Times series forecasting

Varicella data

Julien JACQUES

We start by loading necessary package

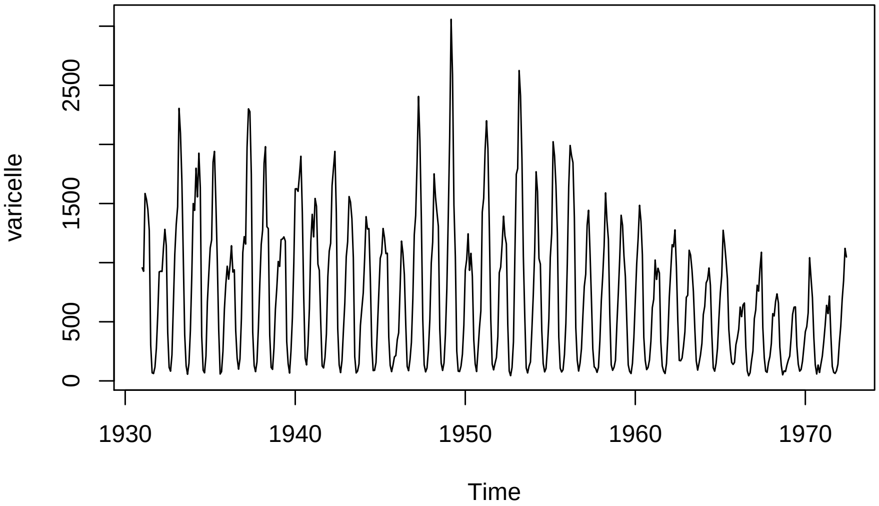
```
library(forecast)
```

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

```
library(ggplot2)
```

We load the data and plot them

```
data=read.csv(file="http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/varicelle.csv")
varicelle<-ts(data$x,start=c(1931,1),end=c(1972,6),freq=12)
plot(varicelle)
```



It seems to be a seasonal pattern. We can check this with the seasonplo

The mean is given by

```
mean(varicelle)
```

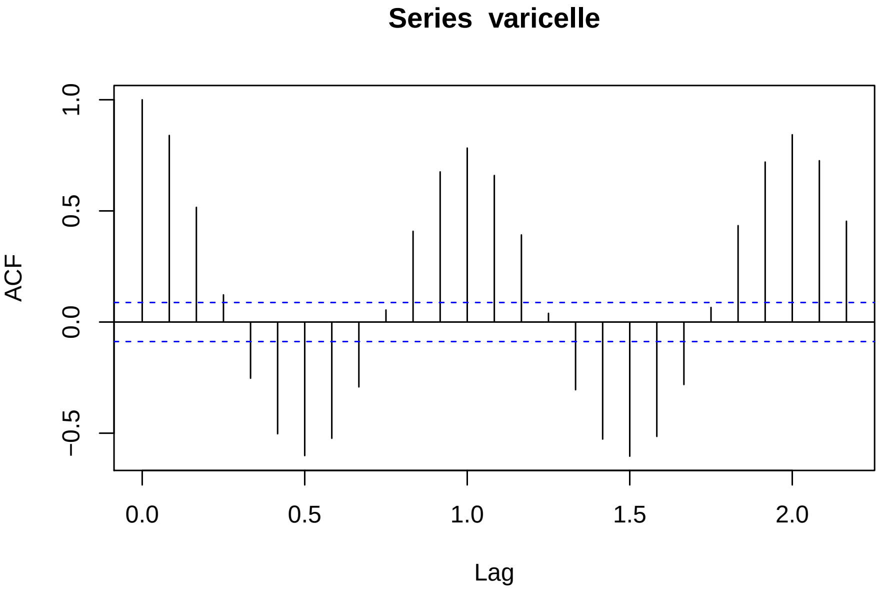
```
## [1] 732.4076
```

And the auto-correlation mean that there is a seasonal pattern in the data

```
tmp=acf(varicelle,type="cor",plot = FALSE)
tmp$acf[1:3,1,1]
```

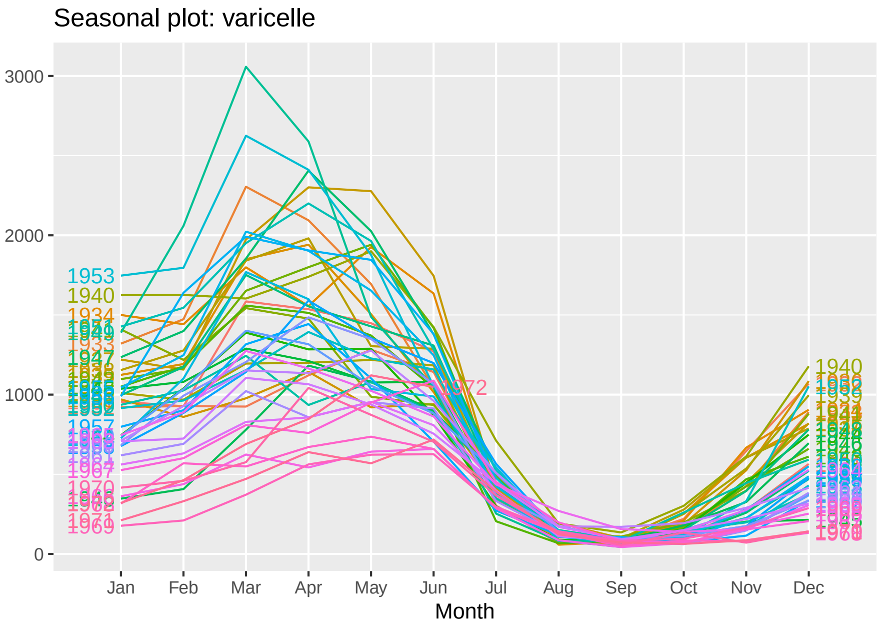
```
## [1] 1.0000000 0.8394105 0.5160841
```

```
plot(tmp)
```



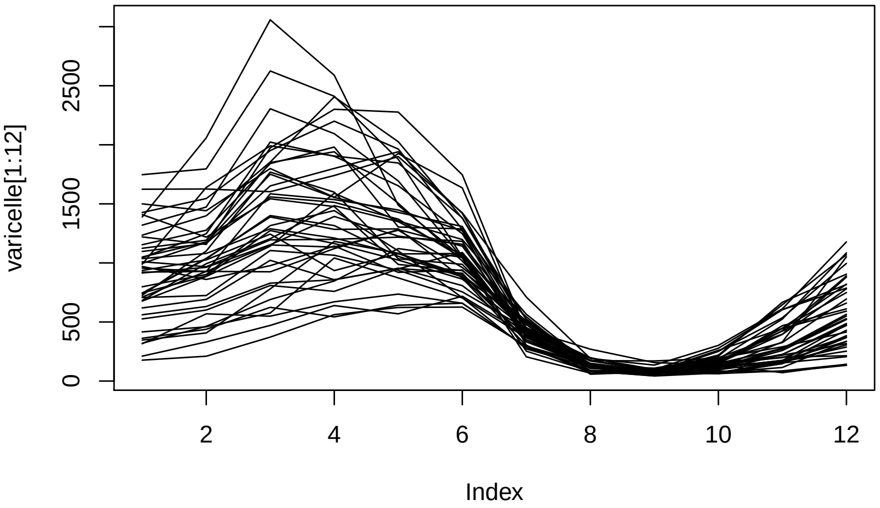
What is confirmed by the seasonal plot

```
ggseasonplot(varicelle,year.labels= TRUE,year.labels.left=TRUE)
```



