SOME EFFECTS OF CAUSAL CONNECTIONS IN SIMULATED TIME-SERIES DATA[1]

Donald C. Pelz and Ray E. Faith

Abstract

By computer simulation, two autoregressive panels x and y are generated, defined recursively by the relations $x_{it} = p_{xx}x_{it-1} + u_{it}$ and $y_{it} = p_{yy}y_{it-1} + c_{yx}x_{it-g} + v_{it}$, where u and v are uncorrelated disturbances, $|p_{xx}| \leq 1$ and $|p_{yy}| \leq 1$ are the autoregressive coefficients, and c_{yx} represents the causal effect on y of a prior x over causal interval g.

With simulated data the correlogram of cross-correlations $r(x_{it}, y_{it+k})$, where k represents any measurement lag, exhibited an expected asymmetric shape: correlations were higher when y was measured later than x. However, it was also observed that greater stability in either variable (higher p_{xx} or p_{yy}) produced higher cross-correlations--an apparently stronger causal connection. Another counter-intuitive observation was that, for certain high levels of p_{xx} and p_{yy}, the cross-correlation was maximum not over causal interval g but rather after some delay beyond g.

By assuming stationary properties, it has been possible to apply the procedures of path analysis to derive expressions which are identical with previous derivations by Lew, at the same time making it easier to grasp the model and explain its peculiar features.

* * * *

Origin of inquiry

The present investigation originated in an inquiry more than six years ago by the senior author and a colleague at the Survey Research Center [5] on the use of lagged correlations between two variables measured at two points in time, as a possible means of detecting direction of causal influence between the variables. It seemed intuitively plausible that if a change in variable x at time 1 produced a change in the same direction in variable y at time 2, then the lagged correlation between x_1 and y_2 should be larger than the corresponding lagged correlation between y_1 and x_2. Independently, Campbell and his associates arrived at the same conjecture [1,2].

Efforts to find mathematical support for this conjecture in the time series literature were unsuccessful, although at a later point Goldberger (personal correspondence, 1969); [and 3, p. 33] indicated a rather simple way of doing so. As a heuristic device, therefore, the senior author decided to simulate artificial time series data having known characteristics for a population of hypothetical individuals, in order to investigate how the introduction of known causal influence would affect the lagged correlations between variables. Spyros Magliveras devised a program to generate two artificial variables x and y for a set of N individuals over 50 periods of time.[2]

A legitimate question is raised as to whether the same results might not have been generated directly by mathematical analysis. In answer, we hope that analysis can eventually do the entire job, but meanwhile the simulation has been useful in demonstrating certain properties that were not obvious even to time series authorities, and which have been demonstrated analytically only with considerable labor.[3]

The simulation model

The present paper will deal with the simplest situation with two variables. For N individuals, independent variables x is determined only by its own prior value and by random outside disturbances; dependent variable y is also generated as a function of its own prior value and a different set of random disturbances, and in addition x exerts a uni-directional influence over a subsequent y after a specified causal interval g.

For each individual i in a population of N individuals, two time series x_{it} and y_{it} are generated at times t = 1,2,3,... in successive operations of the computer program, by the recursion equations

(1) $$x_{it} = p_{xx}x_{it-1} + u_{it}$$

(2) $$y_{it} = p_{yy}y_{it-1} + c_{yx}x_{it-g} + v_{it}$$

where u and v are random disturbances uncorrelated across individuals and across time, $|p_{xx}| \leq 1$ and $|p_{yy}| \leq 1$ are the respective autogregressive coefficients, and c_{yx} is the causal coefficient representing the effect on y of the x value for that individual which occurred g time units earlier, g being the causal interval.[4]

The coefficients p_{xx} and p_{yy} are equal to the theoretical autocorrelations between adjacent values of the Markov series x_t and y_t [4, pp. 405 ff.]. They will be called coefficients of "short-term stability."

The model also permits incorporation of "long-term stability" consisting of individual constants for each individual ζ_{ix} and ζ_{iy}, which remain unchanged through time. There is not space here to discuss the latter; the reader is warned that effects of the two sources of stability are markedly different.[5]

Some unexpected features

From time series data thus generated, one can obtain three sets of correlations between pairs of variables across N individuals, where measurements of the variables in each pair is separated by an interval designated k. Two of the sets are the autocorrelations $r(x_{it},x_{it+k})$ and $r(y_{it},y_{it+k})$; the third set is the cross-correlation $r(x_{it},y_{it+k})$, where measurement lag k can be any positive or negative integer 0, 1, 2, ...

