

Météo Londres

Samuel BALLU Paul LE BRETON Timothée TEMPLIER

2024-04-04

Introduction

Importation des packages

L'analyse des séries temporelles est cruciale pour comprendre et prévoir les phénomènes météorologiques. Dans ce rapport, nous étudions les températures moyennes enregistrées à Londres en utilisant divers outils statistiques et de modélisation. Nous appliquons des techniques de décomposition des séries temporelles pour identifier les tendances et les composantes saisonnières, suivies de la modélisation ARIMA pour capturer les dépendances temporelles. Notre objectif est de comprendre les tendances et les fluctuations des températures moyennes à Londres.

Chargement du jeu de données

```
df <- read.csv("london_weather.csv")

df <- df |> select(date, mean_temp)

# Suppression des valeurs manquantes
df <- na.omit(df) # il y en avait 36

head(df)

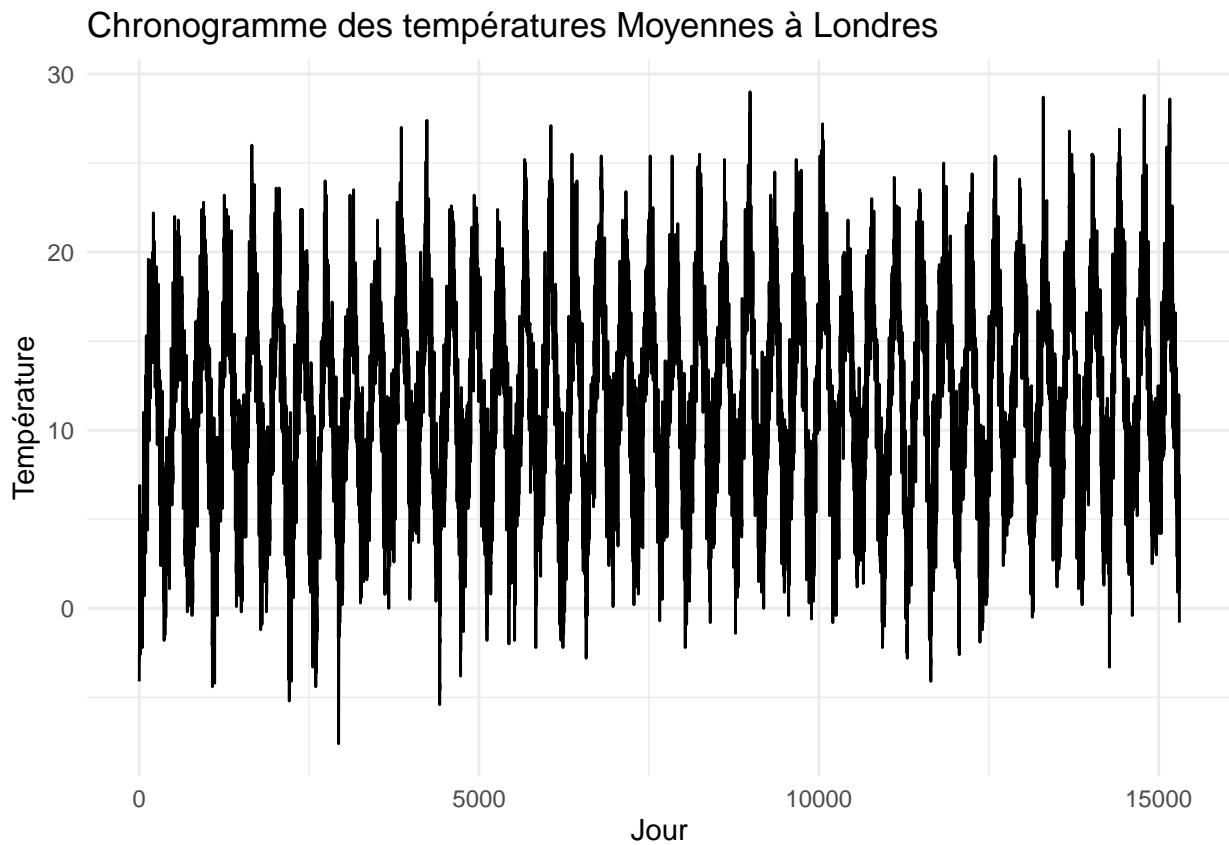
##          date mean_temp
## 1 19790101     -4.1
## 2 19790102     -2.6
## 3 19790103     -2.8
## 4 19790104     -2.6
## 5 19790105     -0.8
## 6 19790106     -0.5
```

Nos données sont composées de deux colonnes : la date et la température moyenne. Nous avons supprimé les valeurs manquantes pour éviter tout problème lors de l'analyse. La température est la température moyenne à Londres en degrés celsius. Les données sont enregistrées quotidiennement, ce qui nous permet d'analyser les tendances et les fluctuations des températures sur une base quotidienne. Les données couvrent une période de 42 ans, de 1979 à 2021.

