



University  
Mohammed VI  
Polytechnic



UNIVERSITÉ MOHAMMED VI  
POLYTECHNIQUE

## EMINES - SCHOOL OF INDUSTRIAL MANAGEMENT

### PROJET STATISTIQUES RAPPORT

---

# Prévision à court terme de la consommation de gaz en Belgique - Secteur retail

---

#### *Réalisé par:*

Hayat AGNAOU  
Anass BOUATRA  
Nezar EL MESSNAOUI  
Nouamane OUBELKASS

#### *Encadré par :*

Saad BENJELLOUN  
Fayçal JAMALI

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Collecte et pré-traitement des données</b>	<b>2</b>
<b>3</b>	<b>Analyse exploratoire des données</b>	<b>3</b>
<b>4</b>	<b>Analyse de régression linéaire</b>	<b>6</b>
<b>5</b>	<b>Analyse de la transformation de Fourier</b>	<b>12</b>
5.1	Transformation de Fourier pour une seule fréquence . . . . .	13
5.2	Généralisation sur les fréquences principales . . . . .	14
<b>6</b>	<b>Discussion et conclusion</b>	<b>17</b>
<b>7</b>	<b>Pistes pour de futures recherches</b>	<b>17</b>
<b>8</b>	<b>Annexe</b>	<b>18</b>

## 1 Introduction

La consommation de gaz dans le secteur de la distribution publique en Belgique représente un élément essentiel de la gestion énergétique quotidienne du pays. Dans cette optique, ce projet vise à élaborer des modèles de prévision à court terme pour anticiper cette consommation avec précision.

Focalisant sur la consommation de gaz dans la zone gaz H retail en Belgique, notre objectif est de développer des outils de prévision efficaces pour répondre aux besoins opérationnels des entreprises du secteur de l'énergie. Nous examinerons attentivement les nominations quotidiennes des allocations de gaz par les expéditeurs au gestionnaire du réseau de transport, dans le but d'améliorer la précision des prévisions et l'efficacité de la gestion des allocations de gaz.

La prévision à court terme de la demande en gaz naturel est une activité fondamentale pour les entreprises du secteur de l'énergie, car elle permet une gestion opérationnelle plus efficace. Cette activité englobe des prévisions horaires et quotidiennes pour un maximum de sept jours à l'avance, facilitant ainsi la planification des achats de gaz au quotidien.

Les avancées technologiques ont permis le développement de modèles de prévision plus sophistiqués, allant des méthodes traditionnelles aux modèles basés sur l'intelligence artificielle. Ces modèles visent à comprendre les tendances de consommation et à améliorer la précision des prévisions pour une gestion énergétique plus efficace.

Dans ce contexte, notre étude se concentre sur l'utilisation de méthodes de régression linéaire et de transformation de Fourier pour modéliser la consommation de gaz en fonction de variables telles que la température, les jours de la semaine et les jours fériés. En combinant ces approches, nous cherchons à obtenir des prévisions plus précises de la consommation de gaz à court terme dans le secteur de la distribution publique en Belgique.

Ce rapport est structuré de manière à présenter d'abord une analyse exploratoire des données, suivie de la méthodologie utilisée pour développer nos modèles de prévision. Nous présentons ensuite les résultats de nos analyses et discutons de leurs implications pour la gestion de la consommation de gaz dans le secteur de la distribution publique.

## 2 Collecte et pré-traitement des données

La source des données utilisées dans cette étude est un ensemble de données détaillant la consommation de gaz dans le secteur de la distribution publique en Belgique, dans la zone gaz H retail. Les données sont collectées à partir de diverses sources, notamment les relevés des compteurs de gaz et les systèmes de suivi des fournisseurs et des gestionnaires de réseau de transport. Ces données comprennent plusieurs colonnes décrivant différents aspects temporels et contextuels de la consommation de gaz.

Les variables incluent des informations sur la date et l'heure de la consommation, avec des colonnes telles que "year\_utc", "month\_utc", "day\_utc", "hour\_utc", ainsi que des variables binaires pour indiquer le jour de la semaine ("Monday" à "Sunday"). De plus, des colonnes binaires spécifient la présence de jours fériés ("Holiday") et de vacances scolaires ("SchoolHoliday"), ainsi que des jours spéciaux tels que les ponts ("BridgeMonday", "BridgeFriday") et les jours fériés individuels comme Noël, Nouvel An, Pâques, etc.

Des indicateurs de périodes de pause ("FallBreak", "ChristmasBreak", "NewYearBreak", "WinterBreak", "EasterBreak", "SummerBreak") sont incluses. De plus, des données météorologiques sont fournies, telles que la température, le taux d'humidité, la

pluviométrie, la vitesse du vent, et le rayonnement solaire, qui sont collectées à différents moments de la journée et dans des délais variables.

Nous avons ajouté des variables saisonnières "winter", "spring", "summer", et "fall" qui n'étaient pas incluses dans le dataset.

La variable dépendante de notre étude est "vol\_mwh", qui indique la consommation de gaz en mégawattheures (MWh). Cette variable, utilisée comme mesure principale pour évaluer la consommation de gaz dans le secteur retail en Belgique et pour développer nos modèles de prévision, a nécessité la gestion des valeurs manquantes. Pour ce faire, nous avons opté pour la méthode de backward filling, qui consiste à remplacer les valeurs manquantes par la dernière valeur observée dans la colonne "vol\_mwh".

Il convient de noter que le dataset que nous avons utilisé était déjà prétraité et ne nécessitait pas de prétraitement supplémentaire. Toutes les variables pertinentes étaient correctement formatées et prêtes à être utilisées dans nos analyses et nos modèles de prévision.

### 3 Analyse exploratoire des données

Notre analyse exploratoire des données a été articulée autour de plusieurs axes pour mieux appréhender les tendances de consommation de gaz. Premièrement, nous avons étudié l'évolution temporelle de la consommation horaire de gaz sur une période s'étalant du 1er janvier 2007 au 1er avril 2014. Les résultats ont révélé une tendance périodique, avec une diminution progressive de l'amplitude de la courbe entre 2011 et 2014, suggérant une dynamique saisonnière.

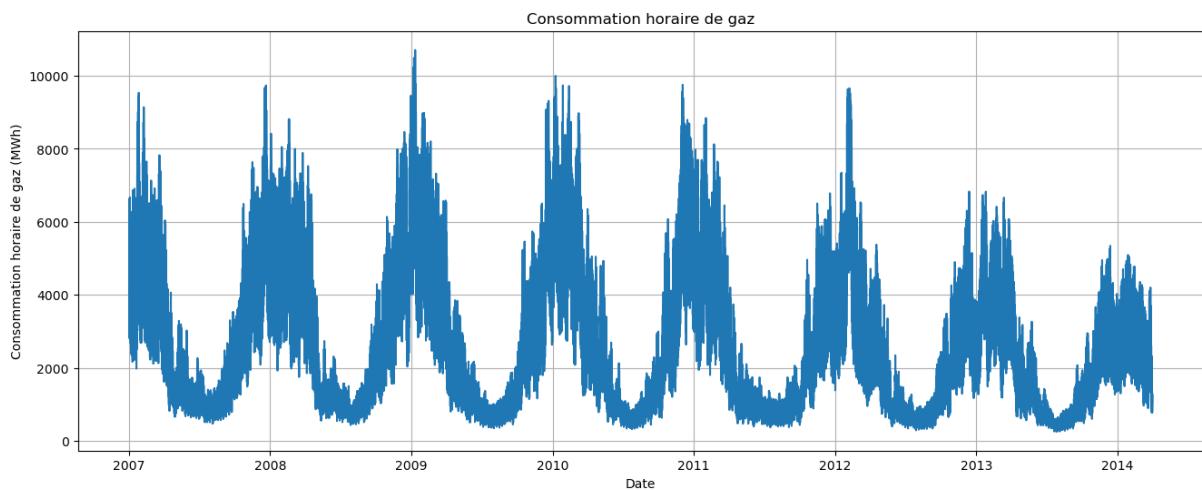


FIGURE 1 – Consommation horaire de gaz

Ensuite, nous avons étudié l'évolution de la consommation moyenne de gaz par heure du jour par chaque année (fig. 2), illustrant davantage les différences entre les années, ainsi que le profil de l'évolution de la consommation de gaz sur un jour.

Après avoir examiné la consommation moyenne de gaz par heure pour chaque année, nous avons poursuivi notre analyse en explorant la relation entre la consommation de gaz et la température. Ce faisant, nous avons pris en compte le tracé du paragraphe précédent et avons eu l'idée de traiter chaque heure de manière séparée. Pour ce faire, nous avons

### 3 ANALYSE EXPLORATOIRE DES DONNÉES

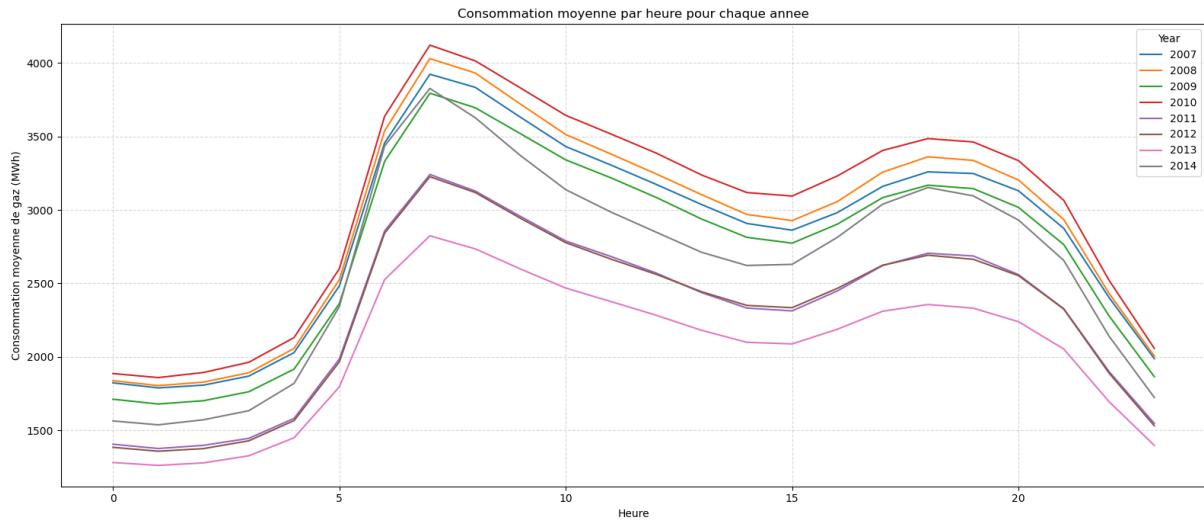


FIGURE 2 – Consommation moyenne par heure pour chaque année

créé 24 nuages de points, un pour chaque heure de la journée (fig. 16 dans l'annexe). Ces graphiques ont mis en évidence une corrélation linéaire entre la consommation de gaz et la température, montrant une stabilisation de la consommation après avoir dépassé un seuil critique de température (un exemple de chaque période du jour dans la figure 3). De plus, visuellement, nous avons remarqué des pentes différentes pour chaque heure, soulignant ainsi l'importance de traiter les données de manière spécifique à chaque période horaire.

