École nationale
de la statistique
et de l'analyse
de l'information

Mastère spécialisé Data Science pour la connaissance client

Module évalué : Séries temporelles

ETUDE DE LA CONSOMMATION DE GAZ DANS DIFFERENTES METROPOLES FRANCAISES

Modèles de prédictions ARMA, SARIMA et SARIMAX

Rapport présenté par :
ADUAYOM MESSAN Messan Daniel
SADIO Ndeye Salimata

Professeur :
Mr. ESSTAFA Youssef

Table des matières

1 Introduction

La transition énergétique et la quête de durabilité environnementale sont au cœur des préoccupations mondiales. Dans le paysage énergétique actuel, la gestion de la consommation de gaz en France se révèle être un enjeu de première importance. Alors que les métropoles françaises, ayant un impératif de transition vers des sources d'énergie plus durables cherchent à réduire leur empreinte carbone, décrypter les schémas de consommation de gaz en vigueur devient essentiel. Dans cette perpective, cette étude approfondie vise à contextualiser les enjeux environnementaux liés à la consommation de gaz dans différentes métropoles françaises, s'inscrivant ainsi dans une démarche globale de transition vers des modes de vie plus durables. Pour ce faire, l'étude s'attelle à explorer et modéliser la consommation de gaz en France en se basant sur des approches spécialisées de séries temporelles.

Les donnée sous-jacentes sont issues de la plate-forme opendata.reseaux-energies.fr, offrant une vision détaillée de l'évolution de la consommation de gaz à l'échelle métropolitaine au fil du temps. L'objectif prédominant de cette analyse consiste à mettre en œuvre des modèles afin de saisir les tendances, saisons et éventuelles influences externes présentes dans les données.

Au travers de cette démarche, nous répondrons à des interrogations cruciales, telles que la détection de tendances saisonnières et la projection de la consommation future. Les résultats issus de cette analyse se présentent comme une source d'information significative pour les décideurs, les planificateurs énergétiques et les chercheurs engagés dans le domaine de l'énergie.
La section suivante détaillera le processus de collecte et de préparation des données, ainsi que les étapes successives d'exploration et de modélisation. En adoptant une approche méthodique, ce rapport ambitionne de fournir une compréhension approfondie de la consommation de gaz en France, ouvrant ainsi la voie à des prises de décision éclairées dans le domaine énergétique.

En outre, cette approche intégrée, combinant la richesse des données historiques à la puissance des modèles de séries temporelles, vise à offrir une vision complète de la consommation de gaz dans les métropoles françaises. À terme, cette recherche aspire à orienter les décideurs vers des choix éclairés et durables, contribuant significativement à la réalisation d'une transition énergétique respectueuse de l'environnement, en harmonie avec les impératifs écologiques de notre époque.

2 Contexte et Objectifs

En France, la pression croissante pour réduire les émissions de gaz à effet de serre a catalysé une transformation profonde du paysage énergétique. En 2021, le gaz naturel représentait 16% du mix énergétique français, soulignant l'importance de comprendre comment cette ressource est utilisée. Alors que les énergies renouvelables ont connu une croissance notable, la dépendance au gaz persiste, notamment dans le secteur résidentiel, où 44% du gaz est dédié au chauffage. La consommation de gaz pose des défis majeurs en matière de durabilité environnementale, principalement en raison de son caractère fossile. Le secteur résidentiel, principal consommateur de gaz, est confronté à la nécessité d'adopter des alternatives plus respectueuses de l'environnement. Les politiques énergétiques, telles que la Programmation Pluriannuelle de l'énergie, visent à encourager la transition vers des sources d'énergie plus propres et à réduire la dépendance au gaz naturel.

Face à ces enjeux, la France a mis en œuvre des initiatives novatrices pour une consommation de gaz plus durable. Les réseaux intelligents, les incitations à l'adoption de technologies écoénergétiques et l'implication des citoyens dans la transition énergétique sont autant de piliers de ces nouvelles dispositions. Le Plan de Relance de 2022, avec ses 30 milliards d'euros dédiés à la transition écologique, illustre l'engagement financier du gouvernement en faveur de ces innovations.
L'étude se positionne donc au croisement de l'urgence environnementale et de la nécessité de repenser la consommation de gaz dans les métropoles françaises. L'analyse des modèles de consommation, à travers l'application de modèles de séries temporelles, offre une opportunité unique de prévoir les évolutions futures et de recommander des stratégies intégrant les impératifs environnementaux. En fusionnant données historiques et nouvelles dispositions énergétiques, cette recherche contribuera à guider les décideurs vers des choix plus durables, favorisant ainsi une transition énergétique respectueuse

de l'environnement.
Pour concrétiser ces ambitions, le rapport détaillera la base de données qui constitue le socle de notre étude, rassemblant avec soin des données historiques sur la consommation de gaz dans diverses métropoles françaises. Ensuite, notre approche méthodologique se centrera sur une analyse globale de cette base de données, visant à extraire des enseignements significatifs en identifiant des motifs récurrents et des tendances émergentes. Ceci servira de base solide pour élaborer une méthodologie adaptée à l'étude des séries temporelles dans le contexte spécifique de la consommation de gaz.

La phase suivante impliquera l'intégration de modèles de séries temporelles, mettant particulièrement l'accent sur ceux liés à la saisonnalité. Les variations saisonnières dans la consommation de gaz fournissent souvent des indices sur les changements de comportement liés aux conditions météorologiques, aux périodes de vacances et à d'autres facteurs environnementaux. L'inclusion de ces modèles permettra une analyse plus précise des schémas de consommation, facilitant ainsi la prédiction des tendances futures et la formulation de recommandations stratégiques. Dans l'ensemble, cette approche intégrée, combinant la richesse des données historiques à la puissance des modèles de séries temporelles, vise à offrir une vision complète de la consommation de gaz dans les métropoles françaises. À terme, cette recherche aspire à orienter les décideurs vers des choix éclairés et durables, contribuant significativement à la réalisation d'une transition énergétique respectueuse de l'environnement, en harmonie avec les impératifs écologiques de notre époque.

3 Présentation de la base de données

3.1 Description du jeu de données

La base de données utilisée dans notre étude compile de manière exhaustive les consommations des réseaux de gaz, notamment GRT (Gestionnaire de Réseau de Transport), GRD (Gestionnaire de Réseau de Distribution), et EDL (Établissements Locaux de Distribution). L'objectif principal de cette base est de fournir une estimation fiable des niveaux de consommation à des échelles géographiques fines, avec un accent particulier sur les Métropoles françaises. La granularité géographique fine, limitée à l'échelle de la Métropole, offre une vision précise de l'évolution de la consommation de gaz au cours des derniers mois.

