

10. Lagged Correlation

Lagged relationships are characteristic of many natural physical systems. Lagged correlation refers to the correlation between two time series shifted in time relative to one another. Lagged correlation is important in studying the relationship between time series for two reasons. First, one series may have a delayed response to the other series, or perhaps a delayed response to a common stimulus that affects both series. Second, the response of one series to the other series or an outside stimulus may be “smeared” in time, such that a stimulus restricted to one observation elicits a response at multiple observations. For example, because of storage in reservoirs, glaciers, etc., the volume discharge of a river in one year may depend on precipitation in the several preceding years. Or because of changes in crown density and photosynthate storage, the width of a tree-ring in one year may depend on climate of several preceding years. The simple correlation coefficient between the two series properly aligned in time is inadequate to characterize the relationship in such situations. Useful functions we will examine as alternative to the simple correlation coefficient are the cross-correlation function and the impulse response function. The cross-correlation function is the correlation between the series shifted against one another as a function of number of observations of the offset. If the individual series are autocorrelated, the estimated cross-correlation function may be distorted and misleading as a measure of the lagged relationship. We will look at two approaches to clarifying the pattern of cross-correlations. One is to remove the persistence from, or prewhiten, the two series before cross-correlation estimation. In this approach, the two series are essentially regarded on an “equal footing.” An alternative is the “systems” approach, in which one series is viewed as input to and the other series as output from a dynamic linear system. The lagged response is estimated through the “impulse response function,” which is the response of the output at current and future times to a hypothetical “pulse” of input restricted to the current time. Estimation of the impulse response function involves application of the same filter to the input and output series, followed by cross-correlation of the filtered series. The filter is designed such that it whitens the input series.

10.1 Cross-Correlation Function

The cross-correlation function (ccf) of two time series is the product-moment correlation as a function of lag, or time-offset, between the series. It is helpful to begin defining the ccf with a definition of the cross-covariance function (ccvf). Consider N pairs of observations on two time series, u_t and y_t . Following Chatfield (2004, p. 158), the sample ccvf is given by

$$\begin{aligned} c_{uy}(k) &= \frac{1}{N} \sum_{t=1}^{N-k} (u_t - \bar{u})(y_{t+k} - \bar{y}) \quad [k = 0, 1, \dots, (N-1)] \\ c_{uy}(k) &= \frac{1}{N} \sum_{t=1-k}^N (u_t - \bar{u})(y_{t+k} - \bar{y}) \quad [k = -1, -2, \dots, -(N-1)] \end{aligned} \quad (1)$$

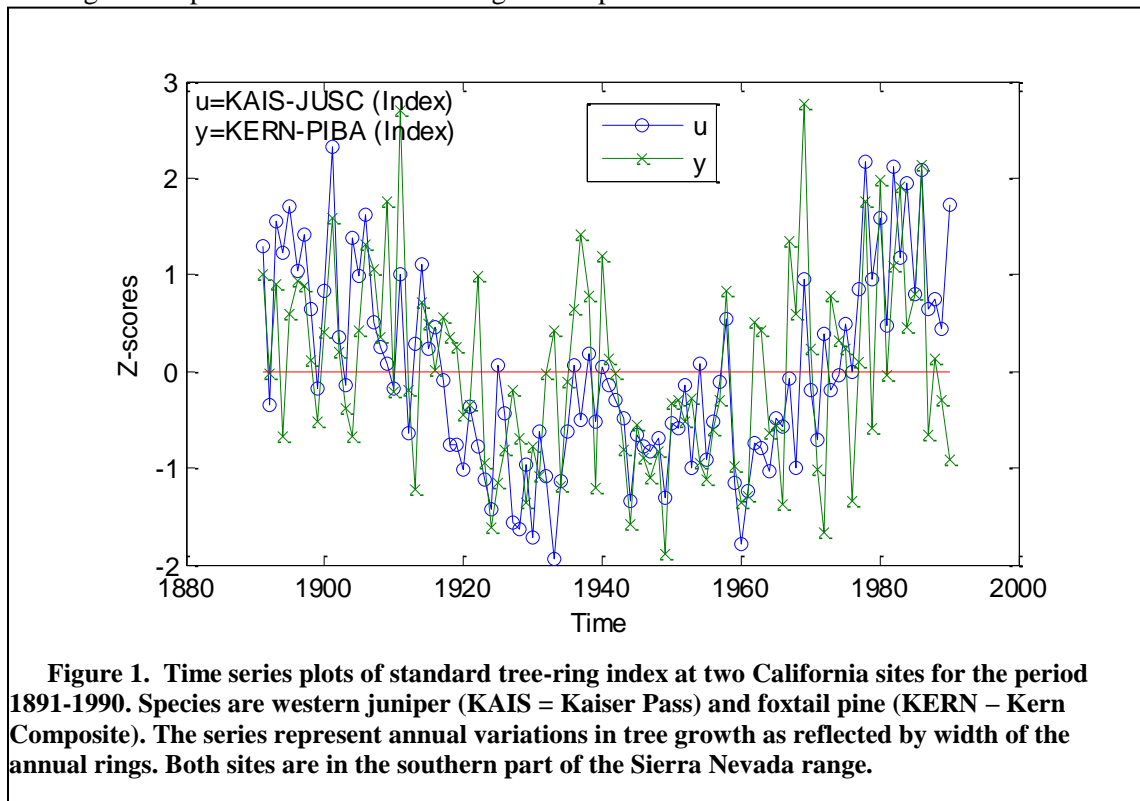
where N is the series length, \bar{u} and \bar{y} are the sample means, and k is the lag. The sample cross-correlation function (ccf) is the ccvf scaled by the variances of the two series:

