EINDRAPPORT
T2009 RAPPORTAGE SCHELDE ESTUARIUM

DIGITALE BIJLAGE 3 “METHODEN EN TECHNIEKEN”

MINISTERIE VAN INFRASTRUCTUUR EN MILIEU
RIJKSWATERSTAAT ZEELAND, UITVOEREND
SECRETARIAAT VAN DE VLAAMS-NEDERLANDSE
SCHELDECOMMISSIE

11 juli 2013
077193992:A - Definitief
C03041.002718.0400

Inhoud

3 Methoden en technieken ... 3
3.1 Inleiding ... 3
3.2 Principes ... 3
3.2.1 Visuele exploratieve data-analyse ... 3
3.2.2 Statistische toetsing op trend ... 4
3.2.3 CrossCorrelatie ... 5
3.2.4 Forecasting ... 5
3.3 Trend analyse tool ... 5
3.3.1 Installatie ... 5
3.3.2 Functionaliteit ... 6
3.3.3 Grafische omgeving ... 6
3.3.4 Invoer en uitvoer ... 9
3.3.4.1 Invoerformaat van tijdreeksen ... 9
3.3.4.2 Uitvoer van resultaten ... 10
3.4 Bepaling van trendbreuken ... 11

3 Methoden en technieken

3.1 INLEIDING

Dit document bevat de digitale bijlage bij hoofdstuk 3 "Methoden en Technieken" van het T2009-rapport. Hierin wordt een overzicht gegeven van de Trend Analyse Tool die gebruikt is voor het uitvoeren van statistische analyses in het kader van de T2009-rapportage.

3.2 PRINCIPES

De (trend)analyses die gevraagd worden in het kader van de T2009-rapportage, omvatten volgende zaken:

- Trends en trendbreuken dienen bepaald te kunnen worden, met statistische toetsing van de waargenomen trends;
- Gekende externe factoren (met name cyclische factoren zoals de 18.6-jarige cyclus in waterstanden) moeten in rekening kunnen gebracht worden;
- Er moet kunnen bepaald worden of er naast deze gekende trends nog andere trends aanwezig zijn in de data, met andere woorden, na detrending moet getest worden of de residuen normaal verdeeld zijn;
- Er moeten voorspellingen kunnen uitgevoerd worden op basis van regressiemodellen;
- De waargenomen trends in de verschillende reken- en verklarende parameters moeten via een (cross)correlatieanalyse onderzocht kunnen worden.

Trendanalyses worden meestal beperkt tot een curve-fitting van een vooropgestelde relatie (exponentieel, polynoom,…) met in het beste geval een betrouwbaarheidsinterval. Hierbij wordt echter uitgegaan van een *subjectief* idee, waarbij men met de data wil bewijzen wat men in diezelfde data reeds heeft menen waar te nemen. Statistisch is dit niet correct.

De hier voorgestelde trendanalyse wordt gebaseerd op een *objectieve* hypothesetoetsing, waarbij wordt nagegaan of de afwezigheid van een trend kan worden verworpen. Indien dit het geval is, wordt in een volgende stap nagegaan of een welbepaalde trend aanvaard kan worden. Tevens worden de kansen bepaald op een foute conclusie.

De trendanalyse kan daarom worden uitgevoerd in een aantal stappen:

- Visuele analyse (herkenning van mogelijke trends)
- Statistische toetsing op trend
- Trendbepaling

3.2.1 VISUELE EXPLORATIEVE DATA-ANALYSE

Een eerste stap omvat het visueel exploreren van de gegevens door het opmaken van een aantal grafiekweergaven en smoothing van de tijdreeksen om meerjarige (mogelijke) trends, trendbreuken of

relaties te herkennen.

Hierbij worden de verschillende variabelen eerst grafisch weergegeven in exploratieve plots :

- Plotten van de tijdreeks: de data wordt afgebeeld zoals in functie van de tijd.
- (Cumulatieve) plot : heeft een louter indicatieve betekenis omtrent plotse gedragswijzingen van toevalsvariabelen. Vooral nuttig is het gelijktijdig plotten van verschillende variabelen op dezelfde figuur, wat op efficiënte wijze inzicht verschaft over al dan niet gekoppeld gedrag.
- First difference plot : deze geeft een aanwijzing over het 'random walk' gedrag. 'Random walk' geeft aan of een systeem evolueert volgens een willekeurige groei. Indien 'random walk' gedrag zichtbaar is (indien de figuur stationair is) wordt een trend minder waarschijnlijk en vice versa. Het 'random walk' model kan getoetst worden getoetst d.m.v. een berekening van de autocorrelatie in de first difference plot : indien deze gelijk kan gesteld worden aan nul, mag men het 'random walk' model aannemen, wat dus een aanwijzing is om de aanwezigheid van een trend te verwerpen.

