Ecole Nationale Supérieure
des Techniques Avancées

Projet séries chronologiques :

STA202

Antoine Germain et Noé Karageorgiou

Évolution temporelle de l'indice du CAC40

Table des matières

1 Introduction

1.1 Sujet d'étude

Dans le cadre de ce projet d'étude d'une série temporelle, nous avons choisi d'étudier l'évolution du nombre de points de l'indice du CAC 40 (Cotation Assistée en Continu 40). Cet indice est le principal indice boursier de la Bourse de Paris, représentant la performance des 40 actions les plus importantes et les plus activement négociées cotées sur Euronext Paris, en les pondérant pas leur capitalisation boursière.

Cet indice est créé le 31 décembre 1987, avec une valeur fixée à 1 000 points à la fin de la publication ce jour-là. L'indice est publié tous les jours ouvrés de 9 heures à 17 h 30 et mis à jour toutes les 15 secondes.

Le CAC 40 est un indicateur de santé économique des plus grandes capitalisations boursières françaises, mais il est calculé sans réinvestissement des dividendes : en prenant en compte cette donnée, l'indice serait à une valeur environ 3 fois plus importante aujourd'hui. C'est un indice, il n'est donc pas échangeable en tant que tel, mais des produits dérivés du CAC 40 existent.

1.2 Base de données

La base de données utilisée provient de ce site. C'est une base de données regroupant, pour chaque jour ouvré entre le 31 décembre 1987 et le 17 mars 2017, la date du jour, le nombre de points respectivement à l'ouverture, au plus bas, au plus haut et à la fermeture de la journée, et le volume échangé. Elle présente donc 6 colonnes et 7389 lignes. Dans la suite du projet, nous considérerons par défaut la valeur de l'indice en fin de journée.

2 Importation et mise en forme de la base de données

2.1 Importation de la base de données

Une fois la base de données téléchargée au format .txt, nous l'avons importée avec la suite de commande suivante :

```
cac <- read_delim("D:/Users/Antoine/Downloads/CAC40/350000.TXT",
delim = "\t", escape_double = FALSE, col_types =
cols(Date = col_date(format = "%d.%m.%Y")), trim_ws = TRUE)
```

Les dates étant au format d.m.Y dans le .txt, elles sont ainsi traduites au format classique de R Y-m-d.

2.2 Mise en forme des données

Notre base de données ne contenant aucun NA, un nettoyage de la base n'est pas nécessaire. Néanmoins, on remarque que la colonne contenant le volume de sous-jacent du CAC 40 échangé par jour, en millions d'euros, est de valeur nulle jusqu'à 2000. Même à partir de 2000, la valeur vaut parfois 0 alors qu'il est évident que si la Bourse de Paris était ouverte ce jour-là, puisque le CAC 40 a été publié, le volume journalier des actions n'était pas nul. Ainsi un traitement spécifique de cette colonne sera nécessaire afin de l'utiliser.

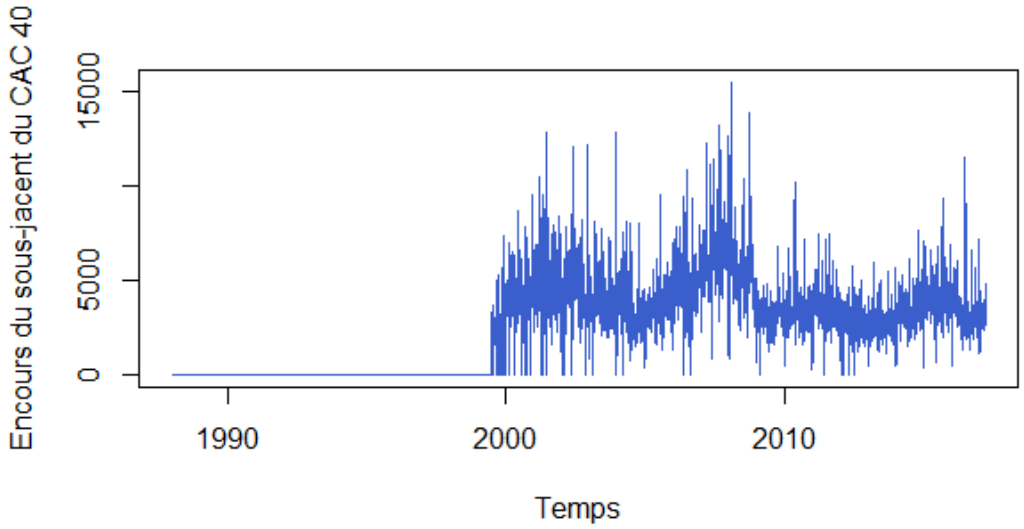


FIGURE 1 – Volume journalier du sous-jacent de l'indice

3 Premiers résultats, série temporelle

3.1 Premiers résultats

Nous pouvons alors visualiser l'entête de la base de données via la fonction *head* et son résumé via la fonction *summary* :

```
> head(cac)
# A tibble: 6 × 6
  Date       Start  High   Low   End  Size
  <date>     <dbl> <dbl> <dbl> <dbl> <dbl>
1 1987-12-31 1007. 1007. 1000  1000      0
2 1988-01-04  985.  985.  985.  985.     0
3 1988-01-05 1017. 1021. 1017. 1021.     0
4 1988-01-06 1032. 1032. 1028. 1028.     0
5 1988-01-07 1025. 1025. 1024. 1024.     0
6 1988-01-08 1030. 1030. 1029. 1029.     0
```

FIGURE 2 – Entête de la base de données

```
> summary(cac)
      Date                Start             High              Low        
 Min.   :1987-12-31   Min.   :   0   Min.   : 900.5   Min.   : 893.8  
 1st Qu.:1995-05-29   1st Qu.:2065   1st Qu.:2074.3   1st Qu.:2051.9  
 Median :2002-09-30   Median :3637   Median :3666.0   Median :3602.3  
 Mean   :2002-09-08   Mean   :3481   Mean   :3507.2   Mean   :3453.7  
 3rd Qu.:2009-12-23   3rd Qu.:4437   3rd Qu.:4467.6   3rd Qu.:4407.6  
 Max.   :2017-03-17   Max.   :6929   Max.   :6944.8   Max.   :6838.7  
      End                Size        
 Min.   : 893.8   Min.   :    0  
 1st Qu.:2064.0   1st Qu.:    0  
 Median :3637.5   Median : 2654  
 Mean   :3481.4   Mean   : 2371  
 3rd Qu.:4439.6   3rd Qu.: 3900  
 Max.   :6922.3   Max.   :15510  
```

FIGURE 3 – Résumé de la base de données

Pour se donner une première idée de l'évolution du CAC40, nous pouvons tracer la valeur de l'indice en fonction du temps. Pour plus de simplicité, nous considérons la valeur de l'indice en fin de journée qui est regroupée dans la colonne "End".

