Google Ngrams

Robert Matsibekker, Kai Li,
Cruz Sanchez Hugo, Thomas Green

Applied Mathematics and Statistics, Stony Brook University

December 6, 2021

Introduction

N-gram is a concept from computational linguistics and probability, and it is defined as a contiguous sequence of n items from a sample of text (called corpora). This sequence can be any combination of phonemes, syllables, letters, words since it has a meaning in the language where it comes from. If it is a single word, it is called unigram, an expression with two words is a bigram, and so on (source: `https://en.wikipedia.org/wiki/N-gram`).

Introduction

In this work, we retrieved our data from Google Ngram Viewer for the word "peace" from the corpora in English. The English corpora are composed of printed books in this language and their data is aggregated in years, from 1500 to 2019. The program is an online search engine that searches for a given ngram and returns the normalized percentage of appearance of the searched ngrams on books published each year (source: `https://en.wikipedia.org/wiki/Google_Ngram_Viewer`).

It can be accessed on this page: `https://books.google.com/ngrams`.

Setting, libraries, and reading data

The series is discontinuous before 1533.

```
df = ngram("peace", year_start = 1533, smoothing = 0,
           count = TRUE, case_ins = TRUE, aggregate = TRUE)

str(df)
```

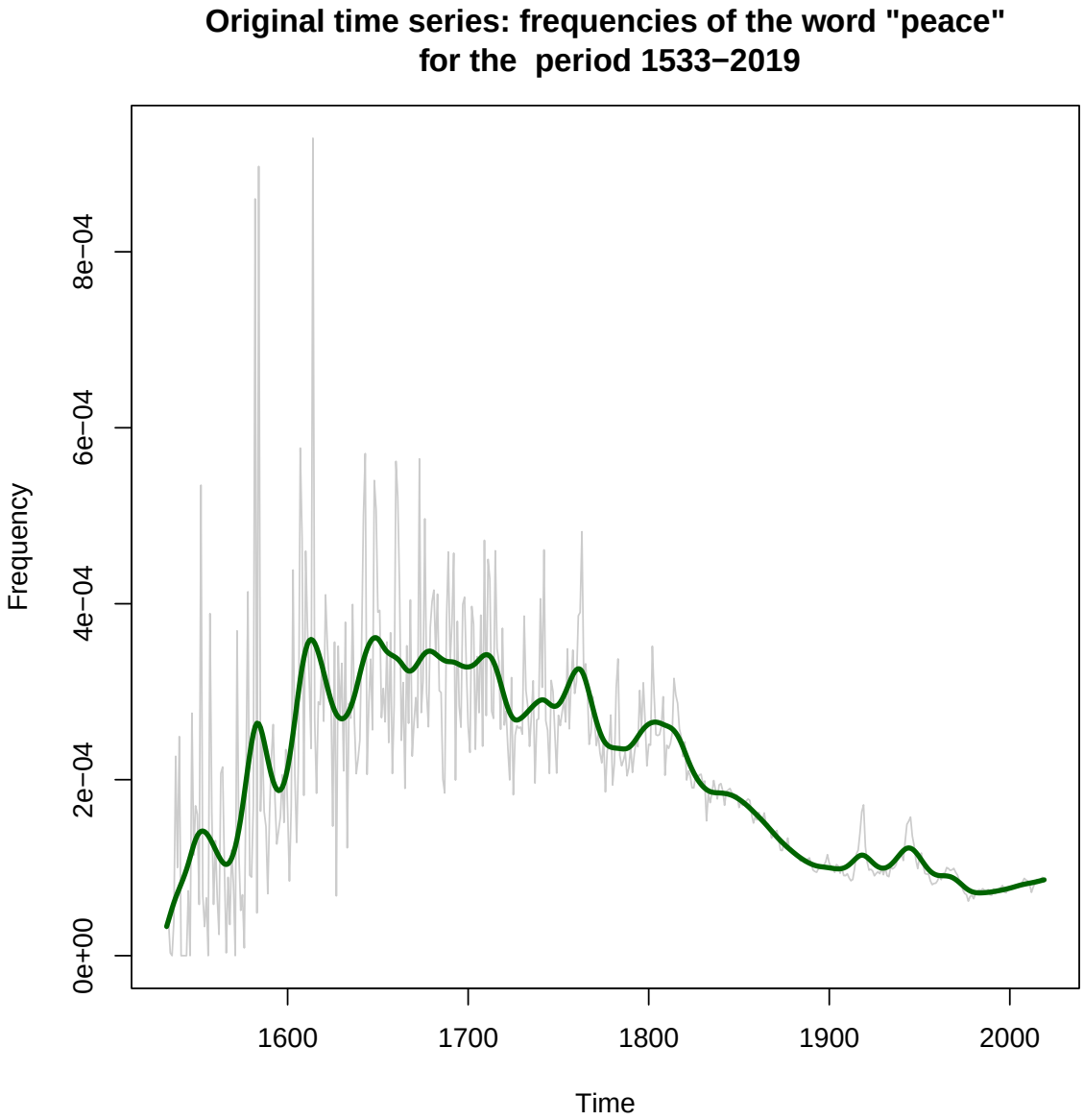
```
## Classes 'ngram' and 'data.frame': 487 obs. of  5 variables:
##  $ Year     : int  1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 ...
##  $ Corpus   : Factor w/ 1 level "eng_2019": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Phrase   : Factor w/ 1 level "peace (All)": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Frequency: num  3.06e-05 4.02e-05 3.16e-06 0.00 3.93e-05 ...
##  $ Count    : num  6 1 1 0 4 86 32 1 0 0 ...
##  - attr(*, "case_sensitive")= logi FALSE
##  - attr(*, "smoothing")= num 0
```

Setting, libraries, and reading data

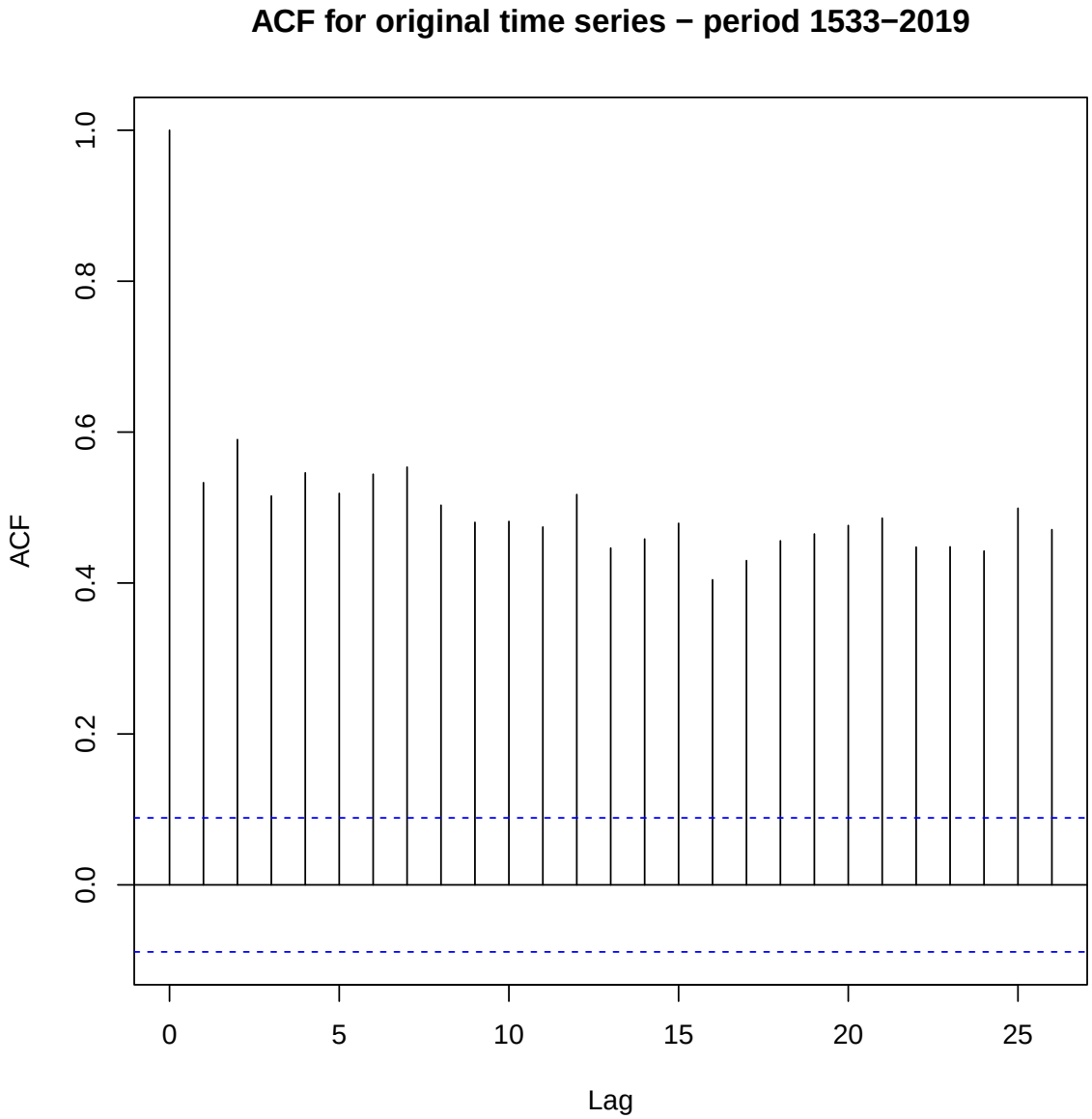
```
summary(df)
```

```
##       Year            Corpus              Phrase      Frequency
##  Min.   :1533   eng_2019:487   peace (All):487   Min.   :0.000e+00
##  1st Qu.:1654                                    1st Qu.:9.903e-05
##  Median :1776                                    Median :1.875e-04
##  Mean   :1776                                    Mean   :2.046e-04
##  3rd Qu.:1898                                    3rd Qu.:2.711e-04
##  Max.   :2019                                    Max.   :9.291e-04
##      Count
##  Min.   :      0.0
##  1st Qu.:    775.5
##  Median :  14546.0
##  Mean   : 384809.5
##  3rd Qu.: 543020.5
##  Max.   :2715105.0
```

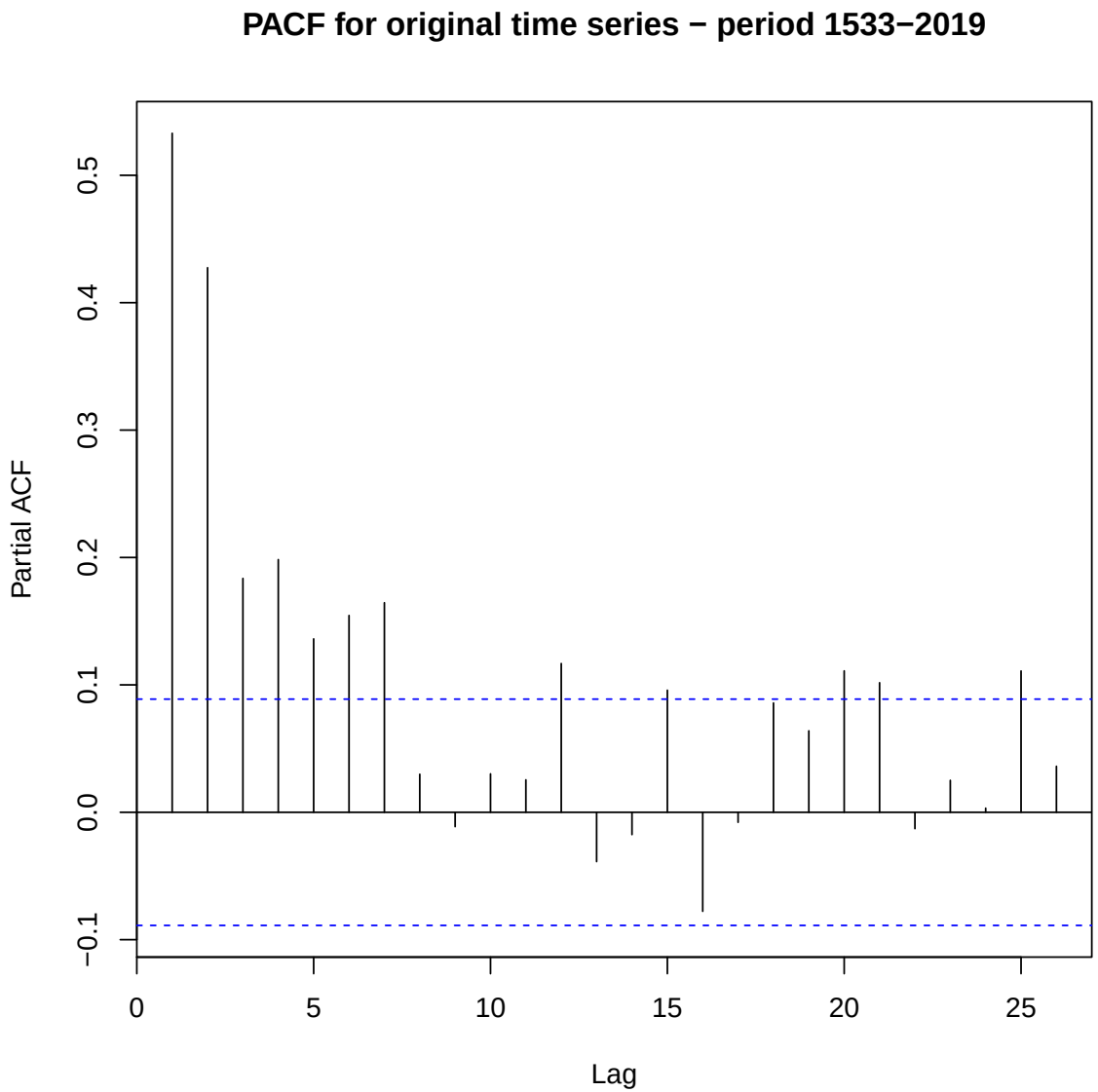
Setting, libraries, and reading data



Setting, libraries, and reading data



Setting, libraries, and reading data



Setting, libraries, and reading data

```
adf.test(ots)

##
##  Augmented Dickey-Fuller Test
##
## data:  ots
## Dickey-Fuller = -3.2849, Lag order = 7, p-value = 0.07334
## alternative hypothesis: stationary
```

It is stationary at 10%, but not at 5%.

