SCHATTING VAN DE KANSVERDELING VAN SIGNIFICANTE GOLFHOOGTE

1. SAMENVATTING

De kansverdeling van de significante golfhoogte H wordt verondersteld overeen te komen met een Weibull verdeling. Omdat deze verdeling vooral de frequentie waarmee hogere waarden voorkomen goed voorstelt, wordt de verdeling enkel van toepassing veronderstelt boven een drempelwaarde h_d die op automatische wijze uit de geobserveerde waarden wordt gekozen

De Weibull verdeling wordt beschreven door twee parameters: u is een schaalfactor en komt overeen met de waarde die met 36.8% kans wordt overschreden. k is een vormfactor en de overschrijdingskans daalt exponentieel met de k^{de} macht van h. Figuur 1 toont enkele voorbeelden van de kansdichtheid voor u=100 en voor verschillende waarden van k. Voor k=1 is de significante golfhoogte exponentieel verdeeld. Voor lagere waarden heeft de verdeling een zwaardere staart, terwijl voor hogere waarden van k de staart korter wordt.

De schatting van de waarden u en k gebeurt door middel van de ML-methode ("maximum likelihood"-methode) toegepast op de geobserveerde waarden hoger dan de drempelwaarde en rekening houdend met het feit dat men het aantal waarden onder de drempelwaarde kent. Initiële schatters worden bepaald door middel van een kleinste-kwadraten benadering van de gegevens in een Weibull kwantielplot. In deze kwantielplot wordt het logarithm van de geordende waarden van de observaties (de orde-statistieken) uitgezet in functie van het logarithm van de theoretische Weibull-kwantiel die men zou vinden als k=u=1. Wanneer H effectief Weibull verdeeld is dan zou men verwachten dat deze plot goed benaderd wordt door een rechte lijn. Een voorbeeld van zulk een kwantielplot wordt getoond in Figuur 2. Dit voorbeeld betreft reële gegevens en toont dat inderdaad de Weibull verdeling enkel geldt voor de hogere waarden.

Wanneer de schattingsmethode wordt toegepast om de kansverdeling te bepalen voor een langere periode (bijvoorbeeld een jaar of een seizoen) dan dient men rekening te houden met het feit dat een over- of onderbemonstering van gegevens in bepaalde seizoenen of maanden kan leiden tot vertekende waarden (indien de kansverdeling verschilt van maand tot maand). Daarom worden aan de metingen gewichten toegekend zodat de steekproefomvang uniform is verdeeld over de verschillende seizoenen of maanden.

Om de standaardfout van de schattingen te bepalen dient men rekening te houden met de correlatie tussen opeenvolgende metingen van H. Zo kan het gebeuren dat de

hogere waarden van H die het resultaat domineren worden genoteerd in één enkele storm. In dit geval is de onzekerheid op het resultaat aanzienlijk hoger dan wanneer deze waarden zouden genoteerd zijn in verschillende stormen en dus effectief onafhankelijk zijn.

Om rekening te houden met deze onzekerheid wordt de spreiding van de schatters bepaald door middel van “non-parametric blocked bootstrapping”, NPB. In deze techniek wordt de schatting herhaaldelijk (B maal) toegepast op onafhankelijke lukrake selecties van tijdsblokken van de gegevens. Wanneer de gegevens gecorreleerd zijn dan is de spreiding die men noteert voor de geschatte parameters (de NPB-spreiding) groter dan de waarden die volgen uit de asymptotische formules voor een lukrake selectie van onafhankelijke gegevens (de AS-spreiding). Om tot consistente resultaten te komen, wordt daarom een equivalente (kleinere) steekproefomvang n_{EQ} berekent zodat de AS-spreiding overeenkomt met de NPB-spreiding voor deze steekproefomvang.

Eénmaal deze steekproefomvang gekend is, is het mogelijk om een parametrische bootstrapping (PB) methode toe te passen: in dit geval worden n_{EQ} onafhankelijke gegevens lukraak geselecteerd uit de Weibull verdeling en voor deze gegevens wordt de schattingsprocedure toegepast. De PB-methode laat toe na te gaan of de “goodness-of-fit” van de kwantielplot van de originele gegevens binnen de grenzen valt die men mag verwachten ten gevolge van toevallige fluctuaties wanneer de Weibull verdeling effectief van toepassing is. Is dit niet het geval dan wordt de drempelwaarde verhoogt om op die wijze tot een betere fit te komen.

Door herhaalde toepassing van de NPB en PB-methode wordt op deze wijze de drempelwaarde bepaald boven de welke de gefitte Weibull verdeling de gegevens goed benadert. De standaardfout van de schatters van u en k en hun correlatie wordt bepaald op basis van de PB-methode met een aangepaste steekproefomvang op basis van de NPB-spreiding. Deze standaardfouten worden ook gebruikt om (benaderende) betrouwbaarheidsintervallen op te stellen voor kwantielwaarden (waarden van H die met een bepaalde kans worden overschreden).

De procedure is toegepast op een aantal gesimuleerde resultaten om na te gaan of de correctie voor correlatie behoorlijk is en of de automatisch bepaalde drempelwaarde goed gekozen wordt. In beide gevallen werden goede resultaten genoteerd.

In dit rapport wordt de schattingsmethode en de verificatie van de schattingsmethode in detail uiteengezet. Dit gebeurt respectievelijk in Hoofdstukken 4 (schattingsmethode) en 5 (verificatie). Ter herinnering wordt eerst in Hoofdstuk 2 uiteengezet hoe men de kansverdeling van de golfhoogte kan gebruiken. Zo is het bijvoorbeeld niet evident om hieruit onmiddellijk de kans te bepalen dat een waarde H wordt overschreden in een bepaalde tijdsperiode. In Hoofdstuk 3 wordt de definitie

van de Weibull verdeling herhaald. Ter afsluiting, worden in Hoofdstuk 6 enkele toepassingen op reële gegevens getoond.

2. INTERPRETATIE EN GEBRUIK VAN DE RESULTATEN

2.1. MARGINALE KANSVERDELING

De marginale kansverdeling toont de relatieve frequentie waarmee verschillende significante golfhoogten worden geobserveerd. In de meest gangbare voorstelling toont men een curve die in functie van de waarde h de proportie voorstelt van de gegevens die deze waarde h overschrijdt. Men noemt dit een complementaire cumulatieve kansverdeling en duidt deze aan als $F_H^*(h)$. Voor eenvoud van notatie gebruiken we hierna echter ook p of p(h).

De proportie p is geldig voor een oneindig lange meetreeks van gegevens op de plaats waar de kansverdeling werd bepaald en voor de tijdsperiode waarop p(h) betrekking heeft. Zo kan men spreken van een jaarlijkse, seizoensafhankelijke (lente, zomer, winter, herfst) en maandelijkse (januari, februari, ...) kansverdelingen.

Belangrijk is bovendien dat de gegevens hetzij equidistant in de tijd hetzij lukraak worden gekozen binnen de voorgestelde periode (jaar, seizoen of maand). Dit is in het bijzonder belangrijk voor de jaarlijkse en seizoensafhankelijke marginale kansverdelingen. Meestal zal p(h) immers veranderen van maand tot maand en voor een samengestelde kansverdeling (een seizoen of een jaar) wordt verondersteld dat de gegevens worden verzameld proportioneel tot de tijdsduur van de verschillende maanden.

De proportie p(h) duidt aan hoe dikwijls een bepaalde waarde h gemiddeld wordt overschreden voor een bepaald aantal waarnemingen (indien deze uniform zijn verdeeld in de tijd binnen de periode) maar geeft geen (rechtstreekse) aanduiding met welke kans een bepaalde waarde h wordt overschreden binnen een bepaalde observatieperiode.

Dit is eenvoudig te begrijpen als volgt. Stel dan men weet dat voor 10% van de H metingen een waarde van 1.5 meter overschrijdt. Voor 10,000 metingen mag men dan inderdaad stellen dat gemiddeld 1,000 metingen de waarde 1.5 meter zullen overschrijden. De eventuele variatie op dit aantal overschrijdingen en de spreiding in de tijd van de overschrijdingen is echter functie van de tijdstap tussen de verschillende metingen en de correlatie tussen de meetwaarden.

Stel bijvoorbeeld dat 100 metingen sequentieel worden uitgevoerd met een tijdstap van 1/100 seconde. Het is zeer waarschijnlijk dat dan ofwel alle gegevens de 1,5 meter zullen overschrijden ofwel niet. De kans dat de waarde 1,5 meter wordt

overschreden binnen de observatieperiode van 1 seconde is in zulk geval 10%. Stel anderzijds dat de elk van de 100 metingen op een willekeurig tijdstip binnen opeenvolgende jaren worden uitgevoerd en dus een periode van 100 jaar bestrijken. In zulk geval verwacht men ieder van de metingen met een kans 10% de waarde 1,5 meter overschrijdt en het is redelijk deze kansen onafhankelijk te veronderstellen. De kans dat een bepaald aantal van deze metingen 1,5 meter overschrijdt wordt dan bepaald door middel van de binomiaalverdeling. Bijvoorbeeld de kans dat geen enkele van de metingen 1,5 meter overschrijdt is $0.9^{100}=2.7\text{x}10^{-5}$ en dus vrijwel nihil.

Hoewel men dus weet dat gemiddeld 10% van de metingen 1,5 meter overschrijdt kan men hieruit niet rechtstreeks de kans bepalen waarmee de waarde 1,5 meter wordt overschreden binnen een bepaalde periode.

Een benaderende rekenwijze is echter mogelijk. Volgende notatie wordt hiertoe geïntroduceerd:

- T(h) is de gemiddelde duur waarbij de significante golfhoogte de waarde h overstijgt;
- S is het tijdsinterval voor hetwelk men de kans van overschrijdingen wenst te bepalen.

