PyConso

Etude de la consommation d'électricité en France de 2013 à 2021

source : ENEDIS

Projet réalisé par :

BOULAHNACHE Ariles,

DIAKITE Hatoumata.

DataScientest

Table des matières

Table des illustrations

Table des abréviations

ACP : Analyse en Composantes Principales

ANOVA : ANalyse Of Variance

CV : Cross-Validation

GES : Gaz à effet de serre

GIEC : Groupe d'Experts Intergouvernemental sur l'Evolution du Climat

IA: Intelligence Artificielle

LASSO : Least Absolute Shrinkage and Selection Operator

ML: Machine Learning

PACA : Provence-Alpes-Côte d'Azur

RMSE : Root Mean Squared Error

RFR : RANDOM FOREST REGRESSOR

RTE : Réseau de transport d'électricité

SGD : Stochastic gradient descent

STEP : Stations de Transfert d'Energie par Pompage

SVR : Support Vector Regression

TCH : le Taux de Charge (TCH) ou facteur de charge (FC) d'une filière représente son volume de production par rapport à la capacité de production installée et en service de cette filière

TCO : Taux de Couverture (TCO) d'une filière de production au sein d'une région représente la part de cette filière dans la consommation de cette région

---

CONTEXTE

La France dispose d'un mix électrique bas carbone grâce à son important parc nucléaire mais de nombreux réacteurs arrivent en fin de vie.

Le pays [1] a entamé une transition énergétique ambitieuse encadrée par la Loi Énergie-Climat de 2019, avec notamment la mise en place la Stratégie Nationale Bas-Carbone qui définit une trajectoire de réduction des émissions de gaz à effet de serre (GES) jusqu'à 2050 et fixe des objectifs à court-moyen terme : les budgets carbone. De plus, La France prévoit de réduire la part d'énergie nucléaire dans le mix énergétique du pays de 75% aujourd'hui à 50% d'ici 2035, ce qui nécessitera des investissements ambitieux dans les énergies renouvelables et l'efficacité énergétique.

D'autre part, pour éviter les pannes, la surcharge du réseau électrique ou les coupures de courant, le RTE (Réseau de transport d'électricité) est le garant de l'équilibre permanent entre la consommation d'électricité et sa production. Il enregistre et surveille en temps réel ces données quels que soient les régions, les sources de production ou les heures. Notre projet se basera sur des jeux de données téléchargés depuis le site internet du RTE.

[1] www.oceole.fr

Machine Learning et Consommation d'Énergie[2] sont aujourd'hui de grands sujets d'actualité qu'on ne peut détacher d'une problématique majeure mondiale : le changement climatique. Avec une population mondiale croissante, la quantité d'énergie dont nous avons besoin ne fait que croître et les émissions de GES augmentent avec elle. De plus, à cause d'une surexploitation globale, la plupart des stocks de ressources naturelles nécessaires à la production d'énergie diminuent. L'accélération et l'intensification d'évènements climatiques accentuant ces diminutions inquiètent donc de plus en plus. Cette année 2022, nous en sommes tous témoins, est par exemple particulièrement impactée par des phénomènes mondiaux de grande sécheresse. La France est particulièrement touchée depuis le mois de janvier[3] mais elle n'est pas la seule. La Chine par exemple a déjà dû mettre en place des restrictions et des coupures d'électricité, sa production dépendant en grande partie de la disponibilité en eau[4]. Si les scientifiques du GIEC (Groupe d'Experts Intergouvernemental sur l'Evolution du Climat) alertent depuis longtemps sur les impacts du changement climatique, cet été nous pousse plus que jamais à réfléchir de manière urgente aux solutions déjà proposées et à créer pour une adaptation de nos sociétés aux changements qui vont s'intensifier[5]. **Les nouvelles technologies** sont un des moyens essentiels utilisés pour répondre aux problèmes posés par le changement climatique. Elles aident à mieux comprendre ses causes, prédire ses impacts et trouver des moyens de réduction des émissions de GES et d'adaptation aux changements[6]. Le **Machine Learning (ML)** est une sous-catégorie d'**Intelligence Artificielle (IA)** qui permet à des algorithmes de s'entraîner sur des ensembles de données et d'apprendre de manière autonome à améliorer leurs performances dans l'exécution de tâches spécifiques comme des prédictions[7]. Chaque jour, des data scientists travaillent à l'amélioration des techniques de ML et à la création de nouveaux algorithmes plus efficaces.

---

[2] https://www.dexma.com/

[3] https://www.lemonde.fr/les-decodeurs/article/2022/08/12/quatre-cartes-et-graphiques-qui-montrent-la-secheresse-exceptionnelle-qui-a-commence-des-janvier_6137873_4355770.html

[4] https://www.lemonde.fr/international/article/2022/08/19/en-chine-une-secheresse-sans-precedent-menace-la-croissance_6138446_3210.html

[5] https://reseauactionclimat.org/6e-rapport-du-giec-quelles-solutions-face-au-changement-climatique/

[6] https://developer.nvidia.com/blog/accelerating-climate-change-mitigation-with-machine-learning-the-case-of-carbon-storage/

[7] https://datascientest.com/machine-learning-tout-savoir

C'est pourquoi de plus en plus d'entreprises se concentrent sur la création de nouveaux moyens de **ML et d'IA** dédiés à l'industrie de l'énergie, en particulier en cherchant à **prévoir avec précision la consommation d'énergie et les performances des sources de production d'énergies renouvelables**.

LE PROJET PYCONSO

Dans le cadre de notre projet fil rouge, PyConso, nous nous sommes fixés trois objectifs :

- **Comparer la consommation d'énergie en France par rapport à sa production**
- **Comparer les filières de production : nucléaire et renouvelables.**
- **Construire un modèle de prédiction robuste de la consommation d'énergie régionale à différentes échelles de temps**

Notre étude s'est basée sur l'analyse d'un jeu de données issues des enregistrements par le RTE, disponible en libre accès sur : https://opendata.reseaux-energies.fr/explore/dataset/eco2mix-national-cons-def/export/?disjunctive.nature. Les données ont été enregistrées de janvier 2013 à février 2022 pour toutes les régions métropolitaines.

Nous avons ajouté à ces données des données complémentaires issues de différentes sources, potentiellement explicatives de la consommation d'énergie afin de construire un modèle de ML robuste et performant pour la prédiction de cette variable cible.

Nous nous sommes donc posé la question suivante : de quelles variables dépend la consommation électrique en France ?

- Au niveau national mais également au niveau régional, on suppose qu'elle est d'abord logiquement fonction des caractéristiques saisonnières, avec une forte thermo sensibilité en hiver (en lien avec l'utilisation des chauffages) et dans une moindre mesure en été. On a donc sélectionné tout un panel de

variables environnementales dont les fluctuations dépendent fortement des saisons ainsi que des variables indicatrices de l'utilisation d'électricité pour le chauffage ou le refroidissement.

- Au niveau local, on suppose que la consommation est également dépendante des cycles d'activités horo-hebdomadaires avec notamment une moindre demande la nuit et les week-ends. On a donc sélectionné des variables en lien avec la densité de population, les secteurs d'activité de la région et les besoins en électricité des transports. On a également intégré des variables en lien avec la baisse d'activité globale : vacances, jours fériés et périodes de confinement.

De plus, la production dépendant évidemment de la consommation, on s'est demandé de quelles filières dépendaient chaque région et on a souhaité comparer la production à la consommation régionale afin de mieux appréhender les problèmes qui risquent de se poser au niveau local.

Pour répondre à nos différents objectifs nous avons suivi les étapes suivantes :

1) ***Exploration des données : comprendre les observations et les variables***
2) ***Preprocessing : préparer les données pour être exploitées***
3) ***Data Visualisation***
4) ***Modélisation***

1. EXPLORATION DES DONNEES

1.1. Découverte des jeux de données

1.1.1. Le jeu de données principal

Le jeu de données principal ***« eco2mix-regional-cons-def_updated.csv*** » est issu de la base de données du RTE et a été téléchargé au format .csv. Il est initialement composé de **1 927 296 observations** enregistrées toutes les demi-heures pour **32 variables** *(fig.1)* :

- Des variables spatio-temporelles : ***'Code INSEE région'***, ***'Région'***, ***'Date'***, ***'Heure'***, ***'Date-Heure'***.

- Une variable précisant si la donnée a été validée après un certain temps : ***'Nature'*** *(modalités « Définitives » ou « Consolidées »).*
- Une variable relative à la consommation d'électricité enregistrée : ***'Consommation (MW)'*** qui sera la variable cible de notre modèle prédictif
- Des variables relatives à la production d'électricité en fonction des filières et aux échanges : ***'Thermique (MW)', 'Nucléaire (MW)',' Eolien (MW)', 'Solaire (MW)', 'Hydraulique (MW)',' Pompage (MW)'*** (La consommation des pompes dans les Stations de Transfert d'Energie par Pompage (STEP)), ***'Bioénergies (MW)', 'Ech. physiques (MW)'*** (Le solde des échanges physiques avec les régions limitrophes), ***'Eolien terrestre', 'Eolien offshore', 'TCO Thermique (%), 'TCH Thermique (%)', ' TCO Nucléaire (%)','TCH Nucléaire (%)','TCO Eolien (%)','TCH Eolien (%)','TCO Solaire (%)','TCH Solaire (%)','TCO Hydraulique (%)','TCH Hydraulique (%)','TCO Bioénergies (%)',' TCH Bioénergies (%)'*** (cf. Table des abréviations*).*
- Des variables avec très peu ou pas de données, relatives à des expérimentations en cours : ***'Stockage batterie', 'Déstockage batterie', 'Column 30'*.** Ces variables seront supprimées dés le début de l'analyse.

	Code INSEE région	Région	Nature	Date	Heure	Date - Heure	Consommation (MW)	Thermique (MW)	Nucléaire (MW)	Eolien (MW)	Solaire (MW)	Hydraulique (MW)	Pompage (MW)	Bioénergies (MW)
0	28	Normandie	Données définitives	2013-01-01	00:00	2013-01-01T00:00:00+01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	27	Bourgogne-Franche-Comté	Données définitives	2013-01-01	00:00	2013-01-01T00:00:00+01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	75	Nouvelle-Aquitaine	Données définitives	2013-01-01	00:00	2013-01-01T00:00:00+01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	11	Île-de-France	Données définitives	2013-01-01	00:00	2013-01-01T00:00:00+01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	53	Bretagne	Données définitives	2013-01-01	00:00	2013-01-01T00:00:00+01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Ech. physiques (MW)	Stockage batterie	Déstockage batterie	Eolien terrestre	Eolien offshore	TCO Thermique (%)	TCH Thermique (%)	TCO Nucléaire (%)	TCH Nucléaire (%)	TCO Eolien (%)	TCH Eolien (%)	TCO Solaire (%)	TCH Solaire (%)	TCO Hydraulique (%)	TCH Hydraulique (%)
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

TCO Bioénergies (%)	TCH Bioénergies (%)	Column 30
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN

Figure 1: Affichage des 5 premières lignes du jeu de données initial

L'affichage du jeu de données initial nous permet déjà de constater la forte présence de valeurs manquantes dans les premières lignes. La création d'une fonction ***valeur_manquante()*** nous permettra d'afficher le nombre de valeurs manquantes par variable (*fig.2)*. De plus, la méthode .info() nous permet d'afficher les informations relatives au jeu de données et de constater que la majorité des variables sont de type numérique (float64, int64) et que seules les variables spatio-temporelles sont de type objet *(fig.3*).

```
Variables avec des valeurs manquantes:

"Nucléaire":"42.0 % valeurs manquantes"
"Pompage":"43.0 % valeurs manquantes"
"Eolien_terr":"93.0 % valeurs manquantes"
"Eolien_off":"93.0 % valeurs manquantes"
"TCO_therm":"76.0 % valeurs manquantes"
"TCH_therm":"76.0 % valeurs manquantes"
"TCO_nuc":"81.0 % valeurs manquantes"
"TCH_nuc":"81.0 % valeurs manquantes"
"TCO_eol":"76.0 % valeurs manquantes"
"TCH_eol":"76.0 % valeurs manquantes"
"TCO_sol":"76.0 % valeurs manquantes"
"TCH_sol":"76.0 % valeurs manquantes"
"TCO_hydro":"89.0 % valeurs manquantes"
"TCH_hydro":"89.0 % valeurs manquantes"
"TCO_bio":"89.0 % valeurs manquantes"
"TCH_bio":"89.0 % valeurs manquantes"
```

