Models and their testing: considerations on the methodology of agricultural research

TH. J. FERRARI

Institute for Soil Fertility, Groningen, Netherlands

Summary

We have experienced that causation, especially in agricultural phenomena, is complex and that the method of analysis used in natural sciences is not satisfactory in all respects. Some directives are given to disentangle this complexity based on the following ideas.

The first point is connected with the thought that also in agricultural research with its applied character the hypothesis expressed in a model and followed by testing should supply the main contribution to new knowledge. As we have experienced, this is frequently forgotten.

The second point is the idea that testing can also be carried out with observational data from experiments without artificial change (non-manipulative experiments).

The third point is the knowledge that the research worker can choose out of many models and functions. In this it is not necessary to confine the choice to functions with few factors and to models in which the *ceteris-paribus* principle must be assumed.

A definite advice which attack and which models and functions should be chosen, can not be given. Each problem requires its own method of analysis, each research worker should follow his own way and chooses his own models.

1. Introduction

It is the task of agronomic research to help the farmer in answering the question which of the different alternatives in technical possibilities he should choose in a certain situation. The agronomist is often in a difficult position because a great number of factors which are difficult to measure are to be taken into account. These two facts cause many difficulties in the research too. The difficulties in particular apply to the agriculturist who, for example, has to study the economics of milk production. One has to realize between which alternatives of possibilities a farmer could choose in his West-European mixed-farming system. Which are the consequences of an increase of grass yield by rising the nitrogen dressing via the chain: soil, crop and cattle for his financial results? To this lifestock farming in particular the words are applicable of the mathematician BROSS in his book "Design for Decision": "It is much more difficult to be a good farmer than a good mathematician because the farmer must deal with so many vague and complex problems". It is the task of research to give solutions for these problems.

It is clear that, for making a justified choice between the alternatives, it is necessary to have at one's disposal a preferably qualitative description as complete as possible of the technical possibilities and their consequences. Whether one has to advice the

Received for publication 5th March, 1965.

farmer on the base of programming models or one tries to analyze a certain farming result, this knowledge is always necessary. In the first case the technical relationships are taken for granted, in the latter case the accent lies on the explanation of certain phenomena. The connection between both procedures is that a representation is made of the relationship or phenomenon by means of models. In the former the model is assumed to be known, in the latter the model is constructed in the form of a hypothesis, the reality value of which has to be tested by an experiment. Consequently, the research as such will have to do with the model mainly as a hypothesis although there are all sorts of nuances. In this paper we only deal with problems met in the investigation into the explanation of relationships. We know, however, that the results of this research can be used for all sorts of programming purposes.

2. Use of models in agricultural research

What do we understand by models and what are their functions in research? Models are simplifying abstractions of reality in which only elements already familiar to us are absorbed. Only those elements of the reality are absorbed which are being studied in the science concerned. The abstraction is expressed in some language, in words or in diagrams, mathematically or materially. Within the given limits we try to describe the reality as completely as possible.

For the research it is of great importance that the models have the quality that conclusions drawn from them are valid for the reality. In other words, the reality value of an assumed model is closely connected to that of the conclusions. It also appears that the hypotheses, so important for progress in science, are suitable to be expressed in models. In this way we get the connection between model and research which takes place as follows. As in all empirical sciences the systematic increase of ouw knowledge in agricultural research is obtained by the formulation of hypotheses which are tested by the reality, *viz.* the observations by means of predictions. The hypothesis is rejected by the absence of agreement between observations and predictions, it is made more acceptable by the presence of agreement. In connection with the complex and practical character of an agricultural object it appears useful to build up models in the form of mathematical equations. In view of this particular character of the object, *e.g.* plant or milk production, one will meet some difficulties in testing and quantifying the parameters of the models. Such difficulties are present also in other sciences as sociology, economy and astronomy.

It is clear that the ultimate criterion in agriculture must be the production expressed in some economic terms. In the model the production, *e.g.* the milk yield in kg per ha will be brought in causal relation with a number of factors. In a simple case the function has the following form: the yield depends on the amount of roughage and concentrates. This is a very simple model, which perhaps applies to stable feeding under certain conditions. It is much more difficult, however, to relate the farm-economic results with the amounts of nitrogen dressings applied. No direct relation between these two factors exists and all kind of factors can interfere. It is clear that the hypothetical model of these relationships becomes complex and that the testing and quantifying will cause difficulties. When we examine which factors can influence the yield or the economic results of an operation, the following groups can be distinguished: –

a. variable factors such as nitrogen dressing and amount of concentrates;

b. hardly or not changeable factors which can be measured previously or predicted *e.g.* soil profile and ground-water level, size of holding, number of cows per ha etc.;
c. hardly or not changeable factors which cannot be predicted, such as weather, diseases and pests, economic state etc.