1 Analyse

1.1 Visualisation générale

```
ggplot(df, aes(x = 1:nrow(df), y = mean_temp)) + geom_line() + labs(title = "Chronogramme des températures Moyennes à Londres")
```



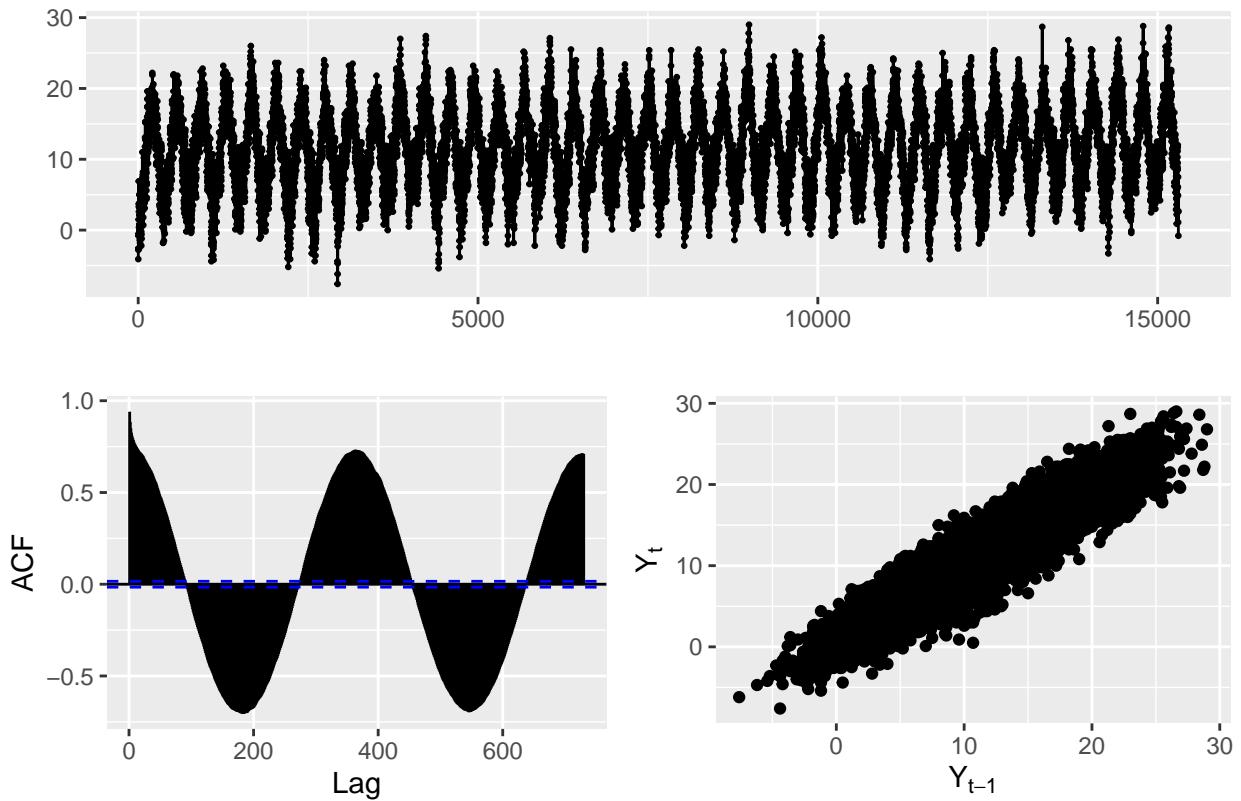
A l'aide de ce premier graphique, il est facile de voir que les températures moyennes augmentent durant les mois d'été, et diminuent l'hiver, formant de grands pics annuels.

1.2 Série, ACF et PACF

La saisonnalité est la variation périodique et régulière des données sur une période donnée, tandis que la tendance globale est le comportement sous-jacent des données sur la période.

Notre premier objectif va être de retirer ces 2 composantes, dans l'objectif d'isoler les composantes résiduelles des données, facilitant ainsi l'analyse des variations irrégulières et des anomalies. Cela aide à mieux comprendre la structure sous-jacente de la série temporelle et à améliorer la précision des prévisions futures.

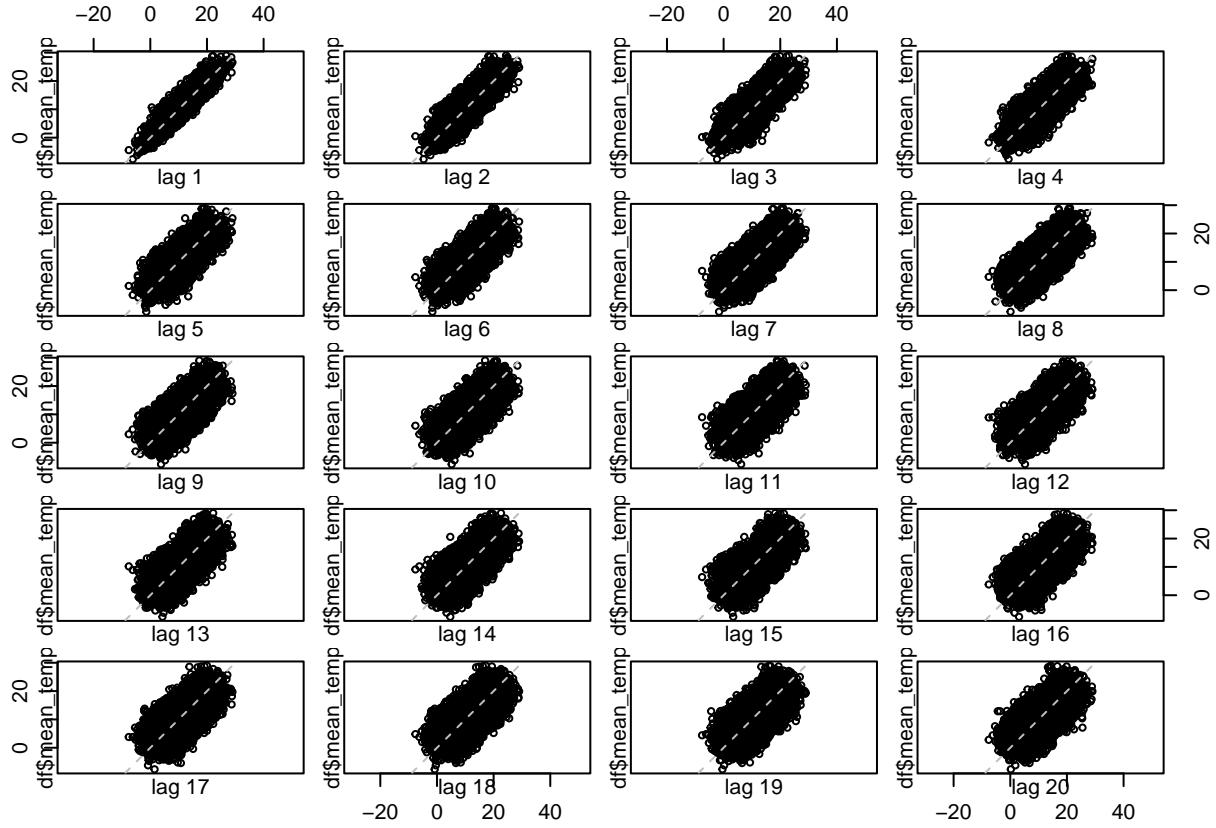
Résumé statistique de la série temporelle



