Séries temporelles et reproductibilité

Andrea Katherín Carranza Díaz

2020-07-29

## Consigne

Les données du fichier hawai.csv comprennent les moyennes des mesures mensuelles de CO2 atmosphérique entre en ppm-volume collectées au Mauna Loa Observatory à Hawaii de mars 1958 à décembre 2001, inclusivement.

## Lecture des données

library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("lubridate")

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:dplyr':  
##   
## intersect, setdiff, union

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library("forecast")

## Warning: package 'forecast' was built under R version 4.0.2

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library("fpp2")

## Warning: package 'fpp2' was built under R version 4.0.2

## Loading required package: fma

## Warning: package 'fma' was built under R version 4.0.2

## Loading required package: expsmooth

## Warning: package 'expsmooth' was built under R version 4.0.2

hawai <- read\_csv("hawai.csv")

## Parsed with column specification:  
## cols(  
## time = col\_double(),  
## CO2 = col\_double()  
## )

glimpse(hawai)

## Rows: 526  
## Columns: 2  
## $ time <dbl> 1958.167, 1958.250, 1958.333, 1958.417, 1958.500, 1958.583, 19...  
## $ CO2 <dbl> 316.1000, 317.2000, 317.4333, 317.4333, 315.6250, 314.9500, 31...

Le tableau de données est lu, qui comporte 526 observations et 2 variables, toutes deux à double caractère, il est donc nécessaire de convertir la variable **time** pour qu’elle soit reconnue comme une date. Ceci est fait au moyen de la fonction lubridate::date\_decimal() et UTC est défini comme le fuseau horaire requis. Il est possible de vérifier cette transformation avec la fonction de base class().

CO2\_hawai <- hawai %>%  
 mutate(time = date\_decimal(hawai$time, tz = "UTC"))  
glimpse(CO2\_hawai)

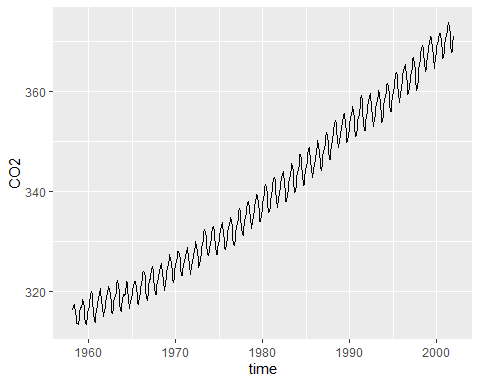
## Rows: 526  
## Columns: 2  
## $ time <dttm> 1958-03-02 20:00:01, 1958-04-02 06:00:00, 1958-05-02 16:00:00...  
## $ CO2 <dbl> 316.1000, 317.2000, 317.4333, 317.4333, 315.6250, 314.9500, 31...

CO2\_hawai %>% pull(time) %>% class()

## [1] "POSIXct" "POSIXt"

Une fois la transformation effectuée, nous procédons à la graduation pour observer comment le CO2 se comporte dans le temps. On observe que les données présentent une tendance croissante et présentent également une fluctuation saisonnière.

CO2\_hawai %>%  
 ggplot(aes(x = time, y = CO2)) +  
 geom\_line()



## Création de la série temporelle

La fonction stats::ts() sera utilisée pour créer la série temporelle. L’argument start, définit la date de la première observation (mars 1958) et l’argument frequency définit le nombre d’observations par unité de temps. Dans ce cas, les mesures ont été prises chaque mois, frequency = 12.

hawai\_ts <- ts(CO2\_hawai %>% dplyr::select(-time),  
 start = c(1958, 3), frequency = 12)

## Séparation de la série en parties d’entraînement

La fonction windows() sera utilisée pour séparer les séries en parties d’entraînement et de test.

hawai\_train <- window(hawai\_ts, start = c(1958, 3), end = c(1988, 12))  
hawai\_test <- window(hawai\_ts, start = c(1989, 1), end = c(2001, 12))

Environ 70 % des mesures sont attribuées à la série d’entraînement (les quelque 370 premières observations, qui ont été faites de mars 1958 à décembre 1988) et les 30 % restants des observations sont attribuées à la série d’essai (de janvier 1989 à décembre 2001) :

## Création d’un modèle prévisionnel sur les données d’entraînement

Le modèle choisi pour les données sur la formation était le modèle ETS (*erreur:* sans tendance - tendance additive - tendance adouci, *tendance :* sans saison - saison additive - saison multiplicative et *saisonnalité :* erreur additive - erreur multiplicative), qui fait partie de la famille SES. L’optimisation du modèle est appliquée aux séries de formation en utilisant la fonction de forecast::ets() qui permet de connaître le type d’erreur du modèle, la tendance et la station, ainsi que de trouver les paramètres de lissage α, β, γ du modèle. Le modèle obtenue correspond à un modèle ETS(A,A,A). Cela signifie qu’une erreur, une tendance et une saison de type additive sont obtenues.

