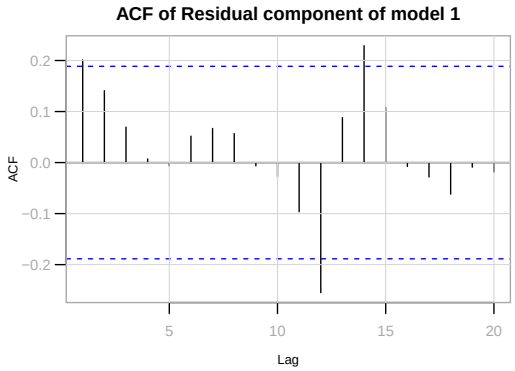
$$
\begin{aligned}
R'_t &= R_t + T_t - T'_t = R_t + \varepsilon_t \\
\hat{R}_t &= R'_t - \bar{R}'_t \qquad \text{et} \qquad \hat{T}_t = T'_t + \bar{R}'_t \\
X_t &= \hat{T}_t + S_t + \hat{R}_t
\end{aligned}
$$



On peut maintenant analyser la stationnarité de la composante résiduelle ajustée.







ADF Test on Residual component of model 1 p-value : 0.0106933
KPSS Test on Residual component of model 1 p-value : 0.1

La série des résidus oscille autour de zéro sans tendance apparente. Sa moyenne glissante reste proche de zéro et stable au fil du temps et sa variance glissante n'augmente pas ou ne diminue pas de manière significative au fil du temps. D'après l'ACF, les autocorrélations chutent à zéro dès le lag 1, cela suggère que les résidus sont indépendants et stationnaires. Les tests ADF et KPSS le confirment, car les p-values sont respectivement inférieures et supérieures à 0.05, ce qui permet de rejeter l'hypothèse d'une racine unitaire et de conclure à la stationnarité des résidus. Par conséquent, nous pouvons considérer que la décomposition a correctement capturé et isolé la tendance et la saisonnalité, et que les résidus peuvent être modélisés séparément à l'aide d'un modèle paramétrique adapté de type ARMA ou potentiellement SARIMA.

3.1.2 Modèle 2 - Décomposition manuelle

En appliquant la méthode STL vue précédemment, on constate que les composantes tendance et saisonnière pourraient être extraites de manière simple et efficace sans utilisation de `stl`. La composante tendance, ayant une forme linéaire, peut être obtenue par une régression linéaire directe sur la série globale. De même, la saisonnalité, étant maintenant considérée comme régulière, peut être isolée en calculant la moyenne de groupe pour chaque mois. En décomposant manuellement la série temporelle avec ces deux techniques, on vise à comparer les résultats obtenus avec ceux de la méthode STL. Cette approche alternative permettra de vérifier la robustesse des prévisions et d'évaluer l'efficacité des deux méthodes de décomposition présentant des caractéristiques similaires.

Cette fois, la régression linéaire se fait sur la série initiale X_t et il n'est pas nécessaire que le modèle soit statistiquement significatif car il ne sert que d'estimateur de la tendance générale laissant les résidus qui incluent les composantes saisonnières et aléatoires :

$$X_t = \beta_0 + \beta_1 \cdot Date_t + \varepsilon_t$$

Elle permet d'isoler la tendance T_t et la série détrendée X_t^d est obtenue en soustrayant la tendance de la

série initiale :

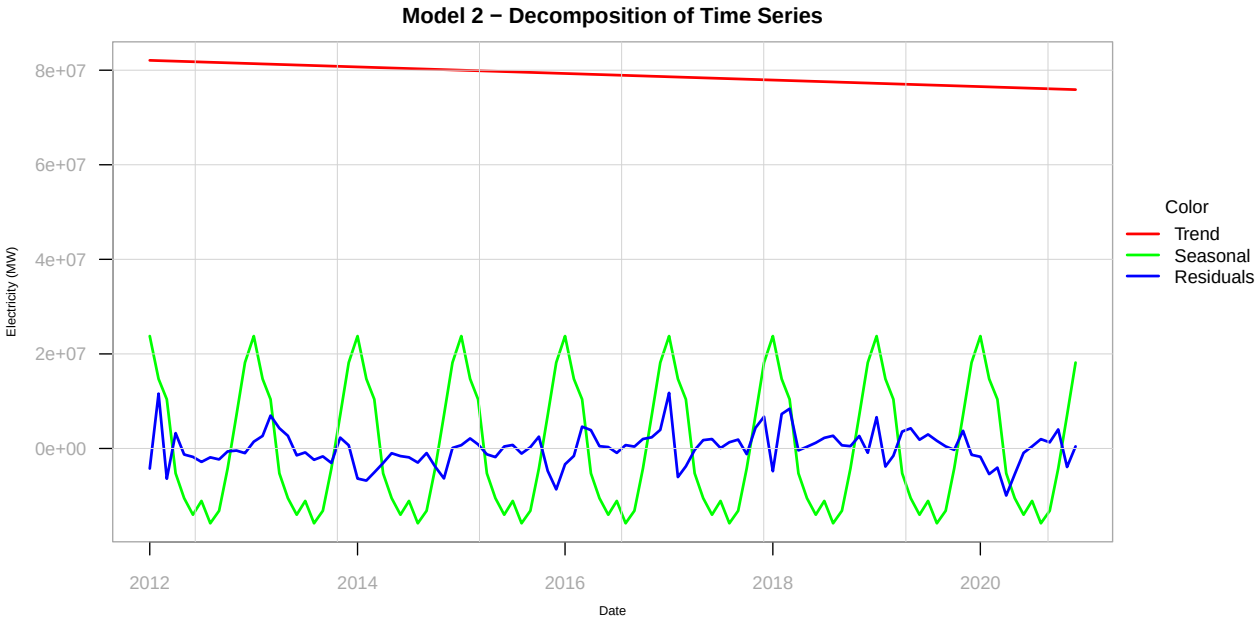
$$T_t = \beta_0 + \beta_1 \cdot Date_t$$
$$X_t^d = S_t + R_t = X_t - T_t$$

La moyenne de groupe, aussi connue sous le nom de moyenne conditionnelle ou moyenne par sous-groupes, consiste à calculer la moyenne des valeurs d'une série temporelle pour des périodes spécifiques (ici les mois) sur plusieurs cycles. Nos données étant mensuelle, le temps t peut être représenté par le mois et l'année. Soit $i = 1, \dots, 12$ les mois de l'année et $j = 1, \dots, N$ où N est le nombre d'années représentées, X_{ij} est la valeur de la série pour le mois i et l'année j. La moyenne de groupe pour le mois i est définie comme :

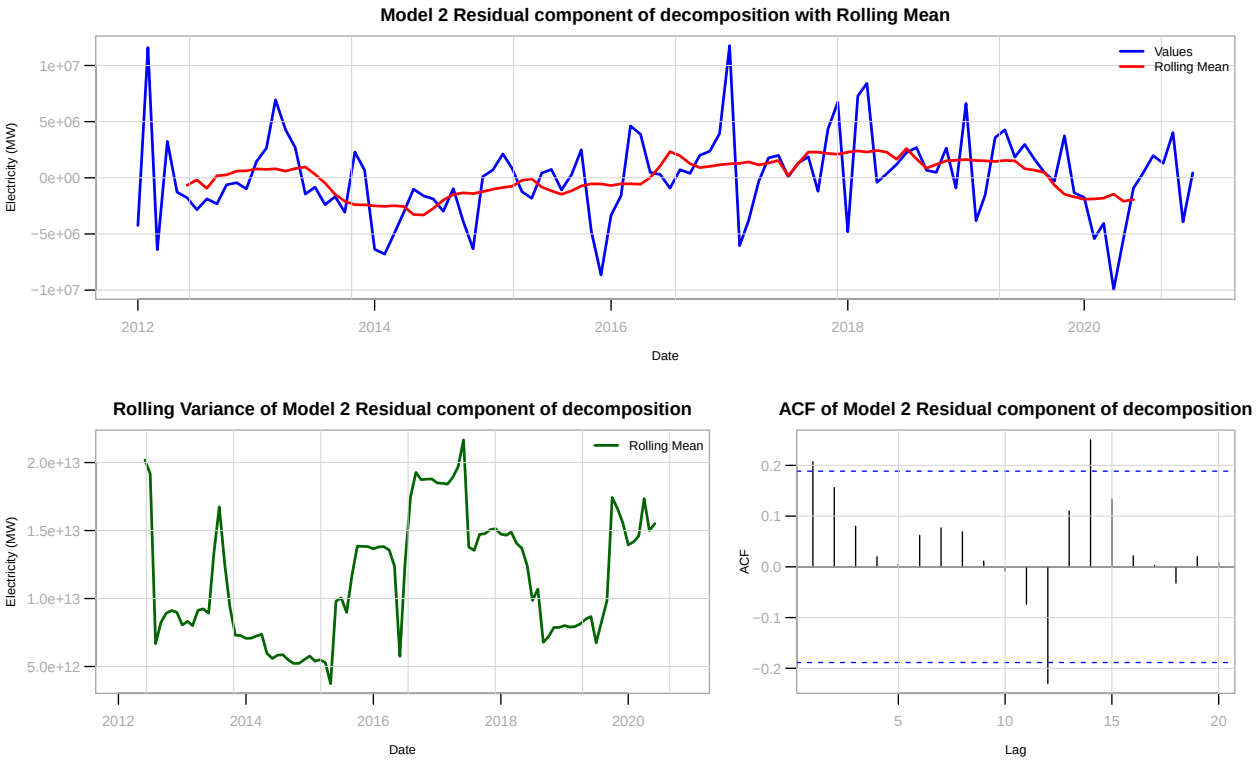
$$\bar{X}_j = \frac{1}{N} \sum_{i=1}^{N} X_{ij}$$

Ainsi le vecteur $(\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{12})$ décrit le motif de saisonnalité S_t et on obtient la composante résiduelle.

$$R_t = X_t^d - S_t$$



On analyse la stationnarité de la composante résiduelle.



ADF Test on Model 2 Residual component of decomposition p-value : 0.01063588
KPSS Test on Model 2 Residual component of decomposition p-value : 0.1

Les résultats du diagnostic sont cohérents avec la méthode de décomposition STL. La composante résiduelle obtenue présente des propriétés similaires en termes de stationnarité. Ceci renforce la robustesse de notre analyse et la validité de notre approche pour isoler la composante tendance et saisonnière de la série temporelle.

3.1.3 Modèle 3 - Décomposition classique

Une autre méthode de décomposition dite classique est elle, basée sur un modèle additif ou multiplicatif. La saisonnalité et la tendance sont extraites de manière fixe par la méthode des moyennes mobiles et la méthode de moyenne de groupe. Elle se fait par la fonction `decompose` de la librarie `stats`. Par défaut, le type de moyenne mobile utilisée est la moyenne mobile centrée :

$$T_t = CMA_t = \frac{1}{k} \sum_{i=t-\lfloor \frac{k}{2} \rfloor}^{t+\lfloor \frac{k}{2} \rfloor} X_i$$

où CMA_t est la moyenne mobile centrée à l'instant t, X_i est la valeur à l'instant i et k est la taille de la fenêtre (doit être impair pour que le calcul soit centré).