Investigating Structural Breaks

Reference: `https://cran.r-project.org/web/packages/strucchange/vignettes/strucchange-intro.pdf`

The first half of the series has a different behavior from the second half.

When we break the series into two pieces we get two stationary models so we decided to investigate for structural breaks.

Investigating Structural Breaks

```
fit = auto.arima(ots, seasonal=FALSE, test="adf", ic="bic",
                 lambda=NULL, stepwise=FALSE,
                 approximation=FALSE, max.p=3, max.q=3)
summary(fit)
```

```
## Series: ots
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2     ar3   mean
##       0.2217  0.3723  0.1886  2e-04
## s.e.  0.0449  0.0428  0.0450  1e-04
##
## sigma^2 estimated as 9.595e-09:  log likelihood=3727.94
## AIC=-7445.89   AICc=-7445.76   BIC=-7425.05
##
## Training set error measures:
##                       ME         RMSE          MAE  MPE MAPE      MASE
## Training set 8.24818e-07 9.754265e-05 5.839932e-05 -Inf  Inf 0.9185002
##                     ACF1
## Training set -0.04057325
```

Investigating Structural Breaks

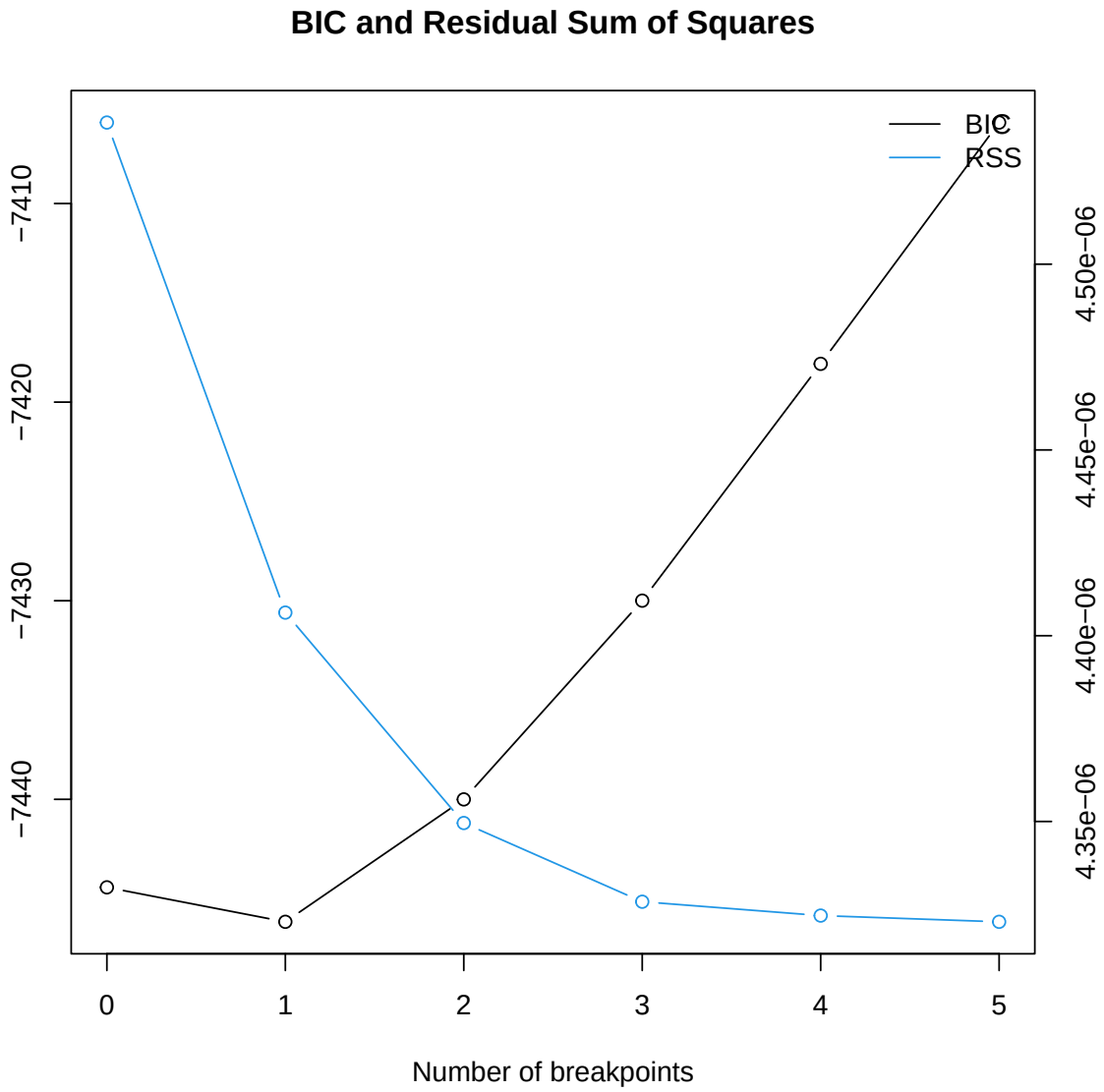
```
as.data.frame(round(confint(fit), 4))
```

```
##              2.5 % 97.5 %
## ar1         0.1338 0.3096
## ar2         0.2884 0.4561
## ar3         0.1004 0.2769
## intercept 0.0001 0.0003
```

We don't have a theoretical model, as it is shown in the reference (the authors use a theoretical macroeconomic relation), so we assume the following hypothesis: a time series with a structural change fitted by only one model (set of parameters) has residuals which show that the model is not suitable. This inadequacy can be detected by testing structural breaks of the level of the residuals.

Investigating Structural Breaks

The next test can detect many structural breaks.

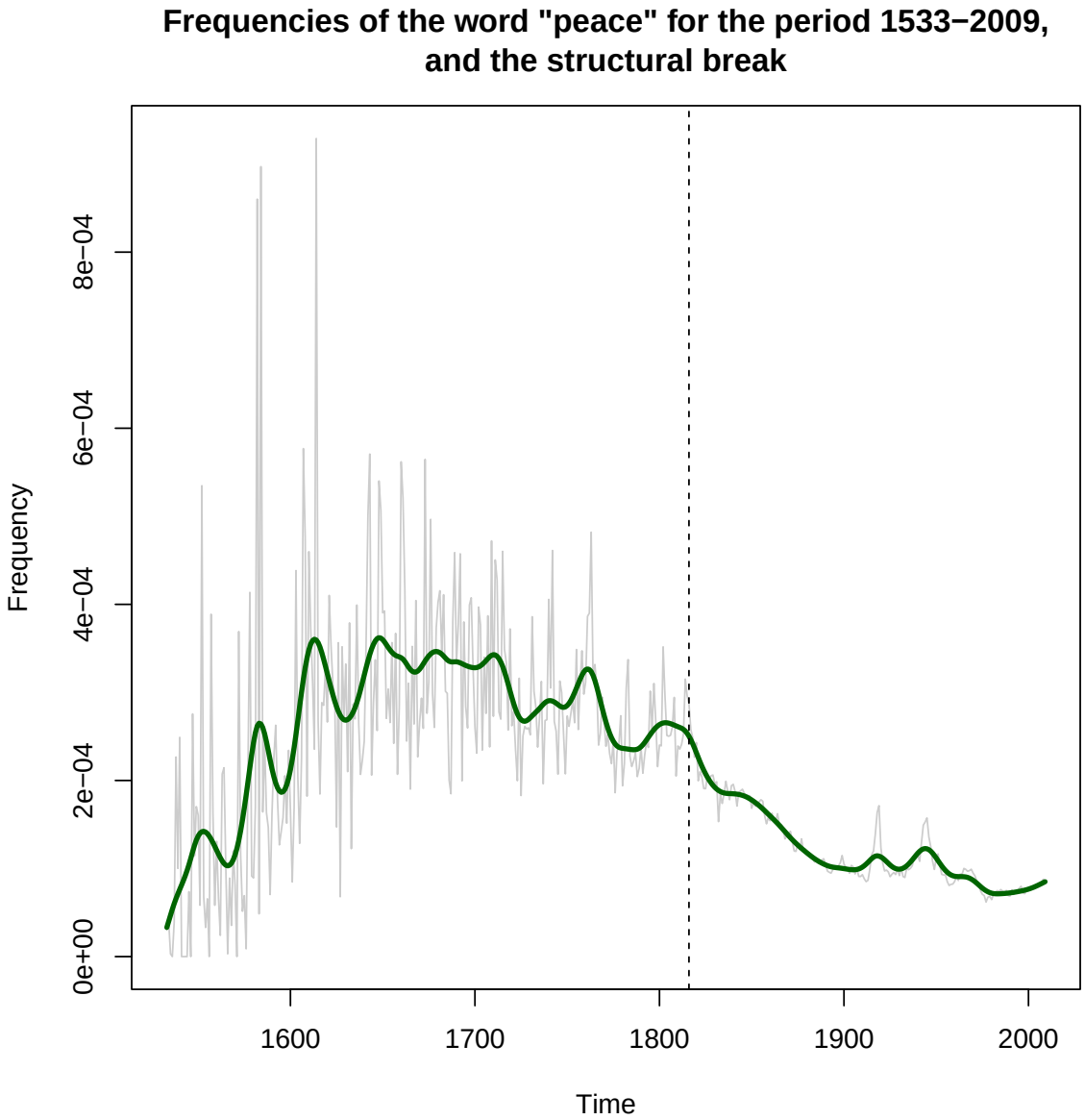


Investigating Structural Breaks

Minimum BIC is at 1. So we have 1 structural break.

```
breakdates(BPtest)

## [1] 1816

#round(min(time(ots)) + breakdates(BPtest)*
# (max(time(ots)) - min(time(ots))))
# use it when it is a fraction
```

Investigating Structural Breaks

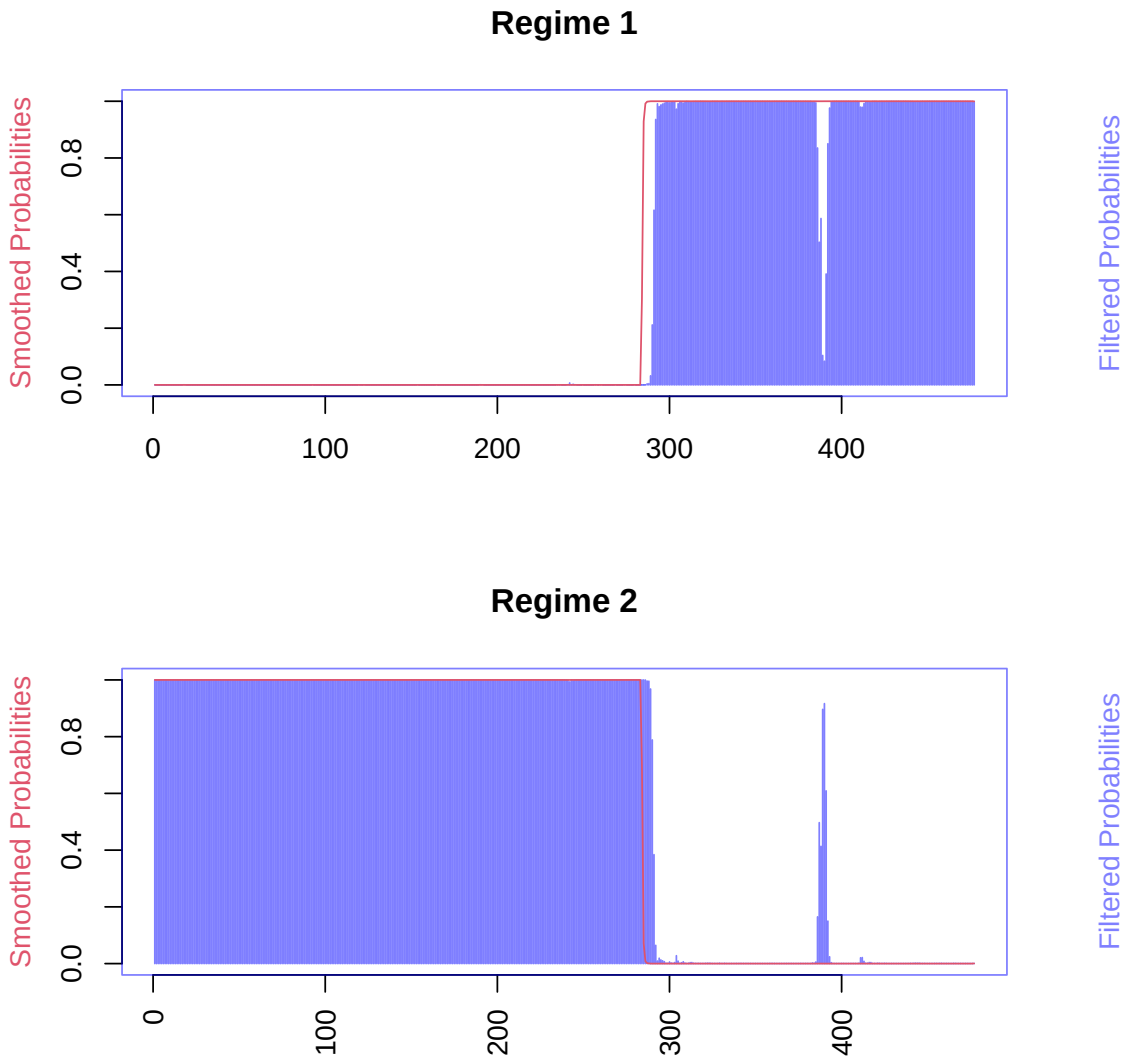


Vertical line at the structural break.

One structural break (hidden states Markov model)

Next, we show that this (hidden states Markov) model of regime-switching finds a structural change at a date similar to the date we formerly found, which corresponds to 284 (the year 1817) in the following plot (the plot shows probabilities for each regime).