$$r_{uy}(k) = \frac{c_{uy}(k)}{\sqrt{c_{uu}(0)c_{yy}(0)}} \quad (2)$$

Where $c_{uu}(0)$ and $c_{yy}(0)$ are the sample variances of u_t and y_t .

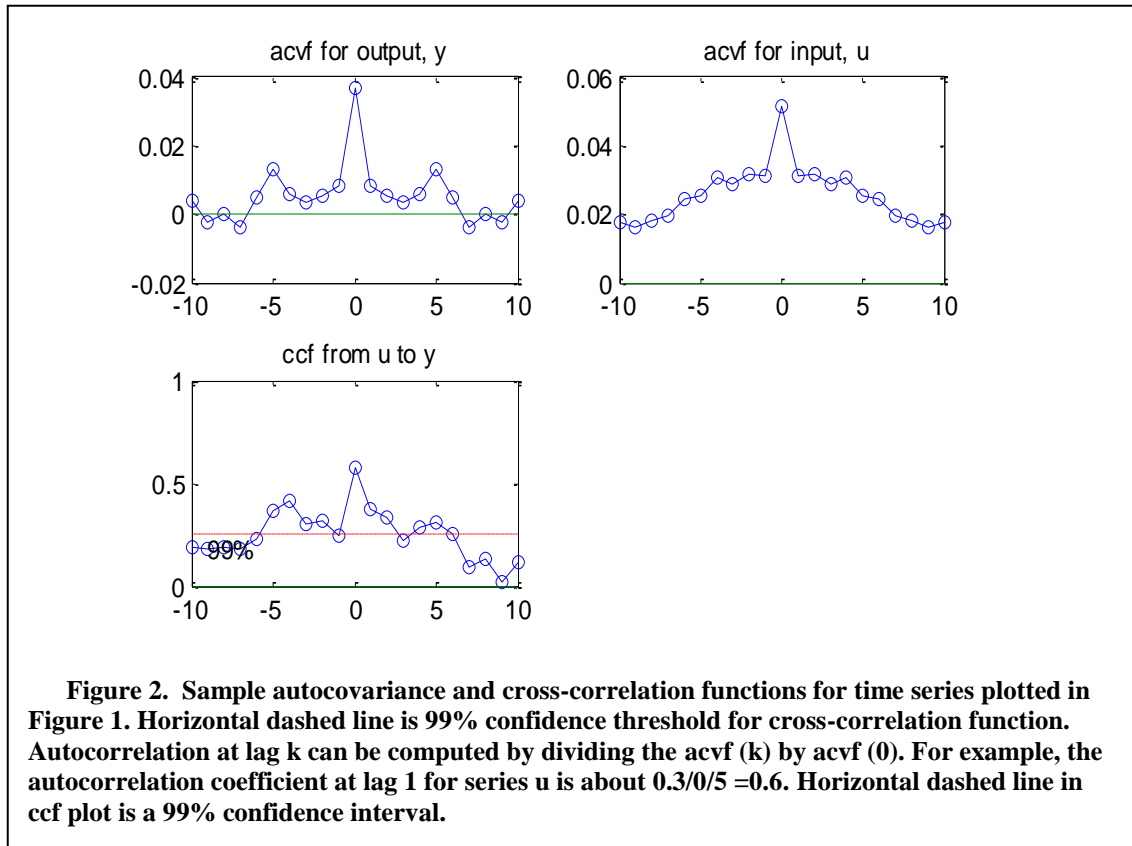
In a previous chapter we studied the autocovariance function (acvf) and autocorrelation function (acf), and learned that these are *symmetrical* functions (value at lag k equals value at lag $-k$). In contrast, the ccvf and ccf are asymmetrical functions. The asymmetry brings about the need for the two parts of equation (1). The cross-correlation function as described by equation (1) can be described in terms of “lead” and “lag” relationships. The first part of the equation applies to y_t shifted forward relative to u_t . With this direction of shift, u_t is said to “lead” y_t . This is equivalent to saying that y_t “lags” u_t . The second part of equation (1) describes the reverse situation, and summarizes lagged correlations when y_t “leads” u_t (u_t “lags” y_t).