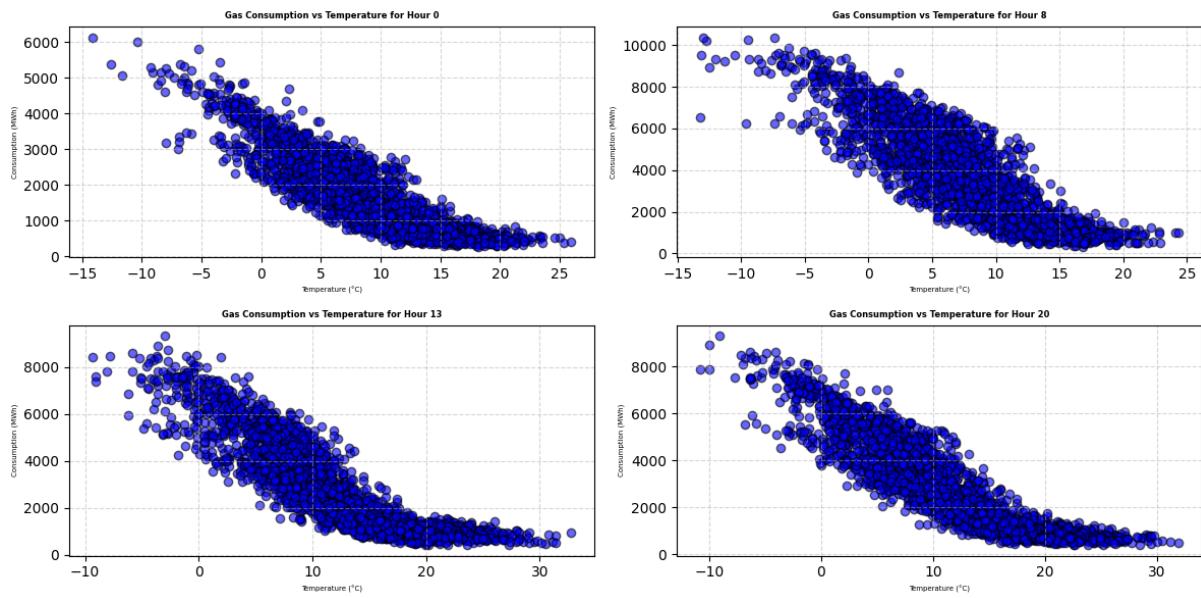


FIGURE 3 – Consommation horaire de gaz vs température par heure

Par la suite, nous avons examiné la consommation de gaz par saison, à travers quatre scatter plots (fig. 4), mettant en évidence des variations saisonnières significatives. En parallèle, nous avons comparé les jours ouvrables et les weekends pour chaque année (fig. 5a), ainsi que si le jour est férié ou non (fig. 5b) et si la variable "heating" est 0 ou 1 (fig. 5c). Ces analyses ont souligné des différences significatives dans les niveaux de consommation en fonction de ces variables, avec une chute marquée de la consommation

entre 2011 et 2014.

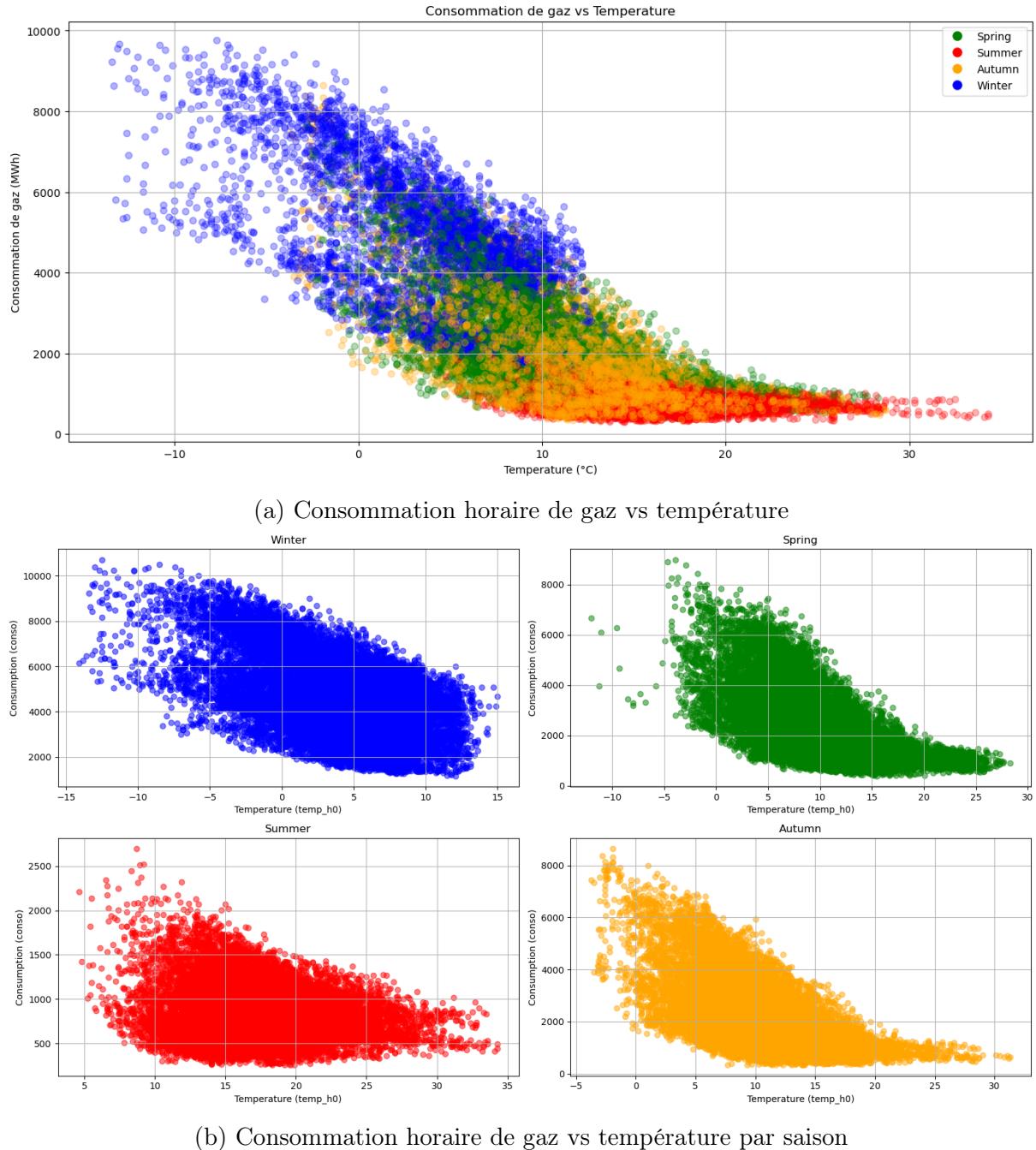


FIGURE 4 – Consommation horaire de gaz vs température

En résumé, notre analyse des données a révélé des tendances saisonnières distinctes et des variations significatives de la consommation de gaz tout au long de la journée. Ces observations soulignent l'importance de tenir compte à la fois des aspects saisonniers et des variations horaires dans notre modèle de prédiction. Ainsi, nous explorerons deux approches distinctes pour modéliser et prédire la consommation de gaz : l'analyse de Fourier pour capturer les tendances saisonnières et la régression linéaire appliquée à chaque heure de la journée, en tenant compte également d'autres paramètres tels que les weekends, les vacances, le chauffage, etc.

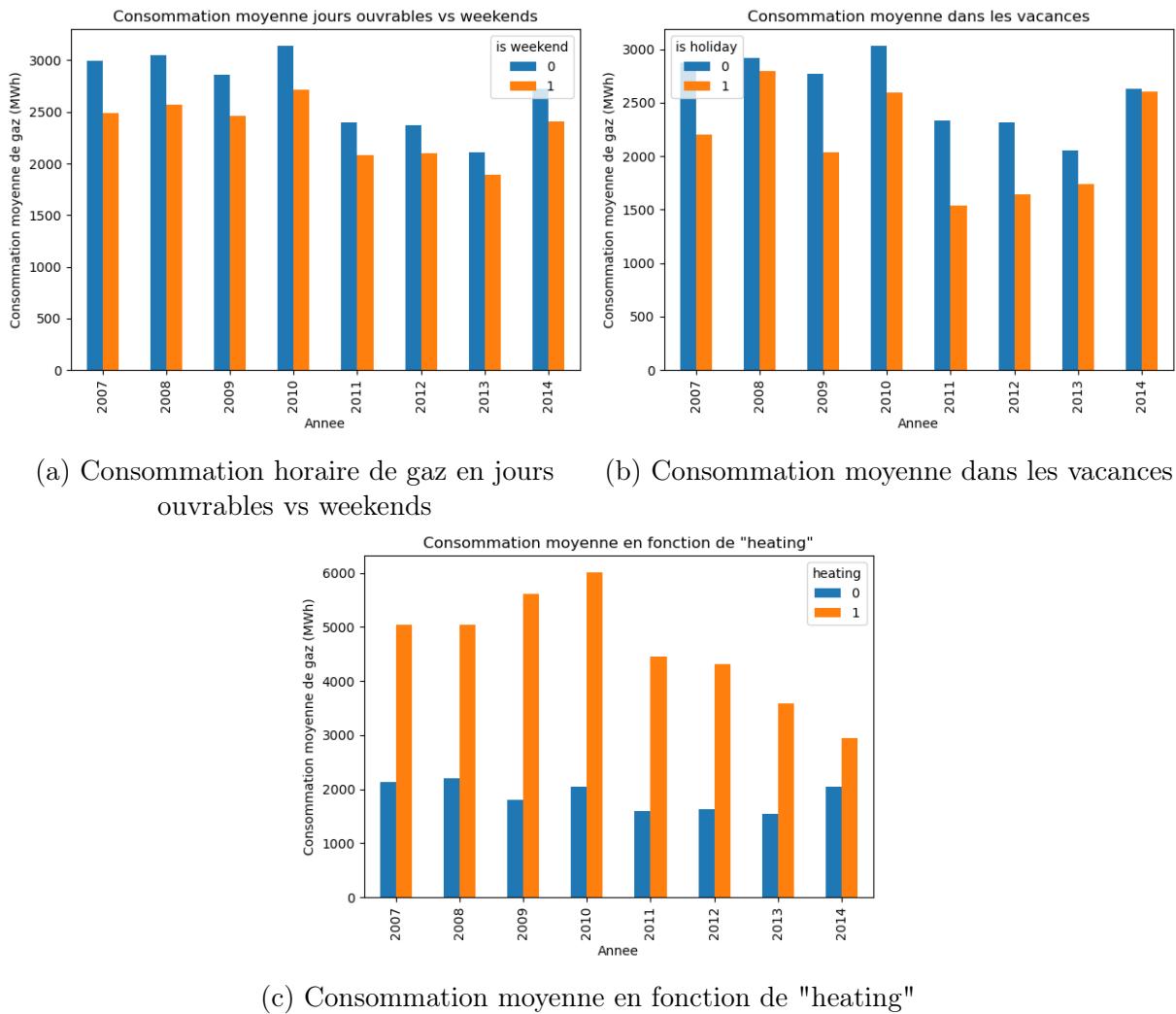


FIGURE 5 – Impact de quelques variables sur la consommation moyenne de gaz pour chaque année

## 4 Analyse de régression linéaire

Nous débuterons notre analyse de régression avec un modèle qui prendra comme variable dépendante la consommation de gaz, et comme variable indépendante la température pour une heure "temp\_h0", qui agira comme baseline. La figure 6 contient un exemple de chaque période du jour (voir figure 17 pour l'intégralité des heures). Notons que notre coefficient de séparation pour l'ensemble d'entraînement et de test est de 0.8, ce qui signifie que les 80% des premières lignes sont incluses dans l'ensemble d'entraînement, tandis que les 20% des dernières lignes sont réservées à l'ensemble de test. Ceci est dû au fait que l'objectif de la régression est la prédiction, et cette division nous permet d'évaluer la performance du modèle sur des données non vues.

