NAVILLI CORRADO

S291251

S291251@studenti.polito.it

Politecnico di Torino

NILE TIME SERIES :ARIMA MODELING USING R SOFTWARE

PROJECT FOR THE COURSE “STOCHASTIC PROCESSES AND TIME SERIES”

A.A 2021/2022

PROF. ENRICO BIBBONA

PROF. DANIELE CAPPELLETTI

1 INTRODUCTION

The *Nile* dataset is a dataset available in the R software package. It consists of a *time series* object containing 100 measurements about the annual flow of the river Nile at Aswan, from 1871 to 1970. The quantities are expressed in $10^8\ m^3$.

The objective of this work is to model the behaviour of such data through an ARIMA(*p,d,q*) model.

For the purpose, the Rstudio IDE has been used with installed R version 4.1.2. The following packages have been used: *forecast*, *urca*.

2 STATIONARITY EVALUATION

The first step in the analysis is to check whether the available time series is a stationary process or not. In the latter case, some transformations are required before being able to build a reliable ARIMA model.

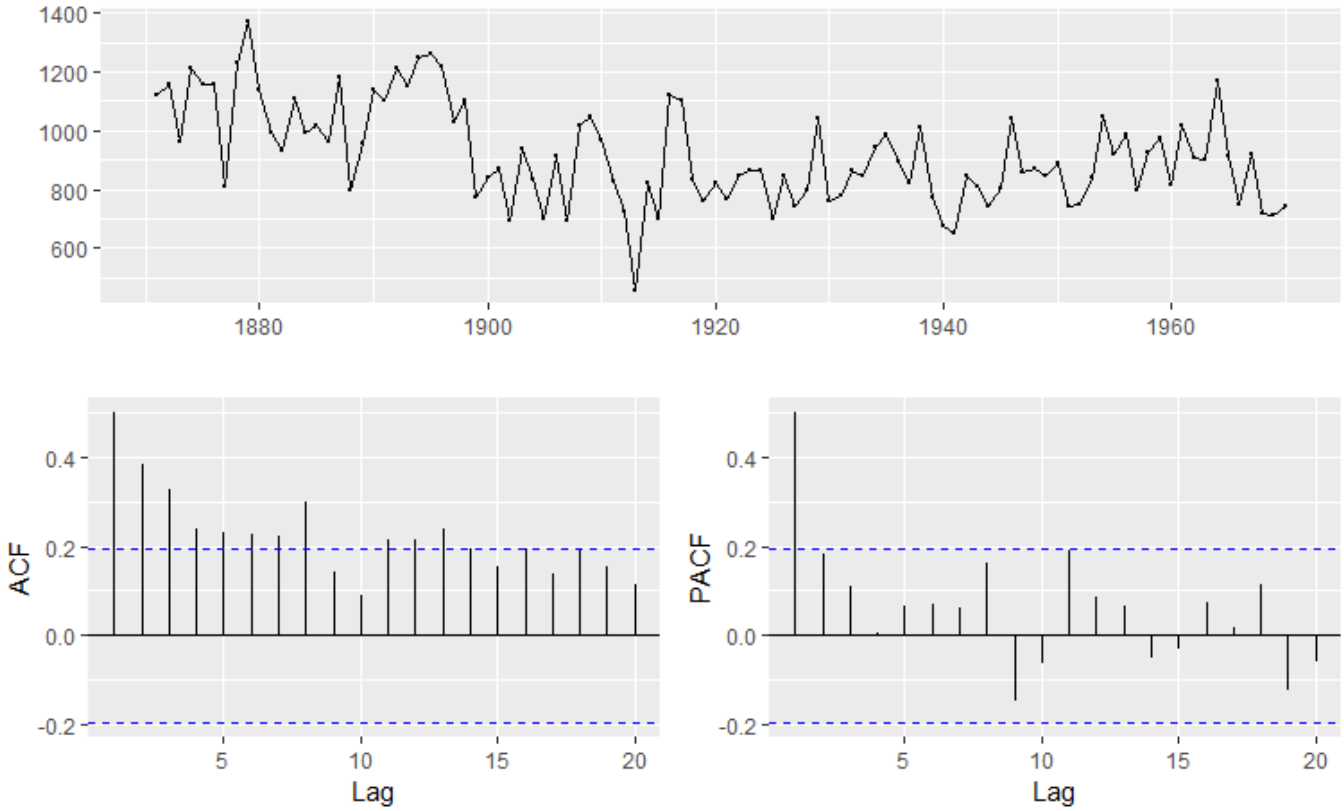


Fig 1. Plots of the time series “Nile” (above) and of its autocorrelation function (bottom left) and partial autocorrelation function (bottom right).

By looking at the shape of the time series, it looks like that the stationarity requirement is almost satisfied. Indeed, the following holds:

- *Seasonality*: it is not visible a seasonal component;
- *Variability*: data seem to have constant variance;

One last element should be checked and it is the presence of a *trend*, that would correspond to a non-constant mean value (and a consequent violation of the stationarity definition). In this case a clear trend is not evident, but if we split the data in two groups at around year 1900 we could notice a higher mean in the first series and a lower in the second, resulting in a light trend on the overall series.
The ACF plot gives another suspect of non-stationarity, since the decrease of the length of the bars is not as fast as it would be expected.

In order to have some numerical confirmation of the suspects, an hypothesis testing approach may be helpful.

The package *urca* (Unit Root and Cointegration Tests for Time Series Data) contains some functions for hypothesis testing and in particular the one for *Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Unit Root* test will be used. It tests the null hypothesis of "stationary data" against the alternative "non stationary data". In order to reject the null hypothesis, the test statistics should be under the threshold of 1% (p-value 0.05).
In this case the result is very close to the threshold, confirming the uncertainty of the situation: 0.9654.
The analysis made so far would suggest to try to differentiate once the data.

