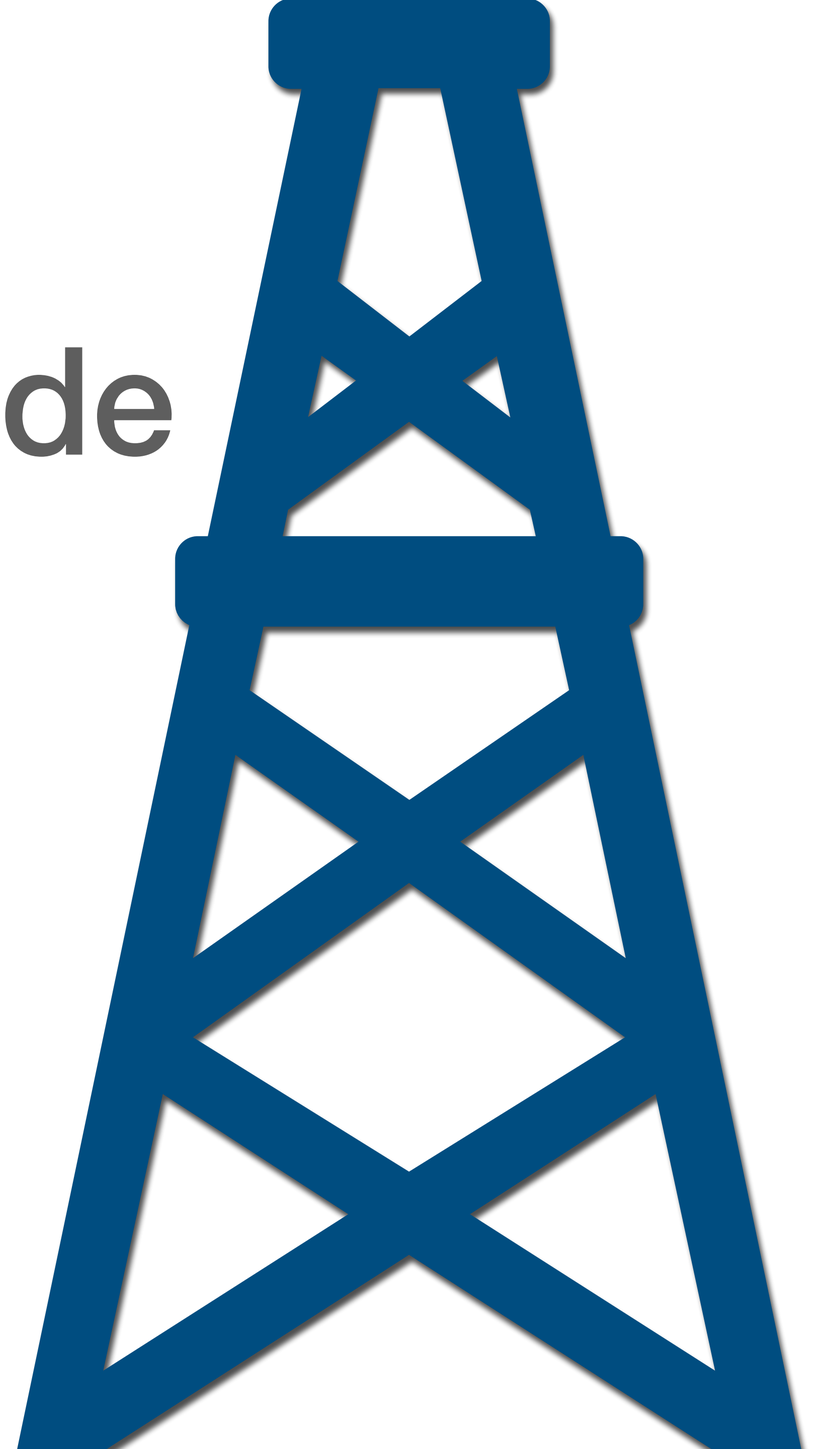
Prédiction de la demande en électricité



OPENCLASSROOMS

ENSAE-ENSAI
Formation continue
(Cepe)

Abderrahmane BOUAKKA - Formation Data Analyst - 2021_2022

SOMMAIRE

1.CONTEXTE

- Enercoop, société coopérative spécialisée dans les énergies renouvelables.

- La plupart de ces énergies sont intermittentes.

- Difficulté à prévoir les capacités de production d’électricité.

- La demande en électricité des utilisateurs varie au cours du temps.

- Dépend de paramètres comme la météo.

- L’enjeu est de mettre en adéquation l'offre et la demande.

2. Présentation des données utilisés

- **On a utilisé deux sources de données :**
    - **RTE (Réseau de Transport d’Electricité français) : pour la consommation totale d’énergie en France en Twh.**
    - **Cegibat qui est le centre d'expertise de GRDF : pour le DJU (Degré Jour Unifié) chauffage par région.**

- **Pré-traitement des données pour RTE:**
    - **Définition de la colonne ‘Mois’ comme index.**
    - **Sélection de la modalité ‘France’ dans la variable Territoire.**
    - **Projection de la variable ‘consommation totale’.**
    - **Prise en compte des valeurs supérieures à ‘0’.**

- **Pré-traitement des données pour CEGIBAT:**

    - **Concaténation des différents fichiers régions dans un data frame ‘DJU’.**

    - **Création d’une colonne année pour faire une moyenne nationale.**

    - **Remplacement des valeurs ‘str’ par une numérotation mensuelle et changement du ‘dtype’ en datetime.**

    - **Réorganisation du data frame ‘DJU’ en un nouveau dataframe ‘chauff_df’ contenant les données chauffage par mois.**

- **Pré-traitement des données RTE et CEGIBAT:**

  - **Jointure gauche et droite des dataframes ‘conso_df’ et ‘chauff_df’.**

Mois	Consommation totale	chauffage
2010-01-01	56342.0	471.142857
2010-02-01	48698.0	355.371429
2010-03-01	48294.0	300.557143
2010-04-01	38637.0	156.542857
2010-05-01	37284.0	119.985714
...	...	...

