K O N I N K L IJ 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. 58-3 (III-222)

P.J. Rijkoort

Homogeniteit, betrouwbaarheid en reductie van klimatologische reeksen.

De Bilt, 1958

I N H O U D

SUMMARY.

In the first part of this report we have traced how homogeneity is defined in climatological literature. It appears that two kinds of definition exist. It is proposed to introduce two names for these two concepts, viz.:

1. reliability; if from a series of observations x_i, the errors $\Delta_i (=x_i - \xi_i$ with ξ_i is true value) have all the same distribution function. (E)
2. homogeneity; if from a series of observations x_i the values x_i have all the same distribution function. (D)

In popular words: A reliable series is a series which reproduces well the true value, except for a random error. A homogeneous series is a series in which no trend, systematic fluctuation or something else can be significantly indicated.

For testing the homogeneity a great number of methods or criteria is at one's disposal. In 1.3 and 1.4 a brief survey of these methods is given. The use of the methods is elucidated by some examples.

Testing reliability with statistical methods is impossible. Some idea on the reliability can only be obtained indirectly. For this purpose the relative homogeneity is introduced. Two series are relative homogeneous if their difference series is homogeneous. The complete definition is:

Two synchronous series $\{x_i\}$ and $\{y_i\}$ (i=1, ..., N) possess relative homogeneity if x_i as well as y_i can be transformed into quantities x_i' and y_i' by means of a series of differences $\{\Delta_i\} = x_i' - y_i'$ which is homogeneous (F).

The following working hypotheses is used: Under certain conditions two series, which possess relative homogeneity are each reliable seperately. (H)

In the second part of this report the reduction of climatological series is treated.

Formula (9) contains the general method to reduce the mean value of a quantity over a period n with the aid of the mean value from another station to the mean value over a period N.

The usefulness of the reduction is investigated. The conditions under which it is useful to reduce the mean value of a series are determined. The accuracy of the reduction is computed.

The special reduction-methods, methods of differences and ratios, enclosed in the general reduction-formula, are considered.

Some other possibilities are treated, in which symmetrical regression and "normal" regression are of importance.

In 2.2.8 the results are elucidated by examples.

Finally, the reduction of separate observations is considered.

1. HOMOGENITEIT EN BETROUWBAARHEID.

1.0 Inleiding.

Voor het bedrijven van klimatologie worden klimatologische waarnemingsreeksen gebruikt. Deze reeksen mogen geen storende onnauwkeurigheden bevatten. In verband met deze vrij vaag geformuleerde eis ziet men het begrip "homogeniteit" in de literatuur opduiken. Hoe wordt dit begrip "homogeniteit" eigenlijk gedefinieerd? In het eerste deel van dit rapport zal getracht worden deze vraag te beantwoorden en tevens na te gaan hoe de gedefinieerde begrippen getoetst kunnen worden.

1.1 Definities.

Conrad en Pollak geven op pag. 223 van Methods in Climatology [1] de volgende definitie:

"Een numerieke reeks die de variaties van een klimatologisch element voorstelt wordt "homogeen" genoemd als de variaties alleen door variaties van weer en klimaat worden veroorzaakt" (A).

Bij R. Sneyers in [2] vinden we:

"Een meetreeks is homogeen als de statistische eigenschappen van de meetfout tijdens het gehele waarnemingstijdvak onveranderd zijn gebleven" (B).

Bij Brooks en Carruthers [3] hebben we tevergeefs naar een directe definitie gezocht. Zij geven eigenlijk alleen een criterium[1]. Wij kunnen hieruit de volgende definitie afleiden:

"Een reeks is homogeen als de gemiddelden van twee willekeurige deelreeksen niet significant verschillen" (C).

Alissow, Drosdow en Rubinstein [4] geven evenmin een directe definitie. Als oorzaken voor inhomogeniteit noemen zij een echte verandering van macro of microklimaat zowel als van een wijziging van een opstelling of instrument of van waarnemings-methode. Dit komt neer op een definitie die we algemeen als volgt willen formuleren:

"Een reeks waarnemingen $x_1, \ldots\ldots, x_n$ is homogeen als de verdelingsfunctie $F_i(\underline{x}_i)$ onafhankelijk is van i" (D).

Ter verduidelijking het volgende: De streep onder x_i geeft aan dat de grootheid stochastisch d.w.z; voordat de i^{de} waarneming gedaan is, kan deze verschillende onbekende waarden gaan aannemen; dit wordt aangegeven door de verdelingsfuncties $F_i(\underline{x}_i)$. Is de waarneming verricht dan is de bekende waarde x_i aangenomen.

Definitie (D) is volkomen in overeenstemming met de definitie van het begrip"homogeen" zoals dat in de mathematische statistiek is ingevoerd. (Zie b.v. Kendall en Buckland [5]). Het aldus gedefinieerde begrip is identiek

---

[1] Onder definitie verstaan we een formulering van het wezen van het beschouwde begrip, onder criterium een voorschrift op grond waarvan besloten kan worden of aan een of andere definitie voldaan is.

met wat in de industriële statistiek een toestand van statistische controle heet (zie [6]). Ook is het eigenlijk hetzelfde als het begrip "stationair" dat b.v. bij de Fourier-analyse van ruis-verschijnselen enz. wordt gebruikt.

De definities (A) en (B) zijn niet identiek met de definities (C) en (D). Volgens de laatste twee kunnen bij klimaat veranderingen of -schommelingen geen homogene reeksen ontstaan, volgens de eerste twee definities kan dit wel.

O.i. is het begrip "homogeniteit" het best gedefinieerd door (D). We zullen deze definitie dus in het vervolg aanhouden. Het begrip dat door (A) en (B) wordt gedefinieerd zouden we dan "betrouwbaarheid" willen noemen. We zullen bij reeksen die aan (A) of (B) voldoen van "betrouwbare reeksen" spreken. We definiëren daarom:
Een reeks is betrouwbaar als van de reeks der meetfouten Δ_i de verdelingsfuncties $F_i(\Delta_i)$ onafhankelijk is van i (E).

Voor de klimatoloog zou de ideale toestand zijn dat hem "betrouwbare reeksen" ter beschikking worden gesteld. Een onderzoek naar de betrouwbaarheid zou dus buiten zijn taak moeten vallen, terwijl daarentegen een onderzoek naar de homogeniteit juist een specifieke taak voor hem is. De begrippen betrouwbaarheid en homogeniteit dekken elkaar dus niet. Een betrouwbare reeks behoeft volstrekt niet homogeen te zijn, terwijl een homogene reeks niet noodzakelijk betrouwbaar behoeft te zijn. Wel is het zo dat het laatste waarschijnlijker is dan het eerste. Immers, als de reeks onbetrouwbaar is maar toch homogeen, dan moet een verandering in waarnemingsmethode of iets dergelijks juist door een klimaatverandering worden gecompenseerd. Dit nu kunnen we stellig als een zeer onwaarschijnlijke gebeurtenis beschouwen.

1.2 Relatieve homogeniteit.

Een exacte toetsing of een reeks betrouwbaar is, is principieel onmogelijk. Slechts bij benadering valt hierover iets te zeggen. Tijdens het ontstaan van de waarnemingsreeks kan dit geschieden door regelmatige ijking van instrumenten en controle van opstelling en waarnemingsmethode. Achteraf geeft vergelijking met een parallelreeks een mogelijkheid om iets van de betrouwbaarheid te zeggen. Hiertoe is het begrip "relatieve homogeniteit" ingevoerd. Als definitie zullen we gebruiken:

Twee synchrone reeksen $\{x_i\}$ en $\{y_i\}$ $(i = 1,..,N)$ zijn relatief homogeen als zowel x_i als y_i getransformeerd kunnen worden tot grootheden x_i' en y_i' waarvan de verschilreeks $\{\Delta_i\} = \{x_i' - y_i'\}$ homogeen is. (F)

Conrad en Pollak geven een definitie, waarin slechts van de verschillen of de verhoudingen van x_i en y_i wordt geëist dat deze een homogene reeks vormen. Ongetwijfeld zijn dit in de praktijk de belangrijkste gevallen (voor de temperatuur de verschillen, voor de neerslag de verhoudingen). Het leek ons echter toch gewenst de definitie zo algemeen mogelijk op te stellen.

Bij de temperaturen zal de transformatie dus de indentiteit zijn en bij de neerslag de logarithmisering. Het spreekt verder vanzelf dat bij een onderzoek van de relatieve homogeniteit de toegepaste transformatie van te voren bekend moet zijn. Deze mag niet uit het, voor het onderzoek gebruikte materiaal worden bepaald, omdat anders de mogelijkheid bestaat, dat een eventuele inhomogeniteit wordt geëlimineerd of althans verdoezeld.

Verder gebruiken Conrad en Pollak voor het homogeniteits-criterium de eis, dat de verschillen (resp. quotienten) aan een normale verdeling moeten voldoen. Levert stelt in [7] een minder zware eis, namelijk: de verschillen moeten symmetrisch verdeeld zijn.

Het betrouwbaarheidsonderzoek is nu gebaseerd op de volgende werkhypothese:

"Twee reeksen die relatief homogeen zijn, zijn elk afzonderlijk betrouwbaar". (H)

In zijn algemeenheid is deze werkhypothese stellig geen these, evenmin als het omgekeerde. Waarnemingsreeksen van twee plaatsen die zeer ver van elkaar verwijderd zijn, kunnen betrouwbaar zijn, maar, tengevolge van klimaats-fluctuaties die niet parallel lopen in beide plaatsen, kan de verschilreeks zeer goed niet homogeen zijn. Het is echter plausibel, dat dit niet het geval zal zijn als de plaatsen dicht bij elkaar liggen. Het is dan ook vanzelfsprekend, dat men betrouwbaarheids-onderzoek door middel van vergelijking van reeksen uit twee plaatsen slechts uitvoert met plaatsen die in eenzelfde klimaatgebied liggen. Voor dit geval is de hypothese (H) als omkeerbare these te beschouwen.

1.3 Homogeniteitscriteria.

Er bestaan verschillende criteria om te toetsen of een reeks $\{x_i\}$ homogeen is. In [2] en [7] worden deze vrij uitvoerig behandeld. Bovendien bestaat er een lange reeks van artikelen van Kivéliovitch en Vialar [8] waarin ook verschillende toetsen worden ontwikkeld.

De toetsen die beschikbaar zijn, kunnen op twee manieren in groepen worden gesplitst, namelijk in die waarbij

A : $F_i(x_i)$ de normale verdeling is.

B : $F_i(x_i)$ symmetrisch is.

C : $F_i(x_i)$ onbekend is.

en die waarbij:

I : de reeks $\{x_i\}$ in 2 of meer deelreeksen wordt gesplitst.

II: de reeks $\{x_i\}$ in zijn geheel wordt behandeld.

De belangrijkste toetsen zijn:

A I : F-toets (bij 2 deelreeksen)
t-toets (" " ")
Enkelvoudige variantie-analyse (bij k deelreeksen)
Toets van Bartlett (bij k deelreeksen).

A II : Criterium van Abbe of toets van Anderson.

B II : Criterium van Helmert.

C I : Contingentie-tabel met χ^2-toets.
Toets van Wilcoxon (2 deelreeksen).
H-toets (k deelreeksen).

C II : "Covariantie"-toets van Wald en Wolfowitz.
Runtoets van Wald en Wolfowitz.
Trendtoets van Mann.
Runtoets van Kivéliovitch en Vialar.
Extremen toets van Kivéliovitch en Vialar.

Een beschrijving van deze toetsen zal hier niet gegeven worden. Hiervoor wordt verwezen naar [2], [7], [8] en [10]. Slechts voor enkele toetsen, die we op een voorbeeld willen toepassen, zullen we de werkwijze behandelen. Alvorens hiertoe over te gaan willen we ons echter eerst in het algemeen afvragen, wat we met bovenstaande toetsen kunnen bereiken.

We moeten ons realiseren, dat definitie (D) een ideaal voorstelt. De homogeniteit die hiermede wordt gedefinieerd, is een ideale homogeniteit waarvan we principieel nooit volledig kunnen toetsen of ze aanwezig is. De toetsen of criteria geven slechts ten dele antwoord op de vraag of de reeks homogeen is in de zin van D. Zo geven b.v. van de groep A I : de F-toets en de toets van Bartlett slechts aan of de spreidingen van de deelreeksen al dan niet identiek zijn. De t-toets en de variantie-analyse daarentegen slechts of de gemiddelden van de deelreeksen uit één universum komen. In het algemeen hebben trouwens de toetsen van groep I het bezwaar, dat eventuele inhomogeniteiten binnen de deelreeksen niet altijd aan het licht zullen komen.

De toetsen van groep II reageren op een geheel ander aspect van de inhomogeniteit en wel in het algemeen op inwendige eigenschappen van de reeks, b.v. persistentie. We moeten namelijk bij de definitie van homogeniteit ook ingesloten achten de eis, dat $\underline{x}_i$ onafhankelijk van $\underline{x}_j$ verdeeld is.

