GROUPE TOURISMECO

RAPPORT DE SCIENCES DE DONNÉES

BRANSOLLE Line, GILIBERT Rémy, GONÇALVES Hugo, MOHAMEDATNI Aya,
SÉNÉCAILLE Cassandra, TRIOZON Lucas

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Avril 2024

Soumis comme contribution partielle
pour le cours Sciences de Données

Table des matières

Introduction

Ce rapport présente une analyse approfondie des tendances mondiales du tourisme en utilisant des techniques d'analyse de données avancées. Notre objectif est d'identifier des relations significatives entre différentes variables et de faire des prédictions précises pour aider à la prise de décisions. Nous commencerons par développer un score unique pour chaque pays. Ensuite, nous utiliserons l'Analyse en Composantes Principales (ACP) pour explorer les corrélations entre les différentes variables de notre ensemble de données. Nous appliquerons également des techniques de clustering pour regrouper les pays en fonction de leurs caractéristiques économiques, environnementales et touristiques. Par la suite, nous utiliserons la régression linéaire pour révéler des relations intéressantes entre différentes variables. Enfin, nous appliquerons le modèle ARIMA pour modéliser et prévoir les séries temporelles, fournissant ainsi des informations précieuses pour la prise de décisions.

Pour faire cette analyse nous avons utilisé différentes variables de notre base, dont une grande partie nous a permis de créer notre score :

- **Émissions de CO2 :** Mesure des émissions de dioxyde de carbone (CO2) dans l'atmosphère, exprimée par habitant et/ou par année.
- **Arrivées Touristiques Totales :** Nombre total d'arrivées de touristes dans le pays au cours d'une année donnée.
- **Arrivées Touristiques par Avion :** Nombre d'arrivées de touristes par avion dans le pays au cours d'une année donnée.
- **Produit Intérieur Brut (PIB) :** Mesure de la production économique totale d'un pays par année.
- **Émissions de Gaz à Effet de Serre (GES) :** Mesure des émissions totales de gaz à effet de serre en tonnes d'équivalent CO2. Il comprend, le dioxyde de carbone, le méthane et le protoxyde d'azote.
- **Indice de GPI :** Le GPI (Global Peace Index), il classe les pays selon leur degré de pacifisme. Plus le score est bas, plus le pays est paisible.
- **Les énergies renouvelables** : mesure en pourcentage la consommation totale du pays en énergie renouvelable par rapport à la consommation totale d'énergie.
- **L'IDH** (Indice de Développement Humain): Evalue le niveau de développement humain du pays sur une échelle de 0 à 1.

Chapitre 1

Score

1.1 Objectif du score

Nous avons créé un score distinct pour chaque pays afin d'ajouter une dimension supplémentaire à notre projet. En effet, en créant un score selon des critères qui suivent l'éthique de notre site, nous permettons à l'utilisateur de s'informer simplement sur sa future destination selon des critères qui lui sont personnels. Pour cela, nous avons mis en place quatre scores. Un score plutôt global, qu'on a appelé le score Tourismeco, qui s'affichera par défaut pour les personnes qui n'ont pas de préférences et qui veulent une vision d'ensemble du pays ou alors pour des personnes qui n'ont pas de compte sur notre site. En effet, en créant un compte et en se connectant, les utilisateurs peuvent personnaliser leurs préférences. Ils ont le choix entre trois scores : le Tourisme moderne, le Tourisme Éco-responsable et le Tourisme d'Exploration et de Découverte.

Le premier score TourismEco se base sur 6 critères, que l'on retrouvera dans les 3 autres scores :

- Les arrivées totales d'avions dans le pays.
- Le GPI
- Le PIB par habitant
- Les GES par habitant
- Les énergies renouvelables
- L'IDH

Ensuite, le score "Tourisme Éco-responsable" vise à apporter plus de précision aux utilisateurs sur son empreinte écologique en se basant sur deux critères : les GES produits par habitant et la quantité d'énergie renouvelable utilisée par le pays par rapport à sa consommation d'énergie générale. Puis, le score "Tourisme Moderne" permet de savoir si le pays est développé, si les conditions humaines sont bonnes et s'il est populaire touristiquement parlant. Pour cela, nous nous sommes basés sur trois critères : les arrivées totales d'avions dans le pays, le PIB par habitant et l'IDH. Pour finir, le score "Tourisme d'Exploration et de Découverte" permet de savoir si le pays est touristique, tout en gardant une conscience écologique et un bon niveau de développement humain. Pour cela, nous nous sommes basés sur trois critères : les arrivées totales d'avions dans le pays, les GES par habitant et l'IDH.

Notre score est compris entre 0 et 1, et il est noté par une lettre allant de A à E, avec 'A' pour dire que le score est parfait et 'E' pour dire que le score est mauvais.

1.2 Mise en place du score

Pour mettre en place nos scores, nous avons utilisé le langage de programmation Python. Nous avons d'abord traité les données en nous connectant à la base de données, puis nous avons créé quatre dataframes, un pour chaque score, en ne conservant que les colonnes pertinentes pour le calcul du score correspondant, en utilisant la fonction `pd.DataFrame()`. Puis nous avons normalisé les données grâce à la méthode multicritère, c'est une méthode de comparaison. Elle permet de déterminer le meilleur élément dans une situation précise. En d'autres termes, elles nous a permis d'apporter un poids différents à chaque critères, certain permettent de faire augmenter le score et d'autres le font diminuer, le résultat est compris entre 0 et 1:

- Normalisation des données
  - Variables qui le font augmenter: arrivées total d'avions, le PIB par habitant, les énergies renouvelables et l'IDH.
  - Exemple:

    ```
    df['pibParHab'] = df['pibParHab'] / df['pibParHab'].max()
    ```

    Plus la valeur de la variable est élevée plus le score est bon.

    $$\frac{\text{valeur du critère de l'élément}}{\text{plus grande valeur sur ce critère de tous les éléments comparés}}$$
  - Variables qui le font diminuer: GPI et les GES par habitant.
  - Exemple:

    ```
    df['GES_hab'] = df['GES_hab'].min() / df['GES_hab']
    ```

    Plus la valeur du critère est basse plus le score est bon.

    $$\frac{\text{plus petite valeur sur ce critère de tous les éléments comparés}}{\text{valeur du critère de l'élément}}$$

\vskip 3em

Ensuite, nous avons attribué des poids à chaque variable, puis nous avons utilisé une moyenne pondérée pour obtenir le score final : \ `df['score'] = df.apply(lambda row: sum(float(row[var]) * poids[var] for var in poids) / sum(poids.values()), axis=1).` \ Puis, pour définir les intervalles du score pour définir les lettres nous l'avons séquencé par la variance, en partant du score minimum (E) et terminant par le score maximum (A). Nous avons appliqué ces intervalles avec la fonction `pd.cut()`. Enfin, pour vérifier si nos scores étaient cohérents, nous avons utilisé une visualisation par une carte du monde avec les fonctions : `px.choropleth()` et `fig.update_geos()`.

1.3 Résultats du score

Grâce à des visualisations par des histogrammes et des cartes mondiales, nous pouvons observer la distribution des scores. Après visualisation, on observe que très peu se situent aux deux extrêmes. Pour le score touristique qui prend en compte toutes les variables, seulement un pays possède un score de catégorie E, et 12 ont un score de catégorie A, sachant que 130 pays sont comptabilisés. De manière plus générale, c'est surtout le score de catégorie A qui comporte très peu de pays pour tous les scores confondus. Cependant, le score de catégorie D est celui

où l'on retrouve le plus de pays, montrant qu'une majorité des pays sont en dessous de C, qui est le score médian. Une majorité a un score plutôt mauvais dans les quatre scores.

Malgré tout, on remarque que le score le plus équilibré est le score touristique (cf. figure 1.1 et figure 1.2).