De tweede stap van de exploratieve data-analyse bestaat uit een smoothing-procedure, waarbij de random variatie van de toevalsvariabelen onderdrukt wordt. Er worden verschillende methodes ter beschikking gesteld in de software:

- Toepassing van glijdende gemiddelden
- Lowess regressie (Robust Locally Weighted Regression and Smoothing ; Cleveland, 1988) : bij beschikbaarheid van een grotere hoeveelheid gegevens is dit een waardevolle techniek. De Lowess regressie is een lokale (lineaire of kwadratische) regressie over een voortschrijdend venster van de data, waarbij telkens een (trikubische) gewichtsfunctie wordt toegepast op het betreffende venster. Deze smoothing procedure wordt standaard gebruikt in meteorologische, hydrologische, geologische en morfologische tijdreeksen, om trends bloot te leggen waarvan men niet a priori weet welke vorm men mag verwachten : deze methode is immers niet-parametrisch, zodat de trend elke denkbare vorm mag aannemen. Bovendien laat de Lowess toe om cyclisch gedrag visueel te herkennen. Voorspelling op basis van deze regressiemodellen is niet mogelijk.

3.2.2 STATISTISCHE TOETSING OP TREND

Wanneer men wil te weten komen of er een trend aanwezig is, is de logische nulhypothese H_0 dat er geen trend aanwezig is. Een statistische test gaat immers op zoek naar het in de metingen aanwezige bewijsmateriaal tegen de nulhypothese. Indien dat bewijsmateriaal er is, is er een bewijs "ten gunste" van het effect dat onderzocht wordt, in dit geval een trend. Een significantietoets is immers ontworpen om de sterkte van het bewijs tegen de H_0 vast te stellen.

Data waar a priori gekende trends aanwezig zijn (bv meerjarige cycli in waterstanden), dienen eerst *gedetrend* worden zodat de resulterende residu's op bijkomende onbekende trends kunnen onderzocht worden. De residu's kunnen eerst aan de visuele data-exploratie onderworpen worden vooraleer verder gegaan wordt. Dit kan aanleiding geven tot het formuleren van alternatieve hypothesen (verwachte trends). H_a, de alternatieve hypothese, geeft deze verwachte trend(s) aan. Dit wil zeggen dat men vooraf moet uitmaken welke richting men van deze trend verwacht (b.v. exponentieel afnemend).

Het significantieniveau bepaalt de bewijssterkte. Als $\alpha = 0.05$, eisen we dat de waargenomen trend zo sterk afwijkt van H_0 (geen trend), dat deze onder omstandigheden dat H_0 geldig is (b.v. verwachte richtingscoëfficiënt = 0), slechts in 5% van de gevallen zal waargenomen worden. Het significantieniveau geeft ook aan wat de kans is dat een foute conclusie wordt getrokken, nl. de verwerping van H_0 terwijl deze geldig zou zijn. Dat heeft ook enig belang bij het veelvuldig herhalen van de toets: in α procent van

de gevallen zal de toets een verkeerde conclusie geven. Hiermee dient men rekening te houden bij de keuze en de interpretatie van α.

De keuzes die moeten gemaakt worden :

- welke parameters worden onderworpen (hiervoor moeten afhankelijkheden in kaart gebracht worden : trends op afhankelijke variabelen geven een vertekend beeld);
- formulering van basishypothese én alternatieve hypothese;
- significantieniveau.

3.2.3 CROSSCORRELATIE

Om na te gaan of trends in verschillende parametersets gelijk lopen, kan van de (cross)correlatie uitgegaan worden.

Gewoonlijk wordt als correlatieparameter de zgn. correlatiecoëfficiënt ρ gekozen, waarmee dan de Pearson coëfficiënt wordt bedoeld (b.v. in Excel). Deze coëfficiënt geeft echter enkel de mate aan van *lineaire* afhankelijkheid, terwijl er in feite geen reden bestaat om deze vorm van afhankelijkheid a priori te veronderstellen. Spearman's ρ en Kendall's τ maken geen a priori veronderstelling omtrent de vorm van afhankelijkheid. Deze correlatiecëfficiënten zijn zgn. rangcorrelatiecoëfficienten en zijn een verdelingsvrije maat voor correlatie, en zijn dus betrouwbaarder.

3.2.4 FORECASTING

Uiteindelijk kan voorspeld worden wat toekomstige waarden zullen zijn. Rond die voorspelling worden predictie-intervallen berekend (b.v. 95% predictie-intervallen). Dit laat toe om onmiddellijk te kunnen vaststellen of de metingen de verwachte trend volgen. Hierbij dient een referentieperiode opgegeven te worden.