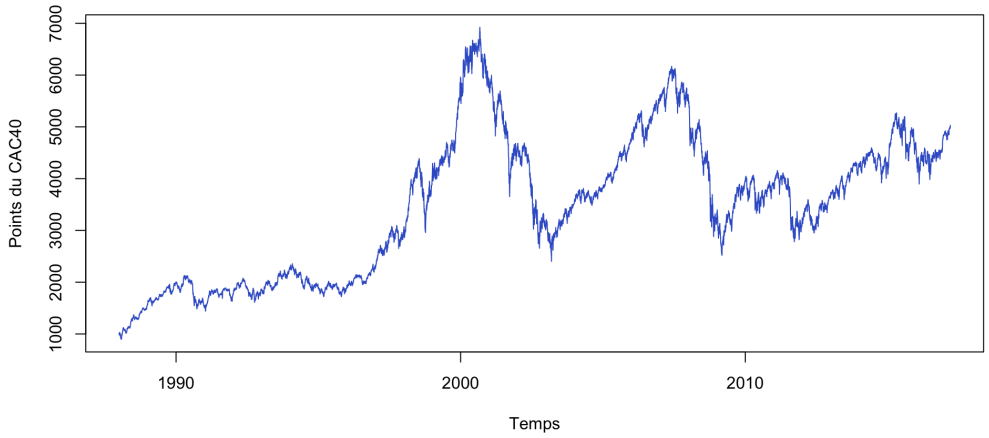


FIGURE 4 – Premier tracé de l'indice en fin de journée

Afin d'avoir une vue plus claire de l'évolution année par année, il est possible d'effectuer l'affichage de la valeur moyenne de l'indice par année :

```
year<-format(cac$Date,"%Y")
mean.year<-tapply(cac$End, as.factor(year),mean)
plot(mean.year, type='b',axes=F,xlab="Temps",ylab="Points du CAC40")
axis(1, c(1:length(mean.year)), labels=names(mean.year))
axis(2)
box()
```

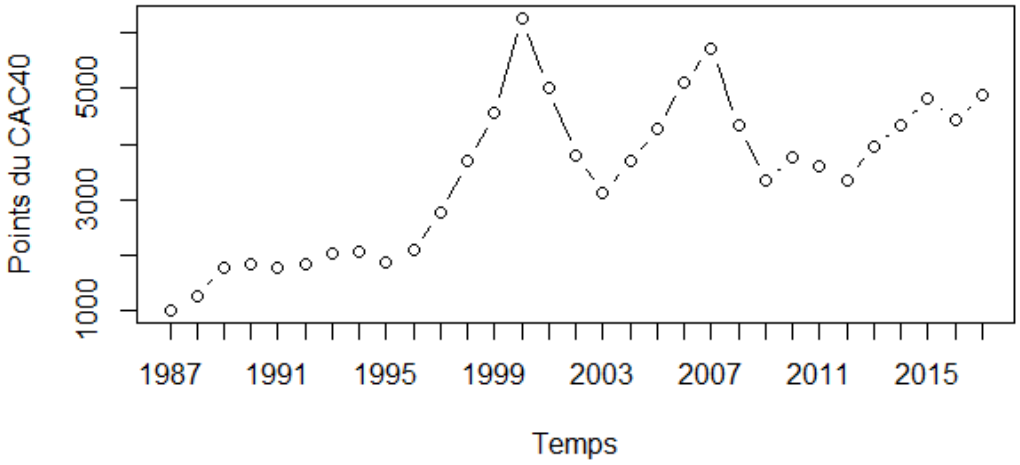


FIGURE 5 – Évolution de l'indice moyenné par an

3.2 Transformation en série temporelle et première analyse

Nous pouvons transformer notre base de données en série temporelle :

```
cac$Date <- as.Date(cac$Date)
cac.xts<-xts(cac[,-1],order.by=cac$Date)
```

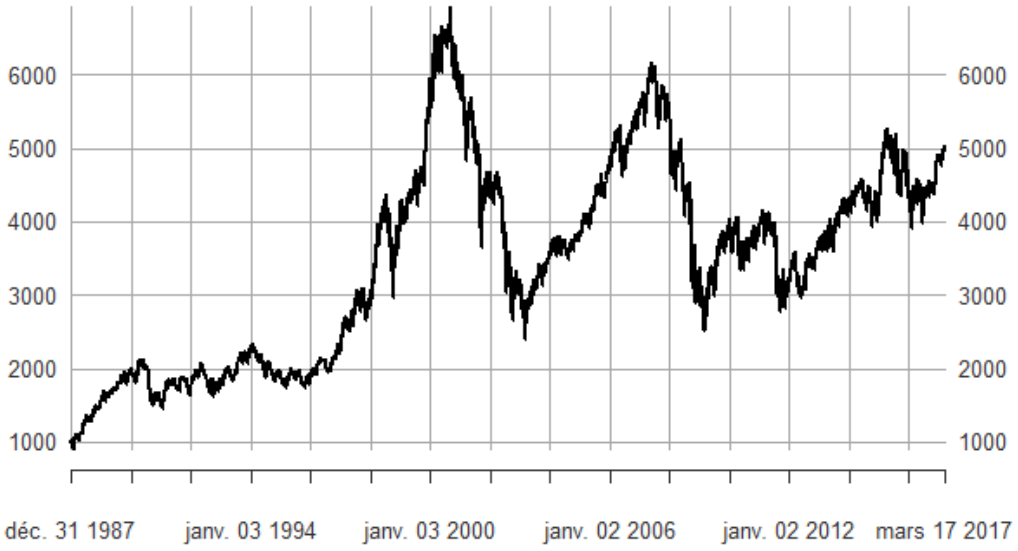


FIGURE 6 – Affichage de la valeur fin de journée de la série temporelle

Nous pouvons alors déjà retrouver sur cette série temporelles des évènements connus de tous :

1. L'entrée de la Bourse de Paris dans Euronext sur lequel est indexé le CAC40, d'où une augmentation considérable de l'investissement du CAC40 en 2000.
2. Le krach boursier de 2001-2022 dû aux faillites de nombreuses entreprises, comme France Telecom, et à l'éclatement de la bulle Internet du début des années 2000.
3. Le krach boursier, dit des subprimes, de 2008 qui a eu ses conséquences en France vers 2009. Cela se traduit par la réticence des investisseurs à investir et donc une forte diminution des investissements pour le CAC40.

On voit une tendance positive se distinguer parmis ces deux pics de forts investissements suivis de krach, montrant que le CAC40 s'est développé par ses investissements multiples qui ont fructifié.