De gemiddelde totale tijdsduur tijdens dewelke de waarde h wordt overschreden binnen S komt overeen met:

$$E[\Sigma_S T_{>h}] = S \times p(h) \tag{1}$$

Het verwachtte aantal A dat H de waarde overschrijdt bij niet-opeenvolgende metingen is dan

$$E[A_{>h}] = \frac{S \times p(h)}{T(h)} \tag{2}$$

De gemiddelde tijd of terugkeerperiode tussen zulke niet-opeenvolgende overschrijdingen, gemeten van het begin van 1 overschrijding tot het begin van de volgende overschrijding is

$$E[R_{>h}] = \frac{T(h)}{p(h)} \tag{3}$$

Wanneer de verwachtte tijd tussen opeenvolgende overschrijdingen aanzienlijk groter is dan de tijd van de overschrijding en men geen regelmaat of groepering van de overschrijdingen verwacht dan is het redelijk om te veronderstellen dat de tijdstippen van de verschillende overschrijdingen onafhankelijk zijn en worden beschreven door

een Poisson proces. Onder die veronderstelling volgt dat de kans dat h minstens éénmaal wordt overschreden binnen de tijdspanne S overeenkomt met

$$P(H > h \text{ in } S) = p(h) + (1 - p(h))\left(1 - \exp^{-\frac{p(h)S}{T(h)}}\right) \quad (4)$$

De eerste term in de rechterzijde verwijst naar de kans dat de eerste waarde van de observatieperiode reeds groter is dan h. De tweede term verwijst naar de kans dat dit niet zo is maar dat ten minste 1 overschrijding gebeurt binnen de observatieperiode S.

Stel bijvoorbeeld dat voor 1,5 meter de gemiddelde tijdsduur van een overschrijding overeenkomt met 24 uur terwijl p(h) nog steeds overeenkomt met 10%. Voor een observatieperiode van 1 seconde is de tweede term in Vergelijking (4) te verwaarlozen en de kans van overschrijding is 10%. Voor een periode van 1 jaar is het verwachtte aantal overschrijdingen gelijk aan 36.5 en vindt men een waarde die zo goed als gelijk is aan 1.

In praktische toepassingen is de observatieperiode waarvan sprake één of meerdere grootteordes groter dan de tijdsduur van de overschrijding en is de kans van overschrijding p(h) bijzonder laag zodat men binnen de observatieperiode gemiddeld minder dan één overschrijding verwacht waar te nemen. In zulk geval wordt Vergelijking (4) nauwkeurig benaderd door de uitdrukking:

$$P(H > h \text{ in } S) = p(h)\frac{S}{T(h)} \quad (5)$$

waarbij dient benadrukt te worden dat deze vergelijking enkel een nauwkeurige benadering is wanneer

$$p(h)\frac{S}{T(h)} < 0.1 \text{ en } T(h) < 0.1 \cdot S \quad (6)$$

Een N-jaarlijkse ontwerpwaarde komt overeen met de waarde h waarvoor de gemiddelde tijd tussen opeenvolgende overschrijdingen overeenkomt met N jaar. Zoals eerder aangeduid bedraagt deze gemiddelde tijd T(h)/p(h). Daaruit volgt dat de N jaarlijkse ontwerpwaarde h_N is gedefineerd als de waarde van h die met kans T(h)/N wordt overschreden in de marginale kansverdeling:

$$p(h_N) = \frac{T(h)}{N} \quad (7)$$

waarbij men T(h) in jaareenheden dient uit te drukken.

Vergelijking (7) is enkel van toepassing voor de jaarlijkse marginale kansverdeling. De aanpassing die nodig is in het geval van maandelijkse of seizoensafhankelijke kansverdelingen wordt besproken in volgende paragraaf.

De berekening van de overschrijdingskans van de golfhoogte veronderstelt dus de kennis van de gemiddelde tijdsduur T(h). De schatting van deze tijdsduur wordt in dit rapport niet besproken maar zal aan bod komen in een latere studie. Voor extreem hoge waarden van h is de resolutie waarmee men de gemiddelde tijdsduur kan bepalen beperkt tot de tijdsresolutie van de metingen: de significante golfhoogte is immers een statistische parameter die de zeetoestand beschrijft binnen een bepaald tijdsinterval (typisch rond de 20 minuten). Wanneer de significante golfhoogte een bepaalde waarde éénmalig overschrijdt dient men T(h) dan ook gelijk te stellen aan de tijdstap van de meting. Dit verklaart meteen de traditionele rekenwijze van de N-jaarlijkse ontwerpwaarde waarbij men de kwantielwaarde van h gebruikt waarvoor de overschrijdingskans overeenkomt met $(\Delta t)/N$ waarbij Δt het tijdsinterval voorstelt tussen opeenvolgende metingen van H en dient uitgedrukt te worden in jaren:

$$p(h_N) = \frac{\Delta t}{N} \quad \text{voor} \quad T(h_N) < \Delta t \tag{8}$$

Hoewel de keuze van T(h) de overschrijdingskans beïnvloedt en dus rechtstreeks de berekening van N-jaarlijkse ontwerpwaarden voor de significante golfhoogte wijzigt, leidt dit niet noodzakelijkerwijze tot reële veranderingen in het ontwerp. Bij werkelijke ontwerpberekeningen is immers niet alleen de ontwerpwaarde, maar ook de duur van overschrijding en de gemiddelde tijd tussen opeenvolgende golfhoogten (de “zero-crossing period”) van belang. Enkel met behulp van deze gegevens kan men het aantal golven berekenen en een verwachtte piekwaarde voor individuele golven bepalen. Een ontwerpwaarde voor de significante golfhoogte van 2.5 meter gebaseerd op een gemiddelde overschrijdingsduur van 20 minuten is bijgevolg niet noodzakelijk ernstiger dan een ontwerpwaarde van 2 meter voor een gemiddelde overschrijdingsduur van 3 uur. In die zin kan men de keuze van T(h) tot op zekere hoogte als een conventie beschouwen en is het gebruik van de tijdstap die door één enkele meting wordt overschreden te verdedigen ook bij lagere waarden van h. Wat ook de berekeningswijze moge zijn, het is in ieder geval belangrijk om de gebruikte waarde van de overschrijdingsduur T(h) te vermelden.

Overigens is het zo dat de schatting van de overschrijdingskans binnen een bepaalde periode beter en meer rechtstreeks kan bepaald worden door een rechtstreekse studie van de piekwaarden van H die worden genoteerd tijdens overschrijdingen boven een bepaald drempelniveau. Deze waarde wordt dan representatief

verondersteld voor de tijdsduur tussen opeenvolgende metingen (Δ t). Zulk een POT ("peak-over-treshold") schatting is het onderwerp van een volgende studie.

2.2. VOORWAARDELIJKE KANSVERDELING

In een aantal gevallen wordt de kansverdeling van H bepaald op basis van die gegevens waarbij een andere variabele (bv. windrichting) binnen een bepaald interval valt. We noemen dit een voorwaardelijke kansverdeling. Ook de maandelijkse en seizoensafhankelijke kansverdelingen kunnen beschouwd worden als voorwaardelijke verdelingen, hoewel de selectie in dit geval op basis van het tijdstip eerder dan op basis van een andere waarneming gebeurt.

De proportie p(h) duidt nu aan hoe dikwijls een bepaalde waarde h gemiddeld wordt overschreden voor een bepaald aantal waarnemingen, indien deze uniform zijn verdeeld in de tijd binnen de periode wanneer de gestelde voorwaarde is voldaan (bv. de windrichting komt overeen met een bepaalde waarde, of het tijdstip valt in een bepaalde maand).

Net zoals voor de marginale kansverdeling is p(h) geen (rechtstreekse) aanduiding van de kans met dewelke een bepaalde waarde h wordt overschreden binnen een bepaalde observatieperiode.

Ten eerste is het, zoals eerder besproken, in principe nodig om de gemiddelde tijdsduur van een overschrijding te bepalen tenzij men voor extreme hoge waarden deze tijdsduur mag gelijk veronderstellen aan de tijdstap tussen de metingen Δ t.

Ten tweede is het nodig aan te duiden hoe vaak de gestelde voorwaarde zich voordoet binnen de totale periode S waarvoor men de kans wenst te berekenen. We duiden deze frequentie aan als p_V. De proportie kan rechtstreeks geschat worden op basis van de gegevens door het aantal keer dat de voorwaarde is voldaan (n_V) te delen door de totale steekproefomvang n:

$$\hat{p}_V = \frac{n_V}{n} \tag{9}$$

In het geval van maandelijkse of seizoens-afhankelijke kansverdelingen is de proportie p_V exact gekend en komt overeen met respectievelijk 1/12 en ¼.

Voor de bepaling van de N-jaarlijkse voorwaardelijke ontwerpwaarde dient men nu onderscheid te maken tussen volgende 2 alternatieve resultaten:

1. ofwel heeft de N-jaarlijkse ontwerpwaarde betrekking tot kalendertijd en in dit geval gebruikt men de kwantielwaarde die overeenkomt met $T(h)/(\hat{p}_V N)$ waarbij men T(h) eventueel kan vervangen door Δt:

$$p(h_N) = \frac{T(h)}{\hat{p}_V N} \quad \text{voor een N-jaarlijkse terugkeerperiode in kalendertijd} \qquad (10)$$

2. ofwel heeft de N-jaarlijke ontwerpwaarde betrekking tot de reële tijd gedurende dewelke de voorwaarde is voldaan. In dit geval gebruikt men T(h)/N of $(\Delta t)/(N)$:

$$p(h_N) = \frac{T(h)}{N} \text{ voor een N-jaarlijkse terugkeerperiode tijdens de voorwaarde} \qquad (11)$$

Zulke ontwerpwaarden zijn uiteraard hoger dan in voorgaande formulering

Rapportering van de ontwerpwaarden volgens kalendertijd, zoals in Vergelijking (10), is het meest gebruikelijk omdat zulke waarden meteen kunnen worden geassocieerd met een bepaalde overschrijdingskans binnen één kalenderjaar (merk op dat de gestelde voorwaarde niet noodzakelijk is voldaan). Formulering 2 is anderzijds meer geschikt voor toepassingen waarbij men risico's dient in te schatten en men weet (bv. hetzij op basis van waarnemingen, hetzij op basis van voorspellingen) dat die voorwaarde is voldaan. Deze formuleringen is ook meer geschikt om de relatieve ernst van verschillende voorwaarden te vergelijken.