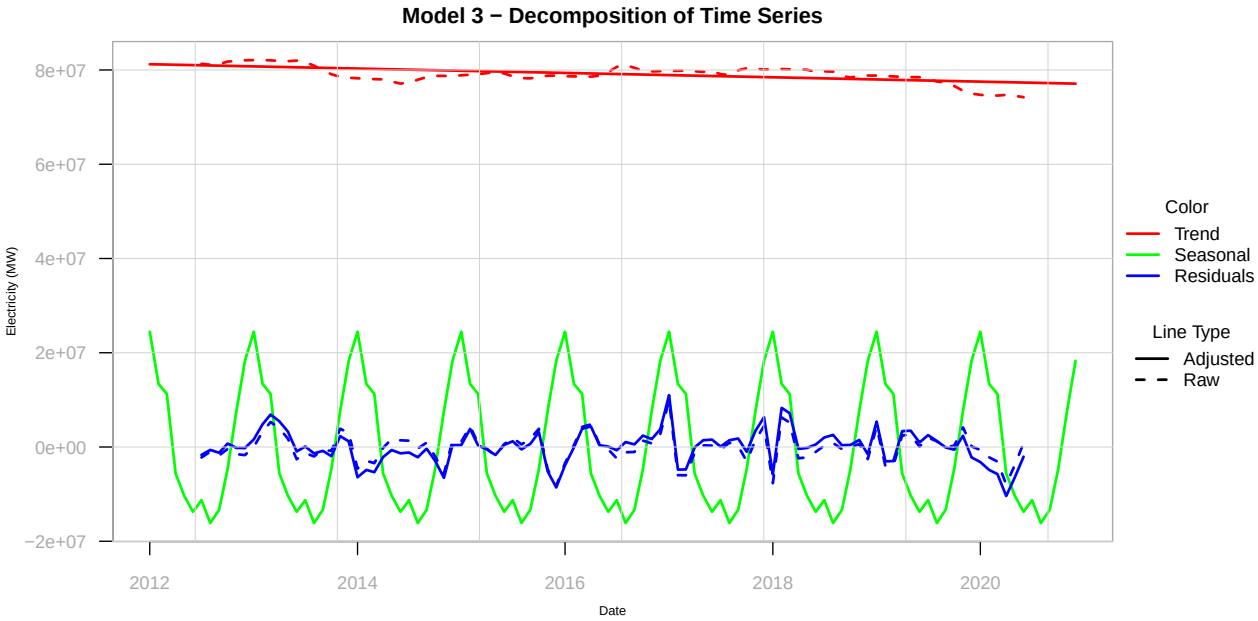
Après extraction de la tendance, la saisonnalité est obtenue par moyenne de groupe puis est centrée pour obtenir la décomposition. La technique des moyennes mobiles ne peut pas capturer des changements subtils dans la saisonnalité ou des tendances non linéaires, ce qui la rend moins flexible que STL mais reste valable dans notre cas d'étude. En revanche, elle est plus rapide et moins coûteuse en calcul. La méthode `decompose` ne fonctionne correctement que si la série couvre des périodes complètes, ce qui est notre cas.

Comme pour la méthode STL, on appliquera une régression linéaire à la composante de la tendance pour

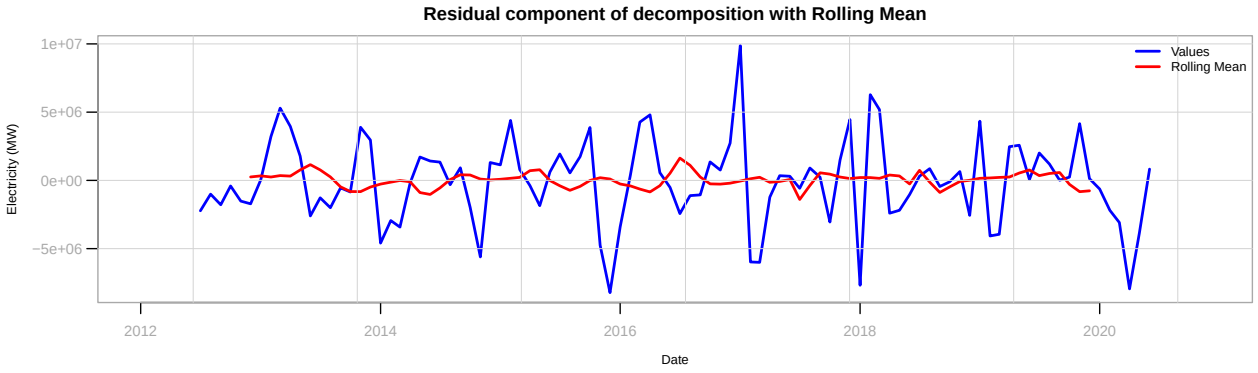
pouvoir faire des prévisions et on ajustera le modèle en conséquence :

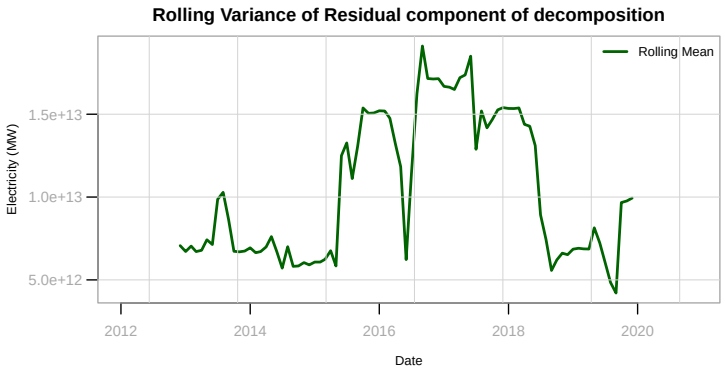
$$
\begin{aligned}
R'_t &= R_t + T_t - T'_t \\
\hat{R}_t &= R'_t - \bar{R}'_t \qquad \text{et} \qquad \hat{T}_t = T'_t + \bar{R}'_t \\
X_t &= \hat{T}_t + S_t + \hat{R}_t
\end{aligned}
$$

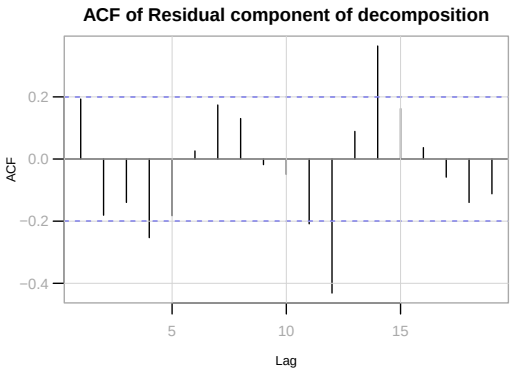
où T'_t est la droite de régression de la tendance, R'_t est la somme des résidus de la décomposition et de la régression, $\hat{R}_t$ est la composante résiduelle ajustée et $\hat{T}_t$ est la composante tendancière ajustée.



Il ne reste qu'à vérifier la stationnarité des résidus.







ADF Test on Residual component of decomposition p-value : 0.01
KPSS Test on Residual component of decomposition p-value : 0.1

On obtient les mêmes conclusions que pour les précédents modèles. Ainsi, les analyses visuelles et statistiques des résidus issues des différentes techniques de décomposition indiquent que les résidus sont stationnaires. Ces résultats valident l'efficacité de nos techniques de décomposition et justifient leur utilisation pour la modélisation de la série de la consommation d'électricité. Les résidus peuvent désormais être modélisés séparément à l'aide de modèles appropriés pour des séries stationnaires, tels que les modèles ARMA ou SARIMA, afin de capturer les dynamiques résiduelles.

3.2 Modélisation de la composante résiduelle

Pour modéliser les séries résiduelles stationnaires de nos modèles, on va analyser leur fonctions ACF et PACF afin de déterminer graphiquement les modèles candidats. Il s'agira alors de les appliquer à nos données et de comparer leurs performances grâce aux critères AIC et BIC. On pourra valider le modèle choisi par analyse de ses résidus. Afin de choisir le modèle adapté, l'introduction de cette section sera dédiée à quelques rappels théoriques sur les fonctions d'autocorrélation, les modèles ARMA et SARIMA et les techniques de validation de modèle qui seront employées.

Fonction ACVF
La fonction d'autocovariance (ACVF) mesure la covariance entre les valeurs d'une série temporelle séparées par un retard k, c'est-à-dire leur dépendance linéaire brute. Soit $(X_t)_t$ avec $t = 1, 2, \dots, T$, une série temporelle stationnaire avec espérance $\mu = \mathbb{E}[X_t]$, la fonction d'autocovariance d'ordre k, notée $\gamma_X(k)$, est définie par :

$$\gamma_X(k) = \mathrm{Cov}(X_t, X_{t+k}) = \mathbb{E}[(X_t - \mu)(X_{t+k} - \mu)]$$

Pour une série temporelle $\{X_1, X_2, \dots, X_T\}$ de moyenne empirique $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$, l'ACVF empirique d'ordre k est donnée par :

$$\hat{\gamma}_X(k) = \frac{1}{T}\sum_{t=1}^{T-k}(X_t - \bar{X})(X_{t+k} - \bar{X})$$

Fonction ACF
La fonction d'autocorrélation (ACF) mesure la corrélation entre les valeurs d'une série temporelle à différents retards (lag). C'est une version normalisée de l'autocovariance. Soit $(X_t)_t$ avec $t = 1, 2, \dots, T$, une série temporelle stationnaire de fonction d'autocovariance γ_X telle que $\gamma_X(0) \neq 0$, la fonction autocorrélation d'ordre k, notée $\rho_X(k)$ est définie par :

$$\rho_X(k) = \frac{\gamma_X(k)}{\gamma_X(0)}$$

Pour une série temporelle $\{X_1, X_2, \dots, X_T\}$ de fonction d'autocovariance empirique $\hat{\gamma}_X$, l'ACF empirique d'ordre k est donnée par :

$$\hat{\rho}_X(k) = \frac{\hat{\gamma}_X(k)}{\hat{\gamma}_X(0)}$$

Fonction PACF
La fonction d'autocorrélation partielle (PACF) est une extension de la fonction d'autocorrélation (ACF) qui mesure la relation entre une valeur d'une série temporelle et ses retards (lags) après avoir retiré l'effet des retards intermédiaires. Mathématiquement, elle correspond au coefficient de X_{t+k} dans une régression linéaire où X_t est expliqué par $X_{t+1}, \dots, X_{t+k}$. Soit $(X_t)_t$ avec $t = 1, 2, \dots, T$ une série temporelle stationnaire, la fonction d'autocorrélation partielle d'ordre k, notée $\phi_{k,k}$, est définie comme suit :

$$\phi_{k,k} = \mathrm{Cor}(X_t, X_{t-k} \mid X_{t-1}, X_{t-2}, \dots, X_{t-(k-1)})$$

Pour une série temporelle $\{X_1, X_2, \dots, X_T\}$ de fonction d'autocorrélation empirique $\hat{\rho}_X$, la PACF empirique d'ordre k est donnée par l'algorithme de Yule-Walker ou des régression successives :

$$\hat{\phi}_{k,k} = \frac{\hat{\rho}_X(k) - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_X(k-j)}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_X(j)}$$

où les $\hat{\phi}_{k-1,j}$ sont les PACF aux retards j pour une série de longueur $k-1$.

