

3 Autocorrelation

Autocorrelation refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called “*lagged correlation*” or “*serial correlation*”, which refers to the correlation between members of a series of numbers arranged in time. Positive autocorrelation might be considered a specific form of “*persistence*”, a tendency for a system to remain in the same state from one observation to the next. For example, the likelihood of tomorrow being rainy is greater if today is rainy than if today is dry. Geophysical time series are frequently autocorrelated because of inertia or carryover processes in the physical system. For example, the slowly evolving and moving low pressure systems in the atmosphere might impart persistence to daily rainfall. Or the slow drainage of groundwater reserves might impart correlation to successive annual flows of a river. Or stored photosynthates might impart correlation to successive annual values of tree-ring indices. Autocorrelation complicates the application of statistical tests by reducing the number of independent observations. Autocorrelation can also complicate the identification of significant covariance or correlation between time series (e.g., precipitation with a tree-ring series). Autocorrelation can be exploited for predictions: an autocorrelated time series is predictable, probabilistically, because future values depend on current and past values. Three tools for assessing the autocorrelation of a time series are (1) the time series plot, (2) the lagged scatterplot, and (3) the autocorrelation function.

3.1 Time series plot

Positively autocorrelated series are sometimes called *persistent* because positive departures from the mean tend to be followed by positive departures from the mean, and negative departures from the mean tend to be followed by negative departures (Figure 3.1). In contrast, negative autocorrelation is characterized by a tendency for positive departures to follow negative departures, and vice versa. Positive autocorrelation might show up in a time series plot as unusually long runs, or stretches, of several consecutive observations above or below the mean. Negative autocorrelation might show up as an unusually low incidence of such runs. Because the “departures” for computing autocorrelation are relative the mean, a horizontal line plotted at the sample mean is useful in evaluating autocorrelation with the time series plot.

Visual assessment of autocorrelation from the time series plot is subjective and depends considerably on experience. Statistical tests based on the observed number of runs above and below the mean are available (e.g., Draper and Smith 1981), though none are covered in this course. It is a good idea, however, to look at the time series plot as a first step in analysis of persistence. If nothing else, this inspection might show that the persistence is much more prevalent in some parts of the series than in others.

3.2 Lagged scatterplot

The simplest graphical summary of autocorrelation in a time series is the lagged scatterplot, which is a scatterplot of the time series against itself offset in time by one to several time steps (Figure 3.2). Let the time series of length N be x_i , $i = 1, \dots, N$. The lagged scatterplot for lag k is a scatterplot of the last $N - k$ observations against the first $N - k$ observations. For example, for lag-1, observations x_2, x_3, \dots, x_N are plotted against observations x_1, x_2, \dots, x_{N-1} .

A random scattering of points in the lagged scatterplot indicates a lack of autocorrelation. Such a series is also sometimes called “random”, meaning that the value at time t is independent

of the value at other times. Alignment from lower left to upper right in the lagged scatterplot indicates positive autocorrelation. Alignment from upper left to lower right indicates negative autocorrelation.

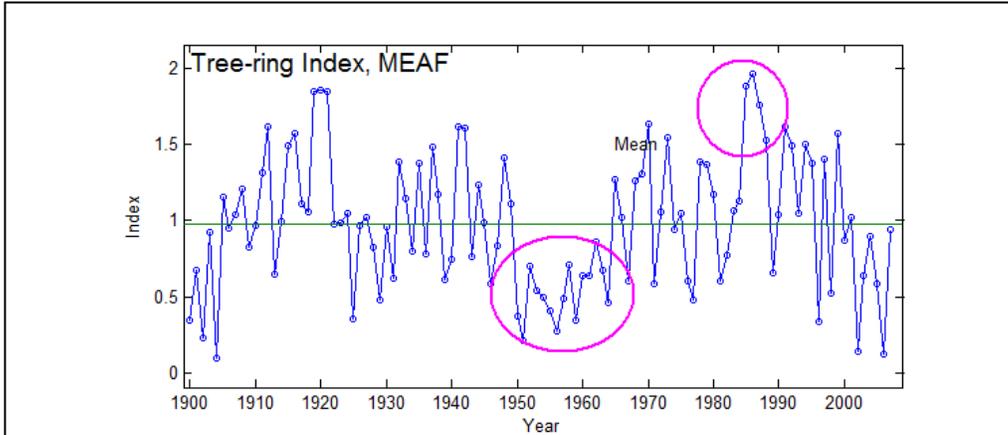


Figure 3.1. Time series plot illustrating signatures of persistence. Tendency for highs to follow highs or lows to follow lows (circled segments) characterize series with persistence, or positive autocorrelation.