It is instructive to compare the behavior of these correlational functions with what was expected in advance. The time-series literature indicated that for independent variable x the autoregressive correlogram (i.e., the plot of autocorrelations $\rho(x_{it}, x_{it+k})$ as a function of measurement interval k, where the symbol ρ designates the theoretical correlation in contrast to the empirical value r) should be a declining exponential function of the form: p_{xx}^k. The literature gave no immediate expectation as to the autocorrelational function of y, although we assumed it would decline exponentially in much the same fashion as x. These expectations were confirmed with the simulated data.[6]

We could find no theoretical expectation for the cross-correlational function. An intuitive conjecture was that as measurement interval k approached causal interval g, the cross-correlation between x and y would rise from zero to a maximum value at g, after which it would decline again toward zero. (These conjectures are illustrated in [7], pp. 6-10.) Under certain circumstances, a cross-correlogram of this shape did appear in simulated data, as illustrated in the bottom curve in Figure 1, where p_{xx} and p_{yy} are both set = .70.

In such a case, the phenomenon we have called the "cross-lagged differential" (following a suggestion by Campbell) will appear. If one compares the cross-correlation when y is measured at g time units __after__ x (in Figure 1, at the point k = +4) with the corresponding correlation when y is measured g time units __before__ x (at the point k = -4) a distinct difference is observed. This differential may disappear, however, if the measurement interval departs substantially from true causal interval. (In the bottom curve of Figure 1, e.g., there is no difference in height of the curve at k = +15 and -15).

In advance of the simulation, we were uncertain what effect an increase in stability would have on the cross-correlations. It seemed likely [7, p. 45] that extremely high stability of __either__ short-term or long-term character (see footnote 5) might cause the cross-correlogram to rise and fall

