KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Verslagen V-66
(R-III-252-1960)

De methode der regressieintegralen ter bepaling van de invloed van weerfaktoren op de opbrengst van gewassen.

door

H. de Hart

Inhoud

1. Inleiding. Het begrip continue regressie.

De opbrengst van cultuurgewassen vertoont van jaar tot jaar vrij grote fluktuaties. Voor een deel kunnen deze fluktuaties worden toegeschreven aan verschil in bemesting, grondbewerking, enz.; voor een groot deel worden ze veroorzaakt door de van jaar tot jaar wisselende weersomstandigheden, waaronder het gewas groeit en rijpt. De eerste twee genoemde faktoren heeft men onder contrôle en de invloed ervan kan vrij eenvoudig worden nagegaan. De faktor "weer" is echter dermate gecompliceerd, dat het vrij moeilijk is de invloed ervan te analyseren. Immers, het weer op een bepaald tijdstip of gedurende een zeker tijdvak wordt gevormd door een combinatie van weerelementen op dat tijdstip of gedurende dat tijdvak, zoals zonneschijnduur of -intensiteit, neerslag, temperatuur, enz. Deze weerelementen kunnen gedurende een zeker tijdvak invloed uitoefenen op de uiteindelijke opbrengst van gewassen en worden daarom weerfaktoren genoemd. Dat deze weerfaktoren invloed uitoefenen zal duidelijk zijn als men bedenkt dat zij stuk voor stuk wel een extreme waarde kunnen bereiken, waarbij zelfs sterke invloed op het gewas wordt uitgeoefend. We denken maar aan harde wind, extreme droogte, zeer veel regen, enz. We zullen echter zeggen, dat er van invloed sprake is, indien langs statistische weg kan worden aangetoond, dat een deel van de spreiding in de opbrengsten werkelijk wordt veroorzaakt door de betreffende weerfaktor. Zo deze invloed bestaat, dan zal ze zich in het algemeen uitstrekken over de gehele groeiperiode van het gewas; misschien zelfs over enige tijd ervoor, doch zeker niet er na. Verder zal de grootte (β) van die invloed niet gedurende de gehele periode constant zijn, doch met de tijd variëren; m.a.w. β kan geschreven worden als een funktie van de tijd:

$$\beta = f(t) \qquad (1.1)$$

Hierbij wordt verondersteld, dat deze funktie voor elk jaar dezelfde is. Dit is niet geheel juist. Het zou beter zijn β als funktie van de fenologische stadia van het gewas te beschouwen. Bij gebrek aan gegevens over de fenologische stadia echter, zullen we ons moeten behelpen met kalenderperioden, in de hoop dat de diverse stadia van het gewas van jaar op jaar niet te veel spreiden. Daar de grootte van de beschouwde weerfaktor (x) eveneens met de tijd varieert, kan deze ook als een funktie van de tijd worden geschreven:

$$x_i = g_i(t) \quad \text{in het jaar } i \; . \qquad (1.2)$$

Het verschil met (1.1) is, dat deze funktie doorgaans voor elk jaar een andere is, aangezien elk jaar de weerfaktor anders varieert. Wordt eenvoudigheidshalve verondersteld, dat het verband tussen de weerfaktor en de opbrengst lineair is, dan geeft de grootte van β in zeker tijdsinterval tussen t en t+dt, vermenigvuldigd met de grootte van x in dat interval de werkelijke invloed weer:

$$\beta x_i = f(t).g_i(t)\,dt \qquad (1.3)$$

En de totale invloed in het jaar i wordt dan voorgesteld door:

$$\int_a^b f(t) \cdot g_i(t)\, dt \qquad (1.4)$$

welke uitdrukking de regressieintegraal wordt genoemd.
Hierin stellen a en b het begin- resp. eindpunt voor van de periode, waarover de invloed zich uitstrekt.
De opbrengst (y) kan dan worden voorgesteld door:

$$y_i = \gamma + \int_a^b f(t) \cdot g_i(t)\, dt \qquad (1.5)$$

De uitdrukking (1.5) noemt men de <u>continue regressiefunktie</u> van y op x.
Het begrip continue regressie is ingevoerd door R.A. Fisher en uitgewerkt bij een onderzoek naar de invloed van neerslag op de opbrengst van tarwe [1] . Het wordt eveneens beschreven door Sanderson [2].

2. <u>Probleemstelling</u>.

Heeft men gegeven:

1) de opbrengst-cijfers per oppervlakte-eenheid van een gewas (y) gedurende n jaren,
2) de weerfaktor (x), waarvan men de invloed op y wil nagaan, over dezelfde n jaren en voor ieder jaar vanaf enige tijd voor de aanvang van het groeiseizoen tot aan de oogst van het gewas.

Het probleem is nu: Hoe kan men uit de onder 2) genoemde gegevens voor elk jaar de beste schatting vinden voor $x = g(t)$ en uit de parameters van deze funkties, gecombineerd met de onder 1) genoemde gegevens de beste schatting van $\beta = f(t)$.

3. <u>Oplossing van het probleem</u>.

Verdeelt men het tijdvak a-b in m tijdvakken van korte duur en gelijke lengte (bijv. dekaden of pentaden), dan kan men voor elk van die tijdvakken de gemiddelde waarde van x bepalen. Dan heeft men dus voor elk jaar m waarden, welke de funktiewaarden zijn van $x = g(t)$ indien men $t = 1, 2, 3 \ldots m$ stelt.
Men kan voor elk jaar i $g_i(t)$ bij benadering voorstellen door een machtreeks in t, waarbij het gebruik van (genormeerde) orthogonale veeltermen nuttig zal blijken te zijn.
We schrijven:

$$g_i(t) \approx g_i^{*}(t) = p_{oi} \cdot \varphi_o(t) + p_{1i} \cdot \varphi_1(t) + \ldots + p_{ki} \cdot \varphi_k(t) \qquad (3.1)$$

Hierin stelt $\varphi_j(t)$ een genormeerde orthogonale veelterm voor in t van de graad j, terwijl zoals hierboven aangegeven t de waarden van 1 t/m m kan aannemen.

De eigenschappen van de genormeerde orthogonale veeltermen $\varphi_j(t)$ zijn per definitie:

$$\sum_{t=1}^{m} \varphi_j(t).\varphi_\ell(t) = 0 \quad \text{voor } j \neq \ell \tag{3.2a}$$

en

$$\sum_{t=1}^{m} \varphi_j(t).\varphi_\ell(t) = 1 \quad \text{voor } j = \ell \tag{3.2b}$$

Het zal duidelijk zijn, dat hoe groter k is in (3.1), hoe nauwkeuriger $g_i(t)$ door $g_i^*(t)$ wordt benaderd. In par.9 wordt nader op de keuze van k ingegaan.

Bij een vooruit gekozen waarde van k wordt als beste benadering gekozen die, waarvoor geldt, dat de som van de kwadraten der m afwijkingen van $g_i^*(t)$ t.o.v. $g_i(t)$, dus:

$$S_i = \sum_{t=1}^{m} \left\{ g_i(t) - g_i^*(t) \right\}^2 = \sum_{t=1}^{m} \left[\left\{ g_i(t) \right\}^2 - 2g_i(t).g_i^*(t) + \left\{ g_i^*(t) \right\}^2 \right] \tag{3.3}$$

minimaal is. Voor (3.3) kunnen we schrijven (de index i weglatende):

$$\sum_{t=1}^{m} \left[\left\{ g(t) \right\}^2 - 2g(t) \left\{ p_o^* \varphi_o(t) + p_1 \varphi_1(t) + \ldots . p_k \varphi_k(t) \right\} + \left\{ p_o \varphi_o(t) + p_1 \varphi_1(t) + \ldots . p_k \varphi_k(t) \right\}^2 \right] \tag{3.3a}$$

Bij uitwerking van de laatste term van (3.3a) vallen wegens de orthogonale eigenschap (3.2a) alle dubbele produkten weg, terwijl wegens (3.2b) alle kwadraten van $\varphi_i(t)$ na sommering overgaan in de eenheid.

