

Sonderdruck
aus

**AKTUELLE PROBLEME
DES LANDWIRTSCHAFTLICHEN VERSUCHSWESENS**

Im Eigenverlag der
Landwirtschaftlich-chemischen Bundesversuchsanstalt Linz/D.
1965

Die Methode der Pfadkoeffizienten

Ferrari Th. J.

Überreicht vom Verfasser

Veröffentl. landwirtsch.-chem. Bundesversuchsanst. Linz/D.
6 (1965) 59-66.

Inhalt

In der Broschüre sind folgende Beiträge veröffentlicht:

	Seite
Ehrendorfer, K.: Die Entwicklung des landw. Versuchswesens.....	1
Ferrari, Th. J.: Prüfen mit und ohne Eingriff	25
Ferrari, Th. J.: Die Methode der Pfadkoeffizienten	59
Lengauer, E.: Das Eliminieren der unabhängigen Variablen aus Regressionen	33
Roppert, J.: Beispiele moderner Verfahren in der Biostatistik ..	21
Schiller, H.: Aktuelle Probleme des landw. Versuchswesens	9
Schiller, H.: Koreferat zu „Probleme bei Anwendung der Faktorenanalyse“	53
Schmetterer, L.: Möglichkeiten der Bearbeitung biolog. Probleme ..	16
Voak, H.: Probleme bei der Anwendung der Faktorenanalyse	46
Allgemeine Diskussion	67

Die Methode der Pfadkoeffizienten

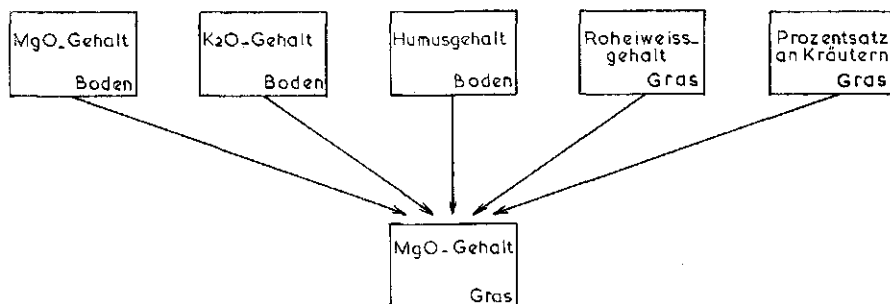
Von Th. J. Ferrari

Einleitung

Eines der meist angewendeten Modelle, um landwirtschaftliche Erscheinungen zu beschreiben, ist das Modell der Regressionsanalyse mit einer oder mit mehreren unabhängigen Variablen. Gegen die Anwendung dieses Regressionsmodells für die Analyse der Zusammenhänge zwischen den Variablen bestehen jedoch große Bedenken. Eines der wichtigsten 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 kann erläutert werden an Hand eines Beispiels aus dem Gebiet der Bodenfruchtbarkeitsforschung, wobei die Zusammenhänge zwischen dem MgO- und dem K₂O-Gehalt des Grases einerseits und den kausalen bodenkundlichen und anderen Faktoren andererseits untersucht wurden (5).

Die Daten, mit einem Experiment ohne Eingriff erhalten, wurden in einem Regressionsmodell analysiert, wobei die MgO- und die K₂O-Gehalte des Grases als die abhängigen Variablen, die MgO- und die K₂O-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 wurden. Schematisch wird dieses Modell in Abb. 1 gezeigt. In diesem Diagramm wird der angenommene kausale Zusammenhang zwischen zwei Faktoren mit einem Pfeile angegeben, wobei der Effekt an der Spitze des Pfeiles 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 K₂O-Gehaltes des Bodens nur eine 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 K₂O-Gehaltes des Bodens, des Roheiweißgehaltes und des Prozentsatzes an Kräutern auf den MgO-Gehalt des Grases nicht richtig geprüft und ausgewertet werden können.

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



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, daß die Variablen Prozentsatz an Kräutern und Roheweißgehalt 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.

