



DataScientest

## Rapport final

### Prévision de la consommation d'électricité en France

**Youssef SERRESTOU**

Tuteur : Kalome BOTOWAMUNGU



# Table de matière

<b>Introduction générale</b>	<b>1</b>
<b>1 Introduction aux séries temporelles</b>	<b>4</b>
1.1 Généralités . . . . .	4
1.1.1 Définitions . . . . .	4
1.1.2 Exemple . . . . .	5
1.2 Décomposition d'une série . . . . .	6
1.2.1 Composantes . . . . .	6
1.2.2 Décomposition additive et décomposition multiplicative . . . . .	6
1.3 Analyse des séries temporelles . . . . .	10
1.3.1 Stationnarité . . . . .	10
1.3.1.1 Définition . . . . .	10
1.3.1.2 Formes de non stationnarité . . . . .	11
1.3.1.3 Tests de stationnarité . . . . .	11
1.3.2 Fonctions d'autocorrélation et d'autocorrélation partielle . . . . .	14
1.3.3 Analyse temps-fréquence . . . . .	17
1.4 Prévision et métriques d'évaluation . . . . .	19
1.4.1 Prévision . . . . .	19
1.4.2 Métriques d'évaluation . . . . .	19
1.5 Conclusion . . . . .	20
<b>2 Construction et analyse de la base de données</b>	<b>21</b>
2.1 Données de consommation d'électricité des utilisateurs du réseau Enedis . . . . .	21
2.1.1 Collection et description générale . . . . .	21
2.1.2 Représentation en séries temporelles . . . . .	23

2.1.3	Volumétrie des données . . . . .	24
2.1.4	Analyse et visualisation . . . . .	25
2.2	Données météorologiques . . . . .	29
2.2.1	Facteurs météorologiques . . . . .	29
2.2.2	Analyse . . . . .	29
2.2.2.1	Ergodicité spatiale . . . . .	30
2.2.2.2	Sur-échantillonnage temporelle . . . . .	32
2.3	Fusion des bases de données . . . . .	32
2.4	Influence des facteurs météorologiques . . . . .	33
2.4.1	Influence de la température . . . . .	35
2.4.2	Influence de l'humidité . . . . .	38
2.4.3	Influence du rayonnement solaire . . . . .	39
2.5	Conclusion . . . . .	41
<b>3</b>	<b>Modèle proposé</b>	<b>42</b>
3.1	Formalisation de notre problème . . . . .	42
3.1.1	Notation . . . . .	42
3.1.2	Formalisation . . . . .	43
3.1.3	Hypothèses validées . . . . .	44
3.2	Modèles proposés dans la littérature . . . . .	44
3.3	Modèle proposé . . . . .	46
3.3.1	Architecture générale . . . . .	46
3.3.2	Étape de détection des saisonnalités par analyse temps-fréquence . . . . .	46
3.3.2.1	Principe . . . . .	46
3.3.2.2	Exemple . . . . .	48
3.3.3	Notre méthode de décomposition des séries temporelles . . . . .	48
3.3.3.1	Implémentation . . . . .	49
3.3.4	Modèles SARIMA pour la prévision des composantes saisonnières . . . . .	50
3.3.5	Modèles basée sur les réseaux LSTM pour les composantes tendance et résidu .	50
3.3.6	Réalisation . . . . .	54
3.3.7	Evaluation des performances . . . . .	56
3.3.7.1	Analyse des résultats par rapport à la métrique MAPE . . . . .	56
3.3.7.2	Analyse des résultats par rapport aux métriques MAE et RMSE . . . . .	56
3.4	Conclusion . . . . .	59

# Liste de Figures

1.1	Consommation d'électricité du profil RES11 (+ RES11WE) avec une puissance ]6-9] kVA, dans la région Auvergne-Rhône-Alpes . . . . .	5
1.2	Exemple de série multivariée . . . . .	6
1.3	Exemple de décomposition multiplicative . . . . .	8
1.4	Exemple de décomposition additive . . . . .	8
1.5	ACF et PACF de la série de la figure 1.1 . . . . .	15
1.6	ACF et PACF de la composante tendance de la série de la figure 1.1 . . . . .	16
1.7	ACF et PACF de la composante saisonnalité de la série de la figure 1.1 . . . . .	16
1.8	ACF et PACF de la composante résiduelle de la série de la figure 1.1 . . . . .	17
1.9	Spectrogramme de la série de la figure 1.1 . . . . .	19
2.1	Entête du DataFrame de la consommation d'électricité . . . . .	22
2.2	Consommation journalière d'électricité en Auvergne-Rhône-Alpes selon la plage de puissance souscrite et le jour de la semaine . . . . .	26
2.3	Consommation journalière d'électricité en Auvergne-Rhône-Alpes selon le profil et le jour de la semaine . . . . .	27
2.4	Saisonalité annuelle de la consommation d'électricité en Auvergne-Rhône-Alpes . . .	28
2.5	Distribution de la consommation d'électricité en Auvergne-Rhône-Alpes selon les jours de la semaine et profile . . . . .	28
2.6	la matrice de corrélation entre les stations de la région Auvergne-Rhône-Alpes . . . . .	31
2.7	Les colonnes de la base de données finale . . . . .	33
2.8	Corrélation des facteurs météorologiques avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	34
2.9	Corrélation des facteurs météorologiques avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes pour pour le profil RES11 (+ RES11WE) et sa plage de puissance P3 : ]6-9] kVA . . . . .	35

2.10	Corrélation de la température avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	36
2.11	Corrélation entre la tendance de la température et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	36
2.12	Corrélation entre la tendance de la température transformée et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	37
2.13	Corrélation de l'humidité avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes pour 2023-2024 . . . . .	38
2.14	Corrélation entre la tendance de l'humidité et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	38
2.15	Corrélation de la température avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	40
2.16	Corrélation entre la tendance du rayonnement solaire global et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	40
2.17	Corrélation entre la tendance du rayonnement solaire global et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	41
3.1	Architecture générale . . . . .	47
3.2	Pipeline de l'architecture générale . . . . .	48
3.3	Spectrogramme de la série temporelle de consommation d'électricité sur toute la région Auvergne-Rhône-Alpes $\left( \sum_{q \in \mathcal{C}} Y_k^{(84,q)} \right)_{kT_s \in \mathbb{T}}$ . . . . .	48
3.4	Exemple de pipeline pour l'analyse temps-fréquence et la décomposition des séries temporelles . . . . .	50
3.5	Prévision par $SARIMA(1,0,1)(1,1,1)_{22}$ de la composante saisonnière de période $P_1$ . . . . .	51
3.7	Les étapes du pipeline pour le modèle de prévision du résidu . . . . .	51
3.6	Prévision par $SARIMA(1,0,1)(1,1,1)_{44}$ de la composante saisonnière de période $P_2$ . . . . .	52
3.8	Transformation en séquences des données pour LSTM . . . . .	52
3.9	Architecture typique de notre modèle basé sur des réseaux LSTM . . . . .	53
3.10	Prévision du résidu de la série $\left( Y_k^{(r,q)} \right)_{kT_s \in \mathbb{T}}$ . . . . .	53
3.11	Prévision de la tendance de la série $\left( Y_k^{(r,q)} \right)_{kT_s \in \mathbb{T}}$ . . . . .	54
3.12	Prévision de la série $\left( Y_k^{(r,q)} \right)_{kT_s \in \mathbb{T}}$ . . . . .	54
3.13	Séparation des données . . . . .	56
3.14	Distribution de la métrique MAPE pour les différents profils en fonction des plages de puissance souscrite. . . . .	57
3.15	Distribution d'erreur MAPE pour chaque plage de puissance souscrite en fonction des profils . . . . .	57

3.16 Distribution de la métrique MAPE pour chaque plage de puissance souscrite en fonction des profils . . . . .	58
3.17 Distribution des métriques MAE et RMSE pour les différents profils en fonction des plages de puissance souscrite. . . . .	58
3.18 Distributions des métriques MAE et RMSE pour chaque plage de puissance souscrite en fonction des profils . . . . .	59

# Liste de Tableaux

2.1	Dictionnaire des données Enedis . . . . .	23
2.2	Profils et plages de puissance souscrite . . . . .	25
2.3	Coefficients de corrélation entre la tendance de la température et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes . . . . .	37
2.4	Corrélation des composantes saisonnières et résiduelles de la température avec les composantes respectives de la consommation d'électricité dans la région Auvergne-Rhône-Alpe	37
2.5	Coefficients de Corrélation entre la tendance de l'humidité et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes. (a) Pour toutes les saisons, (b) En séparant les saisons . . . . .	39
3.1	Liste des régions et leurs codes . . . . .	43

# Liste des algorithmes

1.1	Estimation de la saisonnalité d'une série temporelle . . . . .	9
1.2	Décomposition multiplicative d'une série temporelle . . . . .	10
3.1	Analyse temps-fréquence et détection des périodes dominantes . . . . .	47
3.2	Décomposition en cascade des séries temporelles . . . . .	49
3.3	Le script final . . . . .	55

# Introduction générale

## Contexte

La prévision est capitale autant pour la gestion de la production d'électricité que pour la gestion de sa consommation. Pour les opérateurs et fournisseurs d'électricité, une prévision précise de la quantité d'électricité à produire permettrait de bien équilibrer l'offre et la demande. Pour le consommateur, une prévision fiable de sa consommation permet de mieux gérer celle-ci et de détecter les éventuelles anomalies. Nous distinguons deux catégories de prévisions [11, 26, 34] : la prévision de la charge électrique et la prévision de la consommation d'électricité. La prévision de la charge électrique consiste à prévoir la quantité d'électricité qui sera nécessaire à un moment donné. Cette prévision est utilisée, par les opérateurs et fournisseurs d'électricité, pour produire et transporter la quantité d'énergie suffisante pour répondre aux besoins de consommation, tout en évitant une surproduction ou une sous-production d'électricité. Quant à la prévision de la consommation d'électricité, elle consiste à prévoir la quantité d'électricité qui sera consommée sur une période par un utilisateur ou un groupe d'utilisateurs.

En fonction de la durée de l'horizon temporel concerné, les prévisions de la charge électrique et de la consommation d'électricité peuvent être classées en trois types [11, 41] :

- ❖ Prévision à court terme : elle couvre une période pouvant aller jusqu'à une semaine. Les prévisions à court terme sont particulièrement importantes pour un suivi et une gestion en temps réel. Dans [41], un autre type, dit à très court terme est considéré pour des durées de quelques minutes à quelques heures.
- ❖ Prévision à moyen terme : elle varie d'une semaine à un an et est utilisée pour la planification de la maintenance et la gestion de la production et de la consommation. Cette prévision prend en compte les variations saisonnières de consommation d'électricité ainsi que les interruptions planifiées.
- ❖ Prévision à long terme : elle couvre généralement une période de plus d'un an et prend en compte des facteurs tels que les variations démographiques, le développement urbain à grande échelle, la croissance économique et les impacts de la politique énergétique. La prévision de charge à long terme se concentre sur la planification et l'optimisation des systèmes, aidant les opérateurs et les fournisseurs de l'électricité à prendre des décisions quant à l'investissement dans de nouvelles capacités de production d'énergie et à la manière d'équilibrer les différentes sources d'énergie.

Ces dernières années, les prévisions de la charge électrique ont suscité un grand intérêt tant dans la recherche académique que dans le secteur industriel. Les auteurs des revues de littérature [33, 38] ont montré une grande croissance du nombre de publications de travaux recherche axés sur ce sujet.

Cette tendance s'explique par les enjeux cruciaux liés à gestion de l'énergétique, à l'intégration des énergies renouvelables, à la gestion des systèmes intelligents (smart grids), à la participation active des consommateurs. Le volume croissant de données disponibles (météo, historique de consommation, caractéristiques thermiques des bâtiments, etc.) et la maturité des algorithmes d'apprentissage automatique contribuent aussi à cette dynamique. Cependant, nous avons constaté que la plupart de ces travaux de recherche se concentrent sur la prévision de la charge électrique et rares sont les travaux dédiés à la prévision de la consommation d'électricité. Ce projet a pour objectif l'établissement d'un modèle de prévision de la consommation d'électricité, à court terme, des utilisateurs du réseau Enedis en France.

## Problématique

Initialement, l'objectif était l'établissement d'un modèle de référence pour chaque consommateur, en particulier les logements résidentiels et d'un modèle de détection des anomalies dans la consommation et d'identification des éventuelles causes de celles-ci. Les jeux de données publiés par Enedis [16] présentent des agrégats de consommation d'électricité des utilisateurs du réseau Enedis, et ne contiennent pas de données à l'échelle d'un logement. Suite à plusieurs échanges avec Enedis, nous avons choisi le jeu de données [15] qui restitue la consommation d'électricité au pas d'une demi-heure des points de soutirage raccordés au réseau Enedis, dont la plage de puissance souscrite est inférieur ou égale à 36 kVA. Ce choix est justifié par la compatibilité de ces données avec nos objectifs initiaux et la possibilités d'une adaptation future pour détecter les anomalies dans la consommation d'électricité à l'échelle d'un logement individuel. Ce jeu de données contient les volumes d'énergie soutirés, les courbes de charge moyennes de clients dotés de compteurs communicants et le nombre de clients. Ces agrégats sont disponibles par région, par plage de puissance souscrite et par profil, définis par Enedis dans [14].

Nous avons, donc, cherché dans ce projet à établir un modèle fiable, robuste et précis pour la prévision de la consommation d'électricité, à court termes, en fonction du profil de l'utilisateur, de la plage de la puissance souscrite, de la région et des données météorologiques de celle-ci.

## Contribution

Pour atteindre les objectifs fixés et répondre à notre problématique nous avons :

- ❖ fait un état de l'art des différentes études effectuées sur la prévision de la charge électrique et de la consommation d'électricité,
- ❖ établi une base de données agrégeant les jeux de données d'Enedis et les données météorologiques (température, humidité, rayonnement solaire et vitesse du vent),
- ❖ analysé et étudié les relations entre les différentes variables explicatives et la consommation d'électricité,
- ❖ et établi un nouveau modèle, basé d'une part sur la décomposition par analyse temps-fréquence de la consommation d'électricité et des différentes variables explicatives, et d'autre part sur la combinaison des modèles autorégressifs SARIMA (pour *Seasonal AutoRegressive Integrated Moving Average*) et les réseaux de neurones récurrents à mémoire court-terme et long-terme LSTM (*pour Long Short-Term Memory*).

## **Organisation du document**

Nous adoptons, dans notre approche la modélisation de la consommation d'électricité, ainsi que les variables explicatives par des séries temporelles. Dans un premier chapitre nous présentons les outils mathématiques pour décrire, analyser et faire la prévision des séries temporelles. Le deuxième chapitre présente la construction de la base de données, les transformations appliquées sur celle-ci et une analyse des différentes relations entre les variables explicatives et la variable cible. Le dernier chapitre présente notre modèle. Afin de comparer nos résultats avec d'autre travaux, nous dressons une synthèse de la revue de littérature, qui ne vaut pas être exhaustive, du domaine de la prévision de la consommation d'électricité à court terme.

# Chapitre 1

## Introduction aux séries temporelles

Nous avons opté pour une modélisation en séries temporelles de la consommation d'électricité et des variables exogènes. Dans ce premier chapitre, nous rappelons les outils mathématiques dédiés à cette modélisation. Nous ferons, tout d'abord, quelques rappels succincts de la notion d'une série temporelle, de ses caractéristiques, de ses transformations et décompositions. Dans un second temps nous aborderons les outils d'analyse de séries temporelles, que nous avons utilisés afin d'en extraire des statistiques et des caractéristiques significatives. Puis nous présentons les notions de la prévision et les métriques associées à l'évaluation des performances de celle-ci. La théorie des séries temporelles est riche et large, nous nous limiterons dans notre présentation aux éléments que nous avons utilisés, et nous renvoyons le lecteur intéressé vers les références citées dans notre développement.

### 1.1 Généralités

#### 1.1.1 Définitions

**Définition 1.** On appelle processus stochastique toute famille  $(Y_t)_{t \in \mathbb{T}}$  de variables aléatoires définies sur un même espace de probabilité  $(\Omega, \mathcal{A}, \mathbb{P})$ , indexées par l'espace des temps  $\mathbb{T}$  et à valeurs dans un même espace des états  $(E, \mathcal{B})$ .

**Définition 2.** On appelle série temporelle tout processus stochastique  $(Y_t)_{t \in \mathbb{T}}$  où :

- ❖  $Y_t \in \mathbb{R}^d (d \geq 1)$
- ❖  $\mathbb{T} \subseteq \mathbb{N}$  ou  $\mathbb{T} \subseteq \mathbb{Z}$ ,
- ❖  $\forall t \in \mathbb{T} \mathbb{E}(Y_t^2) < +\infty$  (on parle de processus du second ordre)

**Définition 3.** Autrement dit une série temporelle est une suite d'observations quantitatives ordonnées dans le temps, généralement mesurées à intervalles réguliers [4, 21, 25, 40].

Une série temporelle est dite univariée si  $Y_t \in \mathbb{R}$  i.e.  $d = 1$ . Il s'agit d'une séquence de valeurs observées pour une seule variable au fil du temps.

Une série temporelle est dite multivariée si  $Y_t \in \mathbb{R}^d$  où  $d > 1$ . Il s'agit de plusieurs variables observées simultanément.

### 1.1.2 Exemple

La figure 1.1 montre une série temporelle univariée, qui représente la consommation d'électricité du profil Enedis « *RES11(+RES11WE)* » dont la puissance souscrite est  $[6 - 9] \text{ kVA}$ . Ces observations couvrent la période du 01/01/2023 au 31/12/2024 avec un pas de temps d'une demi-heure. Chaque valeur de cette série temporelle correspond, donc, à la consommation d'électricité de tous les points de soutirage, situés dans la région Auvergne-Rhône-Alpes, et qui sont de même profil et de même plage de puissance souscrite. Cette consommation est mesurée en Watt-heure (Wh).

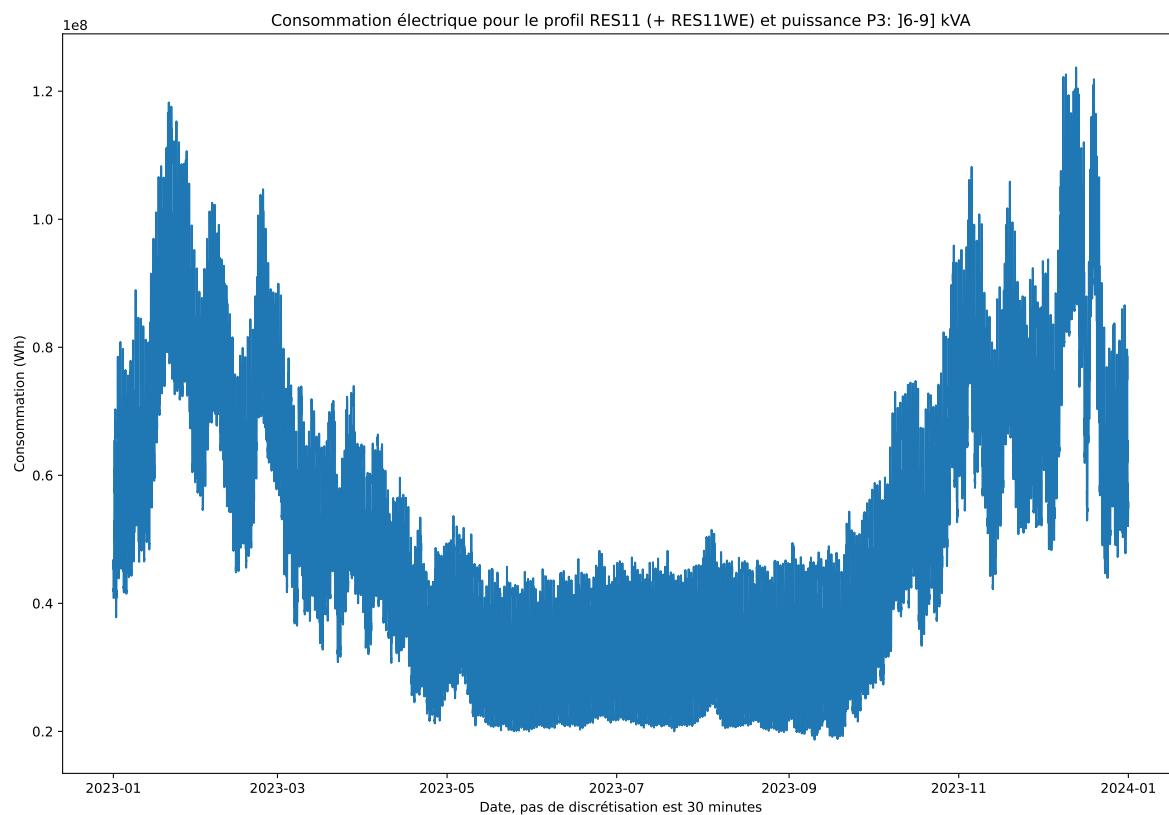


FIGURE 1.1 – Consommation d'électricité du profil RES11 (+ RES11WE) avec une puissance  $[6-9]$  kVA, dans la région Auvergne-Rhône-Alpes

La figure 1.2 montre un exemple de série temporelle multivariée, qui représente la série de l'exemple précédent observée simultanément avec des variables météorologiques. Cette série temporelle multivariée est composée des variables suivantes :

- ❖ la consommation d'électricité de tous les points de soutirage, situés dans la région Auvergne-Rhône-Alpes, et qui sont de même profil RES11 (+ RES11WE) et de même plage de puissance souscrite  $[6-9]$  kVA, exprimée en Wh,
- ❖ la température moyennée sur tous les départements de la région Auvergne-Rhône-Alpes, observée avec un pas de temps d'une demi-heure, sur la période du 01/01/2023 au 31/12/2024, et exprimée en  $^{\circ}\text{C}$ ,
- ❖ l'humidité moyennée sur tous les départements de la région Auvergne-Rhône-Alpes, observée avec un pas de temps d'une demi-heure, sur la période du 01/01/2023 au 31/12/2024, et exprimée en %,
- ❖ la vitesse du vent moyennée sur tous les départements de la région Auvergne-Rhône-Alpes, observée avec un pas de temps d'une demi-heure, sur la période du 01/01/2023 au 31/12/2024,

et exprimée en m/s.

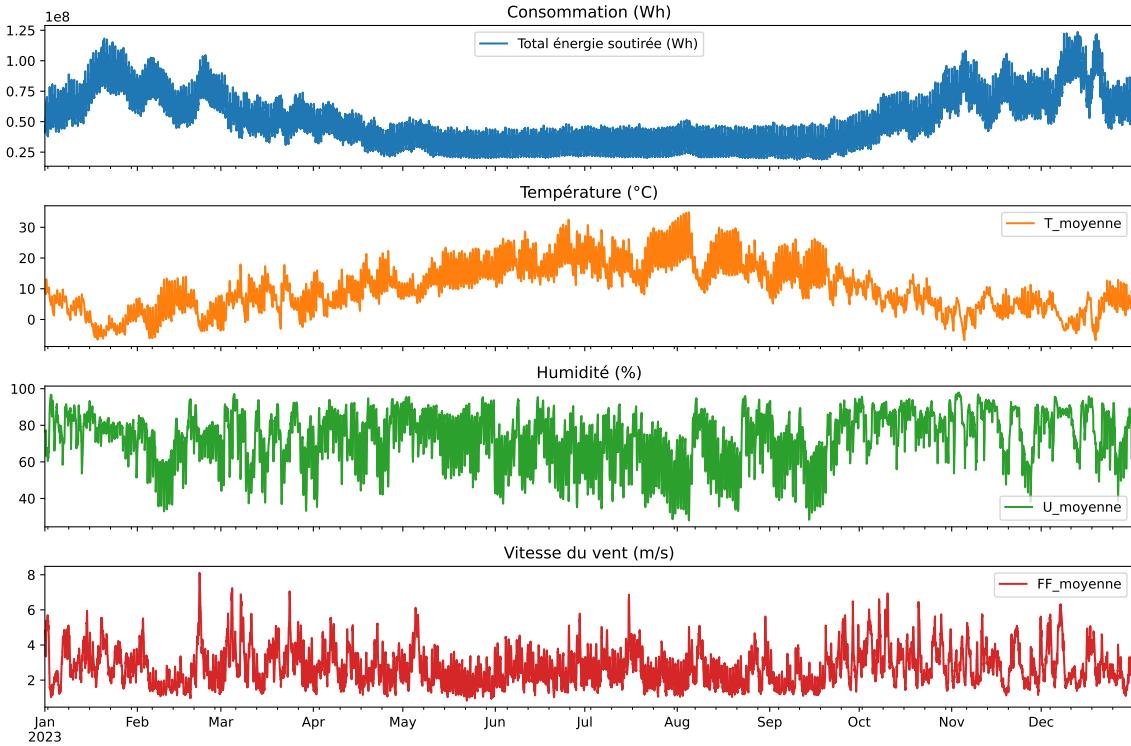


FIGURE 1.2 – Exemple de série multivariée

## 1.2 Décomposition d'une série

### 1.2.1 Composantes

Intrinsèquement une séries temporelle est constituée de plusieurs composantes qui la caractérise [4, 21, 25]. Ces composantes sont :

- ❖ la tendance, qui représente l'évolution à long terme.
- ❖ la ou les composantes saisonnières, qui correspondent à des répétitions régulières de motifs,
- ❖ la ou les composantes cycliques, qui représentent des fluctuations se produisant sur des périodes plus longues et de manière irrégulière,
- ❖ et le résidus, dit aussi bruit, qui correspond aux restes stochastiques après l'élimination des composantes déterministes. Il représente les variations aléatoires et imprévisibles.

En générale les composantes cycliques et la tendance sont regroupées en une seule composante cycle-tendance, qui est souvent appelée la tendance.

### 1.2.2 Décomposition additive et décomposition multiplicative

La décomposition d'une série temporelle consiste à extraire ses composantes principales. Elle peut être additive ou multiplicative, selon la nature des interactions entre les composantes de la série [4, 21, 25] .

Dans une décomposition additive, la série temporelle  $(Y_t)_{t \in \mathbb{T}}$  est exprimée comme la somme de ses composantes :

$$Y_t = T_t + S_t + C_t + R_t \quad (1.1)$$

où :

- ❖  $T_t$  est la tendance,
- ❖  $S_t$  est la saisonnalité,
- ❖  $C_t$  est la cyclicité,
- ❖  $R_t$  est le bruit.

Lorsque la tendance et le cycle sont regroupés en une seule composante, la décomposition additive s'écrit :

$$Y_t = T_t + S_t + R_t \quad (1.2)$$

Dans une décomposition multiplicative, la série temporelle  $(Y_t)_{t \in \mathbb{T}}$  est exprimée comme le produit de ses composantes :

$$Y_t = T_t \times S_t \times C_t \times R_t \quad (1.3)$$

Si la composante cyclique est incluse dans la composante tendance cette décomposition s'écrit :

$$Y_t = T_t \times S_t \times R_t \quad (1.4)$$

Si l'amplitude des fluctuations saisonnières autour de la tendance ne varie pas avec le niveau de la série temporelle, alors la décomposition additive est la plus appropriée. Lorsque la variation du motif saisonnier autour de la tendance semble proportionnelle au niveau de la série temporelle, une décomposition multiplicative est alors plus appropriée [11, 25].

Il faut noter qu'en appliquant une transformation logarithmique (respectivement exponentielle) nous pouvons passer d'une décomposition multiplicative (respectivement additive) à une décomposition additive (respectivement multiplicative). L'utilisation d'une décomposition additive sur une série temporelle qui a d'abord subi une transformation logarithmique de manière à ce que sa variation paraisse stable dans le temps équivaut à utiliser une décomposition multiplicative.

Par exemple, nous constatons pour la série de la figure 1.1 , que l'amplitude des fluctuations varie selon le niveau de la série :

- ❖ en hiver (début et fin de l'année), la consommation est élevée avec une forte variabilité,
- ❖ en été (milieu d'année), la consommation est plus basse et plus stable.

C'est un comportement typique du modèle multiplicatif décrit par l'équation 1.4 où :

- ❖ les fluctuations saisonnières (et les résidus) s'amplifient avec la tendance,
- ❖ la variance n'est pas constante, on parle dans ce cas d'une hétéroscédasticité visuelle.

Une transformation logarithmique ou une transformation de Box-Cox sert à stabiliser la variance avant la modélisation.

La figure 1.3 montre la décomposition multiplicative de la série de la figure 1.1 pour la période du 01/01/2023 au 31/01/2023.

