PROJET PYNERGY

Cursus : Bootcamp Data Scientist, Juillet 2021

Mené par : Romain MOULY et Zéphirin NGANMENI

Sous la supervision de : Mounir

DataScientest

Plan de la présentation

01

Introduction, problématique et objectifs

02

Prétraitement et analyse exploratoire

03

Modélisation et résultats

04

Conclusion, difficultés et perspectives

Introduction, problématique et objectifs

Introduction (1/2)

L'énergie est une ressource incontournable pour le bon fonctionnement des équipements utilisés dans divers secteurs socio-économiques :

- Ménages
- Industrie
- Transports
- Agriculture
- Etc.

Introduction, problématique et objectifs

Introduction (2/2)

La demande et offre dépendent de plusieurs facteurs:

- Marché ouvert à l'international
- Variété des sources d'énergies : Eolien, Solaire, Hydraulique, Thermique, Nucléaire, Bioénergies
- Besoin de limiter les impacts environnementaux

Conséquence : variation de la demande et de l'offre, ce qui conduit à des tensions qu'il faut contenir.

- Techniques : mesures de régulation.
- Economiques : adaptation des prix.

Introduction, problématique et objectifs

Problématique

Comprendre et **anticiper** les **évolutions** de la demande et de l'offre en énergie électrique avec un focus sur les **énergies renouvelables**.

Introduction, problématique et objectifs

Objectifs spécifiques

- Constater le phasage entre la consommation et la production énergétique française (risque de black out notamment).

  => ***Classification.***

- Analyse au niveau départemental et prévision de consommation.

  => ***Séries temporelles.***

- Analyse par filière de production : énergie nucléaire / renouvelable.

  => ***Séries temporelles, régressions***

- Focus sur les énergies renouvelables (lieu d'implantation).

  => ***Clustering.***

Prétraitement et analyse exploratoire
Données

Source de données principale: ODRE (Open Data Réseaux Energies).

- Informations régionales de consommation et production par filière jour par jour (toutes les 1/2 heure) depuis 2013.
- Données météorologiques régionales de l'ODRE depuis 2016 :
  - ✓ ***Températures*** *min, max, moy (quotidiennes);*
  - ✓ ***Vitesse du vent*** *et* ***ensoleillement*** *(fréquence de 3 heures).*

Prétraitement et analyse exploratoire
Données

Données brutes à disposition (échantillon):

Code INSEE région	Région	Nature	Date	Heure	Date - Heure	Consommation (MW)	Thermique (MW)	Nucléaire (MW)	Eolien (MW)	Solaire (MW)	Hydraulique (MW)	Pompage (MW)	Bioénergies (MW)	Ech. physiques (MW)	Flux physiques d'Auvergne-Rhône-Alpes vers Grand-Est	Flux physiques de Bourgogne-Franche-Comté vers Grand-Est	Flux physiques de Bretagne vers Grand-Est	Flux physiques de Centre-Val de Loire vers Grand-Est	Flux physiques de Grand-Est vers Grand-Est	Flux physiques de Hauts-de-France vers Grand-Est	Flux physiques d'Ile-de-France vers Grand-Est
93	Provence-Alpes-Côte d'Azur	Données consolidées	2021-06-30	23:30	2021-06-30T23:30:00+02:00	4623.0	380.0	NaN	30.0	0.0	1039.0	0.0	75.0	3101.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
52	Pays de la Loire	Données consolidées	2021-06-30	23:30	2021-06-30T23:30:00+02:00	2699.0	-12.0	NaN	20.0	0.0	3.0	NaN	52.0	2636.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
53	Bretagne	Données consolidées	2021-06-30	23:30	2021-06-30T23:30:00+02:00	2306.0	6.0	NaN	6.0	0.0	4.0	-1.0	52.0	2240.0	NaN	NaN	-	NaN	NaN	NaN	NaN

```
Nombre de lignes total = 1787328
```

Prétraitement et analyse exploratoire

Traitement ISNA

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
Flux physiques d'Auvergne-Rhône-Alpes vers Grand-Est	1761072
Flux physiques de Bourgogne-Franche-Comté vers Grand-Est	1734816
Flux physiques de Bretagne vers Grand-Est	1734816

- Suppression des colonnes avec trop de N/A ou peu significatives (Flux physiques…)
- Remplacement des N/A des données électriques par 0
- Suppression des 12 lignes sans aucune donnée

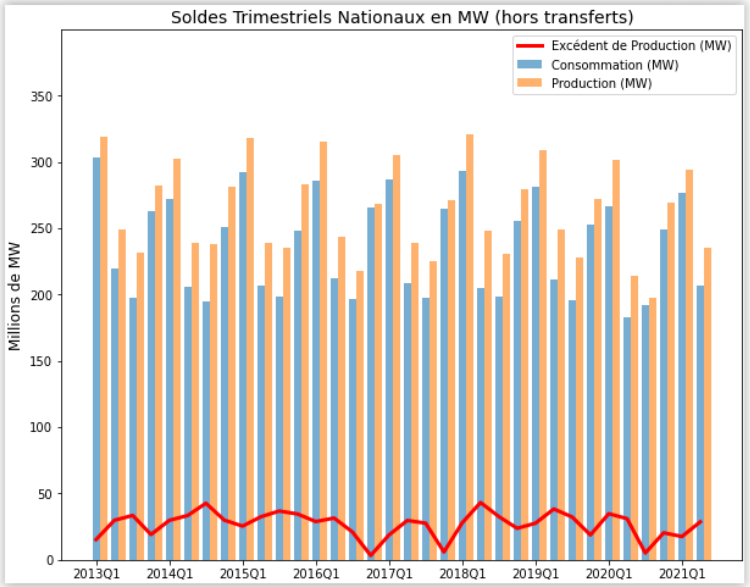
Prétraitement et analyse exploratoire
Feature Engineering