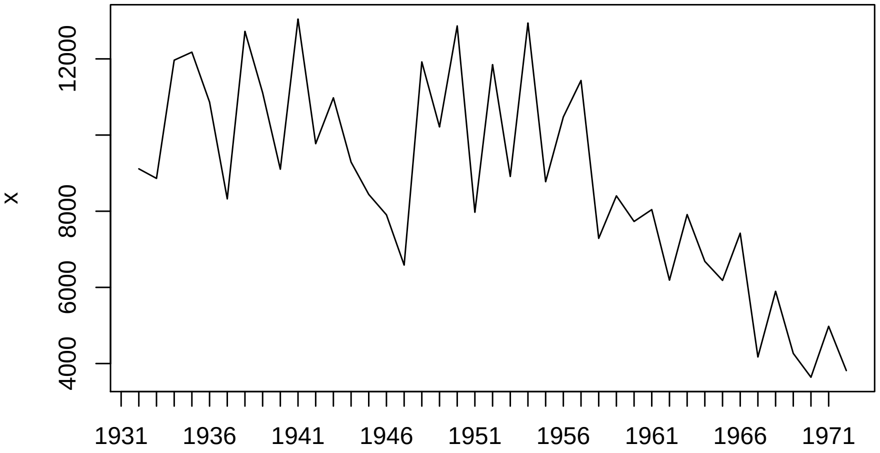
We can also plot manually the seasonal plot

```
plot(varicelle[1:12],type="l",ylim=c(min(varicelle),max(varicelle)))
for (i in 1:41) lines(varicelle[(1+12*i):(12*(i+1))])
```



We now compute and plot the annual evolution

```
x=rep(0,41)
for (i in 0:40) x[i+1]<-sum(varicelle[(1+12*i):(12*(i+1))])
plot(x,type='l',xaxt='n',xlab='')
axis(1,at = 0:40,labels = 1931:1971)
```



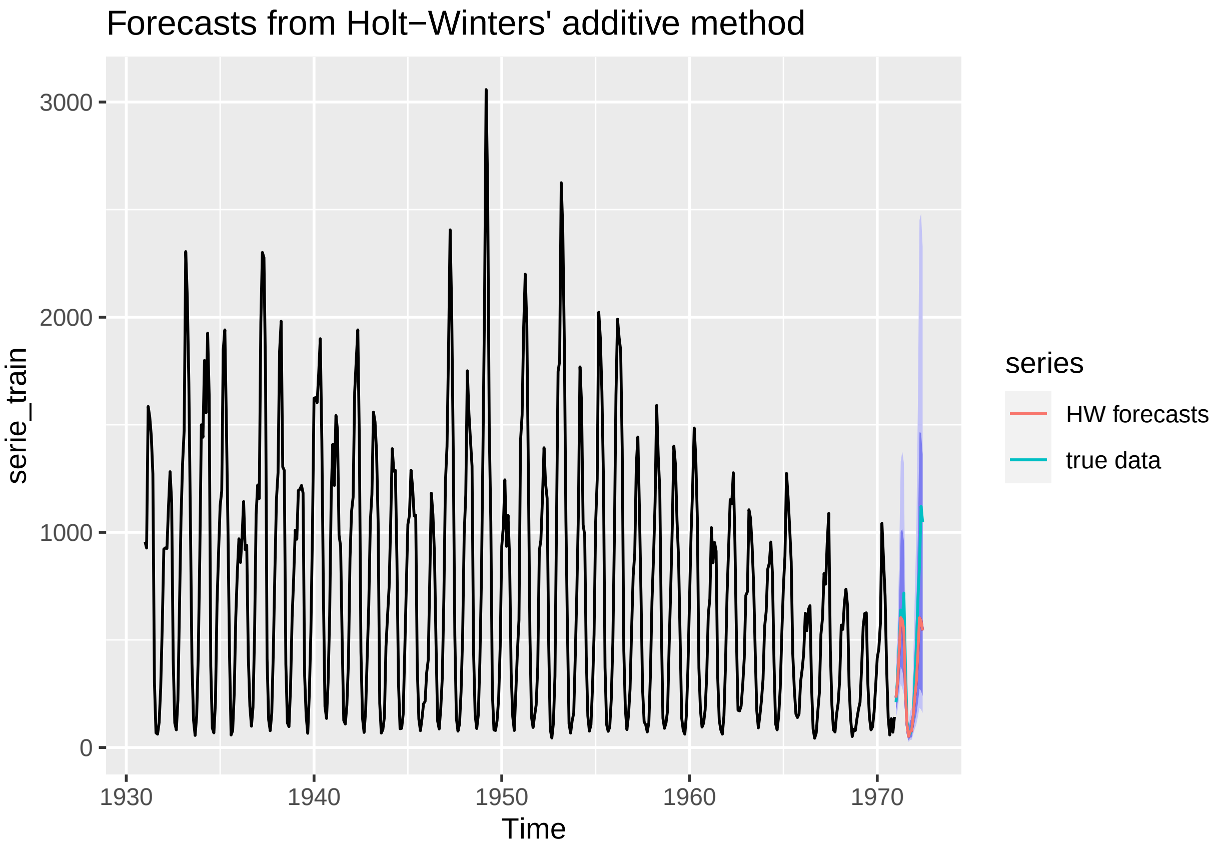
Forecasting

We split the serie into train and test

```
serie_train=window(varicelle,start=c(1931,1),end=c(1970,12))
serie_test=window(varicelle,start=c(1971,1),end=c(1972,6))
```

Forecasting with a **Holt Winters exponential smoothing**

```
fit=hw(serie_train,lambda="auto")
prev=forecast(fit,h=18)
autoplot(prev) + autolayer(serie_test, series="true data")+
  autolayer(prev$mean, series="HW forecasts")
```

