Université de Haute Alsace
Faculté des Sciences et Techniques
Département de Mathémathiques

Semestre 8
UE: projet 01
M1 -IMDS

Analyse et Visualisation des Tendances Énergétiques de l'Union Européenne

EL BOUGHDADI Abdeljalil et HAMDINOU Moulaye Driss

Sous l'encadrement de MADAME Suzy MADDAH

10 mai 2025

Remerciement

Nous tenons à remercier nos parents pour leur amour, leurs sacrifices ainsi que pour leur soutient tout au long de nos études.

Nous exprimons nos plus vive gratitude à notre encadrante de ce projet, Madame **Suzy MADDAH**, pour son encadrement durant toute la durée de ce travail. elle a dirigée notre projet avec beaucoup de patience.

Nous remercions également les membres du jury : Monsieur **Abdenacer MAKHLOUF**, Monsieur **Cornel MUREA**, Monsieur **Nicolas JUILLET** et Monsieur **Zakaria BELHACHMI** et nous espérons pouvoir réussir la tâche qui nous a été accordée.

Nos remerciements vont aussi à l'ensemble des professeurs qui ont assuré avec succès l'encadrement et l'enseignement de la filière Mathématiques, et à tout ceux qui ont participé de loin ou proche à la réalisation de ce projet.

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION ET CONTEXTE

L'énergie constitue aujourd'hui l'un des piliers fondamentaux de nos sociétés modernes. Elle conditionne à la fois la croissance économique, le confort des populations et les dynamiques géopolitiques mondiales. L'accès à une énergie fiable, abordable et durable est au cœur des enjeux du XXI^{e} siècle, notamment dans un contexte de transition énergétique visant à réduire les émissions de gaz à effet de serre et à lutter contre le changement climatique. La crise énergétique récente, les tensions sur les marchés internationaux et la volonté des pays de sécuriser leur approvisionnement ont mis en lumière l'extrême sensibilité des économies contemporaines aux fluctuations du secteur énergétique.

Dans ce cadre, la capacité à collecter, structurer, analyser et visualiser des données énergétiques devient un enjeu stratégique. Comprendre les tendances de consommation, évaluer les performances des politiques énergétiques, identifier les anomalies ou anticiper les évolutions futures sont autant de défis qui nécessitent une approche rigoureuse, fondée sur des données fiables et des outils adaptés.

1.1 Présentation des données

Pour mener à bien notre projet, nous avons utilisé un jeu de données mis à disposition par la plateforme *Our World in Data*, accessible publiquement à l'adresse suivante :

`https://github.com/owid/energy-data`

Le fichier source est un vaste jeu de données brutes, non nettoyées, couvrant une période allant de 1900 à 2023 et contenant plusieurs centaines de colonnes pour plus de 200 pays. Il contient des informations variées telles que la population, le produit intérieur brut (PIB), ainsi que des données détaillées sur la production et la consommation d'énergie selon différentes sources : charbon, gaz, pétrole, nucléaire, biocarburants, énergies renouvelables (solaire, éolien, hydraulique, etc.). Ces données, bien que riches en potentiel, sont à l'origine ni filtrées, ni nettoyées, ni structurées pour un usage immédiat. Elles présentent de nombreuses valeurs manquantes, des redondances, et des formats hétérogènes. Aucune base relationnelle n'était proposée, et aucun tableau de bord prêt à l'emploi n'était disponible.

Dès lors, l'intégralité du processus d'analyse présenté dans ce rapport a été conçu, réalisé et automatisé par nos soins, depuis l'extraction initiale des fichiers CSV jusqu'à la modélisation avancée des données. Nous avons entrepris un véritable travail d'ingénierie de données, en :

- filtrant les données pour ne conserver que les pays membres de l'Union Européenne ;
- évaluant la complétude des séries temporelles et supprimant les lignes ou colonnes trop lacunaires ;
- appliquant des méthodes rigoureuses de traitement des valeurs manquantes (interpolation, extrapolation) ;
- construisant une base de données relationnelle PostgreSQL, avec un schéma normalisé, des clés primaires composées et des relations entre les différentes sources d'énergie ;
- automatisant l'insertion des données et la création des tables via des scripts Python dynamiques.

1.2 Objectifs et structure du projet

Le projet s'articule autour de plusieurs grandes étapes complémentaires, mobilisant des outils modernes de traitement de données, de visualisation et d'analyse prédictive :

— **Étape 1 : Préparation et nettoyage des données**
Les données brutes ont été traitées afin d'en assurer la qualité (gestion des valeurs manquantes, harmonisation des formats, etc.), puis insérées dans une base de données relationnelle PostgreSQL. Cette structuration permet des requêtes efficaces et reproductibles.

— **Étape 2 : Visualisation et storytelling avec Grafana**
Nous avons conçu un tableau de bord interactif à l'aide de Grafana, outil open source de visualisation, afin de représenter graphiquement les principales tendances énergétiques par pays, par source d'énergie et par année. Ce volet vise à produire une analyse visuelle accessible, dynamique et pertinente.

— **Étape 3 : Analyse avancée avec Python**
Dans la dernière partie du projet, nous avons mobilisé Python et plusieurs bibliothèques spécialisées telles que `pandas`, `numpy`, `matplotlib`, `scikit-learn` et `statsmodels` pour mener deux types d'analyses :

- **Détection d'anomalies** à l'aide du Z-score et de la méthode de Tukey, permettant d'identifier des années ou des pays présentant des comportements énergétiques atypiques.
- **Modélisation prédictive** à l'aide de deux modèles : la régression linéaire (modèle simple pour évaluer les tendances) et le modèle ARIMA (modèle temporel plus sophistiqué permettant de faire des prévisions basées sur les valeurs passées).

Le **chapitre 4** du rapport contient des études théoriques des modèles statistiques utilisés, notamment les fondements des modèles de régression linéaire et des modèles ARIMA, ainsi que les hypothèses nécessaires à leur bon fonctionnement.

Le **chapitre 5** regroupe l'ensemble des éléments techniques du projet, incluant tous les scripts Python développés, ainsi que les requêtes SQL utilisées pour créer la base, insérer les données et interroger les indicateurs clés. Cette section garantit la transparence, la reproductibilité et la traçabilité complète de notre démarche.

Ce projet illustre une démarche complète, allant de la collecte des données brutes à leur analyse exploratoire et prédictive. Il s'appuie sur une combinaison cohérente d'outils de gestion de bases de données, de techniques de visualisation et de méthodes statistiques rigoureuses. Contrairement aux jeux de données prétraités disponibles sur des plateformes telles que Kaggle ou l'UCI Machine Learning Repository, les données utilisées ici n'ont fait l'objet d'aucune exploration préalable. Ce travail s'inscrit donc pleinement dans une approche exploratoire authentique, depuis l'ingestion des données jusqu'à la modélisation

CHAPITRE 2

CONCEPTION ET INSERTION DES DONNÉES DANS POSTGRESQL

2.1 Nettoyage des données

Le **nettoyage des données** consiste à traiter ces données bruitées, soit en les supprimant, soit en les modifiant de manière à tirer le meilleur profit. L'**intégration** est la combinaison des données provenant de plusieurs sources (bases de données, sources externes, etc.). Le but de ces deux opérations est de générer des *entrepôts de données* et/ou des *magasins de données spécialisés* contenant les données traitées afin de faciliter leurs exploitations futures.

2.1.1 Suppression des données non pertinentes pour l'étude

Dans le cadre de notre projet, nous disposons d'un fichier Excel contenant l'ensemble des données énergétiques pour plusieurs pays à travers le monde, sur la période allant de 1900 à 2023. Toutefois, notre étude se concentre exclusivement sur l'analyse des tendances énergétiques au sein de l'Union Européenne. Par conséquent, il a été nécessaire de supprimer toutes les données relatives aux pays qui ne font pas partie de l'Union. Cette opération de filtrage a été réalisée automatiquement à l'aide d'un script Python, que vous retrouverez en annexe dans le dernier chapitre de ce rapport. Ce traitement nous a permis de ne conserver que les données pertinentes à notre analyse, d'alléger considérablement le volume de données à manipuler, et de garantir la cohérence du périmètre géographique étudié.

2.1.2 La prise en compte des données manquantes

La donnée manquante est une source de problèmes majeurs dans les méthodes d'analyse de données. Même si nos méthodes d'analyse gagnent en sophistication, nous rencontrons des valeurs manquantes dans les champs, particulièrement dans les bases comportant un grand nombre de champs. L'absence d'information est rarement bénéfique.Toute chose étant égale par ailleurs, plus de détails est toujours préférable. C'est pourquoi nous devons réflichir avec attention à la façon dont nous allons gérer ce problème de données manquantes.
Une méthode courante pour prendre en compte les valeurs manquantes consiste à omettre de l'analyse les enregistrements ou les variables avec des valeurs manquantes. Cependant, cela peut être dangereux puisque la répartition des valeurs manquantes peut être systématique, et éliminer les enregistrements avec des valeurs manquantes conduirait à un sous-ensemble de données biaisées. De plus, cela conduirait à omettre les informations présentes dans les autres champs, simplement parce que la valeur d'un champ particulier serait manquante. En fait **Galit Shmueli** établissent que si seulement 5% des valeurs des données sont manquantes au sein d'un jeu de données de trente variables et que ces valeurs manquantes sont réparties au sein des données, presque 80% des enregistrements auront au moins une valeur manquante. C'est pourquoi les analystes de données ont mis au point des méthodes visant à remplacer les valeurs manquantes par une valeur substituée selon différent critères :

— **Remplacer par une constante :** l'analyste choisit une valeur arbitraire (par exemple, 0, -1, ou "inconnu") pour remplir les champs vides. Cette méthode est simple, mais elle peut fausser les distributions et introduire des biais si la constante choisie est interprétée comme une vraie valeur.

— **Remplacer par la moyenne :** cette méthode est fréquemment utilisée pour les variables numériques. Mais la moyenne peut ne pas être toujours le meilleur choix, **Daniel Larose** montre un jeu de données où la moyenne est plus élevée que le 81^e percentile. De plus si beaucoup de valeurs manquantes sont remplacés par la moyenne, le niveau de confiance en résultant sera trop optimiste, puisque la mesure de la répartition va se réduire artifficiellement.

— **Remplacer les valeurs manquantes par des valeurs générées aléatoirement à partir de la distribution observée de la variable :** il est possible de tirer des valeurs aléatoires dans la distribution observée de la variable. Un avantage de cette méthode est que les mesures de centralité et de dispersion devraient rester plus proches des données originales, comparaticement avec la méthode de remplacement par la moyenne. Cependant, il n'y a pas de garantie que les valeurs trouvées aient du sens.

— **Imputation basée sur des caractéristiques similaires :** c'est une méthode plus sophistiquée, qui consiste à rechercher dans les données des individus proches (sur d'autres variables) et à utiliser leurs valeurs comme substituts. Par exemple, on peut estimer la consommation d'énergie manquante pour un pays en se basant sur les valeurs observées pour d'autres pays de profil économique ou démographique similaire.

Dans notre projet, nous avons écarté les approches fondées sur la moyenne ou la constante, car elles ont tendance à simplifier artificiellement les données et à ignorer leur dynamique temporelle. Nous avons opté pour une méthode d'interpolation linéaire, particulièrement adaptée aux séries chronologiques comme la consommation ou la production d'énergie. L'interpolation estime la valeur manquante à partir des observations voisines dans le temps, ce qui permet de conserver la continuité et la cohérence des évolutions annuelles.

2.1.3 Suppression des lignes par pays

Chaque pays ayant des périodes différentes de couverture des données, il est nécessaire d'analyser la complétude ligne par ligne, par pays. Ainsi, pour chaque pays, seules les lignes présentant un remplissage supérieur à 30% ont été conservées.

Définition 2.1.1 (Complétude d'une ligne)
Soit L une ligne correspondant à une observation annuelle pour un pays P. La complétude (en %) de L est définie par :

$$\textit{Complétude}(L) = \frac{\textit{nombre de valeurs non-nulles dans } L}{\textit{nombre total de colonnes numériques}} \times 100$$

Exemple 2.1.1
Dans notre analyse, nous avons évalué la complétude des données ligne par ligne, chaque ligne représentant les données énergétiques d'un pays européen pour une année donnée entre 1900 *et* 2023*. Pour cela, nous avons calculé le pourcentage de complétude de chaque ligne (nombre de valeurs non manquantes par rapport au nombre total de colonnes), et avons visualisé les résultats sous forme de carte thermique (heatmap) à l'aide d'un script Python (voir chapitre ANNEXE).*

La figure ci-dessous illustre la complétude des données par pays et année :

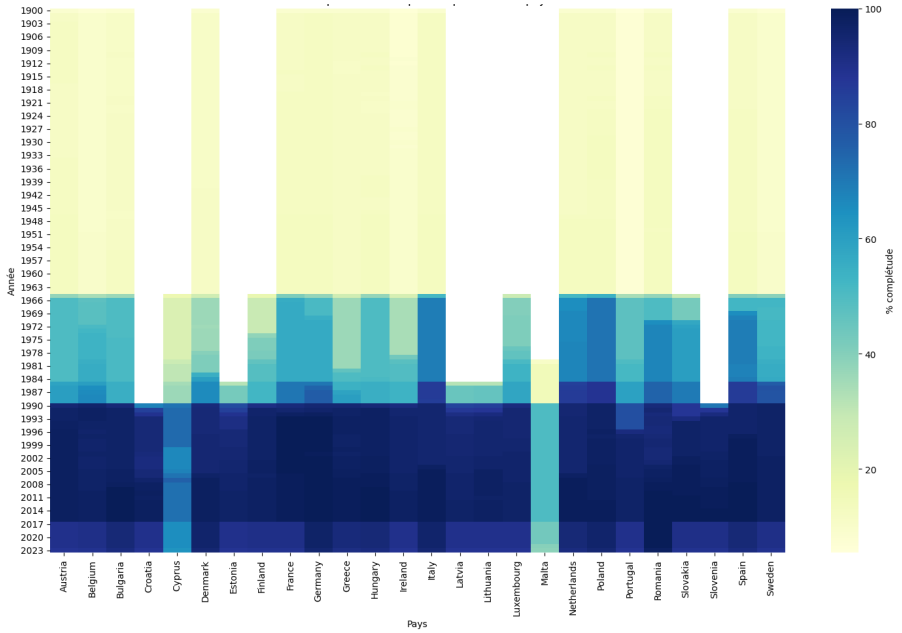

FIGURE 2.1 – Heatmap du pourcentage de complétude par année et par pays avant nettoyage

Remarque 2.1.1
On observe clairement que les données avant 1965 sont très incomplètes, avec des pourcentages de complétude généralement inférieurs à 30%. *Ces lignes contiennent très peu d'informations exploitables. Ainsi, nous avons pris la décision de supprimer toutes les lignes antérieures à* 1966 *et de ne conserver que les données à partir de cette année, qui sont suffisantes et pertinentes pour nos analyses.*

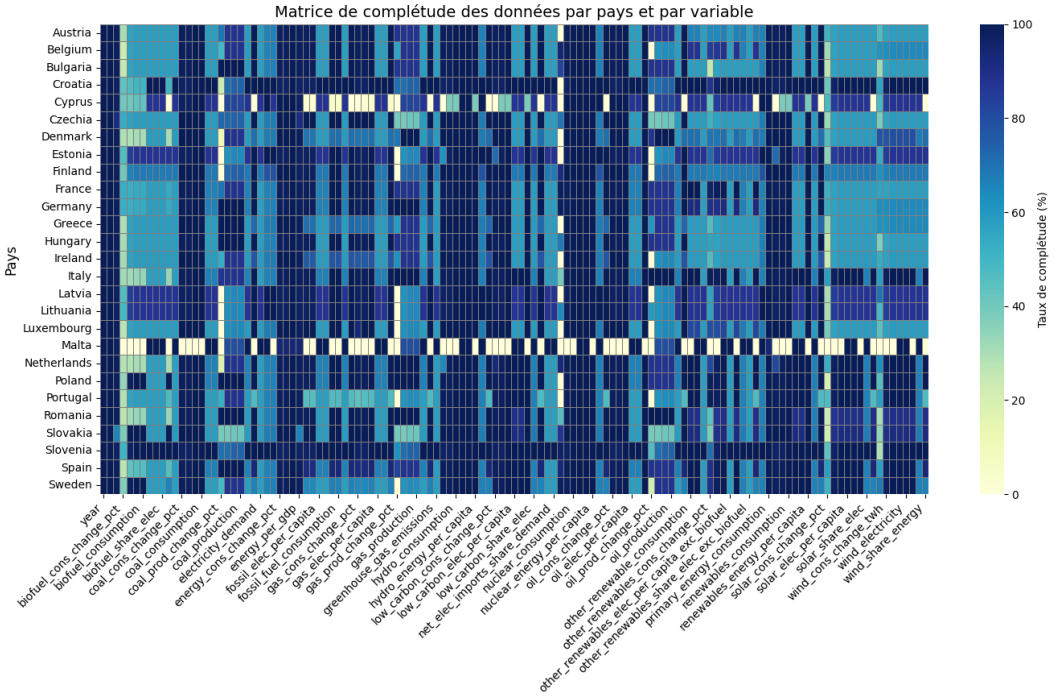

FIGURE 2.2 – Matrice de complétude aprés suppression des colonnes et lignes incomplètes

2.1.4 Interpolation et extrapolation linéaires

Définition 2.1.2
L'interpolation linéaire est une méthode qui consiste à estimer une valeur manquante située entre deux points connus dans une série chronologique, en supposant une relation affine entre eux.
L'extrapolation linéaire, quant à elle, vise à estimer une valeur manquante située en dehors de l'intervalle connu — soit avant le premier point disponible, soit après le dernier. Elle repose également sur l'hypothèse de linéarité, mais est considérée comme plus risquée car elle s'aventure hors des bornes observées.

Proposition 2.1.1
Soient deux points de coordonnées (x_a, y_a) et (x_b, y_b) avec $x_a < x_b$. On peut approximer la valeur de $f(x)$, que ce soit en interpolation ou en extrapolation, par la formule affine suivante :

$$f(x) = y_a + \frac{y_b - y_a}{x_b - x_a} \times (x - x_a)$$

- *Si $x \in [x_a, x_b]$, on parle d'interpolation ;*
- *Si $x < x_a$ ou $x > x_b$, on parle d'extrapolation.*

Remarque 2.1.2
Dans le contexte des séries temporelles, l'interpolation est généralement plus fiable, car elle repose sur des données encadrées. L'extrapolation suppose que la tendance observée se prolonge au-delà de la période mesurée, ce qui peut être discutable, surtout si la dynamique change (crise, transition énergétique, etc.).

Application 2.1.1
Dans notre projet, nous avons utilisé ces deux techniques de manière complémentaire :
- *L'interpolation linéaire a été appliquée pour combler les valeurs manquantes entre deux années connues, en assurant la continuité des séries ;*
- *L'extrapolation linéaire a été utilisée pour remplir les premières ou dernières années, dans les cas où un pays ne disposait pas encore de données au début ou à la fin de la période étudiée.*

Ces méthodes nous ont permis d'obtenir un tableau de données complet, cohérent et exploitable pour les visualisations et analyses à venir.
Vous trouverez le code Python utilisé dans le chapitre ANNEXE.