Het predictievenster kan opgesteld worden in functie van de gestelde vraag, bv ligt de huidige trend binnen de verwachtingen op basis van voorafgaande metingen (historisch perspectief). Zo kan b.v. op basis van trends en random variaties op die trends ingeschat worden wat de kans is dat de ontwikkeling van het laagdynamisch oppervlak in 2009 binnen de verwachting ligt op basis van de voorgaande jaren.

3.3 TREND ANALYSE TOOL

Op basis van de hierboven beschreven principes is een trend analyse programma ter beschikking gesteld door IMDC. Het is ontwikkeld in Matlab.

3.3.1 INSTALLATIE

Het analyseprogramma vereist dat Matlab-functies afzonderlijk worden geïnstalleerd. Dit gebeurt onder de vorm van een Matlab Compiler Runtime (MCR) die meegeleverd wordt met het analyseprogramma. Deze installatie is eenmalig uit te voeren.

Na installatie van deze MCR is het bestand dat het analyseprogramma omvat uitvoerbaar zonder bijkomende installaatie.

3.3.2 FUNCTIONALITEIT

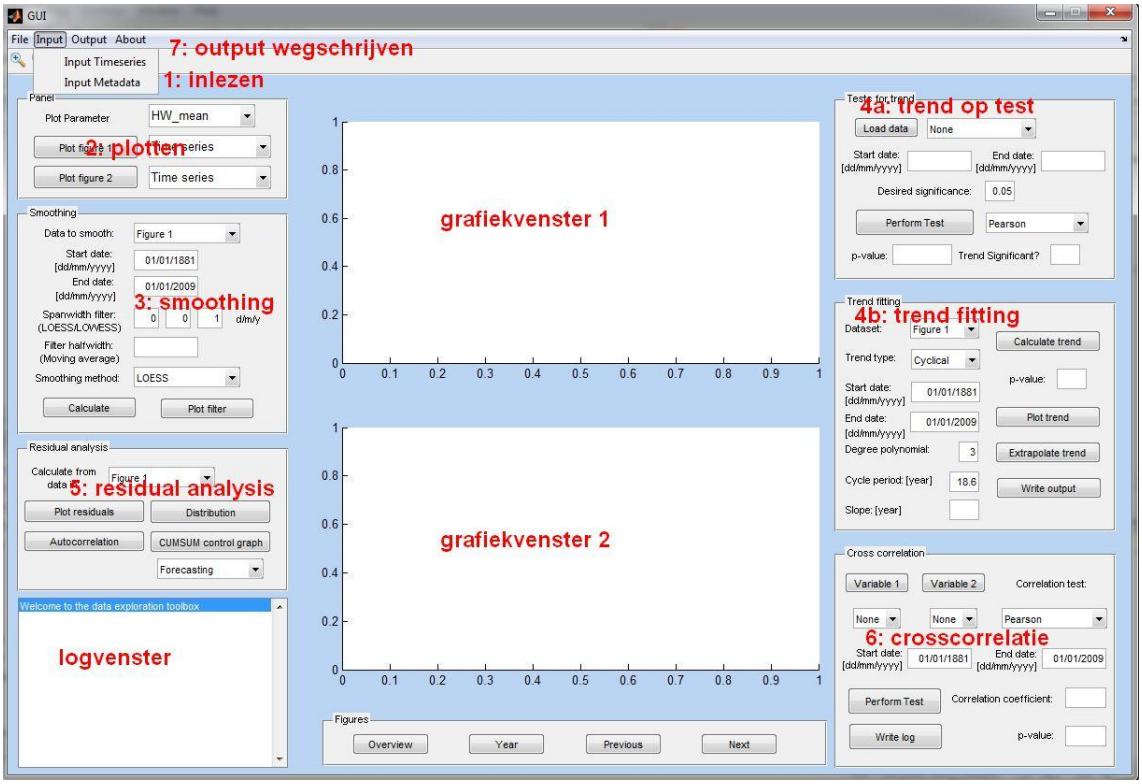
In de trend analyse tool is de volgende functionaliteit voorzien. In de volgende paragrafen worden specifieke elementen verder toegelicht.