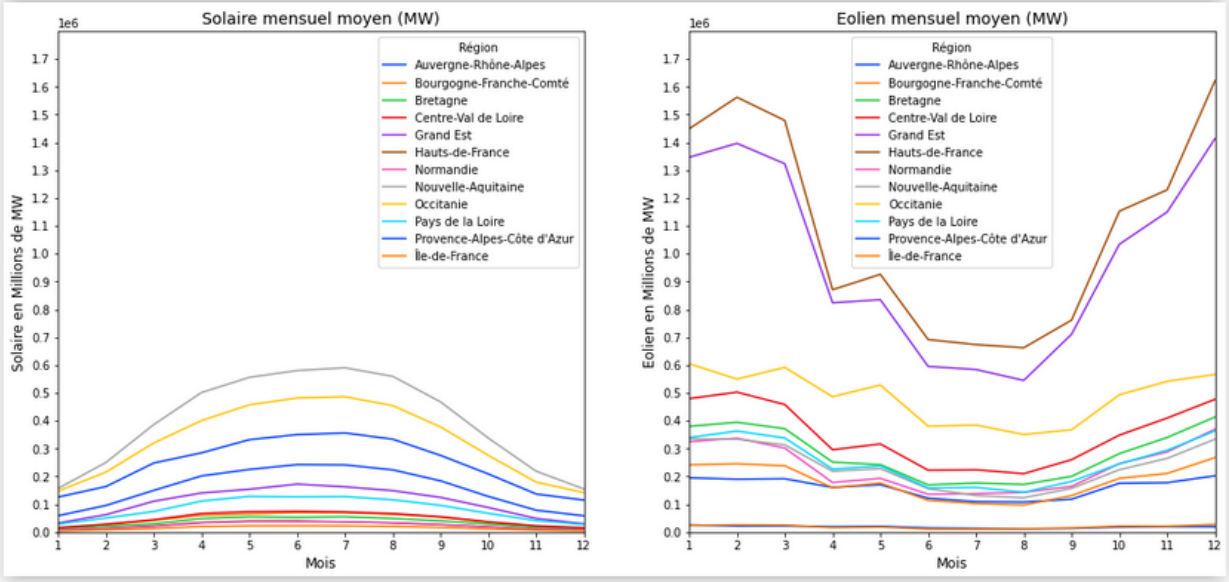
- Création de variables **synthétisant** l'information:
  - ✓ **Solde brut** = production – consommation
  - ✓ **Solde avec transferts** = production – consommation + transferts
  - ✓ **Total des renouvelables** = Eolien + Solaire + Bioénergies + Hydraulique + Pompage
- Création de variables extrayant les données **chronologiques** de chaque ligne :
  - ✓ *Date, Weekday, Jour, Mois, Trimestre, Année, Heure*

Prétraitement et analyse exploratoire

Analyse Exploratoire

- Analyses au niveau national
- Analyses au niveau régional





- Avec des échelles temporelles différentes

Prétraitement et analyse exploratoire

Analyse Exploratoire: insights

- **Saisonnalité** des consommations et productions d'énergie.
- Saisonnalité moins présente en détaillant par sources d'énergie du fait des **disparités régionales**.
- Régions = déterminants importants avec producteurs nets (Centre-Val de Loire, Grand-Est) et consommateurs nets (PACA, Ile-de-France).
- Au niveau national, France = **excédentaire** en énergie, différent au niveau régional.
- **Renouvelables en hausse** progressive dans le mix énergétique français, accélération sur les dernières années.

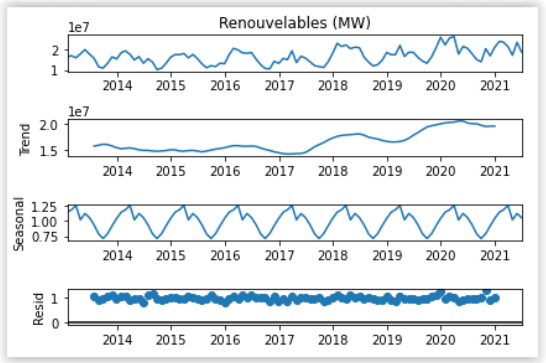
Modèle Séries Temporelles

- **Objectif**: Quelle va être l'évolution du mix énergétique français ?
- **Méthode**: 4 modèles **SARIMAX** différents sur une base quotidienne :
  - ✓ Renouvelables
  - ✓ Nucléaire
  - ✓ Thermique
  - ✓ Consommation

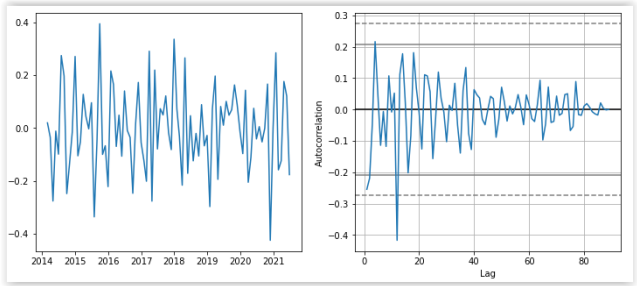
- Processus identique pour chaque modèle

Modèle Séries Temporelles

Workflow

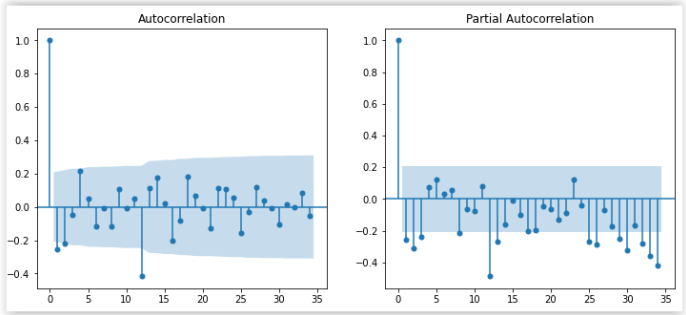


Décomposition

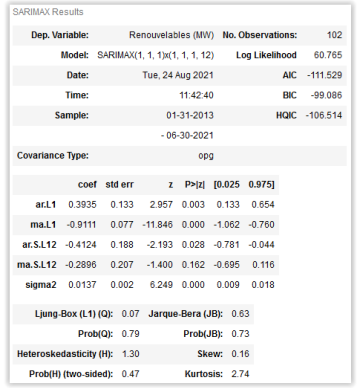


```
P-value AdFuller: 8.597024412577399e-06
```

