Projet ML

YEHOSSOU Kakpo

May 20, 2023

Abstract

Dans ce document, nous allons expliquer notre travail qui consistera à utiliser le modèles ARIMA pour prédire la température en France.

Contents

1 Introduction

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est une méthode de prévision de séries temporelles qui prend en compte à la fois les composantes auto-régressives (AR) et moyennes mobiles (MA) de la série, ainsi que la différenciation intégrée (I) pour rendre la série stationnaire. Les modèles ARIMA sont généralement exprimés en termes de trois paramètres : p (ordre AR), d (ordre de différenciation) et q (ordre MA). Les modèles ARIMA sont largement utilisés dans la prévision de séries temporelles économiques, financières et de production, ainsi que dans d'autres domaines tels que la climatologie et l'épidémiologie.

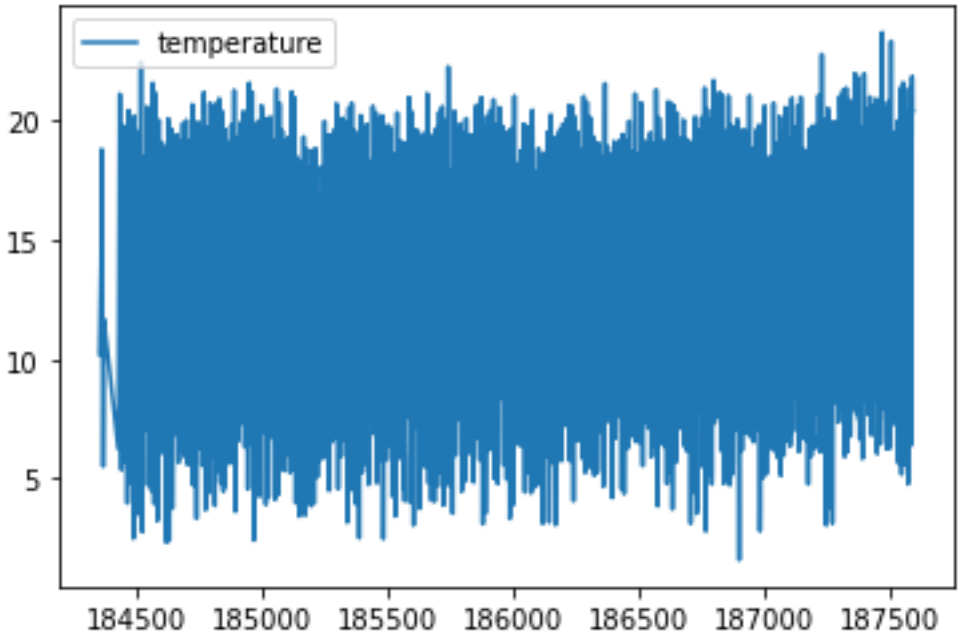

Figure 1: Caption

2 Méthodologie

Pour rendre stationnnaire notre données dans le but de faire notre étude efficacement, nous avons:

- Analyser et nettoyer la données;
- Analyser les résidus, par la méthode de décomposition saisonnier de la bibliothèque statsmodels
- Lissage (méthode de log et différentielle)
- Test de Dickey Fuller Augmenté pour vérifier la stationnarité
- Sélection : Sélection du bon modèle par la méthode de Box Jen Kins
- Choix et mésure
- Evaluation de la performance
- Comclusion

3 Analyser et Néttoyage de la données

Elles nous a permit de voir les éléments nuls, le type de chaque colonne, le nombre de ligne et de colonne. Mieux nous permet de remplacer les éléments nuls par des valeurs proches en utilisant l'interpolation linéaire.

