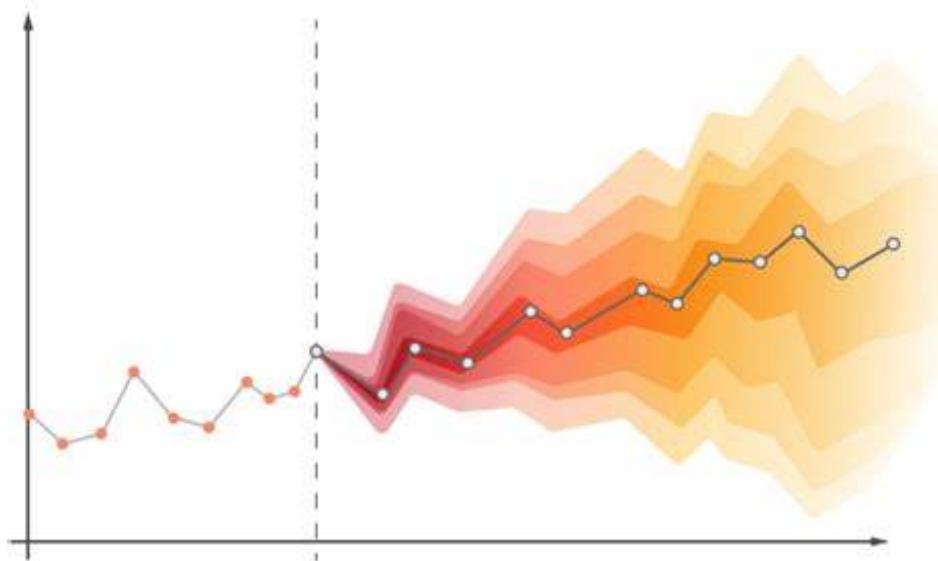


# Université Sultan Moulay Slimane

## École Nationale des Sciences Appliquées - Khouribga -

Filière : Informatique et Ingénierie des Données



## Prédiction du Taux de Chômage aux États-Unis : Analyse Exploratoire et Modélisation Temporelle

---

Réalisé par :  
BOUMLIK YOUNESS

Encadré par :  
Mme. Sara Baghdadi

---

Année universitaire :  
2024/2025

## Résumé

Le taux de chômage est un indicateur économique vital, reflétant la santé du marché du travail et l'état général de l'économie. La capacité à prédire ses fluctuations est essentielle pour la planification politique et économique. Cette étude explore l'évolution historique du taux de chômage mensuel aux États-Unis à partir de données s'étendant de janvier 1948 à janvier 2025. Après un prétraitement des données incluant la conversion des dates et la vérification des doublons, une analyse exploratoire est menée, comprenant une visualisation de la série et sa décomposition en tendance, saisonnalité et résidus. Trois approches de modélisation sont ensuite mises en œuvre et comparées : la régression polynomiale, le modèle Prophet de Facebook, et le modèle ARIMA/SARIMA. Les performances de ces modèles sont évaluées à l'aide de métriques telles que le  $R^2$ , l'erreur absolue moyenne (MAE), la racine de l'erreur quadratique moyenne (RMSE) et l'erreur moyenne absolue en pourcentage (MAPE). Enfin, une brève présentation d'une application Streamlit potentielle pour la visualisation des prévisions est esquissée.

# **Remerciements**

Je tiens à exprimer ma gratitude à Madame Sara Baghdadi pour la clarté de ses explications, la rigueur de son enseignement et sa disponibilité tout au long du cours de Machine Learning. Ses interventions pédagogiques m'ont permis d'acquérir une bonne compréhension des notions fondamentales, ce qui m'a grandement aidé à réaliser ce projet dans de bonnes conditions.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Contexte et Problématique . . . . .	5
1.2	Objectifs de l'étude . . . . .	5
1.3	Structure du rapport . . . . .	5
<b>2</b>	<b>Données et Prétraitement</b>	<b>7</b>
2.1	Description des Données . . . . .	7
2.2	Prétraitement . . . . .	7
2.3	Analyse Exploratoire Initiale . . . . .	8
2.3.1	Évolution Temporelle . . . . .	8
2.3.2	Décomposition de la Série Temporelle . . . . .	8
<b>3</b>	<b>Méthodologie de Modélisation et Métriques d'Évaluation</b>	<b>10</b>
3.1	Régression Polynomiale . . . . .	10
3.1.1	Principe . . . . .	10
3.1.2	Formulation Mathématique . . . . .	10
3.2	Modèle Prophet . . . . .	10
3.2.1	Principe . . . . .	10
3.2.2	Formulation Mathématique . . . . .	11
3.3	Modèle ARIMA/SARIMA . . . . .	11
3.3.1	Principe . . . . .	11
3.3.2	Test de Stationnarité . . . . .	11
3.3.3	Formulation Mathématique . . . . .	12
3.4	Métriques d'Évaluation . . . . .	13
3.4.1	Coefficient de Détermination ( $R^2$ ) . . . . .	13
3.4.2	Erreur Absolue Moyenne (MAE) . . . . .	13
3.4.3	Racine de l'Erreur Quadratique Moyenne (RMSE) . . . . .	13
3.4.4	Erreur Moyenne Absolue en Pourcentage (MAPE) . . . . .	13
<b>4</b>	<b>Implémentation et Résultats des Modèles</b>	<b>14</b>
4.1	Régression Polynomiale . . . . .	14
4.2	Modèle Prophet . . . . .	15
4.3	Modèle ARIMA/SARIMA . . . . .	16
4.3.1	Remarque sur l'Absence de Composante Saisonnière . . . . .	18
<b>5</b>	<b>Comparaison des Modèles et Discussion</b>	<b>19</b>
<b>6</b>	<b>Application Streamlit pour la Visualisation des Prédictions</b>	<b>21</b>
<b>7</b>	<b>Conclusion et Perspectives</b>	<b>24</b>

# Table des figures

2.1	Évolution du taux de chômage aux États-Unis (1948-2025). . . . .	8
2.2	Décomposition saisonnière du taux de chômage. . . . .	9
4.1	Ajustements par régression polynomiale de différents degrés. . . . .	14
4.2	Prévisions du taux de chômage avec Prophet (jusqu'en janvier 2027). . . . .	15
4.3	Décomposition des composantes par Prophet (Tendance, Saisonnalités annuelle, mensuelle, et cycle économique). . . . .	16
4.4	Ajustement du modèle ARIMA(1,1,1) sur les données historiques. . . . .	17
4.5	Prévisions du taux de chômage avec ARIMA(1,1,1) (jusqu'en janvier 2027). . . . .	17
5.1	Comparaison des prévisions du taux de chômage avec Prophet et ARIMA(1,1,1) jusqu'en 2027. . . . .	19
6.1	Vue tabulaire des prévisions dans l'application Streamlit. Affiche les dates, les valeurs prédites par les modèles (Prophet, ARIMA) et leurs intervalles de confiance.	22
6.2	Visualisation graphique interactive des données historiques et des prévisions du taux de chômage. Les différentes couleurs représentent les données réelles, les prédictions de Prophet et celles d'ARIMA. . . . .	22
6.3	Exemple de prévision sur un horizon de 3 ans dans l'application Streamlit. L'utilisateur peut ajuster cette période via les paramètres de prévision, en choisissant un nombre de mois ou d'années. . . . .	23

# Liste des tableaux

2.1 Statistiques descriptives de la variable UNRATE. . . . .	8
5.1 Comparaison des métriques de performance des modèles Prophet et ARIMA. . .	19

# Chapitre 1

## Introduction

### 1.1 Contexte et Problématique

Le taux de chômage est un des indicateurs macroéconomiques les plus scrutés, offrant un aperçu direct de la vitalité du marché de l'emploi et, par extension, de la santé économique globale d'un pays. Ses variations peuvent avoir des implications profondes sur les politiques gouvernementales, les stratégies d'investissement des entreprises, et le bien-être social. Aux États-Unis, le suivi de cet indicateur par des organismes comme le Bureau of Labor Statistics (BLS) est une tradition de longue date, fournissant des séries temporelles riches pour l'analyse.

