ENSEA
Beyond Engineering

Prévision efficace de séries temporelles soutenue par la découverte causale : Une application au prix des actions

Quentin de La Chaise Naël Briand

ENSEA
Beyond Engineering

I - INTRODUCTION

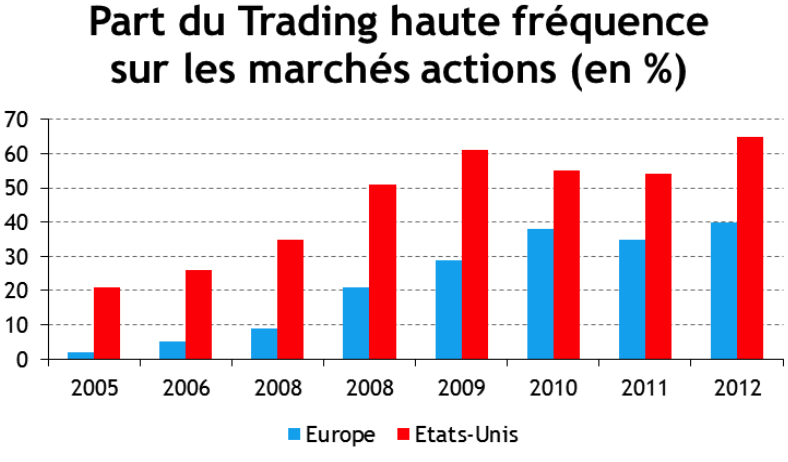
ENSEA
Beyond Engineering

I.A. Contexte du projet

Une quête ancienne pour comprendre et prévoir les marchés financiers
Marchés financiers (XVIIe siècle, 1ére bourses à Amsterdam et Londres). Les investisseurs et économistes cherchent à prévoir les fluctuations des prix des actifs. -> l'actualité !

Pourquoi prévoir les cours des actions aujourd'hui ?

- Prendre des décisions d'investissement : compréhension fine des dynamiques du marché,
- Maîtriser les volumes de données financières : trading haute fréquence, données en temps réel.
- Gérer les risques : la volatilité accrue.



Évolution du CAC40

ENSEA
Beyond Engineering

I.A. Contexte du projet

L'objectif -> modèles capables de lire directement les données financières en grande quantité et rapidement ! :

Optimisation des stratégies d'investissement :
- Des réseaux profonds peuvent capturer des "caractéristiques universelles" des carnets d'ordres. [1]

Amélioration des performances des modèles d'IA :
- Utilisation de méthodes modernes comme les architectures Seq2Seq et Attention, qui permettent de traiter des séquences longues et complexes (LSTM, Transformers). [2]

Réduction des risques financiers :
- Identifier les signaux d'alerte pour des crises potentielles. [3]

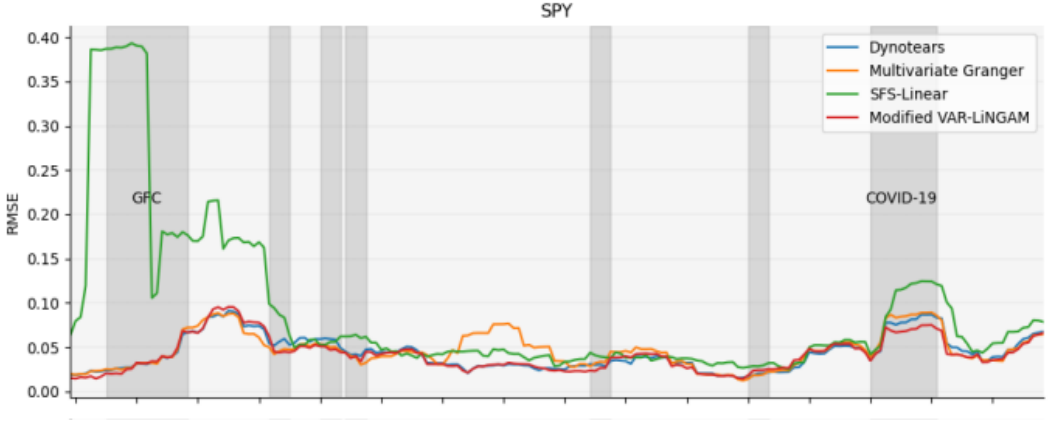
ENSEA
Beyond Engineering

I.B. Définition du problème étudié

Les modèles ne fonctionnent pas.

Difficultés et limites modèles classiques : changements de distribution (Ex : crise financière subprimes 2007-2008, covid 2020-2021,…), nombre excessif de paramètre qui influence le cours de l'action (économie, météo, dates, géopolitiques...)

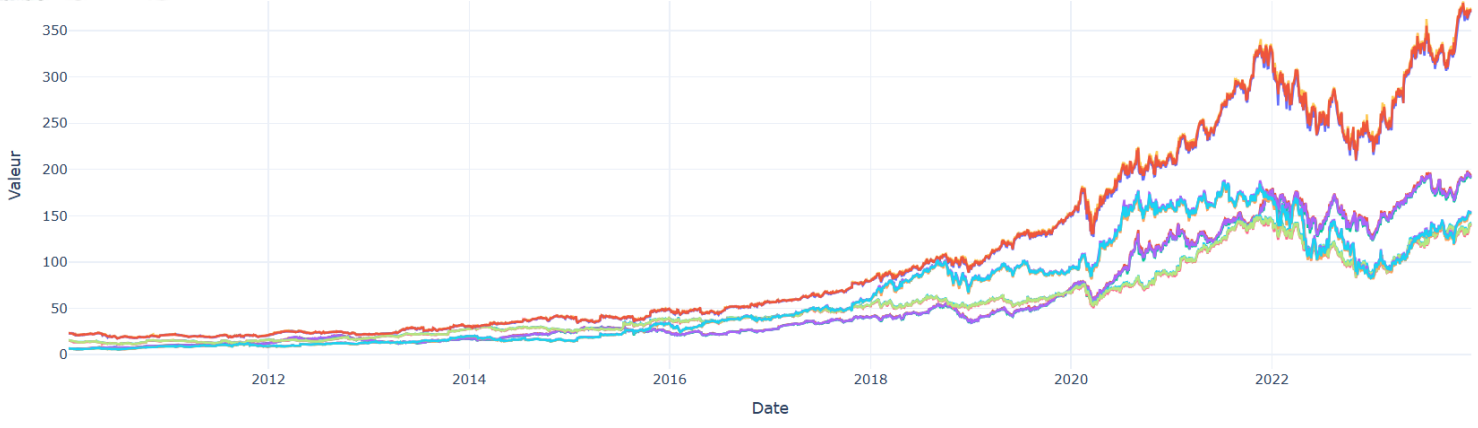
-> problème de variation de corrélation
-> problème de *spurious corrélations*
-> dépendance à beaucoup de paramètre



Comment rendre ces modèles plus robustes au changement de distibution tout en étant sensible a plus de paramétre ?

