Institut für Bodenfruchtbarkeit, Groningen

Auswertung biologischer Kettenprozesse mit Hilfe von Pfadkoeffizienten

Th. J. Ferrari

Einleitung

Die Modelle, die in der biologischen Forschung geprüft und quantitativ geschätzt werden, sind meistens von einfacher Art. Der Forscher bezieht oft nicht mehr als einen oder zwei Faktoren in seinen Versuch. Einerseits hängt dies zusammen mit der Tatsache, daß die Naturwissenschaften auf diese Weise große Erfolge erzielt haben, anderseits fehlen die Methoden, um die Wirkung vieler Faktoren zugleich zu analysieren. In den letzten Jahren tritt die Bedeutung der Forschung über die Zusammenwirkung vieler Faktoren in den biologischen Wissenschaften mehr in den Vordergrund. Dies gilt besonders für jene Disziplinen, in denen die Ergebnisse brauchbar gemacht werden sollen für die Anwendung in der Praxis, und viele Faktoren, die nicht oder sehr schwierig zu ändern sind, eine Rolle spielen. Ein großes Teil der landwirtschaftlichen Forschung, z. B. der Bodenfruchtbarkeit, gehört dazu.

Die Art der bis jetzt ausgeführten Forschung ist gekennzeichnet durch die Methode des Ceteris-Paribus-Prinzips, wobei wie bekannt ein oder mehrere Faktoren künstlich variiert und die anderen als nicht von Bedeutung angenommenen Faktoren möglichst konstant gehalten werden (Experiment mit Eingriff). Dieser Eingriff isoliert das Problem auf einige Faktoren, so daß nur einige Aspekte des Problems untersucht werden und die Synthese erschwert wird. Die Methode bietet wenig Möglichkeiten, wenn die Wirkung von Faktoren, die nicht oder schwierig zu ändern sind, untersucht werden muß. Zur Lösung der gestellten Fragen kann der Forscher nur die Variation benutzen, die in der Natur anwesend ist, um auf diese Weise die Hypothese zu prüfen und quantitativ zu schätzen (Experiment ohne Eingriff). Auch andere Disziplinen wie die Soziologie und die Ökonomie haben solche Probleme. Ein Vergleich zwischen Experiment mit und ohne Eingriff zeigt, daß beide Methoden Vorteile und Nachteile haben, daß es jedoch keine prinzipiellen Unterschiede gibt (Ferrari, 1960). Beide Methoden haben die Möglichkeit, die Richtigkeit der Hypothese negativ zu prüfen.

Die Konsequenz der Methode ohne Eingriff ist, daß der Forscher in der Untersuchung biologischer Zusammenhänge und Prozesse meistens nur die Regressionsanalyse anwenden kann. Gegen die Anwendung des normalen Regressionsmodells für die Analyse der Zusammenhänge zwischen den Veränderlichen bestehen jedoch große Bedenken. Eines der wichtigsten

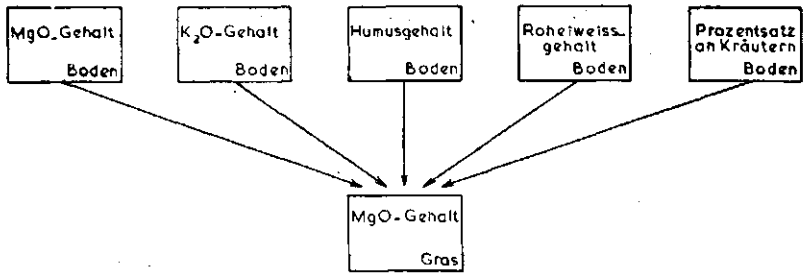

Abb. 1. Regressionsmodell mit MgO-Gehalt des Grases als abhängige Variable, die anderen Variablen sind als kausale Faktoren angenommen

ist mit der Unvollkommenheit und den Beschränkungen des Regressionsmodells verbunden. In diesem Modell wird nämlich angenommen, daß die sogenannten unabhängigen oder erklärenden Faktoren einander nicht beeinflussen. Das heißt, eine Änderung eines Faktors hat keine Änderung eines anderen erklärenden Faktors zur Folge. Eine solche Annahme entspricht jedoch in vielen Fällen nicht der Wirklichkeit. Dies und jenes kann erläutert werden an Hand eines Beispiels aus dem Gebiet der Bodenfruchtbarkeitsforschung (es gibt auch viele aus anderen Disziplinen), wobei die Zusammenhänge zwischen den MgO- und K_2O-Gehalten des Grases einerseits und den kausalen bodenkundlichen und anderen Faktoren anderseits untersucht wurden (SLUIJSMANS, 1962).

