Time series modelling of Annual Maximum Sunspot Numbers

Minoru Tanaka

Department of Network and Information, School of Network and Information,
Senshu University, Kawasaki 214-8580, Japan

Abstract. We consider the prediction of Solar cycles from the sunspot numbers (yearly averages and extreme values) based on an AR model. We focus on a time series of annual maximum sunspot numbers, and we also consider the estimation of a probability distribution function and the prediction of the annual maximum data.

Keywords: AR model, block maximum, Weibull distribution, solar cycles, sunspot numbers.

1. Introduction

We know that solar activity is not uniformly distributed in time and is apparent in the record of the number of sunspots appearing on the solar disc. There were very few sunspots during the period from 1645 to 1715 which is the so-called Maunder Minimum (see Thoms and Weiss [7]). That periods had a harmful effort on the earth. Since the sun has a great influences on the earth like the cold weather affects the crops, it must be very important for us to predict the solar cycles.

A well-known data recorded for yearly averaged of the International relations sunspot number has been analyzed by many authors and researchers. It is well-known that the sunspot has about 11-year period, which was first recognized by Schwabe in 1843 and was reported by Wolf in 1852 (Thoms and Weiss [7]). But the sunspot cycle seems to be not strictly periodic and the cycles vary both in size and length. Therefore the predicting the peaks and valleys may be very difficult. A new prediction for the next solar cycle is reported by NOAA (NASA) that it will peak in May 2013 with a below average number of sunspot of 90. The numbers of yearly averaged sunspot (from 1700 to 2010) is plotted in Figure 1.

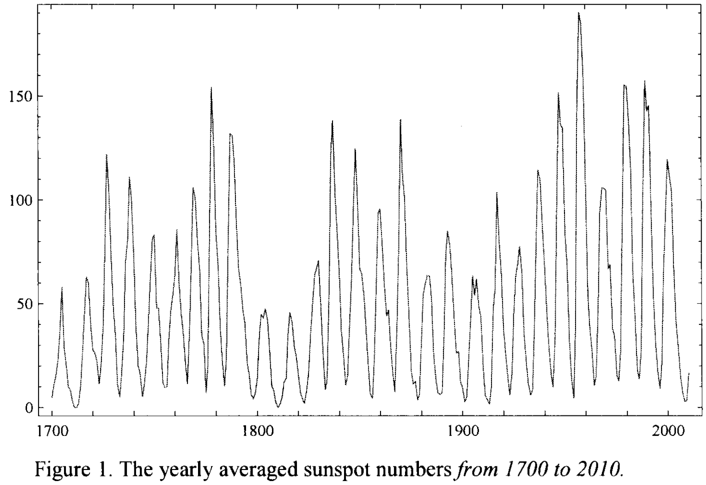

Figure 1. The yearly averaged sunspot numbers*from 1700 to 2010.*

The spectrum (spectral density function) of a stationary time series, which is the Fourier transform of the covariance function of the series, can be calculated by use of Mathematica. The nonparametric estimate of the spectrum is shown in Figure 2. The prominent peak is located at about frequency of 0.573 and then the series will oscillate with an approximate period of 11.

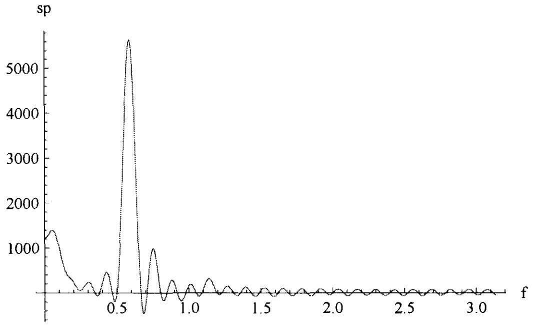

Figure 2. The sample spectrum of annual sunspot numbers.

Here we have a question whether the yearly (annual) maximum data based on the daily sunspot numbers have the similar property for the solar cycle, or not.

In this paper we first consider the prediction of Solar cycles from the sunspot numbers based on an AR time series model in the next section. In Section 3 we focus on a series of annual maximum sunspot numbers obtained from the daily averaged sunspot numbers (for the period 1 May 1749 through 2 February 2011), and we consider the estimation of a probability distribution function and the prediction of the annual maximum data.