Figure 2 : Affichage des valeurs manquantes par variable après suppression des variables relatives aux expérimentations en cours.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1927296 entries, 0 to 1927295
Data columns (total 32 columns):
 #   Column               Dtype
---  ------               -----
 0   Code_région          int64
 1   Région               object
 2   Nature               object
 3   Date                 object
 4   Heure                object
 5   Date - Heure         object
 6   Consommation         float64
 7   Thermique            float64
 8   Nucléaire            float64
 9   Eolien               float64
 10  Solaire              float64
 11  Hydraulique          float64
 12  Pompage              float64
 13  Bioénergies          float64
 14  Echanges             float64
 15  Stockage batterie    float64
 16  Déstockage batterie  float64
 17  Eolien_terr          float64
 18  Eolien_off           float64
 19  TCO_therm            float64
 20  TCH_therm            float64
 21  TCO_nuc              float64
 22  TCH_nuc              float64
 23  TCO_eol              float64
 24  TCH_eol              float64
 25  TCO_sol              float64
 26  TCH_sol              float64
 27  TCO_hydro            float64
 28  TCH_hydro            float64
 29  TCO_bio              float64
 30  TCH_bio              float64
 31  Column 30            float64
dtypes: float64(26), int64(1), object(5)
memory usage: 470.5+ MB
```

Figure 3 : Affichage des informations relative au jeu de données initial

1.1.2. Les jeux de données complémentaires

- **PyMeteo** : jeu de données météorologiques avec des variables potentiellement corrélées à la consommation et à la production d'électricité

telles que la température, les précipitations, le rayonnement solaire ou encore la vitesse des vents. Le jeu de données est au format .csv : **'donnees-meteo-climat.csv'**, et il est issu de https://www.data.gouv.fr/fr/datasets/observation-meteorologique-historiques-france-synop/. Les données sont enregistrées toutes les 15 minutes de 2010 à 2022. Il se compose de 2 066 537 observations pour 82 variables.

- **PyBorne** : jeu de données relatif aux stations de bornes de recharge de voitures électriques. Il est au format .csv : **'bornes-irve.csv'** et est issu de https://opendata.reseaux-energies.fr/explore/dataset/bornes-irve/table/. Les données sont enregistrées à certaines dates (non régulières) de 2017 à 2021 pour 18 141 observations et 21 variables.
- **PyChauffage** : plusieurs jeux de données de consommation, de production et de points de livraison d'énergie en fonction des secteurs chaud ou froid et des filières de production. Les jeux sont au format .csv et couvrent les années 2008 à 2017. Les jeux comportent en tout 7620 observations pour 60 variables**.**
- **PySecteurs :** jeu de données relatives à la consommation annuelle d'électricité et de gaz par département et par secteur d'activité. Il est au format .csv : **'conso-elec-gaz-annuelle-par-secteur-dactivite-agregee-departement.csv'** et est issu de https://www.data.gouv.fr/fr/datasets/consommation-annuelle-delectricite-et-gaz-par-departement-et-par-secteur-dactivite/. Les données sont enregistrées de 2013 à 2020 pour 5022 observations et 30 variables.
- **dfPop :** jeu de données relatif aux densités annuelles de population par région et par tranche d'âge. Il est au format .csv : **'Pop_globale_2013-2022.csv'** et est issu de https://www.insee.fr/fr/statistiques/1893198. Les données sont enregistrées de 2013 à 2022 et comportent 108 observations pour 8 variables.
- **jf :** jeu de données comportant les jours fériés par année. Il est au format .csv : **'jours_feries_metropole.csv'** et est issu de https://www.data.gouv.fr/fr/datasets/jours-feries-en-france/. Il comporte 286 données pour 4 variables enregistrées entre 2002 et 2022.

- **Vacances et confinement :** ces variables sont créées à partir des informations recueillies sur https://www.data.gouv.fr.

1.2. Premier nettoyage des données

Cette étape est importante car elle permet de détecter les valeurs manquantes, doublons, valeurs aberrantes, de créer de nouvelles variables utiles et de supprimer des variables inutiles au projet.

1.2.1. Création de variables

Dans un premier temps, on choisit de créer de nouvelles variables temporelles (‘***Année’, ‘Mois’, ‘Semaine’, ‘Saison’***) nous permettant de visualiser nos données à différentes échelles de temps et de comprendre leurs distributions.

Afin de comparer les différentes filières de production et de comparer la production globale d’énergie face à la consommation, on crée les variables ‘***Energies_renouvelables'*** (la somme des productions des filières dites renouvelables), ‘***Production***' (la somme de toutes les productions d’électricité toutes filières confondues) et '***Bilan***' (la production totale à laquelle on soustrait la consommation totale et on ajoute les échanges).

1.2.2. Suppression de variables et de modalités

Certaines variables contiennent beaucoup de valeurs manquantes (**fig.2)** et on peut déterminer qu’elles ne sont enregistrées que certaines années (**fig.4**). On choisit de supprimer ces variables.

Année	TCO_therm	TCH_therm	TCO_nuc	TCH_nuc	TCO_sol	TCH_sol	TCO_eol	TCH_eol	TCO_hydro	TCH_hydro	TCO_bio	TCH_bio	Eolien_terr	Eolien_off
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2017	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2018	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2019	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2020	210816	210816	122976	122976	210816	210816	210816	210816	210816	210816	210816	210816	0	0
2021	210240	210240	210240	210240	210240	210240	210240	210240	0	0	0	0	122640	122640
2022	33984	33984	33984	33984	33984	33984	33984	33984	0	0	0	0	19824	19824

Figure 4 : Dataframe créé par un groupby affichant le nombre de données par variable comptant plus de 705% de valeurs manquantes

De plus, l'année 2022 n'étant pas complète on choisit de ne pas la garder. Cette modalité étant supprimée toutes les données sont de type « Définitives », la variable '***Nature***' n'est donc plus utile.

Au niveau des jeux de données complémentaires, on ne garde que les variables potentiellement utiles pour le modèle prédictif de la consommation en se basant sur les matrices de corrélations générées après différentes étapes de fusion des jeux de données (**fig 5**). Pour les jeux relatifs aux productions/consommations/distributions d'énergie selon les filières et les secteurs d'activité, on ne garde que les variables donnant un nombre de points de branchement ou de livraison d'énergie. En effet, on ne peut pas prédire la consommation d'électricité totale à partir de consommations partielles ou à partir de variables de production d'électricité.

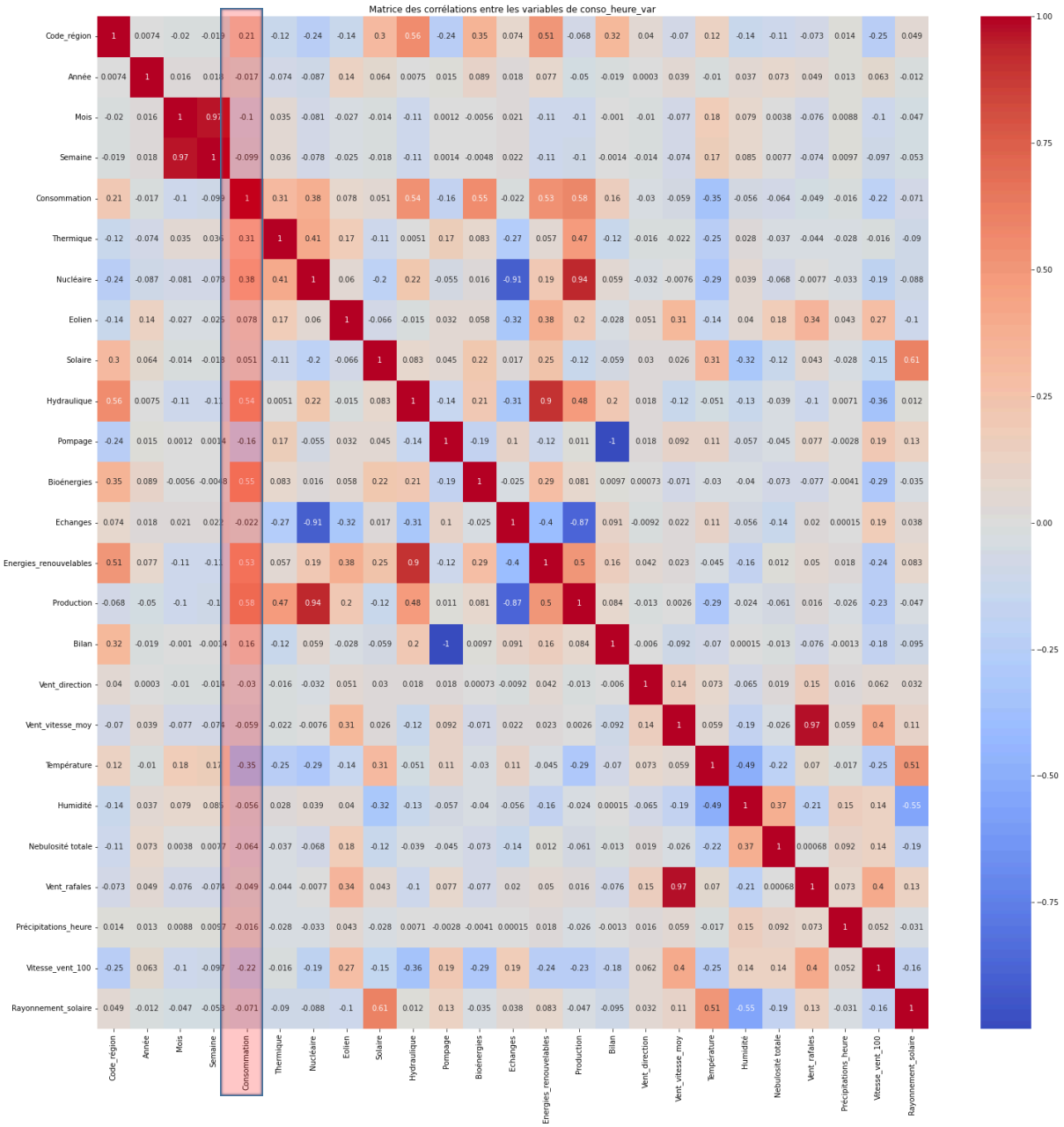


Figure 5 : Matrice de corrélation du jeu de données conso_heure_var après la fusion avec les données météorologiques