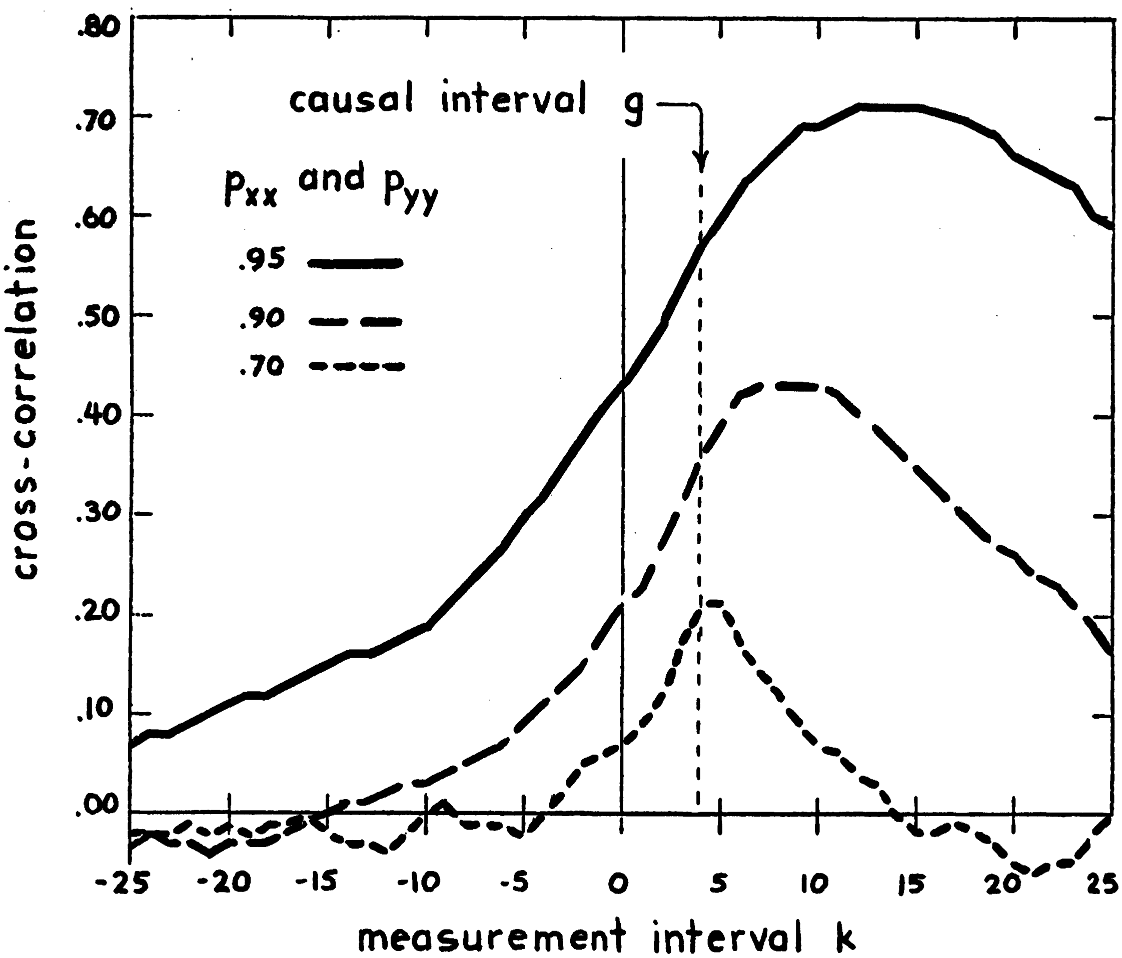

Figure 1. As autoregressive stability (p_{xx} and p_{yy}) increased for simulated variables x and y, the cross-correlations (1) became higher at point g, (2) increased in duration, (3) were increasingly delayed in reaching a maximum height beyond causal lag g. All curves used a constant causal coefficient (c_{yx} = .10). (From [7], p. 36.)

very slowly, and thus would obscure the cross-lagged differential.

When simulated data were generated and cross-correlations examined, some surprising departures from expectation appeared when high autoregressive coefficients p_{xx} or p_{yy} were used. The reasons for these departures were not intuitively obvious, but it is hoped that the mathematical analysis reported below will clarify them.

1. The higher the autoregressive coefficients (that is, the higher the short-term stability) the larger was the cross-correlation which appeared at causal interval g, and the sharper was the cross-lagged differential. Note in Figure 1 the increasing height of the cross-correlograms at $k = g = 4$, as p_{xx} and p_{yy} were increased from .70 to .95.

Thus in terms of the observed correlation between x and y, the causal influence <u>appeared</u> to become stronger, even though the actual coefficient of causal influence (c_{yx}) remained constant.

2. The higher the stability coefficients, the longer the cross-correlations <u>persisted</u> through time, over measurement lags several times longer than the causal interval. Thus in Figure 1 if one measures the cross-lagged differential--the difference in height of the correlogram at points +k and -k respectively--in the bottom curve the differential has disappeared at $k = \pm 15$. But in the top curve the differential is substantial even at $k = \pm 25$, and by extrapolation it appears to persist to $k = \pm 50$ or beyond.

Thus high autoregressive (or short-term) stability served to to <u>magnify and to perpetuate</u> the observable effects of causal influence.

3. Counter to intuitive expectation, the point of maximum cross-correlation was not necessarily at the point of actual causal impact (causal interval g). Rather, the higher the autoregressive coefficients, the greater was the <u>delay</u> between g and the point of maximum cross-correlation. In Figure 1, when the stability coefficients were extremely high (both = .95) the cross-correlation was maximum at an interval of from 10 to 15 time periods after the causal interval.

4. We had expected that a cross-lagged differential would appear only if it took <u>time</u> for the causal effect of x to appear in y--that is, only if the causal interval were several time units (one "time unit" being the interval for the autoregressive effect to occur). This expectation was confirmed in the case of variables with low stability. The effect of shortening g was simply to move the entire cross-correlogram to the left, without altering its shape. Thus, if the bottom curve in Figure 1 were moved 3 or 4 time units to the left, it would become almost symetrical around $k = 0$, and the cross-lagged differential would almost disappear.

Note however that with high stability (middle and upper curves in Figure 1), even if the curves were moved leftward so that $g = 0$ the asymmetry would persist, and so would the cross-lagged differential. An important property emerged: if the two time series were quite stable (in the short-term sense of high autoregressive coefficients), the causal influence of x on y remained apparent <u>even when the causation was almost simultaneous</u>.

<u>Mathematical properties derived by Lew</u>.

Robert Lew undertook the task of deriving mathematical properties of the simulation model. Like the model, he began with a finite starting point at which values for x_{it} and y_{it} were generated randomly for each of N individuals; uncorrelated random terms u_{it} and v_{it} were added; and after g time-periods had passed the causal component $c_{yx}x_{it-g}$ was added to the y variable. When the mathematical equivalent of this model was in operation over a long time certain terms vanished, and the equations could be expressed in asymptotic form.

The basic expressions for autocorrelations of x_{it} and y_{it}, and for the cross-correlations xy, are given in a technical note to Pelz and Lew [6] and in full detail in appendices of the interim report [7]. Lew's expressions generate theoretical autocorrelations and cross-correlations which parallel very closely the empirical curves produced by simulation such as those in Figure 1.

Lew also derived a prediction of phenomenon 3 noted above, that when autoregressive coefficients are high, the maximum cross-correlation will appear with some delay following causal interval g, as illustrated in Figure 2.

