the american journal of sociology

Volume 72 Number 1 July 1966

Path Analysis: Sociological Examples[1]

Otis Dudley Duncan

ABSTRACT

Linear causal models are conveniently developed by the method of path coefficients proposed by Sewall Wright. Path analysis is useful in making explicit the rationale of conventional regression calculations. It may also have special usefulness in sociology in problems involving the decomposition of a dependent variable or those in which successive experiences of a cohort are measured. Path analysis focuses on the problem of interpretation and does not purport to be a method for discovering causes. It may, nevertheless, be invaluable in rendering interpretations explicit, self-consistent, and susceptible to rejection by subsequent research.

The long-standing interest of sociologists in causal interpretation of statistical relationships has been quickened by discussions focusing on linear causal models. The basic work of bringing such models to the attention of the discipline was done by Blalock,[2] drawing upon the writings of Simon[3] and Wold[4] in particular. The rationale of this approach was strengthened when Costner and Leik[5] showed that "asymmetric causal models" of the kind proposed by Blalock afford a natural and operational explication of the notion of "axiomatic deductive theory," which had been developed primarily by sociologists working with verbal formulations. Most recently, Boudon[6] pointed out that the Simon-Blalock type of model is a "special case" or "weak form" of path analysis (or "dependence analysis," as Boudon prefers to call it). At the same time, he noted that "convincing empirical illustrations are missing," since "moderately complicated causal structures with corresponding data are rather scarce in the sociological literature." This paper presents some examples (in the form of reanalyses of published work) which may be interesting, if not "convincing." It includes an exposition of some aspects of path technique, developing it in a way that may make it a

[1] Prepared in connection with a project on "Socioeconomic Background and Occupational Achievement," supported by contract OE-5-85-072 with the U.S. Office of Education. Useful suggestions were made by H. M. Blalock, Jr., Beverly Duncan, Robert W. Hodge, Hal H. Winsborough, and Sewall Wright, but none of them is responsible for the use made of his suggestions or for any errors in the paper.

[2] Hubert M. Blalock, Jr., *Causal Inferences in Nonexperimental Research* (Chapel Hill: University of North Carolina Press, 1964).

[3] Herbert A. Simon, *Models of Man* (New York: John Wiley & Sons, 1957), chap. ii.

[4] Herman Wold and Lars Jureen, *Demand Analysis* (New York: John Wiley & Sons, 1953).

[5] Herbert L. Costner and Robert K. Leik, "Deductions from 'Axiomatic Theory,'" *American Sociological Review,* XXIX (December, 1964), 819–35.

[6] Raymond Boudon, "A Method of Linear Causal Analysis: Dependence Analysis," *American Sociological Review,* XXX (June, 1965), 365–74.

little more accessible than some of the previous writings.

Path coefficients were used by the geneticist Sewall Wright as early as 1918, and the technique was expounded formally by him in a series of articles dating from the early 1920's. References to this literature, along with useful restatements and illustrations, will be found in Wright's papers of 1934, 1954, and 1960.[7] The main application of path analysis has been in population genetics, where the method has proved to be a powerful aid to "axiomatic deductions." The assumptions are those of Mendelian inheritance, combined with path schemes representing specified systems of mating. The method allows the geneticist to ascertain the "coefficient of inbreeding," a quantity on which various statistical properties of a Mendelian population depend. It also yields a theoretical calculation of the genetic correlations among relatives of stated degrees of relationship. Most of Wright's expositions of this *direct* use of path coefficients are heavily mathematical;[8] an elementary treatment is given in the text by Li.[9]

Apart from a few examples in Wright's own work, little use has been made of path coefficients in connection with the *inverse* problem of estimating the paths which may account for a set of observed correlations on the assumption of a particular formal or causal ordering of the variables involved. Of greatest substantive interest to sociologists may be an example relating to heredity and environment in the determination of intelligence.[10] Another highly suggestive study was a pioneer but neglected exercise in econometrics concerning prices and production of corn and hogs.[11] Although the subject matter is remote from sociological concerns, examples from studies in animal biology are instructive on methodological grounds.[12] If research workers have been slow to follow Wright's lead, the statisticians have done little better. There are only a few expositions in the statistical literature,[13] some of which raise questions to which Wright has replied.[14]

PATH DIAGRAMS AND THE BASIC THEOREM

We are concerned with linear, additive, asymmetric relationships among a set of

[7] Sewall Wright, "The Method of Path Coefficients," *Annals of Mathematical Statistics,* V (September, 1934), 161–215; "The Interpretation of Multivariate Systems," in O. Kempthorne *et al.* (eds.), *Statistics and Mathematics in Biology* (Ames: Iowa State College Press, 1954), chap. ii; "Path Coefficients and Path Regressions: Alternative or Complementary Concepts?" *Biometrics,* XVI (June, 1960), 189–202.

[8] Sewall Wright, "The Genetical Structure of Populations," *Annals of Eugenics,* XV (March, 1951), 323–54.

[9] C. C. Li, *Population Genetics* (Chicago: University of Chicago Press, 1955), chap. xii–xiv. See also C. C. Li, "The Concept of Path Coefficient and Its Impact on Population Genetics," *Biometrics,* XII (June, 1956), 190–210.

[10] Sewall Wright, "Statistical Methods in Biology," *Journal of the American Statistical Association,* XXVI (March, 1931, suppl.), 155–63.

[11] Sewall Wright, *Corn and Hog Correlations,* U.S. Department of Agriculture Bulletin 1300 (Washington: Government Printing Office, 1925); also, "The Method of Path Coefficients," pp. 192–204.

[12] Sewall Wright, "The Genetics of Vital Characters of the Guinea Pig," *Journal of Cellular and Comparative Physiology,* LVI (suppl. 1, November, 1960), 123–51; F. A. Davidson *et al.*, "Factors Influencing the Upstream Migration of the Pink Salmon (*Oncorhynchus gorbuscha*)," *Ecology,* XXIV (April, 1943), 149–68.

[13] J. W. Tukey, "Causation, Regression and Path Analysis," in O. Kempthorne *et al., op. cit.,* chap. iii; Oscar Kempthorne, *An Introduction to Genetic Statistics* (New York: John Wiley & Sons, 1957), chap. xiv; Malcolm E. Turner and Charles D. Stevens, "The Regression Analysis of Causal Paths," *Biometrics,* XV (June, 1959), 236–58; Eleanor D. Campbell, Malcolm E. Turner, and Mary Frances Wright, with the editorial collaboration of Charles D. Stevens, *A Handbook of Path Regression Analysis,* Part I: *Estimators for Simple Completely Identified Systems* (Preliminary Ed.; Richmond: Medical College of Virginia, Department of Biophysics and Biometry, 1960); Henri Louis Le Roy, *Statistische Methoden der Populationsgenetik* (Basel: Birkhäuser, 1960), chap. i; P. A. P. Moran, "Path Coefficients Reconsidered," *Australian Journal of Statistics,* III (November, 1961), 87–93.