2.1 DIFFERENCING

In the context of observation in discrete time, differencing means taking the difference between two ("first differencing") or more consecutive measurements instead of the measurements themselves.
Given the original values X_t, the (first) differenced ones can be denoted as:

$$X'_t = X_t - X_{t-1}$$

Consequently, the first differenced series contains one less point. Another notation uses the backshift operator and can be used to better generalize the degree of differencing (here denoted as *d*):

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

Differencing may be used to reach stationarity by removing trend and handling seasonality.
In this case, a first differencing has been applied on original Nile data. The plot of the new data are showed in the following image:

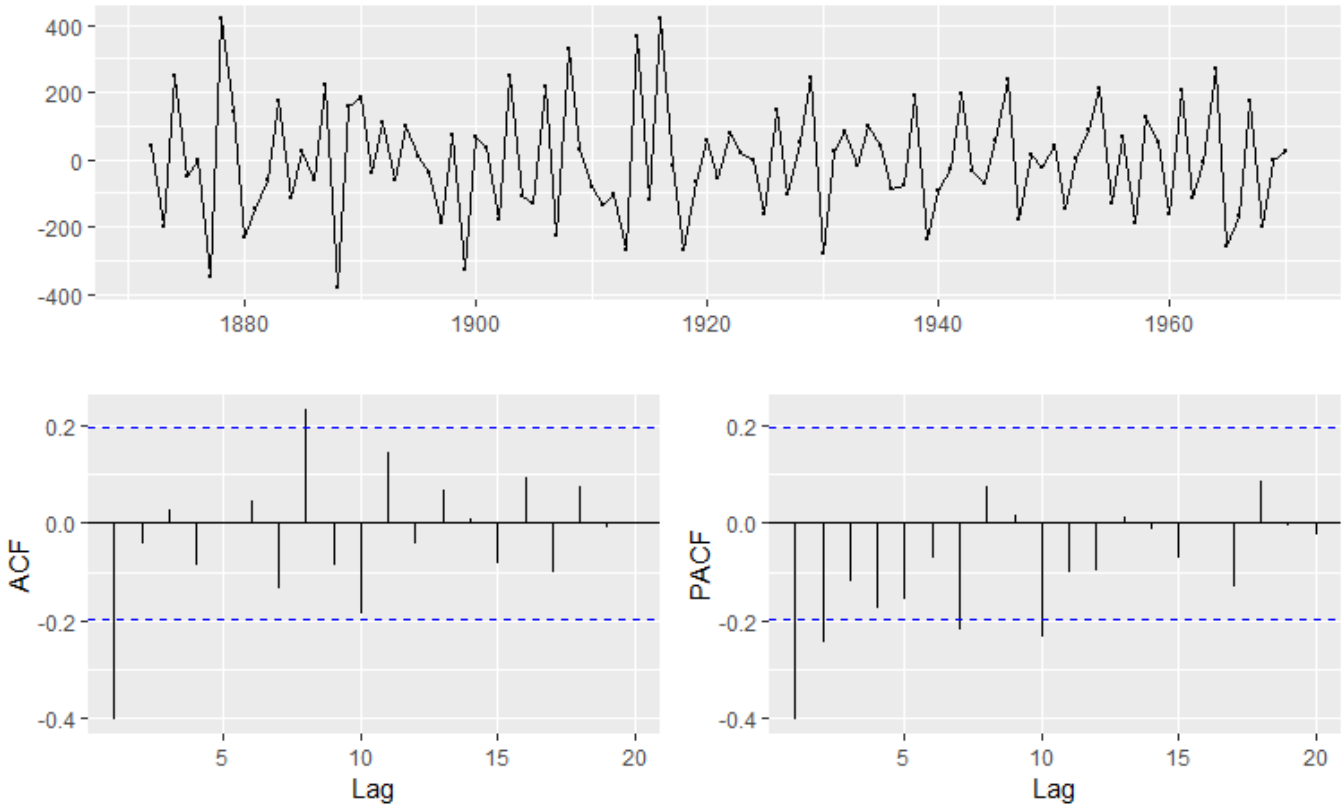


Fig 2. Plots of the differenced Nile series (above) and of its autocorrelation function (bottom left) and partial autocorrelation function (bottom right).

Now the suspects that emerged before from the time series plot and the ACF plot are not visible anymore and the series looks stationary.
The unit root test is used again in order to confirm such hypothesis and this time the test statistics shows a value of 0.0233 that is definitely under the threshold. The series in now stationary and the modelling phase may start.

3 MODEL SELECTION

The choice of the model is not trivial and it requires the trial of different hypothetical models to be then individually checked (for their validity) and compared to each other for finding the best among the valid ones.
As starting point of the discussion, we are looking for the type of model ARIMA(p,d,q) to be trained. An ARIMA model is an ARMA(p,q) with the addition of an integration factor that may be specified when working on not-differenced data (d=0 in our case, since the data have been previously differenced).
The analytical form of ARMA(p,q) process is the following:

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + c$$

Where ε_t is centered (mean=0) white noise.
The process can be also written as:

$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t$$

[1]The choice of *p, q* may be guided by the analysis on the ACF and PACF plots and they regulate the component of "auto regression" and "moving average" inside the model. In this work, the constant element *c* has been set to 0 (include.mean=False).

The ACF plots the correlation coefficient against the lag, therefore it is a graphical representation of the autocorrelation.
If it is sinusoidal or exponential decaying and the PACF shows no significant peaks after a certain lag (or also defined as "go to zero abruptly") an AR model may be more appropriate, with the parameter *p* corresponding to the lag of such last peak.

The PACF plots the partial autocorrelation, that is to say the correlation between the series and a lagged version of itself. Basically when computing the correlation between not adjacent points (lag > 1) the components of the correlation with the points contained between the two is included and therefore there is not a proper measure of the "larger-lagged" correlation. For example, in the correlation computed between X_t and X_{t-2} it would be included the correlations between X_t and X_{t-1} and between X_{t-1} and X_{t-2}. The partial autocorrelation remove such components.
If the PACF plot is sinusoidal or exponential decaying and the ACF goes to zero abruptly a MA model may be more appropriate, with the parameter q corresponding to the lag of such last peak.

