Source:

<http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html#forecasting-using-an-arima-model>

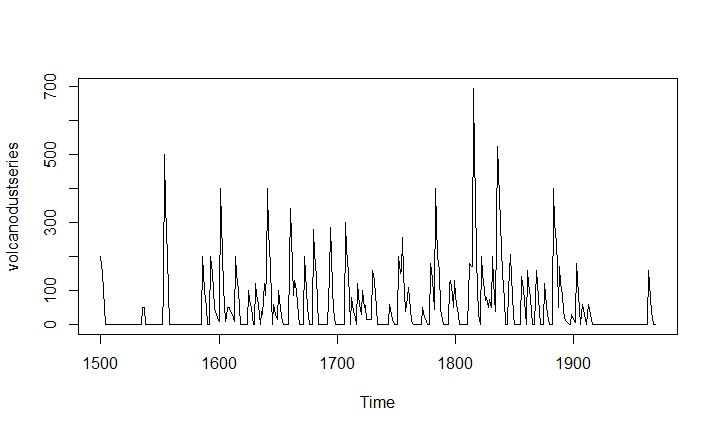
Autoregressive Integral Moving Average (ARIMA) models include a statistical model for the irregular component of a time series, that accounts for non-zero autocorrelations in the irregular component.

As ARIMA models are defined for stationary time series, in case we start off with a non-stationary time-series, we must ‘difference’ it until we obtain a stationary time series.

To demonstrate the selection of an appropriate ARIMA model, we use the volcano dust data set at:

<http://robjhyndman.com/tsdldata/annual/dvi.dat>

## Exploratory Analysis

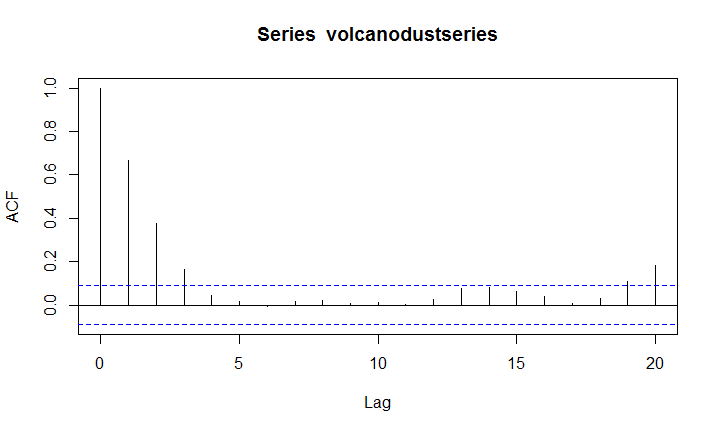
An exploratory plot of the time series:

From the plot, we can see that the random fluctuations are of approximately equal size, so an additive model is appropriate.

Also, the level and variance of the time series are constant over time (i.e. stationary). There is hence no need to difference it, and we can fit an ARIMA model to the time series as-is (order of differencing required = d = 0).

Correlogram, and values of the autocorrelations:

|  |  |
| --- | --- |
| 0 | 1 |
| 1 | 0.666 |
| 2 | 0.374 |
| 3 | 0.162 |
| 4 | 0.046 |
| 5 | 0.017 |
| 6 | -0.007 |
| 7 | 0.016 |
| 8 | 0.021 |
| 9 | 0.006 |
| 10 | 0.01 |
| 11 | 0.004 |
| 12 | 0.024 |
| 13 | 0.075 |
| 14 | 0.082 |
| 15 | 0.064 |
| 16 | 0.039 |
| 17 | 0.005 |
| 18 | 0.028 |
| 19 | 0.108 |
| 20 | 0.182 |