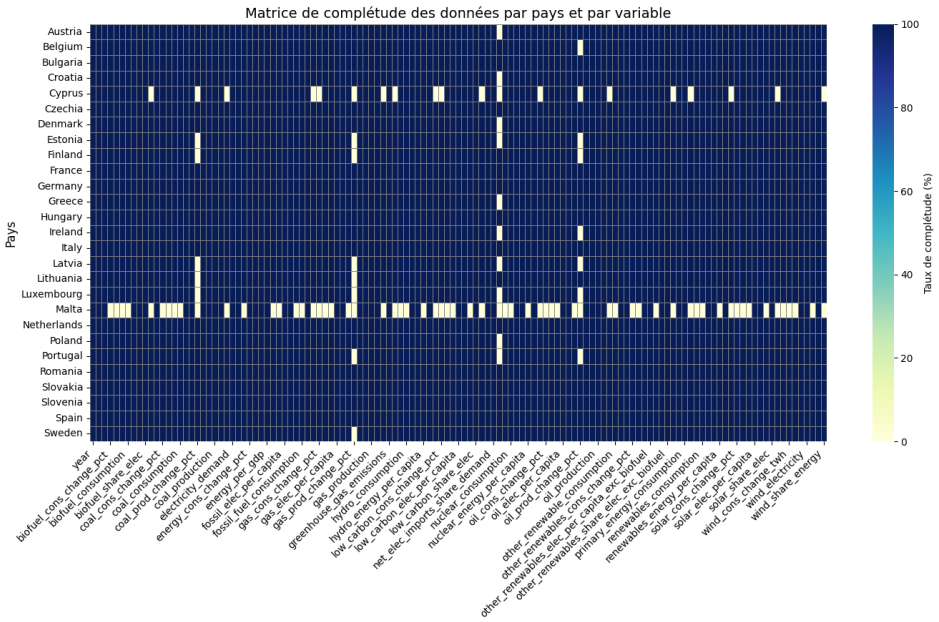

FIGURE 2.3 – Matrice de complétude après interpolation et extrapolation linéaires

Remarque 2.1.3
La figure ci-dessus montre la matrice de complétude des données après traitement par interpolation et extrapolation linéaires. Bien que la majorité des pays de l'Union Européenne présentent désormais des séries temporelles quasi complètes, deux pays se distinguent encore par une couverture lacunaire : ***Malte*** *et* ***Chypre****.*

Ces deux États partagent plusieurs caractéristiques qui expliquent cette situation :

- ***Taille démographique et géographique réduite :*** *Malte et Chypre sont les deux plus petits États membres de l'UE, tant en population qu'en superficie. Cette échelle limitée a souvent pour conséquence une moindre priorité statistique, notamment dans les publications internationales.*
- ***Absence de production d'énergie primaire :*** *historiquement, ces pays ne produisent quasiment aucune énergie fossile ou nucléaire, et sont massivement dépendants des importations. Cela entraîne une absence ou une rareté de certaines variables dans les bases de données, comme la production de charbon ou de gaz naturel.*
- ***Intégration relativement récente à l'UE :*** *bien que Chypre et Malte aient rejoint l'Union Européenne en 2004, les données disponibles avant cette date sont parfois éparses ou non harmonisées avec les standards statistiques européens (Eurostat, ENTSO-E, etc.).*
- ***Limites structurelles dans les systèmes de collecte de données :*** *ces pays disposent d'un réseau statistique national plus restreint, ce qui peut freiner la mise à jour ou la standardisation des données historiques, surtout dans des domaines techniques comme l'énergie.*

Ainsi, même après application de méthodes robustes de complétion, certaines lacunes subsistent, non pas à cause d'un échec du traitement, mais parce que l'information n'existe tout simplement pas ou a été mal collectée à l'époque. Cela souligne une vérité importante en data science : ***le traitement algorithmique ne peut pas toujours compenser l'absence structurelle de données réelles****. C'est une limite qu'il faut reconnaître dans toute démarche d'analyse responsable.*

2.1.5 Modélisation relationnelle préalable

Avant même de procéder à la création effective des tables dans PostgreSQL, nous avons réalisé une phase de conception préalable, indispensable à toute démarche de structuration de données relationnelles. Cette modélisation, réalisée à l'aide de l'outil dbdiagram.io, a permis de définir les entités principales, leurs attributs et les relations entre elles.

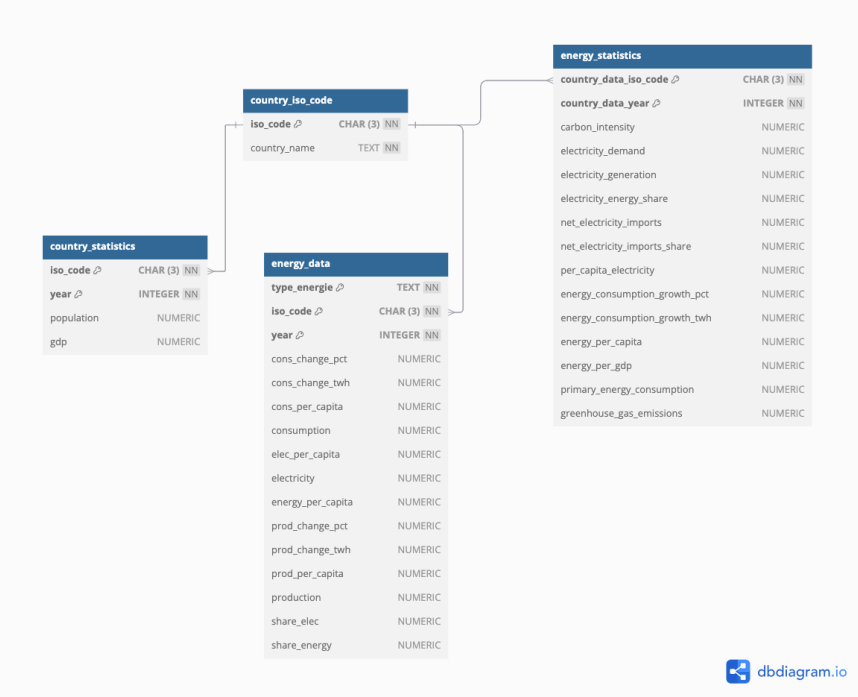

FIGURE 2.4 – Schéma relationnel préliminaire de la base de données énergétique

Le schéma ci-dessus illustre la structure logique que nous avons définie. On y retrouve les entitées suivantes :

— **country_iso_code**, qui contient le référentiel des pays de l'Union Européenne via leurs codes ISO ;
— **country_statistics**, contenant des indicateurs macroéconomiques (population, PIB) par pays et par année ;
— **energy_statistics**, regroupant des données globales sur la consommation et la production énergétique ;
— La table **energy_data** est un exemple de table unifiée, dans laquelle les données de production et de consommation sont fusionnées pour chaque filière énergétique (charbon, gaz, pétrole, nucléaire, solaire, éolien, etc.). En réalité, une table distincte est créée pour chaque type d'énergie.

L'approche retenue repose sur la création d'une seule table normalisée pour les énergies, identifiée par le champ `type_energie`, ce qui permet de regrouper l'ensemble des indicateurs dans une structure commune, tout en évitant la multiplication de tables spécialisées. Cela facilite considérablement les requêtes, les jointures, et les analyses temporelles comparatives entre filières.

Ce travail de modélisation a été une étape clé pour assurer la cohérence, la lisibilité et la maintenabilité de notre base de données relationnelle.

Dans le cadre de notre projet, il était indispensable de structurer les données énergétiques de manière rigoureuse afin de faciliter les analyses et les visualisations ultérieures. Pour cela, nous avons eu recours à une base de données relationnelle PostgreSQL, qui offre des garanties en matière de performance, d'intégrité des données et de normalisation.

La base de données a été construite selon une **modélisation relationnelle** classique respectant les principes de la **Boyce-Codd Normal Form (BCNF)** :

— Chaque table contient des données homogènes portant sur une seule thématique (bioénergie, énergies fossiles, énergie nucléaire, etc.).
— Les redondances sont évitées par la factorisation des entités communes, notamment les informations pays via la table `country_iso_code`.
— Toutes les dépendances fonctionnelles sont correctement gérées, avec des clés primaires bien définies.
— Pour chaque dépendance fonctionnelle $X \to Y$, X est une clé candidate.

Chaque table (sauf la table `country_iso_code`) possède une clé primaire composée {`iso_code, Année`} : le code ISO du pays (`CHAR(3)`) et l'année de référence (`INTEGER`). Cette combinaison unique assure l'unicité de chaque enregistrement temporel pour un pays donné. De plus, nous avons appliqué des `contraintes d'intégrité référentielle` (`FOREIGN KEY`) entre les tables secondaires et la table principale des pays, ce qui garantit que les données sont toujours associées à un pays valide, et supprime automatiquement les lignes orphelines en cas de suppression d'un pays (`ON DELETE CASCADE`).

2.1.6 Approche modulaire et sémantique des données

La structure adoptée repose sur une logique modulaire. Chaque table représente un bloc sémantique clair correspondant à une source d'énergie (charbon, pétrole, gaz, solaire, etc.). Cela facilite :

— la maintenance du schéma relationnel,
— les mises à jour différenciées selon les thématiques,
— l'optimisation des requêtes SQL ciblées,
— et la lisibilité globale du modèle pour tout utilisateur externe.

Cette structuration a également permis de simplifier les jointures SQL futures, en évitant des tables trop larges contenant des centaines de colonnes, ce qui nuit à la clarté et à la performance.

2.1.7 Génération automatisée du schéma via Python

Pour éviter les erreurs humaines et accélérer le processus de création des tables, nous avons développé un script Python dédié. Ce script lit dynamiquement la structure des fichiers CSV (via `pandas`) et génère les commandes `CREATE TABLE` correspondantes, avec les types adaptés, les clés primaires, et les contraintes de clé étrangère. Ce programme permet :

— une reproductibilité du processus de modélisation (important dans un cadre de projet scientifique),
— une adaptabilité à de nouveaux jeux de données structurés de façon similaire,
— et une cohérence stricte dans les noms de colonnes et de tables.

Le script est fourni en annexe du rapport.

2.1.8 Chargement des données avec la commande COPY

Le remplissage des tables a été réalisé grâce à la commande `COPY FROM` de PostgreSQL. Cette commande est particulièrement adaptée pour importer de grands volumes de données à partir de fichiers CSV. Elle est bien plus rapide qu'une série de `INSERT INTO` et s'intègre parfaitement dans un pipeline d'automatisation.

Chaque fichier CSV contient les données correspondant exactement à la structure d'une table. Par exemple :

```
COPY fossil_energy_data FROM '.../fossil_energy_data.csv' DELIMITER ';' CSV HEADER;
```

Ces commandes ont permis de charger efficacement plus de 100 000 lignes de données en quelques secondes. Tous les fichiers sont séparés, propres, et encodés en UTF-8 pour assurer la compatibilité.

L'ensemble des commandes SQL utilisées, y compris les `CREATE TABLE` et `COPY`, sont jointes en annexe pour assurer la transparence et la reproductibilité complète du projet.

CHAPITRE 3

VISUALISATION INTERACTIVE ET STORYTELLING AVEC GRAFANA

3.1 Conception de tableaux de bord interactifs

Dans la continuité du travail de structuration des données sous PostgreSQL, nous avons développé un tableau de bord interactif à l'aide de l'outil Grafana, dans le but de transformer des indicateurs statistiques bruts en visualisations accessibles et intuitives. L'objectif de cette section est de rendre la dynamique énergétique européenne lisible, compréhensible et exploitable, tant pour un public averti que non spécialiste.

Grafana a été choisi pour ses nombreux avantages : il permet une connexion directe à notre base de données PostgreSQL, l'écriture de requêtes SQL personnalisées, la création de visualisations dynamiques (séries temporelles, barres horizontales, courbes comparatives), ainsi que l'intégration d'éléments de storytelling comme les annotations temporelles. Chaque visualisation (ou panel) a été conçue pour répondre à une interrogation précise, basée sur les observations tirées des données.

Dans cette optique, nous avons choisi d'étudier la problématique suivante :

« L'Europe face au défi de la transition énergétique et l'objectif de neutralité carbone en 2050 »

L'Union Européenne s'est engagée dans une profonde transformation de son système énergétique, avec pour ambition de réduire de 55% ses émissions de gaz à effet de serre d'ici 2030, et d'atteindre la neutralité carbone à l'horizon 2050, conformément à la Loi européenne sur le climat adoptée en 2021. Ces objectifs s'inscrivent dans un contexte marqué par d'importantes disparités entre pays membres, une dépendance toujours forte à certaines énergies fossiles, et un besoin urgent d'accélérer le déploiement des énergies renouvelables.

Pour mieux comprendre les dynamiques à l'œuvre et évaluer si l'Europe est véritablement en transition, notre projet cherche à répondre aux questions suivantes :

— Comment la consommation d'énergie a-t-elle évolué dans l'Union Européenne depuis 2000 ?
— Quels pays ont le plus progressé vers une production et une consommation d'énergies renouvelables ?
— Dans quelle mesure les États membres restent-ils dépendants des énergies fossiles ?
— Quels ont été les effets des événements récents, notamment la guerre en Ukraine, sur les choix énergétiques européens ?
— Et surtout : l'UE pourra-t-elle réellement atteindre ses objectifs de 2030 en matière de réduction des émissions ?

Les panels du tableau du bord (ci-dessous) sont interactifs : l'utilisateur peut filtrer les années ou les pays, afficher les valeurs exactes au survol, et interpréter les ruptures de tendance à l'aide des annotations visuelles. L'exemple ci-dessous montre l'aspect général de notre tableau de bord :

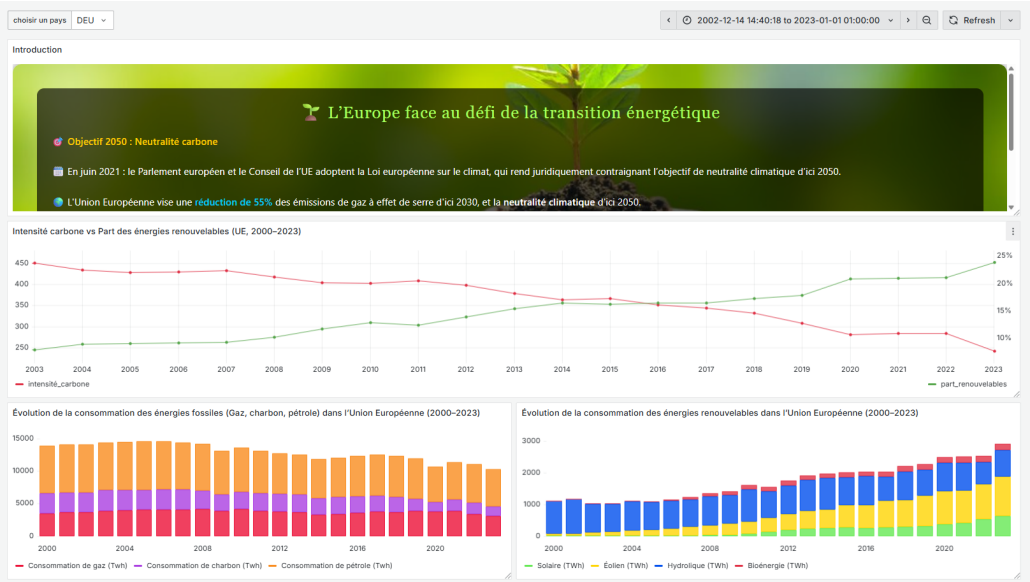

FIGURE 3.1 – Aperçu du tableau de bord principal dans Grafana (vue globale des panels interactifs)

Ce travail de visualisation constitue le socle du chapitre suivant, où nous allons confirmer par des méthodes statistiques rigoureuses les tendances identifiées visuellement ici.

3.2 Présentation du tableau de bord réalisé

Dans la continuité de notre démarche de visualisation, nous avons élaboré un tableau de bord interactif destiné à représenter de manière synthétique et structurée les principales dynamiques énergétiques observées au sein de l'Union Européenne.

Ce tableau de bord est composé de plusieurs panels, chacun ayant pour objectif de contribuer à la réponse globale aux questions soulevées dans la problématique définie en début de chapitre. Chaque panel apporte un éclairage précis sur des indicateurs clés, tels que l'évolution de la consommation d'énergies fossiles, la progression des énergies renouvelables ou encore la réduction de l'intensité carbone, permettant ainsi de mieux comprendre la dynamique énergétique à l'échelle européenne.

Panel 1 – Consommation d'énergies fossiles en Europe (2000–2023)

Ce premier panel illustre l'évolution annuelle de la consommation des principales énergies fossiles utilisées en Europe : le charbon, le gaz et le pétrole. Ces trois sources ont été choisies en raison de leur importance historique dans le mix énergétique européen et de leur impact environnemental majeur.

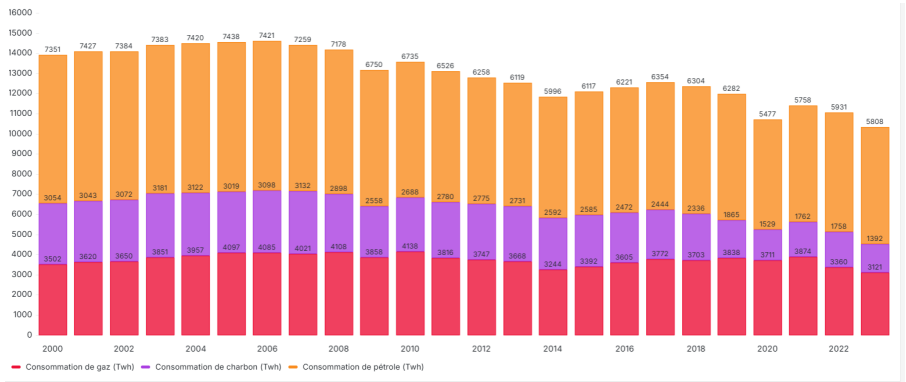

FIGURE 3.2 – Évolution de la consommation d'énergies fossiles (Twh) en Europe (2000–2023)

L'analyse visuelle permet de mettre en évidence une baisse régulière de la consommation de charbon,

traduisant une volonté progressive de désengagement des États vis-à-vis de cette ressource fortement émettrice de CO_2. La consommation de gaz, quant à elle, apparaît relativement stable sur la période, tandis que celle du pétrole reste globalement élevée avec de légères fluctuations.

Cette figure illustre la dynamique temporelle des principales énergies fossiles consommées en Europe sur la période récente. Elle permet de visualiser l'évolution conjointe du charbon, du gaz et du pétrole, en mettant en évidence les ruptures de tendance ou les stabilités relatives.

Panel 2 – Consommation des énergies Renouvelables en Europe (2000–2023)

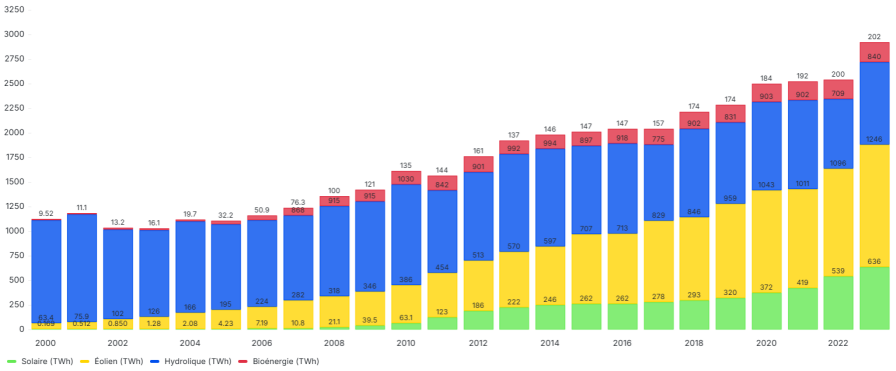

FIGURE 3.3 – Évolution de la consommation des energies renouvelables en Europe (2000–2023)

Ce panel présente l'évolution de la consommation des principales sources d'énergies renouvelables en Europe : solaire, éolien, hydroélectricité et bioénergie. Chaque filière suit une dynamique propre, portée par les avancées technologiques et les choix politiques des États membres.

L'hydroélectricité reste relativement stable sur la période, tandis que le solaire et l'éolien connaissent une forte progression, soutenue par les politiques européennes en faveur de la transition énergétique. Cette croissance répond aux objectifs du Green Deal européen visant à réduire la dépendance aux énergies fossiles et à atteindre la neutralité carbone d'ici 2050. La bioénergie progresse également, mais de manière plus modérée par rapport aux autres sources.

Ce graphique met ainsi en évidence la transformation progressive du mix énergétique européen, marquée par l'essor des énergies renouvelables et la volonté politique d'accélérer la transition vers un système énergétique plus durable.

Panel 3 – Corrélation entre intensité carbone et part des énergies renouvelables (2000–2023)

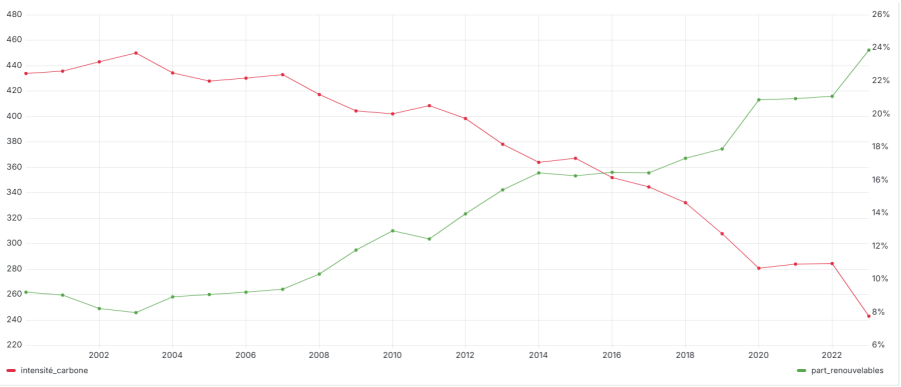

FIGURE 3.4 – Corrélation entre intensité carbone et part des énergies renouvelables (2000–2023)

Ce troisième panel met en relation deux indicateurs fondamentaux de la transition énergétique européenne : d'une part, l'intensité carbone de la production d'électricité (mesurée en grammes de CO_2 par kilowattheure), et d'autre part, la part des énergies renouvelables dans le mix électrique total.