Tenslotte dient vermeld dat het mogelijk is om niet-voorwaardelijke N-jaarlijkse ontwerpwaarden af te leiden op basis van de voorwaardelijke kansverdelingen wanneer die voorwaarden collectief volledig en onderling onafhankelijk zijn (bv. alle mogelijke windrichtingen). Duiden we deze voorwaarden aan door de index k, k=1 tot K, dan komt de samengestelde marginale kansverdeling p(h) overeen met:

$$p(h) = \sum_{k=1}^{K} \hat{p}_{V,k} p_k(h) \qquad (12)$$

waarbij $p_k(h)$ de voorwaardelijke kansverdeling voorstelt. N-jaarlijkse ontwerpwaarden komen dan overeen met die waarden van h waarvoor p(h)=T(h)/n. In principe dient deze ontwerpwaarde overeen te komen met de ontwerpwaarde die wordt afgeleid op basis van de marginale kansverdeling. Dit is echter niet noodzakelijk het geval omwille van schattingsfouten en de inconsistentie van de veronderstelling van de verdelingsvorm voor de voorwaardelijke kansverdelingen en de marginale kansverdeling. De som van verschillende Weibull verdelingen is immers niet Weibull verdeeld, tenzij de parameters u en k dezelfde zijn. Wanneer de voorwaardelijke verdelingen sterk verschillen, is Vergelijking (12) in theorie een betere basis voor de

berekening van de ontwerpwaarden vermits in deze formulering rekening wordt gehouden met de nonstationariteit van de verdelingsvorm. In de praktijk is het echter zo dat het resultaat in Vergelijking (12) leidt tot een grotere statistische fout omdat de deelresultaten (de voorwaardelijke kansverdelingen) worden berekend op basis van een geringer aantal gegevens.

2.3. EMPIRISCHE FREQUENTIEVERDELING EN GEFITTE RESULTATEN

De kansverdeling p(h) dient uiteraard te worden afgeleid op basis van de waargenomen gegevens (binnen een bepaalde periode en eventueel voor een bepaalde voorwaarde). We veronderstellen hierna dat T gegevens zijn waargenomen en duiden de overeenkomstige waarden aan als h_t waarbij de index t varieert van 1 tot T.

2.3.1. EMPIRISCHE FREQUENTIEVERDELING

Binnen het bereik van de metingen kan de waarde p(h) rechtstreeks geschat worden door het aantal waarden $h_t>h$ te berekenen en te delen door de steekproefomvang. Omdat de golfhoogte een continue variabele is, is de kans dat H>h gelijk aan de kans dat H≥h. Nochtans leidt het tellen van het aantal overschrijdingen tot een verschillend resultaat ter plaatse van de waargenomen waarden naargelang men de ene of andere definitie gebruikt. Een eenvoudige correctie hiervoor verloopt als volgt.

Duiden we de geordende waarden van h aan als $h_{(t)}$. $h_{(1)}$ is bijvoorbeeld de minimum waarde. Wanneer verschillende waarnemingen dezelfde waarde hebben (omwille van afrondingen) dan worden deze samengevoegd. Men noemt zulke waarden “ties”. Het aantal metingen voor zulk een waarde wordt verder aangeduid als $n_{(t)}$. Voor individuele waarden is $n_{(t)}=1$. Het totaal aantal orde-statistieken is niet gelijk aan T wanneer gelijke waarden in de steekproef voorkomen. Dit totaal aantal wordt daarom aangeduid door een nieuw symbool T*.

De proportie waarmee binnen de steekproef een orde-statistiek $h_{(t)}$ wordt overschreden wordt dan geschat als:

$$\hat{p}(h_{(t)}) = \frac{0.5n_{(t)} + \sum_{t=(t)+1}^{T^*} n_{(t)}}{N} \qquad (13)$$

Men noemt dit ook de Blomscore van de orde-statistiek. Het gebruik van de waarde $0.5n_{(t)}$ volgt uit het feit dat 50% van de waarden waarvoor $h=h_{(t)}$ wordt verondersteld groter te zijn dan $h_{(t)}$. Een theoretisch betere correctie kan worden voorgesteld wanneer men de werkelijke kansverdeling p(h) kent, maar voor een voldoende grote steekproefomvang en een gering aantal “ties” leidt dit tot vrijwel identieke resultaten.

De empirische frequentieverdeling wordt dan bekomen door de schattingen van p(h) uit te zetten voor iedere orde-statistiek en tussen deze waarden een linear verloop te veronderstellen.

Indien men binnen de observatieperiode nonstationariteit vermoedt van de kansverdeling (bv. afhankelijkheid van maanden of seizoenen) dan is het belangrijk dat de verschillende tijdscategorieën (maanden of seizoenen) proportioneel worden bemonsterd tot de tijdsduur van deze categoriën. Indien de gegevens equidistant worden gemeten in de tijd dan is dit in principe het geval. Maar omdat tengevolge van instrumentdefecten of andere redenen de gegevens kunnen ontbreken tijdens sommige periodes is het nodig om dit na te gaan en eventueel een correctie uit te voeren. Dit gebeurt als volgt.

Veronderstellen we K verschillende tijdscategoriën met een relatieve tijdsduur die wordt uitgedrukt door de serie waarden r_k (bv. het aantal dagen in iedere maand or seizoen). Voor iedere tijdscategorie kan nu op basis van de originele metingen h_t het aantal metingen worden bepaald. Deze worden aangeduid als f_k. Om tot een uniforme verdeling van de metingen binnen de tijdsperiode te komen wordt nu aan een meting in de tijdscategorie k een fictief aantal metingen toegekend n_t:

$$n_t = \frac{r_k}{f_k}\frac{N}{R} \quad \text{waarbij} \quad R = \sum_{k=1}^{K} r_k \tag{14}$$

en k overeenkomt met de tijdscategorie van de t^{de} meetwaarde.

Vervolgens worden de gegevens geordend en voor de verschillende orde-statistieken worden de fictieve aantallen n_t samengevoegd tot het fictief aantal $n_{(t)}$ waarmee de orde-statistiek werd vastgesteld.

Wanneer de correctiefactor $(r_kN)/(f_kR)$ zeer groot is dan wordt het resultaat in sterke mate beïnvloedt door de (relatief weinige) meetwaarden in de overeenkomstige tijdscategorie en is het eindresultaat minder betrouwbaar. Hoewel deze vermindering van de nauwkeurigheid in rekening wordt gebracht bij het bepalen van het betrouwbaarheidsinterval voor de geschatte verdeling, is het zinvol de correctiefactor te beperken tot een maximum waarde. Daarom worden eerst de correctiefactoren c_k berekent voor iedere tijdscategorie en wordt, indien nodig, deze waarde beperkt tot 3. Dan geldt:

$$c_k = \min\left(3, \frac{r_k}{f_k}\frac{N}{R}\right) \quad \text{waarbij} \quad R = \sum_{k=1}^{K} r_k \tag{15}$$

en

$n_t = c_k$ waarbij k de tijdscategorie van de t^{de} meting voorstelt (16)

Wanneer deze beperking van toepassing is, dan is de nieuwe totale steekproefomvang $N^* = \sum_{t=1}^{T} n_t = \sum_{t=1}^{T^*} n_{(t)}$ kleiner dan het aantal effectief waargenomen waarden. Zowel de toegepaste correctiefactoren als de uiteindelijke afwijking ten opzichte van de vooropgestelde verdeling over de verschillende tijdscategoriën maakt echter deel uit van het eindresultaat.

2.3.2. GEFITTE KANSVERDELING

De empirische frequentieverdeling is een beschrijvende statistiek die, afgezien van de correctie voor de tijdscategoriën en de aanpassing van de frequenties voor de discrete natuur van de waarnemingen, de gegevens voorstelt zoals ze zijn waargenomen.

Het bereik van de empirische frequentieverdeling is bijgevolg noodzakelijkerwijze beperkt tot de minimum en maximum waargenomen waarden. Bovendien is de betrouwbaarheid van dit resultaat gering voor de schatting van p(h) bij hoge waarden vermits het aantal overschrijdingen in dit geval klein is. Hetzelfde geldt overigens bij lage waarden voor de schatting van de onderschrijdingskans, 1-p(h), die men uit dit resultaat kan afleiden.

Om tot meer betrouwbare resultaten te komen en extrapolatie mogelijk te maken is het daarom nodig om een bepaalde vorm voor de kansverdeling te veronderstellen. Voor de significante golfhoogte wordt verondersteld dat de verdeling, alleszins voor de hogere waarden, overeenkomt met een Weibull verdeling. De vorm en eigenschappen van deze verdeling worden besproken in volgend hoofdstuk. Voor het fitten van deze kansverdeling wordt uitgegaan van de orde-statistieken en de gecorrigeerde aantallen $n_{(k)}$ zoals hiervoor beschreven zodat de eventuele non-uniformiteit van de meetfrequentie in de tijd in rekening wordt gebracht. De veronderstelling dat de gegevens Weibull verdeeld zijn wordt verder in rekening gebracht door de empirische frequentieverdeling uit te zetten op Weibull waarschijnlijkheidspapier, ook een Weibull Q-Q plot genoemd. In deze voorstelling wordt het logarithm van de orde-statistiek uitgezet in functie van een getransformeerde Blomscore, zoals verder wordt uitgelegd in volgend hoofdstuk. Indien de gegevens Weibull verdeeld zijn, dan verwacht men een lineair verloop tussen de uitgezette waarden. Visueel kan men bijgevolg onmiddellijk nagaan in hoeverre de veronderstelling is voldaan. Bovendien bekomt men bij deze voorstelling, als het verloop inderdaad lineair is, meer nauwkeurige resultaten bij de interpolatie van de empirische frequentieverdeling.

3. EIGENSCHAPPEN VAN DE WEIBULL VERDELING

3.1. VORM VAN DE VERDELING

De significante golfhoogte H wordt Weibull verdeeld verondersteld boven een drempelwaarde h_d. De complementaire cumulatieve kansverdeling wordt dan beschreven als:

$$P(H > h) = \exp\left[-\left(\frac{h}{u}\right)^k\right] \qquad \text{voor } h > h_d \tag{17}$$

waarbij u en k twee parameters voorstellen. u is een schaalfactor en komt overeen met de significante golfhoogte die in 37% van de gevallen wordt overschreden. k is een vormfactor die de zwaarte van de staart bepaald. Voor het bijzondere geval k=1, vindt men de exponentiele verdeling. Naarmate k verhoogt verlaagt de frequentie waarmee extreem hoge waarden voorkomen.

