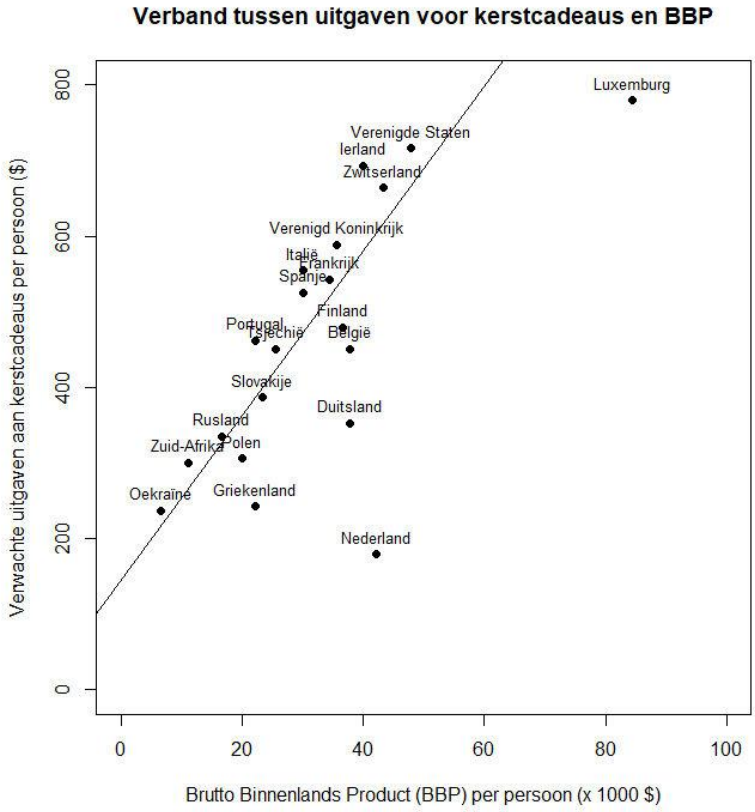
Het spreidingsdiagram opnieuw bekeken



Jelke Bethlehem

juni 2019

1. Over grafieken

Grafieken hebben door de eeuwen heen een belangrijke rol gespeeld bij statistische analyse. Inderdaad, door de eeuwen heen, want honderden jaren geleden werd al gebruikt gemaakt van grafieken. Bekend is het werk van Engelsman John Playfair (1786). Hij tekende tijdreeksen van economische gegevens. Ook het werk van de Fransman Minard mag niet onvermeld blijven. Hij maakte mooie en duidelijke grafieken, die veel informatie bevatten. Bekend is zijn grafisch verslag van de veldtocht van Napoleon naar Rusland in 1812. Tufte (1983) beschrijft deze grafiek. Ook het werk van John Snow mag niet onvermeld blijven. Mede door zijn kaarten van cholerагevallen in Londen in 1854, kon de oorzaak van deze epidemie worden vastgesteld. Tufte (1983) beschrijft ook deze mooie en effectieve grafiek.

Lange tijd was het maken van grafieken een tijdrovend handwerk. Maar de situatie verandert omstreeks de 80er jaren van de vorige eeuw. Er zijn twee belangrijke ontwikkelingen:

- De opkomst van de exploratieve statistiek in de zeventiger jaren van de vorige eeuw. Gangmaker hiervan was Tukey (1977). In zijn boek toont hij aan dat grafische voorstellingen een belangrijk hulpmiddel kunnen zijn bij de speurtocht van een onderzoeker naar patronen en structuren in verzamelde gegevens.
- De opkomst van de computer en vooral die van de microcomputer. Die maken het steeds eenvoudiger om grafieken te maken. Veel gebruikte software zoals Excel heeft faciliteiten voor het maken van veelvoorkomende grafieken. En er komt steeds meer statistische software (*SAS, SPSS, STATA, R*) waarmee je kant-en-klare grafieken kunt maken.

De groeiende belangstelling voor grafieken zorgde er niet automatisch voor dat die grafieken ook van goede kwaliteit waren. Er zijn computerprogramma's die grafieken produceren die, statistisch gezien, niet door de beugel kunnen. Ze zetten minder ervaren gebruikers die niet zijn getraind in het doorzien van deze voetangels en klemmen van grafieken, op het verkeerde been. Dit is des te meer vervelend, aangezien grafieken bedoeld zijn voor het overbrengen van statistische informatie aan een breed publiek. Zowel onderzoekers die zelf grafieken maken als ontwikkelaars van grafische programmatuur doen er goed aan zich te bezinnen op het volgen van richtlijnen voor goede grafieken. Zulke richtlijnen kun je bijvoorbeeld vinden in Schmid (1983) en Bethlehem (2018, hoofdstukken 10 en 11).

Je kunt grafieken op twee manieren gebruiken. In de eerste plaats zijn ze een nuttig hulpmiddel bij de *exploratieve analyse* van gegevens. Grafieken kunnen zeer effectief zijn in zo'n analyse. Ze helpen bij het verkrijgen van inzicht in de gegevens door het onthullen van onverwachte patronen en structuren. Het zijn vooral de onderzoekers zelf die grafieken op deze manier gebruiken. Daarom is de opmaak van de grafieken iets minder belangrijk. Het 'ontdekken' staat voorop.