Nous avons choisi de représenter ces deux séries sur un graphique à double axe Y afin de mettre en évidence leur évolution simultanée sur la période 2000–2023. L'observation visuelle révèle que : plus la part des énergies renouvelables augmente, plus l'intensité carbone diminue. Cette dynamique illustre concrètement l'impact environnemental positif des politiques européennes de décarbonation et justifie l'importance stratégique accordée aux investissements dans les énergies vertes.

Cette figure met en relation deux variables clés de la transition énergétique : l'intensité carbone d'une part, et la part des énergies renouvelables d'autre part. Le graphique illustre une tendance inverse manifeste entre les deux indicateurs, ce qui permet de visualiser l'impact environnemental positif des politiques énergétiques mises en œuvre sur la période.

Panel 4 – Top 5 des pays européens en part d'énergies renouvelables (2023)

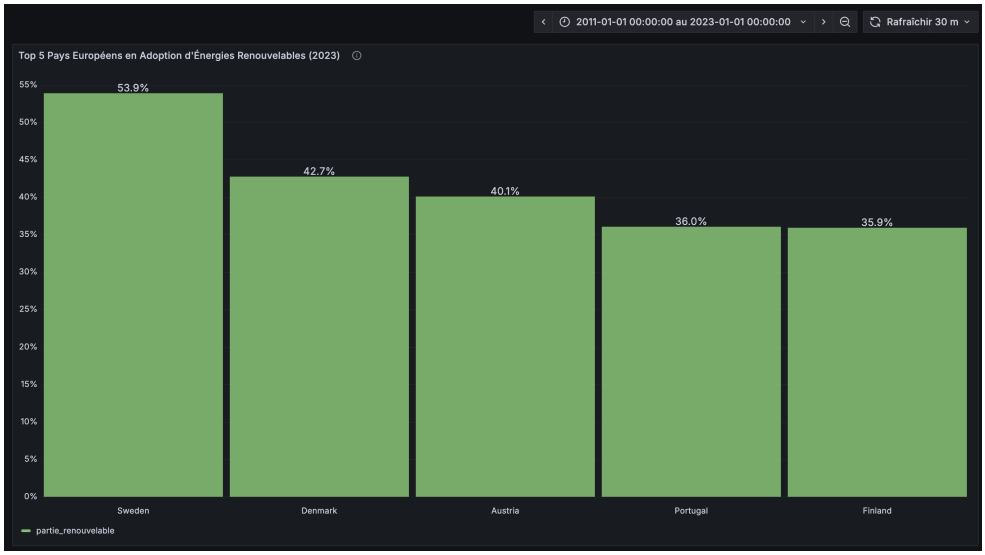

FIGURE 3.5 – Top 5 des pays européens en part d'énergies renouvelables (2023)

Ce panel propose un classement horizontal des cinq pays européens ayant enregistré la plus forte part d'énergies renouvelables dans leur consommation d'énergie en 2023. En choisissant cette année de référence récente, nous avons cherché à mettre en évidence les pays les plus engagés dans la transition énergétique, sur la base de données complètes et à jour.

Cette visualisation permet de comparer rapidement les performances entre États membres et de souligner les efforts exemplaires de certains pays comme la Suède, le Danemark ou le Portugal. Ces résultats illustrent des politiques nationales volontaristes en matière de production d'énergie propre et renforcent la compréhension des disparités régionales au sein de l'Union Européenne.

Cette figure présente un classement horizontal des cinq pays les plus avancés en matière d'énergies renouvelables. Elle permet de comparer rapidement les performances nationales en 2023, et de mettre en lumière les pays moteurs de la transition énergétique au sein de l'Union Européenne.

Panel 5 – Comparaison de la part des énergies renouvelables et fossiles dans la consommation d'énergie de l'Union Européenne (2000–2023)

Maintenant, nous allons faire la comparaison entre la part des énergies renouvelables et celle des énergies fossiles dans la consommation totale d'énergie de l'Union Européenne entre 2000 et 2023.

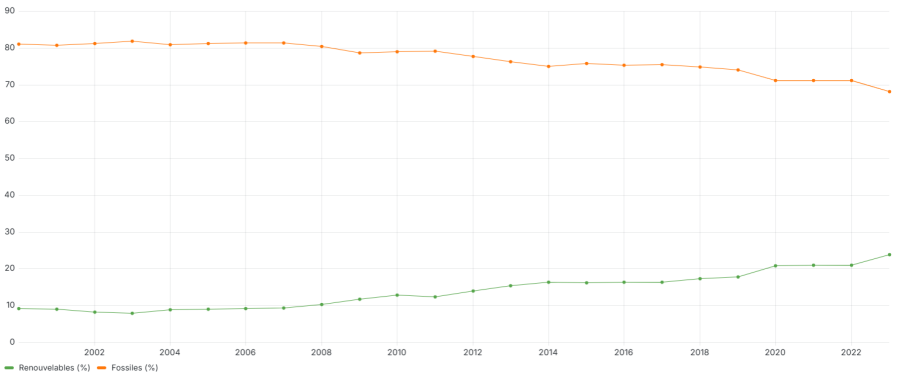

FIGURE 3.6 – Evolution de la part des énergies renouvelables et fossiles dans la consommation d'énergie de l'UE (2000–2023)

Ce panel met en évidence deux dynamiques opposées. La part des énergies fossiles reste longtemps dominante et relativement stable, avant d'entamer un déclin plus marqué à partir des années 2010. Cette baisse reflète les effets progressifs des politiques climatiques européennes et les engagements en faveur de la réduction des émissions de gaz à effet de serre.

À l'inverse, la part des énergies renouvelables affiche une progression continue, témoignant d'une volonté affirmée des États membres de développer des alternatives durables. Cette tendance s'intensifie après 2020, en lien avec les objectifs du Pacte Vert européen, les plans de relance post-COVID et la crise énergétique liée à la guerre en Ukraine.

Ce graphique met ainsi en lumière le rééquilibrage progressif du mix énergétique européen, dans une logique de transition vers un système plus sobre en carbone.

Panel 6 – Réduction des émissions de Gaz à effet de serre (GES) : où en est l'Union Européenne ?

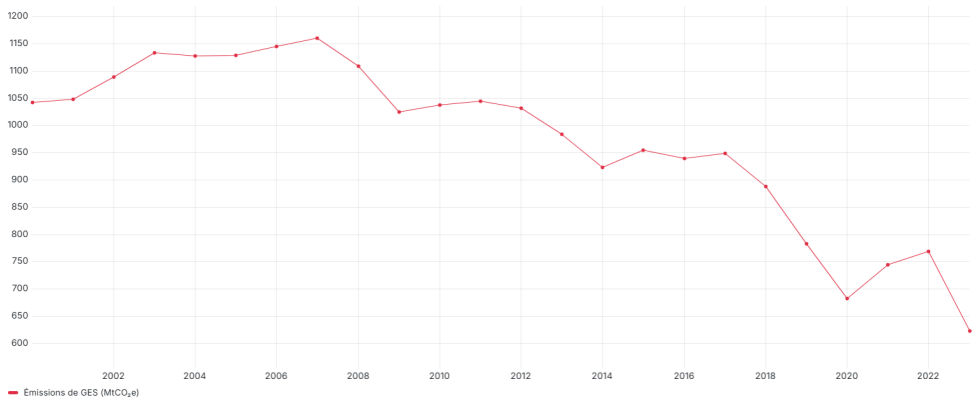

FIGURE 3.7 – Evolution de la part des énergies renouvelables et fossiles dans la consommation d'énergie de l'UE (2000–2023)

Le graphique ci-dessus retrace l'évolution des émissions de gaz à effet de serre (GES) dans l'Union Européenne sur une période de 23 ans. ces émissions sont un indicateur central pour évaluer les efforts de décarbonation du continent.

La tendance générale est à la baisse progressive, avec quelques phases de stagnation ou de légères hausses intermédiaires. Après une période relativement stable au début des années 2000, une première baisse notable apparaît autour de 2008–2009, probablement liée à la crise financière.

À partir de 2014, la trajectoire devient plus clairement descendante, reflétant les impacts combinés des politiques climatiques européennes, du développement des énergies renouvelables, de l'amélioration de l'efficacité énergétique, ainsi que des évolutions industrielles.

La baisse marquée observée autour de 2020 peut être partiellement attribuée aux effets de la pandémie de COVID-19, qui a provoqué un ralentissement économique et une baisse temporaire de la consommation d'énergie. La chute très nette de 2023 semble traduire un renforcement récent des efforts de réduction des émissions, dans un contexte de crise énergétique et d'accélération de la transition écologique.

Ce graphique met en évidence la tendance vers une décarbonation progressive du système énergétique européen, bien que des efforts supplémentaires restent nécessaires pour atteindre les objectifs climatiques à l'horizon 2030 et 2050.

Remarque 3.2.1

Bien que notre jeu de données couvre une période bien plus large (à partir de 1965), nous avons choisi de nous concentrer sur l'intervalle 2000–2023 pour plusieurs raisons. Tout d'abord, cette période correspond à une phase de transition énergétique particulièrement active au sein de l'Union Européenne, marquée par des politiques climatiques ambitieuses, des engagements internationaux (comme les accords de Paris), ainsi que des ruptures géopolitiques majeures (crise sanitaire, guerre en Ukraine).

Par ailleurs, les données récentes sont généralement plus complètes, homogènes et comparables entre les pays membres, ce qui renforce la fiabilité de l'analyse. En se focalisant sur cette période récente, nous avons pu produire des visualisations plus cohérentes et plus pertinentes dans le cadre de notre storytelling visuel.

Conclusion 3.2.1

En résumé, les panels montrent que l'Union Européenne a fait de grands efforts pour avancer vers la transition énergétique. On observe une baisse des émissions de gaz à effet de serre, une part croissante des énergies renouvelables et une réduction progressive de la dépendance aux énergies fossiles.

Mais une question reste ouverte : l'UE atteindra-t-elle vraiment ses objectifs d'ici 2030 ? Et comment des événements comme la crise du COVID-19 ou la guerre en Ukraine ont-ils influencé ces efforts ?

C'est ce que nous allons explorer dans le prochain chapitre.

CHAPITRE 4

ANALYSES STATISTIQUES POUR LA DÉTECTION D'ANOMALIES ET LA PRÉDICTION

Objectifs de chapitre :

Ce chapitre vise à répondre à la problématique posée dans le chapitre précédent à travers une approche statistique et prédictive structurée, articulée en plusieurs étapes :

1. Explorer les données énergétiques de l'UE (2000-2023) pour confirmer les tendances observées.
2. Détecter les anomalies (ruptures, crises) dans les séries temporelles des GES, renouvelables et fossiles.
3. Prédire l'évolution des GES jusqu'en 2030 via des modèles de Machine Learning.

Cette démarche vise à fournir une analyse robuste permettant à la fois de comprendre les dynamiques actuelles du système énergétique européen et d'évaluer les perspectives de réussite des objectifs climatiques à moyen terme.

4.1 Notions en statistiques descriptives

Définition 4.1.1 (Moyenne arithmétique)
Si $(x_1, x_2, ..x_n)$ sont les valeurs d'une variable statistique X, alors la moyenne arithmétique de X est donnée par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Définition 4.1.2 (Variance - Écart-type)
Soit $(x_1, x_2, ..x_n)$ sont les valeurs d'une variable statistique X et $\bar{x}$ sa moyenne, alors on a :
***La variance** de X est définie par :*

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

***L'écart-type** de la variable statistique X est :*

$$\sigma = \sqrt{Var(X)}$$

Définition 4.1.3 (Minimum et maximum)
Le minimum et le maximum correspondent respectivement aux valeurs les plus faibles et les plus élevées de l'échantillon.

Pour centraliser ces résultats, un nouveau schéma nommé `stats` a *été* créé dans PostgreSQL, contenant une table `energy_stats_summary` pour stocker les statistiques calculées (moyenne, *écart-type*, minimum et maximum). Ces calculs ont *été réalisés* par requêtes SQL, garantissant la traçabilité et la reproductibilité des analyses.

Résultats des statistiques descriptives :

Variable	Moyenne	Écart-type	Minimum	Maximum
Émissions GES ($MtCO_2e$)	42.0	68.2	0.08	348
Consommation d'inérgie (TWh)	4.04	40.5	-295	252
Intensité carbone (gCO_2/kWh)	451	237	33.3	968

Définition 4.1.4
*Un **quantile d'ordre** $p \in [0,1]$ est une valeur q_p telle que :*

$$P(X \leq q_p) = p$$

Autrement dit, q_p est la valeur en dessous de laquelle se trouvent $100 \times p$ % des données.

Remarque 4.1.1
- *Médiane = quantile d'ordre* 0,5.
- *Premier quartile (Q_1) = quantile d'ordre* 0,25.
- *Troisième quartile (Q_3) = quantile d'ordre* 0,75.
- *Déciles = quantiles d'ordre* 0,1, 0,2, …, 0,9.
- *Centiles (ou percentiles) = quantiles d'ordre* 0,01,..,0,99.

4.2 Analyse des corrélations entre variables clés

L'analyse des corrélations permet de comprendre les relations linéaires entre les principaux indicateurs *énergétiques*. Pour deux variables quantitatives X et Y, le coefficient de corrélation de Pearson est donné par :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Application 4.2.1
dans le cadre de notre problématique de transition energytique on veut comprendre les relations linéaires entre plusieurs variables clés du système énergétique européen à savoir :
- ***Les émissions de GES ($MtCO_2e$)***
- ***La consommation d'énergie (TWh)***
- ***L'intensité carbone (gCO_2/kWh)***

Les corrélations ont été calculées à partir des données stockées dans les tables PostgreSQL (`energy_statistics`, `renewable_energy_data`, `fossil_energy_data`), en prenant uniquement en compte les pays européens.

Les variables considérées et leurs corrélations sont les suivantes :

Corrélation	*Coefficient de Pearson*
Part renouvelables (%) vs Intensité carbone (gCO_2/kWh)	*-0.587*
Part fossiles (%) vs Émissions GES ($MtCO_2e$)	*0.204*
Consommation énergie (TWh) vs Émissions GES ($MtCO_2e$)	*0.076*
Part renouvelables (%) vs Émissions GES ($MtCO_2e$)	*-0.207*
Part fossiles (%) vs Intensité carbone (gCO_2/kWh)	*0.748*

Interprétation des corrélations :

— *Une forte corrélation positive ($r = 0.748$) est observée entre la part des énergies fossiles et l'intensité carbone, indiquant que l'utilisation accrue des fossiles augmente significativement la quantité de CO_2 en (kWh) par produit.*

— *Une corrélation négative ($r = -0.587$) entre la part des énergies renouvelables et l'intensité carbone confirme que l'accroissement des renouvelables contribue à réduire l'intensité carbone.*

— *Les corrélations entre les émissions de GES et les autres variables (part fossiles, consommation d'énergie) sont faibles, reflétant l'influence combinée de multiples facteurs sur les émissions globales.*

Remarque 4.2.1

Ces résultats sont cohérents avec les panels de visualisation précédemment présentés, qui illustrent la dynamique inverse entre la progression des énergies renouvelables et la réduction de l'intensité carbone dans l'UE.

4.3 Méthodes statistiques pour la détection des anomalies

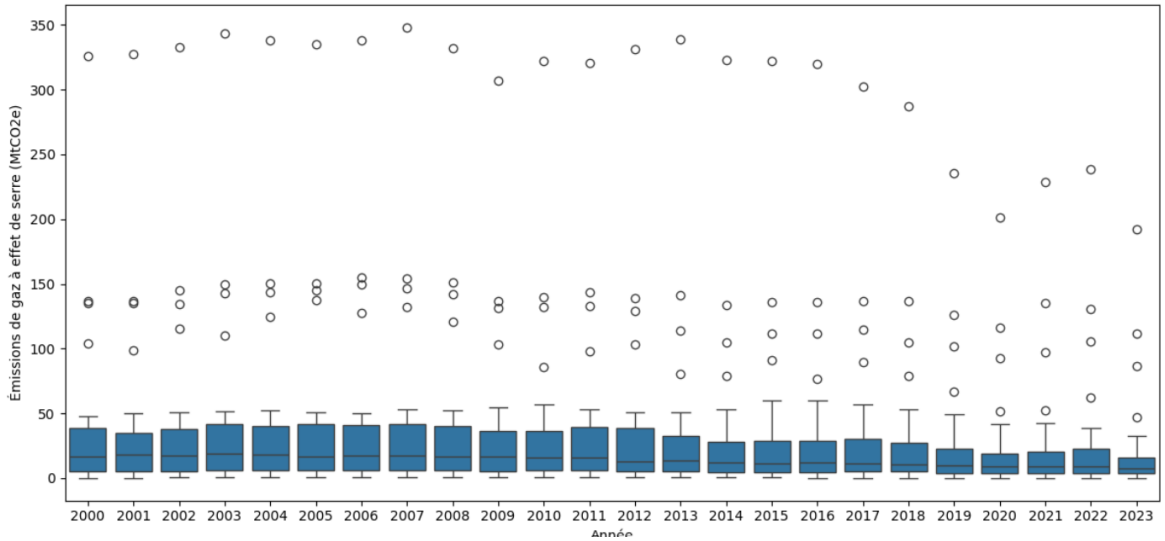

FIGURE 4.1 – Distribution des émissions de gaz à effet de serre (MtCO2e) par année(2000-2023

La figure présenté ci-dessus permet de visualiser la distribution des émissions de gaz à effet de serre (GES) dans l'Union Européenne entre 2000 et 2023. On constate que la majorité des pays ont des émissions relativement faibles (souvent inférieures à 50 $MtCO_2e$), tandis que plusieurs pays affichent des niveaux très élevés, dépassant parfois 300 $MtCO_2e$. Ces observations extrêmes apparaissent sous forme de valeurs aberrantes (outliers). Ce graphique vient renforcer les résultats des statistiques descriptives précédentes : la dispersion importante et l'écart-type élevé trouvent une explication visuelle claire. Il met également en évidence la nécessité d'utiliser des méthodes robustes pour l'analyse statistique, notamment pour la détection d'anomalies, qui constitue l'étape suivante de notre étude.

4.3.1 Méthode de Tukey pour la détection d'anomalies

La méthode de Tukey, introduite par le statisticien américain John W. Tukey en 1977 dans son ouvrage *Exploratory Data Analysis*, est une approche robuste permettant d'identifier les valeurs aberrantes dans un jeu de données numériques. Contrairement aux méthodes basées sur la moyenne et l'écart-type, souvent sensibles aux extrêmes, la méthode de Tukey repose exclusivement sur les quartiles, ce qui lui confère une grande stabilité face aux données dispersées ou asymétriques.

Elle utilise l'écart interquartile (IQR), défini comme la différence entre le troisième quartile (Q_3) et le premier quartile (Q_1). Toute valeur située en dehors de l'intervalle : $[Q_1 - 1{,}5 \times \text{IQR}\,;\ Q_3 + 1{,}5 \times \text{IQR}]$ est consi-

dérée comme valeur aberrante. Cette règle empirique est simple mais efficace pour détecter les observations inhabituelles.

Dans notre projet, nous avons appliqué la méthode de Tukey pour détecter les valeurs aberrantes dans l'évolution des émissions de gaz à effet de serre des pays de l'Union européenne entre 2000 et 2022. Les bornes inférieure et supérieure de cette méthode, définies respectivement comme :

$$Q_1 - 1{,}5 \times \text{IQR} \quad \text{et} \quad Q_3 + 1{,}5 \times \text{IQR},$$

sont représentées dans le graphique suivant par des courbes rouges en pointillés.

