SAE Description et Prévision de séries temporelles

Par Dimitri DERAMOND, Jeanne FOND et Louise THOMAS

IUT de Paris - Rives de Seine
Université Paris Cité

Sommaire

Introduction

Dans le cadre de la SAE série temporelle, il nous a été confié de réaliser différentes analyses sur une série chronologique donnée. Nous réalisons ce travail grâce au logiciel de programmation statistique Rstudio.
Les données fournies proviennent de l'Administration américaine de l'information sur l'énergie (EIA) qui est la principale agence aux Etats-Unis dans le système statistique fédéral, elle est chargée de collecter, d'analyser et de diffuser des informations sur l'énergie. Elles portent sur la production d'électricité entre 2001 et 2021, les données sont mensuelles et sont décomposées par type de production d'électricité, l'unité utilisée est le millier de mégawatts/heures.
Notre série chronologique porte sur la production d'électricité à partir de charbon dans l'État de l'Alaska aux Etats-Unis. Nous allons décrire le code que nous avons groupé en 9 parties.

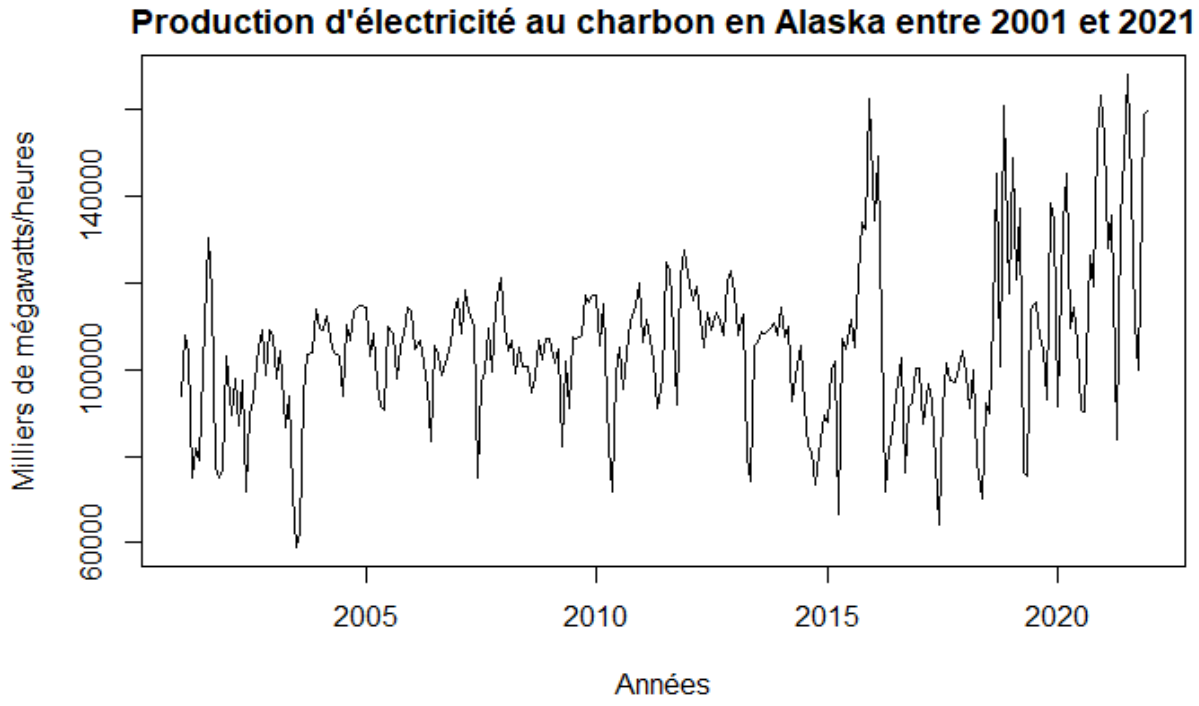
Importation des données

```
load(file = "D://IUT/SD2/SAE/series temporelles/SAE/AK_coal.RData")
data <- ts(AK_coal, frequency = 12, start=c(2001,1))
data_log <- log(data)
```

