Modelleren van Regenbuien in Afrika

Modeling Rain Showers in Africa

I.C.J. Backers

Modelleren van Regenbuien in Afrika

Modeling Rain Showers in Africa

door

I.C.J. Backers

ter verkrijging van de graad van Bachelor of Science
aan de Technische Universiteit Delft,
in het openbaar te verdedigen op donderdag 7 juli om 10:00 uur.

Studentnummer:	4960947	
Projectduur:	25 februari 2022 – 7 juli 2022	
Commissie:	Prof. dr. ir. G. Jongbloed,	TU Delft, begeleider
	Prof. dr. ir. N. C. van de Giesen,	TU Delft
	Dr. J. G. Spandaw,	TU Delft

TU Delft

Voorwoord

Door klimaatveranderingen lijkt het weer steeds extremer te worden en moeilijker te voorspellen. Voor Afrika is erg weinig data beschikbaar om klimaatverandering te meten of in weermodellen te gebruiken. Vanuit de TU Delft is daarom al vanaf 2007 gewerkt aan de opzet van een netwerk van lowcost weerstations om gronddata te verzamelen.

Initator van dit project en voorzitter van Delft Global Initiative Prof. dr. ir. Nick van de Giesen klopte aan bij de afdeling statistiek van wiskunde met de vraag dat hij graag een metriek wilde hebben die de mate van Poissoniteit van de regen kwantificeert.

Door zelf op zoek te gaan naar een onderwerp voor mijn BEP ben ik met een paar mailtjes bij dit project terecht gekomen waar ik met veel plezier aan heb gewerkt. Wat voor mij het project zo leuk maakte, was dat ik kon werken met echte data. Alle rekenkunsten die ik de afgelopen vier jaar heb geleerd kon ik nu eindelijk echt een keer toepassen op een probleem waarvan het antwoord niet al van te voren vast staat. En wat het helemaal bijzonder maakt is dat het een heel klein deelprojectje is van een veel groter project om het klimaat van Afrika beter in kaart te brengen.

Graag wil ik Prof. dr. ir. Geurt Jongbloed heel erg bedanken voor de wekelijkse bijeenkomsten met nuttige gesprekken, suggesties, hulp en tijd. Ook Prof. dr. ir. Nick van de Giesen wil ik bedanken voor de hulp bij het formuleren van dit project, het beschikbaar stellen van de gebruikte data en voor het periodiek aanschuiven bij de bijeenkomsten. Ook wil ik Dr. Jeroen Spandaw bedanken voor het completeren van mijn beoordelingscommissie. Als laatste wil ik Lutz Dümbgen bedankten voor het delen van zijn code om taut string werkende te krijgen.

Irene Backers
Rotterdam, Juni 2022

Samenvatting

Om goed te kunnen voorspellen waar, wanneer en hoeveel het gaat regenen is het belangrijk om satellietbeelden, weermodellen en grondmetingen te combineren. Het in 2007 opgezette project 'The Trans-African Hydro-Meteorological Obsevatory' (TAHMO) heeft de afgelopen jaren 500 goedkope weerstations geplaatst, verspreid over 21 landen in Sub-Sahara Afrika en wil uitbreiden naar 20.000 weerstations. De grondmetingen die met de al geplaatste weerstations zijn gedaan, kunnen gebruikt worden om weermodellen te ontwikkelen en te verbeteren.

Over de te verwachten regenval in Afrika is nog weinig bekend. Vaak wordt regen gemodelleerd met een Poissonproces. Daarom is vanuit 'Delft Global Initiative' de vraag gekomen of het Poissonproces een goed model kan zijn voor regenval in Afrika.

In dit onderzoek hebben we regendata van één weerstation gebruikt dat tijdens de moesson gegevens heeft verzameld op Mafia eiland, een eiland zo'n 20 kilometer voor de kust van Tanzania.

Een Poissonproces kan homogeen of inhomogeen zijn en we hebben de data op beide processen getoetst. Voor het homogene Poissonproces zijn veel verschillende toetsen te vinden in de literatuur. De Monte Carlo simulaties zijn gebaseerd is op de geïntegreerde verdelingsfunctie en gebruikt om de p-waarde te bepalen. Voor alle regen samen in de gehele observatieperiode was de p-waarde 0 en daarom kan de regen niet met een homogeen Poissonproces worden beschreven.

Voor de inhomogeen Poissonproces benadering hebben we ervoor gekozen om de intensiteit van de regen te benaderen met een stuksgewijs constante intensiteit. Met taut string werd de regen opgedeeld in 474 deelperiodes met een constante intensiteit die per deelperiode verschilt. Voor iedere deelperiode is getoetst of de regen met een homogeen Poissonproces kan worden beschreven. De p-waarde is voor iedere deelperiode bepaald met Monte Carlo simulaties en we werken met een significantieniveau van 5%. Voor 448 van de 474 deelperiodes kan de regen met een homogeen Poissonproces worden beschreven.

Een inhomogeen Poissonproces met stuksgewijs constante intensiteit lijkt daarom een redelijk bruikbaar model om de regen op Mafia eiland mee te beschrijven. Of dit model ook bruikbaar is voor alle andere weerstations van TAHMO zullen toekomstige studies moeten uitwijzen.

Lekensamenvatting

Om goed te kunnen voorspellen waar, wanneer en hoeveel het gaat regenen is het belangrijk om satellietbeelden, weermodellen en grondmetingen te combineren. Het in 2007 opgezette project 'The Trans-African Hydro-Meteorological Obsevatory' (TAHMO) heeft de afgelopen jaren 500 goedkope weerstations geplaatst, verspreid over 21 landen in Sub-Sahara Afrika en wil uitbreiden naar 20.000 weerstations. Over het voorspellen van regen in Afrika is nog nauwelijks wat bekend.

In dit onderzoek hebben we regendata van één weerstation gebruikt dat tijdens de moesson gegevens heeft verzameld op Mafia eiland, een eiland zo'n 20 kilometer voor de kust van Tanzania. Het Poissonproces is een gebruikelijk model om regen te modelleren. Uit de analyse van de data blijkt dat dit Poissonproces hier geen goed model is om de regen te beschrijven. Wanneer de regenmeting in kleinere deelperiodes wordt opgedeeld blijkt dat het Poissonproces de regen voor 448 van de 474 deelperiodes wel goed kan beschrijven en bruikbaar is om de regen op Mafia eiland mee te beschrijven. Of dit model, dat een inhomogeen Poissonproces wordt genoemd, ook bruikbaar is voor alle andere weerstations van TAHMO zullen toekomstige studies moeten uitwijzen.

Inhoudsopgave

1

Inleiding

Afrika's economie is voor een groot deel afhankelijk van regen en toch weten we nog maar heel weinig over waar, wanneer en hoeveel het regent in Afrika. Met alle wetenschappelijke kennis van nu is dit onaanvaardbaar (TUDelft, 2019), zeker als je bedenkt dat de grootste sector in Afrika de agrarische sector is en dat de verbouwing van 98% van de gewassen afhankelijk is van regenwater (de Villiers e.a., 2021).

Om goed te kunnen voorspellen waar, wanneer en hoeveel het gaat regenen, moeten satellietbeelden, weer/klimaatmodellen en grondmetingen gecombineerd worden. Het ontbrak hierbij vooral aan de grondmetingen. Daarom is in 2007 'The Trans-African Hydro-Meteorological Observatory' (TAHMO) opgezet. Het doel is om 20.000 goedkope weerstations te plaatsen verspreid over heel Sub-Sahara Afrika en zo het weer, het klimaat en de regen beter in beeld te kunnen brengen (TUDelft, 2019). De opstartfase duurde lang waardoor er pas in 2015 echt begonnen kon worden en er in juli 2019 vijfhonderd weerstations verspreidt over 21 landen stonden.

Als we de regen beter kunnen voorspellen, kan de voedselproductie verbeteren met alle voordelen van dien. Door de weerstations bij scholen te plaatsten en het op te nemen in het schoolprogramma, heeft het project ook educatieve waarde.

De data die met de grondmetingen verzameld is, kan gebruikt worden bij het verbeteren en ontwikkelen van weermodellen. Vaak wordt regen gemodelleerd met een Poissonproces en vanuit Delft Global Initiative (DelftGlobal, 2022) is de vraag gekomen of het Poissonproces een goed model kan zijn voor regenval in Afrika.

Uit een master thesis geschreven in 2019 en het daarop volgende artikel van de Villiers e.a., 2021 blijkt dat dit model voor de regen in Afrika niet goed past. In het artikel wordt het vermoeden geformuleerd dat regen met een lage intensiteit, bijvoorbeeld miezer, wel gemodelleerd kan worden met een Poissonproces, maar dat voor regen met hoge intensiteit, bijvoorbeeld een stortbui, het Poissonproces geen goed model lijkt te zijn.

In het artikel wordt met 'Poisson' verwezen naar een homogeen Poissonproces. De lichte regen lijkt dus wel met een homogeen Poissonproces beschreven te kunnen worden, maar de zwaardere regen niet. In het artikel wordt ook gewezen op het feit dat je een inhomogeen Poissonproces zou kunnen splitsen in kleinere stukken die homogeen Poisson zijn.

Figuur 1.1: Een weerstation bij een school in Afrika. Overgenomen van TUDelft, 2019

Het doel van dit project is om te **onderzoeken of een Poissonproces, homogeen of inhomogeen, een redelijk model is om de regenbuien in Afrika mee te modelleren**. Belangrijk om hierbij te onthouden is de wijsheid van de Britse statisticus George Box:

"All models are wrong, but some are useful"

Om het antwoord te vinden op deze vraag toetsen we eerst of de regen door een homogeen Poissonproces kan worden beschreven. Hiervoor gebruiken we een toets uit de literatuur. Daarna zullen we de regen toetsen op een inhomogeen Poissonproces door de data te benaderen met een stuksgewijs homogeen Poissonproces. We delen de regen op in allemaal kleine homogene Poissonprocessen die samen een inhomogeen Poissonproces vormen.

De structuur van het verslag is als volgt. Als eerste wordt in hoofdstuk 2 de data uitgelegd. Er wordt onder andere ingegaan op de gebruikte meetapparatuur, waar in Afrika de data voor dit project verzameld is, wanneer en waarom de data verzameld is en als laatste wordt uitgelegd hoe de beschikbare data eruit ziet. In hoofdstuk 3 wordt uitgelegd wat een Poissonproces is en hoe de data gebruikt kan worden om te toetsen op een homogeen of inhomogeen Poissonproces. Hoofdstuk 4 geeft een overzicht van (een deel van) de vele toetsen die er in de literatuur zijn voor een homogeen Poissonproces. Vervolgens wordt één van deze toetsen gekozen en uitgevoerd op de regendata. Tot slot wordt in hoofdstuk 5 de regendata getoetst op een inhomogeen Poissonproces. Hiervoor wordt eerst de intensiteit van de regen geschat met Taut String. Dit geeft een stuksgewijs constante intensiteit waardoor de regendata verdeeld wordt in periodes met dezelfde intensiteit. Deze periodes worden vervolgens getoetst op een homogeen Poissonproces.

2

Het hoe, wat, waar en waarom van de dataverzameling

Dit hoofdstuk geeft een overzicht van de dataverzameling. Als eerste kijken we in paragraaf 2.1 naar de meetapparatuur waarmee de data verzameld is. In paragraaf 2.2 wordt uitgelegd waar en waneer de data verzameld is. Paragraaf 2.3 legt uit waarom de data verzameld is. Als laatste staat in paragraaf 2.4 welk deel van de beschikbare data gebruikt is voor dit project en wordt uitgelegd hoe de beschikbare data eruit ziet.

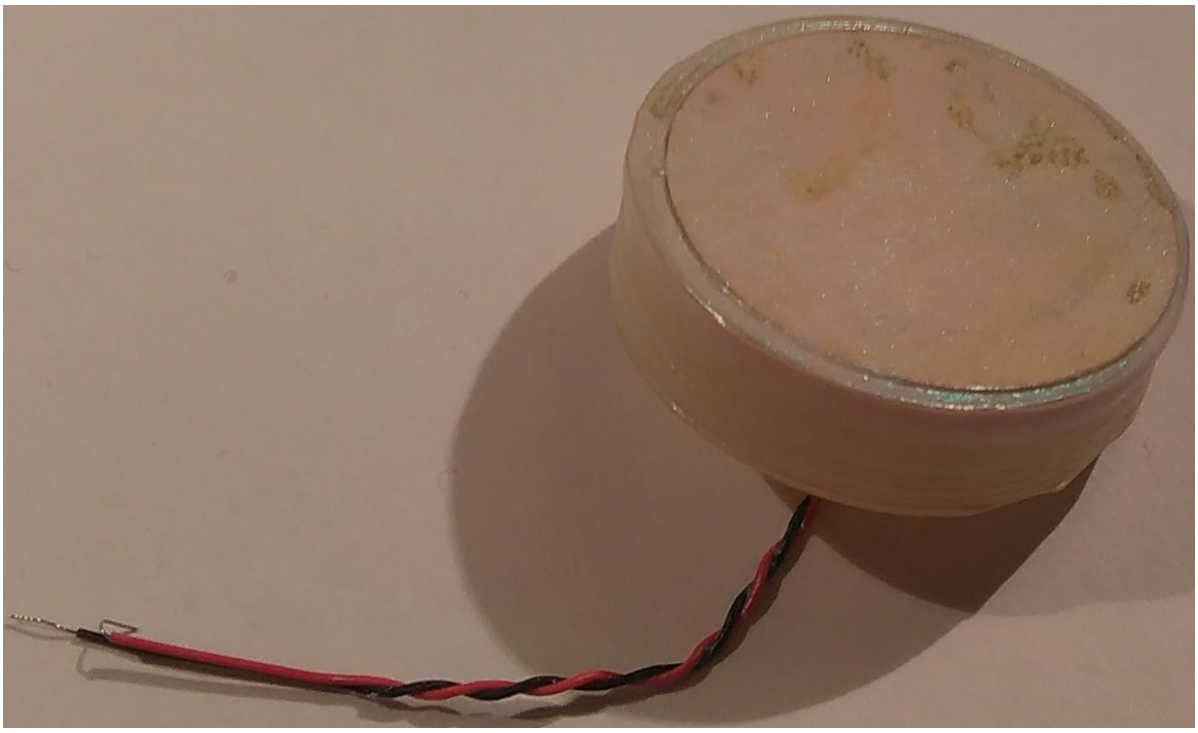
2.1. De meetapparatuur

De data is verzameld met drie verschillende meetinstrumenten: tipping bucket regenmeters gemaakt door Onset in de Verenigde Staten, akoestische disdrometers gemaakt door Disdro in Delft en intervalometers. In figuur 2.1 staat van alle drie een foto.

Tipping buckets meten de hoeveelheid regen, disdrometers registreren de aankomst van regendruppels en berekenen de grootte van de druppels en intervalometers registreren de aankomst van regendruppels (de Villiers, 2019). Voor dit onderzoek is gebruikt gemaakt van metingen gedaan met een intervalometer.

Figuur 2.1: De gebruikte meetinstrumenten. Van links naar rechts: een tipping bucket, een akoestische disdrometer en een intervalometer. Overgenomen van ONSET, 2022, Nieuwenhuizen, 2010 en van de Giesen, 2019.

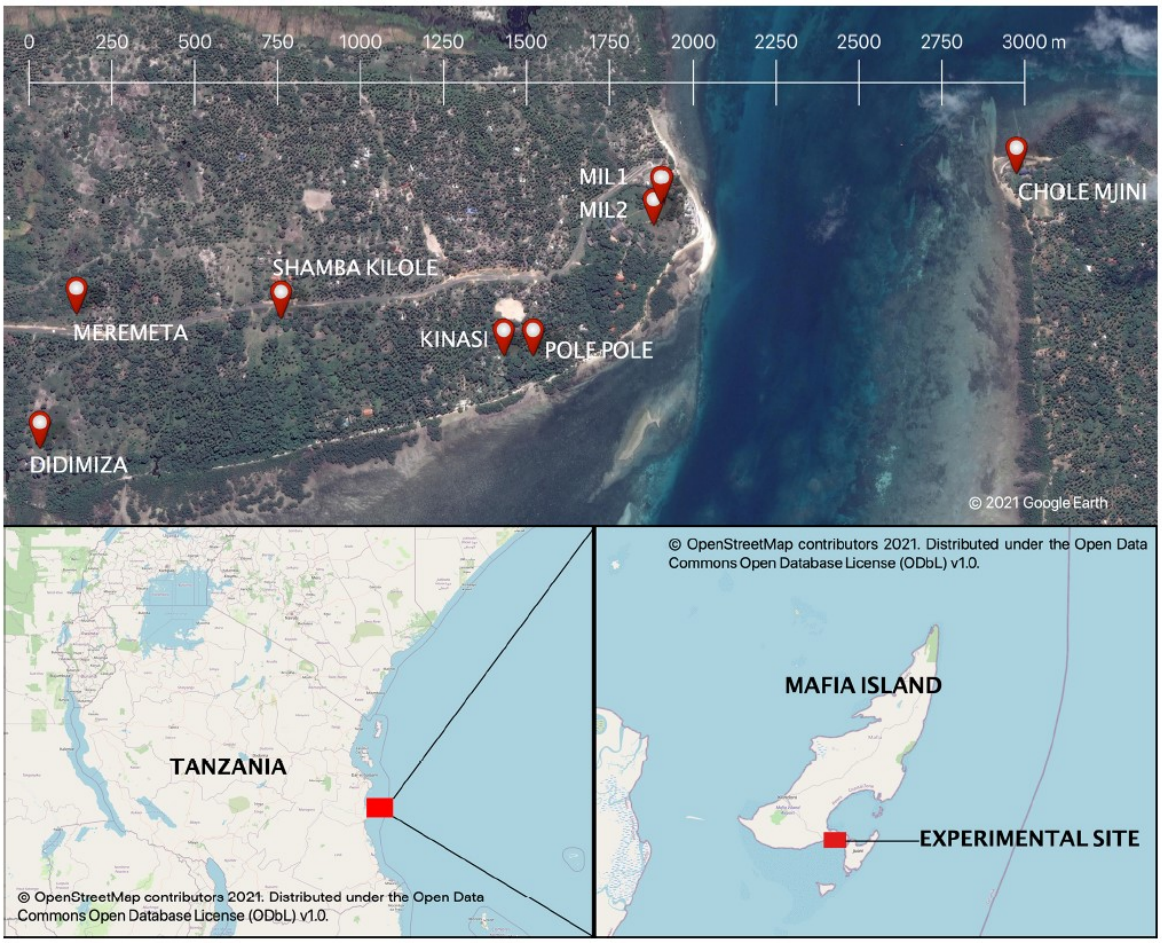
Intervalometers zijn meetinstrumenten die goedkoop zijn in aanschaf, gebruik en onderhoud (TUDelft, 2022a). Ze kosten rond de 200 dollar, terwijl soortgelijke meetinstrumenten voor de Europese markt pas verkrijgbaar zijn vanaf 2000 dollar en meer (TUDelft, 2010). Het maken van dit goedkope meetinstrument kan door al bestaande apparatuur of onderdelen hiervan te gebruiken voor een ander doel dan waar het oorspronkelijk voor ontworpen was (TUDelft, 2010). Zo maakt een intervalometer gebruik van een piëzo-elektrische sensor van ongeveer 3,5 cm groot zoals in figuur 2.2. Als er een druppel op de sensor valt, dan vervormt de sensor en wordt er spanning opgewekt. Door de sensor aan een computer te verbinden, kan geregistreerd worden wanneer de druppels gevallen zijn en bijvoorbeeld ook hoe groot de druppels waren (van de Giesen, 2019).



Figuur 2.2: De Piëzo-elektrische sensor die gebruikt wordt voor een intervalometer. De sensor is zo'n 3,5 cm groot en wekt spanning op als er een druppel op valt. Overgenomen van van de Giesen, 2019.

2.2. Waar en wanneer is de data verzameld?

De data is verzameld op acht locaties langs de zuidkust van Mafia Eiland, een eiland dat ongeveer 20 kilometer voor de kust van Tanzania ligt. De locaties liggen in een gebied van 3,1 km bij 500 m en ongeveer in een lijn zoals de kaart in figuur 2.3 laat zien (de Villiers e.a., 2021). Op elke locatie staat in ieder geval één intervalometer. Pole Pole heeft naast een intervalometer ook nog een disdrometer en een tipping bucket; MIL 1 heeft twee intervalometers en een tipping bucket (de Villiers, 2019). De meetinstrumenten zijn zo geplaatst dat ze qua toegankelijkheid en landschap zo goed mogelijk voldoen aan de eisen van de Wereld Meteorologische Organisatie (de Villiers e.a., 2021), dit is een gespecialiseerde organisatie van de Verenigde Naties op het gebied van weer, klimaat en water ("Our Mandate", 2022).



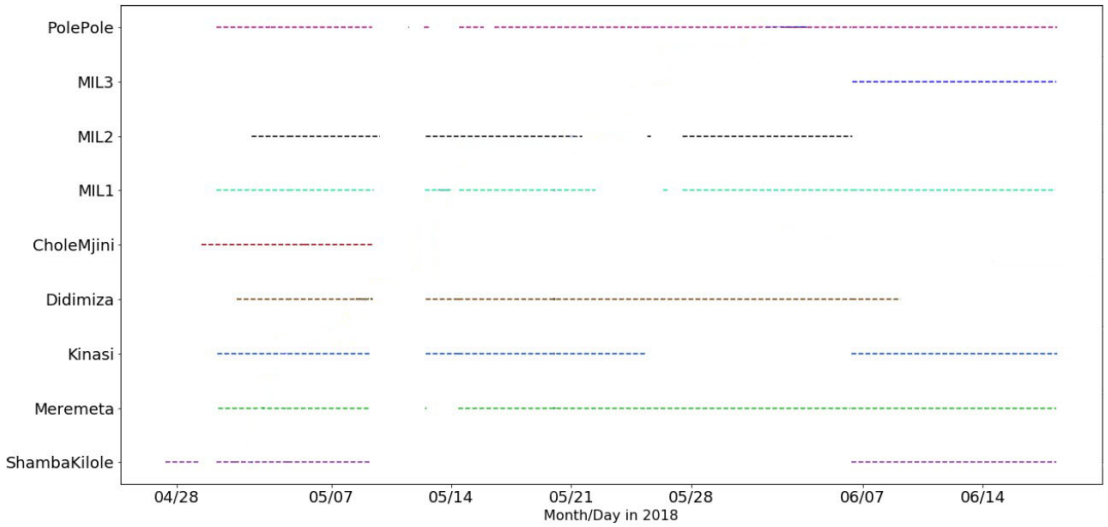
Figuur 2.3: De acht locaties waar de data verzameld is. Op alle locaties staat een intervalometer. Pole Pole heeft naast de intervalometer nog een disrometer en een tipping bucket. MIL 1 heeft nog een extra intervalometer (twee in totaal) en een tipping bucket. Overgenomen van de Villiers e.a., 2021.

De data is verzameld tijdens de moesson in Afrika. De eerste meting is gedaan op 27 april 2018 en de laatste op 18 juni 2018. Helaas is de dataverzameling niet altijd goed gegaan. Zo heeft de disdrometer vanaf 20 mei een oscillerend signaal gemeten waardoor de data niet meer te gebruiken is. Ook hebben een aantal intervalometers waterschade opgelopen tijdens zware buien waardoor ze tijdelijk geen data konden verzamelen. Twee intervalometers zijn dusdanig beschadigd dat ze niet meer te repareren waren.

Figuur 2.4 laat zien welke intervalometer wanneer data heeft verzameld. Elke intervalometer heeft de naam van de locatie behalve MIL 3, dit is de tweede intervalometer die bij MIL 1 staat.

Waarschijnlijk heeft het de ochtend van 9 mei hard geregend. Alle intervalometers die het op dat moment deden, zijn namelijk 9 mei uitgevallen, de eerste om 05:02 uur, de laatste om 09:13 uur. Op 12 mei waren de meeste intervalometers weer gerepareerd, alleen Chole Mjini en Shamba Kilole niet.

Na 12 mei zijn veel intervalometers nog één of meerdere keren uitgevallen. Daarom wordt er gewerkt aan een robuustere versie van de intervalometer



Figuur 2.4: Voor iedere locatie de periodes dat de intervalometers data hebben verzameld. Iedere intervalometer heeft de naam van de locatie behalve MIL 3, dit is de tweede intervalometer die bij MIL 1 staat. Aangepast overgenomen van de Villiers e.a., 2021.

2.3. Waarom wordt de data verzameld?

Momenteel weten we een stuk minder over het weer en klimaat in Afrika dan in bijvoorbeeld Europa (TUDelft, 2010). Dat we zo weinig weten van het weer en klimaat in Afrika komt onder andere door een gebrek aan data. Satellietbeelden waren er al wel maar data waargenomen vanaf de grond was er nauwelijks. In 2015 stond er volgens de Villiers e.a., 2021 minder dan één weerstation per 1 miljoen vierkante kilometer.

Ter vergelijking: in Nederland staan 48 automatische weerstations waarvan er 34 op land staan en 14 op zee (KNMI, 2022). Dit is één automatisch weerstation per 1.222 km^2. Naast deze automatische weerstations staan er verspreid over heel Nederland ook nog ruim 300 neerslagmeters van vrijwilligers (KNMI, 2022).

Met de beschikbare satellietbeelden in Afrika kunnen wel wolken worden waargenomen, maar je weet dan nog niet of het regent, waar het regent en hoeveel het regent. Om dat te kunnen weten, heb je grondwaarnemingen nodig en zo ontstond het idee om 'The Trans-African Hydro-Meteorological Observatory' (TAHMO) op te zetten. Het doel is om 20.000 goedkope weerstations te plaatsten verspreid over Sub-Sahara Afrika (TUDelft, 2019). Deze grondwaarnemingen kunnen gebruikt worden om betere weer- en klimaatmodellen te ontwikkelen. Vervolgens kunnen de grondwaarnemingen, de modellen en de satellietbeelden gecombineerd worden om een goed beeld te krijgen van waar, wanneer en hoeveel het regent.

Het beter kunnen voorspellen van de regen heeft meerdere voordelen (TUDelft, 2022b). Op wetenschappelijk gebied krijgen we meer inzicht in het hydrologisch systeem. Deze kennis kan gebruikt worden op het gebied van waterkracht, irrigatie, drinkwater en overstromingen.