Différenciation



Autocorrélation



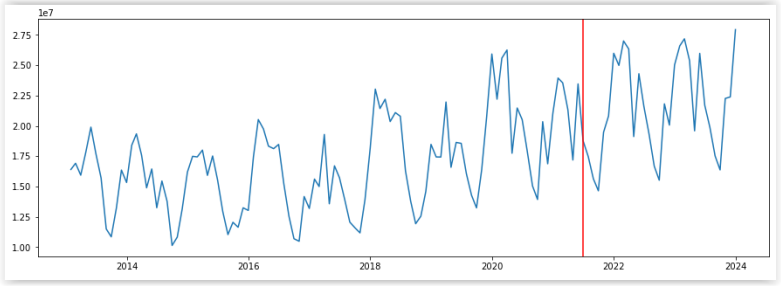
Modélisation

```
pred_20_21 = np.exp(model_fitted.predict(84,101))
real_20_21 = time_serie.loc['2020-1-31':'2021-06-30']

num = np.sum(np.abs(pred_20_21-real_20_21))/12
denom = np.mean(real_20_21)
error = num/denom
print(f'Erreur moyenne relative: {"{:.1f}".format(error*100)}%')
```

```
Erreur moyenne relative: 16.7%
```

Evaluation



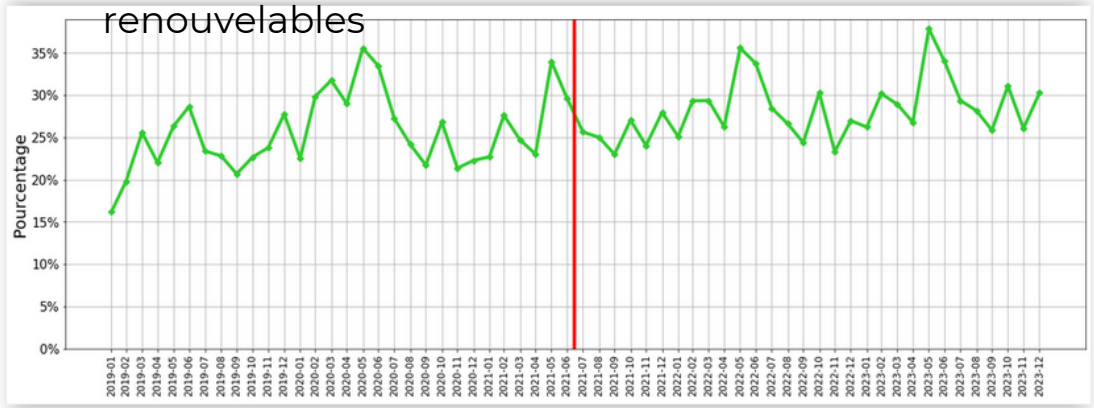
Prévision

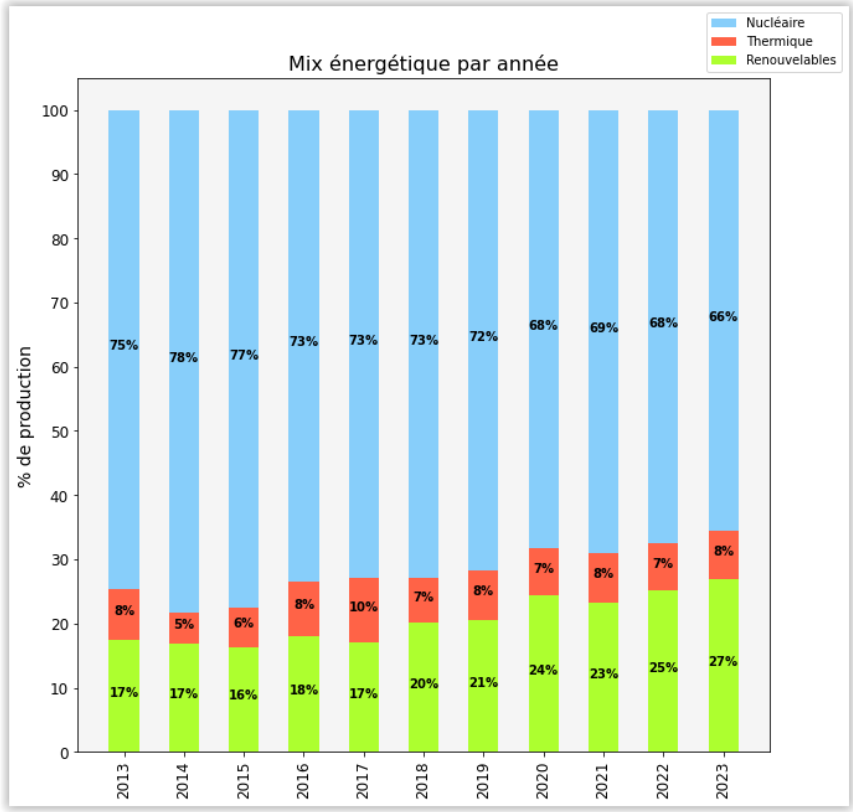
DataScientest

Modèle Séries Temporelles
Conclusions

- Challenge: rupture exogène du Covid-19 => **incertitudes**
- Mais modèles pour l'instant **cohérents** en terme de score et de tendance
- **Insights**:
  - ✓ Saisonnalité confirmée
  - ✓ Hausse prévue de la part des renouvelables
  - ✓ Nucléaire = variable d'ajustement à la consomma

