



L'IMPACT DE LA VARIABILITÉ CLIMATIQUE SUR LE SYSTEME ELECTRIQUE FRANÇAIS

MIG PROSPECTUS
Note de synthèse

AULLEN CHOURAC Yannis, ARNAUD Oriane, BALBZIOUI Ziad,
BONMARCHAND Goulven, CHRISTMANN Raphaëlle, DAURES-BOUVET Eliott,
GIUNTA Romain, HOUIMAIRE Marie, LAUVERGNE Alexis, MAZINGUE Léna,
MRIMI Mouad, NEVES Tiago, TOOFA Keanu

Valentina SESSA, Chargée d'enseignement recherche (CMA)
Damien CORRAL, Ingénieur de recherche (CMA)

Table des matières

1 Introduction

A l'heure où la question du climat et de l'énergie s'impose à nous, le lien entre les deux est de plus en plus étudié et exploré. Nous pouvons en effet désormais trouver des modèles énergétiques qui intègrent la variabilité climatique, ce qui n'était pas le cas il y a quelques années. En particulier, la variabilité climatique est particulièrement intéressante pour le domaine de l'hydroélectricité, qui représente près de 14% de la production française et qui dépend directement des précipitations par exemple.

Au cours de ces trois semaines de MIG, nous avons ainsi étudié **comment intégrer la variabilité climatique dans les barrages hydrauliques au fil de l'eau ?**

En effet, il existe deux principaux types de centrales hydrauliques : les centrales à barrages (hydroelectric dam) de moyenne ou haute chute, caractérisés par une implantation dans des régions montagneuses, un débit faible à moyen et une hauteur de chute moyenne à haute (de 30m à >300m), et les centrales hydraulique au fil de l'eau (Run of River - RoR), caractérisés par une implantation le long de grands fleuves ou grandes rivières, avec un débit très fort et une faible hauteur de chute (moins de 30m). Localisés au sein de la région PACA, nous nous sommes concentrés sur les barrages au fil de l'eau, prépondérants dans la région. (*rajouter deux trois chiffres dessus*) Étant placés directement le long des fleuves et rivières, ils dépendent très fortement des conditions climatiques.

De manière plus concrète, nous avons donc essayé de modéliser, de différentes manières en étudiant le cycle de l'eau, la puissance électrique que pourront fournir les barrages en France, en connaissant le volume de précipitations ainsi que la température. On réalise ces prédictions via le facteur de capacité (CF), définit ainsi :

$$CF = \frac{P_{produite}}{P_{capacité installée}}$$

Après s'être entraîné sur des données pour les années 2015 à 2022, puis validé en testant nos modèles sur les données de 2023, nous avons appliqué cela aux années futures, en essayant de modéliser des situations précises à l'horizon 2050.

2 Méthodologie

Afin de faire ces diverses prévisions, nous avons choisi de faire usage de l'apprentissage automatique, également connu sous le nom de Machine Learning, ainsi que du Deep Learning.

Les données dont nous disposons correspondent aux valeurs de températures et de précipitations au niveau régional, ainsi que les CF au niveau national, pour des données journalières allant du premier janvier 2015 au 31 décembre 2023. On cherche donc à prédire le facteur de capacité national à partir des données météorologiques.

2.1 Préparation et exploration des données

Avant toute modélisation, pour réaliser un modèle de machine learning il est nécessaire de manipuler un set de données et de les trouver. Les données journalières de précipitations et de température sont extraites du projet CLIM2POWER (NUTS-2). Pour la réalisation de notre projet, les données étaient déjà pré-nettoyées. Les données sont explorées et analysées pour trouver de nouvelles *features* qui permettront d'améliorer le modèle.

L'étape de **feature engineering** est cruciale pour améliorer les modèles. Elle consiste à manipuler les données brutes pour extraire des caractéristiques (features) afin de renforcer l'apprentissage du modèle. Pour notre problème, ces caractéristiques sont :

- la saisonnalité analysée sur les courbes de température et de Capacity Factors,
- la fonte des neiges au printemps,
- l'accumulation des précipitations.

2.2 Apprentissage Automatique

Le Machine Learning est une branche de l'Intelligence Artificielle qui permet à un système "d'apprendre", de faire des prédictions et de prendre des décisions, suite à une période d'apprentissage.