One structural break (hidden states Markov model)



One structural break (hidden states Markov model)

Why did these structural breaks happen?

Shifts in the relative frequency of the word "peace" might be related to changes in legislation, technological improvements, social development, and historical events:

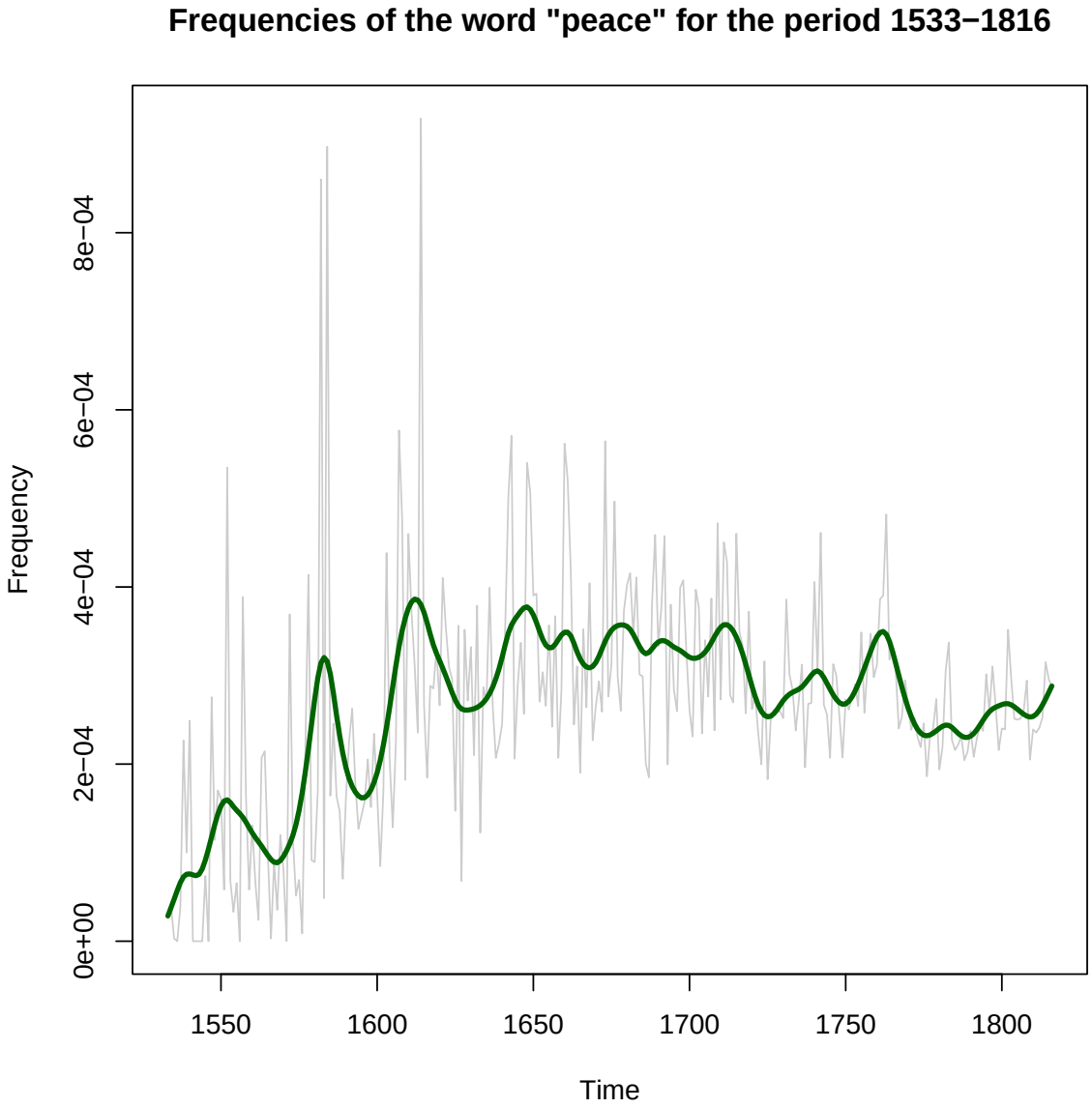
1. Change in legislation: at the end of the XVIII century, changes in legislation (the USA and GB) made it easier to publish dissent texts (source: `https://www.britannica.com/topic/publishing/Spread-of-education-and-literacy#ref28633`).
2. Technological improvements: several innovations related to publishing happened at the beginning of the XIX century and made it much cheaper (source: `https://www.britannica.com/topic/publishing/Spread-of-education-and-literacy#ref28633`).
3. Social development: increase in population size during the XIX century (2x in GB, 5x in the USA) and higher social status associated with reading (source: `https://www.britannica.com/topic/publishing/Spread-of-education-and-literacy#ref28633`).

One structural break (hidden states Markov model)

4. Historical events: several major events in American and British history, associated with peace and war, happened around 1816:
   1. War against Great Britain in 1812 (source: `https://history.state.gov/milestones/1801-1829/war-of-1812`), part of the world scenario of the Napoleonic Wars.
   2. War and against the Barbary States (pirate states in North Africa) in 1816 (source: `https://history.state.gov/milestones/1801-1829/barbary-wars`)
   3. The Rush-Bagot Pact, 1817 and Convention of 1818, between the USA and Great Britain, about patrolling the border with Canada (source: `https://history.state.gov/milestones/1801-1829/rush-bagot`).
   4. Acquisition of Florida: Treaty of Adams-Onis (1819) and Transcontinental Treaty (1821), a series of border conflicts between Spain and the USA, fueled by the support of Great Britain to the Spanish colonies (source: `https://history.state.gov/milestones/1801-1829/florida`).

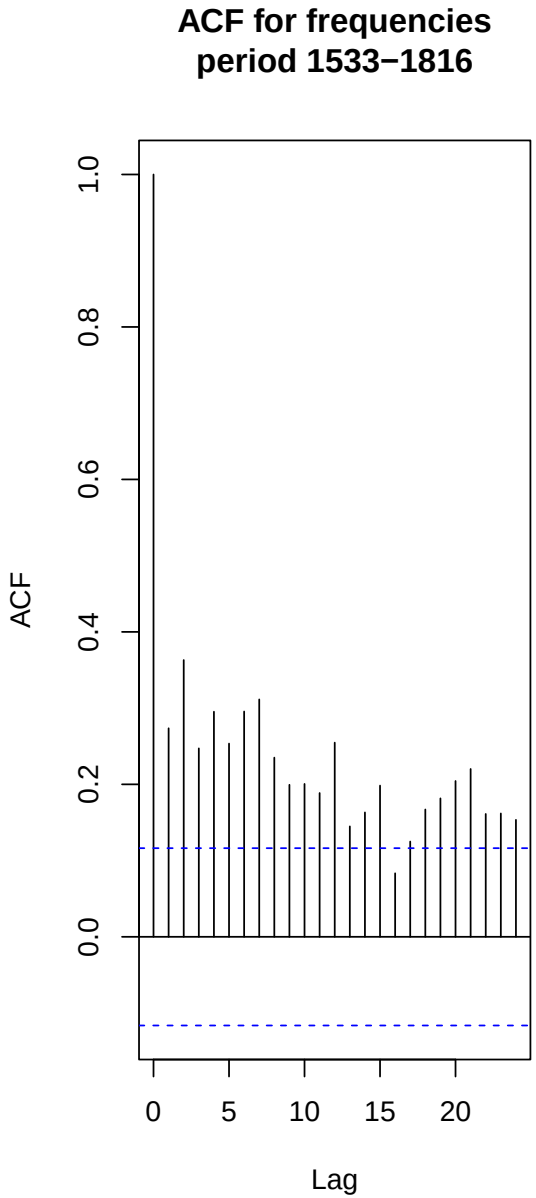
2 different models

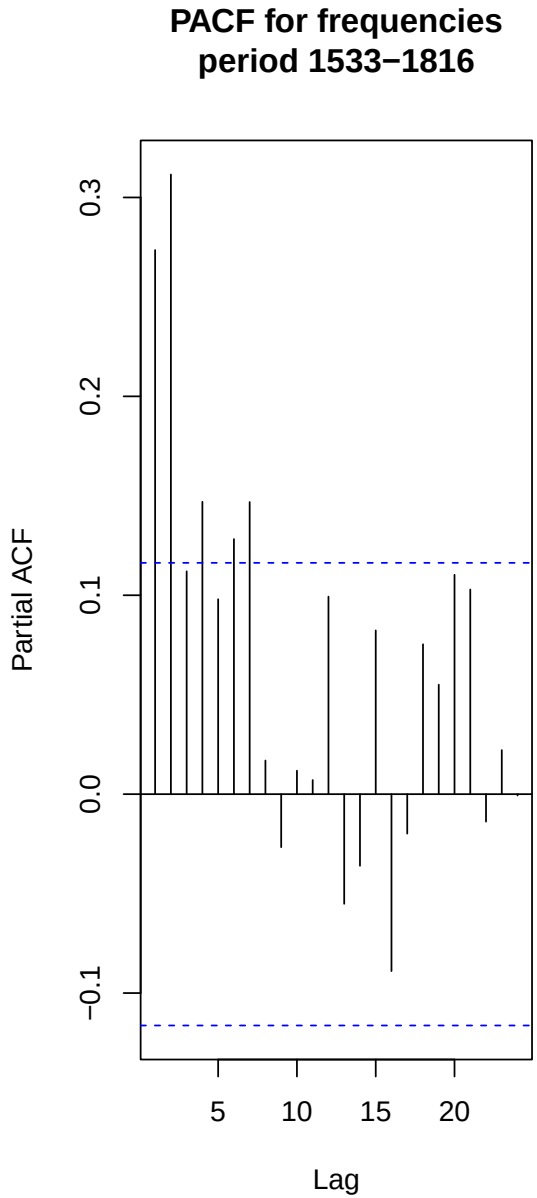
First model - 1533-1816



2 different models

First model - 1533-1816





2 different models

First model - 1533-1816

```
adf.test(ots1)

##
##  Augmented Dickey-Fuller Test
##
## data:  ots1
## Dickey-Fuller = -3.3503, Lag order = 6, p-value = 0.06312
## alternative hypothesis: stationary
```

It is stationary at 10%, but not at 5%.

2 different models

First model - 1533-1816

```
fit_1 = auto.arima(ots1, seasonal=FALSE, test="adf",
                   ic="bic", lambda=NULL, stepwise=FALSE,
                   approximation=FALSE, max.p=3, max.q=3)
summary(fit_1)
```

```
## Series: ots1
## ARIMA(2,0,1) with non-zero mean
##
## Coefficients:
##          ar1     ar2      ma1   mean
##       0.9257  0.0604  -0.8676  2e-04
## s.e.  0.0864  0.0766   0.0672  1e-04
##
## sigma^2 estimated as 1.421e-08:  log likelihood=2164.32
## AIC=-4318.63   AICc=-4318.42   BIC=-4300.39
##
## Training set error measures:
##                       ME         RMSE          MAE  MPE MAPE      MASE
## Training set 7.295323e-06 0.0001183568 7.737577e-05 -Inf  Inf 0.7556539
##                      ACF1
## Training set -0.002287252
```

2 different models

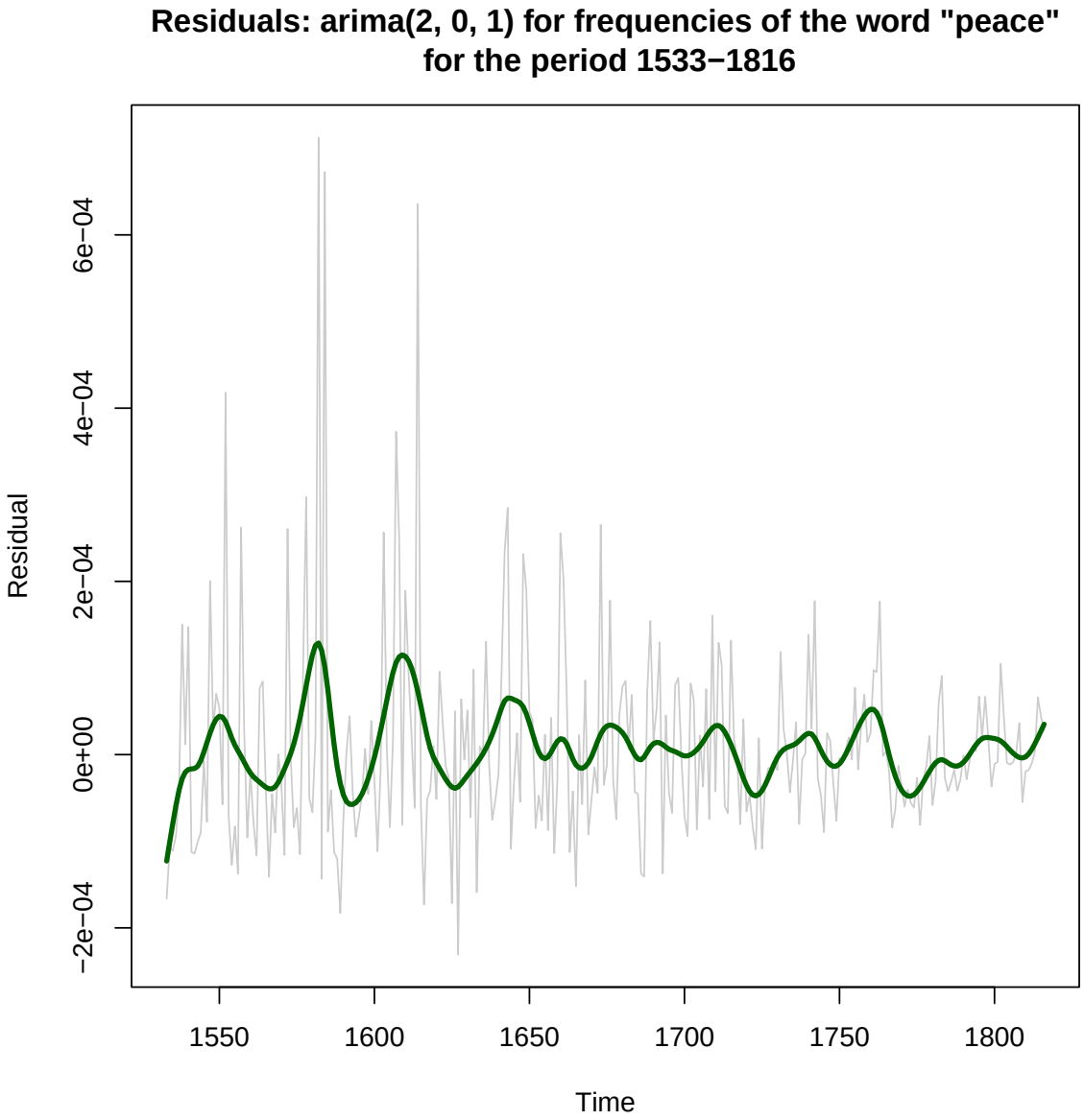
First model - 1533-1816

```
as.data.frame(round(confint(fit_1), 4))
```

```
##              2.5 %  97.5 %
## ar1         0.7563  1.0951
## ar2        -0.0897  0.2104
## ma1        -0.9992 -0.7359
## intercept   0.0000  0.0004
```