De overeenkomstige kansdichtheid is van de vorm:

$$f_H(h) = k\left(\frac{h}{u}\right)^{k-1} \exp\left[-\left(\frac{h}{u}\right)^k\right] \qquad \text{voor } h > h_d \tag{18}$$

Enkele voorbeelden van de vorm van de kansdichtheid worden voorgesteld in Figuur 1.

3.2. WEIBULL Q-Q PLOT

Noemen we h_p de p-bovenkwantielwaarde (d.w.z. de waarde die met kans p wordt overschreden) dan volgt uit (17) dat:

$$k\left[\log(h_p) - \log(u)\right] = \log\left[-\log(p)\right] \quad \text{voor } h_p > h_d \tag{19}$$

of

$$\log(h_p) = \log(u) + \frac{1}{k}\log\left[-\log(p)\right] \quad \text{voor } h_p > h_d \tag{20}$$

-log(p) komt overeen met de p-bovenkwantielwaarde van een Weibull verdeling waarvoor u=1 en k=1. Deze kwantielwaarde noemen we de standaard-Weibull bovenkwantiel. Vergelijking (20) toont dat het logarithm van de kwantielwaarden voor een Weibull verdeling met ongekende parameters u en k linear varieert in functie van het logarithm van de standaard-Weibull bovenkwantielen. Het intercept van deze lijn komt overeen met log(u), terwijl 1/k de helling bepaalt.

Stel nu dat op basis van een meetreeks de orde-statistieken $h_{(t)}$ zijn gekend voor t=1,T*. Het aantal waarnemingen $h_{(t)}$ in de meetreeks is $n_{(t)}$ en de overeenkomstige empirische schatting van de overschrijdingsfrequentie wordt aangeduid als $\hat{p}_{(t)}$. Wanneer men het logarithm van de orde-statistieken $\log(h_{(t)})$ uitzet in functie van het logarithm van de standaard-Weibull bovenkwantielen voor de verschillende schattingen $\hat{p}_{(t)}$ dan verwacht men een linear verloop boven de drempelwaarde h_d. Zulk een grafiek noemt men een Weibull Q-Q plot (zie Figuur 2 voor een voorbeeld). Vermits het intercept en de helling van het linear verloop bepaalt wordt door de parameters u en k, kan men op basis van deze grafiek ook de waarde van u en k schatten zoals verder wordt uiteengezet in volgend hoofdstuk.

4. SCHATTING VAN DE WEIBULL VERDELING

4.1. ALGEMENE BESCHRIJVING VAN DE PROCEDURE

De schattingsprocedure die hier wordt uiteengezet is van toepassing op een meetreeks van h waarden, waarbij voor iedere waarde h de overeenkomstige tijdscategorie k is aangeduid. Het relatieve aantal metingen dat men in elke tijdscategorie verwacht is eveneens gekend en wordt aangeduid door r_k.

In Paragraaf 2.3 is reeds uiteengezet hoe men op basis van zulke gegevens de orde-statistieken $h_{(t)}$, t=1,T* en de overeenkomstige gecorrigeerde aantallen $n_{(t)}$ kan bepalen. Ter herinnering: de correctie zorgt ervoor dat de bemonstering van de h waarden uniform is in de tijd en dus geen vertekend resultaat oplevert indien een deelperiode over- of onderbemonsterd is. In de procedure wordt tevens gebruik gemaakt van de geschatte empirische overschrijdingskans, die wordt aangeduid als $p_{(t)}$.

De originele meetreeks wordt tevens onderverdeeld in consecutieve tijdsblokken van 1 maand. In tegenstelling tot de tijdscategoriën behoren gegevens gemeten in dezelfde maand maar in verschillende jaren tot verschillende tijdsblokken. Het totaal aantal tijdsblokken wordt hierna genoteerd als M. Deze onderverdeling wordt gebruikt om lukraak de M tijdsblokken van de gegevens te herbemonsteren in de niet-parametrische bootstrapping. Door tijdsblokken te herbemonsteren eerder dan individuele gegevens behoudt men de correlatiestructuur in de aldus gesimuleerde tijdreeks. Deze correlatie kan de spreiding van de schatters sterk beïnvloeden: een sterke positieve correlatie tussen opeenvolgende gegevens vermindert de effectieve steekproefomvang van onafhankelijke gegevens die beschikbaar zijn om de schatters te bepalen.

De schattingsprocedure omvat een iteratie om de drempelwaarde te bepalen waarboven de Weibull verdeling de empirische verdeling goed benadert. De volledige procedure omvat dan volgende stappen:

1. de drempelwaarde h_d wordt initiëel gelijk veronderstelt aan 0;

2. orde-statistieken en de overeenkomstige gecorrigeerde aantallen worden bepaald op basis van de meetreeks;

3. de waarde van u en k wordt geschat voor de gekozen drempelwaarde. Bij deze stap worden enkel waarden boven de drempelwaarde gebruikt. Het totaal aantal waarnemingen dat gelijk is

of kleiner dan de drempelwaarde wordt eveneens in rekening gebracht. Deze stap omvat volgende deelberekeningen:

a. schatting van de kleinste-kwadraten schatters ("least-squares" of LS-schatters) van u en k op basis van de Weibull Q-Q plot;

b. iteratie om de ML-schatters (waarden die de L-funktie maximaliseren) van u en k te berekenen. Startwaarden bij de iteratie komen overeen met de LS-schatters;

c. berekening van een "goodness-of-fit" statistiek of G-statistiek. De G-statistiek die gebruikt wordt in deze procedure komt overeen met de correlatie tussen de gefitte bovenkwantielen voor de verschillende orde-statistieken en de waargenomen orde-statistieken.

4. niet-parametrische schatting van de spreiding van de schatters: uit de M tijdsblokken worden M tijdsblokken lukraak geselecteerd met teruglegging en stappen 1 tot 3 worden herhaald. Deze selectie wordt B maal uitgevoerd en voor de B schattingen van u en k wordt de standaarddeviatie bepaald van de schattingen: $\sigma_{\hat{u}}^{NPB}$ en $\sigma_{\hat{k}}^{NPB}$;

5. door vergelijking van de eerder bekomen standaarddeviaties hetzij (bij de eerste iteratie van deze stappen) met de assymptotische standaarddeviaties van de schatters, $\sigma_{\hat{u}}^{AS}$ en $\sigma_{\hat{k}}^{AS}$, hetzij (bij de volgende iteraties) met de standaarddeviaties van de schatters bepaald door parametrische bootstrapping in de vorige iteratie, $\sigma_{\hat{u}}^{PB}(N_{eq})$ en $\sigma_{\hat{k}}^{PB}(N_{eq})$, wordt een nieuwe equivalente steekproefomvang N_{eq} bepaald;

6. parametrische schatting van de spreiding van de schatters gebruikmakend van de benaderende equivalente steekproefomvang: in deze stap worden N_{eq} onafhankelijke gegevens gesimuleerd uit de geschatte Weibull verdeling. Deze simulatie wordt B maal uitgevoerd en op basis van de B schattingen wordt de standaarddeviatie bepaald van de schattingen: $\sigma_{\hat{u}}^{PB}(N_{eq})$ en $\sigma_{\hat{k}}^{PB}(N_{eq})$;

7. de equivalente steekproefomvang wordt herberekend om een betere overeenkomst te maken tussen de niet-parametrische en de parametrische schatting van de standaarddeviaties;

8. herhaling van de parametrische schatting van de spreiding van de schatters, ditmaal gebruikmakend van de verbeterde equivalente steekproefomvang. Naast de standaarddeviaties van de schattingen wordt ook het gemiddelde en de standaarddeviatie van de G-statistiek bepaald: $\mu_G^{PB}(N_{eq})$ en $\sigma_G^{PB}(N_{eq})$;

9. beoordeling van de G-statistiek en keuze van een volgende drempelwaarde: in deze stap wordt nagegaan of de fit voldoet en indien niet dan wordt de iteratie verder gezet bij stap 2 met een nieuwe h_d waarde;

10. wanneer de drempelwaarde is bepaald waarbij de fit voldoet (of zo goed mogelijk is) dan wordt de parametrische bootstrapping B' maal herhaald om meer nauwkeurige resultaten af te leiden. Uiteindelijke resultaten betreffende de spreiding van de schattingen en de G-statistiek zijn gebaseerd op de B+B' simulaties. Naast de standaarddeviaties van de schattingen u en k worden ook benaderende betrouwbaarheidsintervallen berekend voor de gefitte kwantielwaarden die overeenkomen met de overschrijdingskansen $p_{(t)}$.

In de volgende paragrafen worden de rekenmethodes die gebruikt worden in deze stappen verder toegelicht. Meer bepaald:

1. de berekening van LS-schatters van u en k;
2. de berekening van ML-schatters van u en k;
3. de berekening van de G-statistiek;
4. uitvoeren van de niet-parametrische bootstrapping;
5. uitvoeren van de parametrische bootstrap;
6. berekening van de equivalente steekproefomvang;
7. beoordeling van de G-statistiek en keuze van de drempelwaarde;
8. berekening van betrouwbaarheidsintervallen voor de kwantielwaarden van H.

In een laatste paragraaf wordt toegelicht hoe de gemeten waarden, indien nodig, dienen afgerond te worden.

4.2. BEREKENING VAN LS-SCHATTERS VAN U EN K

LS-schatters van u en k worden bepaald op basis van Vergelijking (20) waarbij h_p wordt vervangen door de orde-statistieken $h_{(t)}$ en p wordt vervangen door $p_{(t)}$ voor alle $h_{(t)}$ die groter zijn dan of gelijk aan h_d. Voor de schatting wordt een gewogen kleinste-kwadraten techniek gebruikt: het gewicht voor ieder punt $h_{(t)}$ wordt gelijkgesteld aan het aantal waarnemingen $n_{(t)}$, behalve voor de drempelwaarde h_d waar het gewicht wordt gelijkgesteld aan het aantal waarnemingen dat kleiner is dan of gelijk aan de drempelwaarde.

4.3. BEREKENING VAN ML-SCHATTERS VAN U EN K

De L-funktie ("Likelihood"-functie) van de gegevens boven de drempelwaarde h_d komt overeen met:

$$\ell = \binom{n}{n_d}\left[1-e^{-\left(\frac{h_d}{u}\right)^k}\right]^{n_d}\prod_{t=t_{d+1}}^{T^*}\left[k\left(\frac{\left(h_{(t)}\right)^{k-1}}{u^k}\right)e^{-\left(\frac{h_{(t)}}{u}\right)^k}\right]^{n_{(t)}} \qquad (21)$$

waarbij n_d overeenkomt met het aantal waarden dat kleiner of gelijk is aan de drempelwaarde en t_d de orde aanduidt van de statistiek die kleiner of gelijk is aan de drempelwaarde.