3. Détail du programme

- **Pour la correction avec la régression linéaire : on va simplement corriger notre série de l'effet température par la tendance, on va essayer d'estimer le modèle de régression linéaire suivant avec la méthode des moindres carrés:**

$$\mathbb{Y} = \mathbb{X}\beta + \epsilon$$

$$min_{(\alpha_0,\ldots,\alpha_n)\in\mathbb{R}^n} \sum_{T=1}^{T} (Y_t - T_t)^2$$

- **Pour ce faire, on va obtenir une estimation de $\hat{T}_t$, afin de corriger notre série temporelle :**

$$Y_t^{Ctemp} = Y_t - \hat{T}_t$$

```
|: # Régression linéaire
   reg= smf.ols('Consommation_totale ~ chauffage', data=df_1).fit()
   print(reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Consommation_totale   R-squared:                       0.960
Model:                             OLS   Adj. R-squared:                  0.960
Method:                  Least Squares   F-statistic:                     1827.
Date:                 Wed, 01 Jun 2022   Prob (F-statistic):           6.61e-55
Time:                         13:11:33   Log-Likelihood:                -678.08
No. Observations:                   78   AIC:                             1360.
Df Residuals:                       76   BIC:                             1365.
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3.195e+04    260.307    122.749      0.000    3.14e+04    3.25e+04
chauffage     50.7628      1.188     42.745      0.000      48.398      53.128
==============================================================================
Omnibus:                        1.459   Durbin-Watson:                   1.686
Prob(Omnibus):                  0.482   Jarque-Bera (JB):                0.852
Skew:                          -0.080   Prob(JB):                        0.653
Kurtosis:                       3.487   Cond. No.                         345.
==============================================================================
```

- Pour la désaisonnalisation avec les moyennes mobiles :

On a utilisé une moyenne mobile comme une combinaison linéaire d'instants passés et futurs qui nous a permis de mettre en évidence les différentes composantes de notre série temporelle.

Polynôme caractéristique de la moyenne mobile :

$$\Theta(z) = \sum_{i=0}^{m1+m2} \theta_{i-m1} z^i$$

On applique la moyenne mobile M sur la série :

$$MX_t = MT_t + MS_t + M\varepsilon_t$$

```
# Décomposition saisonière
decomp_x = seasonal_decompose(
    x_cvs,
    model="additive",
    filt=None,
    period=12,
    two_sided=True,
    extrapolate_trend=0);
```

Après avoir estimé la saisonnalité et obtenu les coefficients saisonniers on les a retranchés à notre série pour obtenir notre série temporelle désaisonnalisée :

$$X_t^{CVS} = X_t - \hat{S}_t$$

```
# On retranche les coef saisoniers aux consommation corrigées pour alimenter une nouvelle variable 'desaisonnalisee_
serie_Corrig_df['desaisonnalisee_corrigee']= x_cvs.values - decomp_x.seasonal.values
serie_Corrig_df
```

- Pour la méthode Holte-Winters :

On a supposé que X_t est approximable autour de T par les coefficients α, β et γ ainsi que s la période du cycle saisonnier de la série temporelle.

Soit $\hat{X}_T(\ell)$ la prévision de $X_{T+\ell}$ à l'instant T.

```
# application de la méthode Holt-Winters
h_W = ExponentialSmoothing(y_hW,seasonal_periods=12, trend='add', seasonal='add').fit()
h_W_pred = h_W.forecast(12)
```

Le lissage par la méthode Holte-Winters :

$$\begin{cases} \hat{a}_T = (1-\alpha)(XT - \hat{S}_{T-s}) + \alpha(\hat{a}_{T-1} + \hat{b}_{T-1} \\ \hat{b}_T = (1-\beta)(\hat{a}_T - \hat{a}_{T-1}) + \beta\hat{b}_{T-1} \\ \hat{S}_T = (1-\gamma)(XT - \hat{a}_T) + \gamma\hat{S}_{T-s} \end{cases}$$

La prévision par la méthode Holte-Winters :

$$\begin{cases} \hat{X}_{T(\ell)} = \hat{a}_{T+\ell}\hat{b}_T + \hat{S}_{T+\ell-s} \quad si\ \ell \in 1,\ldots,s \\ \hat{X}_{T(\ell)} = \hat{a}_{T+\ell}\hat{b}_T + \hat{S}_{T+\ell-2s} \quad si\ \ell \in s+1,\ldots,2s \\ \cdots \end{cases}$$

- Pour la méthode SARIMA :

- **Stationnarisation :**

Eu égard à la tendance et à la saisonnalité de notre série (non-stationnaire), on a travaillé non pas sur la série temporelle mais sur des différences de la série avec l'opérateur de différenciation ∇_s^D au lieu de X_t : $\nabla_s^D = (I - B_s)^D$

On a appliqué un filtre de différence première pour éliminer la tendance et un filtre d'ordre 12 pour la saisonnalité afin de tendre vers une stationnarité du processus.

```
# 1ère Differenciation en tendance
y_dif1 = y['Consommation_corrigee'] - y['Consommation_corrigee'].shift(1)

plot_sortie_acf(acf(np.asarray(y_dif1[1:])), y_len)
```

