K O N I N K L I J K N E D E R L A N D S
M E T E O R O L O G I S C H I N S T I T U U T

Wetenschappelijk rapport
W.R. 69-4

C.Levert
en
M.Scharringa

Enkele schijncorrelaties in landbouwmeteorologisch onderzoek

De Bilt, 1969

Publikatienummer: K.N.M.I. W.R. 69-4 (R III - 308 1969)

U.D.C. : 551.586:63 :
519.2 :
551.577.3

ENKELE SCHIJNCORRELATIES IN LANDBOUWMETEOROLOGISCH ONDERZOEK.

door

Dr.C.Levert

en

M.Scharringa

> The best advice that we can give to the man who finds a correlation and starts to say, "It's obvious" is: "Think again. Ten to one there is a catch in it".
>
> Moroney: Facts from Figures.

1.1 Weer en opbrengst van gewassen

Statistisch onderzoek naar het verband tussen het weer en de opbrengst van gewassen heeft in veel gevallen onbevredigende resultaten opgeleverd. Onbevredigend in die zin dat geen of slechts een zeer zwak verband werd gevonden terwijl de ervaring toch leert dat de opbrengst sterk afhankelijk is van het weersverloop tijdens de groeiperiode.

Als verklaring wordt wel aangevoerd dat de invloed van het totaal van alle weerselementen dermate complex is, dat men de invloed van de afzonderlijke elementen niet meer kan aantonen.

Bij correlatieonderzoek dient men echter op enkele gevaren, die zich ook bij het onderzoek naar de relatie weer-opbrengst voordoen, bedacht te zijn zoals uit het volgende zal blijken.

1.2 Verband tussen neerslag en opbrengst op een bepaald perceel.

Als wij ons bepalen tot de invloed van de, over het groeiseizoen gesommeerde, neerslaghoeveelheid op de opbrengst en wij onderstellen daarbij dat wij met een grond te doen hebben die van nature onvoldoende water voor de groei kan leveren, dan zal het gewas weinig opbrengen als er weinig of geen neerslag valt. Wij hebben dit gezien in de jaren 1911, 1921 en 1959. Op dergelijke gronden zijn neerslag en opbrengst positief gecorreleerd. Deze positieve correlatie zal blijven bestaan tot een zeker neerslagniveau is bereikt. Valt er meer neerslag dan zal de extra hoeveelheid niet meer tot een verhoging van de opbrengst bijdragen; anders gezegd, de correlatie zal nul worden. Bij nog meer neerslag zal de opbrengst zelfs nadelig worden beinvloed zodat er sprake is van een negatieve correlatie.

Als wij het verband tussen neerslagsom en opbrengst op een perceel, zoals boven omschreven, in een grafiek voorstellen, dan zal er een kromme ontstaan zoals in figuur 1.

Als wij deze optimumkromme in korte, nagenoeg rechte, stukken verdelen dan mogen wij bij benadering in elk stuk een lineair verband tussen neerslag en opbrengst onderstellen en dus behoort er bij elk stuk een correlatiecoëfficient.

Aanvankelijk (O - A) heeft de correlatiecoëfficient een geringe waarde hetgeen begrijpelijk is, want enkele tientallen millimeters neerslag over een groeiseizoen zullen de opbrengst nauwelijks kunnen beinvloeden. In het traject A - B is er een duidelijke positieve correlatie, die in het traject B - C aanzienlijk hoger wordt.

In het traject C - D neemt de correlatie af om in de buurt van het optimum, dus in het traject D - E nagenoeg nul te worden. In het traject E - F is de correlatie negatief en in het laatste traject zelfs sterk negatief.

De figuur is slechts een model en dient om de gedachtengang aanschouwelijk voor te stellen. Het is een sterk vereenvoudigd model, ook al omdat er geen rekening is gehouden met het feit dat de neerslagsom op verschillende wijzen over het seizoen verdeeld kan zijn.

Als wij een verband trachten te vinden tussen neerslagsom en opbrengst dan kiezen wij een zo lang mogelijke reeks van neerslag- en opbrengstgegevens, want hoe groter de steekproef, hoe betrouwbaarder de uitkomst. In een reeks van jaren zullen neerslagsommen voorkomen die in verschillende trajecten uit figuur 1 thuisbehoren, zodat een dergelijke reeks geacht kan worden te zijn ontstaan door samenvoeging van een aantal reeksen van ongelijke lengte, gestoken uit populaties die, in elk geval voor wat de correlatie tussen de twee grootheden betreft, verschillend zijn.

1.3 <u>Verschillen tussen de relaties neerslag- opbrengst op verschillende percelen</u>.

Als wij de opbrengstgegevens niet van één bepaald perceel doch van een gebied gebruiken, dan dienen wij te bedenken dat de kromme uit figuur 1 voor elk perceel een ander verloop zal (kunnen) hebben. In een gebied kunnen gronden voorkomen die varieren van zeer droogtegevoelig tot die welke van nature reeds voldoende water leveren. In het laatste geval is de opbrengst op geheel andere wijze afhankelijk van de neerslagsom en zelfs zal een verband kunnen voorkomen zoals in figuur 2 is afgebeeld.

Deze inleiding die met meer voorbeelden zou kunnen worden uitgebreid dient slechts om de praktische betekenis van hetgeen volgt, aannemelijk te maken.

2.1 Schijncorrelatie veroorzaakt door menging van steekproeven uit niet identieke populaties.

Stel wij nemen aselect een steekproef van m paren x_i, y_i uit een al of niet normaal verdeelde $\underline{x}$, $\underline{y}$ bipopulatie waarin de correlatiecoëfficient ρ is en de gemiddelden en varianties zijn: μ_x, μ_y, σ_x^2 en σ_y^2.

Een andere aselecte steekproef van n paren u_j, v_j wordt genomen uit een $\underline{u}$, $\underline{v}$ bipopulatie met een correlatiecoëfficient τ en gemiddelden en varianties: μ_u, μ_v, σ_u^2 en σ_v^2.

De m paren x_i, y_i en de n paren u_j, v_j worden tot m+n paren samengevoegd, dat wil zeggen tussen x en u wordt geen onderscheid meer gemaakt, evenmin als tussen y en v en wij berekenen uit de m+n paren de zogenaamde steekproefcorrelatiecoëfficient: r.

Zouden wij dit bij dezelfde waarden van m en n talloos vele malen doen dan zou $\underline{r}$ stochastisch verdeeld zijn rondom een gemiddelde $\mathcal{E}_{\underline{r}}$ en met een variantie $\sigma_{\underline{r}}^2$. Nu is de vraag hoe deze $\mathcal{E}_{\underline{r}}$ en $\sigma_{\underline{r}}^2$ samenhangen met m, n, $\mu_x, \mu_y, \mu_u, \mu_v, \sigma_x^2, \sigma_y^2, \sigma_u^2, \sigma_v^2, \rho$ en τ.

Voor r luidt de definitie:

$$\frac{(\Sigma xy + \Sigma uv) - \frac{(\Sigma x + \Sigma u)(\Sigma y + \Sigma v)}{m+n}}{\sqrt{\left[\Sigma x^2 + \Sigma u^2 - \frac{(\Sigma x + \Sigma u)^2}{m+n}\right]\left[\Sigma y^2 + \Sigma v^2 - \frac{(\Sigma y + \Sigma v)^2}{m+n}\right]}} = \frac{\underline{A}}{\underline{B}} = \frac{\underline{A}}{\sqrt{\underline{B}_1 \underline{B}_2}} \qquad (1)$$

Ofschoon $\mathcal{E}\left(\frac{\underline{A}}{\underline{B}}\right) \neq \frac{\mathcal{E}\underline{A}}{\mathcal{E}\underline{B}}$ achten wij hier de benadering $\mathcal{E}\left(\frac{\underline{A}}{\underline{B}}\right) \simeq \frac{\mathcal{E}\underline{A}}{\mathcal{E}\underline{B}}$

geoorloofd.