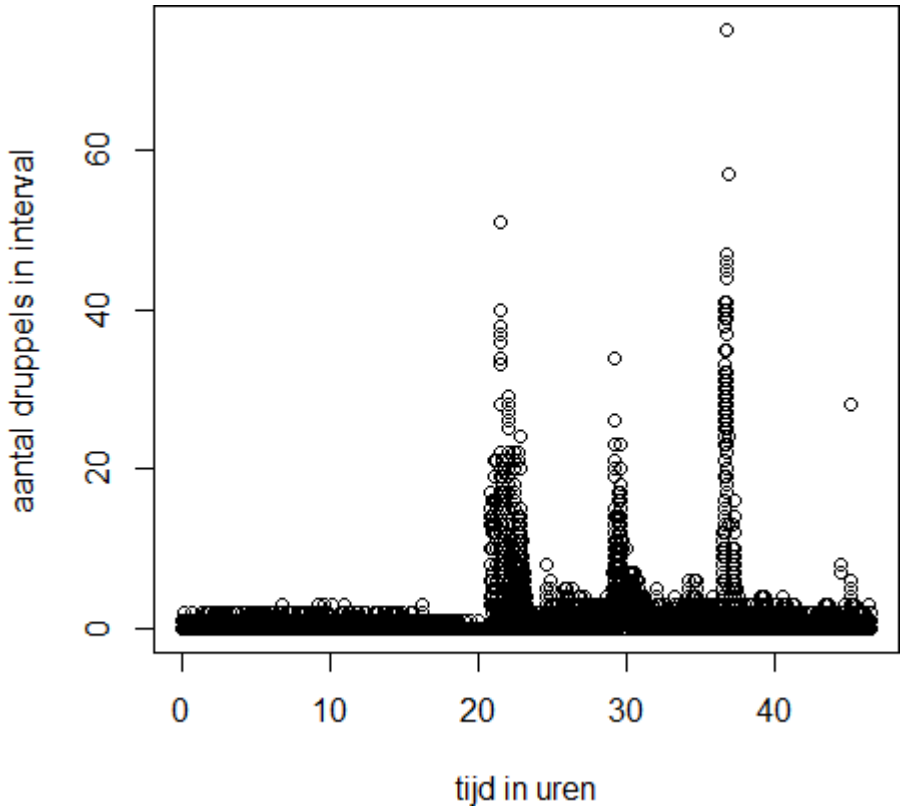
Financieel gezien kunnen regeringen beter beslissen waarin geïnvesteerd gaat worden op het gebied van watermanagement. Door in de juiste dingen te investeren, kan er beter geanticipeerd worden op het weer waardoor oogsten kunnen verbeteren. De voedselproductie kan dus toenemen maar ook de oogstvoorspellingen worden beter. Ook kan de data gebruikt worden door verzekeraars, zo kunnen boeren bijvoorbeeld eerlijker gecompenseerd worden als er oogsten mislukken.

Als laatste zullen er ook educatieve voordelen zijn. De meetinstrumenten worden bij scholen geplaatst en worden opgenomen in het onderwijs. Zo kunnen de kinderen leren over het klimaat in Afrika en hun verbondenheid daarmee. Bij natuurkunde kunnen de meetinstrumenten worden gebruikt om experimenten te doen en het kan gebruikt worden om wiskunde toe te passen.

2.4. Welke data is er gebruikt en hoe ziet die data er uit?

Er is bijna 3 maanden aan regendata beschikbaar van acht verschillende locaties. Op aanraden van Didier heb ik er voor gekozen om data te gebruiken van een regenbui opgenomen op Pole Pole. Het gaat om regen die gevallen is vanaf 1 mei 12 uur tot 3 mei 10:30 uur. Deze periode noemen we de observatieperiode. In figuur 2.4 komt dit overeen met het eerste stukje van de roze stippellijn helemaal linksboven in de figuur.

De intervalometer heeft de tijdstippen opgenomen wanneer er een druppel op de drum valt. Van iedere druppel weten we tot op de milliseconde wanneer deze gevallen is. Ook hebben we een bestand met het aantal gevallen regendruppels per interval van 10 seconden. Een weergave hiervan staat in figuur 2.5. Op de x-as staat de tijd in uren waarbij we 1 mei 12 uur gelijk gesteld hebben aan tijdstip 0; op de y-as staat het aantal regendruppels per interval van 10 seconden. Het aantal druppels per interval heeft te maken met de intensiteit van de regen. Regent het zacht, dan zullen er weinig druppels per interval zijn. Regent het hard, dan zullen er veel druppels per interval zijn. De eerste 20 uur zijn er alleen maar intervallen met 0, 1, 2 of 3 druppels. Het heeft toen dus niet hard geregend. De drie uur daarna vielen er meer druppels per interval en heeft het dus harder geregend.



Figuur 2.5: Een weergave van het aantal druppels per interval gevallen bij Pole Pole van 1 mei 12 uur tot 3 mei 10:30. Op de x-as de tijd in uren met 1 mei 12 uur als tijdstip 0, op de y-as staat het aantal druppels per interval.

3

Het Poissonproces

Dit hoofdstuk zal worden uitgelegd wat een Poissonproces is en hoe dit in de data terug kan komen. In paragraaf 3.1 worden definities en eigenschappen van een Poissonproces gegeven en uitgelegd. Vervolgens worden er drie manieren uitgelegd om naar een Poissonproces te kijken in paragraaf 3.2. Als laatste wordt in paragraaf 3.3 gekeken hoe de data zich moet gedragen als het een Poissonproces volgt en wordt er gekeken naar de mogelijkheden en onmogelijkheden van de data voor het toetsen op een homogeen en inhomogeen Poissonproces.

3.1. Wat is een Poissonproces?

Een Poissonproces is een speciaal soort telproces en telt of registreert het aantal 'gebeurtenissen' in een bepaald tijdsinterval (Pishro-Nik, 2014). Zo'n gebeurtenis wordt een aankomst genoemd. De aankomsten van een Poissonproces gebeuren met een bepaalde intensiteit maar wel volledig willekeurig zonder enige structuur en onafhankelijk van elkaar (Pishro-Nik, 2014). De intensiteit van het proces is het aantal aankomsten dat je verwacht in een tijdsinterval van lengte 1. Een Poissonproces wordt onder andere gebruikt voor het modelleren van de aankomst van mensen in een winkel, de uitstoot van radioactieve deeltjes door radioactief afval en het overgaan van de telefoon bij een klantenservice (Grimmett en Welsh, 2014).

Je kan het aantal aankomsten in een tijdsinterval registeren. Laat de tijd lopen van 0 tot oneindig, dus $t \in [0, \infty)$, en laat N_t het aantal aankomsten tot en met tijdstip t zijn, dus in het interval $[0, t]$. Het proces $N = (N_t : t \in [0, \infty))$ begint op tijdstip $t = 0$ met nul aankomsten $N_0 = 0$, bij elke volgende aankomst neemt de waarde van het proces met 1 toe. Dit betekent dat het proces waardes aanneemt in {0,1,2,3,...} en dat N een discreet stochastisch proces is in continue tijd (Grimmett en Welsh, 2014). $N = (N_t : t \in [0, \infty))$ is een telproces als het voldoet aan de volgende voorwaarden (Pishro-Nik, 2014, Grimmett en Welsh, 2014):

1. N_t is een stochastische variabele die waardes aan neemt in $\{0, 1, 2, \dots\}$,
2. $N_0 = 0$, op tijdstip nul zijn er nul aankomsten,
3. $N_s \leq N_t$ als $s \leq t$, het is dus een stijgend proces,
4. Voor $0 \leq s < t$ is $N_t - N_s$ het aantal aankomsten in het interval $(s, t]$. Dit aantal is onafhankelijk van het aantal aankomst voorafgaand aan s dus in het interval $[0, s)$

Een telproces heeft onafhankelijke aangroeiingen als voor $0 \leq t_1 < t_2 < \cdots < t_n$ de stochastische variabelen $N_{t_2} - N_{t_1}$, $N_{t_3} - N_{t_2}$, …, $N_{t_n} - N_{t_{n-1}}$ onafhankelijk zijn van elkaar. Merk op

dat $N_{t_i} - N_{t_{i-1}}$ het aantal aankomsten zijn in het tijdsinterval $(t_{i-1}, t_i]$. Daarom heeft een telproces onafhankelijk aangroeiingen als de aantalen aankomsten in disjuncte (niet-overlappende) intervallen onafhankelijk zijn van elkaar (Pishro-Nik, 2014).

Een speciaal soort telproces is het Poissonproces. Er zijn meerdere manieren om een Poissonproces te definiëren waarvan één via de Poissonverdeling.

$N = (N_t : t \in [0, \infty))$ is een Poissonproces met intensiteit λ als N voldoet aan de vier bovenstaande eisen van een telproces én als het aantal aankomsten tot en met tijdstip t een Poissonverdeling heeft met parameter λt. λ is het verwachte aantal aankomsten per interval van lengte 1. Dat betekent dat voor $t > 0$ de kans op j aankomsten in het tijdsinterval $[0, t]$ gegeven is door (Grimmett en Welsh, 2014)

$$\mathbb{P}(N_t = j) = \frac{(\lambda t)^j e^{-\lambda t}}{j!} \qquad \text{voor } j = 0, 1, 2, \ldots.$$

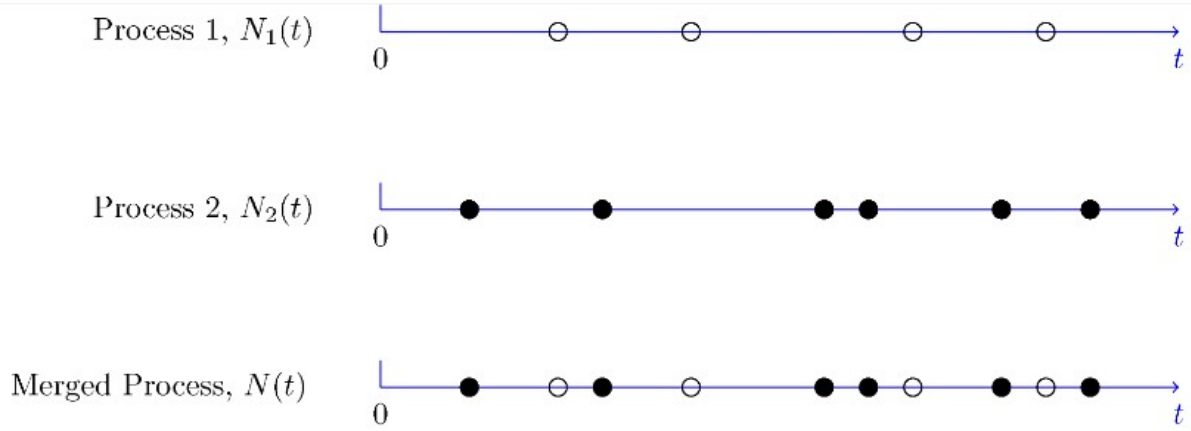
Uit deze definitie van een Poissonproces volgt dat de verwachting en de variantie van het aantal aankomsten tot en met tijdstip t, N_t, gelijk is aan elkaar en lineair toeneemt als t toeneemt, namelijk:

$$\mathbb{E}[N_t] = \lambda t, \qquad \text{var}(N_t) = \lambda t \qquad \text{voor } t > 0.$$

De intensiteit λ bepaalt of het Poissonproces homogeen of inhomogeen is. Als λ constant is in de tijd, dan is het Poissonproces homogeen. Als λ varieert in de tijd, dan is het Poissonproces inhomogeen.

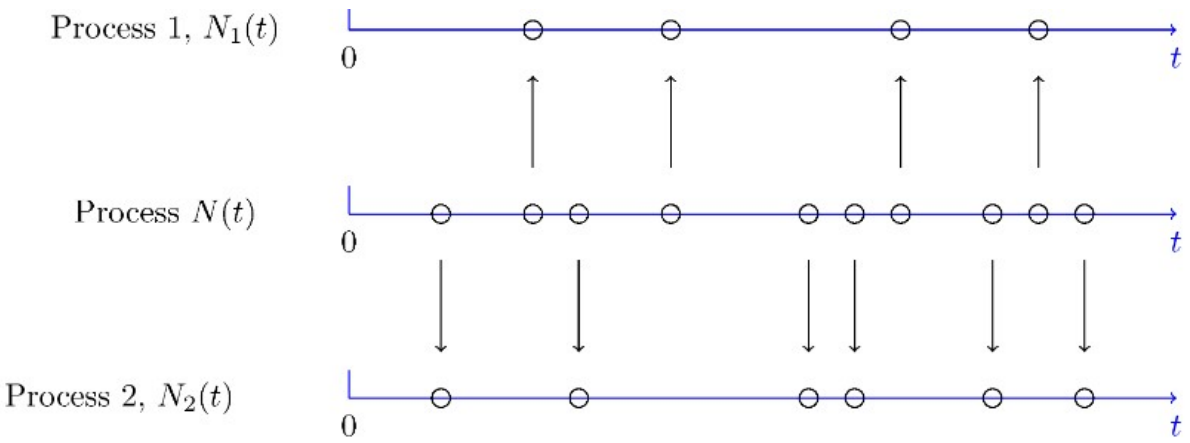
Een homogeen Poissonproces heeft een aantal mooie eigenschappen waaronder het samenvoegen en splitsen van Poissonprocessen.

Neem aan dat N_1 een Poissonproces is met intensiteit λ_1 en dat N_2 een Poissonproces is met intensiteit λ_2 onafhankelijk van N_1. Het samengevoegde proces $N = N_1 + N_2$ is dan een Poissonproces met intensiteit $\lambda_1 + \lambda_2$ (Grimmett en Welsh, 2014). Dit is gevisualiseerd in figuur 3.1.



Figuur 3.1: Het samenvoegen van twee onafhankelijke Poissonprocessen. Overgenomen van Pishro-Nik, 2014.

Een Poissonproces kan ook gesplitst worden. Laat N een Poissonproces zijn met intensiteit λ, het proces kan op de volgende manier gesplitst worden in twee processen N_1 en N_2. Elke aankomst van het Poissonproces N gaat met kans p naar proces N_1 en met kans $1 - p$ naar proces N_2. N_1 en N_2 zijn dan onafhankelijk Poissonprocessen met intensiteit λp voor N_1 en $\lambda(1 - p)$ voor N_2. Dit is gevisualiseerd in figuur 3.2.



Figuur 3.2: Het splitsen van een Poissonproces in twee onafhankelijke Poissonprocessen. Overgenomen van Pishro-Nik, 2014.

Een andere eigenschap en ook definitie van een homogeen Poissonproces heeft te maken met de tussenaankomsttijden. De tussenaankomsttijd is de tijd tussen twee opeenvolgende aankomsten. Voor een homogeen Poissonproces met intensiteit λ zijn de tussenaankomsttijden onafhankelijke stochastische variabelen met de exponentiële verdeling met parameter λ.

Volgens Grimmett en Welsh, 2014 is dit niet alleen een eigenschap van een Poissonproces maar ook een manier om het Poissonproces te definiëren. Namelijk: als de tussenaankomsttijden van een proces onafhankelijk en exponentieel verdeeld zijn met parameter λ, dan is het een Poissonproces met intensiteit λ.

Een kenmerk van de exponentiële verdeling is de geheugenloosheid. Wat er in het verleden gebeurd is, heeft vanwege de geheugenloosheid geen invloed op wat op dit moment gaat gebeuren. Voor een Poissonproces betekent de geheugenloosheid dat de tijd van de eerdere aankomsten geen invloed heeft op de tijd van de volgende aankomst (Grimmett en Welsh, 2014).

Laat T_i de tussenaankomsttijd zijn van aankomsten i en $i-1$, dus de tijd tussen aankomst i en aankomst $i-1$. Gegeven dat we al tijd s wachten op de volgende aankomst, dan is de kans om nog tijd t extra te moeten wachten net zo groot als dat we vanaf het begin tijd t moeten wachten. Wiskundig is dit als volgt in kansen uit te drukken (Grimmett en Welsh, 2014):

$$\mathbb{P}\left(T_i > s + t \mid T_i > s\right) = \mathbb{P}\left(T_i > t\right)$$

Dat de tussenaankomsttijden onafhankelijk en exponentieel verdeeld zijn, geldt alleen voor een homogeen Poissonproces. Voor een inhomogeen Poissonproces zijn de tussenaankomsttijden niet meer onafhankelijk en ook niet meer exponentieel verdeeld (Yakovlev e.a., 2005).

3.2. Drie mogelijkheden om naar een Poissonproces te kijken

Met de Poissonverdeling van het aantal aankomsten en de exponentiële verdeling van de tussenaankomsttijden hebben we twee manieren om een homogeen Poissonproces te definiëren. Via deze definities kunnen we op verschillende manieren naar een Poissonproces kijken, waarvan er hier drie worden uitgelegd.

De eerste manier is via de Poissonverdeling. In de definitie zeiden we dat het aantal aankomsten tot en met tijdstip t Poisson-verdeeld moet zijn met parameter λt. Dit geldt niet

alleen voor het aantal aankomsten in het interval $[0, t]$ maar voor ieder interval van lengte t. Voor een Poissonproces met intensiteit λ moet het aantal aankomsten in een interval van lengte t dus Poisson-verdeeld zijn met parameter λt. Neem je een interval van lengte 1, dan is het aantal aankomsten in dat interval dus Poisson-verdeeld met parameter λ. Belangrijk hierbij is dat het aantal aankomsten in disjuncte intervallen onafhankelijk is van elkaar. Zo heeft het aantal aankomsten in het interval $[0, t]$ bijvoorbeeld geen invloed op het aantal aankomsten in het interval $[3t, 4t]$.

Een tweede manier is via de exponentiële verdeling van de tussenaankomsttijden. Voor een Poissonproces met intensiteit λ, zijn de tussenaankomsttijden onafhankelijk en exponentieel verdeeld met parameter λ.

Een derde manier om naar een Poissonproces te kijken is via de tijd dat het duurt tot er een bepaald aantal aankomsten is. Stel we willen weten hoe lang het duurt tot er n aankomsten zijn. Deze tijd is de tijd tot de eerste aankomst, plus de tijd tussen de eerste en tweede aankomst, plus de tijd tussen de tweede en derde aankomst, en zo door tot de tijd tussen de $(n-1)$-ste en n-de aankomst. Dit is de som van de eerste n tussenaankomsttijden, het is dus een som van n onafhankelijke exponentieel(λ) verdeelde stochasten. Daarom is de tijd tot de n-de aankomst Gamma(n, λ) verdeeld (Kim, 2019).

3.3. Mogelijkheden en onmogelijkheden van de data

De data die beschikbaar is bestaat uit de tijdstippen waarop de druppels zijn gevallen en uit het aantal druppels per interval van 10 seconden. Zoals in paragraaf 2.4 beschreven, hebben we van al deze data maar een klein deel gebruikt. De gebruikte data is verzameld op Pole Pole van 1 mei 12:00 uur tot 3 mei 10:30 uur. Tijdens die periode zijn er 15.545 druppels gevallen, de tijdstippen waarop deze druppels gevallen zijn noemen we de aankomsttijden en geven we aan met $X_1, X_2, \ldots, X_n$ met n =15.545. X_1 is dus de tijd waarop de eerste druppel is gevallen en X_i is de tijd waarop de i-de druppel is gevallen. Verdelen we de tijdsperiode in intervallen van 10 seconden, dan krijgen we 16.734 tijdsintervallen. Het aantal druppels in die intervallen noemen we $Y_1, Y_2, \ldots, Y_m$ met m =16.734. Y_1 is dan het aantal druppels in het interval $[0, 10)$ en Y_i is het aantal druppels in het interval $[10\,(i-1), 10i)$.

Deze data kunnen we gebruiken om te toetsen of de regen met een homogeen of inhomogeen Poissonproces kan worden beschreven. De mogelijkheden en onmogelijkheden van de data worden hieronder besproken.

3.3.1. Homogeen Poissonproces

Toetsen op een homogeen Poissonproces zou vrij makkelijk moeten kunnen. Als de regen namelijk een homogeen Poissonproces is kunnen we twee dingen doen:

Als eerste kunnen we het aantal druppels per interval van 10 seconden $Y_1, \ldots, Y_m$ toetsen op een Poissonverdeling. Neem in dit geval 10 seconden als eenheidslengte zodat ieder interval als het ware lengte 1 heeft. Als de regen een homogeen Poissonproces met intensiteit λ volgt, dan weten we van de definitie van een Poissonproces dat $Y_1, \ldots, Y_m$ onafhankelijk en Poisson-verdeeld moeten zijn met parameter λ. Verwerpen we dat $Y_1, \ldots, Y_m$ Poisson-verdeeld zijn met parameter λ, dan verwerpen we ook dat de regen met een homogeen Poissonproces met intensiteit λ kan worden beschreven. Verwerpen we niet dat $Y_1, \ldots, Y_m$ Poisson-verdeeld zijn met parameter λ, dan kan de regen met een homogeen Poissonproces beschreven worden.

Een tweede mogelijkheid is om de tussenaankomsttijden te toetsen op een exponentiële verdeling. Neem de aankomsttijden $X_1, \ldots, X_n$, de tussenaankomsttijden kunnen we bereke-

nen door het verschil te nemen van één aankomsttijd en de vorige aankomsttijd. Noem de tussenaankomsttijden $T_1, \dots, T_n$. T_1 is dan gelijk aan X_1, T_2 is de tijd tussen aankomst 1 en 2 en is dus gelijk aan $X_2 - X_1$, algemeen geldt dat $T_i = X_i - X_{i-1}$.

Als de regen met een homogeen Poissonproces met intensiteit λ kan worden beschreven, dan weten we van de definitie van een Poissonproces dat de tussenaankomsttijden $T_1, \dots, T_n$ onafhankelijk en exponentieel verdeeld moeten zijn met parameter λ. Verwerpen we dat $T_1, \dots, T_n$ exponentieel verdeeld zijn met parameter λ, dan verwerpen we ook dat de regen met een homogeen Poissonproces met intensiteit λ kan worden beschreven. Verwerpen we niet dat $T_1, \dots, T_n$ exponentieel verdeeld zijn met parameter λ, dan kan de regen met een homogeen Poissonproces worden beschreven.

3.3.2. Inhomogeen Poissonproces

Toetsen op een inhomogeen Poissonproces zal een stuk lastiger zijn. De intensiteit van het proces is niet constant maar varieert in de tijd. Neem je nu het aantal druppels per interval van 10 seconden, dan zal dit niet meer Poisson-verdeeld zijn met parameter λ. Wel kan de data in andere (kleinere) intervallen ingedeeld worden met elk een eigen constante intensiteit (de Villiers e.a., 2021).

Het lastige is nu dat je altijd een intensiteit kan vinden die bij een interval past als je de intervallen maar klein genoeg maakt (de Villiers e.a., 2021). Voor een regenbui zou dit betekenen dat je er altijd een inhomogeen Poissonproces omheen kan bouwen door de intervallen klein genoeg te maken en voor ieder interval de intensiteit te schatten. Om na te gaan of deze schatting realistisch is, zou je het liefst meer data van dezelfde locatie willen hebben.

Zouden er meerdere intervalometers op dezelfde locatie staan, dan zou je de data van één intervalometer kunnen gebruiken om de intensiteit te schatten, en deze schatting vervolgens kunnen verifiëren met de data van de andere intervalometer(s). Als we nu aannemen dat de regen in de ruimte homogeen verdeeld is (niet perse Poisson), dan zou de intensiteit van dezelfde bui bij de verschillende intervalometers hetzelfde moeten zijn.

Behalve bij MIL 1 staat er op ieder locatie één intervalometer waardoor het met de uitgekozen data van Pole Pole niet mogelijk is om op bovenstaande manier de intensiteit te schatten en te verifiëren. Wel heeft Pole Pole naast een intervalometer nog een tipping bucket en een disdrometer.

Deze meetinstrumenten hebben data verzameld van dezelfde buien op dezelfde locatie als de intervalometer. De data van deze verschillende meetinstrumenten zouden we kunnen gebruiken om de intensiteit te schatten en te verifiëren. Waar dan rekening mee gehouden moet worden, is dat het oppervlak waarop de druppels vallen anders is voor de verschillende meetinstrumenten. Op een groter oppervlak kunnen meer druppels vallen waardoor de intensiteiten niet één op één hetzelfde zullen zijn als je de intensiteit van dezelfde bui schat aan de hand van de data van de verschillende meetinstrumenten. De intensiteit van het ene meetinstrument zal waarschijnlijk een constante maal de intensiteit van het andere meetinstrument zijn.

Een andere mogelijkheid zou zijn om intervalometerdata te gebruiken van verschillende locaties die (redelijk) dicht bij elkaar liggen. Als we kunnen aannemen dat een regenbui in de ruimte homogeen verdeeld is over een bepaalde afstand, en als de intervalometers binnen deze afstand staan, dan zullen ze data verzamelen van dezelfde regen. Mogelijke locaties hiervoor zijn MIL 1 en 2 die 72 meter uit elkaar liggen en Pole Pole en Kinasi die 90 meter uit elkaar liggen.

Nog een andere mogelijkheid is om te sampelen uit de intervalometerdata. Zo kun je als het ware simuleren dat je data hebt van meerdere intervalometers van dezelfde regenbui. Die data kan je gebruiken om de intensiteit te schatten en te verifiëren.

We zullen nu eerst in hoofdstuk 4 de data van Pole Pole toetsen op een homogeen Poissonproces. Daarna zullen we ook toetsen op een inhomogeen Poissonproces in 5. Hiervoor zullen we gebruikmaken van een methode om te sampelen uit de beschikbare data.

4

Toetsen voor een homogeen Poissonproces

In dit hoofdstuk wordt getoetst of de regen van de data volgens een homogeen Poissonproces valt. Als eerste geven paragrafen 4.1 t/m 4.5 een uitgebreid overzicht van verschillende toetsen en toetsingsgrootheden uit de literatuur. De toetsen zijn ingedeeld in verschillende categorieën gebaseerd op Gürtler en Henze, 2000 en Karlis en Xekalaki, 2000. In deze paragrafen gaat het echt om het overzicht en is het doel niet om alle verschillende toetsen toe te passen. Daarna wordt in paragraaf 4.6 één van de toetsen gekozen. Als laatste wordt in paragraaf 4.7 deze toets toegepast op de regendata.