Modèle AR(p) (Auto Regressive)
Un modèle autorégressif (AR) est un modèle de série temporelle qui utilise des observations passées (ou retards) de la série pour prédire les valeurs futures. Soit X_t une série temporelle, un modèle autorégressif d'ordre p (noté AR(p)) est défini par l'équation suivante :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

où X_t est la valeur de la série à l'instant t, ϕ_1, ϕ_2, ..., ϕ_p sont les coefficients du modèle, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ sont les valeurs retardées de la série, p est l'ordre du modèle indiquant le nombre de valeurs retardées nécessaires pour prédire la valeur actuelle et ε_t est un terme d'erreur aléatoire à l'instant t, supposé être un bruit blanc, c'est-à-dire avec une moyenne nulle, une variance constante et non autocorrélé.

Dans un modèle AR(p), la dépendance est une combinaison linéaire des valeurs passées, entraînant une décroissance lente de l'ACF. Mais chaque valeur a une dépendance directe avec ses p termes précédents, donc la PACF sera proche de zéro après le lag p.

Modèle MA(q) (Moving Average)
Un modèle de moyenne mobile (MA) est un modèle de série temporelle qui utilise les erreurs passées (ou termes de bruit) pour modéliser les valeurs actuelles de la série. Soit X_t une série temporelle, un modèle de moyenne mobile d'ordre q (noté MA(q)) est défini par l'équation suivante :

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

où X_t est la valeur de la série à l'instant t, μ est la moyenne de la série temporelle (dans le cas où la série est centrée autour de zéro, cette moyenne est souvent omise), $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients du modèle, $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ sont les termes d'erreur retardés, q est l'ordre du modèle indiquant le nombre d'erreurs retardées nécessaires pour prédire la valeur actuelle et ε_t est le terme d'erreur aléatoire à l'instant t, supposé être un bruit blanc, c'est-à-dire avec une moyenne nulle, une variance constante et non autocorrélé.

Dans un modèle MA(q), l'ACF est proche de zéro après le lag q car les erreurs au-delà de q ne sont pas corrélées et la PACF décroit lentement en raison de la dépendance des erreurs passées.

Modèle ARMA(p,q) (Auto Regressive Moving Average)
Un modèle autorégressif de moyenne mobile (ARMA) est un modèle de série temporelle qui combine les aspects des modèles autorégressifs (AR) et des modèles de moyenne mobile (MA). Il est utilisé pour capturer à la fois les dépendances linéaires avec les valeurs passées de la série temporelle et les dépendances avec les erreurs passées. Soit X_t une série temporelle, un modèle ARMA d'ordre p, q (noté ARMA(p, q)) est défini par l'équation suivante :

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

où X_t est la valeur de la série à l'instant t, c est une constante, $\phi_1, \phi_2, \ldots, \phi_p$ sont les coefficients autorégressifs du modèle, $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients de la moyenne mobile du modèle, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ sont les valeurs passées de la série, $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ sont les termes d'erreur retardés et ε_t est le terme d'erreur aléatoire à l'instant t, supposé être un bruit blanc (moyenne nulle, variance constante, non autocorrélé).

Dans un modèle ARMA(p, q), la complexité des termes autorégressifs et de moyenne mobile crée des décroissances sans coupures nettes dans les fonctions ACF et PACF.

Modèle ARIMA(p, d, q) (AutoRegressive Integrated Moving Average)
Un modèle ARIMA est un modèle de série temporelle qui combine les composants auto-régressifs (AR), moyenne mobile (MA), et différenciation intégrée (I). Il est utilisé pour modéliser des séries temporelles qui peuvent présenter une tendance et une non-stationnarité (en différenciant les séries pour les rendre stationnaires). Soit X_t une série temporelle, un modèle ARIMA d'ordre (p, d, q) où p est l'ordre de l'autorégression, q est l'ordre de la moyenne mobile et d est le degré de différenciation nécessaire, est défini par :

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)\varepsilon_t$$

où X_t est la valeur de la série temporelle à l'instant t, $\phi_1, \phi_2, \ldots, \phi_p$ sont les coefficients autorégressifs (AR), ε_t est le terme d'erreur (ou bruit blanc) à l'instant t, $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients de la composante de moyenne mobile (MA) et B est l'opérateur de décalage.

Le modèle ARIMA est donc une combinaison d'un modèle AR pour les dépendances linéaires, d'un modèle MA pour les erreurs et d'une intégration pour rendre la série stationnaire. L'ACF et la PACF montrent des comportements typiques après différenciation et peuvent être utilisés pour identifier les ordres p et q.

Modèle SARIMA$(p,d,q)(P,D,Q)[s]$ (Seasonal ARIMA)
Le modèle SARIMA est une extension du modèle ARIMA pour les séries temporelles qui présentent une saison ou des motifs répétitifs à intervalles réguliers et la capture en ajoutant des composantes saisonnières au modèle ARIMA classique. Soit X_t une série temporelle, un modèle SARIMA d'ordre $(p,d,q)(P,D,Q)[s]$ où (p,s,q) sont les ordres pour la composante saisonnière, (P,D,Q) sont les ordres de la composante saisonnière et s est la périodicité saisonnière, est défini par :

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(1-B)^d X_t = (1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q)\varepsilon_t \times (1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps})(1-B^s)^D$$

où X_t est la série temporelle observée à l'instant t, B est l'opérateur de décalage, ε_t est le terme d'erreur, $\phi_1, \phi_2, \ldots, \phi_p$ sont les coefficients AR non saisonniers, $\theta_1, \theta_2, \ldots, \theta_q$ sont les coefficients MA non saisonniers, $\Phi_1, \Phi_2, \ldots, \Phi_P$ sont les coefficients AR saisonniers, $(1-B)^d$ et $(1-B^s)^D$ représentent respectivement les opérations de différenciation non saisonnière et saisonnière.

Dans un modèle SARIMA, l'ACF montre s'il existe une saisonnalité qui nécessite une différenciation saisonnière. En appliquant cette dernière, les ACF et PACF permettent d'identifier les ordres saisonniers P et Q : si la PACF montre un pic significatif à des multiples de s, cela est caractéristique d'un modèle AR(P) saisonnier, si l'ACF montre un pic significatif à des multiples de s, cela indique un modèle MA(Q) saisonnier.

Sélection du modèle en fonction de l'ACF et la PACF (après stationnarisation)

Modèle	ACF	PACF	Différenciations nécessaires
AR(p)	Décroît lentement	Zéro après p retards	Aucune
MA(q)	Zéro après q retards	Décroît lentement	Aucune
ARMA(p, q)	Zéro après q retards	Zéro après p retards	Aucune
ARIMA(p, d, q)	Zéro après q retards ou décroît lentement	Zéro après p retards ou décroît lentement	d ordinaires
SARIMA$(p, d, q)(P,D,Q)[s]$	Zéro après q retards ou décroît lentement Zéro après Q retards ou décroît lentement (échelle saisonnière)	Zéro après p retards ou décroît lentement Zéro après P retards ou décroît lentement (échelle saisonnière)	d ordinaires D saisonnières de période s

Une fois le type et les ordres du modèles identifiés, les paramètres ϕ et θ peuvent être estimés à l'aide de méthodes telles que la méthode des moindres carrés ordinaires (OLS) ou des algorithmes plus avancés comme l'estimation de maximum de vraisemblance (MLE). Pour se faire numériquement, la fonction `arima` en `R` applique une méthode MLE.

Afin de choisir le meilleur modèle, on pourra les comparer suivant deux critères :

AIC (Akaike Information Criterion, Akaike, 1973)
Le critère d'information d'Akaike (AIC) est défini par la formule suivante :

$$\mathrm{AIC} = -2\ln(L) + 2k$$

où L est la vraisemblance maximale du modèle et k est le nombre de paramètres estimés dans le modèle. Le terme $-2\ln(L)$ mesure la qualité de l'ajustement du modèle, tandis que le terme $2k$ pénalise les modèles avec un plus grand nombre de paramètres pour éviter le surajustement.

BIC (Bayesian Information Criterion, Shwartz, 1978)
Le critère d'information bayésien (BIC) est défini par la formule suivante :

$$\text{BIC} = -2\ln(L) + k\ln(n)$$

où L est la vraisemblance maximale du modèle, k est le nombre de paramètres estimés dans le modèle et n est le nombre d'observations dans les données. Le terme $k\ln(n)$ pénalise les modèles avec plus de paramètres de manière plus sévère que l'AIC, ce qui favorise les modèles plus simples, surtout lorsque la taille de l'échantillon n est grande.

On choisira le modèle avec les plus faibles valeurs d'AIC et/ou BIC :
- L'AIC peut parfois favoriser des modèles plus complexes parce qu'il pénalise moins fortement l'ajout de paramètres (seulement $2k$). Il suggère le modèle qui devrait avoir la meilleure capacité prédictive pour les données observées, mais aussi risque de surajuster les données d'entraînement.
- Le BIC peut aider à éviter le surajustement dans un problème où un modèle trop complexe s'adapte trop étroitement aux données d'entraînement et ne généralise pas bien aux nouvelles données.

La fonction `auto.arima` en `R` est capable à la fois de déterminer automatiquement les ordres p, d et q et d'estimer les paramètres du modèle. Elle utilise entre autres les critères d'information (par défaut l'AIC) pour sélectionner le meilleur modèle en termes de compromis entre ajustement et complexité.

Validation de modèle
Afin de s'assurer que le modèle est approprié pour les données et qu'il fournit des prévisions fiables, il faut analyser les résidus du modèles et vérifier qu'ils correspondent à un bruit blanc. En d'autres termes, il faut vérifier qu'ils soient non autocorrélés, normalement distribués avec une moyenne nulle et une variance constante. Cela se vérifie graphiquement notamment à l'aide de l'ACF pour l'autocorrélation et d'un histogramme pour la distribution, et grâce au test de Ljung-Box.