Eerst wordt de verwachtingswaarde van de teller $(\mathcal{E}\underline{A})$ berekend ')

Daar $\mathcal{E}\, x_i, y_i = \rho\sigma_x\sigma_y + \mu_x\mu_y$; i = 1, 2, 3, ...m

en $\mathcal{E}\, u_j, v_j = \tau\sigma_u\sigma_v + \mu_u\mu_v$; j = 1, 2, 3, ...n

en $\mathcal{E}(\Sigma x_i y_i + \Sigma u_j v_j) = m(\rho\sigma_x\sigma_y + \mu_x\mu_y) + n(\tau\sigma_u\sigma_v + \mu_u\mu_v)$

en $\mathcal{E}\left[(\Sigma x + \Sigma u)(\Sigma y + \Sigma v)\right] = (m\mu_x + n\mu_u)(m\mu_y + n\mu_v) + mn\rho\sigma_x\sigma_y + n\tau\sigma_u\sigma_v$

wordt

$$\mathcal{E}\underline{A} = \frac{(mn\rho\sigma_x\sigma_y + n\tau\sigma_u\sigma_v)(m+n-1) + mn(m+n)(\mu_x - \mu_u)(\mu_y - \mu_v)}{m+n} \qquad (2)$$

De verwachtingswaarde van de noemer $\underline{B}$ berekenen wij via de verwachtingswaarden $\mathcal{E}\underline{B}_1$ en $\mathcal{E}\underline{B}_2$ van $\underline{B}_1$ en $\underline{B}_2$

Daar $\underline{B}_1 = \Sigma x^2 + \Sigma u^2 - \frac{(m\bar{x} + n\bar{u})^2}{m+n}$ of: $\frac{m(\Sigma x^2 - m\bar{x}^2) + n(\Sigma u^2 - n\bar{u}^2) + n\Sigma x^2 + m\Sigma u^2 - 2mn\bar{x}\bar{u}}{m+n}$

wordt $\mathcal{E}\underline{B}_1 = \frac{(m+n-1)(m\sigma_x^2 + n\sigma_u^2) + mn(\mu_x - \mu_u)^2}{m+n}$ (3)

Evenzo is $\mathcal{E}\underline{B}_2 = \frac{(m+n-1)(m\sigma_y^2 + n\sigma_v^2) + mn(\mu_y - \mu_v)^2}{m+n}$ (4)

zodat $\mathcal{E}\underline{B} \simeq \sqrt{\mathcal{E}\underline{B}_1\, \mathcal{E}\underline{B}_2} = \sqrt{(3)(4)}$ (5)

') De volledige afleidingen worden in het aanhangsel gegeven.

Aldus geldt:

$$\mu_{\underline{r}} = \varepsilon_{\underline{r}} \approx \frac{(m\rho\sigma_x\sigma_y + n\tau\sigma_u\sigma_v)(m+n-1) + mn(\mu_x-\mu_u)(\mu_y-\mu_v)}{\sqrt{[(m+n-1)(m\sigma_x^2+n\sigma_u^2)+mn(\mu_x-\mu_u)^2][(m+n-1)(m\sigma_y^2+n\sigma_v^2)+mn(\mu_y-\mu_v)^2]}} \quad (6)$$

Daar in het algemeen $(m + n) \gg 1$ kan de uitdrukking enigszins worden vereenvoudigd tot:

$$\mu_{\underline{r}} \approx \frac{(1+\frac{m}{n})\rho\sigma_x\sigma_y + (1+\frac{n}{m})\tau\sigma_u\sigma_v + (\mu_x-\mu_u)(\mu_y-\mu_v)}{\sqrt{[(1+\frac{m}{n})\sigma_x^2+(1+\frac{n}{m})\sigma_u^2+(\mu_x-\mu_u)^2][(1+\frac{m}{n})\sigma_y^2+(1+\frac{n}{m})\sigma_v^2+(\mu_y-\mu_v)^2]}} \quad (7)$$

Aan deze formule in zijn meest algemene vorm zijn de invloeden van de verschillende grootheden op $\mu_{\underline{r}}$ niet eenvoudig te zien.

Wij specialiseren daarom tot: $\mu_x = \mu_u$; $\mu_y = \mu_v$; $\sigma_x = \sigma_u$; $\sigma_y = \sigma_v$ met $\rho \neq \tau$.

De twee bipopulaties verschillen nu slechts in de correlatiecoëfficienten.

Dan wordt $\mu_{\underline{r}} \approx \frac{m\rho + n\tau}{m+n}$ en als $\frac{m}{n} = f$, dan is $\mu_{\underline{r}} \approx \frac{f\rho+\tau}{f+1}$ (7a)

Wij onderscheiden drie gevallen en wel:

1) $m \gg n$, dan is $f \gg 1$ en $\mu_{\underline{r}} \longrightarrow \rho$

2) $m \ll n$, dan is $f \ll 1$ en $\mu_{\underline{r}} \longrightarrow \tau$

3) $m = n$, dan is $f = 1$ en $\mu_{\underline{r}} = \frac{\rho+\tau}{2}$

Een andere specialisatie van (6) is: $\rho = \tau = 0$, met $\mu_x \neq \mu_u$ en $\mu_y \neq \mu_v$.

Dan is $\mu_{\underline{r}} \neq 0$

De formule voor $\mu_{\underline{r}}$ wordt:

$$\mu_{\underline{r}} \approx \frac{(\mu_x-\mu_u)(\mu_y-\mu_v)}{\sqrt{[(1+\frac{m}{n})\sigma_x^2+(1+\frac{n}{m})\sigma_u^2+(\mu_x-\mu_u)^2][(1+\frac{m}{n})\sigma_y^2+(1+\frac{n}{m})\sigma_v^2+(\mu_y-\mu_v)^2]}} \quad (8)$$

Het is duidelijk dat de teller kleiner is dan de noemer zodat $|\varepsilon_{\underline{r}}| < 1$.

Als $\mu_x \gg \mu_u$ en $\mu_y \gg \mu_v$ dan is $\mu_{\underline{r}} \approx 1$

De hieruit volgende regel luidt:

Is de steekproef x_i, y_i uit een bipopulatie gestoken die andere gemiddelden heeft als die waaruit de steekproef u_j, v_j afkomstig is, dan zal men als men de reeksen samenvoegt, in de combinatie van de steekproeven, gemiddeld genomen, een correlatie $\neq 0$ vinden, ook al zijn ρ en τ beide 0.

2.2 <u>Een kunstmatig voorbeeld van schijncorrelatie veroorzaakt door menging van steekproeven uit niet identieke populaties</u>.

Wij deden viermaal telkens tien worpen met een zelfde dobbelsteen. De eerste maal leverde de uitkomsten: 3,5,4,1,2,4,6,4,3,5, genoemd x_1 t/m x_{10}.

De tweede maal leverde 4,3,2,5,4,1,6,3,1,2, genoemd y_1 t/m y_{10}.

Aldus ontstaan m = 10 paren x_i, y_i.

De derde maal leverde 6,4,1,4,5,4,3,4,6,5, en deze getallen werden met $\Delta_1 = 20$ vermeerderd en worden u_1 t/m u_{10} genoemd.
De vierde maal leverde 1,4,3,4,5,1,3,6,6,1 en deze getallen werden met $\Delta_2 = 30$ vermeerderd en leverden de reeks v_1 t/m v_{10}. Aldus ontstaan $n = 10$ paren u_j, v_j.
De $m = 10$ paren x_i, y_i worden verenigd met de $n = 10$ paren u_j, v_j tot 20 paren, waarin tussen x en u evenmin als tussen y en v onderscheid wordt gemaakt. De correlatiecoëfficient op basis van deze 20 paren bedraagt + 0,984. (zie fig.3).
Welke gemiddelde waarde van r (μ_r) zou men mogen verwachten als dit spel oneindig vele malen zou worden herhaald?
Wij raadplegen daartoe de formule (8).
Bij een dobbelsteen heeft elk der cijfers 1 t/m 6 evenveel kans om boven te komen zodat $\mu_x = 3,5$ en $\sigma_x^2 = 35/12$, zodat in ons geval $\mu_x = \mu_y = 3,5$; $\mu_u = \Delta_1 + 3,5 = 23,5$ en $\mu_v = \Delta_2 + 3,5 = 33,5$.
In de algemene notatie geschreven is:

$$\mu_r \approx \frac{\Delta_1 \Delta_2}{\sqrt{(\Delta_1^2 + 4\sigma_x^2)(\Delta_2^2 + 4\sigma_x^2)}} = \frac{20 \cdot 30}{\sqrt{(20^2 + 4 \cdot 35/12)(30^2 + 4 \cdot 35/12)}} = +0,979$$