Élaborée par les opérateurs de transport tels que GRTgaz et Teréga, cette base de données répond à la nécessité cruciale de rendre compte des évolutions de la consommation de gaz à des échelles locales. Actualisés mensuellement, ces jeux de données permettent aux territoires de surveiller de près l'évolution de leur consommation, fournissant ainsi une vision dynamique des tendances énergétiques.

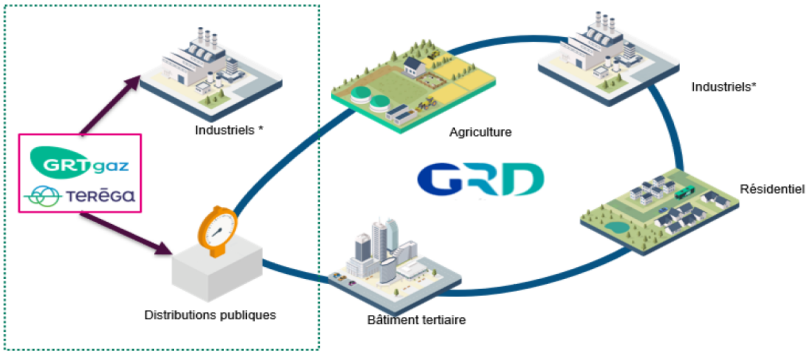


FIGURE 1 – Consommation brute des métropoles

Cette ressource essentielle a été récupérée sur le site de datagouv, assurant ainsi une transparence et une accessibilité accrues. Les données principales incluses dans cette base couvrent la consommation brute des métropoles ainsi que des détails spécifiques sur la consommation de gaz.

Informations clés sur la Base de Données : Présentation des variables

- **Donnée de Consommation de la Zone :** Consommation brute des métropoles en KWh PCS 0°C (Puissance Calorifique Supérieure , Kilowatt-heure à la température de référence 0°C)
- **Granularité Géographique :** Métropoles.
- **Périmètre Géographique :** France métropolitaine, hors Corse.
- **Pas Temporel :** Mensuel.
- **Profondeur Historique :** Depuis le 1er janvier 2017.
- **Producteur de Données :** GRTgaz, Teréga.
- **Date de Dernière Mise à Jour :** Entre le 15 et le 20 du mois suivant.
- **Fréquence de Mise à Jour :** Mensuel.

Cette base de données représente ainsi une source robuste et actualisée, offrant une assise solide pour l'analyse des modèles de consommation de gaz dans les métropoles françaises.

TABLE 1 – Aperçu de la base de données

Date	Code Officiel EPCI	Nom Métropole	Type EPCI	Consommation	Centroid
2021-01	200054807	Métropole d'Aix-Marseille-Provence	ME	4145557583	43.493308249,5.344537348
2021-01	244400404	Nantes Métropole	ME	710077638	47.224910325,-1.593721407
2020-09	200093201	Métropole Européenne de Lille	ME	274871006	50.643899239,3.033354341
2023-07	200093201	Métropole Européenne de Lille	ME	172036797	50.643899239,3.033354341
2017-01	243400017	Montpellier Méditerranée Métropole	ME	285548931	43.615713457,3.863300356
2019-08	243400017	Montpellier Méditerranée Métropole	ME	28810584	43.615713457,3.863300356

3.2 Préparation de la donnée

Dans le cadre de notre étude, des ajustements ont été apportés aux données pour les rendre plus adaptées à nos objectifs de recherche. Les principales étapes de ces travaux incluent la création de nouvelles variables importantes et des modifications de formatage pour assurer la qualité des données.

- **Latitude et Longitude :** Les variables `latitude` et `longitude` ont été générées en séparant la variable existante `centroid`. Cette démarche nous permettra d'utiliser plus efficacement ces informations géographiques dans nos analyses, améliorant ainsi notre compréhension des modèles de consommation de gaz à l'échelle spatiale.

— **Saison :** La `saison` : Hiver, Printemps, Ete et Automne. Cette variable a été générée à partir du mois. Cette démarche nous permettra de voir les baisses et augmentations de consommation suivant les saisons au cours de l'année.

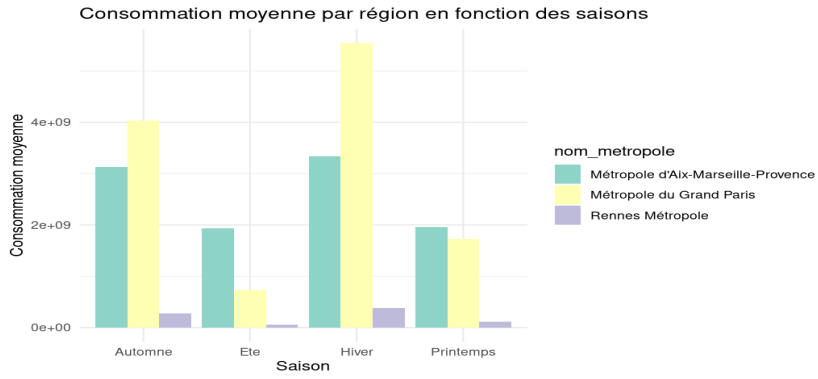


FIGURE 2 – Consommation brute des métropoles en focntion des saisons

— **Formatage de la Date :** La variable temporelle `date` a été formatée de manière à garantir une représentation temporelle homogène et facilitant son utilisation dans nos modèles de séries temporelles.

— **Exclusion de Colonnes non utiles :** Pour simplifier notre analyse, les colonnes `code officiel` et `type epci` ont été retirées du jeu de données. Cette démarche vise à concentrer notre attention sur les variables essentielles pour la modélisation de la consommation de gaz dans les métropoles.

Ces ajustements de données sont cruciaux pour préparer nos informations à une analyse approfondie des modèles de consommation de gaz. Ils renforcent la qualité de notre ensemble de données, nous permettant de tirer des conclusions significatives pour guider nos futures recommandations.

3.3 Exploration des données

Suite à ces prétraitements, nous entamons à présent une analyse de notre base de données. Cette section débutera par une vue d'ensemble des différentes variables. Ultérieurement, une attention particulière sera accordée à la consommation dans les diverses zones.

Table 2 – Statistiques descriptives pour les variables du tableau.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
`date`	2017-01-01	2018-10-01	2020-07-01	2020-06-15	2022-03-01	2023-11-01
`consommation`	1.34e+07	8.624e+07	2.096e+08	4.917e+08	4.263e+08	8.327e+09
`latitude`	43.12	44.88	47.31	46.74	48.57	50.64
`longitude`	-4.471	1.028	3.09	3.026	5.718	7.718

L'examen des statistiques descriptives des variables clés de notre ensemble de données, à savoir la *date* et la *consommation*, offre des perspectives importantes pour mieux comprendre les tendances dans notre étude.

