Systematiek voor de controle van meerjarige reeksen van kustmetingen

19 Juni 2003 (Concept)

in opdracht van RIKZ

Model!T
SOFTWARE SOLUTIONS

Modelit
Rotterdamse Rijweg 126
3042 AS Rotterdam
www.modelit.nl

tel. +31 10 4623 621
fax. +31 10 4623 637

Auteur: N.J. van der Zijpp
zijpp@modelit.nl

Inhoud

1 Inleiding

1.1 Inleiding

De huidige controle functies in de Morfologie Applicaties zijn vooral gericht op het controleren en bewerken van de jaarlijkse kustmetingen en vaklodingen. In principe kunnen met deze functies ook de reeksen uit het verleden worden gecontroleerd, alleen zou gezien het groot aantal te controleren jaargangen en het grote aantal locaties een dergelijke controle een groot aantal manuren vergen, omdat dit neerkomt om het uitvoeren van een visuele controle op alle reeksen. Daarom wordt er door RIKZ onderzocht of het mogelijk is om deze controle op een meer efficiënte wijze uit te voeren. Aan Modelit is de volgende opdracht verstrekt:

1.2 Opdracht

- Het ontwerpen van een systematiek voor het in een beperkt tijdbestek controleren van kustmetingen die over een periode van enkele decennia zijn verzameld;
- Het ontwerpen van een in de Morfologie Applicatie te integreren software module waarmee deze controle kan worden uitgevoerd, vergezeld van een inschatting van de benodigde hoeveelheid tijd om deze module te implementeren.

1.3 Uitgangspunten en aannames

- Er wordt aangenomen dat de te controleren reeksen betrekking hebben op geschematiseerde c.q. sterk uitgedunde reeksen. Hierdoor bestaan deze reeksen in verhouding tot reeksen van ruwe data uit een klein aantal datapunten, en kan een zeer groot aantal van deze reeksen in een werkgebied worden geladen;
- Deze reeksen kunnen eventueel fouten bevatten, welke bijvoorbeeld kunnen samenhangen met een van de volgende oorzaken:
    - Administratieve fouten, zoals het koppelen van een foutieve locatiecode aan een gemeten reeks;
    - Foutieve verwerking van data, in het bijzonder:
        - het uitvoeren van een verkeerde conversie, bijvoorbeeld centimeters worden millimeters;
        - nulpuntverschuiving;
        - foutieve verwerking van de waterstandsreductie
    - Niet waargenomen geulen bij hoogtemetingen als gevolg van water dat achterblijft;
    - Defecten of foutieve instelling van de meetapparatuur;
    - Ongewenste manipulatie van data.
- Er wordt aangenomen dat de reeksen op de volgende wijze kunnen worden gekarakteriseerd:
    - Op basis van locatie;
    - Op basis van type, waarbij onderscheid mogelijk is tussen diepte, hoogte en eventueel gekoppelde reeksen.
    - Op basis van jaartal, waarbij wordt aangenomen dat per locatie, reekstype en jaartal steeds hooguit één reeks aanwezig is.

1.4 Validatiemethodiek

De methode van validatie binnen de Morfologie Applicatie bestaat uit de volgende stappen:

- Het berekenen van voorspelde waarden en de detectie van outliers;
- De presentatie van outliers en hiaten;
- Het beoordelen van de data, hetgeen kan inhouden dat de data wordt geaccepteerd, dat de voorspelde waarde wordt gebruikt, of dat data worden aangemerkt als hiaat.

Bij deze methodiek zal worden aangesloten. Het grote aantal te controleren reeksen stelt ons daarbij voor het probleem om de data overzichtelijk te presenteren, maar biedt ook de mogelijkheid om verbanden te leggen tussen verschillende jaargangen van de data. Dit beidt extra mogelijkheden om outliers te identificeren.

1.5 Inhoud van deze rapportage

Allereerst (hoofdstuk 2) wordt ingegaan op de mogelijkheden die bestaan om de data te controleren, zonder nog in detail in te gaan op de berekeningswijze van voorspelde waarden. Bij deze controle speelt het begrip consistentie een belangrijke rol. Consistentie bestaat zowel op puntniveau als op reeksniveau. Vervolgens wordt geïnventariseerd hoe outliers aan de gebruiker kunnen worden gepresenteerd (hoofdstuk 3). Daarna wordt een voorstel gedaan voor de berekening van voorspelde waarden en outliercriteria (hoofdstuk 4). Tenslotte wordt een indicatieve planning gegeven voor de realisatie van de in dit rapport beschreven validatie methodiek (hoofdstuk 5).

2 Toe te passen controle mechanismen

2.1 Controle op consistentie

Het voornaamste middel om reeksen te controleren is het uitvoeren van een vergelijking tussen de te controleren reeksen (of individuele datapunten) en verwante reeksen (of individuele datapunten), de zogenaamde consistentie check. Het uitgangspunt daarbij is dat het merendeel van de gegevens klopt, zodat de afwijkende reeksen of datapunten gemakkelijk te herkennen zijn, omdat zij ten opzichte van meerdere reeksen afwijken.

Voor wat betreft de consistentiecheck zijn de volgende vergelijkingen zinnig:

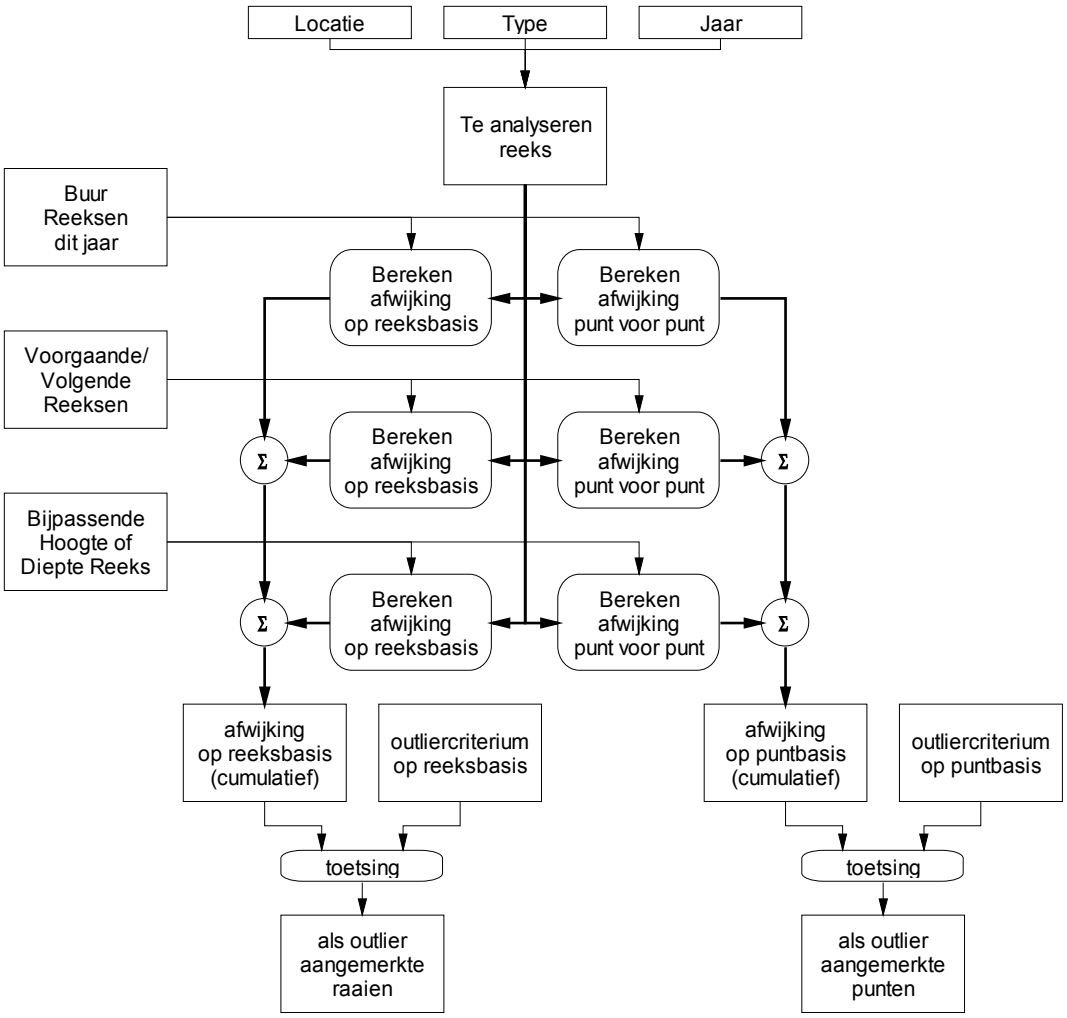
- Een vergelijking van meetgegevens van een bepaald type (hoogte, diepte of gekoppeld) van opeenvolgende jaren voor dezelfde locatie;
- Een vergelijking van meetgegevens van een bepaald type met buurlocaties binnen hetzelfde opnamejaar;
- Een vergelijking van meetgegevens van verschillende types, zoals diepte en hoogte reeksen, voor een bepaalde locatie en een bepaald opname jaar.

De consistentiecheck kan zowel op het niveau van *individuele datapunten* als op het niveau van de *gehele reeks* worden uitgevoerd, afhankelijk van het type fout dat men probeert op te sporen.

Fouten die de gehele reeks betreffen, zoals een verkeerd toegepaste waterstands reductie zijn het duidelijkst aan te wijzen als de consistentiecheck op de hele reeks wordt toegepast: de afwijkingen die optreden zijn misschien klein, maar het gelijktijdig optreden van vele kleine afwijkingen is een duidelijke indicator dat er iets niet klopt.

Het opsporen van afwijkingen van individuele datapunten gebeurt eveneens via een consistentiecheck, maar dan toegepast op het niveau van individuele datapunten. Wanneer er voor één datapunt een grote afwijking optreedt, dan is dit misschien niet genoeg om alle datapunten af te punten in de reeks in één keer af te keuren, maar wel om dit ene datapunt als outlier aan te merken.

In feite moeten dus gelijktijdig een aantal checks worden uitgevoerd, waarbij elke check een aparte drempelwaarde heeft.



Figuur 1: *De consistentiecheck vindt zowel op het niveau van individuele datapunten als op het niveau van de gehele reeks plaats. Voor de maximaal toegestane gemiddelde absolute afwijking tussen reeksen geldt een ander (strenger) criterium dan voor het maximale absolute verschil tussen individuele datapunten.*

2.2 Detectie van niet gemeten geulen in hoogtemetingen

Het ontbreken van afwijkingen in datapunten voor een locatie tussen opeenvolgende jaren kan ook een aanwijzing zijn voor het optreden van fouten. In dit geval hebben we wellicht te maken met een geul waarin water blijft staan en die niet correct door de hoogtemetingen is verwerkt. In dit geval moet dus juist bij het ontbreken van een bepaalde levendigheid van het signaal alarm worden geslagen. Praktijk tests zullen moeten uitwijzen of dit een zinnige manier is om dergelijke fouten op te sporen. Wanneer dit het geval is dan kunnen verdachte gevallen als onderdeel van een meer uitgebreid overzicht worden gepresenteerd (zie sectie 3.1.1).

2.3 Overige controles

Naast de bovengenoemde controles kunnen de reeksen ook nog op fysische criteria worden beoordeeld, zoals maximum diepte, maximum helling en maximum hoogte, maar het is de vraag of deze controle in de praktijk veel toe kan voegen aan de eerstgenoemde controle. Vooralsnog is de noodzaak hiertoe echter niet gebleken.

2.4 Controle op hiaten

Naast de consistentiecheck dient ook een controle op het aantal hiaten te worden uitgevoerd. Het aantal optredende hiaten, alsmede het aantal optredende outliers, inclusief de mogelijkheid daarop te sorteren, maakt ook in de huidige situatie al deel uit van de alfanumerieke lijst waarin de raaigegevens worden weergegeven. Het aantal hiaten hoort zou ook in het uitgebreid overzicht zoals beschreven in sectie 3.1.1moeten worden gepresenteerd..

2.5 Bijhouden reeksstatus

Op dit moment wordt binnen Morfologie Applicatie alleen de status van individuele datapunten bijgehouden. Een dergelijke status zou na beoordeling bijvoorbeeld ‘outlier’ kunnen zijn. Na het accepteren of verwerpen van outliers, verminderd het aantal outliers, hetgeen zijn neerslag heeft in diverse alfanumerieke en grafische overzichten.

Als nieuwe element is in deze sectie aangemerkt dat reeksen als zijn geheel als outlier kunnen worden aangemerkt. Het ligt dan ook voor de hand dat deze reeksgebonden outlier status na inspectie kan worden verwijderd, waardoor in de diverse overzichten de reeks niet meer als outlier wordt aangemerkt.