La problématique centrale de cette étude réside dans la capacité à anticiper les évolutions futures du taux de chômage. Une prédiction fiable permettrait aux décideurs d'ajuster les politiques monétaires et fiscales, aux entreprises d'anticiper les conditions du marché, et aux individus de mieux planifier leur avenir professionnel. Cependant, le taux de chômage est influencé par une multitude de facteurs complexes et souvent interconnectés (cycles économiques, politiques, événements mondiaux, innovations technologiques), ce qui rend sa prédiction particulièrement ardue.

### 1.2 Objectifs de l'étude

L'objectif principal de ce rapport est d'explorer et d'évaluer différentes méthodes de modélisation temporelle pour la prédiction du taux de chômage aux États-Unis. Plus spécifiquement, les objectifs sont les suivants :

- Analyser les données historiques du taux de chômage américain pour identifier les tendances, les composantes saisonnières, et les cycles économiques potentiels.
- Mettre en œuvre un modèle de régression polynomiale comme approche de base pour capturer les tendances non-linéaires.
- Appliquer le modèle Prophet, conçu pour les séries temporelles présentant des saisonsnalités multiples et des points de changement.
- Développer un modèle ARIMA (ou SARIMA si la saisonnalité est significative) basé sur les approches statistiques classiques d'analyse de séries temporelles.
- Évaluer et comparer les performances prédictives de ces modèles en utilisant des métriques statistiques appropriées ( $R^2$ , MAE, RMSE, MAPE).
- Discuter des avantages, des inconvénients, et de l'applicabilité de chaque modèle dans le contexte de la prédiction du chômage.

### 1.3 Structure du rapport

Ce rapport est structuré comme suit :

- Le Chapitre 2 détaille les données utilisées, leur source, ainsi que les étapes de prétraitement effectuées pour préparer les données à l'analyse et à la modélisation.

- Le Chapitre 3 présente les fondements théoriques et les formulations mathématiques des trois approches de modélisation retenues : la régression polynomiale, le modèle Prophet, et le modèle ARIMA/SARIMA.
- Le Chapitre 4 décrit l’implémentation pratique de chaque modèle à l’aide des outils Python, en incluant les paramètres choisis et les principaux résultats graphiques et numériques obtenus sur les données historiques.
- Le Chapitre 5 offre une comparaison quantitative et qualitative des performances des modèles, basée sur les métriques d’évaluation et les prévisions générées.
- Le Chapitre 6 esquisse brièvement une application Streamlit potentielle pour la visualisation interactive des données et des prévisions.
- Enfin, le Chapitre 7 résume les principales conclusions de l’étude, discute des limites des modèles et propose des perspectives pour des travaux futurs.

# Chapitre 2

## Données et Prétraitement

### 2.1 Description des Données

Les données utilisées dans cette étude proviennent de la base de données économique de la Réserve fédérale des États-Unis, connue sous le nom de FRED (Federal Reserve Economic Data), accessible via le site de la Federal Reserve Bank of St. Louis. Le fichier utilisé est un fichier CSV nommé `UNRATE.csv`, téléchargeable depuis la plateforme FRED à l'adresse suivante : <https://fred.stlouisfed.org/series/UNRATE/>.

Ce fichier contient une série temporelle du taux de chômage mensuel aux États-Unis. Le jeu de données comprend 925 observations réparties sur deux colonnes :

- `observation_date` : La date de l'observation, initialement de type `object`. Elle couvre la période de janvier 1948 à janvier 2025.
- `UNRATE` : Le taux de chômage en pourcentage, de type `float64`.

Aucune valeur manquante n'a été détectée dans ces deux colonnes.

### 2.2 Prétraitement

Les étapes de prétraitement suivantes ont été appliquées :

1. **Conversion de la date** : La colonne `observation_date` a été convertie en objets datetim pandas en utilisant `pd.to_datetime()`. Cela permet une manipulation et une analyse temporelle plus aisées.
2. **Gestion des doublons** : La commande `chomage_usa.drop_duplicates()` a été exécutée par précaution pour assurer l'unicité des enregistrements. Aucune observation dupliquée n'a été identifiée et supprimée lors de cette étape.
3. **Création d'une variable numérique de date** : Pour la régression polynomiale, une colonne `date_num` a été créée. Elle représente le nombre de jours écoulés depuis la première date d'observation (`chomage_usa['observation_date'].min()`). Cette transformation est nécessaire car les modèles de régression classiques requièrent des entrées numériques.
4. **Mise en index de la date** : Pour les analyses de séries temporelles avec `statsmodels` et `Prophet`, la colonne `observation_date` (ou sa version renommée 'ds' pour `Prophet`) est utilisée comme index ou comme colonne de date principale.

Les statistiques descriptives montrent que le taux de chômage varie entre 2.5% et 14.8%, avec une moyenne de 5.68%.

TABLE 2.1 – Statistiques descriptives de la variable UNRATE.

UNRATE	
count	925.000000
mean	5.681622
min	2.500000
25%	4.400000
50%	5.500000
75%	6.700000
max	14.800000
std	1.708949

## 2.3 Analyse Exploratoire Initiale

### 2.3.1 Évolution Temporelle

La figure 2.1 illustre l'évolution du taux de chômage de 1948 à 2025.

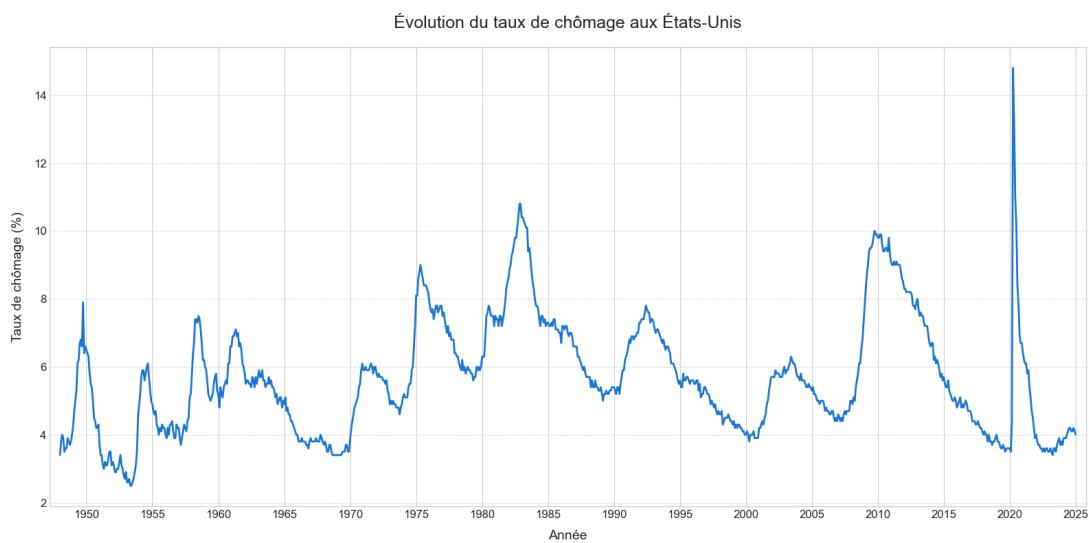


FIGURE 2.1 – Évolution du taux de chômage aux États-Unis (1948-2025).

On observe des fluctuations notables au fil du temps, avec des pics correspondant souvent à des périodes de récession économique (par exemple, début des années 80, crise de 2008, pic de 2020 lié à la pandémie). La série ne semble pas stationnaire à première vue, présentant des variations de moyenne et potentiellement de variance.

### 2.3.2 Décomposition de la Série Temporelle

Une décomposition saisonnière additive de la série temporelle a été effectuée, en supposant une périodicité de 12 mois. La figure 2.2 montre les composantes de cette décomposition.