Met $\Delta_1 = 20$ en $\Delta_2 = 30$ is $\mu_r = + 0,979$ (zie fig.3)
Met $\Delta_1 = 3$ en $\Delta_2 = 5$ is $\mu_r = + 0,545$ (zie fig.4 en 4a)
Met $\Delta_1 = 1$ en $\Delta_2 = 1$ is $\mu_r = + 0,079$
In fig.4 zijn x, y en u, v door punten voorgesteld; in fig.4a werden de u, v paren door een kruisje aangegeven.
De uit de steekproeven berekende correlatiecoëfficienten r zijn:
Met $\Delta_1 = 20$ en $\Delta_2 = 30$ is r = + 0,984
Met $\Delta_1 = 5$ en $\Delta_2 = 3$ is r = + 0,65
Met $\Delta_1 = 1$ en $\Delta_2 = 1$ is r = + 0,15

2.3 <u>Een klimatologisch voorbeeld van mengingscorrelatie</u>

Een dergelijk geval doet zich voor bij het onderzoek naar het verband tussen de gemiddelde dagelijkse maximumtemperatuur (zegge x) en het aantal uren zonneschijn (zegge y) te De Bilt, beide per maand.
Laten de paren $x_1, y_1 \ldots\ldots x_{30}, y_{30}$ betrekking hebben op de oktober- en de paren $u_1, v_1 \ldots\ldots u_{30}, v_{30}$ op de novembermaanden van de jaren 1931 t/m 1960.
Men zal geneigd zijn de correlatiecoëfficient te berekenen op basis van $m + n = 60$ paren, liever dan op basis van 30 paren (x,y dan wel u,v) daar in het eerste geval een grotere nauwkeurigheid bereikt wordt.
De 60 paren leveren r = + 0,68. Toch is deze hoge waarde ten dele

schijn en wel om de in 2.1 besproken redenen. De 30 paren x,y leveren r_{xy} (schatting van ρ) = + 0,15.
De 30 paren u,v: r_{uv} (schatting van ρ) = -0,21.
Voorts is $\bar{x}$ (schatting van μ_x) = 14,1°C, $\bar{y}$ (schatting van μ_y) = 101, $\bar{u}$ (schatting van μ_u) = 8,9°C en $\bar{v}$ (schatting van μ_v) = 50.
Wij constateren: $r \gg r_{xy}$ en r_{uv}. Wij hebben hier te doen met het onder 2.1 behandelde effect.

Tabel 1

Enkele correlatiecoëfficienten, berekend uit maandgegevens over 1931 t/m 1960 van De Bilt, die het verband aangeven tussen de gemiddelde dagelijkse maximumtemperatuur en het aantal uren zonneschijn.

maand/tijdvak	$\bar{t}_x$(°C)	uren zon	corr.c.30 paren	60 paren	360 paren
mrt	9,5	126	+ 0,17		
apr	13,4	163	+ 0,37		
mei	21,0	210	+ 0,37		
apr + mei				+ 0,63	
jun	20,8	222	+ 0,80		
jul	22,1	197	+ 0,84		
aug	21,9	185	+ 0,83		
jul + aug				+ 0,83	
sep	19,2	145	+ 0,74		
okt	14,1	101	+ 0,15		
nov	8,9	50	- 0,21		
okt + nov				+ 0,676	
dec	5,4	41	- 0,40		
jan	4,3	56	- 0,49		
feb	5,3	69	- 0,07		
jan + feb				- 0,21	

Alle maanden: a) zonder eliminatie van de jaarlijkse gang + 0,81
b) na eliminatie van de jaarlijkse gang (zie 2.4) + 0,26

2.4 Eliminatie van de jaarlijkse gang

Het voorgaande heeft ook geleerd, dat de berekening van de correlatiecoëfficient over alle maanden in het tijdvak 1931/60 en aldus op 30x12 = 360 paren, alleen dan zinvol is als men de gegevens eerst ontdoet van de jaarlijkse gang en er aldus voor zorgt dat alle herleide gegevens stammen uit universa met hetzelfde gemiddelde (= 0) en dezelfde standaarddeviatie (= 1).
Als men dit doet vindt men een correlatiecoëfficient die gelijk is aan

het gemiddelde van de twaalf correlatiecoëfficienten der afzonderlijke maanden. De waarde bedraagt + 0,26.
Wij kozen juist dit klimatologische voorbeeld omdat wij twee grootheden zochten die gecorreleerd zijn, terwijl de correlatiecoëfficient zich in de loop van het jaar wijzigt en zelfs van teken verandert.
De gedachte is nl. dat een zonnige zomer als regel een warme zomer, maar een zonnige winter als regel een koude winter is.
Enkele kritische opmerkingen willen wij nog maken. In de eerste plaats doet zich binnen de maand hetzelfde voor als binnen het jaar, want ook in een maand verlopen de maximumtemperaturen en de zonneschijnduren. Men zou dan ook met dagwaarden moeten werken, doch daarvan is afgezien met het oog op de omvangrijke arbeid. In de tweede plaats hebben wij maanden als tijdvakken van gelijke lengte beschouwd en ook dit is niet geheel juist.

2.5 <u>De schijncorrelatie als gewogen-gemiddelde correlatie.</u>

Als een steekproef bestaat uit een mengsel van verschillend gecorreleerde grootheden dan zal, gemiddeld beschouwd, de correlatiecoëfficient in deze gemengde steekproef liggen tussen de grootste en de kleinste der twee correlatiecoëfficienten, die men gemiddeld beschouwd, zou vinden als men voor de twee steekproeven afzonderlijk de berekening zou uitvoeren.
Zie: formule (7a).
Wanneer $\tau < \rho$ dan is zeker $\tau < \mu_r < \rho$. Een zeer extreem geval is $\rho = -\tau$, zodat (als m = n) $\mu_r = 0$. Hier is de correlatie zelfs verdwenen, ofschoon ze voor de menging (positief en negatief) aanwezig was.
Daar in (7a) een gewogen gemiddelde der twee correlatiecoëfficienten ρ en τ gezien mag worden, behoeft het boven vermelde niet te verbazen.
Een numeriek voorbeeld construeerden wij met behulp van twee stationnaire Markoffketens I. Een Markoff-keten van type I: $\ldots t_{m-1}, t_m, t_{m+1} \ldots$ wordt gedefinieerd door de autoregressie: $t_i = a.t_{i-1} + e_i$, waarin $|a| < 1$ en e_i een willekeurig uit een normale verdeling gestoken element is. In zulk een reeks is de autocorrelatiecoëfficient ρ_l van de orde l, dit is de correlatiecoëfficient tussen t-waarden op plaatsen i en i+l gelijk aan a^l.

Wij staken uit een Markoff-keten I met $\rho_1 = + 0,7$ tienmaal twee opeenvolgende getallen; 10 paren x,y. Wij namen een tweede steekproef op dezelfde wijze uit een Markoff-keten I met $\rho_1 = -0,7$; vormende 10 paren u,v. Zie tabel 2.

Tabel 2.

	x	y		u	v
i = 1	0,996	0,976	j = 1	0,080	- 0,299
i = 2	0,240	0,418	j = 2	- 0,780	0,568
i = 3	- 0,352	- 0,347	j = 3	- 0,500	0,521
i = 4	- 1,130	- 0,362	j = 4	- 1,454	0,932
i = 5	- 1,670	- 0,119	j = 5	- 1,388	1,529
i = 6	- 0,586	- 1,803	j = 6	0,217	0,052
i = 7	1,929	1,836	j = 7	0,397	- 1,414
i = 8	0,154	0,436	j = 8	- 0,203	0,620
i = 9	0,602	0,200	j = 9	0,598	- 0,804
i =10	- 0,628	- 0,597	j =10	1,297	- 1,337

De uit deze reeksen berekende waarden zijn:

$r_{x,y} = + 0,75$; $s_x = 1,05$; $s_y = 0,97$.

