Prédiction des séries temporelles

Hocine KADI et Aleksandrina *STOYANOVA-CHRISTEN*

Contexte

En tant que data scientist d'une entreprise anglaise travaillant dans l'énergie, on vous a confié une base de données contenant l'évolution de la consommation énergétique en Angleterre de différentes natures (issue de l'éolien, du charbon, …) et ce en fonction du temps.

Votre rôle est de prédire, sur l'année qui suit, la production énergétique des différentes ressources et de consigner dans un rapport les méthodes utilisées pour parvenir au meilleur résultat ainsi que l'interprétation de vos résultats.

Fort de vos compétences en Time Series vous avez à disposition un certain nombre de modèles d'intelligence artificielle qui pourraient vous être utiles :

- Modèle additif
- Modèle multiplicatif
- Modèle autorégressif
- LSTM
- ...

A l'aide d'une étude des données : stationnarité, corrélogramme etc .. et de ces derniers vous pourrez produire une analyse détaillée de vos prédictions.

Exploration des données :

Description :

Dans ce travail, nous avons eu à travailler avec un jeu de données qui décrit la production d'énergie au Royaume-Uni entre le 01-01-2012 et 20-08-2019 sous les différentes formes d'énergie à savoir le charbon, le nucléaire, l'hydroélectricité, l'éolien et le solaire.

Le jeu de données est donc composé de 6 colonnes et 796453 lignes. Les enregistrements ont été réalisés toutes les 5 minutes.

En raison des politiques publiques et de la sensibilisation des états au réchauffement climatique et la décarbonation des sociétés, l'évolution de la production sous les différentes formes ne suit pas la même évolution et subit des tendances différentes en fonction de la forme d'énergie produite. Pour le charbon, sont enregistrées la production des dernières centrales au charbon britanniques restantes. Sur la partie éolienne, les parcs éoliens non mesurés ne sont pas enregistrés dans le jeu de données et ne s'affichent que sous la forme

d'une petite baisse de la demande plus ou moins proportionnelle aux rendements des parcs éoliens mesurés. Pour la production solaire, il s'agit de la production estimée fournie par l'Université de Sheffield. Ces estimations sont peut-être 10 % plus grandes que la réalité.
Le tableau 1 résume les principales informations sur les différentes colonnes.

Nettoyage :

Le jeu de données est relativement propre toutefois nous avons dû :

- Éliminer 52 doublons où pour la même date et la même heure nous avons plusieurs lignes.
- Éliminer les valeurs zéro de l'énergie nucléaire car il est physiquement impossible que toutes les centrales nucléaires de tous le Royaume-Uni s'arrêtent pour 5 minutes et repartent juste après.
- Deux valeurs aberrantes pour l'énergie solaire à la date du 29-06-2017 et que nous avons interpolées.
- Sur les granularités ; jour, mois et années, en raison du manque de quelques données, nous avons effectué un rééchantillonnage et moyenner sur les périodes choisies.

Tableau 1 : Description des colonnes du jeu de données

Colonne	Type de données	Description
Temps	Datetime	Par intervalle de 5 minutes
Charbon	Entier	La production des dernières centrales au charbon britanniques restantes.
Nucléaire	Entier	La production totale de toutes les centrales nucléaires britanniques.
Hydroélectrique	Entier	La production totale de toutes les centrales hydroélectriques.
Vent	Entier	Production de tous les parcs éoliens mesurés. Les parcs éoliens non mesurés (intégrés) ne sont pas enregistrés ici et ne s'afficheront que sous la forme d'une petite baisse de la demande plus ou moins proportionnelle aux rendements des parcs éoliens mesurés.
Solaire	Décimal	Il s'agit de la production estimée de l'énergie solaire, fournie par l'Université de Sheffield. Il faut souligner qu'il s'agit d'une meilleure estimation, pas d'une sortie réelle mesurée, et il y a des raisons de supposer qu'elle est peut-être 10 % plus grande que la réalité.

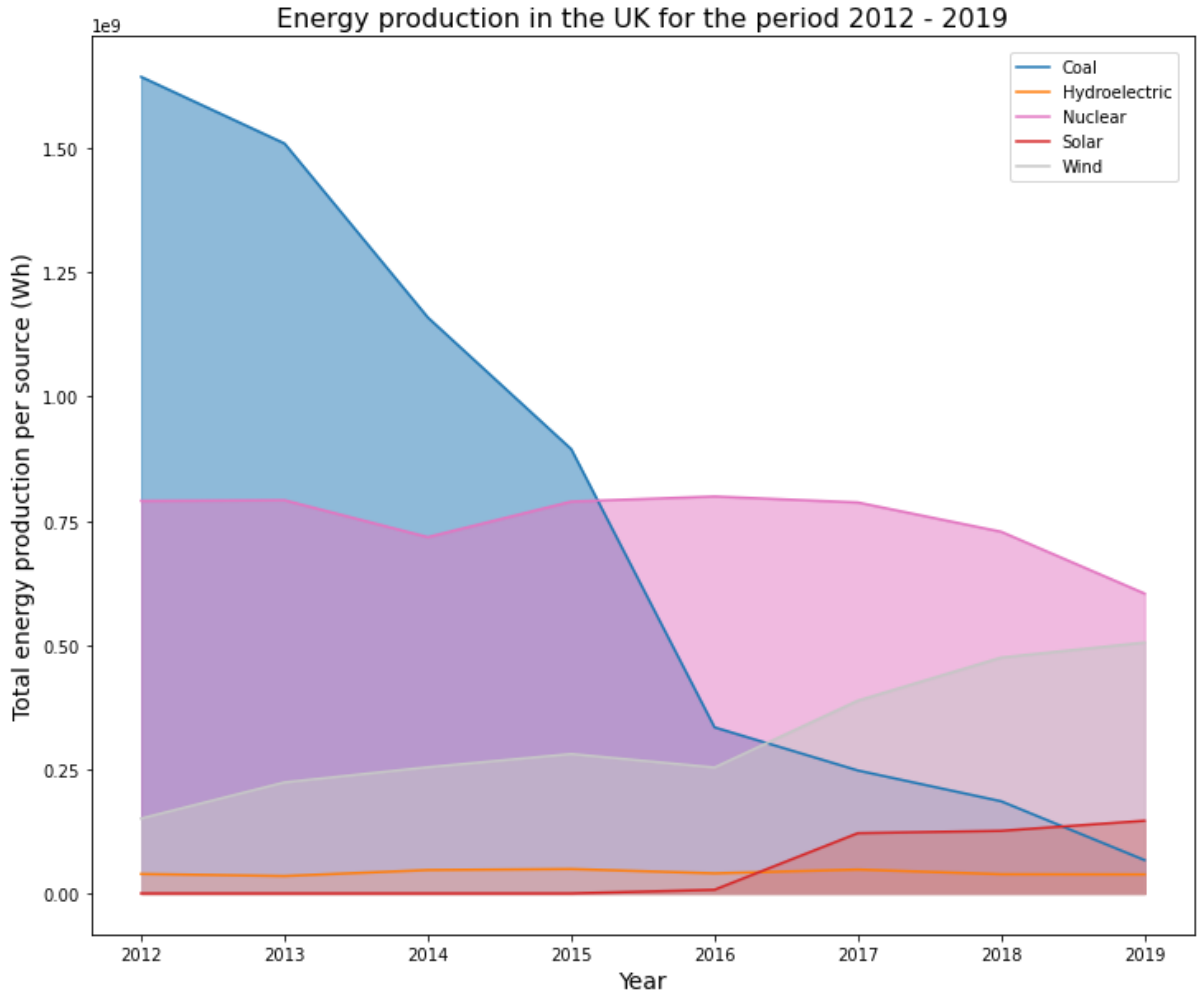
Analyse exploratoire :

La production électrique, contrairement à la consommation qui revêt un caractère moins contrôlable, reste un élément de souveraineté des états et très impacté par la

décision politique et du contexte géopolitique. Toutefois, les énergies intermittentes (non-pilotables) à l'instar de l'éolien et du solaire dépendent fortement des conditions météorologiques et climatiques ce qui rend la prédiction sur ces données plus complexe et nécessite notamment de coupler ces données avec d'autres jeux de données pour compléter une partie de l'information cachée. D'autres raisons peuvent influencer la production électrique solaire et éolienne, principalement politiques dans le contexte de la transition écologique actuelle.

Pour visualiser les tendances de production des différentes sources d'énergie nous avons rapporté sur la figure 1, l'évolution de la production cumulée sur la période 2012-2019 en fonction du type d'énergie.

Figure 1 : Évolution de la production de l'énergie au Royaume-Uni en fonction du type de l'énergie sur la période 2012-2019. Les données n'étant disponibles que jusqu'au



*20-08-2019, nous avons moyenné les différentes productions sur 2019 que nous avons ensuite multiplié par le nombre 12*24*365 (12 fois 5 minutes pour faire 1 heure, puis 24 heures sur les 365 jours).*

On peut noter de cette figure des tendances fortes ont impactées la production de l'énergie au Royaume-Uni sur les 8 années. La production du charbon est en continuelle diminution où elle enregistre près de 90% de baisse. Cette baisse spectaculaire s'explique

par la volonté des états de décarboner la production dans le but de réduction des gaz à effets de serre. Les productions nucléaire et hydroélectrique sont stables sur la période d'étude. Le caractère pilotable de ces deux sources d'énergie permet de les utiliser comme moyen d'ajustement pour les politiques énergétiques. Par ailleurs, depuis 2016 on note une augmentation de la production de l'énergie d'origine renouvelable, une augmentation plus prononcée pour l'éolien.

Afin de comprendre l'évolution de la production entre les différentes sources d'énergie, nous avons rapporté sur la figure 2, l'évolution de la distribution sur la période d'étude 2012-2019.

On note que la baisse de la production d'origine charbon est compensée par le nucléaire entre 2012 et 2016 puis par le trio nucléaire, vent et solaire. De la distribution, nous pouvons suivre les décisions en termes de changement de politique énergétique du Royaume-Uni.Deux éléments sont à noter de ces deux premiers graphes, la diminution de l'utilisation du charbon comme source d'énergie impacte fortement la production globale. d'autre part, comme la part de l'énergie nucléaire restant stable, c'est la production totale qui se trouve impactée.