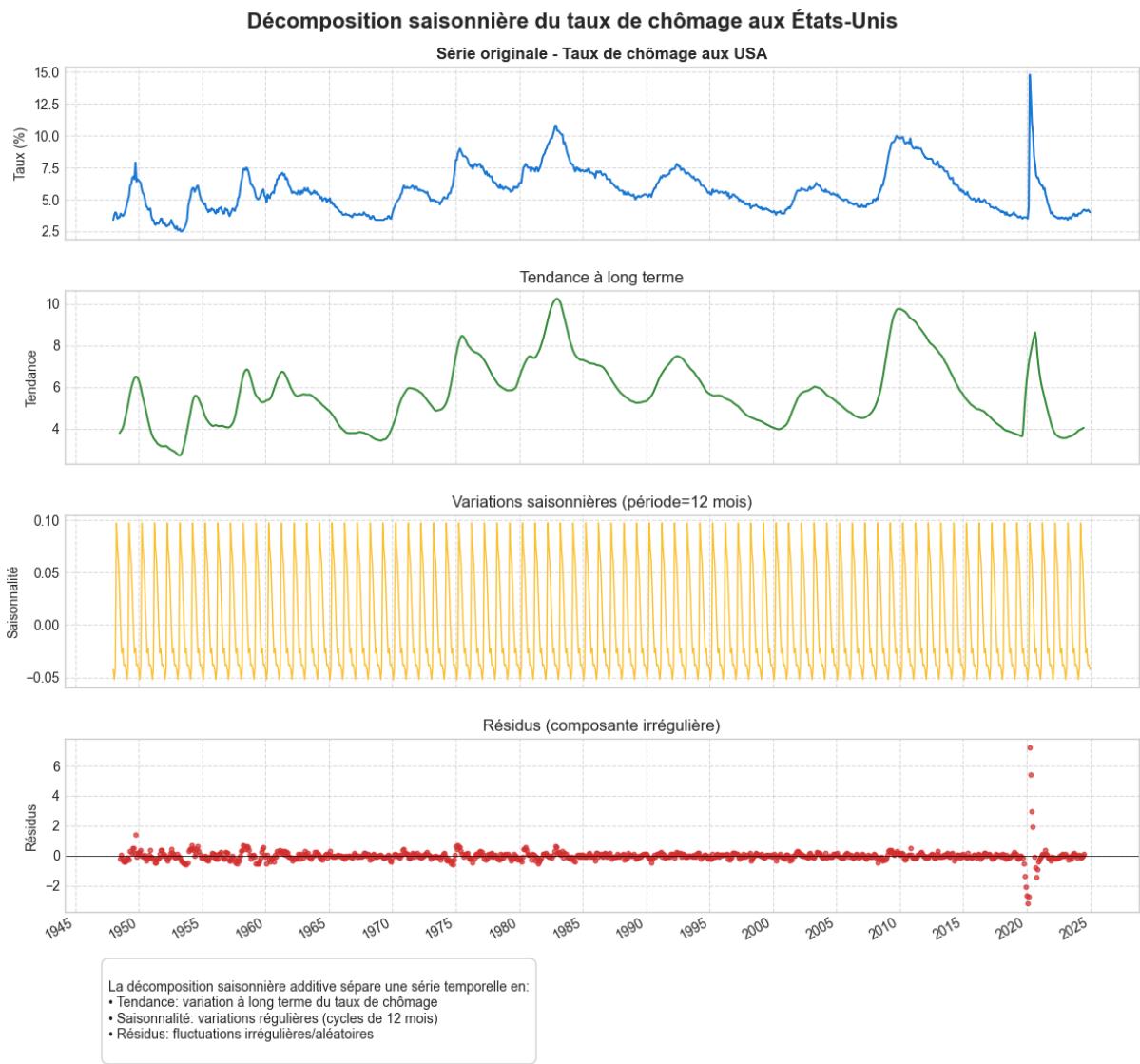


FIGURE 2.2 – Décomposition saisonnière du taux de chômage.

L'analyse de la décomposition a révélé :

- **Tendance** : Une variation à long terme significative, avec une amplitude de 7.54 points de pourcentage (min 2.71, max 10.25).
- **Saisonnalité** : Des variations régulières sur des cycles de 12 mois, mais d'amplitude relativement faible (0.149 points de pourcentage).
- **Résidus** : Des fluctuations irrégulières avec un écart-type de 0.423, indiquant une certaine volatilité non expliquée par la tendance et la saisonnalité simple.

# Chapitre 3

## Méthodologie de Modélisation et Métriques d'Évaluation

Trois approches distinctes ont été utilisées pour modéliser et prédire le taux de chômage.

### 3.1 Régression Polynomiale

#### 3.1.1 Principe

La régression polynomiale est une forme de régression linéaire où la relation entre la variable indépendante  $x$  (temps, dans notre cas `date_num`) et la variable dépendante  $y$  (taux de chômage `UNRATE`) est modélisée comme un polynôme de degré  $n^{ième}$ . Cela permet de capturer des tendances non linéaires dans les données.

#### 3.1.2 Formulation Mathématique

Le modèle de régression polynomiale de degré  $n$  est donné par l'équation :

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_i^n + \epsilon_i \quad (3.1)$$

où :

- $y_i$  est la valeur observée du taux de chômage au temps  $x_i$ .
- $x_i$  est la variable de temps numérique.
- $\beta_0, \beta_1, \dots, \beta_n$  sont les coefficients du polynôme à estimer.
- $\epsilon_i$  est le terme d'erreur.

L'estimation des coefficients  $\beta_j$  se fait généralement par la méthode des moindres carrés. Le choix du degré  $n$  est crucial : un degré trop faible peut sous-ajuster les données (biais élevé), tandis qu'un degré trop élevé peut sur-ajuster (variance élevée) et mal généraliser. Dans cette étude, différents degrés ont été testés et le meilleur a été sélectionné sur la base du score R<sup>2</sup> sur un ensemble de test.

### 3.2 Modèle Prophet

#### 3.2.1 Principe

Prophet est un modèle de prévision de séries temporelles développé par Facebook. Il est conçu pour gérer des caractéristiques courantes des séries temporelles commerciales, telles que les tendances multiples, la saisonnalité annuelle, hebdomadaire et journalière, ainsi que les effets de jours fériés. Il est robuste aux données manquantes et aux changements de tendance.

### 3.2.2 Formulation Mathématique

Prophet modélise la série temporelle  $y(t)$  comme une somme de trois composantes principales et d'un terme d'erreur :

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (3.2)$$

où :

- $g(t)$  : La fonction de **tendance**, modélisant les changements non-périodiques. Prophet utilise un modèle linéaire par morceaux ou logistique pour la tendance. Les points où le taux de croissance de la tendance change sont appelés « changepoints ».
- $s(t)$  : La fonction de **saisonnalité**, modélisant les changements périodiques (par exemple, annuels, mensuels). Elle est approximée par des séries de Fourier :

$$s(t) = \sum_{n=1}^N \left( a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

où  $P$  est la période (par exemple, 365.25 pour la saisonnalité annuelle).

- $h(t)$  : La fonction des **effets de jours fériés** et événements spéciaux. Elle est modélisée comme un ensemble de régresseurs indicateurs.
- $\epsilon_t$  : Le terme d'erreur, supposé être un bruit blanc normally distribué.

#### Remarque sur l'origine du nom « Prophet »

*J'ai cherché à comprendre pourquoi ce modèle s'appelle « Prophet ». Il apparaît que ce nom a été choisi car un prophète (en français) est traditionnellement perçu comme une personne capable de prédire l'avenir. Ainsi, le nom « Prophet » pour cet outil de modélisation vise à souligner sa capacité à effectuer des prédictions sur des séries temporelles, tout en suggérant une certaine facilité d'utilisation pour « voir » les tendances futures des données.*

## 3.3 Modèle ARIMA/SARIMA

### 3.3.1 Principe

ARIMA (AutoRegressive Integrated Moving Average) est une classe de modèles statistiques utilisée pour analyser et prédire des données de séries temporelles. SARIMA (Seasonal ARIMA) étend ARIMA pour modéliser explicitement la saisonnalité.