$r_{u,v} = - 0,91$; $s_u = 0,87$; $s_v = 0,98$.

Beide Markoff-reeksen zijn gestandaardiseerd, dat wil zeggen ze hebben een gemiddelde = 0 en een standaarddeviatie = 1, zodat

$$\mu_x = \mu_y = \mu_u = \mu_v = 0 \; ; \; \sigma_x = \sigma_y = \sigma_u = \sigma_v = 1.$$

terwijl, zoals gezegd: $\rho_{x,y} = + 0,7$ en $\rho_{u,v} = - 0,7$.

Maken wij van de twee steekproeven één steekproef van 20 stuks dan vinden wij een correlatiecoëfficient r = + 0,066, welke waarde niet significant van nul blijkt te verschillen.

3.1 Schijncorrelatie tussen twee variabelen vanwege correlatie met een derde variabele.

Er zijn nog andere oorzaken van schijncorrelatie. Op één daarvan vestigen wij de aandacht aangezien zij dikwijls voorkomt.

Er zijn situaties waarin elk van twee variabelen x en y gecorreleerd is met een derde variabele z, waardoor ook x met y gecorreleerd is, dus "langs een omweg".

Toch vragen wij naar de correlatie tussen x en y nadat beide "van de invloed van z zijn ontdaan".

Er is dan een $\rho_{x,y}$ (hoe ook z is), een $\rho_{x,z}$ (ongeacht y) en een $\rho_{y,z}$ (ongeacht x). Uit het materiaal berekent men een $r_{x,y}$ (schatting van $\rho_{x,y}$) maar men wenst te weten $\rho_{xy/z}$ zijnde de populatiecorrelatiecoëfficient tussen x en y nadat deze van de invloeden van z zijn ontdaan. Er geldt dan de, hier niet bewezen, relatie:

$$\rho_{x,y/z} = \frac{\rho_{x,y} - \rho_{x,z}\rho_{y,z}}{\sqrt{(1-\rho_{x,z}^2)(1-\rho_{y,z}^2)}} \qquad (9)$$

Een voorbeeld vinden wij bij het verband tussen de gemiddelde dagelijkse maximumtemperatuur (x) en de zonneschijnduur (y). Beide zijn gecorreleerd met de daglengte (z). Het zou beter zijn, in plaats van de daglengte, de maandsommen van de globale straling te gebruiken, want de daglengte is geen onafhankelijke variabele.

De 360 maandgegevens leveren $r_{x,y} = + 0{,}81$; $r_{x,z} = + 0{,}87$ en $r_{y,z} = +0{,}89$. Substitueren wij deze waarden als schattingen van $\rho_{x,y}$ respectievelijk $\rho_{x,z}$ en $\rho_{y,z}$ in formule (9) dan vinden wij voor $r_{xy|z}$ een waarde van +0,16. Dat deze uitkomst verschilt van de eerder gevonden waarde voor de correlatiecoëfficient + 0,26 nadat de jaarlijkse gang was geëlimineerd (zie 2.4 en tabel 1) moet worden toegeschreven aan het feit dat de berekende r_{xy}, r_{xz} en r_{yz} betrekkelijk onnauwkeurige schattingen zijn van $\rho_{x,y}$, $\rho_{x,z}$ en $\rho_{y,z}$.

Omdat de nodige gegevens nagenoeg alle voorhanden waren, berekenden wij ook nog de correlatiecoëfficient tussen de gemiddelde dagelijkse maximumtemperaturen (x) en de zonneschijnduur in procenten van de maximaal mogelijke duur (y). Zonder voor de jaarlijkse gang te corrigeren komt er $r_{xy} = + 0{,}73$, zijnde van dezelfde orde van grootte als $r_{xy} = + 0{,}81$, hetgeen te begrijpen is, daar ook de relatieve zonneschijnduur sterk van de daglengte afhankelijk is.

4. Aanhangsel

In dit aanhangsel wordt van sommige uitdrukkingen, waarvan in het voorgaande gebruik is gemaakt, de afleiding gegeven.

4.1 Afleiding van $\mathcal{E}_{\underline{x}_i y_i} = \rho\sigma_x\sigma_y + \mu_x\mu_y$, gebruikt om tot de uitdrukking (2) te komen. Wij schrijven de formule voor ρ als volgt:

$$\rho = \frac{\frac{\Sigma x_i y_i}{n} - \mu_x\mu_y}{\sigma_x\sigma_y} = \frac{\mathcal{E}_{x_i y_i} - \mu_x\mu_y}{\sigma_x\sigma_y}$$

hieruit volgt: $\mathcal{E}_{x_i y_i} = \rho\sigma_x\sigma_y + \mu_x\mu_y$ hetgeen te bewijzen was.

Evenzo is $\mathcal{E}_{u_j v_j} = \tau\sigma_u\sigma_v + \mu_u\mu_v$

4.2 Bij de opbouw van (2) werd gesteld: $\mathcal{E}(\Sigma x_i y_i + \Sigma u_j v_j) = m(\rho\sigma_x\sigma_y + \mu_x\mu_y) + n(\tau\sigma_u\sigma_v + \mu_u\mu_v)$.

Dit volgt uit 4.1.

4.3 Bij dezelfde uitdrukking (2) werd gesteld: $\mathcal{E}[(\Sigma x + \Sigma u)(\Sigma y + \Sigma v)] = (m\mu_x + n\mu_u)(m\mu_y + n\mu_v) + mn\rho\sigma_x\sigma_y + n\tau\sigma_u\sigma_v$.

$$\mathcal{E}[(\Sigma x + \Sigma u)(\Sigma y + \Sigma v)] = \mathcal{E}[\Sigma x\Sigma y + \Sigma x\Sigma v + \Sigma u\Sigma y + \Sigma u\Sigma v] \qquad (10)$$

Hierin is $\Sigma x\Sigma y = \sum_1^m x_i y_i + \sum_1^m x_i y_j \ (i \neq j) = m(\rho\sigma_x\sigma_y + \mu_x\mu_y) + (m^2 - m)\mu_x\mu_y$

waarbij gebruik is gemaakt van het feit dat $\underline{x}_i$ wel gecorreleerd is

met $\underline{y}_i$, maar niet met $\underline{y}_j$, voor $i \neq j$, zodat $\mathcal{E} x_i y_j = \mathcal{E} x_i \cdot \mathcal{E} y_j = \mu_x \mu_y$.
Omdat ook $\underline{x}$ en $\underline{v}$ evenals $\underline{u}$ en $\underline{y}$ niet gecorreleerd zijn, wordt (10):

$$mn\rho\sigma_x\sigma_y + mn\mu_x\mu_y + (m^2-m)\mu_x\mu_y + mn\mu_x\mu_v + mn\mu_y\mu_u + n^2\tau\sigma_u\sigma_v + n^2\mu_u\mu_v + (n^2-n)\mu_u\mu_v = (m\mu_x + n\mu_u)(m\mu_y + n\mu_v) + mn\rho\sigma_x\sigma_y + n^2\tau\sigma_u\sigma_v \quad (11)$$

4.4 De verwachtingswaarden voor B_1 en B_2, gebruikt voor de opbouw van (3), (4) en (5) worden als volgt afgeleid:

$$B_1 = \Sigma x^2 + \Sigma u^2 - \frac{(\Sigma x + \Sigma u)^2}{m+n} = \frac{(m+n)\sum_1^m x_i^2 + (m+n)\sum_1^n u_j^2 - \left(\sum_1^m x_i + \sum_1^n u_j\right)^2}{m+n} \quad (12)$$

Wij noemen de eerste, tweede en derde term in de teller respectievelijk A, B en C zodat B_1 kan worden geschreven als $\frac{A + B - C}{m + n}$.
Wij beginnen met de verwachtingswaarde van C.