[14] "Paths Coefficients and Path Regressions."

variables which are conceived as being measurable on an interval scale, although some of them may not actually be measured or may even be purely hypothetical—for example, the "true" variables in measurement theory or the "factors" in factor analysis. In such a system, certain of the variables are represented to be dependent on others as linear functions. The remaining variables are assumed, for the analysis at hand, to be given. They may be correlated among themselves, but the explanation of their intercorrelation is not taken as problematical. Each "dependent" variable must be regarded explicitly as *completely* determined by some combination of variables in the system. In problems where complete determination by measured variables does not hold, a residual variable uncorrelated with other determining variables must be introduced.

Although it is not intrinsic to the method, the diagrammatic representation of such a system is of great value in thinking about its properties. A word of caution is necessary, however. Causal diagrams are appearing with increasing frequency in sociological publications. Most often, these have some kind of pictorial or mnemonic function without being isomorphic with the algebraic and statistical properties of the postulated system of variables—or, indeed, without having a counterpart in any clearly specified system of variables at all. Sometimes an investigator will post values of zero order or partial correlations, association coefficients, or other indications of the "strength" of relationship on such a diagram, without following any clearly defined and logically justified rules for entering such quantities into the analysis and its diagrammatic representation. In Blalock's work, by contrast, diagrams are employed in accordance with explicit rules for the representation of a system of equations. In general, however, he limits himself to the indication of the sign (positive or negative) of postulated or inferred direct relationships. In at least one instance[15] he inserts zero-order correlations into a diagram which looks very much like a causal diagram, although it is not intended to be such. This misleading practice should not be encouraged.

In path diagrams, we use one-way arrows leading from each determining variable to each variable dependent on it. Unanalyzed correlations between variables not dependent upon others in the system are shown by two-headed arrows, and the connecting line is drawn curved, rather than straight, to call attention to its distinction from the paths relating dependent to determining variables. The quantities entered on the diagram are symbolic or numerical values of *path coefficients,* or, in the case of the bidirectional correlations, the simple correlation coefficients.

Several of the properties of a path diagram are illustrated in Figure 1. The original data, in the form of ten zero-order correlations, are from Turner's study of determinants of aspirations.[16] The author does not provide a completely unequivocal formulation of the entire causal model shown here, but Figure 1 appears to correspond to the model that he quite tentatively proposes. At one point he states, "background affects ambition and ambition affects both IQ and class values; in addition . . . there is a lesser influence directly from background to class values, directly from background to IQ, and directly between IQ and class values."[17] Elsewhere,[18] he indicates that school rating operates in much the same fashion as (family) background. As for the relationship between the two, Turner notes, on the one hand, that "families may choose their place of residence," but also that "by introducing neighborhood, we may only be measuring family background more precisely."[19] Hence, it seems that there is no firm assumption about the causal ordering within

[15] Blalock, *op. cit.,* p. 77.

[16] Ralph H. Turner, *The Social Context of Ambition* (San Francisco: Chandler Publishing Co., 1964), pp. 49 and 52, Tables 11, 17, and 20.

[17] *Ibid.,* p. 107.

[18] *Ibid.,* pp. 54–61.

[19] *Ibid.,* p. 61.

this pair of variables; but since these two precede the remaining ones, it suffices to represent the link between X_1 and X_2 as merely a bidirectional correlation.

Allowing Turner to take responsibility for the causal ordering of the variables (assuming his statements are understood correctly) and deferring the question of how the path coefficients were estimated, let

$$X_3 = p_{32}X_2 + p_{31}X_1 + p_{3u}R_u ,$$
$$X_4 = p_{43}X_3 + p_{42}X_2 + p_{41}X_1 + p_{4v}R_v , \quad (1)$$
$$X_5 = p_{54}X_4 + p_{53}X_3 + p_{52}X_2 + p_{51}X_1 + p_{5w}R_w .$$

The use of the symbol p for the path coefficient is perhaps obvious. Note that the order of the subscripts is significant, the

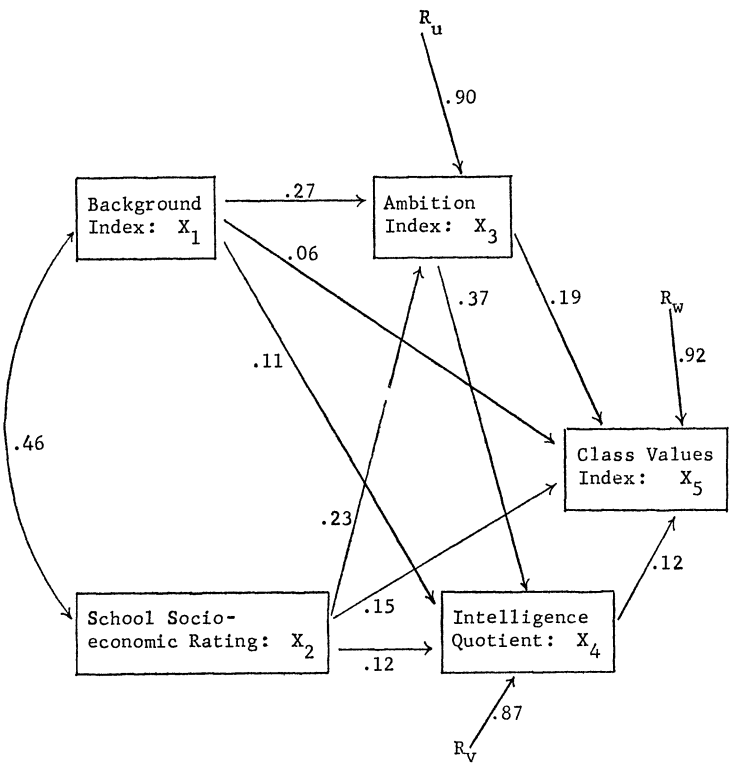

FIG. 1.—Causal model from Turner, *op. cit.*, with path coefficients estimated for male sample

us see what the system represented by Figure 1 is like. Each variable is taken to be in standard form; that is, if V_i is the ith variable as measured, then $X_i = (V_i - \bar{V}_i)/\sigma_{V_i}$. The same convention holds for the residuals, R_u, R_v, and R_w, to which a literal subscript is attached to indicate that these variables are not directly measured. The system represented in Figure 1 can now be written:

convention being the same as that used for regression coefficients: the first subscript identifies the dependent variable, the second the variable whose direct effect on the dependent variable is measured by the path coefficient. The order of subscripts is immaterial for correlations. But note that while $r_{42} = r_{24}$ and $r_{42.123} = r_{24.123}$, $p_{42} \neq p_{24}$; indeed p_{42} and p_{24} would never appear in the same system, given the restriction to

recursive systems mentioned subsequently. Contrary to the practice in the case of partial regression and correlation coefficients, symbols for paths carry no secondary subscripts to identify the other variables assumed to affect the dependent variable. These will ordinarily be evident from the diagram or the equation system.

