EEN EN ANDER OVER DE POLYFAKTORANALYSE

door

DR. IR. TH. J. FERRARI
(Landbouwproefstation en Bodemkundig Instituut T.N.O.)

„It is much more difficult to be a good farmer than a good mathematecian because the farmer must deal with so many vague and complex problems". (IRWIN BROSS, Design for Decision)

Een van de hulpmiddelen om de kausale afhankelijkheid tussen verschillende faktoren vast te stellen, is het experiment. Hierbij wordt doelbewust een variatie bij een of meer faktoren aangebracht. Deze variatie wordt met de eventuele verandering in het te verklaren verschijnsel in verband gebracht.

Deze opzet wordt in het bodemvruchtbaarheidsonderzoek veelvuldig toegepast. De beperktheid van de mogelijkheden die met deze opzet samenhangen, wordt bij dit landbouwkundig onderzoek dikwijls gevoeld. Bij de adviezen aan de boer op het terrein van de bodemvruchtbaarheid moet met vele faktoren rekening worden gehouden. De proefvelden die bij het onderzoek worden gebruikt zijn te klein voor dit doel. Het is dikwijls moeilijk en duur de invloed van faktoren te onderzoeken, die niet of zeer moeilijk te veranderen zijn, zoals grondwaterstand, struktuur, slibgehalte enz. Men wordt dan gedwongen om bij het onderzoek van de in de natuur voorkomende variaties gebruik te maken. Bij het verwerken van de verkregen resultaten treden problemen op betreffende lijnvereffening, wiskundige formulering, invloed van korrelaties tussen onafhankelijke faktoren, enz. Een grafische verwerking, waaraan in Nederland de naam van W. C. Visser verbonden is, kan hierbij worden toegepast. Het onderzoek tracht voornamelijk een antwoord te geven op de vraag, welke faktoren aangegeven kunnen worden, die een invloed uitoefenen; men concentreert zich op een beschrijving van de gegevens. Deze beschrijving kan verder worden gebruikt om de beste schatting van de invloed te krijgen. Men gaat ook na of de gegevens goed beschreven kunnen worden wanneer enkele invloeden niet verondersteld worden.

Men begint met de veronderstelling, dat er een funktioneel verband tussen b.v. de groeifaktoren en de opbrengst bestaat, d.w.z. bij elke kombinatie van waarden van groeifaktoren behoort een bepaalde opbrengst. De verwachtingswaarde is dan een funktie van de groeifaktoren, dus

$$E(y) = \hat{y} = f(x_1, x_2, \ldots x_n).$$

Het doel van het onderzoek is deze funktie met behulp van de gegevens te schatten.

Een tweede veronderstelling is, dat deze funktie er een is van een bepaalde soort, namelijk

$$y = f_1(x_1) + \ldots f_{1.2}(x_1x_2) + f_{1.3}(x_1x_3) + \ldots f_{1.2\ldots n}(x_1x_2\ldots x_n) + c.$$

De funkties $f_i(x_i)$ geven aan, dat bij b.v. twee waarden van een faktor x_i verschillende opbrengsten behoren. Dit verschil hangt eventueel af van de waarde van andere faktoren (interaktie); de interakties zijn afwezig indien de funkties met meer dan een argument gelijk aan nul zijn.

Het doel van het onderzoek is, een regressievergelijking van bovenstaande vorm te vinden, die de opbrengstgegevens voldoende beschrijft. De theoretische kennis over de invloed van een groeifaktor is in de landbouw klein en de vorm van de regressiefunktie is meestal onbekend. Bovendien is het bepalen van de parameters, indien de funktie gegeven is, vaak moeilijk en vraagt veel tijd. Wij gebruiken daarom een grafische vereffening; door het gemiddelde van de puntenzwerm wordt een vloeiende lijn getrokken; de door de gegevens aangebrachte suggestie wordt dus volledig gebruikt.