Décomposition multiplicative de la série de la consommation électrique pour le profil RES11 (+ RES11WE) et la plage de puissance souscrite P3: ]6-9] kVA

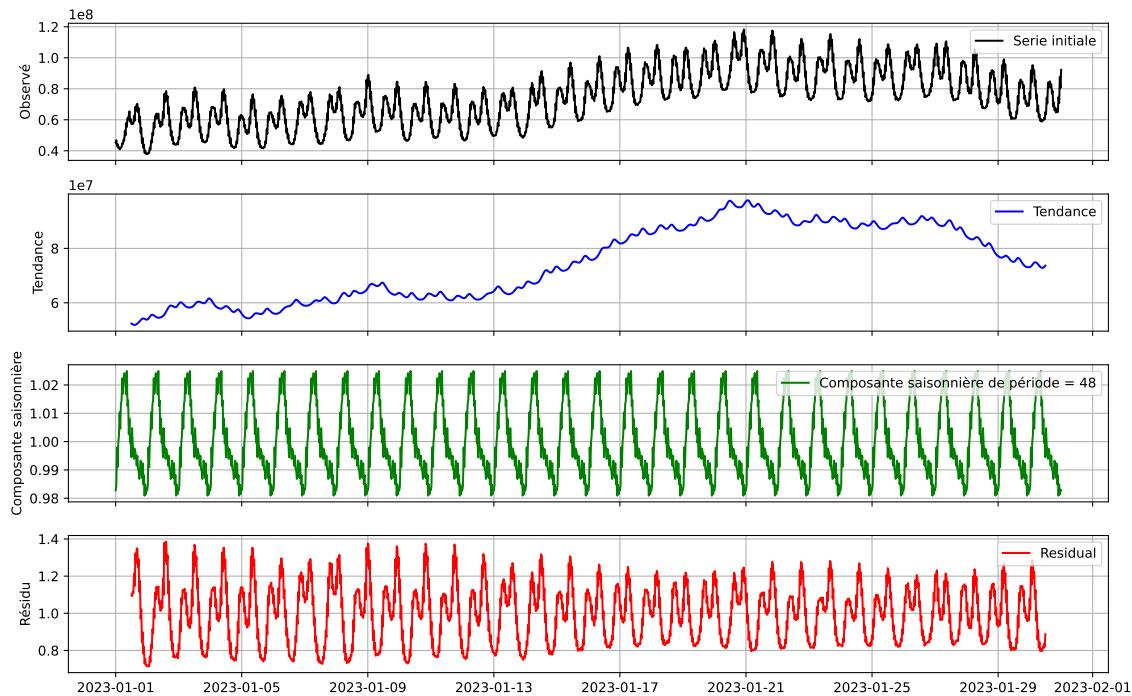


FIGURE 1.3 – Exemple de décomposition multiplicative

La figure 1.3 montre la décomposition additive de la série du figure 1.4 après transformation logarithmique, avec un zoom sur la période 01/01/2023 au 31/01/2023.

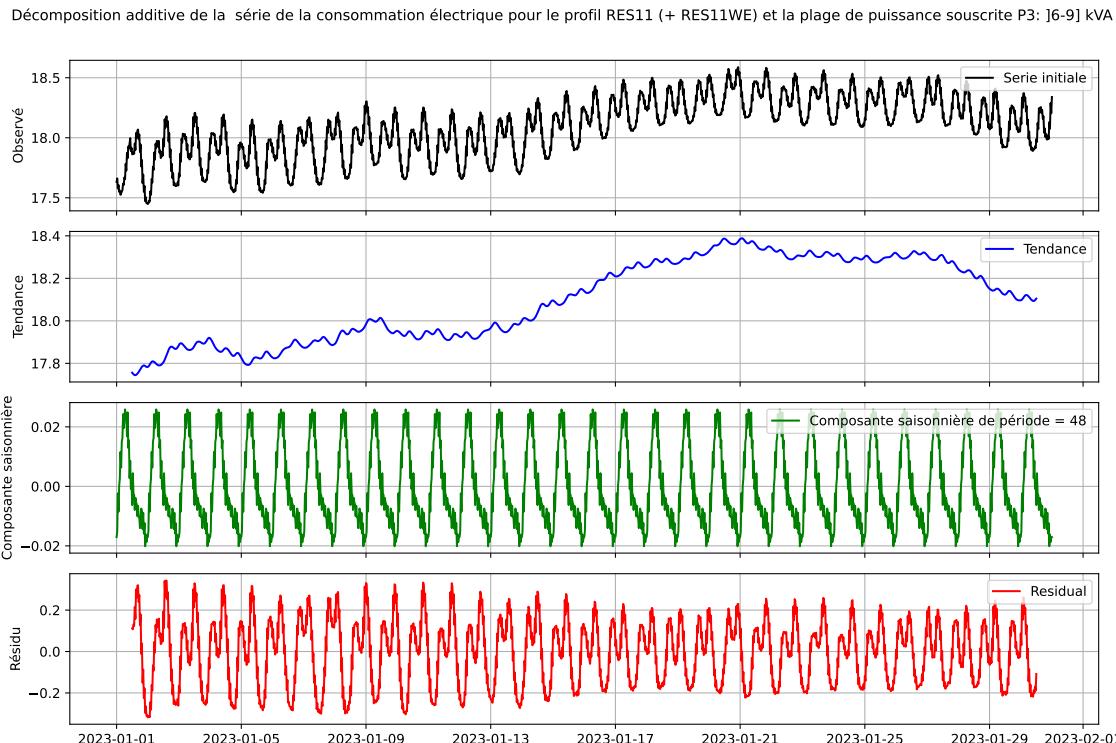


FIGURE 1.4 – Exemple de décomposition additive

**Remarque :** nous pouvons constater que le résiduel présente aussi des composantes saisonnière. En effet notre série est multi-saisonnière. Dans notre approche nous procédons à des décompositions en

cascade.

Dans [25] les auteurs présentent des méthodes d'estimation des composantes d'une série temporelle. Afin de ne pas s'écartez du cœur de notre sujet, nous ne détaillerons pas ici toutes ces méthodes, et nous envoyons le lecteur intéressé à cette ouvrage disponible gratuitement. Dans notre étude nous avons fait usage de la fonction *seasonal\_decompose* de statsmodel [39], qui repose sur un modèle de décomposition classique des séries temporelles, que nous décrivons dans ce qui suit.

On considère une série temporelle  $(Y_t)_{t \in \mathbb{T}}$ , qu'on peut décomposer selon l'équation 1.2 :  $Y_t = T_t + S_t + R_t$ . L'estimation de la composante tendance est basée sur une moyenne glissante. Si la saisonnalité est de période  $p = 2n + 1$  alors la tendance est donnée par l'équation :

$$\widehat{T}_t = \frac{1}{2n+1} \sum_{k=-n}^{k=n} Y_{t+k} \quad (1.5)$$

Si la période est  $p = 2n$ , celle-ci est donnée par l'équation :

$$\widehat{T}_{t+n} = \frac{1}{4n} \sum_{k=0}^{k=2n-1} Y_{t+k} + \frac{1}{4n} \sum_{k=0}^{k=2n-1} Y_{t+k+1} \quad (1.6)$$

Une fois la tendance est estimée, on peut alors estimer la saisonnalité par l'algorithme 1.1 :

---

**Algorithm 1.1** Estimation de la saisonnalité d'une série temporelle

---

❖ Soustraction de la tendance pour isoler la saisonnalité :

$$D_t = Y_t - \widehat{T}_t \quad (1.7)$$

❖ Groupement des  $D_t$  selon le modulo de  $t$  par rapport à la période  $p$ , et calcul de la moyenne sur les  $N$  périodes complets des données pour chaque  $j \in \{0, \dots, p-1\}$  :

$$\widehat{S}_j = \frac{1}{N} \sum_{k=0}^{k=N-1} D_{j+k.p} \quad (1.8)$$

❖ On centre la saisonnalité pour qu'elle ait une somme nulle

$$\widehat{S}_j \leftarrow \widehat{S}_j - \frac{1}{p} \sum_{k=0}^{k=p-1} \widehat{S}_k$$


---

Enfin le résidu s'obtient :

$$\widehat{R}_t = Y_t - \widehat{T}_t - \widehat{S}_t \quad (1.9)$$

Dans le cas d'une décomposition multiplicative, une transformation logarithmique est appliquée, puis la décomposition additive est effectuée, et ensuite une transformation exponentielle est appliquée 1.2.

**Algorithm 1.2** Décomposition multiplicative d'une série temporelle

❖ Transformation logarithmique :

$$Y_t = T_t \times S_t \times R_t \Rightarrow \log(Y_t) = \log(T_t) + \log(S_t) + \log(R_t) \quad (1.10)$$

❖ Décomposition additive de la série  $\log(Y_t)$

$$\log(Y_t) = T_t^1 + S_t^1 + R_t^1 \quad (1.11)$$

❖ Transformation exponentielle :

$$\log(Y_t) = T_t^1 + S_t^1 + R_t^1 \Rightarrow Y_t = \underbrace{\exp(T_t^1)}_{T_t} \times \underbrace{\exp(S_t^1)}_{S_t} \times \underbrace{\exp(R_t^1)}_{R_t} \quad (1.12)$$

## 1.3 Analyse des séries temporelles

Nous distinguons entre deux approches complémentaires d'analyse. L'approche temporelle et l'approche spectrale. La décomposition, décrite précédemment et qui fait parti de l'analyse temporelle, permet d'identifier les composantes de la série analysée, mais ne révèle pas les dépendances entre les valeurs passées et les valeurs présentes. L'étude des fonctions d'autocorrélation et d'autocorrélation partielle permet de mesurer cette dépendance. En complément, l'analyse de Fourier permet de transformer la série pour révéler sa structure fréquentielle. Cependant cette analyse n'est pas adaptée pour les séries temporelles non stationnaires, i.e. dont les caractéristiques spectrales varient au cours du temps. Pour ce type de séries, une analyse temps-fréquence est plus adaptée. Dans cette section, nous présentons la notion de stationnarité, les fonctions d'autocorrélation et d'autocorrélation partielle, ainsi que les différentes méthodes d'analyse temps-fréquence. Nous mentionnons pour chaque outil son intérêt pour notre étude.

### 1.3.1 Stationnarité

#### 1.3.1.1 Définition

Une série temporelle est dite stationnaire au sens strict si toutes ses propriétés statistiques (moyenne, variance, moments statistiques d'ordre supérieurs, autocorrélation, ...) sont invariantes dans le temps. Une série temporelle est dite stationnaire au sens large, ou faiblement stationnaire, si sa valeur moyenne et sa fonction d'autocorrélation sont invariantes dans le temps [18]. Plus généralement, Une série temporelle dont les caractéristiques spectrales ne varient pas au cours du temps est dite stationnaire, a contrario, une série temporelle dont les caractéristiques spectrales varient au cours du temps est dite non stationnaire. De nombreux méthodes d'analyse et de modélisation des séries temporelles reposent sur l'hypothèse de stationnarité, car elle garantit la prédictibilité stable du processus sous-jacent.

Nous donnons ici les définitions formelles de ces deux notions.

**Définition 4. Stationnarité stricte :** une série temporelle  $(Y_t)_{t \in \mathbb{T}}$  est dite « strictement stationnaire » (ou « fortement stationnaire ») si pour tout vecteur  $(t_1, \dots, t_n) \in \mathbb{T}^n$ , avec  $n \geq 1$ , les vecteurs de variables aléatoires  $(Y_{t_1}, \dots, Y_{t_n})$  et  $(Y_{t_1+k}, \dots, Y_{t_n+k})$  ont la même loi et ce pour tout décalage temporel  $k \in \mathbb{Z}$ .

**Définition 5. Stationnarité faible (du second ordre)** : une série temporelle  $(Y_t)_{t \in \mathbb{T}}$  est dite « faiblement stationnaire » (ou « stationnaire au second ordre ») si

- ❖  $\forall t \in \mathbb{T} E(Y_t) = m$ , où  $m$  est une constante.
- ❖  $\forall (t, s) \in \mathbb{T} \forall k \in \mathbb{Z} \text{Cov}(Y_t, Y_s) = \text{Cov}(Y_{t+k}, Y_{s+k})$  où  $\text{Cov}(Y_t, Y_s) = E(Y_t Y_s) - E(Y_t)E(Y_s)$

### 1.3.1.2 Formes de non stationnarité

Il existe deux formes simples de non stationnarité en fonction de leur source. La première, appelée non-stationnarité TS (pour *Trend Stationary* en anglais), est due à la présence dans la série d'une tendance déterministe. Les séries décrites par le modèle de l'équation 1.13 sont un exemple de cette forme.

$$Y_t = c + bt + \epsilon_t \quad (1.13)$$

où  $\epsilon_t$  est un bruit blanc ( $E(\epsilon_t) = 0$ ) et de variance  $V(\epsilon_t) = \sigma_\epsilon^2$ . Dans ce modèle, l'espérance mathématique dépend du temps :  $E(Y_t) = c + bt$ , par contre la variance  $V(y_t) = \sigma_\epsilon^2$  est invariant dans le temps. La suppression de la tendance permet de stationnariser ce type de série. Les séries  $Y_t - c - bt$  et  $Y_t - bt$  sont stationnaires.

La deuxième forme, stationnaire en différence DS (pour *Difference Stationary* en anglais), est due à la présence dans la série d'une tendance stochastique. Les séries décrites par le modèle de l'équation 1.14 sont un exemple de cette forme.

$$Y_t = c + Y_{t-1} + \epsilon_t \quad (1.14)$$

où  $\epsilon_t$  est un bruit blanc d'espérance  $E(\epsilon_t) = 0$  et de variance  $V(\epsilon_t) = \sigma_\epsilon^2$ . Dans ce modèle, l'espérance mathématique dépend du temps :  $E(Y_t) = E(Y_0) + ct$ , et la variance aussi  $V(y_t) = tV(\epsilon_t)$ . La série obtenue par différentiation, i.e.  $\Delta Y_t = Y_t - Y_{t-1}$  est stationnaire.

### 1.3.1.3 Tests de stationnarité

L'identification et la caractérisation de la stationnarité ou la non-stationnarité peuvent être effectuées grâce à des tests statistiques. Ces tests aident à déterminer s'il faut stationnariser la série, en effectuant des transformations. Il existe un grand nombre de tests [4, 21, 27, 28, 36] dont les plus utilisés pour leur simplicité sont les tests de Dickey et Fuller Augmenté (ADF) [12], le test de Phillips-Perron (PP) [36] et le test de Kwiatkowski–Phillips–Schmidt–Shin (KPSS) [28]. Dans cette section nous donnons un aperçu sur ces trois tests que nous avons utilisés dans notre projet. Pour expliquer rigoureusement ces tests, nous allons rappeler quelques notions utiles.

**Differentiation** : la différentiation est une transformation qui permet de stationnariser une série en éliminant la tendance qui cause la non stationnarité. Une différentiation de premier ordre, par exemple, consiste à utiliser la différence entre deux observations consécutives. Il est parfois nécessaire de différencier avec un ordre supérieur la série temporelle afin d'éliminer une tendance polynomiale par exemple. On parlera alors de différentiation d'ordre  $d$ . La différentiation d'ordre  $d$  est décrite par l'équation récursive 1.15, ou de manière équivalente en utilisant l'opérateur de retard  $\mathbb{L}$ , i.e.  $\mathbb{L}Y_t = Y_{t-1}$  par l'équation 1.16 :

$$Y_t^{(d)} = \begin{cases} Y_t - Y_{t-1} & \text{si } d = 1 \\ Y_t^{(d-1)} - Y_{t-1}^{(d-1)} & \text{si } d > 1 \end{cases} \quad (1.15)$$

$$Y_t^{(d)} = \Delta^d Y_t = (1 - \mathbb{L})^d Y_{t-1} \quad (1.16)$$

**Racine unitaire :** pour comprendre cette notion, nécessaire à l'explication des principes de test on considère le modèle décrit par l'équation 1.17 :

$$Y_t = \rho_1 Y_{t-1} + \dots + \rho_p Y_{t-p} + c + bt + \epsilon_t \quad (1.17)$$

où  $\epsilon_t$  est un bruit blanc.

En utilisant l'opérateur de retard  $\mathbb{L}$ , l'équation précédente s'écrit :

$$\underbrace{(1 - \rho_1 \mathbb{L} + \dots + \rho_p \mathbb{L}^p)}_{\Phi(L)} Y_t = c + bt + \epsilon_t \quad (1.18)$$

$\Phi$  est un polynôme de degré  $p$ . On note  $(z_i)_{i \in \llbracket 1, p \rrbracket}$  les racines de ce polynôme. L'équation 1.18 devient :

$$\underbrace{(1 - z_1^{-1} \mathbb{L}) \dots (1 - z_p^{-1} \mathbb{L})}_{\Phi(L) \text{ factorisé}} Y_t = c + bt + \epsilon_t \quad (1.19)$$

Si une racine des  $(z_i)_{i \in \llbracket 1, p \rrbracket}$  est égale à 1 alors on dit que la série a une racine unitaire.

### Test de Dickey-Fuller Augmenté (ADF)

Le test ADF test la présence de racine unitaire dans une série temporelle. L'hypothèse nulle stipule que la série est non stationnaire, i.e. la série a une racine unitaire, tandis que l'hypothèse alternative indique que la série n'a pas de racine unitaire et donc elle est stationnaire.

$$\begin{aligned} \mathcal{H}_0 & : \text{la série a au moins une racine unitaire} \\ \mathcal{H}_1 & : \text{la série a une racine unitaire} \end{aligned}$$

### Test de Phillips-Perron (PP)

Le test PP constitue une extension du test ADF, conçue pour traiter de manière robuste l'autocorrélation et l'hétéroscléasticité des résidus. Il s'appuie sur les mêmes modèles que ceux du test de Dickey et Fuller simple en corrigeant la statistique de test par des méthodes non paramétriques. L'hypothèse nulle demeure la présence d'une racine unitaire (non-stationnarité).

$$\begin{aligned} \mathcal{H}_0 & : \text{la série a au moins une racine unitaire} \\ \mathcal{H}_1 & : \text{la série n'a pas de racine unitaire} \end{aligned}$$

### Test de Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

Le test de KPSS adopte une logique complémentaire en inversant les hypothèses du test ADF. Son hypothèse nulle postule que la série est stationnaire (autour d'une moyenne ou d'une tendance déterministe), tandis que l'hypothèse alternative suggère une non-stationnarité due à la présence d'une racine unitaire.

- $\mathcal{H}_0$  : la série n'a pas de racine unitaire
- $\mathcal{H}_1$  : la série a au moins une racine unitaire

## Exemples

Afin d'illustrer avec un exemple, nous appliquons ces tests sur notre série de la figure 1.1.

### ❖ Test ADF

#### \* Test sur la série originelle

- ◊ Statistique ADF : -2.80
- ◊ P-value : 0.059
- ◊ Seuil critique à 1% : -3.43, à 5% : -2.86 et à 10% : -2.57}
- ◊ **Interprétation** : La statistique ADF est légèrement supérieure (en valeur absolue) au seuil critique de 5%. La *p-value* est proche de 0.05 mais légèrement au-dessus (0.059), ce qui conduit à ne pas rejeter l'hypothèse nulle ( $H_0$ ) de racine unitaire au seuil de 5%.
- ◊ **Conclusion** : La série n'est pas stationnaire. Elle présente probablement une tendance stochastique ou une racine unitaire.

#### \* Test sur la série différenciée d'ordre 1

- ◊ Statistique ADF : -24.79
- ◊ P-value : 0.000
- ◊ Seuil critique à 1% : -3.43
- ◊ **Interprétation** : La statistique ADF est très inférieure aux seuils critiques, et la *p-value* est nulle. On rejette fortement l'hypothèse nulle ( $H_0$ ) de racine unitaire.
- ◊ **Conclusion** : La série différenciée d'ordre 1 est stationnaire.

### ❖ Test KPSS

#### \* Test KPSS sur la série originale :

- ◊ Statistique : 1.5614
- ◊ P-value : 0.0100
- ◊ Lags utilisés : 104
- ◊ Seuils critiques : 10% : 0.347, 5% : 0.463, 2.5% : 0.574, 1% : 0.739
- ◊ **Interprétation** : Plus la statistique KPSS est élevée par rapport aux seuils critiques, plus l'évidence de non-stationnarité est forte. Ici, 1.5614 est bien au-dessus même du seuil le plus strict (1%), ce qui indique une forte présence d'une tendance stochastique ou non stationnarité dans la série. La *p-value* est proche de 0.01 et inférieure à 0.05, ce qui conduit à rejeter l'hypothèse de stationnarité. La série est non stationnaire (on rejette  $H_0$  de stationnarité au seuil de 1%).
- ◊ **Conclusion** : La série n'est pas stationnaire. Elle présente probablement une tendance stochastique ou une racine unitaire.

#### \* Test KPSS pour la série différenciée d'ordre 1 :

- ◊ Statistique : 0.3649 (valeur illustrative adaptée pour la p-value donnée)
- ◊ P-value : 0.1000
- ◊ Lags utilisés : 974
- ◊ Seuils critiques : 10% : 0.347, 5% : 0.463, 2.5% : 0.574, 1% : 0.739
- ◊ **Interprétation** : La statistique (0.3649) se situe entre le seuil de 10% (0.347) et celui de 5% (0.463). La p-value d'environ 0.10 suggère qu'on ne rejette pas  $H_0$  au seuil de 5%, mais on serait proche de la frontière au seuil de 10%.

- ◊ **Conclusion :** la série différenciée est considérée comme stationnaire au seuil de 5 %, bien que cette stationnarité soit « faible » ou « limite » si on utilisait un seuil de 10 %. Ce test confirme la stationnarité de la série différenciée d'ordre 1 établie par le test ADF.

### 1.3.2 Fonctions d'autocorrélation et d'autocorrélation partielle

Les fonctions d'autocorrélation, abrégée en **ACF** (pour *AutoCorrelation Function*), mesure la corrélation entre les deux observations décalées dans le temps d'une série temporelle. Quand à la fonction d'autocorrélation partielle, abrégée en **PACF** (pour *Partial AutoCorrelation Function*), elle mesure cette corrélation directe en éliminant les influences des observations intermédiaires entre les deux observations concernées [5, 25]. Dans ce qui suit nous donne les définitions formelles.

**Définition 6.** La fonction d'autocorrélation d'une série temporelle  $(Y_t)_{t \in \mathbb{Z}}$ , notée  $\rho$ , est définie, pour tout  $(t, k) \in \mathbb{Z}$ , par :

$$\rho(t, k) = \text{Corr}(Y_{t+k}, Y_t) = \frac{\text{Cov}(Y_{t+k}, Y_t)}{\mathbb{V}(Y_t)} = \frac{E(Y_{t+k} Y_t) - \mathbb{E}(Y_{t+k}) \times \mathbb{E}(Y_t)}{E(Y_t^2) - \mathbb{E}(Y_t)^2} \quad (1.20)$$

Pour une série temporelle  $(Y_t)_{t \in \mathbb{T}}$  stationnaire (faiblement ou strictement), cette fonction est indépendant de l'origine du temps  $t$ , et définie pour tout  $k \in \mathbb{Z}$  par :

$$\rho(k) = \text{Corr}(Y_k, Y_0) = \frac{\mathbb{E}(Y_k Y_0) - \mathbb{E}(Y_k) \times \mathbb{E}(Y_0)}{\mathbb{E}(Y_0^2) - \mathbb{E}(Y_0)^2} \quad (1.21)$$

**Définition 7.** La fonction d'autocorrélation partielle d'une série temporelle  $(Y_t)_{t \in \mathbb{Z}}$ , notée  $r$  ou  $\text{Corr}$ , est définie, pour tout  $(t, k) \in \mathbb{Z}$ , par :

$$r(t, k) = \text{Corr}(Y_{t+k}^p, Y_t^p) \quad (1.22)$$

où  $Y_{t+k}^p$  et  $Y_t^p$  sont les projections linéaires des variables  $Y_{t+k}, Y_t$  sur l'espace vectoriel engendré par les variables  $Y_{t+1}, \dots, Y_{t+k-1}$ .

Pour une série temporelle  $(Y_t)_{t \in \mathbb{T}}$  stationnaire (faiblement ou strictement), cette fonction est indépendante de l'origine du temps  $t$ , et définie pour tout  $k \in \mathbb{Z}$  par :

$$r(t, k) = r(k) = \text{Corr}(Y_k^p, Y_0^p). \quad (1.23)$$

**Exemple 8.** La figure 1.5 montre les fonctions ACF et PACF pour la série de la figure 1.1 différenciée.

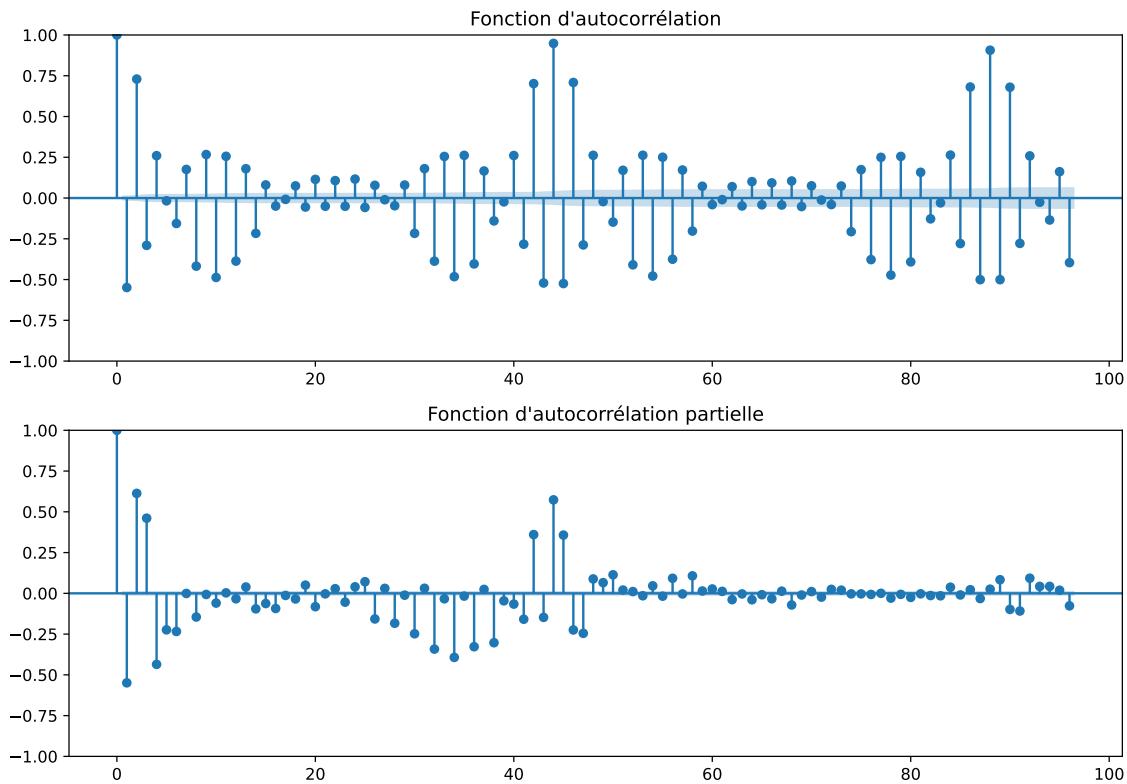


FIGURE 1.5 – ACF et PACF de la série de la figure 1.1

L’analyse des ces graphiques permet d’identifier une saisonnalités multiples (périodes de 22 et 44), une décroissance lente de la fonction d’autocorrélation. La fonction PACF montre un coefficient fort au décalage  $k = 1$  et quelques autres coefficients plus faibles mais significatifs.

Après décomposition de la série, avec une période de 44, nous analysons les fonctions ACF et PACF de ses composantes. Les figures 1.6, 1.7 et 1.8 montrent celles-ci.

Pour la composante résiduelle 1.8, nous constatons que les pics sont généralement proches de zéro mais à plusieurs décalage, les amplitudes des pics dépassent encore la bande de confiance. Ce qui montre que la corrélation n’est pas totalement nulle. Nous concluons qu’il y a encore de la saisonnalité et qu’il faut faire une décomposition en cascade. Nous détaillerons ce point dans le prochain chapitre. La présence d’autres saisonnalités est clairement confirmée par les fonctions ACF et PACF du résidu.

