Sunspot Analysis

Emre Yurtbay

4/14/2019

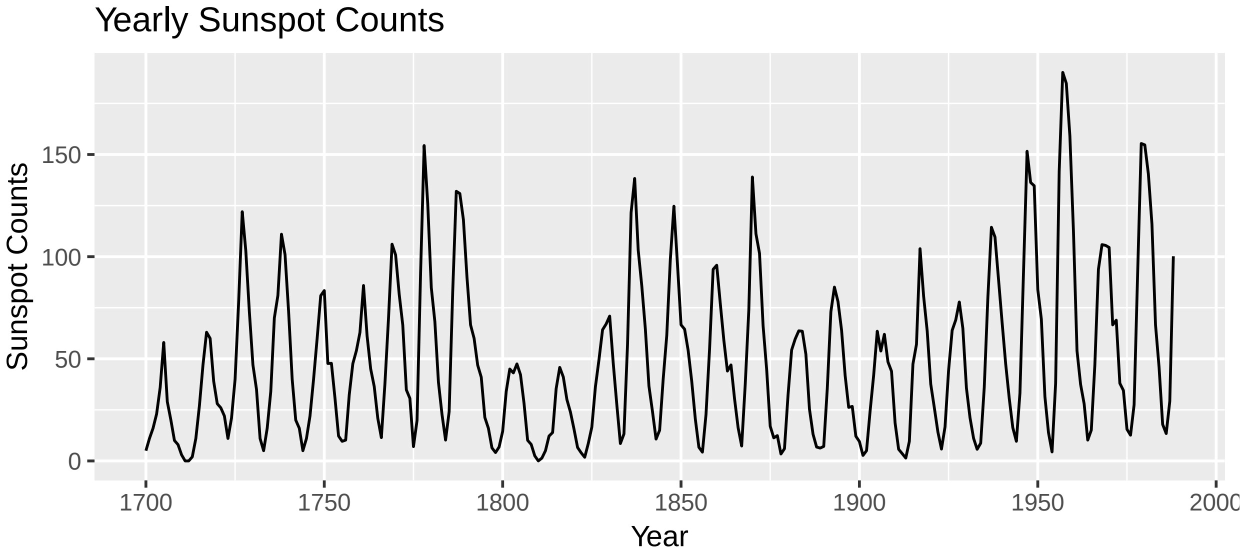
Background and Key Question

Sunspots are temporary phenomena on the sun's photosphere that appear as spots darker than the surrounding areas. Due to their distinct coloring, they are often quite easy to find with simple tools like telescopes, meaning that reliable data of sunspot counts have existed since the 1700s. Intrestingly, sunspot occurrence is linked to solar cycles, which are about 11 earth years in length. The count of sunspots on the sun's surface for any given year is known as "Wolf's Number," and knowing and predicticing this quantity is quite important to both astronomers and meteorologists. Due to its link to other kinds of solar activity, sunspot occurrence can be used to help predict space weather as well as the state of the ionosphere, and hence the conditions of short-wave radio propagation or satellite communications. Solar activity (and the solar cycle) is also linked with global warming. For example, fluctuations in the sunspot counts are associated with the Little Ice Age, a cold period that hit Europe in the 17th century. Due to these reasons, predicting sunspot intensity has some real implications and can be valuable to astronomers and meteorologists.

