CAUSAL INFERENCES FROM DICHOTOMOUS VARIABLES

N.Davidson

ISBN 0 902246 59 3

ISSN 0306-6142

CATMOG

(Concepts and Techniques in Modern Geography)

CATMOG has been created to fill a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for the teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

1. An introduction to Markov chain analysis - L. Collins
2. Distance decay in spatial interactions - P.J. Taylor
3. Understanding canonical correlation analysis - D. Clark
4. Some theoretical and applied aspects of spatial interaction shopping models - S. Openshaw
5. An introduction to trend surface analysis - D. Unwin
6. Classification in geography - R.J. Johnston
7. An introduction to factor analytical techniques - J.B. Goddard & A. Kirby
8. Principal components analysis - S. Daultrey
9. Causal inferences from dichotomous variables - N. Davidson
10. Introduction to the use of logit models in geography - N. Wrigley
11. Linear programming: elementary geographical applications of the transportation problem - A. Hay
12. An introduction to quadrat analysis - R.W. Thomas
13. An introduction to time-geography - N.J. Thrift
14. An introduction to graph theoretical methods in geography - K.J. Tinkler
15. Linear regression in geography - R. Ferguson
16. Probability surface mapping. An introduction with examples and Fortran programs - N. Wrigley
17. Sampling methods for geographical research - C. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C. Dixon & B. Leach

Other titles in preparation

This series, Concepts and Techniques in Modern Geography is produced by the Study Group in Quantitative Methods, of the Institute of British Geographers.
For details of *membership of the Study Group, write to the Institute of British Geographers, 1 Kensington Gore, London, S.W.7.*
The series is published by Geo Abstracts, University of East Anglia, Norwich, NR4 7TJ, to whom all other enquiries should be addressed.

CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY No. 9

CAUSAL INFERENCES FROM DICHOTOMOUS VARIABLES

by

NORMAN DAVIDSON
(Hull University)

CONTENTS

Acknowledgement

Figure 2 is adapted from James A. Davis, ELEMENTARY SURVEY ANALYSIS, (C) 1971, p 87. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey, U.S.A.

I INTRODUCTION

(i) Spurious correlation and the need for causal inferences

The reader of texts on quantitative techniques in geography will be aware how often he is warned of the dangers of inferring causation from association. Finding a good correlation between two sets of observations, we are told, does not necessarily mean that one causes the other for the relationship could have arisen accidentally.

If it is an objective of geographical investigation to derive axioms, simple or otherwise, about spatial distributions, we need to be sure that they are not based on the shifting sands of accidental relationships. We want to be sure that any connection we draw between objects, events or processes is true. How do we set about this? Simon (1954) provides us with a model argument in which he suggests that the acid test of a true relationship is that it should not change regardless of the conditions under which it is observed. Thus

> in investigating spurious correlation we are interested in learning whether the relation between two variables persists or disappears when we introduce a third variable.' (p 469)

The problem in simple two-variable correlation is that we are obliged to ignore other factors which may bear on either or both variables. If, however, we can include and control for relevant third variables, inferences about cause and effect become possible and legitimate.

The purpose of this monograph is to provide a set of tests for establishing whether correlations are true or spurious, and to elaborate some procedures for drawing legitimate causal inferences. The techniques are based on dichotomies both for reasons of simplicity and to show that quite powerful conclusions may be inferred from relatively unsophisticated material. In the conclusions we will note how some of the concepts and procedures may be extended to other kinds of data. The reader interested in the more philosophical aspects of causation is referred to the introduction of Simon's paper above and to the relevant chapters of Harvey (1969).

(ii) Prerequisites

In view of the simplicity of the data used, very little mathematical knowledge is required to follow the methods outlined here though an elementary grasp of probability theory will be useful. More important is a basic understanding of the logic of hypothesis testing in a geographic context (c.f. Harvey (1969) Ch 15; Hammond and McCullagh (1971) Ch 6). The emphasis will, in any case, be on the interpretation of results rather than on the statistical procedures themselves. Since the methods are suitable for a wide range of applications in both human and physical geography, no special knowledge is required, though it will become evident that any particular application will demand a thorough consideration of the objects, events or processes involved in the axiom being tested. Perhaps the greatest utility for the methods will be found in the context of analysing survey material whether from questionnaires or field observation. Some experience with survey methods and with cross-tabulation would certainly provide a useful introduction to this text.

(iii) Geographical applications

Despite common application in the geographic literature of the simple correlation and chi-square tests of association, it is rare to find examples of the explicit control of a third variable through the procedure generally known as partial correlation. Even rarer are attempts to elucidate causal connections. Rather and sometimes regrettably geographers have tended to go for statistically more sophisticated multivariate procedures such as factor analysis. Some of the difficulties of interpretation encountered by these applications may well arise from a failure to consider properly the relations involved.

Some good examples do exist, however. Johnston (1974) in a study of Christchurch, New Zealand, shows that proximity is a major influence on patterns of social contact even when social distance is held constant. Mercer (1975) discusses how the relationship between blacks and poor housing in American cities is affected by consideration of poverty and racial discrimination. He is able to demonstrate that the connection between being black and living in poor housing is indirect - a product of low incomes and segregation. In a discussion of alluvial meanders, Ferguson (1973) suggests that their wavelength is controlled by discharge and bank strength through width and sinuosity using partial correlation to elaborate his model. Gregory, in his presidential address to the IBG (1976), quotes the case of a positive correlation observed for convex slopes between slope angle and soil depth as an example of spurious correlation. To establish the true connection, distance down slope (amongst other factors) would have to be controlled. Two studies adopting a rather different approach to causal relations are Pickvance's (1974) analysis of the determinants of household mobility and Winsborough's (1962) decomposition of population densities which Duncan (1971) further elaborates. Cox (1968) elaborates the factors influencing suburban voting patterns, an analysis which should be read in the light of Taylor's (1969) comments.

These and other examples are fairly advanced in their conceptual discussion, but there is no reason to suppose all applications need be at the research frontier. Indeed as the worked examples used in this monograph will show, there is much to be gained in testing and elaborating causal relations at a much simpler level. It will certainly lead to a more critical appraisal of the axioms about spatial distributions for which geographers strive.

(iv) The survey data for the worked examples.

Throughout the discussion of methods to follow frequent reference will be made to empirical examples. For consistency these are all taken from surveys of journey-to-work habits undertaken by first-year geography students at Hull University. The purpose of this section is to define the variables used in the examples.

The survey area in the proximity of the University was chosen for its varied residential character. It contained nineteenth century terraced housing, owner-occupied housing of rather mixed age and type and part of an inter-war council estate. Two fifths of the houses were selected systematically street-by-street and a straightforward questionnaire administered. The journey-to-work details of 480 economically active persons in the responding households were recorded.

Table 1 contains a list of the variables used in the examples, together with the definition of their dichotomisation. It should be noted that the sample size for particular examples will vary according to the number of individuals omitted for either classification reasons or because of lack of information.

Table 1 : Journey-to-work variables

Personal variables		
DISTANCE	Distance[1] travelled to work	1 = 3km or over 0 = Under 3km
TIME	Time taken[2] travelling to work	1 = 20 mins. or over 0 = Under 20 mins.
MODE1[3]	Mode of transport used	1 = Car 0 = Other
MODE2[3]	Mode of transport used	1 = Car) other omitted 0 = Bus)
SEX		1 = Male 0 = Female
OCCUPATION		1 = Non-manual 0 = Manual
HEAD	Relation to head of household	1 = Head 0 = Other
Household variables[4]		
HOTYPE	Type of house	1 = Detached, semi-detached, bungalow 0 = Terraced, flat, other
TENURE		1 = Owner occupied 0 = Rented (LA and private), other

Notes:

[1] Straight-line map distance between residence and place of work

[2] Persons with extremely variable time omitted

[3] Two alternative classifications of mode; the second omits other categories

[4] Household variables are defined for household as a whole but applied to each economically active person within that household

II TWO VARIABLE ANALYSIS

Let us start from scratch. Suppose we have a sample population for which we have measured two dichotomous variables, i.e. two attributes that individuals in the population either possess or lack, be it red hair, right-wing sentiments, two cars, a southern aspect', a diameter of over 20 microns or whatever. We wish to establish the degree to which the distributions of these attributes are related. The classic method is to cross-tabulate and use the chi-square test of association. For various reasons, however, we cannot use chi-square if we wish to test whether the association persists under different conditions. The primary difficulty is that it is impossible to compare chi-square values unless both sample sizes happen to be identical. It is necessary, therefore, to introduce a less well-known coefficient of association for dichotomies developed by the statistician G.U. Yule and labelled Yule's Q (Goodman & Kruskall, 1954) which has the required properties for three variable analysis.

(i) Logic of Yule's Q

In a test of association between dichotomous variables, we are interested in the relative frequency of occurrence of different combinations of two variables which, following the usual convention, we will call X and Y. In establishing a connection between X and Y, we are in fact implying that a change in X will produce a non-random change in Y (or vice-versa). These changes may be regarded as consistent (where presence of X produces presence of Y) or inconsistent (where presence of X produces absence of Y and vice-versa). For Yule's Q, the critical ratio is:-

$$\frac{\text{consistent changes - inconsistent changes}}{\text{changes in both X and Y}}$$

We can see immediately that if all changes are consistent, this ratio will have a value of +1, and if all the changes are inconsistent, -1. If neither predominates, the ratio will be zero.