3 Presentatie

Gezien het grote aantal te controleren raaien (circa 200 per kustvak) en het grote aantal te controleren jaren (circa 35) is het essentieel om van om van effectieve overzichten gebruik te maken. Voor de presentatie en het vergelijken van raaien zijn na overleg een aantal methodes naar voren gekomen. Deze methodes zijn in te delen in methodes voor het opsporen van de raaien die outliers bevatten (selectie van raaien met outliers) en methodes om de eerder gesignaleerde raaien in meer detail te bekijken en eventueel te corrigeren (inspectie van raaien met outliers).

In dit hoofdstuk staan een aantal methodes beschreven. Deze hoeven niet allemaal geïmplementeerd te worden om een effectief validatie proces mogelijk te maken. Er moet nog een definitieve keus gemaakt worden uit deze methodes.

3.1 Selectie van raaien met outliers

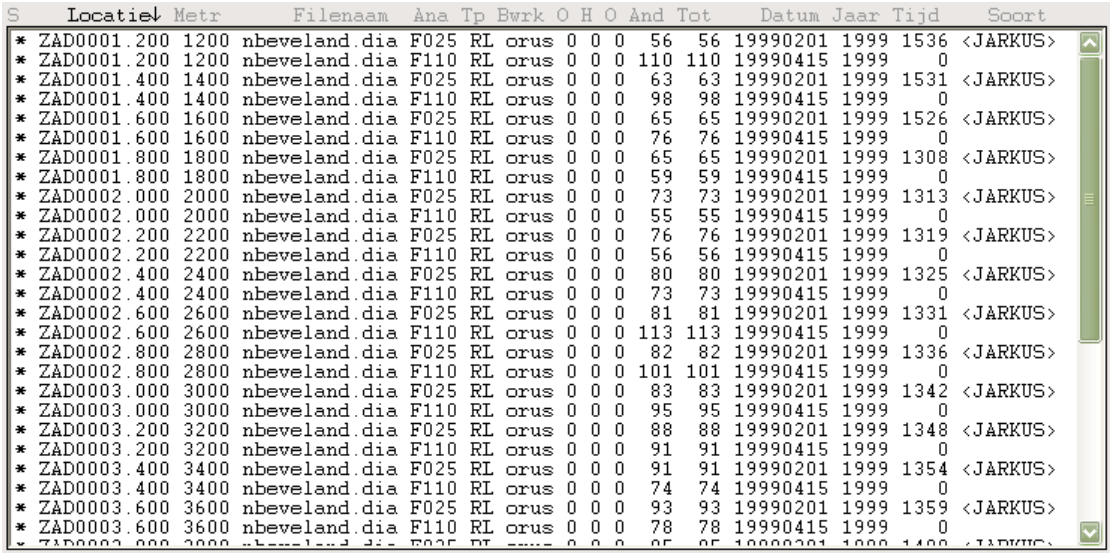
3.1.1 Het samenvattend alfanumeriek overzicht

In de voorgaande secties zijn reeds een aantal velden genoemd die men aan het alfanumeriek overzicht van raaien zou willen toevoegen, teneinde nog effectiever raaien op te kunnen sporen die fouten bevatten. Het gaat om de volgende velden:

- Het aantal outliers in de reeks (geen actie nodig, want al aanwezig in huidig overzicht);
- Het aantal hiaten in de reeks (geen actie nodig, want al aanwezig in huidig overzicht);
- De outlier status van de reeks als geheel. Deze kan zijn ‘Anders’ (nog niet geëvalueerd), ‘Outlier’ (geëvalueerd, maar een te grote afwijking geconstateerd) of ‘Ok’ (voldoet aan drempel waarde of is handmatig goedgekeurd door gebruiker);
- Het jaartal van de voorgaande reeks (indien aanwezig)
- De outlier status van de reeks ten opzichte van de voorgaande reeks;
- Het jaartal van de volgende reeks (indien aanwezig);
- De outlier status van de reeks ten opzichte van de volgende reeks;
- De locatie code van de linkerbuur-raai (indien aanwezig);
- De outlier status van de reeks ten opzichte van de linkerbuur-raai;
- De locatie code van de rechterbuur-raai (indien aanwezig);
- De outlier status van de reeks ten opzichte van de rechterbuur-raai;
- De outlier status van de reeks ten opzichte van de complementaire raai;
- De conditie ‘mogelijk niet gedetecteerde geul in hoogte reeks’.

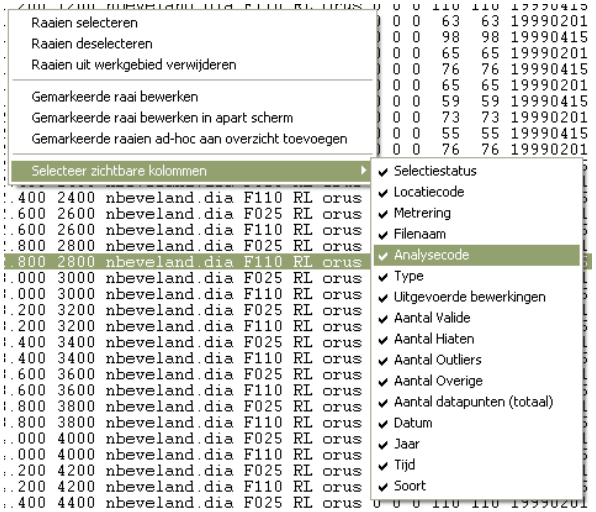
Al deze kolommen kunnen op de gebruikelijke manier aan- en uitgezet worden. Wanneer de gewenste combinatie van weer te geven kolommen eenmaal is bereikt, zou het mogelijk moeten zijn om deze combinatie als een snelmenukeuze op te slaan.

Een andere toe te voegen functionaliteit is dat de alfanumerieke lijst als overzicht ‘inhoud werkgebied’ in een ASCII file bewaard kan worden.