- **AR(p) (AutoRégressif)** : Le modèle utilise la relation de dépendance entre une observation et un certain nombre d'observations passées.  $p$  est l'ordre de la partie autorégressive.
- **I(d) (Intégré)** : Utilise la différenciation des observations brutes (par exemple, en soustrayant l'observation précédente de l'observation actuelle) pour rendre la série temporelle stationnaire.  $d$  est le nombre de différenciations nécessaires.
- **MA(q) (Moyenne Mobile)** : Le modèle utilise la dépendance entre une observation et une erreur résiduelle d'un modèle de moyenne mobile appliqué à des observations passées.  $q$  est l'ordre de la partie moyenne mobile.

La composante saisonnière  $(P, D, Q)_m$  fonctionne de manière similaire mais sur des pas de temps saisonniers ( $m$  est la période de la saisonnalité).

### 3.3.2 Test de Stationnarité

Avant d'appliquer un modèle ARIMA, il est crucial de s'assurer que la série temporelle est stationnaire (c'est-à-dire que sa moyenne, sa variance et son auto-covariance ne varient pas avec le temps). Le test de Dickey-Fuller Augmenté (ADF) est utilisé à cette fin. L'hypothèse nulle ( $H_0$ ) du test ADF est que la série possède une racine unitaire (non stationnaire). Une p-value faible (typiquement  $< 0.05$ ) permet de rejeter  $H_0$ , indiquant que la série est stationnaire.

### 3.3.3 Formulation Mathématique

#### Modèle ARIMA(p,d,q)

Un modèle ARIMA combine trois composantes principales :

**Équation générale :**

$$\text{AR}(L) \times (1 - L)^d \times Y_t = c + \text{MA}(L) \times \epsilon_t \quad (3.3)$$

**Où :**

- $Y_t$  = valeur de notre série temporelle au moment  $t$
- $L$  = opérateur de retard ( $L \times Y_t = Y_{t-1}$ )
- $(1 - L)^d$  = opération de différenciation répétée  $d$  fois
- $c$  = constante (terme d'ajustement)
- $\epsilon_t$  = erreur aléatoire (bruit blanc)

**Les trois composantes :**

1. **AR(p) - Partie Autorégressive :**

$$\text{AR}(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p \quad (3.4)$$

- Utilise les  $p$  valeurs passées pour prédire la valeur actuelle
- $\phi_i$  sont les coefficients autorégressifs

2. **I(d) - Intégration (Différenciation) :**

$$(1 - L)^d \text{ appliqué à } Y_t \quad (3.5)$$

- Rend la série stationnaire en supprimant les tendances
- $d$  = nombre de différenciations nécessaires

3. **MA(q) - Moyenne Mobile :**

$$\text{MA}(L) = 1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q \quad (3.6)$$

- Utilise les  $q$  erreurs passées
- $\theta_j$  sont les coefficients de moyenne mobile

#### Modèle SARIMA(p,d,q)(P,D,Q)<sub>m</sub> - Version Saisonnière

Le SARIMA ajoute une composante saisonnière au modèle ARIMA :

**Équation générale :**

$$\text{AR}_{\text{saison}}(L^m) \times \text{AR}(L) \times (1 - L^m)^D \times (1 - L)^d \times Y_t = c + \text{MA}_{\text{saison}}(L^m) \times \text{MA}(L) \times \epsilon_t \quad (3.7)$$

**Composantes supplémentaires :**

- $m$  = période saisonnière (12 pour des données mensuelles, 4 pour trimestrielles)
- $P$  = ordre AR saisonnier
- $D$  = ordre de différenciation saisonnière
- $Q$  = ordre MA saisonnier
- $L^m$  = opérateur de retard saisonnier ( $L^m \times Y_t = Y_{t-m}$ )

**Parties saisonnières :**

1. **AR saisonnier :**

$$\text{AR}_{\text{saison}}(L^m) = 1 - \Phi_1 L^m - \Phi_2 L^{2m} - \dots - \Phi_P L^{Pm} \quad (3.8)$$

2. **Différenciation saisonnière :**

$$(1 - L^m)^D \quad (3.9)$$

3. **MA saisonnier :**

$$\text{MA}_{\text{saison}}(L^m) = 1 + \Theta_1 L^m + \Theta_2 L^{2m} + \dots + \Theta_Q L^{Qm} \quad (3.10)$$

## Sélection Automatique des Paramètres

La fonction `auto_arima` de la bibliothèque `pmdarima` :

- **Objectif** : Trouver automatiquement les meilleurs paramètres  $(p, d, q)(P, D, Q)_m$
- **Méthode** : Teste différentes combinaisons et sélectionne celle qui minimise l'AIC
- **AIC (Critère d'Information d'Akaike)** : Mesure qui équilibre la qualité de l'ajustement et la complexité du modèle
- **Avantage** : Évite le processus manuel fastidieux de sélection des paramètres

## 3.4 Métriques d'Évaluation

### 3.4.1 Coefficient de Détermination ( $R^2$ )

Le coefficient de détermination  $R^2$  mesure la proportion de la variance de la variable dépendante qui est prévisible à partir des variables indépendantes :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.11)$$

où :

- $y_i$  = valeurs observées
- $\hat{y}_i$  = valeurs prédites
- $\bar{y}$  = moyenne des valeurs observées
- $SS_{res}$  = somme des carrés des résidus
- $SS_{tot}$  = somme totale des carrés

$R^2$  varie entre 0 et 1, une valeur proche de 1 indiquant un modèle qui explique bien la variabilité des données.

### 3.4.2 Erreur Absolue Moyenne (MAE)

La MAE mesure la moyenne des valeurs absolues des erreurs entre les prédictions et les observations :

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.12)$$

Cette métrique est robuste aux valeurs aberrantes et s'exprime dans la même unité que la variable prédite.

### 3.4.3 Racine de l'Erreur Quadratique Moyenne (RMSE)

La RMSE calcule la racine carrée de la moyenne des erreurs au carré :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.13)$$

La RMSE pénalise davantage les grandes erreurs par rapport à la MAE, étant donné l'élévation au carré. Elle s'exprime également dans la même unité que la variable prédite.

### 3.4.4 Erreur Moyenne Absolue en Pourcentage (MAPE)

La MAPE exprime l'erreur moyenne en termes de pourcentage des valeurs observées :

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3.14)$$

Cette métrique permet une interprétation intuitive de l'erreur relative et facilite la comparaison entre différents modèles ou jeux de données.

# Chapitre 4

## Implémentation et Résultats des Modèles

### 4.1 Régression Polynomiale

Différents degrés de polynômes (de 1 à 7) ont été testés pour modéliser la tendance du taux de chômage en fonction de la variable `date_num`. L'ensemble de données a été divisé en un ensemble d'entraînement (80%) et un ensemble de test (20%). La figure 4.1 montre les ajustements des différents modèles polynomiaux.

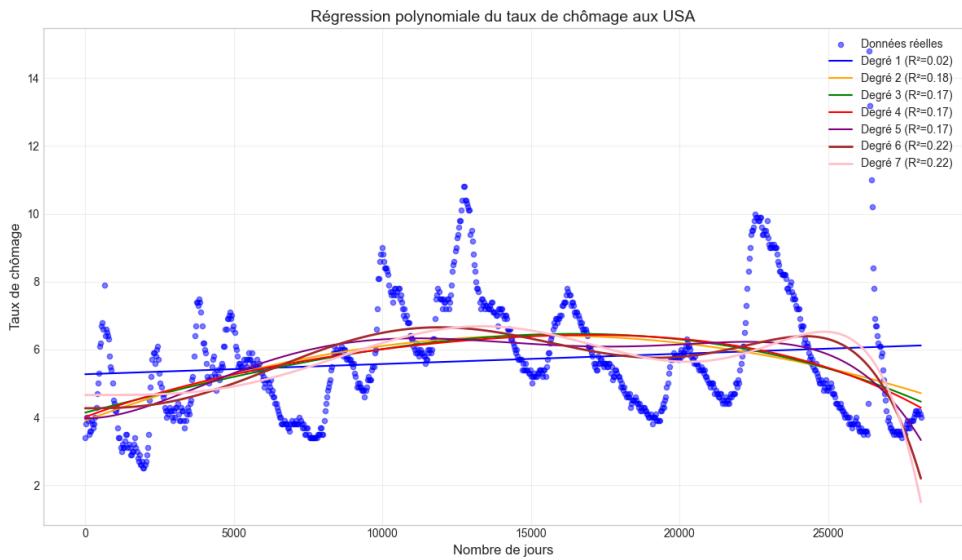


FIGURE 4.1 – Ajustements par régression polynomiale de différents degrés.