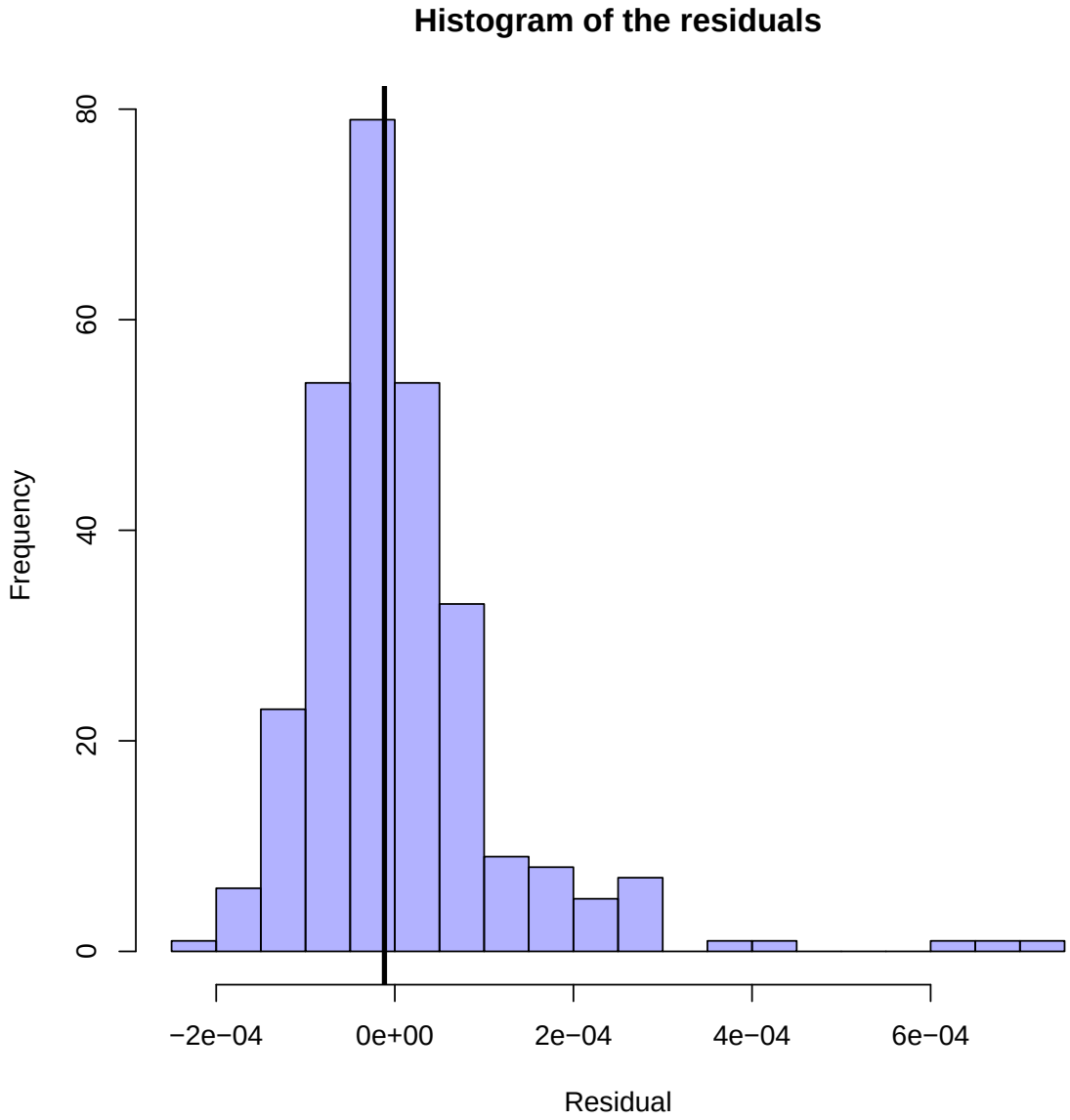
2 different models

First model - 1533-1816



2 different models

First model - 1533-1816



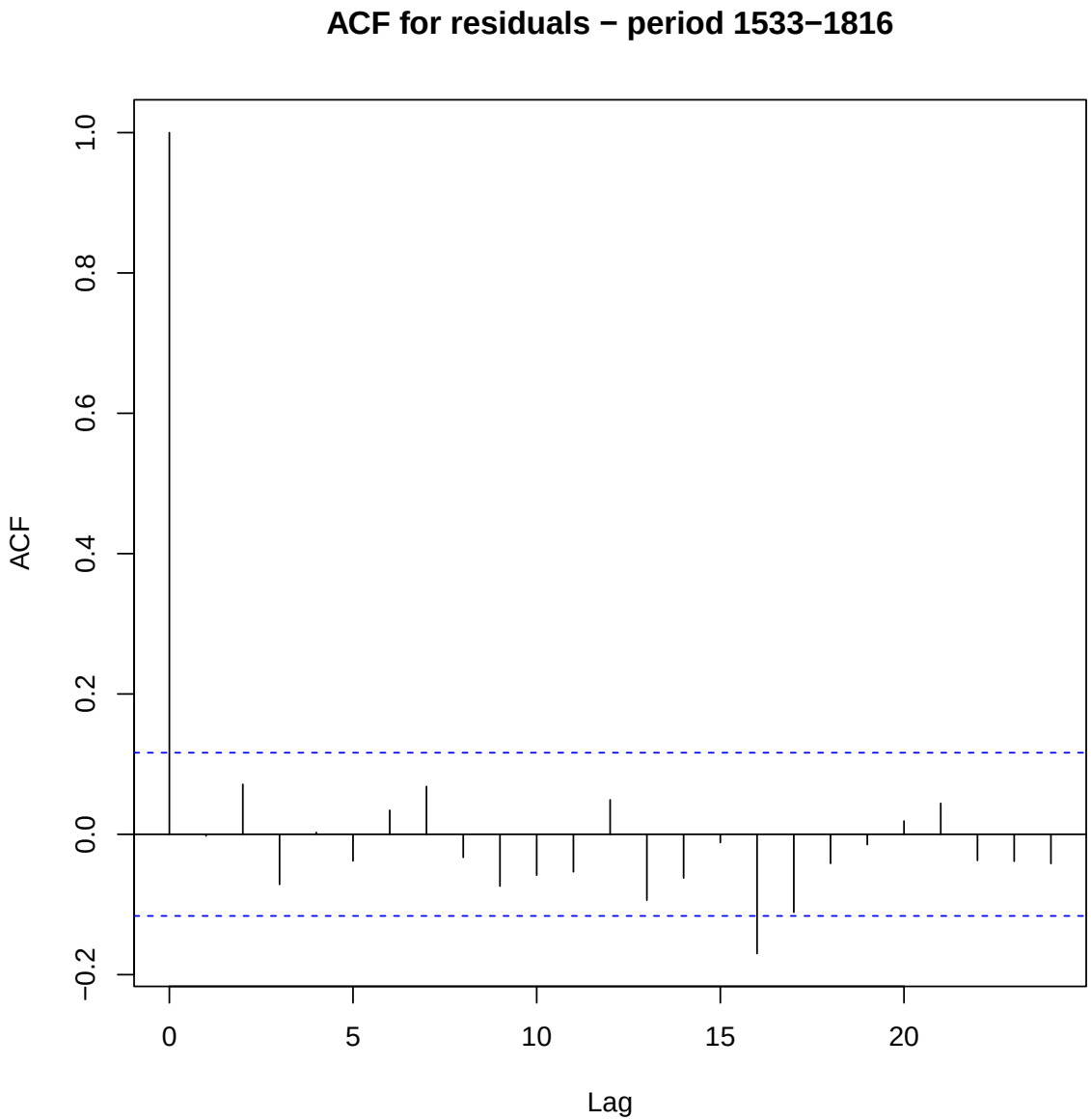
2 different models

First model - 1533-1816

```
shapiro.test(resid)

##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.82044, p-value < 2.2e-16
```

2 different models

First model - 1533-1816



2 different models

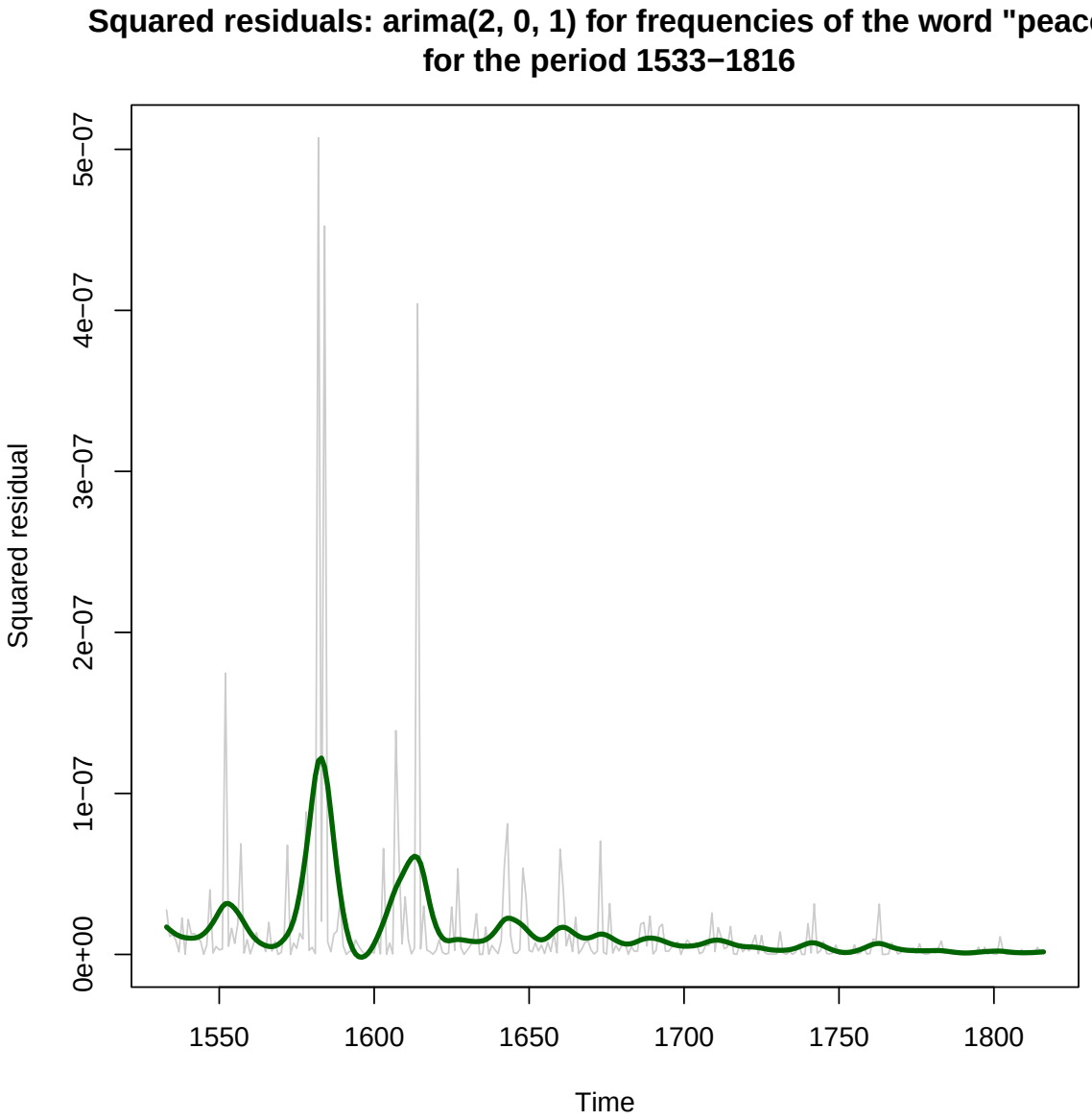
First model - 1533-1816

```
Box.test(resid, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  resid
## X-squared = 0.0015015, df = 1, p-value = 0.9691

# H0: indep./uncorr.
```

Residuals are independent/uncorrelated.