Het zoeken naar de waarden van u en k die de L-funktie maximaliseren (de ML-schatters) gebeurt door middel van een dubbele iteratie. Voor gegeven k is het mogelijk om iteratief de overeenkomstige schatter van u te berekenen en voor deze waarde de correctie te berekenen die aan k moet worden toegebracht. Rekening houdend met de waarde en het teken van deze correctie wordt dan een nieuwe k-waarde bepaald en de schatting van u herhaald tot de correctie voor k voldoende klein is. Initiële waarden voor de schatters worden gelijkgesteld aan de LS-schatters.

Verdere details betreffende de ML-schatters worden uiteengezet in Appendix A.

4.4. BEREKENING VAN DE G-STATISTIEK

De G-statistiek dient een maat te zijn in hoeverre de gefitte Weibull verdeling de oorspronkelijke empirische frequentieverdeling benaderd bij en boven de drempelwaarde. Vermits fouten op de werkelijk voorspelde waarden eerder dan op de logarithmisch getransformeerde variabelen van belang zijn in de praktijk werd in de schattingsprocedure gekozen voor een statistiek die het verschil tussen de werkelijke en gefitte orde-statistieken samenvat.

Bij een perfecte fit zijn de gefitte kwantielwaarden $q_{(t)}$ die men vindt voor de overschrijdingskansen $p_{(t)}$ identiek aan de orde-statistieken $h_{(t)}$. De correlatie tussen $q_{(t)}$ en $h_{(t)}$ is bijgevolg een directe maatstaf voor de “goodness-of-fit”. Deze correlatie wordt berekend rekening houdend met het aantal waarnemingen dat tot elke orde-statistiek behoort als volgt:

$$\rho = \frac{n_d(q_{(t_d)}-\bar{q})(h_{(t_d)}-\bar{h}) + \sum_{t=t_d+1}^{T^*} n_{(t)}(q_{(t)}-\bar{q})(h_{(t)}-\bar{h})}{\sqrt{n_d(q_{(t_d)}-\bar{q})^2 + \sum_{t=t_d+1}^{T^*} n_{(t)}(q_{(t)}-\bar{q})^2}\sqrt{n_d(h_{(t_d)}-\bar{h})^2 + \sum_{t=t_d+1}^{T^*} n_{(t)}(h_{(t)}-\bar{h})^2}} \qquad (22)$$

waarbij $\bar{h}$ en $\bar{q}$ overeenkomen met het gewogen gemiddelde van de waarden gelijk aan of groter dan $h_{(t_d)}$:

$$\bar{h} = \frac{n_d h_{(t_d)} + \sum_{t=t_d+1}^{T^*} n_{(t)} h_{(t)}}{n_d + \sum_{t=t_d+1}^{T^*} n_{(t)}} \tag{23}$$

$$\bar{q} = \frac{n_d q_{(t_d)} + \sum_{t=t_d+1}^{T^*} n_{(t)} q_{(t)}}{n_d + \sum_{t=t_d+1}^{T^*} n_{(t)}} \tag{24}$$

Voorgaande formulering is van toepassing voor een rechte lijn waarvan het intercept niet gekend is. In dit geval verwacht men echter dat het intercept nul is en de lijn door de oorsprong gaat. Om dit in rekening te brengen wordt in Vergelijking (22) de gemiddelde waarden $\bar{h}$ en $\bar{q}$ gelijkgesteld aan 0, eerder dan bepaald op basis van voorgaande vergelijking. Het praktisch effect van deze aanpassing is dat de "goodness-of-fit" bij hogere waarden relatief gezien een grotere invloed hebben op de G-statistiek, hetgeen wenselijk is.

Omwille van de beperkte steekproefomvang verwacht men niet dat deze correlatie effectief gelijk is aan 1. Om de correlatie te beoordelen is het daarom nodig de gemiddelde waarde en de standaarddeviatie te kennen voor de statistiek indien de gegevens effectief Weibull verdeeld zijn. Dit gebeurt in de procedure door de G-statistiek te herrekenen voor de gesimuleerde waarden in de parametrische bootstrap. De correlatie is echter beperkt tot 1 en voor hoge waarden en een kleine steekproefomvang zeker niet normaal verdeeld. Om een normale verdeling beter te benaderen wordt de G-statistiek daarom gelijkgesteld aan de volgende getransformeerde variabele:

$$G = \frac{\log(1-\rho)}{\log(1+\rho)} \tag{25}$$

Het is deze G-statistiek die wordt berekend voor de fit van de oorspronkelijke gegevens en die herhaaldelijk wordt berekend bij het toepassen van de parametrische bootstrap. Noemen we voor deze laatste simulaties μ_G de gemiddelde waarde en σ_G de standaardeviatie dan kan men volgende Z-statistiek berekenen die onder de null-hypothese dat de gegevens effectief Weibull verdeeld zijn benaderend standaard normaal verdeeld is:

$$Z = \frac{G - \mu_G}{\sigma_G} \tag{26}$$

De kans dat de data effectief Weibull verdeeld zijn en de G-statistiek kleiner zou zijn dan de g-waarde berekend voor de oorspronkelijke gegevens noemt men de P-waarde van de test-statistiek. Deze P-waarde is de kans dat men voor een standaardnormaal verdeelde toevalsvariabele een waarde kleiner dan Z vindt.

Hoe de voorgaande statistieken worden gebruikt om de drempelwaarde te kiezen wordt uiteengezet in Paragraaf 4.7.

4.5. NIET-PARAMETRISCHE BOOTSTRAP

Bij de niet-parametrische bootstrap worden uit de steekproef van M tijdsblokken lukraak M tijdsblokken gekozen met teruglegging. In de nieuwe tijdreeks zullen sommige tijdsblokken bijgevolg meermaals voorkomen terwijl andere niet worden gekozen.

Om de statistieken die volgen uit de niet-parametrische bootstrap voldoende nauwkeurig te berekenen is het nodig deze bootstrap meermaals uit te voeren (B maal). Men heeft aangetoond dat de nauwkeurigheid verbetert indien in de globale keuze van de MxB tijdsblokken ieder tijdsblok M maal voorkomt. Men verwijst naar deze selectietechniek als “balanced bootstrapping” en het is deze techniek die is toegepast in de schattingsprocedure. De praktische uitvoering van de selectie gebeurt door B lukrake selecties zonder teruglegging van M blokken uit een vector van MxB blokken waarin ieder block M maal voorkomt.

4.6. PARAMETRISCHE BOOTSTRAPPING

In tegenstelling tot de non-parametrische bootstrap waar de gesimuleerde tijdreeksen uit de oorspronkelijke gegevens worden gekozen en de gegevens dus niet noodzakelijkerwijze Weibull verdeeld zijn en onafhankelijk, wordt in de parametrische bootstrap N onafhankelijke gegevens gesimuleerd uit de geschatte Weibull verdeling.

Omdat enkel de waarden boven de drempelwaarde dienen gekend te zijn, wordt eerst de waarde van n_d (het aantal waarden kleiner of gelijk aan h_d) gesimuleerd. N_d is binomiaal verdeeld met parameters N (het aantal experimenten) en

$$p_d = 1 - \exp^{-\left(\frac{h_d}{u}\right)^k} \tag{27}$$

Wanneer p_dN en $(1-p_d)N$ groter zijn dan 10 mag men de verdeling van N_d benaderen door een normaalverdeling met verwachtingswaarde p_dN en variantie $(1-p_d)p_dN$. Wanneer deze voorwaarde vervuld is wordt n_d gesimuleerd als:

$$n_d = p_dN + \Phi^{-1}(U)\sqrt{(1-p_d)p_dN} \tag{28}$$

waarbij Φ^{-1} overeenkomt met de inverse standaardnormaal verdeling en U een gesimuleerde toevalsvariabele voorstelt met uniforme verdeling binnen het interval (0,1). De N-n_d overige waarden worden dan gesimuleerd door herhaalde toepassing van volgende vergelijking:

$$H = u \cdot \left(-\log(Up_d)\right)^{1/k} \tag{29}$$

waarbij U opnieuw overeenkomt met een toevalsvariabele met uniforme verdeling binnen het interval (0,1).

Indien $p_d N$ of $(1-p_d)N$ kleiner zijn zijn dan 10 dan wordt de simulatie rechtstreeks uitgevoerd voor alle N gegevens als volgt:

$$H = \begin{cases} h_d \quad \text{als} \quad U > p_d \\ u \cdot \left(-\log(U)\right)^{1/k} \quad \text{als} \quad U \le p_d \end{cases} \tag{30}$$

4.7. BEREKENING VAN DE EQUIVALENTE STEEKPROEFOMVANG

De bedoeling van de equivalente steekproefomvang N_{eq} is dat de standaarddeviaties van de schattingen bij de parametrische bootstrap benaderend overeenkomt met de waarden die worden gevonden op basis van de niet-parametrische bootstrap. Verschillen tussen de twee resultaten bij een zelfde steekproefomvang zijn te wijten aan het feit dat bij de parametrische bootstrap de gesimuleerde gegevens effectief Weibull verdeeld zijn en onafhankelijk, terwijl bij de niet-parametrische bootstrap de gegevens worden gekozen uit de originele tijdreeks en bijgevolg niet noodzakelijkerwijze onafhankelijk of Weibull verdeeld zijn. Een andere ongewenste reden voor eventuele verschillen is dat beide bootstrap methodes slechts een beperkt aantal keer (B maal) wordt uitgevoerd en de geschatte standaarddeviaties bijgevolg niet exact zijn.

