Fred Emery

Causal Path Analysis[1]

"Causal path analysis" is the accepted designation for attempts to empirically identify a system of coupled elements. In social systems, we are seeking to identify producer-product relations. The logical requirements for inferring producer-product relations are

1. Correlations between two or more variables cannot, by themselves, establish that any of these variables are producers or co-producers of any others, i.e., that they constitute even necessary conditions—let alone sufficient conditions—for any other variable.
2. Postulating the state of one variable x as a necessary condition for the state of another variable y, which requires that two judgmental conditions are met:
   (a) That x is a *possible* producer of y. Thus we would not regard x as a possible producer of y if it occurred in time after y, if there is no conceivable path and if there is no conceivable mechanism, e.g., for an unassisted mouse to move a mountain.
   (b) That x is also a *probable* producer of y (i.e., observed concomitance); observed r_{xy} is greater than the r_{xy} possible by chance alone.
3. The conditions stated in requirement 2 are sufficient if it can *also* be established that they are true of x regardless whether any other variable z is also a probable producer of y.

To establish that x is a necessary condition for the state of y regardless of other variables, it is necessary to establish that the probability of x producing y is significantly greater than the probability of any other possible producer producing *both* x and y, i.e., that r_{xy} is significantly greater than $r_{zy} \times r_{zx}$. We are assuming that x has met the conditions specified in requirement 2 of possible and probable production. We are not requiring that x is a more probable producer than z (else we would never identify more than one necessary condition and would exclude the whole notion of co-producers).

If the conditions in requirement 2 are met, but not this condition with respect to x and y, then the observed correlation between x and y is a "spurious correlation."

[1] Excerpted from F.E. Emery and C. Phillips, *Living at Work.* Canberra: Australian Government Publishing Service, 1976.

The Problems of Analysis

Let us take as given a set of n variables for which there are indices of concomitance (e.g., correlations) for the states of all variables within the same individuals at the same moment of time (i.e., within a period of time within which the relation of the variables can be assumed not to have changed, e.g., that in doing a battery of tests growing fatigue has not changed performance in a test).

For this set of n variables, the number of possible direct, one-step, relations R between them is given by $(n - 1)$ factorial, or the following equation:

$$R = \frac{n\,(n - 1)}{2}$$

This number rises rapidly as n increases: $n = 2, R = 1$; $n = 6, R = 15$; $n = 12, R = 66$.

Theoretically, all these relations have to be examined to determine whether the conditions specified in requirement 2 are met, i.e., one has to determine of each observed relation whether one is a *potential* producer of the other, a *possible* producer, a *probable* producer. This has to be done for x with respect to y and for y with respect to x. The possibilities that have to be considered for n variables increase threefold.

The major problem is, therefore, one of sheer complexity, not a problem of logic.

Reduction of complexity is, in the first place, aided by setting levels of statistical significance for the correlation coefficients. Correlations below this are regarded as not probable as the observed relation could be due to chance factors alone.

The next step to reduce complexity has usually been to postulate specific causal models and test these for goodness of fit with the observed correlations. This is a subjective step and usually conservative in that the researcher will tend to postulate models that conform to current belief.

This model of "causal path analysis" is usually associated, in the social sciences, with the name of Blalock (1964). It is a carryover into the social sciences of the statistical method developed, decades before, by Sewell Wright (1934; 1960) to track the transmission of characteristics within domestic animal breeding stocks. The carrying over of this method ignored a basic assumption in Sewell Wright's method. He was justified in accepting the logically stronger form of regression equations (relative to correlation coefficients that average the regression of variable A on variable B and of B on A), because he always knew which of his animals were the parents and which the offspring.

He knew which was the independent variable and which the dependent variable. In the social sciences there is rarely such certainty. We cannot pass lightly over this blindness to basic assumptions because we, in the social sciences, have made the earlier error of uncritically adopting R.A. Fisher's (1935) statistical methods for analysis of variance in yields of small, contiguous agricultural plots to different levels and types of fertilizer treatment (cf. Chow [1992] on "the agricultural model of scientific experimentation"). Fisher could assume without doubt that the crop growth was the dependent variable and the fertilizer treatments the independent variable. He did not have to worry about whether it might be the fertilizer that grew or that the crop might act on whether it approved or took umbrage at the fertilizer treatment (Moghaddam and Harre, 1992; Silverman, 1977). Fisher knew beforehand that randomization of contiguous plots in a small area would overcome micro-climate and soil variations. He knew also that other possible sources of variation, e.g., wandering cattle, were definitely fenced off. It is a very lucky statistician in the social sciences who can be so sure about what is under control and what are the independent and the dependent variables. Social scientists have to confront genuine *multi*variate analysis; and the variables they conceptualize are concrete universals, not the abstract universals that mathematicians and formal logicians work with. Concrete universals have members that also have the characteristics of other concrete universals. Simply put, the concepts of the social sciences cannot be assigned, with assurance, to classes of independent, intervening or dependent variables.

A simple example might indicate the difficulties that statisticians face in the social sciences. In a study of smoking and stress (Emery et al., 1968), we were expecting to find patterns like the following:

social class → personal stress → smoking

We expected this because social class is so often the most independent, least easily changed, of all the variables. As it turned out, for young women (not for other sex and age groups) the pattern that emerged from the observed correlations was

personal stress → social class → smoking

What was indicated by this result was that social class, in this instance, was acting as an intervening cultural variable, not as an independent life conditions variable. That is, whether a young woman under stress smoked depended on whether she was in a culture that accepted this behavior or in a middle-class (English) culture that frowned on smoking by young women.

Using the regression-based model of causal paths, we would have tested the