Ensuite, nous considérerons une caractéristique spécifiquement conçue pour capter l'effet de la température sur la consommation de gaz, que nous nommerons "clipped\_temp". À partir de nos observations lors de l'analyse exploratoire des données, nous avons constaté que la consommation de gaz tend à se stabiliser au-delà d'un certain seuil de température. Pour tenir compte de cette relation, nous procéderons à une itération sur toutes les heures

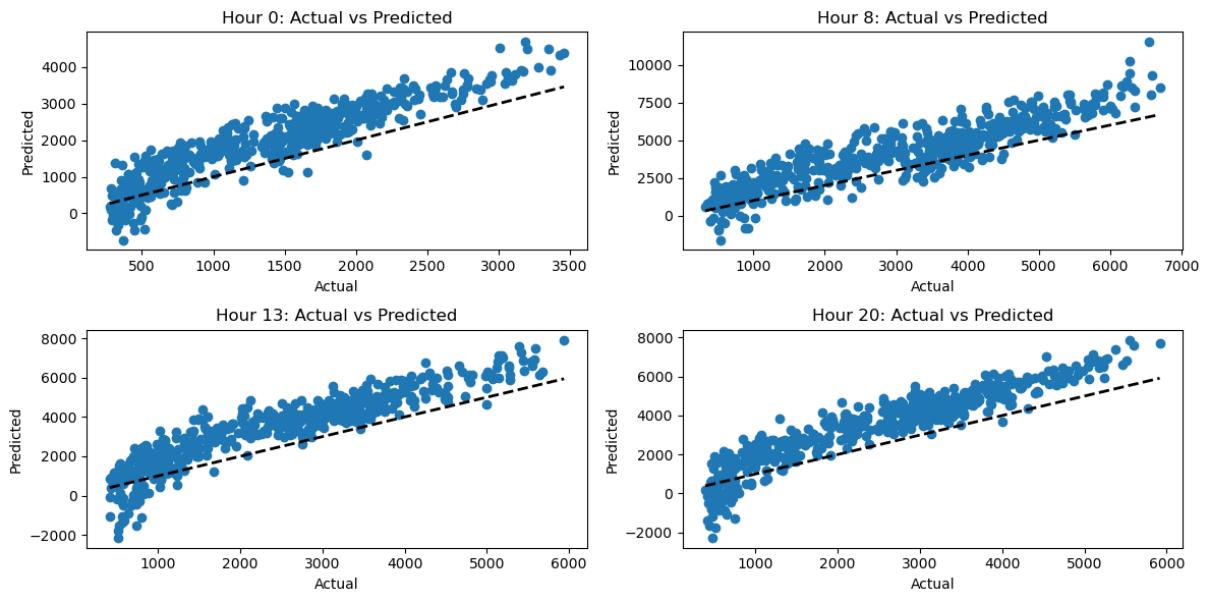


FIGURE 6 – Résultats de la régression linéaire avec "temp\_h0"

Metric	Value
R-squared	0.8194668
Adjusted R-squared	0.8193814
MSE	131867.9

TABLE 1 – Métriques de la régression avec "temp\_h0"

de la journée. Pour chaque heure, nous identifierons la température maximale au-delà de laquelle la consommation de gaz reste constante, en effectuant une boucle sur une plage de valeurs de température. Ensuite, pour chaque valeur de "clipped\_temp", nous exécuterons une régression linéaire avec la consommation de gaz comme variable dépendante. Nous sélectionnerons alors la valeur de "clipped\_temp" qui maximise le coefficient de détermination ( $R^2$ ), indiquant ainsi le modèle de régression linéaire le plus approprié pour chaque heure de la journée. La figure 7 contient un exemple de chaque période du jour (version intégrale : figure 18 en annexe).

Hour	Optimal T	Mean $R^2$
0	19.1	0.780307
8	19.9	0.764950
13	20.7	0.810536
20	22.8	0.827427

TABLE 2 – Température optimale et R-carré moyen

Une fois que nous aurons identifié la valeur optimale de "clipped\_temp" pour chaque heure de la journée, nous procéderons à une régression linéaire multiple pour prédire la consommation de gaz. Notre modèle inclura les variables suivantes en tant que prédicteurs : "clipped\_temp", "weekend" (0 ou 1), "Holiday", la saison (représentée par des

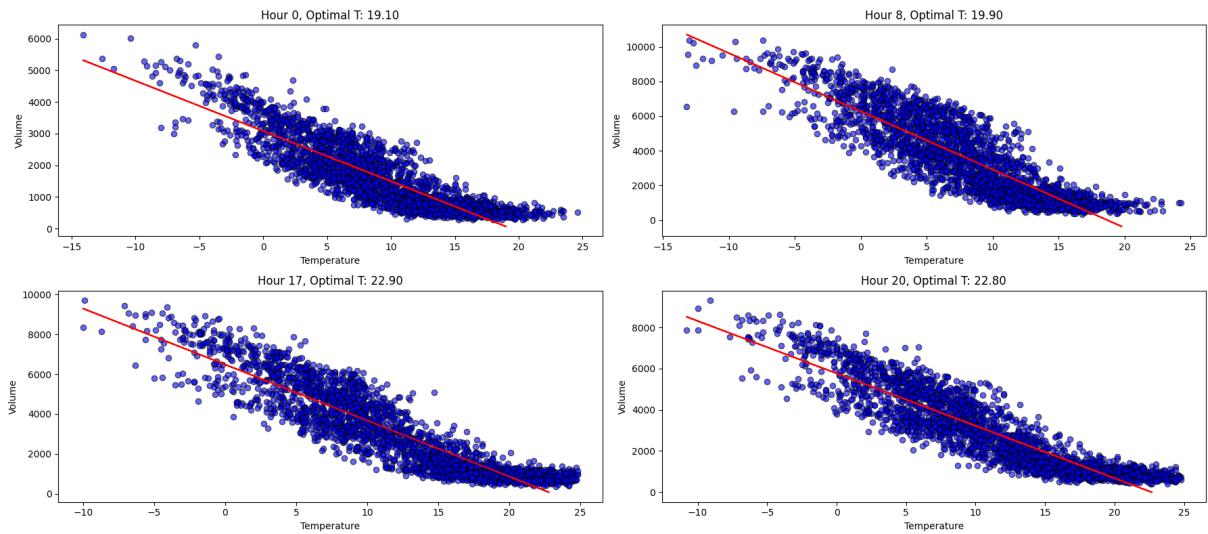


FIGURE 7 – Résultats de la régression linéaire avec "clipped\_temp"

variables binaires pour chaque saison, prenant la valeur 1 si le mois correspond à la saison et 0 sinon), et "heating" (0 ou 1). Nous utiliserons les métriques suivantes pour évaluer la performance de notre modèle : le coefficient de détermination ( $R^2$ ), le coefficient de détermination ajusté (adjusted  $R^2$ ) et l'erreur quadratique moyenne (MSE). De plus, nous introduirons deux autres mesures pertinentes pour ce type d'étude : la racine carrée de l'erreur quadratique moyenne (RMSE) et la pondération de l'erreur absolue moyenne (wMAPE).

Metric	Value
R-squared	0.8991734
Adjusted R-squared	0.8988844
MSE	135523
RMSE	1135.839
WMAPE	35.20167%

TABLE 3 – Moyennes des métriques

Après cette évaluation des performances du modèle on constate qu'il peut être amélioré davantage. Par conséquent, nous devons introduire d'autres caractéristiques. Après avoir identifié une forte dépendance temporelle dans nos données, nous avons décidé d'explorer davantage cette relation en examinant l'auto-corrélation de la consommation de gaz. Pour ce faire, nous avons utilisé à la fois l'auto-corrélation classique et l'auto-corrélation partielle. Initialement, nous avons tracé les fonctions d'auto-corrélation pour chaque heure de la journée, mais nous avons rencontré des difficultés dans l'interprétation des résultats obtenus. Par conséquent, nous avons également calculé les fonctions d'auto-corrélation partielle pour chaque heure.

Les graphiques obtenus (quatre exemples en fig. 8, versions intégrales : figures 19 et 20 en annexe) ont révélé des informations intéressantes : il est apparu que la valeur de la consommation de gaz à une heure donnée est fortement influencée par celle du jour précédent, ainsi que par celle précédent immédiatement le jour précédent (lag 1 et 2). Cette

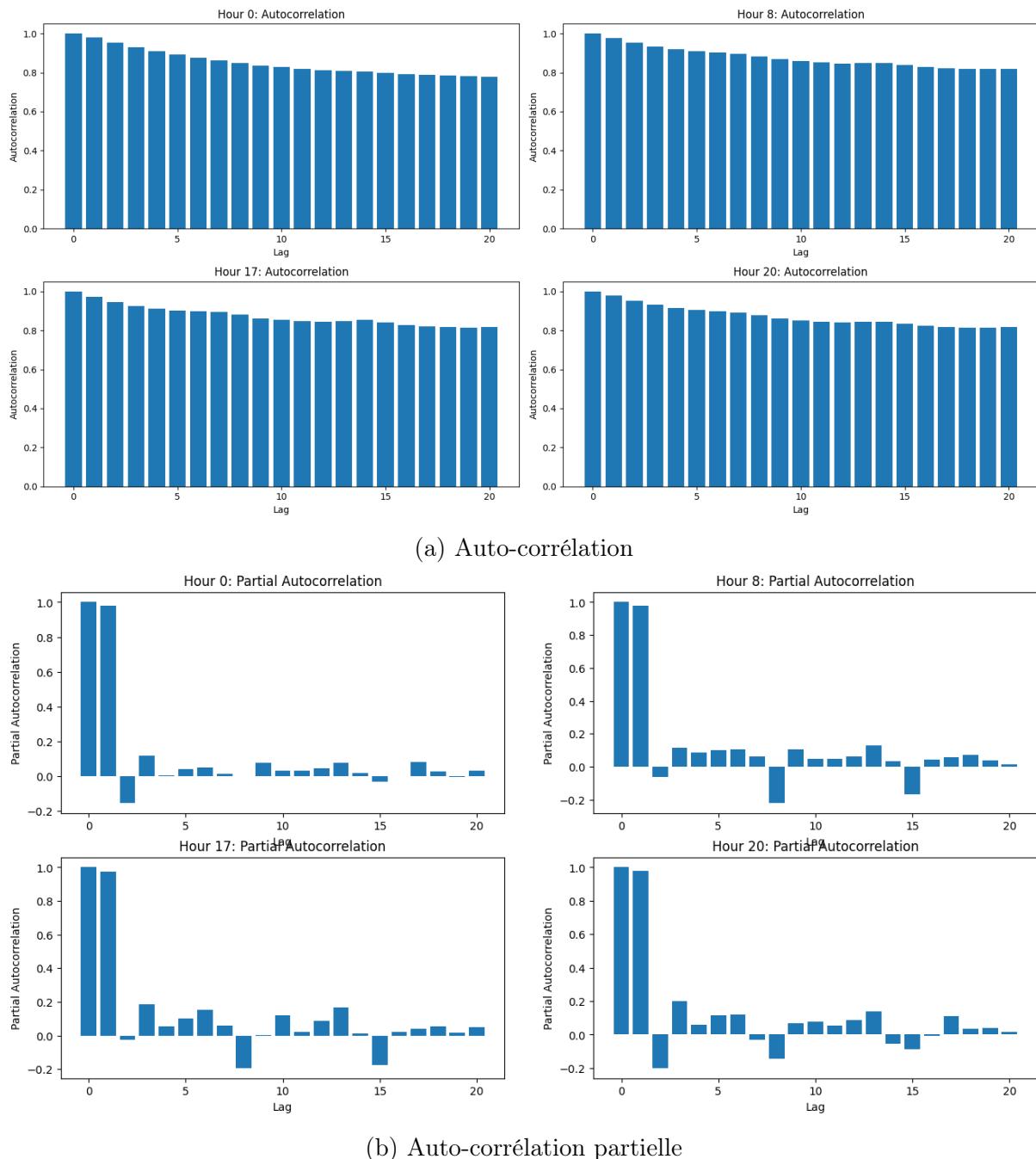


FIGURE 8 – Auto-corrélation et auto-corrélation partielle

observation suggère une forte dépendance temporelle dans les données, où les schémas de consommation semblent être fortement corrélés sur une base horaire. En tenant compte de ces résultats, nous avons décidé d'intégrer des variables de décalage temporel dans notre modèle de prédiction, qui donnent lieu à une amélioration décente dans la performance (table 4).