3.1 Analyser des résidus, par la méthode de décomposition seasonnier de la bibliothèque statsmodels

Cette méthode nous a permis d'identifier si nous somme en présence d'un modèle additif(figure 2) ou multiplicatif(Figure3)

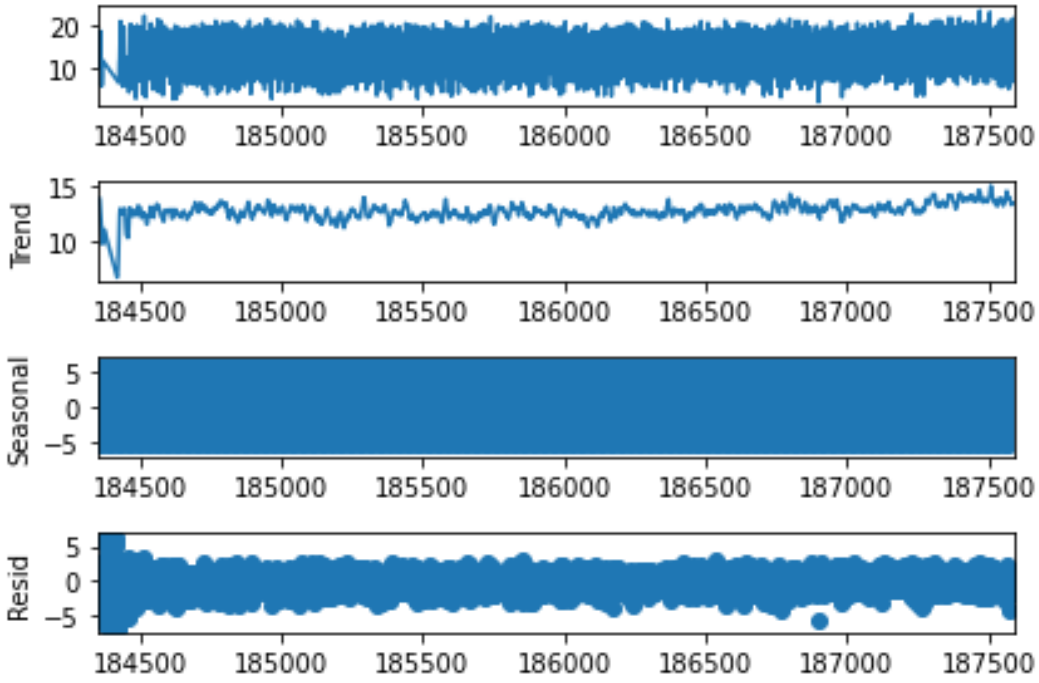

Figure 2: Modèle Additif

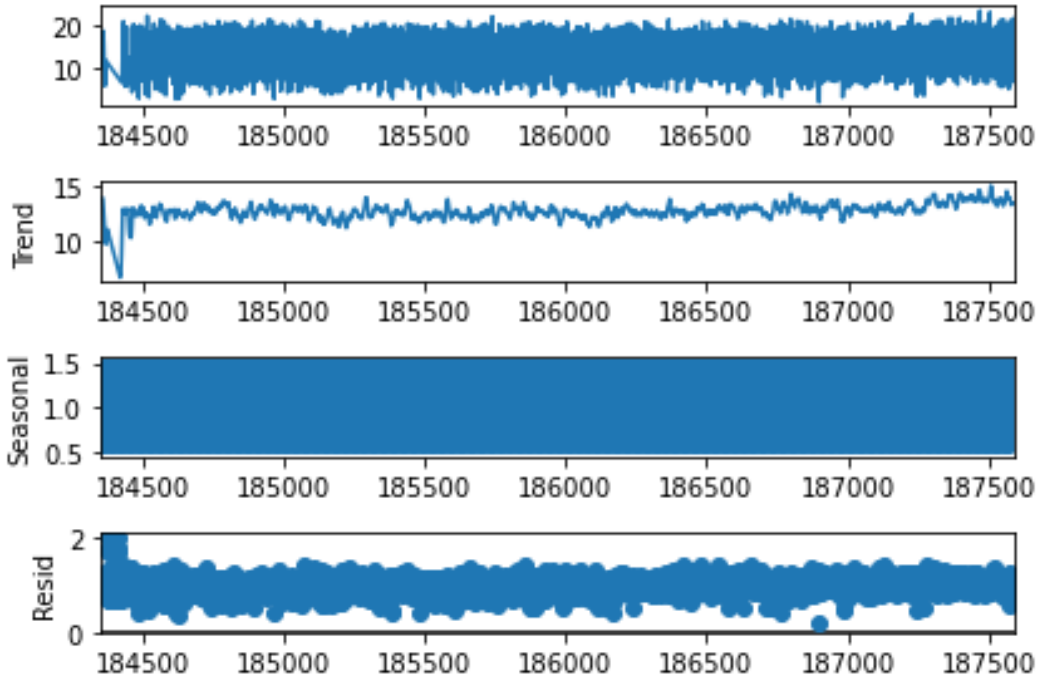

Figure 3: Modèle multiplicatif

3.2 Lissage (méthode de log et différentielle)

Le lissage encore appeler le filtrage, il nous permet d'enlever le bruit.Permet de décomposer la tendance, la saisonarité le et le bruit. ici nous vont utiliser la méthode de log et différentielle. Nous avons les résultats suivant (Figure 4) :