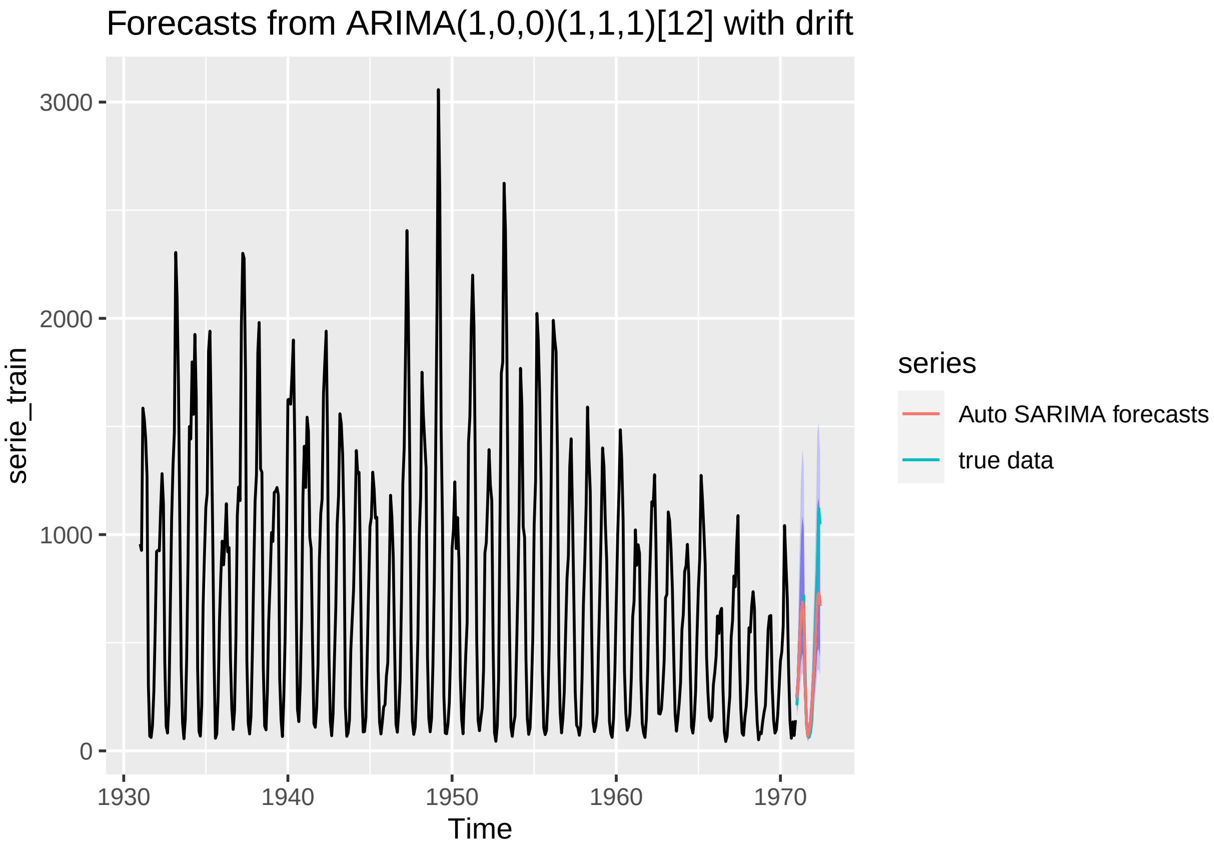


```
cat('RMSE :',sqrt(mean((serie_test-prev$mean)^2)),'\n')
```

```
## RMSE : 203.0989
```

Forecasting with a **SARIMA** model, automatically chosen

```
fit=auto.arima(serie_train,lambda="auto")
prev=forecast(fit,h=18)
autoplot(prev) + autolayer(serie_test, series="true data")+
  autolayer(prev$mean, series="Auto SARIMA forecasts")
```

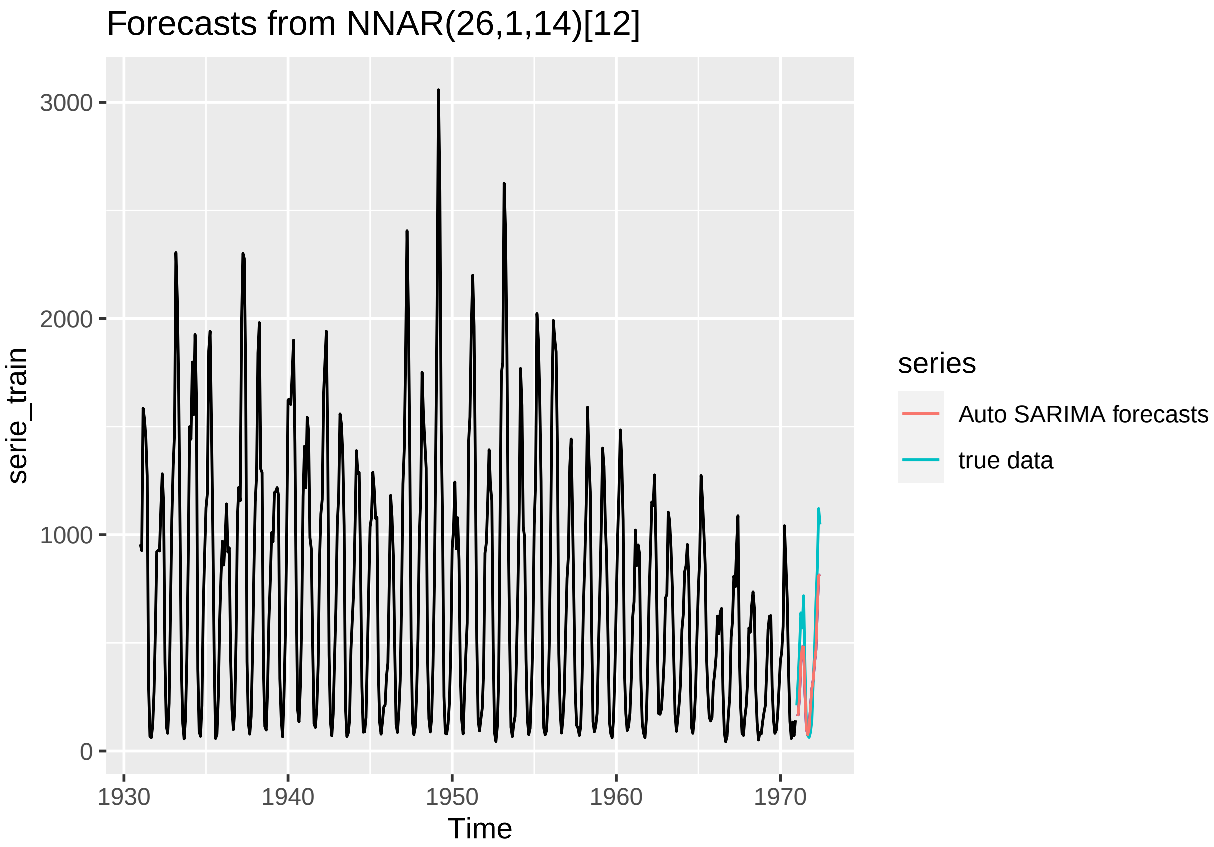


```
cat('RMSE :',sqrt(mean((serie_test-prev$mean)^2)),'\n')
```

```
## RMSE : 142.2667
```

Forecasting with a **auto-regressive neural network**

```
 fit=nnetar(serie_train,lambda = "auto")
prev=forecast(fit,h=18)
autoplot(prev) + autolayer(serie_test, series="true data")+
  autolayer(prev$mean, series="Auto SARIMA forecasts")
```