Concernant la variable *date*, nos données couvrent la période allant du 1er janvier 2017 au 1er novembre 2023. Cette fenêtre temporelle étendue offre une base solide pour examiner l'évolution de la consommation de gaz dans les métropoles au fil du temps.

En ce qui concerne la variable *consommation*, les statistiques indiquent une plage dynamique, variant de 13,400,000 à 8,327,000,000. La consommation minimale représente le niveau le plus bas enregistré, tandis que la consommation maximale représente le pic observé. Les quartiles fournissent des points de repère essentiels pour comprendre la répartition des données.

La médiane, fixée au 1er juillet 2020, peut servir de point central pour diviser notre ensemble de données en deux parties égales. La moyenne, établie au 15 juin 2020, donne une indication de la tendance centrale, bien que la différence entre la médiane et la moyenne puisse signaler l'influence potentielle de valeurs extrêmes.

Ces observations offrent un aperçu préliminaire, mais des analyses plus approfondies, telles que l'application de modèles de séries temporelles, seront nécessaires pour discerner les tendances à long terme et formuler des recom-

mandations stratégiques spécifiques à notre contexte.

3.4 Analyse approfondie de la Consommation Énergétique

Pour pousser l'analyse plus loin, nous envisageons une évaluation visant à identifier les zones présentant une forte consommation énergétique. Notre base de données, adoptant les caractéristiques d'une donnée de panel avec les variables de date et de métropole, sera examinée plus en détail. Dans le cadre de notre étude, nous concentrerons notre attention sur trois métropoles spécifiques, les considérant comme des séries distinctes. Cette approche nous permettra d'explorer d'éventuelles similitudes et différences entre ces métropoles.

Cette méthodologie s'inspire de l'étude de J. Sovacool et D. Brown, "Urban Energy Consumption : Different Cities, Different Patterns"[4]. Cette recherche explore les disparités dans les schémas de consommation énergétique observées dans diverses métropoles, soulignant ainsi l'importance d'une analyse détaillée pour appréhender les tendances propres à chaque ville. En adoptant une perspective similaire, notre étude aspire à révéler les spécificités des modèles de consommation au sein des métropoles ciblées, enrichissant ainsi notre compréhension des dynamiques énergétiques urbaines.

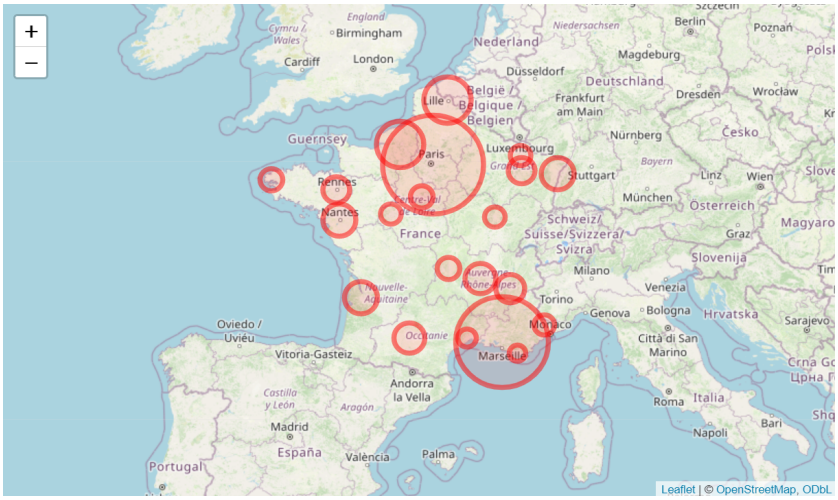


FIGURE 3 – consommation brute des métropoles en 2023

Suite à l'analyse, il est notable que les grandes métropoles, en particulier les métropoles parisiennes et marseillaises, se distinguent le plus en termes de

consommation énergétique. Cependant, dans un souci d'accroître la diversité des profils, nous avons décidé d'inclure la métropole de Rennes dans notre étude. Celle-ci se caractérise par une consommation moyenne plus faible au cours de l'année 2023.

Comme mentionné précédemment, cette approche vise à appliquer une méthodologie détaillée, permettant ainsi de discerner les tendances spécifiques à chaque ville. En adoptant une perspective fine, notre étude s'efforce de dévoiler les nuances dans les modèles de consommation énergétique de ces métropoles sélectionnées, contribuant ainsi à une compréhension plus approfondie des dynamiques énergétiques urbaines.

4 Etudes des séries temporelles : Paris - Marseille - Rennes

4.1 Analyse de l'évolution temporelles

Maintenant que nous avons identifié le cadre dans lequel notre étude va se faire, rappelons que nous amorçons une procédure d'analyse approfondie en adoptant une approche individualisée pour chaque métropole. Nous avons sélectionné les séries temporelles de consommation de gaz des métropoles de Paris, Marseille, et Rennes comme base de notre investigation. Dans cette phase de notre recherche, l'accent sera mis sur l'exploration de l'évolution temporelle de la consommation au sein de ces séries métropolitaines. Cette démarche nous permettra de tirer des conclusions préliminaires, lesquelles constitueront le fondement de nos modèles et de nos analyses ultérieurs.