Consider the two time series of annual tree-ring index plotted in Figure 1. The plots suggest the series are positively correlated, but the year-to-year variations are too noisy to visually judge whether one series tends to lead or lag the other. Both series, however, appear to be positively autocorrelated, as positive departures from the mean tend to follow positive departures, and negative departures tend to follow negative departures.



The autocovariance functions (acvf's) of the two series and the cross-correlation function (ccf) between them are plotted in Figure 2. The acvf is just the autocorrelation function (acf) multiplied by the variance, and so the acvf and acf have exactly the same shape but different scaling. The acvf is symmetric, as evidenced by $\text{acvf}(k)$ equaling $\text{acvf}(-k)$. The plotted acvf's in Figure 2 are consistent with positive autocorrelation in both series, as positive acvf at some lag k would re-scale to positive autocorrelation, but the statistical significance cannot be judged from these plots. The autocovariance at lag $k = 0$ is identically the variance, such that we can see from the top plots in Figure 2 alone that the variance of u is slightly greater than the variance of y .

The ccf as plotted indicates that u_t and y_t are significantly positively correlated at lag $k = 0$. The horizontal dashed line in Figure 2 marks the upper 99% confidence level for significance of the ccf. This confidence interval relies on several simplifying assumptions and can be computed from the sample size alone. For a two-tailed test, the approximate 99% confidence



interval ($\alpha = 0.01$) is $0 \pm 2.58/\sqrt{N}$, where N is the sample size. The value 2.58 is the 0.995 probability-point of the cdf of the normal distribution. This confidence interval relies on assumptions that 1) the processes generating u_t and v_t are uncorrelated, 2) the processes are not autocorrelated, 3) the populations are normally distributed, and 4) the sample size is large. Under those assumptions, the sample cross-correlations are $\sim N(0, 1/N)$, or normally distributed with mean zero and variance $1/N$. A ccf estimate outside the confidence interval is deemed “significant” in that the null hypothesis that the true ccf at that lag is zero must be rejected against the alternative hypothesis that the true ccf is non-zero.

The 99% confidence level for the ccf plotted in Figure 2 is a horizontal line and easy to calculate, but that simple interval relies on the stated assumptions. The first assumption, that the processes generating u_t and v_t are uncorrelated, is central to the null hypothesis, but the assumptions on normality and absence of autocorrelation are problematic. Geophysical time series often come from dynamic systems, which tend to impart autocorrelation. Moreover, the time series themselves may not be normally distributed, such that the sample autocorrelations may not be assumed to come from a normal distribution. Angell (1981) circumvents these two problems by computed an effective sample size, adjusted downward depending on the sample autocorrelations of u_t and v_t , and applying a Fisher’s Z transform (Panofsky 1958) to the sample cross-correlations before testing their significance with a normal distribution. The equation for the effective sample size, derived by simplifying Bartlett’s equation (Box and Jenkins 1976), is attributed by Angell (1981) to Quenouille (1952), and is as follows:

$$N' = N / (1 + 2r_{1,u}r_{1,v} + 2r_{2,u}r_{2,v} + \dots + 2r_{K,u}r_{K,v}), \quad (3)$$

where $r_{i,u}$ and $r_{i,v}$ are the autocorrelations of time series u_t and v_t at lag i , N is the length of the time series, and K is the maximum lag of autocorrelation used in the adjustment. In his example, Angell (1981) sets K to the highest lag for which the product of sample autocorrelations is non-

negligible. Another reasonable approach might be to set K to the highest lag before the autocorrelation of either u_t or v_t becomes non-positive.

Angell (1981) then computes the sample cross-correlations $r_{u,v}(k)$, applies the Fisher Z-transform

$$z(k) = \frac{1}{2} \left[\ln(1 + r_{u,v}(k)) - \ln(1 - r_{u,v}(k)) \right], \quad (4)$$

and tests the significance of z by referring to a normal distribution with mean 0 and variance