Alleen de p_j's blijven over en (3.3a) gaat over in:

$$S = \sum_{t=1}^{m} \left\{ g(t) \right\}^2 - 2 \sum_{t=1}^{m} g(t) \left\{ p_o \varphi_o(t) + p_1 \varphi_1(t) + \ldots . + p_k \varphi_k(t) \right\} + \sum_{j=1}^{k} p_j^2 \tag{3.3b}$$

Beschouwt men S als funktie van p_j, dan bereikt deze haar minimum, indien de partiële afgeleiden van S naar de p_j's nul zijn:

$$\frac{\partial S}{\partial p_j} = 2p_j - 2 \sum_{t=1}^{m} g(t).\varphi_j(t) = 0 \qquad j = 0,1,2, \ldots . k \tag{3.4}$$

Uit (3.4), opgesteld voor elk der n jaren, vindt men dus een uitdrukking voor p_{ij} en wel $\hat{p}_{ij}$:

$$\hat{p}_{ij} = \sum_{t=1}^{m} g_i(t).\varphi_j(t) \qquad (j = 0,1,2,3 \ldots . k \; , \; i = 1,2 \ldots n) \tag{3.5}$$

De p's zijn nu bekend want de m waarden van $g_i(t)$ zijn empirisch gegeven ($t = 1,2, \ldots . m$) en door substitutie in (3.1) is dus tevens $g_i^*(t)$ bekend.

Zonder de algemeenheid te kort te doen, kan men $\beta = f(t)$ benaderd denken door eenzelfde uitdrukking als $g(t)$:

$$f(t) \approx f^{*}(t) = q_o\varphi_o(t) + q_1\varphi_1(t) + \ldots\ldots q_l\varphi_l(t) \qquad (3.6)$$

Hier kan t eveneens de waarden 1 t/m m aannemen. Voor zekere t vindt men de waarde van β behorende bij de x , welke men vindt door in (3.1) deze t te substitueren.

Zoals reeds eerder werd opgemerkt, bestaat er, in tegenstelling tot $x = g(t)$, welke funktie voor elk jaar andere waarden kan aannemen, slechts één funktie $\beta = f(t)$, daar verondersteld wordt, dat deze niet aan een jaar-invloed onderhevig is. M.a.w. in het universum van jaren heeft men bij èlk jaar een funktie $x_i = g_i(t)$, doch slechts één funktie $\beta = f(t)$, waarvan de benadering $f^{*}(t)$ op basis der gegevens uit de ter beschikking staande n jaren zo goed mogelijk wordt geschat.

(1.5) kan nu bij benadering worden voorgesteld door:

$$y_i^{*} = c + \sum_{t=1}^{m} f^{*}(t).g_i^{*}(t) \qquad (3.7)$$

Bij uitwerking van de tweede term van het rechterlid en door toepassing van de orthogonale eigenschappen van $\varphi_j(t)$ gaat deze term over in:

$$p_{o_i}q_o + p_{1_i}q_1 + \ldots\ldots + p_{l_i}q_l \text{ indien } l < k \qquad (3.8a)$$

en

$$p_{o_i}q_o + p_{1_i}q_1 + \ldots\ldots + p_{k_i}q_k \text{ indien } l > k \qquad (3.8b)$$

waarbij l de graad is van $f^{*}(t)$ en k die van $g_i^{*}(t)$. Indien we nu gemakshalve $l = k$ veronderstellen gaat (3.7) over in:

$$y_i^{*} = c + p_{o_i}q_o + p_{1_i}q_1 + \ldots\ldots + p_{k_i}q_k \qquad (3.9)$$

Het blijkt dus, dat de q_j's de regressiecoëfficiënten zijn in het lineaire regressieverband tussen y en de p_j's ($j = 0,1, \ldots k$). M.a.w. c en q_o t/m q_k (dus k + 2 onbekenden) kunnen door middel van de methode der kleinste kwadraten worden bepaald uit de reeks van n jaren, dus n waarden van y met n bijbehorende stellen waarden p_o t/m p_k. Voor de beste schatting van c en de q_j's eisen we ook hier (zie 3.3) dat

$$T = \sum_{i=1}^{n} (y_i - y_i^{*})^2 = \sum_{i=1}^{n} \left\{ y_i^2 + c + p_{o_i}q_o + p_{1_i}q_1 + \ldots\ldots + p_{k_i}q_k)^2 - \right.$$

$$\left. 2\, y_i(c + p_{o_i}q_o + p_{1_i}q_1 + \ldots\ldots p_{k_i}q_k) \right\} \qquad (3.10)$$

minimaal is. Aan deze eis wordt voldaan door de partiële afgeleiden van T achtereenvolgens naar c, q_o, q_1, ... q_k gelijk aan nul te stellen. Dit

geeft een stelsel van k + 2 lineaire vergelijkingen (de z.g. normaalvergelijkingen) in c, q_o, q_1, ... q_k, waaruit deze grootheden zijn op te lossen; oplossingen : $\hat{c}$, $\hat{q}_o$,.... $\hat{q}_k$. Hiermede is de gezochte regressievergelijking geheel bekend.

4. <u>Oplossingsmethode voor het stelsel normaalvergelijkingen.</u>

Voor kleine waarden van k zal het oplossen van de normaalvergelijkingen geen moeilijkheden opleveren. Voor een grotere waarde van k wordt het zeer bewerkelijk om het stelsel door middel van eliminatie op te lossen. Een oplossing m.b.v. determinanten wordt eveneens zeer bewerkelijk. Indien men behalve de weerfaktor (bepaald door p_o t/m p_k) ook nog andere faktoren tegelijkertijd in de regressieberekening wil betrekken (b.v. zaaidatum, trend e.d.), of indien men twee weerfaktoren gelijktijdig wil beschouwen, die elk door een uitdrukking als (3.1) bepaald zijn wordt het oplossen van het stelsel zelfs een omvangrijk werk. Er zijn echter in de lineaire algebra methoden uitgewerkt, welke in dit geval overzichtelijk zijn en sneller tot het doel voeren. De normaalvergelijkingen vormen een bijzonder stelsel vergelijkingen, daar de coëfficiënten- matrix symmetrisch is, d.w.z. de j^e rij is gelijk aan de j^e kolom. In dat geval is voor de oplossing van het stelsel de methode Cholesky de meest doelmatige (zie Fox e.a. [3]).

Voor het aantonen van bovengenoemde eigenschap van de coëfficiënten van de normaalvergelijkingen en een uiteenzetting van de methode Cholesky, gaan we opnieuw uit van de vorm (3.9).

$$y^* = c + p_o q_o + p_1 q_1 + ... p_k q_k \qquad (3.9)$$

Geven we de jaren de index i , dus i = 1,2, ... n en noemen we verder $y_i^* - y_i = v_i$, dan hebben we n vergelijkingen van de vorm:

$$v_i = c + p_{o_i} q_o + p_{1_i} q_1 + ... + p_{k_i} q_k - y_i \qquad (4.1)$$

Gezocht de beste c en q_j's (j = 0,1, ... k).

Met behulp van de volgende matrices (matrices onderstreept):

$$\underline{A} = \begin{bmatrix} 1 & p_{o.1} & p_{1.1} & .... & p_{k.1} & - & y_1 \\ 1 & p_{o.2} & p_{1.2} & .... & p_{k.2} & - & y_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & p_{o.n} & p_{1.n} & .... & p_{k.n} & - & y_n \end{bmatrix} , \quad \underline{q} = \begin{bmatrix} c \\ q_o \\ . \\ . \\ . \\ q_k \\ 1 \end{bmatrix} , \quad \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ . \\ . \\ . \\ v_n \end{bmatrix} \qquad (4.2\ a,b,c)$$

kunnen we schrijven:

$$\underline{v} = \underline{A}.\underline{q} \qquad (4.3)$$

Als beste benadering van y nemen we weer die funktie y^* , waarvoor geldt dat

$$T \equiv \sum_{i=1}^{n} (y_i - y_i^{*})^2 = \sum_{i=1}^{n} v_i^2 \qquad (4.4)$$

minimaal is.