- Inlezen van tijdreeksen in een vast dataformaat
- Testen voor trend (lineair (Pearson) of niet-parametrisch (Spearman's Rho, Mann-Kendall))
- Visualisatie van tijdreeksen
  - Plotten van tijdreeksen
  - Plotten van *first difference* en cumulatieve grafieken
- Weergeven van distributie (histogram), autocorrelatie, normaliteitsplot
- Visuele exploratie op trend
  - Moving average curve (half-width window in samples)
  - LO(W)ESS curve (venster in tijd)
- Regressie met behulp van verschillende functies:
  - Polynomiaal (met vrij te bepalen graad)
  - Cyclisch (met opgave van een gekende periode)
  - Combinatie van polynomiaal en cyclische trends (simulatane regressie)
  - Lineair, met gekende slope (om gekende lineaire trend uit data te verwijderen)
  - Weergave van regressie-vergelijking en significantieniveau (p-waarde)
  - Visualisatie van de regressiecurve met confidentie- en predictie-intervallen.
- Extrapolatie op basis van laatst uitgevoerde regressie
  - Forecast of hindcast
- Analyse van de residuen
  - Op basis van een uitgevoerde regressie (al dan niet met extrapolatie) worden de residuen berekend
  - Distributie, autocorrelatie, normaliteitsplot en CUMSUM controlegrafiek op basis van deze residuen
- Crosscorrelatie (lineair (Pearson) of niet-parametrisch (Spearman's Rho, Kendall's Tau))
- Uitvoer
  - Grafisch: afbeeldingen van de verschillende types figuren kunnen opgeslagen worden
  - Textueel:
    - Er wordt een logbestand bijgehouden dat alle bewerkingen opslaat
    - Manuele opslag van weerhouden regressiegegevens met uitvoer van modelwaarden, confidentie- en predictie-intervallen, residuen
    - De uitvoerbestanden met residuen kunnen opnieuw ingeladen worden om hierop verdere trendanalyses uit te voeren

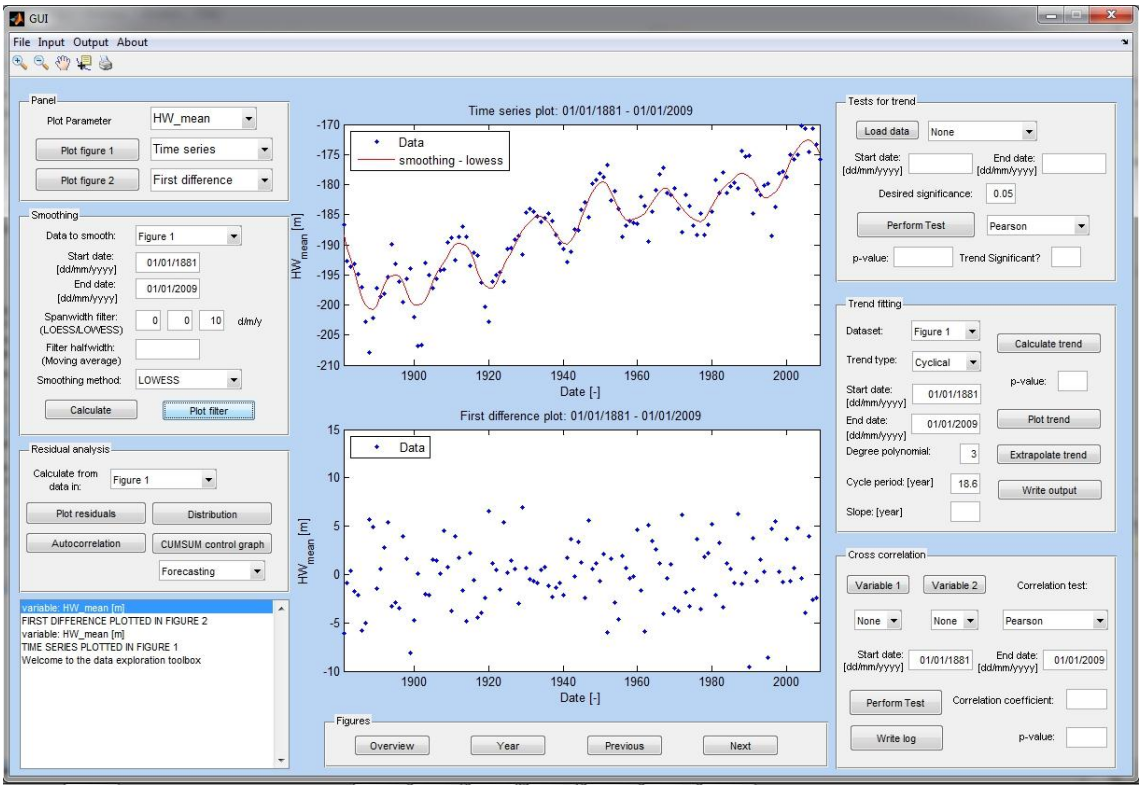
3.3.3 GRAFISCHE OMGEVING

In een aantal figuren wordt de grafische omgeving van de trend analyse tool geïllustreerd. In Figuur 3.3.3.1 worden de verschillende onderdelen van de grafische omgeving aangeduid:

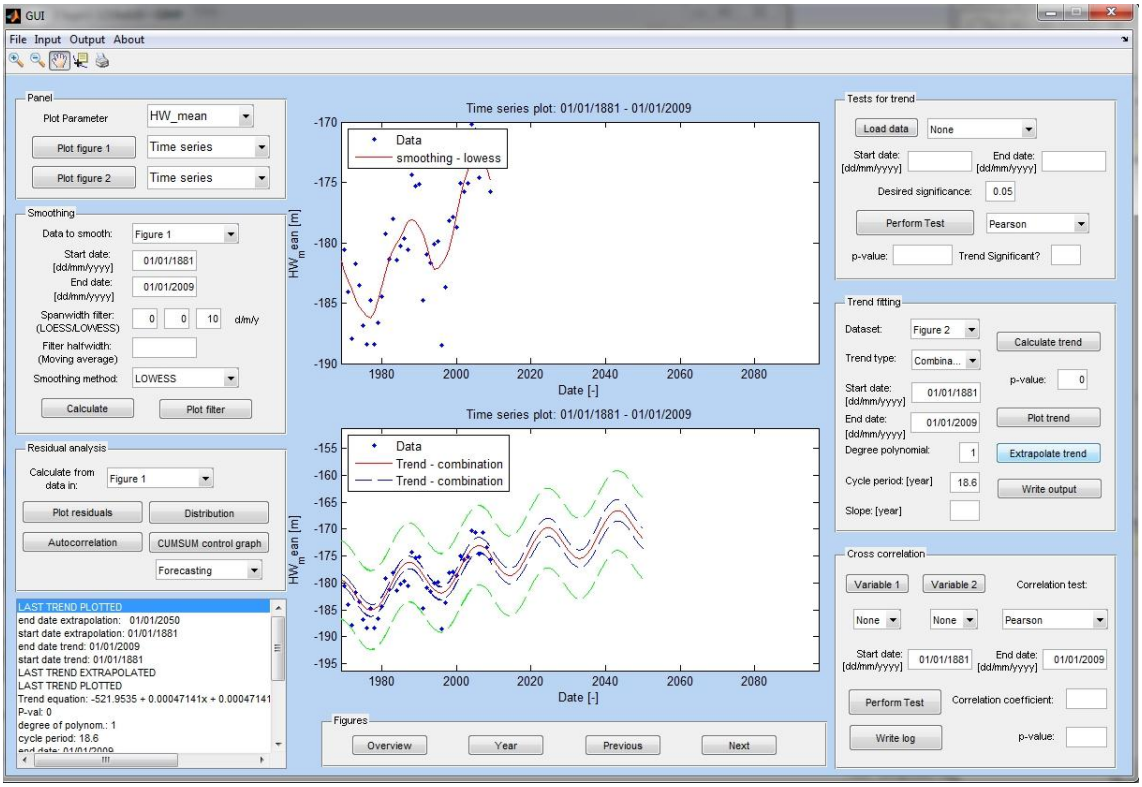
- In het midden van het beeld komen twee grafiekvensters voor. X-assen kunnen gekoppeld worden;
- Linksonder worden de acties in de software geregistreerd en worden resultaten weergegeven
- Items 1, 7 in de menubalk: inlezen van tijdreeksen, uitvoer van logbestanden, …;
- Item 2: Functies voor het afbeelden van tijdreeksen, cumulatieve plot, first difference plot, …;
- Item 3: Functies voor de visuele exploratie op trend: smoothing, Lowess curves;
- Item 4a: Functies voor het inladen van gegevens en testen op aanwezigheid van trend;
- Item 4b: Functies voor het uitvoeren van de regressie-analyse, afbeelden, extrapolatie, uitvoer;
- Item 5: Functies om de residuen ten opzichte van het regressiemodel verder te analyseren;
- Item 6: Functies voor het uitvoeren van cross-correlaties.