A l'issu de ce premier nettoyage nous avons fusionné l'ensemble des variables dans un jeu de données horaire : **conso_heure_final.csv** puis utilisé des groupby en calculant la moyenne des variables pour obtenir des jeux de données journaliers, hebdomadaires, mensuels et annuels : **conso_jour_final.csv**, **conso_semaine_final.csv**, **conso_mois_final.csv** et **conso_annee_final.csv**. L'ensemble des variables que nous avons choisi de conserver ainsi que les informations relatives à ces variables sont présentées dans le fichier excel : template_-_rapport_exploration_des_donne__es___1 DM (1).xlsx.

1.3. Analyse statistique et visuelle

1.3.1. Analyse de la variable cible : 'Consommation'

- **Distribution et valeurs aberrantes**

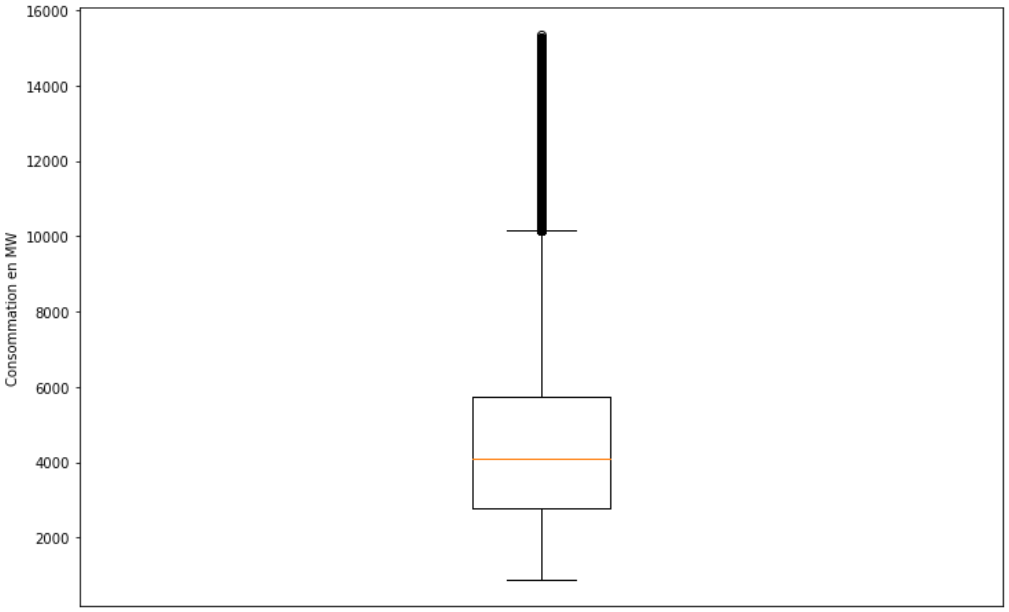


Figure 6 : Boxplot affichant la distribution de la consommation horaire d'électricité en France de 2013 à 2021. En jaune : la médiane, en noir gras : les outliers ou valeurs aberrantes

La distribution de consommation d'électricité enregistrée toutes les demi-heures (dite 'horaire' dans l'étude) montre un grand nombre de valeurs aberrantes. (**Fig. 6**) En regardant au niveau régional on s'aperçoit que ces valeurs supérieures à 10 000 MW sont uniquement enregistrées dans les régions Auvergne-Rhône-Alpes et Ile-de-France (**Fig. 7**). Nous choisissons de les conserver et de ne pas les considérer comme des erreurs.

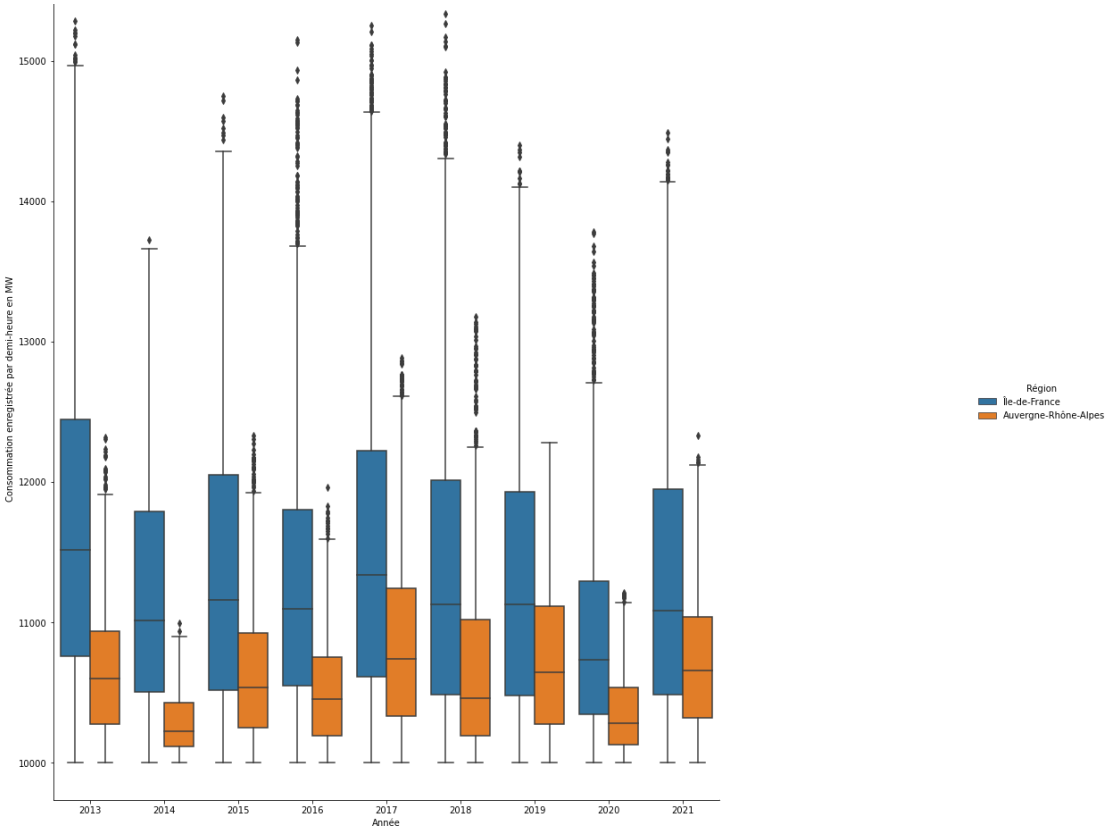
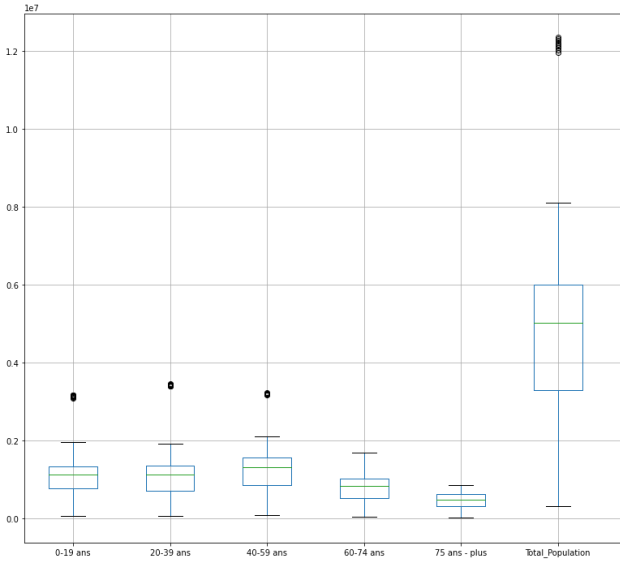


Figure 7 : Evolution de la consommation d'électricité supérieure à 10 000 MW par demi-heure par région et par année.

De plus, en regardant la distribution de la population on retrouve des densités nettement supérieures aux autres régions dans la région Ile-de-France (Fig. 8). Les consommations d'électricité particulièrement élevées dans cette région paraissent donc cohérentes et nous confirme notre choix de les conserver.



	Année	Région	Total_Population
7	2021	Île-de-France	12348605
21	2020	Île-de-France	12308593
35	2019	Île-de-France	12262544
49	2018	Île-de-France	12213447
63	2017	Île-de-France	12174880
77	2016	Île-de-France	12117132
91	2015	Île-de-France	12082144
105	2014	Île-de-France	12027565
119	2013	Île-de-France	11959807

Figure 8 : Distribution de la densité de population par tranche d'âge(à gauche) et densités de population régionales uniquement supérieures à 1 000 000 d'habitants recensés par an.

- **Corrélations avec les autres variables**

La première variable qu'on souhaite tester est la température. On affiche son évolution temporelle par rapport à celle de la consommation horaire et on note clairement la relation anti-synchrone qui en ressort (Fig. 9). Toutes deux oscillent de manière sinusoïdale au cours du temps mais quand l'une augmente l'autre diminue.

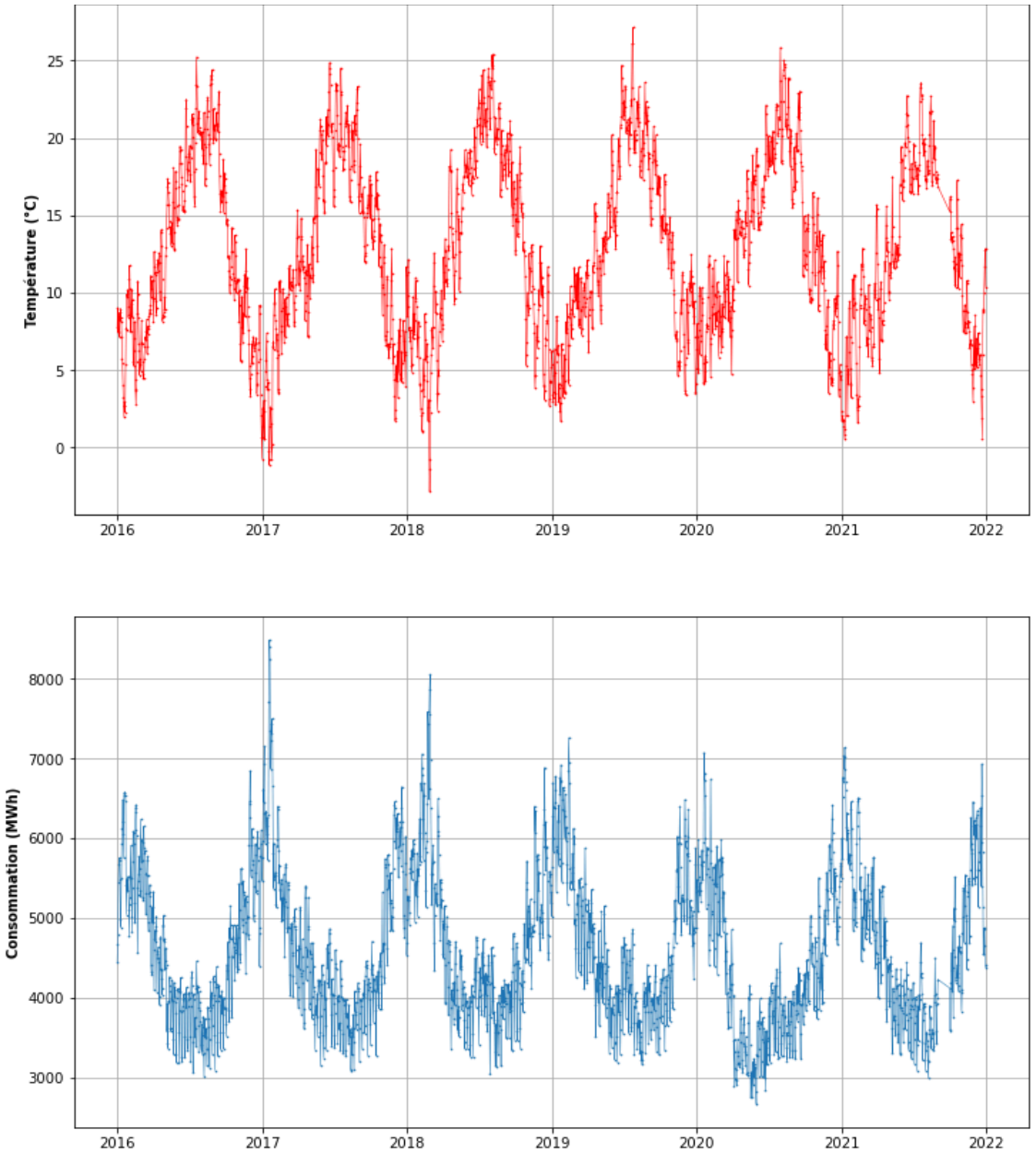


Figure 9 : Evolution de la température (en haut) et de la consommation moyenne horaire par jour entre les années 2016 et 2022

Cette corrélation est confirmée par le test de Pearson avec une p_value de 0<0.05 invalidant l'hypothèse nulle que les variables sont indépendantes. Cependant, on trouve un coefficient de corrélation plus faible qu'attendu de -0.34. La régression locale affichée entre les deux variables montre que la corrélation est moins marquée quand les valeurs de la température sont élevées et que celles de la consommation sont faibles.

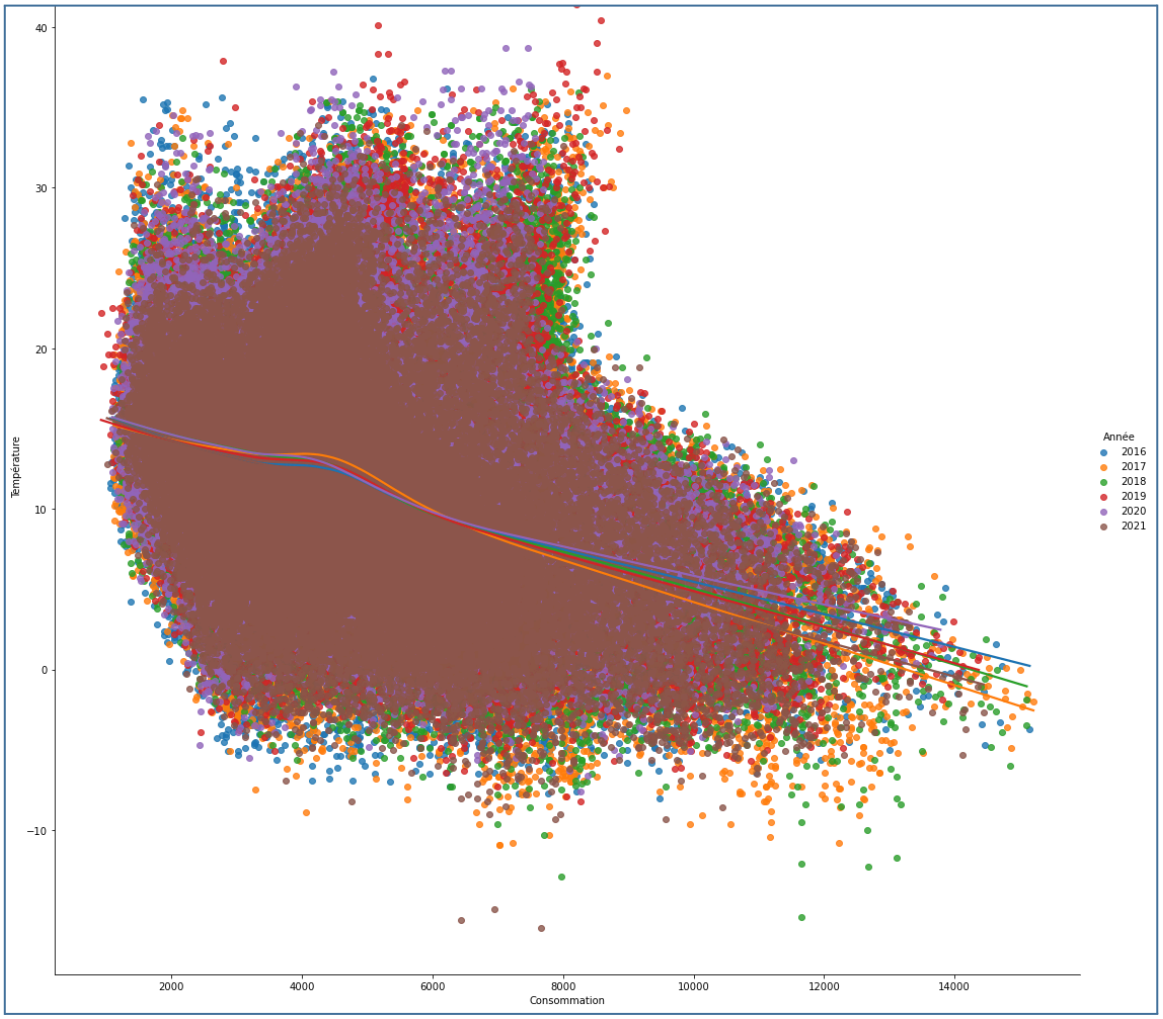


Figure 10 : Régression locale entre Consommation et Température par année

Les régressions avec les autres variables météorologiques ne montrent pas de forte corrélation (Fig. 11).

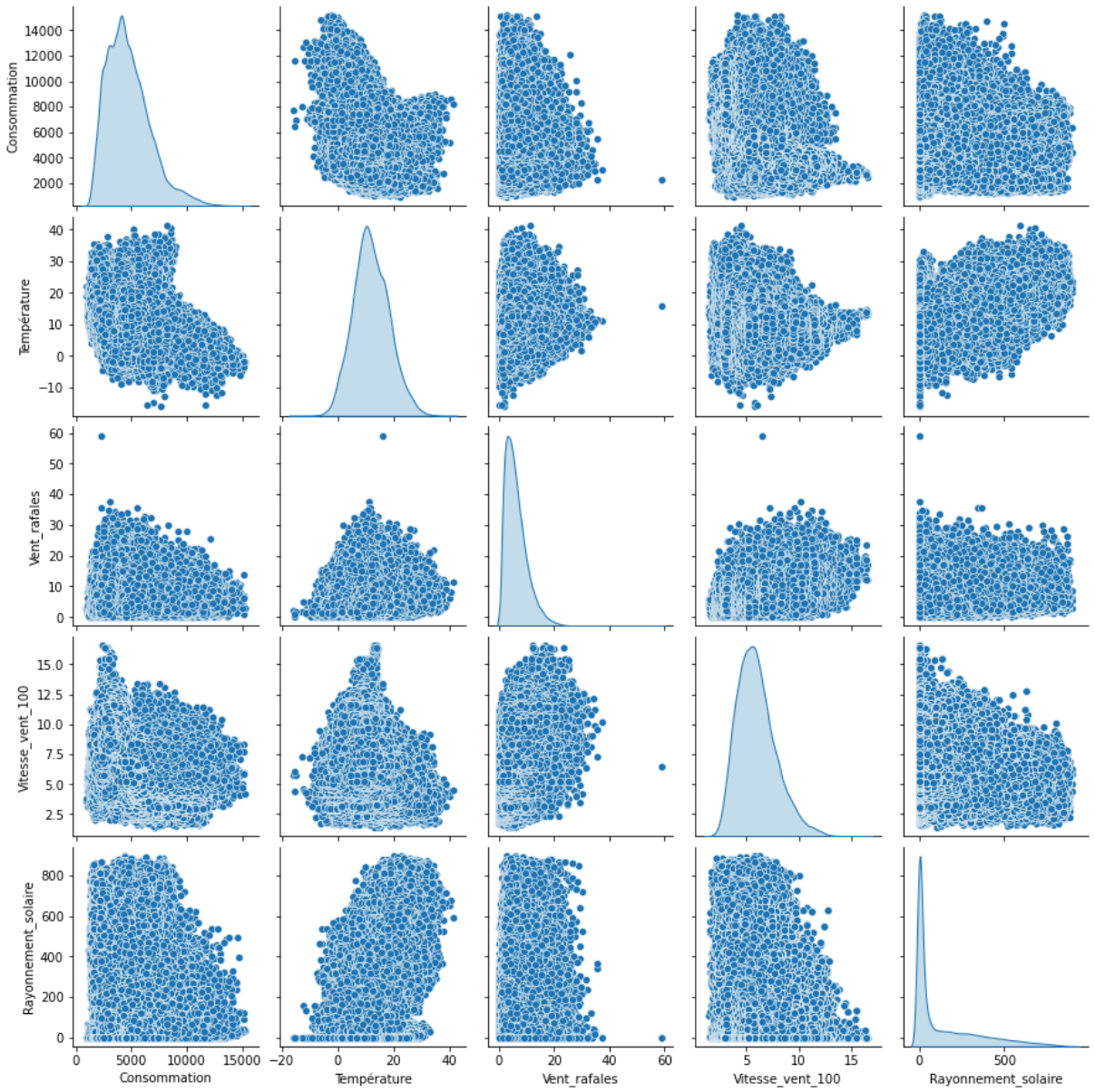
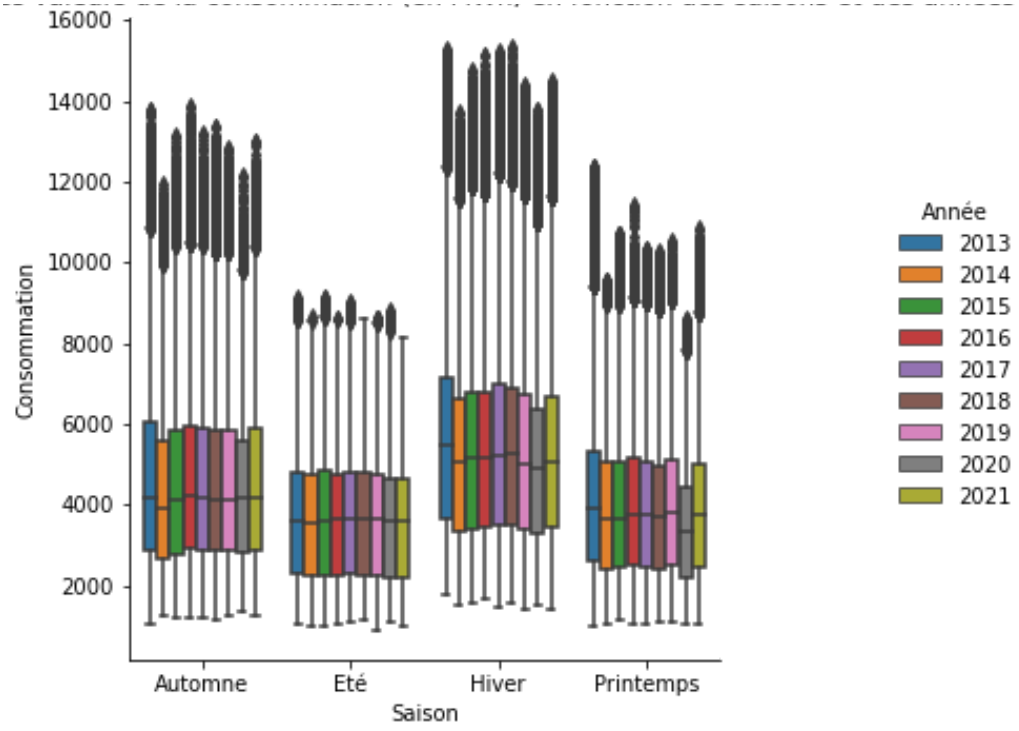


Figure 11 : Relations entre chaque variable météorologique et la consommation d'électricité (nuages de points et courbes de densité)

De plus, on suppose que si la température est une variable explicative, la température évoluant au fil des saisons, les saisons doivent elles aussi expliquer une partie de la variance de la consommation (Fig. 12). Une ANOVA permet de confirmer la dépendance entre les deux variables.



	df	sum_sq	mean_sq	F	PR(>F)
Saison	3.0	1.038881e+12	3.462938e+11	83504.301499	0.0
Residual	1893296.0	7.851531e+12	4.147017e+06	NaN	NaN

Figure 12 : Distribution de la consommation moyenne saisonnière (en MW par demi-heure) (en haut) et résultats de l'ANOVA appliquée aux variables Consommation et Saison

Enfin, on suspecte une forte relation entre les densités de population et la consommation d'électricité, ce qu'on vérifie grâce à la matrice des corrélations (Fig. 13). Quelle que soit la tranche d'âge concernée les coefficients de corrélation entre chaque densité de population est très fortement corrélée à la consommation d'électricité (>0.9). Les différentes variables de densité sont également très fortement corrélées entre elles.

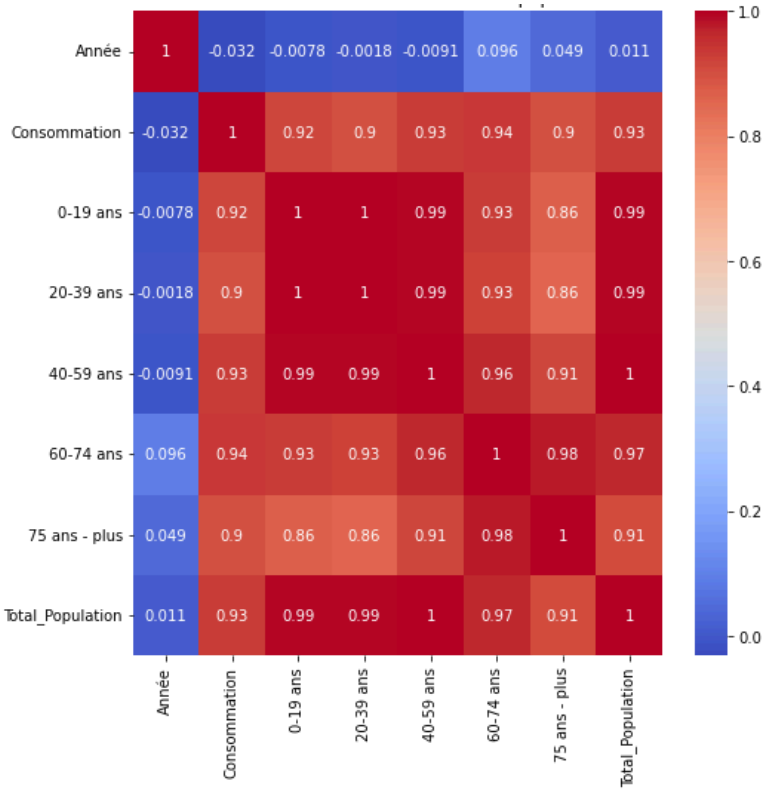


Figure 13 : Matrice de corrélation entre les variables de densité de population et la consommation d'électricité.

1.3.2. Comparaison entre la production et la consommation d'électricité

Depuis 2013, on observe que la consommation comme la production française d'électricité a baissé (**Fig. 14**). On remarque notamment une chute nette des deux variables entre 2019 et 2020 puis une remontée parallèle entre 2020 et 2021. Ces deux phénomènes pourraient peut-être s'expliquer par l'arrêt ou le ralentissement de beaucoup d'activités pendant ces deux années très marquées par la crise du COVID.

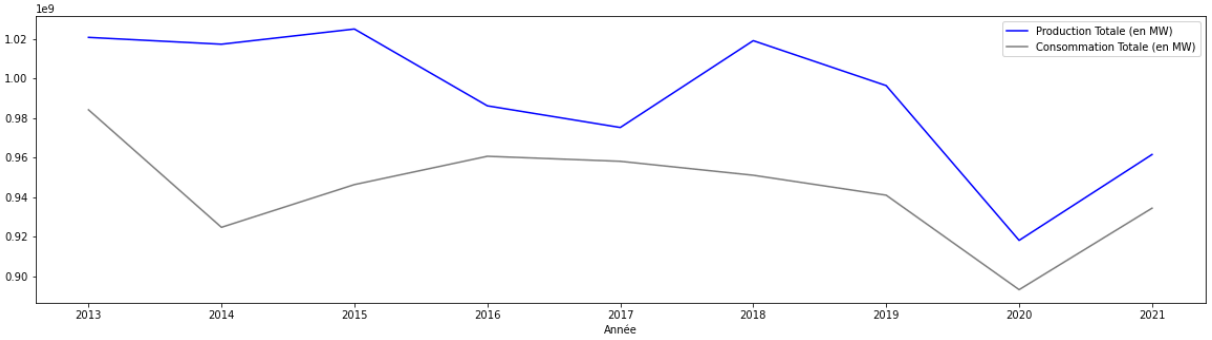


Figure 14 : Evolution de la consommation et la production totale métropolitaine entre 2013 et 2021.

De plus, on remarque que la production est toujours supérieure à la consommation, ce qui permet au pays d'assurer son offre mais en cas de forte demande. C'est notamment important pour les régions qui n'ont pas la capacité de produire d'assez d'électricité mais dont la demande est forte. C'est le cas des régions Ile-de-France, Provence-Alpes-Côte d'Azur (PACA), Pays de la Loire, Bretagne et Bourgogne-Franche-Comté (Fig. 15).

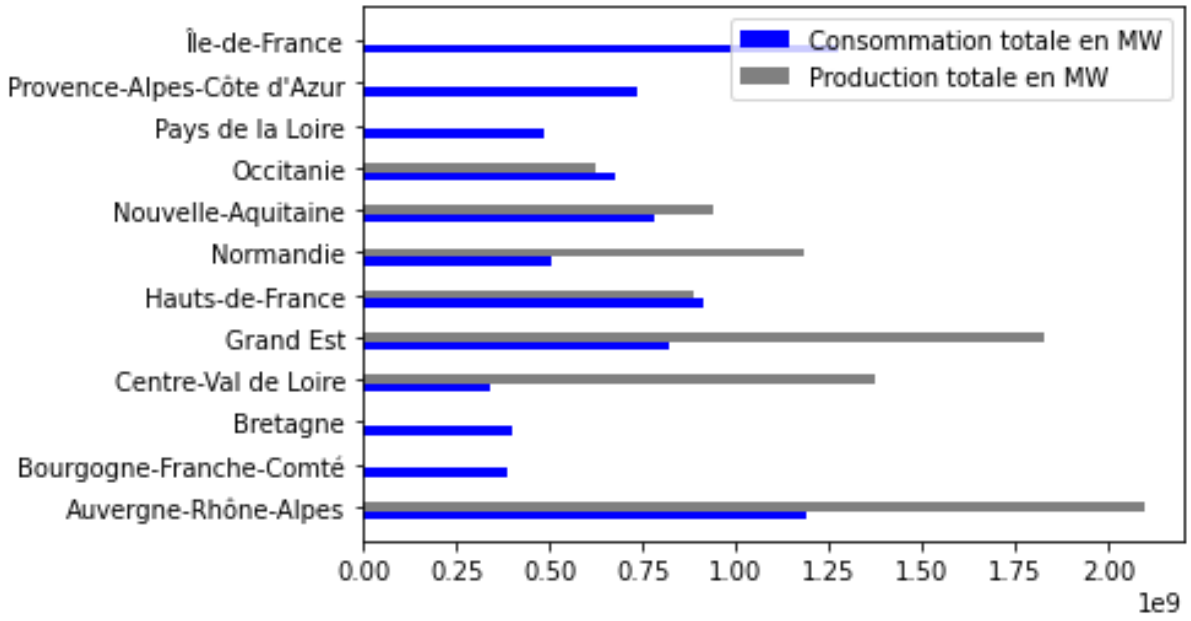


Figure 15 : Consommation et Production totale par région entre 2013 et 2021

1.3.3. Comparaison entre les différentes filières de production

On observe sur les cartes suivantes (Fig. 16) colorées en fonction des échelles de production et de consommation annuelles régionales qu'il existe bien une grande hétérogénéité à l'échelle locale. Toutes les régions ne consomment ni ne produisent

de la même manière. De plus chaque région n'est pas forcément capable d'assurer son autonomie énergétique d'où la nécessité d'échanges inter-régionaux. Par exemple, la région Ile-de-France qui consomme le plus d'électricité n'en produit que très peu et est donc très dépendante de la production dans les autres régions.

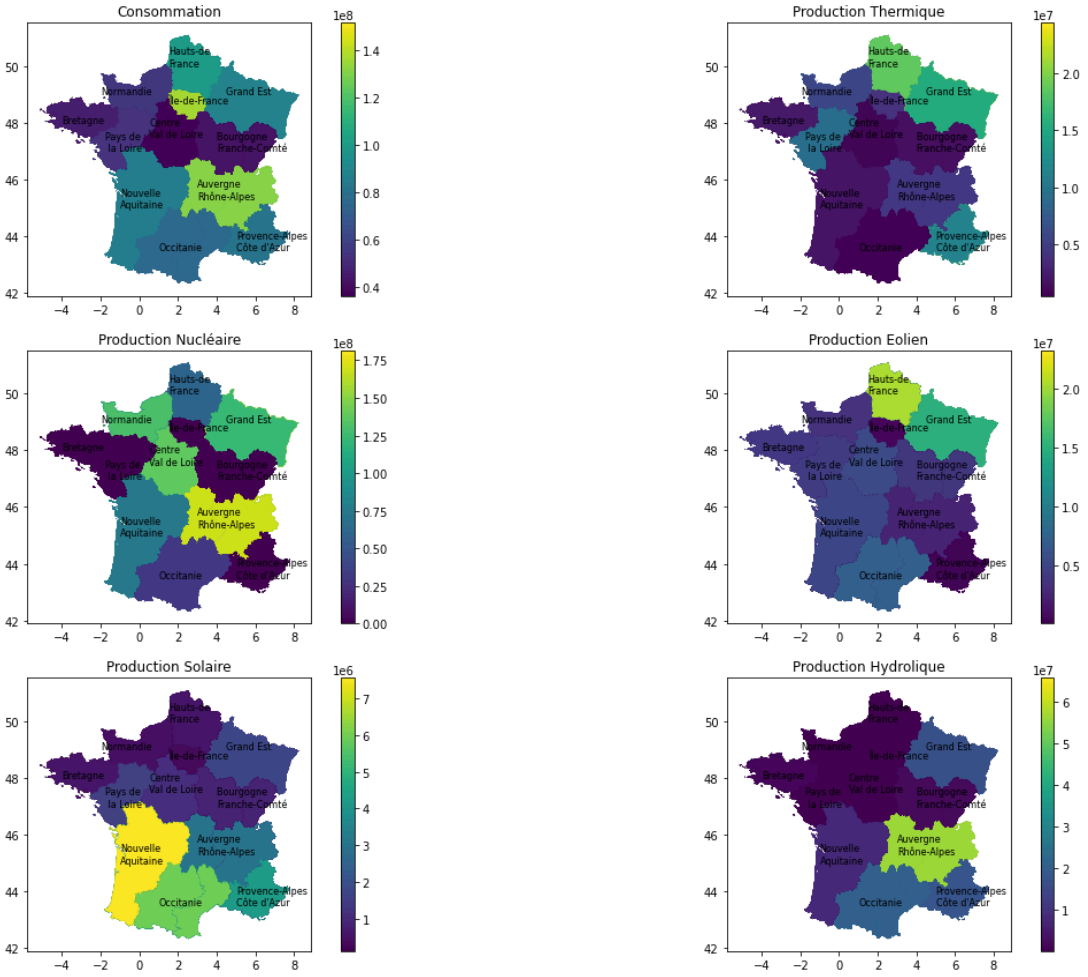


Figure 16 : Distribution régionale de la consommation et des productions d'électricité selon les filières (en MW par demi-heure)