De toetsen ontstaan steeds zo: Er wordt uitgegaan van de onderstelling, dat de homogeniteit aanwezig is. Op grond hiervan wordt berekend hoe een bepaalde toetsingsgrootheid verdeeld is. Heeft de berekende waarde van de toetsingsgrootheid een overschrijdingskans kleiner dan 5%, dan wordt de nulhypothese (de homogeniteits-hypothese)verworpen; dan is er inderdaad duidelijk

geen homogeniteit. Is de overschrijdingskans groter dan 5%, dan zullen we, volgens de regel, de nulhypothese aanvaarden. In dit geval is de zaak minder duidelijk. Het kan zijn, dat er toch inhomogeniteit van de soort waar de toets gevoelig voor is, aanwezig is, maar dat deze te gering is om ontdekt te worden. Het kan echter ook zijn, dat er een inhomogeniteit van een andere soort aanwezig is. In de praktijk is het zo, dat we een bepaald criterium kiezen. Op grond hiervan kunnen we eventueel tot een beperkte, een "praktische" homogeniteit besluiten. Hiermede nemen we doorgaans genoegen.

Welk criterium we kiezen hangt van het probleem af. Hebben we b.v. bij een gegeven tijdreeks x_1 ..... x_N het vermoeden, dat er tussen x_n en x_{n+1} iets gebeurd is, dat mogelijk de waarden van x_i ($i = n+1, \ldots, N$) beïnvloed heeft, dan ligt het voor de hand een toets uit groep I te kiezen. Hebben we daarentegen een reeks x_1 ..... x_N waarvan we niets weten, maar waarvan we in het algemeen de homogeniteit willen onderzoeken dan is groep II te verkiezen, hoewel we hier ook wel een toets uit groep I kunnen toepassen.

1.4 Toepassing.

Als voorbeeld gebruiken we de temperatuurreeksen van de juni gemiddelden van De Bilt en Oudenbosch van 1908 - 1957. We nemen aan, dat het verband tussen de temperaturen in De Bilt (t_b) en Oudenbosch (t_o) lineair is, zodat $t_o = t_b + x$. Op grond hiervan kan dus de transformaties uit de definitie van relatieve homogeniteit de identiteit gebruikt worden. We kunnen dus de verschilreeks $\{x_i\}$ direct voor het onderzoek der relatieve homogeniteit gebruiken.

In de figuren 1, 2 en 3 zijn resp. t_o, t_b en x als functie van de tijd voorgesteld.

We zullen ter illustratie een aantal toetsen toepassen: daarbij vestigen we er nog wel de aandacht op, dat het niet geoorloofd is bij een werkelijk onderzoek een groot aantal toetsen toe te passen en dan eventueel tenslotte die te gebruiken, welke ons het best past. Men moet in de praktijk van te voren tot een bepaalde toets besluiten, tenzij men principieel verschillende nulhypothesen wil toetsen en dus de hierbij passende toetsen kiest.

1.4.1 Berekening van toetsingsresultaten.

a. Contingentietabel met χ^2-toets.

We splitsen het materiaal in 5 deelreeksen van 10 jaren n.l. 1908/1917, enz. Bovendien verdelen we de x_i waarden in twee groepen n.l. $x_i \geq 0{,}2$ en $x_i < 0{,}2$. De keuze 0,2 zorgt ervoor dat de totale aantallen in beide groepen niet te veel verschillen.

De 2 x 5 contingentie-tabel die ontstaat ziet er als volgt uit:

tabel I

	1908/1917	1918/1927	1928/1937	1938/1947	1948/1957	totaal
$x_i \geq 0.2$	7 (5.8)	7 (5.8)	9 (5.8)	5 (5.8)	1 (5.8)	29
$x_i < 0.2$	3 (4.2)	3 (4.2)	1 (4.2)	5 (4.2)	9 (4.2)	21
totaal	10	10	10	10	10	50

De aantallen die we zouden verwachten in iedere cel, als de verdeling over beide x_i groepen voor alle groepen van 10 jaren dezelfde zou zijn, vinden we tussen haakjes aangegeven, n.l. $\frac{10x29}{50} = 5.8$ resp. $\frac{10x21}{50} = 4.2$. Als f_{ij} het waargenomen en φ_{ij} het verwachtte aantal in cel i, j is (i = 1, ... 5 voor de 10-jaar groepen en j = 1,2 voor de x_i verdeling) dan wordt berekend

$$\chi^2 = \sum \frac{(f_{ij} - \varphi_{ij})^2}{\varphi_{ij}} = 14,72$$

Met (2-1) (5-1) = 4 vrijheidsgraden vinden we voor de berekende χ^2 een overschrijdingskans P $\approx$ 0.008. De nulhypothese moet dus verworpen worden. Derhalve kunnen we zeggen: F ($\underline{x}_i$) is stellig niet onafhankelijk van i. We zien trouwens aan de cijfers direct al wel, dat de verdeling in 1948/1957 geheel anders is geweest.

b. Toets van Anderson.

Voor deze toets wordt de autocorrelatie coëfficiënt $r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$ waarbij $x_{n+1} \equiv x_1$ gesteld wordt, berekend.

Onder de nulhypothese, dat alle x_i aan dezelfde normale verdeling voldoen en bovendien onderling onafhankelijk zijn, geldt, dat r bij benadering normaal verdeeld is met gemiddelde $E(r) = -\frac{1}{N-1}$ en standaard deviatie $S(r) = \frac{\sqrt{N-2}}{N-1}$

We vinden uit de waarnemingen:

r = 0,377, E(r) = -0,020 en S(r) = 0.141.

Voor $\frac{r-E(r)}{S(r)}$ vinden we dus 2,82. Dit moet onder de nulhypothese aan een (0,1) normale verdeling voldoen. De tweezijdige overschrijdingskans van 2,82 is 0,005. Derhalve moeten we ook met deze toets tot verwerping der nulhypothese besluiten.

c. Covariantie-toets van Wald-Wolfowitz.

In dit geval behoeven we niet te eisen, dat de $\underline{x}_i$'s normaal verdeeld zijn, maar slechts, dat ze onafhankelijk van i dezelfde verdeling bezitten en bovendien onderling onafhankelijk zijn.

De toetsingsgrootheid is $R = \sum_{i=1}^{N} x_i x_{i+1}$. Onder de nulhypothese is deze bij benadering normaal verdeeld met gemiddelde $E(R) = \frac{S_1^2 - S_2}{N - 1}$

en variantie

$$S^2(R) = \frac{S_2^2 - S_4}{N-1} + \frac{S_1^4 - 4S_1^2 S_2 + 4S_1 S_3 + S_2^2 - 2S_4}{(N-1)(N-2)} - E^2(R)$$

(Waarin $S_k = \sum_1^N x_i^k$)

We berekenen: $R = 4{,}34$ $E(R) = 1{,}67$ en $S(R) = 0{,}93$

Dus $\frac{R - E(R)}{S(R)} = 2{,}87$ met een tweezijdige overschrijdingskans 0,004. Ook dit geval leidt tot verwerping van Ho.

d. Criterium van Helmert.

De nulhypothese, die aan dit criterium beantwoordt, luidt (zoals door Levert in [7] aangetoond is): x_i is symmetrisch verdeeld t.o.v. $\bar{x}$.

Voor deze toets wordt een reeks + en - tekens gevormd uit de reeks $\{x_i\}$ en wel een + als $x_i > \bar{x}$ en een - als $x_i < \bar{x}$. Een opvolging + - of - + heet variatie en een opvolging -- of ++ een sequentie. Het aantal sequenties noemen we S en het aantal variaties V.

H = S - V is bij benadering normaal verdeeld met gemiddelde 0 en variantie N - 1.

We berekenen S = 31 en V = 18. Dus H = 13. Verder is $\sqrt{N-1} = \sqrt{49} = 7$. Behouden we 5% als significantiedrempel dan wordt de nulhypothese verworpen als H buiten -14 ..... +14 valt. We komen dus met dit criterium niet tot verwerping der nulhypothese. Helmert heeft zelf als grenzen $\pm\sqrt{N-1}$ gesteld, wat er dus op neerkomt, dat hij reeds de homogeniteit verwerpt bij een kans van 1 op 3. Dit is wel wat erg snel. Overigens zou dan in ons geval wel tot verwerping der homogeniteit worden besloten.

Verder kunnen we opmerken dat het feit, dat dit criterium dus op 5% niveau niet tot verwerping leidt, in tegenstelling tot de vorige toetsen, misschien veroorzaakt wordt door het feit, dat dit criterium slechts een deel van de beschikbare informatie benut, namelijk alleen het teken van $x_i - \bar{x}$; hoewel dit ook geldt t.o.v. de toetsen van Anderson en van Wald en Wolfowitz.

e. Runtoets van Kivéliovitch en Vialar.

Bij deze toets wordt uit de reeks $\{x_i\}$ ook een teken-reeks gemaakt, n.l. van + als $x_i > x_{i-1}$ en - als $x_i < x_{i-1}$ en een 0 als $x_i = x_{i-1}$. We noemen nu een reeks van k opeenvolgende gelijke tekens (- of +) een k-run. Hierbij wordt een 0 tussen twee plus-tekens als een + en tussen twee min-tekens als een - gerekend. De overige nullen blijven buiten beschouwing. Noemen we n_i het aantal runs van de lengte i en P het totaal aantal runs, dan geldt bij benadering:

$E(n_1)$	= 5/8 P	S_1	= 0,4841 P
$E(n_2)$	= 11/40 P	S_2	= 0,4465 P
$E(n_3+n_4\ldots)$	= 1/10 P	$S_{\geq 3}$	= 0,3000 P

We berekenen $P = 22$. Verder is $n_1 = 7$, $n_2 = 8$, $n_3 = 7$.
De 95% marges voor n_i zijn:

$i = 1$: $13,75 \pm 4,53$
$i = 2$: $6,05 \pm 4,19$
$i = 3$: $2,20 \pm 2,81$

We zien dus dat n_1 significant kleiner is dan we zouden verwachten en n_3 groter dan we zouden verwachten. n_2 valt binnen de 5% grenzen. Ook nu moet de nulhypothese verworpen worden.

1.4.2 Interpretatie van de toetsingsresultaten.

De vraag is thans hoe we de toetsingsresultaten moeten interpreteren. Betekent de verwerping van de nulhypothese steeds hetzelfde? Dit is zeker niet het geval. Indien de nulhypothese niet juist is, moet een andere hypothese wel juist zijn. Deze andere hypothesen, de z.g. alternatieve hypothesen, vormen, bij iedere toets, een bepaalde groep. Deze groepen zijn hier alle dezelfde.

Bij de voor a) gebruikte contingentietabel bestaan de alternatieve hypothesen hoofdzakelijk uit die hypothesen, volgens welke van een of meer der 10-jaar groepen de x_i waarden op een ander niveau liggen. Strikt genomen is alleen te zeggen, dat de alternatieve hypothesen bestaan uit die hypothesen, waarbij $P(x_i \geq 0.2)$ *) niet voor alle 10-jaar groepen gelijk is. Dit zullen in de praktijk wel vooral gevallen zijn waarbij van een niveau-verschuiving sprake is (dus verschil in mediaan of gemiddelde). Echter is dit niet de enige mogelijkheid voor de alternatieven. Het kan namelijk ook zijn, dat voor een of meer der 10-jaar groepen de spreiding van de x_i's afwijkt van die der overige, vooral als bij een verdeling van de x_i's in tweeën de grenswaarde wat ver van gemiddelde of mediaan verwijderd is.

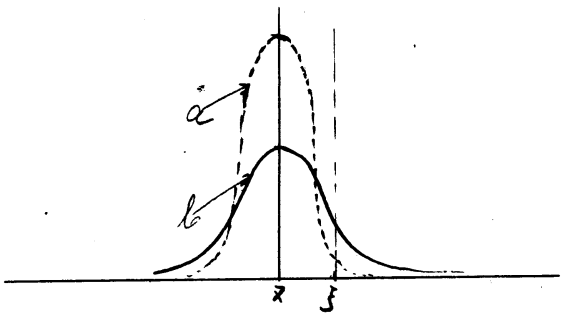


fig. 4

Zie ter verduidelijking fig. 4. Stel, dat een deel der groepen x_i volgens a verdeeld is en een ander deel volgens b, dan is het duidelijk dat $P(x_i > \xi)$ in beide gevallen geheel verschillend is, terwijl toch $\bar{x}$ dezelfde is.

De alternatieven, die bij de toetsen b) van Anderson en c) Wald en Wolfowitz behoren, hebben een enigszins ander karakter. In de eerste plaats denken we hierbij aan al of niet periodieke fluctuaties. Echter vallen hieronder in ruime betekenis ook die gevallen, waarbij een deel van de reeks naar een ander niveau verschoven is. Daarentegen zijn deze toetsen niet gevoelig voor een verandering in spreiding.

*) P = waarschijnlijkheid dat

Bij de toets van Anderson vallen onder de alternatieven ook de niet-normale verdelingen.

Bij toepassing van d) Het criterium van Helmert, omvatten de alternatieven alle gevallen waarbij de x_i's niet symmetrisch om een vast gemiddelde verdeeld zijn. Dit kunnen zowel verschuivingen zijn van een deel der groepen, als ook eenzelfde, maar scheve, verdeling voor alle x_i's.