This paper is supported by the computer software Mathematica V8.0 and its application Time Series Pack for Mathematica.

2. A Model for Annual Sunspot Numbers

The sunspot number has been analyzed by many authors and researchers. Brockwell and Davis[1] analyzed that an ARMA(3,4) with GARH(1,1) noise model is an appropriate model for annual sunspot numbers from 1700 to 1985. We also consider the modelling of the yearly averaged (annual) sunspot numbers for the interval 1700-2010 shown in Figure 1.

In order to make the variance more uniform, we first take the square of the data and subtract out the mean. We can fit the data to a stationary time series model. The behavior of the sample correlation function (correlogram) obtained from the data oscillates with an approximate cycle of period 11. Also the slowly decaying behavior of the correlogram indicates that the series can be fitted to AR model with long order.

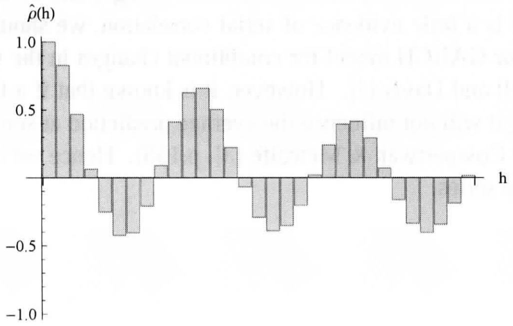

Figure 3. The sample correlation function (correlogram) of the annual sunspot numbers.

An AR model of order p is defined by

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + e_t,$$

where $\{e_t\}$ is white noise with mean zero and variance σ^2. (see, for example, [1], [3], [5])

We can estimate some nearby models to see if they give a lower AIC value. For example, we try AR(2), ... , AR(15), etc. It turns out that AR(9) has a lower AIC value, and the estimated parameters are

$$\{a_1, a_2, ..., a_9\} = \{1.228, -0.499, -0.125, 0.240, -0.224, -0.007, 0.173, -0.226, 0.301\},$$

$$\sigma^2 = 1.0635 .$$

We can see that the behavior of the sample correlation function is consistent with a correlation function of the AR(9) model.

To test the adequacy of the model we calculate the residuals of the fitted AR(9) model and the correlogram of the residuals. If the AR(9) model is suitable the residuals should appear to be a realization of white noise with zero mean.

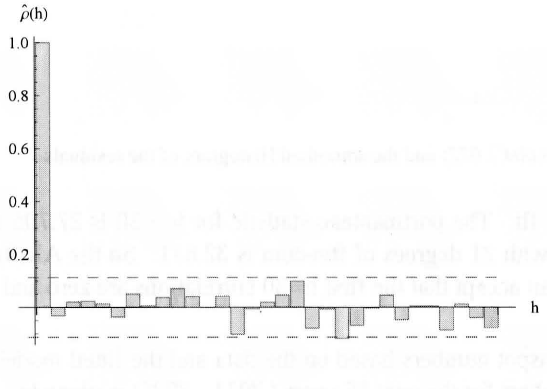

Figure 4. The Correlogram of the residuals and the *95%* bounds.

The correlogram of the residuals (Figure 4) suggests that the residuals behave like white noise. In order to detect volatility (conditional heteroskedastic) we should look at the correlogram of the squared residuals since the squared values are equivalent to the variance. From Figure 5 there is not clear evidence of serial correlation in the squared values, although the values of the correlation function at lag 1 and 10 do not fall within the 95% bounds. If we judge that there is a little evidence of serial correlation, we should fit a model to the residuals such that ARCH model or GARCH model for conditional changes in the variance. A survey for this line can be found in Brockwell and Davis [1]. However, it is known that if a GARCH model is fitted to the residuals of a fitted model, it will not influence the average prediction at some point in time since the mean of the residuals is zero (see Cowpertwait & Metcalfe [3], p.155). Hence we choose the AR(9) model with a white noise residuals for the series.

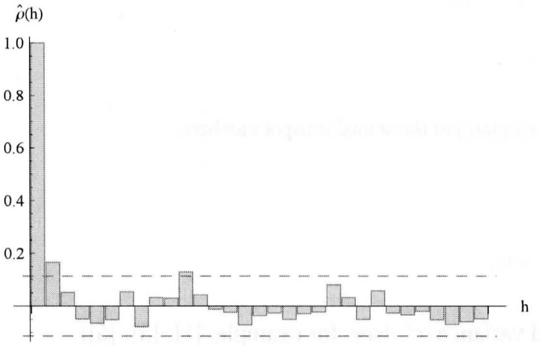

Figure 5. The Correlogram of the squared residuals and the *95%* bounds.

