INTERPRETATION OF ECOLOGICAL DATA BY PATH ANALYSIS

D. SCOTT

Plant Physiology Division, Department of Scientific and Industrial Research, Palmerston North

Introduction

In ecological research one is often faced with analyzing quantitative data. A dilemma that frequently arises is that, in applying many of the existing methods of statistical analysis, we have to assume that certain variables are statistically independent even though we know from biological or physical considerations that they must interact. The purpose of this paper is to acquaint readers with the relatively new technique of path analysis which seems to have great potential. This is easiest done by relating path analysis to multiple regression analysis.

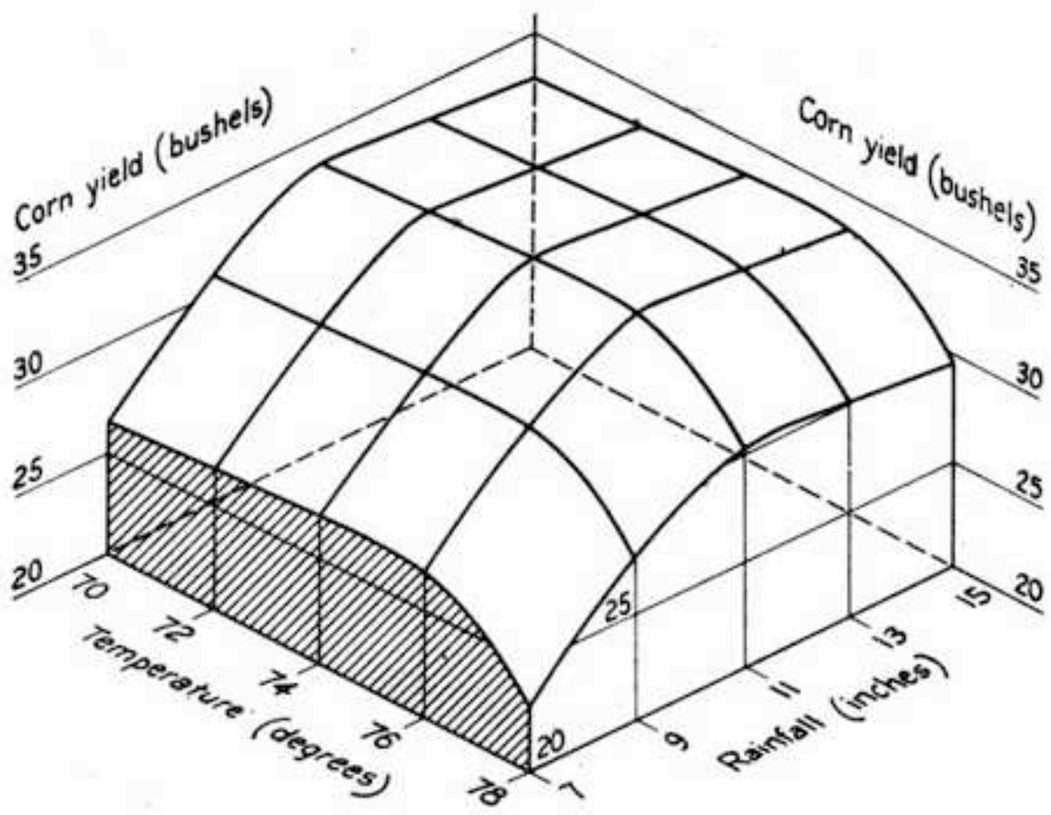


FIGURE 1. *Probable yield of corn for various specific combinations of rainfall and temperature, from multiple curvilinear regression* (from Ezekiel and Fox 1959).

Mulitiple Regression

In multiple regression one variable (dependent variable) is determined as a function of the other variable (independent variable), e.g. the data shown graphically in Figure 1 would be represented by the equation

