Projet ‘Température Terrestre’

Rapport final

Présentation n°1 faite le 15 février 2024
Présentation n°2 faite le 29 février 2024
Rapport final rendu le 07 mars 2024

DataScientest • com

Cohorte Data Analyst de janvier 2024

Equipe projet :

Claire MOREAU

Rahma ATTIA EL HILI

Sébastien L'HOSTE

Mentor du projet :

Tarik ANOUAR

Version du document	Date	Commentaires
0.1	30/01/2024	Initialisation du document.
1.0	15/02/2024	Ajout du contenu du document depuis les notebooks dédiés. Rendu 1 : **rapport d'exploration, de data visualisation et de pre-processing des données.**
2.0	27/02/2024	Ajout du contenu du document depuis les notebooks dédiés. Rendu 2 : **rapport de modélisation.**
2.1	29/02/2024	Modifications suite présentations rendus 1 et 2, ajouts de contenu et relecture par équipe projet et mentor.
3.0	07/03/2024	**Rapport final.**

Table des matières

1. Contextualisation du sujet et objectifs

Le réchauffement climatique est un problème mondial majeur qui résulte de l'accumulation de gaz à effet de serre dans l'atmosphère, principalement causée par les activités humaines telles que la combustion des combustibles fossiles, la déforestation et l'agriculture intensive.

Ce phénomène entraîne une augmentation de la température moyenne de la Terre, ce qui a des conséquences dévastatrices sur notre environnement. Les gaz à effet de serre sont des composés présents dans l'atmosphère qui absorbent et émettent le rayonnement infrarouge, contribuant ainsi au réchauffement de la planète. Les principaux gaz à effet de serre d'origine humaine sont le dioxyde de carbone (CO_2), le méthane (CH_4) et le protoxyde d'azote (N_2O).

Le projet d'étude de la température terrestre à travers des analyses sur Python vise à comprendre les variations de température à l'échelle mondiale et leur corrélation avec les gaz à effet de serre. Les objectifs de ce projet sont multiples. Tout d'abord, il s'agit de collecter et prétraiter les données sur les températures terrestres et les gaz à effet de serre à partir de différentes sources fiables.

Ensuite, l'objectif est d'utiliser des techniques de visualisation pour représenter graphiquement les variations de température au fil du temps et leur relation avec les concentrations de gaz à effet de serre. Cela permettra de mettre en évidence les tendances et les schémas qui se dégagent des données. Une autre étape importante consiste à effectuer des analyses statistiques pour quantifier les relations entre les températures et les gaz à effet de serre.

Le projet vise aussi à développer des modèles de prédiction en utilisant des techniques de modélisation et de machine learning. Ces modèles doivent permettre de prévoir les températures en fonction des concentrations de gaz à effet de serre sur la période considérée. Toutes ces analyses contribueront à une meilleure compréhension du changement climatique.

2. Sources et détails des données utilisées pour ce projet

Our World in Data (OWID) est une organisation qui rassemble et présente des données sur divers sujets mondiaux. Ils mettent notamment à disposition de tous des jeux de données et des analyses sur tout sujet entrant dans leurs préoccupations, allant du changement climatique, la faim dans le Monde, ...

Dans le périmètre de notre projet, deux ensembles de données d'OWID nous ont intéressé :

- Un jeu de données, appelé dataset n°1 dans la suite du document, sur les émissions annualisées de gaz à effet de serre par zone géographique depuis 1850. Les zones géographiques sont majoritairement des pays, mais des zones plus larges ou plus petites y existent aussi. Ce jeu de données comprend 2 fichiers récupérables :

    - Le premier (*owid-co2-data*) disponible au format csv, xlsx et json comprend les données brutes ;
    - Le second au format csv (*owid-co2-codebook.csv*) comprend les métadonnées du premier fichier, telles que le descriptif de chaque colonne, l'unité de mesure, et les sources de données ayant permis de calculer chacune des valeurs des colonnes.

    A noter : Le calcul des émissions de gaz à effets de serre, et celui de leur impact en évolution des températures sont extrapolés via une méthodologie spécifique. Il en découle notamment que toutes les anomalies de températures ne sont pas basées sur des mesures dans le monde réel, et utilise l'année 1850 comme valeur de départ pour les écarts de température.

- Un deuxième jeu de données, appelé dataset n°2 dans la suite du document, présentant les anomalies de températures annuelles de surface par année par pays depuis 1850. Ces données d'évolution de température sont basées sur les anomalies de température de surface mesurées (ou collectées) par le *Hadley Centre et le Climatic Research Unit* de l'Université d'East Anglia. Elles fournissent des informations sur les variations de température à long terme. Ces données sont téléchargeables au format csv.

    A la différence du précédent jeu de données, les écarts et anomalies de températures sont basés sur des mesures réelles, et utilisent la période de 1950 à 1980 comme période de référence, se situant à peu près entre le début de l'ère industrielle et maintenant. La période de référence de 30 ans semble longue, mais elle permet ainsi de lisser les variabilités mesurées d'une année à l'autre.

Liens cités :

- Informations sur l'organisation OWID : site web
- Dataset N°1 : données : GitHub dédié ; Méthodologie associée : publication.
- Dataset N°2 : données : page dédiée.

En conclusion, nous disposons ainsi de 2 jeux de données complémentaires sur une durée remontant au début de l'ère industrielle, largement démontrée comme la période où l'impact des activités humaines sur les températures globales, terrestres comme maritimes, est détectable.

3. Analyse exploratoire des données

Cette étape consiste à explorer les données pour comprendre leur structure et leur contenu, s'approprier les données, en assurer la mise en qualité en préalable des actions futures de représentation graphique, analyse statistique et prédiction par *Machine Learning*.

Note : le détail et code des actions suivantes sont disponibles dans un notebook dédié nommé *'3 Analyse exploratoire des données.ipynb'* fourni avec ce rapport.

Les étapes suivantes ont été suivies séquentiellement pour les 2 datasets :

1. Chargement bibliothèques Python ;
2. Accès au Google Drive hébergeant les sets de données ;
3. Chargement des données brutes depuis le fichier csv, et alimentation d'un dataframe par fichier source ;
4. Affichage des caractéristiques des dataframes ;
5. Recherche et gestion des données manquantes, et éventuelle suppression de données non requises pour les étapes suivantes ;
6. Recherche et gestion de données aberrantes ;
7. Renommer les variables pour un affichage de meilleure qualité dans les graphiques et résultats ;
8. Fusion à terme des 2 dataframes, prérequis des étapes suivantes (analyse statistiques et prédiction par modèle de *Machine Learning*).

- Dans le cas du dataset n°1 (données concernant l'émission des gaz à effets de serre et de leur impact sur l'évolution des températures par pays/zone au fil du temps), ce Dataframe alimenté comprend 48058 lignes et 79 colonnes.
  Il a cette structure (le grand nombre de colonnes empêche une visualisation d'ensemble) :

```
df.head()
```

	country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	...	share_global_other_co2	shar
0	Afghanistan	1850	AFG	3752993.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
1	Afghanistan	1851	AFG	3767956.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
2	Afghanistan	1852	AFG	3783940.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
3	Afghanistan	1853	AFG	3800954.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
4	Afghanistan	1854	AFG	3818038.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	

5 rows × 79 columns

Après avoir exploré les données, nous avons nettoyé les données pour éliminer les données inutiles pour notre analyse.

- Dans le cas du dataset n°2 (données concernant les écarts de température par pays/zone au fil du temps), le Dataframe alimenté comprend 29 566 lignes et 4 colonnes.

  Il a cette structure :

```
df2.head()
```

	Entity	Code	Year	Surface temperature anomaly
0	Afghanistan	AFG	1947	1.93
1	Afghanistan	AFG	1948	0.83
2	Afghanistan	AFG	1949	0.05
3	Afghanistan	AFG	1950	-1.36
4	Afghanistan	AFG	1951	-0.03

Gestion des données manquantes :

- Tout comme pour le **Dataset n°1**, l'absence de données '*Code*' est très probablement dû à la présence de régions ou zones non codifiées dans la norme ISO-3166-1, ou alternativement, que le pays existe bien, mais son code n'a pas été répercuté dans le **Dataset n°2** ;
- Une simple analyse du couple de modalités des colonnes '*Entity*' et '*Code*' devrait permettre de valider si ces hypothèses sont les bonnes ;
- Une recherche confirme que le pays 'Micronésie' existe bien, et est codifié 'FSM' dans la codification à 3 lettres de la norme ISO-3166-1. Sachant qu'une seule entité est affectée, on peut remplacer tous les NaN de la colonne code par le code 'FSM'.

Gestion des données aberrantes ou superflues : cas des zones/régions avec un code ISO-3166-1 fantaisiste :

- Une analyse minutieuse des données, en lien avec celles présentes dans le SataSet n°1, est le présente de codes ISO non présents dans le dataset N°1, ni dans les codes disponibles dans la norme ISO-3166-1 ;
- Pour valider et identifier ces données inutiles pour la suite de l'analyse, il suffit donc d'identifier tout code non présent dans les codes alpha-3 de la norme ISO-3166-1.

Note :

Bien qu'il soit possible d'utiliser le nom des pays à la place des codes, on peut se heurter d'un fichier à l'autre d'avoir des différences, liées à de multiples facteurs, tels que : casse différente, langage et/ou alphabets différents (il existe plusieurs catégories de noms pour un pays dans la norme), problèmes liés aux diacritiques...

Il est donc très recommandé d'utiliser les codes alpha-2 ou alpha-3 de la norme, ou à la valeur numérique associée à un pays. La colonne '*Entity*' n'a donc pas lieu d'être pour la suite des opérations.

Cohérence des données, et éventuelle élimination de colonnes en prévision de la fusion avec le dataset N°1 :

- La finalité de la fusion est de rajouter par pays et par année la valeur d'évolution des températures ;
- Les clés pour réaliser cette fusion, sont le code de l'entité et l'année (ex : 'AFG' et 1987) ;
- La casse du code entité est identique dans les 2 dataframes ;
- Il n'y a donc aucune action supplémentaire à réaliser sur ce dataframe.

