Rapport du Projet DS - Energie

Réalisé par : Romain Mouly et Zéphirin Nganmeni

Sous la supervision de : Mounir

29 septembre 2021

Table des matières

Chapitre Premier

Introduction et préliminaires

1.1 Introduction

Ce document est le rapport du projet réalisé dans le cadre de la formation en data science chez Datascientest. Le projet porte sur le thème « Energie » et est réalisé par Romain Mouly et Zéphirin Nganmeni sur la coordination de Mounir. L'objectif est de : (1) constater le phasage entre la consommation et la production énergétique au niveau national et au niveau départemental (risque de blackout notamment) ; (2) analyse au niveau départemental pour en déduire une prévision de consommation ; (3) analyse par filière de production : énergie nucléaire / renouvelable avec un focus sur les énergies renouvelables (où sont-elles implantées ?). La source de données est celle de l'ODRE (Open Data Réseaux Energies) : on a accès à toutes les informations de consommation et production par filière jour par jour (toutes les 1/2 heure) depuis 2013. Nous disposons également d'autres jeux de données notamment, des données sur les variations : de la température, du rayonnement solaire, de la vitesse du vent, etc. Pour atteindre les objectifs annoncés, notre travail est organisé en 5 grandes parties numérotées de 1 à 5.

Cette présentation introductive est le début de la première partie. Elle se complète par la présentation du contexte et de la problématique, l'analyse de l'architecture des données et quelques opérations de prétraitement des données. La deuxième partie porte sur l'analyse exploratoire. Nous y proposons une analyse globale (à l'échelle nationale) et une analyse régionale. Ce travail réalisé essentiellement pendant la première itération du projet contient beaucoup de graphique illustratifs, des commentaires, des conjectures et quelques analyses préliminaires. La troisième partie du document porte sur la modélisation et l'analyse des

résultats. Nous avons regroupé nos modèles en deux groupes : les approches basées sur les séries temporelles et les approches basés sur le Machine Learning. La quatrième partie porte sur les difficultés rencontrées, la conclusion et les perspectives. L'annexe constitue la dernière partie du document.

1.2 Contexte et problématique

L'énergie est une ressource incontournable pour le fonctionnement des équipements utilisés dans divers secteurs socio-économiques : santé, éclairage, chauffage, transports, agriculture, etc. Pour assurer l'équilibre entre l'offre et la demande énergétique, il faut anticiper et prévoir au quotidien, la demande énergétique qui varie au gré des phénomènes qui affectent le comportement des consommateurs. Par ailleurs, étant donné l'impact du mode de production de l'énergie sur l'environnement, on privilégie les énergies renouvelables. Ainsi, notre problématique porte sur les prévisions des évolutions de la demande et d'offre énergétiques en France en vue d'assurer la disponibilité tout en limitant les impacts environnementaux.

Le but va être dans un premier temps de faire une étude de la situation passée et présente en terme d'adéquation de la production à la consommation et d'analyser les sources de production utilisées. Ensuite, sur la base des données, l'objectif sera d'effectuer des modélisations permettant de prévoir les évolutions des consommations et productions électriques et de repérer potentiellement les points de tension. Egalement, une attention spéciale sera apportée à la situation des énergies renouvelables dans ce contexte.

1.3 Architecture des données

Pour aborder cette problématique, nous avons à disposition les données régionales consolidées de janvier 2013 à juin 2021 issues de l'application éCO2mix. On y trouve en méga watts les données françaises de consommation et de production électrique suivant le type de génération. Les données disponibles sont organisées suivant le type de génération (Eolien, Solaire, Hydraulique, Thermique, Nucléaire, Bioénergies ainsi que la puissance consommée par le pompage des stations de transfert d'énergie). Nous avons également les flux de transfert

reçus ou émis permettant de combler les déficits éventuels en énergie.

Ces données sont détaillées par : région administrative française, date et horaire (nous avons une ligne de données en Mega Watt sur un pas de demi-heure).

1.4 Prétraitement des données

Le pré-traitement des données va consister principalement à réaliser les opérations suivantes :

- rendre utilisable les dates fournies dans le fichier ;
- supprimer les colonnes sans données ou avec énormément de données manquantes (telles que certains taux calculés, ou les flux d'exportation de la région Grand-Est) ;
- remplacer les données manquantes restantes par des 0 dans les données de production ;
- commencer le feature engineering en calculant des colonnes agrégeant les productions de sources renouvelables, donnant le solde énergétique avant et après transfert et les taux de couverture des besoins (toute source/renouvelable uniquement).

Pour avoir un ordre d'idées, voici le nombre de lignes avant traitement, ainsi que le nombre de données manquantes pour les colonnes principales :

Nombre de lignes total = 1787328	
Code INSEE région	0
Région	0
Nature	0
Date	0
Heure	0
Date - Heure	0
Consommation (MW)	12
Thermique (MW)	12
Nucléaire (MW)	744727
Eolien (MW)	108
Solaire (MW)	12
Hydraulique (MW)	12
Pompage (MW)	779767
Bioénergies (MW)	12
Ech. physiques (MW)	12

Après traitement, nous ne supprimons que 12 lignes de la base (les 12 lignes sans aucune donnée que l'on peut percevoir dans le tableau ci-dessus). Pour le reste, nous avons remplacé les données manquantes par des 0.

Chapitre Deuxième

Analyse exploratoire

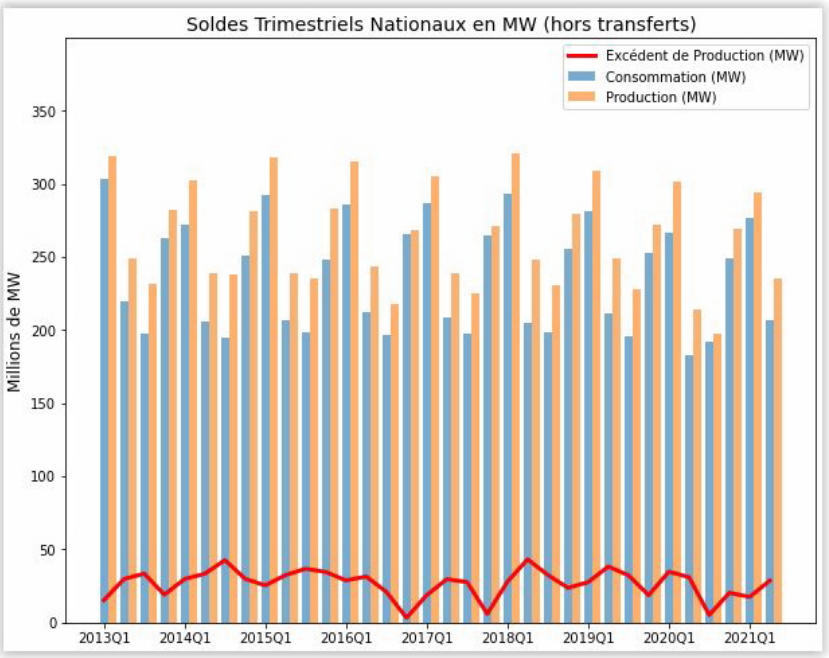
Etant donné que cette base de données peut être approchée selon différents points de vue, nous allons produire des graphiques et analyses variées pour identifier quels angles d'attaque utiliser dans nos modélisations.

2.1 Analyse globale (à l'échelle nationale)

Dans cette sous-section, nous proposons divers types d'analyse de la consommation énergétique sur le plan national (la colonne qui porte sur la région n'est pas prise en compte.). Nous proposons notamment les représentations suivantes : soldes trimestriels nationaux (hors transferts), consommations mensuelles moyennes, taux de couverture de la consommation par sources renouvelables et, mix énergétique par année.

2.1.1 Solde trimestriels nationaux (hors transferts)

Tout d'abord, agrégons à un haut niveau les données pour avoir un aperçu de la situation nationale à un pas trimestriel. Il s'agit d'un très grand angle d'analyse.



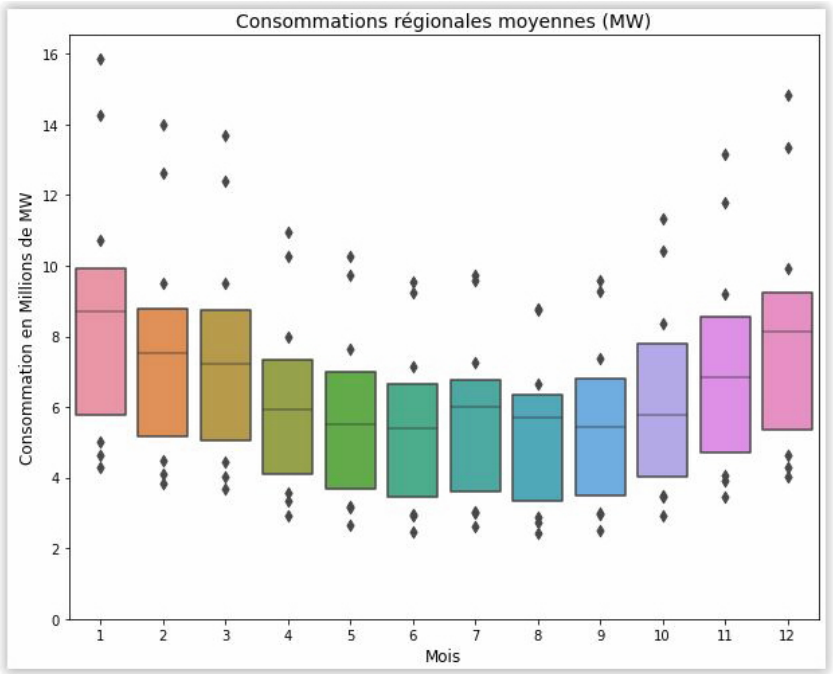
Test ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
Trimestre	3.00	855190461209777.00	285063487069925.69	3.81	0.02
Residual	30.00	2247324147505122.50	74910804916837.42	NaN	NaN

Nous pouvons voir un premier insight essentiel : la France est en permanence excédentaire dans sa capacité de production électrique. Certes, il peut exister certains trimestres en tension comme les derniers trimestres 2016 ou 2017. De plus, avec le test ANOVA ci-dessus, il apparaît clairement que le solde de production est fortement dépendant du trimestre de l'année (la p-value étant inférieure au seuil de 5%).

2.1.2 Consommations mensuelles moyennes

Observons ensuite la consommation mensuelle moyenne régionale. Nous rentrons un peu plus précisément dans les données.



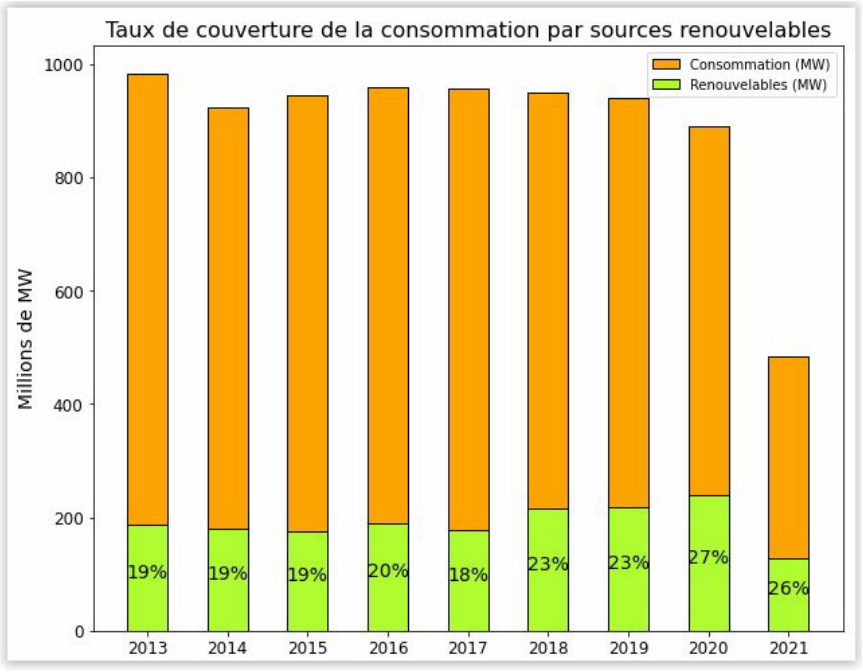
Test ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
Mois	1.00	124069519053615.25	124069519053615.25	13.81	0.00
Residual	1222.00	10978673482968658.00	8984184519614.29	NaN	NaN

Il y a une claire saisonnalité de la consommation électrique qui est à son pic en hiver et à son minimum en été (test de la p-value inférieur à 5%). Cependant, on peut remarquer une grande disparité dans la consommation entre les régions grâce au boxenplot.