De uitkomst van de toepassing van dit kriterium op ons voorbeeld, geeft de gelegenheid iets nader in te gaan op een ander aspect van het geval. Indien men bij het toepassen van een toets tot verwerping van de nulhypothese komt, dan kan men zeggen, dat het wel zeer waarschijnlijk is dat de nulhypothese niet juist is. Komt men echter niet tot verwerping dan kan niet gezegd worden dat het zeer waarschijnlijk is dat de nulhypothese juist is. Naast de nulhypothese kunnen n.l. nog allerlei andere hypothesen geoorloofd zijn. Niet verwerpen van de nulhypothese wil eigenlijk zeggen: er is geen reden de nulhypothese niet te aanvaarden; het is echter nooit een bewijs dat de nulhypothese de juiste is. Bij een bepaalde waarde van de toetsingsgrootheid behoort voor iedere hypothese een zekere overschrijdingskans. Alle hypothesen waarvoor de overschrijdingskans kleiner dan b.v. 5% is komen in aanmerking om geaccepteerd te worden. In de praktijk is het natuurlijk niet mogelijk voor alle hypothesen de overschrijdingskansen te berekenen. Men maakt alleen een verstandige keuze voor de nulhypothese en werkt hiermede verder. Is deze niet acceptabel dan moet een andere keuze gedaan worden. Doet men dit op grond van het beschikbare materiaal dan is toetsing met ditzelfde materiaal niet zinvol maar moet getoetst worden met nieuwe gegevens. Uiteraard is deze procedure bij homogeniteits of betrouwbaarheids-onderzoek van beschikbare reeksen niet mogelijk. De enige weg is dan op gronden los van het materiaal een verstandige alternatieve keuze te doen.

Het is nu wel duidelijk dat het feit, dat het criterium van Helmert niet tot verwerping voert, niet in strijd is met het feit, dat de andere toetsen wel tot verwerping aanleiding geven. Zoals bekend, is de opstelling in De Bilt in 1950 en 1951 veranderd. Ongetwijfeld is dit de oorzaak van de inhomogeniteit van de reeks $\{x_i\}$ zoals uit a), b) en c) blijkt. Laat ons daarom de waarden van x_i van 1950 tot 1957 veranderen zodat $\bar{x}_i$(1950-57) gelijk wordt aan $\bar{x}_i$ (1908-1949). Passen we op deze nieuwe reeks het criterium van Helmert toe, dan vinden we V = 24 en S = 25, dus H = 1. Nu is er dus evenmin sprake van verwerping.

Passen we voor dit geval de 2 x 5 contingentietabel methode toe, dan vinden we $\chi^2 = 7.36$ en V = 4, dus $P \approx 0{,}12$. De toets van Anderson geeft r = -0,04.

De tweezijdige overschrijdingskans is nu P = 0,27. Tenslotte geeft de toets van Wald en Wolfowitz R = 4,07 E(R) = 3,95 en S(R) = 0,59, waarmee P = 0,84 wordt gevonden.

We zien dus, dat de alternatieve hypothese van de verschuiving over 0,6° van het gedeelte 1950-1958 van de x_1 reeks met alle vier toetsen tot aanvaarding voert.

Bij e) de runtoets van Kivéliovitsch en Vialar zijn de alternatieven van een ander karakter. Deze toets is n.l., althans in't algemeen, niet gevoelig voor een verschuiving van een gedeelte der reeks. Vermeerderen we n.l. de waarden $x_{n+1} \ldots x_N$ met een zeker bedrag α dan blijft het teken van $x_{n+1} - x_n$ hetzelfde als $\alpha(x_{n+1} - x_n) < (x_{n+1} - x_n)^2$ is en verandert als $\alpha(x_{n+1} - x_n) > (x_{n+1} - x_n)^2$ is. In het eerste geval verandert niets aan de uitkomst van de toets. In het tweede geval kan er iets veranderen al zal dit meestal niet veel zijn. Als de tekens van, respectievelijk $(x_n - x_{n-1})$ en $(x_{n-2} - x_{n+1})$ dezelfde zijn wordt P met 2 vermeerderd en zullen n_1, n_2 of n_3 met hoogstens 3 veranderen. In ons voorbeeld geeft de verandering van $x_{1950} \ldots x_{1957}$ met 0,6 voor de uitkomst van deze runtoets nu

$n_1 = 10, \quad n_2 = 8, \quad n_3 = 7$

en voor de 95% marges:

n_1 : 15 $\pm$ 4.74

n_2 : 6,6 $\pm$ 4.36

n_3 : 2,4 $\pm$ 2.94

Ook nu blijven n_1 en n_3 nog buiten de marges, hoewel aanzienlijk minder ver dan zonder de verandering.

We hebben dus toch wel reden de nulhypothese te verwerpen. De alternatieven die hier in aanmerking komen, wijzen op een soort periodieke fluctuaties. Het aantal lange runs is groter en het aantal korte kleiner dan te verwachten was. Er zijn dus jaren waarin het verschil tussen Oudenbosch en De Bilt regelmatig toeneemt of afneemt. Het is niet onmogelijk, dat dit samenhangt met opstellings verandering enz. in Oudenbosch. Het blijkt n.l. dat de hut omstreeks 1926 een tijd lang zwart geschilderd is geweest. Verder is er in 1946 nog een kleine verplaatsing geweest in verband met de groei van de bomen in de omgeving. Deze veranderingen zijn niet van zodanige invloed geweest dat een direct duidelijk zichtbaar effect op de waarnemingsresultaten het gevolg is. Ze kunnen echter toch wel tot enige verandering in de verschilreeks aanleiding geweest zijn.

Tot slot zal nog worden nagegaan in hoeverre de temperatuur-reeksen zelf homogeen zijn:

Voor de 2 x 5 contingentie-tabellen vinden we:

De Bilt	1908/17	1918/27	1928/37	1938/47	1948/57
$> \bar{t}_B$	4	3	8	6	5
$< \bar{t}_B$	6	7	2	4	5
($\bar{t}_B$ = 16,95)					
Oudenbosch					
$> \bar{t}_o$	4	2	7	5	6
$< \bar{t}_o$	6	8	3	5	4
($\bar{t}_o$ = 16,76)					

Voor beide gevallen vinden we $\chi^2 = 11{,}84$ wat met $\nu = 4$ een overschrijdingskans P = 0,02 geeft. De afwijking ten gevolge van de verandering in opstelling te De Bilt omstreeks 1950, die aanleiding was tot de inhomogeniteit van de verschilreeks speelt hier helemaal geen rol. De significante inhomogeniteit die we nu vinden ontstaat uit een tegenstelling tussen de jaren 1930-1940 en de overige jaren zoals de figuren 1 en 2 laten zien.

Passen we daarentegen de runtoets van Kinéliowitch en Vialar toe dan vinden we P = 30.

	n_i	$E(n_i) \pm 2S(n_i)$
i = 1	20	18.75 ± 5.30
i = 2	7	8.25 ± 4.88
i ≥ 3	3	3.00 ± 3.28

Dus dit geeft geen reden om de homogeniteit B te verwerpen.

2. REDUCTIE VAN KLIMATOLOGISCHE REEKSEN.

2.0 Inleiding.

Het heeft alleen zin gemiddelden van klimatologische grootheden over veel jaren in twee plaatsen A en B te vergelijken als ze op hetzelfde tijdvak betrekking hebben. In de praktijk is aan deze eis niet altijd voldaan. Willen we b.v. van het station A het gemiddelde over N jaren vergelijken met het gemiddelde van B en zijn in B slechts n van de N jaren waargenomen dan moet het gemiddelde van B over n jaren gereduceerd worden tot het tijdvak van N jaren door vergelijking van de overeenkomstige waarnemingen a_i in A en b_i in B in de n gemeenschappelijke jaren.

Er zijn sedert lang twee methoden in gebruik n.l.:

1° De verschil-methode. Deze methode wordt toegepast als het verschil van de overeenkomstige waarden in A en B van een bepaalde meteorologische grootheid vrijwel constant is. (b.v. in het geval van de temperatuur).

2° De quotienten-methode. Hierbij zijn de quotienten van de waarden in A en B constant (b.v. bij de neerslag).

Met de terminologie van 1.2 kunnen we zeggen dat de verschil-methode wordt toegepast als de beide reeksen relatief homogeen zijn en de transformatie de identiteit is, terwijl de quotienten-methode wordt gebruikt als de reeksen relatief homogeen zijn en de transformatie de logarithmisering is.

In principe is de verschil-methode dus voldoende. Kent men de algemene vorm voor het verband tussen a en b dan kan men a en b transformeren tot a' en b' en hierop de verschil-methode toepassen.

Het is verder duidelijk dat we kunnen stellen:

"Reductie van een klimatologisch gemiddelde $\bar{b}$ over n jaren tot $\bar{b}$ over N jaren met behulp van de overeenkomstige waarnemingen a is slechts dan mogelijk als de reeksen a en b relatief homogeen zijn."

De vraag die we nu moeten stellen is: Wanneer is de reductie doelmatig? Het is n.l. niet onder alle omstandigheden zo, dat het gemiddelde over n jaren na reductie tot N jaren een betere schatting is van het ware gemiddelde over N jaren dan het gemiddelde over n jaren zelf. De oudere literatuur beantwoordt deze vraag met de eis: De variatie van de verschillen (resp. quotienten) moet kleiner zijn dan de variatie der oorspronkelijke grootheid; dus in formule $\frac{\mu(a-b)}{\mu(b)} < 1$; hierbij verstaat men onder $\mu(x)$: $\frac{1}{n}\sum_{i=1}^{N}|x_i|$ (zie b.v. [1] pg. 238). Volgens Conrad is de praktische grenswaarde voor het quotient 2/3. Latere schrijvers gebruiken als maat voor de variabiliteit niet de variatiebreedte μ maar de variantie of eenvoudig de kwadraatsom. Zie b.v. [3] pg.50, waar de eis is

$$\frac{\Sigma (a-b)^2}{\Sigma b^2} < 1$$

en de praktische grens bij 0.8 wordt gelegd.

Een vrij volledige theoretische behandeling van het probleem is te vinden in het werk van Alissow, Drosdow en Rubinstein [4] dat in een Duitse vertaling in Berlijn is uitgegeven. Met een belachelijke zelfverheerlijking wordt, wat dit onderwerp betreft, over het belangrijke werk van de Russen in vergelijking met het "primitieve" in de "kapitalistische" landen geschreven.

(Dem Problem der Reduktion der Beobachtungsreihen verschiedener Längen auf eine Periode sind in verschiedenen Ländern viele Arbeiten gewidmet, aber nur in der UdSSR wurde diese Frage mit genügender Gründlichkeit untersucht. Die Ursache dafür liegt in der immer wachsenderen Bedeutung, welche der Sowjetischen Klimatologie in der Volkswirtschaft zukommt, da bei unserem Aufbau eine möglichst vollständige Klimaberechnung erforderlich ist.)

We moeten echter toegeven, dat de behandeling der betrokken reductie in dit boek inderdaad zeer volledig is. Tot nu toe hebben we dit in de westerse literatuur zo niet kunnen vinden.

Met enige modificaties en uitbreidingen zal deze theorie in het volgende worden behandeld. De uitbreidingen hebben betrekking op de reductie van afzonderlijke maand- of jaar-waarden, terwijl in het bovenstaande alleen de reductie van "normalen" ter sprake komt. De uitbreidingen tot reductie van afzonderlijke jaarwaarden is van belang in verband met de hier en daar toegepaste methode om zeer lange klimatologische reeksen te construeren.

2.1 Definities, notaties en voorwaarden.

We zullen onderstellen dat we van drie stations A, B en C waarnemingen bezitten. De waargenomen meteorologische grootheden noemen we respectievelijk:

$$a_i,\ b_i \text{ en } c_i.$$

De index i geeft het jaar aan.

Het gemiddelde over een periode van N jaar zullen we door $\bar{a}^N$ aangeven, dus

$$\bar{a}^N = \frac{1}{N} \sum_{i=1}^{N} a_i \qquad \text{enz.}$$

We onderstellen verder, dat $\underline{a}_i$ een verdelingsfunctie F ($\underline{a}$) bezit die onafhankelijk is van i. Dus is o.a.:

$$E\,\underline{a}_i = E\,\underline{a}_j = E\,\underline{a} = \qquad (1)$$

en

$$\sigma^2(\underline{a}_i) = \sigma^2(\underline{a}_j) = \sigma_a^2 = E\,(\underline{a}_i - \alpha)^2 = E\,\underline{a}^2 - \alpha^2 \qquad (2)$$

Evenzo onderstellen we, dat b_i een verdelingsfunctie G(b) volgt en C een verdelingsfunctie H(c) met overeenkomstige eigenschappen.

