



---

## Projet BNP Paribas : Commercial Real Estate

---

Étudiants :

Yanis KAAFARANI  
Tom DELANDE  
Sami RAYAN CHEMLAL  
Martin PUPAT

CentraleSupélec  
Mars 2025

# Table des matières

<b>Introduction</b>	<b>1</b>
<b>1 Étude théorique des modèles de prédiction</b>	<b>3</b>
1.1 ARIMA et XARIMA . . . . .	3
1.1.1 Composantes du modèle ARIMA . . . . .	3
1.1.2 Processus d'ajustement du modèle ARIMA . . . . .	3
1.1.3 Composantes du modèle XARIMA . . . . .	3
1.1.4 Application et ajustement du modèle XARIMA . . . . .	4
1.1.5 Comparaison entre ARIMA et XARIMA . . . . .	4
1.2 LSTM (Long Short-Term Memory) . . . . .	4
1.3 Prophet . . . . .	5
<b>2 Réconciliation des données et preprocessing</b>	<b>7</b>
2.1 Chargement et harmonisation des bases de données . . . . .	7
2.2 Analyse exploratoire des données (EDA) . . . . .	7
2.3 Transformation et uniformisation des variables . . . . .	9
2.4 Traitement des valeurs aberrantes et des valeurs manquantes . . . . .	9
<b>3 Approche agrégée pour la modélisation</b>	<b>12</b>
3.1 Définition et motivation de l'approche agrégée . . . . .	12
<b>4 Entraînement et résultats des modèles</b>	<b>14</b>
4.1 Modèle ARIMA . . . . .	14
4.2 Modèle LSTM . . . . .	16
4.3 Modèle Prophet . . . . .	16
4.4 Comparaison des modèles . . . . .	22
<b>5 Optimisation du portefeuille immobilier selon Markowitz</b>	<b>23</b>
5.1 Introduction du problème . . . . .	23
5.2 Cadre théorique : modèle de Markowitz . . . . .	23
5.3 Description du Jeu de Données . . . . .	24
5.4 Explication des Choix de Paramètres dans la Simulation de la Matrice de Covariance . . . . .	24
5.5 Formulation mathématique du problème . . . . .	25
5.6 Implémentation et résultats . . . . .	26
<b>Conclusion</b>	<b>30</b>
<b>Remerciements</b>	<b>31</b>

# Introduction

## Présentation du projet et des enjeux

Le projet que nous allons présenter se concentre sur la prédition des émissions de CO<sub>2</sub> pour des biens immobiliers tertiaires en France, une problématique cruciale dans le cadre des enjeux climatiques actuels. L'objectif principal de ce projet est de modéliser l'intensité des émissions de gaz à effet de serre (GES) exprimée en kgCO<sub>2</sub>/m<sup>2</sup>/an, en prenant en compte plusieurs variables influençant cette intensité. Cette analyse repose sur deux bases de données fournies par l'ADEME, couvrant respectivement les périodes avant et après juillet 2021. La complexité du projet réside dans la nécessité de traiter et de réconcilier ces deux ensembles de données, qui, bien que portant sur un sujet similaire, diffèrent à la fois par les variables disponibles et par la manière dont l'intensité des émissions est mesurée.

L'un des enjeux principaux est la gestion des données manquantes ou aberrantes, qui peuvent altérer les modèles prédictifs si elles ne sont pas traitées correctement. Ce problème est particulièrement pertinent dans le cadre de l'étude des émissions de CO<sub>2</sub>, car des erreurs dans les données peuvent fausser les prévisions et rendre les résultats moins fiables. Une autre difficulté importante réside dans le fait que les données sont classées sous deux systèmes de catégorisation différents, en fonction de la typologie des bâtiments. Il sera donc nécessaire de mettre en place une procédure de cartographie pour harmoniser ces deux classifications et permettre une comparaison pertinente entre les différentes catégories de biens immobiliers.

Le projet implique également une modélisation des émissions de CO<sub>2</sub> à travers l'utilisation de plusieurs techniques d'analyse prédictive. Les approches proposées, telles que les modèles ARIMA ,XARIMA, LSTM et Prophet, offrent des perspectives différentes pour traiter les données temporelles et intégrer des variables explicatives. Chaque modèle présente des avantages spécifiques : ARIMA et XARIMA pour leur capacité à traiter des séries temporelles classiques et leurs extensions pour la prise en compte de variables exogènes, LSTM pour sa puissance dans l'analyse des données séquentielles complexes et Prophet pour sa flexibilité et sa robustesse face aux tendances saisonnières. L'enjeu ici est de sélectionner le modèle le plus adapté aux données tout en prenant en compte les différentes granularités des informations disponibles, qu'elles soient agrégées par temps ou à un niveau plus fin, tenant compte des caractéristiques de chaque bien immobilier.

Enfin, ce projet s'inscrit dans une démarche plus globale de réduction des émissions de GES dans le secteur immobilier, un secteur essentiel dans la lutte contre le réchauffement climatique. En fournissant des outils prédictifs robustes et fiables, il sera possible d'optimiser la gestion des biens immobiliers, en orientant les investissements vers des solutions plus écologiques et en anticipant les actions nécessaires pour réduire l'empreinte carbone des bâtiments. L'enjeu est donc à la fois scientifique, technique et sociétal : il s'agit de contribuer à la transition énergétique en utilisant les données et les modèles pour aider à la prise de décision et à l'élaboration de stratégies efficaces de réduction des émissions de CO<sub>2</sub> dans le secteur tertiaire.

## Description des bases de données utilisées

Dans le cadre de ce projet, nous utilisons deux bases de données provenant de l'ADEME, l'Agence de l'environnement et de la maîtrise de l'énergie, qui recensent les Diagnostics de Performance Énergétique (DPE) des bâtiments tertiaires en France. Ces données sont essentielles pour évaluer l'impact environnemental des bâtiments commerciaux et pour prédire les intensités d'émission de CO des biens tertiaires.

La première base de données couvre les DPE des bâtiments tertiaires avant juillet 2021. Elle contient des informations détaillées sur les caractéristiques des bâtiments ainsi que leurs émissions de gaz à effet de serre (GES), exprimées en  $\text{kgCO}_2/\text{m}^2/\text{an}$ . La colonne clé qui représente l'intensité des émissions est intitulée *estimation\_ges*. Cette base inclut des variables liées aux caractéristiques des bâtiments telles que la surface, l'année de construction, la localisation, et plusieurs autres paramètres techniques influençant les émissions.

La deuxième base de données correspond aux DPE des bâtiments tertiaires après juillet 2021, avec une mise à jour des variables et des critères d'évaluation. Dans cette version, la variable représentant l'intensité des émissions de CO<sub>2</sub> se nomme *Emission\_GES\_kgCO2/m²/an*. Ce fichier présente également des informations similaires à celles de la première base mais avec quelques différences dans la structure des données. Par exemple, certaines colonnes sont réorganisées et de nouvelles informations sont incluses pour affiner les calculs des émissions.

Une étape cruciale de ce projet consiste à réconcilier ces deux bases de données, en prenant en compte les différences de structure et en harmonisant les variables. Par ailleurs, la base après juillet 2021 inclut une nouvelle catégorisation des secteurs d'activité des bâtiments, ce qui nécessite un mapping avec l'ancienne classification pour garantir la compatibilité des données et permettre une analyse homogène.

Enfin, il est important de souligner que ces bases contiennent des valeurs manquantes ou aberrantes qu'il conviendra de traiter par des méthodes appropriées, comme l'imputation ou la suppression des valeurs extrêmes. Les données seront ensuite utilisées pour construire des modèles de prédition des émissions de CO, en prenant en compte différents facteurs influençant ces émissions selon la catégorie et les caractéristiques des bâtiments.

# Chapitre 1

# Étude théorique des modèles de prédition

## 1.1 ARIMA et XARIMA

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est un modèle statistique largement utilisé pour la prévision des séries temporelles. Il combine trois éléments essentiels : l'autocorrélation (AR), l'intégration (I), et la moyenne mobile (MA), permettant ainsi de modéliser des séries temporelles avec des dépendances linéaires et des tendances.

### 1.1.1 Composantes du modèle ARIMA

Le modèle ARIMA se compose de trois principaux paramètres :

- $p$  : Le paramètre de l'autoregression (AR), qui indique le nombre de périodes précédentes utilisées pour prédire la valeur actuelle. Le modèle autoregressif suppose que la valeur de la série à un instant donné dépend linéairement des valeurs passées.
- $d$  : Le paramètre d'intégration (I), qui est utilisé pour rendre la série stationnaire en éliminant les tendances ou la non-stationnarité de la série temporelle. Cela se fait par la différenciation de la série. Le paramètre  $d$  détermine le nombre de différenciations nécessaires pour rendre la série stationnaire.
- $q$  : Le paramètre de la moyenne mobile (MA), qui représente l'effet des erreurs passées sur la valeur actuelle. Autrement dit, ce paramètre utilise une combinaison linéaire des erreurs de prévision passées pour prédire la valeur à l'instant  $t$ .

La formulation générale du modèle ARIMA est donc la suivante :

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

où  $Y_t$  représente la série temporelle,  $\varepsilon_t$  est l'erreur ou le terme résiduel, et  $\alpha, \beta_1, \dots, \beta_p, \theta_1, \dots, \theta_q$  sont les paramètres à estimer.

### 1.1.2 Processus d'ajustement du modèle ARIMA

Le processus d'ajustement du modèle ARIMA commence généralement par l'analyse de la stationnarité de la série temporelle. Si la série n'est pas stationnaire, une ou plusieurs différenciations sont appliquées pour stabiliser la variance et rendre la série stationnaire. Ensuite, les ordres  $p$ ,  $d$ , et  $q$  sont déterminés par une analyse de l'autocorrélation (ACF) et de l'autocorrélation partielle (PACF) de la série, ainsi que par des critères comme l'AIC (Akaike Information Criterion) et le BIC (Bayesian Information Criterion). Une fois ces paramètres définis, le modèle ARIMA peut être estimé et utilisé pour faire des prévisions.

Le modèle XARIMA (Extended ARIMA) est une extension du modèle ARIMA traditionnel qui permet d'intégrer des variables explicatives externes dans le processus de prédition. Cette version est particulièrement utile lorsqu'on dispose d'informations supplémentaires qui peuvent influencer la série temporelle, mais qui ne font pas partie de la série elle-même. Ces variables explicatives sont également appelées *regressors externes*.

### 1.1.3 Composantes du modèle XARIMA

Le modèle XARIMA conserve les mêmes composantes de base que l'ARIMA, mais y ajoute une composante supplémentaire pour les variables explicatives externes. La structure du modèle devient donc :

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \sum_{k=1}^r \gamma_k X_{t-k} + \varepsilon_t$$

Où  $X_{t-k}$  représente les variables externes à la série temporelle,  $\gamma_k$  sont les coefficients associés aux regressors externes, et  $Y_t$ ,  $\varepsilon_t$  sont respectivement les valeurs de la série et les erreurs à l'instant  $t$ .

Le modèle XARIMA est donc plus flexible que l'ARIMA traditionnel car il permet de prendre en compte des facteurs externes qui peuvent influencer la série temporelle. Ces variables explicatives peuvent inclure des informations économiques, environnementales, sociales ou toute autre donnée externe pertinente.

#### 1.1.4 Application et ajustement du modèle XARIMA

Le processus d'ajustement du modèle XARIMA est similaire à celui du modèle ARIMA, mais avec l'ajout de la sélection et de l'intégration des variables explicatives. Il est essentiel d'identifier les variables qui ont une relation significative avec la série temporelle à prédire. Une analyse de corrélation entre la série temporelle et les variables externes peut être effectuée pour sélectionner les meilleures variables explicatives. Une fois ces variables sélectionnées, le modèle peut être ajusté en utilisant des techniques d'estimation similaires à celles utilisées pour ARIMA. Les prévisions peuvent ensuite être effectuées en incluant les variables externes dans le modèle.

#### 1.1.5 Comparaison entre ARIMA et XARIMA

La principale différence entre ARIMA et XARIMA réside dans l'ajout des variables explicatives externes dans le modèle XARIMA. Si la série temporelle à modéliser est fortement influencée par des facteurs externes, le modèle XARIMA est préférable car il prend en compte ces effets. En revanche, si la série temporelle peut être modélisée efficacement sans l'influence de variables externes, un modèle ARIMA standard pourrait être suffisant. Le choix entre ces deux modèles dépend donc de la nature des données et des informations disponibles.

## 1.2 LSTM (Long Short-Term Memory)

Le modèle LSTM (Long Short-Term Memory) est un type de réseau de neurones récurrent (RNN) conçu pour résoudre le problème du gradient qui disparaît dans les réseaux traditionnels. Il est particulièrement adapté aux séries temporelles, car il permet de modéliser des dépendances à long terme. Les LSTM utilisent une architecture composée de cellules de mémoire qui peuvent conserver et oublier des informations sur des périodes longues.

Les formules essentielles qui gouvernent le fonctionnement d'une cellule LSTM sont les suivantes :

- **Porte d'oubli :** Cette porte décide de quelle information doit être oubliée de la cellule de mémoire en fonction de l'entrée à l'instant  $t$  et de la sortie précédente  $h_{t-1}$ .

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

où  $f_t$  est la sortie de la porte d'oubli,  $W_f$  est le poids de la porte d'oubli,  $b_f$  est le biais, et  $\sigma$  est la fonction sigmoïde.

- **Porte d'entrée :** Cette porte décide de quelle information doit être ajoutée à la cellule de mémoire. Elle est composée de deux parties : une fonction sigmoïde et une fonction tangente hyperbolique.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{aligned}$$

où  $i_t$  est la sortie de la porte d'entrée,  $\tilde{C}_t$  est la candidate pour la cellule de mémoire, et  $\tanh$  est la fonction tangente hyperbolique.