Assuming that the frequency distribution of X and Y in our sample is:-

	NOT Y	Y	
X	A	B	
NOT X	C	D	
			N

..... (1)

The proportion of consistent changes can be shown to be:-

$$2(\frac{B}{N} \cdot \frac{C}{N})$$

and the proportion of inconsistent changes:-

$$2(\frac{A}{N} \cdot \frac{D}{N})$$

Thus the critical ratio becomes:-

$$\frac{2(\frac{B}{N} \cdot \frac{C}{N}) - 2(\frac{A}{N} \cdot \frac{D}{N})}{2(\frac{B}{N} \cdot \frac{C}{N}) + 2(\frac{A}{N} \cdot \frac{D}{N})} = \frac{B.C - A.D}{B.C + A.D} = Q \quad ..... (2)$$

This argument, very much abbreviated here, is elaborated by Davis (1971 pp 39-50). The critical ratios of Q and other coefficients of association for dichotomies are reviewed by Goodman (1965).

In summary, Q measures the degree of association between two dichotomou variables. It assumes a value of 1.00 when there is perfect and direct association; -1.00 where there is perfect and inverse association; and 0.00 where there is no association. Its properties, therefore, are closely similar to the product-moment coefficient for interval data. It has the additional property of being invariant to sample size. Thus the percentage distribution will give the same Q as the actual frequencies (though as we shall see in the next section, the latter are required for significance testing).

(ii) Significance of Yule's Q

As with similar coefficients the actual value of Q depends not only on the strength of the association but also on the size of sample on which it is measured. To assess the significance of Q, the effect of sample size must be controlled. Two methods are given below: an exact method which can be laborious to apply and a quick graphical method which will suffice except in certain critical cases

(a) Exact Method

Using the cell frequencies (1) above, the confidence interval of Q is shown (Kendall and Stuart, 1961, p540) to be approximately normally distributed and the upper and lower limits of Q given by

$$Q_{LIMITS} = Q \pm Z \sqrt{\frac{(1 - Q^2)^2(\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D})}{4}} \quad ..... (3)$$

where Z = 1.96 at the 5% significance level
Z = 2.33 at the 1% significance level
Z = 3.09 at the 0.1% significance level

Q is not significant if zero lies between the limits at the chosen significance level (i.e. their sign is different). To be more precise the difference between the observed (sample) Q and a population Q of zero (under the null hypothesis that no relationship exists in the population) is not significant.

If zero lies outside the limits, Q is significant.

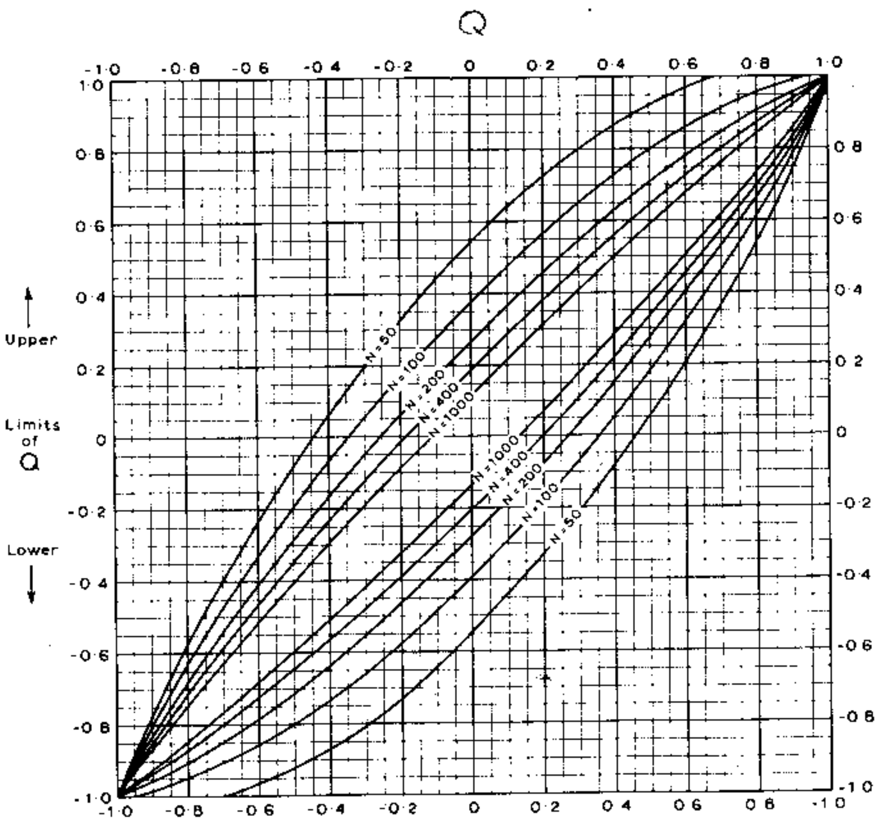


Figure 1: Minimum confidence limits of Q at 5% level.

(b) Graphical Method

In Figure 1 the upper and lower limits for varying values of Q and for varying sample sizes are given at the 5% significance level. An important assumption has been made, however, that the frequency split within both X and Y is equal (i.e.A+B=C+D=A+C=B+ D). This means that the limits given by the graphical method are the minimum limits for a particular Q : more extreme splits will tend to increase the confidence interval.

To establish whether Q is significant or not, draw on the graph the two curves representing the appropriate sample size. Find the value of Q on the Q-scale. The upper and lower limits are given on the limits-scale where the vertical line representing this value intersects the two limits curves. If zero on the limits-scale lies between the limits, Q is not significant at the 5% level. Again to be more precise, the null hypothesis that no relationship exists in the population is accepted. Since the curves represent the minimum limits, such a result may be confidently accepted whatever the split on X and Y. If zero lies outside the limits, Q is significant at the 5% level unless zero is just outside the limits and the split on X or Y is very uneven. If these latter conditions hold, the exact method should be applied. (As a rough-and-ready guide for extreme splits, the critical area will extend to the curves representing sample sizes approximately 1/2 to 2/3 the actual. A little juggling with notional splits will show how the sample size has to be varied to maintain a given confidence level.)

The graphical method may also be used to indicate the critical values of Q at the 5% significance level for a given sample size. These will be given by the values on the Q-scale where the limits curves intersect zero on the limits-scale (see Table 11 for a summary). This facility may be useful when sifting a data set for potential relationships though such a shot-gun approach is not to be recommended when searching for causal connections.

(iii) Data constraints

Since Yule's Q is based on simple dichotomies, it is relatively free from data constraints. However the following observations should be borne in mind:-

(b) Dichotomisation: whenever possible ensure that variables vary. Where there is an element of choice in the criterion of possessing or lacking a characteristic (e.g. if a number of categories are to be grouped, or if a critical level of a numeric variable is to be taken) try to ensure that there is as even a split of the population as possible (i.e. not worse than 70:30, though this is no more than a rough rule). However, beware of grouping categories which should, in theory, be different in order to satisfy this rule. In cases where no choice of definition exists (e.g. sex; presence/absence of botanical species; etc.) it may be necessary to increase the sample size to meet conditions (b) below. If there are alternative ways of dichotomising a particular variable, be prepared to investigate each.

Cell frequencies: the following conditions must be fulfilled:-
(1) no cell may have an ex ected frequency of less than 5 since the test becomes unstable in these conditions (cf. Siegel, 1956, p110). Since the smallest expected frequency will result from the combination of the categories of X and Y with the lower proportionate occurrence, hence the need for ensuring as even a split of dichotomisation as possible. This requirement may be applied in reverse to give some indication of the minimum sample size requirement for Yule's Q:-
suppose P_i ...... are the proportions in the smaller category of each variable, then to test association between variables i and j, $N \geqslant \frac{5}{P_iP_j}$ (in 3 variable case $N \geqslant \frac{5}{P_iP_jP_k}$)

(2) no cell should have an <u>observed</u> frequency of zero. If this is so, by definition Q ± 1.00 indicating perfect association. However, as Blalock (1972, pp 298-299) points out, it is dubious whether this should be so unless a second cell is also zero, i.e. when the condition of either variable <u>mutually</u> excludes one of the conditions of the other. It is probably best to avoid these inferential difficulties by reclassifying one of the variables concerned should this circumstance arise.

(c) <u>Model-building:</u> in a more general way it is worthwhile giving some consideration to theoretical expectations when organising data for analysis. The Yule's Q procedure is more sensitive than the very simple data it uses would suggest. Do the variables chosen fairly reflect the theoretical constructs in the relationship hypothesised? Is the model suggesting that possession of one characteristic is related to possession or lack of another? Which characteristic is likely to be dependent on what? These considerations become absolutely vital in a three-variable analysis, so the strengths and weaknesses of the data should not be ignored.