$\sum v_i^2$ verkrijgt men, indien men de kolommatrix $\{V\}$ (4.2c) voor-vermenigvuldigt met zijn getransponeerde $\{V\}^T$, dus in matrixnotatie:

$$\sum_{i=1}^{n} v_i^2 = \lfloor v_1, v_2, \ldots v_n \rfloor \cdot \{v_1, v_2, \ldots v_n\} = \underline{V}^T \cdot \underline{V} \qquad (4.5)$$

Noot *)

Door gebruikmaking van (4.3) kunnen we hiervoor schrijven:

$$\underline{V}^T \cdot \underline{V} = (\underline{Aq})^T (\underline{Aq}) = \underline{q}^T \cdot \underline{A}^T \cdot \underline{A} \cdot \underline{q} = \underline{q}^T \cdot \underline{B} \cdot \underline{q} \qquad (4.6)$$

waarin $\underline{B} = \underline{A}^T \cdot \underline{A}$. Deze matrix $\underline{B}$ is, aangezien ze het produkt is van de matrix $\underline{A}$ met zijn getransponeerde $\underline{A}^T$, een symmetrische, vierkante matrix van k + 3 bij k + 3. $\underline{B}$ is op de laatste rij en laatste kolom na gelijk aan de matrix van de coëfficiënten der k + 2 normaal-vergelijkingen, welke door differentiatie uit (3.9) worden gevonden. Men kan de methode Cholesky ook direkt toepassen op het stelsel normaalvergelijkingen. De hier gevolgde methode verdient echter de voorkeur, zoals later zal blijken. Daar de(k + 3, k + 3) matrix $\underline{B}$ symmetrisch is, is ze te splitsen (zie blz. 9) in twee (k + 3, k + 3) driehoekmatrices, welke elkanders getransponeerde zijn (een stelling). We schrijven dus:

$$\underline{B} = \underline{D}^T \cdot \underline{D} \qquad (4.7)$$

waarbij $\underline{D}$ de boven driehoekmatrix is, waarvan per definitie alle elementen beneden de hoofddiagonaal nul zijn.

$$\underline{V}^T \cdot \underline{V} = \underline{q}^T \cdot \underline{D}^T \cdot \underline{D} \cdot \underline{q} = \underline{\xi}^T \cdot \underline{\xi} \qquad (4.8)$$

Hierin is $\underline{\xi}$ een kolommatrix en $\underline{\xi}^T$ een rijmatrix, terwijl $\underline{\xi} = \underline{D}\,\underline{q}$, ofwel, indien d_{ij} het element van $\underline{D}$ is:

$$\begin{aligned}
\xi_{11} &= d_{11}c + d_{12}q_0 + d_{13}q_1 + \ldots + d_{1,k+2}q_k + d_{1,k+3} \\
\xi_{21} &= \quad d_{22}q_0 + d_{23}q_1 + \ldots + d_{2,k+2}q_k + d_{2,k+3} \\
&\vdots \\
\xi_{k+2,1} &= \ldots\ldots\ldots\ldots\ldots\ldots d_{k+2,k+2}\cdot q_k + d_{k+2,k+3} \\
\xi_{k+3,1} &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots d_{k+3,k+3}
\end{aligned} \qquad (4.9)$$

Noot *)

$\lfloor v_1, v_2, \ldots v_n \rfloor$ is de notatie voor een rij matrix.
$\{v_1, v_2, \ldots v_n\}$ is de notatie voor een kolommatrix.

Resumerend hebben we:

$$T = \sum_{i=1}^{n} v_i^2 = \underline{V}^T.\underline{V} = (\underline{A}.\underline{q})^T(\underline{Aq}) = \underline{q}^T.\underline{A}^T.\underline{Aq} = \underline{q}^T.\underline{B}.\underline{q} = \underline{q}^T.\underline{D}^T.\underline{Dq} = \underline{\xi}^T.\underline{\xi} = \xi_1^2 + \xi_2^2 + \ldots\ \xi_{k+3}^2$$

Deze T als som van een aantal kwadraten bereikt een minimum, indien elk van deze k+3 kwadraten minimaal, dus gelijk aan nul is, hetgeen voor ξ_1 t/m ξ_{k+2} bij een juiste keuze van c en q_o t/m q_k te bereiken is. Aan ξ_{k+3} echter kan men geen eisen stellen, daar de variabelen, c en q_j's , er in ontbreken. We eisen dus:

$$\xi_1 = \xi_2 = \xi_3 = \ldots = \xi_{k+2} = 0 \qquad (4.10)$$

Er geldt dan: Min.T = Min.$(\underline{V}^T.\underline{V}) = \xi^2_{k+3} = d^2_{k+3,k+3}$ (4.11)

Uit de $(k+2)^e$ vergelijking van (4.9) vinden we: $q_k = -\dfrac{d_{k+2,k+3}}{d_{k+2,k+2}}$

Door substitutie van deze waarde van q_k in de $(k+1)^e$ vergelijking vinden we q_{k-1} enz.

Hiermede is de oplossing bepaald, want de $d_{i\,j}$'s liggen reeds vast via $\underline{B}$; $\underline{B}$ weer via $\underline{A}$ en $\underline{A}$ weer via de $n_{i\,j}$'s uit elk der n jaren.

5. De orthogonale veeltermen.

Door Fisher en Yates [4] zijn ordinaatwaarden van orthogonale veeltermen getabelleerd, behorende bij de absciswaarden t = 1,2, ... m. Men vindt in deze tabellen veeltermen van de 1e t/m de 5e graad en wel voor m = 3 t/m m = 75. Ze voldoen aan de recurrente betrekking:

$$\xi_{j+1} = \xi_1 \xi_j - \frac{j^2(m^2 - j^2)}{4(4j^2 - 1)} \xi_{j-1} \qquad \text{waarin} \qquad (5.1)$$

ξ_j een orthogonale veelterm in t is van de graad j.

$$\xi_o = 1 \text{ en } \xi_1 = t - \bar{t} \quad (\bar{t} = \frac{1}{m} \sum_{t=1}^{m} t = \tfrac{1}{2}(m+1)) \qquad (5.2a,b)$$

Gemakshalve zijn niet de waarden der veeltermen ξ zelf getabelleerd, doch veelvouden ervan, n.l.:

$$\xi' = \lambda \xi \qquad (5.3)$$

waarin λ het kleinste getal is, waarvoor geldt, dat ξ' geheel is.
De door Fisher getabelleerde veeltermen zijn niet genormeerd, zodat de eigenschap (3.2a) wel geldt, doch in (3.2b) het „=1" vervangen moet worden door „ $\neq 0$" . Daarom wordt bij gebruik van deze veeltermen de overgang van (3.3a) op (3.3b) enigszins anders. En als gevolg hiervan wordt

(3.5) in dat geval:

$$p_j = \frac{\sum_{t=1}^{m} g(t) \circ \xi_j(t)}{\sum_{t=1}^{m} \{\xi_j(t)\}^2} = \frac{\sum_{t=1}^{m} g(t) \circ \xi_j'}{\sum_{t=1}^{m} (\xi_j')^2} \qquad (5.4)$$

Er moet dus na sommatie gedeeld worden door de som der kwadraten, welke sommen eveneens getabelleerd zijn.

Voor ξ geldt verder, dat voor even graad de waarden symmetrisch zijn t.o.v. $\frac{1}{2}$ m en voor oneven graad tegengesteld symmetrisch t.o.v. $\frac{1}{2}$ m:

$$\xi_j(c) = \xi_j(m+1-c) \qquad (j \text{ even}, c = 0,1, \ldots n) \qquad (5.5a)$$

$$\xi_j(c) = -\xi_j(m+1-c) \qquad (j \text{ oneven}, c = 0,1, \ldots n) \qquad (5.5b)$$

In verband hiermede zijn de waarden van ξ' voor m oneven getabelleerd vanaf $\frac{m+1}{2}$ t/m m en voor m even vanaf $\frac{m+2}{2}$ t/m m.