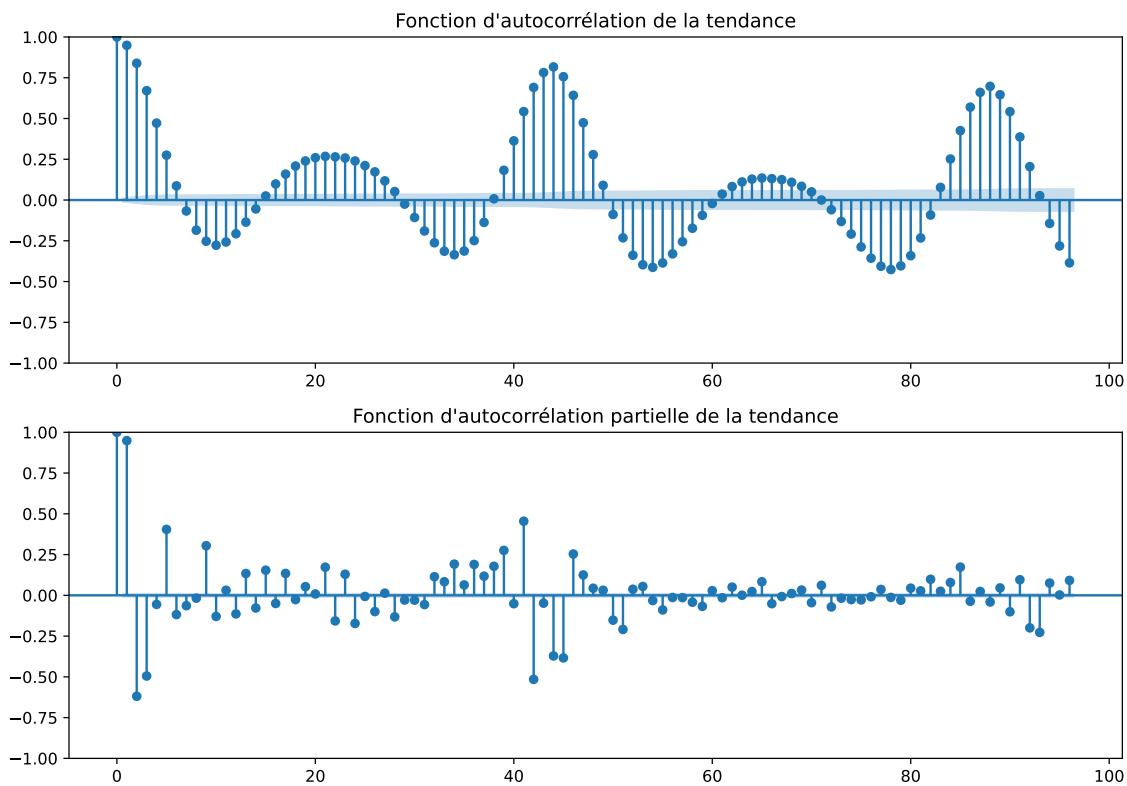


FIGURE 1.6 – ACF et PACF de la composante tendance de la série de la figure 1.1

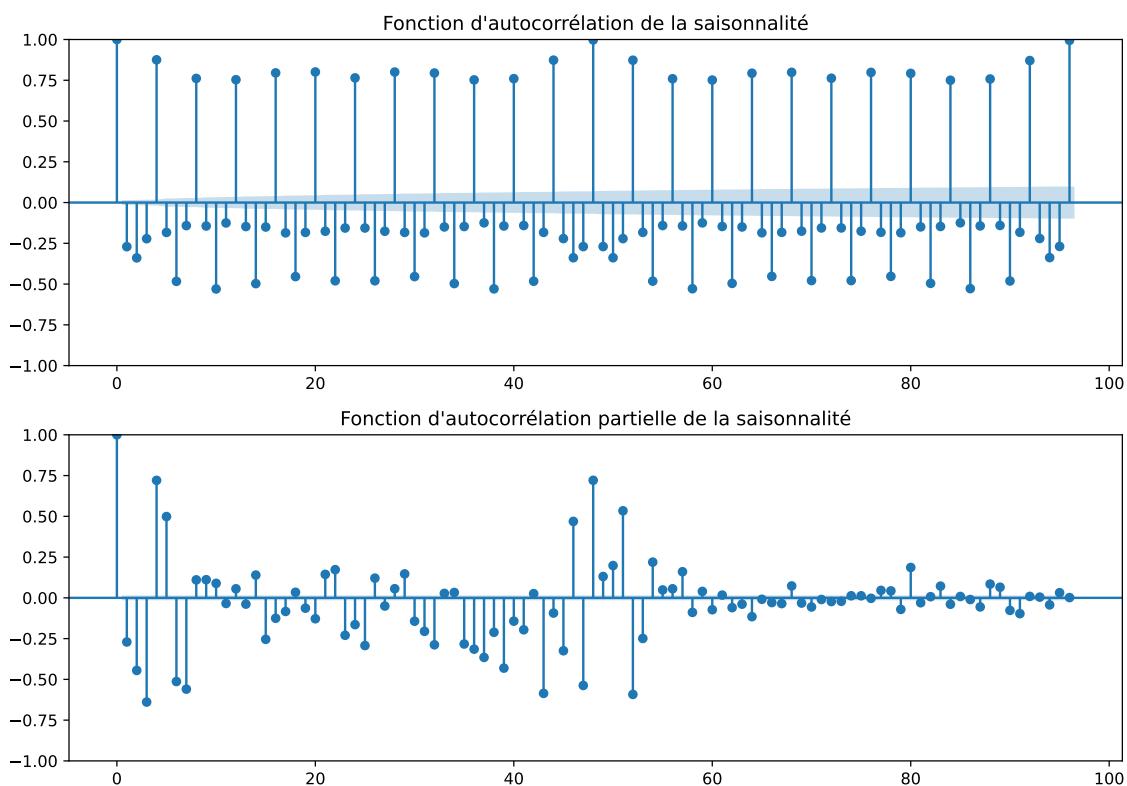


FIGURE 1.7 – ACF et PACF de la composante saisonnalité de la série de la figure 1.1

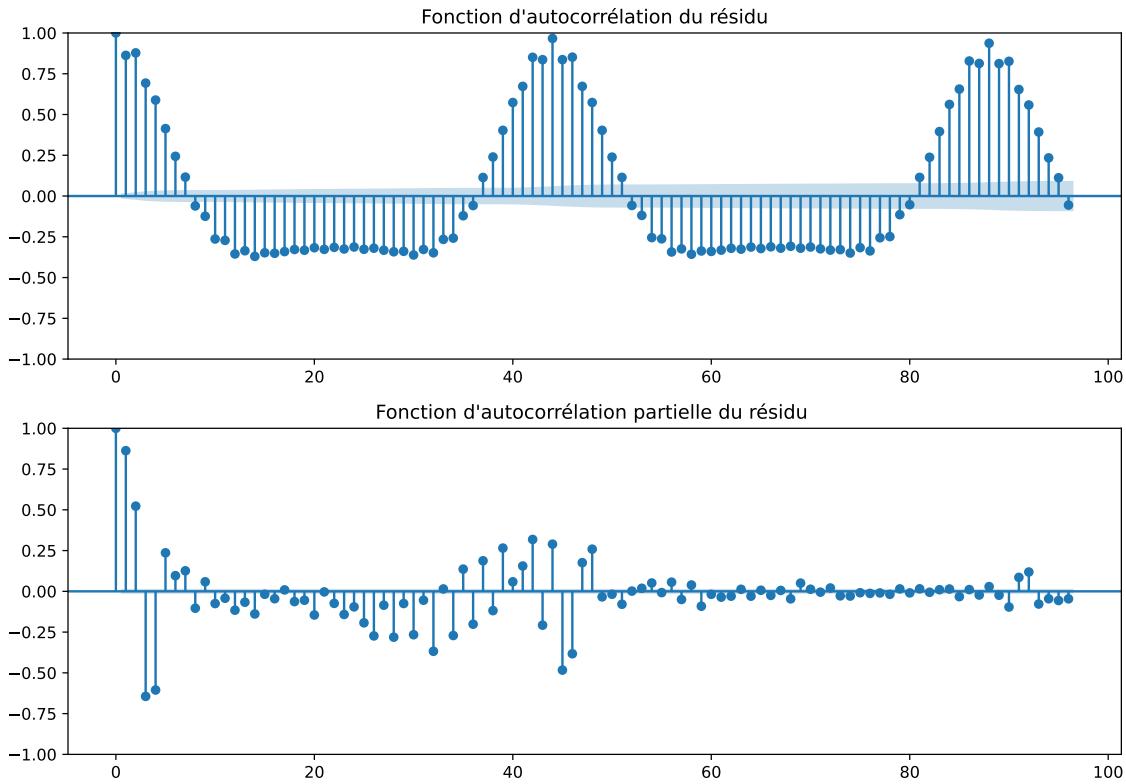


FIGURE 1.8 – ACF et PACF de la composante résiduelle de la série de la figure 1.1

### 1.3.3 Analyse temps-fréquence

Les séries temporelles de notre projet sont des séries non stationnaires. Cette constatation nous a conduit à utiliser des outils spécifiques et adaptés aux signaux non-stationnaires pour analyser ces séries. Parmi ces outils figurent les méthodes d'analyse temps-fréquence, qui permettent de représenter conjointement les informations temporelle et fréquentielle. Elles fournissent une représentation qui permet d'extraire les différentes saisonnalités présentes dans nos séries. Les méthodes les plus utilisées, détaillées dans [18], pour effectuer cette analyse sont :

- ❖ La transformée de Fourier à court terme (Short-Time Fourier Transform (STFT)) qui consiste à diviser la série en fenêtres puis appliquer la transformée de Fourier localement. Cette transformée permet de construire une représentation temps-fréquence, appelée spectrogramme, avec une résolution temps-fréquence fixe.
- ❖ La transformée en ondelettes (Wavelet Transform (CWT)) qui utilise des fonctions de base localisées en temps et en fréquence (ondelettes), en permettant ainsi une résolution adaptative. Elle donne une bonne résolution temporelle pour les hautes fréquences et une bonne résolution fréquentielle pour les basses fréquences.
- ❖ La décomposition en modes empiriques (Empirical Mode Decomposition (EMD)) proposée par Huang dans [24], et qui consiste à décomposer le signal, d'une façon adaptative, en somme de composantes oscillantes, appelées IMF (pour intrinsic mode function en anglais). Cette décomposition est adaptée pour les signaux non-stationnaires. En utilisant cette décomposition nous pouvons estimer l'amplitude et la fréquence instantanées et établir une représentation temps-fréquence de la série [2].

Nous nous limitons ici à l'explication du principe du STFT que nous avons utilisée pour estimer les différentes fréquences présentes dans nos séries.

**Transformée de Fourier à Court Terme** Son principe consiste à diviser la série en  $P$  segments de longueur  $N$  et avec un recouvrement  $\rho$  donnée, les pondérer par une fonction localisée en temps  $w$  (fenêtre), puis appliquer la transformée de Fourier sur chaque segment pondéré. Pour une série temporelle  $(Y_n)_{n \in \mathbb{T}}$ , après fenêtrage on obtient la série  $(Y_{1,n}, \dots, Y_{P,n})_{n \in \mathbb{T}}$  issues de la pondération par  $w$ .  $(Y_{p,n})_{n \in \mathbb{T}}$ , qui correspond à la pondération de la  $p^{\text{ème}}$  segment par  $w$ , est définie par l'équation 1.24.

$$Y_{p,n} = Y_{((1-\rho)N-1) \times p+n} \times w_n \quad (1.24)$$

L'équation 1.25 donne un exemple classique de  $w$ .

$$w_n = a_0 - (1 - a_0) \cos(2\pi n/N) \quad (1.25)$$

Si  $a = 0,56$  il s'agit d'une fenêtre d'Hamming, et si  $a_0 = 0,5$  il s'agit d'une fenêtre de Hann.

La transformée de Fourier à court terme est l'ensemble des segments, notée  $STFT((Y_n)_{n \in \mathbb{T}}) = (X_{1,k}, \dots, X_{P,k})_{k \in \{0, \dots, K\}}$ , où  $(X_{p,k})_{k \in \{0, \dots, K\}}$ , la transformée de Fourier de la  $p^{\text{ème}}$  segment est définie par :

$$X_{p,k} = \sum_{t=0}^{K-1} Y_{p,t} e^{-\frac{j2\pi kn}{K}} \quad (1.26)$$

où  $K$  est le nombre de points pour le calcul de la transformée de Fourier.

Le spectrogramme est

$$\mathbb{S} = \left\{ |X_{p,k}|^2, (p, k) \in \{0, \dots, P-1\} \times \{0, \dots, K-1\} \right\} \quad (1.27)$$

Dans notre cadre, l'objectif de cette analyse est d'extraire les périodes dominantes pour identifier toutes les saisonnalités. L'ensemble de ces périodes pour un seuil de détection  $\delta$  est donnée par l'équation 1.28 :

$$\mathbb{P}_\delta = \left\{ \frac{K}{k} T_s \text{ tel que } \forall p \in \{0, \dots, P-1\} \frac{|X_{p,k}|^2}{\max_{l \in \{0, \dots, K\}} \frac{1}{P} \sum_{m=0}^{P-1} |X_{m,l}|^2} \geq \delta \right\} \quad (1.28)$$

où  $T_s = \frac{1}{F_s}$  est la période d'échantillonnage, ou le pas de discréttisation temporelle de la série, et  $F_s$  dénote sa fréquence d'échantillonnage.

La figure 1.9 montre le spectrogramme de la série de la figure 1.1. Le pas de discréttisation temporelle de la série est de 30 minutes. La fréquence d'échantillonnage est  $F_s = 1/1800 \text{ Hz}$ . Ce spectrogramme est obtenu par STFT avec des fenêtres de longueur de 7 jours, i.e.  $L = 7 * 48$ , et avec un recouvrement de deux jours ( $2 \times 48 = 96$ ). La pondération est faite par une fenêtre de Hann.

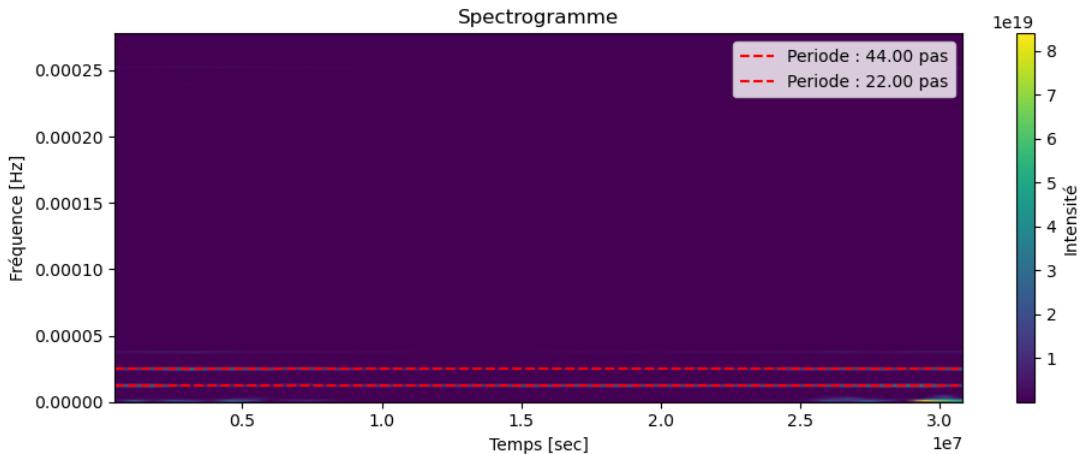


FIGURE 1.9 – Spectrogramme de la série de la figure 1.1

Ce spectrogramme confirme la multi-saisonnalité, montrée par les fonctions ACF et PACF. Les périodes détectées sont  $T_1 = 44 \text{ pas} \simeq 0.92 \text{ jours}$  et  $T_2 = 22 \text{ pas} \simeq 0.46 \text{ jour}$

## 1.4 Prévision et métriques d'évaluation

### 1.4.1 Prévision

Dans [25], Hyndman & Athanasopoulos donnent pour définition à la prévision (**forecasting**) : "Forecasting is about predicting the future as accurately as possible, given all the information available, including historical data and knowledge of any future events that might impact the forecasts."

Autrement dit la prévision d'une série temporelle univariée consiste à estimer ou prédire ses valeurs futures, en utilisant les informations disponibles sur ses valeurs passées. Si la série est multivariée, la prévision consiste à estimer ses valeurs futures, en utilisant ses valeurs passées et les valeurs des variables explicatives.

Étant donnée une série temporelle  $(Y_t)_{t \in \mathbb{T}} = (Y_{1,t}, Y_{2,t}, \dots, Y_{d,t})_{t \in \mathbb{T}}$ , notons  $\tau = \max(\mathbb{T})$ ,  $\mathcal{F}_\tau$  l'ensemble d'information disponible jusqu'au l'instant  $\tau$ , incluant les valeurs passées de la variable cible et des variables explicatives. On suppose que la variable cible est  $(Y_{1,t})_{t \in \mathbb{T}}$ . La prévision consiste à établir un modèle  $\mathcal{M}$  qui permet d'estimer le vecteur  $(Y_{1,\tau+1}, Y_{1,\tau+2}, \dots, Y_{1,\tau+h})$  en fonction de  $\mathcal{F}_\tau$  :

$$\left( \hat{Y}_{1,\tau+1}, \hat{Y}_{1,\tau+2}, \dots, \hat{Y}_{1,\tau+h} \right) = \mathcal{M}(\mathcal{F}_\tau) \quad (1.29)$$

$\mathcal{F}_\tau$ , constitue l'ensemble de caractéristiques à utiliser pour la prévision par le modèle  $\mathcal{M}$ . Le nombre  $h$ , dit horizon, détermine le type de prévision, i.e. court terme, moyen terme ou long terme.

### 1.4.2 Métriques d'évaluation

Il existe de nombreuses métriques d'évaluation applicables aux approches de prévision de séries temporelles. Nous adoptons dans notre projets sont les métriques les plus utilisées [33, 38, 42].

- ❖ **MAPE ( Mean Average Percentage Error :** l'erreur percentile absolue moyenne est très utilisée dans la cadre des prévisions de la consommation d'électricité. En reprenant les notations de la section précédente, le score MAPE s'exprime par la formule suivante :

$$MAPE = \frac{100}{k} \sum_{i=1}^k \left| \frac{\hat{Y}_{1,t+i} - Y_{1,t+i}}{Y_{1,t+i}} \right| \quad (1.30)$$

où  $Y_{1,t+i}$  est la vraie valeur à l'instant  $t + i$ ,  $\hat{Y}_{1,t+i}$  est la prévision à l'instant  $t + i$ , et  $k$  est le nombre de prévisions évaluées.

- ❖ **MAE (Mean absolute error) :** l'erreur absolue moyenne est également très populaire dans la cadre des prévisions de la consommation d'électricité car. Cette métrique est facile à comprendre et à calculer. le score MAE s'exprime par la formule suivante :

$$MAE = \frac{1}{k} \sum_{i=1}^k \left| \hat{Y}_{1,t+i} - Y_{1,t+i} \right| \quad (1.31)$$

- ★ Simple à interpréter.
- ★ Sensible aux échelles des données.
- ★ Même unité que la variable cible.

- ❖ **RMSE (Root mean squared error) :**

$$MAE = \sqrt{\frac{1}{k} \sum_{i=1}^k \left| \hat{Y}_{1,t+i} - Y_{1,t+i} \right|^2} \quad (1.32)$$

- ★ Même unité que la variable prédite.
- ★ Souvent utilisé pour comparer des modèles sur une même échelle.

## 1.5 Conclusion

Dans ce chapitre nous avons posé les bases théoriques indispensables, à la compréhension, à l'analyse et à la modélisation, pour le travail réalisé pendant ce projet. Nous avons ainsi défini et illustré à travers des exemples de notre projet les concepts suivants :

1. la notion d'une série temporelle,
2. les composantes d'une série temporelle,
3. la décomposition d'une série temporelle
4. les fonctions d'autocorrélation et d'autocorrélation partielle,
5. la notion de stationnarité, qui caractérise la l'invariance ou non des propriétés statistiques dans le temps,
6. l'analyse temps-fréquence d'une série temporelle,
7. et les métriques d'évaluation.

## Chapitre 2

# Construction et analyse de la base de données

La consommation d'électricité est influencée, d'une part par des facteurs météorologiques tels que la température, la vitesse du vent, le rayonnement solaire, l'humidité, la couverture nuageuse, le type et l'intensité des précipitations [20]. Et d'autre part elle peut être influencée par autres facteurs, tels que les comportements des consommateurs (télétravail, autoproduction, contrôle actif de la consommation, etc.), les informations calendaires (jour de la semaine, jours spéciaux et fériés), les propriétés thermiques des bâtiments et le prix de l'électricité [7, 11, 20].

Notre bases de données est issues des jeux de données publiés par Enedis [15]. Ces données sont agrégées par région, par plage de puissance souscrite et par profil [14]. Elles contiennent également des informations calendaires. Cependant elles ne contiennent pas de données météorologiques. Pour compléter notre base de données avec les autres facteurs potentiellement influents sur la variable cible, nous avons collecté, re-échantillonné et nettoyé les données météorologiques [32] afin de les fusionner avec les données de consommation pour une période de deux années consécutives (2023 et 2024).

Dans ce chapitre, nous présentons notre base de données en expliquant toutes les étapes : téléchargement, construction de la base de données, nettoyage de celle-ci, fusion, analyse et sélection des variables exogènes à garder pour la modélisation.

## 2.1 Données de consommation d'électricité des utilisateurs du réseau Enedis

### 2.1.1 Collection et description générale

Enedis publie les données de consommation et de production d'électricité en France trimestriellement (30 jours après la fin de chaque trimestre) sur deux années glissantes<sup>1</sup>. Les jeux de données publiés présentent des agrégats de consommation et de production d'électricité des utilisateurs du réseau Enedis au pas  $1/2h$  [13]. Ces agrégats sont fournis à la maille nationale et régionale (12 régions administratives métropolitaines).

1. Pour répondre à l'obligation (L111-73-1) créée dans le cadre de la Loi Pour une République Numérique, et conféremment au décret 2017-486 et son arrêté d'application paru le 29 décembre 2017, qui fixent les modalités détaillées de mise en Open Data de données de consommation et production par les GRD et GRT, Electricité et Gaz.

Ces jeux de données contiennent 3 familles d'agrégats :

- ❖ Le nombre de points de soutirage / d'injection
- ❖ L'énergie totale soutirée / injectée
- ❖ Les courbes moyennes de consommation / production

Leur publication est faite au travers de trois jeux de données principaux contenant les croisements attendus pour les différents segments d'utilisateurs du réseau Enedis :

- ❖ Consommateurs  $\leq 36$  kVA : agrégats publiés par Profil et Plage de puissance souscrite
- ❖ Consommateurs  $> 36$  kVA : agrégats publiés par Profil, Plage de puissance souscrite et Secteur d'activité
- ❖ Producteurs : agrégats publiés par Filière de production et Plage de puissance d'injection

Les profils types définis par Enedis sont également publiés et détaillés dans [14].

Comme nous l'avons mentionné dans l'introduction générale, nous avons collecté parmi ces données, le jeu le plus compatible avec nos objectifs initiaux et dont l'usage pour notre modèle offre la possibilité d'une adaptation future pour détecter d'éventuelles anomalies dans la consommation d'électricité à l'échelle résidentielle.

Nous avons, donc, collecté le jeu de données restituant **l'énergie totale soutirée au pas  $1/2 h$  des points de soutirage** pour des **consommateurs** avec une puissance totale **inférieure à 36kVA**. Le tableau 2.1 résume les champs de ce jeu avec une description détaillée. Le tableau 2.1 donne un aperçu de l'entête du DataFrame correspondant avec les différentes colonnes avant la fusion avec les données météorologique.

	AAAAMMMJJHH	Région	Profil	Plage de puissance souscrite	Nb points soutirage	Total énergie soutirée (Wh)	Courbe Moyenne n°1 (Wh)	Indice représentativité Courbe n°1 (%)	Courbe Moyenne n°2 (Wh)	Indice représentativité Courbe n°2 (%)	Courbe Moyenne n°1 + n°2 (Wh)	Indice représentativité Courbe n°1 + n°2 (%)	Jour max du mois (0/1)	Semaine max du mois (0/1)
0	2023010100	Auvergne-Rhône-Alpes	ENT3 (+ ENT4 + ENT5)	P0: Total <= 36 kVA	1374.0	2081460.0	1734.0	49	1264.0	48	1499.0	98	0.0	0.0
1	2023010100	Auvergne-Rhône-Alpes	ENT3 (+ ENT4 + ENT5)	P0: Total <= 36 kVA	1374.0	2081460.0	1734.0	49	1264.0	48	1499.0	98	0.0	0.0
2	2023010101	Auvergne-Rhône-Alpes	ENT3 (+ ENT4 + ENT5)	P0: Total <= 36 kVA	1374.0	2081460.0	1734.0	49	1264.0	48	1499.0	98	0.0	0.0
3	2023010101	Auvergne-Rhône-Alpes	ENT3 (+ ENT4 + ENT5)	P0: Total <= 36 kVA	1374.0	2081460.0	1734.0	49	1264.0	48	1499.0	98	0.0	0.0
4	2023010102	Auvergne-Rhône-Alpes	ENT3 (+ ENT4 + ENT5)	P0: Total <= 36 kVA	1374.0	2081460.0	1734.0	49	1264.0	48	1499.0	98	0.0	0.0

FIGURE 2.1 – Entête du DataFrame de la consommation d'électricité

Nom de la colonne	Format	Description
Horodate	Objet	Date-Heure au pas $1/2\ h$
Région	Objet	Nom de la région
Profil	Objet	Profil type au sens de la RecoFlux
Plage de puissance souscrite	Objet	La puissance électrique souscrite par l'utilisateur dans son contrat de fourniture.
Nb points soutirage	int64	Nombre de sites avec un contrat actif sur le réseau Enedis
Total énergie soutirée (Wh)	float64	Volume d'électricité consommée sur la $1/2\ h$ donnée par l'ensemble des sites du profil et de la plage de puissance considérée.
Courbe Moyenne n°1 (Wh)	float64	La moyenne des volumes d'électricité consommés sur la $1/2\ h$ donnée par des sites équipés de compteurs communicants et faisant partie du groupe dont le ratio (Conso 8h-20h)/(Conso totale) est le plus élevé
Indice représentativité Courbe n° 1 (%)	float64	Le ratio entre le nombre de points sur la Courbe Moyenne 1 et le nombre de points total de la même catégorie de client (même profil, même plage de puissance souscrite)
Courbe Moyenne n°2 (Wh)	float64	La Courbe Moyenne 2 correspond à la moyenne des volumes d'électricité consommés sur la $1/2\ h$ , donnée par des sites équipés de compteurs communicants et faisant partie du groupe dont le ratio (Conso 8h-20h)/(Conso totale) est le plus bas
Courbe Moyenne n°1 + n°2 (Wh)	float64	La Courbe Moyenne n°1 + n°2 correspond à la moyenne des volumes d'électricité consommés sur la $1/2\ h$ donnée par tous les sites équipés de compteurs communicants pris en compte dans les courbes 1 et 2.
Indice représentativité Courbe n° 1 + n° 2 (%)	float64	Le ratio entre le nombre de points sur lesquels a été basé le calcul de la Courbe Moyenne n°1 + n°2 et le nombre de points total de la même catégorie de client (même profil, même plage de puissance souscrite)
Jour Max du mois (0/1)	float64	Elle indique si la $1/2\ h$ considérée fait partie du jour qui a vu le pic de consommation d'énergie du mois, en France métropolitaine.
Semaine Max du mois (0/1)	float64	Elle indique si la $1/2\ h$ considérée fait partie de la semaine qui contient le jour max du mois.

TABLE 2.1 – Dictionnaire des données Enedis

### 2.1.2 Représentation en séries temporelles

Pour notre problématique de prévision nous avons besoin des colonnes : *Horodate*, *Région*, *Profil*, *Plage de puissance souscrite*, *Nb points soutirage* et *Total énergie soutirée (Wh)*. Pour une région donnée, un profil et une plage de puissance souscrite de ce profil, les données de la colonne *Total énergie soutirée (Wh)* constituent une série temporelle. La colonne *Nb points soutirage*, qui est aussi une série temporelle, représente le nombre de points de soutirage, i.e. le nombre de sites avec un contrat actif sur le réseau Enedis, pour toutes les  $1/2\ h$  d'une même journée. Celle-ci peut être utilisée pour calculer la consommation moyenne d'un profil au pas d'une  $1/2\ h$  et faire la prévision de celle-ci à la place de la consommation totale. Ce moyennage permet de réduire la dynamique de la série temporelle et améliorer la précision du modèle. Nous discuterons ce point dans le chapitre dédié à la prévision.

Pour une région donnée, un profil et une plage de puissance souscrite, nous avons donc :

- ❖ un ensemble d'instants d'observation, issue de la colonne *Horodate*, que nous représentons par  $\mathbb{T} = \{kT_s, k \in \{0, \dots, L\}\}$  où :
  - ★  $T_s = 30 \text{ min}$  est la période d'échantillonnage
  - ★  $L$  le nombre d'observation, pour 2023 et 2024 nous avons :  $L = (365 + 366) * 48 = 35088$
- ❖ une série temporelle qui correspond à la colonne *Total énergie soutirée (Wh)* , et que nous notons  $(Y_k)_{kT_s \in \mathbb{T}}$
- ❖ une série temporelle qui représente *Nb points soutirage* , que nous notons  $(N_k)_{kT_s \in \mathbb{T}}$
- ❖ une série temporelle qui représente consommation moyenne des consommateurs de même profil et même puissance souscrite  $(\bar{Y}_k)_{kT_s \in \mathbb{T}} = \left(\frac{Y_k^g}{N_k}\right)_{kT_s \in \mathbb{T}}$

Notre objectif est de faire un modèle de prévision pour la série  $(\bar{Y}_k)_{kT_s \in \mathbb{T}}$  puis en déduire  $(Y_k)_{kT_s \in \mathbb{T}}$ .

### 2.1.3 Volumétrie des données

Pour donner une idée sur la volumétrie de cette base de données, nous analysons les colonnes *Région*, *Profil* et *Plage de puissance souscrite*.

**Région** Les données sont publiées par région pour les 12 régions administratives métropolitaines : Bretagne, Normandie, Pays de la Loire, Nouvelle Aquitaine, Centre Val de Loire, Île de France, Hauts de France, Grand Est, Bourgogne-Franche-Comté, Auvergne-Rhône-Alpes, Occitanie, Provence-Alpes-Côte d'Azur.

**Profil et Plage de puissance souscrite** L'ensemble des profils et les méthodes utilisées pour profiler les consommateurs sont détaillés dans le document réglementaire RTE [1]. Dans notre base, il y a 12 profils, faisant partie de trois catégories : résidentiel, professionnel et ENT. Le tableau 2.2 résume ces profils et les plages de puissances souscrites associées. Il y a 73 configurations en total. Nous avons donc  $73 \times 12 = 876$  séries temporelles dont chacune est de longueur  $L = (365 + 366) * 48 = 35088$ .