CAUSALITÉ > CORRÉLATION

I.C. Définition des termes du sujet



- Série temporelle (financières)
- Multivariée
- Instantanée
- Acyclique

Exemple cycle : Open(t)-> Close(t)-> Open(t) => Boucle cyclique

- Distribution non-gaussienne
- Série temporelle discrete

ENSEA
Beyond Engineering

II - MÉTHODOLOGIE

ENSEA
Beyond Engineering

II.A. Méthode de découverte causale

Qu'est-ce que la découverte causale ?

- Identifier les relations de cause à effet (application : neuroscience [9], finance, économie, marketing)

Causalité VS Corrélation

- **Causalité** : Un évènement A entraîne directement un changement dans un évènement B

Ex : augmentation taux d'intérêt → diminution des investissements.

- **Corrélation** : Deux évènements statistiquement liés, varient ensemble.

Ex (étude) : Corrélation taille des pieds et compétence en orthographe. [8]

→ Spurious corrélation

Modèles de découvertes causales

- Multivariate Granger Causality, VAR-LiNGAM, Dynotears, PCMCI.

→ VAR-LiNGAM [3], PCMCI.

ENSEA
Beyond Engineering

II.B. Algorithme VAR-LiNGAM

Définition :

Vector Autoregressive Linear Non-Gaussian Acyclic Model (VAR + LiNGAM)
VAR
LiNGAM

Hypothèses du modèle :

Linéarité
Non-Gaussianité
Acyclicité
Pas de variables latentes : aucune influence cachée ou non observée

Modèle mathématique :

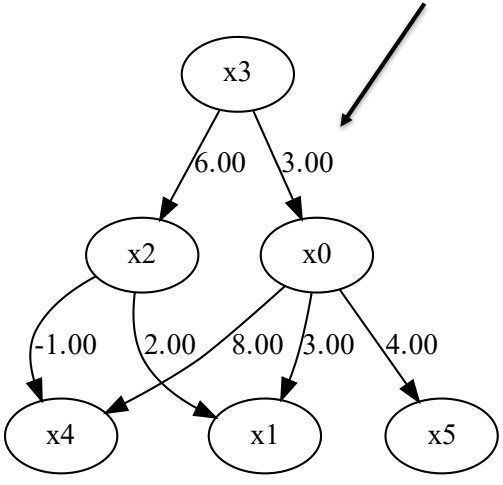
$x(t) = \sum_{\tau=0}^{N} B_\tau x(t-\tau) + e(t)$

- $x(t)$: Vecteur des variables observées à l'instant t.
- B_τ : Matrice de coefficients décrivant l'effet des variables à l'instant $t-\tau$ sur les variables à t.

$b_{i,j}^{(\tau)}$: $\text{variable\_j}(t-\tau) \rightarrow \text{variable\_i}(t)$

- $e(t)$: Vecteur des termes d'erreur résiduels à t, supposés non gaussiens, indépendants et sans causes communes.

Coefficient de causalité : indiquant la force et la direction de l'influence d'une variable sur une autre =/= p_values

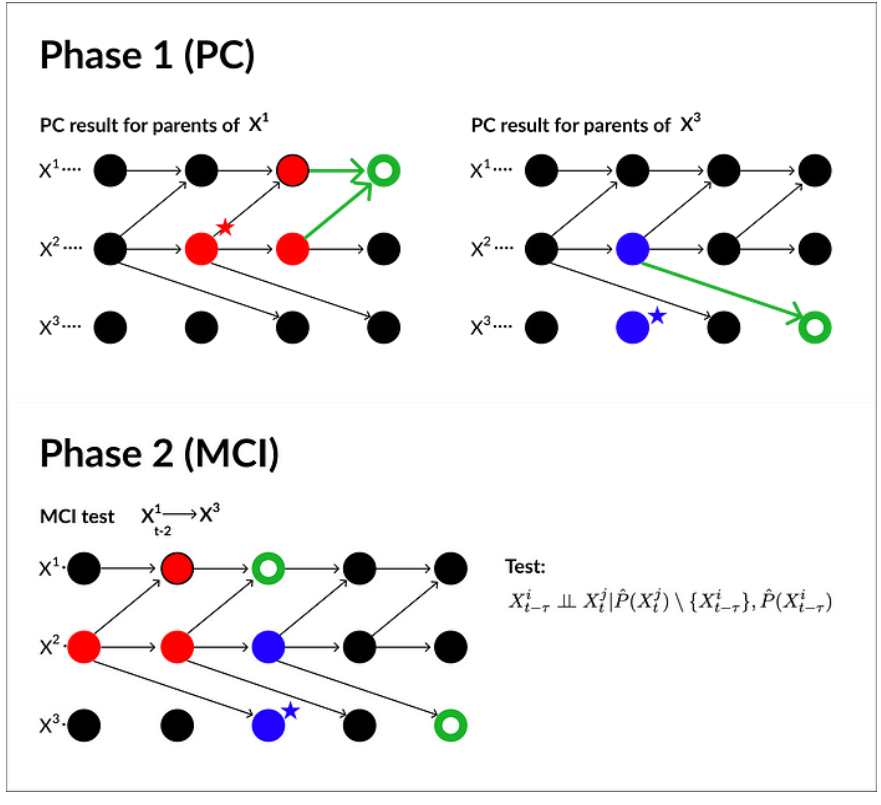


II.C. Algorithme PCMCI

PCMCI : Découverte causale basée sur des contraintes

PC1 : Variante robuste de l'algorithme PC
- Identification de la structure causale initiale
- Tests d'indépendance conditionnelle
- Contraintes temporelles

