Exploration des données

CM3 : Séries temporelles

Mickaël Martin Nevot

V1.0.0

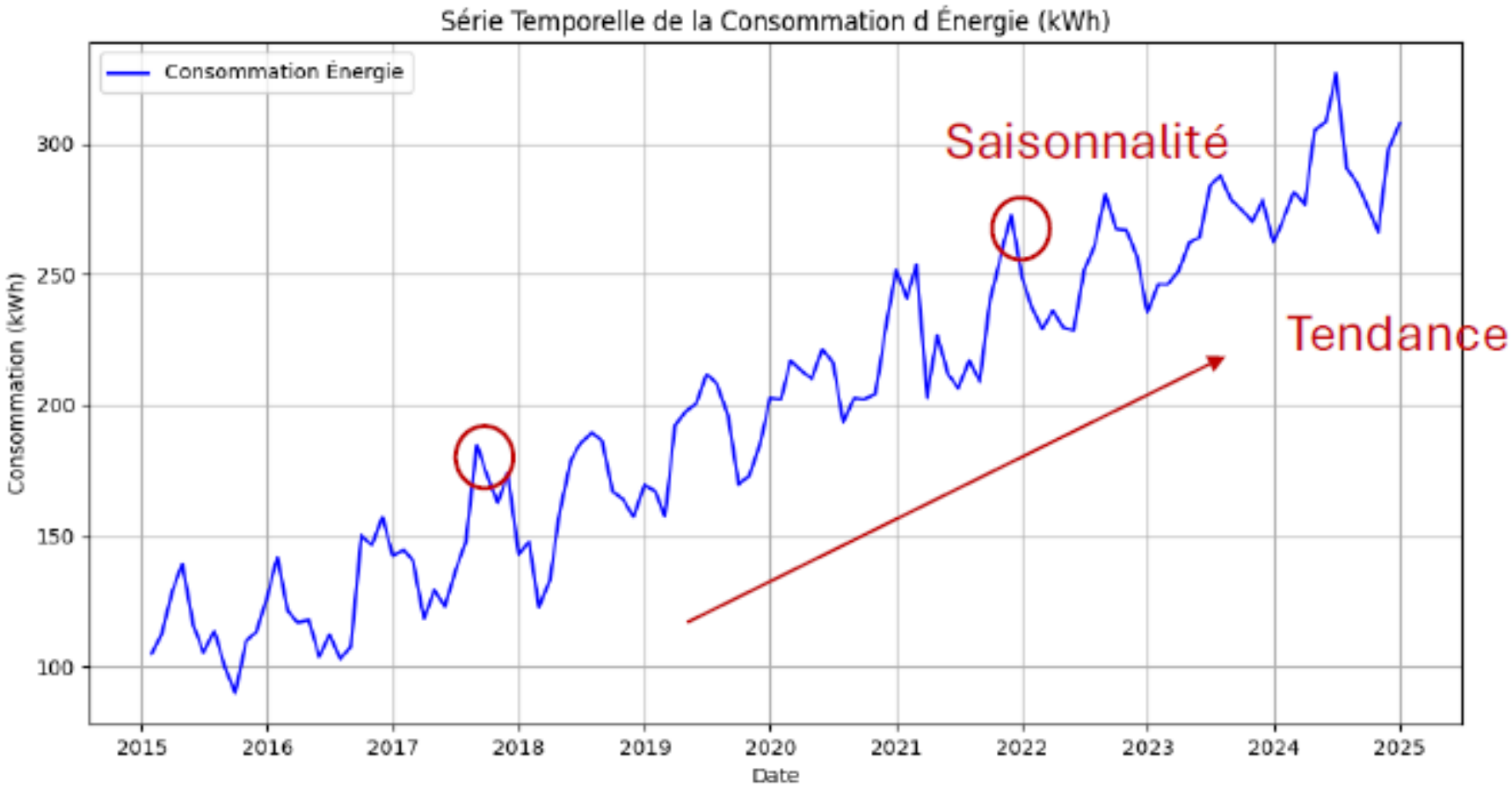
Exploration des données

Séries temporelles

- Suite d'observations répétées d'un même phénomène à des dates différentes
- Exemple :
  - Économie : chômage, croissance, actions en bourse, etc.
  - Environnement : pression, température, niveau d'eau, etc.
  - Santé : électrocardiogramme, etc.
- Utilité :
  - Compréhension du phénomène représenté par la série
  - Prédictions

Dépendance dans le temps et comportements saisonniers

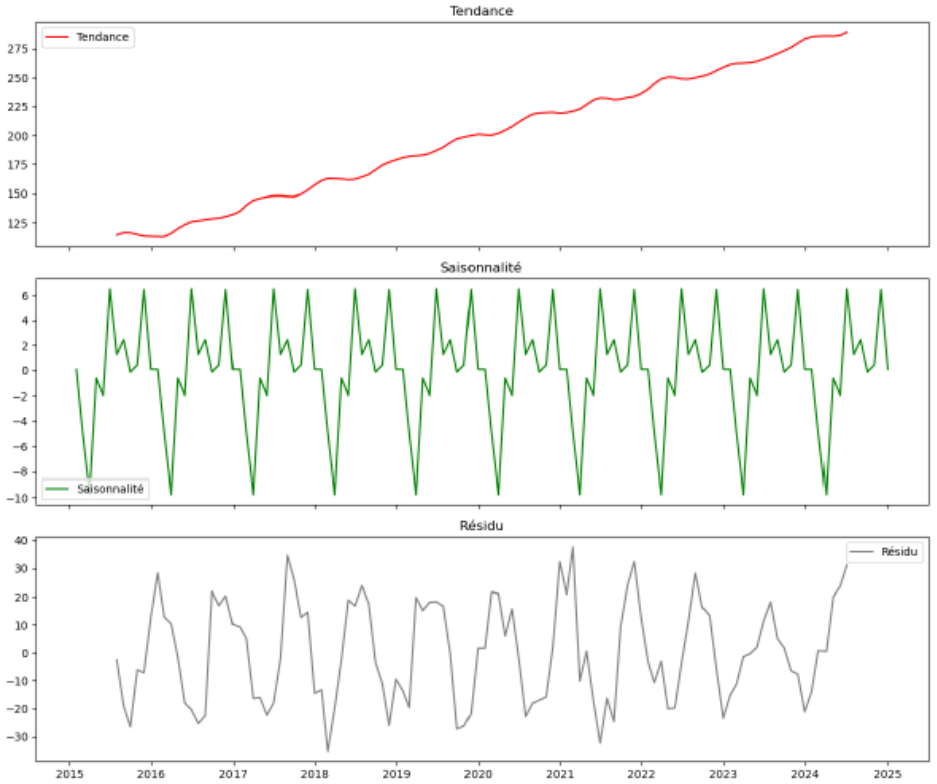
Exemple



Décomposition

En ses composantes

- Composants :
  - **Tendance** (T) : stabilité / augmentation / diminution dans le temps
  - **Saisonnalité** (S) : variations périodiques (facteurs saisonniers ?)
  - **Résidus** (R) : variations aléatoires inexplicables par T / S (bruit) / erreur entre valeur observée et valeur prédite