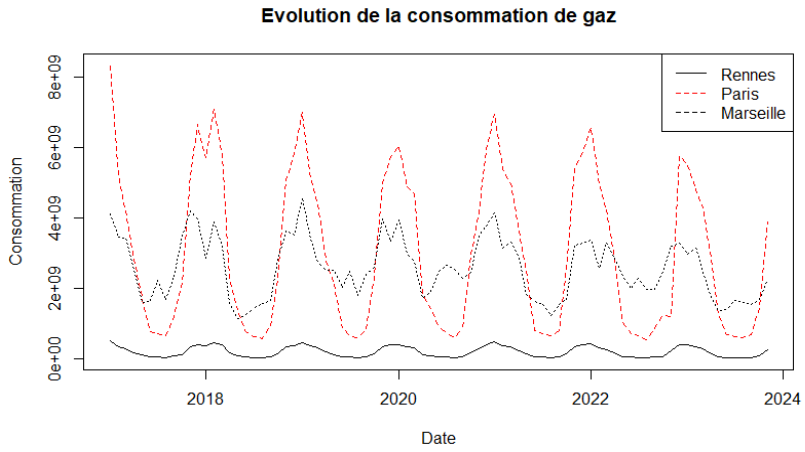


FIGURE 4 – consommation gaz de 2017 à 2023

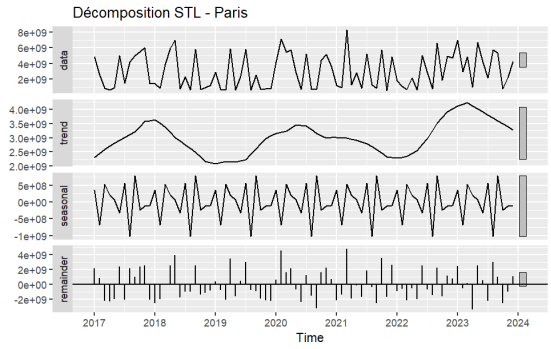
Dans la figure 4, l'observation des données révèle une disparité significative dans la consommation de gaz entre les métropoles étudiées. La ville de Paris se distingue par une consommation nettement plus élevée, suivie par la métropole de Marseille, tandis que Rennes affiche une consommation sensiblement inférieure. Ces différences peuvent être attribuées en partie à la taille et à la densité de population de chaque métropole. Paris, en tant que métropole majeure et centre d'activités professionnelles intensives, présente naturellement une demande énergétique plus élevée.

Les chiffres clés soutiennent cette observation, soulignant la densité de population, l'activité économique et le nombre de foyers comme des facteurs déterminants. Par exemple, la région parisienne est reconnue pour abriter un grand nombre d'entreprises, une forte intensité des transports et une population considérable, tous contribuant à une consommation de gaz plus élevée.

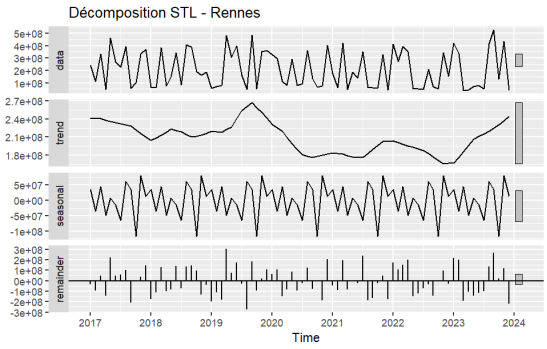
Au-delà de cette distinction, une analyse temporelle révèle des variations saisonnières marquées. Toutes les métropoles enregistrent des baisses significatives de consommation pendant les périodes estivales, suivies d'augmentations notables pendant les périodes hivernales. Ces fluctuations suggèrent fortement un effet saisonnier influant sur la consommation de gaz. Il serait intéressant d'approfondir cette observation en considérant des facteurs tels que les besoins de chauffage saisonniers, les comportements de consommation, et les politiques énergétiques en vigueur, y compris les initiatives de réduction de la consommation énergétique adoptées par la France, telles que la coupure temporaire de l'alimentation des radiateurs pendant certaines périodes.

4.2 Décomposition temporelles

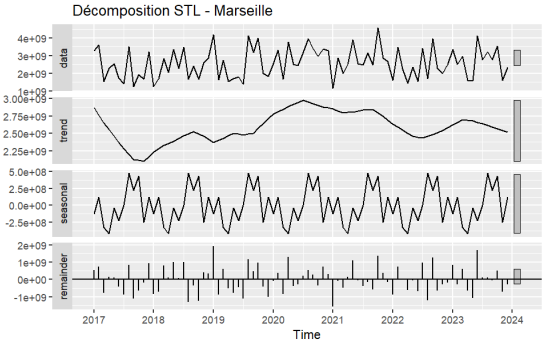
Nous avons précédemment constaté que nos séries sont susceptibles d'être influencées par divers facteurs tels que la densité de population et une forme de saisonnalité dans les séries temporelles. À ce jour, les données dont nous disposons ne nous permettent pas de prendre en compte directement l'effet de la population et des activités. Cependant, nous pouvons éventuellement obtenir des informations sur la taille de la métropole en nous basant sur son nom. Pour l'instant, ce qui nous est le plus accessible est l'identification d'une saisonnalité ou son absence. Pour ce faire, nous allons procéder à une décomposition temporelle afin de récupérer la tendance et l'effet saisonnier dans nos différentes séries.



(a) Décomposition série Paris



(b) Décomposition série Rennes



(c) Décomposition série Marseille

FIGURE 5 – Décomposition des séries des différentes métropoles

Suite à la décomposition de nos séries, les résultats sont présentés dans la Figure 5. Dans cette première analyse, l'observation initiale indique une stabilité de la variabilité autour de la moyenne pour toutes les métropoles étudiées. Cette observation préliminaire suggère la possibilité d'une certaine stationna-

rité dans nos séries, une caractéristique que nous chercherons à confirmer avec des tests appropriés dans la prochaine section de notre étude.

En examinant de plus près la composante saisonnière, nous constatons que toutes les séries présentent une saisonnalité prononcée qui se répète régulièrement au fil du temps. De plus, il semble que cette saisonnalité soit relativement similaire entre les différentes métropoles. Cette uniformité pourrait être attribuée à l'étude d'un phénomène unique, où la seule distinction réside dans la nature spécifique de chaque métropole.

Prenons par exemple la métropole de Paris, où la saisonnalité semble suivre des tendances similaires à celles observées dans les métropoles de Marseille et Rennes. Cette cohérence dans les motifs saisonniers peut être liée à des facteurs climatiques, à des comportements de consommation généraux ou à d'autres influences externes.

4.3 Tests de stationnarité

La visualisation graphique de nos séries, comme précédemment mentionné, semble indiquer une tendance vers la stationnarité. Toutefois, pour valider cette hypothèse, nous entreprendrons une approche plus formelle dans cette section en utilisant des tests statistiques appropriés. En effet, l'évaluation de la stationnarité est une étape cruciale dans notre analyse, impliquant la vérification de la constance des propriétés statistiques d'une série temporelle dans le temps.

La stationnarité d'une série temporelle est un concept crucial dans l'analyse des séries temporelles, soulignant la stabilité des propriétés statistiques de la série au fil du temps. Cette notion revêt une importance significative, car elle garantit la constance des caractéristiques de la série, renforçant ainsi la fiabilité des prévisions et des modèles construits sur ces données. Afin d'évaluer la stationnarité de nos séries, nous allons utiliser le test de Dickey-Fuller, une méthode fréquemment employée pour cette vérification formelle.