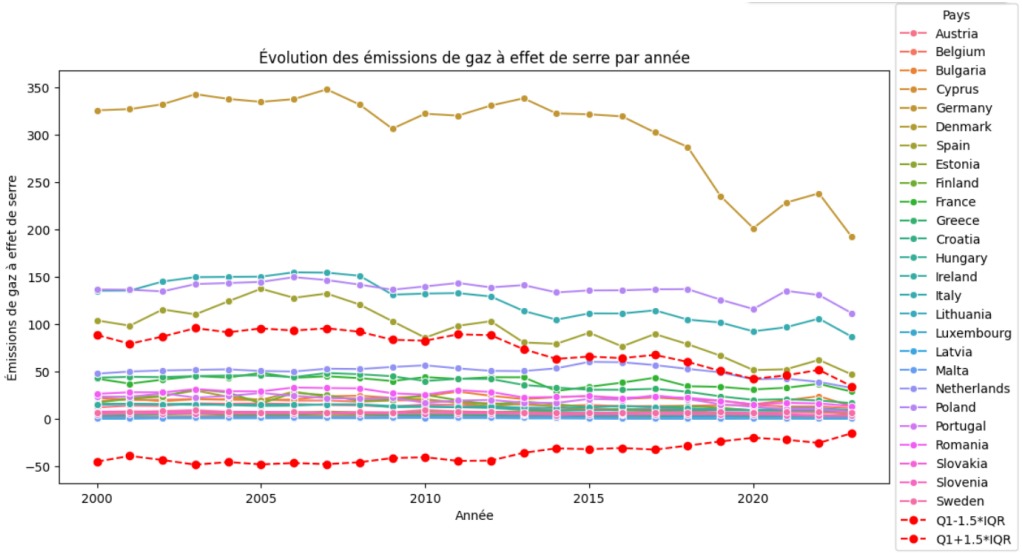

FIGURE 4.2 – Détection des anomalies sur la distribution des émissions de gaz à effet de serre ($MtCO_2$) par pays.

L'analyse du graphique montre que quatre pays dépassent régulièrement la borne supérieure de Tukey : **l'Allemagne**, **l'Italie**, **l'Espagne** et **la Pologne**. Ces pays sont considérés comme des **valeurs aberrantes supérieures**, car leurs niveaux d'émissions sont nettement plus élevés que ceux de la majorité des États membres.

Ce constat peut être expliqué par plusieurs facteurs :

— **L'Allemagne** se distingue par son statut de première puissance industrielle d'Europe et sa dépendance historique au charbon, ce qui en fait le plus grand émetteur de l'UE sur l'ensemble de la période.

— **L'Italie** et **l'Espagne**, en tant que pays fortement peuplés et industrialisés, conservent des niveaux élevés d'émissions, notamment en raison de leur mix énergétique encore partiellement fossile.

— **La Pologne** reste particulièrement dépendante du charbon pour sa production d'électricité, ce qui entraîne des émissions durablement élevées.

On note également une tendance à la baisse des émissions, en particulier en Allemagne à partir de 2019, traduisant l'effet des politiques climatiques, des efforts de transition énergétique, et de la réduction temporaire de l'activité économique liée à la crise du COVID-19. En parallèle, les courbes de Tukey se resserrent progressivement, ce qui indique une réduction des écarts entre pays au fil du temps.

4.3.2 Méthode Z-score (approche mobile) pour la détection d'anomalies

Le **Z-score** est une méthode statistique efficace permettant d'identifier des anomalies dans une série temporelle en mesurant l'écart d'une observation par rapport à la moyenne, en unités d'écart-type. Dans le cadre de l'analyse des données énergétiques de l'Union Européenne, une version adaptée dite **Z-score mobile asymétrique** est utilisée afin de détecter des ruptures locales dans des tendances souvent stables.

Définition 4.3.1 (Moyenne mobile)
*Soit une série de données $x_1, x_2, \ldots, x_n$, la **moyenne mobile** de taille $2k+1$, centrée autour du point x_i, est définie par :*

$$\mu_i = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} x_j$$

pour tout i tel que $k \leq i \leq n-k$, c'est-à-dire là où la fenêtre est entièrement contenue dans la série.

Remarque 4.3.1
Cette moyenne permet de lisser les fluctuations locales de la série en calculant, pour chaque point, la moyenne de ses k voisins précédents, k suivants, et lui-même.

dfdf

Définition 4.3.2 (Écart-type mobile)
Soit une série $x_1, x_2, \ldots, x_n$, et une fenêtre de taille $2k+1$. L'écart-type mobile autour du point x_i est défini par :

$$\sigma_i = \sqrt{\frac{1}{2k+1} \sum_{j=i-k}^{i+k} (x_j - \mu_i)^2}$$

où μ_i est la moyenne mobile centrée sur x_i.

Remarque 4.3.2
*L'**écart-type mobile** est une mesure de la variabilité locale d'une série de données.*

Pour chaque point x_i, on calcule l'écart-type des valeurs contenues dans une fenêtre centrée autour de ce point. Cela permet de détecter les fluctuations locales et les zones de forte ou faible dispersion.

Définition 4.3.3 (Z-score mobile)
Soit une série temporelle $(x_1, x_2, \ldots, x_n)$, la version mobile du Z-score à l'instant t s'exprime ainsi :

$$Z_t = \frac{x_t - \mu_t}{\sigma_t}$$

où :
- *x_t est la valeur observée à l'instant t*
- *μ_t est la moyenne mobile autour de x_t*
- *σ_t est l'écart-type mobile autour de x_t*

Seuil de détection : Dans cette étude, une observation est considérée comme une anomalie si :

$$|Z_t| > 2.5$$

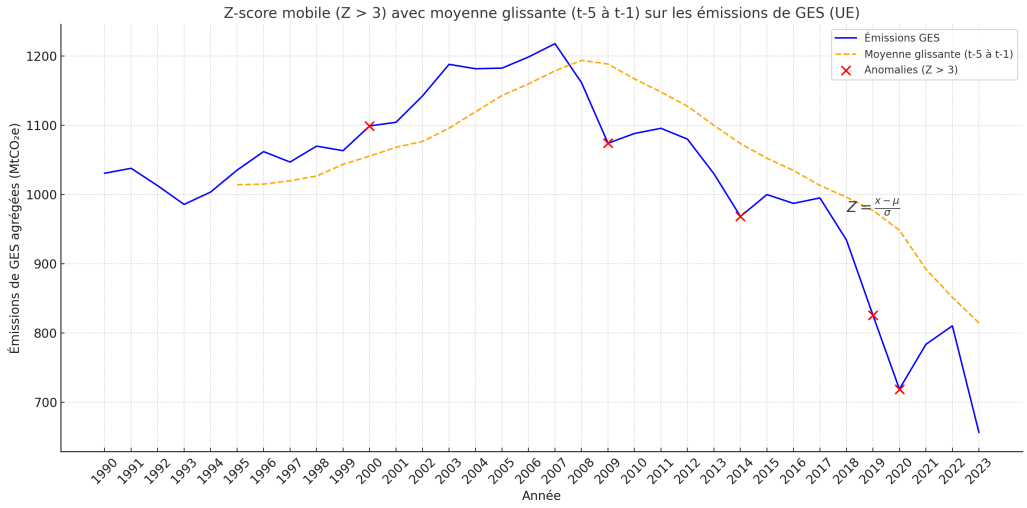

FIGURE 4.3 – Illustration visuelle de la détection d'une anomalie avec flèche Z-score et moyenne locale

TABLE 4.1 – Détails des anomalies détectées par Z-score mobile ($|Z| > 3$) sur les émissions de GES (UE, 2000–2023)

Année	Émissions GES	Z-score	Moyenne mobile	Écart-type
2000	1098,94	3,09	1055,44	14,07
2009	1074,04	-5,44	1188,33	21,02
2014	968,39	-4,09	1073,51	25,69
2019	826,03	-5,66	977,02	26,66
2020	719,20	-3,13	948,55	73,34

Remarque 4.3.3 (Interprétation visuelle)
Une forte valeur de Z-score ne correspond pas nécessairement au pic ou au creux absolu. Elle reflète une ***variation soudaine par rapport aux années précédentes****. Cela explique pourquoi certaines anomalies ne coïncident pas avec les extrêmes visibles sur la courbe.*

4.4 Modélisation prédictive

4.4.1 Le modèle de regression linèaire

Commençons par un exemple afin de fixer les idées. Dans le contexte de la lutte contre le changement climatique, on s'intéresse à l'évolution des **émissions de gaz à effet de serre** (exprimées en millions de tonnes équivalent CO_2) au sein de l'Union européenne. En particulier, on cherche à savoir s'il est possible d'expliquer la tendance générale des émissions au cours du temps à l'aide d'un modèle simple basé sur l'année d'observation. Les données utilisées sont les suivantes :

Année	2000	2001	2002	2003	2004	...
Émissions ($MtCO_2e$)	1043.02	1048.49	1089.21	1133.87	1127.91	...

Tableau des émissions totales de gaz à effet de serre dans l'UE (extrait 2000–2004).

D'un point de vue pratique, le but de cette régression est double :
— Ajuster un modèle pour expliquer l'évolution des émissions en fonction du temps ;
— Prédire les émissions futures à partir de la tendance observée.

Avant toute analyse quantitative, il est pertinent de représenter graphiquement les données, comme illustré dans la figure suivante :

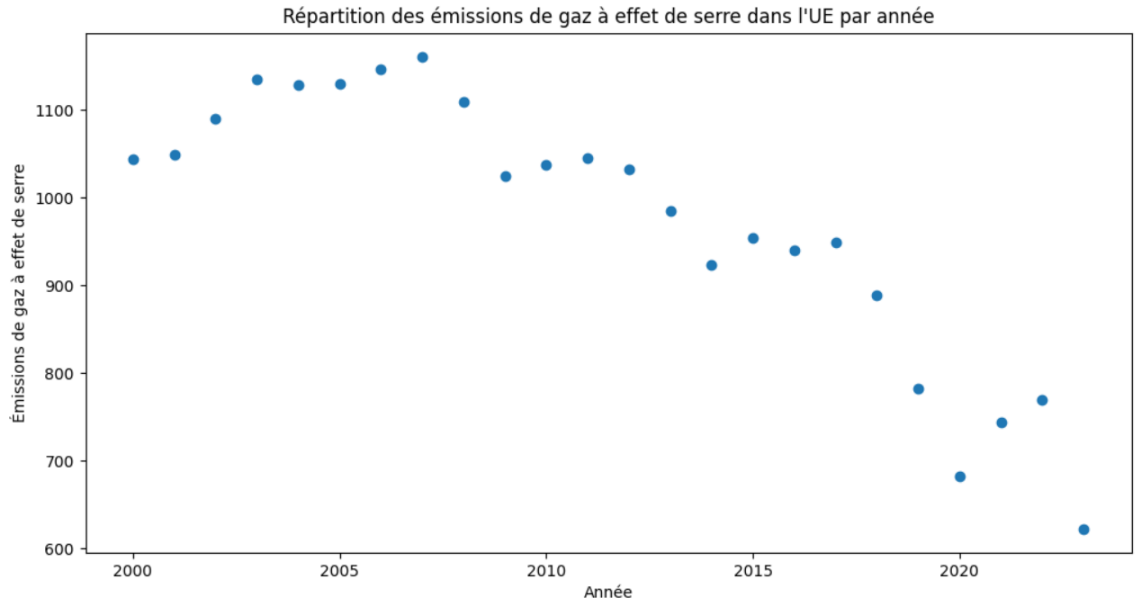

FIGURE 4.4 – Répartition des émissions de gaz à effet de serre ($MtCO_2$) dans lÚE par année

Pour analyser la relation entre les x_i (les années) et les y_i (les émissions de gaz à effet de serre), nous allons chercher une fonction f telle que :

$$y_i \approx f(x_i).$$

Pour préciser le sens de $\approx$, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données observées. Il convient également de définir une classe de fonctions $\mathcal{F}$ dans laquelle est supposée vivre la fonction réelle, encore inconnue, qui relie l'évolution temporelle aux émissions.

Le problème mathématique peut alors s'écrire de la façon suivante :

$$\arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} L(y_i - f(x_i)),$$

où n représente le nombre d'années disponibles dans l'échantillon, et $L(\cdot)$ désigne une fonction de perte (*Loss* en anglais), qui mesure l'écart entre la valeur prédite par le modèle et la valeur observée.

4.4.2 Modélisation

Dans de nombreuses situations, en première approche, une idée naturelle est de supposer que la variable à expliquer y est une fonction affine de la variable explicative x, c'est-à-dire de chercher f dans l'ensemble $\mathcal{F}$ des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}$. C'est le principe de la régression linéaire simple. On suppose dans la suite disposer d'un échantillon de n points (x_i, y_i) du plan.

Définition 4.4.1 (Modèle de régression linéaire simple)
Soit $\beta_1, \beta_2 \in \mathbb{R}$, un modèle de régression linéaire simple est défini par une équation de la forme :

$$\forall i \in \{1,\dots,n\} \qquad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Les quantités ε_i viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. On les appelle les erreurs (ou bruits) et elles sont supposées aléatoires. Pour pouvoir dire des choses pertinentes sur ce modèle, il faut néanmoins imposer des hypothèses les concernant. Voici celles que nous ferons dans un premier temps :

$$(H)\begin{cases} (H_1) : \mathbb{E}[\varepsilon_i] = 0 & \text{pour tout indice } i \\ (H_2) : \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2 & \text{pour tout couple } (i,j) \end{cases}$$

Les erreurs sont donc supposées centrées, de même variance (homoscédasticité) et non corrélées entre elles (δ_{ij} est le symbole de Kronecker, i.e. $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ si $i \neq j$).

Notons que le modèle de régression linéaire simple de la définition précédente peut encore s'écrire de façon vectorielle :

$$Y = \beta_1 \mathbb{1} + \beta_2 X + \varepsilon,$$

où :
— le vecteur $Y = [y_1, \ldots, y_n]^t$ est aléatoire de dimension n ;
— le vecteur $\mathbb{1} = [1, \ldots, 1]^t$ est le vecteur de $\mathbb{R}^n$ dont les composantes valent toutes 1 ;
— le vecteur $X = [x_1, \ldots, x_n]^t$ est un vecteur de dimension n (non aléatoire) ;
— les coefficients β_1 et β_2 sont les inconnues à estimer du modèle ;
— le vecteur $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^t$ contient les erreurs aléatoires.

Moindres Carrés Ordinaires

Les points (x_i, y_i) étant donnés, le but est maintenant de trouver une fonction affine f telle que la quantité $\sum_{i=1}^n L(y_i - f(x_i))$ soit minimale. Pour pouvoir déterminer f, encore il faut préciser la fonction de coût L. Deux fonctions sont classiquement utilisées :
— le coût absolu $L(u) = |u|$;
— le coût quadratique $L(u) = u^2$.

Les deux ont leurs vertus, mais on privilégiera dans la suite la fonction de coût quadratique. On parle alors de méthode d'estimation par moindres carrés (terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes).

Définition 4.4.2 (Estimateurs des Moindres Carrés Ordinaires)
*On appelle estimateurs des **Moindres Carrés Ordinaires** (en abrégé **MCO**) $\widehat{\beta}_1$ et $\widehat{\beta}_2$ les valeurs minimisant la quantité :*

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $y = \widehat{\beta}_1 + \widehat{\beta}_2 x$.

Calcul des estimateurs de β_1 et β_2

La fonction de deux variables S est une fonction quadratique et sa minimisation ne pose aucun problème, comme nous allons le voir maintenant.

Proposition 4.4.1 (Estimateurs $\widehat{\beta}_1$ et $\widehat{\beta}_2$)
Les estimateurs des MCO ont pour expressions :

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x},$$

et :

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Avec $(\bar{x}, \bar{y})$ est le point moyen.

Preuve :
La première méthode consiste à remarquer que la fonction $S(\beta_1, \beta_2)$ est strictement convexe, donc qu'elle admet un minimum en un unique point $(\widehat{\beta}_1, \widehat{\beta}_2)$, lequel est déterminé en annulant les dérivées partielles de S. On obtient les "équations normales" :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) = 0 \\ \dfrac{\partial S}{\partial \beta_2} = -2\sum_{i=1}^{n} x_i(y_i - \beta_1 - \beta_2 x_i) = 0 \end{cases}$$

La première équation donne :

$$\widehat{\beta}_1 n + \widehat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

d'où l'on déduit immédiatement :

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}. \tag{1.1}$$

où $\bar{x}$ et $\bar{y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i. La seconde équation donne :

$$\widehat{\beta}_1 \sum_{i=1}^{n} x_i + \widehat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

et en remplaçant $\widehat{\beta}_1$ par son expression (1.1), nous avons :

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1.2}$$

L'expression (1.2) de $\widehat{\beta}_2$ suppose que le dénominateur $\sum_{i=1}^{n}(x_i - \bar{x})^2$ est non nul. Or ceci ne peut arriver que si tous les x_i sont égaux, situation sans intérêt pour notre problème et que nous excluons donc a priori dans toute la suite.

Remarque 4.4.1

1. *La relation* $\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}$ *montre que la droite des MCO passe par le point moyen* $(\bar{x}, \bar{y})$.
2. *L'estimateur* $\widehat{\beta}_2$ *peut aussi s'écrire comme suit :*

$$\widehat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}. \tag{1.3}$$

Si cette décomposition n'est pas intéressante pour le calcul effectif de $\widehat{\beta}_2$ *puisqu'elle fait intervenir les quantités inconnues* β_2 *et* ε_i*, elle l'est par contre pour démontrer des propriétés théoriques des estimateurs (biais et variance). Son avantage est en effet de mettre en exergue la seule source d'aléa du modèle, à savoir les erreurs* ε_i.

Avant de poursuivre, notons que le calcul des estimateurs des moindres carrés est purement déterministe : il ne fait en rien appel aux hypothèses (H_1) et (H_2) sur le modèle. Celles-ci vont en fait servir dans la suite à expliciter les propriétés statistiques de ces estimateurs.

Quelques propriétés des estimateurs $\widehat{\beta}_1$ et $\widehat{\beta}_2$

Sous les seules hypothèses (H_1) et (H_2) de centrage, décorellation et homoscédasticité des erreurs ε_i du modèle, on peut déjà donner certaines propriétés des estimateurs $\widehat{\beta}_1$ et $\widehat{\beta}_2$ des moindres carrés.

Théorème 4.4.1 (Estimateurs sans biais)
$\widehat{\beta}_1$ *et* $\widehat{\beta}_2$ *sont des estimateurs sans biais de* β_1 *et* β_2.

Preuve :
Partons de l'écriture (1.3) pour $\widehat{\beta}_2$:

$$\widehat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}.$$

Dans cette expression, seuls les bruits ε_i sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que $\mathbb{E}[\widehat{\beta}_2] = \beta_2$. Pour $\widehat{\beta}_1$, on part de l'expression :

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x},$$

d'où l'on tire :

$$\mathbb{E}[\widehat{\beta}_1] = \mathbb{E}[\bar{y}] - \bar{x}\mathbb{E}[\widehat{\beta}_2] = \beta_1 + \bar{x}\beta_2 - \bar{x}\beta_2 = \beta_1.$$

■

On peut également exprimer lles variances et la covariance de nos estimateurs.

Théorème 4.4.2 (Variances et covariance)
Les variances des estimateurs sont :

$$Var(\widehat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) \quad et \quad Var(\widehat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2},$$

tandis que leur covariance vaut :

$$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Démonstration
On part à nouveau de l'expression de $\hat{\beta}_2$ utilisée dans la preuve du non-biais :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2},$$

or les erreurs ε_i sont décorrélées et de même variance σ^2 donc la variance de la somme est la somme des variances :

$$Var(\hat{\beta}_2) = \frac{\sum(x_i - \bar{x})^2\sigma^2}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Par ailleurs, la covariance entre $\bar{y}$ et $\hat{\beta}_2$ s'écrit :

$$Cov(\bar{y}, \hat{\beta}_2) = Cov\left(\frac{\sum y_i}{n}, \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) = \frac{\sigma^2}{n}\frac{\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = 0,$$

d'où il vient pour la variance de $\hat{\beta}_1$:

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum y_i}{n} - \hat{\beta}_2\bar{x}\right) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}Cov(\bar{y}, \hat{\beta}_2),$$

c'est-à-dire :

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2}.$$

Enfin, pour la covariance des deux estimateurs :

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = Cov(\bar{y}, \hat{\beta}_2) - \bar{x}Var(\hat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}.$$

Théorème 4.4.3 (Gauss-Markov)
Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.