Taux de couverture de la conso par les renouvelables





DataScientest

Modèle Clustering: Relations entre régions

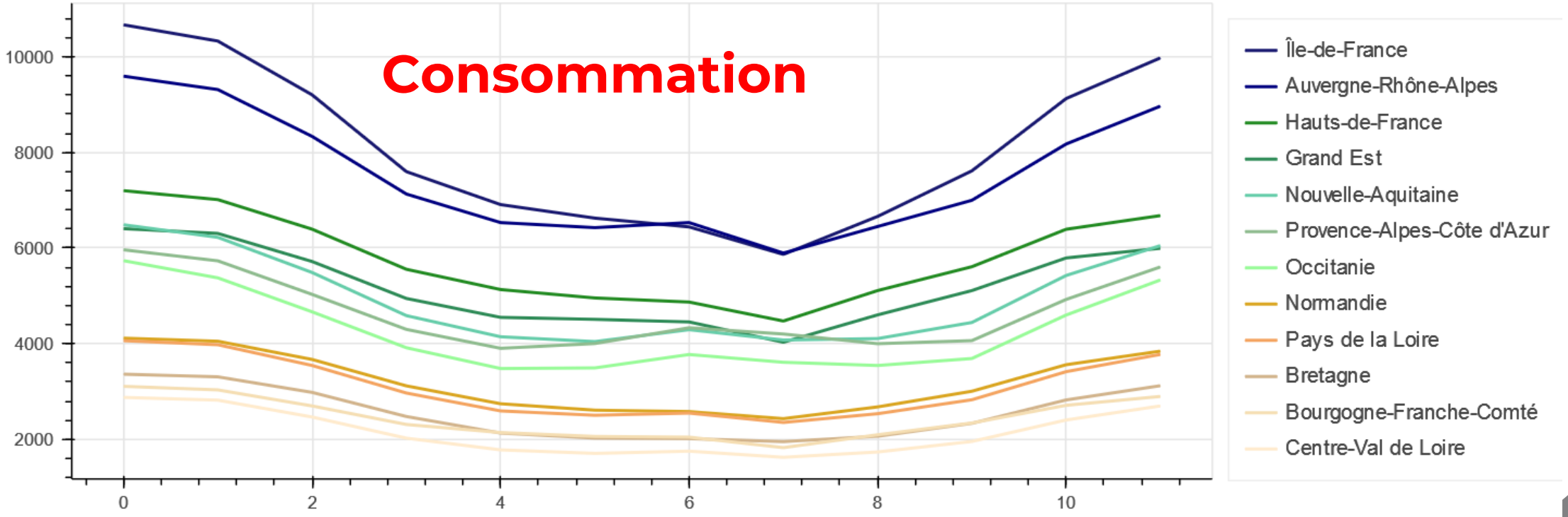
- **Préoccupation**
  - ✓ Regrouper les régions en fonction des séries temporelles issues de la **consommation** ou de la **production** (par type d'énergie).

- **Quelques enjeux**
  - ✓ Dégager les tendances spatiales : mise en évidence d'éventuelles relations complexes qui peuvent exister entre les séries temporelles.
  - ✓ Constituer des classes de données plus homogènes en vue de l'optimisation des modèles et traitements.
  - ✓ Localisation des sites de production des différents types d'énergie.

Modèle Clustering: Relations entre régions

Etude exploratoire spécifique

Variation des séries temporelles par région

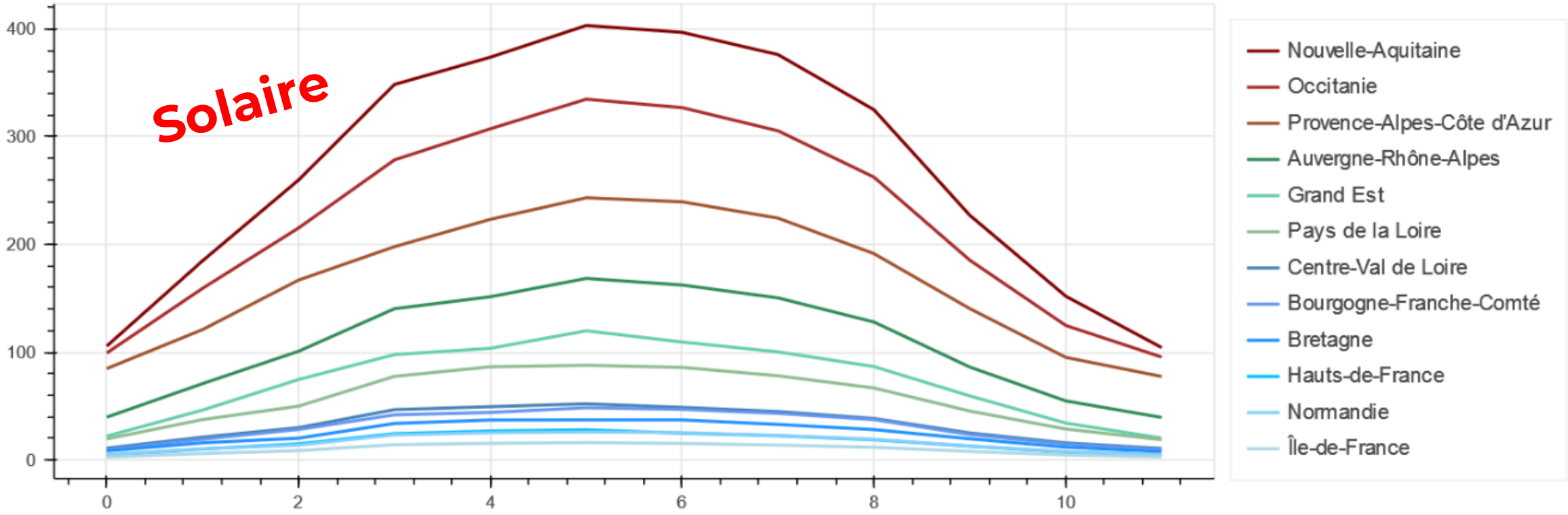


Observation : les courbes sont plus/moins parallèles dans le temps.