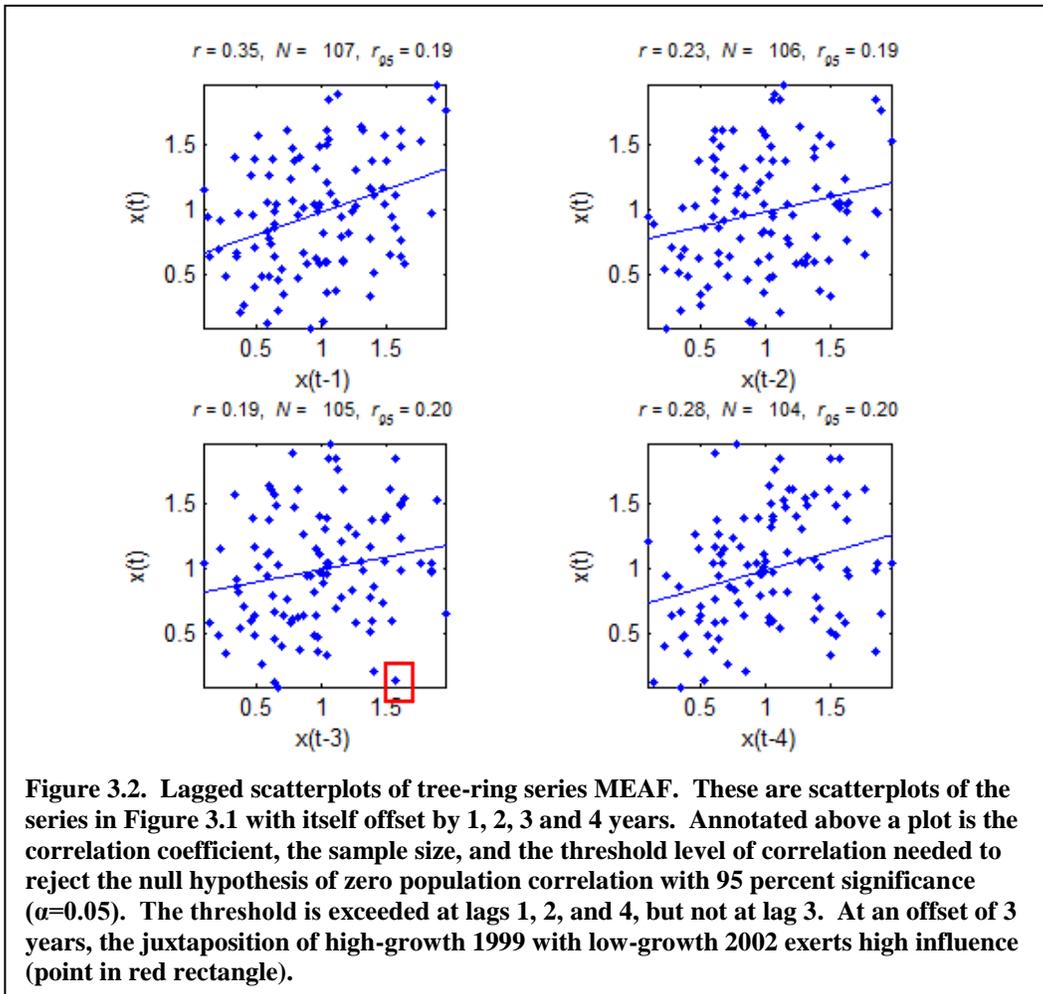
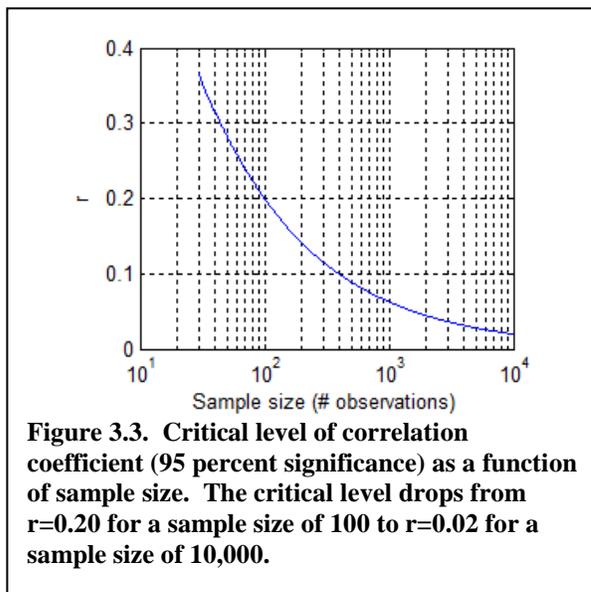