The parametric estimate of the probability density function is shown as a solid curve in Figure 5 together with a dotted curve which is a smoothed histogram of the residuals derived from the AR(9) model. It turns out that the density function is estimated by a normal distribution N(-0.004 , 1.072).

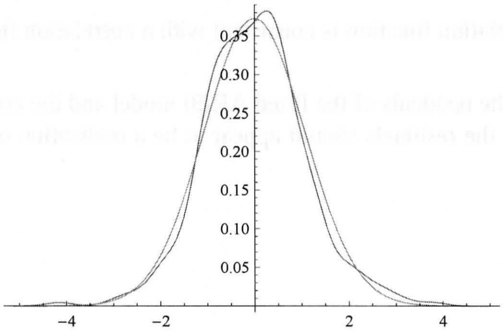

Figure 6. Density plot of normal distribution *N(-0.004,1.072)* and the smoothed Histogram of the residuals

We check the AR(9) model for goodness-of-fit. The portmanteau statistic for h = 30 is 27.705 and the quantile for 0.95 of Chi-square distribution with 21 degrees of freedom is 32.671. So the AR(9) model passes the portmanteau test. Therefore we can accept that the first h=30 correlations are zero and the fitted model AR(9) is adequate.

Next we forecast the future values of the sunspot numbers based on the data and the fitted model AR(9). The best linear prediction of the sunspot numbers for the next 15 years (2011 - 2025) is given by

```
{47.3686, 76.2924, 80.8924, 71.3874, 51.7563, 31.5554, 16.9579, 9.18953,
  8.07493, 13.7802, 29.1596, 50.5363, 68.7123, 74.6002, 66.3624}
```

We plot the annual sunspots data along with the next 15 predicted values in Figure 7.

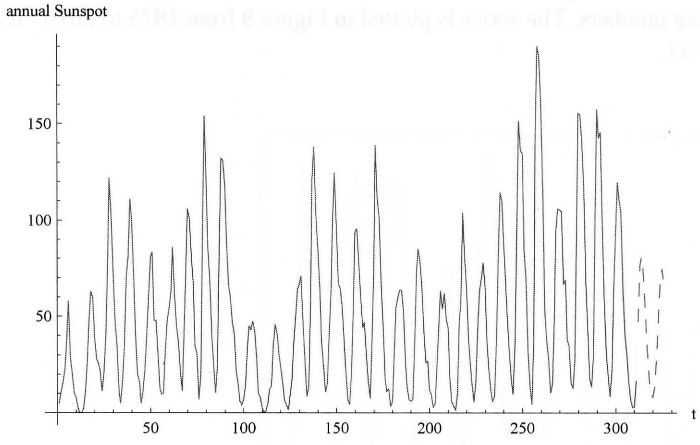

Figure 7. Time plot for last *15* years with added predicted values (dotted line)

This result shows that the next peak of the solar cycle will be in 2013 with a less sunspot number of 81. This is almost consistent with the previous report of the prediction given by NOAA (NASA).

3. Blocked Maximum Data

Here we consider a daily averages of the International relative sunspot number (published in Solar Geophysical Data available from NOAA/SEL, Boulder, Colorado). We should note that daily values for years prior to 1849 are missing. Hence we use the some daily averages plotted in Figure 8 for the period 1 May 1875 through 2 February 2011.

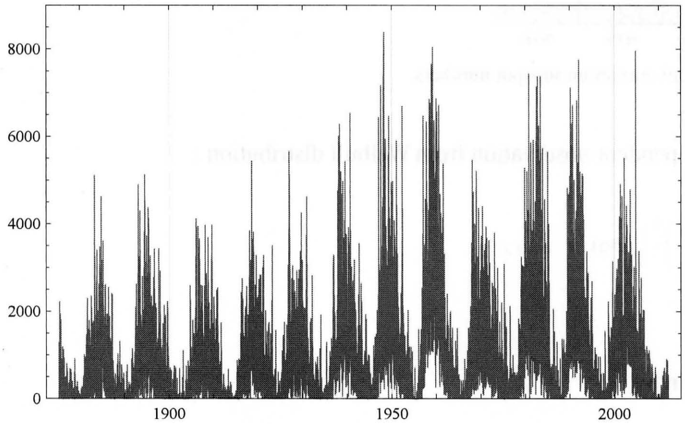

Figure 8. The series of daily sunspot numbers from *1 May 1875 through 2 Feb. 2011.*