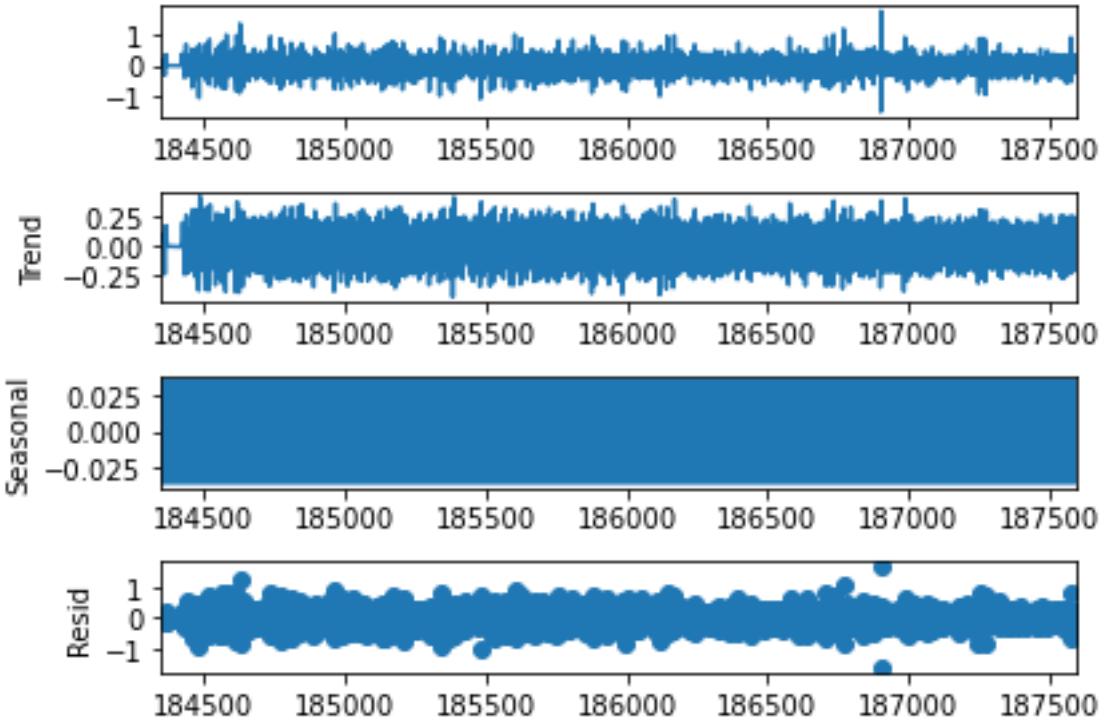

Figure 4: Lissage

3.3 Test de Dickey Fuller Augmenté pour vérifier la stationnarité

Pour avoir de la stationnarité, nous devrons avoir :

$$\mathrm{E}(x_t) = \nu \tag{1}$$

$$\mathrm{Var}(x_t) = \gamma^2(2)$$

$$\mathrm{Cov}(x_t, x_{t+h}) = \lambda(h) \tag{3}$$

Qui ne dépendent donc pas du temps. Ce pendant le test de Dickey Fuller Augmenté(ADF), nous aide à vérifier la stationnarité d'un modèle. Et il est important de retenir que les hypothèses de la stationnarité doivent être respecter pour appliquer "AR", "ARMA" et autres. Intéressons nous à ARIMA :