Figure 3.2. Lagged scatterplots of tree-ring series MEAF. These are scatterplots of the series in Figure 3.1 with itself offset by 1, 2, 3 and 4 years. Annotated above a plot is the correlation coefficient, the sample size, and the threshold level of correlation needed to reject the null hypothesis of zero population correlation with 95 percent significance ($\alpha=0.05$). The threshold is exceeded at lags 1, 2, and 4, but not at lag 3. At an offset of 3 years, the juxtaposition of high-growth 1999 with low-growth 2002 exerts high influence (point in red rectangle).

An attribute of the lagged scatterplot is that it can display autocorrelation regardless of the form of the dependence on past values. An assumption of linear dependence is not necessary. An organized curvature in the pattern of dots might suggest nonlinear dependence between time-separated values. Such nonlinear dependence might not be effectively summarized by other methods (e.g., the autocorrelation function [acf], which is described later). Another attribute is that the lagged scatterplot can show if the autocorrelation is characteristic of the bulk of the data or is driven by one or more outliers. The scatter plot in Figure 3.2 for lag-3 (lower left plot), for example, has a distinct lower-left to upper-right slant supporting positive lag-3 autocorrelation, but an outlier (highlighted) probably keeps the lag-3 autocorrelation from reaching statistical significance. Influence of outliers would not be detectable from the acf alone.

Fitted line. A straight line can be fit to the points in a lagged scatterplot to facilitate evaluation linearity and strength of relationship of current with past values. A series of lagged scatterplots at increasing lags (e.g., $k = 1, 2, \dots, 8$) helps in assessing whether dependence is restricted to one or more lags.

Correlation coefficient and 95% significance level. The correlation coefficient for the scatterplot summarizes the strength of the **linear** relationship between present and past values. It is helpful to compare the computed correlation coefficient with critical level of correlation required to reject the null hypothesis that the sample comes from a population with zero correlation at the indicated lag. If a time series is completely random, and the sample size is large, the lagged-correlation coefficient is approximately normally distributed with mean 0 and variance $1/N$ (Chatfield 2004). It follows that the approximate threshold, or critical, level of correlation for 95% significance ($\alpha = 0.05$) is $r_{.95} = 0 \pm 2 / \sqrt{N}$, where N is the sample size. Accordingly, the required level of correlation for “significance” becomes very small at large sample size (Figure 3.3).



3.3 Autocorrelation function (correlogram)

An important guide to the persistence in a time series is given by the series of quantities called the sample autocorrelation coefficients, which measure the correlation between observations at different times. The set of autocorrelation coefficients arranged as a function of separation in time is the sample autocorrelation function, or the acf. An analogy can be drawn between the autocorrelation coefficient and the product-moment correlation coefficient. Assume N pairs of observations on two variables x and y . The correlation coefficient between x and y is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2 \right]^{1/2} \left[\sum (y_i - \bar{y})^2 \right]^{1/2}} \quad (1)$$

where the summations are over the N observations.

A similar idea can be applied to time series for which successive observations are correlated. Instead of two different time series, the correlation is computed between one time series and the same series lagged by one or more time units. For the first-order autocorrelation, the lag is one time unit. The first-order autocorrelation coefficient is the simple correlation coefficient of the first $N - 1$ observations, $x_t, t = 1, 2, \dots, N - 1$ and the next $N - 1$ observations, $x_t, t = 2, 3, \dots, N$. The correlation between x_t and x_{t+1} is given by

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x}_{(1)})(x_{t+1} - \bar{x}_{(2)})}{\left[\sum_{t=1}^{N-1} (x_t - \bar{x}_{(1)})^2 \right]^{1/2} \left[\sum_{t=2}^N (x_t - \bar{x}_{(2)})^2 \right]^{1/2}} \quad (2)$$

where $\bar{x}_{(1)}$ is the mean of the first $N - 1$ observations and $\bar{x}_{(2)}$ is the mean of the last $N - 1$ observations. As the correlation coefficient given by (2) measures correlation between successive observations, it is called the autocorrelation coefficient or serial correlation coefficient.