Nous pouvons alors tracer l'autocorrélation en fonction du décalage :

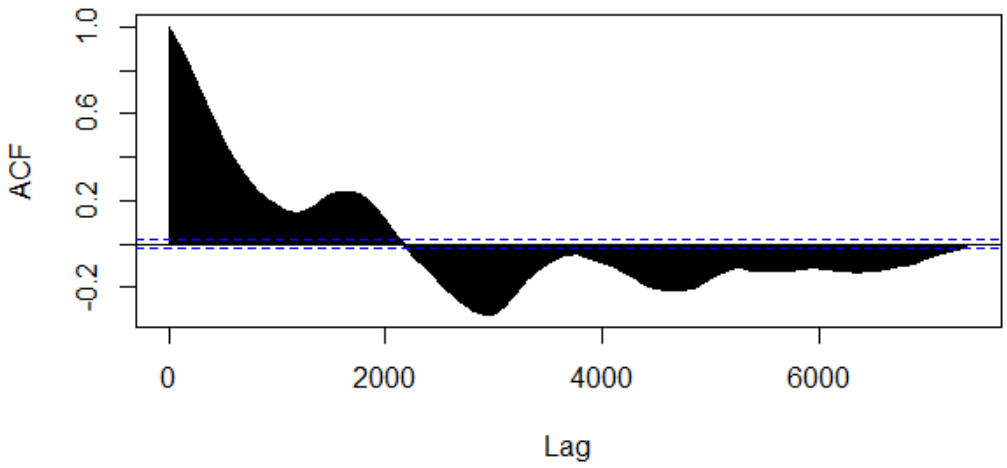


FIGURE 7 – Autocorrélogramme de notre série temporelle

La courbe obtenue est assez prévisible : si l'autocorrélation du CAC 40 avec un décalage temporel faible, inférieur à un an, est largement positive, elle devient beaucoup moins exploitable dès que le décalage devient important. On remarque par exemple un pic de corrélation négative aux environs d'un décalage de 10 ans, ce qui est principalement explicable par la coincidence de la chute de 2009 et la montée de 2000, bien plus que par un réel lien de causalité.

MAX MOINS MIN

4 Modélisation de la série temporelle

Une série temporelle est communément décomposable en trois composantes : la tendance qui correspond à l'évolution à long terme de la série temporelle, la saisonnalité qui correspond à un phénomène périodique de période identifiée, et enfin l'erreur, qui représente la partie aléatoire de la série temporelle.

Nous pouvons utiliser un modèle additif : la série notée Y_t s'écrit alors : $Y_t = T_t + S_t + \epsilon_t$
Nous nous intéressons ici à la modélisation de la composante déterministe de la série : T_t et S_t.

4.1 Estimation de la tendance de la série temporelle

Il existe différents procédés permettant d'estimer la tendance d'une série temporelle. Nous allons dans cette partie les découvrir, puis sélectionner la meilleure méthode.

4.1.1 Estimation par régression linéaire

Voilà le résultat de la régression linéaire de la tendance de la série temporelle :

FIGURE 8 – Approche de la tendance par régression linéaire

(a) Série temporelle et tendance linéaire (b) Série temporelle avec tendance soustraite

Cette estimation de la tendance semble très grossière, il est donc intéressant d'explorer d'autres possibilités.

4.1.2 Estimation par moyenne mobile

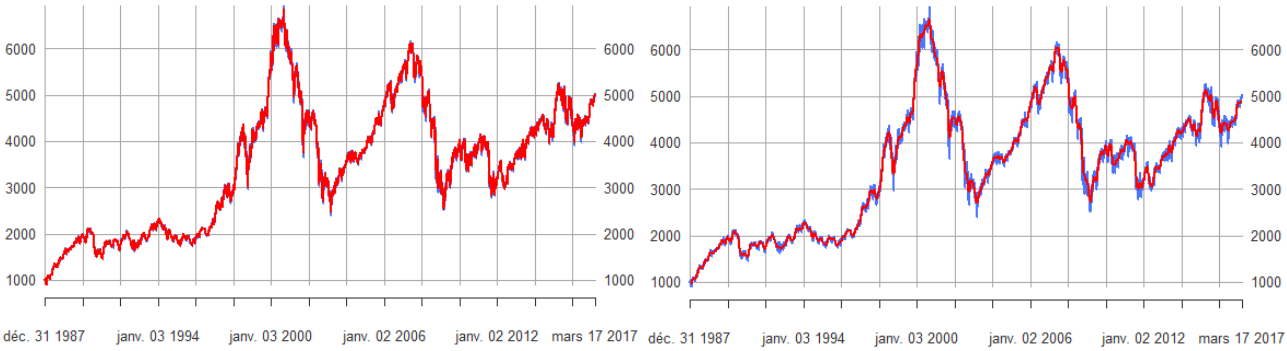
L'estimation par moyenne mobile permet de lisser des fluctuations de la série temporelle en effectuant à un voisinage de la valeur considérée une moyenne. La moyenne mobile se calcule de la sorte :

$$\hat{y}_t = \frac{1}{2l+1} \sum_{i=t-l}^{t+l} y_t$$

où la longueur du voisinage considéré est $2l$.

Nous traçons également 4 estimations de la tendance par moyenne mobile avec 4 fenêtres de temps différentes 9a, 9b, 10a, 10b.

FIGURE 9 – 2 exemples d'approche de la tendance par moyenne mobile



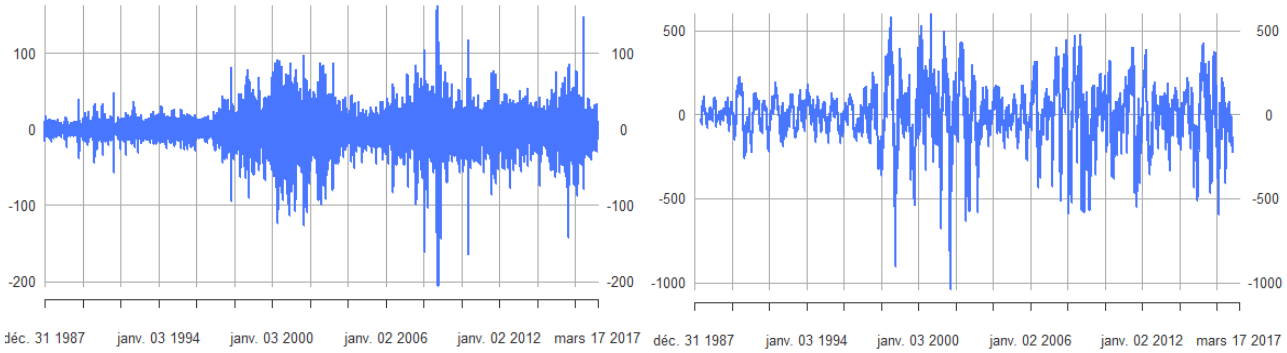
(a) Moyenne mobile avec une fenêtre de 3 jours (b) Moyenne mobile avec une fenêtre de 30 jours

FIGURE 10 – 2 exemples d'approche de la tendance par moyenne mobile

(a) Moyenne mobile avec une fenêtre de 6 mois (b) Moyenne mobile avec une fenêtre de 2 ans

On remarque que plus la fenêtre est petite, plus l'approximation est fidèle.
Nous pouvons alors soustraire la tendance à la série temporelle 11a, 11b.