Les hypothèses de stationnarité pour une série temporelle X_t sont définies comme suit :

1. **Hypothèse de constance de la moyenne :** $\mu_{X_t} = \text{constante}$
2. **Hypothèse de constance de la variance :** $\sigma^2_{X_t} = \text{constante}$
3. **Hypothèse de l'indépendance temporelle :** $Cov(X_t, X_{t+k}) = 0$ pour tout $k \neq 0$

Table 3 – Résultats des tests de Dickey-Fuller

Métropole	**Statistique Dickey-Fuller**	**Lag order**	**p-value**
Paris	-4.8226	4	0.01
Marseille	-4.2155	4	0.01
Rennes	-3.8489	4	0.02067

Table 4 – Tests de stationnarité KPSS

Métropole	**KPSS Level**	**Truncation lag**	**p-value**
Paris	0.60431	3	0.02224
Marseille	0.22192	3	0.1
Rennes	0.13985	3	0.1

Les résultats des tests de Dickey-Fuller et de stationnarité KPSS pour nos

séries indiquent ce qui suit :

— Pour Paris, la statistique Dickey-Fuller est -4.8226 avec un ordre de retard de 4 et une p-value de 0.01, suggérant la stationnarité.

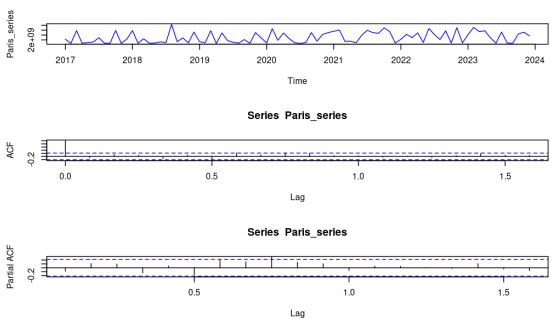
— Marseille présente une statistique Dickey-Fuller de -4.2155, un ordre de retard de 4 et une p-value de 0.01, confirmant également la stationnarité.

— Rennes montre une statistique Dickey-Fuller de -3.8489, un ordre de retard de 4 et une p-value de 0.02067, indiquant une possible stationnarité.

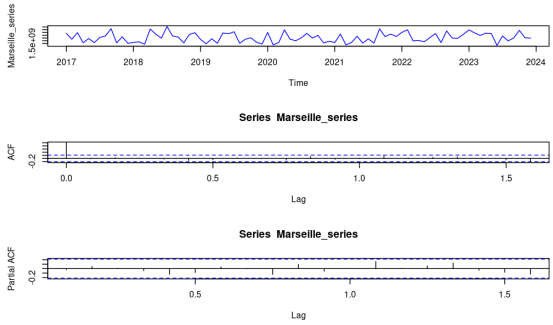
— Les tests KPSS confirment la stationnarité pour Paris, suggèrent une légère non-stationnarité pour Marseille, et indiquent une possible non-stationnarité pour Rennes.

4.4 Autocorrélation et autocorrélation partielle

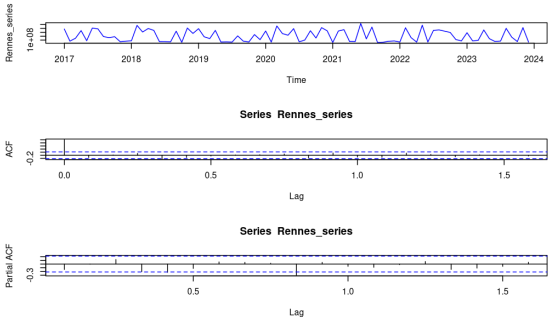
Dans cette section de l'étude, nous aborderons l'analyse des autocorrélations et autocorrélations partielles des séries. L'objectif est de comprendre les relations entre les observations successives et d'identifier d'éventuels schémas de dépendance temporelle. L'analyse des autocorrélations et autocorrélations partielles offre des perspectives importantes sur la structure temporelle des données, permettant d'identifier des tendances, des cycles ou des phénomènes saisonniers. Cela s'avère crucial pour une modélisation précise des séries temporelles.



(a) ACF et PACF serie Paris



(b) ACF et PACF serie Marseille



(c) ACF et PACF serie Rennes

FIGURE 6 – ACF et PACF series

En observant l'Analyse des Corrélations (ACF), on constate que les décalages après le décalage 0 ne montrent pas de significativité. Cela suggère qu'il n'y a pas de corrélation systématique entre les observations successives une fois que les influences immédiates ont été prises en compte. Dans le contexte de la modélisation de la consommation de gaz, cela pourrait indiquer une absence de tendance ou de structure temporelle évidente après avoir considéré

les variations immédiates.

Quant à l'Analyse des Corrélations Partielles (PACF) présentant un aspect sinusoïdal, cela suggère une dépendance temporelle périodique dans les données. Cette périodicité peut être associée à des schémas saisonniers ou cycliques dans la consommation de gaz. Il serait donc pertinent d'explorer davantage ces schémas pour mieux comprendre les facteurs saisonniers influençant la consommation de gaz dans les différentes métropoles.

En résumé, l'ACF ne montre pas de corrélation significative au-delà du décalage 0, indiquant une absence de corrélation systématique à long terme. D'autre part, la PACF suggère une dépendance temporelle périodique, mettant en lumière des patterns saisonniers potentiels dans les séries temporelles.

5 Modélisation

Force est de noter que dans un soucis d'aménagement des calculs et du fait que la consommation est mesurée en millions, nous avons travaillé sur des données normalisées avec la formule zscore et que nous procédons par la suite à une dénormalisation, ce qui pourrait provoquer des frottements et des variabilités dans les prédictions.

5.1 Sélection de modèles

Notre série est stationnaire, ses propriétés sont constantes. Elle présente également une saisonnalité. La stationnarité fait penser au modèle ARMA et la saisonnalité fait penser aux modèles SARIMA(Seasonal AutoRegressive Integrated Moving Average) et SARIMAX(Seasonal AutoRegressive Integrated Moving Average with eXogenous factors).

Soit $(\varepsilon_t)_{t\in\mathbb{Z}}$ un bruit blanc faible de variance σ^2.

On dit qu'un processus $(X_t)_{t\in\mathbb{Z}}$ est un processus ARMA (AutoRegressive Moving Average) d'ordre (p, q), noté ARMA(p, q), si :

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

où $(\phi_1, \ldots, \phi_p) \in \mathbb{R}^p$, $\phi_p \neq 0$, $(\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$, et $\theta_q \neq 0$.