- **Mise à jour de la cellule de mémoire :** La cellule de mémoire est mise à jour en fonction des informations retenues par la porte d'entrée et oubliée par la porte d'oubli.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

où  $C_t$  est la cellule de mémoire à l'instant  $t$ , et  $C_{t-1}$  est la cellule de mémoire à l'instant  $t-1$ .

- **Porte de sortie** : Cette porte détermine quelle information de la cellule de mémoire doit être envoyée comme sortie. Elle utilise également une fonction sigmoïde et une fonction tangente hyperbolique pour moduler la sortie.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

où  $o_t$  est la sortie de la porte de sortie, et  $h_t$  est la sortie du LSTM à l'instant  $t$ .

LSTM est généralement composé d'une séquence d'étapes dans lesquelles chaque porte est alimentée avec les entrées, les sorties et les valeurs précédentes pour produire une sortie en fonction des informations pertinentes pour la tâche.

L'architecture de la cellule LSTM est représentée ci-dessous :

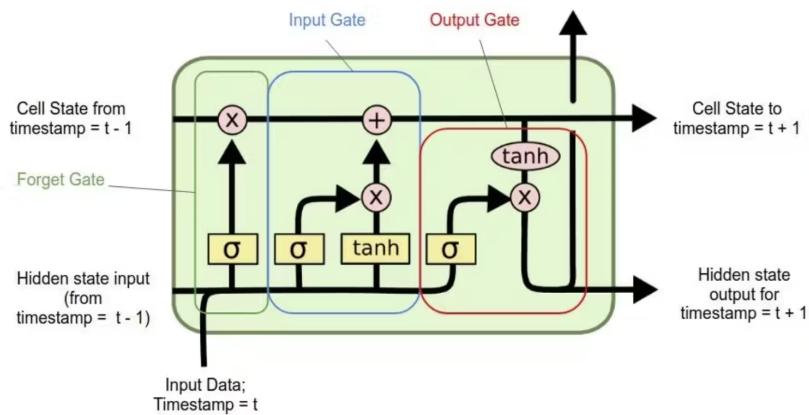


FIGURE 1.1 – Architecture d'une cellule LSTM

Les LSTM sont capables de modéliser des séquences complexes et peuvent prendre en compte des dépendances temporelles longues grâce à leur capacité de conserver l'information pendant de nombreuses étapes temporelles.

### 1.3 Prophet

Le modèle Prophet est un modèle de prévision développé par Facebook qui est particulièrement adapté aux séries temporelles comportant des tendances saisonnières et des effets de jours fériés ou de jours spéciaux. Prophet est conçu pour être flexible, robuste et capable de gérer des séries temporelles avec des données manquantes et des changements de tendance. Ce modèle est également très efficace dans les cas où les données sont irrégulières ou non stationnaires.

Le modèle Prophet se base sur la décomposition de la série temporelle en trois composantes principales :

- **Tendance** : La tendance générale des données, qui peut être croissante, décroissante ou constante. Prophet permet de capturer les changements dans la tendance de manière souple grâce à une fonction de lissage.
- **Saisonnalité** : Les effets saisonniers récurrents à différentes échelles temporelles, tels que la saisonnalité quotidienne, hebdomadaire ou annuelle.
- **Effets des jours fériés** : L'impact des jours fériés ou des événements spéciaux sur les séries temporelles, en modélisant l'influence de ces jours spécifiques sur les données.

La modélisation des séries temporelles dans Prophet repose sur la décomposition additif ou multiplicatif des données :

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

où : -  $y(t)$  est la valeur observée de la série à l'instant  $t$ , -  $g(t)$  est la fonction de tendance à l'instant  $t$ , -  $s(t)$  est la fonction de saisonnalité à l'instant  $t$ , -  $h(t)$  est l'effet des jours fériés à l'instant  $t$ , -  $\epsilon_t$  est le bruit résiduel, supposé être normalement distribué.

La fonction de tendance  $g(t)$  dans Prophet est définie de manière flexible, permettant d'avoir soit une tendance linéaire, soit une tendance logistique pour des séries temporelles avec des limites asymptotiques. La fonction de saisonnalité  $s(t)$  est modélisée par des sinusoïdes qui capturent les effets saisonniers périodiques.

Le modèle fait appel à une série de paramètres pour ajuster les composantes de la tendance et de la saisonnalité. Par exemple, la tendance est modélisée comme suit :

$$g(t) = \text{Piecewise Linear or Logistic Growth Model}$$

où un changement dans la tendance est capturé par un seuil ou une rupture (changements dans les paramètres à un moment donné de la série temporelle).

La saisonnalité est modélisée par la décomposition en fonction périodique, telle que :

$$s(t) = \sum_{i=1}^k \alpha_i \cos\left(\frac{2\pi i t}{T_i}\right) + \beta_i \sin\left(\frac{2\pi i t}{T_i}\right)$$

où  $k$  est le nombre de composantes saisonnières,  $T_i$  est la période saisonnière, et  $\alpha_i$  et  $\beta_i$  sont les paramètres associés.

Pour les jours fériés, Prophet ajuste une fonction de type indicateur qui permet d'ajouter des effets spécifiques à certains jours :

$$h(t) = \sum_{j=1}^m \gamma_j \cdot \mathbb{I}(t \in \text{holiday}_j)$$

où  $\gamma_j$  est l'impact du  $j$ -ième jour férié et  $\mathbb{I}(t \in \text{holiday}_j)$  est une fonction indicatrice valant 1 si le jour  $t$  est un jour férié et 0 sinon.

L'optimisation du modèle s'effectue par ajustement des paramètres  $\alpha_i, \beta_i, \gamma_j$ , ainsi que des paramètres de la fonction de tendance et de saisonnalité.

Une fois les paramètres ajustés, le modèle peut effectuer des prévisions sur de nouvelles données temporelles. Prophet est flexible dans son approche, permettant d'incorporer des changements de tendance, des effets saisonniers multiples et des jours fériés spécifiques.

Le modèle Prophet est robuste et facile à utiliser, et il est conçu pour être accessible aux utilisateurs non-experts tout en permettant des ajustements fins pour les utilisateurs plus avancés. Grâce à sa flexibilité, il peut être appliqué à un large éventail de séries temporelles, y compris celles comportant des anomalies ou des irrégularités dans les données.

## Chapitre 2

# Réconciliation des données et preprocessing

### 2.1 Chargement et harmonisation des bases de données

L'analyse commence par l'importation des bibliothèques essentielles telles que **NumPy**, **Pandas** et **Matplotlib**, qui permettent de manipuler et visualiser les données. Les bases de données utilisées sont chargées à partir de fichiers CSV : **dpe-tertiaire.csv** et **dpe-v2-tertiaire-2.csv**, qui contiennent des informations sur les diagnostics de performance énergétique (DPE) des bâtiments tertiaires en France.

Les coordonnées géographiques ne sont pas sous le même format dans les deux bases de données. Dans la première base, elles sont dans le système WGS 84, alors que dans la deuxième, elles sont dans le système Lambert 93. Une première étape consiste donc à convertir les coordonnées cartographiques des bâtiments de la deuxième base, du système Lambert 93 vers le système WGS 84 (latitude/longitude). Pour cela, la bibliothèque **pyproj** est utilisée afin de transformer les coordonnées en appliquant une conversion sur chaque ligne de la base de données.

Par la suite, le nombre et le nom des colonnes n'étant pas identiques entre les deux bases, un processus d'harmonisation des colonnes a été réalisé. Les noms des colonnes de la deuxième base ont été remplacés par ceux correspondant aux noms des colonnes de la première base grâce à un mapping. Nous avons ainsi conservé uniquement les colonnes présentes de manière équivalente dans les deux bases.

Ces différentes étapes ont alors rendu possible la concaténation des deux bases pour obtenir une seule base sur laquelle travailler.

Une conversion des dates est également appliquée. Les colonnes sont converties au format **datetime** afin de permettre des analyses temporelles plus précises. Ces dates sont ensuite exploitées pour extraire l'année, le mois et le jour de la visite diagnostique.

Nous avons décidé d'uniquement garder la colonne '**date\_visite\_diagnostiqueur**' car c'est elle qui correspond physiquement à la date de l'émission. En effet, nous avons remarqué que l'établissement du DPE pouvait se réaliser bien après la visite, donc la date de la visite semble avoir plus de sens physiquement.

Concernant les types de bâtiments, un regroupement est mis en place pour classifier les différents secteurs d'activité en catégories plus cohérentes. Les secteurs d'activités sont remplacés par des appellations uniformes, permettant d'obtenir une variable **asset\_cre**, qui regroupe les typologies d'usage des bâtiments tertiaires de manière plus structurée.

### 2.2 Analyse exploratoire des données (EDA)

L'exploration des données se poursuit par une analyse des valeurs manquantes. Une visualisation est réalisée à l'aide d'une **matrice de chaleur (heatmap)** avec la bibliothèque **Seaborn**, permettant d'identifier rapidement la distribution des valeurs manquantes.

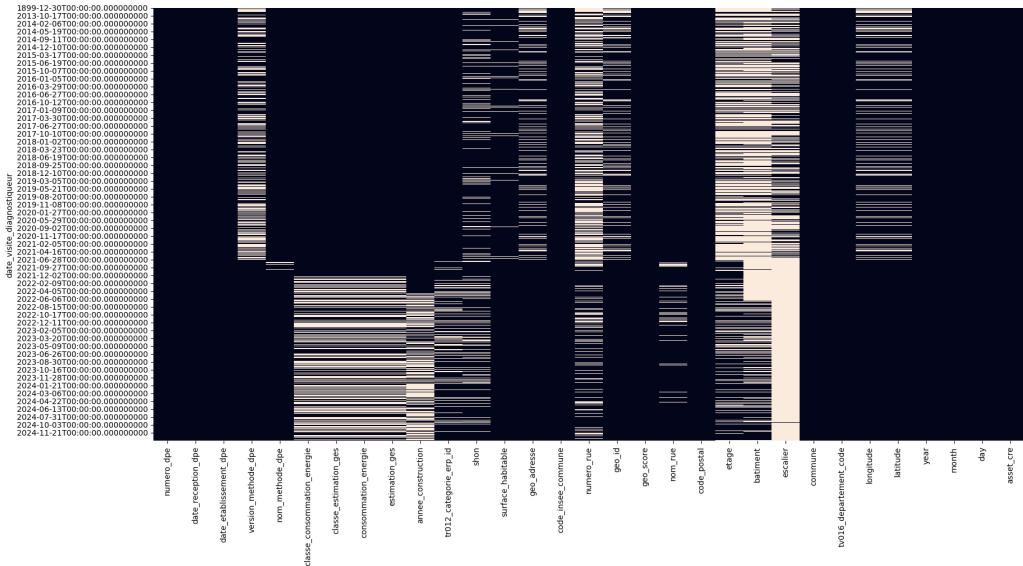


FIGURE 2.1 – Matrice des valeurs manquantes

Nous avons également calculé le taux de valeurs manquantes par colonne afin de déterminer l'ampleur du problème : nous avons supprimé les colonnes qui avaient plus de 40% de valeurs manquantes.

Nous avons aussi supprimé les colonnes inutiles à la prédiction de futurs intensités carbone, dans la mesure où les données de ces colonnes sont obtenus lors de l'établissement d'un dpe, or la prédiction se fait sans avoir accès à ces données : nous pouvons donc supprimer les colonnes numero\_dpe, classe\_estimation\_ges, classe\_consommation\_énergie, consommation\_énergie, nom\_méthode\_dpe et version\_méthode\_dpe.

Enfin, une analyse des types de données est effectuée, suivie d'une conversion de certaines colonnes au format numérique lorsque nécessaire.

Une représentation graphique sous forme de **diagramme circulaire** est générée pour visualiser la répartition des types de données dans l'ensemble du jeu de données.

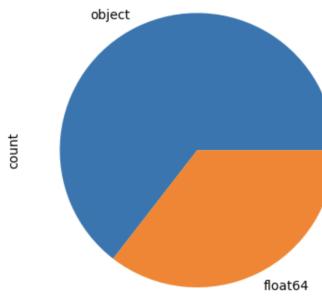


FIGURE 2.2 – Répartition des types de données

Par exemple, les colonnes **shon** et **surface\_habitable**, qui contiennent parfois des valeurs non numériques, sont d'abord nettoyées avant d'être converties en float.

Une représentation graphique sous forme de **diagramme circulaire** est générée pour visualiser la répartition des types de données dans l'ensemble du jeu de données.

## 2.3 Transformation et uniformisation des variables

Certaines variables ont été retravaillées afin de les rendre plus exploitables :

Le champ geo-id présentait une variabilité de format. Une transformation a été appliquée pour ne conserver que les neuf premiers caractères utiles, supprimant ainsi les suffixes variables et garantissant une meilleure standardisation.

Le code INSEE des communes et l'identifiant géographique ont été analysés. Il a été déterminé que geo-id incluait un niveau de granularité plus fin. Cependant, compte tenu du nombre de valeurs manquantes dans cette variable, celle-ci a été supprimée au profit du code INSEE. De plus commune peut être vu comme une version encodée de code\_insee\_commune, nous avons donc supprimé cette colonne.

Les codes des départements ont été harmonisés en traitant les exceptions pour la Corse, dont les codes '2A' et '2B' ont été convertis en '96' et '97', respectivement.

Les variables catégorielles ont été traitées par un encodage numérique afin d'être intégrées dans les modèles prédictifs. Par exemple, la variable asset-cre a été remplacée du target encoding selon sa médiane de sur la colonne estimation\_ges pour sein de chaque catégorie. Nous avons utilisé la médiane car cela est plus propice à la distribution non symétrique de estimation\_ges.

Cette étape d'harmonisation permet d'obtenir une base de données propre et standardisée, facilitant les analyses ultérieures ainsi que l'application de modèles statistiques et d'apprentissage automatique.