$C = (x_1 + x_2 + x_3 \ldots\ldots x_m + u_1 + u_2 + u_3 \ldots\ldots u_n)^2$

Uitwerken geeft:

m termen x_i^2

$m^2 - m$ termen x_i, x_j $(i \neq j)$

2 mn termen $x_i u_j$ $(i = \text{of} \neq j)$

n termen u_j^2

$n^2 - n$ termen $u_i u_j$ $(i \neq j)$

Ondersteld wordt, dat x_i en x_j $(i \neq j)$ niet gecorreleerd zijn, evenmin als u_i en u_j $(i \neq j)$ en dat x_i met u_j ook niet gecorreleerd is $(i = \text{of} \neq j)$.

De verwachtingswaarden zijn (zie 4.1 en 4.3):

$\mathcal{E} x_i^2 = \sigma_x^2 + \mu_x^2$

$\mathcal{E} u_j^2 = \sigma_u^2 + \mu_u^2$

$\mathcal{E} x_i x_j = \mu_x^2 \quad (i \neq j)$

$\mathcal{E} x_i u_j = \mu_x \mu_u \quad (i = \text{of} \neq j)$

$\mathcal{E} u_i u_j = \mu_u^2 \quad (i \neq j)$

De verwachtingswaarde van C wordt dan:

$$m(\sigma_x^2 + \mu_x^2) + m(m-1)\mu_x^2 + 2mn\mu_x\mu_u + n(\sigma_u^2 + \mu_u^2) + n(n-1)\mu_u^2$$

De verwachtingswaarde van A wordt: $m(m+n)(\sigma_x^2 + \mu_x^2)$

en die van B: $n(m+n)(\sigma_u^2 + \mu_u^2)$

De verwachtingswaarde van (12) wordt:

$$\frac{m(m+n)(\sigma_x^2 + \mu_x^2) + n(m+n)(\sigma_u^2 + \mu_u^2) - m(\sigma_x^2 + \mu_x^2) - m(m-1)\mu_x^2 - 2mn\mu_x\mu_u - n(\sigma_u^2 + \mu_u^2) - n(n-1)\mu_u^2}{m+n}$$

of

$$\frac{(m+n-1)(m\sigma_x^2 + n\sigma_u^2) + mn(\mu_x - \mu_u)^2}{m+n}$$

hetgeen te bewijzen was.

Op analoge wijze vinden wij de verwachtingswaarde voor B_2.

4.5 Elimineren van de jaarlijkse gang uit en standaardisering van de gegevens van de gemiddelde dagelijkse maximumtemperatuur en de relatieve zonneschijnduur.

De jaarlijkse gang wordt als volgt geëlimineerd.

Zij $x_{i,k}$ de gemiddelde dagelijkse maximumtemperatuur in de maand k van het jaar i; k = 1,2...12 en i = 1,2....30. Voor iedere k berekenen wij $\bar{x}_k$ en s_k^2, het gemiddelde en de variantie van de 30 waarden $x_{i,k}$.
Dan is $c_{i,k} = (x_{i,k} - \bar{x}_k)/s_{xk}$ een grootheid waarvan voor elke k geldt dat het gemiddelde en de variantie respectievelijk 0 en 1 zijn. Dit heet gestandaardiseerd. Op overeenkomstige wijze voeren wij voor de relatieve zonneschijnduur in: $d_{i,k} = (y_{i,k} - \bar{y}_k)/s_{yk}$.
Voor elk van de twaalf maanden zal $r_{xy} = r_{cd}$ zijn, daar de correlatiecoëfficient invariant is voor lineaire transformatie.
Vervolgens worden de 30 paren c, d bij k = 1 (jan) en de 30 paren c,d bij k = 2 (feb).....k = 12 (dec) verenigd tot 360 paren, genoemd p,q,
De 12 populaties hebben zeker een gemeenschappelijk gemiddelde, te weten 0, en een gemeenschappelijke variantie, te weten 1, en mochten zij alle normaal verdeeld zijn dan zijn ze zelfs identiek.
De correlatiecoëfficient r_{pq} van de 360 paren p,q is het gemiddelde van de 12 correlatiecoëfficienten (r_{pq}) .....$(r_{pq})_{12}$ van de twaalf afzonderlijke maanden.
Wij bewijzen dit als volgt:
Per definitie geldt:

$$r_{p,q} = \frac{\sum_1^{360} pq - \left(\sum_1^{360} p \sum_1^{360} q\right)/360}{\sqrt{\left[\sum_1^{360} p^2 - \left(\sum_1^{360} p\right)^2/360\right]\left[\sum_1^{360} q^2 - \left(\sum_1^{360} q\right)^2/360\right]}} = \frac{\sum_1^{360} pq}{\sqrt{\sum_1^{360} p^2 \sum_1^{360} q^2}}$$, daar $\left(\sum_1^{30} p\right)_{k=1} = \left(\sum_1^{30} p\right)_{k=2} \text{---} = 0$

Voor q geldt hetzelfde.
Daar voor elk der 12 maanden afzonderlijk $\sum_1^{30} p^2 = 29 = \sum_1^{30} q^2$.
(want $\sigma_p^2 = 1 = \frac{\sum p^2 - \frac{(\sum p)^2}{n}}{n-1}$; $\sum p = 0$; $n = 30$)
terwijl

$$(r_{p,q})_k = \frac{\left(\sum_1^{30} pq\right)_k}{\sqrt{\left(\sum_1^{30} p^2\right)_k \left(\sum_1^{30} q^2\right)_k}} \quad (k = 1, 2, 3 \text{---} 12)$$

wordt

$$r_{p,q} = \frac{\left(\sum_1^{30} pq\right)_{k=1} + \left(\sum_1^{30} pq\right)_{k=2} + \text{----} \left(\sum_1^{30} pq\right)_{k=12}}{\sqrt{\left[\left(\sum_1^{30} p^2\right)_{k=1} + \text{---} \left(\sum_1^{30} p^2\right)_{k=12}\right]\left[\left(\sum_1^{30} q^2\right)_{k=1} + \text{----} \left(\sum_1^{30} q^2\right)_{k=12}\right]}} = \frac{(r_{p,q})_{k=1} + \text{--} (r_{p,q})_{k=12}}{12} =$$

$\overline{(r_{p,q})_k}$ hetgeen te bewijzen was

4.6 De onder 4.5 behandelde methode geeft aanleiding tot het volgende. Voor alle maanden van het jaar werd de correlatiecoëfficient berekend uit 30 paren x_1y_1 ; 30 paren x_2y_2 .....30 paren $x_{12}y_{12}$. Neemt men nu van elk van de 30 paren de sommen a en b van x en y over 12 maanden; dus $a = x_1 + x_2 \ldots\ldots x_{12}$ en $b = y_1 + y_2 \ldots\ldots y_{12}$, dan heeft men het materiaal van de jaarlijkse gang ontdaan en men kan zich afvragen tot welke uitkomsten de berekening van de correlatiecoëfficient uit de 30 paren a,b leidt.

