Time Series Analysis

Christian FRANCQ Jean-Michel ZAKOIAN

CREST

Chapter 3: Using ARMA and SARIMA Models

Outline

Preliminary transformations

Before fitting an ARMA model to a real time series, one has to check that stationarity is a plausible assumption.

Preliminary transformations are called for if the data exhibit visible deviations from stationarity (trend, seasonality...).

Ordinary and seasonal differentiations are used to suppress trends and seasonalities.
The logarithmic transformation, $Y_t = \log X_t$ for $X_t > 0$, may allow to reduce some nonlinearities (e.g. heteroskedasticity). For macroeconomic (financial) series, growth rates (log-returns) $Y_t = \log X_t / X_{t-1}$ are often used. More generally, one can use the Box-Cox transformation:

$$f_\lambda(X_t) = \begin{cases} (X_t^\lambda - 1)/\lambda, & X_t \geq 0,\ \lambda > 0 \text{ or } X_t > 0,\ \lambda \neq 0 \\ \log(X_t) & X_t > 0,\ \lambda = 0 \end{cases}$$

Different steps

ARMA(p,q):

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t - \psi_1 \epsilon_{t-1} - \cdots - \psi_q \epsilon_{t-q} + c.$$

i) A priori identification of orders p and q (acf, pacf, ...);
ii) Estimation of the parameters (LS, ML);
iii) Diagnostic checking
iv) Model choice (AIC, BIC...).

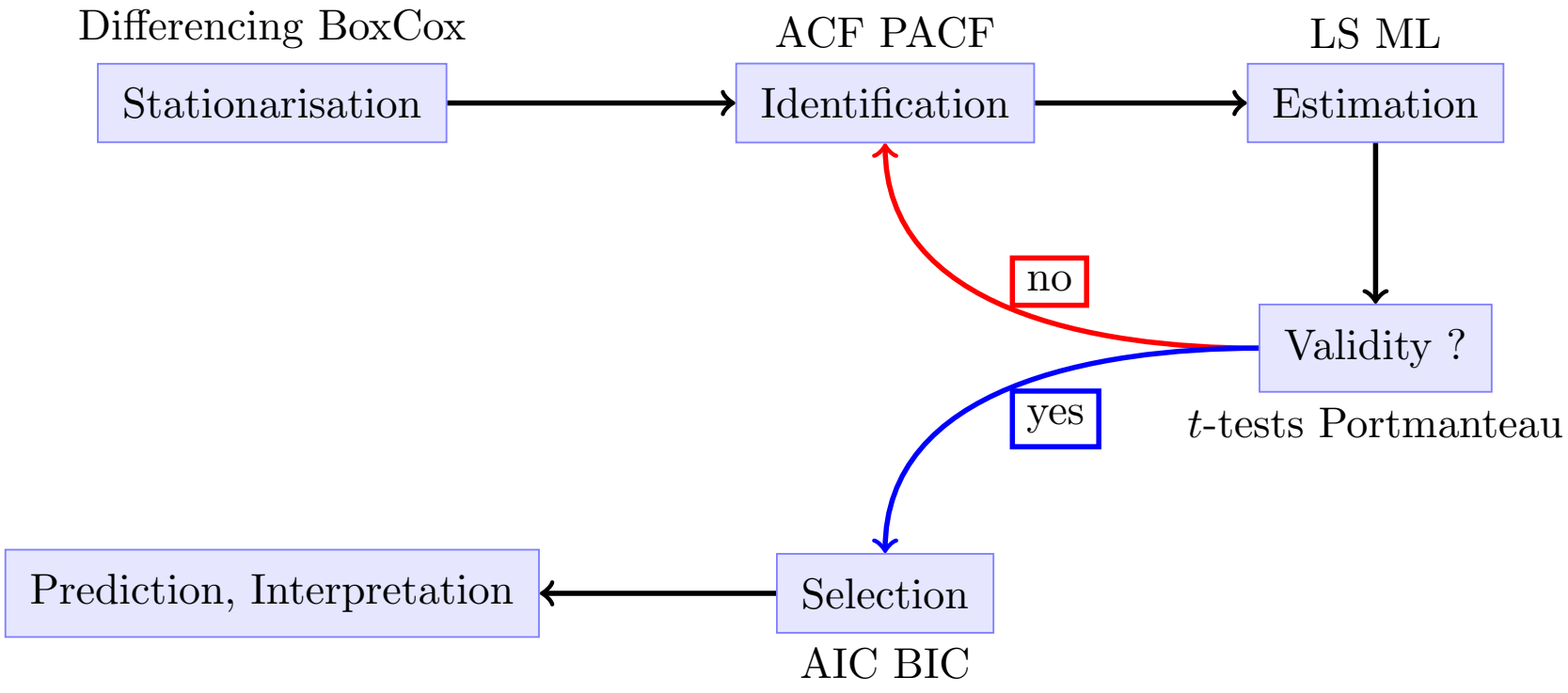


Figure: Box-Jenkins methodology

Ordinary differentiation

If no visible deviation from stationarity occurs and if the empirical autocorrelations decrease rapidly, an ARMA model can be fitted on the series after mean correction.

Otherwise we look for a transformation of the series achieving such properties. If the series exhibits a trend but no seasonality, differences can be used.

Differencing leads to consider the class of autoregressive-integrated moving average, or ARIMA models.

Definition

Let d a nonnegative integer. A series (X_t) is an ARIMA(p,d,q) if

$$Y_t = (1-B)^d X_t$$

is a causal ARMA(p,q), possibly with a mean μ.

We thus have

$$\phi(B)(1-B)^d X_t = \psi(B)\epsilon_t + c, \quad \epsilon_t \text{ WN } (0,\sigma^2)$$

where ϕ is a polynomial of degree p with all its roots of modulus >1 and ψ is a polynomial of degree q.

(X_t) is stationary only if $d=0$.

Remarks

If $d \geq 1$, any polynomial of degree $d-1$ can be added to X_t without changing the model.* In particular, ARIMA are thus appropriate for series with a trend.

ARIMA can also be appropriate for series without a deterministic trend (a random walk for instance).

Estimation of the parameters is thus based on the differenced series $(1-B)^d X_t$. Supplementary assumptions are required for prediction of the original series.

*By induction on d, check that $(1-B)^d(a_0 + a_1 t + \cdots + a_{d-1} t^{d-1}) = 0$

Prediction based on ARIMA models

Suppose that

$$(I-B)^d X_t = Y_t, \quad t \geq 1$$

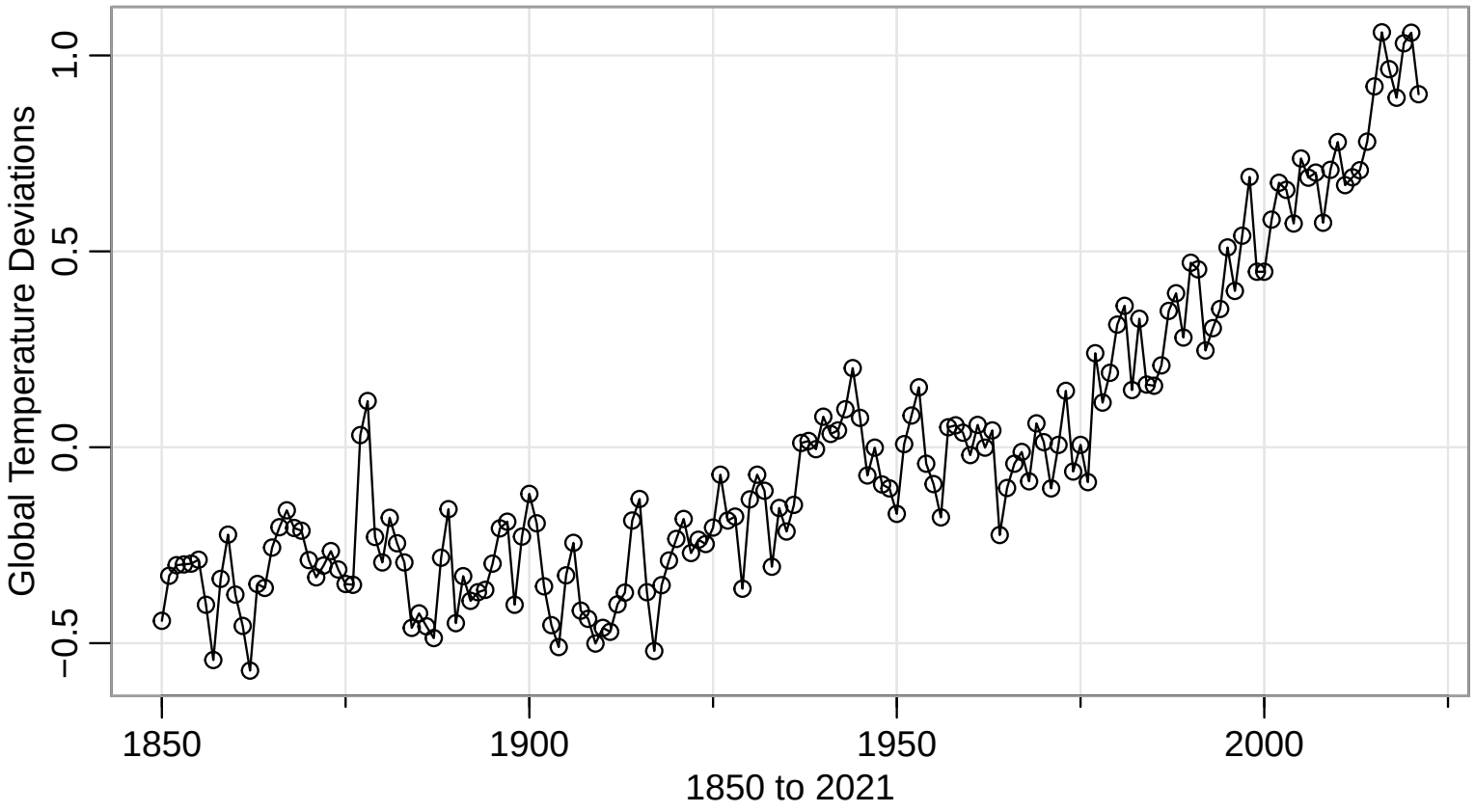
where (Y_t) is a causal ARMA(p,q) and that the vector $(X_{1-d},\ldots,X_0)$ is non correlated with $(Y_t, \quad t \geq 1)$. We thus have

$$X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j}, \quad t \geq 1.$$

It can be assumed that $(X_{1-d},\ldots,X_n)$ are observed (and thus also $(Y_1,\ldots,Y_n)$). We then have

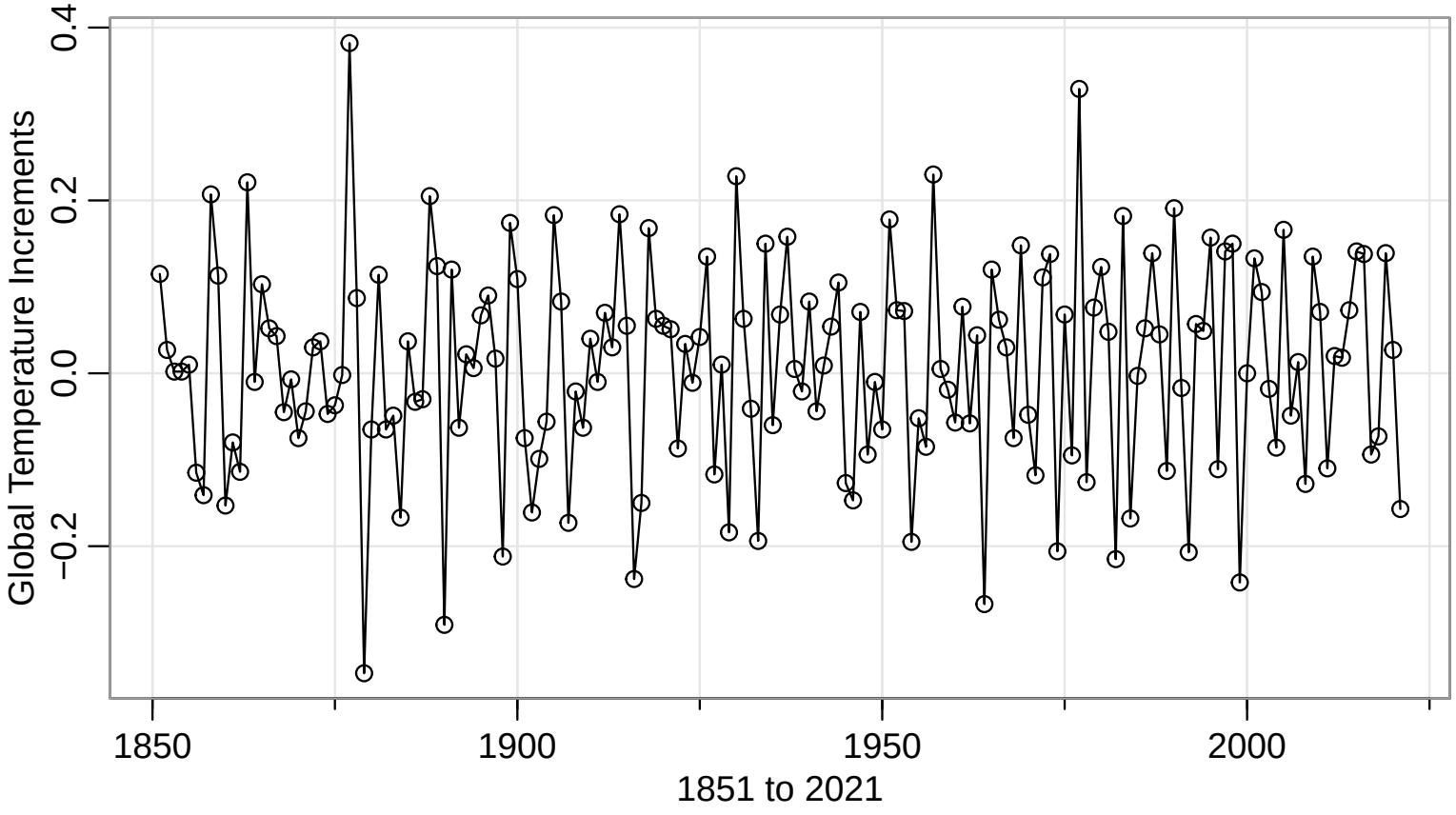
$$\hat{X}_{n+1|n} = \hat{Y}_{n+1|n} - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{n+1-j}.$$

Global temperature deviations (from 1951-1980 average)