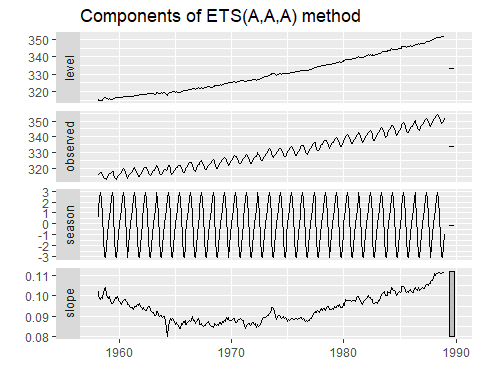
la prévision est réalisée par modélisation additive en fonction du niveau, de la tendance et de la saison. Le paramètre α décrit la distribution des poids. Dans ce cas, on a obtenu un α=0.6698, ce qui signifie que les événements plus récents du modèle ont un poids plus important dans la prévision.Pendant que, le valeur de β = 0.003 (s’approche de zéro) indique que la tendance change à une faible vitesse. γ correspond à la portion saisonnière. Généralement, ce paramètre fluctue autour de zéro, dans ce cas, γ = 2e-04.Le modèle ne présente pas de valeur ϕ, qui adouci la pente. Il est donc probable que l’adoucissement ne se justifie pas.

Le graphique du niveau en fonction du temps indique que la tendance de la concentration de CO2 dans l’atmosphère augmente avec le temps. En outre, dans le graphique de la station au fil des ans, on constate que la station est de nature additive puisque l’effet saisonnier fluctue autour de zéro.

hawai\_ets <- hawai\_train %>% ets()  
hawai\_ets

## ETS(A,A,A)   
##   
## Call:  
## ets(y = .)   
##   
## Smoothing parameters:  
## alpha = 0.6698   
## beta = 0.003   
## gamma = 2e-04   
##   
## Initial states:  
## l = 315.2483   
## b = 0.102   
## s = 0.6164 0.0116 -0.9384 -2.0207 -3.0899 -2.8853  
## -1.2381 0.7047 2.2379 2.8344 2.416 1.3513  
##   
## sigma: 0.3404  
##   
## AIC AICc BIC   
## 1408.202 1409.940 1474.731

autoplot(hawai\_ets)

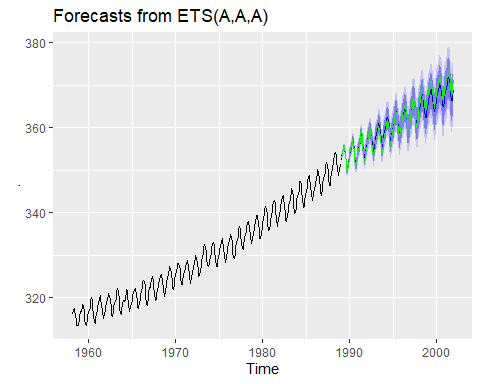


## Prévision de CO2 atmosphérique pour comparer aux données test

La fonction forecast: : forecast() a été utilisée pour obtenir la prédiction du modèle. L’argument h (nombre de périodes pour les prévisions) de 156 qui correspond au nombre d’observations, qui composent la série de test (longeur de la série de test).

Le graphique suivant présente les performances du modèle de prévision : la ligne bleue indique les valeurs obtenues, tandis que la ligne verte indique les valeurs obtenues à partir de la série de tests. Les régions bleues et transparentes font référence aux intervalles prévisionnels du modèle.

hawaifc <- hawai\_ets %>% forecast(h = length(hawai\_test))  
autoplot(hawaifc) + autolayer(hawai\_test, color = "green")

 On observe que l’approximation du modèle est plus élevée pour certaines années, tandis que pour d’autres, il y a une sous-estimation ou une surestimation, cependant, il est pertinent de faire une analyse de précision et de résidus pour la vérifier.

Dans le tableau suivant, on observe que l’erreur d’échelle absolue moyenne (MASE) de la prévision est de 0,7198956 et celle du modèle est de 0,2034533, ce qui indique que le modèle fournit une bonne prévision puisque plus cette valeur est proche de zéro, meilleure est la capacité de prévision du modèle.

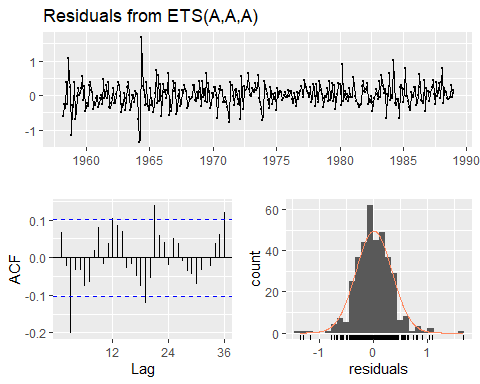
accuracy(hawaifc, hawai\_ts)

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.009127538 0.3329635 0.2496935 0.002272597 0.07593717 0.2034533  
## Test set 0.258792487 1.1125984 0.8835110 0.067546737 0.24278880 0.7198956  
## ACF1 Theil's U  
## Training set 0.06670592 NA  
## Test set 0.94372350 0.8673732

## Analyse des résidus

Afin de vérifier ces propriétés, la fonction forecast::checkresiduals() est utilisée comme suit :

checkresiduals(hawai\_ets)



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,A,A)  
## Q\* = 54.438, df = 8, p-value = 5.676e-09  
##   
## Model df: 16. Total lags used: 24

Il est à noter que la p-value obtenue est de 5.676e-09.Il est statistiquement significative. Elle indique une preuve solide contre l’hypothèse nulle, car il y a moins de 5% de probabilité que l’hypothèse nulle soit correcte (et les résultats sont aléatoires).