Van de eigenschappen (5.5a) en (5.5b) kan men met voordeel gebruik maken bij de berekeningen. Hiervoor zie men par.8 .

6. De matrices B en D.

Een element van $\underline{A}$ (4.2a) stellen we voor door a_{ij} en het overeenkomstige element van $\underline{B}$ (4.6) door b_{ij}. Hierin is de eerste index de rij-index en de tweede de kolom-index.

$\underline{B}$ ontstaat door voorvermenigvuldiging van $\underline{A}$ met $\underline{A}^T$. Het element b_{ij} is dus gelijk aan het inwendige produkt van de i^e rij van $\underline{A}^T$ en de j^e kolom van $\underline{A}$. En daar de i^e rij van $\underline{A}^T$ gelijk is aan de i^e kolom van $\underline{A}$, is b_{ij} dus gelijk aan het inwendige produkt van de i^e en de j^e kolom van $\underline{A}$, waaruit direkt volgt dat $b_{ij} = b_{ji}$

$$b_{ij} = b_{ji} = \sum_{l=1}^{n} a_{li} \circ a_{lj} \qquad (6.1)$$

De bepaling van $\underline{D}$ uit $\underline{B}$ lichten we toe aan de hand van een voorbeeld (k=o)

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \underline{D}^T \circ \underline{D} = \begin{bmatrix} d_{11} & o & o \\ d_{12} & d_{22} & o \\ d_{13} & d_{23} & d_{33} \end{bmatrix} \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ o & d_{22} & d_{23} \\ o & o & d_{33} \end{bmatrix} =$$

$$= \begin{bmatrix} d_{11}^2 & d_{11}d_{12} & d_{11}d_{13} \\ d_{11}d_{12} & d_{12}^2+d_{22}^2 & d_{12}d_{13}+d_{22}d_{23} \\ d_{11}d_{13} & d_{12}d_{13}+d_{22}d_{23} & d_{13}^2+d_{23}^2+d_{33}^2 \end{bmatrix} \qquad (6.2)$$

Hieruit vinden we de volgende betrekkingen:

$$
\begin{aligned}
d_{11}^2 &= b_{11} & d_{11} &= \sqrt{b_{11}} \\
d_{11}\,d_{12} &= b_{12} & d_{12} &= \frac{b_{12}}{d_{11}} \\
d_{11}\,d_{13} &= b_{13} & d_{13} &= \frac{b_{13}}{d_{11}} \\
d_{12}^2 + d_{22}^2 &= b_{22} & d_{22} &= \sqrt{b_{22} - d_{12}^2} \\
d_{12}d_{13} + d_{22}d_{23} &= b_{23} & d_{23} &= \frac{b_{23} - d_{12}d_{13}}{d_{22}} \\
d_{13}^2 + d_{23}^2 + d_{33}^2 &= b_{33} & d_{33} &= \sqrt{b_{33} - d_{13}^2 - d_{23}^2}
\end{aligned}
\qquad (6.3)
$$

En algemeen geldt indien d_{ij} het element is van $\underline{D}$:

$$i > j \qquad d_{ij} = 0 \qquad (6.4a)$$

$$i = j \qquad d_{ij} = \sqrt{b_{ii} - (d_{1i}^2 + d_{2i}^2 + \ldots + d_{i-1,i}^2)} \qquad (6.4b)$$

$$i < j \qquad d_{ij} = \frac{b_{ij} - (d_{1i}d_{1j} + d_{2i}.d_{2j} + \ldots + d_{i-1,i}d_{i-1,j})}{d_{ii}} \qquad (6.4c)$$

7. Signifikantie der regressie.

Algemeen geldt:

$$s^2_{yres} = (1-R^2)\, s_y^2 \qquad (7.1a)$$

of ook

$$s^2_{yres} = \frac{\sum_{i=1}^{n}(y_i - y_i^{\mathrm{x}})^2}{\nu} \qquad (7.1b)$$

Wegens (4.4) en (4.11) kan men voor (7.1b) schrijven:

$$s^2_{yres} = \frac{\sum_{i=1}^{n} v_i^2}{\nu} = \frac{d^2_{k+3,k+3}}{\nu} \qquad (7.1c)$$

Hier komt nu nog een voordeel van de methode Cholesky naar voren: het laatste element van $\underline{D}$, dus $d_{k+3,k+3}$, is de wortel uit de kwadraatsom van de restvariantie. Het levert dus, gekwadrateerd en gedeeld door het aantal vrijheidsgraden (ν) , direkt de restvariantie op. Het aantal graden van vrijheid is hier gelijk aan het aantal jaren (n) verminderd met het aantal van alle in de regressieberekening betrokken variabelen (afhankelijke en onafhankelijke); dit zijn hier y en p_o t/m p_k, dus k+2 stuks, d.w.z. $\nu = n-k-2$.
Uit (6.4b) volgt:

$$d^2_{k+3,k+3} = b_{k+3,k+3} - d^2_{1,k+3} - d^2_{2,k+3} \ldots\ldots - d^2_{k+2,k+3} \qquad (7.2)$$

Daar $b_{k+3,k+3} = \sum_1^n y_i^2$ en $d^2_{1,k+3} = \left(\frac{b_{1,k+3}}{\sqrt{b_{11}}}\right)^2 = \frac{(\sum_1^n y_i)^2}{n}$, kunnen we voor (7.2) schrijven:

$$d^2_{k+3,k+3} = \sum_1^n y_i^2 - \frac{(\sum_1^n y_i)^2}{n} - d^2_{2,k+3} \ldots\ldots - d^2_{k+2,k+3} \qquad (7.3)$$

ofwel: $$\sum_1^n (y_i-\bar{y})^2 = d^2_{2,k+3} + d^2_{3,k+3} + \ldots\ldots + d^2_{k+2,k+3} + d^2_{k+3,k+3} \qquad (7.4)$$

d.i. $$\sum_1^n (y_i-\bar{y})^2 = d^2_{2,k+3} + d^2_{3,k+3} + \ldots\ldots + d^2_{k+2,k+3} + \sum_1^n (y_i-y_i^*)^2 \qquad (7.5)$$

Hierdoor is de kwadraatsom van de variantie van y met n-1 vrijheidsgraden gesplitst in de kwadraatsom van de restvariantie van y met n-k-2 vrijheidsgraden en k+1 kwadraatsommen elk met 1 vrijheidsgraad.

Wij spreken van een signifikante invloed als in de populatie (d.i. $n \to \infty$) voor minstens één $d^2_{j,k+3}$-waarde geldt:

$$d^2_{j,k+3} > \sigma^2_{yres} \qquad (j = 2,3, \ldots\ldots k+2) \qquad (7.6)$$

Wij bepalen het quotiënt $w_j = \frac{d^2_{j,k+3}}{s^2_{yres}}$ (7.7)

Onder de voorwaarde, dat deze varianties onafhankelijk zijn en de n coëfficiënten p_j van $x = g(t)$ een steekproef uit een normale verdeling zijn ($j = 1,2, \ldots k$, dus k steekproeven), volgt dit quotiënt een F-verdeling met 1 en n-k-2 vrijheidsgraden, geschreven $F_{1,n-k-2}$ of $F_{1,\nu}$.

Zal nu aan de eis (7.6) met een waarschijnlijkheid van 95% zijn voldaan, dan moet gelden:

$$w_j = \frac{d^2_{j,k+3}}{s^2_{yres}} > F_{1,n-k-2}\ (5\%) \qquad (7.8)$$

Wordt niet door tenminste één der $d^2_{j,k+3}$ aan (7.8) voldaan, dan kan men met een waarschijnlijkheid van 95 % hieruit concluderen, dat er geen invloed van de beschouwde weerfaktor aanwezig is, althans, dat uit het gebruikte cijfermateriaal bij de gekozen signifikantie-drempel van 5% niets hieromtrent blijkt.