On charge la base de données contenant les données de production d'électricité au charbon en Alaska (AK_coal) et on les convertit en une série temporelle mensuelle. La conversion en une série temporelle est essentielle pour l'analyse des données chronologiques.
Pour convertir nos données en série une série temporelle on utilise la fonction suivante : `ts(AK_coal, frequency = 12, start=c(2001,1))`. La fréquence correspond au nombre de mois dans l'année car on a des données mensuelles soit 12, pour "`start`" on commence à l'année 2001 et au 1er mois de l'année soit Janvier (1).
On remarque que notre série temporelle se trouve être multiplicative. Ne sachant que traiter des séries additives nous appliquons la fonction "`log`" à nos données afin qu'elles se comportent comme une série additive : `data_log <- log(data)`. Pour la suite, une fois les analyses faites, nous appliquons la fonction "`exp`" afin d'annuler le "`log`" et d'avoir des résultats cohérents avec nos données initiales.

Graphique de la production d'électricité au charbon

```
plot(data, main = "Production d'électricité au charbon en Alaska entre 2001
et 2021", ylab="Milliers de mégawatts/heures", xlab="Années")
```

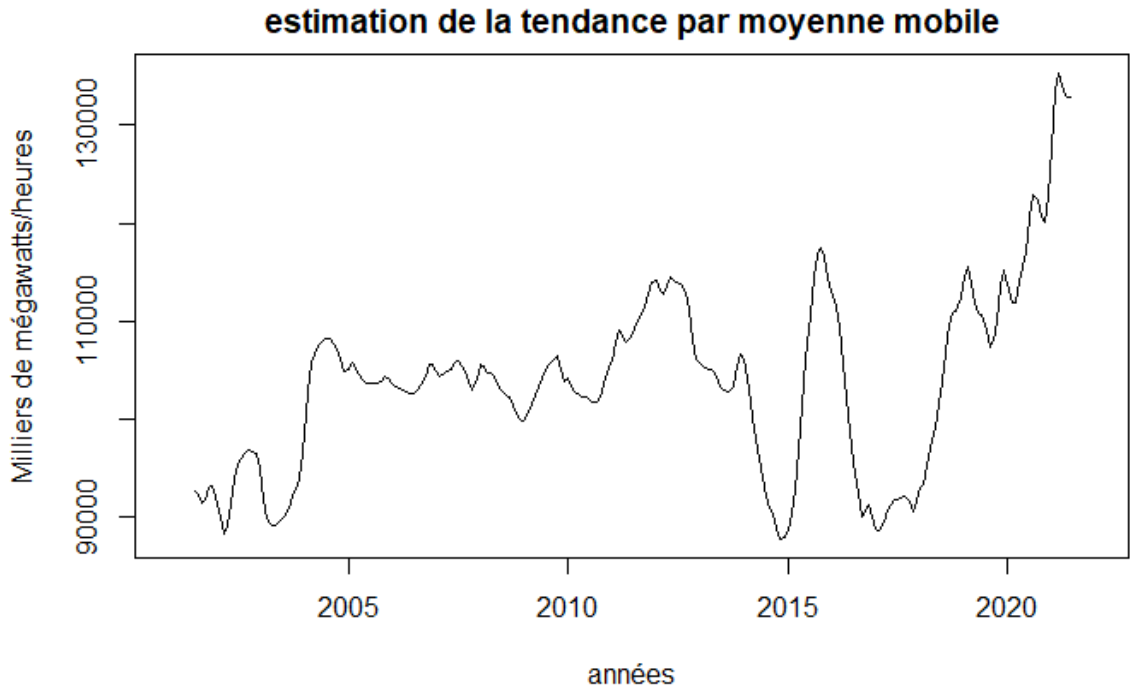


Ce code crée un graphique de la production d'électricité au charbon au fil du temps en Alaska grâce à la fonction "`plot`". Le graphique permet une visualisation rapide de la série chronologique, mettant en évidence les tendances et les motifs saisonniers.

Ce graphique nous montre que la quantité d'électricité produite au charbon est très variable, on note une tendance à une plus forte production depuis 2019. On remarque également qu'à une période de chaque année il y a une forte chute de la production.