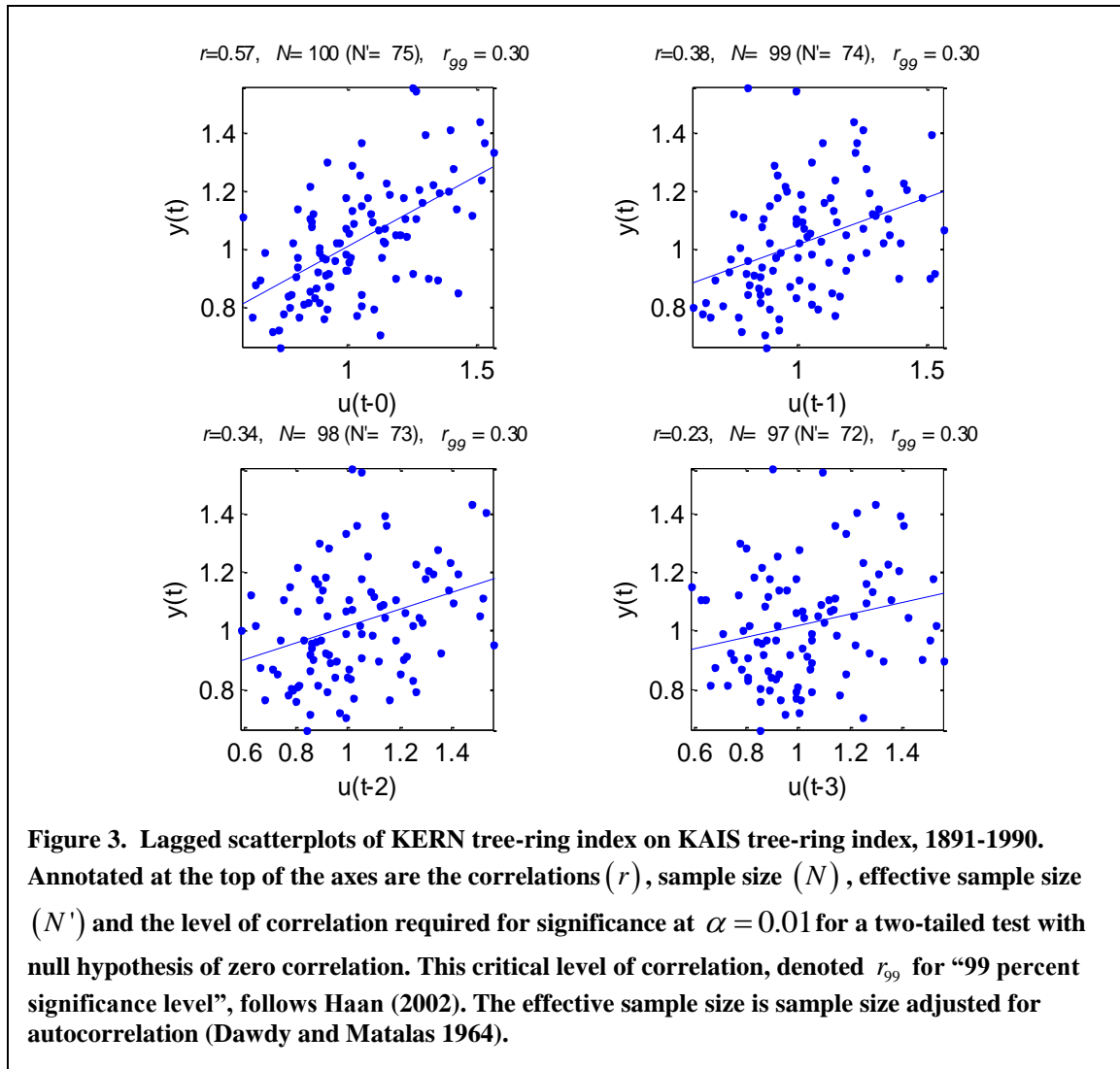
$$\sigma_z^2 = \frac{1}{N' - 4} \quad (5)$$

Non-normality can alternatively be handled by transformation of the series to normality, followed by application of techniques suited for normally distributed variables. Violations of normality, and of other assumptions, can also be dealt with by Monte Carlo techniques, which do not rely on an assumed theoretical distribution of the lagged correlations. A Monte Carlo approach would consist of generating many (e.g., 10,000) simulated time series of u_t with the same distribution and autocorrelation properties as the observed u_t , and using the cdf of lagged correlations of v_t with those simulated series to place empirical confidence intervals on the lagged correlations of observed u_t and v_t . For more information on Monte Carlo methods, which are not addressed in this chapter, the reader is referred elsewhere (e.g., Martinez and Martinez 2002, p 191; Salas et al. 1980, p 97; Haan 202, p. 331).

The significant cross-correlations in Figure 2 extend over several lags, such that u_t is significantly correlated with y_t shifted several lags forward and backwards. If the ccf in Figure 2 were significant at only zero and positive lags, it could be said that y_t tended to “lag” u_t . Such a pattern with input u_t and output y_t would be consistent with a “causal” system. The plotted ccf in Figure 2 is significant different from zero at negative as well as non-negative lags, and consequently is not consistent with a causal system. This is not surprising for these data, which are two tree-ring series with no physically causal relationship to imply that variations in one series should in any sense “drive” variations in the other.

The ccf at a particular lag can be regarded as a correlation coefficient between two time series, one of which just happens to be shifted some number of time units. Just as the bivariate relationship for any two variables can be examined with a scatterplot, the bivariate relationship posed by the ccf can be examined with lagged scatterplots. Such plots might indicate oddities in the data – for example, that the ccf at a particular lag is driven by one or two outlier values of the time series. Lagged scatterplots for the two time series used as an example above are plotted in

Figure 3. The plots show (y_t) in year t plotted against (u_t) in years $t, t-1, t-2$, and $t-3$.