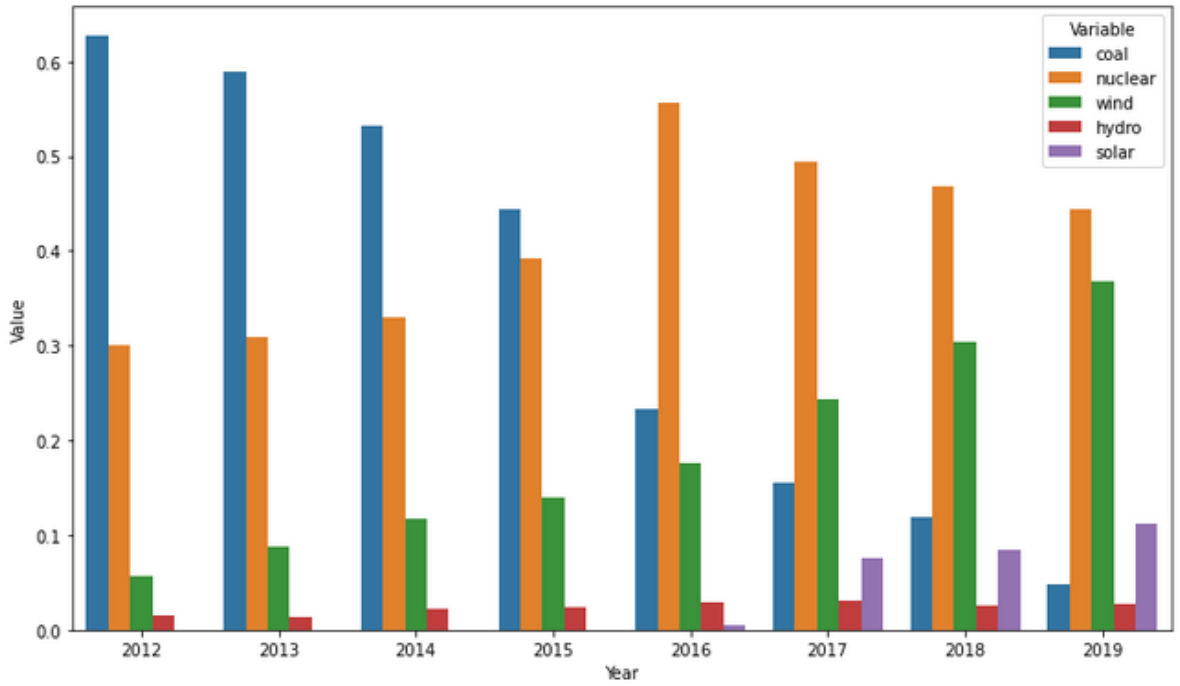


Figure 2 : Évolution de la distribution entre les différentes sources d'énergie entre 2012 et 2019.

Un autre élément intéressant à analyser est la production de l'énergie en fonction des mois de l'année. Nous rapportons sur la figure 3 les distributions cumulées sous forme de boîtes à moustaches sur la période 2012-2019 en fonction des mois de l'année.

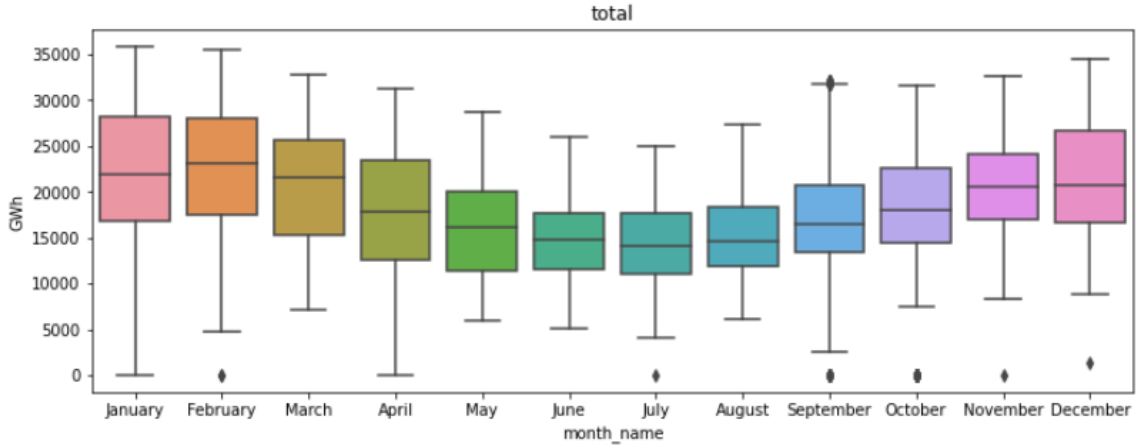


Figure 3 : Distributions cumulées de la production totale sur la période 2012-2019 en fonction des mois de l'année.

Nous notons un effet saisonnier important où la production est plus grande en hiver (novembre-février) par rapport à la période estivale (mai-août). La production étant tractée par la consommation qui en raison du chauffage augmente notablement sur la période hivernale.

De cette analyse exploratoire, nous prenons comme objectifs d'étude :

- ☐ De prédire la production totale en fonction des mois sur une année.
- ☐ D'essayer de prédire l'évolution de la distribution en fonction du temps.

Ces objectifs se justifient à notre sens par la forte dépendance de chaque source d'énergie à des facteurs extérieurs dont on ignore les impacts. Par contre, la production totale étant pilotable par l'état, elle est moins soumise à une forte variabilité. Nous complétons cette étude par la prédiction de la distribution.

Nous rapportons sur la figure 4, l'évolution de la production totale en fonction sur la période 2012-2019 en faisant la somme sur le mois. Nous notons à partir de cette figure des maximas qui correspondent à la période hivernale et des minimas liés à la saison estivale.

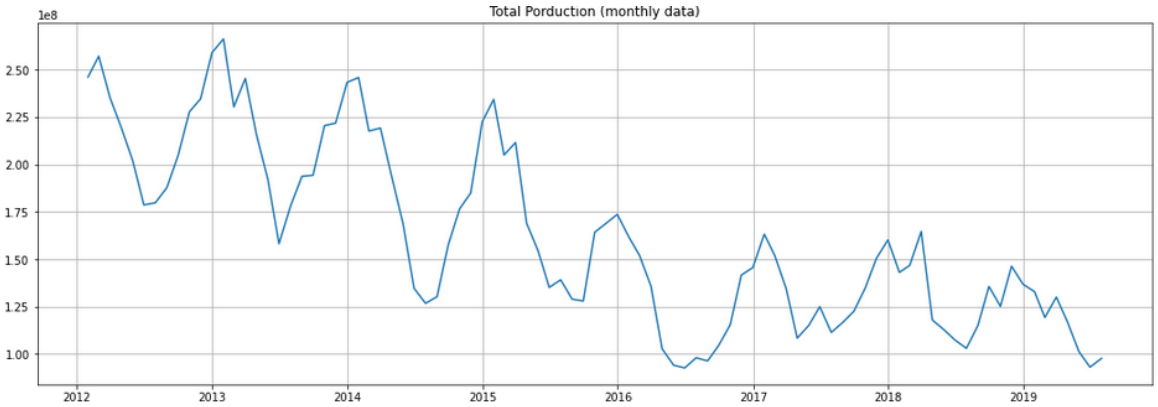


Figure 4 : Évolution mensuelle de la production totale de l'énergie au Royaume-Uni sur la période 2012-2019.

Prédiction et modèles :

Nous allons appliquer dans cette section différents modèles statistiques ARIMA et SARIMA ainsi que des modèles de deep learning pour la prédiction à l'année de la production mensuelle totale de l'énergie du Royaume-Uni.

Modèles statistiques :

Nous commençons par l'analyse de la série temporelle (Figure 4) en utilisant un modèle simple dit *Auto-Régressive Integrated Moving Average* [ref arima]. Le modèle suppose que la série temporelle ne présente pas une saisonnalité. Une simple observation de la série montre que ce n'est pas le cas : la série temporelle présente à la fois une tendance et une saisonnalité. Deux approches sont possibles :

- ☐ Retirer la composante saisonnière, i.e., désaisonnaliser, et appliquer le modèle ARIMA
- ☐ Utiliser le modèle SARIMA [ref sarima] qui est un modèle ARIMA prenant aussi en compte la composante saisonnière

Nous décrivons explicitement dans ce document seulement les résultats de la première approche. Les résultats du modèle SARIMA sont présentés et commentés dans le notebook UK_production_energy_SARIMA.ipynb.

ARIMA:

Tout d'abord en décomposant notre série temporelle représentant l'évolution de la production totale de l'énergie en composantes de niveau, bruit, saisonnalité et tendance (un modèle multiplicative est utilisé car la variation saisonnière n'est pas constante dans le temps), nous confirmons la présence d'une tendance décroissante et d'une saisonnalité de période égale à 12 mois (voir, Figure 5).

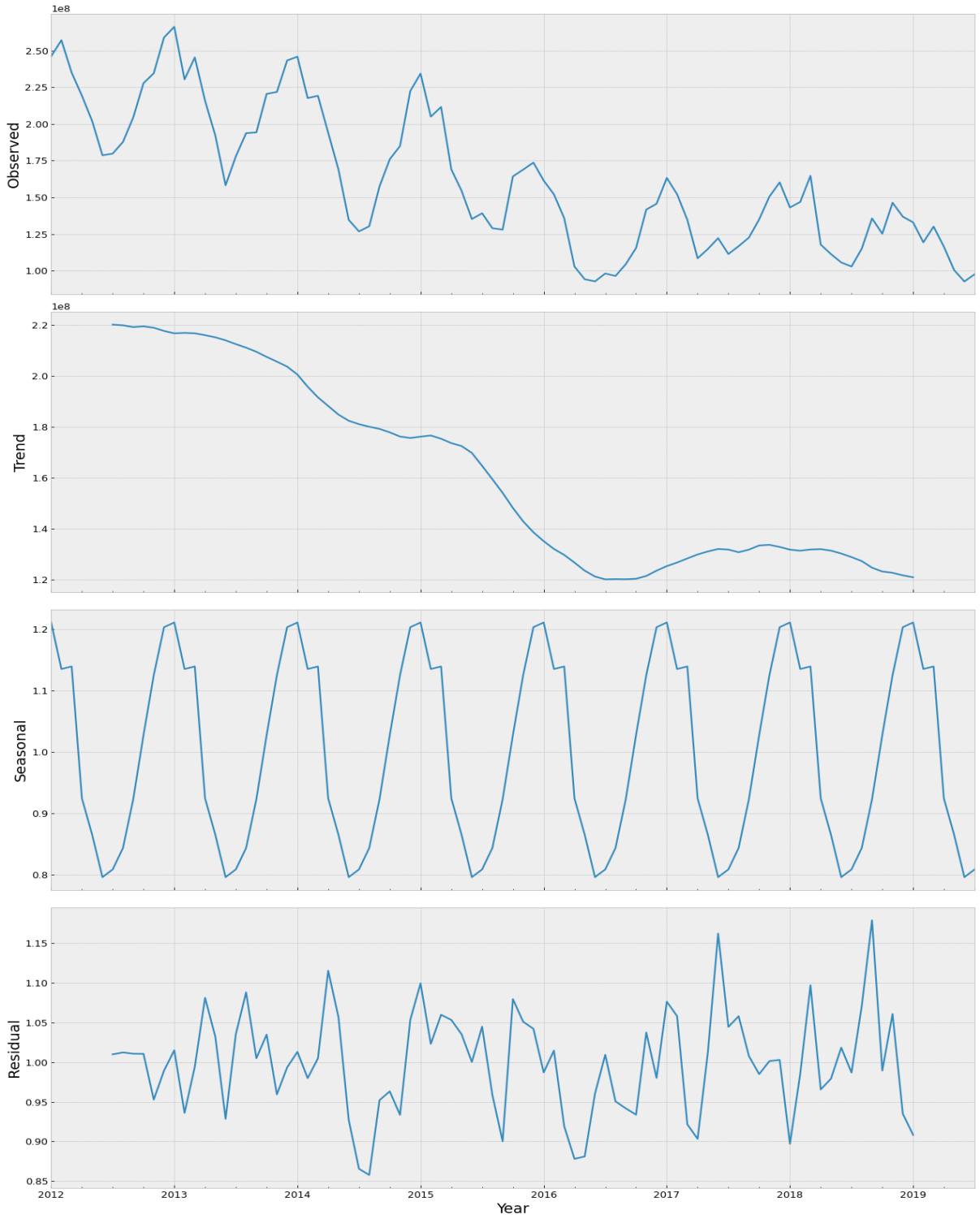
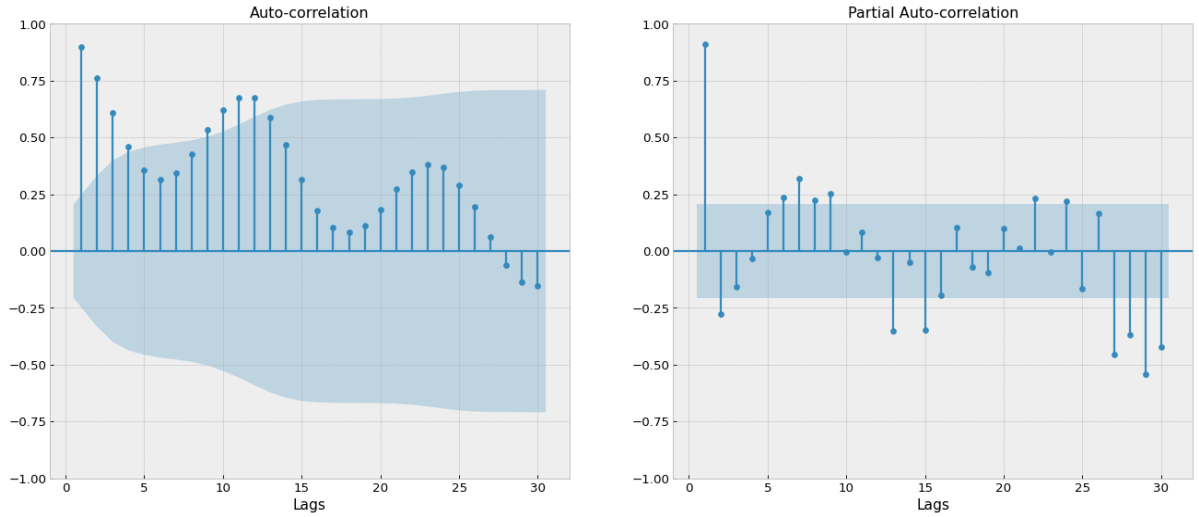