Démonstration
L'estimateur des MCO s'écrit

$$\hat{\beta}_2 = \sum_{i=1}^{n} p_i y_i, \quad avec\ p_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Considérons un autre estimateur $\tilde{\beta}_2$ *linéaire en* y_i *et sans biais, c'est-à-dire :*

$$\tilde{\beta}_2 = \sum_{i=1}^{n} \lambda_i y_i.$$

Montrons que $\sum \lambda_i = 0$ *et* $\sum \lambda_i x_i = 1$. *L'égalité*

$$\mathbb{E}(\tilde{\beta}_2) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i + \sum \lambda_i \mathbb{E}(\varepsilon_i) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i$$

est vraie pour tout β_2. *L'estimateur* $\tilde{\beta}_2$ *est sans biais donc* $\mathbb{E}(\tilde{\beta}_2) = \beta_2$ *pour tout* β_2, *c'est-à-dire que* $\sum \lambda_i = 0$ *et* $\sum \lambda_i x_i = 1$.

Montrons que $Var(\tilde{\beta}_2) \geq Var(\hat{\beta}_2)$ *:*

$$Var(\tilde{\beta}_2) = Var(\tilde{\beta}_2 - \hat{\beta}_2 + \hat{\beta}_2) = Var(\tilde{\beta}_2 - \hat{\beta}_2) + Var(\hat{\beta}_2) + 2Cov(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2).$$

Or :

$$Cov(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2) = Cov(\tilde{\beta}_2, \hat{\beta}_2) - Var(\hat{\beta}_2) = \frac{\sigma^2 \sum \lambda_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} - \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = 0,$$

la dernière égalité étant due aux deux relations $\sum \lambda_i = 0$ *et* $\sum \lambda_i x_i = 1$. *Ainsi :*

$$Var(\tilde{\beta}_2) = Var(\tilde{\beta}_2 - \hat{\beta}_2) + Var(\hat{\beta}_2).$$

Une variance est toujours positive, donc :

$$Var(\tilde{\beta}_2) \geq Var(\hat{\beta}_2).$$

Le résultat est démontré. On obtiendrait la même chose pour $\hat{\beta}_1$.

Prévision :

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative x. Soit donc x_{n+1} une nouvelle valeur, pour laquelle nous voulons prédire y_{n+1}. Le modèle est toujours le même :

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $\text{Var}(\varepsilon_{n+1}) = \sigma^2$ et $\text{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. Il est naturel de prédire la valeur correspondante via le modèle ajusté :

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}.$$

Deux types d'erreurs vont entacher notre prévision : la première est due à la non-connaissance de ε_{n+1}, la seconde à l'incertitude sur les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$.

Proposition 4.4.2 (Erreur de prévision)
Ł'erreur de prévision $\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1}$ *satisfait les propriétés suivantes :*

$$\begin{cases} \mathbb{E}(\hat{\varepsilon}_{n+1}) = 0 \\ Var(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right) \end{cases}$$

Preuve

Pour l'espérance, il suffit d'utiliser le fait que ε_{n+1} est centrée et que les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais :

$$\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbb{E}[\beta_1 - \hat{\beta}_1] + \mathbb{E}[\beta_2 - \hat{\beta}_2]x_{n+1} + \mathbb{E}[\varepsilon_{n+1}] = 0.$$

Nous obtenons la variance de l'erreur de prévision en nous servant du fait que y_{n+1} est fonction de ε_{n+1} seulement tandis que $\hat{y}_{n+1}$ est fonction des autres erreurs $(\varepsilon_i)_{1\leq i\leq n}$:

$$Var(\hat{\varepsilon}_{n+1}) = Var(y_{n+1} - \hat{y}_{n+1}) = Var(y_{n+1}) + Var(\hat{y}_{n+1}) = \sigma^2 + Var(\hat{y}_{n+1}).$$

Calculons le second terme :

$$\begin{aligned} Var(\hat{y}_{n+1}) &= Var\left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) = Var(\hat{\beta}_1) + x_{n+1}^2 Var(\hat{\beta}_2) + 2x_{n+1}Cov(\hat{\beta}_1, \hat{\beta}_2) \\ &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\left(\frac{\sum x_i^2}{n} + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\left(\frac{(\sum(x_i - \bar{x})^2)}{n} + \bar{x}^2 + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right). \end{aligned}$$

Au total, on obtient bien :

$$Var(\hat{\varepsilon}_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Ainsi la variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Autrement dit, faire de la prévision lorsque x_{n+1} est "loin" de $\bar{x}$ est périlleux, puisque la variance de l'erreur de prévision est alors très grande ! Ceci s'explique naturellement par le fait que plus une observation x_{n+1} est éloignée de la moyenne $\bar{x}$ et moins on a d'information sur elle.

Le coefficient de détermination R^2

Définition 4.4.3 ((Coefficient de détermination R^2))

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y} - \bar{y}.\mathbf{1}\|^2}{\|Y - \bar{y}.\mathbf{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}.\mathbf{1}\|^2} = 1 - \frac{SCR}{SCT}.$$

où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle). Ceci peut se voir comme une formule typique de décomposition de la variance. Elle permet en outre d'introduire le coefficient de détermination de façon naturelle.

On peut différencier les cas suivants :

- *Si $R^2 = 1$, Les points de l'échantillon sont parfaitement alignés sur la droite des moindres carrés ;*
- *Si $R^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i. Le modèle de régression linéaire est inadapté puisqu'on ne modélise rien de mieux que la moyenne ;*
- *Si R^2 est proche de zéro, le modèle de régression linéaire est inadapté, la variable x n'explique pas bien la variable réponse y (du moins pas de façon affine).*

De façon générale, l'interprétation est la suivante : le modèle de régression linéaire permet d'expliquer $100 \times R^2\%$ de la variance totale des données.

Remarque 4.4.2

On peut aussi voir R^2 comme le carré du coefficient de corrélation empirique entre les x_i et les y_i :

$$R^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right)^2 = r_{X,Y}^2.$$

Application : Prédiction des émissions de gaz à effet de serre dans l'UE jusqu'à 2030 en utilisant la regression linéaire :

Dans cette application, nous avons développé un code Python permettant de construire une droite de régression linéaire représentant l'évolution des émissions de gaz à effet de serre dans l'Union Européenne.

Les données utilisées couvrent la période allant de l'an 2000 à 2023. Le modèle ainsi ajusté permet de projeter l'évolution des émissions jusqu'en 2030, en prolongeant la tendance linéaire observée.

La figure ci-dessus illustre la droite de régression obtenue à partir des données historiques, ainsi que la prévision pour les années futures. Cette visualisation permet de mieux apprécier la tendance générale des émissions et d'anticiper leur évolution si les conditions actuelles se maintiennent.

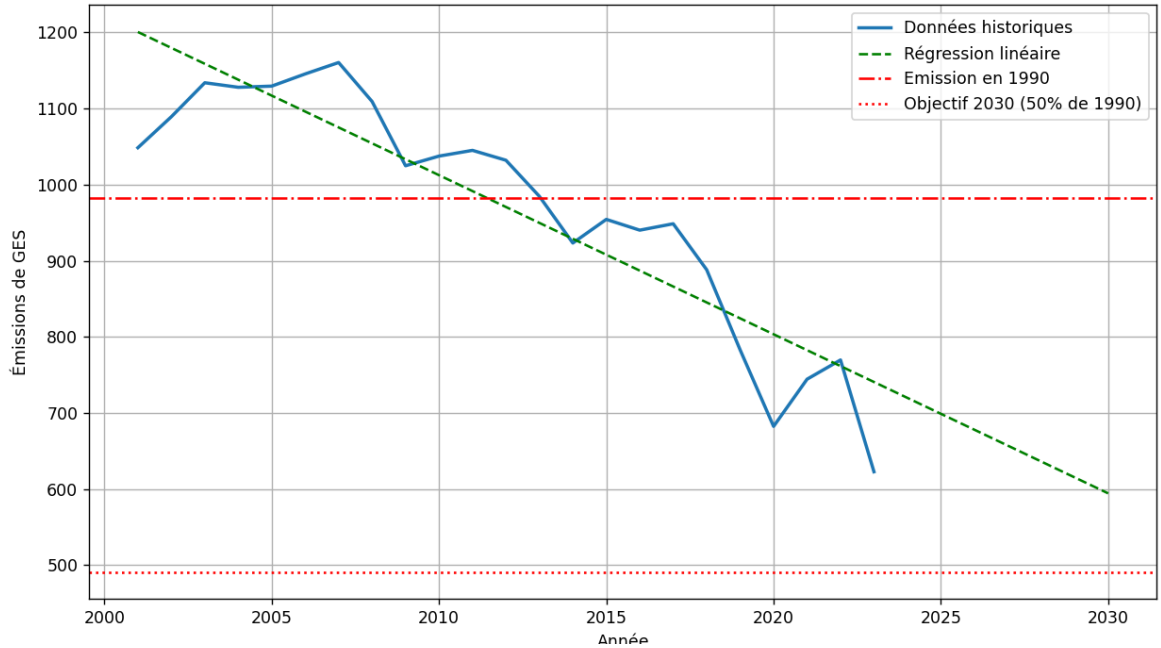

FIGURE 4.5 – Modèle de regression lineaire pour prédire les émissions GES ($MtCO_2$) dans lÚE avec données d'apprentissage, les données hisorique 2000-2023

La figure ci-dessus illustre l'évolution des émissions de gaz à effet de serre (GES) dans l'Union Européenne depuis l'an 2000 jusqu'en 2023, ainsi que la projection linéaire jusqu'en 2030.

- La courbe bleue représente les données historiques observées. On y constate une tendance générale à la baisse, bien que certaines fluctuations soient visibles d'une année à l'autre.

- La droite verte en pointillés correspond à la régression linéaire ajustée sur la période 2000–2023. Elle permet de modéliser cette tendance et de faire une prédiction pour l'année 2030.

- La ligne rouge en pointillés-traits indique le niveau d'émissions en 1990, utilisé comme référence dans les objectifs climatiques de l'UE.

- La ligne rouge pointillée marque l'objectif 2030, correspondant à une réduction de 50 % des émissions par rapport au niveau de 1990.

Ce premier modèle, basé sur toutes les valeurs historiques depuis 2000, prédit que l'Union Européenne n'atteindra pas son objectif en 2030. Toutefois, il ne prend pas en considération les accords et politiques climatiques majeurs signés au cours des dernières années, qui ont eu un impact significatif sur la dynamique des émissions.

On observe en effet une baisse marquée des émissions à partir de 2015, probablement liée à l'intensification des efforts de transition énergétique. Pour refléter cette évolution récente, un second modèle linéaire a été ajusté en ne prenant en compte que les données de 2015 à 2023.

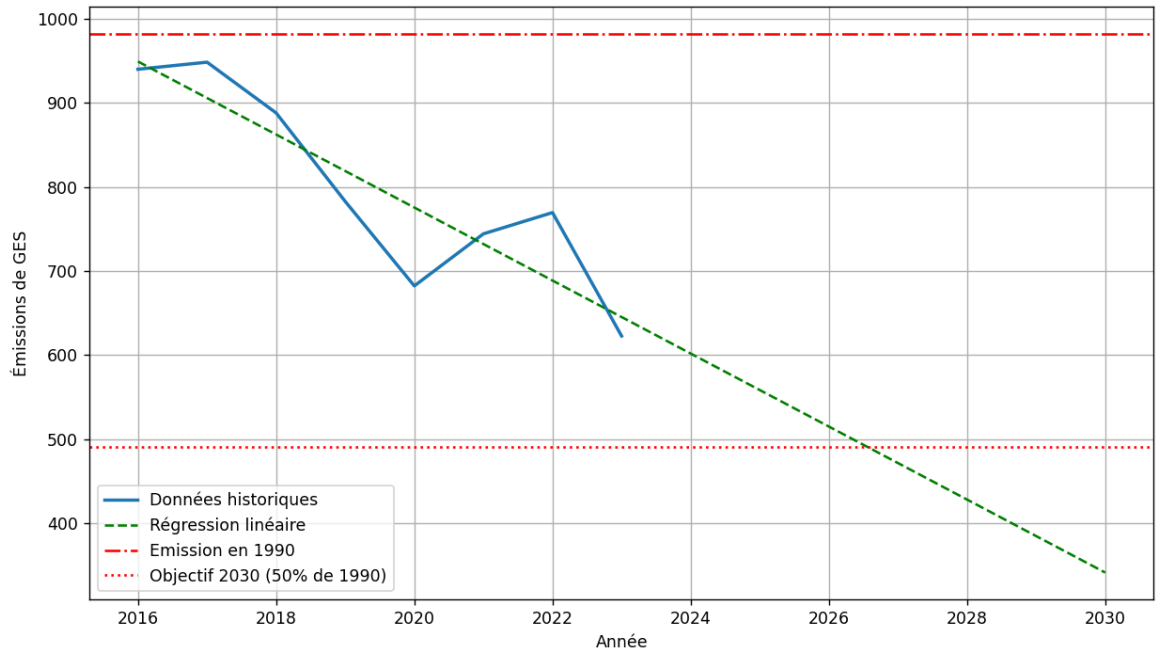

FIGURE 4.6 – Modèle de regression lineaire pour prédire les émissions GES ($MtCO_2$) dans lÚE avec données d'apprentissage, les données hisorique 2015-2023

La figure ci-dessus montre ce second ajustement : la pente de la régression est plus prononcée, ce qui confirme que, si la tendance post-2015 se maintient, l'UE pourrait effectivement atteindre son objectif climatique de 2030.

Enfin, il est important de noter que la prévision reste sensible aux aléas économiques, politiques et technologiques des prochaines années. Ces résultats doivent donc être interprétés comme des scénarios indicatifs plutôt que comme des certitudes.

Remarque 4.4.3
*Le **coefficient de détermination** R^2 permet d'évaluer la qualité d'ajustement du modèle linéaire. Pour le modèle basé sur les données de 2000 à 2023, on obtient $R^2 = 0{,}7697$, ce qui indique que 76,97 % de la variance des émissions est expliquée par la régression. En revanche, pour le modèle basé uniquement sur les données de 2015 à 2023, on obtient un coefficient plus élevé : $R^2 = 0{,}8289$, montrant une meilleure qualité d'ajustement. Cela confirme la pertinence d'un modèle concentré sur la période post-2015 pour capturer la dynamique récente.*

4.4.3 Introduction aux séries temporelles

L'analyse des séries temporelles occupe une place importante dans de nombreux domaines d'application tels que l'économie, la finance, la météorologie ou encore l'ingénierie. Une série temporelle est une suite de données chronologiques mesurées à intervalles réguliers, dont l'objectif est souvent de comprendre le comportement passé d'un phénomène afin d'en prévoir l'évolution future.

Avant toute modélisation, il est essentiel de bien comprendre la structure d'une série temporelle. De manière générale, une série peut être décomposée en plusieurs composantes caractéristiques :

— **La tendance** : elle représente l'évolution globale à long terme de la série. Par exemple, la consommation des ménages ou le PIB présentent une tendance croissante au fil du temps.
— **La saisonnalité** : ce sont des fluctuations périodiques qui reviennent à intervalles réguliers, souvent liées à des effets calendaires ou climatiques (comme les ventes de champagne en fin d'année).
— **Les fluctuations irrégulières** : ce sont des variations aléatoires et imprévisibles, souvent assimilées au bruit.

Ces composantes peuvent être combinées de manière additive ou multiplicative selon la nature de la série étudiée. L'observation des graphiques permet souvent de détecter visuellement la présence de ces différentes structures, ce qui guide le choix du modèle approprié pour l'analyse et la prévision.

4.4.4 Types de décomposition

Une fois les composantes d'une série temporelle identifiées, il est courant de représenter la série comme une combinaison de ces éléments. Deux formes principales de décomposition sont utilisées :

— **Modèle additif** :

$$y_t = f_t + s_t + e_t$$

Ce modèle suppose que la tendance (f_t), la saisonnalité (s_t) et les fluctuations irrégulières (e_t) s'ajoutent indépendamment. Il est adapté lorsque l'amplitude des variations saisonnières reste constante au cours du temps.

— **Modèle multiplicatif** :

$$y_t = f_t \times (1 + s_t) \times (1 + e_t)$$

Dans ce cas, les composantes interagissent proportionnellement. Ce modèle est pertinent lorsque l'amplitude des variations saisonnières dépend du niveau de la série (ex. : plus la tendance augmente, plus les oscillations saisonnières sont grandes).

Le choix entre ces deux modèles dépend des caractéristiques observées dans les données. En pratique, on peut souvent rendre un modèle multiplicatif additif en appliquant une transformation logarithmique à la série.

4.4.5 Notion de stationnarité

La stationnarité est une propriété fondamentale en analyse des séries temporelles. Une série est dite *stationnaire au sens large* si ses propriétés statistiques (moyenne, variance, autocorrélation, etc.) ne dépendent pas du temps.

Plus rigoureusement, une série $\{X_t\}$ est *stationnaire au sens strict* si, pour tout entier k et tout ensemble d'indices $t_1, \ldots, t_k$, la distribution conjointe de $(X_{t_1}, \ldots, X_{t_k})$ est identique à celle de $(X_{t_1+h}, \ldots, X_{t_k+h})$ pour tout décalage h.

Dans la pratique, on se contente souvent de la *stationnarité au second ordre*, qui impose deux conditions :

— La moyenne $\mathbb{E}[X_t] = \mu$ est constante dans le temps.

— La fonction d'autocovariance $\gamma(h) = \mathrm{Cov}(X_t, X_{t+h})$ dépend uniquement du décalage h et non de t.

De nombreuses séries économiques ou environnementales ne sont pas stationnaires : elles présentent des tendances, des ruptures ou des variations de variance. Pour modéliser ces séries, il est souvent nécessaire de les transformer en séries stationnaires.

La méthode la plus courante consiste à appliquer la **différenciation** d'ordre d, qui consiste à calculer des écarts successifs :

$$Y_t = X_t - X_{t-1} \quad \text{ou plus généralement} \quad \nabla^d X_t = (1 - B)^d X_t$$

où B est l'opérateur de retard tel que $BX_t = X_{t-1}$.

La série transformée $\nabla^d X_t$ est alors plus susceptible de satisfaire les conditions de stationnarité, ce qui permet d'appliquer des modèles statistiques appropriés.

4.4.6 Le bruit blanc

L'exemple le plus simple de modèle stochastique est le bruit blanc discret, la structure « revêtue » des résidus qui restent après qu'on enlève la tendance ou la moyenne d'un processus.

Définition 4.4.4
Un processus ε_t, $t \in \mathbb{T}$, où $\mathbb{T}$ est un ensemble dénombrable quelconque, est appelé bruit blanc stationnaire *si les variables ε_t sont i.i.d. (indépendantes et identiquement distribuées) avec une espérance $\mathbb{E}[\varepsilon_t] = 0$. Il est appelé* bruit blanc Gaussien *si la distribution de chaque variable aléatoire ε_t est gaussienne.*

Un bruit blanc a une fonction de covariance :

$$\gamma(s,t) = \mathbb{E}[\varepsilon_s \varepsilon_t] = 0, \quad \forall s \neq t$$

et donc un coefficient de corrélation :

$$\rho(s,t) = \frac{\gamma(s,t)}{\sigma_s \sigma_t} = \delta(s-t)$$

où $\delta(s-t)$ est le symbole de Kronecker.

Définition 4.4.5
Un processus ε_t, $t \in \mathbb{T}$, est appelé bruit blanc de second ordre *s'il a une moyenne nulle, une variance constante $\mathbb{E}[\varepsilon_t^2] = \sigma^2$, et une covariance nulle pour $s \neq t$, soit $\gamma(s,t) = \mathbb{E}[\varepsilon_s \varepsilon_t] = 0$.*

Remarques

1. Le bruit blanc gaussien est une hypothèse probabiliste très naturelle. La distribution gaussienne possède plusieurs propriétés importantes, notamment son invariance par rotation, ce qui en fait un bon modèle de bruit aléatoire.
2. Le bruit blanc stationnaire est une idéalisation utilisée pour modéliser les résidus d'une régression linéaire. Même si on suppose souvent qu'ils sont indépendants, il est difficile de le vérifier rigoureusement. Des tests empiriques, comme le test des turning points ou des corrélations empiriques, peuvent indiquer si l'indépendance est plausible.
3. Si les données ne suivent pas un bruit blanc, notamment en présence de corrélation dans les résidus, il faut adopter des approches plus complexes (par exemple : le krigeage en géostatistique).