Chapitre 2

ACP

2.1 Analyse en Composantes Principales (ACP)

Après avoir constitué notre base de données, nous avons entrepris de comprendre et d'analyser nos variables. Nous disposons d'un ensemble de données comprenant des variables écologiques, économiques et touristiques pour les pays du monde sur une période d'environ une trentaine d'années. L'objectif est de comprendre les relations entre ces variables et d'identifier les tendances significatives qui pourraient influencer notre compréhension du tourisme mondial. Pour cela, nous avons décidé d'effectuer une Analyse en Composantes Principales (ACP), une méthode statistique abordée au semestre 5 et 6, qui vise à explorer les corrélations entre les variables.

L'ACP est une technique analytique puissante qui permet d'explorer les corrélations entre les différentes variables d'un ensemble de données. En réduisant la dimension des données tout en préservant au mieux les informations importantes, l'ACP nous offre la possibilité de découvrir les structures sous-jacentes et les relations cachées entre les variables. Chaque composante principale est associée à un vecteur propre et à une valeur propre. Le cercle de corrélation est une représentation graphique qui permet d'interpréter les variables par rapport à ces composantes principales. Dans le cercle de corrélation, chaque variable initiale est représentée par un vecteur, et l'angle entre ces vecteurs indique leur corrélation avec chaque composante principale. Les variables ayant des vecteurs proches de l'axe de la composante principale sont fortement corrélées avec cette composante, tandis que celles dont les vecteurs sont perpendiculaires ont une corrélation faible (cf. figure 1.3).

2.2 Variables Utilisées dans l'Analyse

Dans cette analyse, nous examinons plusieurs variables clés qui couvrent un large spectre des aspects écologiques, économiques et touristiques des pays du monde.

- **Émissions de CO2 :**
  - ***Interprétation*** : Des valeurs élevées indiquent une forte empreinte carbone du pays, souvent associée à une utilisation intensive des combustibles fossiles et des activités industrielles.
- **Arrivées Touristiques Totales :**
  - ***Interprétation*** : Des valeurs élevées indiquent une attractivité touristique élevée du pays, généralement associée à une infrastructure touristique développée.
- **Arrivées Touristiques par Avion :**
  - ***Interprétation*** : Des valeurs élevées peuvent indiquer une accessibilité internationale accrue du pays et un tourisme de loisirs ou d'affaires important.

- **Produit Intérieur Brut (PIB) :**
  - ***Interprétation*** : Un PIB élevé reflète une économie forte et diversifiée, avec un niveau de vie généralement plus élevé pour ses habitants.
- **Émissions de Gaz à Effet de Serre (GES) :**
  - ***Interprétation*** : Des valeurs élevées indiquent une forte empreinte carbone du pays, avec des implications potentielles pour le changement climatique et l'environnement.
- **Indice de GPI :**
  - ***Interprétation*** : Un GPI faible indique un niveau pacifisme élevé.

Il est important de noter que le choix des variables peut influencer les résultats de l'ACP. En incluant ou excluant certaines variables, nous pourrions observer des variations dans les structures des composantes principales et dans les interprétations qui en découlent. Nous avons choisi de réaliser une ACP sur ces variables pour obtenir une vue d'ensemble des relations entre les aspects écologiques, économiques et touristiques des pays du monde. Cela nous permet d'identifier les tendances générales et les structures sous-jacentes des données.

2.3 Calcul du Ratio et Visualisation

Nous avons calculé le ratio des deux premières valeurs propres par rapport à la somme totale des valeurs propres, ce qui nous donne un ratio de 65.82%. Ce ratio indique la proportion de variance des données qui est expliquée par les deux premières composantes principales. Nous avons également représenté graphiquement les corrélations entre les variables dans un cercle des corrélations. Ce graphique permet de visualiser les relations entre les variables et les deux premières composantes principales.

2.4 Résultats de l'ACP et Interprétations

- **Corrélation positive entre les émissions de CO2, les arrivées touristiques totales et les arrivées touristiques par avion :**

Cette corrélation suggère qu'il existe une tendance où les pays qui accueillent un plus grand nombre de touristes, notamment par avion, ont tendance à générer davantage d'émissions de CO2. Cela pourrait être dû à l'utilisation des transports et à d'autres activités associées au tourisme.

- **Corrélation positive entre le PIB et les émissions de gaz à effet de serre (GES) :**

Cette corrélation indique que les pays avec un PIB plus élevé ont tendance à émettre plus de GES. Cela peut être attribué à une grande production et consommation dans les secteurs industriels et commerciaux, qui sont souvent associés à des émissions de GES.

- **Corrélation négative entre l'Indice de GPI et les émissions de GES par habitant :**

Cette corrélation suggère que les pays avec un Indice de GPI plus élevé ont tendance à avoir des émissions de GES par habitant plus faibles.

- **Corrélation négative entre l'Indice de GPI et le PIB par habitant :**

Cette corrélation met en évidence une relation inverse entre l'Indice de GPI et le PIB par habitant. Cela suggère que les pays avec un GPI élevé ont tendance à avoir un PIB par habitant plus faible.

- **Absence de corrélation entre l'Indice de GPI et les arrivées touristiques totales :**

L'absence de corrélation entre l'Indice de GPI et les arrivées touristiques totales suggère que les touristes ne choisissent pas nécessairement leurs destinations en fonction du pacifisme du pays. D'autres facteurs tels que l'attrait touristique, les infrastructures et les coûts peuvent jouer un rôle plus important dans le choix des destinations.

- **Absence de corrélation entre les émissions de GES par habitant et les émissions de GES :**

Cette absence de corrélation peut sembler contre-intuitive, mais cela peut indiquer que certaines variables non incluses dans l'analyse pourraient influencer les émissions totales de GES. Par exemple, des facteurs tels que la taille de la population ou les politiques environnementales pourraient jouer un rôle dans la quantité totale d'émissions de GES, indépendamment des émissions par habitant.

2.5 Conclusion de l'ACP

Cette analyse initiale nous offre un aperçu significatif des relations entre les variables écologiques, économiques et touristiques à l'échelle mondiale. Les découvertes présentées ici fournissent une base solide pour des analyses plus approfondies et des recommandations futures dans le domaine de la politique internationale, du développement économique et de la durabilité environnementale.

2.6 Perspectives Futures

Pour approfondir notre compréhension, des analyses supplémentaires pourraient explorer les tendances temporelles, les interactions non linéaires entre les variables et les implications politiques et économiques des résultats présentés ici.

Chapitre 3

Clustering

3.1 Clustering

Le clustering est une technique d'apprentissage non supervisé qui consiste à regrouper des ensembles de données similaires ou proches les uns des autres. Après avoir constitué notre base de données contenant des informations sur divers pays, nous avons entrepris une analyse de clustering pour découvrir les similarités entre eux. Ce rapport explique le processus de clustering appliqué à notre base de données pour identifier des pays similaires. L'objectif était de regrouper les pays en fonction de leurs caractéristiques communes.

3.2 Préparation des données

Les données ont été extraites de la base de données en utilisant Python. Nous avons sélectionné les variables pertinentes pour chaque pays pour l'année 2019:

- PIB par habitant
- Indice de Global Peace
- Indice de Développement Humain
- Arrivées totales
- Émission de CO2
- Émissions de GES par habitant

Les valeurs manquantes ont été supprimées pour assurer la qualité des données. Ensuite, les données ont été normalisées à l'aide de la méthode Min-Max pour les rendre comparables.

3.3 Clustering avec KMeans et Visualisation

En utilisant l'algorithme KMeans avec différents nombres de clusters, nous avons regroupé les pays en fonction de leurs similitudes dans ces caractéristiques. Nous avons visualisé les clusters en utilisant deux méthodes. Tout d'abord, une carte a été générée pour montrer la répartition géographique des clusters à travers le monde (cf. figure 1.5 et figure 1.7). Ensuite, une figure tridimensionnelle a été créée pour représenter les clusters dans un espace tridimensionnel défini par le PIB par habitant, les émissions de CO2 par habitant et le GPI (cf. figure 1.4 et figure 1.6).