Similar to the annual sunspot numbers, the level of the series seems to oscillate with an approximate period of 11. But the series of daily sunspot numbers are fluctuating widely and sharply, so it looks very difficult to get an appropriate model for the prediction of the series .

We here focus on annual maximum sunspot numbers, which is the block maximum (extreme value) data obtained from the daily sunspot numbers. The series is plotted in Figure 9 from 1875 to 2010. It also seems to oscillate with the period of 11.

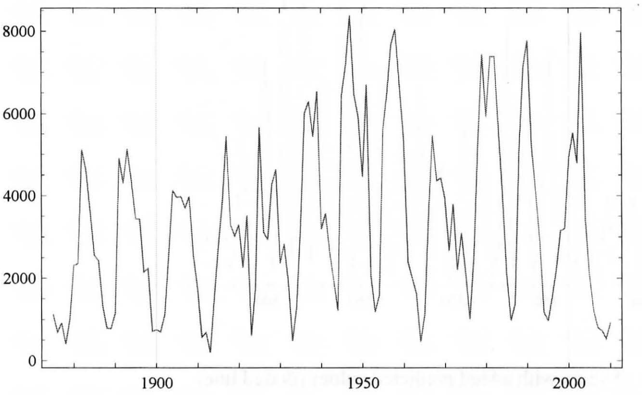

Figure 9. Annual maximum sunspot numbers from *1875 to 2010.*

We first consider the estimation of a probability density function to the series.

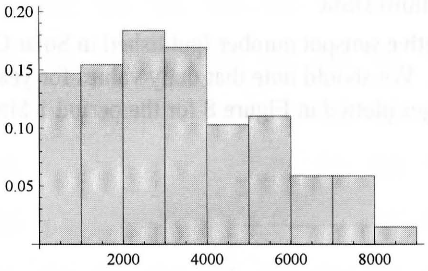

Figure 10. Histogram of annual maximum sunspot numbers.

Now we model the data as independent observation from Weibull distribution ;

$$F\ (x) = 1 - \mathrm{Exp}\left[-\left(\frac{x-\mu}{\beta}\right)^{\alpha}\right] \ for\, x \geq \mu,$$

$$= 0\ for\, x < \mu\ ;$$

and the density function is given by

$$f\ (x) = \frac{\alpha}{\beta}\ Exp\left[-\left(\frac{x-b}{\beta}\right)^{\alpha}\right]\left(\frac{x-b}{\beta}\right)^{-1+\alpha} \ for\, x \geq \mu,$$

$$= 0\ for\, x < \mu\ ;$$

where the parameters are shape $\alpha > 0$, scale $\beta > 0$ and location μ. (see, for example, [2], [4], [6]). (Note that this distribution is the Weibull family for minima.)

Maximization of the Weibull distribution log-likelihood for these data leads to the estimates of the parameters $\{\alpha, \beta, \mu\} = \{1.497, 3534.42, 148.1\}$. The density function is plotted in Figure 11.

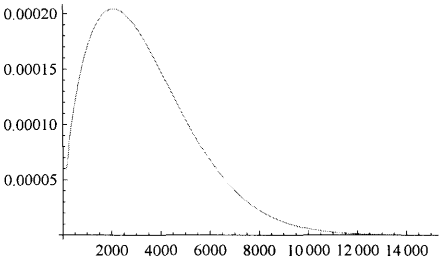

Figure 11. Density plot of the Weibull distribution with $\{\alpha, \beta, \mu\} = \{1.497, 3534.42, 148.1\}$.

Also the estimate of the density is shown as a solid curve in Figure 12 together with a curve which is a smoothed histogram of the series.

