THE EXPERIMENTAL DESIGN WITH SINGLE PLOTS AND THE GRAPHICAL METHOD STATISTICAL ANALYSIS OF DATA

TH. J. FERRARI

Agricultural Experiment Station and the Institute for Soil Research
T.N.O., Groningen, Netherlands

In natural science the research worker often tries to overtake the relations between dependent and independent factors by introducing into the research a variation of one or more factors *artificially*. He concentrates particularly on the question whether a *supposed* influence of a factor really exists. The investigation is limited to some most important factors.

As agricultural science is a practical science many factors have to be considered for a right prediction. It will be clear that the usual experimental fields are too small for this purpose. Moreover, it is difficult and expensive to investigate the influence of factors which cannot be changed or are very hard to change, such as groundwater level, structure, clay content. No experimental results are, therefore, obtained for many interesting combinations of factor values.

VISSER projected a way of research, quitting the usual experimental fields and using single plots distributed over a wide area. A graphical method of statistical analysis, the *polyfactor analysis*, has been worked out to unravel the influence of all these growth factors. The analysis tries mostly to give an answer to the problem of which factors can be indicated as causing an influence; one concentrates on the problem of *describing* the data. Besides, this description is used to obtain the best (in a certain sense) estimate of the influence of a factor. There is also investigation into whether the experiment reasonably can be described omitting some influences.

In the analysis one starts from the hypothesis that there is a *functionel* relation between the growth factors and the yield, i.e., to any combination of values of the growth factors belongs a particular value of the yield. In practice one will find yields which do not coincide with this yield, because the actual yield forms a sample of a probability distribution.

The *expectation value* of the yield (the probability distribution) $E(y)$ is the mentioned function of the growth factors: $E(y) = \hat{y} = f(x_1, x_2, \dots . x_n)$. The purpose of the research will be to estimate this function with the aid of the data.

A second hypothesis is that this function is of the particular form (a *joint* function of EZEKIEL, 1947):

$$y = f_1(x_1) + \dots . f_n(x_n) + f_{1\cdot2}(x_1x_2) + f_{1\cdot3}(x_1x_3) + f_{1\cdot2 \dots . n}(x_1x_2 \dots . x_n) + C$$

The functions $f_i(x_i)$ indicate that to, e.g., 2 values of a certain factor x_i yields of a different amount belong. This difference in yield depends furthermore on the values of other factors (*interaction*). The interactions are absent if the functions with more than one argument in this formula are zero. The interactions found in practice could be satisfactorily described with this formula.

The purpose of the research is to draw up a regression equation of the above-mentioned type, which describes the experiments satisfactorily. It is, therefore, necessary that the *mathemcticcl* form of this regression equation be known or reasonably assumed, after which the parameters in this equation with the aid of the experiments can be determined. The assumption of the form of the equation is essentially not a mathematical but an agricultural question. The data obtained by the experiment may give a *suggestion* about the possible form.

The theoretical knowledge of the influence of the growth factors is very small and consequently the form of the regression function is mostly unknown. Moreover, if a form is given, the process of the determination of the parameters is mostly very difficult and takes much time. A graphical method of statistical taking together is here used: a fluent line is drawn by hand, joining as well as possible the averages of the columns. Consequently the suggestion given by the data is fully used.

The figures 1, 2a, 2b, 3a, 3b and 4 show schematically the relation between the yield and respectively 1, 2 and 3 growth factors. These relations are represented mathematically by the equations

$y=f(x_1)$, $y=f(x_1, x_2)$ and $y=f(x_1, x_2, x_3)$.