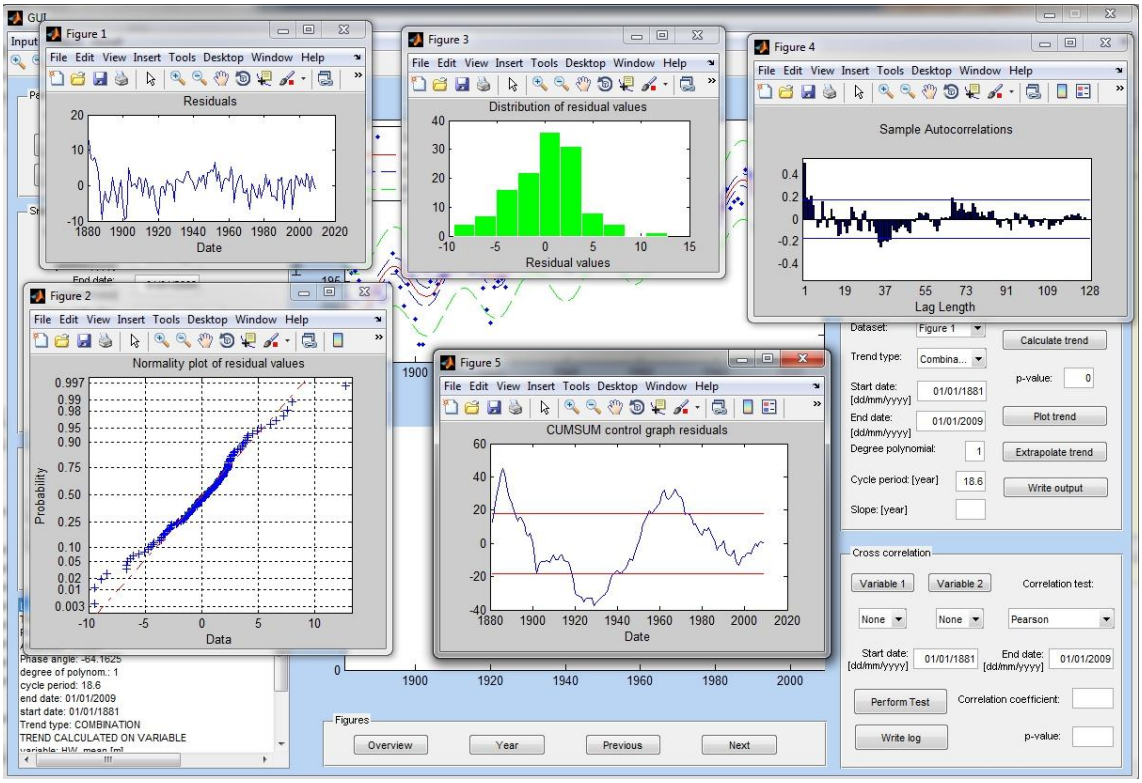
Figuur 3.1: Onderdelen van de grafische omgeving



Figuur 3.2: Voorbeeld van de afbeelding van een tijdreeks met Lowess curve (bovenste venster) en een first difference plot van de zelfde data (onderste venster)



Figuur3.3: Voorbeeld van de afbeelding van een tijdreeks met Lowess curve (boven) en een trendfitting met voorspelling (onder).



Figuur 3.4: Exploratie van de residuen

3.3.4 INVOER EN UITVOER

Er wordt in de software met een vast formaat van databestanden gewerkt. Deze worden hieronder toegelicht.

3.3.4.1 *INVOERFORMAAT VAN TIJDREEKSEN*

Het databestand (ASCI) dient strikt de volgende indeling te hebben:

- Header
  - Vrije tekst, invoegen metadata met uitzondering van de volgende sleutelwoorden:
  - EXCEPTION -9999 *"Geen data" (matlab NaN) waarde*
  - LOCATION VLISSINGEN *Vrije tekst die in titel van de figuren wordt geplot*
  - DATEFORMAT dd/mm/yyyy *String met definitie datumformaat. Toegelaten zijn de door matlab gekende formaten, waaronder: dd/mm/yyyy, mm/yyyy, yyyy, …*
  - TIMEUNIT *de eenheid van de datums in de tijdreeks: zijn de stappen in dagen, weken, maanden, jaren… Dit bepaalt de grootte van bv de slope van een linaire regressie.*
- Aanduiding begin data
  - Er word *letterlijk* 5 keer een % teken afgebeeld op een nieuwe lijn: %%%%%
- Datablok
  - Opmaak is in kolommen, met TAB als delimiter
  - Eerste twee rijen bevatten de Grootheid en Eenheid.
  - Daarna volgen per rij de waarden van variabelen voor verschillende tijdstippen.
- Voorbeeld:

Jaargemiddeld Laag- en hoogwater en getijslag, vanaf 1881 te VLIS
EXCEPTION -9999
DATEFORMAT yyyy
TIMEUNIT year
LOCATION VLISSINGEN
%%%%%%%

YEAR	LW	HW	Getijslag
[-]	[cm]	[cm]	[cm]
1881	-186.62	184.87	371.49
1882	-192.68	183.11	375.79
1883	-193.58	182.81	376.38
1884	-193.16	181.03	374.19
1885	-194.95	179.12	374.07
1886	-197.1	178.85	375.95
1887	-202.86	170.49	373.35
1888	-207.86	167.97	375.83

…

3.3.4.2 *UITVOER VAN RESULTATEN*

De uitvoer van de resultaten van een regressie volgt dezelfde opbouw van het hierboven weergegeven invoerformaat.

