KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Verslagen V-61
(R III-242-1960)

Verslag van eerste resultaten met betrekking tot verwachtingsmogelijkheden van de zomerneerslag in het Deltagebied

door

Dr C. Levert

Vraag: Is het mogelijk (langs statistische weg) een redelijk betrouwbare verwachting van de totale hoeveelheid zomerneerslag in het Deltagebied te geven enige maanden vóór de zomer begint?

Neven-vraag: zo neen, misschien wel in de toekomst, b.v. over 10 jaren?; zo ja, zou deze verwachting over bijv. 10 jaren veel beter zijn dan nu?

A. Na overleg met Heren Volker en Valken werd besloten de definities van de begrippen als
"langs statistische weg (de methodiek), redelijk betrouwbaar, de zomer, het Deltagebied, "enige" maanden van te voren, betere verwachting" aan het K.N.M.I. over te laten.

Dit voorlopig rapport is verdeeld in:

1. Soorten verwachtingen.
   - 1.1 Een klimatologische verwachting
   - 1.2 Een regressie-verwachting
   - 1.3 Een tijdreeks-extrapolatie
   - 1.4 Een weersverwachting op lange termijn.
2. Enige theoretische details.
   - 2.1 De associatie en de correlatie coëfficiënt.
   - 2.2 De lineaire regressie als beschrijving van het stochastische verband
   - 2.3 De lineaire regressie bij het gebruik in signifikantie-onderzoekingen
   - 2.4 De lineaire regressie bij het gebruik bij "forecasting". Drie belangrijke problemen
   - 2.5 Opmerkingen, van belang vooral als het aantal paren x, y klein is.
3. De eerste numerieke resultaten
   - 3.1 Correlatie-coëfficiënten
   - 3.2 Regressie-rechten; frequentieverdelingen op waarschijnlijkheidspapier
   - 3.3 Puntenwolken
   - 3.4 Betere correlatie over 10 jaren?
4. Wat nu verder te doen?

1. Soorten verwachtingen

Na gedachtenwisselingen over deze vraag, gehouden op het K.N.M.I., kwamen 4 "methodieken" naar voren; bij een keuze daaruit zou gelet moeten worden op het belang snel een voorlopige indruk van de oplosbaarheid van het probleem te verkrijgen.

1.1 Een klimatologische verwachting

Het woord "verwachting" is hier nauwelijks op zijn plaats.
Wanneer de frequentieverdeling van de zomerneerslag (bijv. de som over juli en aug., bijv. gemiddeld over drie stations in het Delta-gebied) zeer goed bekend zou zijn, is het altijd mogelijk op basis daarvan voor de komende zomer te beweren (een "verwachting" te geven), dat er een kans k_1 (H) is, dat de hoeveelheid $\geq$ H zal zijn of dat er een kans k_2 (H) is op een hoeveelheid $\leq$ H of dat er een kans k_3 is, dat $H' \leq$ hoeveelheid $\leq H''$, waarbij wij H, H' en H'' willekeurig mogen kiezen. Deze k's zullen des te onbetrouwbaarder zijn, naarmate H groter of kleiner of het H'., H'' — gebied breder gekozen wordt (en des te meer naarmate de basisperiode minder lang is). Deze onbetrouwbaarheid kan worden berekend. De vraag of zulk een "verwachting" (die dus geen gebruik maakt van wat er in de maanden voor de zomer in kwestie viel) over bijv. 10 jaren "beter" (betrouwbaarder) zal zijn dan nu, moet bevestigend beantwoord worden (althans als intussen het "klimaat" niet verandert), louter omdat wij dan over een groter basismateriaal, dat ons de genoemde frequentieverdeling levert, kunnen beschikken. Maar als nu reeds dit basismateriaal "groot" is (bijv. 50 jaren) zal deze verbetering over bijv. 10 jaren (50 → 60 jaren) zeker niet bijzonder markant zijn.

1.2 Een regressie-verwachting

Wij denken nu aan een regressie-onderzoek, voorafgegaan door een associatie-analyse. Bestaat er een niet-toevallige (z.g. statistisch significante) associatie tussen de zomerneerslag in het Delta-gebied en de neerslag (of een ander element) in ditzelfde Delta-gebied (of elders in de wereld) in een maand (of in een combinatie van successieve maanden) vóór deze zomer? Zo ja, dan moet een regressie-analyse volgen, die deze associatie met een regressie-vergelijking zo goed mogelijk tracht te beschrijven (lineaire, kromlijnige regressie?), waarna deze regressie voor het maken van verwachtingen zal worden gebruikt.

Echter behoeft een signifikante associatie (signifikante correlatie-coëfficiënt) niet per se een in de praktijk bruikbare regressie te impliceren. Rondom de regressierechte (eventueel kromme) behoort een betrouwbaarheidsband getekend te worden. De lineairiteitscoëfficiënt (de z.g. correlatie-coëfficiënt) moet namelijk niet alleen signifikant, maar ook "zeer groot" zijn, opdat de z.g. rest-variantie (residu) voldoende klein is en de genoemde betrouwbaarheidsgordel daarvoor voldoende smal zal zijn, hetgeen zal betekenen, dat alleen zéér strakke (maar toch stochastische) verbanden bruikbaar zullen zijn. Verder behoren wij te definiëren wanneer wij een lineaire regressie "beter" noemen dan een klimatologische; het zal dan blijken, dat het weinig zinvol (althans gevaarlijk) is om de regressierechte te gebruiken bij waarden van de verklarende variabele buiten een bepaald gebied, hetgeen betekent dat, als de verklarende variabele (zeer) klein of (zeer) groot is (waarschijnlijk is ze juist dan interessant) de verwachting alle (althans veel) waarde verliest.

1.3 Tijdreeks-extrapolatie

Methode van Craddock? Hierbij wordt uitgebreid rekening gehouden (volgens een bepaald rekenschema) met de waarden der elementen in voorbije maanden. Deze analyse eist zeer veel rekenwerk.

1.4 Weersverwachting op lange termijn (physische basis)

Deze methodiek, welke nog onvoldoende ontwikkeld is, wordt verder buiten beschouwing gelaten.

2. Enige theoretische details

Hoewel de resultaten van de berekeningen het belangrijkst zijn, is het goed terwille van een verantwoorde en juiste interpretatie van deze resultaten, vooral enige aandacht aan mathematisch-statistische details te schenken.

2.1 De associatie en de correlatie-coëfficiënt.

Dikwijls ziet men in de op klassieke wijze (via produkten en kwadraten) in de steekproef berekende correlatie-coëfficiënt r tussen y en x een maat voor de sterkte van de associatie tussen x en y in het universum. Na de berekening van r volgt raadpleging van de signifikantie-tabel voor de r.

Indien de gemeten r de uit de tabel afgelezen drempel r_d overtreft, heet de correlatie signifikant. Eerst dan heeft het zin om er mee verder te werken. Te bedenken is hierbij, dat dit alles slechts geldt voor <u>geassocieerde normale</u> verdelingen $N_x(\mu_x;\sigma_x)$ en $N_y(\mu_y;\sigma_y)$. De associatiegraad heet ρ, met $0 \leq |\rho| \leq 1$. Als $\rho = 0$ dan heten N_x en N_y <u>onafhankelijk</u>, d.w.z. gegeven een x, dan volgt y de normale verdeling N_y, die niets met deze x te maken heeft.

Is $|\rho| = 1$, dan is er een <u>functioneel verband</u> tussen x en y.

Ligt $|\rho|$ tussen 0 en 1 dan volgt, bij gegeven x, de y een normale verdeling rondom een gemiddelde $E(y \mid x) = \beta_2 x + \alpha_2$ met $\beta_2 = \rho\sigma_y/\sigma_x$ en $\alpha_2 = \mu_y - \beta_2\mu_x$ en met $\sigma_2(x) = \sigma_y\sqrt{1-\rho^2}$ als standaarddeviatie (die niets met x te maken heeft). Men kan bewijzen, dat associatie, d.w.z. niet-onafhankelijkheid, van twee normaal-verdeelde variabelen slechts het bovenstaande impliceren kan.

Het "werken" met de klassieke correlatie-coëfficiënt houdt daarom in: <u>het weten, althans het onderstellen dat x én y normaal verdeeld zijn</u>! Natuurlijk zal men noch μ_x, noch μ_y, noch σ_x, noch σ_y, noch ρ kennen; de beperkte steekproef (n paren x, y) levert <u>schattingen</u> $\bar{x}$, $\bar{y}$, s_1, s_2 en r, waarvan de juistheid (zegge betrouwbaarheid) met n samenhangen zal.

<u>Opm</u>: omgekeerd, als y gegeven is, volgt x een normale verdeling met $E(x \mid y) = \beta_1 y + \alpha_1$ en $\sigma_1(y) = \sigma_x\sqrt{1-\rho^2}$, met $\beta_1 = \rho\sigma_x/\sigma_y$ en $\alpha_1 = \mu_x - \beta_1\mu_y$.