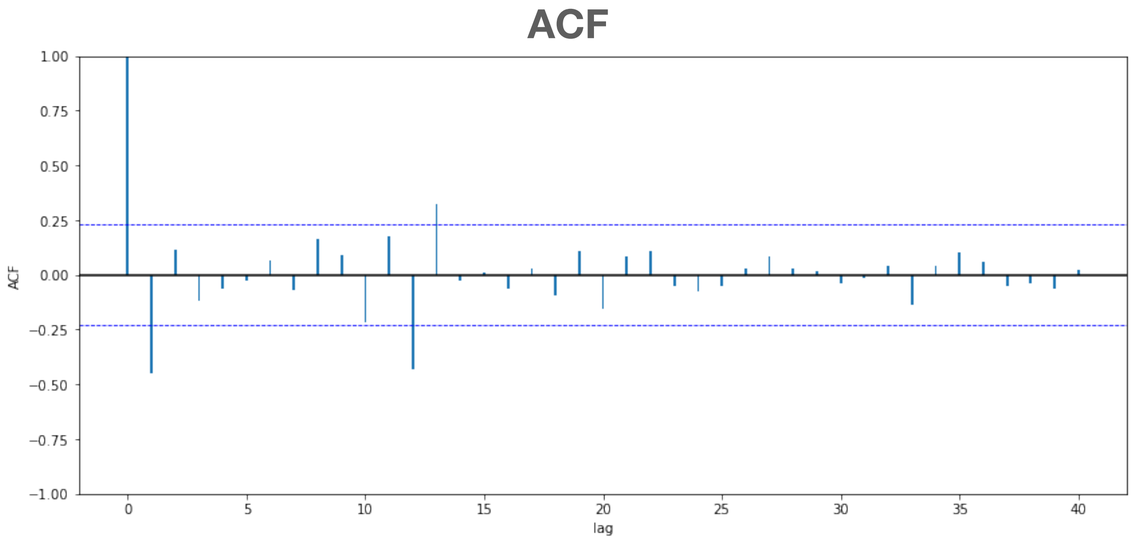
```
# 2 ème Différenciation  en saisonnalité
y_dif_1_12 = y_dif1 - y_dif1.shift(12)

plot_sortie_acf(acf(np.asarray(y_dif_1_12[13:])), y_len)
```

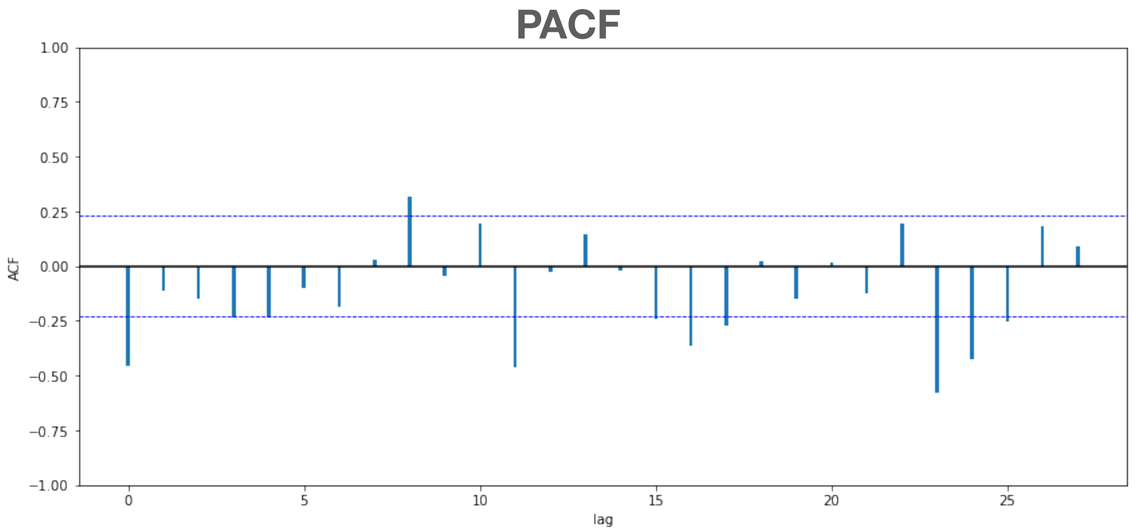
- **Identification des modèles :**

Malgré une série doublement différenciée, il restait une structure dans la série temporelle :

- **la fonction d'auto-corrélation (ACF) de la différence saisonnière d'ordre 12 présentait des pics aux retards 1, 12 et une décroissance sinusoïdale amortie.**

- **l'auto-corrélogramme partiel (PACF) de la différence saisonnière d'ordre 12 présentait des pics aux retards 1, 12 et 24 .**





- **Estimation et validation des modèles :**

Au regard des auto-corrélogrammes empiriques simples et partiels :

- **On a estimé un modèle** $SARIMA(1,1,1)(1,1,1)_{12}$

Soit un modèle :

$$(I-\varphi_1 B)(I-\varphi'_1 B^{12})(I-B)(I-B^{12})ln(X_t) = (I+\theta_1 B)(I+\theta'_1 B^{12})\varepsilon_t$$

```
# importation
from statsmodels.tsa.statespace.sarimax import *
from statsmodels.stats.diagnostic import acorr_ljungbox

model1 = SARIMAX(np.asarray(y['Consommation_corrigee']), order=(1,1,1), seasonal_order=(1,1,1,12))
results1 = model1.fit()                          # composante non saisonière  # composante saisonière
print(results1.summary())

print('Retard : p-value')
for elt in [6, 12, 18, 24, 30, 36]:
    print('{} : {}'.format(elt, acorr_ljungbox(results1.resid, lags=elt)[1].mean()))
```

- **Ensuite, on a estimé un deuxième modèle** $SARIMA(1,1,0)(0,1,0)_{12}$ **:**

Soit un modèle auto-régressif (AR) d'ordre 1:

$$(I-\varphi'_1 B)ln(X_t) = (I+\theta_1 B)(I+\theta'_1 B^{12})\varepsilon_t$$

```
model2 = SARIMAX(np.asarray(y['Consommation_corrigee']), order=(1,1,0), seasonal_order=(0,1,0,12))
results2 = model2.fit()                          # composante non saisonière  # composante saisonièr
print(results2.summary())

print('Retard : p-value')
for elt in [6, 12, 18, 24, 30, 36]:
    print('{} : {}'.format(elt, acorr_ljungbox(results2.resid, lags=elt)[1].mean()))
```

Le modèle 2 présente des tests de significativité , des paramètres de blancheur et de normalité du résidu satisfaisants, contrairement au 1er modèle.