In one respect, the equation system (1) is less explicit than the diagram because the latter indicates what assumptions are made about residual factors. Each such factor is assumed by definition to be uncorrelated with any of the immediate determinants of the dependent variable to which it pertains. In Figure 1, the residuals are also uncorrelated with each other, as in the Simon-Blalock development.[20] We shall see later, however, that there are uses for models in which some residuals are intercorrelated, or in which a residual is correlated with variables antecedent to, but not immediate determinants of, the particular dependent variable to which it is attached. Where the assumption of uncorrelated residuals is made, deductions reached by the Simon-Blalock technique of expanding the product of two error variables agree with the results obtained by the formulas mentioned below, although path analysis involves relatively little use of the partial correlations which are a feature of their technique.

Equation system (1), as Blalock points out, is a recursive system. This discussion explicitly excludes non-recursive systems, involving instantaneous reciprocal action of variables, although Wright has indicated ways of handling them in a path framework.[21] Thus we shall not consider diagrams showing a direct or indirect feedback loop.

The principle that follows from equations in the form of (1) is that the correlation between any pair of variables can be written in terms of the paths leading from common antecedent variables. Consider r_{35}. Since $X_3 = (V_3 - \bar{V}_3)/\sigma_3$ and $X_5 = (V_5 - \bar{V}_5)/\sigma_5$ we have $r_{35} = \Sigma(V_3 - \bar{V}_3) \times (V_5 - \bar{V}_5)/N\sigma_3\sigma_5 = \Sigma X_3 X_5/N$.

We may expand this expression in either of two ways by substituting from (1) the expression for X_3 or the one for X_5. It is more convenient to expand the variable which appears later in the causal sequence:

$$r_{35} = \Sigma X_3 X_5/N = \frac{1}{N}\Sigma X_3(p_{54}X_4 + p_{53}X_3 + p_{52}X_2 + p_{51}X_1 + p_{5w}R_w) \quad (2)$$
$$= p_{54}r_{34} + p_{53} + p_{52}r_{23} + p_{51}r_{13},$$

making use of the fact that $\Sigma X_3 X_3/N = 1$ and the assumption that $r_{3w} = 0$, since X_3 is a factor of X_5. But the correlations on the right-hand side of (2) can be further analyzed by the same procedure; for example,

$$r_{34} = \frac{1}{N}\Sigma X_3 X_4 = \frac{1}{N}\Sigma X_3(p_{43}X_3 + p_{42}X_2 + p_{41}X_1 + p_{4v}R_v) \quad (3)$$
$$= p_{43} + p_{42}r_{23} + p_{41}r_{13},$$

and

$$r_{32} = \frac{1}{N}\Sigma X_2 X_3 = \frac{1}{N}\Sigma X_2(p_{32}X_2 + p_{31}X_1 + p_{3u}R_u) \quad (4)$$
$$= p_{32} + p_{31}r_{12}.$$

Note that r_{12}, assumed as a datum, cannot be further analyzed as long as we retain the particular diagram of Figure 1.

These manipulations illustrate the basic theorem of path analysis, which may be written in the general form:

$$r_{ij} = \sum_q p_{iq}r_{jq}, \quad (5)$$

where i and j denote two variables in the system and the index q runs over all vari-

[20] Blalock, *op. cit.*, p. 64; Boudon, *op. cit.*, p. 369.

[21] Sewall Wright, "The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis," *Biometrics,* XVI (September, 1960), 423–45.

ables from which paths lead directly to X_i. Alternatively, we may expand (5) by successive applications of the formula itself to the r_{jq}. Thus, from (2), (3), (4), and a similar expansion of r_{13}, we obtain

$$\begin{aligned} r_{53} = {} & p_{53} + p_{51}p_{31} + p_{51}r_{12}p_{32} + p_{52}p_{32} \\ & + p_{52}r_{12}p_{31} + p_{54}p_{42}p_{32} \\ & + p_{54}p_{42}r_{12}p_{31} + p_{54}p_{43} \\ & + p_{54}p_{41}p_{32}r_{12} + p_{54}p_{41}p_{31} . \end{aligned} \qquad (6)$$

Such expressions can be read directly from the diagram according to the following rule. Read *back* from variable *i, then forward* to variable *j*, forming the product of all paths along the traverse; then sum these products for all possible traverses. The same variable cannot be intersected more than once in a single traverse. In no case can one trace back having once started forward. The bidirectional correlation is used in tracing either forward or back, but if more than one bidirectional correlation appears in the diagram, only one can be used in a single traverse. The resulting expression, such as (6), may consist of a single direct path plus the sum of several compound paths representing all the indirect connections allowed by the diagram. The general formula (5) is likely to be the more useful in algebraic manipulation and calculation, the expansion on the pattern of (6) in appreciating the properties of the causal scheme. It is safer to depend on the algebra than on the verbal algorithm, at least until one has mastered the art of reading path diagrams.

An important special case of (5) is the formula for complete determination of X_i, obtained by setting $i = j$:

$$r_{ii} = 1 = \sum_q p_{iq}r_{iq} , \qquad (7)$$

or, upon expansion,

$$r_{ii} = \sum_q p_{iq}^2 + 2\sum_{q,q'} p_{iq}r_{qq'}p_{iq'} , \qquad (8)$$

where the range of q and q' $(q' > q)$ includes all variables, measured and unmeasured. A major use for (8) is the calculation of the residual path. Thus we obtain p_{3u} in the system (1) from

$$p_{3u}^2 = 1 - p_{32}^2 - p_{31}^2 - 2p_{32}r_{12}p_{31} . \qquad (9)$$

The causal model shown in Figure 1 represents a special case of path analysis: one in which there are no unmeasured variables (other than residual factors), the residuals are uncorrelated, and each of the dependent variables is directly related to all the variables preceding it in the assumed causal sequence. In this case, path analysis amounts to a sequence of conventional regression analyses, and the basic theorem (5) becomes merely a compact statement of the normal equations of regression theory for variables in standard form. The path coefficients are then nothing other than the "beta coefficients" in a regression setup, and the usual apparatus for regression calculations may be employed.[22] Thus, the paths in Figure 1 are obtained from the regression of X_3 on X_2 and X_1, setting $p_{32} = \beta_{32.1}$ and $p_{31} = \beta_{31.2}$; the regression of X_4 on X_3, X_2, and X_1, setting $p_{43} = \beta_{43.12}$, $p_{42} = \beta_{42.13}$, and $p_{41} = \beta_{41.23}$; and the regression of X_5 on the other four variables, setting $p_{54} = \beta_{54.123}$, $p_{53} = \beta_{53.124}$, and so on. Following the computing routine which inverts the matrix of intercorrelations of the independent variables, one obtains automatically the standard errors of the β coefficients (or b^*-coefficients, in the notation of Walker and Lev). In the present problem, with sample size exceeding 1,000, the standard errors are small, varying between .027 and .032. All the β's are at least twice their standard errors and thus statistically significant.

In problems of this kind, Blalock[23] has been preoccupied with the question of whether one or more path coefficients may be deleted without loss of information. As compared with his rather tedious search

[22] Helen M. Walker and Joseph Lev, *Statistical Inference* (New York: Holt, Rinehart & Winston, 1953), chap. xiii.