Figure 5 : Décomposition de la série temporelle de la Figure 4 en composantes.

Afin d'employer le modèle d'AR(I)MA pour la modélisation de la série nous avons besoin de retirer cette saisonnalité et, plus généralement, de rendre la série stationnaire. La série n'est pas stationnaire car la moyenne et la variance ne sont pas constantes dans le temps. Nous avons aussi effectué le test augmenté de Dickey-Fuller afin de confirmer la non-stationnarité (voir notebook



“UK_production_energy_ARIMA.ipynb”). De plus, le diagramme d’auto-corrélation nous permet de visualiser la dépendance des valeurs présentes de la série des valeurs passées. Nous avons des corrélations prononcées avec des valeurs proches (lag = 1, 2, 3, 4) ainsi que avec quelques valeurs plus distantes dans la passée (lag = 9, 10, 11, 12) confirmant de nouveau la saisonnalité de 12. Des autres tests statistiques sont montrés dans le notebook qui valident la non-stationnarité de la série.

Afin de retirer la saisonnalité de notre série et de la stationnariser nous avons procédé par différenciation :

- ☐ Pour retirer la saisonnalité, nous prenons la différence entre deux points séparés par un intervalle appelé lag. La valeur du lag est choisie être égale à 12 (la saisonnalité).
- ☐ Une différenciation était nécessaire afin de rendre la série désaisonnalisée stationnaire [voir Figure 6, troisième graphe du haut au bas] (critère utilisé - test augmenté de Dickey-Fuller), ce qui détermine la valeur de **d** = 1 dans le modèle ARIMA. Nous insistons sur le fait que le modèle ARIMA peut gérer les séries non-stationnaires [il peut les stationnariser avant d’appliquer le modèle ARMA]. Néanmoins nous choisisson de trouver l’ordre de différenciation **d** afin de simplifier la recherche de ce paramètre (ne pas inclure ce paramètre dans le Grid Search)

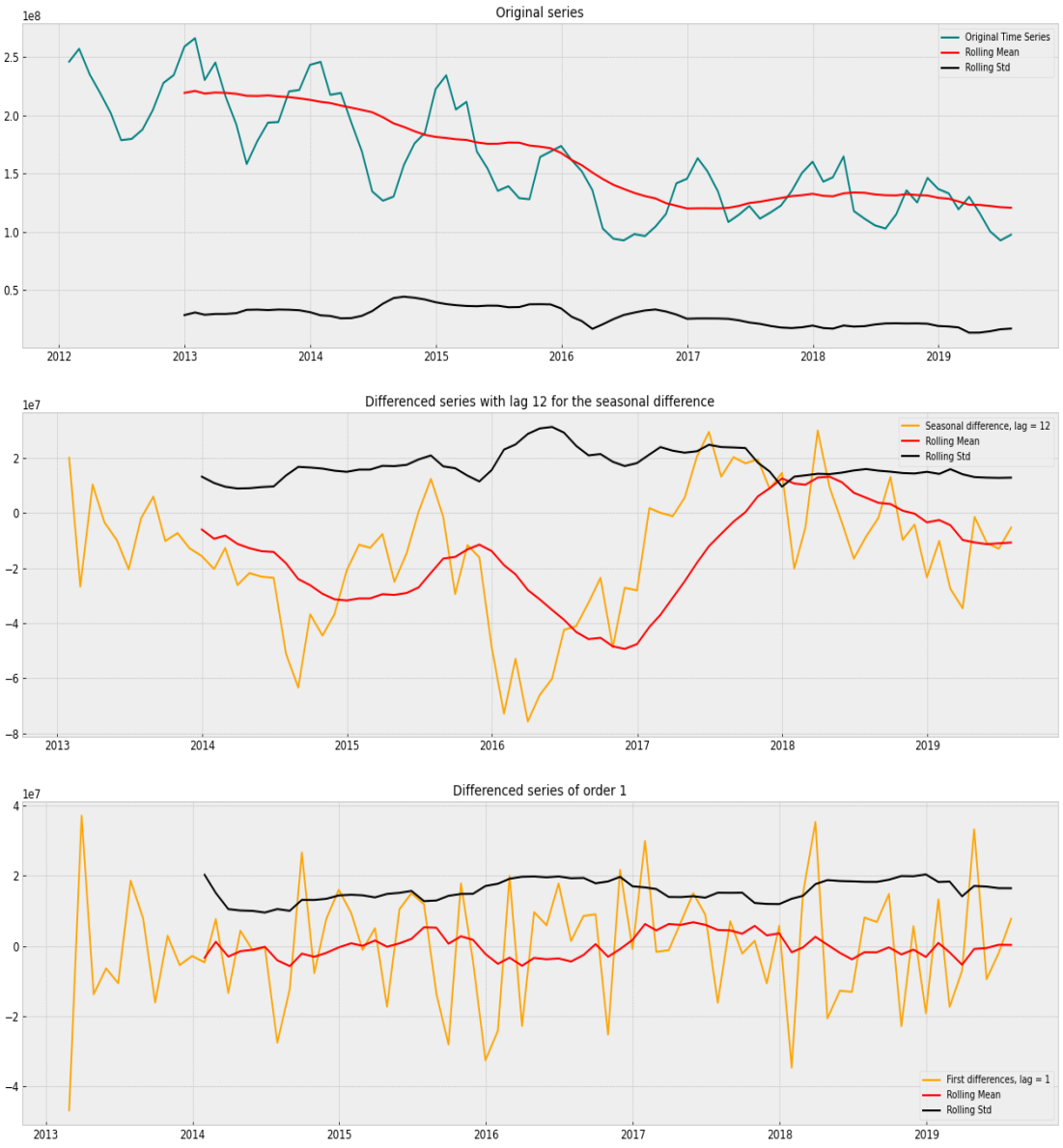


Figure 6 : *Résultats de la différenciation de la série temporelle avec un lag de 12 pour retirer la saisonnalité. L'ordre de la différenciation suivante pour rendre la série stationnaire est d = 1.*

Suite à la désaisonnalisation de la série, nous avons examiné les diagrammes d'auto-corrélation (ACF) et d'auto-corrélation partielle (PACF). Nous constatons que la saisonnalité n'est plus présente, néanmoins nous avons encore plusieurs lags de magnitude importante (voir Figure 7). De plus, le test augmenté de Dickey-Fuller indique que la série n'est pas stationnaire [voir Figure 6, deuxième graphe du haut au bas].

Finalement, afin d'enlever ces lags de magnitude importante et de stationnariser la série, nous avons pris les premières différences en soustrayant la série d'elle-même avec un lag = 1. En examinant de nouveau les diagrammes d'auto-corrélation et d'auto-corrélation partielle (Figure 8), nous remarquons que la valeur de **p** déduite du diagramme PACF est probablement 2 car c'est le dernier lag important après lequel les autres lags ne sont pas signifiants.