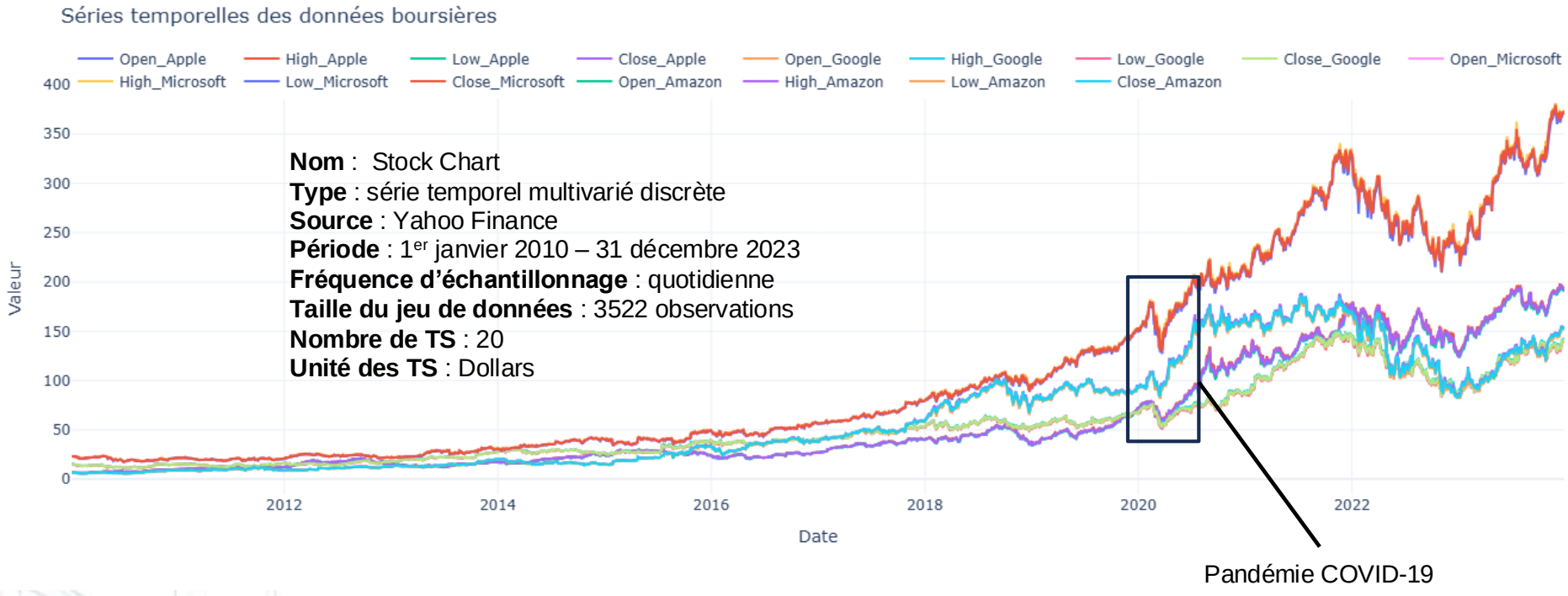
MCI (Momentary Conditional Independence)
- Affinement des liens causaux à des décalages temporels spécifiques
- Contrôle des auto-corrélations et faux positifs
- P-values ajustées (FDR)



III – DESCRIPTION DE LA DONNÉE

ENSEA
Beyond Engineering

III.A. Dataset et variables



Nom : Stock Chart
Type : série temporel multivarié discrète
Source : Yahoo Finance
Période : 1er janvier 2010 – 31 décembre 2023
Fréquence d'échantillonnage : quotidienne
Taille du jeu de données : 3522 observations
Nombre de TS : 20
Unité des TS : Dollars

Série temporelle cible : Close_Google
Horizon de prédiction : 1 Mois
Features possibles : Open, High, Low, Close, Volume de chaque action – Close_Google