## 2.4 Traitement des valeurs aberrantes et des valeurs manquantes

Dans toute analyse de données, la gestion des valeurs aberrantes et des valeurs manquantes est une étape cruciale pour garantir la robustesse des modèles prédictifs et l'intégrité des résultats.

Dans un premier temps, certaines variables ont été soumises à des critères de filtrage afin d'éliminer les valeurs manifestement erronées ou aberrantes. Les règles appliquées sont les suivantes :

- **Remplacement des valeurs nulles par des NaNs** : Les valeurs manquantes ont été remplacées par des NaNs.
- **Année de construction** : seules les valeurs supérieures ou égales à 1900 ont été conservées.
- **Estimation des émissions de GES** : les valeurs inférieures à 0 ont été supprimées.
- **Score géographique** : seules les valeurs supérieures ou égales à 0.3 ont été retenues.
- **Coordonnées géographiques** :
  - La latitude a été contrainte entre 40 et 55.
  - La longitude a été restreinte entre -6 et 11.

Pour les valeurs manquantes, plusieurs approches ont été adoptées en fonction de la nature des données. Certaines variables ont été jugées non exploitables en raison d'un trop grand nombre de valeurs absentes et ont donc été supprimées. Pour d'autres, une imputation par des valeurs médianes ou des statistiques pertinentes a été réalisée.

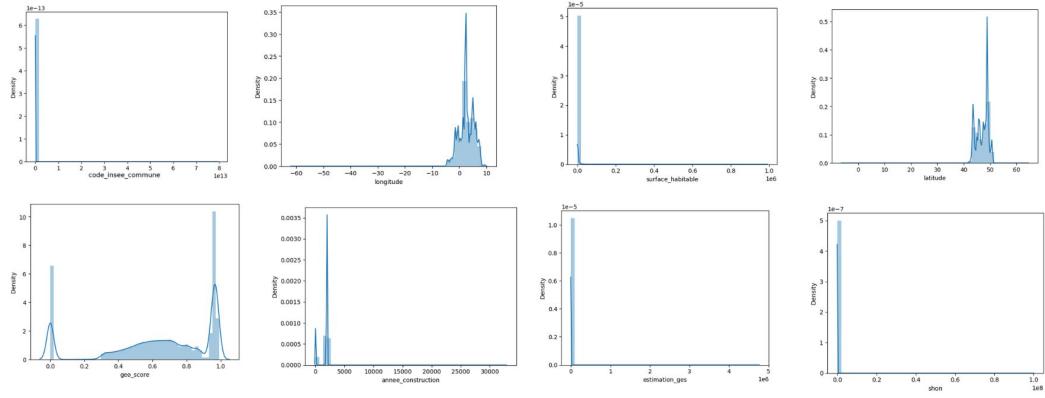


FIGURE 2.3 – Distribution avant preprocessing

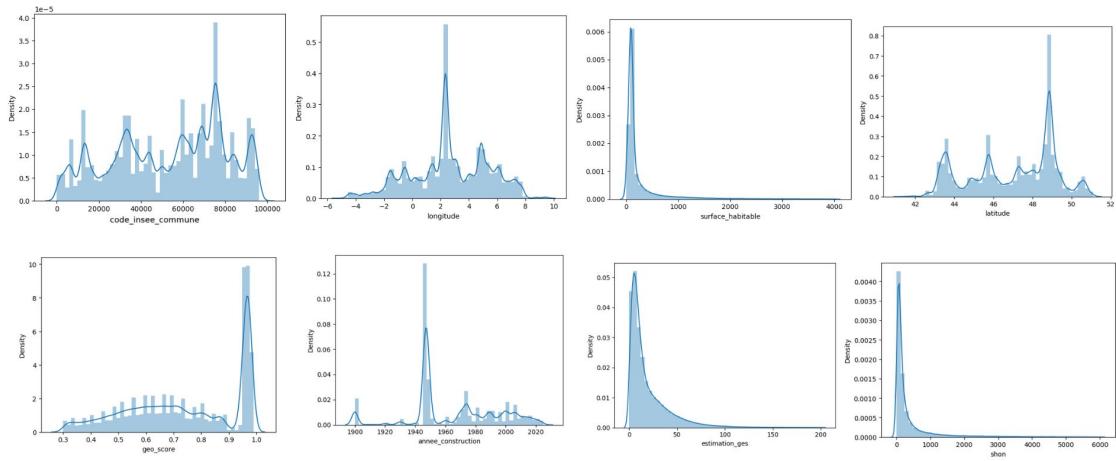


FIGURE 2.4 – Distribution après preprocessing

Une approche plus avancée a été utilisée pour certaines variables critiques : l'imputation par les  $k$  plus proches voisins (K-Nearest Neighbors, KNN). Cet algorithme repose sur l'idée que les valeurs manquantes peuvent être estimées à partir des observations les plus similaires dans l'espace des caractéristiques. Plus précisément, on identifie, pour chaque donnée incomplète, les  $k$  observations les plus proches selon une métrique de distance, typiquement la distance euclidienne, et on impute la valeur manquante par la moyenne ou la médiane des voisins identifiés. Ce procédé permet de conserver la structure globale des données et d'éviter des biais qui pourraient être introduits par des imputations simples.

L'algorithme KNN a été appliqué à la variable estimation\_ges du jeu de données en sélectionnant les caractéristiques pertinentes pour garantir une estimation cohérente des valeurs manquantes. Cette approche a permis d'améliorer la complétude des données tout en minimisant la perte d'information, préparant ainsi la base à des analyses plus approfondies.

Finalement nous obtenons un jeu de données de 400 000 lignes au lieu d'environ 120 000 si nous avions directement supprimé les valeurs manquantes.

Enfin nous avons tracé la matrice de corrélation du jeu de données pour comprendre les relations importantes entre ces données, utiles pour nos futurs modèles.

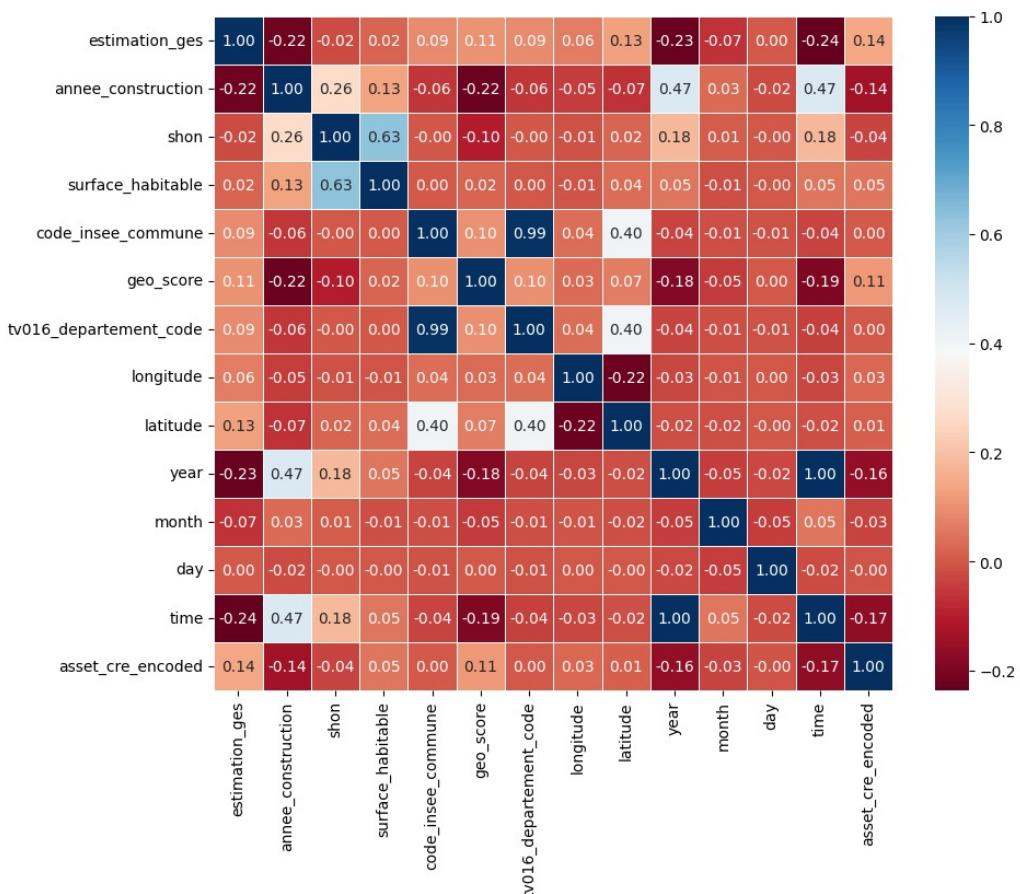


FIGURE 2.5 – Matrice de corrélation après preprocessing

Cette phase d'exploration a permis d'obtenir des données en quantité suffisante, pouvant ensuite être exploitées dans des modèles de machine learning.

## Chapitre 3

# Approche agrégée pour la modélisation

### 3.1 Définition et motivation de l'approche agrégée

L'approche agrégée consiste à calculer les intensités d'émission de CO de manière pondérée, en fonction des catégories de biens et des différentes périodes de temps, à savoir par jour, mois et année. Cette approche vise à fournir des estimations globales des émissions de CO, en tenant compte des surfaces des biens et de leurs différentes caractéristiques. Le processus d'agrégation se déroule en trois étapes principales, chacune correspondant à un niveau temporel différent.

La première étape de l'agrégation consiste à regrouper les données par jour. Pour chaque bien, les valeurs d'intensité d'émission sont pondérées par la surface du bien, en utilisant la variable `surface_habitable` pour la base avant juillet 2021 et `surface_utile` pour la base après juillet 2021. Une fois cette pondération effectuée, les intensités d'émission pondérées sont agrégées par catégorie de bien. Ce calcul se fait en prenant la moyenne pondérée des intensités d'émission pour chaque catégorie, sur la période quotidienne. Ce niveau d'agrégation permet d'observer les fluctuations d'émission à court terme pour chaque type d'actif.

La deuxième étape de l'agrégation consiste à regrouper les données sur une période hebdomadaire. Les intensités d'émission de chaque bien sont à nouveau pondérées par leur surface respective, et l'agrégation est effectuée sur l'ensemble de la semaine. Cette approche permet de calculer l'intensité d'émission moyenne pondérée pour chaque catégorie de bien, sur une période d'une semaine.

La dernière étape consiste à regrouper les données par mois. À l'instar de l'agrégation journalière, les intensités d'émission sont pondérées par la surface des biens et agrégées sur une période mensuelle. Les résultats obtenus sont les moyennes pondérées des intensités d'émission mensuelles pour chaque catégorie de bien. Cette agrégation permet de mieux capturer les tendances saisonnières et d'obtenir des données mensuelles globales, qui offrent une vue d'ensemble des émissions sur le long terme.