Ergebnisse, mit einem Experiment ohne Eingriff erhalten, wurden in einem Regressionsmodell analysiert, wobei die MgO- und K_2O-Gehalte des Grases als die abhängigen Variablen, die MgO- und K_2O-Gehalte des Bodens, der Gehalt an organischer Substanz (Humus), der Prozentsatz an Kräutern*) und der Roheiweißgehalt des Grases als die unabhängigen oder erklärenden kausalen Faktoren angenommen worden sind. Schematisch wird dieses Modell in Abb. 1 gegeben. In diesen Diagrammen wird der angenommene kausale Zusammenhang zwischen zwei Faktoren mit einem Pfeile angegeben, wobei der Effekt an der Spitze gelegen ist. Durch die statistische Auswertung wird der Einfluß der Korrelationen zwischen den unabhängigen Faktoren eliminiert. In dieses Modell wird also die Annahme gelegt, daß eine Änderung z. B. des K_2O-Gehaltes des Bodens nur eine

*) Kräuter = alle „Nichtgräser"

Änderung des MgO-Gehaltes des Grases und nicht eine Änderung des Roheiweißgehaltes des Grases oder des Prozentsatzes an Kräutern zur Folge hat. Diese Annahme ist wie bekannt vermutlich falsch, so daß die Einflüsse des K_2O-Gehaltes des Bodens, des Roheiweißgehaltes und des Prozentsatzes an Kräutern auf den MgO-Gehalt des Grases nicht richtig geprüft und ausgewertet werden können.

Ein mehr mit der Wirklichkeit übereinstimmendes Modell dieser Einflüsse wird in Abb. 3 gegeben. In diesem Modell sind die heutigen Einsichten und Kenntnisse besser verarbeitet worden. Das Modell ist so aufgebaut worden, daß die Variablen Prozentsatz an Kräutern und Roheiweißgehalt sowohl Ursache als auch Effekt sind. Es ist klar, daß das Modell durch die Aufnahme von sogenannten Kettenprozessen mehr in Übereinstimmung mit der Wirklichkeit ist. Die Methode des einfachen Regressionsmodells kann hierauf nicht mehr angewendet werden. Die Methode der Pfadkoeffizienten bietet die Möglichkeit, solche Modelle mit kausalen Kettenprozessen als Ganzes zu prüfen und quantitativ zu schätzen.

Das Prinzip der Methode der Pfadkoeffizienten (engl. path coefficient) ist von WRIGHT (1921) eingeführt worden. Er hat in späteren Veröffentlichungen seine Ideen erläutert und mathematisch weiter ausgearbeitet. Anwendung fand sie hauptsächlich nur in der Vererbungslehre, obwohl WRIGHT auch mögliche Anwendungen in anderen Disziplinen wie der Physiologie besprochen hat. Weitere Anwendungen in der Biologie außerhalb des Gebietes der Vererbungslehre sind wenig bekannt (TURNER und STEVENS, 1959; FERRARI, 1963). Die Methode ist übrigens lange Zeit bei den Mathematikern und Genetikern nicht anerkannt worden (KEMPTHORNE, 1957). In den letzten Jahren wird die Methode in der Vererbungslehre immer mehr angewendet (LE ROY, 1959). Moderne Einsichten sind weiter zu finden in den Arbeiten von LI (1956), WRIGHT (1934, 1960) und von WRIGHT und TUKEY im Buch „Statistics and Mathematics in Biology" (KEMPTHORNE, 1954).

Auch in der Wirtschaftslehre hat man oft mit Prozessen zu tun, worin Variable sowohl Ursache als Effekt sind, und wobei man sehr schwierig künstliche Änderungen durchführen kann, um diese Kettenprozesse zu untersuchen. Die Wirtschaftslehre hat die Methode der simultanen Gleichungen entwickelt, um diese Prozesse zu analysieren. Es zeigt sich, daß die Methoden der Pfadkoeffizienten und der simultanen Gleichungen mathematisch identisch sind, so daß eine Kenntnisnahme der ökonometrischen Literatur nicht nur empfehlenswert, sondern auch notwendig ist, um die Methode der Pfadkoeffizienten zu verstehen. Gute Übersichten geben die Bücher von TINTNER (1952), KLEIN (1956), VALAVANIS (1959), und besonders THEIL (1961), und die Veröffentlichungen von THEIL und

KLOEK (1959), BASMANN (1960), SIMON (1953), KOOPMANS (1953) und MEULENBERG (1962). Das Lesen dieser Literatur ist ziemlich schwer, wei Mathematik und Ökonomie stark verknüpft miteinander behandelt werden

Das alles bedeutet, daß das Prinzip der Methode der simultanen Gleichungen zunächst in der Biologie entstanden, aber nicht so weit zur Entwicklung gekommen ist. Die Wirtschaftslehre hat am Ende der dreißiger Jahre aufs neue die Methode entdeckt und ist einen eigenen Weg gegangen. Die Biologie kann jetzt mit großen Vorteilen die mathematischen Fortschritte in der Ökonometrie benützen.