Is voor zekere index b.v. voor $j = \ell$ ($0 \leq \ell \leq k+2$) aan (7.8) voldaan, dan wijst dit op signifikante invloed van p_ℓ en daar p_ℓ een polynoom van de graad ℓ in de variabele t vertegenwoordigt, waarin dus t's van de 0^e t/m de ℓ^e graad voorkomen, zal het duidelijk zijn, dan men dan ook alle p_j met $j<\ell$ mee moet nemen in het regressieverband.

Voor toetsing gebruiken we de betrekkingen (4.9) en beginnen onderaan:

$$w_{k+2} = \frac{d^2_{k+2,k+3}}{s^2_{yres}}$$

Geldt: $w_{k+2} \geqslant F_{1,\nu}$ (5%) dan moet men dus, i.v.m. de zojuist gemaakte opmerking over p_j, alle q_o t/m q_k en c uitrekenen.

Is $w_{k+2} < F_{1,\nu}$ (5%) dan is er geen invloed van p_k op y aanwezig en toetsen wij op gelijke wijze:

$$w_{k+1} = \frac{d^2_{k+1,k+3}}{s^2_{yres}}$$

enz. tot er een w_j optreedt, welke groter is, dan de vaste signifikantiedrempel $F_{1,\nu}$ (5%). Is dit het geval voor $d^2_{\ell,k+3}$, dan schrapt men van het stelsel (4.9) de $(\ell+1)^e$ t/m $(k+2)^e$ rij en kolom weg en lost daarna c en q_o t/m $q_{\ell-1}$ op.

Terzijde zij hier nog opgemerkt, dat indien men het op prijs stelt de totale correlatie-coëfficiënt te kennen, deze op eenvoudige wijze uit (7.1a) en (7.1c) is af te leiden; men vindt:

$$R = \sqrt{1 - \frac{(n-k-2)\, s_y^2}{d^2_{k+3,k+3}}} \qquad (7.9)$$

De voor de berekening van s_y^2 benodigde gegevens vindt men in de matrix $\underline{B}$.

8. De berekeningen.

Hieronder wordt van enige onderdelen een voorbeeld van berekening gegeven.

a) Benadering door orthogonale veeltermen (ξ') van een funktie waarvan de waarde voor een aantal discrete waarden van de onafhankelijke variabele bekend is.

Stel, dat de waarde van $x = g(t)$ voor 18 aequidistante waarden van t gegeven is, b.v. x_1 t/m x_{18}, dan moeten deze volgens (5.4) vermenigvuldigd worden met de overeenkomstige waarden van ξ'. We vinden in de tabellen van Fisher de waarden van ξ' voor $\frac{m+2}{2} = 10$ t/m 18; deze waarden zijn hieronder weergegeven.

In verband met de eigenschappen (5.5a) en (5.5b) vormen we kolommen a en b, zoals hieronder eveneens is aangegeven, waarna men direkt kan vermenigvuldigen met $\xi'(10)$ t/m $\xi'(18)$.

a	b
$x_9 + x_{10}$	$x_{10} - x_9$
$x_8 + x_{11}$	$x_{11} - x_8$
$x_7 + x_{12}$	$x_{12} - x_7$
$x_1 + x_{18}$	$x_{18} - x_1$

ξ_1'	ξ_2'	ξ_3'	ξ_4'
1	- 40	- 8	44
3	- 37	- 23	33
5	- 31	- 35	13
7	- 22	- 42	- 12
9	- 10	- 42	- 36
11	5	- 33	- 51
13	23	- 13	- 47
15	44	20	- 12
17	68	68	- 68

$\Sigma(\xi_0')^2 = 18$
$\Sigma(\xi_1')^2 = 1938$
$\Sigma(\xi_2')^2 = 23256$
$\Sigma(\xi_3')^2 = 23256$
$\Sigma(\xi_4')^2 = 28424$

Numeriek voorbeeld:

Gegeven van Eelde de dekadengemiddelden van de overdag-temperatuur $(\frac{8+14+19}{3})$ over de maanden maart t/m augustus 1955. We willen het verloop hiervan benaderen door een veelterm van de 4e graad.

	I	II	III
mrt	- 1.3	0.4	0.5
apr	8.1	7.5	10.1
mei	12.3	8.5	10.8
jun	14.5	13.4	16.7
jul	16.2	21.2	18.5
aug	16.9	20.4	21.3

Uit deze gegevens berekenen we de kolommen a en b en de produkten van de waarden van a met de overeenkomstige waarden van ξ_2' en ξ_4' en van b met ξ_1' en ξ_3'.

De sommen der kolommen, gedeeld door $(\xi')^2$, geven de p's. Zie (5.4). We vinden dan onderstaande waarden:

	a	b	$b\xi_1'$	$a\xi_2'$	$b\xi_3'$	$a\xi_4'$
	25.3	3.7	3.7	-1012.0	- 29.6	1113.2
	21.9	4.9	14.7	- 810.3	- 112.7	722.7
	29.0	4.4	22.0	- 899.0	- 154.0	377.0
	26.3	6.1	42.7	- 578.6	- 256.2	- 315.6
	28.7	13.7	123.3	- 287.0	- 575.4	-1033.2
	26.6	10.4	114.4	133.0	- 343.2	-1356.6
	21.9	11.9	154.7	503.7	- 154.7	-1029.3
	20.8	20.0	300.0	915.2	400.0	- 249.6
	20.0	22.6	384.2	1360.0	1536.8	1360.0
som	220.5	-	1159.7	- 675.0	311.0	- 411.4
deler	18	-	1938	23256	23256	28424
quotiënt	12.25 = p_0	-	0.598 = p_1	- 0.029 = p_2	0.0134 = p_3	-0.0145 = p_4

We hebben dus gevonden: $g^*(t) = 12.25\,\xi_0' + 0.598\,\xi_1' - 0.029\,\xi_2' + 0.0134\,\xi_3' - 0.0145\,\xi_4'$ voor 1955.

Wil men dit omwerken tot een machtreeks in t, dan moet men voor ξ' de analytische uitdrukking substitueren. Deze vindt men uit (5.1), (5.2a,b) en (5.3).

b) Berekening van de matrices B en D.

Stel, dat de matrix A uit 5 kolommen (N, p_0, p_1, p_2, y) van 8 getallen (8 jaren) bestaat. Hieraan voegen we de kolom S toe, door de getallen horizontaal op te tellen. Voor de berekeningen is het niet nodig de kolom y met tegengesteld teken te noteren, zoals in (4.2a) is aangegeven.

Matrix A:

	N	p_0	p_1	p_2	y	S
	1	2	9	4	18	34
	1	7	16	3	17	44
	1	4	14	2	12	33
	1	8	7	1	10	27
	1	5	8	2	15	31
	1	2	10	3	15	31
	1	4	11	4	16	36
	1	4	17	3	17	42
totaal:	8	36	92	22	120	278

De matrix B wordt nu verkregen door de kolommen onderling te vermenigvuldigen. Het getal, dat in rij p_0 en kolom p_2 ligt, is als volgt verkregen:

$$2 \times 4 + 7 \times 3 + 4 \times 2 + 8 \times 1 + 5 \times 2 + 2 \times 3 + 4 \times 4 + 4 \times 3 = 89 .$$

Dit getal staat wegens de symmetrie ook in het "snijpunt" van de rij p_2 en de kolom p_0. Het getal, in rij p_1 en kolom p_1, is als volgt ontstaan:

$$9^2 + 16^2 + 14^2 + 7^2 + 8^2 + 10^2 + 11^2 + 17^2 = 1156 .$$

De kolom, gevormd door de produkten van S met resp. N, p_0, p_1, p_2 en y wordt niet ook nog eens als rij genoteerd, daar S uitsluitend ter contrôle dient. De som van de elementen in een rij van B moet gelijk zijn aan het element van S dat in die rij staat.