2.2 <u>De lineaire regressie als beschrijving van het stochastische verband</u>

De opgave is: Bij <u>elke</u> waarde van de verklarende variabele x behoort een waarschijnlijkheidsverdeling van y, met zeker gemiddelde $E(y \mid x)$, met zekere variantie $\sigma_2^2(x)$, met zeker derde moment etc. Het is de bedoeling <u>deze y-waarschijnlijkheidsverdeling</u> (in zijn geheel een functie van x) <u>te vinden</u>. Dat betekent (althans zo gaat men gewoonlijk te werk) in de eerste plaats:

van welke aard is de functie $E(y \mid x) = f(x)$? Hebben we een sterk vermoeden (gesteund door vroegere ervaringen of verkregen na een speciaal statistisch onderzoek), dat het om een <u>lineaire functie</u> $f(x) = \alpha x + \beta$ zal gaan, dan willen we <u>alles</u> te weten komen over de numerieke waarden van deze α en β en wel op basis van de beperkte

steekproef (de statistische waarde der schattingen; zuiverheid; doeltreffendheid; efficiency; gemiddelde waarden; betrowubaarheidsmarges, etc.). Dit zal slechts gelukken door óók nog onderstellingen te maken over genoemde tweede, derde enz. momenten t.a.v. hun verband met x (meestal neemt men dan géén verband aan). Bovendien komen er moeilijkheden i.v.m. het feit, dat x en y nooit foutloos gemeten zijn. Verder rijzen er problemen zowel ten aanzien van de aard der kansverdeling, waaraan x gehoorzaamt, als wel van die der kansverdeling van y, als x gegeven is.

De moderne theorie behandelt uitvoerig de waarde van de alom bekende "Methode der Kleinste Kwadraten", waarbij blijkt, dat zeer bepaalde voorwaarden aan de verdelingen van x en y (x) gesteld moeten worden, alsmede aan de waarnemingsfouten, opdat deze M.d.K.K. tot zuivere en meest doeltreffende schattingen der in de lineaire regressie optredende parameters zal leiden.

Het zijn deze voorwaarden, die men zo licht over het hoofd ziet. Helaas kan men meestal (door gebrek aan materiaal) niet verifiëren of er aan wordt gehoorzaamd, hetgeen niet mag rechtvaardigen het bestaan ervan te negeren.

Het verst ontwikkeld is de theorie der lineaire regressie voor het geval x niet aan een waarschijnlijkheidsverdeling gehoorzaamt (dus discreet te kiezen is, zoals in landbouwexperimenten dikwijls het geval is).

Helaas hebben wij in de klimatologie juist dikwijls wèl te maken met een x, die een verdeling volgt (de maandsom neerslag, de dagelijkse neerslag, het aantal regendagen per maand, bijv.). Voor dat geval is de theorie óók uitgewerkt, echter mits x voldoet aan een verdeling uit een klasse van scherp aangeefbare typen, waartoe, gelukkig, ook de normale distributie behoort.

2.3 De lineaire regressie bij het gebruik in signifikantie-onderzoekingen

Vragen

a. Helt de ware regressierechte statistisch zeker?

b. Als wij in twee onderzoekingen twee beste regressierechten vinden, die verschillend hellen, zullen de richtingscoëfficiënten van de twee ware regressierechten dan statistisch verschillen?

c. Verschilt de constante term in de uitdrukking voor de ware regres-

sierechte statistisch zeker van nul, als die in de uitdrukking voor de beste regressierechte (on)gelijk aan nul is?

d. Tussen welke grenzen ligt, met een waarschijnlijkheid van bijv. 0,95, de ware regressiecoëfficiënt?

e. Behoort een nieuw koppel x_{n+1}, y_{n+1} tot dezelfde bipopulatie, als waartoe de andere n paren behoren?

2.4 De Lineaire regressie bij het gebruik bij "forecasting", interpolatie en extrapolatie.

Aan dit gebruik denken wij in het bijzonder in het onderhavige probleem.

Voorwaarden: Bij elke waarde van x behoort een kansverdeling van y, geschreven $f(y \mid x)\,dy$, die volgens gausz is en wel met een gemiddelde η dat lineair met x samenhangt, d.i. $\eta = \alpha + \beta x$, en met een variantie σ^2, welke niet met x samenhangt. Met het stellen van deze voorwaarde wordt bedoeld, dat tot nu toe slechts dan de theorie goed uitgewerkt is. De theorie is aanvankelijk uitgewerkt voor x - waarden, die "instelbaar" zijn, d.w.z. niet een kansdistributie volgen, en foutloos zijn. Voor zover wel met fouten behept, moeten de kansverdelingen der fouten aan bepaalde voorwaarden voldoen. Voor zover we de x - waarden niet vrij in de hand hebben (in de klimatologie), is de theorie ook uitgewerkt voor zeer bepaalde kansverdelingen (weinig scheef!), waartoe die van gausz ook behoort.

De volgende drie belangrijke vragen a, b en c kunnen gesteld worden:

a. Welk betrouwbaarheidsgebied kunnen wij voor de theoretische regressie-waarde η_i bij gegeven x_i berekenen?

Antw: Er zijn n x-waarden $x_1, x_2 \ldots x_n$ (gemakshalve in volgorde van toeneming geschreven) gemeten; bij elk ervan behoort één y (natuurlijk kunnen er bijv. 3 verschillende y's gemeten zijn bij één enkele x-waarde, welke x dan drievoudig geteld wordt). Bij bijv. x_3 behoort $\hat{y}_3 = \hat{\alpha} + \hat{\beta} x_3$, waarin $\hat{\alpha}$ en $\hat{\beta}$ de beste (volgens de M.d.K.K. berekende) schattingen van α en β voorstellen. Hoe goed representeert deze de ware, doch onbekende $\eta_3 = \alpha + \beta x_3$?

Het ligt voor de hand, dat bij een tweede stel, volkomen dezelfde x-waarden als boven andere y's gemeten zullen worden (gedeeltelijk of volkomen andere), zodat er dan een andere, beste, regressierechte resulteert en dezelfde x_3 een andere $\hat{y}_3$ bepalen zal. Toch is ook

deze nieuwe $\hat{y}_3$ een beste puntschatting van dezelfde, onbekende, η_3.

Het blijkt, dat deze $\hat{y}_3$ een kansverdeling volgt, waarvan het gemiddelde natuurlijk η_3 is en de variantie is:

$$\hat{\sigma}^2(\hat{y}_3) = \sigma^2\left[\frac{1}{n} + \frac{(x_3-\bar{x})^2}{\sum_1^n (x_i-\bar{x})^2}\right] \quad ; \bar{x} = \frac{1}{n}\sum_1^n x_i$$

Hierin is σ^2 (meestal óók onbekende) variantie van de toevallige component van y.

Deze wordt het beste geschat met

$$\hat{\sigma}^2 = \sum_1^n (y_i - \hat{y}_i)^2/(n-2)$$

Zodoende wordt de beste schatting $\{\hat{\sigma}(\hat{y}_3)\}^2$ van $\sigma^2(\hat{y}_3)$

$$\{\hat{\sigma}(\hat{y}_3)\}^2 = \frac{\sum_1^n (y_i-\hat{y}_i)^2}{n-2}\left[\frac{1}{n} + \frac{(x_3-\bar{x})^2}{(n-1)s_x^2}\right] \quad \text{als } s_x^2 = \frac{\sum_1^n (x_i-\bar{x})^2}{n-1}$$

Het quotiënt $(\hat{y}_3 - \eta_3)/\hat{\sigma}(\hat{y}_3)$ blijkt een t-verdeling (Student-verdeling) te volgen met n-2 graden van vrijheid, g.v.v. (hoe groter de n, hoe minder verschilt deze t-verdeling van een normale)

De factor $[\quad]$ in bovenstaande uitdrukking is altijd $\geq \frac{1}{n}$. Hij is het grootst voor de "uiteinden" x_1 en x_n en het kleinst voor die x_i (uit de n x_i's), waarvoor $|x_i - \bar{x}|$ het kleinst is (als er één x_i toevallig gelijk aan $\bar{x}$ is, dan is $[\quad] = \frac{1}{n}$)

Let wel: deze procedure bepaalt een gordel rondom de regressierechte tussen x_1 en x_n. De gordel is aan de bovenzijde door een hyperbooltak begrensd, aan de onderzijde eveneens. De gordel is het smalst bij die x_i, die het dichtst bij $\bar{x}$ ligt en des te smaller naarmate n groter is.

Uitspraak: op grond van de gemeten n punten, mogen we zeggen, dat de ware regressie-waarde van y, behorende bij x_3, met een waarschijnlijkheid 0.95 gelegen is tussen $\hat{y}_3 - t_{n-2}\cdot\hat{\sigma}(y_3)$ en $\hat{y}_3 + t_{n-2}\cdot\hat{\sigma}(y_3)$ waarbij $\hat{y}_3 = \hat{\alpha} + \hat{\beta}x_3$ en t_{n-2} de in de t-tabel bij n-2 g.v.v. en P = 0.95 afgelezen drempel is.

b. Binnen welk gebied ligt met een waarschijnlijkheid van bijv. 0.95 de onbekende regressierechte ofwel binnen welk gebied liggen 95 % der rechten, die men zou verkrijgen als men bij herhaling de y-waarden

behorende bij een gegeven stel dezelfde x-waarden beschouwen zal?

Antw: Ook dit gebied (eveneens een gordel) wordt begrensd door twee hyperbooltakken, de ene boven, de andere beneden de beste regressierechte $y = \hat{\alpha} + \hat{\beta}x$ gelegen. Men moet vanuit elk der n punten x_i, y_i naar boven en naar beneden eenzelfde segment uitzetten, groot $\hat{\sigma}\sqrt{2F}\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}\right]$. F stelt voor de F-waarde, afgelezen in de z.g. F-P-tabel voor 2 en n-2 g.v.v. bij P = 0.95. Men ziet, dat dit segment weer van x_i afhangt, maar anders dan onder vraag a. Eerst stond er vóór [ ] de factor t_{n-2}; thans $\sqrt{2F} > t_{n-2}$. De "ruimte" tussen deze twee hyperbooltakken is groter dan die onder vraag a.

c. Welke is de gemiddeld bij een n + 1 ste x-waarde x_{n+1} behorende y en binnen welke grenzen mag men met een waarschijnlijkheid van bijv. 0.95 de door x_{n+1} bepaalde y verwachten, alles louter en alleen op basis van de op de gepasseerde n paren x, y berekende regressierechte?