Prinzip und Methode der Pfadkoeffizienten

Ausgangspunkt ist ein Modell, das aus einem geschlossenen, kausalen, linearen System mit L primären Ursachen x_j (exogene Variablen der Wirtschaftslehre) und M Effekten y_i (endogene Variablen) besteht. Weiter wird angenommen, daß diese L und M Variablen durch ein Netzwerk von kausalen Pfaden miteinander verbunden sind. Unter einem geschlossenen kausalen linearen System wird ein Netzwerk verstanden, in dem jede Variable entweder eine lineare Kombination einer oder mehrerer Variablen dieses Systems ist oder eine der Variablen, welche durch keine der Variablen des Systems bestimmt ist. Diese letzten Variablen sind dann die primären Ursachen x_j oder die exogenen Faktoren der Ökonometrie. Die Tatsache, daß eine Variable als eine lineare Kombination einer oder mehrerer Variablen betrachtet wird, bedeutet, daß dieser Effekt y_a als eine lineare Funktion einer oder mehrerer Variablen x_j und y_i auszudrücken ist. Für die primären Ursachen x_j ist dies, wenigstens im gewählten Modell, nicht der Fall. Der Parameter, welcher in diesen Funktionen die Größe des Einflusses angibt, wird Pfadkoeffizient genannt und gibt ebenso wie der Regressionskoeffizient die Größe der Änderung des Effektes an, wenn eine Ursache um eins wächst. Die Ökonometrie spricht nur von Parametern. Die Linearität ist übrigens eine nicht-notwendige Annahme, Erfahrungen mit nicht-linearen Funktionen sind bis jetzt gering (TURNER, MONROE und LUCAS, 1961).

Ein Kettenprozeß wird gekennzeichnet durch die Wirkung einer Ursache auf einen mehr oder weniger entfernten Effekt über eine bestimmte Reihe von Variablen. Ausgedrückt in Pfadkoeffizienten, ist der Gesamteffekt über diese Reihe dem Produkt der betreffenden Pfadkoeffizienten (zusammengesetzter Pfadregression) in dieser Reihe gleich. Theoretisch erzielt man den Gesamteffekt eines Kettenprozesses dadurch, daß dieser Effekt y_a, welcher erklärt werden muß, ausgedrückt wird als eine Funktion der Variablen, deren direkter Erfolg y_a ist. Diese Ausdrücke werden Struktur-

gleichungen genannt, weil sie zusammen die Struktur des Modells angeben. Wenn diese letzten Variablen keine primären Ursachen sind, werden sie auch als Funktion von Variablen ausgedrückt. Dieser Eliminierungs-Prozeß geht weiter, bis der erstgenannte Effekt y_a als eine Funktion von nur primären Ursachen x_j ausgedrückt worden ist (reduzierte Strukturgleichungen). Man kann diese Strukturgleichungen auch als Teilregressionsgleichungen betrachten, worin der Regressionskoeffizient als eine Summe der schon genannten zusammengesetzten Pfadregressionen ausgedrückt ist. Diese zusammengesetzten Pfadregressionen müssen für alle Effekte aufgestellt oder ausgerechnet werden. Man erhält hiermit ein System von Gleichungen, worin die Regressionskoeffizienten bekannt und woraus die Pfadkoeffizienten bisweilen zu lösen sind.

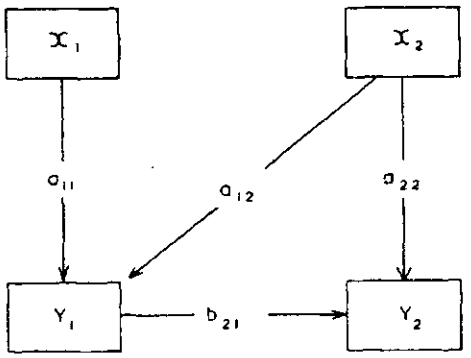

Abb. 2. Geschlossenes kausales Modell mit x_1 und x_2 als primäre kausale Faktoren und mit einem indirekten Einfluß von x_2 auf y_4 über y_3

Ein einfaches Beispiel, entnommen aus Turner und Stevens (1959), kann dies und jenes verdeutlichen. Das Modell in Abb. 2 zeigt, daß y_1 durch die primären Ursachen x_1 und x_2 bestimmt wird. Der Effekt y_2 steht direkt nur unter Einfluß von x_2 und y_1. Die primäre Ursache x_1 übt ihren Einfluß auf y_2 nur indirekt über y_1 aus. Die Strukturgleichungen, die aus dem Modell aufgestellt werden können, sind die folgenden:

$$y_1 = a_1 + a_{11}\, x_1 + a_{12}\, x_2 \tag{1}$$

$$y_2 = a_2 + b_{21}\, y_1 + a_{22}\, x_2 . \tag{2}$$

Hiervon ist Gleichung (1) schon eine reduzierte Strukturgleichung. Die zweite erzielt man, wenn (1) in (2) substituiert wird, also:

$$y_2 = a_2 + a_1\, b_{21} + a_{11}\, b_{21}\, x_1 + (a_{12}\, b_{21} + a_{22})\, x_2 . \tag{3}$$

Die Ausdrücke (1) und (3) für y_1 und y_2 sind auch als Regressionsgleichungen für y_1 bzw. y_2 auf x_1 und x_2 zu betrachten. Wenn die ausgerechneten Regressionskoeffizienten den entsprechenden Gliedern der Gleichungen (1) und (3) gleichgesetzt werden, erzielt man ein System von Gleichungen, aus dem die vier unbekannten Pfadkoeffizienten in diesem Falle zu bestimmen sind. Solche Systeme von Gleichungen sind nicht immer lös-

bar. Im besprochenen Beispiel ist es der Fall. Hierauf wird im nächster Abschnitt noch zurückgekommen.