FIGURE 11 – Série temporelle avec tendance soustraite



(a) Tendance soustraite, 3 jours (b) Tendance soustraite, 6 mois

Nous remarquons que la série temporelle avec tendance soustraite s'apparente à un bruit blanc, d'amplitude d'autant plus faible que la fenêtre de la moyenne mobile est petite.

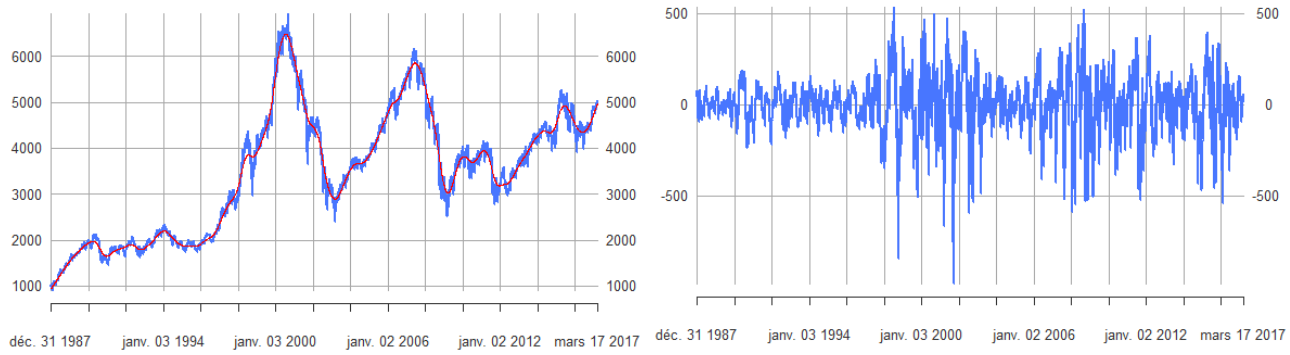
4.1.3 Estimation par noyau gaussien

Une autre méthode estimation de la tendance est l'estimation par noyau gaussien. Au lieu de faire une moyenne sur la fenêtre, on donne plus d'importance aux valeurs proches grâce à une répartition gaussienne centrée sur la valeur considérée. Ainsi l'estimateur à noyaux d'une fonction f est calculée selon :

$$\hat{f}_h(x) = \frac{\sum_{t=1}^{n} y_t K(\frac{x-t}{h})}{\sum_{t=1}^{n} K(\frac{x-t}{h})}$$

où $K = \frac{1}{2\pi} e^{-\frac{x^2}{2}}$ dans le cas d'un noyau gaussien.

FIGURE 12 – Approche de la tendance par noyau gaussien



(a) Série temporelle et tendance par noyau gaussien (b) Série temporelle avec tendance soustraite

Cette méthode d'estimation de la tendance donne des résultats à première vue semblable à ceux obtenus avec la moyenne mobile avec fenêtre de 6 mois.

4.1.4 Estimation par polynômes locaux

Nous procédons ensuite à une estimation de la tendance par polynômes locaux de degré 1. Le principe est que, pour chaque valeur de temps, on estime une fonction polynomiale approximant au mieux les données de la fenêtre. Il s'agit d'estimer une fonction f par l'estimateur :

$$\hat{f}_h(x) = argmin_P \sum_{t=1}^{n} W_t(x) \|y_t - P(x_t - x)\|$$

où P est un polynôme de degré q et $W_t(x) = \frac{K(\frac{x-t}{h})}{\sum_{t=1}^{n} K(\frac{x-t}{h})}$

FIGURE 13 – Approche de la tendance par polynômes locaux

(a) Série temporelle et tendance par polynômes locaux (b) Série temporelle avec tendance soustraite

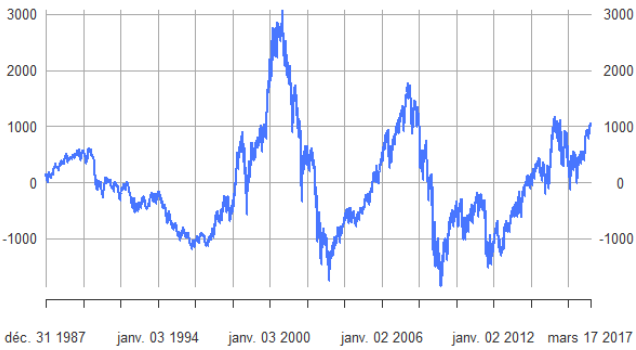
Cette estimation de la tendance indique une tendance quasi constante de l'indice depuis les années 2000.

4.1.5 Estimation par régression sur base de splines

Enfin, nous pouvons essayer une dernière technique vue en cours, l'estimation de la tendance par régression sur base de splines. Il s'agit alors de projeter la fonction f sur une base de fonctions adaptées : par exemple des fonctions splines polynômiales constantes par morceaux.

FIGURE 14 – Approche de la tendance par régression sur base de splines

(a) Série temporelle et tendance par régression sur base de splines



(b) Série temporelle avec tendance soustraite

Cette dernière approche aboutit à une tendance décroissante sur ces denières années, contrairement aux autres estimations jusqu'à présent.