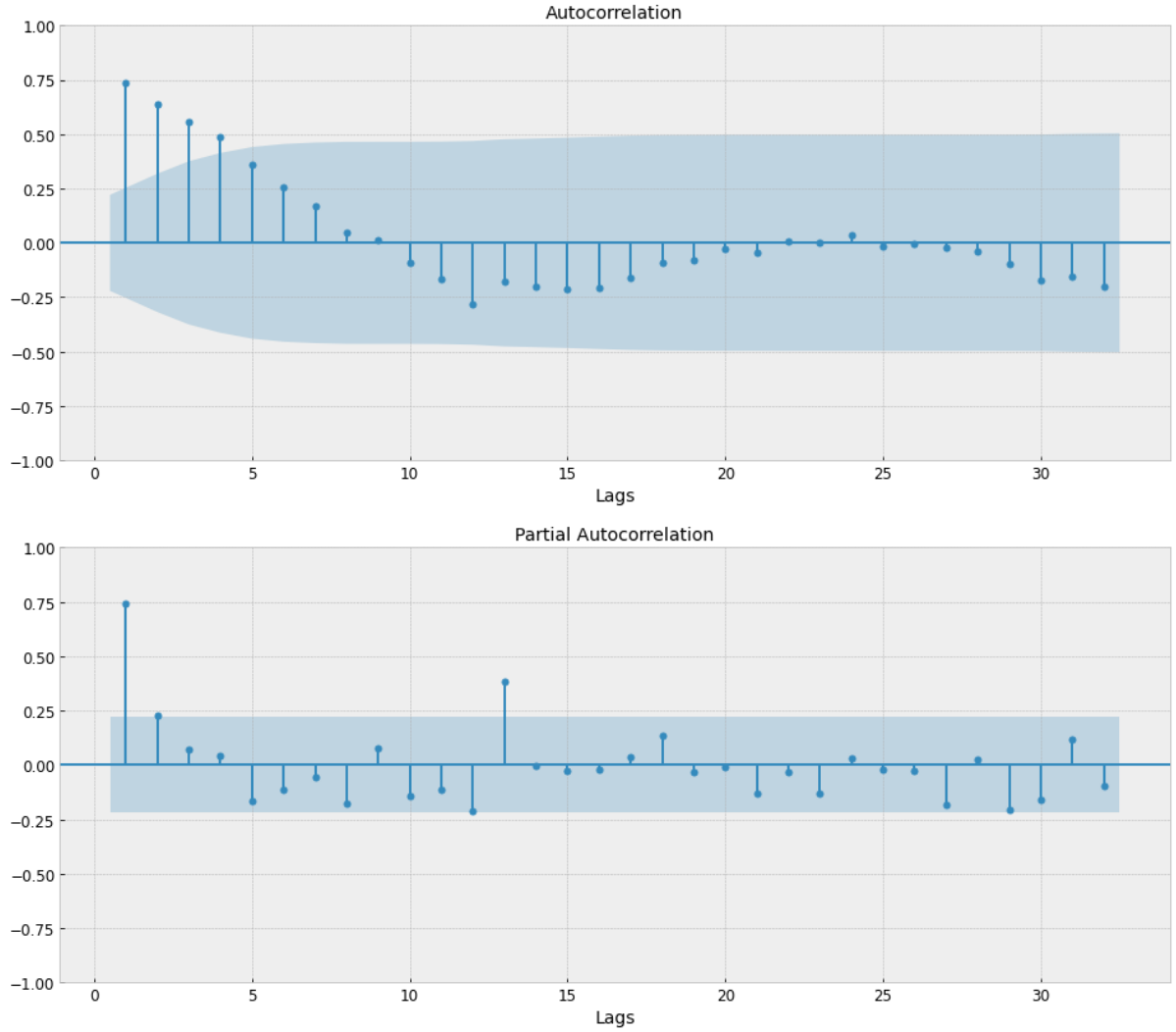


Figure 7 : Autocorrélation et auto-corrélation partielles obtenues après la désaisonnalisation de la série.

La valeur de **q** est aussi déduite égale à 2 à partir du graphe ACF.

Les résultats des deux graphes ACF et PACF n'étant pas concluants, nous avons effectué un Grid Search afin de trouver des valeurs optimales pour **p** et **q** (d = 1). Le critère utilisé dans ce Grid Search était le *critère* d'information d'Akaike (AIC). En variant les valeurs des paramètres dans certains intervalles nous cherchons la combinaison qui minimise l'AIC. Néanmoins, différents critères peuvent être utilisés comme, par exemple, l'erreur quadratique moyenne ou la racine carrée de l'erreur quadratique moyenne qui sont aussi minimisés. Ces critères peuvent amener aux différentes valeurs optimales pour **p** et **q** et donc à une ambiguïté dans leur détermination. Nos résultats de l'analyse basé sur le critère d'AIC pour **p** et **q** sont les valeurs suivantes **p** = 2 et **q** = 5 (voir notebook "UK_production_energy_ARIMA.ipynb").

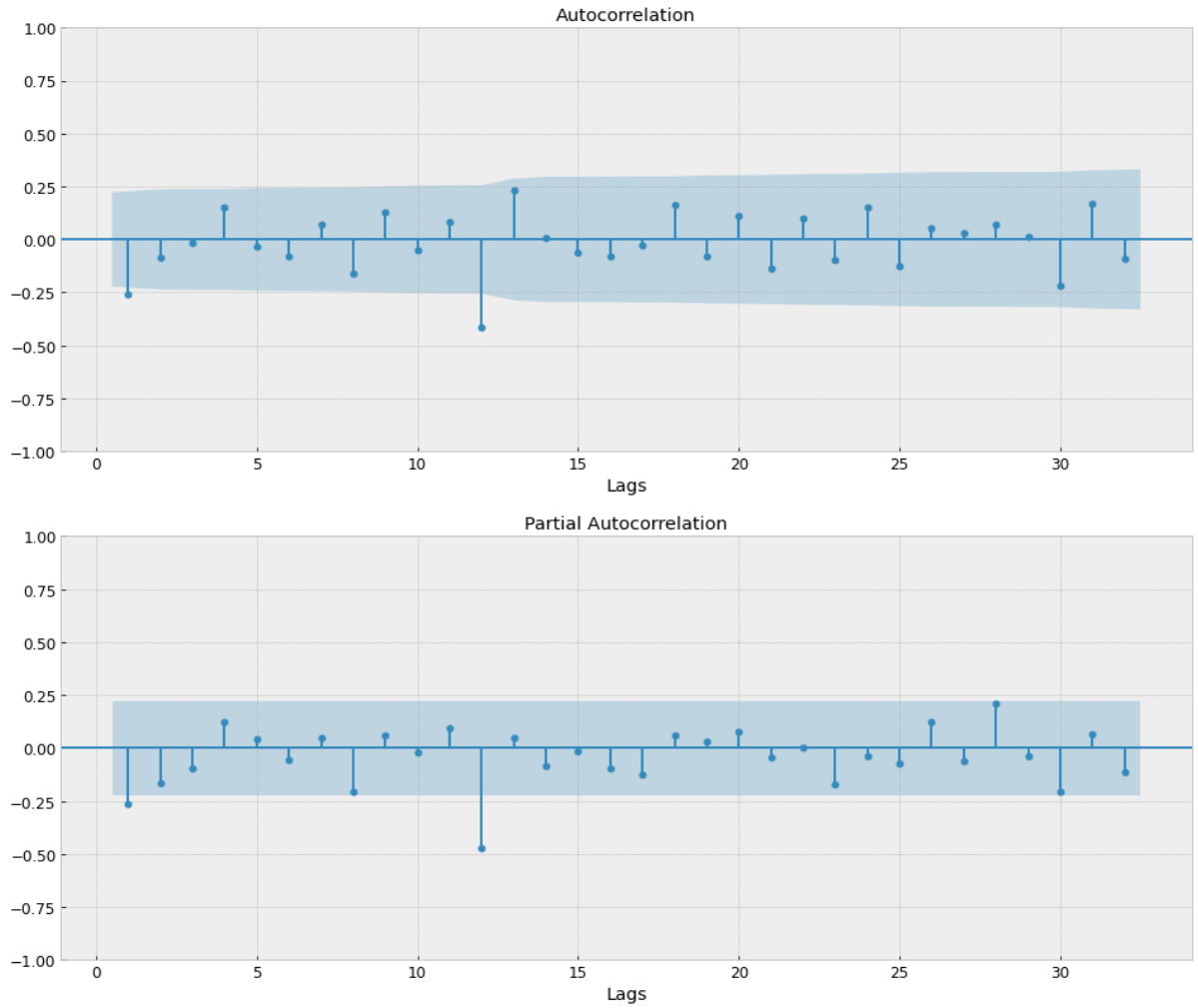


Figure 8 : Auto-corrélation et auto-corrélation partielles obtenues après la désaisonnalisation de la série puis la différenciation de la série une fois.

Parce que nous utilisons la série désaisonnalisée et différenciée (stationnaire) la construction des modèles auto-régressifs se résume aux modèles ARMA (d = 0 pour les série stationnaires, voir section “Modèle ARMA” dans le notebook). Discussion sur l’importance statistique des lags AR et MA du modèle est disponible dans le notebook (modèle summary). Nous avons aussi utilisé la série désaisonnalisée mais non-stationnaire et construit un modèle ARIMA avec **p** = 2, **d** = 1 et **q** = 5 (voir section “Modèle ARIMA dans le notebook “UK_production_energy_ARIMA.ipynb” et graphes dans l’annexe). Comme anticipé, le modèle montre le même taux de précision. Il n’est pas discuté davantage dans ce document.

Les données étaient séparées en données d'entraînement (données du Février 2013 au Décembre 2018) et données de test (l’année 2019 ou 7 mois). La performance du modèle entraîné est illustrée dans Figure **9 (a, b, c)**. Figure **9 (a) deuxième figure** montre une bonne prédiction de la courbe de production d’énergie mensuelle avec une erreur de l’ordre de 10 % (sur les données de train). En ce qui concerne les données de test, le modèle AR(I)MA (2, 0, 5) donne une prédiction de la même précision que celui sur les données de train. De plus, les valeurs observées actuellement (à l’exception d’une) se trouvent dans l’intervalle de confiance de 95 %.

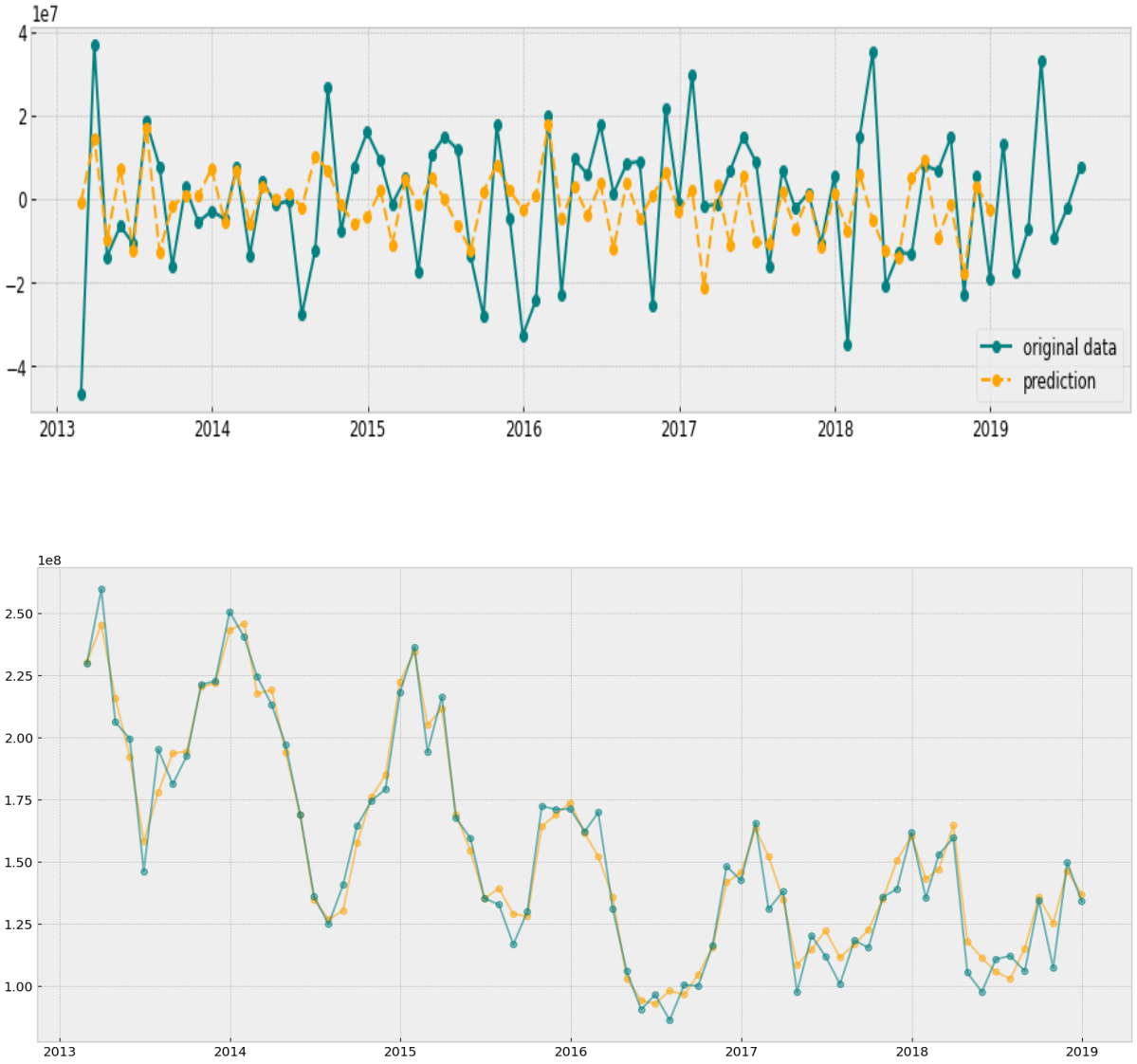


Fig. **9 (a) Première figure** : série désaisonnalisée et différenciée d'ordre 1, modèle *AR(I)MA (2, 0, 5)* : comparaison entre la vérité terrain (jeu d'entraînement) et la prédiction (MAE (train) = 11905669.453, MAPE (train) : 1.42226). **Deuxième figure** : Série originale (réel et prédiction) MAE (train) : 6209636.127, MAPE (train) : 0.04336.