Die allgemeine Form der Strukturgleichungen, der reduzierten Strukturgleichungen und der Auswertung nach den Pfadkoeffizienten kann jetzt abgeleitet werden. Hierbei wird angenommen, daß die Stichprobe aus N Beobachtungen besteht und (n) eine beliebige Beobachtung angibt. Weiter wird angenommen, daß die x-Variablen fehlerlos sind. Die Fehler der y-Variablen haben eine Normalverteilung und sind nicht miteinander und mit den x-Variablen korreliert. Unter diesen Annahmen darf die statistische Auswertung mit Hilfe der Methode der kleinsten Quadrate durchgeführt werden. In der Ökonometrie wird noch gefordert, daß das System komplett ist, d. h. daß die Anzahl Gleichungen der Anzahl endogener Variablen gleich ist. Mit einem Modell als Ausgangspunkt ist diese Forderung selbstverständlich immer befriedigt.

Die allgemeine, aber ausführliche Form der Strukturgleichungen mit Zufallsgliedern u_i kann für die Beobachtung (n) wie folgt ausgeschrieben werden:

$$\begin{aligned} b_{11}\, y_1^{(n)} + \cdots + b_{1M}\, y_M^{(n)} + a_{11}\, x_1^{(n)} + \cdots + a_{1L}\, x_L^{(n)} &= u_1^{(n)} \\ \vdots \qquad\qquad & \qquad \vdots \\ b_{M1}\, y_1^{(n)} + \cdots + b_{MM}\, y_M^{(n)} + a_{M1}\, x_1^{(n)} + \cdots + a_{ML}\, x_L^{(n)} &= u_M^{(n)}. \end{aligned} \tag{4}$$

Die Konstanten der Gleichungen werden in dem Parameter eines x, das konstant ist, gefunden.

Es ist klar, daß für reelle Modelle verschiedene Parameter a priori gleich Null angenommen werden. Eine bestimmte Gleichung kann also – wenn die Parameter normiert sind, so daß der Parameter eines y gleich 1 wird – wie folgt ausgeschrieben werden:

$$y_1^{(n)} = b_{12}\, y_2^{(n)} + \cdots + b_{1,m+1}\, y_{m+1}^{(n)} + a_{11}\, x_1^{(n)} + \cdots + a_{1l}\, x_l^{(n)} + u_1^{(n)}. \tag{5}$$

Diese Gleichung für y_1 hat also $(m+1) \leqq M$ Effekte y_i und $l \leqq L$ primäre Ursachen x_j.

Das System (4) für N Beobachtungen wird in der Matrizenform:

$$YB + XA = U, \tag{6}$$

geschrieben, worin Y die $N \times M$-Matrix der Variablen y_i ist, X die $N \times L$-Matrix der Variablen x_j, B und A die $M \times M$- bzw. $L \times M$-Matrizen der Pfadkoeffizienten und U die Matrix der Zufallselemente. Es wird angenommen, daß die Matrix B nichtsingulär ist.

Da also die Kehrmatrix von B besteht, ist es möglich, diese Strukturgleichungen mit Hilfe einer Matrizenmultiplikation auf die reduzierten Strukturgleichungen zu reduzieren gemäß:

$$Y = -XAB^{-1} + UB^{-1} = XP + V. \tag{7}$$

Im Gegensatz zu den Strukturgleichungen (6) darf auf das Gleichungssystem (7) die Methode der kleinsten Quadrate angewendet werden, die x-Variablen auf der rechten Seite sind nicht-stochastisch. Die Schätzung von $P: (X'X)^{-1}X'Y$ ist erwartungstreu und asymptotisch treffend (konsistent).

Zur Erlangung der Pfadkoeffizienten müssen darauf aus diesen geschätzten reduzierten Gleichungen die Strukturgleichungen zurückgerechnet werden. Zur Auswertung z. B. der Strukturgleichung von y_1 in (5) werden die $(m+1)$ reduzierten Strukturgleichungen der Effekte $y_1 \ldots y_{m+1}$ gemäß (7) in der ausführlichen Form, die MEULENBERG (1962) gibt, mit Ersatz der rechten Seiten der Gleichungen durch h_i niedergeschrieben:

$$\begin{aligned}
&1y_1 - p_{1,l+1}x_{l+1} - \cdots - p_{1,L}x_L &&= p_{1,1}x_1 + \cdots + p_{1,l}x_l &&= h_1\\
&0y_1 - p_{2,l+1}x_{l+1} - \cdots - p_{2,L}x_L &&= p_{2,1}x_1 + \cdots + p_{2,l}x_l - y_2 &&= h_2\\
&\vdots &&&&\vdots\\
&0y_1 - p_{m+1,l+1}x_{l+1} - \cdots - p_{m+1,L}x_L &&= p_{m+1,1}x_1 + \cdots + p_{m+1,l}x_l - y_{m+1} &&= h_{m+1}.
\end{aligned} \tag{8}$$