Dit is het "echte" forecasting-probleem, waarbij men allicht veel interesse heeft voor het geval, dat x_{n+1} ver ligt t.o.v. $\bar{x}$, in het bijzonder $x_{n+1} \ll x_1$ of $x_{n+1} \gg x_n$.

Wij onderstellen weer, dat de bij x_{n+1} behorende y_{n+1} een normale waarschijnlijkheidsverdeling zal volgen (dit is een noodzakelijk te maken hypothese!) en wel rondom het gemiddelde $\eta_{n+1} = \alpha + \beta x_{n+1}$, met variantie σ^2, maar noch α, noch β, noch σ zijn exact bekend; we hebben er slechts schattingen van. Het blijkt, dat $(y_{n+1} - \hat{y}_{n+1}) : s_d$ een t-verdeling volgt met n-2 g.v.v.

Hierbij is $$s_d^2 = \frac{\sum_1^n (y_i - \hat{y}_i)^2}{n-2}\left[\frac{n+1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)s_x^2}\right] = s_y^2(1-r^2)\frac{n-1}{n-2}[g]$$

N.B. Eerst nu voeren we de steekproefcorrelatiecoëfficiënt in:

$$r = \Sigma(x-\bar{x})(y-\bar{y}) / s_x \cdot s_y$$

We zagen: onder a: $\left\{\hat{\sigma}(\hat{y}_i)\right\}^2 = s_y^2(1-r^2)\frac{n-1}{n-2}f$; met $f = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} \geqslant \frac{1}{n}$

en thans $s_d^2 = s_y^2(1-r^2)\frac{n-1}{n-2}g$, met $g = f + \frac{1}{n} \geqslant 1 + \frac{1}{n}$

zodat $\boxed{s_d^2 = \left[\hat{\sigma}(\hat{y}_{n+1})\right]^2 + \hat{\sigma}^2}$

Gebruik: zoek in de t-tabel de drempel t_{n-2} bij bijv. P = 0.95 en n-2 g.v.v. op. De bewering, dat de bij x_{n+1} "behorende" y zal liggen tussen

$\hat{y}_{n+1} - t_{n-2} \cdot s_d$ en $\hat{y}_{n+1} + t_{n-2} \cdot s_d$
heeft een betrouwbaarheid 0.95.

Eveneens:

De bewering, dat de gemiddeld bij x_{n+1} te verwachten y zal liggen tussen $\hat{y}_{n+1} - t_{n-2} \hat{\sigma}$ en $\hat{y}_{n+1} + t_{n-2} \hat{\sigma}$, is in 95 % der gevallen juist.

2.5 Opmerkingen

1. men mag strikt genomen niet werken met de sterke vereenvoudiging $s_d^2 = (1-r^2) s_y^2 \frac{n-1}{n-2}$ (alsof g = 1), waarbij vele leerboeken bovendien nog $\frac{n-1}{n-2}$ ongeveer 1 nemen. (De vereenvoudiging is des te meer fout naarmate én n kleiner is én x_i verder van $\bar{x}$ ligt)

2. men behoort met de t-verdeling (n-2 g.v.v.) te rekenen en niet met de normale verdeling ($n = \infty$), d.w.z. het is niet waar, dat y_{n+1} een normale verdeling volgt rondom $\hat{y}_{n+1}$ met $s_y^2(1-r^2)$ als variantie, maar wel, dat $(y_{n+1} - \hat{y}_{n+1}) : s_d$ aan een t-verdeling op n-2 g.v.v. gehoorzaamt.
N.B.: bij n = 10 is $t_{10} = 2.31$; n = 20, dan $t_{20} = 2.10$; n = ∞ (norm. verd.) dan $t_\infty = 1.96$ (alles bij 0.95 betrouwbaarheid; bij geringere betrouwbaarheid wordt het verschil tussen t_n en t_∞ groter)

3. Als men met de x_{n+1} ergens tussen x_1 en x_n zit en des te meer naarmate $|x_{n+1} - \bar{x}|$ kleiner is, én als n niet te klein is, is $g \approx 1$. Doch we moeten ook belangstelling hebben voor grote $|x_{n+1} - \bar{x}|$, bijv. $|x_{n+1} - \bar{x}| : s_x \geq 2$ (extreme waarden van x_{n+1}), d.w.z. wij moeten zeker ook beschouwen $x_{n+1} < x_1$ en $x_{n+1} > x_n$. Dan is zeker $f > \frac{1}{n}$ en $g > 1 + \frac{1}{n}$.

4. Een veel voorkomend misverstand bestaat hierin, dat men de "standaardfout van schatting", berekend via de steekproef, als een maat beschouwt van de "standaard-deviatie van de forecast" (in onze formules: resp. de wortels van $s_y^2(1-r^2)$ en $s_d^2 = s_y^2 \frac{n-1}{n-2} g$) Deze interpretatie gaat alleen op voor (zeer) grote steekproeven, via welke de regressiekarakteristieken berekend zijn. Men behoort eigenlijk altijd te bedenken (en speciaal als n klein is, bijv. n $\leq$ 10, maar wat klein is hangt ook samen met s_x en s_y), dat er niet slechts is een "error variance of estimate" rondom de regressielijn, maar ook mog een steekproef-fout van de regressierechte zelf.

In verband met het vorige heeft het gebruiken van de regressierechte om ermee te "forecasten" geen zin meer als s_d, vanwege een ver van $\bar{x}$ ver-

wijderde x_{n+1}, toch weer boven s_y uitkomt. Met toenemende $|x_{n+1} - \bar{x}|$ nl. neemt f toe en zal $(1-r^2)\frac{n-1}{n-2}g$ tenslotte (eerst kleiner dan 1 zijnde) toch groter dan 1 worden en s_d boven s_y uitkomen. Wij zullen dus zulk een gebied voor x_{n+1} moeten berekenen in alle gevallen van een signifikante regressie. De gebruiker moet zeggen of hij dit gebied onvoldoende breed vindt. Wij kunnen ons nl. voorstellen, dat hij ten zeerste teleurgesteld is als hij verneemt dat de regressie juist als x_{n+1} zeer klein of zeer groot is haar forecastings-waarde volkomen verliest.

Verder kan de gebruiker zelfs bij x_{n+1}-waarden binnen bovenbedoeld gebied, ontevreden zijn over de verwachting, nl. de 95 % betrouwbaarheidsband veel te breed is voor nuttig gebruik in de praktijk. Wanneer nl. op grond van een totale hoeveelheid neerslag van bijv. 100 mm over september + oktober voor de komende zomersom (juli + augustus) een gemiddeld te verwachten waarde van 105 mm resulteert, echter met een betrouwbaarheidsmarge lopende van 10 tot 200 mm, dan zal deze uitkomst wel van weinig betekenis geacht worden. De band is zo breed door de niet zeer grote correlatie-coëfficiënt (0.42). Wij kunnen het ook zo zeggen: de klimatologische verwachting zou met een s_y = 55 mm werken; door de signifikante correlatie en de lineaire regressie wordt deze s_y verlaagd tot $s_y\sqrt{1-r^2}$ = 50 mm, dus slechts net 10 %. Alhoewel de correlatie signifikant is, leidt ze toch slechts tot 10 % verlaging van de s_y, m.a.w. met de afhankelijke variabele is te "weinig verklaard". Men zou eigenlijk wel 10 % onverklaard willen zien, d.i. $\sqrt{1-r^2} \leq 0.10 \therefore r \geq 0.99$. Natuurlijk is dit voor een deel een kwestie van smaak, maar het lijkt toch logisch de eis te stellen om d.m.v. de lineaire regressie tot een rest-st.dev. te komen, die ≤ 0.4 van de uitgangst.deviatie brdraagt. Dit leidt tot een eis: $r \geq 0.9$. Dit zijn hoge correlaties, die men in klimatologisch werk niet veel tegenkomt. Natuurlijk kan men, naar eigen smaak, met minder betrouwbaarheid genoegen willen nemen; bijv. 0.75 i.p.v. 0.95. Bij (zeer) grote n behoort dan een t_{n-2} = 1.15 en de 0.95-betrouwbaarheidsband is $105 \pm 1.96 \cdot 48.5$, d.i. loopt van 10 tot 200 mm, moet dan vervangen worden door de 0.75-betr. band $105 \pm 1.15 \cdot 48.5$, d.i. loopt van 49 tot 161 mm.

De band is smaller geworden, maar het voordeel is slechts schijnbaar, aangezien de betrouwbaarheid van de uitspraak, dat de komende y-waarde binnen

de berekende band liggen zal, gedaald is van 95 tot 75 %. Wil men "scherper" voorspellen, dan moet men met minder betrouwbaarheid genoegen nemen.

3. De eerste numerieke resultaten

Voor een eerste oriëntatie omtrent de mogelijkheid en de practische waarde van zulk een forecast d.m.v. lineaire regressie en teneinde de eerste resultaten snel (zij het daarmede ten koste van enige nauwkeurigheid) te verkrijgen, waarom wij direct beschikbare geponste gegevens machinaal lieten behandelen en een zeer simpele approximatieve statistische methodiek kozen, is het volgende geschied:

3.1 Berekening van correlatie-coëfficiënten voor Vlissingen tussen de hoeveelheid neerslag in de zomer (zowel juli + augustus als mei + juni + juli + augustus, genoemd z resp. Z) en de totale hoeveelheid over een tijdvak van 2, 4, 6, 8 of 10 sucessieve, voorafgaande, maanden, ook te Vlissingen, welke tijdvakken op afstanden van een even aantal maanden t.o.v. de zomer gelegen waren.