Le modèle polynomial de degré 6 a obtenu le meilleur score  $R^2$  sur l'ensemble de test, avec une valeur de 0.2179. Les scores  $R^2$  pour les différents degrés sont :

- Degré 6 :  $R^2 = 0.2179$
- Degré 7 :  $R^2 = 0.2160$
- Degré 2 :  $R^2 = 0.1766$
- Degré 5 :  $R^2 = 0.1745$
- Degré 3 :  $R^2 = 0.1723$
- Degré 4 :  $R^2 = 0.1653$
- Degré 1 :  $R^2 = 0.0228$

Bien que le degré 6 capture une certaine courbure, le score  $R^2$  global reste faible, indiquant que ce modèle n'explique qu'une petite partie de la variance du taux de chômage. La régression polynomiale simple est limitée pour prédire les dynamiques complexes inhérentes aux séries temporelles économiques.

## 4.2 Modèle Prophet

Le modèle Prophet a été appliqué à la série temporelle du taux de chômage. Les paramètres suivants ont été utilisés lors de l'initialisation du modèle pour tenter d'améliorer les performances :

- `changepoint_prior_scale=0.5` (flexibilité accrue de la tendance)
- `changepoint_range=0.95` (changepoints détectés sur 95% de l'historique)
- `n_changepoints=50` (nombre de points de changement potentiels)
- `seasonality_prior_scale=10` (force de la saisonnalité)
- `seasonality_mode='multiplicative'` (saisonnalité multiplicative)
- `yearly_seasonality=20` (ordre de Fourier pour la saisonnalité annuelle)

De plus, des saisonnalités mensuelles (période 30.5 jours, ordre 5) et un cycle économique de 8 ans (ordre 3) ont été ajoutés explicitement. Le modèle a été entraîné sur l'ensemble des données historiques, puis utilisé pour prédire les 24 prochains mois.

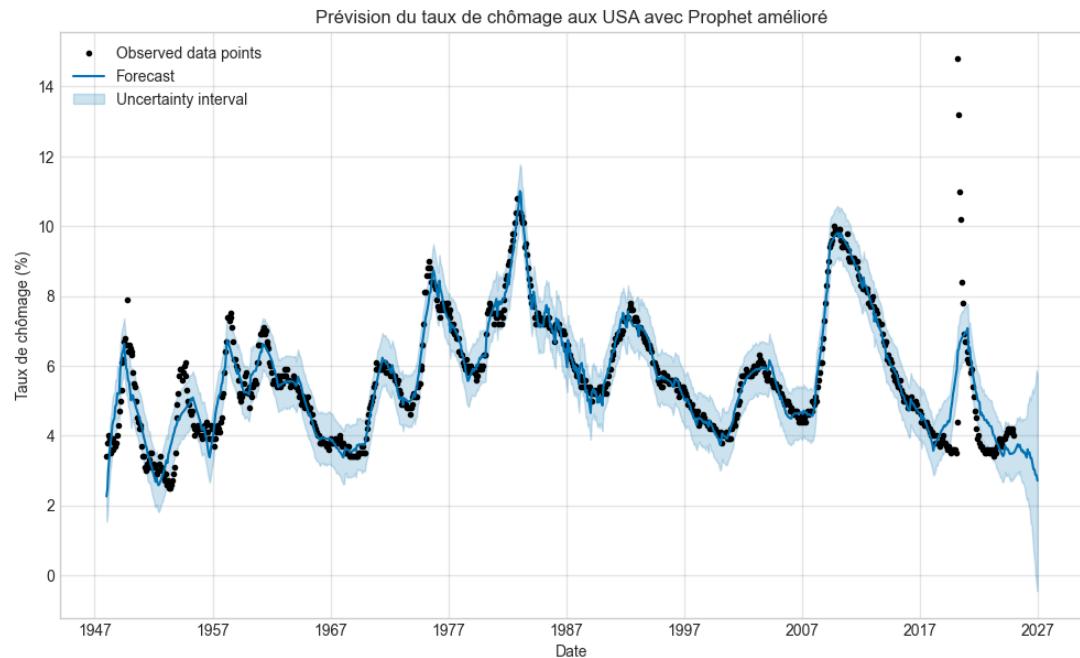


FIGURE 4.2 – Prévisions du taux de chômage avec Prophet (jusqu'en janvier 2027).

La figure 4.2 montre les données historiques, les prévisions du modèle, ainsi que l'intervalle d'incertitude. La figure 4.3 détaille les composantes identifiées par Prophet.

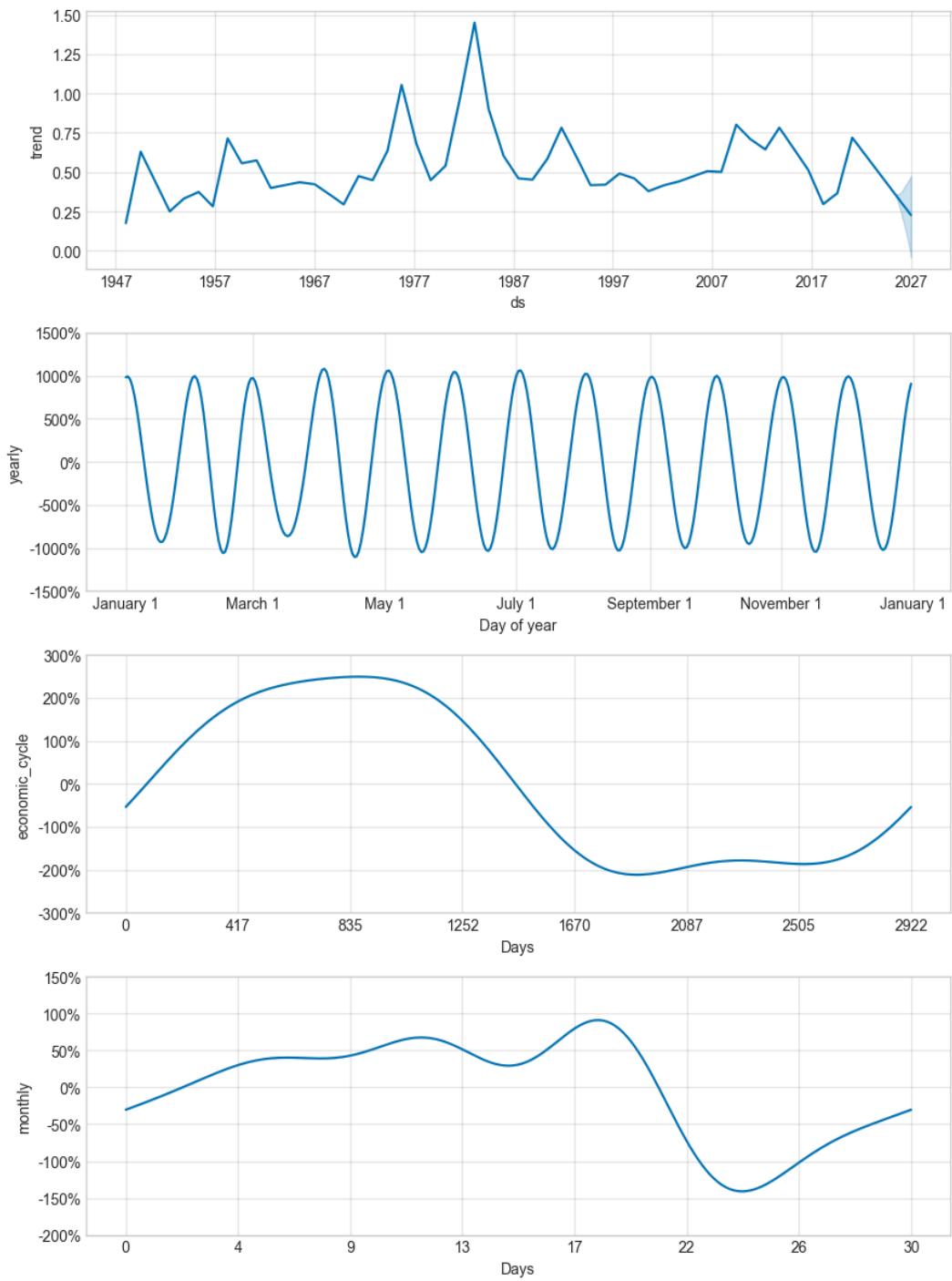


FIGURE 4.3 – Décomposition des composantes par Prophet (Tendance, Saisonnalités annuelle, mensuelle, et cycle économique).