S	Locatie	Metr	Filenaam	Ana	Tp	Bwrk	O	H	O	And	Tot	Datum	Jaar	Tijd	Soort
*	ZAD0001.200	1200	nbeveland.dia	F025	RL	orus	0	0	0	56	56	19990201	1999	1536	<JARKUS>
*	ZAD0001.200	1200	nbeveland.dia	F110	RL	orus	0	0	0	110	110	19990415	1999	0	
*	ZAD0001.400	1400	nbeveland.dia	F025	RL	orus	0	0	0	63	63	19990201	1999	1531	<JARKUS>
*	ZAD0001.400	1400	nbeveland.dia	F110	RL	orus	0	0	0	98	98	19990415	1999	0	
*	ZAD0001.600	1600	nbeveland.dia	F025	RL	orus	0	0	0	65	65	19990201	1999	1526	<JARKUS>
*	ZAD0001.600	1600	nbeveland.dia	F110	RL	orus	0	0	0	76	76	19990415	1999	0	
*	ZAD0001.800	1800	nbeveland.dia	F025	RL	orus	0	0	0	65	65	19990201	1999	1308	<JARKUS>
*	ZAD0001.800	1800	nbeveland.dia	F110	RL	orus	0	0	0	59	59	19990415	1999	0	
*	ZAD0002.000	2000	nbeveland.dia	F025	RL	orus	0	0	0	73	73	19990201	1999	1313	<JARKUS>
*	ZAD0002.000	2000	nbeveland.dia	F110	RL	orus	0	0	0	55	55	19990415	1999	0	
*	ZAD0002.200	2200	nbeveland.dia	F025	RL	orus	0	0	0	76	76	19990201	1999	1319	<JARKUS>
*	ZAD0002.200	2200	nbeveland.dia	F110	RL	orus	0	0	0	56	56	19990415	1999	0	
*	ZAD0002.400	2400	nbeveland.dia	F025	RL	orus	0	0	0	80	80	19990201	1999	1325	<JARKUS>
*	ZAD0002.400	2400	nbeveland.dia	F110	RL	orus	0	0	0	73	73	19990415	1999	0	
*	ZAD0002.600	2600	nbeveland.dia	F025	RL	orus	0	0	0	81	81	19990201	1999	1331	<JARKUS>
*	ZAD0002.600	2600	nbeveland.dia	F110	RL	orus	0	0	0	113	113	19990415	1999	0	
*	ZAD0002.800	2800	nbeveland.dia	F025	RL	orus	0	0	0	82	82	19990201	1999	1336	<JARKUS>
*	ZAD0002.800	2800	nbeveland.dia	F110	RL	orus	0	0	0	101	101	19990415	1999	0	
*	ZAD0003.000	3000	nbeveland.dia	F025	RL	orus	0	0	0	83	83	19990201	1999	1342	<JARKUS>
*	ZAD0003.000	3000	nbeveland.dia	F110	RL	orus	0	0	0	95	95	19990415	1999	0	
*	ZAD0003.200	3200	nbeveland.dia	F025	RL	orus	0	0	0	88	88	19990201	1999	1348	<JARKUS>
*	ZAD0003.200	3200	nbeveland.dia	F110	RL	orus	0	0	0	91	91	19990415	1999	0	
*	ZAD0003.400	3400	nbeveland.dia	F025	RL	orus	0	0	0	91	91	19990201	1999	1354	<JARKUS>
*	ZAD0003.400	3400	nbeveland.dia	F110	RL	orus	0	0	0	74	74	19990415	1999	0	
*	ZAD0003.600	3600	nbeveland.dia	F025	RL	orus	0	0	0	93	93	19990201	1999	1359	<JARKUS>
*	ZAD0003.600	3600	nbeveland.dia	F110	RL	orus	0	0	0	78	78	19990415	1999	0	

Figuur 2: *Het alfanumerieke overzicht in zijn huidige vorm. Toe te voege functionaliteit: extra velden, en de mogelijkheid de lijst af te drukken*



Figuur 3: *Selectie menu voor het kiezen van de zichtbare kolommen. Aangezien het instellen van de juiste combinatie bewerkelijk kan zijn, is het gewenst een menu aan te brengen waarmee een eenmaal ingestelde configuratie aan het voorkeuze menu kan worden toegevoegd.*

3.1.2 Het gridoverzicht

Wanneer een raai sterk afwijkt van zijn buurman, dan is dit een aanwijzing dat één van de twee raaien fouten bevat, maar het is uit deze ene vergelijking nog niet duidelijk welke van de twee dat is. Bij elke raai zijn er echter meerdere reeksen waarmee een vergelijking kan worden uitgevoerd. Zo zijn er zoals al eerder opgemerkt:

- een linker en een rechter buurman raai;
- een voorgaande en volgende jaargang;
- een complementaire raai (hoogte raai of diepte raai).

Wanneer een raai sterk afwijkende opnamegegevens bevat, dan uit zich dat in meerdere inconsistenties. Deze inconsistenties kunnen middels rood gearceerde lijnen in een grid worden gevisualiseerd, waarbij elke gridknoop de combinatie van een locatie en jaargang

representeert. De complementaire raaien kunnen desgewenst in het grid worden weergegeven door deze steeds schuin onder een knoop te tekenen (zie Figuur 4 en Figuur 5).

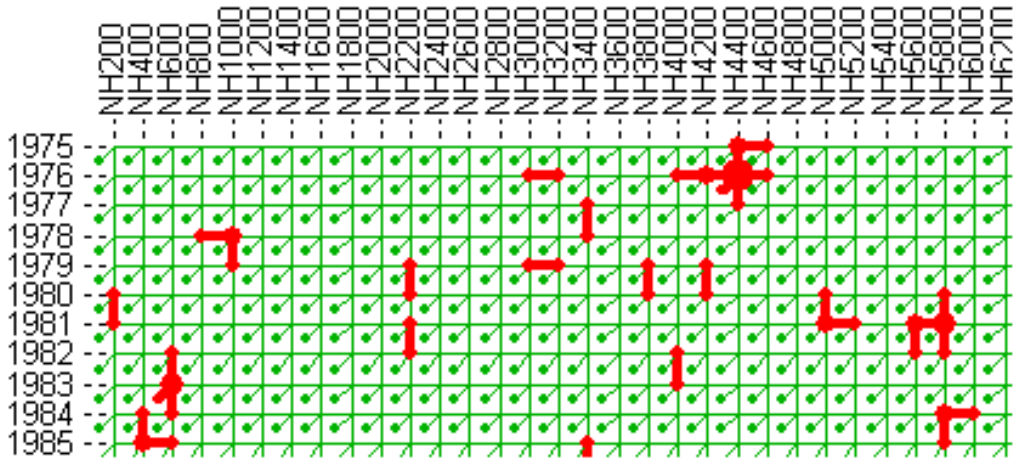
Met behulp van het gridoverzicht kan in één oogopslag de consistentie voor een groot aantal locaties en jaargangen worden gecontroleerd. Doordat alle consistentierelaties tegelijk zichtbaar zijn wordt het mogelijk om geconstateerde inconsistenties te herleiden tot de vermoedelijke aanwezigheid van fouten in een bepaalde raai. Aan het overzicht kunnen interactieve eigenschappen worden toegevoegd zodat bijvoorbeeld het raai-inspectiescherm kan worden geselecteerd voor een bepaalde raai door met de muis op een knoop te dubbelklikken.

Omdat een gemiddeld kustvak circa 200 raaien kan bevatten en het aantal jaargangen tot 35 kan oplopen dient het overzicht wel over zoom functionaliteit te beschikken (waarbij de headers van het overzicht goed zichtbaar dienen te blijven).