(iv) <u>A two-variable worked example</u>

<u>Hypothesis:</u> that the time taken and the distance travelled to work are related. This is a straightforward assertion, the theoretical justification for which is not too taxing. We would expect that the further an individual travels, the longer time he is likely to take. Note, however, that we are already even at this stage implying conditions that we may wish to examine - e.g. 'likely to take' suggests that the relationship is not going to be perfect and moreover may be subject to some systematic distortion - a third variable. We are, even at this stage, asserting that time is dependant on distance and not vice-versa. One does not get off the bus when a given time has elapsed.

Employing the usual convention of defining Y as the dependent variable, X becomes distance and Y time.

Data: since both time and distance are measured on a continuous scale, the dichotomisation is designed to ensure as even a split as possible on each variable. In the journey-to-work survey (see I(iv)) the following frequencies were observed:-

Table 2 : Distance and time travelled to work

	TIME (minutes)		
DISTANCE (km)	Under 20	20 or over	
3 or over	77	146	223
under 3	192	48	240
	269	194	463

Results: $$Q = \frac{146.192 - 77.48}{146.192 + 77.48} = +0.767$$

Limits (graphical method) upper = + 0.84 (approx)
lower = + 0.69 (approx)

Since the split is not extreme on either X nor Y and the lower limit is well above zero, the null hypothesis may be confidently rejected at the 5% level. The positive association between time and distance is demonstrated.

(Exact method)	5% limits:	upper = +0.853
		lower = +0.681

	0.1% limits:	upper = +0.903
		lower = +0.631

III <u>THE EFFECT OF CONTROLLING FOR A THIRD VARIABLE</u>

(i) <u>Introduction: possible outcomes</u>

We now wish to test the stability of our relationship by examining how it is affected by variations in a third variable (which we will call T, the test variable). In order to do this, we must introduce some further coefficients. First let us now call the straightforward Q between X and Y the <u>zero-order</u> coefficient (Z). If only those individuals possessing the T characteristic are selected and we measure the strength of association between X and Y for them alone, this value is called a <u>conditional</u> coefficient (i.e. it is conditional on possessing T). A second conditional may be obtained for those individuals lacking T. The weighted average of the two conditionals is called the <u>partial</u> coefficient (P), which is a measure of association when the third variable is held constant - either present or absent. In the partial, therefore, variations in T may be said to be controlled. When T is allowed to vary, the coefficient is called the <u>differential</u> (D). Again since we are dealing with dichotomies, we will find that the weighted average of the partial and the differential is the zero-order. The weights are proportionate to the relative frequency with which the criteria occur (see Davis, 1971, p85).

So we may decompose the zero-order correlation between two variables into a partial and a differential for a third variable. The former will be of more interest since it allows the third variable to be controlled. It may itself be decomposed into two further conditionals.

To understand the effect of a third variable on a relationship, the critical comparison is between the zero-order and the partial coefficients. In fact we are simply asking ourselves the question 'What happens to the relationship when T is controlled?' The possible answers are no change', 'gets stronger' or 'gets weaker'. What do these mean for the relation between zero-order and partial? No change' would refer to the situation where P = Z, i.e. there is no difference in the relationship whether or not T is controlled. This outcome is generally called <u>no effect</u> - the third variable has no bearing on the XY relationship. If $|P|<|Z|$ ('got weaker') then the relationship is tending to disappear when T is controlled. Such an outcome is called <u>explanation</u> : it is variation in T which explains the existence of the XY relationship. Finally IPl>IZI suggests that the relationship is being concealed by variations in T. This outcome is called <u>suppression</u> and refers to the situation where the existence of a relationship is revealed or greatly strengthened by controlling for T. Note that the relational indicators used here should be interpreted in terms of significance rather than numerical equality.

There is one further outcome which rather complicates matters. It arises when there is a <u>different</u> relationship among individuals possessing T than among those lacking it, i.e. the conditionals differ in sign or very greatly in value. The effect of T is to specify which of two different relationships holds, hence this outcome is called <u>specification</u> (sometimes known as interaction). The complication is that specification overlaps the other outcomes. While it is rare for explanation or suppression to involve specification, a no effect situation can arise in this way when two conditionals, very much different from the zero-order in size and/or sign, cancel each other when combined in the partial. We will have to check for specification before concluding the effect of T.

(ii) <u>Formulae for partial, differential and conditional coefficients.</u>

Let the cell frequencies in the 3-way cross-tabulation be:-

<u>Table 3:</u> Cell notation for 3-way tables

		NOT Y	r
T	X	A	B
	NOT X	C	D
NOT T	X	E	F
	NOT X	G	H

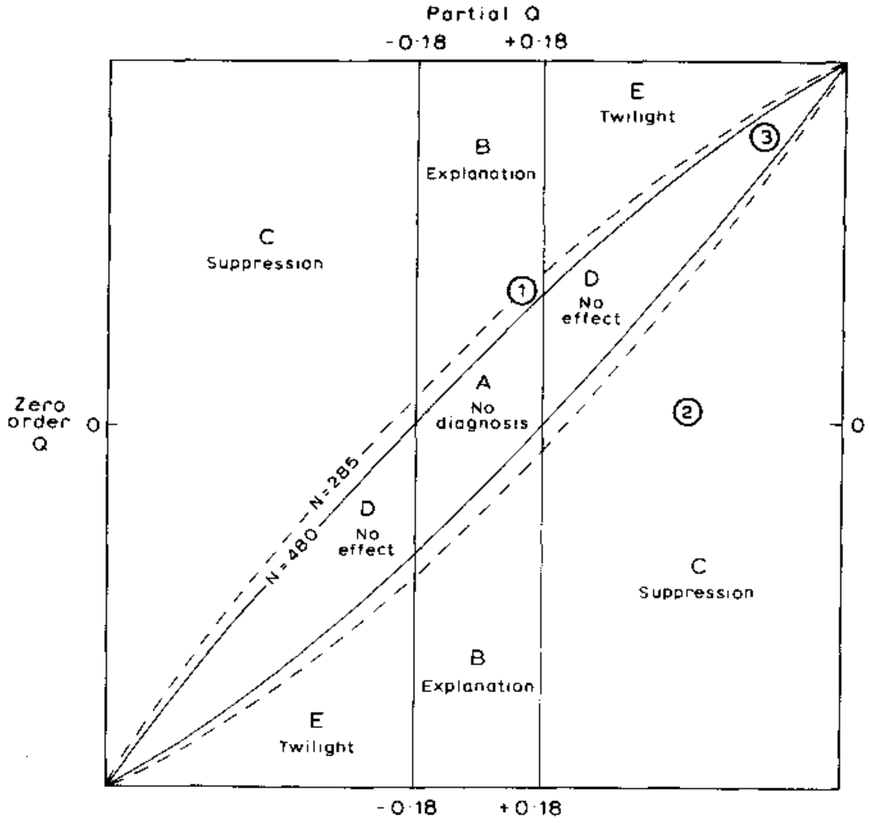
(a) <u>Zero-order</u> (Q_{XY}): from a 3-way table, the two sub-tables for T are merged, giving

$$Q_{XY} = \frac{\left[(B + F) \quad (C + G)\right] - \left[(A + E) \quad (D + H)\right]}{\left[(B + F) \quad (C + G)\right] + \left[(A + E) \quad (D + H)\right]} \quad .... (4)$$

The simple symbol Q is not used in 3-variable analyses to avoid confusion.

(b) <u>Partial</u> ($Q_{XYTIEDT}$) : this is defined as the ratio of consistent to inconsistent changes among those exhibiting no change in T. In this way, attention is directed at a particular combination of changes among the individuals. The argument is developed quite simply from the two variable case given in Section II(i) without the statistical derivation (see Davis, 1971, p84) to give the formula

$$Q_{XYTIEDT} = \frac{\left[(B.C) + (F.G)\right] - \left[(A.D) + (E.H)\right]}{\left[(B.C) + (F.G)\right] + \left[(A.D) + (E.H)\right]} \quad .....(5)$$



① Sex & mode effect of head of household (N.480)

② Distance 8. mode : effect of time (N =285)

③ Distance &time : effect of occupation (N=463)

Figure 2: Effect of third variable - outcome regions.

(c) Differential ($Q_{XYDIFFT}$) : this is likewise developed from the ratio of consistent to inconsistent changes, but this time among individuals differing on T. i.e.

$$Q_{XYDIFFT} = \frac{[(B.G) + (C.F)] - [(A.H) + (D.E)]}{[(B.G) + (C.F)] + [(A.H) + (D.E)]} \quad \ldots\ldots (6)$$

(d) Conditional T (Q_{XYT}) : the zero-order for individuals possessing T is

$$Q_{XYT} = \frac{B.C - A.D}{B.C + A.D} \quad \ldots\ldots (7)$$

(do not confuse this formula with the simple zero-order (formula 2))

(e) Conditional Not T (Q_{XYNOTT}) : is the zero-order for individuals lacking T

$$Q_{XYNOTT} = \frac{F.G - E.H}{F.G + E.H} \quad (8)$$

(f) Coefficients for the other relationships

From the same frequency table it is possible to define the coefficients corresponding to the other two relationships - between X and T, and Y and T. As we shall see in the next section, these coefficients are required before the causal system linking the three variables can be fully validated.

