Comment anticiper la consommation en énergie électrique d'une région en fonction de sa température?

Eliot Morard, Enzo Thomas

June 10, 2024

1 Introduction

Prédire la consommation énergétique est essentiel pour assurer la fiabilité, l'efficacité économique et la durabilité du système énergétique. En effet, les gestionnaires de réseau doivent maintenir un équilibre constant entre la production et la consommation d'électricité pour éviter les coupures de courant ou les surcharges. Une prévision précise permet de mieux ajuster la production en temps réel. Les acteurs du marché peuvent négocier des contrats futures avec une meilleure connaissance des besoins futurs, ce qui peut réduire les coûts et les risques financiers. Enfin les consommateurs peuvent bénéficier de tarifs plus bas si les fournisseurs optimisent leur production et leurs achats sur la base de prévisions précises.

Une mauvaise anticipation a notamment produit, l'hiver dernier, une multiplication par trois de la facture d'électricité pour de nombreux citoyens et entreprises. Ceci n'était pas directement dû à une pénurie d'énergie, mais plutôt à un effet bien connu des marchés financiers où certains acteurs s'engagent sur des contrats futurs et subissent une lourde perte suite à un mouvement imprévu de leur côté.

Nous allons dans un premier temps nous intéresser à la modélisation de la consommation électrique de la région Île-de-France entre 2016 et 2024. Nous ferons premièrement l'hypothèse que la consommation électrique est seulement fonction de la température. Ensuite, nous chercherons à modéliser les périodicités de la consommation en intégrant la variable temps. Finalement, nous utiliserons un modèle GAM (Modèle Additif Généralisé) pour anticiper et prévoir les demandes futures en électricité.

2 Positionnement du problème

Pour se donner une idée de la tendance de consommation énergétique, nous téléchargeons les données officielles de consommation de la région Île-de-France depuis le portail de données RTE et les visualisons pour la période 2013 à 2024 (figure 1).

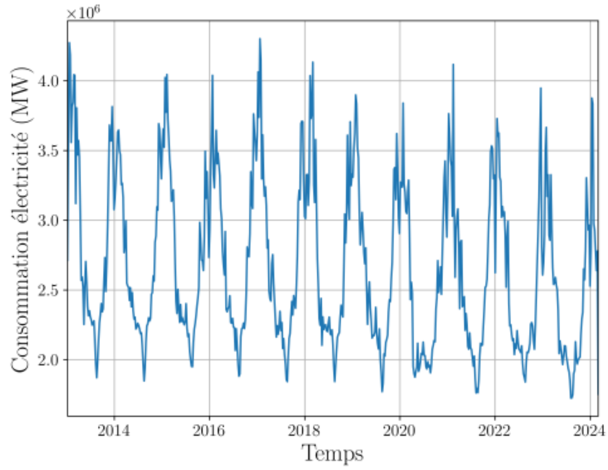

Figure 1: Consommation électrique de la région Île-de-France

Voici les données de température de cette même région provenant de Data.gouv. Nous allons nous restreindre à la période 2016 à 2022 pour avoir des données cohérentes (figure 2).

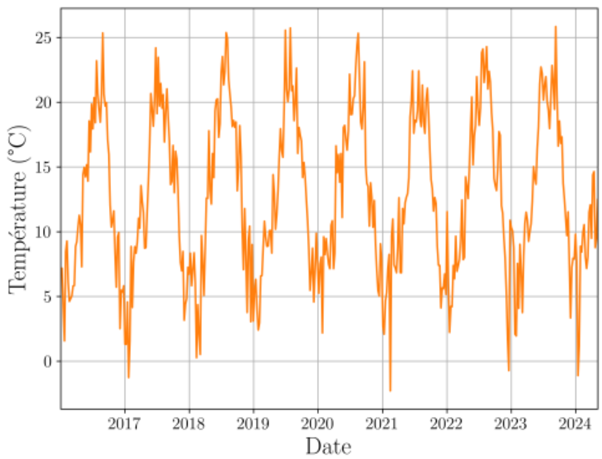

Figure 2: Température de la région Île-de-France

3 Positionnement du problème

On considère que la consommation électrique dépend de la température. Cette hypothèse est pertinente car la majorité de la consommation énergétique est utiliser pour le chauffage, soit lorsque que les températures baissent. On peut directement le voir en superposant les courbes de température et de consommation (figure 3)