Le graphique ci-dessus nous donne une distribution des valeurs des températures, ainsi que des autocorrélations et des corrélations partielles de notre série temporelle, avec un nombre maximum de décalages de 730 jours, aidant à identifier les schémas de dépendance temporelle. Comme prévu, ici sur l'ACF on remarque bien une saisonnalité avec un rythme et des pics tous les 365 jours, on va donc retirer celle-ci dans la suite.

1.3 Lag-plots

```
lag.plot(df$mean_temp, lags = 20)
```



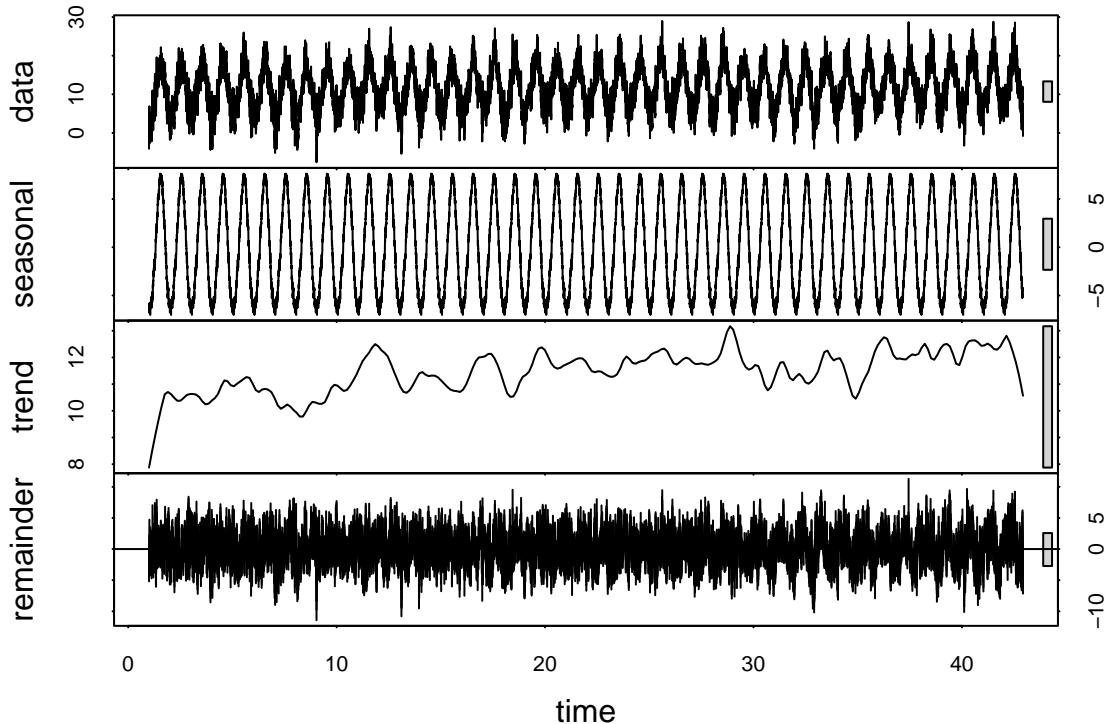
Le graphique ci-dessus montre les corrélations entre les données actuelles et ses valeurs décalées jusqu'à 20 jours en arrière, révélant des motifs de dépendance temporelle.

Un décalage de moins en moins corrélé sur les lag-plots signifie que les températures d'aujourd'hui sont de moins en moins liées à celles d'il y a 20 jours, ce qui pourrait indiquer des changements météorologiques ou saisonniers. Si cette tendance continue pour des décalages plus longs, cela peut suggérer des fluctuations aléatoires ou des changements climatiques sur une plus longue période.

On ne remarque plus trop de changement à partir du lag numéro 3 ou 4. Nous pouvons donc supposer que l'une de ces valeurs serait la période de saisonnalité. On ne doit pas prendre un lag trop élevé pour ne pas faire augmenter la variance.

2 Analyse de la saisonnalité et de la tendance

```
ts_temp <- ts(df$mean_temp, frequency = 365)
decomposed <- stl(ts_temp, s.window = "periodic")
plot(decomposed)
```



Le graphique obtenu nous donne des informations sur les 3 composantes principales suivantes : la tendance, la saisonnalité et les résidus.

Saisonnalité : La composante de saisonnalité nous confirme bien des fluctuations régulières et récurrentes des données tous les ans.