It will be noted that all the formulae above have a common form:-

$$\frac{P - Q}{P + Q}$$

The elements, P and Q, of all 15 coefficients are given in Table 4.

(iii) Diagnosing the effect of a third variable

Kendall and Lazarsfeld (1950) outline a scheme for categorising the outcomes of controlling for a third variable. Davis (1971, Figure 4.1, p87) presents a simple and explicit solution to the problem of comparing the zero-order and partial coefficients. Figure 2 is an adaptation of Davis' solution using the limits curves of Figure 1, rather than his arbitrary limits. Figure 1 may in fact be used for this purpose by plotting the value of the partial on the Q-scale and the zero-order on the limits-scale. By adding to the two limits curves (already drawn for the appropriate sample size) two further vertical lines through the points where the limits curves intersect the horizontal zero line, five outcome regions are defined (Figure 2).

It should be stressed that it is the partial which is now plotted on the same scale as the zero-order when the latter's significance was previously being established. Regrettably the graph cannot be turned on its side since the limits curves are not quite symmetrical. In a sense what we are doing here is suggesting that the partial is the 'true' relationship and treating its significance in the same way as we previously treated the zero-order.

Table 4 : Q coefficients for 3-variable analysis

Coefficients : all $\frac{P - Q}{P + Q}$		P	Q
Zero-orders	Q_{XY}	(B + F) (C + G)	(A + E) (D + H)
	Q_{XT}	(A + B) (G + H)	(E + F) (C + D)
	Q_{YT}	(B + D) (E + G)	(A + C) (F + H)
Partials	$Q_{XYTIEDT}$	(B.C) + (F.G)	(A.D) + (E.H)
	$Q_{XTTIEDY}$	(B.H) + (A.G)	(D.F) + (C.E)
	$Q_{YTTIEDX}$	(B.E) + (D.G)	(A.F) + (C.H)
Differentials	$Q_{XYDIFFT}$	(B.G) + (C.F)	(A.H) + (D.E)
	$Q_{XTDIFFY}$	(B.G) + (A.H)	(C.F) + (D.E)
	$Q_{YTDIFFX}$	(B.G) + (D.E)	(A.H) + (C.F)
Conditionals	Q_{XYT}	B.C	A.D
	Q_{XYNOTT}	F.G	E.H
	Q_{XTY}	B.H	D.F
	Q_{XTNOTY}	A.G	C.E
	Q_{YTX}	B.E	A.F
	Q_{YTNOTX}	D.G	C.H

Outcome regions and effect of T

(A) No diagnosis

This region in the middle of the diagram refers to a situation where the partial is not significant and the zero-order lies within its limits. This is a special case of no effect where there is no relationship between X and Y whether T is controlled or not. Such an outcome is always a little disappointing and may lead us to question our initial hypothesis.

(B) Explanation

These two areas of the graph relate to the situation where the partial is not significant, but the zero-order is (either positive or negative). The relationship disappears when I is controlled. By way of illustration, consider the potential relationship between sex and mode of travel to work (the latter in this case taken as MODEL, whether or not the individual travels by car). It may be hypothesised that the much-discussed dominance of males in our society will lead to a significant and positive association here - men tending to travel by car, women by other means. A frequency count in the journey-to-work survey revealed the following figures:

Table 5 : Sex and mode of transport

	Other modes	Car
Males	172	137
Females	124	47

The zero-order coefficient is +0.36, with limits of +0.64 and +0.08 at the 0.1% level - a strong association. At this point the believer in sex discrimination may retire with yet another fact for their armoury. The alert student, however, may ponder the implications awhile. Who grabs the car? Is it always or even usually the man? Or is it the individual who has contributed most to its acquisition? Let us control for head of household (on the grounds that it is he or she who is most likely to command household resources) and observe its effect on the association between sex and travelling to work by car.

Table 6 : Sex and mode and head of household

		Other modes	Car
Heads of household	Males	132	118
	Females	10	6
Other members	Males	40	19
	Females	114	41

The partial is +0.16, which lies within the 5% limits of zero association (±0.18). The relationship is therefore seen to disappear when controlled for head of household (this outcome marked as 1 on Figure 2). This may be confirmed by observing the two conditionals. For heads of households the association is +0.20 and for other members of the household, +0.14. Note that the confidence interval for the conditionals is larger since the sample size in the sub-tables is smaller: the limits are ±0.24 and ±0.27 respectively for no association. Thus the relationship exists for neither heads nor other members of households. The zero-order relationship between sex and mode may now be said to be spurious - the product of variation in some exogenous factor. A quick inspection of the table will suggest that this arose because most heads of households are male (the reader might like to test whether this is a true relation).

(C) Suppression

Consider now the situation where the partial is greater than the zero-order. This can happen when we find no association between X and Y unless T is controlled or when a weak association (either positive or negative) is significantly improved by controlling for T.

A

An illustration of this outcome is provided by a consideration of mode of transport and distance travelled. In this case, MODE2 - travelling by car or bus - is taken to highlight the effect. We may suppose that being able to travel by car would allow a greater distance to be covered, or conversely that a more distant and inaccessible workplace would predispose towards using a car rather than a bus. Either way we expect a positive association between mode and distance. The frequencies from the survey were:-

Table 7 : Distance and mode

	Bus	Car
3 km and over	64	105
Under 3 km	46	70

N = 285 (users of other modes omitted)

and the zero-order coefficient works out at +0.04 - no association. Before rejecting the hypothesis outright, it is worth again searching for a reason for the inconsistency. Why should the individual who has further to travel choose to travel by car? It may be that, amongst other reasons, he makes a material saving in time. If this were not so, the additional expense of the car journey may not be worthwhile. Let us therefore control for the time taken for the journey. The frequencies are now:-

Table 8 : Distance and mode and time

		Bus	Car
20 mins and over	3 km and over	57	55
	Under 3 km	27	4
under 20 mins	3 km and over	7	50
	Under 3 km	19	66

and the partial is +0.56, highly significant. By controlling for time we have turned an apparently insignificant relationship into a strong positive one in line with the original hypothesis (position 2 on Figure 2). The two conditionals are +0.73 (for long journeys) and +0.34 (for short journeys), showing that time is particularly critical if it is longer rather than shorter. The effect of time is to suppress the genuine relationship between mode and distance. To understand how this comes about, observe the different ways time divides the 4 cells of Table 7 into Table 8.

(D) No effect

Such an outcome arises when the zero-order and the partial are both significant and of the same magnitude. Controlling for T does not alter the strength of the association. On the face of it, this may seem an uninteresting result, but as we shall see in Section IV (ii) it may arise in some contrasting ways.

For an example of no effect, consider the relationship between time and distance travelled to work, shown to be strong and positive ($Q_{XY} = +0.77$) in the two-variable worked example. We might hypothesise that one of the reasons this association is not even stronger is the different occupation (and by implication, social status) of the individuals involved. It may be that higher occupational status allows greater command of commuting resources which in turn may allow a greater distance to be exchanged for a given time, producing a stronger relationship. Likewise with lower occupational status, more time may more regularly be required for a given distance, also producing a stronger relationship. Controlling for occupation elaborates the frequencies of Table 2 into:-

Table 9 : Distance and time and occupation

		Under 20 mins	20 mins and over
Non-manual occupations	3 km and over	31	50
	Under 3 km	46	15
Manual occupations	3 km and over	46	96
	Under 3 km	130	33

In this case the partial is +0.77, identical to two places with the zero-order. We are forced to conclude that controlling for occupation has in fact no effect on time and distance (position 3 on Figure 2).

(E) Twilight

Two small areas of the graph now remain. These cover situations where the magnitude of the partial is significantly smaller than the zero-order, but yet not sufficiently small to be non-significant. Davis (1971) calls this the twilight region: it may be construed more explicitly as a region of partial explanation, for the XY relationship weakens but does entirely disappear when T is controlled. Such outcomes are perhaps more common though less helpful to theorising than pure explanations. As it happens, none of the possible effects on the journey-to-work relationships produced a straightforward example so an illustration is omitted here.

(F) Specification

The final possible outcome does not appear on the outcome diagram. In some ways it is the nigger in the woodpile to a neat exposition of the effect of a third variable. It is best introduced by an example: among the respondents to the journey-to-work survey was noted a relationship between house type and tenure. People living in detached and semi-detached houses tend to be owner-occupiers. When occupation was controlled, the relationship disappeared, suggesting quite reasonably, that occupation explained the relationship:-

Table 10 : House type and tenure and occupation

		Renting	Owner-occup.
Non-manual occupations	Det. and semi-det.	11	68
	Terrace, etc.	36	49
Manual occupations	Det. and semi-det.	34	14
	Terrace, etc.	149	119

The zero-order is +0.33 and the partial -0.01. However when the conditionals are examined they are found to differ greatly. That for non-manual workers is +0.64, for manual workers -0.32. Averaging those into the partial simply conceals the true situation. With specification, the effect of T cannot be determined by consideration of the zero-order and partial: it is in fact doubtful if further progress can be made without dividing the population into two groups on the basis of the specifying variable and continuing the search for causal connections within each. At the same time, the reason for the specification may become apparent if the frequencies are reconsidered, especially if reduced to row percentages. In the example above, the large number of manual owner-occupiers who live in terraced houses is striking and requires elaboration (probably older property).

