DataScientest • com

**Rapport du projet « eC2F_py »**

# Prévision de la consommation d'électricité en France

**Aurélien BOYER**

**Reda KAOUA**

**Geoffroy LADONNE**

*Février 2021*

# Table des matières

# *Introduction générale*

L'électricité est produite à partir de plusieurs sources d'énergies dites « primaires », disponibles dans la nature. Ces sources si elles ne sont pas utilisées directement, doivent être transformées en source d'énergie secondaire pour être exploitées et transformées en électricité.

En France, il existe trois filières de production d'électricité, classées en fonction de l'énergie primaire utilisée pour sa production :

- Les énergies fossiles : obtenues essentiellement grâce à la combustion d'hydrocarbures, charbon et gaz.

- L'énergie nucléaire : obtenue lors des réactions de fission nucléaire des noyaux atomiques au sein d'un réacteur nucléaire.

- Les énergies renouvelables : obtenues par des sources dont le renouvellement naturel est assez rapide pour qu'elles puissent être considérées comme inépuisables, comme le soleil, le vent, l'hydraulique et les bioénergies.

A l'échelle régionale, du fait des spécifications météorologiques et du parc de production installé, le réseau électrique permet d'assurer les échanges avec les régions limitrophes puisque certaines régions produisent plus qu'elles ne consomment.

Il en va de même pour les pays limitrophes où la France reste le 1$^{er}$ exportateur d'électricité européen en 2019.

Une des principales problématiques d'un réseau de production électrique, est le stockage de l'électricité produite. En effet, actuellement le stockage électrique utilise des stations de pompage-turbinage entre deux retenues d'eau situées à deux altitudes différentes, où l'eau est pompée vers le bassin supérieur pendant les heures de faible consommation alors que pendant les heures de pointe, l'eau passe dans une turbine qui produit un appoint d'électricité sur le réseau.

Concrètement, cette technique de stockage consomme plus d'énergie pour le pompage de l'eau que le turbinage n'en crée, et engendre des pertes d'énergie allant de 15 à 30% [1].

[1] https://www.ecosources.info/dossiers/Station_stockage_transfert_pompage_turbinage

Afin de répondre à cette problématique de stockage, il est nécessaire de pouvoir adapter la production en fonction de la consommation.

Et c'est dans cette démarche que s'inscrit notre projet « **eC2F_Py** », qui a pour but d'analyser et de prédire la consommation électrique afin d'adapter la production d'électricité au niveau régional.

Pour la réalisation de ce projet, nous nous sommes fixés les objectifs suivants :

- Analyse de la consommation énergétique au niveau national et régional.
- Analyse de la production énergétique par filière de production.
- Analyse approfondie sur la production des énergies renouvelables.
- Mise en place d'un modèle de prédiction de la consommation énergétique au niveau régional.

Dans ce document, nous commencerons par présenter dans la partie I, l'analyse exploratoire de nos données, où on présentera le jeu de donnée utilisé et les différentes variables, ainsi que les différentes étapes de prétraitement que nous avons réalisé.

Dans la 2e partie, nous détaillerons les différentes analyses statistiques et les visualisations sur la consommation et la production électrique au niveau national et régional.

La 3e partie sera entièrement consacrée aux différents modèles de Machine Learning que nous avons implémenté et les différents résultats obtenus, et ainsi nous expliquerons notre choix sur le modèle retenu.

# *Partie I : Analyse exploratoire des données*

*Dans cette partie d'analyse exploratoire, nous présenterons le jeu de donnée utilisé et les différentes variables, ainsi que les différentes étapes de prétraitement que nous avons réalisé.*

## I.1 – Présentation du jeu de données

Le jeu de données utilisé, provient de l'Open Data Réseaux Energie (ODRE), il représente les données énergétiques régionales (au pas de 30 minutes) consolidées depuis janvier 2020 et définitives de janvier 2013 à décembre 2019, elles sont issues de l'application éCO2mix [2].

Le jeu de données présente les données énergétiques régionales telles que :

- La consommation réalisée
- La production selon les différentes filières composant le mix énergétique (thermique, nucléaire, dites 'renouvelables', pompage)
- Le solde des échanges avec les régions limitrophes

Ces données sont élaborées à partir des comptages et complétées par des forfaits. Les données sont dites consolidées lorsqu'elles ont été vérifiées et complétées (livraison en milieu de M+1). Elles deviennent définitives lorsque tous les partenaires ont transmis et vérifié l'ensemble des comptages, (livraison deuxième trimestre A+1).

Les données publiées sur le portail « www.rte-france.com » sont publiques et leur réutilisation est permise sous réserve de mentionner la source.

Le Data Set utilisé lors de notre étude, regroupe les données définitives depuis le 1er Janvier 2013 jusqu'au 31 décembre 2019, et consolidées depuis le 1er Janvier au 30 Novembre 2020 pour chacune des régions en France métropolitaine (hors Corse).

Le Data Set ainsi utilisé contient 1.665.216 lignes et 66 colonnes.

## I.2 – Description des différentes variables du data set

Avec ses 66 colonnes, le data set se compose des variables suivantes :

| COLONNES | NOM DE LA VARIABLE | FORMAT | DESCRIPTION |
|---|---|---|---|
| 0 | **Code INSEE Région** | Entier | Code INSEE de la région |
| 1 | **Région** | Texte | Nom de la région |
| 2 | **Nature** | Texte | Nature de la donnée (définitive ou consolidée) |
| 3 - 5 | **Date, Heure et Date – Heure** | aaaa-mm-jj HH:MM | Date du jour, heure de l'échantillon et le fuseau horaire |
| 6 | **Consommation (MW)** | entier | Consommation en MW |
| 7 | **Thermique (MW)** | Entier | Production thermique en MW |
| 8 | **Nucléaire (MW)** | Entier | Production nucléaire en MW |
| 9 | **Eolien (MW)** | Entier | Production éolienne en MW |
| 10 | **Solaire (MW)** | Entier | Production solaire en MW |
| 11 | **Hydraulique (MW)** | Entier | Production hydraulique en MW |
| 12 | **Pompage (MW)** | Entier | Pompage hydraulique en MW |
| 13 | **Bioénergies (MW)** | Entier | Production Bioénergies en MW |
| 14 | **Ech. Physiques (MW)** | Entier | Solde imports/exports (flux physiques) en MW |

[2] éCO2mix : projet de RTE-France (https://www.rte-france.com/eco2mix)

| | | | |
|---|---|---|---|
| **15-52** | **Flux physiques entre régions** | Entiers | Sommes des lignes électriques importatrices entre régions |
| **53-64** | **TCO et TCH de chaque type d'énergie** | Décimaux | TCO : Taux de Couverture<br>TCH : Taux de charge |
| **65** | **Column 64** | Vide | Vide |

Le choix des variables les plus pertinentes à notre analyse a été fait en fonction des besoins de chaque objectif :

- Pour l'analyse de la consommation et la production énergétique (par filière) au niveau national et régional, notre choix s'est porté sur les variables : **Régions**, **Date**, **Heure**, **Consommation** et **productions énergétiques**.

- Pour la partie modélisation par machine Learning, notre choix s'est naturellement porté sur la variable « **Consommation (MW)** » qui représente notre variable cible, ainsi que les variables : **Régions**, **Date** et **Heure**.

## I.3 – Analyse des valeurs manquantes

Le tableau ci-dessous résume le nombre de valeurs manquantes et leurs proportions pour chaque variable du data set :

| VARIABLES | NOMBRE DE NaN | PROPORTION DES NaN |
|---|---|---|
| Consommation (MW) | 12 | < 0.1% |
| Thermique (MW) | 12 | < 0.1% |
| Nucléaire (MW) | 693.847 | 42 % |
| Eolien (MW) | 108 | < 1% |
| Solaire (MW) | 12 | < 0.1% |
| Hydraulique (MW) | 12 | < 0.1% |
| Pompage (MW) | 728.887 | 44 % |
| Bioénergies (MW) | 12 | < 0.1% |
| Ech. physiques (MW) | 12 | < 0.1% |
| Flux physiques entre régions | 1.633.056 | 98 % |
| TCO et TCH de chaque type d'énergie | 1.472.256 | 88 % |
| Column 64 | 1.665.216 | 100 % |

Effectivement, plusieurs variables montrent un nombre assez élevé de valeurs manquantes. Afin de mieux visualiser ces valeurs manquantes, la figure ci-dessous réalisé avec une heatmap, nous permet d'avoir la répartition des valeurs manquantes pour chaque variable :

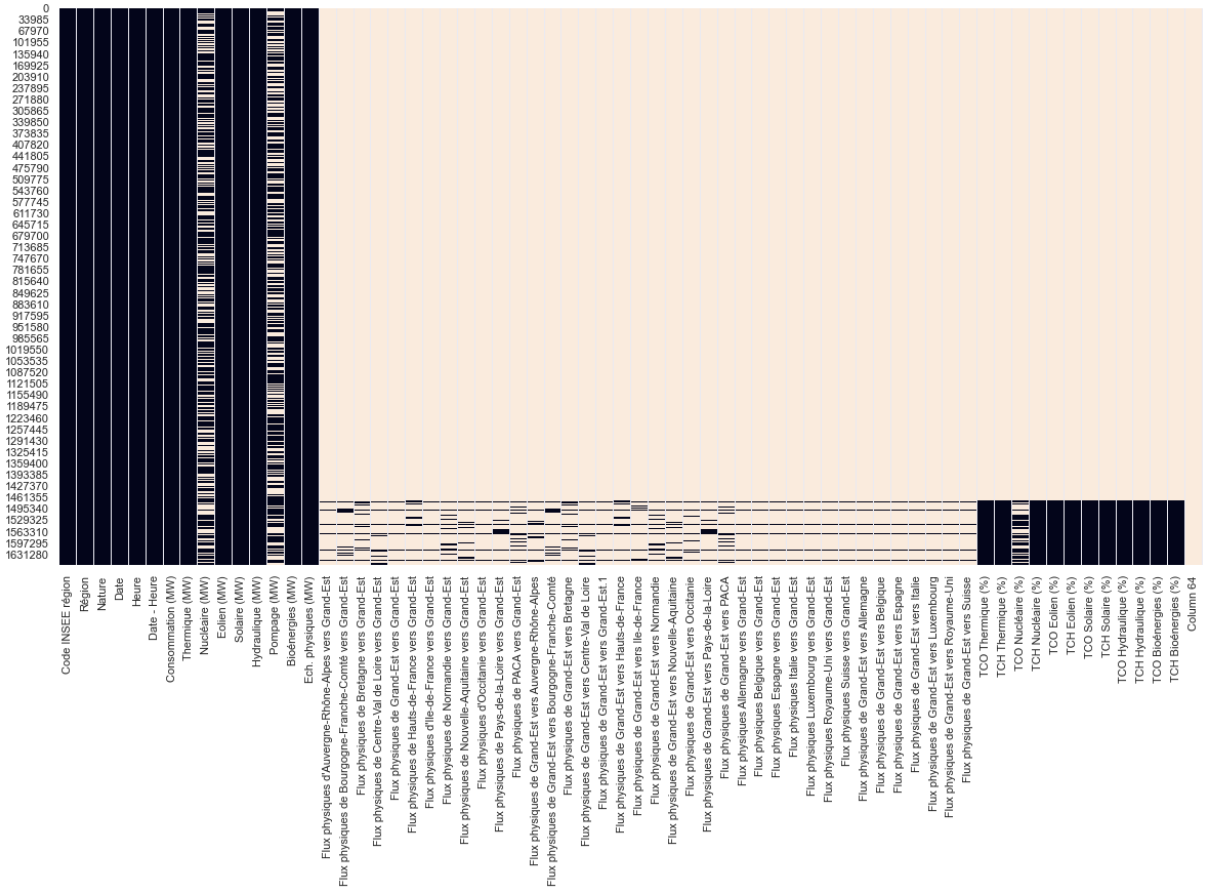



*Figure 1 - Analyse visuelle des valeurs manquantes*

Cette visualisation nous permet de constater :

- Les variables de Flux physiques entre régions et celles de TCO et TCH, ne sont disponibles qu'à partir d'une certaine date.
- La variable « Column 64 » elle ne représente que des valeurs manquantes.
- Les valeurs manquantes des variables « Nucléaire (MW) » et « Pompage (MW) » sont réparties tout au long du data set.

Une analyse approfondie des valeurs manquantes a été effectué sur chaque variable, et il a été constaté :

- Les 12 valeurs manquantes de la variable cible « Consommation (MW) » représente les 12 premières lignes du Data Set correspondant à l'enregistrement de minuit du 1er Janvier 2013. Il s'agit vraisemblablement d'un problème d'acquisition de données.
- Les valeurs manquantes des variables 'Nucléaire (MW)' et 'Pompage (MW)' ne sont présentes que dans les régions qui ne produisent pas ce type d'énergie.
- Les variables de Flux physiques entre régions et celles de TCO et TCH, ne sont disponibles que pour l'année 2020.

## I.4 – Preprocessing des données

Le prétraitement des données a été réalisé en deux phases :

- La gestion des valeurs manquantes
- La réorganisation du Data Set

### I.4.a – Gestion des valeurs manquantes

Le traitement des valeurs manquantes a été réalisé en plusieurs étapes :

- Suppression de la variable « Column 64 »
- Suppression des 12 lignes manquantes de la variable cible « Consommation (MW) »
- Remplacement des valeurs manquantes des variables de productions énergétiques par des 0
- Suppression des variables de Flux entre régions et celles de TCO et TCH du Data Set principal

### I.4.b – Réorganisation du Data Set

La réorganisation du Data Set a été réalisée comme suit :

- Remplacement du nom de la variable « Nature » par « Données définitives » et remplacement de ses deux modalités par 0 et 1. -> Données consolidées = 0 -> Données définitives = 1.
- Dissociation de la variable « Date » en variables : « **Année** », « **Mois** », « **Jour** » et « **Jour_semaine** ».
- Suppression de la colonne « Date-Heure ».
- Création d'une variable « **Prod_renouvelables_MW** » qui représente la somme des variables de productions d'énergie renouvelable, à savoir : « Eoliens », « Solaire », « Hydraulique » et « Bioénergies ».
- Création d'une nouvelle variable « **Prod_totale_MW** » qui représente la somme de toutes les variables de productions.
- Ré-identification des variables « **Thermique** », « **Nucléaire** », « Pompage » en « **Prod_fossiles** », « **Prod_nucleaire** », et « **Prod_step** ».
- Ré-identification de l'ensemble des variables en éliminant les accents, espaces et caractères spéciaux.

