# IA et Applications

## Projet Data Augmentation - Air Liquide

réalisé par Julian AKUESON, MAHENDRAN Sujeeban et TRAN Michael

**Introduction à l'analyse des résultats**

Dans cette section, nous analysons l'impact des différentes techniques de **Data Augmentation** appliquées aux séries temporelles de prix. L'objectif est d'évaluer dans quelle mesure ces méthodes influencent la performance du modèle de prédiction et d'identifier les approches les plus pertinentes pour améliorer la robustesse des prévisions.

Cinq techniques ont été testées :

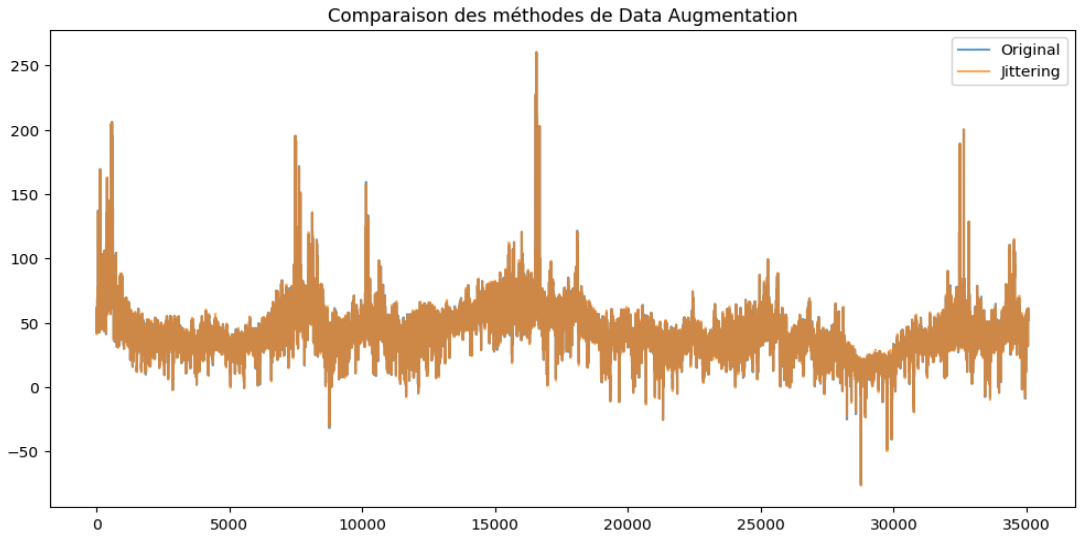
1. **Jittering** : Ajout de bruit gaussien pour simuler des variations réalistes sans altérer la structure des données.
2. **Time Warping** : Modification temporelle des séries, introduisant des déformations pour tester la résilience du modèle face aux variations temporelles.
3. **Window Slicing** : Découpage aléatoire des sous-séries pour enrichir le dataset et accroître la diversité des échantillons utilisés pour l'entraînement.
4. **Permutation :** Réarrangement aléatoire des segments d'une série temporelle afin de perturber l'ordre original tout en conservant la distribution des valeurs, ce qui permet d'évaluer la robustesse du modèle face à des séquences ordonnées.
5. **Rotation :** Transformation circulaire des données, où une partie de la série est déplacée en fin ou en début de séquence, conservant ainsi l'intégrité des valeurs mais modifiant leur alignement temporel pour tester la capacité du modèle à reconnaître des motifs décalés.

Nous allons observer comment ces techniques modifient les données d'origine, analyser leurs avantages et inconvénients, et déterminer lesquelles sont les plus adaptées à notre problématique. Enfin, nous exclurons les méthodes inefficaces sur la base des résultats observés avant de poursuivre les tests sur notre modèle de prédiction.