In het algemeen is

$$\alpha \neq \beta \neq \gamma \qquad (3)$$

en

$$\sigma_a \neq \sigma_b \neq \sigma_c \qquad (4)$$

We onderstellen verder nog, dat de variabelen van verschillende jaren ongecorreleerd zijn, dus dat

$$\rho(a_i\, a_j) \equiv \rho_{a_i a_j} = 0 \qquad i \neq j \quad \text{enz.} \tag{5}$$

en

$$\rho_{a_i l_j} \begin{cases} = 0 & \text{voor } i \neq j \\ = \rho_{ab} \neq 0 & \text{voor } i = j \end{cases} \tag{6}$$

Tenslotte maken we omtrent het verband tussen a_i, b_i en c_i de onderstelling, dat het stochastisch lineair is, dus dat:

$$b_i = ka_i + x_i \quad \text{of} \quad \tilde{b}_i = k\tilde{a}_i + \tilde{x}_i \tag{7}$$

en $\quad c_i = la_i + y_i \quad \text{of} \quad c_i = l\tilde{a}_i + \tilde{y}_i$

of $\quad c_i = \frac{l}{k}b_i + z_i \quad \text{of} \quad c_i = \frac{l}{k}\tilde{b}_i + \tilde{z}_i$

Hierin is $\tilde{a}_i = a_i - \alpha$

2.2 Reductie van normalen.

We nemen in de eerste plaats aan, dat a_i beschikbaar is voor $i = 1, 2, \ldots, N$; b_i voor $i = p, (p+1), \ldots, (p+n-1)$ en C_i voor $i = q, (q+1), \ldots, (q+m-1)$ waarbij $1 < p < q$ en $(q+m-1) < (p+n-1) < N$. (Zie fig. 5)

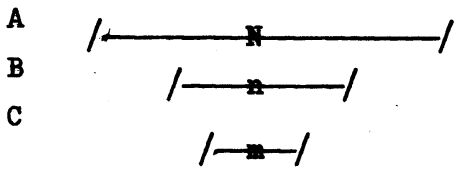


fig. 5.

2.2.1 De algemene reductieformule.

Stel nu, dat we de gemiddelde waarde ("de normaal") over N jaren in B willen weten. We beschikken in B slechts over n jaren waarnemingen en dus over $\bar{b}^n$. Deze $\bar{b}^n$ zelf kunnen we in de eerste plaats opvatten als een schatting van $\bar{b}^N$. Het is echter ook mogelijk met behulp van de waarnemingen a_i een "gereduceerde" waarde te bepalen, die we met $\bar{b}'^N$ zullen aangeven en die een betere schatting van $\bar{b}^N$ is. Uit (7) volgt n.l.

$$\bar{b}^N = k\,\bar{a}^N + \bar{x}^N \tag{8}$$

en $\quad \bar{b}^n = k\,\bar{a}^n + \bar{x}^n$

dus $\quad \bar{b}^N - \bar{b}^n = k\,(\bar{a}^N - \bar{a}^n) + \bar{x}^N - \bar{x}^n$

of $\quad \bar{b}^N = \bar{b}^n + k\,(\bar{a}^N - \bar{a}^n) + \bar{x}^N - \bar{x}^n$

Nu zijn de x_i's onderling onafhankelijk en als n en N maar voldoende groot zijn, is $\bar{x}^N - \bar{x}^n$ zeer kleine en verwaarloosbaar. We kunnen dan

$$\bar{b}'^N = \bar{b}^n + k\,(\bar{a}^N - \bar{a}^n) \tag{9}$$

als schatting van $\bar{b}^N$ gebruiken.

Onder reduceren verstaan we dus de bewerking die in (9) vervat is en waarbij $\bar{b}^n$ wordt vermeerderd met k-maal het verschil van de gemiddelden van a_i

over N en n jaren om zodoende een verbeterde schatting $(\bar{b}'^{N})$ van $\bar{b}^{N}$
In het geval k = 1 is, gaat (9) over in

$$\bar{b}'^{N} = \bar{a}^{N} + \bar{b}^{n} - \bar{a}^{n} = \bar{a}^{N} + \overline{(b - a)}^{n} \qquad (10)$$

Dit is de bekende vorm van de reductieformule voor de z.g. verschilmethode, die o.a. voor temperatuurreeksen vaak wordt toegepast. Deze gaat er dus van uit dat het verschil quasi-constant is.
Stellen we in (9) $k = \frac{\bar{b}^{n}}{\bar{a}^{n}}$, dan ontstaat:

$$\bar{b}'^{N} = \frac{\bar{b}^{n}}{\bar{a}^{n}}\ \bar{a}^{N} \qquad (11)$$

Dit is de reductieformule voor de quotientenmethode. Bij deze methode gaat men ervan uit, dat de quotienten b_i/a_i quasi-constant zijn.
Opmerking:

1° In formule (10) maakt het geen verschil of men de eerste of de tweede vorm neemt; immers altijd is $\bar{a}^{n} - \bar{b}^{n} = \overline{a - b}^{n}$. Bij de quotientenmethode maakt het wel verschil, want in 't algemeen is $\bar{b}^{n}/\bar{a}^{n} \neq \overline{b/a}^{n}$; alleen als er echte constantheid i.p.v. quasi-constantheid zou zijn, zou $\bar{b}^{n}/\bar{a}^{n} \equiv \overline{b/a}^{n}$ (en $\bar{b}^{N}/\bar{a}^{N} \equiv \overline{b/a}^{N}$ zijn.

2° In principe kan men de quotienten-methode in de verschil-methode omzetten door van de grootheden de logarithme te nemen. Hierdoor gaat het quasi-constante quotient b_i/a_i over in het quasi-constante verschil $\ln b_i - \ln a_i$. Passen we op de nieuwe grootheden $b_i^{*}(= \ln b_i)$ en $a_i^{*}(= \ln a_i)$ de verschil-methode toe, dan krijgen we een schatting van $\bar{b}^{*}$. Deze schatting kan beschouwd worden als de logarithme van een schatting van $\bar{b}^{N}$. We krijgen op deze wijze een schatting van $\bar{b}^{N}$ die niet identiek is aan die volgens (11). Beide schattingen behoren echter nog van een betrouwbaarheidsband voorzien te zijn. In het algemeen zullen de schattingen wel binnen elkaars betrouwbaarheidsinterval liggen. De bepaling van een betrouwbaarheidsinterval is bij niet normale verdelingen zeer lastig. De mogelijkheid bestaat dat door transformatie normaal verdeelde grootheden ontstaan waarvoor het bepalen van betrouwbaarheidsbanden eenvoudiger is en dan is de omweg via een transformatie dus te verkiezen.

Er zijn tenslotte voor k nog enkele andere mogelijkheden die later ter sprake zullen komen.

2.2.2 Het kriterium der doelmatigheid bij enkelvoudige reductie.

Het heeft uiteraard alleen dan zin tot reductie over te gaan als de gereduceerde waarde $\bar{b}'^{N}$ nauwkeuriger is dan de ongereduceerde $\bar{b}^{n}$. Het ligt voor de hand de standaarddeviatie van het verschil tussen schatting en $\bar{b}^{N}$ te gebruiken als maat voor de nauwkeurigheid en dus te eisen:

$$\sigma(\bar{b}'^{N} - \bar{b}^{N}) < \sigma(\bar{b}^{n} - \bar{b}^{N})$$

of

$$\sigma(\approx^{N} - \approx^{N}) < \sigma(\approx^{n} - \approx^{N}) \qquad (12)$$

We kunnen hiervoor schrijven:

$$E(\bar{b}'^N - \bar{b}^N)^2 < E(\bar{b}^n - \bar{b}^N)^2$$

We berekenen

$$(\bar{b}^n - \bar{b}^N)^2 = (\tfrac{1}{n}\sum^{n} b_i - \tfrac{1}{N}\sum^{N} b_i)^2 = \left\{(\tfrac{1}{n} - \tfrac{1}{N})\sum^{n} b_i - \tfrac{1}{N}\sum^{N-n} b_i\right\}^2 =$$
$$(\tfrac{1}{n} - \tfrac{1}{N})^2 \sum^{n} b_i^2 + (\tfrac{1}{n} - \tfrac{1}{N})^2 \sum_{i\neq j}^{n} b_i b_j + \tfrac{1}{N^2}\sum^{N-n} b_i^2 + \tfrac{1}{N^2}\sum_{i\neq j}^{N-n} b_i b_j - 2\cdot\tfrac{1}{N}(\tfrac{1}{n} - \tfrac{1}{N})\sum b_i \sum^{N-n} b_i .$$

Wegens (2) en (5):

$$E(\bar{b}^n - \bar{b}^N)^2 = n(\tfrac{1}{n} - \tfrac{1}{N})^2 E\, b_i^2 + \tfrac{1}{N^2}(N-n) E\, b_i^2 = (\tfrac{1}{n} - \tfrac{1}{N}) E\, b_i^2 = (\tfrac{1}{n} - \tfrac{1}{N})\sigma_b^2$$

Voor $E(\bar{b}'^N - \bar{b}^N)^2$ vinden we uit:

$$\bar{b}'^N - \bar{b}^N = k\bar{a}^N + \bar{b}^n - k\bar{a}^n - \bar{b}^N = \tfrac{k}{N}\sum^{N} a_i + \tfrac{1}{n}\sum^{n} b_i - \tfrac{k}{n}\sum^{n} a_i - \tfrac{1}{N}\sum^{N} b_i =$$
$$-\tfrac{1}{N}\sum^{N} x_i + \tfrac{1}{n}\sum^{n} x_i = (\tfrac{1}{n} - \tfrac{1}{N})\sum^{n} x_i - \tfrac{1}{N}\sum^{N-n} x_i .$$

dat $E(\bar{b}'^N - \bar{b}^N)^2 = (\tfrac{1}{n} - \tfrac{1}{N})^2 E(\sum^{n} x_i)^2 + \tfrac{1}{N^2} E(\sum^{N-n} x_i)^2 =$

$$n(\tfrac{1}{n} - \tfrac{1}{N})^2 \sigma_x^2 + (N-n)\tfrac{1}{N^2}\sigma_x^2 = (\tfrac{1}{n} - \tfrac{1}{N})\sigma_x^2 .$$

uit (5) volgt n.l. ook $\rho_{x_i x_j} = 0 \quad (i \neq j)$

(12) gaat dus over in $\sigma_x\sqrt{(\tfrac{1}{n} - \tfrac{1}{N})} < \sigma_b\sqrt{(\tfrac{1}{n} - \tfrac{1}{N})}$

of

$$\sigma_x < \sigma_b \qquad (13)$$

Hiermede is het kriterium der reductiedoelmatigheid teruggevonden, dat volgens [3] reeds door Hann aan het eind van de vorige eeuw werd geformuleerd. Zie ook [3] pg. 50 of [1] pg. 238. De oudere onderzoekers gebruikten echter i.p.v. de standaarddeviatie als maat voor de variabiliteit de z.g. gemiddelde afwijking ($\mu(a-b) = E|a-b|$).

Met behulp van de formule

$$\sigma_x^2 = k^2\sigma_a^2 + \sigma_b^2 - 2k\rho_{ab}\sigma_a\sigma_b \qquad (14)$$

kunnen we voor (13) ook schrijven

$$\rho_{ab} > \frac{k}{2}\frac{\sigma_a}{\sigma_b} \qquad (15)$$

2.2.3 <u>De fout van de reductie</u>.

De fout die we bij de reductie maken t.o.v. $\bar{b}^N$, is dus:

$$\sigma_{\bar{b}'^N} = \sqrt{(\tfrac{1}{n} - \tfrac{1}{N})}\,\sigma_x \qquad (16)$$

Stellen we ons op het standpunt, dat het ons niet gaat om $\bar{b}^N$ maar om de universumwaarde β waarvan $\bar{b}^N$ zelf ook een schatting is, dan kunnen we de fout t.o.v. β als volgt berekenen met de formules (7) en (9):

$$E(\bar{b}'^N)^2 = E(\tfrac{k}{N}\sum^{N} a + \tfrac{1}{n}\sum b - \tfrac{k}{n}\sum^{n} a)^2 = E\left\{\tfrac{1}{n}\sum^{n} x - \tfrac{1}{N}\sum^{N} x + \tfrac{1}{N}\sum^{N} b\right\}^2 =$$
$$E\left\{(\tfrac{1}{n} - \tfrac{1}{N})\sum^{n} x - \tfrac{1}{N}\sum^{N-n} x + \tfrac{1}{N}\sum^{N} b\right\}^2 =$$

$$\left(\tfrac{1}{n}-\tfrac{1}{N}\right)^2 E\left(\sum^{n}\tilde{x}\right)^2+\tfrac{1}{N^2}E\left(\sum^{N-n}\tilde{x}\right)^2+\tfrac{1}{N^2}E\left(\sum^{N}\tilde{b}\right)^2+\tfrac{2}{N}\left(\tfrac{1}{n}-\tfrac{1}{N}\right)E\left(\sum^{n}\tilde{x}\right)\left(\sum^{n}\tilde{b}\right)$$
$$-\tfrac{2}{N^2}E\left(\sum^{N-n}\tilde{x}\right)\left(\sum^{N-n}\tilde{b}\right)=$$
$$n\left(\tfrac{1}{n}-\tfrac{1}{N}\right)^2\sigma_x^2+\tfrac{1}{N^2}(N-n)\sigma_x^2+\tfrac{1}{N}\sigma_b^2+\tfrac{2}{N}\left(\tfrac{1}{n}-\tfrac{1}{N}\right)n\left(\sigma_b^2-k\rho_{ab}\sigma_a\sigma_b\right)$$
$$-\tfrac{2}{N^2}(N-n)\left(\sigma_b^2-k\rho_{ab}\sigma_a\sigma_b\right)$$

Met (14) vinden we tenslotte:

$$\sigma_{\bar{b}''}^N=\sqrt{\left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_x^2+\tfrac{1}{N}\sigma_b^2} \tag{17}$$

2.2.4 Tweetrapsreductie.

We gaan nu een stap verder.