The complex character of the production, specially under farming conditions, and the peculiar attributes of the just-mentioned factor groups have certain consequences in the research for the construction and testing of models.
This applies first of all to the testing. It is a well-known fact that testing of a hypothesis in natural sciences takes place mainly by means of an artificial variation, *ceteris paribus,* according to the assumption that the change of a factor assumed to be a cause must also result in a corresponding change of its effect. In this the *ceteris-paribus* assumption is very important.
The introduction of a variation is difficult or even impossible when we are dealing with factors of the second or third group for they are not changeable. Astronomy shows that after all it is possible to obtain important results without artificial change. Apart from that it is doubtful whether the *ceteris-paribus* principle with an artificial change can be maintained in many cases. Changes in ground-water level or nitrogen dressing, for instance, cause a whole series of changes of other factors which affect the production in their turn, and the result is that conclusions with regard to such a factor causing a phenomenon cannot be drawn. It is also clear that it is difficult to investigate effects of certain changes under farming conditions. Restrictions by farming conditions, cost, *etc.* limit the introduction of experiments with artificial changes in farm-economic research. There are perhaps possibilities in certain production branches in which the feeding takes place in the stable and with purchased feeding stuffs only.
A second difficulty is connected with the great number of factors influencing the production, and with their interdependence. In view of the wanted usefulness in practice it has the consequence that the researcher always has to investigate many factors together. The normal experiment by which the influence of one or two factors is investigated, is less suitable to solve practical questions. It is a well-known fact that increasing the number of factors to be investigated soon becomes impossible; an increase of the number of factors increases the size of the experiment, by which the rest variance, certainly of field experiments, becomes the main factor. Statisticians have tried to eliminate this drawback by introducing the principle of confounding; a satisfying solution however has not yet been obtained.
The limitation remains that the results of experiments are valid only for the special case with special conditions of soil, climate, care, *etc.* It is therefore experienced that the results of the different investigations can diverge strongly. The investigator can try to solve this difficulty by carrying out a large number of experiments in order to grasp a great number of production circumstances, but after all only an average result is obtained. A subdivision according to geographical units usual in sociology does not satisfy either. Without a more thorough analysis of the causing the differences factors an extrapolation from the average result to the future individual cases remains risky.
Such an analysis is possible however, as it not necessary at all to test the hypothesis by means of empirical data obtained by an artificial change only. Under the influence of natural sciences many research-workers are of the opinion that the so-called exper-

iment with an artificial change (*manipulative experiment*) [1] is the only correct method. However, it is quite possible to test a hypothesis by means of data from an experiment without this artificial change (*non-manipulative experiment*) [1], in which the variation of nature is used. As far as the logic of experimentation is concerned, this distinction is of no account at all. The testing of the hypothesis by means of deduced predictions is deciding. The word "experiment", derived from the latin verb *experiri, i.e.* to test, expresses this already. However, by the methods and results of the physical and chemical sciences the word "experiment" has got a quite other sense, *viz.* artificial change, and the original sense is often forgotten. Of course it must be said that an non-manipulative experiment also involves certain difficulties, of which the difficulty to obtain a sufficient separation between the possible causal factors should be mentioned in particular. For the rest, the difficulties of a manipulative experiment should not be underrated either. We have mentioned already the unreal assumption of the *ceteris-paribus* principle. In a previous paper we compared the advantages and disadvantages of both methods. It is evident that an experiment without artificial change gives the possibility to test and quantify models in which factors of the second and third group (see p. 367) are absorbed, *i.e.* factors which are not or hardly changeable and by which differences between experimental results can be explained.

2.1. The use of two-variable models with one equation

Which models and which functions are generally used in agricultural research? In the following discussion we shall illustrate our statements with examples derived from soil-fertility studies. It will be clear that the statements also apply to other parts of the agricultural research. For the present we restricts ourselves to models which can be described with one equation with one or more factors.