[23] *Op. cit.*, chap. iii.

procedure, the procedure followed here seems more straightforward. Had some of the β's turned out both non-significant and negligible in magnitude, one could have erased the corresponding paths from the diagram and run the regressions over, retaining only those independent variables found to be statistically and substantively significant.

As statistical techniques, therefore, neither path analysis nor the Blalock-Simon procedure adds anything to conventional regression analysis as applied recursively to generate a system of equations, rather than a single equation. As a *pattern of interpretation,* however, path analysis is invaluable in making explicit the rationale for a set of regression calculations. One may not be wholly satisfied, for example, with the theoretical assumptions underlying the causal interpretation of Turner's data provided by Figure 1, and perhaps Turner himself would not be prepared to defend it in detail. The point is, however, that *any* causal interpretation of these data must rest on assumptions—at a minimum, the assumption as to ordering of the variables, but also assumptions about the unmeasured variables here represented as uncorrelated residual factors.[24] The great merit of the path scheme, then, is that it makes the assumptions explicit and tends to force the discussion to be at least internally consistent, so that mutually incompatible assumptions are not introduced surreptitiously into different parts of an argument extending over scores of pages. With the causal scheme made explicit, moreover, it is in a form that enables criticism to be sharply focused and hence potentially relevant not only to the interpretation at hand but also, perchance, to the conduct of future inquiry.

Another useful contribution of path analysis, even in the conventional regression framework, is that it provides a calculus for indirect effects, when the basic equations are expanded along the lines of (6). It is evident from the regression coefficients, for example, that the direct effect of school on class values is greater than that of background, but the opposite is true of the indirect effects. The pattern of indirect effects is hardly obvious without the aid of an explicit representation of the causal scheme. If one wishes a single summary measure of indirect effect, however, it is obtained as follows: indirect effect of X_2 on $X_5 = r_{52} - p_{52} = .28 - .15 = .13$; similarly, indirect effect of X_1 is $r_{51} - p_{51} = .24 - .06 = .18$. These summations of indirect effects include, in each case, the effects of one variable via its correlation with the other; hence the two are not additive. Without commenting further on the substantive implications of the direct and indirect effects suggested by Turner's material, it may simply be noted that the investigator will usually want to scrutinize them carefully in terms of his theory.

DECOMPOSITION OF A DEPENDENT VARIABLE

Many of the variables studied in social research are (or may be regarded as) composite. Thus, population growth is the sum of natural increase and net migration; each of the latter may be further decomposed, natural increase being births minus deaths and net migration the difference between in- and out-migration. Where such a decomposition is available, it is of interest (1) to compute the relative contributions of the components to variation in the composite variable and (2) to ascertain how causes affecting the composite variable are transmitted via the respective components.

An example taken from work of Winsborough[25] illustrates the case of a variable with multiplicative components, rendered additive by taking logarithms. Studying variation in population density over the seventy-four community areas (omitting the central business district) of Chicago in 1940, Winsborough noted that density, de-

[24] *Ibid.,* pp. 46–47.

[25] Hal H. Winsborough, "City Growth and City Structure," *Journal of Regional Science,* IV (Winter, 1962), 35–49.

fined as the ratio of population to area, can be written:

$$\frac{\text{Population}}{\text{Area}} = \frac{\text{Population}}{\text{Dwelling Units}} \times \frac{\text{Dwelling Units}}{\text{Structures}} \times \frac{\text{Structures}}{\text{Area}}.$$

Let $V_0 = \log$ (Population/Area), $V_1 = \log$ (Population/Dwelling Units), $V_2 = \log$ (Dwelling Units/Structures), and $V_3 = \log$ (Structures/Area); then

$$V_0 = V_1 + V_2 + V_3 .$$

If each variable is expressed in standard form, we obtain,

$$\frac{V_0 - \bar{V}_0}{\sigma_0} = \frac{V_1 - \bar{V}_1}{\sigma_1} \cdot \frac{\sigma_1}{\sigma_0} + \frac{V_2 - \bar{V}_2}{\sigma_2} \cdot \frac{\sigma_2}{\sigma_0} + \frac{V_3 - \bar{V}_3}{\sigma_3} \cdot \frac{\sigma_3}{\sigma_0},$$

or

$$X_0 = p_{01}X_1 + p_{02}X_2 + p_{03}X_3 ,$$

where $X_0, \ldots, X_3$ are the variables in standard form and p_{01}, p_{02}, p_{03} are the path coefficients involved in the determination of X_0 by X_1, X_2, and X_3. Observe that the path coefficients can be computed in this kind of problem, where complete determination by measured variables holds as a consequence of definitions, without prior calculation of correlations:[26]

$p_{01} = \sigma_1/\sigma_0 = .132$ $\sigma_0 = .491$ $\sigma_1 = .065$

$p_{02} = \sigma_2/\sigma_0 = .468$ $\sigma_2 = .230$

$p_{03} = \sigma_3/\sigma_0 = .821$ $\sigma_3 = .403.$

The intercorrelations of the components, shown in Table 1, are used to complete the diagram, Figure 2, *a*. The correlations of the dependent variable with its components may now be computed from the basic theorem, equation (5).

$r_{01} = p_{01} + p_{02}r_{12} + p_{03}r_{13} = -.419;$

$r_{02} = p_{01}r_{12} + p_{02} + p_{03}r_{23} = .636;$ and

$r_{03} = p_{01}r_{13} + p_{02}r_{23} + p_{03} = .923.$

The analysis has not only turned up a clear ordering of the three components in terms of relative importance, as given by the path coefficients, it has also shown that one of the components is actually correlated negatively with the composite variable, owing to its negative correlations with the other two components.

TABLE 1

CORRELATION MATRIX FOR LOGARITHMS OF DENSITY AND ITS COMPONENTS AND TWO INDEPENDENT VARIABLES: CHICAGO COMMUNITY AREAS, 1940

Variable	X_1	X_2	X_3	W	Z
X_0 density (log)	−.419	.636	.923	−.663	−.390
X_1 persons per dwelling unit (log)	.....	−.625	−.315	.296	.099
X_2 dwelling units per structure (log)	.....	.....	.305	−.594	−.466
X_3 structures per acre (log)	.....	.....	.....	−.517	−.226
W distance from center	.....	.....	.....	.....	.549
Z recency of growth	.....	.....	.....	.....	.....

Source: Winsborough, *op. cit.*, and unpublished data kindly supplied by the author.

Winsborough considered two independent variables as factors producing variation in density: distance from the city center and recency of growth (percentage of dwelling units built in 1920 or later). The diagram can be elaborated to indicate how these factors operate via the components of log density (see Fig. 2, *b*).

The first step is to compute the path coefficients for the relationships of each component to the two independent variables. (The requisite information is given in Table 1.) For example, the equations,

$$r_{1W} = p_{1W} + p_{1Z}r_{ZW} ,$$

$$r_{1Z} = p_{1W}r_{ZW} + p_{1Z} ,$$

[26] Based on data kindly supplied by Winsborough.

may be used to solve for p_{1Z} and p_{1W}. (This is, of course, equivalent to computing the multiple regression of X_1 on W and Z, with all variables in standard form.) Substantively, it is interesting that distance, W, has somewhat larger effects on each component of density than does recency of growth, Z, while the pattern of signs of the path coefficients is different for W and Z.