Estimation de la tendance par moyenne mobile

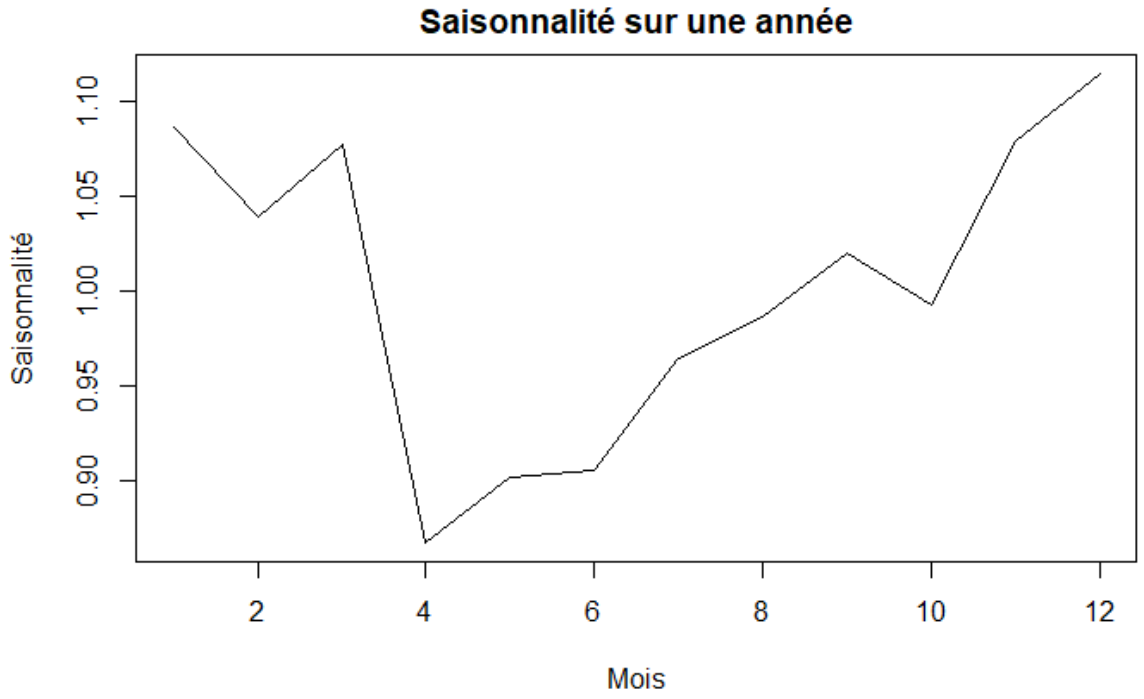
```
infos <- decompose(data_log)
FI<- exp(infos$trend)
plot(FI, type='l', ylab="Milliers de mégawatts/heures", xlab="années",
main="estimation de la tendance par moyenne mobile")
```



La visualisation de la tendance aide à identifier les changements à long terme dans la série chronologique. Ce code utilise la décomposition de la série temporelle. Pour se faire on exécute le code suivant : `infos <- decompose(data_log)`. On crée ainsi un dataframe contenant différentes informations sur notre série comme la saisonnalité avec "`seasonal`", la tendance avec "`trend`" ou encore les résidus avec "`random`". Pour estimer la tendance à l'aide de la moyenne mobile on exécutera donc le code suivant : `FI <- exp(infos$trend)`.
Ce graphique nous montre la tendance de production d'électricité par le charbon de 2001 à 2021. On peut voir plus clairement comme dit précédemment qu'il y a une production plus importante d'électricité depuis 2019. On remarque également des périodes avec une baisse notable de la production comme en 2015 et en 2017-2018.