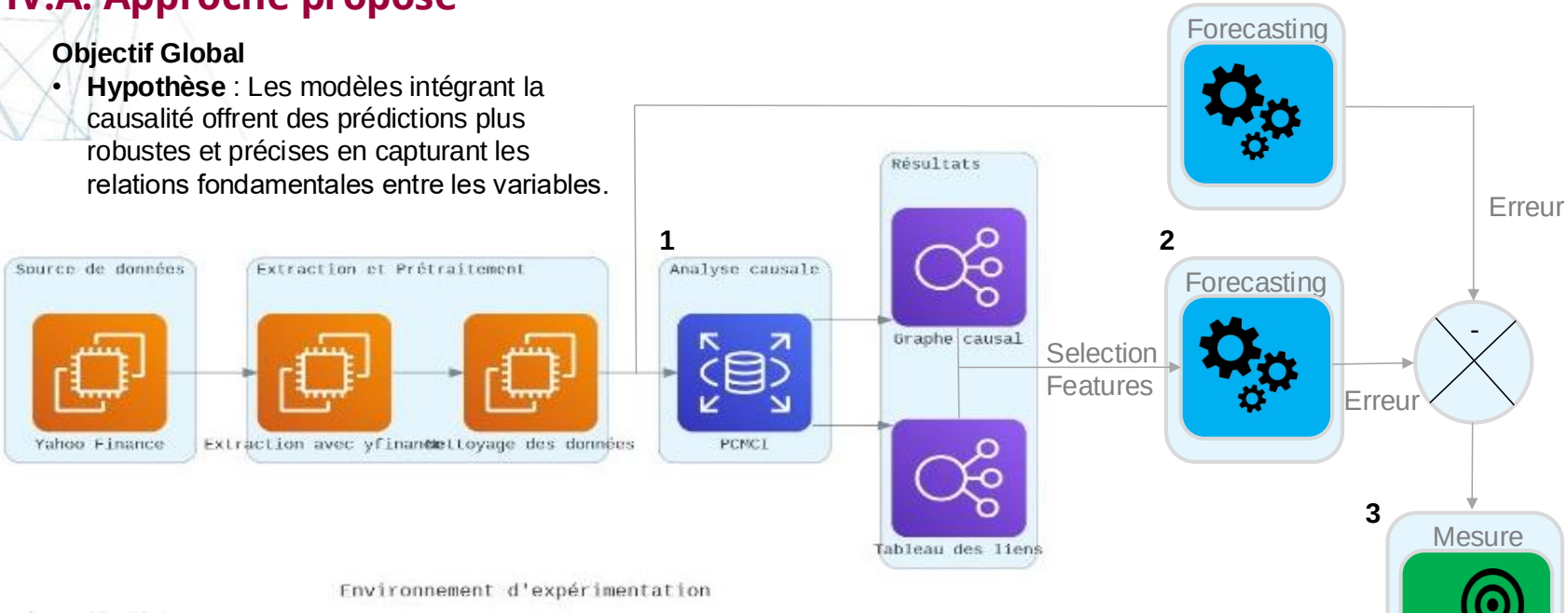
IV – ENVIRONNEMENT D'EXPÉRIMENTATION

ENSEA
Beyond Engineering

IV.A. Approche proposé

Objectif Global

- **Hypothèse** : Les modèles intégrant la causalité offrent des prédictions plus robustes et précises en capturant les relations fondamentales entre les variables.



1. Découverte Causale
relations de causalité entre les variables

Outils Utilisés :

- **VAR-LiNGAM**
- **PCMCI**

2. Prévision Basée sur la Causalité

- **Idée Principale** : Selection des features par leurs impact causal

Modèles de Prédiction :

- Approches naïve: Random Forest.
- Approches avancées : LSTM

3.Comparaison
Métriques Utilisées :

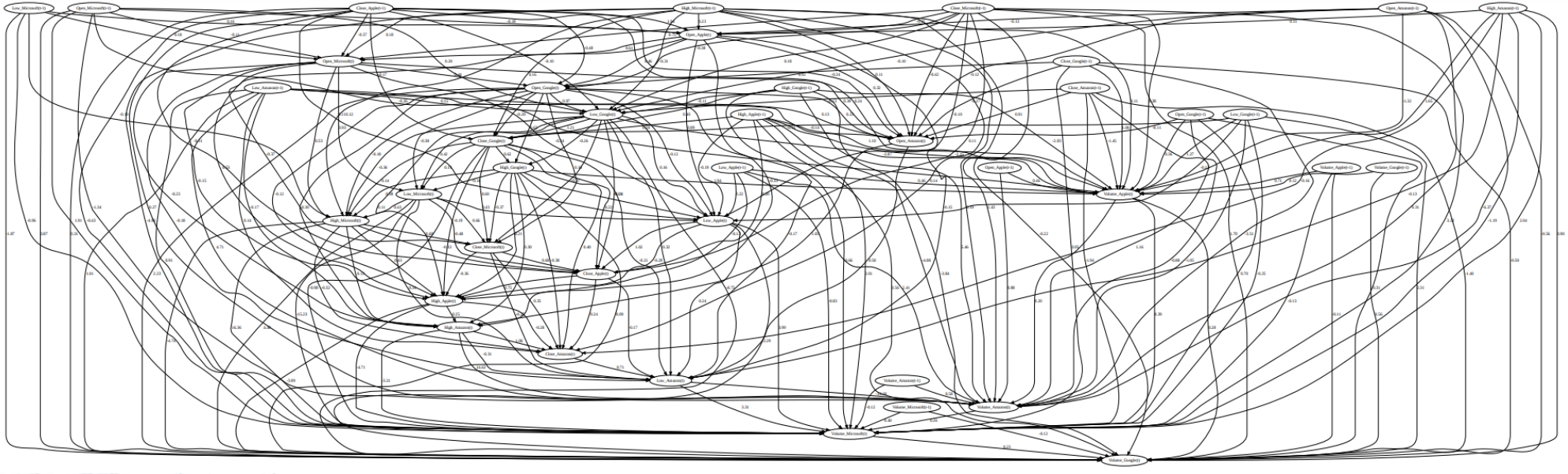
- Erreur MSE.
- Erreur MAE.

Configurations Testées :

- Avec/Sans causalité.
- Horizons temporels courts/longs.

IV.B. Expérience VAR-LiNGAM
→ Analyse Causale

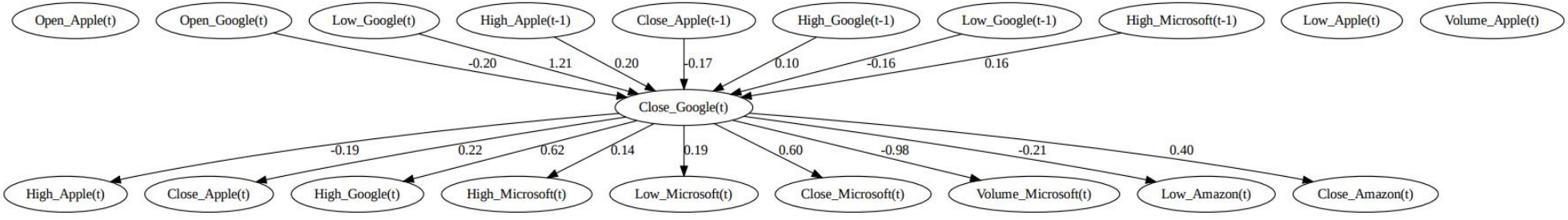
```
model = VARLiNGAM(lags=30, criterion='bic')
```



Graphe de l'analyse causale résultant de l'entraînement de VAR-LiNGAM avec les hyperparamètres ci-dessus.