The most simple model is the hypothesis that the yield differences can be explained by one or more factors without a further description of the functional form. This is the point of view of the analysis of variance. The drawback of course is that a possibility to interpolate and to extrapolate is difficult because of the absence of a function. Economic interpretations are difficult in that case.

More possibilities are given by the model of which the function is a linear equation: a one-unit increase of the independent variable increases the effect with a constant amount, no matter which value the first variable has. We know that this assumption is not real in many cases. The linearity can be useful in a limited region of the production, but according to experiences of agricultural research it would be more useful to utilize non-linear functions reaching a maximum. It has advantages to choose the most simple function in this case. In the litterature many equations have been proposed. The most well-known is the MITSCHERLICH equation, afterwards with a depression. Some more functions are: –

$y = A(1 - 10^{-cx})$ .... (MITSCHERLICH)

$y = ax^b$ ..... (COBB-DOUGLAS)

$y = A \cdot 10^{-z\left(\log \frac{x+i}{a+i}\right)^n}$ ... (VON BOGUSLAWSKI-SCHNEIDER)

$y = bx - cx^2$

$y = b\sqrt{x} - cx$

We renounce discussing the general and particular properties of these functions for

[1] In Dutch language: – *proef met ingreep* and *proef zonder ingreep*, in German: – *Experiment mit Eingriff* and *Experiment ohne Eingriff*.

which we refer to the book of HEADY and DILLON (1961) and to the paper of HOFFMANN and DÖRFEL (1963).

The just-mentioned equations have one thing in common: they have been developed mainly heuristic and their theoretical base is very small. By this we mean to say that there is no preference for one of them from a physiological or biochemical point of view. The only theoretical derivation we know is the one for the MITSCHERLICH equation by LINSER and KAINDL (1951) in the domain of plant nutrition. It is striking indeed that so little basic research on the production functions has been done. This does not apply to all parts of agricultural research. In soil science for instance a number of processes have been described by functions derived from basic chemical or physical knowledge. It appears that we are in urgent need for more biologically-derived equations, especially in view of the great possibilities which the computers have for the solution of these equations.

Personal preference decides at present which equation is chosen ultimately. The choice is often made by the suggestion which is made by the observational data. A study on the milk production by HEADY, SCHNITTKER, JACOBSON and BLOOM (1956) leaves the choice between three functions, *viz.* the logarithmic, the quadratic or the square-root equations. It is clear that the function ultimately chosen should be taken again as a hypothesis in the next investigation. Experience shows, however, that this has often been omitted. The uncertainty about the function to be taken and the impossibility to compute — formerly we did not yet have these calculating machines — have been the background to develop the graphic method in soil-fertility research some 30 years ago, always using the suggestion given by the observational data. The same application has also been used in the economic research in the U.S.A. As an example we show in FIG. 1 the results of an investigation into the relationship between potash status of the soil and the loss of potato yield without potash dressing, expressed in percentages of the maximal yield. Each point represents the result of one field experiment, the differences in potash status being acquired without artificial change by taking natural situations. We may expect that the differences between the graphic and numerical methods will be small.

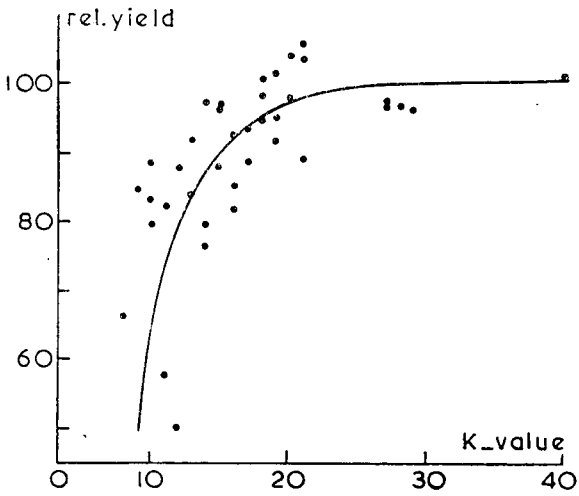


FIG. 1
Relationship between potash status of the soil and potato-yield loss in the case that no potash dressing is administered

2.2. The use of more-variable models with one equation

We know that the two-variable models mostly do not meet the needs of a complete or satisfying description of the processes. With a view to this, description functions with more factors have been developed such as: –