Saisonnalité sur une année

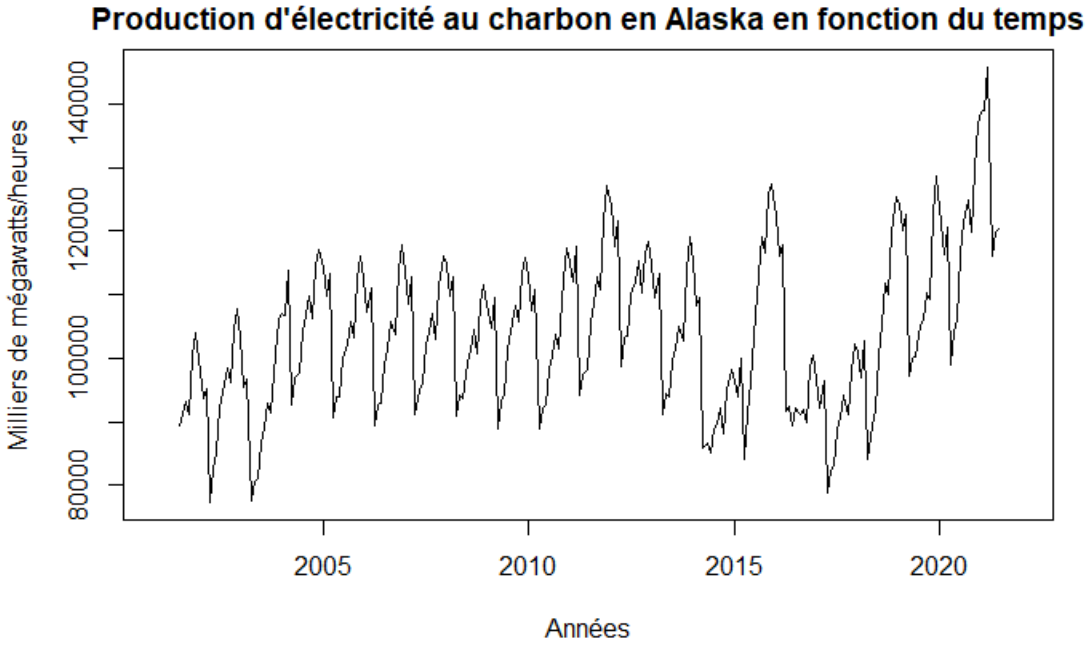
```
SI <- exp(infos$seasonal)
plot(x=c(1:12), y=SI[1:12], type='l', xlab="Mois", ylab="Saisonnalité",
main="Saisonnalité sur une année")
```



Ce code représente graphiquement les coefficients saisonniers sur une année. Il permet d'observer les variations saisonnières mensuelles dans la série temporelle. On calcul la saisonnalité grâce au code suivant : `SI <- exp(infos$seasonal).` On utilise la fonction “`exp`” afin d’annuler le “`log`” et d’avoir des résultats cohérents avec nos données initiales.
Grâce au tracé de la saisonnalité sur une année on peut observer comme dit précédemment une baisse de la production au cours de chaque année. On voit effectivement qu’au mois d’Avril il y a une baisse d’environ 20 000 milliers de mégawatt/heure par rapport aux autres mois.

Prédictions lissées

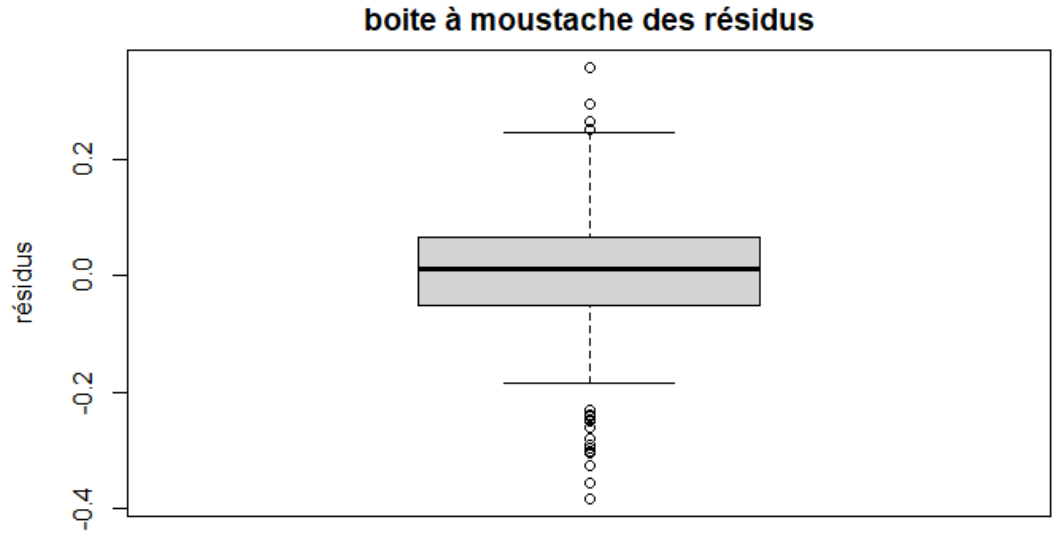
```
x<-as.vector(time(data_log))
y<-as.vector(data_log)

plot(x=x, y=(FI)*(SI),main="Production d'électricité au charbon en Alaska
en fonction du temps", xlab="Années", ylab="Milliers de mégawatts/heures",
type = "l")
lines(x=x, y=as.vector(data_log), type='l',col="red")
```



Ce code génère un graphique de la production d'électricité au charbon avec la tendance lissée en utilisant la multiplication des composantes de tendance et saisonnière. Il offre une vision plus claire de la série temporelle, en mettant en évidence la tendance et en comparant les prédictions lissées avec les données réelles. Pour se faire on utilise le code suivant : `plot(x=x, y=(FI)*(SI),main=... .` Où `x` correpond au temps et `y` correspond à la tendance (`FI`) fois la saisonnalité (`SI`).