Aus diesem System kann mit Anwendung der Cramerschen Regel y_1 in $y_2 \ldots y_{m+1}, x_1 \ldots x_l$ ausgedrückt werden gemäß:

$$y_1 = |C_1| / |C|, \tag{9}$$

worin C die Matrix der Koeffizienten der Variablen auf der linken Seite ist, also:

$$|C| = \begin{vmatrix} 1 & -p_{1,l+1} & \cdots & -p_{1,L} \\ \vdots & & & \vdots \\ 0 & -p_{m+1,l+1} & \cdots & -p_{m+1,L} \end{vmatrix}$$

und

$$|C_1| = \begin{vmatrix} h_1 & -p_{1,l+1} & \cdots & -p_{1,L} \\ \vdots & & & \vdots \\ h_{m+1} & -p_{m+1,l+1} & \cdots & -p_{m+1,L} \end{vmatrix}.$$

Eine eindeutige Lösung für y_1 ist nur möglich, wenn die C-Matrizen quadratisch und die Determinanten dieser Matrizen nicht gleich Null sind (genau identifiziert). Die erste Forderung bedingt, daß $L - l + 1 = m + 1$

oder $L - l = m$, und kann im Modell immer leicht kontrolliert werden. Sie ist eine notwendige, doch nicht hinreichende Bedingung. Die zweite Forderung ist dagegen eine hinreichende Bedingung und betrifft den Rang der Matrizen; der Rang soll $m + 1$ sein (THEIL und KLOEK, 1959). Wir unterscheiden weiter zwei Fälle.

Ist $L - l < m$, dann ist die Gleichung (5) unteridentifiziert (nichtidentifizierbar); y_1 kann nicht in den Variablen $y_2 \ldots y_{m+1}$, $x_1 \ldots x_l$ ausgedrückt werden. Ein eindeutiger Ausdruck würde erzielt werden, wenn $m - (L - l)$ Reihen, also $m - (L - l)$ Variablen y_i aus der Matrix C_1 entfernt werden. Abgesehen von der Frage, welche Reihen in diesem Falle entfernt werden müssen, ist das große Bedenken dieser Handlungsweise, daß das Modell wesentlich geändert wird. Ein extremes Beispiel dieser Unteridentifizierung ist das Modell der Faktorenanalyse (FERRARI, PIJL und VENEKAMP, 1957).

Eine weniger beschwerliche Situation ist die sog. Überidentifizierung, wobei $L - l > m$ ist. Die Matrizen sind hierbei auch singulär und es ist nicht möglich, y_1 in den Variablen $y_2 \ldots y_{m+1}$, $x_1 \ldots x_l$ eindeutig auszudrücken. Man kann dies ermöglichen durch die Entfernung von $L - l - m$ nicht-stochastischen Variablen $x_{l+1} \ldots x_L$ aus dem Modell. Der Forscher wird sich nicht leicht dazu entschließen, weil auch hierdurch das Modell stark geändert wird und Information verloren geht. Die Ökonometrie hat jedoch einige Methoden entwickelt, wobei die Entfernung nicht nötig ist und doch eine Lösung erzielt wird. Es gibt hierbei Methoden, welche für die Schätzung der Pfadkoeffizienten die Information verwenden, die in nur einer, einiger oder allen Gleichungen vorhanden ist (Methoden mit und ohne Beschränkung der Informationsverwendung). Für eine ausführliche Besprechung siehe z. B. THEIL (1961).

Eine einfache und viel gebrauchte Methode ist die der kleinsten Quadrate in zwei Runden von THEIL, worin jedesmal die a priori gegebenen Beschränkungen einer Gleichung verwendet werden[1]. Die Methode verläuft wie folgt. Unter der Annahme, daß die Parameter der Gleichung (5) geschätzt werden müssen, wird die Methode der kleinsten Quadrate auf die Gleichungen für $y_2 \ldots y_{m+1}$ in reduzierter Form angewendet. Die auf diese Weise erzielten Schätzungen von $y_2 \ldots y_{m+1}$ sind erwartungstreu und asymptotisch treffend. Die Variablen $y_2 \ldots y_{m+1}$ in Gleichung (5) werden dann durch diese Schätzungen der ersten Runde ersetzt. Danach wird die Methode der kleinsten Quadrate auf diese Gleichung mit neuen Werten für $y_2 \ldots y_{m+1}$ angewendet[2]. Es ist möglich zu beweisen, daß die auf diese Weise erzielte

[1] NAGAR hat hieraus eine iterative Methode entwickelt, die die Beschränkungen aller Gleichungen berücksichtigt (THEIL und KLOEK, 1959).

[2] ZELLNER und THEIL (1962) geben ein einfacheres Schema für die Auswertung.

Schätzung von y_1 asymptotisch erwartungstreu ist (MEULENBERG, 1962). Die Schwierigkeit eines Systems mit Überidentifizierung ist dann auch, daß der Stichprobenumfang theoretisch unendlich groß sein soll.