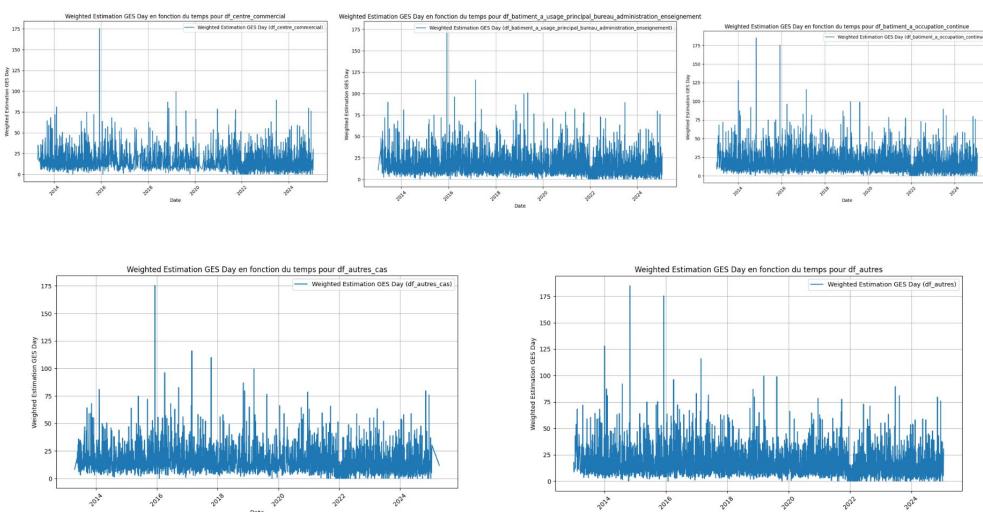


FIGURE 3.1 – Agrégation par jour

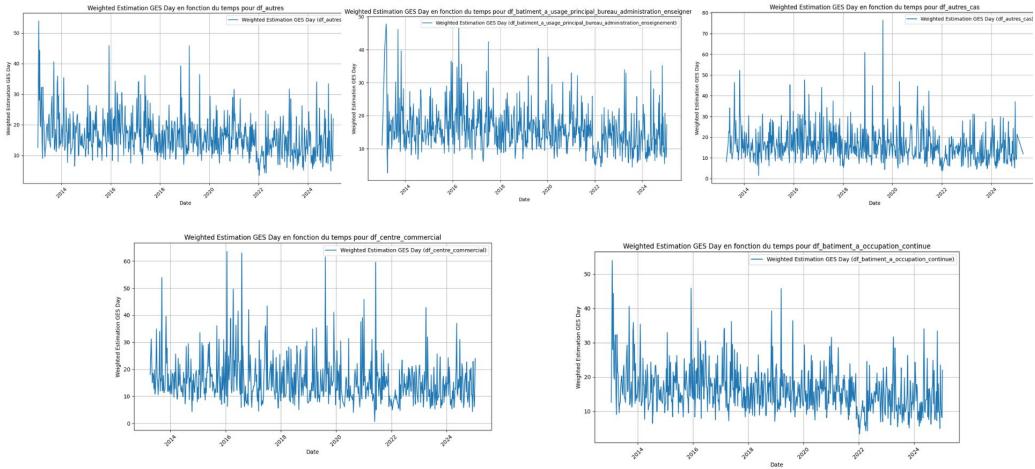


FIGURE 3.2 – Agrégation par semaine

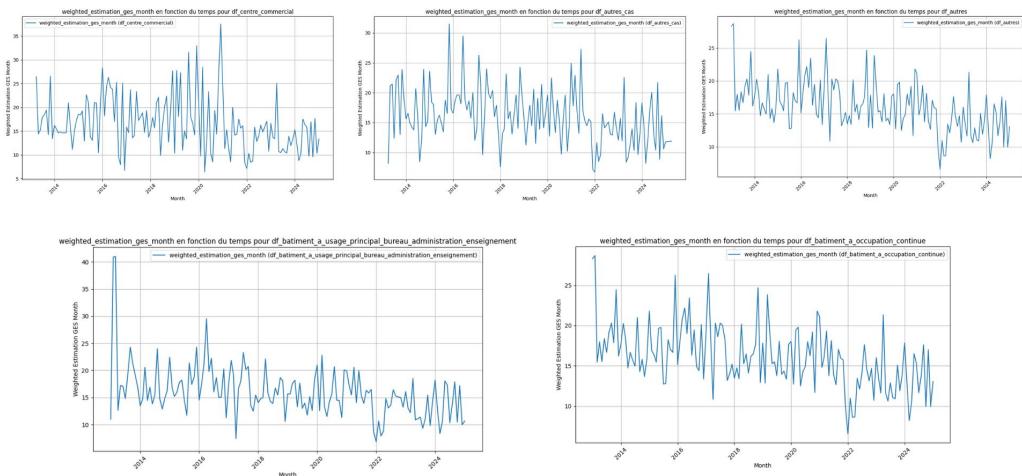


FIGURE 3.3 – Agrégation par mois

Nous avons décidé de nous concentrer sur l'aggrégation par mois car il y avait beaucoup moins de variance sur les émissions.

En résumé, chaque étape de l'agrégation permet de réduire la granularité des données tout en offrant une vue d'ensemble sur les émissions de CO<sub>2</sub> par catégorie de biens. L'agrégation par jour, sem et année permet de capturer différentes dynamiques temporelles, depuis les fluctuations à court terme jusqu'aux tendances de fond. Ces séries temporelles agrégées fournissent une base solide pour la modélisation des émissions de CO<sub>2</sub>, en facilitant l'analyse des tendances et la prédition des évolutions futures des émissions pour chaque type de bien.

# Chapitre 4

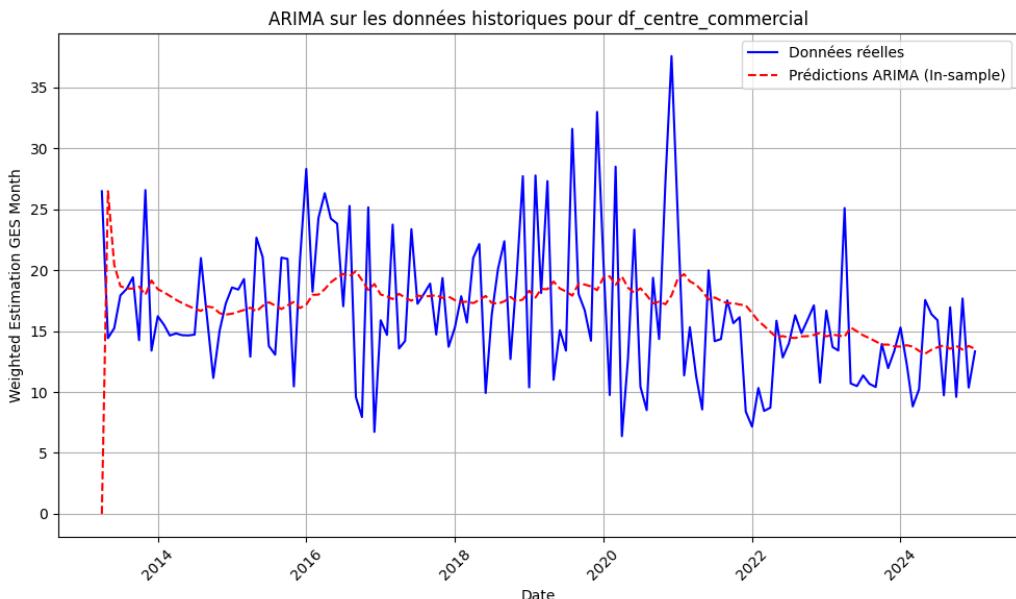
## Entraînement et résultats des modèles

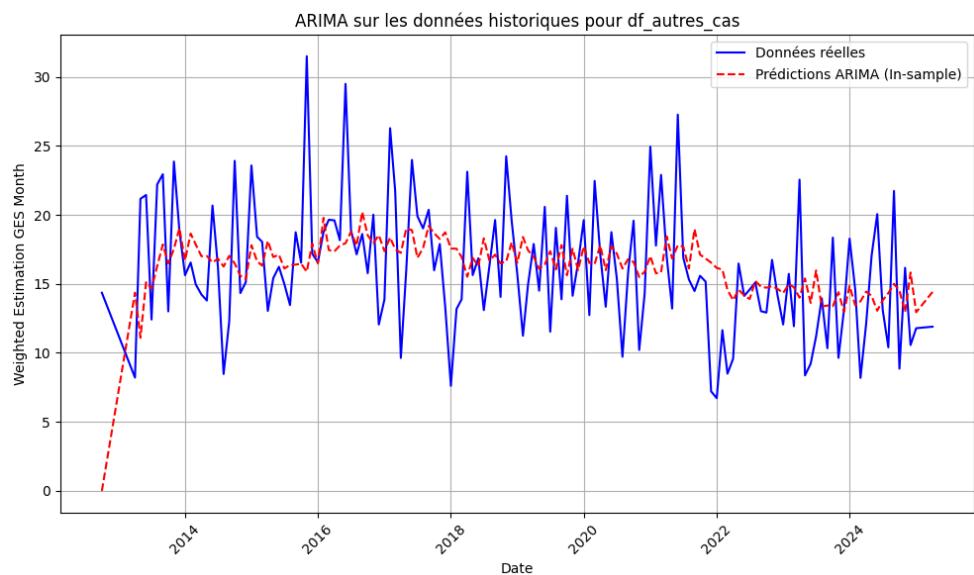
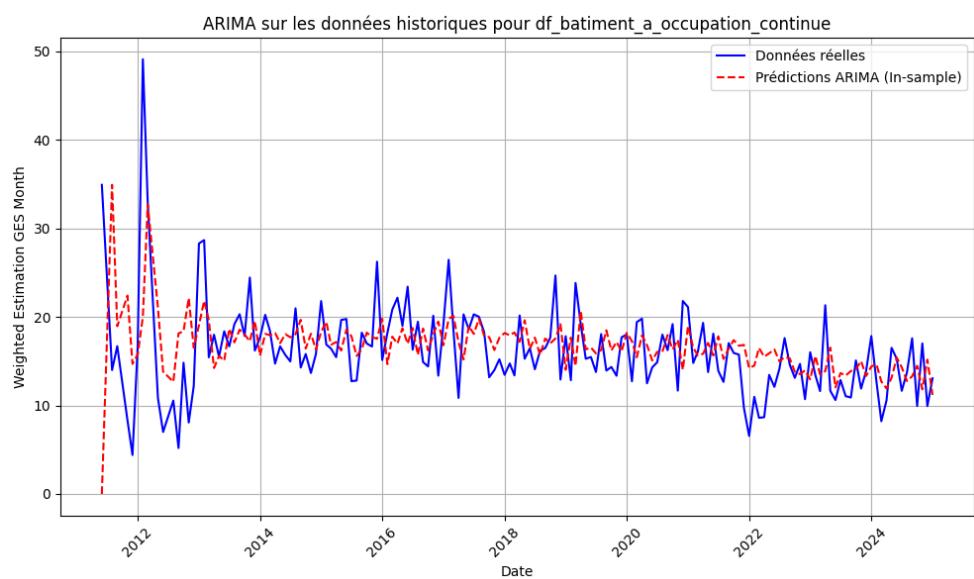
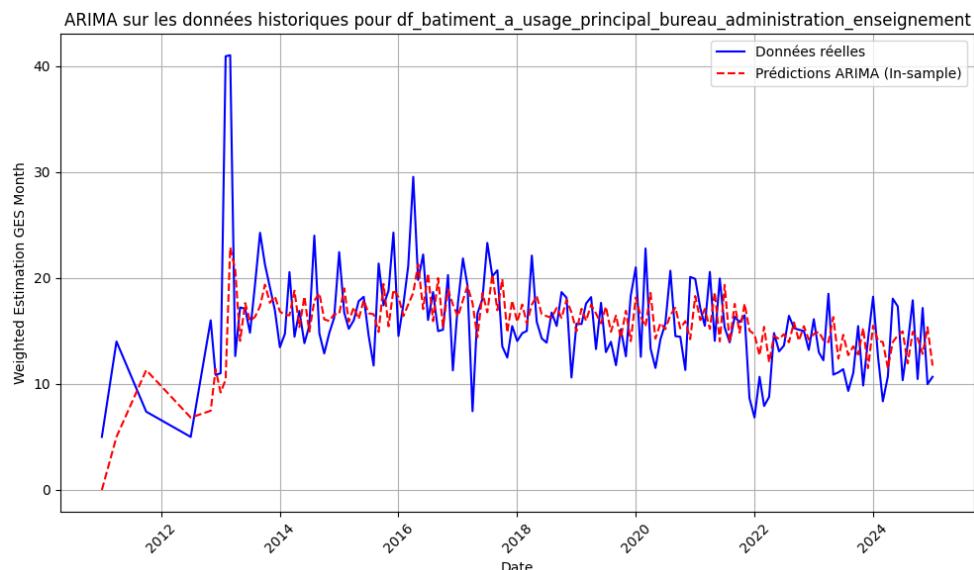
### 4.1 Modèle ARIMA

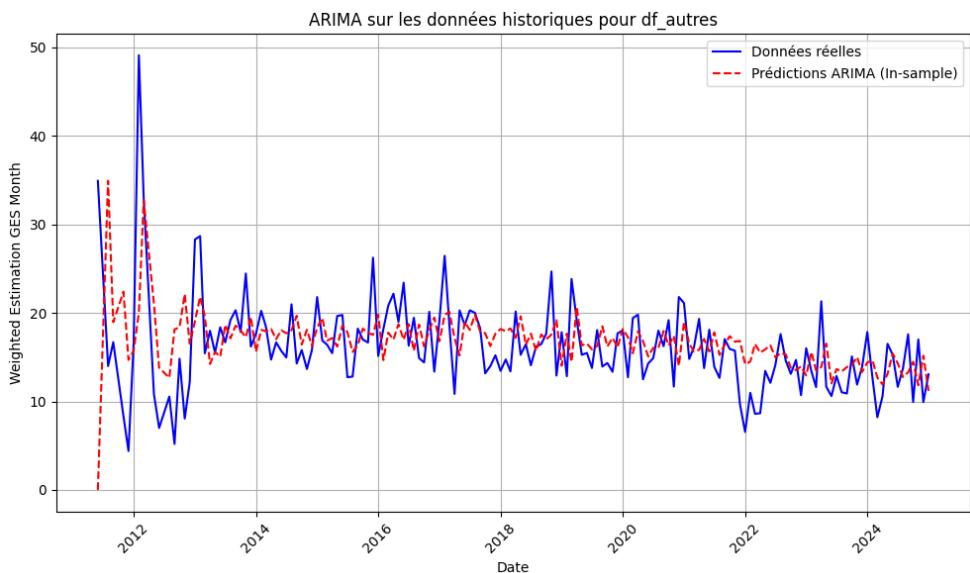
L'entraînement du modèle ARIMA repose sur une approche de séries temporelles où l'intensité des émissions de CO<sub>2</sub> est modélisée en fonction du temps et d'autres variables explicatives. Les données ont d'abord été préparées en convertissant les dates d'inspection en format datetime, en créant une variable *month* basée sur les périodes mensuelles et en ajoutant une colonne *days\_elapsed* pour capturer l'évolution temporelle des émissions. Une première analyse exploratoire a révélé une distribution asymétrique des émissions de CO, suggérant une forte variabilité entre les actifs.

L'entraînement du modèle a nécessité une transformation des données en moyennes pondérées par surface habitable, afin de lisser les fluctuations extrêmes. La sélection automatique des hyperparamètres a été réalisée à l'aide de la fonction *auto\_arima* de la bibliothèque PMDARIMA, qui optimise les ordres ( $p, d, q, P, D, Q$ ) du modèle ARIMA saisonnier en minimisant un critère d'information (AIC ou BIC). Une fois les paramètres déterminés, le modèle a été ajusté sur les données agrégées mensuellement en prenant en compte la dynamique temporelle et les tendances sous-jacentes. Une validation a été réalisée par une approche de type *walk-forward validation*, où les prédictions sont comparées aux valeurs réelles sur une période de test distincte.

Les résultats obtenus montrent que le modèle XARIMA capture efficacement la tendance des émissions de CO et permet d'anticiper leur évolution future avec une précision satisfaisante. La comparaison des valeurs prédites et des valeurs réelles met en évidence une bonne corrélation, bien que certaines fluctuations locales restent difficiles à modéliser en raison de la variabilité intrinsèque des actifs immobiliers. L'analyse des résidus du modèle montre que ceux-ci sont globalement stationnaires et suivent une distribution normale, confirmant la validité du modèle pour cette problématique.