Voor de ML-methode toegepast op niet-gecensureerde gegevens (dat wil zeggen voor een drempelwaarde gelijk aan 0) beschikt men over volgende asymptotische formules voor de varianties van u en k die geldig zijn bij grote steekproefomvang:

$$\left(\sigma_u^{AS}\right)^2 = \frac{1087}{N}\left(\frac{u}{k}\right)^2 \tag{31}$$

$$\left(\sigma_k^{AS}\right)^2 = \frac{0.608}{N}(k)^2 \tag{32}$$

Voor de drempelwaarde 0 kan de equivalente steekproefomvang bijgevolg berekend worden als:

$$N_{eq} = \text{Min}\left(\frac{1087}{\left(\sigma_u^{NPB}\right)^2}\left(\frac{u}{k}\right)^2, \frac{0.608}{\left(\sigma_k^{NPB}\right)^2}(k)^2 \right) \tag{33}$$

Zoals eerder vermeld is omwille van het beperkt aantal bootstraps B dat wordt uitgevoerd de schatting van de NPB-spreiding van u en k onzeker. Om te voorkomen dat door toevallige variatie de steekproefomvang wordt veranderd, wordt daarom de benedengrens gebruikt van een 90% eenzijdig betrouwbaarheidsinterval. Voor een variantie σ^2 bepaald op basis van B gegevens komt deze benedengrens overeen met:

$$\sigma_{min}^2 = \frac{N-1}{\chi_{0.1}^2(N-1)}\sigma^2 \tag{34}$$

waarbij $\chi_{0.1}^2(N-1)$ overeenkomt met de 10% bovenkwantiel van een chi-kwadraat verdeelde toevalsvariabele met N-1 vrijheidsgraden.

In Vergelijking (23) worden de NPB-schattingen van de standaarddeviaties van u en k beide gelijkgesteld aan de minimum waarden berekend op basis van Vergelijking (34). Uiteraard wordt de equivalente steekproefomvang verder beperkt tot de oorspronkelijke steekproefomvang N, zodat

$$N_{eq} = \text{Min}\left(\frac{1087}{\left(\sigma_{u,min}^{NPB}\right)^2}\left(\frac{u}{k}\right)^2, \frac{0.608}{\left(\sigma_{k,min}^{NPB}\right)^2}(k)^2, N \right) \tag{35}$$

De asymptotische formules houden echter geen rekening met het feit dat gegevens kleiner of gelijk aan de drempelwaarde worden gecensureerd en de formules zijn enkel geldig voor voldoende grote steekproefomvang. Daarom wordt een verbeterde berekening uitgevoerd van de equivalente steekproefomvang door rechtstreeks de spreiding te vergelijken van de parametrische en niet-parametrische bootstrap en te veronderstellen dat de variantie van de schatters inverse proportioneel is tot de steekproefomvang. Dit leidt tot volgende gecorrigeerde steekproefomvang:

$$N_{eq}^c = N_{eq}F \tag{36}$$

waarbij voor u:

$$F = \frac{\left(\sigma_u^{PB}(N_{eq})\right)^2}{\left(\sigma_u^{NPB}\right)^2} \tag{37}$$

en voor k:

$$F = \frac{\left(\sigma_k^{PB}(N_{eq})\right)^2}{\left(\sigma_k^{NPB}\right)^2} \tag{38}$$

Bij het gebruik van voorgaande formules dient men nu echter rekening te houden zowel met de onzekerheid op de schatting van de NPB-spreiding als die van de PB-spreiding. Om toevallige reducties van de steekproefomvang te vermijden wordt daarom de bovengrens gebruikt van een 90% éénzijdig betrouwbaarheidsinterval. Deze waarde komt overeen met de 10% bovenkwantielwaarde van een F-verdeling met B_{NPB}-1 en B_{PB}-1 vrijheidsgraden:

$$F_{max} = F \times F_{0.1}(B_{NPB} - 1, B_{PB} - 1) \tag{39}$$

B_{NPB} en B_{PB} komen overeen met het respectieve aantal simulaties uitgevoerd bij de niet-parametrische en parametrische bootstrap.

De uiteindelijke gecorrigeerde steekproefomvang wordt dan gelijkgesteld aan het minimum van beide resultaten en begrensd door de oorspronkelijke steekproefomvang:

$$N_{eq}^c = \text{Min}(N_{eq,u}^{c,max},\ N_{eq,k}^{c,max},\ N) \tag{40}$$

Wanneer de drempelwaarde niet nul is dan zijn de assymptotische formules niet van toepassing. Daarom wordt voor het bepalen van de equivalente steekproefomvang ook in dit geval Vergelijking (40) toegepast, waarbij N_{eq} in Vergelijking (36) overeenkomt met de gecorrigeerde equivalente steekproefomvang die werd berekend bij de voorgaande drempelwaarde. Deze berekening levert een betere benadering op, vermits de drempelwaarde naar de uiteindelijke drempelwaarde convergeert.

4.8. BEOORDELING VAN DE G-STATISTIEK EN KEUZE VAN DE DREMPELWAARDE

Zoals eerder uiteengezet in Paragraaf 4.3 kan men door middel van het resultaat van de parametrische bootstrap de P-waarde van de G-statistiek bepalen.

Bovendien wordt voor de drempelwaarde van toepassing de MSE-fout ("mean-square error) van de G-statistiek berekend:

$$MSE_G = \begin{cases} (G-\mu_G)^2 + \sigma_G^2 & \text{indien } G < \mu_G \\ \sigma_G^2 & \text{indien } G \geq \mu_G \end{cases} \quad (41)$$

Wanneer de G-statistiek groter is dan gemiddeld wordt verwacht, dan wordt enkel de variantie van de G-statistiek in rekening gebracht; is dit niet zo dan wordt ook de kwadratische vertekening ten opzichte van deze waarde toegevoegd. De MSE waarde vermindert naarmate G dichter bij de verwachtingswaarde ligt maar vemeerdert wanneer de variantie stijgt zodat het gebruik van een kleinere steekproefomvang wordt gepenalizeerd.

Indien bij de drempelwaarde van 0 de P-waarde groter is dan 5% dan wordt deze fit als voldoende beschouwd. Is dit niet het geval, dan wordt de minimum drempelwaarde gelijkgesteld aan 0 en de bijhorende P- en MSE-waarden worden genoteerd. We noemen deze waarden h_{min}, P_{min}, MSE_{min}. De maximum waarde van de drempel h_{max} wordt aangeduid als onbepaald. Vervolgens wordt de nieuwe drempelwaarde gelijkgesteld aan de mediaan van de steekproefgegevens (de waarde die met frequentie 50% wordt overstegen) en volgend iteratieschema wordt toegepast:

Indien de P-waarde voor de nieuwe drempelwaarde groter is dan 50% dan wordt deze waarde als maximum waarde voor de drempel genoteerd.

Is de P-waarde kleiner dan 50% dan wordt nagegaan of de P-waarde groter is dan 5%. Wanneer deze voorwaarde is voldaan en indien P_{min} nog steeds lager is dan 5% dan vervangt de nieuwe drempelwaarde de minimum waarde h_{min} en de overeenkomstige statistieken worden aangepast (P_{min} en MSE_{min}).

Is de P-waarde kleiner dan 5% of is de P-waarde bij het minimum reeds groter dan 5% dan wordt nagegaan of de MSE-waarde kleiner is dan de MSE-waarde bij het drempelminimum. Is dit zo, dan wordt het drempelminimum vervangen door de huidige drempelwaarde. Is dit niet zo, dan wordt het drempelmaximum vervangen door de huidige drempelwaarde.

De keuze van de volgende drempelwaarde wijzigt zich naargelang een maximum drempelwaarde al dan niet reeds is bepaald.

Wanneer h_{max} is gekend dan wordt de nieuwe drempelwaarde gelijkgesteld aan het gemiddelde van de minimum en maximum waarde. De iteratie wordt stopgezet indien het verschil tussen het aantal steekproefgegevens dat de minimum en maximum drempelwaarde overstijgt kleiner is dan 5% van de steekproefgegevens of wanneer het verschil tussen de minimum en maximum drempelwaarde overeenkomt met de nauwkeurigheid van de h-metingen.

Is h_{max} niet gekend dan wordt de nieuwe drempelwaarde gelijkgesteld aan de mediaan van de steekproefgegevens die de huidige drempelwaarde overstijgen. Indien dit aantal kleiner is dan 50 dan wordt de iteratie stopgezet.

Voorgaand algoritme convergeert in principe naar de drempelwaarde waarbij de MSE-waarde van de G-statistiek minimaal is terwijl de P-waarde 5% overstijgt. Omdat de P- en MSE-waarde berekend worden met slechts een beperkte nauwkeurigheid is het echter mogelijk dat minder optimale drempelwaarden worden gekozen. Bovendien dient men volgende twee uitzonderingen te noteren: wanneer de P-waarde groter is 5% bij de drempelwaarde 0 dan wordt geen censoring toegepast; indien geen P-waarde wordt bereikt die hoger is dan 5% bij censoring voor opeenvolgende mediaanwaarden (50%, 25%, 12.5% , 6.25%, ...) dan wordt het drempelniveau gelijkgesteld aan de hoogste van die reeks mediaanwaarden waarbij meer dan 50 steekproefgegevens de waarde overschrijden.

Om toevallige fluctuaties zoveel mogelijk uit te sluiten wordt de niet-parametrische bootstrapping simulatie enkel toegepast bij de drempelwaarde 0 en wordt bij de daaropvolgende simulaties dezelfde tijdsblokken gebruikt. Deze afhankelijkheid tussen de gesimuleerde resultaten bij verschillende drempelwaarden verbetert de relatieve nauwkeurigheid van de variatie van de NPB-spreiding van de schatters, terwijl de nauwkeurigheid bij 1 enkele drempelwaarde dezelfde blijft. Bovendien wordt door deze operatie de schattingsprocedure aanzienlijk versnelt. Voor de parametrische bootstrap is een gelijkaardige operatie niet mogelijk omdat de equivalente steekproefomvang varieert naargelang de drempelwaarde en het aantal te simuleren waarden varieert.

4.9. BEREKENING VAN BETROUWBAARHEIDSINTERVALLEN VOOR DE KWANTIELWAARDEN VAN H

Om de nauwkeurigheid van de resultaten te verbeteren wordt, éénmaal de optimale drempelwaarde is bereikt, de parametrische bootstrap methode nogmaals B' keer toegepast. Statistieken betreffende de spreiding van de eindresultaten zijn gebaseerd op deze simulaties samengevoegd met de simulaties in de laatste iteratiestap. Voor eenvoud van de notatie wordt het totaal bootstrap simulaties verder opnieuw aangeduid als B.