Résidus

```
boxplot(infos$random, ylab="résidus", main="boîte à moustache des résidus")
```



Ce code affiche la boîte à moustache des résidus des prédictions effectuées. Il permet de mesurer à quel point les prédictions diffèrent des observations réelles. Pour le tracer on utilise la fonction suivante : `boxplot(infos$random)`. “`random`” contient les données

sur les résidus de notre série temporelle. Cette boîte à moustache se trouve plutôt centrée et les moustaches sont à peu près de la même longueur des deux côtés, ce qui suggère que les résidus sont distribués symétriquement autour de zéro. On remarque néanmoins de nombreuses valeurs aberrantes suggérant une différence importante avec nos observations pouvant avoir un impact sur notre modèle.

Prévisions pour l'année 2022

Méthode ARIMA

```
#prevision avec ARIMA

z <- arima(data, order=c(1,1,1), seasonal=list(order=c(1,1,1), period=12))
prev <- predict(z,n.ahead=12)
prev
```

On fait un ajustement du modèle ARIMA grâce à "arima", avec comme paramètres `order=c(1,1,1)`, donc un modèle ARIMA(1,1,1). De plus, on a pris en compte la saisonnalité grâce à "seasonal=list(order=c(1,1,1), period=12)".
On génère des prévisions pour les 12 prochains mois grâce à la fonction "predict". En résumé, on effectue l'ajustement d'un modèle ARIMA, l'évaluation de sa qualité, l'examen des résidus, et la génération de prévisions pour les mois à venir.

Méthode Holt Winter

```
#prevision avec Holtwinter

hw <- HoltWinters(data)
prev_hw <- predict(hw,n.ahead=12)
```

Ici, on ajuste un modèle Holt-Winters grâce à la fonction HoltWinters(), et on génère des prévisions pour les 12 prochains mois avec la fonction predict, puis on met ces prévisions dans le graphique final.

Modèle polynomial et linéaire

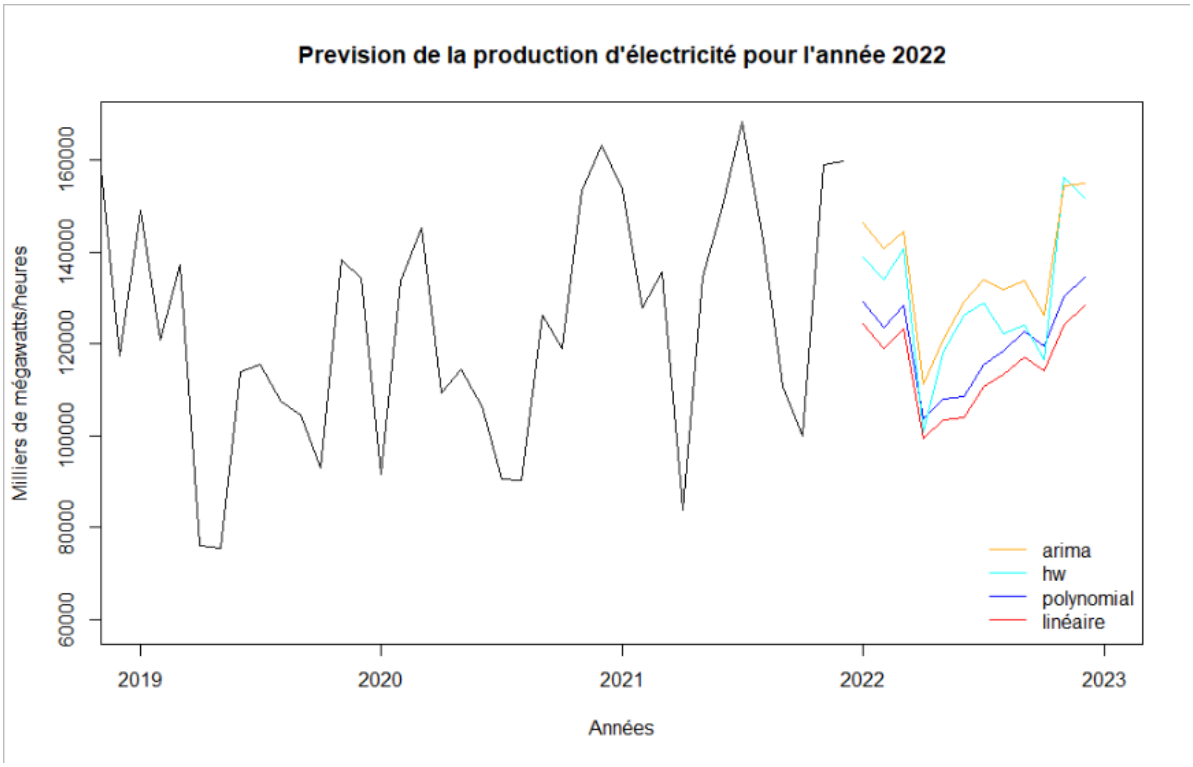
```
#prevision par modèle polynomial et linéaire

CVS_mul <- exp(data_log) / exp(infos$seasonal)
CVS_mul

modele_lineaire_mul <- lm(CVS_mul ~ x)

#les coeffs estimés : modele_lineaire_mul$coefficients

tend_lin_mul<-modele_lineaire_mul$coefficients[1]+modele_lineaire_mul$coeff
icients[2]*x
```