There are two likely origins of specification. Firstly, there may be perfectly valid substantive reasons for the different relationships to exist. Sex is a common specifying variable, especially in behavioural studies, because men and women may respond very differently to the same stimulus, or the laws or norms regulating a particular action may vary according to sex. However specification can arise, secondly, from a weak classification of a variable. The obvious example is when one cell of the 3-way table is void. By definition, one conditional must be ±1.00 and specification may follow. In these circumstances it is best to redefine the basis for the dichotomisation (if this is possible).

If specification applies, it is not legitimate to proceed to elaborate the causal system between the 3 variables (the purpose of the next two sections), since specification implies at least two different systems operating within the data. We now give two methods of checking for specification.

Exact test of significance of difference in conditionals

Goodman (1965) gives the confidence limits for the difference between two conditionals (cell frequencies as Table 3) as:

$$\text{LIMITS}_{\text{DIFF}} = Q_{XYT} - Q_{XYNOTT}$$

$$\pm Z \sqrt{\frac{(1 - Q_{XYT}{}^2)^2(\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D})}{4} + \frac{(1 - Q_{XYNOTT}{}^2)^2(\frac{1}{E} + \frac{1}{F} + \frac{1}{G} + \frac{1}{H})}{4}} \quad ..(9)$$

where Z = 1.96 at the 5% significance level
Z = 2.33 at the 1% significance level
Z = 3.09 at the 0.1% significance level

If zero lies between the two limits, the difference is not significant and specification does not apply. To be more precise, the null hypothesis that the two conditionals are drawn from a common population has to be accepted. If zero lies outside the limits (i.e. they have the same sign), specification is occurring.

Rule-of-thumb check for specification

Since the exact test is somewhat cumbersome in application, let us make some simplifying assumptions that will indicate the likely magnitude of the difference between conditionals that should be sought for specification in a given sample.

An even split between the T's and NOTT's (as well as on X and Y) will minimise the confidence interval, so take

$$A = B = C = D = E = F = G = H = N/8$$

This implies $Q_{XYT} = Q_{XYNOTT} = 0$

and therefore sets the limits of the null hypothesis of no difference when both conditionals are zero. This may sound crazy since we are unlikely to be testing for specification in such circumstances, but it does give a limiting value for the difference. Eauation (9) is thus reduced to:

$$\text{LIMITS}_{(Q_{XYT} = Q_{XYNOTT} = 0)} = \pm Z.\frac{4}{\sqrt{N}}$$

For varying sample size we may now derive the limiting magnitude of differences between conditionals. These are given in Table 11, together with limits for the zero-order under similar assumptions.

Table 11 : Magnitude of differences in conditionals

Sample size (N) (N.B. whole sample incl. both T's and NOTT's)	5% confidence limits	
	Difference between conditionals $Q_{XYT} = Q_{XYNOTT} = 0$	Zero-order $Q_{XY} = 0$
100	0.78	±0.39
150	0.64	±0.32
200	0.55	±0.28
300	0.45	±0.23
400	0.41	±0.20
500	0.35	±0.18
600	0.32	±0.16
800	0.28	±0.14
1000	0.25	±0.12

Note that the confidence limits of the zero-order under the same assumptions are exactly half and their values are given by the two vertical lines drawn on the outcome diagram (Figure 2) for the appropriate sample size.

A simple check for specification, therefore, is to read off the value of the right-hand vertical line on your outcome diagram, double it, and see whether the conditionals differ by more than this amount.

Circumstances where this check may be inadequate are:-

(1) If the split on T is extreme.
(2) If either conditional is large (approaching ±1.00)

In either case, or if the difference is near the rule-of-thumb value use the exact test to be sure if specification applies.

If it does apply, remember to consider whether there is a good theoretical reason for different relationships within the specifying variable, or whether it is the dichotomisation of the specifying variable which is at the root of the problem.

IV THREE VARIABLE MODELS

(i) Causal assertions

Thus far we have proceeded to elaborate the effect which a third variable T may have on the relationship between X and Y. By so doing we can check, within the confines of this three variable system, whether or not this relationship is spurious. We may also wish to observe the effect of Y on the XT relationship and X on YT to complete our investigation into the connections between all three variables. This amounts to being able to define the true causal connections operating within the three variable system as opposed to the simple pairwise associations.

A moment's consideration will reveal that this task is more complicated than it seems. Each of the three relationships may or not exist; if it exists it may be positive or negative; whether it is positive or negative, the dependence of one variable on the other may work one way, or the other way, or both ways (mutual dependence). There are therefore seven possible connections between 2 variables and $7^3 = 343$ between 3. To identify the correct model by inspection may be a rather hit-and-miss operation.

A much more effective (and in some cases, the only) way to proceed is to use our a priori knowledge of the likely existence, sign and direction of the relationships to set up a model which can be tested against our data. We impute a set of causal assertions and proceed to validate them. If we are wrong we reconsider the assertions and test the validity of the new model.

The key element in this procedure is the identification of the effect of a particular assertion if it is true. We have to be able to predict the behaviour of the coefficients of association under our a priori assumptions. If the coefficients do behave as we expect them to we may consider the assertions true. This is Simon's (1954) rationale in discussing spurious correlation.

In IV (iii) below, we will show that the behaviour of the coefficients is systematic and that certain rules may be developed to predict this behaviour from the assertions. However, first some conventions for representing causal assertions are given (Table 12) and the more common three variable systems are reviewed to provide a context for the discussion.

Table 12 : Conventional representation of causal assertions

Assertions about X and Y (apply to XT,YT as appropriate)	Existence, sign and direction (causal diagrams)	Sign (predictions)
No causal connection[1]	X Y	0
Positive relationships[1]		
X depends on Y	$X \longleftarrow Y$	+
Y depends on X	$X \longrightarrow Y$	
X,Y mutually dependent	$X \longleftrightarrow Y$	
Negative relationships[1]		
X depends on Y	$X \dashleftarrow Y$	-
Y depends on X	$X \dashrightarrow Y$	
X,Y mutually dependent	$X \dashleftarrow\dashrightarrow Y$	

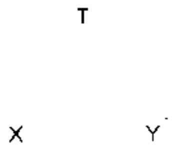
[1] Significant or not at whatever significance level is adopted.

(ii) Some common three-variable models

In all the models below, X, Y and T may be rearranged without altering the basic structure.

A. NO ASSERTIONS:

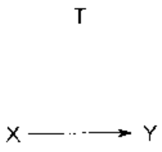
(1) Complete independence



As with no diagnosis the outcome of introducing T, this is the rather uninteresting case where there are no assertions of dependence between any of the variables. We would expect none of the variables to affect the relationships between the others.

B. ONE ASSERTION:

(2) Single relationship

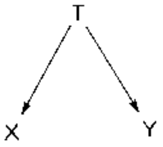


In this case we assert that Y is dependent on X and that T is related to neither X nor Y. We would expect therefore that T will not affect the XY relationship.

C. TWO ASSERTIONS:

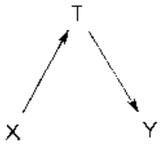
If there are two true causal connections between the variables, it is highly probable that the third link will appear to exist also. For example if T depends on X and Y depends on T, then Y should at least appear to have a connection with X even though there may be no justification for imputing causation directly. This is of course our old friend spurious correlation and it is here in the various types of two-assertion model that we are likely to find him lurking.

(3) Antecedent T



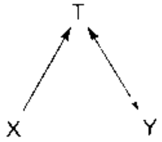
This model produces the classic case of spurious relationship. If both X and Y depend on T, then there is no way causation may be imputed indirectly between X and Y.

(4) Intervening



In this case a spurious relationship between X and Y arises from indirect causation through T.

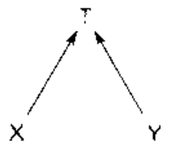
(5) One or two mutual



In many two-assertion models either or both assertions may be mutual and this tends to reduce the clarity of the indirect causal connection. Nevertheless these cases are still examples where a spurious third link will appear.

Cases 3 to 5 have one feature in common. We would expect the spurious link between X and Y to disappear when T is controlled, i.e. they are all cases of explanation. Note that a similar situation would occur if both assertions were negative. One further two-assertion model does not quite fit this pattern and should be treated separately:

(6) Consequent T

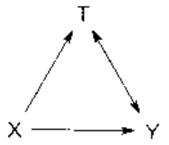


The problem here is that since T depends on both X and Y, it cannot by definition have any effect on their relationship. Any attempt to control for T is therefore theoretically invalid and the diagnosis should be discounted (the case of spurious explanation of a spurious relationship?).

D. THREE ASSERTIONS:

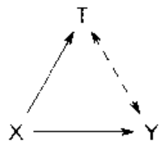
Three assertion models are those in which both direct and indirect causation is asserted. There are four common types: in two of these the indirect effect through T serves to reinforce the direct connection between X and Y; in the other two the direct effect is reduced or suppressed by the indirect.