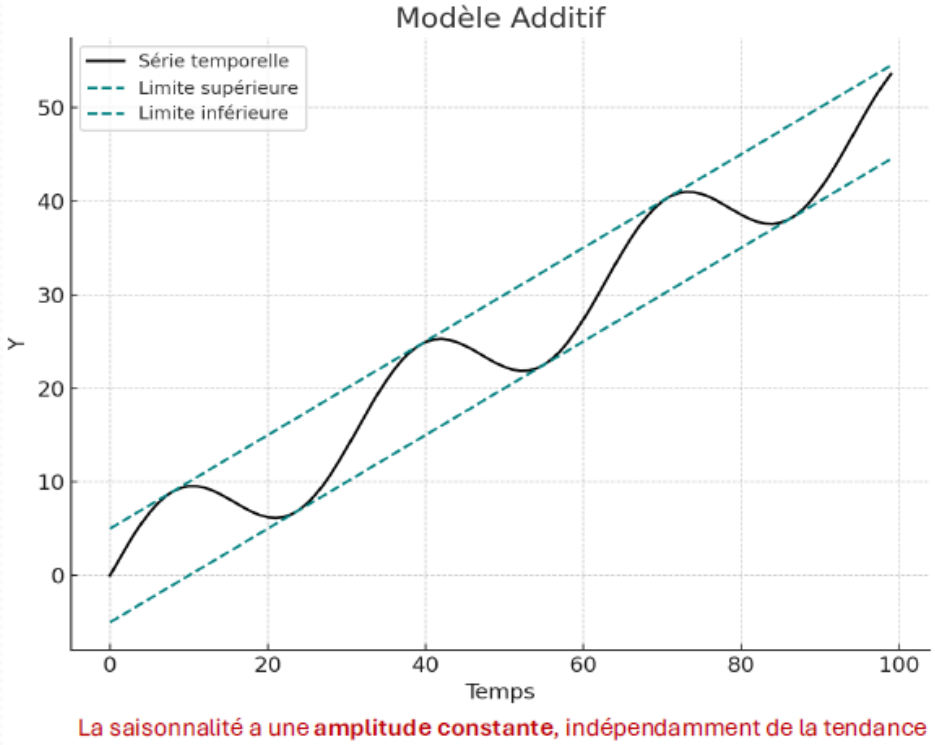
Modèles de décomposition

Modèle additif

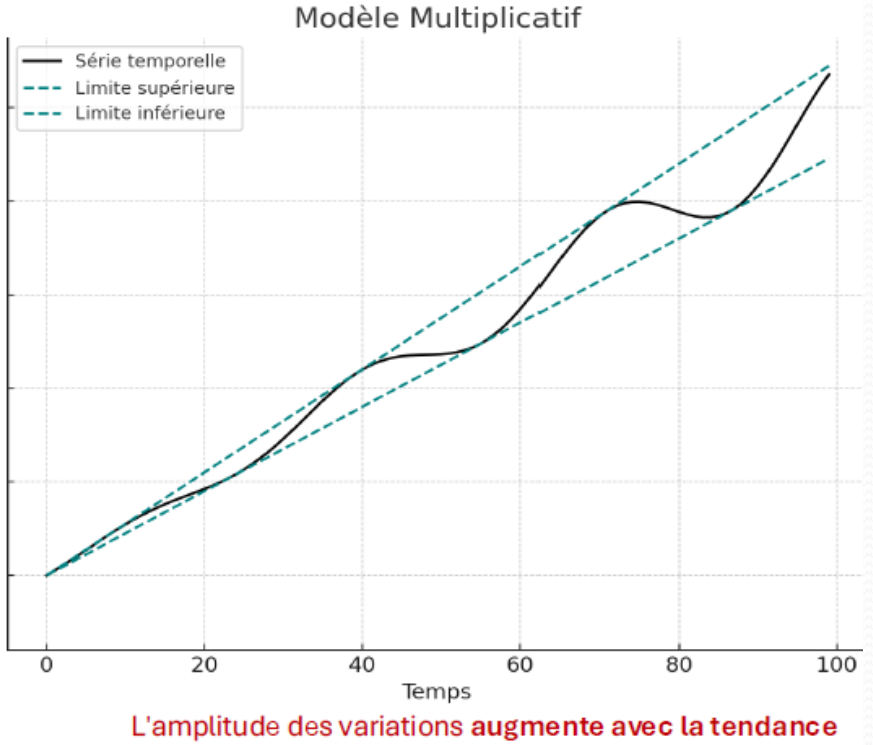
- Amplitude : $Y_t = T_t + S_t + R_t$
- Ex. : temp. journalière en hiver

Modèle multiplicatif

- Amplitude : $Y_t = T_t \times S_t \times R_t$
- Ex. : taux de pluie par mois

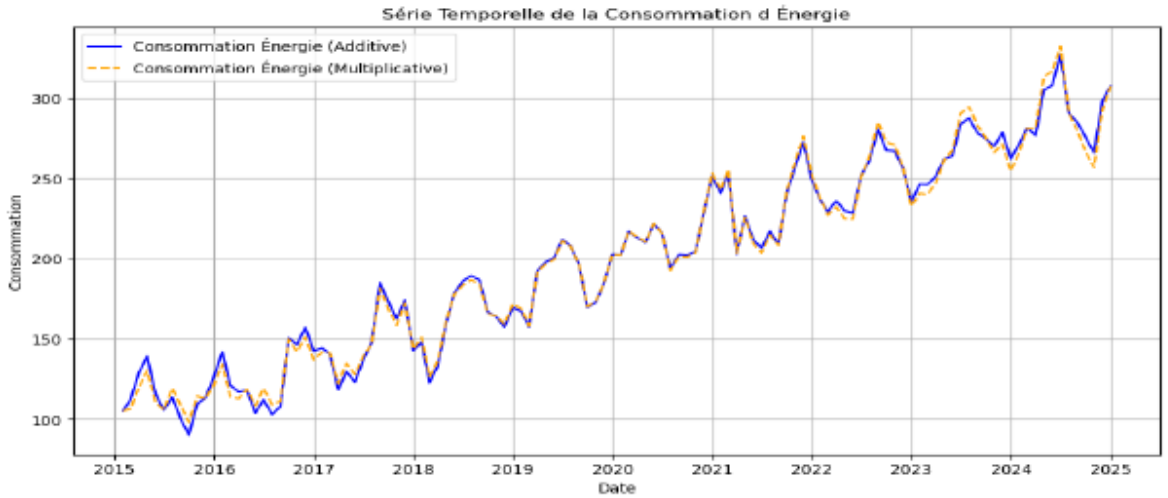


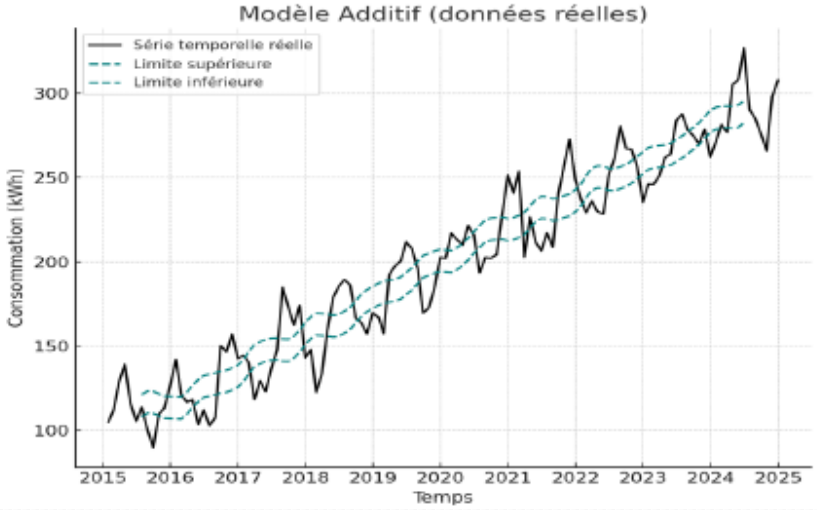
La saisonnalité a une **amplitude constante**, indépendamment de la tendance