$y = b_1x_1 + b_2x_2 + \ldots.$

$y = A(1—10^{-c_1x_1})\ (1—10^{-c_2x_2} \ldots.$ MITSCHERLICH

$y = ax_1^{b_1}\ x_2^{b_2} \ldots\ldots\ldots\ldots\ldots$ COBB-DOUGLAS

$y = b_1x_1 + b_2x_2 + b_{12}x_1x_2$

The properties of these equations will not be discussed either, although they are important in connection with terms as isoclines, substitution rates *etc.* We only point to the possibility to absorb in these equations terms for interaction. In the last equation the product term represents the interaction. Although the interaction in our opinion is mostly nothing else than a word to mark our lack of knowledge, the researcher is often forced to absorb these terms of interactions. FIG. 2 shows a tested model with interaction in which the effect of potash dressing depends on the potash status of the soil.

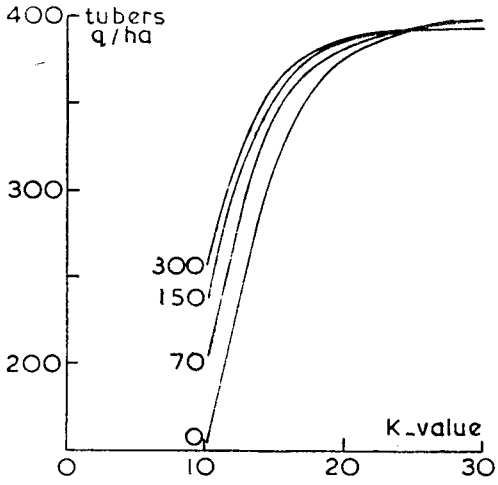


FIG. 2
Relationship between potash status of the soil and potato yield at four different potash dressings

The extreme consequences of the possibilities of an experiment without artificial change and of more-variable models are the investigations in which he research-worker tries to find in a graphical or numerical way an explanation for the differences in yield or economic farming results by means of single plots or farms respectively. A model has been drawn, which aims to give an explanation of the variance present in nature. The building-up of the model with many factors goes rather far. As distinct from the design of the analysis of variance in which the rest variance is made as small as possible, these multi-factorial investigations are interested especially in a great starting variance. FIG. 3 shows the possibilities of such an analysis by means of the correlation between actual and estimated yields. This analysis was based on a model with 13 factors, of which 9 had a statistically-significant influence. FIG. 4 shows the decrease of the yield variance by successive eliminations of the factor influences. The diagram also shows a probably more general phenomenon: many factors have a small, only few have a relatively great influence. We shall return