Basismateriaal: 47 jaren, 1906-1953.

Deze tijdvakken waren

z = jul + aug	Z = mei + jun + jul + aug	k = a + b + c = sep t/m feb
a = sep + okt	f = a + b = sep t/m dec	l = b + c + d = nov t/m apr
b = nov + dec	g = b + c = nov t/m feb	m = c + d + e = jan t/m jun
c = jan + feb	h = c + d = jan t/m apr	
d = mrt + apr	j = d + e = mrt t/m jun	
e = mei + jun		

n = a + b + c + d = sep t/m apr p = a + b + c + d + e = sep t/m jun

o = b + c + d + e = nov t/m jun

3.2 Berekening van de regressierechte alleen voor die z. of Z. combinaties, waarbij de correlatiecoëfficiënt op basis van een 0.05-onbetrouwbaarheid signifikant zijn.

De correlatiecoëfficiënt r (steeds in een steekproef van 47 paren) werd niet op de correcte wijze via $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$ berekend, maar via een 4 velden-tableau, met mediaanscheidingen: de z.g. mediaan-correlatie, die slechts op aantallen berust

Deze meridiaan-correlatie (gekozen, juist omdat ze gemakkelijk en snel, louter machinaal kan worden berekend) is alleen dan een goede approximatie van de klassieke (en des te meer naatmate N, hier 47, groter is) als de geassocieerde variabelen normaal verdeeld zijn. Het was dus zaak ook hierover snel een indruk te verkrijgen, hetgeen geschiedde door van elk der 17 genoemde variabelen de frequentieverdeling op lineair-waarschijnlijkheidspapier uit te zetten en daarna op het oog de lineariteit der ligging der 47 punten te beoordelen. Het blijkt dat er voor sommige der tijdvakken van weinig (2, 4) maanden nogal wat hapert aan deze normaliteit (het zou te veel tijd gevergd hebben hierover meer finesses via normaliteitstoetsen te verkrijgen). Er werd in elk van deze 17 prentjes op het oog een rechte tussen de punten door gelegd(soms ook enkele), waarmede een grove schatting van σ berekend kon worden. Deze is nodig in de regressievergelijking.

De mediaan-correlatie $\tilde{r}$ is een gedefinieerd a.v.

$\tilde{a}$ \ $\tilde{z}$	<	>
<	A	B
>	C	D

47

Hiernaast is het 4-velden tableau voor de z, a combinaties getekend; $\tilde{z}$ = mediaan= 142 mm en $\tilde{a}$ = mediaan = 131 mm; A = totale aantal paren z, a waarbij zowel $z < \tilde{z}$ als $a < \tilde{a}$. N = 47 A bleek 15.

Hoe strenger de correlatie, hoe voller is elk der A en D-cellen, ten koste van de C en B-cellen. Bij een absolute non-correlatie zou het in het meest ideale geval elk der 4 cellen een aantal $(\frac{1}{2})^2$ N moeten bevatten (doch het steekproefeffect veroorzaakt vullingen der cellen, die er tóch van verschillen). Er geldt p.d.

$$\tilde{r} = \cos 180 \left(1 - \frac{2A}{47}\right)$$; hier A = 15 en $\tilde{r}$ (z, a) = 0.424.

De 95 % drempel is 0.29 (de 99 % drempel is 0.37). Deze 0.424 verschilt derhalve signifikant van nul. Natuurlijk heeft deze r een sterk discreet karakter (des te erger naarmate N kleiner is), aangezien A nu eenmaal slechts de hele waarden 0, 1, .., 47 aannemen kan. We moeten dit goed in het oog houden als wij ons over de gelijkheid van sommige correlatiecoëfficiënten verbazen, terwijl we ongelijkheid zouden hebben verwacht.

Een volgende beperktheid in onze aanpak is gelegen in de enkelvoudigheid van de correlatie. Natuurlijk zou men ook een tweevoudige, drievoudige corr. moeten proberen. Om der wille van de tijd is dit (nog) niet gebeurd. (zie onder 4.3)

De eerste resultaten zijn a.v. samengevat:

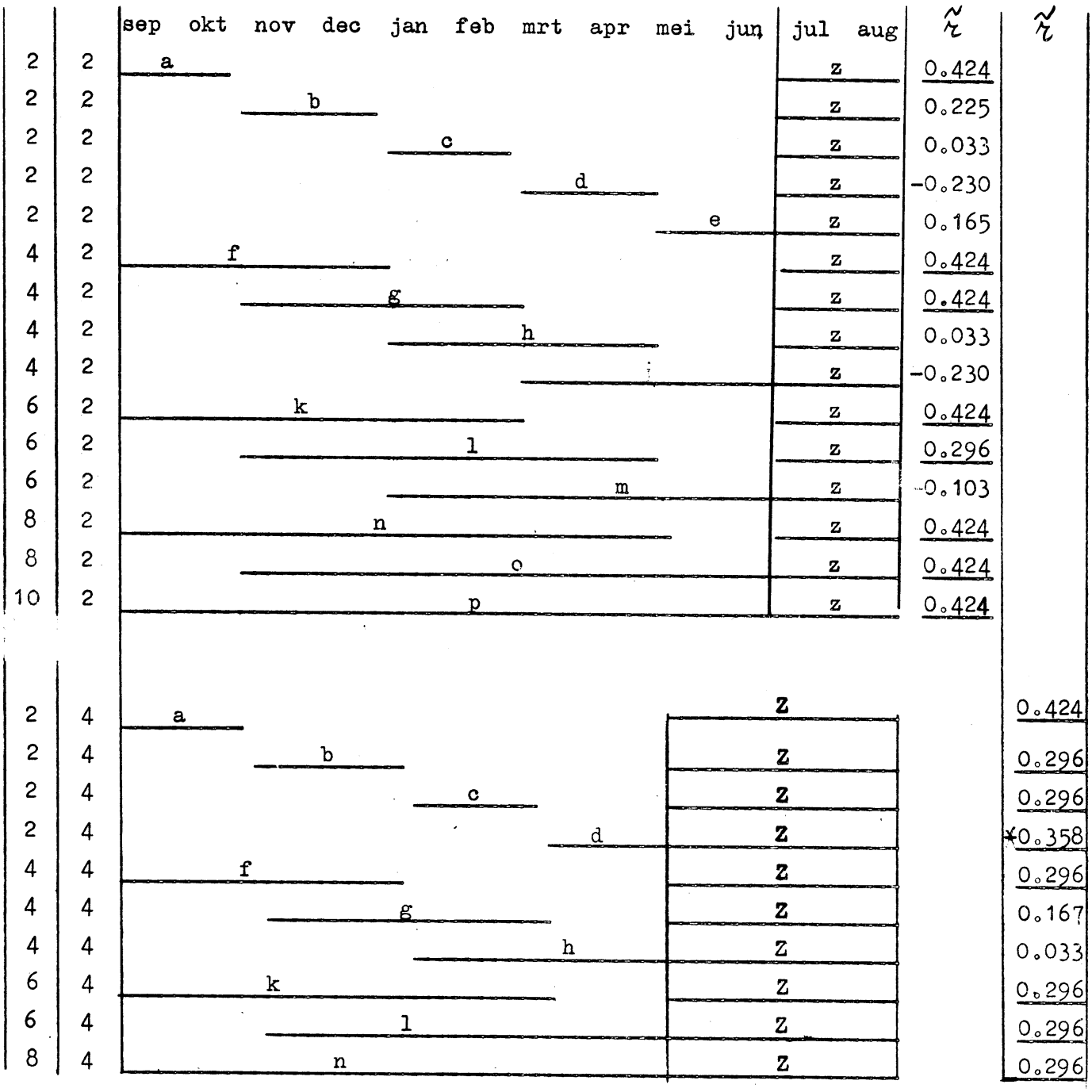
Vlissingen 47 jaren (1906-1953)

tijdvak	aantal mnd in tijdvak	extremen in mm m	M	mediaan mm $\tilde{x}$	gemidd mm $\bar{x}$	geschatte s mm	corr.coëff z met	Z met
z = jul + aug	2	13	242	142	130	55		
Z = mei t/m aug	4	116	395	230	231	98		
a = sep + okt	2	38	233	131	135	50(43à60)	0.424	0.424
b = nov + dec	2	43	256	136	140	62	0.225	0.296
c = jan + feb	2	37	200	102	100	45	0.033	0.296
d = mrt + apr	2	33	147	79	86	34	-0.230	0.358
e = mei + jun	2	39	177	102	100	37	0.165	
f = sep t/m dec	4	94	479	271	277	83	0.424	0.296
g = nov t/m feb	4	93	380	247	240	87	0.424	0.167
h = jan t/m apr	4	73	224	178	186	50	0.033	0.033
j = mrt t/m jun	4	94	307	186	187	56	-0.230	
k = sep t/m feb	6	167	566	390	378	109	0.424	0.296
l = nov t/m apr	6	143	323	323	328	100	0.296	0.296
m = jan t/m jun	6	137	292	292	288	97	-0.103	
n = sep t/m apr	8	223	452	452	454	127(98-136)	0.424	0.296
o = nov t/m jun	8	223	685	434	428	107	0.424	
p = sep t/m jun	10	262	817	563	563	124	0.424	

Significante r-waarden (95 % drempel) zijn onderstreept.