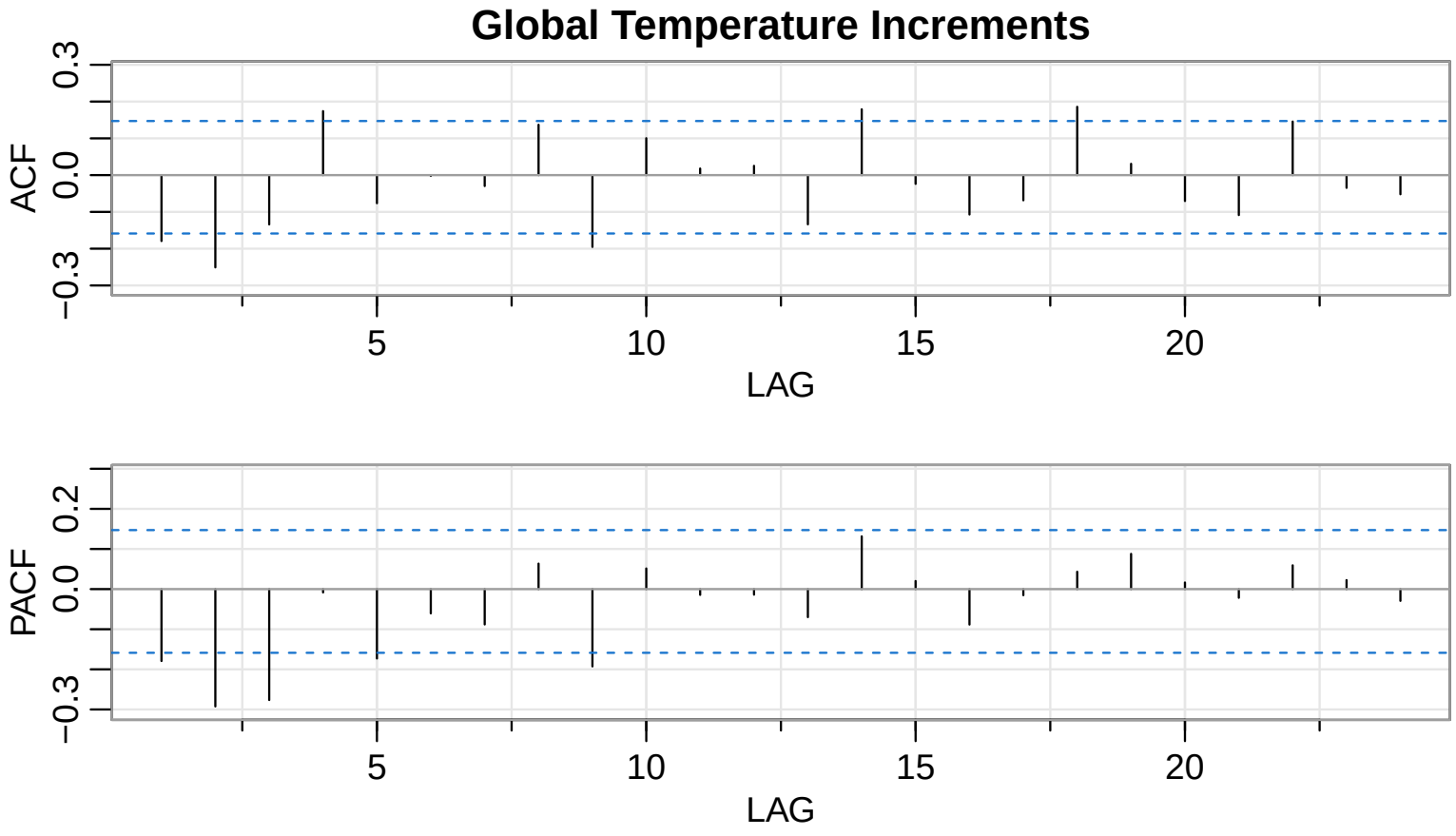


Source: `http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_summary.txt`

The differentiated series seems stationary



The ACF and PACF suggest trying MA(2), MA(4), AR(3) and mixed ARMA models



Using the R packages astsa, forecast and portes

An ARMA(3,3) seems too complicated, all the ARMA(1,3) are significant

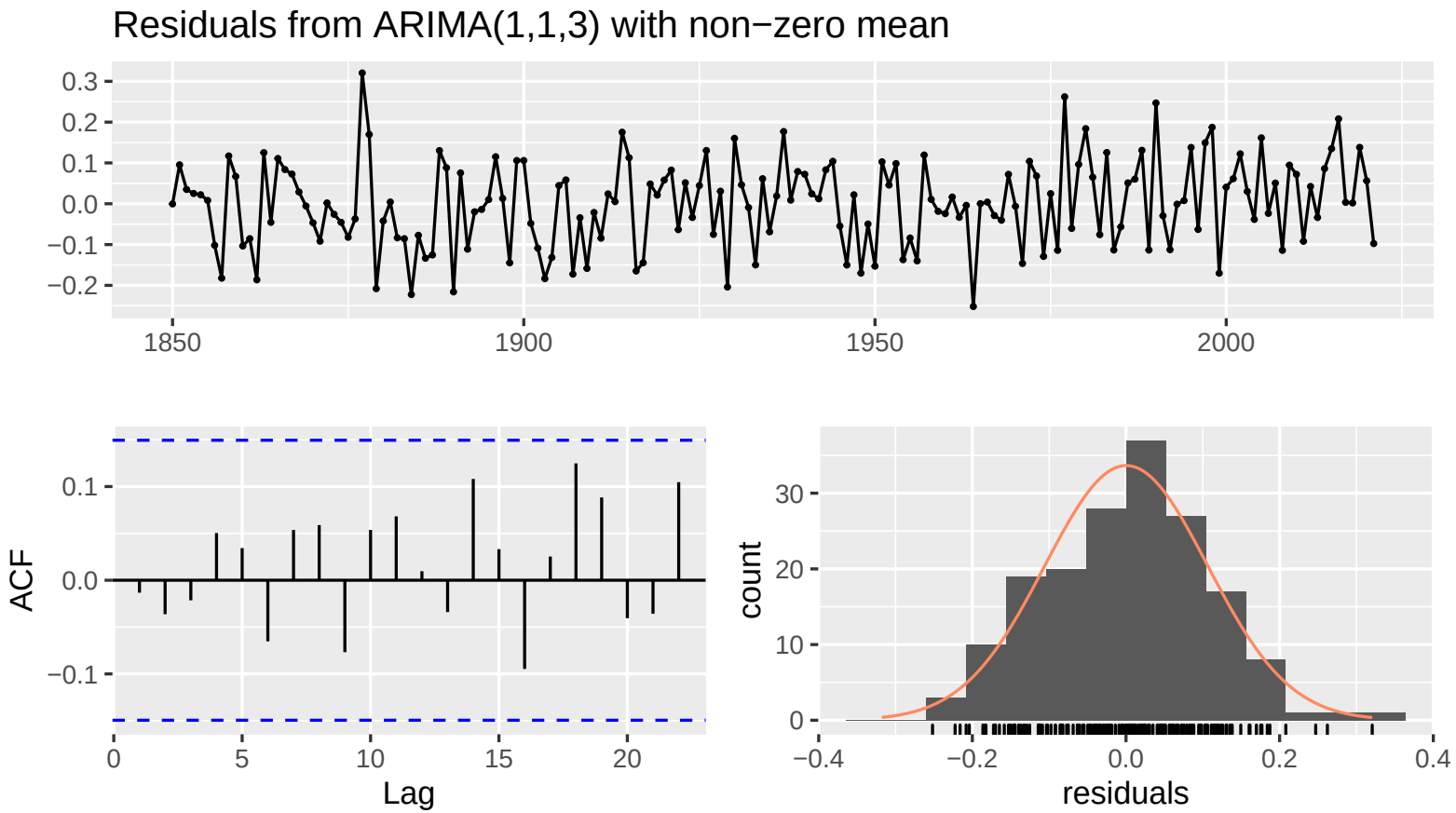
```
> res33<-sarima(GlobalTemp,3,1,3,details=FALSE)
> res33$ttable
         Estimate     SE  t.value p.value
ar1       -0.8165 0.2748 -2.9719  0.0034
ar2       -0.0558 0.1806 -0.3092  0.7575
ar3       -0.1730 0.1775 -0.9746  0.3312
ma1        0.5033 0.2770  1.8169  0.0711
ma2       -0.5601 0.1132 -4.9486  0.0000
ma3       -0.2276 0.2306 -0.9870  0.3251
constant   0.0078 0.0029  2.7233  0.0072
>
> res13<-sarima(GlobalTemp,1,1,3,details=FALSE)
> res13$ttable
         Estimate     SE  t.value p.value
ar1       -0.9458 0.0426 -22.2209  0.0000
ma1        0.6492 0.0799   8.1298  0.0000
ma2       -0.6137 0.0666  -9.2146  0.0000
ma3       -0.3966 0.0674  -5.8868  0.0000
constant   0.0078 0.0027   2.8684  0.0047
```

Using the R packages astsa, forecast and portes

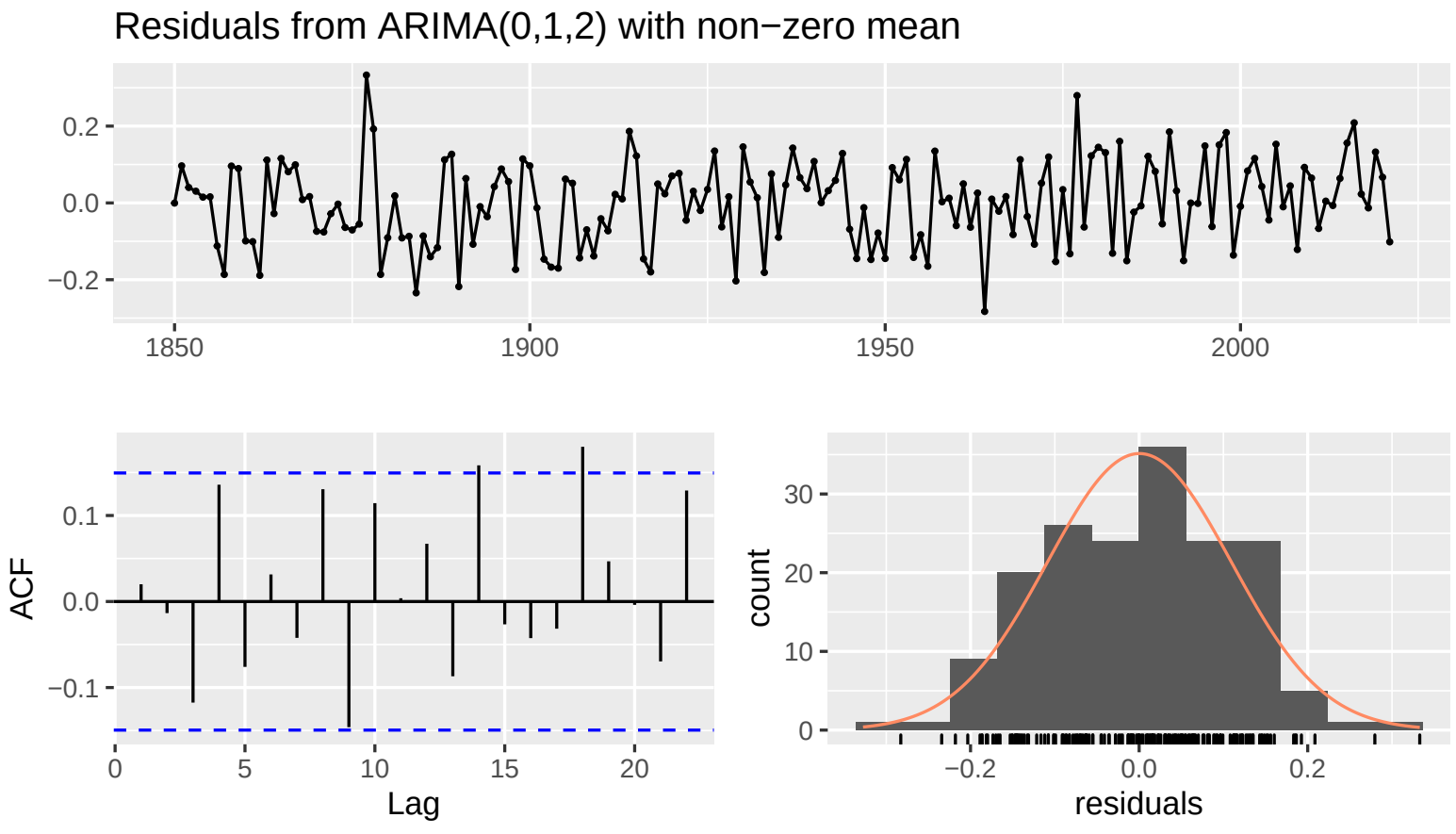
The portmanteau tests do not reject the ARMA(1,3), but the MA(2) is rejected

```
> LjungBox(res13$fit)
 lags  statistic  df   p-value
    5  1.007065   1  0.3156069
   10  4.543316   6  0.6035682
   15  8.067863  11  0.7072122
   20 14.784980  16  0.5404397
   25 18.647001  21  0.6077732
   30 23.059241  26  0.6295934
> res02<-sarima(GlobalTemp,0,1,2,details=FALSE)
> res02$ttable
          Estimate     SE  t.value p.value
ma1        -0.3873 0.0677  -5.7242   0.000
ma2        -0.2822 0.0625  -4.5190   0.000
constant    0.0078 0.0028   2.7844   0.006
> LjungBox(res02$fit)
 lags  statistic  df    p-value
    5  6.863331   3  0.07638385
   10 16.806416   8  0.03218918
   15 23.953359  13  0.03156180
   20 31.212052  18  0.02720280
   25 36.211838  23  0.03924452
   30 40.775812  28  0.05629976
```

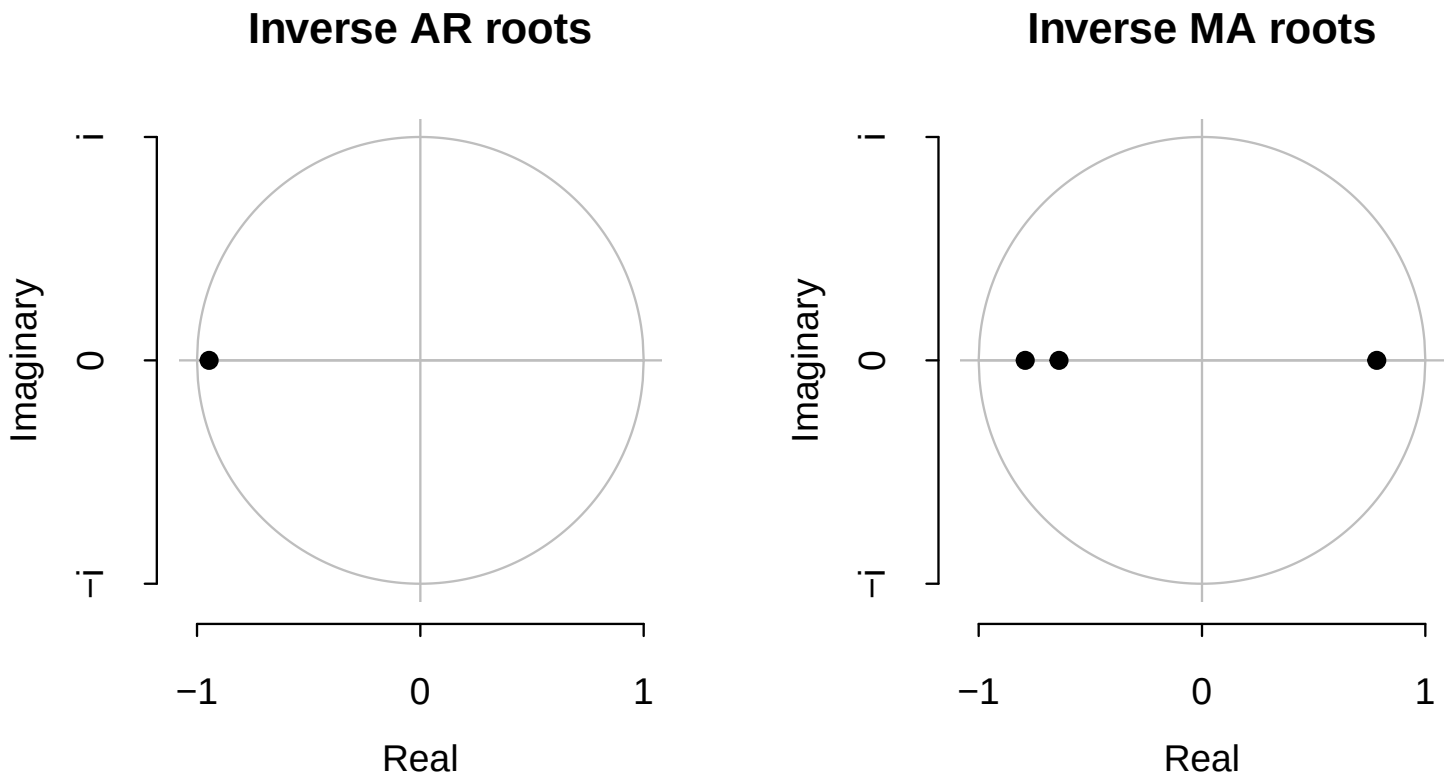
The ARMA(1,3) residuals resemble a white noise



