Raport de projet

Chen Chaofan, Peng haochuan

February 2025

1 Introduction

La prévision des prix de l'électricité (Electricity Price Forecasting, EPF) est un problème clé dans l'exploitation du marché de l'électricité et la gestion de l'énergie. En raison des caractéristiques non linéaires, périodiques et fortement bruitées des prix de l'électricité, la capacité de généralisation des modèles est essentielle. L'augmentation des données, en tant que moyen important d'améliorer la précision des prévisions, permet d'enrichir les caractéristiques des données et d'atténuer le problème de rareté des données.

Cette étude compare l'effet de l'absence d'augmentation des données, de l'augmentation basique des données, de l'augmentation avancée des données et des données simulées d'Air Liquide sur la prévision des prix de l'électricité aux horizons de 6, 12, 24, 48, 72 et 168 heures. Les expériences sont menées en utilisant le même modèle afin d'évaluer l'impact des différentes stratégies d'augmentation des données sur la performance des prévisions. Les résultats contribuent à une meilleure compréhension du rôle de l'augmentation des données dans la prévision des prix de l'électricité et fournissent un appui pour améliorer la précision des prévisions.

2 Méthodologie

2.1 Augmentation des données (Data Augmentation)

Cette étude compare trois stratégies de traitement des données avec les données simulées d'Air Liquide :

- **Données sans augmentation (Baseline)** : Utilisation directe des données brutes des prix de l'électricité pour l'entraînement du modèle, sans traitement supplémentaire.
- **Augmentation avancée des données (TimeGAN)** : TimeGAN combine un autoencodeur et un réseau antagoniste génératif (GAN) pour apprendre les caractéristiques temporelles et la distribution des séries chronologiques. Il génère des données synthétiques avec des dynamiques réalistes afin d'améliorer la capacité de généralisation du modèle.
- **Augmentation basique des données (Basic Augmentation)** : Application d'opérations courantes d'augmentation aux séries temporelles, notamment :
  - **Perturbation par bruit (Jittering)** : Ajout d'un bruit aléatoire de faible amplitude aux données.
  - **Distorsion temporelle (Time Warping)** : Modification de la distribution temporelle des données par interpolation ou compression de l'axe du temps.
  - **Translation (Shifting)** : Déplacement des données le long de l'axe temporel pour simuler des variations périodiques.
  - **Mise à l'échelle (Scaling)** : Agrandissement ou réduction des données afin d'améliorer leur robustesse.

2.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) est un type de réseau de neurones récurrents (RNN) conçu pour capturer les dépendances à long terme dans les séries temporelles. Grâce à son mécanisme de portes

(entrée, oubli et sortie), il régule le flux d'information et atténue les problèmes de gradient rencontrés dans les RNN classiques.

Dans la prévision des prix de l'électricité, LSTM est utilisé pour modéliser les tendances temporelles complexes et améliorer la précision des prédictions en apprenant les relations non linéaires entre les prix historiques.

2.3 Apprentissage par transfert (Transfer Learning)

L'apprentissage par transfert est une technique qui permet de réutiliser un modèle pré-entraîné sur un ensemble de données source pour une nouvelle tâche sur un ensemble de données cible. Cette approche est particulièrement utile lorsque les données disponibles pour la tâche cible sont limitées.

Dans cette étude, nous appliquons l'apprentissage par transfert en utilisant un modèle entraîné sur des données simulées pour améliorer la prédiction des prix de l'électricité à partir de données réelles. Cette stratégie permet d'exploiter les connaissances acquises sur les données simulées tout en s'adaptant aux dynamiques des données réelles, améliorant ainsi la précision des prévisions.

3 Experimental

3.1 Augmentation des données

L'augmentation des données est utilisée pour enrichir l'ensemble d'entraînement et améliorer la robustesse du modèle. Dans cette étude, nous appliquons deux stratégies d'augmentation : l'augmentation basique et l'augmentation basée sur TimeGAN.

3.1.1 Augmentation basique (Basic Augmentation)

L'augmentation basique vise à enrichir les séries temporelles en appliquant des transformations simples. Dans cette étude, nous effectuons un décalage temporel aléatoire des données de quelques jours pour simuler des variations périodiques, une mise à l'échelle en multipliant les prix par un facteur compris entre 0.95 et 1.05 afin de refléter des fluctuations du marché, ainsi qu'un ajout de bruit gaussien de faible amplitude ($\sigma = 0.02$) pour améliorer la robustesse du modèle face aux variations aléatoires. Les données ainsi augmentées sont ensuite combinées aux données originales pour renforcer l'apprentissage du modèle.

