Lecture Notes in Quantitative Biology

Autocorrelation. Autoregression

Chapter 20.3 (from 25 November 1993)

ReCap
Autocorrelation
Time series. Spatial series
Separation vs frequency
Statistics of association
Application: independent residuals?
Models

Handouts. Serial.ref
Codacf.out

Autocorrelation.

Introduction.

Autocorrelated data often encountered in environmental biology
Example: Counts of plants in adjacent quadrats.
If count high in a quadrat, count in neighbor likely to be high.
If count low, count in neighbor likely to be low.
Autocorrelated data often in laboratory work.
Example: Behaviour of pigeon on successive trials.

Wrap-up.

Autocorrelated data has become increasingly common in biology, due largely to automated recording of data, taken at high temporal resolution (measurements close together in time) or at high spatial resolution (measurements taken close together in space).

Autocorrelation can be quantified either relative to separation or frequency of measurement. Models of autorcorrelation can also be developed.

Autocorrelation

The basic idea behind autocorrelation is that if we take a series of measurements in time or space, then we often expect an observation to be related in some way to the immediately preceding observation. For example, if we take a series of measurements of photosynthetically active radiation (PAR) reaching the forest floor, we will find that the measurements are autocorrelated. If a lot of light is reaching the forest floor during the present hour, then on average we expect a lot of light to reach the forest floor during the next succeeding hour. The factors responsible for this are many: day-night variation in light levels, variation due to clouds brought in by weather systems, and seasonal variation in light levels.

Similarly, if we measure the amount of chlorophyll within adjacent quadrats up a mountain, we will expect the amount of chlorophyll in any one quadrat to be related to its neighbors, rather than being completely independent. Again, many factors are responsible for this: reduction of plant biomass in areas of steep slope, reduction in biomass due to landslides, reduction in plant biomass at high altitudes.

Autocorrelated data is a frequent problem in biology.
Measurements are often taken in a temporal sequence,
with one measurement being correlated with the prior measurement.
Similarly, measurements in space can be correlated, especially
over short distances.

Autocorrelated data is becoming more frequent. This is due to automated collection of data. For example, hourly measurements of PAR were not possible over extended periods of time until the development of digital recording devices to automatically store PAR readings. Chlorophyll amount per hectare in a series of adjacent hectares up a mountain was not possible until the advent of satellite imagery.

The topic is typically not covered by introductory texts, despite its frequency in biological data sets. There are many specialized books. The handout (serial.ref) lists several texts, ranging from the highly mathematical to more accessible.

The analysis of autocorrelated data can be elaborate, as is evident from some of these books. The purpose of this lecture is to

introduce the concept,
show how to recognize autocorrelation in its various forms
show how to identify and remove the effects of autocorrelation
from an analysis

Time series. Spatial series

As always, we are working with defined quantities. Here are two examples.

Catch of fish (cod) over a series of years.
Y_t = total landings in year t
Units are metric tonnes per year from a specific area off Newfoundland

For a number of reasons we expect Y_{t+1} to be related to Y_t but not vice versa. Catches depend on investment in boats and equipment, which changes at the scale of years. If catch is small this year due to small investment in equipment, then it will not suddenly quadruple the next year. Similarly, if there are large numbers of people catching fish, this will not normally change very much from one year to the next.

The handout shows a graph of inshore catches over a period of 30 years. The graph shows clear trends. It is evident that the catch in any one year is related to the catch in the previous year, on average. The difference in catch between successive years is, on average, less than the difference over many years.

Here is spatial example.

N_x = number of Gem clams, *Gemma gemma*, in 18 core samples collected at 2 m intervals along a straight line across an intertidal sand flat.

Units are number of clams per 10 cm diameter core, taken to depth of 10 cm into the substrate.

Graph shows trends in density along the transect. As with the temporal series, the difference between adjacent samples is on average less than the difference between samples at greater separations. Another way of looking at this is that the density is on average higher at one end of the transect than at the other.