Ici, on ajuste des modèles linéaire et polynomial aux données avec la fonction lm, et on calcule les prévisions pour les 12 prochains mois en utilisant les coefficients estimés. On intègre ces prévisions dans le graphique final.

Pour la réalisation du graphique des prévisions 2022 avec les différentes méthodes on exécute ce code :

```
# Prévisions pour l'année 2022

x_fut <- seq(2022, 2022 + 11/12, by = 1/12)
tend_lin_fut <- modele_lineaire_mul$coefficients[1] +
modele_lineaire_mul$coefficients[2] * x_fut
tend_poly_fut <- modele_polynomial_mul$coefficients[1] +
modele_polynomial_mul$coefficients[2] * x_fut +
modele_polynomial_mul$coefficients[3] * x_fut^2

plot(data, main="Prevision de la production d'électricité pour l'année 2022
(vue zoomée)", xlab="Années", ylab="Milliers de mégawatts/heures",
xlim=c(2019,2023))

lines(x_fut, tend_lin_fut * exp(infos$seasonal)[1:12], type="l", col="red")
lines(x_fut, tend_poly_fut * exp(infos$seasonal)[1:12], type="l",
col="blue")
lines(prev_hw, col="cyan")
lines(prev$pred, col="orange")
legend("bottomright", lty=c(1:1), col=c("orange", "cyan", "blue", "red"),
legend=c("arima","hw","polynomial","linéaire"), bty="n")
```

Grâce à ces trois méthodes différentes, nous obtenons le graphique suivant :



Ce graphique nous donne un aperçu des prévisions de production d'électricité au charbon en Alaska pour l'année 2022, en utilisant les différentes méthodes que nous avons vu juste avant. Les prédictions issues des modèles polynomial et linéaire sont représentées en bleu et rouge. On remarque que ces deux prédictions se ressemblent beaucoup, mais diffèrent considérablement des prévisions des deux autres modèles (ARIMA et HoltWinters), qui présentent également des similitudes entre eux.

Prévisions pour l'année 2021

On reproduit ces trois méthodes en supprimant du jeu de données initiales l'année 2021. Nous appelons cet échantillon : "`apprentissage`". Il nous permet de prédire l'année 2021 en n'ayant pas connaissance de cette année. Nous pouvons ainsi comparer les prédictions aux données réelles et se rendre compte de l'efficacité des modèles.