In de tweede plaats kunnen je grafieken gebruiken in *publicaties* met de uitkomsten van statistisch onderzoek. Deze grafieken moeten in staat zijn de boodschap in de gegevens ook aan minder statistische onderlegde gebruikers duidelijk te maken. Daarom is grote zorgvuldigheid vereist bij het ontwerpen van dit soort grafieken. De grafiek moet de boodschap onthullen en niet verhullen.

Een grafiek is dus een krachtig hulpmiddel om de informatie in de gegevens te tonen en begrijpelijk te maken voor een groot publiek. Grafieken zijn vaak zinvoller dan platte tekst of tabellen. Het is daarom niet verbazingwekkend dat je grafieken veel tegenkomt in onderzoeksrapporten en ook in de populaire media (websites, kranten en televisie). Dat is mooi, maar er kleven ook gevaren aan het gebruik van grafieken. Slecht ontworpen grafieken kunnen je al snel op het verkeerde been zetten, waardoor je verkeerde conclusies trekt. Er zijn talloze voorbeelden van zulke foute grafieken. Waar komen ze vandaan? Vaak zijn deze grafieken gemaakt door grafisch ontwerpers die onvoldoende kennis hebben

van statistische technieken. Ze zijn meer gericht een aantrekkelijke vormgeving dan op de statistische inhoud ervan.

In dit rapport gaat het over één type grafiek en dat is het *spreidingsdiagram*. Soms spreken we ook wel van een *puntenwolk* genoemd. Het spreidingsdiagram is een veelgebruikt instrument voor het onderzoeken en weergeven van de samenhang tussen twee variabelen. Om te komen tot een goede interpretatie is een correcte vormgeving van het spreidingsdiagram belangrijk. Een aantal aspecten komt in dit rapport aan bod. Speciaal wordt ingegaan op het weergeven van overlappende punten door middel van vlekken, zonnebloemen en zeepbellen, en op het grafisch onderzoeken van samenhang met kruis-medianen, splines en lokaal gewogen regressie.

Dit rapport is een bewerkte versie van een rapport dat al in 1987 verscheen (met dezelfde titel). Zie Bethlehem (1987). Het besturingssysteem *Windows* bestond toen nog niet. De grafieken in die publicatie zijn gemaakt met een Turbo Pascal programma dat draaide onder MS-DOS 2.0. De computer moest een grafische beeldbuis hebben. De grafieken zijn afgedrukt met een HP 7475A plotter. Ondertussen draaien de microcomputers onder Windows en dat heeft veel grafische mogelijkheden. De grafieken in de nieuwe versie van het rapport zijn gemaakt met de Open Source analyse- en programmeeromgeving *R*.

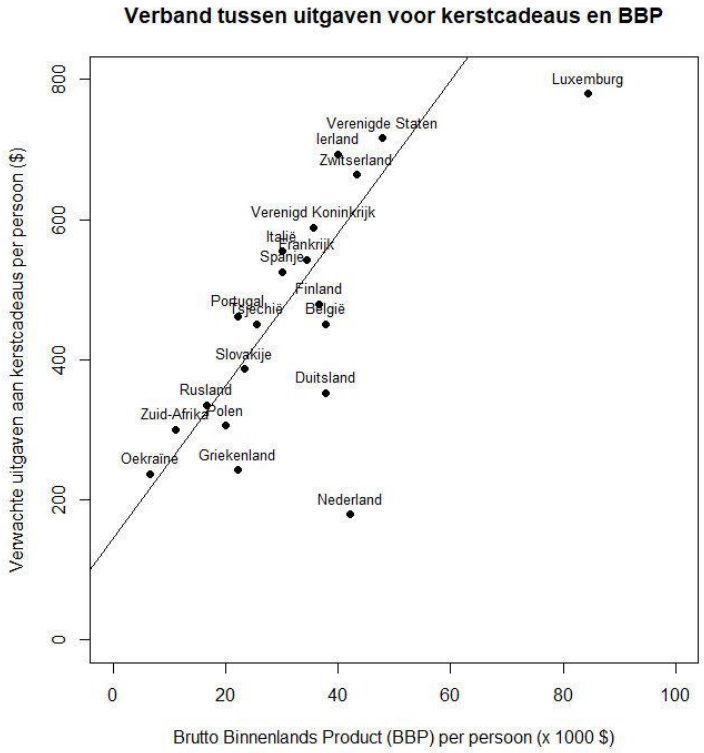
Een *spreidingsdiagram* (of(*puntenwolk*) is een diagram dat de gezamenlijke spreiding van twee variabelen X en Y weergeeft. Ieder waarnemingspaar (X_i, Y_i) wordt gerepresenteerd door een punt waarvan de coördinaten (in een gewoon rechthoekig assenstelsel) gelijk zijn aan de waarden X_i en Y_i. Een verzameling van n waarnemingsparen (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) geeft zo n punten in het diagram en de verspreiding of samenklontering van de punten geeft weer hoe de variabelen X en Y samenhangen. Dit is ongeveer de definitie zoals die is gegeven in Kendall en Buckland (1960). Kortom, een spreidingsdiagram bestaat uit een assenstelsel met een reeks punten.