In het begin van 2.2 hebben we het station C reeds genoemd. Als we nu het gemiddelde van C_i over de m-jaren willen reduceren tot het tijdvak N, dan kunnen we dit rechtstreeks op A doen, maar we kunnen ook eerst B op A reduceren (B') en dan C op B'. De vraag is nu: Onder welke omstandigheden is de reductie in twee stappen beter dan de rechtstreekse reductie?

De reductie van C op B' bestaat uit

$$\begin{aligned}\bar{c}''^N&=\bar{c}^m+\ell/k\left(\bar{b}'^N-\bar{b}^m\right)\\&=\bar{c}^m+\ell/k\left\{\bar{b}^n+k\left(\bar{a}^N-\bar{a}^n\right)-\bar{b}^m\right\}\\&=\bar{c}^m+\ell/k\left(\bar{b}^n-\bar{b}^m\right)+\ell\left(\bar{a}^N-\bar{a}^n\right)\end{aligned} \tag{18}$$

terwijl de rechtstreekse is:

$$\bar{c}'^N=\bar{c}^m+\ell\left(\bar{a}^N-\bar{a}^m\right). \tag{19}$$

Volledigheidshalve zullen we naast $\bar{c}''^N$ en $\bar{c}'^N$ als schattingen van $\bar{c}^N$ ook nog beschouwen de reductie van C op de reeks n van B n.l.: $\bar{c}'^n$ en de ongereduceerde $\bar{c}^m$.

Analoog aan de reeds eerder uitgevoerde berekeningen kunnen achtereenvolgens de berekeningen van de standaarddeviatie van de vier schattingen t.o.v. $\bar{c}^N$ worden uitgevoerd.

Resultaat:

$$\left.\begin{aligned}1)\quad &\sigma^2\left(\bar{\bar{c}}''^N-\bar{\bar{c}}^N\right)=\left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_y^2+\left(\tfrac{1}{m}-\tfrac{1}{n}\right)\sigma_z^2\\2)\quad &\sigma^2\left(\bar{\bar{c}}'-\bar{\bar{c}}^N\right)=\left(\tfrac{1}{m}-\tfrac{1}{N}\right)\sigma_y^2\\3)\quad &\sigma^2\left(\bar{\bar{c}}'^N-\bar{\bar{c}}^N\right)=\left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_c^2+\left(\tfrac{1}{m}-\tfrac{1}{n}\right)\sigma_z^2\\4)\quad &\sigma^2\left(\bar{\bar{c}}^m-\bar{\bar{c}}^N\right)=\left(\tfrac{1}{m}-\tfrac{1}{N}\right)\sigma_c^2\end{aligned}\right\} \tag{20}$$

Onderlinge vergelijking van de doelmatigheid der schattingen levert de volgende voorwaarden:

$$
\left.\begin{array}{lll}
\text{a/} & \overline{C''}^{N} \text{ beter dan } \overline{C'}^{N} & \text{als } \sigma_z^2 < \sigma_y^2 \\
\text{b/} & \overline{C''}^{N} \text{ " " } \overline{C'}^{n} & \text{" } \sigma_y^2 < \sigma_c^2 \\
\text{c/} & \overline{C''}^{N} \text{ " " } \overline{C}^{m} & \text{" } \left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_y^2+\left(\tfrac{1}{m}-\tfrac{1}{n}\right)\sigma_z^2 < \left(\tfrac{1}{m}-\tfrac{1}{N}\right)\sigma_c^2 \\
\text{d/} & \overline{C'}^{N} \text{ " " } \overline{C'}^{n} & \text{" } \left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_y^2 < \left(\tfrac{1}{n}-\tfrac{1}{N}\right)\sigma_c^2+\left(\tfrac{1}{m}-\tfrac{1}{n}\right)\sigma_z^2 \\
\text{e/} & \overline{C'}^{N} \text{ " " } \overline{C}^{m} & \text{" } \sigma_y^2 < \sigma_c^2 \\
\text{f/} & \overline{C'}^{n} \text{ " " } \overline{C}^{m} & \text{" } \sigma_z^2 < \sigma_c^2
\end{array}\right\} \quad (21)
$$

Voor de onderlinge volgorde van de drie standaard-deviaties $\sigma_y, \sigma_z, \sigma_c$ zijn zes mogelijkheden. Ieder impliceert een bepaalde volgorde in de vier schattingen n.l.:

$$
\left.\begin{array}{lll}
\text{I} & \sigma_y < \sigma_z < \sigma_c & \text{geeft } \overline{C'}^{N} \cdot \overline{C''}^{N} \cdot \overline{C'}^{n} \cdot \overline{C}^{m} \\
\text{II} & \sigma_c < \sigma_z < \sigma_y & \text{" } \overline{C}^{m} \cdot \overline{C'}^{n} \cdot \overline{C''}^{N} \cdot \overline{C'}^{N} \\
\text{III} & \sigma_z < \sigma_c < \sigma_y & \text{" } \overline{C'}^{n} \cdot (\overline{C''}^{N} \text{ of } \overline{C}^{m}) \cdot \overline{C'}^{N} \\
\text{IV} & \sigma_y < \sigma_c < \sigma_z & \text{" } \overline{C'}^{N} \cdot (\overline{C''}^{N} \text{ of } \overline{C}^{m}) \cdot \overline{C'}^{n} \\
\text{V} & \sigma_c < \sigma_y < \sigma_z & \text{" } \overline{C}^{m} \cdot (\overline{C'}^{N} \text{ of } \overline{C'}^{n}) \cdot \overline{C''}^{N} \\
\text{VI} & \sigma_z < \sigma_y < \sigma_c & \text{" } \overline{C''}^{N} \cdot (\overline{C'}^{N} \text{ of } \overline{C'}^{n}) \cdot \overline{C}^{m}
\end{array}\right\} \quad (22)
$$

De eerste van deze schattingen is dus de beste.

Voor de gevallen III t/m VI hangt de volgorde van de schattingen tussen haakjes ervan af of de verhoudingen $\frac{\sigma_c^2-\sigma_z^2}{\sigma_y^2-\sigma_c^2}$ en $\frac{\sigma_y^2-\sigma_z^2}{\sigma_c^2-\sigma_y^2}$ $>$ of $<$ $\frac{m(N-n)}{N(n-m)}$ zijn.

Uit (22) kunnen we concluderen, dat alleen VI voert tot een volgorde waarbij $\overline{\sigma''}^{N}$ de beste schatting is.

We vinden derhalve als voorwaarde voor de doelmatigheid van tweetrapsreductie:

$$\sigma_z < \sigma_y < \sigma_c \qquad (22^1)$$

Opmerking:

1° De fouten die bij de reductie gemaakt worden zijn direct met (20) gegeven. Willen we de onnauwkeurigheid niet t.o.v. $\overline{C}^{N}$ maar t.o.v. γ weten, dan moet bij ieder der fouten $\frac{1}{N}\sigma_c^2$ worden bijgeteld (analoog aan (17)).

2° Het is bij de tweetrapsreductie blijkbaar niet noodzakelijk dat de reductie $\overline{B}^{N}$ doelmatig is.

2.2.5 Reductie van B op A als extra parallel-waarnemingen in A en B beschikbaar zijn.

Het kan voorkomen, dat buiten het tijdvak N waarvoor we B op A willen reduceren nog een aantal parallel-waarnemingen gedurende p jaren beschikbaar zijn.

De vraag is nu of het zin heeft hiervan bij de reductie gebruik te maken. Willen we ervan gebruik maken, dan luidt de reductieformule:

$$\overline{b}^{*N} = \overline{b}^{n+P} + k\left(\overline{a}^{N} - \overline{a}^{n+P}\right) \tag{23}$$

We berekenen weer de standaard-deviatie t.o.v. $\overline{b}^{N}$ en vinden

$$\sigma^2\left(\overline{b}^{*N} - \overline{b}^{N}\right) = \frac{N-n+P}{N(n+P)}\,\sigma_x^2 \tag{24}$$

De reductie $\overline{b}^{*N}$ is beter dan de gewone reductie $\overline{b}^{N}$ als

dus als $$\frac{N-n+P}{N(n+P)} < \frac{1}{n} - \frac{1}{N}$$

of $$n(N-n+P) < (N-n)(n+P)$$

waaruit $$n < \tfrac{1}{2}N. \tag{25}$$

2.2.6 <u>Reductie tot een standaardperiode N als zowel A als B niet volledig zijn.</u>

Stel nu, dat zowel in A als in B geen volledige reeks voor de periode N bestaat. Maar b.v. een reeks van n_1 jaren alleen in A, n_2 jaren in A en B en n_3 jaren alleen in B terwijl $n_1 + n_2 + n_3 = N$.

Willen we nu $\overline{b}^{N}$ zowel als $\overline{a}^{N}$ schatten dan gaan we als volgt te werk. We reduceren $\overline{b}^{n_2}$ tot het tijdvak $n_1 + n_2$ met behulp van $\overline{b}^{n_1+n_2} = \overline{b}^{n_2} + k(\overline{a}^{n_1+n_2} - \overline{a}^{n_2})$

Nu middelen we $\overline{b}^{n_1+n_2}$ en $\overline{b}^{n_3}$ en krijgen dan de volgende schatting van $\overline{b}^{N}$:

$$\begin{aligned}
\overline{b}^{N} &= \frac{n_1+n_2}{N}\,\overline{b}^{n_1+n_2} + \frac{n_3}{N}\,\overline{b}^{n_3} \\
&= \frac{n_1+n_2}{N}\left\{\overline{b}^{n_2} + k\left(\overline{a}^{n_1+n_2} - \overline{a}^{n_2}\right)\right\} + \frac{n_3}{N}\,\overline{b}^{n_3} \\
&= \frac{n_1+n_2}{N}\,\overline{b}^{n_2} + \frac{n_3}{N}\,\overline{b}^{n_3} + k\frac{n_1+n_2}{N}\left(\overline{a}^{n_1+n_2} - \overline{a}^{n_2}\right) \\
&= \frac{n_1+n_2}{N}\,\overline{x}^{n_2} + \frac{n_3}{N}\,\overline{b}^{n_3} + k\frac{n_1+n_2}{N}\,\overline{a}^{n_1+n_2}
\end{aligned}$$

De fout die we t.o.v. $\overline{b}^{N}$ maken is

$$\begin{aligned}
\sigma^2\left(\overline{b}^{*N} - \overline{b}^{N}\right) &= E\left(\overline{b}^{*N} - \overline{b}^{N}\right)^2 = E\left(\overline{b}^{*N} - \overline{b}^{N}\right)^2 = \\
&E\left\{\frac{n_1+n_2}{N}\,\overline{x}^{n_2} + \frac{n_3}{N}\,\overline{b}^{n_3} + k\frac{n_1+n_2}{N}\,\overline{a}^{n_1+n_2} - \overline{b}^{N}\right\}^2 = \\
&E\left\{\frac{n_1+n_2}{N}\,\overline{x}^{n_2} - \frac{n_1}{N}\,\overline{x}^{n_1} - \frac{n_2}{N}\,\overline{x}^{n_2}\right\}^2 = E\left\{\frac{n_1}{N}\,\overline{x}^{n_2} - \frac{n_1}{N}\,\overline{x}^{n_1}\right\}^2 = \\
&\left\{\frac{n_2}{n_2^2}\left(\frac{n_1}{N}\right)^2 + \frac{n_1}{n_1^2}\left(\frac{n_1}{N}\right)^2\right\}\sigma_x^2 = \frac{n_1(n_1+n_2)}{n_2 N^2}\,\sigma_x^2
\end{aligned}$$

We kunnen geheel analoog een schatting maken van $\overline{a}^{N}$, waarbij we $b_i = ka_i + x_i$ moeten veranderen in $a_i = \frac{1}{k}b_i - \frac{1}{k}x_i = \frac{1}{k}b_i + y_i$ met $y_i = -\frac{1}{k}x_i$

we vinden dan $$\overline{a}^{*N} = \frac{n_2+n_3}{N}\,\overline{y}^{n_2} + \frac{n_1}{N}\,\overline{a}^{n_1} + \frac{1}{k}\,\frac{n_2+n_3}{N}\,\overline{b}^{n_2+n_3}$$

en

$$\sigma^2\left(\overline{a}^{*N} - \overline{a}^{N}\right) = \frac{n_2(n_2+n_3)}{n_2 N^2}\,\sigma_y^2 \tag{26}$$