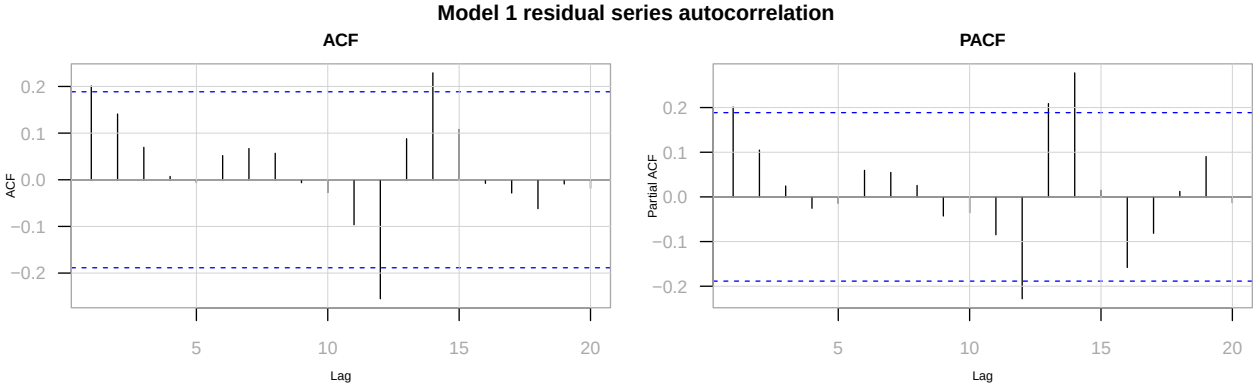
Test de Ljung-Box
Le test de Ljung-Box est un test statistique utilisé pour évaluer si les résidus d'un modèle sont indépendants et aléatoires, donc l'absence d'autocorrélation. Il est basé sur la statistique suivante :

$$Q = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{n-k}$$

où n est la taille de l'échantillon (nombre d'observations), r_k est l'autocorrélation des résidus à l'ordre k et h est le nombre de lags que l'on souhaite tester. La somme est effectuée pour chaque lag k de 1 à h.
- Hypothèse nulle H_0 : Les résidus ne présentent pas d'autocorrélation significative à l'ordre k, c'est-à-dire qu'ils sont indépendants.
- Hypothèse alternative H_1 : Les résidus présentent une autocorrélation significative, indiquant que le modèle n'a pas capturé toute la structure d'autocorrélation.

3.2.1 Modèle 1

Analysons les fonctions ACF et PACF de la composante résiduelle du modèle 1.



Les graphes montrent une autocorrélation faible au lag 1, c'est le seul lag significatif à court terme. Les autocorrelations significatives aux lags 12 et 14 peuvent suggérer une saisonnalité résiduelle ou une structure cyclique non capturée par la décomposition, et peuvent être liées à des dépendances temporelles complexes ou la présence de valeurs atypiques. Les autres lags ne montrent pas de significativité, ce qui confirme la stationnarité de la composante résiduelle.

Les autocorrélations à peine significatives au lag 1 peuvent indiquer une composante autorégressive faible d'ordre 1 (AR(1)) ou une composante de moyenne mobile d'ordre 1 (MA(1)), ou les deux (ARMA(1, 1)). Les autocorrélations significatives aux lags 12 et 14 suggèrent une composante saisonnière potentielle d'ordre saisonnier ($S = 12$), pouvant être ajoutée.

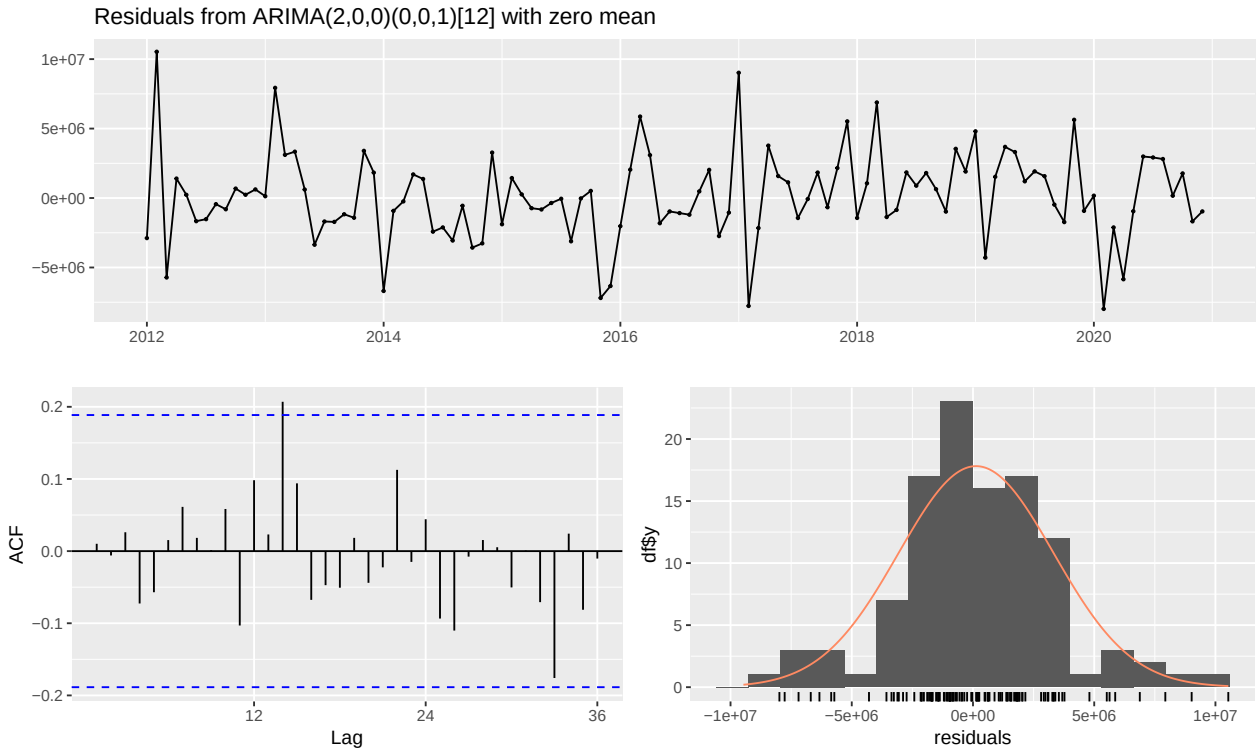
On ajuste chacun de ces modèles pour comparer leur critères d'information.

Table 5 – Comparaison de modèles ARIMA pour le Modèle 1

Model	AIC	BIC
AR(1)	3573.078	3578.443
MA(1)	3573.868	3579.232
ARMA(1, 1)	3578.413	3586.459
SARIMA(1, 0, 0)(0, 0, 1)[12]	3559.684	3567.730
SARIMA(0, 0, 1)(0, 0, 1)[12]	3561.518	3569.564
SARIMA(1, 0, 1)(0, 0, 1)[12]	3558.051	3568.779
SARIMA(2, 0, 0)(0, 0, 1)[12]	3555.676	3566.405

Parmi eux, le modèle avec la meilleure performance BIC est le modèle AR(1) avec composante saisonnière de période $S = 12$ et celui avec la meilleure performance AIC est le modèle ARMA(1,1) avec composante saisonnière. Cela nous donne une première indication du meilleur modèle et en particulier qu'un modèle avec saisonnalité est peut être plus adapté. C'est pourquoi on lance également une recherche exhaustive avec `auto.arima` en configurant une exploration dans les modèles saisonniers. La fonction sélectionne un modèle SARIMA(2,0,0)(0,0,1)[12] (ligne 7) qui a des valeurs AIC et BIC significativement inférieures à celles du modèle choisi à priori. Ce modèle capture donc mieux des subtilités dans les données, il sera donc retenu.

Pour valider ce modèle, on procède à un diagnostic de ses résidus.

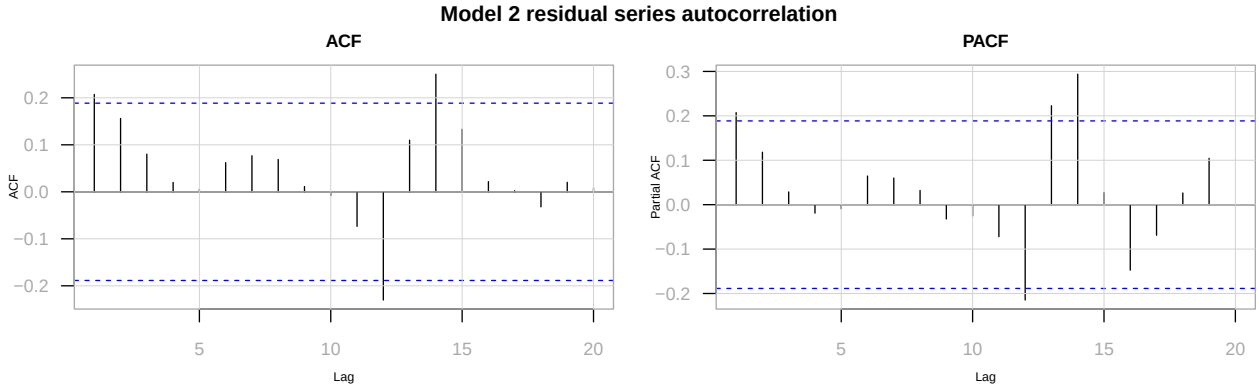


Ljung-Box test on Residuals from SARIMA(2, 0, 0)(0, 0, 1)[12] p-value : 0.7573

Les résidus oscillent autour de zéro sans tendance apparente ni saisonnalité, et leur distribution est normale. Le test de Ljung-Box (p-value = 0.7573 > 0.05) indique que les résidus sont indépendants. Bien que le lag 14 soit à peine significatif dans l'ACF, on peut considérer qu'il soit négligeable car le modèle est globalement performant et possède les meilleurs critères (AIC/BIC et erreur de prédiction). Dans une analyse plus approfondie, on pourrait vérifier si cette corrélation est dûe à un bruit aléatoire (en modifiant la fenêtre d'entrainement par exemple), à une variable exogène et si la capturer augmenterait considérablement la qualité des prévisions.

3.2.2 Modèle 2

Pour la composante résiduelle du modèle 2, les fonctions ACF et PACF sont très similaires au modèle précédent.