3.1.2 Augmentation via TimeGAN

TimeGAN est un modèle génératif basé sur les réseaux adversaires génératifs (GAN) combiné à un autoencodeur récurrent, permettant d'apprendre les dépendances temporelles des séries chronologiques. Il génère des données synthétiques ayant des dynamiques similaires aux données d'origine.

Dans cette étude, TimeGAN est entraîné sur les données historiques des prix de l'électricité en utilisant un réseau encodeur pour extraire des représentations latentes, un superviseur pour capturer la structure temporelle et un générateur pour produire de nouvelles séquences réalistes. Un discriminateur est ensuite utilisé pour distinguer les séquences réelles des séquences générées, améliorant ainsi la qualité des données synthétiques. Ces données augmentées sont ensuite intégrées à l'ensemble d'entraînement pour améliorer la robustesse du modèle de prévision. Le modèle est entraîné avec 200 époques et un taux d'apprentissage de 0.001 en utilisant la fonction `train_timegan(model, dataloader, epochs=200, lr=0.001)`.

3.2 Prétraitement des données

Afin d'assurer une équité expérimentale entre les différentes sources de données, nous avons harmonisé la taille des ensembles de données utilisés. Les méthodes d'augmentation de données ont chacune généré 70 128 heures de prix de l'électricité, et nous avons extrait une séquence de même longueur à partir des données simulées fournies par Air Liquide. En revanche, les données réelles étant limitées à environ 30 000 heures, elles ont été utilisées dans leur totalité.

Chaque série temporelle a été transformée en un ensemble d'échantillons avec une fenêtre glissante, où chaque entrée contient une séquence de *window_size* heures précédentes et la sortie correspond aux prévisions des prix pour plusieurs horizons (*forecast_steps* de 6, 12, 24, 48, 72 et 168 heures). Pour

optimiser le stockage et l'accès, les données ont été converties en fichiers `memmap`. Par ailleurs, nous avons appliqué une normalisation basée sur la moyenne et l'écart-type du jeu d'entraînement pour assurer la cohérence des distributions. Les ensembles ont ensuite été divisés en données d'entraînement, de validation et de test, avec des proportions adaptées aux volumes disponibles pour chaque source de données.

3.3 Modèle utilisé

Pour garantir une comparaison équitable entre les différentes sources de données, nous utilisons un modèle identique pour tous les ensembles d'entraînement. Il s'agit d'un réseau neuronal basé sur **Long Short-Term Memory (LSTM)**, conçu pour capturer les dépendances temporelles dans les séries chronologiques.

Le modèle prend en entrée des séquences de *window_size* heures avec 5 caractéristiques, puis passe par deux couches LSTM avec 64 et 128 unités respectivement. La sortie est ensuite traitée par une couche dense intermédiaire de 8 neurones avec activation ReLU, suivie d'une couche de sortie linéaire avec 6 neurones correspondant aux horizons de prévision définis ($forecast\_steps = 6$, 12, 24, 48, 72 et 168 heures).

3.4 Entraînement du modèle

Tous les modèles sont entraînés avec les mêmes hyperparamètres afin d'assurer une comparaison équitable des performances. L'objectif est de prédire les prix de l'électricité aux horizons de 6, 12, 24, 48, 72 et 168 heures.

Dans un premier temps, nous entraînons un modèle uniquement sur les données réelles avec 10 époques et un taux d'apprentissage de 0.001. Ensuite, nous pré-entraînons trois autres modèles sur des ensembles augmentés issus de l'augmentation basique, des données synthétiques d'Air Liquide et des données générées par TimeGAN, en utilisant les mêmes paramètres d'entraînement. Pour évaluer l'impact de l'apprentissage par transfert, ces modèles pré-entraînés sont affinés sur les données réelles avec un taux d'apprentissage réduit à 0.0001 sur 10 époques supplémentaires.

Chaque modèle est sauvegardé après son entraînement afin d'être utilisé pour l'évaluation des performances.

4 Résultats

5 Résultats

Les performances des modèles sont évaluées en termes d'erreur quadratique moyenne (*Mean Squared Error*, MSE) sur différentes horizons de prévision (6, 12, 24, 48, 72 et 168 heures). Le tableau 1 présente les résultats obtenus pour les modèles entraînés avec les différentes stratégies de données : sans augmentation, augmentation basique, augmentation via TimeGAN et données synthétiques d'Air Liquide.