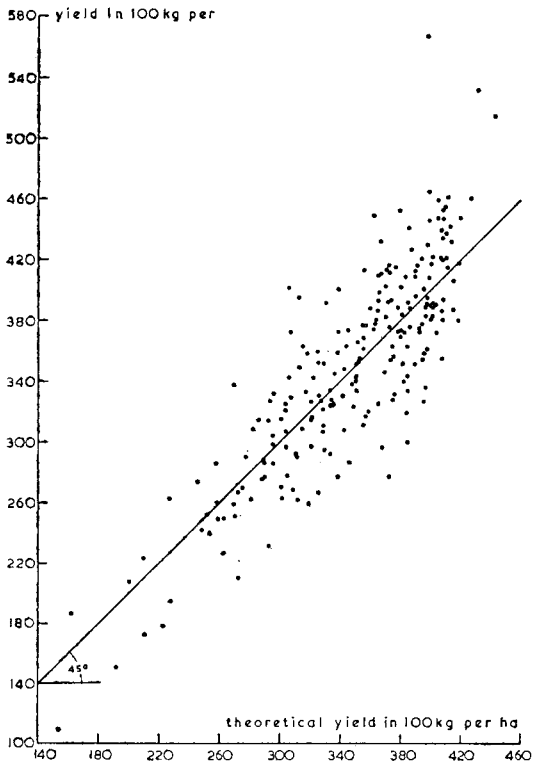


FIG. 3
Correlation between estimated and actual potato yield

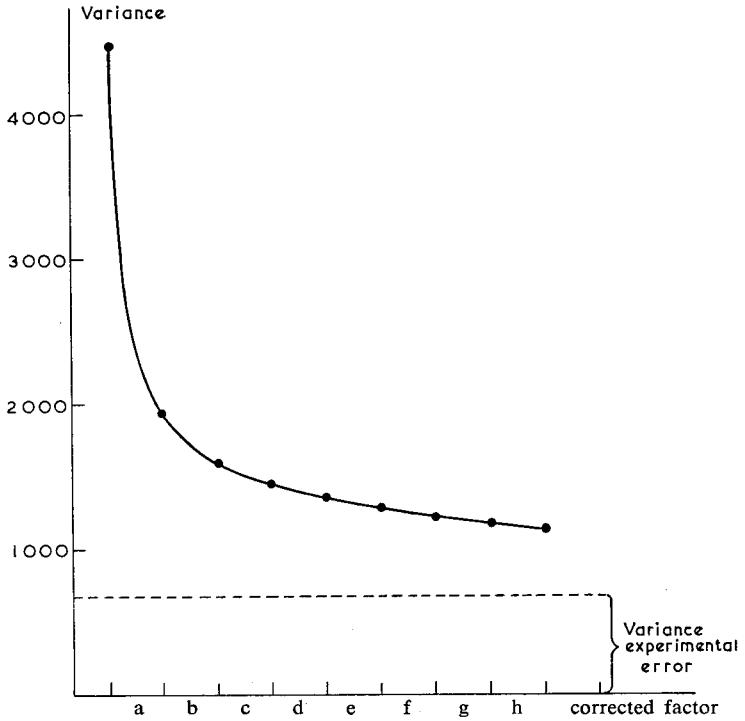
to this subject later in connection with the choice of simple-structure rotation in factor analysis.

2.3. Models with more equations, chain processes

The equations of the models discussed up till now especially are normal regression equations. The regression model is characterized by the hypothesis that a causal relationship exists between the s.c. independent or causal factors and the dependent factor or effect. It is also assumed that a change of an independent factor affects the dependent variable only and does not affect the other factors. The same assumption must also be made in the experiment with artificial change according to the *ceteris-paribus* principle. We find however that these assumptions only sometimes are in agreement with the facts both in the experiment with artificial change and in that without such a change. This means that the assumed model is incorrect and can not be applied.

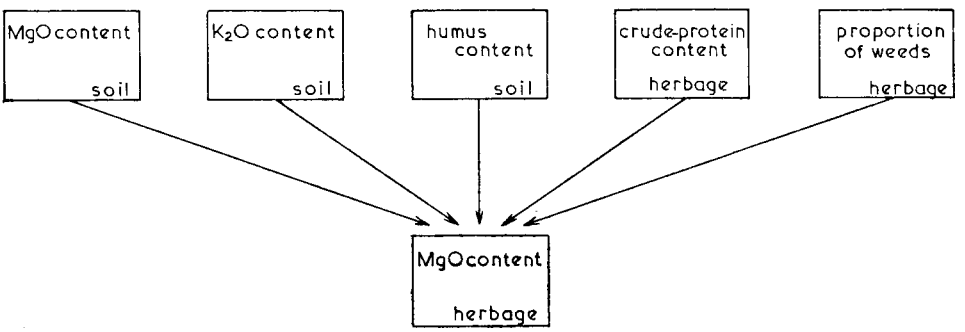
This can be illustrated by means of an example from an investigation into the factors affecting the magnesium content of herbage. At first a normal regression model was built up and tested by observations of a non-manipulative experiment. The diagram of FIG. 5 shows the hypothetic model. In this model the magnesium content of the herbage is the dependent variable or effect. Further it is assumed that the factors magnesium, potash and humus content of the soil, crude-protein content and proportion of weeds in the herbage will influence causally the magnesium content of the

FIG. 4. Decrease in yield variance after successive eliminations of the factor influences



a = potash content and lime status.
b = potash manuring.
c = organic-matter content (%).
d = distance to farmstead.
e = clay content of topsoil ($< 16\ \mu$).
f = date of planting.
g = depth of reduction.
h = ground-water level.

FIG. 5. Regression model with magnesium content of the herbage as dependent variable and other variables as dependent causal variables



