GUIDE COMPLET (PAS A PAS) — ANALYSE DE SERIES TEMPORELLES

Portée générale des séries temporelles

Les méthodes de séries temporelles sont souvent associées uniquement aux données financières (cours boursiers, taux d'intérêt, taux de change). Cette perception est restrictive. En réalité, le critère fondamental d'application des séries temporelles n'est pas le domaine étudié, mais la présence d'une dimension temporelle dans les données.

Toute variable observée de manière ordonnée dans le temps peut être analysée comme une série temporelle, qu'elle provienne de l'économie, de la santé, de l'énergie, du climat, de l'industrie ou encore des plateformes numériques. Par exemple, le nombre quotidien d'hospitalisations, la consommation électrique horaire, la température journalière ou les ventes quotidiennes d'un produit relèvent tous de l'analyse de séries temporelles.

Ainsi, les séries temporelles constituent un cadre méthodologique général destiné à analyser, modéliser et prévoir l'évolution de phénomènes observés dans le temps, indépendamment de leur nature financière ou non.

1) Objectif du travail

Avant toute modélisation, définis ce que tu veux obtenir : (1) prévoir la variable cible, (2) comprendre la dynamique (tendance/saison), ou (3) détecter des anomalies. Tu dois aussi fixer un horizon de prévision h (ex : 7 jours, 6 mois).

2) Comprendre la structure des données

Une série temporelle est une suite d'observations indexées dans le temps :

$y_1, y_2, \ldots, y_t$

Ce que tu vérifies : (a) une colonne date/temps, (b) une colonne valeur, (c) des dates triées, et (d) une fréquence claire (journalière, mensuelle, etc.).

3) Nettoyage & préparation (pourquoi c'est indispensable)

3.1 Trier, vérifier la fréquence et les trous

Pourquoi : ARIMA/SARIMA et les tests statistiques supposent une série à pas de temps régulier. Si tu as des dates manquantes ou des doublons, il faut corriger avant de tester ou modéliser.

3.2 Valeurs manquantes (missing values)

Pourquoi : les NA peuvent casser les tests (ADF/KPSS) et provoquer des erreurs ou des paramètres biaisés. Résultat attendu : une série sans trous, ou une stratégie documentée (suppression/interpolation).

3.3 Stabiliser la variance (si l'amplitude augmente avec le niveau)

Quand : si la série « grossit » (variabilité plus forte quand le niveau est haut).

Transformation fréquente :

$z_t = \log(y_t)$, $t = 1, 2, \ldots, T$

4) Décomposition : tendance, saisonnalité, bruit (comment reconnaître une série saisonnière)

Pourquoi : savoir s'il faut un modèle saisonnier (SARIMA) ou non, et quelles transformations appliquer. La saisonnalité est un motif qui se répète toutes les s périodes (ex : mensuel avec cycle annuel → $s = 12$).

Deux formes classiques :

Additif : $y_t = T_t + S_t + e_t$

Multiplicatif : $y_t = T_t \times S_t \times e_t$

Interprétation : modèle additif si l'amplitude saisonnière est à peu près constante. Multiplicatif si l'amplitude saisonnière grandit avec le niveau (souvent après log, on revient à l'additif).

5) Stationnarité (étape clé avant ARIMA)

Une série est stationnaire si sa moyenne et sa variance ne changent pas au cours du temps, et si la dépendance dépend seulement du retard (lag).

Pourquoi : AR/MA/ARIMA supposent une structure stable à travers le temps.

5.1 Décider si c'est stationnaire : tests + logique

ADF (Augmented Dickey-Fuller) : H_0 = « non stationnaire ». Si p-value < 0.05 → on rejette H_0 → stationnaire.

KPSS : H_0 = « stationnaire ». Si p-value < 0.05 → on rejette H_0 → non stationnaire.

Bon réflexe : utiliser ADF et KPSS ensemble (ils se complètent).

6) Rendre la série stationnaire : différenciation (d) et différenciation saisonnière (D)

6.1 Différence simple (enlève souvent la tendance)

Formule :

$\Delta y_t = y_t - y_{t-1}$

Si la tendance reste, on peut différencier une 2e fois :

$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$

Le nombre de différences simples appliquées correspond au paramètre d dans ARIMA(p,d,q).

6.2 Différence saisonnière (enlève la saisonnalité)

Si la série est saisonnière de période s (ex : s = 12), on utilise :

$\Delta_s y_t = y_t - y_{t-s}$

Le nombre de différences saisonnières correspond au paramètre D dans SARIMA.

7) Identifier p et q : ACF / PACF (les fameux « pics » expliqués)

ACF : corrélation entre la série et ses retards.

Définition : Corr(y_t, y_{t-k}).

PACF : corrélation « directe » avec le lag k en retirant l'effet des lags intermédiaires.

Règles pratiques :

• PACF coupe net à p → candidat AR(p)

• ACF coupe net à q → candidat MA(q)

• ACF et PACF décroissent lentement → candidat ARMA/ARIMA

Résultat attendu : 2 à 4 modèles candidats (pas 20).

8) Modèles et formules (AR, MA, ARIMA, SARIMA)

8.1 AR(p) :

$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$

8.2 MA(q) :

$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta\_q \varepsilon_{t-q}$

8.3 ARIMA(p,d,q) (forme opérateur) :

$\Phi(L)(1 - L)^d \, y_t = \Theta(L)\, \varepsilon_t$

où $\Phi(L)$ et $\Theta(L)$ sont des polynômes de retards (lags).

8.4 SA RIMA(p,d,q)(P,D,Q,s) :

$\Phi(L)\Phi_s(L^s)(1 - L)^d(1 - L^s)^D \, y_t = \Theta(L)\Theta_s(L^s)\, \varepsilon_t$

Cette forme ajoute une structure AR/MA saisonnière et une différenciation saisonnière.

9) Choisir le meilleur modèle : AIC/BIC + validation temporelle

AIC/BIC : plus petit = meilleur compromis qualité/complexité.

Important : validation temporelle (train = passé, test = futur).

On évalue sur le test avec MAE / RMSE / MAPE.

10) Diagnostics (résidus = bruit blanc)

Après estimation, les résidus doivent ressembler à du bruit blanc : moyenne proche de 0, pas d'autocorrélation, variance stable.

Test de Ljung-Box : H_0 = résidus indépendants (p-value > 0.05 → OK).

11) Prévision

On produit des prévisions $\hat{y}_{t+h}$ avec intervalles de confiance.

Les horizons courts sont généralement plus fiables.

Toujours commenter l'incertitude (intervalles).