4.4.7 Les modèles MA(q) : le cas des processus à composantes dépendantes

On considère ici le modèle de moyenne mobile d'ordre q, ou MA(q), comme un exemple simple de processus présentant une dépendance temporelle.

Définition 4.4.6
Un processus Y_t est appelé linéaire *en ε_t s'il peut être représenté dans la forme :*

$$Y_t = \sum_{i=-\infty}^{\infty} \psi_i \varepsilon_{t-i}$$

où ε_t est un bruit blanc.

Définition 4.4.7
Un processus linéaire Y_t est appelé causal *s'il peut être représenté dans la forme :*

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

où ε_t est un bruit blanc et $\sum \psi_i^2 < \infty$.

Définition 4.4.8
On appelle processus MA(q) un processus linéaire Z_t, $t \in \mathbb{Z}$ vérifiant une relation :

$$Z_t = \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \quad et \quad \theta_0 = 1$$

En utilisant la notation des polynômes de retard, cette équation s'écrit :

$$Z_t = \theta(B)\varepsilon_t$$

Théorème 4.4.4
Un processus linéaire

$$Y_t = \sum_{i=-\infty}^{\infty} \psi_i \varepsilon_{t-i}$$

avec $\sum \psi_i^2 < \infty$ *est :*
— *de variance constante stationnaire :* $\mathrm{Var}(Y_t) = \sigma_\varepsilon^2 \sum_{i=-\infty}^{\infty} \psi_i^2$
— *d'autocovariance donnée par :*

$$\gamma(t, t+k) = \sigma_\varepsilon^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+k}$$

Remarque : la stationnarité ici est d'ordre deux.

Les fonctions de covariance ou de corrélation $\gamma(k)$ et $\rho(k)$ d'un processus MA(q) s'annulent pour $k > q$, ce qui permet d'identifier empiriquement une structure de type MA(q).

Plus précisément, pour accepter l'hypothèse que la série est MA(q) pour un q donné, on vérifie que toutes les corrélations pour $k > q$ satisfont :

$$|\hat{\rho}_k| \leq z_\alpha \hat{\sigma}_k$$

où

$$\hat{\sigma}_k^2 = \frac{1 + 2(\hat{\rho}(1)^2 + \hat{\rho}(2)^2 + \cdots + \hat{\rho}(q)^2)}{n}$$

(formule de Bartlett) et z_α est la fractile de la loi normale standard associée au niveau de confiance souhaité (par exemple, $z_{0.05} \approx 2$).

4.4.8 Les modèles AR(p)

On appelle **modèle autorégressif d'ordre** p, ou **AR(p)**, un processus aléatoire $\{Y_t\}$ défini par l'équation :

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varepsilon_t$$

où ε_t est un bruit blanc.

Ce type de modèle décrit une variable Y_t comme une combinaison linéaire de ses propres valeurs passées et d'un terme aléatoire ε_t.

En utilisant la notation de l'opérateur de retard B, on peut écrire :

$$Y_t = \varphi(B) Y_t + \varepsilon_t, \quad \text{avec} \quad \varphi(B) = 1 - \sum_{i=1}^{p} \varphi_i B^i$$

Définition 4.4.9
Un processus AR(p) est dit stationnaire *s'il existe une solution stationnaire à l'équation ci-dessus.*

Cas particulier : AR(1) Considérons le cas particulier :

$$Y_t = \varphi Y_{t-1} + \varepsilon_t$$

Ce modèle admet une solution stationnaire si et seulement si $|\varphi| < 1$.

Dans ce cas, la solution peut être exprimée comme une série infinie convergente :

$$Y_t = \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j}$$

ce qui montre que Y_t est une combinaison linéaire des valeurs passées du bruit. La série est donc dite *causale* : elle dépend uniquement du passé.

Théorème 4.4.5
Le processus AR(1) défini par $Y_t = \varphi Y_{t-1} + \varepsilon_t$ *admet une représentation causale si et seulement si* $|\varphi| < 1$.

Remarque importante : Lorsque $|\varphi| = 1$, on parle de *marche aléatoire*. Dans ce cas, la série $\{Y_t\}$ n'est pas stationnaire. Toutefois, les **incréments** $\Delta Y_t = Y_t - Y_{t-1} = \varepsilon_t$ sont stationnaires. Cette propriété est fondamentale pour la suite de l'analyse des séries non stationnaires.

Causalité des modèles AR(p)

Nous avons vu qu'il y a des problèmes d'existence de solutions stationnaires et de non-causalité avec le modèle AR(1), lorsque la racine $\lambda = \varphi^{-1}$ de son polynôme $\varphi(z) = 1 - \varphi z$ n'est pas à l'extérieur du cercle unitaire.

La même situation se produit pour les modèles AR(p).

Théorème 4.4.6
Un processus AR(p) est causal, c'est-à-dire qu'il peut être représenté sous la forme :

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \quad avec \quad \sum \psi_i^2 < \infty$$

si et seulement si toutes les racines du polynôme caractéristique $\varphi(z)$ sont à l'extérieur du cercle unitaire. Les coefficients ψ_i sont alors ceux de la série de Taylor de $\psi(z) = \frac{1}{\varphi(z)}$.

Définition 4.4.10
La fonction $\psi(z) = \frac{1}{\varphi(z)}$, intervenant dans la représentation

$$Y_t = \psi(D)\varepsilon_t$$

*est appelée **fonction de transfert** du modèle AR(p).*

Remarque 4.4.4
La fonction de transfert permet de représenter le modèle AR(p) comme un modèle MA(∞).

On peut calculer les coefficients de la fonction de transfert par plusieurs méthodes. La méthode la plus simple consiste à identifier les coefficients dans :

$$\psi(z)\varphi(z) = 1$$

Cette méthode est utilisée, entre autres, dans les approximations de Padé.

On en déduit une récurrence fondamentale qui apparaîtra également dans l'étude des modèles ARMA.

4.4.9 Causalité des modèles AR(p)

Nous avons vu qu'il y a des problèmes d'existence de solutions stationnaires et de non-causalité avec le modèle AR(1), lorsque la racine $\lambda = \varphi^{-1}$ de son polynôme $\varphi(z) = 1 - \varphi z$ n'est pas à l'extérieur du cercle unitaire.

La même situation se produit pour les modèles AR(p).

Théorème 4.4.7
Un processus AR(p) est causal, c'est-à-dire qu'il peut être représenté sous la forme :

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \quad avec \quad \sum \psi_i^2 < \infty$$

si et seulement si toutes les racines du polynôme caractéristique $\varphi(z)$ sont à l'extérieur du cercle unitaire. Les coefficients ψ_i sont alors ceux de la série de Taylor de $\psi(z) = \frac{1}{\varphi(z)}$.

Définition 4.4.11
La fonction $\psi(z) = \frac{1}{\varphi(z)}$*, intervenant dans la représentation*

$$Y_t = \psi(D)\varepsilon_t$$

est appelée ***fonction de transfert*** *du modèle AR(p).*

Remarque 4.4.5
La fonction de transfert permet de représenter le modèle AR(p) comme un modèle MA(∞*).*

On peut calculer les coefficients de la fonction de transfert par plusieurs méthodes. La méthode la plus simple consiste à identifier les coefficients dans :

$$\psi(z)\varphi(z) = 1$$

Cette méthode est utilisée, entre autres, dans les approximations de Padé.

On en déduit une récurrence

4.4.10 Récurrence de Yule-Walker

On cherche à déterminer les coefficients ψ_i de la fonction de transfert $\psi(z) = \frac{1}{\varphi(z)}$.

Par définition, on a :

$$\psi(z)\varphi(z) = 1$$

En identifiant les coefficients dans le produit de séries formelles :

$$(1 - \varphi_1 z - \varphi_2 z^2 - \cdots - \varphi_p z^p)(1 + \psi_1 z + \psi_2 z^2 + \cdots) = 1$$

on obtient la relation de récurrence suivante :

$$\psi_k = \sum_{i=1}^{\min(k,p)} \varphi_i \psi_{k-i}$$

Exemple : cas $p = 2$ Supposons que le polynôme caractéristique soit $\varphi(z) = 1 - \varphi_1 z - \varphi_2 z^2$.

Alors :

$$\psi_1 = \varphi_1, \quad \psi_2 = \varphi_1\psi_1 + \varphi_2, \quad \psi_3 = \varphi_1\psi_2 + \varphi_2\psi_1, \quad \text{etc.}$$

Ce système de récurrence permet de calculer tous les ψ_k de manière récursive. On en déduit que la suite $\{\psi_k\}$ satisfait une relation de type :

$$\psi_k = \varphi_1\psi_{k-1} + \varphi_2\psi_{k-2}$$

Ce qui est équivalent à dire que les coefficients ψ_k obéissent à la même équation caractéristique que celle du modèle AR(2).

Ainsi, si λ_1 et λ_2 sont les racines de $\varphi(z)$, la solution générale est :

$$\psi_k = A\lambda_1^k + B\lambda_2^k$$

où A et B sont des constantes déterminées par les conditions initiales.

Cette représentation joue un rôle fondamental dans les prévisions linéaires et l'analyse fréquentielle.

4.4.11 Inversibilité des modèles MA(q)

L'inversibilité est une propriété symétrique de la causalité. Elle concerne cette fois la possibilité d'exprimer le bruit ε_t comme une combinaison linéaire des valeurs passées du processus Y_t.

Considérons un processus MA(q) de la forme :

$$Y_t = \theta(B)\varepsilon_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\,\varepsilon_t$$

On cherche à inverser cette relation afin d'écrire :

$$\varepsilon_t = \sum_{i=0}^{\infty} \pi_i Y_{t-i}$$

c'est-à-dire :

$$\varepsilon_t = \pi(B) Y_t \quad \text{où} \quad \pi(B) = \frac{1}{\theta(B)}$$

Définition 4.4.12
Un processus MA(q) est dit ***inversible*** *si la série* $\pi(B)$ *est absolument convergente, c'est-à-dire si* $\sum_{i=0}^{\infty} |\pi_i| < \infty$.

definition

Théorème 4.4.8
Un processus MA(q) est inversible si et seulement si toutes les racines du polynôme $\theta(z)$ *sont situées à l'extérieur du cercle unité.*

Cette condition garantit l'unicité de la représentation MA d'un processus. Dans le cas contraire, plusieurs séries MA(q) peuvent générer la même série observée, ce qui pose problème pour l'interprétation et l'estimation des modèles.

La condition d'inversibilité permet donc de s'assurer que le bruit blanc ε_t peut être reconstitué de manière unique à partir des observations $\{Y_t\}$.

Comme pour la causalité, la fonction inverse $\pi(B)$ peut être calculée par identification dans le produit :

$$\pi(B)\theta(B) = 1$$

4.4.12 Équations de Yule-Walker

Considérons un processus AR(p) défini par :

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$$

où ε_t est un bruit blanc centré de variance σ^2.

On suppose que le processus est stationnaire d'ordre 2. On cherche alors à établir une relation entre les coefficients φ_i et la fonction d'autocovariance $\gamma(k)$.

En utilisant l'opérateur de retard B, le modèle s'écrit :

$$\varphi(B) Y_t = \varepsilon_t, \quad \text{avec} \quad \varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p$$

Par stationnarité, la covariance croisée entre Y_t et Y_{t-k} ne dépend que de k. En multipliant chaque membre de l'équation par Y_{t-k} et en prenant l'espérance, on obtient le système d'équations :

$$\gamma(k) = \varphi_1 \gamma(k-1) + \varphi_2 \gamma(k-2) + \cdots + \varphi_p \gamma(k-p), \quad \forall k \geq 1$$

Ce système est appelé système des **équations de Yule-Walker**.

En particulier, en prenant $k = 1, 2, \ldots, p$, on obtient un système linéaire d'ordre p de la forme matricielle suivante :

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}$$

Soit en notation compacte :

$$R_p \varphi = r$$

où R_p est la matrice de Toeplitz formée des autocovariances, φ est le vecteur des coefficients AR et r est le vecteur des autocovariances décalées.

Une fois les autocovariances estimées empiriquement, on peut résoudre ce système pour obtenir les coefficients du modèle AR.

Dans la pratique, on utilise souvent les **autocorrélations** $\rho(k) = \gamma(k)/\gamma(0)$, ce qui donne le système :

$$R_p^{(\rho)} \varphi = \rho$$

où $R_p^{(\rho)}$ est la matrice formée des autocorrélations, et ρ le vecteur des autocorrélations décalées.

Ces équations jouent un rôle central dans l'estimation des modèles AR via la méthode des moindres carrés ou la méthode des moments.

4.4.13 Modèle ARMA(p,q)

Définition 4.4.13
*Un processus **ARMA(p,q)** est un processus stationnaire (Y_t) qui vérifie la relation :*

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \quad \forall t \in \mathbb{Z}$$

avec ε_t un bruit blanc de variance σ^2, et φ_i, θ_j des réels.

En utilisant les opérateurs de retard, cette relation s'écrit :

$$\varphi(B) Y_t = \theta(B) \varepsilon_t$$

où :

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p$$
$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

Représentations équivalentes. Si $\varphi(B)$ et $\theta(B)$ ont toutes leurs racines strictement à l'extérieur du cercle unité, alors :

— le processus ARMA(p,q) admet une représentation MA(∞) :

$$Y_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}, \quad \psi(B) = \frac{\theta(B)}{\varphi(B)}$$

— il admet aussi une représentation AR(∞) :

$$\varepsilon_t = \sum_{k=0}^{\infty} \pi_k Y_{t-k}, \quad \pi(B) = \frac{\varphi(B)}{\theta(B)}$$

Ces deux représentations sont utiles pour la prévision et l'interprétation du processus.

Théorème 4.4.9
Un processus ARMA(p,q) est causal et inversible si et seulement si toutes les racines de $\varphi(z)$ et $\theta(z)$ sont à l'extérieur du cercle unité.

Dans ce cas, les espaces vectoriels engendrés par $\{Y_{t-i}\}$ et $\{\varepsilon_{t-i}\}$ sont identiques :

$$\text{span}\{Y_{t-i}, i = 0, 1, ...\} = \text{span}\{\varepsilon_{t-i}, i = 0, 1, ...\}$$

Démonstration :

a) Si $\varphi(z)$ a ses racines à l'extérieur du cercle unité, alors on peut écrire

$$\psi(B) = \frac{\theta(B)}{\varphi(B)} = \sum_{i=0}^{\infty} \psi_i B^i$$

et donc $Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$, soit Y_t est causal.

b) Si $\theta(z)$ a ses racines à l'extérieur du cercle unité, alors on peut écrire

$$\pi(B) = \frac{\varphi(B)}{\theta(B)} = \sum_{i=0}^{\infty} \pi_i B^i$$

et donc $\varepsilon_t = \sum_{i=0}^{\infty} \pi_i Y_{t-i}$, soit ε_t est causal.

4.4.14 Prévision linéaire avec les modèles ARMA(p,q)

Pour une série (X_t) satisfaisant un modèle ARMA(p,q), la prévision linéaire d'ordre k de X_{t+k} au temps t est notée :

$$\hat{X}_t(k) = \mathbb{E}[X_{t+k} \mid \mathcal{F}_t]$$

où $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \dots)$ est l'information disponible au temps t.

Quand le processus est stationnaire et que ε_t est un bruit blanc Gaussien, cette prévision est la meilleure possible au sens quadratique.

Pour les modèles AR(p), la prévision s'obtient à partir de la relation de Yule-Walker :

$$\varphi(B)X_t = \varepsilon_t \quad \Rightarrow \quad \hat{X}_t(k) = \sum_{j=1}^{p} \varphi_j \hat{X}_t(k-j)$$

La solution se construit itérativement à partir des valeurs initiales.

Pour un AR(2), par exemple :

$$\hat{X}_t(1) = \varphi_1 X_t + \varphi_2 X_{t-1}, \quad \hat{X}_t(2) = \varphi_1 \hat{X}_t(1) + \varphi_2 X_t, \quad \text{etc.}$$

Remarque 4.4.6
Dans le cas causal ($|\lambda_i| < 1$), la prévision $\hat{X}_t(k)$ converge vers 0 quand $k \to \infty$.

Théorème 4.4.10
(Génératrice des prévisions) *Soit $X(z) = \sum_{k=0}^{\infty} z^k \hat{X}_t(k)$ la série génératrice des prévisions d'un processus AR(p), alors :*

$$X(z) = \frac{X_t + z(\varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p})}{1 - \varphi(z)}$$

Exemple 4.4.1
Considérons le modèle AR(2) suivant :

$$(1 - 0.7B)(1 + 0.5B)X_t = \varepsilon_t$$

On a :

$$\hat{X}_t(1) = (\varphi_1 + \varphi_2)X_t, \quad \hat{X}_t(2) = (\varphi_1^2 + \varphi_2)X_t + \varphi_1\varphi_2 X_{t-1}, \text{etc.}$$

Toutes ces prévisions sont obtenues à partir de la récurrence de Yule-Walker en remplaçant les valeurs initiales.

Remarque 4.4.7
Dans le cas causal $|\lambda| < 1$, on a toujours $\lim_{k\to\infty} \hat{X}_t(k) = 0$.

4.4.15 Les modèles ARIMA (p,d,q)

Définition 4.4.14
*On appelle **processus ARIMA**(p,d,q) un **processus non stationnaire** (X_t) pour lequel le processus différencié d'ordre d,*

$$Y_t = (1-B)^d X_t, \quad t \in \mathbb{Z}$$

est stationnaire, et vérifie une relation de récurrence ARMA(p,q) :

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}, \quad \forall t \in \mathbb{Z}$$

La notation des polynômes de retard ramène cette équation à la forme :

$$\varphi(B)(1-B)^d X_t = \varphi(B) Y_t = \theta(B)\varepsilon_t,$$

où $\varphi(B), \theta(B)$ sont des polynômes relativement premiers dans l'opérateur de retard B, à ordres p, q, avec coefficient libre 1, et avec racines dehors le cercle unitaire.

Remarque importante. Le processus X_t n'est en général **pas stationnaire**, mais si on applique l'opérateur de différenciation d'ordre d, on obtient un processus $Y_t = (1-B)^d X_t$ qui est stationnaire et suit un modèle ARMA(p,q) :

$$\varphi(B) Y_t = \theta(B)\varepsilon_t$$

Tous les résultats classiques établis pour les processus ARMA(p,q) peuvent donc être appliqués à Y_t, et par conséquent indirectement à X_t.

Forme explicite. On peut réécrire le modèle ARIMA(p,d,q) sous la forme :

$$\varphi(B)(1-B)^d X_t = \theta(B)\varepsilon_t \quad \text{ou encore} \quad Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}$$

où l'on rappelle que $Y_t = (1-B)^d X_t$.

Prévision linéaire. Le but est de construire une estimation $\widehat{X}_{t+k|t}$ de X_{t+k} à l'instant t. On procède en prédisant le processus stationnaire différencié Y_t, et en réintégrant ensuite les valeurs prédites pour retrouver X_{t+k}.

On utilise pour cela la relation de récurrence :

$$\widehat{X}_{t+k|t} = \sum_{i=1}^{p} \varphi_i \widehat{X}_{t+k-i|t} + \sum_{i=1}^{q} \theta_i \widehat{\varepsilon}_{t+k-i}$$

avec :
— $\widehat{X}_t = X_t$ pour $t \leq 0$ (valeurs initiales),
— $\widehat{\varepsilon}_t = X_t - \widehat{X}_{t|t-1}$ ou calculée par estimation,
— les φ_i et θ_i estimés par maximum de vraisemblance ou moindres carrés.

Points pivots. La prévision à horizon k utilise les $p+d$ dernières valeurs observées de X_t (pivots) :

$$X_t, X_{t-1}, \ldots, X_{t-(p+d)+1}$$

En effet, le polynôme $\varphi(B)(1-B)^d$ étant de degré $p+d$, la récurrence correspondante passe par ces derniers points.

Forme équivalente et notation opératorielle

Soit $\{X_t\}$ un processus ARIMA(p,d,q), c'est-à-dire que la série différenciée $Y_t = \nabla^d X_t = (1-B)^d X_t$ suit un modèle ARMA(p,q) :

$$\varphi(B)Y_t = \theta(B)\varepsilon_t,$$

où $\varphi(B)$ et $\theta(B)$ sont les polynômes d'autorégression et de moyenne mobile respectivement définis par :

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q,$$

et $\{\varepsilon_t\}$ est un bruit blanc gaussien de variance σ^2.