La production nucléaire est majoritairement assurée par les régions Auvergne-Rhône-Alpes, Centre-Val de Loire, Nouvelle-Aquitaine, Grand-Est et Normandie. La production solaire se concentre principalement dans le sud de la France, le climat s'y prête en effet alors que la production éolienne se concentre plus au nord de la France. La production thermique est assez homogène sur tout le territoire. Enfin la production hydraulique est principalement assurée par la région Auvergne-Rhône-Alpes.

La production nucléaire reste de loin la première production d'électricité en France devant les énergies renouvelables (toutes filières confondues) qui produisent elles-mêmes plus du double de la production thermique. On ne constate pas de grandes différences au niveau des répartitions en fonction des années (Fig. 17).

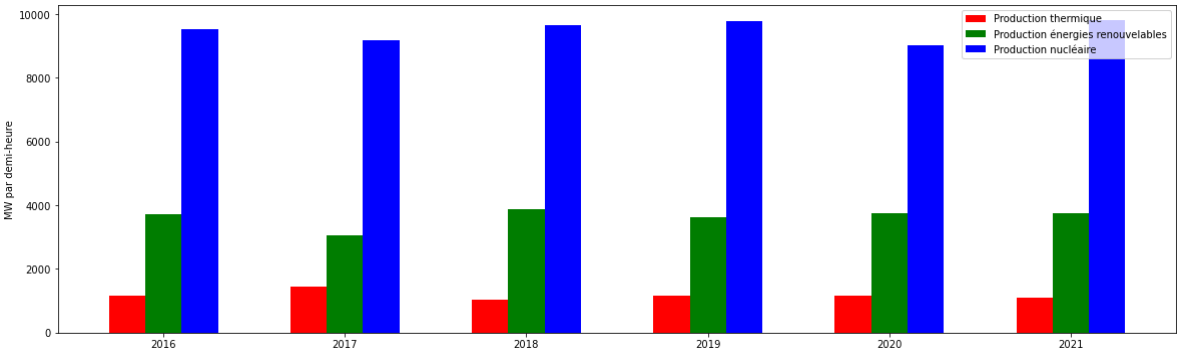


Figure 17 : Production d'électricité moyenne annuelle produite par demi-heure en fonction des trois grandes filières de production.

2. MODELISATION

Afin de choisir le modèle le plus adapté à la prédiction de la consommation d'électricité en France nous avons choisi d'entrainer différents modèles de régression sur trois jeux de données aux échelles temporelles différentes :

- **Les données dites 'horaires' (conso_heure_final.csv)**
- **Les données moyennes hebdomadaires (conso_semaine_final.csv)**
- **Les données moyennes mensuelles (conso_mois_final.csv)**

Tous affichent des données sur l'intervalle de temps 2016 -2021.

2.1. Préparation finale des données (pre-processing)

Pour chaque jeu de données, nous avons d'abord supprimé toutes les variables non-explicatives de la consommation ('Nature', 'Nucléaire', 'Eolien', 'Solaire', 'Hydraulique', 'Pompage', 'Bioénergies', 'Echanges', 'Energies_renouvelables', 'Production', 'Bilan') et conservé seulement la variable temporelle nous servant d'échelle et d'index (Fig. 18)

Heure	Code_région	Région	Année	Saison	Consommation	Température	Humidité	Vent_rafales	Vitesse_vent_100	Rayonnement_solaire
04:00	11	Île-de-France	2016	Hiver	7174.0	4.9	95.0	2.3	9.43	0.00
07:00	11	Île-de-France	2016	Hiver	7348.0	3.9	97.0	0.9	8.90	0.00
10:00	11	Île-de-France	2016	Hiver	7746.0	4.9	99.0	3.5	8.25	12.12

jour_ferie	Vacances	confinement	nbre_pdc	points_livraison_chaud	points_livraison_froid	nb_elec_agri	nb_elec_indus	nb_elec_ter
1	1	0	120.0	13470.0	937.0	822.0	11962.0	84912.0
1	1	0	120.0	13470.0	937.0	822.0	11962.0	84912.0
1	1	0	120.0	13470.0	937.0	822.0	11962.0	84912.0

nb_elec_resi	nb_elec_autre	0-19 ans	20-39 ans	40-59 ans	60-74 ans	75 ans - plus	Total_Population
6409602.0	483.0	3132488	3425345	3200438	1547105	811756	12117132
6409602.0	483.0	3132488	3425345	3200438	1547105	811756	12117132
6409602.0	483.0	3132488	3425345	3200438	1547105	811756	12117132

Figure 18 : Affichage du jeu de données horaires prêt pour le train-split

Pour le jeu de données conso_heure_final nous comptons 349030 pour 27 variables, le jeu de données conso_semaine_final : 4405 pour 28 variables et pour le jeu conso_mois_final : 857 pour 27 variables.

Les données de chaque jeu sont séparées en deux ensembles : l'ensemble d'entraînement X_train qui contient les données enregistrées entre 2016 et 2020 et l'ensemble test X_test qui contient les données enregistrées en 2021. A la suite de ce 'spilt', la variable Saison seule variable catégorielle est encodée et toutes les variables numériques sont centrées-réduites par la méthode StandardScaler().

2.2. Méthodes

Dans le cadre de notre projet nous avons testés différents modèles afin de trouver le plus robuste et le plus performant applicable en entreprise :

- **LASSOCV**
- **RIDGE REGRESSION**
- **EasticNetCV**

- **SVR KERNEL**
- **RANDOM FOREST REGRESSOR**
- **SGD**

Nous avons lancé plusieurs itérations pour chaque modèle entre lesquelles nous nous sommes aidés de trois méthodes pour sélectionner les variables les plus explicatives et donc réduire le surapprentissage du ML et améliorer sa performance.

- **L'Analyse en Composantes Principales (ACP)**
- **La matrice de corrélation**
- **SelectKBest**

Enfin afin d'évaluer les performances de chaque modèle et de pouvoir l'interpréter nous nous sommes servis des métriques **score** et **RMSE** et de la bibliothèque **XgBoost.**

2.2.1. Les modèles[8]

- **LASSO**

La régression LASSO (Least Absolute Shrinkage and Selection Operator) est un type d'analyse de régression dans laquelle la sélection et la régulation des variables ont lieu simultanément. Cette méthode utilise une pénalité nommée alpha (α) qui affecte leur valeur des coefficients de régression. Plus la pénalité augmente, plus les coefficients s'approchent de zéro et vice versa. La régression de LASSO est une méthode que nous pouvons utiliser pour adapter un modèle de régression lorsque la « multicolinéarité » est présente dans les données. La régression des moindres carrés tente de trouver des estimations de coefficients qui minimisent la somme des résidus au carré. LASSO va donc permettre une sélection des variables les plus explicatives en leur donnant à chacune un poids, celles dont le poids sera égal à zéro seront supprimer.

- **LASSOCV**

Avec LASSOCV on va en plus réaliser une validation croisée (CV), c'est-à-dire une division du jeu de données en un nombre de sous-ensemble qu'on définit afin de créer un modèle sur chaque échantillon et de retourner un jeu de statistiques pour

[8] https://scikit-learn.org

chaque échantillon. Le résultat retourné sera celui qui aura obtenu le meilleur score en fonction des valeurs d'alpha testées, c'est-à-dire le meilleur coefficient de détermination R^2.

- **RIDGE REGRESSION[9]**

La régression des crêtes est une technique pour analyser les données de régression multiples qui souffrent de multicolinéarité. Quand la multicolinéarité se produit, les estimations des moindres carrés (utilisées par LASSO) sont sans biais, mais leurs variances sont importantes, de sorte qu'elles peuvent être loin de la valeur réelle. En ajoutant un certain biais noté lambda (λ) aux estimations de régression, la régression de crête réduit les erreurs-types. On espère que l'effet net sera de fournir des estimations plus fiables.

- **ElasticNetCV**

Elastic Net est une combinaison des deux modèles précédents LASSO et RIDGE qui consiste à éviter la sélectivité trop forte de LASSO et conserver les variables fortement corrélées. Nous devons régler les paramètres pour identifier les meilleures valeurs alpha et lambda et pour cela, nous devons utiliser le pack caret. Nous ajustons le modèle par validation croisée en itérant sur un certain nombre de paires alpha et lambda afin d'obtenir la paire avec la plus faible erreur associée.

- **SVR [10]**

SVR un modèle de regression construit sur la base du concept de Support Vector Machine ou SVM. C'est l'un des modèles populaires de ML qui peut être utilisé dans les problèmes de classification ou d'attribution de classes lorsque les données ne sont pas linéairement séparables. Afin de trouver un modèle qui donne des résultats cohérents nous procédons différemment avec d'autres modèles et un réglage d'hyperparamètres. La construction du modèle nécessite :