Matrix B:

	N	p_0	p_1	p_2	y	S
N	8	36	92	22	120	278
p_0	36	194	414	89	520	1253
p_1	92	414	1156	260	1407	3329
p_2	22	89	260	68	347	786
y	120	520	1407	347	1852	4246

Met de formules (6.4a,b,c) berekenen we de elementen van $\underline{D}$.

d_{11} wordt $\sqrt{b_{11}} = \sqrt{8} = 2.828$

$d_{12} = \frac{b_{12}}{d_{11}} = \frac{36}{2,828} = 12,728$; $d_{13} = \frac{b_{13}}{d_{11}} = \frac{92}{2,828} = 32,527$ enz.

De eerste rij van D wordt dan: 2,828 ; 12,728 ; 32,527 ; 7,778 ; 42,426 ; 98,287.

Contrôle: de som van de eerste 5 elementen is ook gelijk aan 98,287.

$d_{21} = o$, daar $i > j$

$d_{22} = \sqrt{b_{22} - d_{12}^2} = \sqrt{194 - 12,728^2} = 5,657$

$d_{23} = \frac{b_{23} - d_{12} \cdot d_{13}}{d_{22}} = \frac{414 - 12,728 \times 32,527}{5,657} = -0,001$ enz.

Door afrondingsfouten is het mogelijk, dat de contrôle niet exact is. De som van de derde rij b.v. geeft 13,334 i.p.v. 13,337. Deze afwijkingen zijn echter zo gering, dat ze te verwaarlozen zijn.

Het uiteindelijk resultaat wordt:

Matrix D:

	N	p_o	p_1	p_2	$\bar{y}$	S
N	2,828	12,728	32,527	7,778	42,426	98,287
p_o	o	5,657	- 0,001	-1,767	- 3,535	0,354
p_1	o	o	9,899	0,707	2,728	13,337
p_2	o	o	o	1,970	4,485	6,457
$\bar{y}$	o	o	o	o	3,463	3,461

Is de gezochte regressievergelijking: $y_i^* = c + q_o p_{oi} + q_1 p_{1i} + q_2 p_{2i}$, dan lost men uit de matrix $\underline{D}$ op:

$$\hat{q}_2 = \frac{d_{4.5}}{d_{4.4}} = \frac{4,485}{1,970} = 2,277$$

Daar men in de matrix $\underline{A}$ de kolom y niet in tegengesteld teken heeft genoteerd, heeft men $\hat{q}_2 = \frac{4,485}{1,970}$ en niet $\hat{q}_2 = -\frac{4,485}{1,970}$.

$$\hat{q}_1 = \frac{2,728 - 0,707 \cdot a_3}{9,899} = \frac{2,728 - 1,610}{9,899} = 0,113$$

$$\hat{q}_o = \frac{-3,535 + 1,767 \cdot a_3 + 0,001 \cdot a_2}{5,657} = \frac{0,4886}{5,657} = 0,086$$

$$\hat{c} = \frac{42,426 - 7,778 \cdot a_3 - 32,527 \cdot a_2 - 12,728 \cdot a_1}{2,828} = \frac{19,9453}{2,828} = 7,053$$

De restvariantie wordt gevonden uit $\frac{d_{5,5}^2}{n} = \frac{3,463^2}{8}$. De wortel hieruit, dus de restspreiding, wordt $\frac{3,463}{2,828} = 1,225$. Met een waarschijnlijkheid van 95 % ligt het verschil tussen de werkelijke opbrengst en de berekende binnen $\pm$ 2.$S_y$rest. Wil men de opbrengst voor het jaar i met behulp van de regressievergelijking voorspellen dan geldt dus:

$$y_i = 7{,}053 + 0{,}086\ p_{oi} + 0{,}113\ p_{1i} + 2{,}277\ p_{2i} \pm 2{,}450$$

9. Enige algemene opmerkingen over de beschreven methode.

Uiteraard is de beschreven methode het best bruikbaar voor geleidelijk verlopende weerfaktoren. Dit als gevolg van het feit, dat een polynoomaanpassing van lage graad slecht voldoet voor een abrupt verlopende grootheid. Een voorbeeld van een abrupt verlopende weerfaktor is b.v. neerslagsommen per dekade.

Een groot voordeel van deze methode is, dat ze ook gebruikt kan worden voor een grootheid met een discrete verdeling, b.v. neerslagdagen.

De gehele bewerking die nodig is lijkt erg omslachtig. Bij een reeks van b.v. 30 jaar heeft men 30 maal een polynoomaanpassing te berekenen, alvorens tot het berekenen van de regressie kan worden overgegaan. Speciaal als daarna het verondersteld verband niet aanwezig blijkt te zijn, is er veel rekenwerk verricht eer men tot die conclusie komt.

Door gebruikmaking van (3.8a) en (3.8b) kan men echter een orienterend onderzoek opzetten, waardoor men voorkomt, dat er een uitgebreide hoeveelheid rekenwerk moet worden verricht, met de kans dat uiteindelijk blijkt, dat van de onderzochte weerfaktor geen enkele invloed op de opbrengst is aan te tonen.

Uit genoemde eigenschappen (3.8a) en (3.8b) blijkt dat de graad van $x = g(t)$ (de weerfaktor) en $\beta = f(t)$ (de regressiefunktie) aan elkaar gebonden zijn. Gebruikt men voor $g(t)$ een 4e graads-benadering, dan vindt men ook voor $f(t)$ een benadering van de 4e graad. Is omgekeerd $f(t)$ een funktie van de 2e graad, dan behoeft men voor $g(t)$ ook slechts een benadering van de 2e graad te gebruiken. Dit is zeer belangrijk. Indien de aanwezigheid van een relatie tussen x en y wordt verondersteld, waarom zou men dan β in eerste instantie niet door een lineaire funktie benaderen? Een benadering van de 0e graad (= gemiddelde) is niet aan te bevelen, daar het niet ondenkbaar is dat indien $\beta \neq 0$, toch de gemiddelde waarde van β gelijk aan nul is. De kans dat in zo'n geval de 1e graadsterm eveneens gelijk aan nul is, is wel bijzonder klein.

Het komt er dus op neer, dat in eerste instantie alleen p_o en p_1 voor x worden berekend en daarna de regressie van y op p_o en p_1, hetgeen dus de coëfficiënten q_o en q_1 oplevert, welke de 1e graads-benadering van β bepalen.

Wel is bij zo'n oriënterend onderzoek aan te bevelen de signifikantiedrempel voor q_o en q_1 te verschuiven van de gebruikelijke 5 % naar 10 %. Het is n.l. heel goed mogelijk, dat men bij zo'n eerste opzet slechts een grove benadering krijgt van het regressieverband (zo dit werkelijk bestaat)

en met de signifikantie-toetsing dus een grotere tolerantie in acht moet nemen.
Zijn de gevonden coëfficiënten q_o en q_1 niet signifikant, dan kan men direkt stoppen. Zijn q_o en q_1 wel signifikant, dan is er dus de mogelijkheid aanwezig, dat er nog q's van hogere orde zijn welke ook signifikant zijn. De volgende stap is in dat geval: p_2 berekenen en opnieuw nu q_o, q_1 en q_2 bepalen, waarna q_2 getoetst kan worden, enz. Bij elke stap kan men genoegen nemen met een signifikantie-drempel van 10 % om dezelfde reden als dit voor q_o en q_1 gebeurde. Aan het einde dient men er echter rekening mee te houden, dat minstens één q_j een overschrijdingskans moet hebben, welke kleiner is dan 5 % en dat deze j tevens de graad van β bepaalt. M.a.w. is j = 3 dan berekent men q_o t/m q_3 .