Tendance : La tendance nous montre comment les données évoluent à long terme. Ici, on peut voir que celle-ci est ascendante, cela indique une légère augmentation générale des températures au fil du temps (ce qui est sûrement du au changement climatique).

Résidus : Ici, les résidus semblent être aléatoires et distribués de manière uniforme autour de 0, indiquant qu'il n'y a pas de modèle sous-jacent non capturé.

2.1 Retrait de la saisonnalité

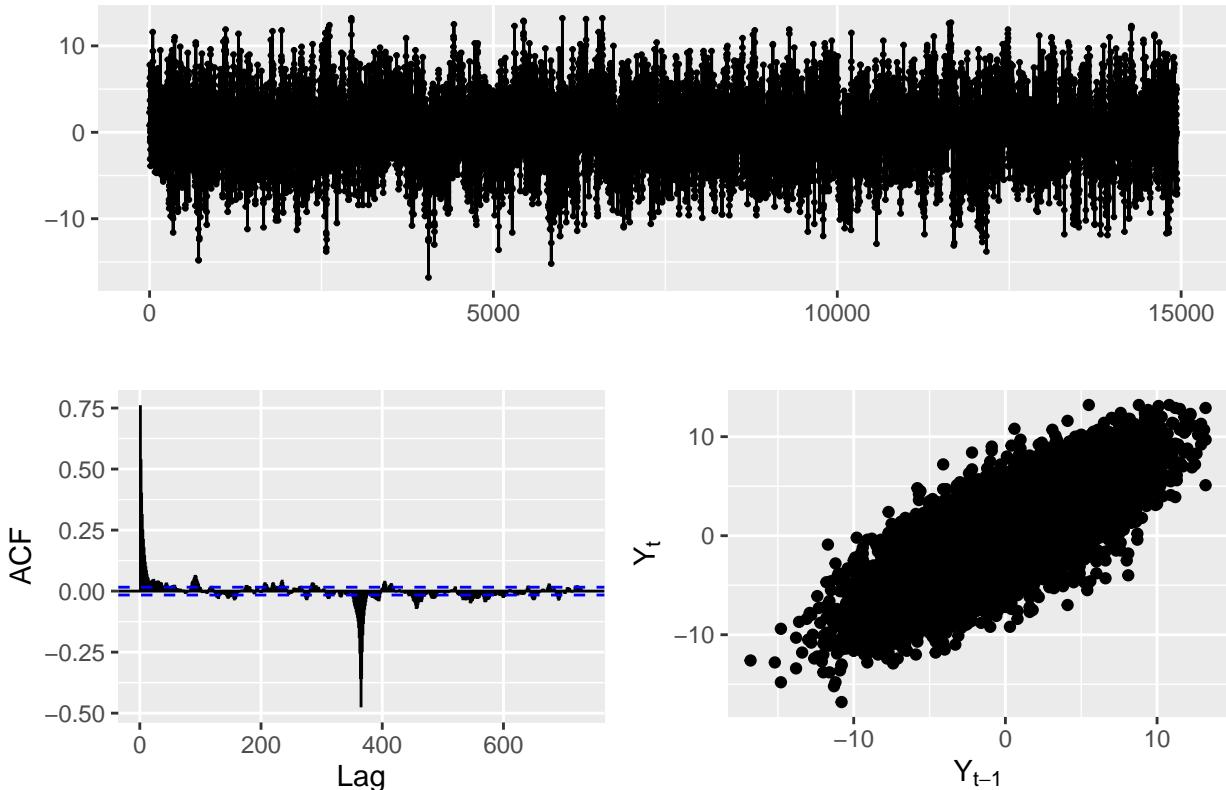
Comme nous nous trouvons dans un cas de saisonnalité annuelle, nous allons retirer cette saisonnalité en utilisant une différence de lag 365.25 (pour tenir compte des années bissextilles).

```

serie_lissee <- df$mean_temp |> diff(lag = 365.25)

ggtsdisplay(serie_lissee, plot.type = "scatter", lag.max = 730)

```



Le graphique des autocorrélations de la série lissée montre de fortes valeurs initiales, indiquant une dépendance à court terme même après avoir retiré la saisonnalité annuelle. Ensuite, les autocorrélations tournent autour de zéro, suggérant que la saisonnalité a été efficacement enlevée.

Donc bien que la saisonnalité annuelle ait été retiré, il reste une dépendance à court terme à considérer pour la suite.

Le nuage de points confirme ceci avec des points qui semblent assez alignés et non aléatoires, on en déduit une auto-corrélation des résidus.

Au vu de notre nouveau chronogramme nous pouvons confirmer que la saisonnalité a été efficacement supprimée, grâce aux variations saisonnières qui ne sont plus visible dans la série lissée.

Test de Dickey-Fuller augmenté

```

adf.test(serie_lissee)

## Warning in adf.test(serie_lissee): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
##  data:  serie_lissee

```

```
## Dickey-Fuller = -20.103, Lag order = 24, p-value = 0.01
## alternative hypothesis: stationary
```