(7) Reinforcing system (all positive assertions)



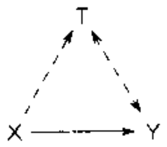
The positive relationships between X and T and Y and T provide an indirect connection between X and Y. Controlling for T will by removing this effect reduce but not eliminate the XY relationship. This case therefore corresponds to partial explanation by T. The other two relationships in the model would be likewise affected.

(8) Suppressing system (two positive assertions)



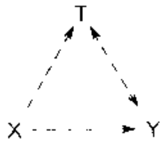
Here the two assertions involving T are opposite in sign and will therefore tend to cancel one another out. The indirect effect is thus to weaken the true connection between X and Y. Controlling for T will strengthen the relationship

(9) Reinforcing system (one positive assertion)



The double negative assertions involving T produce a reinforcing system similar to (7)

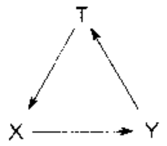
(10) Suppressing system (no positive assertions)



If all the assertions are negative another suppressing system is implied.

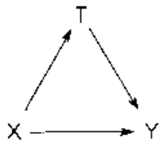
Note that in each of these four cases, one assertion is taken to be mutual. If this were not so, two further models emerge.

(11) Feedback system



This model is in fact impossible unless the time sequence of X, Y and T is ignored. Causal connections normally imply a sequence of events - a temporal ordering of cause and effect.

(12) Reinforcing or suppressing system with a consequent variable



This is a sub-type of all models 7-10 in which one variable is a consequence of the other two (i.e. has two one-way arrows running into it). As in model 6, the effect of this variable must be discounted and the system regarded as suppressing or reinforcing only on the basis of the indirect effects of the other two.

(iii) Validation of three-variable models

Now let us attempt to provide some rules whereby the validity of our assertions based on *a priori* knowledge may be tested. In order to elaborate the system fully we need to consider two things: firstly whether our assertions are individually true or not (the direct effect); and secondly whether the combination of assertions produces the appropriate indirect effect. To predict the direct effect we use the partial coefficient. For the indirect effect we compare the partial and differential.

A. The direct effect

We have already shown that partial may be regarded as a measure of the true association between two variables (Section III(iii)). If an assertion is true, i.e. there is a direct causal connection, then the partial must correspond to the assertion in both existence and sign. There is, however, one important exception to this - consequence. Controlling for a consequent variable will produce a spurious partial which we should ignore. Thus the first two rules for predicting the size and sign of coefficients from our *a priori* assertions may be put forward, using the sign convention for prediction given in Table 12.

Rule 1 : *The partial will have the sign of the causal assertion unless the test variable is consequent* (i.e. has two one-way arrows running into it).

Rule 2 : *If the test variable is consequent, the sign of the partial may be ignored when validating the model.*

B. The indirect effect

The indirect effects work through the causal connections involving the third variable. Let us consider the outcome of three possible arrangements of the assertions involving T - one or both zero; one positive and one negative; both positive or negative.

If one or both of them is zero, then an indirect connection cannot be achieved and the third variable will have no effect on the relationship, i.e. if XT or YT or both are zero, then P = Z (within confidence limits rather than exactly equal).

If both assertions are non-zero but opposite in sign, then the effect of the third variable will be to make the true relationship less positive or more negative than it appears when T is not considered. This means that T either:-

(1) weakens a true positive relationship between X and Y; or
(2) explains no relationship between X and Y; or
(3) strengthens a true negative relationship between X and Y, i.e:-
 if XT and YT are opposite in sign, then P>Z.

Finally, if both assertions are non-zero and have the same sign, the effect of T will be to make the XY relationship more positive or less negative than when T is not considered. The indirect effect of T is to either:-

(1) strengthen a true positive relationship between X and Y; or
(2) explain no relation between X and Y; or
(3) weaken a true negative relationship between X and Y, i.e:-
 if XT and YT have the same sign, then P<Z.

We have here no more than an elaboration of the effect of T - none, suppression or explanation. In Section II(iii) we gave a graphical method of diagnosing this effect making certain assumptions. Let us now be more precise. One of the assumptions was that the split on T should be as even as possible. If this is not so, and in particular if the number of pairs tied and differing on T is not equal, the numerical difference between the zero-order and the partial will have no exact interpretation, since the zero-order is the *weighted* average of the partial and the differential. To guage the effect of I more precisely we should therefore use the numerical difference between the partial and the differential (D-P). The zero-order will of course lie between the two, but its exact location will depend on the weighting.

Summarising the possible combinations of assertion, we get:-

Assertions involving T		Effect of T	Coefficient relations	Sign of D-P
One or both zero	(00,0+,0-,+0,-0)	None	P = Z = D	0
both non-zero, opposite sign	(+-,-+)	Suppression	P > Z > D	-
both non-zero, same sign	(++,--)	Explanation	P<Z<D	+

from which we may derive a straightforward algebraic rule for predicting the effect of a third variable.

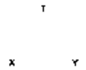
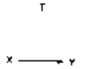
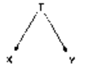
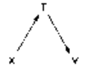
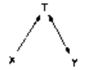
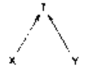
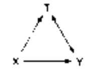
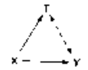
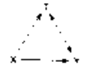
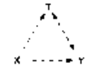
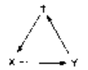
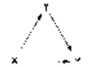
Rule 3 : *To predict the effect of a test variable, predict the sign of D-P by multiplying the signs of the causal assertions involving the third variable.*

In practice, Rule 3 turns out to be rather weak, particularly with 3 assertion models. If the indirect effects are relatively small, the degree of reinforcement or suppression will be small also and the D-P values not as large as the rule suggests. This, however, is usually easily spotted. For various reasons (cf. Davis (1971), Appendix), this rule should not be regarded as a strict function of the outcomes.

These three rules allow us to set out predictions (or expectations) about the coefficients from our hypothesised model. In Table 13 these are given for the 12 common causal systems previously discussed.

The partial and differential for each pair of variables may now be calculated and checked against the predictions. One tricky little problem remains: when is the value of the partial or D-P large enough to say that it is non-zero? Unfortunately the sampling distributions of the partial and differential are not known, so there is no exact significance test. Davis (1971) suggests ±0.10 as critical values, but as a rule-of-thumb this is only really satisfactory with largish samples (over 1000). We may slide around this issue by assuming that the partial and differential have approximately the same sample distribution as the zero-order - an assumption already made in the graphical solution to the effect of T. We can use the limits graph to read off the limits of the partial (plotted on the Q-scale, Figure 1)

Table 13 : Predictions of coefficients in common three-variable models

			Predictions					
			Direct (Partials)			Indirect (D - Ps)		
	Model	Type	XY	XT	YT	XY	XT	YT
(1)		Complete independence	0	0	0	0	0	0
(2)		Single relationship	+	0	0	0	0	0
(3)		Antecedent T	0	+	+	+	0	0
(4)		Intervening T	0	+	+	+	0	0
(5)		One or two mutual	0	+	+	+	0	0
(6)		Consequent T	Sp	+	+	+	0	0
(7)		Reinforcing - all positive	+	+	+	+	+	+
(8)		Suppressing - 2 positive	+	+	-	-	-	+
(9)		Reinforcing - I positive	+	-	-	+	-	-
(10)		Suppressing - no positive	-	-	-	+	+	+
(11)		Feedback	+	+	+	+	+	+
(12)		Reinforcing - consequent Y	+	Sp	+	+	+	+

Sp spurious

for our particular sample, to provide us with checks that the calculated values match our predictions.

Check 1 : If zero lies outside its limits (i.e. they have the same sign), the partial is significantly non-zero at the 5% level (the sign of the partial is as calculated).

Check 2 : If the value of the differential lies outside the limits of the partial, the value of D-P is significantly non-zero at the 5% level (the sign is given by the subtraction).
Both these checks are subject to the caveats of the graphical method.

If the coefficients calculated correspond to the predictions we may say that our a riori causal assertions are valid as to existence and sign. If not, we will have to think anew. The procedure does not allow, however, the direction of the assertion to be itself validated, hence the importance of being clear about this beforehand.

V CAUSAL ANALYSIS

From a research application point of view much of the foregoing discussion is necessarily obfuscated by considerations of statistical technique and operational logic. The purpose of this section is to summarise the procedure by suggesting a strategy for causal enquiry and working through two examples - one straightforward, the other less so.

(i) A strategy for causal analysis

The procedure is outlined in Figure 3, where reference is made to the sections relevant to each heading. While this diagram is largely self-explanatory it is worthwhile emphasising one or two points.

(A) Hypothesis : Do try to start with a well thought-out hypothesis. It may be that on occasion Yule's Q is used to dredge survey data for possible relationships: being so simple to calculate it can be quite efficient in this role. Such a strategy, however, is often a substitute for hard thinking about likely cause and effect relationships. Too much geographical research has a weakness for number-crunching instead of grafting for conceptual and theoretical utility. Start by defining the variable most in need of explanation and select its most likely explanator. As the analysis proceeds, the variable set and the matrix of connections can be enlarged step by step.