Figuur 1.1 bevat een voorbeeld van een spreidingsdiagram. De gegevens zijn afkomstig uit een artikel in het tijdschrift *The Economist* op 12 december 2011. De punten stellen landen voor. Op de horizontale as (de X-as) staat het Bruto Binnenlands Product (BBP) per persoon. Het BBP per persoon van een land geeft aan hoe rijk een land is. Naarmate een punt meer naar rechts ligt, is het land rijker. Op de verticale as (de Y-as) staan de verwachte uitgaven aan kerstcadeaus. Naarmate een punt hoger ligt, zijn de uitgaven aan kerstcadeaus groter. Door het bestuderen van het patroon van punten in de grafiek kun je vaststellen of er verband bestaat tussen deze twee variabelen, En als er verband is, kun je de aard van dit verband bepalen.

In de grafiek is een globale trend te zien: inwoners van rijkere landen kopen meer kerstcadeaus dan inwoners van arme landen. Die trend kun je beschrijven door een rechte lijn. Linksonder liggen de armere landen Oekraïne en Griekenland die maar weinig aan kerstcadeaus uitgeven. Rechtsboven liggen landen als de VS, Ierland en Zwitserland die veel geld aan kerstcadeaus uitgeven.

Er zijn twee uitschieters te zien: Luxemburg en Nederland. Deze twee landen gedragen zich anders dan alle andere landen. Ze volgen niet de trend zoals weergegeven met de rechte lijn. Beide landen geven minder uit aan kerstcadeaus dan je zou verwachten op grond van hun BBP per persoon. Zulke uitschieters roepen vragen op. Wat is er met deze landen aan de hand? Zijn Nederlanders misschien krenteriger dan inwoners van de andere landen? Een andere mogelijke verklaring voor Nederland zou kunnen zijn dat de Nederlanders al veel geld uitgeven aan Sinterklaascadeaus, zodat er dan minder geld overblijft voor kerstcadeaus. Het is duidelijk dat het beeld in de grafiek oproept tot nader onderzoekt.

Figuur 1.1. Spreidingsdiagram van BBP per persoon tegen uitgaven aan kerstcadeaus. Bron: The Economist, 2011.



Hoe eenvoudig de definitie van een spreidingsdiagram ook is, er zitten best nog wat haken en ogen aan het maken ervan. In paragraaf 2 wordt ingegaan op de vormgeving van het spreidingsdiagram en de aspecten waarop je daarbij moet letten. Paragraaf 3 is gewijd aan het weergeven van overlappende punten in een spreidingsdiagram. Paragraaf 4 behandelt technieken die kunnen helpen bij het onderzoeken en beschrijven van de samenhang.

2. De vormgeving

Als je een spreidingsdiagram wilt tekenen, dan begin je met de *X*-as en de *Y*-as. En daar hebben we gelijk het eerste probleem: welke verhouding neem je voor de *X*-as en de *Y*-as? Welke verhouding van breedte en hoogte leent zich het beste voor het onderzoek naar samenhang? Hoger-dan-breed of breder-dan-hoog? Zie ook figuur 2.1.

Figuur 2.1. De verhouding tussen breedte en hoogte: hoger-dan-breed of breder-dan-hoog?

Tukey (1977) suggereert dat een hoger-dan-breed grafiek beter bruikbaar is voor verschijnselen waarin een steeds snellere (bijvoorbeeld exponentiële) groei optreedt. Bij de exploratie van veel verschijnselen kan de vorm van de samenhang echter wel eens bergen en dalen vertonen, terwijl er ook

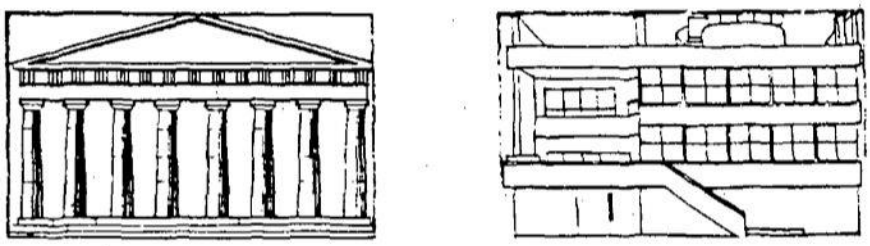
veel spreiding in de punten aanwezig kan zijn. Dergelijke verschijnselen lenen zich beter voor weergave in een breder-dan-hoog grafiek. Breder-dan-hoog grafieken zijn ook beter te volgen met het oog. Tufte (1983) wijst op de analogie met de horizon. Het menselijk oog is beter ingesteld op het waarnemen van afwijkingen ten opzichte van de horizon. Daarom moet een grafiek lijken op een landschap met een horizon, een panorama. Dit betekent dat de grafiek breder dan hoog moet zijn. Een praktisch argument is bovendien dat het plaatsen van teksten (die wel breed, maar niet hoog zijn) beter gaat in een breder-dan-hoog grafiek. In Chambers et al. (1983) en Cleveland & McGill (1984) zijn alle spreidingsdiagrammen vierkant.