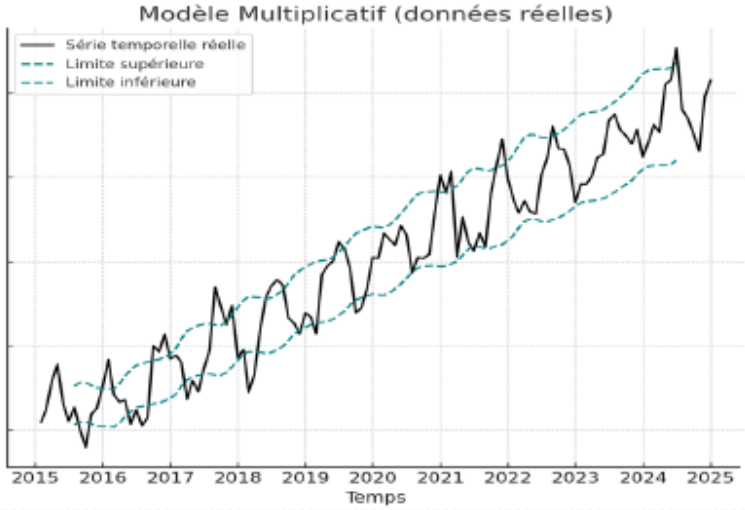


L'amplitude des variations **augmente avec la tendance**

Modèles de décomposition





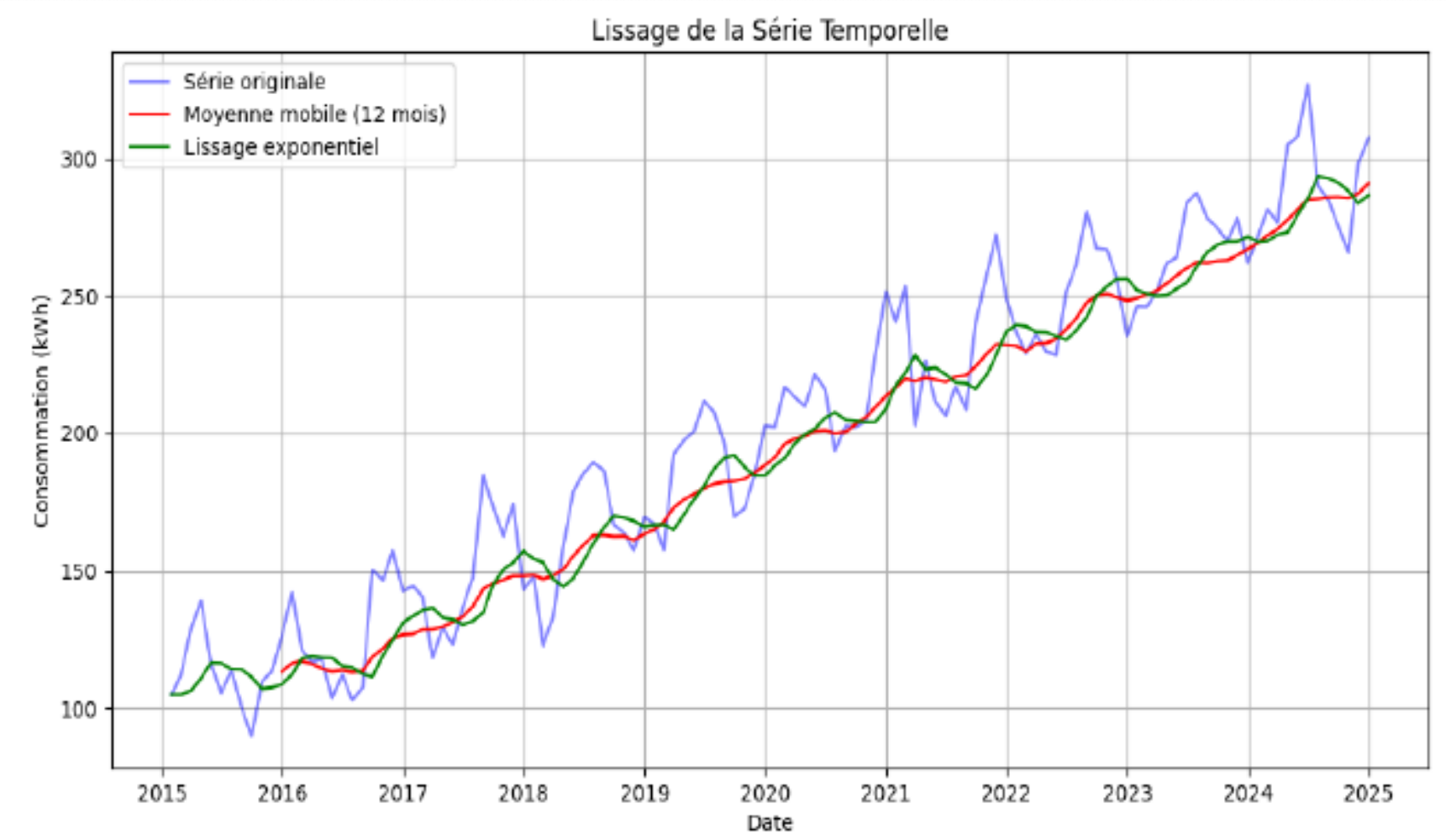


Lissage / filtrage

- Application d'une transformation pour :
  - Réduire le bruit
  - Déterminer la structure (T et S)
  - Révèle les composantes non-stationnaires :
    - Tendance : non-stationnarité en moyenne
    - Variation d'amplitude : non-stationnarité en variance
- Méthodes de lissage :
  - **Moyenne mobile** : remplace chaque valeur par la moyenne des k valeurs précédentes
  - **Lissage exponentiel** : observations récentes ont plus de poids que les plus anciennes (sensible aux fluctuations)

Méthode de Holt-Winters : lissage exponentiel avec prise en compte de T et S

Lissage / filtrage



Stationnarité

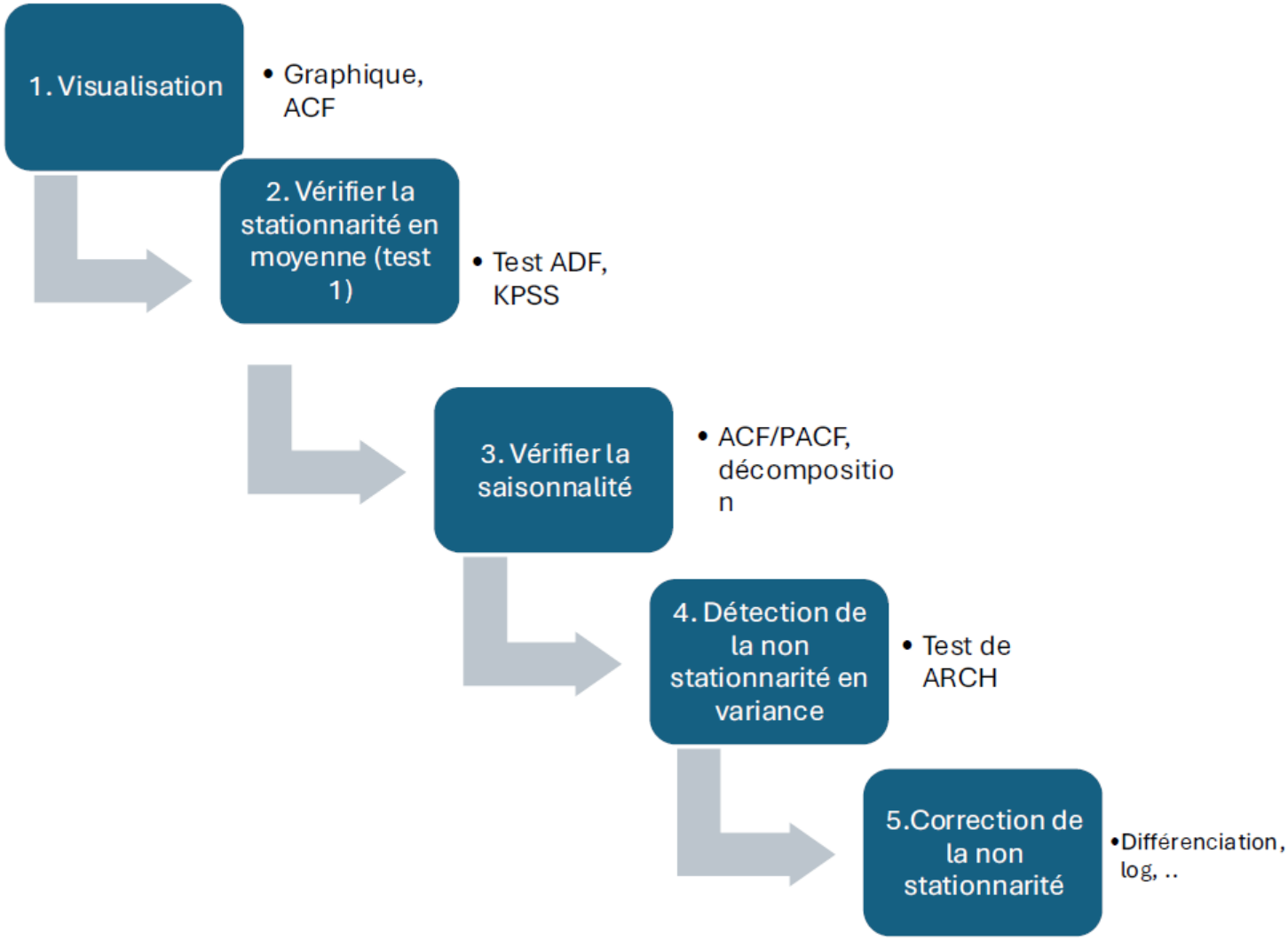
- Problèmes de la non-stationnarité :
  - Résultats biaisés
  - Mauvaise interprétation
  - Incohérences dans les résultats des tests statistiques
  - **Imprédictibilité**
- Tester la stationnarité :
  - Méthodes visuelles : graphique, **ACF**, **PACF**
  - Méthodes statistiques : **ADF**, **KPSS**, **PP**