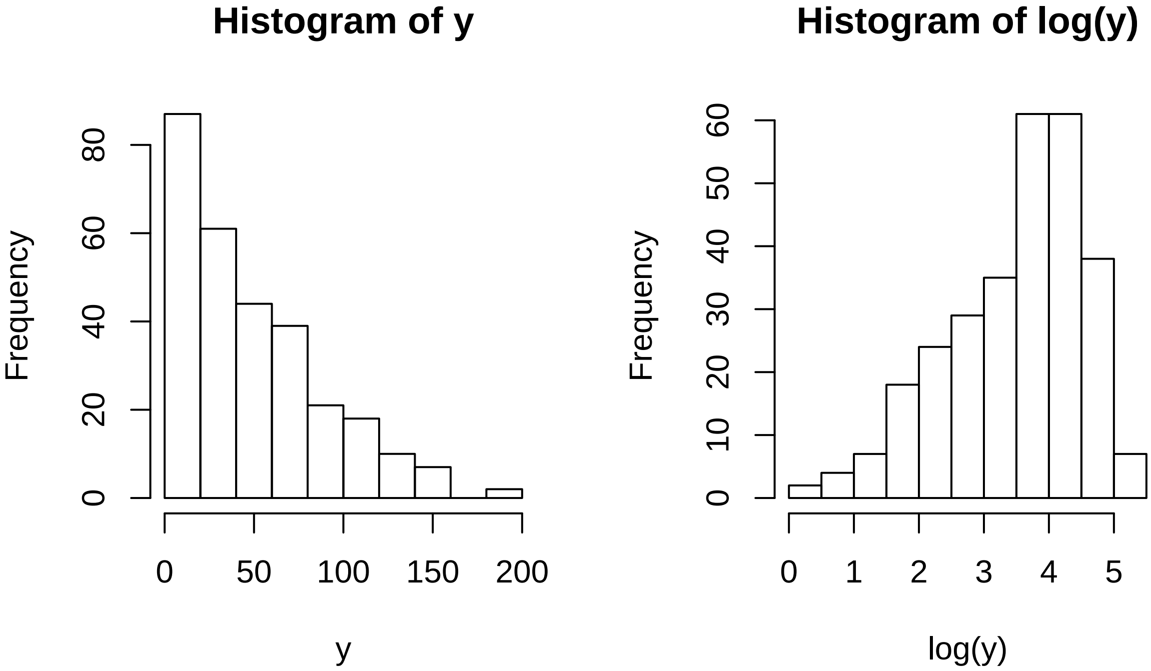
Some questions we would like answer include the following: Can we create a forecast that can help us accurately predict sunspot cycles? Can we build a Hidden Markov model to label years as having low, moderate, and high sunspot activity? Can we find the transition probability from one of these states to another? Can we confirm the proposed length of sunspot cycles?

Data: Definition, Aquisition and Exploration

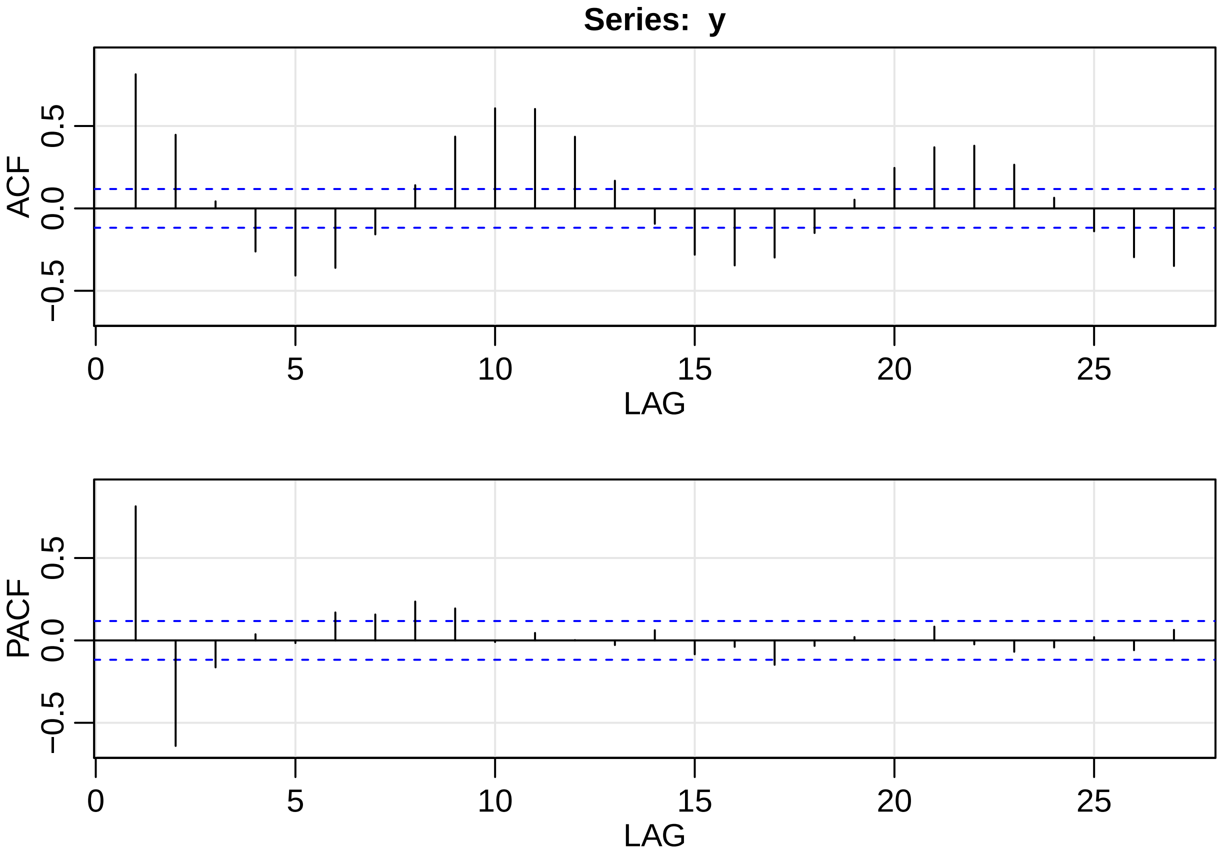
Yearly sunspot data from 1700 to 1988 is obtained from the SIDC, or the Solar Influences Data Analysis Center. The counts of observed sunspots is recorded and updated regularly on the SIDC website as the data become available. For more information, visit http://www.sidc.be/silso/datafiles. A csv of my dataset will be included in the submission. For my analysis, I define the data as a *ts* object. The time series plot of the sunspot counts below shows a highly correlated series with very regular oscillations. Sunspot cycles are supposed to last 11 years, and the data roughly confirm this. Between 1700 and 1750, there were roughly 4.5 cycles, which roughly corresponds to an 11 year cycle on average.