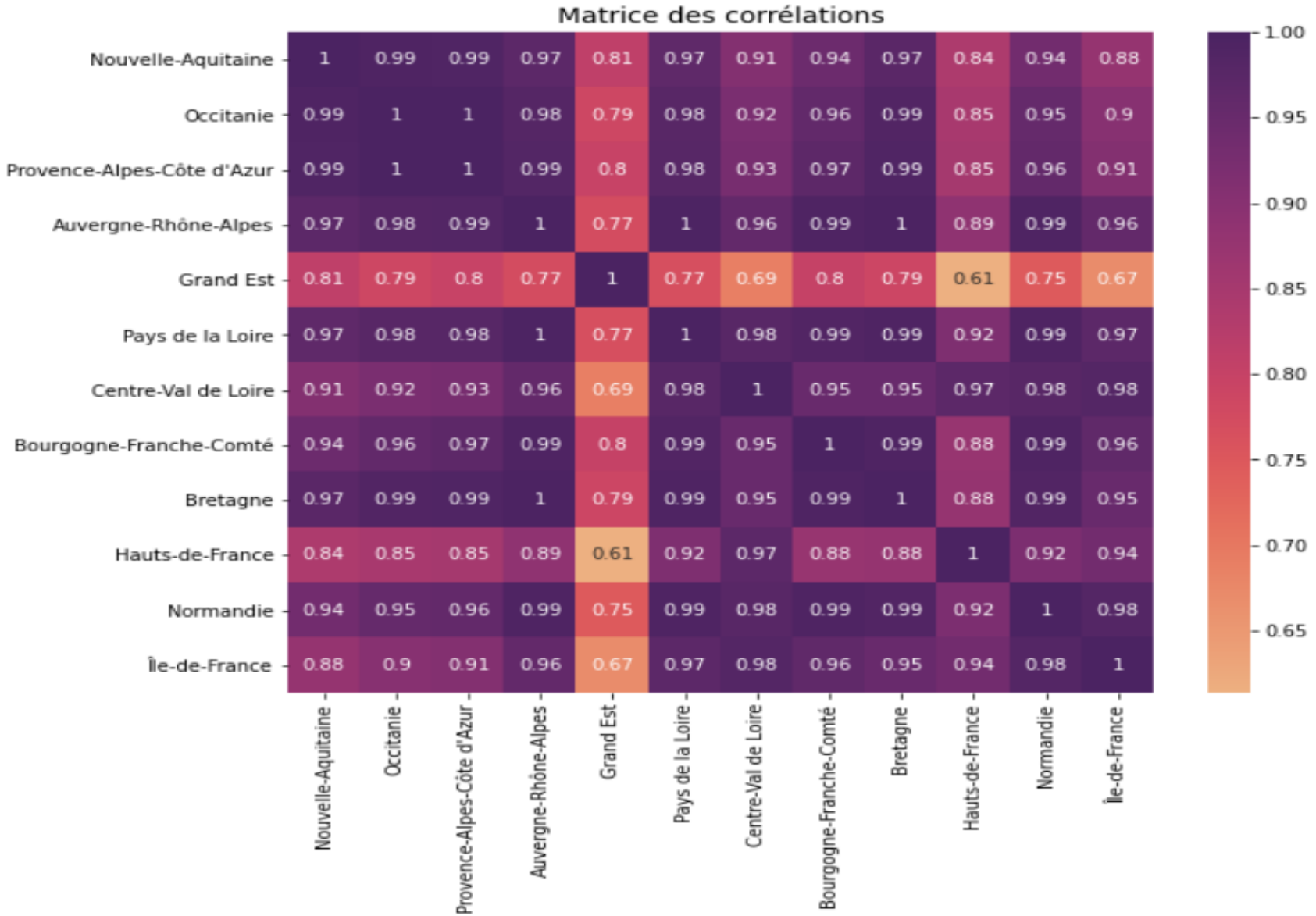
Modèle Clustering: Relations entre régions

Etude exploratoire spécifique

Variation des séries temporelles par région



Hypothèse 1 : Forte corrélation entre les séries temporelles régionales.



Solaire

Confirmation de l'hypothèse 1, matrice de corrélation.