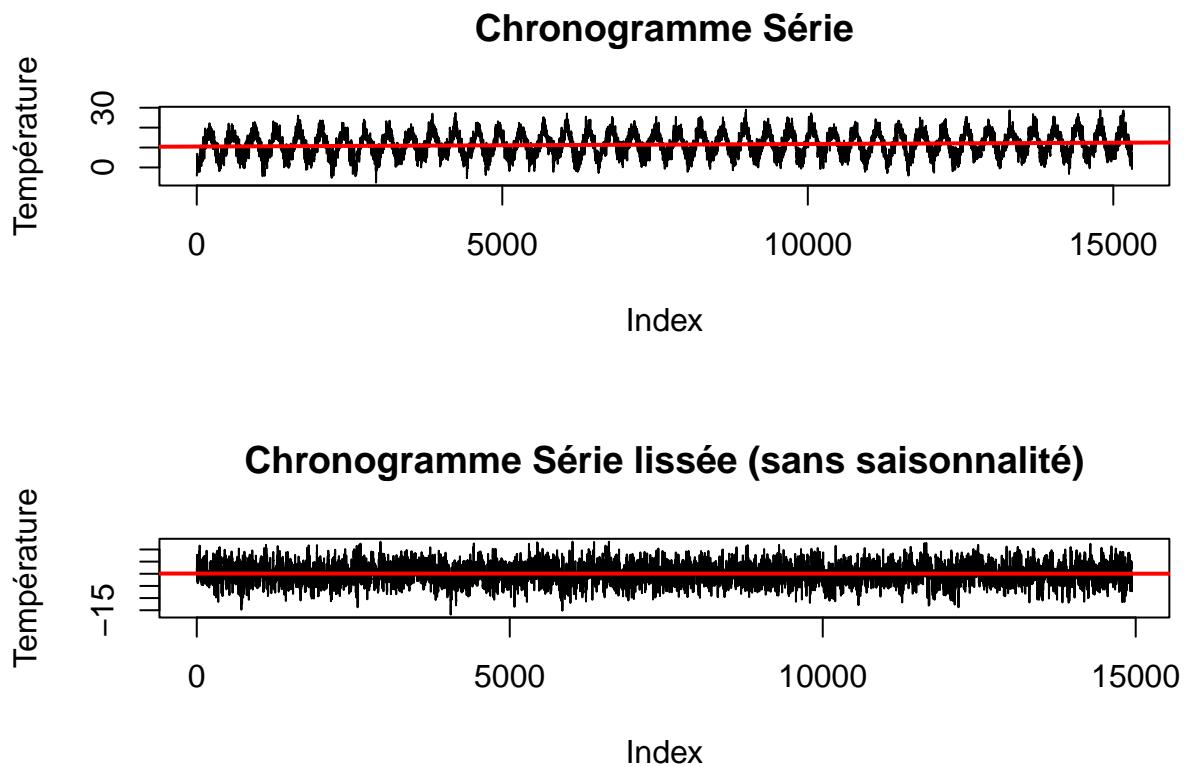
Le test de Dickey-Fuller augmenté sur les résidus de la série lissée indique une p-value de 0.01, ce qui est inférieur au seuil de significativité communément utilisé de 0.05, suggérant que les résidus sont stationnaires et qu'il n'y a pas de saisonnalité significative dans la série lissée.

2.2 Retrait de la tendance

```
par(mfrow = c(2, 1))
plot(df$mean_temp, type = "l", main = "Chronogramme Série", xlab = "Index", ylab = "Température")
abline(lm(df$mean_temp ~ seq_along(df$mean_temp)), col = "red", lwd = 2)

df_serie_lisse <- data.frame(serie_lisse)

plot(df_serie_lisse$serie_lisse, type = "l", main = "Chronogramme Série lissée (sans saisonnalité)", x
abline(lm(df_serie_lisse$serie_lisse ~ seq_along(df_serie_lisse$serie_lisse)), col = "red", lwd = 2)
```



En comparant ces 2 graphiques, on peut voir que la tendance globale semble avoir été supprimée de la série, en même temps que la saisonnalité. Cependant, même si une tendance claire n'est pas visible, nous allons vérifier qu'elle n'est pas présente dans la suite.

2.3 Régression linéaire

```
modele <- lm(serie_lissee ~ seq_along(serie_lissee))

summary(modele)

##
## Call:
## lm(formula = serie_lissee ~ seq_along(serie_lissee))
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -16.8770 -2.6779 -0.0453  2.7239 13.1325 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 9.220e-02 6.604e-02 1.396   0.163    
## seq_along(serie_lissee) -3.752e-06 7.656e-06 -0.490   0.624  
## 
## Residual standard error: 4.036 on 14938 degrees of freedom
## Multiple R-squared:  1.608e-05, Adjusted R-squared:  -5.086e-05 
## F-statistic: 0.2402 on 1 and 14938 DF, p-value: 0.624
```

Le coefficient associé à la variable temporelle n'est pas significativement différent de zéro (p-value = 0.624), indiquant qu'il n'y a pas de tendance linéaire discernable dans vos données. En résumé, notre série temporelle semble être stable sans direction claire de croissance ou de décroissance au fil du temps.

3 Stationnarité

3.1 La stationnarité

La stationnarité est une propriété d'une série temporelle. Ses statistiques, comme la moyenne et la variance, ne changent pas au fil du temps. Si notre série suit un processus stationnaire, on pourra donc affirmer que sa structure reste la même, elle n'évolue pas avec le temps. La stationnarité est essentielle pour appliquer de nombreux modèles prédictifs de manière fiable. Nous allons tester la stationnarité de notre série pour voir si nous avons besoin d'effectuer des modifications sur celle-ci.

Test de Dickey-Fuller

```
kpss.test(df_serie_lisse$serie_lissee)

## Warning in kpss.test(df_serie_lisse$serie_lissee): p-value greater than printed
## p-value

##
## KPSS Test for Level Stationarity
##
## data: df_serie_lisse$serie_lissee
## KPSS Level = 0.037308, Truncation lag parameter = 13, p-value = 0.1
```

Le test augmenté de Dickey-Fuller (ADF) montre que l'on rejette l'hypothèse nulle, ce qui suggère que la série est stationnaire.