3.4 Interprétation des Clusters

Carte avec 3 Clusters :

- Cluster 1 : Regroupe l'Amérique du Sud, une partie de l'Asie et quelques pays d'Afrique du Nord (y compris l'Afrique du Sud).
- Cluster 2 : Inclut l'Amérique du Nord, une grande partie de l'Europe et l'Océanie.
- Cluster 3 : Rassemble les pays d'Afrique centrale, d'Amérique centrale et d'Asie du Sud. Les clusters démontrent des regroupements géographiques et économiques intéressants, bien visualisés dans la figure tridimensionnelle.

Carte avec 8 Clusters :

- Les clusters sont encore plus détaillés, répartissant les pays en groupes plus spécifiques selon leurs caractéristiques communes.
- Par exemple, un cluster peut regrouper l'Australie, le Canada et les pays d'Europe de l'Est, tandis qu'un autre peut inclure certains pays d'Amérique du Sud, l'Algérie et des pays de l'Europe du Nord avec la Russie.

3.5 Conclusion du Clustering

Le clustering des pays en fonction de leurs caractéristiques économiques, environnementales et touristiques offre des perspectives intéressantes sur les similarités et les différences entre les nations. Ces analyses fournissent des informations utiles pour la compréhension des tendances mondiales et peuvent être utilisées pour recommander un pays à partir d'un autre pays. Cette analyse se base sur un ensemble spécifique de variables, mais il est important de noter que l'ajout ou la suppression de certaines variables peuvent affiner ou élargir les résultats, les rendant ainsi plus pertinents pour d'autres domaines ou perspectives.

3.6 Perspectives Futures

Pour approfondir notre compréhension, des analyses supplémentaires pourraient être explorées. Ces analyses pourraient également tester différentes combinaisons de variables pour mieux comprendre les facteurs sous-jacents qui influencent les regroupements des pays. Une telle approche permettrait une meilleure contextualisation des résultats et une application plus précise dans divers domaines de recherche. En examinant de près les relations entre les variables et en explorant des méthodes analytiques avancées telles que l'apprentissage automatique et l'analyse de réseaux, nous pourrions découvrir des schémas et des tendances encore plus nuancés.

Chapitre 4

Régression linéaire

4.1 Introduction

La régression linéaire est une approche utile pour prédire ou estimer des valeurs en se basant sur une relation linéaire entre une variable dépendante et une ou plusieurs variables explicatives.

Notre approche consiste à trouver des relations entre différentes variables de notre base de données. Pour qu'elles soient des variables de prédictions intéressantes, à mettre en place. Pour cela nous verrons dans un premier temps, le PIB comme variable explicative, puis, nous explorerons d'autres variables supplémentaires comme variable explicative.

Pour mettre en place cette régression linéaire, nous avons utilisé Python. Nous avons utilisé des fonctions et méthodes similaires pour chaque régression, en prenant à chaque fois des données sur un intervalle de 2 ans, pour certaines données que nous avions jusqu'à 2022 nous avons pris de 2020 à 2022, dans l'autre cas nous avons pris des données de 2018 à 2020 :

Tout d'abord, nous utilisons la fonction `DataFrame()` pour créer un tableau à partir des données fournies. Ensuite, afin de garantir la qualité de nos données, nous utilisons la méthode `dropna()` pour supprimer les valeurs NULL. Dans certains cas, il est nécessaire de prétraiter les données pour les adapter au modèle. Cela peut impliquer de convertir des colonnes de type chaîne de caractères en type flottant à l'aide de la fonction `astype(float)`, ou de remplacer des virgules par des points avec la fonction `replace()`. Une fois les données prétraitées, nous sélectionnons les colonnes pertinentes et nous les divisons en ensembles d'entraînement et de test en utilisant la fonction `train_test_split()`. Nous créons ensuite notre modèle de régression linéaire en utilisant `model = LinearRegression()` et nous l'entraînons sur les données d'entraînement en utilisant `model.fit(X_train, y_train)`. Une fois le modèle entraîné, nous utilisons la fonction `predict()` pour effectuer des prédictions sur la variable dépendante (y). Pour évaluer la performance de notre modèle, nous utilisons la fonction `score()`. Enfin, nous visualisons nos données en créant un graphique qui représente à la fois les données réelles et la droite de régression linéaire. Pour cela, nous utilisons les fonctions `scatter()` et `plot()`.

4.2 Le PIB comme variable explicative

Nous avons utilisé le PIB par habitant comme variable explicative dans notre régression linéaire. Nous avons observé une corrélation positive entre le PIB par habitant et le Revenu National Brut (RNB) par habitant, ce qui suggère que le PIB par habitant est un prédicteur efficace du RNB par habitant ($R^2 = 0.98$). (cf.figure 1.8)

Nous avons également constaté une corrélation positive entre le PIB par habitant et l'IDH (Indice de Développement Humain), indiquant que lorsque les conditions de vie s'améliorent, cela peut entraîner une augmentation des revenus et de la richesse au niveau individuel et national ($R^2 = 0.68$).

Enfin, nous avons observé une corrélation positive, mais plus faible, entre le PIB par habitant et les émissions de GES par habitant. Cela suggère que d'autres variables doivent être prises en compte pour une prédiction plus précise des émissions de GES ($R^2 = 0.187$).

En conclusion, le PIB par habitant peut être une variable intéressante pour définir des prédictions, mais il a ses limites. Par exemple, pour les GES, le PIB permet de prédire en partie les GES, mais pas entièrement. Dans certains cas, la régression linéaire n'est pas suffisante, et il faut utiliser la régression multidimensionnelle.

4.3 Exploration des variables supplémentaires

Nous avons exploré d'autres variables explicatives pour notre régression linéaire.

1. **IDH et espérance de vie** : Nous avons observé une forte corrélation positive entre l'IDH et l'espérance de vie, confirmée par un score R^2 de 0.794. Cela suggère que l'IDH est un bon prédicteur de l'espérance de vie.(cf. figure 1.9)
2. **IDH et GPI** : Nous avons constaté une corrélation négative entre l'IDH et le GPI, indiquant que plus le pays est paisible (GPI faible), plus le développement du pays (IDH) est important. Cependant, le score R^2 de 0.376 montre que cette relation est moins forte.
3. **Énergie renouvelable et GES par habitant** : Nous avons observé une corrélation négative entre l'énergie renouvelable et les GES par habitant, indiquant que plus il y a d'énergie renouvelable mise en place dans le pays, plus le GES par habitant est faible. Cependant, cette relation est faible (score R^2 de 0.125), en partie à cause de deux valeurs aberrantes.

Ces résultats montrent que, bien que le PIB par habitant soit une variable intéressante pour définir des prédictions, d'autres variables peuvent également être utiles, malgré certaines limites."

4.4 Conclusion

En conclusion, l'analyse des relations entre le PIB par habitant et différentes variables explicatives, ainsi que l'exploration de variables supplémentaires, ont fourni des résultats significatifs.

Premièrement, le PIB par habitant s'est révélé être un prédicteur efficace du Revenu National Brut par habitant et de l'Indice de Développement Humain. Cela souligne son importance en tant qu'indicateur clé de la richesse et du développement économique d'un pays. Bien que sa performance dans la prédiction des émissions de Gaz à Effet de Serre (GES) par habitant soit légèrement moins robuste, le PIB par habitant a néanmoins présenté une corrélation positive avec ces émissions.

Deuxièmement, l'exploration de variables supplémentaires telles que l'IDH, l'espérance de vie, le GPI (Global Peace Index) et l'énergie renouvelable a permis de mettre en lumière des relations intéressantes. L'IDH s'est avéré être fortement corrélé à l'espérance de vie, renforçant ainsi l'idée que le développement humain est étroitement lié à des aspects tels que la santé et l'éducation. De plus, malgré une corrélation négative relativement faible entre l'IDH et le GPI, cette relation offre néanmoins des pistes de réflexion pour une compréhension plus approfondie des facteurs influençant le développement socio-économique.