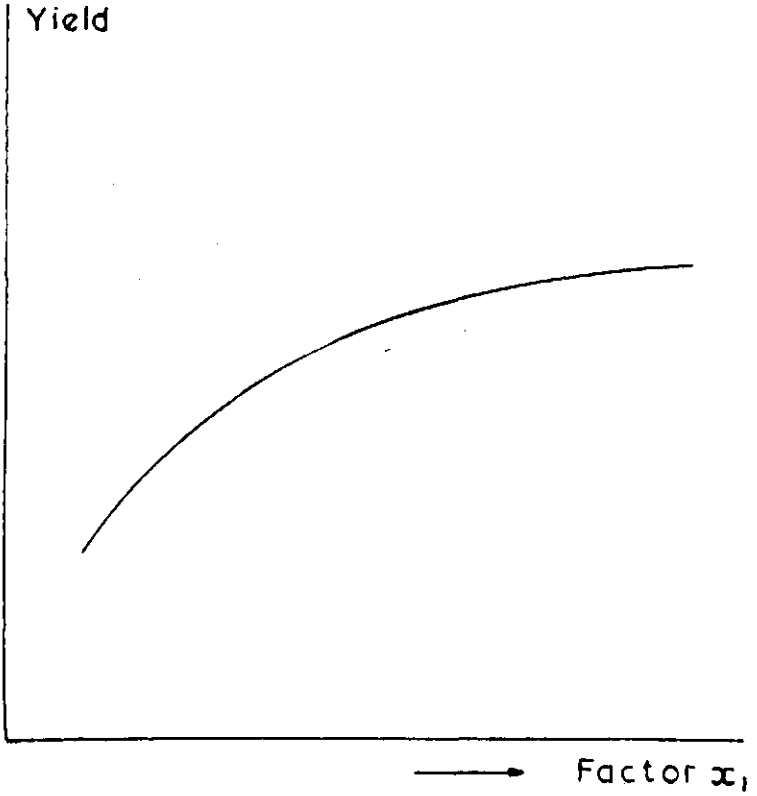

Fig. 1—Graph of the influence of the variation of a factor x *on the yield.*

The relation e.g., $y=f(x_1, x_2)$ is visualized in the three-dimensional diagram of the figures 2a and 3a. In the analysis a two-dimensional representation is used (figures 2b and 3b) : a graph of the function $y=f(x_1, x_2)$ (x_2 constant) is given at different values of x_2.

The relation $y=f(x_1, x_2)$ in the figures 2a and 2b is supposed to be $y=f_1(x_1)+f_2(x_2)$, in which $f(x_1, l^i) - f(x_1, l^k)$ is constant. The relation $y=f(x_1, x_2)$ in the figures 3a and 3b, however, is supposed to be $y=f_1(x_1)+f_2(x_2)+f_{1\cdot 2}(x_1\ x_2)$, in which $f(x_1, l^i) - f(x_1, l^k)$ is not constant (interaction).

In an analogous way it is possible to show the influences of n factors by using geometrical terms, which, when bounded by a three-dimensional space, can be represented geometrically. In the analysis one determines each time the influence of a growth factor on the yield, the other factors being held (approximately) constant at different values.

The more factors are studied at the same time, the more data are necessary, so that in complicated multi-dimensional problems the supposition must be made that interactions of higher degrees are absent. This supposition is probably permissible from the agricultural viewpoint as the frequency of the parcels, which are in an extreme situation for several factors at the same time, is relatively low.

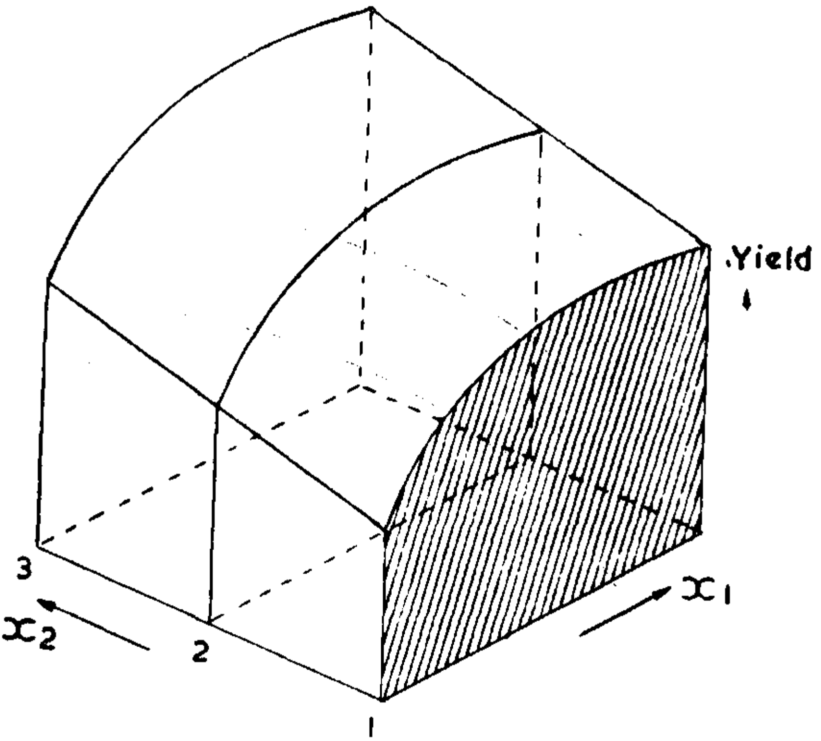

Fig. 2a—Graph of the influence of the variation of a factor x_1 *on the yield, at different values of a factor* x_2*, without interactions. Fig. a—3-dimensional, fig. b—2-dimensional.*