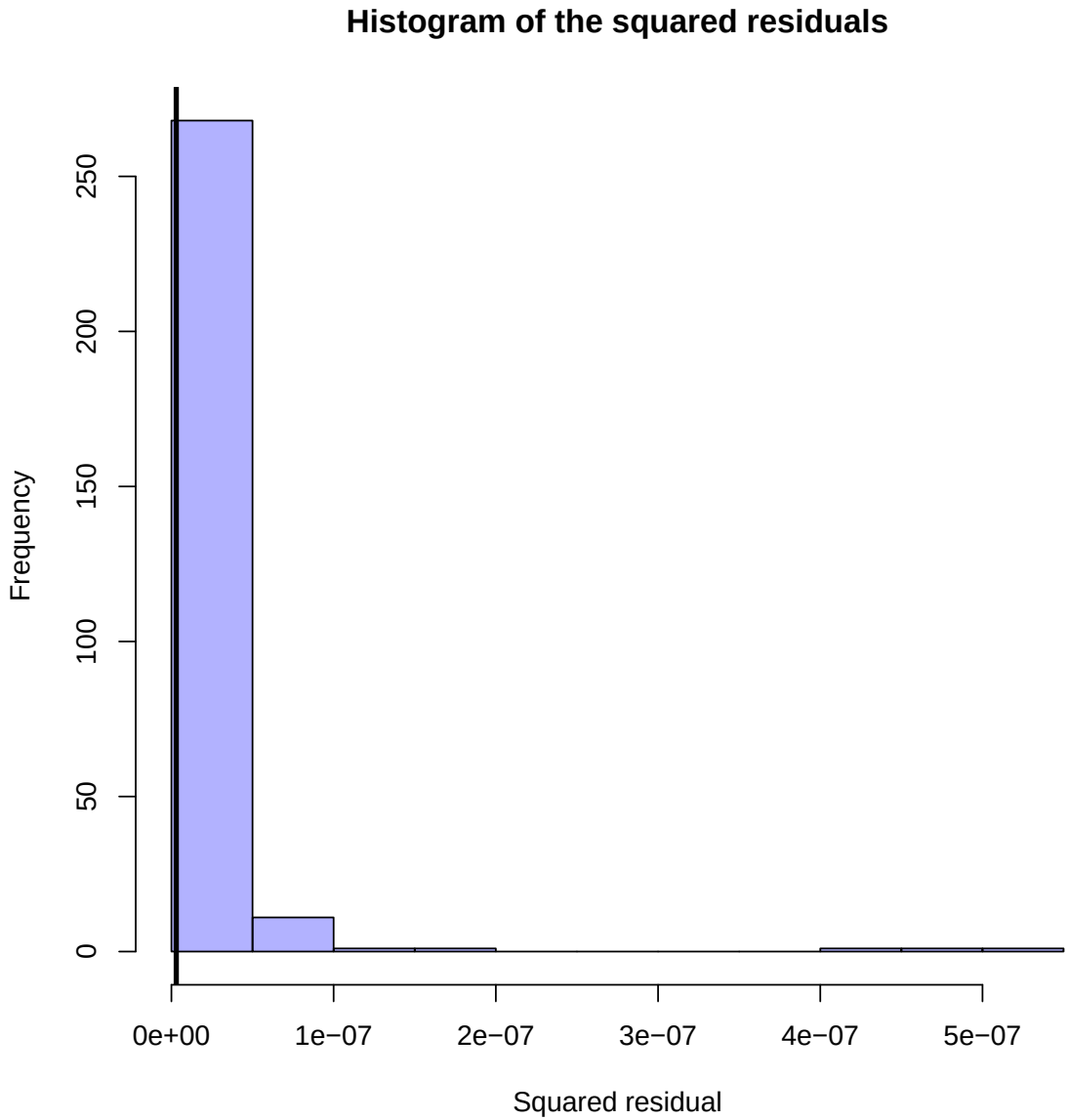
2 different models

First model - 1533-1816



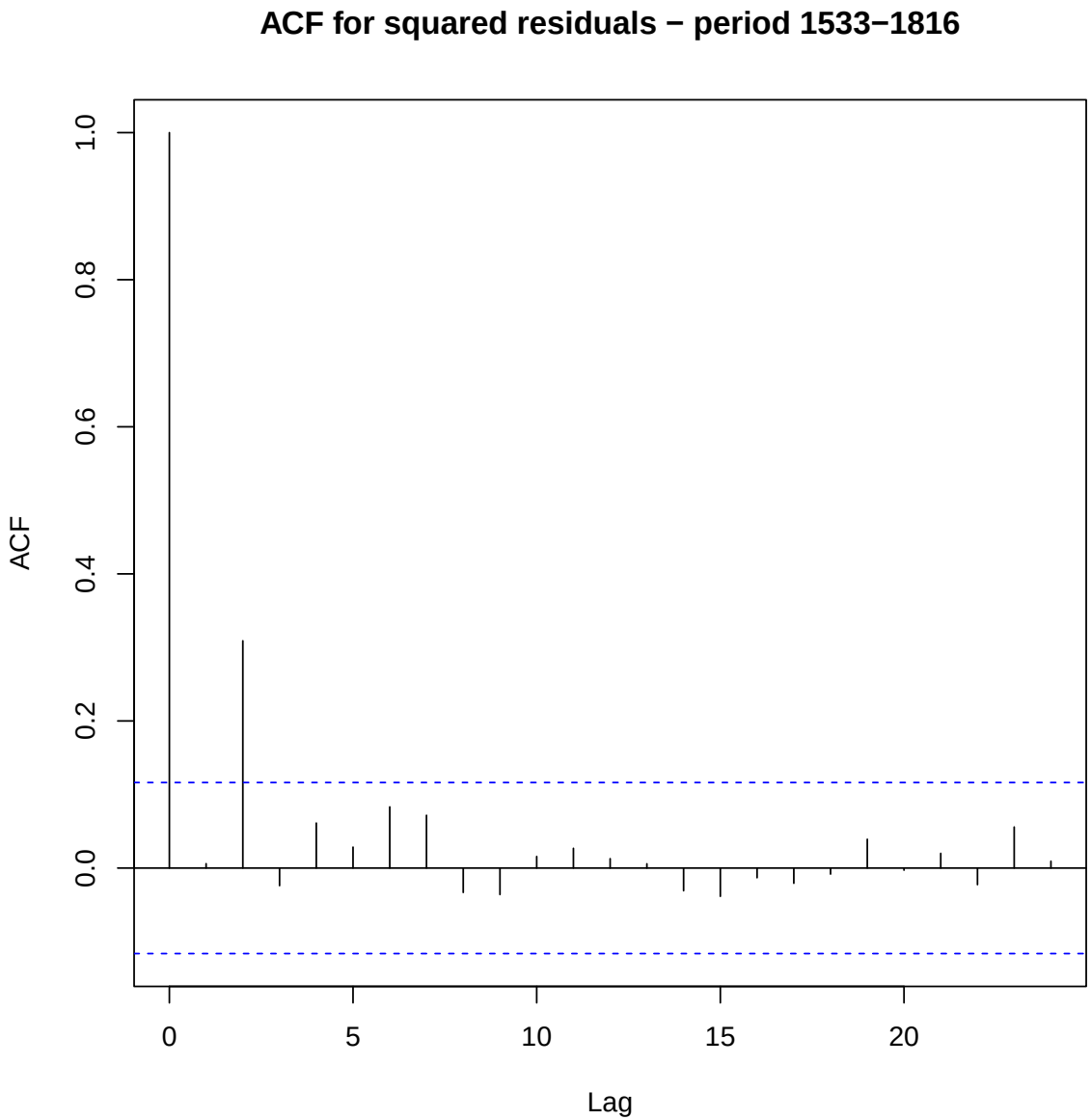
2 different models

First model - 1533-1816



2 different models

First model - 1533-1816



2 different models

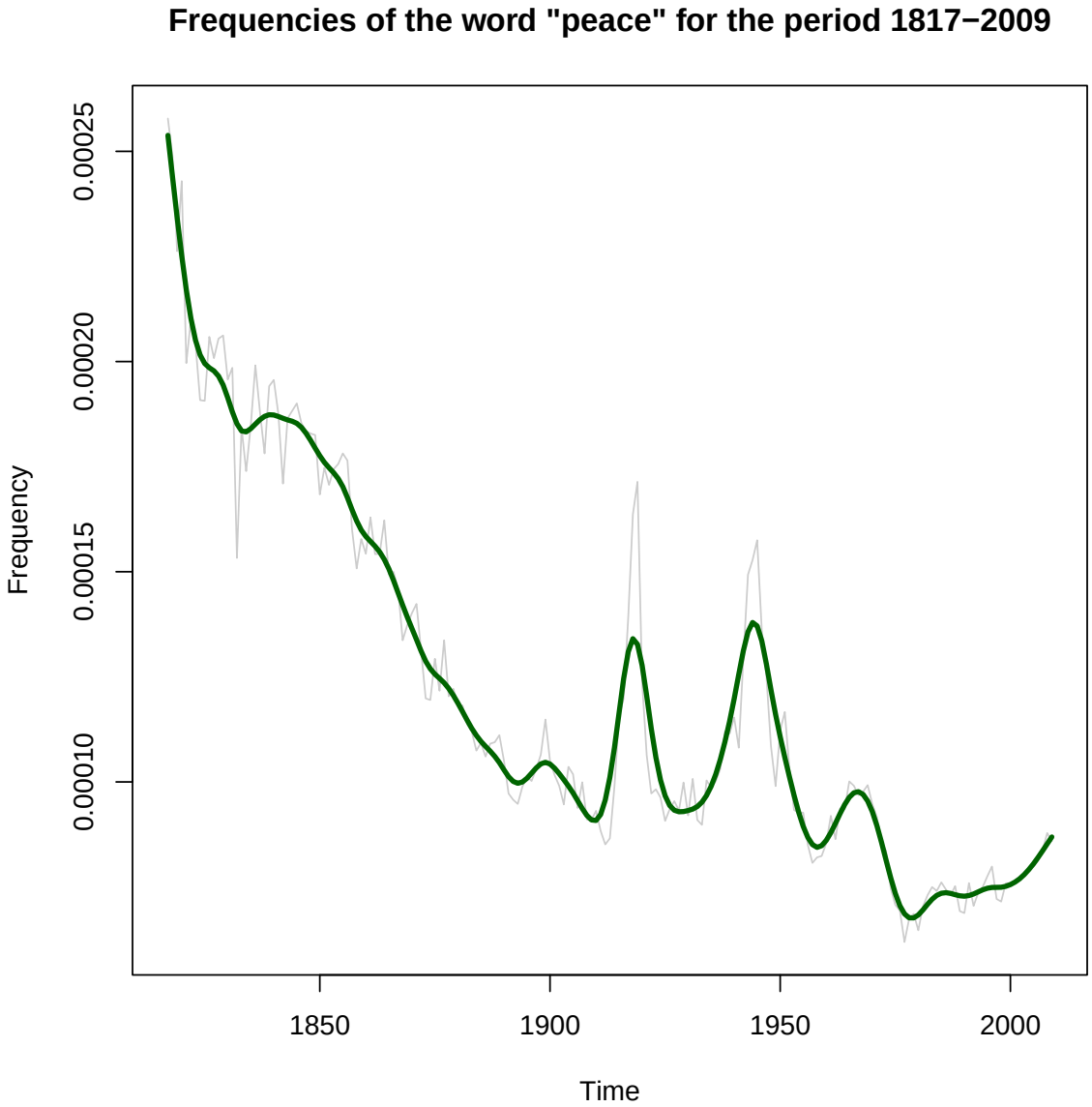
First model - 1533-1816

```
Box.test(abs(resid)^2, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  abs(resid)^2
## X-squared = 0.010222, df = 1, p-value = 0.9195

# H0: indep./uncorr.
```

Squared residuals are independent/uncorrelated.