<u>Theoretical properties derived</u>
<u>from application of path analysis</u>

Co-author Faith has conceptualized our time series model in terms of a path analysis model which is fully described in the technical appendix.[7] The path diagram is given in Figure 3. Variables x_0 and y_{g-1} are specified as the correlated inputs. From these, together with the residual disturbances u and v, new variables x_t and y_t are defined by the following recursive path formulas:

$$(3)\quad x_{t+1} = p_{xx}x_t + p_{xu}u_{t+1} + 0\cdot(\text{all variables } x_s,\ y_{s+t} \text{ where } s \quad t)$$

$$(4)\quad y_{t+g} = p_{yy}y_{t+g-1} + p_{yx}x_t + p_{yv}v_{t+g} + 0\cdot(\text{all variables } x_s,\ y_{s+g-1} \text{ where } s \quad t)$$

for $t = 0,1,\ldots$

These are similar to the basic simulation equations (1) and (2) given above, when operating over an extended time period.

Contrary to Lew's approach of assuming an arbitrary set of initial conditions which gradually stabilize, the path model assumes a system that is stationary from the outset. This assumption is achieved by a suitable restriction of the parameters of the model (notably the causal path coefficient p_{yx} in expression (4), and the correlation between the variables regarded as inputs.)

It is deduced that under the specified conditions, the autocorrelations and cross-correla-