En général, les graphiques ci-dessus montrent que la moyenne des résidus est proche de zéro et qu’il n’y a pas de corrélation significative entre eux. La variation présentée dans le graphique des résidus en fonction du temps tend à être constante. Alors que l’histogramme des résidus indique qu’il n’y a pas de distribution normale, la normalité des résidus sera vérifiée par un test de Shapiro-Wilk.

shapiro.test(residuals(hawai\_ets))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(hawai\_ets)  
## W = 0.96983, p-value = 6.061e-07

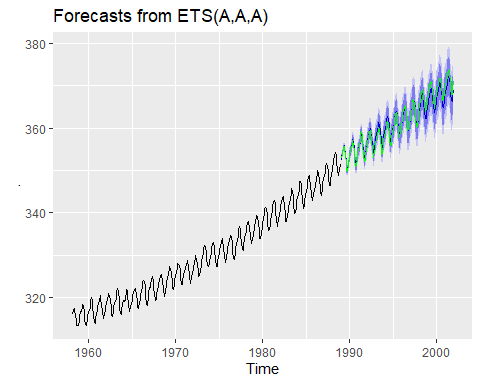
La p-value (~ 6.061e-07) indique que la distribution des résidus n’est pas normale.

Une méthode de prévision est considérée bonne lorsque ses résidus présentent les propriétés suivantes : absence de corrélation entre les résidus, la moyenne des résidus est de zéro, la variance des résidus est constante, les résidus sont normalement distribués. Bien que de ces quatre paramètres, le dernier n’est pas rempli, la méthode appliquée génère de bonnes prévisions. Cependant, l’application d’une transformation de Box-Cox peut parfois contribuer à améliorer la normalité de la distribution des résidus.

### Application de la transformation de Box-Cox

La modélisation des séries temporelles avec la transformation Box-Cox est effectuée de la même manière que pour les séries temporelles sans transformation, mais en utilisant le paramètre lambda dans l’application de la méthode ETS, qui est estimé avec la fonction de prévision::BoxCox.lambda().

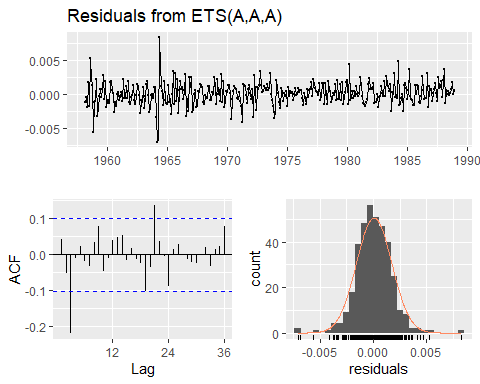
hawai\_BC <- hawai\_train %>% ets(lambda = BoxCox.lambda(hawai\_train))  
  
hawaifc\_BC <- hawai\_BC %>%  
 forecast(h = length(hawai\_test))  
  
autoplot(hawaifc\_BC) + autolayer(hawai\_test, color = "green")



accuracy(hawaifc\_BC, hawai\_ts)

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01317536 0.3201071 0.2414950 0.003562256 0.07340163 0.1967731  
## Test set 0.17310078 1.0530018 0.8432859 0.044216546 0.23196266 0.6871197  
## ACF1 Theil's U  
## Training set 0.03928175 NA  
## Test set 0.94442197 0.8221877

checkresiduals(hawai\_BC)



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,A,A)  
## Q\* = 42.618, df = 8, p-value = 1.037e-06  
##   
## Model df: 16. Total lags used: 24

shapiro.test(residuals(hawai\_BC))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(hawai\_BC)  
## W = 0.96958, p-value = 5.491e-07

Après l’application de la transformation, il est possible de constater que les paramètres statistiques ont diminué, mais pas de manière significative. En ce qui concerne les résidus, ils ne présentent pas d’amélioration, conservant les mêmes caractéristiques que les précédents et la distribution des résidus n’est pas normale.

## Conclusion

On peut voir que les données montrent une tendance à la croissance. Le modèle appliqué montre de bonnes prévisions, il n’y a pas de corrélation entre les résidus, leur moyenne est nulle et la variance essaie de rester constante. Une transformation Box-Cox a été effectuée afin de savoir si les données pouvaient avoir une distribution normale, cependant, cela n’a pas été le cas et la modification des résultats des autres paramètres n’a pas été significative, c’est pourquoi il est décidé de continuer avec le modèle initialement proposé. Un autre modèle pourrait être essayé afin de comparer les résultats obtenus.