Les métriques de performance du modèle Prophet sur l'ensemble des données historiques sont :

- Erreur absolue moyenne (MAE) : 0.323
- Erreur quadratique moyenne (RMSE) : 0.581
- Erreur moyenne absolue en pourcentage (MAPE) : 6.276%

Ces métriques indiquent un ajustement raisonnable aux données historiques.

## 4.3 Modèle ARIMA/SARIMA

L'application d'un modèle ARIMA (ou SARIMA) a débuté par un test de stationnarité.

- **Test de Dickey-Fuller Augmenté (ADF)** : La p-value obtenue est de 0.001966. Étant inférieure au seuil de significativité de 0.05, l'hypothèse nulle (présence d'une racine unitaire, non-stationnarité) est rejetée. La série est donc considérée comme stationnaire ou

devenant stationnaire après une différenciation d'ordre faible.

Ensuite, la fonction `auto_arima` de la bibliothèque `pmdarima` a été utilisée pour identifier automatiquement les meilleurs paramètres ( $p,d,q$ ) et ( $P,D,Q,m$ ) pour un modèle SARIMA.

- **Meilleur modèle sélectionné par `auto_arima`** : ARIMA(1,1,1) sans composante saisonnière ( $m=0$ ). Le paramètre  $d=1$  indique qu'une différenciation d'ordre 1 a été jugée nécessaire par `auto_arima` pour atteindre la stationnarité, malgré le résultat initial du test ADF sur la série brute.

Le modèle ARIMA(1,1,1) final a été ajusté sur l'ensemble des données. La figure 4.4 montre l'ajustement du modèle aux données historiques, et la figure 4.5 présente les prévisions pour les 24 prochains mois avec les intervalles de confiance.

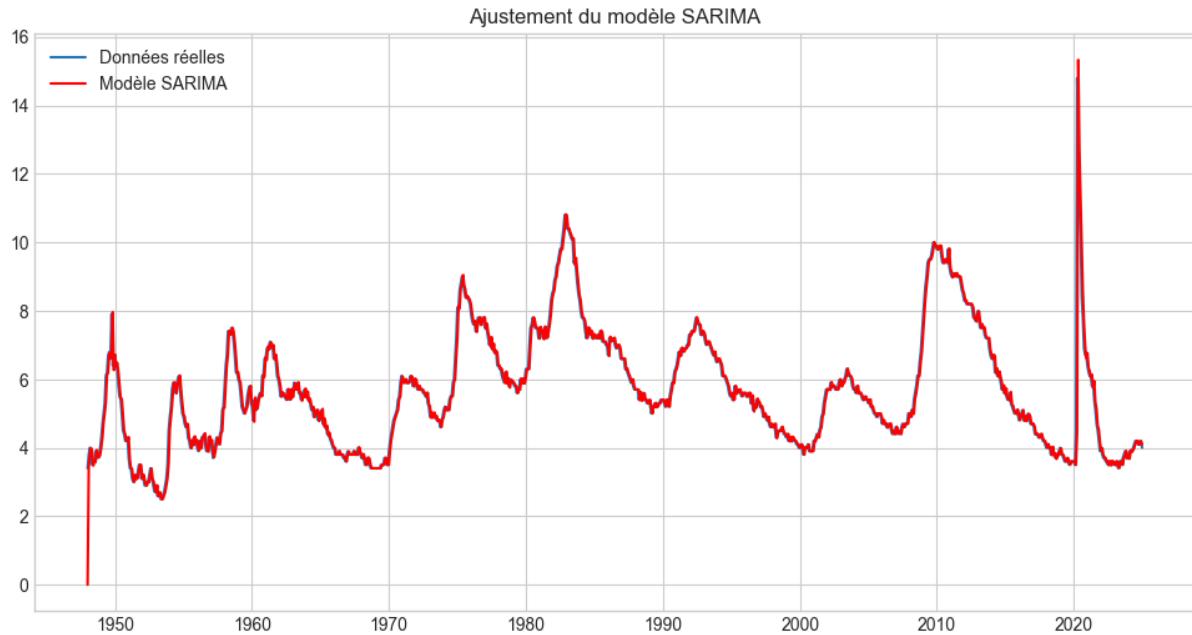


FIGURE 4.4 – Ajustement du modèle ARIMA(1,1,1) sur les données historiques.

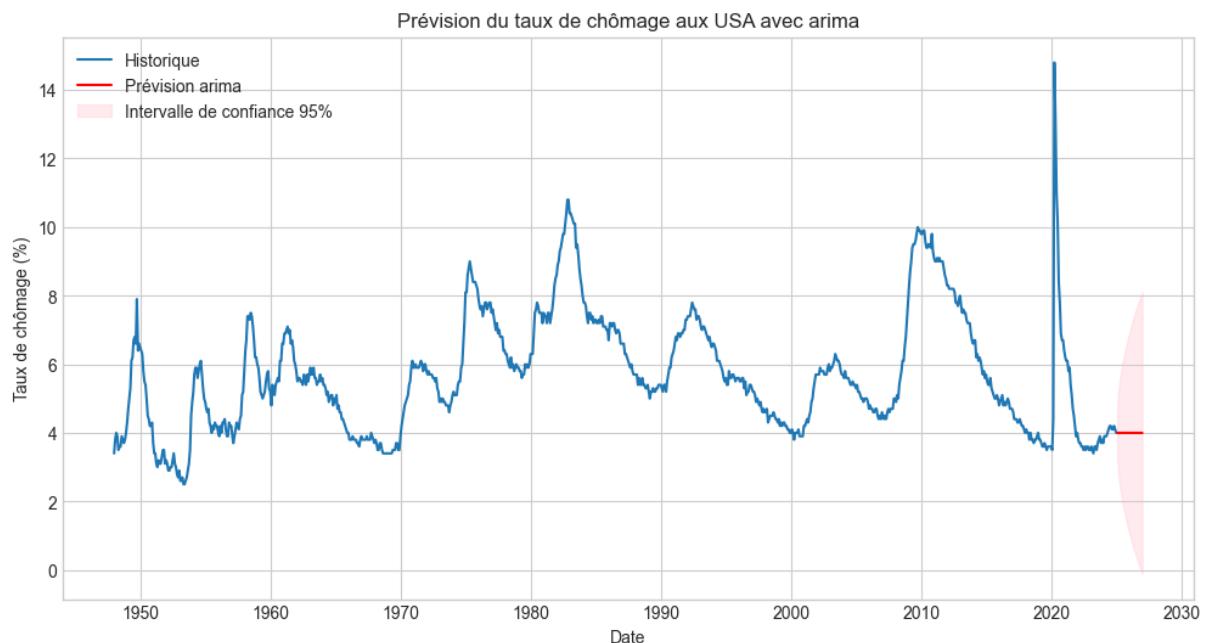


FIGURE 4.5 – Prévisions du taux de chômage avec ARIMA(1,1,1) (jusqu'en janvier 2027).

Les métriques de performance du modèle ARIMA(1,1,1) sur les données d'entraînement (après différenciation) sont :

- Erreur absolue moyenne (MAE) : 0.167
- Erreur quadratique moyenne (RMSE) : 0.415
- Erreur moyenne absolue en pourcentage (MAPE) : 2.89%
- Coefficient de détermination ( $R^2$ ) : 0.941

Ces valeurs indiquent un très bon ajustement du modèle ARIMA aux données historiques.