Profile		Plage de puissance souscrite
ENT3(+ENT4+ENT5)		P0 : Total <= 36 kVA
PRO1(+PRO1WE)		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P2 : ]3-6] kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA, P8 : ]24-30] kVA P9 : ]30-36] kVA
PRO2(+PRO2WE+PRO6)		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA, P8 : ]24-30] kVA P9 : ]30-36] kVA
PRO3		P0 : Total <= 36 kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA
PRO4		P0 : Total <= 36 kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA
PRO5		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P2 : ]3-6] kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA, P9 : ]30-36] kVA
RES1(+RES1WE)		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P2 : ]3-6] kVA,
RES11(+RES11WE)		P0 : Total <= 36 kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA, P9 : ]30-36] kVA
RES2(+RES5)		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA, P7 : ]18-24] kVA, P8 : ]24-30] kVA P9 : ]30-36] kVA
RES2WE		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P3 : ]6-9] kVA, P4 : ]9-12] kVA
RES3		P0 : Total <= 36 kVA, P1 : ]0-3] kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA
RES4		P0 : Total <= 36 kVA, P4 : ]9-12] kVA, P5 : ]12-15] kVA, P6 : ]15-18] kVA

TABLE 2.2 – Profils et plages de puissance souscrite

#### 2.1.4 Analyse et visualisation

Dans cette section, nous présentons l'analyse et la visualisation des données, que nous avons effectuées lors de la première phase du projet. Celle-ci a permis d'identifier l'importance de chaque facteur et la manière dont elle influence sur la variable cible. Par souci de synthèse, nous ne présentons pas toute l'analyse effectuée mais seulement une partie qui nous semble la plus pertinente. Pour toutes les courbes que nous montrons ici, nous avons utilisé la consommation moyennée par rapport au nombre de points de soutirage  $(\bar{Y}_k)_{kT_s \in \mathbb{T}} = \left( \frac{Y_k}{N_k} \right)_{kT_s \in \mathbb{T}}$  pour la région Auvergne-Rhône-Alpes. Nous rappelons que pour chaque profil et chaque plage de puissance souscrite, la série temporelle  $(\bar{Y}_k)_{kT_s \in \mathbb{T}}$  reflète la consommation moyenne de cette catégorie de clients Enedis.

**Saisonnalité intra-journalière** Les courbes de la figure 2.2 montrent l'évolution par heure de la consommation d'électricité pour chaque jour de la semaine et chaque plage de puissance souscrite. Les courbes de la figure 2.3 montrent l'évolution par heure de la consommation d'électricité pour chaque jour de la semaine et chaque profil.

Ces figures nous permettent de constater la présence de saisonnalité multiple qui dépend de la plage de puissance souscrite et du profil. Ceci est en accord avec l'analyse temps-fréquence que nous avons montrée dans le premier chapitre . Elle montrent aussi l'effet hebdomadaire pour certains profils, i.e. la consommation dépend du jours de la semaine (jours ouvrés et week-ends). Cependant dans la plupart des cas la cyclicité intra-journalière ne dépend pas du jour de la semaine.

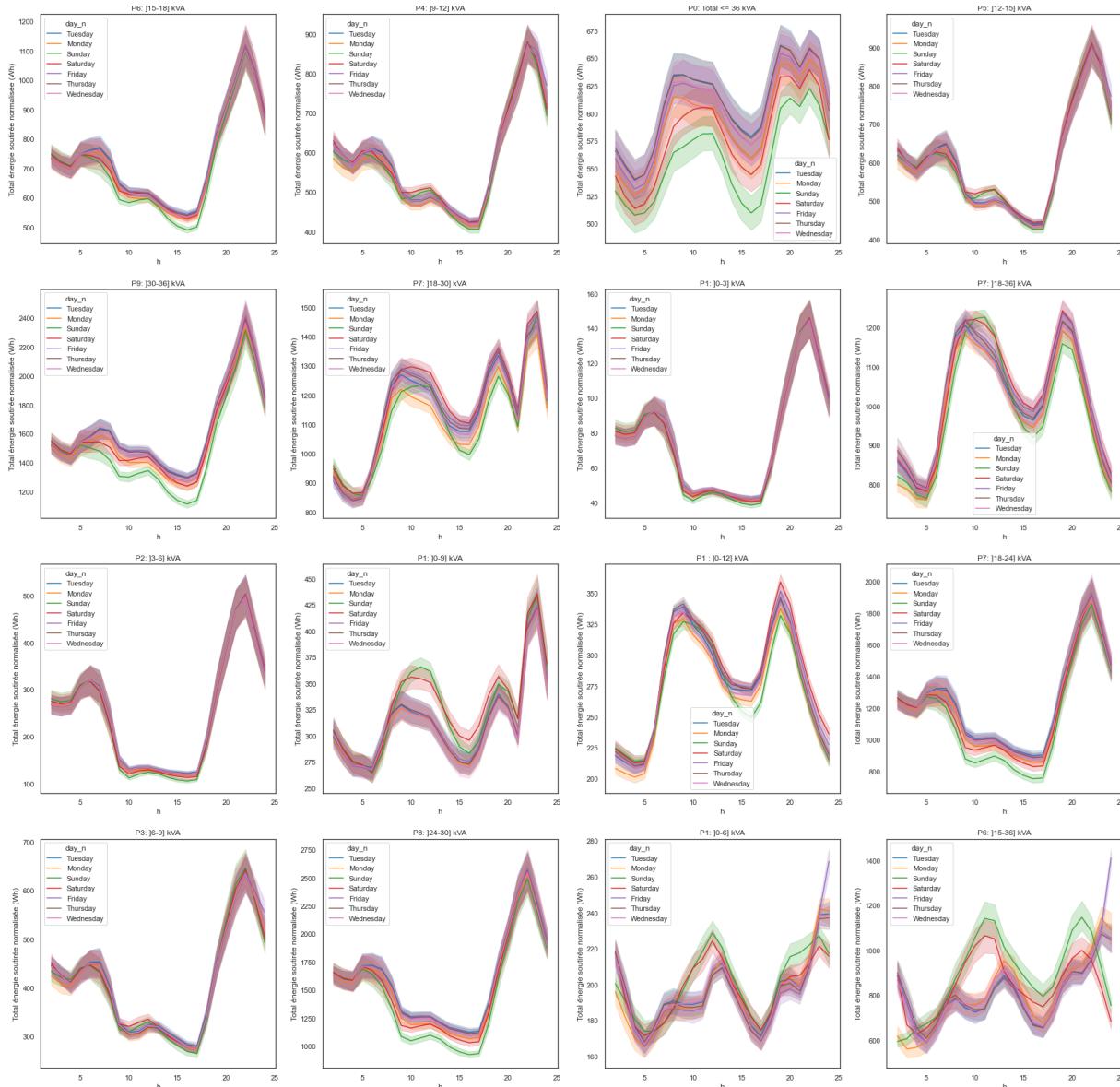


FIGURE 2.2 – Consommation journalière d'électricité en Auvergne-Rhône-Alpes selon la plage de puissance souscrite et le jour de la semaine

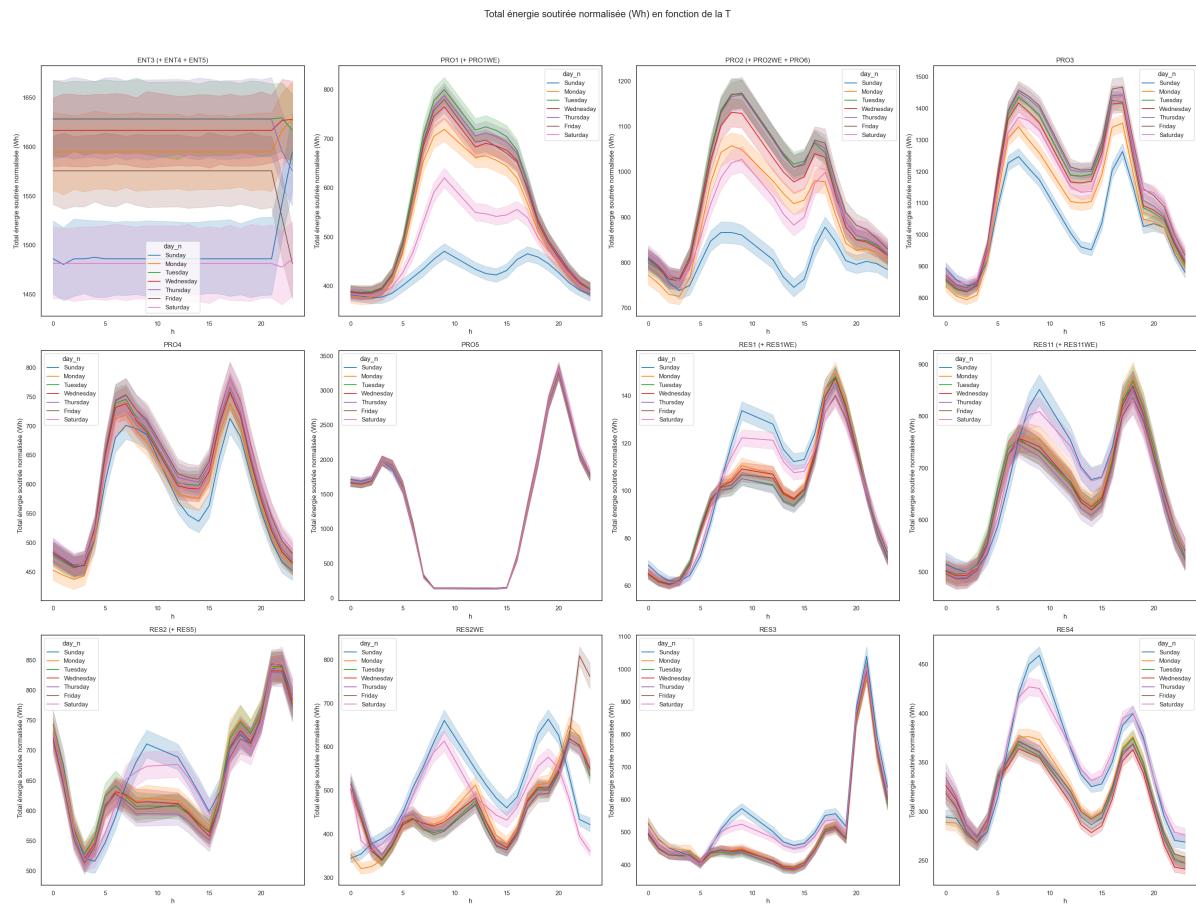


FIGURE 2.3 – Consommation journalière d'électricité en Auvergne-Rhône-Alpes selon le profil et le jour de la semaine

Ces analyses justifient la modélisation multi-niveaux (par région, par profil et par plage de puissance souscrite) basées sur des séries temporelles multivariées et hiérarchiques.

**Saisonnalité annuelle** La figure 2.4 montre la saisonnalité annuelle de nos séries temporelles. Ce graphique montre une saisonnalité intra-annuelle prononcée, différenciant l'été et l'hiver. Pour une catégorie résidentielle, nous stipulons que cette saisonnalité est principalement causée par la consommation d'électricité des appareils de chauffage et de climatisation, et dépend donc principalement de la température [11].

**Influence du jour de la semaine** Certaines études affirment que la distinction entre jours ouvrés, week-end et jours spéciaux (jours fériés, jours de fêtes,...) est importante dans le domaine de la prévision de la consommation d'électricité à court terme [7,8,33] [7,8,33]. Nous avons constaté que la différence entre jours ouvrés et week-end n'est significative que pour certains profils de catégorie professionnelle ou ENT. La figure illustre le fait que les profils résidentiels conservent un usage plus stable contrairement aux profils professionnels. De plus la sélection et le traitement des jours spéciaux étant un sujet assez large pour pouvoir former un projet à part, nous n'avons pas intégrée cette variable dans notre modèle.

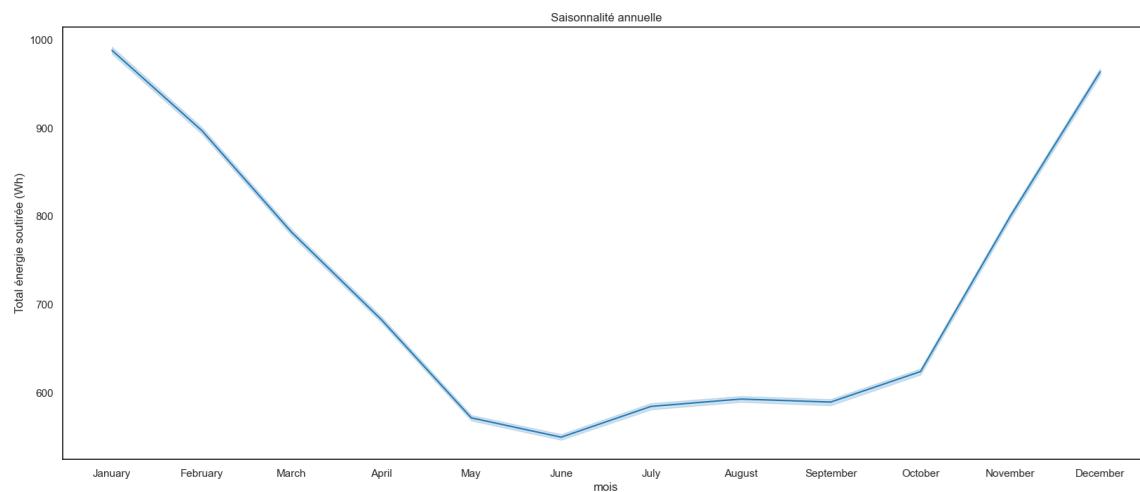


FIGURE 2.4 – Saisonnalité annuelle de la consommation d'électricité en Auvergne-Rhône-Alpes

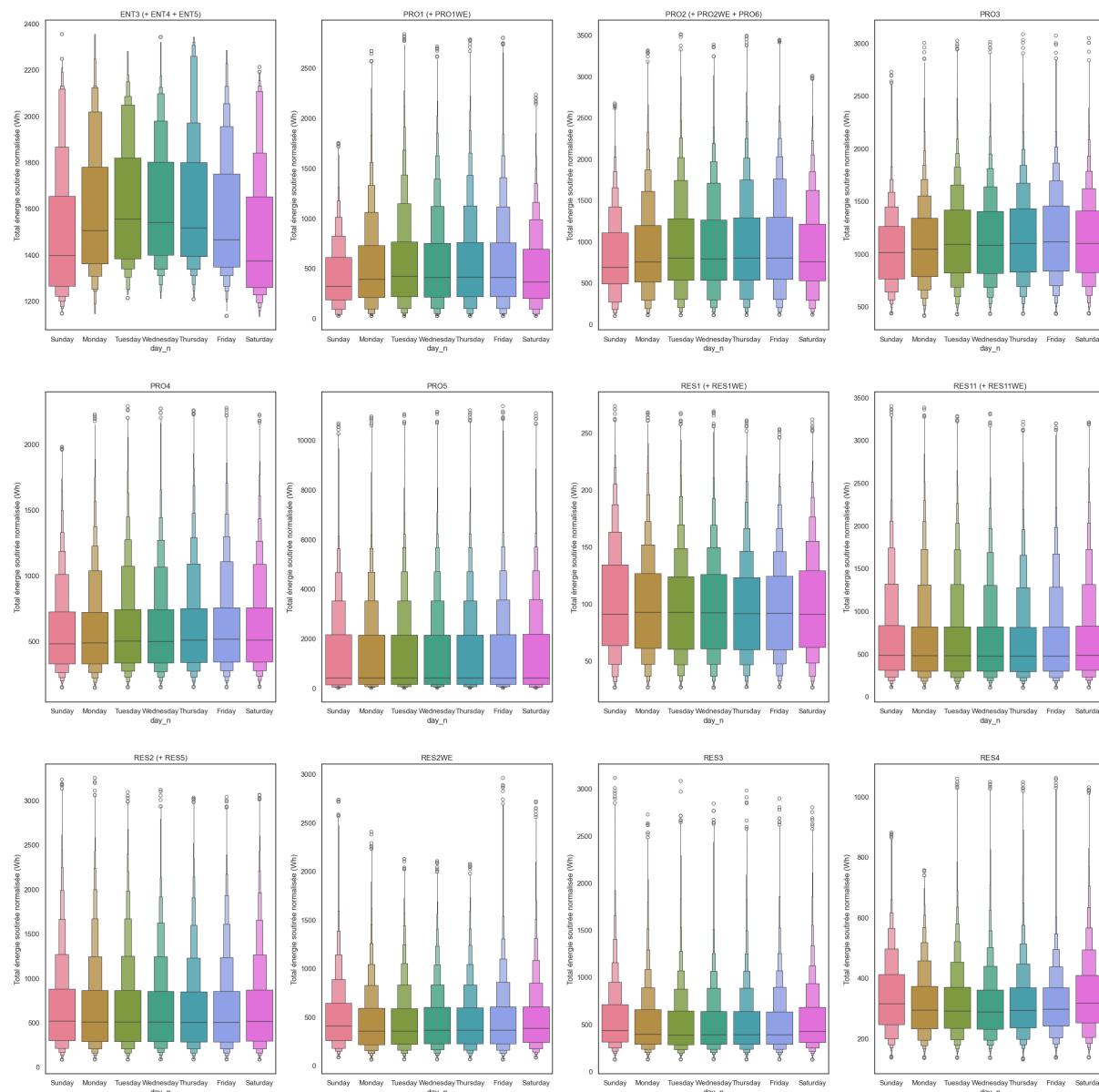


FIGURE 2.5 – Distribution de la consommation d'électricité en Auvergne-Rhône-Alpes selon les jours de la semaine et profile

## 2.2 Données météorologiques

### 2.2.1 Facteurs météorologiques

Plusieurs études [3, 9, 11, 20, 22, 23, 35, 45, 47] montrent que les facteurs météorologiques qui influent sur la consommation d'électricité sont la température, l'humidité, la vitesse du vent et la nébulosité (couverture nuageuse). Ces études s'accordent sur le fait que la température est le facteur le plus influant. Dans [3, 17] les auteurs montrent l'impact de la température et de l'humidité sur la consommation résidentielle. L'étude [11] montre l'impact de la température, de la vitesse du vent et de la nébulosité sur la consommation d'électricité au Québec. Il faut noter que la plupart de ces études sont faite sur des données de pays étrangers (Canada, Espagne, Angleterre, ..). Pour étudier l'impact de ces facteurs sur la consommation d'électricité en France, et établir un nouveau modèle, nous avons chercher à les collecter à partir des données publiés par Météo-France [32]. Ces données météorologiques regroupent les mesures effectuées par toutes les stations de métropole et outre-mer depuis leur ouverture. Les différents facteurs météorologiques disponibles sont décrit dans [30]. Il s'avère que les données sur la nébulosité ne sont pas disponibles, nous avons choisi d'étudier l'impact du rayonnement solaire global [31]. Nous avons, donc, à notre disposition :

- ❖ la température exprimée en °C,
- ❖ la vitesse du vent exprimée en m/s (ou en km/h),
- ❖ l'humidité exprimée en %,
- ❖ le rayonnement solaire global exprimée  $W/m^2$

La températures, la vitesse du vent et l'humidité sont mesurées au pas d'une heure et ceci avec un maillage départementale. Quand au rayonnement solaire global, les données disponibles sont régionales mais avec un pas horaire de 3 heures. Pour fusionner ces données avec notre base de donnés de consommation d'électricité, qui est donnée au pas d'une demi-heure, nous avons réglés les deux problèmes spatio-temporelles.

Pour une région  $i$  et une station météo  $j$  de celle-ci, notons les processus aléatoires, qui correspondent respectivement aux mesures de la température, de l'humidité, de la vitesse du vent par :

$$\left(T_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}, \left(U_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}, \left(V_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}$$

et pour le rayonnement

$$(R_k^i)_{kT_s \in \mathbb{T}_2}$$

où  $T_s = 1800s$ ,  $\mathbb{T}_1 = \{2kT_s, k \in \{0, \dots, L/2\}\}$ ,  $\mathbb{T}_2 = \{6kT_s, k \in \{0, \dots, L/6\}\}$  et  $L = (365 + 366) * 24 = 17544$  (période de 01/01/2023 au 31/12/2024)

### 2.2.2 Analyse

Pour résoudre le problème spatiale, nous devons trouver une solution pour passer de l'échelle d'une station météo à l'échelle d'une région pour les données météorologiques : températures, vitesse du vent et humidité. Pour y remédier à ce problème nous avons étudié :

- ❖ les distributions des processus  $\left(T_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}$ ,  $\left(U_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}$ ,  $\left(V_k^{(i,j)}\right)_{kT_s \in \mathbb{T}_1}$ ,

- ❖ les distributions des moyennes spatiales au niveau du centre géographique d'une région. Pour la température par exemple ce processus est défini par :

$$T_k^{C_i} = \frac{1}{N_{C_i}} \sum_{j=1}^{N_{C_i}} T_k^{(i,j)}$$

où  $C_i$  est le département centre géographique de la région  $i$ , et  $N_{C_i}$  est le nombre de postes de ce département.

- ❖ les distributions des moyennes spatiales au niveau de toute la région. Pour la température par exemple ce processus est défini par :

$$T_k^i = \frac{1}{N_i} \sum_{j=1}^{N_i} T_k^{(i,j)}$$

où  $N_i$  est le nombre de postes météo de la région  $i$

- ❖ la stationnarité et l'ergodicité de ces processus et des processus issues de la différentiation du premier et du second ordre.
- ❖ les corrélations intrarégionales et interrégionales entre ces processus.

### 2.2.2.1 Ergodicité spatiale

L'ergodicité spatiale suppose qu'on peut estimer des propriétés statistiques globales en utilisant uniquement les mesures d'un grand nombre de stations météo dispersées. L'intérêt de cette hypothèse est de savoir si nous pouvons considérer que les processus moyennes, décrites précédemment, sont représentatives de toutes les mesures des stations d'une région. Cette hypothèse n'est pas démontrée théoriquement, mais en pratique, la moyenne spatiale des températures sur un réseau de stations est utilisé comme approximation de leur comportement temporel moyen [6, 10, 29, 37]. Dans notre étude nous admettons cette hypothèse en la soutenant par les corrélations élevées entre les stations d'une même région. La figure 2.6 montre à titre d'exemple la matrice de corrélation entre les processus  $(T_k^{(1,j)})_{kT_s \in \mathbb{T}_1}$  issues des différentes stations météorologiques de la région Auvergne-Rhône-Alpes (numéroté par  $i = 1$ ).

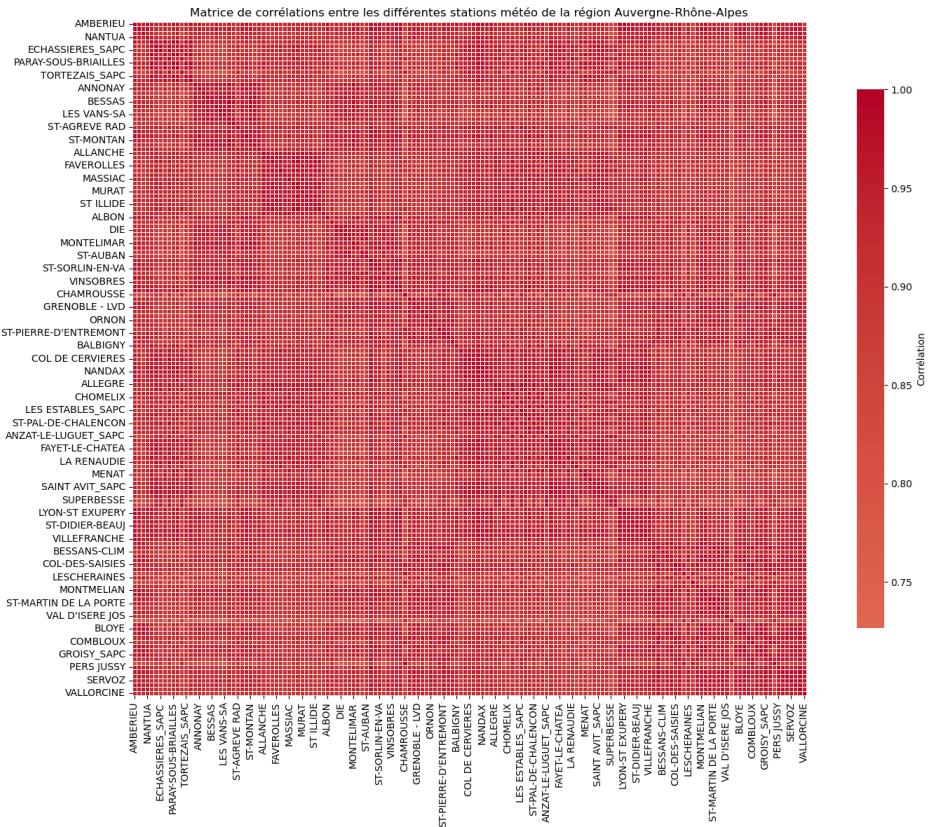


FIGURE 2.6 – la matrice de corrélation entre les stations de la région Auvergne-Rhône-Alpes

En admettant l'hypothèse d'ergodicité spatiale, nous avons calculé et retenu pour chaque facteur (température, humidité et vitesse du vent) des statistiques spatiales de sa distribution au niveau d'une région. Ces statistiques sont : les moyennes, les écart-types, les quartiles, les extrema, le coefficient d'asymétrie (*skewness en anglais*) et le coefficient d'aplatissement (*kurtosis en anglais*). Nous gardons la notation  $kT_s \in \mathbb{T}_1$  pour l'instant d'observation,  $i \in 1, \dots, 12$  pour la région,  $N_i$  pour le nombre de postes de la région  $i$ , et  $j$  pour numérotter les postes (stations) météo.

❖ Moyennes spatiales :

$$T_k^i = \frac{1}{N_i} \sum_{j=1}^{N_i} T_k^{(i,j)}, \quad U_k^i = \frac{1}{N_i} \sum_{j=1}^{N_i} U_k^{(i,j)}, \quad V_k^i = \frac{1}{N_i} \sum_{j=1}^{N_i} V_k^{(i,j)}, \quad R_k^i \quad (2.1)$$

❖ Écart-type :

$$\begin{aligned} \sigma T_k^i &= \sqrt{\frac{1}{N_i-1} \sum_{j=1}^{N_i} (T_k^{(i,j)} - T_k^i)^2}, \\ \sigma U_k^i &= \sqrt{\frac{1}{N_i-1} \sum_{j=1}^{N_i} (U_k^{(i,j)} - U_k^i)^2}, \\ \sigma V_k^i &= \sqrt{\frac{1}{N_i-1} \sum_{j=1}^{N_i} (V_k^{(i,j)} - V_k^i)^2} \end{aligned} \quad (2.2)$$

❖ Extrema :

$$\begin{aligned} \min T_k^i &= \min_{j \in \{1, \dots, N_i\}} (T_k^{(i,j)}), \quad \min U_k^i = \min_{j \in \{1, \dots, N_i\}} (U_k^{(i,j)}), \quad \min V_k^i = \min_{j \in \{1, \dots, N_i\}} (V_k^{(i,j)}) \\ \max T_k^i &= \max_{j \in \{1, \dots, N_i\}} (T_k^{(i,j)}), \quad \max U_k^i = \max_{j \in \{1, \dots, N_i\}} (U_k^{(i,j)}), \quad \max V_k^i = \max_{j \in \{1, \dots, N_i\}} (V_k^{(i,j)}) \end{aligned} \quad (2.3)$$

❖ Le coefficient d'asymétrie évalue le défaut de symétrie de la distribution. Il est nul pour une distribution symétrique. Il est positif pour une distribution "étalée à droite".

$$\begin{aligned} ST_k^i &= \frac{1}{(\sigma T_k^i)^3 N_i} \sum_{j=1}^{N_i} \left( T_k^{(i,j)} - T_k^i \right)^3 \\ SU_k^i &= \frac{1}{(\sigma U_k^i)^3 N_i} \sum_{j=1}^{N_i} \left( U_k^{(i,j)} - U_k^i \right)^3 \\ SV_k^i &= \frac{1}{(\sigma V_k^i)^3 N_i} \sum_{j=1}^{N_i} \left( V_k^{(i,j)} - V_k^i \right)^3 \end{aligned} \quad (2.4)$$

- ❖ Le coefficient d'aplatissement évalue la dispersion des valeurs "extrêmes" par référence à la loi normale. Il est nul pour une distribution normale, négatif pour une distribution moins "aplatie" qu'une distribution normale et positif sur une distribution plus aplatie qu'une distribution normale.

$$\begin{aligned} KT_k^i &= \frac{1}{(\sigma T_k^i)^4 N_i} \sum_{j=1}^{N_i} \left( T_k^{(i,j)} - T_k^i \right)^4 - 3 \\ KU_k^i &= \frac{1}{(\sigma U_k^i)^4 N_i} \sum_{j=1}^{N_i} \left( U_k^{(i,j)} - U_k^i \right)^4 - 3 \\ KV_k^i &= \frac{1}{(\sigma V_k^i)^4 N_i} \sum_{j=1}^{N_i} \left( V_k^{(i,j)} - V_k^i \right)^4 - 3 \end{aligned} \quad (2.5)$$

- ❖ Notons  $F_{T_k^i}$ ,  $F_{U_k^i}$  et  $F_{V_k^i}$  les fonction de répartition spatiales de la température, de l'humidité et la vitesse du vent à l'instant  $kT_s$ . Les quartiles (premier quartile, médiane, et troisième quartile) sont données par l'équation 2.6.

$$\begin{aligned} Q_{0.25} T_k^i &= \inf \left\{ x : F_{T_k^i}(x) \geq 0.25 \right\}, \quad Q_{0.5} T_k^i = \inf \left\{ x : F_{T_k^i}(x) \geq 0.5 \right\}, \quad Q_{0.75} T_k^i = \inf \left\{ x : F_{T_k^i}(x) \geq 0.75 \right\} \\ Q_{0.25} U_k^i &= \inf \left\{ x : F_{U_k^i}(x) \geq 0.25 \right\}, \quad Q_{0.5} U_k^i = \inf \left\{ x : F_{U_k^i}(x) \geq 0.5 \right\}, \quad Q_{0.75} U_k^i = \inf \left\{ x : F_{U_k^i}(x) \geq 0.75 \right\} \\ Q_{0.25} V_k^i &= \inf \left\{ x : F_{V_k^i}(x) \geq 0.25 \right\}, \quad Q_{0.5} V_k^i = \inf \left\{ x : F_{V_k^i}(x) \geq 0.5 \right\}, \quad Q_{0.75} V_k^i = \inf \left\{ x : F_{V_k^i}(x) \geq 0.75 \right\} \end{aligned} \quad (2.6)$$

### 2.2.2.2 Sur-échantillonnage temporelle

Pour l'échantillonnage temporelle, nous avons gardé le pas  $T_s = 30min = 1800s$  de référence. Puis nous avons sur-échantillonné par interpolation linéaire les données météorologiques. On considère  $\mathbb{T} = \{kT_s, k \in \{0, \dots, L\}\}$  avec  $L = (365 + 366) * 48 = 35088$ .