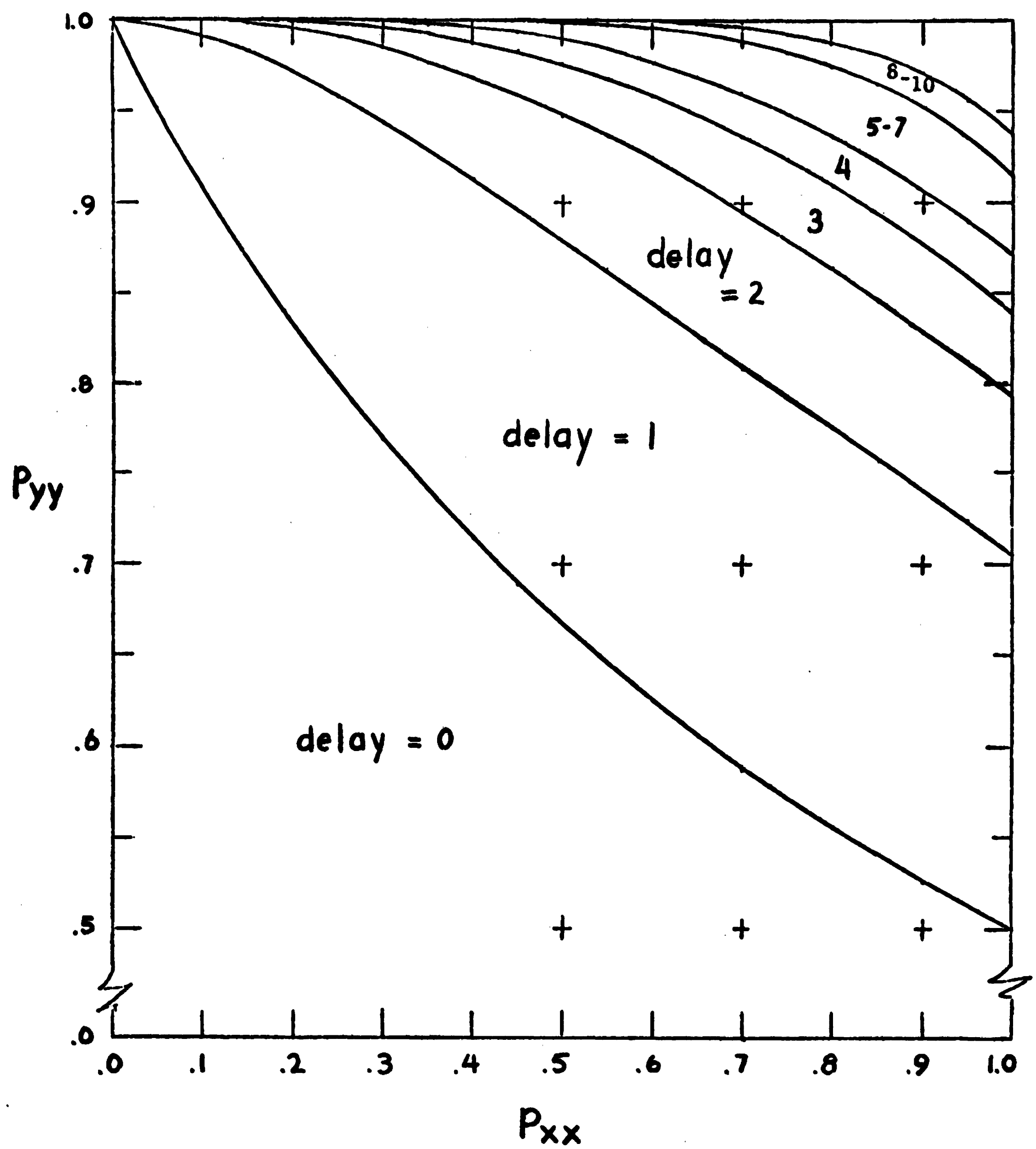

Figure 2. As autoregressive stability of the two variables increases, so does the theoretical expectation of delay between the causal lag and the measurement lag at which the xy cross-correlation is maximum. Thus, where p_{xx} and p_{yy} both = .9, a delay of 4 time units can be expected. (From [7], p. 41).

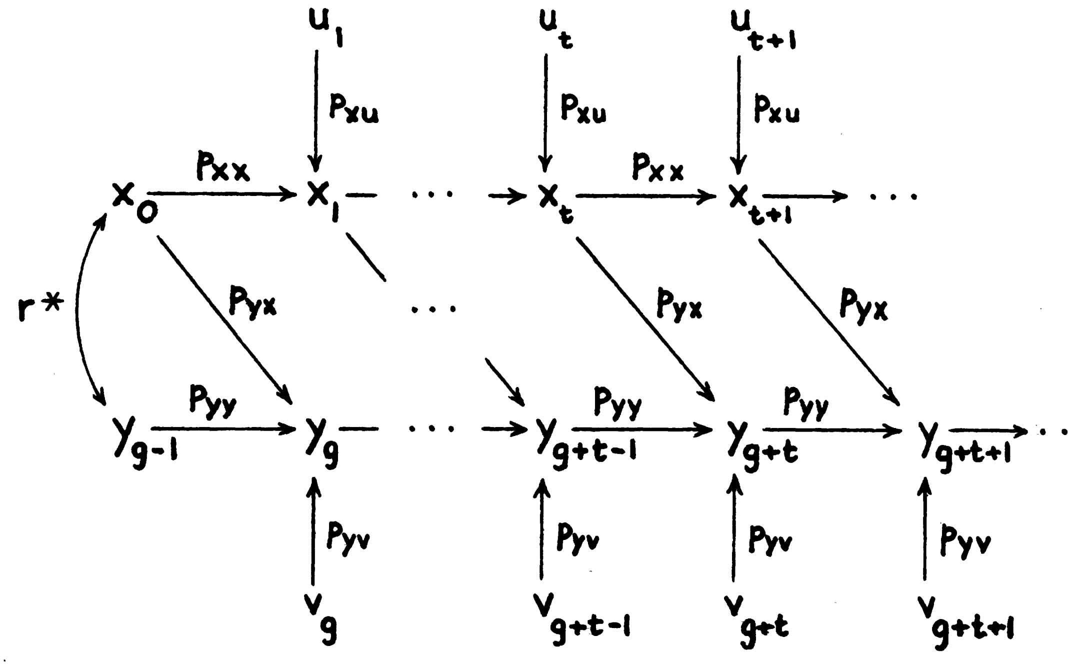

Figure 3. Path diagram of independent variable x and dependent variable y, governed by recursive relations (3) and (4) in technical appendix. Although y_g is plotted below x_1, g can take any value.

tions for the x and y variables are depéndent only on the time interval k over which the measurements are taken.

The values of these correlations are predicted from the path coefficients by the rules of path analysis. The computation of the cross-correlations $\rho(x_t, y_{t+k})$ is of particular interest. We find that these are governed not by a single formula but rather by two. One formula applies to cross-correlations for which measurement interval k is less than the causal interval g, and the other applies where k is greater than g, as illustrated in Figure 4. (The two expressions give the same correlation for k = g.) The derivation of the cross-correlation formulas is facilitated if one considers not the actual measurement interval k, but rather the amount by which k differs from g, defined as $k' = |k-g|$.