Er zijn enkele toevoegingen:

- Naam van het bronbestand
- Start- en einddatum van de tijdreeks
- Regressiegegevens:
    - Variabele waarop gewerkt is
    - Begin- en einddatum waarop de regressie is uitgevoerd
    - p-waarde (indien p=0 is deze kleiner dan de ingebouwde tolerantie)
    - Trendvergelijking
- Resultaten met de gemeten waarde, de modelwaarde, residuen, predictie en confidentie-intervallen voor p<0.05.
- Voorbeeld:

OUTPUT GENERATED BY IMDC TATOO v1.7 AT 09:20:44 15/01/2013

data read from file: LWHWGS99pHW_antl_comb_xls.txt
start date: 1901
end date: 2009

EXCEPTION -9999
DATEFORMAT yyyy
TIMEUNIT year
LOCATION ANTL

REGRESSION DATA
Trend determined on variable: HW
Trend type:COMBINATION ANGLE-POLYNOMIAL
Start date of trend: 1980
End date of trend: 2009
P-value: 0.0023743
Degree polynom.: 1
cycle period: 18.613
Phase angle: 95.7793
Amplitude:3.7906
Trend equation: -138.8947 + 0.33456x + 3.7906 cos(w1x - 95.7793)

%%%%%%%

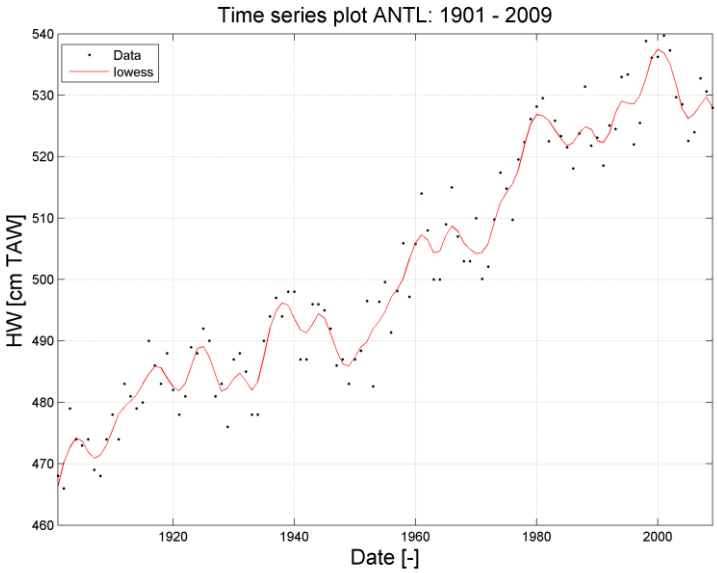
Date	HW	TrendData	HW_residuals	P.I.max	P.I.min	C.I.max	C.I.min
[-]	cm TAW	cm TAW	[-]	[-]	[-]	[-]	[-]
1980	528.19	526.4724	1.7176	537.4747	515.4702	529.5312	522.6173
1981	529.51	525.8448	3.6652	536.6884	515.0012	528.3552	522.4978
1982	522.51	524.9964	-2.4864	535.7154	514.2773	527.2334	522.0156
1983	525.89	524.0591	1.8309	534.7218	513.3964	526.1966	521.3741
1984	523.36	523.1764	0.18361	533.8554	512.4974	525.7202	520.5934
1985	521.53	522.4839	-0.95386	533.224	511.7437	525.2352	519.4023
1986	518.1	522.1011	-4.0011	532.9001	511.3021	525.1196	518.8332
1987	523.8	522.1068	1.6932	532.9183	511.2952	525.1755	518.8011
1988	531.42	522.5379	8.8821	533.2939	511.7818	525.3984	519.4213
1989	521.81	523.3864	-1.5764	534.0283	512.7445	525.9729	520.6782

Tijdens een analyse worden alle stappen weergegeven in het log-venster linksonder de grafische omgeving, maar dezelfde informatie wordt ook opgeslagen in een logbestand. Daarin worden, naast alle verwerkingsstappen, ook beschrijvende statistieken van de ingelezen bestanden opgeslagen.