Enfin, la corrélation négative entre l'énergie renouvelable et les émissions de GES par habitant, bien que légère, souligne l'importance croissante des politiques énergétiques durables dans la réduction des impacts environnementaux.

Cette analyse met en évidence les différentes variables utilisées dans notre projet et montre comment ces variables peuvent avoir un impact significatif dans une prédiction. Cela nous permet de mieux comprendre les tendances actuelles et futures.

Chapitre 5

Arima

5.1 Introduction aux séries temporelles et ARIMA

5.1.1 Introduction générale aux séries temporelles

Les séries temporelles représentent un aspect intéressant de l'analyse des données, offrant un aperçu des phénomènes qui évoluent avec le temps. Avant de plonger dans les détails du modèle ARIMA, il est important de comprendre les fondamentaux des séries temporelles. Les séries temporelles sont des ensembles de données indexées par le temps, reflétant ainsi l'évolution d'un phénomène au fil du temps. Ces données jouent un rôle important dans de nombreux domaines, de la finance à la météorologie, en passant par l'économie et la santé publique. Leur analyse et leur prédiction ont une importance pour la prise de décisions. Les séries temporelles sont décomposées en trois composantes principales :

- Tendance (Tt) : Représente la variation à long terme de la série, pouvant être croissante, décroissante ou stable.
- Saisonnalité (St) : Indique les variations périodiques qui se répètent à des intervalles réguliers, comme les saisons ou les cycles économiques.
- Résidu ou erreur (εt) : Comprend les fluctuations aléatoires non expliquées par la tendance ou la saisonnalité.

Cette décomposition permet une compréhension approfondie de la structure des données temporelles, facilitant ainsi la modélisation et la prédiction.

5.1.2 Introduction à ARIMA

L'approche ARIMA (AutoRegressive Integrated Moving Average) est un modèle statistique puissant utilisé pour analyser et prévoir les séries temporelles. ARIMA tire son efficacité de la combinaison de trois concepts principaux : l'autorégression (AR), l'intégration (I) et la moyenne mobile (MA). Comprendre ces composantes est essentiel pour maîtriser ARIMA et exploiter ses capacités dans l'analyse des séries temporelles.

- Autorégression (AR) met l'accent sur la dépendance des observations actuelles avec les observations précédentes dans la série temporelle. Un modèle utilise les 'p' observations précédentes pour prédire la prochaine observation.
- Intégration (I) est une étape cruciale pour rendre la série temporelle stationnaire. Une série temporelle stationnaire est caractérisée par des propriétés constantes au fil du temps telles que la moyenne et la variance. La différenciation, qui est une forme d'intégration, est appliquée à la série temporelle brute pour éliminer les tendances et les structures de dépendance temporelle. Le degré de différenciation, noté par 'd', indique le nombre de différences successives nécessaires pour rendre la série temporelle stationnaire.

- Moyenne mobile (MA) modélise la relation entre une observation et une moyenne mobile de ses erreurs résiduelles précédentes. L'ordre de la moyenne mobile, noté par 'q', spécifie la taille de la fenêtre de la moyenne mobile utilisée dans le modèle. Un modèle utilise les 'q' erreurs résiduelles précédentes pour prédire la prochaine observation. En définissant les paramètres 'p', 'd' et 'q', le modèle ARIMA peut s'adapter aux structures temporelles spécifiques des données, capturant à la fois les tendances à long terme et les variations aléatoires. Cette flexibilité permet à ARIMA de modéliser une grande variété de séries temporelles avec précision. En comprenant ces concepts fondamentaux, nous sommes prêts à explorer plus en profondeur le fonctionnement et les applications pratiques du modèle ARIMA dans l'analyse et la prévision des séries temporelles.

5.2 Configuration et évaluation des modèles ARIMA

Le processus de configuration et d'évaluation des modèles ARIMA est essentiel pour obtenir des prévisions précises et fiables des séries temporelles. Cette partie aborde les étapes clés de ce processus avec le paramétrage initial du modèle et l'évaluation de sa performance.

5.2.1 Paramétrage et évaluation du modèle ARIMA

La configuration d'un modèle ARIMA implique de choisir les valeurs appropriées pour les paramètres p, d et q, qui déterminent l'ordre de l'auto régression, le degré de différenciation et l'ordre de la moyenne mobile, respectivement. Cependant, estimer ces paramètres peut être un défi, nécessitant souvent des méthodes d'essais et d'erreurs itératifs. Cela implique d'ajuster différents modèles ARIMA avec des combinaisons de paramètres et de sélectionner celui qui minimise les erreurs de performance, telles que l'erreur quadratique moyenne (RMSE) ou l'erreur absolue moyenne (MAE).

5.2.2 Recherche en grille des hyperparamètres ARIMA

Pour simplifier et accélérer le processus de configuration des modèles ARIMA, l'automatisation à l'aide de la recherche en grille est souvent utilisée. Cette approche implique de spécifier une grille de valeurs pour les paramètres p, d et q, puis d'évaluer la performance de chaque combinaison de paramètres à l'aide d'une métrique de performance définie. Une fois que les modèles ARIMA ont été configurés, il est essentiel d'évaluer leur performance avant de les utiliser pour faire des prévisions. Cela implique de diviser les données en ensembles de formation et de test, d'ajuster les modèles sur l'ensemble de formation, de faire des prévisions sur l'ensemble de test et d'évaluer la précision des prévisions à l'aide de mesures d'évaluation appropriées. L'évaluation des modèles ARIMA peut être réalisée en utilisant des bibliothèques Python telles que statsmodels et pandas. Ces bibliothèques offrent des fonctionnalités pour ajuster les modèles ARIMA, faire des prévisions et évaluer la performance des modèles à l'aide de différentes métriques. La recherche en grille des hyperparamètres ARIMA est une approche systématique pour trouver les valeurs optimales des paramètres p, d et q. En résumé cette procédure implique de spécifier une grille de valeurs pour chaque paramètre, d'évaluer les performances de chaque combinaison de paramètres et de sélectionner celle qui donne les meilleures performances prévisionnelles.

5.3 Application et prédictions avec ARIMA

5.3.1 Mise en place du modèle ARIMA

Une fois que le modèle ARIMA a été configuré et ajusté aux données d'entraînement, il est prêt à être utilisé pour faire des prédictions sur l'échantillon de test. Les données portent sur le pourcentage de production d'énergies renouvelables aux États-Unis de 1995 à 2020. L'utilisation du modèle ARIMA pour faire des prédictions implique plusieurs étapes. La première étape consiste à importer les bibliothèques nécessaires. Dans ce cas, la bibliothèque sklearn.metrics est utilisée pour calculer le score R^2 et l'erreur quadratique moyenne (RMSE), qui sont des mesures couramment utilisées pour évaluer la précision des prédictions. Les données sont divisées en un ensemble d'entraînement et un ensemble de test. L'ensemble d'entraînement contient environ 80 % des données, tandis que l'ensemble de test contient les 20 % restants. Cette division permet d'évaluer la performance du modèle sur des données qu'il n'a pas vues pendant l'entraînement. Deux listes sont initialisées : history, qui contient les valeurs de la série temporelle jusqu'à présent, et predictions, qui contiendra les prédictions du modèle. Ces listes sont utilisées pour stocker les données réelles et prédites au fur et à mesure que le modèle est évalué sur l'ensemble de test.