- Un dictionnaire contenant les valeurs possibles pour les hyperparamètres
- Un classifieur[11] à partir de la grille de paramètres
- Entrainer ce classifieur sur X_train_scaled

[9] https://ncss-wpengine.netdna-ssl.com
[10] https://stackoverflow.com
[11] Un classifieur linéaire représente une famille d'algorithmes de classement statistique.

- Afficher toutes les combinaisons possibles d'hyperparamètres et la performance moyenne du modèle associé
- Afficher les meilleurs paramètres de la grille pour le modèle

- **RANDOM FOREST REGRESSOR[12] (RFR)**

La régression forestière aléatoire est une technique d'apprentissage d'ensembles. Dans l'apprentissage d'ensembles, on prend plusieurs algorithmes ou le même algorithme plusieurs fois et on construit un modèle qui est plus puissant que l'original. La prédiction basée sur les arbres est plus précise parce qu'elle tient compte de nombreuses prédictions. Un « Random Forest »[13] a besoin de trois hyper-paramètres principaux (paramètres fixes), qui doivent être définis avant l'entraînement :

- la taille des arbres (le nombre de nœuds maximal)
- le nombre d'arbres à utiliser
- le nombre de caractéristiques échantillonnées (nombre de variables aléatoires choisies à chaque mélange depuis les variables explicatives).

À partir de là, le modèle peut être utilisé pour résoudre les problèmes de régression ou de classification.

1 - La première étape consiste à appliquer le principe du bagging, c'est-à-dire créer de nombreux sous-échantillons aléatoires de notre ensemble de données avec possibilité de sélectionner la même valeur plusieurs fois. (Saison, population …)

2 - Des arbres de décision individuels sont ensuite construits pour chaque échantillon. Chaque arbre est entraîné sur une portion aléatoire afin de recréer une prédiction. Notons bien que ces modèles-là sont très peu corrélés, et chaque arbre de décision fonctionne individuellement et indépendamment des autres. La combinaison de tous ces modèles indépendants permet de réduire la variance du modèle d'ensemble (plus stable, moins chaotique).

3 - Enfin, chaque arbre va prédire un résultat (target). Le résultat avec le plus de votes (le plus fréquent) devient le résultat final de notre modèle. Dans le cas de régression, on prendra la moyenne des votes de tous les arbres.

[12] https://scikit-learn.org
[13] https://blent.ai

- **SGD « Regressor »**

Le « Stochastic gradient descent »[14] SGD Regressor est une méthode itérative pour optimiser une fonction objective avec des propriétés de lissage appropriées.

Elle peut être considérée comme une approximation stochastique[15] de l'optimisation de la descente en gradient puisqu'elle remplace le gradient réel (calculé à partir de l'ensemble de données) par une estimation de celui-ci (calculée à partir d'un sous-ensemble de données choisi au hasard).

En particulier dans les problèmes d'optimisation de haute dimension, cela réduit la charge de calcul très élevée, obtenant des itérations plus rapides (pour un taux de convergence plus faible.)

2.2.2. La sélection des variables

- **ACP**

La spécificité de l'ACP (Analyse en Composantes Principales) est de parvenir à expliquer une partie de la variance avec un minimum de facteurs.

En fait, l'ACP est une méthode bien connue de réduction de dimension qui va permettre de transformer des variables très corrélées en nouvelles variables décorrélées les unes des autres.

Le principe est simple : Il s'agit en fait de résumer l'information qui est contenue dans une large base de données en un certain nombre de variables synthétiques appelées : Composantes principales.

L'idée est ensuite de pouvoir projeter ces données sur l'hyperplan le plus proche afin d'avoir une représentation simple de nos données. Ainsi, les nouvelles variables, notamment les composantes principales, sont des combinaisons des variables d'origine qui sont extraites d'une manière spécifique. Elles sont non corrélées et de variance maximale.

- **Matrice de corrélation**

[14] https://towardsdatascience.com
[15] Probabilités appliquées aux statistiques (ou « en ligne »)

La corrélation est une mesure de la relation de linéarité qui unit deux variables :

- Corrélation positive : lorsque deux variables augmentent ou diminuent de la même manière.
- Corrélation négative : lorsque deux variables augmentent ou diminuent de manière opposée.

Les méthodes de corrélation « SciPy, NumPy » et Pandas sont rapides, complètes et bien documentées. (Pearson, Spearman et Kendall coefficients de corrélation). Grâce à la méthode ***seaborn.heatmap*** on peut afficher une matrice colorée des coefficients de corrélation calculés entre chaque paire de variable.

- **SelectKBest**[16]

« Scikit-learn API »[17] fournit la classe SelectKBest pour extraire les meilleures fonctionnalités d'un ensemble de données donné.

La méthode SelectKBest sélectionne les fonctionnalités en fonction du score k le plus élevé. En modifiant le paramètre 'score_func', nous pouvons appliquer la méthode pour les données de classification et de régression. Le choix des meilleures fonctionnalités est un processus important lorsque nous préparons un grand ensemble de données. Il nous aide à éliminer une partie moins importante des données et à réduire le temps d'exécution.

2.2.3. Méthode d'évaluation et d'interprétation

- **RMSE (Root Mean Squared Error)**

$$\mathbf{RMSE} = \sqrt{\Sigma(P_i - O_i)^2 / n}$$

Où :

- P_i est la valeur de la prédiction de la $i^{ème}$ observation
- O_i est la valeur de la $i^{ème}$ observation
- n est la taille de l'échantillon

[16] https://scikit-learn.org
[17] https://www.datatechnotes.com

Il s'agit d'un indicateur pertinent. Cet indice fournit une indication par rapport à la dispersion ou la variabilité de la qualité de la prédiction. Le RMSE[18] peut être relié à la variance du modèle. La comparaison entre sa valeur pour l'ensemble d'apprentissage et celle pour l'ensemble test est un bon indicateur d'un effet de sur-apprentissage ou de sous-apprentissage qui rend le modèle moins robuste.

- **SCORE**[19]

Le score ou R^2 score ou coefficient de détermination, est l'une des métriques les plus utilisées pour la régression linéaire. C'est en fait une version "normalisée" de la MSE (Mean Squared Error). Le R^2 peut se définir simplement comme l'erreur du modèle qu'on teste divisé par l'erreur d'un modèle basique qui prédit tout le temps la moyenne de la variable à prédire. Plus le modèle est performant plus il est élevé. Il vaut au maximum 100%, lorsque toutes les prédictions sont exactes. Plus sa valeur est proche de 0 plus il se rapproche d'un modèle simple de régression linéaire prédisant tout le temps la valeur moyenne. S'il est négatif alors les prédictions sont moins bonnes que si l'on prédisait systématiquement la valeur moyenne.

- **Xgboost**[20]

XGBoost, qui signifie Extreme Gradient Boosting, est une bibliothèque d'apprentissage automatique évolutive et distribuée. Elle fournit une arborescence parallèle et c'est la principale bibliothèque d'apprentissage automatique pour les problèmes de régression, de classification et de classement. Ses algorithmes s'appuient sur un apprentissage automatique supervisé, des arbres de décision, un apprentissage d'ensemble et un boosting de gradient.