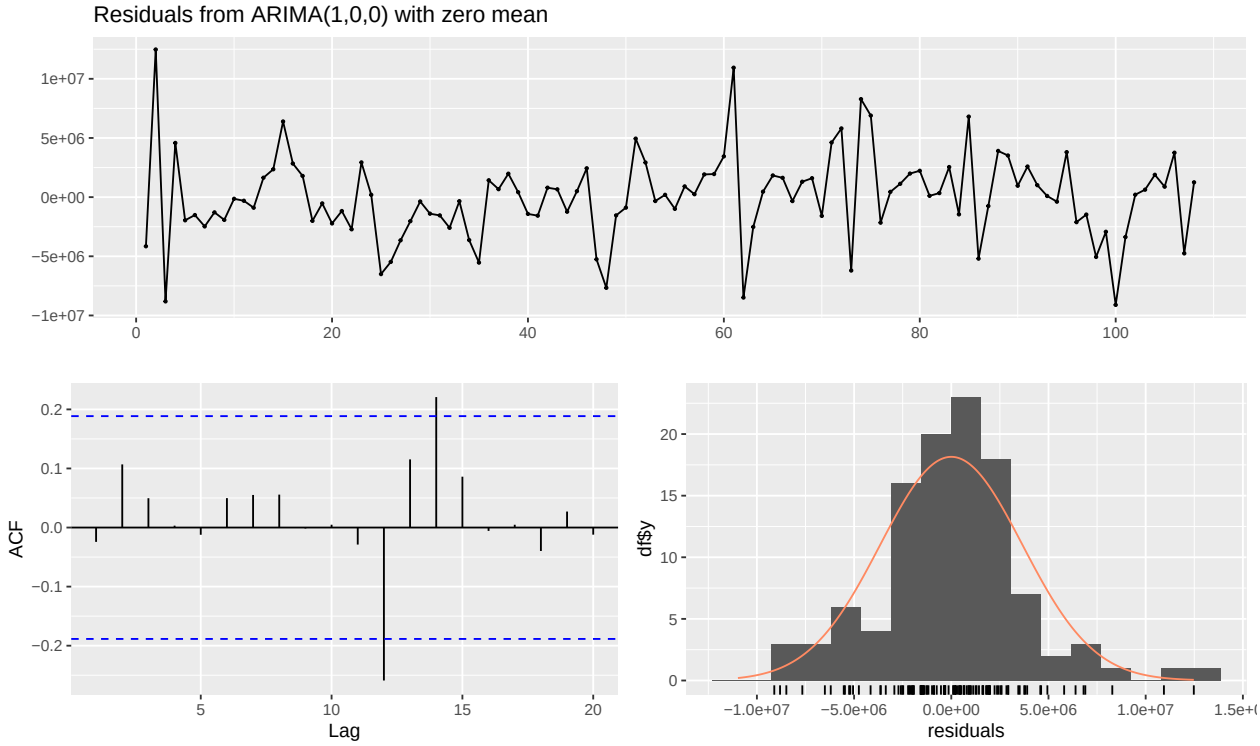
On peut par conséquent sélectionner les mêmes modèles candidats :

Table 6 – Comparaison de modèles ARIMA pour le Modèle 2

Model	AIC	BIC
AR(1)	3573.890	3579.254
MA(1)	3574.813	3580.177
ARMA(1, 1)	3574.521	3582.567
SARIMA(1, 0, 0)(0, 0, 1)[12]	3563.848	3571.895
SARIMA(0, 0, 1)(0, 0, 1)[12]	3566.371	3574.417
SARIMA(1, 0, 1)(0, 0, 1)[12]	3560.213	3570.942
SARIMA(2, 0, 0)(0, 0, 1)[12]	3558.359	3569.087

Les résultats des critères de performances indiquent le modèle SARIMA(1,0,1)(0,0,1)[12] comme le meilleur en termes d'AIC et BIC. De même que pour le modèle 1, la recherche exhaustive fourni par `auto.arima` (ligne 7) sélectionne le modèle SARIMA(2,0,0)(0,0,1)[12]. Cela est cohérent avec les techniques de décomposition qui ont été utilisées (régressions et moyennes de groupe pour les deux modèles), rendant les décompositions des modèles 1 et 2 similaires jusqu'à présent. Puisque le modèle 2, se veut plus simple et rapide, nous pouvons envisager de modéliser la composante résiduelle avec un modèle simple sans composante saisonnière. L'idée est de considérer ici que la saisonnalité des résidus n'est pas significative et que malgré les performances de prévisions d'un modèle saisonnier, opter pour un modèle plus général pourrait éviter un problème de surajustement. On utilisera alors le meilleur modèle simple, le modèle AR(1).

On procède au diagnostic des résidus.

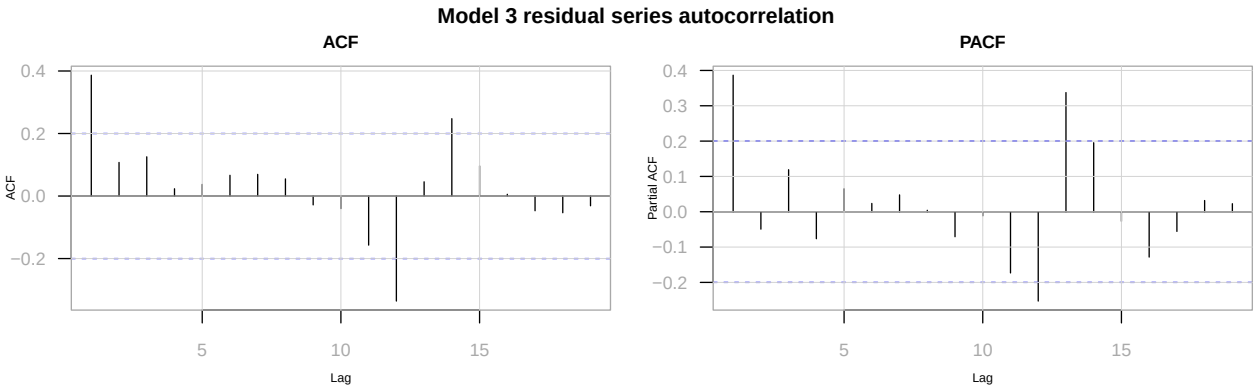


Ljung-Box test on Residuals from AR(1) p-value : 0.9762

Le lag significatif (lag 12) anticipé suggère une saison non complètement capturée mais son amplitude de corrélation reste raisonnablement faible. Les autres tests montrent l'indépendance et la normalité des résidus, renforçant la validité du modèle. Puisque l'objectif est la prévision, on peut envisager de valider ce modèle si le lag significatif n'affecte pas la capacité prédictive du modèle, ce que l'on analysera sur l'échantillon de test.

3.2.3 Modèle 3

Les fonctions ACF et PACF pour le modèle 3 montrent également un signe de saisonnalité mais une plus forte corrélation au lag 1 que les prédécents :

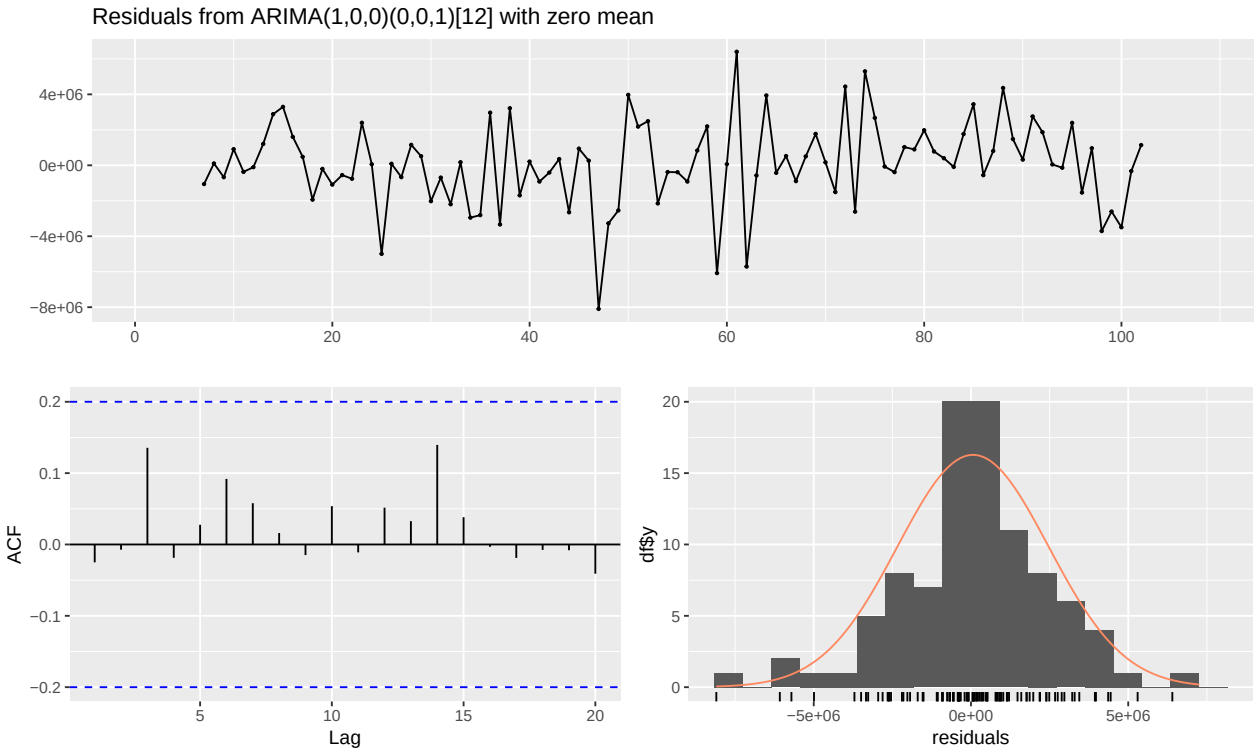


Comparons les modèles candidats ainsi que le modèle choisi par `auto.arima` (ligne 7).

TABLE 7 – Comparaison de modèles ARIMA pour le Modèle 3

Model	AIC	BIC
AR(1)	3156.567	3161.696
MA(1)	3155.950	3161.078
ARMA(1, 1)	3157.922	3165.615
SARIMA(1, 0, 0)(0, 0, 1)[12]	3124.642	3132.335
SARIMA(0, 0, 1)(0, 0, 1)[12]	3128.526	3136.219
SARIMA(1, 0, 1)(0, 0, 1)[12]	3125.807	3136.064
SARIMA(1, 0, 0)(2, 0, 0)[12]	3135.544	3145.802

D'après ces résultats, on gardera le modèle SARIMA(1,0,0)(0,0,1)[12] et on vérifie les résidus :



Ljung-Box test on Residuals from SARIMA(1, 0, 0)(0, 0, 1)[12] p-value : 0.8878

La validation du modèle montre que tous les tests sont satisfaisants : les résidus suivent une distribution normale, ils sont indépendants (aucun lag significatif dans l'ACF des résidus), et le test de Ljung-Box confirme l'absence d'autocorrélation significative. Par conséquent, le modèle peut être considéré comme valide.

3.3 Prévisions et erreurs

Afin de faire des prévisions, il suffit d'additionner les prévisions des modèles pour chaque composantes. La composante saisonnière étant stable et régulière, on prolonge le pattern saisonnier observé dans les données pour prévoir les futures fluctuations saisonnières. On compare ensuite les résultats des prévisions avec les réelles données de l'échantillon de test (non utilisées pour ajuster le modèle). Cela permettra d'évaluer les capacités prédictives de nos modèles en comparant les métriques d'erreur telles que le RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), etc.

Modèle 1

$$X_t^1 = T_t^1 + S_t^1 + R_t^1$$

où T_t^1 est une droite de régression linéaire, S_t^1 est un motif saisonnier stable de période 12 et R_t^1 est un modèle SARIMA(2,0,0)(0,0,1)[12].