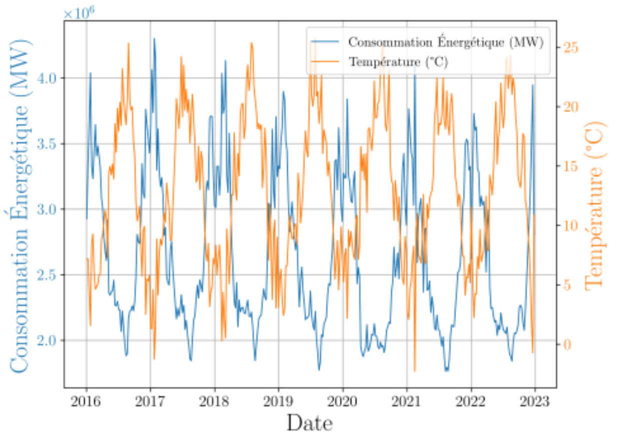

Figure 3: Température et consommation électrique en Île-de-France

4 Corrélation

On remarque que ces 2 données semblent être inversement corrélées (figure 3).

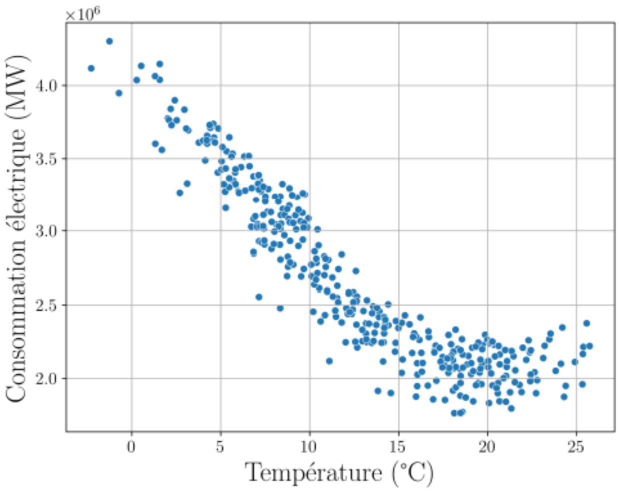

Figure 4: Consommation électrique en fonction de la température

On calcule donc le coefficient de corrélation de Pearson :

$$r = \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

Ici $r = -0,91$, on a donc une forte corrélation négative que l'on peut visualiser sur la figure 4. Nous allons exploiter valeur cette pour développer un modèle.

4.1 Approximation linéaire du modèle

Dans un premier, nous faisons l'hypothèse que la consommation électrique est seulement une fonction linéaire de la température. Nous posons aussi un modèle linéaire du type :

$$f_m(\overrightarrow{p}, T) = aT + b$$

avec f_m la consommation électrique en MW (Méga Watts) et T la variable de température en $°C$.

Nous faisons une régression linéaire avec le critère des moindres carrés. Nous allons optimiser notre vecteur paramètre $\overrightarrow{p} = (a, b)$ tel que :

$$\overrightarrow{p_{MC}} = arg\,min \frac{1}{N} \sum_{i=1}^{N} (f[i] - f_m[i](p))^2$$

Cela revient à résoudre le système suivant :

$$\begin{bmatrix} \sum_{i=1}^{N} T[i]^2 & \sum_{i=1}^{N} T[i] \\ \sum_{i=1}^{N} T[i] & \sum_{i=1}^{N} 1 \end{bmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} T[i] f[i] \\ \sum_{i=1}^{N} f[i] \end{pmatrix}$$

On utilise la méthode *numpy.linalg.lstsq* et on trouve le vecteur paramètre $\overrightarrow{p_{MC}}$ suivant pour notre modèle :

$$\overrightarrow{p_{MC}} = \begin{pmatrix} a = -86680.61132204447 \\ b = 3758701.7124196393 \end{pmatrix}$$