The MA(2) residuals have few autocorrelations outside the significance bounds



Not sure the ARMA(1,3) can be simplified

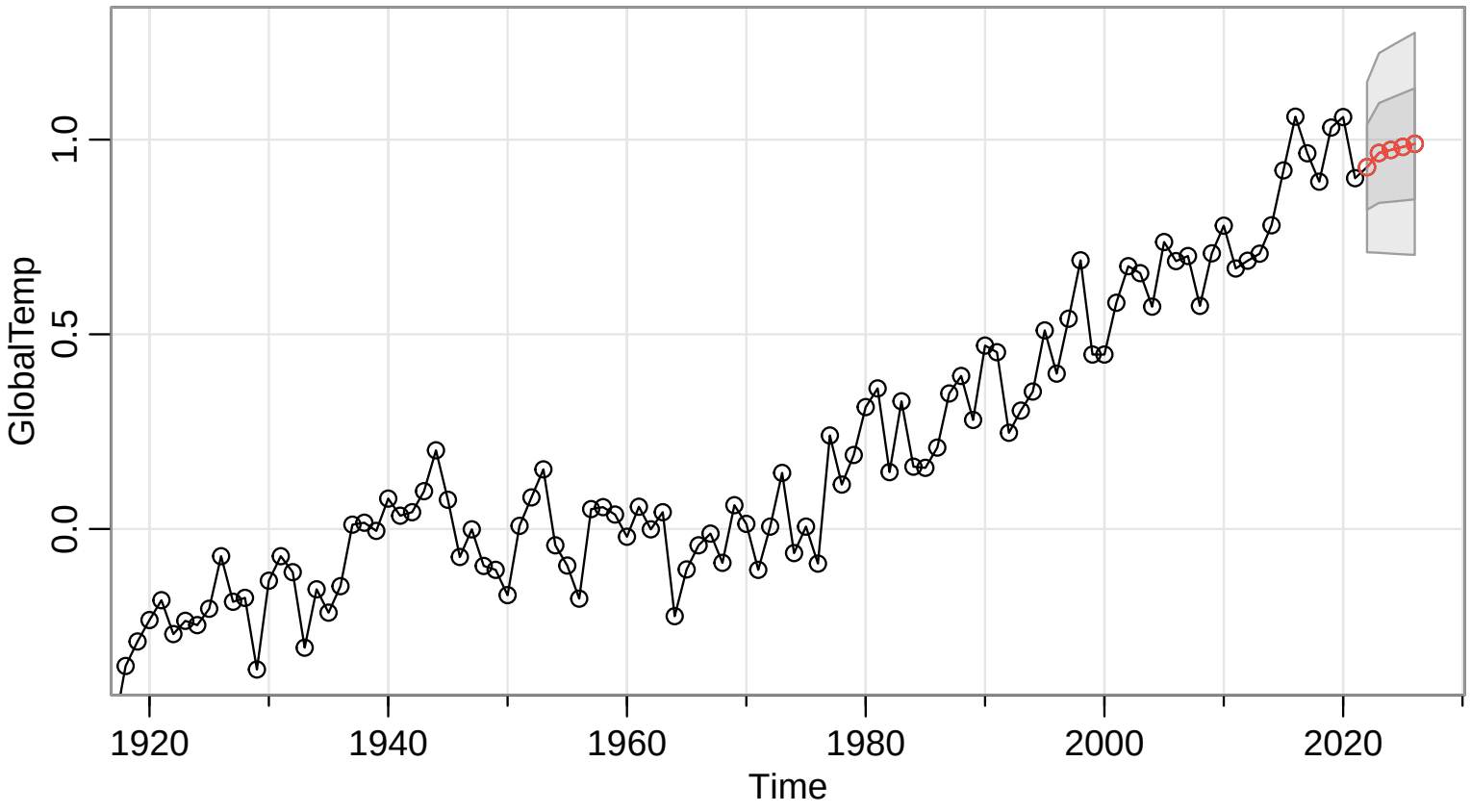


AIC selects the ARMA(1,3) and BIC the MA(2)

```
> AIC(res02$fit,res04$fit,res13$fit,res33$fit,res23$fit,res12$fit,
           df       AIC
res02$fit  4 -263.2663
res04$fit  6 -264.6961
res13$fit  6 -269.1751
res33$fit  8 -266.0326
res23$fit  7 -263.7410
res12$fit  5 -261.7083
res30$fit  5 -261.6468
res24$fit  8 -268.2935
>
> BIC(res02$fit,res04$fit,res13$fit,res33$fit,res23$fit,res12$fit,
           df       BIC
res02$fit  4 -250.6997
res04$fit  6 -245.8461
res13$fit  6 -250.3251
res33$fit  8 -240.8993
res23$fit  7 -241.7494
res12$fit  5 -246.0000
res30$fit  5 -245.9385
res24$fit  8 -243.1602
```

Forecasts of the ARIMA(0,1,2) model

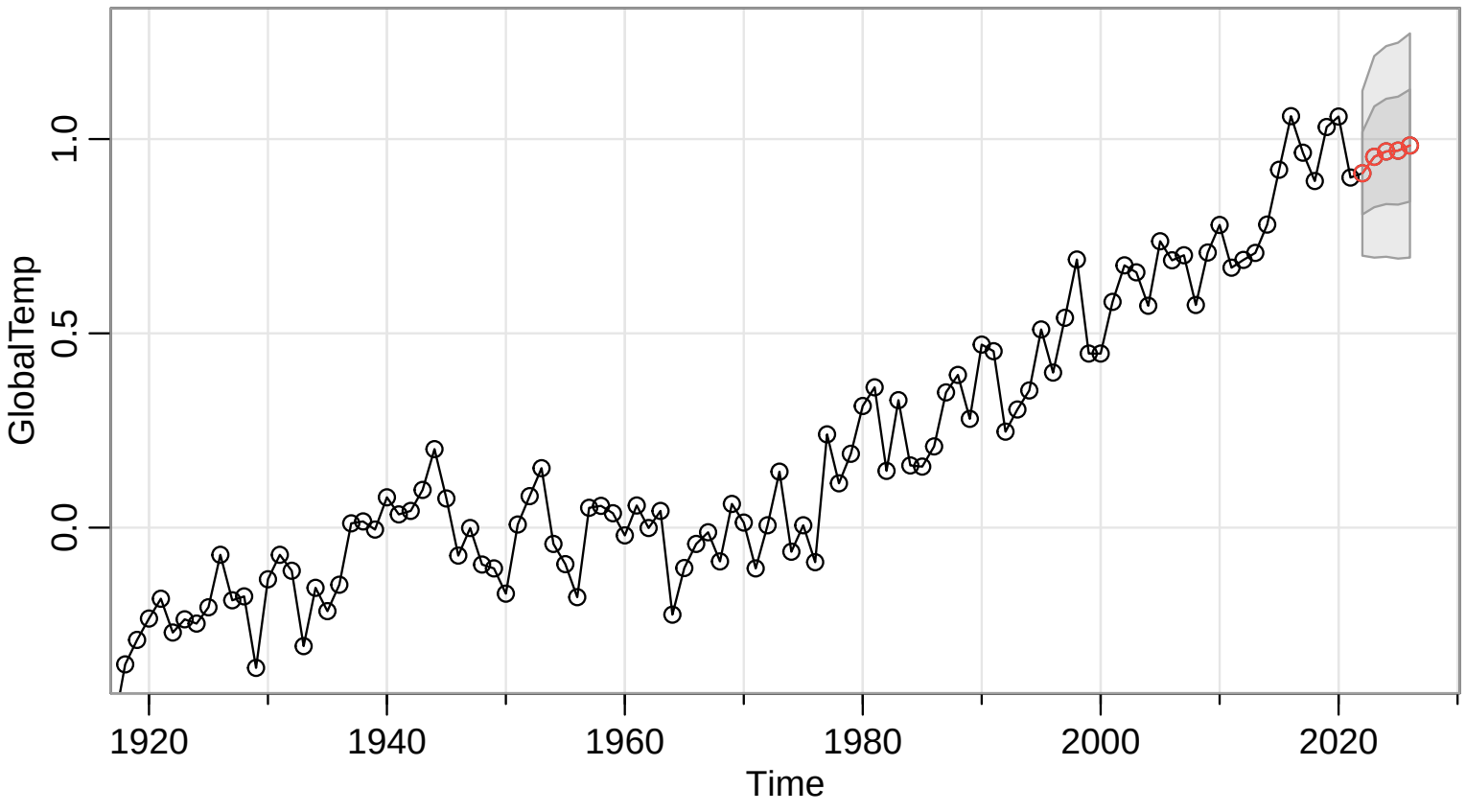
$\hat{X}_{2022} = 0.929(0.109)$



$$Y_t := X_t - X_{t-1} = \epsilon_t - \underset{(0.07)}{0.39}\epsilon_{t-1} - \underset{(0.06)}{0.28}\epsilon_{t-2} + \underset{(0.0028)}{0.0078}$$

Forecasts of the ARIMA(1,1,3) model

$\hat{X}_{2022} = 0.912(0.106)$



Non stationary seasonal series

Many monthly (Leisure employment) and quarterly series present strong seasonality, respectively of order $s = 12$ and $s = 4$.

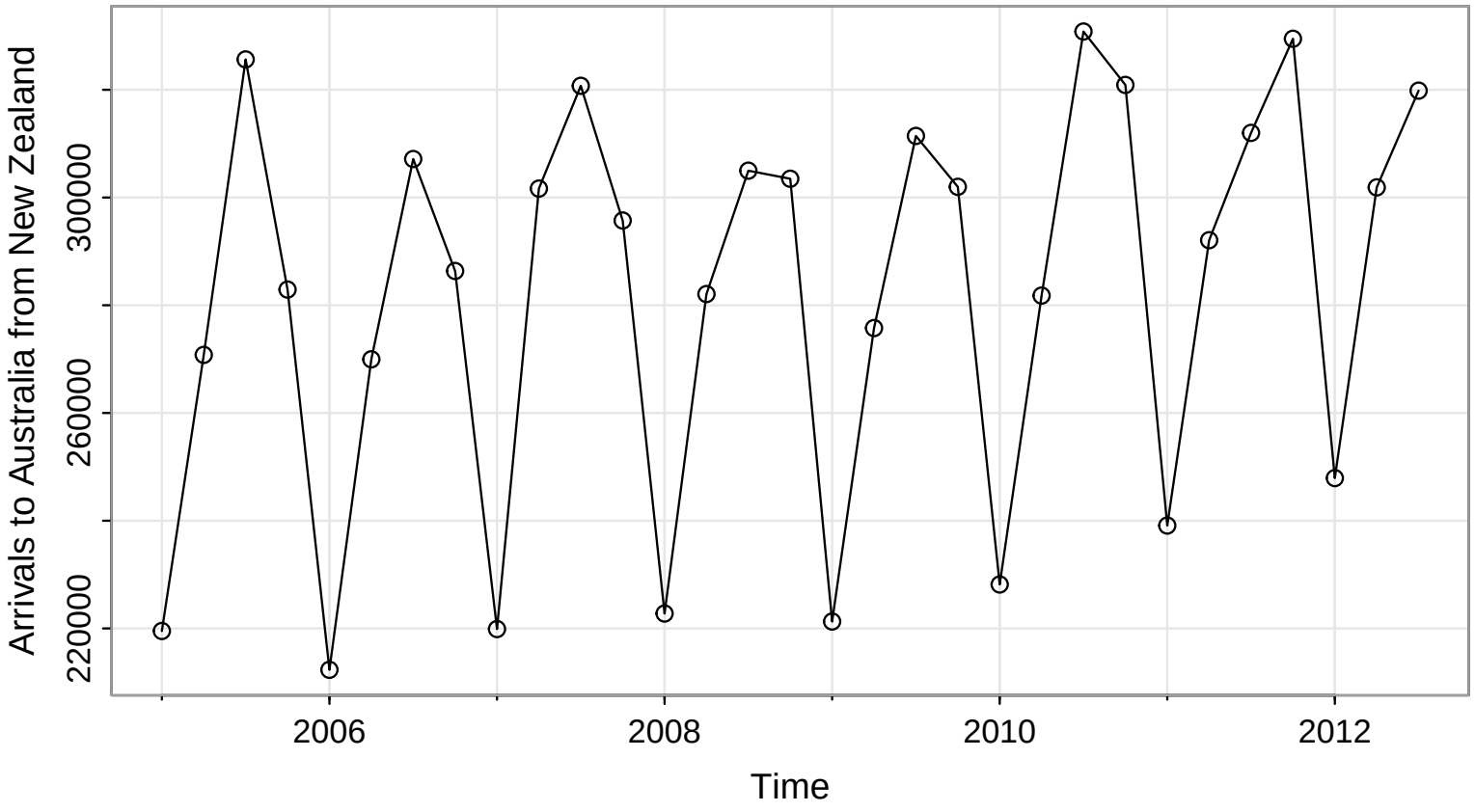
Each date $t = ks + \nu$ correspond to a cycle/year k and a season/month $\nu \in \{1, \ldots, s\}$. If the distribution of X_t depends on s, the series (X_t) is not stationary.

A SARIMA model for the non-stationary series X_t is an ARMA model on the stationary series

$$Y_t = (1 - B)^d (1 - B^s)^D X_t \sim ARMA.$$

For parsimony reasons, the AR and MA polynomials are generally product of polynomials in B and B^s.