Par exemple pour la température les valeurs d'indices impaires sont calculées par l'équation 2.7 :

$$T_{2k+1}^i = \frac{T_{2k}^i + T_{2k+2}^i}{2}, \quad k \in \{0, \dots, [L/2]\} \quad (2.7)$$

Le calcul est le même pour les deux autres variables : humidité et vitesse du vent. Pour le rayonnement la formule est différente :

$$R_{6k+j}^i = T_{6k} + j \frac{T_{6k+6}^i - T_{6k}^i}{2}, \quad j \in \{1, \dots, 5\}, \quad k \in \{0, \dots, [L/6]\} \quad (2.8)$$

## 2.3 Fusion des bases de données

Après avoir analysée, nettoyé, traité et construit les deux bases de données, i.e. la base de données de consommation d'électricité et la base de données météorologiques, nous avons fusionnées les deux. Nous avons également ajouté des champs nécessaires à notre analyse : jour de la semaine, mois, année, heure et minutes. La figure 2.7 montre les différents colonnes de la base de données construites. Dans

la section suivante nous allons étudier l'influences de chaque facteur sur la consommation d'électricité. A noter que les colonnes de nom débutant par *FF* sont les statistiques de la vitesse du vent que nous avons noté *V* dans nos équations.

#	Column	Dtype	
0	Région	object	29 U_moyenne
1	Code région	float64	30 U_STD
2	Profil	object	31 U_min
3	Plage de puissance souscrite	object	32 U_q25
4	Nb points soutirage	float64	33 U_q50
5	Total énergie soutirée (Wh)	float64	34 U_q75
12	Jour max du mois (0/1)	float64	35 U_max
13	Semaine max du mois (0/1)	float64	36 date
14	AAAAMJJHH	int64	37 year
15	T_moyenne	float64	38 month
16	T_STD	float64	39 month_n
17	T_min	float64	40 day
18	T_q25	float64	41 day_n
19	T_q50	float64	42 h
20	T_q75	float64	43 mn
21	T_max	float64	44 s
22	FF_moyenne	float64	45 Rayonnement solaire global (W/m2)
23	FF_STD	float64	
24	FF_min	float64	
25	FF_q25	float64	
26	FF_q50	float64	
27	FF_q75	float64	
28	FF_max	float64	

FIGURE 2.7 – Les colonnes de la base de données finale

## 2.4 Influence des facteurs météorologiques

La figure 2.8 présente les effets des facteurs météorologiques sur la consommation d'électricité totale pour la région Auvergne-Rhône-Alpes. La figure 2.9présente les effets de ces facteurs sur la consommation d'électricité pour un profil et une plage puissance spécifiques dans la région Auvergne-Rhône-Alpes. Les valeurs en ordonnée correspondent à la consommation en Wh et en abscisse aux différents facteurs. Ces figures confirment le fait que la température est le facteur le plus influant. Elles montrent aussi que certaines statistiques, comme l'écart type spatiale des facteurs n'ont pas beaucoup d'intérêt. La vitesse du vent n'a pas non plus un grand impacte sur la consommation.

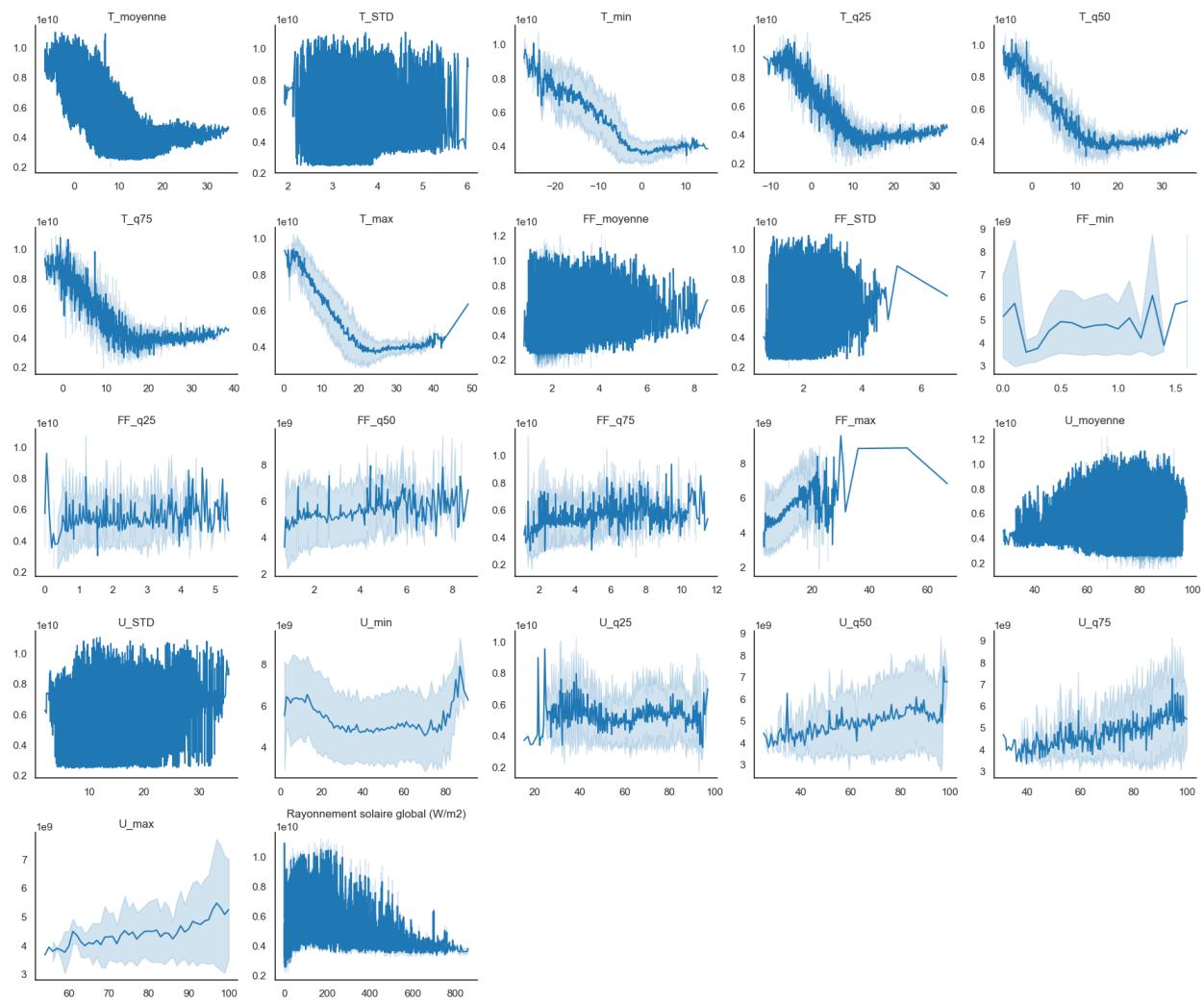


FIGURE 2.8 – Corrélation des facteurs météorologiques avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes

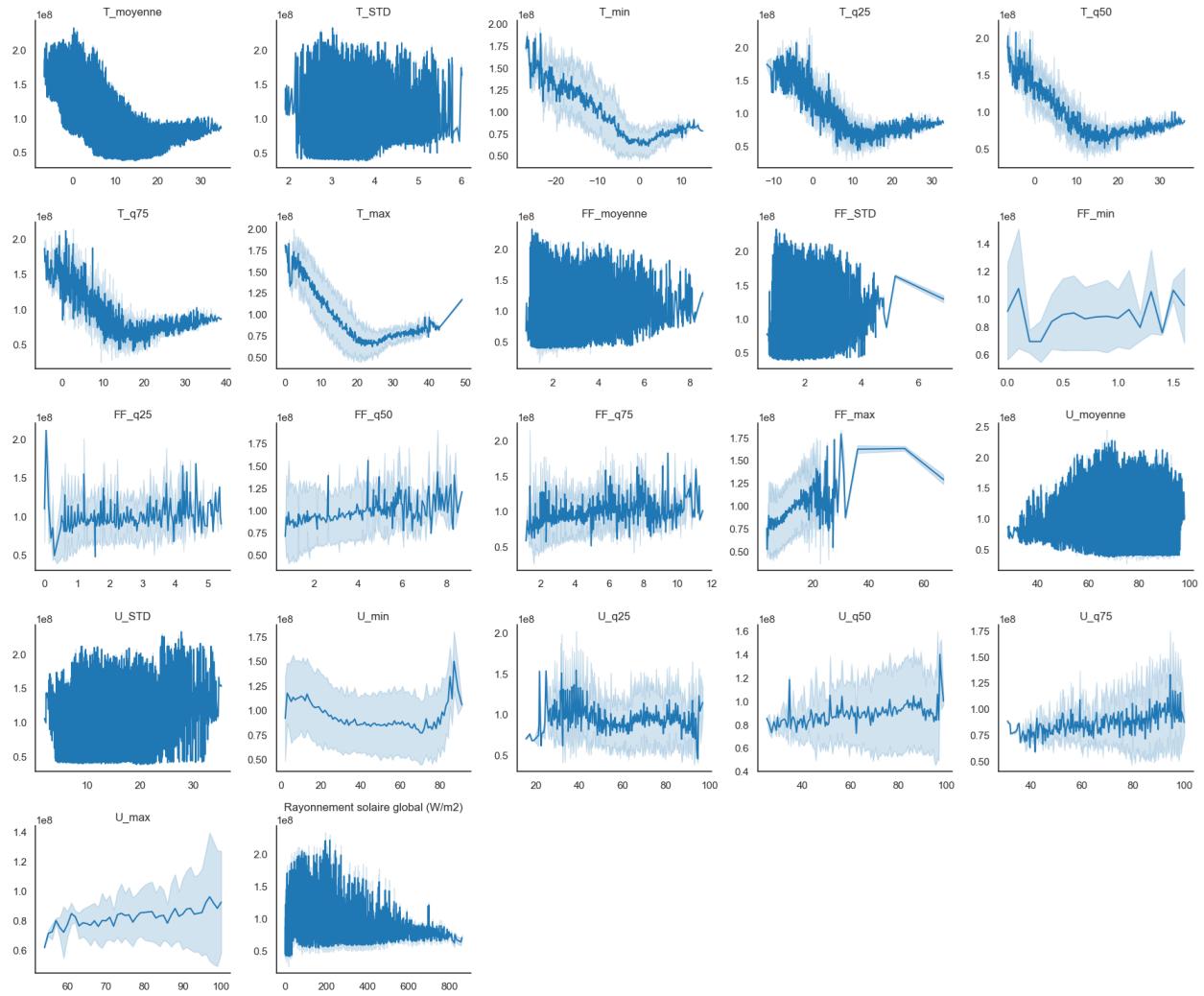


FIGURE 2.9 – Corrélation des facteurs météorologiques avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes pour le profil RES11 (+ RES11WE) et sa plage de puissance P3 : ]6-9] kVA

#### 2.4.1 Influence de la température

La figure 2.10 montre la corrélation entre la température moyenne et la consommation d'électricité observée entre 2023 et 2024. Elle présente une segmentation linéaire par morceaux des effets de la température moyenne et la température maximale. Ces courbes confirment que la consommation d'électricité est fortement liée au facteur température, et que cette relation est non linéaire. Les principaux points de rupture détectés sont 11.27°C et 15.54°C pour les températures moyennes et 17.3°C et 22.3°C pour températures maximales. Quand les températures maximales sont inférieures à 17.3°C, la pente de la droite de régression est forte, ce qui s'explique par l'utilisation de l'électricité pour se chauffer. De même, au-dessus de 22.3°C, la consommation d'électricité commence à augmenter, probablement à cause de l'utilisation de la climatisation. Ces points de ruptures sont très utiles pour l'élaboration d'une spécification linéaire par morceaux appropriée [8, 11].

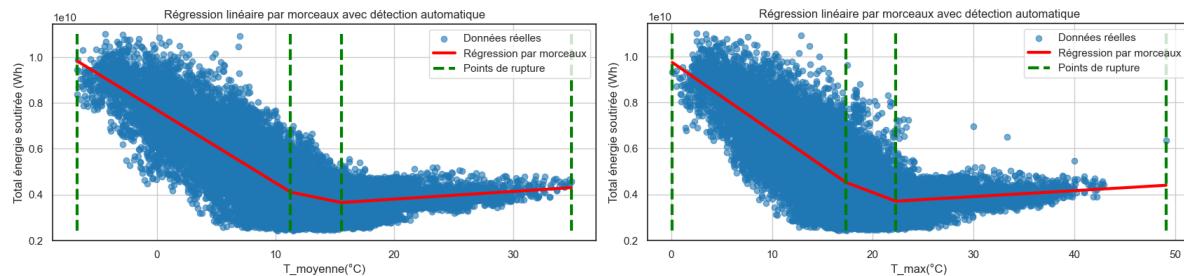


FIGURE 2.10 – Corrélation de la température avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes

Nous avons raffiné notre analyse en observant les relations entre les composantes respectives des deux séries temporelles représentant la température et la consommation d'électricité. La figure 2.10 montre la corrélation entre la tendance de la température et la tendance de la consommation d'électricité observées entre 2023 et 2024 pour la région Auvergne-Rhône-Alpes.

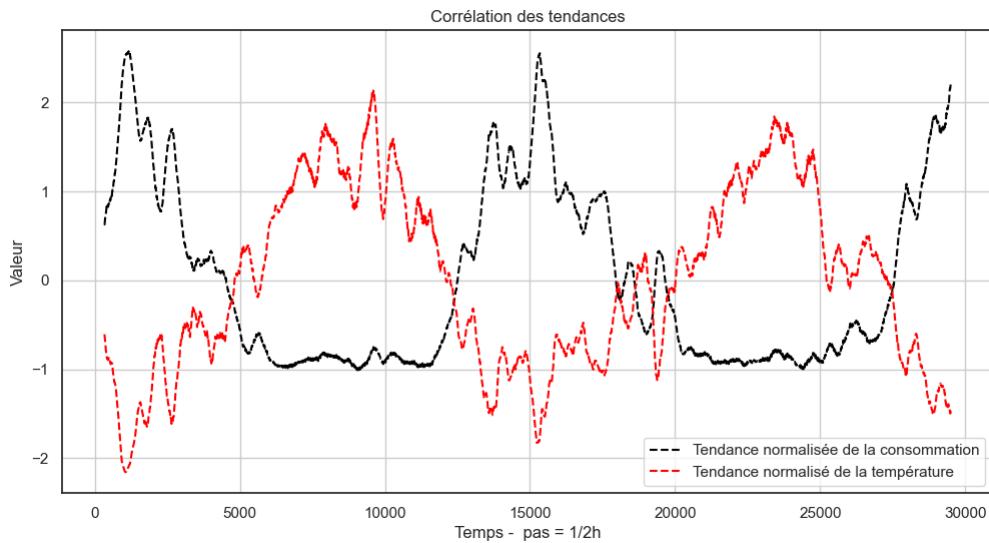


FIGURE 2.11 – Corrélation entre la tendance de la température et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

Ces tendances sont obtenues par une moyenne glissante sur une semaine puis centrées et réduites. Nous constatons que la tendance de la consommation d'électricité est corrélée avec l'inverse de la tendance de la température. La figure 2.12 et le tableau 2.3 montrent cette corrélation après inversion et standardisation.

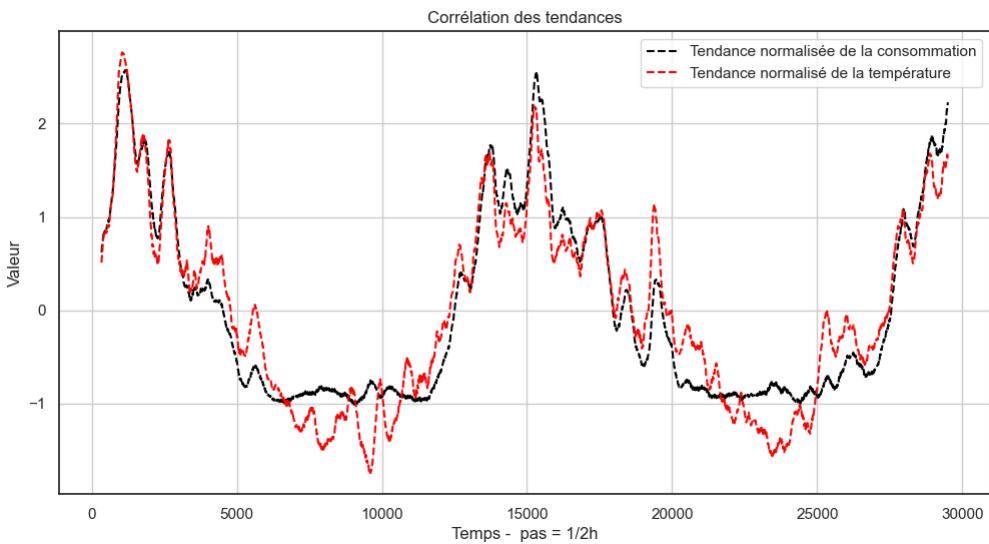


FIGURE 2.12 – Corrélation entre la tendance de la température transformée et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

	Tendance de la consommation	Inverse de la tendance de la température
Tendance de la consommation	1.00	0.91
Inverse de la tendance de la température	0.91	1.00

TABLE 2.3 – Coefficients de corrélation entre la tendance de la température et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

Pour éviter les dévissons par zéros nous avons introduit une translation de la tendance de la température. La transformation effectuée est décrite par l'équation 2.9.

$$f : \begin{array}{ccc} \mathbb{R} & \xrightarrow{g} & \mathbb{R} \\ T & \rightarrow g(T) = \frac{1}{T+C} & \rightarrow f(T) = \frac{g(T) - \mu_{g(T)}}{\sigma_{g(T)}} \end{array} \quad (2.9)$$

où  $\mu_X$  et  $\sigma_X$  désignent la moyenne et l'écart-type de la variable  $X$ .

Les composantes saisonnières sont décorrélées et les composantes résiduelles sont faiblement corrélées comme le montre le tableau 2.4.

	Résidus de la consommation	Résidus de la température
Résidus de la consommation	1.00	0.39
Résidus de la température	0.39	1.0
	Saisonnalité de la consommation	Saisonnalité de la température
Saisonnalité de la consommation	1.00	-0.03
Saisonnalité de la température	-0.03	1.00

TABLE 2.4 – Corrélation des composantes saisonnières et résiduelles de la température avec les composantes respectives de la consommation d'électricité dans la région Auvergne-Rhône-Alpe

## 2.4.2 Influence de l'humidité

La figure 2.13 montre la corrélation entre l'humidité et la consommation d'électricité observée entre 2023 et 2024. Elle présente une segmentation linéaire par morceaux des effets de l'humidité moyenne et maximale. Ces courbes révèlent une légère dépendance de la consommation d'électricité au facteur d'humidité. Les principaux points de rupture détectés sont 93% et 96% pour l'humidité moyenne et 93% et 95% pour l'humidité maximale. Quand l'humidité dépasse le seuil de 95%, la consommation tende vers une grande augmentation.

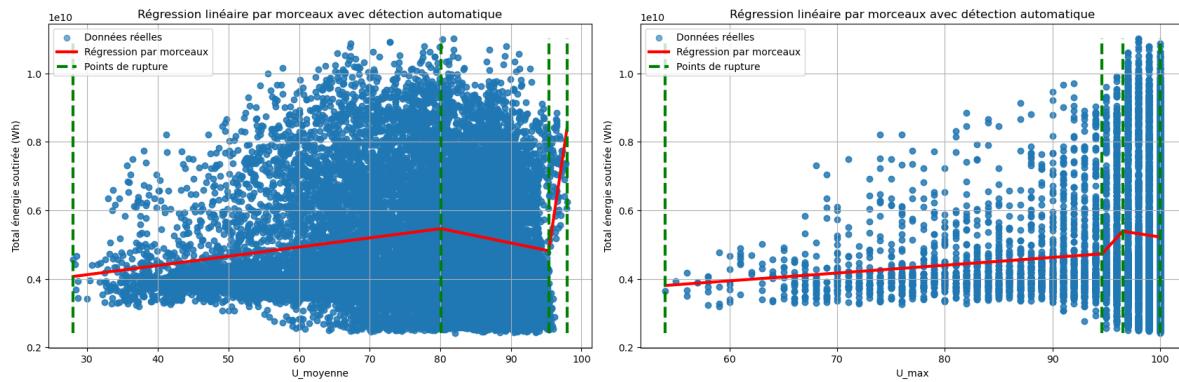


FIGURE 2.13 – Corrélation de l'humidité avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes pour 2023-2024

La figure 2.14 montre les tendances de l'humidité et de la consommation d'électricité observée entre 2023 et 2024 pour la région Auvergne-Rhône-Alpes. Ces tendances sont obtenues par une moyenne glissante sur une semaine puis centrées et réduites.

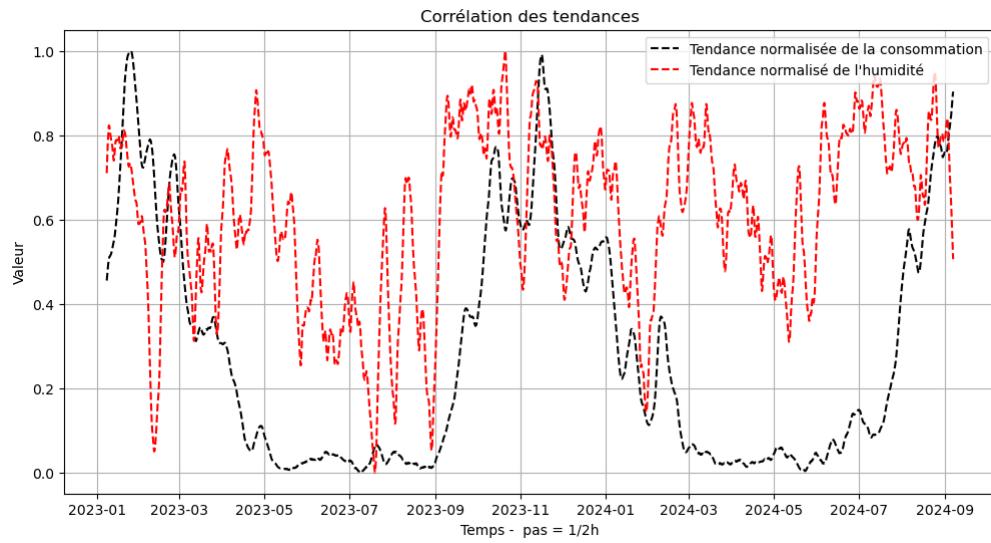


FIGURE 2.14 – Corrélation entre la tendance de l'humidité et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

Nous constatons que la tendance de la consommation d'électricité est corrélée avec la tendance de l'humidité pendant la période hivernale, et elle est corrélée avec l'inverse de la tendance de l'humidité pendant la période estivale. les tableau 2.3 donne les coefficients de corrélation, calculés par la méthode

de Pearson. Le premier tableau donne ces coefficients pour toutes les saisons et le deuxième en séparant les saisons.

(a)	Tendance de la consommation	Tendance de l'humidité
Tendance de la consommation	1.00	0.19
Tendance de l'humidité	0.19	1.00
(b)	Tendance de la consommation	Tendance de l'humidité
Tendance de la consommation	1.00	0.29
tendance de l'humidité	0.29	1.00

TABLE 2.5 – Coefficients de Corrélation entre la tendance de l'humidité et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes. (a) Pour toutes les saisons, (b) En séparant les saisons

La composante saisonnière de la consommation d'électricité est décorrélée de la composante saisonnière de l'humidité, avec un coefficient de corrélation de 0.032. Par contre leurs composantes résiduelles sont faiblement corrélées avec un coefficient de corrélation de -0.34.

#### 2.4.3 Influence du rayonnement solaire

La figure 2.15 montre la relation entre le rayonnement solaire global et la consommation d'électricité observée entre 2023 et 2024, avec une segmentation linéaire par morceaux des effets. Ces courbes confirment que la consommation d'électricité est fortement liée au rayonnement solaire global. Cette relation s'explique par le besoin d'éclairage quand le rayonnement solaire est faible ou nul. Le principal point de rupture détecté est  $63W/m^2$ . Quand le rayonnement est inférieur à cette valeur l'utilisation d'électricité pour l'éclairage augmente la consommation, et plus le rayonnement augmente plus la consommation diminue. Ce lien entre le rayonnement solaire et la consommation d'électricité, pourtant évident, n'est pas étudié dans la littérature à notre connaissance contrairement aux autres facteurs.

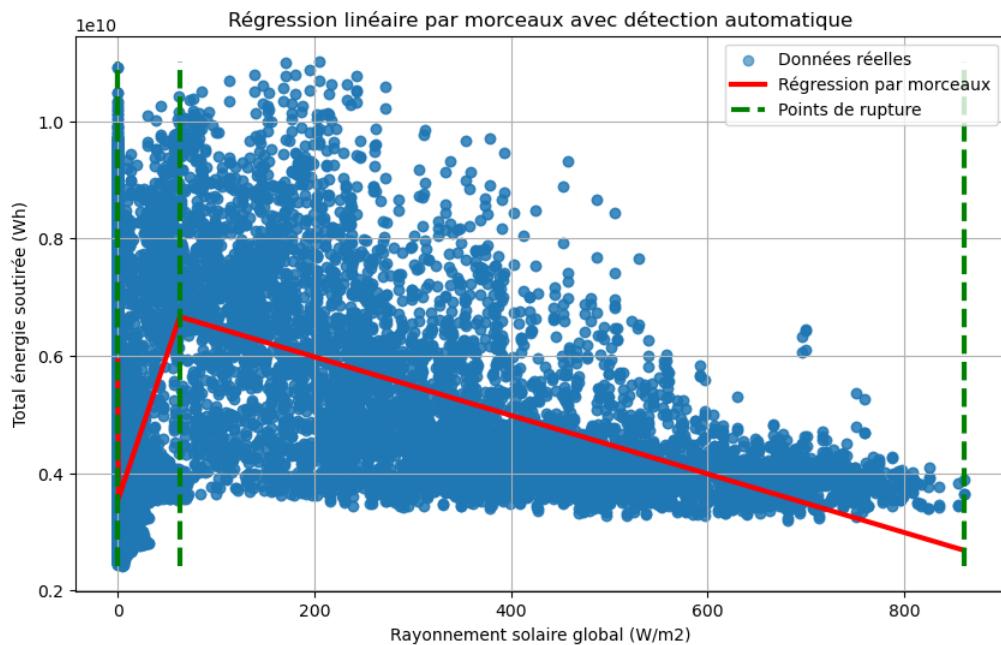


FIGURE 2.15 – Corrélation de la température avec la consommation d'électricité dans la région Auvergne-Rhône-Alpes

La figure 2.16 montre la relation entre la tendance du rayonnement solaire global et la tendance de la consommation d'électricité observée entre 2023 et 2024 pour la région Auvergne-Rhône-Alpes. Nous constatons que la tendance de la consommation d'électricité est fortement corrélée avec l'inverse de la tendance du rayonnement.