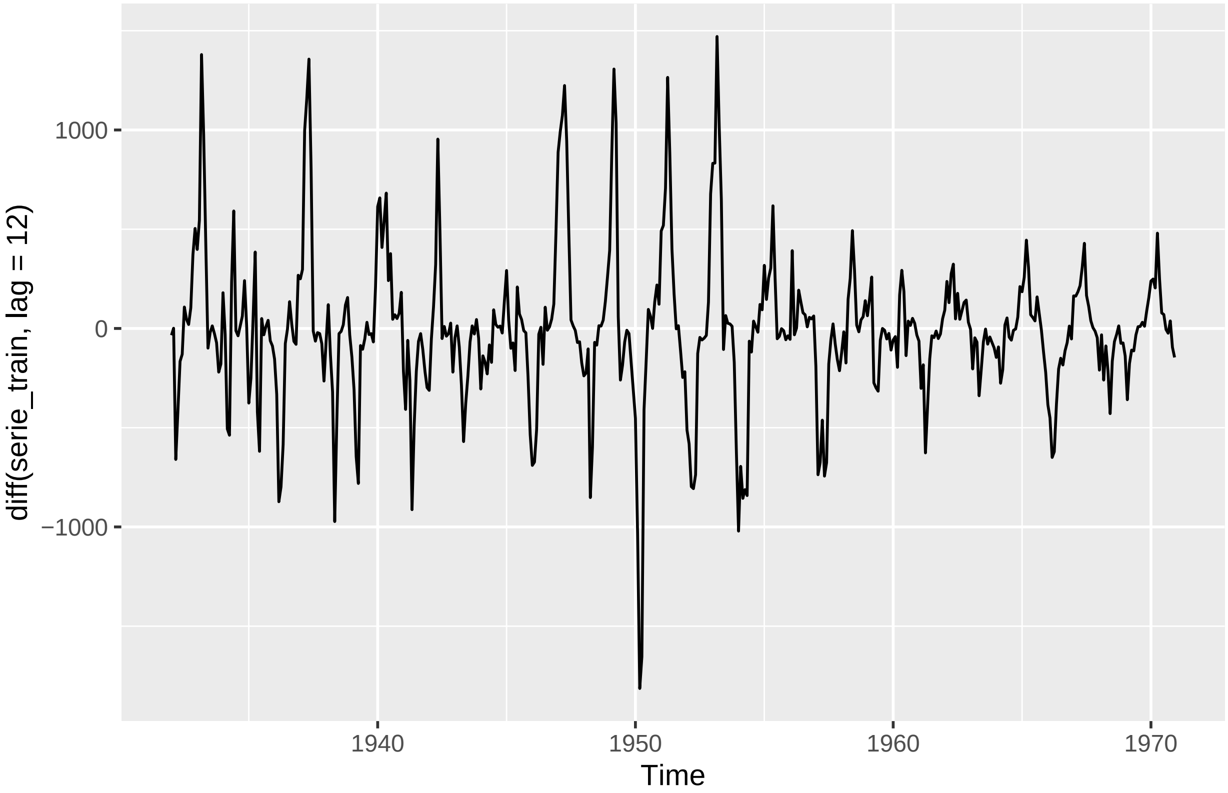


```
cat('RMSE :',sqrt(mean((serie_test-prev$mean)^2)),'\n')
```

```
## RMSE : 163.9794
```

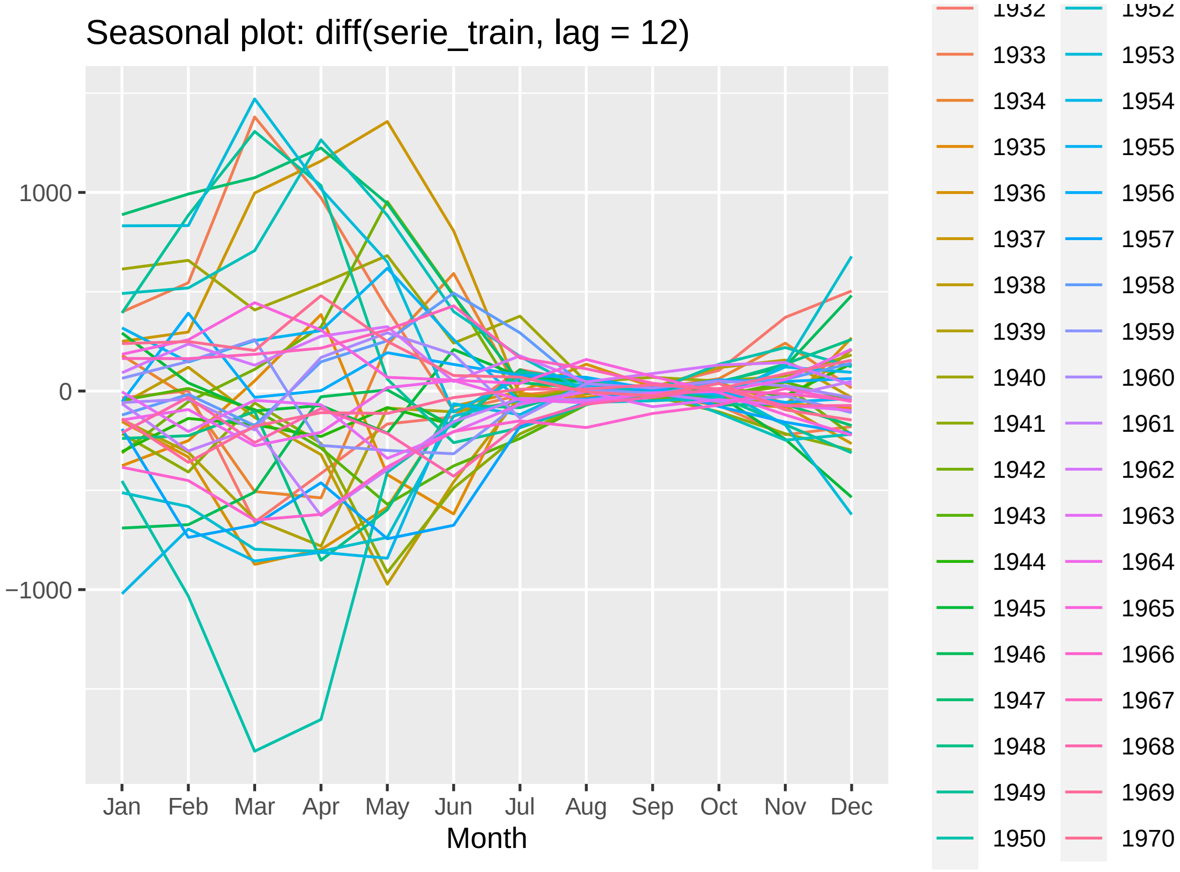
The best forecast is the SARIMA model, we can try to improve it. Let's remove the trend

```
autoplot(diff(serie_train,lag = 12))
```



It seems that there is something still periodic ?