Les propriétés statistiques (moyenne, variance, autocovariance, autocorrélation) d'une série temporelle stationnaire ne varient pas dans le temps

Une série temporelle non-stationnaire devrait être transformée

Bruit blanc : série temporelle purement aléatoire

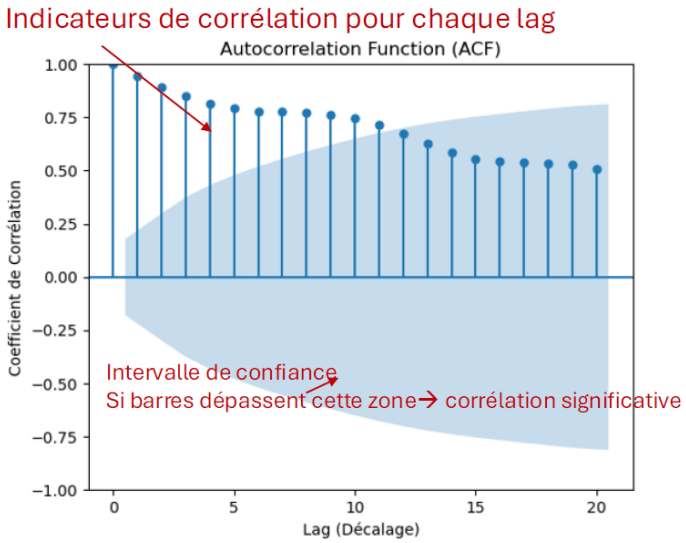
Tester la stationnarité

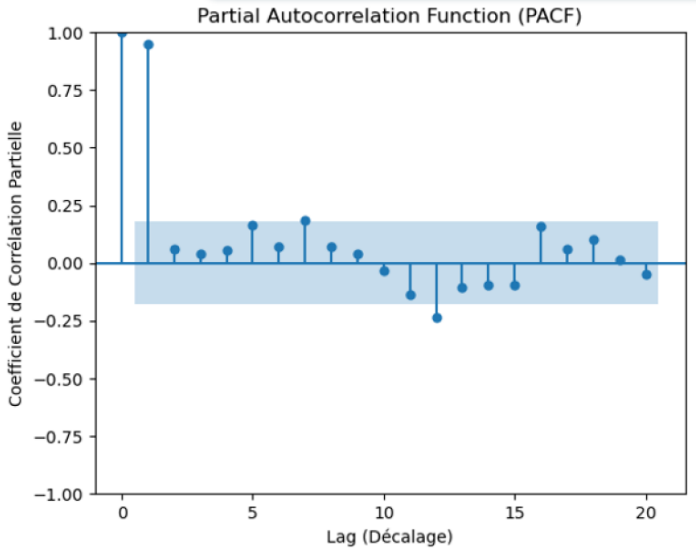


Stationnarité : visuellement

Type non-stationnarité	ACF	PACF
Tendance (moyenne)	Diminue lentement, ne coupe pas à zéro	Forte valeur au lag n°1 Décroît lentement
Saisonnière (périodicité)	Pics périodiques	Pics périodiques aux mêmes lags
Variance (hétéroscédasticité)	Instable	Variable selon les périodes

Lag : nb de périodes de retard



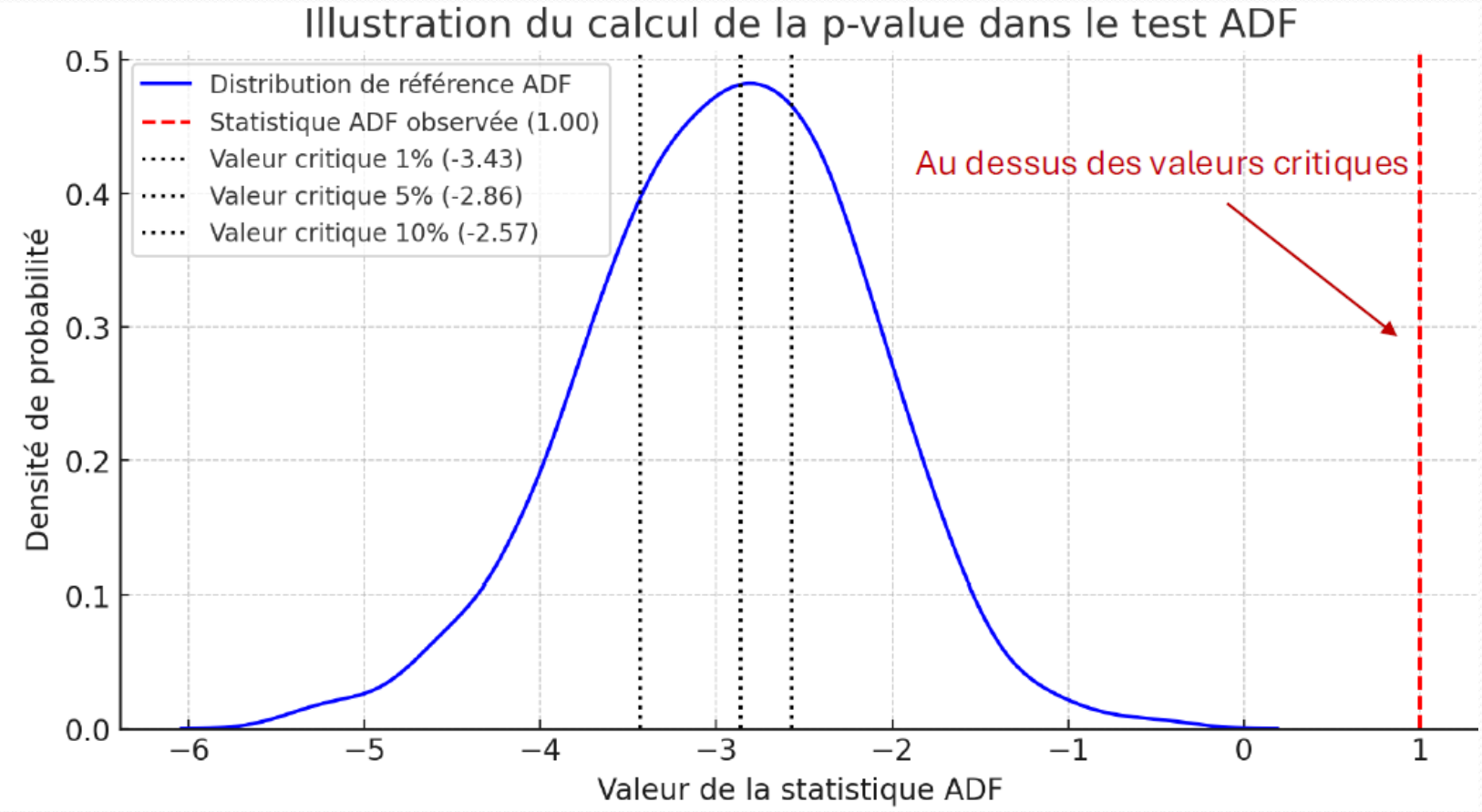


ADF

- **Hypothèse nulle** : série **non-stationnaire**
- Résultats : Générée par un processus présentant une racine unitaire
  - Statistique ADF : si < valeurs critiques, série stationnaire
  - *p-value* (probabilité d'erreur) : si ≤ 0.05, série stationnaire
  - Valeurs critiques : seuils de référence pour décision