The distribution of sunspot counts is rather right skewed, meaning there are few years with very high intensity. If we take the log of y, we fix the skewness slightly, but we do not solve the problem.



Below, we see the ACF and PACF plots of the sunspot data. The PACF cuts off after lag 2, which suggests maybe an AR(2) model may fit the data. The ACF seems to tail off more slowly than expected however, which may affect the modeling capability of the AR(2).



Modeling Wolf's Number using a Hidden Markov Model

An Hidden Markov Model is a statistical framework in which the system being modeled (sunspot counts or Wolf's Number) is assumed to be a Markov process with unobserved (i.e. hidden) states. HMMs are used in all sorts of fields, including part of speech tagging, error decoding, bioinformatics, and speech recognition. In the time series context, HMMs are rather useful in dealing with real world data, since they can handle complex interactions and non-stationarity.

Sunspots are a good phenomenon to model with a Markov chain because the present is heavily dependent on the past. We can fit Markov models of many different orders as well. An order 1 markov chain looks 1 year to the past to predict the present, and an order 2 markov chain looks 2 years into the past to predict the present, and so on. Here, we want to fit a HMM to label years as having low, moderate, and high sunspot activity. In the spirit of model parsimony, an order 1 markov model seems appropriate.

After fitting our Hidden Markov Model, we want to find the transition probabilities matrix, which can help us create one step ahead forecasts and describe the dynamics of our system. Additionally, we would like to visualize the model as well.