### I.4.c – Corrélation entre les variables

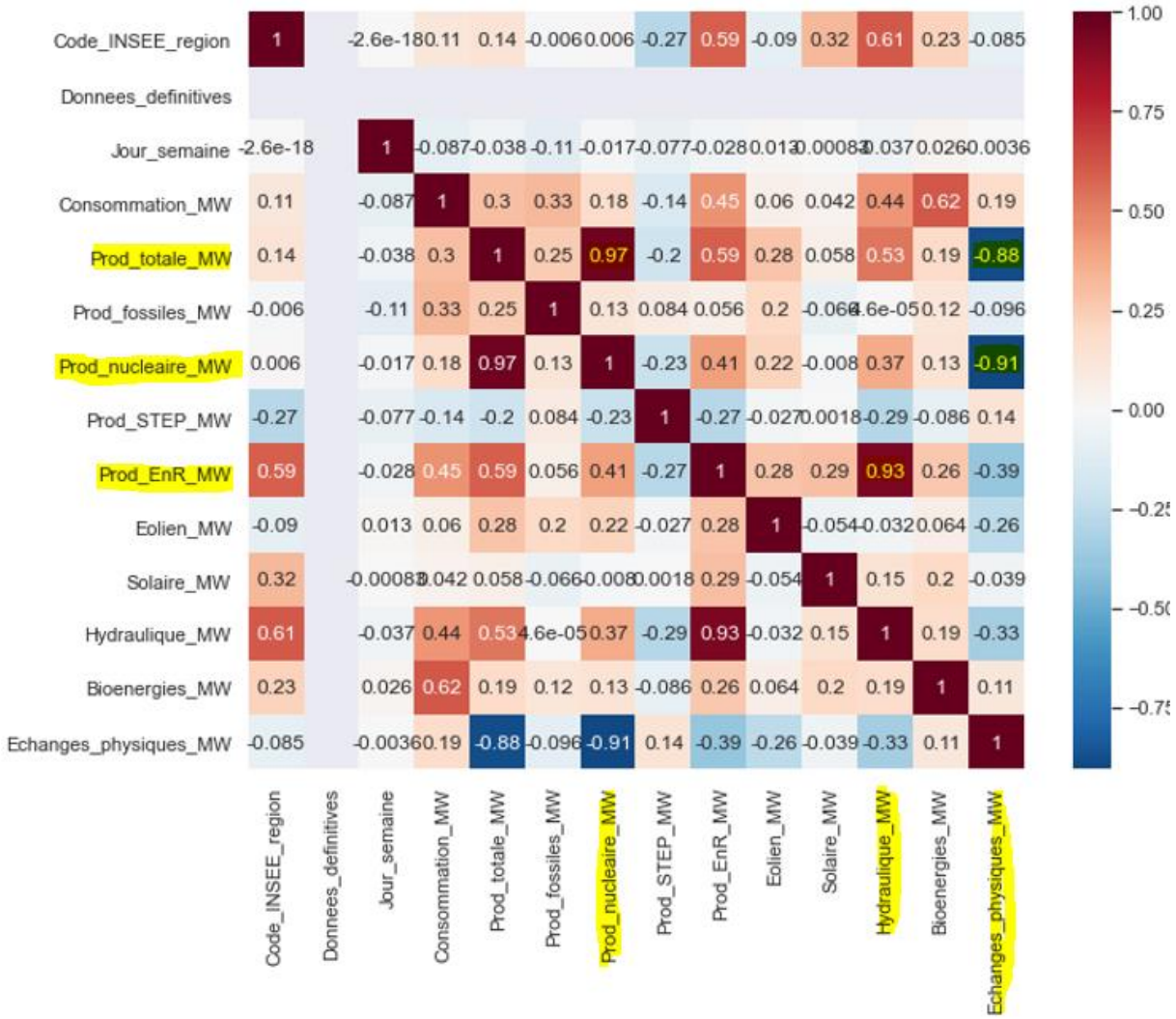



*Figure 2 - Table de corrélation entre les variables*

La Figure 4 montre la corrélation entre les variables du dataset.

Elle met en évidence 4 corrélations :

- Corrélation positive de 97% entre « Prod_nucleaire_MW » et « Prod_totale_MW »
- Corrélation positive de 93% entre « Prod_hydraulique_MW » et « Prod_EnR_MW »
- Corrélation négative de 91% entre « Prod_nucleaire_MW » et « Échanges_physiques_MW »
- Corrélation négative de 88% entre « Prod_totale_MW » et « Échanges_physiques_MW »

# *Partie II :*
# *Analyses statistiques et Data Visualisation*

*Dans cette partie nous détaillerons les différentes analyses statistiques et les visualisations sur la consommation et la production électrique au niveau national et régional.*

## II.1 – Analyse de la distribution des variables

Pour cette analyse visuelle, nous avons divisé notre data set principal en plusieurs data sets régionaux, afin d'éviter un surplus de valeurs aberrantes sur la distribution des variables.

Afin d'analyser la distribution des variables de notre data set, nous avons opté pour une visualisation des variables dans un graphique en boites à moustaches pour chaque région.

Les figures ci-dessous représentent la distribution des variables pour les régions « Ile-de-France » et « Pays de la Loire »

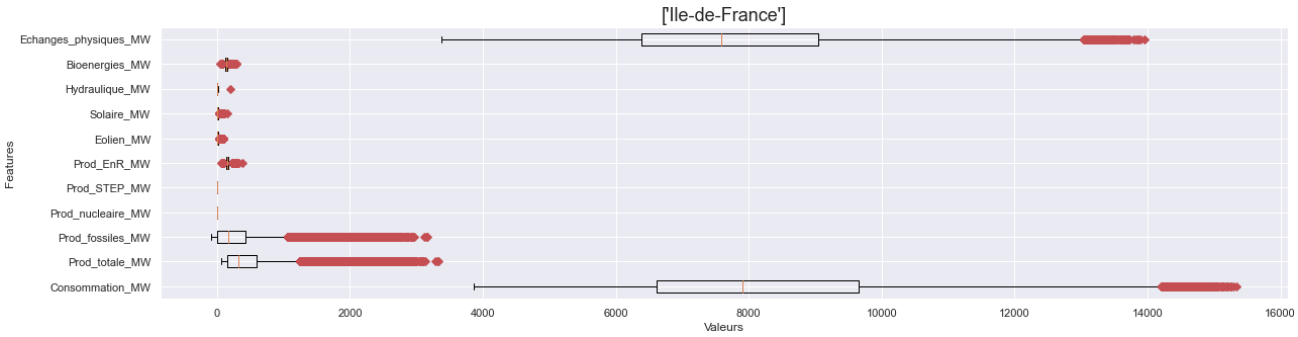



*Figure 3 - Distribution des variables de la région Ile-de-France*

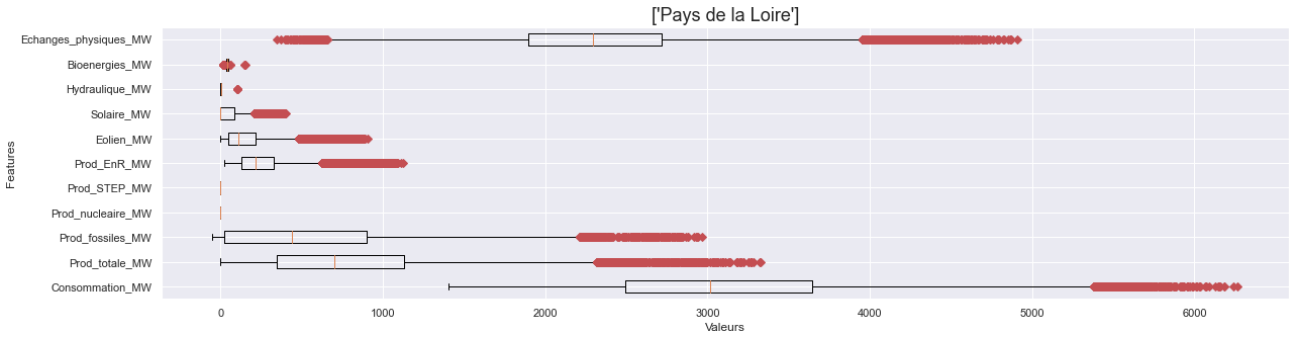
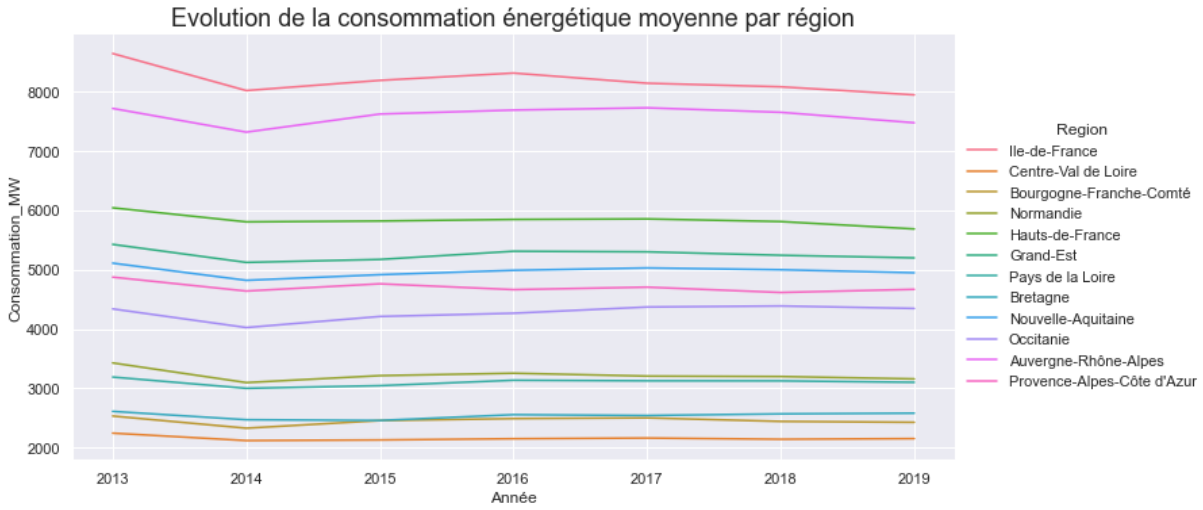



*Figure 4 - Distribution des variables de la région Pays de la Loire*

L'analyse visuelle de la distribution des variables indique la présence de beaucoup de valeurs aberrantes sur les variables de notre data set. Ceci est dû au fait que chacune des variables varie en fonction des saisons des années, avec une évolution annuelle.

## II.2 – Analyse de la consommation énergétique

### II.2.a – Evolution de la consommation énergétique moyenne par région




Le graphique de l'évolution de la consommation énergétique moyenne par région nous permet de constater :

- La consommation énergétique est différente d'une région à une autre, ce qui s'explique par la différence de démographie et nombre de foyer dans chaque région, mais aussi du nombre d'industries et d'entreprises implantés.

- Une chute de la consommation moyenne annuelle dans toutes les régions en 2014, ceci s'explique par la montée de la température annuelle nationale de 1.2°C [3].

- Un retour en 2015 à un niveau de consommation moyen similaire à celui de 2013

- Une certaine stabilité de consommation entre 2015 et 2018

- Une chute de la consommation moyenne dans certaines régions

[3] Article du Parisien du 5 Janvier 2014: Chaleur : 2014, année record en France et en Europe depuis 1900

### III.2.b – Répartition mensuelle de la consommation énergétique nationale

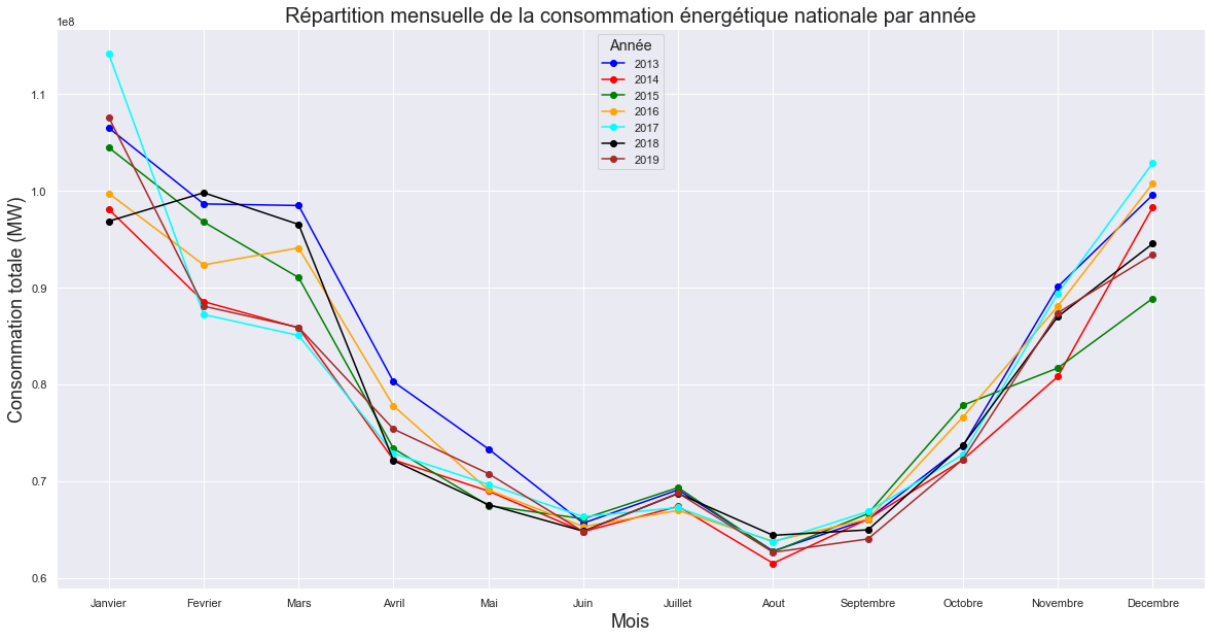



*Figure 5 - Répartition mensuelle de consommation électrique de 2013 à 2019*

Le graphique de répartition mensuelle de la consommation énergétique nationale nous permet de constater que :

- La consommation énergétique varie en fonction des saisons.
- Le maximum de consommation est atteint pendant l'hiver
- Le minimum de consommation est atteint pendant le mois d'Aout
- Une certaine symétrie entre le printemps et l'automne.

Ce graphique est très riche en information car il montre la répétitivité et la saisonnalité de la variable cible **Consommation** en fonction des mois de l'année, et donc permet d'affirmer une importante corrélation entre ces deux variables.