Het is mogelijk de invloed van n factoren weer te geven door gebruik te maken van meetkundige termen, die bij snijding van de driedimensionale ruimte meetkundig voor te stellen zijn. Men zal er dan toe moeten overgaan dit te doen door de invloed van slechts één of twee faktoren op de opbrengst weer te geven bij verschillende toestanden van andere faktoren. Voor 2 faktoren is dit in figuur 1 weergegeven, waarbij de veronderstelde regressievergelijking is:

$$y = f_1(x_1) + f_2(x_2) + f_{1.2}(x_1 x_2)$$

In de twee-dimensionale grafiek wordt een grafische voorstelling $y = f(x_1,x_2)$ gemaakt bij verschillende waarden van x_2. In geval van optreden van interakties vindt men dat $f(x_1,b^l) - f(x_1, b^k)$ niet konstant is (a, b enz. geven konstante waarden van x_1, x_2 enz. aan). De bewerking van meer-dimensionale gevallen is analoog, zodat om de funktie $y = f(x_1,x_2, \ldots x_n)$ vast te stellen principiëel meer-dimensionaal gewerkt moet worden. De werkwijze bestaat hierin, dat allereerst $y = f(x_1,b, c \ldots)$ wordt nagegaan, zodat de invloed van de eerste faktor bij verschillende kombinaties van konstante waarden van andere faktoren wordt bepaald. Vervolgens kan dan $y = f(a, x_2, c \ldots)$ bepaald worden enz.

Hoe meer faktoren in het onderzoek worden opgenomen, des te meer gegevens zijn noodzakelijk, zodat in ingewikkelde meer-dimensionale problemen de veronderstelling moet worden gemaakt dat interakties van hogere orde dan b.v. twee afwezig zijn. In landbouwkundig opzicht is deze veronderstelling wel toegestaan, omdat het percentage percelen, dat tegelijk in extreme toestanden voor verschillende faktoren verkeert, betrekkelijk laag is.

Het is mogelijk, wanneer eenmaal een invloed $y = f(x)$ gevonden is, deze invloed te elimineren (= korrektie), waardoor de variantie van de opbrengst verkleind wordt. De korrektie is het bedrag gelijk aan het verschil tussen het korrektieniveau en de verwachtingswaarde van de invloed behorende bij de waarde van de groeifaktor(en) in elk geval. Als korrektieniveau wordt de opbrengst genomen waarbij $\Sigma\,(f(x^k) - y^{\text{korrektieniveau}})$ gelijk aan nul is.

Wij hebben er reeds op gewezen, dat er tussen de verschillende onafhankelijke faktoren min of meer sterke korrelaties bestaan. Door de invloed van een faktor bij binnen zekere grenzen konstante waarden van een of meer faktoren na te gaan, zoals dat ook bij een onderzoek naar de interakties gebeurt, wordt de fout, die kan ontstaan