De algemene formule voor ρ_{ab} kan worden opgesteld als volgt:

$$\rho_{a,b} = \frac{\mathcal{E}\,\underline{a}\,\underline{b} - \mathcal{E}\,\underline{a}\,\mathcal{E}\,\underline{b}}{\sigma_a\,\sigma_b} = \frac{\mathcal{E}(\underline{x}_1 + \underline{x}_2 \cdots + \underline{x}_{12})(y_1 + y_2 \cdots + y_{12}) - \mathcal{E}(\underline{x}_1 + \underline{x}_2 \cdots + \underline{x}_{12})\,\mathcal{E}(y_1 + y_2 \cdots + y_{12})}{\sigma(\underline{x}_1 + \underline{x}_2 \cdots + \underline{x}_{12})\,\sigma(y_1 + y_2 \cdots + y_{12})}$$

De teller wordt aldus uitgewerkt:

De eerste term van deze teller kunnen wij schrijven als:

$$\mathcal{E}\left[\sum_1^{12} \underline{x}_i y_i + \sum_{\substack{i,j=1 \\ (i \neq j)}}^{12} \underline{x}_i y_j\right] \qquad (i = 1 \text{ t/m } 12)$$

Voorts is:

$$\rho_{x_i y_i} = \frac{\mathcal{E}\,\underline{x}_i y_i - \mu_{\underline{x}_i}\mu_{y_i}}{\sigma_{x_i}\,\sigma_{y_i}} \qquad \text{zodat } \mathcal{E}\,x_i y_i = \rho_{x_i y_i}\,\sigma_{x_i}\,\sigma_{y_i} + \mu_{x_i}\mu_{y_i}$$

Gemakshalve schrijven wij voor $\rho_{x_i y_i} : \rho_i$
en voor $\sigma_{x_i} ; \sigma_i$; $\sigma_{y_i} ; \tau_i$ en voor $\mu_{x_i} ; \mu_i$; $\mu_{y_i} ; \eta_i$

De eerste term van de teller wordt dan:

$$\sum_1^{12}(\rho_i\,\sigma_i\,\tau_i + \mu_i\,\eta_i) + \sum_{\substack{i,j=1 \\ (i \neq j)}}^{12} \mu_i\,\eta_j$$

De tweede term kunnen wij schrijven als

$$\sum_1^{12}\mu_i \sum_1^{12}\eta_i = \sum_1^{12}\mu_i\,\eta_i + \sum_{\substack{i,j=1 \\ (i \neq j)}}^{12} \mu_i\,\eta_j$$

De gehele teller wordt dan:

$$\sum_1^{12}(\rho_i\,\sigma_i\,\tau_i + \mu_i\,\eta_i) + \sum_{\substack{i,j=1 \\ (i \neq j)}}^{12}\mu_i\,\eta_j - \sum_1^{12}\mu_i\,\eta_i - \sum_{\substack{i,j=1 \\ (i \neq j)}}^{12}\mu_i\,\eta_j = \sum_1^{12}\rho_i\,\sigma_i\,\tau_i$$

De noemer: $\sigma(x_1 + x_2 \cdots x_{12})\,\sigma(y_1 + y_2 \cdots y_{12})$ wordt, aannemende dat er geen autocorrelatie is van maand op maand, noch voor x noch voor y:

$$\sigma^2(x_1 + x_2 \cdots + x_{12}) = \sigma_1^2 + \sigma_2^2 \cdots \sigma_{12}^2 \quad \text{en} \quad \sigma^2(y_1 + y_2 \cdots + y_{12}) = \tau_1^2 + \tau_2^2 \cdots + \tau_{12}^2 .$$

Bijgevolg wordt:

$$\rho_{a,b} = \frac{\sum_1^{12} \rho_i \sigma_i \tau_i}{\sqrt{\sum_1^{12} \sigma_i^2 \cdot \sum_1^{12} \tau_i^2}} = \sum_1^{12} g_i \rho_i$$

als wij g_i schrijven voor $\frac{\sigma_i \tau_i}{\sqrt{\sum_1^{12} \sigma_i^2 \sum_1^{12} \tau_i^2}}$

In de tabel 3 vinden wij de waarden voor ρ_1 t/m ρ_{12}; σ_1 t/m σ_{12}; τ_1 t/m τ_{12} en g_1 t/m g_{12}

Tabel 3

Waarden voor:

$i=$	ρ_i	σ_i	τ_i	g_i	$g_i \rho_i$
1	- 0,49	2,46	24,00	0,083	- 0,0408
2	- 0,07	3,00	19,05	0,081	- 0,0056
3	+ 0,17	2,00	31,01	0,087	+ 0,0166
4	+ 0,37	1,79	42,47	0,107	+ 0,0397
5	+ 0,37	1,42	29,05	0,058	+ 0,0215
6	+ 0,80	1,55	35,99	0,079	+ 0,0629
7	+ 0,84	1,61	45,49	0,103	+ 0,0868
8	+ 0,83	1,83	39,25	0,101	+ 0,0841
9	+ 0,74	1,73	30,80	0,075	+ 0,0556
10	+ 0,15	1,20	29,43	0,050	+ 0,0075
11	- 0,21	1,27	15,62	0,028	- 0,0059
12	- 0,41	2,04	15,00	0,043	- 0,0177

Hieruit volgt voor $\rho_{a,b}$ een waarde + 0,305.

De rechtstreeks uit het materiaal berekende waarde van $r_{a,b}$ bedraagt + 0,47. De waarde van $\rho_{a,b}$ verschilt praktisch niet van de eerder gevonden waarde + 0,26 (zie tabel 1). Het verschil tussen $r_{a,b}$ en $\rho_{a,b}$ moet worden toegeschreven aan het feit dat wij slechts beschikken over schattingen van de diverse grootheden, verkregen uit een betrekkelijk kleine steekproef.

4.7 Als men een correlatiecoëfficient gevonden heeft, die wijst op een verband tussen twee grootheden dan kan men, als men slechts over waarnemingen van één van de grootheden beschikt, trachten de andere grootheid te schatten met behulp van een regressieformule.

Wil men bijvoorbeeld y schatten uit waarnemingen van x dan heeft deze formule de gedaante:

$$\hat{y} = ax + b.$$

De factoren a en b zijn constanten die men als volgt berekent.

$$a = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = r_{xy} \frac{s_y}{s_x}$$

$$b = \frac{\Sigma x \Sigma xy - \Sigma y \Sigma x^2}{(\Sigma x)^2 - n \Sigma x^2}$$

Alle schattingen van y ($\hat{y}$) liggen op een rechte, de regressielijn van y op x. De werkelijke waarden van y spreiden om deze rechte en als wij deze spreiding uitdrukken als variantie - de restvariantie - dan is deze gegeven door

$$s^2_{\hat{y}} = s^2_y (1 - r^2).$$

Deze restvariantie is kleiner naarmate r groter is. Het spreekt vanzelf dat hoe kleiner deze restvariantie is, des te beter is onze schatting. Als wij terugkeren naar ons klimatologisch voorbeeld, dan vonden wij een hoge correlatie tussen de gemiddelde dagelijkse maximumtemperatuur (maandgegevens) en de zonneschijnduur (eveneens per maand) nl. + 0,81 als wij de jaarlijkse gang niet elimineren. Na eliminatie van de jaarlijkse gang bleef er slechts een correlatiecoëfficient ter waarde van + 0,26 bestaan.

De vraag ligt voor de hand of er met behulp van de hoge correlatie niet een betere schatting van de zonneschijnduur uit de maximumtemperatuur mogelijk is dan met behulp van de lage.

In het volgende zal worden aangetoond dat de lagere coëfficient een betere schatting geeft.

Wij geven hier nog eens de uitdrukking voor de mengcorrelatiecoëfficient (7).

$$\mu_r \approx \frac{(1+\frac{m}{n})\rho\sigma_x\sigma_y + (1+\frac{n}{m})\tau\sigma_u\sigma_v + (\mu_x-\mu_u)(\mu_y-\mu_v)}{\sqrt{\left[(1+\frac{m}{n})\sigma_x^2 + (1+\frac{n}{m})\sigma_u^2 + (\mu_x-\mu_u)^2\right]\left[(1+\frac{m}{n})\sigma_y^2 + (1+\frac{n}{m})\sigma_v^2 + (\mu_y-\mu_v)^2\right]}}$$

We onderstellen $\mu_x = \mu_y = 0$ (hiervoor kan men altijd zorgen)
$\mu_u = \mu_v = \Delta$

$\sigma_x = \sigma_y = \sigma_u = \sigma_v = \sigma$; $\rho = \tau$; $m = n$

In dit geval is $\mu_r = \frac{4\rho\sigma^2 + \Delta^2}{4\sigma^2 + \Delta^2}$ en als wij voor $\frac{\Delta}{2\sigma}$ = k schrijven (k > 0), wordt de uitdrukking $\mu_r = \frac{\rho + k^2}{1 + k^2}$ (13)

Dan is:

als $\rho = 0$: $\mu_r = \frac{k^2}{1+k^2}$ en als $k \longrightarrow \infty$ dan zal $\mu_r \longrightarrow 1$

als $\rho = -1$: $\mu_r = \frac{k^2-1}{k^2+1}$ en als $k \longrightarrow \infty$ dan zal $\mu_r \longrightarrow 1$

als $\rho = +1$: $\mu_r = \frac{k^2+1}{k^2+1}$ = 1

De reeks $y_1, y_2 \ldots\ldots y_m, v_1, v_2 \ldots\ldots v_n$ schrijven wij als $w_1, w_2 \ldots\ldots w_{m+n}$

Dan is: $s^2_w = \frac{\Sigma w^2 - \frac{(\Sigma w)^2}{m+n}}{m+n-1} = \frac{\sum_1^m y^2 + \sum_1^n v^2 - \frac{(\sum_1^m y + \sum_1^n v)^2}{m+n}}{m+n-1}$ (14)

In 4.4 werd de verwachtingswaarde afgeleid van de uitdrukking (12):

De uitdrukking (14) is aan (12) gelijk als wij x en u vervangen door y resp.v, met dit verschil dat wij $\mathcal{E}$ (12) nog moeten vermenigvuldigen met een factor 1/(m+n-1).