Fusion des données :

- Après ces dernières opérations de mise en qualité de données, la fusion des 2 datasets a été réalisée par la méthode merge en jointure interne (option *how = 'inner'*), en utilisant les données années et code Pays alpha-3 ISO-3166.

4. Data Visualization

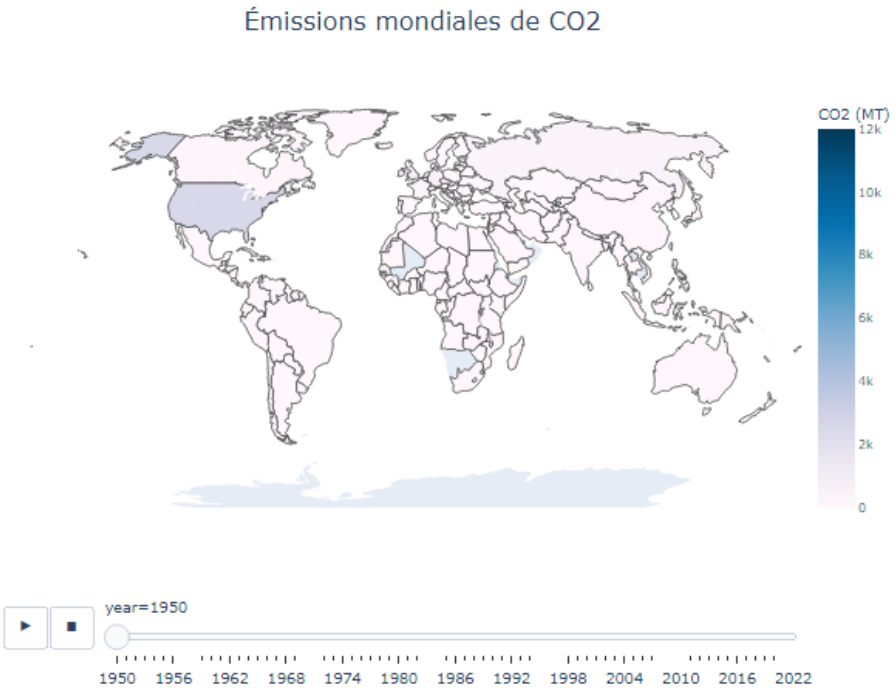
Cette partie consiste à représenter graphiquement les données pour faciliter leur compréhension et mettre en évidence les tendances, les modèles et les relations entre les différents éléments. Pour cela on a utilisé la bibliothèque Plotly, pour avoir des visualisations interactives.

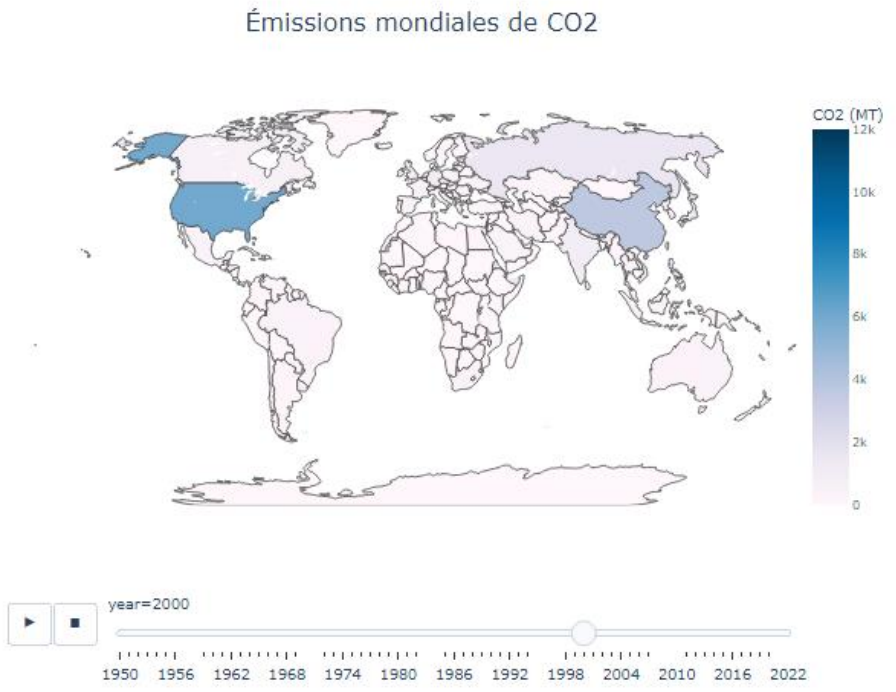
Note : le détail et code des actions suivantes, ainsi que les graphiques interactifs sont disponibles dans un notebook dédié nommé *'4 et 5 Data Visualization et Analyses stats.ipynb'* fourni avec ce rapport.

Note² : les graphiques utilisent généralement la donnée année en abscisse, cependant les données les plus anciennes sont très souvent et de plus en plus manquantes au fur et à mesure que l'on s'éloigne de de la période actuelle. La période sélectionnée sera donc spécifiée au cas par cas, et motivée pour la raison ci-dessus, afin de produire l'analyse la plus fiable possible.

4.1. Emission de CO_2 par pays au fil du temps

Les deux graphiques suivants représentent la contribution relative de chaque pays à la production annuelle de CO_2 depuis l'année 1950 :





Analyse :

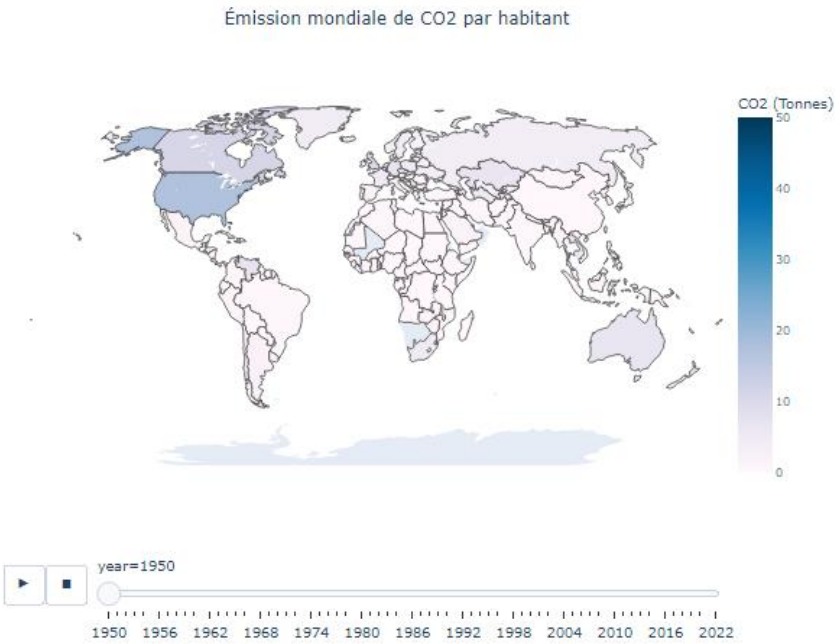
- En observant la carte interactive, nous pouvons remarquer que les émissions de CO_2 ont augmenté de manière constante depuis 1950. Les pays ayant les émissions les plus élevées sont les États-Unis, la Chine, l'Inde, la Russie et le Japon. Ces pays ont des émissions de CO_2 plusieurs fois plus élevées que d'autres pays de la carte, ce qui indique que la majeure partie de l'émission de CO_2 dans le monde provient de ces pays.

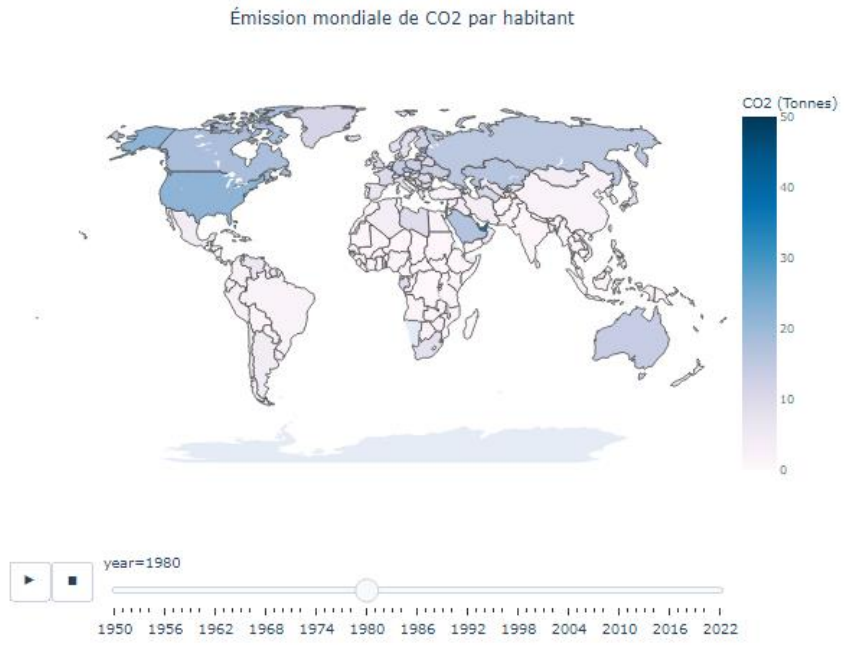
- Nous pouvons également observer que les émissions de CO_2 ont connu une augmentation rapide à partir du milieu des années 1950 et ont augmenté de manière plus prononcée à partir des années 2000. Cette augmentation rapide peut être liée à l'augmentation de la population mondiale et à l'augmentation de la demande d'énergie pour les transports, la production manufacturière et l'électricité.