## Analyse des résultats :

Comparons comment chaque méthode modifie les données d'origines :

**Jittering (Ajout de bruit gaussien)**




**Observation :**

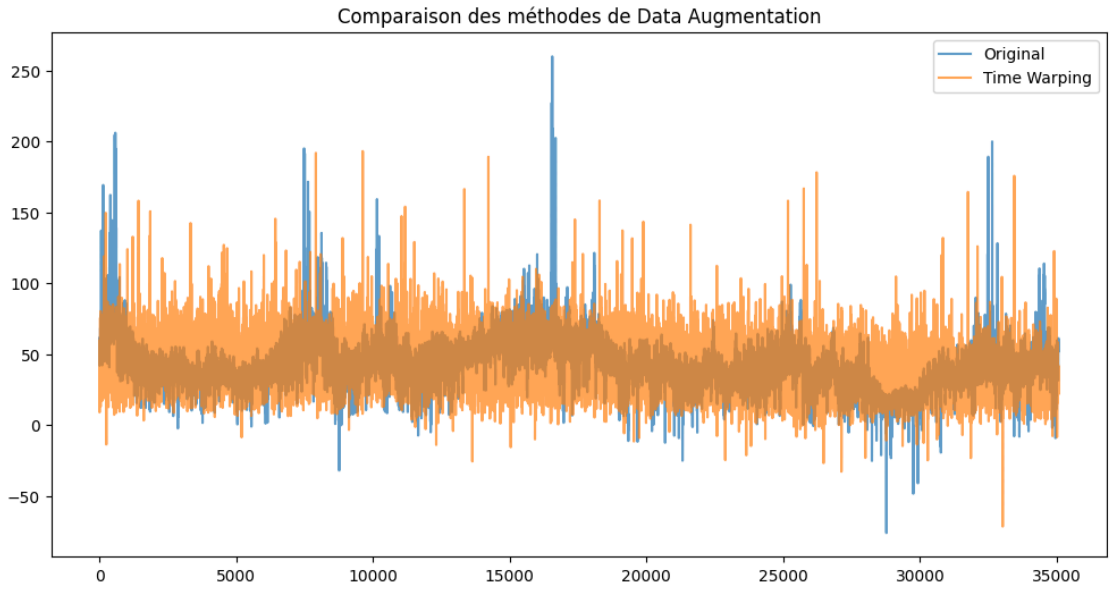
- La série transformée est superposée à la série originale.
- Elle conserve très bien la structure et les tendances.
- Peu de variations visibles, juste un léger bruit ajouté.

**Interprétation :**

- **Avantage** : Bonne pour renforcer la robustesse du modèle sans altérer les tendances.
- **Limite** : N'ajoute pas énormément de diversité.

---

**Time Warping (Déformation temporelle)**



**Observation :**

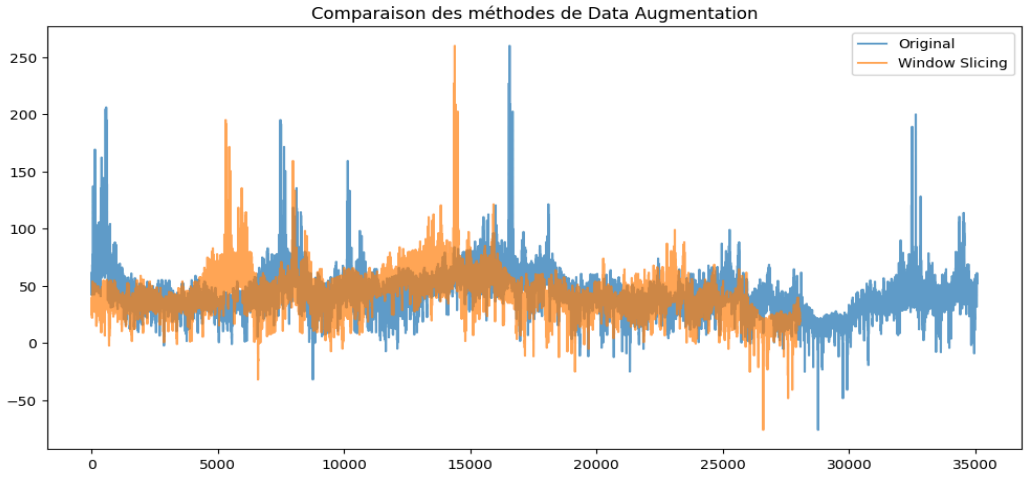
- Plus de variations visibles par rapport à l'original.
- Certaines zones ont été décalées dans le temps, introduisant des distorsions.
- Moins de pics alignés avec la série originale.

**Interprétation :**

- **Avantage** : Utile pour entraîner un modèle à gérer des variations temporelles.
- **Limite** : Peut perturber la structure saisonnière des données.