4.1.6 Analyse des différentes estimations de la tendance

Nous pouvons comparer les estimations de la tendance en affichant toutes les tendances estimées, avec une fenêtre de moyenne mobile de 2 ans :

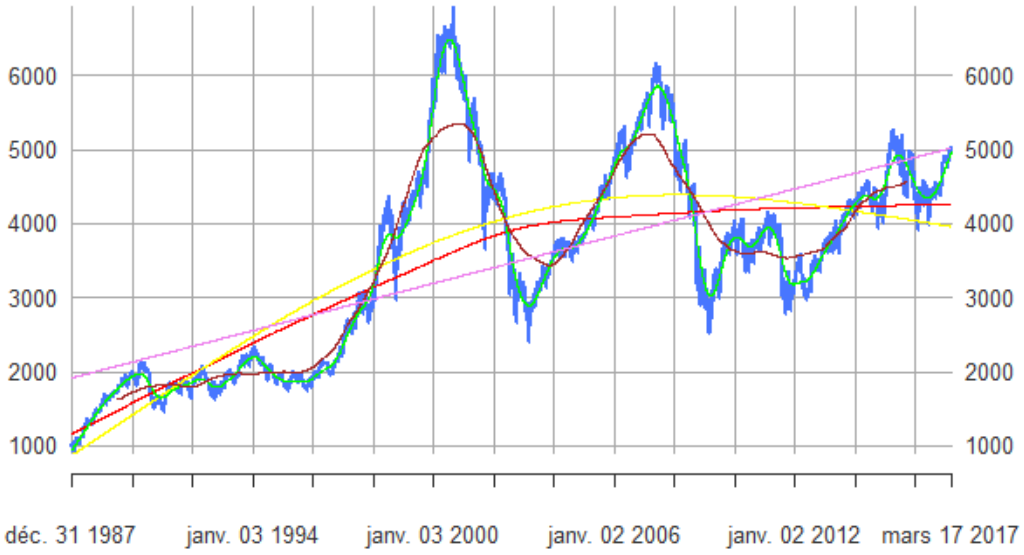


FIGURE 15 – Superposition de toutes les estimations et de la série temporelle

Nous pouvons maintenant déterminer l'estimation la plus fidèle de la tendance de notre série, en calculant l'écart de la moyenne de chaque série soustraite de sa tendance estimée à zéro :

```
moyerr <- (cac.xts$End-linear$fitted)/7389
moyerr<-sum(moyerr)
```

Ou, pour la moyenne mobile sur 3 jours :

```
moyerr <- (cac.xts$End-mobile)/7389
moyerr<-sum(moyerr[2,-2])
```

L'approche par régression linéaire donne un écart en valeur absolu de $1.91 * 10-13$. L'approximation de la tendance par la moyenne mobile sur 3 jours génère une différence $2,34*10^{-3}$ et celle sur 6 mois $-3,63*10^{-3}$. La méthode par le noyau gaussien donne un écart de $1.35*10-12$, tandis que celle des polynômes locaux aboutit à une différence de 88.0. Enfin, l'estimation par la régression sur base de splines engendre un écart absolu de $9,24 * 10^{-12}$.

La méthode d'estimation de la tendance la plus efficace est donc la régression linéaire. La moyenne mobile, qui est pourtant très proche de la courbe, n'estime pas efficacement sa tendance.

4.2 Estimation de la saisonnalité de la série temporelle

Notre série temporelle n'étant pas décomposable par la fonction decompose de R, elle ne présente a priori aucune saisonnalité : $S_t = 0$.

4.3 Estimation des résidus de la série temporelle

4.3.1 Tracé des résidus

La série étant sans saisonnalité, $\epsilon_t = Y_t - T_t$. Ainsi, nous pouvons tracer les résidus en fonction du temps en utilisant la meilleure estimation de la tendance, la régression linéaire :

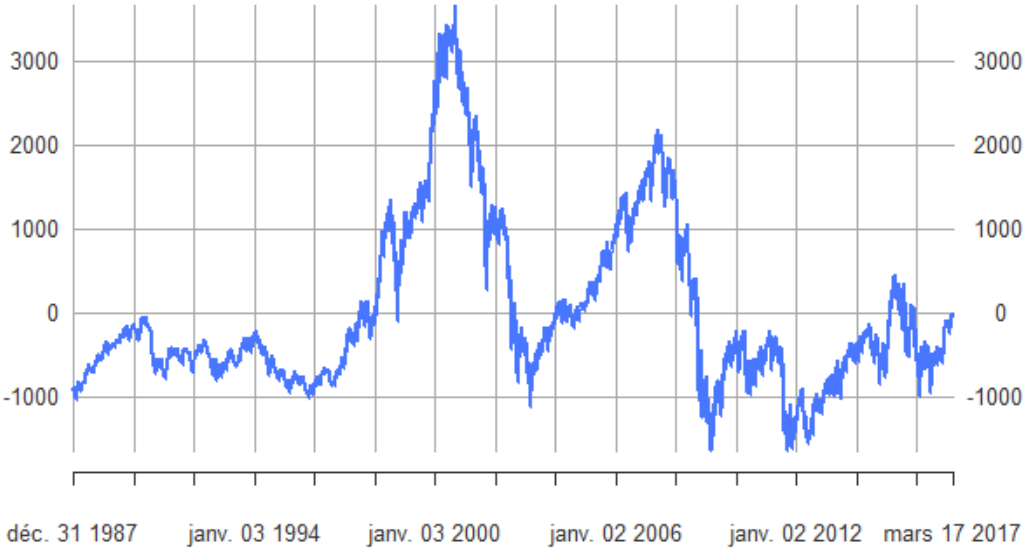


FIGURE 16 – Résidus avec estimation de la tendance par régression linéaire

4.3.2 Autocorrélogramme des résidus

Nous pouvons alors tracer l'autocorrélogramme des résidus afin de déterminer si la série des résidus est stationnaire.

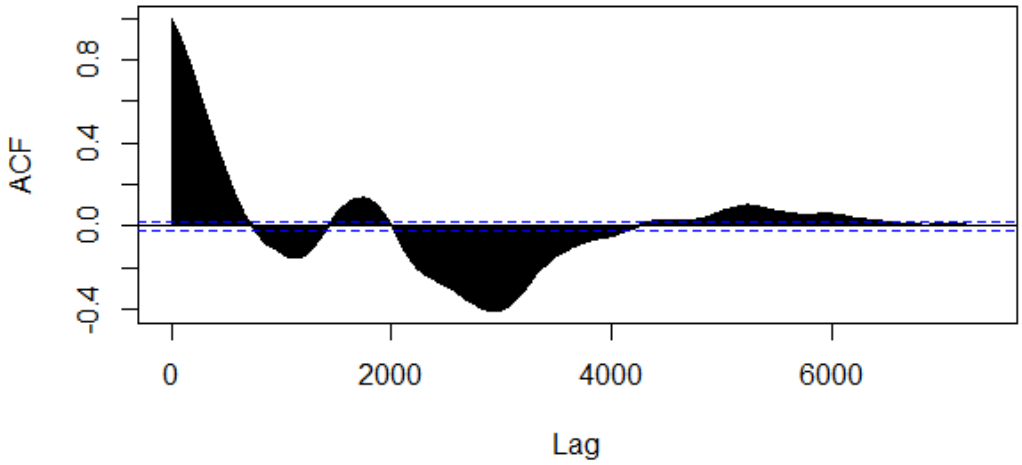


FIGURE 17 – Autocorrélogramme de la série des résidus

Nous en concluons que les résidus ne peuvent pas être considérés comme purement stationnaires, l'autocorrélation ne tendant pas vers 0 très rapidement.