### III.2.c – Analyse de la consommation électrique en 2020

L'année 2020 a été marquée par la crise sanitaire liée au Covid-19, la figure ci-dessous représente la consommation électrique nationale de l'année 2020 en comparaison avec celle de 2019, avec mise en évidence des deux périodes de confinement décrété par le gouvernement en Mars et Octobre 2020.

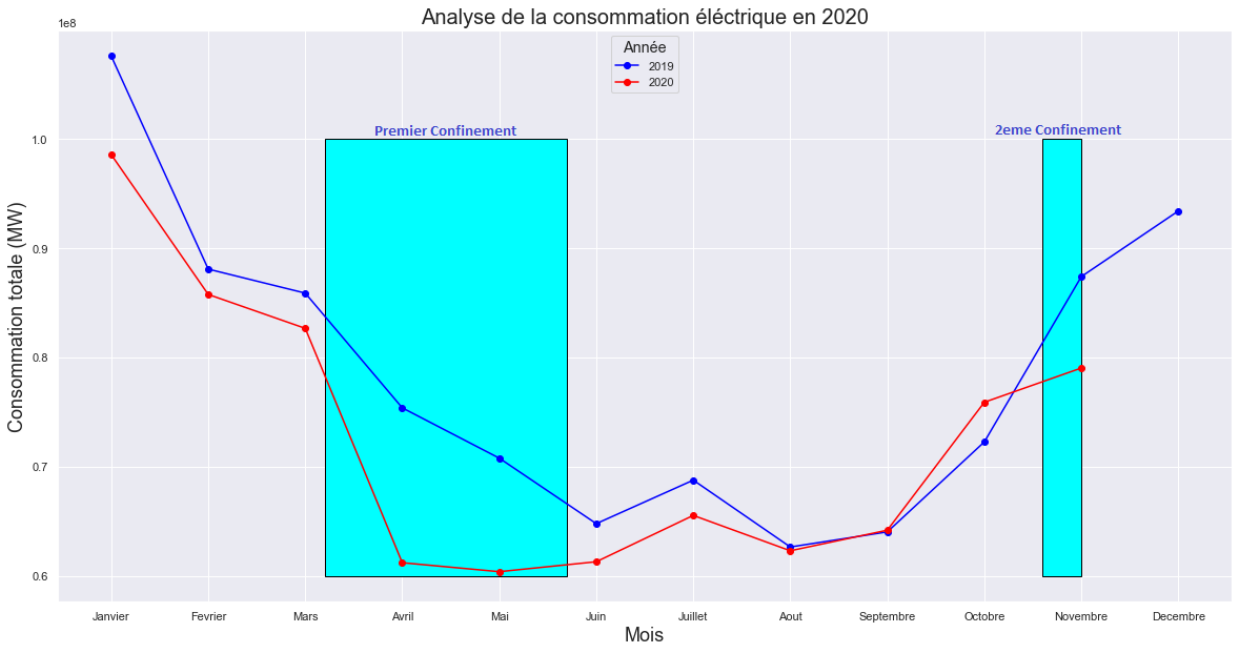



*Figure 6 - Analyse de la consommation électrique en 2020*

A partir de cette figure, nous constatons que la crise sanitaire liée au virus COVID-19, a eu un impact majeur et immédiat sur la consommation d'électricité en France, dès que les mesures de confinement ont été adoptées.

Ainsi, dès les premiers jours de confinement, une baisse importante de la consommation a été enregistrée. Au plus fort de la crise (deuxième et troisième semaines de confinement), les mesures de confinement ont pu entraîner un impact sur la consommation d'électricité supérieur à 15 % [4].

Cet impact s'est par la suite réduit sur les semaines suivantes, du fait d'une reprise partielle de l'activité, notamment dans le secteur industriel.

[4] https://assets.rte-france.com/prod/public/2020-11/Rapport_hiver%202020-2021_novembre%202020%20DEF_0.pdf

## II.3 – Analyse de la production énergétique

### II.3.1 – Production annuelle d'électricité au niveau national

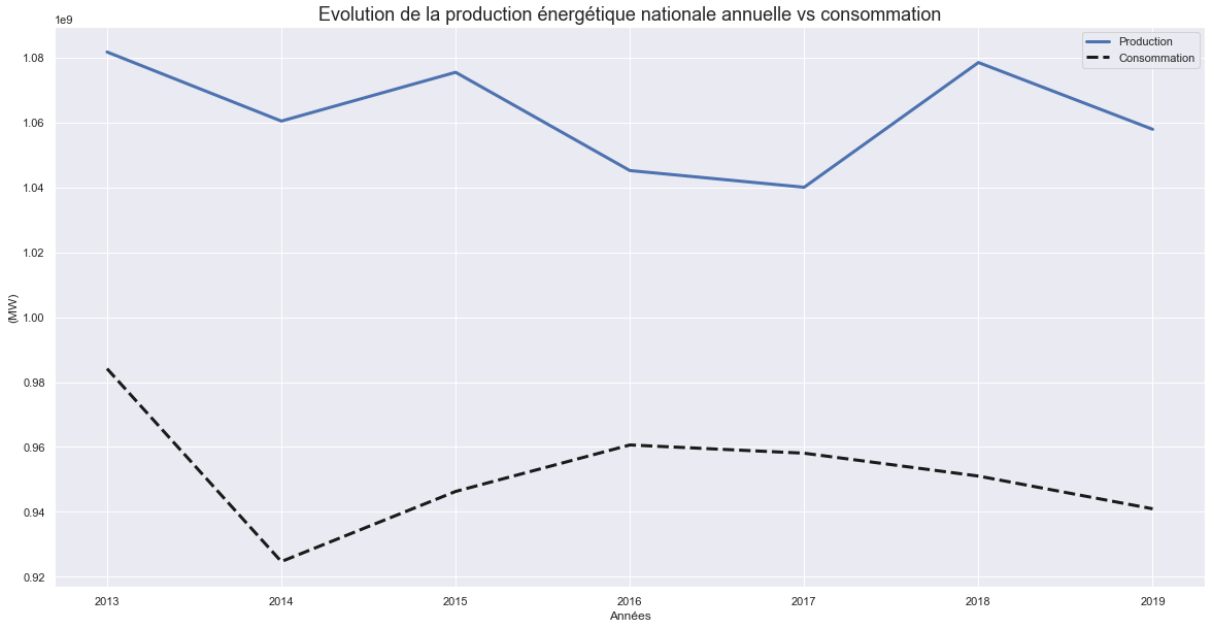



*Figure 7 - Evolution de la production énergétique nationale*

Le graphique ci-dessus nous permet de constater que la production énergétique est dans une certaine mesure proportionnelle à la consommation énergétique, avec un certain phasage de sécurité, afin d'éviter les risques de Blackout.

Il est également à noter, qu'une partie de l'électricité produite est destinée à l'exportation vers les pays limitrophes. En 2019, la France a exporté 57.7 TW vers l'Europe, ce qui lui vaut d'être le pays le plus exportateur d'Europe [5].

Cette production annuelle représente un mix des différentes filières de production. Par exemple pour l'année 2019 la production est répartie comme suit :

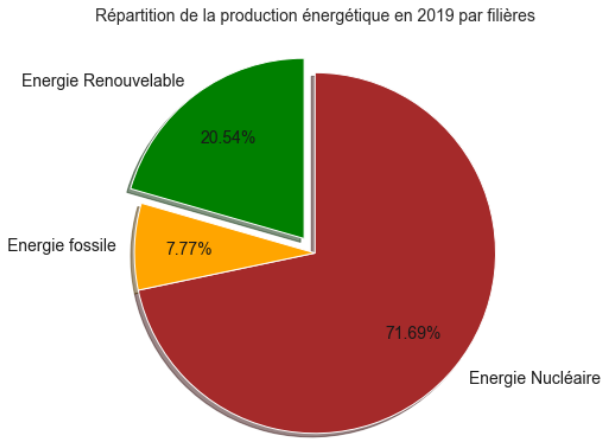



*Figure 8 - Répartition de la production énergétique en 2019*

[5] https://bilan-electrique-2019.rte-france.com/prix-echanges-solde-france-echanges/

Nous remarquons que la production nucléaire représente plus 71% de la production totale (comme représenté dans la matrice de corrélation dans la figure 2), suivie par les énergies renouvelables avec 20% et enfin les énergies fossiles avec moins de 8%.

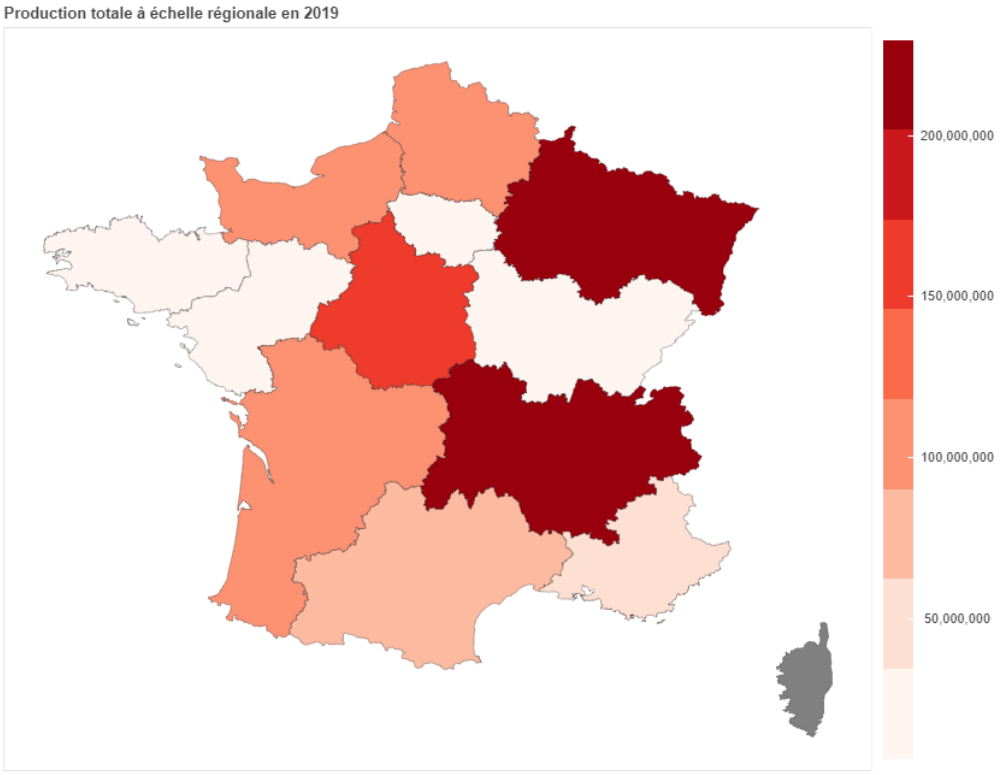



*Figure 9 - Production énergétique en région pour l'année 2019*

La figure 9 montre que la région Auvergne-Rhône-Alpe arrive en tête des productions pour l'année 2019 avec plus de 130TW, suivie de la région Grand-est avec plus 91 TW.

La production d'électricité ramenée à la maille régionale permet non seulement de couvrir les besoins de la région productrice mais contribue également à la couverture de la demande émanant de régions limitrophes et au-delà [6]. Le réseau de transport d'électricité assure la solidarité interrégionale à deux niveaux :

- D'abord d'un point de vue géographique : les régions Centre-Val de Loire ou Grand-Est qui produisent beaucoup plus qu'elles ne consomment contribuent fortement à cette solidarité. De cette façon les régions dépendant fortement de l'électricité produite dans les régions limitrophes telles que l'Île-de-France, la Bourgogne Franche-Comté ou la Bretagne ont l'assurance de pouvoir couvrir leurs besoins de consommation.

- Ensuite d'un point de vue temporel : chaque région est amenée à recourir à des productions en dehors de son territoire pour couvrir ponctuellement ses besoins.

### II.3.2 – Production annuelle des énergies renouvelables au niveau national

Afin de pouvoir évaluer l'évolution de la production annuelle des énergies renouvelables, la figure ci-dessous représente le taux annuel de la production des énergies renouvelables par rapport à la production totale :

[6] https://bilan-electrique-2019.rte-france.com/territoires-et-regions-equilibre-entre-production-et-consommation/

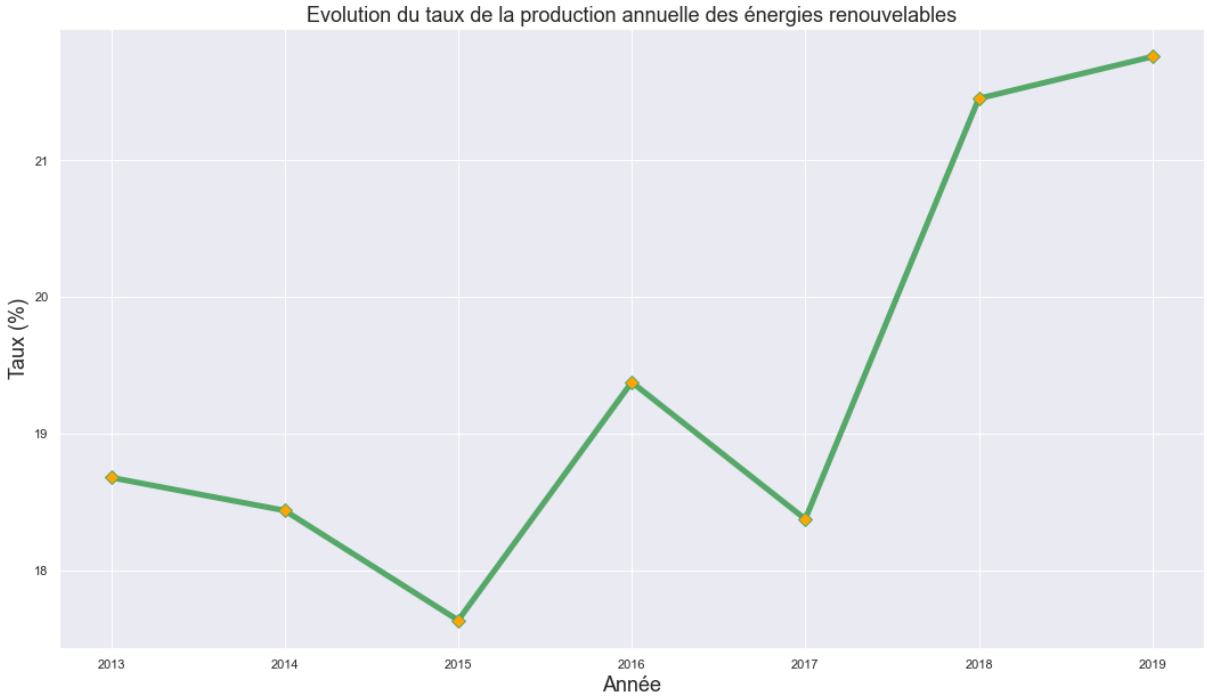



*Figure 10 - Evolution du taux de production totale des énergies renouvelables*

Ce graphique révèle une hausse de production de plus de 15% entre 2013 et 2019 avec de fortes chutes de production en 2015 et en 2017.