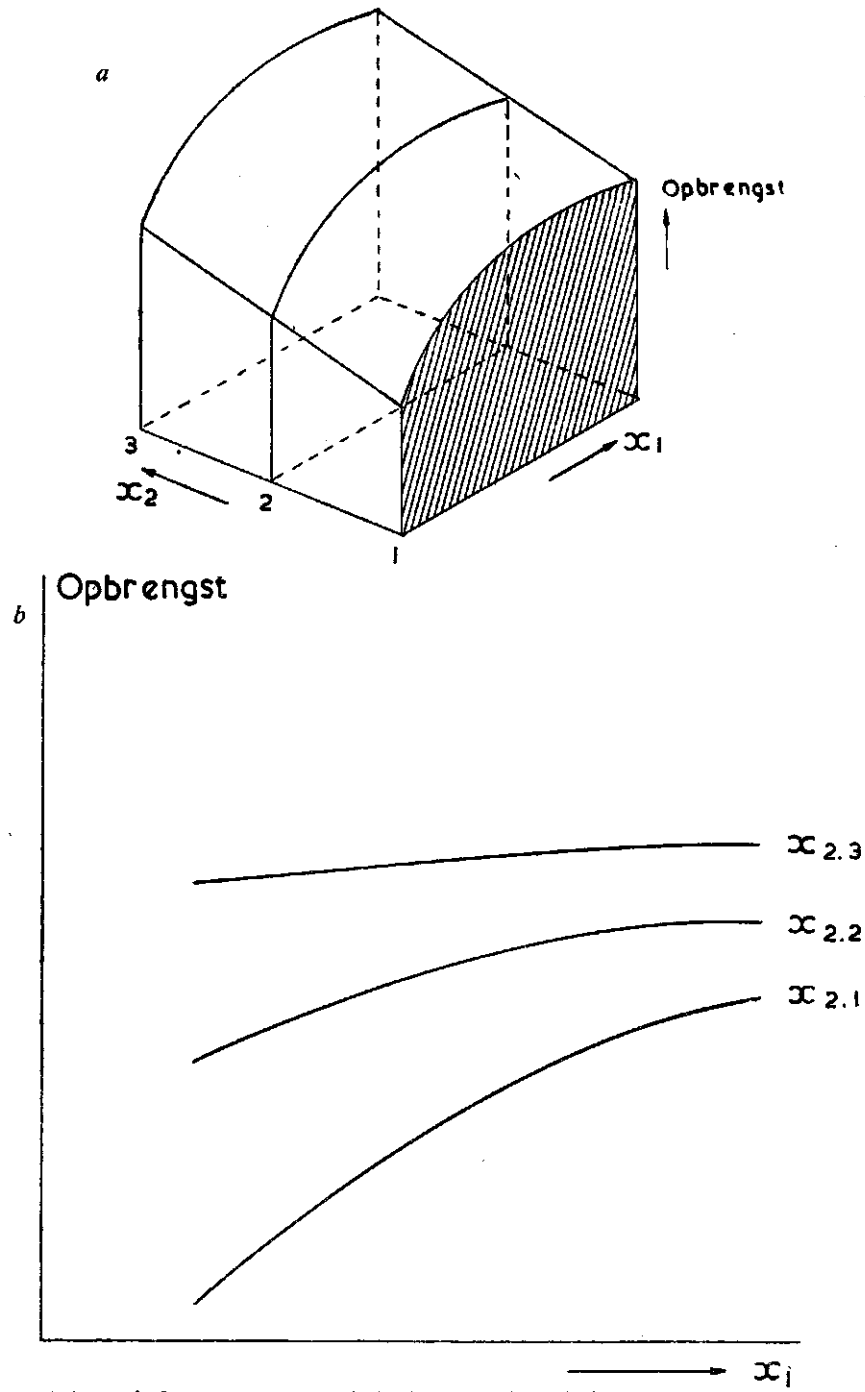

FIG. 1. Schematische weergave van de invloed van de variatie van faktor x_1 op de opbrengst bij verschillende toestanden van faktor x_2.
Fig. *a* 3-dimensionaal, fig. *b* 2-dimensionaal.

door het aanwezig zijn van korrelaties, geëlimineerd en wordt de zuivere invloed gevonden. Het aantal groepen van konstante waarden hangt af van de sterkte van de korrelaties. Over het algemeen is het aantal gegevens niet voldoende voor een volledig meer-dimensionale bewerking, waardoor een iteratieproces moet worden toegepast.

Deze iteratie kan met het volgende voorbeeld verduidelijkt worden. Wij veronderstellen, dat de invloed van b.v. 2 faktoren op de opbrengst door de vergelijking

$$y = f_1(x_1) + f_2(x_2)$$

wordt weergegeven; in de bewerking moet $\Sigma\,(y - f(x_1,x_2))^2$ een minimum zijn. In het iteratieproces wordt de funktie $y = f_1{}^1(x_1)$ eerst bepaald, waarbij $\Sigma\,(y - f_1{}^1(x_1))^2$ een minimum is. Na een korrektie voor $y = f_1{}^1\,(x_1)$ wordt de funktie $y = f_2{}^1(x_1)$ bepaald; hier geldt weer dat $\Sigma\,(y - f_1{}^1(x_1) - f_2{}^1(x_1))^2$ een minimum is. In de tweede ronde wordt de funktie $y = f_1{}^2(x_1)$ bepaald; nu moet $\Sigma\,(y - f_1{}^2(x_1) - f_2{}^1(x_2))^2$ een minimum zijn, enz. De funkties $f_1{}^r(x_1)$ en $f_2{}^r(x_2)$ zijn de voorlopig beste schattingen van de invloeden, indien in de $(r + 1)$ - de ronde de schatting van $f_1(x_1)$ geen verandering vertoont. De uiteindelijke schatting verkrijgt men door uit alle funkties $a_1.f_1(x_1) + a_2.f_2(x_2)$ weer de beste te kiezen.

Het vaststellen van de wiskundige betrouwbaarheid van de verkregen resultaten geschiedt als volgt. Wij wezen reeds op de mogelijkheid om met behulp van de experimentele gegevens de uiteindelijk beste schatting van de invloeden te verkrijgen. Het bepalen van deze beste schatting komt overeen met een bewerking, waarbij een in gedaante ontworpen, als hypothese gestelde regressievergelijking nader kwantitatief gepreciseerd wordt door met behulp van de experimentele gegevens de konstanten door de beste schattingen te vervangen. De in getallen vastgelegde hypothese kan daarna aan de experimentele gegevens worden getoetst.