---

**Window Slicing (Découpage aléatoire des sous-séries)**




**Observation :**

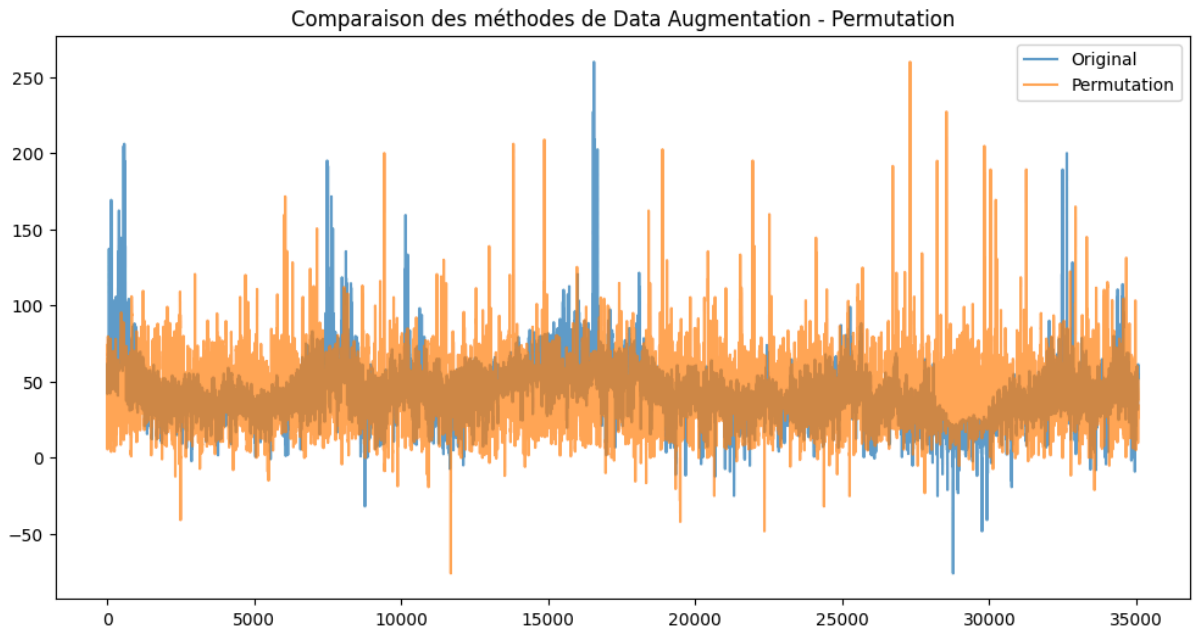
- La série transformée conserve bien la forme générale, mais les pics ne correspondent plus exactement.
- L'amplitude des variations semble légèrement modifiée.
- Cette technique semble créer des sous-séries différentes mais cohérentes.

**Interprétation :**

- **Avantage** : Excellente méthode pour augmenter la taille du dataset.
- **Limite** : Peut supprimer des informations importantes sur le long terme.

---

**Permutation**




**Observation :**

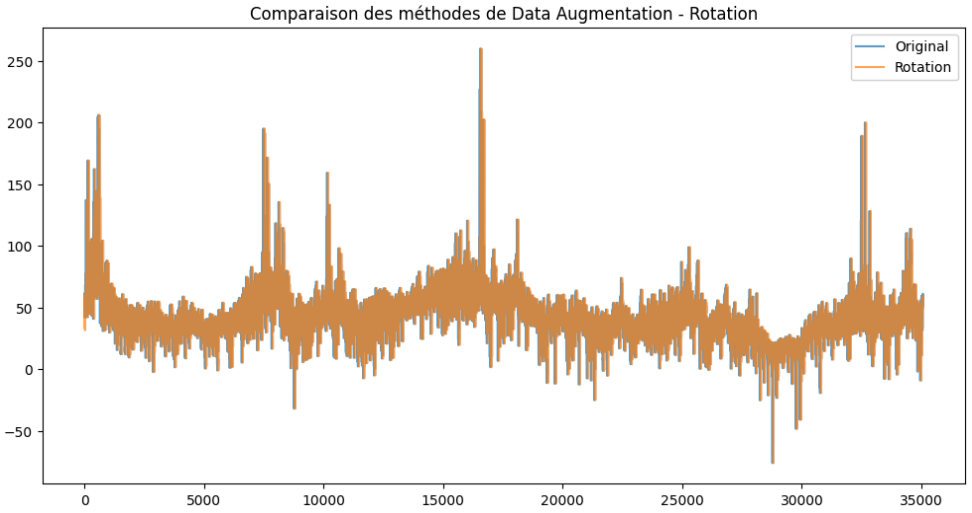
- La série transformée a une variation supérieure et les pics ne correspondent plus exactement.
- Cette technique essaie de créer des un graphique qui correspond mais dans la généralité, les différences sont grandes.