In the literature on the subject three adjustment lines are known : (*a*) and (*b*), the two regression lines, one of the coordinates is supposed to be errorless, (*c*) the line obtained by an adjustment perpendicular to this line, which has to be taken if both coordinates are equally inaccurate. VISSER pointed out that regression lines may be of no use ; usually the error of the first coordinate is not small with respect of the error of the second coordinate. He, therefore, introduced a so-called direction of averaging, which takes into account the error of all coordinates and which is based on the relation between the gradients of two conjugate diameters of an ellipse.

Once a relation $y=f(x)$ is found, it is possible to eliminate this influence (*correction*), by which the variance of the yield data is decreased. The correction value is the amount equal to the difference between the correction level and the expectation value of the yield corresponding to the values of the growth factor(s) on every plot. That yield is taken as the correction level for which $\Sigma(f(x^k) - {}_{y}$ correction level) is zero.

More or less high correlations between the growth factors are introduced into the research, because they are not avoided in this method. It is necessary to eliminate the error in the conclusions, caused by these correlations. This may be done in the same way

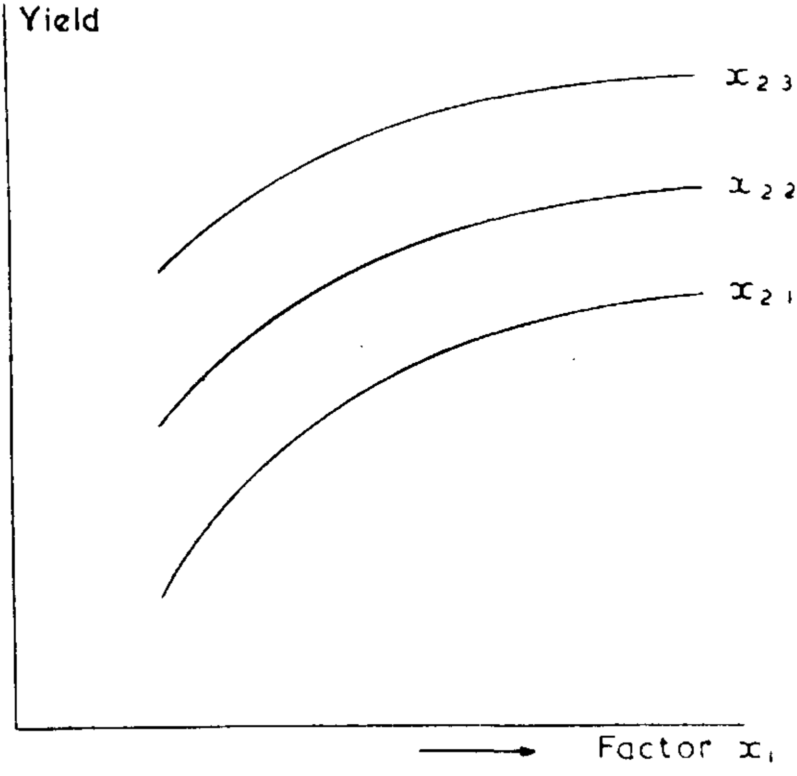

Fig. 2b

as has already been described showing the existence of the interactions; the number of constant values (groups) depends on the highness of the correlations. In our research each time three groups were used.

If the number of growth factors is n and if the number of groups, in which the extent of variation of a factor is divided, is s, the number of necessary data is proportional with $c/n \times s^{n-1}$; here is $c < 1$. Obviously a complete more-dimensional treatment demands a very large number of data. Generally the available number of data is not sufficient for a complete more-dimensional treatment. In such cases a process of *iteration* can be applied.