- Cependant, il existe également des différences régionales dans les émissions de CO_2. L'Amérique du Nord, l'Europe et l'Asie ont des émissions plus élevées que l'Afrique et l'Amérique latine. Cela peut être dû à des facteurs tels que les niveaux de développement économique, les politiques environnementales et les sources d'énergie employées.

4.2. Emission de CO_2 par pays per capita au fil du temps

Les deux graphiques suivants représentent la contribution relative de chaque habitant de chaque pays à la production annuelle de CO_2 depuis l'année 1950 :





Analyse :

- En observant la carte interactive pour l'émission de CO_2 par habitant de 1951 à 2022, nous pouvons remarquer rapidement que les pays ayant les émissions de CO_2 par habitant les plus élevées sont les États-Unis, le Canada, l'Australie, mais aussi - il vaut mieux zoomer

pour les voir - les pays du golfe persique (Qatar, Koweït, Emirats Arabes Unis, Oman et Arabie saoudite). Cependant, si l'on compare les émissions de CO_2 par habitant entre les pays à différents moments du temps, il est également possible de voir des changements importants. Par exemple, nous pouvons constater que certains pays ont réduit leurs émissions de CO_2 par habitant au fil du temps. Cette baisse peut être due à des efforts pour favoriser la production d'énergie renouvelable et réduire la consommation d'énergie fossile.

- En revanche, certaines régions ont connu une augmentation rapide de leurs émissions de CO_2 par habitant au cours des dernières décennies. C'est le cas notamment de l'Asie, où plusieurs économies en développement ont connu une croissance rapide de leur émission de CO_2 par habitant depuis les années 1980. Cette évolution reflète une industrialisation rapide et une croissance économique rapide dans la région.

- Cependant, si l'on compare les émissions de CO_2 par habitant des différentes régions du monde, on peut noter que l'Amérique du Nord, l'Europe et l'Asie ont tendance à avoir des émissions de CO_2 par habitant plus élevées que d'autres régions, du moins jusqu'à récemment. Le niveau de développement économique, les modes de consommation, les habitudes de transport, la composition de l'énergie et les politiques environnementales sont des facteurs qui peuvent expliquer ces différences régionales.

4.3. Sélection des 15 pays les plus gros émetteurs de CO_2

Le graphique suivant représente les 15 plus gros pays contributeurs historiques à la production annuelle de CO_2 par combustion de ressources fossiles sur la période 1950 à 2017 :

	country	co2
215	United States	4622.852233
43	China	3540.094781
167	Russia	1548.342781
80	Germany	894.600753
102	Japan	872.883082
94	India	791.822822
214	United Kingdom	550.266890
38	Canada	414.535877
212	Ukraine	396.886082
75	France	379.668356
100	Italy	325.566918
162	Poland	315.744479
188	South Africa	279.836616
128	Mexico	272.661959
190	South Korea	267.329438

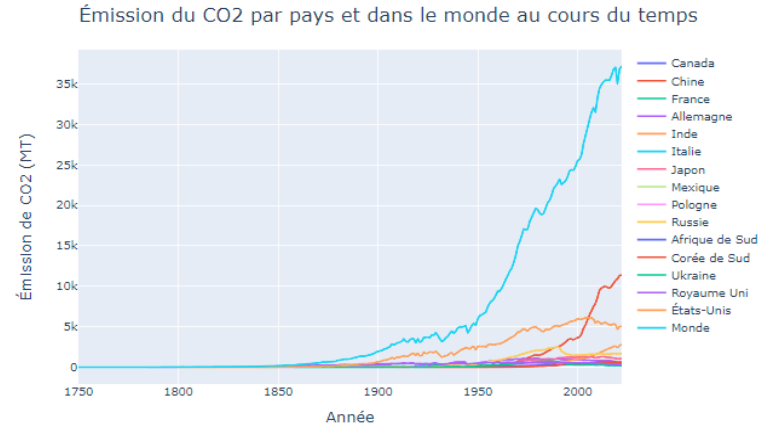
Deux données expliquent ce choix de sélection :

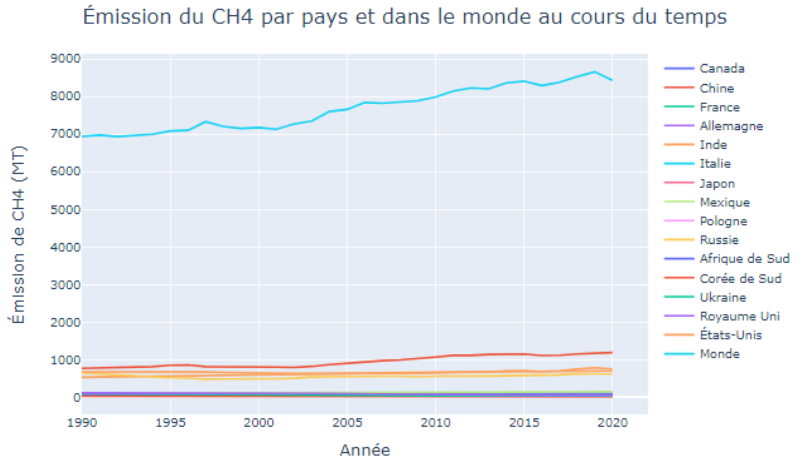
- Les 3 plus gros pays émetteurs combinés contribuent à **47 %** des émissions mondiales ;
- Les 15 plus gros pays émetteurs combinés contribuent à **75 %** des émissions mondiales.

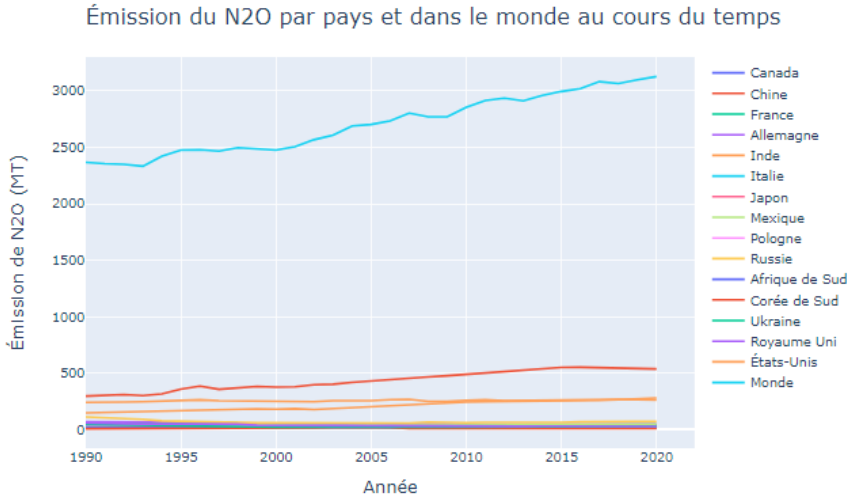
Dans la suite de ce document, ces 15 plus gros contributeurs seront appelés Top15.

4.4. Représentation des 15 plus gros pays émetteurs de gaz à effet de serre

Les graphiques suivants représentent la contribution relative de chaque pays à la production annuelle depuis l'année 1990 de trois gaz à effet de serre (CO_2, CH_4 et N_2O) :







- **Analyse sur les émissions de CO_2 :**

  - Depuis les années 1900, le niveau mondial d'émissions de CO_2 a augmenté très rapidement. Cette augmentation s'explique par la croissance économique, l'industrialisation et l'augmentation de la population mondiale.

  - Les États-Unis ont produit la grande majorité des émissions de CO_2 jusqu'aux années 2000, date à laquelle la Chine est devenue le plus pollueur du monde. Cependant, les émissions de CO_2 de la Chine ont augmenté de manière très rapide, dépassant celles des États-Unis.
  - Le Japon, l'Allemagne, le Royaume-Uni, le Canada et la France sont également responsables d'une quantité importante d'émissions de CO_2.

  - La courbe montre que les pays ont des trajectoires différentes pour les émissions de CO_2. Par exemple, les émissions des États-Unis ont commencé à stagner depuis les années 2000, tandis que celles de la Chine ont continué à augmenter de manière prononcée.
  - Nous pouvons également observer une tendance à la baisse pour les pays Européens où les émissions de CO_2 ont connu une réduction depuis les années 1990, cela peut être attribué à un développement plus sain avec une utilisation accrue de sources d'énergie renouvelable.

- **Analyse sur les émissions de CH_4 :**

  - Tout comme pour les émissions de CO_2, la Chine est responsable de la majeure partie des émissions de CH_4, suivie par les Etats Unis, la Russie et l'Inde. Les émissions de CH_4 ont également augmenté de manière importante au Mexique et en Australie.

  - Les émissions de CH_4 sont associées à l'agriculture, à la production de combustibles fossiles et à la production de déchets.

  - Les émissions varient considérablement dans le temps. Par exemple, au cours des deux dernières décennies, les émissions de CH_4 ont diminué en Allemagne et en Italie, mais ont augmenté en Chine et en Inde. Cela est également associé à certaines pratiques agricoles, comme l'élevage de bétail et la culture de riz, qui sont plus fréquentes dans certains pays que dans d'autres.