In Gürtler en Henze, 2000 gaan ze uit van $X_1, \dots, X_n$ onafhankelijke waarnemingen van de stochastische variabele X met onbekende verdelingsfunctie F. De verdelingsfunctie van een Poisson-verdeelde stochast met intensiteit λ is $F(j;\lambda) \coloneqq e^{-\lambda} \sum_{l=0}^{j} \lambda^l / l!,\ j = 0, 1, 2 \dots$. De hypothese die in dit artikel getoetst wordt, is

$$\mathrm{H}_0 : X_1, \dots, X_n \text{ zijn Poisson}(\lambda) \text{ verdeeld, ofwel } \ F = F(\cdot;\lambda) \quad \text{voor een } \lambda > 0$$

tegen een algemeen alternatief.

In Karlis en Xekalaki, 2000 wordt H_0: de data komt van één Poissonverdeling, getoetst tegen H_1: de data komt niet van een Poissonverdeling. 'De data' is hier niet gespecificeerd.

Zij $Y_1, \dots, Y_m$ met m =16.734 het aantal druppels per interval van 10 seconden net als in paragraaf 3.3.

Als de regen op Pole Pole met een homogeen Poissonproces met intensiteit λ kan worden beschreven, dan weten we van de definitie van een Poissonproces dat $Y_1, \dots, Y_m$ onafhankelijk en Poisson-verdeeld moeten zijn met parameter λ.

Wordt H_0 verworpen dan zijn $Y_1, \dots, Y_m$ geen onafhankelijke Poisson(λ)-verdeelde stochastische variabelen en kunnen we concluderen dat de regenval op Pole Pole niet met een homogeen Poissonproces met parameter λ kan worden beschreven.

Wordt H_0 niet verworpen, dan hoeft dat niet te betekenen dat H_0 waar is. H_0 niet verwerpen is dus niet hetzelfde als concluderen dat de regen op Pole Pole met een homogeen Poissonproces kan worden beschreven.

Om H_0 te kunnen toetsen hebben we een toetsingsgrootheid nodig. In de volgende paragrafen staat een uitgebreid overzicht van verschillende toetsingsgrootheden uit de literatuur.

4.1. Toetsen gebaseerd op de empirische verdelingsfunctie

Laat de empirische verdelingsfunctie van $X_1, \ldots, X_n$ gegeven zijn door $F_n(j) \coloneqq n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq j\}}$ met $\mathbb{1}_{\{X_i \leq j\}}$ de indicatorfunctie van de gebeurtenis $\{X_i \leq j\}$

4.1.1. Kolmogorov-Smirnov toetsen

Gürtler en Henze, 2000 geven als toetsingsgrootheid

$$K_n \coloneqq \sup_{j \geq 0} \sqrt{n} \left| F_n(j) - F(j; \hat{\lambda}) \right|$$

met $\hat{\lambda}$ een schatting voor de parameter λ.

4.1.2. Cramér-von Mises toetsen

Gürtler en Henze, 2000 geven als toetsingsgrootheid

$$C_n \coloneqq n \sum_{j=0}^{\infty} \left[F_n(j) - F(j; \hat{\lambda}) \right]^2 f(j, \hat{\lambda})$$

met $f(j, \hat{\lambda})$ de sprong van $F(\cdot, \hat{\lambda})$ in j.

Karlis en Xekalaki, 2000 geven vier verschillende Cramér-von Mises toetsingsgrootheden:

$$\mathrm{CVM}_1 = n^{-1} \sum_{j=0}^{M} Z_j^2 p_j^{(\hat{\lambda})}$$

$$\mathrm{CVM}_2 = n^{-1} \sum_{j=0}^{M} \frac{Z_j^2 p_j^{(\hat{\lambda})}}{H_j (1 - H_j)}$$

$$\mathrm{CVM}_3 = n^{-1} \sum_{j=0}^{M} Z_j^2$$

$$\mathrm{CVM}_4 = n^{-2} \sum_{j=0}^{M} Z_j^2 O_j$$

met $Z_j = \sum_{i=0}^{j} \left(O_i - E_i^{(\hat{\lambda})} \right)$, O_i het waargenomen aantal waarnemingen gelijk aan i, $E_i^{(\hat{\lambda})}$ het verwachte aantal waarnemingen gelijk aan i onder de nulhypothese, $H_j = \sum_{i=0}^{j} p_i^{(\hat{\lambda})}$ en $p_i^{(\hat{\lambda})}$ de kans op waarde i onder de nulhypothese. De bovengrens van de som, M, kan zo gekozen worden dat $p_M < 0.0001$. CVM_1, CVM_2 en CVM_3 zijn van Spinelli en Stephens, 1997 en CVM_4 is van Henze, 1996.

4.1.3. Overig

Als laatste stellen Gürtler en Henze, 2000

$$L_n \coloneqq \sum_{j \geq 0} \sqrt{n} \left| F_n(j) - F(j; \hat{\lambda}) \right|$$

voor als toetsingsgrootheid.

4.2. Toetsen gebaseerd op de geïntegreerde empirische verdelingsfunctie

De geïntegreerde verdelingsfunctie Ψ van de variabele X die Gürtler en Henze, 2000 gebruiken, is gedefinieerd als $\Psi(t) \coloneqq E(X-t)^+ = \int_t^\infty \big(1-F(x)\big)\mathrm{d}x$. De geïntegreerde empirische verdelingsfunctie die Gürtler en Henze, 2000 gebruiken, is gegeven door $\Psi_n(t) \coloneqq \int_t^\infty \big(1-F_n(x)\big)\mathrm{d}x = n^{-1}\sum_{i=1}^n (X_i - t)\mathbb{1}_{\{X_i>t\}}$. Een voordeel van deze definitie is dat de functie begrensd is als $\mathbb{E}\,[X]$ bestaat.

Een schatter voor de geïntegreerde verdelingsfunctie onder Poissoniteit is dan $\hat{\Psi}_n(t) \coloneqq \int_t^\infty \big(1-F(x;\hat{\lambda})\big)\mathrm{d}x$. Gürtler en Henze, 2000 en Karlis en Xekalaki, 2000 geven een type Kolmogorov-Smirnov toetsingsgrootheid namelijk

$$I_n \coloneqq \sup_{t\geq 0} \sqrt{n}\left|\Psi_n(t) - \hat{\Psi}_n(t)\right|$$

Beide artikelen verwijzen naar het artikel van Klar, 1999 waarin de toetsingsgrootheid ook voor discrete data wordt uitgewerkt:

$$I_n = \max_{0\leq k\leq M} \sqrt{n}\left|\sum_{j=0}^{k-1}\big(F(j,\hat{\lambda}) - F_n(j)\big)\right|$$

met $M = \max_{1\leq i\leq n} X_i$.

4.3. Toetsen gebaseerd op de kansgenererende functie

De kansgenererende functie van X is $g(u) \coloneqq \mathbb{E}[u^X] = \sum_{j=0}^\infty P(X=j)u^j$ en kan geschat worden met de empirische kansgenererende functie $g_n(u) \coloneqq n^{-1}\sum_{i=1}^n u^{X_i}$. Als X Poisson(λ) verdeeld is, dan is de kansgenererende functie $g(u) = e^{\lambda(u-1)}$

Om H_0 te toetsen kan gebruik worden gemaakt van het 'empirische kansgenererende functie'-proces met geschatte parameter

$$G_n(u) \coloneqq \sqrt{n}\big(g_n(u) - g(u;\hat{\lambda})\big) = \sqrt{n}\left(g_n(u) - e^{\hat{\lambda}(u-1)}\right)$$

voor $0 \leq u \leq 1$ en waar de tweede $=$ volgt uit de kansgenererende functie van een Poissonproces met parameter λ. Toetsingsgrootheden voorgesteld door Gürtler en Henze, 2000 zijn:

- $R_n \coloneqq \int_0^1 G_n^2(u)\mathrm{d}u$
- $R_{n,a} \coloneqq \int_0^1 G_n^2(u)u^a\mathrm{d}u$, een gewogen type Cramér-von Mises grootheid
- $T_n \coloneqq n\int_0^1 [\overline{X}_n g_n(u) - g_n'(u)]^2\mathrm{d}u$, gebaseerd op de differentiaal vergelijking $g'(u) = \lambda g(u)$
  Karlis en Xekalaki, 2000 hebben ook een toetsingsgrootheid gebaseerd op deze differentiaalvergelijking $T_n = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\left(\frac{\overline{X}^2}{X_i+X_j+1} + \frac{X_iX_j}{X_i+X_j-1}\right) - (n-f_0)\overline{X}$, met f_0 het deel van de waarnemingen gelijk aan nul

- $T_{n,a} \coloneqq n \int_0^1 [\overline{X}_n g_n(u) - g_n'(u)]^2 u^a \mathrm{d}u$

- $V_n \coloneqq \frac{1}{n^3} \sum_{i,j,k,l=1}^{n} X_i(X_i - X_j - 1)X_k(X_k - X_l - 1)\mathbb{1}_{\{X_i+X_j=X_k+X_l\}}$, gebaseerd op de differentiaal vergelijking $\partial^2/\partial u^2 \left[\log\left(g(u;\lambda)\right)\right] = 0$

- $V_n^* \coloneqq V_n/(\overline{X}_n)^{1,45}$
  Deze toetsingsgrootheid wordt ook door Karlis en Xekalaki, 2000 genoemd.

Toetsingsgrootheden voorgesteld door Karlis en Xekalaki, 2000, deels overeenkomende met de hierboven genoemde grootheden, zijn:

- $K = \frac{G_n(u)}{\exp\{\hat{\lambda}(u^2-1)\} - \exp\{2\hat{\lambda}(u-1)\}\{1+\hat{\lambda}(u-1)^2\}}$. Een nadeel volgens Karlis en Xekalaki, 2000 is dat K afhangt van u.

- Een toetsingsgrootheid die niet van u afhangt is

$$d_n(\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{X_i + X_j + 1} - 2\exp(-\hat{\lambda})\sum_{i=1}^{n} T(X_i, \hat{\lambda}) + n\frac{1-\exp(-2\hat{\lambda})}{2\hat{\lambda}},$$

  met $T(x,\hat{\lambda}) = \int_0^1 u^x \exp(\hat{\lambda}u)\mathrm{d}u$. $T(x,\hat{\lambda})$ kan ook recursief berekend worden (zie p.364 van Karlis en Xekalaki, 2000).

- T_n in vorige opsomming al genoemd

- V_n^* in vorige opsomming al genoemd

4.4. Toetsen gebaseerd op Fisher's dispersie-index

Zoals in hoofdstuk 3 is uitgelegd, heeft een Poisson-verdeelde stochast gelijke verwachting en variantie. Delen we de variantie door de verwachting, dan zou daar 1 uit moeten komen. Dit is waar Fisher's dispersie-index op gebaseerd is.

Fisher's dispersie-index is door Gürtler en Henze, 2000 gedefinieerd als

$$D_n \coloneqq \sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\overline{X}} = (n-1)\frac{S^2}{\overline{X}}$$

met $\overline{X}$ het gemiddelde en S^2 de variantie van de steekproef. H_0 wordt verworpen voor grote of kleine waarden van D_n. In Karlis en Xekalaki, 2000 noemen ze dit de variantie toets en geven ze een geschaalde versie met toetsingsgrootheid

$$S^* = \frac{\overline{X}(D_n - n)^2}{\sum_{i=1}^{n}\left((X_i - \overline{X})^2 - X_i\right)}$$

Een andere toetsingsgrootheid die H_0 verwerpt voor alleen grote waarden is

$$\hat{U}_{n2}^2 \coloneqq \left(\frac{1}{\sqrt{2n}}(D_n - n)\right)^2$$

De toetsingsgrootheid

$$Z_j = \frac{k_j - \overline{X}}{\sqrt{\operatorname{var}\left(k_j|X\right)}}, \qquad \text{voor } j = 2, 3, 4$$

gebruikt het feit dat de cumulanten van de Poissonverdeling optellen tot de Poissonparameter λ.

De cumulanten k_j zijn de coefficenten van de Taylorreeks van de cumulant genererende functie rond de oorsprong. De cumulant genererende functie is de natuurlijke logaritme van de moment genererende functie ("Cumulants", 2013):

$$K_X(t) = \ln M_X(t) = \ln \mathbb{E}\left[e^{tX}\right] = \sum_{j=1}^{\infty} k_j \frac{t^j}{j!}$$

In de toetsingsgrootheid is k_j de j-de cumulant en $\operatorname{var}\left(k_j|X\right)$ de variantie van de j-de cumulant gegeven de som van alle observaties. Uitwerkingen van Z_2, Z_3 en Z_4 staan op p.359 van Karlis en Xekalaki, 2000.

Andere toetsingsgrootheden genoemd door Karlis en Xekalaki, 2000 zijn:

- Een algemene twee-parameter gamma-verdeling waarbij de eerste twee momenten van de dispersie-index overeenkomen met de momenten van de Gamma-verdeling.
- Om $\mathrm{H}_0 : \sigma^2 = \mu$ te toetsen, heb je toetsingsgrootheid $O_2 = \sqrt{\left(\frac{n-1}{2}\right)\left(\frac{S^2}{\overline{X}} - 1\right)} = \frac{\sum_i (X_i - \overline{X})^2}{\overline{X}\sqrt{2(n-1)}} - \frac{\sqrt{n-1}}{\sqrt{2}}$.
- Om een één-parameter Poissonverdeling te toetsen tegen een gemengde Poissonverdeling met dezelfde verwachting heb je $Z = \frac{\sum_i (X_i - \overline{X})^2}{\overline{X}\sqrt{2n}} - \sqrt{\frac{n}{2}}$,
- Om de Poissonaanname te toetsen tegen een gegeneraliseerde Poissonverdeling met kansdichtheid $P(X = x) = \lambda\left(\lambda + x\theta\right)^{x-1} \exp\left\{-\lambda + x\theta\right\} / x!$, is er de toetsingsgrootheid $W = \left(\frac{S^2}{\overline{X}} - 1\right)^2 \frac{n}{2}$.
- $S_k = \sum_{i=2}^{k} V_i^2$, met V_i de Poisson-Charlier polynomen van orde k. Zie p.361 van Karlis en Xekalaki, 2000 voor een definitie van V_i en een recurrente relatie die daaruit volgt. Voor $k = 2$ is de toetsingsgrootheid hetzelfde als Z die eerder in deze opsomming staat.
- Een toets die de coëfficiënten van scheefheid en kurtosis gebruiken met als toetsingsgrootheid $T = \frac{1}{2}\sqrt{\frac{n}{1 + 24\overline{X} + 6\overline{X}^2}} \frac{m_2(m_4 - 3m_2^2) - m_3}{\overline{X^2}}$, met $m_k = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
- Cox (1983) heeft de variantie toets veralgemeniseerd zodat hij geldig is voor alle exponentiële families zowel discreet als continu.
- Gelfand en Dalal (1990) hebben een toets voor overdispersie die lijkt op die van Cox. Ze hebben ook een grafische toets voor overdispersie gevonden.
- Voor andere manieren om grafisch de Poisson aanname te checken verwijzen Karlis en Xekalaki, 2000 naar Ord, 1967, Lindsay, 1986, Santner en Duffy, 1989, Hoagline, 1980 en Lindsay en Roeder, 1992.

- Pettigrew en Mohler hebben een toets gebaseerd op de range.
- Lee heeft een toets gemaakt die gebaseerd is op een uitbreiding van de verdelingsklasse afhangend van een parameter. De Poissonverdeling heeft parameter 0.
- De vorige toets van Lee leidt tot een variant van de variantie toets die een dispersie parameter toevoegt aan de exponentiële familie. De toetsingsgrootheid voor de Poissonverdeling om te toetsen of de dispersie parameter 0 is, is $\mathrm{EF} = 2\sum_{i=1}^{n}\left\{\overline{X} - X_i + X_i \ln\left(\frac{X_i}{\overline{X}}\right)\right\} = 2\sum_{i=1}^{n} X_i \ln\left(\frac{X_i}{\overline{X}}\right)$

4.5. Overige toetsen

Karlis en Xekalaki, 2000 geven nog vijf verschillende goodness-of-fit toetsen die niet in één van de vorige vier categorieën passen. De eerste is de X^2-toets met toetsingsgrootheid

$$\mathrm{X}^2 = \sum_{i=0}^{k} \frac{\left(O_i - E_i^{(\hat{\lambda})}\right)^2}{E_i^{(\hat{\lambda})}}$$

met O_i de waargenomen aantallen van i en $E_i^{(\hat{\lambda})}$ de verwachte aantallen van i onder de nulhypothese. Een voordeel van de toetsingsgrootheid is dat deze makkelijk te berekenen is.

De tweede toets is de (multinomial) likelihood ratio test (LRT) met een toetsingsgrootheid die ook asymptotisch X^2-verdeeld is. Een groot nadeel is dat de benodigde berekeningen onbetaalbaar zijn.

Een derde toets is ontworpen door Nass in 1959 met toetsingsgrootheid

$$N = \frac{\sum_{i=0}^{m} \frac{O_i^2}{E_i^{(\hat{\lambda})}} - n - (m-1)}{\sqrt{\left[\frac{m-1}{m}\left\{2m - \frac{(m+1)^2+2m}{n} + \sum_{i=0}^{m} \frac{1}{E_i^{(\hat{\lambda})}}\right\}\right]}}$$

met n de grootte van de steekproef en m het grootst waargenomen aantal.

Een vierde goodness-of-fit toets is een groep toetsen die de power divergent toetsingsgrootheid van Read en Cressie (1988) gebruikt

$$I^{\lambda} = \frac{1}{\lambda\,(\lambda+1)} \sum_{i=0}^{m} E_i^{(\hat{\lambda})}\left\{\left(\frac{O_i}{E_i^{(\hat{\lambda})}}\right)^{\lambda+1} - 1\right\}, \qquad \lambda \in \mathbb{R}$$

Bijzondere gevallen van de power divergent toetsingsgrootheid zijn:

1. $\lambda = 1$ is de X^2-toetsingsgrootheid
2. $\lambda = 0$ is de multinomial likelihood ratio toetstingsgrootheid als de limiet voor λ naar 0 wordt genomen.
3. $\lambda = -2$ is de Neyman X^2-toets

4. $\lambda = \frac{1}{2}$ is de Freeman-Tukey toets

volgens Read en Cressie is $\lambda = \frac{2}{3}$ een erg goed compromis.

De vijfde en laatste goodness-of-fit toets is de minimum disparity goodness-of-fit toets van Basu en Sarkar (1994). Deze toets is gerelateerd aan de vorige toetsen.

Een andere toets die niet in één van de vorige vier categorieën passen is de Likelihood ratio toets. De likelihood ratio toets kan volgens Karlis en Xekalaki, 2000 goed gebruikt worden als de nulhypothese een speciaal geval is van de alternatieve hypothese. In hun artikel gebruiken ze deze toets om een Poissonverdeling te toetsen tegen een gemengde Poissonverdeling met twee parameters. De toetsingsgrootheid is

$$L = 2(L_1 - L_0)$$

met $L_i,\ i = 0, 1$ de gemaximaliseerde log-likelihoods van de hypotheses H_i.

4.6. Toets kiezen

Om te bepalen of de regen met een homogeen Poissonproces kan worden beschreven, is het belangrijk om de juiste toets te kiezen. Bij het kiezen van de juiste toets maakt het uit hoe je naar een Poissonproces kijkt (zie hoofdstuk 3) en wat je wil toetsen. Wil je weten of het aantal druppels per tijdsinterval Poisson-verdeeld is, wil je weten of de tussenaankomsttijden exponentieel verdeeld zijn of wil je weten of de som van de tussenaankomsttijdens gamma verdeeld is?

De toetsen hierboven zijn allemaal gericht op het toetsen van een Poissonverdeling van het aantal druppels per tijdsinterval. Sommige van deze toetsen zijn ook te gebruiken om te toetsen op een exponentiële verdeling van de tussenaankomsttijden. Dit zijn onder andere de Chi-kwadraat toets en de Kolmogorov-Smirnov toets (Winton, 2009).

In Karlis en Xekalaki, 2000 en Gürtler en Henze, 2000 worden in de conclusie de ‘beste’ toetsingsgrootheden genoemd. Welke toetsingsgrootheden het beste zijn, wordt bepaald aan de hand van het onderscheidend vermogen van de toetsingsgrootheid bij verschillende alternatieven, hoe makkelijk de waarde van de toetsingsgrootheid te berekenen is en soms ook in hoeverre de verdeling van de toetsingsgrootheid onder de nulhypothese bekend is. Het onderscheidend vermogen van de toetsingsgrootheden wordt in beide artikelen bepaald met simulaties.

Kijken we naar de conclusies van beide artikelen, dan worden de volgende toetsingsgrootheden genoemd als goede toetsingsgrootheden, ingedeeld in de categorieën die eerder dit hoofdstuk ook gebruikt zijn. Voor details met betrekking tot de afwegingen van de auteurs verwijzen we naar de artikelen Gürtler en Henze, 2000 en Karlis en Xekalaki, 2000.

1. Toetsen gebaseerd op de empirische verdelingsfunctie:
    - Cramér-von Mises toetsingsgrootheden krijgen volgens Karlis en Xekalaki, 2000 de voorkeur boven Kolmogorov-Smirnov toetsingsgrootheden. Karlis en Xekalaki, 2000 verwijzen voor de verschillende Cramér-von Mises toetsingsgrootheden naar Spinelli en Stephens, 1997 en volgens hun gaat de voorkeur uit naar CVM_2.
    - $L_n := \sum_{j \geq 0} \sqrt{n} \left| F_n(j) - F(j; \hat{\lambda}) \right|$, krijgt volgens Gürtler en Henze, 2000 de voorkeur boven andere toetsen gebaseerd op de empirische verdelingsfunctie.
2. Toetsen gebaseerd op de geïntegreerde empirische verdelingsfunctie:

- $I_n \coloneqq \sup_{t\geq 0} \sqrt{n}\left|\Psi_n(t) - \hat{\Psi}_n(t)\right|$, is een type Kolmogorov-Smirnov toetsingsgrootheid en krijgt de voorkeur boven L_n.

3. Toetsen gebaseerd op de kansgenererende functie:

   - $T_{n,5} \coloneqq n\int_0^1 \left[\overline{X}_n g_n(u) - g_n'(u)\right]^2 u^5 \mathrm{d}u$, krijgt volgens Gürtler en Henze, 2000 de voorkeur boven L_n.
   - $R_{n,5} \coloneqq \int_0^1 G_n^2(u)u^5\mathrm{d}u$, een gewogen type Cramér-von Mises grootheid. $R_{n,5}$ krijgt volgens Gürtler en Henze, 2000 de voorkeur boven L_n.
   - $V_n^* \coloneqq V_n/\left(\overline{X}_n\right)^{1,45}$, is volgens Gürtler en Henze, 2000 goed als het alternatief een uniforme verdeling is.

4. Toetsen gebaseerd op Fisher's dispersie-index:

   - Alle toetsingsgrootheden gebaseerd op de variantie-gemiddelde ratio doen het goed bij alternatieve verdelingen met een variantie groter dan de verwachting (overdispersie) en bij alternatieven verdelingen met een variantie kleiner dan dan verwachting (onderdispersie). Bij alternatieve verdelingen met een variantie gelijk aan de verwachting (equidispersie) doen deze toetsingsgrootheden het niet goed.
   - $Z_4 = \frac{m_4 - 3S^2 - \overline{X}}{\sqrt{2n\overline{X}\left(n\overline{X}-1\right)\left(49 + \frac{108\left(n\overline{X}-2\right)}{n-2} + \frac{12(n+1)\left(n\overline{X}-2\right)\left(n\overline{X}-3\right)}{n(n-2)(n-3)}\right)}} n\sqrt{n-1}$, met $m_k = n^{-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$, lijkt het goed te doen bij een kleine steekproef en equidispersie.
   - $O_2 = \sqrt{\left(\frac{n-1}{2}\right)\left(\frac{S^2}{\overline{X}} - 1\right)} = \frac{\sum_i (X_i - \overline{X})^2}{\overline{X}\sqrt{2(n-1)}} - \frac{\sqrt{n-1}}{\sqrt{2}}$, is goed voor zowel over- als onderdispersie
   - $T = \frac{1}{2}\sqrt{\frac{n}{1+24\overline{X}+6\overline{X}^2}} \frac{m_2(m_4 - 3m_2^2) - m_3}{\overline{X^2}}$ lijkt het net als Z_4 goed te doen bij een kleine steekproef en equidispersie.

Ik kies er voor om de toetsingsgrootheid I_n te gebruiken. Volgens Gürtler en Henze, 2000 en Karlis en Xekalaki, 2000 krijgt I_n de voorkeur boven L_n en L_n krijgt dan weer de voorkeur boven Cramér-von Mises toetsingsgrootheden. Daarom krijgt I_n de voorkeur boven zowel L_n als Cramér-von Mises toetsingsgrootheden. I_n krijgt de voorkeur omdat voor de meeste alternatieven gebruikt in Gürtler en Henze, 2000, het onderscheidend vermogen van I_n groter is dan het onderscheidend vermogen van L_n en Cramér-von Mises toetsingsgrootheden. Het gaat dan om binomiale, negatief binomiale, gemengde en gegeneraliseerde Poisson-, zeta-, logaritmische en twee uniforme verdelingen. Tegen de andere drie gebruikte uniforme verdelingen was het onderscheidend vermogen van I_n lager dan het onderscheidend vermogen van L_n en Cramér-von Mises toetsingsgrootheden.

Een andere reden om voor I_n te kiezen is dat de toetsingsgrootheid volgens Gürtler en Henze, 2000 makkelijk en extreem snel te implementeren is. Dit is handig als je zelf simulaties wilt uitvoeren.

Een derde reden om voor I_n te kiezen is dat het een type Kolmogorov-Smirnov toetsingsgrootheid is en daarom ook gebruikt kan worden om te toetsen op een exponentiële verdeling van de tussenaankomsstijden Winton, 2009. Echter gelden bovenstaande argumenten over het onderscheidend vermogen van I_n tegen verschillende alternatieven dan niet meer. Het onderscheidend vermogen is namelijk bepaald met een Poissonverdeling in de nulhypothese en niet een exponentiële verdeling.