In this specific case, it is not easy to state what are the best options also because the dataset is not very large, therefore such behaviors in the shape of the plots are not evident.
The ACF plot does not have significant peaks after lag 1, therefore as MA parameter only *q*=1 will be considered.
The PACF plot has instead a slower decay in the shape, and the last significant peak is at lag 2 (excluding some peaks at lag 7 and 10). Therefore as AR parameters, *p*=1,2 will be considered.
As last model, one containing both AR and MA components will be considered (with *p*=1 and *q*=1).

The following sub-paragraph will be structured as follows:

- Analytical form of the model (in our specific case, X_t actually represent with the differenced data X'_t)
- Estimates of the parameters;
- Evaluation of the residuals distribution through plots: they are supposed to be normally distributed;
- Evaluation of the autoregressive components: they are supposed to stay within the significance intervals (meaning they are not significant different from zero);
- Plot of the AR / MA inverse roots: in order for the model to be stationary the *p* complex roots of the polynomial $\Phi(B)$ should lie outside the unit circle, while the invertibility conditions are that the *q* complex roots of the polynomial $\Theta(B)$ lie outside the unit circle (the plots show the opposite: the inverse root may therefore lie within the unit circle).

A final recap with the results will be used to compare the different models in order to choose the best.

3.1 MA(1)

The following is the analytical form of a MA(1) model:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

The current observation X_t is described as a moving average of the error related to the previous observation.

The training gave the following results:

$$\theta_1 = -0.732$$

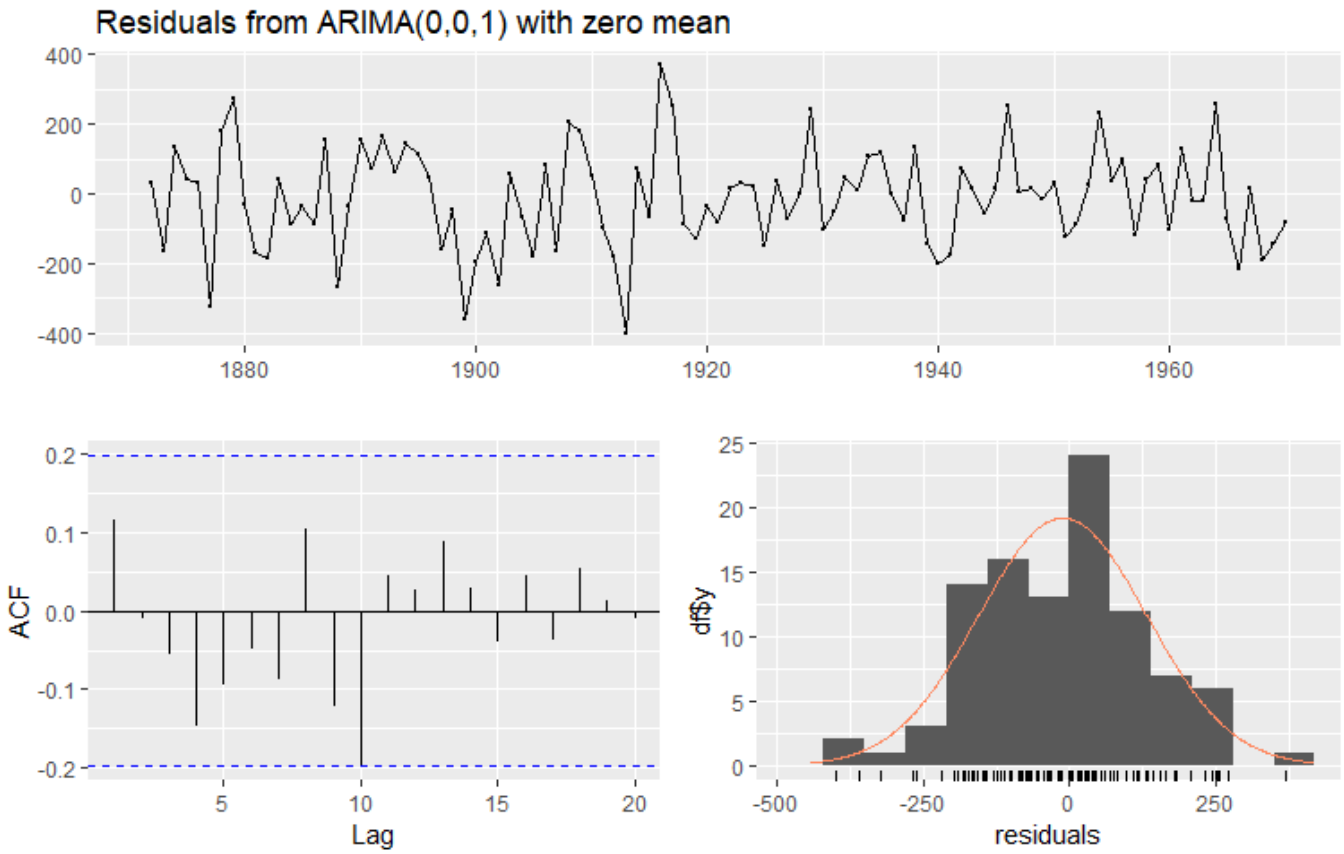


Fig 3. The residuals for MA(1) look normally distributed. The bars in the ACF plot are all within the boundaries.

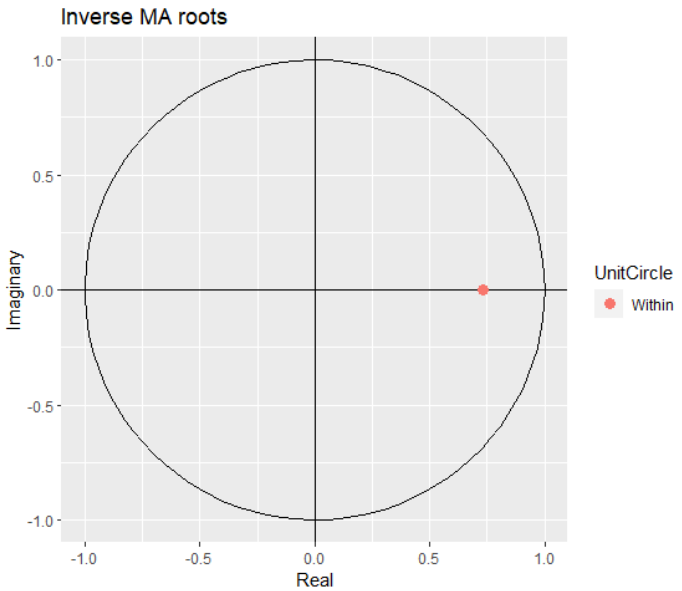


Fig 4. Inverse MA root lies within the unit circle

3.2 AR(1)

The following is the analytical form of a AR(1) model:

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t$$

The current observation is correlated to the value of the previous one, plus an error component.
The training gave the following results:

$$\varphi_1 = -0.3975$$

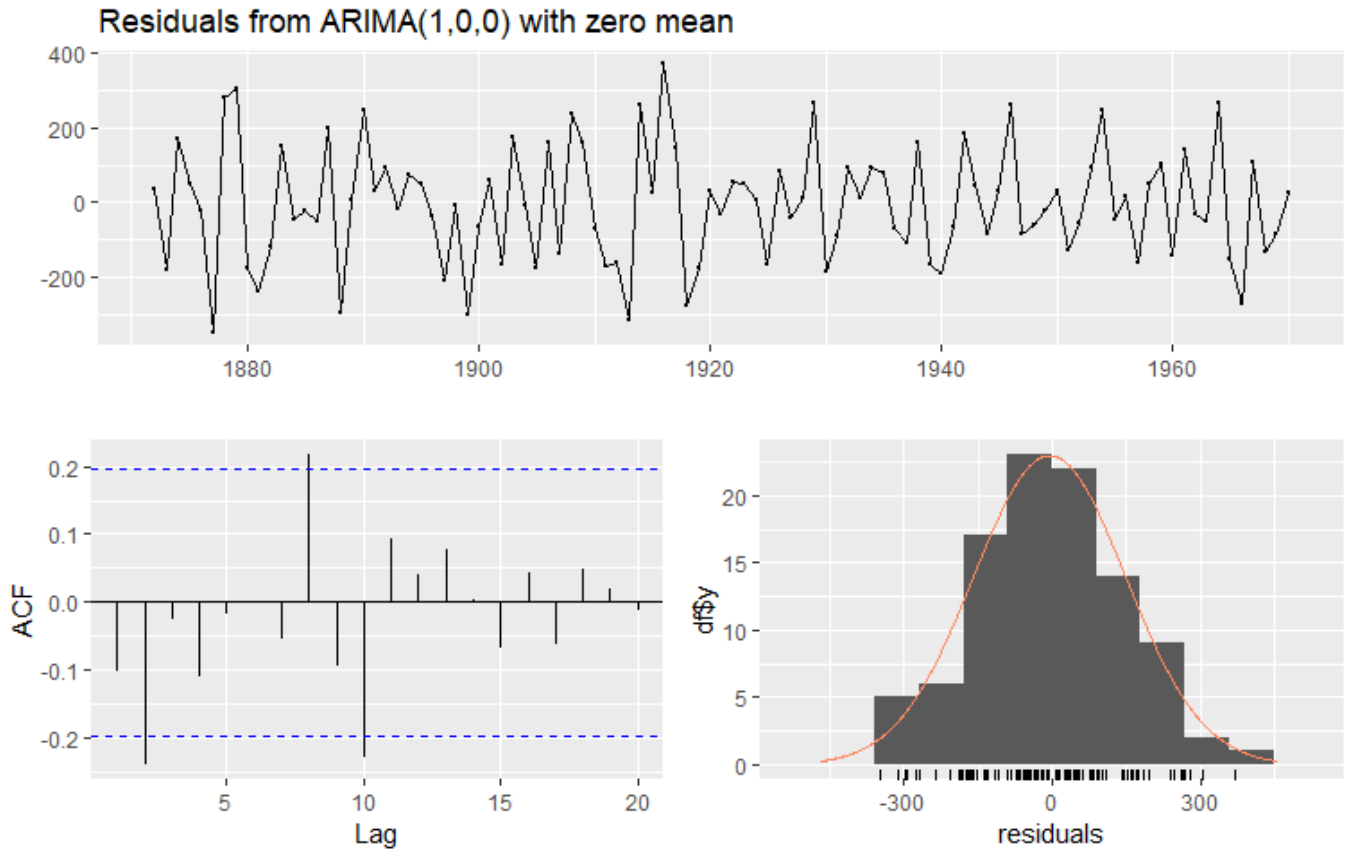


Fig 5. The check on the residuals for AR(1) is positive. About the ACF, bar at lags 2, 8, 10 exceed the bounds but the total number should still be ok.

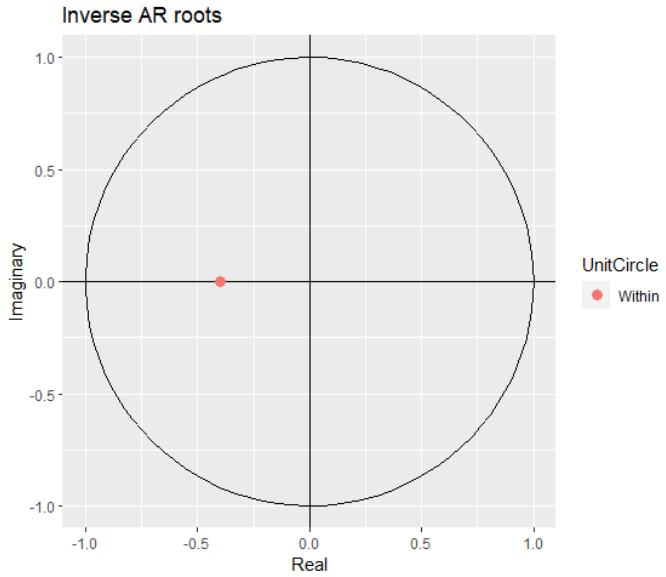


Fig 6. Inverse AR root lies within the unit circle

3.3 AR(2)

The following is the analytical form of a AR(2) model:

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \varepsilon_t$$

The current observation is correlated to the value of the previous one and the one before it, plus an error component. The training gave the following results:

$$\varphi_1 = -0.4949$$
$$\varphi_2 = -0.2417$$

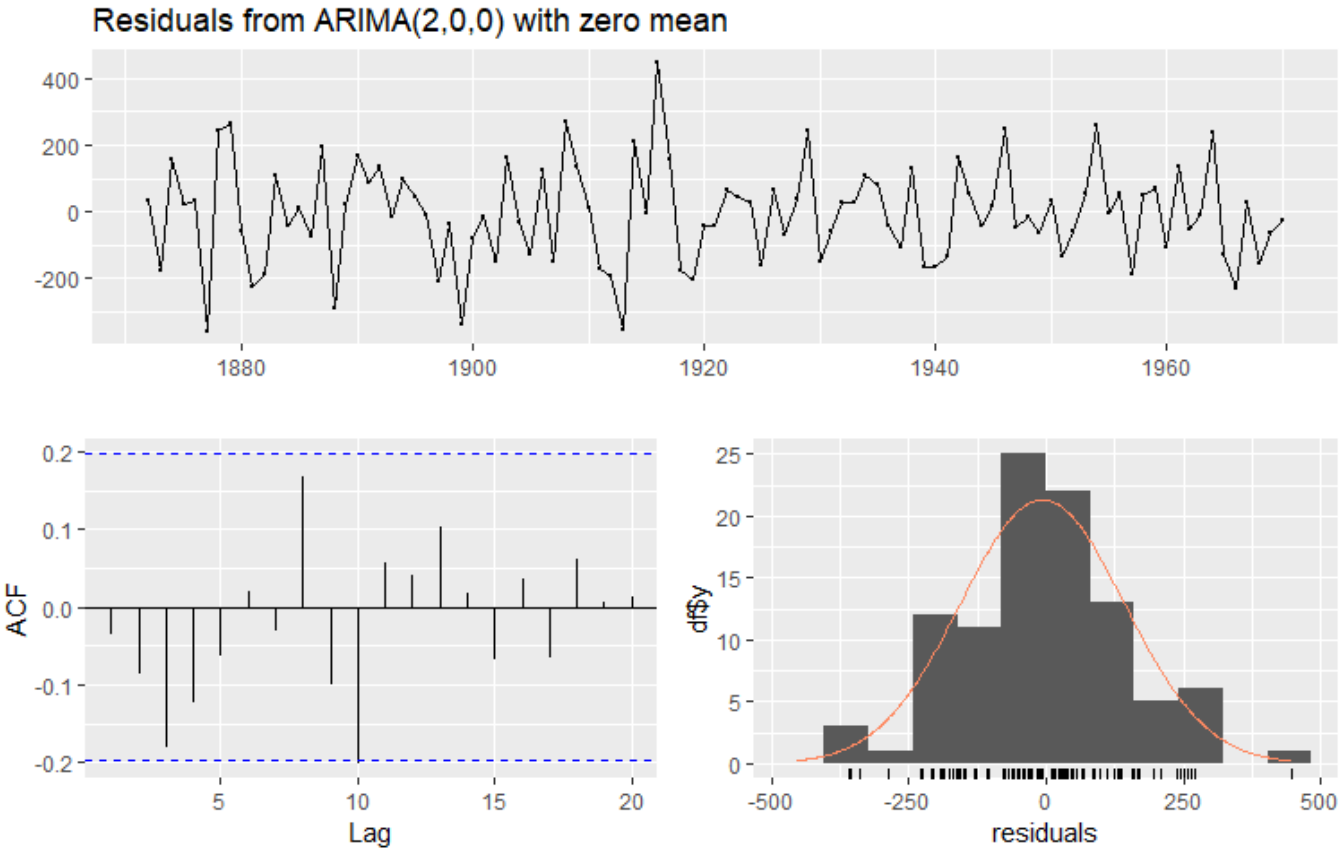


Fig 7. The check on the residuals for AR(2) is positive. The bars in the ACF plot are all within the boundaries

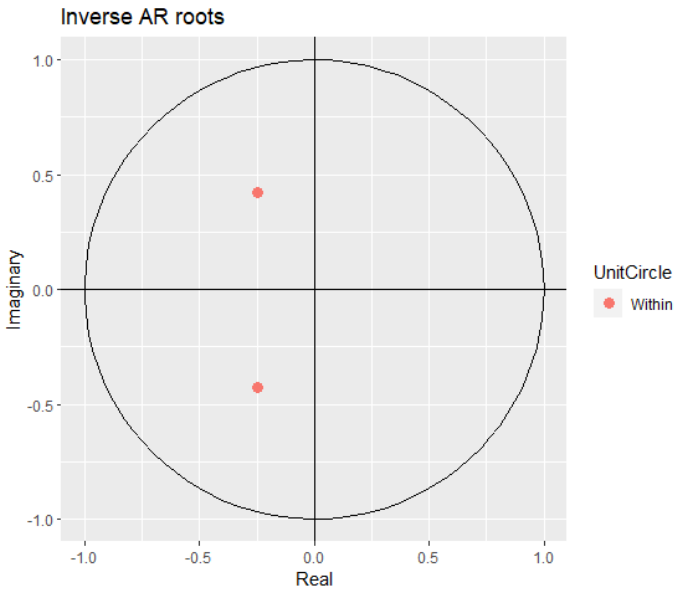


Fig 8. Inverse AR roots lie within the unit circle

3.4 ARMA(1,1)

The following is the analytical form of a ARMA(1,1) model:

$$X_t - \varphi_1 X_{t-1} = \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