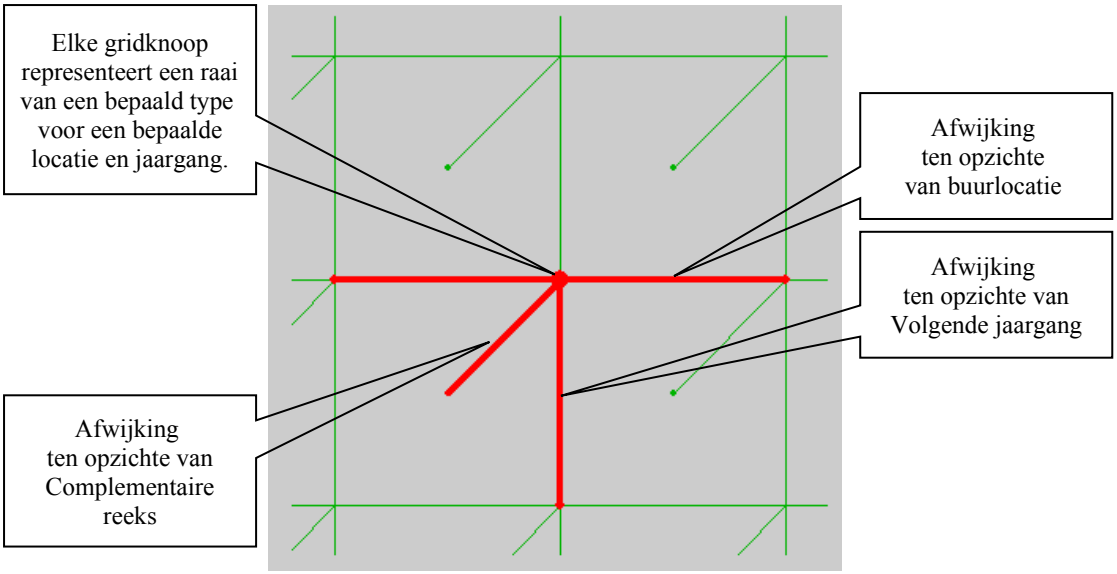
In het grid kunnen zowel de inconsistente datapunten als de inconsistente reeksen getoond worden. Naar keuze zou echter ook moeten kunnen worden ingesteld dat alleen één van beide inconsistenties getoond wordt.

Het grid overzicht dient ook een aantal invulvelden te bevatten waarin de drempel waarden voor de outliercriteria kunnen worden geconfigureerd (voor details hierover zie hoofdstuk 4). Ook de button waarmee de consistentie analyse wordt gestart bevindt zich in dit window.

Eventueel zouden de consistentierelaties tussen de complementaire raaien ook kunnen worden weergegeven, maar het is de vraag of het figuur dan nog voldoende overzichtelijk blijft.



Figuur 4: *Het gridoverzicht. Inconsistentie-relaties worden weergegeven door rode lijnen, consistentie-relaties worden weergegeven door groene lijnen. Knopen die niet voldoen aan het samengestelde outliercriterium worden met behulp van een rood bolletje weergegeven (zie de onderstaande figuur voor details)*



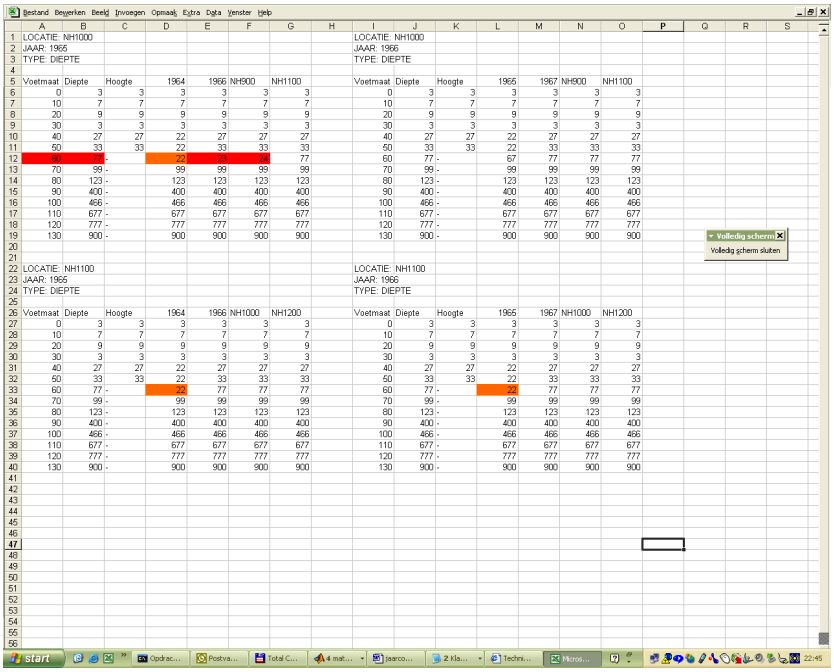
Figuur 5: *Gridoverzicht (detail)*

3.1.3 De vergelijkingstabel

Deze tabel bevat kolommen voor

- de voetmaat
- de gemeten waarden
- de gemeten waarden voor de twee buurraaien
- de gemeten waarden voor het voorgaande en het volgende jaar
- de gemeten waarde voor de complementaire reeks

Afwijkingen kunnen door middel van een kleurcode worden aangegeven. Om dit in Matlab te realiseren zal een zogenaamde text-tabel utility moeten worden geschreven, zodat verschillende elementen op één regel met een verschillende font en kleur kunnen worden weergegeven. Om hetzelfde overzicht met behoud van kleurcodering naar een document te exporteren kan niet volstaan worden met een gewoon ASCII bestand, maar zou bijvoorbeeld gebruik gemaakt kunnen worden van een gegenereerd HTML bestand. Nog mooier zou het zijn wanneer een Excel sheet inclusief de kleurcoderingen zou kunnen worden gegenereerd, maar hier is, althand bij Modelit, nog geen ervaring mee opgedaan, zodat niet met absolute zekerheid gezegd kan worden dat dit mogelijk is.