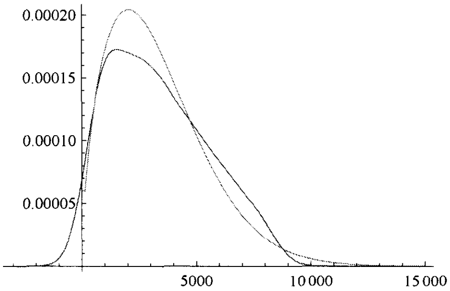

Figure 12. Density plots of the Weibull distribution and the smoothed histogram of the series.

The corresponding density function estimate seems to be almost consistent with the histogram of the data. It is seen that the Weibull distribution model is adequate for the data. This is confirmed by the standard diagnostic graphical check of the quantile plot shown in Figure 13.

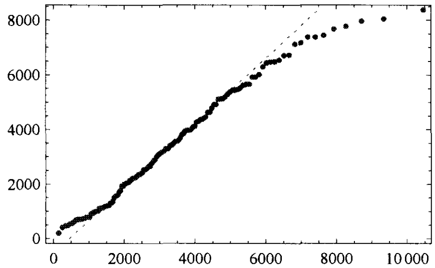

Figure 13. Quantile plot for Weibull distribution function fitted to annual maximum sunspot numbers.

There is no evidence of a linear trend in Figure 9, and then we fit data to a stationary time series model. Its sample correlation and sample partial correlation functions are plotted in Figure 14a-b. The slowly decaying behavior seems to imply that the series may be non stationary.

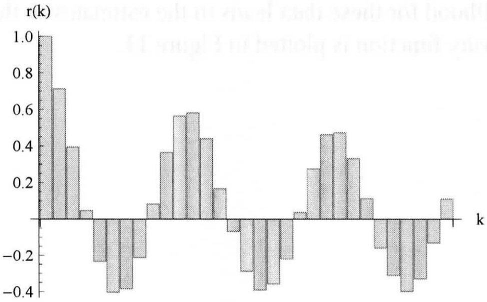

Figure 14a. The correlogram of the annual maximum sunspot numbers.

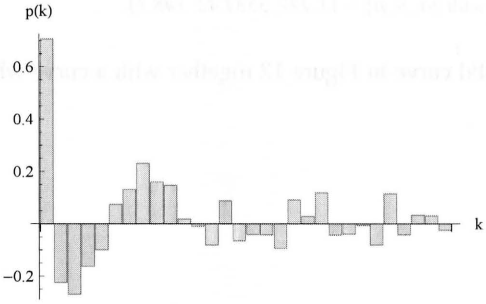

Figure 14b. The sample partial correlation function of the series.

We can model the series as observation from the AR model, and the smallest AIC model is AR(10). It is seen that the maximum likelihood estimates of the AR(10) parameters are given by

$$\{a_1,\ a_2,\ ....,\ a_{10}\} = \{0.644, -0.048, -0.059, -0.019, -0.078, -0.040, 0.016, -0.021, 0.078, 0.253\},$$

$$\sigma^2 = 124.035\,.$$

The parametric estimate of the spectral density based on the AR(10) model is shown as a solid curve together with a sample spectrum in Figure 15 .

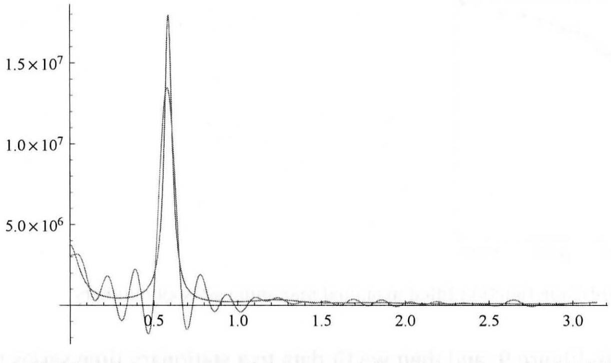

Figure 15. Sample spectrum and the spectral density based on the AR(10) model.

To test the adequacy of the model we calculate the residuals of the fitted AR(10) model and the correlogram of the residuals. The correlogram of the residuals shown in Figure 18 suggests that the residuals behave like white noise.