Toujours dans la dépendance temporelle, il est plausible que la consommation de gaz d'une heure soit liée à celle de l'heure précédente, comme l'indique la figure 9. Pour tester l'impact de ce paramètre, nous avons intégré la consommation de gaz de l'heure précédente dans notre modèle de régression linéaire. Nous évaluons ensuite la performance

Metric	Value
R-squared	0.898482
Adjusted R-squared	0.897331
MSE	265616.347058
RMSE	506.027390
WMAPE	15.8333 %

TABLE 4 – Métriques de la régression avec les décalages

de ce modèle pour déterminer son adéquation aux données (table 9 et figure 10, figure 21 pour la version intégrale).

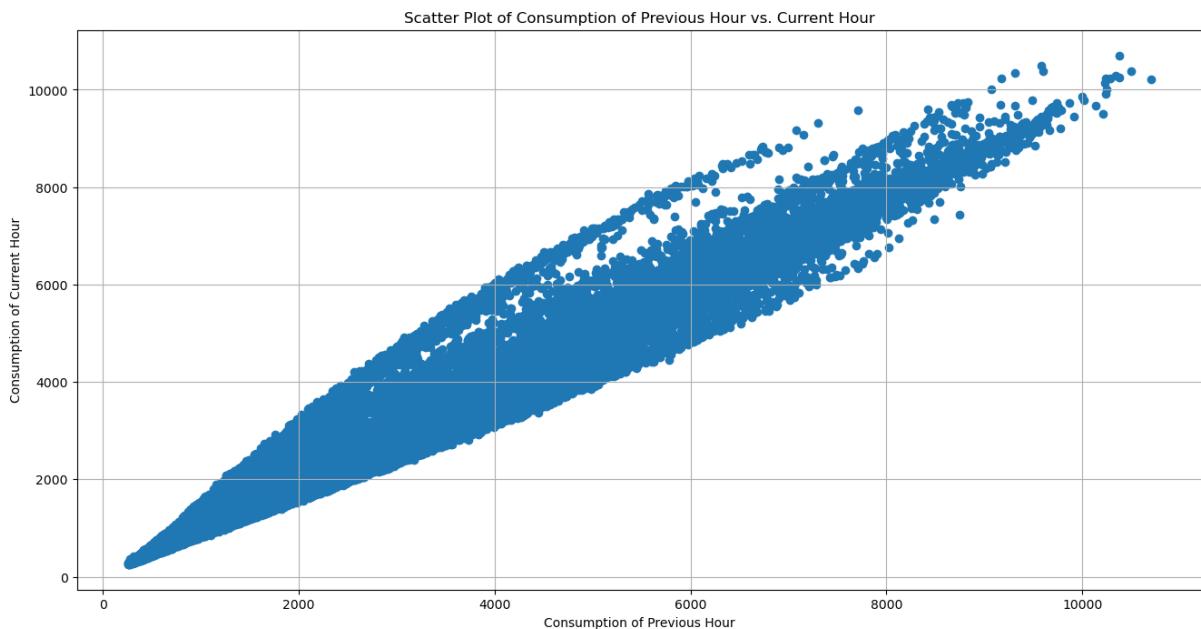


FIGURE 9 – Consommation de l'heure actuelle vs l'heure précédente

Metric	Value
R-squared	0.971585
Adjusted R-squared	0.971223
MSE	57325.875678
RMSE	227.767300
WMAPE	0.070000

TABLE 5 – Métriques de la régression avec l'heure précédente

D'après ces métriques, nous constatons qu'il existe encore des marges d'amélioration. Notre objectif étant de réduire le wMAPE à 2%. Après avoir observé que notre modèle présentait une baisse de performance pendant les heures comprises entre 11 heures et 17 heures, nous avons pris l'initiative de segmenter les données horaires en quatre parties distinctes de la journée (nuit, matin, après-midi, soir). Nous avons ensuite entrepris

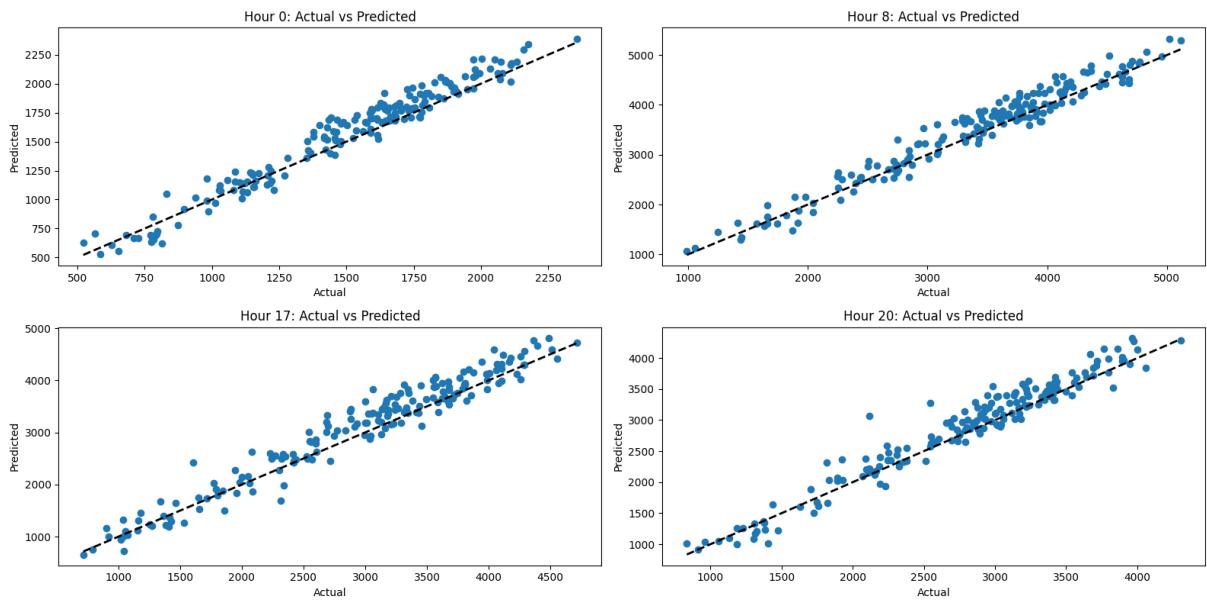


FIGURE 10 – Résultats de la régression linéaire avec l'heure précédente

d'optimiser nos caractéristiques en explorant plus de 195 colonnes pour identifier celles qui influencerait le plus notre variable cible. Notre démarche a débuté par la sélection des colonnes présentant une corrélation de 0.6 ou plus avec la variable cible. Ensuite, à l'aide d'une boucle, nous avons testé différentes combinaisons de 10 caractéristiques afin d'atteindre le meilleur score possible en termes de WMAPE

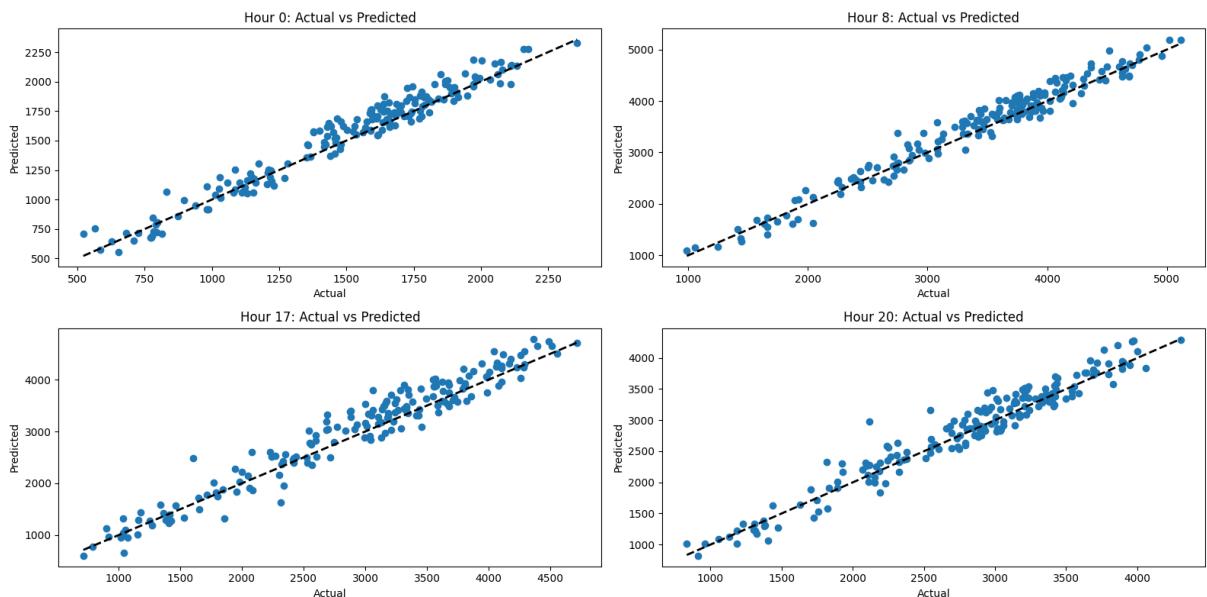


FIGURE 11 – Résultats de la régression linéaire finale

Metric	Value
R-squared	0.977820
Adjusted R-squared	0.977288
MSE	49375.033569
RMSE	207.880688
WMAPE	6.337500

TABLE 6 – Moyennes des métriques de la régression finale

## 5 Analyse de la transformation de Fourier

L’expansion en séries de Fourier d’un signal périodique est un outil bien connu pour l’analyse mathématique. Sous des hypothèses assez douces, une fonction périodique  $f(t)$  de période  $\tau$  peut être représentée comme une somme de fonctions sinusoïdales de périodes  $\tau/k$  ( soit de fréquences  $fk$ ), où  $k = 1, \dots, \infty$ , dans le sens où la série converge vers  $f(t)$  ponctuellement, sauf aux points de discontinuité de  $f(t)$ . Les sinusoïdes dont les périodes sont  $\tau/k$  sont appelées les harmoniques de rang  $k$ . Si la forme de  $f(t)$  est loin d’être une fonction sinusoïdale, les harmoniques supérieures peuvent être plus fortes que la variation principale. La formule pour l’expansion en séries de Fourier d’une fonction  $f(t)$  est donnée par :

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega t) + \sum_{n=1}^{\infty} b_n \sin(n\omega t)$$

avec les paramètres  $a_0$ ,  $a_n$  et  $b_n$  sont des paramètres à déterminer et

$$\omega = 2\pi/\tau = 2\pi f$$

Bien que théoriquement la somme soit infinie, en pratique, la force des harmoniques supérieurs est généralement négligeable.

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{N} [a_n \cos(n\omega t) + b_n \sin(n\omega t)]$$

Nous allons donc déterminer les coefficients  $a_0$ ,  $a_n$  et  $b_n$  pour la fréquence dominante en utilisant une régression linéaire, en optimisant l’ordre harmonique. Ensuite, nous généraliserons ces résultats aux dix fréquences les plus dominantes. Et pour faire ceci on crée un base de donnée de larne ci-dessous ( Table 1 )

$\cos(\omega_1 t)$	$\sin(\omega_1 t)$	$\cos(2\omega_1 t)$	$\sin(2\omega_1 t)$	...	$\cos(N\omega_{10} t)$	$\sin(N\omega_{10} t)$	consommation
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

TABLE 7 – Tableau pour faire la régression

Comme expliqué précédemment, nous avons commencé par déterminer les coefficients  $a_0$ ,  $a_n$  et  $b_n$  pour la fréquence dominante en utilisant une régression linéaire, en optimisant l’ordre harmonique.