This process may be demonstrated by the following example. The influence of e.g., 2 factors is assumed to be expressible in the form $y=f(x_1, x_2)=f_1(x_1)+f_2(x_2)$; in the analysis $\Sigma(y-f(x_1, x_2))^2$ must be a minimum. In the process of iteration the function $y=f_1^1(x_1)$ is first determined in such a way that $\Sigma(y-f_1^1(x_1))^2$ is a minimum. After the correction for $y-f_1^1(x_1)$ the function $f_2^1(x_2)$ is determined; this postulates that $\Sigma(y-f_1^1(x_1)-f_2^1(x_2))^2$ is a minimum. In the second tour the function $y=f_1^2$ is determined; now $\Sigma(y-f_1^2(x_1)-f_2^1(x_2))^2$ must be a minimum. And so on. The functions $f_1^r(x_1)$ and $f_2^r(x_2)$ are the best preliminary estimates of the influences, if in the $(r+1)$-th tour the estimate of $f_1(x_1)$ shows no change. It is possible to get an ultimate best estimate by choosing the best one out of all the function $c_1 . f_1(x_1)+c_2 . f_2(x_2)$ (see further).

With what factors the analysis should be started, depends on:

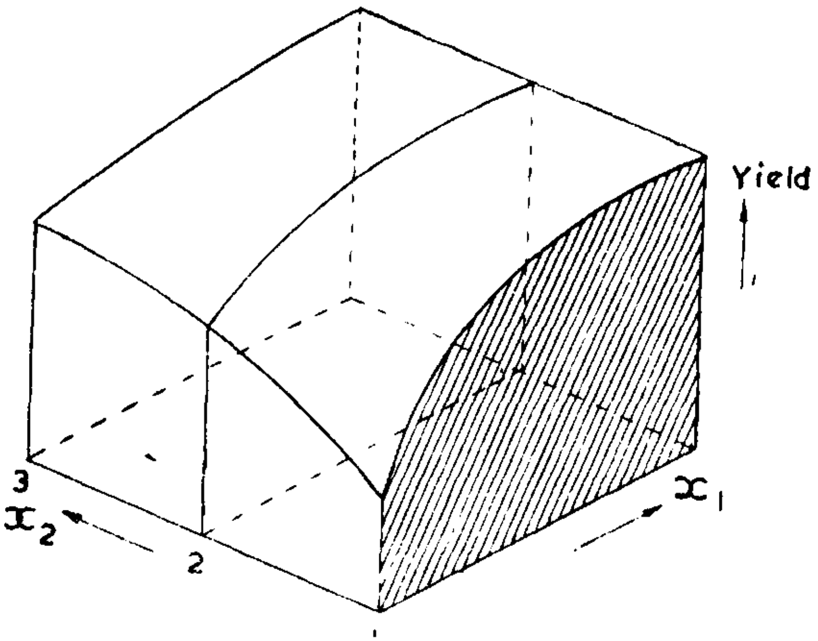

Fig. 3a—Graph of the influence of the variation of a factor x_1 *on the yield, at different values of a factor* x_2*, with interactions. Fig. a—3-dimensional, fig. b—2-dimensional.*

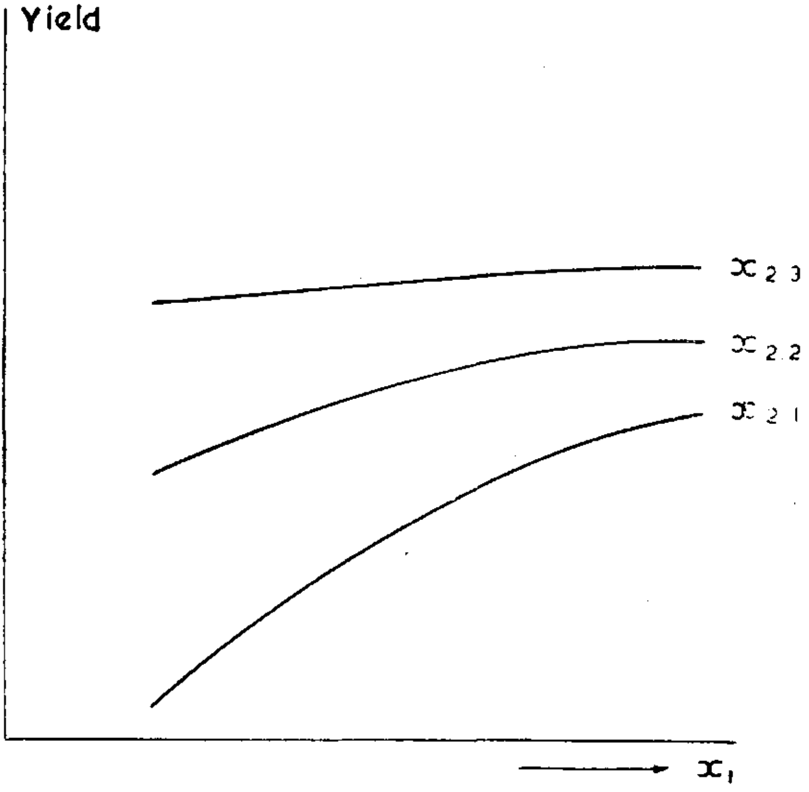

Fig. 3b

(*a*) the degree of the correlation between the several growth factors;

(*b*) the degree of the influence of each separate factor;

(*c*) the degree of the interaction.