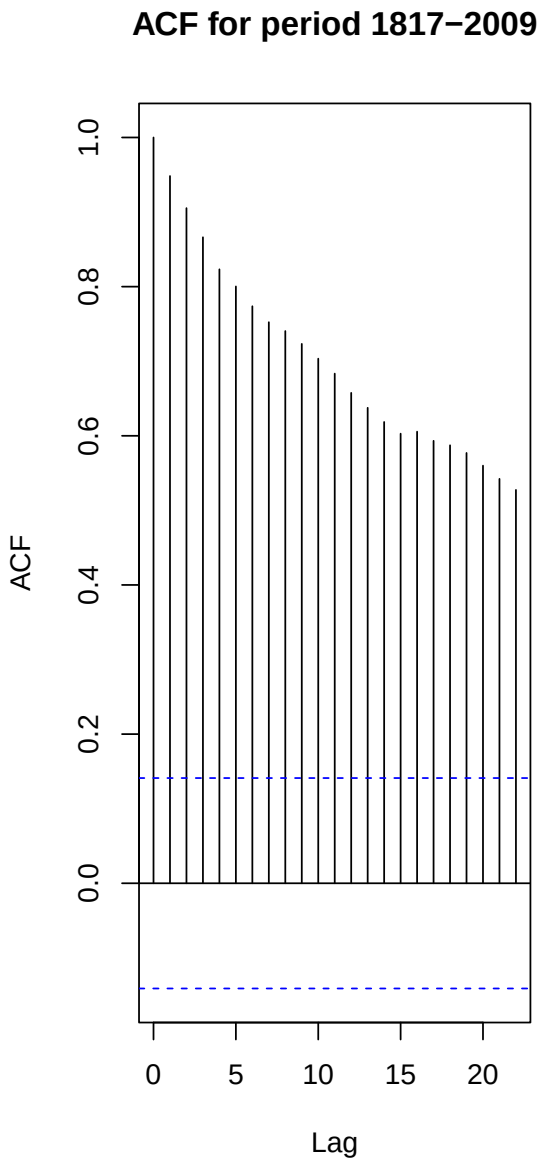
2 different models

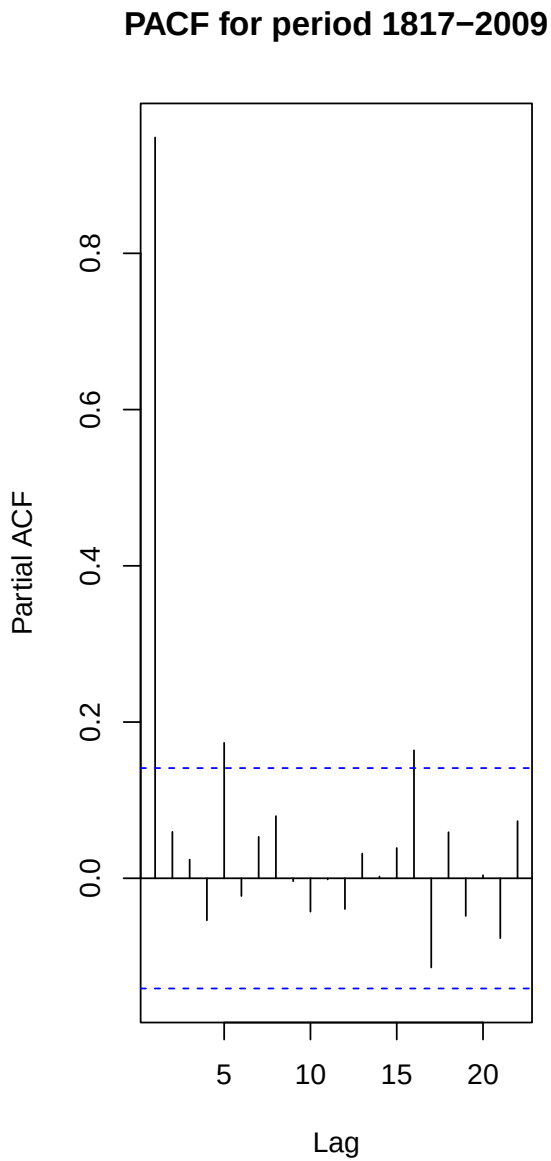
Second model - 1817-2009



2 different models

Second model - 1817-2009





2 different models

Second model - 1817-2009

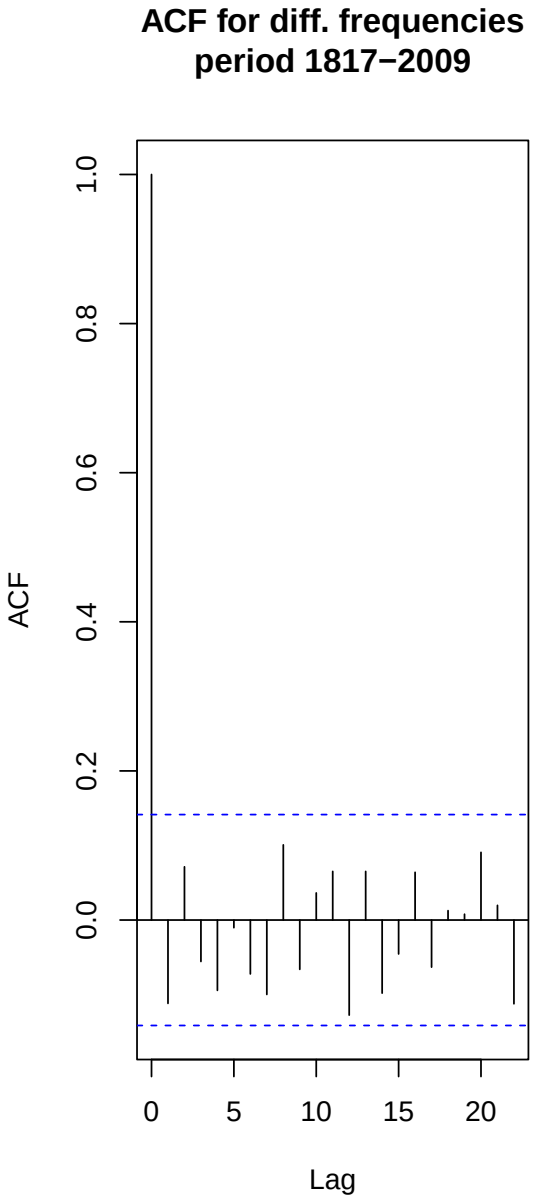
```
adf.test(ots2)

##
##  Augmented Dickey-Fuller Test
##
## data:  ots2
## Dickey-Fuller = -2.7838, Lag order = 5, p-value = 0.2481
## alternative hypothesis: stationary
```

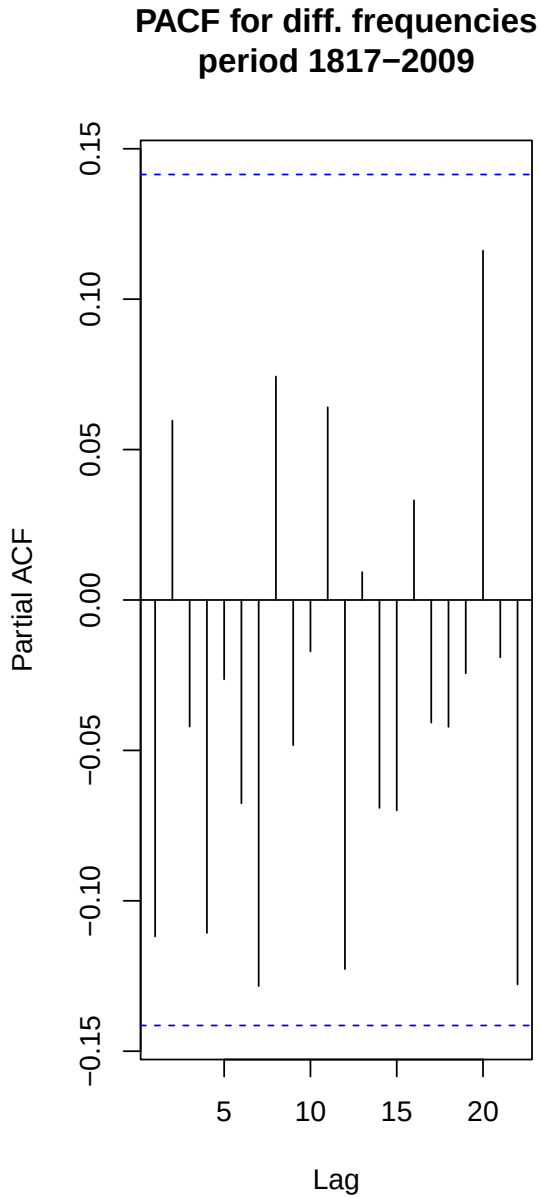
ADF test does not reject non-stationarity even at 10%. But the best model for this series with auto.arima is an AR(1). We prefer to difference the series.

```
dots2=na.omit(diff(ots2))
```

2 different models

Second model - 1817-2009





2 different models

Second model - 1817-2009

```
fit_2 = auto.arima(dots2, seasonal=FALSE, test="adf",
                   ic="bic", lambda=NULL, stepwise=FALSE,
                   approximation=FALSE, max.p=3, max.q=3)
summary(fit_2)
```

```
## Series: dots2
## ARIMA(0,0,0) with zero mean
##
## sigma^2 estimated as 9.689e-11:  log likelihood=1941.08
## AIC=-3880.17   AICc=-3880.15   BIC=-3876.91
##
## Training set error measures:
##                        ME         RMSE          MAE MPE MAPE      MASE
## Training set -8.954279e-07 9.843036e-06 6.549165e-06 100  100 0.6769678
##                     ACF1
## Training set -0.1118436
```

2 different models

Second model - 1817-2009

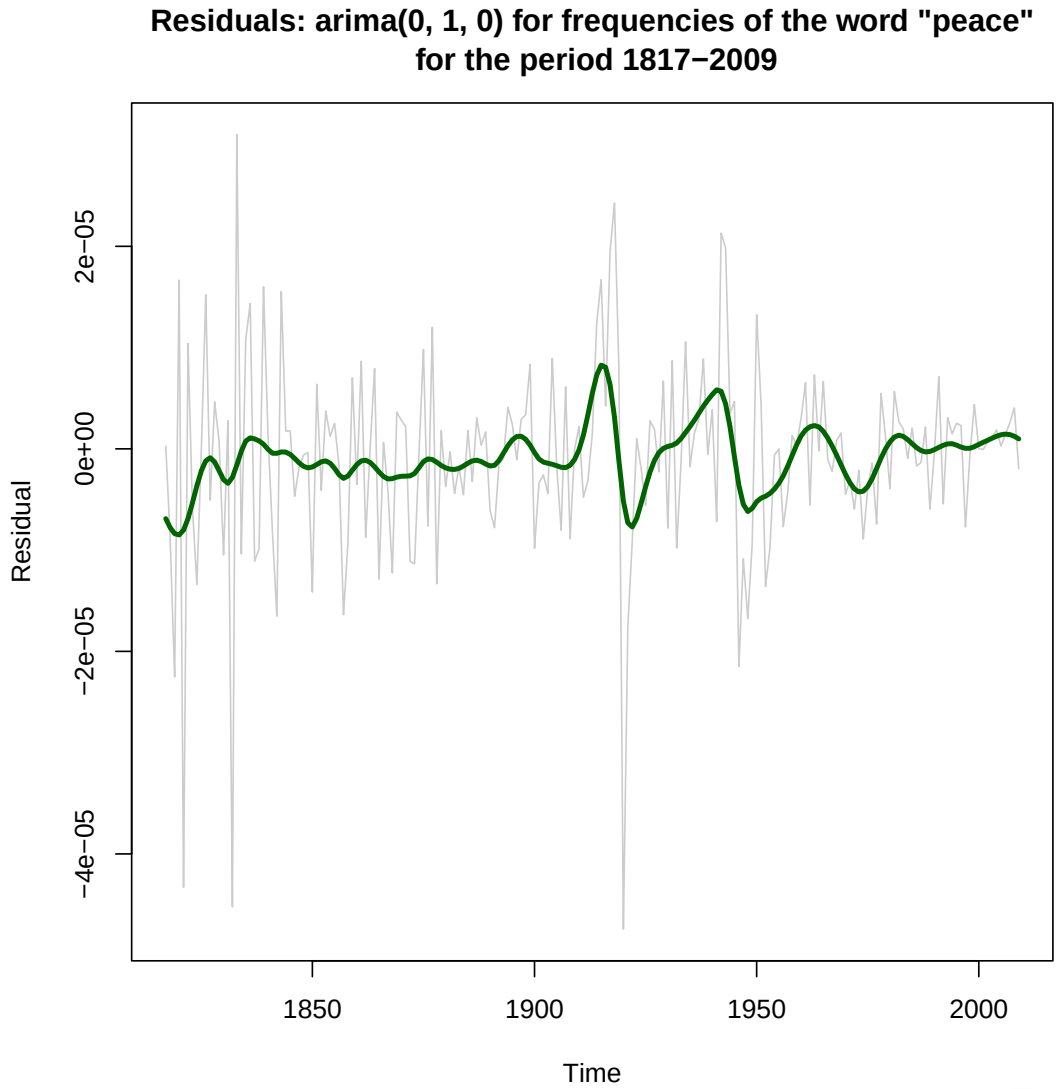
```
fit_2 = Arima(ots2, order=c(0,1,0))
summary(fit_2)
```

```
## Series: ots2
## ARIMA(0,1,0)
##
## sigma^2 estimated as 9.689e-11:  log likelihood=1941.08
## AIC=-3880.17   AICc=-3880.15   BIC=-3876.91
##
## Training set error measures:
##                        ME        RMSE           MAE        MPE     MAPE
## Training set -8.894525e-07 9.81752e-06 6.516567e-06 -0.8246778 5.25744
##                    MASE       ACF1
## Training set 0.9950226 -0.1123449
```