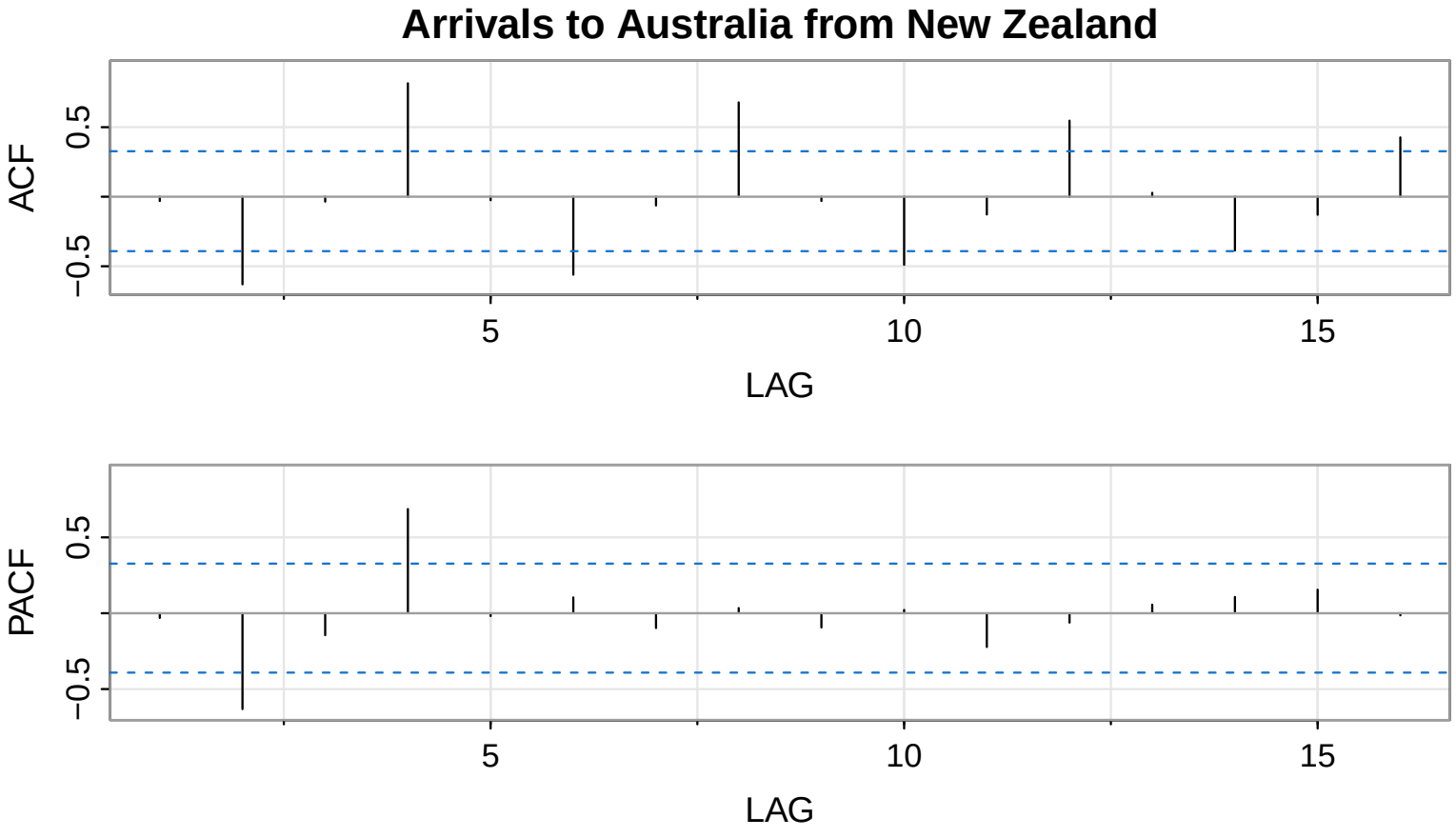
An example of seasonal quarterly series



Source: "fpp3" R package

The theoretical ACF and PACF do not exist

Strong empirical autocorrelations at lags multiples of s



Seasonal ARIMA models

SARIMA models allow to handle a random seasonal component.

Suppose that observations of a monthly time series are available over r years

	Month			
Year	1	2	...	12
1	X_1	X_2	...	X_{12}
2	X_{13}	X_{14}	...	X_{24}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$X_{1+12(r-1)}$	$X_{2+12(r-1)}$	...	$X_{12+12(r-1)}$

Same ARMA model for each column

Each column is viewed as a series generated by the same ARMA(P,Q) model. For the jth month we have:

$$\begin{aligned} X_{j+12t} &= \phi_1 X_{j+12(t-1)} + \ldots + \phi_P X_{j+12(t-P)} \\ &\quad + U_{j+12t} - \psi_1 U_{j+12(t-1)} + \ldots + \psi_Q U_{j+12(t-Q)}. \end{aligned}$$

We thus have

$$\begin{aligned} X_t &= \phi_1 X_{t-12} + \ldots + \phi_P X_{t-12P} \\ &\quad + U_t - \psi_1 U_{t-12} + \ldots + \psi_Q U_{t-12Q}, \end{aligned}$$

that is,

$$\Phi(B^{12}) X_t = \Psi(B^{12}) U_t.$$

Links between the columns

It would not be realistic to assume that the 12 series corresponding to the months are non correlated.
The process (U_t) is not a WN in general: $Cov(U_t U_{t+h}) \neq 0$.

To account for these correlations, the process (U_t) is modeled by an ARMA(p,q):

$$\phi(B)\, U_t = \psi(B)\epsilon_t$$

where $\{\epsilon_t,\ t \in Z\}$ is a WN with variance σ^2.

Global model

Combining the different models - and differencing if necessary - we have

Definition:

A SARIMA$(p,d,q)(P,D,Q)_s$ process with period s is any process (X_t) such that

$$Y_t = (I-B)^d(I-B^s)^D X_t$$

is a causal ARMA:

$$\phi(B)\Phi(B^s)\,Y_t = \psi(B)\Psi(B^s)\epsilon_t + c$$

where ϕ, Φ, ψ, Ψ are polynomials of degrees p, P, q, Q respectively.

Identifying a SARIMA from the autocorrelations

1. Search d **(0, 1 or 2)** and D **(0 or 1)** such that

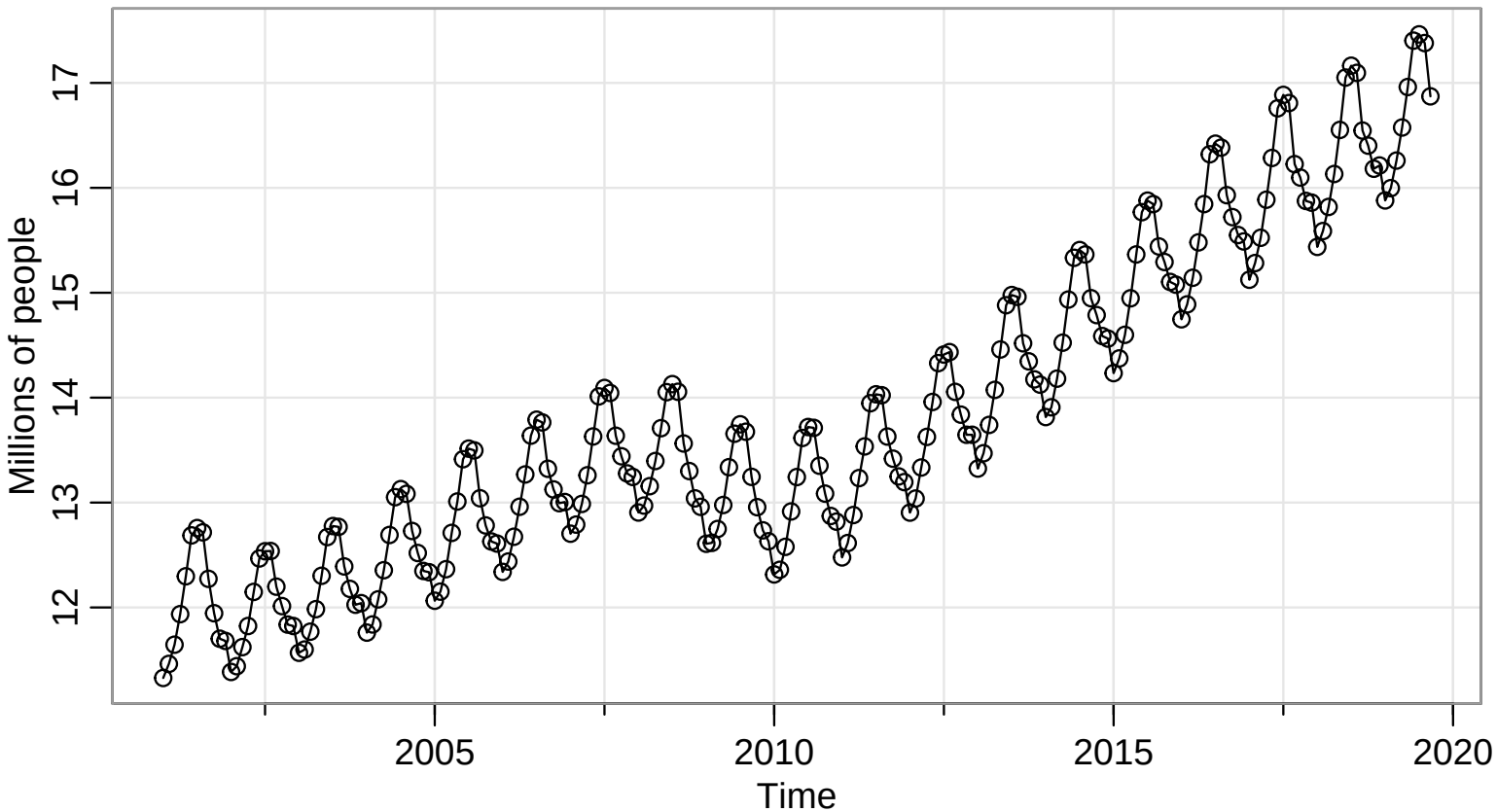
$$Y_t = (I - B)^d (I - B^s)^D X_t$$

looks stationary.

2. Examine the empirical autocorrelations and partial autocorrelations of (Y_t) at lags multiple of s. Identify the orders P and Q.
3. Select the orders p and q by comparing $\hat{\rho}(1), \ldots, \hat{\rho}(s-1)$ to the autocorrelations of an ARMA(p, q)
4. For p, P, q, Q and d fixed, estimate ϕ, Φ, ψ, Ψ and σ^2 by LS. The series (Y_t) is an ARMA$(p+sP, q+sQ)$ with some coefficients equal to zero. The parameters are thus estimated by specifying the multiplicative relations of the model.
5. Use information criteria and validity checks to choose the most appropriate SARIMA models.

Monthly Leisure and Hospitality employments in US

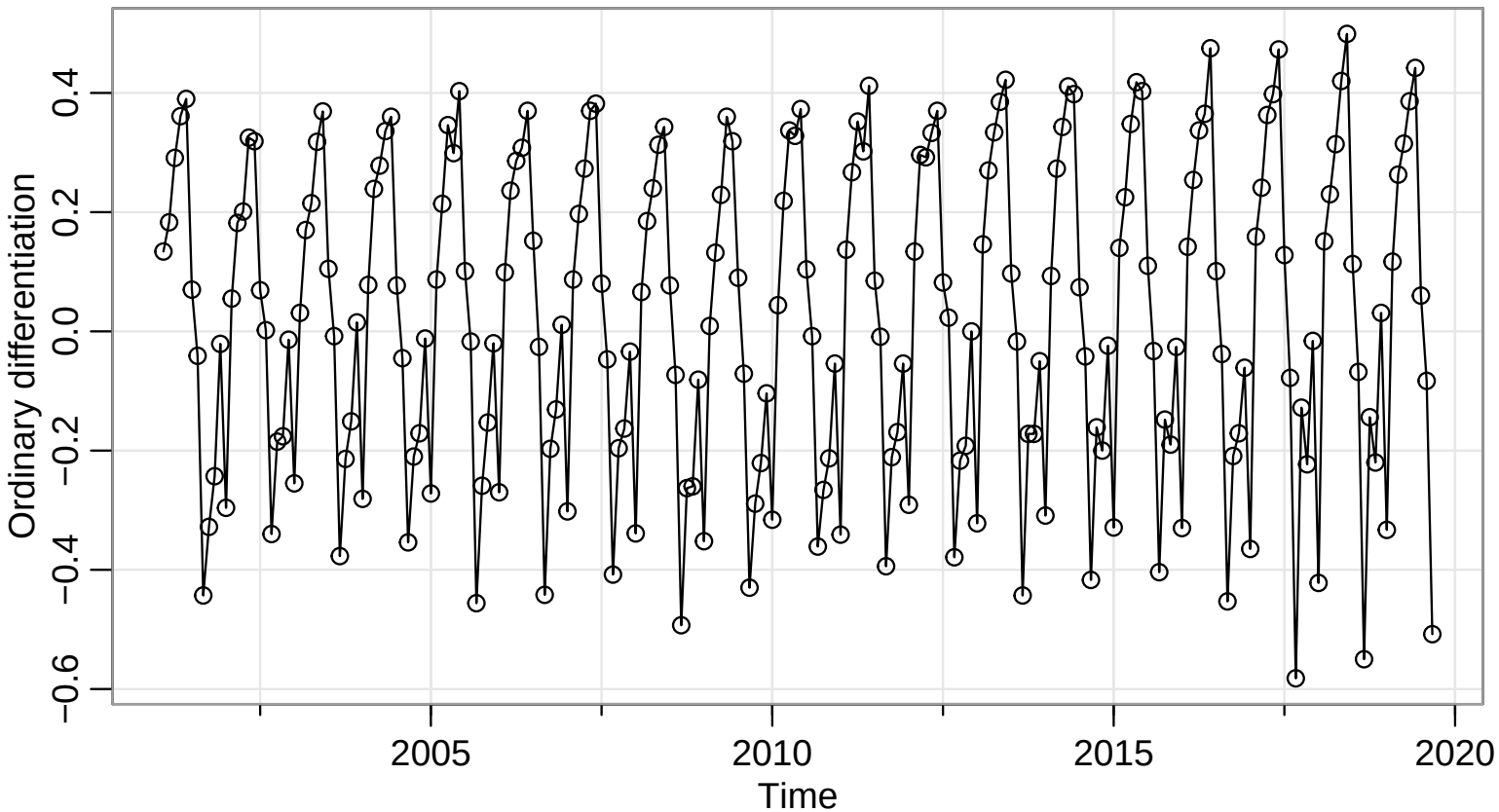
Due to trend and seasonality, the series is not stationary



Source: "fpp3" R package

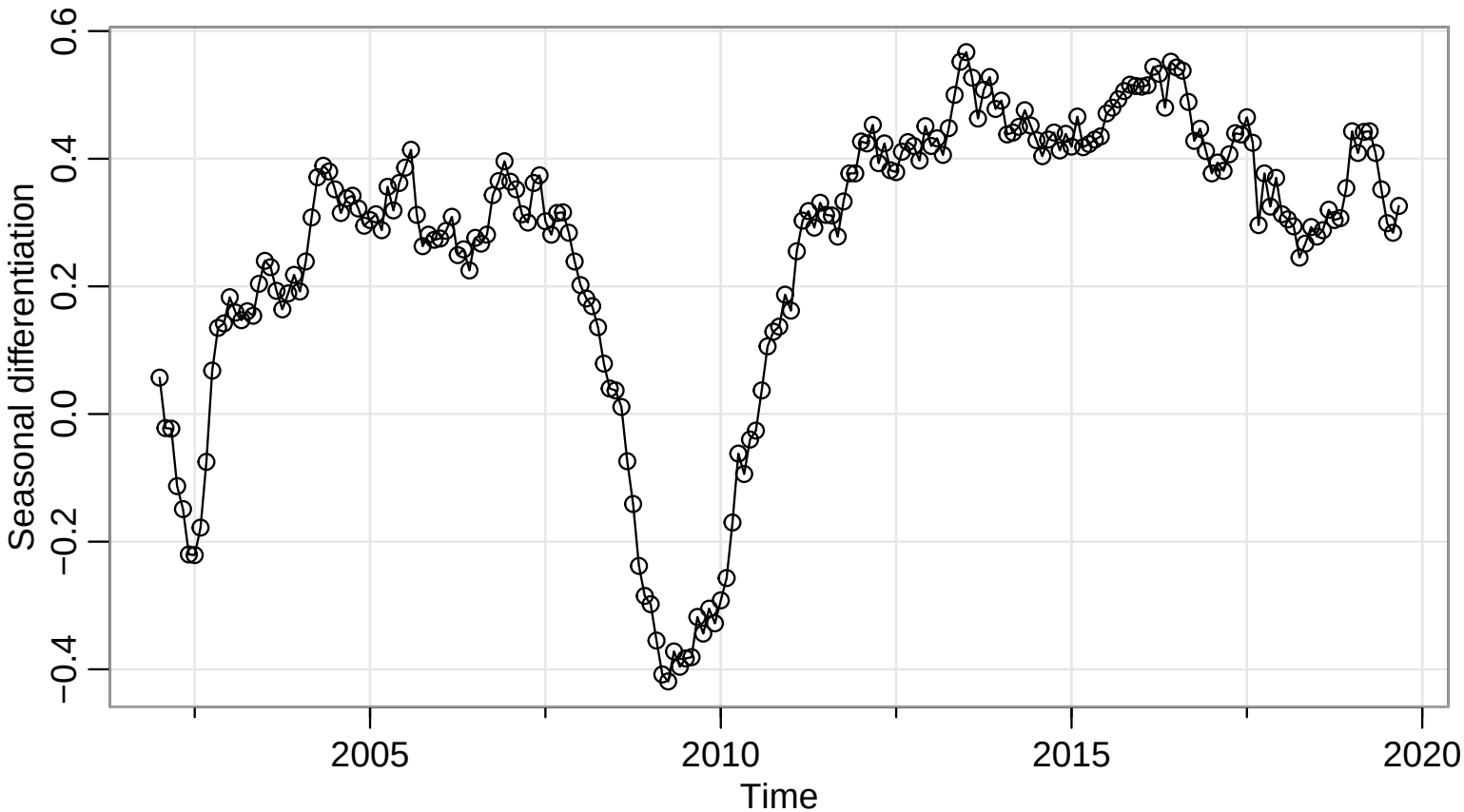
Differentiated series $\nabla X_t = X_t - X_{t-1}$