Een reden om te willen kunnen toetsen op een Poissonverdeling van het aantal regendruppels per tijdsinterval en op een exponentiële verdeling van de tussenaankomsttijden van de druppels is de volgende: het kan zo zijn dat je wel verwerpt als je een dataverzameling toetst op een Poissonverdeling van het aantal regendruppels per interval maar dat je niet verwerpt als je dezelfde dataverzameling toetst op een exponentiële verdeling van de tussenaankomsttijden.

4.7. Toets uitvoeren

Zowel Gürtler en Henze, 2000 als Karlis en Xekalaki, 2000 beschrijven in hun artikel kort de toetsingsgrootheid I_n en verwijzen allebei naar het artikel van Klar, 1999. De dataset in dit artikel bestaat uit onafhankelijke gelijk verdeelde stochastische variabelen $X_1, X_2, \dots, X_n$ met verdelingsfunctie F. Zij $\mathscr{F}$ een familie discrete verdelingen, de nulhypothese die getoetst wordt is $\mathrm{H}_0 : F \in \mathscr{F}$ tegen $\mathrm{H}_1 : F \notin \mathscr{F}$.

We willen toetsen op een Poissonverdeling van het aantal druppels per interval. Daarom gebruiken we als data $Y_1, \dots, Y_m$ met Y_i het aantal druppels gevallen in het interval $[10(i-1), 10i)$. $\mathscr{F}$ bestaat uit de verdelingsfuncties van de Poissonverdeling, $\mathscr{F} = \{F(j;\lambda) \coloneqq e^{-\lambda}\sum_{l=0}^{j}\lambda^l/l!,$ voor $\lambda > 0$ en $j = 0, 1, 2 \dots\}$. λ is hier de parameter en omdat we te maken hebben met een Poissonverdeling is λ gelijk aan het verwachte aantal druppels per interval van 10 seconden.

De nulhypothese wordt $\mathrm{H}_0 : F \in \mathscr{F}$ tegen $\mathrm{H}_1 : F \notin \mathscr{F}$. In woorden is dit H_0 : het aantal regendruppels per tijdsinterval volgt één Poissonverdeling met niet nader gespecificeerde waarde van λ, tegen H_1 : het aantal regendruppels per tijdsinterval volgt geen Poissonverdeling.

4.7.1. De toetsingsgrootheid

De toetsingsgrootheid I_n is gebaseerd op het verschil tussen de geïntegreerde verdelingsfunctie van Y, $\Psi(t) \coloneqq E(Y-t)^+ = \int_t^\infty \big(1 - F(y)\big)\mathrm{d}y$ en de empirische geïntegreerde verdelingsfunctie $\Psi_m(t)$.

$$I_n \coloneqq \sup_{t\geq 0}\sqrt{m}|\Psi_m(t) - \hat{\Psi}(t)|$$

Met

$$\Psi_m(t) = \int_t^\infty \big(1 - F_m(y)\big)\mathrm{d}y = \frac{1}{m}\sum_{i=1}^{m}(Y_i - t)\mathbb{1}_{\{Y_i > t\}}$$

$$\hat{\Psi}(t) = \int_t^\infty \left(1 - F\left(y, \hat{\lambda}\right)\right)\mathrm{d}y$$

Hierin is $F_m(y)$ de empirische verdelingsfunctie van $Y_1, \dots, Y_m$ en $\hat{\lambda}$ een schatter voor de parameter λ. $\mathbb{1}_{\{Y_i > t\}}$ is de indicatorfunctie van de gebeurtenis $\{Y_i > t\}$, de functie is dus 1 als Y_i groter is dan t en 0 als Y_i kleiner dan of gelijk is aan t. De toetsingsgrootheid kan als volgt worden omgeschreven:

$$
\begin{aligned}
I_n &= \sup_{t\geq 0} \sqrt{m}\left|\Psi_m(t) - \hat{\Psi}(t)\right| \\
&= \max_{0\leq k\leq M} \sqrt{m}\left|\Psi_m(k) - \hat{\Psi}(k)\right| \\
&= \max_{0\leq k\leq M} \sqrt{m}\left|\int_k^\infty \left(1 - F_m(y)\right)\mathrm{d}y - \int_k^\infty \left(1 - F\left(y;\hat{\lambda}\right)\right)\mathrm{d}y\right| \\
&= \max_{0\leq k\leq M} \sqrt{m}\left|\int_k^\infty \left(F\left(y;\hat{\lambda}\right) - F_m(y)\right)\right| \\
&= \max_{0\leq k\leq M} \sqrt{m}\left|\sum_{j=k}^\infty \left(F\left(j;\hat{\lambda}\right) - F_m(j)\right)\right| \\
&= \max_{0\leq k\leq M} \sqrt{m}\left|\sum_{j=0}^{k-1} \left(F\left(j;\hat{\lambda}\right) - F_m(j)\right)\right|
\end{aligned}
$$

met $M = \max_{1\leq i\leq m} Y_i$. De tweede gelijkheid volgt uit het feit dat Y een discrete stochast is. De laatste gelijkheid krijgen we door vergelijkingen 4.1 en 4.2 van elkaar af te halen:

$$\sum_{j=0}^{\infty}(1 - F_m(j)) = \overline{Y} \implies \sum_{j=0}^{k-1}(1 - F_m(j)) = \overline{Y} - \sum_{j=k}^{\infty}(1 - F_m(j)) \tag{4.1}$$

$$\sum_{j=0}^{\infty}\left(1 - F\left(j;\hat{\lambda}\right)\right) = \overline{Y} \implies \sum_{j=0}^{k-1}\left(1 - F\left(j;\hat{\lambda}\right)\right) = \overline{Y} - \sum_{j=k}^{\infty}\left(1 - F\left(j;\hat{\lambda}\right)\right) \tag{4.2}$$

$$(4.1) - (4.2) \implies \sum_{j=0}^{k-1}\left(F\left(j;\hat{\lambda}\right) - F_m(j)\right) = \sum_{j=k}^{\infty}\left(F_m(j) - F\left(j;\hat{\lambda}\right)\right) \tag{4.3}$$

In vergelijking 4.3 zijn F en F_m weliswaar omgedraaid ten opzichte van de toetsingsgrootheid maar omdat je de absolute waarde neemt, maakt de volgorde niet uit. $|F - F_m|$ is namelijk gelijk aan $|F_m - F|$.

$F\left(j,\hat{\lambda}\right)$ is de verdelingsfunctie van een Poissonverdeling met $\hat{\lambda}$ een schatter voor λ, daarom geldt $F\left(j,\hat{\lambda}\right) = e^{-\hat{\lambda}}\sum_{l=0}^{j}\hat{\lambda}^l/l!$. Klar, 1999 schrijft in zijn artikel dat $\hat{\lambda}$ een geschikte schatter moet zijn voor λ maar schrijft niks over welke schatters wel en niet geschikt zijn. Veel gebruikte schatters zijn de momentschatter (method of momements) en de maximum-likelihood-schatter. Om te bepalen welke schatter we in dit geval het beste kunnen gebruiken, rekenen we eerst beide schatters uit. Daarna vergelijken we de schatters aan de hand van het gemiddelde van de kwadratische afwijkingen (mean square error), deze wil je zo klein mogelijk hebben.

4.7.2. Schatter voor lambda

Bijma e.a., 2017 leggen in hun boek uit hoe je de momentschatter kan vinden. Zij $X_1, \ldots, X_n$ een steekproef van een verdeling met onbekende parameter θ. De momentschatter voor θ is de waarde $\hat{\theta}$ waar het j-de moment overeenkomt met het j-de steekproefmoment: $\mathbb{E}_{\hat{\theta}}(X^j) = \overline{X^j}$.

Als $X_1, \ldots, X_n$ gelijk verdeelde onafhankelijke stochastische variabelen van een steekproef zijn, dan is het j-de steekproefmoment $\overline{X^j} = n^{-1} \sum_{i=1}^{n} X_i^j$. Bij het berekenen van de momentschatter gaat de voorkeur uit naar de kleinste mogelijke j. Voor 1-dimensionale parameters is $j = 1$ vaak al genoeg.

Onder de aanname dat de regendata $Y_1, \ldots, Y_m$ Poisson-verdeeld zijn met parameter λ kunnen we de momentschatter $\hat{\lambda}$ als volgt vinden. We nemen $j = 1$ waardoor we $\mathbb{E}_{\hat{\lambda}}(Y) = \overline{Y}$ moeten oplossen. Omdat voor een Poisson-verdeelde stochast de verwachting gelijk is aan de parameter wordt deze vergelijking $\hat{\lambda} = \overline{Y}$. De schatter voor λ is dus gelijk aan het gemiddelde van de steekproef.

In Bijma e.a., 2017 wordt ook uitgelegd hoe je de maximum-likelihood-schatter kan vinden. Zij X een stochastische vector met kansdichtheid p_θ die afhankelijk is van een parameter $\theta \in \Theta$. Voor vaste x is de functie $\theta \mapsto L(\theta; x) \coloneqq p_\theta(x)$ de likelihood functie.

De maximum-likelihood-schatter voor θ is de waarde van $T(x) \in \Theta$ waarvoor de likelihood functie maximaal is.

Een andere manier om de maximum-likelihood-schatter te vinden is door de log-likelihood functie $\theta \mapsto \log L(\theta; x) = \log p_\theta(x)$ te maximaliseren. Als de kansdichtheid p_θ een product is van (marginale) dichtheden kan de log-likelihood de berekeningen een stuk eenvoudiger maken.

Laat $(y_1, \ldots, y_m)$ waarnemingen zijn van de regendata $Y = (Y_1, \ldots, Y_m)$ en neem aan dat $Y_1, \ldots, Y_m$ onafhankelijk en Poisson-verdeeld zijn met parameter λ. De kansdichtheid is dan $p_\lambda(y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$. De maximum-likelihood-schatter voor λ kunnen we als volgt vinden. De likelihood functie is $\theta \mapsto L(\lambda; y_1, \ldots, y_m) = \prod_{i=1}^{m} p_\lambda(y_i)$ omdat dit een product is van marginale dichtheden werken we met de loglikelihood.

$$
\begin{aligned}
\lambda \mapsto \log\left(L(\lambda; y_1, \ldots, y_m)\right) &= \log \prod_{i=1}^{m} p_\lambda(y_i) \\
&= \log \prod_{i=1}^{m} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
&= \sum_{i=1}^{m} \log \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\
&= \sum_{i=1}^{m} \left[\log\left(e^{-\lambda}\right) + \log\left(\lambda^{y_i}\right) - \log\left(y_i!\right)\right] \\
&= \sum_{i=1}^{m} \left[-\lambda + y_i \log \lambda - \log\left(y_i!\right)\right] \\
&= -m\lambda + \log \lambda \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} \log\left(y_i!\right)
\end{aligned}
$$

De log-likelihood willen we nu maximaliseren. Dit doen we door de afgeleide naar λ gelijk te stellen aan nul.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log\left(L\left(\lambda; y_1, \ldots, y_m\right)\right) = 0$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[-m\lambda + \log\lambda \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} \log\left(y_i!\right)\right] = 0$$

$$-m + \frac{1}{\lambda}\sum_{i=1}^{m} y_i = 0$$

$$\hat{\lambda} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

$$\hat{\lambda} = \overline{Y}$$

De log-likelihood heeft dus een extreme waarde voor $\lambda = \overline{Y}$. Om te bepalen of dit een minimum of een maximum is berekenen we de tweede afgeleide:

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log\left(L\left(\lambda; y_1, \ldots, y_m\right)\right) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[-m + \frac{1}{\lambda}\sum_{i=1}^{m} y_i\right] = -\frac{1}{\lambda^2}\sum_{i=1}^{m} y_i$$

De tweede afgeleide is altijd kleiner dan 0 dus de log-likelihood heeft een maximum voor $\lambda = \overline{Y}$ en daarom is de maximum-likelihood-schatter voor λ gelijk aan $\overline{Y}$.

Zowel de momentenschatter als de maximum-likelihood-schatter voor λ zijn $\overline{Y}$, het gemiddelde van de steekproef. Beide zullen dus dezelfde gemiddelde kwadratische afwijking hebben, namelijk:

$$\begin{aligned}
\mathrm{MSE}(\lambda; \overline{Y}) &= \mathrm{E}_\lambda \left|\left|\overline{Y} - \lambda\right|\right|^2 \\
&= \mathrm{var}_\lambda\left(\overline{Y}\right) + \left(\mathrm{E}_\lambda \overline{Y} - \lambda\right)^2 \\
&= \mathrm{var}_\lambda\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) + \left(\mathrm{E}_\lambda\left[\frac{1}{m}\sum_{i=1}^{m} y_i\right] - \lambda\right)^2 \\
&= \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{var}_\lambda\left(y_i\right) + \left(\frac{1}{m}\sum_{i=1}^{m} \mathrm{E}_\lambda\left[y_i\right] - \lambda\right)^2 \\
&= \frac{1}{m^2} \cdot m \cdot \mathrm{var}_\lambda\left(y_1\right) + \left(\frac{1}{m} \cdot m \cdot \mathrm{E}_\lambda\left[x_m\right] - \lambda\right)^2 \\
&= \frac{1}{m} \cdot \lambda + \left(\lambda - \lambda\right)^2 \\
&= \frac{1}{m} \cdot \lambda
\end{aligned}$$

4.7.3. Verwerpen van de nulhypothese

We toetsen de nulhypothese H_0 : het aantal regendruppels per tijdsinterval volgt een Poissonverdeling met constante intensiteit, tegen het alternatief H_1 : het aantal regendruppels per tijdsinterval volgt geen Poissonverdeling met constante intensiteit.

We verwerpen H_0 voor grote waarden van de toetsingsgrootheid. Wat groot is, hangt van de situatie af en moeten we dus bepalen. Met behulp van Monte Carlo simulaties kunnen we de kritieke waardes en de p-waardes bepalen waarvoor we H_0 verwerpen.

Als de waarde van de toetsingsgrootheid groter is dan de kritieke waarde, dan verwerp je dat de data van een Poissonverdeling met parameter λ komt. Voor welke p-waarde de nulhypothese wordt verworpen, hangt af van het significantieniveau. Hier nemen we een significantieniveau van 5% en wordt de nulhypothese verworpen voor een p-waarde kleiner dan 0,05. Een significantieniveau van 5% betekent dat naar verwachting in 5% van de gevallen de nulhypothese onterecht wordt verworpen (Bijma e.a., 2017).

We toetsen een dataset op een Poissonverdeling. Om de kritieke waardes en p-waardes voor deze dataset te bepalen met een Monte Carlo simulatie, bepaal je als eerste het aantal runs van je simulatie. Dit getal mag je zelf kiezen maar hoe meer runs, hoe dichter bij de echte kritieke waarde en dus hoe kleiner de fout. Daarna trek je per run net zoveel punten uit een Poissonverdeling als de dataset punten heeft. Stel het aantal runs is 500 en de dataset heeft 100 punten, dan trek je 500 keer 100 punten uit een Poissonverdeling met parameter λ. De waarde van de parameter van de Poissonverdeling kan je zelf kiezen of bepalen aan de hand van de data. We beginnen met een simulatie waarbij de waardes van de parameter zelf zijn gekozen, daarna voeren we ook nog een simulatie uit waarbij de waardes van de parameter geschat zijn aan de hand van de data. Dit laatste heet een parametrische bootstrap.

4.7.4. Simulatie met zelfgekozen parameters

In tabel 4.1 staan voor vijf verschillende gekozen waardes van λ de kritieke waardes van tien simulaties van elk 1000 runs en 2500 trekkingen uit een Poissonverdeling met parameter $\lambda = 0,1;\ 0,2;\ 0,5;\ 1;\ 2$.

Te zien is dat met 1000 runs de kritieke waardes voor $\lambda = 0,1;\ 0,2;\ 0,5;\ 1$ in het tweede decimaal verschillen en voor $\lambda = 2$ zelf in het eerste decimaal. Het is daarom beter om meer runs uit te voeren zodat de kritieke waarde uit de simulatie dichter bij de echte kritieke waarde zit.

Tabel 4.1: Kritieke waardes van 10 simulaties van elk 1000 runs en 2500 trekkingen uit een Poissonverdeling voor verschillende waardes van de parameter λ met een significantie niveau van 5%.

$\lambda \rightarrow$	0,1	0,2	0,5	1	2
Kritieke waardes	0.125	0.245	0.443	0.663	0.933
	0.128	0.225	0.467	0.682	0.911
	0.131	0.229	0.444	0.668	0.931
	0.127	0.232	0.496	0.657	0.914
	0.126	0.254	0.457	0.646	0.934
	0.122	0.224	0.446	0.669	0.934
	0.125	0.232	0.440	0.665	0.878
	0.127	0.250	0.464	0.654	0.908
	0.124	0.247	0.451	0.653	0.931
	0.131	0.223	0.477	0.655	0.957

Bij 10.000 runs was er voor $\lambda = 0,5;\ 1;\ 2$ in 12 van de 30 simulaties een verschil in het tweede decimaal, voor $\lambda = 0,1;\ 0,2$ waren bij alle 10 de simulaties de eerste twee decimalen hetzelfde.

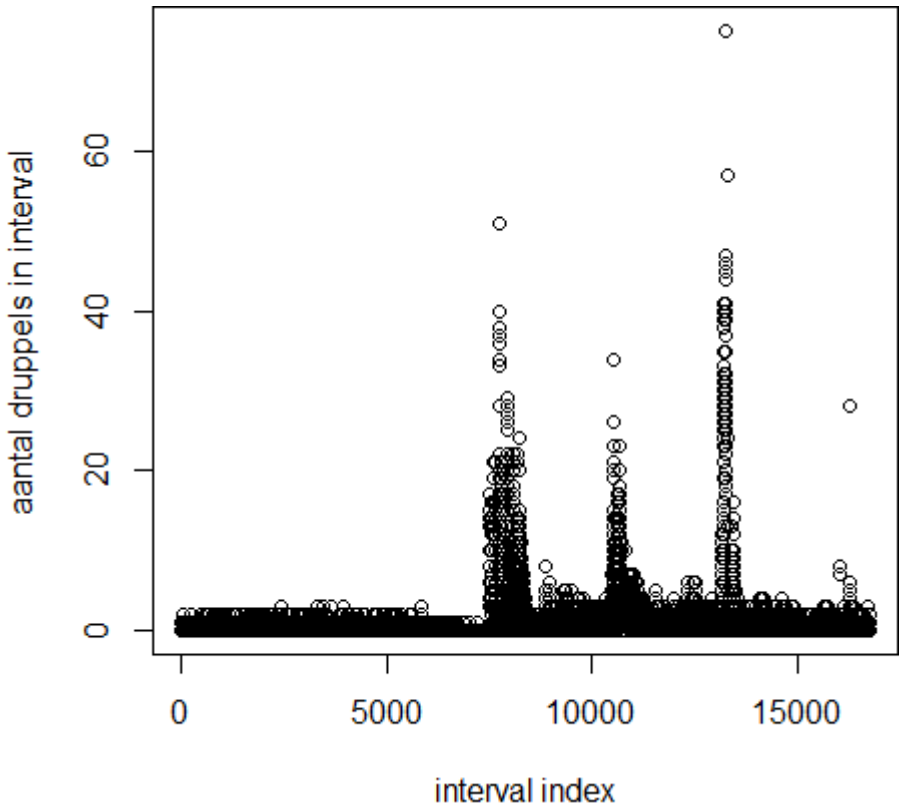
Bij 100.000 runs was er voor $\lambda = 0,5;\ 1;\ 2$ in 3 van de 30 simulaties een verschil in het

tweede decimaal, voor $\lambda = 0,1;\ 0,2$ waren bij alle 10 de simulaties de eerste twee decimalen hetzelfde. Omdat de computer bij 100.000 runs er soms al 25 à 30 seconden over doet om de kritieke waardes uit te rekenen en er maar in 3 van de 50 simulaties een verschil is in het tweede decimaal, kiezen we er voor om verder te werken met 100.000 runs per simulatie.

4.7.5. Simulatie met geschatte parameters

Nu het aantal runs per simulatie is gekozen, gaan we simulaties aan de hand van de data uitvoeren. De nulhypothese stelt dat de data een homogene Poissonverdeling met intensiteit $\lambda > 0$ volgt. De waarde van λ is nog niet bekend en kan alles zijn zolang het maar groter dan 0 is. De nulhypothese is dus samengesteld en daarom gaan we een parametrische bootstrap uitvoeren. Bij een parametrische bootstrap gebruik je in plaats van een zelfgekozen parameter een parameter die geschat wordt aan de hand van de oorspronkelijke data (Pennsylvania State University, 2018). Voordat we simulaties gaan uitvoeren, kijken we eerst wat beter naar de data die we hebben.

Figuur 4.1 is een plot van $Y_1, \ldots, Y_m$, het aantal druppels in de intervallen $1, \ldots, m$. Op de x-as staat de index van het interval en op de y-as staat het aantal druppels in dat interval, bij bijvoorbeeld x-waarde interval-index 1 hoort als y-waarde het aantal druppels in het interval $[0, 10)$.



Figuur 4.1: Het aantal druppels per interval $Y_1, \ldots, Y_m$. Op de x-as staat de index van het interval, op de y-as het aantal druppels in het interval.

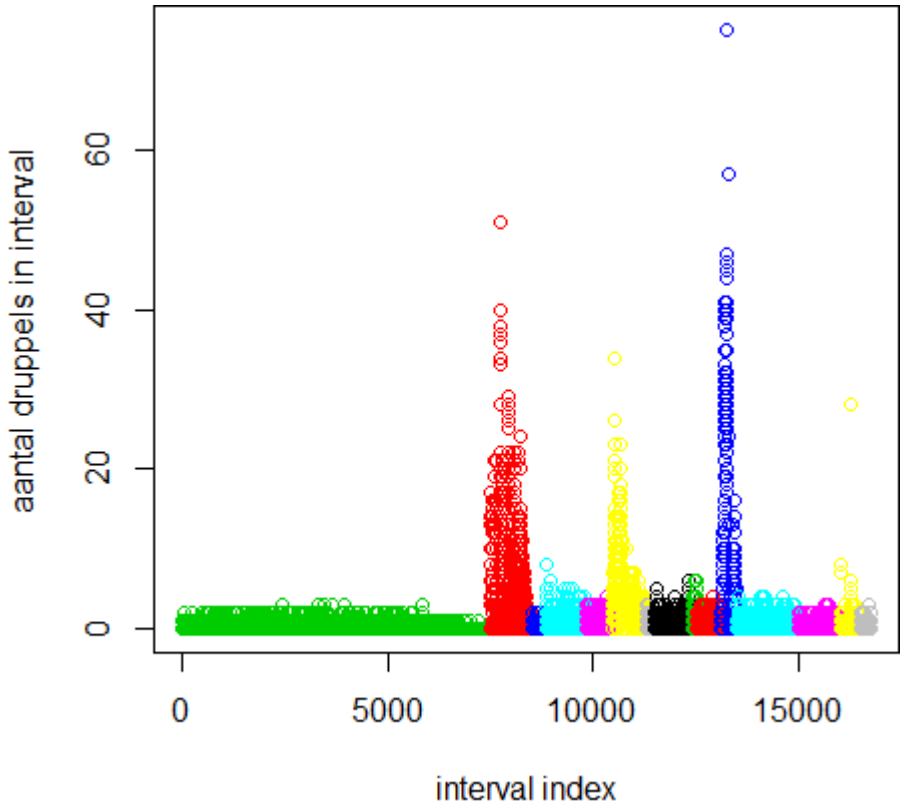
Kijken we naar het aantal druppels per interval in figuur 4.1 dan lijkt de intensiteit niet constant gedurende de gehele observatieperiode. Er zitten drie hoge pieken in het aantal druppels en ook drie wat lagere pieken.

Dit blijkt ook als we een parametrische bootstrapsimulatie uitvoeren. Daar komt een p-waarde uit van 0. Dat betekent dat we voor ieder significantieniveau verwerpen dat het aantal

druppels per interval homogeen Poisson-verdeeld is.

In het artikel van de Villiers e.a., 2021 wordt geconstateerd dat vooral regen met een lage intensiteit een Poissonverdeling lijkt te volgen en regen met een hoge intensiteit niet. Het zou daarom kunnen dat er deelperiodes zijn waarvoor het aantal druppels per interval wel homogeen Poisson-verdeeld kan zijn.

Om dit te onderzoeken delen we de data $Y_1, \ldots, Y_m$, op het oog, op in deelperiodes zodat in die deelperiodes de intensiteit redelijk constant is. We kunnen bijvoorbeeld $Y_1, \ldots, Y_{7500}$ zien als één deelperiode. De observatieperiode is opgedeeld in vijftien deelperiodes. Iedere deelperiode is in figuur 4.2 aangegeven met een eigen kleur.



Figuur 4.2: De vijftien deelperiodes waarin de data op het oog is verdeeld.

Voor iedere deelperiode voeren we een parameterische bootstrapsimulatie uit.

Als eerste schatten we de intensiteit van de deelperiodes met het steekproefgemiddelde.

Iedere simulatie bestaat uit 100.000 runs, in iedere run trekken we net zoveel punten uit een Poissonverdeling met parameter gelijk aan het steekproefgemiddelde als de deelperiode intervallen heeft. Voor iedere trekking berekenen we de waarde van de toetsingsgrootheid en zo kunnen we de kritieke waarde en de p-waarde berekenen.

De kritieke waarde met significantieniveau 5% is het 0,95-kwantiel, H_0 wordt verworpen als de waarde van de toetsingsgrootheid van de dataset groter is dan het 0,95-kwantiel van de gerealiseerde bootstrapwaarde.

De p-waarde wordt berekend door de fractie te nemen van de waardes van de toetsingsgrootheid van de trekking die groter zijn dan de waarde van de toetsingsgrootheid van de deelperiode.

We berekenen de p-waarde omdat die meer informatie geeft over de significantie dan de kritieke waarde. Is de p-waarde bijvoorbeeld 0 dan zou H_0 bij ieder significantieniveau ver-

worpen worden, maar is de p-waarde bijvoorbeeld 0,03 dan wordt H_0 wel verworpen bij een significantieniveau van 5% maar niet bij een significantieniveau 1%. Dit zou je niet kunnen concluderen als alleen de kritieke waarde bekend is.