Pour expliquer cette tendance, nous allons nous intéresser à l'évolution de la production de chaque type d'énergie renouvelable :

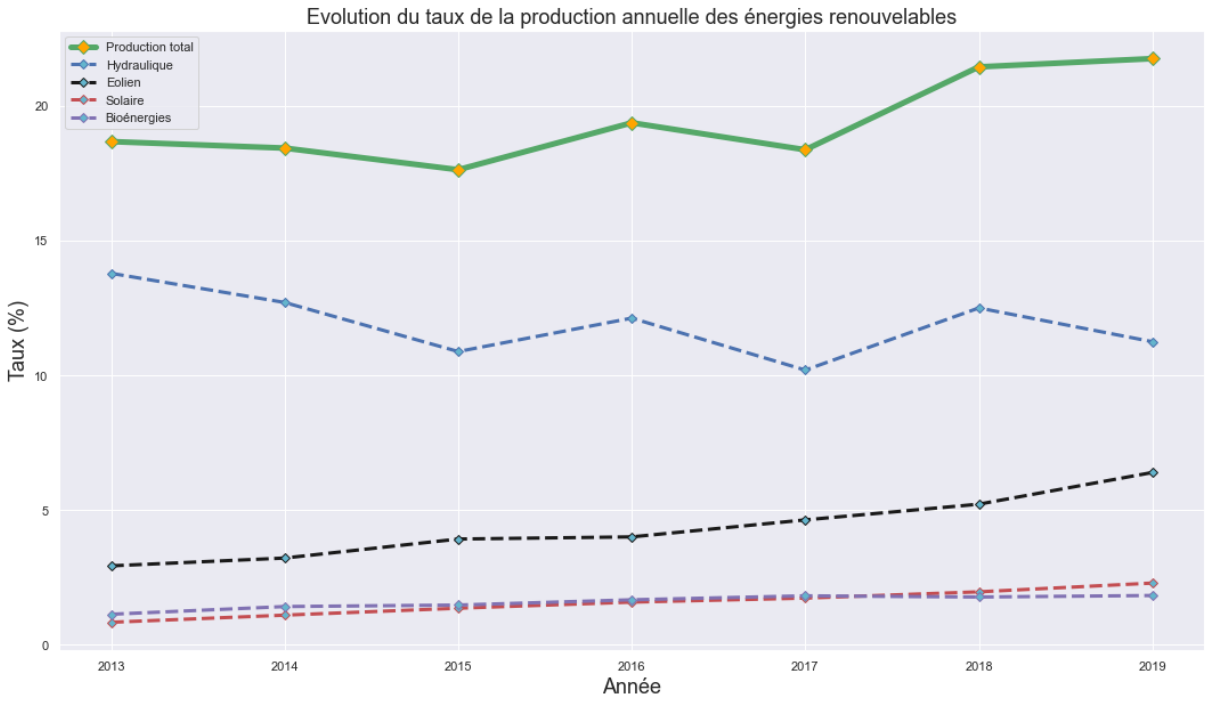



*Figure 11 - Evolution du taux de production par type d'énergie renouvelable*

Le graphique ci-dessus démontre la forte influence de la production hydroélectrique sur la production totale, car elle représente plus de 70% (comme représenté dans la matrice de corrélation dans la figure

2) de la production totale en 2013 et plus de 50% en 2019, et donc les chutes de production totale de 2015 et 2017 sont directement liées à la chute de la production hydroélectrique.

En effet, en moyenne et sur l'année 2015, la pluviométrie a été inférieure à la normale de plus de 15%[7] ce qui a impacté la production hydraulique d'un déficit de production de plus de 11%.

En 2017, le cumul de précipitations a été déficitaire de plus de 10%, plaçant 2017 parmi les années les plus sèches sur la période 1959-2017 [8], ce qui a contraint la production hydroélectrique à une baisse de 16,3%

Inversion de tendance pour l'année 2018, le cumul des précipitations a été légèrement excédentaire en moyenne sur l'année [9], ce qui a engendré un bond de production de 27.5% par rapport à 2017.

La baisse de production hydraulique en 2019 par rapport à 2018 a ainsi été en partie compensée par une augmentation des productions éolienne et solaire, portée à la fois par des conditions météorologiques plus favorables et par un parc qui continue de croître [10].

Le graphique révèle aussi une augmentation sur la période 2013-2019 de la production éolienne de plus de 87%, de l'énergie solaire de près de 162% et des bioénergies de près de 50%

### II.3.3 – Production régionale des énergies renouvelables pour l'année 2019

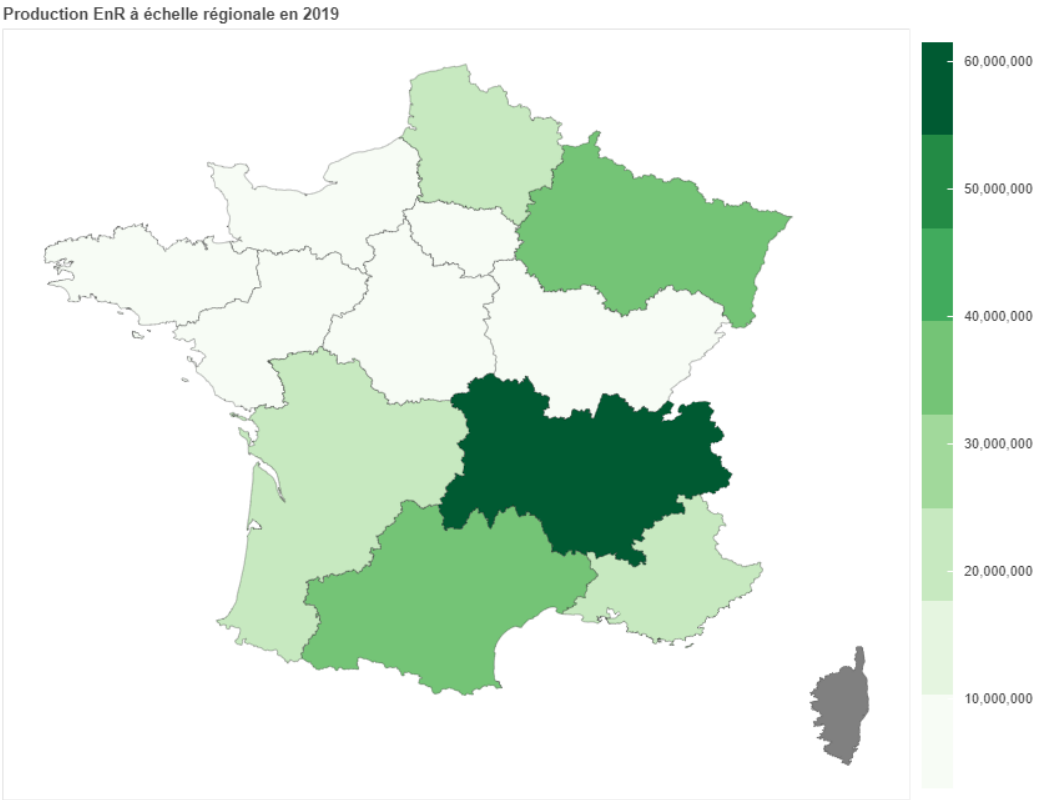



*Figure 12 - Production des énergies renouvelables en région pour l'année 2019*

7 http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques/bilan-2015/bilan-climatique-de-l-annee-2015
8 http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques/bilan-2017/bilan-climatique-de-l-annee-2017
9 http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques/bilan-2018/bilan-climatique-de-l-annee-2018
10 https://bilan-electrique-2019.rte-france.com/production-renouvelable/

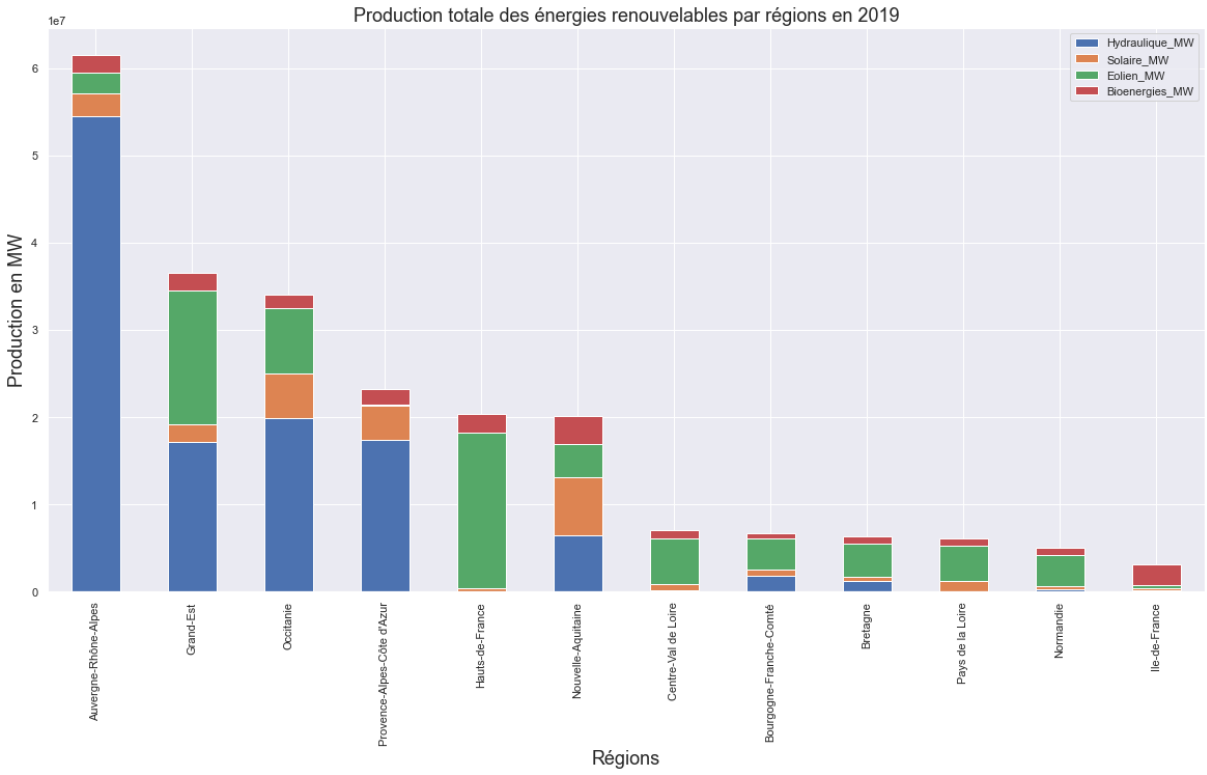



*Figure 13 - Classement des régions par production des énergies renouvelables*

On observe sur ces deux graphiques que la région Auvergne-Rhône-Alpes arrive en tête du classement avec une production totale de plus de 60 TW d'énergies renouvelables en 2019.

En regardant de plus près sur la production énergétique en région Auvergne-Rhône-Alpes (figure 14), nous constatons que la production d'Energie renouvelable ne représente 17% de la production totale de la région, contre 80% pour l'énergie Nucléaire et de 2% pour les énergies fossiles.

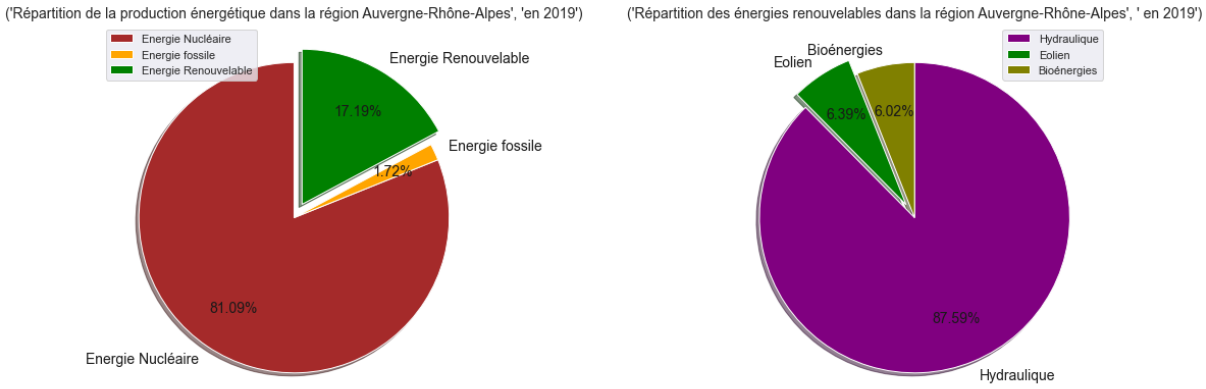



*Figure 14 - Production énergétique en région Auvergne-Rhône-Alpes*

On retrouve également une diversité dans le type de production d'énergie renouvelable avec plus 87% de production d'énergie hydroélectrique, et un peu plus de 6% pour les énergies éoliennes et les bioénergies.

# *Partie III : Modélisation*

*Dans cette partie, nous présenterons différents modèles de Machine Learning que nous avons implémenté et les différents résultats obtenus, et ainsi nous expliquerons notre choix sur le modèle retenu.*

## Introduction

Compte tenu du type de données, les premiers modèles expérimentés ont été ceux faisant partie de la famille des « régressions ». Quatre modèles ont été implémentés :

- Régression « SGD »
- Régression « RIDGE »
- Régression « LASSO »
- Régression « ELASTIC NET CV »

Ensuite, deux modèles « Séries Temporelles » ont été testées :

- Modèle SARIMAX
- Modèle ARIMA

Enfin, un modèle de réseaux de neurones a été mis en place en dernière étape.

Afin de respecter une cohérence sur les modèles et les spécificités propres aux différentes régions, les modélisations ont été réalisées par région.

**Note : Seule la région « Pays de la Loire » (Code INSEE = 52) sera présentée dans ce rapport.**

Les dates utilisées pour la modélisation ont été réparties selon le Tableau 3 :

| | Date début | Date fin |
|---|---|---|
| **Entrainement des modèles** | 01/01/2013 | 31/12/2018 |
| **Prédictions** | 01/01/2019 | 30/11/2020 (fin du dataset) |

*Tableau 1 - Dates utilisées pour l'entrainement et pour les données à prédire*

Le jeu de données a été étudié en entrainant les modèles avec les données regroupées par mois et par semaine.