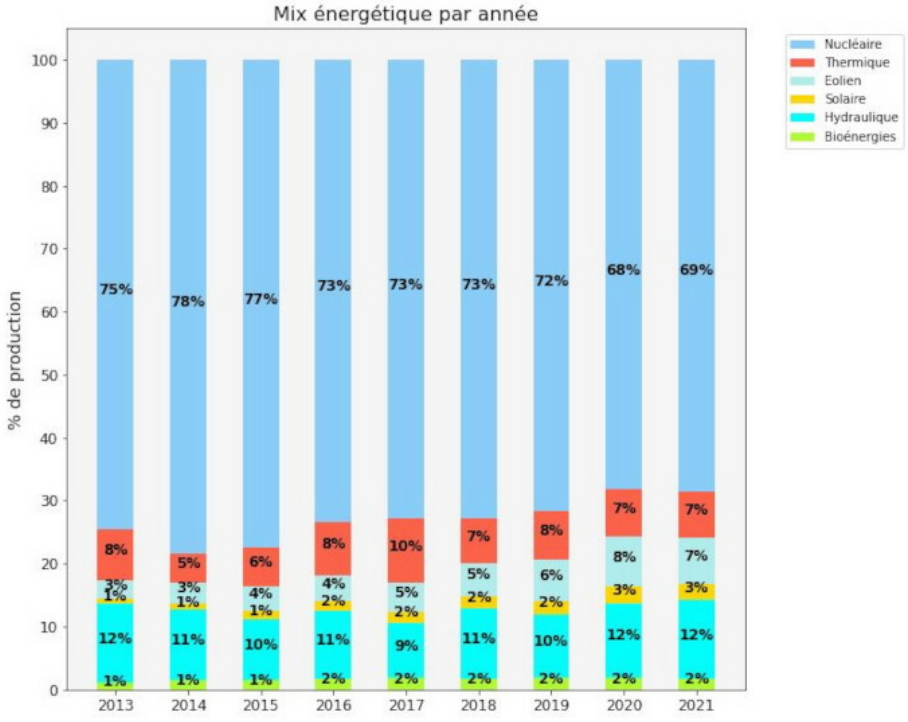
2.1.3 Taux de couverture de la conso par sources renouvelables

Maintenant que nous avons identifié des saisonnalités, nous allons rentrer dans le détail des consommations et productions. Le premier point d'intérêt est la production d'énergies renouvelables.



Nous constatons que la couverture de la consommation par les sources renouvables a augmenté sensiblement à partir de 2018 et se stabilise maintenant aux alentours des 27%.

2.1.4 Mix énergétique par année

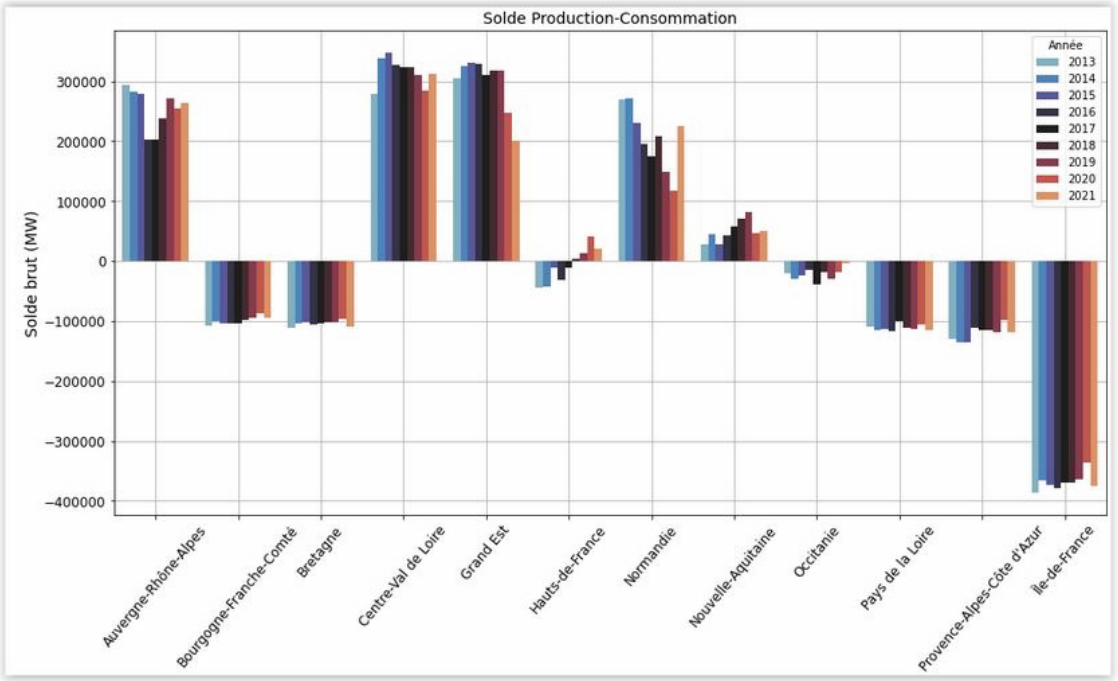


Il apparait que sur la période considérée, la part des énergies renouvelables croit. En particulier, l'énergie éolienne connait une hausse très significative. Sa proportion dans le mix énergétique est passée de 1% en 2013-2015 à 8% en 2020. Sur la même période, la proportion de l'énergie solaire est passée de 1% à 3%. La proportion des autres types d'énergies renouvelables est restée quasiment constante. La hausse de la proportion des énergies renouvelables se traduit naturellement par la baisse de la proportion de l'énergie nucléaire. Celle-ci est passée de 75% en 2013 à 68 % en 2020. L'énergie nucléaire reste donc utilisée de manière prépondérante.

2.2 Analyse par région

Nous allons à présent mener une étude comparative entre les différentes régions.

2.2.1 Solde production-conso en fonction de la région

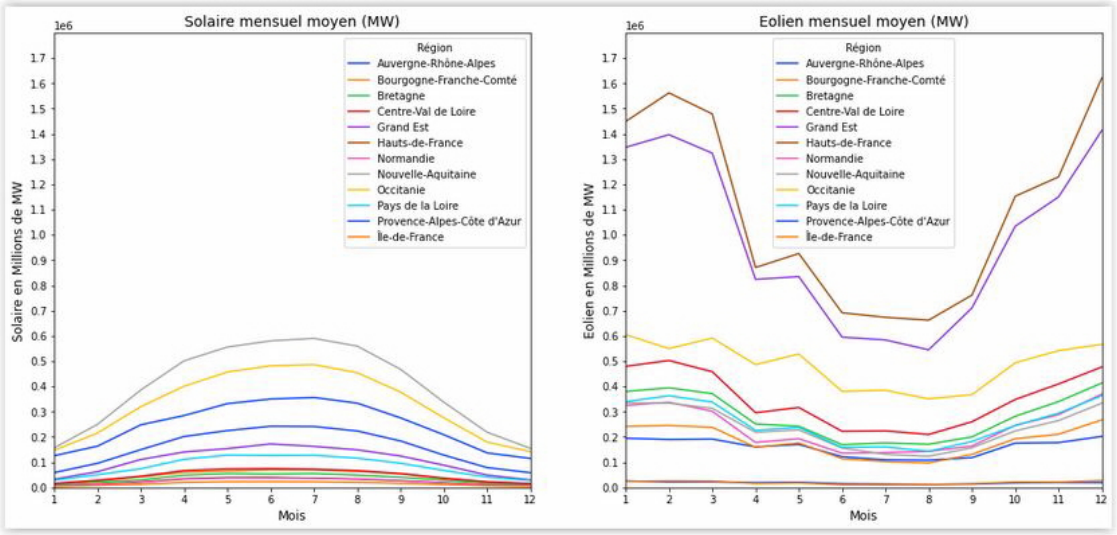


Test ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
Région	11.0	4.275360e+12	3.886691e+11	580.751464	1.276866e-82
Residual	96.0	6.424820e+10	6.692521e+08	NaN	NaN

2.2.2 Solaire mensuel moyen, énergie éolien mensuel moyen

En détaillant le solaire et l'éolien ci-dessous, nous pouvons voir un lien entre la saisonnalité et ces productions d'énergie (solaire en été, éolien en hiver). Cependant, les tests des p-value ne sont pas concluants avec des p-values extrêmement élevées : un grand nombre de régions ne sont pas impliquées dans ces productions.



Test ANOVA Solaire

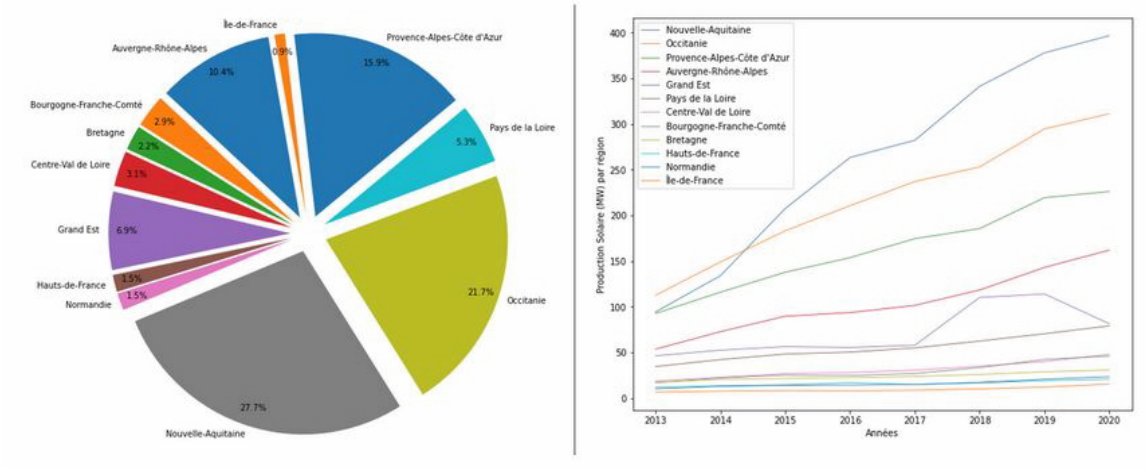
	df	sum_sq	mean_sq	F	PR(>F)
Mois	1.0	2.367429e+09	2.367429e+09	0.114168	0.735947
Residual	142.0	2.944564e+12	2.073637e+10	NaN	NaN

Test ANOVA Eolien

	df	sum_sq	mean_sq	F	PR(>F)
Mois	1.0	6.326570e+10	6.326570e+10	0.479229	0.489902
Residual	142.0	1.874622e+13	1.320157e+11	NaN	NaN

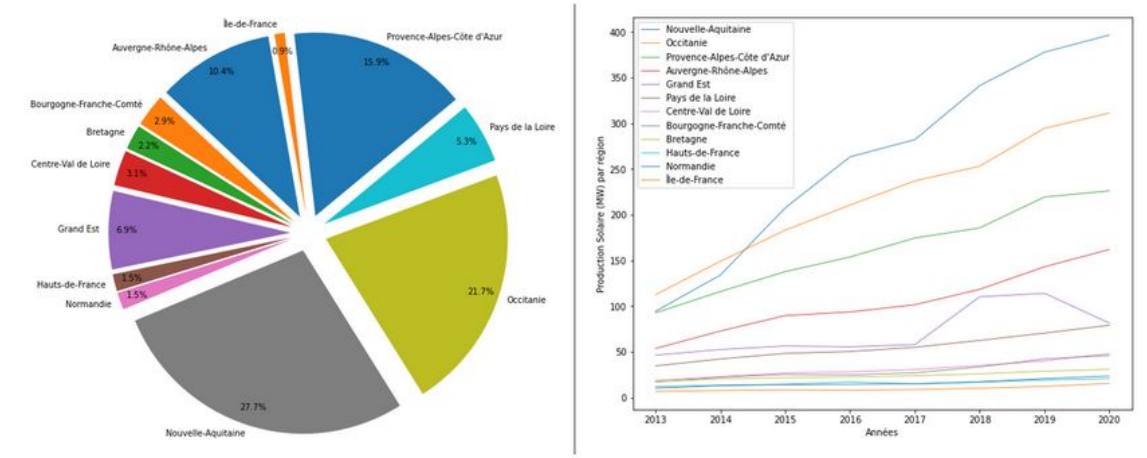
Au niveau géographique, nous pouvons voir que la situation est très différenciée avec certaines clairement excédentaires et d'autres en demande d'électricité. Naturellement, la p-value du test effectué ci-dessus confirme la relation entre la région et le solde électrique (p-value inférieure à 5%)

2.2.3 Production moyenne solaire en fonction de la région



La croissance de la production de l'énergie solaire est plus importante pour certaines régions telles que Nouvelle-Aquitaine, l'Occitanie, Provence-Alpes-Côte d'Azur et Auvergne-Rhône-Alpes. Le niveau de production est très faible et reste stagnant depuis 2013 pour les régions telles que : Île-de-France, Normandie, Hauts-de-France, Bretagne, Bourgogne-Franche-Comté, Centre-Val de Loire et Pays de la Loire. Après une période de croissance, on note depuis 2019 une décroissance de la production de la région de Grand Est.

