# TemporalMLPSCM
## Synthetic Temporal Data Generator based on SCM Dynamics

Noé Amar

Neuralk AI

November 19, 2025

# Sommaire


1. Contexte & Motivation
2. Rappel : MLPSCM (TabICL)
3. Contribution : TemporalMLPSCM
4. Analyse temporelle (1 dataset)
5. Séparation multi-datasets
6. Résultats : bruit vs hyperparamètres
7. Limitations & Perspectives
8. Conclusion

# Motivation

**Problème :**

- Peu de datasets temporels **réalistes** disponibles.
- Difficile de **pré-entraîner** un modèle fondation temporel tabulaire.
- TabICL / TabPFN $\Rightarrow$ efficacité prouvée mais **pas de dynamique séquentielle**.

**Objectif :** créer un **générateur temporel contrôlable** permettant :

- Autoregression ($\alpha$)
- Périodicité ($\beta$, period)
- Bruit gaussien $\Rightarrow$ incertitude / réalisme
- Structure causale implicite (héritée du SCM)

# Rappel : MLPSCM (TabICL)

**Modèle original TabICL :**

$$h_k = MLP([h_{k-1}]) + \epsilon \Rightarrow X = h[\text{idx\_X}], \quad y = h[\text{idx\_y}]$$

**Forces :**

- Très efficace pour le tabulaire statique
- Diversité générée via causes + poids + bruit
- Peut générer des "mondes causaux" différents

**Limite majeure :**

$$X_t \quad \rightarrow \quad X_{t+1} \quad \text{(aucune dépendance temporelle)}$$

**Impossible à utiliser** pour forecasting / séquence / signaux réels.

# Objectif : introduire une dynamique temporelle

**Idée centrale :** conserver le SCM de TabICL et **y injecter du temps**

$$h_t^{(k)} = MLP([h_{t-1}^{(k)},\ \text{causes}_t]) + \alpha \cdot h_{t-1}^{(k)} + \beta \cdot h_{t-\text{period}}^{(k)} + \epsilon_t$$

**Éléments conservés du MLPSCM :**

- Architecture MLP (ModuleList)
- `XSampler` $\rightarrow$ causes latentes
- Sélection de features via idx_X / idx_y
- Bruit gaussien injecté à chaque timestep

# Tests appliqués à un dataset

**Analyse statistique de la temporalité :**

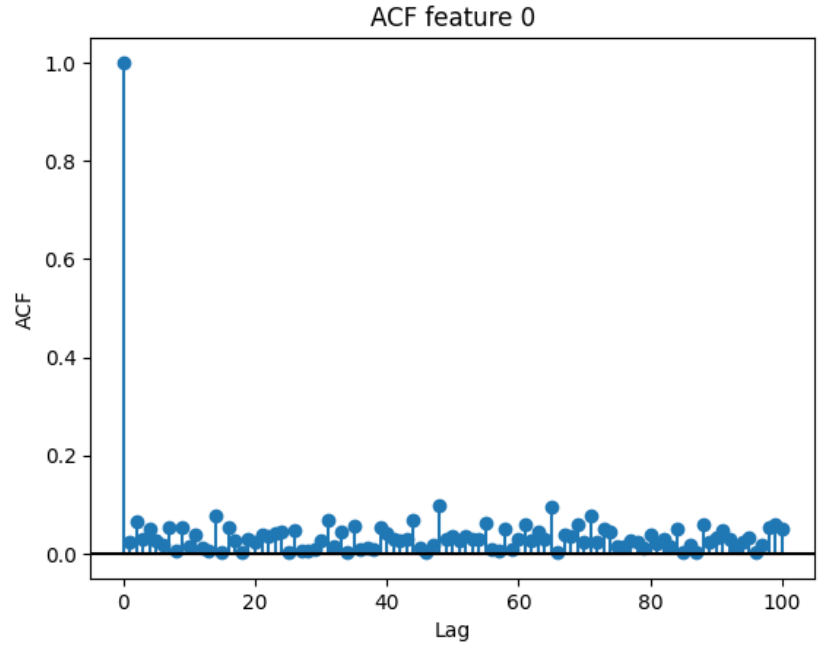
- **ACF** – mémoire temporelle
- **CCF** – influence entre features
- **ADF** – test de stationnarité
- **FFT / spectre** – fréquence dominante