On peut donc réécrire le modèle ARIMA(p,d,q) dans sa forme complète :

$$\varphi(B)(1-B)^d X_t = \theta(B)\varepsilon_t.$$

Cette écriture met en évidence l'opérateur de différenciation $(1-B)^d$ qui transforme le processus non stationnaire $\{X_t\}$ en un processus stationnaire $\{Y_t\}$ satisfaisant un modèle ARMA(p,q).

4.4.16 Représentation causale et inversible des modèles ARIMA(p,d,q)

Considérons un processus X_t suivant un modèle ARIMA(p,d,q), soit :

$$\varphi(B)(1-B)^d X_t = \theta(B)\varepsilon_t, \tag{4.1}$$

avec ε_t un bruit blanc gaussien, et $\varphi(B)$, $\theta(B)$ deux polynômes de degré p et q tels que :

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

Stationnarisation par différentiation. Le processus X_t n'est pas stationnaire mais sa série différenciée $Y_t = (1-B)^d X_t$ l'est. Ainsi, Y_t suit un modèle ARMA(p,q) :

$$\varphi(B)Y_t = \theta(B)\varepsilon_t. \tag{4.2}$$

Causalité. Le processus ARMA(p,q) (et donc Y_t) est **causal** si toutes les racines du polynôme $\varphi(z)$ sont à l'extérieur du cercle unité :

$$|z| > 1 \quad \text{pour tout } z \text{ tel que } \varphi(z) = 0.$$

Dans ce cas, Y_t admet la représentation linéaire causale suivante :

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \tag{4.3}$$

avec $\psi_0 = 1$ et les ψ_i obtenus par l'expansion de la fraction rationnelle $\theta(B)/\varphi(B)$.

Inversibilité. Le processus est **inversible** si toutes les racines du polynôme $\theta(z)$ sont à l'extérieur du cercle unité :

$$|z| > 1 \quad \text{pour tout } z \text{ tel que } \theta(z) = 0.$$

Dans ce cas, le bruit blanc peut être exprimé en fonction du passé observé du processus :

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j}, \tag{4.4}$$

avec $\pi_0 = 1$ et les π_j déduits de l'expansion de $\varphi(B)/\theta(B)$.

Représentation finale. Sous hypothèses de causalité et d'inversibilité, le processus X_t admet donc la représentation :

$$(1-B)^d X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \quad \text{avec } \varepsilon_t \sim \mathcal{WN}(0, \sigma^2), \tag{4.5}$$

ce qui permet le développement théorique des méthodes de prévision et d'estimation liées au modèle ARIMA(p,d,q).

Méthode de Box-Jenkins
La méthode de Box-Jenkins est une démarche structurée en 4 étapes :

1. **Identification** : déterminer p, d, q via ACF, PACF, tests de stationnarité ;
2. **Estimation** : calcul des paramètres du modèle ARIMA(p,d,q) ;
3. **Diagnostic** : tests sur les résidus (autocorrélation, bruit blanc...) ;
4. **Prévision** : utilisation du modèle pour prédire les valeurs futures.

Remarque : C'est une méthode itérative : si le diagnostic échoue, on recommence.

4.4.17 Identification du modèle *ARIMA*(p,d,q)

Avant d'estimer les paramètres du modèle ARIMA(p,d,q), il est essentiel de déterminer les ordres p (autorégressif), d (différenciation) et q (moyenne mobile). Cette phase s'appelle **l'identification du modèle** et fait partie de la méthode de modélisation Box-Jenkins.

1. Détermination de l'ordre d (stationnarisation). Le paramètre d est choisi pour rendre la série stationnaire. On analyse pour cela :
— Le graphe de la série : une tendance visible ou une variance croissante suggère une non-stationnarité.
— Le test de racine unitaire (Dickey-Fuller augmenté ou ADF).
— Le test KPSS (stationnarité sous l'hypothèse nulle).
Si la série n'est pas stationnaire, on applique la différenciation une ou plusieurs fois jusqu'à obtenir une série stationnaire $Y_t = (1-B)^d X_t$.

2. Analyse de l'ACF et de la PACF de la série stationnarisée.
— **ACF (Autocorrelation Function)** : donne des indications sur q, l'ordre du MA.
— **PACF (Partial Autocorrelation Function)** : donne des indications sur p, l'ordre de l'AR.
On observe les coupes nettes ou les décroissances progressives selon le modèle :

Modèle	ACF	PACF
AR(p)	Décroissance progressive	Coupe après p
MA(q)	Coupe après q	Décroissance progressive
ARMA(p,q)	Décroissances des deux côtés	Décroissances des deux côtés

Exemple pratique : Identification visuelle avec une série AR(2)

Illustration : Série temporelle simulée (AR(2))

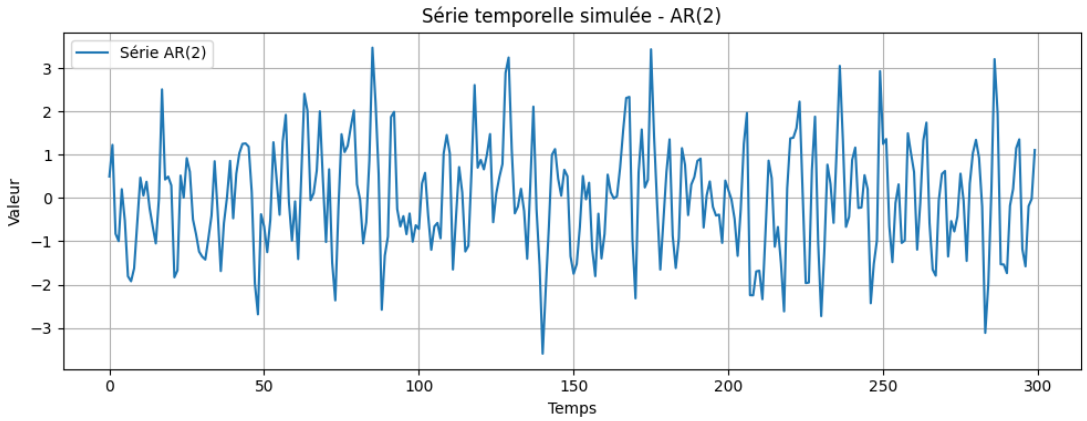

FIGURE 4.7 – Exemple de série temporelle simulée selon un modèle AR(2).

Avant d'identifier les paramètres d'un modèle ARIMA, il est utile de visualiser la série pour détecter visuellement une tendance ou des irrégularités éventuelles. Dans cette simulation AR(2), la série ne présente pas de tendance marquée, ce qui suggère une stationnarité.

Pour illustrer l'analyse de l'ACF et de la PACF, on simule une série temporelle suivant un processus AR(2). On observe ensuite ses fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF).

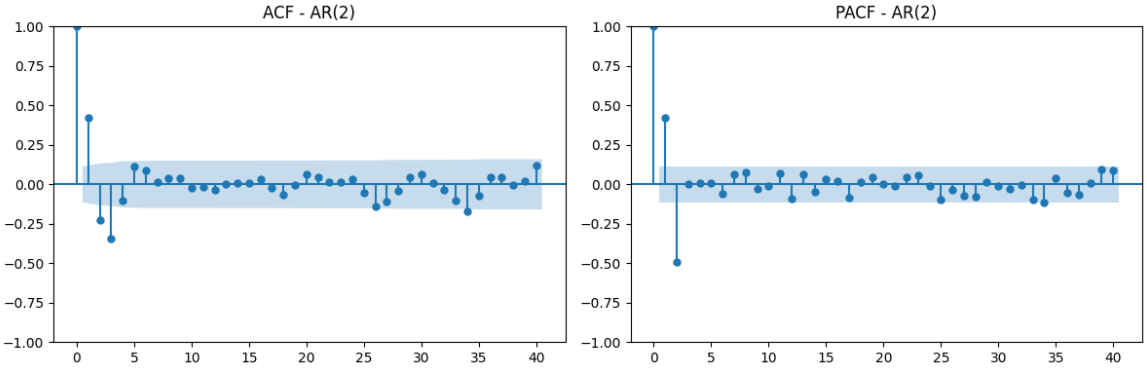

FIGURE 4.8 – Série AR(2) simulée avec ses fonctions ACF et PACF

On observe :
— Une décroissance progressive dans l'ACF (caractérique des modèles AR).
— Une coupure nette après le lag 2 dans la PACF, confirmant la structure AR(2).

Conclusion de l'exemple AR(2) La série simulée montre un comportement stationnaire sans tendance visible. L'ACF présente une décroissance progressive, tandis que la PACF s'annule brusquement après le lag 2. Cela confirme la structure autorégressive d'ordre 2 (AR(2)), conformément au modèle utilisé pour générer la série.

Application aux émissions de GES de l'UE (2000–2023) Dans notre étude, nous avons modélisé la série des **émissions de GES totales de l'Union Européenne**.

Le modèle optimal ARIMA a été automatiquement sélectionné via un algorithme de recherche par AIC :

`Best model: ARIMA(0,1,2)` $(\text{AIC} = 252.91)$

Données utilisées :
— Période : 2000–2023

— Unité : $MtCO_2e$ (mégatonnes d'équivalent CO_2)

Exemple 4.4.2
*Le modèle entraîné permet de **prédire l'évolution jusqu'en 2030** des émissions de GES, en tenant compte des tendances historiques.*

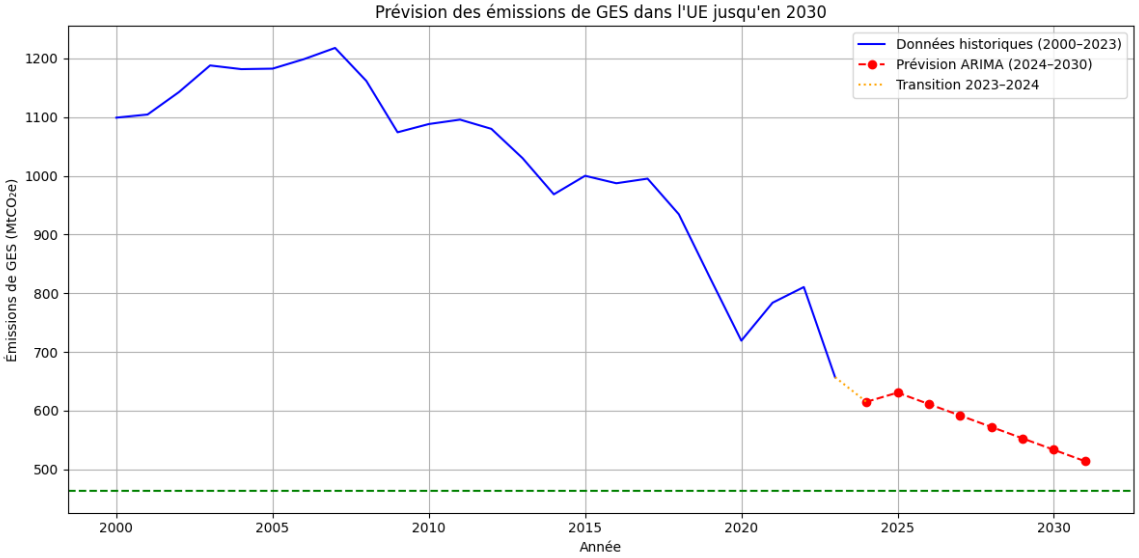

FIGURE 4.9 – Prévision des émissions de GES dans l'UE jusqu'en 2030 à l'aide du modèle ARIMA optimal

4.4.18 Estimation des paramètres

Une fois les ordres (p,d,q) identifiés, on estime les paramètres du modèle ARIMA à l'aide de la méthode du maximum de vraisemblance. Cette estimation est effectuée automatiquement par les logiciels comme `statsmodels` en Python.

— Les coefficients AR (ϕ_i) et MA (θ_j) sont ajustés pour maximiser la vraisemblance.
— La qualité du modèle est évaluée via des critères comme l'AIC (Akaike Information Criterion) ou le BIC (Bayesian Information Criterion).
Dans notre cas, l'algorithme a choisi le modèle ARIMA(0,1,2) avec un AIC de 252.91.

Remarque 4.4.8 (Choix automatique du modèle ARIMA)
*Contrairement à une approche manuelle où les paramètres (p,d,q) du modèle ARIMA sont fixés à l'avance, nous avons ici laissé le **modèle déterminer automatiquement la meilleure configuration** via un algorithme de sélection basé sur le critère d'information AIC (Akaike Information Criterion). Cela garantit que le modèle est **optimal pour les données historiques** disponibles, sans sur-ajustement.*

Remarque 4.4.9 (Interprétation de la figure de prévision)
*La figure montre une tendance continue à la baisse des émissions de GES jusqu'en 2030, ce qui est **cohérent avec les politiques de transition énergétique** engagées dans l'UE. Cependant, le modèle ARIMA reste **purement statistique** : il **n'intègre pas explicitement les ruptures structurelles ou les accélérations dues aux politiques climatiques**, comme l'interdiction des voitures thermiques ou le pacte vert européen.*

Remarque 4.4.10 (Limite de la prévision)
*L'ARIMA se base uniquement sur les valeurs passées pour extrapoler le futur. Ainsi, la prévision est **lisse** et **n'anticipe pas les changements brusques**. Cela signifie que les **ruptures liées à des innovations technologiques, à des décisions politiques ou à des crises géopolitiques** ne sont pas capturées dans la projection. Ce comportement est normal et attendu pour un modèle univarié comme ARIMA.*

4.4.19 Comparaison des approches : Régression linéaire vs ARIMA

Objectif de comparaison : Comparer la capacité de la régression linéaire et du modèle ARIMA à prédire les émissions de GES dans l'Union Européenne jusqu'en 2030, en lien avec l'objectif climatique (*–50% par rapport à 1990*).

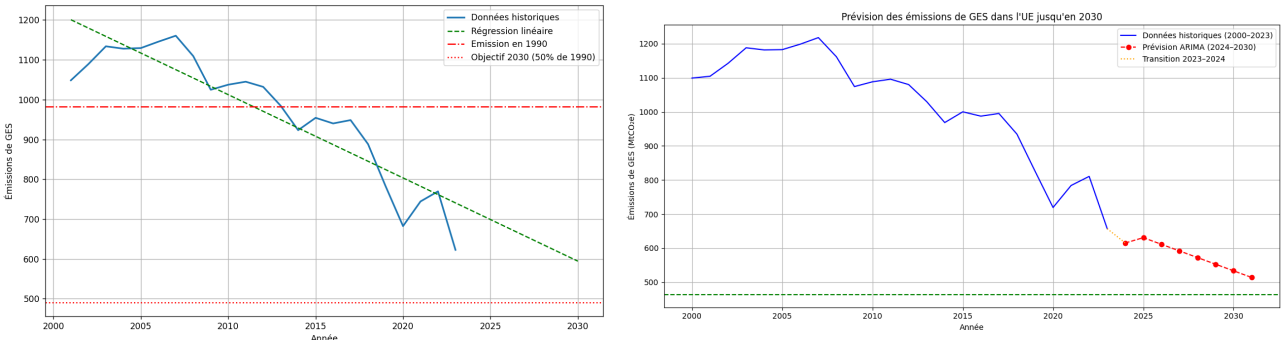

FIGURE 4.10 – Comparaison des prévisions : Régression linéaire (gauche) vs ARIMA (droite)

Analyse comparative :

- **Tendance générale** : les deux modèles prévoient une baisse continue des émissions.
- **Régression linéaire** :
  - Hypothèse d'une tendance constante.
  - Ne capture pas les chocs ou ruptures (ex. : crise de 2020, politiques récentes).
  - Prédiction rigide.
- **ARIMA** :
  - Intègre les variations internes à la série (retards, inertie).
  - Prend en compte la dynamique temporelle et les effets de mémoire.
  - Prédiction plus douce, mais potentiellement plus réaliste à court terme.

Conclusion : La régression linéaire fournit une estimation directe et interprétable mais simpliste. Le modèle ARIMA, bien que plus complexe, semble mieux capter les dynamiques passées et fournit une prévision plus crédible. Toutefois, aucun des deux ne modélise explicitement les politiques futures.

CHAPITRE 5

ANNEXE : SCRIPTS ET REQUÊTES

5.1 Extraction des données de l'Union Européenne

```python
import pandas as pd

# Charger le fichier CSV original
df = pd.read_csv("owid-energy-data.csv")

# Liste des codes ISO des pays de l'Union Européenne
eu_iso_codes = [
    'AUT', 'BEL', 'BGR', 'HRV', 'CYP', 'CZE', 'DNK', 'EST', 'FIN', 'FRA', 'DEU',
    'GRC', 'HUN', 'IRL', 'ITA', 'LVA', 'LTU', 'LUX', 'MLT', 'NLD', 'POL', 'PRT',
    'ROU', 'SVK', 'SVN', 'ESP', 'SWE'
]

# Filtrer les lignes correspondant aux pays de l'UE
df_eu = df[df["iso_code"].isin(eu_iso_codes)].copy()

# Sauvegarder les données filtrées
df_eu.to_csv("eu_energy_data.csv", index=False)

# Message de confirmation
print("Fichier 'eu_energy_data.csv' généré avec succès.")
```

Listing 5.1 – Filtrage des pays de l'Union Européenne

5.2 Traitement des valeurs manquantes

```python
import pandas as pd

# Chargement du fichier filtré pour l'UE
df_eu = pd.read_csv("eu_energy_data.csv")

# Calcul du pourcentage de valeurs manquantes
missing_percent = df_eu.isnull().mean() * 100

# Sélection des colonnes ayant au moins une valeur manquante
missing_percent = missing_percent[missing_percent > 0].sort_values(ascending=
    False)

# Affichage des résultats
```

```
print("Pourcentage de valeurs manquantes par colonnes :")
print(missing_percent)
```

Listing 5.2 – Calcul du pourcentage de valeurs manquantes par colonne

5.3 Suppression des colonnes incomplètes

```
import pandas as pd

# Chargement du fichier pré-nettoyé
df = pd.read_csv("eu_energy_data.csv")

# Calcul du taux de valeurs manquantes
missing_percent = df.isnull().mean() * 100

# Identification des colonnes à supprimer
cols_to_drop = missing_percent[missing_percent > 70].index

# Suppression des colonnes sélectionnées
df_cleaned = df.drop(columns=cols_to_drop)

# Sauvegarde du fichier nettoyé
df_cleaned.to_csv("eu_energy_data_cleaned_step1.csv", index=False)

# Résumé
print(f"{len(cols_to_drop)} colonnes supprimées.")
print(f"Dimensions du fichier final : {df_cleaned.shape[0]} lignes    {
    df_cleaned.shape[1]} colonnes")
```

Listing 5.3 – Suppression des colonnes avec plus de 70% de valeurs manquantes

5.4 Heatmap de complétude des lignes par pays et par année

```
# Ajouter une colonne avec le pourcentage de complétude de chaque ligne
import pandas as pd
import matplotlib.pyplot as plt

df["completeness_percent"] = (df.notnull().sum(axis=1) / df.shape[1] * 100).
    round(2)

# Afficher les colonnes clés avec la complétude
print(df[["country", "year", "completeness_percent"]].head())

# Créer une matrice (pivot) avec year en lignes, country en colonnes
completeness_pivot = df.pivot_table(index='year', columns='country', values='
    completeness_percent')

# Tracer la heatmap
plt.figure(figsize=(15, 10))
sns.heatmap(completeness_pivot, cmap="YlGnBu", annot=False, cbar_kws={'label': '
    % complétude'})
plt.title("Heatmap du % de complétude par année et pays")
plt.xlabel("Pays")
plt.ylabel("Année")
plt.tight_layout()
plt.show()
```

Listing 5.4 – Heatmap du pourcentage de complétude par année et par pays avant nettoyage

5.5 Matrice de complétude par pays

```python
import pandas as pd

# Charger le fichier CSV nettoyé après suppression des colonnes incomplètes
df = pd.read_csv("eu_energy_data_cleaned_step1.csv")

# Sélection des colonnes numériques uniquement
numeric_cols = df.select_dtypes(include='number').columns

# Calcul du taux de complétude (non-missing) par pays
completude_par_pays = df.groupby("country")[numeric_cols].apply(
    lambda group: group.notna().mean() * 100
)

# Arrondi pour lisibilité
completude_par_pays = completude_par_pays.round(1)

# Export au format Excel
completude_par_pays.to_excel("matrice_completude_par_pays.xlsx")

# Résumé
print("Matrice de complétude générée avec succès.")
print("Fichier enregistré : matrice_completude_par_pays.xlsx")
```

Listing 5.5 – Calcul de la complétude des colonnes numériques par pays

5.6 Visualisation de la complétude sous forme de heatmap

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Chargement de la matrice de complétude depuis un fichier Excel
df_completude = pd.read_excel("matrice_completude_par_pays.xlsx", index_col=0)

# Création de la heatmap
plt.figure(figsize=(14, 10))
sns.heatmap(
    df_completude,
    cmap="YlGnBu",
    linewidths=0.5,
    linecolor='gray',
    cbar_kws={'label': 'Taux de complétude (%)'},
    annot=False  # Passer à True pour afficher les pourcentages dans chaque cellule
)

# Personnalisation du graphique
plt.title("Matrice de complétude des données par pays et par variable", fontsize=14)
plt.xlabel("Variables", fontsize=12)
plt.ylabel("Pays", fontsize=12)
plt.xticks(rotation=45, ha="right")
plt.tight_layout()

# Affichage du graphique
plt.show()
```

Listing 5.6 – Affichage graphique de la matrice de complétude

5.7 Interpolation temporelle des colonnes numériques

```
import pandas as pd

# Charger le fichier nettoyé (colonnes fortement incomplètes déjà supprimées)
df = pd.read_csv("eu_energy_data_cleaned_step1.csv")

# Trier par pays et année pour assurer une interpolation cohérente
df = df.sort_values(by=["country", "year"])

# Appliquer l'interpolation sur chaque colonne numérique, groupe par groupe
for col in df.select_dtypes(include='number').columns:
    df[col] = df.groupby("country")[col].transform(lambda group: group.
        interpolate())

print("Interpolation réussie pour toutes les colonnes numériques par pays et par
    année.")