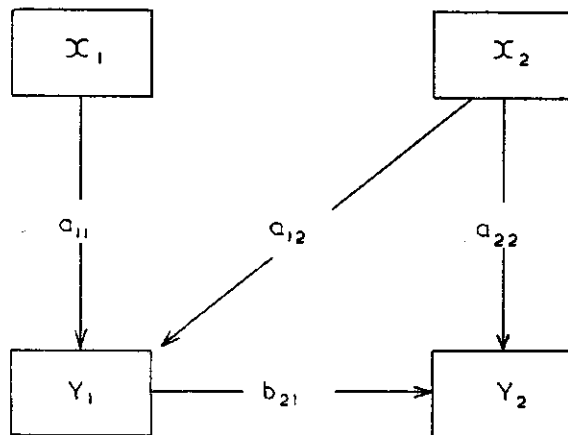
Prinzip der Pfadkoeffizienten

Ausgangspunkt ist ein Modell, das aus einem geschlossenen kausalen linearen System mit primären Ursachen x_j und Effekten y_i besteht. Weiters wird angenommen, daß diese 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 Variable, welche durch keine der übrigen Variablen des Systems bestimmt ist. Diese letzteren Variablen sind dann die primären Ursachen x_j . 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 Linearität ist übrigens eine nicht notwendige Annahme (6).

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 (zusammengesetzte 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 Strukturgleichungen genannt, weil sie zusammen die Struktur des Modells angeben. Wenn diese letzteren Variablen keine primären Ursachen sind, werden sie auch als Funktion von Variablen ausgedrückt. Dieser Eliminierungsprozeß 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, in dem die Regressionskoeffizienten bekannt und woraus die Pfadkoeffizienten 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_2 über y_1 .



Ein einfaches Beispiel kann dies 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:

$$y_1 = a_1 + a_{11}x_1 + a_{12}x_2 \quad (1)$$

$$y_2 = a_2 + b_{21}y_1 + a_{22}x_2 \quad (2)$$

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

$$y_2 = a_2 + a_1b_{21} + a_{11}b_{21}x_1 + (a_{12}b_{21} + a_{22})x_2. \quad (3)$$

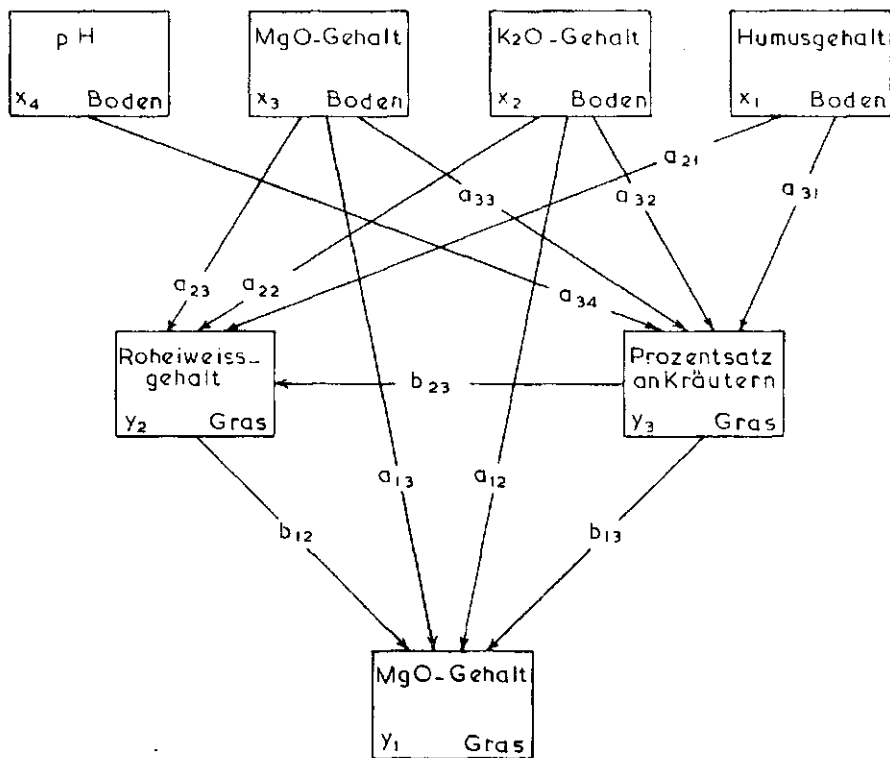
Die Ausdrücke (1) und (3) für y_1 und y_2 sind auch als Regressionsgleichungen für y_1 bzw. y_2 auf x_1 und x_2 zu betrachten. Wenn die ausgerechneten Regressionskoeffizienten den entsprechenden Gliedern der Gleichungen (1) und (3) gleichgesetzt werden, erzielt man ein System von Gleichungen, aus dem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

Es wird deutlich sein, daß ein System von Gleichungen, das die Wirklichkeit beschreibt, im allgemeinen komplexer ist. Diese Systeme sind dann nicht immer lösbar. Im soeben beschriebenen Beispiel würde es möglich sein, die Pfadkoeffizienten zu berechnen, weil die Gleichungen identifiziert sind. Man hat meistens mit Systemen zu tun, die über- oder unteridentifiziert sind. Dies hängt zusammen mit dem Verhältnis zwischen den in jeder Gleichung aufgenommenen primären Ursachen- und Effektanzahlen. Methodisch ist die Überidentifizierung kein Problem, sie gibt nur mathematische Schwierigkeiten. Unteridentifizierung ist nur zu beseitigen durch eine Änderung des Modells; hiergegen bestehen große methodische Bedenken. Ein extremes Beispiel der Unteridentifizierung ist das Modell der Aspektenanalyse.