4 ARIMA

Lorsqu'un modèle n'est pas stationnaire, on peut différentier notre série de telle sorte à ce qu'elle soit stationnaire. le I dans le ARIMA fait référence à l'ordre d'intégration, il correspond à l'ordre de différentiation nécessaire pour obtenir la stationnarité. elle obéir aux équations suivantes:

$$Z_t = \alpha_0 + \alpha_1(x_{t-1} - x_{t-2}) + \epsilon_t + \theta_1\epsilon_{t-1} \tag{4}$$

Donc ça devient:

$$Z_t = \alpha_0 + \alpha_1 Z_{t-1} + \epsilon_t + \theta_1\epsilon_{t-1} \tag{5}$$

4.1 Présentation de la distribution

L'évolution de la distribution se présente comme suit à travers les différente méthodes :

4.1.1 Présentation de la distribution avant le log

Au départ la distribution se présentait comme suit :

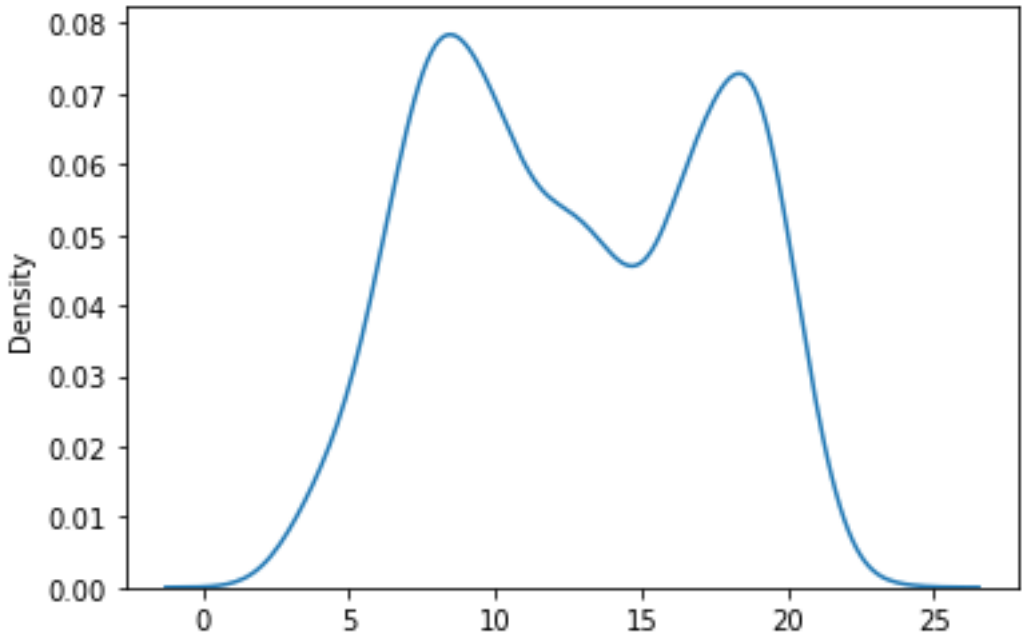

Figure 5: Présentation de la distribution la la donnée initiale

4.1.2 Évolution de la distribution avec la méthode log

La distribution se présente comme suit :