- **Analyse sur les émissions de N_2O :**

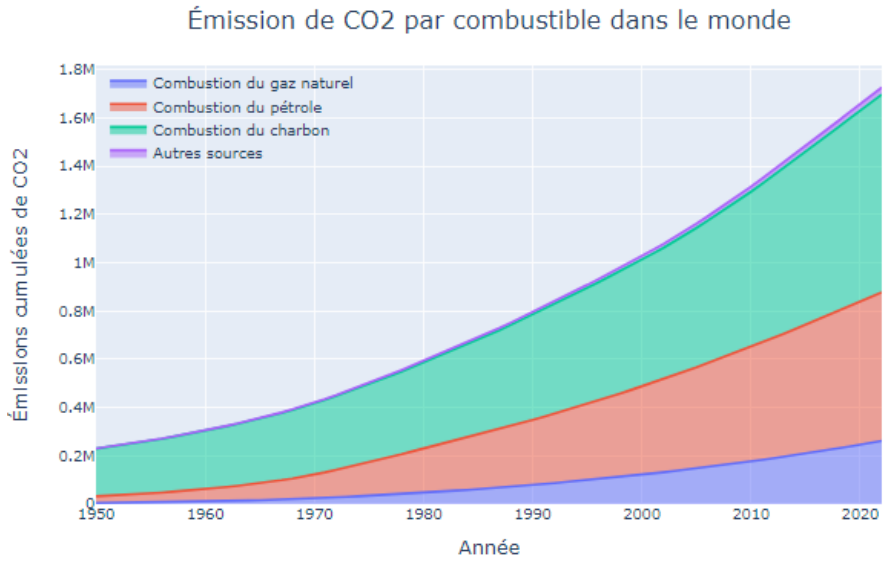
  - Les émissions de N_2O sont fortement liées à l'utilisation accrue de fertilisants dans l'agriculture, ainsi qu'à l'utilisation croissante de combustibles fossiles. La Chine est le plus grand producteur d'émissions de N_2O, suivis par les États-Unis et l'Inde.

  - Les émissions de N_2O sont particulièrement difficiles à diminuer car elles sont associées à l'utilisation de fertilisants azotés, qui sont essentiels à la production alimentaire.

  - Les émissions varient considérablement dans le temps et en fonction des politiques nationales. Les pays qui ont mis en place des politiques visant à réduire leur utilisation de fertilisants, comme l'Allemagne et la France, ont des émissions plus faibles de N_2O.

4.5. Evolution des ressources émettrices de CO_2

Le graphique suivant représente la contribution relative de chaque ressource fossile à la production annuelle mondiale de CO_2 depuis l'année 1950 :



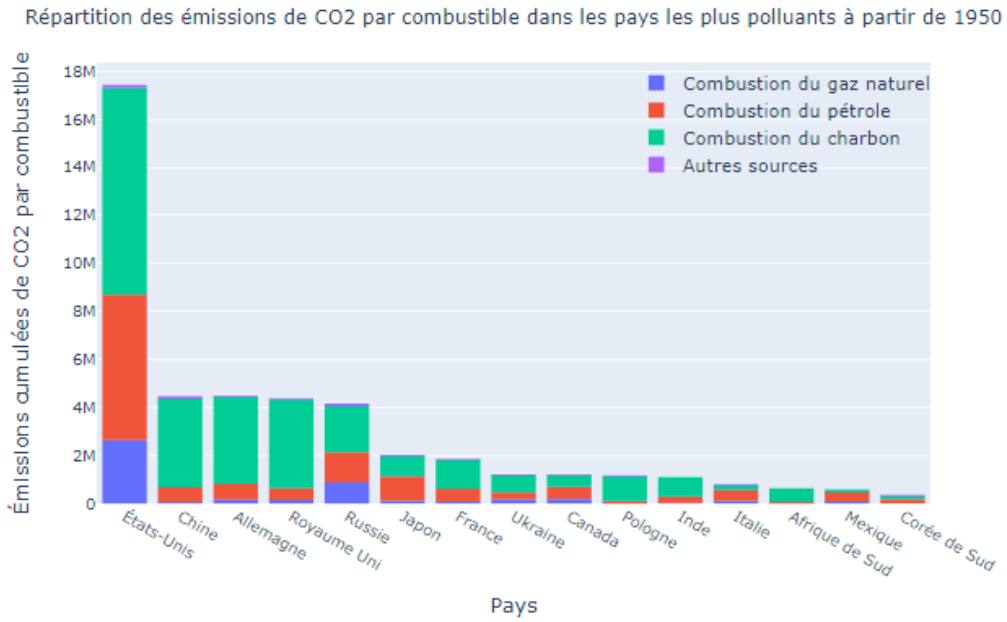
Analyse :

- À partir de ce graphique, nous pouvons observer que le charbon est le plus grand contributeur aux émissions de CO_2, suivi du pétrole et du gaz naturel. Au fil des années, la contribution relative de chaque combustible aux émissions de CO_2 a augmenté de manière significative. Les émissions de CO_2 liées au gaz naturel ont également augmenté à un rythme plus lent que celles du pétrole.

- Les autres sources d'émissions de CO_2 (telles que la production industrielle) ont tendance à être stables au fil du temps.

- Le charbon est le plus grand contributeur aux émissions de CO_2 car il est historiquement la source d'énergie fossile la plus largement utilisée pour produire de l'électricité et de la chaleur. La combustion du charbon produit environ deux fois plus de dioxyde de carbone (CO_2) que la combustion de gaz naturel ou de pétrole pour la même quantité d'énergie produite.

Le graphique suivant représente la contribution relative de chaque ressource fossile à la production annuelle de CO_2 depuis l'ère préindustrielle, pour les pays du Top15 identifié au §4.1 :



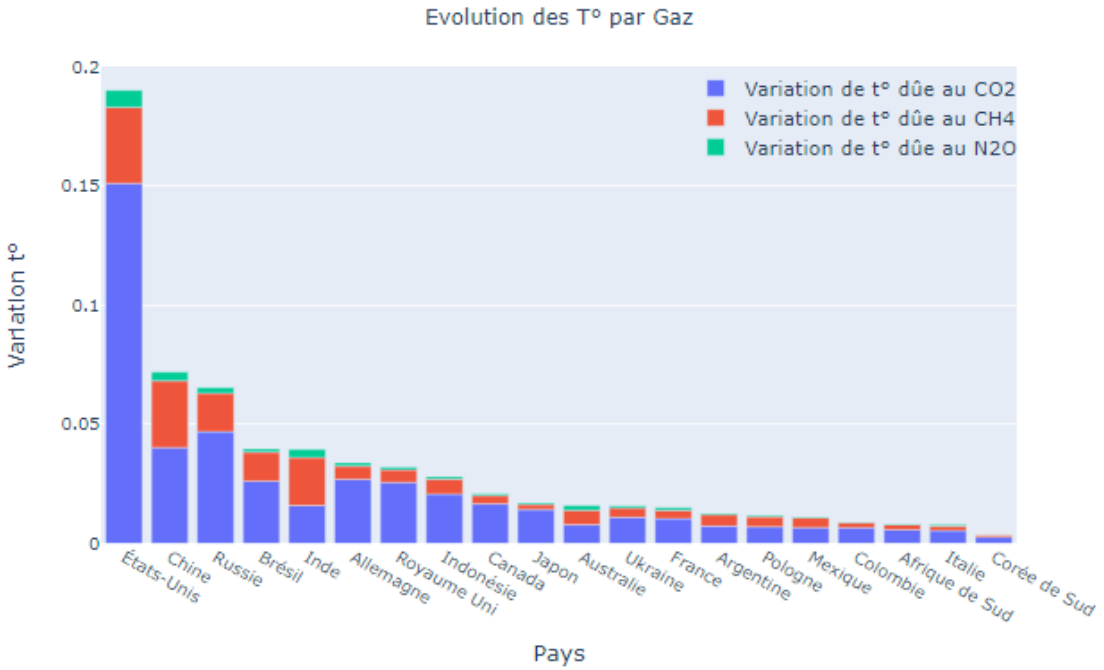
Analyse :

- L'histogramme de la répartition des émissions de CO_2 par combustible dans les pays les plus polluants, nous permet de constater plusieurs tendances intéressantes. Le charbon est le principal combustible utilisé dans tous les pays examinés, excepté le Japon, le Canada, l'Italie et le Mexique. L'utilisation du pétrole est la plus élevée aux États-Unis tandis que la Chine utilise principalement le charbon. Le gaz naturel est plus utilisé en Russie et au Japon.

- En analysant cette répartition, il est clair que la dépendance au charbon persiste malgré les efforts mondiaux pour réduire les émissions de gaz à effet de serre. Les pays économiquement puissants comme les États-Unis et la Chine sont les plus grands émetteurs de CO_2, principalement en raison de leur utilisation intensive du charbon.

4.6. Contribution des gaz à effet de serre à l'évolution des températures

Le graphique suivant représente la contribution relative de chaque ressource fossile à l'évolution globale de températures pour chacun des gaz à effet de serre, depuis l'année 1980 pour les pays du Top15 identifié au §4.1 :



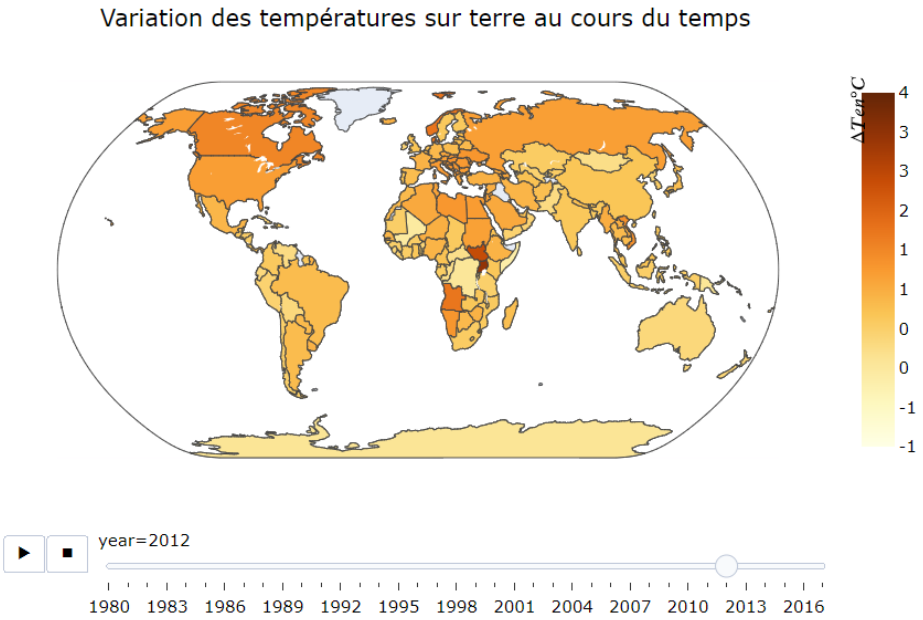
Analyse :

- Ici, on remarque à quel point les Etats-Unis ont agi sur le réchauffement des températures depuis 1950 via les émissions de gaz à effet de serre (et surtout via le CO_2).