Modèle Clustering: Relations entre régions

Hypothèse

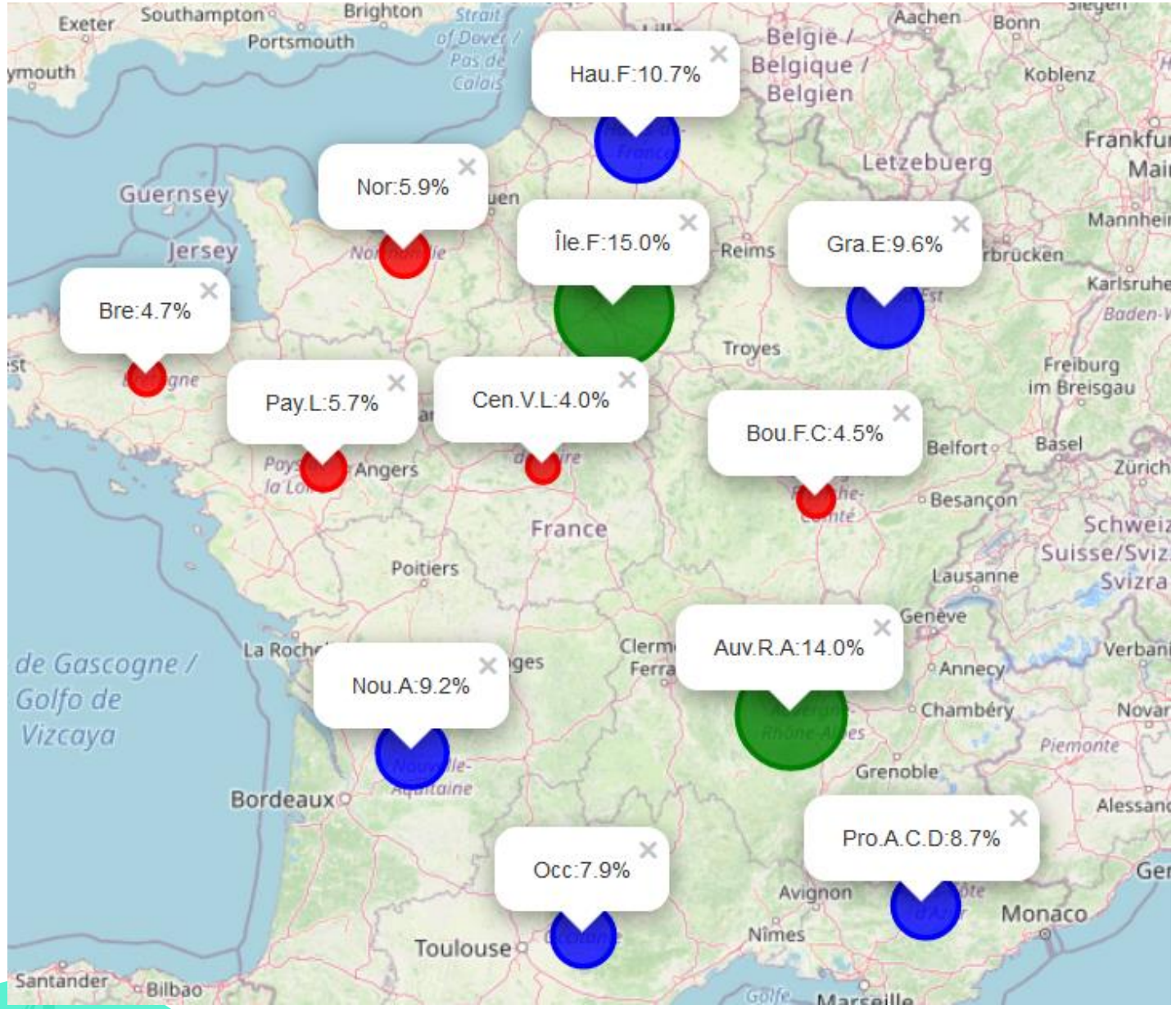
L'ordre de grandeur des valeurs est un critère discriminatoire déterminant.

- Faire ressortir les groupes sur la base de ce critère discriminatoire.

  ⇒ Modèle clustering non supervisé.
  ⇒ Métrique : Dynamic time warping (DTW).
  ⇒ Nombre optimal de classes est 3 (trouvé par la méthode du coude).

- Représentation spatiale des régions

  ⇒ Données de localisations: georef-france-region issue *« Référentiel géographique » Opendatasoft*
  ⇒ Proportion marginale régionale par rapport aux valeurs cumulées
  ⇒ Distinction des clusters à l'aide des couleurs.



➢ Consommation

Région	%
Île-de-France	15.0 %
Auvergne-Rhône-Alpes	14.0 %
Hauts-de-France	10.7 %
Grand Est	9.6 %
Nouvelle-Aquitaine	9.2 %
Provence-Alpes-Côte d'Azur	8.7 %
Occitanie	7.9 %
Normandie	5.9 %
Pays de la Loire	5.7 %
Bretagne	4.7 %
Bourgogne-Franche-Comté	4.5 %
Centre-Val de Loire	4.0 %

➢ Eolien

Région	%
Hauts-de-France	25.4 %
Grand Est	22.8 %
Occitanie	11.3 %
Centre-Val de Loire	8.2 %
Bretagne	6.6 %
Normandie	5.95%
Pays de la Loire	5.9 %
Nouvelle-Aquitaine	5.5 %
Bourgogne-Franche-Comté	4.2 %
Auvergne-Rhône-Alpes	3.7 %
Île-de-France	0.4 %
Provence-Alpes-Côte d'Azur	0.4 %