Like the ccf, the scatterplots show the strongest relationship at lag 0. Correlation at lag 0 ($r=0.57$) is by definition the value of the cross-correlation function at lag 0 (compare correlation coefficient for lag 0 in Figure 3 with lag-0 cross-correlation in Figure 2). The computed correlation coefficients (r) in Figure 3 can be compared with the threshold correlation required for significance at the 99% confidence level, r_{99} . The scatterplots indicate significant correlations at lags $k = 0, -1, -2$, which is consistent with the ccf plot in Figure 2. An exact match of lags with significant correlation will not generally be found in comparing the lagged scatterplots with the ccf plot because the significance thresholds in the scatterplots have been adjusted for autocorrelation in the individual series (see previous lecture). For example, the 99% confidence level is at $r_{u,y}(k) = 0.26$ in the ccf (Figure 2), and at $r_{99} = 0.30$ in the scatterplots of Figure 3

10.2 Prewhitening to clarify lagged relationships

As already mentioned, lagged correlations between time series can be misleading as evidence of lagged relationships and dependence, especially if the individual time series are

autocorrelated. Jenkins and Watts (1968) suggest that to clarify lagged relationships the individual series both be prewhitened before estimating the ccf. In contrast to the systems approach described later, this approach to lagged relationships treats time series on an “equal footing” (Chatfield 2004, p. 155). The systems approach, on the other hand, treats one series as an input to a hypothetical system and the other series as an output.

Prewhitening in the context of estimation of the ccf is intended to deal with the complicating effects of autocorrelation on the estimated ccf and its standard deviations. Referring to the estimators of the ccvf and ccf ($c_{uy}(k)$ and $r_{uy}(k)$), Chatfield (2004, p. 158) says:

It can be shown that these estimators are asymptotically unbiased and consistent. However it can also be shown that estimators at neighbouring lags are themselves autocorrelated. Furthermore, it can be shown that the variances of sample cross-correlations depend on the autocorrelation functions of the two components. In general, the variances will be inflated. Thus even for moderately large values of N up to about 200, or even higher, it is possible for two series, which are actually unrelated, to give rise to apparently ‘large’ cross-correlation coefficients which are spurious, in that they arise solely from autocorrelations within the two series. Thus, of a test is required for non-zero correlation between two time series, then (at least) one of the series should first be filtered to convert it to (approximate) white noise.

Chatfield goes on to suggest the systems approach, in which one time series, regarded as “input,” is prewhitened by a time series model. The other series is filtered by the same model, and correlations between the prewhitened input and filtered output are plotted and examined. He also mentions the alternative approach of prewhitening *both* time series by fitting them individually to time series models, and then examining the cross-correlations between the prewhitened series. This latter approach is suggested by Jenkins and Watts (1968) and Brockwell and Davis (2002), and consists of a non-systems, or “equal-footing” approach.