The two independent variables by no means account for all the variation in any of the components, as may be seen from the size of the residuals, p_{1a}, p_{2b}, and p_{2c}, these being computed from the formula (7) for complete determination. It is possible, nevertheless, for the independent variables to account for the intercorrelations of the components and, ideally, one would like to

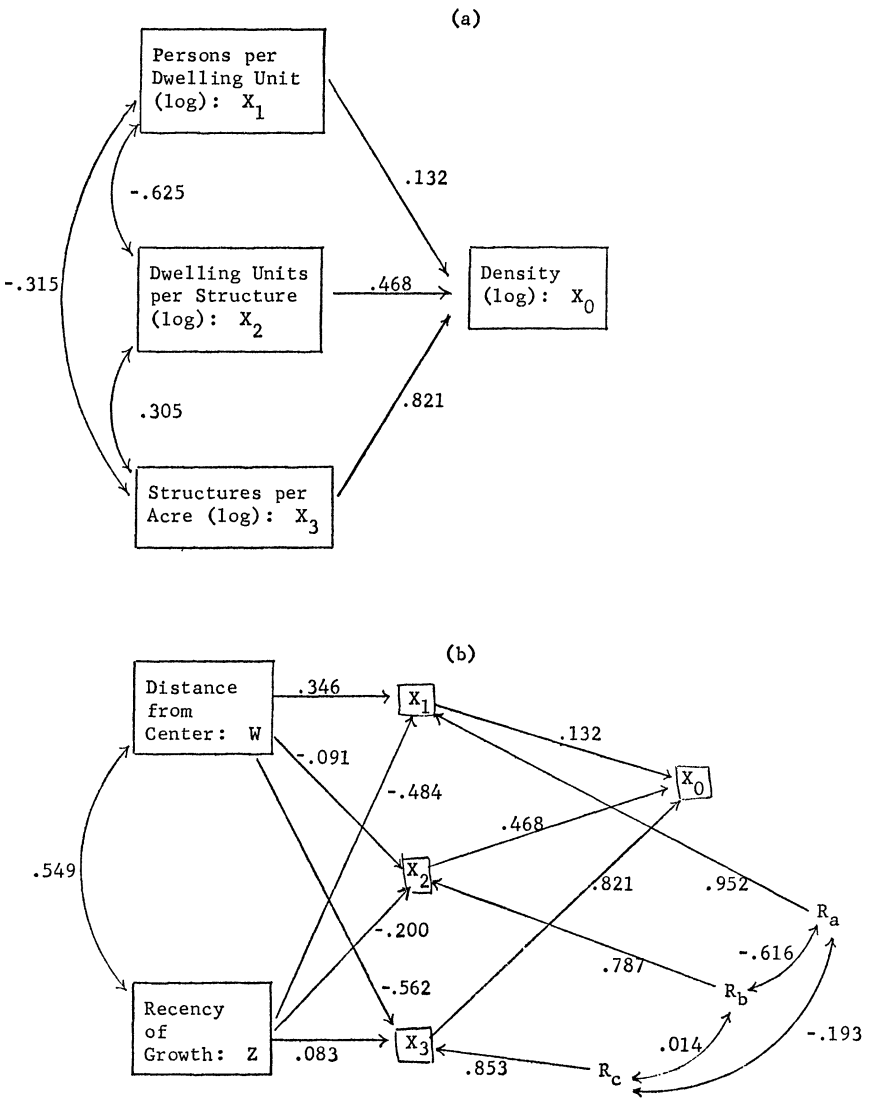

FIG. 2.—*a*, decomposition of log density (X_0) into components; *b*, effects of distance and recency of growth on log density via components. (Source: Winsborough, *op. cit.*, and unpublished calculations kindly supplied by the author.)

discover independent variables which would do just that. The relevant calculations concern the correlations between residuals. These are obtained from the basic theorem, equation (5), by writing, for example,

$$r_{23} = p_{2W}r_{3W} + p_{2Z}r_{3Z} + p_{2b}p_{3c}r_{bc},$$

which may be solved for $r_{bc} = .014$. In this setup, the correlations between residuals are merely the conventional second-order partial correlations; thus $r_{ab} = r_{12.WZ}$, $r_{ac} = r_{13.WZ}$, and $r_{bc} = r_{23.WZ}$. Partial correlations, which otherwise have little utility in path analysis, turn out to be appropriate when the question at issue is whether a set of independent variables "explains" the correlation between two dependent variables. In the present example, while $r_{23} = .305$, we find $r_{bc} = r_{23.WZ} = .014$. Thus the correlation between the logarithms of dwelling units per structure (X_2) and structures per acre (X_3) is satisfactorily explained by the respective relationships of these two components to distance and recency of growth. The same is not true of the correlations involving persons per dwelling unit (X_1), but fortunately this is by far the least important component of density.

Although the correlations between residuals are required to complete the diagram and, in a sense, to evaluate the adequacy of the explanatory variables, they do not enter as such into the calculations bearing upon the final question: How are the effects of the independent variables transmitted to the dependent variable via its components? The most compact answer to this question is given by the equations,

$$r_{0W} = p_{01}r_{1W} + p_{02}r_{2W} + p_{03}r_{3W}$$
$$= .039 - .278 - .424 = -.663,$$

and

$$r_{0Z} = p_{01}r_{1Z} + p_{02}r_{2Z} + p_{03}r_{3Z}$$
$$= .013 - .218 - .185 = -.391.$$

Density is negatively related to both distance and recency of growth, but the effects transmitted via the first component of density are positive (albeit quite small). Distance diminishes density primarily via its intermediate effect on structures per acre (X_3), secondarily via dwelling units per structure (X_2). The comparison is reversed for recency of growth, the less important of the two factors. More detailed interpretations can be obtained, as explained earlier, by expanding the correlations r_{1W}, r_{2W}, etc., using the basic theorem (5). For further substantive interpretation, the reader is referred to the source publication, which also offers an alternative derivation of the compound paths.

The density problem may well exemplify a general strategy too seldom employed in research: breaking a complex variable down into its components before initiating a search for its causes. One egregious error must, however, be avoided: that of treating components and causes on the same footing. By this route, one can arrive at the meaningless result that net migration is a more important "cause" of population growth than is change in manufacturing output. One must take strong exception to a causal scheme constructed on the premise, "If both demographic and economic variables help explain metropolitan growth, then we may gain understanding of growth processes by lumping the two together."[27] On the contrary, "understanding" would seem to require a clear distinction between demographic *components* of growth and economic *causes* which may affect growth via one or another of its components.

A CHAIN MODEL

Data reported by Hodge, Siegel, and Rossi[28] seem to fit well the model of a *simple causal chain* (see Fig. 3, *a*). These

[27] George L. Wilber, "Growth of Metropolitan Areas in the South," *Social Forces*, XLII (May, 1964), 491.