Consider first that part of the cross-correlation to the right of g (i.e., $k>g$). In the technical appendix it is shown that the expression for this part of the cross-correlation is given by appendix formula (7c):

$$r_{xy}(k') = p_{yx}p_{xx}r_{xx}(k'-1) + p_{yy}r_{xy}(k'-1)$$

Let us examine the two components after the equal sign in this expression. If the cross-correlation $r_{xy}(k')$ were determined only by the second component, it would be lower than the cross-correlation over the previous lag ($r_{xy}(k'-1)$), since, the latter term is multiplied by p_{yy} which is less than unity. However, the cross-correlation is increased at the same time by the first component, which depends on the size of the causal coefficient p_{yx} and also on the size of p_{xx}. When the first component adds more than the second component removes, the cross-correlation will <u>rise</u> from one time period to the next. Otherwise it will fall.

The technical appendix specifies the conditions (see expressions (16) and (17)) under which the cross-correlation will rise following g.

Let us return to that part of the cross-correlation to the left of g (i.e., $k<g$). Expression (13b) in the appendix shows that this portion is given by a relatively simple exponential function

$$r_{xy}(-k') = r_{yx}(k') = \frac{p_{yx}}{1 - p_{xx}p_{yy}} p_{xx}^{k'}$$

which rises steadily as measurement interval k approaches g.

<u>Relative influence of stability in x or y</u>

We saw in Figure 1 that as the autogreressive stability of both x and y increased, the xy cross-correlation became more asymmetrical around causal interval g. It is appropriate to ask whether this asymmetry is produced more by stability in x or by stability in y, or equally by both.

A partial answer was seen in Figure 2, showing the relative effect of the two autoregressive coefficients on the amount of delay in the maximum cross-correlation. The reader will observe that if variable y has only moderate stability (such as p_{yy} = .70), then even a very high stability in x will produce a delay no greater than 1 time unit. In contrast, if variable x has the same moderate

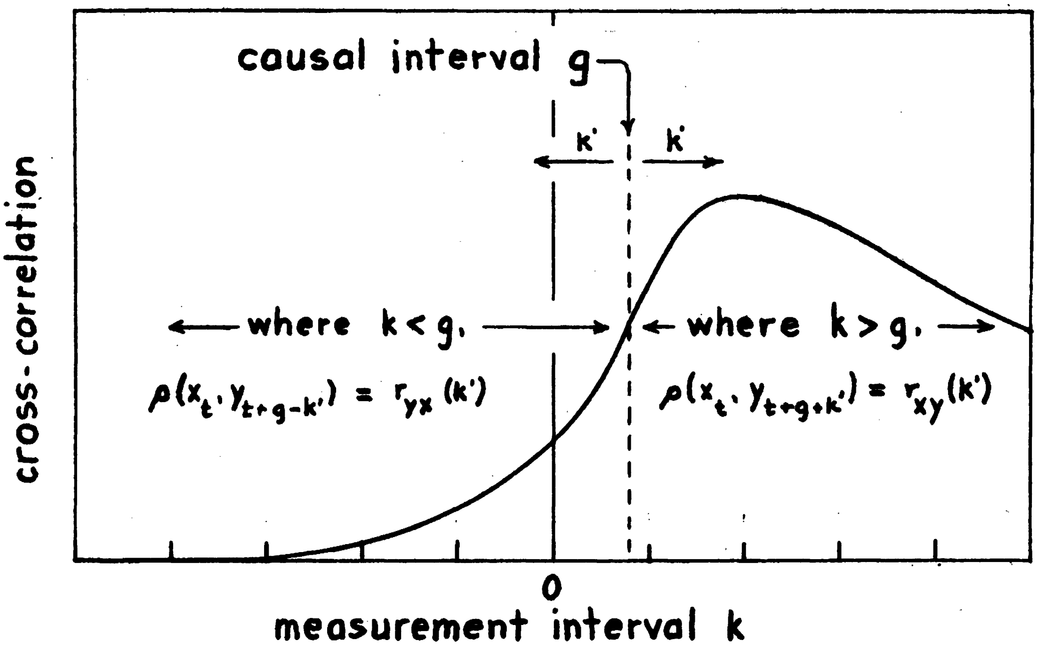

Figure 4. Formula for cross-correlation is composed of two parts, one where measurement interval $k <$ causal interval g, the other where $k > g$. Expressions for $r_{yx}(k')$ and $r_{xy}(k')$ are given in text and appendices.

stability ($p_{xx} = .70$), then increases in stability of y can produce delays up to 5 or more time units.

Hence it appears that the stability of dependent variable y is more critical than that of independent variable x, in producing asymmetry and delay in cross-correlations. (It is also true that for much longer delays of 8 or more time units, both x and y must be highly stable.)

A more complete answer can be given with the theoretical expressions derived in the appendix. These have been used to generate four cross-correlograms plotted in Figure 5. The bottom curve uses moderate stability in both variables ($p_{xx} = p_{yy} = .70$) and is very similar to the bottom curve in Figure 1 for simulated variables using the same parameters. The top curve, with very high stability in both variables ($p_{xx} = p_{yy} = .95$), is again similar to the top curve in Figure 1 with the same parameters.