For N reasonably large, the denominator in equation (2) can be simplified by approximation. First, the difference between the sub-period means $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ can be ignored. Second, the difference between summations over observations 1 to $N-1$ and 2 to N can be ignored. Accordingly, r_1 can be approximated by

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2} \quad (3)$$

where $\bar{x} = \sum_{t=1}^N x_t$ is the overall mean.

Equation (3) can be generalized to give the correlation between observations separated by k time steps:

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (4)$$

The quantity r_k is called the autocorrelation coefficient at lag k . The plot of the autocorrelation function as a function of lag is also called the *correlogram*.

Link between acf and lagged scatterplot. The correlation coefficients for the lagged scatterplots at lags $k = 1, 2, \dots, 8$ are equivalent to the acf values at lags $1, \dots, 8$.

Link between acf and autocovariance function (acvf). Recall that the variance is the average squared departure from the mean. By analogy the autocovariance of a time series is defined as the average product of departures at times t and $t+k$

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \quad (5)$$

where c_k is the autocovariance coefficient at lag k . The autocovariance at lag zero, c_0 , is the variance. By combining equations (4) and (5), the autocorrelation at lag k can be written in terms of the autocovariance:

$$r_k = c_k / c_0 \quad (6)$$

Alternative equation for autocovariance function. Equation (5) is a biased (though asymptotically unbiased) estimator of the population covariance. The acvf is sometimes computed with the alternative equation

$$c_k = \frac{1}{(N-k)} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \quad (7)$$

The acvf by (7) has a lower bias than the acvf by (5), but is conjectured to have a higher mean square error (Jenkins and Watts 1968, chapter 5).

3.4 Testing for randomness with the correlogram

The first question that can be answered with the correlogram is whether the series is random or not. For a random series, lagged values of the series are uncorrelated and we expect that $r_k \cong 0$. It can be shown that if x_1, \dots, x_N are independent and identically distributed random variables with arbitrary mean, the expected value of r_k is

$$E(r_k) = -1/N \quad (8)$$

the variance of r_k is

$$\text{Var}(r_k) \cong 1/N \quad (9)$$

and r_k is asymptotically normally distributed under the assumption of weak stationarity. The 95% confidence limits for the correlogram can therefore be plotted at $-1/N \pm 2/\sqrt{N}$, and are often further approximated to $0 \pm 2/\sqrt{N}$. Thus, for example, if a series has length 100, the approximate 95% confidence band is $\pm 2/\sqrt{100} = \pm 0.20$. Any given r_k has a 5% chance of being outside the 95% confidence limits, so that one value outside the limits might be expected in a correlogram plotted out to lag 20 even if the time series is drawn from a random (not autocorrelated) population.

Factors that must be considered in judging whether a sample autocorrelation outside the confidence limits indicates an autocorrelated process or population are (1) how many lags are being examined, (2) the magnitude of r_k , and (3) at what lag k the large coefficient occurs. A very large r_k is less likely to occur by chance than a smaller r_k barely outside the confidence bands. And a large r_k at a low lag (e.g., $k = 1$) is more likely to represent persistence in most physical systems than an isolated large r_k at some higher lag.

3.5 Large-lag standard error

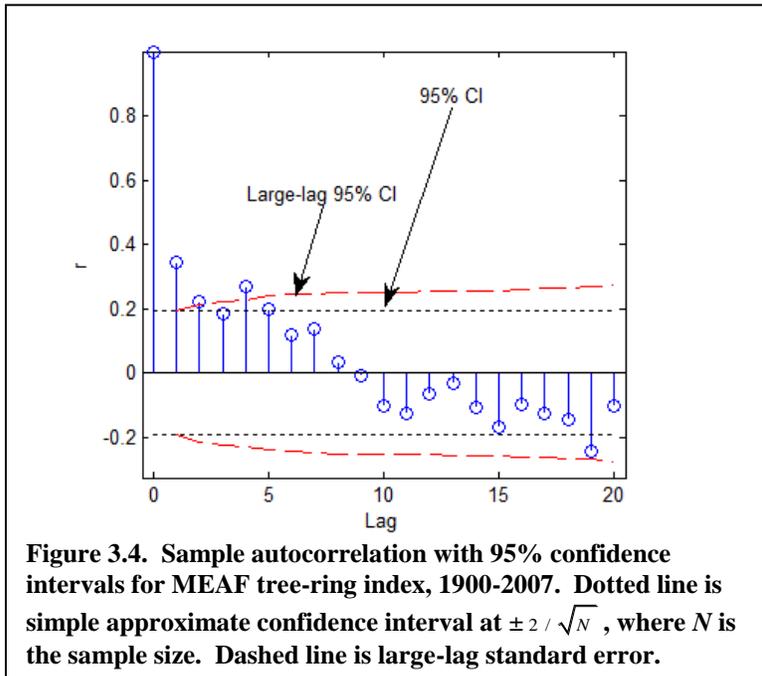
While the confidence bands described above are horizontal lines above and below zero on the correlogram, the confidence bands you see in the assignment script may appear to be narrowest at