De B schattingen van u en k laten toe om de spreiding en de correlatie van de schatters te bepalen. Vergelijking (20) toont dat het logarithm van de gefitte kwantielwaarden linear verloopt met de waarden van log(u) en 1/k. Daarom worden ook de varianties van deze getransformeerde schatters en hun covariantie bepaald op basis van de gesimuleerde resultaten. Voor een gegeven p waarde wordt de standaarddeviatie van het logarithm dan berekend als:

$$\sigma_{\log(h_p)} = \sqrt{\mathrm{VAR}(\log(u)) + \mathrm{VAR}\left(\frac{1}{k}\right)(\log(-\log(p))^2 + 2 \times \mathrm{COVAR}(\log(u), \frac{1}{k})} \quad (42)$$

Een benaderend 95% tweezijdig betrouwbaarheidsinterval voor de schatting van h_p komt dan overeen met:

$$\left(h_p e^{-1.96\sigma_{\log(h_p)}}, h_p e^{1.96\sigma_{\log(h_p)}}\right) \quad (43)$$

waarbij h_p de ML-schatting voorstelt.

4.10. AFRONDING VAN DE MEETWAARDEN

Om het aantal gegevens te beperken is het nuttig om de oorspronkelijke meetgegevens af te ronden en dus te discretiseren. Noemen we Δh de stap tussen opeenvolgende metingen (bv. 5 cm). Voor de juiste toepassing van voorgaand algoritme dient men volgende punten te noteren.

Omdat de empirische kans van overschrijding wordt geschat er vanuitgaande dat 50% van de gegevens boven de genoteerde waarde liggen dient men voor een gegeven h-waarde de dichtsbijzijnde waarde noteren. Om te vermijden dat frequenties worden genoteerd bij de 0-waarde, dient men daarom de h-waarden te groeperen in intervallen $(0, \Delta h]$, $(\Delta h, 2\Delta h]$, ... en voor ieder interval de gemiddelde waarde te noteren. De orde-statistieken zijn bijgevolg van de vorm $0.5\Delta h$, $1.5\Delta h$, ...

Voor de drempelwaarde h_d wordt anderzijds in de schattingsprocedure een schatting gemaakt van de kans dat deze waarde wordt onderschreden door alle aantallen voor h-waarden kleiner dan de drempel op te tellen. Door de drempelwaarden steeds gelijk te stellen aan veelvouden van Δh, is dit een juiste schatting.

5. VERIFATIE VAN DE SCHATTINGSMETHODE

Om de schattingsmethode te verifieren werd de methode toegepast of 200 gesimuleerde steekproeven elk met 1500 datapunten. Volgende 3 gevallen worden beschouwd:

1. de gegevens zijn lukraak en onafhankelijk gekozen uit een Weibull verdeling waarvoor u=60 en k=1.15. De bedoeling is na te gaan of de schattingsprocedure de juiste waarde van u en k vindt en of de geschatte spreiding overeenkomt met de spreiding van de schatters;

2. zoals in 1. maar in dit geval worden slechts 300 datapunten gesimuleerd en elk punt wordt 5 maal gedupliceerd. De bedoeling is na te gaan of de schattingsmethode deze extreme vorm van correlatie juist kan detecteren;

3. zoals in 1. maar waarden kleiner dan 80 worden vervangen door een toevalsvariabele waarvan de kansdichtheid overeenkomt met een lognormale verdeling met mediaanwaarde 40 en standaarddeviatie van het logarithm gelijk aan 1 die wordt afgeknot tussen de waarden 20 en 80. De gemengde verdeling die hieruit volgt wordt getoond in Figuur 3. De bedoeling is na te gaan of de schattingsprocedure de drempelwaarde van 80 goed detecteert en tot correcte schattingen van u en k leidt.

5.1. GESIMULEERDE WEIBULL VERDELING

Resultaten van deze simulatie worden getoond in Appendix B en worden hier overlopen aan de hand van de samenvattende statistieken:

Het aantal meetpunten is 1500. In ongeveer 90% van de gevallen wordt geen correlatie gevonden en blijft de steekproefomvang dezelfde (zie Plot 20). Wanneer de steekproefomvang wordt verminderd is dit eerder beperkt, zoals men zou verwachten.

Slechts in enkele gevallen (ongeveer 3%, zie plots 18 en 19) wordt een drempelwaarde hoger dan 0 geschat. Wanneer dit echter gebeurt, dan is het mogelijk dat een redelijk hoge drempelwaarde wordt aangenomen. Vermoedelijk gebeurt dit voor gesimuleerde steekproeven waarbij de hoogste orde-statistieken toevallig sterk afwijken van de theoretische waarden en in dit geval zal het algoritme trachten om deze hoge waarden beter te benaderen. Anderzijds wordt natuurlijk in zulke gevallen een grotere standaardfout berekend voor de schattingen.

De resultaten voor de schattingen u en k tonen dat deze goed overeenkomen met de voorspelde asymptotische waarden. Waarden MLU0 en MLK0 verwijzen naar de schattingen zonder aanpassing van de steekproefomvang en voor drempelwaarde 0. EMLU en EMLK verwijzen naar de uiteindelijke schattingen. De gemiddelde waarden

komen goed overeen en benaderen zeer goed de werkelijke waarden, u=60 en k=1.15. Omdat het algoritme eventueel de steekproefomvang vermindert en/of de drempelwaarde verhoogt worden wel iets hogere standaarddeviaties genoteerd. Zoals in de volgende paragrafen wordt getoond wordt deze lichte verhoging echter meer dan gecompenseerd door de kleinere vertekening in de schatting van de juiste waarden en hun standaardfout wanneer correlatie en/of een drempelwaarde werkelijk voorkomt.

SMLU en SMLK is de geschatte standaardfout van de schattingen. Gemiddeld komen deze waarden goed overeen met de asymptotische waarden. Plot 14 en 16 tonen echter dat in enkele gevallen relatief hoge waarden worden geschat. Dit gebeurt voor die gevallen waarbij de drempelwaarde wordt verhoogd of de steekproefomvang wordt verminderd.

Ter vergelijking worden SMLU0 en SMLK0, de geschatte standaardfouten voor drempelwaarde 0 en zonder vermindering van de steekproefomvang, getoond. Deze waarden tonen een merkelijk lagere spreiding en benaderen gemiddeld ook zeer juist de asymptotische waarden. De spreiding op deze resultaten is uiteraard niet gewenst en zou verder kunnen verminderd worden door het opvoeren van het aantal bootstrap. De relatieve standaardfout is echter slechts een 10% bij het gebruikte bootstrap aantal hetgeen een redelijke nauwkeurigheid is.

De correlatie tussen u en k komt overeen met de waarde die theoretisch wordt verwacht.

Resultaten voor de G-statistiek (GFIT) en het overeenkomstig significantieniveau (PFIT) en de gestandaardizeerde G-statistiek (ZFIT) tonen dat de fits gemiddeld gezien juist gebeuren. In het bijzonder ZFIT zou in principe dienen overeen te komen met een standaardnormaal verdeling (zie PLOT 6). Dit is niet helemaal het geval omdat voor kleine ZFIT waarden het algoritme de drempelwaarde automatisch verhoogt om een betere fit te bekomen.

Tenslotte wordt de maximale fout getoond bij de ML-schatting van k en dit voor alle schattingen die in het algoritme worden uitgevoerd (alle bootstraps en schattingen op de originele verdeling). De nauwkeurigheid die wordt bereikt is in 98% van de gevallen bijzonder hoog (zie PLOT 5). In enkele gevallen blijkt het algoritme minder goed te convergeren (de iteratie voor de ML-waarde van k wordt stopgezet na 50 stappen). De nauwkeurigheid is echter nog steeds meer dan aanvaardbaar indien men deze fout vergelijkt met de standaardfout van k.

5.2. GESIMULEERDE WEIBULL VERDELING MET DUPLICATEN

Detailresultaten voor deze simulatie worden getoond in Appendix C. We bespreken hier voornamelijk de afwijkingen ten opzichte van het vorige geval.

Deze simulatie is een voorbeeld van een extreme correlatie. Eerder dan 1500 meetpunten zijn in feite slechts 300 onafhankelijke meetpunten aanwezig.

De samenvattende statistieken in Appendix C tonen dat dit goed wordt gedetecteerd door het algoritme. In alle gevallen wordt de steekproefomvang vermindert. Het gemiddelde komt overeen met 392 hetgeen hoger is dan het juiste aantal 300. Dit is echter te verwachten omdat in het algoritme een 90% boven-betrouwbaarheidsgrens wordt gebruikt voor de equivalente steekproefomvang. PLOT 19 toont dat de 10% benedenkwantiel van NEQV inderdaad dicht bij de waarde 300 ligt.

Het algoritme blijkt in dit opzicht dus goed te werken. De variatie van NEQV is opnieuw ongewenst en zou verminderd kunnen worden door het opvoeren van het aantal bootstrap. De relatieve spreiding van NEQV is ongeveer 30%. De fout bij de schatting van de standaardfout varieert echter met de vierkantswortel van NEQV en men mag dus een nauwkeurigheid van ongeveer 15% verwachten bij de schatting van de standaardfout.

De resultaten voor u en k tonen dat de fout voor de schatting eerder 20 tot 30% bedraagt. Dit is te verwachten vermits naast de foute schatting van NEQV ook de bootstrap zelf een fout genereert op de schatting van de standaardfout.

De schatting van de standaardfouten is in ieder geval merkelijk beter dan het geval waarbij geen rekening wordt gehouden (zie SMLU0 en SMLK0). In dit geval worden de resultaten systematisch als meer nauwkeurig beoordeeld dan in werkelijkheid het geval is.

De overige resultaten zijn gelijkaardig aan de vorige gevallen. Bij de schatting van u en k wordt wel één enkele uitbijter genoteerd (zie PLOT 20) hetgeen vermoedelijk te wijten is aan een toevallige maar extreme afwijking van de vooropgestelde Weibull verdeling.

5.3. GESIMULEERDE LOGNORMAAL-WEIBULL VERDELING

Resultaten van deze simulatie worden getoond in Appendix D.

Gemiddeld over de 200 simulaties is de drempelwaarde 81.9 hetgeen de theoretische waarde van 80 dicht benadered. De variatie van simulatie tot simulatie is redelijk groot (van 45 tot 170) maar vermits de lognormaalverdeling de Weibull dicht benaderd in dit gebied is dit niet verwonderlijk.