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^N [a_n \cos(n\omega t) + b_n \sin(n\omega t)]$$

## 5.1 Transformation de Fourier pour une seule fréquence

Dans le cadre de notre étude, nous entamons notre analyse en nous focalisant initialement sur la fréquence présentant la plus grande magnitude, soit  $f = 0.000110$ . Cette fréquence correspond à une période de 1 an, ce qui en fait un élément central dans notre investigation. En examinant cette fréquence en premier lieu, nous cherchons à comprendre son impact et son rôle dans le phénomène étudié, jetant ainsi les bases essentielles pour une analyse approfondie.

Nous avons amorcé l'analyse en adoptant un ordre harmonique initial de 1. Ainsi, la fonction sous-jacente peut être représentée par la forme

$$f(t) = \frac{1}{2}a_0 + a_1 \cos(\omega t) + b_1 \sin(\omega t)$$

où  $a_0$ ,  $a_1$  et  $b_1$  sont les coefficients correspondants. En procédant de la sorte, nous avons appliqué une régression linéaire spécifiquement sur ces coefficients pour obtenir une estimation précise de leur valeur. Cette étape préliminaire revêt une importance cruciale, car elle nous permet de capturer les tendances initiales et les comportements fondamentaux du système étudié.

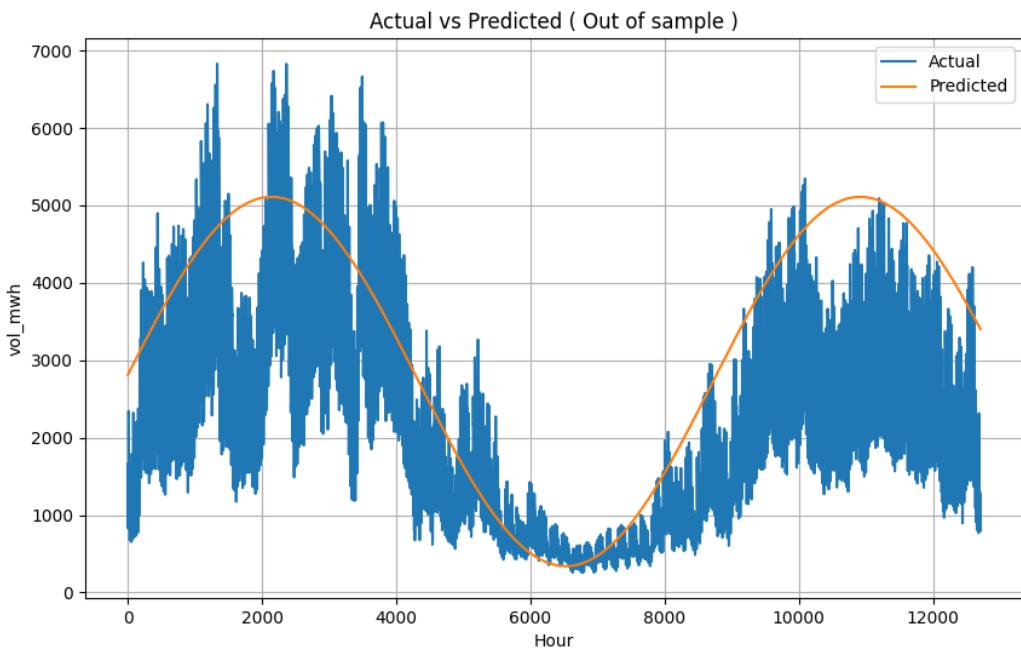


FIGURE 12 – Résultats de la régression linéaire sur les coefficients de Fourier

Une fois cette étape achevée, nous procédons à l'optimisation de l'ordre harmonique. Ce processus implique une évaluation systématique de la racine carrée de l'erreur quadratique moyenne (RMSE) tout en ajustant l'ordre harmonique. En intégrant cette mesure

de performance dans notre processus d'optimisation, nous sommes en mesure de sélectionner l'ordre harmonique optimal qui minimise l'erreur de prédiction tout en préservant la précision du modèle.

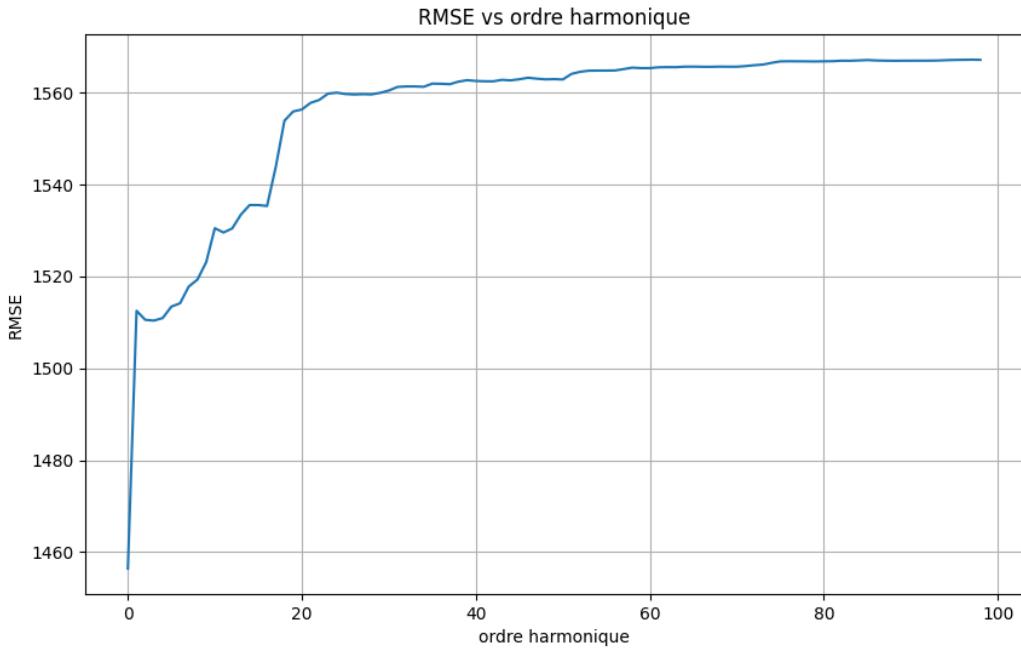


FIGURE 13 – Résultats de l'optimisation de l'ordre harmonique

Lors de notre analyse, nous avons obtenu une prédiction avec une erreur quadratique moyenne (RMSE) de 1456. Cette mesure de performance nous donne une indication de l'écart moyen entre les valeurs observées dans nos données et les valeurs prédites par notre modèle.

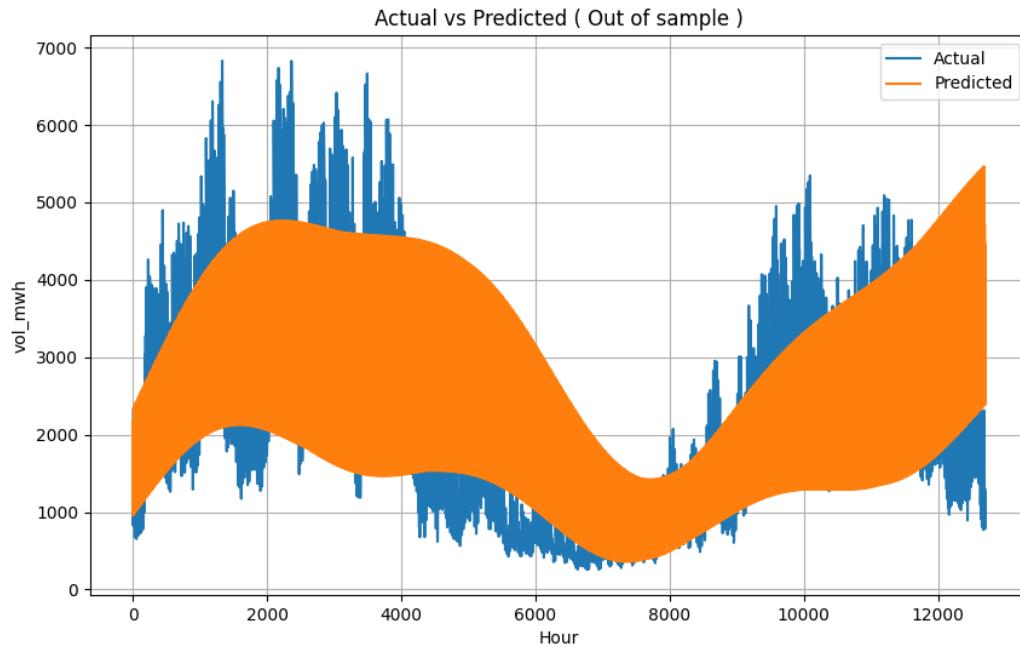
Metric	Value
MSE	57325.875678
RMSE	1456.4
WMAPE	66.2

TABLE 8 – Métriques de la régression des coefficients de fourier

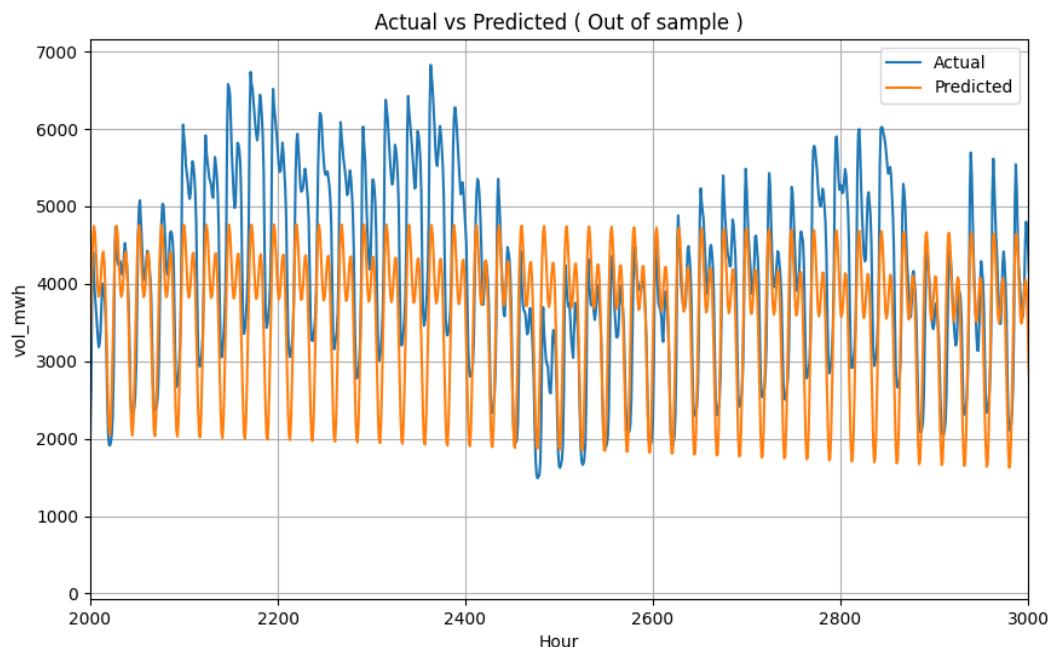
## 5.2 Généralisation sur les fréquences principales

Après avoir effectué cette procédure pour la fréquence correspondant à une année, nous étendrons notre analyse aux dix premières fréquences les plus significatives. De la même manière que précédemment, nous optimiserons l'ordre harmonique pour chaque fréquence, en ajustant les coefficients  $a_0$ ,  $a_n$  et  $b_n$  par le biais d'une régression linéaire. En intégrant cette optimisation dans notre méthodologie, nous cherchons à capturer de manière exhaustive les tendances et les comportements complexes présents dans les données,

afin de parvenir à une modélisation précise et robuste du phénomène étudié.