ADF sert aussi à tester la stationnarité des résidus

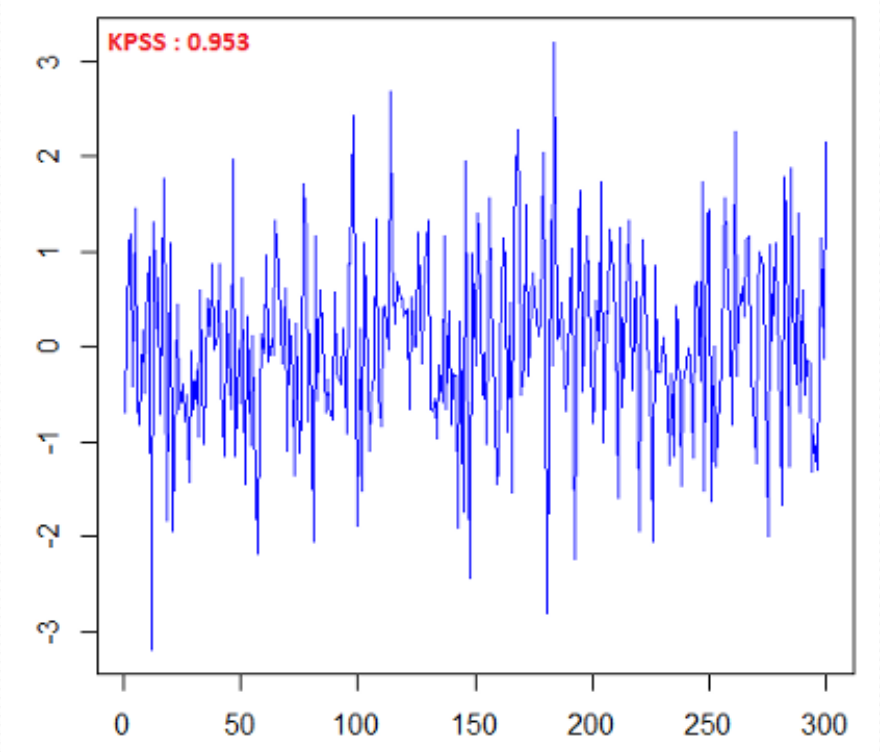
ADF



Stat. ADF	*p-value*	Valeur critique 1%	Valeur critique 5%	Valeur critique 10%
1.0018	0.994	-3.4918	-2.888	-2.5811

KPSS

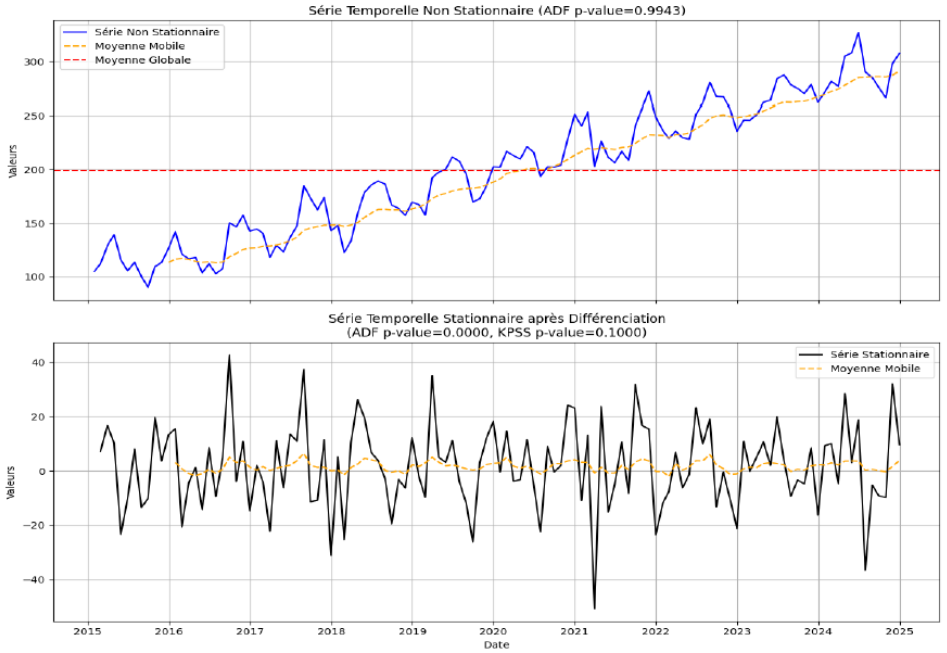
- Similaire à ADF avec **hypothèse nulle** : série **stationnaire**
- Résultats :
  - *p-value* : si > 0.05 : série stationnaire



ADF + KPSS

ADF	KPSS	Interprétation
Non-stationnaire	Stationnaire	Stationnaire (autour d'une tendance)
Non-stationnaire	Non-stationnaire	Série à transformer
Stationnaire	Stationnaire	Stationnaire
Stationnaire	Non-stationnaire	Incohérent, à compléter avec ACF/PACF

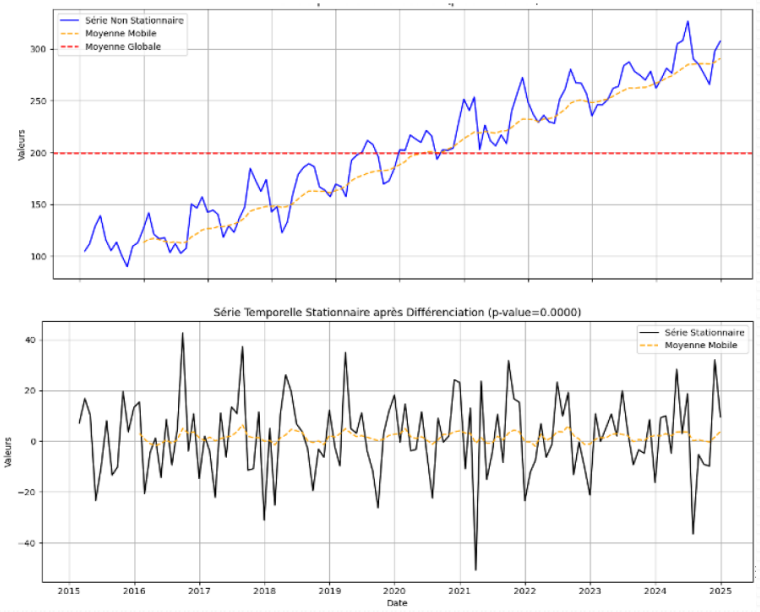
Test	Statistic (Original)	p-value (Original)
ADF	1.001846	0.994285
KPSS	1.790778	0.010000