De als hypothese gestelde vergelijking is nu

$$y = a_1.f_1(x_1) + a_2.f_2(x_2) + \ldots a_n.f_n(x_n)$$

zodat m vergelijkingen

$$y^k = a_1.f_1(x_1{}^k) + a_2.f_2(x_2{}^k) + \ldots a_n.f_n(x_n{}^k)$$

gegeven zijn. In de vergelijkingen vertegenwoordigt y^k de op het veldje k gevonden opbrengst en geeft $f_i(x_i{}^k)$ het effekt aan, dat een groeifaktor of een kombinatie van groeifaktoren op de opbrengst van het veldje k heeft. Met behulp van de normaalvergelijkingen worden de beste schattingen a_i en bijbehorende standaardafwijkingen $s_{\bar{a}_i}$ uitgerekend. De resultaten van een dergelijke berekening in een bepaald geval zijn in de volgende tabel weergegeven.

groeifaktor	$\bar{a}_i$	$s_{\bar{a}_i}$	a_i'	a_i''	$p(a_i'=0)$
kaligehalte en zuurgraad	1.089	0.04	1.009	1.169	0
kalibemesting	0.964	0.11	0.774	1.184	0
slibgehalte bovengrond	1.153	0.18	0.793	1.513	0
afstand tot boerderij	0.965	0.18	0.605	1.325	0
pootdatum	0.807	0.25	0.307	1.307	0.001
humusgehalte	1.263	0.31	0.643	1.883	0
grondwaterstand	1.779	0.38	1.019	2.539	0
visuele struktuur	0.525	0.49	0.455	1.505	1.36