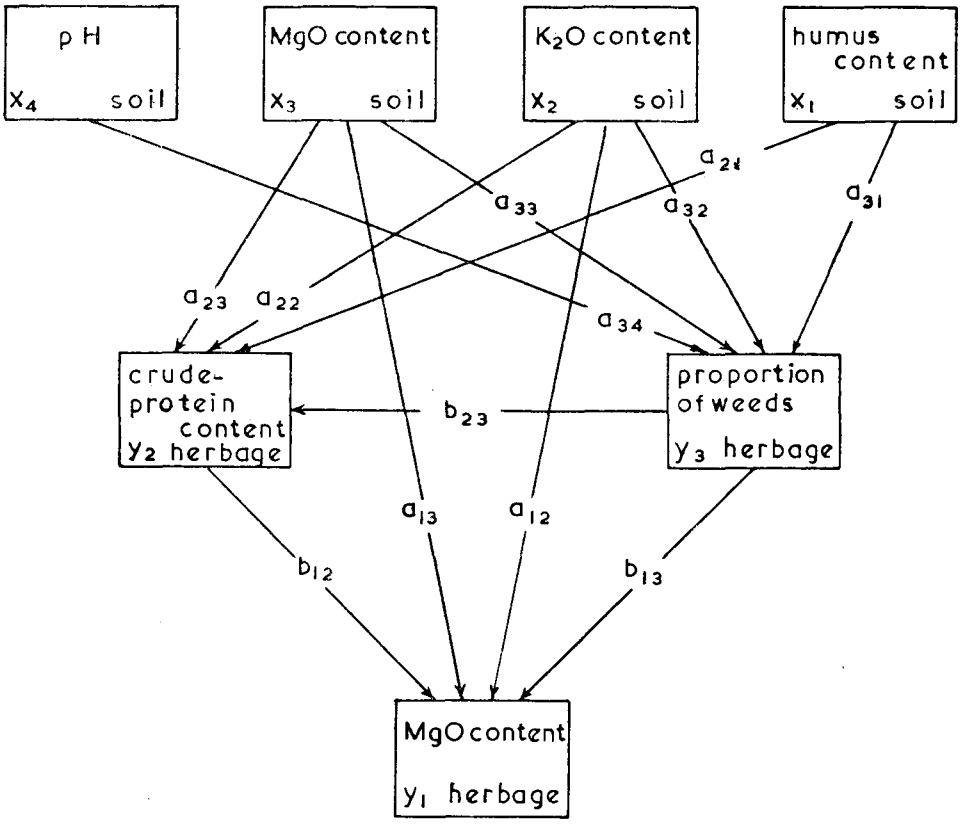
herbage. In the diagram these influences are marked with an arrow, the rate and the direction being calculated from the observations. Thus we assume that a change of the magnesium content of the soil only effects the magnesium content of the herbage but does not effect the crude-protein content and the proportion of weeds. We know however from other investigations that this is not true and the model is therefore not acceptable. Essentially we meet in this case a s.c. chain process which is not describable by means of one equation.

The diagram of FIG. 6 gives a model of these relationships probably more in agreement with reality. The variables crude-protein content and proportion of weeds are not only taken as independent variables, both variables now being cause as well as effect. A change of the magnesium content of the soil affects the magnesium content of the herbage not only directly but also indirectly via the chain: – proportion of weeds and crude-protein content. The model of FIG. 5 without these processes in the chain can be represented by one equation: –

$$y_1 = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5,$$

the second model needing a system of the following 3 equations: –

FIG. 6. Direct and indirect influences of the 4 causal factors on the magnesium content of the herbage



$$y_1 = b_{12}y_2 + b_{13}y_3 + a_{12}x_2 + a_{13}x_3$$
$$y_2 = b_{23}y_3 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3$$
$$y_3 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4$$

Such systems of equations can be solved by the *method of path coefficients*. The term "path" has something to do with pathes along which the influence is affected. By this method the hypothesis formulated in a model concerning these relationships is tested and quantified. The influence is represented by the path coefficient, giving the rate and direction of the effect change for every unit change of the causal variable. TABLE 1 gives the results of the analysis of the model shown in FIG. 6.

TABLE 1. Computed values of the 12 path coefficients of the model of FIG. 6

Effect / Cause	Humus content (x_1)	K_2O-content soil (x_2)	MgO-content soil (x_3)	pH (x_4)	Proportion of weeds (y_3)	Crude-protein content (y_2)
Proportion of weeds (y_3)	1.67	—0.23	—0.031	5.26		
Crude-protein content (y_2)	—0.74	0.11	0.011		0.20	
MgO-content of herbage (y_1) ..		—0.0038	0.0004		0.0041	0.0083

The general form of a system of equations describing a chain process is as follows: -