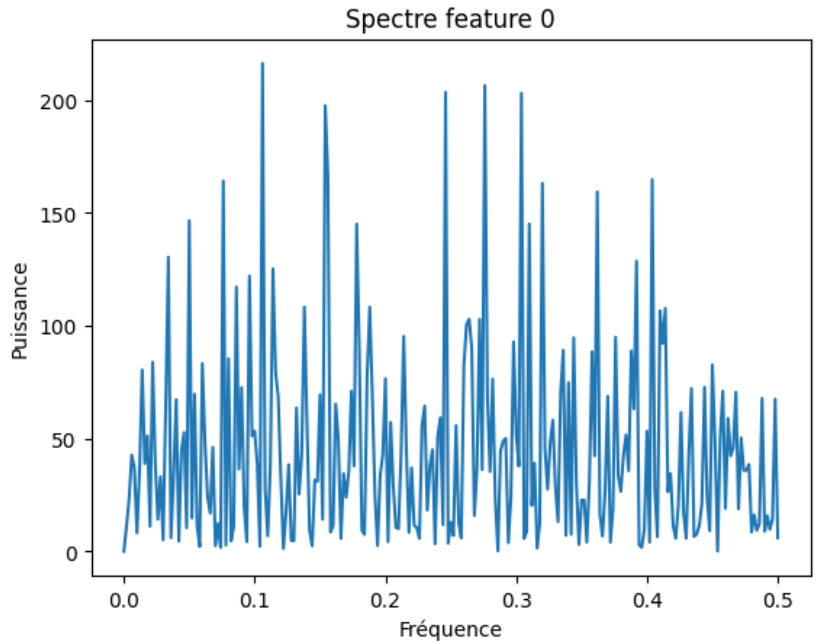
# MLPSCM Classique (sans temporalité)

**Propriétés du modèle original :**

- Pas de dépendance entre $t$ et $t-1$
- ACF $\approx 0$ dès le lag 1
- Spectre plat ( bruit blanc)
- **Dataset non exploitable pour forecasting**

$$h^{(k)} = MLP([h^{(k-1)}]) + \epsilon \quad \Rightarrow \quad X_t \to X_{t+1}$$