## III.1 – Préparation du jeu de données

Afin d'implémenter un modèle de Machine Learning, une étape de modification du jeu de données a été nécessaire afin qu'il soit correctement entrainé par la suite par les différents modèles. Les différentes modifications apportées lors de cette étape sont listées ci-dessous :

- **Regroupement des heures en créneaux de 3h**

- **Opérations de dichotomie sur les variables suivantes** :
  - Créneaux horaires
  - Jour de la semaine ('Jour_semaine'), Jour du mois ('Jour'), Mois ('Mois'), Année ('Annee')
  - Région ('Region_*')

- **Suppression des variables non utiles pour la modélisation** :
  - Heure ('Heure')
  - Variables des autres régions + Code INSEE des régions ('Code_Insee_region')
  - Variables de production + Echanges physiques ('Prod_totale_MW', 'Prod_fossiles_MW', 'Prod_nucleaire_MW', 'Prod_STEP_MW', 'Prod_EnR_MW', 'Eolien_MW', 'Solaire_MW', 'Hydraulique_MW', 'Bioenergies_MW', 'Echanges_physiques_MW')

## III.2 – Modèles de régression

### III.2.a – Premières observations

Les Figures 15 et 16 représentent les données utilisées pour entrainer le modèle, les données à prédire ainsi que les prédictions regroupées respectivement par semaine et par mois.

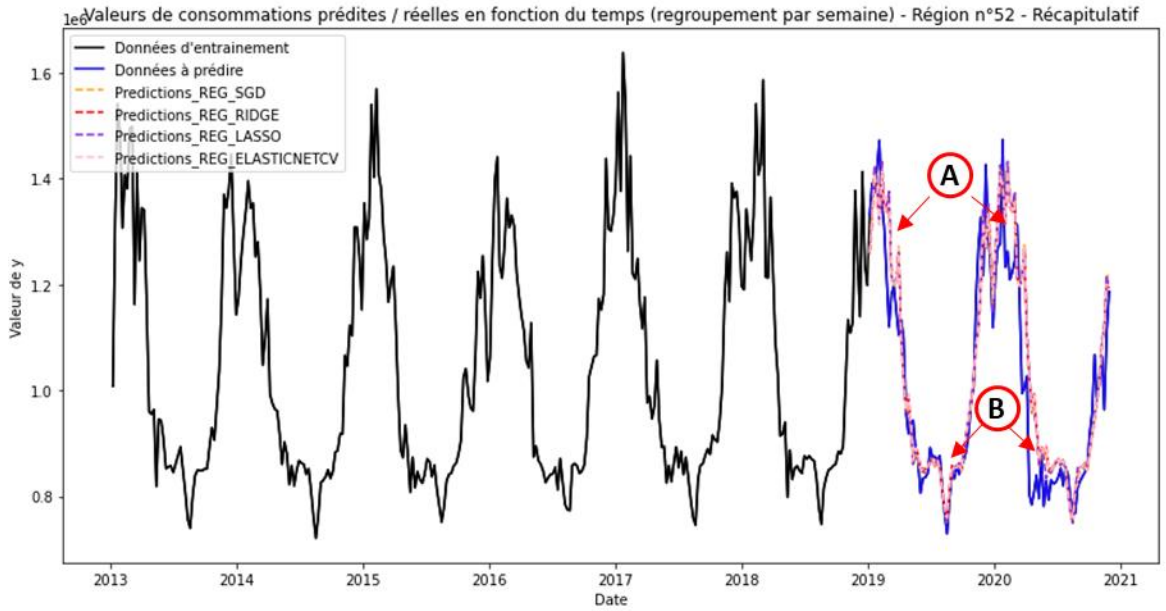



*Figure 15 - Données d'entrainement (2013 à 2018), données à prédire (2019 à 2020) et prédictions des 4 modèles de régression (SGD, RIDGE, LASSO, ELASTIC NET CV) regroupées par semaine*

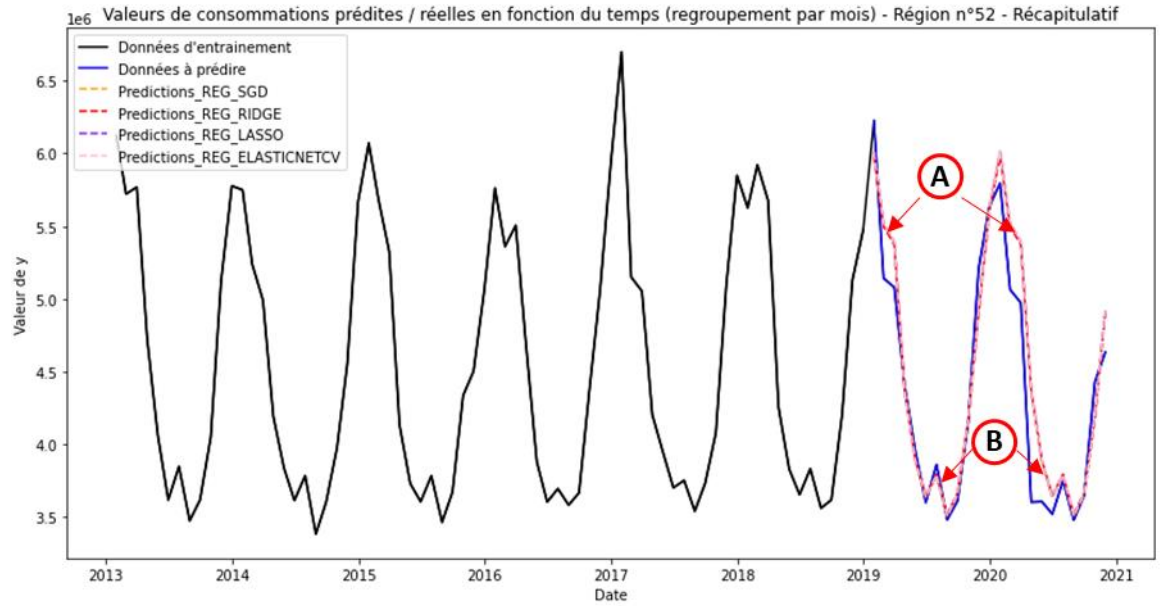



*Figure 16 : Données d'entrainement (2013 à 2018), données à prédire (2019 à 2020) et prédictions des 4 modèles de régression (SGD, RIDGE, LASSO, ELASTIC NET CV) regroupées par mois*

La représentation mensuelle permet un lissage des données et met en évidence deux spécificités des années 2019 et 2020 concernant les prédictions.

Tout d'abord au niveau du repère « A » où l'écart avec les prédictions est le plus important. Ensuite au niveau du repère « B » une différence est plus importante sur l'année 2020.

Afin d'expliquer ces 2 observations, les températures moyennes de cette région ont été intégrées.

La Figure 17 représente les données d'entrainement et à prédire avec la modélisation « Régression Ridge » ainsi que les températures moyennes mensuelles[11].

Cette figure met en évidence qu'une même observation pourrait être faite pour l'année 2017 où l'on observe le même comportement pour le mois de mars également.

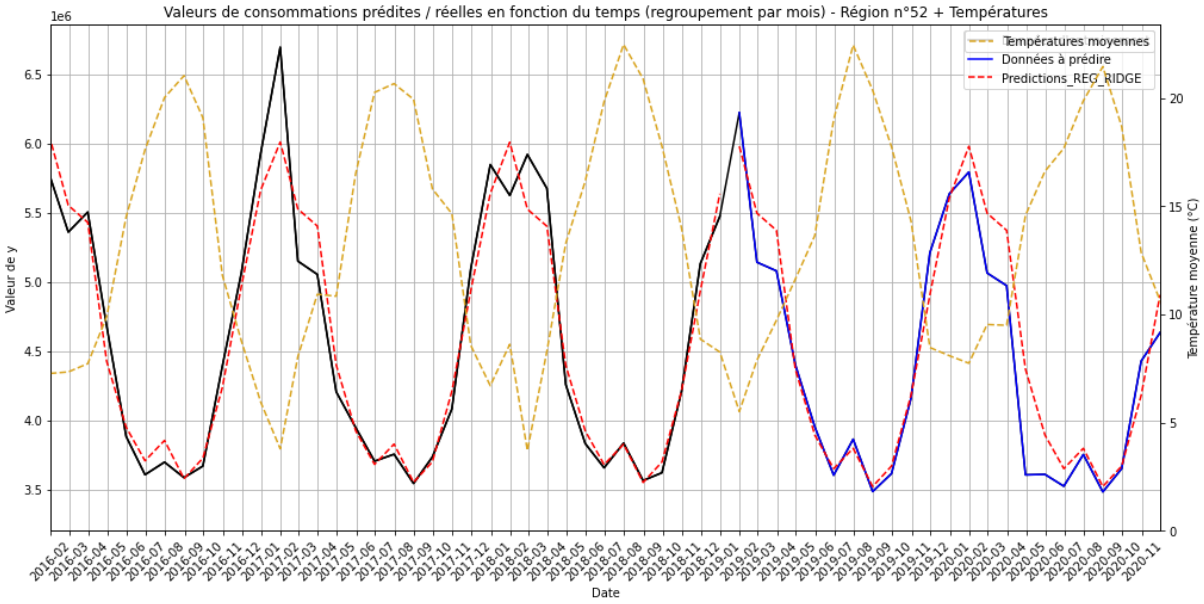



*Figure 17 - Données d'entrainement (2013 à 2018), données à prédire (2019 à 2020) et prédictions du modèle RIDGE pour l'ensemble des données regroupées par mois + Températures moyennes*

Une première explication pourrait provenir de températures plus clémentes. La Figure 18, qui montre les températures moyennes des mois de Mars de 2016 à 2020, va en ce sens en indiquant que ces 3 années possèdent des mois de Mars les plus chauds (1 à 2°C supplémentaires par rapport aux autres années).

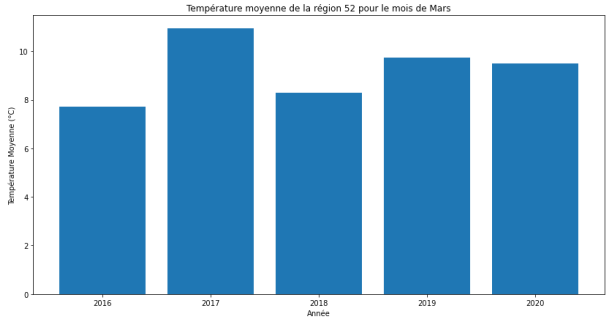



*Figure 18 - Températures moyennes du mois de Mars de la région 52 de 2016 à 2020*

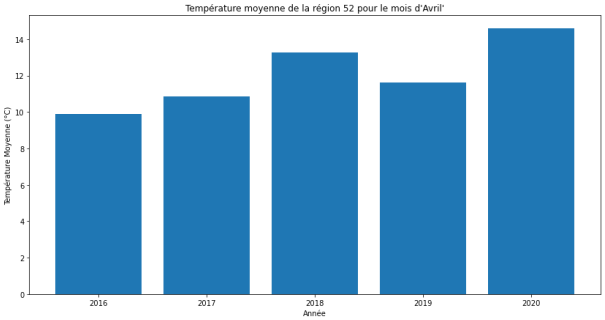



*Figure 19 - Températures moyennes du mois d'Avril de la région 52 de 2016 à 2020*

Concernant l'observation du repère « B », les Figures 15 et 16 montrent clairement une importante différence de consommation entre les mois de Mars et Avril.

Encore une fois, les températures clémentes peuvent expliquer en partie ce comportement compte tenu du fait que le mois d'Avril 2020 a montré une température moyenne plus élevée que les 4 dernières années sur ce même mois (cf. Figure 19).

[11] https://opendata.reseaux-energies.fr/explore/dataset/temperature-quotidienne-regionale/information/?disjunctive.region

Ces 2 mois semblent être particuliers et un effet additionnel de la situation sanitaire due au COVID 19 (activités au ralenti, confinement) peut également avoir eu une influence non négligeable.

Dans son rapport « Hiver 2020-2021 », le gestionnaire du réseau de transport d'électricité (RTE) attribut explicitement ce comportement à l'effet du 1er confinement de la crise COVID 19[12].

### III.2.b – Comparaisons des modèles de régressions (à échelle mensuelle)

La Figure 20 présente les valeurs des années 2019 à 2020 à prédire ainsi que les différentes prédictions. Elle montre que les 4 prédictions sont quasi-identiques. C'est également ce que démontrent les Figures 21 et 22 qui affichent les valeurs des scores obtenues, respectivement RMSE (Root Mean Squared Error) et R2 (coefficient de détermination).

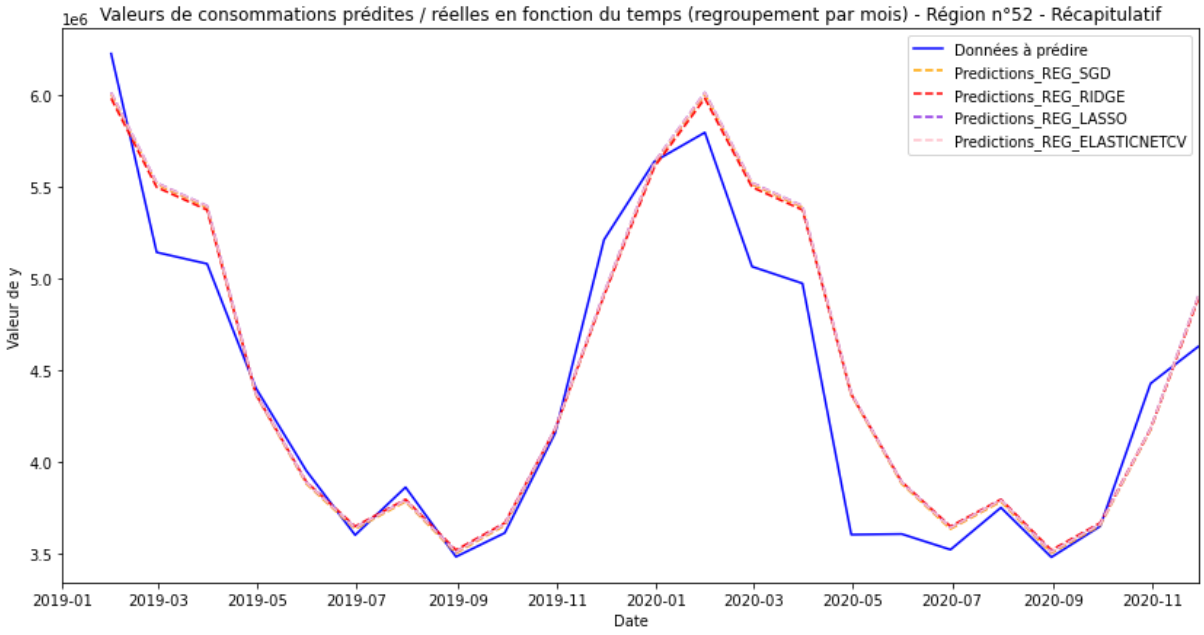



*Figure 20 - Valeur de consommation à prédire (2019 à 2020) et prédictions des 4 modèles de régressions (regroupement mensuel)*

L'écart de valeurs entre les scores RMSE des données d'entrainement et ceux des valeurs prédites sont d'un même ordre de grandeur pour l'ensemble des modèles. Le score RMSE des données d'entrainement est plus faible (donc résultat plus précis) que celui des prédictions : la différence est d'environ 70 000 pour les 4 modèles.