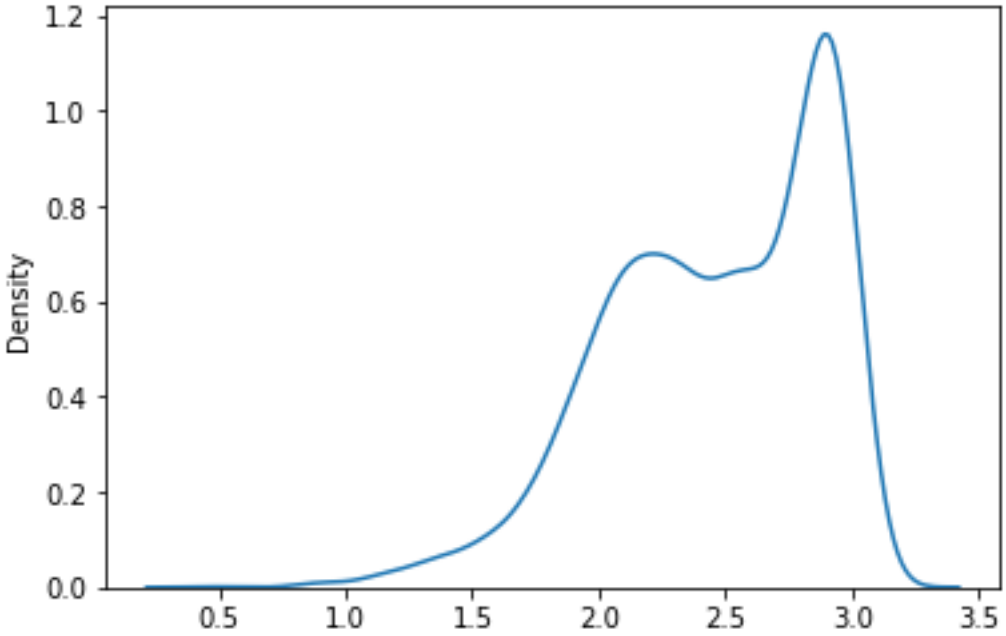

Figure 6: Distribution avec le log

4.1.3 Présentation de la distribution avec la méthode de différentiation

En différenciant notre variable de données, nous avons sur la Figure 6 une distribution centrée sur 0 :

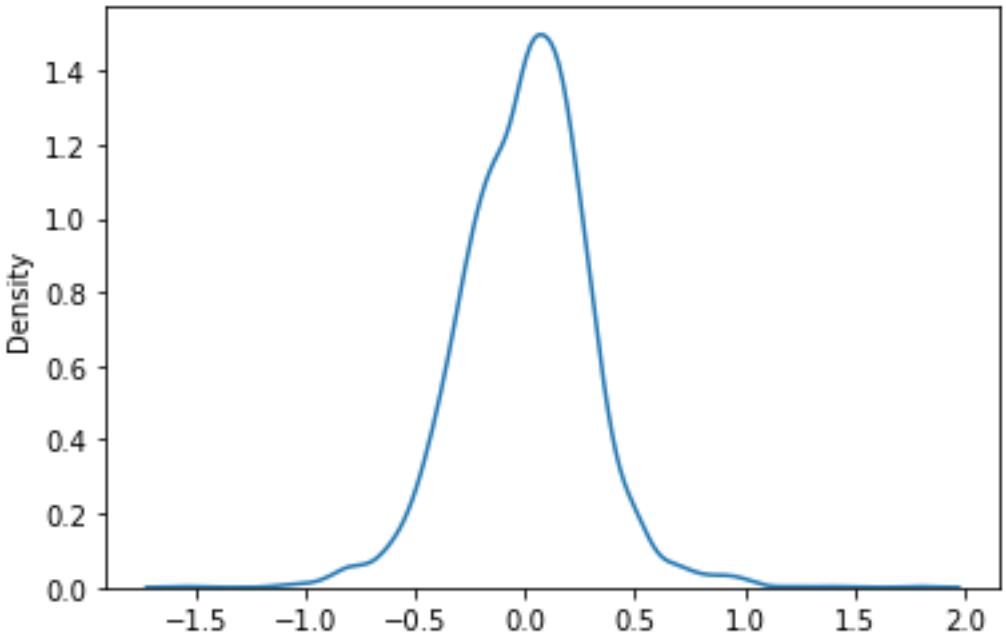

Figure 7: Distribution centrer

5 Sélection du bon modèle, méthode de Box-Jenkins

5.1 ACF

Communément appelée Fonction d'autocorrélation : C'est la corrélation de Pearson entre les valeurs d'une série. Par exemple entre X_t et X_{t-2}