Nous pouvons essayer d'améliorer davantage le modèle en expérimentant avec les valeurs des paramètres **p** et **q** afin de voir si nous pouvons obtenir un modèle avec tous les termes AR et MA statistiquement importants ($p < 0.05$ dans le modèle summary) tout en gardant une petite valeur de l'AIC. Néanmoins, notre Grid Search a déjà trouvé des valeurs optimales de **p** et **q** basés sur le critère de l'AIC.

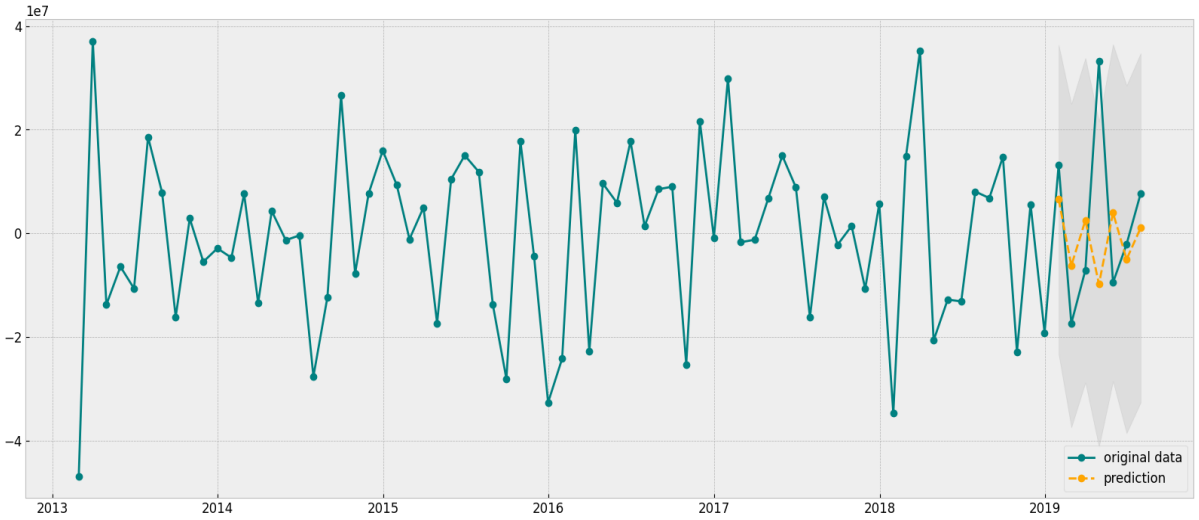


Fig. **9 (b)** Série désaisonnalisée et différenciée d'ordre 1, *modèle AR(I)MA (2, 0, 5)* : Comparaison entre la vérité terrain (jeu de test) et la prédiction. MAE (test) = 13355308.890, MAPE (test) = 1.069126.

Finalement, nous avons fait un forecast pour la période du 2019/08 au 2020/07 (Figure **9 (c)**).

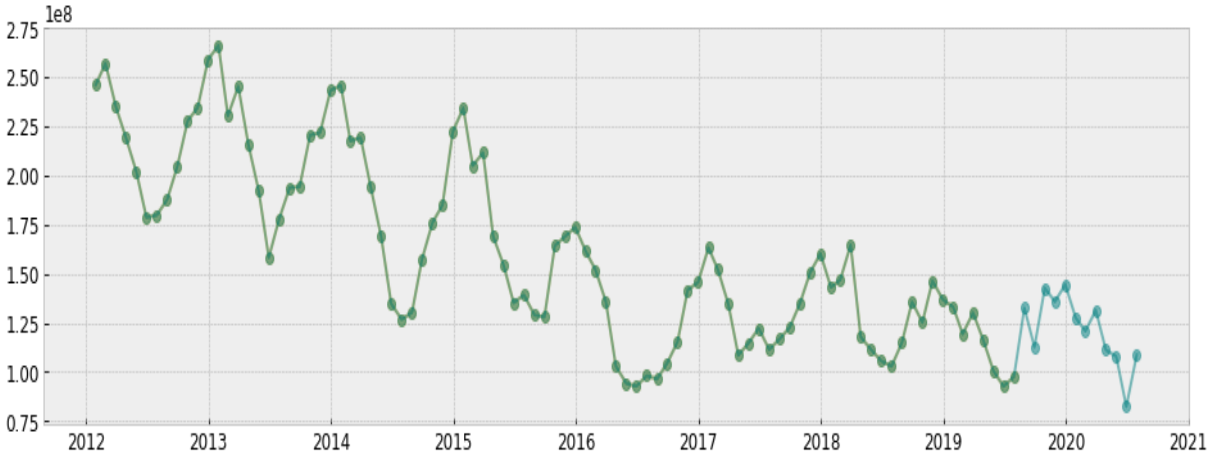


Fig. **9 (c)** Série originale (couleur verte) : Forecast du 2019/08 au 2020/07 (couleur turquoise).

La prédiction sur la production d'énergie au Royaume-Uni pour la période de l'Août 2019 au Juillet 2020 (Figure **9 (c)**) montre une stabilité de cette dernière. La courbe prédite préserve la tendance décroissante globale dans la production d'énergie totale mensuelle et démontre une structure similaire aux années précédentes. Néanmoins, la quantité des données mensuelles n'est pas suffisante pour un forecast plus fiable. Nous pouvons envisager soit une granularité plus fine, par exemple, hebdomadaire, soit l'utilisation d'un jeu de données portant sur plus des années. Il faut aussi remarquer le fait que la courbe de production d'énergie totale a une structure très différente au début de la période considérée (jusqu' au milieu de l'année 2016). Nous pouvons parler de deux séries temporelles distinctes.

Deep learning :

Bayesian LSTM et incertitudes

Une estimation fiable de l'incertitude pour la prévision des séries temporelles est essentielle notamment dans les domaines de gestion qui touche à un nombre important de personnes ce qui le cas de la production d'énergie. Les modèles de séries temporelles classiques sont souvent utilisés en conjonction avec une formulation probabiliste pour l'estimation de l'incertitude. Dans ce travail nous utilisons des modèles de réseaux de neurones à base de LSTM bayésien de bout en bout qui fournissent une prédiction de séries temporelles avec une estimation de l'incertitude[1]. À l'aide de la technique de dropout Monté-Carlos et de la distribution des erreurs de spécification du modèle, le modèle permet de prédire la production d'énergie mensuelle avec une estimation de l'incertitude robuste.

Modèles :

Nous avons utilisé deux modèles simples de réseaux de neurones basés sur les LSTM avec une architectures avec deux couches cachées et derrière chaque couche un dropout à 50%.
Le caractère aléatoire du dropout permet en générant une simulation de type Monté-Carlos de calculer un intervalle de confiance à 99% fiable.
Modèle 1 :
Couche 1 : 128 cellules LSTM,
Couche 2 : 32 cellules LSTM

Modèle 2 :
Couche 1 : 96 cellules LSTM,
Couche 2 : 24 cellules LSTM

Ces deux modèles ont été sélectionnés après un très grand nombre d'essais.

Expérimentations et résultats :

Afin de tester les modèles nous avons réalisé un split des données en train et test avec un ratio de 73%. Le jeu de données d'entraînement : 2012 au 06-2018 et le test de 07-2018 au 07-2019 dans le but de garder une année pour le test. Afin de trouver les meilleurs paramètres, du moins d'entrevoir des pistes pour les trouver, et éventuellement d'évaluer le surapprentissage, nous avons testé 2 batch-size et 2 nombre d'epochs différents, et pour chaque expérimentation nous avons calculé les MAE sur le train et le test. Nous avons également un learning rate de 0,001.

Modèle	Batch size	nombre d'epochs	MAE train	MAE test
Modèle 1	12	500	5157828	13908246
	12	1000	4583844	13466826

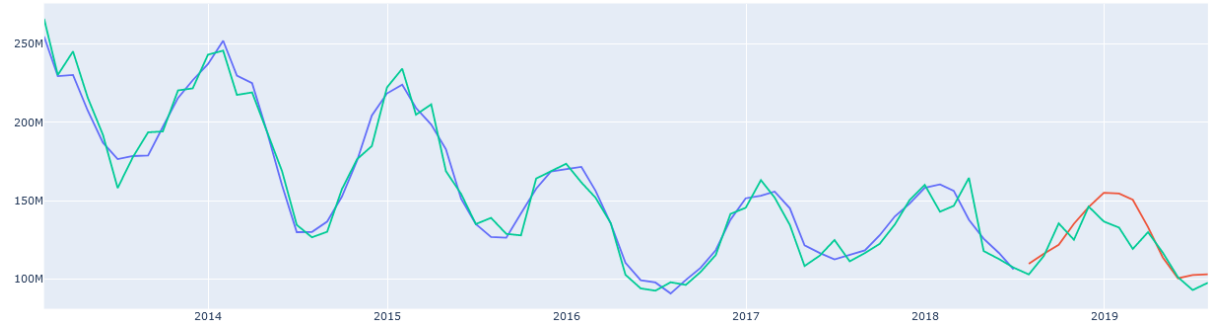
[1] Deep and Confident Prediction for Time Series at Uber : https://arxiv.org/pdf/1709.01907.pdf
https://colab.research.google.com/drive/1XSouHjY0ImMKNDYqIZtHPnDs_MESinwk#scrollTo=3FQQioKQ-Tkr

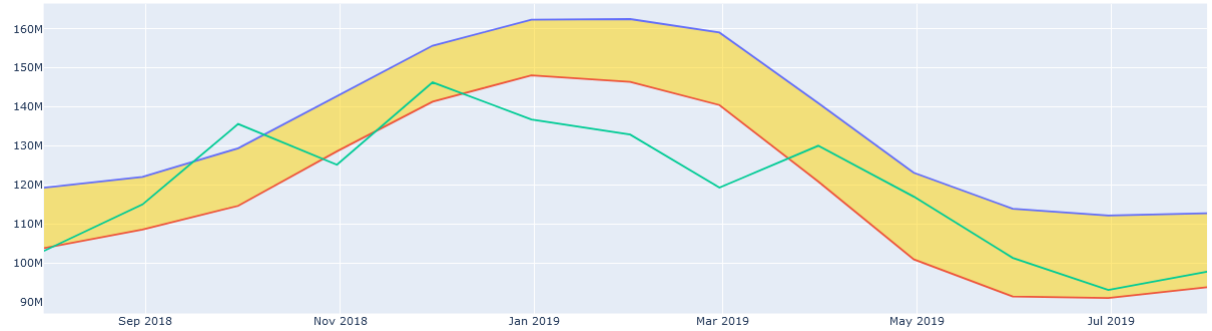
	24	500	6832084	9663458
	24	1000	6110157	9990941
Modèle 2	12	500	6430481	9997836
	12	1000	4939364	12292542
	24	500	8992399	9934823
	24	1000	6027939	10868592

Tableau 2 : Résumé des différentes expérimentations avec les deux modèles retenus et les résultats obtenus de la MAE sur les jeux de données de train et de test.