[12] https://assets.rte-france.com/prod/public/2020-11/Rapport_hiver%202020-2021_novembre%202020%20DEF_0.pdf

Le modèle de régression « Ridge » obtient une valeur de RMSE légèrement inférieure aux 3 autres modèles.

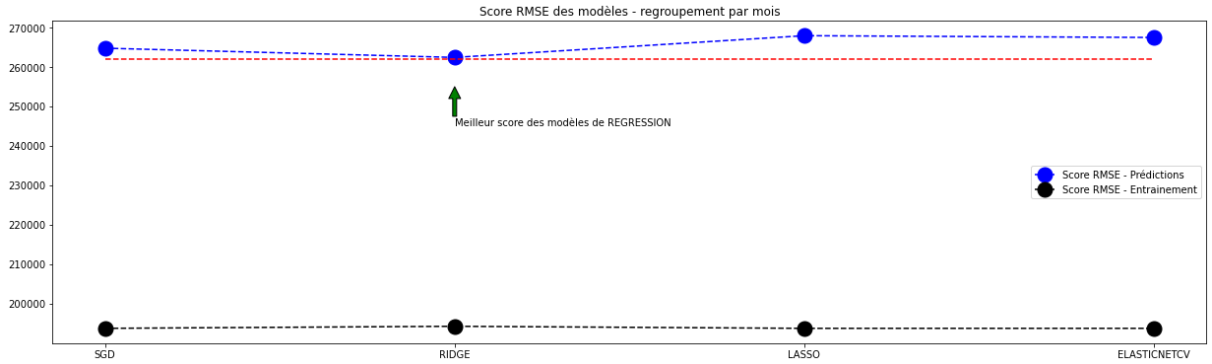



*Figure 21 - Score RMSE des valeurs d'entrainement et prédites*

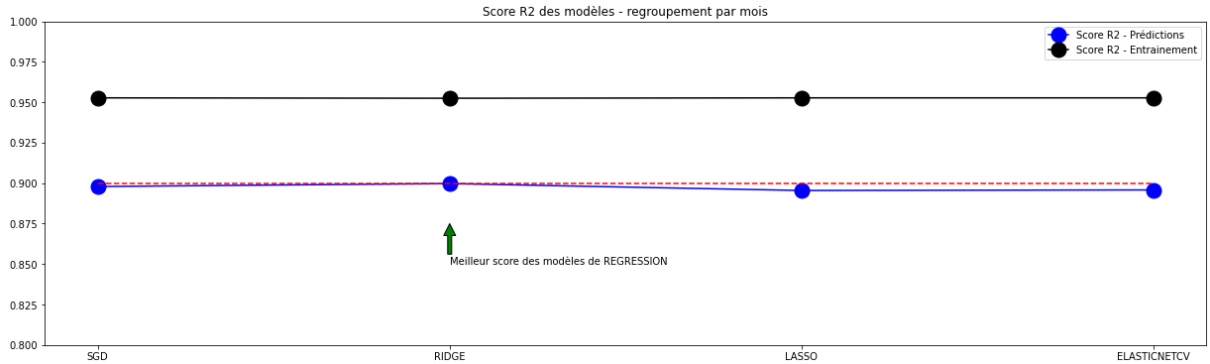



*Figure 22 - Valeur des coefficients de détermination R2 pour chaque modèle de régressions et pour les données d'entrainement et les prédictions (regroupement mensuel)*

La Figure 22 confirme les mêmes observations que la Figure 21 :

- La valeur des coefficients de détermination R2 est plus élevée pour les données d'entrainement que pour les prédictions : le modèle décrit 95% des variations pour les données utilisées pour l'entrainement contre 89% pour les données prédites.

- Le modèle de régression « Ridge » possède un coefficient de détermination R2 légèrement plus élevé que les 3 autres modèles

### III.2.c – Comparaisons des modèles de régressions (à échelle hebdomadaire)

La Figure 23 montre un zoom sur 2019 avec les informations de température associées et la Figure 24 montre celui de 2020.

A cette échelle aussi, les modèles semblent présenter des résultats suivants la même tendance.

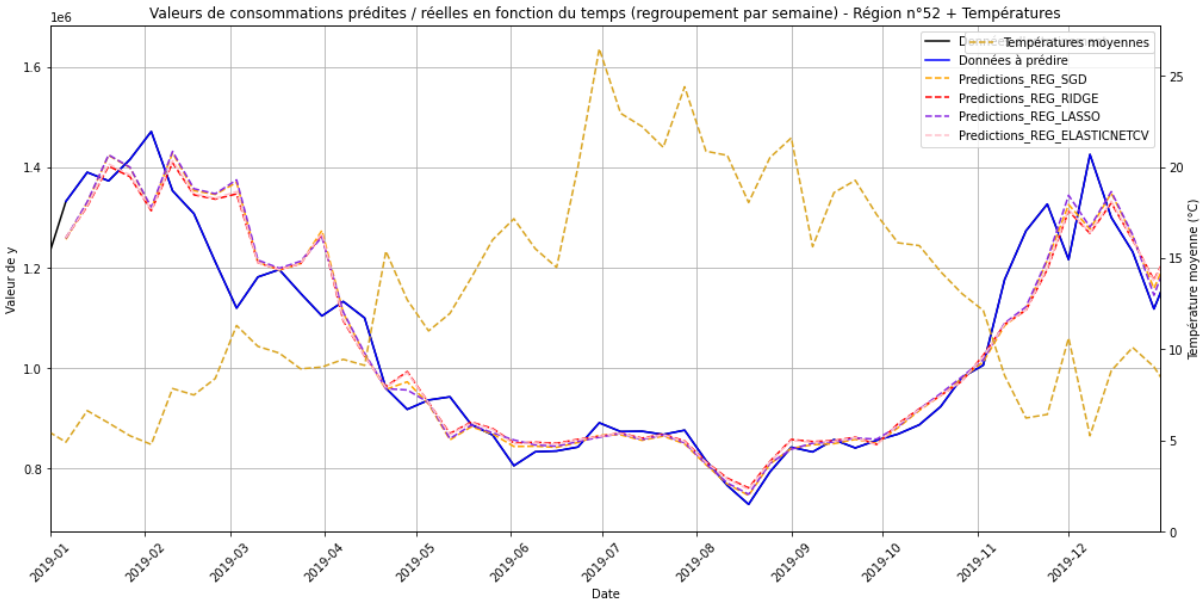



*Figure 23 - Valeur de consommation à prédire pour l'année 2019, prédictions des 4 modèles de régressions et valeurs des températures (regroupement hebdomadaire)*

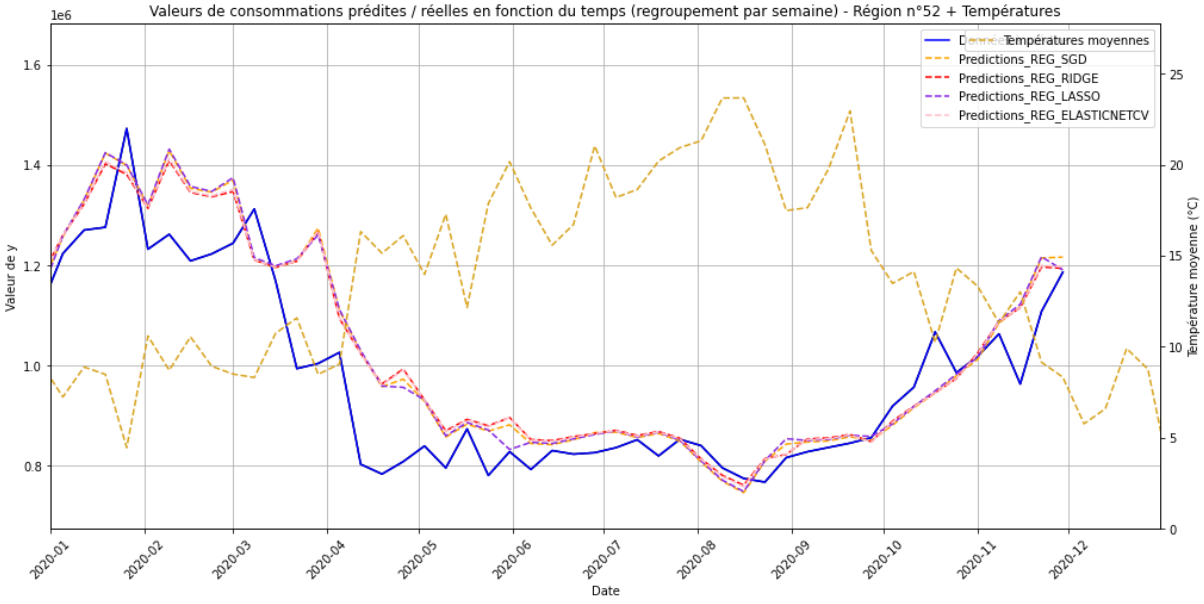



*Figure 24 - Valeur de consommation à prédire pour l'année 2020, prédictions des 4 modèles de régressions et valeurs des températures (regroupement hebdomadaire)*

| | MAE | MSE | RMSE | Median_AE | MAPE | Score_R2 |
|---|---|---|---|---|---|---|
| SGD | 58948.183103 | 7.198618e+09 | 84844.669028 | 32936.363107 | 5.272595 | 0.825192 |
| RIDGE | 58949.795867 | 6.936466e+09 | 83285.450795 | 34396.830959 | 5.341588 | 0.831558 |
| LASSO | 58461.969830 | 7.137702e+09 | 84484.921484 | 33722.252911 | 5.214530 | 0.826672 |
| ELASTICNETCV | 58991.777477 | 6.995325e+09 | 83638.058499 | 33754.627284 | 5.331186 | 0.830129 |

*Tableau 2 : Récapitulatif des différents scores pour les 4 modèles de régression pour les données prédites*

| | MAE | MSE | RMSE | Median_AE | MAPE | Score_R2 |
|---|---|---|---|---|---|---|
| SGD | 47003.870774 | 4.399177e+09 | 66326.292876 | 30118.479418 | 4.228739 | 0.910526 |
| RIDGE | 48589.062172 | 4.710901e+09 | 68636.003119 | 31006.686974 | 4.388141 | 0.904186 |
| LASSO | 46005.046460 | 4.318353e+09 | 65714.173183 | 28778.148035 | 4.129855 | 0.912170 |
| ELASTICNETCV | 48345.062603 | 4.652536e+09 | 68209.504468 | 30806.628964 | 4.363452 | 0.905373 |

*Tableau 3 : Récapitulatif des différents scores pour les 4 modèles de régression pour les données d'entrainement (regroupement hebdomadaire)*

Les mêmes commentaires que la partie précédente peuvent être formulés à une échelle différente. Le modèle de régression « Ridge » est également le modèle présentant de meilleurs scores (RMSE et R2).

## III.3 – Modèles de « Séries Temporelles »

Les Figures 25 et 26 présentent les prédictions des modèles SARIMAX et ARIMA.

Les mêmes observations peuvent être faites que pour les modèles de régressions pour le modèle SARIMAX.

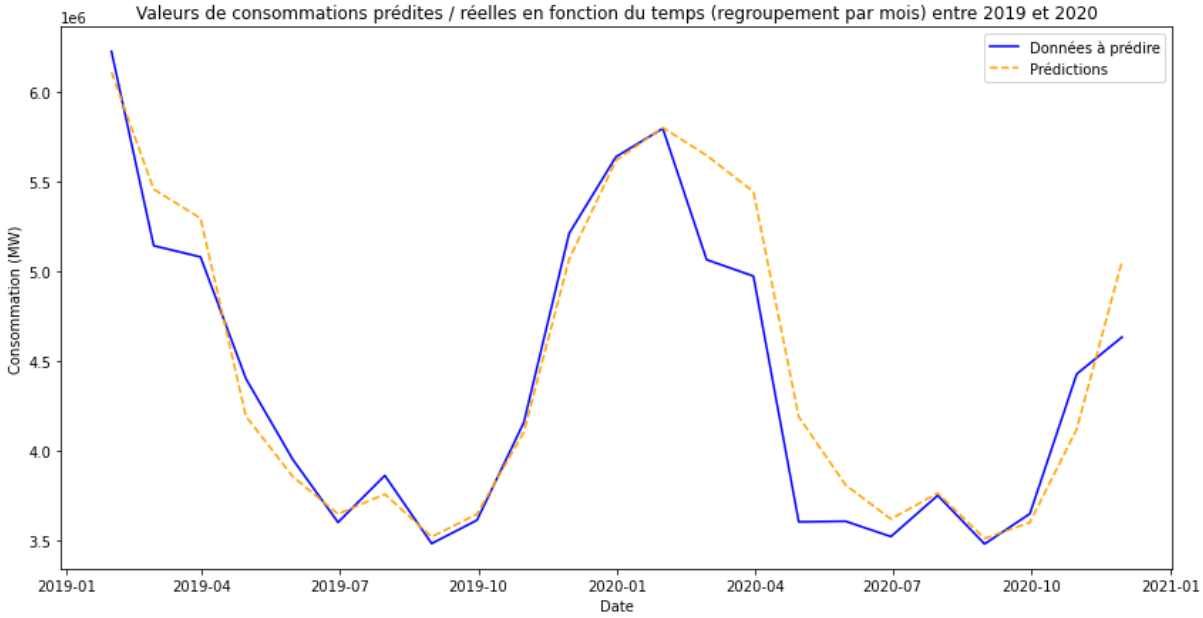



*Figure 25 : Valeur de consommation à prédire (2019 à 2020) et prédictions avec le modèle SARIMAX (regroupement mensuel)*

Le modèle ARIMA mis en place présente quant à lui des écarts plus importants.