2.2.4 Production moyenne de l'énergie solaire par heure, par jour



La production de l'énergie solaire est concentrée sur le créneau horaire 7h-19h : c'est la période de la journée où on dispose du rayonnement solaire nécessaire à la production. Les variations relatives observées dans la production ne dépendent ni de la région ni du jour de

la semaine. En d'autres termes, il y'a une corrélation entre les productions. Cela se vérifie à travers une table de corrélation.

Observations De l'analyse exploratoire, nous pouvons constater les élements suivants :

- il existe une saisonnalité des consommations et productions d'énergie au niveau régional et national
- cette saisonnalité peut être moins présente en détaillant par sources d'énergie du fait des disparités régionales
- les régions sont des déterminants importants des données d'énergie, avec des producteurs nets (Centre-Val de Loire, Grand-Est) et des consommateurs nets (PACA, Ile-de-France)
- cela signifie que même si au niveau national, la France est en permanence excédentaire en énergie, la situation est différente au niveau régional
- les renouvelables augmentent progressivement dans le mix énergétique français, avec une accélération sur les dernières années.

Chapitre Troisième

Modélisation et analyse des résultats

Nous allons utiliser différents modèles de Machine Learning pour confirmer et infirmer les pistes que nous avons identifié mais également pour obtenir des informations supplémentaires de cette base.

3.1 Approches basées sur les séries temporelles

Le premier type de modèles utilisé va être une analyse de séries temporelles. Cette base se prête particulièrement à ce type d'analyse car indexée par date. Le but va être de tirer profit de certaines saisonnalités identifiées pour effectuer des prévisions de production et consommation sur les prochaines années. Ainsi, si les modèles sont satisfaisants, nous pourrons effectuer des prévisions de tendance.

Les modèles sur séries temporelles vont tous suivre la même méthodologie :

- en premier, une décomposition saisonnière des données pour vérifier la tendance, et la stabilité éventuelle des résidus non-saisonniers ;
- ensuite, une analyse des auto-corrélations pour décomposer les séries de manière pertinentes ;
- enfin, l'application des paramètres trouvés dans un modèle Sarimax qui permettra d'effectuer des prévisions dans le temps

Dans ce cadre, l'objectif se situe au niveau national et au niveau des grandes tendances saisonnières. Nous allons donc regrouper nos données par mois et sans prendre en compte les régions.

A noter : tous les modèles Sarimax suivant la même méthodologie, nous allons détailler

le processus sur le premier cas puis passerons plus vite sur les cas suivants pour arriver à l'analyse des résultats.

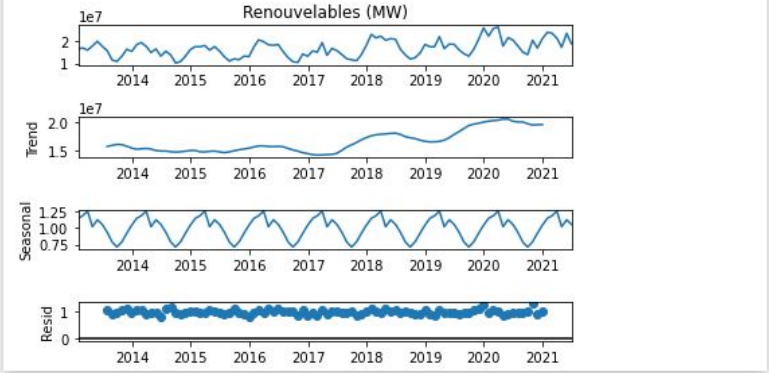
3.1.1 Prévision de la production des énergies renouvelables

Commençons avec les énergies renouvelables où nous allons regrouper :*

- le solaire
- l'éolien
- l'hydroélectrique (et le pompage lié à cette énergie)
- les bioénergies

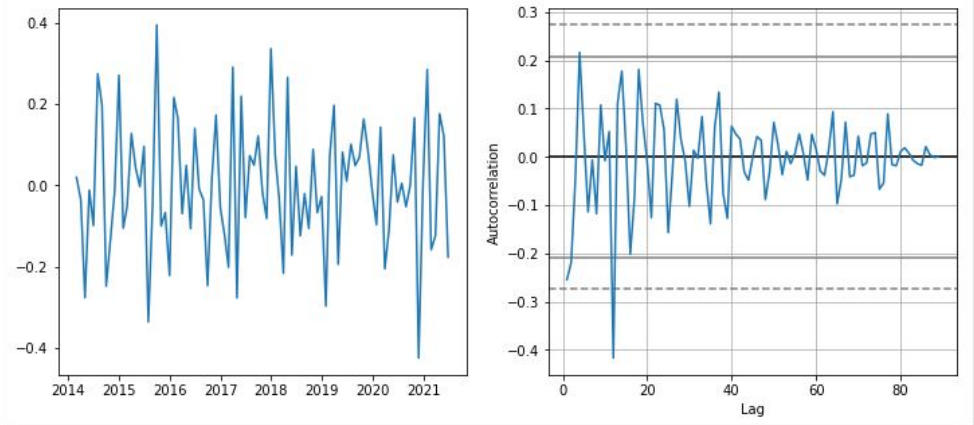
Avec une décomposition saisonnière, nous constatons que l'utilisation d'un modèle logarithmique permet de stabiliser les résidus et de les rendre aléatoires. On remarque la tendance qui est à la croissance depuis 2017. Egalement, l'effet saisonnier est marqué avec des pics de production lors des hivers et des minimums à l'été.

```
time_serie = temp_df['Renouvelables (MW)']
season_effect = seasonal_decompose(time_serie, model='multiplicative')
season_effect.plot()
plt.show();
```



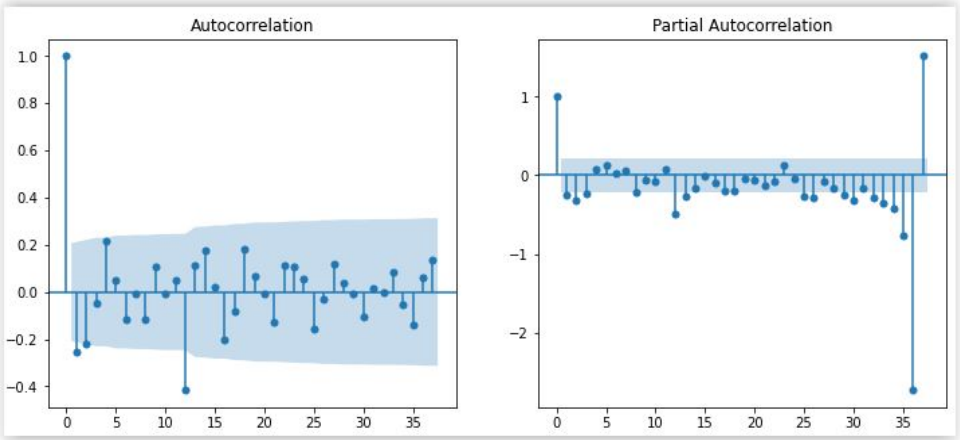
En procédant ensuite à la différenciation des données, nous arrivons sur des résultats stables en procédant à une première différenciation avec un pas de 1, puis une seconde avec un pas de 12 (donc sur la durée d'une année).

```
time_serie_log2 = time_serie_log1.diff(12).dropna()
fig, (ax1,ax2) = plt.subplots(1,2, figsize=(12,5))
ax1.plot(time_serie_log2)
pd.plotting.autocorrelation_plot(time_serie_log2, ax=ax2);
```



La valeur du test Augmented Dickey-Fuller est alors de 0.000009, ce qui est largement inférieur à 5% : nous pouvons considérer cette série comme stationnaire et donc fiable pour une modélisation Sarimax.

En traçant les graphiques d'autocorrélation, nous allons alors déterminer les paramètres du modèle : Pour la partie non saisonnière, on peut voir que les 2 graphiques convergent vers 0, nous allons partir vers p = q = 1, sachant que d = 1 puisque nous avions effectuer une première différenciation. Pour la partie saisonnière, D = 1 et la période est de 12 grâce à la seconde différenciation. Au niveau des autres paramètres, on va prendre p = q = 1 car les points 12/24 convergent vers 0 même si on voit un pic au rang 12. Il y a cependant un pic important à la fin de l'autocorrélation partielle des données mais qui peut être dû aux limites de notre échantillon.



Avec ces paramètres, nous obtenons un modèle Sarimax où la p-value du paramètre MA saisonnier est de 0.162, nous allons le supprimer (Q=0).

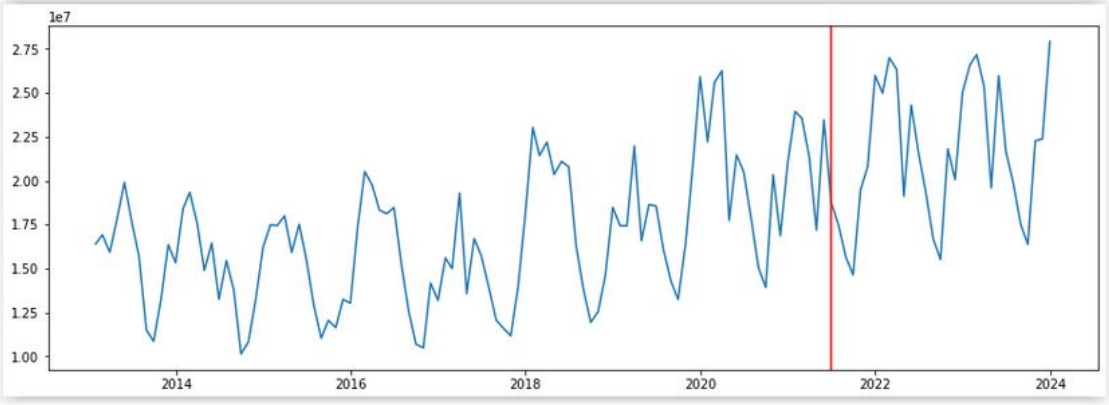
```
model = sm.tsa.SARIMAX(time_serie_log,order=(1,1,1), seasonal_order=(1,1,0,12))
model_fitted = model.fit()
model_fitted.summary()
```

Dep. Variable:	Renouvelables (MW)	**No. Observations:**	102
Model:	SARIMAX(1, 1, 1)x(1, 1, [], 12)	**Log Likelihood**	59.942
Date:	Tue, 24 Aug 2021	**AIC**	-111.883
Time:	11:42:41	**BIC**	-101.929
Sample:	01-31-2013	**HQIC**	-107.871
	- 06-30-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3985	0.128	3.110	0.002	0.147	0.650
ma.L1	-0.9288	0.063	-14.713	0.000	-1.053	-0.805
ar.S.L12	-0.5931	0.085	-6.940	0.000	-0.761	-0.426
sigma2	0.0141	0.002	6.422	0.000	0.010	0.018