10.3 “Equal-Footing” Approach

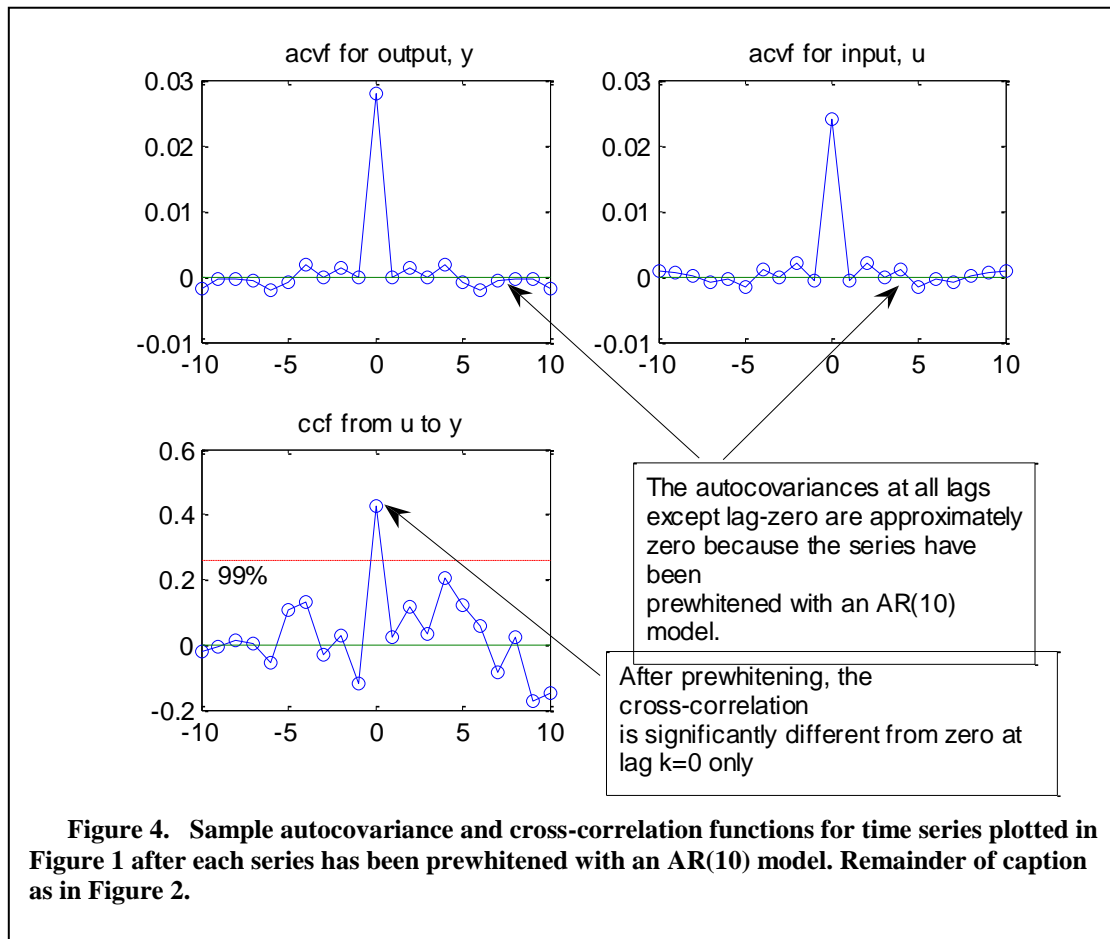
As in section 10.1, we consider two time series, u_t and y_t . Each series is individually fit to a high-order autoregressive model (say, AR(10)), and each is then prewhitened by its respective model to generate the two prewhitened series α_t and β_t . The ccf is computed for these prewhitened series. Because neither series is autocorrelated after prewhitening, we can apply the conventional non-adjusted confidence interval estimation for a correlation coefficient to test the significance of the ccf at various lags. The approximate variance of the cross-correlations is $1/N$, where N is the number of observations. An approximate 95% confidence interval for the ccf is therefore $0 \pm 1.96 / \sqrt{N}$, and an approximate 99% interval is $0 \pm 2.58 / \sqrt{N}$.

For the tree-ring example discussed previously, the ccf of the prewhitened series indicates that the series are significantly correlated at lag zero only (Figure 4). This result is quite different from the ccf on the original series (Figure 2), and suggests no lag in the relationship between the two tree-ring series. The significant ccf values at non-zero lags in Figure 2 might therefore be considered artifacts resulting from autocorrelation in the individual series

10.4 Systems Approach -- Impulse Response Function (IRF)

As discussed above, autocorrelation in the individual time series makes direct use of the ccf to study lagged relationships problematic. Essentially, the questions asked of the ccf are 1) how strongly is one series related to another, 2) is the relationship simultaneous or distributed over several time steps, and 3) if distributed, how many lags are involved and what is their relative importance?

The prewhitening approach described in the preceding section treated each series on an equal footing in dealing with autocorrelation. These questions can also be addressed by a systems



approach in which the series are regarded as input to and output from a *linear dynamic system* (Figure 5). *Dynamic* refers to the possible dependence of the output at time t on the input signal at many previous times. For such a system, the hypothetical response to a unit pulse of input at time $t = 0$ is given by the *impulse response function* (IRF).

For the tree-ring example, the IRF gives the hypothetical response of series KERN at times t , $t+1$, $t+2$, ... to a pulse of anomalous (above or below mean) "input" from series KAIS at time t . (Of course in a tree-ring context this is an artificial conceptualized system, and is used here only to illustrate methods. A more natural input-output system might have precipitation as input and tree-growth as output). The system ideally has the following properties:

- *time invariant*: response to an input signal does not depend on absolute time
- *linear*: output response to a linear combination of inputs is the same as the linear combination of the output responses to the individual inputs
- *causal*: output at a certain time depends on the input up to that time only

The system sketched in Figure 5 is described by the equation

$$y(t) = \sum_{k=1}^{\infty} g(k)u(t-k) + v(t)$$

where

$y(t)$ = output in year t (6)

$u(t-k)$ = input in year $t-k$

$g(k)$ = impulse response function at lag k

$v(t)$ = noise term in year t

The summation as written, following the notation in Ljung (1987), does not allow a contemporaneous response, or a response at time t to an input at time t . This can be remedied with no loss of generality by shifting the input one time step relative to the output (re-aligning the series) so that the summation effectively starts with $k=0$. Realignment can also insure that for a system with nominally assigned input and output, the output response does not precede the input stimulus.