296
358
424

Overzichtsschema



N.B. We hebben de indruk, dat hier de N (= 47) toch wel wat klein was om er de r mee te berekenen. Het markante discrete karakter van $\check{r}$ blijkt lastig te zijn. Zie het volgende tabelletje, dat dit illustreert.

N	A	$\tilde{r}$	A	$\tilde{r}$
47	11	-0.103	16	0.537
47	12	0.033	17	0.648
47	13	0.225	18	0.740
47	14	0.296	19	0.827
47	15	0.424	20	0.891

Ondanks alle gebreken aan onze aanpak, die wij bewust hebben aanvaard, omdat wij snel een eerste oriëntatie wilden verkrijgen, mogen we wel zeggen, dat deze lineaire regressie tot lage correlatie-coëfficienten leidde; zelfs de meest signifikante zijn veel te laag, als wij bedenken, dat zij in de buurt van 0.7 à 0.8 zouden moeten liggen om er voor een redelijk bruikbare voorspelling wat aan te hebben.

2.3 Grafieken

De prentjes, elk met 47 punten , z tegen a, b, enz. of Z tegen a, b, enz. geven duidelijk te zien hoe weinig strak de verbanden zijn; vele puntenwolken zijn zeer bolvormig, andere zijn wat meer sigaarvormig, doch te breed. In enkele ervan hebben wij de lineaire regressie-rechten getekend, bijv. $z=\bar{z} + r \frac{Sz}{S_a} (a - \bar{a})$ waarbij $\bar{z}$, $\bar{a}$, r (z,a), s_z en s_a berekend of grafisch geschat moesten worden. Ook zijn de betrouwbaarheidsgordels aangebracht, berekend met de in de inleiding genoemde formules. Eveneens zijn aangebracht, op enkele grafieken, de grenzen (op de hor-as), waarbuiten de variabele "niet mag komen", d.w.z. komt daar de verklarende variabele te liggen, dan heeft de forecast-st-deviatie een grootte, gelijk aan of zelfs groter dan de s_z (of s_Z) (d.i. de lineaire-regressie-verwachting is dan "niets" beter dan de klimatologische).

Wanneer de correlatiecoëfficiënten voldoende groot uitgevallen zouden zijn, zouden wij ons de achtereenvolgende verwachtingen a.v. gedacht hebben:

1. Sep + okt is gepasseerd; direct daarna volgt op grond daarvan de verwachting: juli + aug zal gemiddeld $\bar{z}$ zijn en de z_1 zal liggen tussen z_{11} en z_{21} met 0.9 waarschijnlijkheid.
2. Daarna wordt nov. + dec. afgewacht; op basis van sep t/m dec. volgt dus een nieuwe verwachting $\bar{z}_2$, met z_2 tussen z_{21} en z_{22}.
3. Dan wordt jan + feb afgewacht.

   De som sep t/m febr zou leiden tot $\bar{z}_3$ en z_{31} à z_{32} etc.

Dit alles mislukt (althans lijkt ons geen practische bruikbaarheid te hebben) door de te lage correlatie-coëfficiënten, waardoor de marges namelijk z_{11} à z_{12} à z_{21} à z_{22} etc. veel te breed zijn. Anders gezegd: de s_z (s_Z) wordt door de lineaire-regressie wel is waar verkleind, en wel tot $s_z\sqrt{1-r^2}$, doch te weinig, zelfs al is de r= 0.42, immers dan is $\sqrt{1-0.42^2} = 0.9$ zodat er maar 10 % "winst" is.

Een numeriek voorbeeld: stel sep + okt = 135 mm (dat is juist de gemiddelde waarde over 47 jaren). De regressie rechte en de 95 % band leren dan: gemiddeld [1] is hierbij een jul + aug van 130 mm te verwachten, en met 95 % waarschijnlijkheid zal de zomerhoeveelheid liggen tussen ongeveer 130 - 2 x 50 = 30 en 130 + 2 x 50 = 230 mm, waarbij $50 = 55\sqrt{1 - r^2}$, met s (z) = 55 mm en r = 0.42.

3.4 Betere correlatie over 10 jaren?

In verband met de "nevenvraag" of wij over bijv. 10 jaren in staat zullen zijn een "betere" verwachting te maken, valt het volgende op te merken: Het correlatienomogram van Pawlowski, dat berust op de onderstelling, dat elk der twee eventueel geassocieerde universa (onbekende correlatie-coëfficiënt ρ) een normale verdeling heeft, zegt: op basis van N=47 paren en een steekpr. corr. 0.42, is met 95 % waarschijnlijkheid $\rho \geqslant 0.15$, doch eveneens: $\rho \leqslant 0.63$ terwijl ook met 95 % waarschijnlijkheid geldt: $0.10 \leqslant \rho \leqslant 0.65$. Wanneer wij over 10 jaren, d.w.z. met N=57, toevallig eenzelfde steekproef-correlatie 0.42 zouden berekenen, zouden wij krijgen een marge $0.15 \leqslant \rho \leqslant 0.63$ (een geringe vernauwing van de marge derhalve).

Het valt op hoe breed de marge is!

Voor de beneden-grens ρ=0.10 is $\sqrt{1-\rho^2}$ =0.995, d.w.z. er is maar 0.5 % verklaard.

Voor de boven-grens ρ=0.65 is $\sqrt{1-\rho^2}$ =0.76, d.w.z. er is maar 24 % verklaard.

---

1) Zelfs deze gemiddeld te verwachten hoeveelheid 130 mm is zeer onbetrouwbaar, want de ware waarde ervan ligt, met een 95 % waarschijnlijkheid, ergens tussen 130-15=115 en 130+15=145 mm. Hierbij is $2 \times 55 \times \sqrt{1-0.42^2} \times \frac{1}{\sqrt{47}} \cong 15$. Men lette op én de kleine r én de kleine N. De marge wordt zelfs nog breder als de totale hoeveelheid over sep. en okt. boven 135 mm uitkomt.

Wanneer wij afspreken (een kwestie van smaak) van een werkelijke verbetering van de verwachting d.m.v. een lineaire regressie te mogen spreken als de standaarddeviatie van de verklaarde variabele daardoor met meer dan de helft vermindert (d.i. $\sqrt{1-\rho_o^2} \geq \frac{1}{2}$), zou dit eisen $\rho_o \geq 0.865$. Wij vragen ons dus af of het mogelijk zou zijn op basis van de over 10 jaren in een steekproef van 57 paren berekende correlatie-coëfficiënt r' een ρ-marge te vinden, met een bovengrens van ten minste 0.865. Het nomogram raadplegende, blijkt dan nodig dat $r' \geq 0.76$. Bij $N' = 57$ en $r' = 0.76$ is dan $0.59 \leq \rho \leq 0.865$. Maar zou (thans) N = 47 met r = 0.42 "verenigbaar" zijn met (in de toekomst) N' = 57 met r' = 0.76, d.w.z. zou het verschil tussen 0.42 en 0.76 geheel aan het steekproefeffect toegeschreven kunnen worden?

Antw.: indien deze steekproeven onafhankelijk zouden zijn, zou er een toevalskans 0.08 zijn op zulk een combinatie van r-waarden, terwijl toch beide steekproeven stammen uit eenzelfde bipopulatie. Echter zijn de steekproeven niet onafhankelijk (de 47 "zit" in de 57), waardoor de 0.08 te groot zal zijn (hoeveel te groot?), waarop inderdaad reeds de geringe overlapping der marges $0.10 \leq \rho \leq 0.65$ en $0.59 \leq \rho \leq 0.86$ zou wijzen. Wij hebben dit statistische aspect niet verder bestudeerd, doch menen te mogen stellen, dat er een slechts kleine kans zal zijn om over 10 jaren een r = 0.76 te vinden (en een zeker nog kleinere kans op een nog grotere r).

Conclusie:

Al met al is een en ander weinig hoopvol.

4. Wat nu verder te doen?

1. Wij zouden kunnen overgaan van één station op het gemiddelde van een aantal stations, die redelijk het Delta-gebied representeren. Bijv. de stations Vlissingen, Oudenbosch, Kerkwerve, Axel.
   Het voordeel is dat de 2, 4... maandsommen zeker veel beter normaal verdeeld zullen zijn dan dezelfde per station en dit feit zal er toe leiden, dat de mediaan-correlaties, beter de op correcte wijze berekenbare correlatie-coëfficiënten zullen benaderen dan voor één station het geval is. Overigens zijn echter de gegevens der 4 stations niet volkomen onafhankelijk.
2. We zouden de correlatie-coëfficiënten wèl correct kunnen berekenen. Veel werk.

3. We zouden op meervoudige regressies kunnen overstappen, bijv. niet afzonderlijk $z = \alpha_a a + \beta_a$, $\alpha_b b + \beta_b$, $\alpha_c c + \beta_c$, $\alpha_d d + \beta_d$, $\alpha_e e + \beta_e$, maar een vijfvoudige lineaire regressie $z = \alpha a + \beta b + \gamma c + \delta d + \varepsilon e + \zeta$ Dit zou tot zeer veel rekenwerk leiden (producten ab, ac, ad, ae, bc, etc, kwadraten a^2, b^2....) en een 6-tal lineaire vergelijkingen (via M.d.K.K.) in 6 onbekenden. Bovendien moeten partiële correlatie-coëfficiënten berekend worden, waarbij thans de mediaan-correlatie-methode niet meer geoorloofd zou zijn. Tenslotte moet dan de rest-variantie $\frac{1}{N-6} \sum (z_g - z_t)^2$ berekend worden, waarin z_g = gemeten, z_t = regressie, en dan moeten wij maar hopen, dat de "rest-standaarddeviatie" zeer veel kleiner dan s(z) zal zijn!