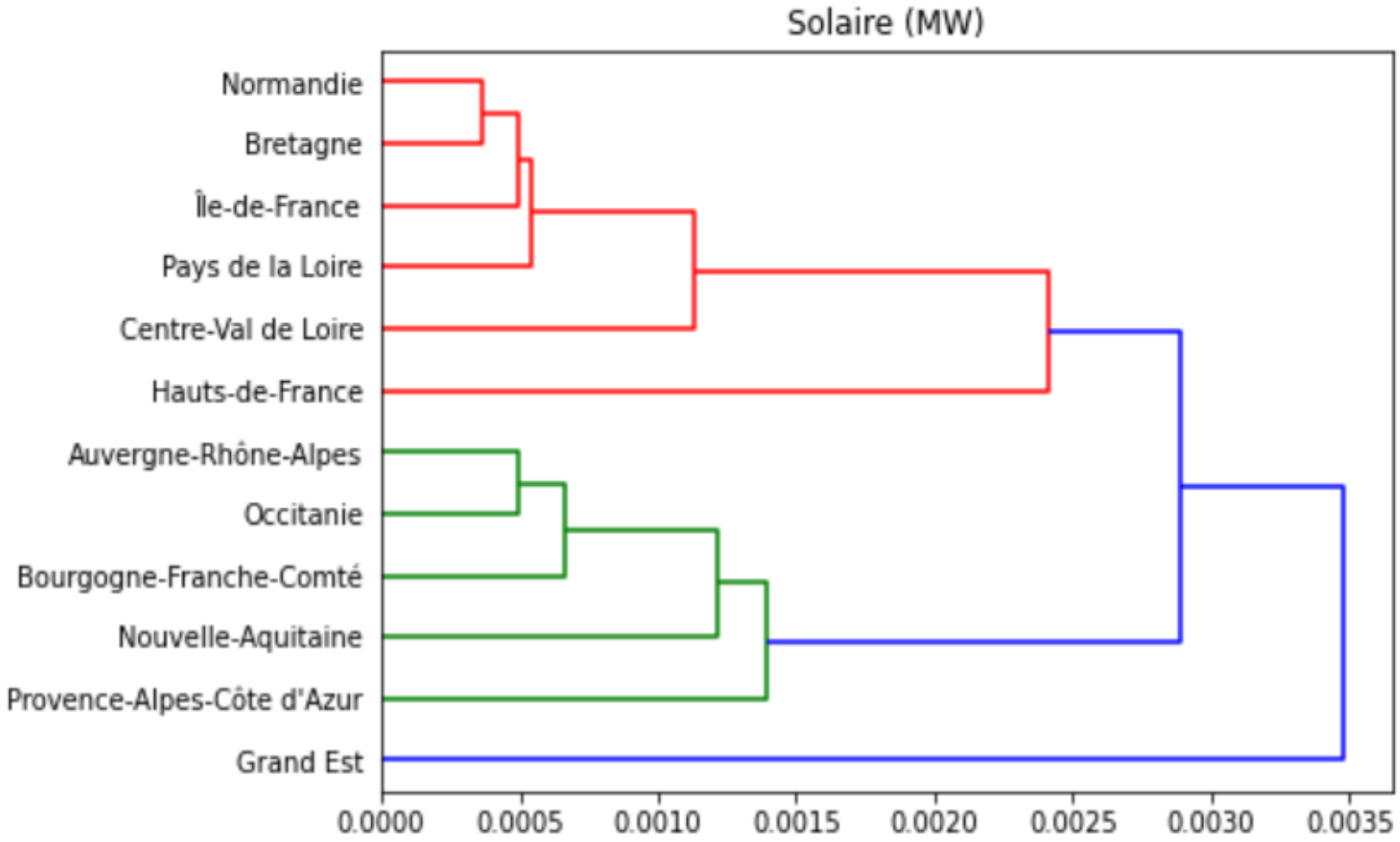
➢ Solaire

Région	%
Nouvelle-Aquitaine	26.3 %
Occitanie	21.8 %
Provence-Alpes-Côte d'Azur	16.2 %
Auvergne-Rhône-Alpes	10.4 %
Grand Est	7.1 %
Pays de la Loire	5.5 %
Centre-Val de Loire	3.2 %
Bourgogne-Franche-Comté	3.0 %
Bretagne	2.3 %
Hauts-de-France	1.6 %
Normandie	1.6%
Île-de-France	1.0 %

Modèle Clustering: Relations entre régions

Remarques finales

- ✓ Confirmation de l'hypothèse sur l'importance des ordres de grandeur.
- ✓ Observation des tendances spatiales (éolien et solaire).
- ✓ Observations corroborées par la classification hiérarchique.



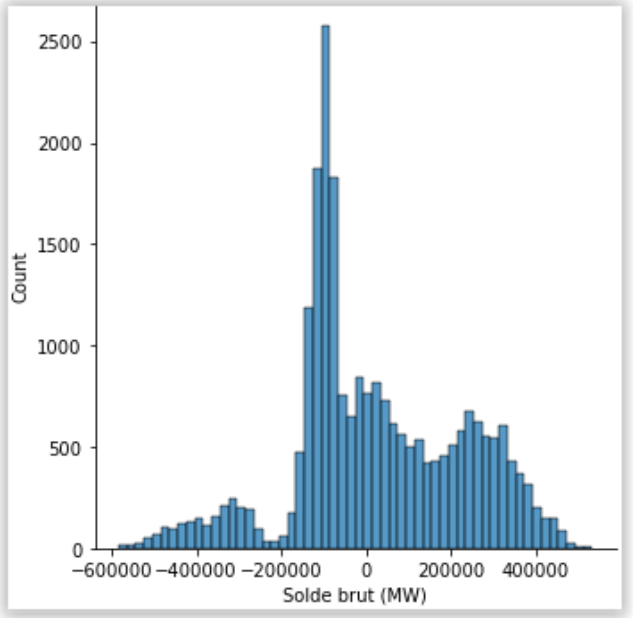
Modèle Classification

➢ **Objectif**: peut-on identifier les risques de blackout au niveau régional?

➢ **Méthode**: combinaison des bases électriques et météorologiques pour une classification supervisée

➢ Sur la base de la distribution des données, 6 buckets de classification des soldes bruts quotidiens:

- ✓ ***Déficit 3*** *= -600,000 MW à -400,000 MW*
- ✓ ***Déficit 2*** *= -400,000 MW à -200,000 MW*
- ✓ ***Déficit 1*** *= -200,000 MW à 0 MW*
- ✓ ***Excédent 1*** *= 0 MW à +200,000 MW*
- ✓ ***Excédent 2*** *= +200,000 MW à +400,000 MW*
- ✓ ***Excédent 3*** *= +400,000 MW à +600,000 MW*



Modèle Classification
Workflow

- Premières itérations sur base complète:
  - ✓ En apparence, **bons résultats globaux** sur les modèles (RandomForest, XGB, Bagging)
  - ✓ Mais **qualité inégale entre classes** (Excédent3 notamment).