Pfadkoeffizienten in einem Bodenfruchtbarkeitsproblem

Die Möglichkeiten der Methode der Pfadkoeffizienten in der biologischen Forschung sollen erläutert werden an Hand einer Anwendung auf ein Problem der Bodenfruchtbarkeit, das schon besprochen worden ist. Wir bringen

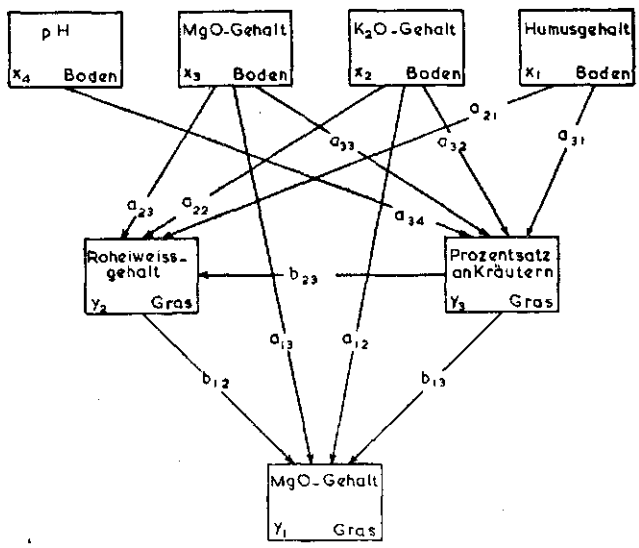

Abb. 3. Pfadkoeffizientenmodell mit direkten und indirekten Einflüssen der vier primären kausalen Faktoren auf den MgO-Gehalt des Grases

noch einmal in Erinnerung, daß wir die Methode der Pfadkoeffizienten vorgeschlagen haben, um die Schwierigkeit zu beseitigen, daß in dem Regressionsmodell der Prozentsatz an Kräutern und der Roheiweißgehalt nur als unabhängige erklärende Faktoren aufgenommen werden konnten. Das Modell aus Abb. 3, worauf die Methode angewendet worden ist, hat diese Schwierigkeit nicht. Es ist so aufgebaut worden, daß beide Variablen sowohl Ursache als auch Effekt sind. Als primäre Ursache sind in das Modell die K_2O- und MgO-Gehalte des Bodens, der Humusgehalt und das pH aufgenommen worden. Zwei dieser Faktoren üben direkt ihren Einfluß auf den MgO-Gehalt des Grases aus, nämlich die MgO- und K_2O-Gehalte des Bodens über die Pfadkoeffizienten a_{13} und a_{12}. Der Humusgehalt und das pH be-

einflussen den MgO-Gehalt des Grases nur indirekt, nämlich über den Prozentsatz an Unkräutern und den Roheiweißgehalt. Das Fehlen eines direkten Einflusses von pH und Humus ist eine Folge der Erwägung, daß wir uns nicht vorstellen konnten, auf welche Weise dies stattfinden würde. Dieselben Erwägungen gelten auch für das Fehlen einer kausalen Verbindung zwischen dem pH und dem Roheiweißgehalt. Verbindungen zwischen Humusgehalt und dem Prozentsatz an Kräutern und dem Roheiweißgehalt sind auf einen möglichen Einfluß mittels Wasserversorgung bzw. Stickstofflieferung gegründet. Der Roheiweißgehalt ist auch abhängig von der botanischen Zusammensetzung der Grasnarbe. Es ist weiter klar, daß der Einhalt solcher kausalen Modelle eng mit Kenntnis, Intuition und Einsicht des Forschers zusammenhängt.

Der Einfluß des MgO-Gehaltes des Bodens auf den MgO-Gehalt des Grases geht über verschiedene Pfade. Erstens haben wir die schon genannte direkte Wirkung, ausgedrückt durch den Pfadkoeffizienten a_{13}. Der indirekte Einfluß entsteht, weil eine Änderung des MgO-Gehaltes des Bodens direkt eine Änderung in dem Prozentsatz an Kräutern und in dem Roheiweißgehalt über a_{33} bzw. a_{23} hervorruft, die selbst auch wieder den MgO-Gehalt des Grases beeinflussen. Außerdem beeinflußt der Prozentsatz an Kräutern auch den Roheiweißgehalt. Der MgO-Gehalt des Grases hängt also auf eine direkte und auf drei indirekte Weisen von dem MgO-Gehalt des Bodens ab. Mittels Pfadkoeffizienten kann man diesen totalen Einfluß ausdrücken als: $a_{13} + a_{33}\,b_{13} + a_{23}\,b_{12} + a_{33}\,b_{23}\,b_{12}$. Entsprechende Ausdrücke sind auch für die anderen Faktoren zu geben. Diese werden erzielt entweder aus den abzuleitenden reduzierten Strukturgleichungen oder direkt aus dem Modell durch Inspektion. Es zeigte sich, daß alle Strukturgleichungen dieses Systems genau identifiziert sind.

Zum Schluß möchten wir noch einige Ergebnisse der Auswertung anführen. In der Tabelle 1 werden die Werte der berechneten Pfadkoeffizienten dieses Modells gegeben.

Eine ausführlichere Besprechung der Ergebnisse hat in einer anderen Arbeit (FERRARI, 1963) stattgefunden; an dieser Stelle wird hierauf nicht eingegangen. Nur zum Vergleich werden in Tabelle 2 noch die Regressionskoeffizienten des ersten Modells gegeben, obwohl ein idealer Vergleich wegen des Fehlens des pH in diesem Modell nicht möglich ist.