Initial State Probabilities Model

pr1	pr2	pr3
0	0	1

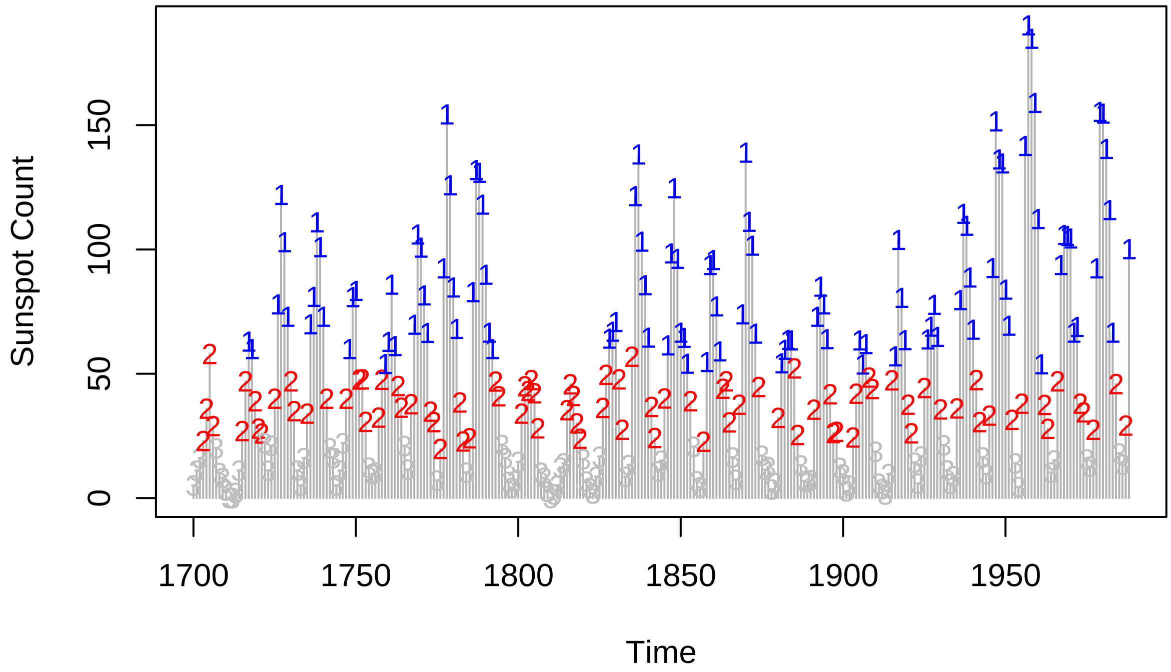
The initial state probabilities do not have too much importance, since we have only given the model a single set of training data, and this set starts at a low intensity. In the conventional setup for a Hidden Markov Model, one often has multiple training examples and therefore the initial state probabilities are very important in prediction. This does not quite extend to the time series setting however.

Transition Matrix

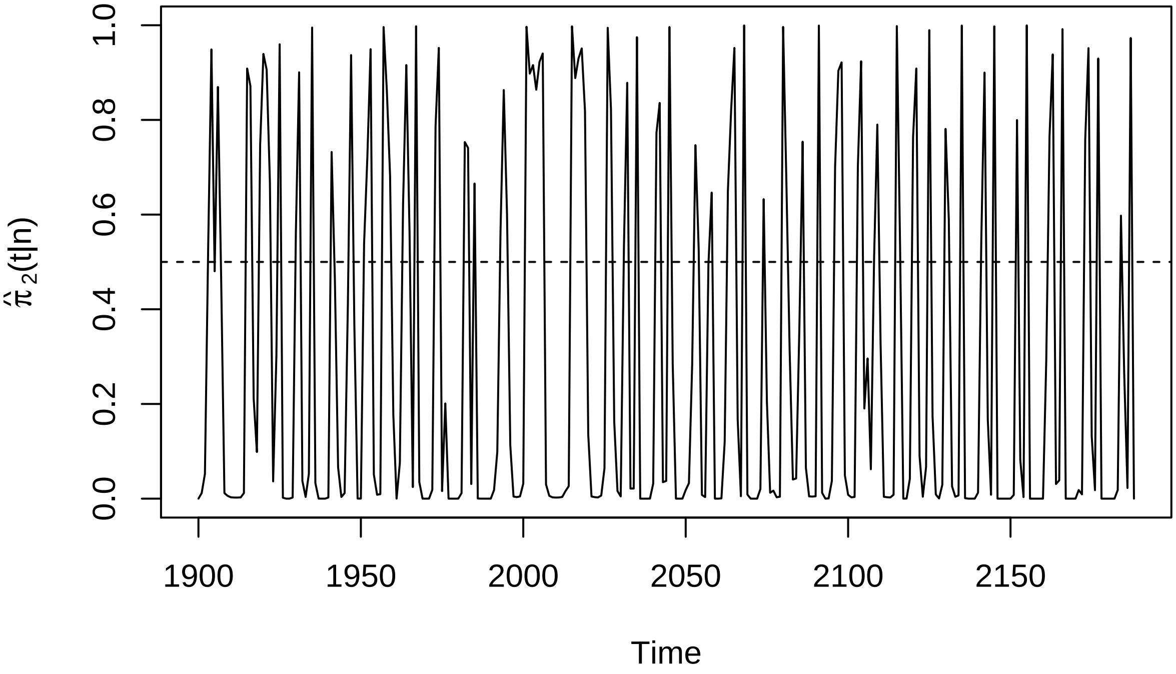
	toS1	toS2	toS3
fromS1	0.799	0.221	0
fromS2	0.283	0.425	0.292
fromS3	0	0.303	0.697