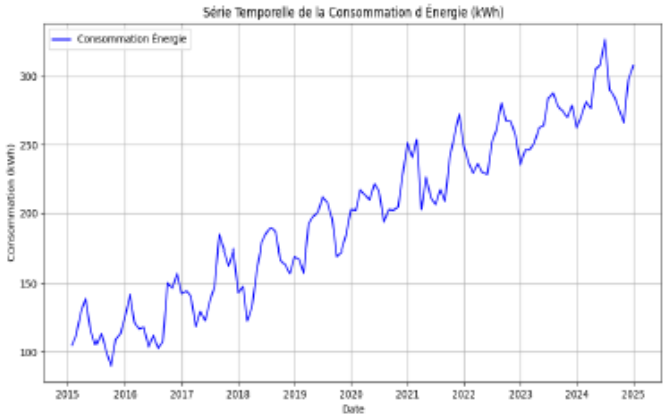
Transformations pour stationnarité

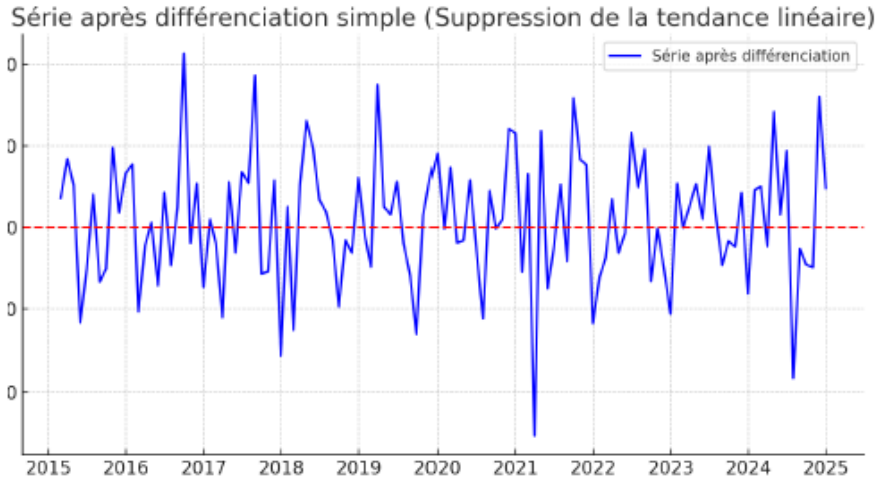
- Transformation pour non-stationnarité :
  - En moyenne (tendance) : différenciation simple
    - $Y_t - Y_{t-1}$ ← Supprime les tendances linéaires
  - De saisonnalité : différenciation saisonnière
    - $Y_t - T_{t-s}$ ← Supprime les cycles saisonniers
  - En variance : trans. logarithmique
    - $Y'_t = \log Y_t$

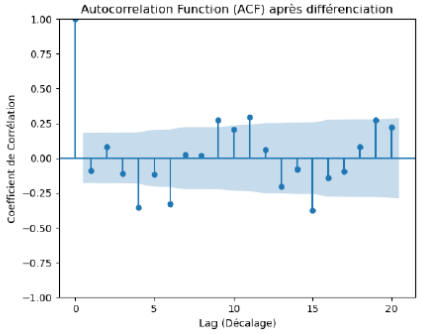
Plusieurs transformations à la suite peuvent être nécessaires

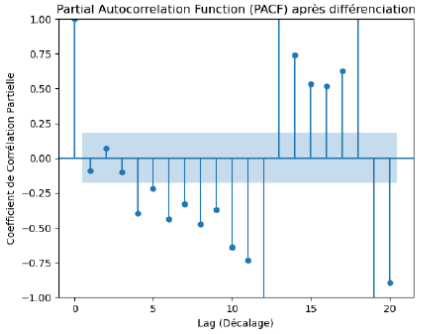


Transformations pour stationnarité









Méthode	Statistique	*p-value*	Stats. (diff.)	*p-value* (diff.)
ADF	1.0018	0.9942	-9.6598	0
KPSS	1.79077	0.01	0.0123	0.1

Prédiction de séries temporelles

- Série stationnaire : **AR**, **MA**, **ARMA**
- Série non-stationnaire (avant) : **ARIMA**, SARIMA
- Critères d'information :
  - AIC *(Akaike info. criterion)* : $AIC = -2 \log L + 2k$
  - BIC *(Bayesian info. criterion)* : $BIC = -\log L + k \log N$

Univariées

Mesure qualité d'un modèle

Prévision court terme quand infos corrélées dans le temps

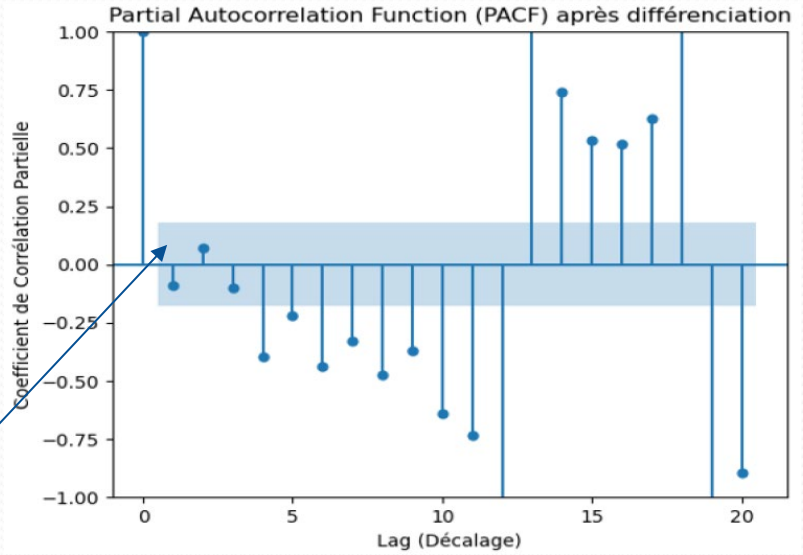
Un bon modèle doit générer des résidus stationnaires : sans tendance ni saisonnalité restante

Les résidus doivent ressembler à un bruit blanc : aléatoire et structure détectable

Modèle AR

- **Autorégressif** :
  - Prédire futures valeurs avec valeurs passées
  - $AR(p)$: valeur actuelle dépend de p valeurs (T) passées
  - $Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t$
  - Choix de p : PACF
  - Estimation des ϕ_p :
    - OLS, Lasso, Ridge

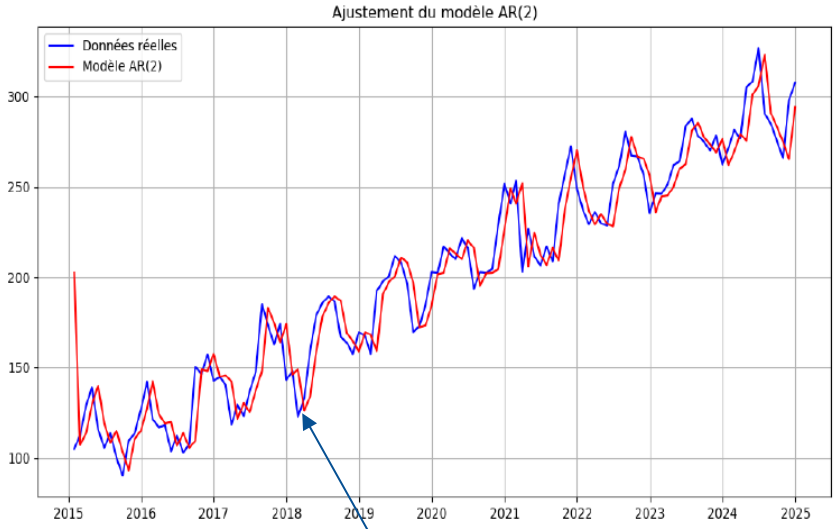
Avec $AR(p)$ la PACF coupe nettement après le *lag* p (ici : p = 1)