**Interprétation :**

- **Avantage** : Enrichit la diversité des données.
- **Limite** : Peut dégrader les relations temporelles importantes, ce qui est mauvais si la précision est essentielle pour le modèle.

---

**Rotation**



**Observation :**

- La série transformée semble obtenir une courbe relativement similaire à la courbe de base.
- Cette technique essaie de créer des un graphique qui correspond mais dans la généralité, les différences sont grandes.

**Interprétation :**

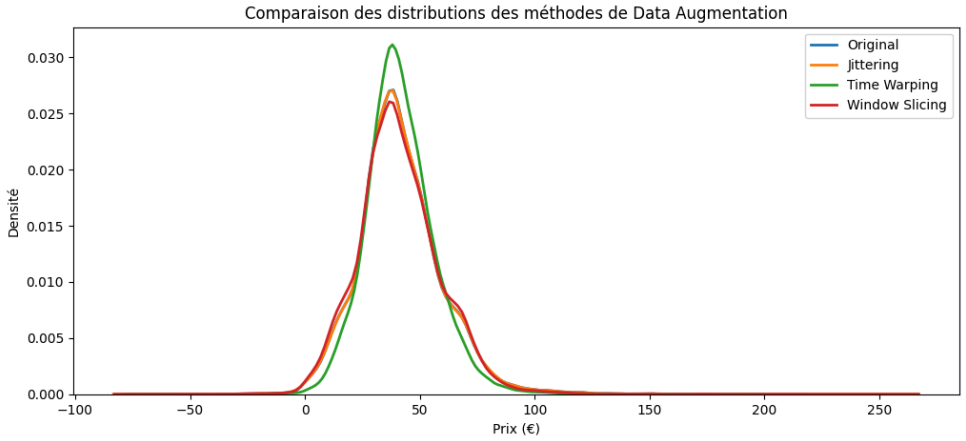
- **Avantage** : Conserve les valeurs et enrichit les données.
- **Limite** : Il semble y avoir un overfitting par rapport aux données de base.

---

## Bilan :

A partir des courbes nous pouvons dores et déjà exclure le **Time Warping** de nos tests, car la courbe générée est très éloignée de la courbe originale.

Pour appuyer nos propos vous trouverez ci-joint un graphique illustrant la distribution des prix :




## Prochaine étape : Tester avec un modèle

**Test sur plusieurs horizons** `[6, 12, 24, 48, 72, 168]` heures.

Entraînement avec un **LSTM pour chaque horizon** avec et sans augmentation de données dans un but comparatif.

- Techniques de Data Augmentation appliquée
  - Jittering
  - Time Warping
  - Window Slicing
  - Permutation
  - Rotation

- Construction d'un modèle **LSTM**
    - 2 couches LSTM pour capturer les patterns temporels.
    - Entraînement avec l'optimiseur **Adam** et la perte **MSE**.

- Metrics d'entraînement :
    - **MAE** (Mean Absolute Error) et **RMSE** (Root Mean Squared Error) sont utilisés comme métriques.
        - **Si MAE et RMSE baissent après Data Augmentation**, cela signifie que la technique a aidé le modèle.

Tableau comparatif des résultats du training en fonction des techniques de data augmentation et de l'horizon:

| Horizon (h) | Originale (pas d'augmentation) | | Jittering | | Window Slicing | | Rotation | | Permutation | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** |
| 6 | 6.5115 | 9.2008 | 6.5516 | 9.212 | 25.7033 | 30.8505 | 10,7555 | 13,4077 | 9,9699 | 12,3413 |
| 12 | 6.9892 | 10.1036 | 7.4277 | 10.4063 | 12.2539 | 15.6456 | 9,9593 | 12,5291 | 11,5451 | 13,5734 |
| 24 | 9.9483 | 13.4996 | 8.39 | 11.558 | 16.9365 | 20.9105 | 9,9282 | 12,4822 | 10,3749 | 13,6041 |
| 48 | 12.2843 | 16.2189 | 13.5776 | 17.3424 | 19.174 | 23.6295 | 9,9107 | 12,1374 | 16,1105 | 17,502 |
| 72 | 11.3358 | 14.9763 | 12.545 | 16.5349 | 17.7705 | 22.0924 | 10,2368 | 12,6635 | 12,1022 | 13,3844 |
| 168 | 8.2276 | 11.4979 | 7.9782 | 11.3015 | 20.7989 | 25.5755 | 10,0511 | 12,2844 | 9,867 | 12,1049 |