lag 1 and to widen slightly at higher lags. That is because the confidence bands produced by the script are the so-called “large-lag” standard errors of r_k (Anderson 1976, p. 8). Successive values of r_k can be highly correlated, so that an individual r_k might be large simply because the value at the next lower lag, r_{k-1} , is large. This interdependence makes it difficult to assess just at how many lags the correlogram is significant. The large-lag standard error adjusts for the interdependence. The variance of r_k , with the adjustment, is given by

$$\text{var}(r_k) \approx \frac{1}{N} \left(1 + 2 \sum_{i=1}^K r_i^2 \right) \quad (10)$$

where $K < k$. The square root of the variance quantity given by (10) is called the *large-lag standard error* of r_k (Anderson 1976, p. 8). Comparison of (10) with (9) shows that the adjustment is due to the summation term, and that the variance of the autocorrelation coefficient at any given lag depends on the sample size as well as on the estimated autocorrelation coefficients at shorter lags. For example, the variance of the lag-3 autocorrelation coefficient, $\text{var}(r_3)$, is greater than $1/N$ by an amount that depends on the autocorrelation coefficients at lags 1 and 2. Likewise, the variance of the lag-10 autocorrelation coefficient, $\text{var}(r_{10})$, depends on the autocorrelation coefficients at lags 1-9. Assessment of the significance of lag- k autocorrelation by the large-lag standard error essentially assumes that the theoretical autocorrelation has “died out” by lag k , but does not assume that the lower-lag theoretical autocorrelations are zero (Box and Jenkins 1976, p. 35). Thus the null hypothesis is NOT that the series is random, as lower-lag autocorrelations in the generating process may be non-zero.

An example for a tree-ring index time series illustrates the slight difference between the confidence interval computed from the large-lag standard error and computed by the rough approximation $\pm 2/\sqrt{N}$, where N is the sample size (Figure 3.4). The alternative confidence intervals differ because the null hypotheses differ. Thus, the autocorrelation at lag 5, say, is judged significant under the null hypothesis that the series is random, but is not judged significant if the theoretical autocorrelation function is considered to not have died out until lag 5.



3.6 Hypothesis test on r_1

The first-order autocorrelation coefficient is especially important because for physical systems dependence on past values is likely to be strongest for the most recent past. The first-order autocorrelation coefficient, r_1 , can be tested against the null hypothesis that the corresponding population value $\rho_1 = 0$. The critical value of r_1 for a given significance level (e.g., 95%) depends on whether the test is one-tailed or two-tailed. For the one-tailed hypothesis, the alternative hypothesis is usually that the true first-order autocorrelation is greater than zero:

$$H_1 : \rho > 0 \quad (11)$$

For the two-tailed test, the alternative hypothesis is that the true first-order autocorrelation is different from zero, with no specification of whether it is positive or negative:

$$H_1 : \rho \neq 0 \quad (12)$$

Which alternative hypothesis to use depends on the problem. If there is some reason to expect positive autocorrelation (e.g., with tree rings, from carryover food storage in trees), the one-sided test is best. Otherwise, the two-sided test is best.

For the one-sided test, the World Meteorological Organization recommends that the 95% significance level for r_1 be computed by

$$r_{1,.95} = \frac{-1 + 1.645\sqrt{N - 2}}{N - 1} \quad (13)$$

where N is the sample size.

More generally, following Salas et al. (1980), who refer to Andersen (1941), the probability limits on the correlogram of an independent series are

$$r_k (95\%) = \frac{-1 + 1.645\sqrt{N - k - 1}}{N - k} \quad \text{one sided} \quad (14)$$

$$r_k (95\%) = \frac{-1 \pm 1.96\sqrt{N - k - 1}}{N - k} \quad \text{two sided}$$

where N is the sample size and k is the lag. Equation (13) comes from substitution of $k=1$ into equation (14).