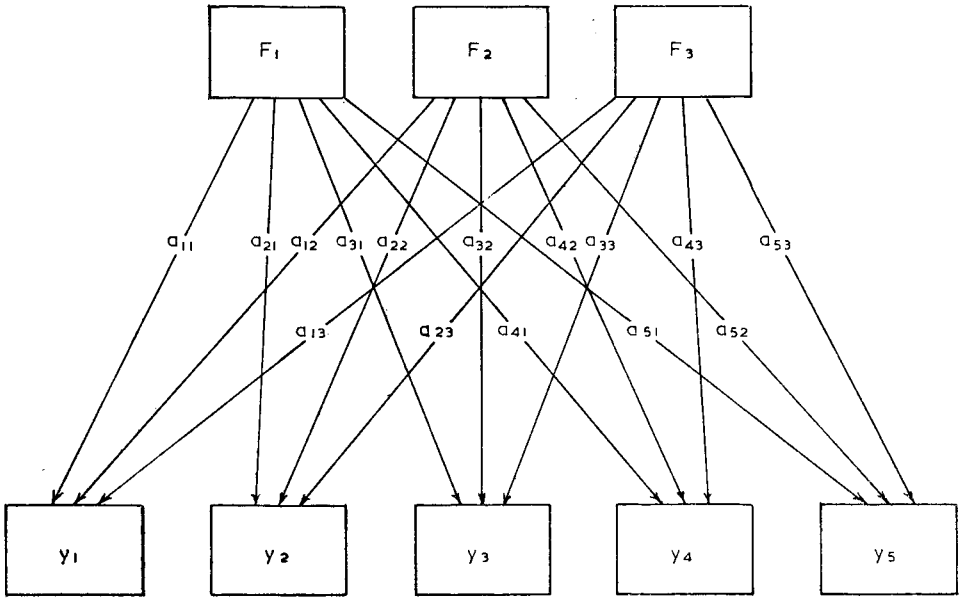
$$b_{11}y_1 + \ldots + b_{1M}y_M + a_{11}x_1 + \ldots + a_{1L}x_L = u_1$$
$$b_{21}y_1 + \ldots + b_{2M}y_M + a_{21}x_1 + \ldots + a_{2L}x_L = u_2$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$b_{M1}y_1 + \ldots + b_{MM}y_M + a_{M1}x_1 + \ldots + a_{ML}x_L = u_M$$

It is clear that some path coefficients a and b *a priori* may be assumed to be zero in real models. By means of this method it is also possible to investigate models in which feedback systems are absorbed. In our opinion such models should be preferred to the normal regression models, especially by their closer correspondence with the reality. It is possible to use non-linear functions in these systems. The method is closely related to the method of simultaneous equations out of the econometry.
An extreme case of such models is the model upon which the factor analysis is based. The number of limiting conditions in a model of the factor analysis is small, by which fact the system of equations has become s.c. unidentifiable; an exact solution can not be obtained by mathematical arguments only. The schematic diagram of such a model is given in FIG. 7. The causal x-variables (here named F) are unknown. Next the analysis tries to calculate these F-variables as s.c. aspects. This factor analysis is not only important for the testing of such models, it also can be used to give ideas for drawing up a more limited hypothesis. The possibilities of this analysis are many. The method is very suitable, for example to indicate and quantify the ecological properties of grasses grown under natural conditions. Starting point for the analysis is the matrix of correlation coefficients between soil factors and sociological characteristics, in this case the frequency percentages of grasses. The

TABLE 2. Interdependences of soil factors and frequency percentages of grasses; aspect values after rotation to simple structure