   We hebben een sterk vermoeden, dat de verbetering van de regressie weinig naam zal hebben en lang niet het vele werk zou compenseren.

4. We zouden de neerslag met de luchtdruk kunnen correleren, in Vlissingen zelf of elders in de wereld. Maar zal dit tot veel grotere correlaties leiden? Wij betwijfelen het, gezien vooral de resultaten die Prof. Visser in zijn bekend werk over "Weersverwachting op lange termijn" noemt.

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Aanhangsel
Verslag V-61
(RIII-242-1960)

Toelichting bij de figuren 1 t/m 8.

1) Op blz. 12 van het aangehaalde verslag worden 17 frequentieverdelingen genoemd. Bijgaande figuren 1, 2, 3 en 4 brengen vier van deze zeventien in beeld. Zij stellen voor, op lineair-waarschijnlijkheidspapier, de cumulatieve frequentieverdeling van de totale hoeveelheid neerslag.
in fig. 1 a in het twee-maanden tijdvak sep. + okt.
in fig. 2 z in het twee-maanden tijdvak jul. + aug.
in fig. 3 f in het vier-maanden tijdvak sep. + okt. + nov. + dec.
in fig. 4 Z in het vier-maanden tijdvak mei + jun. + jul. + aug.

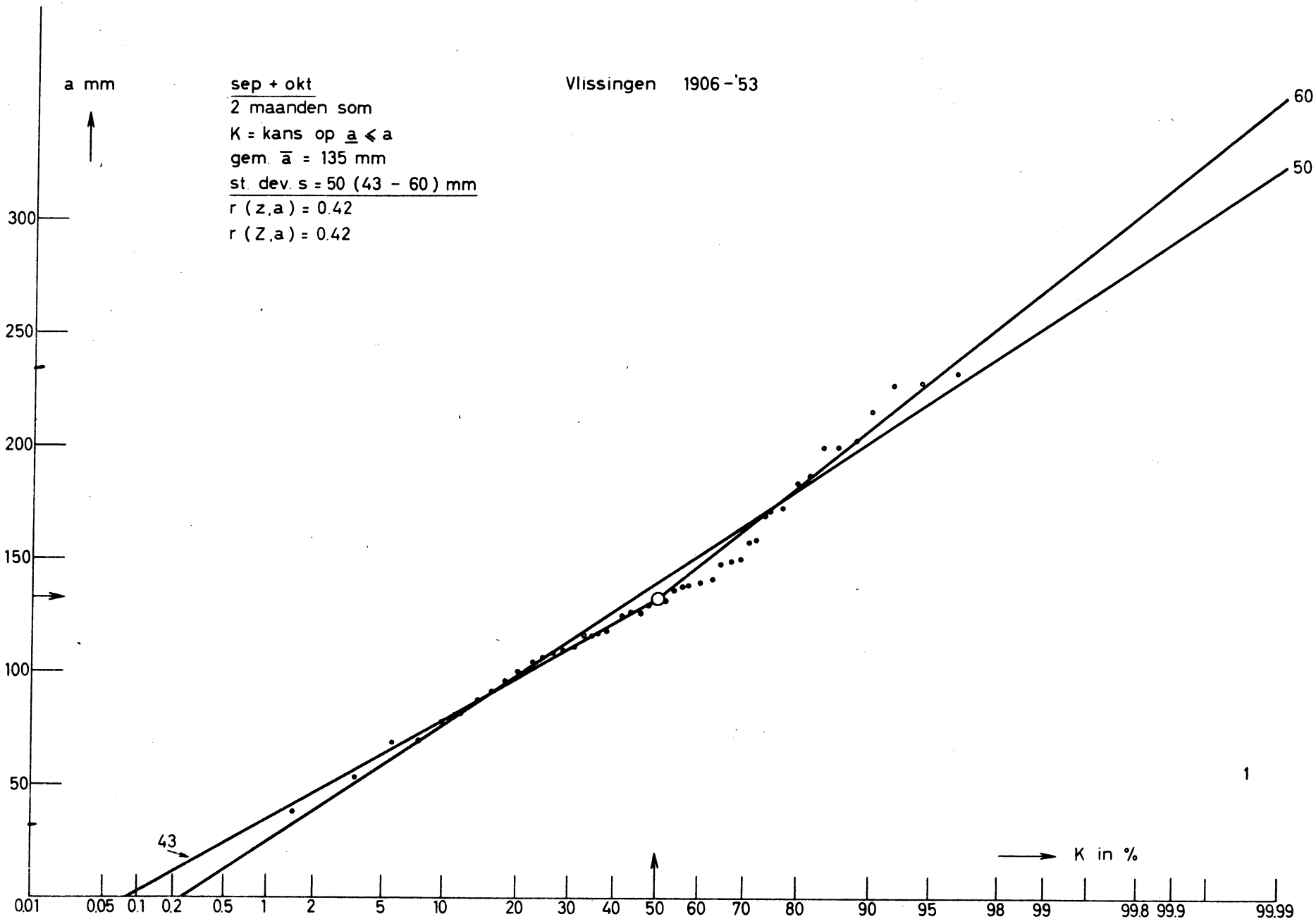
2) Op blz. 15 van het rapport wordt gesproken van de lineaire regressies z tegen a, z tegen b, enz., Z tegen a, Z tegen b, enz. Bijgaande figuren 5 en 6 brengen het verband in beeld.
in fig. 5 tussen z en a,
in fig. 6 tussen z en f.
De zg. 95%-betrouwbaarheidsband is in beide figuren gelegen tussen de gestreepte 5%-krommen. Bovendien is op elk der twee horizontale assen een tweetal gebieden gearceerd en wel
in fig. 5 $a < 0$ en $a > 280$ mm,
in fig. 6 $f < 30$ en $f > 510$ mm.
In de tekst wordt uiteengezet, dat voor a-waarden in deze gebieden of voor f-waarden in deze gebieden, de verwachting door middel van de lineaire regressie "niets beter" is dan de zuiver klimatologische.

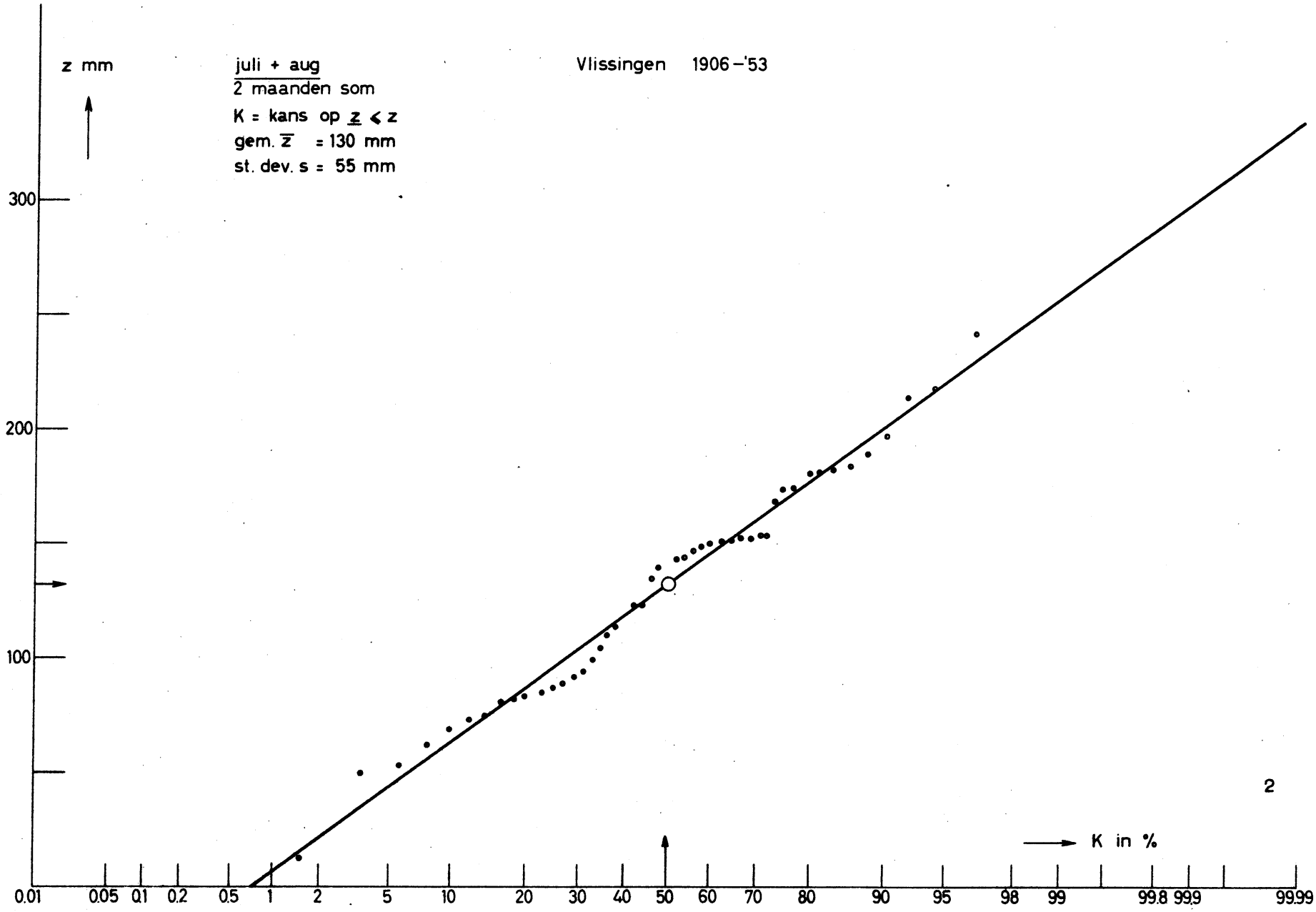
3) In de figuren 7 en 8 zijn de correlatie-coëfficiënten nog eens samengenomen.
Fig. 7 brengt in beeld hoe de correlatie-coëfficiënt r(z,i) tussen z en i afhangt van de "afstand" tussen de gecorreleerde tijdvakken; i is daarbij a, b, c, d, e; f, g, h, j; k, l, m; n, o en p.
Men kan uit de vrije hand curven met parameters 2, 4, 6, 8 tekenen.
Fig. 8 brengt in beeld hoe de correlatie-coëfficiënt r(Z,i) afhangt van de "afstand" tussen Z en i; i als in fig. 7.
Nadere bijzonderheden, zie tekst blz. 14.