# Sauvegarde du résultat
df.to_csv("eu_energy_data_cleaned_step2_interpolated.csv", index=False)
print("Fichier enregistré : eu_energy_data_cleaned_step2_interpolated.csv")
```

Listing 5.7 – Interpolation des valeurs manquantes par pays et par année

5.8 Extrapolation des données manquantes en début et fin de série

```
import pandas as pd

# Charger le fichier déjà interpolé
df = pd.read_csv("eu_energy_data_cleaned_step2_interpolated.csv")

# Trier les données pour respecter l'ordre chronologique
df = df.sort_values(by=["country", "year"])

# Extrapolation bidirectionnelle (début et fin de série) par pays
for col in df.select_dtypes(include='number').columns:
    df[col] = df.groupby("country")[col].transform(
        lambda group: group.interpolate(limit_direction='both')
    )

print("Extrapolation terminée.")

# Sauvegarde du fichier extrapolé
df.to_csv("eu_energy_data_cleaned_step4_extrapolated.csv", index=False)
print("Fichier enregistré : eu_energy_data_cleaned_step4_extrapolated.csv")
```

Listing 5.8 – Extrapolation bidirectionnelle des colonnes numériques par pays

5.9 Schéma de la base de données PostgreSQL

5.9.1 Structure des tables principales

```sql
CREATE TABLE country_iso_code (
    iso_code CHAR(3) NOT NULL,
    country_name TEXT NOT NULL,
    PRIMARY KEY (iso_code)
);
```

Listing 5.9 – Structure de la table des pays

```sql
CREATE TABLE country_statistics (
    iso_code CHAR(3) NOT NULL,
    year INTEGER NOT NULL,
    population NUMERIC,
    gdp NUMERIC,
    PRIMARY KEY (iso_code, year),
    CONSTRAINT fk_iso_code FOREIGN KEY (iso_code) REFERENCES country_iso_code(
        iso_code) ON DELETE CASCADE
);
```

Listing 5.10 – Structure de la table country_statistics

```sql
CREATE TABLE energy_statistics (
    country_data_iso_code CHAR(3) NOT NULL,
    country_data_year INTEGER NOT NULL,
    carbon_intensity NUMERIC,
    electricity_demand NUMERIC,
    electricity_generation NUMERIC,
    electricity_energy_share NUMERIC,
    net_electricity_imports NUMERIC,
    net_electricity_imports_share NUMERIC,
    per_capita_electricity NUMERIC,
    energy_consumption_growth_pct NUMERIC,
    energy_consumption_growth_twh NUMERIC,
    energy_per_capita NUMERIC,
    energy_per_gdp NUMERIC,
    primary_energy_consumption NUMERIC,
    greenhouse_gas_emissions NUMERIC,
    PRIMARY KEY (country_data_iso_code, country_data_year),
    CONSTRAINT fk_country_data
        FOREIGN KEY (country_data_iso_code) REFERENCES country_iso_code(iso_code
            )
        ON DELETE CASCADE
);
```

Listing 5.11 – Structure de la table energy_statistics

5.9.2 Exemples de tables par type d'énergie

```sql
CREATE TABLE biofuel_energy_data (
    country_data_iso_code CHAR(3) NOT NULL,
    country_data_year INTEGER NOT NULL,
    biofuel_cons_change_pct NUMERIC,
    biofuel_cons_change_twh NUMERIC,
    biofuel_cons_per_capita NUMERIC,
    biofuel_consumption NUMERIC,
    biofuel_elec_per_capita NUMERIC,
    biofuel_electricity NUMERIC,
    biofuel_share_elec NUMERIC,
    biofuel_share_energy NUMERIC,
    PRIMARY KEY (country_data_iso_code, country_data_year),
```

```
    CONSTRAINT fk_country_data
        FOREIGN KEY (country_data_iso_code) REFERENCES country_iso_code(iso_code
            )
        ON DELETE CASCADE
);
```

Listing 5.12 – Exemple : biofuel_energy_data

```
CREATE TABLE solar_energy_data (
    country_data_iso_code CHAR(3) NOT NULL,
    country_data_year INTEGER NOT NULL,
    solar_cons_change_pct NUMERIC,
    solar_cons_change_twh NUMERIC,
    solar_consumption NUMERIC,
    solar_elec_per_capita NUMERIC,
    solar_electricity NUMERIC,
    solar_energy_per_capita NUMERIC,
    solar_share_elec NUMERIC,
    solar_share_energy NUMERIC,
    PRIMARY KEY (country_data_iso_code, country_data_year),
    CONSTRAINT fk_country_data
        FOREIGN KEY (country_data_iso_code) REFERENCES country_iso_code(iso_code
            )
        ON DELETE CASCADE
);
```

Listing 5.13 – Exemple : solar_energy_data

5.10 Visualisation des données par Grafana

```
    <div style="
  background-image: url('https://raw.githubusercontent.com/Jalil-El/public-files
      /main/pacte-vert-europeen-1024x683.jpg');
  background-size: cover;
  background-position: center;
  padding: 40px;
  border-radius: 15px;
  box-shadow: 0 0 15px rgba(0,0,0,0.5);
">

  <div style="
    background-color: rgba(0, 0, 0, 0.65);
    padding: 25px;
    border-radius: 12px;
    color: white;
    font-size: 16px;
    font-family: 'Segoe UI', Tahoma, Geneva, Verdana, sans-serif;
  ">

    <h2 style="
      color: #a8ff60;
      text-align: center;
      font-size: 28px;
      letter-spacing: 1px;
      font-family: 'Georgia', serif;
      margin-top: 0;
      margin-bottom: 20px;
    ">
              L Europe  face au d fi de la transition  nergtique
```

```
    </h2>

    <b style="color:#ffcc00;">       Objectif 2050 : Neutralit  carbone</b><br
       ><br>

            En juin 2021 : le Parlement europ en et le Conseil de  l UE
       adoptent la Loi europ enne sur le climat, qui rend juridiquement
       contraignant  l objectif  de neutralit  climatique  d ici  2050.<br><br>

          L'Union Europ enne vise une <span style="color:#00cfff;"><b>
       r duction de 55%</b></span> des  missions  de gaz  effet de serre d'
       ici 2030, et la <b>neutralit  climatique</b> d'ici 2050.<br><br>

          Ce tableau de bord explore l' volution  de la consommation d' nergie
        dans l'UE, la mont e des  nergies  renouvelables, et la d pendance
       persistante aux  nergies  fossiles.<br><br>

          <i>Faites d filer les graphiques ci-dessous pour voir si la
       transition est r ellement en cours.</i>

  </div>

</div>
```

Listing 5.14 – Code HTML pour la première fenêtre du dashboard (Representation du dashboard)

```
    SELECT
    (g.year_ || '-01-01')::DATE AS "time",
    SUM(g.gas_consumption) AS "Consommation de gaz (Twh)",
    SUM(c.coal_consumption) AS "Consommation de charbon (Twh)",
    SUM(o.oil_consumption) AS "Consommation de p trole (Twh)"
FROM
    gas_energy_data g
JOIN
    coal_energy_data c ON g.iso_code = c.iso_code AND g.year_ = c.year_
JOIN
    oil_energy_data o ON g.iso_code = o.iso_code AND g.year_ = o.year_
WHERE
    g.year_ >= 2000
GROUP BY
    g.year_
ORDER BY
    g.year_;
```

Listing 5.15 – Requête SQL pour construire le diagramme en barre d'évolution de la consomation d'énergies fossiles en europe

5.11 Statistiques descriptives et corrélations (schéma `stats`)

```
-- Cr ation du sch ma s'il n'existe pas
CREATE SCHEMA IF NOT EXISTS stats;

-- Table pour stocker les statistiques descriptives globales
CREATE TABLE IF NOT EXISTS stats.energy_stats_summary (
    variable TEXT PRIMARY KEY,
    moyenne NUMERIC,
    ecart_type NUMERIC,
    minimum NUMERIC,
    maximum NUMERIC
```

```
);

-- Vider la table pour insertion propre
TRUNCATE TABLE stats.energy_stats_summary;

-- Insertion des valeurs statistiques
INSERT INTO stats.energy_stats_summary (variable, moyenne, ecart_type, minimum,
    maximum)
VALUES
('  missions   GES (MtCO2e)',
     (SELECT AVG(greenhouse_gas_emissions) FROM energy_statistics),
     (SELECT STDDEV(greenhouse_gas_emissions) FROM energy_statistics),
     (SELECT MIN(greenhouse_gas_emissions) FROM energy_statistics),
     (SELECT MAX(greenhouse_gas_emissions) FROM energy_statistics)
),
('Consommation   nergie   (TWh)',
     (SELECT AVG(energy_consumption_growth_twh) FROM energy_statistics),
     (SELECT STDDEV(energy_consumption_growth_twh) FROM energy_statistics),
     (SELECT MIN(energy_consumption_growth_twh) FROM energy_statistics),
     (SELECT MAX(energy_consumption_growth_twh) FROM energy_statistics)
),
('Intensit   carbone (gCO2/kWh)',
     (SELECT AVG(carbon_intensity) FROM energy_statistics),
     (SELECT STDDEV(carbon_intensity) FROM energy_statistics),
     (SELECT MIN(carbon_intensity) FROM energy_statistics),
     (SELECT MAX(carbon_intensity) FROM energy_statistics)
);
```

Listing 5.16 – Création des tables pour les statistiques et corrélations

```
-- Table pour stocker les coefficients de corr lation
CREATE TABLE IF NOT EXISTS stats.energy_correlation_summary (
    variable_x TEXT NOT NULL,
    variable_y TEXT NOT NULL,
    correlation_value NUMERIC,
    PRIMARY KEY (variable_x, variable_y)
);

-- Nettoyer la table avant insertion
TRUNCATE TABLE stats.energy_correlation_summary;

-- Insertion des corr lations
INSERT INTO stats.energy_correlation_summary (variable_x, variable_y,
    correlation_value)
VALUES
('  missions   GES (MtCO2e)', 'Intensit   carbone (gCO2/kWh)',
     (SELECT corr(greenhouse_gas_emissions, carbon_intensity) FROM
        energy_statistics)
),
('  missions   GES (MtCO2e)', 'Consommation   nergie   (TWh)',
     (SELECT corr(greenhouse_gas_emissions, energy_consumption_growth_twh) FROM
        energy_statistics)
),
('Intensit   carbone (gCO2/kWh)', 'Consommation   nergie   (TWh)',
     (SELECT corr(carbon_intensity, energy_consumption_growth_twh) FROM
        energy_statistics)
);
```

Listing 5.17 – Création et remplissage de la table des corrélations

5.12 Détection des anomalies

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

#Charger les données
df_2000 = pd.read_csv('df_2000.csv')

# Calcul de l'IQR par ann\'ee
iqr_par_annee = df_2000.groupby('year')['emission_ges'].agg(
    Q1=lambda x: x.quantile(0.25),
    Q3=lambda x: x.quantile(0.75)
).reset_index()

# Ajout d'une colonne IQR = Q3 - Q1
iqr_par_annee['borne_inf'] = iqr_par_annee['Q1']-1.5*(iqr_par_annee['Q3'] - iqr_par_annee['Q1'])
iqr_par_annee['borne_sup'] = iqr_par_annee['Q3']+1.5*(iqr_par_annee['Q3'] - iqr_par_annee['Q1'])

# Affichage des résultats
iqr_par_annee.head()

plt.figure(figsize=(12,6))
sns.lineplot(df_2000, x='year', y='emission_ges', hue='country',marker='o', linestyle='-')
plt.plot(iqr_par_annee['year'], iqr_par_annee['borne_inf'], marker='o', linestyle='--', color='r', label='Q1-1.5*IQR')
plt.plot(iqr_par_annee['year'], iqr_par_annee['borne_sup'], marker='o', linestyle='--', color='r', label='Q1+1.5*IQR')
plt.xlabel('Année')
plt.ylabel(' missions  de gaz à effet de serre')
plt.legend(title="Pays", loc='center left', bbox_to_anchor=(1, 0.5))
plt.title(' volution  des émissions de gaz à effet de serre par année')
plt.show()
```

Listing 5.18 – Script python pour détecter les anomalies en utilisant la méthode de Tukey

5.13 Prédiction

```python

import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
import numpy as np
import psycopg2

#Charger les données
df = pd.read_sql(query, conn)
conn.close()

# Préparation des données
X = df[['year']]
y = df['total_emissions']

# Régression linéaire
```

```
model_lin = LinearRegression()
model_lin.fit(X, y)

# Prédiction jusqu à 2030
years_future = pd.DataFrame({'year': list(range(2016, 2031))})
pred_lin = model_lin.predict(years_future)

# Objectif : 50 % des émissions de 1990
emissions_1990 = 982.29
objectif_2030 = emissions_1990 * 0.5

# Affichage graphique
plt.figure(figsize=(10,6))

#Evolution des emissions de Gaz a effet de serre depuis la base de données
plt.plot(df['year'], df['total_emissions'], label="Données historiques",
    linewidth=2)

#Prédiction par la regression lineaire
plt.plot(years_future['year'], pred_lin, label="Régression linéaire", linestyle=
    '--', color = 'green')

#Emission GES en 1990
plt.axhline(emissions_1990, color='red', linestyle='-.', label="Emission en 1990
    ")

#Emission GES visé en 2030
plt.axhline(objectif_2030, color='red', linestyle=':', label="Objectif 2030 (50%
     de 1990)")
plt.xlabel("Année")
plt.ylabel(" missions  de GES")
plt.title("Prédiction linéaire des émissions de GES en UE (après 2000)")
plt.legend()
plt.grid()
plt.tight_layout()
plt.show()

# Affichage de la prédiction pour 2030
emissions_2030_lin = model_lin.predict([[2030]])[0]
print(f" Prédiction 2030 (régression linéaire) : {emissions_2030_lin:.2f}")
if objectif_2030:
print(f" Objectif 2030 (50% de 1990) : {objectif_2030:.2f}")
```

Listing 5.19 – Méthode de regression lineaire pour prédire les émissions GES jusqu'á 2030

5.14 Prévision des émissions de GES avec `auto_arima`

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from pmdarima import auto_arima

# Chargement
df = pd.read_csv("/Users/hamdinoumoulayedriss/Desktop/P-Master/Donne es tableau/
    energy_statistics.csv", delimiter=";")

# Nettoyage des colonnes
df.columns = df.columns.str.strip()

# Filtrage pour l'ann e 1990
```

```
df_1990 = df[df['year'] == 1990]

# Somme des  missions  de GES pour 1990
emissions_1990 = df_1990['greenhouse_gas_emissions'].sum()

# Calcul de l'objectif 2030 (-55 %)
objectif_2030 = emissions_1990 * 0.45
print(objectif_2030)

# Filtrer la p riode 2000  2023
df = df[(df['year'] >= 2000) & (df['year'] <= 2023)]

# Agr ger les  missions  de GES pour l'UE
df_ue = df.groupby('year')['greenhouse_gas_emissions'].sum().reset_index()

# S rie temporelle
ts = df_ue['greenhouse_gas_emissions'].values
years = df_ue['year'].values

# Pr vision avec auto_arima
model = auto_arima(ts, seasonal=False, trace=True, suppress_warnings=True)

# Pr vision pour 2024    2030
n_forecast = 8
forecast = model.predict(n_periods=n_forecast)
years_forecast = np.arange(2024, 2024 + n_forecast)

# Visualisation
plt.figure(figsize=(12, 6))
plt.plot(years, ts, label='Donn es historiques (2000  2023  )', color='blue')
plt.plot(years_forecast, forecast, label='Pr vision ARIMA (2024  2030  )', linestyle='--'
    , marker='o', color='red')
plt.plot([2023, 2024], [ts[-1], forecast[0]], linestyle=':', color='orange', label='
    Transition 2023  2024  ')

# Mise en forme
plt.title("Pr vision des  missions  de GES dans l'UE jusqu'en 2030")
plt.xlabel("Ann e")
plt.ylabel(" missions  de GES ( M t C O e )")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.axhline(y=objectif_2030, color='green', linestyle='--', label='Objectif 2030 (2250
    M t C O e )')
plt.show()
```

Listing 5.20 – Prévision ARIMA des émissions de GES dans l'UE jusqu'en 2030

BIBLIOGRAPHIE

[1] **Daniel T. Larose / Chantal D. Larose**, *Data mining : Découverte de connaissances dans les données*.

[2] **Stéphane Tufféry**, *Data Mining et Statistique décisionnelle : L'intelligence des données*.

[3] **Wes McKinney**, *Python for Data Analysis*.

[4] **Aileen Nielsen**, *Practical Time Series Analysis : Prediction with Statistics Machine Learning*.

[5] **Arnaud Guyader**, *Régression linéaire*, Université Rennes 2, Master de Statistique, 2012–2013.

[6] **Pierre-André Cornillon, Eric Matzner-Lober, Laurent Rouvière**, *Régression avec R*.

[7] **Florin Avram**, *Séries temporelles : régression et modélisation ARIMA(p,d,q)*, 2012. Disponible en ligne : `https://avram.perso.univ-pau.fr/sertemp/ser.pdf`

[8] **Frédéric Sur**, *Modélisation des séries temporelles – Séance 2 : Les processus ARIMA*. Disponible en ligne : `https://members.loria.fr/FSur/enseignement/modseries/seance2.pdf`

[9] **ENSAI**, *Cours de séries temporelles*. Disponible en ligne : `https://ensai.fr/wp-content/uploads/2019/06/Polyseriestemp.pdf`

[10] **Philippe Marchand**, *Modèles ARIMA pour les séries temporelles*, Notes de cours ECL8202. Disponible en ligne : `https://pmarchand1.github.io/ECL8202/notes`$_{c}$*ours*/11 − *Series*$_{t}$*emporelles.html*

[11] **Jean-Jacques Ruch**, *Statistique : Estimation*, Préparation à l'Agrégation, Université Bordeaux 1, 2012–2013. Disponible en ligne : `https://www.math.u-bordeaux.fr/ mchabano/Agreg/ProbaAgreg1213-COURS2-Stat1.pdf`

[12] **Didier Delignières**, *Séries temporelles – Modèles ARIMA*, Séminaire EA "Sport – Performance – Santé", Mars 2000. Disponible en ligne : `https://didierdelignieresblog.wordpress.com/wp-content/uploads/2019/09/arimacomplet.pdf`

[13] **Agnès Lagnoux**, *Séries chronologiques*, Université Toulouse – Jean Jaurès. Disponible en ligne : `https://www.math.univ-toulouse.fr/ lagnoux/Poly`$_{S}$*C.pdf*

[14] **IBM**, *Le modèle ARIMA expliqué*. Disponible en ligne : https://www.ibm.com/fr-fr/think/topics/arima-model

[15] **IA**, *ChatGPT*.