## 4.2 Modèle LSTM

Le modèle LSTM a été envisagé pour prévoir les émissions de GES en exploitant sa capacité à modéliser des dépendances temporelles complexes. Nous avons d'abord transformé les données en séquences d'entrée et de sortie en respectant leur ordre chronologique, puis construit un modèle comportant plusieurs couches LSTM ainsi que des couches de dropout pour limiter le surapprentissage. L'objectif était de tirer parti des caractéristiques non linéaires et de la saisonnalité présentes dans les données.

Pour optimiser le modèle, nous avons mis en place une recherche sur grille (grid search) qui explore différentes combinaisons d'hyperparamètres, notamment le nombre d'unités dans les couches LSTM, les taux de dropout, le nombre d'époques, la taille des batches et le taux d'apprentissage, tout en utilisant une validation croisée adaptée aux séries temporelles pour préserver l'ordre chronologique.

Cependant, en raison de la taille importante de nos données et de la complexité du modèle, nous avons rencontré d'importants problèmes de convergence, le modèle n'ayant pas réussi à converger de manière satisfaisante malgré plusieurs ajustements d'hyperparamètres. Ce travail est présenté en détail dans le notebook associé.

## 4.3 Modèle Prophet

Les prévisions fournies par XARIMA constituent un outil robuste pour anticiper les émissions et ajuster les stratégies d'investissement en conséquence. Une analyse plus fine pourrait être réalisée en intégrant des variables explicatives supplémentaires telles que les tendances macroéconomiques ou les politiques environnementales spécifiques à certains secteurs d'activité.

Le dernier modèle auquel nous nous sommes intéressés est donc le modèle Prophet. Après un import des bibliothèques pertinentes, nous avons créé une fonction `prophet_hyperparameter_tuning` prenant en paramètres un nombre d'époques et une vitesse d'apprentissage et qui optimise les deux hyperparamètres pertinents de Prophet, `changepoint` et `seasonality`, en faisant un apprentissage par descente de gradient avec la bibliothèque `sklearn`. Notons que les paramètres sont bornés à la main ; cela permet d'assurer que le modèle Prophet s'exécute correctement, car des paramètres négatifs ou trop grands conduisent à des erreurs.

Ensuite, les bases de données partielles (une pour chaque catégorie de bâtiment) `df`, agrégées par mois, sont formatées pour le modèle Prophet, en standardisant le format de dates et en renommant les colonnes `ds` pour la date et `y` pour l'ordonnée (ici l'estimation de GES). Ensuite, un modèle Prophet est optimisé en hyperparamètres sur chacune d'entre elles, avant de faire une prédition (dont les résultats sont présentés plus bas). Enfin, nous traçons ces résultats à l'aide de `matplotlib`.

Le modèle Prophet, même s'il est assez obscure dans sa manière de fonctionner, nous donne quand même des indicateurs intéressants avec lesquels travailler, et qui nous permettent de valider la bonne précision du modèle. Nous pouvons faire des prédictions sur les données historiques pour voir si le modèle est bien calibré, et sur les données futures pour bien prédire les informations suivantes. Enfin, nous pouvons extraire la tendance de la prédition (c'est-à-dire la linéarisation de la courbe de prédictions sur le long terme) ainsi que sa saisonnalité (c'est-à-dire les motifs répétitifs à intervalles réguliers, ici annuels).

Ces prédictions et indicateurs, discriminés entre les différents types de bâtiments, sont présentés dans les divers graphiques dès la page suivante.

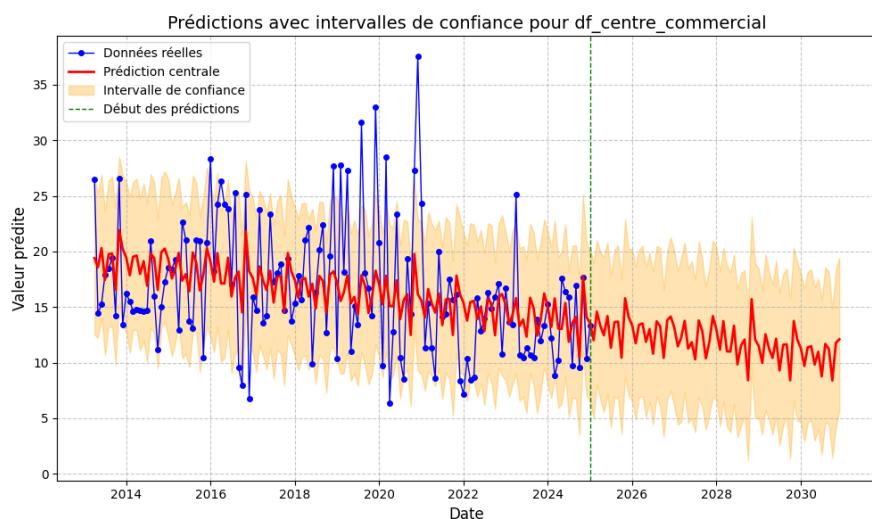


FIGURE 4.1 – Prédictions pour centre\_commercial

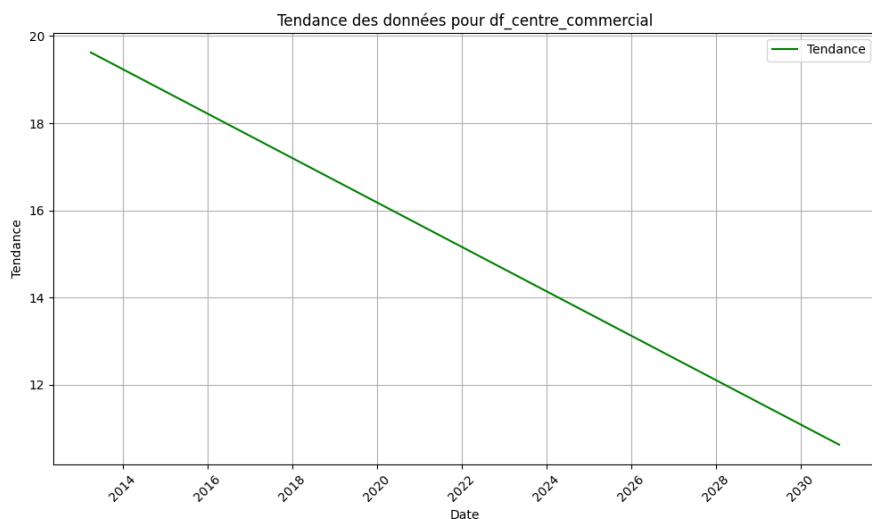


FIGURE 4.2 – Tendance pour centre\_commercial

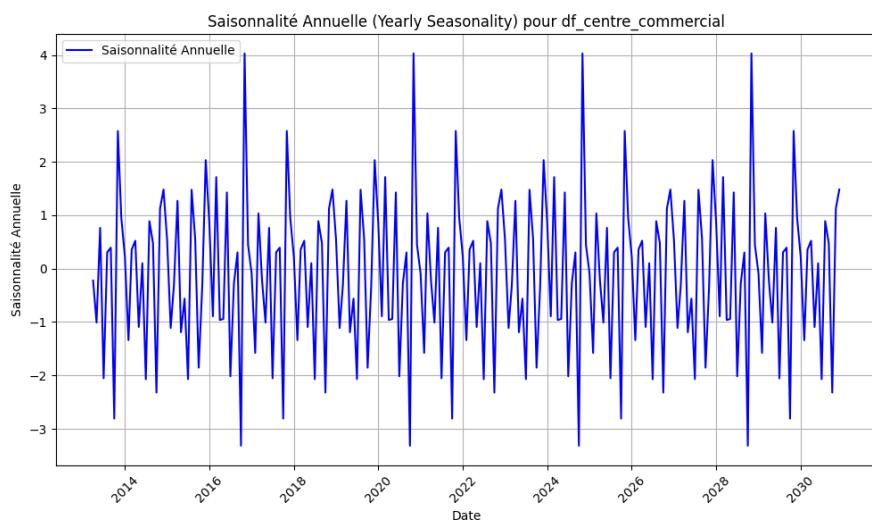


FIGURE 4.3 – Saisonalité pour centre\_commercial

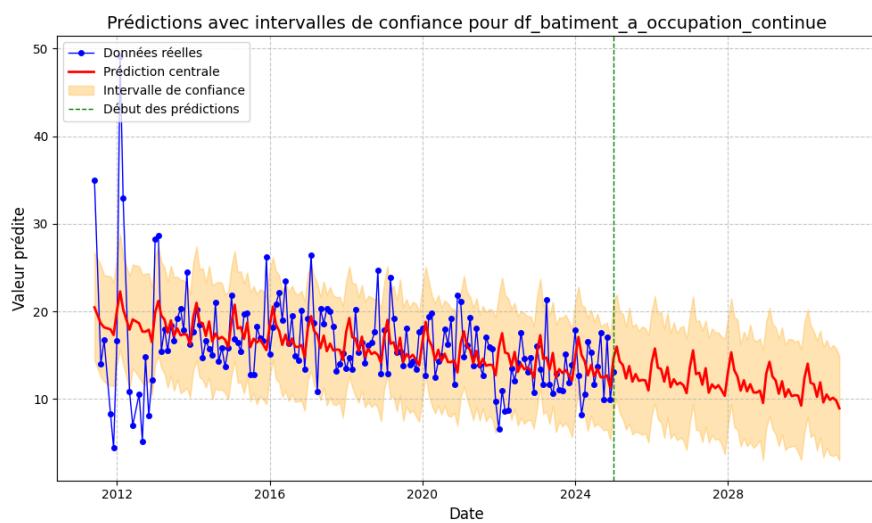


FIGURE 4.4 – Prédictions pour batiment\_a\_occupation\_continue

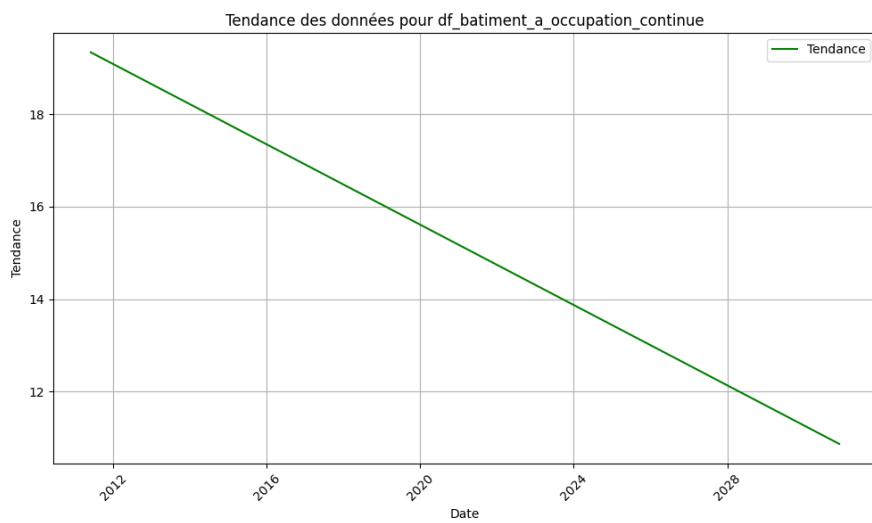


FIGURE 4.5 – Tendance pour batiment\_a\_occupation\_continue

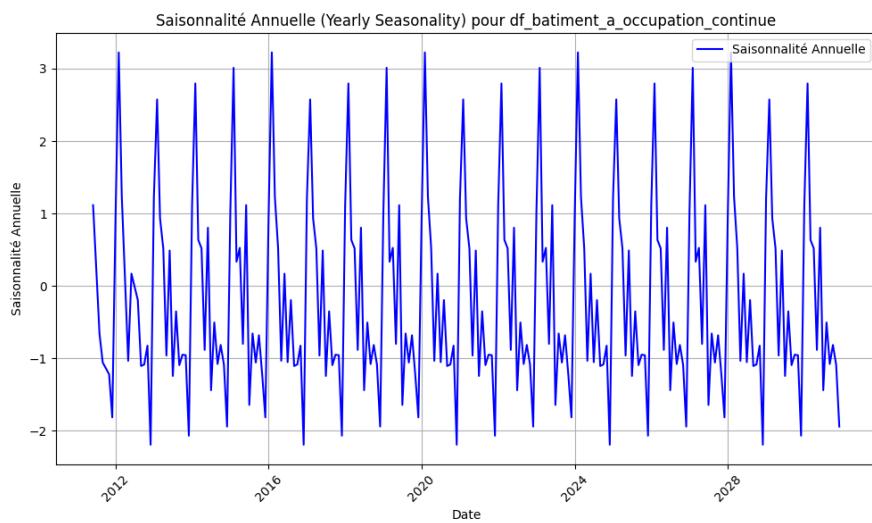


FIGURE 4.6 – Saisonnalité pour batiment\_a\_occupation\_continue

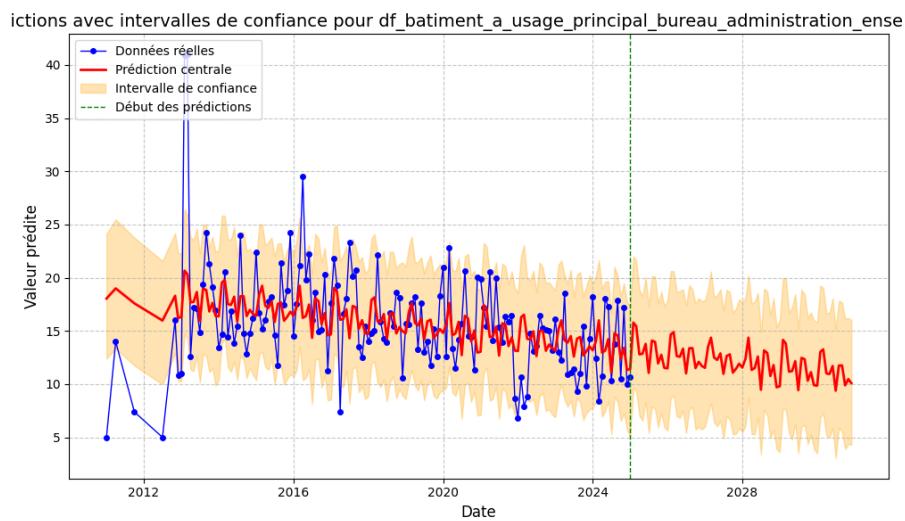


FIGURE 4.7 – Prédiction pour batiment\_a\_usage\_principal\_bureau\_administration\_enseignement