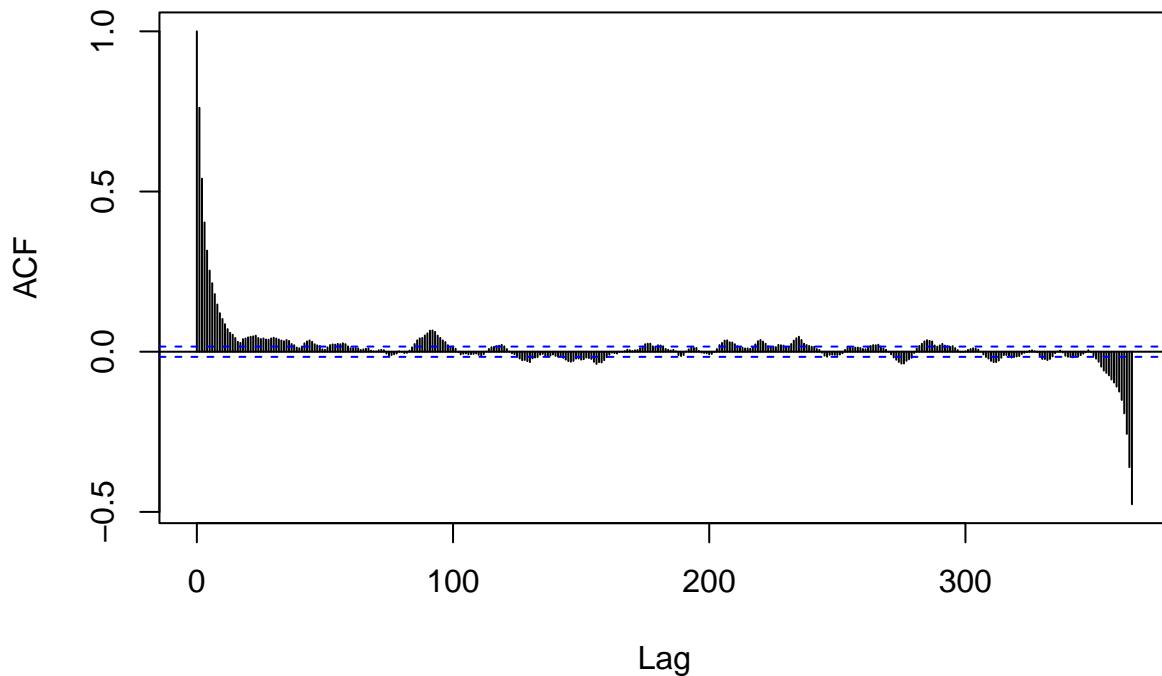
C'est une bonne nouvelle car nous n'aurons pas besoin de différencier la série pour poursuivre notre analyse. En effet, une série stationnaire indique que ses caractéristiques statistiques restent constantes au fil du temps, renforçant la fiabilité des analyses et des prévisions.

3.2 Bruit blanc

Un bruit blanc est une série temporelle aléatoire où chaque valeur est indépendante des autres et a une moyenne et une variance constantes, sans structure discernable. Nous allons maintenant étudier notre série pour voir si celle-ci est un bruit blanc ou non, pour cela, commençons par regarder le graphique de l'ACF.

```
acf(df_serie_lisse$serie_lissee, lag.max = 365, main="ACF série lissée")
```

ACF série lissée



D'après le graphique ACF, il est possible que la série soit un bruit blanc, mais nous ne pouvons pas conclure de manière définitive ici et nous allons devoir analyser les résultat des tests. Pour essayer de clarifier nos hypothèses, nous allons réaliser un test de Box-Pierce

```
Box.test(df_serie_lisse$serie_lisse)
```

```
##  
##  Box-Pierce test  
##  
##  data:  df_serie_lisse$serie_lisse  
##  X-squared = 8661.3, df = 1, p-value < 2.2e-16
```

Le test de Box-Pierce montre une statistique de test élevée et une p-value faible. Cela suggère que la série ne soit pas un bruit blanc. Mais ce n'est pas grave car nous avons déjà vu que la série est stationnaire et que nous pouvons continuer notre analyse.

4 Modélisation de la série

Pour modéliser notre série, nous allons utiliser le modèle ARIMA car il peut capturer efficacement les structures de dépendance dans les données via ses composantes autoregressives, intégrée et moyenne mobile. Il est donc essentiel pour analyser et prévoir de manière fiable.

4.1 Choix du modèle ARIMA

Un modèle ARIMA est un modèle de série temporelle qui combine des composantes autorégressives (AR), des différences intégrées (I) et des moyennes mobiles (MA). On va l'utiliser pour capturer des motifs et des tendances dans nos données, dans l'objectif de créer une prévision.

```
fit <- auto.arima(serie_lisse)
summary(fit)

## Series: serie_lisse
## ARIMA(3,0,1) with zero mean
##
## Coefficients:
##          ar1      ar2      ar3      ma1
##        1.4186 -0.6238  0.1104 -0.5833
##  s.e.  0.0896  0.0722  0.0090  0.0901
##
## sigma^2 = 6.752: log likelihood = -35463.75
## AIC=70937.49  AICc=70937.5  BIC=70975.55
##
## Training set error measures:
##           ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
## Training set 0.01416751 2.598089 2.07439 NaN  Inf  0.9380695 0.0008119546
```

Cette fonction sélectionne automatiquement le meilleur modèle ARIMA en fonction de différents critères tel que l'AIC, le BIC, etc.