It is possible to investigate whether the results obtained by this graphical method are reliable in a mathematical-statistical respect and whether all important growth factors have been really found.

An analysis of variance can be made for the sake of an investigation into the statistical reliability; this done, the ratio of the variances of the corrections and of the rest could be calculated. The theoretical expectation value of this ratio in a *O*-hypothesis is equal to 1; the F-table gives the probability of the ratio $\geq$the ratio found.

A second method of investigating the reliability of the results is to choose with the aid of the experimental data the best one out of all the regression equations $y=a_1.f_1(x_1)+a_2.f_2(x_2)+....+a_n.f_n(x_n)$. In these equations a_1, a_2, ... a_n are constants the best estimates of which must be calculated and $f_1(x_1)$, $f_2(x_2)$, ... $f_n(x_n)$ are the functions found graphically. The number of plots being *m* there are m equations $y^k = a_1.f_1(x^k{}_1) + a_2.f_2(x^k{}_2) + ... + a_n f_n(x^k{}_n)$. The normal equations give the best estimates $\bar{a}_1$, $\bar{a}_2$, ... $\bar{a}_n$. The following table gives the results of such a calculation (FERRARI, 1952).

Growth factor	$\bar{a}_1$	$s\bar{a}_1$	$\bar{a}_i'$	a''_i	$p(a'_i = o)$
Potash content and acidity	1,089	0,04	1,009	1,169	0
Potash dressing	0,964	0,11	0,744	1,184	0
Clay content of topsoil ...	1,153	0,18	0,793	1,513	0
Distance to farmstead ..	0,965	0,18	0,605	1,325	0
Data of planting	0,807	0,25	0,307	1,307	0,001
Organic matter content ...	1,263	0,31	0,643	1,883	0
Groundwater level ...	1,779	0,38	1,019	2,539	0
Depth of reduction ...	0,763	0,29	0,183	1,343	0,005
Structure	0,525	0,49	−0,455	1,505	0,136
Distribution of water-stable aggregates ...	0,137	0,68	−1,223	1,497	0,421
Groundwater fluctuation	1,508	0,95	−0,392	3,408	0,065
Clay content of subsoil ...	0,492	0,40	−0,308	1,292	0,115

The best estimate of a_i and the standard error $s\bar{a}_1$ are calculated from the equations $y^k = a_1.f_1(x_1{}^k) + a_2.f_2(x_2{}^k) + ... + a_{12} f_{12}(x_{12}{}^k)$ in which $k = 1, 2, \ldots\ldots 222$. In these equations, y^k represents the yield on plot k and $f_i(x_i{}^k)$ the effect of a growth factor or of a combination of growth factors (=correction value) on a plot k. Furthermore