4. Présentation et interprétation des graphiques

4.1. Correction des données de consommation mensuelles de l'effet température avec la régression linéaire:

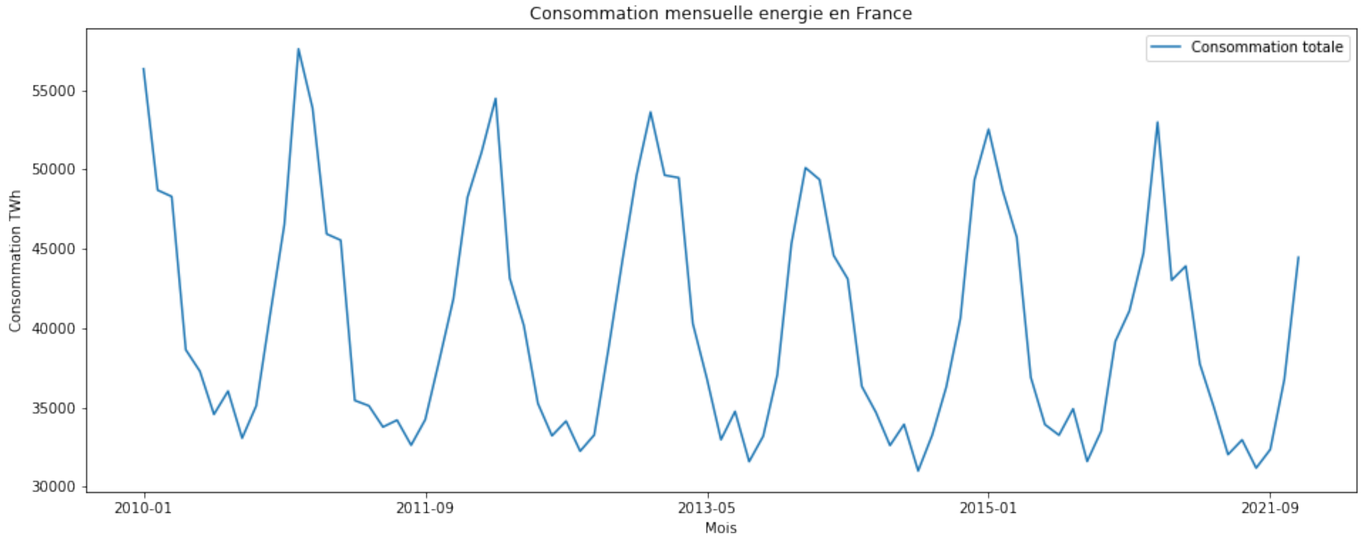
Le graphique est une série temporelle qui nous indique la consommation mensuelle d'énergie en France en Twh.

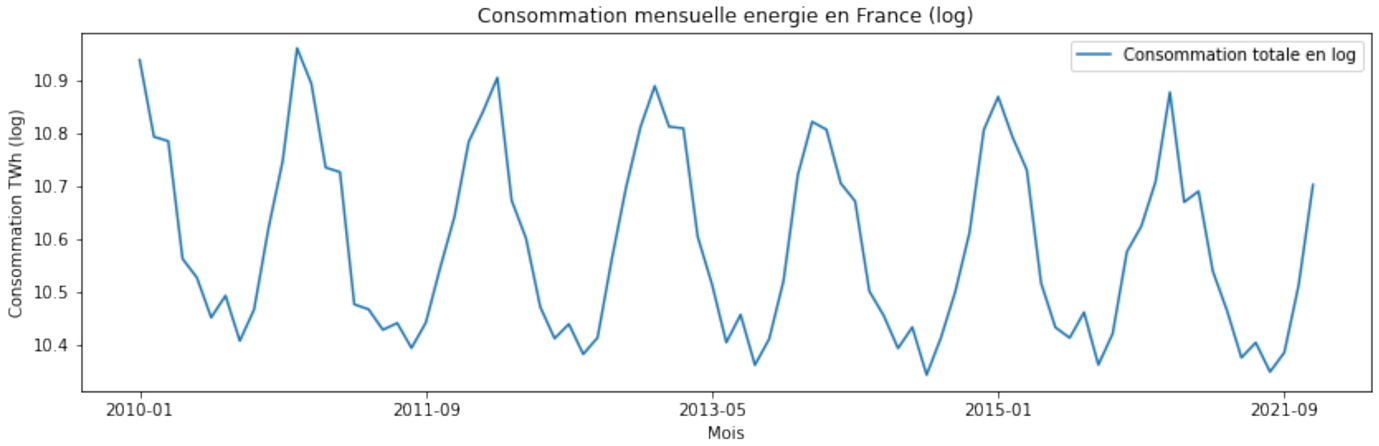
Notre série à une tendance stable, avec des pics annuels qui représentent une saisonnalité.