[28] Robert W. Hodge, Paul M. Siegel, and Peter H. Rossi, "Occupational Prestige in the United States, 1925–1963," *American Journal of Sociology*, LXX (November, 1964), 286–302.

authors give correlations between the occupational prestige ratings of four studies completed at widely separated dates: Counts (1925), Smith (1940), National Opinion Research Center (1947), and NORC replication (1963). In a simple causal chain, the correlations between temporally adjacent variables are the path coefficients (this is an immediate consequence of the definition of path coefficient). Using these three correlations as reported by Hodge *et al.*, we may infer that the correlation between NORC (1963) and Smith is (.990)(.982) = .972; between NORC (1963) and Counts is (.990)(.982)(.968) = .942; and between NORC (1947) and Counts is (.982)(.968) = .951. The observed values of these correlations (with differences from the inferred values in parentheses) are $r_{YS} = .971$ (—.001), $r_{YC} = .934$ (—.008), and $r_{XC} = .955$ (.004). Acceptance of this causal chain model is consistent with the conclusion of Hodge *et al.* that the amount of change in the relative positions of occupations in a prestige hierarchy is a direct function of elapsed time.

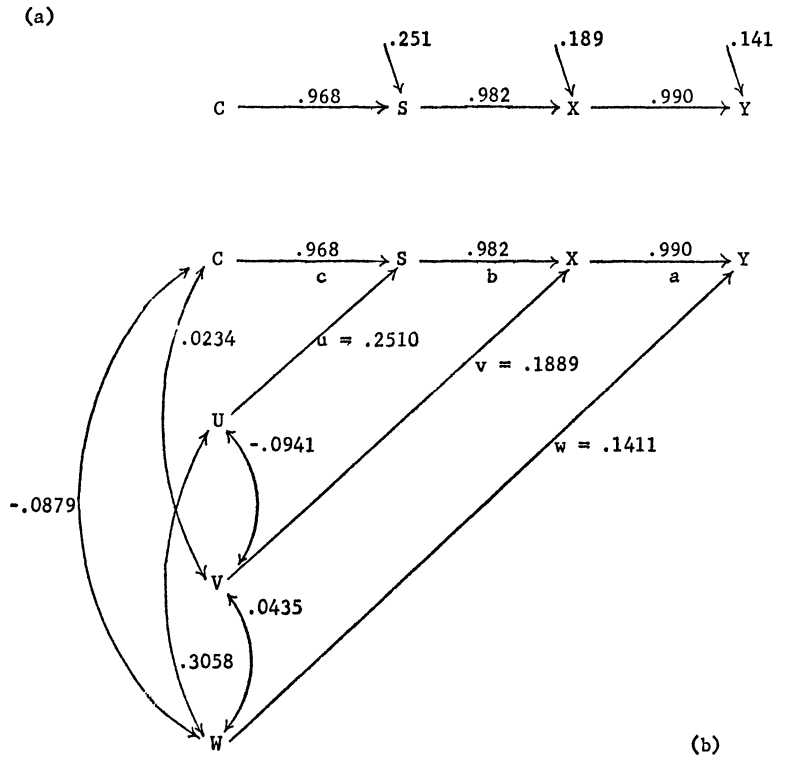

FIG. 3.—Causal chain: *a*, correlations taken from Hodge *et al.*, *op. cit.* (*C* = Counts, 1925; *S* = Smith, 1940; *X* = NORC, 1947; *Y* = NORC, 1963); *b*, intercorrelations of residuals implied by acceptance of chain hypothesis for the data in *a*.

Although the discrepancies between inferred and observed correlations seem trivial, it is worth noting that acceptance of the estimates shown in Figure 3, *a*, along with the assumption of a simple causal chain, requires us to postulate a complex pattern of correlations (most of them negligible in size) among the residuals or errors. This pattern is shown in Figure 3, *b*. In obtaining this solution, we assume that

each residual is uncorrelated with the immediately preceding variable in the chain but not necessarily with variables two or more links behind it. In the present example, then, the crucial assumptions are that $r_{VS} = r_{WX} = 0$. We can then, using equation (5) or the verbal algorithm, write the number of equations required to solve for the quantities to be entered on the diagram (for convenience, lower-case letters designate paths):

$$
\begin{aligned}
&r_{YX} = a = .990\,,\\
&r_{XS} = b = .982\,,\\
&r_{SC} = c = .968\,,\\
&r_{YY} = 1 = a^2 + w^2\,,\\
&r_{XX} = 1 = b^2 + v^2\,,\\
&r_{SS} = 1 = c^2 + u^2\,,\\
&r_{XC} = .955 = bc + vr_{VC}\,,\\
&r_{YC} = .934 = abc + avr_{VC} + wr_{CW}\,,\\
&r_{YS} = .971 = ab + cwr_{CW} + uwr_{UW}\,,\\
&r_{VS} = 0 = ur_{UV} + cr_{CV}\,,\\
&r_{WX} = 0 = vr_{VW} + br_{SW}\\
&\qquad (\text{where } r_{SW} = cr_{CW} + ur_{UW}).
\end{aligned}
\tag{11}
$$

In general, if we are considering a k-variable causal chain, we shall have to estimate $k - 1$ residual paths, $(k - 1)(k - 2)/2$ correlations between residuals, $k - 1$ paths for the links in the chain, and $k - 2$ correlations between the initial variable and residuals 2, 3, . . . , k in the chain. This is a total of $(k^2 + 3k - 6)/2$ quantities to be estimated. We shall have at our disposal $k(k - 1)/2$ equations expressing known correlations in terms of paths, $k - 1$ equations of complete determination (for all variables in the chain except the initial one), and $k - 2$ equations in which the correlation of a residual with the immediately preceding variable in the chain is set equal to zero. This amounts to $(k^2 + 3k - 6)/2$ equations, exactly the number required for a solution. The solution may, of course, include meaningless results (e.g., $r > 1.0$), or results that strain one's credulity. In this event, the chain hypothesis had best be abandoned or the estimated paths modified.

In the present illustration, the results are plausible enough. Both the Counts and the Smith studies differed from the two NORC studies and from each other in their techniques of rating and sampling. A further complication is that the studies used different lists of occupations, and the observed correlations are based on differing numbers of occupations. There is ample opportunity, therefore, for correlations of errors to turn up in a variety of patterns, even though the chain hypothesis may be basically sound. We should observe, too, that the residual factors here include not only extrinsic disturbances but also real though temporary fluctuations in prestige, if there be such.

What should one say, substantively, on the basis of such an analysis of the prestige ratings? Certainly, the temporal ordering of the variables is unambiguous. But whether one wants to assert that an aspect of social structure (prestige hierarchy) at one date "causes" its counterpart at a later date is perhaps questionable. The data suggest there is a high order of persistence over time, coupled with a detectable, if rather glacial, drift in the structure. The calculation of numerical values for the model hardly resolves the question of ultimate "reasons" for either the pattern of persistence or the tempo of change. These are, instead, questions raised by the model in a clear way for further discussion and, perhaps, investigation.