ENSEA
Beyond Engineering

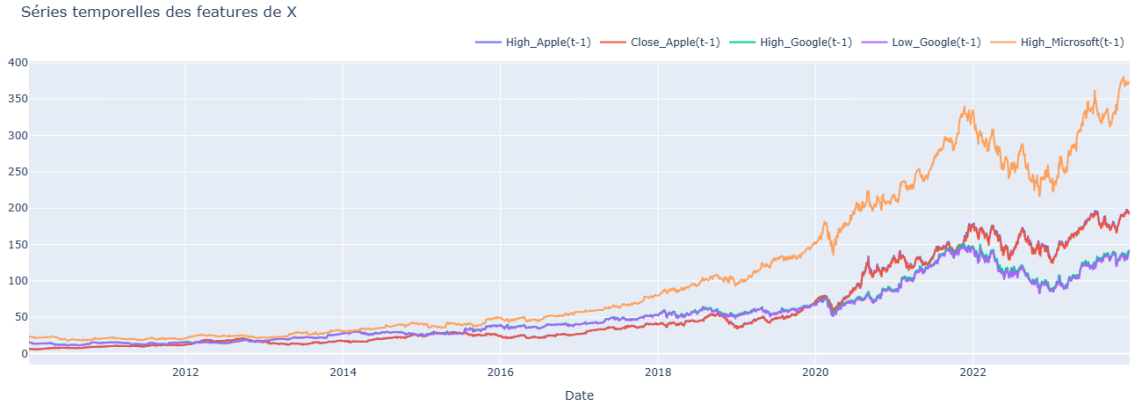
IV.B. Expérience VAR-LiNGAM
→ Analyse Causale



Features sélectionées :
- High_Apple
- Close_Apple
- High_Google
- Low_Google
- High_Microsoft

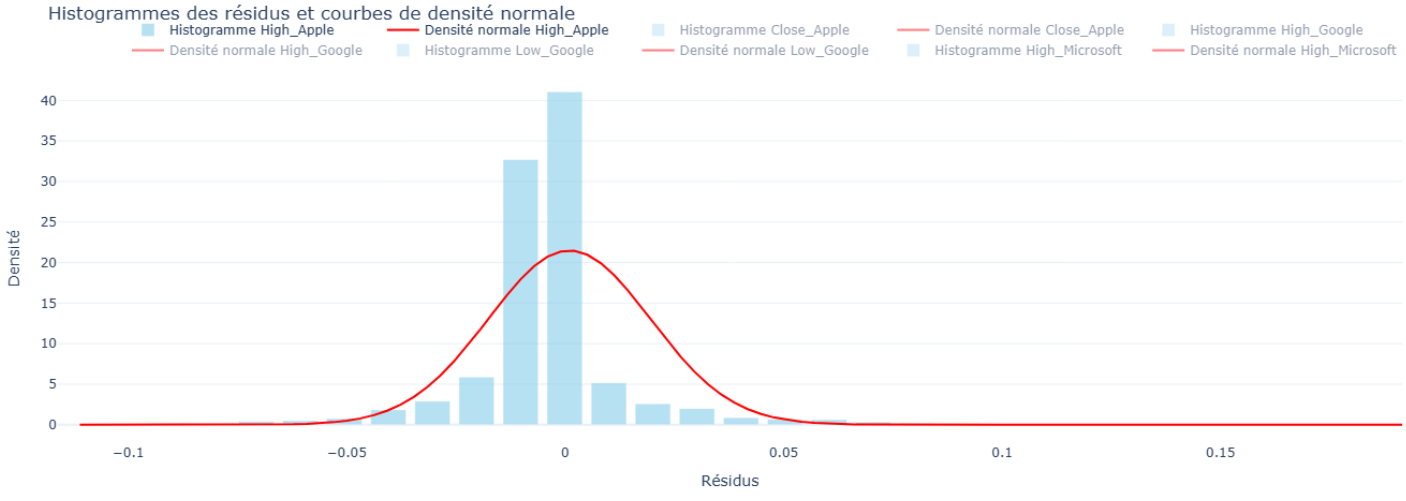
Pourquoi ne pas sélectionner Open_Google(t) et Low_Google(t) ?

→ Data leakage → Résultats faussés
Modèle trop optimiste



IV.B. Expérience VAR-LiNGAM
→ Vérifications des hypothèses

Rappels des hypothèses du modèle : Linéarité, Non-Gaussianité, Acyclicité, Pas de variables latentes.

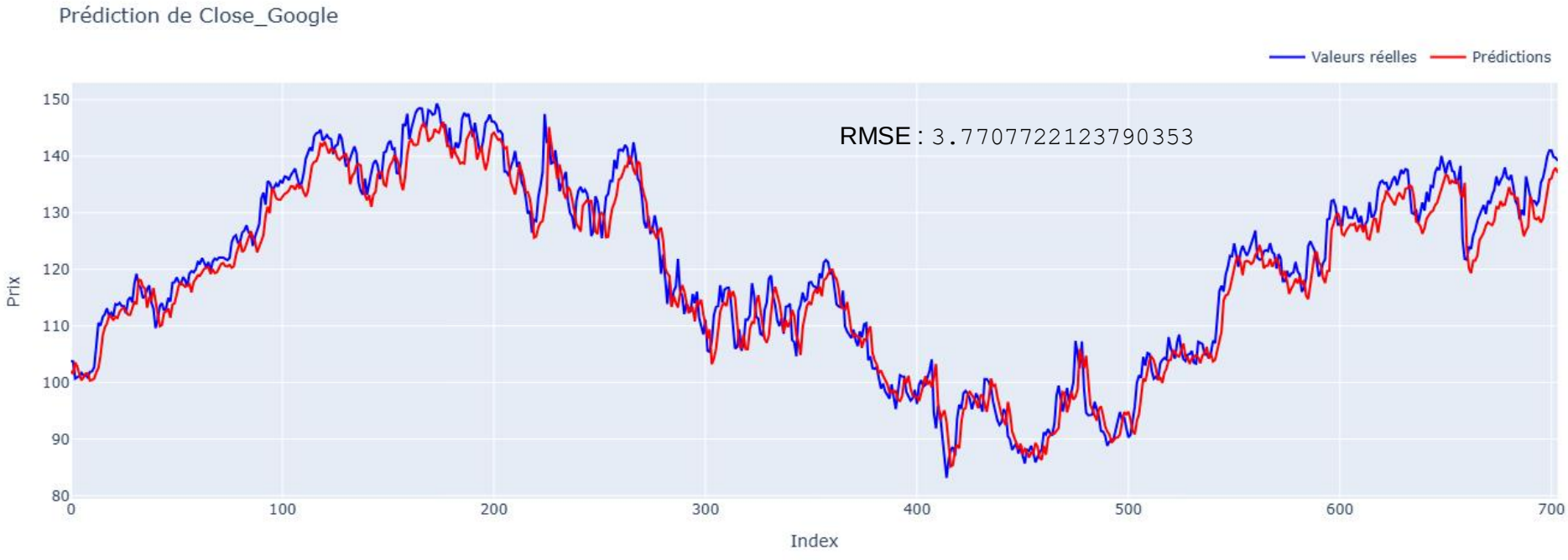


Résultats des tests de normalité

Variable	Shapiro-Wilk Statistique	Shapiro-Wilk p-valeur	Kolmogorov-Smirnov Statistique	Kolmogorov-Smirnov p-valeur
High_Apple	0.7587	0.0000	0.2078	0.0000
Close_Apple	0.7826	0.0000	0.1945	0.0000
High_Google	0.7326	0.0000	0.1922	0.0000
Low_Google	0.7671	0.0000	0.1837	0.0000
High_Microsoft	0.7432	0.0000	0.2124	0.0000