Due to seasonality, the series is not stationary



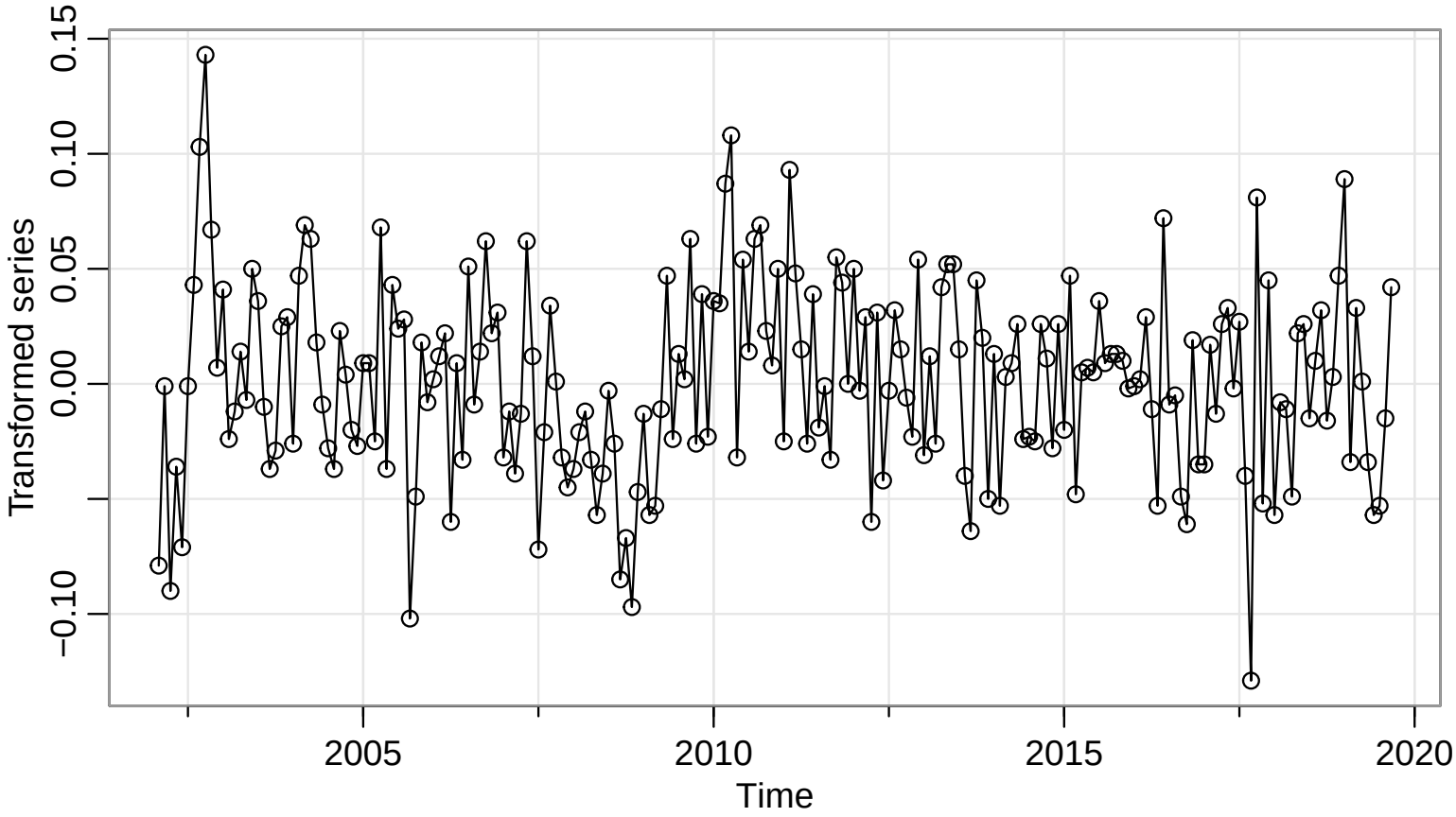
Seasonally differentiated series $\nabla_{12} X_t = X_t - X_{t-12}$

The series looks like a random walk (formal test in Chapter 4)

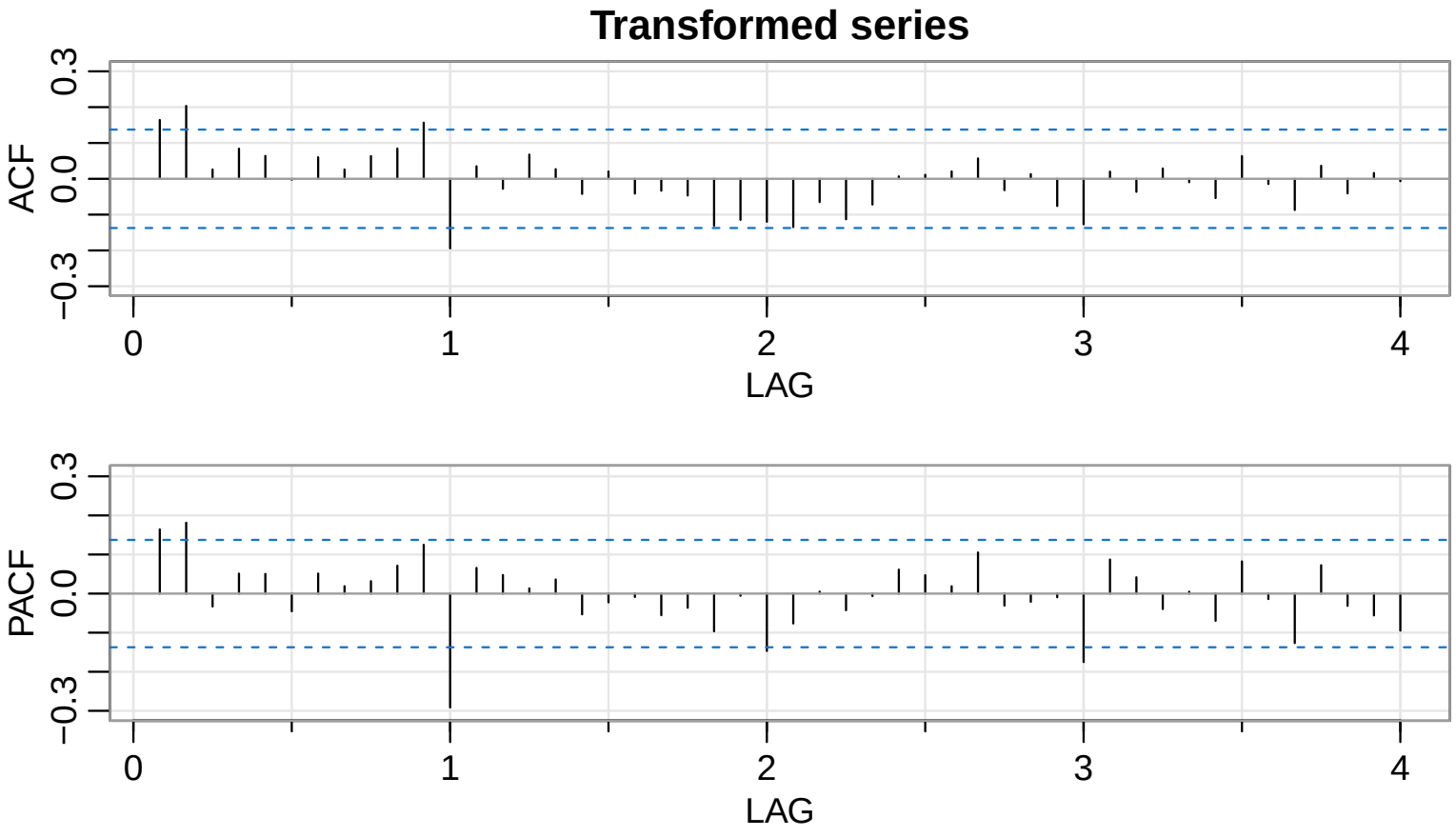


Series transformed by an ordinary and a seasonal differentiation $(1-B)(1-B^{12})X_t$

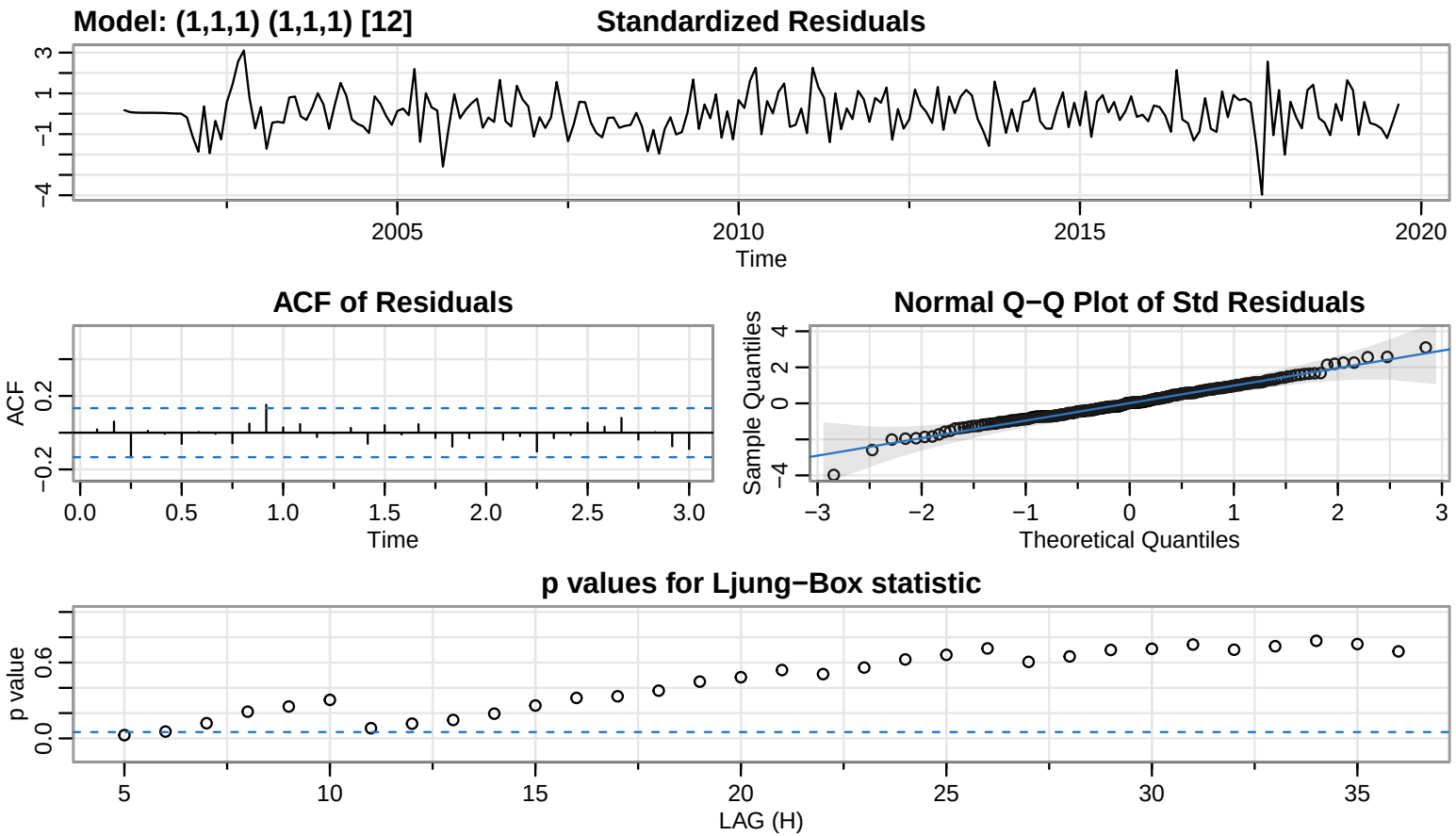
The series no longer shows obvious signs of non-stationarity



ACF and PACF suggest trying mixed ARMA models with polynomials in B^{12} and B



This model is not so bad. Note few outliers in the residuals.