Vaak is er bij de variabelen sprake van een *responsvariabele*. Dat is een variabele waarvan je het gedrag wilt verklaren uit een *verklarende variabele*. De responsvariabele zet je traditioneel op de *Y*-as (de verticale as), en de verklarende-variabele op de *X*-as (de horizontale as). In een breder-dan-hoog grafiek kun je het effect van de verklarende variabele op de respons-variabele beter onderzoeken.

Tufte (1983) heeft een groot aantal erkend goede grafieken (89) van John Playfair uit de 17-de eeuw bekeken. De breedte-hoogte verhouding van deze grafieken lag steeds tussen 1,4 en 1,8. Ze waren dus breder-dan-hoog.

Er lijkt een voorkeur te bestaan voor breder-dan-hoog grafieken. Maar welke verhouding moet je dan precies hanteren? Een aardig suggestie zou kunnen zijn de *Gulden Snede* te gebruiken. Die dateert uit de vijfde eeuw voor Christus en komt overeen met een breedte-hoogte-verhouding van 1 staat tot 1,618. Deze verhouding is ook terug te vinden in de architectuur. Figuur 2.2 bevat twee voorbeelden: het Parthenon in Athene en een villa van Le Corbusier in Parijs. Dit is kennelijk een prettig ogende verhouding.

Figuur 2.2. De Gulden Snede in de architectuur: het Parthenon in Athene en een villa van Le Corbusier in Parijs. Bron: Bergami (1969).

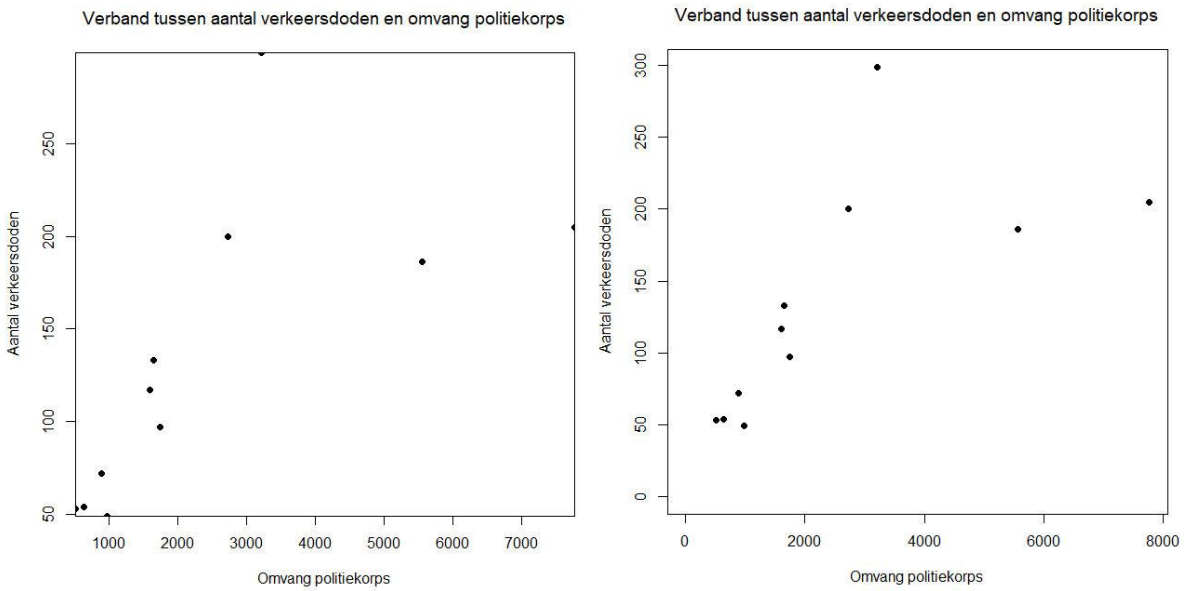


Tufte (1983) komt op grond van zijn onderzoek naar de beste breedte-hoogte-verhouding tot de volgende conclusie: Als de aard van de gegevens een bepaalde verhouding suggereert, neem die dan over in de grafiek. Neem in alle overige gevallen de breedte ruwweg anderhalf maal zo groot als de hoogte.

Een volgende stap in het ontwerp van het spreidingsdiagram is het vastleggen van het waardebereik van beide assen. De meest voor de hand liggende techniek is het waardebereik van de assen gelijk nemen aan het waardebereik van de gegevens. Dit is echter niet verstandig. Het heeft namelijk tot gevolg dat een aantal punten op de randen van het grafiekgebied komt te liggen. Daardoor zijn ze slecht zichtbaar en kunnen dus makkelijk over het hoofd worden gezien.

Figuur 2.3 toont een voorbeeld. Hierin is gebruikt gemaakt van CBS-cijfers uit de 80er jaren over de omvang van de politiekorpsen en het aantal dodelijk ongelukken in de (toen nog) 11 provincies. De omvang van het politiekorps staat op de *X*-as en het aantal dodelijke ongelukken op de *Y*-as.

Figuur 2.3. Spreidingsdiagram van de omvang van het politiekorps tegen het aantal dodelijke ongelukken. Bron CBS, 1987.