(B) Introducing a test variable : Whether or not the initial relationship proves significant, try controlling for other variables. If you fail to do this, you may be guilty of hanging your hat on the wrong peg if the relationship is spurious, or conversely rejecting a perfectly valid relationship that happens to be hidden by something else. If your initial hypothesis was soundly conceived, have faith in it by trying to control for yet more variables.

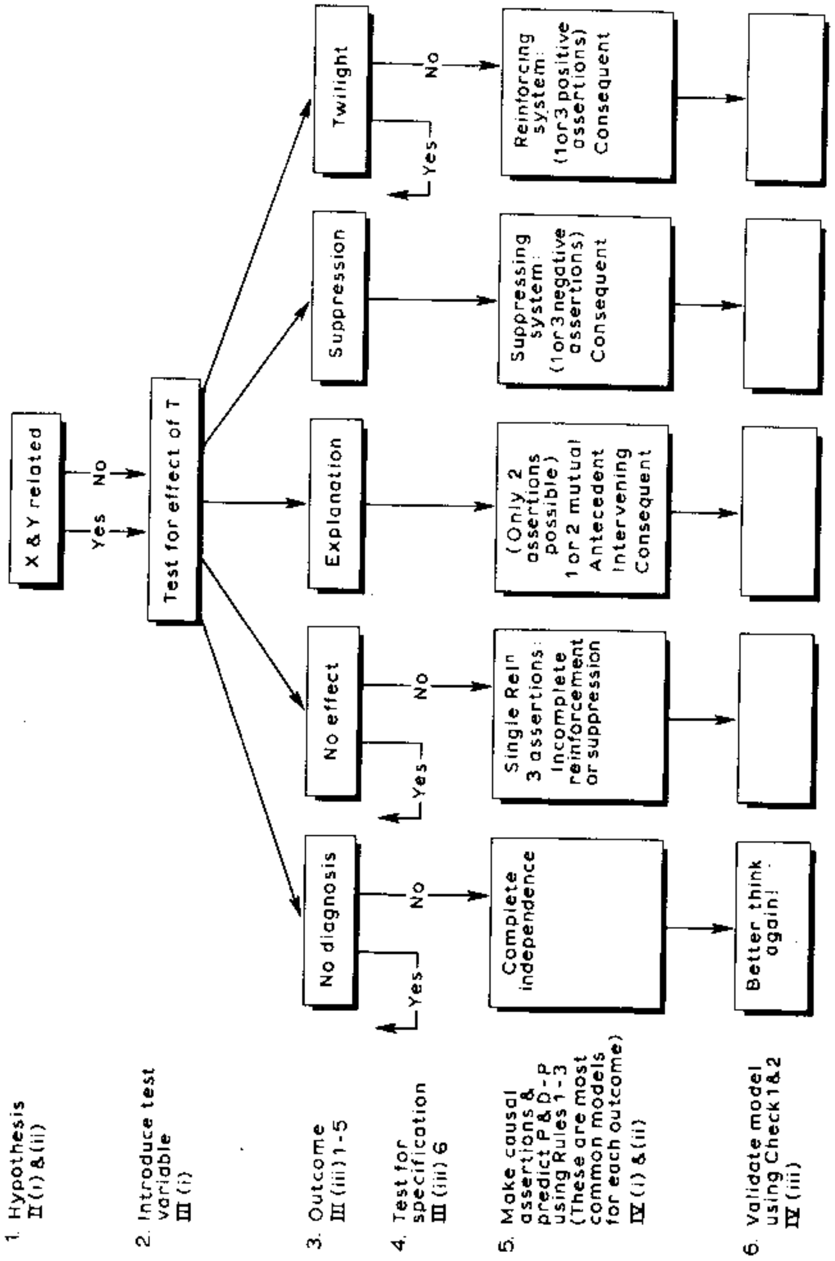

X & Y related
Yes
No
Test for effect of T
No diagnosis
No effect
Explanation
Suppression
Twilight
Yes
No
Yes
No
Yes
No
Complete independence
Single Relⁿ 3 assertions: Incomplete reinforcement or suppression
(Only 2 assertions possible) 1 or 2 mutual Antecedent Intervening Consequent
Suppressing system: (1or3 negative assertions) Consequent
Reinforcing system: (1or3 positive assertions) Consequent
Better think again!
1. Hypothesis II (i) & (ii)
2. Introduce test variable III (i)
3. Outcome III (iii) 1-5
4. Test for specification III (iii) 6
5. Make causal assertions & predict P & D-P using Rules 1-3 (These are most common models for each outcome) IV (i) & (ii)
6. Validate model using Check 1&2 IV (iii)

Figure 3: A strategy for causal analysis

(C) Testing for specification : remember that specification implies two models within the data - one law for the rich and another for the poor. If you find specification, there is no point in proceeding beyond this point with the data as it stands since the method cannot cope with two simultaneous models. If you can postulate some theoretical reason for the two models, split the sample on the specifying variable and continue the search for causation in each part separately. If there is no theoretical reason, try redefining the specifying variable.

(D) Common models : an attempt is made here to suggest the most likely models under each outcome. It is not mathematically impossible for some of these models to appear under a different outcome, but there are good grounds for not expecting this. The three-assertion (all or one positive) systems are the most controversial. To be properly reinforcing, all the relationships should be reinforced by the indirect effects of the remaining variable, and should therefore appear in the twilight or partial explanation region. However, if only one or two of the variables have a reinforcing effect, the others may be diagnosed as having no effect. Such incomplete reinforcement (or suppression, if one or all the assertions are negative) is by no means exceptional.

(ii) Three-variable worked examples

The first example is straightforward: the second is chosen to highlight some of the problems of validation.

(A) Example 1 : Distance, time and mode of transport

Table 14 : Example 1 - frequencies

X = Distance
Y = Time
T = Mode 1
	N = 463	Less than 20 mins (NOTY)	20 mins and over (Y)

Travel by car (T)	3 km and over (X) Under 3 km (NOTX)	50 66	55 4
Travel by other mode (NOTT)	3 km and over (X) Under 3 km (NOTX)	27 126	91 44

The initial hypothesis is that time and distance travelled to work are related. The zero-order is significant at the 5% level (Q_{XY} = +0.77).

We now speculate that this relationship may be affected by mode of transport on the grounds that some modes will be faster than others. In this example we take Mode 1, i.e. whether or not the journeys are by car.

From the frequencies in Table 14, the partial is calculated and found to be significant also, but higher than the zero-order ($Q_{XYTIEDT}$ = +0.83).

When plotted on the outcome diagram with the limit curves for N = 463, the combination of partial and zero-order is found to lie in the suppression region. By not considering mode of transport we significantly underestimate the true strength of the relationship between distance and time.

Next we check that specification does not apply (very unlikely with suppression). We find that both conditionals are higher than the zero-order (Q_{XYT} = +0.90; Q_{XYNOTT} = +0.81) but that the difference between them is too small for specification. We may also check at this stage that there is no specification on the other variables : there is none (see Table 16).

Now we can proceed to make the causal assertions about the three relationships on a priori grounds, viz:-

Time depends on distance : we choose, through our choice of residence or workplace, to traveT a certain distance to work and the time taken will depend on that distance (amongst other things). It is clear we cannot assert the opposite - that distance depends on time - since this would demand that the journey be completed after a certain time. The relationship should be positive.

Distance and mode are mutually related : we may say that we choose to work further from home because of the freedom from the constraints of public transport conferred by travelling by car. Conversely we may be encouraged, even forced, to travel by car because of a distant and inaccessible workplace. The causal connection must therefore be regarded as mutual: it is also positive since we would expect the longer trips to be more likely to be made by car.

Time and mode are mutually related : since travelling by car is generally quicker than other modes (train and tube not being relevant for this particular survey), it is likely that time depends on mode. We may also conceivably choose to travel by car because we have less time available for the trip. The mutuality here is perhaps more debatable than in the previous case, but will not affect the validation of the model. The relationship will, however, be negative since car trips should take less time than others for a given distance.

These assertions may be summarised by the model given in Figure 4.

Figure 4:
Example 1 - assertions

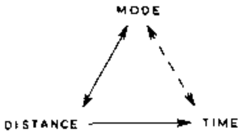


Table 15 : Example 1 - predictions and validation

	Partials			D - Ps		
	XY	XT	YT	XY	XT	YT
Predictions	+	+	-	-	-	+
Computed values	+0.83	+0.61	-0.58	-0.13	-0.41	+0.50
Validity	Yes	Yes	Yes	Yes	Yes	Yes

Note that by having 2 positive assertions, it corresponds to the suppressing system suggested as the most like model from the suppression region. There is no consequence involved in this model.

On the basis of the assertions, predictions are made by applying Rules 1 and 3 (Section IV (iii)). These are given in Table 15 where they are compared to the computed values of the coefficients (listed in full in Table 16) using the two checks also given in Section IV (iii).

Table 16 : Example 1 - Yule's Q coefficients

	Distance and Time (XY)		Distance and Mode (XT)		Time and Mode (YT)	
Zero-orders		+0.77		+0.37		-0.27
Partials		+0.83		+0.61		-0.58
Differentials		+0.70		+0.20		-0.08
Conditionals	T	+0.90	Y	+0.74	X	-0.51
	NOTT	+0.81	NOTY	+0.56	NOTX	-0.70

All the coefficients behave as predicted by the model and we may accept the validity of our assertions. Distance, time and mode are linked in a suppressing system which effectively conceals the true strength of all the causal connections between the variables.