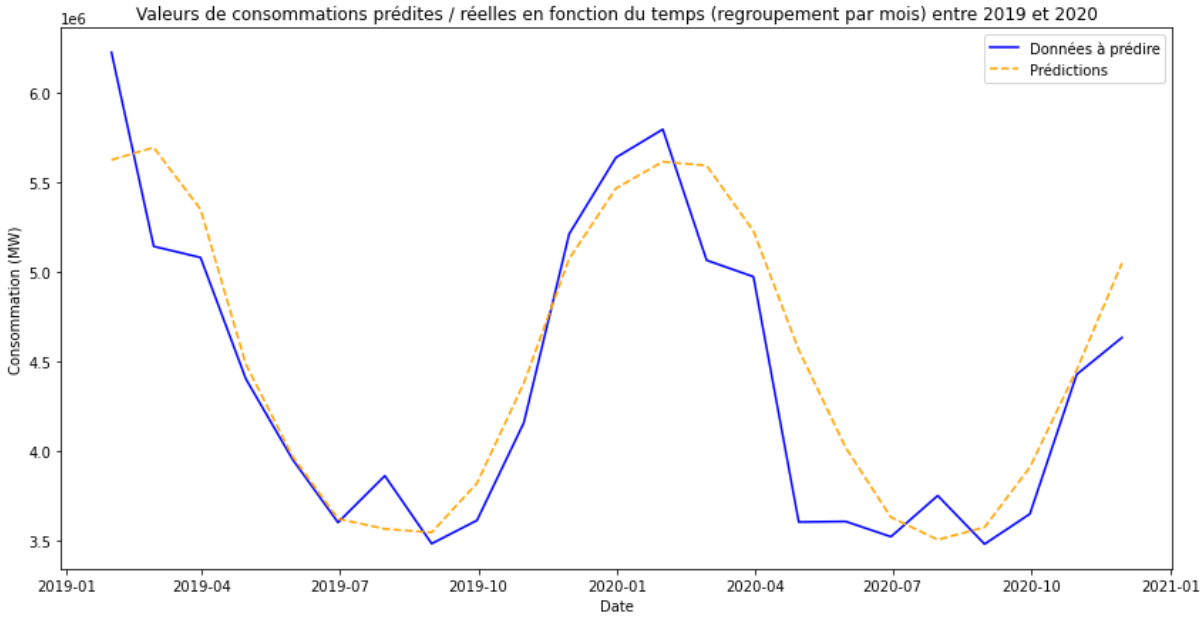



*Figure 26 : Valeur de consommation à prédire (2019 à 2020) et prédictions avec le modèle ARIMA (regroupement mensuel)*

Les Figures 27, 28 et le Tableau 4 comparent l'ensemble des modèles pour les prédictions.

Ils confirment que le modèle ARIMA présente des scores inférieurs aux autres modèles. Le modèle SARIMAX s'avère être quant à lui le modèle représentant le mieux les variations.

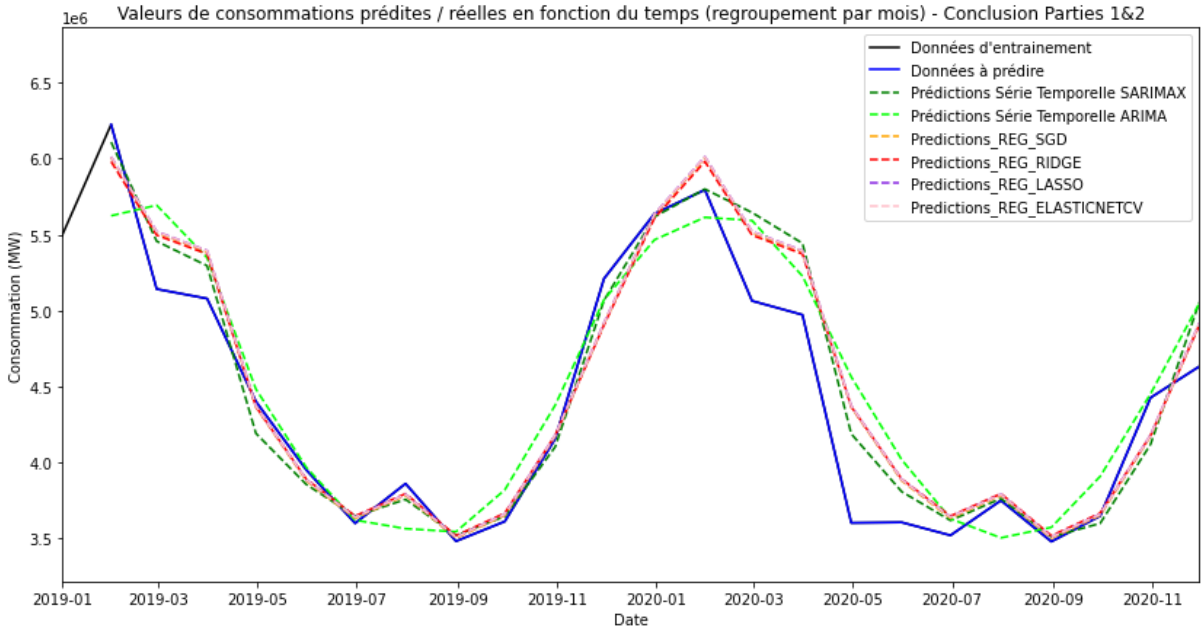



*Figure 27 : Valeur de consommation à prédire (2019 à 2020) et prédictions des 4 modèles de régressions et des 2 modèles de Séries Temporelles (regroupement mensuel)*

| | MAE | MSE | RMSE | Median_AE | MAPE | Score_R2 |
|---|---|---|---|---|---|---|
| **SGD** | 187527.160461 | 7.014346e+10 | 264846.112765 | 111941.835781 | 3.926028 | 0.897883 |
| **RIDGE** | 189695.731201 | 6.886756e+10 | 262426.300693 | 126948.345856 | 4.004761 | 0.899740 |
| **LASSO** | 190517.266971 | 7.178957e+10 | 267935.760930 | 119857.391889 | 3.993462 | 0.895486 |
| **ELASTICNETCV** | 190116.729820 | 7.154518e+10 | 267479.301639 | 118897.796264 | 3.984656 | 0.895842 |
| **SARIMAX** | 180542.655761 | 6.430695e+10 | 253588.145258 | 103705.989911 | 3.873346 | 0.906380 |
| **ARIMA** | 266184.617773 | 1.202856e+11 | 346822.133741 | 220557.435002 | 20.516286 | 0.824884 |

*Tableau 4 : Récapitulatif des différents scores pour les 4 modèles de régression et des 2 modèles de Séries Temporelles pour les données d'entrainement (regroupement mensuel)*

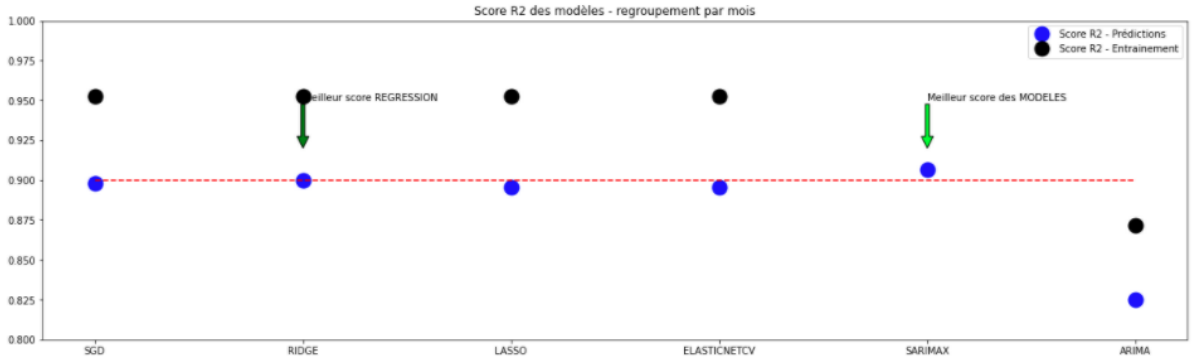



*Figure 28 : Valeur des coefficients de détermination R2 pour chaque modèle de régressions et chaque modèle de Séries Temporelles pour les données d'entrainement et les prédictions*

## III.4 – Modèle de réseau de neurones récurrents

### III.4.a – Choix du modèle

Après l'élaboration des différents modèles présentés, nous avons essayé d'aborder la problématique de prédiction de la consommation électrique sous un nouvel angle.

Nous avons constaté que notre variable cible variait principalement en fonction du temps et qu'on pouvait s'inspirer des méthodes que les traders en bourse utilisent pour prédire le prix des cours des actions.

Nous nous sommes donc penchés vers un modèle de réseau de neurones récurrents RNN (Recurrent Neural Network). Bien adapté aux données de séries chronologiques, les RNN traitent une série chronologique étape par étape, en conservant un état interne d'un pas de temps à l'autre.

Bien qu'ils soient construits spécifiquement pour gérer des séquences, les RNN simples ont certaines limites, notamment dans le traitement de séquences longues (modèle à mémoire courte).

Pour remédier à cette problématique, nous avons choisi d'implémenter un modèle de type RNN plus complexe, le LSTM (pour Long Short-Term Memory), sa structure introduit un mécanisme de mémoire des entrées précédentes qui persiste dans les états internes du réseau et peut ainsi impacter toutes ses sorties futures.

### III.4.b – Implémentation du modèle

Pour des raisons de performance, nous avons décidé d'implémenter un modèle LSTM pour chaque région en utilisant les data set régionaux que nous avions créé dans l'étape d'Analyse statistique.

L'implémentation de chaque modèle LSTM s'est déroulé suivant les étapes suivantes :

- Convertir les données en array Numpy
- Uniformiser les données avec un MinMaxScaler
- Création des données d'entrainement, en prenant les données uniformisées entre le 1er Janvier 2013 et 31 Décembre 2018
- Création d'un ensemble d'entrainement X_train et y_train représentant pour chaque série de 50 valeurs de y_train une série X_train contenant les 100 valeurs précédentes de y_train.
- Création du modèle LSTM avec :
    - Une couche LSTM de 100 neurones
    - Une couche Dense de 100 neurones
    - Une couche Dense de 50 neurones
    - Une couche Dense de 1 neurone
- Compilation du modèle avec les paramètres suivants :
    - Optimiseur = 'SGD'
    - Erreur = 'Erreur Quadratique Moyenne'
    - Métrique = 'précision'
- Entrainement du modèle sur l'ensemble d'entrainement.
- Création des données de tests, en prenant les données uniformisées entre le 1er Janvier et le 31 Décembre 2019.
- Création d'un ensemble de test X_test et y_test
- Prédiction du modèle sur l'ensemble de test.
- Récupérer les valeurs de prédiction et les reconvertir vers leurs valeurs réelles à l'aide de la méthode inverse_transform() du MinMaxScaler

- Enregistrer les données prédites dans un fichier csv.

L'implémentation, l'entrainement et la prédiction des modèles pour chaque région a nécessité un temps d'exécution de : 1 jour, 5 heures, 51 minutes et 38 secondes.

### III.4.c – Résultat du modèle

Avant de passer à la visualisation des prédictions du modèle, la figure ci-dessous illustre les valeurs de l'erreur RMSE (Root Mean Squared Error) calculé pour chaque région :

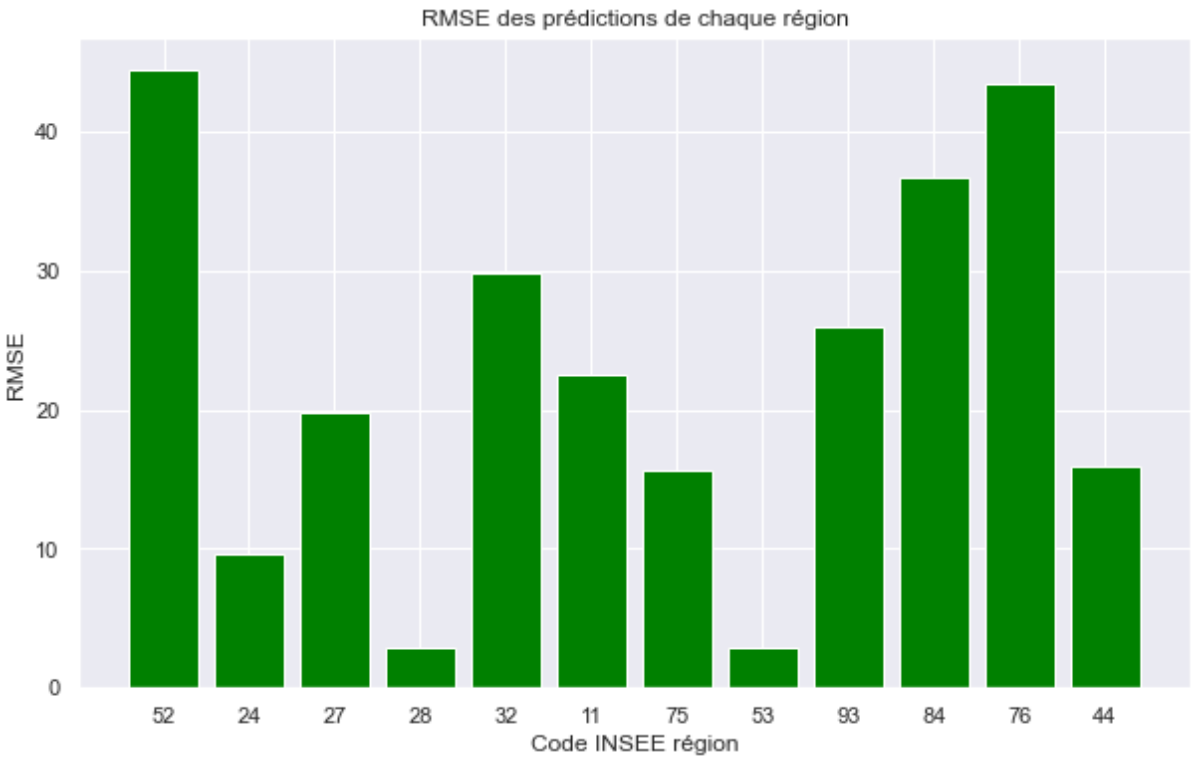



*Figure 29 - RMSE des prédictions de chaque région*

Nous constatons de ce graphique que la RMSE est très faible dans chaque région, par rapport à la distribution des valeurs de la consommation.