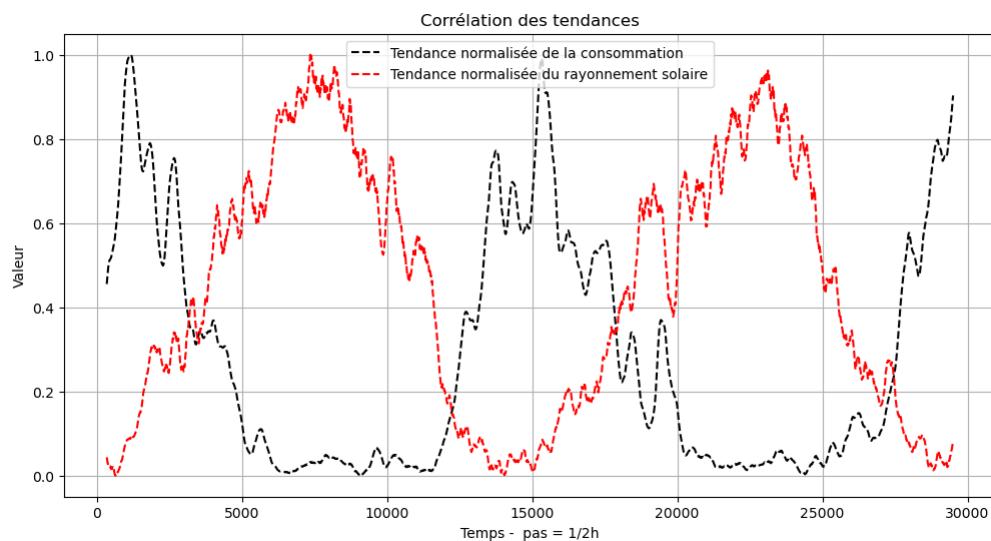


FIGURE 2.16 – Corrélation entre la tendance du rayonnement solaire global et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

La figure 2.17 montre cette corrélation après inversion et standardisation.

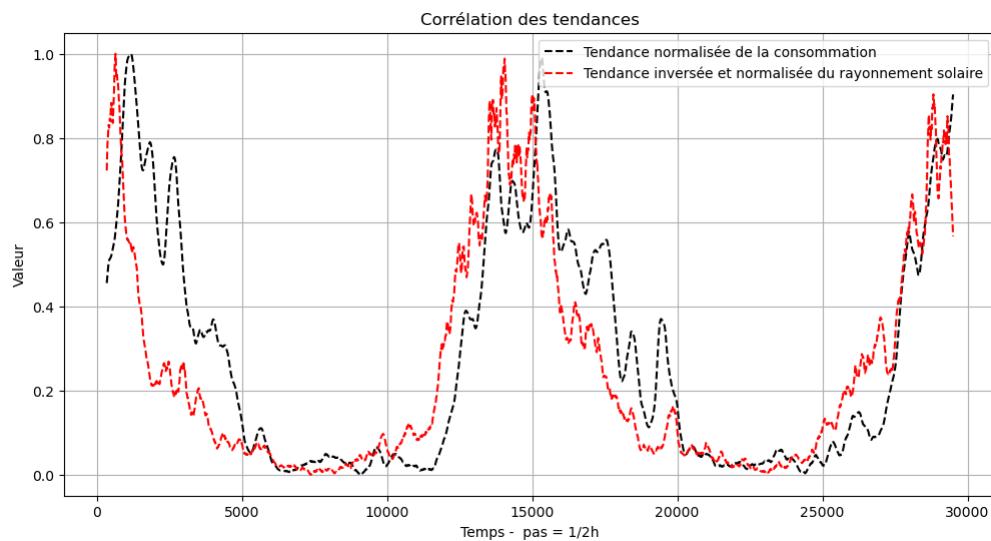


FIGURE 2.17 – Corrélation entre la tendance du rayonnement solaire global et la tendance de la consommation d'électricité dans la région Auvergne-Rhône-Alpes

## 2.5 Conclusion

Dans ce chapitre nous avons présenté notre démarche pour collecter, analyser, fusionner et traiter les bases de données dont nous avions besoin pour mener ce projet. Nous avons étudié les relations entre les variables exogènes et la variable cible. Nous avons ainsi démontré que les composantes saisonnières de la consommation d'électricité sont indépendantes des variables exogènes. Nous avons aussi formalisé la relation non linéaire qui existe entre la tendance (respectivement le résidu) de la consommation d'électricité et les tendances (respectivement les résidus) des variables météorologiques.

# Chapitre 3

## Modèle proposé

Dans ce chapitre nous décrivons l'approche proposée pour effectuer la prévision de la consommation d'électricité. Dans le premier chapitre nous avons formalisé, dans le cas général, la problématique de prévision pour une série multivariée. Nous transposons celle-ci sur notre série, définie et décrite dans le chapitre 2. Afin de comparer notre approche avec d'autres travaux, nous dressons ici une synthèse de la revue de littérature générale, qui ne vaut pas être exhaustive. Nous renvoyons le lecteur intéressé aux récents travaux [33, 38, 43] qui ont présentés des méta-analyses des articles publiés sur le sujet. A la suite de cette revue de littérature nous présentons notre modélisation en détails et nous terminons par une analyse de ses performances. Nous tenons à noter que dans notre cas, nous nous intéressons à la prévision de la consommation réelle des clients Enedis et non pas à la prévision de la charge électrique.

### 3.1 Formalisation de notre problème

#### 3.1.1 Notation

Avant de décrire formellement le problème posé nous rappelons les notations utilisées dans ce chapitre :

- ❖  $\mathbb{T} = \{kT_s, k \in \{0, \dots, L\}\}$  l'ensemble d'instants d'observation disponibles dans notre base. Cette ensemble couvre la période du 01/01/2023 au 31/12/2024. On rappelle que :
  - ★  $T_s = 30 \text{ min}$  est la période d'échantillonnage
  - ★  $L = (365 + 366) * 48 = 35088$  est le nombre d'observations disponibles
- ❖  $\mathbb{T}_m = \{kT_s, k \in \{m, \dots, \tau\}\}$  l'ensemble d'instants d'observation pour l'entraînement qui débute à  $mT_s$ , où  $m \in \{0, \dots, L/2 - 1\}$  est choisi aléatoirement à chaque entraînement du modèle, et se termine à  $\tau = m - 1 + L/2$ .
- ❖  $\mathbb{T}_{\text{test}} = \{kT_s, k \in \{\tau + 1, \tau + h\}\}$  l'ensemble d'instants d'observation pour le test des performances du modèle à l'horizon  $h$ .
- ❖  $\mathbb{T}_1 = \mathbb{T}_m \cup \mathbb{T}_{\text{test}}$  l'ensemble d'instants d'observation pour l'entraînement et le test.
- ❖  $\mathcal{R} = \{11, 24, 27, 28, 32, 44, 52, 53, 75, 76, 93, 84\}$  désigne l'ensemble des régions métropolitaines encodées par leur code officiel :

Région	code
Auvergne-Rhône-Alpes,	84
Bourgogne-Franche-Comté	27
Bretagne	53
Grand-Est	44
Centre-Val-de-Loire	24
Hauts-de-France	32
Normandie	28
Nouvelle-Aquitaine	75
Occitanie	76
Pays-de-la-Loire	52
Provence-Alpes-Côte-d'Azur	93
Île-de-France	11

TABLE 3.1 – Liste des régions et leurs codes

- ❖  $\mathcal{C}$  désigne l'ensemble de configurations associant profil Enedis et plage de puissance souscrite. Cet ensemble est décrite dans le tableau 2.2.
- ❖  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$  est la série temporelle de la consommation d'électricité (en Wh) de la configuration profil - plage de puissance, notée  $q \in \mathcal{C}$ , dans la région  $r \in \mathcal{R}$ .
- ❖  $(N_k^{(r,q)})_{kT_s \in \mathbb{T}}$  est la série temporelle qui représente le nombre de point de soutirage de la configuration profil - plage de puissance, notée  $q \in \mathcal{C}$ , dans la région  $r \in \mathcal{R}$ .
- ❖  $(\bar{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}} = \left(\frac{Y_k^q}{N_k}\right)_{kT_s \in \mathbb{T}}$  est la série temporelle qui qui représente la consommation moyenne des utilisateurs de la configuration profil - plage de puissance  $q \in \mathcal{C}$  dans la région  $r \in \mathcal{R}$ .
- ❖  $(T_k^{(r)})_{kT_s \in \mathbb{T}}$  est la série temporelle qui correspond à la température moyenne de la régions  $r$ .
- ❖  $(U_k^{(r)})_{kT_s \in \mathbb{T}}$  est la série temporelle qui correspond à l'humidité moyenne de la régions  $r$ .
- ❖  $(R_k^{(r)})_{kT_s \in \mathbb{T}}$  est la série temporelle qui correspond au rayonnement solaire global de la régions  $r$ .
- ❖  $\mathbb{S}^{(r,q)}(N, \rho, w)$  le spectrogramme de la série  $(Y_k^{(r,q)})_{r \in \mathcal{R}, q \in \mathcal{C}, kT_s \in \mathbb{T}}$  obtenu en utilisant la fenêtre  $w$  de longueur  $N$  et avec un recouvrement  $\rho$ .
- ❖  $\mathbb{P}_\delta^{(r,q)}$  l'ensemble des périodes détectés par analyse spétrale de la série  $(Y_k^{(r,q)})_{r \in \mathcal{R}, q \in \mathcal{C}, kT_s \in \mathbb{T}}$  avec un seuil  $\delta$ .

### 3.1.2 Formalisation

Pour toute configuration  $q \in \mathcal{Q}$  (profile - plage de puissance ) dans une région  $r \in \mathcal{R}$ , nous cherchons un modèle  $\mathcal{M}^{(q,r)}$  qui permet d'estimer les valeurs futures  $(Y_k)_{kT_s \in \mathbb{T}_{\text{test}}}$ , pour l'horizon  $h$ , en fonction de  $\mathcal{F}_\tau$  l'ensemble d'information disponible jusqu'au l'instant  $\tau$  sur les valeurs passées de la série cible et sur les variables exogènes.

$$\mathcal{F}_\tau = \left\{ \left( Y_k^{(r,q)} \right)_{kT_s \in \mathbb{T}_m}, \left( T_k^{(r)} \right)_{kT_s \in \mathbb{T}_1}, \left( U_k^{(r)} \right)_{kT_s \in \mathbb{T}_1}, \left( R_k^{(r)} \right)_{kT_s \in \mathbb{T}_1} \right\}$$

$$\left( \hat{Y}_{\tau+1}, \hat{Y}_{\tau+2}, \dots, \hat{Y}_{\tau+h} \right) = \mathcal{M}^{(q,r)} (\mathcal{F}_\tau) \quad (3.1)$$

### 3.1.3 Hypothèses validées

Dans les précédents chapitres nous avons validé statistiquement les hypothèses suivantes :

- ❖ les séries temporelles de la consommation d'électricité sont des séries multi-saisonnieres,
- ❖ l'analyse temps-fréquence permet de déterminer les saisonnalités dominantes de celles-ci,
- ❖ les séries temporelles de la consommation d'électricité sont non stationnaires et ceci est dû à leurs tendances,
- ❖ les composantes saisonnières et résiduelles sont stationnaires,
- ❖ Les composantes saisonnières sont indépendantes des variables exogènes et dépendent uniquement de la configuration profil - plage de puissance souscrite,
- ❖ il existe une relation non linéaire entre la tendance de la consommation d'électricité et les tendances des variables exogènes,
- ❖ la tendance de la consommation d'électricité est fortement corrélée aux séries obtenues par translation et inversion des tendances des variables exogènes,
- ❖ il existe une relation non linéaire entre le résidus de la consommation d'électricité et les résidus des variables exogènes,
- ❖ le résidus de la consommation d'électricité est fortement corrélée aux séries obtenues par transformation des résidus des variables exogènes,

## 3.2 Modèles proposés dans la littérature

Avant de faire la synthèse de notre revue de littérature, nous tenons à souligner que la majorité des études publiées traitent de la prévision de la charge électrique qui consiste à prévoir la quantité d'électricité demandée à un moment donné. Alors que dans notre étude nous intéressons à la prévision de la consommation d'électricité, qui consiste à prévoir la quantité d'électricité qui sera consommée sur une période par un agrégat d'utilisateurs. Malgré cette différence de point de vue (production vs consommation) le paradigme reste le même.

Notre synthèse s'appuie sur les dernières revues de littérature publiés [33, 38, 43]. La première [33] est une méta-analyse de 240 articles publiés sur la prévision de la charge électrique à court terme (*STLF pour Short-Term Load Forecasting*) entre 2000 et 2019. Selon cette étude, 87% des études traitent ce problème de prévision comme un problème de régression alors que 13% le traitent comme un problème de classification. L'utilisation de réseaux de neurones artificiels (ANN) représente 21 % des modèles proposés dans la littérature, et en combinaison avec d'autres modèles ce taux passe à 45 %. Ils sont suivis des modèles de série temporelle, incluant les modèles autorégressifs à moyenne mobile (ARIMA), qui représentent 10% des publications analysées. Les modèles de régression représentent 9 % des modèles proposés, alors que les modèles basés sur la logique floue représentent 4% et la régression à vecteurs de support (SVR) est utilisée dans 4 % des publications. L'étude cite également d'autre modèle comme les modèles d'optimisation par essaims particulaires (particle swarm optimization ou PSO), les vecteurs autorégressifs bayésiens (Bayesian vector autoregression ou BVAR), les modèles de décomposition, les filtres de Kalman, les cartes autoadaptatives ou cartes de Kohonen (self organizing maps ou SOM), Grey Prediction, les algorithmes de colonies de fourmis (ant colony optimization ou ACO), et les algorithmes génétiques (GA). Dans l'ensemble, les techniques d'apprentissage automatique ont été utilisées dans environ 43 % des travaux, les techniques hybrides employant plusieurs

modèles dans 44 %, et les méthodes statistiques dans environ 13 %. Les métriques les plus utilisées sont dans l'ordre :

- ❖ MAPE (Mean Absolute Percentage Error) utilisée dans 62 % des cas,
- ❖ MSE (Mean Squared Error – MSE ) utilisée dans 34% des cas,
- ❖ RMSE(Root Mean Squared Error) utilisée dans 26% des cas,
- ❖ MAE (Mean Absolute Error) utilisée dans 25 % des cas,
- ❖ et l'APE (Average Percentage Error) utilisée dans 6% des cas.

L'étude souligne que pour une prévision à court terme, d'horizon d'un jour et avec un pas d'une heure, les meilleurs score sont un MAPE entre 1% et 2%, et dans quelque rare cas ce score passe sous la barre de 1%.

L'étude [38] analyse 38 travaux de recherche, publiés entre 2011 et 2022, sur la prévision de la charge d'électricité à court terme (STLF) appliquées au secteur résidentiel. Elle affirme que dans le monde, la consommation d'électricité des bâtiments résidentiels représente 39 % de la consommation mondiale d'électricité , et qu'en Europe, cette proportion passe à 40 % . Le nombre d'articles publiés au cours des dernières décennies sur le STLF pour le secteur résidentiel témoigne de l'intérêt croissant de la communauté scientifique pour ce sujet. Les approches présentées dans ces 38 articles se fondent sur des méthodes d'apprentissage automatique (ANN, LSTM, SVR, ANN, CFNN, MLP, CNN-LSTM), des méthodes statistiques (ARIMA, SARIMA, Optimisation bayésienne , Processus Gaussiens (GP)) ou sur les combinaisons des deux. L'apprentissage automatique est utilisée dans 57.9% des articles analysés, les méthodes statistiques sont utilisées dans 18.4 % et la combinaison des deux dans 23.7%. Les auteurs soulèvent le fait que les performances des méthodes statistiques sont limités à cause de la non-linéarité du comportement de la consommation d'électricité dans les bâtiments résidentiels et que les algorithmes d'apprentissage automatique, tels que la régression du vecteur de support (SVR) et les réseaux neuronaux artificiels (ANN), peuvent remédier à cette non-linéarité. Cependant pour la régression du vecteur de support, une sélection inappropriée de la fonction noyau ou des hyperparamètres dégrade les performances de la prévision. Quand aux ANN le problème de convergence vers des optimums locaux entraîne une des grands écarts entre les valeurs prédites et les valeurs réelles. Les auteurs soulignent que des améliorations significatives des performances ont été constatées dans les études utilisant l'apprentissage profond. Il affirme que les solutions les plus performantes et les plus réussies sont les modèles qui intègrent des réseaux neuronaux à convolution (CNN) et des réseaux LSTM. Les approches basées sur des modèles CNN ont de meilleures performance pour des prévisions à court terme avec un horizon d'une heure à un jour, tandis que les architectures LSTM permettent de faire des prévisions à court terme avec des horizons plus longs allant de 7 jours à 1 mois, grâce à leur capacité de capturer les dépendances à long terme dans les séries temporelles. Les approches hybrides combinent les avantages d'une ou plusieurs techniques pour réduire les erreurs de prévision. Ces modèles surmontent les inconvénients présentés par les structures non hybrides dans la recherche d'une meilleure précision de prévision en augmentant la robustesse et l'efficacité d'un modèle hybride. La plupart des modèles hybrides présentés combinent des modèles linéaires et non linéaires pour une prévision efficace. Les auteurs relèvent l'intérêt croissant pour ces approches hybrides, combinant modèles statistiques et réseaux de neurones pour pallier leurs faiblesses respectives. L'étude insiste sur l'importance des variables exogènes, notamment météorologiques, dont l'intégration améliore significativement la qualité des prévisions. En conclusion, les auteurs affirment qu'aucun modèle universel ne se distingue nettement : le choix dépend du contexte, de l'horizon de prévision et des données disponibles. Cette étude confirme que les métriques les plus utilisées sont MAPE, RMSE et MAE.

Autres modèles ont eu le jour comme Prophet, NeuralProphet et le modèle deep learning N-BEATS, qui mérite d'être exploré. Nous les avons évalués et écartés sans les étudier en profondeur. En effet à cause des problèmes de compatibilité des packages, des problèmes de dépendances, et l'instabilité de l'environnement engendré, nous les avons écartés de notre étude. En plus, les résultats obtenus pour quelques tests n'étaient pas convaincantes.

### 3.3 Modèle proposé

#### 3.3.1 Architecture générale

L'analyse que nous avons effectuée et les hypothèses que nous avons validées montre qu'il faut effectuer une analyse temps-fréquence de la variable cible pour générer ses périodes, décomposer la variable cible ainsi que les variables exogènes selon ces périodes, puis établir un modèle adapté pour chaque composante.

Les études [4, 25, 46] montrent que le modèle autorégressif intégré à moyenne mobile saisonnière SARIMA<sup>1</sup>, est le plus adapté pour les séries cycliques/saisonnières avec des saisonnalités stables. L'analyse temps-fréquence des séries temporelles de la consommation d'électricité montre qu'elles sont multi-saisonnières et que leurs saisonnalités sont stables dans le temps. Autrement dit, l'analyse temps-fréquence par transformée de Fourier à court terme, montre la présence de périodes dominantes et que celles-ci sont invariantes dans le temps. Ces périodes, détectées par cette analyse, sont utilisées pour décomposer la série temporelle de la consommation d'électricité (en Wh) ainsi que les séries temporelles représentant les variables météorologiques.

Nous avons constaté que la tendance de la consommation d'électricité, dans sa relation avec les tendances des variables météorologiques, présente un effet dynamique reflétant le fait que les individus s'adaptent aux conditions météorologiques changeantes avec un certain retard. La tendance de la consommation ne dépend pas uniquement de la température à l'instant présent, mais aussi des températures des jours précédents. Ceci est également confirmé par d'autres études [9, 19, 25, 44]. Nous avons fait les mêmes constations pour la composante résiduelle. Les valeurs de la tendance de la série de consommation dépendent de ses valeurs passées et des valeurs présentes et passées des variables météorologiques. Le modèle le plus adapté pour capter cet effet de mémoire doit être basée sur des architectures LSTM (Long Short Term Memory). Nous adoptons donc ce modèle pour les composantes tendance et résidus. L'architecture général de notre approche est donnée par le schéma de la figure 3.1.

La figure 3.2 illustre l'implémentation de ce schéma en pipeline. Chaque partie de cette architecture sera décrite dans une section dédiée.

#### 3.3.2 Étape de détection des saisonnalités par analyse temps-fréquence

##### 3.3.2.1 Principe

L'objectif de cette étape est d'extraire les périodes dominantes pour identifier toutes les saisonnalités des séries  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_m}$ . Cette extraction est faite par analyse temps-fréquence, basée sur la Transformée de Fourier à Court Terme (Short-Time Fourier Transform (STFT)). Les fondements théoriques

---

1. Seasonal AutoRegressive Integrated Moving Average

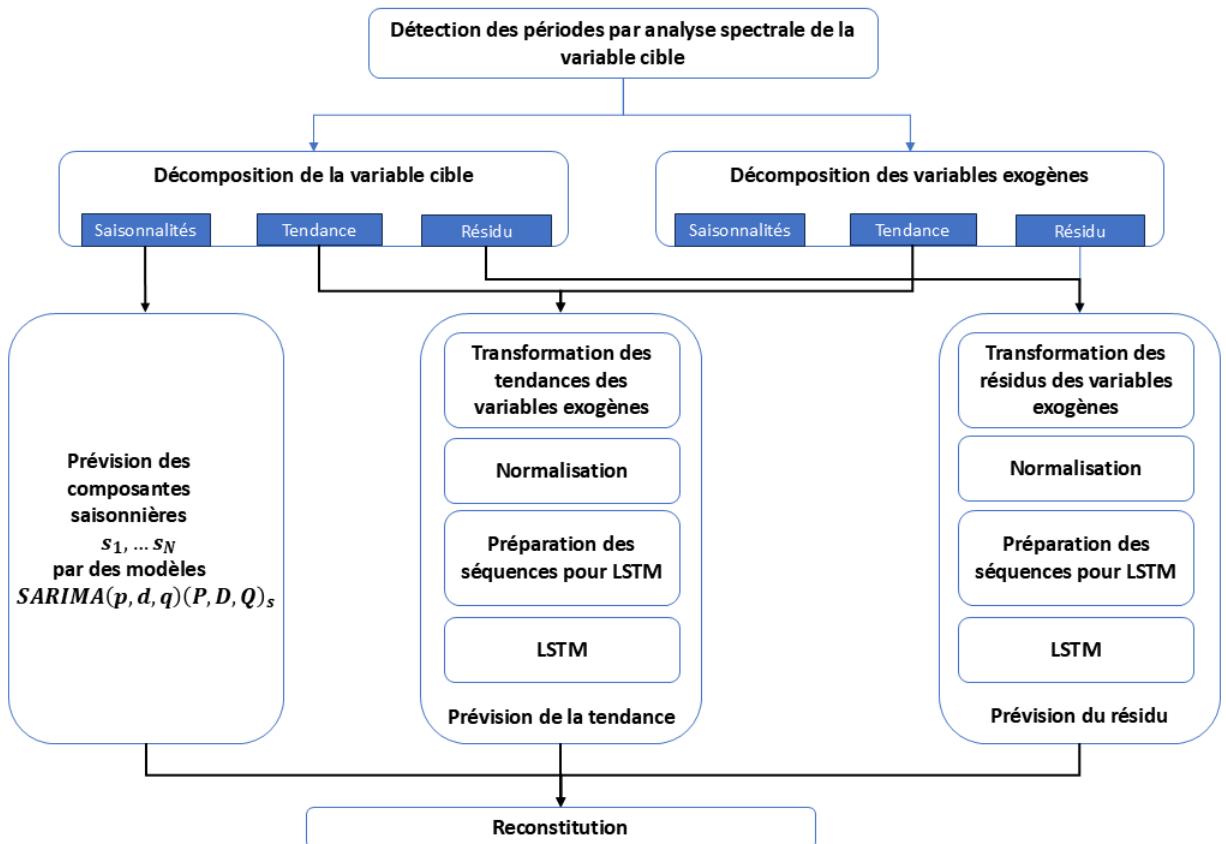


FIGURE 3.1 – Architecture générale

de cette étape ont été décrites dans le chapitre de formalisation. Le principe de l'implémentation est décrit par l'algorithme 3.1.

---

**Algorithm 3.1** Analyse temps-fréquence et détection des périodes dominantes

---

- ❖ Analyse temps-fréquence (SpectrogramAnalysis)
    - \* Entrée :
      - ◊  $w$  : fenêtre de lissage, par défaut *hann*
      - ◊  $N$  : longueur de la fenêtre
      - ◊  $\rho$  : longueur du recouvrement
      - ◊  $\delta$  : seuil de détection
      - ◊  $F_s$  : fréquence d'échantillonnage
    - \* Sortie :
      - ◊  $\mathbb{P}_\delta$  : vecteur de périodes dominantes relativement au seuil  $\delta$
    - \* fit :
      - ◊ Entrée : une série temporelle  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_m}$
      - ◊ Action :
        - Calcul du spectrogramme  $\mathbb{S}^{(r,q)}(N, \rho, w)$  (voir l'équation 1.27)
        - Détection des périodes dominantes  $\mathbb{P}_\delta^{(r,q)}$  au seuil  $\delta$  (voir l'équation 1.28)
    - \* transform :
      - ◊ Entrée : une série temporelle  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_m}$
      - ◊ Action : retourne les périodes dominantes  $\mathbb{P}_\delta^{(r,q)}$  au seuil  $\delta$
    - \* fit\_transform :
      - ◊ Combinaison de fit et transform
    - \* plot :
      - ◊ Affichage du spectrogramme avec marquage des périodes détectées
-

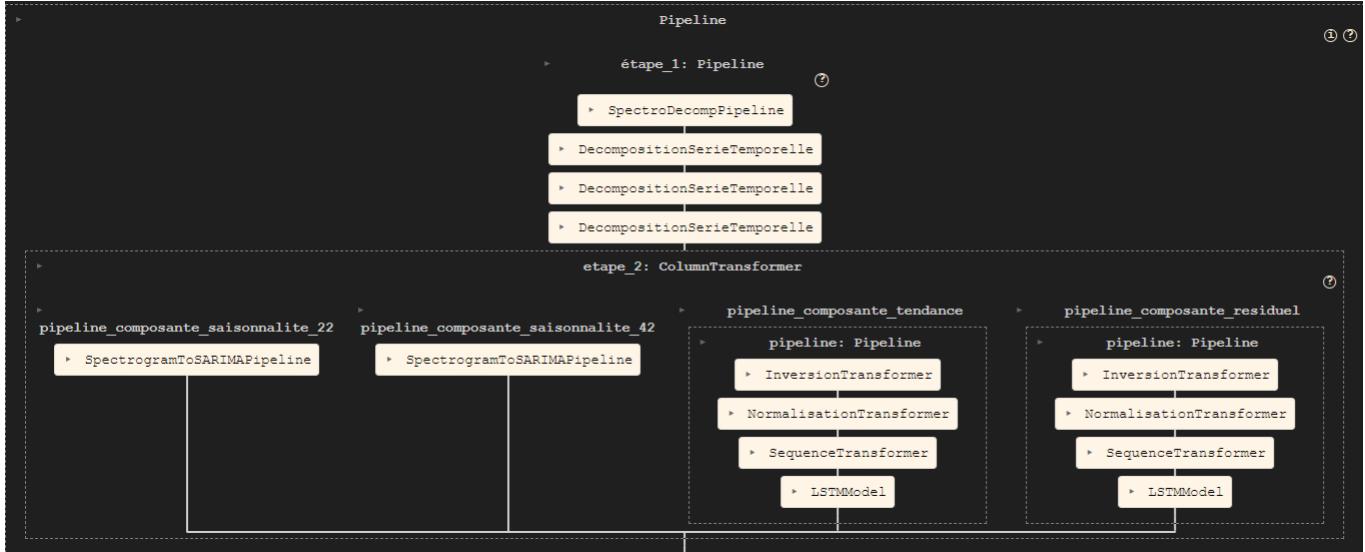


FIGURE 3.2 – Pipeline de l'architecture générale

### 3.3.2.2 Exemple

La figure 3.3 illustre un exemple d'analyse temps-fréquence où trois saisonnalités :  $P_1 = 21 \times T_s = 10,5 \text{ heures}$ ,  $P_2 = 42 \times T_s = 21 \text{ heures}$  et  $P_3 = 336 \times T_s = 7 \text{ jours}$  sont détectés.