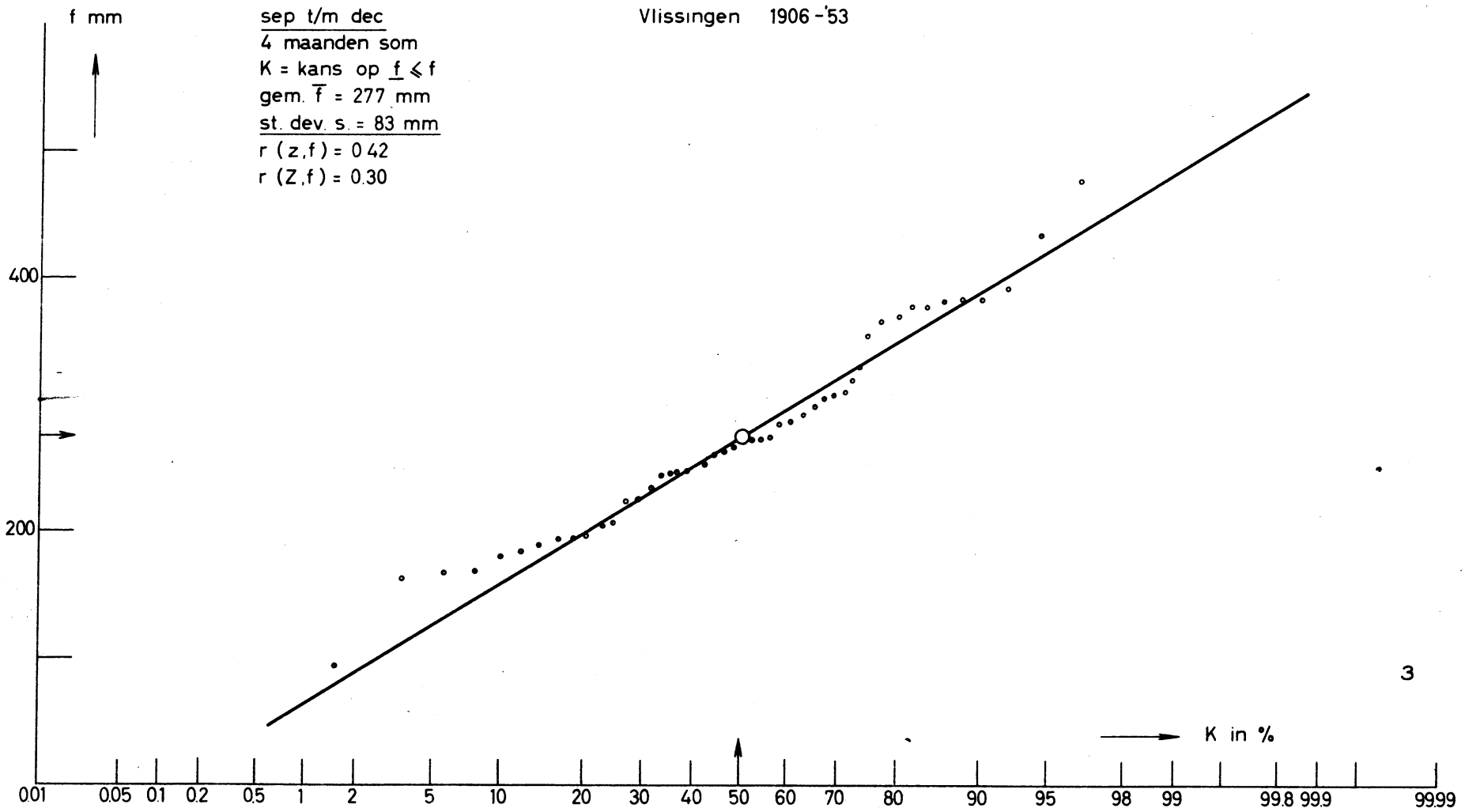
Het discrete karakter van r, veroorzaakt (zie tekst blz. 12) door de benaderingsmethode, volgens welke r berekend werd (mediaancorrelatie), is er oorzaak van, dat de curven niet zo gemakkelijk getrokken kunnen

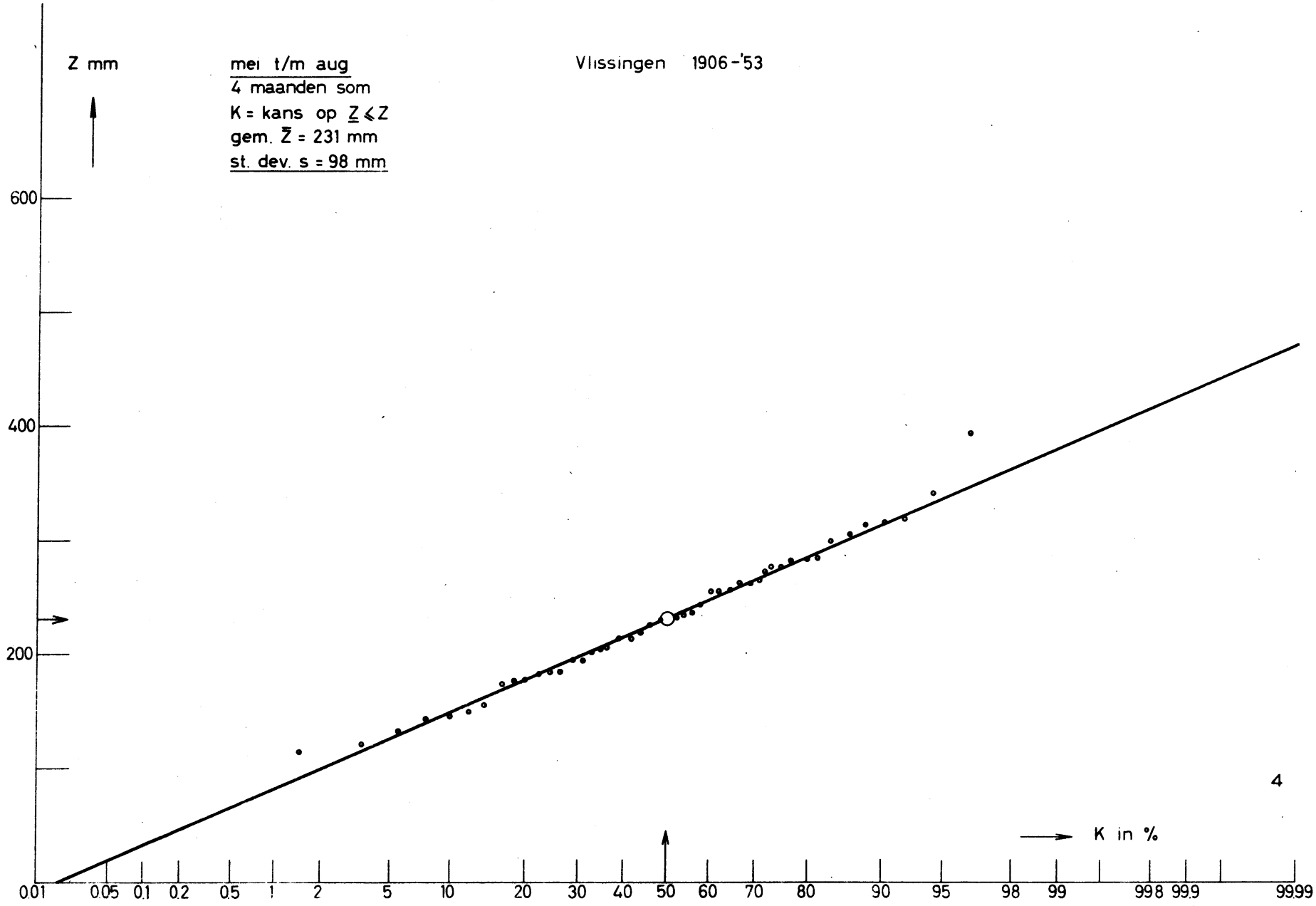
worden.

Een r-waarde gelegen boven de bovenste of beneden de onderste 5%-rechte, verschilt niet signifikant van nul, behoudens een onbetrouwbaarheid van 0.05.