The model represented by (6) gives the output as a linear combination of past (and possibly current) input. The numbers $\{g(k)\}$ are called the impulse response function (IRF). Generally, the IRF is unknown, and must be estimated from the data -- the input signal

$$u(t), t = 1, \dots, N \quad (7)$$

and the output signal

$$y(t), t = 1, \dots, N. \quad (8)$$

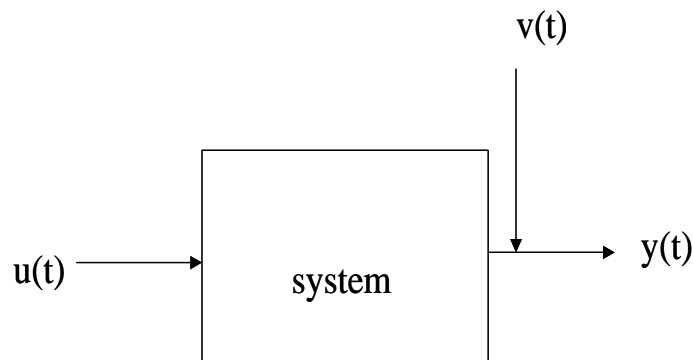


Figure 5. Input-output system with superposed noise. For example, input $u(t)$ might be rainfall, output $y(t)$ a tree-ring index, and $v(t)$ a disturbance variable incorporating all other influences on the tree-ring index. In the example used in class, one tree-ring chronology (KAIS) is used as the nominal input, and another chronology (KERN) as the nominal output (see Figure 1).

10.5 Estimation of the IRF

The “systems” method for estimating the IRF to be described is based on reducing one of the series to white noise before computing its correlation with the other series. Unlike in the “equal footing” approach described in section 10.3, only the input series is explicitly prewhitened.

The systems method for clarifying the lagged relationships in the presence of autocorrelation amounts to passing the input and output series through a filter before computing the cross-correlation function. The filter is chosen such that it reduces the input series to white noise (removes the autocorrelation). The filtered input series is therefore called the “prewhitened” input. The output is passed through the same filter, but the filtered output will generally not be white noise because the filter has been designed specifically to prewhiten the input, not the output. The ccf between the prewhitened input and filtered output is an estimate of the IRF of the system.

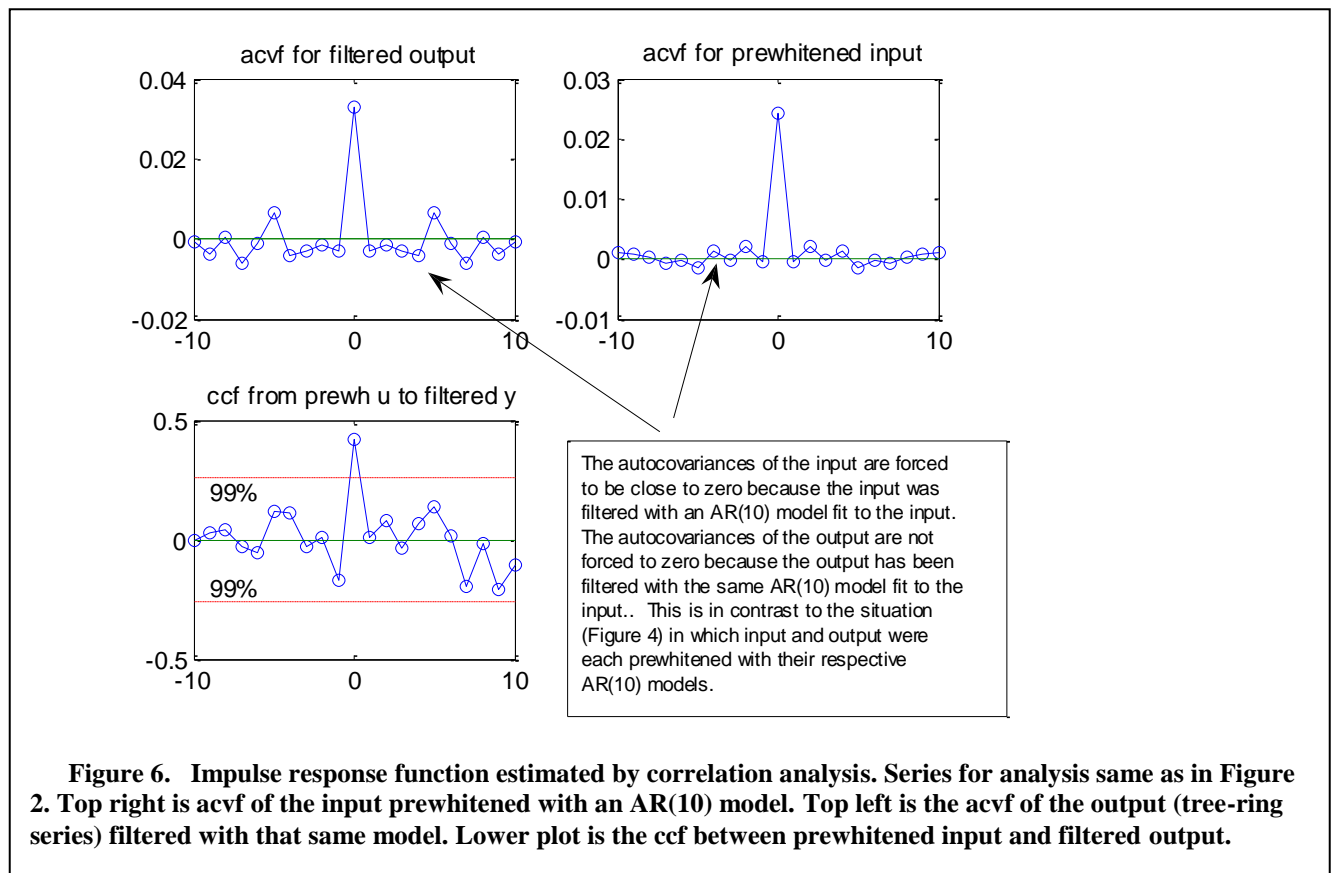
The filter for this operation is a high-order (e.g., order 10) autoregressive (AR) model fit to the input series. Figure 6 illustrates estimation of the IRF by this method for the tree-ring example. Before computing the ccf, we have prewhitened the input series with a 10th order autoregressive model and filtered the output series with that same model. The 10th order AR model has of course modified the input series so that its autocovariance is essentially zero through lag 10. The acvf of the output filtered by this same model is clearly affected, but has not been reduced to zero. In other words, the filtered output is not white noise.