In de linker grafiek zijn begin- en eindpunt van de assen gelijk genomen aan de kleinste en grootste waarden van de variabele. Daardoor komen vier punten op de rand van het vierkant te liggen. Dat zijn Noord-Brabant (politie=3216, doden=299) op de bovenrand, Zuid-Holland (politie=7775, doden=205) op de rechter rand, Groningen (politie=972, doden=49) op de onderrand en Zeeland (politie=504, doden=53) op de linker rand. Die punten zijn maar amper zichtbaar. Vooral Zeeland is vrijwel niet te zien.

In de rechter grafiek zijn de tekortkomingen weggewerkt. Er liggen geen punten meer op de randen. Als een variabele een natuurlijk nulpunt heeft, dan hoort de as te beginnen bij dat nulpunt. Dat is het geval in de rechter grafiek van figuur 2.3.

Sommige computerprogramma's bieden de mogelijkheid om de assen automatisch te schalen. Dan kun je een spreidingsdiagram krijgen zoals de linker grafiek in figuur 2.3. Dat is dus niet zo'n goede grafiek. Volgens Cleveland en McGill (1984) is het dan beter om automatisch een marge in te bouwen van 5% tot 10% van het waardebereik.

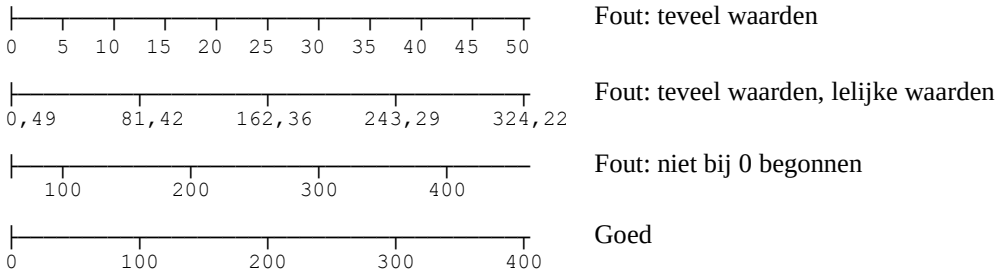
Als je het begin- en eindpunt van de as handmatig instelt, dan kun je meteen ook goed het nulpunt afhandelen. Als een variabele een natuurlijk nulpunt heeft, dan is het beter om de as bij dat nulpunt te laten beginnen. Beginnen bij een waarde groter dan 0 kan een vertekend beeld opleveren. De verschillen worden dan groter voorgesteld dan ze in werkelijkheid zijn.

De assen mogen niet onbenoemd blijven. Minimaal moeten je de namen van de variabelen bij de assen zetten. Maar die namen zijn vaak te kort en daarom niet duidelijk genoeg. De programmatuur voor het maken van spreidingsdiagrammen moet daarom de mogelijkheid bieden om langere teksten bij de vakken te zetten. Die teksten moeten ook melden in welke eenheden de variabelen zijn gemeten.

Naast informatie over de variabele moet ook het waardebereik van de variabelen goed zichtbaar zijn op de bijbehorende assen. Op een aantal plaatsen op de as moet je kleine dwarsstreepjes (*ticks*) zetten en bij deze ticks moet je de waarden van de variabele vermelden. Gebruik niet teveel ticks en waarden. Doe je dat wel dan raak je al gauw het overzicht kwijt. Gebruik ook niet te weinig ticks, want dat is weinig informatief.

Het is gangbaar om in ieder geval ticks bij het begin en het eind van de as te zetten. In de ruimte ertussen worden op gelijke afstanden nog een aantal andere ticks gezet. Dit kan onhandig zijn bij zelf-schalende programma's. Het is immers heel goed mogelijk, en zeer waarschijnlijk, dat met deze ticks geen ‘mooie’ waarden corresponderen. Cleveland en McGill (1984) merken op dat er geen enkele noodzaak is om de ticks op het hoekpunt te laten beginnen. In tegendeel, de ticks moeten overeenkomen met ‘mooie’ waarden. In de praktijk kun je het probleem van de mooie ticks oplossen met een programma dat een goed werken algoritme heeft om de mooie waarden te bepalen. Uiteraard kun je het ook handmatig doen. Figuur 2.4 geeft enkele voorbeelden van goede en slechte schaalverdelingen.

Figuur 2.4. Voorbeelden van het vermelden van ticks en schaalwaarden bij assen.



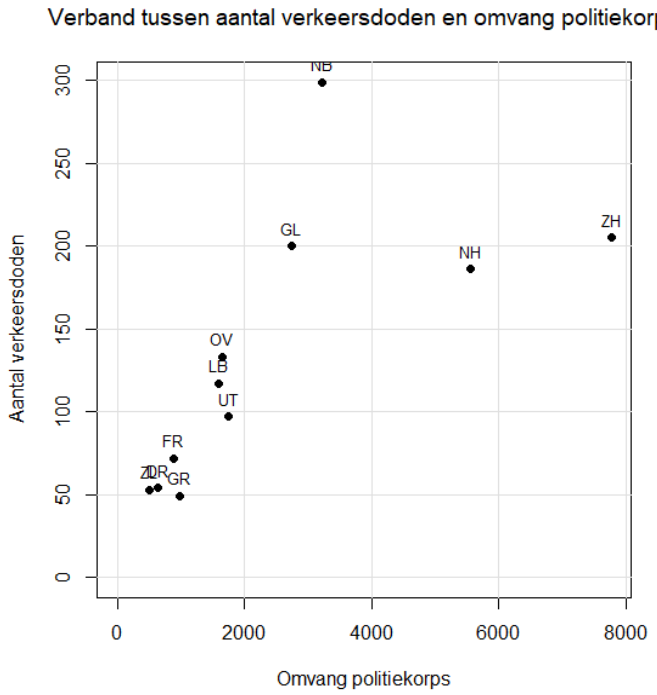
Vroeger, toen er nog geen computers beschikbaar waren, maakten we onze grafieken op grafiekenpapier. Dat is papier met een rooster van vakjes van 1 mm bij 1 mm. Dat is heel nuttig als je een grafiek met de hand wilt maken, maar het vertroebelt de presentatie en exploratie van de gegevens. Een rooster met zoveel horizontale en verticale en horizontale lijnen laat de punten in de puntenwolk te weinig naar voren springen. Het leidt af en bemoeilijkt daardoor de interpretatie van de grafiek.