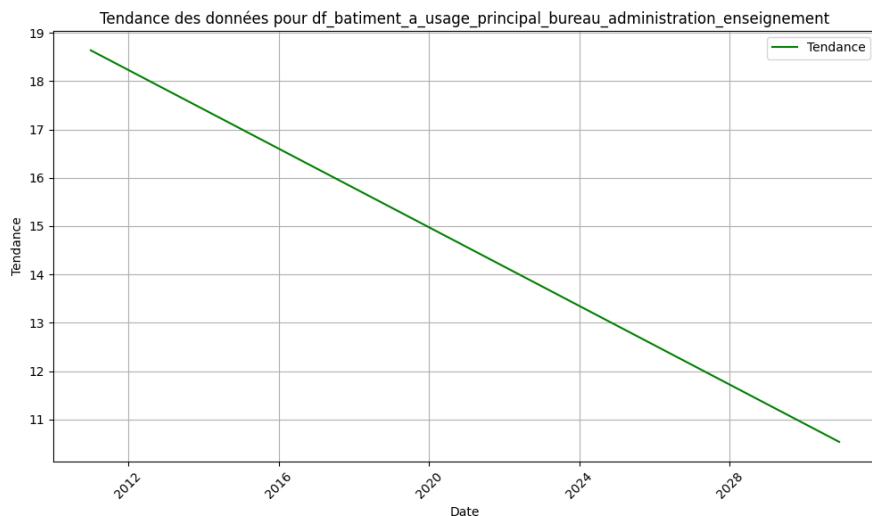


FIGURE 4.8 – Tendance pour batiment\_a\_usage\_principal\_bureau\_administration\_enseignement

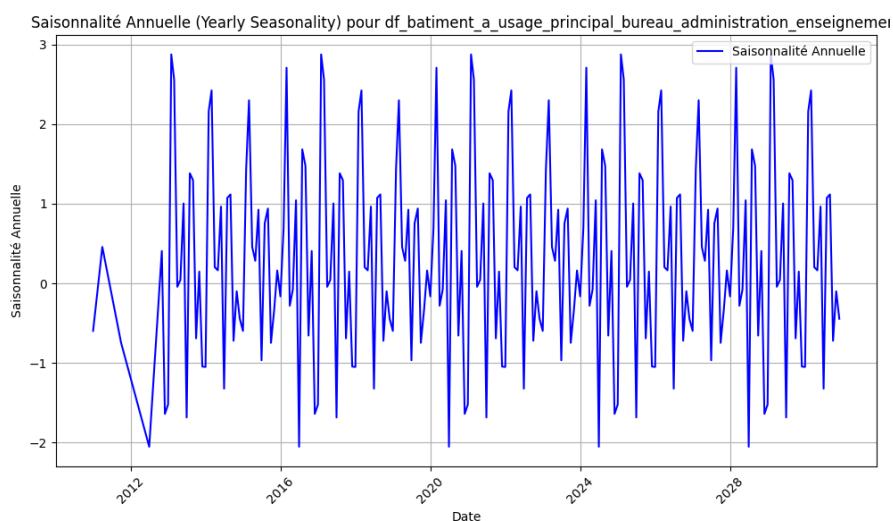


FIGURE 4.9 – Saisonnalité pour batiment\_a\_usage\_principal\_bureau\_administration\_enseignement

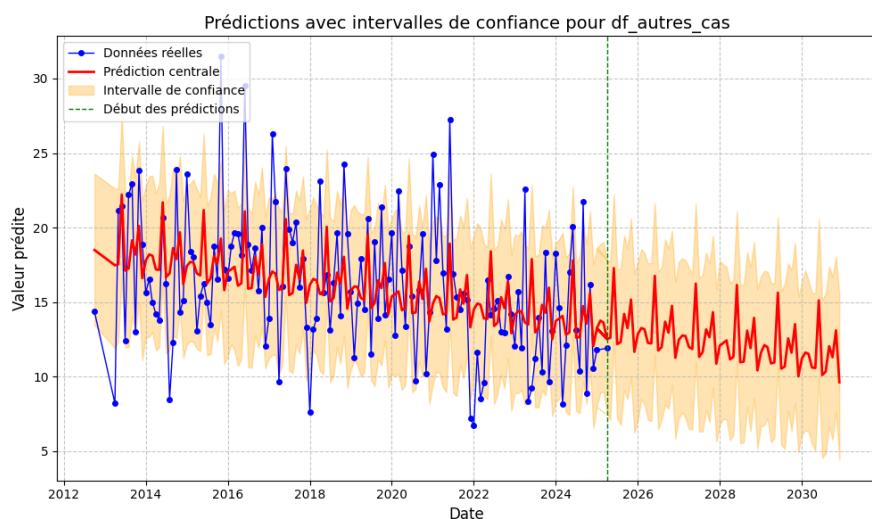


FIGURE 4.10 – Prédictions pour autres\_cas

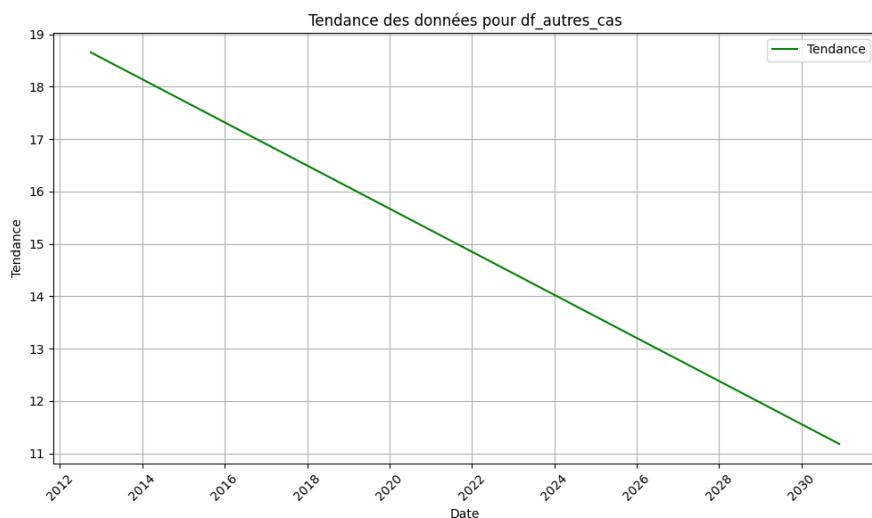


FIGURE 4.11 – Tendance pour autres\_cas

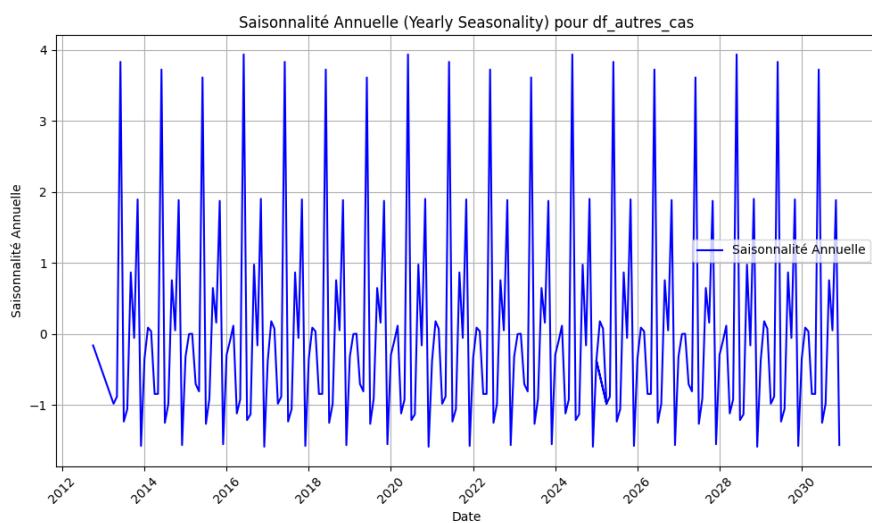


FIGURE 4.12 – Saisonalité pour autres\_cas

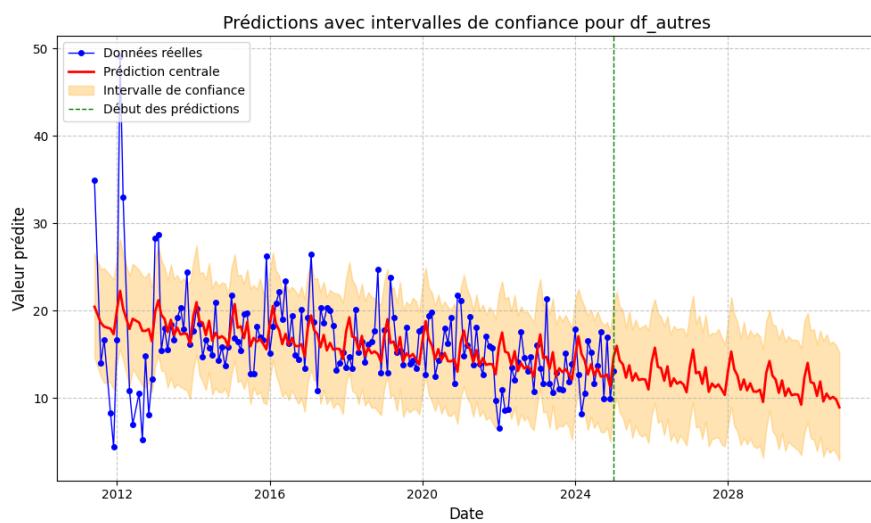


FIGURE 4.13 – Prédictions pour autres

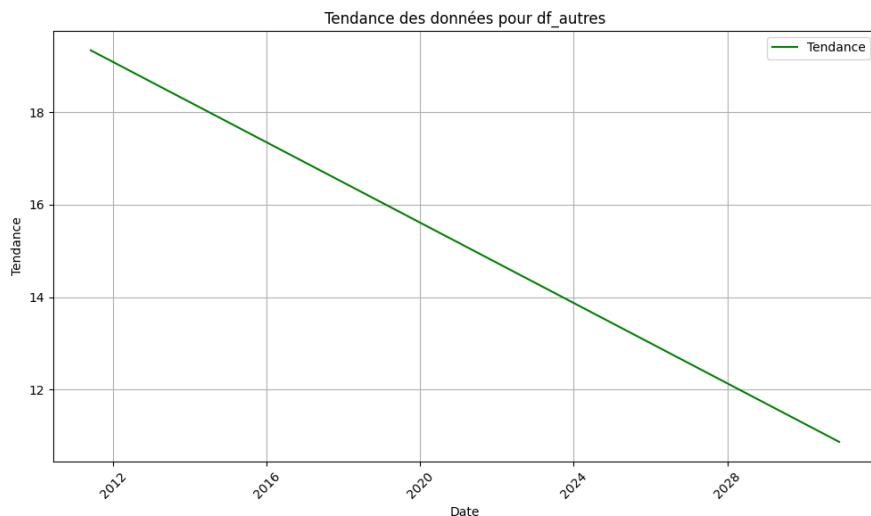


FIGURE 4.14 – Tendance pour autres

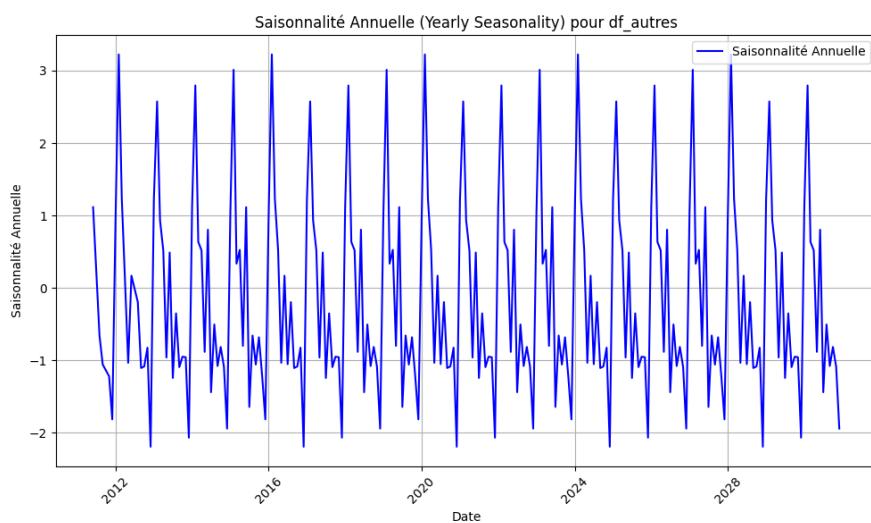


FIGURE 4.15 – Saisonnalité pour autres

## 4.4 Comparaison des modèles

Après avoir lancé les modèles ARIMA et Prophet sur nos données, nous avons calculés les pertes  $R^2$  des deux modèles sur les bases de données afin de comparer leurs efficacités. Voici nos résultats :

$R^2$ en fonction du type de bâtiment	ARIMA	Prophet
Centre commercial	0.014	0.271
Batiment à usage principal	0.023	0.284
Batiment à occupation continue	0.108	0.322
Autres cas	0.035	0.274
Autre	0.108	0.322

Le modèle Prophet est bien meilleur qu'ARIMA, c'est donc lui que nous préconisons d'utiliser.