Factor	Aspects			
	1	2	3	4
pH(KCl)	0.655	—0.246	—0.209	—0.074
Humus content	0.684	—0.098	0.003	—0.240
Silt content	0.811	—0.298	—0.113	0.003
Sand content	—0.881	0.242	0.074	0.110
Specific surface sand	0.671	—0.261	—0.258	—0.028
Magnesium-content soil	0.575	—0.385	0.266	—0.152
Phosphate-content soil (water)	—0.137	0.550	0.255	0.010
Phosphate-content (citric acid)	0.650	0.184	0.112	0.243
Potash content	—0.049	0.691	0.396	—0.463
Copper content (Asp.)	0.647	—0.340	—0.096	0.055
Distance farm	0.318	0.029	—0.493	0.020
Depth clay layer	0.197	0.360	—0.040	0.380
Thickness humus layer	0.023	—0.004	—0.038	0.568
Moisture content	0.611	—0.134	—0.062	0.400
Ground-water level	0.626	—0.381	—0.075	—0.409
Fluctuation	—0.495	—0.214	0.059	—0.023
Nitrogen dressing	—0.151	0.352	0.357	0.320
Phosphate dressing	0.007	0.461	—0.037	0.059
Potash dressing	0.023	0.538	0.252	0.116
Poa pratensis L.	—0.401	0.052	0.103	0.248
Festuca rubra L.	0.383	—0.131	—0.412	0.298
Agrostis tenuis SIBTH.	—0.341	—0.259	—0.213	—0.446
Lolium perenne L.	—0.254	—0.217	0.282	—0.166
Poa annua L.	—0.219	0.252	0.563	—0.069
Alopecurus geniculatus L.	0.321	—0.301	0.336	—0.357
Agropyron repens P.B.	—0.246	0.075	0.372	0.326
Festuca pratensis HUDS.	0.787	—0.024	—0.005	0.154
Poa trivialis L.	0.314	—0.726	0.034	—0.181
Agrostis stolonifera L.	—0.105	—0.186	—0.365	—0.107
Dactylis glomerata L.	—0.111	0.069	0.156	0.342
Achillea millefolium L.	—0.270	0.072	0.227	0.181
Ranunculus repens L.	0.051	—0.427	—0.086	—0.218
Cardamine pratensis L.	0.346	—0.521	—0.025	0.004
Carex stolonifera L.	0.716	0.099	—0.165	0.043
Glyceria maxima HOLMB.	0.807	—0.018	—0.138	0.214
Ranunculus acer L.	0.297	—0.344	—0.489	—0.181
Rumex acetosa L.	0.393	—0.152	—0.414	0.272
Holcus lanatus L.	0.274	—0.183	—0.568	—0.134
Anthoxanthum odoratum L.	0.172	—0.246	—0.676	—0.020
Centaurea jacea L.	0.174	—0.201	—0.251	0.202
Bellis perennis L.	0.099	—0.482	—0.101	0.307
Cynosurus cristatus L.	0.046	—0.300	—0.232	—0.373
Alopecurus pratensis L.	0.141	—0.047	0.022	0.252
Luzula camprestis LAM. et D.C.	—0.137	0.033	—0.556	—0.023
Trifolium repens L.	—0.104	—0.203	—0.041	—0.545
Bromus mollis L.	—0.069	0.031	—0.216	0.337
Phleum pratense L.	0.024	0.011	0.309	0.124
Taraxacum officinale WEB.	—0.116	—0.083	0.129	0.671
Leontodon autumnalis L.	—0.229	—0.199	0.056	0.080
Phalaris arundinacea L.	—0.207	0.176	0.062	—0.015
Quality figure grass	0.050	—0.196	0.266	—0.103

FIG. 7. Diagram of a factor-analysis model



factor analysis with a following rotation of the aspect axes resulted in a number of aspects given in TABLE 2, of which the first aspect represents the reaction of grass to water supply. The differences in numbers running from +1 to —1 are a measure for this reaction. The positive numbers show the hydrophile character, the negative ones the drought resistence.

The most remarkable result is that it is obtained by a mathematical analysis followed by a rotation to the simple structure only. The choice of a rotation to the simple structure is based on the phenomenon already mentioned, that many factors have a small, and only few a great influence. A rotation of the model to simple structure tries to reach the same situation.

REFERENCES

FERRARI, TH. J. 1960 Vergelijking tussen proeven met en zonder ingreep. *Landbk. Tijdschr. 72*, 792—801.

——— 1963 Causal soil-plant relationships and path coefficients. *Plant and Soil. 19*, 81—96.

HEADY, E. O., J. A. SCHNITTKER, N. L. JACOBSON and S. BLOOM 1956 Milk production functions, hay/grain substitution rates and economic optima in dairy cow rations. Agr. Exp. Sta. Iowa State Coll. *Res. Bull.* No. 444.

HEADY, E. O., and J. L. DILLON 1961 Agricultural production functions. *Iowa Agric., Ames.*

HOFFMANN, E., und H. DÖRFEL 1963 Die funktionale Betrachtungsweise bei der Auswertung von Feldversuchen. *Landw. Vers.- und Untersuch.w. 9*, 75—107.

LINSER, H., und K. KAINDL 1951 Versuch einer trefferstatistischen Deutung des Mitscherlichen Ertraggesetzes. *Z. Plz. Ernähr., Düng. u. Bodenk. 53*, 47—63.