Je kunt wel een bescheiden rooster gebruiken als hulpmiddel bij het aflezen van de waarden (*X*-waarden en *Y*-waarden) die horen bij de punten. Teken dan niet teveel roosterlijnen en maak ze grijs in plaats van zwart.

Tot slot komen de punten zelf ook nog even aan bod. Je zou natuurlijk punten kunnen zetten in de grafiek, maar wellicht is het beter om wat meer substantie aan te brengen door cirkeltjes, vierklantjes, driehoekjes te tekenen. Dat voorkomt bijvoorbeeld dat je het stof op het glas van de kopieermachine analyseert in plaats van de echte puntenwolk. Het gebruik van dit soort symbolen heeft bovendien het voordeel dat je verschillende groepen kunt onderscheiden door elke groep zijn eigen symbool te geven.

Figuur 2.5 toont een voorbeeld van een versie van de grafiek van het aantal verkeersdoden tegen de omvang van het politiekorps. De grafiek heeft ‘mooie’ assen. Er is een bescheiden rooster dat helpt bij het aflezen. En wat natuurlijk ook helpt bij het interpreteren van de grafiek is het zetten van namen bij de punten.

Figuur 2.5. Spreidingsdiagram van de omvang van het politiekorps tegen het aantal dodelijke ongelukken. Met mooie assen, roosterlijnen en labels. Bron CBS, 1987.



3. Overlappende punten

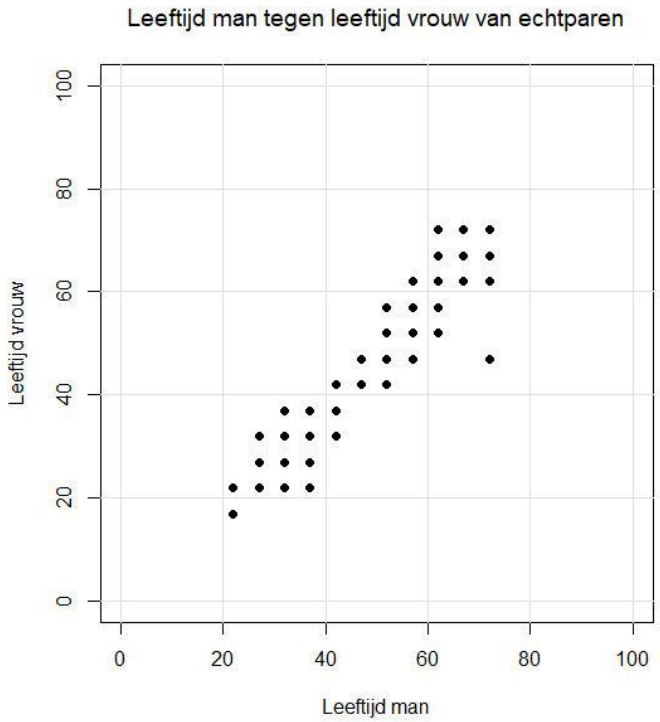
Alle gegevens moeten zichtbaar zijn in het spreidingsdiagram. Is dit niet het geval, dan kan een verkeerd beeld ontstaan waardoor je een verkeerde conclusie kunt trekken. In deze paragraaf bespreken we een van die problemen en dat is het probleem van de overlappende punten.

We spreken van *overlappende punten* als twee of meer punten dezelfde coördinaten hebben of als hun coördinaten zo dicht bij elkaar liggen dat het computerprogramma de punten op dezelfde locatie afbeeldt. Is dat erg? Jazeker, want overlappende punten zie je niet terug in de grafiek. Op het scherm of het papier staat een punt, maar je weet niet hoeveel waarden dat punt vertegenwoordigt. Dat kan ernstige gevolgen hebben voor de interpretatie van de puntenwolk. De dichtheid van een groep punten lijkt dan veel lager dan hij in werkelijkheid is.

We geven een voorbeeld. Als gegevens gebruiken we de leeftijd van 200 echtparen. De gegevens dateren uit de jaren 80 van de vorige eeuw. In het spreidingsdiagram zetten we voor elke echtpaar de leeftijd van de man af tegen de leeftijd van de vrouw. Zo op het oog (zie figuur 3.1) lijkt er een sterk verband te bestaan tussen de leeftijd van de man en de leeftijd van de vrouw. De puntenwolk heeft een mooie sigaarvorm. Dat betekent hier dat de man van een echtpaar ongeveer dezelfde leeftijd heeft als de vrouw. Er is sprake van één uitschieter: er is een echtpaar waarvan de man ruim 70 is en de vrouw onder de 50.