$\mathcal{E}$ (14) wordt dan:

$$\mathcal{E}\, s_w^2 = \frac{m}{m+n}\sigma_y^2 + \frac{n}{m+n}\sigma_v^2 + \frac{mn(\mu_y-\mu_v)^2}{(m+n)(m+n-1)} \qquad (15)$$

Specialisatie tot m = n levert (als (m + n) $\gg$ 1):

$$\mathcal{E}\, s_w^2 = \tfrac{1}{2}\sigma_y^2 + \tfrac{1}{2}\sigma_v^2 + \tfrac{1}{4}(\mu_y-\mu_v)^2 \qquad (16)$$

In ons bijzonder geval ($\sigma_y = \sigma_v = \sigma$; $\mu_y - \mu_v = \Delta$) dus:

$$\mathcal{E}\, s_w^2 = \sigma^2 + \tfrac{1}{4}\Delta^2 \qquad (17)$$

De variantie van w is dus groter dan die van y (of v) en wel meer naarmate Δ groter is.

Nu terug naar de restvariantie.

$\sigma_w^2(\text{rest}) = (1-\mu_I^2)\sigma_w^2$ met $\mu_I = \frac{\rho+k^2}{1+k^2}$ (13) en $\sigma_w^2 = \sigma^2 + \frac{1}{4}\Delta^2$ (17)

Is $\sigma_w^2(\text{rest})$ kleiner dan $\sigma_y^2(\text{rest})$?

Wij hebben afgeleid (17) dat $\sigma_w^2 = \sigma^2 + \frac{1}{4}\Delta^2$ of $\sigma^2(1+k^2)$ want $k = \frac{\Delta}{2\sigma}$, zodat $\Delta = 2\sigma k$

De vraag wordt dus:

Is $\left[1-\left(\frac{\rho+k^2}{1+k^2}\right)^2\right](1+k^2)\sigma^2$ kleiner dan $(1-\rho^2)\sigma^2$?

Stel $1 + k^2 = q$ ($q > 1$ want $k > 0$)

De vorm wordt dan:

$$q\left[1-\left(\frac{q-(1-\rho)}{q}\right)^2\right] < (1-\rho^2) \;?$$

Voor een aantal waarden van q en ρ geven de termen links en rechts van het teken $<$ de volgende uitkomsten.

Tabel 4

Waarden van $q\left[1-\left(\frac{q-(1-\rho)}{q}\right)^2\right]$ voor een aantal waarden van q en ρ . Tussen haakjes de waarden van $1-\rho^2$

q	$\rho = 0$	$\rho = +0{,}5$	$\rho = -0{,}5$	$\rho = +0{,}9$	$\rho = -0{,}9$
1,1	1,09 (1,00)	0,772 (0,75)	0,954 (0,75)	0,1909 (0,19)	0,518 (0,19)
1,5	1,33 (")	0,833 (")	1,500 (")	0,193 (")	1,394 (")
2,0	1,50 (")	0,875 (")	1,875 (")	0,195 (")	1,995 (")
2,5	1,60 (")	0,900 (")	2,100 (")	0,196 (")	2,356 (")
3,0	1,66 (")	0,916 (")	2,250 (")	0,197 (")	2,597 (")
10,0	1,90 (")	0,975 (")	2,775 (")	0,200 (")	3,439 (")

In alle gevallen is de restvariantie van w groter dan de restvariantie van y (of v).

De verhouding tussen de beide restvarianties geeft:

$$\frac{2\left[1-\left(\frac{2-(1-\rho)}{2}\right)^2\right]}{1-\rho^2} = \frac{2q-(1-\rho)}{2(1+\rho)}$$

q is groter dan 1 en voor enkele waarden van q en ρ wordt de verhouding tussen de restvarianties:

$q=$	$\rho = -0{,}99$	$\rho = -0{,}5$	$\rho = 0$	$\rho = +0{,}5$	$\rho = +0{,}99$
1,01	2,97	1,03	1,01	1,003	1,00005
2	100,5	2,50	1,50	1,166	1,0025
3	133,7	3,00	1,66	1,222	1,0033
10	180,1	3,70	1,90	1,300	1,004

Dat de verhouding tot 1 nadert bij zeer hoge positieve waarden van ρ is duidelijk. Dat de verhouding tot oneindig nadert als $\rho \longrightarrow -1$ vindt zijn oorzaak in Δ die wij positief namen.

5. Slotbeschouwing

In het voorgaande werd een uitdrukking ontwikkeld die de verwachtingswaarde geeft van de correlatiecoëfficient die ontstaat als men twee steekproeven uit verschillende bi-populaties tot één steekproef verenigt.

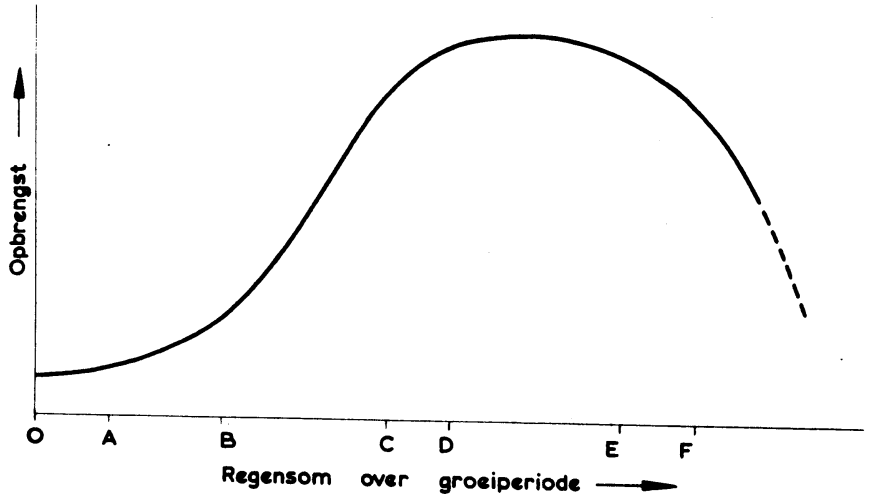
Een dergelijk geval doet zich dikwijls voor, maar in de klimatologie en de landbouwmeteorologie zal het ook voorkomen dat de steekproeven uit meer verschillende bi-populaties afkomstig zijn.

Wij hebben daarmee o.a. te maken als wij verbanden tussen weer en opbrengst van een gebied trachten op te sporen (zie 1.3).