The IRF, or the ccf between the prewhitened input and filtered output describes the lagged correlation structure disentangled from the influence of autocorrelation. The IRF in Figure 6 indicates significant response is restricted to lag 0. This result is in agreement with the results of the “equal footing” approach in section 10.3.

The 99% confidence interval for the IRF is once again (as in Figure 2) a horizontal line at $0 \pm 2.58/\sqrt{N}$. A constant CI is deemed applicable this estimated IRF because one of the series is now approximately white (Ljung 1987).

10.6 Summary Comments

Several alternative methods of examining lagged bivariate relationships have been presented. A simple graphical approach is the lagged scatterplot, but this approach may become cumbersome when many scatterplots must be examined to cover the possibility of relationship at higher lags. Moreover, for autocorrelated series, the lagged correlations at various lags are interdependent. The cross-correlation function is a the basic tool for studying correlation as a function of lag, but the pattern of weights in the ccf of the original time series is again distorted when the individual series are autocorrelated. The autocorrelation can be dealt with by removing it prior to computation of the ccf. In the “equal footing” approach, both time series are prewhitened, such that separate AR models are fit to each series. In the “systems” approach, just the input is prewhitening, while the output is filtered with the same AR model fit to the input.



An example with two tree-ring chronologies from the same geographical region was used to illustrate the methods. These chronologies are highly autocorrelated, and it is apparent from the results that the autocorrelation distorts the estimated “raw” cross-correlation function sufficiently that a cross-correlation analysis on the original series is not suitable for identifying lagged relationships between the series. The “equal footing” and “systems” approaches to studying the lagged relationship agree for this example in pointing to a contemporaneous (no lag) relationship.

The more highly autocorrelated the time series (especially the input), the more important prewhitening and filtering become in studying lagged relationships between time series. If the input already happens to be white noise, the “systems” approach becomes pointless: an AR model fit to the input will have little effect, and the ccf of the original data will be virtually the same as the ccf and FIR (except for scaling) after filtering the data with the AR model. Likewise, if both series are white noise, the prewhitened series in the “equal footing” method will be approximately the same as the original series, and the ccf of the original data will differ negligibly from the ccf of the prewhitened data.

10.7 References

- Angell, J. K. (1981), Comparison of variations in atmospheric quantities with sea surface temperature variations in the equatorial eastern pacific, *Mon. Weather Rev.*, 109:230–243.
- Box, G.E.P., and Jenkins, G.M., 1976, *Time series analysis: forecasting and control*: San Francisco, Holden Day, p. 575 pp.
- Brockwell, P. J., and Davis, P. J., 2002, *Introduction to Time Series and Forecasting*, 2nd edition. Springer, New York.
- Chatfield, C., 2004, *The Analysis of Time series, An Introduction*, sixth edition. Chapman & Hall/CRC, New York. 333 pp.
- 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. (sample size adjustment for first-order autocorrelation)
- Haan, C.T., 2002, *Statistical Methods in Hydrology*, second edition: Ames, Iowa, Iowa State University Press.
- Ljung, L., 1987, *System Identification: Theory for the User*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 07632, 519 pp. [*Impulse response function*]
- Martinez, W. L., and Martinez, A. L., 2002. *Computational Statistics Handbook with Matlab*, Chapman & Hall/CRC, 591 pp.
- Quenouille, M. H. (1952), *Associated Measurements*, Butterworths, 241 pp.
- Salas, J. D., Delleur, J. W., Yevjevich, V., and Lane, W. L., 1980. *Applied Modeling of Hydrologic Time Series*, Water Resources Publications, P. O. Box 2841, Littleton, CO80161, USA.