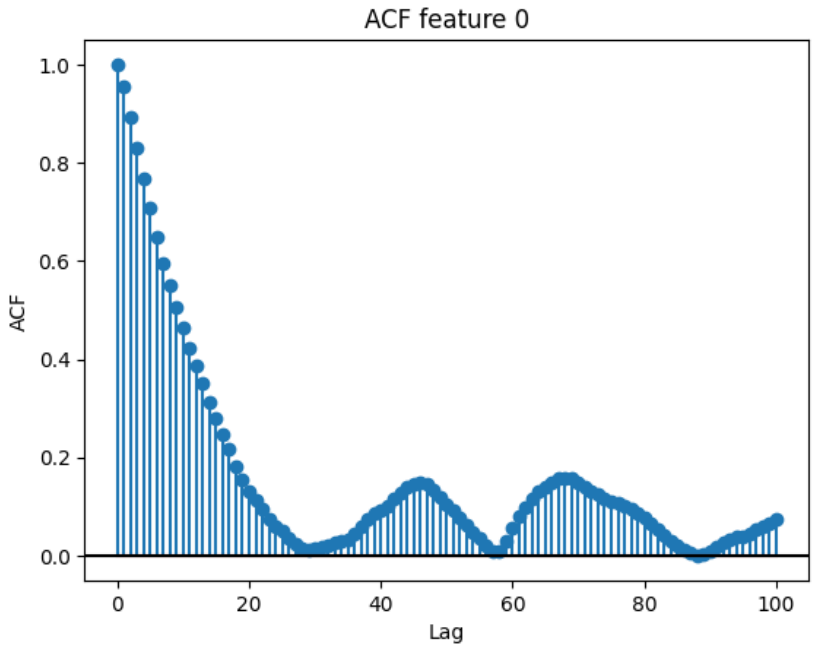



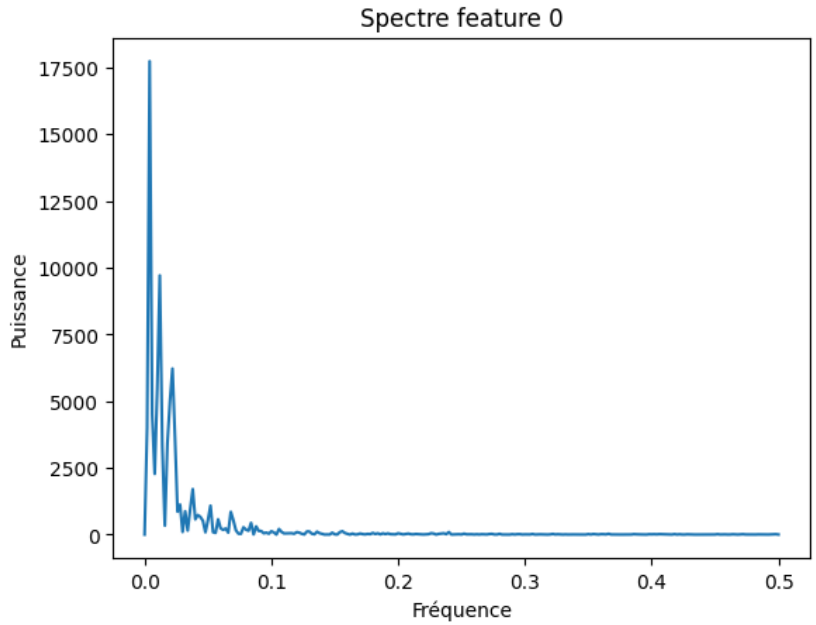
# Autoregression ($\alpha$)