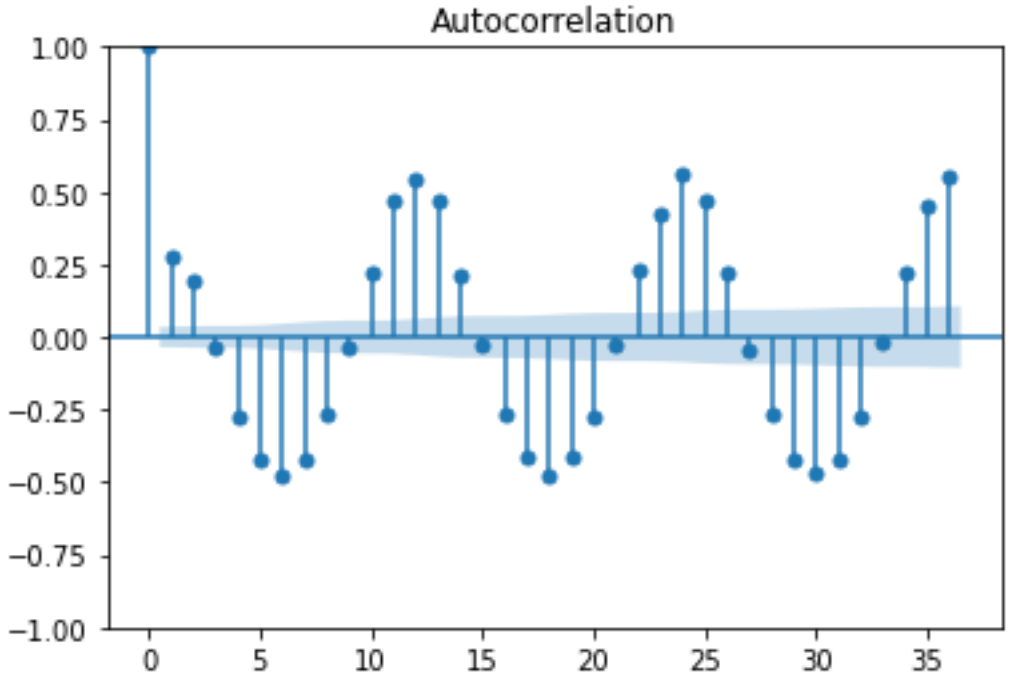

Figure 8: Courbe ACF

5.2 PACF

Le PACF qui est le partial autocorrélation function est aussi une corrélation de Pearsn entre les valeurs d'une serie isolée de limpact des autres valeurs de la série.

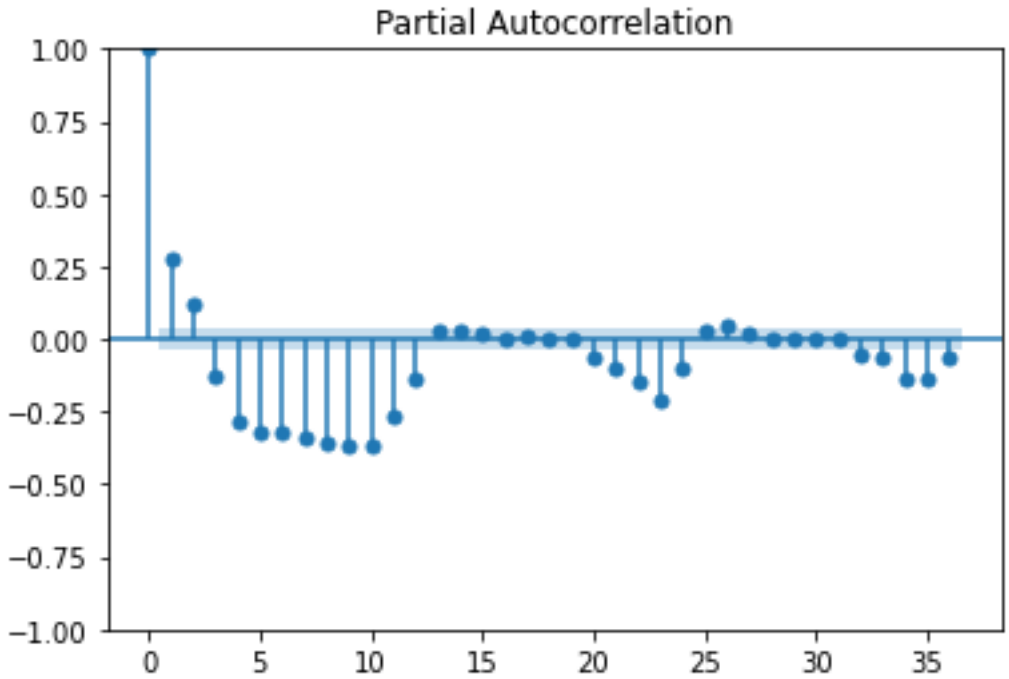

Figure 9: Courbe PACF

Méthodes	AR(p)	MA(q)	ARMA(p,q)
ACF	Tends vers 0	s'annule après l'ordre de q	Tends vers 0
PACF	S'annule après l'ordre de p	Tends vers 0	Temps vers 0