5.3.2 Prédictions avec ARIMA

Après que le modèle ARIMA soit ajusté aux données d'entraînement, il peut être utilisé pour faire des prédictions. Pour ce faire, nous pouvons utiliser la fonction `ARIMA.predict()` pour générer des prévisions. Pour chaque donnée dans l'ensemble de test, le modèle ARIMA génère une prévision pour le prochain point de données, et cette prévision est ajoutée à la liste des prédictions. Ensuite, la valeur réelle correspondante est ajoutée à la liste d'historique. Ce processus se poursuit jusqu'à ce que toutes les données de l'ensemble de test soient prédites. Une fois que toutes les prédictions ont été générées, l'erreur quadratique moyenne (RMSE) et le score R^2 sont calculés entre les prédictions et les valeurs réelles de l'ensemble de test. Ces mesures fournissent une indication de la précision du modèle par rapport aux données observées. Le coefficient de détermination R^2 est utilisé pour évaluer la qualité de l'ajustement du modèle aux données. Nous obtenons un R^2 de 0,88 ce qui indique que le modèle explique environ 88 % de la variance dans les données. En d'autres termes, il montre que le modèle ARIMA est assez bon pour représenter et prédire la tendance pour cet ensemble de donnée. Enfin, les résultats sont affichés à l'aide d'un graphique qui compare les valeurs réelles de l'ensemble de test aux prédictions du modèle. Cela permet de visualiser la performance du modèle (cf. figure 1.10 et figure 1.11).

5.3.3 Prédictions hors échantillon

Une fois que les paramètre du modèle ARIMA ont été configuré et que les prédictions sur l'échantillon de test ont été jugé bon à l'aide du calcul des erreurs métriques. Nous pouvons faire des prédictions hors échantillon, ce sont des prévisions faites pour des périodes futures qui ne font pas partie de l'ensemble de données d'entraînement initial. Ces prédictions sont importantes pour évaluer la capacité du modèle ARIMA à généraliser de nouvelles données et à maintenir sa performance prédictive dans des conditions réelles. Pour faire des prédictions hors échantillon, nous pouvons utiliser la fonction `ARIMA.forecast()` dans les bibliothèques Python comme statsmodels. Cette fonction permet de générer des prédictions pour un certain nombre de pas de temps dans le futur. Pour cela de nombreuses étapes sont utilisées, la

fonction `difference(dataset)` calcule la différence entre chaque point de données et le point de données précédent, nous l'avons utilisée pour rendre une série temporelle stationnaire. Un modèle ARIMA est créé avec l'ordre déterminé lors du processus de paramétrage du modèle et la fonction `model_fit = model.fit()` est utilisée pour l'apprentissage supervisé du modèle ARIMA sur les données ajustées. Des prévisions sont générées pour un nombre spécifié de pas de temps dans le futur avec `forecast = model_fit.forecast(steps=7)` . Chaque prévision est inversée avec la fonction `inverse_difference(history, prediction)` qui inverse la différenciation de départ appliquée à la série temporelle, transformant ainsi les prédictions faites sur la série différenciée en prédictions sur la série originale. Les prévisions sont ensuite ajoutée à l'historique, et une nouvelle année est crée. Finalement, les résultats sont affichés à l'aide d'un graphique où l'historique qui comprend maintenant les prévisions, est tracé en bleu et la série temporelle de départ est tracé en rouge.

En conclusion, le graphique présente en rouge le pourcentage de production d'énergie renouvelable aux États-Unis de 1995 à 2020, tandis que les prévisions du pourcentage d'énergie renouvelable pour les sept années suivantes sont représentées en bleu (cf. figure 1.12). En outre, il est important de noter que ces données peuvent être étendues pour inclure d'autres pays et d'autres années. Cela signifie que le modèle ARIMA peut être adapté pour étudier et prédire les tendances de la production d'énergies renouvelables à l'échelle mondiale et sur des périodes plus étendues.

Conclusion

Ce projet a souligné l'importance de l'analyse des données pour comprendre les tendances mondiales du tourisme. Nous avons utilisé des techniques d'analyse de données avancées, comme la régression linéaire, le clustering et l'ARIMA, pour identifier des relations significatives entre différentes variables et faire des prédictions précises.

La régression linéaire a révélé des relations intéressantes entre le PIB par habitant et d'autres variables, comme le RNB par habitant, l'IDH et les émissions de GES par habitant. L'exploration de variables supplémentaires, comme l'IDH, l'espérance de vie, le GPI et l'énergie renouvelable, a également mis en lumière des relations intéressantes.

L'ACP a permis d'explorer les corrélations entre les différentes variables de notre ensemble de données, offrant ainsi une compréhension approfondie de la structure des données.

Le clustering a permis de regrouper les pays en fonction de leurs caractéristiques économiques, environnementales et touristiques, offrant des perspectives intéressantes sur les similarités et les différences entre les nations.

Enfin, l'ARIMA a permis de modéliser et de prévoir les séries temporelles, fournissant des informations précieuses pour la prise de décisions.

Grâce à cette analyse nous avons pu mieux comprendre l'impact de chaque variable pour ensuite créer notre score.

Dans l'ensemble, ce projet a démontré l'importance de l'analyse des données pour comprendre les tendances mondiales du tourisme et fournit une base solide pour des analyses plus approfondies et des recommandations futures dans différents domaines.