The training gave the following results:

$$\varphi_1 = 0.2544$$
$$\theta_1 = -0.8741$$

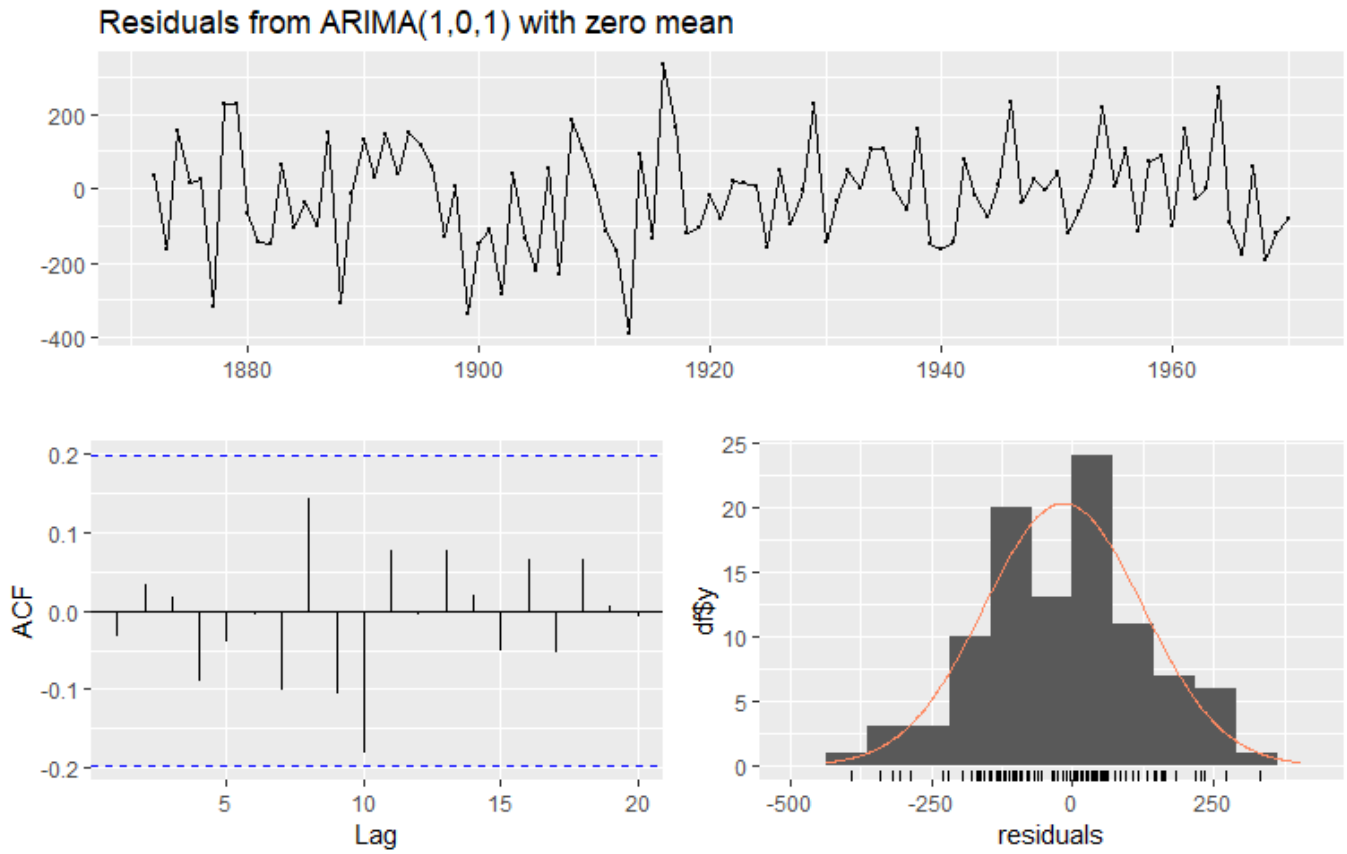


Fig 5. The check on the residuals for ARMA(1,1) is positive. The bars in the ACF plot are all within the boundaries.

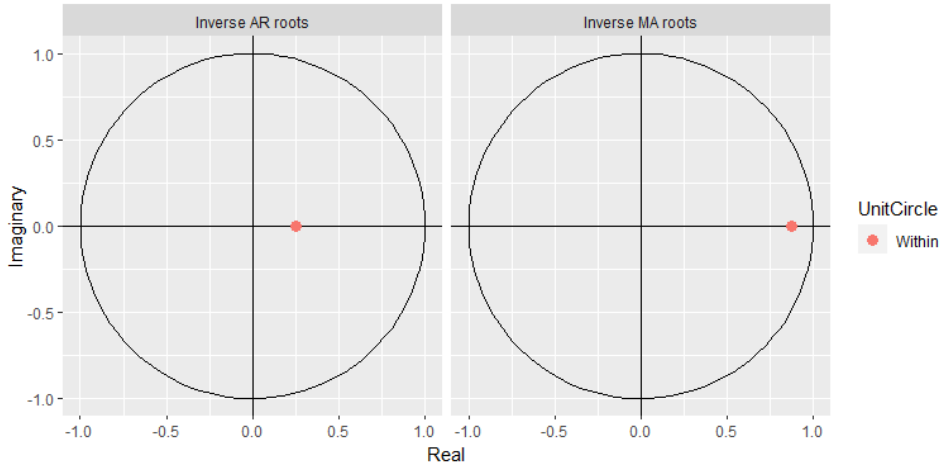


Fig 10. Inverse AR root and inverse MA root lie within the unit circle

3.5 MODELS COMPARISON

When comparing different models, it is necessary to establish some quantifiable criteria for an objective choice. In this case the choice is made on the results provided by the building functions themselves under two well-known measusers.

- **Log likelihood**: measure of the model fit, desirable as higher as possible;
- **AIC**: Akaike's information criterion, representing a trade-off between model complexity (number of parameters) and model fit (maximum likelihoos), desirable as lower as possible.