3.7 Effective Sample Size

If a time series of length N is autocorrelated, the number of *independent* observations is fewer than N . Essentially, the series is not random in time, and the information in each observation is not totally separate from the information in other observations. The reduction in number of independent observations has implications for hypothesis testing.

Some standard statistical tests that depend on the assumption of random samples can still be applied to a time series despite the autocorrelation in the series. The way of circumventing the problem of autocorrelation is to adjust the sample size for autocorrelation. The number of independent samples after adjustment is fewer than the number of observations of the series. Below is an equation for computing so-called “effective” sample size, or sample size adjusted for autocorrelation. More on the adjustment can be found elsewhere (WMO 1966; Dawdy and Matalas 1964). The equation was derived based on the assumption that the autocorrelation in the series represents *first-order* autocorrelation (dependence on lag-1 only). In other words, the governing process is *first-order autoregressive*, or *Markov*. Computation of the effective sample

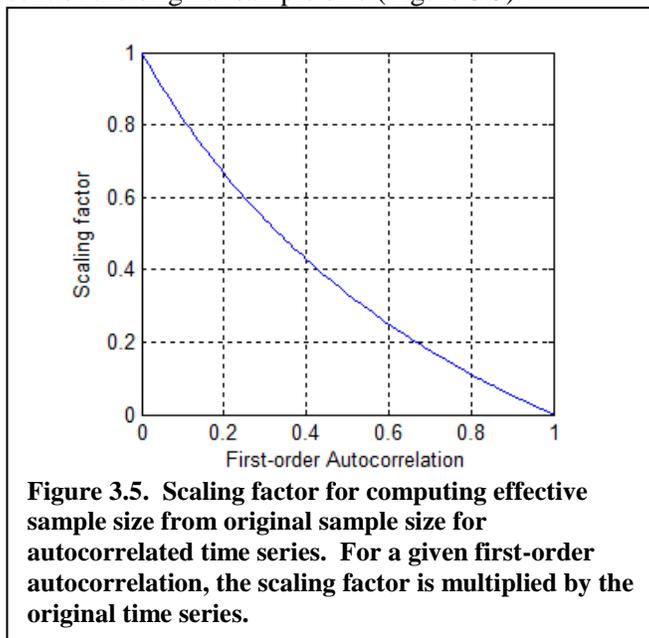
size requires only the sample size and first-order sample autocorrelation coefficient. The “effective” sample size is given by:

$$N' = N \frac{(1 - r_1)}{(1 + r_1)} \quad (15)$$

where N is the sample size, N' is the effective samples size, and r_1 is the first-order autocorrelation coefficient. The ratio $(1 - r_1)/(1 + r_1)$ is a scaling factor multiplied by the original sample size to compute the effective sample size. For example, an annual series with a sample size of 100 years and a first-order autocorrelation of 0.50 has an adjusted sample size of

$$N' = 100 \frac{(1 - 0.5)}{(1 + 0.5)} = 100 \frac{0.5}{1.5} \approx 33 \text{ years}$$

The adjustment to effective sample size becomes less important the lower the autocorrelation, but a first-order autocorrelation coefficient as small as $r_1=0.10$ results in a scaling to about 80 percent of the original sample size (Figure 3.5).



References

Anderson, R.L., 1941, Distribution of the serial correlation coefficients: *Annals of Math. Statistics*, v. 8, no. 1, p. 1-13.

Anderson, O., 1976, *Time series analysis and forecasting: the Box-Jenkins approach*: London, Butterworths, p. 182 pp.

Box, G.E.P., and Jenkins, G.M., 1976, *Time series analysis: forecasting and control*: San Francisco, Holden Day, p. 575 pp.

Chatfield, C., 2004, *The analysis of time series, an introduction*, sixth edition: New York, Chapman & Hall/CRC.

Dawdy, D.R., and Matalas, N.C., 1964, Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series, in Ven Te Chow, ed., *Handbook of applied hydrology, a compendium of water-resources technology*: New York, McGraw-Hill Book Company, p. 8.68-8.90.

Jenkins, G.M., and Watts, D.G., 1968, *Spectral analysis and its applications*: Holden-Day, 525 p.

Salas, J.D., Delleur, J.W., Yevjevich, V.M., and Lane, W.L., 1980, *Applied modeling of hydrologic time series*: Littleton, Colorado, Water Resources Publications, 484 pp.

World Meteorological Organization, 1966, Technical Note No. 79: Climatic Change, WMO-No, 195.TP.100, Geneva, 80 pp.