LOCATIE: NH1000						
JAAR: 1965						
TYPE: DIEPTE						
Voetmaat	Diepte	Hoogte	1964	1966	NH900	NH1100
0	3	3	3	3	3	3
10	7	7	7	7	7	7
20	9	9	9	9	9	9
30	3	3	3	3	3	3
40	27	27	22	27	27	27
50	33	33	22	33	33	33
60	77	-	22	33	33	77
70	99	-	99	99	99	99
80	123	-	123	123	123	123
90	400	-	400	400	400	400
100	466	-	466	466	466	466
110	677	-	677	677	677	677
120	777	-	777	777	777	777
130	900	-	900	900	900	900

LOCATIE: NH1000						
JAAR: 1966						
TYPE: DIEPTE						
Voetmaat	Diepte	Hoogte	1965	1967	NH900	NH1100
0	3	3	3	3	3	3
10	7	7	7	7	7	7
20	9	9	9	9	9	9
30	3	3	3	3	3	3
40	27	27	22	27	27	27
50	33	33	22	33	33	33
60	77	-	67	77	77	77
70	99	-	99	99	99	99
80	123	-	123	123	123	123
90	400	-	400	400	400	400
100	466	-	466	466	466	466
110	677	-	677	677	677	677
120	777	-	777	777	777	777
130	900	-	900	900	900	900

LOCATIE: NH1100						
JAAR: 1965						
TYPE: DIEPTE						
Voetmaat	Diepte	Hoogte	1964	1966	NH1000	NH1200
0	3	3	3	3	3	3
10	7	7	7	7	7	7
20	9	9	9	9	9	9
30	3	3	3	3	3	3
40	27	27	22	27	27	27
50	33	33	22	33	33	33
60	77	-	22	77	77	77
70	99	-	99	99	99	99
80	123	-	123	123	123	123
90	400	-	400	400	400	400
100	466	-	466	466	466	466
110	677	-	677	677	677	677
120	777	-	777	777	777	777
130	900	-	900	900	900	900

LOCATIE: NH1100						
JAAR: 1966						
TYPE: DIEPTE						
Voetmaat	Diepte	Hoogte	1965	1967	NH1000	NH1200
0	3	3	3	3	3	3
10	7	7	7	7	7	7
20	9	9	9	9	9	9
30	3	3	3	3	3	3
40	27	27	22	27	27	27
50	33	33	22	33	33	33
60	77	-	22	77	77	77
70	99	-	99	99	99	99
80	123	-	123	123	123	123
90	400	-	400	400	400	400
100	466	-	466	466	466	466
110	677	-	677	677	677	677
120	777	-	777	777	777	777
130	900	-	900	900	900	900

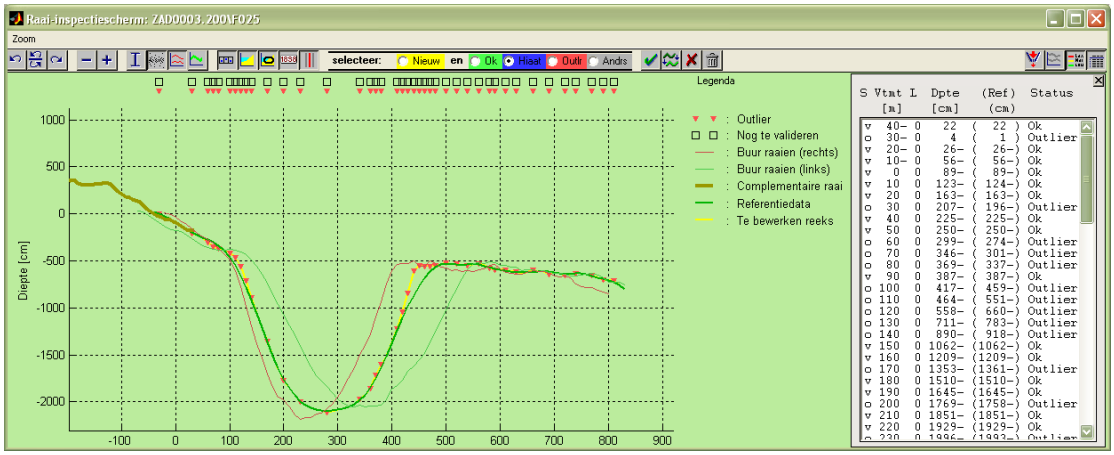
Figuur 6: *Alle datapunten opgenomen in spreadsheet. Naast elkaar: de verschillende jaargangen. Boven elkaar: de verschillende locaties. De outliers zijn via een rode achtergrondkleur gemarkeerd*

3.2 Inspectie van raaien met outliers

3.2.1 De vergelijkingsgrafiek

In dit geval worden de gegevens uit de vergelijkingstabel in een grafiek weergegeven. Dit gebeurt steeds voor één raai (met bijbassende raaien) tegelijk. Voor dit doel kan het huidige raai-inspectiescherm grotendeels volstaan. In de huidige situatie kunnen de volgende raaien al automatisch worden weergegeven bij het selecteren van een raai:

- De buurraaien voor hetzelfde jaar
- De complementaire raai voor hetzelfde jaar
- De historische raaien, met de beperking dat alle historische raaien worden getoond, in plaats van de voorgaande en de volgende



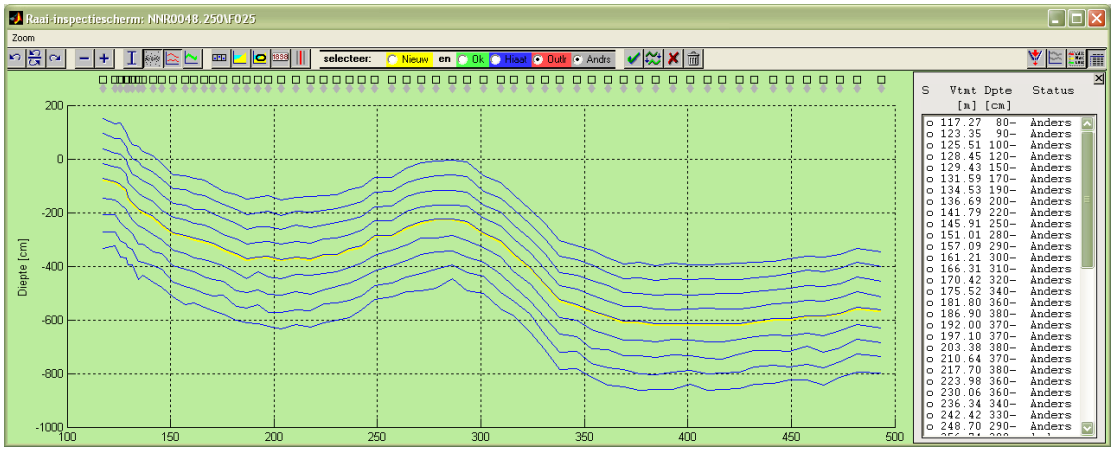
Figuur 7: *Voorbeeld van een vergelijkingsgrafiek. De geselecteerde raai wordt samen met buurraaien, voorgaande, volgende, en complementaire raai weergegeven*