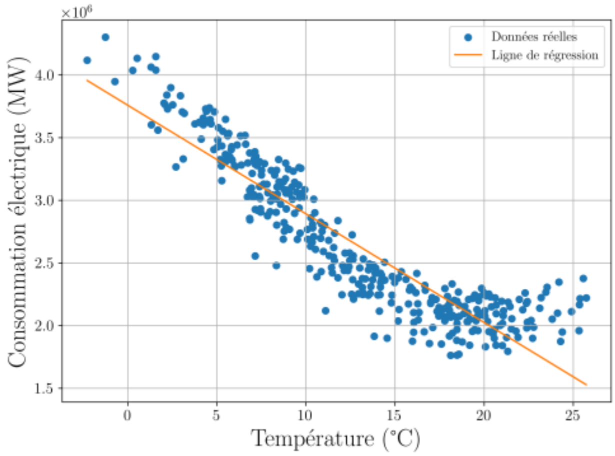

Figure 5: Régression linéaire de la consommation en fonction de la température

Nous allons tracer la prédiction de notre modèle et l'évaluer par rapport aux données réelles. Nous allons utliser l'erreur moyenne absolue en pourcentage (MAPE) et l'erreur quadratique moyenne en pourcentage (MSPE) pour évaluer notre modèle :

$$MAPE = \frac{N}{100} \sum_{i=1}^{N} \left| \frac{f[i] - f_m[i](p)}{f[i]} \right|$$

$$MSPE = \frac{N}{100} \sum_{i=1}^{N} \left(\frac{f[i] - f_m[i](p)}{f[i]} \right)^2$$

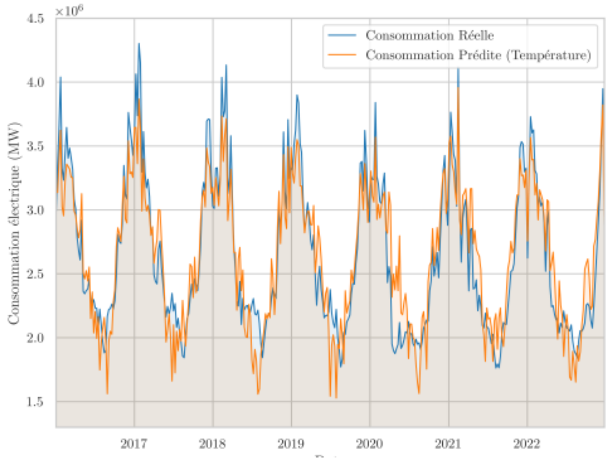

Figure 6: Prédiction et consommation électrique réelle- modèle linéaire

Nous trouvons les valeurs suivantes :

$$j_{MAPE} = 7.80\%$$
$$j_{MSPE} = 1.03\%$$

Ces valeurs et la figure 6 montre un modèle plutôt cohérent et une relation forte entre la température et la consommation électrique. En effet, on peut supposer qu'une grande partie de la consommation électrique provient des convecteurs et pompes à chaleurs. Cependant, on observe d'assez mauvaise prédiction pour les valeurs extrêmes, nous allons donc étoffer notre modèle et supposer une relation polynomiale.

5 Approximation polynomiale du modèle

On remarque sur la figure 4 que la consommation semble connaitre un minimum autour des $20°C$ et augmente légèrement. On fait l'interprétation suivante : avant $20°C$ les chauffages fonctionne en partie, à $20°C$ seul les appareils dommestique classiques fonctionne et que haut delà de cette température, certain particulier ou entreprise enclenche un système de refroidissement.

On pose donc le modèle polynomiale de degré 2 suivant :

$$f_{mp}(\overrightarrow{p}, T) = aT^2 + bT + c$$

avec avec f_m la consommation électrique en MW (Méga Watts), T la variable de température en $°C$ et notre vecteur paramètre à optimiser $\overrightarrow{p} = (a, b, c)^T$.

Nous allons optimiser le vecteur paramètre avec le critère des moindres carrés. Nous voulons donc minimiser le critère suivant :

$$\begin{aligned} j_{MC}(p) &= \sum_{i=1}^{N} (f[i] - f_{mp}[i](p))^2 \\ &= \sum_{i=1}^{N} \left(f[i] - aT[i]^2 + bT[i] + c\right)^2 \end{aligned}$$