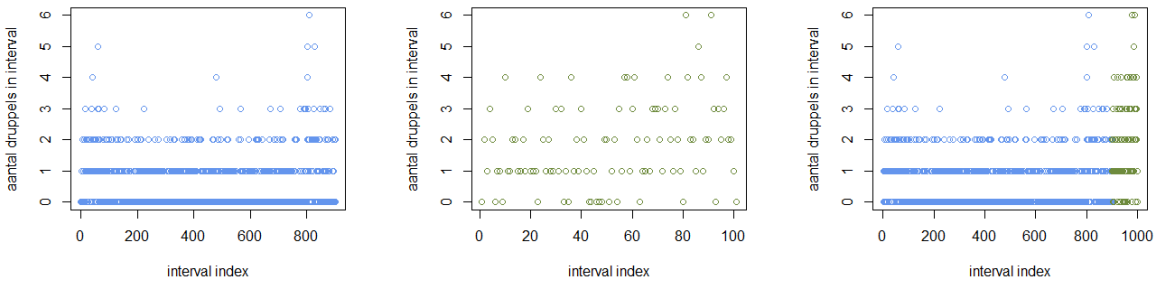
De resultaten van de parametrische bootstrapsimulaties staan in tabel 4.2. Voor iedere deelperiode bevat de tabel de geschatte parameter $\hat{\lambda}$ onder de aanname dat de data een Poissonverdeling volgt, de kritieke waarde, de p-waarde, de waarde van de toetsingsgrootheid van de deelperiode en of de nulhypothese, dat het aantal regendruppels per tijdsinterval een Poissonverdeling volgt, wel of niet wordt verworpen. H_0 wordt verworpen als de waarde van de toetsingsgrootheid groter is dan de kritieke waarde en/of als de p-waarde kleiner is dan het significantieniveau van 0,05.

Tabel 4.2: Voor verschillende deelperiodes de waarde van $\hat{\lambda}$, de kritieke waarde, de p-waarde, de waarde van de toetsingsgrootheid en of H_0 wordt verworpen. De kritieke waardes en p-waardes zijn berekend met een parametrische bootstrap, een Monte Carlo simulatie van 100.000 runs waarbij wordt getrokken uit een Poissonverdeling met parameter gelijk aan $\hat{\lambda}$. $\hat{\lambda}$ is het steekproefgemiddelde van het aantal druppels per interval voor de intervallen in een deelperiode. We verwerpen H_0 als de waarde van de toetsingsgrootheid groter is dan de kritieke waarde of als de p-waarde kleiner is dan het significantieniveau van 0,05.

interval-index	$\hat{\lambda}$	kritieke waarde	p-waarde	toetsingsgrootheid	verwerpen ja/nee
0:7500	0.1825	0.2159	0.0698	0.2000	nee
7500:8500	4.0819	1.3072	0	43.592	ja
8500:8850	0.2963	0.3207	0.1904	0.2139	nee
8850:9850	0.7702	0.5816	0	1.3073	ja
9850:10.450	0.6639	0.5373	0.5991	0.1868	nee
10.450:11.300	3.1974	1.1572	0	21.1696	ja
11.300:11.500	0.6318	0.5190	0.0182	0.6245	ja
11.500:12.400	0.6604	0.5364	0.1714	0.3829	nee
12.400:12.500	1.8020	0.8703	0.7821	0.2638	nee
12.500:13.100	0.5790	0.4978	0.1459	0.3743	nee
13.100:13.500	8.4239	1.8806	0	77.9782	ja
13.500:15.000	0.5929	0.5028	0.8167	0.1061	nee
15.000:16.000	0.2657	0.2955	0.8127	0.0503	nee
16.000:16.500	0.4731	0.4436	0	2.1884	ja
16.500:16.734	0.3191	0.3398	0.1513	0.2473	nee

Opvallend is wat er gebeurt bij het aantal druppels $Y_{11.500}, \ldots, Y_{12.500}$. In figuur 4.3 staan van links naar rechts plots van het aantal druppels $Y_{11.500}, \ldots, Y_{12.400}$ (4.3a), $Y_{12.400}, \ldots, Y_{12.500}$ (4.3b) en $Y_{11.500}, \ldots, Y_{12.500}$ (4.3c).

Als je in tabel 4.2 de p-waardes van de deelperiodes $Y_{11.500}, \ldots, Y_{12.400}$ en $Y_{12.400}, \ldots, Y_{12.500}$ apart bekijkt, dan zijn de p-waardes respectievelijk 0,17 en 0,78 en wordt voor beide deelperiodes apart de nulhypothese niet verworpen. Bekijk je echter de p-waarde van beide deelperiodes samen ($Y_{11.500}, \ldots, Y_{12.500}$) dan is de p-waarde 0,0003 en wordt de hypothese wel verworpen. In de drie plotjes in figuur 4.3 is te zien dat het verwerpen van de nulhypothese niet te wijten is aan een hoge piek in de intensiteit. Hetzelfde gebeurt bij $Y_{13.500}, \ldots, Y_{15.000}$ en $Y_{15.000}, \ldots, Y_{16.000}$. Dit zou kunnen duiden op een inhomogeen Poissonproces.



(a) $Y_{11.500}, \ldots, Y_{12.400}$ (b) $Y_{12.400}, \ldots, Y_{12.500}$ (c) $Y_{11.500}, \ldots, Y_{12.500}$

Figuur 4.3: Het aantal druppels per interval voor verschillende deelperiodes. De linker plot is de deelperiode met aantal druppels $Y_{11.500}, \ldots, Y_{12.400}$, de middelste plot is de deelperiode met aantal druppels $Y_{12.400}, \ldots, Y_{12.500}$. Voor deze deelperiodes wordt de nulhypothese niet verworpen. De rechter plot is het aantal druppels van beide deelperiodes samen ($Y_{11.500}, \ldots, Y_{12.500}$), hiervoor wordt de nulhypothese wel verworpen.

Wat erg belangrijk is om te beseffen, is dat we aan data snooping of p-hacking doen door de regendata op deze manier te onderzoeken. We kiezen bewust bepaalde deelperiodes uit om toetsen op uit te voeren om zo een significant resultaat te krijgen (Lechner e.a., 2021). Dit kan tot gevolg hebben dat we een patroon in de data vinden dat statistisch significant is terwijl er eigenlijk geen patroon is. We moeten dus voorzichtig zijn met het trekken van conclusies gebaseerd op de resultaten in tabel 4.2.

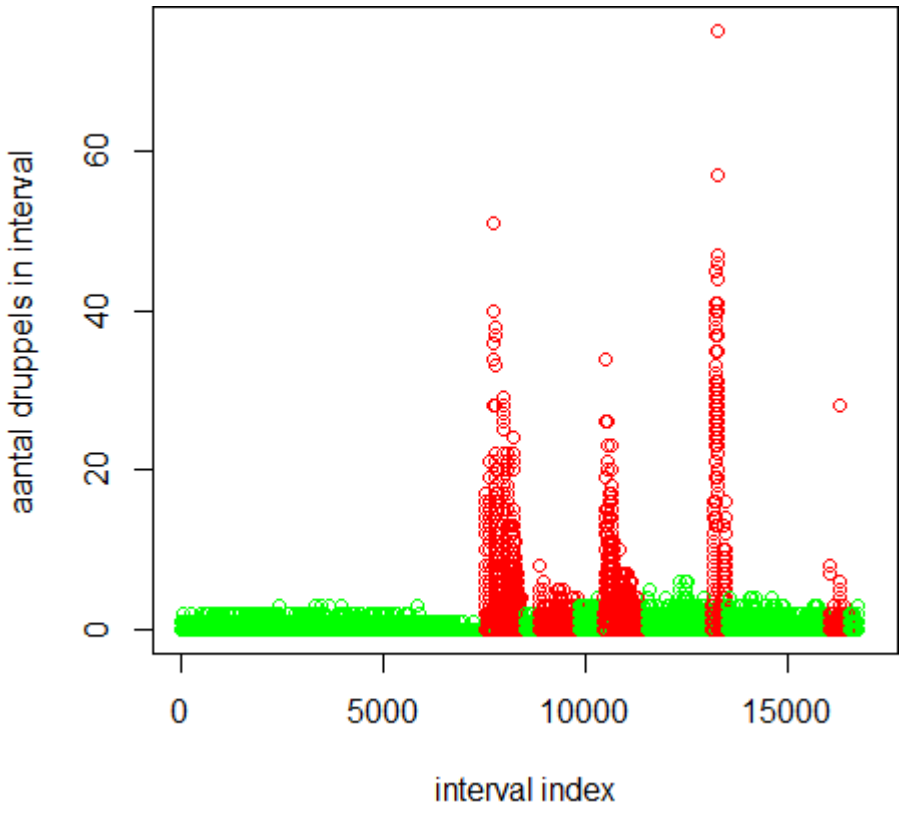
Ook goed om te beseffen, is dat we met schattingen en simulaties te maken hebben. Eerst schatten we de waarde van λ van de deelperiode en daarna voeren we met die geschatte waarde simulaties uit. We zullen dus nooit de exacte kritieke waarde en de exacte p-waarde krijgen. Als je met een grensgeval te maken hebt, kan het zo zijn dat je bij de ene simulatie een p-waarde krijgt van 0,049 en bij de volgende simulatie voor dezelfde dataset een p-waarde van 0,051. Bij een significantie-niveau van 5% zou je bij de eerste simulatie wel verwerpen en bij de tweede niet terwijl het dezelfde dataset is.

Wat we daarom doen, is een interval bepalen waarin de p-waarde zich bevindt. Hiervoor berekenen we voor iedere deelperiode eerste het 95%-betrouwbaarheidsinterval voor de schatting van λ. Met de boven- en ondergrens van die betrouwbaarheidsintervallen voeren we simulaties uit en zo krijgen we voor iedere deelperiode een interval voor de p-waarde. De resultaten hiervan staan in tabel 4.3. Dit zijn dezelfde deelperiodes als in tabel 4.2. In de derde kolom staan de intervallen voor de p-waardes. Te zien is dat er geen deelperiodes zijn die zich bij een significantieniveau van 5% op de grens bevinden van wel of niet verwerpen, maar er zijn wel intervallen met een flink verschil tussen de boven en ondergrens. Zo is bij de deelperiode met interval-indices 8500:8850 het verschil tussen de boven- en ondergrens 0,1316. Dat is dus een verschil van 13%. Bij de laatste deelperiode met interval-indices 16.500:16.734 is dit verschil zelfs ruim 14%.

Ondanks dat de intervallen van de p-waardes bij deze indeling van deelperiodes geen verschil zouden maken tussen wel of niet verwerpen van de nulhypothese, kan er dus wel een groot verschil zijn tussen de boven- en ondergrens van de p-waardes wat misschien bij een andere indeling in deelperiodes zou kunnen leiden tot een andere conclusie of er wel of niet verworpen wordt.

Tabel 4.3: Voor verschillende deelperiodes het 95%-betrouwbaarheidsinterval voor $\hat{\lambda}$ en een interval voor de p-waarde. De p-waardes zijn berekend met een parametrische bootstrap, een Monte Carlo simulatie van 100.000 runs waarbij wordt getrokken uit een Poissonverdeling met parameter gelijk aan de boven- en ondergrens van het betrouwbaarheidsinterval van $\hat{\lambda}$.

interval-index	95%-betrouwbaarheidsinterval voor $\hat{\lambda}$	interval voor p-waarde
0:7500	[0.1729, 0.1922]	[0.0586, 0.0839]
7500:8500	[3.9568, 4.2071]	[0, 0]
8500:8850	[0.2393, 0.3532]	[0.1223, 0.2539]
8850:9850	[0.7159, 0.8246]	[0, 0]
9850:10.450	[0.5988, 0.7290]	[0.5576, 0.6390]
10.450:11.300	[3.0773, 3.3176]	[0, 0]
11.300:11.500	[0.5219, 0.7417]	[0.0088, 0.0288]
11.500:12.400	[0.6073, 0.7134]	[0.1495, 0.1952]
12.400:12.500	[1.5402, 2.0638]	[0.7306, 0.8198]
12.500:13.100	[0.5182, 0.6399]	[0.1198, 0.1753]
13.100:13.500	[8.1399, 8.7080]	[0, 0]
13.500:15.000	[0.5540, 0.6319]	[0.7968, 0.8358]
15.000:16.000	[0.2338, 0.2977]	[0.7693, 0.8452]
16.000:16.500	[0.4128, 0.5333]	[0, 0]
16.500:16.734	[0.2469, 0.3914]	[0.0792, 0.2215]



Figuur 4.4: Het aantal druppels per interval. Op de x-as staat de index van het interval, op de y-as het aantal druppels in het interval. Voor de groene punten wordt de nulhypothese, de data is homogeen Poisson-verdeeld, niet verworpen; voor de rode punten wel.

Om een beter beeld te krijgen voor welke deelperiodes de nulhypothese wel en niet verworpen wordt, staat in figuur 4.4 een plot van het aantal druppels per interval waarbij de intervallen

groen en rood gekleurd zijn. Voor het aantal druppels in de groene intervallen wordt de nulhypothese niet verworpen op basis van de simulatie in tabel 4.2, voor het aantal druppels in de rode intervallen wel. Uit de figuur komt duidelijk naar voren dat regen met lage intensiteit homogeen Poisson kan zijn en dat de hoge pieken niet homogeen Poisson zijn.

4.7.6. Conclusie wel of niet homogeen Poisson

Uit bovenstaande simulaties kunnen we concluderen dat alle regendata over de observatieperiode niet met een homogeen Poissonproces kan worden beschreven, de nulhypothese wordt hiervoor verworpen. Wel zijn er deelperiodes van de observatieperiode waarvoor de nulhypothese niet wordt verworpen en waarvoor het regenvalproces dus wel met een homogeen Poissonproces kan worden beschreven. Er zijn ook twee deelperiodes naast elkaar waarvoor de nulhypothese niet wordt verworpen als je de periodes apart bekijkt, maar waarvoor de nulhypothese wel wordt verworpen als je de periodes samen bekijkt. Dit zou er op kunnen wijzen dat de regendata met een inhomogeen Poissonproces kan worden beschreven, al moeten we vanwege data snooping voorzichtig zijn met het trekken van conclusies. In het volgende hoofdstuk wordt op een objectieve manier de data in deelperiodes verdeeld met Taut String en wordt onderzocht of de regen met een inhomogeen Poissonproces kan worden beschreven.

In paragraaf 4.6 staat uitgelegd dat het nuttig zou kunnen zijn om zowel te toetsen op een Poissonverdeling van het aantal druppels per interval als op een exponentiële verdeling van de tussenaankomsttijden. Dat was ook een reden om voor de toetsingsgrootheid I_n te kiezen. In paragraaf 4.7 hebben we veel aandacht besteed aan het toetsen op een Poissonverdeling van het aantal druppels per interval en helaas hebben we geen tijd meer gehad om ook te toetsen op een exponentiële verdeling van de tussenaankomsttijden.

5

Toetsen voor een inhomogeen Poissonproces

In het vorige hoofdstuk hebben we aan de hand van Monte Carlo simulaties kunnen concluderen dat de regen in de gehele observatieperiode niet met een homogeen Poissonproces kan worden beschreven. Wel zijn er deelperiodes waarop de regendata wel door een homogeen Poissonproces kan worden beschreven, de nulhypothese wordt hier niet verworpen. Dit zou er op kunnen wijzen dat de regendata inhomogeen Poisson-verdeeld is. Het indelen van de regendata in verschillende deelperiodes hebben we echter op het oog gedaan. Vanwege datasnooping moeten we dan ook voorzichtig moeten zijn met het trekken van conclusies op basis van deze indeling.

Dit hoofdstuk gaan we de indeling van de regendata in verschillende deelperiodes objectiever doen met behulp van taut string. In paragraaf 5.1 wordt uitgelegd wat taut string is en hoe het werkt. Omdat er bij taut string gewerkt wordt met een zelfgekozen penalty parameter, wordt in paragraaf 5.2 uitgelegd hoe we de optimale penalty parameter hebben gekozen. In paragraaf 5.3 wordt deze optimale penalty parameter gebruikt om de regendata met taut string op te delen in deelperiodes met een constante intensiteit. Omdat we werken met het aantal druppels per interval van 10 seconden, zal deze intensiteit een schatting zijn voor het aantal druppels per interval van 10 seconden. Deze indeling in deelperiodes met constante intensiteit wordt vervolgens in paragraaf 5.4 gebruikt om de data te toetsen op een inhomogeen Poissonproces. Belangrijk is om te realiseren dat we toetsen op een inhomogeen Poissonproces waarvan de intensiteit een stuksgewijs constante functie is.

5.1. Taut string methode

De taut string methode wordt gebruikt om data te benaderen met een stuksgewijs constante functie en werkt als volgt. De beschikbare data is een vector $\vec{y} = (y_1, y_2, \dots, y_m)$ van lengte m. Als eerste wordt de data geïntegreerd door de cumulatieve som te berekenen, dit geeft de vector $\vec{Y} = (Y_1, Y_2, \dots, Y_m)$. Daarna wordt er een bovengrens $Y.upp = (Y_1, Y_2 + \beta, Y_3 + \beta, \dots, Y_{m-2} + \beta, Y_{m-1} + \beta, Y_m)$ en een ondergrens $Y.low = (Y_1, Y_2 - \beta, Y_3 - \beta, \dots, Y_{m-2} - \beta, Y_{m-1} - \beta, Y_m)$ bepaald waar de taut string tussen moet blijven. Neem nu in gedachte een stuk touw vanaf het punt Y_1 tot het punt Y_m en trek het touw strak tussen de boven- en ondergrens. Dit geeft een stuksgewijs lineaire functie door de punten $\vec{X} = (X_1, X_2, \dots, X_m)$. De afgeleide van $\vec{X}$ geeft de unieke vector $\vec{x} = (x_1, x_2, \dots, x_m)$ die de totale variatie minimaliseert. Dit betekent dat het touw zo kort mogelijk is en wiskundig komt dit overeen met het minimaliseren van de volgende

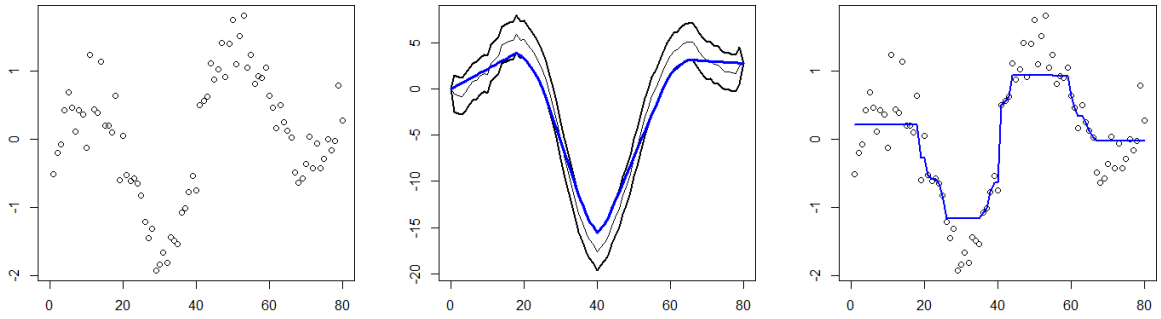
uitdrukking (Dümbgen en Kovac, 2009):

$$\frac{1}{2}\sum_{i=1}^{m}(y_i - x_i)^2 + \beta\sum_{i=2}^{m}|x_i - x_{i-1}| \tag{5.1}$$

De vector $\vec{x}$ is de benadering van de data. Deze benadering is stuksgewijs constant omdat het de afgeleide is van een stuksgewijs lineaire functie.

De eerste fidelity term in uitdrukking 5.1 is de l_2-afstand tussen de data $\vec{y}$ en de schatting $\vec{x}$, de tweede penalty term is een parameter β keer de totale variatie van de schatting $\vec{x}$. β wordt ook wel de penalty parameter genoemd.

In figuur 5.1 staan drie plots van een voorbeelddataset y_{vb} waarop taut string is uitgevoerd met een penalty parameter van 2. In figuur 5.1a staat een scatterplot van de voorbeelddata. In figuur 5.1b staat de geïntegreerde voorbeelddata Y_{vb} (dunne zwarte lijn), de boven- en ondergrens $Y.upp_{vb}$ en $Y.low_{vb}$ (dikke zwarte lijnen) en de taut string X_{vb} (blauwe lijn). Te zien is dat de taut string start in het eerste punt $Y_{vb,1}$, eindigt in het laatste punt $Y_{vb,m}$ en daartussen strak getrokken is tussen de boven- en ondergrens. In figuur 5.1c staat nogmaals een scatterplot van de voorbeelddata y_{vb} met dit keer de schatting x_{vb} van de data in het blauw erbij. Te zien is dat deze schatting inderdaad stuksgewijs constant is.



(a) Voorbeelddata y_{vb}

(b) Taut string X_{vb} met een penalty parameter van 2

(c) Voorbeelddata en de geschatte intensiteit x_{vb} voor een penalty parameter van 2

Figuur 5.1: Taut string met een penalty parameter van 2 uitgevoerd op voorbeelddata. Links staat een scatterplot van de voorbeelddata; in het midden staat de geïntegreerde data (dun zwart), de boven- en ondergrens (dik zwart) en taut string (blauw); rechts staat een scatterplot van de voorbeelddata (zwart) met de stuksgewijs constante benadering in het blauw.

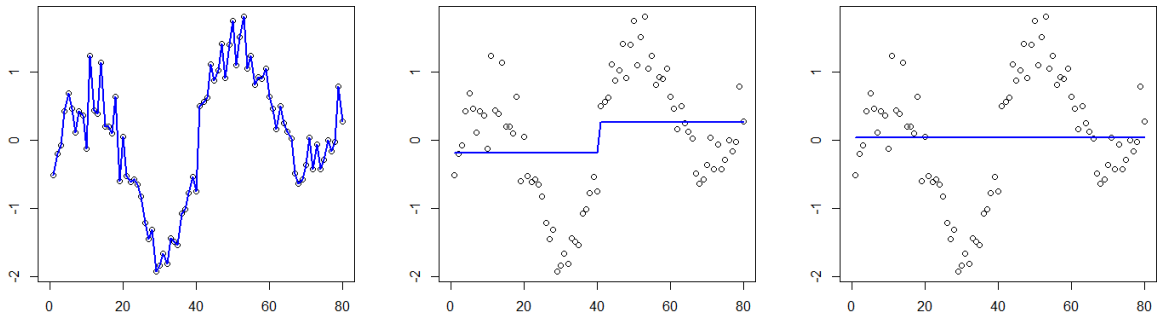
Hoe nauwkeurig de schatting x is, hangt af van de penalty parameter β. De penalty parameter kan zelf gekozen worden en bepaalt de afstand tussen de boven- en ondergrens, deze afstand is namelijk twee keer de penalty parameter. Bij een penalty parameter van 0 volgt de schatting x precies de data, bij een grote penalty parameter wordt de benadering grof en op een gegeven moment zelfs constant. Dit is te zien in figuur 5.2.

Op de voorbeelddata is taut string uitgevoerd met de penalty parameters 0, 10 en 20. In de bovenste rij van figuur 5.2 staan scatterplots van de voorbeelddata y_{vb} in het zwart, met de benadering x_{vb} in het blauw, in de onderste rij staan plots van de geïntegreerde voorbeelddata Y_{vb} (dunne zwarte lijn), de boven- en ondergrens (dikke zwarte lijn) en de taut string X_{vb} (blauwe lijn).

In de linker scatterplot 5.2a is te zien dat bij een penalty parameter van 0 de schatting precies de punten van de voorbeelddata volgt. In de plot van de geïntegreerde data 5.2d is

ook geen verschil te zien tussen de geïntegreerde data, de boven- en ondergrens en de taut string. In de middelste figuren is de penalty parameter 10 en bestaat de schatting X_{vb} uit twee constante stukken (5.2b) en de taut string X_{vb} uit twee lineaire stukken (5.2e). In de rechter figuren 5.2c en 5.2f is de penalty parameter 20 en is de schatting constant.

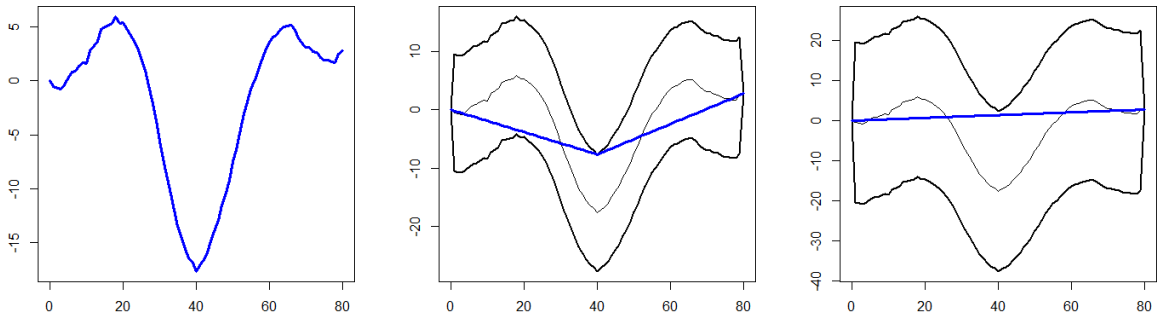
Te zien is dat bij een (te) kleine penalty parameter de schatting heel erg gaat schommelen tussen alle datapunten zoals in de scatterplot 5.2a en dat bij een (te) grote penalty parameter de schatting constant wordt zoals in scatterplot 5.2c. Tussen deze twee extremen wil je een balans zien te vinden. Hoe deze balans gevonden kan worden, wordt uitgelegd in de volgende paragraaf.



(a) Voorbeelddata en de geschatte intensiteit voor een penalty parameter van 0

(b) Voorbeelddata en de geschatte intensiteit voor een penalty parameter van 10

(c) Voorbeelddata en de geschatte intensiteit voor een penalty parameter van 20



(d) Taut string met een penalty parameter van 0

(e) Taut string met een penalty parameter van 10

(f) Taut string met een penalty parameter van 20

Figuur 5.2: Taut string met een penalty parameter van 0, 10 en 20 uitgevoerd op de voorbeelddata. Boven staan scatterplots van de voorbeelddata in het zwart met de benadering in het blauw, onder staan plots van de geïntegreerde data (dun zwart), de boven- en ondergrens (dik zwart) en taut string (blauw). In de linker plots is de penalty parameters 0, in het midden 10 en rechts 20.