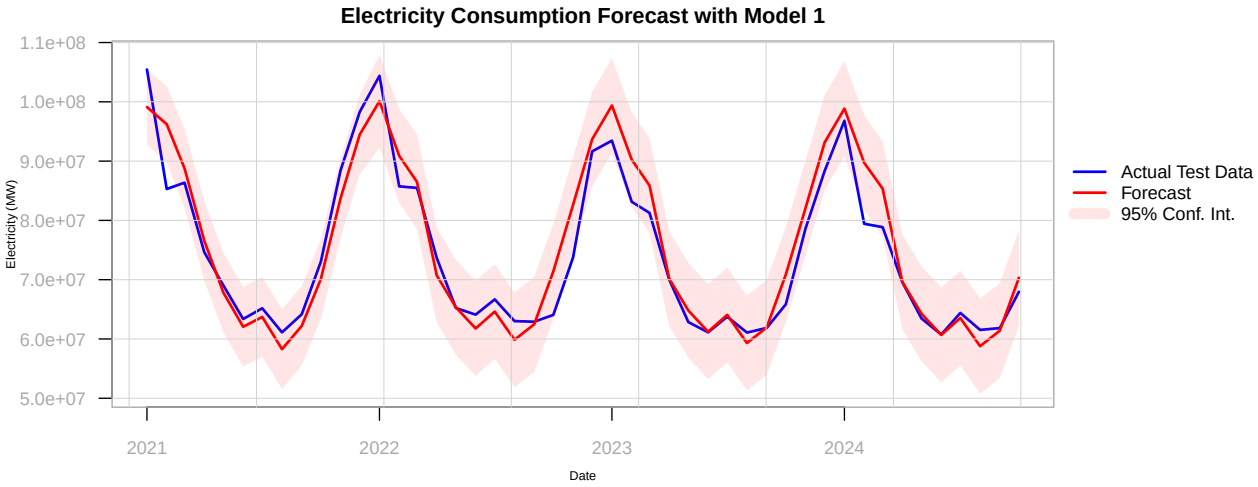


TABLE 8 – Erreur de prévision sur l'échantillon de test pour le Modèle 1

	ME	RMSE	MAE	MPE	MAPE
Test set	1036595	4127547	3107528	1.013748	3.86606

Modèle 2

$$X_t^2 = T_t^2 + S_t^2 + R_t^2$$

où T_t^2 est une droite de régression linéaire, S_t^2 est un motif saisonnier stable de période 12 et R_t^2 est un modèle AR(1).

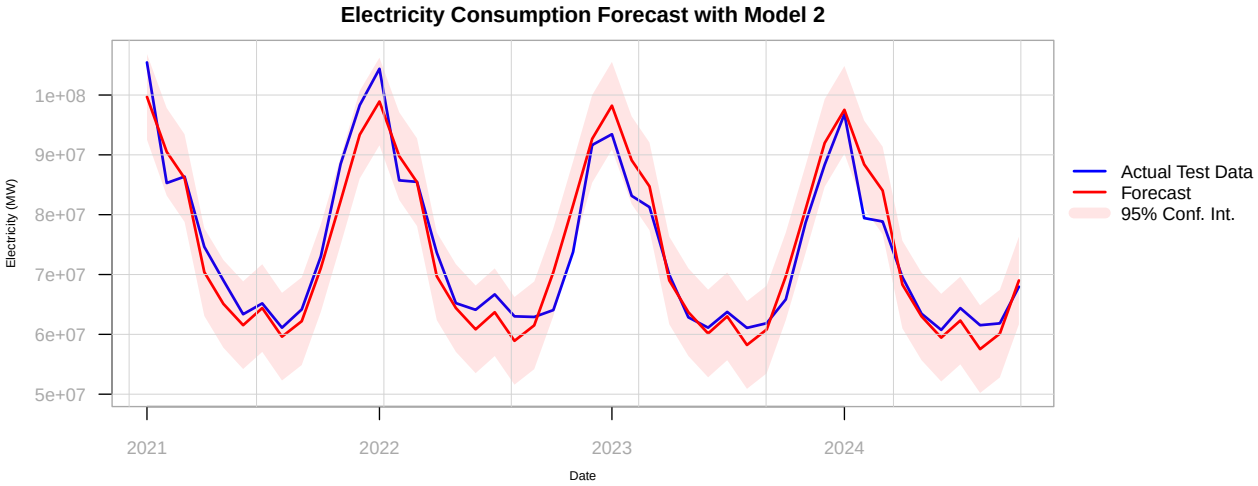


TABLE 9 – Erreur de prévision sur l'échantillon de test pour le Modèle 2

	ME	RMSE	MAE	MPE	MAPE
Test set	-164163.7	3684022	2995918	-0.5698338	3.95015

Modèle 3

$$X_t^3 = T_t^3 + S_t^3 + R_t^3$$

où T_t^3 est une droite de régression linéaire, S_t^3 est un motif saisonnier stable de période 12 et R_t^3 est un modèle SARIMA(1,0,0)(0,0,1)[12].

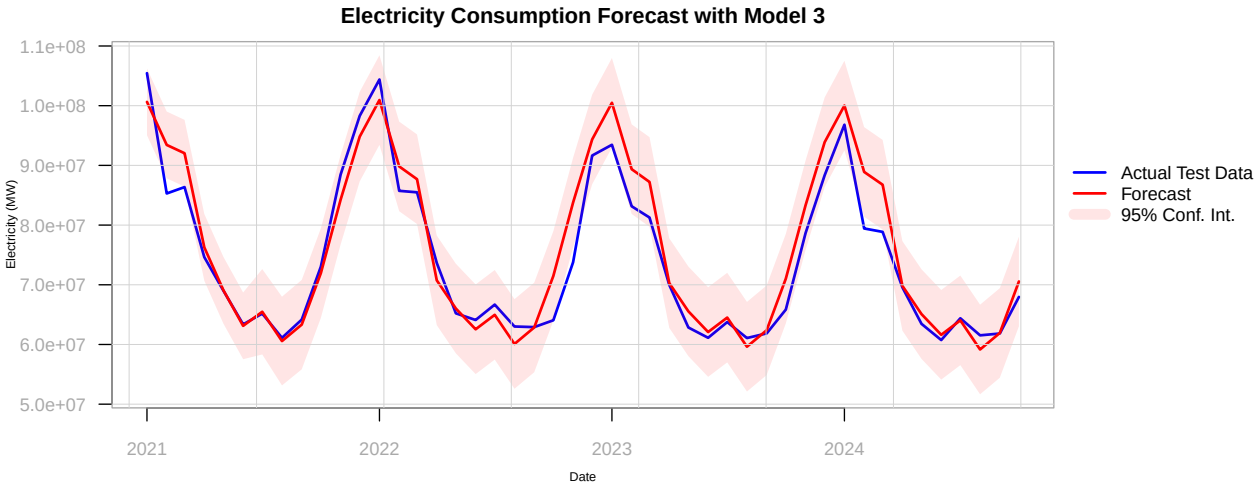


TABLE 10 – Erreur de prévision sur l'échantillon de test pour le Modèle 3

	ME	RMSE	MAE	MPE	MAPE
Test set	1671967	4085274	3062581	1.905789	3.738464

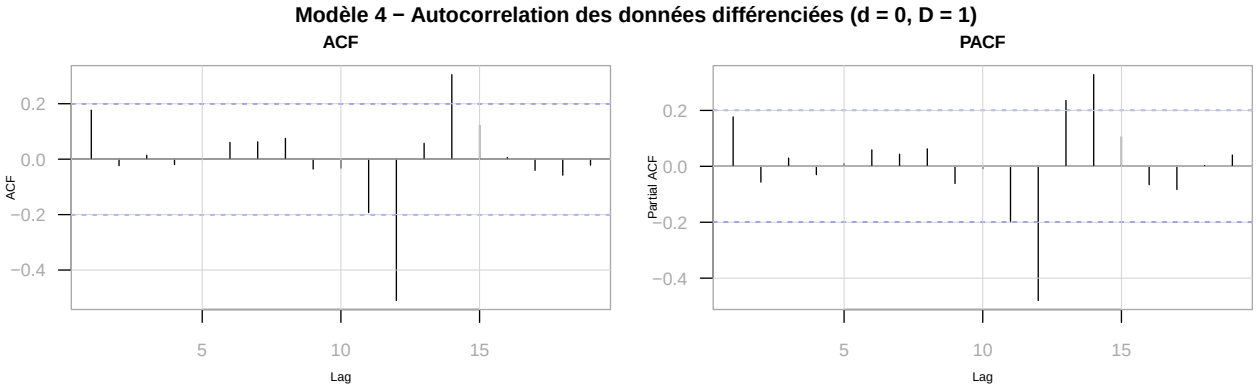
Chacun des modèles semble émettre des prévisions très satisfaisantes sur les données de tests. En effet, la courbe prévisionnelle suit la même allure que les données réelles et l'intervalle de confiance 95% englobe la quasi-totalité de cette dernière. Sur la base des métriques d'erreur (ME, RMSE, MAE, MPE, MAPE), le modèle 2 est clairement le plus performant en termes de prévision. Il minimise les erreurs absolues, quadratiques et relatives par rapport aux autres modèles. C'est par conséquent le modèle à privilégier, d'autant plus qu'il est plus simple et plus rapide que les autres puisqu'il n'applique pas de lissage LOESS ou de décomposition préliminaire pour la tendance et que sa composante résiduelle est modélisée par un simple modèle AR(1) avec peu de paramètres. Bien que son ajustement aux données de l'échantillon d'entraînement n'était pas à la hauteur des autres modèles, il parvient à mieux prédire les valeurs futures et constitue l'exemple qu'un modèle simple permet d'éviter un surajustement dans l'objectif de faire des prévisions. De plus, les tests complémentaires (analyse des résidus, etc.) confirment la validité de ce modèle pour les données analysées.

4 Modélisation SARIMA

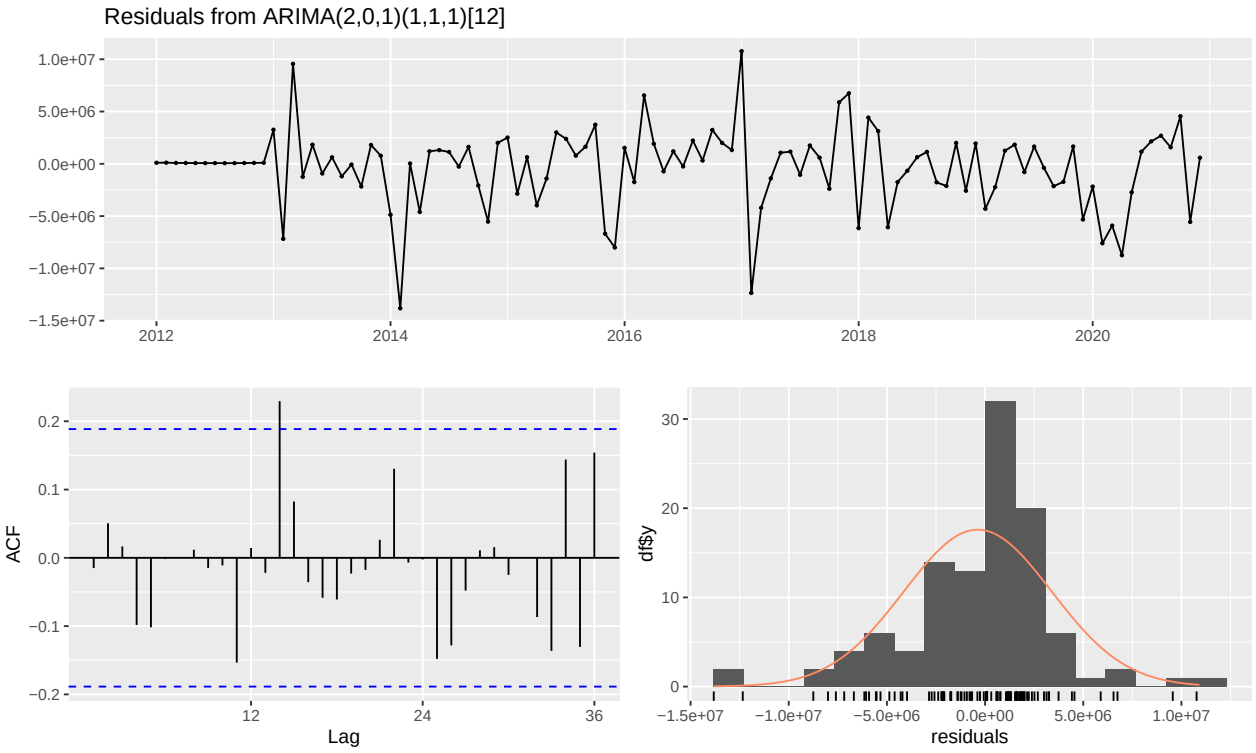
Comme expliqué précédemment, la série temporelle n'est pas stationnaire. Nous allons essayer un modèle SARIMA avec une saisonnalité de 12.