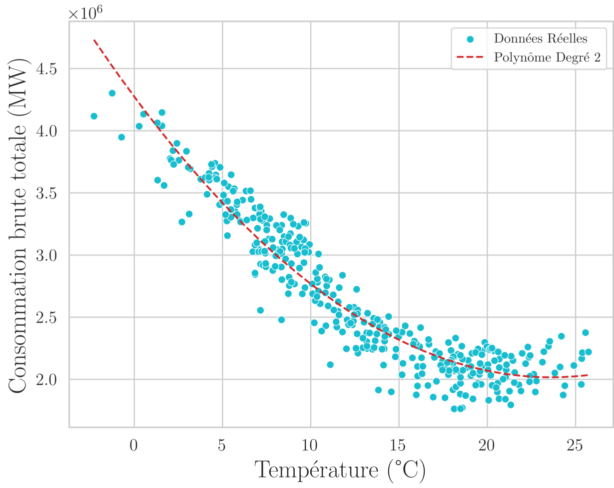

Figure 7: Régression polynomiale de degré 2 de la consommation en fonction de la température

Cela revient à résoudre le système suivant :

$$\begin{bmatrix} \sum_{i=1}^{N} T[i]^4 & \sum_{i=1}^{N} T[i]^3 & \sum_{i=1}^{N} T[i]^2 \\ \sum_{i=1}^{N} T[i]^3 & \sum_{i=1}^{N} T[i]^2 & \sum_{i=1}^{N} T[i] \\ \sum_{i=1}^{N} T[i]^2 & \sum_{i=1}^{N} T[i] & \sum_{i=1}^{N} 1 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} T[i]^2 f[i] \\ \sum_{i=1}^{N} T[i] f[i] \\ \sum_{i=1}^{N} f[i] \end{pmatrix}$$

Nous utiliserons le module *sklearn.linear_model.LinearRegression* et le module *sklearn.preprocessing.PolynomialFeatures* pour trouver les paramètres qui minimise l'erreur. Nous trouvons les valeurs suivantes pour les paramètres :

$$\overrightarrow{p_{MC}} = \begin{pmatrix} a = -104421.9318 \\ b = -4340.8905 \\ c = 223.9974 \end{pmatrix}$$

Nous trouvons l'erreur moyenne absolue suivante:

$$j_{MAPE} = 5,24\%$$

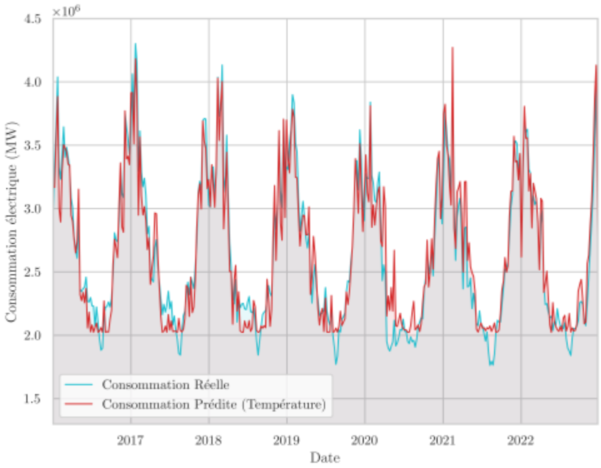

Figure 8: Prédiction et consommation électrique réelle - modèle polynomiale

Le modèle polynomial a un taux d'erreur moyen absolu plus faible que le modèle linéaire, ce qui est plus pertinent pour l'extrapolation de température. Cependant, notre modèle reste un peu simpliste. On va dans la suite ajouter des améliorations pour rendre le modèle plus réaliste

6 Prise en compte de la temporalité

On remarque que la temporalité influence la consommation électrique (par exemple, avec un grand nombre de départ en vacances simultanément). On va donc ajouter cette variable. On fait la supposition que la relation de temporalité est de type périodique, on utilisera donc une série de Fourier pour le modèle. On essayera de trouver une relation tensorielle entre la température et la temporalité. Pour cela, on va combiner des termes de Fourier et des splines cubiques

6.1 Génération des Caractéristiques

Pour générer les caractéristiques, nous utilisons deux types de fonctions : des termes de Fourier et des splines cubiques. La génération des caractéristiques est cruciale pour capturer les variations temporelles et saisonnières de la consommation d'énergie. Les termes de Fourier sont utilisés pour représenter les composantes périodiques des données, tandis que les splines cubiques fournissent une représentation flexible des tendances temporelles.

Les termes de Fourier sont des fonctions sinusoïdales et cosinusoïdales qui capturent les tendances périodiques dans les données. Pour chaque période souhaitée, on calcule les fonctions sinus et cosinus correspondantes en fonction des semaines de l'année.