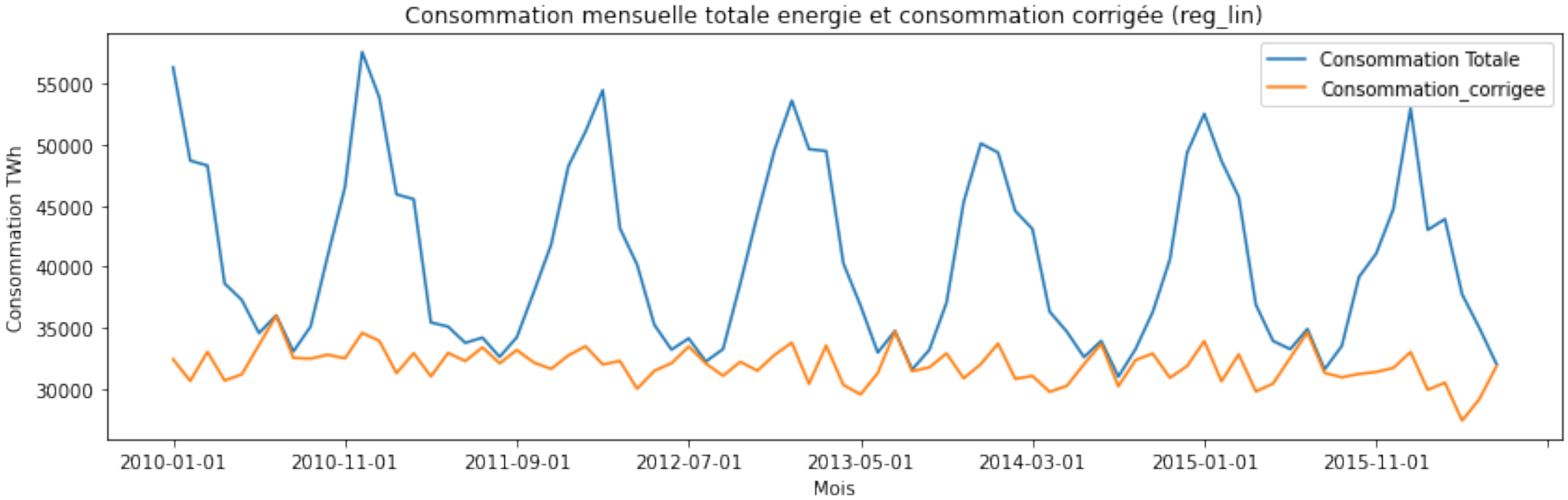
Elle semble stationnaire en moyenne et stationnaire en variance. Lorsqu'on opère une transformation logarithmique, les résultats sont identiques.

Aussi, eu égard, à la tendance stable, à la saisonnalité de notre série temporelle ainsi qu'aux écarts entre les pics et les creux stables sur une période au cours du temps, on peut affirmer que nous sommes face à un modèle additif :

$$X_t = T_t + S_t + \epsilon_t.$$



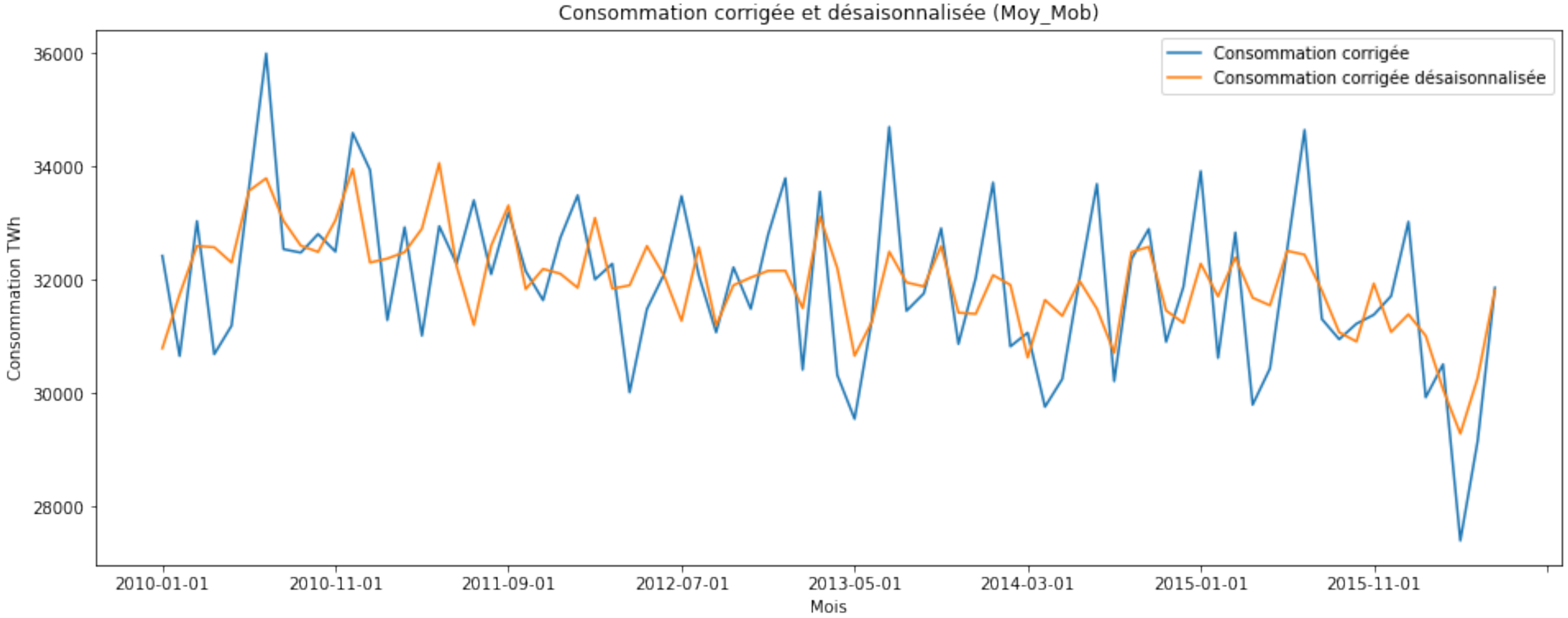




Le graphique nous indique les niveaux de la consommation mensuelle d'énergie et la consommation corrigée de l'effet de température (chauffage)en TWh entre janvier 2010 et décembre 2015.

On obtient une tendance de la consommation corrigé qui traduit un niveau moyen stable et une saisonnalité qui est centrée.