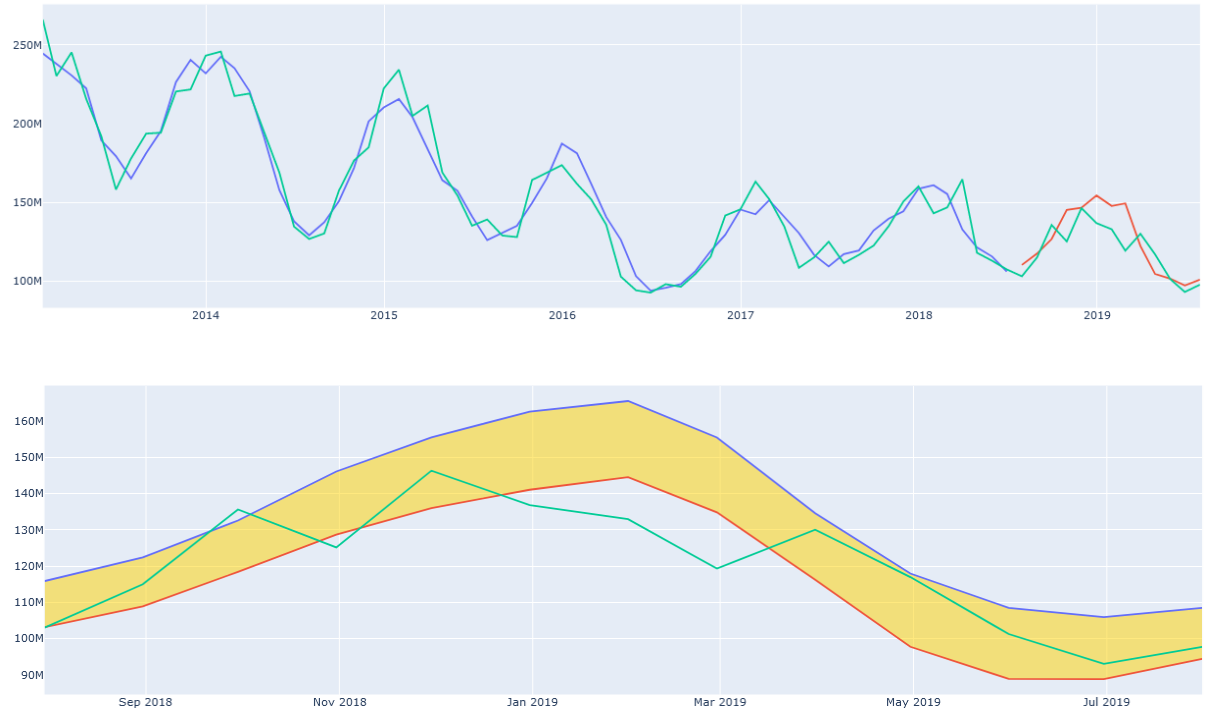
Nous avons testé les deux modèles avec des valeurs de batch size et de nombre d'epochs différents et après avoir calculé les MAE de train et de test, nous retenons les modèle 1 et 2 avec un batch size 24 et un nombre d'epochs de 500. Au vus des résultats du train les meilleurs modèles sont ceux avec un batch size 12 et un nombre d'epochs 1000, toutefois on note un écart important entre les MAE train et test ce qui suggère un surapprentissage de ces modèles.

Modèle 1 (Batch size = 24, epochs number = 500)





Modèle 2 (Batch size = 24, epochs number = 500)



Première figure : comparaison entre la vérité terrain (train et test) (couleur verte) avec les prédictions du modèle (train en bleu et test en rouge).
Deuxième figure : comparaison entre la vérité terrain test entourée des deux limités d'incertitude à 99% obtenues avec une simulation de type Monté-Carlos

Ces premiers tests sont à titre d'exemple car pour déterminer les paramètres optimaux, il est plus intéressant de réaliser une Grid Search. Toutefois de cette première étude, on voit l'effet de ces paramètres sur les résultats.

Conclusion

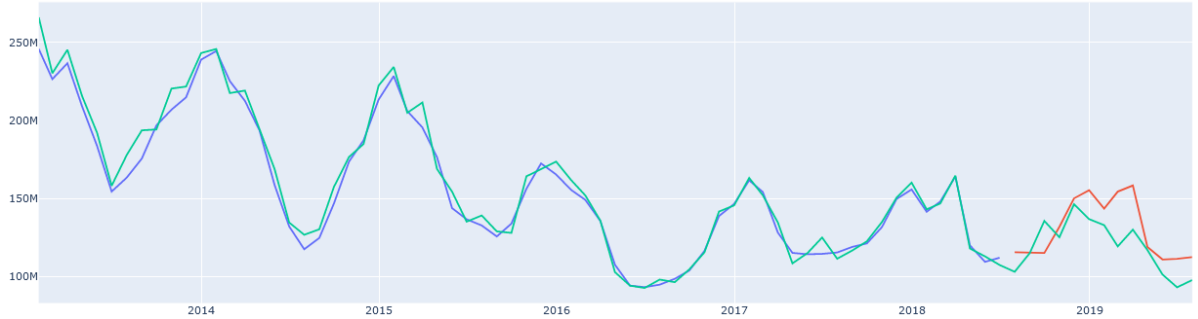
Nous avons présenté deux approches principales à l'analyse et la prédiction de séries temporelles : méthodes autorégressives telles que AR(I)MA et SARIMA et l'approche de Bayesian LSTM. Les deux approches sont très sensibles au tuning de leurs hyper-paramètres respectifs. Nous avons choisi de considérer l'énergie totale (sum des sources différentes) dans nos travaux prédictifs car les proportions des sources différentes dépendent des facteurs multiples : économiques, météorologiques, sociétaux et politiques, parmi les autres. Ces facteurs ne sont pas inclus dans les données, mais leur effet est implicitement reflété dans les corrélations observées entre les sources énergétiques différentes. Cependant ces corrélations ne démontrent pas une causalité. Pour cette raison, étudier la production de l'énergie provenant de chaque ressource séparément dans l'absence d'information concernant les facteurs (variables exogènes dans une série temporelle multivariate) mentionnés ci-dessus ne semble pas être sensé. En revanche, nous pouvons étudier et prédire la distribution des composantes énergétiques provenant de ressources différentes dans l'énergie totale à moment instantané.

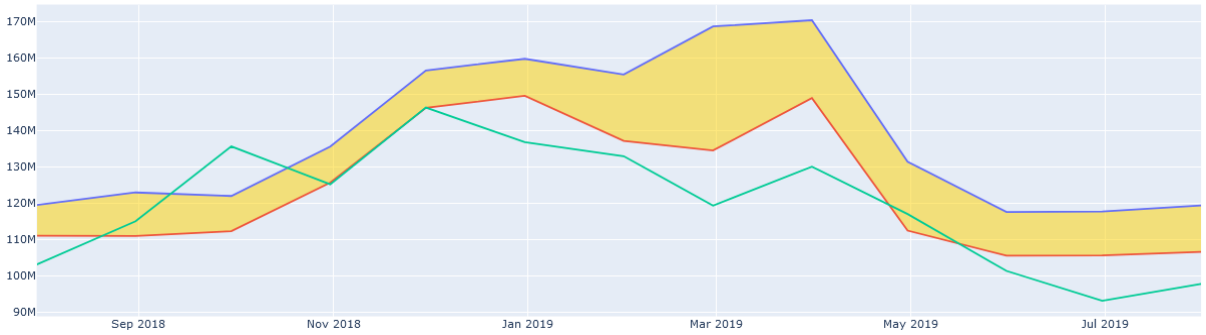
Finalement, nous mentionnons avoir obtenu des prédictions pour la production d'énergie totale au Royaume-Uni avec une erreur de l'ordre de 10 %.

Annexe

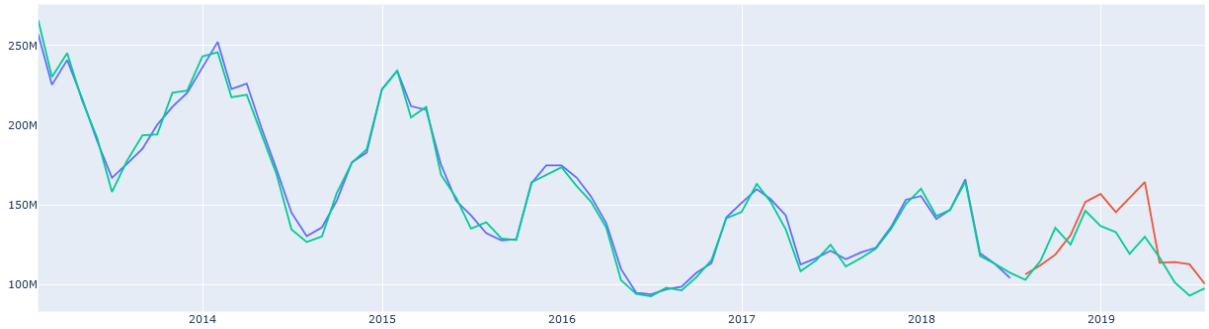
Deep Learning

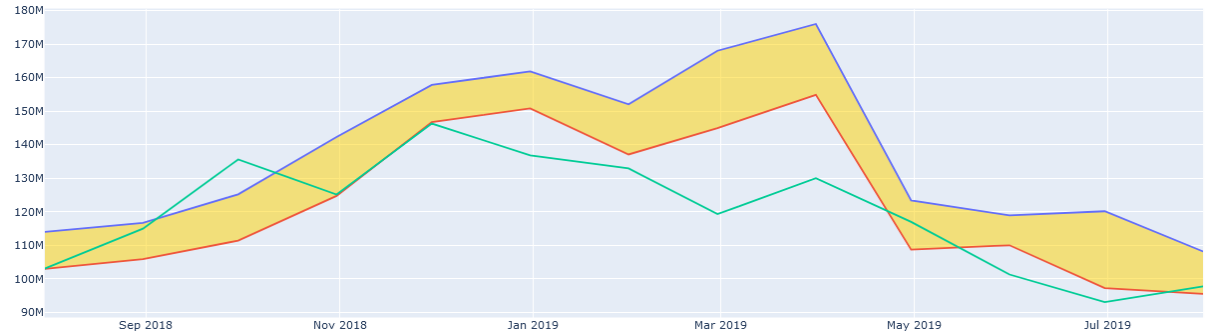
Résultats des différentes expérimentations avec les deux modèles LSTM Bayésien pour les différents batch size et nombre d'epochs retenus :
Première figure : comparaison entre la vérité terrain (train et test) (couleur verte) avec les prédictions du modèle (train en bleu et test en rouge).
Deuxième figure : comparaison entre la vérité terrain test entourée des deux limités d'incertitude à 99% obtenues avec une simulation de type Monté-Carlos
Modèle 1 (Batch size = 12, epchs number = 500)