De schatting van de Weibull verdeling zonder de gegevens beneden de drempelwaarde te verwijderen leidt duidelijk tot foute resultaten (zie MLU0, MLK0, SMLU0 en SMLK0) die bovendien als te nauwkeurig worden geschat. Met de automatische schatting van de drempelwaarde wordt gemiddeld gezien de juiste waarde gevonden. De standaardfout van deze schatters is echter hoger zoals men

zou mogen verwachten. De geschatte standaardfout is gemiddeld gezien iets lager dan de standaarddeviatie die men vindt bij de 200 simulaties. Dit is te wijten aan het feit dat in het eerste geval de drempelwaarde gekend wordt verondersteld, terwijl in de simulaties de drempelwaarde varieert van geval tot geval hetgeen tot een bijkomende spreiding aanleiding geeft.

Het significantieniveau (PFIT) is gemiddeld gezien iets te laag (t.o.v. de waarde 0.5). Dit is echter te verwachten vermits het algoritme naar de beste fit zoekt (d.w.z. met de kleinste MSE van de G-statistiek) maar, wanneer het significantieniveau groter is dan 0.5, de drempelwaarde niet verder verhoogt.

6. TOEPASSING OP WERKELIJKE GEGEVENS

Het algoritme is toegepast op de volledige meetreeks van HM0 waarden genoteerd ter plaatse van Westhinder. In dit geval werd vereist dat de meetreeks dezelfde steekproefomvang zou omvatten voor de verschillende kwartalen van het jaar. De meetreeks werd verder onderverdeeld in maandblokken, waarbij verschillende maanden echter werden samengevoegd tot 1 enkel blok wanneer het tijdsblok een ongewoon laag aantal metingen bevat (afwijkingen tussen het aantal metingen per blok worden hierdoor beperkt).

De resultaten van de schatting worden getoond in Appendix E.

De totale meetreeks bevat 92,339 gegevens. Het algoritme bepaalt echter dat in werkelijkheid de meetreeks representatief is voor slechts 757 onafhankelijke gegevens. Dit is verrassend weinig, maar vermoedelijk te verklaren door sterke nonstationariteit van de Weibull verdeling en een effectief zeer hoge correlatie tussen de gegevens. De drempelwaarde wordt gelijkgesteld aan 75 cm.

De resulterende fit werd reeds getoond in Figuur 2 (Plot 6 in Appendix E). Het fitten van de rechterstaart wordt verder verduidelijkt in Plot 5 die voor elke orde-statistiek het gefitte resultaat toont. De empirische verdeling wordt zeer goed benaderd behalve voor de hoogste orde-statistiek die wordt overschat. Klaarblijkelijk is deze afwijking echter statistisch niet significant, vermits PFIT overeenkomt met 74%.

De samenvattende statistieken tonen verder dat voor deze reële gegevens een aanpassing voor de verschillende kwartalen nodig is. Blijkbaar werden in het tweede en derde kwartaal relatief meer gegevens genoteerd en in het algoritme wordt het aantal gegevens in deze kwartalen verminderd, terwijl het aantal gegevens in de andere kwartalen wordt vermeerderd.

APPENDIX A
ML-SCHATTING VAN PARAMETERS U EN K

De ML-schatters van u en k worden bepaald door de L-funktie te maximaliseren:

$$\ell = \binom{n}{n_d}\left[1-e^{-\left(\frac{h_d}{u}\right)^k}\right]^{n_d}\prod_{t=t_{d+1}}^{T^*}\left[k\left(\frac{(h_{(t)})^{k-1}}{u^k}\right)e^{-\left(\frac{h_{(t)}}{u}\right)^k}\right]^{n_{(t)}} \qquad (1)$$

Het logarithm van de L-functie varieert met u en k als volgt:

$$\log \ell \propto n_d \log\left(1-e^{-\left(\frac{h_d}{u}\right)^k}\right) + n_e \log(k) - n_e k \log(u) + (k-1)\sum_{t=t_d+1}^{T^*} n_{(t)} \log(h_{(t)}) - \sum_{t=t_d+1}^{T^*} n_{(t)}\left(\frac{h_{(t)}}{u}\right)^k$$

waarbij n_e het totaal aantal waarden aanduidt dat groter is dan de drempelwaarde:

$$n_e = \sum_{t=t_d+1}^{T^*} n_{(t)}$$

en dus ook overeenkomt met n-n_d.

Partiële afleiding van de log-likelihood naar u leidt tot volgende ML-vergelijking:

$$-n_d k \frac{(h_d)^k}{(u)^{k+1}} \frac{e^{-\left(\frac{h_d}{u}\right)^k}}{1-e^{-\left(\frac{h_d}{u}\right)^k}} - n_e k \frac{1}{u} + \frac{k}{(u)^{k+1}} \sum_{t=t_d+1}^{T^*} n_{(t)} (h_{(t)})^k = 0 \qquad (2)$$

of

$$n_e u^k = \sum_{t=t_d+1}^{T^*} n_{(t)} (h_{(t)})^k - n_d (h_d)^k \frac{e^{-\left(\frac{h_d}{u}\right)^k}}{1-e^{-\left(\frac{h_d}{u}\right)^k}} \qquad (3)$$

Noemen we p_d de kans dat H kleiner is dan h_d en p_e de kans dat H groter is dan h_d:

$$p_d = 1 - e^{-\left(\frac{h_d}{u}\right)^k} \tag{4}$$

$$p_e = e^{-\left(\frac{h_d}{u}\right)^k} \tag{5}$$

De ML-vergelijking kan nu worden herschreven als:

$$n_e u^k = \sum_{t=t_d+1}^{T^*} n_{(t)} \left(h_{(t)}\right)^k - n_d \left(h_d\right)^k \frac{p_e}{p_d} \tag{6}$$

De partiele afgeleide van de log-likelihood naar k leidt tot de volgende ML-vergelijking voor k:

$$n_d \left(\frac{h_d}{u}\right)^k \log\left(\frac{h_d}{u}\right) \frac{e^{-\left(\frac{h_d}{u}\right)^k}}{1 - e^{-\left(\frac{h_d}{u}\right)^k}} + \frac{n_e}{k} + \sum_{t=t_d+1}^{T^*} n_{(t)} \log\left(\frac{h_{(t)}}{u}\right) - \sum_{t=t_d+1}^{T^*} n_{(t)} \left(\frac{h_{(t)}}{u}\right)^k \log\left(\frac{h_{(t)}}{u}\right) = 0$$

De ML-schatter van k dient dus te voldaan aan:

$$+\frac{n_e u^k}{k} = -n_d \left(h_d\right)^k \log\left(\frac{h_d}{u}\right) \frac{e^{-\left(\frac{h_d}{u}\right)^k}}{1 - e^{-\left(\frac{h_d}{u}\right)^k}} - u^k \sum_{t=t_d+1}^{T^*} n_{(t)} \log\left(\frac{h_{(t)}}{u}\right) + \sum_{t=t_d+1}^{T^*} n_{(t)} \left(h_{(t)}\right)^k \log\left(\frac{h_{(t)}}{u}\right)$$

of

$$k = \frac{n_e u^k}{-n_d \left(h_d\right)^k \frac{p_e}{p_d} \log\left(\frac{h_d}{u}\right) - u^k \sum_{t=t_d+1}^{T^*} n_{(t)} \log\left(\frac{h_{(t)}}{u}\right) + \sum_{t=t_d+1}^{T^*} n_{(t)} \left(h_{(t)}\right)^k \log\left(\frac{h_{(t)}}{u}\right)} \tag{7}$$

Vermits uit Vergelijking (6) volgt dat:

$$-n_d \left(h_d\right)^k \frac{p_e}{p_d} \log\left(\frac{1}{u}\right) - u^k \sum_{t=t_d+1}^{T^*} n_{(t)} \log\left(\frac{1}{u}\right) + \sum_{t=t_d+1}^{T^*} n_{(t)} \left(h_{(t)}\right)^k \log\left(\frac{1}{u}\right) = 0 \tag{8}$$

kan men Vergelijking (7) herschrijven als:

$$k = \frac{n_e u^k}{-n_d (h_d)^k \frac{p_e}{p_d} \log(h_d) - u^k \sum_{t=t_d+1}^{T^*} n_{(t)} \log(h_{(t)}) + \sum_{t=t_d+1}^{T^*} n_{(t)} (h_{(t)})^k \log(h_{(t)})} \quad (9)$$

Noteren we nu de volgende statistieken die enkel functie zijn van k en de steekproefgegevens:

$$S_{ke} = \frac{\sum_{t=t_d+1}^{T^*} n_{(t)} (h_{(t)})^k}{n_e} \quad (10)$$

$$S_{kd} = \frac{n_d (h_d)^k}{n_e} \quad (11)$$

$$S_{le} = \frac{\sum_{t=t_d+1}^{T^*} n_{(t)} \log(h_{(t)})}{n_e} \quad (12)$$

$$S_{kle} = \frac{\sum_{t=t_d+1}^{T^*} n_{(t)} (h_{(t)})^k \log(h_{(t)})}{n_e} \quad (13)$$

Gebruik makend van deze statistieken kan men de ML-vergelijkingen herschrijven als:

$$u^k = S_{ke} - S_{kd} \frac{p_e}{p_d} \quad (14)$$

$$k = \frac{u^k}{-S_{kd} \frac{p_e}{p_d} \log(h_d) - u^k S_{le} + S_{kle}} \quad (15)$$

Voor drempelwaarde 0, is $(S_{kd} p_e)/p_d = 0$. In dit geval vereenvoudigen de ML-vergelijkingen tot:

$$u^k = S_{ke} \quad (16)$$

$$\frac{1}{k} = \frac{S_{kle}}{S_{ke}} - S_{le} \qquad (17)$$

Oplossen van Vergelijkingen (14) en (15), of (16) en (17), naar u en k gebeurt door middel van een dubbele iteratie.

Voor gegeven waarde van k, kan de ML-schatter van u iteratief worden opgelost uit Vergelijking (14). Substitutie van deze waarde in Vergelijking (15) laat dan toe een nieuwe waarde voor k te berekenen, waarbij in het rechterlid de voorgaande waarde van k wordt ingevuld. Op deze wijze kan men bepalen of k verhoogd of verlaagd dient te worden. Verdere iteratie naar k is dan mogelijk tot convergentie optreedt.