4.2. Désaisonnalisation de la consommation après correction, grâce aux moyennes mobiles :

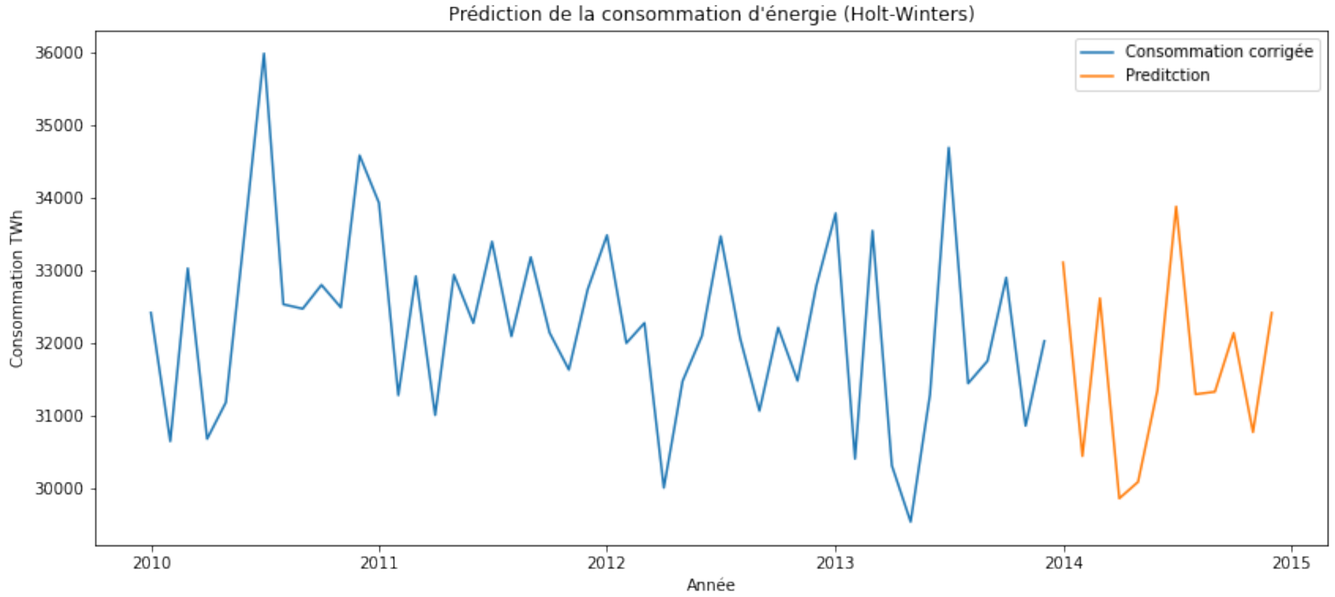


Le graphique nous indique les niveaux de consommation corrigée et la consommation corrigée désaisonnalisée en TWh entre janvier 2010 et décembre 2015.

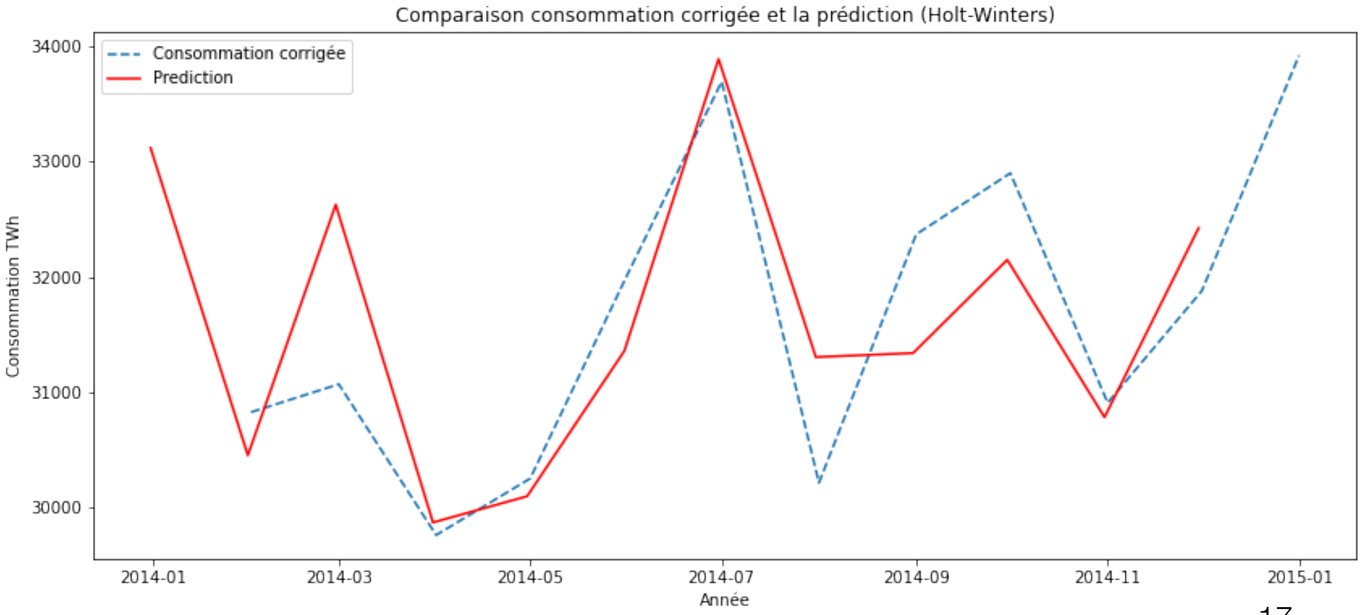
Notre série temporelle est invariante et absorbée à savoir rendue nulle par les moyennes mobiles. En somme, notre série temporelle conserve sa tendance.

Quant au pouvoir de réduction de variance entre entrée et sortie, à savoir atténuer la perturbation (bruit blanc faible), on note une atténuations de la variance.