Récapitulatif:

| Horizon (h) | **Méthode** | **Justification** |
|---|---|---|
| 6 | Originale | Plus précis, peu de variance. |
| 12 | Originale | Écart faible avec Jittering mais reste plus fiable. |
| 24 | Jittering | Améliore légèrement la performance par rapport à l'original. |

| | | |
|---|---|---|
| 48 | Originale | Jittering est légèrement moins performant. |
| 72 | Originale | Jittering reste proche mais pas meilleur. |
| 168 | Jittering | Meilleur que l'original et Window Slicing. |

### Conclusion : Quelle méthode privilégier ?

Les résultats obtenus permettent d'identifier la technique de **Data Augmentation** la plus adaptée en fonction de l'horizon de prédiction :

- **Pour les prévisions à court terme (moins de 48 heures)** : La méthode **originale** demeure la plus fiable, offrant les meilleures performances en termes de précision et de stabilité des prévisions. L'ajout de bruit ou la modification des structures temporelles ne semble pas apporter d'amélioration significative sur ces horizons.
- **Pour les prévisions à long terme (au-delà de 168 heures)** : La technique de **Jittering** présente un léger avantage en rendant le modèle plus robuste aux variations des données, tout en conservant la structure des séries temporelles. Elle peut donc être privilégiée pour les prévisions à horizon étendu.
- **Concernant la méthode Window Slicing** : Elle montre des performances inférieures, notamment sur les courtes prévisions, avec une dégradation significative des erreurs de prédiction. Cette méthode peut toutefois être envisagée pour des expérimentations sur des horizons très longs, mais elle ne constitue pas une approche optimale dans le cadre de cette étude.

Ainsi, le choix de la méthode de **Data Augmentation** dépend fortement de l'horizon de prédiction visé. Pour les besoins de ce projet, l'utilisation de **Jittering** est recommandée pour les longues prévisions, tandis que la méthode originale reste la plus performante pour les horizons courts.

## Techniques Avancées de Data Augmentation : Expérimentation avec TimeGAN

### Introduction

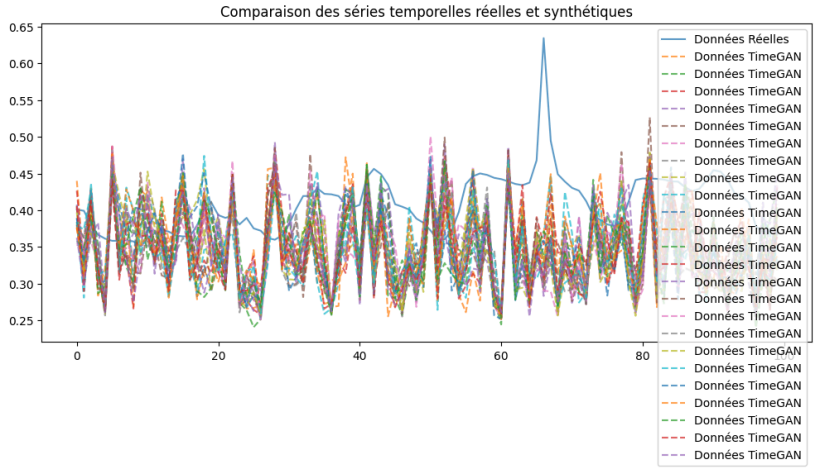
Dans cette section, nous explorons l'utilisation de TimeGAN (Time-series Generative Adversarial Networks) comme technique avancée de data augmentation pour générer des séries temporelles synthétiques. Contrairement aux méthodes de transformations classiques (Jittering, Window Slicing, etc.), TimeGAN permet de générer des séquences temporelles tout en préservant les dynamiques sous-jacentes des données originales.