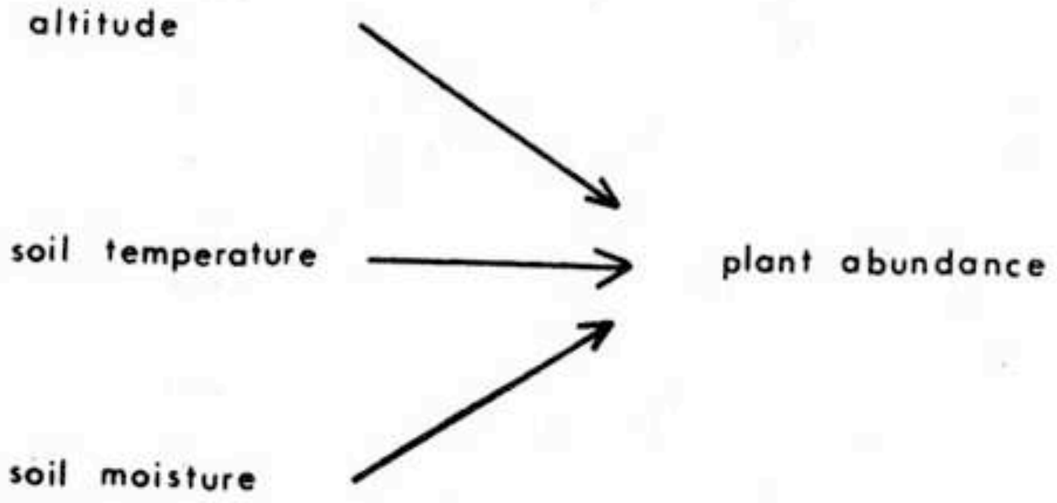
$$y = b_0 + b_1x_1 + b_2x_2$$

where y is corn yield, x_1 and x_2 functions of temperature and rainfall respectively, and where b_0, b_1 and b_2 are fitted coefficients.

The advantages of multiple regression analysis are: (i) that it predicts the value of one variable from measurement of the others; (ii) that there are tests of significance for the b coefficients thereby determining the relative significance of each of the variables in the prediction equation; and (iii) that it can cope with curvilinear relationships between the dependent and independent variables.

The main disadvantage is that the method usually assumes that all factors except the dependent one vary independently of each other.

For instance, in the diagram below, which shows the effect of environmental factors on plant abundance, it might be necessary to assume that there was no interaction between altitude and soil temperature even though we know temperature decreases about 1.8°C. per 1,000 ft.



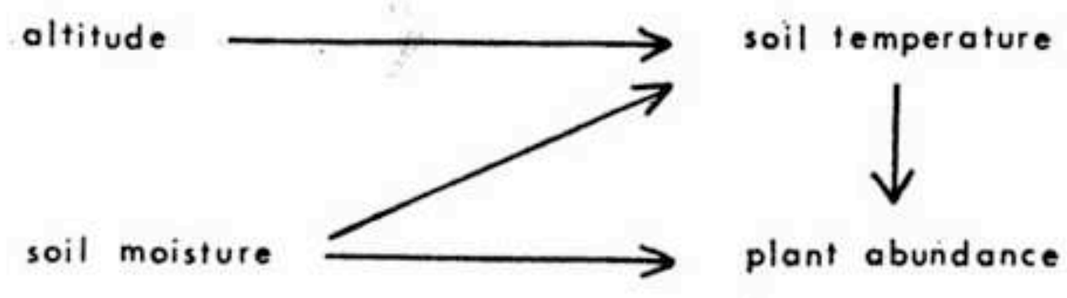
This is a serious limitation when dealing with biological data where interaction between factors is the rule rather than the exception.

Path Analysis

The technique of path analysis has all the advantages of multiple regression analysis, but in addition was developed for dealing with interacting factors. It does this in a manner which gives a convenient method of looking at a problem even if the mathematical aspects of a technique are not used. The method first developed 30 years ago (Wright 1934) for correlation problems in genetics, has only more recently been extended to regression problems (Tukey 1954, Turner and Stevens 1959, Ferrari

1964). Ferrari's paper (1963) is probably the first example of its use with ecological data. The method is also being used in economics under the name of structural analysis.

The starting point is the fact that in any particular problem some of the variables can be designated as causes and others as effects. In this way a network of causal relationships may be built up. Some of the variables may be designated as primary causes in that they are not influenced by changes of other variables within the scheme. For example, the relationship between altitude, soil moisture, temperature, and plant abundance may be represented as follows. Altitude and soil moisture are considered the primary causes. Soil temperature is determined by altitude and soil moisture, and plant abundance by temperature and soil moisture, thus:—



The mathematical theory of the method then allows regression equations to be determined between each of the effects and the primary causes. The important point is that this causal scheme is constructed on the basis of the understanding of the expected ecological relationships either known or hypothesized. Such a scheme may be simple or very complex and the solution will depend on the particular causal scheme chosen and may not always be solvable.