Verder nog een opmerking over de in par. 5 besproken orthogonale veeltermen. Zoals reeds opgemerkt, zijn de door Fisher getabelleerde orthogonale veeltermen niet genormeerd, d.w.z. ze voldoen niet aan (3.2b) (wel $\neq 0$, maar niet = 1). Als gevolg hiervan gaat bij gebruik ervan (3.8) over in:

$$y^* = c + q_o p_o \Sigma \xi_o^2 + q_1 p_1 \Sigma \xi_1^2 + q_2 p_2 \Sigma \xi_2^2 + \ldots . \qquad (9.1)$$

Uit de regressieberekening met p_i vindt men dus geen q_i doch $q_i \Sigma \xi_i^2$.
Men kan ook in de regressieberekening $p_i \Sigma \xi_i^2$ gebruiken en q_i vinden. De keuze is vrij en het doet weinig terzake, mits men bij eventueel uitschrijven van $\beta \approx f^*(t)$ er maar rekening mee houdt, of men met q_i dan wel met $q_i \Sigma \xi_i^2$ te doen heeft.

10. Kwadratische regressie.

Met de in het voorgaande besproken methode der continue regressie vindt men een lineaire uitdrukking voor het verband tussen de variabelen y en de p_j's. In het algemeen zal het verband tussen weerfaktoren en opbrengst echter niet lineair zijn en men kan zich dus afvragen, of deze methode van Fisher ook te gebruiken is, om tot een niet-linaire uitdrukking van de regressie te komen, dus b.v. een kwadratische. In dat geval is het analogon voor de uitdrukking (1.3):

$$\beta_1 x_i + \beta_2 x_i^2 \qquad (10.1)$$

waarin: $\beta_1 = f_1(t)$ en $\beta_2 = f_2(t)$ (10.2)

De uitdrukking (1.5) wordt in dat geval:

$$y_i = \gamma + \int_a^b \{ f_1(t) + f_2(t) . g_i(t) \} g_i(t)\, dt \qquad (10.3)$$

Dit is de meest algemene vorm voor het kwadratisch verband tussen de weer-

faktor x en de opbrengst y. De regressie-coëfficiënt $\beta = f(t)$ uit (1.5) is hier:

$$\beta_i = f_1(t) + f_2(t).g_i(t) \qquad (10.4)$$

Deze regressiecoëfficiënt is een funktie van de tijd èn van $x = g_i(t)$. Bij uitwerking van (10.3) kan slechts ten dele worden geprofiteerd van de eigenschappen der orthogonale veeltermen, waaruit $f_1(t)$, $f_2(t)$ en $g_i(t)$ zijn opgebouwd. Deze moeilijkheid kan worden verholpen door te veronderstellen, dat de invloed van x op de grootte van de regressiecoëfficiënt onafhankelijk is van de tijd. Deze restrictie houdt in:

$$\beta_2 = f_2(t) = \text{constant} = d \qquad (10.5)$$

Dan gaat (10.3) over in :

$$y_i = \gamma + \int_a^b \left\{ f_1(t) + d\, g_i(t) \right\} g_i(t)\, dt \qquad (10.6)$$

Uitgewerkt geeft dit:

$$y^* = c + p_o q_o + p_1 q_1 + ... + p_k q_k + d(p_o^2 + p_i^2 + ... p_k^2) \qquad (10.7)$$

Vergelijk (3.9).

Onder de voorwaarde weergegeven door (10.5) voert het kwadratisch verband tussen y en x aldus tot een lineair verband tussen de variabelen p_o t/m p_k en $\sum_o^k p_i^2$ met k + 3 regressiecoëfficiënten, te weten c, q_o t/m q_k en d.

In de benadering van $x = g(t)$ door $g^*(t) = p_o \varphi_o(t) + p_1 \varphi_1(t) + ... + p_k \varphi_k(t)$ kan men vrij algemeen stellen, dat de rij $\bar{p}_i$ snel naar nul convergeert, zodat geldt: $\bar{p}_o \gg \bar{p}_1 \gg \bar{p}_2 \gg ... \gg \bar{p}_k$ ($\bar{p}_i$ is het gemiddelde van p_i over n jaren, $n \to \infty$). En dit geldt dan in nog sterker mate voor p_i^2. Men kan hiervan gebruik maken bij het berekenen van de kwadratische regressie. Daar $\sum_o^k p_i^2$ vrijwel geheel bepaald wordt door p_o^2, als gevolg van het feit, dat p_1 t/m p_k in het algemeen maar zeer weinig gewicht in de schaal leggen, is het juist in vele gevallen voldoende om niet met $\sum_o^k p_i^2$, maar met p_o^2 alleen als extra variabele te werken.

Indien er sprake is van een kwadratisch verband, dan zal dit bij extreme waarden van de onafhankelijk variabele het duidelijkst tot uiting komen, terwijl het "midden gebied" op redelijke wijze is voor te stellen door een lineaire uitdrukking voor de regressie.
Daar in het midden gebied van de onafhankelijke variabele doorgaans veel meer waarden liggen dan er buiten, zal het veelal niet meevallen de kwadratische regressie betrouwbaar vast te stellen.
Het is dan ook zeker aan te bevelen éérst de lineaire regressie te berekenen en de q's te toetsen. Blijken dan b.v. q_o t/m q_j signifikant te

zijn, dan kan men daarna voor kwadratische regressie de variabele $\sum_{0}^{j} p_i^2$ of alleen p_0^2 nog toevoegen en de daaruit gevonden $(j+2)^e$ regressiecoëfficiënt toetsen.

11. Twee of meer weerfaktoren tegelijk.

Voor het geval men meerdere weerfaktoren gelijktijdig in de berekening wil betrekken, volgt hieronder een voorbeeld voor 2 weerfaktoren, naar analogie waarvan het theoretisch ook mogelijk is een willekeurig aantal faktoren tegelijk te beschouwen.

Gesteld de weerfaktoren x_1 en x_2, opnieuw afhankelijk van t, zijn in jaar i:

$$x_{1i} = f_i(t) \approx a_{oi}\varphi_o(t) + a_{1i}\varphi_1(t) + \ldots\ldots a_{ki}\varphi_k(t) \qquad (11.1a)$$

en

$$x_{2i} = g_i(t) \approx b_{oi}\varphi_o(t) + b_{1i}\varphi_1(t) + \ldots\ldots b_{ji}\varphi_j(t) \qquad (11.1b)$$

Hierbij behoren resp. de regressiecoëfficiënten β_1 en β_2, welke worden voorgesteld door:

$$\beta_1 = \psi_1(t) \approx p_o\varphi_o(t) + p_1\varphi_1(t) + \ldots\ldots + p_k\varphi_k(t) \qquad (11.2a)$$

en

$$\beta_2 = \psi_2(t) \approx q_o\varphi_o(t) + q_1\varphi_1(t) + \ldots\ldots + q_j\varphi_j(t) \qquad (11.2b)$$

Dan is de opbrengst in het jaar i bij benadering voor te stellen door:

$$y_i^* = c + \sum_{t=1}^{m}\psi_1(t).f_i(t) + \sum_{t=1}^{m}\psi_2(t).g_i(t) \qquad (11.3)$$

Dit gaat na uitwerking en toepassing van de orthogonale eigenschappen over in (zie de overgang van (3.7) naar (3.9)):

$$y_i^* = c + a_{oi}p_o + a_{1i}p_1 + \ldots\ldots a_{ki}p_k + b_{oi}q_o + b_{1i}q_1 + \ldots\ldots + b_{ji}q_j \qquad (11.4)$$

met regressiecoëfficiënten c, p_i's en q_i's (k + j + 3 stuks).

Hierbij is de graad k van x_1 niet gebonden aan de graad j van x_2.

De verdere berekening is gelijk aan die welke voor één weerfaktor moet worden uitgevoerd. Alleen de toetsing is iets gecompliceerder.

Zie hierover par. 12.

Voor een efficiënte werkwijze is het echter nooit aan te bevelen meerdere weerfaktoren tegelijk te beschouwen, alvorens men van elke weerfaktor afzonderlijk de invloed heeft nagegaan.