The following table summarize the results:

Model	Log likelihood	AIC
MA(1)	-632,55	1269,09
AR(1)	-638,74	1281,48
AR(2)	-635,74	1277,48
ARMA(1,1)	-630,63	1267,25

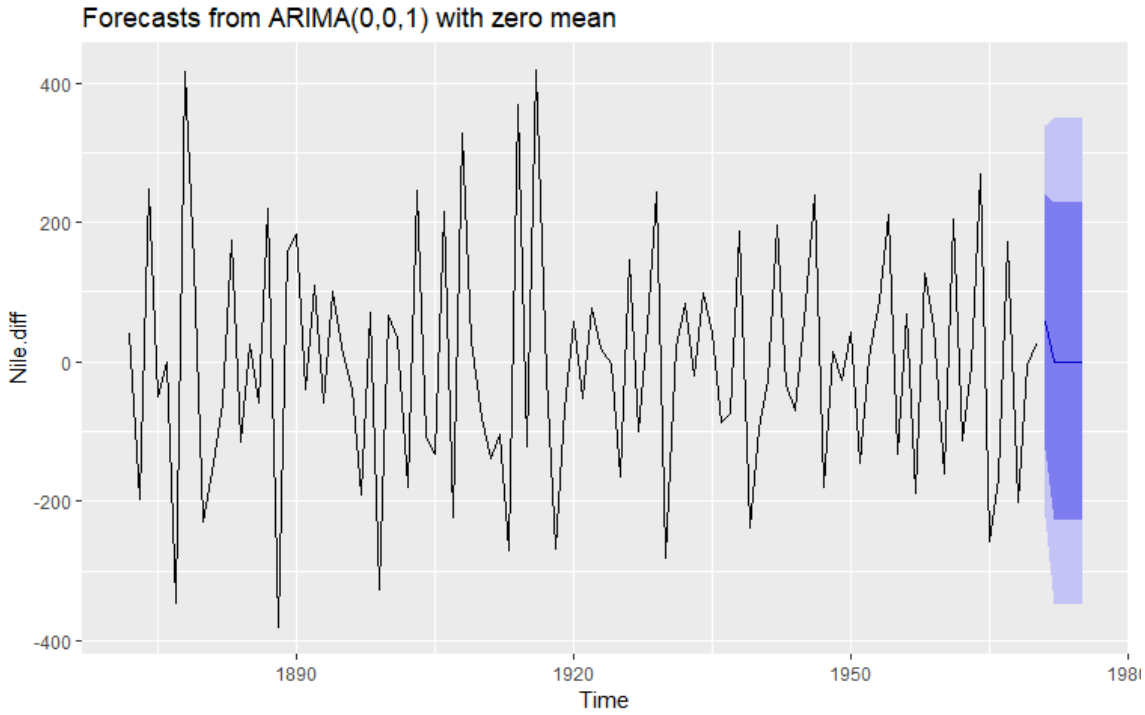
Table 1. Evaluation metrics: ARMA(1,1) is the best model

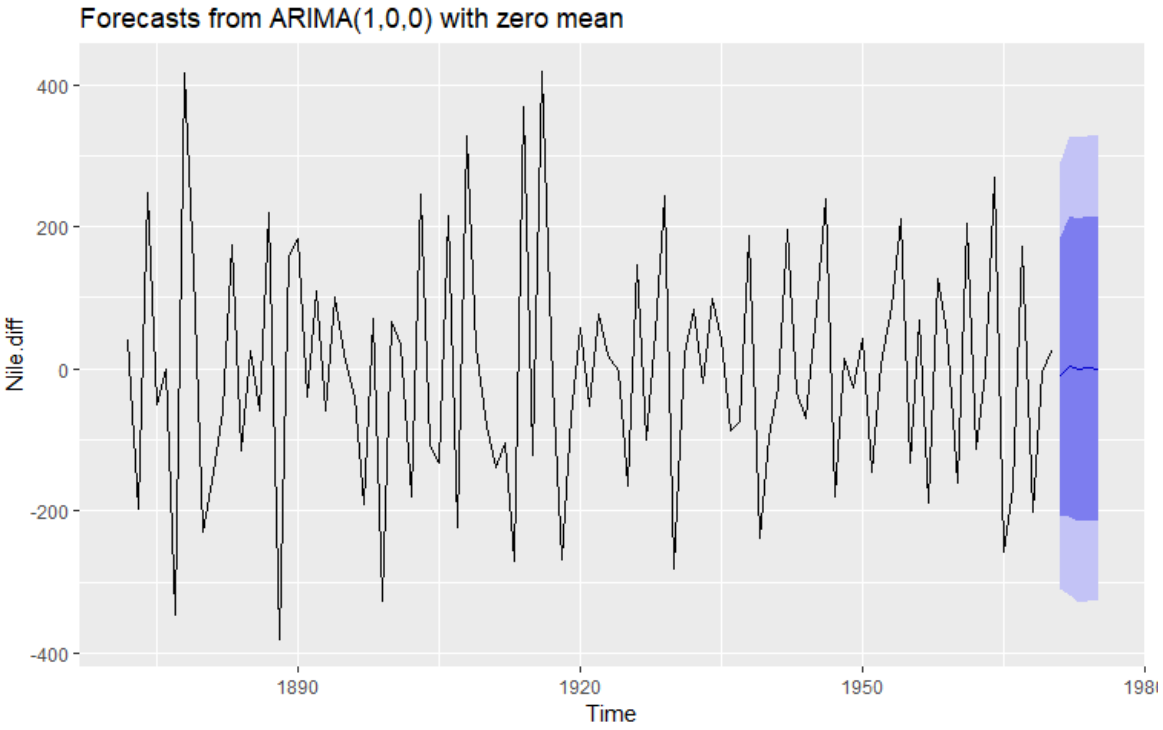
As it can be seen from the table, the best model among the four proposed is the one with both auto regression and moving average components included. This is also consistent with the difficulty found in deciding from the ACF and PACF plots the type of model: this kind of approach works better when the data are distributed according to only one of the two components.