4.3.3 Caractère normal de la série des résidus

En tracant un histogramme de la série des résidus, nous observons que le caractère normal de la série des résidus n'est pas clairement établi :

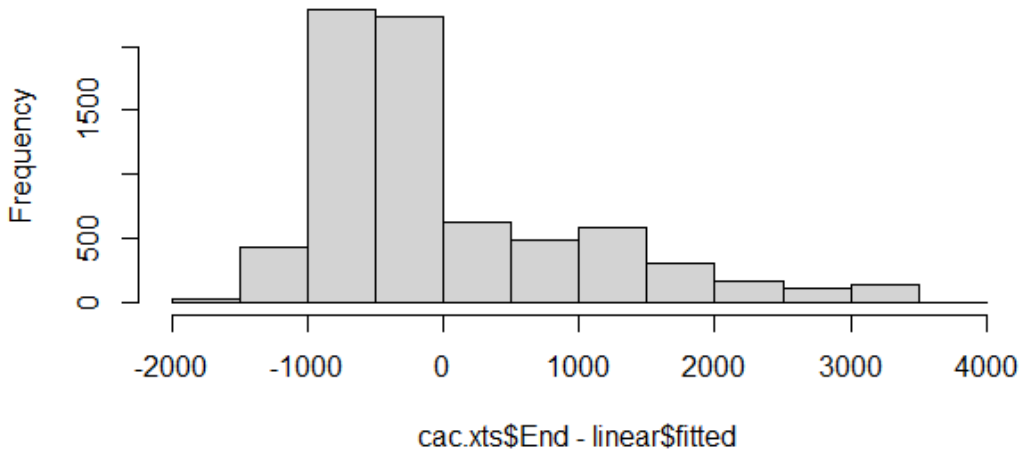


FIGURE 18 – Histogramme de la série des résidus

5 Lissage exponentiel

L'objectif de cette partie est d'appliquer les différents lissages exponentiels vus en cours à notre base de données afin d'obtenir des prévisions sur le comportement des données. Pour garantir une meilleure lisibilité lors des tracés, nous avons seulement utilisé la base de données entre les années 2014 et 2017 pour les prévisions.

5.1 Lissage exponentiel simple

Le lissage exponentiel simple permet de prédire le comportement de la série temporelle à l'instant suivant. Si notre série temporelle est notée y_t, le lissage exponentiel simple de

paramètre $\alpha \in [0, 1]$ de cette série est noté $\hat{y}_t$ et est définit de telle sorte :

$$\hat{y}_{t+1/t} = \alpha y_t + (1 - \alpha)\hat{y}_{t/t-1}$$

d'où

$$\hat{y}_{t+1/t} = \sum_{t=0}^{t-1} \alpha(1 - \alpha)^i y_{t-i}$$

La valeur du paramètre α permet de plus ou moins prendre en compte les valeurs récentes : si α est proche de 1, ce sont les valeurs récentes qui influent majoritairement sur la prévision tandis que quand α est proche de 0, la prévision prend en compte des valeurs plus anciennes.

Nous avons en premier lieu appliqué la méthode du lissage exponentiel simple pour le paramètre α=0.05, nous souhaitons donc prendre en compte les valeurs plus anciennes.

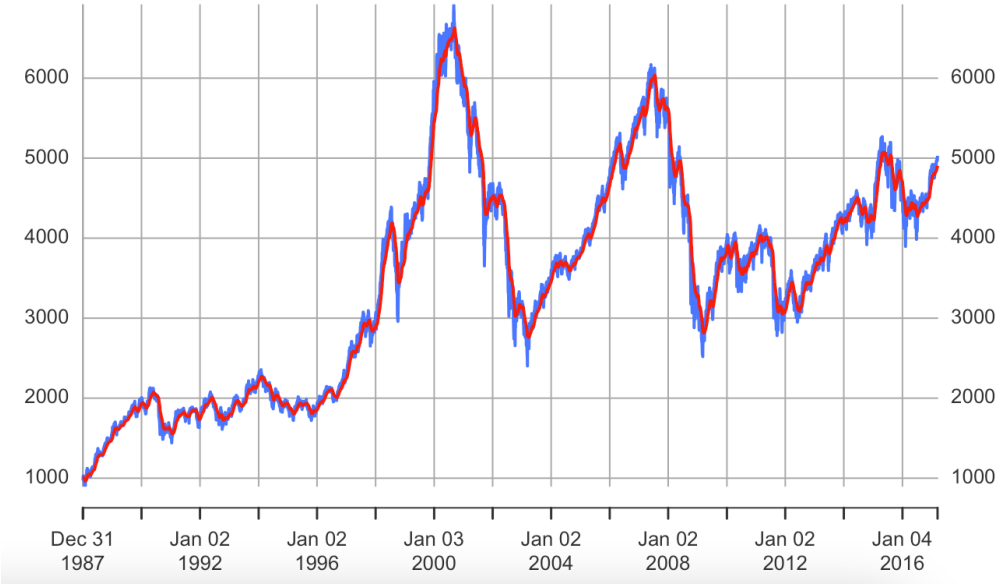


FIGURE 19 – Lissage Exponentiel Simple où α=0.05

Comparons l'efficacité de ce lissage avec un α proche de 1 :

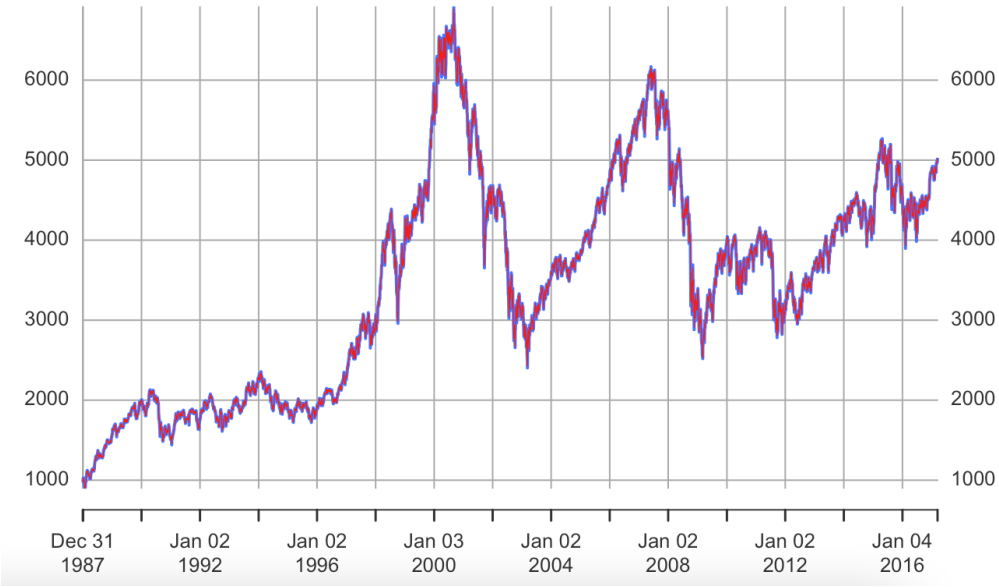


FIGURE 20 – Lissage Exponentiel Simple où α=0.8

Nous remarquons que le lissage exponentiel avec α=0.8 est plus précis que celui avec α=0.05. Cela est cohérent avec le mode de fonctionnement d'une bourse : les investissements sont faits en fonction de données récentes majoritairement.