State 1 represents a high intensity year, state 2 represents a medium intenisty year, and state 3 represents low intensity year. From the transition matrix, we see that it is highly probable that next year's state is identical to this year's state. That is, if this year is a high intensity sunspot year, next year is likely to also be one of high intensity. This also holds true for moderate and low intenisty years. We also want to examine the probability of moving between states. Since state 2 is moderate intensity, it is possible to travel from moderate to high or moderate to low, depending on where the year falls in relation to the solar cycle. Because of this, these the probabilities are all above zero. However, it is impossible to travel from a low intensity state to a high intensity state (as far as we know), hence the 0 transition probability from state 1 to state 3.

The transition probability matrix is very helpful in making forecasts for upcoming years. Given this year's information, we can make forecasts for the coming year by reading off the transition probability matrix. If we are in a high intensity year, our model says that there is a 80% chance next year will also be a high intensity year and about a 20% chance next year will be a moderate intensity year.



The above plot is a graphical representation of the transition probability matrix from above. As we can see, the years labeled 1 are years of intense sunspot activity, the years labeled 3 are years of low sunspot activity, and the years labeled 2 are years of moderate sunspot activity.



Above, we see the probability that any given year will be in state 2. The peaks in this graph match years of medium intensity in the plot above, as is expected.

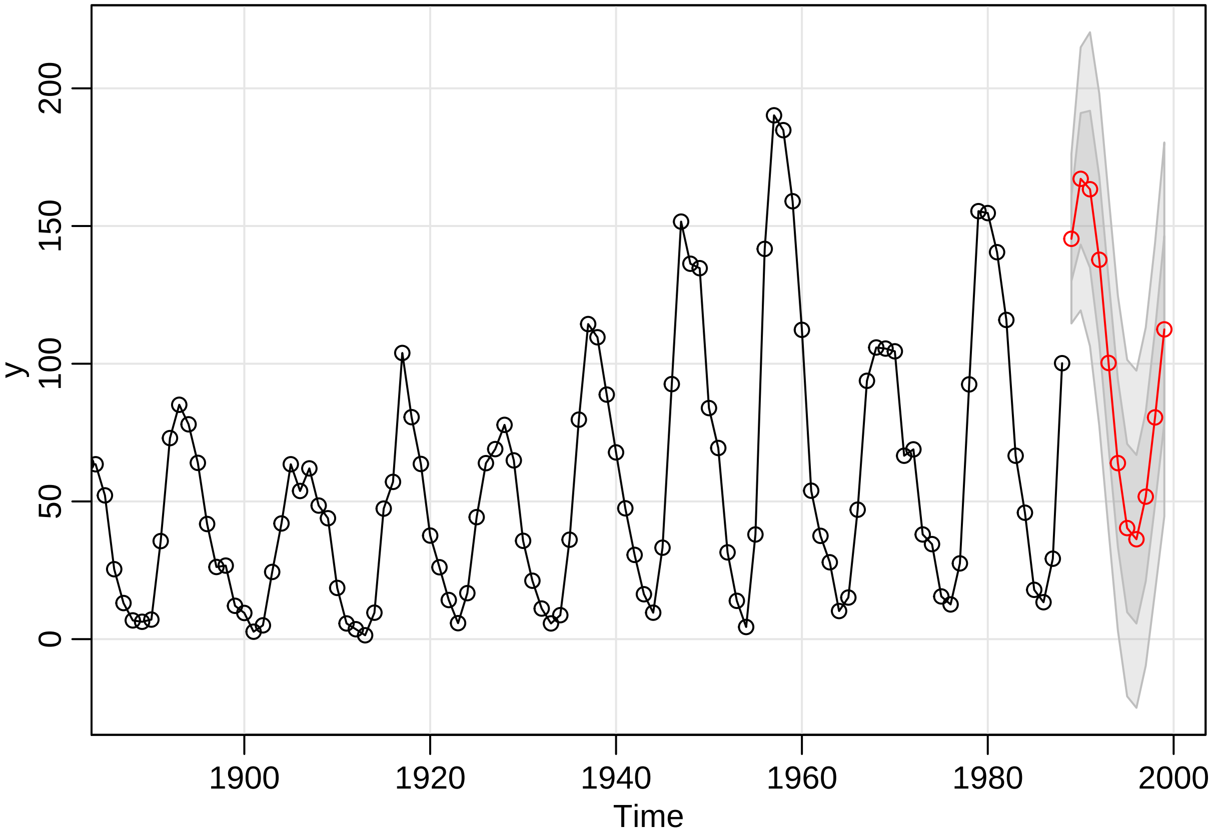
One thing to note is that a Hidden Markov model of a higher order may be more appropriate in this context. Looking back further into the past may give you more information as to which state the next year will be in. Say you are in state 2 - moderate intensity. There is a non-zero probability that you might transition to state 3 or you might transition to state 1. If you were just in state 1 last year, and you are in state 2 this year, the probability you will be in state 3 next year is much higher than the probability you will be back in state