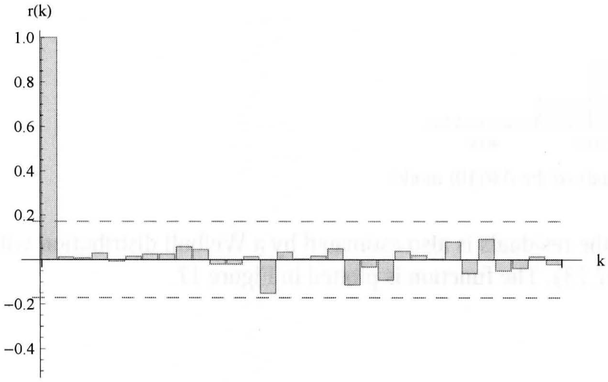

Figure 18. The Correlogram of the residuals and the 95% bounds.

In order to detect volatility (conditional heteroskedastic) we look at the correlogram of the squared residuals. From Figure 18 there is no evidence of serial correlation in the squared values since the correlation function falls within the 95% bounds. The AR(10) model also passes the portmanteau test. Hence the AR(10) model is apropriate for the series.

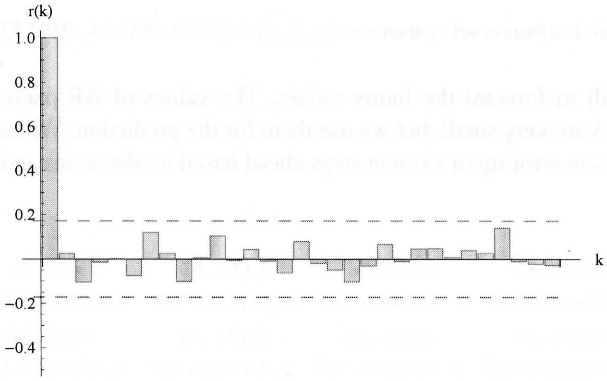

Figure 19. The Correlogram of the squared residuals and the 95% bounds.

Here we consider the estimation of a probability density function of the residuals. A histogram of the residuals is shown in Figure 16.

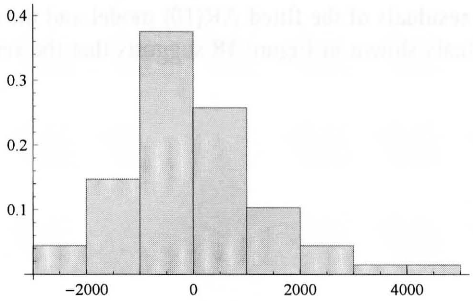

Figure 16. Histogram of the residuals of the AR(10) model

It is seen that the density function of the residuals is also estimated by a Weibull distribution with parameters $\{\alpha, \beta, \mu\} = \{2.036, 2815.44, \text{-}2507.73\}$. The function is plotted in Figure 17.

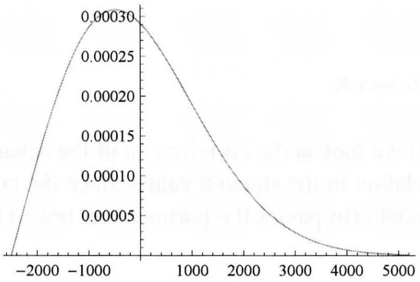

Figure 17. Density plot of the Weibull distribution with parameters $\{\alpha, \beta, \mu\} = \{2.036, 2815.44, \text{-}2507.73\}$

Then we use the fitted model AR(10) to forecast the future values. The values of AR-parameters a_2, ..., a_9 of the AR(10) model are very small, but we use them for the prediction. We calculate the best linear prediction and its mean square error up to 12 time steps ahead based on the series and the model AR(10) :

$$\begin{pmatrix} 3149.57 & 4423.04 & 5399.31 & 4830.88 & 3974.66 & 2904.81 & 1983.19 & 1299.97 & 1132.37 & 1526.56 & 2571.02 & 3657.36 \\ 1.63858\times10^{6} & 2.27045\times10^{6} & 2.53448\times10^{6} & 2.58448\times10^{6} & 2.58854\times10^{6} & 2.60899\times10^{6} & 2.65741\times10^{6} & 2.71624\times10^{6} & 2.72596\times10^{6} & 2.73429\times10^{6} & 2.91799\times10^{6} & 3.17779\times10^{6} \end{pmatrix}$$

We plot the data along with the next 12 predicted values in Figure 20. From this result it is seen that the next peak of the annual maximum sunspot data is also in 2013 with a sunspot number 5399, and the time for the peak is exactly consistent with the results for the annual sunspot numbers discussed in Section 2.