5.1.1 Prévision

En appliquant la méthode, nous obtenons le tracé suivant pour la prévision pour l'année 2017 :

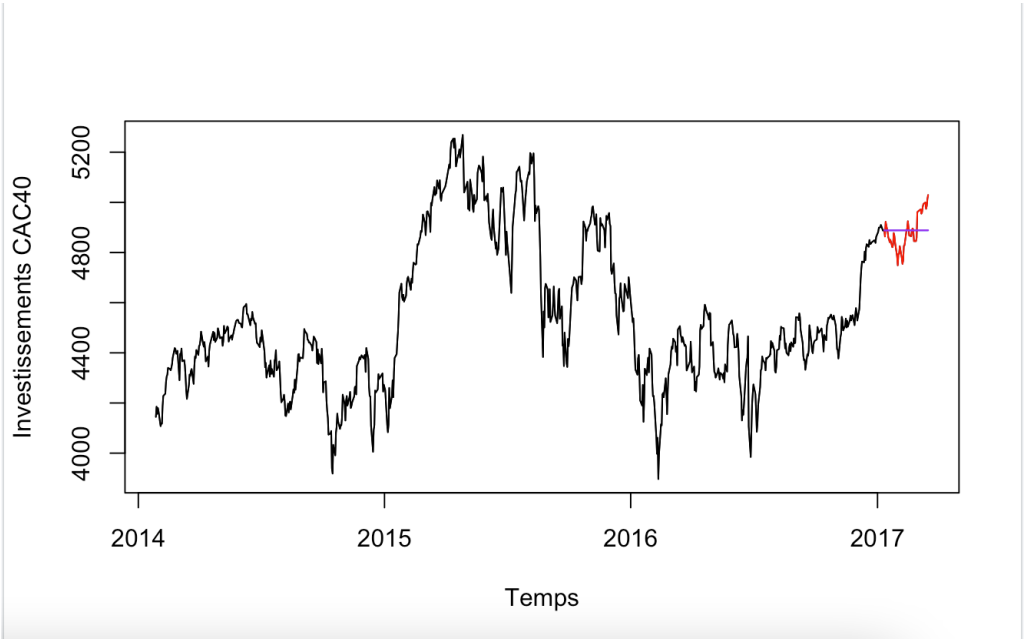


FIGURE 21 – Prévision sur l'année 2017 avec le lissage exponentiel simple

Nous remarquons qu'une telle méthode n'est pas adaptée pour la prévision d'une série temporelle puisqu'elle prédit une valeur constante à la dernière mesurée.

5.2 Lissage exponentiel double

Le principe du lissage exponentiel double est non pas d'ajuster la série temporelle par une constante mais par une droite. Cela se traduit par les équations suivantes :

$$\hat{y}_{t+1/t} = l_t + b_t$$

où $\begin{cases} l_t = l_{t-1} + b_{t-1} + (1-(1-\alpha)^2)(y_t - \hat{y}_{t/t-1}) \\ b_t = b_{t-1} + \alpha^2(y_t - \hat{y}_{t/t-1} \end{cases}$

La première étape est de chercher le α optimal pour le lissage double. Cela se fait en minimisant l'erreur entre les valeurs réelles et celles obtenues par lissage. On retrouve un α proche de 1.

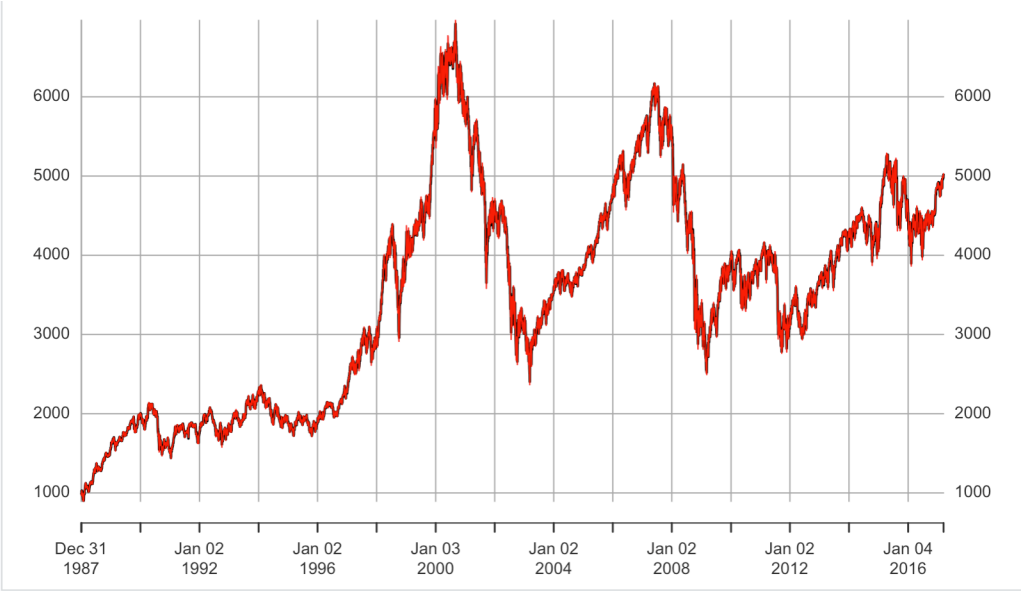


FIGURE 22 – Lissage Exponentiel Double pour α optimal

Nous remarquons que le lissage double présente tout de même des imprecisions, notamment en terme d'amplitude.

5.2.1 Prévision

Nous pouvons également faire une prévision de nos données avec un lissage exponentiel double. Etant difficle d'effectuer une prévision sur un horizon sur un laps de temps particulier et de la rendre "esthétique" lors du tracé, nous avons choisi de ne considérer que les années 2014 à 2017 ce qui représente environ 800 valeurs.