Pour aller plus loin, vu que notre série présente des motifs saisonniers (des variations saisonnières de consommation de gaz), un modèle SARIMA peut être également approprié. On dit qu'un processus $(X_t)_{t\in\mathbb{N}}$ est un processus SARIMA (Seasonal AutoRegressive Integrated Moving Average) d'ordre $(p, d, q)(P, D, Q)_s$, noté SARIMA$(p, d, q)(P, D, Q)_s$, si :

$$\Phi(B)\Phi'(B_s)\nabla^d\nabla_s^D X_t = \Theta(B)\Theta'(B_s)\varepsilon_t$$

où :

$$\nabla^d = (I - B)^d$$

$$\nabla_s^D = (I - B_s)^D$$

$$\Phi(B) = I - \phi_1 B - \ldots - \phi_p B^p$$

où $(\phi_1, \ldots, \phi_p) \in \mathbb{R}^p$ et $\phi_p \neq 0$,

$$\Phi'(B) = I - \phi'_1 B - \ldots - \phi'_P B^P$$

où $(\phi'_1, \ldots, \phi'_P) \in \mathbb{R}^P$ et $\phi'_P \neq 0$,

$$\Theta(B) = I + \theta_1 B + \ldots + \theta_q B^q$$

où $(\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$ et $\theta_q \neq 0$,

$$\Theta'(B) = I + \theta'_1 B + \ldots + \theta'_Q B^Q$$

où $(\theta'_1, \ldots, \theta'_Q) \in \mathbb{R}^Q$ et $\theta'_Q \neq 0$. Ensuite, nous pouvons approfondir notre

étude en exploitant la modélisation SARIMAX) qui est une extension du modèle SARIMA qui permet d'intégrer des variables exogènes dans la modélisation des séries temporelles.

5.1.1 Evaluation du modèle

Dans notre cas de figure, vu que nous prédisons des consommations, afin d'évaluer la précision du modèle, nous allons utiliser le RMSE(Root Mean Squared Error). Elle représente la racine carrée de la moyenne des carrés des erreurs entre les valeurs prédites par le modèle et les valeurs réelles observées. Sa formule est définie comme suit où :

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

où n est le nombre total d'observations, y_i est la valeur réelle observée, et $\hat{y}_i$ est la valeur prédite par le modèle.

5.1.2 ARMA

Dans le cas d'une série stationnaire,un modèle ARMA peut être efficace pour capturer les structures temporelles sans nécessiter de transformation supplémentaire comme la différenciation. Cependant, la série présente des saisonnalités donc ARMA n'est pas adéquat. Cependant nous allons quand même l'appliquer pour en avoir la confirmation. Le choix de l'ordre des composantes AR et MA (c'est-à-dire le choix des termes p et q respectivement) peut être déterminé en utilisant des techniques telles que la fonction d'autocorrélation partielle (PACF) pour AR et la fonction d'autocorrélation (ACF) pour MA.

Dans notre approche, nous avons choisi de procéder à des tests pour évaluer différentes combinaisons d'ordre, en comparant les critères d'information d'Akaike (AIC). L'objectif était de sélectionner la combinaison d'ordre présentant le meilleur AIC. Cette configuration d'ordre optimale sera ensuite maintenue et appliquée dans les étapes ultérieures pour les modèles plus complexes que

nous développerons.

Table 5 – Marseille

AR	MA	AIC
1	1	244.5035
2	1	246.3526
3	1	247.3941
1	2	246.2763
2	2	243.4365
3	2	244.4080
1	3	247.4633
2	3	244.7290
3	3	246.2660

Table 6 – Paris

AR	MA	AIC
1	1	244.4613
2	1	246.4324
3	1	248.3010
1	2	246.4517
2	2	248.3941
3	2	245.6580
1	3	248.2822
2	3	245.2843
3	3	247.2487

Table 7 – Rennes

AR	MA	AIC
1	1	243.6453
2	1	245.0181
3	1	246.2978
1	2	244.1071
2	2	242.1170
3	2	242.1381
1	3	245.4563
2	3	241.6344
3	3	245.1430

Nous avons identifié les configurations d'ordres optimales pour nos séries, à savoir 2,0,2 pour Marseille, 1,0,1 pour Paris, et 2,0,3 pour Rennes. Nous débuterons notre modélisation ARMA en utilisant ces configurations spécifiques. Cette approche nous permettra de minimiser la répétition et le calcul des critères d'information d'Akaike (AIC) pour différentes combinaisons d'ordres, simplifiant ainsi le processus tout en utilisant des configurations préalablement sélectionnées. De plus, cette procédure nous facilitera l'exploration d'autres modèles de séries temporelles dans la suite de notre analyse.

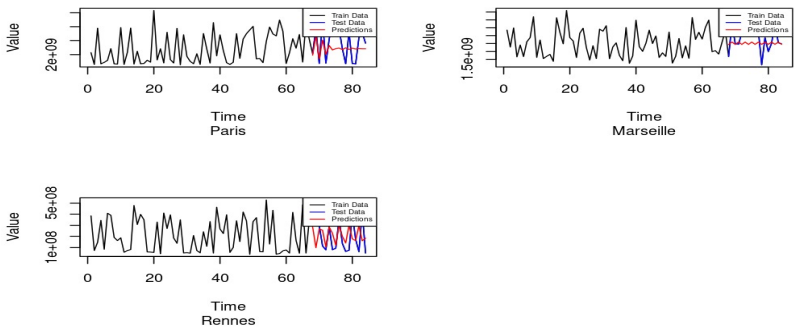


Figure 7 – Modélisation ARMA

— RMSE for Paris : 2282729980
— RMSE for Marseille : 765995756
— RMSE for Rennes : 171753025

Comme anticipé, les résultats des prédictions ne se révèlent pas particulièrement pertinents. Le modèle ARMA appliqué à la série temporelle de Paris et de

Marseille produit des prédictions à long terme qui demeurent stables sans présenter de fluctuations significatives. En revanche, la série de Rennes se distingue avec le plus faible Root Mean Square Error (RMSE) et des prédictions relativement meilleures.

Il est intéressant de noter que, dans le contexte des métropoles à forte consommation de gaz, le modèle ARMA montre ses limites en ne parvenant pas à capturer les valeurs élevées, ce qui se traduit par une augmentation significative de l'erreur de prédiction. Cette observation souligne la nécessité d'explorer des approches alternatives pour améliorer la performance du modèle. Par la suite, nous comparerons ces résultats à ceux du modèle intégrant la saisonnalité, à savoir le modèle SARIMA.

5.1.3 SARIMA

Suite à nos analyses antérieures, nous avons identifié une saisonnalité marquée dans nos données, manifeste notamment lors des périodes de forte consommation de gaz et pendant les mois d'été, caractérisés par une faible consommation de gaz dans les différentes métropoles. Dans ce contexte, les modèles préconisés sont les modèles SARIMA.