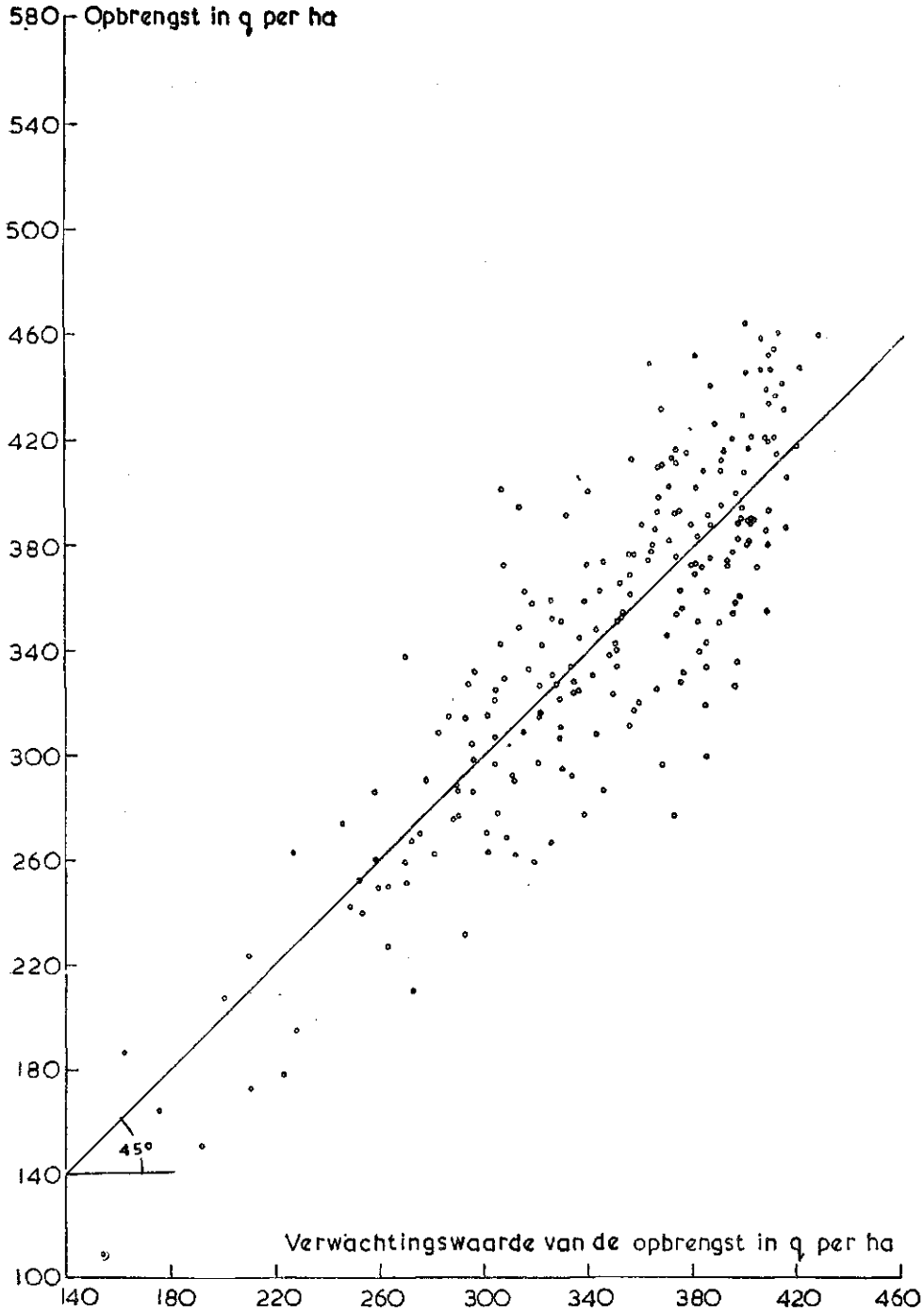

FIG. 2. Het verband tussen de verwachtingswaarde van de opbrengst en de waargenomen opbrengst.

aggregaat-analyse	0.137	0.68	–1.223	1.497	0.421
fluktuatie	1.508	0.95	–0.392	3.408	0.065
zwaarte ondergrond	0.492	0.40	–0.308	1.292	0.115
reduktie	0.763	0.29	0.183	1.343	0.005

De waarden a_i' en a_i'' zijn de grenzen waarbuiten a_i met een waarschijnlijkheid van 0.05 ligt. Ligt 0 binnen deze grenzen, dan is $f_i(x_i) \equiv 0$ ook een aannemelijke schatting en kan men deze veronderstelde invloed laten vervallen. Men kan ook de waarschijnlijkheid berekenen, dat $a_i' = 0$ en $a_i'' = \infty$. Wanneer deze waarde > 0.05 is, dan is $f_i(x_i) \equiv 0$ ook een aannemelijke schatting met dezelfde konklusie als hierboven.

Wij verwachten een waarde voor $\bar{a}_i$ die in de buurt van 1 ligt. Indien $\bar{a}_i > 1$ is, dan is de invloed onderschat, indien $\bar{a}_i < 1$ is, dan is de invloed overschat. In beide gevallen kan door een verandering in de opbrengstschaal van de grafiek de beste schatting worden aangegeven.

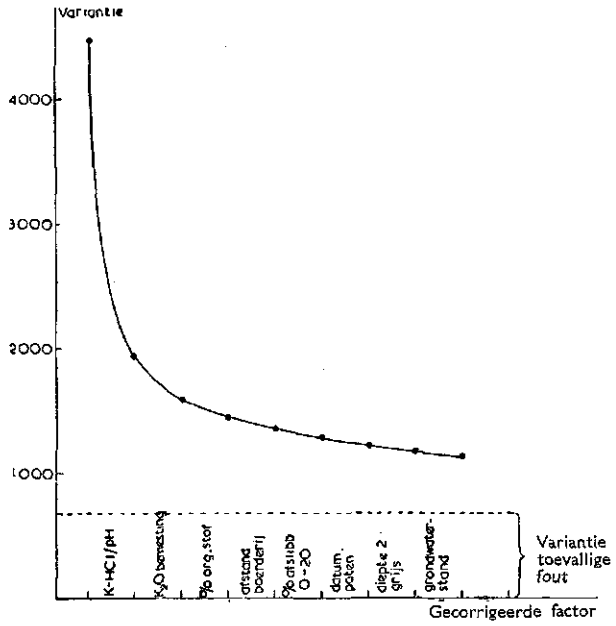

FIG. 3. De grootte van elke variantie na successievelijke korrekties.

De figuren 2 en 3 geven het resultaat van een hierboven beschreven grafische bewerking weer. Bij een onderzoek met aardappelen werden dertien faktoren aangeduid (zie tabel), die van betekenis waren voor de grootte van de opbrengst. Het was nu mogelijk voor elk geval (222 gevallen) de opbrengst te berekenen en deze berekende opbrengst met de werkelijk gevondene te vergelijken. De overeenkomst wordt in figuur 2 gedemonstreerd; de korrelatiekoëfficiënt bedraagt 0.86, welk bedrag in verband met het optreden van toevallige fouten hoogstens 0.92 had kunnen zijn. De betekenis van de verschillende faktoren wordt in figuur 3 weergegeven.