De gewenste uitbreidingen ten opzichte van de huidige situatie zijn derhalve:

- dat een geschatte raai kan worden aangemaakt op basis van de buur-raaien en de voorgaande en volgende raai. Dit maakt het mogelijk de outliers te identificeren en markeren;
- dat er meer controle komt over het weergeven van historische raaien. In de huidige situatie kan de weergave van historische raaien enkel aan- of uitgezet worden. Naar analogie van het weergeven van buur-raaien zou men willen kunnen selecteren dat er maximaal M voorgaande en N volgende jaargangen worden getoond;

3.2.2 De gestapelde grafiek

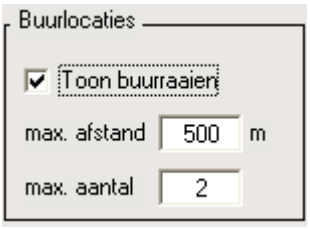
In de gestapelde grafiek wordt een raai tegelijk met zijn buurraaien en/of volgende/voorgaande jaargangen weergegeven. Daarbij wordt voor elk jaar dat de opnamedatum verschilt, of voor elke 100 meter dat de metrering verschilt een instelbare ophoging toegepast (bijvoorbeeld 50 centimeter).



Figuur 8: *Voorbeeld van een gestapelde grafiek. Voor elk jaar dat de opnamedatum verschilt wordt een (instelbare) ophoging toegepast. Hetzelfde principe kan ook worden toegepast bij het weergeven van buurraaien.*

De weergave van de gestapelde grafiek kan het beste met het bestaande raai-inspectie scherm geïntegreerd worden. Ten opzichte van de huidige situatie zijn dan de volgende uitbreidingen nodig:

- Bij de selectiebox voor de weergave van de buurraaien (zie figuur) moet een veld worden toegevoegd waarin kan worden aangegeven wat de offset per 100 meter verschil in metrering dient te zijn. Door voor deze offset 0 in te vullen verkrijgt men het equivalent van de huidige situatie;
- De labels in de legenda en de eventueel interactief aan te brengen labels dienen aangepast te worden;
- Analoog aan de selectiebox voor buurraaien is een meer uitgebreide selectiebox voor historische raaien nodig. Merk op dat hiermee tevens een groot deel van de bij sectie 3.2.1 genoemde functionaliteit wordt gerealiseerd;
- De labels bij de historische raaien dienen aangepast te worden aan de ingestelde offset



Figuur 9: *Selectiebox voor de weergave van de buurraaien (huidige situatie)*

4 Berekening van voorspelde waarden en outliercriteria

Doordat voor de meeste locaties data van vele jaren beschikbaar zijn, wordt het mogelijk om een statistische analyse van enige betekenis toe te passen op deze data. Ten behoeve van de eenvoud zullen we ons daarbij beperken eenvoudige aannames.

4.1 Berekening van voorspelde waarden

Een eenvoudig voorspellend model is de gewogen som. Hierbij wordt de diepte voor de te controleren raai (de hoofdraai) bepaald als een gewogen som van de raaien die aan de te analyseren raai gerelateerd zijn (de nevenraaien). Als weegfactoren worden de reciprokes van de varianties van de afzondelijke verschillen gebruikt:

$$\hat{d}(v,j) = \left(\sum_{b \in B} \frac{1}{\sigma_b^2} d_b(v,j) \right) / \left(\sum_{b \in B} \frac{1}{\sigma_b^2} \right)$$

met:

v	Index voor voetmaat	
j	Index voor jaar	
B	De verzameling bij deze raai horende nevenraaien, zijnde de 2 buur-raaien, de raai voor het voorgaande en volgende jaar en de complementaire raai	(0)
b	Index in de verzameling nevenraaien	
$\hat{d}(v,j)$	Geschatte diepte voor de hoofdraai	
$d_b(v,j)$	Waargenomen diepte voor nevenraai	
σ_b^2	Variantie van het verschil tussen nevenraai en hoofdraai, te berekenen volgens formule (0)	

Dit model gaat er feitelijk vanuit dat de hoofdraai op een aantal onafhankelijke manieren kan worden voorspeld, De variantie is steeds maatgevend voor de kwaliteit van deze voorspellingen.

Hierbij gelden nog de volgend opmerkingen:

- Er wordt (nog) niet expliciet rekening gehouden met systematische verschillen tussen hoofdraai en nevenraaien. Met systematisch verschil wordt een verschil bedoeld dat elk jaar terug te vinden is. Met name bij buurraaien zou dit kunnen optreden;
- De variantie van het verschil tussen neven- en hoofdraai wordt voor alle voetmaten gezamenlijk berekend, alhoewel de data het in principe toestaan deze variantie per voetmaat te berekenen. Het voordeel van één waarde voor de variantie is, naast de grotere eenvoud, dat dit beter aansluit met de werkwijze binnen Morfologie Applicatie. Daar geldt dat per raai ook steeds één drempelwaarde wordt ingesteld voor het identificeren van outliers.

$$\sigma_b^2 = \frac{1}{N}\sum_{j\in J}\sum_{v\in V_j}\left(d_b(v,j) - d(v,j)\right)^2$$

met:

J	De verzameling jaren met waarnemingen	(0)
V_j	De verzameling voetmaten voor jaar j waarvoor beide reeksen geen hiaat vertonen	
N	Het totaal aantal waarnemingen, oftewel: $N = \sum_{j\in J}\sum_{v\in V_j} 1$	
$d(v,j)$	Waargenomen diepte voor de hoofdraai	

4.2 Outlier criteria voor datapunten

Door het combineren van meerde onafhankelijke voorspellende modellen (voor elke nevenraai wordt één voorspelling gegenereerd) wordt de kwaliteit van de voorspelling verbeterd.

Bij onafhankelijkheid van de voorspellingen laat de gemiddelde kwadratische fout zich iteratief berekenen als:

$$\sigma_{1,1}^2 = \sigma_1^2$$
$$\sigma_{1,2..b}^2 = \frac{\sigma_{1,2..b-1}^2\sigma_b^2}{\sigma_{1,2..b-1}^2 + \sigma_b^2} \qquad (0)$$

Bij elke nevenraai die wordt toegevoegd neemt de variantie dus af.

Er is sprake van een outlier wanneer de werkelijke kwadratische afwijking van voorspelde en waargenomen waarde significant groter is als de bovengenoemde variantie, of equivalent: indien de absolute afwijking groter is dan de wortel van de variantie (de standaardafwijking). Niet elke outlier wordt veroorzaakt door een fout, indien een correcte waarneming toch een outlier veroorzaakt, spreken we van een *vals alarm*.