5.2. Taut string bepalen penalty parameter

Bij het uitvoeren van taut string moet er een penalty parameter gekozen worden die invloed heeft op de nauwkeurigheid van de schatting van de onderliggende functie. Het is daarom belangrijk om een goede waarde voor de penalty parameter te kiezen.

Om een goede penalty parameter te bepalen, volgen we het volgende stappenplan:

1. Kies een penalty parameter om mee te beginnen. Dit is β_{start}, ook wel 'pilot' genoemd.
2. Schat met taut string met de gekozen penalty parameter β_{start} de intensiteit van de data $\vec{y} = (y_1, y_2, \ldots, y_m)$. Dit geeft de vector $\vec{x}_{\beta_{\text{start}}} = \left(x_1^{(\beta_{\text{start}})}, x_2^{(\beta_{\text{start}})}, \ldots, x_m^{(\beta_{\text{start}})}\right)$, met voor

ieder punt in $\vec{y}$ een schatting van de intensiteit. Neem dit als begin-intensiteit.

3. Genereer nieuwe data $\vec{y}^* = (y_1^*, \ldots, y_m^*)$ door te trekken uit Poissonverdelingen met als parameter steeds één waarde van de begin-intensiteit $\vec{x}_{\beta_{\text{start}}}$. We trekken uit een Poissonverdeling omdat we uiteindelijk de regendata willen toetsen op een Poissonverdeling.

4. Schat de intensiteit van de gegenereerde data $\vec{y}^* = (y_1^*, \ldots, y_m^*)$ met taut string voor verschillende penalty parameters $\beta_1, \beta_2, \ldots, \beta_N$. Dit geeft de vectoren $\vec{x}^*_{\beta_1}, \ldots, \vec{x}^*_{\beta_N}$ elk van lengte m, met voor iedere penalty parameter β_i $(i = 1, \ldots, N)$ een schatting voor de intensiteit van y_j^* met $j = 1, \ldots, m$.

5. Voor iedere penalty parameter β_i berekenen we de l_2-afstand tussen de schatting $\vec{x}^*_{\beta_i}$ en de begin-intensiteit $\vec{x}_{\beta_{\text{start}}}$.
Bij een penalty parameter van 0 zal deze afstand groot zijn, daarna neemt de afstand af voor groeiend penalty parameter maar op een zeker moment zal de afstand weer gaan stijgen als de penalty parameter nog verder toeneemt. De penalty parameter β_i waarvoor deze afstand minimaal is, is de optimale penalty parameter.

6. Herhaal stap 3, 4 en 5 honderd keer. Dit doe je omdat iedere keer dat je een nieuwe dataset $\vec{y}^* = (y_1^*, \ldots, y_m^*)$ genereert, de gegenereerde data net anders is omdat je willekeurige punten uit een Poissonverdeling trekt.

7. Door stap 3, 4 en 5 honderd keer te herhalen, krijgen we honderd optimale penalty parameters. Het gemiddelde van deze honderd optimale penalty parameters gebruiken we uiteindelijk om de intensiteit van de data $\vec{y} = (y_1, \ldots, y_m)$ te schatten met taut string.

Eerst zullen we bovenstaande stappen verduidelijken met een voorbeeld, daarna zullen we de stappen uitvoeren voor de regendata.

Stel we hebben een voorbeelddataset $\vec{y}_{\text{voorbeeld}} = (1, 0, 4, 7, 4, 0, 2, 5, 13, 6)$ en in stap 1 kiezen we $\beta_{\text{start}} = 2$.

In stap 2 schatten we met taut string met een penalty parameter van 2 de intensiteit van $\vec{y}_{\text{voorbeeld}}$. Dit geeft de begin-intensiteit $\vec{x}_2 = (1,5;\ 1,5;\ 3,67;\ 3,67;\ 3,67;\ 3;\ 3;\ 5;\ 9;\ 8)$. Omdat taut string een stuksgewijs constante functie geeft kan een intensiteit meerdere keren achter elkaar voorkomen.

In stap 3 genereren we een nieuwe dataset door met het commando `rpois(n, lambda)` willekeurig te trekken uit Poissonverdelingen met steeds één waarde van $\vec{x}_2$. Zo krijgen we $\vec{y}^*_{\text{voorbeeld}}$ = [`rpois(1; 1,5), rpois(1; 1,5), rpois(1; 3,67), rpois(1; 3,67), rpois(1; 3,67), rpois(1; 3), rpois(1; 3), rpois(1; 5), rpois(1; 9), rpois(1; 8)`]

In stap 4 schatten we met taut string met penalty parameters 0, 1, 2,…, 10 de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$. Zo krijgen we elf vectoren met schattingen voor de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$:
$\vec{x}^*_0$ is de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$ geschat met taut string met een penalty parameter van 0,
$\vec{x}^*_1$ is de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$ geschat met taut string met een penalty parameter van 1,
$\vec{x}^*_2$ is de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$ geschat met taut string met een penalty parameter van 2,
tot en met $\vec{x}^*_{10}$, de intensiteit van $\vec{y}^*_{\text{voorbeeld}}$ geschat met taut string met een penalty parameter van 10.

In stap 5 berekenen we de l_2-afstand tussen $\vec{x}_2$ en $\vec{x}_0^*$, tussen $\vec{x}_2$ en $\vec{x}_1^*$, tussen $\vec{x}_2$ en $\vec{x}_2^*$, tot en met de l_2-afstand tussen $\vec{x}_2$ en $\vec{x}_{10}^*$.
De waarde van β waarvoor deze afstand het kleinst is, is de optimale penalty parameter. In dit voorbeeld is de afstand tussen $\vec{x}_2$ en $\vec{x}_4^*$ het kleinst dus de optimale penalty parameter is hier 4.

Stap 3, 4 en 5 herhalen we honderd keer omdat $\vec{y}^*_{\text{voorbeeld}}$ steeds anders is en we daardoor op een andere optimale penalty parameter kunnen uitkomen.

Het gemiddelde van de honderd optimale penalty parameters gebruiken we om met taut string de intensiteit van de data $\vec{y}_{\text{voorbeeld}}$ te schatten.

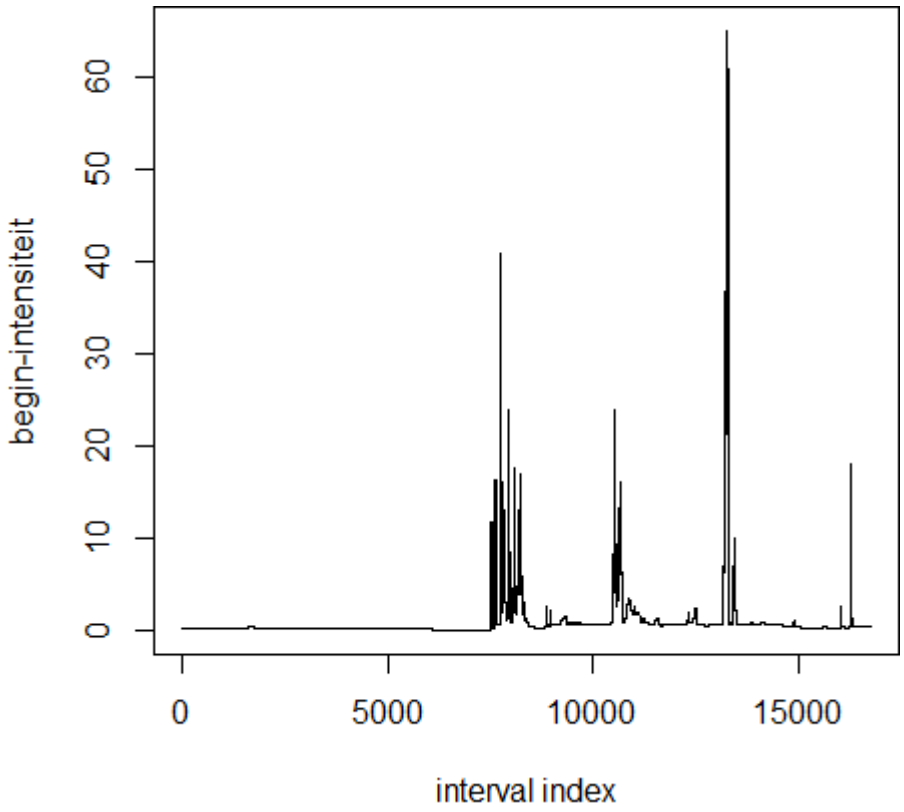
Nu gaan we de optimale penalty parameter bepalen voor de regendata $\vec{y}_{\text{regen}}$. Dit is een vector van lengte $m = 16.734$ en bestaat uit het aantal druppels per interval van 10 seconden.

5.2.1. Stap 1: Kies penalty parameter

Voor de regendata $\vec{y}_{\text{regen}}$ beginnen we met een penalty parameter β_{start} van 5. Dus $\beta_{\text{start}} = 5$

5.2.2. Stap 2: Schat begin-intensiteit

Taut string uitvoeren op $\vec{y}_{\text{regen}}$ geeft een vector $\vec{x}_{\beta_{\text{start}}} = \vec{x}_5$ met voor ieder interval van 10 seconden een schatting voor de intensiteit, dit zijn 16734 intensiteiten. Taut string geeft een stuksgewijs constante functie dus het zullen niet 16734 verschillende intensiteiten zijn, er zullen ook deelperiodes zijn met dezelfde intensiteit. De begin-intensiteit $\vec{x}_5$ voor de regendata $\vec{y}_{\text{regen}}$ is weergegeven in figuur 5.3.

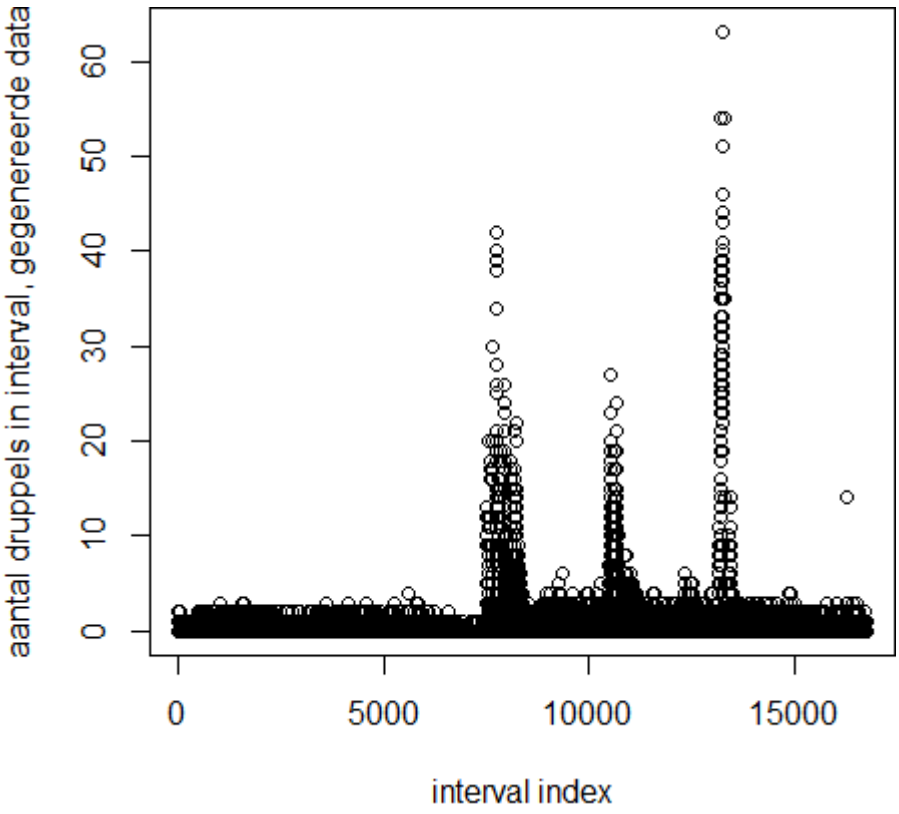


Figuur 5.3: De begin-intensiteit $\vec{x}_5$ van de regendata $\vec{y}_{\text{regen}}$ gevonden met taut string met een gekozen penalty parameter β_{start} van 5.

5.2.3. Stap 3: Trek uit Poissonverdelingen met parameters uit begin-intensiteit

Voor iedere intensiteit in $\vec{x}_5$ trekken we met het commando `rpois(n, lambda)` willekeurig 1 (= `n`) punt uit een Poissonverdeling met `lambda` gelijk aan één van de intensiteiten van $\vec{x}_5$. We beginnen met de eerste intensiteit in $\vec{x}_5$ en dan nemen we steeds de volgende.

De trekkingen zetten we allemaal achter elkaar in een vector $\vec{y}^*_{\text{regen}}$ van lengte m =16.734. De waardes in deze vector zijn het aantal druppels in een interval van 10 seconden van de gegenereerde data. Figuur 5.4 is een plot van de gegenereerde data. Zouden we deze plot nog een keer maken, dan ziet de plot er iets anders uit. Dit komt omdat we voor het maken van $\vec{y}^*_{\text{regen}}$ willekeurige punten trekken uit een Poissonverdeling. Dus iedere keer dat we een punt trekken kan die anders zijn dan de vorige trekking.

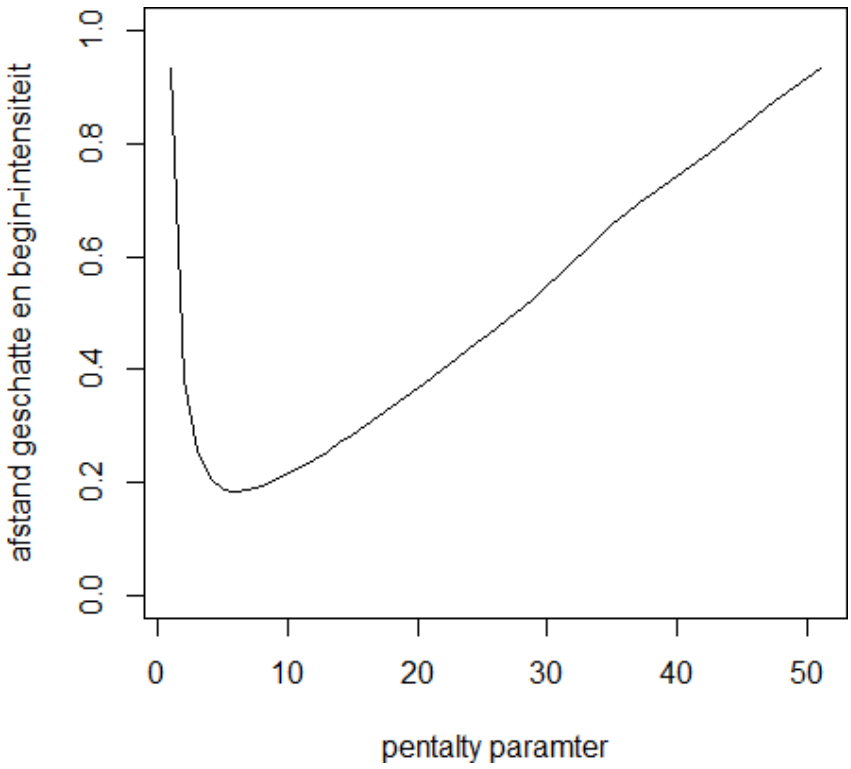


Figuur 5.4: Plot van de gegenereerde dataverzameling $\vec{y}^*_{\text{regen}}$.

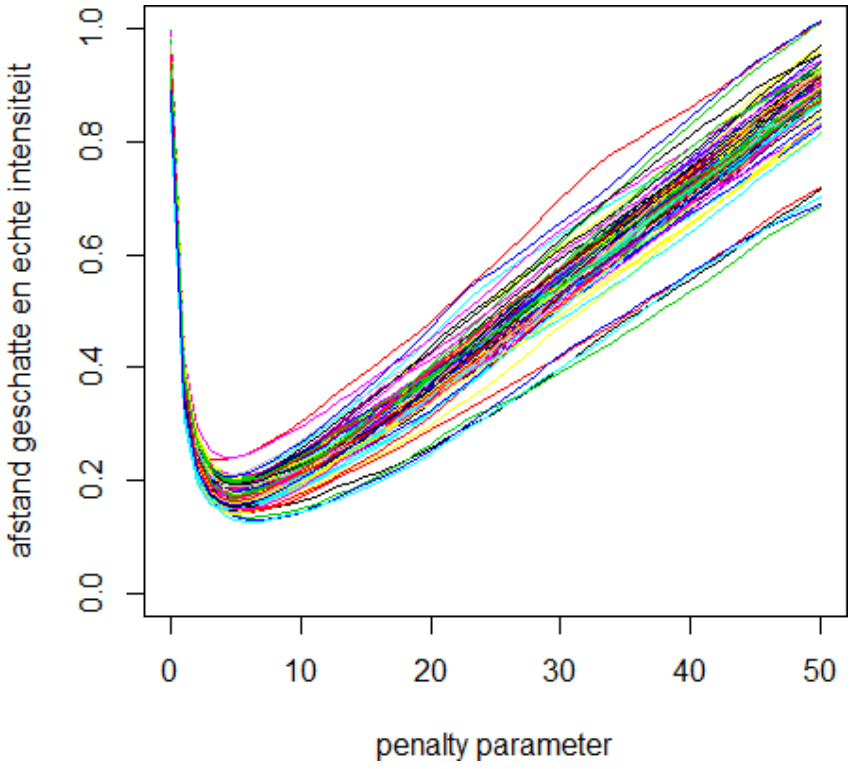
5.2.4. Stap 4: Bepaal penalty parameter zodat afstand begin-intensiteit en geschatte intensiteit minimaal is

Nu kunnen we de gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ gebruiken om te bepalen bij welke penalty parameter β de afstand tussen de begin-intensiteit $\vec{x}_5$ en de intensiteit van $\vec{y}^*_{\text{regen}}$ geschat met taut string $\vec{x}^*_\beta$, het kleinste is. Aanvankelijk zal de totale afstand kleiner worden voor groter wordende penalty parameters maar op een zeker moment wordt de totale afstand weer groter. Dit is te zien in figuur 5.5. Op de x-as staan de penalty parameters 0, 1, 2,…, 50, op de y-as staat de totale afstand tussen de begin-intensiteit $\vec{x}_5$ en de geschatte intensiteit van $\vec{y}^*_{\text{regen}}$, geschat met taut string met de bijbehorende penalty parameter op de x-as. Bij x=0 bijvoorbeeld, is de penalty parameter 0 en is de y-waarde de l_2-afstand tussen $\vec{x}_5$ en $\vec{x}^*_0$.

Omdat $\vec{y}^*_{\text{regen}}$ elke keer anders is, berekenen we de totale afstand voor 100 verschillende gegenereerde dataverzamelingen. In figuur 5.6 is elke lijn de totale afstand tussen de begin-intensiteit $\vec{x}_5$ en de intensiteit van één van de 100 gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$.



Figuur 5.5: De totale afstand tussen de begin-intensiteit $\vec{x}_5$ en de intensiteit van één gegenereerde dataverzameling $\vec{y}^*_{\text{regen}}$ geschat met taut string voor penalty parameters 0, 1, 2,…, 50 (x-as). Bij x=0 bijvoorbeeld, is de penalty parameter 0 en is de y-waarde de l_2-afstand tussen $\vec{x}_5$ en $\vec{x}^*_0$.

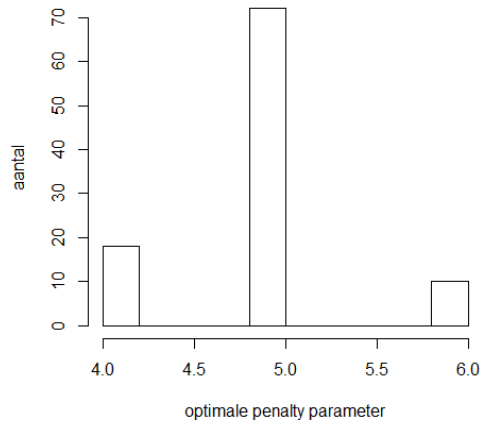


Figuur 5.6: Voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ de totale afstand tussen de begin-intensiteit $\vec{x}_5$ en de intensiteit van één van de gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ geschat met taut string voor penalty parameters 0, 1, 2,…, 50.

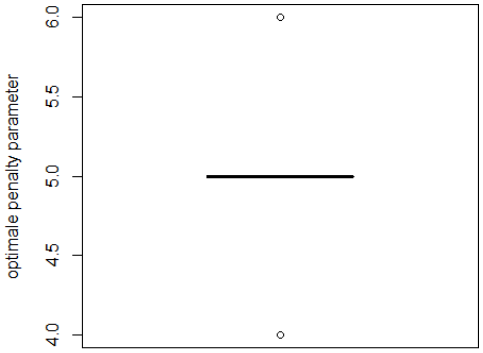
Voor elke gegenereerde dataverzameling $\vec{y}^*_{\text{regen}}$ in figuur 5.6 lijkt de minimale afstand zich rond een penalty parameter van 5 te bevinden Dit zou dus betekenen dan de optimale penalty parameter rond de 5 ligt.

In figuren 5.7b en 5.7a staan een boxplot en een histogram van de optimale penalty parameter van elk van de 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$. Hieruit blijkt inderdaad dat de optimale penalty parameter voor de verschillende individuele gesimuleerde datasets $\vec{y}^*_{\text{regen}}$ zich concentreert rond 5. Het gemiddelde van de optimale penalty parameter van elk van de 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ is 4,92.

Omdat we nu de afstand hebben bepaald voor penalty parameters 0, 1, 2,..., 50, is de optimale penalty parameter altijd een geheel getal. Daarom heeft figuur 5.7 alleen waardes bij de gehele getallen.



(a) Histogram van de penalty parameter waarvoor de afstand minimaal is voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$. De linker staaf is het aantal keer dat de optimale penalty parameter 4 is, de tweede is voor penalty parameter 5 en de laatste voor 6.



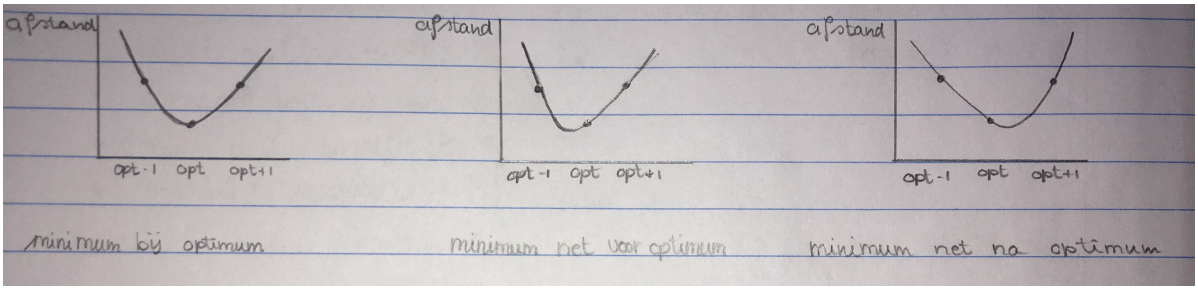
(b) Boxplot van de penalty parameter waarvoor de afstand minimaal is voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$.

Figuur 5.7: Histogram en boxplot van de optimale penalty parameters voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$. Omdat we nu werken met gehele getallen als penalty parameters, is de optimale penalty parameter ook altijd een geheel getal waardoor voor de regendata alleen de waardes 4, 5 en 6 voorkomen.

Nu we op helen nauwkeurig hebben bepaald wat de optimale penalty parameters voor de regendata $\vec{y}_{\text{regen}}$ zijn, zijn er drie mogelijkheden waar het minimum van de afstand zich kan bevinden, namelijk: op de gevonden waarde; net voor de gevonden waarde; of net na de gevonden waarde. Dit is gevisualiseerd in figuur 5.8 en belangrijk om te weten als we de optimale penalty parameter op één of meerdere decimalen nauwkeurig willen bepalen.

Uit figuren 5.7a en 5.7b wordt duidelijk dat de afstand tussen de begin-intensiteit en de geschatte intensiteit voor de 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ minimaal is voor penalty parameters van 4, 5 en 6. Dat betekent dat de optimale penalty parameters zich tussen 3 en 7 kunnen bevinden.

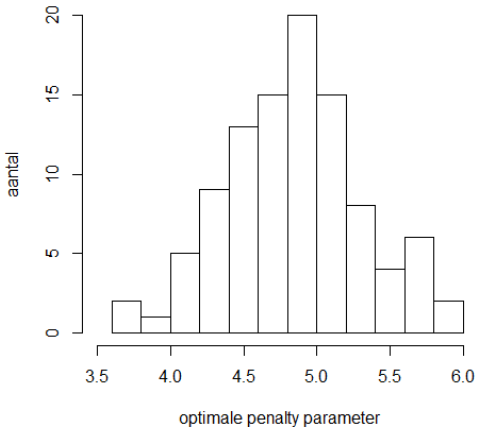
Om op één decimaal nauwkeurig te bepalen waar de optimale penalty parameter zich bevindt, gaan we als volgt te werk. We weten dat de optimale penalty parameter zich tussen de 3 en 7 bevindt. Daarom gaan we voor dezelfde 100 gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$ als eerder de intensiteit schatten met taut string maar nu met penalty parameters van 3,0, 3,1, 3,2,..., 6,8, 6,9 en 7,0. Voor iedere gegenereerde dataverzameling $\vec{y}^*_{\text{regen}}$ en iedere penalty parameter β berekenen we de afstand tussen de begin-intensiteit $\vec{x}_5$ en de geschatte intensiteit



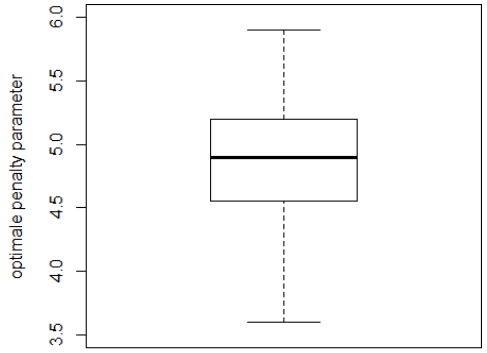
Figuur 5.8: Drie mogelijkheden waar het minimum zich kan bevinden. V.l.n.r.: op de laatste waarde, net voor de laatste waarde, net na de laatste waarde.

$\vec{x}^*_\beta$. De penalty parameter waarvoor deze afstand minimaal is, nemen we als optimale penalty parameter.