De procedure welke het best gevolgd kan worden is:

a) Van de, op basis van buiten het statistisch onderzoek gelegen gronden, in aanmerking komende weerfaktoren wordt van elk afzonderlijk de invloed op de opbrengst nagegaan; deze invloed wordt op signifikantie getoetst.

b) Daarna worden hieruit die weerfaktoren genomen waarvan de invloed signifikant is. Elk met de erbij behorende graad van de polynoom-ontwikkeling,

welke voor betreffende weerfaktor nodig is. Deze benodigde graad werd dus voor elke weerfaktor bij de toetsing van de regressie bepaald.

c) De parameters van de onder b) genoemde weerfaktor (dus de coëfficiënten van de polynoomaanpassingen) worden allen gezamenlijk voor de uiteindelijke regressieberekening gebruikt.

Volgt men deze procedure, dan is men altijd verantwoord. De tegenwerping kan worden gemaakt, dat een niet signifikante invloed van een weerfaktor wel signifikant kan blijken te zijn bij gelijktijdige beschouwing van meerdere weerfaktoren. Men bedenke echter, dat signifikantie nog niet altijd wil zeggen dat de invloed praktische waarde heeft. In het algemeen zal het bij dergelijke onderzoeken toch de bedoeling zijn om te trachten de opbrengst binnen vrij nauwe grenzen te bepalen en zo mogelijk een verwachtingsformule voor komende jaren op te stellen. Wil dit mogelijk zijn, dan moet de signifikantie zeer sterk zijn. Zodoende blijft men aan de veilige-kant indien per weerfaktor een signifikantie-drempel van 5 % wordt gebruikt. Nadat de onder c) genoemde bewerking zover is uitgevoerd tot de matrix D is verkregen, kan men de dan nog zwak signifikante faktoren wegschrappen (afhankelijk van de gestelde eisen), alvorens de uiteindelijke regressie-vergelijking wordt opgesteld.

Dat voor een betrouwbare regressie-vergelijking sterke signifikantie (d.i. grote R) is vereist, blijkt uit de betrekking:

$$S_y\text{rest} = \sqrt{1 - R^2}\,.\,S_y \qquad (11.5)$$

Immers, eist men, dat door de regressie-vergelijking de spreiding teruggebracht wordt tot b.v. de helft van de oorspronkelijke, dan wordt deze eis voorgesteld door:

$$S_y\text{rest} = 0{,}5\ S_y$$

en in verband met (11.5) geldt dan:

$$\sqrt{1 - R^2} = 0{,}5 \longrightarrow R \approx 0{,}87$$

Om aan deze gestelde eis te kunnen voldoen is dus een multipele correlatiecoëfficiënt vereist van 0,87 , welke hoge waarde alleen bij zeer sterke signifikantie van de lineaire regressie kan optreden.

12. Enige opmerkingen over de toetsing van de regressie.

De toetsing, zoals deze in par. 7 is behandeld, is afkomstig van het Centrum voor Landbouwwiskunde. Het lijkt mij meer verantwoord deze toetsing enigszins anders uit te voeren.

In par. 7 is gezegd, dat, als b.v. de variantie veroorzaakt door p_3 signifikant is, dan ook p_0, p_1 en p_2 in de berekening moeten worden meegenomen. Dit op grond van het feit, dat de beste polynoom-aanpassing van de derde

graad voor de weerfaktor gebaseerd is op p_o t/m p_3. Is in zo'n geval echter de invloed van p_3 signifikant, d.w.z. draagt p_3 een werkelijk aandeel bij voor de verklaarde variantie, dan is het toch nog mogelijk dat p_o, p_1 en p_2 slechts zeer weinig bijdragen voor deze "verklaarde" variantie. Zodoende kan het voorkomen, dat dan de invloed van p_o t/m p_3 gezamenlijk niet signifikant is.

In verband hiermede kan de toetsing beter als volgt worden uitgevoerd:
Bij het oriënterend onderzoek (zoals voorgesteld in par. 9), toetst men de gezamenlijke variantie van p_o en p_1 tegen de restvariantie, dus:

$$\frac{d^2_{2,k+3} + d^2_{3,k+3}}{2} \text{ tegen } \frac{d^2_{k+3,k+3} - (d^2_{1,k+3} + d^2_{2,k+3} + d^2_{3,k+3})}{n-3}$$

Het quotiënt heeft dan een F-verdeling met 2 en n-3 vrijheidsgraden.
Geeft de uitkomst hiervan aanleiding tot verder onderzoek, dan toetst men vervolgens:

$$\frac{d^2_{2,k+3} + d^2_{3,k+3} + d^2_{4,k+3}}{3} \text{ tegen } \frac{d^2_{k+3,k+3} - (d^2_{1,k+3} + d^2_{2,k+3} + d^2_{3,k+3} + d^2_{4,k+3})}{n-4}$$

met eventueel nog verdere stappen. Hierbij wordt dus steeds de signifikantie van de gehele funktie $x = g(t)$ getoetst en wordt boven omschreven moeilijkheid vermeden.

Er kan bewezen worden, dat $d^2_{2,k+3}$ de invloed van p_o op de oorspronkelijke variantie weergeeft, $d^3_{3,k+3}$ de invloed van p_1, nadat de invloed van p_o reeds geëlimineerd is; $d^2_{4,k+3}$ de invloed van p_2 nadat de invloed van $p_o + p_1$ geëlimineerd is, enz.
Zo is dus, bij beschouwing van meerdere weerfaktoren tegelijk, de invloed die men van de weerfaktor vindt, welke achteraan geplaatst is, bij de berekeningen, dié invloed, welke overblijft nadat de invloed van de eraan voorafgaande weerfaktoren reeds geëlimineerd is.
Het beste kunnen daarom in beschouwingen met meerdere weerfaktoren, in de eindberekening die faktoren, waarvan de invloed het zwakst is, achteraan geplaatst worden. Dan blijkt of deze, in afhankelijkheid van de voorgaande weerfaktoren, werkelijk nog signifikante invloed hebben, of dat ze verworpen kunnen worden.
Worden, zonder voorafgaand onderzoek per weerfaktor, direkt meerdere weerfaktoren in de berekening betrokken, dan wordt dus steeds de signifikantie van een weerfaktor getoetst in afhankelijkheid van de weerfaktoren welke ervoor geplaatst zijn. Alleen van de weerfaktor, welke als eerste in de

rij staat, kan men de invloed onafhankelijk van de andere weerfaktoren toetsen. Dit is geen groot bezwaar, echter zij men bedacht op het volgende: Is bij 3 weerfaktoren, in volgorde a, b en c genoemd, c signifikant, b niet, a echter weer wel, dan mag men in de matrix D niet zonder meer de rijen en kolommen welke betrekking hebben op b wegstrepen en daarna de regressiecoëfficiënten van de overige weerfaktoren uit het stelsel D oplossen (zie par. 7). Immers, de rijen en kolommen welke betrekking hebben op c zijn ook afhankelijk van de rijen en kolommen welke op b betrekking hebben. Het enige juiste is dan, om dit wegstrepen uit te voeren in de matrix B , waarna het laatste gedeelte van D opnieuw wordt berekend. Bij D mogen, indien nodig, alleen de laatste rijen en kolommen (uitgezonderd de kolom waarin de toetsingsgrootheden staan) worden weggestreept, aangezien geen der voorgaande elementen hiervan afhankelijk is.

Literatuur:

1 R.A. Fisher, "The Influence of Rainfall on the Yield of Wheat at Rothamsted". Phil.Trans.of the Royal Society of London. Series B, 213 (1924).

2 F.H. Sanderson, "Methods of Crop Forecasting".

3 L. Fox, H.D.Huskey and J.H. Wilkinson, "Notes of the Solution of algebraic simultaneous linear Equations". The Quart.J.of Mech.and Appl. Math., Vol.1, 1948.

4 R.A. Fisher and F. Yates "Stat.Tables for Biol., Agric. and Med. Research".

mei, 1960.