2.2.7 De waarde van de regressie-coëfficient k.

We onderzoeken thans nader welke waarden k aannemen kan. In 2.2.1 hebben we reeds de mogelijkheden $k = 1$ (α) en $k = \overline{b^n}/\overline{a^n}$ (β) of $k = \left(\overline{b^n}/\overline{a^n}\right)^{\frac{1}{n}}$ (β') genoemd. Naast deze twee mogelijkheden kunnen we voor k ook kiezen: σ_i/σ_a (γ)

en $\rho_{ab}\,\sigma_b/\sigma_a$ (δ)

Volgens Drosdow is $k = \sigma_b/\sigma_a$ een generalisatie van een reductieformule volgens Wild. Door de laatste onderzoeker is echter i.p.v. de verhouding der standaard-afwijkingen de verhouding der gemiddelde absolute afwijkingen gebruikt. De regressielijn $b = \sigma_i/\sigma_a \,.\, a + x$ is de z.g. symmetrische of diagonale regressielijk van Frisch (zie o.a. [7]).
De keuze $k = \rho_{ab} = \sigma_b/\sigma_a$ voert tot de bekende regressie van b op a:

$$b = \beta_a\, a + x \qquad \text{met} \qquad \beta_a = \rho_{ab}\,\sigma_b/\sigma_a \tag{27}$$

De vraag is natuurlijk: Welke van de waarden van k moeten we in de praktijk gebruiken? Het is duidelijk dat de mogelijkheden (α) en (β) eigenlijk kunnen worden opgevat als bijzondere gevallen van (γ) en (δ). De keuze tussen (γ) en (δ) is in zekere zin een kwestie van smaak. Men kan redeneren: Het station A is alleen maar A genoemd omdat het een volledige reeks bezit. Dit heeft niets te maken met de klimatologische omstandigheden. Het a of b zijn is dus gelijkwaardig en het verband tussen a en b kan men het best symmetrisch zien. Men kan zich echter ook op het standpunt stellen, dat B tot het volledige tijdvak N gereduceerd moet worden met behulp van A en dat het dus logisch is de waarden a_i als primaire variabelen te beschouwen en b_i te zien in afhankelijkheid van a_i, wat dus tot de regressielijn (27) voert.

Een minder fraaie consequentie van het laatste standpunt treedt aan het licht in het geval van de tweetrapsreductie. In dit geval nemen we voor het verband tussen c en a:

$$c_i = \rho_{ac}\,\sigma_c/\sigma_a\, a + y_i$$

Voor het verband tussen c en b volgt hieruit:

$$c_i = \frac{\rho_{ac}}{\rho_{ab}}\frac{\sigma_c}{\sigma_b}\, b_i + z_i \qquad \text{met} \qquad z_i = y_i - \frac{\rho_{ac}\,.\,\sigma_c}{\rho_{ab}\,.\,\sigma_b}\, x_i$$

Dit is nu een ander soort regressie dan die tussen b en a en die tussen c en a. De regressie van hetzelfde type zou zijn:

$$c_i = \rho_{bc}\,\sigma_a/\sigma_b\; b_i + z_i$$

In 't algemeen is echter

$$\rho_{bc} \neq \rho_{ac}/\rho_{ab}$$

Uit praktisch oogpunt is het misschien geen bezwaar (het gaat er tenslotte om een zo nauwkeurig mogelijke schatting van $\overline{c}^N$ te krijgen), maar het is toch niet erg elegant.

Bij de keuze van de symmetrische regressielijn is de overeenstemming in soort regressie ten volle aanwezig.

Laten we nu nog even nader ingaan op het geval $k = 1$ de z.g. verschil-methode.

De motivering voor het gebruik van de verschil-methode is steeds geweest, dat het verschil veel minder varieert dan de meteorologische grootheid zelf. Conrad en Pollak spreken dan van "quasi-constantheid" van het verschil. Deze, in de praktijk waargenomen eigenschap is dus in wezen het voldoen aan de voorwaarde (13). Men kan het ook anders zeggen. Het woord "quasi-constant" zou kunnen betekenen, dat het verschil onafhankelijk is van de tijd; de tijdreeks der verschillen vertoont geen trend, geen periode fluctuatie enz.; echter is het ook mogelijk "quasi-constantheid" te zien in de betekenis: het verschil is onafhankelijk van de grootheid zelf. Op de laatste opvatting wijst wat Conrad en Pollak zeggen op pg. 226 van [1]: "There is no reason why variations of differences (or ratios) should be systematically influenced by average weather events".

Houden we ons aan de laatste opvatting, dan moet dus de correlatie-coëfficiënt van het verschil en het element zelf nul zijn. Wat moeten we onder het element zelf verstaan? Drie mogelijkheden: a, b of $\frac{1}{2}$ (a + b).

Kiezen we eerst $\frac{1}{2}$ (a + b), dus eisen we:

$$\rho\{(a-b),\ \tfrac{1}{2}(a+b)\} = 0$$

dan is

$$\frac{\Sigma(\tilde{a}-\tilde{b})\cdot\frac{1}{2}(\tilde{a}+\tilde{b})}{\sqrt{\Sigma(\tilde{a}-\tilde{b})^2\cdot\Sigma\frac{1}{4}(\tilde{a}+\tilde{b})^2}} = \frac{\Sigma\tilde{a}^2-\Sigma\tilde{b}^2}{\sqrt{\Sigma(\tilde{a}-\tilde{b})^2\cdot\Sigma(\tilde{a}+\tilde{b})^2}} = 0$$

of $\Sigma\tilde{a}^2 = \Sigma\tilde{b}^2$

of

$$\sigma_a^2 = \sigma_b^2 \qquad (28)$$

m.a.w.:

Blijkt in de praktijk, dat a - b onafhankelijk is van $\frac{1}{2}$ (a + b) en hebben we $k = 1$ gekozen, dan betekent dit, dat we gebruik hebben gemaakt van de symmetrische regressie als verband tussen a en b.

Kiezen we $\rho(a-b, a) = 0$ dan is $\Sigma\tilde{a}^2 = \Sigma\tilde{a}\tilde{b}$

of

$$\rho_{ab} = \sigma_a/\sigma_b \qquad (29)$$

d.w.z.:

Blijkt a - b onafhankelijk van a, dan is keuze van $k = 1$ equivalent met aanvaarding van de "normale regressie van b op a als verband tussen a en b.

Evenzo blijkt onafhankelijkheid van a - b en b te geven dat keuze van $k = 1$ equivalent is met aanvaarding van de regressie van a op b als verband tussen a en b.

Voor de voorwaarde (15) bij verschillende keuze van K vinden we:

$k = 1$ $\qquad \rho_{ab} > \sigma_a/2\sigma_b$

$k = \bar{b}/\bar{a}$ $\qquad \rho_{ab} > \bar{b}\sigma_a/2\bar{a}\sigma_b$

$k = \sigma_b/\sigma_a$ $\qquad \rho_{ab} > \frac{1}{2}$

$k = \rho_{ab}\sigma_b/\sigma_a$ $\qquad \rho_{ab} > \frac{1}{2}\rho_{ab}$

We zien dus, dat aan de laatste voorwaarde steeds is voldaan. De methode waarbij voor het verband tussen a en b de normale regressie van b op a gebruikt wordt is dus steeds bruikbaar. Bij $\rho < \frac{1}{2}$ is het dus, bij de keuze tussen (γ) en (δ), de enige mogelijkheid.

Welke keuze is de beste in geval $\rho > \frac{1}{2}$? We berekenen hiertoe de verhouding tussen de onnauwkeurigheid van het ongereduceerde gemiddelde tot die van het gereduceerde gemiddelde; dus $V = \dfrac{\sigma(\bar{b}^n - \bar{b}^N)}{\sigma(\bar{b}^{n'} - \bar{b}^{N'})} = \sigma_b/\sigma_x$

Uit (14) volgt: $V = (k^2\sigma_a^2/\sigma_b^2 + 1 - 2k\rho_{ab}\sigma_a/\sigma_b)^{-\frac{1}{2}}$

Geval (γ): $k = \sigma_b/\sigma_a$: $V_\gamma = (2 - 2\rho_{ab})^{-\frac{1}{2}}$

Geval ( ) : $k = \rho_{ab}\sigma_b/\sigma_a$: $V_\delta = (1 - \rho_{ab}^2)^{-\frac{1}{2}}$

Tabel 1 (overgenomen uit [4] pag. 393) geeft V voor bepaalde waarden van ρ volledigheidshalve ook voor $\rho < \frac{1}{2}$.

TABEL I

ρ	0.0	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
V_δ	1.00	1.02	1.05	1.09	1.16	1.25	1.40	1.67	2.29	3.20
V_γ	0.71	0.79	0.85	0.91	1.00	1.12	1.29	1.58	2.24	3.16

Dus is $V_\delta > V_\gamma$ voor alle waarden van ρ m.a.w. de onnauwkeurigheid in de reductie is voor bij gebruik van de normale regressie van b op a kleiner dan bij gebruik van de symmetrische regressie. Het verschil tussen beide methoden wordt echter kleiner naarmate ρ toeneemt. In de praktijk is het meestal zo, dat we reductie toepassen op reeksen van stations die vrij dicht bij elkaar liggen en de correlatie tussen deze reeksen is meestal zeer groot (b.v. 0.9). In deze gevallen is het verschil tussen beide methoden verwaarloosbaar.

We zien uit tabel 1 tevens dat, hoevel het in theorie mogelijk en doelmatig is om te reduceren, bij kleine ρ -waarden de verbetering die men bereikt zo klein is, dat deze geen praktische betekenis meer heeft.

2.2.8 Enkele opmerkingen.

2.2.8.1

1. De bovenstaande resultaten zijn strikt genomen niet helemaal correct in het geval dat $k \neq 1$ gekozen wordt. In de praktijk moet men namelijk, als men voor k een der grootheden b/a , σ_b/σ_a of $\rho_{ab}\sigma_b/\sigma_a$ wil gebruiken, voor deze grootheden, die als universumwaarden genoteerd zijn, schattingen gebruiken uit de steekproef en deze schattingen zijn zelf stochastisch. Bij de afleiding van de formules is gebruikt

$$E b_i = k E a_i + E x_i$$

Als k echter zelf stochastisch is moet men $E\, b_i = E\, k.a_i + E x_i$ gebruiken, wat niet precies hetzelfde is.
Nu blijkt het verschil in één geval (n.l. $k = \bar{b}^n/\bar{a}^n$) dat doorgerekend is, verwaarloosbaar te zijn.

We geven voor dit geval de afleiding. We gaan uit van de benaderingsformules voor gemiddelde en standaard-deviatie van een functie van stochastische variabelen, zoals o.a. te vinden is in [11].
Gegeven de stochastische variabelen

$$\underline{x}_{1i} \text{ en } \underline{x}_{2i} \quad (i = 1, \ldots, N')$$

met

$$m_j = \frac{1}{N}\sum_{i=1}^{N} x_{ji} \quad \text{en} \quad s_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ji} - m_j)^2$$

van de functie $F_i = \frac{x_{1i}}{x_{2i}}$ is

$$M = \frac{1}{N}\sum_{i=1}^{N} F_i \approx \frac{m_1}{m_2}\left[1 + \left(\frac{s_2}{m_2}\right)^2\left(1 - r\frac{m_2}{m_1}\frac{s_1}{s_2}\right)\right] \tag{30}$$

en

$$s_F^2 = \frac{1}{N-1}\sum_{i=1}^{N}(F_i - M)^2 \approx \left(\frac{m_1}{m_2}\right)^2\left[\left(\frac{s_1}{m_1}\right)^2 + \left(\frac{s_2}{m_2}\right)^2 - 2r\left(\frac{s_1}{m_1}\right)\left(\frac{s_2}{m_1}\right)\right]. \tag{31}$$

We passen dit toe op het geval $\bar{b}^N = \frac{\bar{b}^n}{\bar{a}^n}\bar{a}^N$

We bepalen eerst gemiddelde en standaard-deviatie van $\frac{\bar{b}^n}{\bar{a}^n}$. We zullen daarbij als universum van $\frac{\bar{b}^n}{\bar{a}^n}$ beschouwen alle grepen van n paren (a_i, b_i) uit het universum van N paren $\{a_i, b_i\}$. Er zijn $N^* = \binom{N}{n}$ zulke grepen.
Dus $x_{1i} = \bar{b}^n$ en $x_{2i} = \bar{a}^n$ met $i\,(1, \ldots, N^*)$
We willen de eindformules uitdrukken in de grootheden van het oorspronkelijke universum van N jaren $\{a_i, b_i\}$
We berekenen

$$m_1 = \frac{1}{N^*}\sum{}^*\bar{b}^n = \frac{n!\,(N-n)!}{N!}\cdot\frac{1}{n}\sum{}^*\sum{}^* b_i = \frac{(n-1)!\,(N-n)!}{N!}\binom{N-1}{n-1}\sum_{i}^{N} b_i = \bar{b}^N \qquad 1)$$

Evenzo $m_2 = \bar{a}^N$
Verder met $\tilde{b}_i = b_i - \bar{b}_i^N$:

$$s_1^2 = \frac{1}{N^*-1}\sum{}^*\left(\bar{b}^n - \bar{b}^N\right)^2 = \frac{n!\,(N-n)!}{\{N! - n!\,(N-n)!\}\,n^2}\sum{}^*\left(\sum{}^n \tilde{a}_i\right)^2 =$$

---

1) We schrijven Σ^* kortheidshalve voor $\sum_{i=1}^{N^*}$ enz.

$$\frac{(n-1)!\,(N-n)!}{n\{N!-n!\,(N-n)!\}}\left[\left\{\binom{N-1}{n-1}\sum^{N}\tilde{b}_i^2+\binom{N-2}{n-2}\sum_{i\neq j}^{N}\tilde{b}_i\,\tilde{b}_j\right\}=\right.$$

$$\frac{(n-1)!\,(N-n)!}{n\{N!-n!\,(N-n)!\}}\left\{\frac{(N-1)!}{(n-1)!\,(N-n)!}-\frac{(N-2)!}{(n-2)!\,(N-n)!}\right\}\sum^{N}\tilde{b}_i^2+\binom{N-2}{n-2}\sum^{N}\tilde{b}_i\sum^{N}\tilde{b}_j\Big]=$$

$$\frac{(n-1)!\,(N-n)!}{n\{N!-n!\,(N-n)!}\,\frac{(N-2)!\,(N-n)}{(n-1)!\,(N-n)!}(N-1)\,S_b^2=\frac{N-n}{n(N-n!\,\frac{(N-n)!}{(N-1)!}}\,S_b^2\approx\frac{N-n}{n\,N}S_b^2.$$

Evenzo $S_2^2\approx\frac{N-n}{n\,N}S_a^2$.