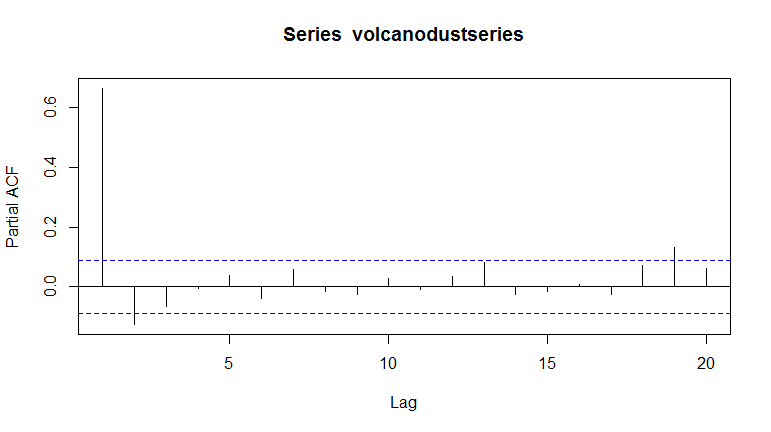


We see from the correlogram that the autocorrelations for lags 1, 2 and 3 exceed the significance level (indicated by the dotted blue line), and that the autocorrelations decrease to zero after lag 3. The autocorrelations for lags 1, 2, 3 are positive, and decrease in magnitude with increasing lag (lag 1: 0.666, lag 2: 0.374, lag 3: 0.162).

The autocorrelation for lags 19 and 20 also exceed significance level – but only just. It is likely noise.

Partial autocorrelogram and partial autocorrelation values:

|  |  |
| --- | --- |
| 1 | 0.666 |
| 2 | -0.126 |
| 3 | -0.064 |
| 4 | -0.005 |
| 5 | 0.04 |
| 6 | -0.039 |
| 7 | 0.058 |
| 8 | -0.016 |
| 9 | -0.025 |
| 10 | 0.028 |
| 11 | -0.008 |
| 12 | 0.036 |
| 13 | 0.082 |
| 14 | -0.025 |
| 15 | -0.014 |
| 16 | 0.008 |
| 17 | -0.025 |
| 18 | 0.073 |
| 19 | 0.131 |
| 20 | 0.063 |



Since the correlogram decreases to zero after lag 3, and the partial correlogram after lag 2, the following ARIMA models are possible:

* an ARMA(2,0) model, since the partial autocorrelogram is zero after lag 2, and the correlogram tails off to zero after lag 3, and the partial correlogram is zero after lag 2
* an ARMA(0,3) model, since the autocorrelogram is zero after lag 3, and the partial correlogram tails off to zero (although perhaps too abruptly for this model to be appropriate)
* an ARMA(p,q) mixed model, since the correlogram and partial correlogram tail off to zero (although the partial correlogram perhaps tails off too abruptly for this model to be appropriate)

The ARMA(2,0) and ARMA(p,q) are equally good models – and better than ARMA(0,3) since they need one fewer parameter (2 rather than 3).

An ARMA(2,0) model is an autoregressive model of order 2:

X\_t - mu = (Beta1 \* (X\_t-1 - mu)) + (Beta2 \* (Xt-2 - mu)) + Z\_t

where

X\_t : stationary time series (volcanic dust veil index)

mu : estimated mean of time series X\_t

Beta1, Beta2 : parameters to be estimated

## Fitting an ARIMA model

If an ARMA(2,0) model is used – where p = 2, q = 0

it implies that an ARIMA(2,0,0) can be used, where p = 2, d = 0, q = 0

(d: order of differencing required)

On fitting an ARIMA model, we get the following coefficients:

Coefficients:

ar1 ar2 intercept

0.7533 -0.1268 57.5274

s.e. 0.0457 0.0458 8.5958

sigma^2 estimated as 4870: log likelihood = -2662.54, aic = 5333.09

## Forecasting

On forecasting for the next 31 years (1970-2000) we get the following set of values:

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

1970 21.48131 -67.94860 110.9112 -115.2899 158.2526

1971 37.66419 -74.30305 149.6314 -133.5749 208.9033

1972 47.13261 -71.57070 165.8359 -134.4084 228.6737

1973 52.21432 -68.35951 172.7881 -132.1874 236.6161

1974 54.84241 -66.22681 175.9116 -130.3170 240.0018

1975 56.17814 -65.01872 177.3750 -129.1765 241.5327

1976 56.85128 -64.37798 178.0805 -128.5529 242.2554

1977 57.18907 -64.04834 178.4265 -128.2276 242.6057

1978 57.35822 -63.88124 178.5977 -128.0615 242.7780

1979 57.44283 -63.79714 178.6828 -127.9777 242.8634

1980 57.48513 -63.75497 178.7252 -127.9356 242.9059

1981 57.50627 -63.73386 178.7464 -127.9145 242.9271

1982 57.51684 -63.72330 178.7570 -127.9040 242.9376

1983 57.52212 -63.71802 178.7623 -127.8987 242.9429

1984 57.52476 -63.71538 178.7649 -127.8960 242.9456

1985 57.52607 -63.71407 178.7662 -127.8947 242.9469

1986 57.52673 -63.71341 178.7669 -127.8941 242.9475

1987 57.52706 -63.71308 178.7672 -127.8937 242.9479

1988 57.52723 -63.71291 178.7674 -127.8936 242.9480

1989 57.52731 -63.71283 178.7674 -127.8935 242.9481

1990 57.52735 -63.71279 178.7675 -127.8934 242.9481

1991 57.52737 -63.71277 178.7675 -127.8934 242.9482

1992 57.52738 -63.71276 178.7675 -127.8934 242.9482

1993 57.52739 -63.71275 178.7675 -127.8934 242.9482

1994 57.52739 -63.71275 178.7675 -127.8934 242.9482

1995 57.52739 -63.71275 178.7675 -127.8934 242.9482

1996 57.52739 -63.71275 178.7675 -127.8934 242.9482

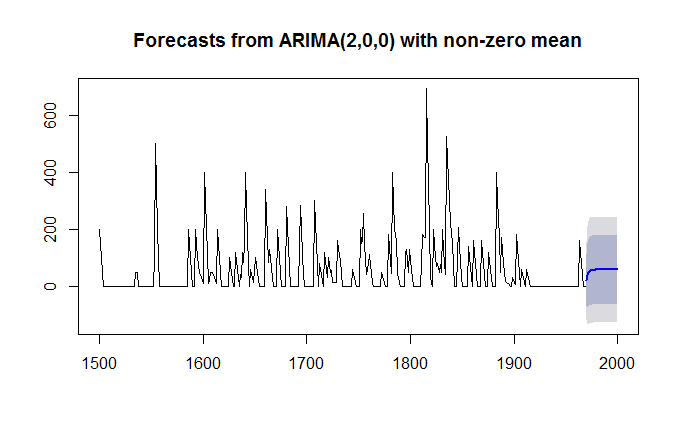
1997 57.52739 -63.71275 178.7675 -127.8934 242.9482

1998 57.52739 -63.71275 178.7675 -127.8934 242.9482

1999 57.52739 -63.71275 178.7675 -127.8934 242.9482

2000 57.52739 -63.71275 178.7675 -127.8934 242.9482

We can compare the dataset given with the forecasts by plotting them together:

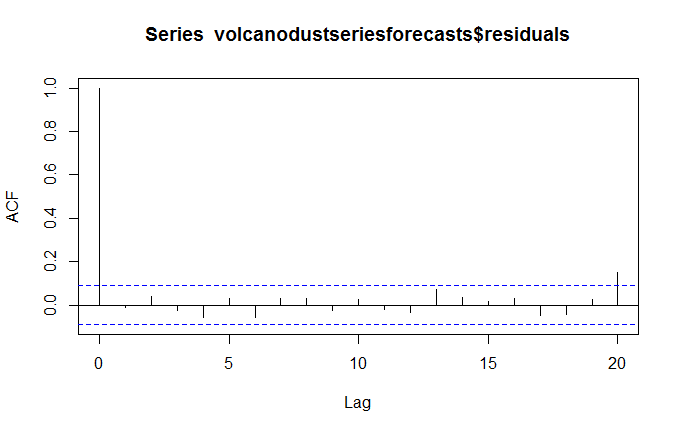


The model predicts positive as well as negative values because the non-negativity of the predicted variable is not accounted for, in this model. This is a limitation of the model.

## Error Diagnostics

We examine whether the errors are correlated, and whether they are normally distributed with mean zero.

For correlation between successive forecast errors we make a correlogram and use the Ljung-Box test.