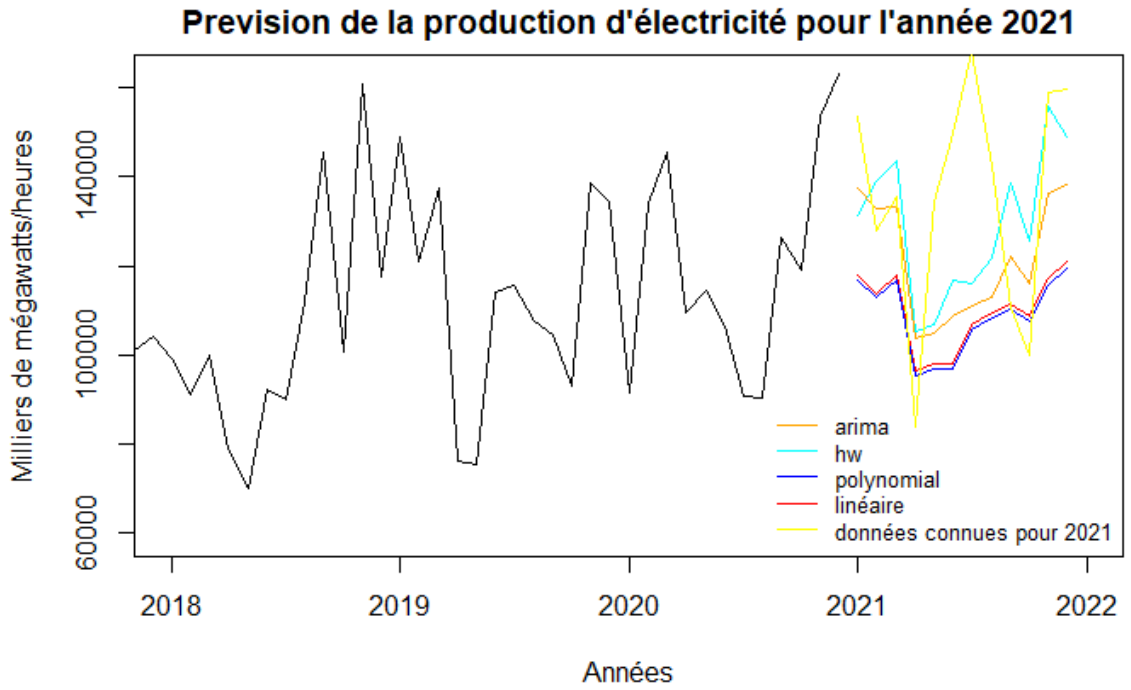
```
apprentissage <- ts(data[1:240], frequency = 12, start=c(2001,1)) #toutes
les années sauf la dernières
ech_test <- ts(data[241:252], frequency= 12, start=c(2021,1)) #dernière
année

#ARIMA
z1 <- arima(apprentissage, order=c(1,1,1), seasonal=list(order=c(1,1,1),
period=12))
prev_ech <- predict(z1, n.ahead = 12)

mse_arima <- mean((prev_ech$pred-ech_test)^2)

#Holt-Winters
hw_app <- HoltWinters(apprentissage)
prev_hw_app <- predict(hw_app,n.ahead=12)

mse_hw <- mean((prev_hw_app-ech_test)^2)

#prevision par modèle polynomial et linéaire

apprentissage_log <- log(apprentissage)
infos_ech <- decompose(apprentissage_log)
x1=x[1:240]

CVS_mul_ech <- exp(apprentissage_log) / exp(infos_ech$seasonal)

modele_lineaire_mul_ech <- lm(CVS_mul_ech ~ x1)
tend_lin_mul_ech <-
modele_lineaire_mul_ech$coefficients[1]+modele_lineaire_mul_ech$coefficient
s[2]*x1

mse <- sqrt(mean((exp(apprentissage_log) - exp(tend_lin_mul_ech))^2))
```

On construit notre graphique avec le code suivant :

```
#prévisions année 2021

x_fut_ech <- seq(2021, 2021 + 11/12, by = 1/12)
tend_lin_fut_ech <- modele_lineaire_mul_ech$coefficients[1] +
modele_lineaire_mul_ech$coefficients[2] * x_fut_ech
tend_poly_fut_ech <- modele_polynomial_mul_ech$coefficients[1] +
modele_polynomial_mul_ech$coefficients[2] * x_fut_ech +
modele_polynomial_mul_ech$coefficients[3] * x_fut_ech^2

plot(apprentissage, main="Prevision de la production d'électricité pour
l'année 2021", xlab="Années", ylab="Milliers de mégawatts/heures",
xlim=c(2018,2022))

lines(x_fut_ech, tend_lin_fut_ech * exp(infos_ech$seasonal)[1:12],
```

```
type="l", col="red")
lines(x_fut_ech, tend_poly_fut_ech * exp(infos_ech$seasonal)[1:12],
type="l", col="blue")
lines(prev_hw_app, col="cyan")
lines(prev_ech$pred, col="orange")
lines(ech_test, col="yellow")
legend("bottomright", lty=c(1:1), col=c("orange", "cyan", "blue", "red",
"black"), legend=c("arima","hw","polynomial","linéaire", "données connues
pour 2021"), bty="n", cex=0.8)
```