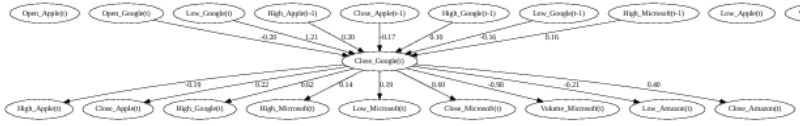
P_values < 0.05 et Statistique élevé donc Rejet systématique de l'hypothèse de normalité des résidus : Hypothèses vérifiées !

IV.B. Expérience VAR-LiNGAM
→ Entraînement d'un LSTM avec découverte causale

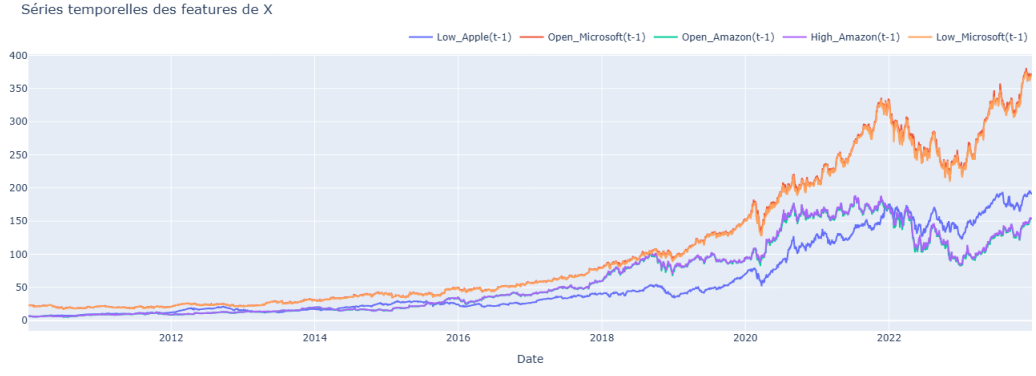


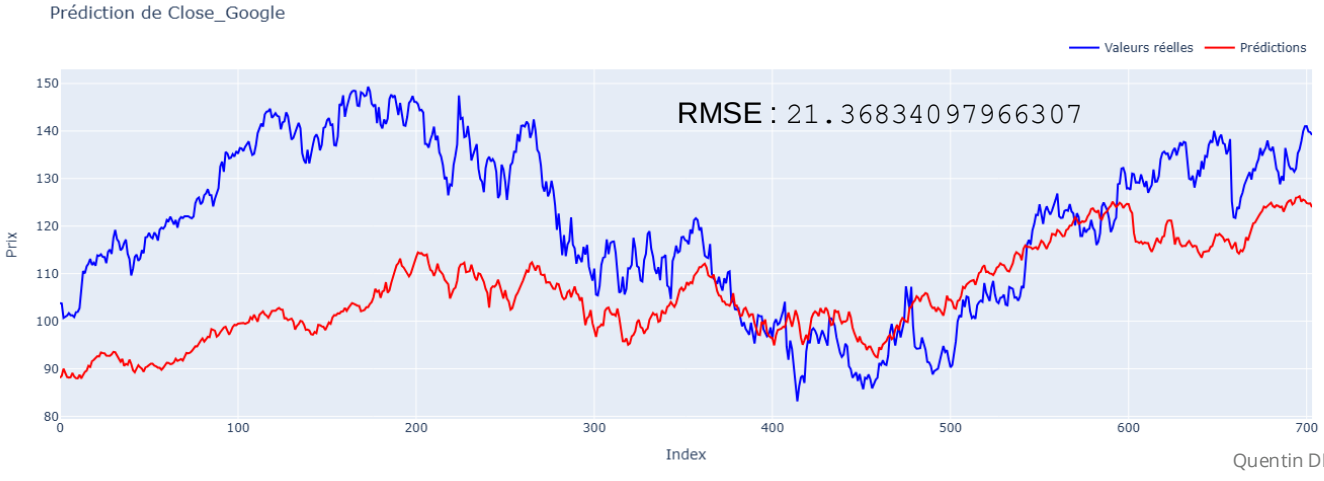
Erreur de 4 dollars : Modèle précis !

IV.B. Expérience VAR-LiNGAM
→ Entraînement d'un LSTM sans découverte causale



Features sélectionnées :
- Low_Apple(t-1)
- Open_Microsoft(t-1)
- Open_Amazon(t-1)
- High-Amzon(t-1)
- Low_Microsoft(t-1)





Amélioration de 82% grâce à la causalité

IV.B. Expérience PCMCI

Paramètres de PCMCI :

- Retard maximal analysé (**tau_max**) : 30 (jours)
- Métrique de dépendance : **ParCorr** (Partial Correlation) : Analytic
- Seuils de signification statistique (**p-value**) : 0.05.