This can be extended to two spatial dimensions N_{xy} or even to three N_{xyz}.

820
1178
876
887
675
651
829
851
867
810
1100
822
1165
957
1065
1101
1073
1130

Separation vs frequency

There are a number of statistics for quantifying the degree of association in serial data. These statistics are expressed either in term of separation, or in terms of frequency.

Separation. Also called distance specification.

What is the average change in catch, from one year to the next ?
To compute this we compute $Y_t - Y_{t-1}$ through the series, sum these values, and take the average. For the fish catch data, the average difference at lag 1 is:

$$\bar{D}(1) = n^{-1} \Sigma Y_t - Y_{t-1} = -1985$$

Similarly, what is the average change in catch at separations of two, three and four years ?

$$\bar{D}(2) = n^{-1} \Sigma Y_t - Y_{t-2} = -4861$$

$$\bar{D}(3) = n^{-1} \Sigma Y_t - Y_{t-3} = -6781$$

$$\bar{D}(4) = n^{-1} \Sigma Y_t - Y_{t-4} = -9297$$

etc

This shows that on average, catches have declined more than they have increase, and that this decline grows larger with increasing time scale (lag).

Those who have taken calculus will recognize the operations here as similar to taking a differential. If we have a function that describes speed at any point in time, then we can use differential calculus to obtain the change in speed (the deceleration) from one point in time to another.

Separation vs frequency (continued)

Frequency. Also called interval specification.

The same information that was expressed as a function of separation can also be expressed as a function of frequency of measurement. For the fish catches, we again start with the average differences from year to year.

$$\bar{D}(30/30) = n^{-1} \Sigma Y_t - Y_{t-1} = -1985 \text{ tonnes/year}$$

The frequency of measurement is thirty times for thirty years, or 30/30.

Next we group the catches into two year averages, and then compute the average differences between catches at time scales of two years. The frequency of measurement is 15 times over 30 years or 15/30.

$$\bar{D}(15/30) = n^{-1} \Sigma Y_{2t} - Y_{2t-1} = -4861 \text{ tonnes/2 year}$$

Then we group the catches into three year averages, computing the average difference at this time scale. The frequency is 10/30

$$\bar{D}(10/30) = n^{-1} \Sigma Y_{3t} - Y_{3t-1} = -6781 \text{ tonnes/3 year}$$

When the catches are grouped at 5 year intervals, the frequency is 6/30.

$$\bar{D}(6/30) = n^{-1} \Sigma Y_{3t} - Y_{3t-1} = -11542 \text{ tonnes/5 year}$$

This is easily continued for even multiples of 30: 6/30 5/30 3/30 2/30.
We can also make estimates at uneven multiples, though this will be harder, and we won't attempt it here.

This analysis of change in catch as a function of increasing time scale (every year, ever two year period, every three year period, etc) appears to be different than our analysis of change in catch with increasing separation in time. In fact it is exactly the same information. It has been expressed in different form. But the information about pattern is exactly the same. In fact we get exactly the same numbers when the lag (1,2,3) matches one of the frequency groupings (30/30 15/30 and 10/30). It takes a while to get used to the idea that analysis as a function of separation (lag) is equivalent to analysis as a function of frequency. Most people who work with serial association end up thinking either in terms of separation, or in terms of frequency. They also end up being unable to recognize that the other form of analysis as equivalent. If they think in terms of separation, they are unable to recognize an equivalent analysis in terms of frequency.

Separation vs frequency (continued)

The key point is that separation (lag) and frequency (interval) specifications are different ways of looking at same information. They are not completely different analyses.