Le résultat nous indique que le modèle ARIMA(3,0,1) sélectionné présente des coefficients significatifs, indiquant une bonne capture des dynamiques temporelles de la série. Nous retrouvons bien un lag de 3 comme nous le prévoyions au début de notre analyse. Les critères AIC, AICc et BIC sont raisonnables, suggérant une adéquation appropriée du modèle. Les erreurs de formation, telles que le RMSE et le MAE, sont faibles, ce qui reflète une bonne performance prédictive.

4.2 Diagnostic du modèle

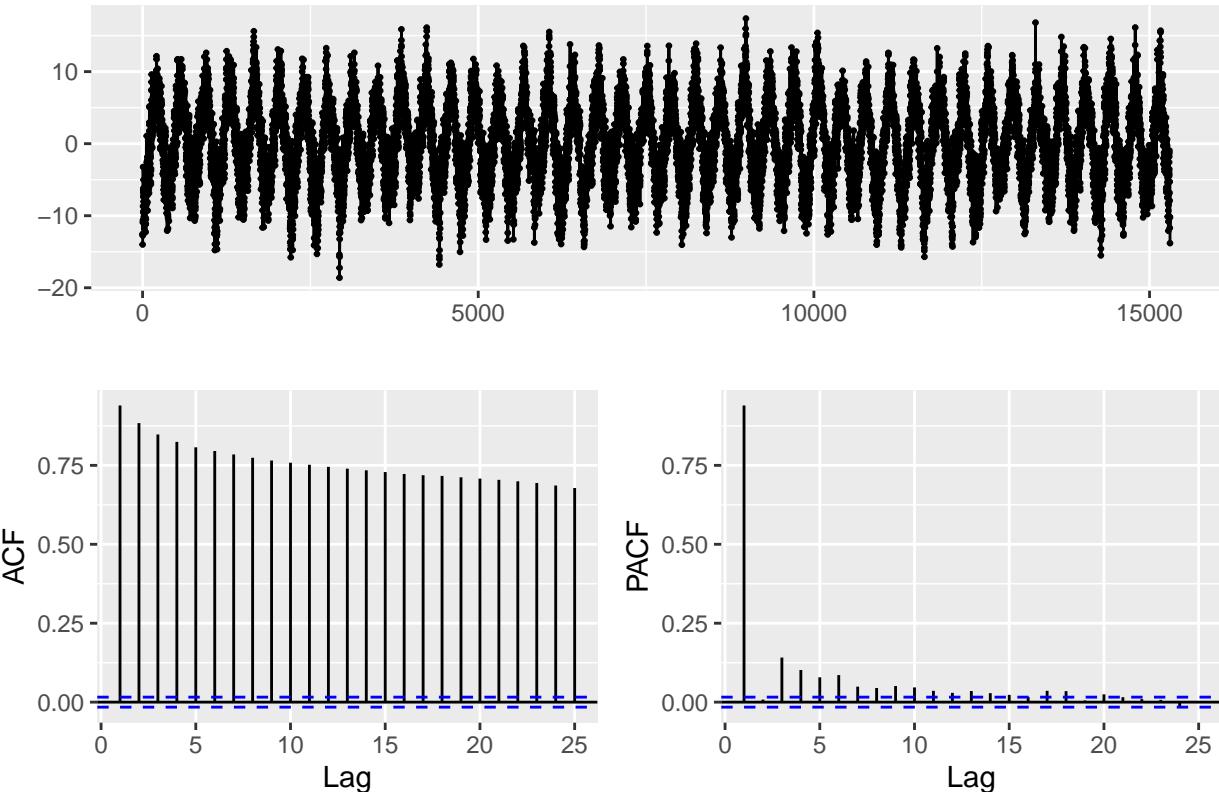
```
Box.test(fit$residuals, lag = 3) # > 5% -> pas d'auto-corrélation
```

```
##
## Box-Pierce test
##
## data: fit$residuals
## X-squared = 1.3826, df = 3, p-value = 0.7096
```

Nous obtenons une p-value égale à 0.7096 ce qui indique que l'hypothèse que nous ne rejetons pas l'hypothèse nulle selon laquelle les résidus sont non autocorrélés, indiquant que le modèle ARIMA capture bien les dépendances temporelles des données.

4.3 Etude des résidus

Nous allons maintenant étudier les résidus pour vérifier l'adéquation du modèle aux données, en s'assurant que les résidus ne présentent pas de motifs ou de structures significatifs, ce qui garantit la fiabilité des prévisions du modèle.



Sur l'étude des résidus ci-dessus, nous avons utilisé un modèle ARIMA(3,0,1) pour ajuster les résidus du modèle de régression linéaire. Le graphique de l'ACF des résidus montre qu'il n'y a pas d'autocorrélation significative, ce qui est confirmé par le test de Box-Pierce. Cela signifie que le modèle ARIMA a bien capturé les dépendances temporelles des résidus, ce qui renforce la fiabilité de notre modèle de régression linéaire.

```
# Ajustement complet avec les variables explicatives et le modèle ARIMA
ajustement_complet <- Arima(meteo.ts, c(3,0,1), xreg = as.matrix(df[,5:16]))

# Test de Box-Ljung sur les résidus de l'ajustement complet
Box.test(ajustement_complet$residuals, lag = 1)
```