2 different models

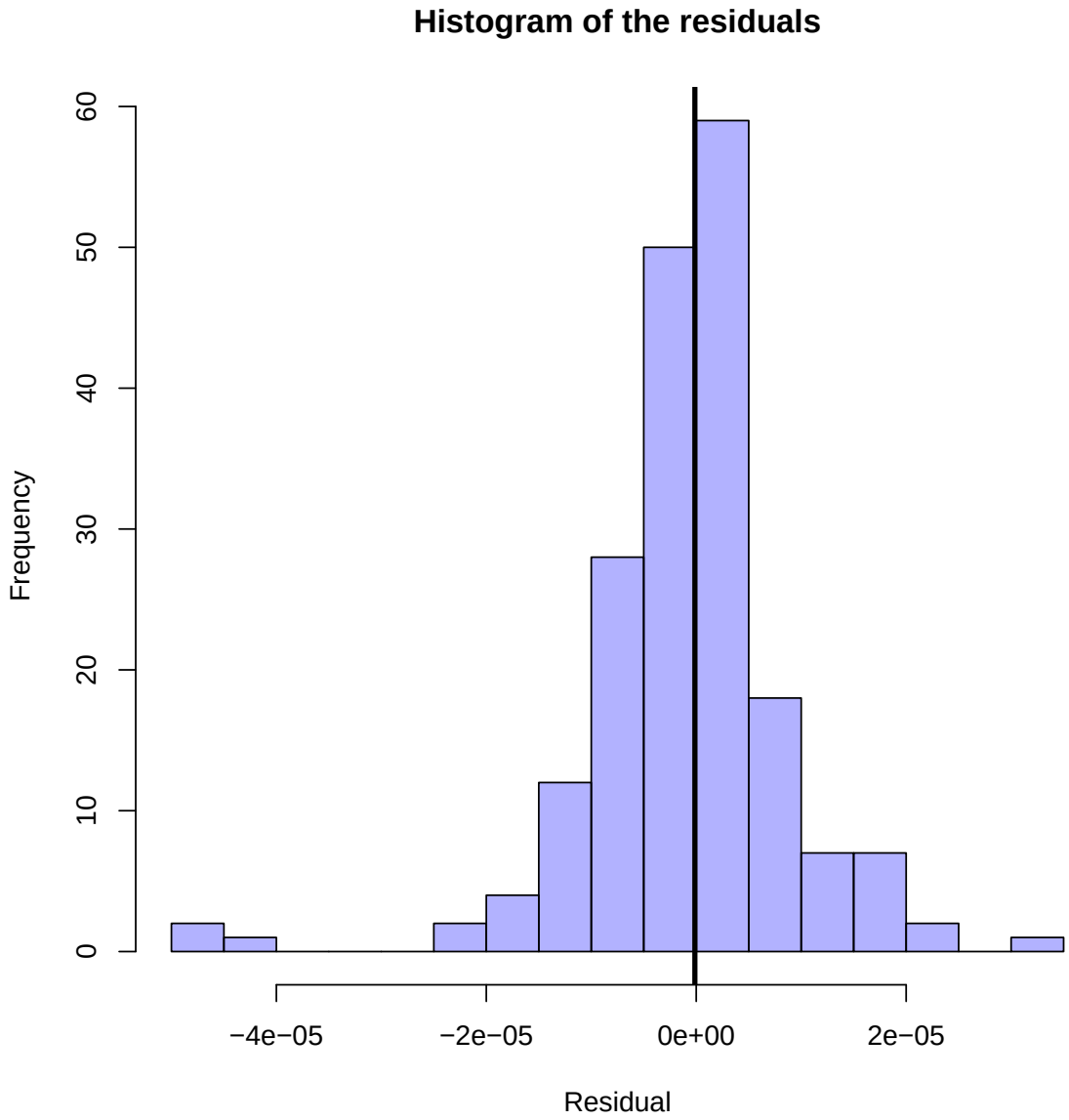
Second model - 1817-2009

Checking the residuals



2 different models

Second model - 1817-2009



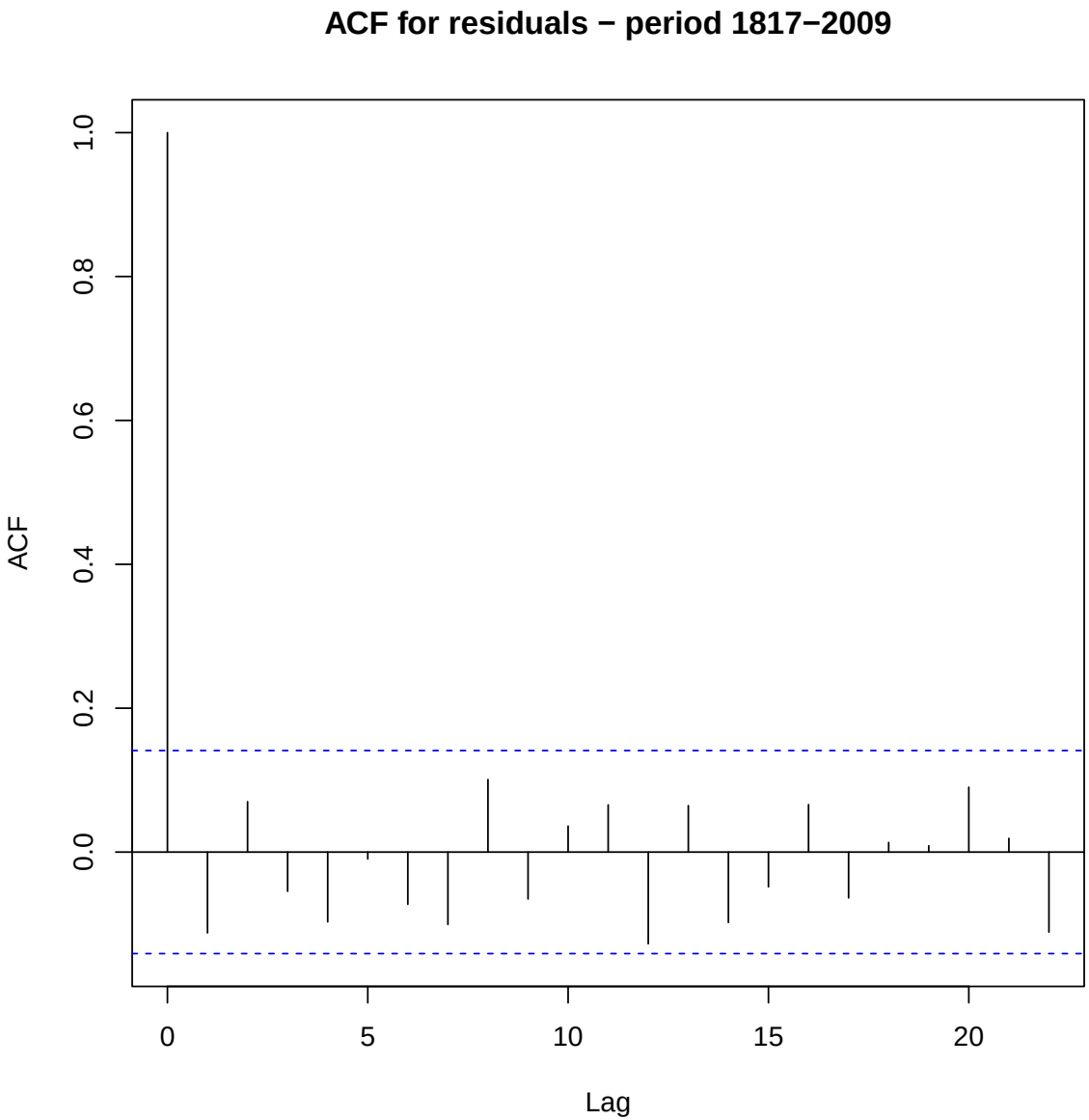
2 different models

Second model - 1817-2009

```
shapiro.test(resid)

##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.89939, p-value = 3.876e-10
```

2 different models

Second model - 1817-2009



2 different models

Second model - 1817-2009

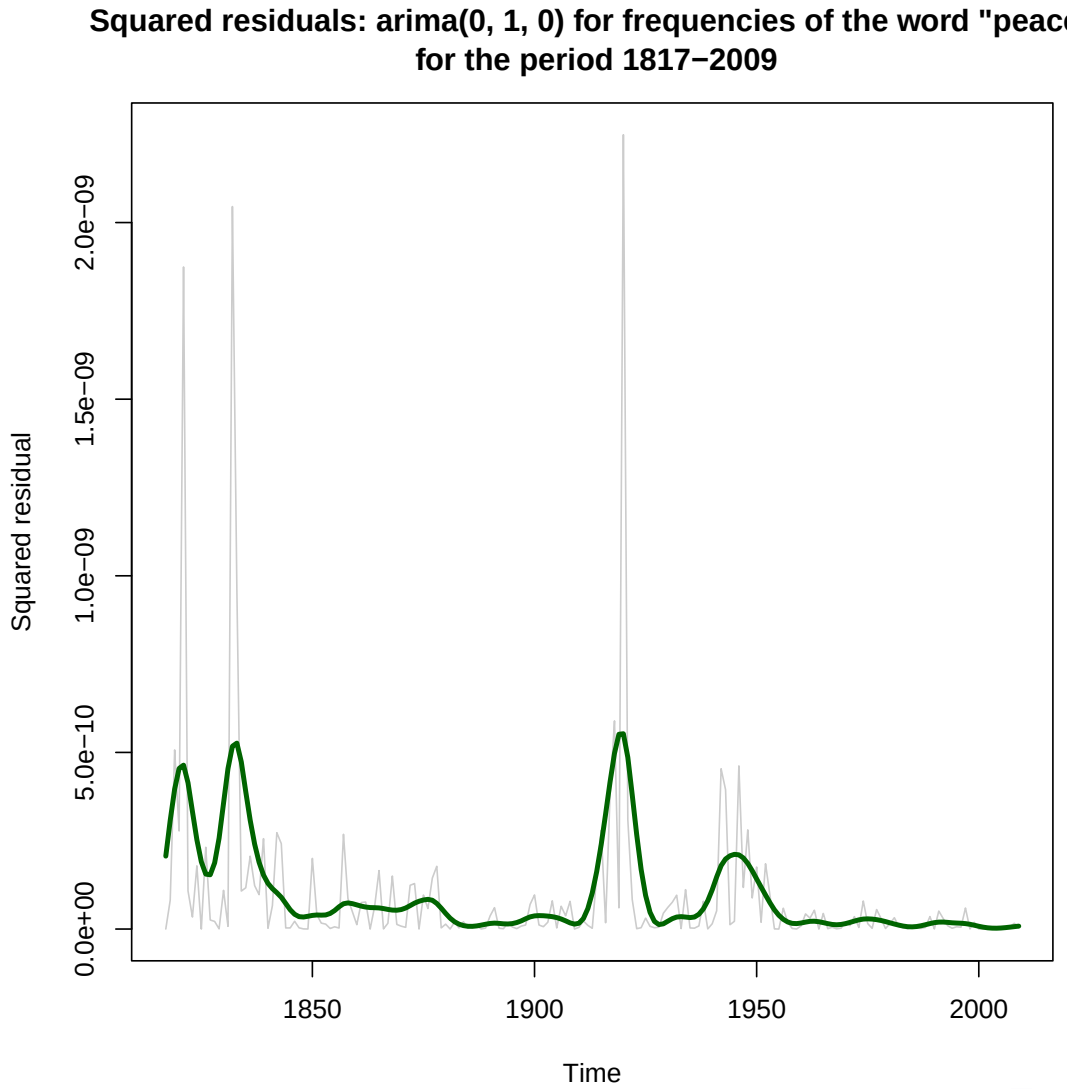
```
Box.test(resid, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  resid
## X-squared = 2.474, df = 1, p-value = 0.1157

# H0: indep./uncorr.
```

Residuals are independent/uncorrelated.