#### 4.3.1 Remarque sur l’Absence de Composante Saisonnière

Il est important de noter que bien que la fonction `auto_arima` ait été configurée pour détecter automatiquement les composantes saisonnières potentielles (avec le paramètre `seasonal=True` et `m=12` pour une saisonnalité mensuelle), le modèle optimal sélectionné est un ARIMA(1,1,1) sans composante saisonnière.

Cette absence de composante saisonnière dans le modèle final peut s’expliquer par plusieurs facteurs :

- **Saisonnalité faible** : Comme observé lors de la décomposition saisonnière (Figure 2.2), l’amplitude de la composante saisonnière est relativement faible (0.149 points de pourcentage) par rapport à la variabilité totale de la série.
- **Critère d’optimisation** : La fonction `auto_arima` utilise le critère AIC (Akaike Information Criterion) qui pénalise la complexité du modèle. L’ajout de paramètres saisonniers (P,D,Q) n’améliore pas suffisamment l’ajustement pour compenser l’augmentation de complexité.
- **Prédominance de la tendance** : Les variations du taux de chômage américain sont principalement dictées par les cycles économiques et les chocs exogènes plutôt que par des variations saisonnières régulières.

Des tests manuels avec des modèles SARIMA de différents ordres ont confirmé que l’ajout de composantes saisonnières n’apportait pas d’amélioration significative des métriques de performance, justifiant ainsi le choix du modèle ARIMA(1,1,1) plus parcimonieux.

## Chapitre 5

# Comparaison des Modèles et Discussion

La comparaison des modèles se base sur leurs performances sur les données historiques (période d'entraînement) et la nature de leurs prévisions. La figure 5.1 superpose les prévisions de Prophet et ARIMA jusqu'à la fin de 2027.

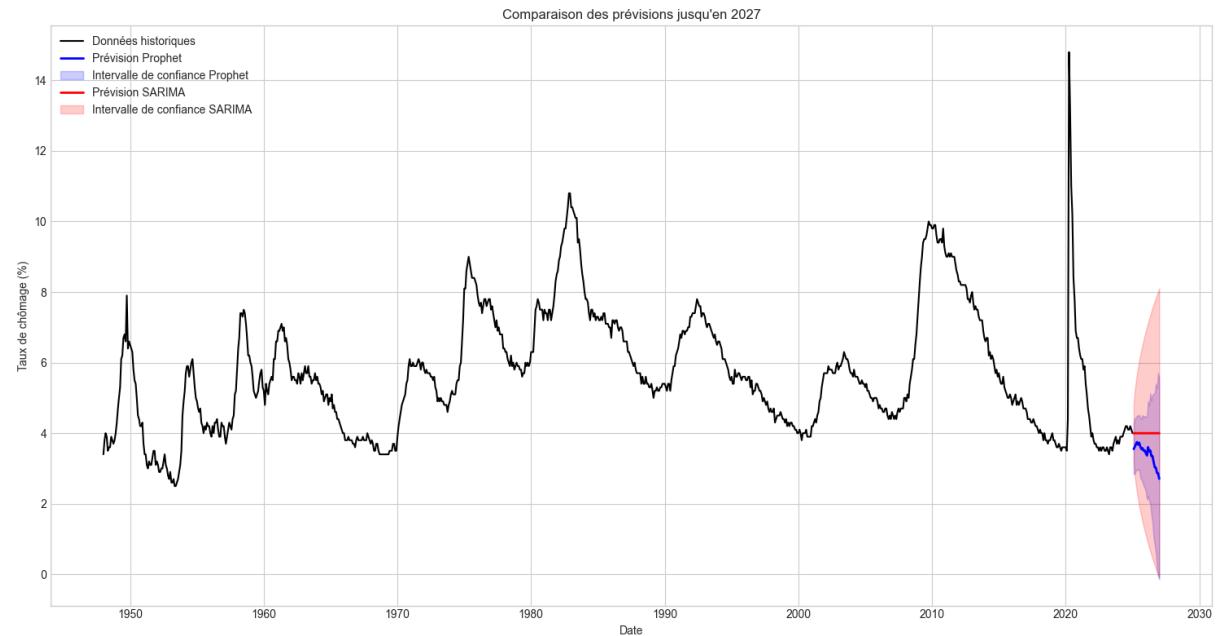


FIGURE 5.1 – Comparaison des prévisions du taux de chômage avec Prophet et ARIMA(1,1,1) jusqu'en 2027.

Le tableau 5.1 résume les métriques de performance des deux modèles.

TABLE 5.1 – Comparaison des métriques de performance des modèles Prophet et ARIMA.

Métrique	Prophet	ARIMA	Meilleur Modèle
MAE (Erreur Absolue Moyenne)	0.3228	0.1669	ARIMA
RMSE (Racine de l'Erreur Quadratique Moyenne)	0.5810	0.4147	ARIMA
MAPE (Erreur Absolue en Pourcentage)	6.2756%	2.8853%	ARIMA
Coefficient de Détermination ( $R^2$ )	0.8843	0.9410	ARIMA

Analyse des résultats :

- **Précision** : Le modèle ARIMA(1,1,1) surpassé significativement le modèle Prophet sur toutes les métriques d'erreur (MAE, RMSE, MAPE) et présente un meilleur score R<sup>2</sup>. Cela indique un ajustement plus précis aux données historiques.
- **Prévisions futures** :
  - Les deux modèles prévoient une relative stabilité du taux de chômage autour de 4% pour les deux prochaines années.
  - Les prévisions de Prophet (ligne bleue pointillée) tendent à revenir vers une moyenne à long terme après avoir capturé les cycles initiaux, et son intervalle de confiance s'élargit plus rapidement, reflétant une plus grande incertitude.
  - Les prévisions ARIMA (ligne rouge tiretée) montrent également une tendance à la stabilisation, avec un intervalle de confiance qui s'élargit plus lentement, suggérant une plus grande confiance dans la persistance de la tendance actuelle.
- **Interprétabilité** : Prophet offre une meilleure interprétabilité grâce à sa décomposition explicite en tendance, saisonnalités, et cycles. ARIMA, bien que statistiquement robuste, est plus une « boîte noire » en termes d'explication des composantes sous-jacentes.
- **Complexité et robustesse** : Prophet est conçu pour être robuste aux données manquantes et aux changements structurels, et il est relativement facile à mettre en œuvre même sans une expertise approfondie en séries temporelles. ARIMA requiert une analyse plus poussée (stationnarité, identification des ordres p,d,q) mais peut être très performant si la série correspond bien à ses hypothèses.

En conclusion, pour la prédiction purement quantitative sur cette série de données, ARIMA(1,1,1) semble être le modèle supérieur. Cependant, Prophet reste un outil précieux pour l'analyse des composantes et pour des scénarios où l'interprétabilité et la robustesse aux changements sont prioritaires.

## Chapitre 6

# Application Streamlit pour la Visualisation des Prédictions

Pour rendre les résultats de la modélisation plus accessibles, interactifs et exploitables par un public plus large, une application web simple a été esquissée à l'aide de la bibliothèque Python Streamlit. Cette application a pour objectif principal de permettre une exploration dynamique des données historiques et des prévisions générées par les modèles.

L'interface utilisateur est conçue pour être intuitive. Elle se compose typiquement d'un panneau latéral (sidebar) permettant de configurer les paramètres de la prévision et d'une zone principale affichant les résultats sous forme tabulaire et graphique.

Les fonctionnalités clés envisagées pour cette application incluent :

- **Choix de l'horizon de prévision :** L'utilisateur peut spécifier le nombre de périodes futures à prédire. Cette durée peut être définie en nombre de mois ou en nombre d'années, offrant une flexibilité pour des analyses à court ou moyen terme.
- **Visualisation des données historiques :** Un graphique interactif affiche l'évolution passée du taux de chômage.
- **Affichage des prévisions :**
  - **Tableau des prévisions :** Un tableau détaillé présente les valeurs prédites pour chaque période future, incluant potentiellement les intervalles de confiance (limites inférieure et supérieure). La figure 6.1 illustre un exemple de cette vue tabulaire, où les prédictions des modèles Prophet et ARIMA sont affichées avec leurs dates correspondantes.
  - **Graphique des prévisions :** Les prédictions sont superposées aux données historiques sur un graphique. Ce graphique est interactif, permettant de zoomer, de se déplacer, et d'inspecter les valeurs. La figure 6.2 montre une vue d'ensemble de ce graphique, combinant les données réelles et les prévisions des modèles.
- **Comparaison des modèles :** En affichant simultanément les prévisions de différents modèles, l'utilisateur peut visuellement comparer leurs performances et leurs trajectoires futures.