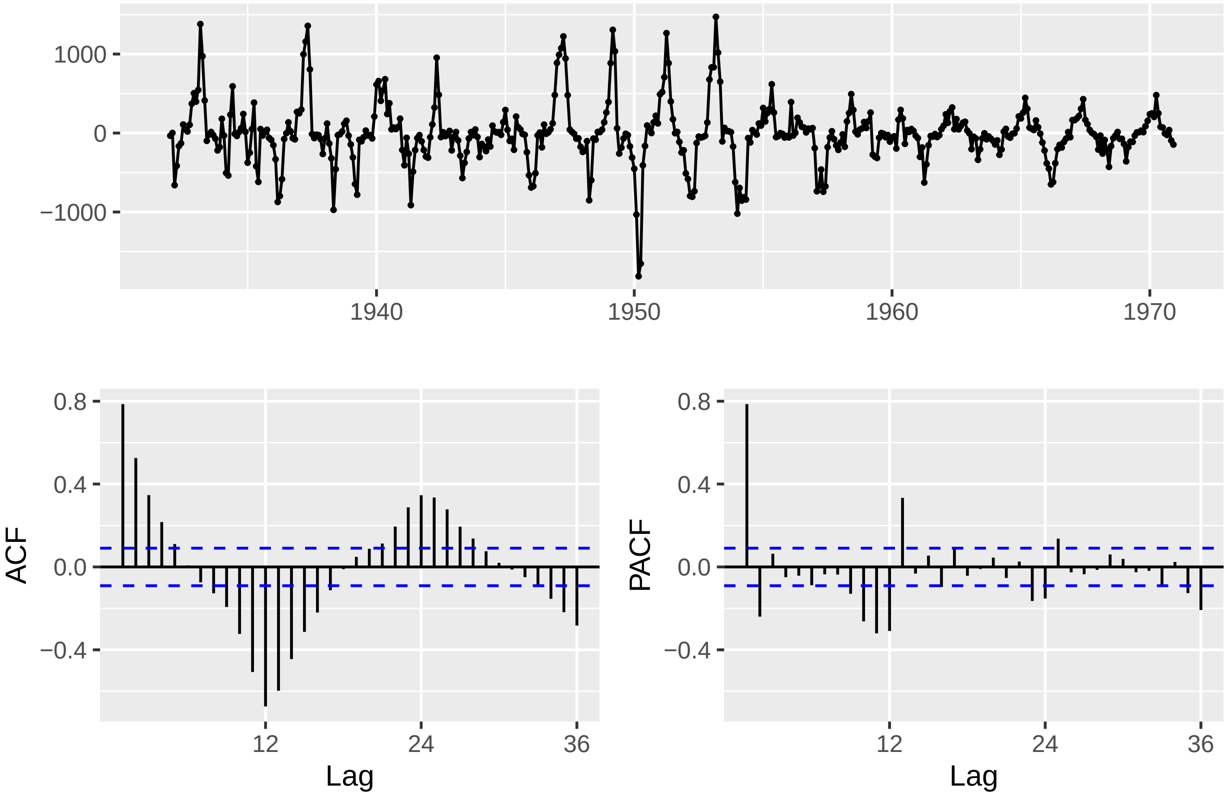
```
ggseasonplot(diff(serie_train,lag=12))
```



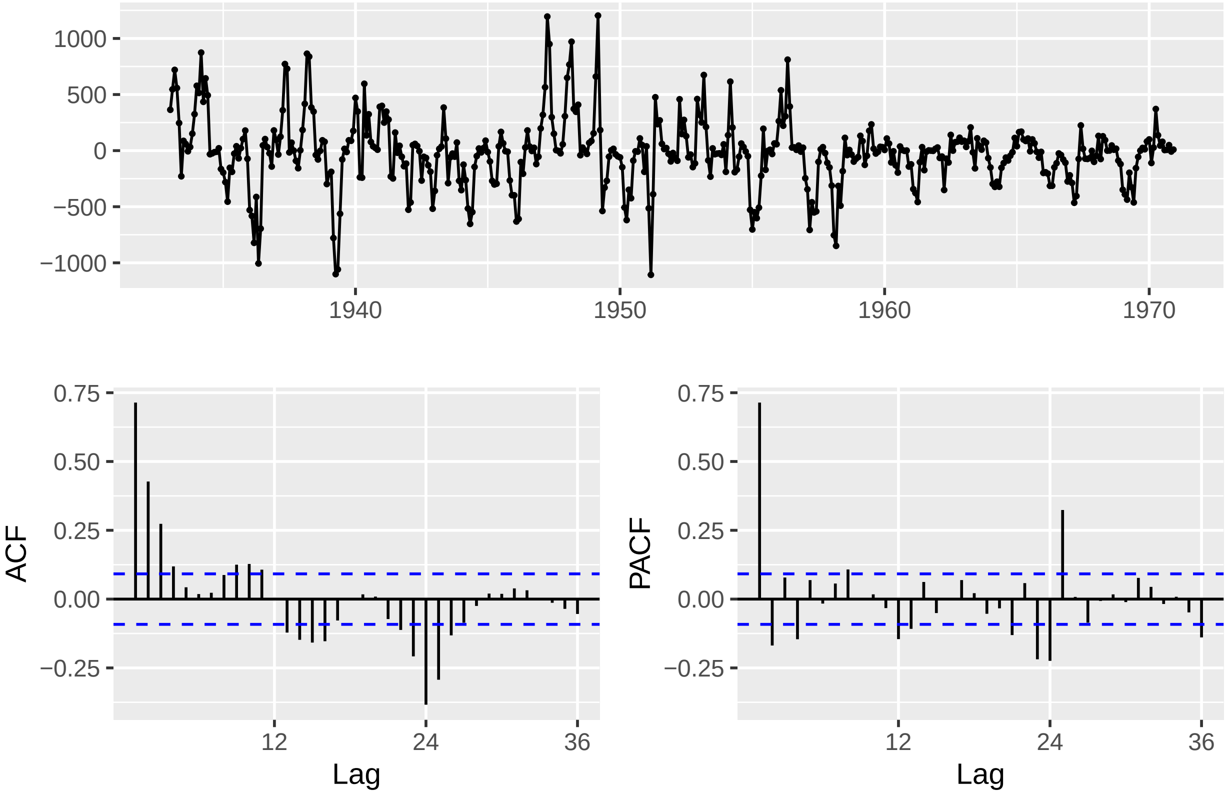
In fact not really: it is just the variance which is higher in spring and lower in fall, but no deterministic seasonal pattern.

Have a look to the ACP / PACF. It seems as there still a seasonal pattern of period 24

```
ggtsdisplay(diff(serie_train,lag = 12))
```