[18] https://www.aspexit.com
[19] https://train.datascientest.com/
[20] https://www.nvidia.com/

L'apprentissage automatique supervisé utilise des algorithmes pour former un modèle et trouver des modèles dans un ensemble de données avec des étiquettes et des caractéristiques, puis utilise le modèle formé pour prédire les étiquettes sur les caractéristiques d'un nouvel ensemble de données.

Les arbres de décision créent un modèle en évaluant un arbre de questions de caractéristiques si-alors-vrai/faux, et en estimant le nombre minimum de questions nécessaires pour évaluer la probabilité de prendre une bonne décision.

Les arbres de décision peuvent être utilisés pour la classification pour prédire une catégorie, ou la régression pour prédire une valeur numérique continue.

2.3. Résultats et interprétation

Dans cette partie nous allons présenter les résultats et l'interprétation obtenu, grâce à l'application des modèles scikit-learn – régression, sur nos 3 jeux de données.

2.3.1. L'échelle horaire

- **1ère Itération**

Sur le jeu de données Heure on retrouve les données Horaires de 2016 à 2021, nous avons 349 030 observations et 27 variables. Les modèles à privilégier sont Lasso ou ElasticNet si on trouve peu de variables explicatives ou Ridge Regressor, SVR et Ensemble Regressors.

Avec ce jeu de données nous obtenons les résultats ci-dessous (Fig. 19) :

Résultats	LassoCV	ElasticNetCV	Random Forest Regressor	Ridge_cv	SGD Regressor
Rmse_train	827.622367	831.500335	153.140848	827.064206	827.401537
Rmse_test	1597.919243	1248.855443	648.150540	1750.129827	1679.792855
Score_train	0.820000	0.821554	0.993947	0.823453	0.823308
Score_test	0.565706	0.596266	0.891252	0.207112	0.269563

Figure 19 : Bilan des métriques d'évaluation des modèles après la 1ère itération pour la consommation horaire.

Sur tous les modèles on constate un effet important de sur-apprentissage. Tous les scores train sont bons (> 0,8) mais seul le score test du RFR est bon. Les scores des modèles LassoCV et ElasticNetCV étant corrects (>0.5), on choisit de sélectionner ces 3 derniers modèles pour les itérations suivantes. Afin d'améliorer ces modèles et de diminuer les effets de sur-apprentissage, nous testons plusieurs itérations selon trois méthodes de sélection de variables :

1. La Matrice des corrélations

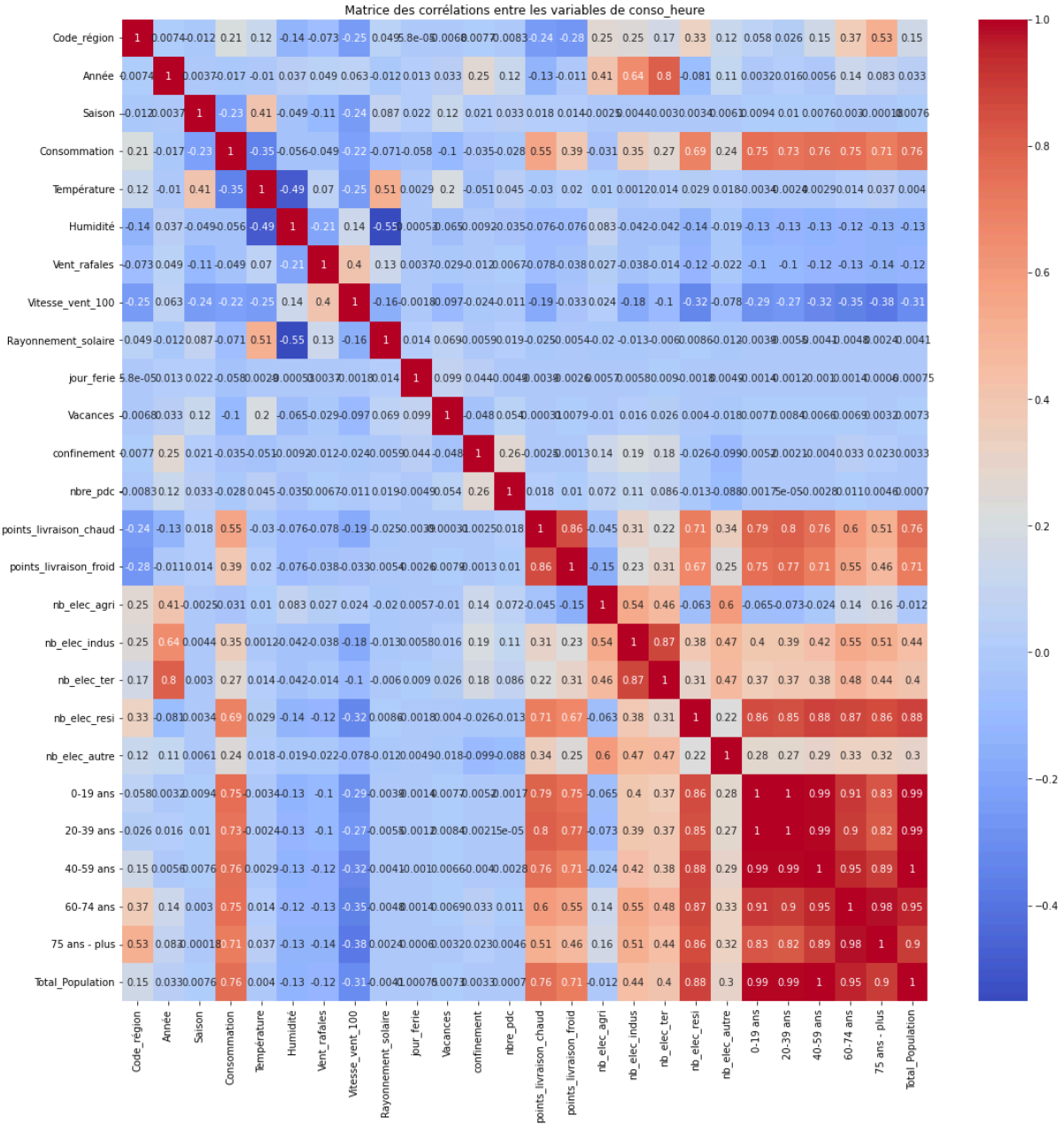


Figure 20 : Matrice des corrélations sur le jeu de données horaire

D'après la matrice des corrélations (Fig. 20) on a gardé les variables dont le coefficient de corrélation à la consommation est supérieur à 0,2 (sauf pour la variable

« Année » qu'on garde de toute façon qui nous sert à séparer les jeux de données. Pour les variables très corrélés entre elles, on a conservé la variable la plus corrélé à la consommation. On conserve donc : **Code_région, Année, Saison, Total_Population, nb_elec_autre, nb_elec_ter, nb_elec_indus, points_livraison_froid, points_livraison_chaud, Température et Vitesse_vent_100.**

2. L'ACP

80% de la variance est expliquée par les 4 premiers axes (Fig. 21)

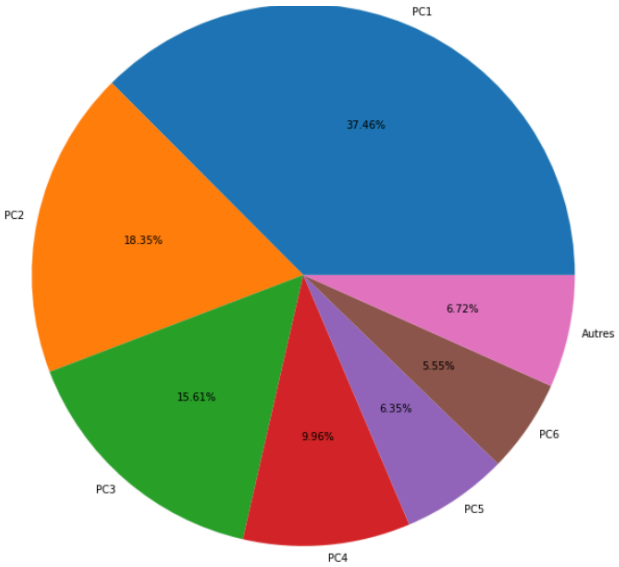


Figure 21 : Camembert affichant la proportion de variance portée par chaque composante principale (ou axe)

D'après l'analyse des cercles de corrélation (Fig. 22) réalisés pour les 4 premières composantes principales, on choisit de conserver les variables : **'Saison', 'Vitesse_vent_100','nb_elec_ter', 'nb_elec_indus', 'nb_elec_autre', 'Total_Population', 'Code_région' et 'points_livraison_froid.**

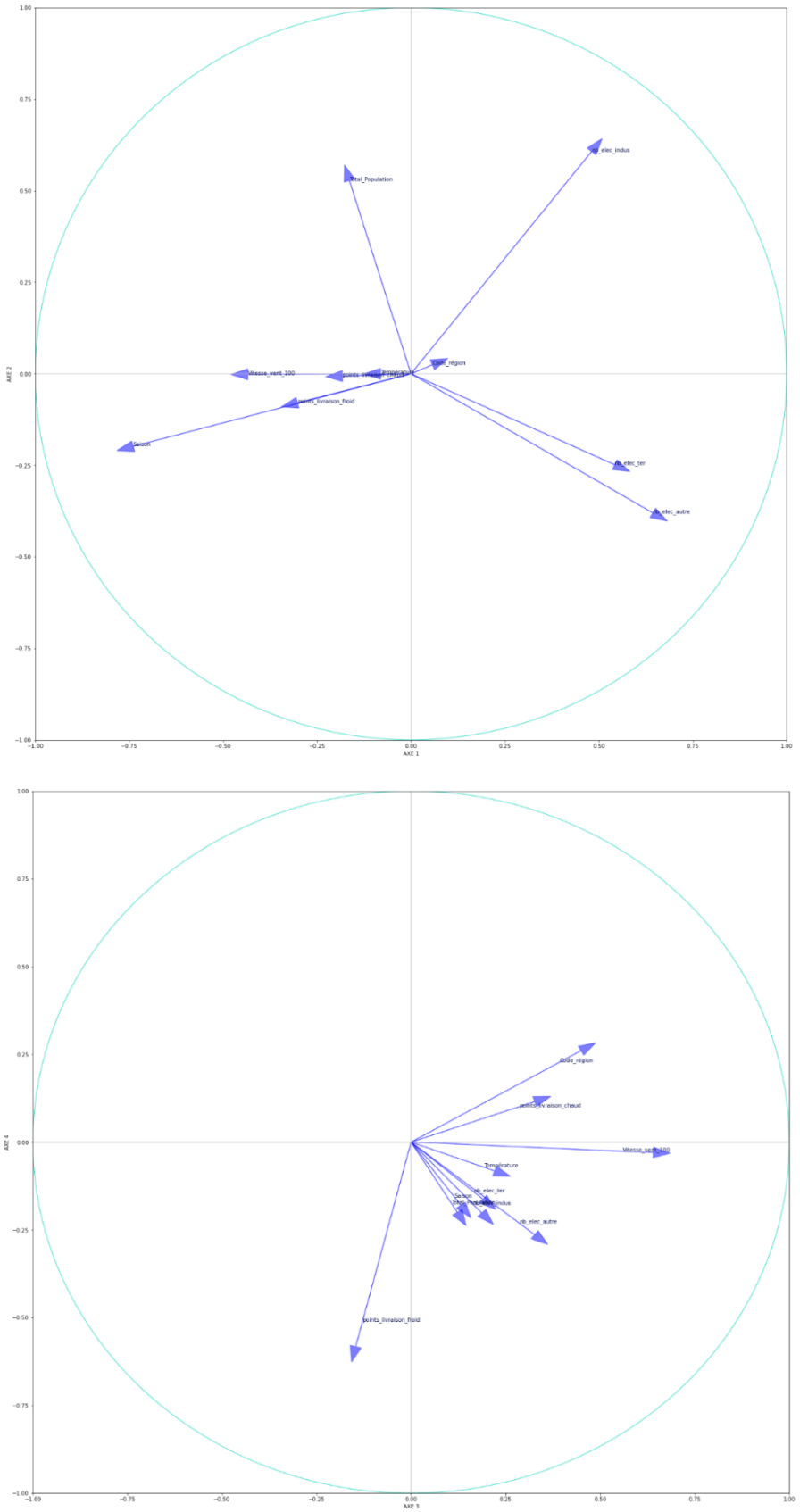


Figure 22 : Cercles de corrélation affichant les 4 premières composantes principales

3. Le SelectKBest

Pour finir on utilise le transformeur SelectKBest de Scikit Learn basé sur un test de Fisher pour obtenir les 5 variables les plus explicatives. Les variables sélectionnées sont : **Code region, Saison, Total _Population, Température, Vitesse_vent_100.**

- **Itération retenue**

Nous avons testé les modèles avec ces 3 sélections de variable et c'est la sélection du SelectKBest qui a obtenu les meilleurs résultats (Fig. 23)

Résultats	LassoCV	ElasticNetCV	Random Forest Regressor
Rmse_train	827.622367	1020.436765	251.205494
Rmse_test	1597.919243	1001.720903	881.034387
Score_train	0.731246	0.731246	0.983713
Score_test	0.740102	0.740244	0.799064

Figure 23 : Bilan des métriques d'évaluation des modèles pour l'itération retenue.

ElasticNetCV ne montre pas d'effet de surapprentissage et donne globalement de bons scores de prédictions. Cependant, il ne fonctionne pas bien pour prédire les consommations en région Ile-de-France, ce qui affecte les prédictions de la consommation moyenne (Fig. 24).

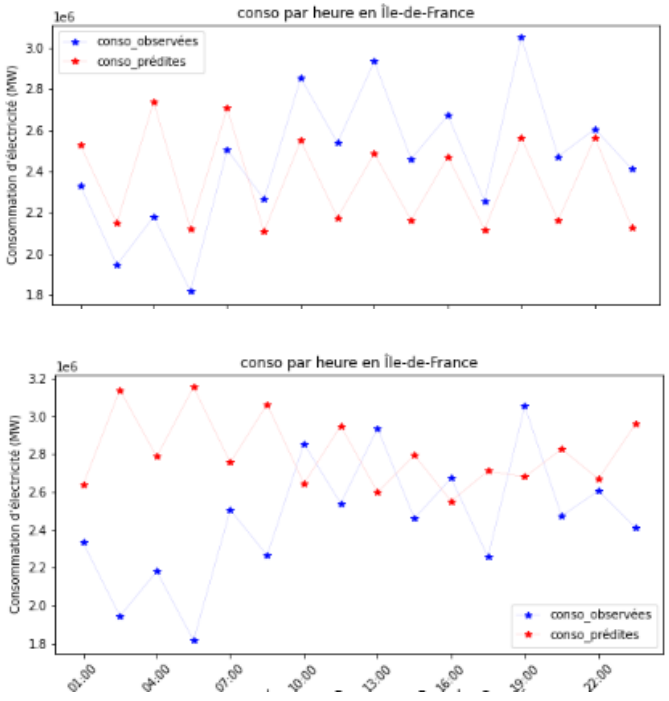


Figure 24 : Consommation horaire moyenne en Ile-de-France obsevée et prédite en 2021 (Modèle RFR en haut et ElasticNetCV en bas)

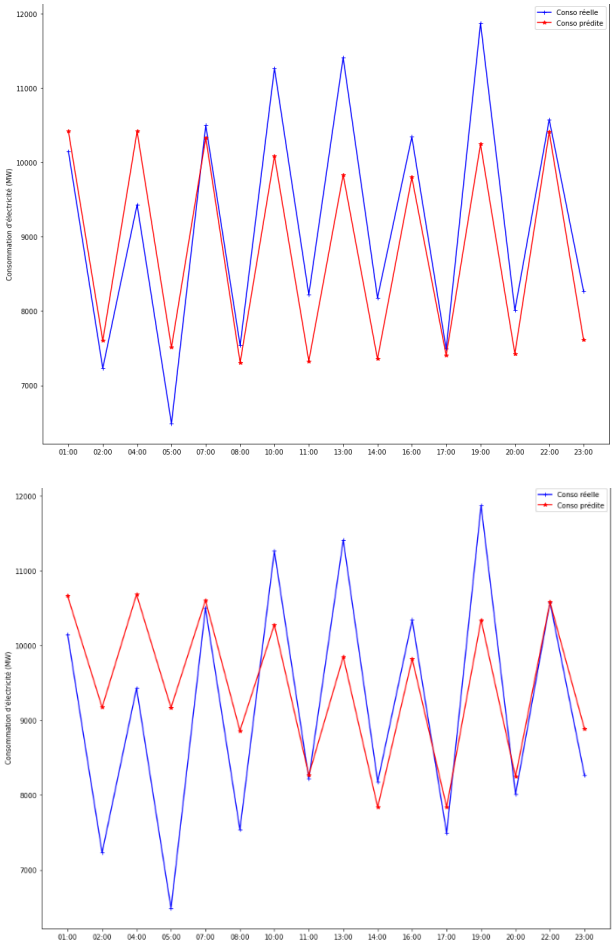


Figure 25 : Consommation moyenne observée et prédite par heure par le modèle RFR (à gauche) et ElasticNetCV (à droite) pour 2021

Alors que le modèle Random Forest Regressor donne de trés bons scores pour toutes les régions dont l'Ile de France mais montre toujours un effet de surapprentissage qui le rend moins robuste que l'autre modèle. Le choix métier se porte toujours sur le modèle le plus robuste applicable à un jeu de données régulièrement enrichi donc ici ElasticNetCV. Cependant, il serait peut-être intéressant de choisir un modèle spécifique aux données de la région Ile-de-France.

- **Interprétabilité**

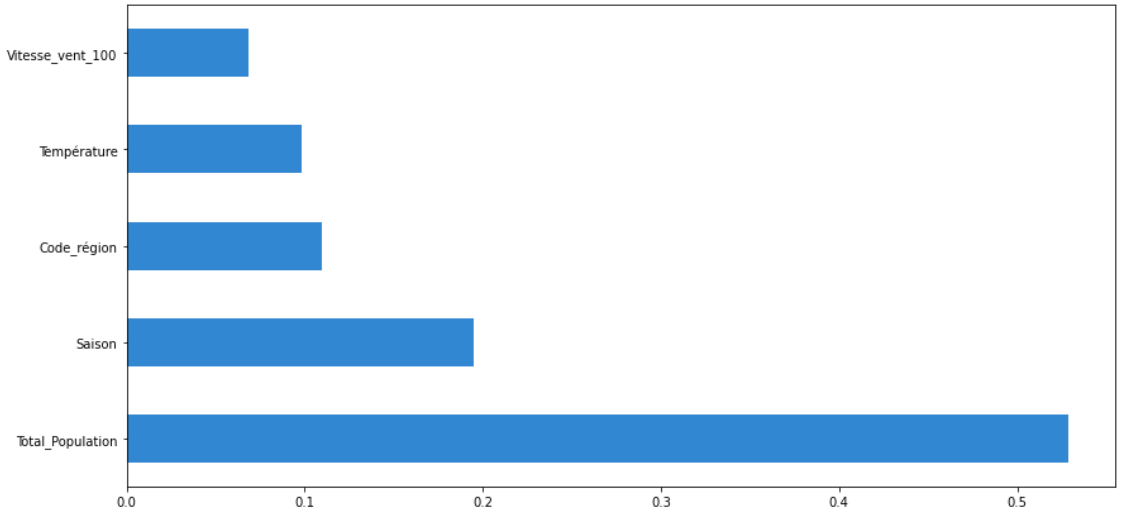


Figure 26 : Importance des variables dans la construction des modèles de ML pour la prédiction de la consommation horaire.

Les résultats de XgBoost exécuté avec le package Skater montrent que 'Total_Population' est la variable la plus explicative de la consommation. Son poids dans le modèle est de plus de 50% par rapport aux autres variables sélectionnées. Elle est suivie par la Saison puis la Région et la Température et la vitesse des vents à 100m.

2.3.2. L'échelle hebdomadaire

- **1ère Itération**

Sur le jeu de données Semaine on retrouve les données hebdomadaires de 2016 à 2021, nous avons 4 405 observations et 28 variables. Les modèles à privilégier sont Lasso, ElasticNet si on trouve peu de variables explicatives ou Ridge Regressor, SVR et Ensemble Regressors.

Avec ce jeu de données nous obtenons les résultats ci-dessous (Fig. 27) :

Résultats	LassoCV	ElasticNetCV	SVR linear	Random Forest Regressor	Ridge_cv	SGD Regressor
Rmse_train	554.122002	559.253455	577.639916	107.902637	556.657565	561.006529
Rmse_test	1388.566051	1265.416338	1260.019343	488.232862	1335.520590	1262.928427
Score_train	0.927033	0.926958	0.922076	0.997281	0.927634	0.926499
Score_test	0.514135	0.611437	0.614744	0.942157	0.567191	0.612963

Figure 27 : Bilan des métriques d'évaluation des modèles après la 1ère itération pour la consommation hebdomadaire