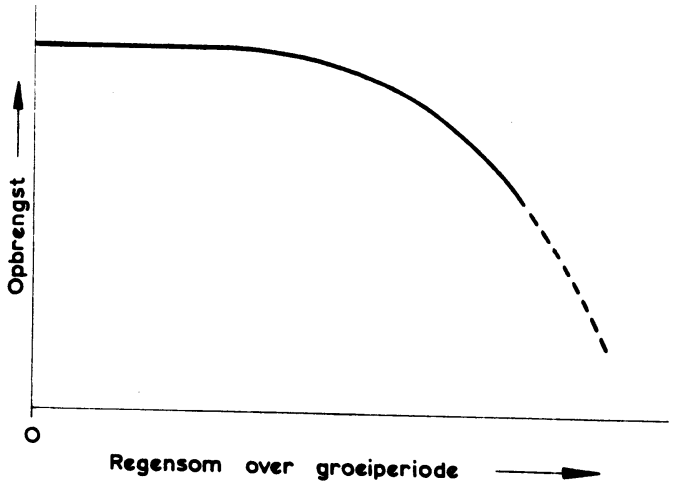
Men kan een uitdrukking ontwikkelen voor de verwachtingswaarde van de correlatiecoëfficient uit drie samengevoegde steekproeven uit verschillende universa, maar reeds dan wordt de uitdrukking praktisch onhanteerbaar. Om het gevaar te signaleren en aan te tonen tot welke consequenties de menging kan leiden lijkt ons de uitdrukking (7) echter voldoende.

6. Literatuur

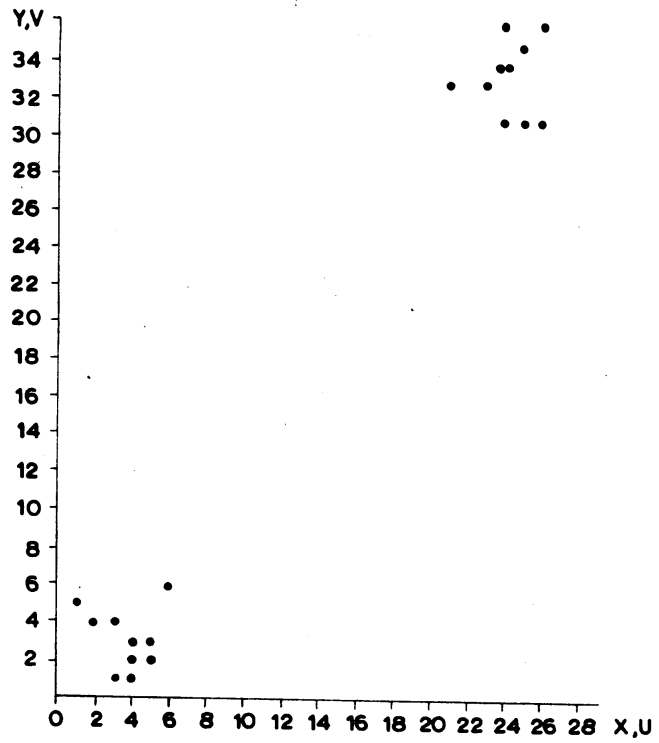
1. Statistische Tabellen en Nomogrammen van de Nederlandse Vereniging voor Statistiek.
2. Pfanzagl.Allgemeine Methodenlehre der Statistik. Sammlung Göschen. 1962.



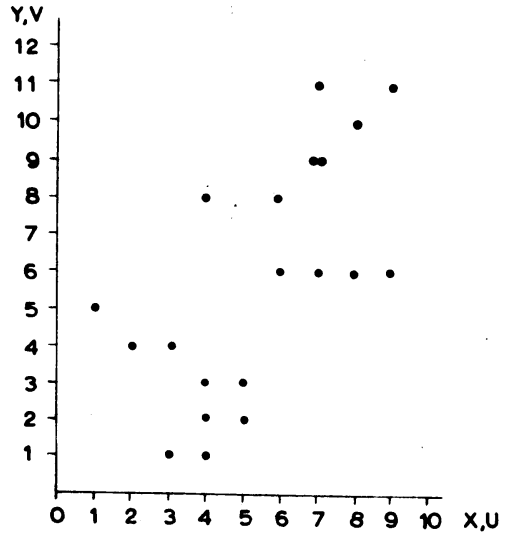
Figuur 1



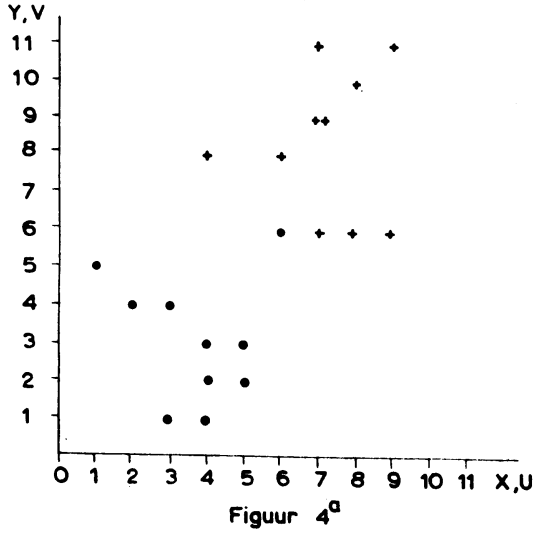
Figuur 2



Figuur 3



Figuur 4



Figuur 4[a]

SUMMARY

Some spurious correlations in agricultural meteorological research.

Research on the relation between weather and crop yields by means of statistical methods most times did not produce satisfactory results. Assuming the normal case of a soil which does not contain enough water from its own for a whole growing season, yields will be very low if there is no or little precipitation. On this soil the amounts of precipitation and yields are positively correlated.
This correlation exists up to a certain amount of rain but if more rain falls during the growing season the extra amount will not cause higher yields. In this range there is no correlation left. If rainfall is superabundant, yields will decrease and the correlation becomes negative. Assuming another case of a soil which contains enough water for the whole growing season, small amounts of precipitation will not influence the yields but abundant rainfall will reduce yields and correlation becomes negative again.

Using data of crop yields and rainfall from a whole area and a long series of years we are dealing with different soils and different amounts of precipitation and as a consequence with different relationships between precipitation and yields.
Mingling of samples from different populations often produces spurious correlations.
The effect of mingling two samples of m pairs $\underline{x}_i, \underline{y}_i$ - from a bi-population $\underline{x}_i, \underline{y}_i$ with a population correlationcoëfficient ρ, means μ_x, μ_y and variances σ_x^2, σ_y^2 - with n pairs $\underline{u}_j, \underline{v}_j$ - from a bipopulation $\underline{u}, \underline{v}$ with a population correlationcoëfficient τ, means μ_u, μ_v and variances σ_u^2, σ_v^2 - is shown by the expression for the expected value of r - the correlationcoëfficient of the mixed sample-being:

$$\mu_r \simeq \frac{\left(1+\frac{m}{n}\right)\rho\sigma_x\sigma_y + \left(1+\frac{n}{m}\right)\tau\sigma_u\sigma_v + (\mu_x-\mu_u)(\mu_y-\mu_v)}{\sqrt{\left[\left(1+\frac{m}{n}\right)\sigma_x^2 + \left(1+\frac{n}{m}\right)\sigma_u^2 + (\mu_x-\mu_u)^2\right]\left[\left(1+\frac{m}{n}\right)\sigma_y^2 + \left(1+\frac{n}{m}\right)\sigma_v^2 + (\mu_y-\mu_v)^2\right]}}$$

In a special case, for instance: $\mu_x=\mu_u$; $\mu_y=\mu_v$; $\sigma_x=\sigma_u$; $\sigma_y=\sigma_v$; $m=n$; $\rho \neq \tau$ one obtains: $\mu_r = \frac{\rho+\tau}{2}$

If two (or more) samples - of equal size or not - from bipopulations with different means, variances or correlationcoëfficients are joined in one sample, the correlationcoëfficient calculated from this sample is different from the correlationcoëfficients of the populations from which the samples were taken.

A good example of the effect of mingling was found in the relation between the monthly values of the daily maximum temperature and those of the sunshine duration from the period 1931-'60 at De Bilt - The Netherlands.

The correlationcoëfficient calculated from the 360 pairs was + 0.81. Both variables however show an annual course and so the sample is coming from 12 different populations.

Correlationcoëfficients calculated for each of the 12 months separately (30 pairs) range from -0.49 for January, up to +0.84 for July.

After eliminating the annual course, the correlationcoëfficient from the 360 pairs decreases to +0.26.

It is proved that the predictive value of the smaller coëfficient is better than this value of the greater one, due to the so-called residual variance or variance of prediction.