3.4 BEPALING VAN TRENDBREUKEN

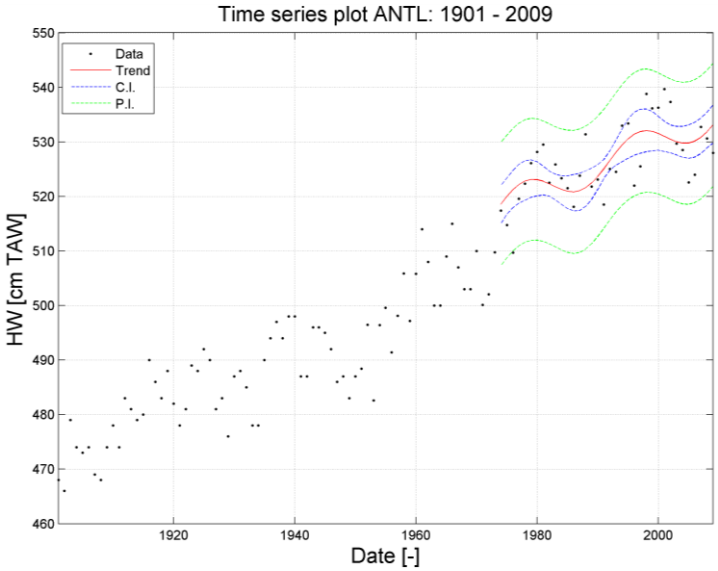
Het identificeren van een trendbreuk in een set jaargemiddelde data werd gedaan op zowel visuele als statistische manier. Er werd bijvoorbeeld bij de analyse van de waterstandsgegevens als volgt te werk gegaan:

1. Visuele identificatie van mogelijke trendbreuken in de datareeks (cf. Figuur3.5);



Figuur 3.5: Gemeten jaargemiddelde hoogwaterstanden te Antwerpen en LOWESS filter van 3 jaar.

2. Kalibreren van een trend op de dataset vanaf de visueel bepaalde trendbreuk, rekening houdend met een zowel een lineair stijgende (of dalende) component als een 18.6-jarige cyclus (waarvan de fase werd vastgelegd te Vlissingen voor heel het estuarium) (Figuur 3.6);

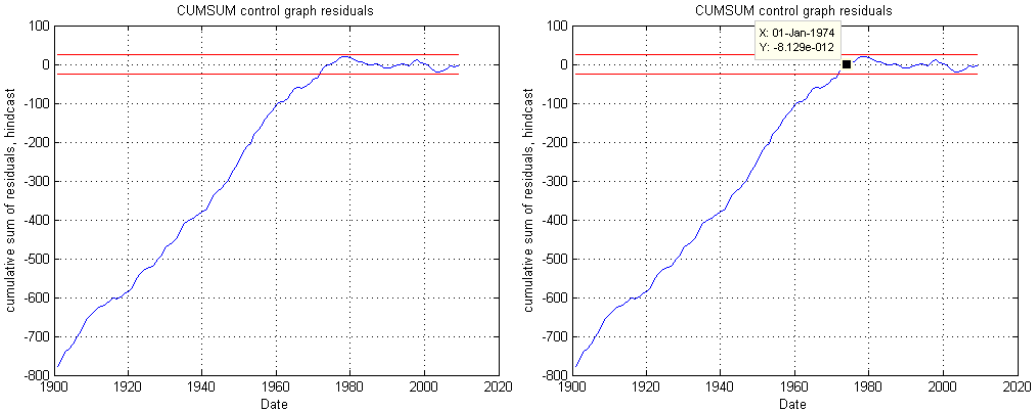


Figuur3.6: Kalibratie van regressiemodel met lineair stijgende component en 18.6-jarige nodale cyclus voor periode 1974-2009.

3. Extrapolatie van de trend naar het startjaar van de dataset (cf. Figuur 3.7);

Figuur3.7: Extrapolatie van het regressiemodel naar 1901.

4. Berekenen van de cumulatieve som van de residuen van de volledige dataset op de (geëxtrapoleerde) trend volgens de hindcasting methode (startjaar van de cumulatieve som is het eindjaar van de trend en de sommatie van de residuen wordt uitgevoerd richting het beginjaar van de dataset) (cf. Figuur 3.8, links);



Figuur3.8: Links: Cumulatieve som van de residuen van de data op het regressiemodel van 2009 tot 1901 (hindcast), of dus cumulatieve som controle grafiek. Rechts: indicatie van het jaar van de trendbreuk (één van de vroegste jaren dat nog binnen de controlegrenzen van de CUSUM controle grafiek ligt).

5. Controle dat het startjaar van de trend (of dus het jaar van de trendbreuk) binnen de controlegrenzen van de controle grafiek ligt. Niet enkel dat, maar ook dat het startjaar dichtbij het punt ligt waarvoor de cumulatieve som buiten één van haar controlegrenzen treedt (cf. Figuur 3.8, rechts).