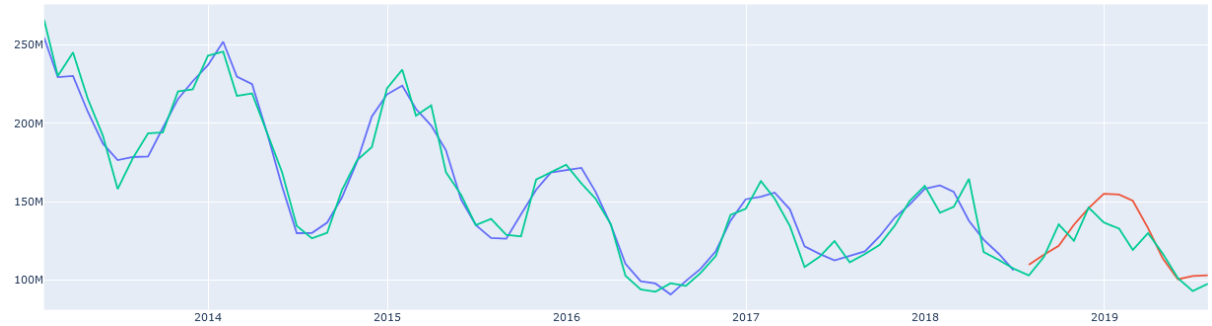


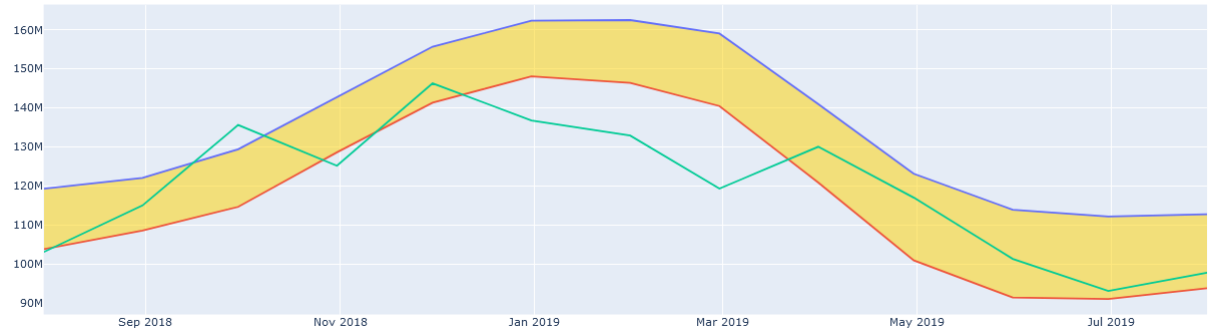
Modèle 1 (Batch size = 12, epochs number = 1000)



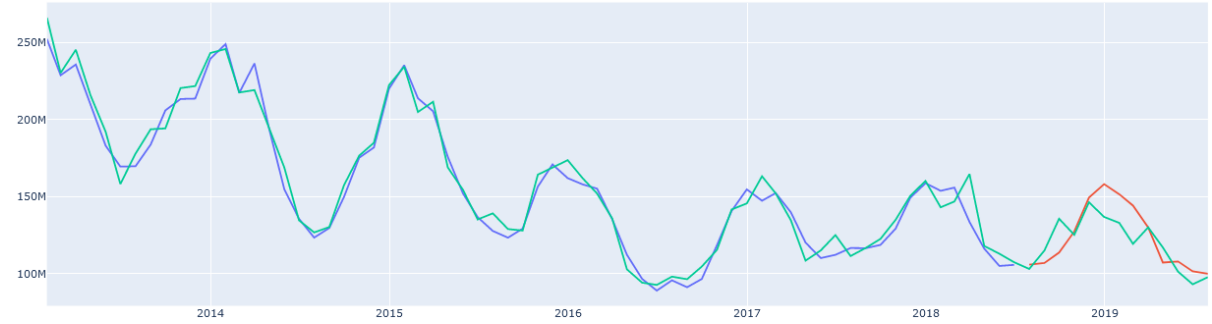


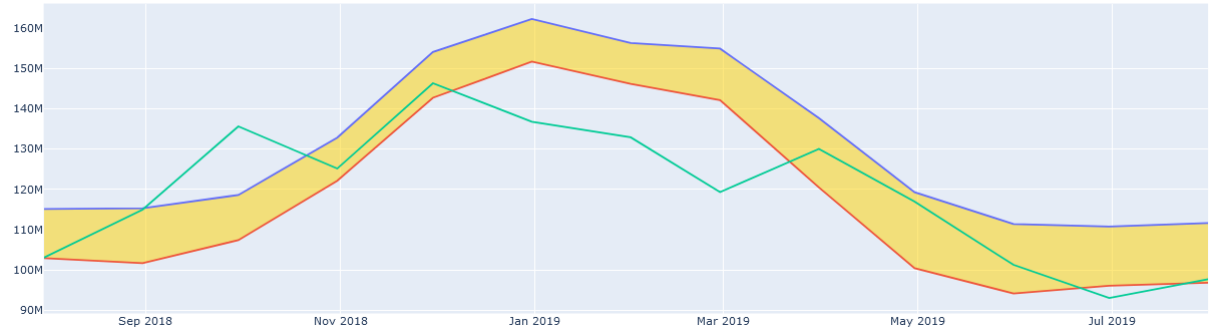
Modèle 1 (Batch size = 24, epochs number = 500)



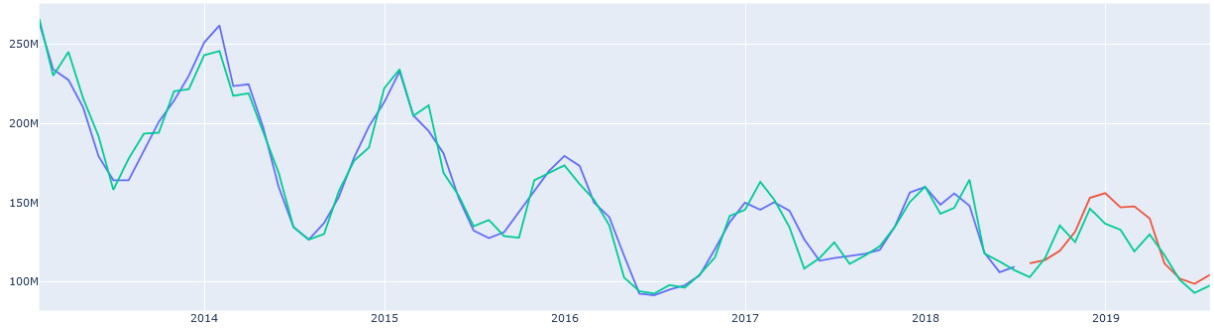


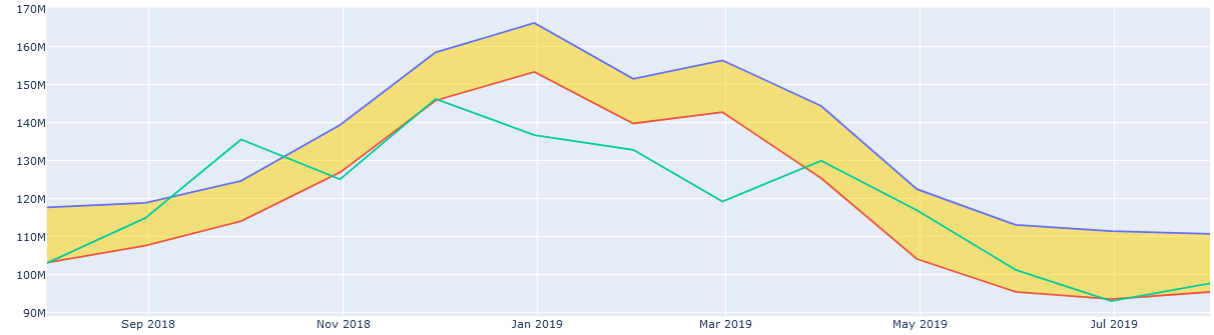
Modèle 1 (Batch size = 24, epochs number = 1000)



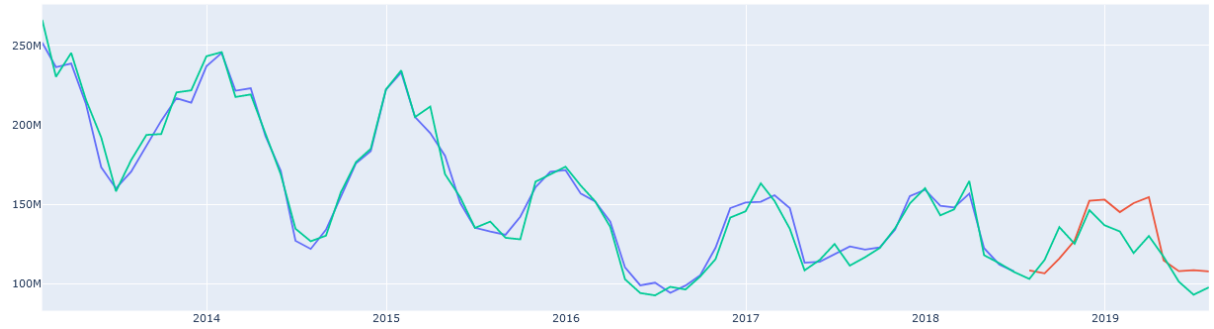


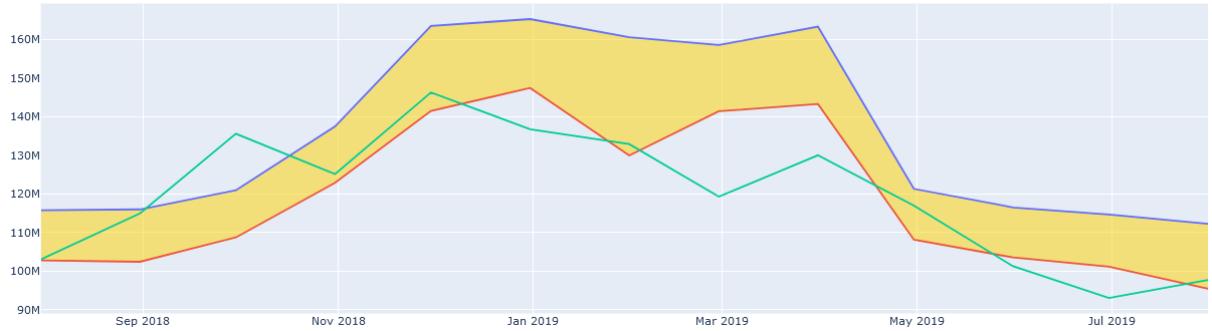
Modèle 2 (Batch size = 12, epochs number = 500)