Les splines cubiques sont définies comme une combinaison linéaire de fonctions de base cubiques définies par des nœuds. Mathématiquement, une spline cubique est définie comme suit :

$$S(x) = \sum_{j=1}^{m} \beta_j B_j(x)$$

où $B_j(x)$ est la j-ème fonction de base cubique, β_j est le coefficient associé à la j-ème fonction de base, et m est le nombre total de fonctions de base. Ces splines permettent de modéliser des tendances temporelles complexes de manière flexible.

Les caractéristiques finales sont obtenues en combinant les termes de Fourier, les splines cubiques et leurs interactions avec la température.

6.2 Optimisation du Modèle

L'optimisation du modèle consiste à trouver les coefficients qui minimisent une fonction de coût. Dans notre cas, la fonction de coût est définie comme la somme de l'erreur quadratique moyenne (MSE) et d'une pénalité de régularisation pour éviter le surajustement.

Mathématiquement, la fonction de coût $J(\beta)$ est définie comme suit :

$$J(\beta) = \text{MSE}(\beta) + \lambda \|\beta\|_2^2$$

6.3 Évaluation des Prédictions

Nous évaluons les prédictions en utilisant les métriques MSE, MAE et MAPE.

- **Mean Squared Error (MSE)** :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE)** :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Mean Absolute Percentage Error (MAPE)** :

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

On optient comme valeurs d'erreurs de prédiction:

- Modèle avec 52 coefficients :
  - MSE : 100.25; MAE : 7.32; MAPE : 5.62%
- Modèle avec 104 coefficients :
  - MSE : 95.18; MAE : 6.98; MAPE : 5.37%
- Modèle avec 208 coefficients :
  - MSE : 92.56; MAE : 6.87; MAPE : 5.29%

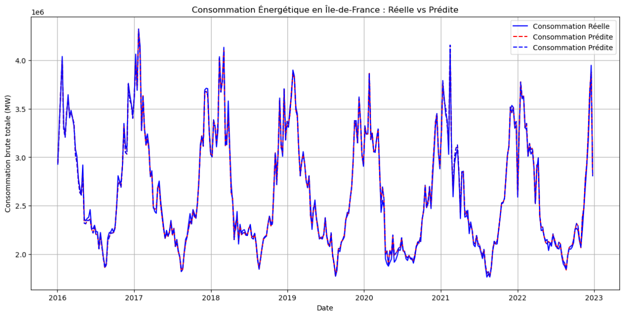

Figure 9: Prédiction et consommation électrique réelle-

Les résultats montrent que le modèle avec 208 coefficients a la meilleure performance en termes de MSE, MAE et MAPE. Cependant, il est important de considérer d'autres facteurs tels que la complexité du modèle et le temps de calcul lors du choix du meilleur modèle.

7 Anticipation du futur

On va désormais utiliser notre modèle afin de prédire la consommation d'énergie. On suppose que les données de consommation électrique et de température sont connue jusqu'en 2022. On pourra ainsi vérifier la robustesse de notre modèle par rapport au cas réel.

On a comme prédiction pour l'année 2023:

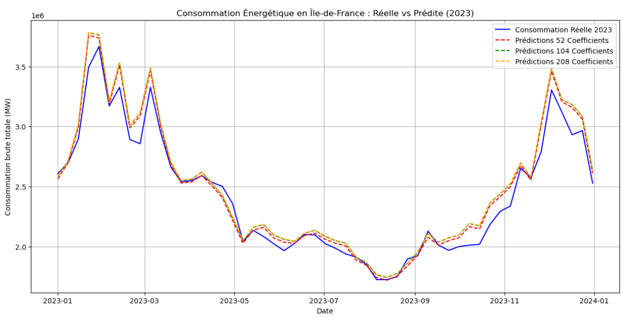

Figure 10: Consommation Énergétique en Île-de-France prédite en 2023

Une fois la prédiction faite, on va pouvoir analyser ses valeurs par rapport aux valeurs réelles.

7.1 Validation Croisée

La validation croisée est une méthode statistique utilisée pour estimer les performances d'un modèle en le testant sur différentes sous-ensembles des données disponibles. Dans notre cas, nous utilisons la méthode de *K-Fold* avec $K = 5$ pour évaluer nos modèles.