Table 1: Comparaison des performances des modèles en fonction des différentes stratégies d'augmentation de données (MSE).

Temps (h)	**Sans aug.**	**Aug. base**	**TimeGAN**	**Synth. Air Liquide**
6	78.17	61.93	60.41	66.71
12	97.68	78.16	81.42	88.41
24	135.20	116.91	106.05	120.88
48	184.83	138.06	149.64	164.40
72	208.80	172.83	171.80	199.99
168	226.54	197.29	203.15	229.04

Les résultats montrent que les méthodes d'augmentation des données permettent d'améliorer significativement la précision des prévisions par rapport à l'entraînement sur les seules données réelles. L'augmentation basique et l'augmentation via TimeGAN offrent les meilleures performances, réduisant

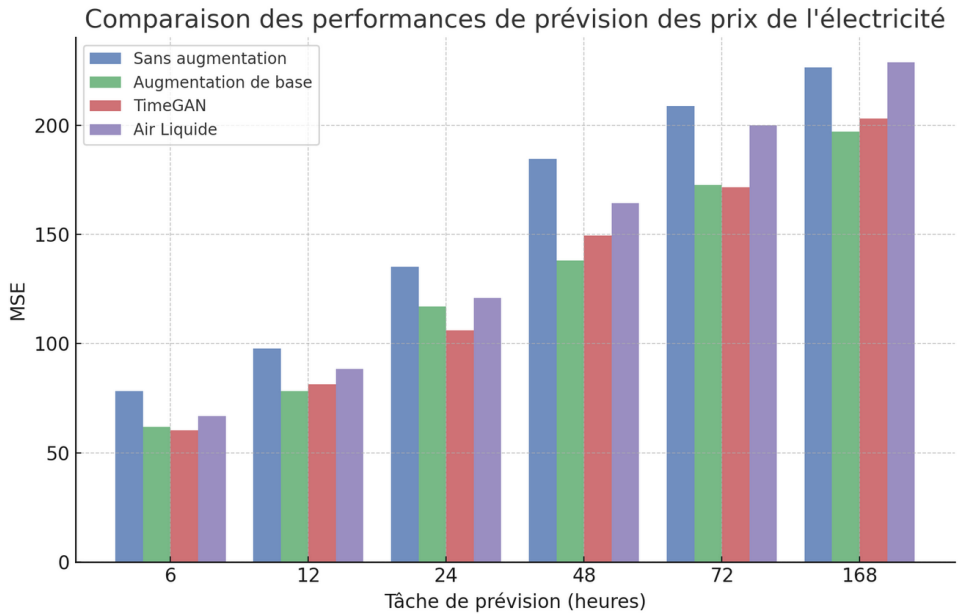

Figure 1: Comparaison des performances de prévision des prix de l'électricité en fonction des différentes stratégies d'augmentation de données (MSE).

l'erreur MSE de manière notable, en particulier sur les courtes échéances (6 à 24 heures). L'utilisation de données synthétiques d'Air Liquide améliore légèrement la précision par rapport aux données réelles seules, mais reste inférieure aux autres méthodes d'augmentation.

De manière générale, les résultats indiquent que l'augmentation de données est une stratégie efficace pour améliorer la robustesse du modèle, et que TimeGAN permet de mieux capturer les dynamiques temporelles des séries chronologiques.

6 Conclusion

Dans cette étude, nous avons évalué l'impact de différentes stratégies d'augmentation de données sur la prévision des prix de l'électricité. Les résultats montrent que l'augmentation basique et l'augmentation via TimeGAN permettent d'améliorer significativement la précision des prévisions, notamment pour les horizons courts (6 à 24 heures). L'utilisation des données synthétiques fournies par Air Liquide, bien que moins performante que les méthodes d'augmentation des données, présente l'avantage d'un volume important. Cela suggère que l'exploitation d'un plus grand ensemble de ces données pourrait encore améliorer la précision du modèle.

Pour les travaux futurs, deux pistes d'amélioration peuvent être envisagées : d'une part, entraîner des modèles distincts pour chaque horizon de prévision afin d'optimiser les performances spécifiques à chaque tâche, et d'autre part, exploiter un volume encore plus large de données simulées pour améliorer la robustesse et la généralisation des modèles de prévision.