Beispiel mit Pfadkoeffizienten

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 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, auf das die Methode angewendet worden ist, hat diese Schwierigkeiten nicht. Es ist so aufgebaut worden, daß beide Variablen sowohl Ursache als auch Effekt sind. Als primäre Ursachen sind in das Modell die K_2O - und die 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 beeinflussen den MgO -Gehalt des Grases nur indirekt, nämlich über den Prozentsatz an Unkräutern und den Roheiweißgehalt. Ein direkter Einfluß von pH und Humus fehlt, da wir uns einen solchen nicht vorstellen konnten. Dasselbe gilt auch für das Fehlen einer kausalen Verbindung zwischen dem pH und dem Roheiweißgehalt. Die Verbindungen zwischen Humusgehalt und dem Prozentsatz an Kräutern und dem Roheiweißgehalt sind auf deren möglichen Einfluß mittels Wasserversorgung bzw. Stickstofflieferung gegründet. Der Roheiweißgehalt ist auch abhängig von der botanischen Zusammensetzung der Grasnarbe. Es ist weiters klar, daß der Inhalt solcher kausaler Modelle eng mit Kenntnis, Intuition und Einsicht des Forschers zusammenhängt.

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



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_{12} . Der indirekte Einfluß entsteht, weil eine Änderung des MgO-Gehaltes des Bodens direkt eine Änderung in dem Prozentsatz an Kräutern und in dem Rohweißgehalt 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 Rohweißgehalt. Der MgO-Gehalt des Grases hängt also auf eine direkte und auf drei indirekte Weisen vom 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 der Betrachtung des Modelles. Es zeigte sich, daß alle Strukturgleichungen dieses Systems genau identifiziert sind.

Tabelle 1: Berechnete Werte der 12 Pfadkoeffizienten des Modells aus Abb. 3.

Ursache Effekt	Humus- gehalt (x_1)	K ₂ O-Gehalt des Bodens (x_2)	MgO-Gehalt des Bodens (x_3)	pH (x_4)	Prozentsatz an Kräutern (y_3)	Roheiweiß- gehalt (y_2)
Prozentsatz an Kräutern (y_3)	1.67	—0.23	—0.031	5.26		
Roheiweiß- gehalt (y_2)	—0.74	0.11	0.011		0.20	
MgO-Gehalt des Grases (y_1)		—0.0038	0.0004		0.0041	0.0083

Zum Schluß möchten wir noch einige Ergebnisse der Auswertung anführen. In der Tabelle 1 werden die Werte der berechneten Pfadkoeffizienten dieses Modells gezeigt. Zum Vergleich werden in Tabelle 2 die Regressionskoeffizienten des ersten Modells wiedergegeben, obwohl ein idealer Vergleich wegen des Fehlens des pH in diesem Modell nicht möglich ist.

Tabelle 2: Regressionskoeffizienten berechnet aus dem Modell von Abb. 1

Ursache: Effekt	Humus- gehalt	K ₂ O-Gehalt des Bodens	MgO-Gehalt des Bodens	Prozentsatz an Kräutern	Roheiweiß- gehalt
MgO-Gehalt des Grases	0.0001	—0.0038	0.0004	0.0029	0.0059

Es zeigt sich, daß die Regressionskoeffizienten und die Pfadkoeffizienten a_{12} und a_{13} als Maß für den direkten Einfluß des K₂O-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.

Die Auswertung gab noch andere interessante Ergebnisse. Wir nennen den negativen Einfluß des K₂O- und MgO-Gehaltes des Bodens auf den Prozentsatz an Kräutern, den positiven Einfluß auf den Roheiweißgehalt, den hier nicht gegebenen positiven totalen Einfluß (direkten und indirekten über die botanische Zusammensetzung) auf den Roheiweißgehalt, den unerwartet negativen Einfluß des Humusgehaltes auf den Roheiweißgehalt, den negativen Einfluß des K₂O-Gehaltes des Bodens auf den MgO-Gehalt des Grases usw. Der negative Einfluß des Humusgehaltes deutet darauf hin, daß die Wirkung der organischen Substanz sicher nicht (nur) in einer Stickstofflieferung gesucht werden muß.