1, since Wolf's number must go down before it goes up again. This gives us much more information while forecasting, but the model is also rather more complex, and vulnerable to the perils of overfitting. Still, an order 1 model has plenty of descriptive power as well as forecasting potential.

Predicting Long Term Wolf's Numbers using an ARIMA Model

The Hidden Markov Model we built earlier is very helpful in building 1 step ahead forecasts by looking at the present, but sometimes we want to look farther into the future. By using the *auto.arima* function, we can find the optimal ARIMA model based on order selection criteria like BIC. The *auto.arima* function returned an ARIMA(3, 1, 2) as the optimal model, Below, we see the estimates for the coefficeints of the ARIMA model and their standard errors. Two assumptions made that are critical in ARIMA modeling are that model parameters are constant over time and the error process is homoscedastic (constant) over time, but the data seem to confirm that these are reasonable. Using *sarima.for*, I created a 11 step ahead forecast and plotted the predictions and error bands. As we can see, the ARIMA forecast follows the 11 year cycle we are expecting. As the forecast goes out farther into the future, the predictions from *sarima.for* will converge toward the mean of the series.

term	estimate	std.error
ar1	1.8167656	0.0880911
ar2	-1.2640546	0.1372992
ar3	0.1930356	0.0814785
ma1	-1.6259166	0.0614456
ma2	0.7405454	0.0605047



Below, we show the exact values of the forecast created by *sarima.for* for the next 11 years.

index	value
1989	145.35737
1990	167.13361
1991	163.35829
1992	137.75556
1993	100.28694
1994	63.91728
1995	40.32576
1996	36.26565
1997	51.74714
1998	80.50906
1999	112.46905

Conclusions

The temporal dynamics of Wolf's Number seems to be well described by a HMM of order 1, with transition probabilities as described in the matrix given. Describing sunspot cycles in three states may be a bit simplistic, but it is nonetheless helpful in understanding how the solar cycle works. By knowing Wolf's number in the present, both meteorologists and astronomers can use the transition matrix of the Hidden Markov Model to predict next year's sunspot intensity, which can affect climate forecasts here on earth. Furthermore, if an astronomer wanted produce long term forecasts, the ARIMA(3,1,2) model would be very useful. To dive deeper into the data, I might want to explore other kinds of models, like Stochastic Volatility Models, as well as Hidden Markov Models of higher orders.