On voit que sur les modèles il y a beaucoup trop de surapprentissage avec un écart assez conséquent entre les scores train et test. On gardera pour les itérations suivantes les modèles ElasticNetCV, SVR linear, Random Forest Regressor et SGD Regressor qui possèdent un score test au-dessus de 0,6. Comme pour la consommation horaire, on exécute d'autres itérations après sélection de variables.

1. La Matrice des corrélations

D'après la matrice des corrélations, on conserve les mêmes variables que pour la consommation horaire mais avec une variable de plus : le rayonnement solaire.

2. L'ACP

Dans le cas hebdomadaire, 80% de la variance est expliquée par les 3 premiers axes et d'après l'analyse des cercles de corrélation on choisit de conserver les variables : **'Saison', 'nb_elec_autre', 'Total_Population', 'nb_elec_ter', 'points_livraison_chaud', 'Rayonnement_solaire' et 'points_livraison_froid.**

3. Le SelectKBest

Les variables sélectionnées sont : **Saison, Total_Population, nb_elec_autre, Rayonnement_solaire et Température.**

- **Itération retenue**

Nous avons testé les modèles avec ces 3 sélections de variable et c'est la sélection du SelectKBest qui a obtenu les meilleurs résultats (Fig. 28)

Résultats	SVR linear	Random Forest Regressor	SGD Regressor	ElasticNetCV
Rmse_train	889.489405	123.149292	856.313122	905.885003
Rmse_test	921.988197	425.309399	865.268490	897.331277
Score_train	0.815227	0.996458	0.828754	0.808353
Score_test	0.793725	0.956106	0.818324	0.804611

Figure 28 : Bilan des métriques d'évaluation des modèles pour l'itération retenue pour la consommation hebdomadaire.

Pour l'ensemble des modèles que nous avons gardés pour l'itération finale, on obtient de meilleurs scores avec moins de sur-apprentissage. SGD Regressor semble être le meilleur modèle : il n'est plus altéré par un effet de sur-apprentissage avec un score train de 0.83 et un score test de 0.81. Il est performant les prédictions

sont généralement bonnes (Fig. 29). C'est par contre dans les régions Ile-de-France et PACA (Fig. 30) en été qu'on note le plus gros écart entre prédictions et observations réelles en 2021. On peut supposer qu'il manque au modèle d'autres variables explicatives comme le nombre de touristes dont l'impact peut jouer à certaines saisons.

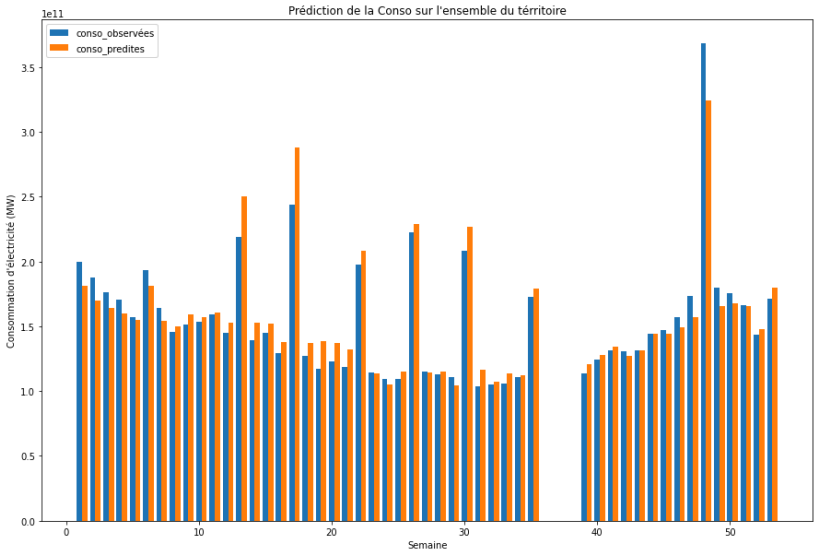


Figure 29: Consommation moyenne hebdomadaire observée et prédite en 2021 par le modèle SGD Regressor

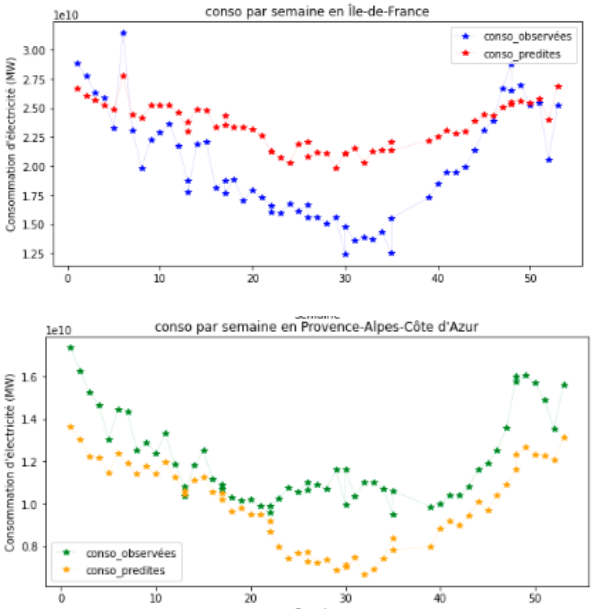


Figure 30 : Consommations hebdomadaires observées e prédites en Ile de France et en PACA

- **Interprétabilité**

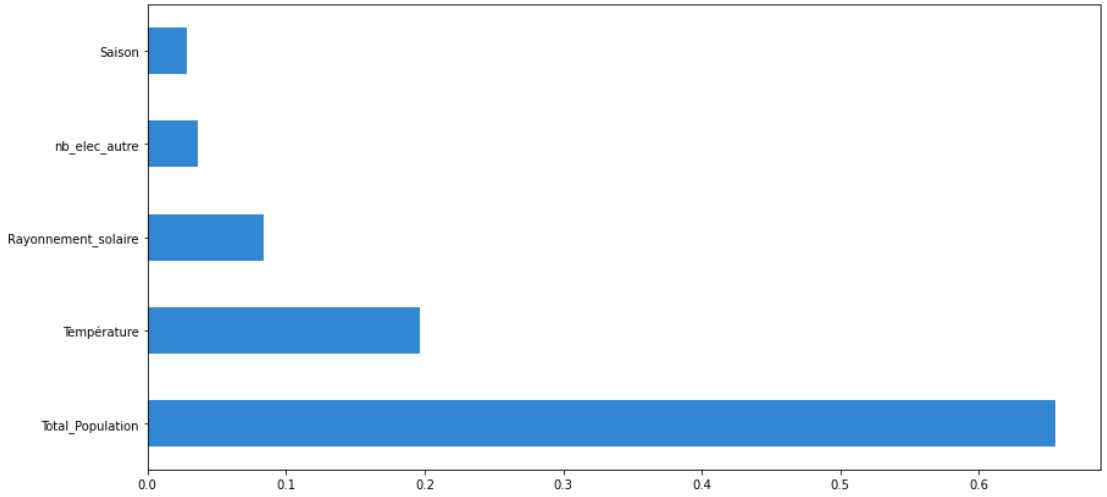


Figure 31 : Importance des variables dans la construction des modèles de ML pour la prédiction de la consommation hebdomadaire

Les résultats de XgBoost exécuté avec le package Skater montrent que 'Total_Population' est toujours la variable la plus explicative de la consommation. Son poids dans le modèle est de plus de 60% par rapport aux autres variables sélectionnées. Elle est suivie par la Température puis le Rayonnement Solaire, le nb_elec_autre (qui représente le nombre de branchements électriques liés aux transports) et enfin la Saison.

2.3.3. L'échelle mensuelle

- **1ère Itération**

Sur le jeu de données Mois on retrouve les données mensuelles de 2016 à 2021, nous avons 857 observations et 27 variables. Les modèles à privilégier sont Lasso ou ElasticNet si on trouve peu de variables explicatives ou Ridge Regressor, SVR et Ensemble Regressors.

Avec ce jeu de données nous obtenons les résultats ci-dessous (Fig. 32) :

Résultats	LassoCV	ElasticNetCV	SVR linear	Random Forest Regressor	Ridge_cv	SGD Regressor
Rmse_train	443.846696	451.769447	543.965757	83.209759	449.298107	481.369703
Rmse_test	1478.180398	1318.360824	975.152615	506.250619	1370.904638	1081.707169
Score_train	0.951549	0.951359	0.929480	0.998350	0.951890	0.944776
Score_test	0.486911	0.563375	0.761117	0.935617	0.527877	0.706059

Figure 32 : Bilan des métriques d'évaluation des modèles après la 1ère itération pour la consommation mensuelle

Cette fois-ci on retient les modèles SVR linear, Random Forest Regressor et SGD Regressor qui obtiennent les scores les plus élevés mais on tentera de réduire l'effet de sur-apprentissage avec une sélection des variables les plus explicatives.

1- La Matrice des corrélations

On conserve les mêmes variables que pour la consommation hebdomadaire à savoir : **Année, Saison, Total population, nb_elec_autre, nb_elec_ter, ne_elec_indus, point_livraison_froid, points, points_livraison_chaud, Rayonnement_solaire, Température et Code_région.**

2- L'ACP

De même que précédemment, 80% de la variance est expliquée par les 3 premiers axes et d'après l'analyse des cercles de corrélation on choisit de conserver les variables : **'Saison', 'nb_elec_autre', 'Total_Population', 'nb_elec_ter', 'points_livraison_chaud', 'Rayonnement_solaire', 'points_livraison_froid.**

3- Le SelectKBest

Les variables sélectionnées sont : **Saison, Total_Population, nb_elec,ter, Rayonnement_solaire et Température.**

- **Itération retenue**

Nous avons testé les modèles avec ces 3 sélections de variable et c'est la sélection du SelectKBest qui a obtenu les meilleurs résultats (Fig. 33)

Résultats	SVR linear	Random Forest Regressor	SGD Regressor
Rmse_train	847.169553	89.348828	815.350950
Rmse_test	858.998010	396.344211	830.084000
Score_train	0.828955	0.998097	0.841562
Score_test	0.814636	0.960537	0.826905

Figure 33 : Bilan des métriques d'évaluation des modèles pour l'itération finale pour la consommation mensuelle

Les modèles SVR Linear et SGD Regressor donnent des résultats quasi-identiques dont l'effet de sur-apprentissage a été supprimé. Les scores des prédictions sont bons : 0.83 pour SGD Regressor sur le train contre 0.81 pour SVR linear. Ils semblent donc à la fois robustes, performants et adaptés aux données mensuelles.

Cependant, on remarque tout de même sur les graphiques des différences selon les régions avec plus ou moins d'écart entre les consommations réelles et prédites, notamment en été. La région Ile-de-France est encore une fois la région la moins bien prédite. Le modèle Random Forest Regressor est le seul des modèles testés à parvenir à de bonnes prédictions pour toutes les régions mais l'effet de sur-apprentissage le rend moins robuste que les autres modèles quelle que soit l'échelle de temps considérée (Fig. 34, 35).

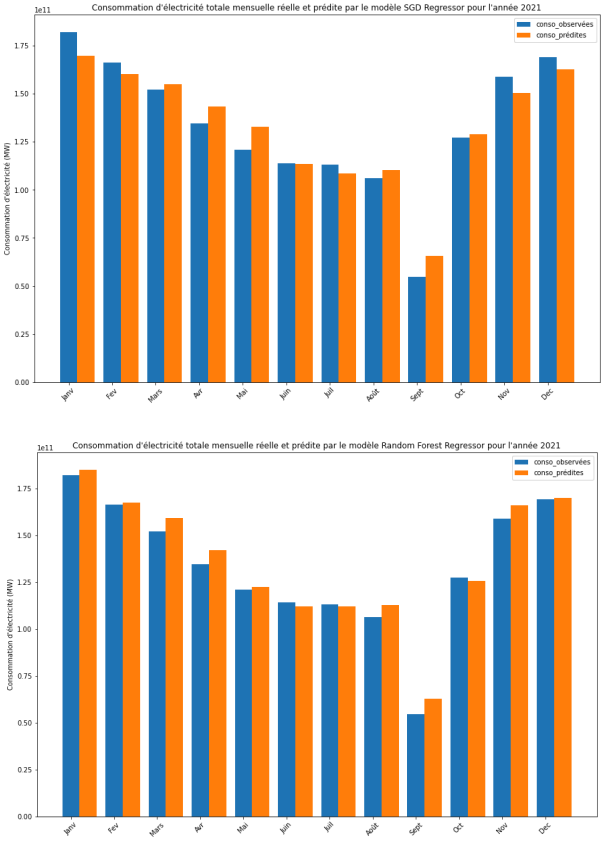


Figure 34 : Consommation moyenne mensuelle observée et prédite pour 2021 par SGD Regressor à gauche et RFR à droite

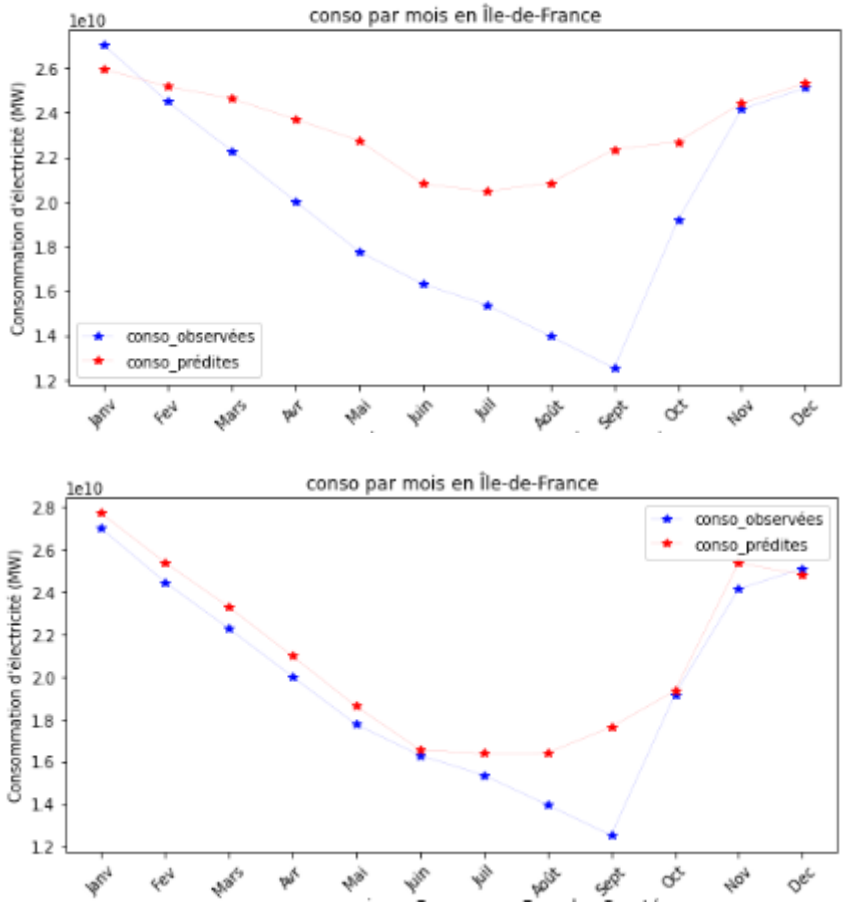


Figure 35 : Consommations mensuelles observées et prédites pour 2021 en Ile-de-France par SGD Regressor en haut et RFR en bas

- **Interprétabilité**

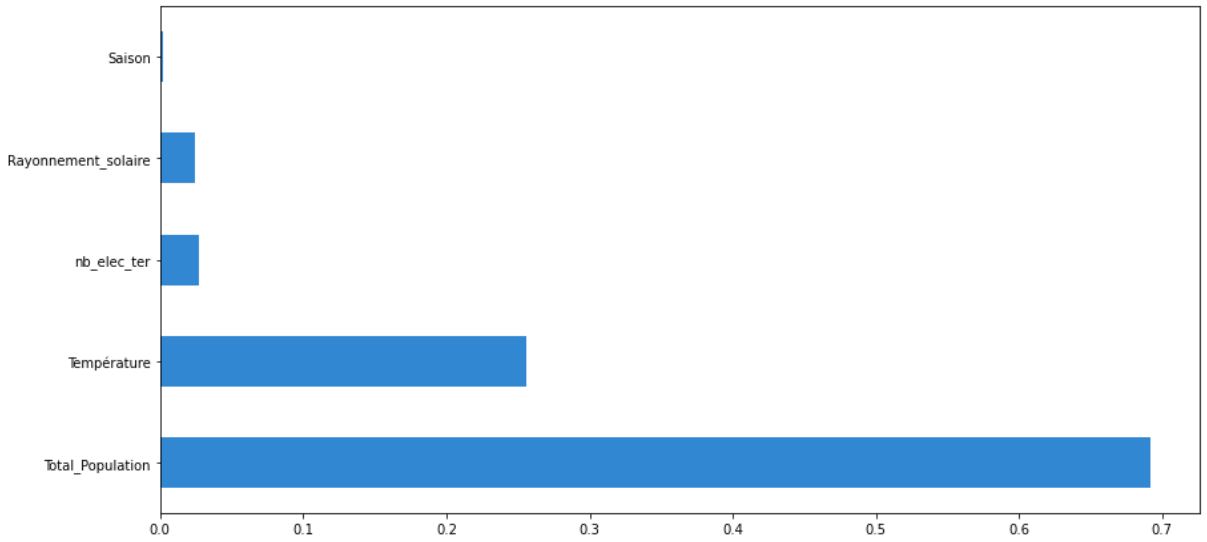


Figure 37 : Importance des variables dans la construction des modèles de ML pour la prédiction de la consommation hebdomadaire