Verder is op analoge wijze aan te tonen dat

$$r_{\bar{a}^n\bar{b}^n}=r_{ab}$$

We vinden tenslotte:

$$M=\frac{\bar{b}^N}{\bar{a}^N}\left[1+\frac{N-n}{nN}\left(\frac{S_a}{\bar{a}^n}\right)^2\left(1-r_{ab}\,\frac{\bar{a}^N}{\bar{b}^N}\,\frac{S_b}{S_a}\right)\right]$$

en

$$S_F^2=\left(\frac{\bar{b}^N}{\bar{a}^N}\right)^2\left[\frac{N-n}{nN}\left(\frac{S_b}{\bar{b}^N}\right)^2+\frac{N-n}{nN}\left(\frac{S_a}{\bar{a}^N}\right)^2-2\left(\frac{S_b}{\bar{b}^N}\right)\left(\frac{S_a}{\bar{a}^N}\right)r_{ab}\cdot\frac{N-n}{nN}\right]=$$

$$\frac{N-n}{nN}\left(\frac{\bar{b}^N}{\bar{a}^N}\right)^2\left[\left(\frac{S_b}{\bar{b}^N}\right)^2+\left(\frac{S_a}{\bar{a}^N}\right)^2-2r_{ab}\,\frac{S_a\,S_b}{\bar{a}^N\,\bar{b}^N}\right].$$

Nu gaat het ons om $\bar{b}^N$ **en de fout hiervan t.o.v.** $\bar{b}^N$**, dus om** $E\,\bar{b}^N$ **en** $E(\bar{b}^N-\bar{b}^N)^2$ **. Hiervoor vinden we:**

$$E\bar{b}^N=E.\frac{\bar{b}^n}{\bar{a}^n}\,\bar{a}^N=\bar{b}^N\left[1+\frac{N-n}{nN}\left(\frac{S_a}{\bar{a}^N}\right)^2\left(1-r_{ab}\frac{\bar{a}^N}{\bar{b}^N}\,\frac{S_b}{S_a}\right)\right]$$

en

$$E(\bar{b}^N-\bar{b}^N)^2=E\left\{\left(\frac{\bar{b}^n}{\bar{a}^n}-M\right)\bar{a}^N+\left(M\bar{a}^N-\bar{b}^N\right)\right\}^2=$$

$$=\bar{a}^{N2}\,E\left(\frac{\bar{b}^n}{\bar{a}^n}-M\right)^2+2\left(M\bar{a}^N-\bar{b}^N\right)\bar{a}^N\,E\left(\frac{\bar{b}^n}{\bar{a}^n}-M\right)+\left(M\bar{a}^N-\bar{b}^N\right)^2=$$

$$=\frac{N-n}{nN}\,\bar{b}^{N2}\left\{\left[\left(\frac{S_b}{\bar{b}^N}\right)^2+\left(\frac{S_a}{\bar{a}^N}\right)^2-2\,r_{ab}\,\frac{S_a\,S_b}{\bar{a}^N\,\bar{b}^N}\right]+\left(\frac{S_a}{\bar{a}^N}\right)^4\left(1-r_{ab}\,\frac{\bar{a}^N}{\bar{b}^N}\,\frac{S_b}{S_a}\right)^2\right\}$$

(32)

Stellen we $x=\frac{\bar{b}^N}{\bar{a}^N}\,a-b.$

dan kunnen we ook schrijven

$$\sigma^2_{\bar{b}^N}\approx\left(\frac{1}{n}-\frac{1}{N}\right)\left\{S_x^2+\frac{\bar{b}^{N2}}{\bar{a}^{N2}}\left(\frac{S_a^2}{\bar{a}^N}\right)^2\left(1-r_{ab}\,\frac{\bar{a}^N\,S_b}{\bar{b}^N\,S_a}\right)^2\right\}$$

(33)

In de praktijk blijkt de tweede term klein t.o.v. S_x^2 **, zodat bij benadering geldt**

$$\sigma^2_{\bar{b}^N}=\left(\frac{1}{n}-\frac{1}{N}\right)S_x^2$$

Dit is dus dezelfde formule, als die we in het algemene geval hebben afgeleid.

2.2.8.2

Het aangeven van de standaard-deviatie als maat voor de onnauwkeurigheid suggereert dat we met een normale verdeling te maken hebben; dit is nu stellig niet het geval als we b.v. met de quotienten-methode werken. In de praktijk is

dit echter waarschijnlijk niet van belang en bij benadering kan de dubbele standaard-deviatie toch wel als maat voor de onnauwkeurigheid worden gebruikt.

2.2.9 Voorbeelden.

Het eerste voorbeeld ontlenen we aan de gegevens van de temperatuur te De Bilt en Oudenbosch n.l. 3 x daagse gemiddelden voor de maand juni over de jaren 1908 t/m 1949.

Stel dat we Oudenbosch slechts ter beschikking hebben van 1929 t/m 1949. We hebben dus

a_i = juni gemiddelde van De Bilt

b_i = " " " Oudenbosch

n = 21 en N = 42.

We berekenden

$\overline{a}^n = 17{,}48$ $\quad S_a = 1{,}147$

$\overline{b}^n = 17{,}17$ $\quad S_b = 1{,}305$

$r_{ab} = 0{,}959$

Verder is $\overline{a}^N = 16{,}94$

Voor k kunnen we 1 kiezen maar ook $\frac{1{,}305}{1{,}147} = 1{,}14$ of $0{,}959 \times \frac{1.305}{1.147} = 1{,}09$.

De bijbehorende waarden van S_x, als $x = ka - b$, zijn de volgende:

k:	S_x	en	$S_x\sqrt{\frac{1}{n} - \frac{1}{N}}$
1,00	0,384		0,059
1,14	0,376		0,058
1,09	0,372		0,057

Aan de eis (13) voor de doelmatigheid der reductie is voldaan.

Tenslotte berekenden we voor de 3 gevallen de gereduceerde waarde $\overline{b'}^N$. In de volgende tabel zijn deze aangegeven met de dubbele standaard-deviatie waarden.

k=	$\overline{b'}^N$
1	$16{,}63 \pm 0{,}12$
1.14	$16{,}56 \pm 0{,}12$
1,09	$16{,}58 \pm 0{,}11$

De ware waarde van van $\overline{b}^N$ kennen we in dit geval. Deze is n.l. 16,65. Zoals men ziet ligt deze in alle gevallen binnen de marge van 95%.

We vergelijken tenslotte nog de ongereduceerde schatting $\overline{b}^n = 17{,}17$ met de ware waarde $\overline{b}^N = 16{,}65$ waarbij we voor de onnauwkeurigheid van $\overline{b}^N$ moeten gebruiken $\sigma_b\sqrt{\frac{1}{n} - \frac{1}{N}} = 0{,}204$; dus de schatting is $17{,}17 \pm 0{,}41$. Het blijkt dus, dat in dit geval de ware waarde $\overline{b}^N$ niet binnen de marge ligt.

De mogelijkheid bestaat dat de reeksen niet voldoende homogeen zijn en misschien is dit de oorzaak van de discrepantie.

Het tweede voorbeeld ontlenen we eveneens aan de juni-temperaturen van De Bilt en Oudenbosch. We passen nu de mogelijkheid van 2.2.5 toe. We willen het gemiddelde van De Bilt voor de jaren 1951-1957 reduceren tot de normaal der jaren 1921-1950 met behulp van Oudenbosch. We hebben dus N = 30, n = 0 en p = 7.

Verder is $\overline{a}^{N} = 16,82$ $\overline{b}^{n+h} = 16,99$ $\overline{a}^{n+h} = 16,63$

$S_x^2 = 0,052$

Formule (24) geeft:

$$S^2\left(\overline{a^*}^{N} - \overline{a}^{N}\right) = \frac{N-n+r}{N.(h+n)} S_x^2 = \frac{37}{210} \times 0,052$$

dus $S_{\overline{a^*}^{N}} = 0,101$

De gereduceerde waarde met onbetrouwbaarheidsmarge wordt nu

$$\overline{a^*}^{N} = 16,46 \pm 0,20$$

Berekenen we $\overline{a}^{N}$ zelf uit de gegevens dan vinden we hiervoor 17,11. Dit ligt ver buiten de marge van de schatting. We kunnen echter in dit geval er zeker van zijn, dat de gereduceerde waarde beter is dan de berekende $\overline{a}^{N}$ zelf, gezien het feit dat we weten dat omstreeks 1951 de opstelling in De Bilt geheel is veranderd (zie ook het voorbeeld in 1.4).

Als derde voorbeeld passen we de quotienten-methode toe op jaarsommen van de neerslag te Groningen en Maastricht van 1853 t/m 1953.

We veronderstellen, dat we de neerslag te Maastricht slechts van 1894 t/m 1953 ter beschikking hebben.

We hebben in dit geval

a_i = jaarsom neerslag te Groningen

b_i = " " te Maastricht

N = 101 n = 60.

Berekend werd

$\overline{a}^{n} = 737,8$ $\overline{b}^{n} = 624,7$

$S_a = 100,6$ $S_b = 116,0$

$r_{ab} = 0,548$

verder is $\overline{a}^{N} = 709,3$

We reduceren nu met (11) en gebruiken (16) om de onnauwkeurigheidsmarge te bepalen. Hiervoor is eerst S_x berekend, waarbij $x = \frac{624,7}{737,8} a - b$ is.

Resultaat $S_x = 99,2$.

De schatting wordt nu $\overline{b}^{N} = 600,6 \pm 16,4$

De waarde $\overline{b}^{N}$ is in werkelijkheid 610,1. Dit ligt weer volkomen bevredigend binnen de marge.

Gebruiken we deze gegevens in formule (33) dan blijkt de tweede term tussen de accoladen 10 te bedragen. Dit is volkomen verwaarloosbaar t.o.v. S_x^2 die ongeveer 10.000 is.

2.3 Reductie van afzonderlijke maandwaarden.

Bij het construeren van zeer lange klimatologische reeksen wordt gebruik gemaakt van reductie van afzonderlijke maandwaarden.

Stel dat in B waarnemingen b_i over de jaren 1 ..... N beschikbaar zijn en in A waarnemingen a_i over de jaren N - n + 1, ... N, N + 1, ... N + p. We willen b_i schatten voor elk der jaren N + 1, N + p.

Als schatting gebruiken we in analogie met (9)

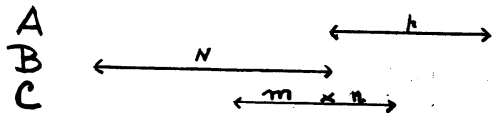
$$b_i' = \bar{b}^n + k(a_i - \bar{a}^n)$$

We verwaarlozen dus $x_i - \bar{x}^n$. De fout die we hierdoor maken is per definitie:

$$\sigma^2(a_i - a_i') = E\left(b_i - \bar{b}^n - k a_i + k\bar{a}^n\right)^2 = E\left(x_i - \bar{x}^n\right)^2 = \left(1 + \frac{1}{n}\right)\sigma_x^2.$$

Nadat de b-reeks op deze wijze is voortgezet tot i = N + p kan met behulp van een eventueel aanwezig derde station C, dat b.v. tot i = N+p+q beschikbaar is, op overeenkomstige wijze de b reeks tot i = N + p + q worden voortgezet waarbij de foutenvariantie $\left(1 + \frac{1}{m}\right)\sigma_y^2$ zal zijn als m het aantal parallel-waarnemingen tussen c en a' (resp. a) is en $y_i = a_i' - c$. Aangezien y dus (althans gedeeltelijk) op gereduceerde a' waarden berust is het duidelijk dat S_y^2 in het algemeen groter dan S_x^2 zal zijn.