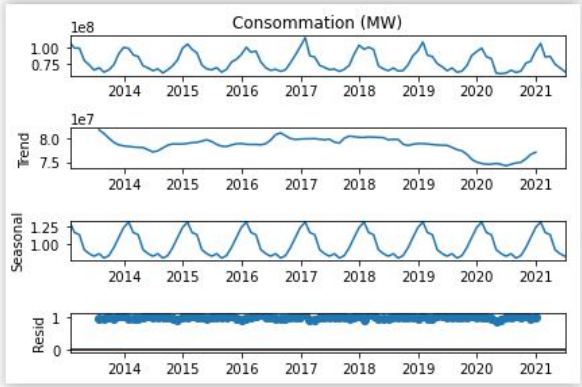
Ljung-Box (L1) (Q):	0.02	**Jarque-Bera (JB):**	0.37
Prob(Q):	0.89	**Prob(JB):**	0.83
Heteroskedasticity (H):	1.20	**Skew:**	0.13
Prob(H) (two-sided):	0.62	**Kurtosis:**	2.83

Nous avons alors un modèle qui a priori fonctionne correctement avec des tests de Ljung-Box et Jarque-Bera indiquant la blancheur des résidus ainsi que leur normalité. Au niveau des métriques d'erreur, nous obtenons une erreur moyenne relative de 16.7% et un RMSE de 2899247 quand le réel moyen est de 16761913 (soit un rapport de 17.3%). Le modèle fonctionne mais les variations mois à mois sont suffisamment importantes et erratiques pour rendre difficile une régression Sarimax. Voici les prévisions jusqu'à 2024 :



3.1.2 Prévision de la consommation globale

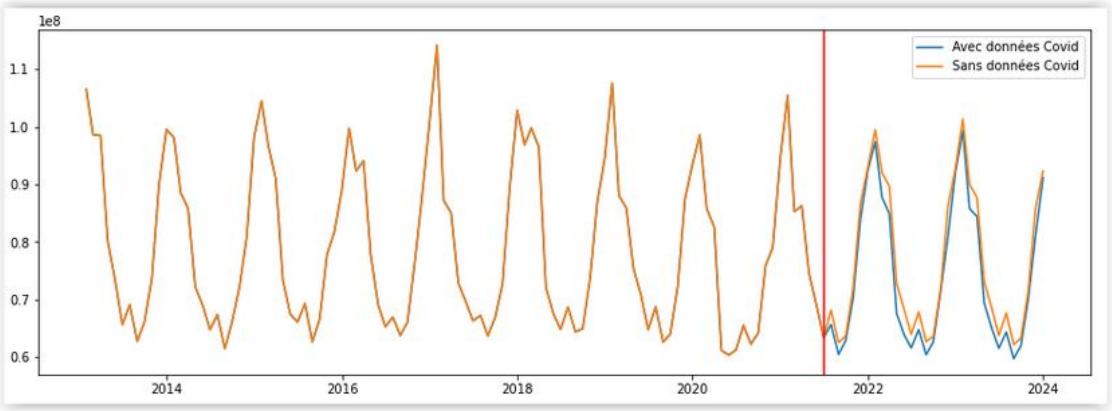
En ce qui concerne la consommation, nous prenons une échelle logarithmique et nous avons une stabilité des résidus. Nous constatons un changement de tendance à la baisse à partir de début 2020 et cela coincide avec le début de la pandémie du Covid-19.



Nous avons différencié les données comme précédemment et nous obtenons une P-value de 0.00018 pour le test de Dickey-Fuller, nous pouvons donc présumer que ce set de données est stationnaire. Grâce aux autocorrélations, nous définissons un SARIMAX "order=(1,1,1), seasonal_order=(1,1,1,12)"

Les résultats sont problématiques car le test de Jarque-Bera rejette la normalité. Le problème pourrait se situer en partie avec la rupture impliquée par la baisse de consommation liée au Covid qui rend moins claire la tendance future. De plus, au niveau des métriques, l'erreur moyenne relative représente environ 6% de la réalité, ce qui est assez efficace. Cependant, comme le modèle prend en compte l'année de Covid, cela veut dire que le modèle est paramétré pour prendre en compte cet effet pandémique. Cela peut être un souci pour effectuer des prévisions fiables dans l'avenir. Nous allons donc tenter de créer un modèle similaire mais en l'entraînant jusqu'à fin 2019 pour exclure les données liées à la pandémie et limiter les effets de ce choc exogène sur la consommation.

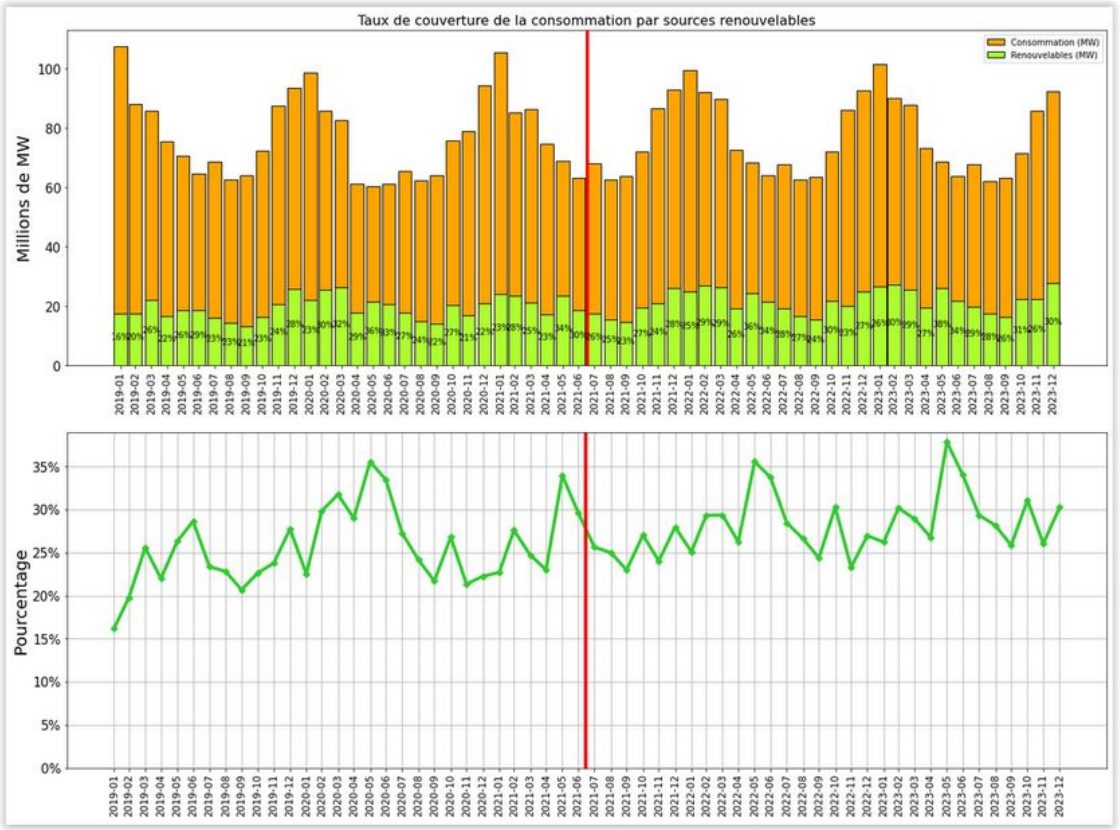
Ce modèle devient alors satisfaisant, que ce soit au niveau des p-values que des tests de Ljung-Box et Jarque-Bera. L'erreur moyenne relative représente 7.2% de la réalité, ce qui reste satisfaisant. C'est plus élevé que lors du Sarimax prenant toutes les données jusqu'à mi-2021 mais cela paraît normal car le modèle va être moins ajusté au choc exogène qu'a été le Covid. Dans notre cas, cela peut être même préférable.



Comme on peut le voir, le modèle sans les données Covid va effectuer des prévisions plus élevées que le modèle les prenant en compte. Considérant que le Covid est un choc exogène est en principe non permanent sur la série temporelle (car lié aux confinements successifs qui ne sont pas voués à se prolonger indéfiniment), nous allons conserver le modèle sans les données Covid.

3.1.3 Prévision du taux de couverture des énergies renouvelables

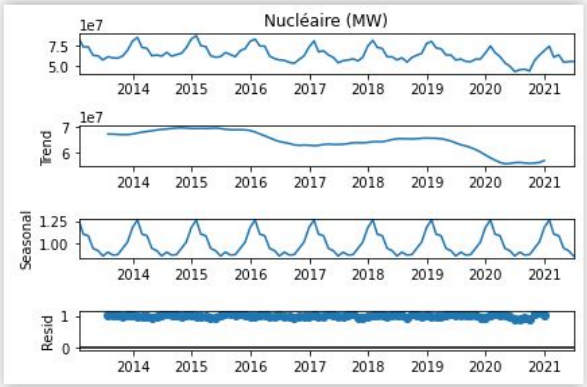
En combinant les 2 prévisions effectuées ci-dessus, nous pouvons donc produire un graphique équivalent à celui vu lors de l'analyse exploratoire pour prévoir l'évolution future du taux de couverture de la consommation par les renouvelables :



Le modèle prévoit une continuation de la tendance à la hausse de la couverture des renouvelables, d'une part grâce à une hausse régulière de la production de renouvelables, et d'autre part par une consommation qui va rester stable (malgré une remontée post-Covid).

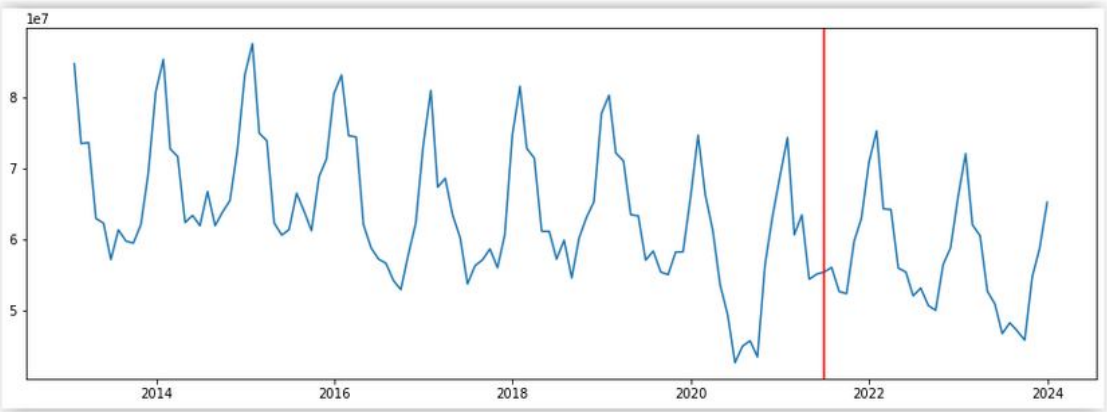
Prévision de la production nucléaire

Nous allons maintenant compléter nos prévisions avec des modèles pour les sources de production restantes en commençant par le nucléaire. Au niveau de la décomposition saisonnière logarithmique, on peut voir que la production nucléaire a connu un renversement de tendance à la baisse en 2016 puis en 2020, le dernier moment coincidant avec le début de la pandémie du Covid-19. On peut supposer que la production nucléaire est une variable d'ajustement à la consommation.



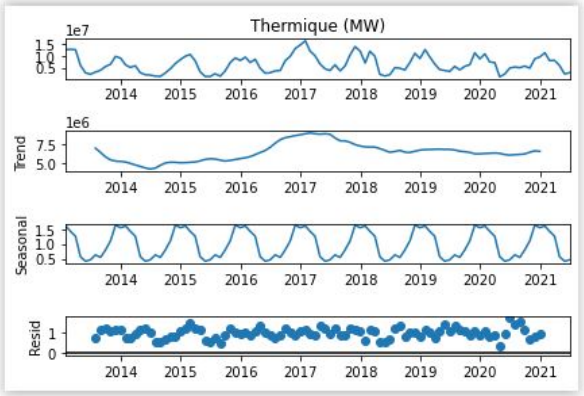
Nous effectuons une double différenciation et nous arrivons à une p-value du test de Dickey-Fuller de 0.088, une valeur légèrement trop élevée. Cependant, il y a une saisonnalité claire de 12 mois dans les données. Nous allons le prendre en compte dans les paramètres du modèle. Grâce aux autocorrélations, nous définissons un modèle SARIMAX "order=(1,1,1), seasonal_order=(2,1,0,12).