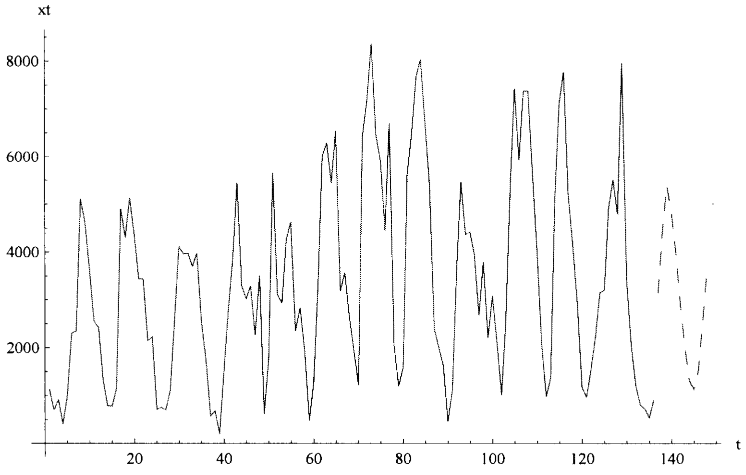

Figure 20. Time plot for last 12 years with added predicted values (dotted line)

4. Monte Carlo Simulation Study

We carried out a Monte Carlo simulation study to see a simulated realization of the fitted model AR(10) for the original block maximum sunspot data given in the previous section. Figure 21 below shows a random samples of size n=138 from the Weibull distribution with parameters $\{\alpha, \beta, \mu\}$ = {2.036, 2815.44, -2507.73}. We generate a time series of length n=138 according to the AR(10) model, where $\{a_1, a_2, \ldots, a_{10}\}$ = {0.644, -0.048, -0.059, -0.019, -0.078, -0.040, 0.016, -0.021, 0.078, 0.253} with the Weibull random sample as the noise series. Figure 22 shows the time plot of the data from the AR(10) model, and a histogram of the series is shown in Figure 23.

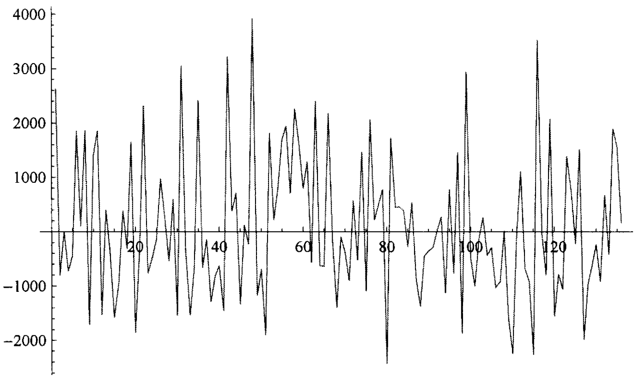

Figure 21. A random noise series from Weibull distribution with parameters $\{\alpha, \beta, \mu\}$ = {2.036, 2815.44, -2507.73}.

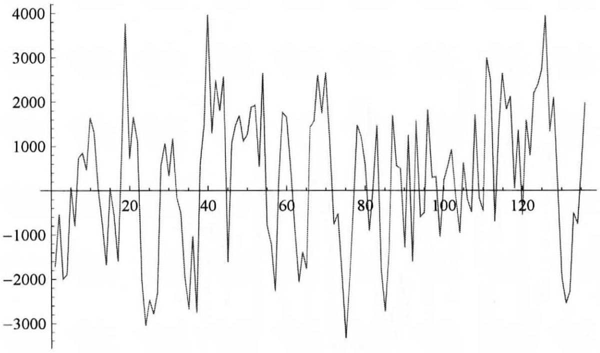

Figure 22. A generated AR(10) series based on the Weibull random noise.

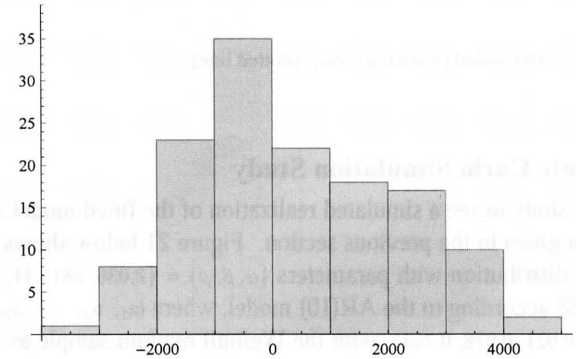

Figure 23. Histogram of the AR(10) series based on the Weibull random noise.