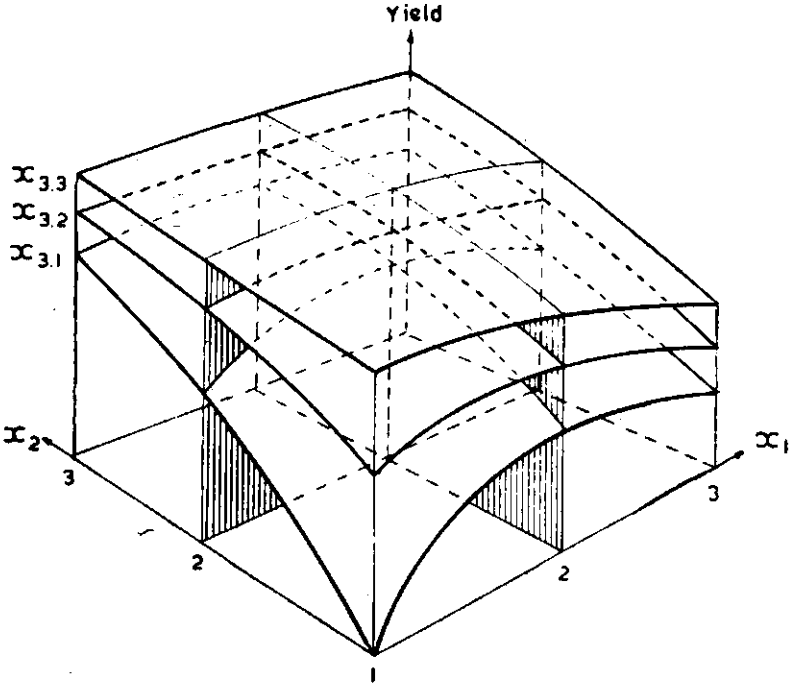

Fig. 4—Graph of the influence of the variation of 3 factors on the yield, with all interactions.

a'_i and a''_i are limits beyond which a_i is lying with a probability p=0,05. Also with the aid of the data the probability p can be calculated if a'_i=0 and a''_i=$\sim$ *is*. This coefficient p being >0,05 then $f_i(x_i)$=0 may be an acceptable estimate and the influence of the factor, found graphically, is not reliable. We expect a value $\bar{a}_i$ *is* about 1. If $\bar{a}_i$>1 *is*, the influence is underestimated; if $\bar{a}_i$<1, the influence is overestimated; the graph of the ultimately best estimate of the influence may be found by changing the yield ordinate with a factor $1/\bar{a}_i$.

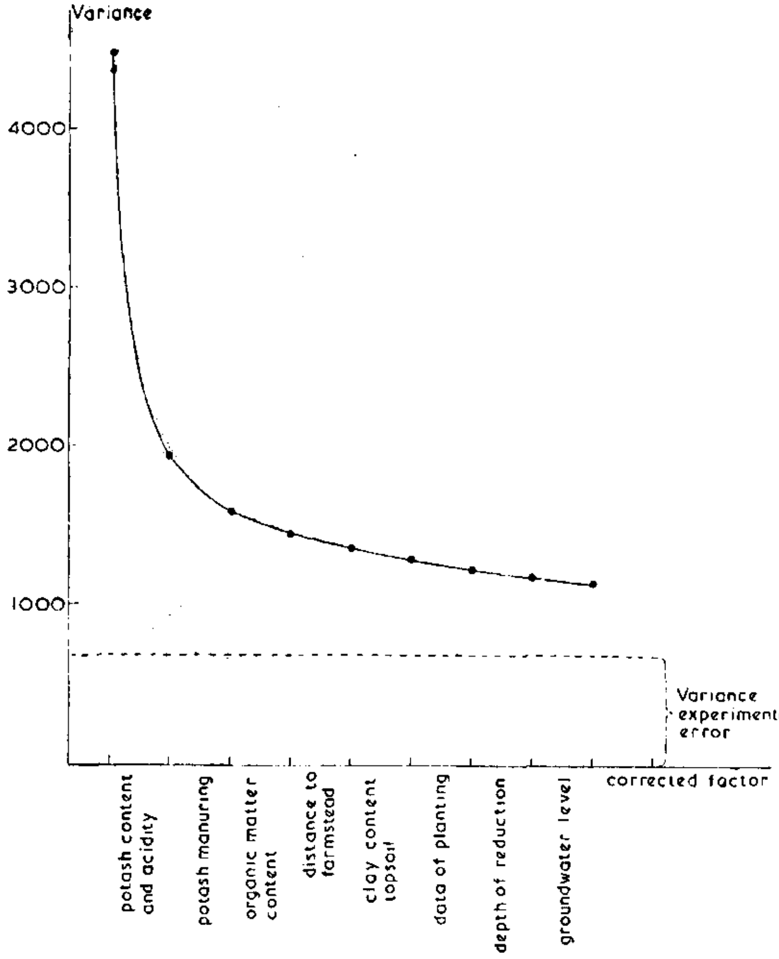

Fig. 5—*Size of each variance after successive corrections.*

Figure 5 gives an answer to the question whether all important factors have been really found. It appears from the figure that about 75% of the variance in the above-mentioned research can be explained from the nine significant factors. If the variance caused by the experimental error and amounting to 676 is taken into account, it appears moreover that 88% of the explicable variance has been explained. This makes it plausible that important factors have not been neglected.

LITERATURE

EZEKIEL, M.: Methods of correlations. New York, 1947.

FERRARI, Th. J.: An agronomic research with potatoes on the river ridge soils of the Bommelerwaard. *Versl. Landbouwk. Onderz.*, **58**, 1 (1952).