Annexe

1.1 Score

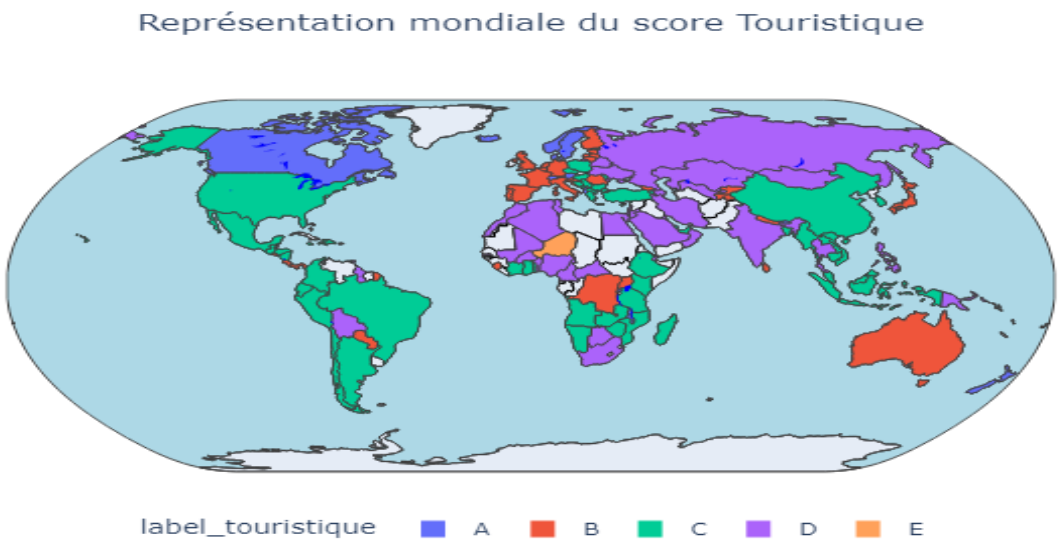

Figure 1.1: Représentation mondiale du Score Touristique

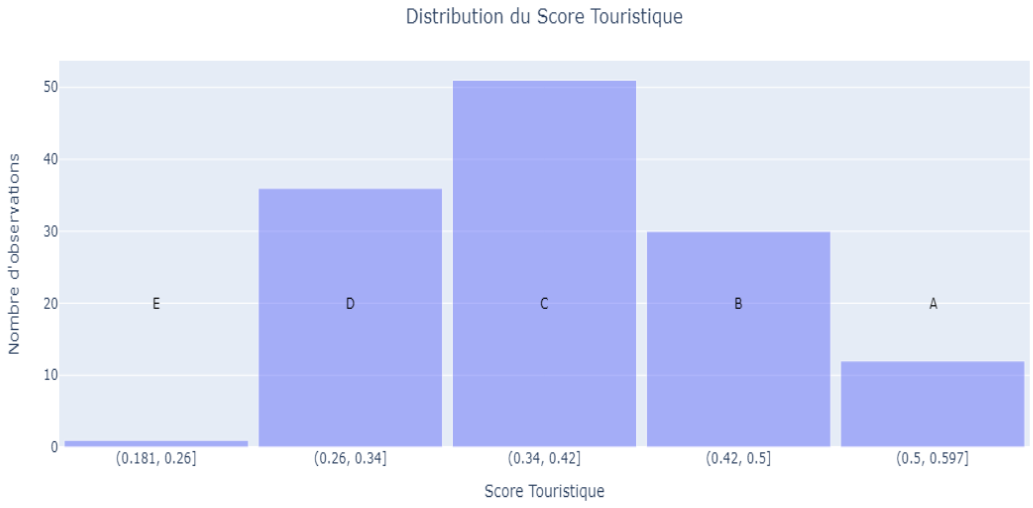

Figure 1.2: Distribution du Score Touristique

1.2 Analyse en Composantes Principales (ACP)

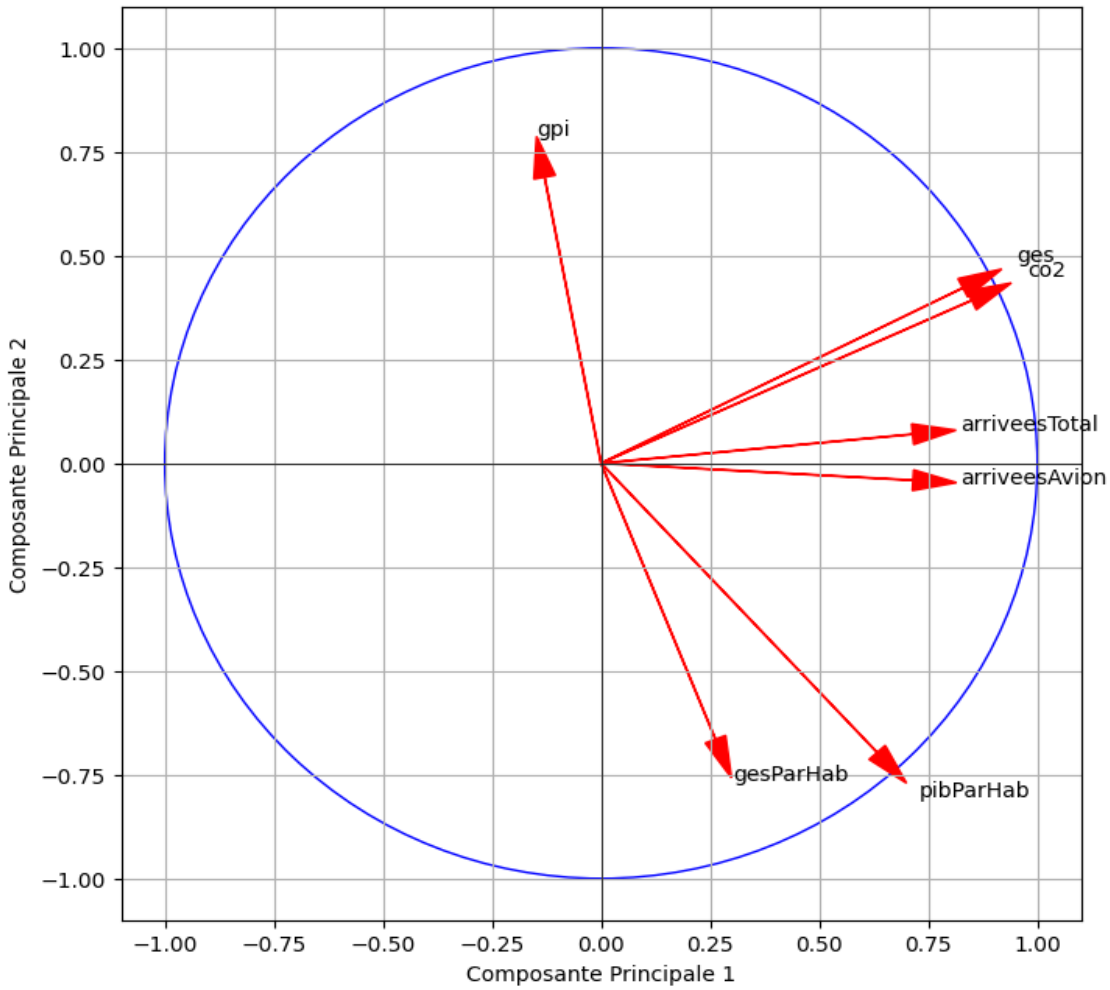

Figure 1.3: Cercle de corrélation

1.3 Clustering

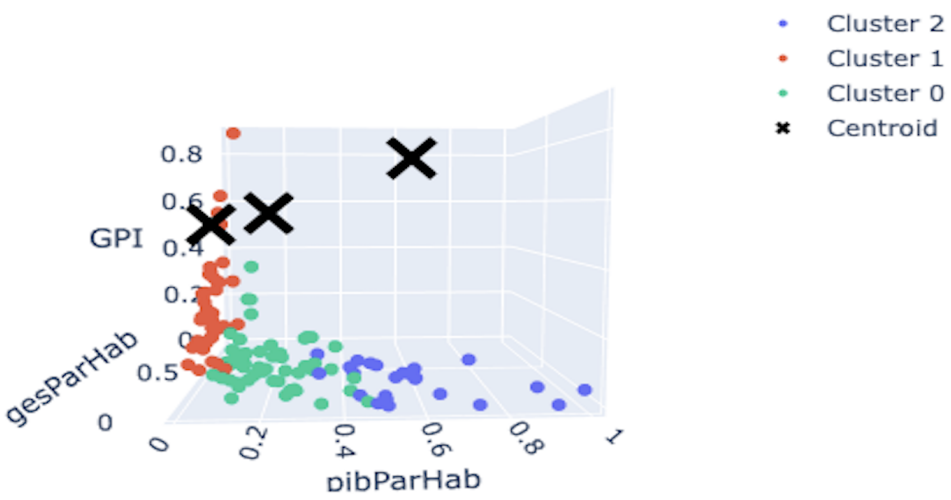

Figure 1.4: Représentation des clusters et des centroïdes dans un espace tridimensionnel défini par le PIB par habitant, les émissions de CO2 par habitant et le GPI