Box-Ljung test

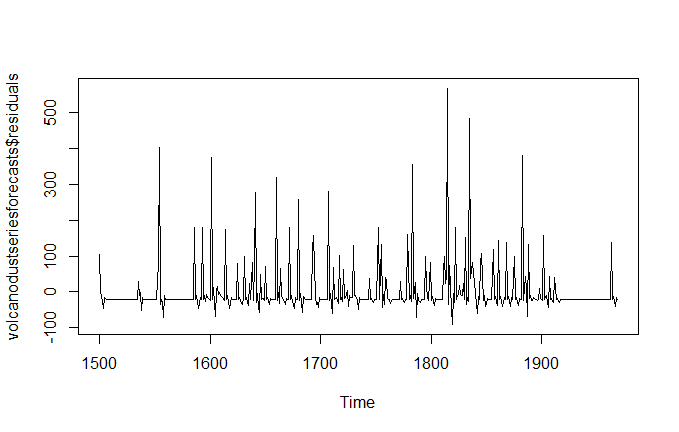
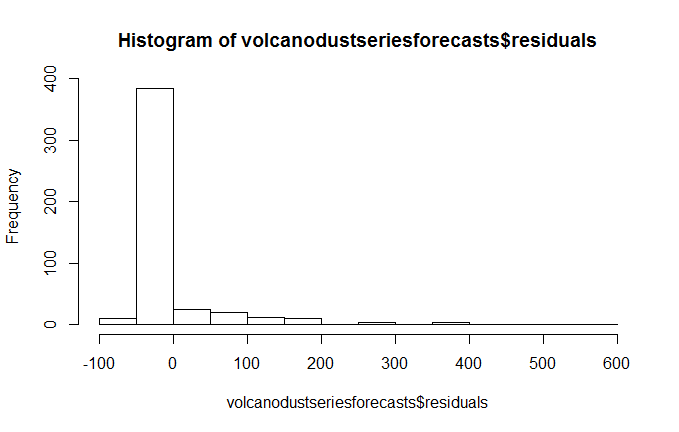
data: volcanodustseriesforecasts$residuals

X-squared = 24.364, df = 20, p-value = 0.2268

As seen in the correlogram, the sample autocorrelation at lag 20 exceeds the significance level. But this is permissible under 95% significance – one of 20 autocorrelations can exceed the significance level.

The p-value for the Ljung-Box test is 0.2 - hence, there is little evidence for non-zero autocorrelations in the forecast errors for lags 1-20.

To check that the errors have normal distribution and mean zero:

The time plot of forecast errors show that they have approximately constant variance over time. Further, the histogram of forecast errors show that their mean is approximately zero (on calculation: -0.22)

Appendix: R code

install.packages("forecast")

library(forecast)

#Reading the dataset

volcanodust <- scan("http://robjhyndman.com/tsdldata/annual/dvi.dat", skip=1)

#creating and plotting the volcano dust veil levels as a time series

volcanodustseries <- ts(volcanodust,start=c(1500))

plot.ts(volcanodustseries)

#autocorrelation values and plot

acf(volcanodustseries, lag.max=20) # plot a correlogram

acf(volcanodustseries, lag.max=20, plot=FALSE) # get the values of the autocorrelations

#partial autocorrelation and plot

pacf(volcanodustseries, lag.max=20)

pacf(volcanodustseries, lag.max=20, plot=FALSE)

#fitting an ARIMA(2,0,0) model

volcanodustseriesarima <- arima(volcanodustseries, order=c(2,0,0))

volcanodustseriesarima

#forecasting for 31 years with the model obtained

volcanodustseriesforecasts <- forecast(volcanodustseriesarima, h=31)

volcanodustseriesforecasts

#comparing past and forecast values

plot.forecast(volcanodustseriesforecasts)

#error diagnostics - correlation

acf(volcanodustseriesforecasts$residuals, lag.max=20)

Box.test(volcanodustseriesforecasts$residuals, lag=20, type="Ljung-Box")

#error diagnostics - normal distribution and mean

plot.ts(volcanodustseriesforecasts$residuals) # make time plot of forecast errors

hist(volcanodustseriesforecasts$residuals) # make a histogram