- En Inde, ce n'est pas le CO_2 le plus gros responsable du réchauffement mais le méthane. Il est très important aussi en Chine.

- On remarquera également qu'on a remis quelques pays dont le nom n'apparaissait pas en tant qu'émetteurs mais qui sont ici importants quand on regarde les évolutions de température. On pourra en particulier remarquer le Brésil et l'Indonésie.

Le graphique suivant représente la contribution relative de chaque ressource fossile à l'évolution globale de températures pour chacun des gaz à effet de serre, depuis l'année 1980 :

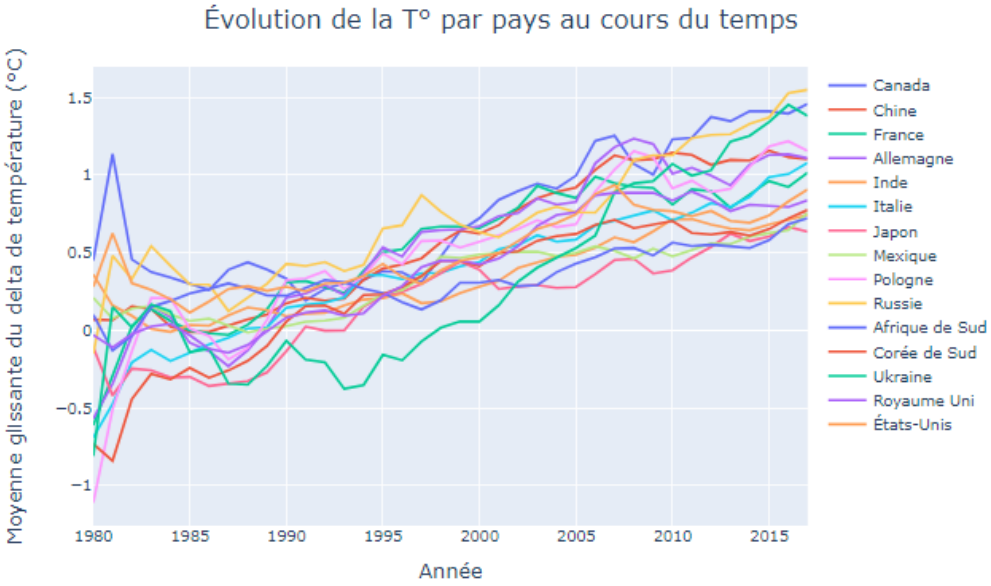


Analyse :

- Cette carte permet de visualiser les évolutions de température par année et par pays. On note qu'après les années 2010, les températures semblent augmenter sur l'ensemble du globe.

- Cependant, ça n'est pas très facile à constater "à l'œil" dans les années précédentes.

La représentation suivante présente la contribution relative de chaque pays du Top 15 à l'évolution globale de températures, depuis l'année 1980 :

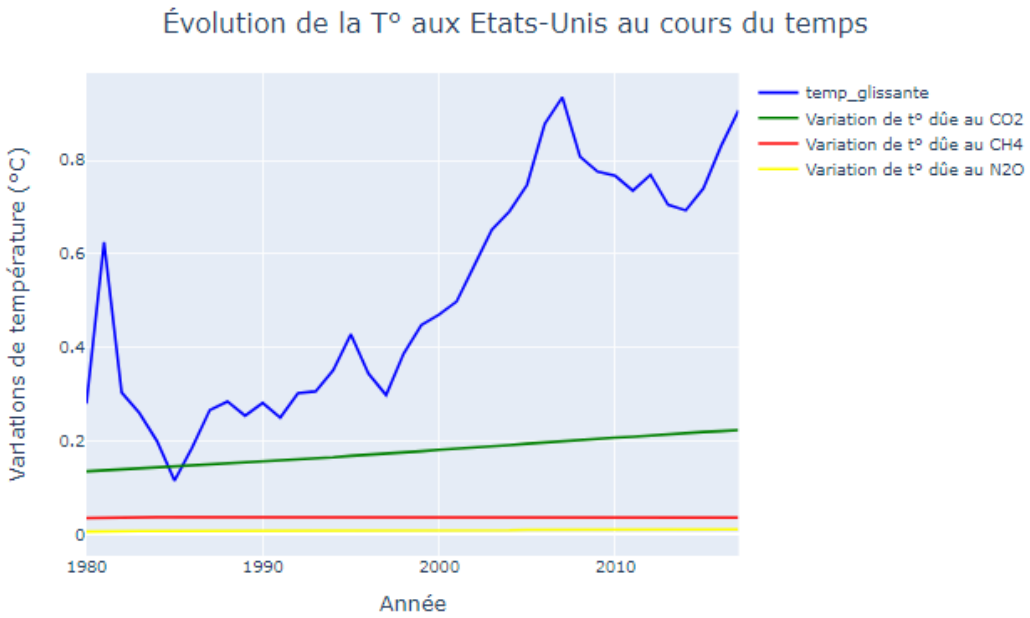


Analyse :

- Il n'apparaît pas de manière évidente ici quels sont les pays qui ont les plus fortes variations de température.

- En revanche, on voit très nettement une tendance à la hausse pour chacun des pays pour une variation à la hausse entre 0.5°C et 1.5°C.

La représentation suivante présente l'évolution globale de températures (telle que mesurée par des stations au sol et maritimes) aux Etats-Unis comparée aux évolutions calculées de températures liées à la combustion des trois ressources fossiles, depuis l'année 1980 :



Analyse :

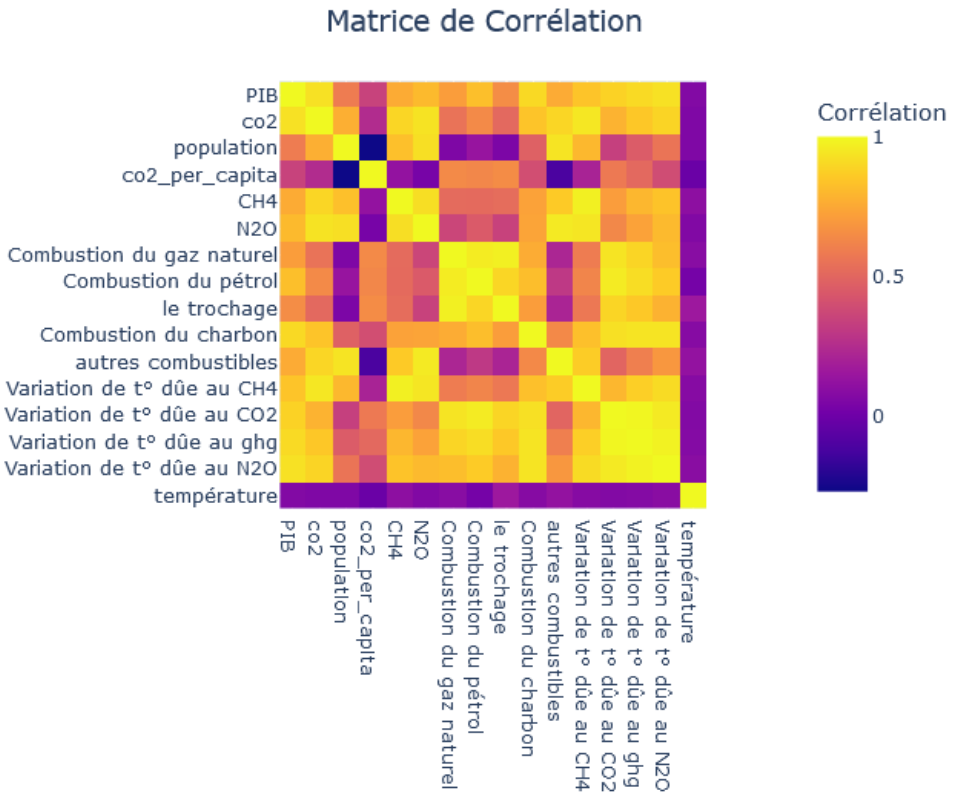
- On ne voit pas de façon évidente ici une corrélation entre l'évolution de la température aux Etats-Unis et la variation de température due aux émissions (cumulées ou pas) des gaz à effet de serre.

5. Analyses statistiques

Rappel : le détail et code des actions suivantes, ainsi que les graphiques interactifs sont disponibles dans un notebook dédié nommé *'4 et 5 Data Visualization et Analyses stats.ipynb'* fourni avec ce rapport.

5.1. Matrice de corrélation entre émission des gaz à effet de serre et évolution des températures

Cette matrice cherche à identifier une corrélation entre les variables présentes dans le dataframe :



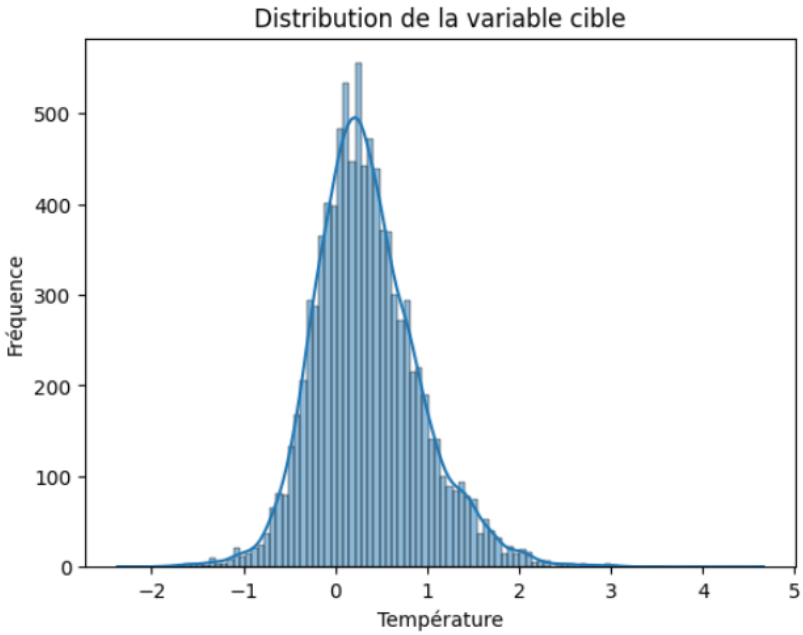
Analyse :

- A partir de la matrice de corrélation on note la présence de corrélation positive entre les températures dues aux gaz à effet de serre et leurs émissions. Ce résultat est probablement compréhensible vu que les températures ont été calculées à partir des données d'émissions.