## Chapitre 5

# Optimisation du portefeuille immobilier selon Markowitz

### 5.1 Introduction du problème

Dans cette deuxième partie du projet, l'objectif est d'optimiser la répartition d'une enveloppe budgétaire de 2 milliards d'euros sur un portefeuille d'actifs immobiliers tertiaires afin de minimiser l'intensité d'émission de CO<sub>2</sub> tout en maximisant les gains. Cette problématique s'inscrit dans un contexte de transition énergétique et de durabilité où la prise en compte des émissions de gaz à effet de serre devient un enjeu majeur pour les institutions financières et les investisseurs.

La base de données fournie contient un échantillon de la base tertiaire après 2021, avec des informations essentielles telles que :

**N°DPE** : clé de jointure avec la base tertiaire après 2021.

**secteur\_activite** : secteur d'activité de l'actif.

**asset\_type\_cre** : catégorie de l'actif.

**Emission\_GES\_kgCO2\_m2\_an** : intensité d'émission exprimée en KgCO<sub>2</sub>/m<sup>2</sup>/an.

**Etiquette\_GES** : étiquette associée à l'intensité d'émission.

**Taux crédit** : taux du crédit utilisé pour financer l'actif.

L'optimisation de la sélection des actifs repose sur l'utilisation de l'algorithme de Markowitz, un cadre théorique permettant de construire un portefeuille optimal en minimisant le risque pour un niveau donné de rendement. Dans ce cas précis, l'objectif est d'adapter ce modèle en prenant en compte non seulement la rentabilité financière des actifs, mais également leur impact environnemental en termes d'émissions de CO<sub>2</sub>.

Ainsi, cette étude vise à proposer une allocation optimale des ressources financières permettant d'atteindre un équilibre entre performance économique et empreinte carbone, en s'appuyant sur des méthodologies quantitatives robustes.

### 5.2 Cadre théorique : modèle de Markowitz

Le modèle de Markowitz, aussi appelé théorie moderne du portefeuille, repose sur l'idée d'optimisation du rendement et du risque d'un portefeuille d'actifs. Ce modèle considère qu'un investisseur rationnel cherche à maximiser le rendement espéré tout en minimisant le risque mesuré par la variance du portefeuille.

Soit un portefeuille composé de  $n$  actifs financiers, avec  $w_i$  représentant la proportion de l'actif  $i$  dans le portefeuille. Le rendement espéré du portefeuille est donné par :

$$E(R_p) = \sum_{i=1}^n w_i E(R_i) \tag{5.1}$$

où  $E(R_i)$  est le rendement espéré de l'actif  $i$ .

Le risque du portefeuille est quantifié par sa variance, définie comme :

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij} \quad (5.2)$$

où  $\sigma_{ij}$  est la covariance entre les actifs  $i$  et  $j$ .

L'optimisation repose sur la minimisation du risque sous contrainte de rendement cible  $R$  :

$$\min_w \quad w^T \Sigma w \quad \text{sous contrainte} \quad \sum_{i=1}^n w_i E(R_i) = R \quad (5.3)$$

Une contrainte supplémentaire impose que la somme des poids soit égale à 1 :

$$\sum_{i=1}^n w_i = 1 \quad (5.4)$$

Ce modèle sera ici adapté pour intégrer les émissions de CO<sub>2</sub> afin de proposer un portefeuille d'investissement respectueux de l'environnement.

Dans le cadre de notre projet, l'optimisation porte sur la recherche d'un équilibre entre performance financière et impact environnemental. Plus précisément, il s'agit de maximiser les gains générés par les actifs immobiliers tout en minimisant leurs émissions de CO<sub>2</sub>. Dans le modèle de Markowitz, le risque est traditionnellement mesuré par la variance des rendements, mais ici, nous assimilons les émissions de CO<sub>2</sub> à une source de risque à réduire. Ainsi, la problématique se formule comme une double optimisation où l'objectif est de sélectionner un portefeuille d'actifs qui maximise la rentabilité attendue tout en minimisant l'empreinte carbone globale. Cette approche permet d'intégrer des critères de durabilité dans la prise de décision et d'aligner la stratégie d'investissement avec les enjeux environnementaux actuels.

### 5.3 Description du Jeu de Données

Le jeu de données fourni regroupe des informations sur un ensemble de biens immobiliers tertiaires après 2021. Il constitue la base sur laquelle sera appliqué l'algorithme de Markowitz afin d'optimiser la répartition d'une enveloppe budgétaire en fonction des émissions de CO<sub>2</sub> et des gains financiers.

Les principales variables contenues dans le dataset sont les suivantes :

- **N°DPE** : Identifiant unique permettant de lier chaque bien immobilier à la base tertiaire.
- **secteur\_activite** : Secteur d'activité auquel appartient l'actif (bureaux d'entreprise, établissements de soins, administrations, etc.).
- **asset\_type\_cre** : Type spécifique de bâtiment, détaillant son usage principal (bureaux, établissements de soins, etc.).
- **Emission\_GES\_kgCO2\_m2\_an** : Intensité des émissions de gaz à effet de serre exprimée en kgCO<sub>2</sub>/m<sup>2</sup>/an. Cette variable est essentielle pour l'optimisation car elle représente le critère environnemental à minimiser.
- **Etiquette\_GES** : Classification qualitative de l'intensité des émissions sous forme d'étiquette (A, B, C, etc.).
- **taux\_credit** : Taux d'intérêt appliqué au crédit utilisé pour financer l'actif, influençant directement la rentabilité des investissements.

L'analyse de ces données permettra d'identifier les caractéristiques des différents actifs et d'optimiser leur sélection afin de minimiser les émissions de CO<sub>2</sub> tout en maximisant les gains financiers. L'intégration de ces variables dans l'algorithme de Markowitz constitue la base de notre approche quantitative d'optimisation du portefeuille.

### 5.4 Explication des Choix de Paramètres dans la Simulation de la Matrice de Covariance

Dans ce modèle, nous n'avons pas accès à une série temporelle réelle des rendements (ici représentés par le **taux\_credit** utilisé comme proxy pour le **Rendement\_Composite**). Pour pallier cette absence de données historiques, nous simulons un historique artificiel en générant **T = 1000** observations pour chaque actif. Voici les principales hypothèses et justifications :

## 1. Interprétation de $T = 1000$

- $T$  représente le nombre d'observations simulées, que nous interprétons ici comme des **données journalières** sur environ 1000 jours.
- Cette durée permet d'obtenir un échantillon suffisamment grand pour estimer de manière stable la **matrice de covariance** entre les rendements simulés des différents actifs.
- L'idée est de modéliser la variabilité potentielle du rendement (ici, le `taux_credit`) sur une période d'environ 1000 jours, ce qui correspond approximativement à 3 ans de données journalières, bien que ce ne soit qu'une approximation.

## 2. Choix du Scale = 0.05 pour la Simulation

- Pour chaque actif, nous générerons des observations suivant une distribution normale avec :
  - Une **moyenne** égale à la valeur observée du `taux_credit` (notre proxy de rendement composite).
  - Un **écart-type** fixé à **0.05**.
- Le paramètre **0.05** est choisi en fonction de l'échelle des valeurs observées dans notre DataFrame. Ce choix suppose que la volatilité journalière des taux de crédit est relativement faible, ce qui est cohérent si les taux varient peu d'un jour à l'autre.
- Cette hypothèse permet de simuler des fluctuations journalières autour de la moyenne, fournissant ainsi un échantillon de données permettant de calculer une matrice de covariance robuste pour l'optimisation du portefeuille.

## 3. Utilisation de la Simulation

- La simulation génère un ensemble de rendements "artificiels" pour chaque actif, organisés dans une matrice de dimensions  $(T, n)$  où  $n$  est le nombre d'actifs.
- En appliquant la fonction `np.cov(..., rowvar=False)`, nous calculons la **matrice de covariance** qui quantifie :
  - La **variance** de chaque actif (éléments diagonaux).
  - La **covariance** entre les actifs (éléments hors diagonale).
- Cette matrice de covariance est essentielle dans le modèle de Markowitz, car elle permet d'évaluer le **risque** du portefeuille et d'optimiser la répartition des actifs en minimisant la variance globale tout en respectant les contraintes (rendement minimal et émission maximale).

### En résumé

- **T = 1000** : Nous simulons 1000 jours de données journalières pour obtenir un échantillon suffisamment grand pour estimer la covariance.
- **Scale = 0.05** : Cet écart-type est choisi pour représenter la volatilité journalière des taux de crédit (notre proxy de rendement composite), en se basant sur l'hypothèse que les taux varient peu d'un jour à l'autre.
- **But de la simulation** : Générer une matrice de covariance robuste, indispensable pour évaluer le risque du portefeuille dans le cadre du modèle de Markowitz.

Ces hypothèses, bien que simplificatrices, fournissent une base pour modéliser la variabilité des rendements et effectuer l'optimisation du portefeuille en l'absence de séries temporelles historiques réelles.

## 5.5 Formulation mathématique du problème

L'objectif principal de ce projet est d'optimiser l'allocation d'un portefeuille d'actifs immobiliers en minimisant l'intensité des émissions de CO<sub>2</sub> tout en maximisant les gains financiers, sous contrainte budgétaire. Ce problème peut être formulé mathématiquement en s'inspirant du modèle de Markowitz d'optimisation de portefeuille.

Soit un ensemble de  $n$  actifs immobiliers indexés par  $i \in \{1, \dots, n\}$ . Pour chaque actif, nous définissons les variables suivantes :

- $x_i$  : proportion du budget total alloué à l'actif  $i$ , avec  $\sum_{i=1}^n x_i = 1$ .
- $r_i$  : rendement attendu de l'actif  $i$ , basé sur le taux de crédit.
- $\sigma_{ij}$  : covariance entre les rendements des actifs  $i$  et  $j$ , représentant la variabilité des gains.
- $e_i$  : intensité des émissions de CO<sub>2</sub> de l'actif  $i$  (exprimée en kgCO<sub>2</sub>/m<sup>2</sup>/an).
- $E_p = \sum_{i=1}^n x_i e_i$  : intensité d'émission du portefeuille, que nous cherchons à minimiser.
- $B$  : budget total disponible (2 milliards d'euros).

L'optimisation repose alors sur la double minimisation du risque financier et de l'impact environnemental, tout en maximisant le rendement. Cela se traduit par le problème d'optimisation suivant :

$$\begin{aligned} \max \quad & R_p = \sum_{i=1}^n x_i r_i \\ \min \quad & \sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n x_i x_j \sigma_{ij} \\ \min \quad & E_p = \sum_{i=1}^n x_i e_i \end{aligned}$$

Sous les contraintes suivantes :

$$\begin{aligned} \sum_{i=1}^n x_i &= 1, \quad x_i \geq 0, \quad \forall i \\ \sum_{i=1}^n x_i C_i &\leq B \end{aligned}$$

où  $C_i$  représente le coût d'acquisition de chaque actif  $i$ .

L'objectif est donc de trouver la meilleure allocation  $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$  permettant d'atteindre un compromis optimal entre rentabilité et impact environnemental, en tenant compte des contraintes budgétaires.

## 5.6 Implémentation et résultats

L'optimisation du portefeuille d'actifs immobiliers a été réalisée à l'aide de Python en utilisant CVXPY, une bibliothèque spécialisée dans la résolution de problèmes d'optimisation convexes. L'implémentation repose sur la formulation mathématique établie précédemment, qui cherche à maximiser les gains tout en minimisant les émissions de CO sous des contraintes budgétaires et de diversification des actifs.

Les données ont d'abord été prétraitées afin de garantir la cohérence du modèle. Les observations contenant des valeurs manquantes sur des variables essentielles telles que le taux de crédit, l'intensité des émissions de CO, le secteur d'activité et le type d'actif ont été supprimées. Une colonne représentant la rentabilité des actifs a été définie en reprenant le taux de crédit comme indicateur principal de rendement financier. Ensuite, une matrice de covariance des rendements a été estimée en générant des variations autour des valeurs moyennes des taux de crédit des actifs. Cette matrice est essentielle pour l'optimisation, car elle permet d'évaluer la volatilité des rendements et de minimiser le risque global du portefeuille.

L'optimisation a été définie en respectant plusieurs contraintes. Tout d'abord, la somme des pondérations allouées aux actifs devait être égale à 1, garantissant une allocation complète du budget de 2 milliards d'euros. Ensuite, la contrainte de non short-selling a été appliquée, interdisant les pondérations négatives afin d'éviter des positions spéculatives sur certains actifs. Une contrainte supplémentaire a été imposée sur le rendement global du portefeuille, qui devait être supérieur à la médiane des rendements des actifs, garantissant ainsi une performance minimale acceptable. Enfin, une contrainte limitant l'intensité des émissions de CO a été introduite, imposant un plafond basé sur le 50e percentile des émissions observées dans le dataset.

La résolution du problème a abouti à une allocation optimale du budget sur les différents actifs, minimisant le risque tout en maintenant un rendement supérieur au seuil fixé. Les résultats ont montré une rentabilité moyenne du portefeuille proche de la médiane des taux de crédit des actifs, tout en réduisant significativement la variance des rendements grâce à la diversification. L'intensité des émissions de CO du portefeuille a également été contrôlée en maintenant la valeur sous le seuil fixé, illustrant la possibilité d'une stratégie d'investissement plus durable.

L'analyse a été complétée par la construction de la frontière efficiente, représentant l'ensemble des allocations possibles offrant le meilleur compromis entre risque et rendement, tout en intégrant la contrainte environnementale. Les résultats montrent que les actifs les plus rentables sont souvent aussi ceux avec les émissions de CO les plus élevées, soulignant un arbitrage inévitable entre performance financière et impact environnemental. La frontière efficiente obtenue permet d'identifier des solutions d'investissement adaptées en fonction du niveau de risque accepté et des objectifs de réduction des émissions.