Au final, les résidus de notre modèle ne sont pas normalement répartis selon le test de Jarque-Bera. Cependant, leur blancheur est assurée par celui de Ljung-Box. Les résultats des tests d'erreurs sont quant à eux plutôt satisfaisants avec une erreur relative moyenne inférieure à 10% et un rmse ne représentant que 7% des observations. Voici les prévisions du modèle jusqu'à 2024. Nous allons conserver ces données dans notre analyse mais il faut garder en tête la p-value du test de Dickey-Fuller qui peut signaler une faiblesse dans la modélisation.



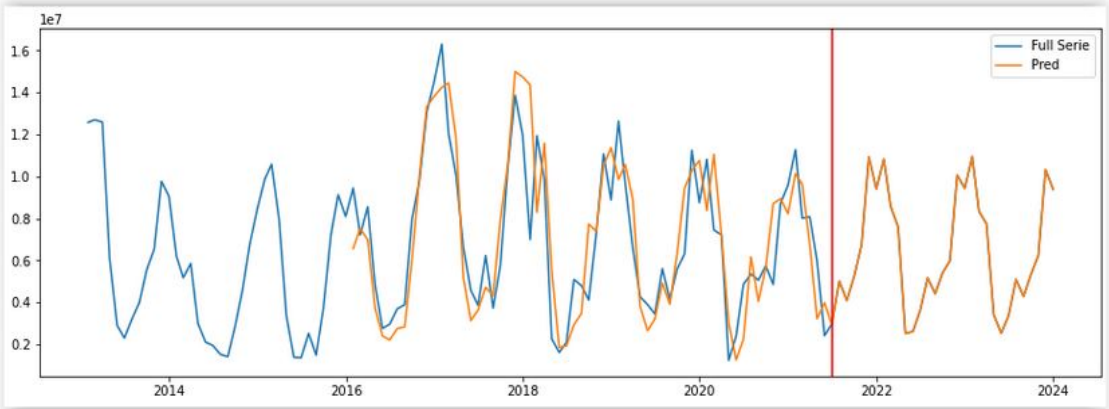
3.1.4 Prévision de la production de l'énergie thermique

Enfin, pour l'énergie thermique, nous avons cette décomposition avec échelle logarithmique :



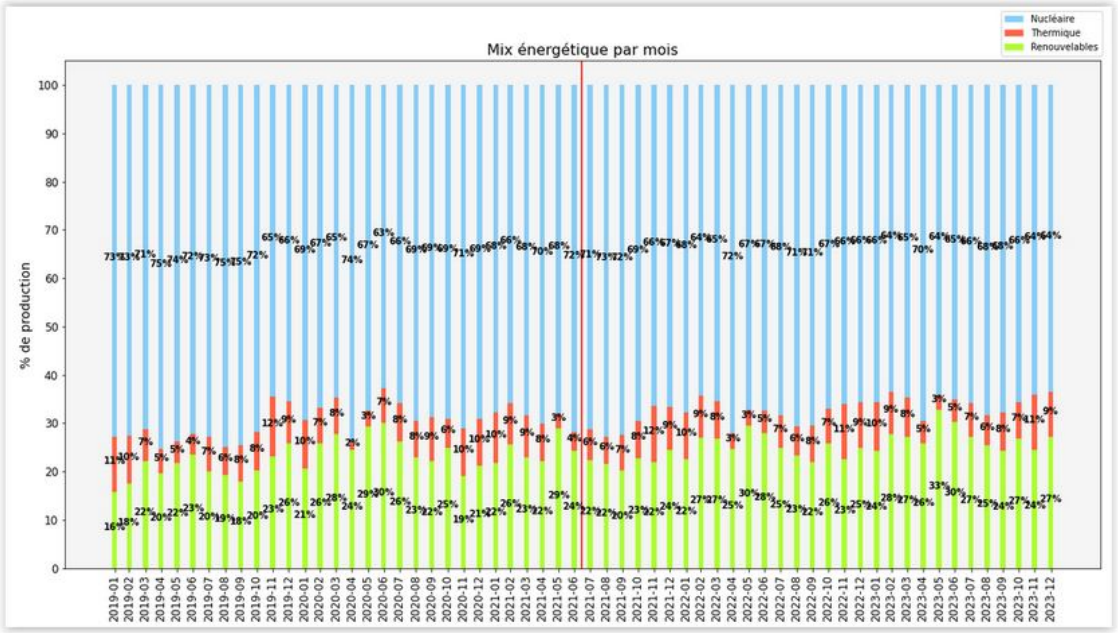
Les résidus sont regroupés ci-dessus autour de 1 mais ne sont pas vraiment stabilisés pour autant. Cependant, après des différenciations sur les pas de 1 puis 12, nous arrivons à une p-value du test Dickey-Fuller de 0.011. Il s'avère que la série est stationnaire selon ce test. Nous pouvons donc continuer le modèle en passant aux autocorrélations qui nous définissent un SARIMAX "order=(1,1,1), seasonal_order=(1,1,1,12)"

Au niveau des résultats du modèle, la normalité des résidus n'est pas assurée, leur blancheur oui. Nous arrivons par contre à un point difficile de ce modèle, les métriques d'erreur. Les résultats sont assez élevés (erreur moyenne relative de 36.6%) mais la courbe de prévision montre que la saisonnalité globale est respectée. Les erreurs sont dûes à l'aspect assez aléatoire de la production thermique, avec une saisonnalité moins rigoureuse. Le modèle paraît cohérent au niveau global.

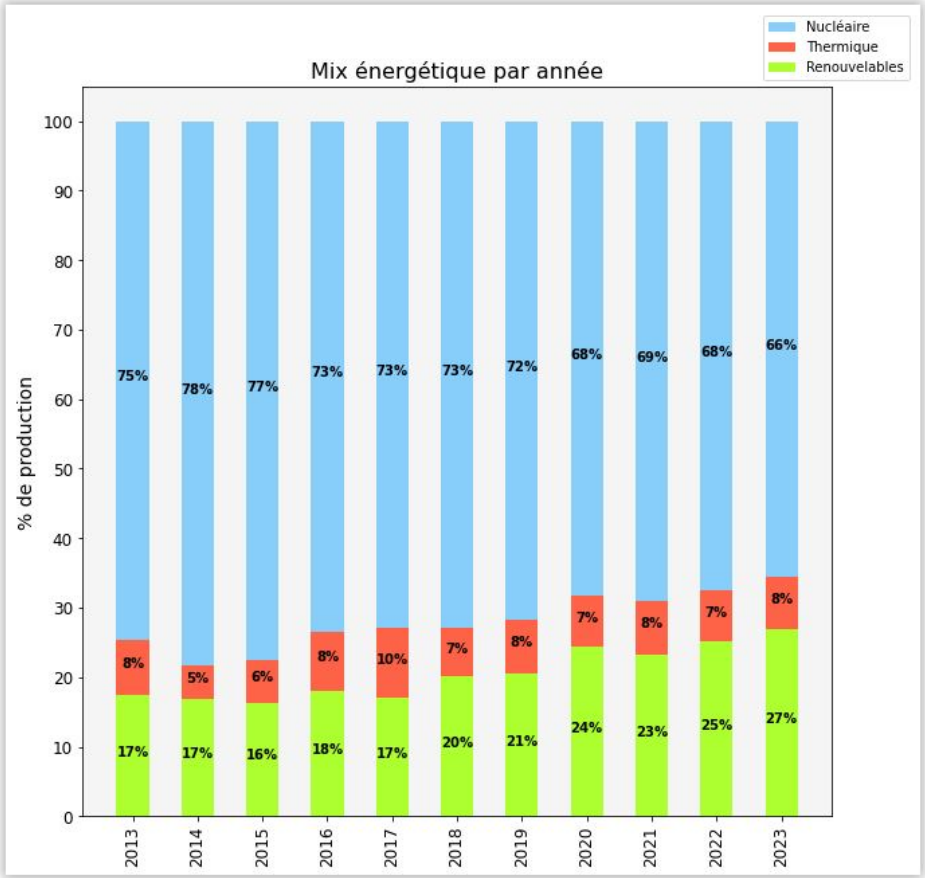


3.1.5 Visualisation des prévisions sur le mix énergétique français

Maintenant que nous disposons de l'ensemble des prévisions sur les productions, nous pouvons donc les visualiser pour anticiper l'évolution du mix énergétique français sur les prochaines années.



L'aspect le plus significatif de ces modèles est en agrégant les résultats des prévisions de manière annuelle. Nous gommons les erreurs et variations dans nos prévisions et arrivons à des prévisions plus interprétables :



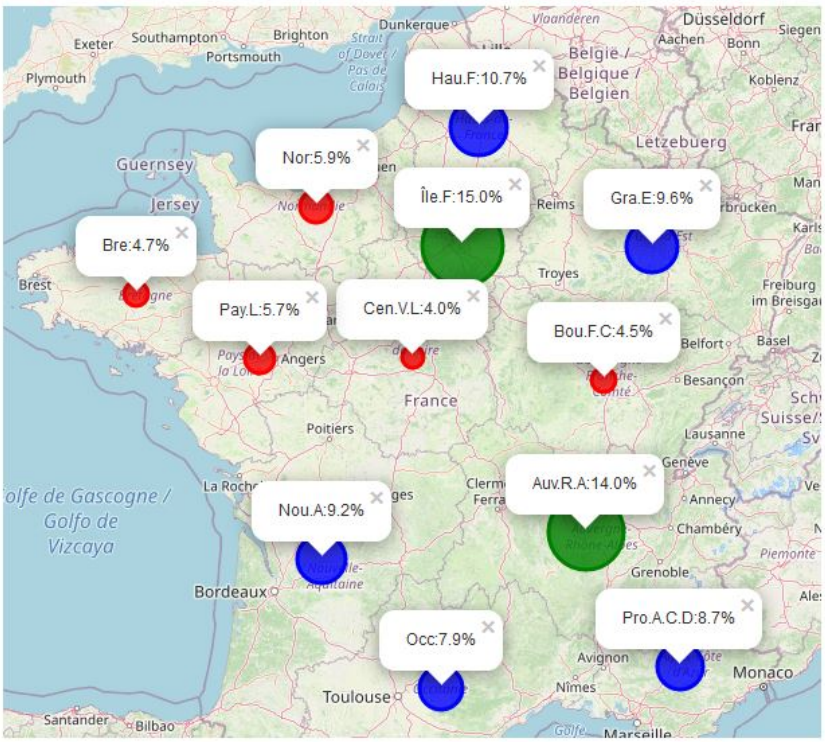
Nos modèles Sarimax prévoient une augmentation régulière de la part des renouvelables dans le mix énergétique français sur la base de la tendance des dernières années. Nous pouvons également remarquer que cela prend en compte le choc de la pandémie de 2020 : le nucléaire étant la variable d'ajustement à la consommation, il a baissé plus fortement en proportion. Nous devons garder en mémoire les limites de nos modèles mais la tendance agrégée au niveau de l'année reste pertinente et répond à une partie de la problématique exposée en introduction.

3.1.6 Classification des conso/région : répartition spatiale

A partir des différentes séries chronologiques scindées par régions, nous proposons une classification non supervisée. Les observations correspondent aux moyennes mensuelles. Dans un premier temps, nous utilisons la méthode du coude pour trouver une valeur optimale de k (nombre de cluster). Le résultat (que nous ne montrons pas dans ce rapport) montre que ce nombre est de 3.

Nous utilisons la métrique de déformation temporelle dynamique, en anglais Dynamic Time Warping (DTW). Cette métrique est plus appropriée pour l'étude des séries temporelles. Nous utilisons trois familles de couleurs (bleue, rouge, verte) pour représenter les trois classes de régions. Nous affichons également le pourcentage de la valeur cumulée de chaque région par rapporte à la valeur cumulée totale. Nous pouvons constater que les classes sont constituées suivant les ordres de grandeurs des valeurs cumulées : classe des valeurs inférieures (en rouge)– classe des valeurs intermédiaires (en bleu)– classe des valeurs supérieures (en vert). Une classification hiérarchique est également faite mais les résultats ne sont pas repris dans ce rapport.