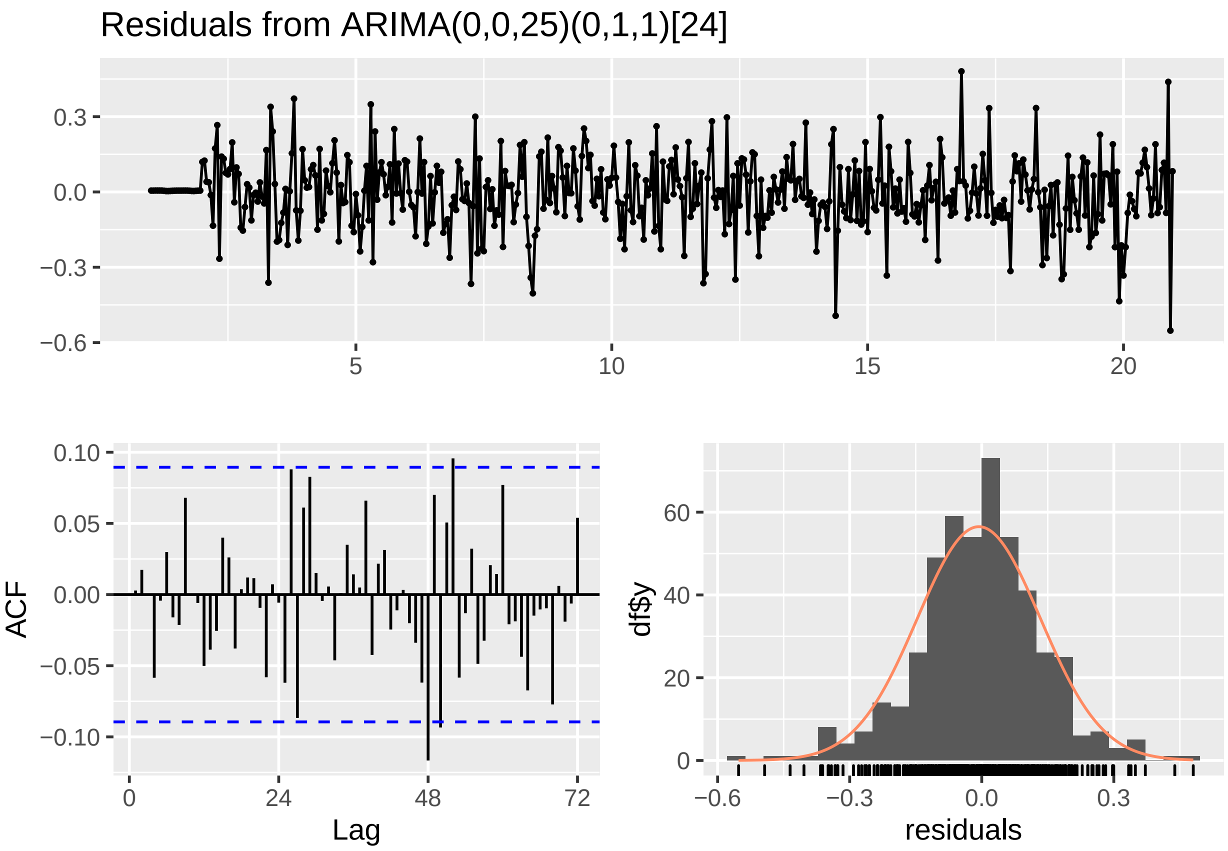


```
ggtsdisplay(diff(serie_train,lag = 24))
```



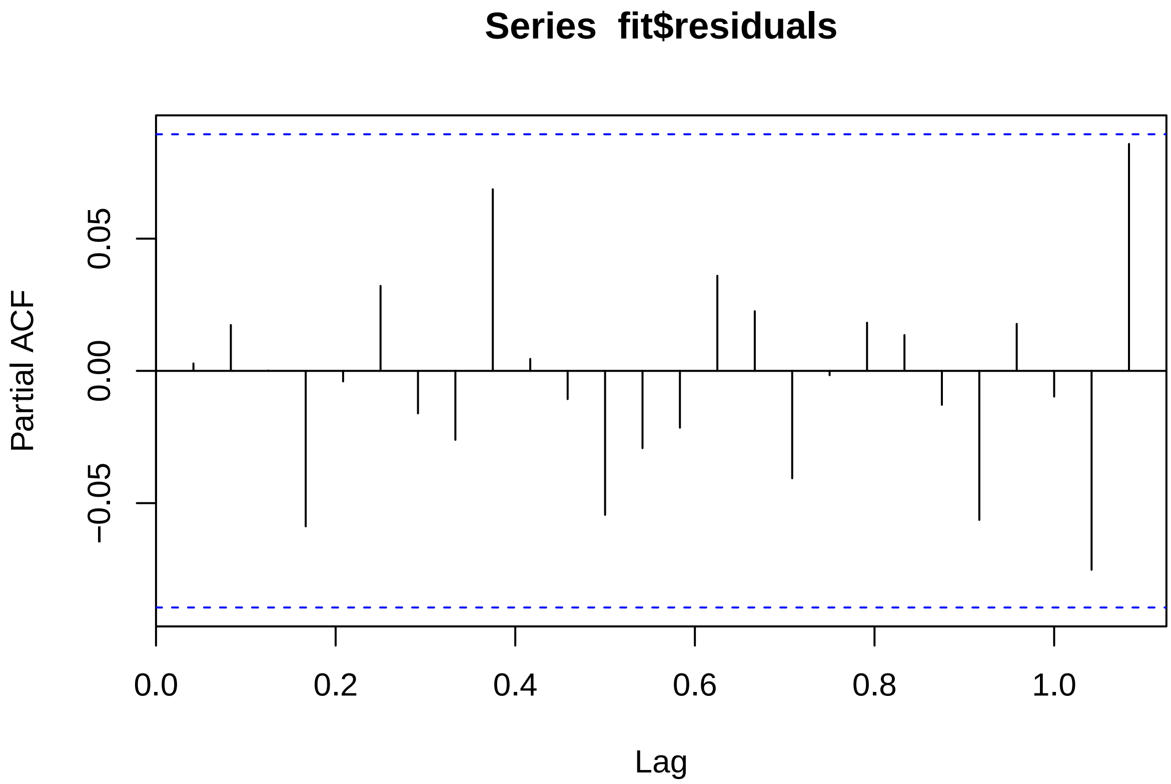
There is a significant ACF at lag 24 and 25... We can test a $SARIMA_{(0,0,25)(0,1,1)24}$

```r
tmp=ts(serie_train,frequency = 24)
tmp_test=ts(serie_test,frequency = 24,start=c(21,1))
fit=Arima(tmp, order=c(0,0,25), seasonal=c(0,1,1),lambda = "auto")
checkresiduals(fit)
```