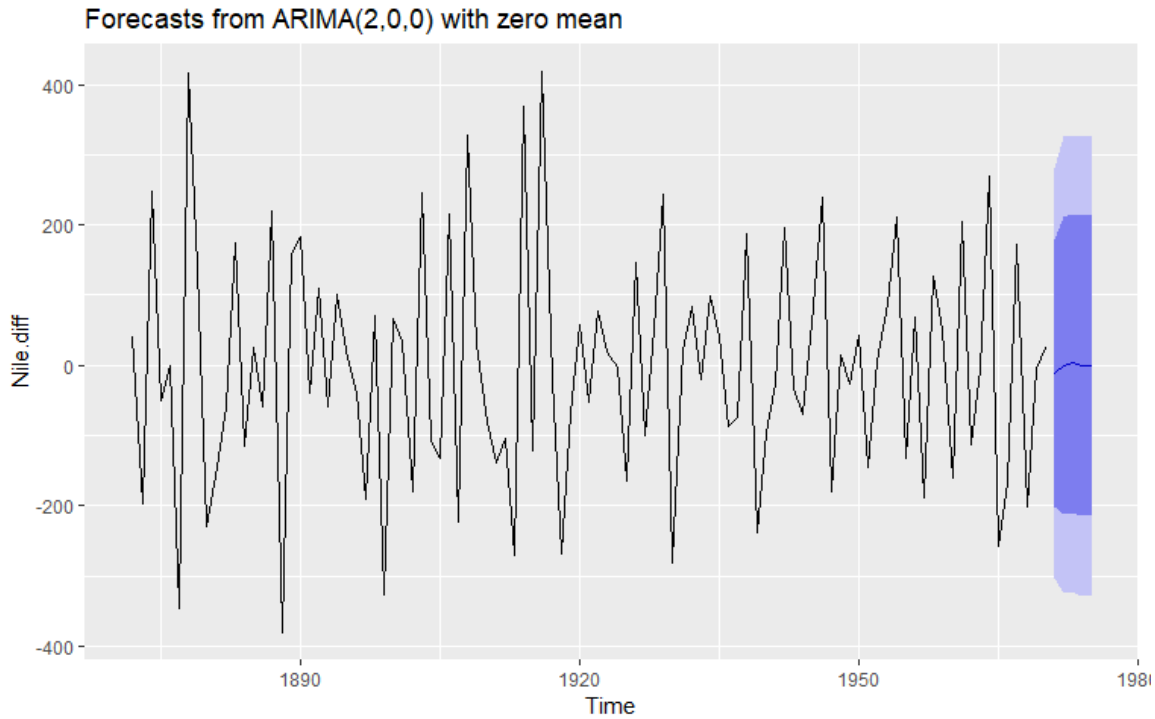
4 CONCLUSION

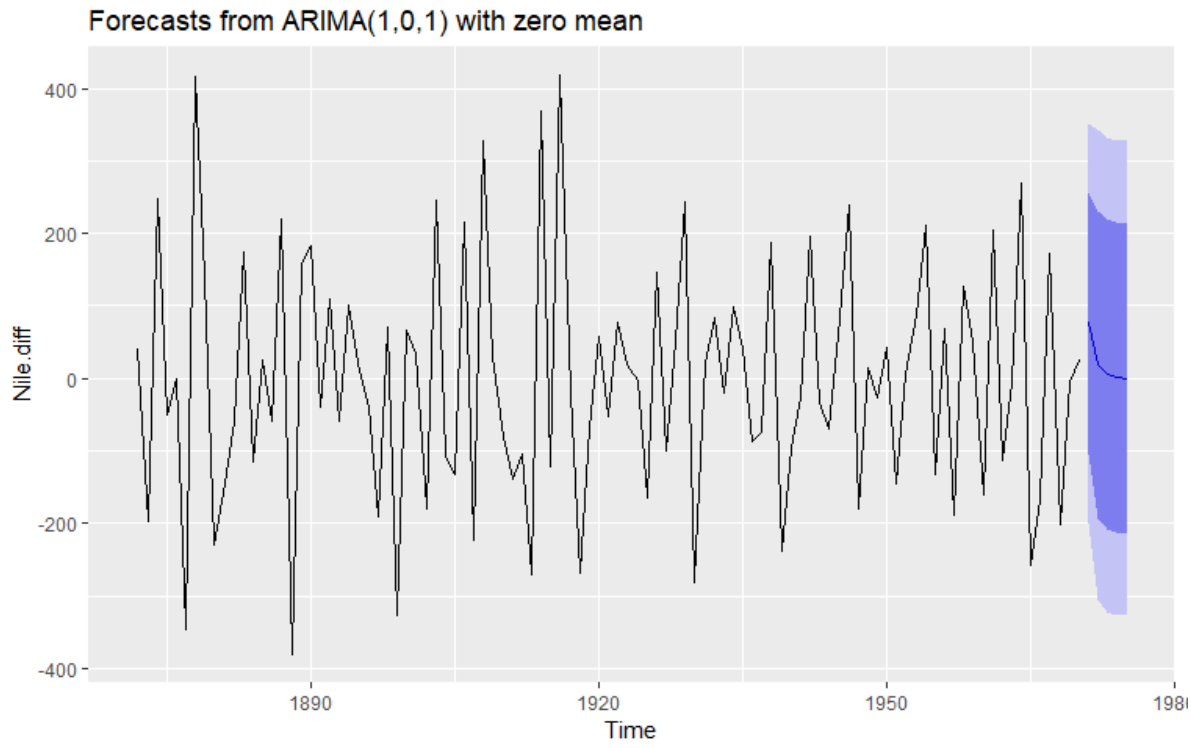
The modeling of a time series through an ARIMA model requires previous steps of pre-processing for making the series stationary.
The most challenging phase is definitely the choice of the model, not in terms of choosing among different models but in terms of choosing the hypothesis on which they rely upon (it is to say the value of the parameters *p, d, q*).

The fitted models can then be used to forecast future values, but this goes beyond the scope of this work. Here below the plots of the prediction intervals, produced thanks to a built-in function in R, in order to make some qualitative reasoning. It can be noticed that when going far away from the last observed value, the amplitude of the intervals increases (meaning that the uncertainty increases, up to becoming simple guessing).




Forecasts from ARIMA(1,0,0) with zero mean
Nile.diff
400
200
0
-200
-400
1890
1920
1950
Time


Forecasts from ARIMA(2,0,0) with zero mean
Nile.diff
400
200
0
-200
-400
1890
1920
1950
Time



5 REFERENCES

The proposed approach has followed the on-line available book: https://otexts.com/fpp2/arima.html

[1]https://medium.com/@ooemma83/how-to-interpret-acf-and-pacf-plots-for-identifying-ar-ma-arma-or-arima-models-498717e815b6#:~:text=The%20basic%20guideline%20for%20interpreting,q%20for%20MA(q)