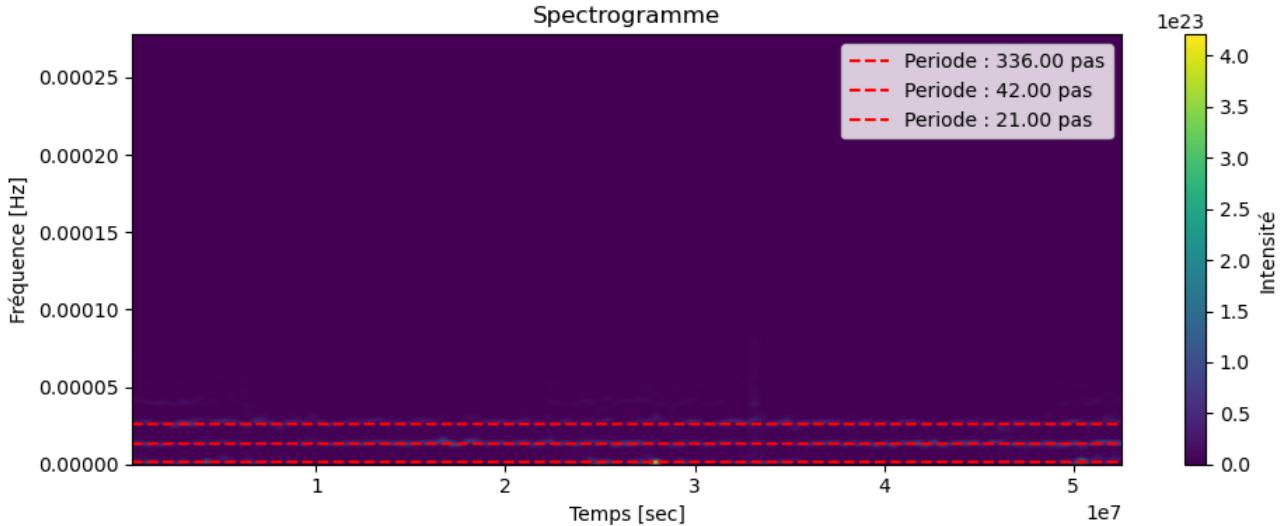


FIGURE 3.3 – Spectrogramme de la série temporelle de consommation d'électricité sur toute la région Auvergne-Rhône-Alpes  $\left( \sum_{q \in \mathcal{C}} Y_k^{(84,q)} \right)_{kT_s \in \mathbb{T}}$

### 3.3.3 Notre méthode de décomposition des séries temporelles

Cette étape est basée sur le principe de décomposition de séries temporelles, décrit dans le chapitre 1. L'algorithme 3.2 explicite l'adaptation de ce principe pour effectuer une décomposition en cascade et extraire les différentes composantes.

---

**Algorithm 3.2** Décomposition en cascade des séries temporelles
 

---

❖ Décomposition (DecompositionSerieTemporelle)

★ Entrée :

◊  $\mathbb{P}_\delta^{(r,q)} = \{p_1, p_2, \dots, p_M\}$  : périodes dominantes d'une série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$

◊ forme : additive ou multiplicative

◊ colonne : la colonne cible de la dataframe d'entrée à décomposer

★ Sortie :

◊  $\mathcal{D}$  : un dataframe contenant les composantes saisonnières, la tendance et le résidus

★ fit :

◊ Entrée : un dataframe contenant les colonnes  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_1}, (T_k^{(r)})_{kT_s \in \mathbb{T}_1}, (U_k^{(r)})_{kT_s \in \mathbb{T}_1}, (R_k^{(r)})_{kT_s \in \mathbb{T}_1}$

◊ fait rien

★ transform :

◊ Entrée : un dataframe contenant les colonnes  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_1}, (T_k^{(r)})_{kT_s \in \mathbb{T}_1}, (U_k^{(r)})_{kT_s \in \mathbb{T}_1}, (R_k^{(r)})_{kT_s \in \mathbb{T}_1}$

◊ Action :

- Sélection de la colonne à décomposer  $(C_k)_{kT_s \in \mathbb{T}}$

- Pour chaque  $p \in \mathbb{P}_\delta^{(r,q)}$  :

Extraire la composante saisonnière  $(S_{p,k})_{kT_s \in \mathbb{T}}$  de période  $p$ , la tendance  $(\Gamma_{p,k})_{kT_s \in \mathbb{T}}$  et le résidu  $(\epsilon_{p,k})_{kT_s \in \mathbb{T}}$  associés

$$\text{forme multiplicative} : C_k = \Gamma_{p,k} \times S_{p,k} \times \epsilon_{p,k} \quad (3.2)$$

$$\text{forme additive} : C_k = \Gamma_{p,k} + S_{p,k} + \epsilon_{p,k} \quad (3.3)$$

Extraire la tendance  $(\Gamma_k)_{kT_s \in \mathbb{T}}$  et le résidu  $(\epsilon_k)_{kT_s \in \mathbb{T}}$  de  $(C_k)_{kT_s \in \mathbb{T}}$

$$\text{forme multiplicative} : \epsilon_k = \frac{\Gamma_{p_m,k}}{\Gamma_k \times \prod_{p \in \mathbb{P}_\delta^{(r,q)}} S_{p,k}} \quad (3.4)$$

$$\text{forme additive} : \epsilon_k = C_k - \Gamma_k - \sum_{p \in \mathbb{P}_\delta^{(r,q)}} S_{p,k} \quad (3.5)$$

◊ Retourne un dataframe contenant les composantes saisonnières  $(S_{p,k})_{kT_s \in \mathbb{T}}$ , la tendance  $(\Gamma_k)_{kT_s \in \mathbb{T}}$  et le résidu  $(\epsilon_k)_{kT_s \in \mathbb{T}}$  de la colonne cible

★ fit\_transform :

◊ Combinaison de fit et transform

---

### 3.3.3.1 Implémentation

L'analyse temps-fréquence, la détection des périodes dominantes et la décomposition de la variable cible et des variables exogènes est pipeliné comme le montre la figure 3.4.

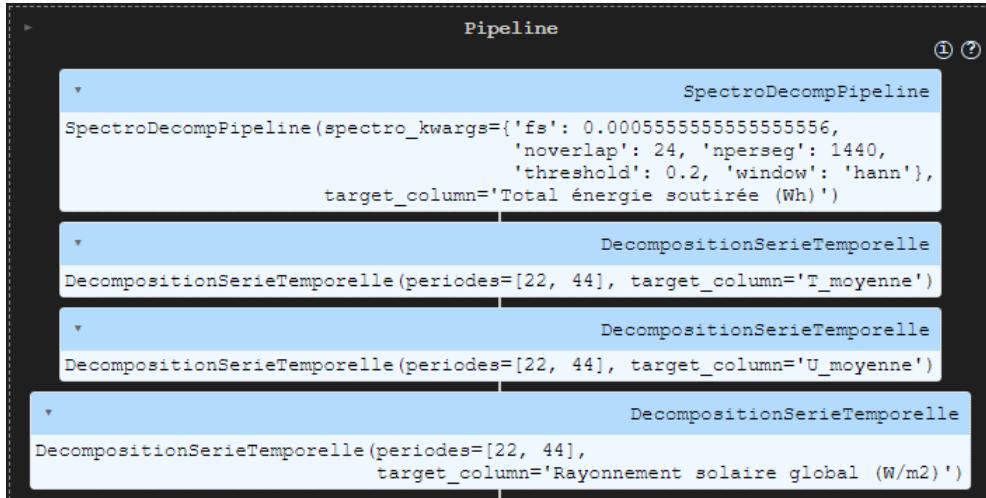


FIGURE 3.4 – Exemple de pipeline pour l’analyse temps-fréquence et la décomposition des séries temporelles

### 3.3.4 Modèles SARIMA pour la prévision des composantes saisonnières

Nous utilisons un modèle  $SARIMA(p, d, q)(P, D, Q)_s$  pour la prévision de chaque composante saisonnière, de période  $s$ . Une recherche du meilleur modèle, en utilisant une cross-validation adaptée aux séries temporelles, est intégrée au modèle. Avant d’entraîner le modèle, une vérification par analyse temps-fréquence de la saisonnalité est effectuée. En effet, après décomposition certaines composantes saisonnières n’ont pas une seule période pure, l’intérêt de cette nouvelle analyse est d’entraîner SARIMA avec la période fondamentale de la composante. Parmi les modèles permettant une meilleure prévision nous avons  $SARIMA(1, 0, 1)(1, 1, 1)_s$  qu’on peut décrire par l’équation :

$$Y_k - Y_{k-s} - \phi(Y_{k-1} - Y_{k-s-1}) - \Phi_1(Y_{k-1} - Y_{k-2s}) + \phi_1\Phi_1(Y_{k-s-1} - Y_{k-2s-1}) = b_k + \theta_1 b_{k-1} + \Theta_1 b_{k-s} + \theta_1\Theta_1 b_{k-s-1} \quad (3.6)$$

où  $\phi_1, \Phi_1, \theta_1, \Theta_1$  sont les paramètres à estimer pour la prévision et  $b_k$  est un bruit blanc gaussien.

Ceci est cohérent avec le fait que les composantes saisonnières traitées sont cyclo-stationnaires. La différenciation saisonnière permet de rendre la série stationnaire afin d’appliquer le modèle SARIMA. Dans notre implémentation nous avons intégré aussi le traitement des cas de séries non stationnaires.

Les figures 3.5 et 3.6 montrent un exemple de prévision des composantes saisonnières de périodes  $P_1 = 22 \times T_s = 11$  heures, de  $P_2 = 44 \times T_s = 22$  heures de la série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$  où  $r = 84$  (région Auvergne-Rhône-Alpes) et la configuration profile\_plage de puissance est  $q = (RES11(+RES11WE), P4 : ]9 - 12] kVA)$ . L’erreur dans ces cas est nulle.

### 3.3.5 Modèles basée sur les réseaux LSTM pour les composantes tendance et résidu

La prévision de la tendance et du résidus est effectuée grâce à un modèle basée sur des LSTM(Long Short Term Memory). La figure 3.7 donne une vue globale de ce modèle avec un exemple de pipeline de la composante résiduelle.

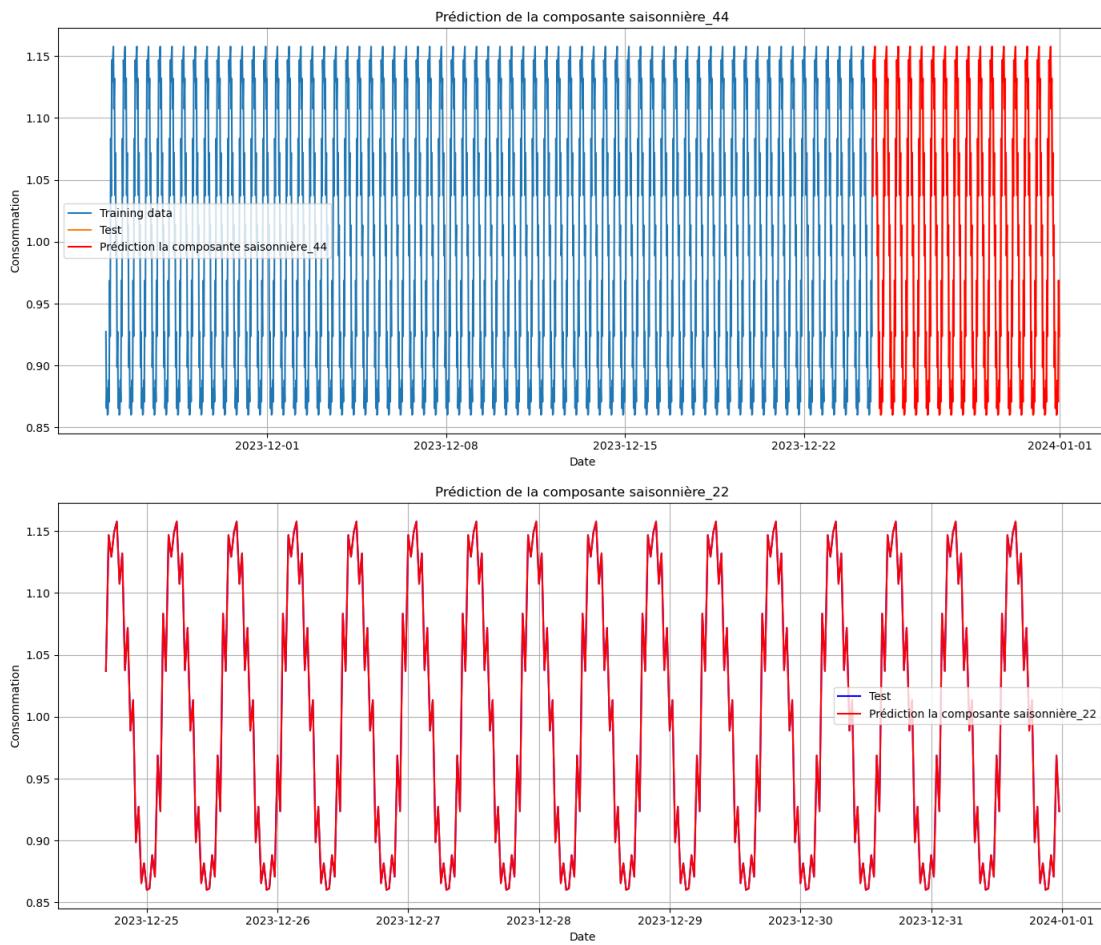


FIGURE 3.5 – Prévision par  $SARIMA(1, 0, 1)(1, 1, 1)_{22}$  de la composante saisonnière de période  $P_1$

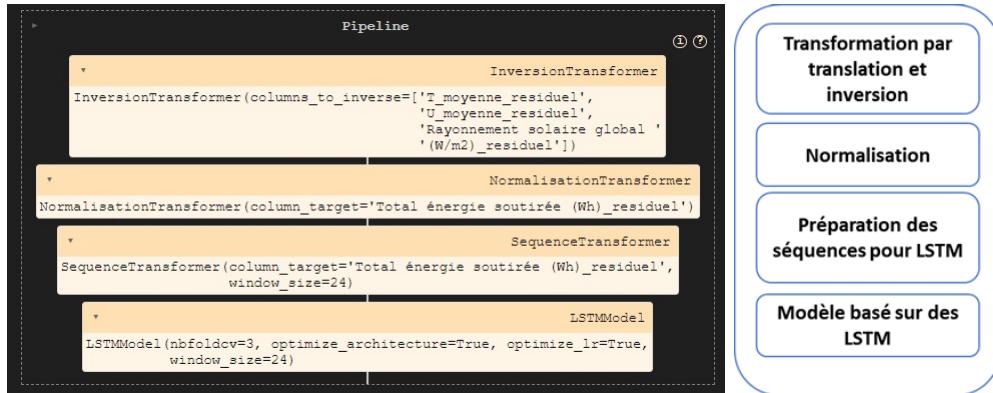


FIGURE 3.7 – Les étapes du pipeline pour le modèle de prévision du résidu

Nous avons démontré dans le chapitre 2 que la tendance (respectivement le résidus) de la variable cible est corrélée avec les inverses des tendances (respectivement les résidus) des variables exogènes. Cette transformation opérée sur les tendances et les résidus est argumentée dans le chapitre 2. Nous rappelons que celle-ci consiste à appliquer une translation afin d'éviter des déviations par zéros, suivie d'une inversion :

$$f : \begin{array}{ccc} \mathbb{R} & \rightarrow & R \\ X & \rightarrow & f(X) = \frac{1}{X+C} \end{array} \quad (3.7)$$

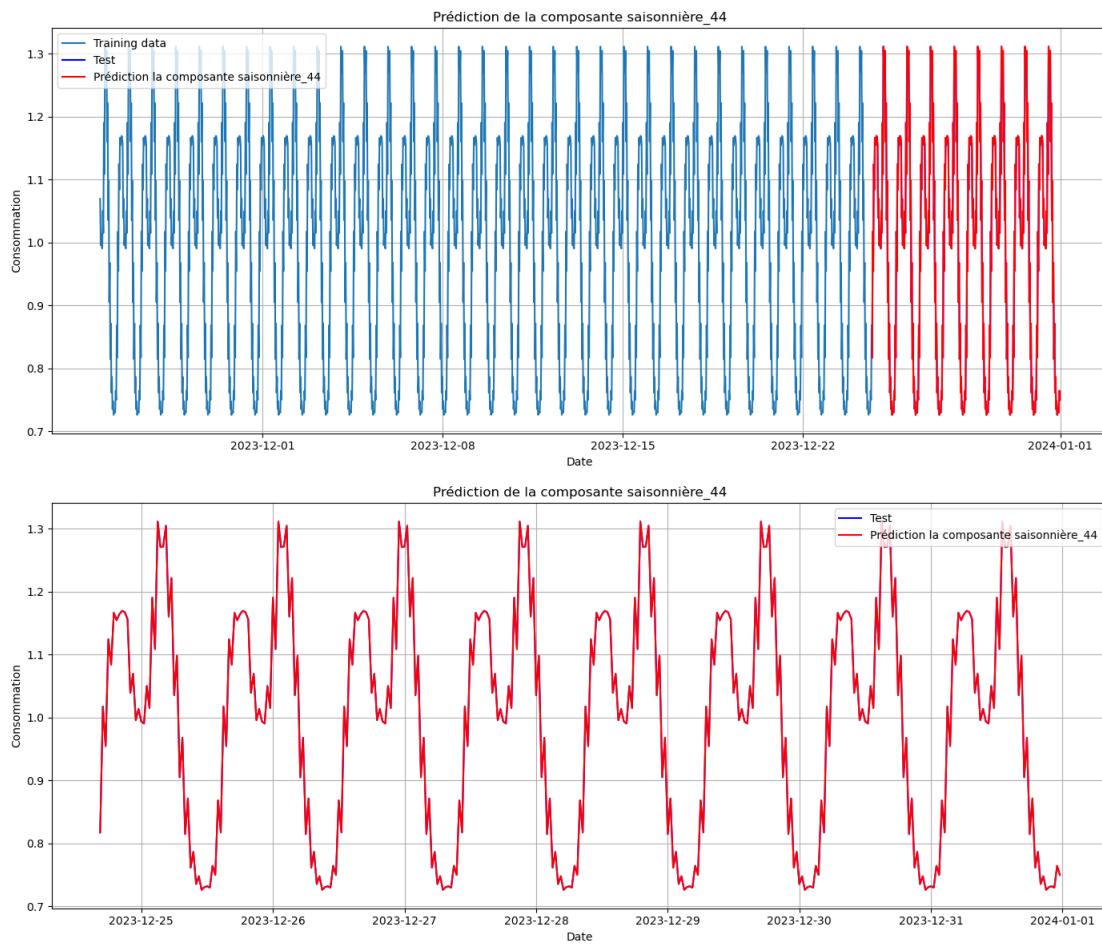


FIGURE 3.6 – Prévision par  $SARIMA(1, 0, 1)(1, 1, 1)_{44}$  de la composante saisonnière de période  $P_2$

Après transformation les nouvelles composantes (inverse des tendances ou inverses des résidus) et la composante de la variable cible sont normalisées. Ces données normalisées nécessitent une transformation en séquences adaptées au traitement séquentiel. Cette étape consiste à les structurer en sous-séquences d'entrée de longueur fixe (fenêtres glissantes) associées aux valeurs cibles. La figure 3.8 illustre cette structuration.

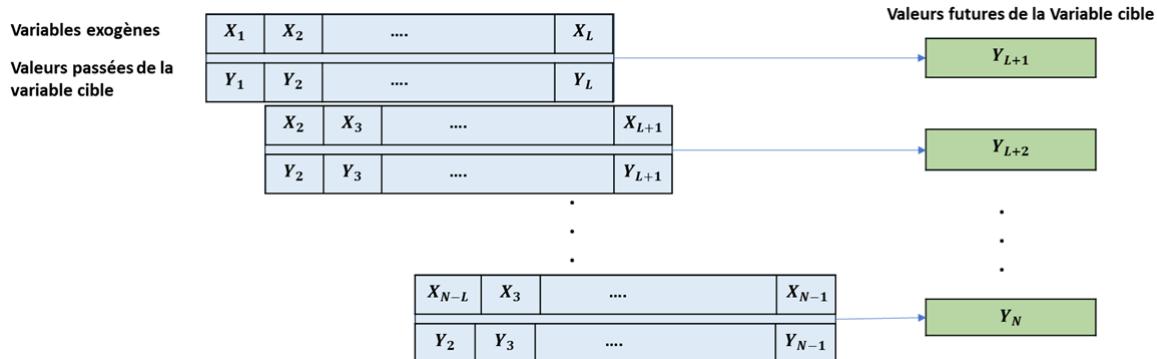


FIGURE 3.8 – Transformation en séquences des données pour LSTM

Enfin le modèle est basée sur l'empilement de plusieurs couche LSTM. Pour déterminer la meilleure architecture, le nombre de couches ainsi que le nombre de neurones par couche ont été intégrés aux hyperparamètres du modèle. L'architecture typique que nous avons identifiées est illustrées dans la

figure 3.9.

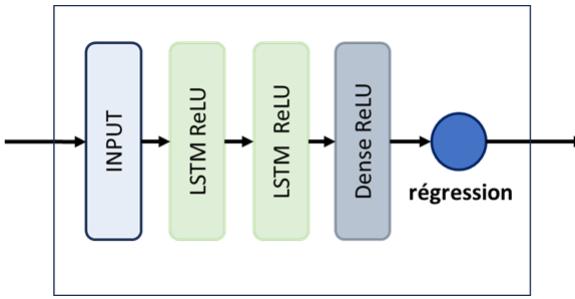
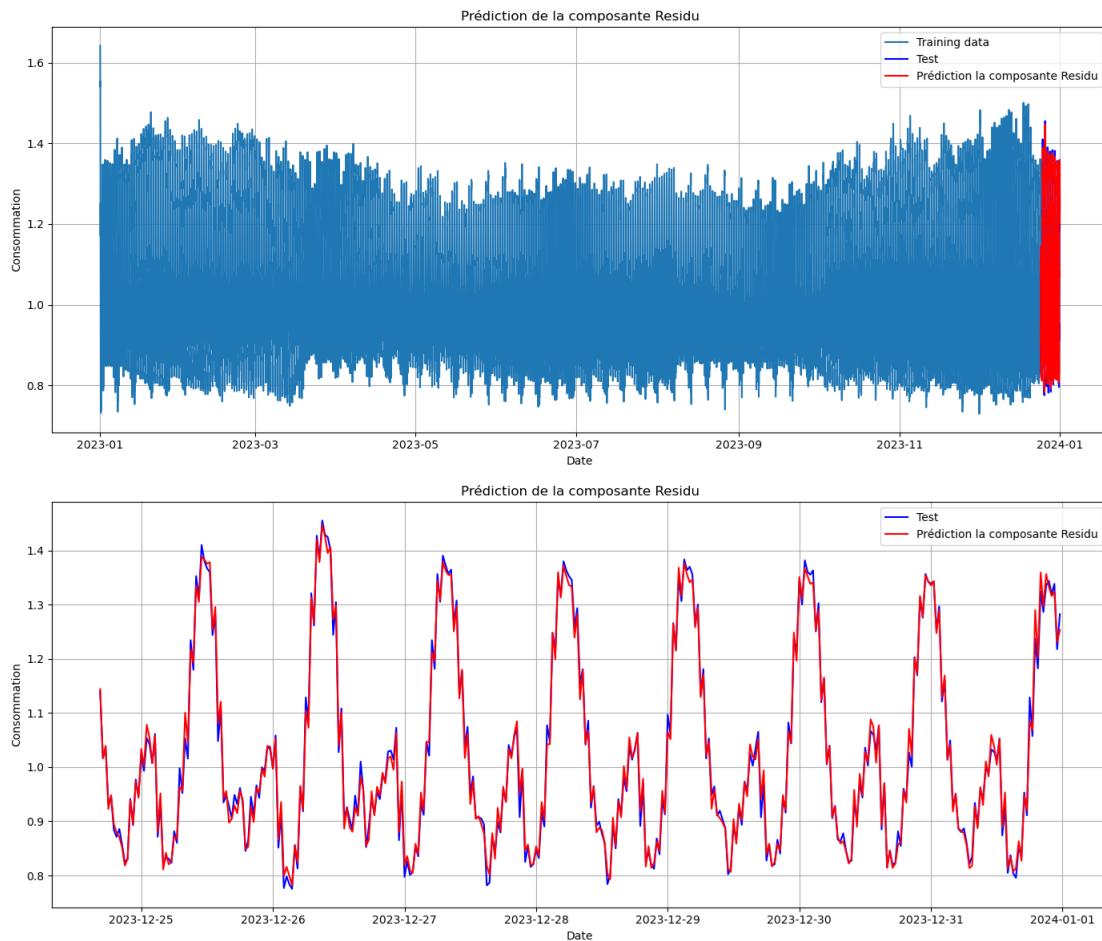


FIGURE 3.9 – Architecture typique de notre modèle basé sur des réseaux LSTM

Les figures 3.5 et 3.6 montrent un exemple de prévision de la tendance et le résidu de la série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$  où  $r = 84$  (région Auvergne-Rhône-Alpes) et la configuration profile - plage de puissance est  $q = (RES11(+RES11WE), P4 : ]9 - 12]kVA)$ .

FIGURE 3.10 – Prévision du résidu de la série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$ 

Les figures 3.5 montrent la prévision de la série utilisée dans les exemples précédents après recompilation.

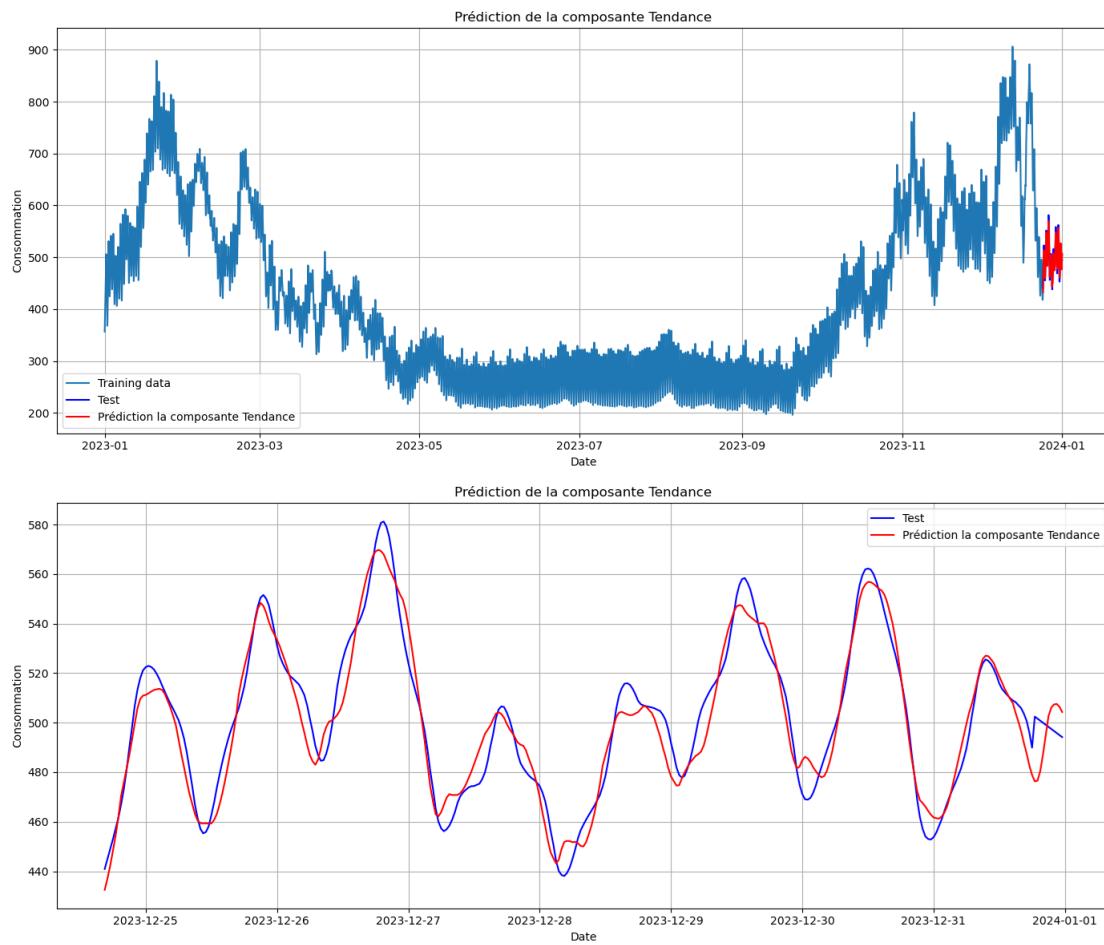


FIGURE 3.11 – Prévision de la tendance de la série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$

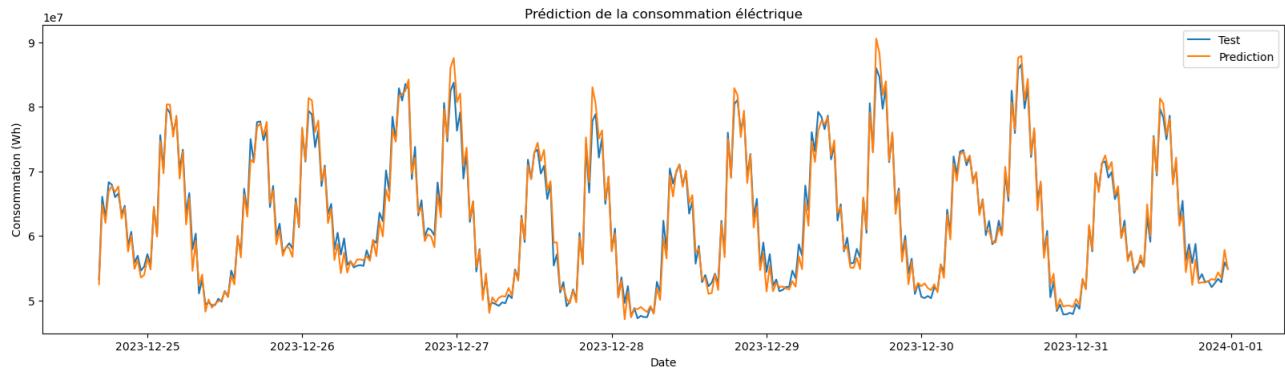


FIGURE 3.12 – prévision de la série  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}}$

### 3.3.6 Réalisation

Le script finale que nous avons développé est décrit par l'algorithme 3.3.