The two middle curves show what happend when stability of the two variables is allowed to differ. When x is the more stable ($p_{xx} = .95$, $p_{yy} = .80$), the correlogram rises and falls gradually, with only slight asymmetry. When the two parameters are reversed and y is the more stable ($p_{xx} = .80$, $p_{yy} = .95$), the correlogram rises abruptly and is markedly non-symmetrical. Again it is clear that stability in the dependent variable is mainly responsible for the peculiar properties of asymmetry around g, and delay.

In conclusion

Simulated time series were created in which independent variable x was allowed to exert a causal influence on dependent variable y over causal interval g. Certain puzzling features of the simulated data are explained by application of a path analysis model. Specifically it is shown that when autoregressive stability of both variables is increased:

1. The cross-correlation between x and y becomes stronger[8] at the interval of causal influence, g.
2. The cross-correlogram becomes asymmetrical around g--that is, the correlations remain higher following g than preceding it, so that the cross-lagged differential persists over a longer measurement interval.
3. The point of maximum correlation between x and y is increasingly delayed beyond g. (Features 2 and 3 are found to result especially from stability in the dependent variable y.)

Finally, from the path equations, it is clear that when autoregressive stability is high (especially in the dependent variable):

4. The asymmetric effect of causal influence remains the same regardless of whether the causal interval is long or short; hence causal influence could be detected even when simultaneous.

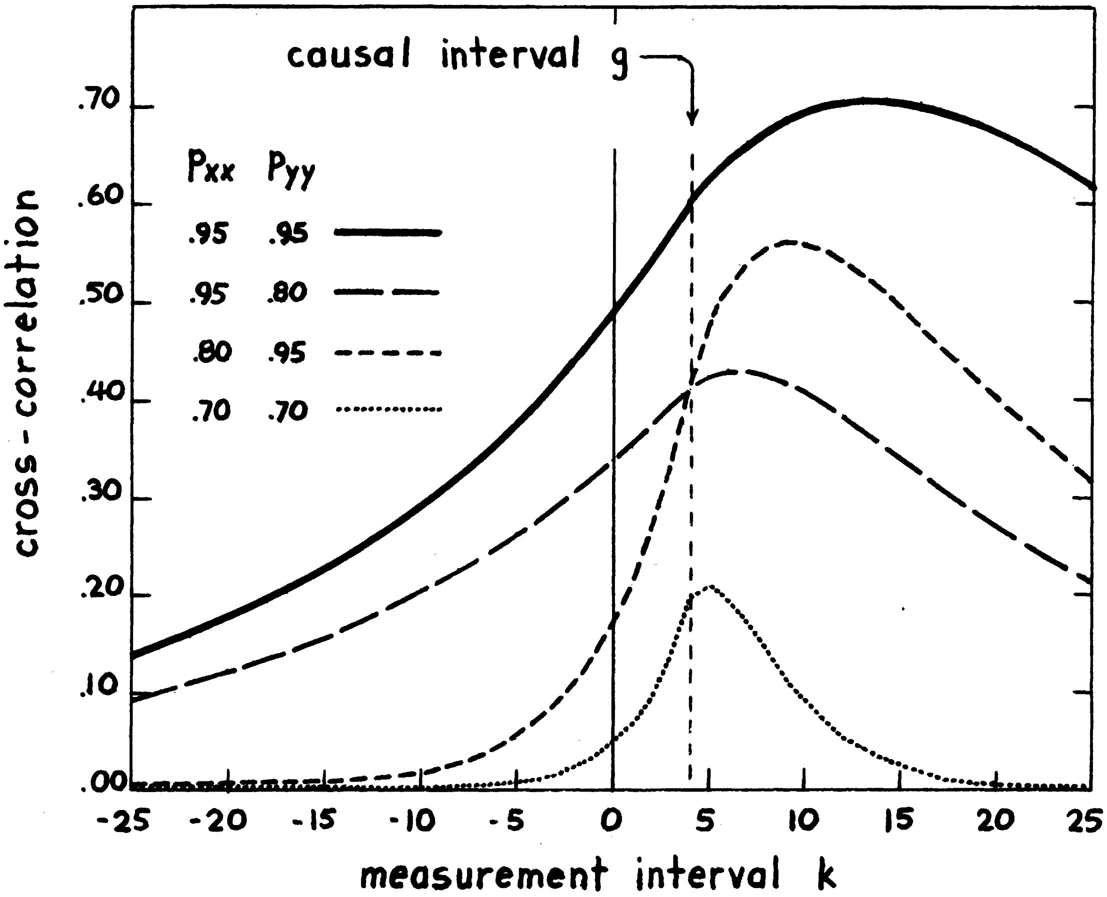

Figure 5. Theoretical cross-correlations based on path model. High stability in the dependent variable y (third curve) produces sharper asymmetry than does an equally high stability in the independent variable x (second curve). (The causal coefficients for the bottom three curves were set at p_{yx} = .10; for the top curve p_{yx} = .06.)