Stel nu, dat A en B geen parallel-waarnemingen hebben (dus n = 0). Men kan dan toch b_i voor i = N + 1,....,N + p schatten, als er een derde station C beschikbaar is, waar zowel met A als met B gedurende enige jaren gelijktijdige waarnemingen zijn verricht. Laten we aannemen, dat C en A n jaren en C en B m jaren parallel-waarnemingen bezitten. Als schatting van a_i nemen we (voor het geval k = 1): $b_i' = a_i - (\bar{a}^n - \bar{c}^n) - (\bar{c}^m - \bar{b}^m)$



De fout in b_i' is nu

$$\sigma^2(b_i' - b_i) = E\left\{b_i - a_i - (\bar{a}^n - \bar{c}^n)(\bar{c}^m - \bar{b}^m)\right\}^2.$$

met x = b - a en y = b - c

wordt dit

$$\left(1 - \frac{1}{n}\right)\sigma_x^2 + \left(\frac{1}{n} + \frac{1}{m}\right)\sigma_y^2$$

Helaas is het in dit geval onmogelijk van σ_x^2 een schatting te berekenen. Hoogstens zou men een idee van σ_x kunnen krijgen als men twee stations ter beschikking heeft, die in overeenkomstige klimatologische omstandigheden

verkeren, op gelijke afstand van elkaar liggen als A en B. Bevindt B zich op ongeveer gelijke afstand van A als van C dan zou men $\sigma_x - \sigma_y$ kunnen stellen.

Voorbeeld:

Gegeven:

a_i =	juni temperatuur	gemiddelde	De Bilt (1931-1950)
b_i =	"	"	Winterswijk (1908-1930)
C_i =	"	"	Oudenbosch (1921-1940)

Gevraagd: schat a_i voor 1908-1930.

We berekenen:

$$\bar{b}^{21-30} = 16{,}28 \qquad \bar{c}^{21-30} = 15{,}87$$
$$\bar{a}^{31-40} = 17{,}82 \qquad \bar{c}^{31-40} = 17{,}49$$

Dus $a_i' = b_i - (\bar{b}^{21-30} - \bar{c}^{21-30}) - (\bar{c}^{31-40} - \bar{a}^{31-40})$

$$= b_i - 0{,}41 + 0{,}33 \approx b_i - 0{,}1$$

Verder werd berekend $S_y^2 = 0{,}13$.

Stellen we ook in dit geval $\sigma_x = \sigma_y$ dan is de standaard fout van de schattingen $\sqrt{0{,}14} = 0{,}38$

Dus worden de schattingen met onbetrouwbaarheids marge $b_i - 0{,}1 \pm 0{,}8$.

We hebben in werkelijkheid in dit geval wel parallel-waarnemingen a en b beschikbaar. Berekenen we hieruit S_x^2 dan vinden we 0,12. Hierbij blijkt dus inderdaad, dat $S_x = S_y$ vrijwel juist is.

3. SAMENVATTING.

In het eerste deel van dit rapport wordt nagegaan hoe men in de klimatologische literatuur het begrip homogeniteit definieert. Het blijkt dat er twee soorten definities zijn. Voorgesteld wordt nu twee benamingen in te voeren, te weten ·

1. "betrouwbaarheid" voor een reeks waarnemingen x_i, waarvan de meetfout Δ_i ($=x_i - \xi_i$, met ξ_i is ware waarde) een verdelings functie bezit die onafhankelijk is van i. (E)
2. "homogeniteit" voor een reeks x_i waarvan x_i zelf een verdelingsfunctie bezit die onafhankelijk is van i. (D)

Populair gezegd: Een betrouwbare reeks is een reeks die de ware grootheid, behoudens een toevallige meetfout goed weergeeft. Een homogene reeks is een reeks waarin geen trend, systematische schommeling of iets dergelijks duidelijk aanwezig is.

Voor het toetsen van de homogeniteit is een groot aantal methoden, kriteria of toetsen beschikbaar. In 1.3 en 1.4 wordt hiervan iets behandeld en met voorbeelden toegelicht.

Toetsing van de betrouwbaarheid langs statistische weg is niet mogelijk. Indirect is over de betrouwbaarheid wel iets te zeggen. Hiervoor wordt het begrip relatieve homogeniteit ingevoerd. Twee reeksen zijn relatief homogeen als hun verschilreeks homogeen is. In de volledige definitie (F) wordt dit begrip nog gegeneraliseerd door toe te laten, dat de reeksen eerst op een of andere wijze worden getransformeerd.

Als werkhypothese (H) is te gebruiken, dat onder zekere omstandigheden twee reeksen die relatief homogeen zijn ieder afzonderlijk betrouwbaar zijn.

In het tweede deel van dit rapport wordt de reductie van klimatologische reeksen behandeld.

Formule (9) bevat de algemene methode om het gemiddelde van een grootheid over een kort tijdvak n met behulp van het gemiddelde van een ander station te reduceren tot het gemiddelde over een tijdvak N.

Nagegaan wordt wanneer het zin heeft tot reductie over te gaan; wanneer reductie doelmatig is. De onnauwkeurigheid van de reductie methode wordt bepaald. De bijzondere reductie-methoden, zoals verschil- en quotienten-methode, die in de algemene reductie-formule zijn opgesloten, worden ter sprake gebracht. Tevens worden enkele andere mogelijkheden, waarbij de symmetrische regresse en de "normale" regressie een rol spelen, behandeld en onderling vergeleken. Enkele voorbeelden in 22.8 lichten de resultaten toe.

Tot slot wordt de reductie van afzonderlijke waarnemingen in 't kort ter sprake gebracht.

LITERATUUR.

[1] V. Conrad en L.W. Pollak: Methods in Climatology. Cambridge-Massachusetts. 1950.

[2] R. Sneyers: Sur la détermination de l'homogénéité des séries climatologiques: Inst.Royal Met. de Belgique. Contributions no. 34. 1957.

[3] C.E.P. Brooks and N. Carruthers: Handbook of Statistical methods in Meteorology. London 1953.

[4] B.P. Alissow, O.A. Drosdow, E.S. Rubinstein: Lehrbuch der Klimatologie. Berlin 1956.

[5] H.G. Kendall and W.R. Buckland: A dictionary of statistical Terms. London 1957.

[6] A. Hald: Statistical theory with engeneering applications. New York 1952.

[7] C. Levert: Some theoretical considerations with regard to the homogeneity-criteria of Helmert and Abbe. Arch. für Met.Geoph. and Biokl. 8-2. 1957.

[8] M. Kivéliovitch et J. Vialar: Étude statistique des séries chronologiques. Journal Scient. de la Met. V 17,18,19,20; VI 21,23,24; VII 27.

[9] C. Levert: Betrekkingen tussen de regressie coëfficienten van Galton en Frisch bij lineaire transformatie. Statistica 5 - 33 - 1951.

[10] P.J. Rijkoort: Statistische toetsingsmethoden. K.N.M.I. R III 120 - 1953.

[11] C. Levert: Functies van variabelen waartussen correlaties bestaan. K.N.M.I. R III 153 - 1955.

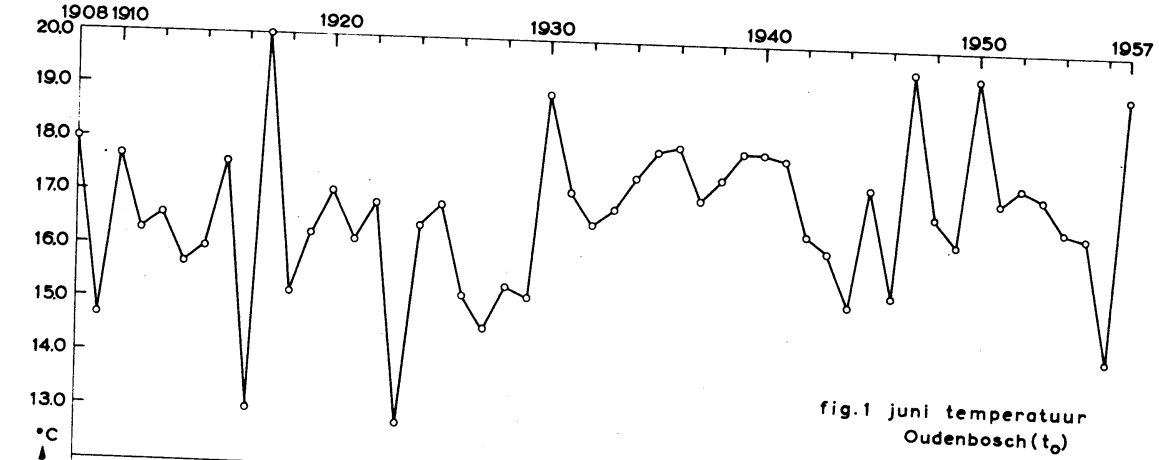


fig.1 juni temperatuur Oudenbosch (t_o)

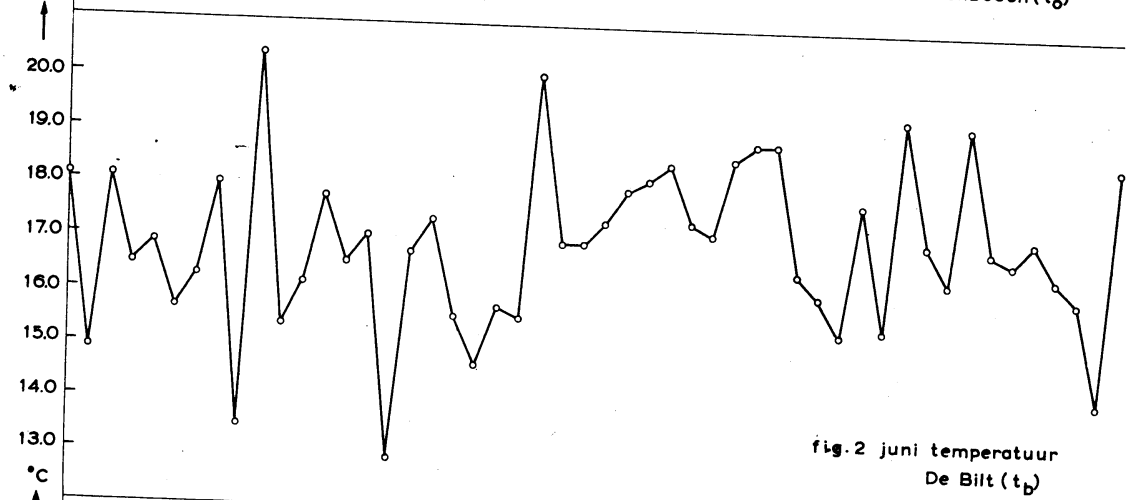


fig.2 juni temperatuur De Bilt (t_b)

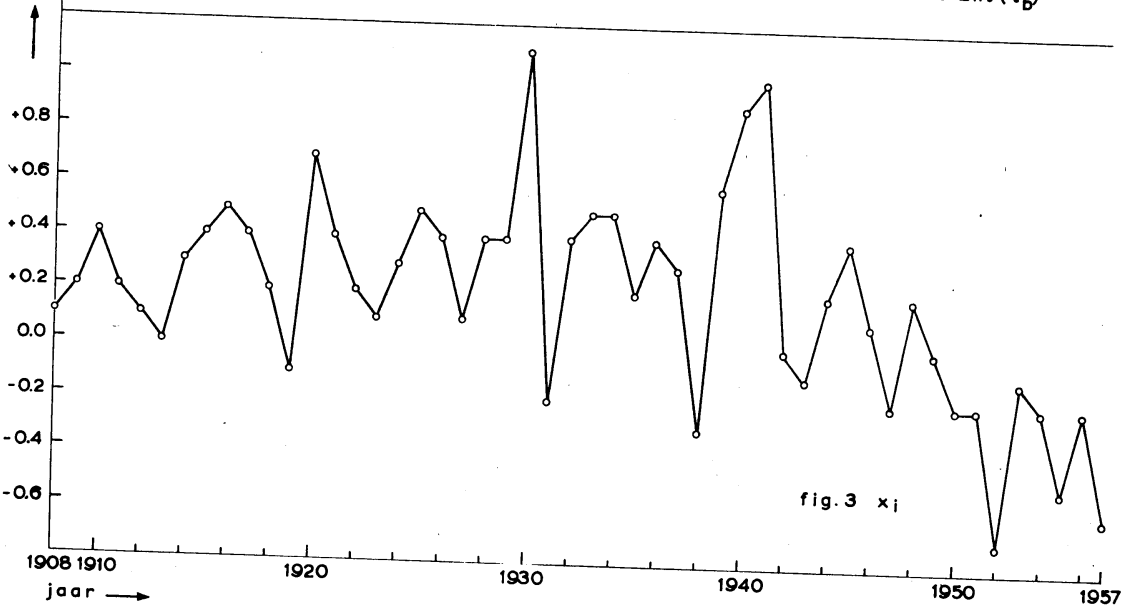


fig.3 x_i