- En revanche, la matrice ne montre pas une corrélation positive entre la variable température et le reste des variables cela signifie qu'il n'y a pas de relation linéaire entre ces deux variables

5.2. Distribution des données de température

Pour vérifier la normalité de la variable température, on a opté pour le test Shapiro–Wilk.



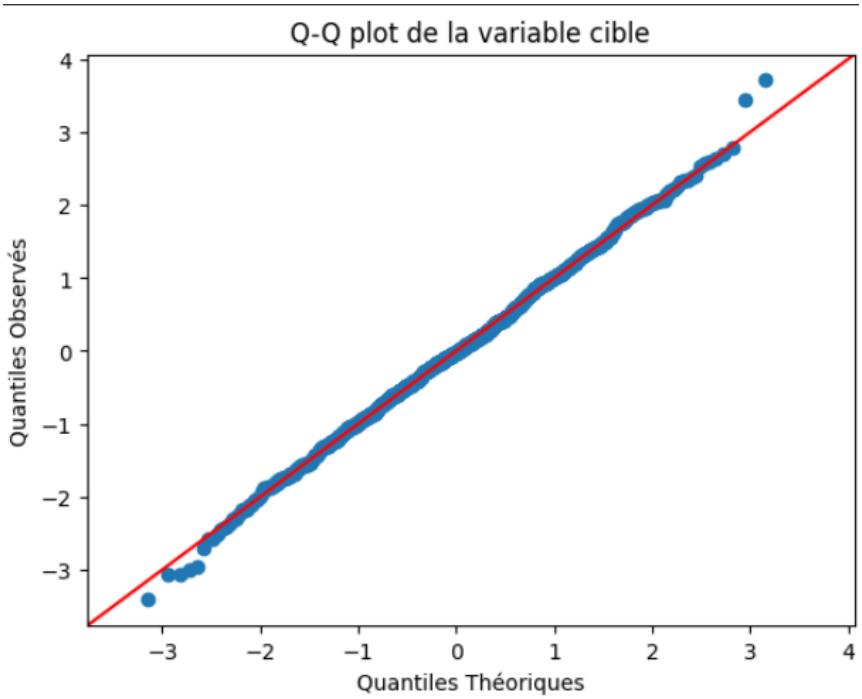
ShapiroResult(statistic=0.9987658262252808, pvalue=0.5599465370178223)

Analyse :

- La p-value (=0.55) est supérieure à 0.05 donc la variable température suit une distribution normale.

5.3. Normalité de la variable d'évolution des températures terrestres

La représentation suivante permet de déterminer si une variable suit une loi de distribution gaussienne normalisée. Dans le cas de la variable température sur notre ensemble de données, le résultat est le suivant :



Analyse :

- Le Q-Q plot nous a permis de montrer que les points de la variable température sont alignés sur la première bissectrice. Cette figure confirme que la distribution de cette variable suit une loi de distribution gaussienne normalisée.

5.4. Conclusions sur l'analyse statistique des données

Il peut y avoir des retards entre le changement de la température et le changement des émissions de CO_2, car les effets de l'augmentation du CO_2 peuvent prendre du temps pour se manifester pleinement sur la température.

Investigation et explication :

Cette absence de corrélation entre les anomalies de température globales ou localisées et les émissions de gaz à effet de serre depuis le début de l'ère industrielle semble étonnante au premier abord, cependant elle s'explique par différents éléments.

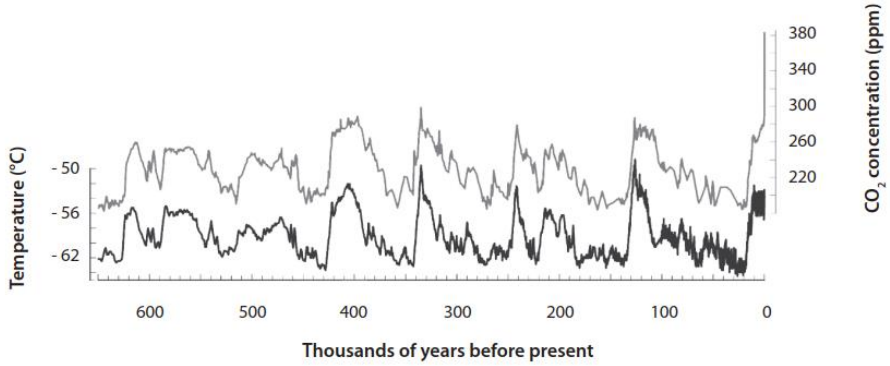
On peut par ailleurs constater que la représentation des données mesurées de température vis-à-vis des émissions de gaz à effet de serre est absente des analyses faites par l'organisation OWID ou sur d'autres sites institutionnels faisant autorité sur ce sujet.

De fait, une corrélation démontrée est bien documentée dans la littérature entre températures globales et gaz à effet de serre, mais elle concerne ces températures et la CONCENTRATION ATMOSPHERIQUE (généralement défini en parts par million ou ppm) de ces différents gaz à un temps T, et reproductible depuis le début de l'air industrielle et même bien au-delà sur des périodes géologiques.

Un exemple sur de longues périodes de temps de cette recherche de corrélation et représentation associée est ainsi :

Correlation Is Not Causation

Figure 2.1: Estimated temperature and carbon dioxide levels in Antarctica over the past 650,000 years



Source: Physics Institute, University of Bern, Switzerland. Adapted from Fretwell, Holly (2007). *The Sky's Not Falling: Why It's Okay to Chill about Global Warming*. World Ahead Media.

A noter que bien que nous disposions des évolutions de températures induites ou non par les effets des gaz à effet de serre, nous ne disposons pas des évolutions de températures mesurées locales par pays sur la période considérée, et non globales (par hémisphère ou mondiale par exemple). En l 'état, il est donc normal que nous n'ayons pas cherché à étudier cette corrélation à l'échelle de la planète, et que nous nous soyons concentrés sur une évolution locale des évolutions de températures.

Publication citée : https://www.fraserinstitute.org/sites/default/files/UnderstandingClimateChange-Lesson2.pdf

6. Pre-processing des données

Rappel : le détail et code des actions suivantes sont disponibles dans un notebook dédié nommé '*6 et 7 Preprocessing et Modélisation.ipynb*' fourni avec ce rapport.

A la suite de l'analyse statistique, une approche prédictive par modèle de *Machine Learning* est entamée. Au préalable, différentes étapes de traitement des données sont cependant requises :

- Éliminer les colonnes qui sont la résultante d'un calcul sur les autres colonnes,
- Gérer les nombreuses valeurs manquantes (données NaN) :
  - Pour le PIB, la population et les émissions de CO_2, on interpole les valeurs manquantes quand on a ses voisines dans le même pays,
  - Pour les autres valeurs de CO_2 manquantes, on supprime la ligne correspondante
  - On élimine tous les pays n'ayant pas d'info de PIB,
  - Supprimer les colonnes méthane et oxydes nitreux pour qu'il nous reste suffisamment de lignes dans notre base de données,
- Vérifier qu'il n'y a pas des valeurs extrêmes (*outliers*) aberrantes.

Nous obtenons ainsi une table de 7610 lignes par 10 colonnes :

zone_geo	pays	iso_code	continent	année	population	pib	co2	delta_T°_dû_aux_ghg	temperature
11	Canada	CAN	1	1970	21434580.0	4.177520e+11	341.177	0.015	-0.23
11	Canada	CAN	1	1971	21888686.0	4.410566e+11	352.287	0.016	-0.05
11	Canada	CAN	1	1972	22222228.0	4.643421e+11	380.792	0.016	-1.55
11	Canada	CAN	1	1973	22502026.0	4.976174e+11	381.273	0.016	0.50
11	Canada	CAN	1	1974	22812430.0	5.179518e+11	389.617	0.017	-0.65

Nous disposons de **données géographiques** à exploiter puisque nos données sont par pays. Un regroupement des pays à 2 niveaux, par zones géographiques variable '*zone_géo*') ou par continent (variable '*continent*'), a été réalisé (depuis une classification internationale et publique). Pour exemple, le Sénégal est ainsi membre de la zone géographique Afrique de l'Ouest, elle-même membre du continent Afrique.

Ces trois niveaux de maillages (par pays, zone géographique et continent) ont ensuite été soumis pour nos modèles de *Machine Learning*.
Suite à quelques essais, la variable '*zone_géo*' apparait comme le meilleur compromis.
On a ainsi 19 zones géographiques dans le monde regroupant l'ensemble des plus de 200 pays du Monde. Chaque continent est typiquement divisé en 4 ou 5 zones géographiques.

Lien cité : Table de regroupement géographique M49 de l'ONU.

Le modèle d'apprentissage sera donc sur 6 variables d'apprentissage (*features*) :

- *'zone_géo',*
- *'année',*
- *'population',*
- *'pib',*
- *'co2',*
- *'delta_T°_dû_aux_GHG',*

et la variable cible (*target*) *'température'* (évolution de température en °C).