FOOTNOTES

[1] The initial simulation was performed under a grant from the National Science Foundation (GS-1873), with supplementary aid from the National Broadcasting Company. The original simulation program was written by Spyros Magliveras, and was revised by Robert A. Lew, assisted by George Gluski. Lew also derived the main mathematical properties of the model. The mathematical extensions reported here were supported by grants from the U.S. Office of Education (Grant No. OEG-5-9-239459-0076) and National Science Foundation (GS-2710). Omitted here is the technical appendix by Faith which was handed out at the ASA session. A copy of the full paper including appendix may be obtained by writing the senior author.

[2] Fruitful guidance was given by Graham Kalton, visiting lecturer in sociology and sampling statistics from the London School of Economics, in helping to structure the simulated model and determine some of its properties.

[3] Subsequently a more complex simulation program has been prepared which can generate up to 10 variables, each of which can exert a causal influence on any other, over any specified combination of causal intervals. This program was developed by Robert A. Lew with the assistance of George Gluski, primarily with support from the research department of the National Broadcasting Company.

[4] The u and v components are not the same as measurement error; unlike the latter they become incorporated in x_{it} and y_{it} respectively and hence enter into subsequent values. The two-variable simulation model does not allow for measurement error; all values are assumed to be perfectly reliable or "true scores." To maintain stationary variance in x_{it} over time, $|p_{xx}| \leq 1$, and variance of $u_{it} = (1-p_{xx}^2)\,\sigma_{x_{it}}^2$; corresponding limitations are imposed for p_{yy} and v_{it}.

[5] With many real social data it is inappropriate to assume a simple Markov process in which the value at time t depends only on the immediately prior value at t-1. Rather, each individual is likely to have an enduring tendency to persist at

the same level through time--equivalent to the effect of individual differences in intellectual ability, personality, or socio-economic background. The contrast between the "short-term stability" of the simple Markov process and "long-term stability" introduced by stable individual differences is demonstrated briefly in Pelz and Lew [6] and in more detail in [7].

[6] The autocorrelations were in fact higher for the dependent variable than predicted from the value of p^k_{yy}; the exact function was subsequently derived by Lew.

[7] The technical appendix is omitted here for lack of space. A copy may be obtained by writing the senior author.

[8] This property is governed equally by stability in either variable. The reader may observe in Figure 5 that the two middle curves cross at g. That is, a given pair of autoregressive coefficients assigned to either x or y will have the same effect on the size of cross-correlation at g.

REFERENCES

[1] Campbell, D.T. "From Description to Experimentation: Interpreting Trends as Quasi-experiments." Pp. 212-242 in C.W. Harris (ed.), Problems in Measuring Change. University of Wisconsin Press, Madison, 1963.

[2] Campbell, D.T. and Stanley, J.S. "Experimental and Quasi-experimental Designs for for Research on Teaching." Pp. 171-246 in N.G. Gage (ed.), Handbook of Research on Teaching. Rand McNally, Chicago, 1963.

[3] Goldberger, Arthur S. "Econometrics and Psychometrics: A Survey of Communalities." Paper EME 7013 (March, 1970), Social Systems Research Institute, University of Wisconsin, Madison. Also Psychometrica, 36 (June 1971).

[4] Kendall, M.G. and Stuart, A. The Advanced Theory of Statistics, Vol. 3, Hafner, New New York, 1966.

[5] Pelz, D.C. and Andrews, F.M. "Detecting Causal Priorities in Panel Study Data." American Sociological Review, 29 (1964), 836-848.

[6] Pelz, D.C. and Lew, R.A. "Heise's Model Applied." In Edgar F. Borgatta and George W. Bohrnstedt (eds.), Sociological Methedology--1970, 1970.

[7] Pelz, D.C., Magliveras, S., and Lew, R.A. "Correlational Properties of Simulated Panel Data with Causal Connections between Two Variables." Causal Analysis Project, Interim Report No. 1, 110 pp. Survey Research Center, University of Michigan, Ann Arbor, 1968.