# Analyse de données - TP 6

ISEP – 8 Décembre 2020

Instructions : Préparez un rapport incluant le code source et vos résultats, et déposez-le sur Moodle. Pas plus de 2 personnes par groupe. N'oubliez pas de mettre les 2 noms sur le rapport, ou de faire 2 rendus.

Dans ce TP, nous allons nous étudier l'évolution du chiffre d'affaires des supermarchés en France entre 2016 et 2020. Nous allons plus particulièrement nous intéresser à la vente de produits alimentaires et de carburant.

## Bibliothèques

Installez si nécessaire les biliothèques suivantes : pandas, numpy, matplotlib, sklearn, scipy et statsmodels. Importez les bibliothèques qui vous seront utiles en début de script :

```
#Imports
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from scipy import stats
import statsmodels.api as sm
from statsmodels.tsa.arima_model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
```

## A Analyse de la stationnarité d'un bruit blanc gaussien

1. Simulez un bruit blanc gaussien contenant 1000 échantillons, par exemple à l'aide de la commande *random.normal* de Numpy.
2. Affichez la série obtenue. Vous semble t-elle stationnaire ? Justifiez.
3. Affichez et décrivez la fonction d'autoccorélation (ACF en anglais) et d'autocorrélation partielle (PACF en anglais). Vous pouvez adapter le code suivant :

```
    fig = plt.figure(figsize=(12,8))
    ax1 = fig.add_subplot(211)
    fig = sm.graphics.tsa.plot_acf(x, lags=20, ax=ax1)
    ax2 = fig.add_subplot(212)
    fig = sm.graphics.tsa.plot_pacf(x, lags=20, ax=ax2)
```

4. Appliquez les tests d'Augmented Dickey Fuller (ADF) et de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) sur la série. En utilisant les **deux** résultats obtenus, concluez sur la stationnarité de la série testée. Vous trouverez des informations sur ces deux tests sur le tutoriel associé sur la page web de Statsmodels.

# B Analyse d'une série réelle

## B.1 Analyse de la stationnarité

Nous allons étudier la stationnarité du chiffre d'affaires de carburant et produits alimentaires entre 2016 et 2019. Pour ce faire, nous allons appliquer deux tests statistiques : le test de Box-Pierce, et celui de Shapiro-Wilk. Le test de Box-Pierce détermine si une série se compose principalement de bruit blanc ou non (i.e. $\forall t \quad \epsilon_t \sim i.i.d.(0, \sigma^2)$). Le test Shapiro-Wilk teste l'hypothèse nulle selon laquelle un échantillon est issu d'une population normalement distribuée.

1. Chargez le fichier *sales.csv* à l'aide de Pandas.
2. Appliquez les tests de Shapiro-Wilk (*stats.shapiro()*) et de Box-Pierce (*acorr_ ljungbox* de *statsmodels.stats.diagnostic*) aux chiffres d'affaires de carburant et produits alimentaires entre janvier 2016 et décembre 2019. En vous appuyant sur la réponse obtenue précedemment, que pouvez-vous dire sur la stationnarité de chacune de ces deux séries ? Si vous ne pouvez pas directement conclure, appliquez les questions 2 à 4 de la section A sur la série concernée.

Remarque : La fonction *acorr_ ljungbox* est habituellement appliquée aux **résidus** d'un modèle. Si vous l'appliquez sur la série brute, c'est donc la p-value d'ordre 0 pour le test de BoxPierce qui vous intéresse.

## B.2 Influence de la différenciation sur la stationnarité

1. Différenciez la **série non stationnaire** étudiée à la section B.1, *cad* créez une nouvelle série définie comme la différence entre l'observation originale et celle prise à l'instant précédent.
2. Affichez la série obtenue. Visuellement, vous semble t-elle stationnaire ?
3. Affichez et commentez les ACF et PACF de la série obtenue.
4. Appliquez les tests de ADF et KPSS sur la série obtenue. Concluez.

# C Ajustement d'un modèle ARIMA à une série réelle

Dans cette section, nous allons essayer d'ajuster un modèle ARMA(p, q) à la série temporelle du chiffre d'affaires du carburant. On suppose que celui-ci est seulement influencé par les ventes de carburant au cours des douze derniers mois.

1. Créez une variable $x_{train}$ contenant les valeurs du chiffre d'affaires du carburant entre janvier 2016 et décembre 2019, et une variable $x_{test}$ contenant les valeurs du chiffre d'affaires du carburant entre janvier 2020 et février 2020.
2. Si cela n'a pas déjà été fait, affichez les courbes ACF et PACF du chiffre d'affaire du carburant entre janvier 2016 et décembre 2019. Fixez bien les valeurs des paramètres. Quelles sont les valeurs de p et q admissibles d'après ces deux figures ?

Nous allons maintenant tester trois modèles : ARMA(1, 1), ARMA(1, 2) et ARMA(8, 2).

3. Pour chacun de ces modèles :
   — entraînez le modèle ARMA sur $x_{train}$ en utilisant la fonction *ARIMA*,
   — calculez le critère d'information bayésien (BIC en anglais),
   — calculez le critère d'information d'Akaike (AIC en anglais),
   — calculez l'erreur type,
   — calculez le log-vraisemblance.
4. Quel est le meilleur modèle selon les valeurs BIC et AIC ?
5. Appliquez un test de Shapiro-Wilk et de Box-Pierce sur les **résidus** des trois modèles. Pouvez-vous conclure sur la stationnarité de ces résidus en vous appuyant sur la réponse donnée à la section A.1 ? Dans le cas contraire, utilisez les tests de ADF et KPSS sur les résidus en question, et concluez sur leur stationnarité. Choisissez le meilleur modèle selon les p-valeurs obtenues.
6. Testez le meilleur modèle obtenu sur les données $x_{test}$. Affichez les prédictions et les valeurs attendues sur la même figure. Le résultat vous semble t-il satisfaisant visuellement ? Evaluez les performances quantitatives de votre modèle en calculant l'erreur quadratique moyenne entre les prédictions et les valeurs attendues.
7. Recommencez la question précédente en testant votre meilleur modèle sur le chiffre d'affaires du carburant entre janvier 2020 et septembre 2020. Concluez.