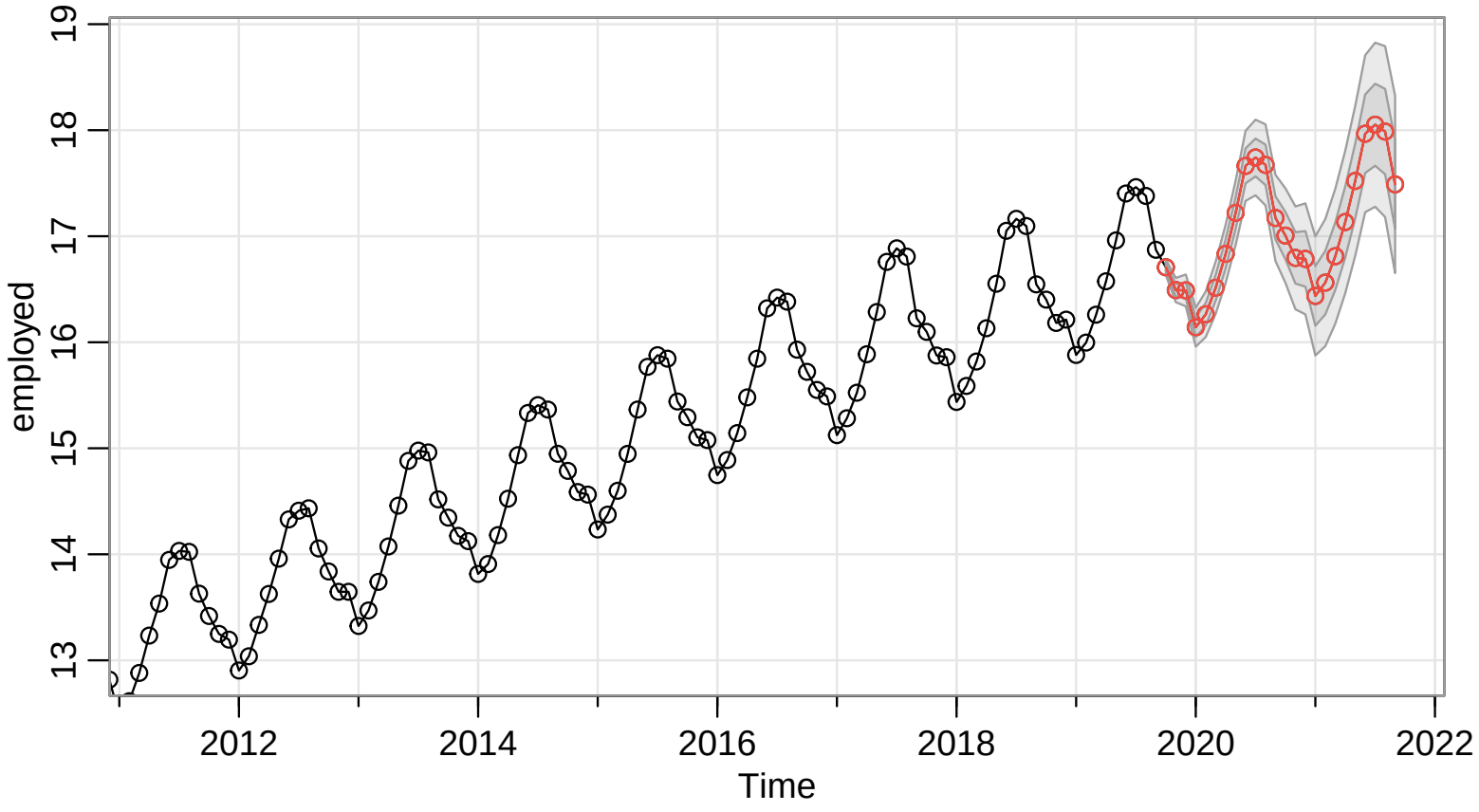


The model is (barely) rejected by the portmanteau test

based on 5 autocorrelations, but this can be explained by the outliers. Moreover, AIC and BIC prefer this model than others.

```
> res$ttable
Estimate     SE t.value p.value
ar1    0.8380 0.1369  6.1217  0.0000
ma1   -0.6742 0.1897 -3.5550  0.0005
sar1   0.2904 0.1308  2.2204  0.0275
sma1  -0.7272 0.0954 -7.6195  0.0000
> LjungBox(res$fit)
lags statistic df    p-value
5   4.972862  1 0.02574801
10  7.176958  6 0.30479197
15 13.530177 11 0.26008867
20 15.559497 16 0.48410393
25 17.810129 21 0.66098609
30 21.626869 26 0.70892867
>
> AIC(res212212$fit,res111111$fit,res212112$fit,res212012$fit,res012012$fit,res012112$fit)
df       AIC
res212212$fit  9 -773.2012
res111111$fit  5 -779.1890
res212112$fit  8 -773.8965
res212012$fit  7 -776.1073
res012012$fit  5 -778.6243
res012112$fit  6 -777.5403
```

Completely wrong forecasts (because of Covid 19!)



$$(1-B)(1-B^{12})(1-\underset{(0.14)}{0.84}B)(1-\underset{(0.13)}{0.29}B^{12})X_t = (1-\underset{(0.19)}{0.67}B)(1-\underset{(0.10)}{0.73}B^{12})\epsilon_t$$

Data Generating Process for the Monte Carlo simulation

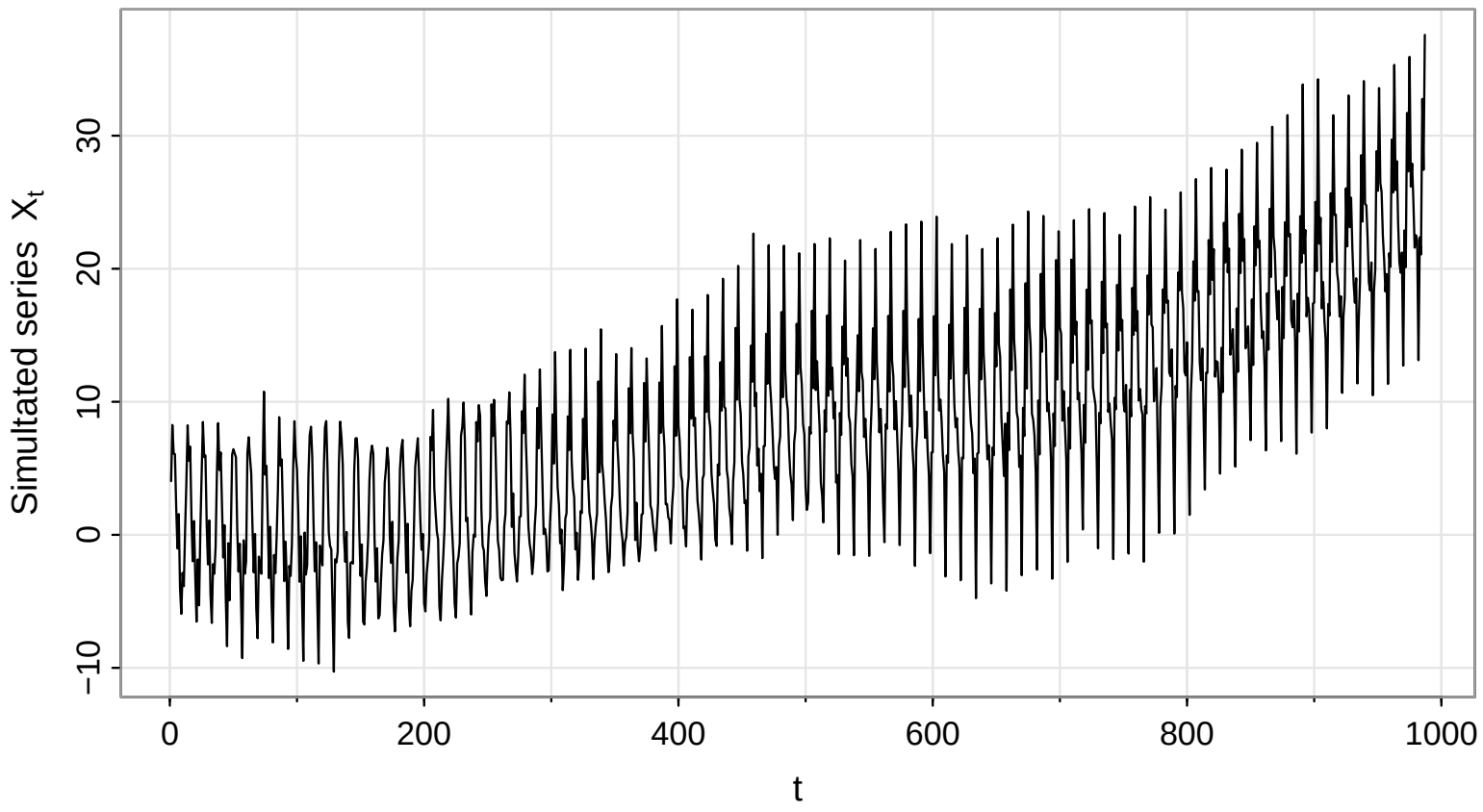
1000 observations simulated from a SARIMA $(1,1,2)(1,1,1)_{12}$ model: $X_t = Z_t + 0.01t + 5\sin(2\pi t/12)$ where

$$(1-0.2B)(1-B)(1-0.4B^{12})(1-B^{12})Z_t$$

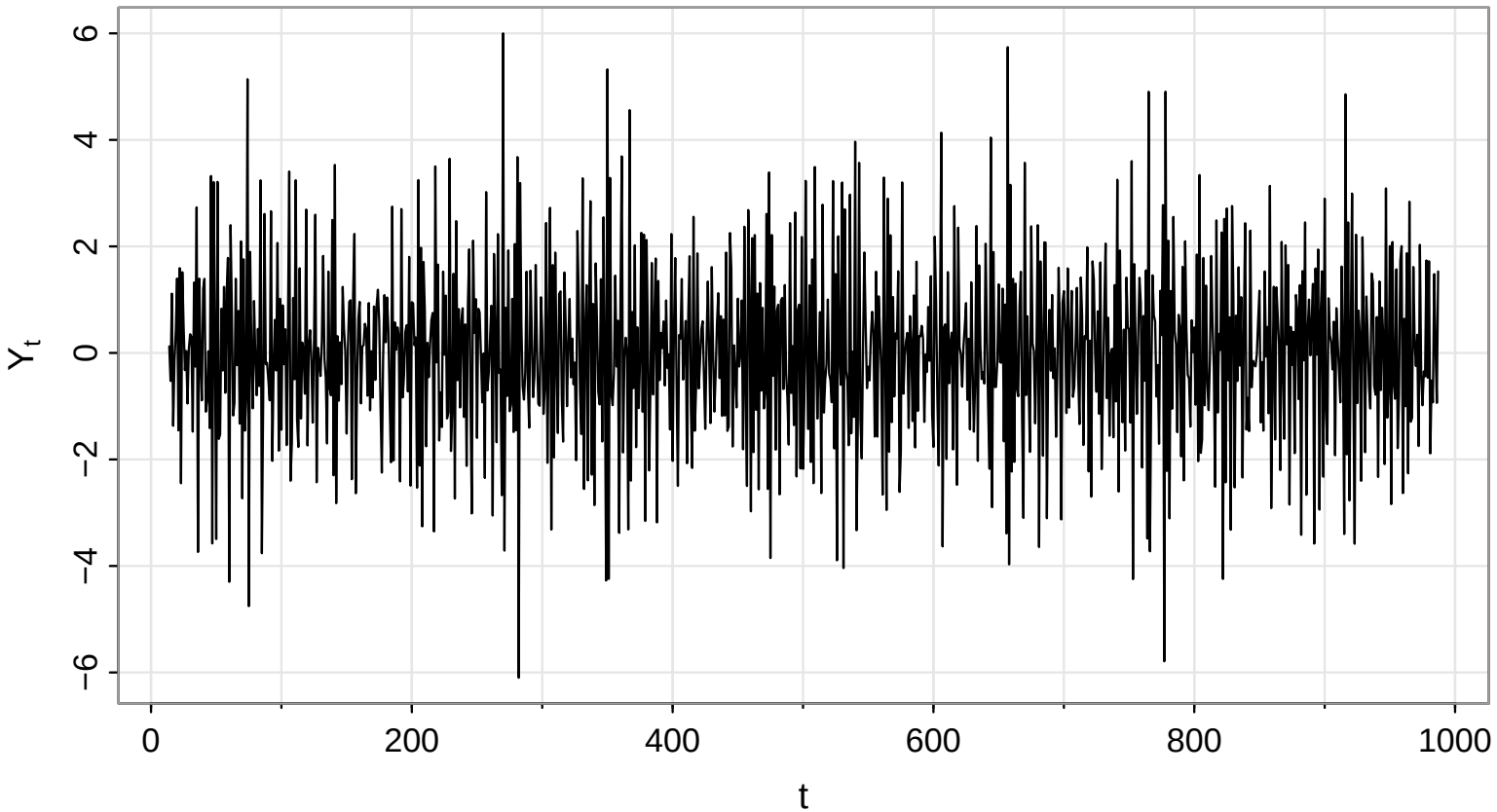
$$= \quad (1-1.4B+0.45B^2)(1-0.7B^{12})\epsilon_t$$

Note that Z_t and X_t satisfy the same model. This model is under canonical form: $1-1.4B+0.45B^2 = (1-0.9B)(1-0.5B)$

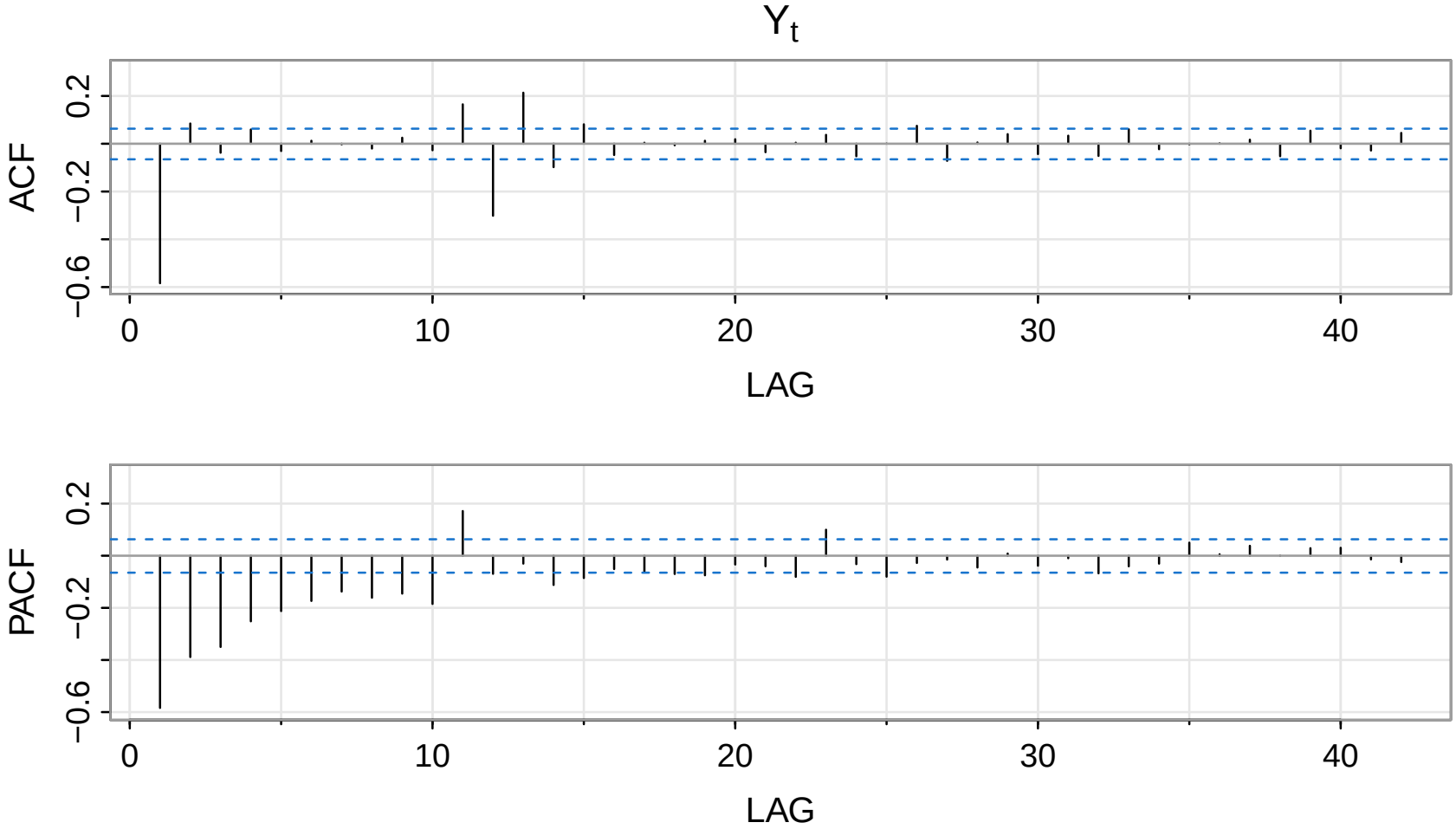
The simulation shows a trend and a seasonality of period 12



The transformed series $Y_t = (1-B)(1-B^{12})X_t$ is stationary



ACF and PACF of Y_t: polynomials in B and B^{12} are necessary.