Es zeigt sich, daß die Regressionskoeffizienten und die Pfadkoeffizienten a_{12} und a_{13} als Maß für den direkten Einfluß des K_2O-Gehaltes und des MgO-Gehaltes des Bodens auf den MgO-Gehalt des Grases von gleicher Größe sind. Die direkten Einflüsse des Prozentsatzes an Kräutern und des Roheiweißgehaltes unterscheiden sich dagegen stark von den entsprechenden Regressionskoeffizienten; das Verhältnis ist ungefähr 3 : 2.

Tabelle 1

Berechnete Werte der 12 Pfadkoeffizienten des Modells aus Fig. 3

Effekt \ Ursache	Humusgehalt (x_1)	K_2O-Gehalt des Bodens (x_2)	MgO-Gehalt des Bodens (x_3)	pH (x_4)	Prozentsatz an Kräutern (y_3)	Roheiweißgehalt (y_2)
Prozentsatz an Kräutern (y_3)	1,67	−0,23	− 0,031	5,26		
Roheiweißgehalt (y_2)	−0,74	0,11	0,011		0,20	
MgO-Gehalt des Grases (y_1)		−0,0038	0,0004		0,0041	0,0083

Tabelle 2

Regressionskoeffizienten, berechnet aus dem Modell von Fig. 1

Effekt \ Ursache	Humusgehalt	K_2O-Gehalt des Bodens	MgO-Gehalt des Bodens	Prozentsatz an Kräutern	Roheiweißgehalt
MgO-Gehalt des Grases	0,0001	−0,0038	0,0004	0,0029	0,0059

Schlußwort

Die Methode der Pfadkoeffizienten ist an den Ergebnissen einer Auswertung erläutert worden, deren Data mit einem Experiment ohne Eingriff gewonnen waren. Die Methode ist unzweifelhaft auch sehr gut anwendbar zur Auswertung von Experimenten mit Eingriff. Es kann dann von Vorteil sein, diese Methode, wobei viele Variablen im Modell aufgenommen sind, anzuwenden. In einem Experiment mit Eingriff wird ja die Annahme gemacht, daß die nicht in Betracht gezogenen Faktoren sich nicht ändern. Die Möglichkeit und darum der Zwang, ein umfassenderes kausales Modell aufzustellen und lösen zu können, verschaffen dem Forscher die Gelegenheit, diese Annahme zu prüfen. Die Abwägung der Notwendigkeit, bestimmte Variablen wohl oder nicht aufzunehmen, macht den Forscher aufmerksam auf die Frage, welche Messungen er vornehmen muß, um ein möglichst wahres Bild der Wirklichkeit zu erzielen. Das Modell macht dem Forscher deutlich, daß die Wirkung eines Eingriffes auf verschiedene Weisen und über verschiedene Pfade stattfinden kann. Die Auswirkung eines Eingriffes kann dann erzielt werden, ohne daß eine direkte kausale Verbindung zwischen den beiden Variablen besteht, und umgekehrt. Hieraus gehen auch

die Beschränkung eines Experimentes mit Eingriff und die große Bedeutung der Eigenschaften der Stichprobe für das Versuchsergebnis hervor (FERRARI, 1960).

Die Methode soll ohne Zweifel von großer Bedeutung für die mehr synthetisch eingestellte biologische Forschung sein. In einer anderen Veröffentlichung haben wir hierfür schon einige Beispiele aus dem Gebiete der Bodenfruchtbarkeitsforschung gegeben (FERRARI, 1963). Dasselbe gilt auch für die Probleme aus dem Gebiete der Physiologie, Ökologie, Biochemie usw. Der große Vorteil hierbei ist auch, daß Prozesse mit Rückkoppelungssystemen (positiven und negativen) aufgenommen werden können. TURNER und STEVENS (1959) und FERRARI (1963) geben hierfür Beispiele. Es ist aus guten Gründen z. B. wünschenswert, daß das Modell in Abb. 3 mit einer Variablen K_2O-Gehalt des Grases (y_4) erweitert wird. Die Wirkung des K_2O-Gehaltes des Bodens auf den MgO-Gehalt des Grases geht dann über die Pfade a_{42} und b_{14}. Aus denselben guten Gründen darf man annehmen, daß der MgO-Gehalt des Grases den K_2O-Gehalt des Grases beeinflußt. Das heißt, man soll auch einen Pfadkoeffizienten b_{41} prüfen und schätzen. Es zeigte sich, daß dieses Rückkoppelungssystem negativ war (b_{41} negativ). Man darf erwarten, daß die Methode der Pfadkoeffizienten von großer Bedeutung für die Lösung dieser Rückkoppelungssysteme sein wird.

Die Methode zwingt den Forscher, sich immer zu vergegenwärtigen, auf welche Weise bestimmte Prozesse in Wirklichkeit zustande kommen. Kenntnis, Intuition, Kombinationsfähigkeit usw. des Forschers werden hierdurch auf die Probe gestellt, was nur ein Vorteil sein kann. Bei der Aufstellung des Modells soll der Forscher sich durch seine Kenntnisse leiten lassen und sich nicht um Identifizierungsverhältnisse kümmern. Erst danach muß er nachprüfen, inwieweit die Identifizierung Probleme gibt. Wir haben schon gesehen, daß Überidentifizierung keine unüberwindlichen Schwierigkeiten bietet. Auch für die Unteridentifizierung gibt es Möglichkeiten (VALAVANIS, 1959). Die Möglichkeit besteht jedoch immer, daß die Lösung nur erzielt werden kann durch eine Änderung des Modells. Es soll dann die Regel sein, diese Änderung zu erwähnen.