4.1 Modèle 4 - Identification du modèle via l'ACF et la PACF

Les graphiques de l'Autocorrélation (ACF) et de l'Autocorrélation Partielle (PACF) des données différenciées aident à identifier les ordres potentiels des composants AR (autorégressif) et MA (moyenne mobile) du modèle ARIMA.



Afin de déterminer les paramètres p, q, P et Q du modèle SARIMA, une différentiation saisonnière est effectuée. Ainsi, grâce à ces deux graphes, nous pouvons estimer ces paramètres, bien que différents choix soient possibles. Étant donné l'ACF et le PACF, nous pouvons supposer que le modèle SARIMA(2,0,1)(1,1,1)[12] est un modèle raisonnable. Voici une simulation de ce modèle :



Ljung-Box test on Residuals from SARIMA(2, 0, 1)(1, 1, 1)[12] p-value : 0.4642

On observe que le modèle SARIMA(2,0,1)(1,1,1)[12] semble acceptable. Les graphiques suggèrent que les résidus ressemblent à un bruit blanc : leur moyenne semble être proche de 0, et le test de Ljung-Box indique qu'ils ne présentent aucune autocorrélation. On pourrait même ajouter que l'histogramme laisse penser que les résidus suivent une loi normale.

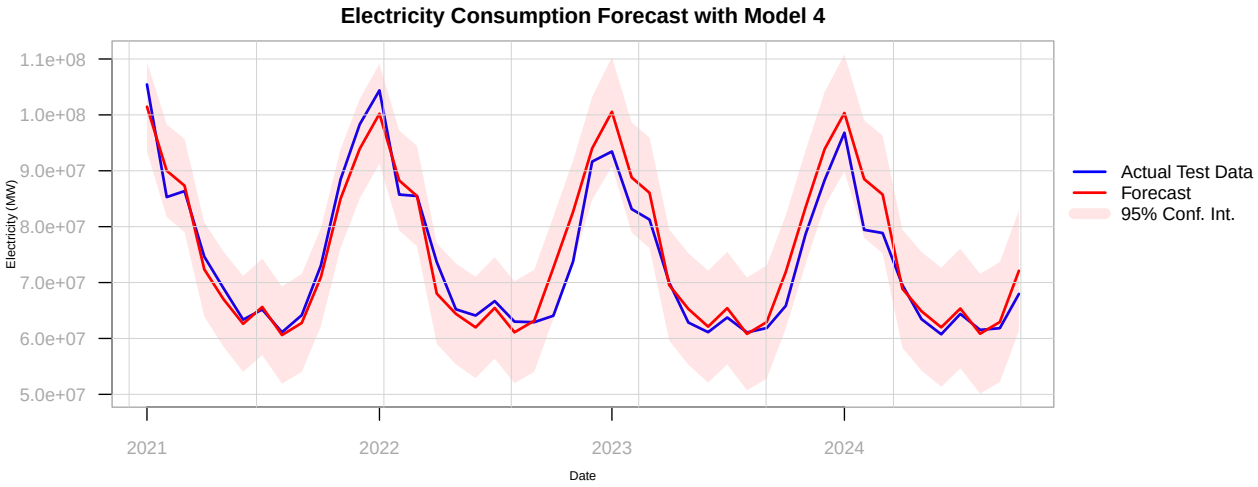


TABLE 11 – Erreur de prévision sur l'échantillon de test pour le Modèle 4

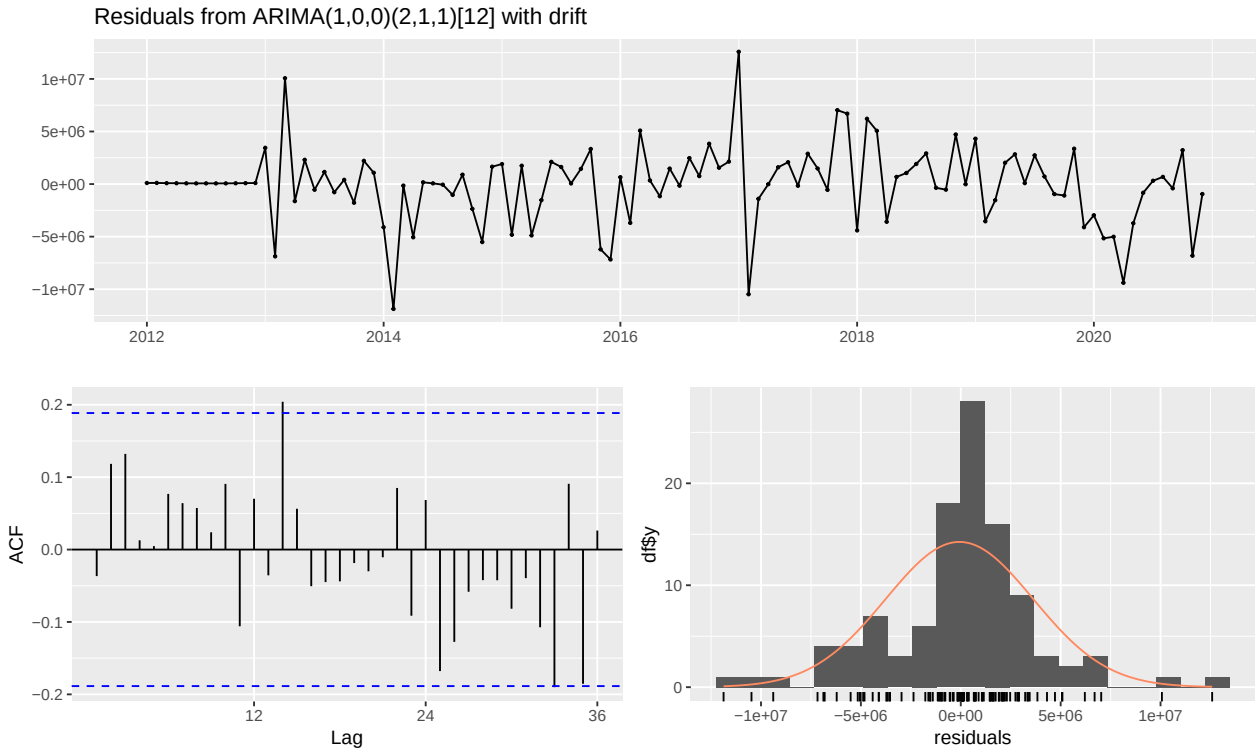
	ME	RMSE	MAE	MPE	MAPE
Test set	1271538	3855871	2956860	1.47323	3.702234

4.2 Modèle 5 - Ajustement automatique

La fonction `auto.arima` du package `forecast` sélectionne automatiquement le meilleur modèle SARIMA en fonction des critères d'information (comme AIC, BIC). Les paramètres spécifiés permettent de contrôler la complexité du modèle, notamment les ordres maximaux des termes autorégressifs (p, P) et de moyenne mobile (q, Q), ainsi que la complexité totale (`max.order`).

Le résumé fournit des détails sur les coefficients estimés, les erreurs standard, le `sigma^2` (variance des résidus), le log-vraisemblance, et les critères d'information. Ces informations sont cruciales pour évaluer la qualité du modèle ajusté.

Encore une fois, vérifions que les résidus sont aléatoires et ne présentent pas de structure systématique à travers le test de Ljung-Box pour l'autocorrélation des résidus.



Ljung-Box test on Residuals from SARIMA(1, 0, 0)(2, 1, 1)[12] p-value : 0.579

Une fois le modèle ajusté et diagnostiqué, l'étape suivante consiste à générer des prévisions. L'horizon de prévision est déterminé par la taille de l'ensemble de test, assurant que les prévisions couvrent la même période que les données réelles pour une comparaison équitable. Le niveau de confiance 95% fournit un intervalle autour des prévisions, offrant une idée de l'incertitude associée.

La visualisation permet de comparer visuellement les prévisions du modèle SARIMA avec les valeurs réelles de consommation d'électricité. En définissant correctement le début de la période de prévision, on s'assure que les séries temporelles sont alignées dans le temps. Le tracé affiche les prévisions en rouge et les valeurs réelles en bleu, avec l'intervalles de confiance, fournissant une évaluation intuitive de la performance du modèle.

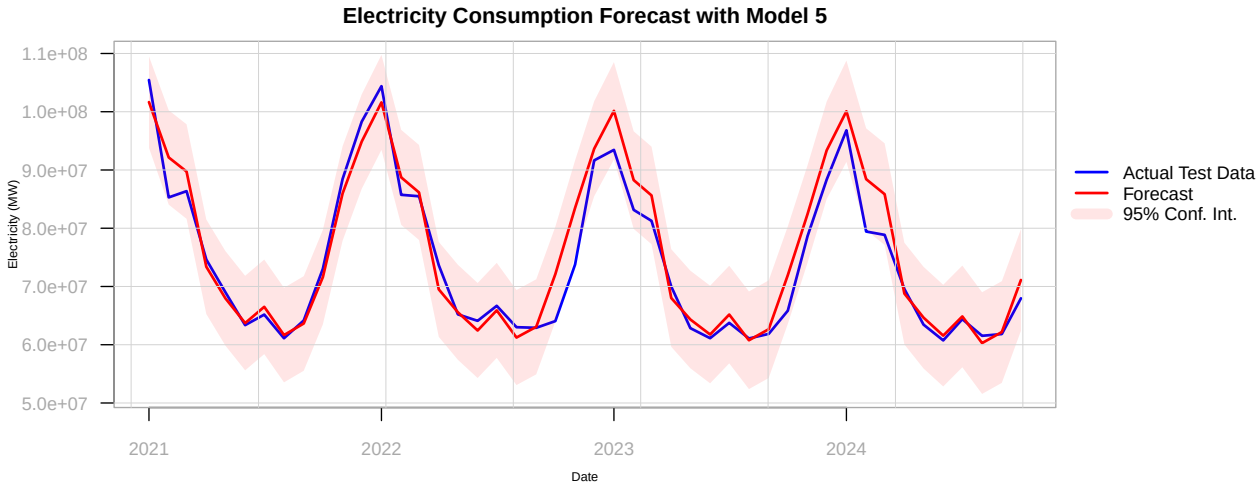


Table 12 – Erreur de prévision sur l'échantillon de test pour le Modèle 5

	ME	RMSE	MAE	MPE	MAPE
Test set	1470909	3710745	2750043	1.727791	3.396159

Grâce à ce graphique, et surtout à cette table des erreurs, nous comparerons ce modèle aux autres.