2 different models

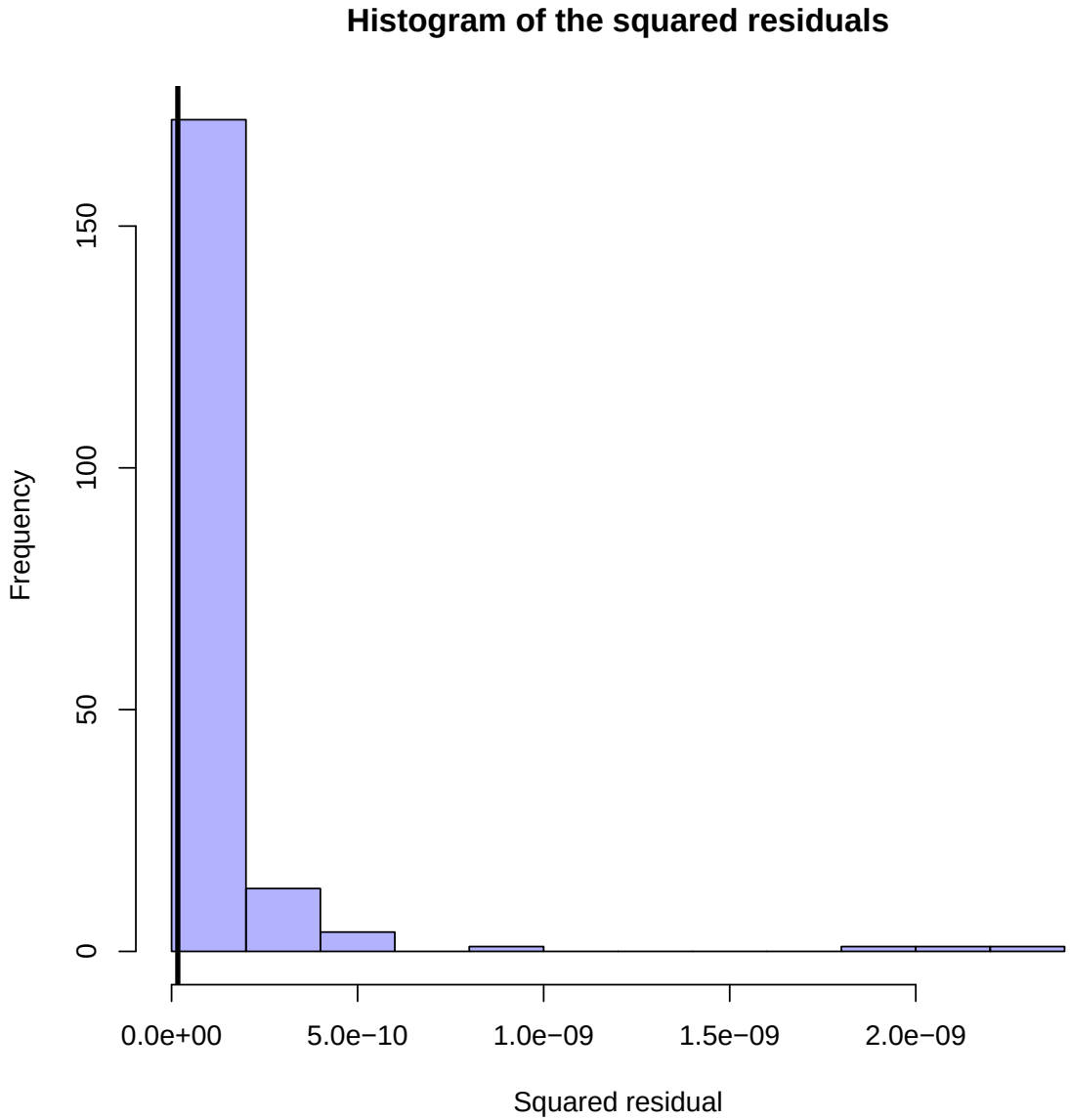
Second model - 1817-2009

Checking the squared residuals



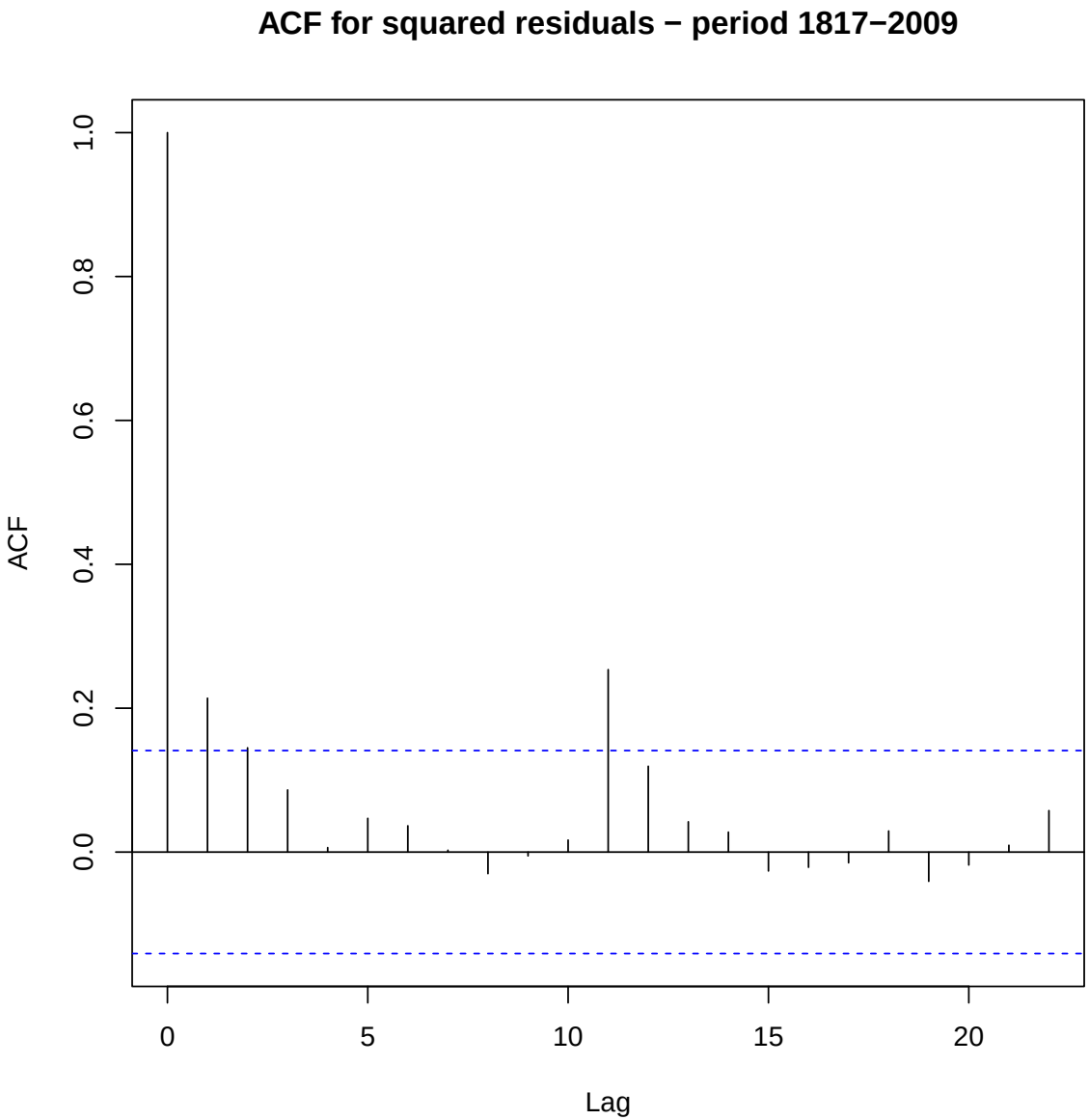
2 different models

Second model - 1817-2009



2 different models

Second model - 1817-2009



2 different models

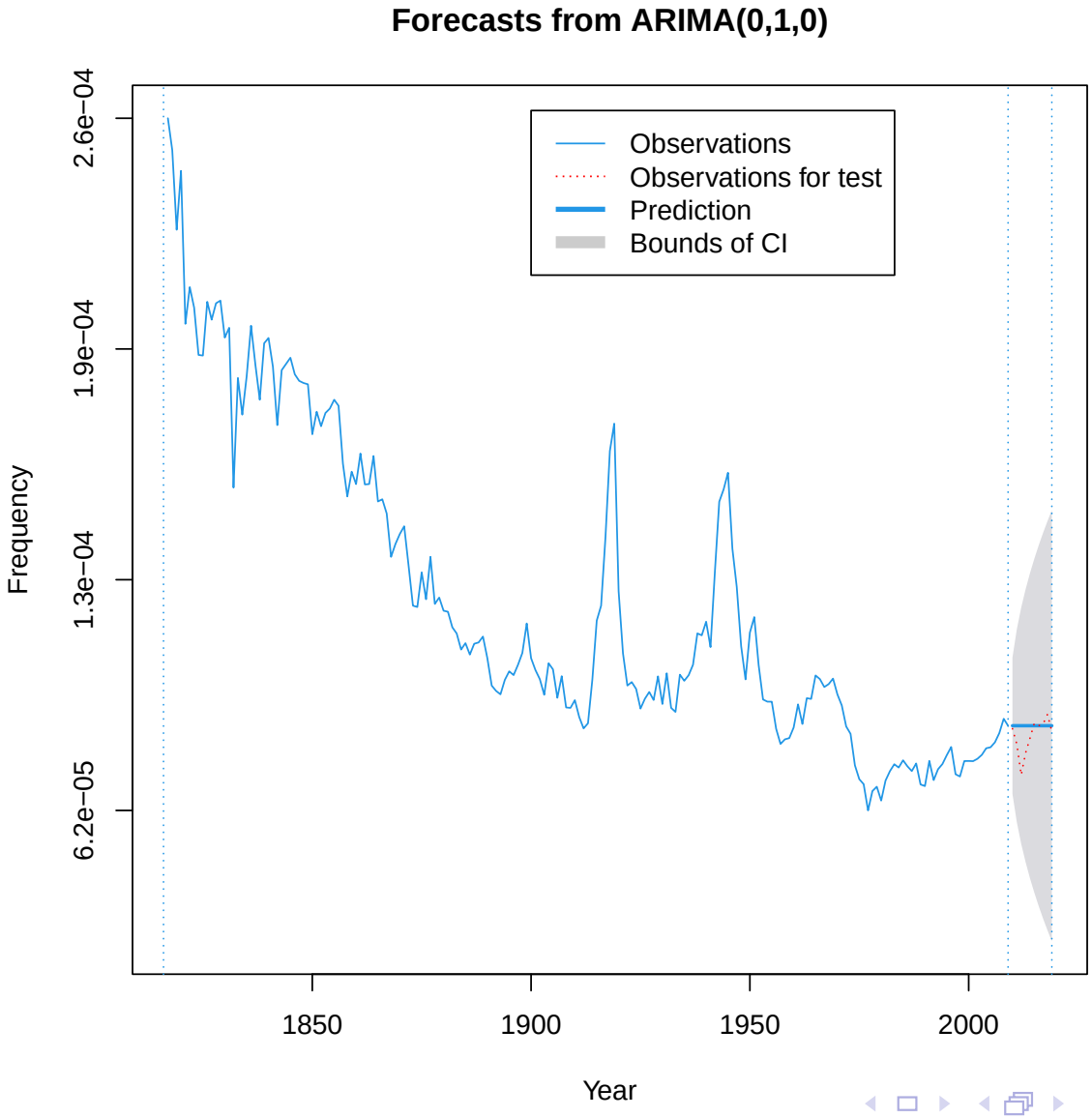
Second model - 1817-2009

```
Box.test(abs(resid)^2, type = "Ljung-Box")

##
##  Box-Ljung test
##
## data:  abs(resid)^2
## X-squared = 8.9696, df = 1, p-value = 0.002745

# H0: indep./uncorr.
```

Squared residuals are not independent/uncorrelated.

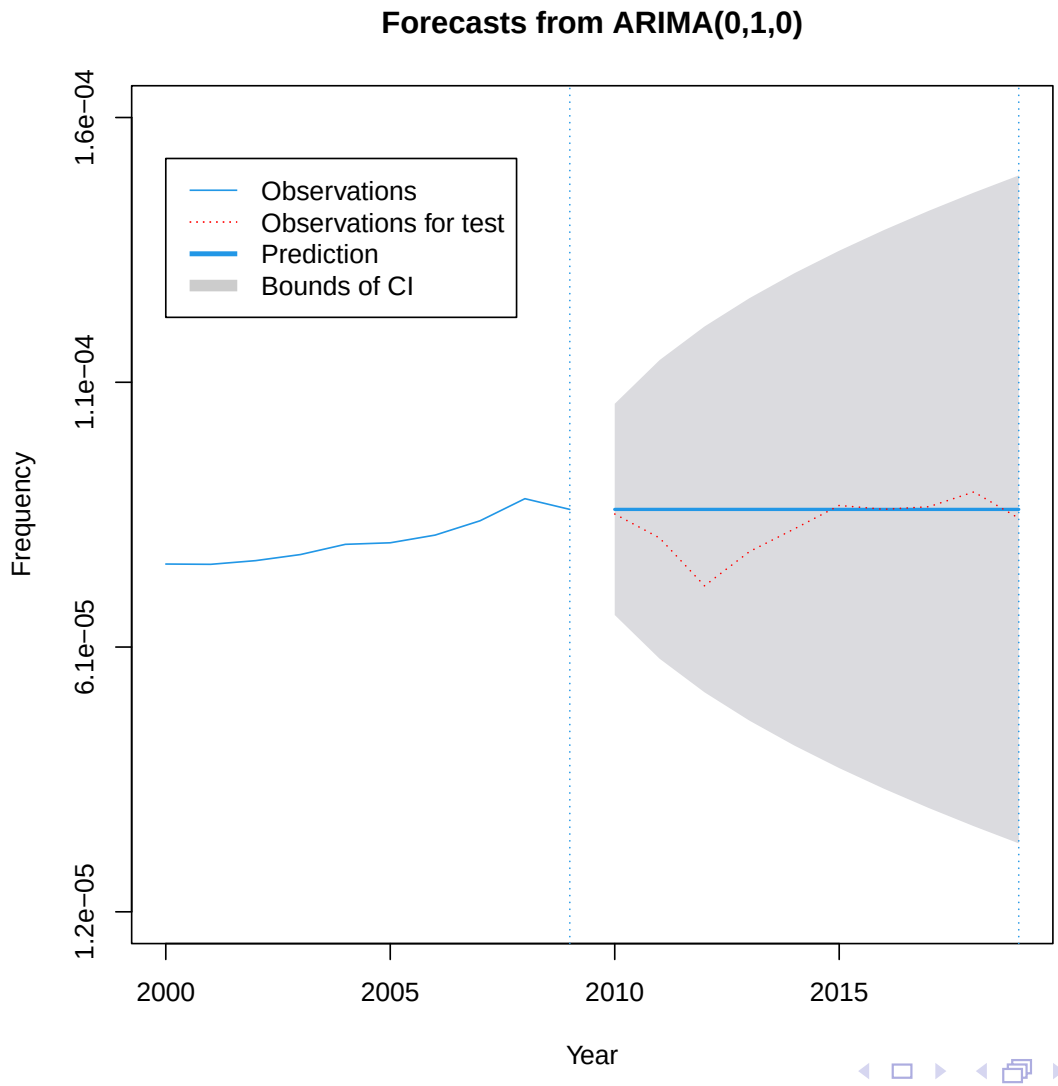
Squared residuals are correlated, but the mean model is a random walk. So, we decided not to extend the modeling to a more complex model like GARCH for example.

Forecast using the model for the second regime - period 1817-2009 -, forecasting 2010-2019



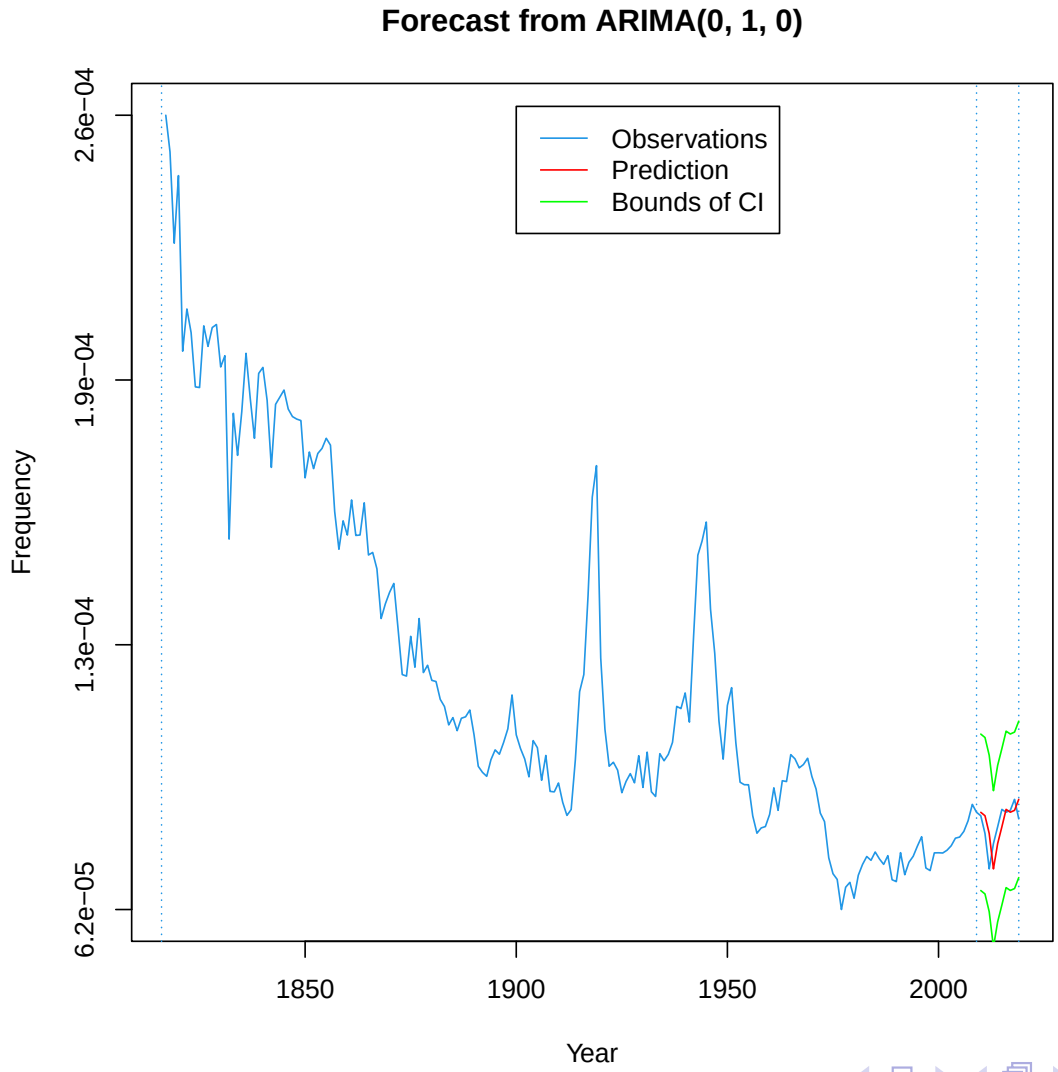
Forecast using the model for the second regime - period 1817-2009 -, forecasting 2010-2019

A closer look at the forecast



Forecast using the model for the second regime - period 1817-2009 -, forecasting 2010-2019

One-day ahead forecasting



Forecast using the model for the second regime - period 1817-2009 -, forecasting 2010-2019

One-day ahead forecasting

A closer look at the forecast.