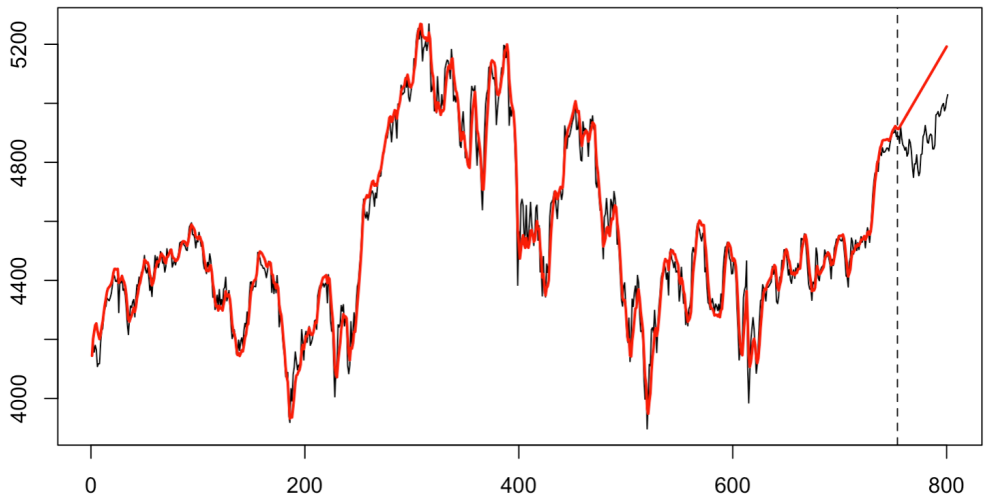


FIGURE 23 – Prévision sur l'année 2017 par un lissage exponentiel double

Encore une fois la prévision n'est pas fidèle aux valeurs réelles mais cela était encore une fois prévisible puisque le lissage exponentiel double approxime les valeurs par des droites et donc la prévision est une droite de coefficient directeur celui de la dernière valeur mesurée.

5.3 Prévision par la méthode de Holt-Winters

Nous avons efin utilisé le lissage d'Holt-Winters qui permet entre autres une prévision des données. Cette méthode se base sur les équations suivantes :

$$\hat{y}_{t+1/t} = (l_t + b_t)s_t$$

où $\begin{cases} l_t = \alpha \frac{y_t}{s_{t-T}} + (1-\alpha)(l_{t-1} + b_{t-1}) \\ b_t = \beta(l_t + l_{t-1}) + (1-\beta)l_{t-1} \\ s_t = \delta \frac{y_t}{l_t} + (1-\delta)s_{t-T} \end{cases}$

Dans R, une fonctions HoltWinters est déjà implémentée et ajuste automatiquement les paramètres. Nous obtenons alors comme prévision pour l'année 2017 :

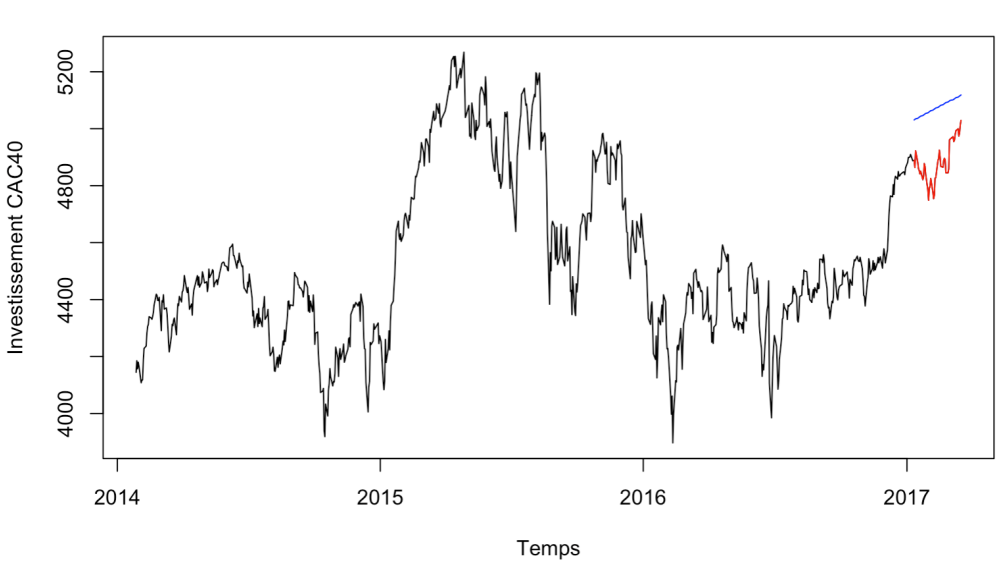


FIGURE 24 – Code conduisant au tracé précédent

La prévision obtenue est totalement absurde : la prédiction ne suit en aucun cas la courbe réelle et même la première valeur ne correspond pas à celle réelle. Dans le code fournit ci-joint, on s'intéresse à la période 2014-2017 correspondant approximativement aux lignes 6589-7389 (seulement 46 dates sont renseignées pour 2017). Cela peut être du à la faible saisonnalité puique la méthode de Holt-Winters la prend en compte.

```
## Holt-Winters
cac$Date <- as.Date(cac$Date)
cac.xts<-xts(cac[,-1],order.by=cac$Date)

cac_0.xts<-cac.xts$End[6589:7342]
cac_1.xts<-cac.xts$End[7343:7389]

hw<-HoltWinters(cac.xts$End, gamma=F)
hw.forecast<-predict(hw, n.ahead=length(cac_1.xts))
plot(cac$Date[6589:7389],c(cac_0.xts,cac_1.xts),type='l',
     xlab="Temps", ylab="Investissement CAC40",)
lines(cac$Date[7343:7389],cac_1.xts,col='red')
lines(cac$Date[7343:7389],  hw.forecast, col='blue')
```

FIGURE 25 – Prévision sur l'année 2017 par un lissage exponentiel double

6 Conclusion

Ce projet nous a permis de mettre en application les différents procédés vus en cours et appliqués à une base de donnée réelle. Il s'agissait également de s'adapter à une base de donnée et d'en interpréter les différentes spécificités. Notre base de donnée n'était pas simple puisqu'il n'y a pas de saisonnalité évidente mais nous avons tout de même pu l'exploiter et relier les informations que nous connaissions avec celles induites par l'étude statistique de la série temporelle. De plus, nous avons pu essayer les différentes méthodes de prévision sur notre base de donnée. C'est en lien direct avec la finance en général puisque les bons investissements se font selon la prévision de la valeur des actions.

Afin d'exploiter au maximum les données présentes dans la base de données, nous aurions pu aussi étudier les différences d'investissement entre le début et la fin de la journée.