**Effet de $\alpha$:**

- Mémoire temporelle
- Processus AR(1) implicite
- Décroissance lente de l'ACF

$$h_t^{(k)} = h_{\text{new}} + \alpha \cdot h_{t-1}^{(k)} + \epsilon_t$$



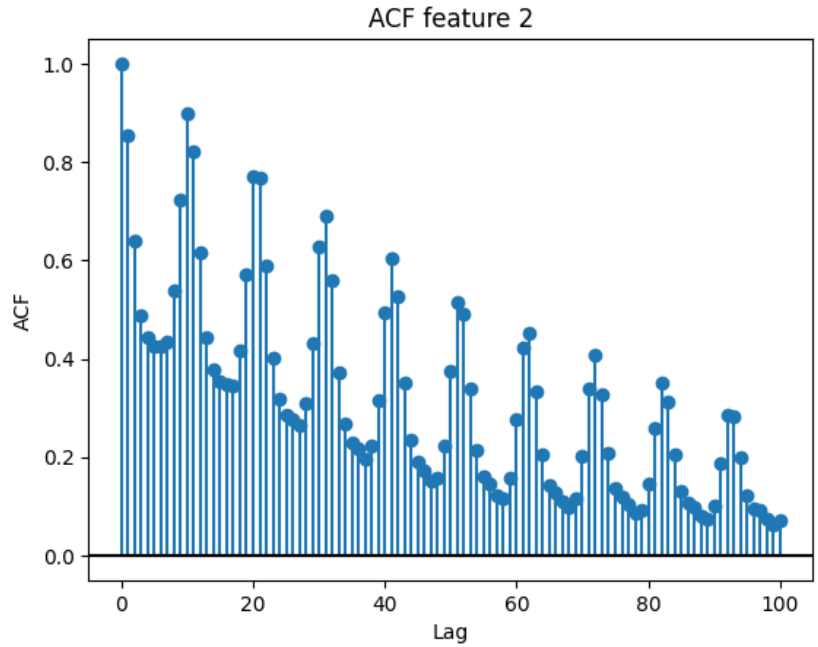



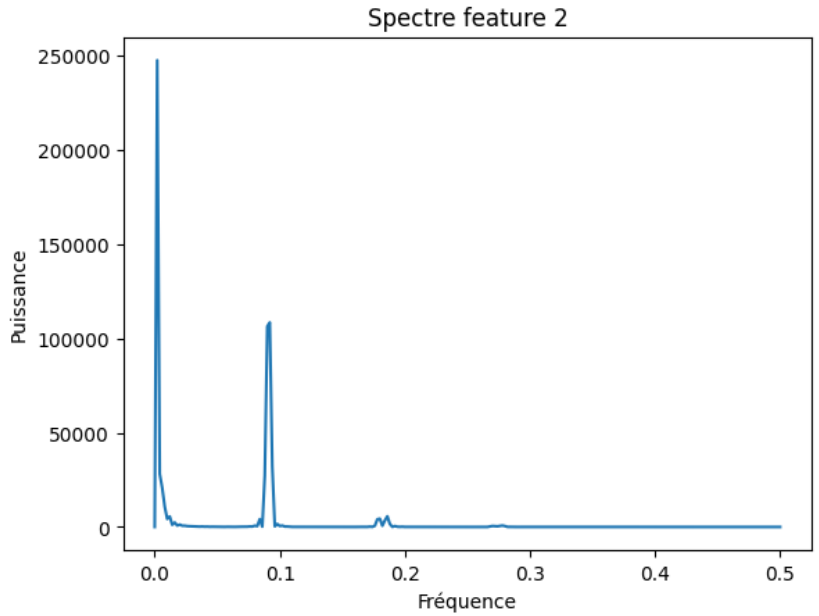
# Périodicité ($\beta$, period)

$$h_t^{(k)} = \cdots + \beta \cdot h_{t-\text{period}}^{(k)}$$

**Apport :**

- Capture d'un phénomène **cyclique**
- Pics clairs dans l'ACF (t = period)



# Pourquoi extraire des signatures ?

**Idée centrale :** Chaque dataset généré doit posséder une **identité propre**. Cela permet de constituer un **corpus riche et diversifié** pour pré-entraîner un *Time Series Foundation Model*.

**Une signature = empreinte mathématique du dataset**

- **Marginal** : forme des distributions (mean, std, skewness, kurtosis, percentiles)
- **Temporal** : dynamique du signal (ACF / decoherence time / fréquence dominante)
- **Structurelle** : interactions entre variables (corrélations inter-features)

**But final :** *Quantifier la diversité entre datasets → vérifier que le générateur crée plusieurs "mondes différents".*

# Comment comparer deux datasets ?

**Idée :** chaque dataset = une *empreinte mathématique* (signature). On peut donc mesurer leur **distance** pour détecter s'ils représentent **des mondes différents**.

$$\boxed{D[i,j] = \| s_i - s_j \|_2}$$

$\Rightarrow$ *plus D est grand, plus les datasets sont différents.*

**Utilisation :**

- Construction d'un **corpus varié** pour pré-training.
- Visualisation via **heatmaps** $\rightarrow$ **clusters de comportements**.
- Validation que le générateur ne produit pas toujours le même dataset.

# Diversité induite par le bruit

**Hyperparamètres fixés — seule stochasticité : bruit + init**

| **Type** | Mean | Min | Max | Std |
|---|---|---|---|---|
| Marginal | 4.57 | 2.63 | 8.81 | 1.34 |
| Temporal | 3.75 | 1.43 | 7.56 | 1.73 |
| Corr. | 0.22 | 0.16 | 0.34 | 0.04 |

# Diversité induite par les hyperparamètres

**Hyperparamètres perturbés :**

- $\alpha, \beta$, `period` (dynamique temporelle)
- activation (Tanh / ReLU / GELU)
- hidden_dim, dropout, bruit

**Résultats :**

| **Type** | Mean | Min | Max | Std |
|---|---|---|---|---|
| Marginal | 13.95 | 3.56 | 33.96 | 7.13 |
| Temporal | 18.95 | 3.88 | 51.63 | 12.79 |
| Corr. Structure | 0.42 | 0.23 | 0.62 | 0.09 |

**Interprétation :** Les hyperparamètres influencent fortement la *temporalité* des comportements, et modérément la *structure multivariée* — ce qui est idéal pour pré-entraîner un modèle fondation robuste.

# Limitations actuelles

- MLP uniquement — on aurait pu essayer d'autres architectures (trees, LSTM, Transformers...)
- Normalisation / clipping peut casser la dynamique
- Pas encore de downstream forecasting

# Conclusion

- Extension du MLPSCM vers le **temporel** avec AR et périodicité.
- **Deux sources de diversité** : bruit & hyperparamètres