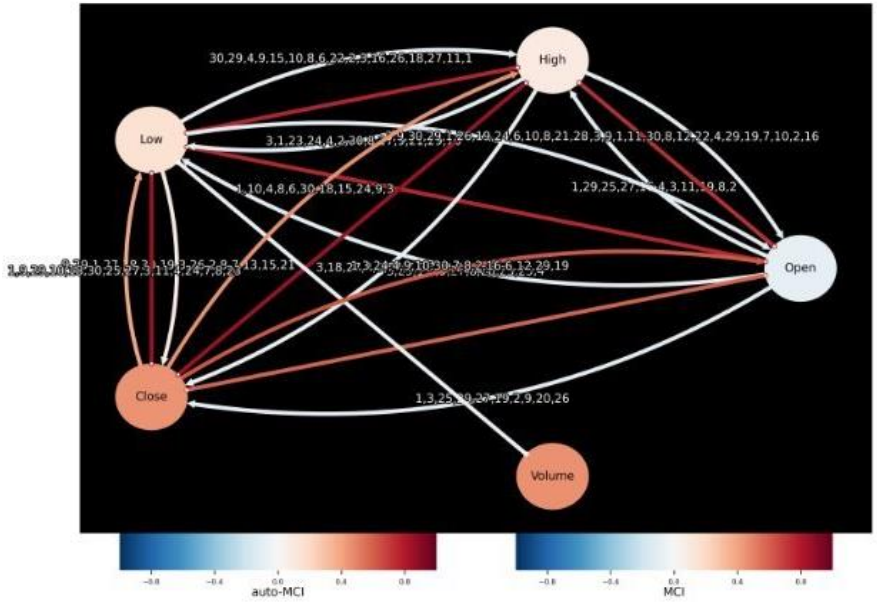
Résultats de l'Analyse Causale

Graphe Causal :

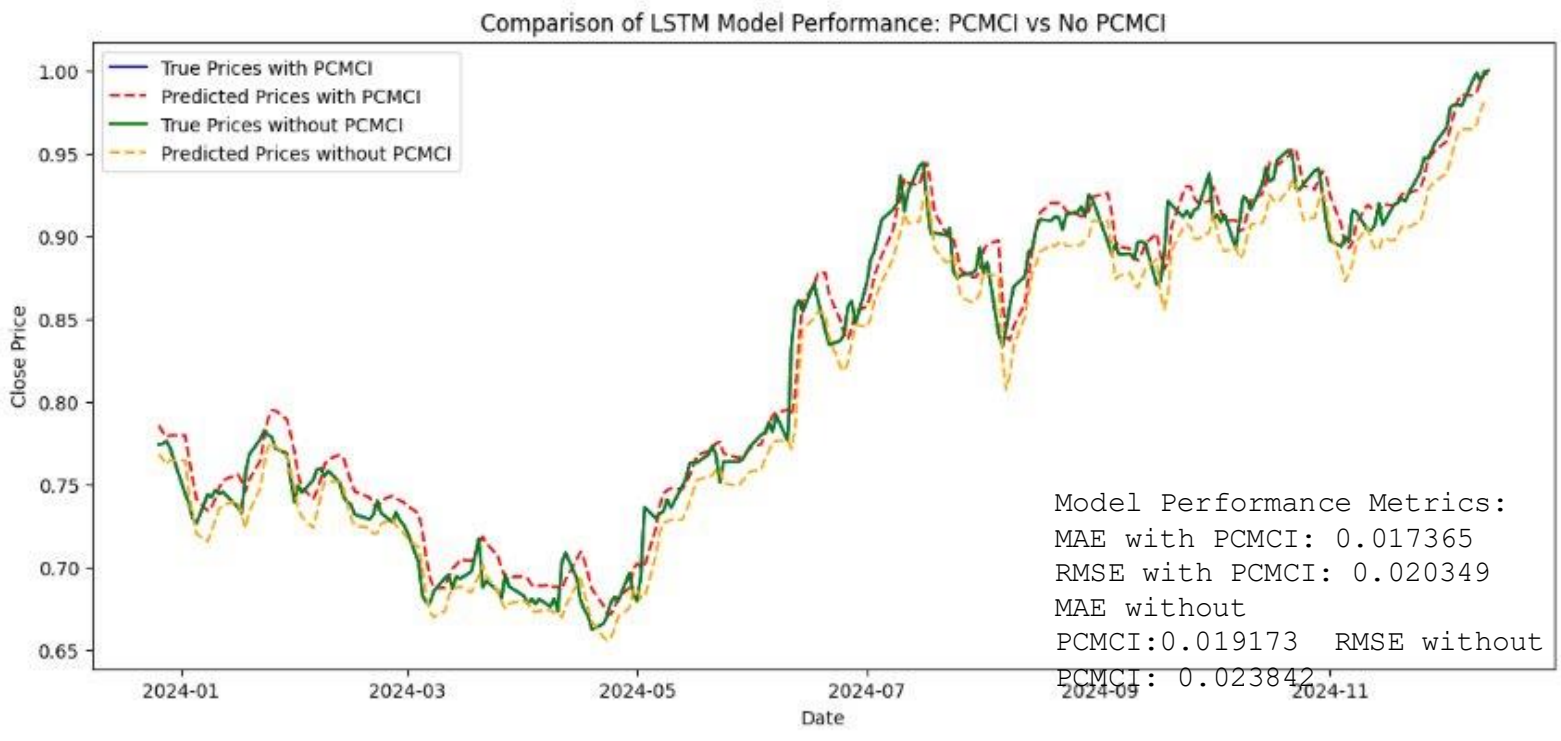
- Les Volumes sont retirés
- Low_Apple
- Open_Microsoft
- Open_Amazon
- High-Amzon
- Low_Microsoft (20 features --> 5 features)