Zusammenfassung

Die Methode der Pfadkoeffizienten bietet die Möglichkeit, biologische Modelle mit kausalen Kettenprozessen und mit Rückkoppelung zu prüfen und zu schätzen. Die Möglichkeiten der Methode werden an einem Beispiel aus dem Gebiete der Bodenfruchtbarkeitsforschung erläutert. Die schon 1921 eingeführte Methode erweist sich als mathematisch identisch mit der Methode der simultanen Gleichungen aus der Ökonometrie.

Die allgemeine mathematische Beschreibung solcher kausalen Modelle durch Gleichungen wird gegeben, wonach abgeleitet wird, auf welche Weise die Lösung erzielt werden kann. Es wird angegeben, wann Über- und Unteridentifizierung auftreten. Die Methode der kleinsten Quadrate in zwei Runden für die Lösung der Überidentifizierung wird besprochen.

Summary

The method of path coefficients gives the possibility to test and to estimate biological models with causal chain processes and feedback. The potentialities of the method are illustrated on a problem of soil-fertility research. The method already introduced in 1921 turns out to be identical with the simultaneous-equations technique in econometrics.

The general mathematical description of such causal models by equations is given. Thereafter it is derived in which way the solution of the system of equations can be obtained. It is shown in which cases over- and under-identification appear. The method of the two-stage least squares estimation for the solution of over-identified equations is discussed.

Literatur

BASMANN, R. L., 1960: An expository note on estimation of simultaneous structural equations. Biometrics **16**, 464–480.

FERRARI, TH. J., H. PIJL and J. T. N. VENEKAMP, 1957: Factor analysis in agricultural research. Neth. J. Agr. Sci. **5**, 211–221.

FERRARI, TH. J., 1960: Vergelijking tussen proeven met en zonder ingreep. Landbouwk. Tijdschr. **72**, 792–801.

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

KEMPTHORNE, O., 1957: An introduction to genetic statistics. New York.

KLEIN, L. R., 1956: A textbook of econometrics. Evanston.

KOOPMANS, T., C. 1953: Identification problems in economic model construction. Chapter II, Studies in Econometric Method. Cowles Commission Monograph no. 14, ed. W. C. Hood and T. C. Koopmans.

LE ROY, H. L., 1959: Grundlage und Anwendungsmöglichkeiten der Methode des Pfadkoeffizienten. Biometrische Zeitschrift 1, 30–43.

LI, C. C., 1956: The concept of path coefficients and its impact on population genetics. Biometrics **12**, 190–210.

MEULENBERG, M. T. G., 1962: Vraaganalyse voor landbouwprodukten uit tijdreeksen. Mededelingen van de Landbouwhogeschool, Wageningen **62**, (3), 1–133.

SIMON, H. A., 1953: Causal ordering and identifiability. Chapter III, Studies in Econometric Method. Cowles Commission Monograph no. 14, ed. W. C. Hood and T. C. Koopmans.

SIMON, H. A., 1954: Spurious correlation: a causal interpretation. Am. Stat. Ass. **49**, 467–479.

SLUIJSMANS, C. M. J., 1962: Magnesium- en kaliumgehalten van gras in afhankelijkheid van bodem- en andere factoren. Tijdschr. Diergeneesk. **87**, 547–556.

THEIL, H., und T. KLOEK, 1959: The statistics of systems of simultaneous economic relationships. Statistica Neerlandica **13**, 56–89.

THEIL, H., 1961: Economic Forecasts and Policy. Amsterdam.

TINTNER, G., 1952: Econometrics. New York.

TUKEY, J. W., 1954: Causation, regression and path analysis. Chapter 3, Statistics and Mathematics in Biology, ed. Kempthorne.

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

TURNER, M. E., R. J. MONROE und H. J. LUCAS, 1961: Generalized asymptotic regression and non-linear path analysis. Biometrics **17**, 120–143.

VALAVANIS, S., 1959: Econometrics. An introduction to maximum likelihood methods. New York.

WRIGHT, S., 1921: Correlation and causation. J. Agr. Res. **20**, 557–585.

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

WRIGHT, S., 1954: The interpretation of multivariate systems. Chapter 2, Statistics and Mathematics in Biology, ed. Kempthorne.

WRIGHT, S., 1960: Path coefficients and path regressions: alternative or complementary concepts? Biometrics **16**, 189–202.

WRIGHT, S., 1960: The treatment of reciprocal interaction, with or without lag, in path analysis. Biometrics **16**, 423–445.

ZELLNER, A., und H. THEIL, 1962: Three-stage least squares: simultaneous estimation of simultaneous equations. Econometrica **30**, 54–78.

Manuskript-Eingang: 26. 8. 63

Anschrift des Verfassers: Dr. Th. J. Ferrari
Instituut voor Bodemvruchtbaarheid
van Hallstraat 3, Groningen, Holland