MARIE-ANGE DIENG

Rapport Synthétique : Analyse de la Taux de Fécondité par Pays

1. Problématique étudiée

La problématique de cette analyse concerne l'étude des tendances temporelles du taux de fécondité total (en nombre de naissances vivantes par femme) pour un ensemble de pays sélectionnés. Ce taux est un indicateur clé pour comprendre les dynamiques démographiques à l'échelle mondiale. L'objectif principal de ce projet est d'examiner l'évolution de ce taux au fil du temps pour **la France** afin d'en dégager des tendances, d'étudier la saisonnalité, et de prédire les futures évolutions du taux de fécondité en utilisant des modèles statistiques.

Les données, provenant des Nations Unies, couvrent la période de 1950 à 2023.

2. Outils, Méthodes et Techniques Utilisées

2.1. Prétraitement des Données

Les premières étapes du projet ont consisté à préparer les données pour l'analyse. Cela inclut :

- **Suppression** de la colonne CODE
- **Conversion de la colonne YEAR** en format datetime pour permettre une analyse temporelle appropriée.
- **Transformation du dataset :** Mettre la colonne YEAR en index, la colonne COUNTRY en noms de colonnes et la colonne PERCENTAGE comme valeurs avec :

  ```
  df=df.pivot(index="Year", columns="Country", values="Percentage")
  ```

- **Vérification de l'échelle temporelle** : Annuelle dans ce cas, de 1950 à 2023
- **Traitement des valeurs manquantes et des valeurs aberrantes** : Pas de valeurs manquantes, et 8 valeurs aberrantes

2.2. Analyse Exploratoire

Une analyse statistique descriptive a été réalisée pour mieux comprendre les données, avec des visualisations du taux de fécondité pour chaque pays au fil des années, une étude de corrélation et de distribution des différentes variables.

Nous avons choisi d'étudier le

2.3. Stationnarité et Transformation des Données

Avant de pouvoir appliquer un modèle ARIMA, la stationnarité des séries temporelles a été vérifiée à l'aide du test de Dickey-Fuller. Il a été nécessaire de différencier la variable France **une fois** pour la rendre stationnaire.

2.4. Construction et Optimisation du Modèle

L'objectif était de prédire l'évolution du taux de fécondité en utilisant des modèles **ARIMA auto_arima et SARIMA** :

- L'**interprétation de l'ACF** et de la **PACF** a permis de déterminer les paramètres initiaux du modèle ARIMA (p=3 ; d=1 ; q=2)
- Le modèle ARIMA a été ensuite ajusté et entraîné.
- Le modèle **auto_arima** a été également testé, permettant d'automatiser la sélection des paramètres optimaux.

2.5. Évaluation du Modèle

L'évaluation a été réalisée par la comparaison entre les valeurs prédites et réelles, en calculant les indicateurs AIC, BIC, et RMSE (Root Mean Squared Error). Les tests de **normalité** et de **blancheur des résidus** ont été effectués pour valider la qualité de l'ajustement du modèle.

Pour ARIMA :	**Pour auto_arima :**
RMSE: 0.38956953627500546	AIC = -168.631
AIC = -166.376	BIC = -162.450

BIC = -154.013	HQIC = -166.224
HQIC = -161.560	Pour auto_arima :

Le modèle auto_arima a des valeurs AIC, BIC et HQIC plus faibles, ce qui signifie qu'il s'ajuste mieux au modèle tout en évitant la sur-adaptation.

Test de Ljung-Box (Q) :

auto_arima : Q = 0.32 (p-value = 0.57)

ARIMA : Q = 0.29 (p-value = 0.59)

Les deux modèles montrent des p-values élevées (> 0.05), ce qui signifie que les résidus des deux modèles semblent être indépendants et que les modèles ne souffrent pas de problèmes d'autocorrélation.

Le modèle auto_arima semble toutefois plus optimal.

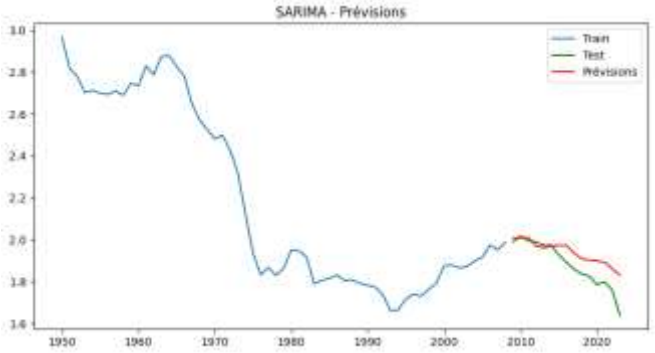
2.6. Bonus : SARIMAX

Un modèle SARIMAX a été construit pour prendre en compte à la fois la saisonnalité de la série temporelle et les effets exogènes. Le modèle spécifié est un **SARIMAX(3, 1, 2)(1, 1, 1, 5)**, où la composante saisonnière a une période de 5 ans.

AIC	-135.279
BIC	-119.516
HQIC	-129.217

L'**AIC** et le **BIC** sont relativement faibles, ce qui suggère que le modèle est relativement bien ajusté aux données tout en évitant le sur-ajustement.

Graphiquement, SARIMAX présente la courbe la plus proche des valeurs test :



3. Difficultés Rencontrées

Les principales difficultés rencontrées ont été dans le choix du dataset et le prétraitement, la détermination des paramètres des modèles à partir des courbes acf et pacf et l'évaluation de ces modèles.

4. Conclusion

À travers ce projet, il a été possible de mettre en évidence plusieurs tendances démographiques importantes pour la France. Les modèles ARIMA et auto_arima ont permis de réaliser des prédictions raisonnablement précises des taux de fécondité à venir. Cependant, l'intégration d'un modèle **SARIMAX**, prenant en compte la saisonnalité a fournis des prévisions encore plus robustes.

Les principaux enseignements tirés de l'analyse sont les suivants :

- Le **modèle auto_arima** s'avère plus optimal que le modèle ARIMA classique, notamment en raison de la sélection automatique des meilleurs paramètres.
- Le **modèle SARIMAX** est aussi plus performant au vu de l'influence de la saisonnalité.