Avec La différentiation, le faite que p et q s'annule avec une périodicité, on a choisir ARIMA, on aurait pu choisir SARIMA pour faire intervenir la saisonnarité mais en tenons nous en sur ARIMA. Avec un peux habileté on choisissions comme paramètre p =12, d=1, q= 12.

- Dans le main.py : nous avons l'appel des différentes classe et fonctions
- Dans le prepro.py nous avons tous ce qui concerne le préprocessing de la données
- Dans le prepare.py il y a la fonction qui divise la données en données d'entraînement, teste et validation
- Un fichier summary regroupe toute les informations possible de l'entraînement.
- Dans le lire.py nous avon la classe du modèle ARIMA : ARIMAModel, où nous avons des fonction de fit et de predict qui nous donne le graphique suivant :

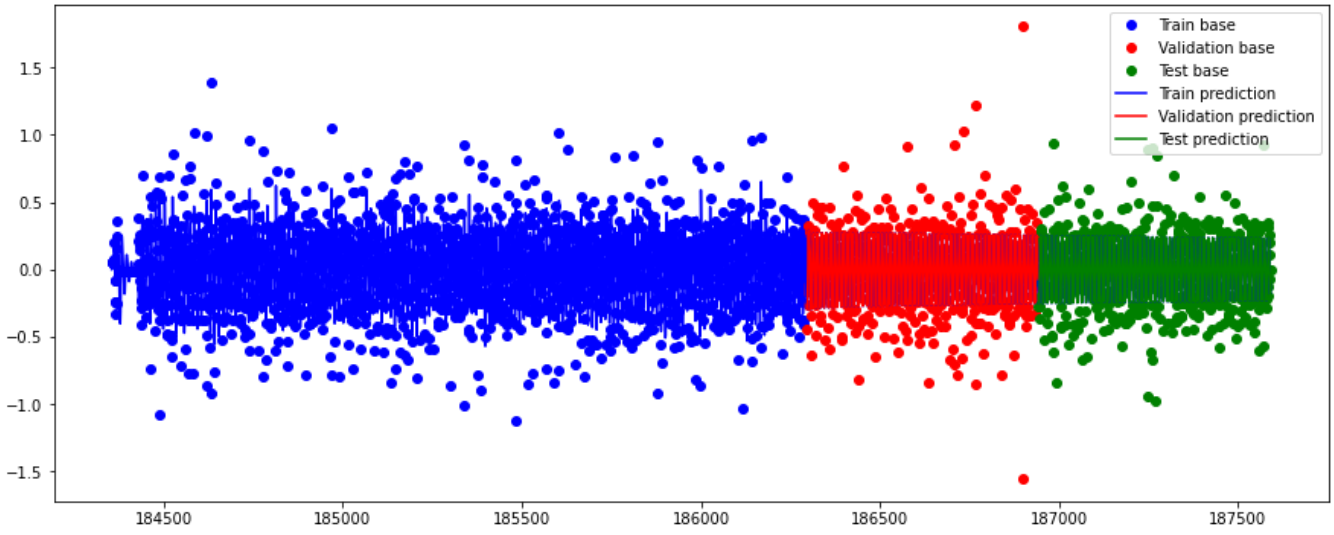

Figure 10: Resultat final du modèle ARIMA

6 Performance

Pour la performance du SEmodèle, nous avon préférer le RMSE qui est : 0.17.
Ce qui suggère que les prédictions du modèle ont en moyenne une différence de 0.17 avec les valeurs réelles. Cela indique une performance assez précise du modèle.

7 Conclusion

Nous aurons pu introduire, la notion de saisonnalité et utiliser le SARIMAX.