A noter que le graphique correspond aux séries issues des Consommations en MW, nous présentons en annexe du document, les résultats pour les séries issues des productions par type d'énergie (Thermique, Nucléaire, Eolien, Solaire, Hydraulique, Bioénergies).



3.2 Approches Machine Learning

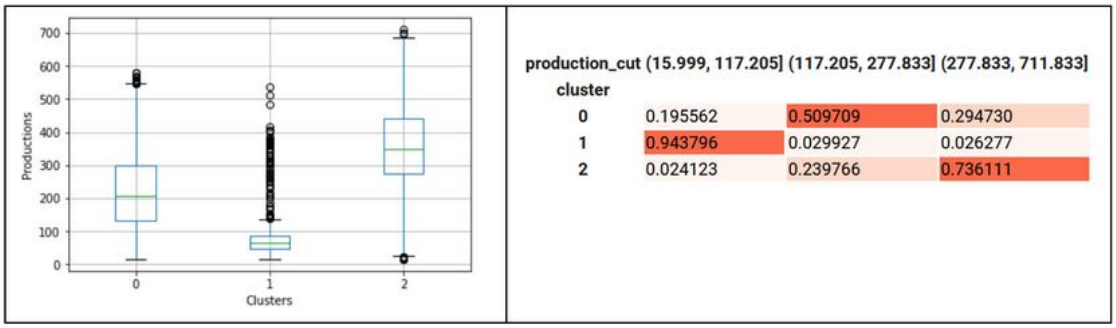
3.2.1 Prévision de la production de l'énergie solaire par classification non supervisée

Dans cette sous-section, nous effectuons un clustering non supervisé. Pour cela, nous disposons : (1) des données sur la production d'énergie solaire au pas de demi-heure, depuis le mois de janvier 2013 et ; (2) des données sur le rayonnement solaire global au pas de 3 heures (0h, 3h, 6h, 9h, 12h, 15h, 18h 21h), depuis le mois de janvier 2016. Ces données sont issues du site Open Data Réseaux Énergies (ODRÉ).

Les données sur le rayonnement solaire global sont des estimations à partir des points de grille du modèle de prévision déterministe du Centre Européen de Prévisions Météorologiques à Moyen Terme (CEPMMT). Comme indiqué sur le site ODRÉ, les moyennes de rayonnement solaire global ne sont pas pondérées en fonction de l'emplacement des parcs (éoliens ou solaires). Cela explique au moins en partie, les différences observées dans le temps, entre les productions associées à un même niveau de rayonnement solaire.

Nous allons organiser les valeurs des productions en 3 classes délimitées par les quantiles 33% et 66%, ce sont dont des classes de même effectif. Nous allons prévoir la classe de la production moyenne nationale pour un instant t donné. Les variables explicatives pour prédire la production à l'instant t sont (a) les valeurs des productions moyennes aux huit instants qui précèdent l'instant t, (b) les valeurs des rayonnements moyens aux huit instants qui précèdent l'instant t et, (c) la valeur moyenne du rayonnement à l'instant t.

Nous utilisons l'algorithme des k-means pour regrouper les productions solaires en 3 classes. Nous pouvons comparer les classes issues de l'algorithme au regroupement par tranche de productions que nous avons envisagé. Sur la figure de gauche, nous avons les boxplot des productions par clusters. Sur la figure de droite, nous avons la table des profils ligne par tranche de production d'énergie, pour chaque cluster.



cluster	production_cut (15.999, 117.205]	(117.205, 277.833]	(277.833, 711.833]
0	0.195562	0.509709	0.294730
1	0.943796	0.029927	0.026277
2	0.024123	0.239766	0.736111

Si chaque ligne (jeu de test) est affectée à la classe dominante du cluster alors, nous avons :

- Accuracy : 0.6631
- Recall : 0.7272
- Score F1 : 0.6731

On constate que les trois clusters proposés par l'algorithme correspondent plus ou moins à nos classes de productions : le cluster 0 correspond à la classe inférieure, le cluster 1 correspond à la classe supérieure et le cluster 2 correspond à la classe intermédiaire. Si nous classons chaque cluster selon la classe dominante, nous abstenons un score de 66%, ce qui est tout à fait satisfaisant.

3.2.2 Prévision de la production de l'énergie solaire par classification supervisée

Les données sont celle présentées dans la sous-section précédente. Nous avons effectué trois types de Classification supervisée : K plus Proches Voisins - Forêt Aléatoire pour la Classification - Multi Layer Perceptron Classifier. La variable cible c'est la classe de production de l'énergie solaire et les variables explicatives sont les productions des 7 derniers jours avant et les prévisions du rayonnement solaire.

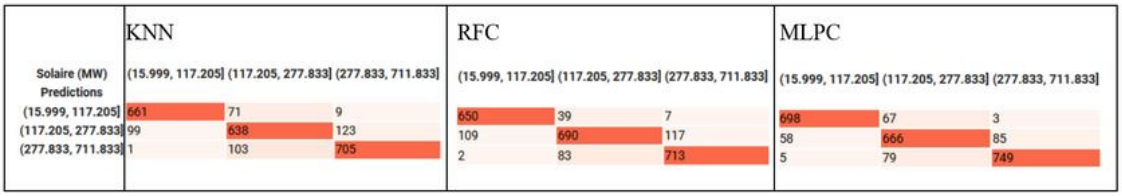
Les scores sont données ci-dessous :

Plus proches voisins	Random Forest Classifier	Multi Layer Perceptron Classifier
- Accuracy : 0.8315	- Accuracy : 0.8519	- Accuracy : 0.8768
- Recall : 0.8351	- Recall : 0.8602	- Recall : 0.8771
- Score F1 : 0.8333	- Score F1 : 0.8543	- Score F1 : 0.8772

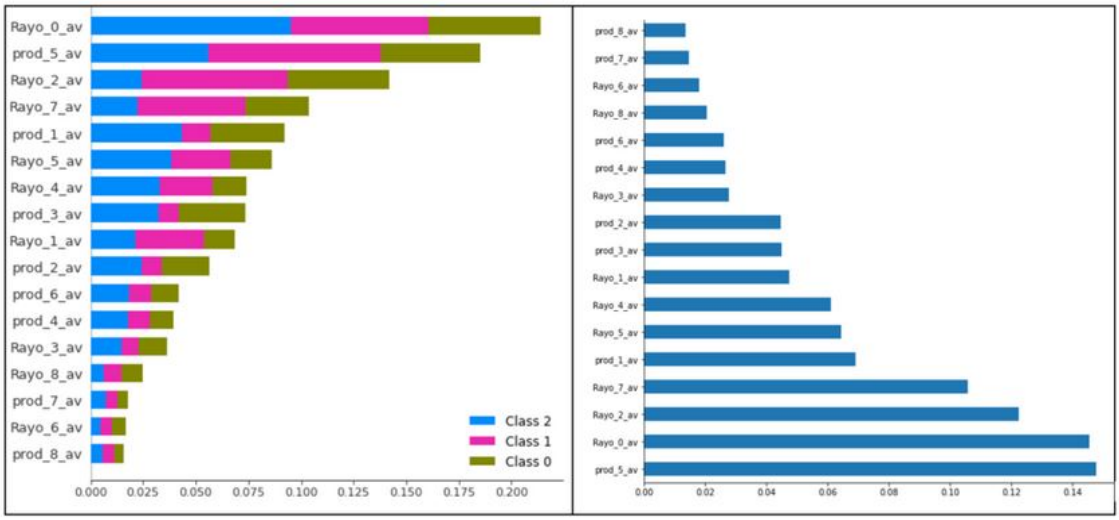
Afin de s'assurer que les performances de nos modèles ne sont pas "dues à la chance", nous allons effectué une cross-validation avec 5 folds, les résultats sont les suivantes :

- Accuracy KNN : 0.82 (+/- 0.04)
- Accuracy RFC : 0.84 (+/- 0.04)
- Accuracy MLPC : 0.86 (+/- 0.03)

Ci-dessous, nous affichons les différentes matrices de confusion.



Interprétabilité



3.2.3 Prévision de la production de l'énergie solaire par régression

Pour ce qui concerne la modélisation par régression nous avons effectué trois types de regression : Régression LASSO, Random Forest Regression et Multi Layer Perceptron Regression. Les scores sont données ci-dessous :

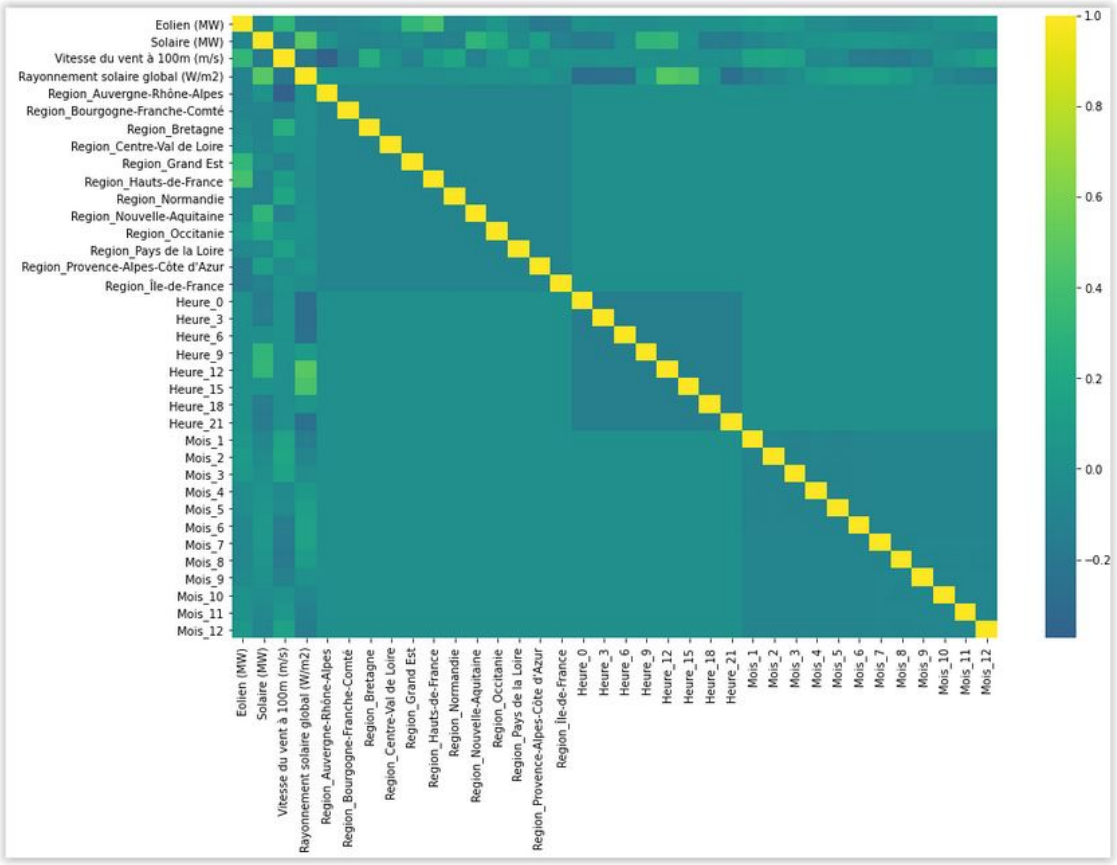
LASSO	Random Forest Regression	Multi Layer Perceptron Regression
- R2 : 0.8203 - RMSE : 55.8127	- R2 : 0.9014 - RMSE : 41.9146	- R2 : 0.9315 - RMSE : 36.2606

Cross-validation

LASSO	Random Forest Regression	Multi Layer Perceptron Regression
- R2 : 0.81 (+/- 0.08) - RMSE : 60.02 (+/- 14.11)	- R2 : 0.88 (+/- 0.04) - RMSE : 47.12 (+/- 12.14)6	- R2 : 0.92 (+/- 0.04) - RMSE : 38.19 (+/- 6.41)

3.2.4 Prévision de la production de l'énergie éolienne par régression

Le but est de prévoir grâce à des régressions la production d'énergie éolienne en MW par région et par intervalle de 3 heures. Cet intervalle provient des données que nous allons ajouter au modèle, qui sont les données de vitesse du vent et de rayonnement solaire régional depuis 2016. En tant que variables explicatives, nous avons donc les données météorologiques citées mais également la région, l'heure ainsi que le mois de l'année, ces 3 variables ayant été encodées car catégorielles. Le tableau des corrélations nous permet de vérifier quelques hypothèses essentielles pour nos modèles :



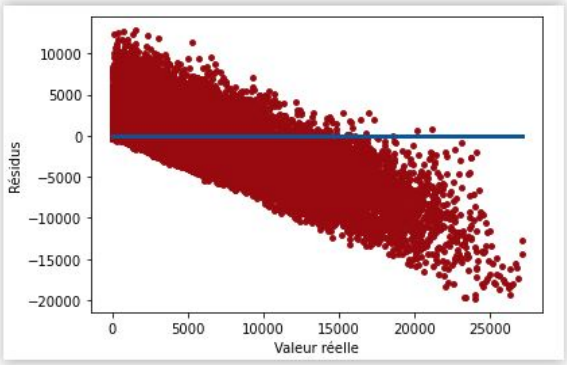
- la vitesse du vent est corrélée avec la production éolienne
- certaines régions sont plus propices à la production éolienne que d'autres
- l'heure n'a pas d'influence sur la production éolienne
- le mois de l'année peut lui éventuellement jouer en hiver : on a une corrélation entre vitesse du vent et mois d'hiver

Accessoirement, on remarque que les régions avec les corrélations avec la vitesse du vent constatée les plus élevées ne sont pas les régions où il y a le plus de production éolienne, ce qui fait émerger des questions sur la localisation du parc éolien français.