+theoretical specification

Next, the spatial example. What is the average change in density from one sample to the next ?

$$\Delta N(1) = n^{-1} \Sigma N_x - X_{x-1} = 18.235 \text{ clams}$$

If we divide the difference in density by the separation we obtain the spatial gradient. This measures how rapidly the density changes per unit distance laterally.

$$\text{gradN}(1) = n^{-1} \Sigma N_x - X_{x-1} = 18.235 \text{ clams} / 2 \text{ m}$$

Both the average difference and the average gradient can be calculated at increasingly large separations or lags.

$$\Delta N(2) = n^{-1} \Sigma N_x - X_{x-2} = 12.812 \text{ clams}$$

$$\Delta N(3) = n^{-1} \Sigma N_x - X_{x-3} = 28.667 \text{ clams}$$

$$\Delta N(4) = n^{-1} \Sigma N_x - X_{x-4} = 43.429 \text{ clams}$$

$$\Delta N(5) = n^{-1} \Sigma N_x - X_{x-5} = 68.462 \text{ clams}$$

etc

$$\text{gradN}(2) = n^{-1} 2^{-1} \Sigma N_x - X_{x-2} = 12.812 / 4 \text{ m} = 3.203 \text{ m}^{-1}$$

$$\text{gradN}(3) = n^{-1} 3^{-1} \Sigma N_x - X_{x-3} = 28.667 / 6 \text{ m} = 4.778 \text{ m}^{-1}$$

$$\text{gradN}(4) = n^{-1} 4_{-1} \Sigma N_x - X_{x-4} = 43.429 / 8 \text{ m} = 5.428 \text{ m}^{-1}$$

$$\text{gradN}(5) = n^{-1} 4_{-1} \Sigma N_x - X_{x-5} = 68.422 / 10 \text{ m} = 6.842 \text{ m}^{-1}$$

The gradient increases as separation increases, which indicates the presence of pattern.

Separation vs Frequency (continued)
The same information that was expressed as a function of separation can also be expressed as a function of frequency. For the clam densities, we again start with the average differences from on location to the next.

$$\Delta N(1/20) = n^{-1} \Sigma N_x - X_{x-1} = 18.234 \text{ clams}$$

The frequency of measurement is once every meter for 20 meters, or 1/20.

Next we group the density into averages over two meter blocks, and then compute the average differences between blocks at a spatial scale of two meters. The frequency of measurement is once every 2 meters over 20 years or 2/20.

$$\Delta N(2/20) = n^{-1} \Sigma N_{2x} - X_{2x-1} = 12.812 / 4 \text{ m}$$

Then we group the data into 8 m blocks. The frequency is 4/20

$$\Delta N(4/20) = n^{-1} \Sigma N_{4x} - X_{4x-1} = 43.429 / 8 \text{ m}$$

This again appears to differ from the analysis based on spatial separation but in fact it is the same information in different form.

Statistics of Association

Temporal and spatial association or pattern can be measured either in relation to separation or in relation to frequency. The correlation coefficient measures association as a function of separation. For the fish catch data we can calculate the correlation between each measurement and its nearest neighbor in time.

$$\text{Corr}(Y_t\ Y_{t-1}) = +0.816$$

Similarly, we can calculate the correlation between each measurment and its second nearest neighbor in time, third nearest neighbor, etc.

$$\text{Corr}(Y_t\ Y_{t-2}) = +0.636$$

$$\text{Corr}(Y_t\ Y_{t-3}) = +0.537$$

$$\text{Corr}(Y_t\ Y_{t-4}) = +0.401$$

This is relatively easy to compute in Minitab.
MTB> Acf 'inshore'

This command automatically calculates the correlation in the data at lags 1 on up to half the length of the data series. Looking at the handout (codacf.out) we see that the inshore catch data is strongly correlated with itself at lags of one year. It becomes somewhat less strongly associated at longer time lags. It becomes negatively associated at time lags on the order of 10 years. That is, high catches were preceded by low catches roughly 10 years before.

Similar information can be extracted by working with frequencies. This is accomplished by computing the variance in catch as a function of frequency of measurement. This will run from low frequency (2 cycles per 30 observations) to high frequency (15 cycles per 30 observations). This is equivalent to running from long lags (lag 15) to short lags (lag 2).

This equivalence is hard enough to see for means. It is even harder to see for variances. In fact for many years ecologists analyzed spatial data for pattern by computing variance as a function of frequency with no indication (or perhaps knowledge) that this was equivalent to working with lagged autocorrelations.