De kans op een vals alarm hangt samen met het aantal maal dat de standaardafwijking moet worden overschreden voordat een outlier wordt getriggerd. Als de afwijkingen normaal verdeeld zijn, dan is de kans dat een absolute afwijking van 3.1 σ of groter optreedt kleiner dan 1 op 500. Met andere woorden: als we het outliercriterium instellen op een maximaal acceptabele afwijking van 3.1 σ dan genereren we minder dan 1 vals alarm op de 500 datapunten. De onderstaande tabel bevat de vals-alarm kansen voor andere instellingen.

Tabel 1: *Drempelwaarden met bijbehorende vals alarm kansen*

Ingestelde drempelwaarde voor de acceptabele afwijking (maal σ)	kans op een onterechte kwalificatie als outlier (%)
D=1.3	20
D=1.6	10
D=2.0	5
D=2.3	2
D=2.6	1
D=3.1	0.2
D=3.3	0.1
D=3.9	0.01
D=4.4	0.001

Er moet een goede trade-off gevonden worden tussen het optreden van valse alarmen en het detecteren van fouten. Wanneer een outlier pas wordt getriggerd bij een overschrijding van 10 maal de standaardafwijking, dan kunnen we er vrijwel zeker van zijn dat elke outlier een fout betreft, maar zullen er tevens vele foutieve waarnemingen zijn die geen outlier triggeren. Een praktijk waarde voor individuele datapunten zou tenminste 3.3 moeten bedragen, want anders worden er, bij een gemiddelde van 100 datapunten per raai, te veel valse alarmen gegenereerd, en moet alsnog het merendeel van alle raaien worden geïnspecteerd.

4.3 Outlier criteria voor reeksen

Veel fouten zoals genoemd in de inleiding beïnvloeden alle datapunten in een raai tegelijk en zouden dus vrij gemakkelijk op te sporen moeten zijn.

Als criterium voor de afstand tussen twee raaien hanteren we nu de gesommeerde kwadratische afstand:

$$\Delta = \sum_{v \in V_j} \left(\frac{d(v,j) - \hat{d}(v,j)}{\sigma} \right)^2 \qquad (0)$$

Als we mogen aannemen dat voor elke voetmaat een onafhankelijke bijdrage aan deze afstandsmaat wordt geleverd, dan geldt dat de verwachte waarde van Δ is gegeven door:

$$E[\Delta] = \sum_{v \in V_j} E\left[\left(\frac{d(v,j) - \hat{d}(v,j)}{\sigma} \right)^2 \right] = \sum_{v \in V_j} 1 \qquad (0)$$

Een kleinere waarde van Δ dan deze verwachting betekent data de data zeer veel lijken op de voorspelde waarde. Een (veel) grotere waarde betekent dat de data slecht overeen komen.

Om een outlier conditie te kunnen definiëren moeten we aannames doen omtrent de verdeling van Δ. De meest voor de hand liggende aanname is dat de verschillen $d(v,j) - \hat{d}(v,j)$ normaal verdeeld zijn met verwachting 0 variantie σ^2. Als we deze verschillen normeren met de standaard afwijking σ ontstaat een standaard normale verdeling. Het kwadrateren en sommeren van deze verdeling voor alle voetmaten (zoals gebeurd in vergelijking (0)) resulteert in een zogenaamde Chi-Square verdeling met N graden van vrijheid, waarbij N

overeen komt met het aantal gesommeerde termen. N.B.: de verwachting van een dergelijke verdeling is N en de variantie 2.N.

Analoog aan Tabel 1 kunnen ook nu drempelwaarden met bijbehorende vals alarm kansen worden berekend, zodat een drempelwaarde kan worden gekozen waarbij een acceptabele vals alarm kans geldt. Als N echter voldoende groot is kan voor een simpeler oplossing worden gekozen, zijnde dat de verdeling van Δ/N wordt benaderd door een normale verdeling met verwachting 1 en variantie 2/N.

Als outliercriterium voor reeksen geldt derhalve dat een reeks als outlier moet worden aangeduid indien:

$$\sum_{v \in V_j} \left(d(v,j) - \hat{d}(v,j) \right)^2 > N.\sigma + D.\sigma \sqrt{2.N} \qquad (0)$$

Hierin geldt D als drempel waarde parameter waarvan de vals alarm kansen zijn gegeven in Tabel 1 met dien verstande dat de vals alarm kansen gehalveerd mogen worden omdat het hier een eenzijdig begrensd acceptatie interval betreft.

5 Planning

In deze sectie volgt een indicatieve planning voor de realisatie van software modules die in dit rapport beschreven zijn, uitgedrukt in te besteden manuren

Functie	**Zie sectie**	**Realisatie tijd [uur]**
Ondersteuning outlierstatus voor reeksen	2.5	4
Het samenvattend alfanumeriek overzicht	3.1.1	4
Het gridoverzicht	3.1.2	24
De vergelijkingstabel	3.1.3	>24
De vergelijkingsgrafiek	3.2.1	4
De gestapelde grafiek	3.2.2	4
Berekening voorspelde waarden	4.1	12
Berekening reeksgebonden en puntgebonden outlierstatus	4.2 en 4.3	12

6 Advies

De vergelijkingstabel kan makkelijk gemist worden bij het valideren van meerjarige kustmetingen omdat dezelfde functionaliteit (selectie en presentatie van data) ook en op meer bruikbare wijze aanwezig is in het gridoverzicht en de vergelijkingsgrafiek. Realisatie hiervan wordt niet aanbevolen.

Het samenvattend alfanumeriek overzicht, de vergelijkingsgrafiek, de gestapelde grafiek en in iets mindere mate het gridoverzicht bieden ook buiten de validatie van meerjarige reeksen een nuttige bijdrage aan de Morfologie Applicatie. Realisatie wordt vanwege deze redenen en vanwege hun noodzakelijkheid voor de voorgestelde validatie methodiek aanbevolen.

De ondersteuning outlierstatus voor reeksen is niet strikt noodzakelijk, maar maakt het systematisch werken wel makkelijker.

De berekening voorspelde waarden en de berekening van reeksgebonden en puntgebonden outlierstatus vormt het hart van de voorgestelde methodiek. Realisatie wordt vanwege deze reden aanbevolen.

Het totale pakket volgens dit advies komt derhalve neer op een tijdbesteding van 64 uur, waarbij naast de functionaliteit voor het valideren van meerjarige reeksen ook een aantal nuttige functies in meer algemene zin aan de Morfologie Applicatie worden toegevoegd.