La modélisation SARIMA présente plusieurs avantages :

— **Capture des tendances saisonnières :** SARIMA intègre des paramètres saisonniers pour saisir les variations régulières dans la série temporelle à des intervalles fixes, permettant au modèle de s'ajuster aux motifs spécifiques liés aux saisons.

— **Prédiction précise :** L'inclusion de composantes saisonnières améliore la précision des prédictions pour les points temporels futurs en prenant en compte les schémas saisonniers historiques.

— **Gestion des effets saisonniers complexes :** SARIMA peut être configuré pour traiter des effets saisonniers complexes, adaptés à des variations saisonnières changeantes au fil du temps.

— **Différenciation saisonnière :** L'intégration saisonnière dans SARIMA

permet de différencier la série temporelle à un niveau saisonnier, stabilisant ainsi la variance saisonnière et rendant la série plus stationnaire.

— **Inclusion de termes d'ordre élevé :** SARIMA peut incorporer des termes d'ordre élevé pour les composantes saisonnières (p, d, q, P, D, Q), permettant de modéliser des variations saisonnières avec une plus grande complexité.

Nous allons donc mettre en place le modèle SARIMA en conservant les ordres qui ont minimisé l'AIC pour le modèle ARMA précédemment établi pour chaque série. Pour déterminer les ordres de saisonnalité, nous utiliserons le package `autoarima`, spécifiant notre intérêt pour la saisonnalité. Ce package nous fournira les combinaisons de saisonnalité les plus optimales pour notre modélisation, que nous utiliserons ensuite pour instaurer le modèle SARIMA sur les différentes séries.

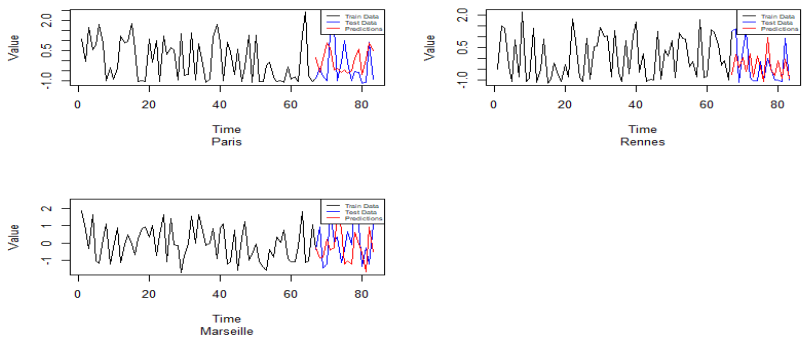


FIGURE 8 – Modélisation SARIMA

— RMSE for Paris : 2554245309
— RMSE for Marseille : 182224581
— RMSE for Rennes : 892764832

Les séries temporelles de Paris et de Marseille, analysées avec le modèle SARIMA, présentent des fluctuations sur l'ensemble de la période. Cependant, en ce qui concerne les prédictions, nous ne parvenons toujours pas à obtenir des résultats convaincants. Contrairement au modèle ARMA, cette fois-ci, c'est la série Marseille qui affiche les erreurs de prédiction les plus faibles. Ces résultats sont toutefois surprenants, car le modèle SARIMA est supposé

intégrer la composante saisonnière dans ses prédictions et mieux modéliser les séries par rapport à un modèle ARMA.

Cette constatation souligne la nécessité d'une évaluation plus approfondie des facteurs qui pourraient influencer les performances du modèle SARIMA dans le contexte spécifique des séries temporelles de consommation de gaz de Paris et de Marseille. Nous poursuivrons notre analyse en examinant attentivement ces résultats et en identifiant d'éventuelles améliorations à apporter pour renforcer la capacité prédictive de notre modèle.

5.1.4 SARIMAX

Suite à l'observation des résultats peu concluants du modèle SARIMA, nous avons pris la décision d'introduire une complexité supplémentaire en optant pour une modélisation qui intègre des variables exogènes.

En effet, le modèle SARIMAX offre la possibilité d'incorporer des variables exogènes, qui sont des facteurs extérieurs à la série temporelle que l'on souhaite modéliser, mais qui peuvent influencer son comportement. Les avantages de SARIMAX incluent :

- **Intégration des variables exogènes :** SARIMAX permet d'inclure des variables exogènes dans le modèle, améliorant ainsi sa capacité à capturer des variations non expliquées par les seules données temporelles.
- **Amélioration de la prédiction :** En incorporant des informations supplémentaires à partir de variables exogènes, SARIMAX peut fournir des prédictions plus précises en prenant en compte des facteurs externes qui influent sur la série temporelle.
- **Gestion de la causalité :** SARIMAX s'avère utile lorsque des relations de causalité existent entre la série temporelle d'intérêt et d'autres variables exogènes, permettant ainsi une modélisation plus précise de l'impact de ces variables sur la série.
- **Adaptation à des modèles plus complexes :** SARIMAX offre une flexibilité pour modéliser des séries temporelles complexes en incluant

des composantes saisonnières ainsi que des variables exogènes, adaptant ainsi le modèle à des situations réelles plus variées.

— **Modélisation de réponses à des interventions externes :** SARIMAX peut être utilisé pour modéliser les réponses de la série temporelle à des événements externes en intégrant des variables exogènes liées à ces événements.

Afin de mettre en œuvre cette approche, nous introduirons une variable discrète "saison" basée sur le mois. Cette variable nous permettra d'identifier la saison dans laquelle nous nous trouvons, avec une catégorisation telle que "Hiver", "Printemps", "Été" et "Automne". Cette démarche s'inscrit dans la logique de l'observation précédente de la surconsommation de gaz en hiver, permettant au modèle d'identifier efficacement cette saison.

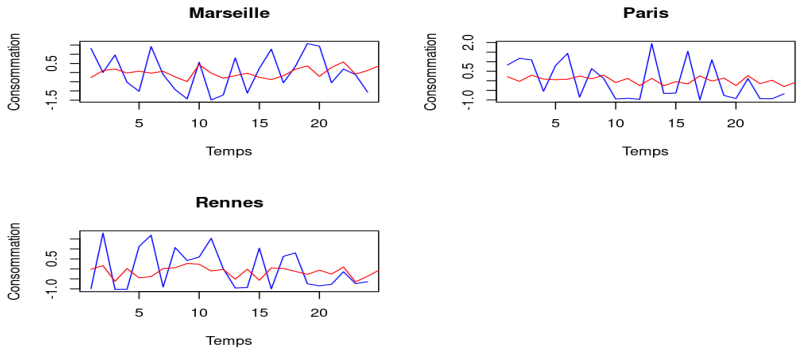


FIGURE 9 – Modélisation SARIMAX

Dans ce modèle, nous n'adopterons pas le critère de RMSE. En effet, la visualisation graphique seule peut déjà nous indiquer que nos valeurs prédites ne sont pas suffisamment fiables. À ce jour, les modèles SARIMA et ARMA semblent être les plus performants en termes de concordance entre les valeurs prédites et les valeurs réelles.