Application: independent residuals?

Another application of autocorrelation is testing whether residuals from an analysis of data meet the important assumption that these residuals be independent. As with the assumptions about normality, it is the residuals that must meet the test, not the data. We should or course examine the residuals but if we are dealing with autocorrelated data, then we have every reason to expect that the residuals will also be autocorrelated. This is easily tested with the Minitab ACF command.

Here is an example, returning to the data on fish catch by the inshore sector. The question is whether the inshore catch is related to the offshore catch. Do inshore catches go down when offshore catches go up ? If we test this by regression of inshore catch against offshore catch, we will need to check the residuals to see if they are autocorrelated. When we do this, we find that the residuals are strongly autocorrelated, hence not independent of one another. This is a problem because it means that each data point is less free to vary than we thought. It means that we do not really have 28 degrees of freedom for our regression analysis. One common remedy is to eliminate auotcorrelation within the response variable by taking differences. That is, we take the difference between year 1 and year 2, year 2 and 3, etc then use this series in our analysis, rather than the inshore catch itself. The handout shows how to do this using the Minitab command

MTB> differences

The handout shows that this substantially reduces the autocorrelation within the inshore catch data. When we test for association of inshore with offshore catch using the differenced data, we find that the residuals are now free also of substantial autocorrelation. The residuals are independent and so we can use 28 degrees of freedom to compute a p-value based on the differenced data. The p-value is 0.352 a good deal less than the p-value of 0.833 with autocorrelation present. It is still not significant, but this time the test is better one.

Application Pattern analysis.

Autocorrelation has a number of applications. One that we have touched on already is discovering pattern in data. In particular, the idea of autocorrelation allows us to describe pattern as a function of spatial or temporal scale. The inshore catch data varies primarily at time scales of decades. There is little variation or pattern at shorter time scales.

With spatial data we can also describe pattern as a function of scale. Patterns become evident at particular scales. For example spatial variation in density of pelagic fish may be very low at small separations, due to even spacing within schools. At slightly larger separations there may be substantial variation due to presence or absence of large schools. At still larger scales there may be additional variation due to differences in the size and frequency of schools at the centre and the edge of the species range. Autocorrelation is a useful way of describing this dependence of pattern on scale.

Models	Autoregressive	(regression type explanatory variable)
	Moving average	(ANOVA type explanatory variable)
	Autocorelation	(X unknown)

Example of autocorrelative model. $r = e^{-k*lag}$ correlation decays exponentially
Examples of autoregressive process regress series on itself at 1 or more lags
Examples of moving average process
regress series against average of preceeding values,
average over 2 or more values

Extra: Deciding on autoregressive versus moving average models.

Diagnosed by examining the autocorrelation and partial autocorrelation.

MTB> ACF c1.
This computes the autocorrelation at series of lags, then plots these as a function of lag.
MTB> PACF c1
This computes the partial autocorrelation coefficient at lag s, controlled for all lesser lags. The PACF is also plotted as a function of lag.

PACF tapering and ACF spike : autoregressive $Y_t = f(Y_{t-s})$
ACF tapering and PACF spike : moving average
Y_t = f(several previous t)
Variance as function of frequency (hierarchical ANOVA)
flat = no autocorrelation.
Rising indicates larger scale correlation.

Diagnosis in distance domain more suitable for
locally important processes,
not sensitive to large scale or long term effects.
Diagnosis in frequency domain more suitable for
larger scale temporal or spatial variation.

Extra:

The variance at measurement frequency f
Aggregate the units into increasingly larger groups
Chop the series into increasingly finer sections
Plot MS among versus group size.
Executed by setting up series of dummy variables in Minitab,
then carry out ANOVA

Spectral analysis. Better estimate of the same thing as above.
Plot spectral density as function of frequency (bottom) and period (On top). These are calculated in the frequency domain.
Cycles per measurement unit on bottom.
Period on top to facilitate reading.