### Objectif de l'Expérimentation

L'objectif principal de cette expérimentation est de comparer la qualité des données générées par TimeGAN avec celles des données réelles et d'évaluer si ces données synthétiques peuvent améliorer les prédictions des modèles de machine learning, notamment LSTM.

## Résultats

### 1. Visualisation des Séries Temporelles

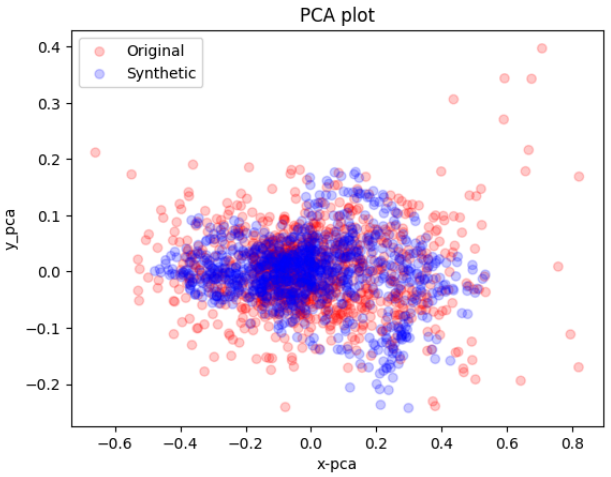



L'image ci-dessous présente la superposition des séries temporelles réelles et des séries temporelles générées par TimeGAN.

**Analyse :**

- Les séries synthétiques conservent la tendance globale des séries réelles.
- Toutefois, elles montrent une plus grande variabilité, suggérant que TimeGAN introduit une diversité qui pourrait être utile pour renforcer la robustesse des modèles de prédiction.

### 2. Analyse en Composantes Principales (PCA)

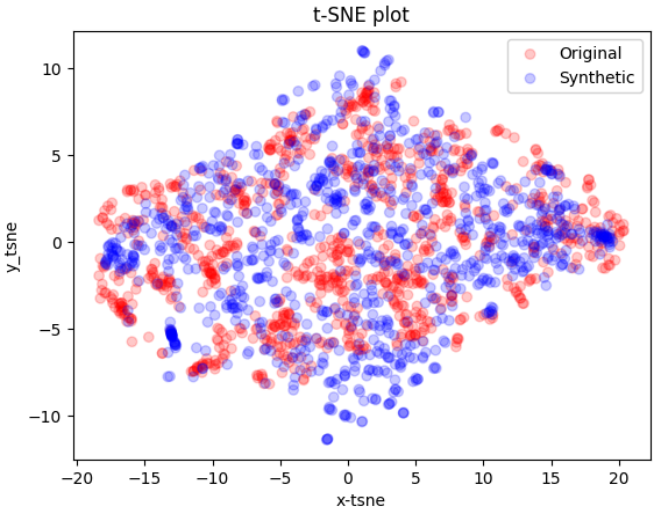


L'analyse PCA permet de visualiser la distribution des données réelles et synthétiques dans un espace réduit.

**Analyse :**

- La distribution des points synthétiques (en bleu) est globalement proche de celle des points réels (en rouge), indiquant que TimeGAN capture bien la structure des données originales.
- Cependant, certaines données synthétiques s'éloignent des clusters principaux, ce qui pourrait refléter une génération imparfaite ou une plus grande diversité.

**3. Analyse t-SNE**




Le t-SNE est utilisé pour réduire la dimensionnalité et visualiser les similarités entre les données réelles et synthétiques.

**Analyse :**

- La représentation t-SNE montre que les données TimeGAN sont bien mélangées avec les données réelles.
- On observe toutefois des regroupements différents qui pourraient être liés à des nuances que TimeGAN n'a pas parfaitement capturées.

## Conclusion et Perspectives

L'expérimentation avec TimeGAN a montré qu'il s'agit d'un outil puissant pour générer des séries temporelles synthétiques proches des données réelles.

**Points positifs :**

- Capacité à préserver les dynamiques temporelles.

- Augmentation de la variabilité des données, ce qui peut potentiellement améliorer la généralisation des modèles de prédiction.

**Limitations :**

- Une plus grande dispersion dans les données générées peut parfois perturber l'entraînement des modèles.
- Certains patterns spécifiques sont mal capturés.