L'encodage des différentes variables est effectué comme suit :

- La variable catégorielle '*zone_géo'* est encodée avec un *OneHotEncoding*,
- Les variables '*population*', '*pib*', '*co2*', '*delta_T°_dû_aux_GHG*' et *'température'* sont transformées avec un *RobustScaling*, pour tenter de gérer nos *outliers*,
- La variable '*année*' est transformée via une normalisation. Ces données étant temporelles, on a étudié les séries temporelles, mais il s'avère qu'on ne peut pas les utiliser ici vu que nous n'avons ni saisonnalité, ni cycle.

7. Modélisations par *Machine Learning* et prédictions

Rappel : le détail et code des actions suivantes sont disponibles dans un notebook dédié nommé '*6 et 7 Preprocessing et Modélisation.ipynb*' fourni avec ce rapport.

7.1. Finalité

En absence de corrélation entre évolution locale des températures et production des gaz à effet de serre, les prédictions sont donc réalisées sur la période de temps de 1970 à 2017 (données les plus récentes disponibles).

7.2. Méthodologie

Un premier test a permis de vérifier si la taille des jeux d'entraînement et test lors du Test-Train-Split, avec des valeurs de 20, 25 et 30%. Aucune différence détectable sur l'impact de ces valeurs n'a été détectée sur les métriques testées lors de nos tests préliminaires.

Les modèles de *Machine Learning* **de régression** suivants ont été testés, selon le cas :

- Manuellement,
- Via la classe *GridSearchCV* de la bibliothèque *scikit-learn*, qui permet une optimisation automatisée des hyperparamètres.

Le résumé des tests combinatoires utilisés, et résultats analysés, est le suivant :

Modèles de Machine Learning	Paramètres à tester	Méthode de tests	Métriques testées	Visualisation(s) utilisée(s)
LinearRegression	Par défaut	Manuel	MSE, RMSE, MAE, R^2 sur les jeux de test et d'apprentissage	Résidus Courbe d'apprentissage
DecisionTreeRegressor	Par défaut	Manuel	MSE, RMSE, MAE, R^2 sur les jeux de test et d'apprentissage	Histogramme d'importance, Résidus, Courbe d'apprentissage
RandomForestRegressor	n_estimators, max_depth	Manuel + GridSearchCV avec n_estimators : arange(10,310,30) max_depth : arange(3,15,3)	MSE, RMSE, MAE, R^2 sur les jeux de test et d'apprentissage	Histogramme d'importance, Résidus, Courbe d'apprentissage
XGBoostRegressor	n_estimators, learning_rate, max_depth, min_child_weight, gamma	Manuel + GridSearchCV dans n_estimators : [100, 200,300], learning_rate : [0,01, 0.05, 0,1], max_depth : [3, 5, 7, 9], min_child_weight : [1, 3, 5], gamma : [0, 0.1, 0.3]	MSE, RMSE, MAE, R^2 sur les jeux de test et d'apprentissage	Histogramme d'importance, Résidus, Courbe d'apprentissage

Cela représente par exemple, pour les modèles suivants :

- ***'RandomForestRegressor'*** : 40 tests sur 5 jeux différents de données ;
- ***XGBoostRegressor*** : 75 tests puis 18 tests sur 5 jeux de données à chaque fois.

7.3. Indicateurs obtenus

Les résultats obtenus sont les suivants :

Métriques calculées en fonction du modèle choisi	MSE	RMSE	MAE	R^2 sur jeu de test	R^2 sur jeu d'apprentissage
LinearRegression	0,21	0,46	0,35	0,4	0,4
DecisionTreeRegressor	0,26	0,51	0,36	0,25	1
RandomForestRegressor	0,13	0,37	0,26	0,62	0,95
RandomForestRegressor (max_depth=12, n_estimators=280)	0,14	0,38	0,28	0,59	0,79
XGBoostRegressor	0,14	0,37	0,27	0,61	0,88
XGBoostRegressor (n_estimators=300, max_depth=9, learning_rate=0.05)	0,13	0,36	0,26	0,64	0,94

Comme on peut le voir dans le tableau avec les différentes métriques, les modèles *LinearRegression* et *DecisionTreeRegressor* ne sont pas adaptés à notre problématique.

Des essais additionnels avec les algorithmes *RandomForestRegressor* et *XGBoostRegressor*. Les tests avec *GridSearchCV* permettant de tester de nombreuses combinaisons de paramètres nous ont permis d'affiner un peu notre modèle, mais ne modifient pas énormément les résultats.

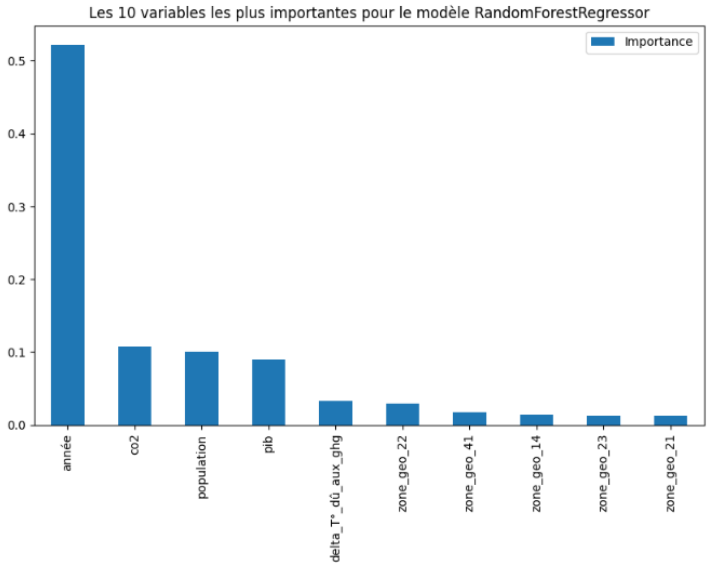
Pour la suite, on s'intéresse à *RandomForestRegressor* avec ses paramètres par défaut (*n_estimators =100, max_depth=None*), et *XGBoostRegressor* (*n_estimators=300, max_depth=9, learning_rate=0.05*).

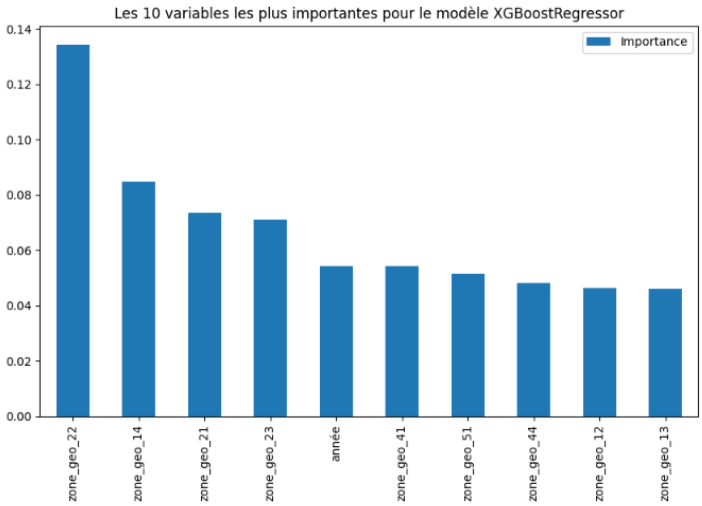
Pour rappel :

- **MSE** : Erreur quadratique moyenne (ou variance des résidus) ;
- **RMSE** : Racine de l'erreur quadratique moyenne (écart-type des résidus) ;
- **MAE** : Erreur absolue moyenne ;
- **R^2** : coefficient de détermination.

7.4. Interprétation des 2 modèles les plus prometteurs

En analysant les résultats de chaque modèle, on a choisi de se focaliser plus sur les modèles *RandomForestRegressor* et *XGBoostRegressor* car ils ont des performances assez bonnes, avec un R^2 supérieur à 0,6 pour les deux modèles. Cela signifie que les deux modèles peuvent expliquer environ 60% de la variance observée dans les données.

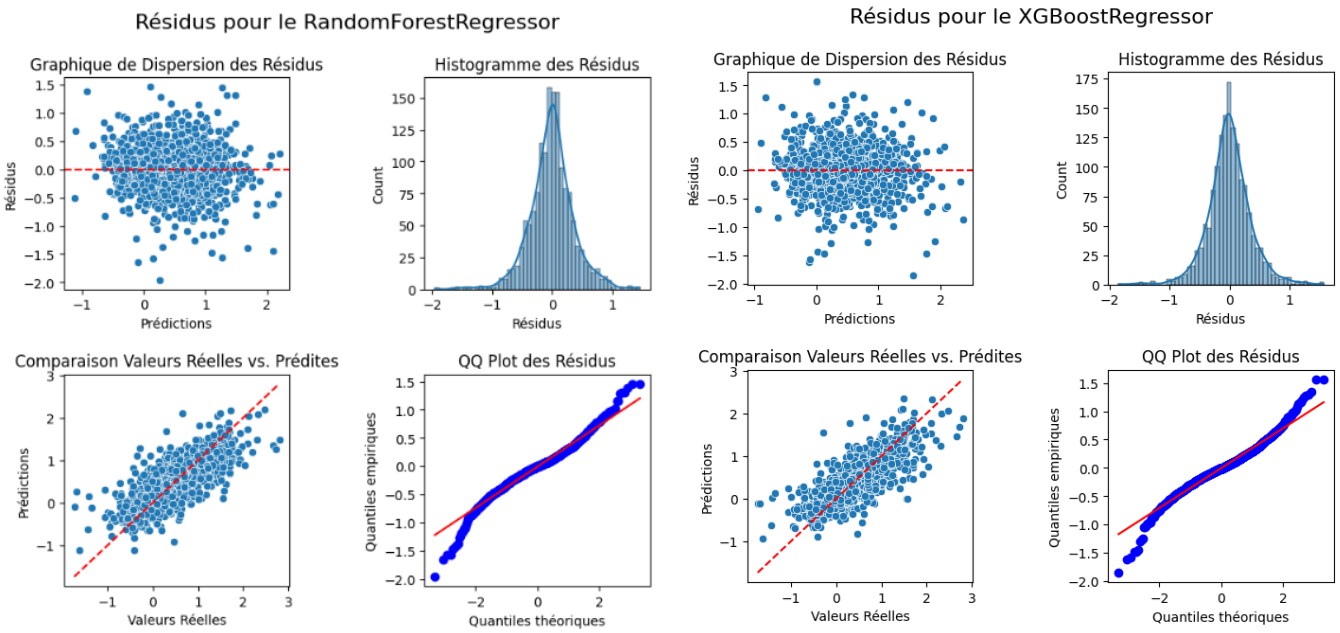




En termes d'importance des variables (*Feature Importance*), les résultats montrent que l'année est la variable la plus importante pour prédire la température pour le modèle *RandomForestRegressor*. Cependant, cela ne signifie pas que l'année est la cause principale de l'augmentation de la température. L'importance de l'année dans le modèle est plutôt due au fait que la température de la Terre a tendance à augmenter au fil des années en raison du changement climatique. Le CO_2, la population et le PIB ont également une certaine importance pour prédire la température, mais leur impact est relativement faible.