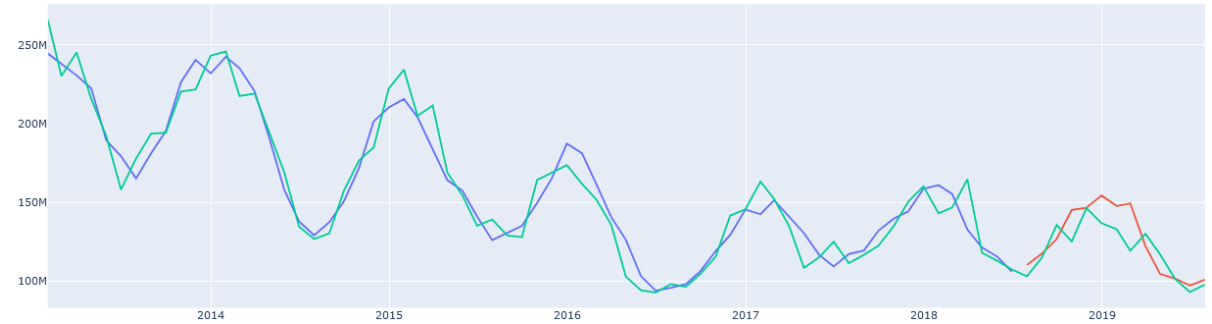


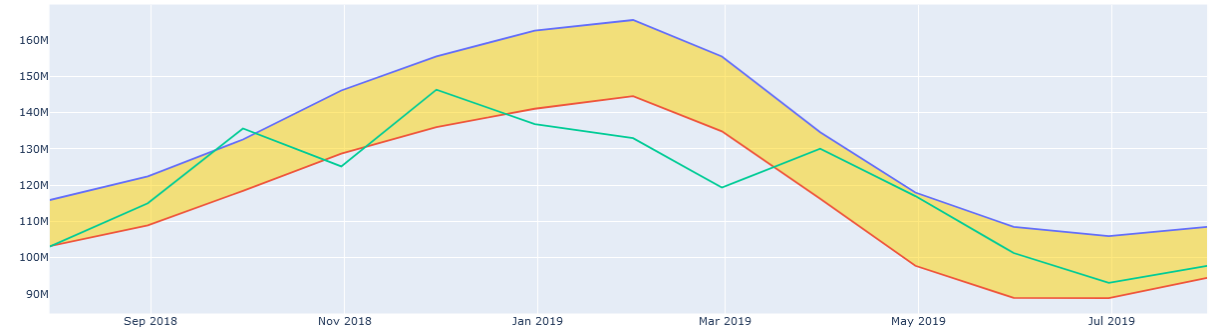
Modèle 2 (Batch size = 12, epochs number = 1000)



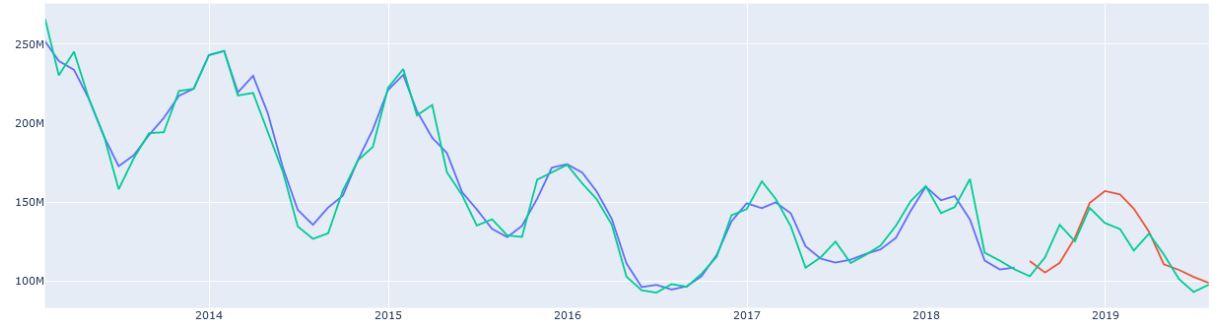


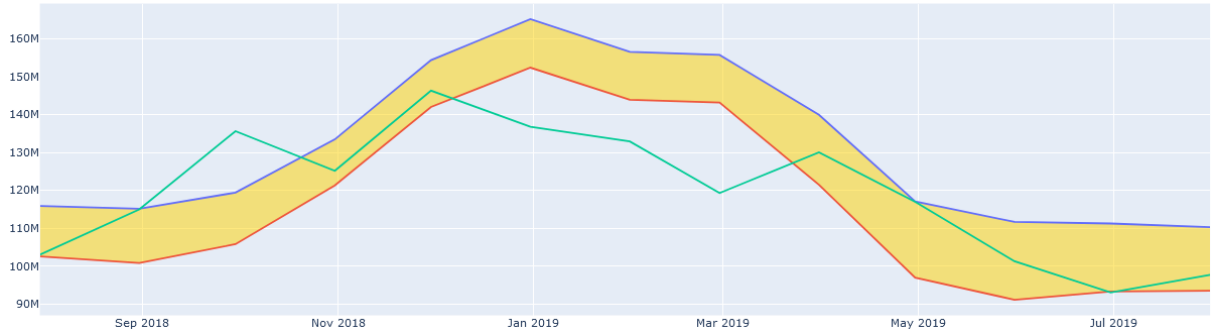
Modèle 2 (Batch size = 24, epochs number = 500)



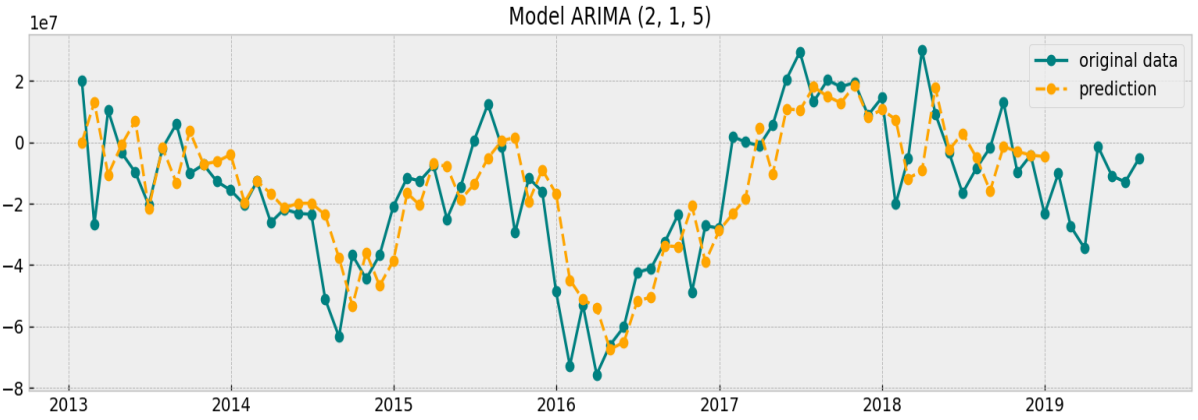


Modèle 2 (Batch size = 24, epochs number = 1000)

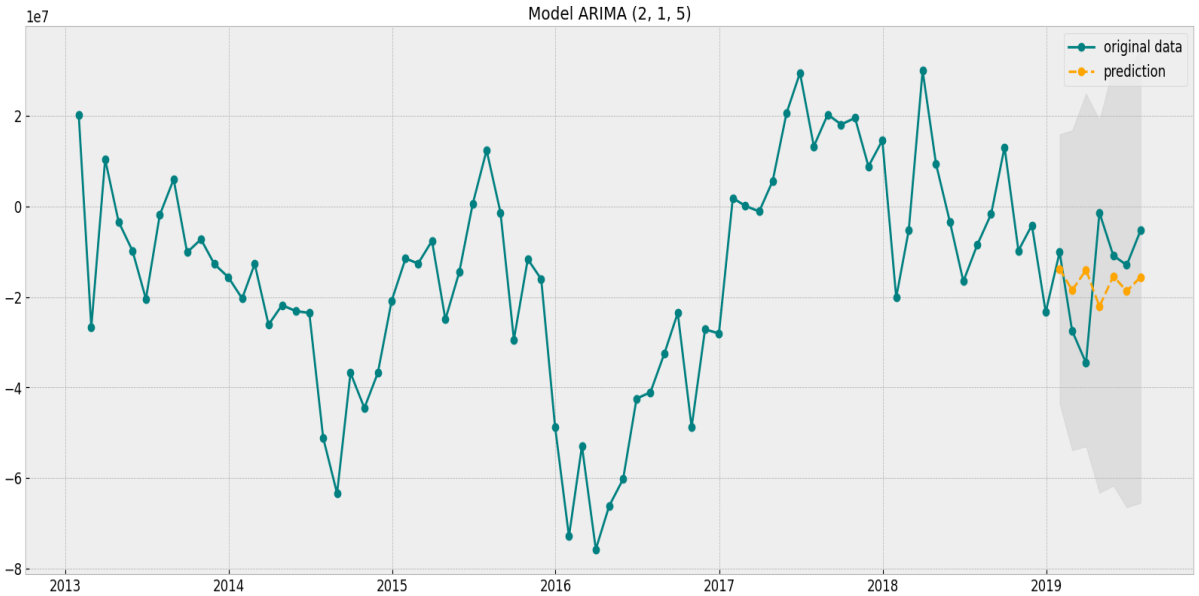




ARIMA



Train MAE = 11481686.226, Train MAPE = 3.663



Test MAE : 10645052.047, Test MAPE 2.727

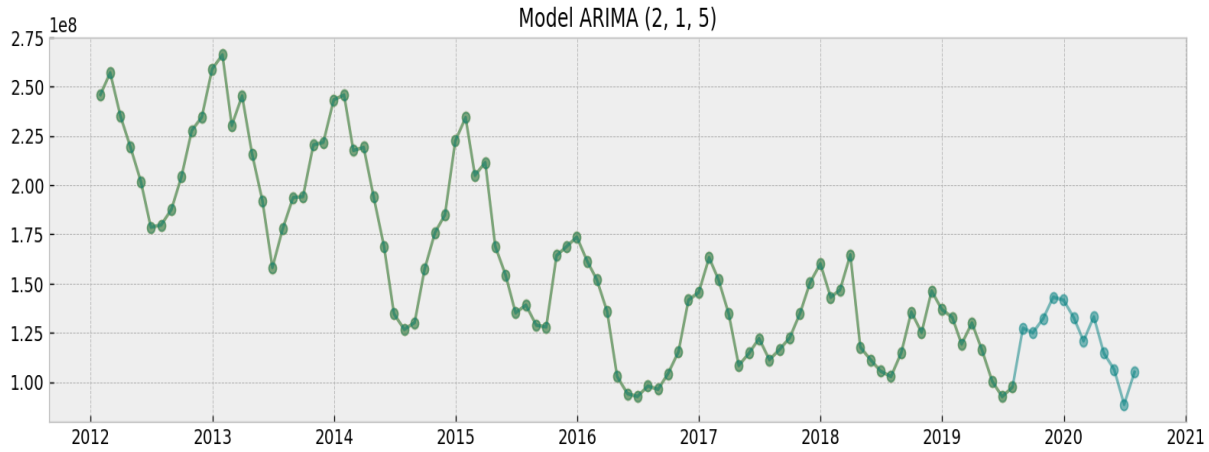


Fig. A Modèle ARIMA : Série originale (couleur verte) : Forecast du 2019/08 au 2020/07 (couleur turquoise).