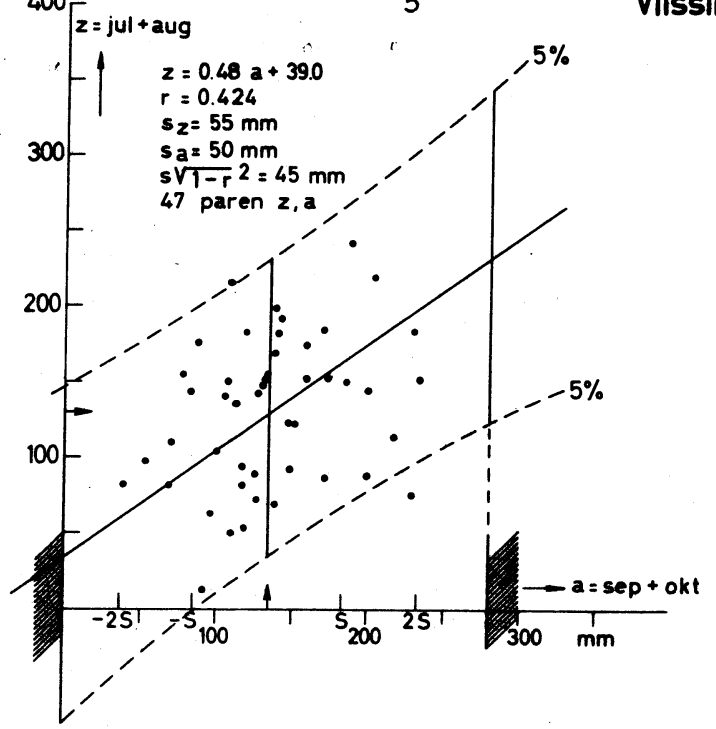

a mm
sep + okt
2 maanden som
K = kans op a ≤ a
gem. ā = 135 mm
st. dev. s = 50 (43 - 60) mm
r (z,a) = 0.42
r (Z,a) = 0.42
Vlissingen 1906 - '53
60
50
300
250
200
150
100
50
43
1
K in %
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99


z mm
juli + aug
2 maanden som
K = kans op z ≤ z
gem. z̄ = 130 mm
st. dev. s = 55 mm
Vlissingen 1906–'53
300
200
100
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
K in %
2

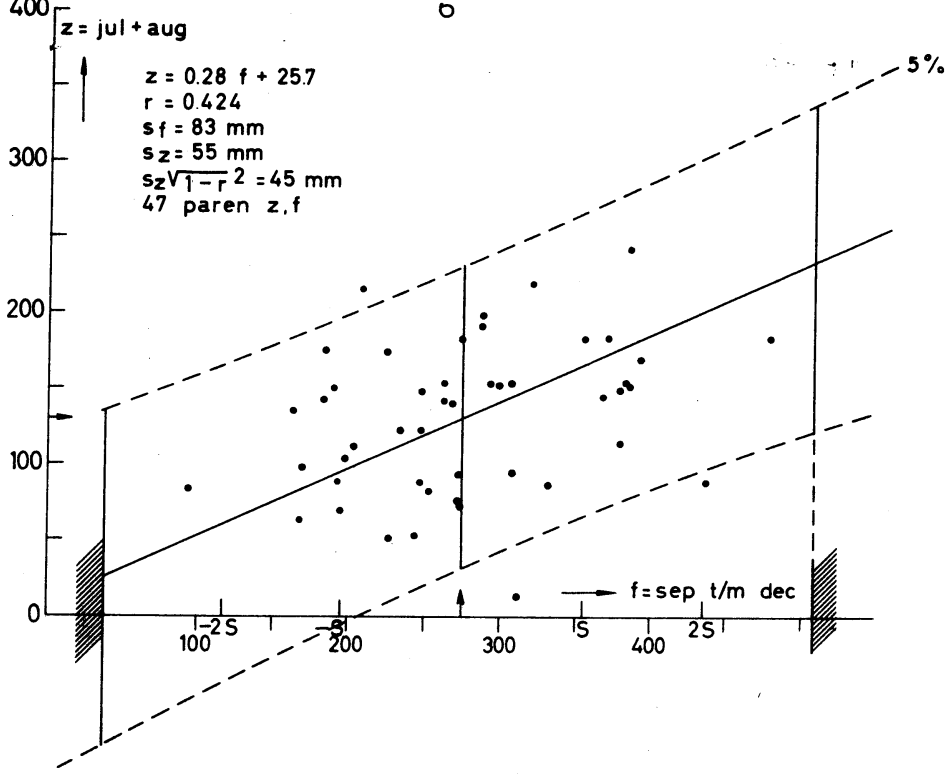

f mm
sep t/m dec
4 maanden som
K = kans op f ≤ f
gem. f̄ = 277 mm
st. dev. s. = 83 mm
r (z,f) = 0.42
r (Z,f) = 0.30
Vlissingen 1906 -'53
400
200
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
K in %

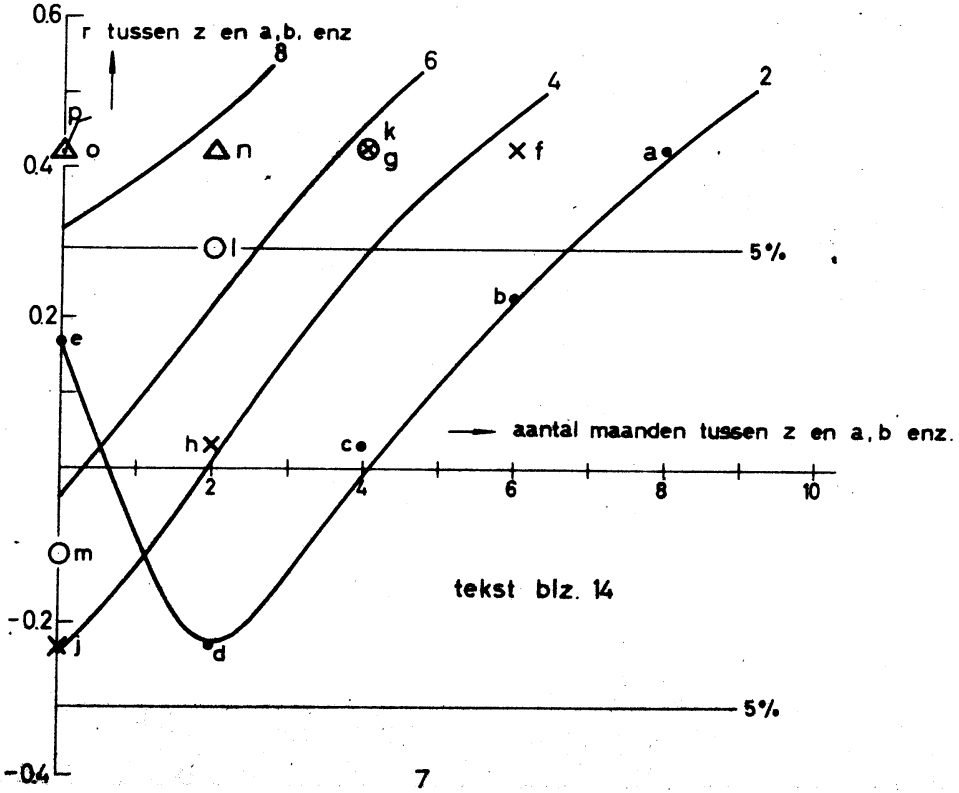

Z mm
mei t/m aug
4 maanden som
K = kans op Z ≤ Z
gem. Z̄ = 231 mm
st. dev. s = 98 mm
Vlissingen 1906-'53
600
400
200
0.01
0.05
0.1
0.2
0.5
1
2
5
10
20
30
40
50
60
70
80
90
95
98
99
99.8
99.9
99.99
K in %

5



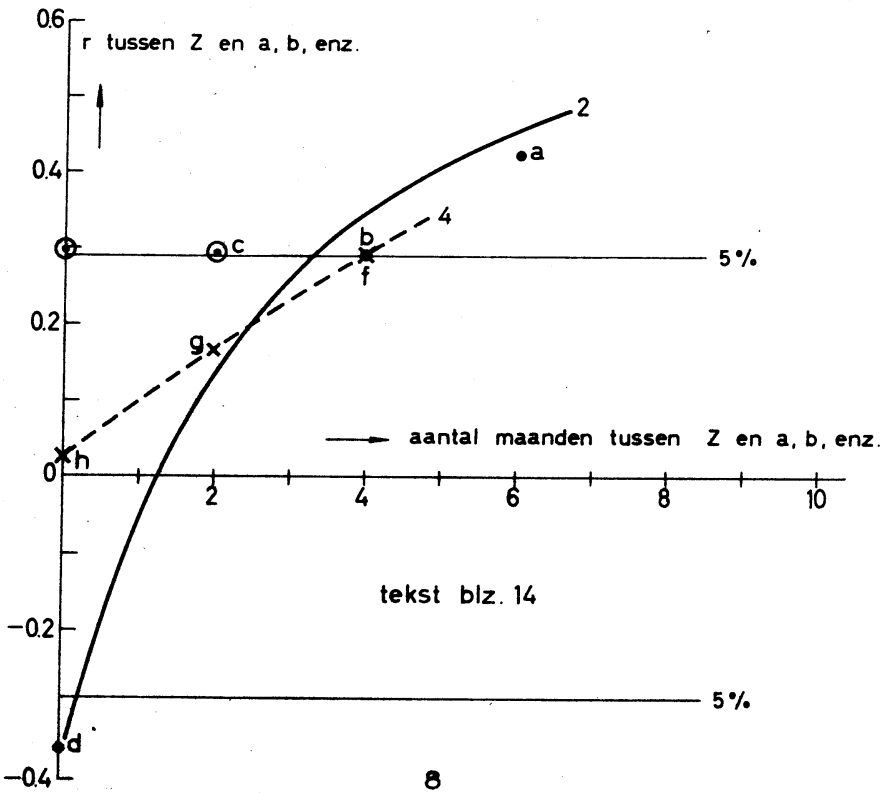
6





tekst blz. 14

7



tekst blz. 14

8