Zo krijgen we 100 optimale penalty parameters met één decimaal. In figuur 5.9a staat een histogram van de penalty parameter waarvoor de afstand minimaal is voor de 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$. De linker staaf is het aantal keer dat de optimale penalty parameter 3,6, 3,7 en 3,8 is, de tweede is voor 3,9 en 4,0, de derde voor 4,1 en 4,2 en zo door tot de laatste staaf die bij penalty parameters 5,9 en 6,0 hoort. Te zien is dat een optimale penalty parameter van 4,9 of 5,0 het vaakst (namelijk 20 keer) voorkomt. Uit figuur 5.9b kunnen we afleiden dat de mediaan van de optimale penalty parameters 4,9 is. Berekenen we het gemiddelde van de 100 optimale penalty parameters dan komen we uit op een waarde van 4,895.



(a) Histogram van de penalty parameter waarvoor de afstand minimaal is voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$. De linker staaf is het aantal keer dat de optimale penalty parameter 3,6, 3,7 en 3,8 is, de tweede is voor 3,9 en 4,0, de derde voor 4,1 en 4,2 en zo door tot de laatste staaf die bij penalty parameters 5,9 en 6,0 hoort.



(b) Boxplot van de penalty parameters waarvoor de afstand minimaal is voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$

Figuur 5.9: Histogram en boxplot van de optimale penalty parameters voor 100 verschillende gegenereerde dataverzamelingen $\vec{y}^*_{\text{regen}}$.

Bij de optimale penalty parameters van figuur 5.7a komen we uit op een gemiddelde van 4,92 en bij de penalty parameters van 5.9a op een gemiddelde van 4,895. Om de intensiteit van de regendata $\vec{y}_{\text{regen}}$ te schatten met taut string lijkt een penalty parameter van 4,9 daarom een goede keuze.

5.2.5. Effect van de gekozen penalty parameter in stap 1

In stap 1 kiezen we een waarde voor de penalty parameter β_{start} om mee te starten. Met die penalty parameter voeren we taut string uit op de regendata $\vec{y}_{\text{regen}}$. De intensiteit die hier uit komt, gebruiken we als begin-intensiteit $\vec{x}_{\beta_{\text{start}}}$ om vervolgens de optimale penalty parameter te bepalen. Wat is het effect van deze gekozen penalty parameter β_{start} op de uiteindelijke optimale penalty parameter?

Om dit te onderzoeken, hebben we vijf verschillende waardes gekozen voor de penalty parameter β_{start}, namelijk 1, 5, 10, 20 en 50. Voor deze 5 verschillende waardes van β_{start} hebben we stap 1 tot en met 7 die aan het begin van deze paragraaf 5.2 staan, uitgevoerd. Zo krijgen we voor iedere waarde van β_{start} honderd optimale penalty parameters, voor iedere gegenereerde dataverzameling $\vec{y}^*_{\text{regen}}$ één. Voor deze 100 penalty parameters kijken we naar de gemiddelde waarde, de minimale waarde en de maximale waarde. Dit staat in tabel 5.1. Te zien is dat de gekozen waarde van β_{start} zeker een effect heeft op de uiteindelijke optimale penalty parameter.

Wat positief opvalt is dat voor kleine waardes van β_{start}, zoals 1, de optimale penalty parameter groter is dan β_{start} en dat voor grote(re) waardes van β_{start}, zoals 10, 20 en 50, de optimale penalty parameter kleiner is dan β_{start}. De optimale penalty parameter neemt dus toe voor grotere start penalty parameter β_{start}, maar de groei lijkt wel minder hard te gaan voor groter waardes van β_{start}.

Wat ook opvalt is dat voor $\beta_{\text{start}} = 5$ de waarde van de optimale penalty parameter 4,9 is en daarmee bijna gelijk aan de start penalty parameter.

Tabel 5.1: Voor vijf verschillende waardes van β_{start} en voor honderd verschillende gegenereerde datasets $\vec{y}^*_{\text{regen}}$ het gemiddelde, het minimum en het maximum van de optimale penalty parameters.

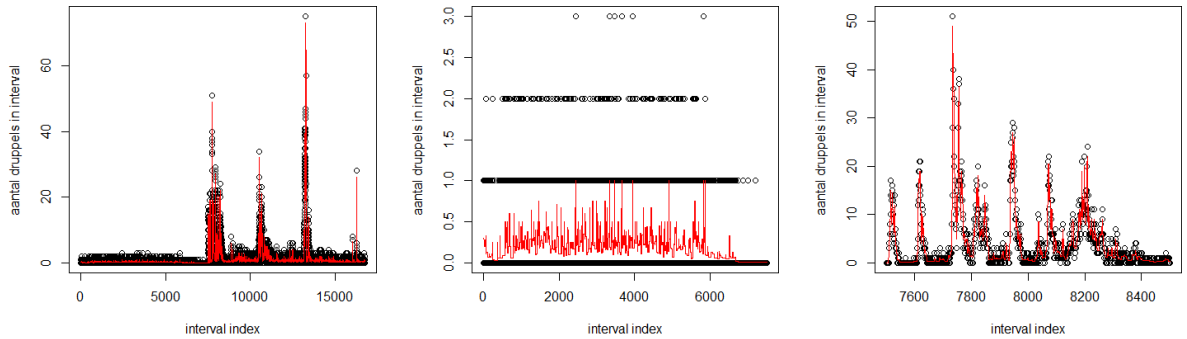
β_{start}	gemiddelde	minimum	maximum
1	2,9	2,2	3,9
5	4,9	3,6	7,1
10	6,2	4,8	7,9
20	7,5	5,5	9,9
50	10,5	7,5	13,4

Uit tabel 5.1 komt niet heel duidelijk één waarde voor de start penalty parameter β_{start} naar voren. Daarom bepalen we op het oog voor welke waarde van β_{start} de fit van de intensiteit redelijk is en voor welke waarde de fit slecht is. Met taut string schatten we de intensiteit van de regendata $\vec{y}_{\text{regen}}$ voor penalty parameters 1, 5, 10, 20 en 50. Dit geeft de vijf vectoren $\vec{x}_1$, $\vec{x}_5$, $\vec{x}_{10}$, $\vec{x}_{20}$ en $\vec{x}_{50}$. Met plotjes van $\vec{y}_{\text{regen}}$ en deze verschillende intensiteiten kunnen we op het oog bepalen voor welke penalty parameter de schatting redelijk is en voor welke penalty parameters de schatting slecht is.

Voor de vijf verschillende schattingen van de intensiteit van de regendata plotten we in de linker figuren van de figuren 5.10 t/m 5.14 alle regendata $\vec{y}_{\text{regen}}$ in het zwart en de geschatte intensiteit $\vec{x}_\beta$ met β =1, 5, 10, 20 en 50 in het rood; in de middelste figuren is het eerste redelijk constante stuk van de regendata $y_1, \ldots, y_{7500}$ (de eerste 7500 intervallen) geplot met de intensiteit; in de rechter figuren is de eerste piek van de regendata $y_{7500}, \ldots, y_{8500}$ (intervallen met index 7500 tot 8500) geplot met de intensiteit.

In figuur 5.10 is te zien dat met een penalty parameter van 1 de schatting van de intensiteit veel heen en weer springt tussen de datapunten. Ook voor het constante stuk van de data, schommelt de intensiteit behoorlijk (zie figuur 5.10b). Een penalty parameter van 1 lijkt dus te

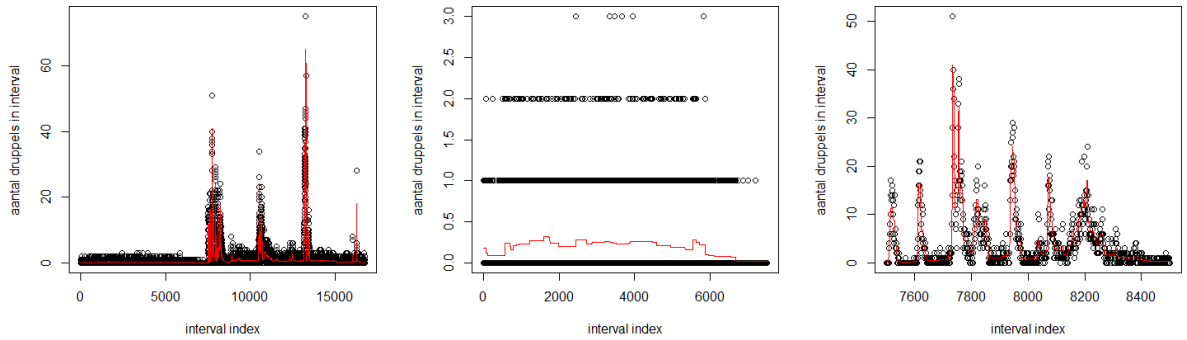
klein voor een goede fit van de schatting.



(a) Alle data $\vec{y}_{\text{regen}}$ (b) Constante stuk, $y_1, \ldots, y_{7500}$ (c) Eerste piek, $y_{7500}, \ldots, y_{8500}$

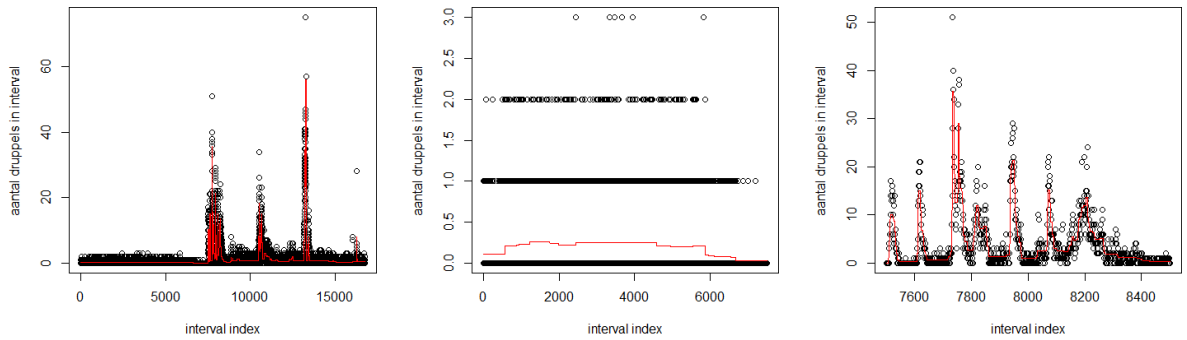
Figuur 5.10: Intensiteit van de data (rood) geschat met taut sting met een penalty parameter van 1.

Voor de intensiteiten geschat met een penalty parameter van 5, 10 en 20 ziet de fit er redelijk goed uit. Op het constante stuk van de data schommelt de intensiteit voor een penalty parameter van 5 (figuur 5.11b) nog wel wat meer dan voor 10 en 20 (figuren 5.12b en 5.13b) maar niet buitenproportioneel veel. Voor de eerste piek (figuren 5.11c, 5.12c en 5.13c) ziet de schatting van de intensiteit er ook redelijk uit voor alle drie de penalty parameters. De fit van de geschatte intensiteit lijkt dus redelijk te zijn voor penalty parameters 5, 10 en 20.



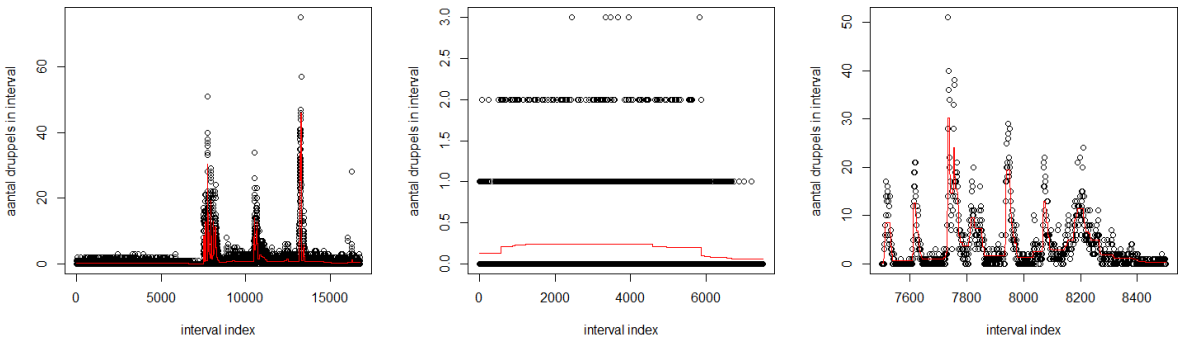
(a) Alle data $\vec{y}_{\text{regen}}$ (b) Constante stuk, $y_1, \ldots, y_{7500}$ (c) Eerste piek, $y_{7500}, \ldots, y_{8500}$

Figuur 5.11: Intensiteit van de data (rood) geschat met taut sting met een penalty parameter van 5.



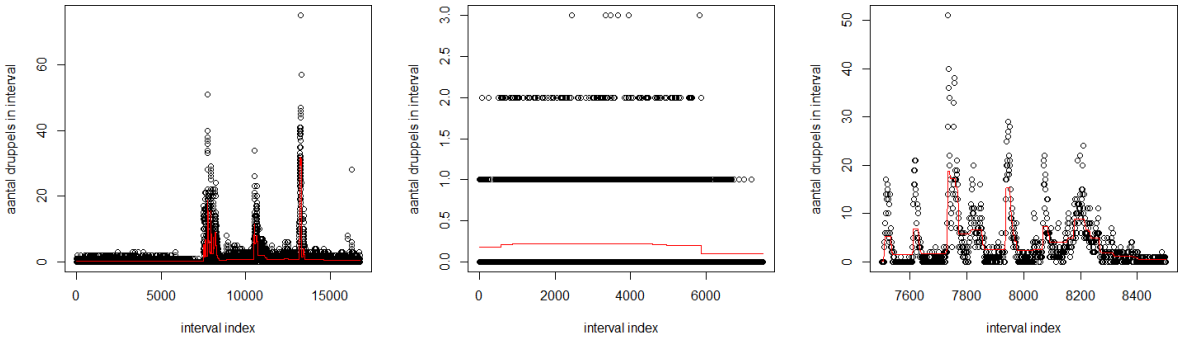
(a) Alle data $\vec{y}_{\text{regen}}$ (b) Constante stuk, $y_1, \ldots, y_{7500}$ (c) Eerste piek, $y_{7500}, \ldots, y_{8500}$

Figuur 5.12: Intensiteit van de data (rood) geschat met taut sting met een penalty parameter van 10.



(a) Alle data $\vec{y}_{\text{regen}}$ (b) Constante stuk, $y_1, \ldots, y_{7500}$ (c) Eerste piek, $y_{7500}, \ldots, y_{8500}$

Figuur 5.13: Intensiteit van de data (rood) geschat met taut sting met een penalty parameter van 20.



(a) Alle data $\vec{y}_{\text{regen}}$ (b) Constante stuk, $y_1, \ldots, y_{7500}$ (c) Eerste piek, $y_{7500}, \ldots, y_{8500}$

Figuur 5.14: Intensiteit van de data (rood) geschat met taut sting met een penalty parameter van 50.

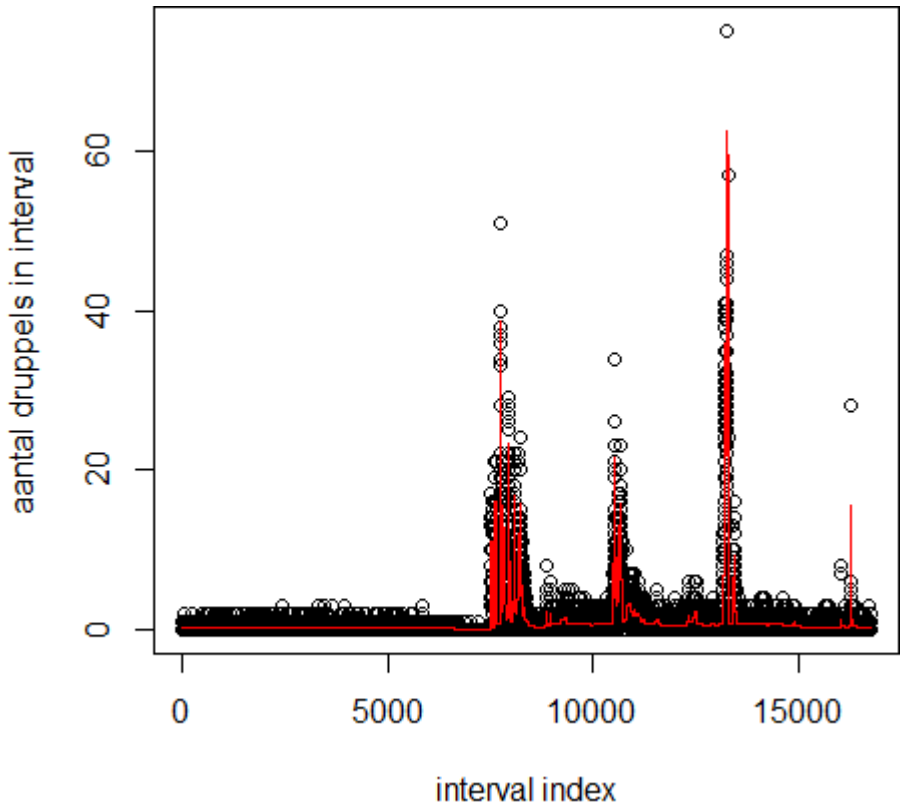
Bij een penalty parameter van 50 wordt de schatting wat grof. Voor de intensiteit van het constante stuk in figuur 5.14b is er misschien weinig verschil te zien met de schatting met een penalty parameter van 20 in figuur 5.13b maar kijken we naar de schatting van de intensiteit van de eerste piek dan is er wel een duidelijk verschil tussen een penalty parameter van 50 in figuur 5.14c en 20 in figuur 5.13c. Zo zijn er net voor interval index 7800 twee pieken (het lijkt bijna één piek) in het aantal druppels per interval. Bij een penalty parameter van 50 wordt de intensiteit van deze pieken geschat met één piek (zie figuur 5.14c) terwijl bij een penalty parameter van 20 (en kleiner) de intensiteit wordt geschat met twee pieken (zie figuur 5.13c. Een penalty parameter van 50 lijkt dus te groot voor een goede fit van de schatting van de intensiteit.

Schatten we de intensiteit van de regendata met taut string met een penalty parameter van 5, 10 of 20 dan lijken we een redelijke fit te hebben maar voor penalty parameters 1 en 50 lijkt de fit niet goed te zijn. Daarom focussen we ons nu op penalty parameters 5, 10 en 20.

Nemen we β_{start} gelijk aan 5, 10 en 20 dan krijgen we als gemiddelde optimale penalty parameters 4,9, 6,2 en 7,5 (zie tabel 5.1). Nemen we het gemiddelde van de drie gemiddelde optimale penalty parameters dan komen we op een penalty parameter van 6,2. Daarom kiezen we ervoor om de intensiteit van de regendata te schatten met taut string met een penalty parameter van 6,2.

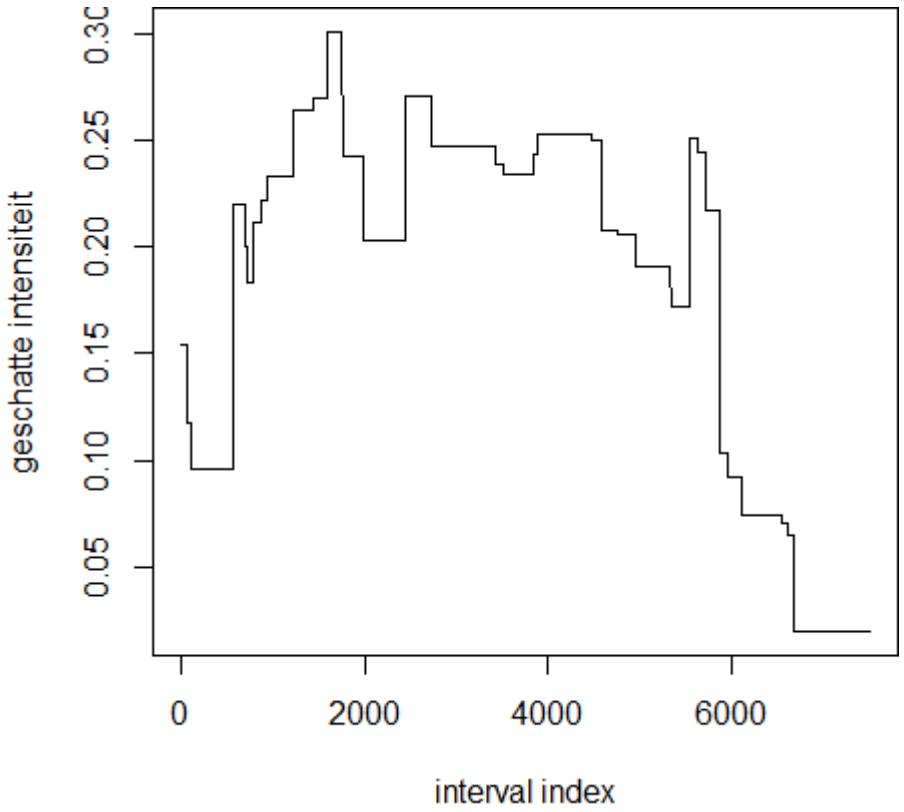
5.3. Intensiteit van de regendata schatten met taut string

In de vorige paragraaf hebben we de keuze gemaakt om met een penalty parameter van 6,2 de intensiteit van de regendata $\vec{y}_{\text{regen}}$ te schatten. Daarvoor voeren we taut string met de gekozen penalty parameter uit op de data. Dit geeft per interval van 10 seconden een schatting $x_1^{(6,2)}, x_2^{(6,2)}, \ldots, x_m^{(6,2)}$ voor de intensiteit. In figuur 5.15 staat een plot van de regendata in het zwart en de geschatte intensiteit in het rood.



Figuur 5.15: De regendata $\vec{y}_{\text{regen}}$ (zwart) en de geschatte intensiteit $\vec{x}_{6,2}$ (rood). De intensiteit is geschat met taut string met een penalty parameter van 6,2.

In figuur 5.15 lijkt het misschien alsof de geschatte intensiteit vanaf het begin tot aan de eerste piek constant is. Vergroten we de intensiteit in de eerste 7500 intervallen echter uit, zoals in figuur 5.16, dan is duidelijk te zien dat de intensiteit een stuksgewijs constante functie is en wel degelijk van waarde verandert in het eerste gedeelte van de data.



Figuur 5.16: De intensiteit van de eerste 7500 intervallen van 10 seconden van de data, geschat met taut string met een penalty parameter van 6,2

5.4. De data toetsen op een inhomogeen Poisson proces

Omdat taut string een stuksgewijs constante functie geeft, zullen er meerdere intervallen achter elkaar zijn met dezelfde intensiteit. Zo ontstaan er deelperiodes in de regendata met dezelfde intensiteit. Deze deelperiodes kunnen we toetsen op het hebben van een Poissonverdeling met intensiteit gelijk aan de intensiteit geschat met taut string.

Net als in paragraaf 4.7 is de nulhypothese H_0 : het aantal regendruppels per tijdsinterval volgt een Poissonverdeling met constante intensiteit. Omdat we met een Poissonverdeling werken, is het verwachte aantal druppels per interval gelijk aan de intensiteit van de Poissonverdeling.

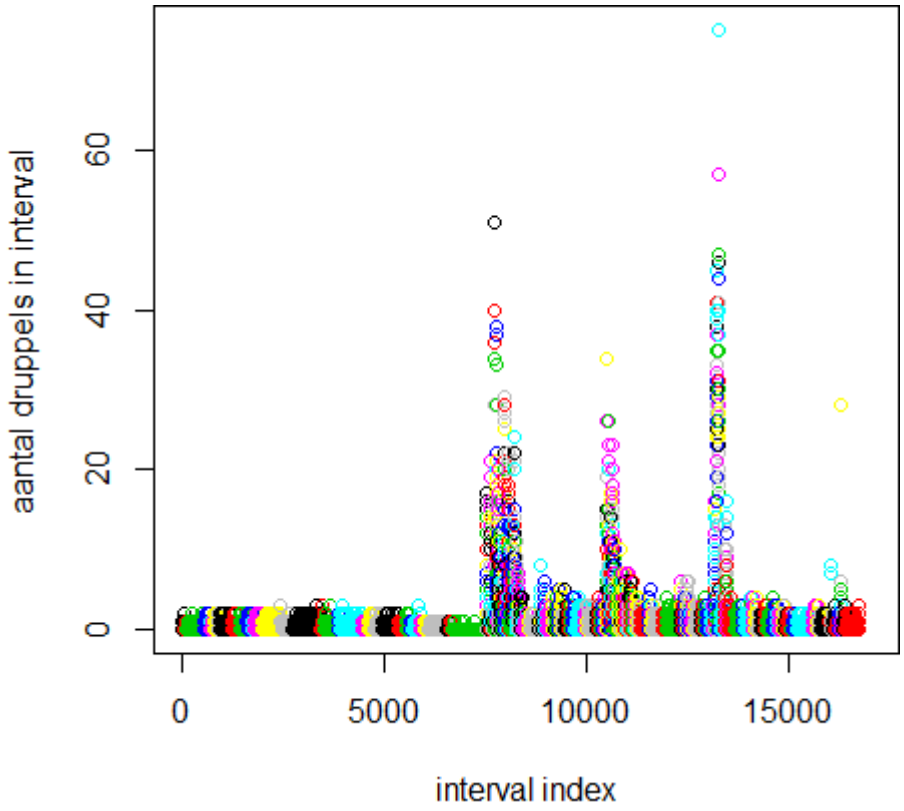
Als de nulhypothese niet wordt verworpen, dan kan het aantal druppels per interval homogeen Poisson-verdeeld zijn met parameter gelijk aan de geschatte intensiteit. Dit betekent dat de data dan met een homogeen Poissonproces kan worden beschreven met dezelfde parameter als die van de Poissonverdeling.

Als we de nulhypothese dus niet verwerpen voor een bepaalde deelperiode van de regendata, dan kan die deelperiode van de regendata met een homogeen Poissonproces worden beschreven. Wordt de nulhypothese voor meerdere deelperiodes achter elkaar niet verworpen, dan hebben we meerdere deelperiodes achter elkaar die met een homogeen Poisson proces kunnen worden beschreven maar met verschillende intensiteiten. Al deze deelperiodes samen kunnen dan met een inhomogeen Poissonproces met stuksgewijs constante intensiteit worden beschreven.