(a) Résultats de la régression linéaire



(b) Zoom entre 2000 et 3000

FIGURE 14 – Résultats de la régression linéaire sur les coefficients de Fourier

Lors de notre analyse, nous avons obtenu une prédition avec une erreur quadratique moyenne (RMSE) de 1046. Qui reste toujours très haut par rapport a ce qu'on a eu dans

l'autre méthode.

Metric	Value
RMSE	1046.2
WMAPE	62.05

TABLE 9 – Métriques de la régression des coefficients de fourier

De manière similaire à l'analyse effectuée pour une seule fréquence, nous avons également entrepris une optimisation de l'ordre harmonique. Cette étape visait à déterminer le niveau d'harmonicité le plus approprié pour représenter de manière précise la variation temporelle de notre phénomène.

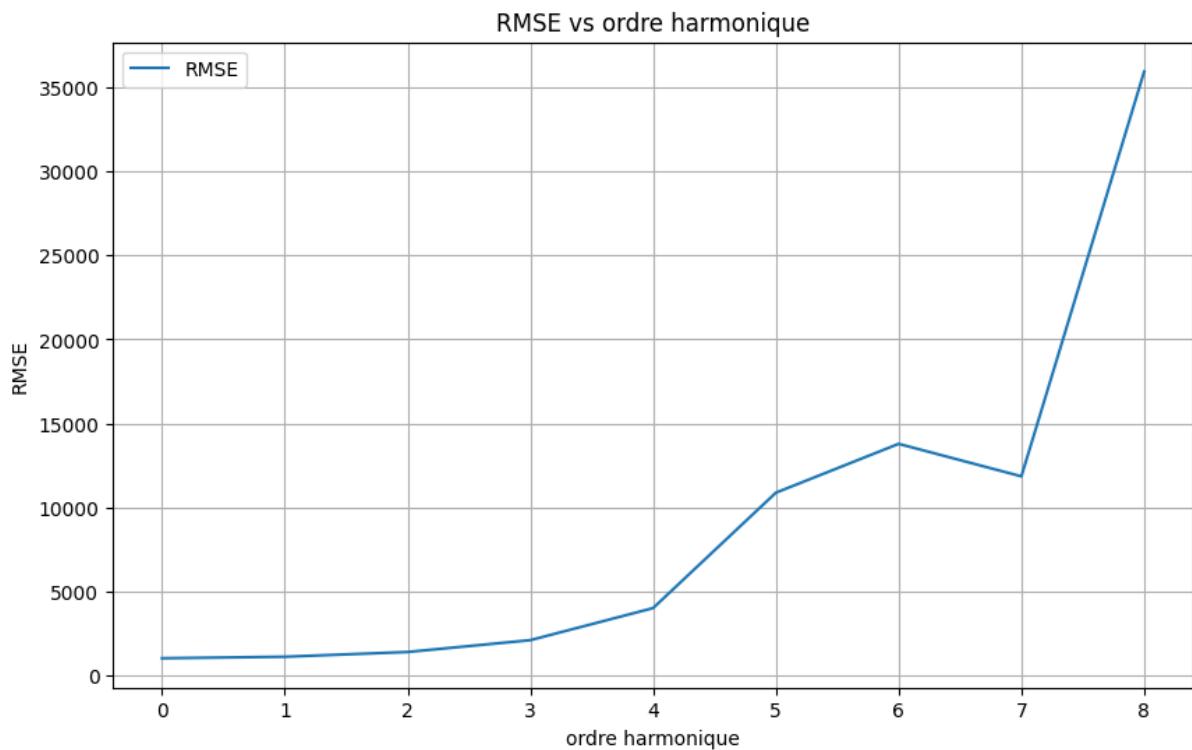


FIGURE 15 – Résultats de l'optimisation de l'ordre harmonique

Cependant, malgré l'utilisation d'une méthodologie rigoureuse et la recherche à travers un espace de paramètres variés, nous avons constaté que le meilleur résultat était toujours obtenu avec un ordre harmonique de 1. Cette observation peut suggérer plusieurs hypothèses, telles que la simplicité intrinsèque du phénomène étudié ou des limitations dans la détection de structures plus complexes dans nos données. L'incorporation de ces résultats dans notre méthodologie souligne l'importance de comprendre la nature spécifique du phénomène et la nécessité de s'adapter aux caractéristiques uniques de nos données pour obtenir des modèles précis et robustes.

## 6 Discussion et conclusion

Dans le cadre de notre projet de prédiction de la consommation de gaz en Belgique, nous avons exploré deux méthodes : la régression linéaire basée sur les coefficients de Fourier et la régression linéaire sur les valeurs à prédire, en tenant compte de facteurs tels que la température, les weekends et les vacances. Voici ce que nous avons appris :

1. *Performance actuelle* : Notre modèle actuel présente un MAPE (Mean Absolute Percentage Error) de 6,3 %. C'est un résultat solide, mais compte tenu du contexte spécifique de la prédiction de la consommation de gaz à court terme, nous pouvons encore améliorer la précision.
2. *Limitations identifiées* :
  - La complexité des 24 régressions linéaires distinctes peut rendre la gestion difficile.
  - Les données manquantes peuvent affecter la qualité des prédictions.
  - La stabilité des coefficients de Fourier peut être remise en question lors d'événements imprévus.
  - Les interactions entre variables doivent être soigneusement gérées.
  - La régression linéaire suppose une relation linéaire, ce qui peut limiter la précision.
  - La période historique couverte par les données peut influencer la robustesse des modèles.
3. *Perspectives d'amélioration* :
  - Explorer des modèles non linéaires pour capturer des relations plus complexes.
  - Approfondir l'analyse de la fréquence pour détecter des tendances saisonnières.
  - Segmenter les utilisateurs pour des modèles spécifiques.
  - Intégrer des données géospatiales pour tenir compte de la localisation.

En résumé, malgré nos bons résultats, nous devons rester vigilants face aux changements imprévisibles dans la consommation de gaz. Nous continuons à affiner nos modèles pour répondre aux exigences spécifiques de ce domaine. Si vous avez besoin de plus d'informations ou d'autres analyses, n'hésitez pas à nous solliciter !

## 7 Pistes pour de futures recherches

Nous avons identifié plusieurs pistes de recherche prometteuses pour améliorer nos modèles de prédiction de la consommation de gaz. Voici deux axes d'exploration :

1. **Évaluation de l'impact des facteurs externes** : Nous recommandons d'approfondir l'analyse de l'impact des événements externes sur la consommation de gaz. Cela inclut des éléments tels que les politiques énergétiques, les fluctuations des prix du gaz et d'autres facteurs économiques. En intégrant ces variables dans notre modèle, nous pourrions mieux comprendre leurs effets et améliorer la précision de nos prédictions.
2. **Validation croisée** : Pour évaluer la performance de nos modèles, nous devrions systématiquement utiliser la validation croisée. Cette approche nous permettra d'estimer leur capacité à généraliser à de nouvelles observations, en utilisant des données non vues. Une validation rigoureuse nous aidera à identifier les modèles les plus robustes.

3. **Segmentation des utilisateurs** : On peut Explorer la possibilité de segmenter les utilisateurs en groupes homogènes. Chaque segment peut avoir des comportements de consommation différents, ce qui pourrait nécessiter des modèles spécifiques.
4. **Inclusion de données géospatiales** : Si on dispose de données géospatiales (par exemple, localisation des utilisateurs), on peut les intégrer dans nos modèles. Notre prédiction va certainement dépendre de la région où l'on se trouve en Belgique.

## 8 Annexe

Dans l'annexe, vous trouverez les versions complètes des figures présentées dans ce rapport. Ces versions incluent tous les sous-graphiques des régressions linéaires effectuées pour chaque heure de la journée, offrant ainsi une vue plus détaillée des analyses réalisées. De plus, un lien vers le dossier OneDrive contenant le notebook utilisé pour l'analyse, le dataset original et toutes les figures est également fourni pour référence supplémentaire et pour permettre une reproductibilité totale des résultats.