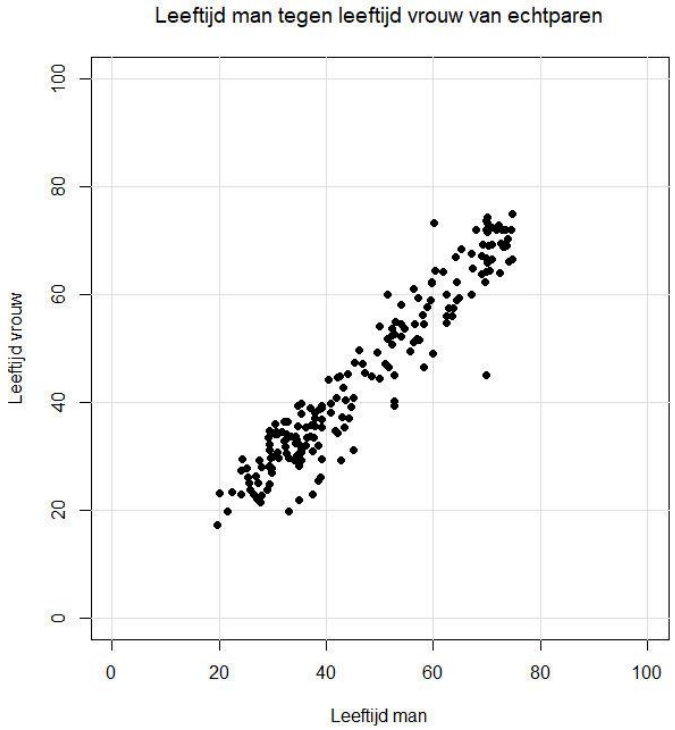
De leeftijden zijn afgerond op veelvouden van 5 jaar. Door dat afronden is er sprake van veel overlappende punten. Dat is goed te zien in figuur 3.1. In de grafiek zijn 38 punten te zien terwijl er 200 echtparen zijn. Dus zijn 162 punten onzichtbaar. Ze zijn verborgen achter andere punten.

Figuur 3.1. De leeftijd van mannen tegen die van vrouwen. Er zijn veel overlappende punten,



Er zijn verschillende manieren om de overlappende punten uit elkaar te trekken en zo zichtbaar te maken. We bespreken drie technieken om dat te doen: jitter, zeepbellen en zonnebloemen. We beginnen met *jitter*. Deze techniek komt erop neer dat we aan alle *X*-waarden en *Y*-waarden een klein beetje ruis toevoegen. Voor elke waarde doen we een trekking uit een homogene verdeling en de uitkomst tellen we op bij de waarde. Dat heeft tot gevolg dat overlappende punten niet meer overlappen. Ze liggen nu verspreid in een klein vierkantje.

Figuur 3.2. De leeftijd van mannen tegen die van vrouwen. Met 5% jitter



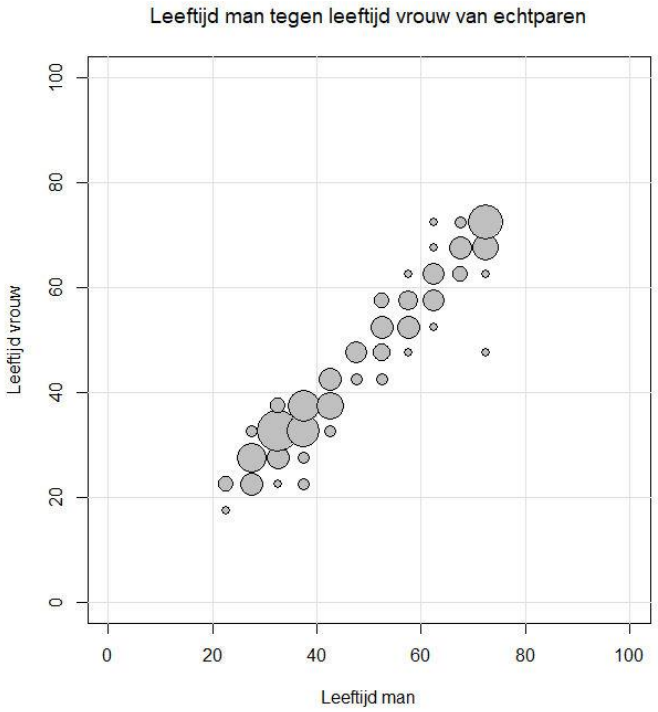
Figuur 3.2 toont een voorbeeld. In het spreidingsdiagram zijn dezelfde gegevens gebruikt als in figuur 3.1. Er zijn nu veel meer punten te zien. De samenhang lijkt nog sterker omdat er veel punten zijn die vlakbij de as van de sigaar liggen. Ook lijkt het erop dat de dichtheid van de punten linksonder groter is dan in de rest van de puntenwolk.