Ainsi, on obtient ce graphique :



Ce graphique nous permet de visualiser les prévisions de production d'électricité par le charbon en Alaska pour l'année 2021 selon les différentes méthodes expliquées précédemment.
En rouge et bleu on observe les prédictions faites à partir du modèle linéaire et polynomial, on peut voir que ces deux prédictions sont assez similaires entre elles mais très différentes des données réelles de 2021.
Pour les méthodes ARIMA en orange et HoltWinter en bleu clair on constate là aussi une forte différence de prédiction pour 2021 par rapport aux réelles de cette année-là.
Ainsi ce graphique permet de voir que les prédictions pour 2021 à partir de ces différentes méthodes sont très éloignées des données réelles, ce qui peut signifier que nos prévisions sont mauvaises.

Erreur quadratique moyenne de prévision

```
mse_arima <- mean((prev_ech$pred-ech_test)^2)

mse_hw <- mean((prev_hw_app-ech_test)^2)
```

L'erreur quadratique moyenne de prévision est une mesure de la précision d'un modèle de prévision.

Pour calculer l'erreur quadratique moyenne de prévision de la méthode ARIMA on utilise le code suivant : `mse_arima <- mean((prev_ech$pred-ech_test)^2)`.
On calcule donc les différences entre les données prédites (`prev_ech$pred`) et les données réelles (`ech_test`) le tout au carré (`^2`) puis on fait la moyenne (`mean`). On obtient donc une erreur quadratique moyenne égale ici à : 731360801.
On fait le calcul sur le même principe pour la méthode HW : `mse_hw <- mean((prev_hw_app-ech_test)^2)`. L'erreur quadratique moyenne ici est égale à : 644828001.
On remarque que pour les méthodes ARIMA et HW l'erreur quadratique moyenne de prévision est extrêmement grande ce qui signifie que nos modèles ne sont pas précis. Les données à notre disposition pourraient être trop complexes pour les modèles que nous avons effectués ou bien les paramètres de nos modèles manquent de précisions par rapport à nos données.

Conclusion

Dans le cadre de la SAE série temporelle, notre mission consistait à analyser une série chronologique portant sur la production d'électricité à partir du charbon en Alaska, couvrant la période de 2001 à 2021, avec des données fournies par l'Administration américaine de l'information sur l'énergie (EIA).
La série temporelle obtenue s'est avérée multiplicative, nécessitant une transformation logarithmique pour faciliter les analyses.
L'analyse de la saisonnalité a souligné des variations mensuelles, notamment une baisse en avril. Ensuite, nous avons fait des prévisions pour l'année 2022 en utilisant trois méthodes différentes : ARIMA, Holt-Winters, et des modèles polynomial et linéaire. Le graphique final a illustré la diversité des prédictions entre ces méthodes, soulignant des similitudes entre les modèles polynomial et linéaire, ainsi qu'avec ARIMA et Holt-Winters, mais différentes entre elles.
En conclusion, bien que les méthodes aient permis d'explorer la série temporelle et de générer des prévisions, les résultats soulignent la nécessité d'une approche plus approfondie pour améliorer la précision des modèles, notamment en ajustant les paramètres et en explorant d'autres méthodes de modélisation.

Résumé en anglais

In the context of the Time Series SAE, the analysis focused on the coal electricity production time series in Alaska from 2001 to 2021. Monthly data, categorized by production type, were provided by the U.S. Energy Information Administration. Initial visualizations revealed significant variability and an upward trend since 2019. Trend estimation, seasonality analysis, and forecasts for 2022 were conducted using ARIMA, Holt-Winters, and polynomial/linear models. Results indicated diverse predictions among methods, with noticeable discrepancies for 2021. Mean Square Error calculations suggested limited accuracy, emphasizing the need for further model refinement and exploration of alternative approaches.