L'objectif principal est que le modèle entraîné puisse fonctionner sur des données nouvelles, inconnues, pour prédire. Ici, l'idée est donc d'entraîner le modèle sur les données des années précédentes, pour essayer de prévoir les facteurs de capacité futurs.

Pour garantir cette généralisation, on divise l'ensemble total des données en trois parties distinctes et chronologiques.

- **Training Set** : c'est le plus grand ensemble de données (souvent 60% à 80% du total), utilisé pour ajuster les paramètres (les poids) du modèle.
- **Validation Set** : il est utilisé pour optimiser les hyperparamètres du modèle et permet de trouver la configuration qui donne la meilleure performance avant de voir les données de test.
- **Test Set** : c'est le plus petit ensemble. Il est utilisé une seule fois, à la toute fin du processus, pour obtenir une estimation non biaisée de la performance du modèle sur des données complètement nouvelles.

Dans notre projet, le *training set* et le *validation set* correspondent aux données de 2015 à 2022, et le *test set* correspond aux données de 2023.

2.3 Erreurs

Afin d'optimiser nos modèles, nous essayons de minimiser plusieurs types d'erreurs. On note y le résultat réel, \hat{y} le résultat prédit, et \bar{y} la moyenne des y_i .

- MAE : MSE : moyenne de l'erreur absolue, qui mesure la moyenne des résidus dans le data set

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- MSE : Moyenne de l'erreur au carré, qui mesure la variance des résultats

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- R^2 : mesure la proportion de la variance de la variable dépendante (la cible, ici le facteur de capacité CF) qui est expliquée par les variables indépendantes (nos features, TA et TP) dans le modèle de régression.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

2.4 Minimisation et Optimisation

A la vue de ces différentes erreurs, nous avons mis en place plusieurs stratégies pour réduire les erreurs.

Sur le traitement des prévision météorologiques, nous avons effectué des moyennes glissantes sur 10 ans, séparé et codé différents scénarios, et supprimé de la modélisation les régions avec une trop faible corrélation entre les données météorologiques et les CF.

Sur le traitement des données lors de l'entraînement, nous avons ainsi procédé à la normalisation des données, au calcul du biais, au moyennage. Nous avons également mis en place des fonctions rendant compte des saisons (ici les fonctions cosinus et sinus) et le lag (utilisation de la valeur d'une variable dans une région voisine pour prédire ou expliquer la valeur de cette variable dans la région, avec des régions temporelles ou spatiales).

3 Modèles de Machine Learning

3.1 Approche Linéaire

L'approche linéaire consiste à modéliser le système à l'aide de fonctions linéaires : c'est une approche assez simple à implémenter et à concevoir à l'aide de la descente de gradient, mais ce n'est pas la plus performante.

En mettant en place les différentes stratégies, nous avons pu obtenir : [RESULTAT MAIS PAS GRAPHIQUE]

3.2 Random Forest

Random Forest est une technique basée sur les arbres de récursion, qui améliore les arbres baggués pour construire des arbres décorrélés.

3.3 Gradient Boosting

Le Gradient Boosting est également une technique utilisant les arbres et la récursion. Néanmoins, contrairement au Random Forest qui construit ses arbres en parallèle et de manière indépendante, le Gradient Boosting construit ses arbres de manière séquentielle et additive. Chacun de ces arbres peut être relativement petit et, en général, cela améliore progressivement la précision du modèle.

4 Deep Learning

Nous nous sommes ensuite penchés sur une approche différente : le Deep Learning. C'est un autre sous-domaine de l'apprentissage automatique qui utilise des réseaux de neurones artificiels composés de nombreuses couches pour analyser des données et prendre des décisions.

En particulier, nous avons utilisé le modèle LSTM (Long Short-Term Memory, soit Mémoire à Long Terme et à Court Terme), qui est un type de réseau de neurones récurrents (RNN) particulièrement adapté pour traiter et prédire les séries temporelles et les séquences de données. (nb : savoir expliquer correctement toutes les gate etc.)

Les résultats obtenus sont les suivants :

[RESULTATS]

5 Comparaison des Résultats

Modèle	R ²	MAE	MSE	RMSE
Ridge Regression				
Random Forest				
Gradient Boosting				
LSTM				

TABLE 1 – Tableau récapitulatif des meilleurs résultats par modèle

6 Prospective

7 Conclusion