(B) Example 2 : Distance, mode and head of household

In this example, we are proceeding to elaborate further the connection established in Example 1 between distance and travelling by car (Mode 1) on the grounds that heads of households may effectively commandeer the household car.

The zero-order (Q_{XY} = +0.37) is significant though not large. When plotted with the partial ($Q_{XYTIEDT}$ = +0.34), we find that the outcome lies in the no effect region of the diagram. The relationship between distance and mode is not affected by consideration of whether or not individuals are heads of households. Again there is no specification (conditionals in Table 19).

Table 17 : Example 2 - frequencies

X = Distance
Y = Mode
T = Head of household
	N = 463	Other modes (NOTY)	Travel by car (Y)

Heads of house-holds (T)	3 km and over (X) Under 3 km (NOTX)	64 74	72 45
Other members (NOTT)	3 km and over (X) Under 3 km (NOTX)	54 96	33 25

A priori assertions about the relationships might be as follows:-

Distance and mode mutually dependent (as Example 1)

Distance depends on head of household: there can certainly be no determination of head of household by any journey-to-work variable. Heads of households may, however, be obliged to travel further to seek work, i.e. a positive relationship.

Mode depends on head of household: heads of households may be more likely to travel by car because of their ability to command use of the household car (if the household has one - a further assertion that could be tested).

So we obtain the model in Figure 5 from which the predictions in Table 18 are derived.

Figure 5:
Example 2 - assertions

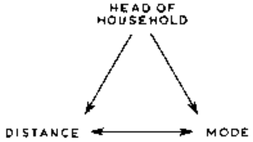


Table 18 : Example 2 - predictions and validation

	Partials			D - Ps		
	XY	XT	YT	XY	XT	YT
Predictions	+	+	+	+	+	+
Computed values	+0.34	+0.18	+0.35	+0.06	+0.09	+0.05
Validity	Yes	border line	Yes	No	No	No

Table 19 : Example 2-Yule's Q coefficients

	Distance and Mode (XY)		Distance and head (XT)		Mode and head (YT)	
Zero-orders		+0.37		+0.23		+0.37
Partials		+0.34		+0.18		+0.35
Differentials		+0.40		+0.27		+0.40
Conditionals	T	+0.30	Y	+0.10	X	+0.30
	NOTT	+0.40	NOTY	+0.21	NOTX	+0.40

There is clearly something wrong with the assertions, for the expected effects (D - Ps) all fail to validate. However the pattern of failure does give a clue to possible amendments to the assertions. We can see that the XT relationship (between distance and head) is borderline. Suppose we now assert that it does not exist, that being head of household does not determine how far is travelled to work. Note that if we have to abandon one of the assertions, this is the most likely candidate. The new model will be:-

Figure 6:
Example 2 - new assertions

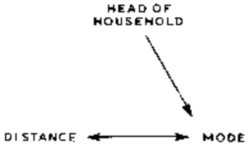


Table 20: Example 2 - new predictions and validation

	Partials			D - Ps		
	XY	XT	YT	XY	XT	YT
Predictions	+	0	+	0	+	0
Computed values	+0.34	+0.18	+0.35	+0.06	+0.09	+0.05
Validity	Yes	border line	Yes	Yes	No	Yes

Apart from the borderline XT partial, the remaining disagreement is the XT D-P, though this is the largest of the three. We may attempt to revise the model still further, but no closer fit than this will be obtained. We have here some weak relationships resulting in a weak model. This example illustrates how we may amend assertions by re-examining a priori assumptions to produce a better fit with empirical findings. It is both a strength and a weakness of this inferential method that it does not necessarily produce a single clear-cut model which matches the data. If it did it would be neater, but might encourage us to think that the real world is simpler than it is.

VI CONCLUSIONS

(i) Some do's and don'ts

One of the important objectives of this monograph is to promote, through its methodological discussion, a critical evaluation of the basis on which conclusions are reached about cause and effect in geographical relationships. It is never easy to be definitive on this subject, but it is worthwhile re-emphasising some of the crucial phases in the process of making causal inferences.

(1) Define your problem.

(2) Never take an empirically-derived relationship at its face value.

(3) Ensure that the theoretical grounds for asserting causation are thoroughly explored before erecting a causal model.

(4) Understand your data and its limitations; define your variables carefully and make sure that, within the demands of substantive comparability, they vary.

(5) Never allow statistical neatness to obscure the meaning of results.

Most of this is just sound commonsense, but it is amazing how quickly commonsense can evaporate when confronted with a battery of survey results.

(ii) Four or more variables

With 15 coefficients and 343 possible models, 3-variable causal analysis can be quite intimidating. Extend the structure to include one further variable and it becomes positively frightening - 7" models and a proliferation of first- and second-order coefficients. Davis (1971, Ch. 6) outlines some methods for attempting to cope with this situation. Beyond four variables, the problems become fairly intransigent, particularly of operationalisation. Another fact worth bearing in mind is that a simultaneous analysis of 5 variables will push the minimum sample size requirement towards 500.

Very briefly, two strategies are recommended for causal structures involving 4 or more variables. The simpler is to elaborate the system as a series of triangles using Yule's Q, preferably working backwards in causal terms from the variable whose explanation is most desired. The disadvantage here is that indirect effects through 2 or more variables are not accounted for and these may be important. Secondly, the structure may be investigated by the method of path analysis discussed below, using the appropriate measure of association (not Yule's Q in this case).

(iii) Other methods of causal analysis

This monograph has been concerned only with dichotomies. The principles involved, however, are capable of application to other kinds of data. Depending on the level of measurement, the following coefficients are capable of being partialled to control for a third variable:-

Somer's d	nominal variables with more than two categories but only if the categories can be put in some sort of rank order
Kendall's tau	rank order variables
Pearson's r	interval or ratio data

The derivation of partials for these coefficients are given in most standard statistical texts (e.g. Blalock, 1972). The identification of causal structures does, however, become more difficult as the power of the information grows.

An alternative method for more powerful data is that of path analysis (c.f. Duncan, 1971) which seeks to evaluate the degree of explanation provided for a single dependent variable by a set of causally antecedent independent variables. It is a derivative of multiple regression but is capable of extension to other levels of measurement which Hawkes (1971) reviews the potential coefficients.

BIBLIOGRAPHY

Technical and methodological basis

Blalock, H.M. (1972), *social statistics* (2nd Edn) McGraw-Hill, N.Y.

Davis, J.A. (1971), *Elementary Survey Analysis,* Prentice-Hall, Englewood Cliffs, New Jersey.

Goodman, L.A. (1965), On the multi-variable analysis of 3 dichotomous variables. *American Journal of Sociology,* (71), 290-301.

Goodman, L.A. and Kruskall, W. (1954) Measures of association for cross classifications. *Journal American Statistical Association,* (49), 732-764.

Hawkes, R.K. (1971), The multivariate analysis of ordinal measures. *American Journal Sociology, PO,* 908-926.

Kendall, P.L. and Lazarsfeld, P.F. (1950), Problems of survey analysis. In R.K. Merton and P.F. Lazarsfeld (eds) *Continuities in Social Research,* Free Press, N.Y. pp 135-167.

Kendall, M.G. and Stuart, A. (1961), *The advanced theory of statistics,* 2, Griffin & Co., London.

Siegal, S. (1956), *Non-parametric statistics for the behavioural sciences,* McGraw-Hill, N.Y.

Simon, H.A. (1954), Spurious correlation: a causal interpretation. *Journal American Statistical Association,* (49), 467-479.

Geographical applications

Cox, K.R. (1968), Suburbia and voting behaviour in the London Metropolitan Area. *Annals Association American Geographers, (58),* 111-127.

Ferguson, R.I. (1973), Channel pattern and sediment type. *Area,* 5(1), 38-41.

Johnston, R.J. (1971), Social distance, proximity and social contact. *Geografiska Annaler, (56-B),* 57-67.

Mercer, J. (1975), Metropolitan housing quality and an application of causal modelling. *Geographical Analysis,* 7(3), 295-302.

Taylor, P.J. (1969), Causal models in geographic research. *Annals Association American Geographers,* (59), 402-404.

Winsborough, H.H. (1962), City growth and city structure. *Journal Regional Science,* (4), 35-49.

Further Reading

Blalock, H.M. (1964), *Causal inferences in non-experimental research,* Univ. of N. Carolina Press, Chapel Hill.

Blalock, H.M. (ed) (1971), *Causal models in the social sciences,* Macmillan, London.

Duncan, O.D. (1971), Path analysis: sociological examples. In H.M. Blalock (ed) *Causal models in the social sciences,* Ch. 7, Op. Cit.

Hammond, R. and McCullagh, P.S. (1971), *Quantitative techniques in geography,* Oxford Univ. Press, London.

Harvey, D. (1969), *Explanation in geography,* E. Arnold, London.

Pickvance, C.G. (1974), Life cycle, housing tenure and residential mobility: a path analytic approach. *Urban Studies,* (11), 171-188.