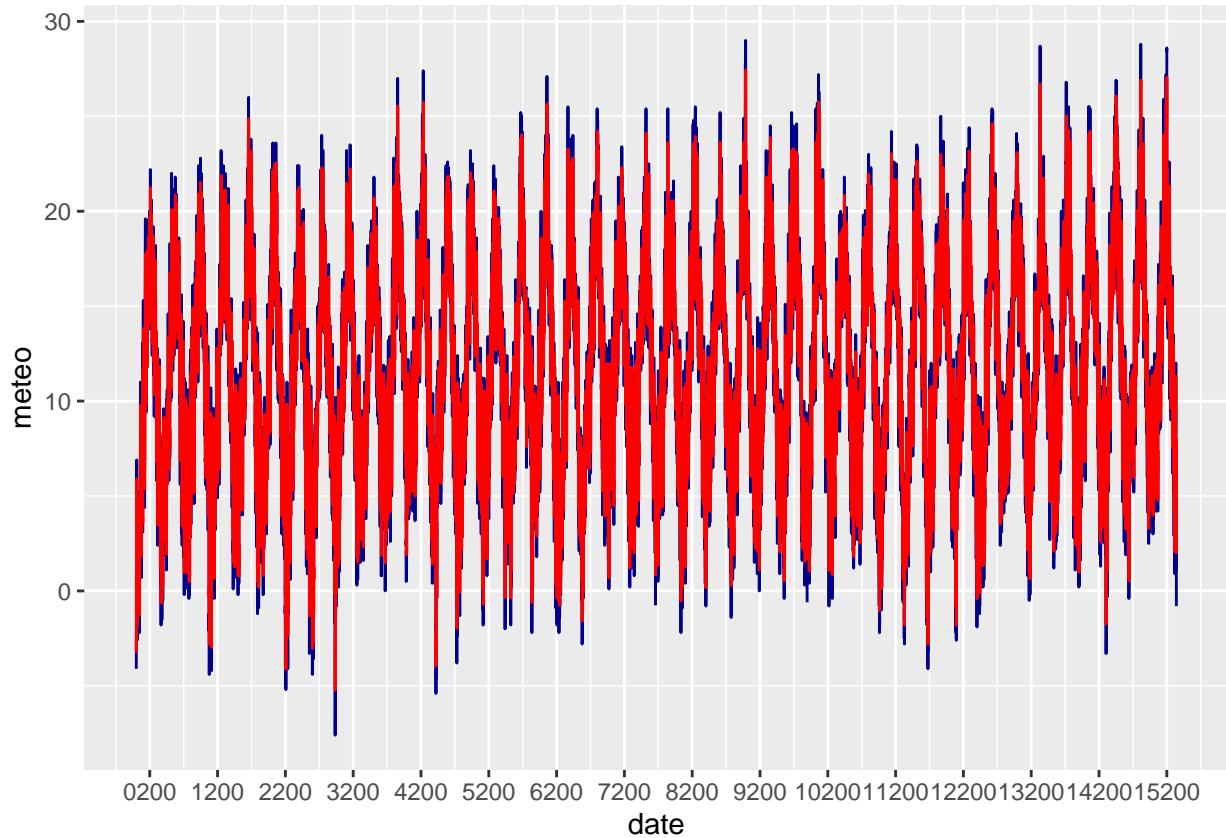
```
##
##  Box-Pierce test
##
##  data:  ajustement_complet$residuals
##  X-squared = 0.033911, df = 1, p-value = 0.8539
```

Le modèle ARIMA(3, 0, 1) avec des variables explicatives supplémentaires a bien modélisé les données de température. Les tests de diagnostic (comme le test de Box-Ljung) montrent que les résidus ne présentent pas d'autocorrélation significative, ce qui confirme la qualité de l'ajustement. Cela signifie que le modèle est adéquat pour capturer les dynamiques et la saisonnalité présentes dans les données.

4.4 Visualisation des données ajustées

```
# Création d'un dataframe pour les données ajustées
df2 <- data.frame(meteo = as.numeric(meteo.ts),
                   date = date_decimal(as.numeric(time(meteo.ts))),
                   ajustement = ajustement_complet$fitted)

# Visualisation des données ajustées
df2 %>% ggplot() +
  geom_line(aes(x = date, y = meteo), color = "blue4") +
  geom_line(aes(x = date, y = ajustement), color = "red")
```



On peut voir sur le graphique ci-dessus que les données ajustées (en rouge) suivent bien les données observées (en bleu), ce qui confirme la qualité de notre modèle ARIMA avec des variables explicatives supplémentaires pour prédire la température à Londres.

Conclusion

À travers cette analyse des températures moyennes à Londres, nous avons démontré la capacité des modèles ARIMA à capturer efficacement les dynamiques temporelles et saisonnières des données météorologiques. L'étude a débuté par une visualisation et une analyse préliminaire des données, révélant une saisonnalité annuelle claire et une légère tendance à la hausse, probablement liée au changement climatique. En décomposant la série temporelle, nous avons pu isoler ces composantes et confirmer la stationnarité de la série une fois la saisonnalité retirée.

La modélisation avec ARIMA(3,0,1), y compris des variables explicatives supplémentaires, s'est avérée être un choix judicieux. Les diagnostics effectués, tels que les tests de Dickey-Fuller et de Box-Pierce, ont validé l'absence d'autocorrélation significative dans les résidus, confirmant la robustesse de notre modèle. Les résidus de notre modèle sont apparus aléatoires, ce qui indique que toutes les structures sous-jacentes ont été bien capturées.

En conclusion, cette étude démontre que l'utilisation de méthodes de séries temporelles, et en particulier le modèle ARIMA, est une approche efficace pour analyser et prévoir les températures moyennes à Londres. Le modèle développé peut être utilisé pour des prévisions météorologiques futures, contribuant ainsi à une meilleure compréhension et gestion des conditions climatiques.