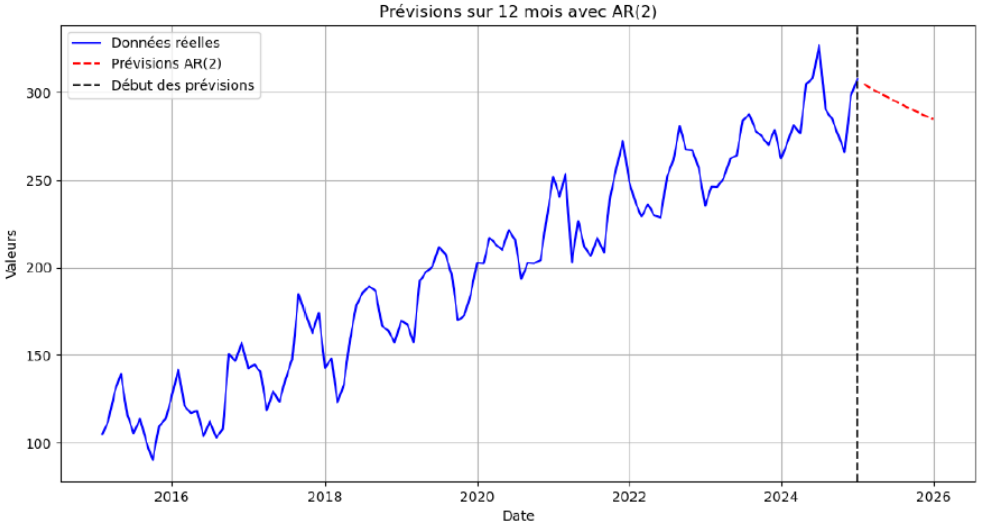


AR ne prend pas en compte les effets saisonniers, sensible aux bruits et données manquantes

Modèle AR

Série temporelle transformée avec différenciation



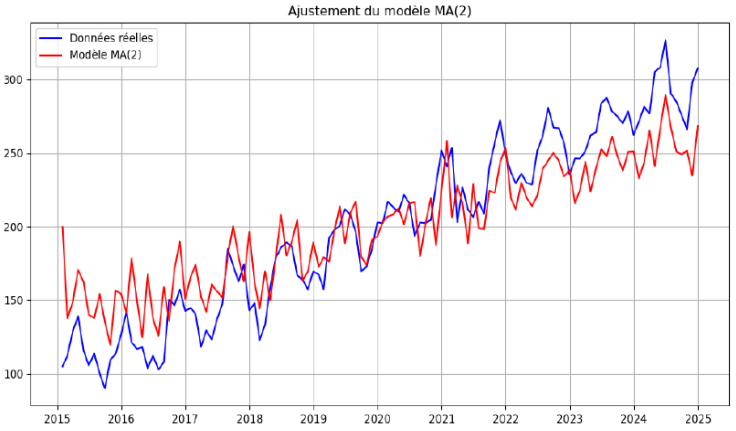


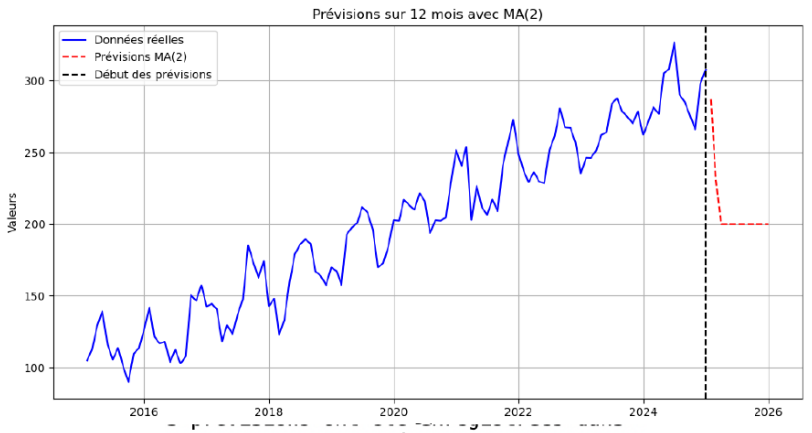
Évaluation du modèle AR(2) :
AIC : 1011.81
BIC : 1022.96
MAE (Erreur absolue moyenne) : 13.07
RMSE (Erreur quadratique moyenne) : 17.90

$AR(2)$ est un bon modèle car courbe rouge suit la bleue

Modèle MA

- **Moyenne mobile** ← ≠ outil de lissage
  - Prédire futures valeurs avec erreurs (R) passés
  - $MA(q) : Y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$
  - Choix de q : ACF





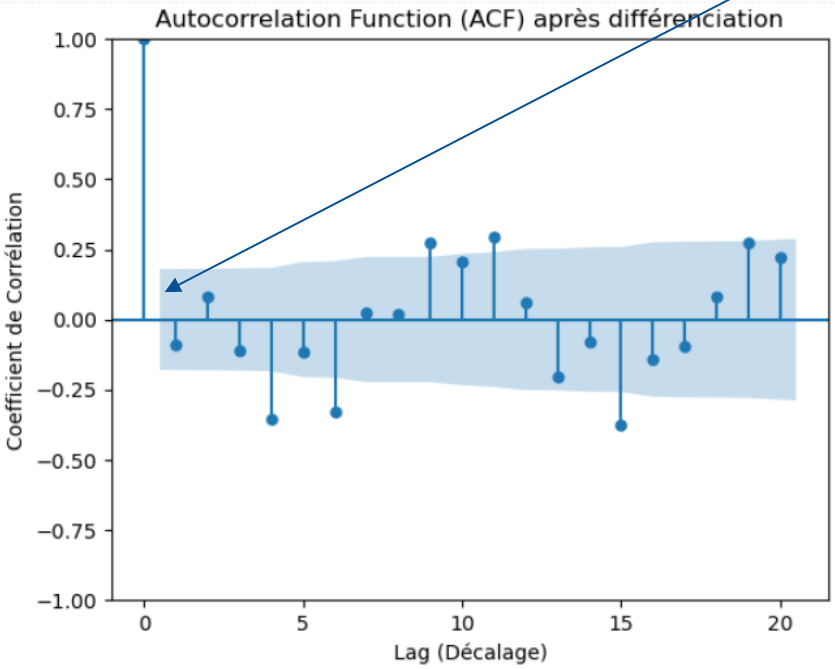
Métrique	Valeur
AIC	1144.25706
BIC	1155.407027
MAE	23.116971
RMSE	28.275765
Stationnarité des résidus	Non

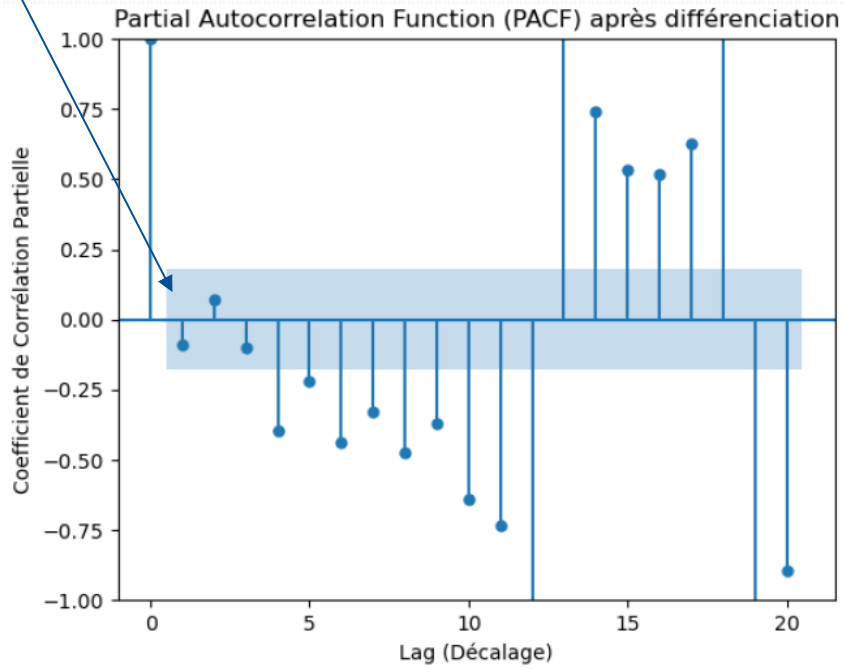
Prédit des valeurs qui convergent rapidement vers une moyenne constante