The following are the stages in a complete solution of a problem and are illustrated by an example from the author's work on determining the relationship between environmental factors and the frequency of various species on different sites above timber line in the Tongariro National Park:

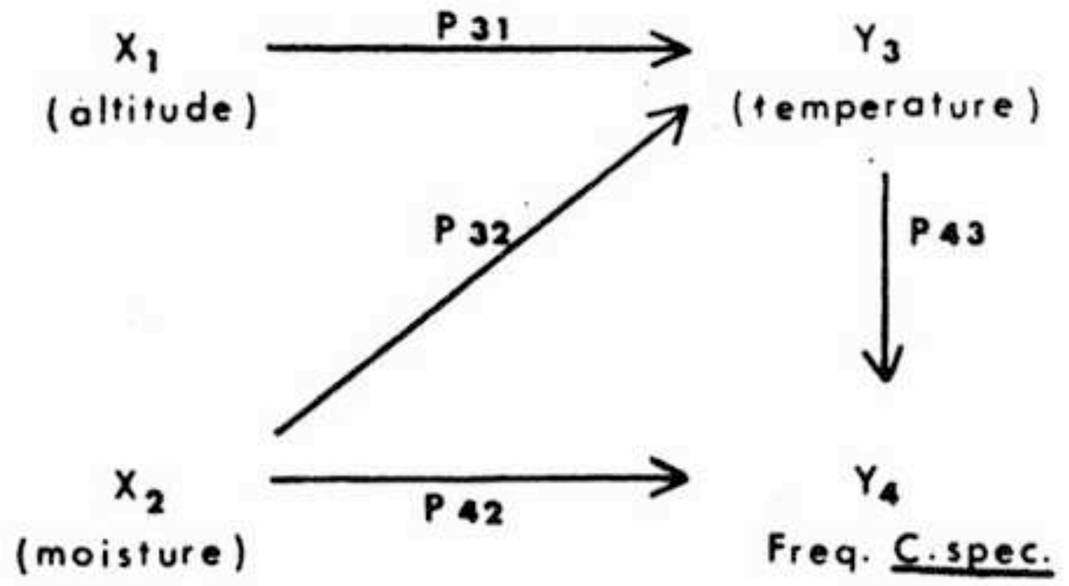
The particular example is the frequency of *Celmisia spectabilis* on 20 different sites in relation to altitude, soil temperature and soil moisture (the latter being determined by gypsum blocks at 10 cm depth.) The soil temperatures and soil moisture were spot measurements taken in the first week in February—the gypsum blocks having been placed in position several months earlier. Altitude has been used as an environmental variable influencing soil temperature in that it is the chief parameter that would be used in calculating the influence of the atmosphere on the solar and long-wave radiation exchange of the site and of the mean air temperature. Data are summarized in Table 1 and give the range of values of the four variables.

TABLE 1. *Mean and range of variables from 20 sites.*

Variable		Mean	Range
Altitude above 4,000′ (ft.)	(x_1)	465	−90 to 1,130
% Moisture in gypsum blocks	(x_2)	120	70 to 206
30 cm soil temperature (°C.)	(y_3)	10.6	8.3 to 12.2
Frequency of *C. spectabilis*	(y_4)	13	0 to 44

(a) *Construct causal network.*

The first stage is to draw a diagram similar to the above showing the known or hypothesized relationships between the variables. It is convenient to designate all primary causes as x's and all effects as y's and to number all x's and y's in sequence. Hence x_1, x_2, y_3, y_4. On this diagram a variable at the head of one or more arrows is interpreted as being a function of just those variables at the tail of the arrows. The p's are called path coefficients and their meaning will become apparent later. For the sake of keeping the demonstration simple the four interactions shown will be assumed the only ones present:—



(b) *Write structural linear equations for each of the effects (y's) in terms of each of their immediately preceding causes.*

This is easiest done by referring to the diagram and for each effect variable writing down equations which include only the variables at the tail of arrows leading to that particular effect. The path coefficients are the coefficients in these equations.

$$y_3 = a_3 + p_{31}x_1 + p_{32}x_2$$
$$y_4 = a_4 + p_{43}y_3 + p_{42}x_2$$

(c) *Structural equations are reduced until each of the effects (y's) are functions of the primary causes (x's).*

This can be done either by substituting equations in each other, or by writing down

the equations by direct reference to path coefficients of the causal diagram and applying several rules which are fully dealt with in most texts on the method.

$$y_3 = a_3 + p_{31}x_1 + p_{32}x_2$$
$$y_4 = (a_4 + a_3.p_{43}) + (p_{31}.p_{43})x_1 + (p_{42} + p_{32}.p_{43})x_2$$

These equations are called the reduced structural equations.