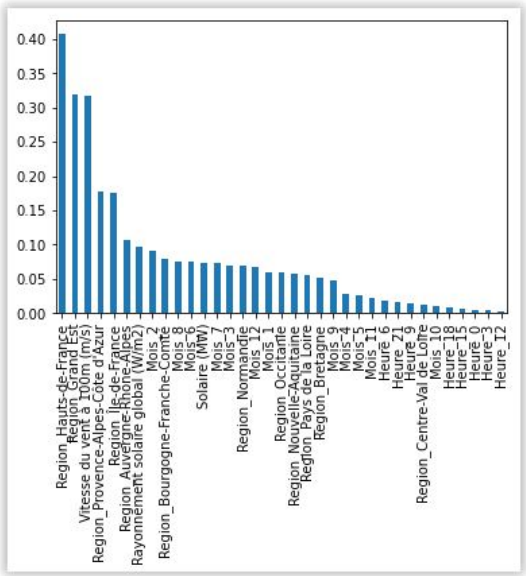
Après avoir créé une base d'entraînement et de test (split 80/20), nous allons alors tester de multiples modèles avec des grilles de validations croisées pour certains et voici leurs scores :

Model	R^2 train	R^2 test
Rég. Linéaire	0.454	0.450
Rég. Polynomiale Degré 2	0.583	0.578
RidgeCV	0.454	0.450
LassoCV	0.454	0.450
ElasticNetCV / SelectKBest (k=16)	0.441	0.436
XGBRegressor	0.617	0.586

Le meilleur modèle est le XGBRegressor avec des paramètres : {'learning_rate' : 0.07, 'max_depth' : 5}



Conclusion de ces régressions : le modèle XGB donne les meilleurs résultats parmi tous les modèles testés (même si la régression polynomiale donne des résultats similaires). Cependant, ce n'est pas optimal car les résidus ne sont pas aléatoires (croissants avec les valeurs réelles) et il y a dispersion des résultats. Le modèle prévoit mal les grandes valeurs. Pourquoi ? Si nous reprenons les corrélations à l'origine de ce modèle, nous nous rendons compte que peu de variables sont réellement significatives : seulement 2 régions ainsi que la vitesse du vent ont une corrélation supérieure à 0.20.



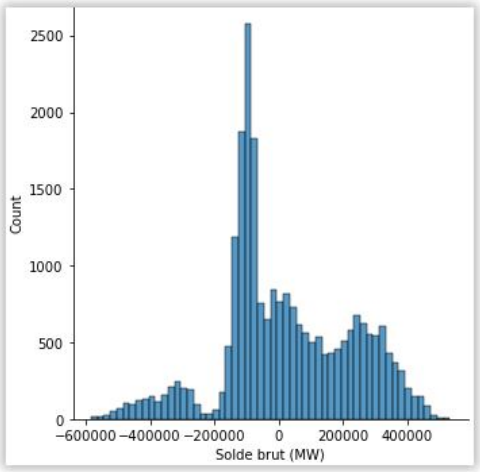
En repartant de l'analyse exploratoire, rappelons-nous que seulement 2 régions étaient productrices d'énergie éolienne de manière importante (les Hauts-de-France et le Grand-Est)

et ceci seulement en hiver. Nous disions également que la vitesse du vent n'avait pas une corrélation claire avec la production éolienne car les régions avec les vents les plus importants ne produisaient pas d'énergie éolienne. La conclusion de cette analyse est que cela trouble toute corrélation potentielle et donne des régressions inefficaces. Nous retombons donc sur la question du placement des éoliennes en France.

3.2.5 Classification des occurrences de déficits et d'excédents

Ce dernier modèle va chercher à utiliser des techniques de classification pour identifier les occurrences régionales de déficits et d'excédents par niveau. Un modèle réussi permettrait d'identifier les jours avec des déficits/excédents importants ainsi que les variables déterminants ce résultat. Le preprocessing va être de cumuler plusieurs databases : les données électriques mais également les données météorologiques vues précédemment ainsi que les données de température. Ces dernières données étant quotidiennes, nous allons agréger nos données par jour depuis 2016.

En premier lieu, il nous faut discrétiser la variable cible, le solde brut quotidien. En observant sa distribution, il semble correct d'avoir des buckets tous les 200000 MW pour obtenir 6 classes. Nous aurons dans l'ordre croissant : Déficit3, Déficit2, Déficit1, Excédent1, Excédent2, Excédent3. Les classes extrêmes sont donc les classes nommées 3.



Après le split train/test ainsi que l'application d'un Standard Scaler, nous passons aux modèles, du plus simple au plus complexe. Nous commençons par la création d'une grille de

validation croisée pour évaluer une régression logistique et une random forest. La régression a un score moyen de 82.13, la Random Forest de 84.66. Nous la conservons, voici le score de la Random Forest optimale :

Best params : {'max_features' : 'log2', 'min_samples_leaf' : 5, 'n_estimators' : 100} ; Accuracy train : 90.36 ; Accuracy test : 85.07

Et le classification report ainsi que les variables plus importantes du modèle :

	precision	recall	f1-score	support
Déficit1	0.93	0.93	0.93	2086
Déficit2	0.95	0.91	0.93	270
Déficit3	0.87	0.90	0.88	131
Excédent1	0.78	0.71	0.74	1200
Excédent2	0.77	0.89	0.82	1033
Excédent3	0.58	0.19	0.28	97
accuracy			0.85	4817
macro avg	0.81	0.75	0.76	4817
weighted avg	0.85	0.85	0.85	4817

	Importance
Region_Île-de-France	0.14
Region_Centre-Val de Loire	0.11
Region_Nouvelle-Aquitaine	0.09
Vitesse du vent à 100m (m/s)	0.09
Region_Grand Est	0.09
Region_Normandie	0.08
tmoy	0.07
Region_Auvergne-Rhône-Alpes	0.07
Rayonnement solaire global (W/m2)	0.06
Region_Bourgogne-Franche-Comté	0.05

Il apparaît clairement que la région ainsi que les conditions météorologiques sont les facteurs principaux. On peut remarquer que ce modèle est déjà satisfaisant pour repérer les déficits, même dans les cas les plus graves qui ont pourtant peu d'occurrences. Si nous restons sur une problématique d'identification des risques de blackout, nous pourrions considérer ce modèle comme efficace. Cependant, ce modèle ne fonctionne pas sur l'identification des excédents importants. Nous allons donc continuer à tester d'autres modèles.

Un modèle de Bagging avec une PCA à 0.9 (réduction d'un quart du nombre de variables) nous fait apparaître un overfitting et nous l'abandonnons vite, un XGBClassifier avec une grille de validation croisée donne lui des résultats très similaires à la Random Forest.

Best params : {'gamma' : 0.1, 'learning_rate' : 0.2, 'max_depth' : 6, 'n_estimators' : 50} ; Score train : 89.47 ; Score test : 84.47

Avec une analyse poussée des résultats, nous faisons toujours le même constat : les cas extrêmes d'excédent sont difficilement identifiés. Cependant, il y a toujours une bonne qualité de prédiction sur les déficits. Pour remédier à tout cela, il apparaît opportun de procéder à un rééquilibrage des données avec un RandomOverSampler pour accroître le pool des possibilités, chaque classe ayant 8499 occurrences.

Random Forest : Best params : {'max_features' : 'sqrt', 'min_samples_leaf' : 1, 'n_estimators' : 50} ; Score train : 99.99 ; Score test : 84.58

XGB : Best params : {'gamma' : 0.1, 'learning_rate' : 0.2, 'max_depth' : 10, 'n_estimators' : 50} ; Score train : 95.32 ; Score test : 83.39

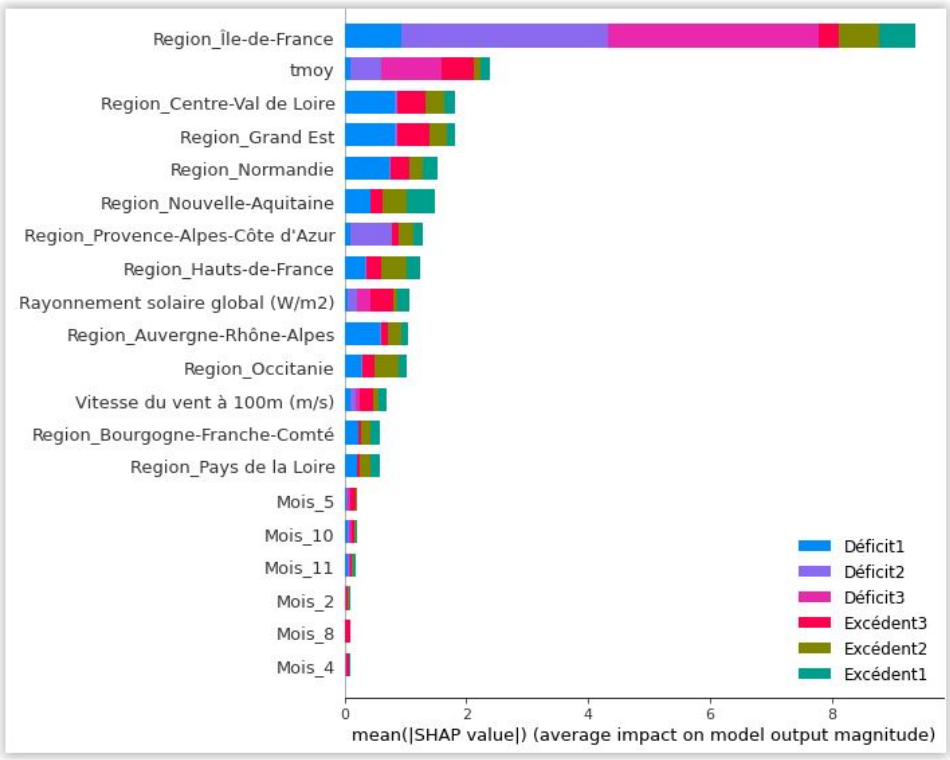
Les résultats ne se sont pas substantiellement améliorés avec en outre l'apparition d'overfitting. Regardons les classification report du modèle XGB qui a le surapprentissage le plus faible.

	precision	recall	f1-score	support
Déficit1	0.95	0.90	0.92	2086
Déficit2	0.91	0.91	0.91	270
Déficit3	0.87	0.89	0.88	131
Excédent1	0.74	0.76	0.75	1200
Excédent2	0.79	0.77	0.78	1033
Excédent3	0.38	0.78	0.51	97
accuracy			0.83	4817
macro avg	0.77	0.83	0.79	4817
weighted avg	0.85	0.83	0.84	4817

Prédit	Déficit1	Déficit2	Déficit3	Excédent1	Excédent2	Excédent3
Réel						
Déficit3	0	15	116	0	0	0
Déficit2	6	246	18	0	0	0
Déficit1	1871	8	0	206	1	0
Excédent1	97	0	0	913	189	1
Excédent2	0	0	0	114	795	124
Excédent3	0	0	0	0	21	76

Ces derniers tableau nous font apparaître un gain réel de ce dernier modèle : les recall sont tous supérieurs à 75%, la classification de la catégorie Excédent 3 s'est améliorée considérablement par rapport aux autres modèles, la contrepartie étant de perdre un peu de

qualité dans la classification de la catégorie Excédent2. Nous pouvons conserver ce modèle en tant que meilleur modèle car fonctionnant le mieux sur toutes les sous-catégories et notamment les plus extrêmes, qui sont la cible de notre analyse. Utilisons les shap_values pour l'interprétation de ce modèle en extrayant les variables déterminantes dans la classification :



- La région Ile-de-France est déterminante dans le classement. En effet, étant souvent en déficit, elle effectue un premier tri essentiel parmi les données.
- Les régions dans leur ensemble sont les principaux facteurs de classification.
- Les variables météorologiques ont un impact, et plus spécifiquement la température moyenne qui est le second facteur de classification.