5 Conclusion

Table 13 – Comparaison d'erreurs de prévision sur l'échantillon de test

	ME	RMSE	MAE	MPE	MAPE
Modèle 1	1036595.0	4127547	3107528	1.0137482	3.866060
Modèle 2	-164163.7	3684022	2995918	-0.5698338	3.950150
Modèle 3	1671967.0	4085274	3062581	1.9057888	3.738464
Modèle 4	1271538.1	3855871	2956860	1.4732304	3.702234
Modèle 5	1470909.4	3710745	2750043	1.7277914	3.396159

Notre analyse de la consommation d'électricité en France nous a permis de mettre en évidence des dynamiques importantes liées aux tendances, aux variations saisonnières et aux résidus. Les résultats obtenus montrent une forte composante saisonnière, caractéristique des variations des besoins énergétiques, principalement influencées par des facteurs climatiques. Nous avons pu utiliser deux approches, décomposition additive et SARIMA, afin de les comparer et mesurer leur pertinence pour la prévision de la consommation.

Parmi les mesures d'erreur, le MAE et RMSE sont considérés comme les plus fiables, l'un étant le plus stable et robuste et l'autre étant le plus sensible aux valeurs extrêmes. Ces critères mettent en évidence le modèle basé sur une décomposition simple avec un ajustement AR(1) et le modèle saisonnier SARIMA(1, 0, 0)(2, 1, 1)[12]. Ainsi le modèle 2 s'est révélé le plus efficace selon le RMSE, combinant simplicité et précision prédictive. Sa simplicité a permis d'éviter de tomber dans le surajustement et ainsi de mieux généraliser la série que les 4 autres modèles. La saisonnalité fortement marquée de la série de données a également contribué à la performance de la décomposition.

Ces prévisions offrent des perspectives utiles pour la planification énergétique et l'optimisation des ressources. En effet, elles permettent d'anticiper les variations de consommation et de mieux répondre aux besoins futurs.

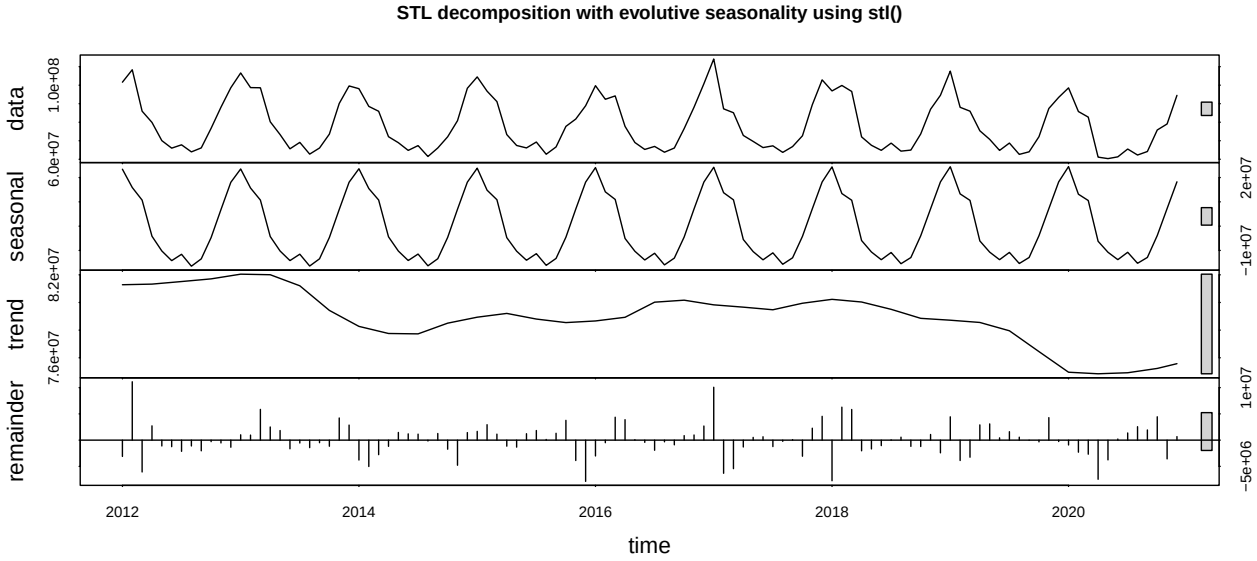
Une limite importante de notre analyse réside dans le fait que les modèles ont été entraînés sur des données collectées avant la crise sanitaire. Cet événement a potentiellement provoqué des transformations majeures dans les habitudes de consommation d'électricité. Les modèles ne capturent donc pas ces nouvelles dynamiques, ce qui affecte la précision des prévisions.

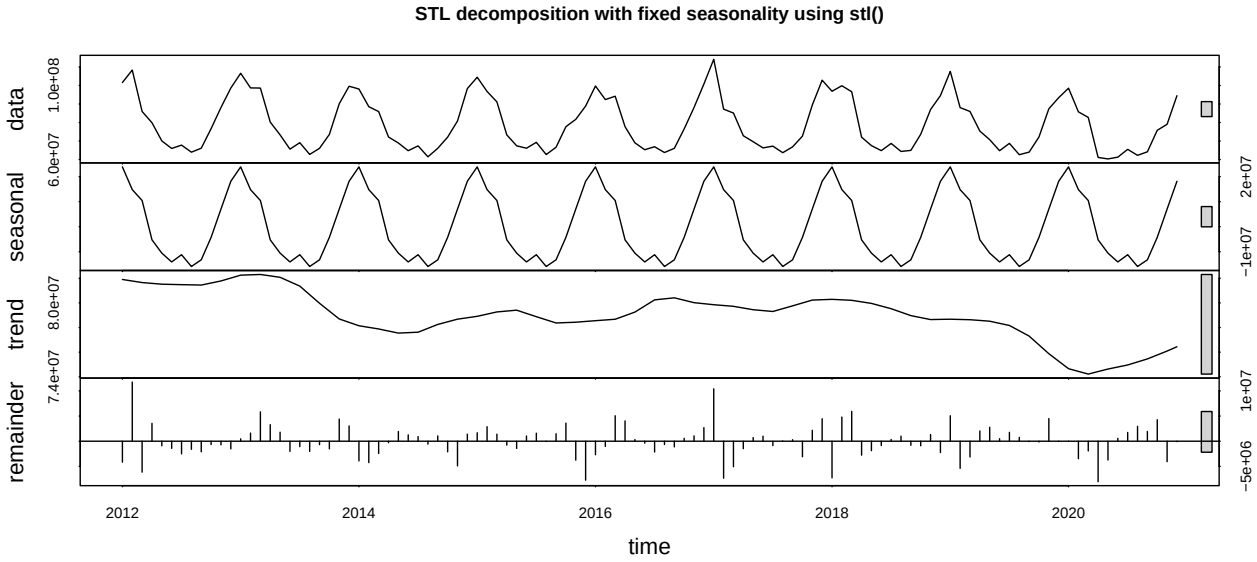
Pour aller plus loin, il serait pertinent d'intégrer des variables exogènes telles que des données météorologiques ou économiques afin d'améliorer la précision des modèles. La présence de telles variables pourrait préciser l'évolution de la saisonnalité ainsi qu'expliquer la tendance.

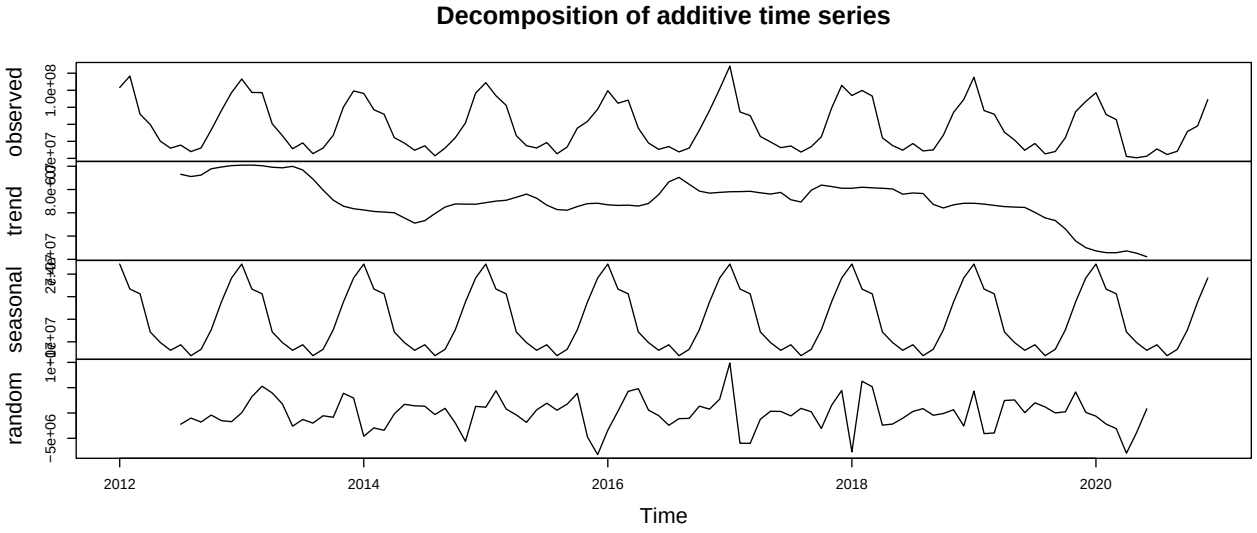
6 Bibliographie

— Séries temporelles, M2 IFMA (Sorbonne Université, 2024)
— Analyse des séries financières, J-M. Bardet (M2MO, Université Paris Cité, 2019)
— Introduction to Time Series and Forecasting, Borckwell-Davis (2nd edition)

7 Annexes


STL decomposition with evolutive seasonality using stl()
data
seasonal
trend
remainder
2012
2014
2016
2018
2020
time


STL decomposition with fixed seasonality using stl()
data
seasonal
trend
remainder
2012
2014
2016
2018
2020
time


Decomposition of additive time series
observed
trend
seasonal
random
2012
2014
2016
2018
2020
Time