Estimation of a SARIMA$(1,1,1)(1,1,1)_{12}$ for the series X_t

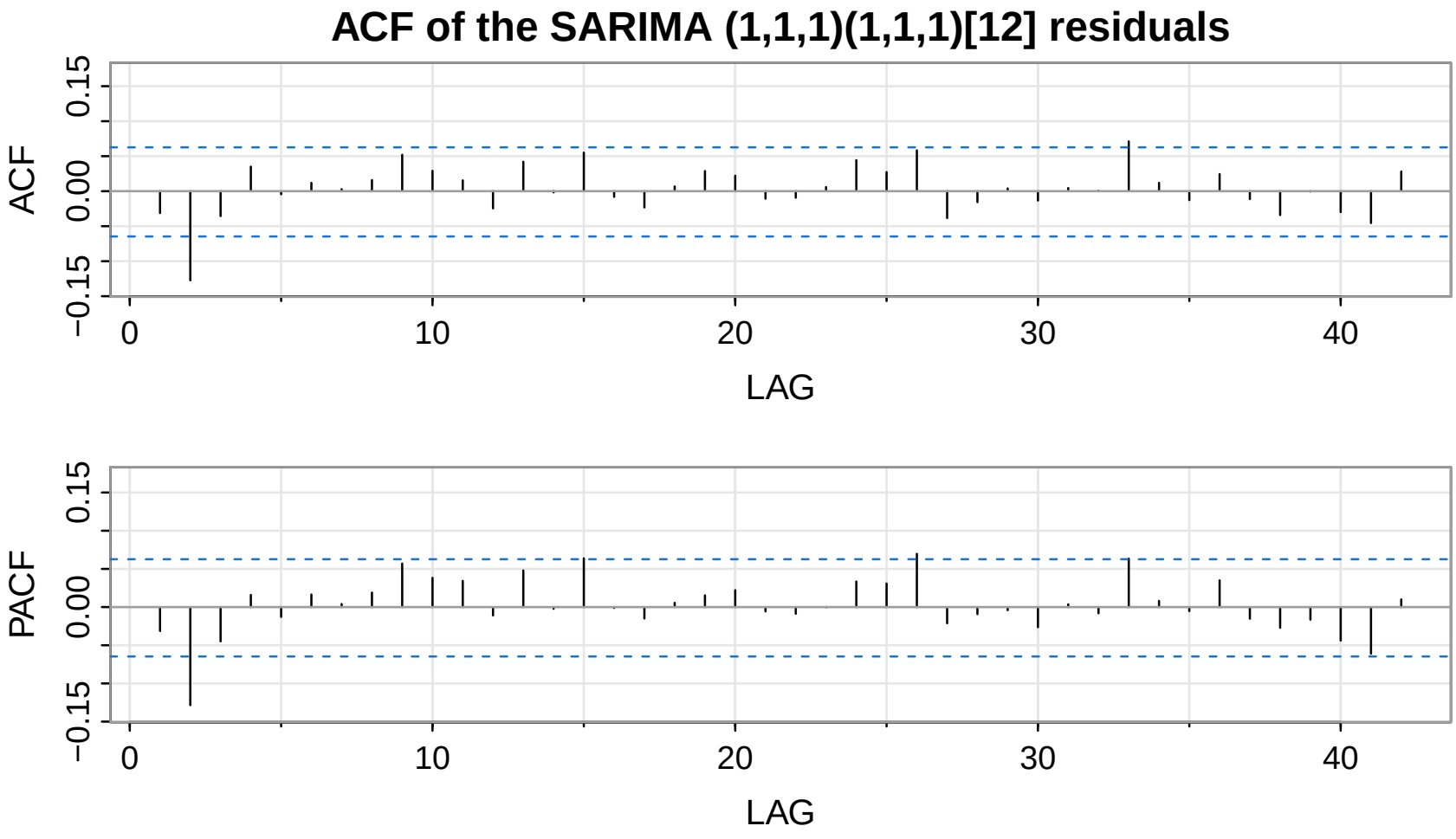
Estimated model:

$$(1+0.25B)(1-0.42B^{12})(1-B^{12})(1-B)X_t = (1-0.93B)(1-0.75B^{12})\epsilon_t$$

```
> res<-sarima(X,1,1,1,1,1,1,12)
> res$ttable
      Estimate   SE   t.value p.value
ar1   -0.2483 0.0318  -7.8059       0
ma1   -0.9347 0.0102 -92.0507       0
sar1   0.4205 0.0645   6.5156       0
sma1  -0.7492 0.0480 -15.6017       0
> LjungBox(res$fit)
lags statistic df      p-value
 5  20.05887  1 7.509460e-06
10  23.53685  6 6.352403e-04
15  29.26971 11 2.061116e-03
20  30.97957 16 1.353731e-02
25  33.96430 21 3.655941e-02
30  39.24337 26 4.620540e-02
```

The residuals don't pass the portmanteau tests

A significant autocorrelation at order 2 remains



Estimation of a SARIMA$(2,1,2)(1,1,1)_{12}$

Estimated model:

$$\begin{aligned} & (1-0.38B-0.035B^2)(1-0.38B^{12})(1-B^{12})(1-B)X_t \\ = \quad & (1-1.6B+0.64B^2)(1-0.73B^{12})\epsilon_t \end{aligned}$$

The AR coefficient has little significance:

```
> res<-sarima(X,2,1,2,1,1,1,12,details=FALSE)
> res$ttable
Estimate     SE  t.value p.value
ar1    0.3800 0.1255   3.0284  0.0025
ar2    0.0346 0.0563   0.6146  0.5390
ma1   -1.6071 0.1202 -13.3699  0.0000
ma2    0.6462 0.1116   5.7925  0.0000
sar1   0.3805 0.0670   5.6762  0.0000
sma1  -0.7306 0.0498 -14.6736  0.0000
> LjungBox(res$fit)
lags statistic df   p-value
5  1.816171  0        NA
10  2.666649  4 0.6150631
15  5.848449  9 0.7549851
20  7.675727 14 0.9055267
25 10.406809 19 0.9421868
30 15.267584 24 0.9126842
```

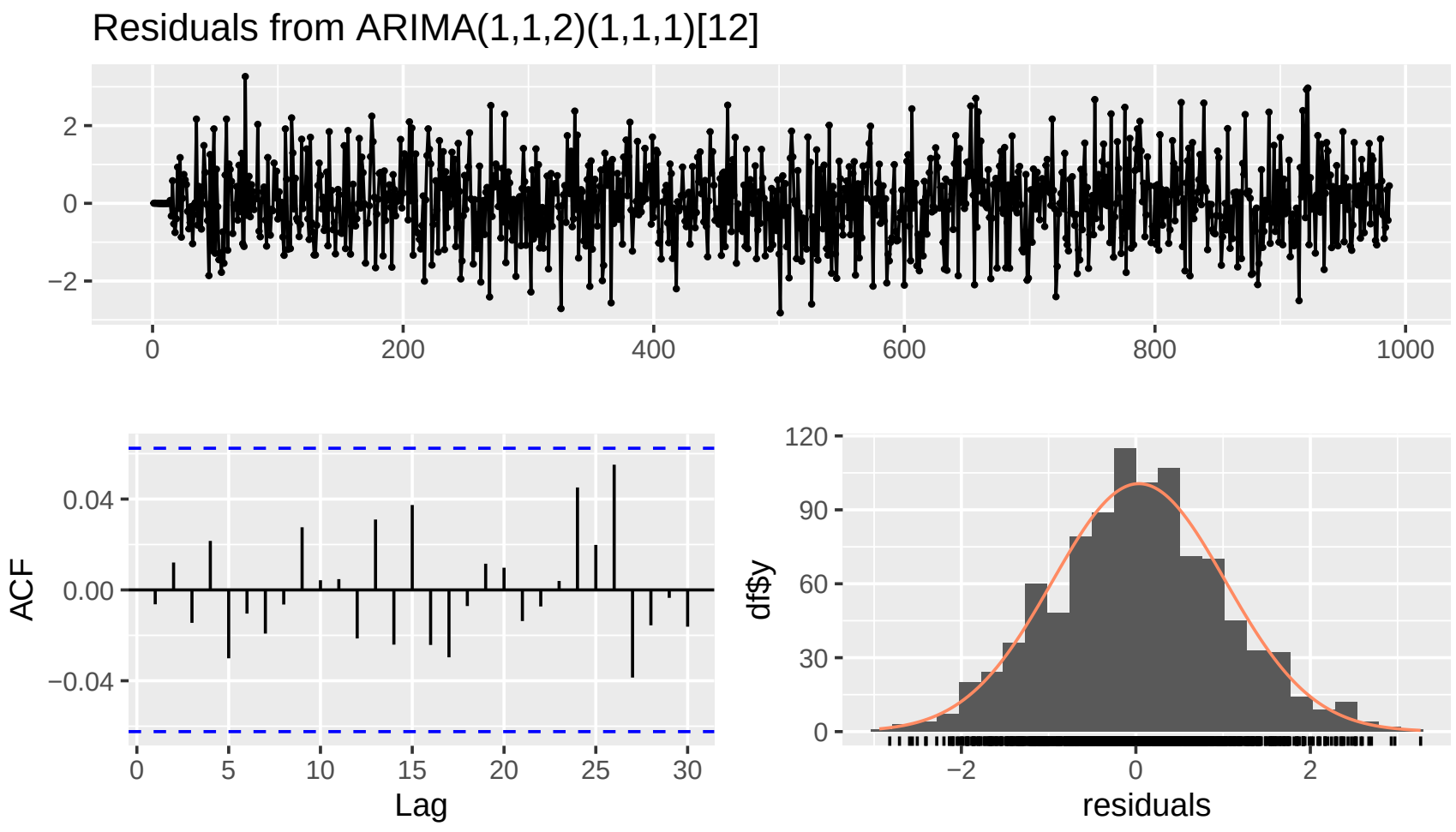
Estimation of a SARIMA$(1,1,2)(1,1,1)_{12}$

All the estimated coefficients are significant (note a bug in `LjungBox`)

```
> res<-sarima(X,1,1,2,1,1,1,12,details=FALSE)
> res$ttable
Estimate     SE  t.value p.value
ar1    0.3191 0.0956   3.3387   9e-04
ma1   -1.5429 0.0818 -18.8577   0e+00
ma2    0.5872 0.0779   7.5328   0e+00
sar1   0.3825 0.0672   5.6938   0e+00
sma1  -0.7295 0.0501 -14.5485   0e+00
> LjungBox(res$fit)
lags statistic df   p-value
 5  2.058168  0 0.0000000
10  3.052468  5 0.6918969
15  6.674195 10 0.7558031
20  8.259399 15 0.9129623
25 10.953618 20 0.9474165
30 15.993518 25 0.9150217
> checkresiduals(res$fit)
Ljung-Box test
data:  Residuals from ARIMA(1,1,2)(1,1,1)[12]
Q* = 3.0525, df = 5, p-value = 0.6919
Model df: 5.   Total lags used: 10
```

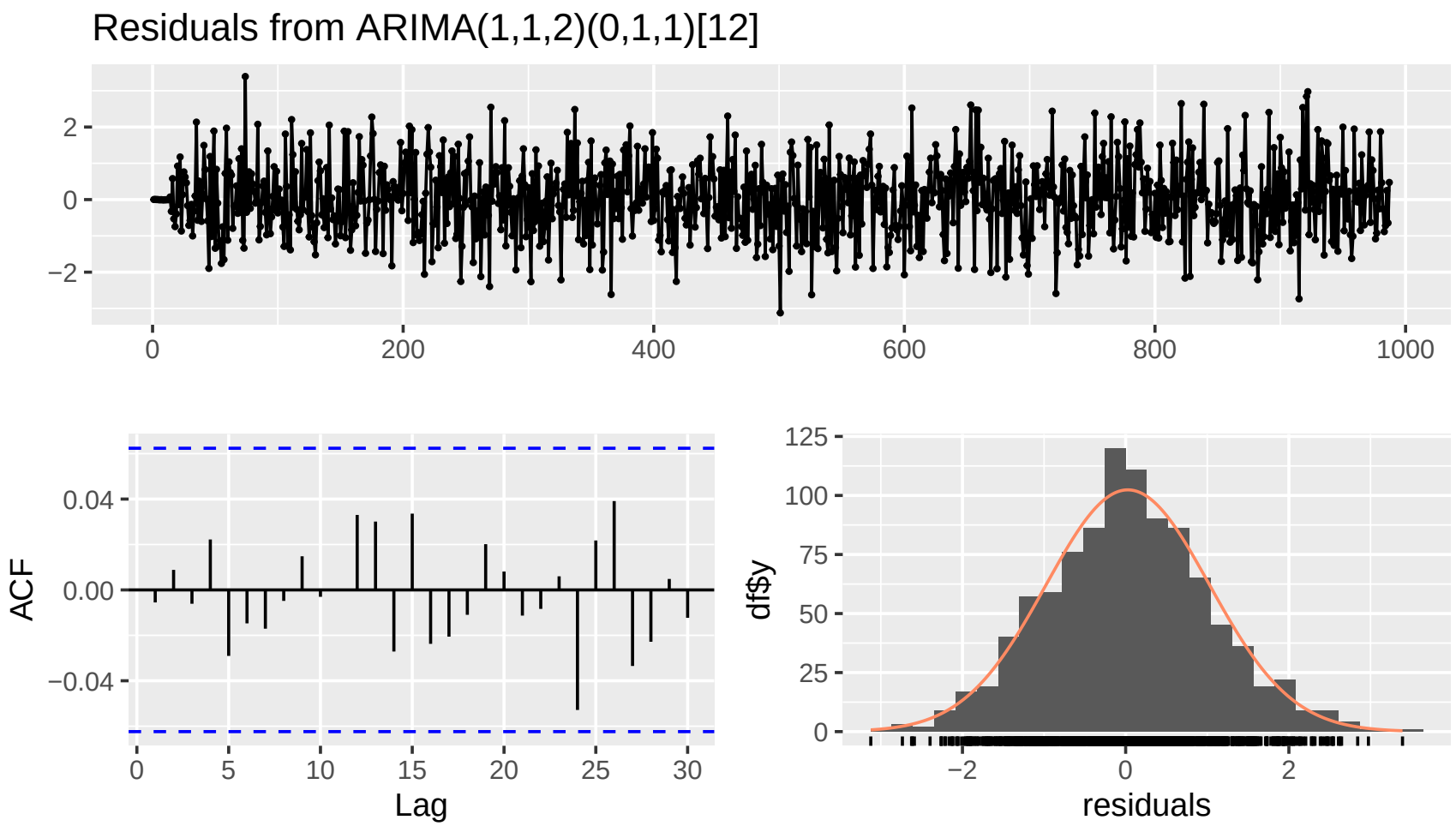
Estimation of a SARIMA$(1,1,2)(1,1,1)_{12}$

The diagnostic tests are good



Let us try a SARIMA$(1,1,2)(0,1,1)_{12}$

The diagnostic tests do not reject the model



Model comparison with information criteria

AIC and BIC select the SARIMA$(1,1,2)(1,1,1)_{12}$

```
>
> AIC(res111111$fit, res212111$fit, res112111$fit, res112011$fit)
          df      AIC
res111111$fit  5 2841.684
res212111$fit  7 2819.248
res112111$fit  6 2817.623
res112011$fit  5 2833.846
> BIC(res111111$fit, res212111$fit, res112111$fit, res112011$fit)
          df      BIC
res111111$fit  5 2866.158
res212111$fit  7 2853.511
res112111$fit  6 2846.991
res112011$fit  5 2858.319
>
```