De manière plus détaillée, les déficits extrêmes sont déterminés principalement par la région Ile-de-France qui concentre ce type de situations. La température quant à elle va avoir une relation importante et inverse : plus il fait chaud, moins on aura de chances d'être en déficit électrique. Les autres variabels sont quasi négligeables (y compris les autres régions). Au final, ce modèle fonctionne assez efficacement pour déterminer les occurrences de cas extrêmes dans les soldes électriques. Il pourrait être amélioré en trouvant d'autres sources de données explicatives mais en l'état, il constitue une base de travail raisonnablement fiable.

Chapitre quatrième

Difficultés, conclusion et perspectives

4.1 Difficultés rencontrées

Au niveau des difficultés communes à tous nos modèles, il est apparu au fur et à mesure des modélisations qu'il existe d'autres variables expliquant le marché électrique français et dont nous ne disposons pas : par exemple, les productions solaires ou électriques dépendent autant de la capacité des centrales que de l'ensoleillement ou de la vitesse du vent. Or nous ne pouvons que supposer où elles se situent et cela apparait dans les conclusions de nos modélisations. De même, la finesse de certaines données météorologiques est parfois insuffisante pour être déterminante.

En ce qui concerne les difficultés spécifiques, nous pouvons citer la crise du COVID-19 qui trouble fortement la qualité des modélisations de séries temporelles. Cette rupture exogène ne peut que mettre en doute la qualité des prévisions basées sur la tendance passée. Nous avons essayé de le prendre en compte.

4.2 Conclusion

Nous avons testé dans le cadre de ce projet de multiples types de modélisations, certaines donnant des résultats concluants. Pour répondre clairement aux enjeux initiaux du projet :

- il est possible de créer un modèle satisfaisant pour classifier et identifier les situations potentielles de tension (blackout) au niveau régional
- les modèles confirment la tendance à la hausse de la production d'énergies renouvelables mais le futur sera déterminé par le maintien au non des évolutions causées en partie

par la rupture du Covid-19
- les saisonnalités et différences régionales sont confirmées par les modèles, ce qui peut servir à clarifier et orienter le futur des politiques régionales
- le nucléaire et le thermique sont bien les variables d'ajustement à la consommation électrique française, leur baisse dans le mix énergétique dépend de la montée en puissance du renouvelable et du contrôle de la consommation future

4.3 Perspectives

Là où pour certains modèles nous avons conclu à la faible qualité de la modélisation (régression sur l'éolien par exemple), d'autres modèles peuvent évoluer et conserver une certaine pertinence dans l'avenir. Les séries temporelles peuvent être utiles pour des confirmations de tendance sur le moyen terme, le clustering régional est pertinent dans la mise en évidence des disparités locales, la classification des situations de tension est d'une qualité suffisante pour être poursuivie.

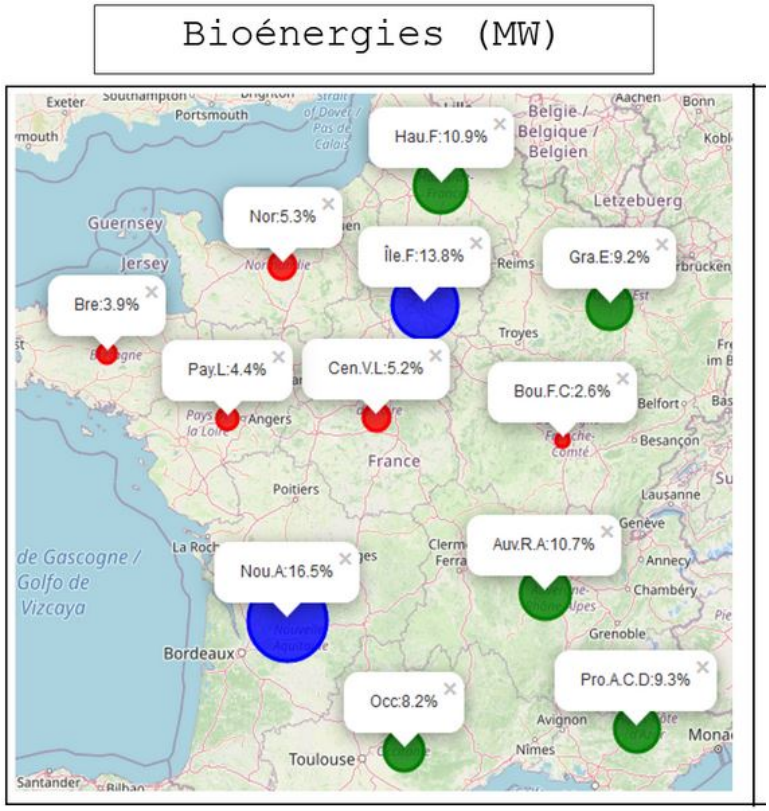
Il faudra certainement enrichir la base avec de nouvelles variables et avoir un niveau de détail supérieur pour améliorer les modélisations dans l'avenir.

Chapitre cinquième

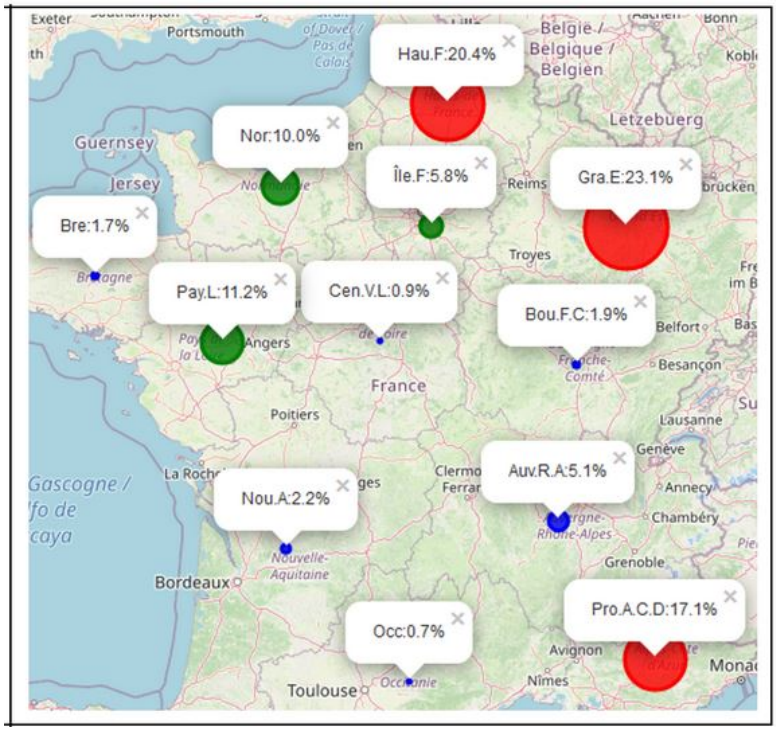
Annexe

5.1 Synthèse du Clustering régions par type d'énergie

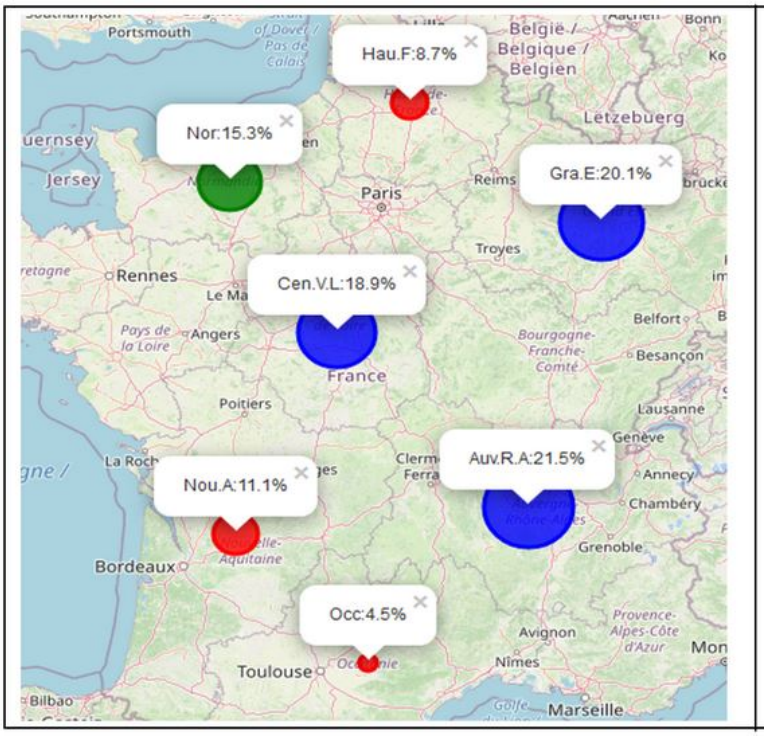
Classification des séries issues des différentes productions par région : répartition spatiale. Nous apportons ainsi une réponse concrète à une importante préoccupation clairement spécifiée dans les spécifications du projet : il s'agit de faire une analyse par filière de production en faisant ressortir les lieux d'implantation des sites de production. Les graphiques sont assez clairs et précis.



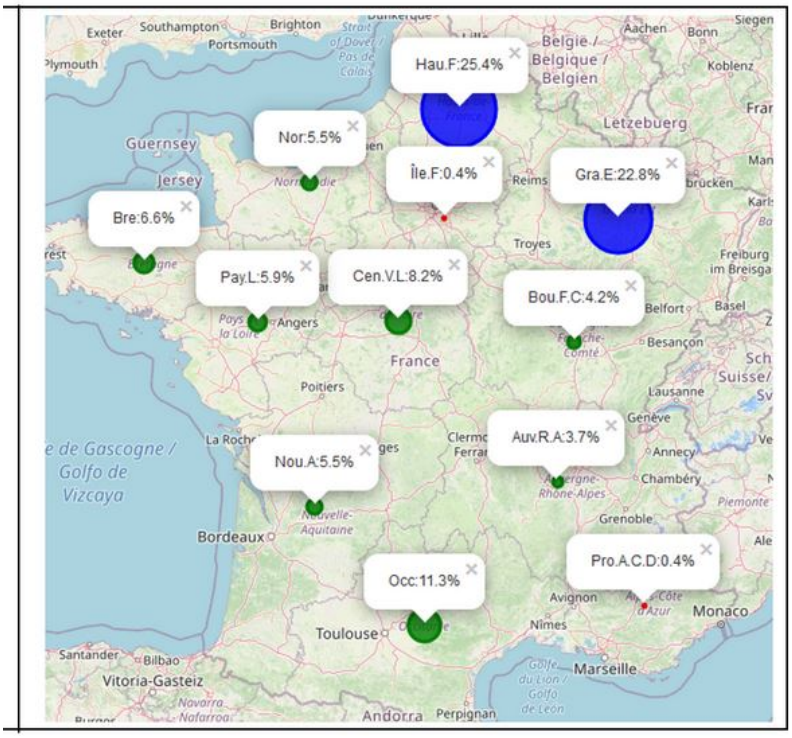
Thermique (MW)



Nucléaire (MW)



Eolien (MW)



Solaire (MW)

Hydraulique (MW)