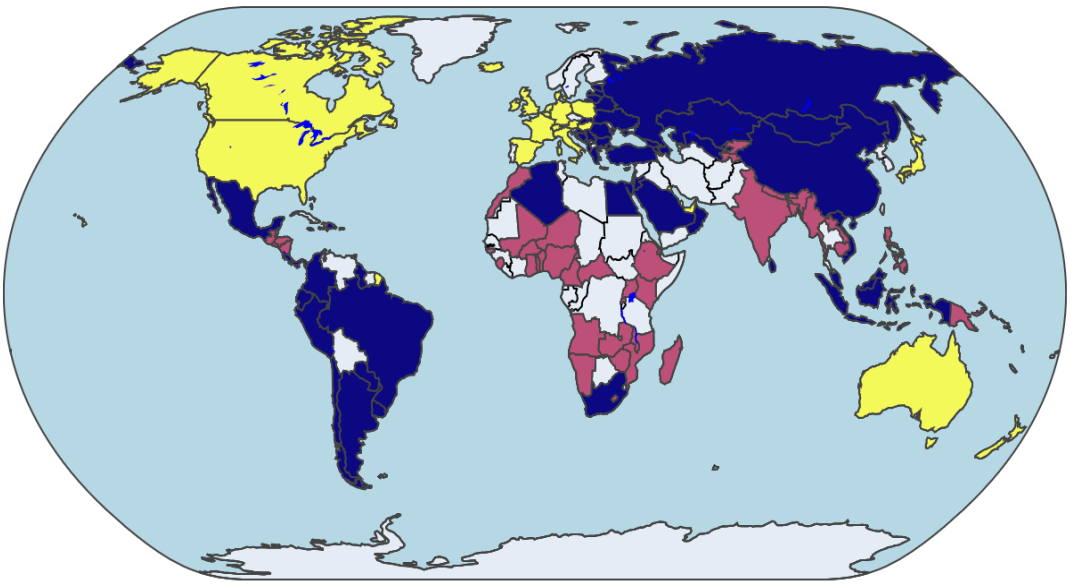

Figure 1.5: Carte Géographique des 3 Clusters

1.4 Clustering+

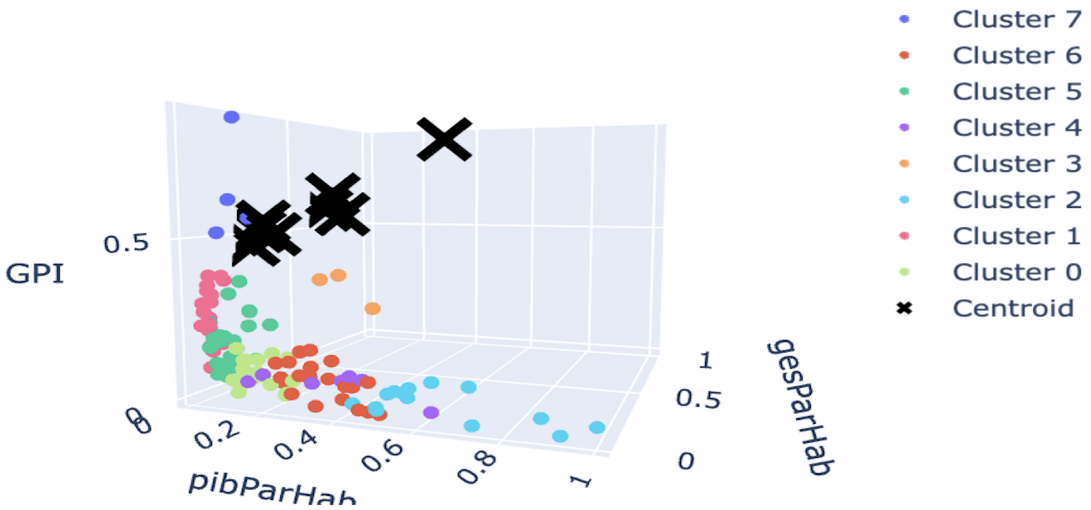

Figure 1.6: Représentation des clusters et des centroïdes dans un espace tridimensionnel défini par le PIB par habitant, les émissions de CO2 par habitant et le GPI