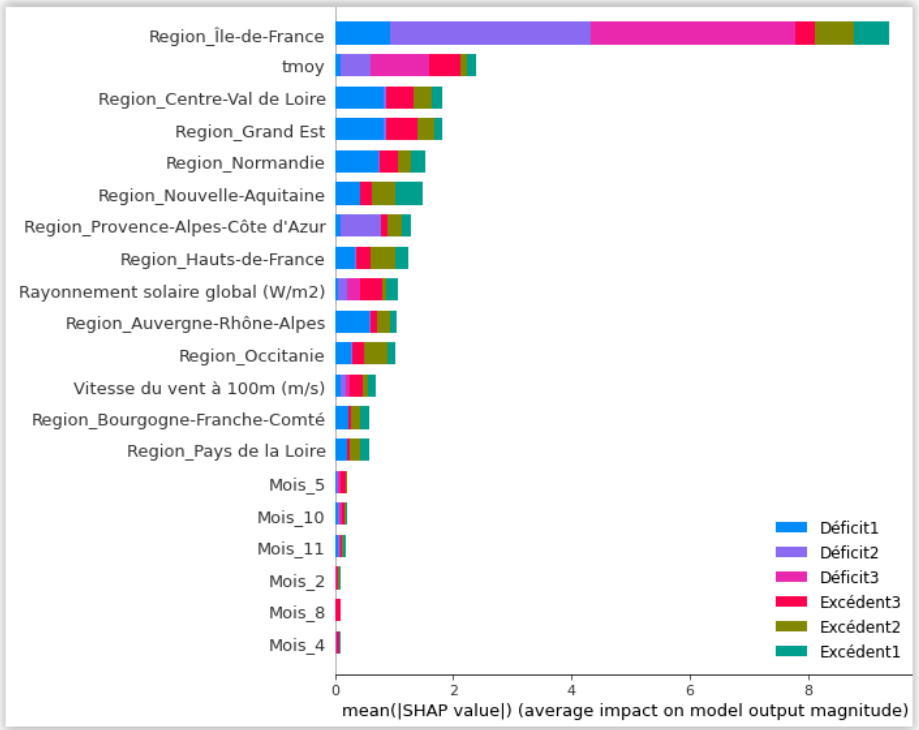
	precision	recall	f1-score	support
Déficit1	0.93	0.93	0.93	2086
Déficit2	0.95	0.91	0.93	270
Déficit3	0.87	0.90	0.88	131
Excédent1	0.78	0.71	0.74	1200
Excédent2	0.77	0.89	0.82	1033
Excédent3	0.58	0.19	0.28	97
accuracy			0.85	4817
macro avg	0.81	0.75	0.76	4817
weighted avg	0.85	0.85	0.85	4817

- Utilisation d'un **RandomOverSampler** pour rééquilibrer les classes et adaptation des modèles précédents

Modèle Classification
Conclusions

- XGBoost avec OverSampling = bons résultats avec un rappel toujours > 75%
  - ✓ Meilleur modèle car fonctionnant le mieux sur l'ensemble des sous-catégories
  - ✓ En cas d'erreur, l'erreur se fait avec une catégorie adjacente

- Régions = principaux facteurs de classification.
- Les variables météo. ont un impact, notamment la température moy.
- Les mois n'ont aucun effet
- Peut permettre de donner une 1re alerte sur les tensions
- Doit être affiné avec plus de variables pour améliorer la précision en production



	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 / Réel	Déficit1	Déficit2	Déficit3	Excédent1	Excédent2	Excédent3
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

DataScientest

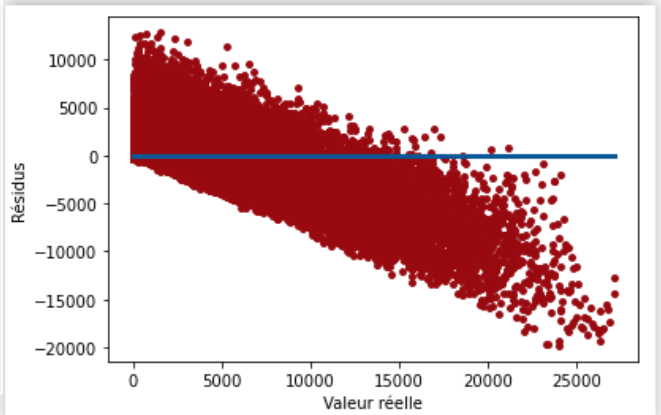
Modèle Régression Eolien

Résumé

- **Objectif**: Peut-on prévoir la production Eolienne en fonction des autres variables (hors tendance temporelle?)
- **Méthode**: plusieurs modèles de régression avec SelectKBest et grille de validation croisée.
- **Conclusions**:
  - ✓ XGBRegressor est le meilleur modèle mais **qualité insuffisante**
  - ✓ Modèle incapable de prédire les **grandes valeurs**

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

- **Explications**:
  - ✓ **Peu de variables significativement corrélées** avec la cible
  - ✓ Seulement 2 régions productrices d'énergie éolienne de manière importante (Hauts-de-France / Grand-Est) et ceci seulement en hiver
  - ✓ **Régions avec vents les plus importants = pas d'éolienne**
  - ✓ **Question du placement des éoliennes en France:** variable des capacités de production manquantes dans le modèle



Conclusion, difficultés et perspectives

➢ Conclusion

Les modèles peuvent répondre raisonnablement aux problématiques exposées :

- ✓**Prévisions** des productions et consommations futures
- ✓Evaluation du **rythme de croissance des renouvelables** dans le mix
- ✓Identification des **points de tension régionaux** par classification supervisée
- ✓Mise en évidence des **disparités** régionales

➢ Difficultés

- ✓ Rupture exogène du **Covid-19** = incertitudes sur toutes les tendances
- ✓ Autres **variables manquantes** intéressantes (capacités de production régionales, finesse des données météorologiques)
- ✓ L'absence de variables empêche d'avoir des régressions de qualité (Eolien, …)

➢ Perspectives

- ✓ Suivi des classifications et clustering, tendances à confirmer pour séries temporelles
- ✓ Amélioration de la finesse des variables