THE SYNTHETIC COHORT AS A PATTERN OF INTERPRETATION

Although, as the example from Turner indicates, it is often difficult in sociological analysis to find unequivocal bases for causal ordering, there is one happy exception to this awkward state of affairs. In the life

cycles of individuals and families, certain events and decisions commonly if not universally precede others. Despite the well-known fallibility of retrospective data, the investigator is not at the mercy of respondents' recall in deciding to accept the completion of schooling as an antecedent to the pursuit of an occupational career (exceptions granted) or in assuming that marriage precedes divorce. Some observations, moreover, may be made and recorded in temporal sequence, so that the status observed at the termination of a period of observation may logically be taken to depend on the initial status (among other things). Path analysis may well prove to be most useful to sociologists studying actual historical processes from records and reports of the experience of real cohorts whose experiences are traced over time, such as a student population followed by the investigator through the first stages of post-graduation achievement.[29]

The final example, however, concerns not real cohorts but the usefulness of a hypothetical synthesis of data from several cohorts. As demographers have learned, synthetic cohort analysis incurs some specific hazards;[30] yet the technique has proved

[29] For example, Bruce K. Eckland, "Academic Ability, Higher Education, and Occupational Mobility," *American Sociological Review*, XXX (October, 1965), 735–46.

[30] P. K. Whelpton, "Reproduction Rates Adjusted for Age, Parity, Fecundity, and Marriage,"

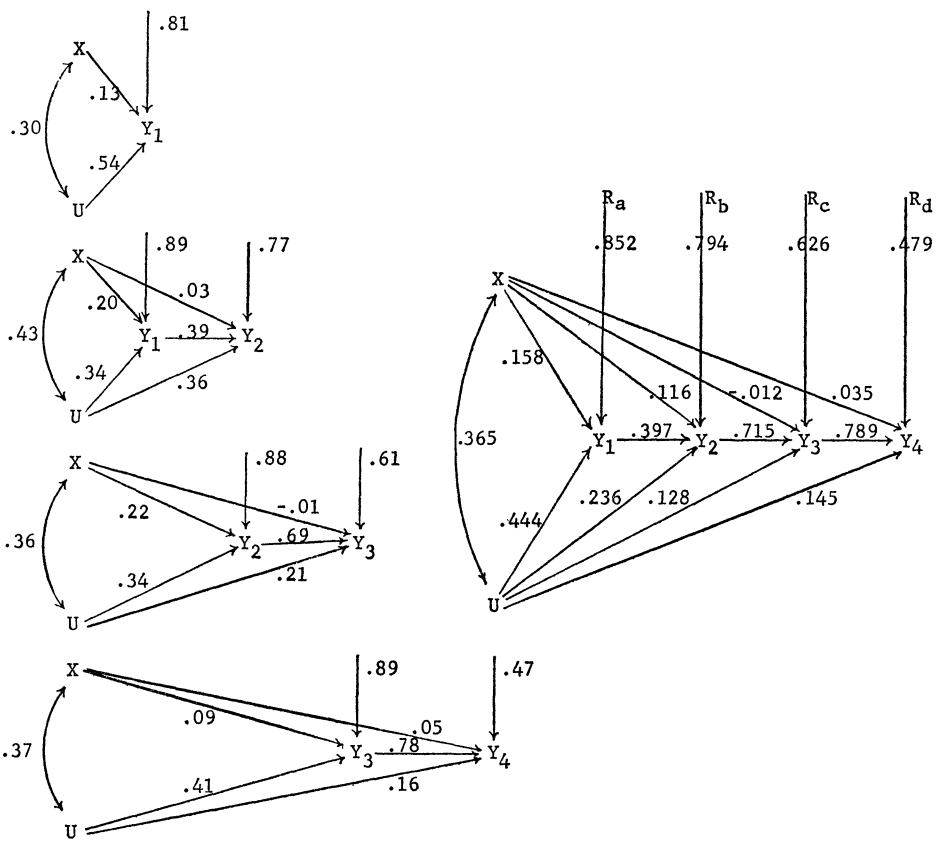

FIG. 4.—Respondent's occupational status (Y) at successive ages, in relation to father's occupational status (X) and respondent's educational attainment (U). Occupational status at age 25–34 = Y_1; at 35–44 = Y_2; at 45–54 = Y_3; at 55–64 = Y_4. (Source: Duncan and Hodge, *op. cit.*, and unpublished calculations kindly supplied by the authors.)

invaluable for heuristic purposes. Pending the execution of full-blown longitudinal studies on real cohorts, the synthetic cohort is, at least, a way of making explicit one's hypotheses about the sequential determination of experiences cumulating over the life cycle.[31]

In a study of the social mobility of a sample of Chicago white men with nonfarm backgrounds surveyed in 1951, Duncan and Hodge,[32] used data on father's occupational status, respondent's educational attainment, and respondent's occupational status in 1940 and 1950 for four cohorts: men 25–34, 35–44, 45–54, and 55–64 years old on the survey date. Their main results, somewhat awkwardly presented in the source publication, are compactly summarized by the first four diagrams in Figure 4. (The superfluous squared term in their equations has been eliminated in the present calculations. The amount of curvilinearity was found to be trivial, and curvilinear relations cannot be fitted directly into a causal chain by the procedure employed here.)

These data involve partial records of the occupational careers of the four cohorts and thus depict only segments of a continuous life history. In the original analysis, it was possible to gain some insights from the interperiod and intercohort comparisons on which that analysis was focused. Here, attention is given to a different use of the same information. Suppose we thought of the four sets of data as pertaining to a single cohort, studied at four successive points in time, at decade intervals. Then, all the data should fit into a single causal or processual sequence.

It is obvious that one cannot achieve perfect consistency on this point of view. The initial correlation, r_{UX}, varies among cohorts, for example. Moreover, age-constant intercohort comparisons of the other correlations (the Y's with X and U) suggest that some variations result from genuine differences between the conditions of 1940 and 1950. But if one is willing to suppress this information for the sake of a necessarily hypothetical synthesis, it is possible to put all the data together in a single model of occupational careers as influenced by socioeconomic origins.

The four correlations r_{UX} were averaged. The remaining correlations for adjacent cohorts were likewise averaged; for example, r_{1U} based on 1950 data for men 25–34 years old was averaged with r_{1U} based on 1940 data for men 35–44 in 1951, and so on. Only r_{4X}, r_{4U}, and the three intertemporal correlations, r_{21}, r_{32}, and r_{43}, had to be based on the experience of just one cohort. (In deriving this compromise one does, of course, lose the temporal specificity of the data by smoothing out apparently real historical fluctuations.) When the correlations had been averaged, the results shown in the "composite" model on the right of Figure 4 were obtained. The estimates of path coefficients here are simply the partial regression coefficients, in standard form, of Y_1 on X and U; Y_2 on Y_1, X, and U; Y_3 on Y_2, X, and U; and Y_4 on Y_3, X, and U.