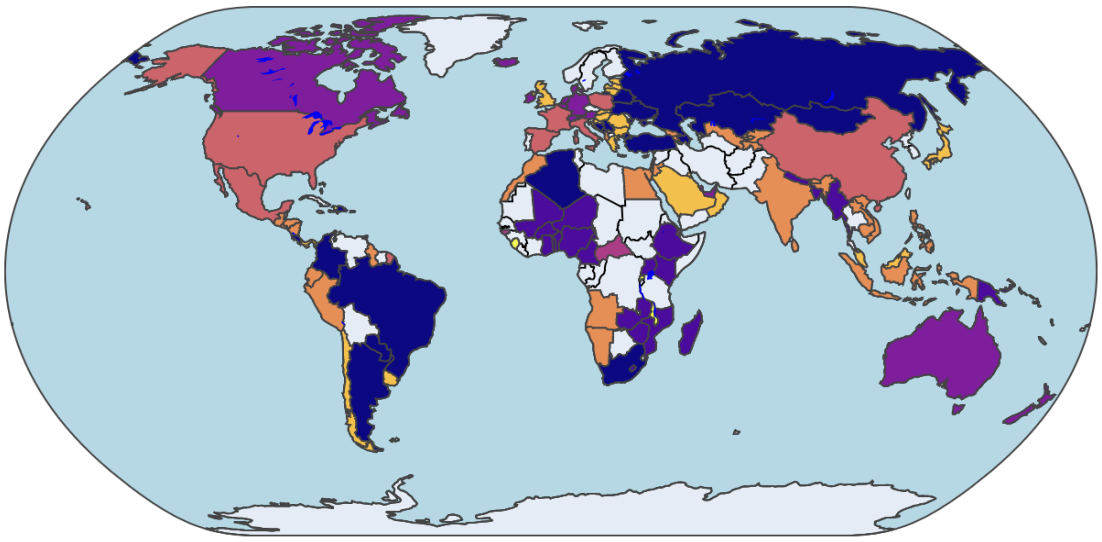

Figure 1.7: Carte Géographique des 8 Clusters

1.5 Régression linéaire

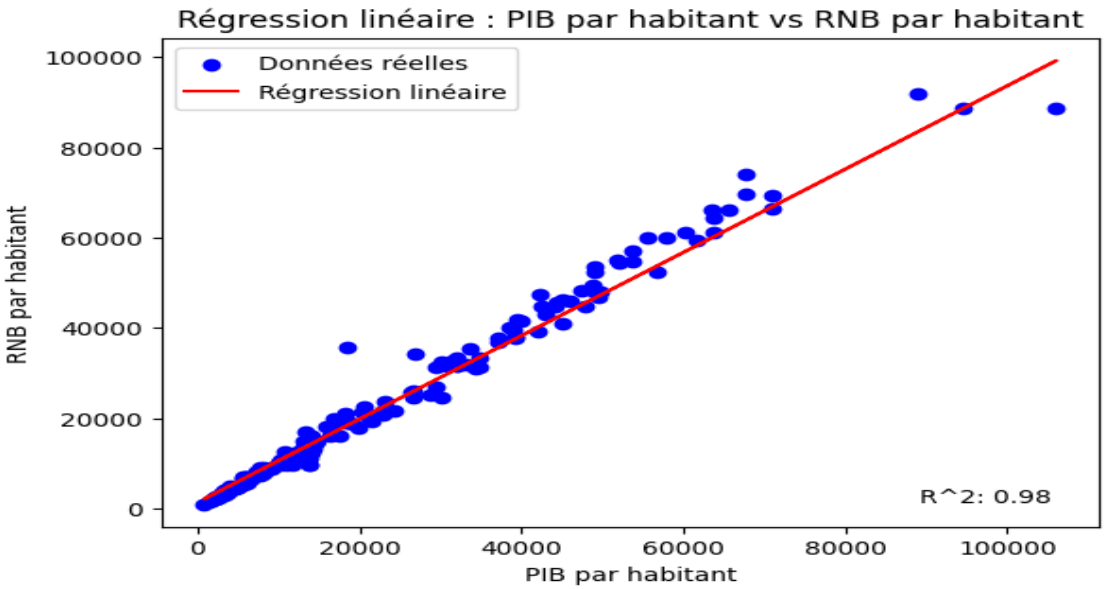

Figure 1.8: Régression linéaire 1

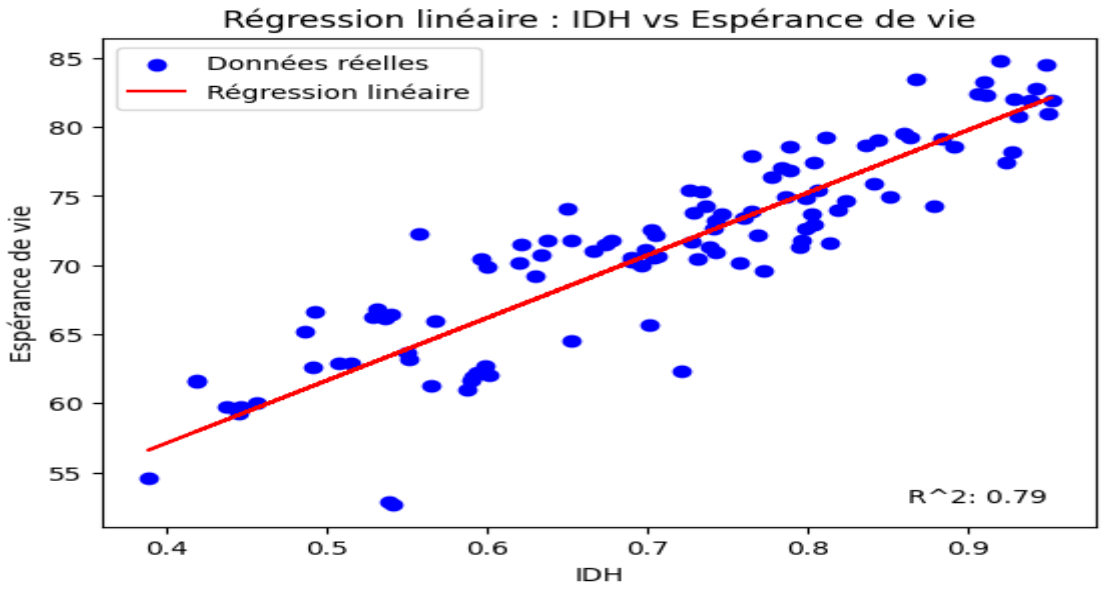

Figure 1.9: Régression linéaire 2

1.6 ARIMA

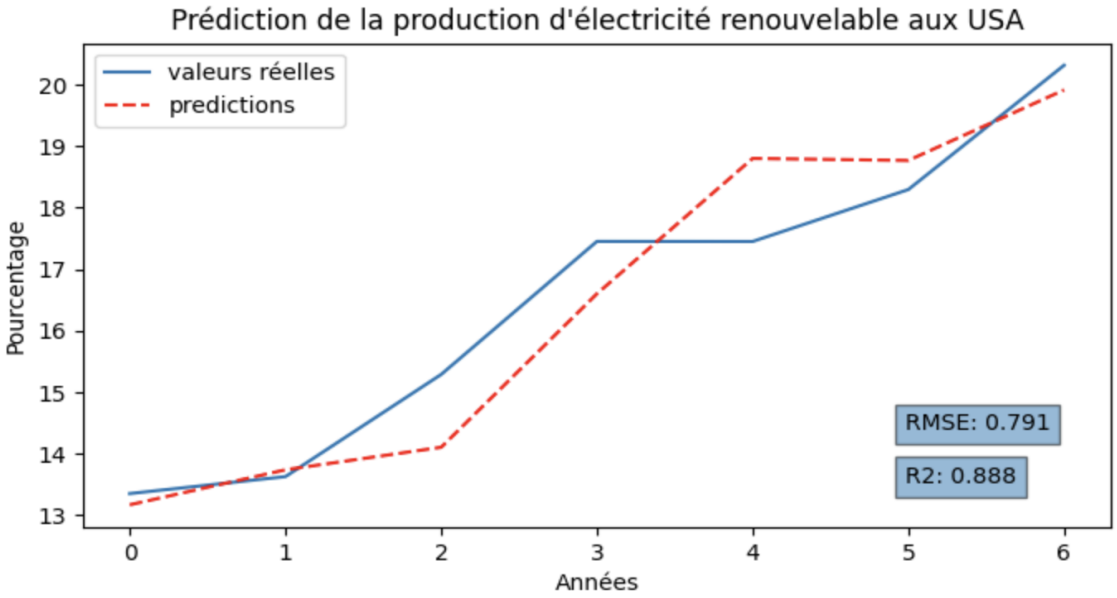

Figure 1.10: Prédiction sur l'échantillion

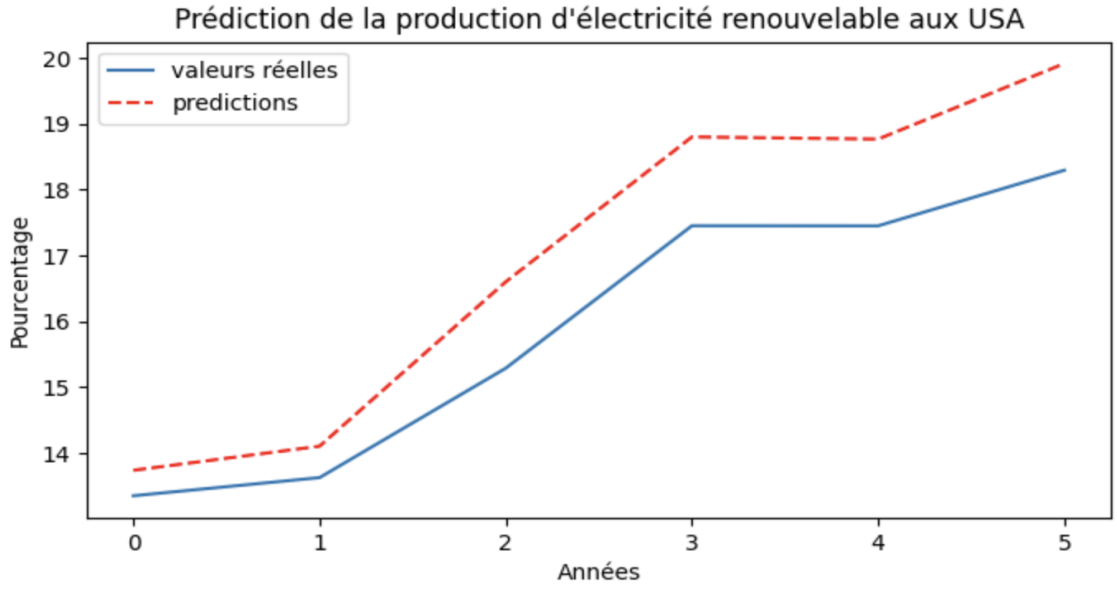

Figure 1.11: Prédiction sur l'échantillion pour l'année n-1

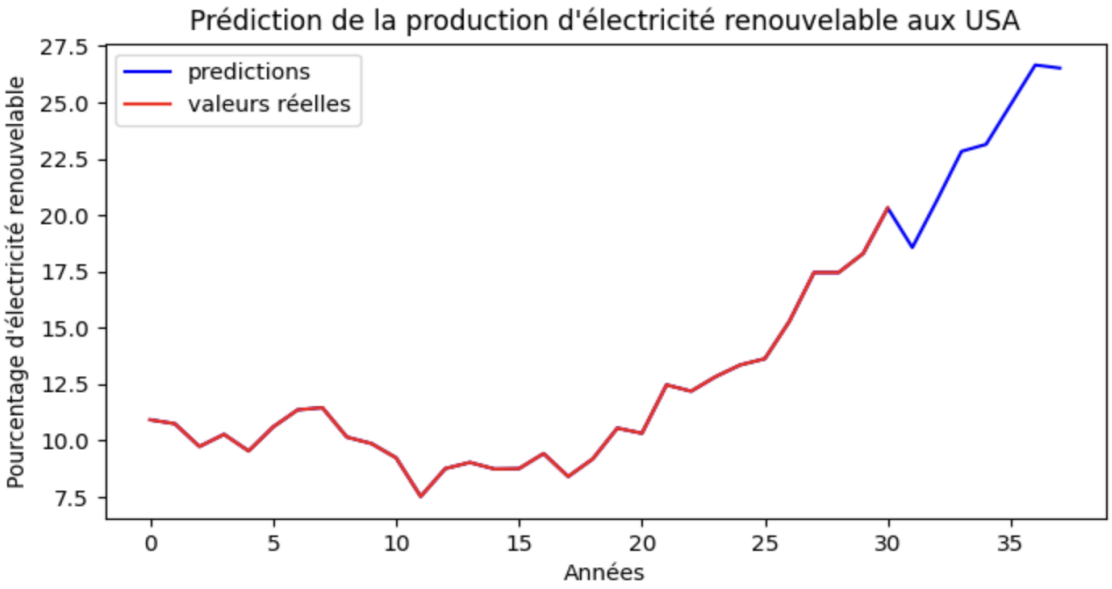

Figure 1.12: Prédiction hors échantillion