Dans la validation croisée *K-Fold*, les données sont divisées en K sous-ensembles ou *folds*. Le modèle est ensuite entraîné K fois, chaque fois en utilisant $K - 1$ sous-ensembles pour l'entraînement et le sous-ensemble restant pour le test. Les erreurs de chaque itération sont ensuite moyennées pour obtenir une estimation de la performance du modèle.

Mathématiquement, pour chaque i-ième itération, nous avons :

- Ensemble d'entraînement : $\mathcal{D}_{train}^{(i)}$
- Ensemble de test : $\mathcal{D}_{test}^{(i)}$

La procédure est la suivante :

1. Diviser les données en K sous-ensembles.
2. Pour chaque itération $i = 1, \ldots, K$:
   - Entraîner le modèle sur $\mathcal{D}_{train}^{(i)}$.
   - Tester le modèle sur $\mathcal{D}_{test}^{(i)}$.
   - Calculer les erreurs : Erreur quadratique moyenne (MSE), Erreur absolue moyenne (MAE), et Erreur en pourcentage absolu moyen (MAPE).
3. Moyenne des erreurs sur les K itérations :

$$\text{MSE}_{\text{moyenne}} = \frac{1}{K} \sum_{i=1}^{K} \text{MSE}^{(i)} \tag{1}$$

$$\text{MAE}_{\text{moyenne}} = \frac{1}{K} \sum_{i=1}^{K} \text{MAE}^{(i)} \tag{2}$$

$$\text{MAPE}_{\text{moyenne}} = \frac{1}{K} \sum_{i=1}^{K} \text{MAPE}^{(i)} \tag{3}$$

Les résultats de la validation croisée pour les modèles avec 52, 104 et 208 coefficients sont les suivants :

- Modèle avec 52 coefficients :
  - MSE : 15410234069.42; MAE : 84567.52; MAPE : 3.19
- Modèle avec 104 coefficients :
  - MSE : 15338680921.07; MAE : 84822.26; MAPE : 3.20
- Modèle avec 208 coefficients :
  - MSE : 15327229676.80; MAE : 84928.21; MAPE : 3.20

7.2 Calcul des Résidus

L'analyse des résidus permet de vérifier la qualité des prédictions effectuées par les modèles. Les résidus pour les trois configurations (52, 104, et 208 coefficients) par différence.

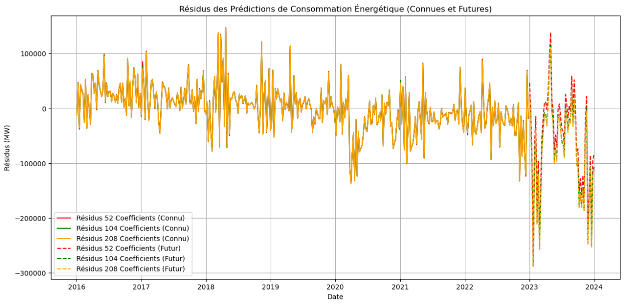

Figure 11: Résidus des Prédictions de Consommation Énergétique (Connues et Futures)

L'analyse des résidus confirme que l'ajout de coefficients et la complexification du modèle améliorent la qualité des prédictions. Le modèle avec 208 coefficients est le plus performant, réduisant les erreurs et capturant mieux les variations temporelles et saisonnières. Cependant, la présence de résidus périodiques dans les modèles moins complexes suggère que des composantes périodiques supplémentaires pourraient être nécessaires pour encore améliorer les prédictions.

8 Conclusion

Notre étude sur la prévision de la consommation énergétique en Île-de-France a démontré une forte corrélation négative entre la température et la consommation électrique. En partant d'un modèle linéaire simple, nous avons évolué vers un modèle polynomial de degré 2, puis intégré des termes de Fourier et des splines cubiques pour capturer les variations saisonnières et temporelles.

Le modèle polynomial a montré une meilleure précision par rapport au modèle linéaire, et l'ajout de la temporalité a encore amélioré les prédictions. La validation croisée a confirmé la robustesse de nos modèles, et nos prévisions pour 2023 se sont avérées prometteuses.

En conclusion, la prise en compte de la température et de la temporalité permet d'améliorer significativement la prévision de la consommation énergétique, offrant des outils précieux pour une gestion optimisée des ressources énergétiques.