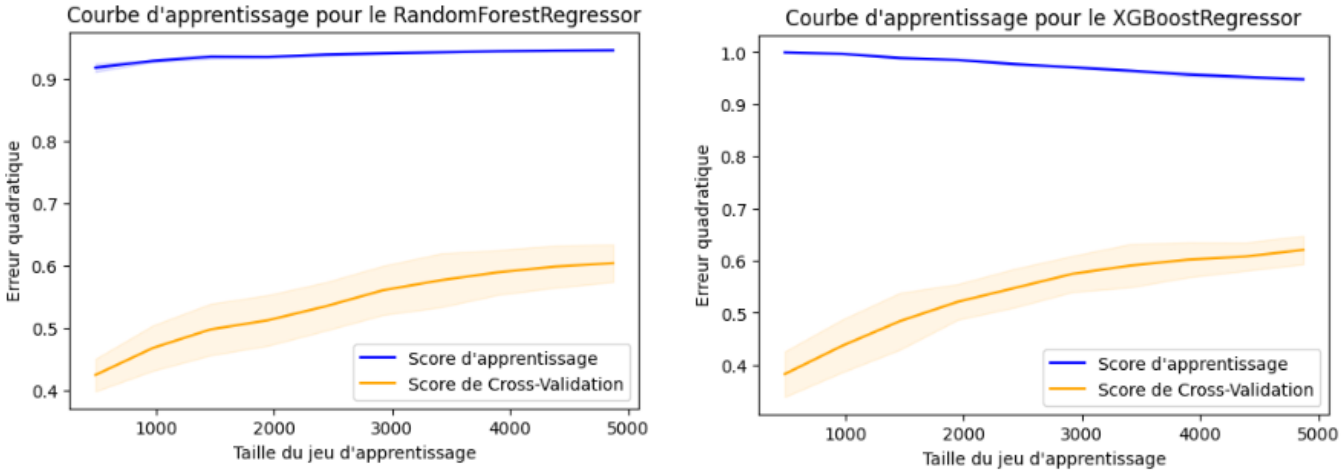
En revanche, le modèle *XGBoostRegressor* a accordé une importance plus grande à quelques zones géographiques, qui ont eu plus d'impact sur les résultats que les autres variables. Cela peut être le résultat des différences climatiques entre les différentes régions géographiques.

En somme, les résultats des deux modèles semblent assez prometteurs, mais il est important de souligner que la relation entre la température de la terre et les variables étudiées est complexe et peut être influencée par de nombreux autres facteurs externes.



Concernant les graphiques de dispersion des résidus pour les deux modèles, on a pu observer des écarts homogènes entre les valeurs prédites et les valeurs réelles, ce qui indique que les erreurs sont homogènes. De même, la répartition des points autour de la ligne de tendance (la ligne en diagonale) dans la comparaison entre les valeurs réelles et prédites indique également une homogénéité des erreurs. En d'autres termes, cela signifie que les erreurs de prédiction sont distribuées uniformément autour de la ligne ‘0’ et ne présentent pas de tendances particulières à mesure que les valeurs cibles augmentent ou diminuent, ce qui est un bon indicateur de l'exactitude des prédictions.

Pour les histogrammes des résidus des deux modèles, la distribution des écarts entre les valeurs prédites et les valeurs réelles suit une loi normale, cela indique que les résidus sont répartis uniformément autour de zéro. De même, dans la représentation *Q-Q plot* des résidus des deux modèles, les points sur le graphique tombent sur la ligne en diagonale, cela indique une distribution normale.



Afin de mieux comprendre la performance de chaque modèle en fonction de la taille de l'échantillon, on a opté pour la courbe d'apprentissage. Le but de cette analyse est de déterminer si le modèle a besoin de plus ou moins de données pour prédire avec précision. Dans chaque courbe d'apprentissage, on a deux courbes, une courbe pour l'ensemble d'apprentissage et une pour l'ensemble de test. La courbe d'apprentissage pour l'ensemble d'apprentissage montre comment la précision du modèle évolue à mesure que la taille de l'ensemble d'apprentissage augmente. La courbe d'apprentissage pour l'ensemble de test montre comment la précision évolue à mesure que la taille de l'ensemble d'apprentissage augmente.

Pour le modèle *RandomForestRegressor*, la courbe d'apprentissage pour l'ensemble d'apprentissage s'approche rapidement d'un score de 1 (qui indique une prédiction parfaite), cela peut indiquer que le modèle fait l'objet d'un sur-apprentissage - c'est-à-dire qu'il s'est trop bien adapté aux données d'apprentissage et qu'il ne généralise probablement pas bien pour les nouvelles données. Cela signifie qu'il est possible que le modèle soit moins précis lorsqu'il est confronté à des données qu'il n'a jamais vues auparavant. Les courbes d'apprentissage pour l'ensemble de test à leur tour augmentent avec une augmentation supplémentaire de la taille de l'échantillon pour les deux modèles. Cela peut indiquer que les modèles n'ont pas encore atteint leur limite de performance. Dans ce cas, l'ajout de données supplémentaires fournira peut-être des gains significatifs en termes de précision de prédiction.

7.5. Conclusions sur la modélisation par *Machine Learning*

En conclusion, on a bien raison de se concentrer sur le *RandomForestRegressor* et le *XGBoostRegressor* en raison de leurs bonnes performances. Cependant, il est difficile de dire si le modèle *XGBoostRegressor* est meilleur que le modèle *RandomForestRegressor*. Le modèle *XGBoostRegressor* accorde une importance plus grande aux zones géographiques, qui ont eu plus d'impact sur les résultats que les autres variables, tandis que le modèle *RandomForestRegressor* a accordé une plus grande importance à l'année en raison de la tendance à l'augmentation de la température de la Terre au fil des années. Le choix entre les deux modèles dépendra normalement des variables les plus importantes pour notre prédiction. Alors que dans notre cas les relations entre notre variable cible et les autres variables sont trop complexes pour pouvoir faire un choix précis.

Il est intéressant aussi de noter que le modèle *RandomForestRegressor* montre une tendance au sur-apprentissage. Cependant, les courbes d'apprentissage pour l'ensemble de test continuent de progresser avec l'augmentation de la taille de l'échantillon, ce qui indique que les modèles n'ont pas encore atteint leur limite de performance.

8. Conclusions du projet et perspectives

Le sujet de l'évolution des températures en corollaire de la production des gaz à effet de serre (GHG) est très largement étudié depuis des décennies, mettant en évidence un impact des activités humaines sur ce qu'il est commun d'appeler le réchauffement climatique.

Nous disposons de sources de données publiques fiables, combinant d'un côté les productions annuelles par pays des différents GHG et l'évolution relative de la température terrestre qui en découle, et d'autre part les données d'évolution relative de la température terrestre par pays. Ces données sont par ailleurs mesurées ou extrapolées depuis l'ère préindustrielle jusqu'à nos jours.

Via le traitement et analyse de ces données, nous pouvons mettre en avant les conclusions suivantes :

- Une augmentation des températures globales et locales est détectée sur la période considérée, de l'ordre de 1,5 °C ;
- Une augmentation massive des productions de GHG depuis le début de l'ère industrielle. Bien que le phénomène soit global sur la planète, cette production est très majoritairement liée à une quinzaine de pays qui représentent à eux-seuls les ¾ des émissions de CO_2 ;
- Dans les dernières décennies, une inflexion de la production de GHG par certains pays est détectée, sans impact détectable sur l'évolution globale ou locale de la température terrestre ;
- La production **locale** de GHG et l'évolution **locale** relative de la température terrestre ne sont pas corrélées ;
- Nos prédictions par *Machine Learning* de l'évolution **locale** relative de la température terrestre en fonction de la production **locale** de GHG sur la période considérée ont montré un succès relatif, avec les modèles *RandomForestRegressor* et *XGBoostRegressor*.

Au-delà de cette étude, il est possible d'envisager de multiples développements :

- Utiliser les mesures atmosphériques disponibles des différents GHG pour recherche de corrélation et prédiction d'évolution de températures terrestres ;
- Utiliser des regroupements de pays sur d'autres critères que géographiques (par le niveau de développement, profils de production des GHG, …) ;
- Obtenir une prédiction meilleure pour les GHG hors CO_2, uniquement pour les périodes récentes et à venir, ces données étant largement manquantes par pays à part pour les dernières décennies ;
- Prendre en compte de multiples scénarios de stabilisation de la température terrestre globale (tels que ratifiés dans l'accord de Paris), que ce soit sur les années ou décennies futures comme récentes (cf. inflexion détectée depuis quelques décennies de production de certains GHG par certains pays).