4.3. Prévision de la consommation corrigée de l'effet température sur un an avec la méthode de Holt Winters et SARIMA :

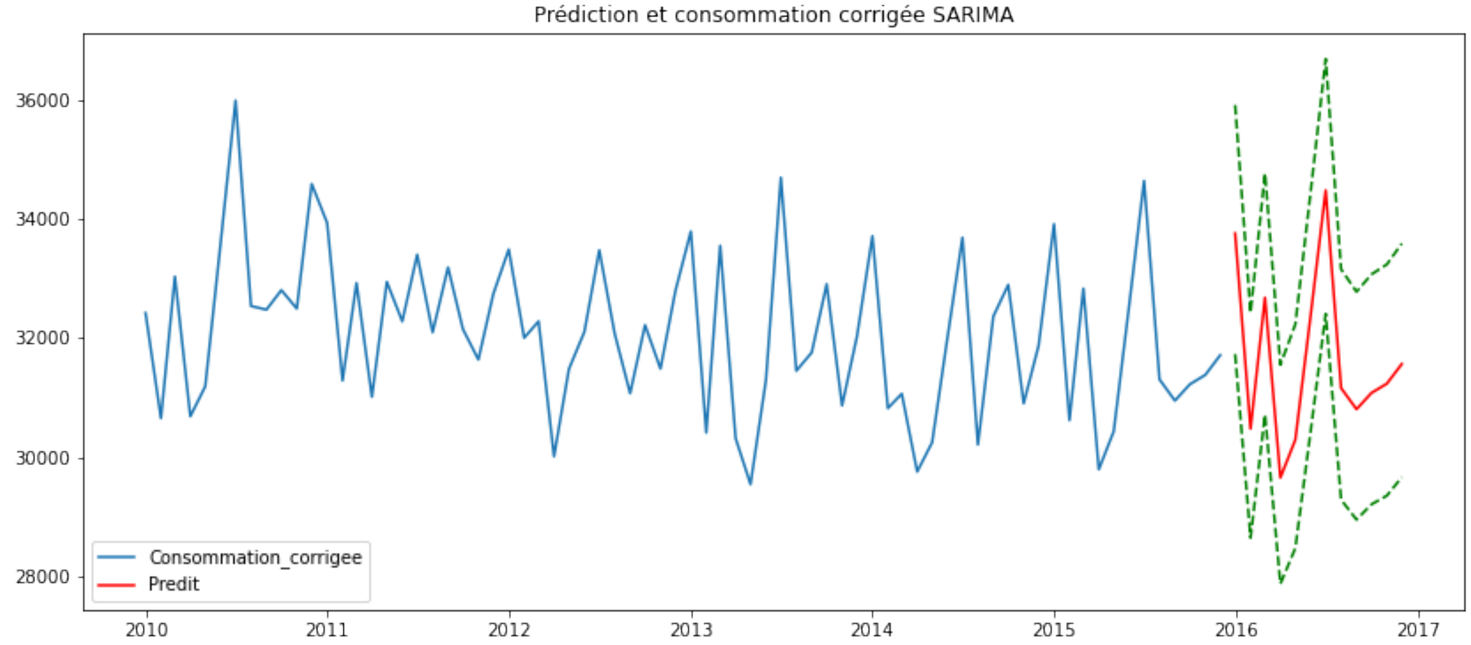


On note sur ce graphique une prédiction qui s'inscrit dans la continuité d'une tendance stable de la consommation corrigée.

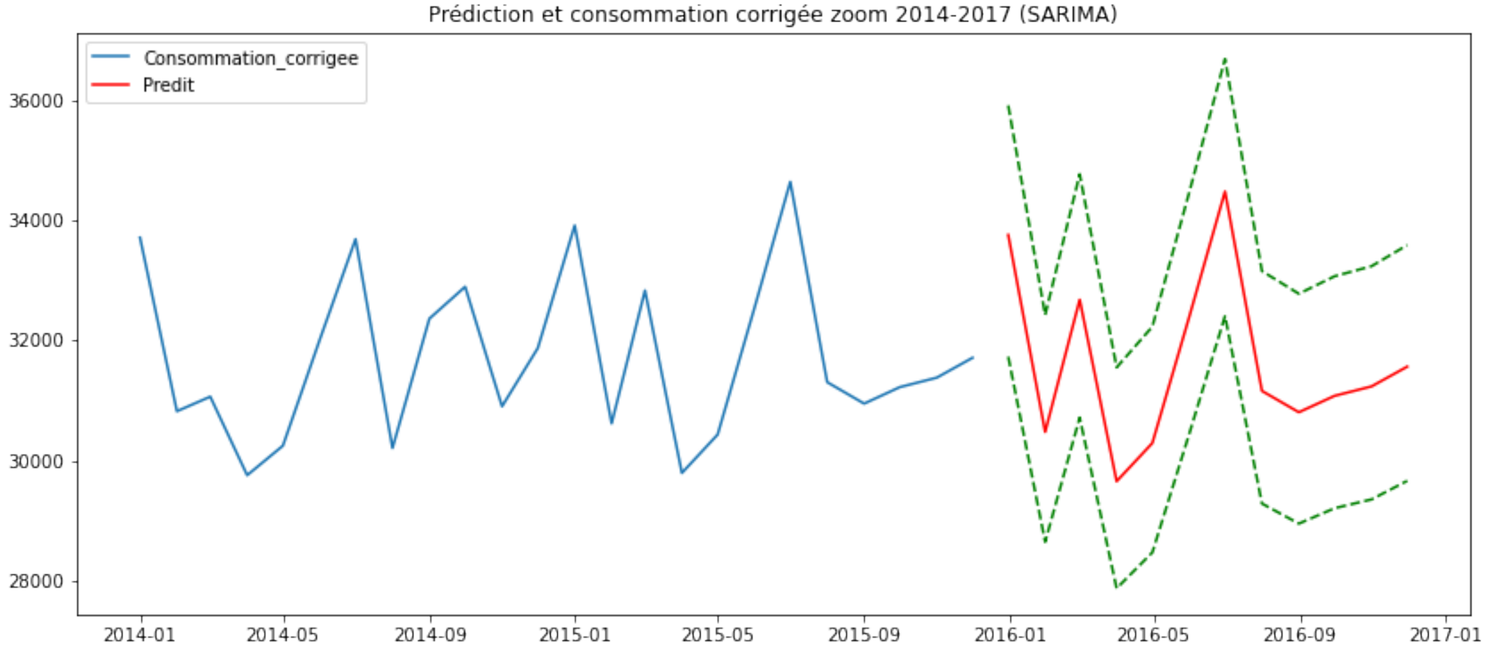


On note sur ce graphique annuel de comparaison une tendance stable pour la prédiction et une tendance plutôt croissante pour la consommation corrigée pour l'année 2014.