The screenshot shows a Streamlit application interface. On the left, a sidebar titled "Paramètres de Prévision" (Forecasting Parameters) includes dropdowns for "Unité de temps" (Months) and "Nombre à prévoir" (12), and a red "Lancer la prévision" (Launch Forecast) button. The main area is titled "Prévision du Taux de Chômage US" (US Unemployment Rate Forecast). It displays a table titled "Tableau des Prévisions" with columns: Date, Prophet, Prophet\_low, Prophet\_high, ARIMA, ARIMA\_low, and ARIMA\_high. The table contains 10 rows of data from February 2025 back to January 2025.

	Date	Prophet	Prophet_low	Prophet_high	ARIMA	ARIMA_low	ARIMA_high
0	2025-02-01	3.5503	2.7724	4.2602	3.9979	3.4664	4.5294
1	2025-03-01	3.6063	2.8785	4.3884	3.9995	3.2269	4.7721
2	2025-04-01	3.683	2.8987	4.4782	3.9983	3.0568	4.9398
3	2025-05-01	3.7476	2.972	4.5168	3.9992	2.906	5.0925
4	2025-06-01	3.6742	2.9116	4.4515	3.9985	2.7783	5.2187
5	2025-07-01	3.7322	2.9687	4.5171	3.9991	2.6596	5.3386
6	2025-08-01	3.6284	2.773	4.469	3.9986	2.5527	5.4445
7	2025-09-01	3.5475	2.6708	4.3675	3.999	2.4518	5.5462
8	2025-10-01	3.5765	2.7048	4.4391	3.9987	2.358	5.6394
9	2025-11-01	3.502	2.5617	4.3885	3.9989	2.2686	5.7292

FIGURE 6.1 – Vue tabulaire des prévisions dans l’application Streamlit. Affiche les dates, les valeurs prédites par les modèles (Prophet, ARIMA) et leurs intervalles de confiance.



FIGURE 6.2 – Visualisation graphique interactive des données historiques et des prévisions du taux de chômage. Les différentes couleurs représentent les données réelles, les prédictions de Prophet et celles d’ARIMA.

La figure 6.3 illustre un cas d’usage spécifique où l’utilisateur a choisi de générer des prévisions pour les 3 prochaines années (soit 36 mois). Le graphique montre clairement la continuation de la série avec la zone de prédiction et les intervalles de confiance associés. Cette flexibilité dans le choix de l’horizon de prévision est un atout majeur de l’application.

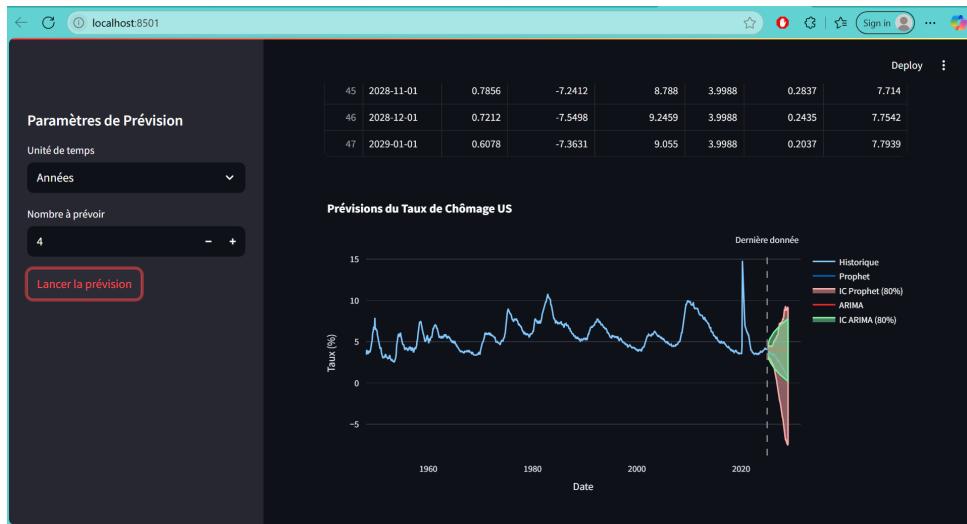


FIGURE 6.3 – Exemple de prévision sur un horizon de 3 ans dans l’application Streamlit. L’utilisateur peut ajuster cette période via les paramètres de prévision, en choisissant un nombre de mois ou d’années.

Une telle application transforme une analyse de modélisation statique en un outil dynamique d’aide à la décision. La sauvegarde préalable des modèles entraînés (Prophet et ARIMA) à l’aide de bibliothèques comme `joblib` est une étape cruciale. Cela permet à l’application Streamlit de charger rapidement les modèles pré-entraînés et de générer des prévisions à la demande, sans nécessiter un ré-entraînement coûteux en temps à chaque interaction de l’utilisateur.

## Chapitre 7

# Conclusion et Perspectives

Cette étude a exploré trois méthodes différentes pour la prédiction du taux de chômage aux États-Unis.

La **régression polynomiale**, bien que simple à mettre en œuvre, s'est avérée peu performante pour capturer la complexité de la série temporelle du chômage, avec un  $R^2$  maximal de 0.2179 pour un polynôme de degré 6. Ce modèle sert principalement de baseline et illustre les limites des approches purement tendancielles sans prise en compte des dynamiques temporelles.

Le modèle **Prophet** a offert un ajustement plus satisfaisant (MAE 0.323, RMSE 0.581, MAPE 6.28%), démontrant sa capacité à modéliser la tendance, les saisonnalités (annuelle, mensuelle) et les cycles économiques à plus long terme. Son principal atout réside dans sa flexibilité et l'interprétabilité de ses composantes.

Le modèle **ARIMA**, après sélection automatique des paramètres aboutissant à un ARIMA(1,1,1), a fourni les meilleures performances prédictives sur les données historiques (MAE 0.167, RMSE 0.415, MAPE 2.89%,  $R^2$  0.941). Cela suggère que les dynamiques autorégressives et de moyenne mobile, après une différenciation, capturent bien le comportement de la série du taux de chômage.

En comparant Prophet et ARIMA, ce dernier s'est montré plus précis pour l'ajustement aux données passées. Les prévisions des deux modèles pour les deux prochaines années sont relativement similaires, indiquant une stabilisation du taux de chômage. Cependant, l'intervalle de confiance de Prophet s'élargit plus vite, suggérant une plus grande prudence pour les prévisions à long terme.

### Limites et Perspectives :

- Les modèles de séries temporelles sont, par nature, basés sur les données passées et peuvent mal performer face à des chocs exogènes imprévus (crises économiques majeures, pandémies, changements politiques drastiques).
- L'inclusion de variables exogènes (par exemple, indicateurs économiques avancés, taux d'intérêt, politiques fiscales) dans des modèles comme SARIMA ou Prophet avec régresseurs pourrait améliorer la robustesse et la précision des prévisions.
- D'autres approches de modélisation, telles que les modèles d'apprentissage automatique (LSTM adaptés aux séries temporelles) pourraient être explorées.
- Une validation croisée temporelle plus rigoureuse et une évaluation sur des périodes hors échantillon distinctes seraient nécessaires pour confirmer la robustesse des modèles.

En définitive, cette étude démontre l'utilité des modèles Prophet et ARIMA pour la prédiction du taux de chômage, ARIMA offrant une meilleure précision quantitative et Prophet une plus grande richesse interprétative.

**Note :** L'ensemble du code et des ressources de ce projet sont disponibles sur le dépôt GitHub dédié : <https://github.com/Younessboumlik/unemployment-forecasting>.