Om de regendata te toetsen op een inhomogeen Poissonproces met stuksgewijs constante

intensiteit gaan we als volgt te werk. Met taut string wordt de data opgedeeld in meerdere deelperiodes met dezelfde intensiteit. Voor deze deelperiodes kunnen we op eenzelfde manier als in subparagraaf 4.7.5 toetsen of het aantal druppels per interval homogeen Poisson-verdeeld kan zijn of niet.

Met een penalty parameter van 6,2 wordt de regendata opgedeeld in 474 deelperiodes met een constante intensiteit. In figuur 5.17 is de regendata geplot en is iedere deelperiode met een eigen kleur aangegeven. In de eerste 7500 intervallen zijn de verschillende deelperiodes nog redelijk goed te onderscheiden maar in de pieken worden de deelperiodes korter en zijn ze minder goed van elkaar te onderscheiden.

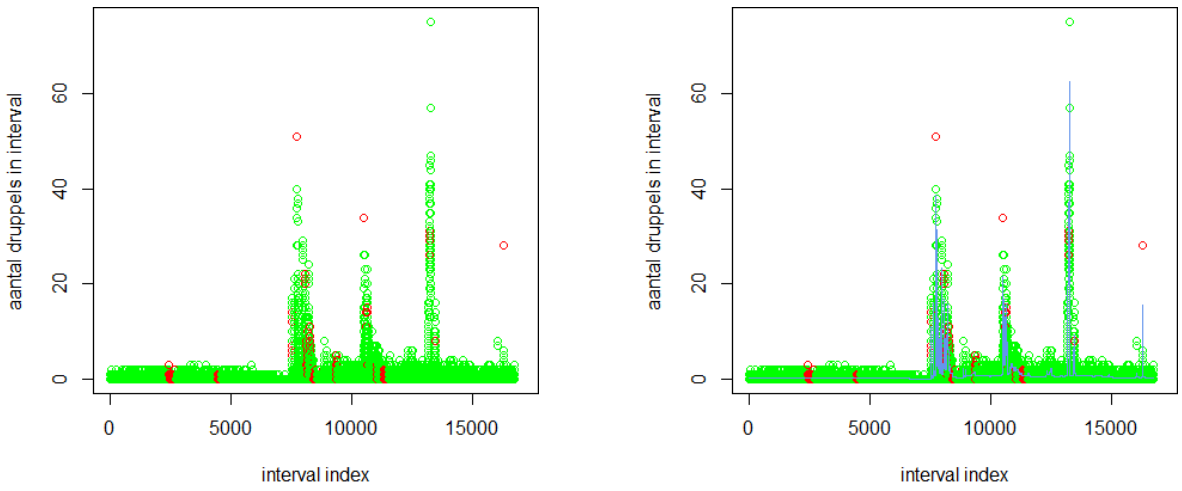


Figuur 5.17: De 474 deelperiodes waain de regendata met taut string met een penalty parameter van 6,2 is verdeeld.

Voor elke deelperiode voeren we net als in paragraaf 4.7.5 een parametrische bootstrapsimulatie uit. Als eerste schatten we de intensiteit van de deelperiode met het steekproefgemiddelde. Daarna trekken we net zoveel punten uit een Poissonverdeling met parameter gelijk aan het steekproefgemiddelde als de deelperiode intervallen heeft. Dit doen we 100.000 keer voor iedere deelperiode omdat we er voor gekozen hebben om simulaties uit te voeren met 100.000 runs. De p-waarde kunnen we nu berekenen door de fractie te nemen van de waardes van de toetsingsgrootheid van de trekking die groter zijn dan de waarde van de toetsingsgrootheid van de deelperiode.

We werken weer met een significantieniveau van 5% en daarom wordt de nulhypothese verworpen als de p-waarde kleiner is dan 0,05. Voor 26 van de deelperiodes wordt de nulhypothese verworpen, voor 448 deelperiodes wordt de nulhypothese niet verworpen. In figuur 5.18 staat een weergave van de regendata. In figuur 5.18a staat alleen de regendata, in figuur 5.18b staat de regendata en de intensiteit geschat met taut string met een penalty parameter van 6,2. Voor het aantal druppels in de groene intervallen wordt de nulhypothese niet verwor-

pen op basis van de gevonden p-waardes in de simulatie, voor het aantal druppels in de rode intervallen wel.



(a) (b)

Figuur 5.18: Het aantal druppels per interval. Op de x-as staat de index van het interval, op de y-as het aantal druppels in het interval. Voor de groene punten wordt de nulhypothese niet verworpen; voor de rode punten wel. Links is alleen de data weergegeven, rechts is naast de data ook de geschatte intensiteit weergegeven. De intensiteit is de intensiteit van de regendata geschat met taut string met een penalty parameter van 6,2.

Ten opzichte van figuur 4.4 zijn er meer stukken van de data die homogeen Poisson-verdeeld kunnen zijn. Zelfs van de hoge pieken zijn er stukken die homogeen Poisson-verdeeld kunnen zijn. Opvallend is dat voor het hoogste punt van de eerste, tweede en vierde piek de nulhypothese wel wordt verworpen maar voor het hoogste punt van de derde piek niet. Waar dat aan ligt is niet helemaal duidelijk. Alle vier de hoogste intervallen van de pieken vormen namelijk een ‘eigen deelperiode’ in de regendata. De intervallen hebben een bepaalde geschatte intensiteit en de intervallen daar net voor en net na hebben een andere intensiteit. De hoogste intervallen van de vier pieken hebben dus allemaal een eigen intensiteit.

In principe verwerp je voor één interval nooit dat het aantal druppels in dat interval Poisson-verdeeld is maar voor de regendata gebeurt dit toch. Het is dus best bijzonder dat we voor de regendata verwerpen dat het aantal druppels in één interval Poisson-verdeeld is maar wat het nog merkwaardiger maakt is dat we voor sommige interval wel en voor andere intervallen niet verwerpen dat het aantal druppels Poisson-verdeeld is. Nou kan dat misschien te maken hebben met het verschil tussen het aantal druppels in een interval en de geschatte intensiteit voor dat interval. In het hoogste interval van de eerste piek zijn er bijvoorbeeld 51 regendruppels gevallen en is de geschatte intensiteit 38,6, dit is een verschil van 12,4. Kijken we voor de drie andere pieken ook naar het verschil tussen het aantal druppel in het hoogste interval en de geschatte intensiteit dan is dat ook allemaal 12,4. Het verschil is dus overal hetzelfde maar het quotiënt van het aantal druppels en de geschatte intensiteit is wel verschillend voor de vier pieken. Voor de eerste piek is dit 1,321, voor de tweede 1,574, voor de derde 1,198 en voor de vierde 1,794. De grootte van het quotiënt is dus waarschijnlijk de reden waarom voor piek één, twee en vier wel wordt verworpen dat het aantal druppels homogeen Poisson-verdeeld is en voor piek drie niet.

5.4.1. Conclusie wel of niet inhomogeen Poisson

De regendata bestaat in totaal uit 16.734 intervallen van 10 seconden. Met taut string worden deze intervallen verdeeld in 474 deelperiodes met een constante intensiteit die per deelperiode verschilt. Voor 26 van deze deelperiodes wordt verworpen dat het aantal druppels per interval homogeen Poisson-verdeeld is, voor de andere 448 deelperiodes wordt niet verworpen dat het aantal druppels per interval homogeen Poisson-verdeeld is. Zoals in figuur 5.18 te zien is, zijn er meerdere deelperiodes (de groene) met verschillende intensiteit achter elkaar waarvoor het aantal druppels per interval homogeen Poisson-verdeeld kan zijn. Dit betekent dat voor die deelperiodes de regenval met een inhomogeen Poissonproces met stuksgewijs constante intensiteit kan worden beschreven.

Belangrijk is om te beseffen dat we er nu voor gekozen hebben om de intensiteit van de regen te schatten met een stuksgewijs constante functie met taut string met een penalty parameter van 6,2. De gemiddelde lengte van de deelperiodes met constante intensiteit is 36 intervallen van 10 seconden, maar van de 474 deelperiodes zijn er ook 297 die bestaan uit tien intervallen of minder. Voor deze deelperiodes zal de nulhypothese niet snel worden verworpen. Het zou dus kunnen dat voor bepaalde deelperiodes niet wordt verworpen dat het aantal druppels per interval homogeen Poisson-verdeeld is terwijl dat eigenlijk wel zou moeten.

Het zou ook kunnen dat een andere intensiteit, bijvoorbeeld een gladde functie, beter bij de regendata past. Dat hebben we niet onderzocht en is mogelijk nog wel interessant als vervolg.

6

Conclusie en Discussie

In deze scriptie is onderzocht of een Poissonproces, homogeen of inhomogeen, een redelijk model is om regenbuien in Afrika mee te modelleren.

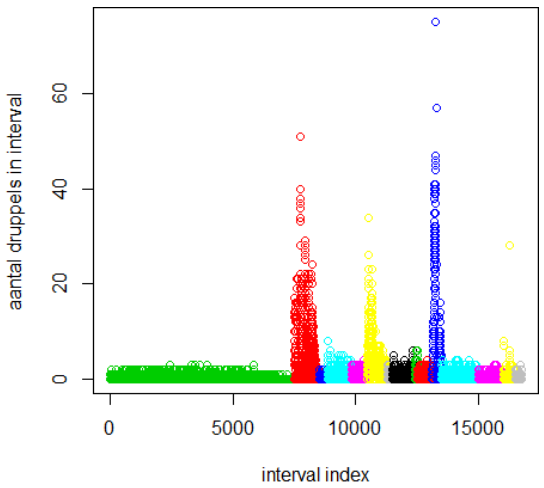
De onderzochte regenbui is opgenomen op Pole Pole vanaf 1 mei 12 uur tot 3 mei 10:30 uur. Deze periode noemen we de observatieperiode. Als eerste is onderzocht of de regen in de gehele observatieperiode kan worden beschreven met een homogeen Poissonproces oftewel een Poissonproces met constante intensiteit.

Met een parametrische bootstrapsimulatie is gevonden dat voor alle regen in de gehele observatieperiode samen de p-waarde 0 is. Dit betekent dat bij ieder significantieniveau wordt verworpen dat de regen in de gehele observatieperiode met een homogeen Poissonproces kan worden beschreven. Een homogeen Poissonproces is dus niet een redelijk model om regenbuien in Afrika mee te modelleren.

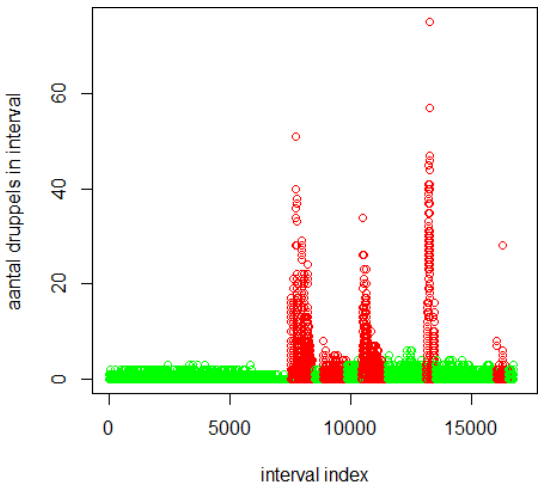
Er zijn wel deelperiodes van de observatieperiode waarin de regen met een homogeen Poissonproces kan worden beschreven. Om dit te onderzoeken is de observatieperiode op het oog opgedeeld in vijftien deelpriodes met een redelijk constant-lijkende intensiteit zoals te zien is in figuur 6.1a.

Voor iedere deelperiode voeren we een parametrische bootstrapsimulatie uit zoals ook voor de gehele observatieperiode gedaan is. Met een significatieniveau van 5%, wordt voor zes van de vijftien deelperiodes verworpen dat de regen met een homogeen Poissonproces kan worden beschreven, voor de andere negen deelperiodes wordt dit niet verworpen en kan de regen dus wel met een homogeen Poissonproces worden beschreven. In figuur 6.1b kan de regen in de groene stukken wel worden beschreven met een homogeen Poissonproces en de regen in de rode stukken niet. Regen met een lage intensiteit kan dus wel met een homogeen Poissonproces worden beschreven maar regen met een hoge(re) intensiteit niet.

Dat er in figuur 6.1 meerdere deelperiodes zijn waarin de regen met een homogeen Poissonproces kan worden beschreven, is een aanwijzing dat de regen misschien met een inhomogeen Poissonproces kan worden beschreven. Om dit objectiever te onderzoeken is de data met taut string opgedeeld in 474 deelperiodes met een constante intensiteit die per deelperiode anders is. Deze deelperiodes zijn in figuur 6.2a weergegeven met een eigen kleur.

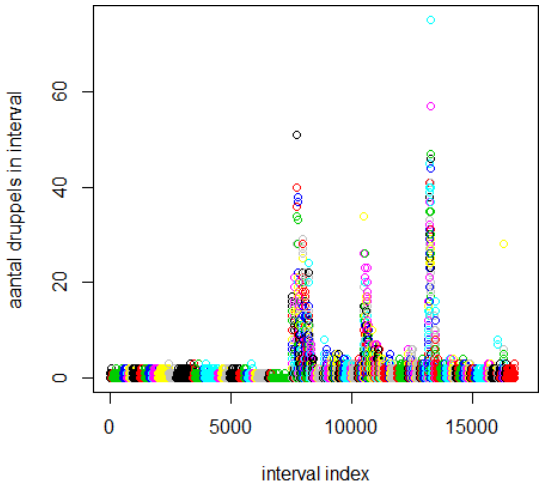


(a) De vijftien deelperiodes waarin de regendata op het oog is verdeeld.

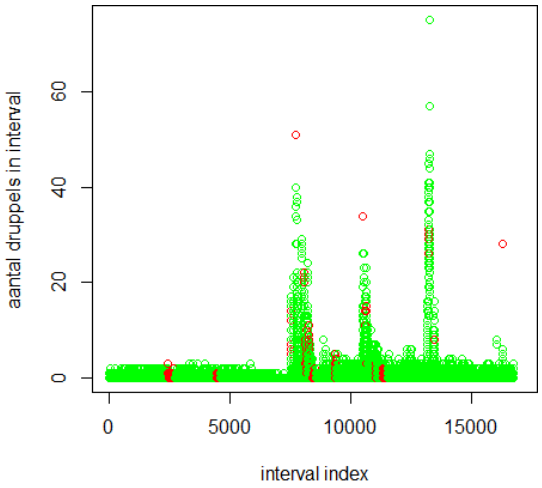


(b) De deelperiodes waarin de regen wel met een homogeen Poissonproces kan worden beschreven in het groen en de deelperiodes waarin de regen niet met een homogeen Poissonproces kan worden beschreven in het rood.

Figuur 6.1



(a) De 474 deelperiodes waarin de regendata met taut string met een penalty parameter van 6,2 is verdeeld.



(b) De deelperiodes waarin de regen wel met een homogeen Poissonproces kan worden beschreven in het groen en de deelperiodes waarin de regen niet met een homogeen Poissonproces kan worden beschreven in het rood.

Figuur 6.2

Voor iedere deelperiode wordt weer een parametrische bootstrapsimulatie uitgevoerd waarmee de p-waarde wordt bepaald. Met een significantieniveau van 5% wordt voor 26 van de 474 deelperiodes verworpen dat de regen met een homogeen Poissonproces kan worden beschreven (de rode stukken in figuur 6.2b), voor de andere 448 deelperiodes wordt dit niet verworpen en kan de regen wel met een homogeen Poissonproces worden beschreven (de groene stukken in figuur 6.2b).

Er zijn dus meerdere deelperiodes met verschillende constante intensiteit achter elkaar

waarin de regen met een homogeen Poissonproces kan worden beschreven. Nemen we de regen van deze deelperiodes samen dan kunnen we deze regen beschrijven met een inhomogeen Poissonproces met stuksgewijs constante intensiteit. Een inhomogeen Poissonproces met stuksgewijs constante intensiteit lijkt dus een redelijk model om regenbuien in Afrika mee te modelleren.

Discussie

Bij het onderzoeken of de regen met een homogeen Poissonproces kan worden beschreven, zijn resultaten uit eerder onderzoek van de Villiers e.a., 2021 bevestigd. Alle regen van de gehele observatieperiode kan niet goed met een homogeen Poissonproces worden beschreven maar er zijn wel deelperiodes, vooral met een lage intensiteit, waarin de regen wel met een homogeen Poissonproces kan worden beschreven.

De data hebben we echter op het oog ingedeeld in deelperiodes waardoor we vanwege p-hacking of data snooping voorzichtig moeten zijn met het trekken van conclusies. Door vervolgens gebruik te maken van taut string hebben we de regendata op een objectievere manier ingedeeld in deelperiodes. Deze methode is objectiever maar er is nog steeds een gedeelte dat 'op het oog' gebeurt. Er moet namelijk een penalty parameter gekozen worden. Door eerst voor honderd gegenereerde datasets de optimale penalty parameter te bepalen en daarna het gemiddelde van die optimale penalty parameters te gebruiken om de intensiteit van de data te schatten zijn we wel een laag dieper te werkt gegaan dan wanneer we zomaar een penalty parameter hadden gekozen en die hadden gebruikt om de intensiteit van de data te schatten.

Door gebruik te maken van taut string zijn er twee zaken waar we rekening mee moeten houden.

De eerste is dat de observatieperiode wordt ingedeeld in deelperiodes met constante intensiteit. De gemiddelde lengte van deze deelperiodes is 36 intervallen van 10 seconden, maar van de 474 deelperiodes zijn er ook 297 die bestaan uit tien intervallen of minder. Deze deelperiodes bestaan dus uit een gering aantal observaties waardoor de nulhypothese niet snel verworpen zal worden. Het zou dus kunnen dat we voor sommige deelperiodes concluderen dat de regen met een homogeen Poissonproces kan worden beschreven terwijl dit eigenlijk niet zo is.

De tweede is dat we aannemen dat de intensiteit van een inhomogeen Poissonproces te benaderen is met een stuksgewijs constante functie. Het zou echter ook kunnen dat een andere functie, bijvoorbeeld een gladde functie, de intensiteit van een inhomogeen Poissonproces beter benadert.

Een laatste punt waar rekening mee gehouden moet worden, is dat in deze scriptie data gebruikt is, die is verzameld gedurende twee dagen op één locatie op Mafia eiland. Het feit dat een inhomogeen Poissonproces met stuksgewijs constante intensiteit voor deze regen een redelijk model lijkt te zijn, hoeft niet te betekenen dat dit voor de rest van het jaar en voor de andere landen in Afrika ook het geval is.

Aanbevelingen

Voor vervolg onderzoek zou het interessant kunnen zijn om meerdere intervalometers op dezelfde locatie te plaatsen. Zo kan je van dezelfde regenbui gegevens verzamelen met meerdere intervalometers. De gegevens van één intervalometer zou je kunnen gebruiken om de intensiteit van de regen te schatten en de gegevens van de andere intervalometer(s) zou je kunnen gebruiken om de intensiteit te verifiëren in plaats van dat je de intensiteit verifieert via gesimuleerde data.

Wat ook nog interessant kan zijn, is om te onderzoeken of er voor een deelperiode een minimum aantal observaties is zodat we kunnen voorkomen dat we onterecht de nulhypothese niet verwerpen.

Voor een deelperiode met weinig observaties zal je niet snel verwerpen dat de regen te beschrijven is met een homogeen Poissonproces. Ook niet als de regen niet te beschrijven is met een homogeen Poissonproces. We verwerpen de nulhypothese dus onterecht niet. Als we zouden weten vanaf hoeveel observaties per deelperiode dit onterecht niet verwerpen van de nulhypothese weinig meer voorkomt, dan kunnen we daar rekening mee houden. We kunnen bijvoorbeeld bij taut string de penalty parameter zo kiezen dat bijna alle deelperiodes meer observaties bevatten dan het minimum aantal. Zo beperken we het aantal deelperiodes waarvoor we de regen classificeren als Poisson terwijl die niet Poisson is, en zo kunnen we betere conclusies trekken of het Poissonproces, homogeen of inhomogeen, en redelijk model is om regenbuien in Afrika mee te modelleren.

Bibliografie

Bijma, F., Jonker, M. & van der Vaart, A. (2017). Method of Moments Estimators. In F. Bijma, M. Jonker & A. van der Vaart (Red.), *An Intoruction to Mathematical Statistics* (pp. 72–74). Amsterdam University Press.

Cumulants. (2013). Verkregen 23 juni 2022, van http://www.stat.uchicago.edu/~pmcc/courses/stat306/2013/cumulants.pdf

DelftGlobal. (2022). *TU Delft Global Initiative*. Verkregen 23 juni 2022, van https://www.tudelft.nl/global

de Villiers, D. (2019). Something fishy going on! Evaluating the Poisson hypothesis for rainfall estimation using intervalometers: first results from an experiment in Tanzania.

de Villiers, D., Schleiss, M., ten Veldhuis, M.-C., Hut, R. & van de Giesen, N. (2021). Something fishy going on? Evaluating the Poisson hypothesis for rainfall estimation using intervalometers: results from an experiment in Tanzania. *Atmospheric Measurement Techniques*, *14*(8), 5607–5623. https://doi.org/https://doi.org/10.5194/amt-14-5607-2021

Dümbgen, L. & Kovac, A. (2009). Extensions of smoothing via taut strings. *Electronic Journal of Statistics*, *3*(none), 41–75. https://doi.org/https://doi.org/10.1214/08-EJS216

Grimmett, G. & Welsh, D. (2014). *Probability: An Introduction* (Second). Oxford University Press.

Gürtler, N. & Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *Journal of Statistical Planning and Inference*, *90*(2), 207–225. https://doi.org/https://doi.org/10.1016/S0378-3758(00)00114-2

Henze, N. (1996). Empirical-distribution-function goodness-of-fit tests for discrete models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *24*(1), 81–93. https://doi.org/https://doi.org/10.2307/3315691

Hoagline, D. C. (1980). A poissonness plot. *The American Statistician*, *34*, 146–149.

Karlis, D. & Xekalaki, E. (2000). A Simulation Comparison of Several Procedures for Testing the Poisson Assumption. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 355–382. https://doi.org/https://doi.org/10.1111/1467-9884.00240

Kim, A. (2019). *Sum of Exponential Random Variables*. Verkregen 30 juni 2022, van https://www.researchgate.net/publication/1844713_Inter-arrival_time_distribution_for_the_non-homogeneous_Poisson_process

Klar, B. (1999). Goodness-of-fit tests for discrete models based on the integrated distribution function. *Metrika*, *49*, 53–69. https://doi.org/https://doi.org/10.1007/s001840050025

KNMI. (2022). *Automatische Weerstations*. Verkregen 20 juni 2022, van https://www.knmi.nl/kennis-en-datacentrum/uitleg/automatische-weerstations

Lechner, I., Evans, N., Markakis, V. & Vidak, M. (2021). *P-value hacking*. The Embassy of good science. Verkregen 16 mei 2022, van https://embassy.science/wiki/Theme:6b584d4e-2c9d-4e27-b370-5fbdb983ab46

Lindsay, B. G. (1986). Exponential Family Mixture Models (with Least-Squares Estimators). *The Annals of Statistics*, *14*(1), 124–137. http://www.jstor.org/stable/2241270

Lindsay, B. G. & Roeder, K. (1992). Residual Diagnostics for Mixture Models. *Journal of the American Statistical Association*, *87*(419), 785–794. https://doi.org/https://doi.org/10.2307/2290216

Nieuwenhuizen, M. (2010). *Luisteren naar regendruppels*. Verkregen 18 juni 2022, van https://www.nemokennislink.nl/publicaties/luisteren-naar-regendruppels/

ONSET. (2022). *HOBO Rain Gauge Data Logger*. Verkregen 18 juni 2022, van https://www.onsetcomp.com/products/data-loggers/rg3/

Ord, J. K. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society: Series A*, *130*, 232–238.

Our Mandate. (2022). Verkregen 20 juni 2022, van https://www.tudelft.nl/citg/over-faculteit/afdelingen/watermanagement/medewerker/staff-water-resources-management/headsecretary/prof-dr-ir-nick-van-de-giesen

Pennsylvania State University, S. (2018). *Bootstrapping*. PennState Eberly College of Science. Verkregen 3 mei 2022, van https://online.stat.psu.edu/stat555/node/119/

Pishro-Nik, H. (2014). Poisson processes [toegankelijk via https://www.probabilitycourse.com/]. In H. Pishro-Nik (Red.), *Introduction to probability, statistics and random processes*. Kappa Research LLC.

Santner, T. & Duffy, D. (1989). *The Statistical Analysis of Discrete Data*. Springer.

Spinelli, J. J. & Stephens, M. A. (1997). Cramér-von Mises Tests of Fit fot het Poisson Distribution. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *25*(2), 257–268. https://doi.org/https://doi.org/10.2307/3315735

TUDelft. (2010). *Trans-African Hydro-Meteorological Observatory* [Video]. Youtube. Verkregen 6 april 2022, van https://youtu.be/Nh7GDD3Ssr8

TUDelft. (2019). *20.000 weerstations in Afrika*. Verkregen 20 juni 2022, van https://www.tudelft.nl/citg/onderzoek/stories-of-science/20000-weerstations-in-afrika

TUDelft. (2022a). *INtervalometer*. Verkregen 7 april 2022, van https://www.tudelft.nl/citg/over-faculteit/afdelingen/watermanagement/medewerker/staff-water-resources-management/headsecretary/prof-dr-ir-nick-van-de-giesen

TUDelft. (2022b). *Trans African Hydro Meteorological Observatory (TAHMO)*. Verkregen 7 april 2022, van https://www.tudelft.nl/citg/over-faculteit/afdelingen/watermanagement/onderzoek/chairs/water-resources/water-resources-management/research/projects/trans-african-hydro-meteorological-observatory-tahmo

van de Giesen, N. (2019). *Intervalometer*. Verkregen 18 juni 2022, van https://github.com/nvandegiesen/Intervalometer/wiki/Intervalometer

Winton, C. (2009). Lecture notes in Computer Assisted Diagnosis.

Yakovlev, G., Rundle, J., Shcherbakov, R. & Turcotte, D. (2005). Inter-arrival time distribution for the non-homogeneous Poisson process. https://www.researchgate.net/publication/1844713_Inter-arrival_time_distribution_for_the_non-homogeneous_Poisson_process