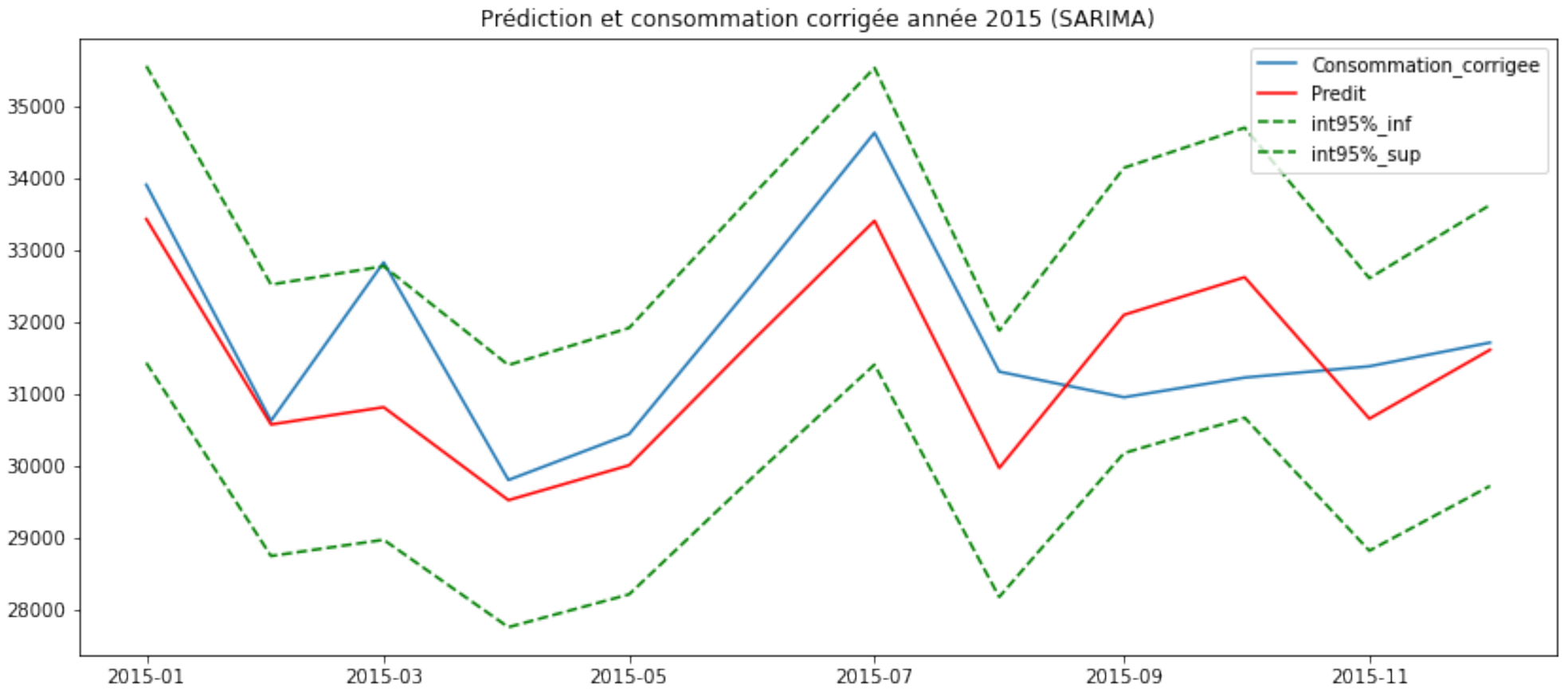
Il semble cohérent car la saisonnalité est mieux prise en compte par la méthode Holt-Winters contrairement à la consommation réelle corrigée qui approxime la saisonnalité (Primo-estimation) et n'est pas tout à fait désaisonnalisée.



Nos graphiques nous informent sur la prédiction de notre modèle sur une année par rapport au passée de notre série temporelle.



L'allure de nos graphiques semble cohérente au vu du passé de notre série temporelle. On obtient une prédiction qui semble convenable.



On note sur ce graphique une prédiction annuelle qui est très proche de la réalisation en bleu et elle s'inscrit dans l'intervalle de confiance. On obtient une prédiction convenable.

5. Conclusion

Avec la méthode SARIMA on a identifié, puis estimé et validé notre modèle.

Ensuite, on a utilisé notre modèle pour effectuer des prévisions à priori acceptables eu égard à l'analyse effectuée à posteriori.

Néanmoins, la prudence est de mise car notre modèle s'appuie sur un nombre faible d'observations.

6. LES RÉFÉRENCES

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_excel.html

https://stackoverflow.com/questions/16504975/error-unsupported-format-or-corrupt-file-expected-bof-record

https://stackoverflow.com/questions/47405628/bokeh-utf8-codec-cant-decode-byte-0xe9-unexpected-end-of-data

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.asfreq.html

https://stackoverflow.com/questions/64617482/valueerror-you-must-specify-a-period-or-x-must-be-a-pandas-object-with-a-dateti

https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles/5001626-decouvrez-des-algorithmes-de-traitement-des-moyennes-mobiles

https://perso.math.univ-toulouse.fr/jydauxoi/files/2017/04/poly_eleves.pdf

https://eric.univ-lyon2.fr/~jjacques/Download/Cours/ST-Cours.pdf

https://searchcode.com/codesearch/view/86129185/

https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/