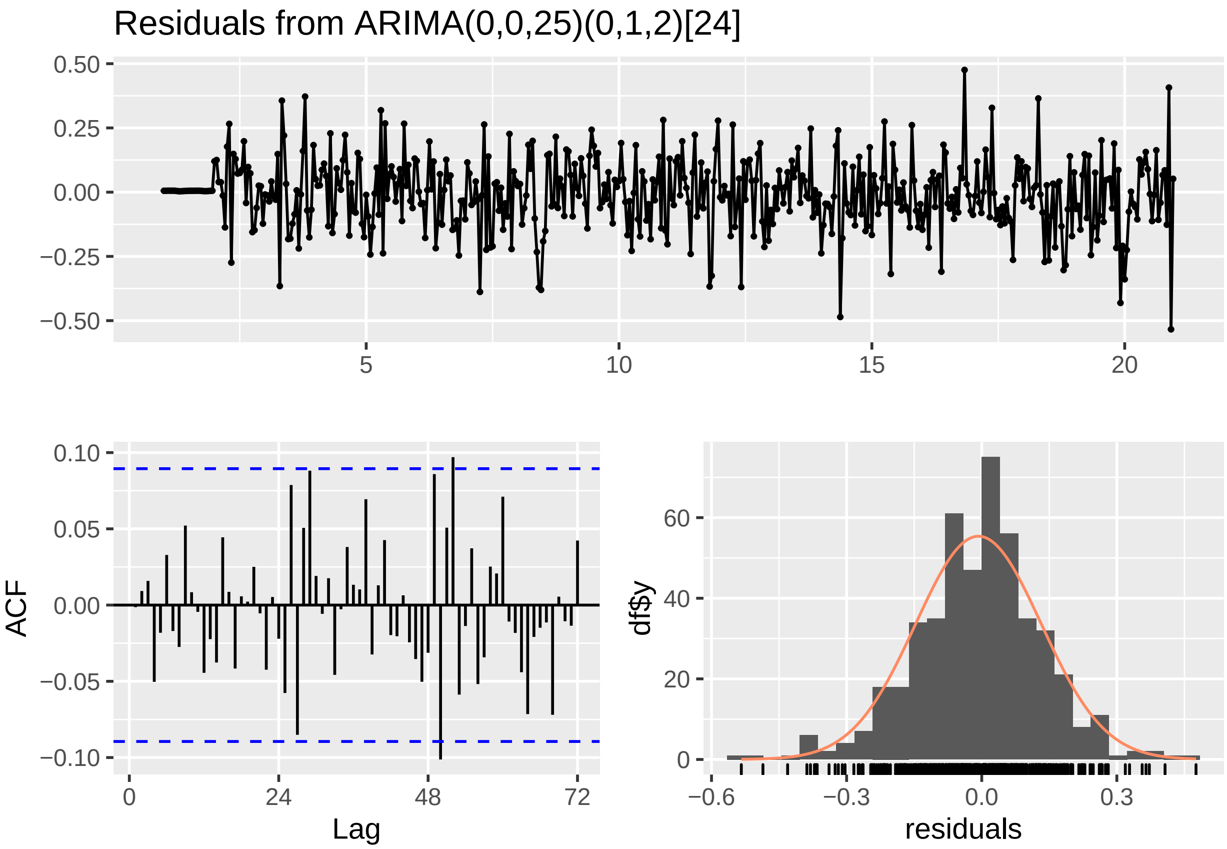
```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,25)(0,1,1)[24]
## Q* = 42.687, df = 22, p-value = 0.005156
##
## Model df: 26.   Total lags used: 48
```

```r
pacf(fit$residuals)
```



There is a significant ACF at lag 48 . . . We can test a $SARIMA_{(0,0,25)(0,1,2)24}$

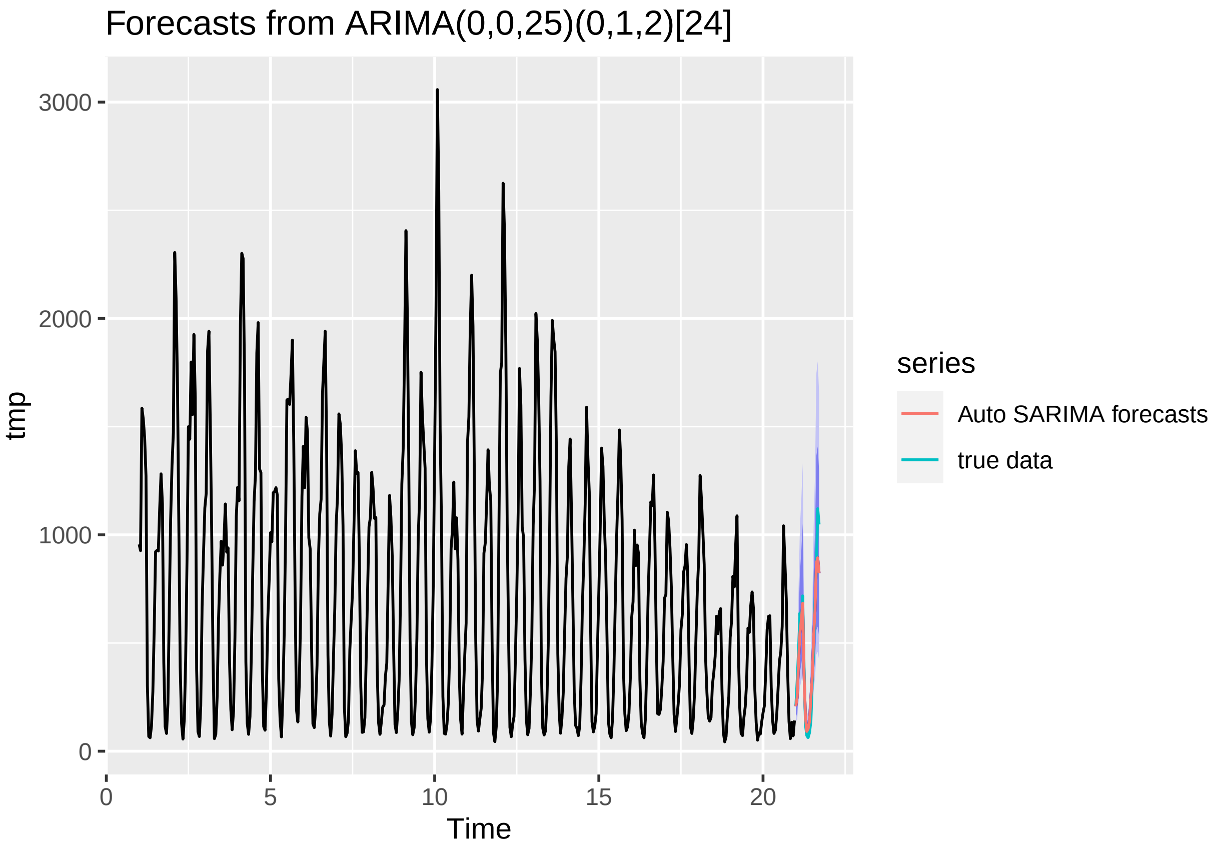
```
tmp=ts(serie_train,frequency = 24)
tmp_test=ts(serie_test,frequency = 24,start=c(21,1))
fit=Arima(tmp, order=c(0,0,25), seasonal=c(0,1,2),lambda = "auto")
checkresiduals(fit)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,25)(0,1,2)[24]
## Q* = 32.932, df = 21, p-value = 0.04698
##
## Model df: 27.   Total lags used: 48
```

We almost capture all correlation. We can use this model for forecasting

```
prev=forecast(fit,h=18)
autoplot(prev) + autolayer(tmp_test, series="true data")+
  autolayer(prev$mean, series="Auto SARIMA forecasts")
```