(d) *Multiple regression equations are fitted to the same data to get corresponding equations between each of the effects and primary causes.*

$$y_3 = b_{30} + b_{31}x_1 + b_{32}x_2$$
$$y_4 = b_{40} + b_{41}x_1 + b_{42}x_2$$

In the example:

Temp. = 11.41 −0.00148 (Alt) −0.008 (% moisture)

Frequency = −9.8 −0.0037 (Alt) +0.201 (% moisture)

(e) *The partial regression coefficients of the fitted regression equations are equated with the corresponding compound path coefficients of the reduced structural equations and these series of equations are then solved to obtain the individual path coefficients.*

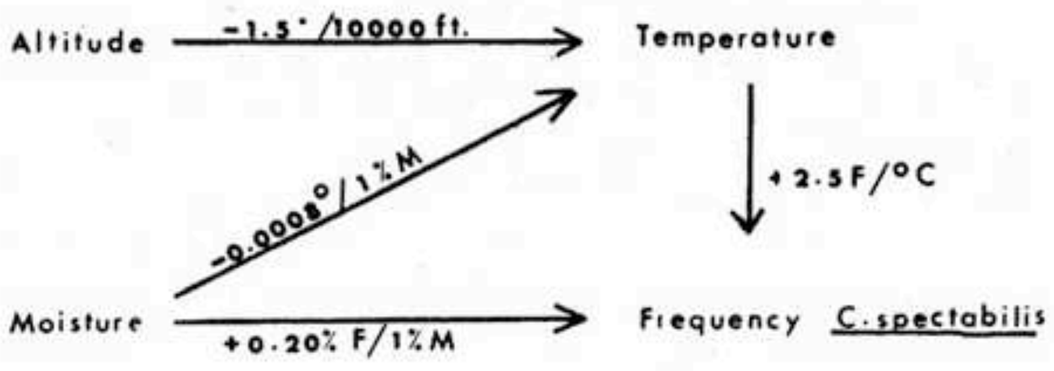
$$b_{31} = p_{31}$$
$$b_{32} = p_{32}$$
$$b_{41} = p_{31}.p_{43}$$
$$b_{42} = p_{42} + p_{32}.p_{43}$$

From which for the example

$$p_{31} = -0.0015$$
$$p_{32} = -0.0008$$
$$p_{43} = +2.5$$
$$p_{42} = +0.20$$

Whether all the path coefficients can be determined from solution of these equations will depend on the particular causal scheme. The confidence intervals of the path coefficients may be determined.

The completed results for the example are as follows:—



In this, the path coefficients are all of the expected sign; temperature decreasing with altitude, temperature decreasing with soil moisture because of the increased specific heat of the soil, and increase of plant frequency with increase of both temperature and moisture. The 1.5° decrease per 1,000 ft. is close to the 1.7 per 1,000 ft. lapse rate given in textbooks on climatology. The 1.5° per 1.000 ft. lapse rate was the only statistically significant path coefficient on the basis of the data used.

The complete solution enables us to do two things. Firstly, we can estimate the value of the effect variables for particular values for the cause variables, e.g. from the equations in section (c) the estimated values of soil temperature and plant frequency when the altitude is 4,070 ft. and the moisture 121% are 11.2°C. and 14 respectively. Secondly, the path coefficients give an estimate of how changes in one factor will influence the others, e.g. with a 20% increase in moisture we would expect a (20 x .20 =) 4% change in frequency caused by direct effect, with an indirect effect of (20 x 0.008 =) 0.2°C. decrease in soil temperature resulting in a 0.2 x 2.5 =) 0.5% decrease in frequency—a net change of a 3½% increase in frequency of *Celmisia spectabilis*.