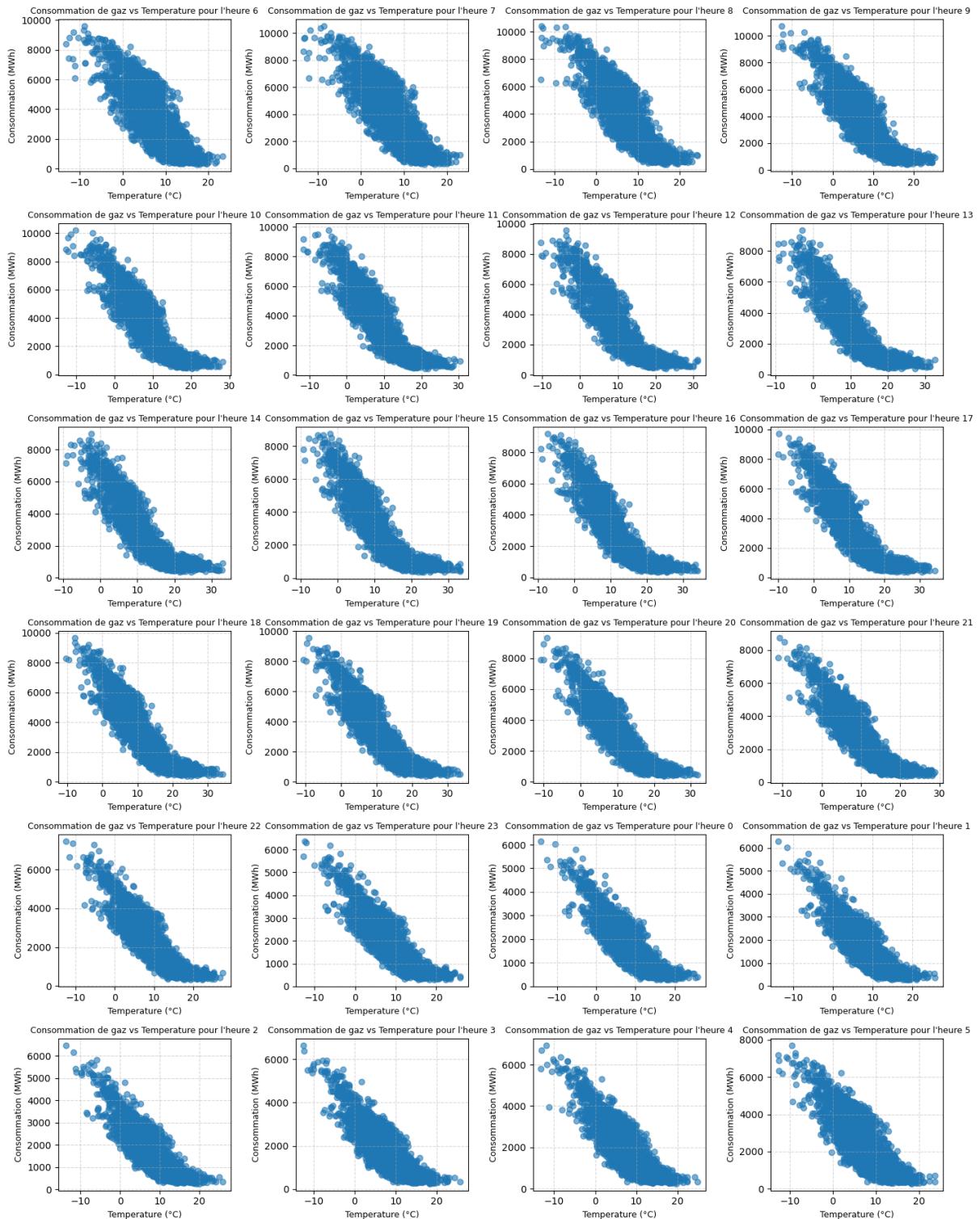


FIGURE 16 – Consommation horaire de gaz vs température par heure

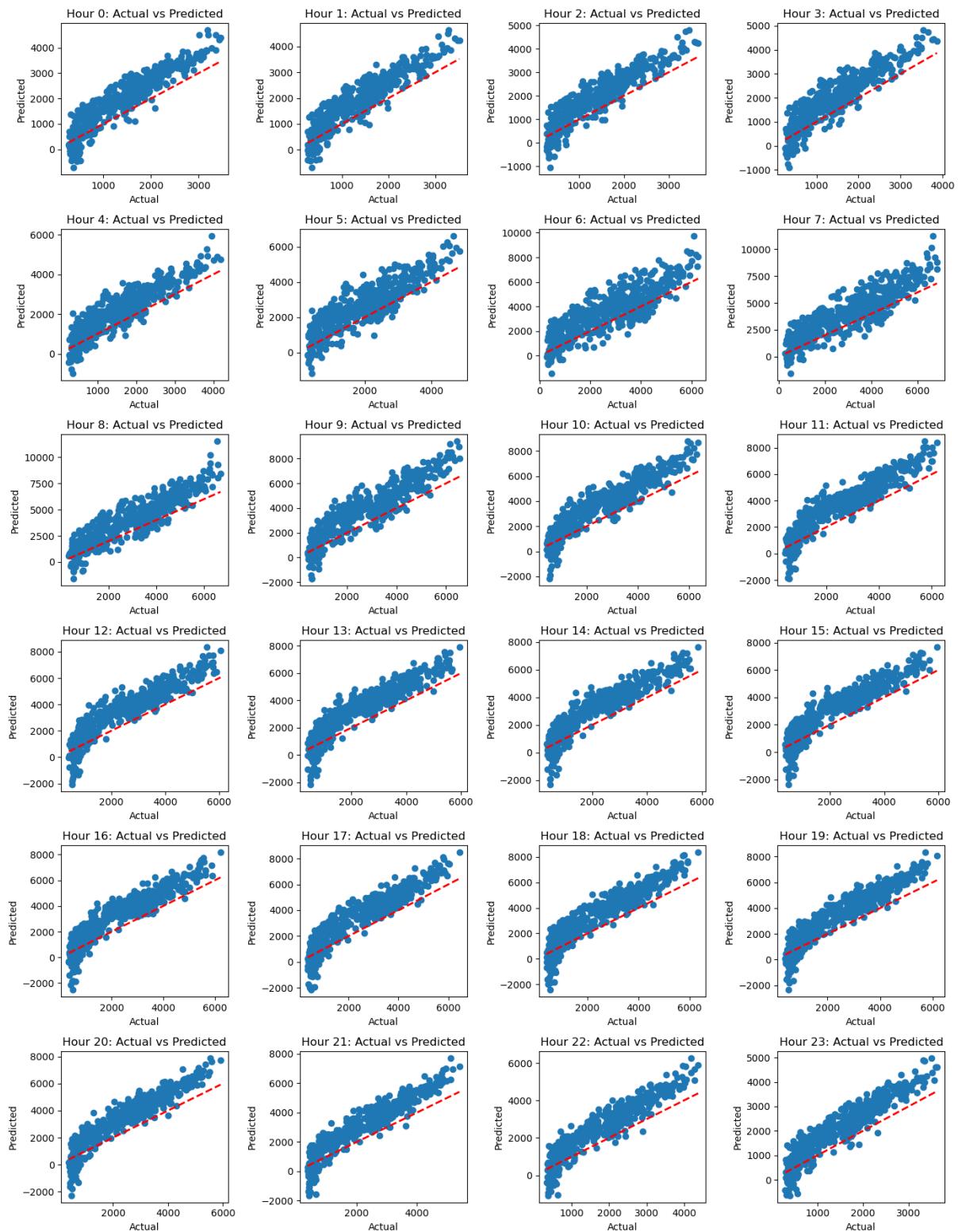


FIGURE 17 – Résultats de la régression linéaire avec "temp\_h0"

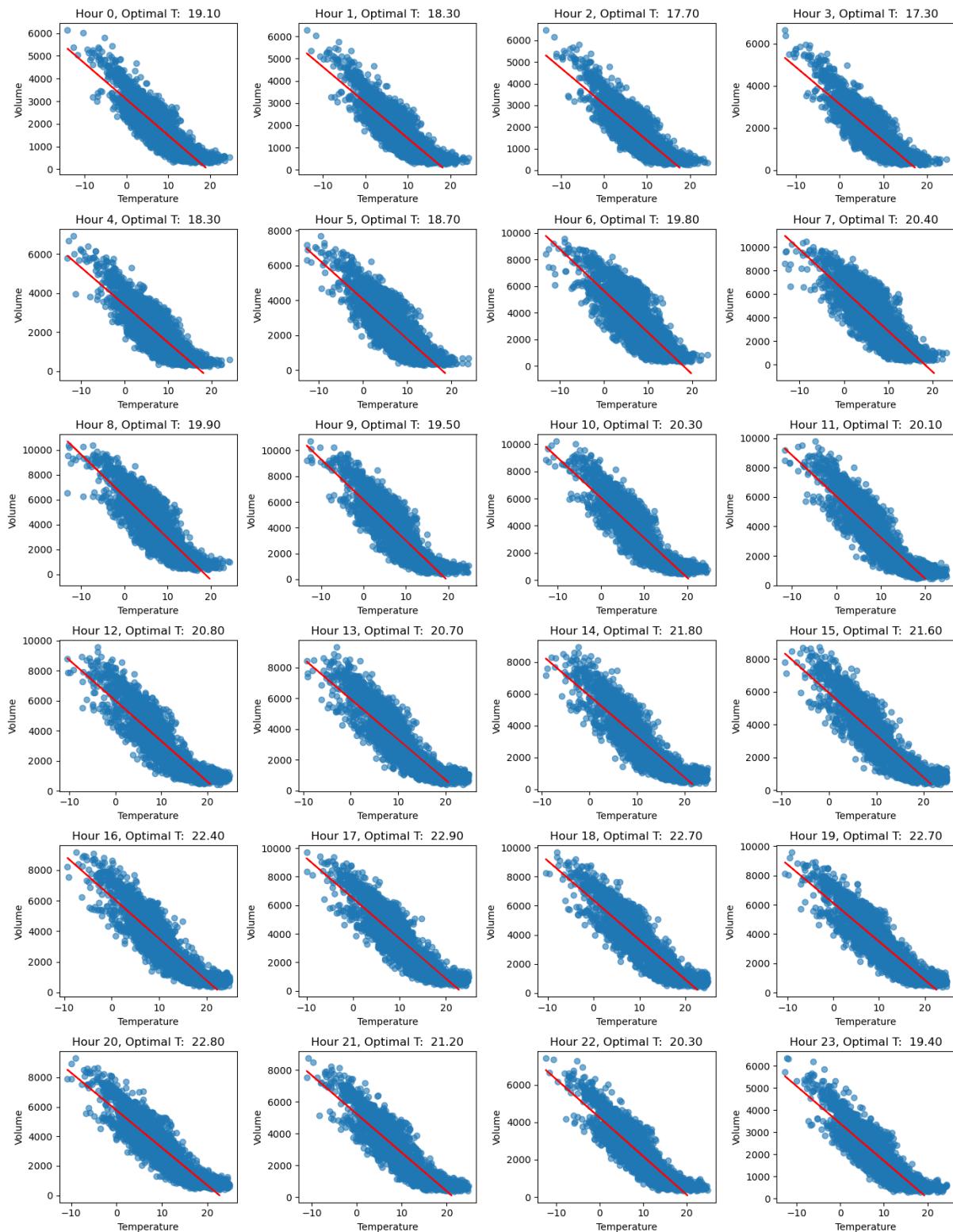


FIGURE 18 – Résultats de la régression linéaire avec "clipped\_temp"