Der große Vorteil der Methode ist auch, daß Prozesse mit Rückkopplungssystemen (positiven und negativen) aufgenommen werden können. Es ist aus guten Gründen z. B. wünschenswert, daß das Modell in Abb. 3 um die Variable 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 Rückkopplungssystem negativ war (b_{41} negativ). Man darf erwarten, daß die Methode der Pfadkoeffizienten von großer Bedeutung für die Lösung dieser Rückkopplungssysteme sein wird.

Schlußwort

Die Methode der Pfadkoeffizienten ist an den Ergebnissen einer Auswertung erläutert worden, deren Daten mit einem Experiment ohne Eingriff gewonnen worden waren. Die Methode ist unzweifelhaft auch sehr gut anwendbar zur Auswertung von Experimenten mit Eingriff. Es kann dabei von Vorteil sein, viele Variable in das Modell aufzunehmen. In einem Experiment mit Eingriff, wie man es jetzt üblicherweise anlegt, wird ja die Annahme gemacht, daß die außer Betracht gelassenen Faktoren sich nicht ändern. Die Möglichkeit und der Zwang, ein umfassenderes kausales Modell aufzustellen und zu lösen, verschaffen dem Forscher die Gelegenheit, diese Annahme zu prüfen. Die Erwägung, ob bestimmte Variable aufzunehmen sind oder nicht, 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. Ein Eingriff kann sich auswirken, ohne daß eine direkte kausale Verbindung zwischen den beiden Variablen besteht und umgekehrt.

Die Methode der Pfadkoeffizienten zwingt den Forscher, sich immer zu vergewissern, 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 aufwirft. Wir haben schon gesehen, daß Überidentifizierung keine unüberwindlichen Schwierigkeiten bietet. Auch für die Unteridentifizierung gibt es Möglichkeiten. Es kann jedoch immer vorkommen, daß die Lösung nur erzielt werden kann durch eine Änderung des biologischen Modells. Es soll dann die Regel sein, diese Änderung zu erwägen.

Literatur

- (1) Ferrari, Th. J.: Vergelijking tussen proeven met en zonder ingreep. Landbouwk. Tijdschrift 72, 792—801, 1960.
- (2) Ferrari, Th. J.: Causal soil-plant relationships and path coefficients. Plant and Soil 19, 81—96, 1963.
- (3) Ferrari, Th. J.: Auswertung biologischer Kettenprozesse mit Hilfe von Pfadkoeffizienten. Biometrische Zeitschrift 6, 89—102, 1964.
- (4) Ferrari, Th. J.: Prüfung mit und ohne Eingriff. In vorliegender Broschüre, S. 25, 1965.
- (5) Sluijsmans, C. M. J.: Magnesium- en kaliumgehalten van gras in afhankelijkheid van bodem- en andere factoren. Tijdschr. Diergeneeskunde 87, 547—556, 1962.
- (6) Turner, M. E., Monroe, R. J. und Lucas, H. J.: Generalized asymptotic regression and non-linear path analysis. Biometrics 17, 120—143, 1961.

Diskussion

Dr. Schiller: Die größte Schwierigkeit der Pfadanalyse besteht darin, das Schema der Zusammenhänge zu finden. Denn wenn ich weiß, wie die Zusammenhänge liegen, kann ich die Pfadanalyse ansetzen. Aber wenn ich es nicht weiß, was mache ich dann?

Dr. Ferrari: Man darf nicht meinen, daß man mit den Pfadkoeffizienten alle Schwierigkeiten lösen könnte. Man muß stufenweise vorgehen, sich in das Problem hineindenken, sich mit Spezialisten beraten und danach das Schema oder Modell aufstellen. Das ist das Schwierigste. Danach folgt die Prüfung dieser Hypothese mit Daten, die mit oder ohne Eingriff erhalten worden sind. Die landwirtschaftliche Forschung geht meist umgekehrt vor: Sie legt zuerst die Versuche an, bekommt Ergebnisse, geht zum Mathematiker und fängt jetzt erst zu überlegen an.

Wenn man natürlich gar nichts über die Zusammenhänge weiß, ist es sehr schwierig, eine Pfadanalyse zu machen. Aspektanalysen zum Beispiel könnten als Grundlage für das Pfadschema dienen.