Les résultats de XgBoost exécuté avec le package Skater montrent que 'Total_Population' est toujours la variable la plus explicative de la consommation et de loin. Son poids dans le modèle est de presque 70% par rapport aux autres variables sélectionnées. Elle est suivie par la Température (25%) puis en moindre proportion par le nb_elec_ter (les points de branchement dans le tertiaire), le Rayonnement Solaire et la Saison.

4- CONCLUSION

Aujourd'hui, la modélisation des données est d'une importance capitale dans la prise de décision pour les entreprises comme pour les gouvernements. Elle établit la structure des données disponibles tout en établissant un processus de limitation des erreurs et permet de prédire l'évolution de ces données. La construction et le choix des modèles sont donc des étapes à ne pas négliger. Il s'agit de garantir l'équité du modèle, la responsabilité et la fiabilité du modèle et la transparence du modèle.

Au cours de notre projet fil rouge nous avions plusieurs objectifs et pour y répondre nous avons exploré, nettoyé, remanié de grands jeux de données auxquels nous avons appliqués plusieurs modèles de Machine Learning à plusieurs échelles de temps. Les résultats obtenus nous ont montré que la prédiction de la consommation d'électricité à l'échelle d'un pays mais également à l'échelle d'une région était très complexe. Les variables explicatives de cette consommation varient en fonction de l'échelle de temps ciblée mais on retrouve toujours trois variables parmi les plus explicatives auxquelles on s'attendait au départ : la densité de population totale, la température et la saison. De plus, quelle que soit l'échelle de temps ciblée, on retrouve toujours la même difficulté pour les modèles à prédire la consommation en région Ile-de-France et encore plus particulièrement en été. Il faudrait maintenant tester chaque modèle sur cette région uniquement et voir si la sélection des variables est la même que sur le jeu de données national et peut-être enrichir le jeu régional d'autres variables comme la fréquentation touristique de la région.

Ce que nous retenons de ce projet c'est qu'il n'est pas simple de prédire l'avenir mais que ces modèles de prédiction sont essentiels à notre préparation et adaptation aux changements qu'ils soient climatiques, économiques, politiques ou sociétaux., d'autant plus dans le secteur de l'énergie dont la consommation mais également la

production et les échanges sont dépendants de nombreux facteurs environnementaux, sociétaux et géopolitiques. Ces trois dernières années en sont les parfaites témoins entre confinements, aléas climatiques et guerres.

Le dernier bilan mensuel livré par le gestionnaire du RTE[21] indique que la production française d'électricité hydraulique était en baisse de 29% en avril, sur un an, retombant ainsi à son plus bas niveau depuis 1976. Depuis le 1er janvier, la production a baissé de 22,9% comparativement à 2010. Habituellement, les centrales hydroélectriques fournissent 12% de l'électricité française. Avec la sécheresse, lorsque les débits des cours d'eau diminuent les centrales nucléaires doivent également tourner au ralenti pour respecter les normes concernant la protection des milieux aquatiques. Toutefois, 14 centrales sur les 58 que compte le territoire français "sont situées en bord de mer, ce qui permet d'assurer une base d'approvisionnement non affectée par la baisse de débit des rivières"[22]

Ces sécheresses record nous amènent à n interroger : **Quels sont les enjeux de la transition énergétique aujourd'hui ?** Les énergies renouvelables ont une incidence sur les cinq variables du développement durable, à savoir le social, l'environnemental, l'économique, la politique et la géopolitique. Moins nocives pour l'Homme, les énergies renouvelables permettraient de tendre vers : plus d'indépendance énergétique vis à vis des pays exportateurs d'énergies fossiles, un impact environnemental moindre, des économies locales revalorisées.

5- CRITIQUES

La première étape où nous aurions pu accumuler des erreurs est l'étape de fusion des différents jeux de données et leurs transformations sur différentes échelles de temps. Cette étape nous a pris beaucoup de temps, elle a d'abord créé beaucoup de doublons et de valeurs manquantes, les échelles de temps de chaque jeu de données étant différentes. Nous avons mis du temps à comprendre comment éviter les erreurs de fusion et à obtenir trois jeux de données propres. Cependant, nous avons surement perdu beaucoup d'information dans cette étape (notamment en

[21] https://bilan-electrique-2021.rte-france.com/
[22] https://www.maxisciences.com

termes d'années) et transformé de l'information en remplaçant des valeurs manquantes.

L'autre étape qui a pu exercer une influence sur la performance de nos modèles est notre gestion des valeurs aberrantes. Nous n'avons pas regardé toutes les distributions nationales et régionales de chaque variable et donc probablement laissé trop de valeurs aberrantes dans les jeux de données. On a par exemple remarqué lors de nos itérations de ML que la consommation en région Ile-de-France était souvent mal modélisée. Or, quand on regarde plus en détail la distribution de la consommation horaire par région, on s'aperçoit qu'il y a deux régions où les valeurs aberrantes sont beaucoup plus élevées que dans les autres régions (> 10 000 MW) : la région Ile-de-France (70%) des valeurs aberrantes et la région Auvergne-Rhône-Alpes (30%) (**Fig. 38**). Il aurait peut-être été intéressant de séparer ces régions et notamment la région Ile-de-France et de les modéliser comme des cas particuliers. Les variables les plus corrélées à la consommation sont peut-être différentes.

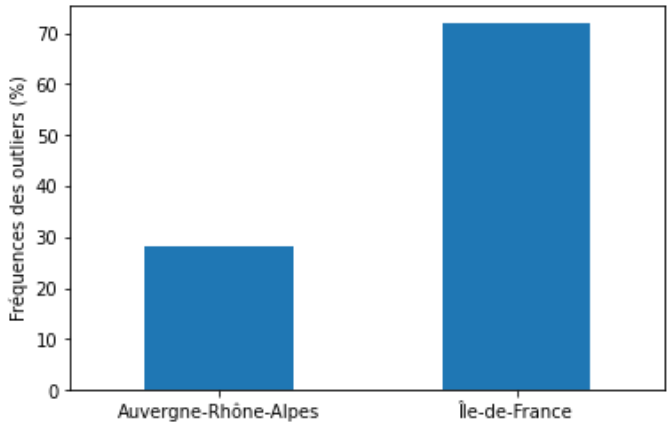


Figure 38 : Fréquence des consommations par demi-heure supérieures à 10 000 MW par région

Enfin, une autre variable qui nous amène à questionner la fiabilité de nos données est la température. En effet, quand on compare son évolution temporelle à celle de la consommation, quelle que soit l'échelle de temps, on observe qu'elles sont en presque parfaite opposition. Quand la température augmente la consommation diminue et vice-versa. Or, les matrices de corrélations comme les

ACP montrent des corrélations négatives certes mais assez faibles entre les deux variables. Les régressons linéaires nous expliquent que le coefficient de corrélation est altéré par les valeurs extrêmement hautes des températures (**Fig. 10**). Il aurait donc peut-être été pertinent d'afficher la distribution de la température et de comprendre ses outliers. Le SelectKBest (basé sur un test de Fisher) nous a confirmé l'importance de cette variable dans nos modèles de prédiction et l'écarter a en effet diminué la performance de nos modèles. Le contexte climatique actuel nous pousse tout particulièrement à comprendre la relation complexe qu'il existe entre température et consommation d'énergie. C'est pourquoi nous présenterons lors de notre soutenance un modèle adapté aux séries temporelles pour prédire la consommation à partir de la température.