The results for the synthetic cohort make explicit the following interpretations: (1) The background factors, father's education (X) and respondent's education (U), have an important direct impact during early stages of a cohort's life cycle; after age 35–44 their direct effects become small or negligible, although they exert indirect effects via preceding achieved statuses (Y_1 and Y_2). (2) Careers tend to stabilize after age 35–44, as indicated by the sharp rise in the path coefficients representing persistence of status over a decade (compare p_{21} with p_{32} and p_{43}) and by the decreasing magnitudes

---

Journal of the American Statistical Association, XLI (December, 1946), 501–16.

[31] See, for example, A. J. Jaffe and R. O. Carleton, *Occupational Mobility in the United States: 1930–1960* (New York: King's Crown Press, 1954), p. 53 (n. 6) and Table 13.

[32] Otis Dudley Duncan and Robert W. Hodge, "Education and Occupational Mobility," *American Journal of Sociology,* LXVIII (May, 1963), 629–44.

of the residual paths from $R_a, \ldots, R_d$. (3) During the life cycle, many circumstances essentially independent of background factors affect occupational mobility, so that achievement in the later stages of the career becomes more and more dependent upon intervening contingencies while continuing to reflect the indirect influence of conditions determinate at the outset. Thus, for example, r_{4c}—the correlation of occupational status at age 55–64 with residual for age 45–54—may be computed as (.789)(.626) = .494, and the residual path to

TABLE 2

OBSERVED AND IMPLIED (*) CORRELATIONS FOR SYNTHETIC COHORT MODEL OF OCCUPATIONAL ACHIEVEMENT

VARIABLE (AGE AND OCCUPATIONAL STATUS)	VARIABLE		
	Y_2	Y_3	Y_4
25–34 (Y_1)	.552	.455*	.443*
35–44 (Y_2)	.....	.772	.690*
45–54 (Y_3)	.....	.....	.866
55–64 (Y_4)	.....	.....	.....

Source: Duncan and Hodge, *op. cit.*, and calculations from model in Figure 4.

Y_4 itself is $P_{4d} = .479$. These are comparable in size with the correlations $r_{4X} = .301$ and $r_{4U} = .525$. The residuals are, by definition, uncorrelated with X and U and represent, therefore, the influence of factors quite unrelated to social origins and schooling. The prevailing impression that the United States enjoys a rather "loose" stratification system is thus quantified by a model such as this one. (4) While the data include observed interannual correlations of occupational statuses separated by a decade (r_{43}, r_{32}, and r_{21}), the synthetic cohort model also implies such correlations for statuses separated by two or three decades. These may be computed from the following formulas based on equation (5):

$$r_{42} = p_{4X}r_{2X} + p_{43}r_{32} + p_{4U}r_{2U},$$

$$r_{31} = p_{3X}r_{1X} + p_{32}r_{21} + p_{3U}r_{1U},$$

$$r_{41} = p_{4X}r_{1X} + p_{43}r_{31} + p_{4U}r_{1U},$$

inserting the value of r_{31} obtained from the second equation into the third. The observed and implied correlations are assembled in Table 2. The latter represent, in effect, hypotheses to be checked whenever data spanning twenty or thirty years of the occupational experience of a cohort become available. In the meantime, they stand as reasonable estimates, should anyone have use for such estimates. If forthcoming evidence casts doubt on these estimates, the model will, of course, be called into question. It is no small virtue of a model that it is capable of being rejected on the basis of evidence.

This last example, since it rests on an explicit fiction—that of a synthetic cohort—perhaps makes clearer than previous examples the point that the role of path analysis is to *render an interpretation* and not merely to provide a format for presenting conventional calculations. In all the examples the intention has been to adhere to the purpose of path analysis as Wright formulated it:

> . . . the method of path coefficients is not intended to accomplish the impossible task of deducing causal relations from the values of the correlation coefficients.[33] . . . The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain, the method can be used to find the logical consequences of any particular hypothesis in regard to them.[34] . . . Path analysis is an extension of the usual verbal interpretation of statistics not of the statistics themselves. It is usually easy to give a plausible interpretation of any significant statistic taken by itself. The purpose of path analysis is to determine whether a proposed set of interpretations is consistent throughout.[35]

[33] "The Method of Path Coefficients," p. 193.

[34] "Correlation and Causation," *Journal of Agricultural Research*, XX (1921), 557–85 (quotation from p. 557).

[35] "The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis," p. 444.

NEGLECTED TOPICS

This paper, for the lack of space and especially for lack of "convincing" examples, could not treat several potentially important applications of path analysis: (1) Models incorporating feedback were explicitly excluded. Whether our present techniques of social measurement are adequate to the development of such models is perhaps questionable. (2) The problem of two-wave, two-variable panel analysis, recently discussed by Pelz and Andrews,[36] might well be formulated in terms of path coefficients. The present writer, however, has made little progress in attempts to clarify the panel problem by means of path analysis. (3) The pressing problem of the disposition of measurement errors[37] may perhaps be advanced toward solution by explicit representation in path diagrams. The well-known "correction for attenuation," where measurement errors are assumed to be uncorrelated, is easily derived on this approach.[38] It seems possible that under very special conditions a solution may also be obtained on certain assumptions about correlated errors. (4) Wright has shown[39] how certain ecological models of the interaction of populations can be stated in terms of path coefficients. The inverse method of using path analysis for studies of multiple time series[40] merits consideration by sociologists. (5) Where the investigation involves unmeasured variables, path analysis may be helpful in deciding what deductions, if any, can be made from the observed data. Such unmeasured variables may, in principle, be observable; in this case, path analysis may lead to hypotheses for testing on some future occasion when measurements can be made. If the unmeasured variable is a theoretical construct, its explicit introduction into a path diagram[41] may well point up the nature of rival hypotheses. Ideally, what are sometimes called "validity coefficients" should appear explicitly in the causal model so that the latter accounts for both the "true causes" under study and the ways in which "indicator variables" are thought to represent "underlying variables." A particular case is that of factor analysis. As Wright's work demonstrates,[42] a factor analysis is prone to yield meaningless results unless its execution is controlled by explicit assumptions which reflect the theoretical structure of the problem. An indoctrination in path analysis makes one skeptical of the claim that "modern factor analysis" allows us to leave all the work to the computer.

UNIVERSITY OF MICHIGAN

[36] Donald C. Pelz and Frank M. Andrews, "Detecting Causal Priorities in Panel Study Data," *American Sociological Review*, XXIX (December, 1964), 836–54.

[37] H. M. Blalock, Jr., "Some Implications of Random Measurement Error for Causal Inferences," *American Journal of Sociology*, LXXI (July, 1965), 37–47; Donald J. Bogue and Edmund M. Murphy, "The Effect of Classification Errors upon Statistical Inference: A Case Analysis with Census Data," *Demography*, I (1964), 42–55.

[38] Wright, "The Method of Path Coefficients" and "Path Coefficients and Path Regressions."

[39] "The Treatment of Reciprocal Interaction."

[40] *Ibid.*

[41] H. M. Blalock, Jr., "Making Causal Inferences for Unmeasured Variables from Correlations among Indicators," *American Journal of Sociology*, LXIX (July, 1963), 53–62.

[42] "The Interpretation of Multivariate Systems."