Conclusions

We have considered the prediction of Solar cycles from the sunspot numbers (yearly averages) based on an AR model. Also we focused on a time series of annual maximum (extreme values) sunspot numbers, and considered a fitting of an extreme value distribution function to the series. We adopt the AR(9) model with a white noise residuals for the series of annual sunspot numbers. It is seen that the next peak of the solar cycle will be in 2013 with a less sunspot number of 81. This result is almost consistent with the previous report of the prediction given by NOAA (NASA).

On the series of annual maximum (extreme values) sunspot numbers, we adopt the AR(10) model. From the best linear prediction up to 12 time steps ahead based on the series and the model AR(10), it is seen that the next peak will be also in 2013 with a sunspot number 5399, and also the peak is exactly consistent with the previous results for the annual sunspot numbers. The density functions of the series and the residuals from the fitted AR(10) model are estimated by a Weibull family for minima.

Appendix

This paragraph is excerpt from Coles ([2], p.155).

We consider the classical extreme value theory and models which focuses on the statistical behavior of

$$M_n = \max\{X_1, ..., X_n\},$$

where $X_1, \ldots, X_n$, is a sequence of independent random variables having a common distribution function F. When n is the number of observations in a year, M_n corresponds to the annual maximum.
If we consider a linear renormalization of the variable M_n for appropriate sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$M_n^* = \frac{M_n - b_n}{a_n},$$

then it is known that the location and scale of M_n^* are stabilized as n increases. The limit distributions for M_n^* are given by the following three classes of distribution termed the extreme value distributions, Gumbel, Fréchet and Weibull families respectively.

Theorem A.
If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr((M_n - b_n) / a_n < z) \rightarrow G(z) \text{ as } n \rightarrow \infty,$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\text{I}: \quad G(x) = Exp\left\{-Exp\left[-\left(\frac{x-b}{a}\right)\right]\right\} \text{ for } -\infty < x < \infty;$$

$$\text{II}: \quad G(x) = Exp\left[-\left(\frac{x-b}{a}\right)^{-\alpha}\right] \text{ for } x \geq b, = 0 \text{ for } x < b;$$

$$\text{III}: \quad G(x) = Exp\left\{-\left[-\left(\frac{x-b}{a}\right)\right]^{\alpha}\right\} \text{ for } x \leq b, = 1 \text{ for } x > b,$$

for parameters a>0, b and, in the case of families II and III, $\alpha > 0$.

We should note that these three families can be combined into a single family, the generalized extreme value (GEV) family of distributions with the form

$$G(x) = Exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

defined on the set $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

The GEV family is useful for modeling the distribution of maxima of long sequences such that the daily averages of sunspot number. The approach for modeling the extremes of a series of independent observations is that the data are blocked into sequences of observations of length n, for some large value of n, generating a series of block maxima, $M_{n,1}, \ldots, M_{n,m}$, to which the GEV distribution can be fitted. When we consider the estimates of extreme quantiles of the annual maximum distribution such that $G(z_p) = 1 - p$ for $p > 0$, z_p is the return level associated with the return period 1/p. The level z_p is expected to be exceeded on the average once every 1/p years.

References

[1] Blockwell,P.J., Davis,R.A. (2002), *Introduction to Time Series and Forecasting*, Springer Verlag, New York.

[2] Coles, S.G. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, New York.

[3] Cowpertwait.P.S.P., Metcalfe.A.V. (2009), *Introductory Time Series With R*, Springer Verlag, New York.

[4] Embrechts,P., Kluppelberg,G., & Mikosch,T. (1998), *Modelling Extremal Events for Insurance and Finance*, Springer Verlag, New York.

[5] He,Y. (1995), *Time Series Pack for Mathematica*, Wolfram Research.

[6] Leadbetter,M.R., Lindgren,C., & Rootzen,H. (1983), *Extremes and Related Properties of Random Sequences and Series*, Springer Verlag, New York.

[7] Thomas, J.H., Weiss, N.O. (2008), *Sunspots and Starspots*, Cambridge University Pree.