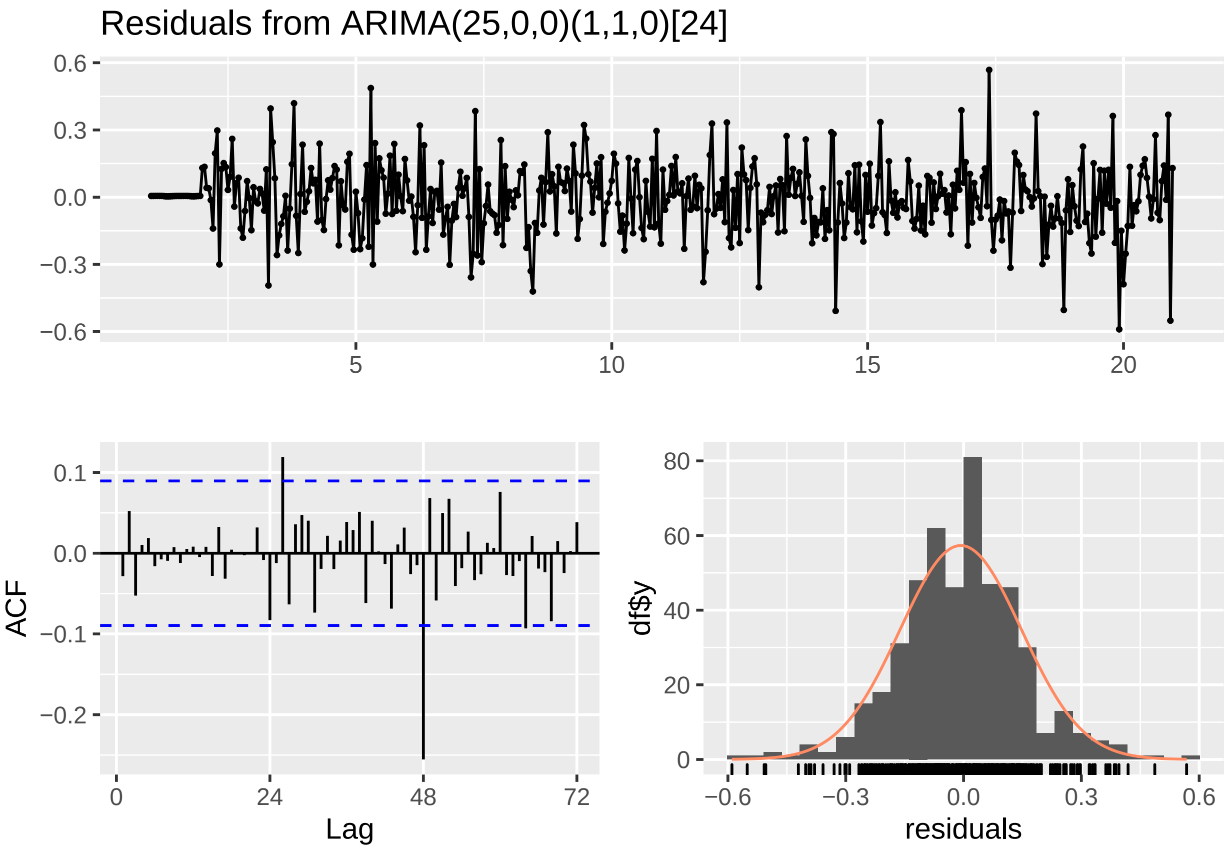


```
cat('RMSE with a SARIMA(25,0,0)(1,1,0)24 :',sqrt(mean((tmp_test-prev$mean)^2)),'\n')
```

```
## RMSE with a SARIMA(25,0,0)(1,1,0)24 : 91.37834
```

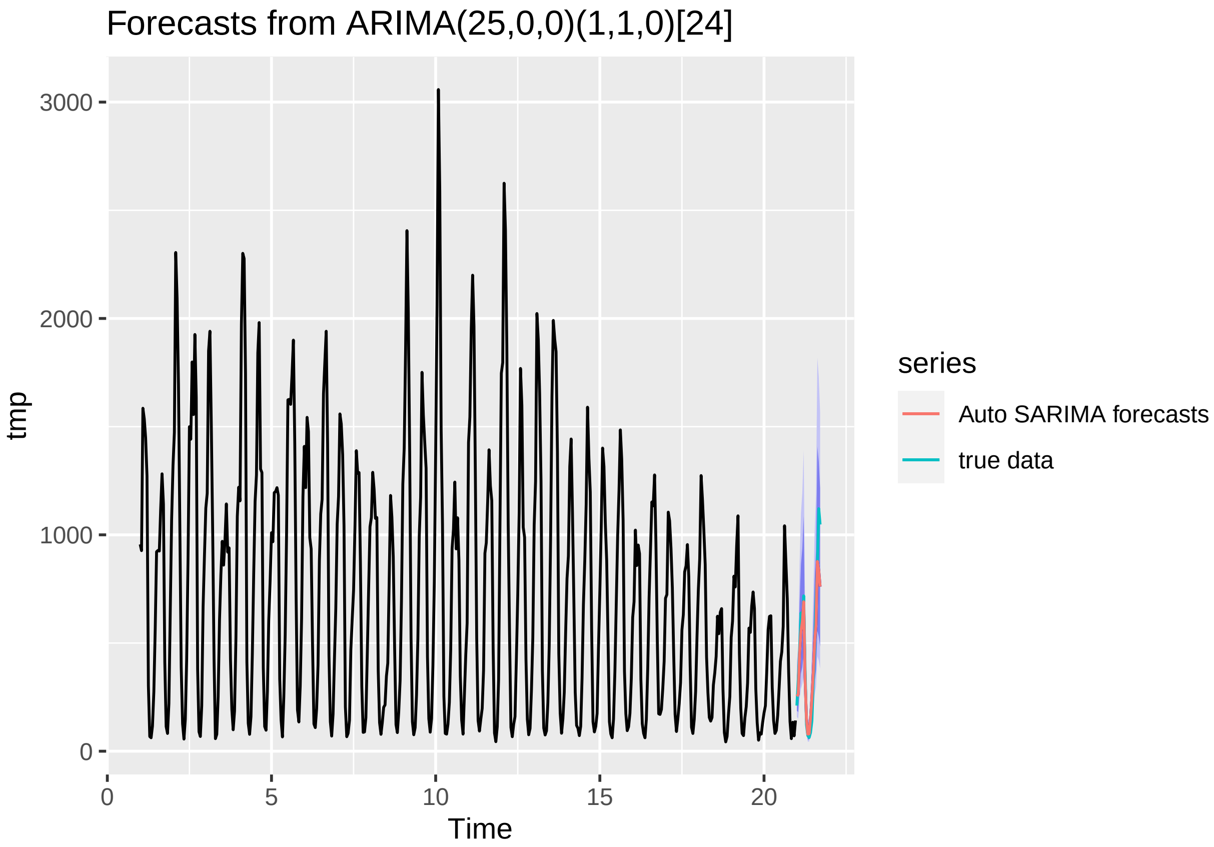
Similarly, we could explore there is a $SARIMA_{(25,0,0)(1,1,0)24}$ since there was also significant PACF at lag 24 and 25…

```
tmp=ts(serie_train,frequency = 24)
tmp_test=ts(serie_test,frequency = 24,start=c(21,1))
fit=Arima(tmp, order=c(25,0,0), seasonal=c(1,1,0),lambda = "auto")
checkresiduals(fit)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0)(1,1,0)[24]
## Q* = 68.702, df = 22, p-value = 1.057e-06
##
## Model df: 26.   Total lags used: 48
```

```
prev=forecast(fit,h=18)
autoplot(prev) + autolayer(tmp_test, series="true data")+
  autolayer(prev$mean, series="Auto SARIMA forecasts")
```



```
cat('RMSE with a SARIMA(25,0,0)(1,1,0)24 :',sqrt(mean((tmp_test-prev$mean)^2)),'\n')
```

```
## RMSE with a SARIMA(25,0,0)(1,1,0)24 : 109.066
```