Modèle ARMA

- $AR(p) + MA(q)$

q=1, p =1

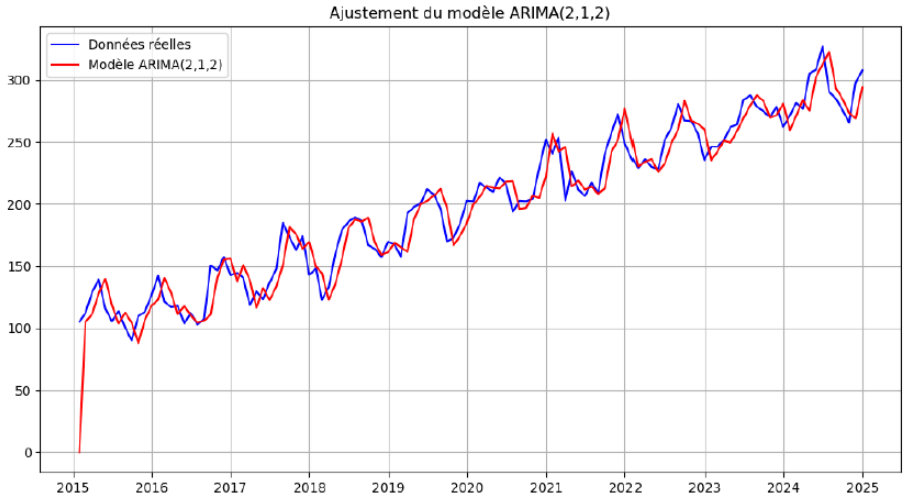


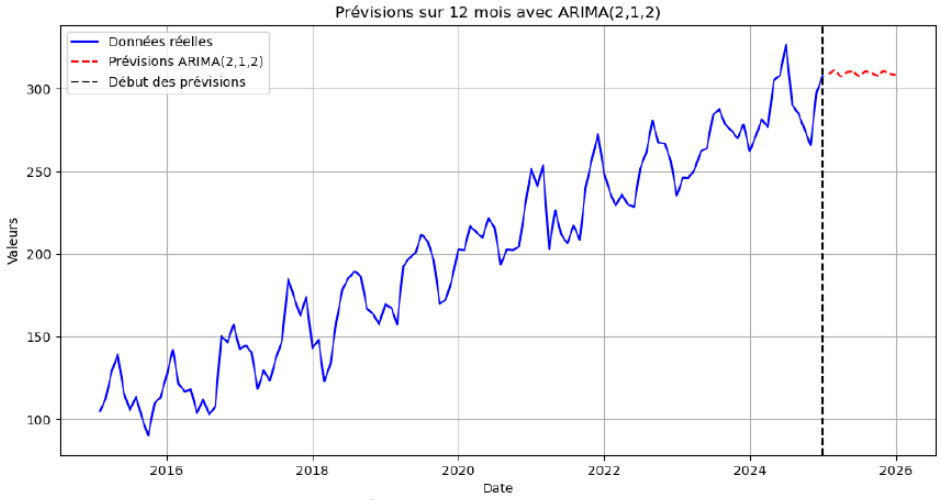


Modèle ARIMA

- *Autoregressive integrated moving average*
- $AR(p) + I(d) + MA(q)$
- **Intégration** : Saisonnalité
  - $I(d)$: d = nb de différenciations pour rendre stationnaire
  - Choix de d : ADF, KPSS

Modèle ARIMA





Métrique	Valeur
AIC	996.18597
BIC	1010.081588
MAE	12.931678
RMSE	17.95582
Stationnarité des résidus	oui

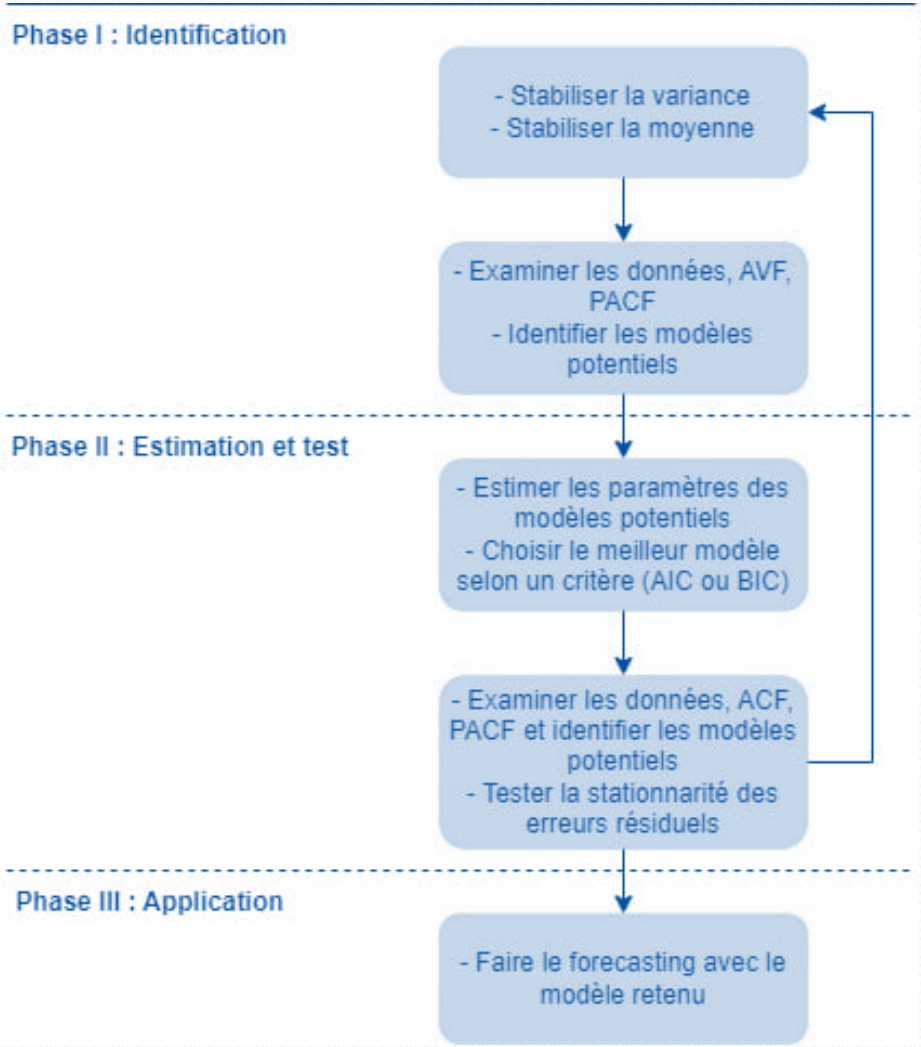
Comparaison des modèles

Modèle	AIC	BIC	MAE	RMSE	Stationnarité
$AR(2)$	1011.81	1022.96	13.07	17.9	Non-stationnaire
$MA(2)$	1144.26	1115.41	23.12	28.28	Non-stationnaire
$ARIMA(2,1,2)$	996.19	1010.08	12.93	17.96	Stationnaire

$ARIMA(2,1,2)$ est le modèle le plus performant

Choix d'un modèle de prédiction

- Minimiser AIC, BIC
- Méthode Box-Jenkins :
  - Protocole statistique de modélisation, d'analyse et d'estimation de modèle



Projet de contributions

- *Outliers* et *skylines* :
  - *Outliers* ⇨ *skyline** ⇨ classement avec méthode de SkyRank
  - *Outliers* + *skyline* + SkyRank
  - *Outlier cube* : toutes les combinaisons de dimensions

Skylines sensibles aux *outliers*

Gain d'efficacité

Connaître dimension(s) à l'origine d'un *outlier*

Aller plus loin

- LCOF *(local correlation outlier factor)*
- Sélection des variables
- Isolation Forest
- SVM
- HQC *(Hannan-Quinn information criterion)*
- SARIMA
- XGBoost
- Random Forest
- LSTM
- Séries temporelles multivariées

Liens

- Document classique :
  - Sana Sellami. *Exploration de données*.
  - Ricco Rakotomalala. *Détection des anomalies*.
  - Breunig , Kriegel, Ng and Sander. *LOF: identifying density based local outliers*.

Crédits

Auteur

Mickaël Martin Nevot
mmartin.nevot@gmail.com

Carte de visite électronique

Relecteurs

Cours en ligne sur : www.mickael-martin-nevot.com