5.1.5 SARIMAX : Intégration d'une variable exogène

Dans notre tentative précédente de modélisation SARIMAX, les résultats obtenus ne se sont pas révélés significativement intéressants. Une piste potentielle pour améliorer ces résultats réside dans la sélection des variables

exogènes intégrées dans notre modèle. Dans la modélisation précédente, une seule variable saisonnière, de nature discrète qui plus est, était utilisée. Une alternative à considérer serait d'incorporer une autre variable, cette fois-ci continue.

Pour explorer cette approche, nous avons identifié la température comme une variable potentiellement fortement corrélée à la quantité de gaz consommée. Dans ce contexte, nous avons choisi de mettre en œuvre cette méthodologie spécifiquement pour la série temporelle associée à Paris. En agrégeant les données des températures moyennes par mois dans la ville de Paris de 2017 à 2022 avec celles de la quantité de gaz consommé, nous obtenons ainsi une nouvelle base de données. Cette dernière nous offre des indications plus riches et introduit une nouvelle variable exogène, à savoir la température moyenne maximale du mois.

TABLE 8 – Statistiques descriptives

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Date	2017-01-01	2018-06-23	2019-12-16	2019-12-16	2021-06-08	2022-12-01
Consommation	5.493×10^8	8.522×10^8	2.410×10^9	3.082×10^9	5.170×10^9	8.327×10^9
Saison	1.00	1.75	2.50	2.50	3.25	4.00
TX	3.700	9.925	16.300	16.543	23.075	29.200

Nous réinstaurons cette fois-ci le modèle SARIMAX en utilisant les mêmes ordres pour les composantes ARMA et saisonnières que nous avions déterminés précédemment. Nous procédons à une nouvelle comparaison entre les variables prédites et les valeurs réellement observées.

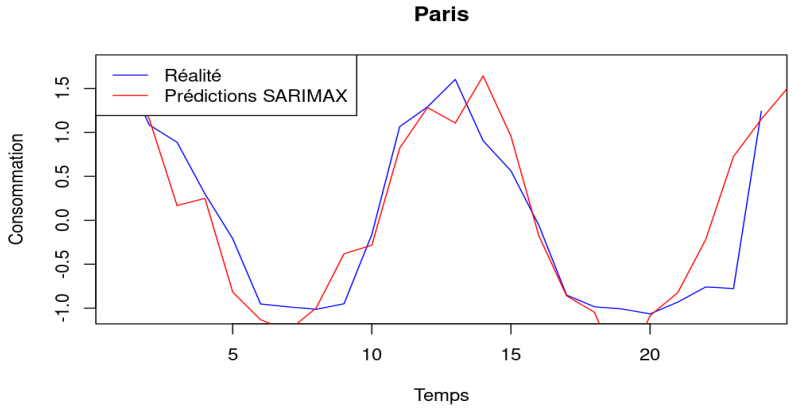


FIGURE 10 – Modélisation SARIMAX avec tempérture

Dans le cadre de notre étude, le modèle SARIMAX, récemment mis en place, a présenté des résultats encourageants, surpassant ainsi nos attentes en termes de précision des prédictions. Cependant, une observation importante émerge de cette analyse : l'efficacité du modèle SARIMAX a été limitée par l'absence de variables exogènes.

Dans l'éventualité où des données supplémentaires, telles que des informations sur la population totale et le nombre de personnes par foyer, seraient disponibles, il serait possible d'améliorer de manière significative la robustesse du modèle. L'intégration de ces variables exogènes pourrait fournir des informations cruciales pour mieux comprendre les dynamiques sous-jacentes du système étudié, contribuant ainsi à affiner nos prédictions. Cette mise en perspective souligne l'importance de la collecte de données complémentaires en vue d'une modélisation plus approfondie et précise.

6 Conclusion

Cette analyse de la consommation de gaz dans différentes métropoles a permis de tirer plusieurs conclusions significatives. Nous avons observé une diminution de la consommation globale en 2020, probablement attribuable aux répercussions de la pandémie de COVID-19. L'analyse exploratoire a mis en évidence des différences notables entre les métropoles, avec Paris et Marseille étant les plus consommatrices. La tendance à la baisse de la consommation pendant l'été suggère une corrélation saisonnière, qui a été confirmée par la création de variables de saisonnalité (Hiver, Printemps, Été, Automne). La décomposition des séries temporelles a révélé une tendance générale à la baisse de la consommation, avec des variations saisonnières claires. Les modèles SARIMA et SARIMAX ont été utilisés pour modéliser ces séries temporelles, et les prédictions ont été évaluées sur des données de test. Le modèle SARIMA nous donne des résultats intéressants mais en terme de prédiction, il manque de l'optimisation.

Enfin, la dénormalisation des prédictions a permis de comparer directement les résultats du modèle avec les données réelles, et le calcul du RMSE a fourni une mesure quantitative de la précision du modèle. En résumé, cette analyse fournit des informations approfondies sur les tendances de consommation de gaz, avec des modèles de prévision robustes qui peuvent être utilisés pour des projections futures, tout en tenant compte des spécificités de chaque métropole. Le modèle sarimax avec la variable saison n'est pas en reste . Des indications telles que l'analyse des activités professionnelle (nombre d'entreprises, nombre de personnes qui se déplacent avec des véhicules alimentés en gaz, le nombre d'habitants par métropole... et le nombre de personnes par foyer) pourraient améliorer les prédictions. Mais c'est variables sont très difficiles à suivre à une granularité mensuelle du a un cout de suivi élevé. Ces données sont suivies généralement à la granularité annuelle donc il nous faudrait agréger notre donnée à une granularité annuelle et passer à une observation sur 7 periodes : 2017 à 2024.

Références

[1] Agence Internationale de l'Énergie. Statistiques mondiales de l'énergie - 2021, 2022.

[2] Gouvernement Français. Plan de relance : Les principales mesures pour la transition écologique, 2022.

[3] Ministère de la Transition Écologique. Bilan énergétique de la france pour 2021, 2022.

[4] J. Sovacool and D. Brown. Urban energy consumption : Different cities, different patterns. *Energy Policy*, 2018.

[1] [3] [2]