Estimated and simulated models

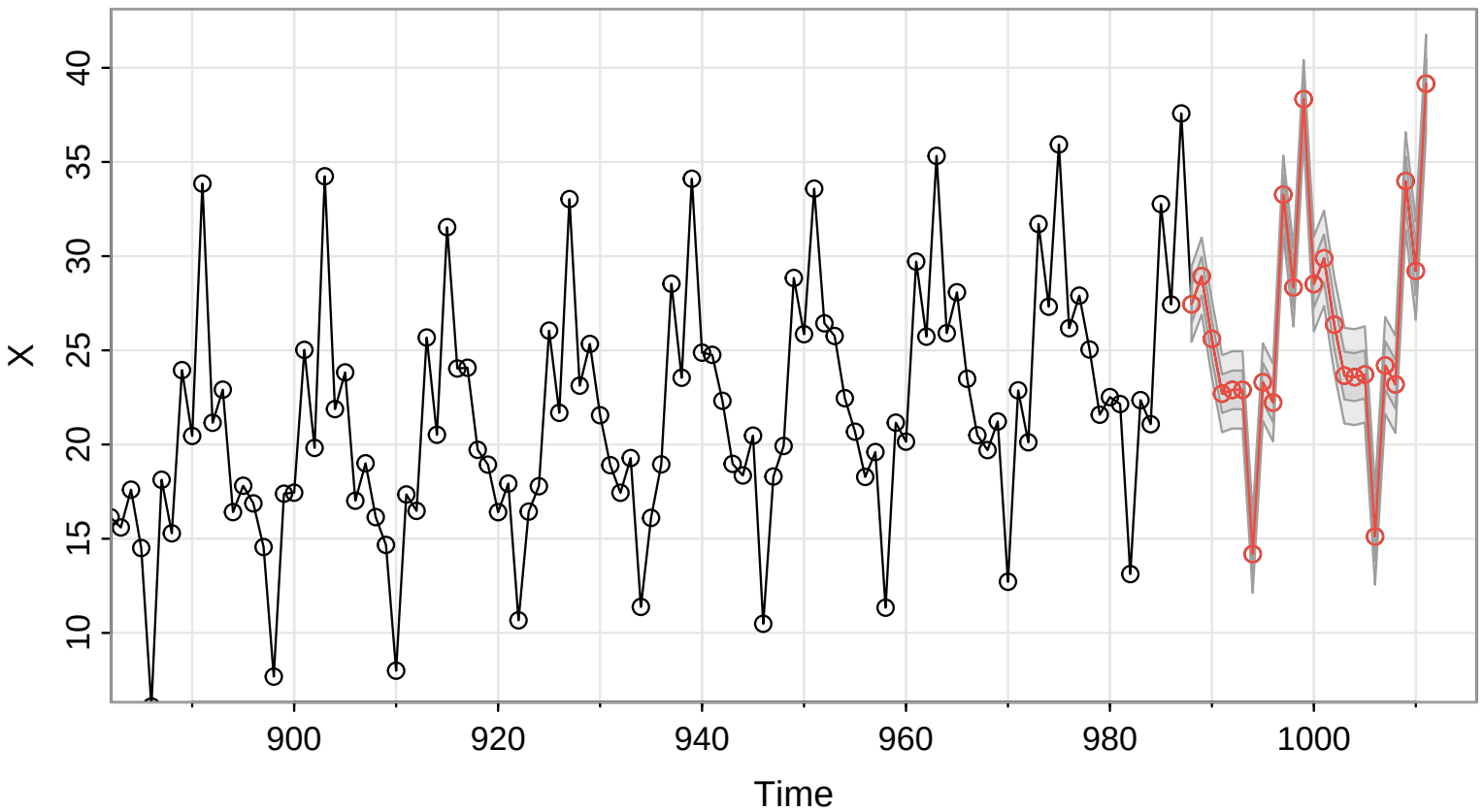
True model:

$$\begin{aligned} & (1-0.2B)(1-B)(1-0.4B^{12})(1-B^{12})X_t \\ = \quad & (1-1.4B+0.45B^2)(1-0.7B^{12})\epsilon_t \end{aligned}$$

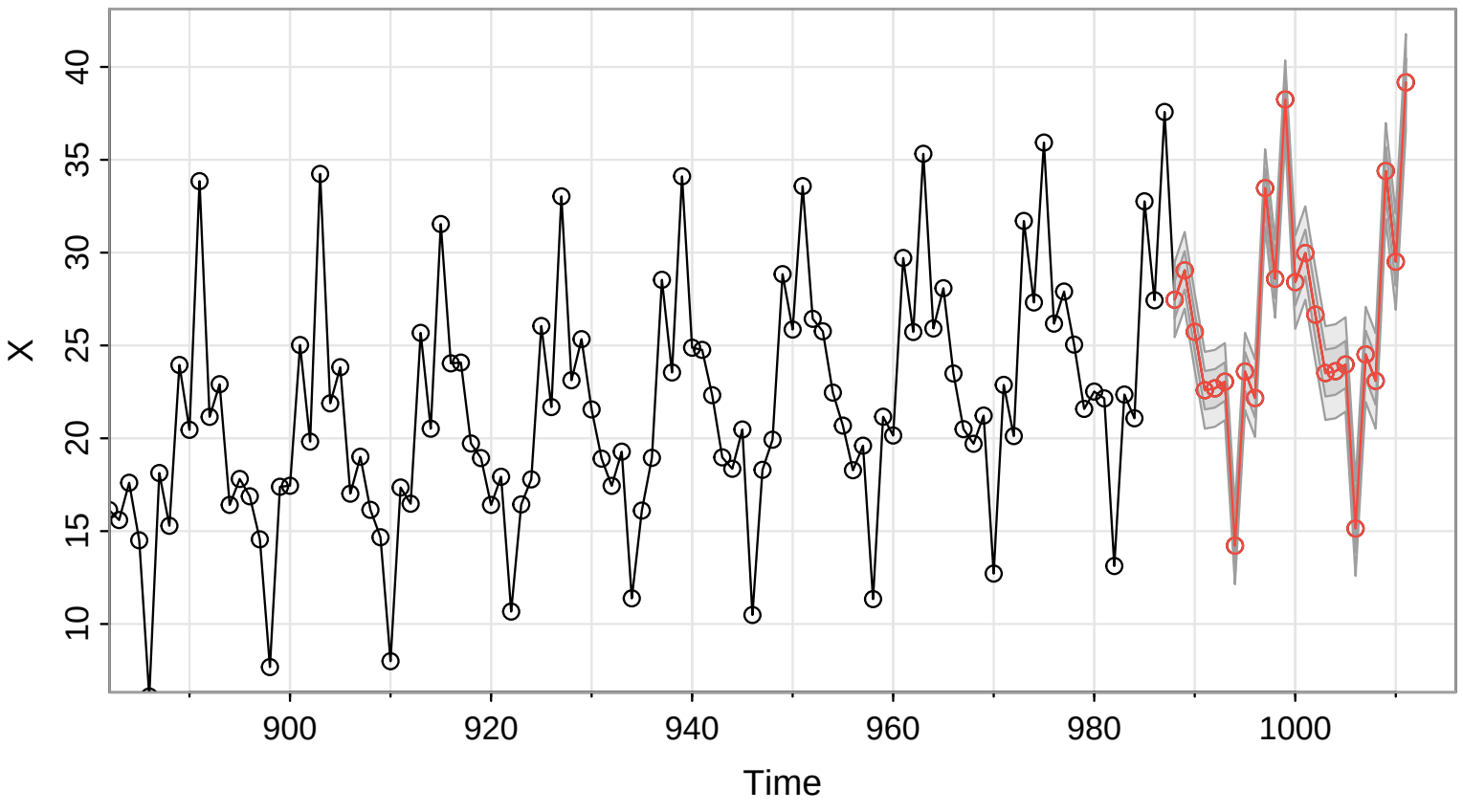
Estimated model:

$$\begin{aligned} & (1-\underset{(0.0956)}{0.3191}B)(1-B)(1-\underset{(0.0672)}{0.3825}B^{12})(1-B^{12})X_t \\ = \quad & (1-\underset{(0.0818)}{1.5429}B+\underset{(0.0779)}{0.5872}B^2)(1-\underset{(0.0501)}{0.7295}B^{12})\epsilon_t \end{aligned}$$

Forecasts of 24 future values using the estimated SARIMA$(1,1,2)(1,1,1)_{12}$ model

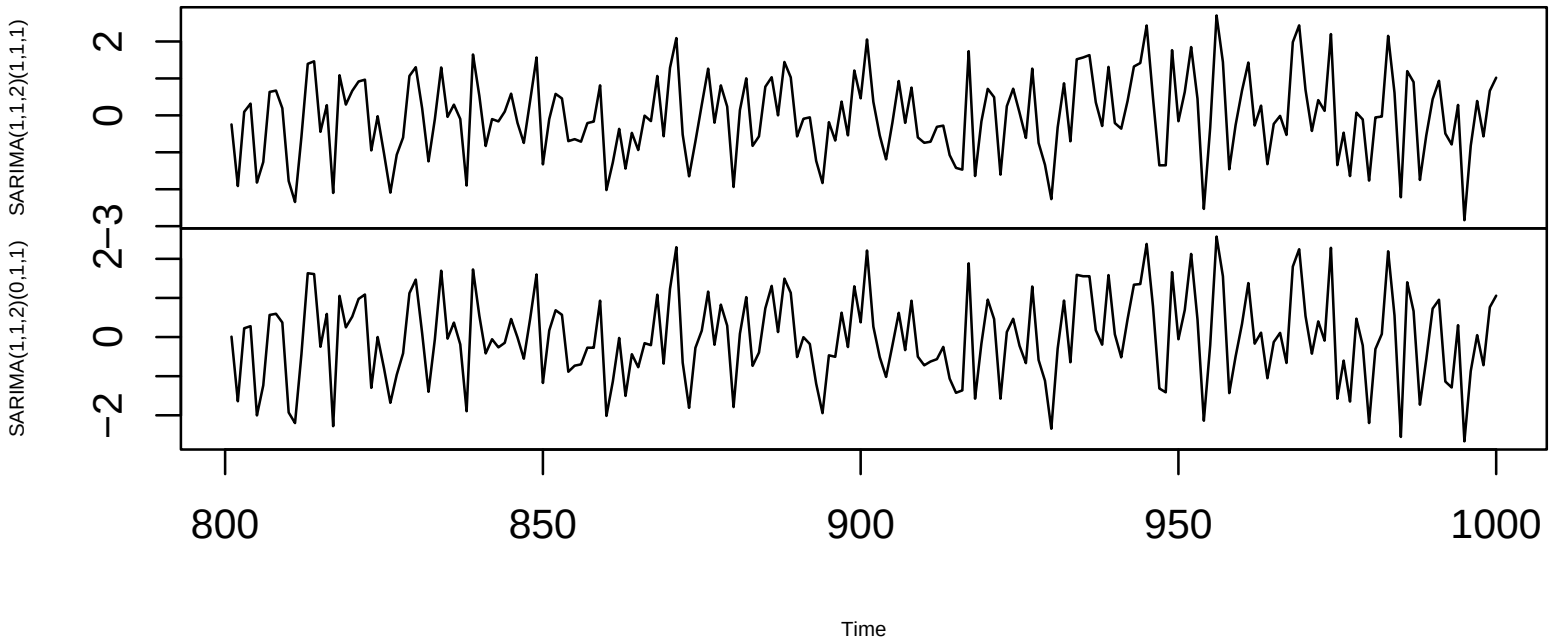


Almost the same results with the SARIMA$(1,1,2)(0,1,1)_{12}$ model



One-step ahead forecast errors of the 2 SARIMA

Models estimated on 800 observations and backtested on 200 observations

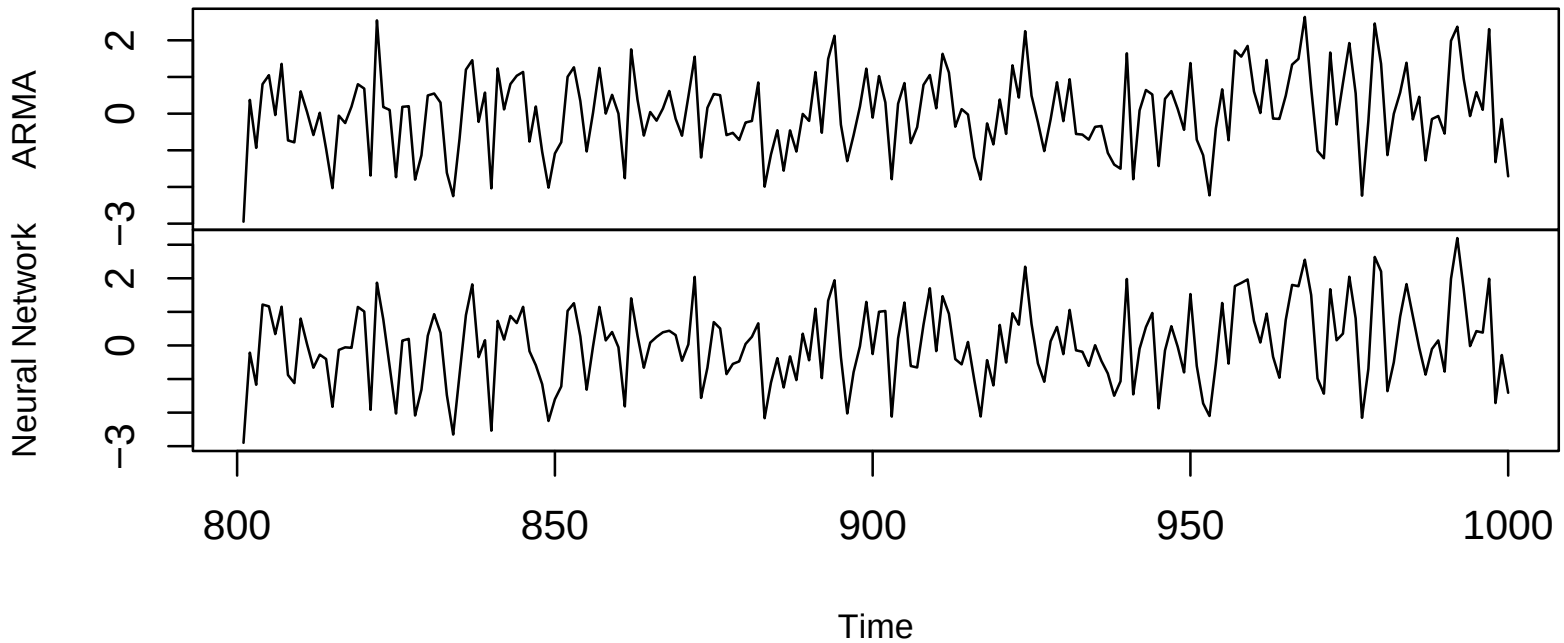


Prediction RMSE of the two models: 1.098011 and 1.11246

Diebold-Mariano test (`dm.test()` in the `R` package `forecast`):
H_0 : same forecast accuracy; H_1 : different forecast accuracy;
p-value = 0.2671

Comparing an ARMA(1,1) and a Neural Network

Function `nnetar()` of `R` package `forecast`. The DGP is ARMA(1,3)



Prediction RMSE of the two methods: 1.082311 and 1.175724

Diebold-Mariano test: p-value = 6.926e-05

End of Chapter 3 ☺ !