Limites Observées

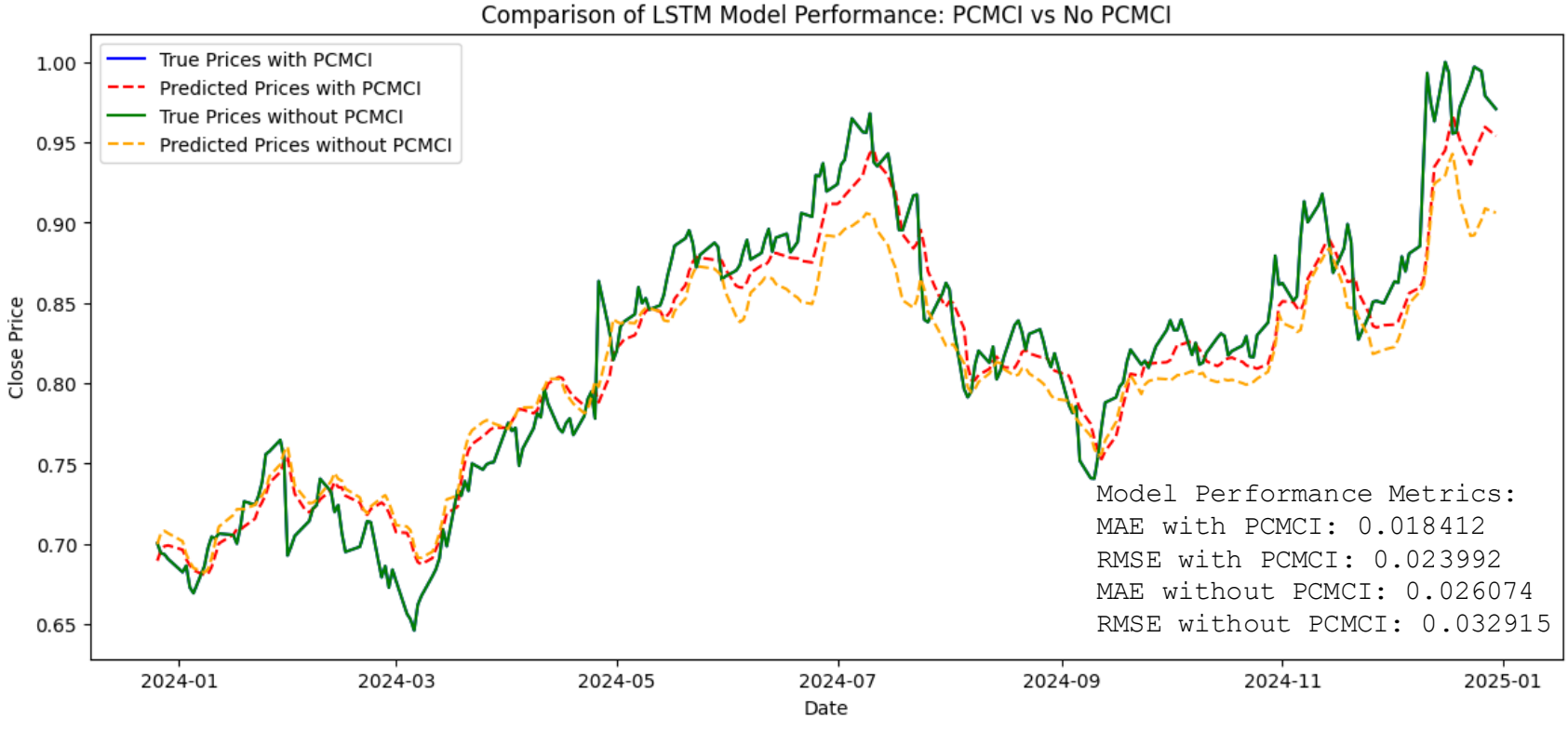
•**PCMCI** : Complexité augmentant avec le nombre de variables et de lags.



IV.B. Expérience PCMCI



IV.B. Expérience PCMCI



ENSEA
Beyond Engineering

V – RÉSULTATS

ENSEA
Beyond Engineering

V. Resultat

1. Comparaison des Performances Prédictives.

- **Métriques Utilisées** :
- **Erreur quadratique moyenne (RMSE)**

Modèle	Horizon court (VAR-LiNGAM)	Horizon long (PCMCI)
Modèle sans causalité	21,36 (RMSE)	0.032915 (RMSE)
Modèle avec causalité	3,78 (RMSE)	0.023992 (RMSE)
Gain (%) avec causalité	82%	37%

2. Remarques

- Les modèles intégrant la causalité (PCMCI, VAR-LiNGAM) surpassent les modèles standards en précision, surtout pour **Horizon court.** Les 2 modèles présentent une meilleure robustesse par rapport à la volatilité.

ENSEA
Beyond Engineering

VI – CONCLUSION ET PERSPECTIVES

ENSEA
Beyond Engineering

VI. CONCLUSION ET PERSPECTIVES

Résumé des Contributions :
- Découverte causale sur des séries temporelles financières.
- Amélioration des performances sur des horizons courts et longs.
- Visualisation et interprétation des graphes causaux pour mieux comprendre les dépendances entre les variables.

Résultats Important :
- **Avec causalité** : Réduction des erreurs de prédiction (jusqu'à 82% pour les horizons courts).
- **Sans causalité** : Modèles plus sensibles au bruit et aux corrélations spurielles.
- Réduction du nombre de features : taux de compression de 75%

Limites Observées :
- Performances réduites sur les horizons longs (dynamique imprévisible des marchés financiers).
- Dépendance à la qualité des données et des graphes causaux.

Perspectives
- **Améliorations Méthodologiques** :
- Modèles non linéaires pour mieux capturer les relations complexes (Attention, Transformers).
- Automatisés la génération de graphes causaux robustes.
- **Chronoepilogi**

ENSEA
Beyond Engineering

RÉFÉRENCES

[1] Zhang, Z., Zohren, S., & Roberts, S. (2019). DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. *IEEE Transactions on Signal Processing*, 67(11), 3001-3012.
[2] Tang, Y., Yang, P., & Zhang, Y. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, 518, 425-436.
[3] Oliveira, D. C., Lu, Y., & Lin, X. (2024). Causality-Inspired Models for Financial Time Series Forecasting. *arXiv preprint arXiv:2408.09960*.
[4] Zhang, Y., & Zohren, S. (2021). Multi-Horizon Forecasting for Limit Order Books: Novel Deep Learning Approaches and Hardware Acceleration using Intelligent Processing Units. *arXiv preprint arXiv:2106.01988*.
[5] Arsac, L., & Spies, T. (2021). Causal Discovery for Time Series: PMINE. *Proceedings of the 2021 ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 1234-1242.
[6] Börjesson, S., & Ul Hassan, M. (2020). Forecasting Financial Time Series through Causal and Dilated Convolutions. *Entropy*, 22(11), 1234.
[7] Zaremba, A., & Shemer, K. (2023). Assessing Causality in Financial Time Series. *Journal of Financial Econometrics*, 21(2), 345-367.
[8] Herbelot, T. (2021). TD Causalité
[9] Shimizu, S., et al. (n.d.). LiNGAM: Applications and Tailor-Made Methods. Shimizu Lab.

ENSEA
Beyond Engineering

quentin.delachaise@ensea.fr

nael.briand@ensea.fr