DISCUSSION

In applying statistics to any problem an investigator has to decide which particular techniques will accomplish what he wants before he becomes embroiled in the computational complexities necessary to achieve this end. Path analysis seems capable of achieving many of the objectives often required in ecological problems. Firstly, the method was developed for studying interacting factors and, if nothing else, the diagram is a convenient way of showing these interactions. Secondly, not only does the method show whether the factor is significant in the statistical sense, but as in regression analysis, it shows the quantitative effects of each factor. Thus, as in the example above, the structural equations can be used to predict the probable values of each of the effects for particular values of the causes and also the path coefficients can show the rates at which one factor changes as the result of changes in other factors.

In the example we have assumed that there was a linear relationship between the variables, but path analysis can also take into account curvilinear relationships. Mention should also be made of two other situations with which

path analysis can deal: (i) where there is feedback betwen the variables and (ii) where some of the variables cannot be measured and yet their path coefficients can be inferred from the equations. These will not be discussed.

In conclusion it is as well to emphasise that the results obtained are dependent on both the underlying causal relationships and on the accuracy and appropriateness of the actual data used. Both of these will be influenced by the investigator's knowledge of the probable causal relationships between the various factors and his skill in measuring the aspects of that factor which is appropriate to the particular problem, e.g. the appropriate temperature measurement for relating to plant response. This will require recourse to single-factor experiments in the field, laboratory or growth cabinets. However the utilization of all such data in a causal network offers one method whereby ecological data could be synthesized.

REFERENCES

EZEKIAL, M., and Fox, K. A., 1959. *Methods of correlation and regression analysis.* 3rd edit., J. Wiley & Son, New York.

FERRARI, T. J., 1963. Causal soil-plant relationships and path coefficients. *Plant and Soil* 19: 81–96.

FERRARI, T. J., 1964. Auswertung biologischer Kettenprozesse mit Hilfe von Pfadoeffizienten. *Biomet. Zeit.* 6: 89–102.

TUKEY, J. W., 1954. Causation, regression, and path analysis. In *Statistics and Mathematics in Biology*, O. Kempthorne, T. A. Bancroft, J. W. Gowen, and J. L. Lush (eds.). Iowa State College Press, Ames.

TURNER, M. E., and STEVENS, C. D., 1959. The regression analysis of causal paths. *Biometrics* 15: 236–258.

WRIGHT, S., 1934. The method of path coefficients. *Ann. Math. Statist.* 5: 161–215.

DRY SPELLS IN NEW ZEALAND AS A FACTOR IN PLANT ECOLOGY

J. D. COULTER

New Zealand Meteorological Service

Although most of New Zealand receives a rainfall which is sufficient in total amount to meet the needs of plants, in most places spells of deficient rainfall of varying lengths have significant effects on vegetation. For example, after a few weeks without rain in summer or autumn the growth of pasture is frequently reduced to such an extent that dairy production declines. In the Kaingaroa area, pine trees were adversely affected after four very dry months in the summer of 1945-46 (Rawlings 1961) although they were apparently not harmed in the 1963-64 summer when two periods of approximately five weeks with negligible rain caused severe farm losses. The distribution of indigenous species and of natural plant communities depends in part on the incidence of dry spells—either directly, on account of differences in drought tolerance; or indirectly, through such effects as increased risk of forest fire when summer droughts are frequent. For full understanding, such topics must be examined in terms of water relationships of individual plants or of plant associations in their particular climatic environment. Nevertheless, a broad-scale comparative study of the occurrence of dry spells should provide a useful background for any detailed ecological study.

Precipitation is very irregularly distributed in space and time, and the statistical treatment of the data poses many problems quite apart from that of obtaining representative samplings in mountain areas. A frequency curve of rainfall totals approaches a symmetrical normal distribution as the unit period of observations is increased. In New Zealand, annual rainfall distributions are only slightly skew, monthly rainfalls show a moderate degree of skewness which increases with the degree of variability, whereas daily falls and the interval between rain days have a highly skew or L-shaped distribution. Representative annual rainfall distributions for a number of New Zealand stations demonstrate that in the driest parts of Central Otago annual rainfall is always well below the year's water requirements of vegetation. In low rainfall areas such as Hastings, the rainfall total is deficient in about 50 per cent. of years, but in most of the country the annual totals almost invariably exceed the year's water need.