L'optimisation réalisée démontre qu'il est possible d'intégrer une contrainte environnementale stricte dans un modèle de sélection de portefeuille sans pour autant sacrifier complètement la performance financière. L'approche

développée permet ainsi aux investisseurs d'adopter une stratégie plus durable tout en assurant une rentabilité satisfaisante.

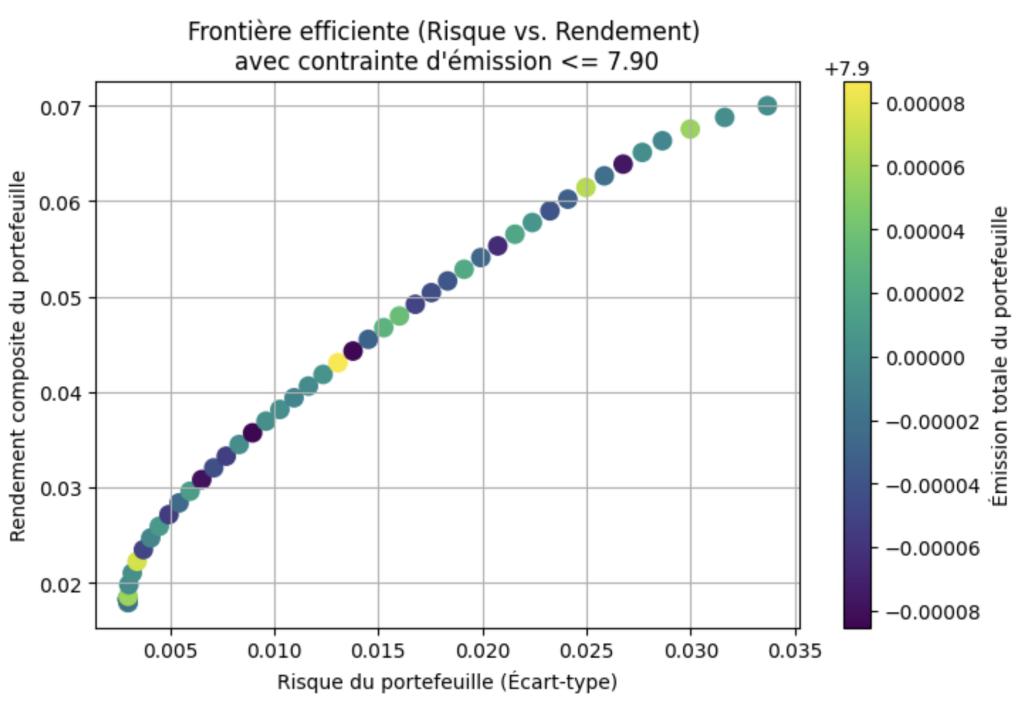


FIGURE 5.1 – Frontière efficiente

Ce graphique représente la distribution du rendement composite des actifs du portefeuille. On observe une forte concentration des valeurs autour de 0,01 et 0,02, ce qui indique que la majorité des actifs ont un rendement relativement faible. Quelques actifs affichent des rendements plus élevés, mais ils sont moins fréquents. Cette distribution asymétrique suggère une hétérogénéité dans les opportunités d'investissement, avec une majorité d'actifs offrant des rendements modestes et une minorité présentant des rendements plus élevés. Cela souligne l'importance d'une sélection rigoureuse des actifs pour optimiser le portefeuille tout en maintenant un niveau de risque maîtrisé.

Ce graphique illustre la distribution des émissions de gaz à effet de serre (GES) des actifs du portefeuille. On remarque une forte asymétrie à droite, avec une majorité d'actifs présentant de faibles niveaux d'émissions, concentrés en dessous de 10 unités. Cependant, quelques actifs affichent des émissions nettement plus élevées, atteignant même des valeurs supérieures à 100 voire 250. Cette distribution indique que la plupart des actifs sont relativement peu polluants, mais qu'une minorité contribue de manière disproportionnée aux émissions totales. Ce constat est crucial pour l'optimisation du portefeuille, car il suggère qu'il est possible de réduire significativement l'empreinte carbone globale en réallouant le capital vers des actifs à faibles émissions tout en maintenant un rendement compétitif.

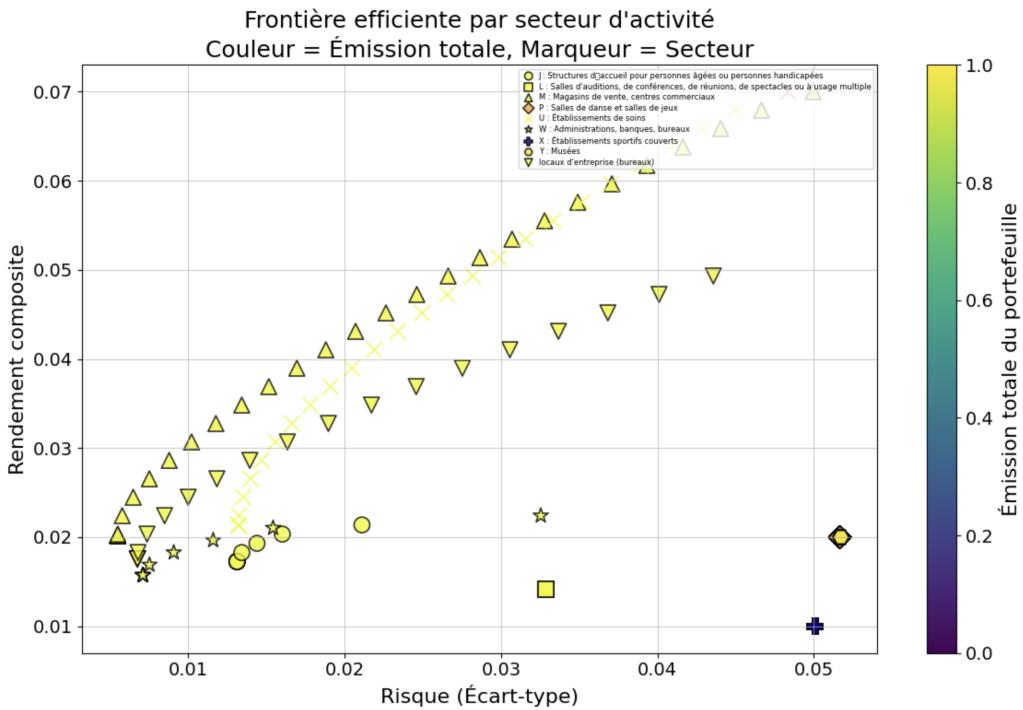


FIGURE 5.2 – Frontière efficiente par secteur d'activité

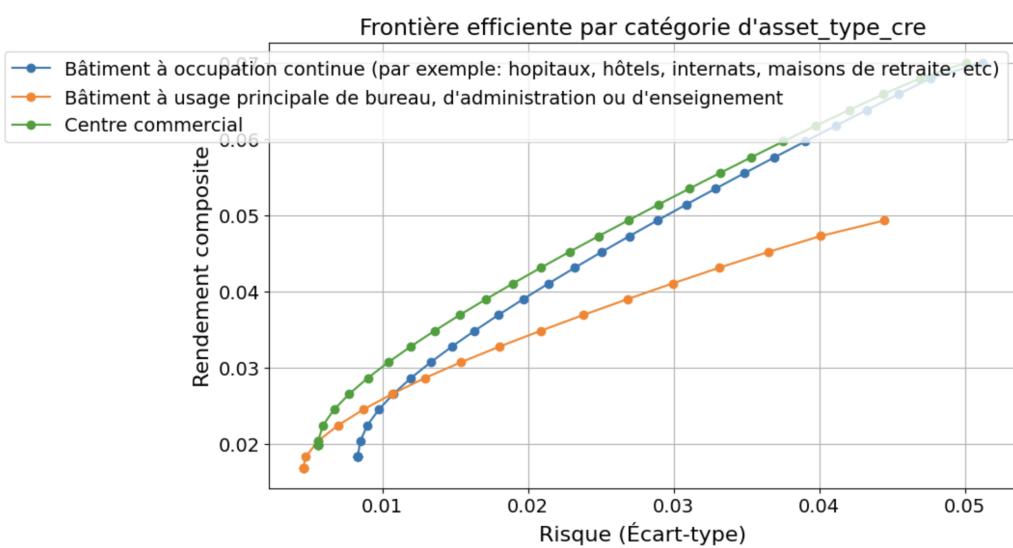


FIGURE 5.3 – Frontière efficiente par catégorie d'asset type cre

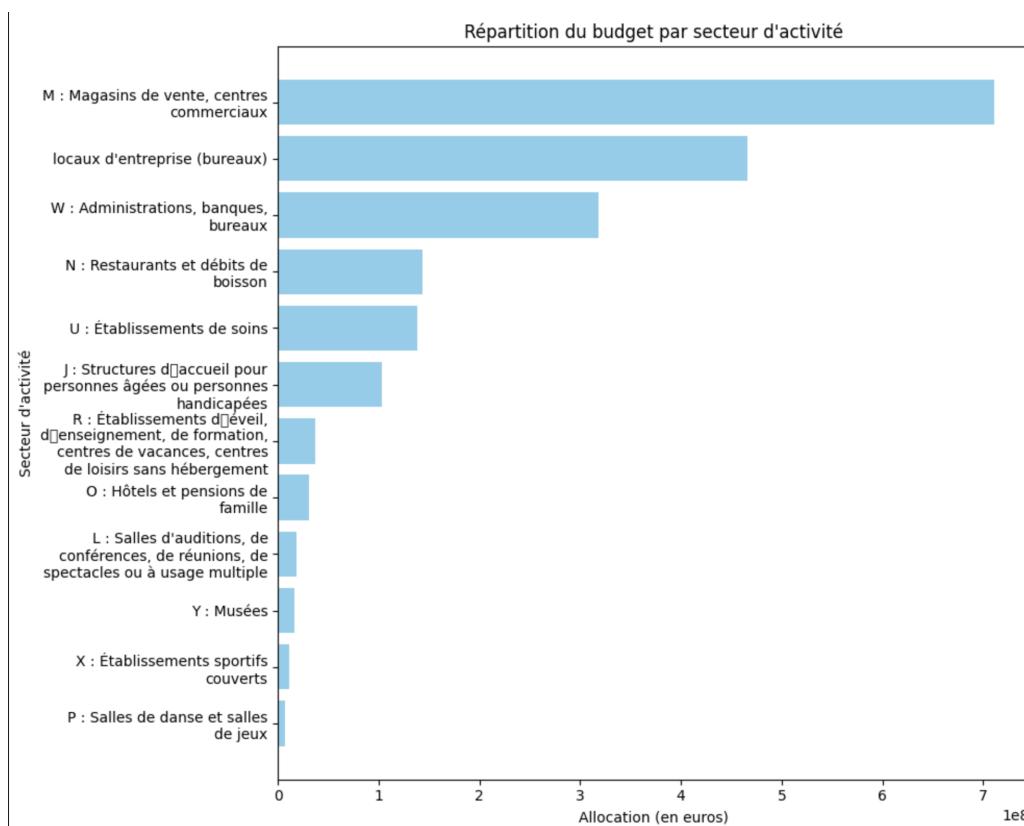


FIGURE 5.4 – Répartition du budget

# Conclusion

## Pour clore le projet

Ce projet a permis de développer une approche quantitative d'optimisation d'un portefeuille d'actifs immobiliers en intégrant à la fois des objectifs financiers et environnementaux. L'utilisation du modèle de Markowitz a permis d'élaborer une stratégie d'allocation du capital visant à maximiser les rendements tout en minimisant l'empreinte carbone du portefeuille. La formulation mathématique du problème a été adaptée afin d'inclure les émissions de CO comme une contrainte supplémentaire dans l'optimisation, assurant ainsi une approche plus durable des investissements.

L'analyse des modèles de séries temporelles, notamment XARIMA et Prophet, a offert une vision détaillée de l'évolution des émissions de CO des actifs immobiliers. Le modèle XARIMA s'est révélé performant pour capturer les tendances et les variations saisonnières, permettant ainsi d'anticiper les évolutions futures des émissions avec une précision satisfaisante. Prophet a complété cette analyse en apportant une flexibilité supplémentaire dans la modélisation des tendances et des changements structurels.

Les résultats obtenus ont mis en évidence la possibilité de réduire significativement les émissions du portefeuille sans compromettre la rentabilité. La frontière efficiente obtenue à partir des simulations d'optimisation a montré que des arbitrages existent entre performance financière et impact environnemental, nécessitant une prise de décision éclairée de la part des investisseurs.

En perspective, une amélioration du modèle pourrait être envisagée en intégrant des variables explicatives supplémentaires, telles que des facteurs macroéconomiques ou des réglementations environnementales, afin d'affiner les prévisions et d'améliorer la robustesse de l'optimisation. De plus, l'exploration d'autres approches d'optimisation multi-objectifs, comme les méthodes basées sur l'optimisation robuste ou les algorithmes évolutionnaires, pourrait apporter une meilleure gestion des incertitudes dans la prise de décision.

Ce projet ouvre ainsi la voie à une gestion plus durable et optimisée des portefeuilles immobiliers, conciliant impératifs financiers et enjeux environnementaux pour une stratégie d'investissement responsable.

# Remerciements

Nous tenons à exprimer notre profonde gratitude à toute l'équipe de BNP Paribas pour leur soutien et leur accompagnement précieux tout au long de ce projet. Leur expertise, disponibilité et bienveillance ont grandement contribué à la réussite de ce travail.

Un remerciement particulier va à Monsieur Anas Majji et Monsieur Yann Lefevre pour leur implication constante, leurs conseils avisés et leur accompagnement de chaque instant. Leur soutien a été essentiel pour mener à bien ce projet et nous leur en sommes profondément reconnaissants.

Nous leur adressons nos plus sincères remerciements, au nom de toute l'équipe.