---

**Algorithm 3.3** Le script final
 

---

- ❖ Script final
  - ★ Entrée :
    - ◊ Base de données :  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_m}, (T_k^{(r)})_{kT_s \in \mathbb{T}_m}, (U_k^{(r)})_{kT_s \in \mathbb{T}_m}, (R_k^{(r)})_{kT_s \in \mathbb{T}_m}, (N_k^{(r,q)})_{kT_s \in \mathbb{T}}$
    - ◊ l'horizon  $h$  pour la prévision, par défaut  $h = 7 * 48 = 336$
    - ◊ Les hyperparamètres
  - ★ Sortie :
    - ◊  $\mathcal{D}$  : un DataFrame contenant les résultats pour chaque configuration  $q \in \mathcal{C}$  et chaque région  $r \in \mathcal{R}$
  - ★ Pour chaque  $(r, q) \in \mathcal{R} \times \mathcal{C}$ 
    - ◊ Séparation des données pour l'entraînement et le test
      - Tirer aléatoirement un  $m \in \{0, \dots, L/2 - 1\}$ 
        - $\mathbb{T}_m = \{kT_s, k \in \{m, \dots, \tau = m - 1 + L/2\}\}$  : l'ensemble d'instants d'observation pour l'entraînement
        - $\mathbb{T}_{\text{test}} = \{kT_s, k \in \{\tau + 1, \tau + h\}\}$  : l'ensemble d'instants d'observation pour le test
      - Extraire les données d'entraînement :  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_m}, (T_k^{(r)})_{kT_s \in \mathbb{T}_m}, (U_k^{(r)})_{kT_s \in \mathbb{T}_m}, (R_k^{(r)})_{kT_s \in \mathbb{T}_m}$
      - Extraire les données de test :  $(Y_k^{(r,q)})_{kT_s \in \mathbb{T}_{\text{test}}}, (T_k^{(r)})_{kT_s \in \mathbb{T}_{\text{test}}}, (U_k^{(r)})_{kT_s \in \mathbb{T}_{\text{test}}}, (R_k^{(r)})_{kT_s \in \mathbb{T}_{\text{test}}}$
    - ◊ Calculer  $(\bar{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}_m} = \left( \frac{Y_k^{(r,q)}}{N_k^{(r,q)}} \right)_{kT_s \in \mathbb{T}_m}$
    - ◊ Analyse temps-fréquence et extraction des périodes dominantes de  $(\bar{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}_m}$
    - ◊ Décomposition des séries  $(\bar{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}_m}, (T_k^{(r)})_{kT_s \in \mathbb{T}_m}, (U_k^{(r)})_{kT_s \in \mathbb{T}_m}$  et  $(R_k^{(r)})_{kT_s \in \mathbb{T}_m}$
    - ◊ Pour chaque composante saisonnière
      - Si la recherche automatique du meilleur modèle est activé, chercher le meilleur modèle  $SARIMA(p, d, q)(P, D, Q)_s$
      - Sinon appliquer le modèle par défaut 3.1
      - Sauvegarder le modèle
    - ◊ Pour la tendance
      - Si la recherche automatique du meilleur modèle est activé, chercher la meilleure architecture
      - Sinon appliquer le modèle par défaut 3.6
      - Sauvegarder le modèle
    - ◊ Pour le résidu
      - Si la recherche automatique du meilleur modèle est activé, chercher la meilleure architecture
      - Sinon appliquer le modèle par défaut 3.6
      - Sauvegarder le modèle
    - ◊ Recomposition et calcul de la prévision  $(\hat{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}_{\text{test}}} = (N_k^{(r,q)} \times \bar{Y}_k^{(r,q)})_{kT_s \in \mathbb{T}_{\text{test}}}$
    - ◊ Calcul des métriques MAPE (équation 1.30), MAE (équation 1.31) et RMSE (équation 1.32)
    - ◊ Sauvegarde des résultats
- 

Le figure 3.13 schématisé la séparation des données d'entraînement et de test décrite dans l'algorithme 3.3.



FIGURE 3.13 – Séparation des données

Nous avons tout d'abord développé une première version sur notebook. Celle-ci nous a permis de valider le concept. Dans celle-ci chaque étape est modélisée par un pipeline. Puis nous avons développées une versions notebook avec un pipeline intégrale. Nous avons constaté une dégradation dans les performances que nous n'avons pas su expliquer. Nous avons donc développée le script décrit précédemment, qui traite toutes les configurations en se basant sur la version avec pipelines séparées. Nous avons intégré du parallélisme pour accélérer le calcul. Cependant nous étions affronté aux limitations de la puissance de calcul de notre machine. Nous avons traité une région et nous espérons pouvoir traiter les autres régions avant la présentation finale.

### 3.3.7 Evaluation des performances

Dans cette section nous présentons et analysons les résultats de notre modèle sur les données de la région Auvergne-Rhône-Alpes. Nous avons choisi un horizon d'une semaine pour la prévision. Nous avons effectué une simulation avec l'option d'optimisation et d'ajustement des hyperparamètres. Malgré le traitement en parallèle des configurations profil - plage de puissance, les temps sont de l'ordre de 20 à 40 minutes par configuration.

#### 3.3.7.1 Analyse des résultats par rapport à la métrique MAPE

Les figures 3.14 et 3.15 montrent pour chaque profil la distribution de la métrique MAPE en fonction des plages de puissance souscrite.

La figure 3.15 montre pour chaque plage de puissance souscrite la distribution de la métrique MAPE en fonction des profils.

Nous constatons que pour les plages de puissance souscrite P6 : [15-18[ KVA et P3 :]6-9[ KVA pour toutes les puissances, le MAPE dépassent dans certain cas les 2% Contrairement aux autres plages. La figure montre en nuage de point la métrique MAPE pour chaque configuration  $q \in \mathcal{C}$  dans la région Auvergne-Rhône-Alpes ( $r = 84$ ). Ce graphique nous permet d'identifier les configurations nécessitant un examen de près pour identifier les raisons des valeurs élevées de la métrique MAPE.

#### 3.3.7.2 Analyse des résultats par rapport aux métriques MAE et RMSE

La figure 3.17 montre pour chaque profil la distribution des métriques MAE et RMSE en fonction des plages de puissance souscrite.

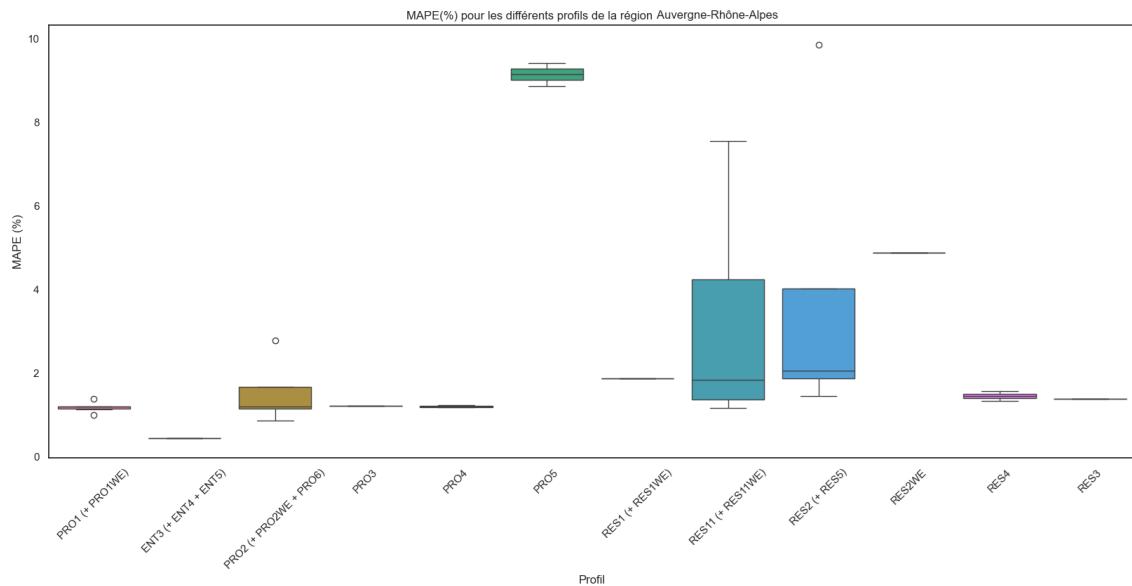


FIGURE 3.14 – Distribution de la métrique MAPE pour les différents profils en fonction des plages de puissance souscrite.

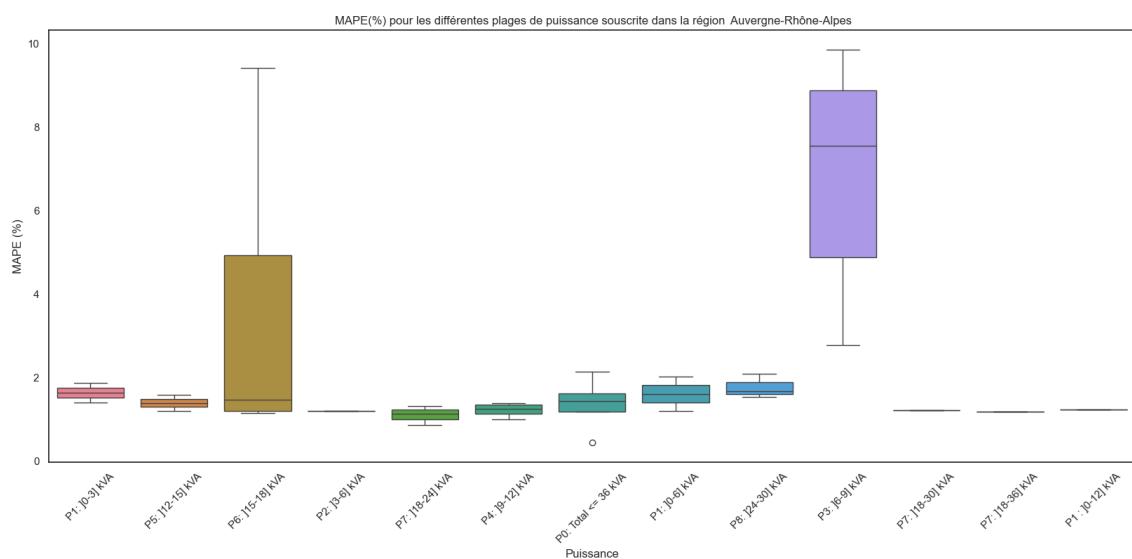


FIGURE 3.15 – Distribution d'erreur MAPE pour chaque plage de puissance souscrite en fonction des profils

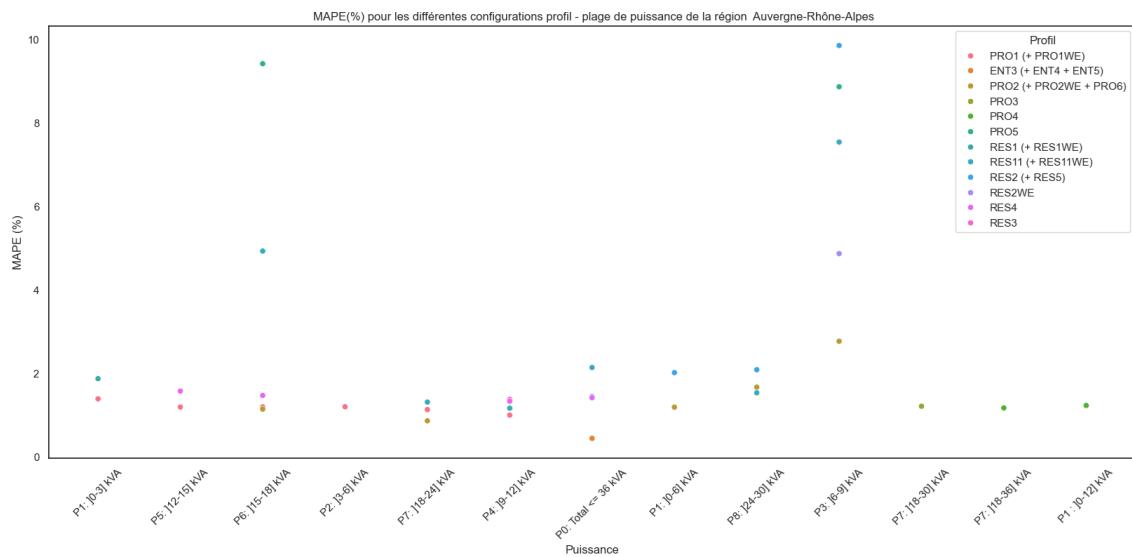


FIGURE 3.16 – Distribution de la métrique MAPE pour chaque plage de puissance souscrite en fonction des profils

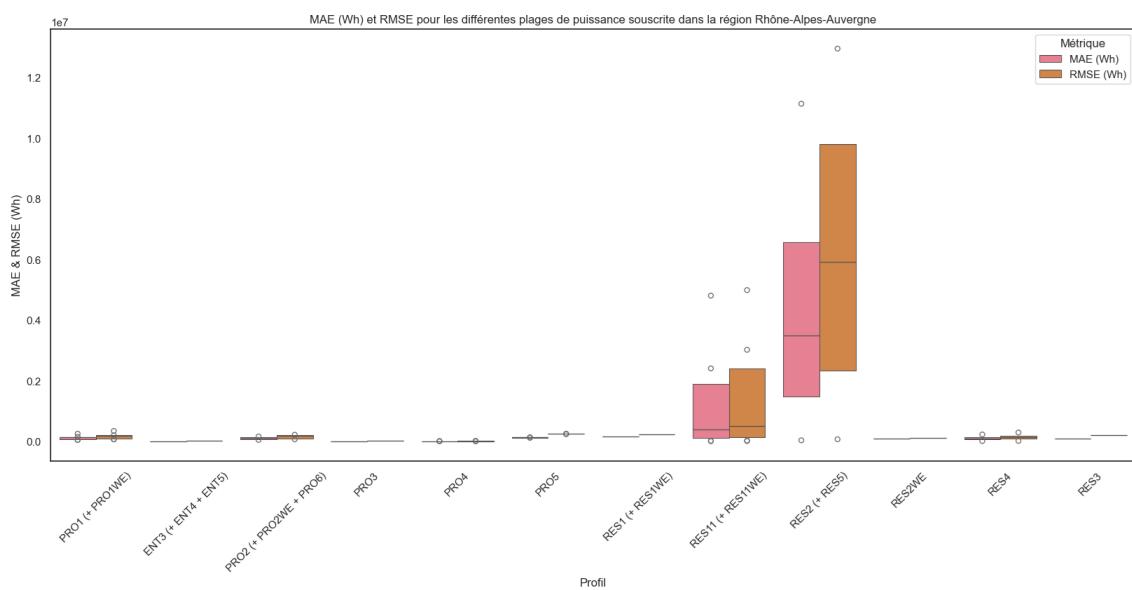


FIGURE 3.17 – Distribution des métriques MAE et RMSE pour les différents profils en fonction des plages de puissance souscrite.

La figure 3.18 montre pour chaque plage de puissance souscrite la distribution des métriques MAE et RMSE en fonction des profils.

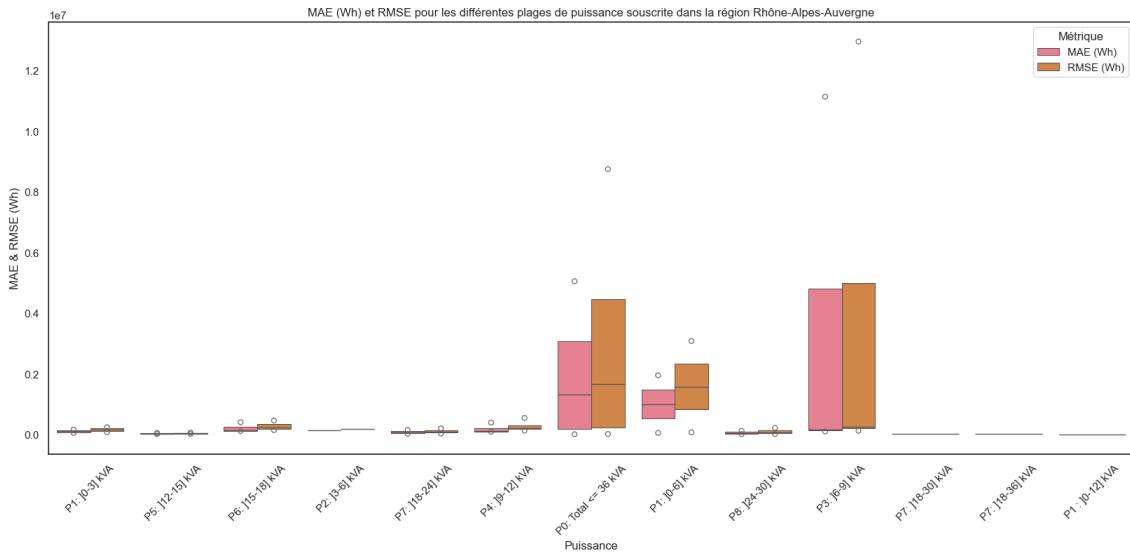


FIGURE 3.18 – Distributions des métriques MAE et RMSE pour chaque plage de puissance souscrite en fonction des profils

Ces graphiques confirment le fait que les résultats des configurations RES11(+RES11WE) - P3 et RES11(+RES11WE) - P6, PRO5 - P3 et PRO5- P6, RES2(+RS5)-P3 sont très mauvais. Pour toutes les autres configurations nous avons des MAPE inférieurs à la barre de 2% . Pour certaines configurations le MAPE est inférieur à 1%. Dans la suite de notre travail, nous allons examiner les causes de ces faibles scores pour ces configurations.

### 3.4 Conclusion

Dans ce chapitre nous avons présenté une nouvelle approche pour la prévision de la consommation d'électricité à court-terme. Cette présentation était accompagnée d'un exposé de notre démarche scientifique qui nous a conduit à élaborer cette approche. Nous avons également montré la réalisation de cette approche et son évaluation avec les bases de données que nous avions construites. Les premiers résultats obtenus sont très encourageants pour un grand ensemble de configurations.

# Conclusion générale

Au cours de ce projet, nous nous sommes intéressés à la prévision de la consommation d'électricité à court terme, pour les utilisateurs du réseau Enedis. Étant donnée que ce sujet suscite un grand intérêt dans le monde académique et industriel, nous l'avons abordé comme un sujet de recherche scientifique. Les travaux menés dans ce cadre, et présentés dans ce rapport, ont pour objectif la recherche d'une nouvelle approche de prévision à court terme de la consommation d'électricité en France au demi-heure près. Pour répondre à ce problème, nous avons adopté une modélisation en séries temporelles des données de la consommation d'électricité, dite variable cible, et des données météorologiques, faisant partie des variables exogènes. Puis nous avons proposé une approche basée sur la décomposition de ces séries par analyse temps-fréquence et sur la prévision, de chaque composante de la variable cible en fonction des composantes des variables exogènes, par un modèle spécifique.

Nous avons introduit, dans le premier chapitre, les concepts généraux de la théorie des séries temporelles indispensables à notre travail. Nous avons illustré ces concepts à travers des exemples tirés de nos données. Cette étude théorique nous a permis de formaliser notre problème et de valider les hypothèses de travail. Une première contribution, tirée de cette première étude, est l'utilisation de l'analyse temps-fréquence pour extraire les périodes dominantes d'une série temporelle et décomposer celle-ci, selon une nouvelle méthode, en utilisant ces périodes.

Dans le second chapitre, nous avons tout d'abord présenté notre démarche de collecte, de nettoyage, de construction et de traitement des données. Puis nous avons exposé l'analyse des relations entre les variables exogènes et la variable cible. Ce travail nous a permis de définir des transformations à opérer sur les composantes des variables exogènes afin d'augmenter leur corrélation avec les composantes correspondantes de la variable cible. L'exploration de ces relations nous a conduit à identifier les modèles les plus appropriés à la prévision de chaque composante.

Dans le dernier chapitre, nous avons exposé l'approche proposée. Nous avons présenté son architecture, détaillé sa réalisation, puis présenté et analysé ses performances. La première étape de cette solution effectue la décomposition des séries par notre méthode. La deuxième consiste à combiner les transformations des composantes avec modèles autorégressifs, pour la prévision des composantes saisonnières, et avec des modèles de régression, dont l'architecture est une mise en cascade de réseaux de neurones récurrents à mémoire court-terme et long terme LSTM et de couches denses, pour la prévision de la tendance et du résidu. Les premiers résultats obtenus sont encourageants et dans la plupart des configurations profil - plage de puissance souscrite les performances sont très satisfaisantes.

Notre approche reste perfectible afin d'améliorer les temps d'exécution et les performances pour certaines configurations. Pour le temps d'exécution, l'apprentissage par transfert est une piste à explorer,

vu la similarité entre des profils d'utilisateurs. Et pour les utilisateurs dont les performances sont perfectibles, l'intégration d'autres caractéristiques, comme les jours spéciaux, peut améliorer la précision de la prévision.

Enfin ce projet était l'occasion de mettre en œuvre l'ensemble des méthodes et techniques apprises dans les cours. Ce projet nous a permis également de découvrir le domaine de la prévision de la consommation d'électricité.

# Bibliographie

- [1] Règles de marché ? chapitre 3 : Dispositif de responsable d'Équilibre. Document réglementaire RTE France, Version 01 en vigueur au 1er avril 2024, 2024. Disponible sur le site de RTE France.
- [2] A. Ahmed, Y. Serrestou, K. Raoof, and J.-F. Diouris. Empirical mode decomposition-based feature extraction for environmental sound classification. *Sensors*, 22(20) :7717, October 2022.
- [3] M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia. Short-term prediction of household electricity consumption : Assessing weather sensitivity in a mediterranean area. *Renewable and Sustainable Energy Reviews*, 12(8) :2040–2065, 2008.
- [4] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis : Forecasting and Control*. John Wiley & Sons, 5th edition, 2015.
- [5] Peter J. Brockwell and Richard A. Davis. *Time Series : Theory and Methods*. Springer Series in Statistics. Springer, 2nd edition, 2016.
- [6] Manuela Brunet, Philip D. Jones, Javier Sigró, and Oriol Saladie. Spatial analysis of french monthly temperature trends. *International Journal of Climatology*, 27(2) :205–223, 2007.
- [7] J. R. Cancelo, A. Espasa, and R. Grafe. Forecasting the electricity load from one day to one week ahead for the spanish system operator. *International Journal of Forecasting*, 24(4) :588–602, 2008.
- [8] A. E. Clements, A. Hurn, and Z. Li. Forecasting day-ahead electricity load using a multiple equation time series approach. *European Journal of Operational Research*, 251(2) :522–530, 2016.
- [9] D. M. L. Comte and H. E. Warren. Modeling the impact of summer temperatures on national electricity consumption. *Journal of Applied Meteorology and Climatology*, 20(12) :1415–1419, 1981.
- [10] Richard C. Cornes, Gerard van der Schrier, Evert JM van den Besselaar, and Phil D. Jones. An ensemble version of the e-obs temperature and precipitation datasets. *Journal of Geophysical Research : Atmospheres*, 123(17) :9391–9409, 2018.
- [11] Charline David. Prévision à court terme du besoin électrique québécois. Mémoire de maîtrise en informatique, Université du Québec à Montréal, Montréal, Québec, 2023.
- [12] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366) :427–431, 1979.
- [13] Enedis. Note externe - jdd article 23-4.pdf. <https://data.enedis.fr/explore/dataset/nb-clients-inf-36/information/>, 2018. 2018-05-17.

- [14] Enedis. Coefficients des profils de consommation électrique. [https://data.enedis.fr/explore/dataset/coefficients-des-profil/information/?disjunctive.sous\\_profil](https://data.enedis.fr/explore/dataset/coefficients-des-profil/information/?disjunctive.sous_profil), 2025. Consulté le 18 juin 2025.
- [15] Enedis. Consommation électrique régionale ? clients résidentiels en puissance inférieure à 36 kva. <https://data.enedis.fr/explore/dataset/conso-inf36-region/information/>, 2025. Consulté le 18 juin 2025.
- [16] Enedis. Portail de données ouvertes - enedis. <https://data.enedis.fr/explore/?source=shared&sort=modified&exclude.theme=Divers&exclude.theme=Donn%C3%A9es%20sp%C3%A9cifiques%20Corse%20et%20outre-mer>, 2025. Consulté le 18 juin 2025.
- [17] S. Fan and R. J. Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1) :134–141, 2012.
- [18] P. Flandrin. Temps-fréquence. 1998.
- [19] P. G. Gould, A. B. Koehler, J. K. Ord, R. D. Snyder, R. J. Hyndman, and F. Vahid-Araghi. Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research*, 191(1) :207–222, 2008.
- [20] M. Grenier. Short-term load forecasting at hydro-québec transÉnergie. In *2006 IEEE Power Engineering Society General Meeting*, pages 5–pp. IEEE, 2006.
- [21] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
- [22] A. Henley and J. Peirson. Non-linearities in electricity demand and temperature : Parametric versus non-parametric methods. *Oxford Bulletin of Economics and Statistics*, 59(1) :149–162, 1997.
- [23] T. Hong, P. Pinson, and S. Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30(2) :357–363, 2014.
- [24] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, pages 903–995, 1998.
- [25] Rob J. Hyndman and George Athanasopoulos. *Forecasting : Principles and Practice*. OTexts, 3rd edition, 2021. Disponible gratuitement en ligne.
- [26] IBM. What is load forecasting ? <https://www.ibm.com/topics/load-forecasting>, 2024. Accessed : 2025-06-18.
- [27] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root ? *Journal of Econometrics*, 54(1-3) :159–178, 1992.
- [28] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1-3) :159–178, 1992.
- [29] Météo-France. Les normales climatiques 1991?2020 en france métropolitaine. <https://meteofrance.com/actualites/les-nouvelles-normales-climatiques-1991-2020>, 2021. Accessed : 2024-06-25.

- [30] Météo-France. Données climatologiques de base horaires. <https://www.data.gouv.fr/fr/datasets/donnees-climatologiques-de-base-horaires/>, 2024. Accès aux données horaires issues du réseau synoptique de Météo-France, incluant température, humidité, vent et autres paramètres météo en France métropolitaine. Consulté en 2024.
- [31] Météo-France. Rayonnement solaire global et vitesse du vent à 100 mètres tri-horaire régional depuis janvier 2016. <https://www.data.gouv.fr/fr/datasets/rayonnement-solaire-global-et-vitesse-du-vent-a-100-metres-tri-horaire-regionaux-depuis-2016/>. Données tri-horaires régionales issues du modèle atmosphérique de Météo-France : rayonnement global et vitesse du vent à 100 ?m pour la France métropolitaine. Consulté en 2024.
- [32] Météo-France. Données météo issues de stations synoptiques - france métropolitaine et outre-mer. <https://meteo.data.gouv.fr/datasets/6569b4473bedf2e7abad3b72>, 2025. Accédé en juillet 2025.
- [33] A. B. Nassif, B. Soudan, M. Azzeh, I. B. Attilli, and O. AlMulla. Artificial intelligence and statistical techniques in short-term load forecasting : A review. *CoRR*, abs/2201.00437, 2022. Available at <https://arxiv.org/abs/2201.00437>.
- [34] Isaac Osunmakinde and Olayinka Alani. Short-term multiple forecasting of electric energy loads using granular data-driven models. *International Journal of Energy and Power Engineering*, 11(2) :107 ?115, 2017.
- [35] A. Pardo, V. Meneu, and E. Valor. Temperature and seasonality influences on spanish electricity load. *Energy Economics*, 24(1) :55–70, 2002.
- [36] P. C. B. Phillips and P. Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2) :335–346, 1988.
- [37] Pere Quintana-Seguí, Patrick Le Moigne, Yves Durand, Eric Martin, Florence Habets, Marie Baillon, Cécile Canellas, Laurent Franchisteguy, and Sophie Morel. Analysis of near-surface atmospheric variables in france from the safran analysis. *International Journal of Climatology*, 28(8) :1119–1131, 2008.
- [38] Filipe Rodrigues, Carlos Cardeira, Jo ?o M. F. Calado, and Rui Melicio. Short-term load forecasting of electricity demand for the residential sector based on modelling techniques : A systematic review. *Energies*, 16 :4098, 2023.
- [39] Skipper Seabold, Josef Perktold, et al. *statsmodels.tsa.seasonal.seasonal\_decompose*. statsmodels developers, 2024. Accessed : 2025-06-26.
- [40] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications : With R Examples*. Springer, New York, 4th edition, 2017.
- [41] Soliman Abdel-Hady Soliman and Ahmad M. Al-Kandari. *Electric Load Forecasting : Fundamentals and Best Practices*. Springer, Boston, MA, 2010.
- [42] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Expert Systems with Applications*, 215 :119292, 2023.
- [43] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58 :80, 2025. Accepted : 29 November 2024 / Published online : 6 January 2025.

- [44] J. Taylor. An evaluation of methods for very short-term load forecasting using minute-by-minute british data. *International Journal of Forecasting*, 24 :645–658, 2008.
- [45] J. W. Taylor and R. Buizza. Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1) :57–70, 2003.
- [46] J. W. Taylor, L. M. de Menezes, and P. E. McSharry. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22(1) :1–16, 2006.
- [47] E. Valor, V. Meneu, and V. Caselles. Daily air temperature and electricity load in spain. *Journal of Applied Meteorology*, 40 :1413–1421, 2001.