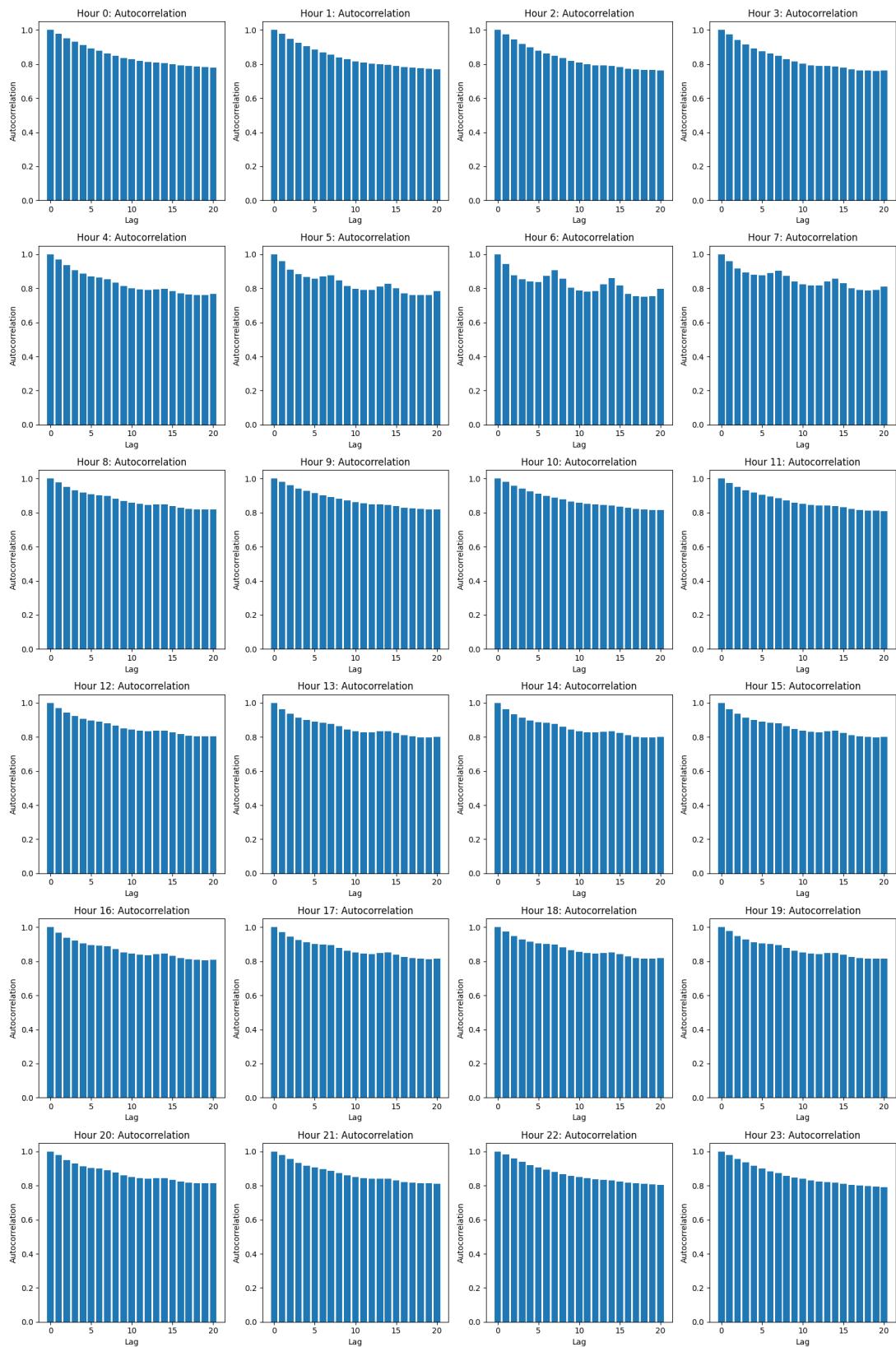


FIGURE 19 – Auto-corrélation

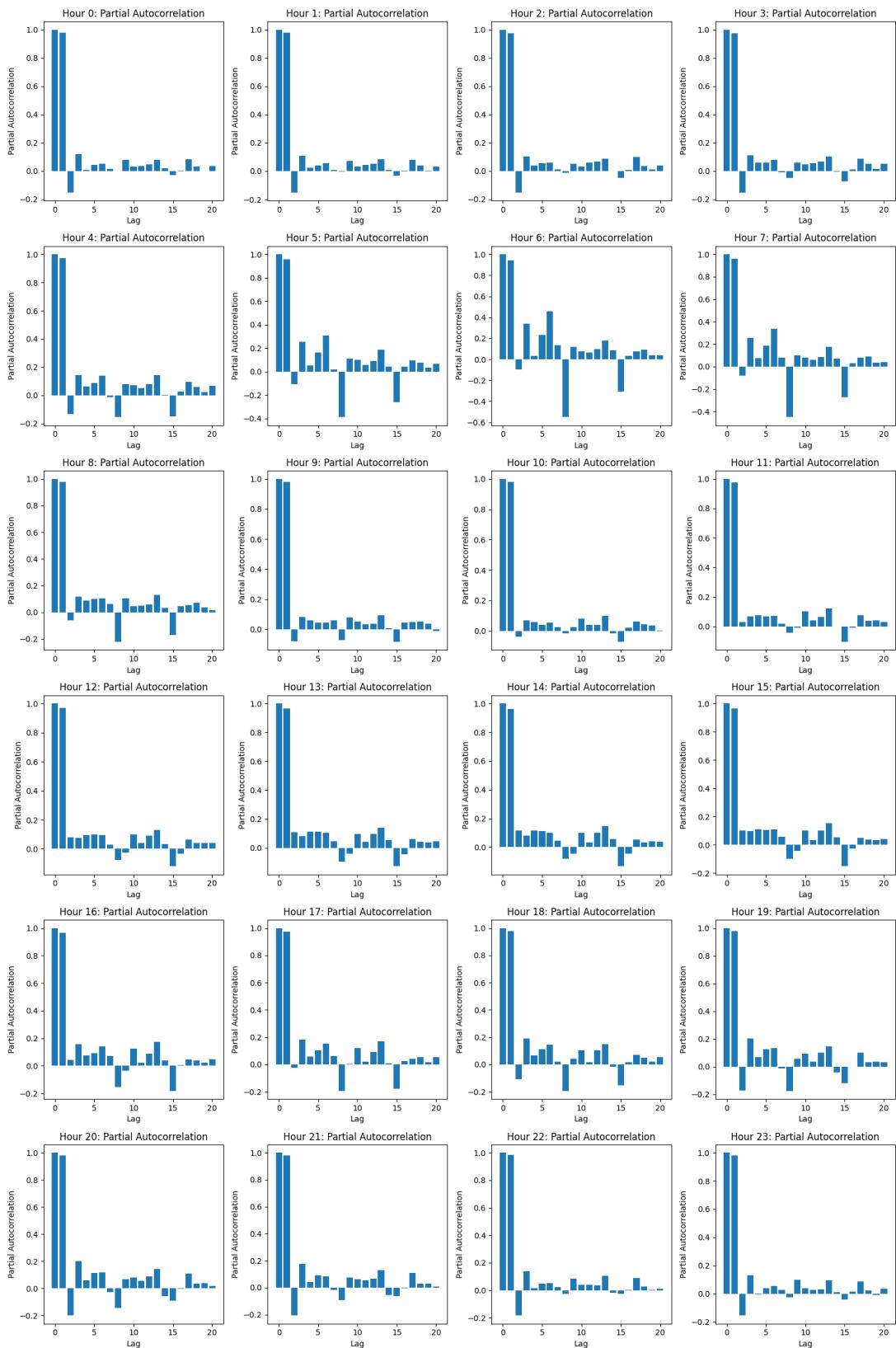


FIGURE 20 – Auto-corrélation partielle

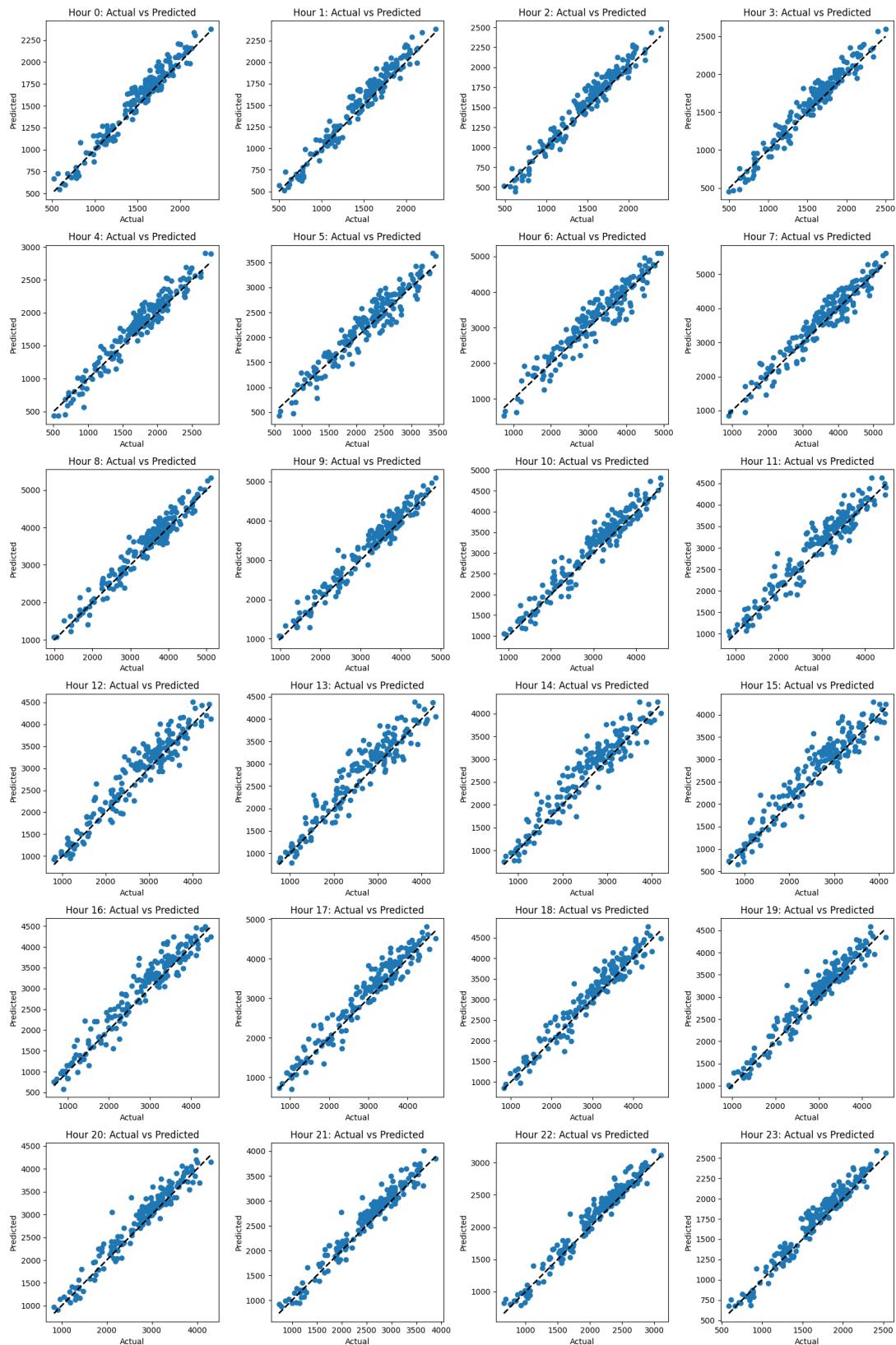


FIGURE 21 – Résultats de la régression linéaire avec l'heure précédente

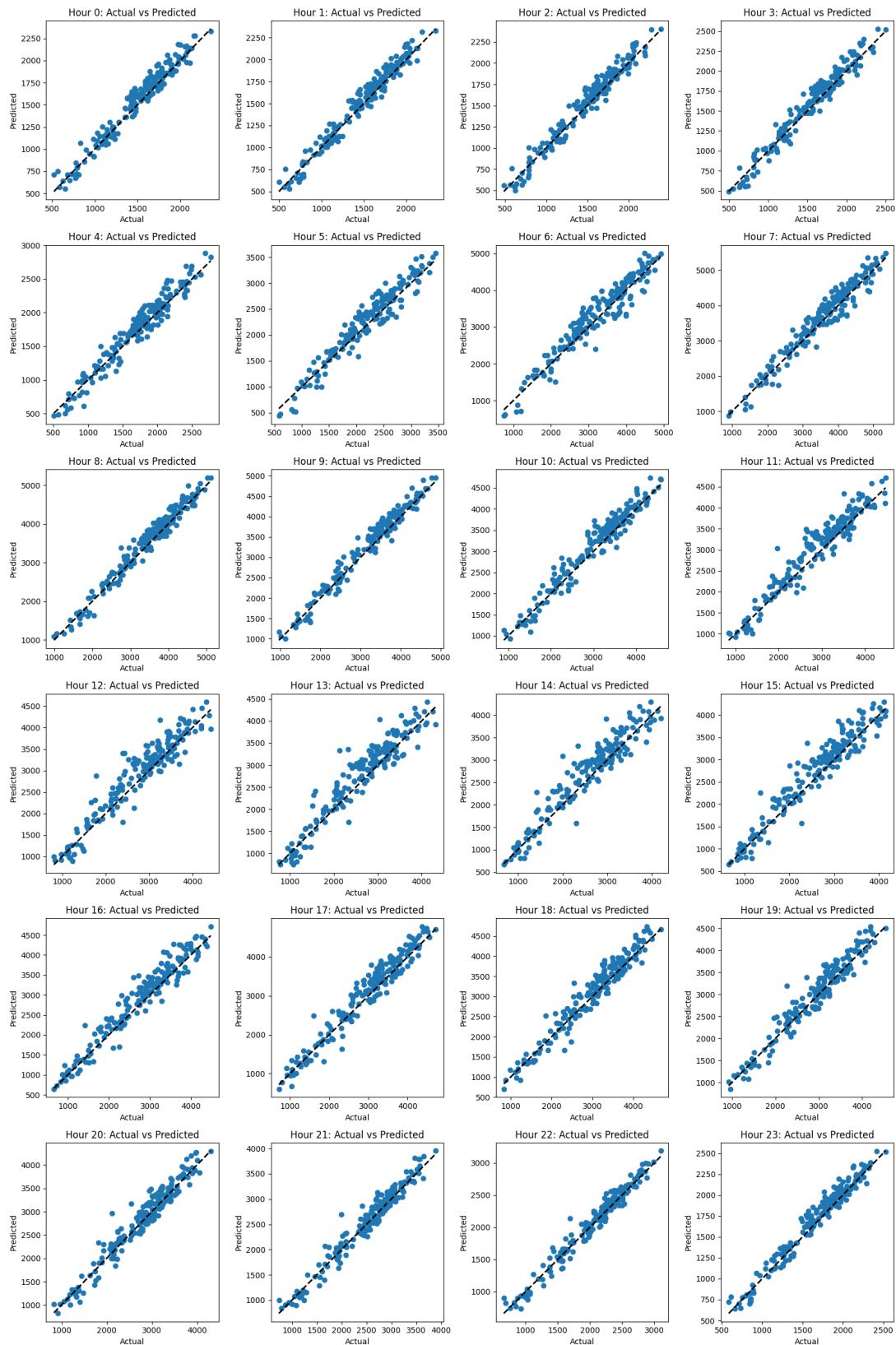


FIGURE 22 – Résultats de la régression linéaire finale