Afin de visualiser en détail les prédictions du modèle, les figures ci-dessous illustrent les prédictions pour la région 52 (Pays de la Loire) sur une semaine au hasard sur les 4 saisons :

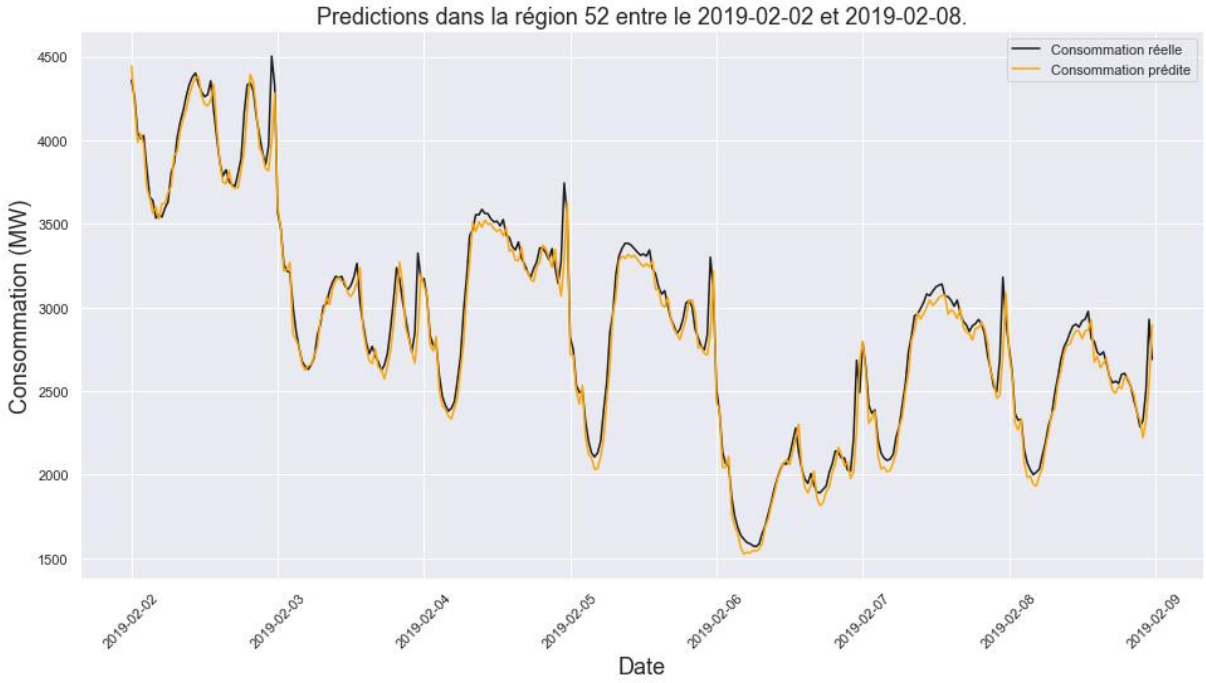



*Figure 30 – Prédictions dans la région 52 en Hiver*

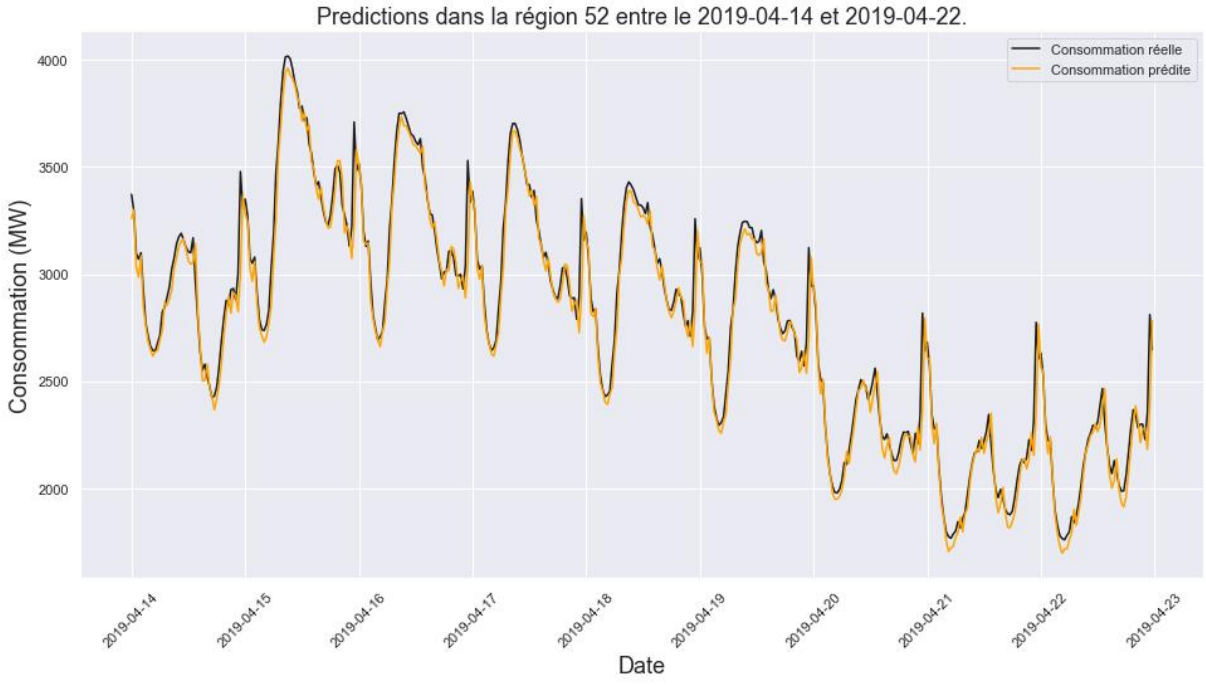



*Figure 31 - Prédictions dans la région 52 au Printemps*

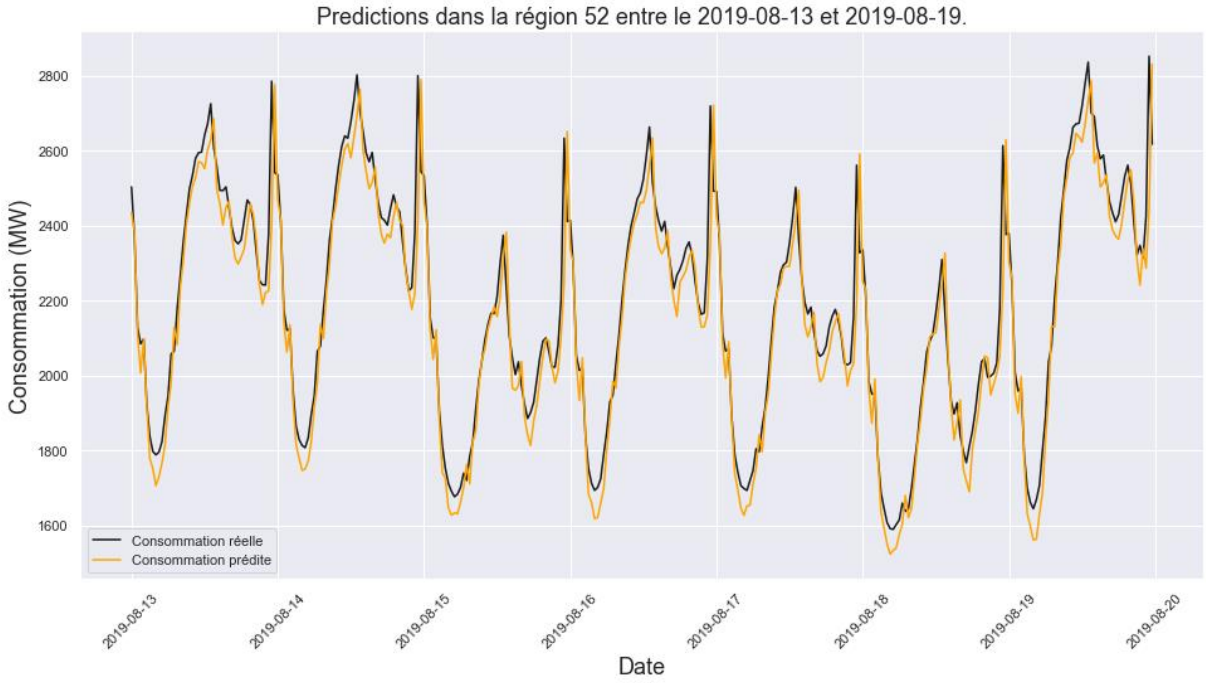



*Figure 32 - Prédictions dans la région 52 en été*

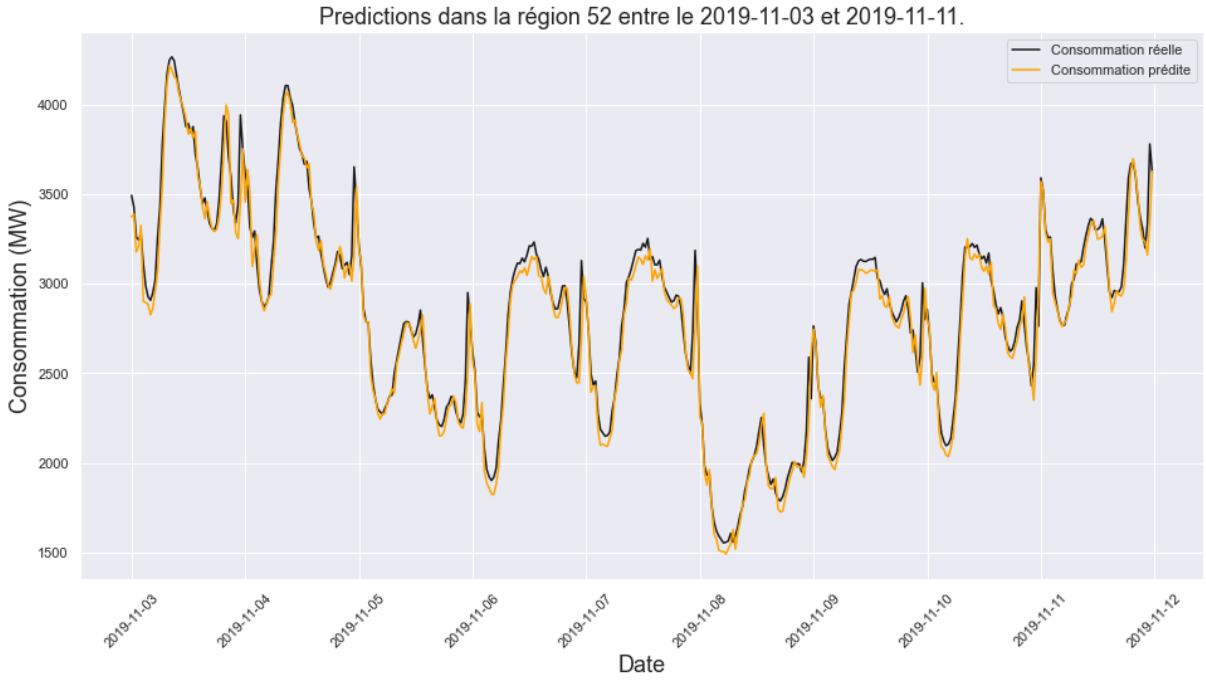



*Figure 33 - Prédictions dans la région 52 en Automne*

Ces modèles LSTM donnent des résultats absolument remarquables en comparaison des autres modèles, les prédictions suivent exactement les mêmes courbes que les valeurs réelles, avec un certain décalage que nous avons essayé de mesurer afin de déterminer un intervalle de confiance pour chaque modèle.

Par exemple pour la région 52 (Pays de la Loire), en calculant le Delta entre les valeurs prédites et les valeurs réelles nous obtenons la distribution suivante :

| | Consommation_MW | Predictions | Delta |
|---|---|---|---|
| count | 17544.000000 | 17544.000000 | 17544.000000 |
| mean | 3103.762711 | 3059.275675 | 67.400795 |
| std | 786.058934 | 785.111158 | 71.474365 |
| min | 1555.000000 | 1492.600300 | 0.003418 |
| 25% | 2499.750000 | 2454.869300 | 25.500214 |
| 50% | 3022.000000 | 2974.985950 | 50.132690 |
| 75% | 3629.000000 | 3589.099350 | 83.270386 |
| max | 5501.000000 | 5341.640600 | 531.396973 |

*Figure 34 - Analyse de la variable Delta*

De ce tableau, nous pouvons conclure que la différence moyenne entre les valeurs prédites et les valeurs réelles est de 97.4 MW sur l'année 2019, avec une différence maximale de 531.39 MW.

### III.4.d – Conclusion sur le modèle LSTM

Le modèle LSTM offre des résultats de prédictions exceptionnels, néanmoins il représente un désavantage conséquent qui est le coût de son exécution. En effet l'implémentation de ce modèle nécessite beaucoup de ressources, et ne convient pas pour une utilisation en temps réelle.

# *Conclusion générale*

La mission du projet, qui est d'apporter une solution à une problématique de stockage électrique, en implémentant un modèle de prédiction de la consommation électrique afin de réguler la production et d'éviter les pertes du réseau, nous a permis de développer en nous, une nouvelle vision des données et une nouvelle méthodologie d'appréhension de problématiques complexes, en utilisant de nouveaux outils et en mettant en pratique de nouvelles compétences.

Pour la réalisation de ce projet, nous sommes passés par plusieurs étapes, de l'exploration des données, leurs analyses, leurs visualisations, jusqu'à l'implémentation de modèles de prédiction.

Les résultats de modélisation nous permettent d'affirmer que c'est les modèles « Ridge » (régression) et SARIMAX (série temporelle) qui ont donné les résultats les plus fiables, et qui anticipent le mieux les variations de la consommation électrique.

Même si l'écart entre tous les modèles implémentés reste faible, le modèle ARIMA ne semble pas apporter de bons résultats.

Quant au modèle LSTM, ses résultats de prédictions sont exceptionnels. Néanmoins il représente un désavantage conséquent qui est le coût de son exécution, car ce modèle est très gourmand en ressources et en temps de calcul.

Les prédictions sur l'année 2019 semblent plus pertinentes que pour l'année 2020 qui s'avère être une année particulière (réchauffement climatique, effet de la pandémie de Covid-19, …) ce qui rajoute de la difficulté pour prédire la courbe de consommation électrique de cette année-là. Ceci est également renforcé par le fait que les données pour l'année 2020 ne sont pas encore définitives.

# *Perspectives*

Pour la continuité du projet, et dans l'optique d'améliorer les performances des modèles de prédictions, nous pensons qu'il serait pertinent de rajouter une variable température au data set, et pouvoir étudier sa corrélation avec la consommation électrique et ainsi l'utiliser dans les données d'entrée des futurs modèles.

Avec l'impact du Covid19, il serait peut être utile de prendre en considération une variable 'Confinement' et/ou 'Couvre-feu' pour pouvoir anticiper des scénarios futurs liés à d'éventuelles nouvelles mesures.

# *Table des illustrations*