Hoe groot moet de omvang van de jitter zijn? Aan de ene kant moet hij niet te klein zijn. Voeg je te weinig jitter toe, dan heeft dit geen effect. Er zijn dan nog steeds overlappende punten. Het oppervlak van de jitter moet ook niet te groot zijn. Met teveel jitter tast je immers de structuur van de puntenwolk aan. Chambers et al. (1983) adviseren voor de jitter een diameter van 4% tot 10% van het waardebereik van de variabele. In geval van afgeronde gegevens zou je ook kunnen overwegen de diameter gelijk te nemen aan de lengte van het afrondingsinterval. Figuur 3.2 is het resultaat van het toevoegen van 5% jitter.

Een tweede manier om overlappende punten zichtbaar te maken, is het maken van *zeepbellen* (bubbles). Je verandert dan niets aan de coördinaten van de punten, maar brengt het aantal overlappende punten tot uiting in de omvang van het symbool. Je tekent bijvoorbeeld cirkels waarvan de oppervlakte evenredig is met het aantal overlappende punten (en dus de diameter evenredig is met de wortel uit het aantal overlappende punten).

Figuur 3.3 geeft een voorbeeld van het gebruik van zeepbellen. Dezelfde 200 echtparen zijn weer gebruikt. De grafiek bevat twee gebieden met een hoge concentratie van punten: linksonder en helemaal rechtsboven. De uitschieter blijft goed zichtbaar.

Figuur 3.3. De leeftijd van mannen tegen die van vrouwen. Met 5% jitter



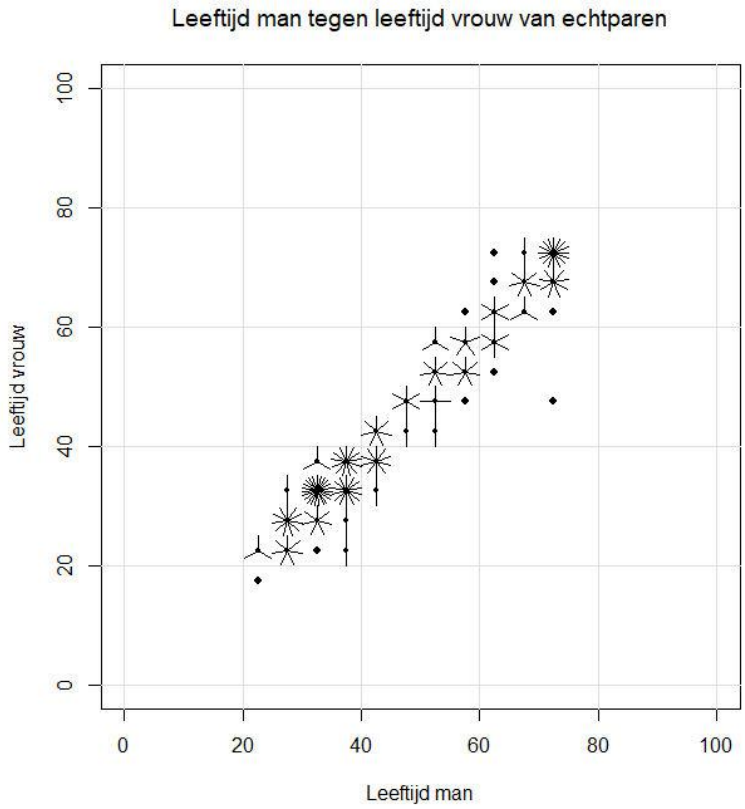
Zeepbellen zijn niet altijd de perfecte oplossing voor overlappende punten. Als de zeepbellen te groot zijn, gaan ze ook weer overlappen met andere zeepbellen. Als dit zich voordoet, dan lijkt de dichtheid lager dan hij in werkelijkheid is. In figuur 3.3 lijkt er sprake van een klein beetje overlap, maar die is niet zo erg dat dit kan leiden tot verkeerde interpretatie.

De derde manier om overlappende punten weer te geven is die van de *zonnebloemen* (*sunflowers*). Zie bijvoorbeeld Chambers et. al. (1983) en Cleveland & McGill (1984). We geven een los punt aan door

een punt. Zijn er overlappende punten, dan trekken we vanuit het desbetreffende punt streepjes, en wel evenveel streepjes als er overlappende punten zijn. Bij twee overlappende punten heeft de zonnebloem twee blaadjes, bij drie punten drie plaatjes, enz. Hoe groter het aantal overlappende punten, des te meer blaadjes de zonnebloem heeft, en des te groter de intensiteit is zoals het menselijk oog die waarneemt.

Figuur 3.4 bevat een voorbeeld van het gebruik van zonnebloemen. Dit spreidingsdiagram laat goed zien waar de dichtheid van de punten hoog en laag is. Ook de uitschieter blijft goed zichtbaar.

Figuur 3.4. De leeftijd van mannen tegen die van vrouwen. Met zonnebloemen.



Gezien de goede indruk die zonnebloemen geven van de dichtheidsverdeling van de punten, zijn er onderzoekers die zonnebloemen aanbrengen in hun spreidingsdiagram ook al zijn er helemaal geen overlappende punten. Ze brengen een denkbeeldig rooster van vakjes aan en tellen het aantal punten in elk vakje. Van die aantallen maken ze dan zonnebloemen.

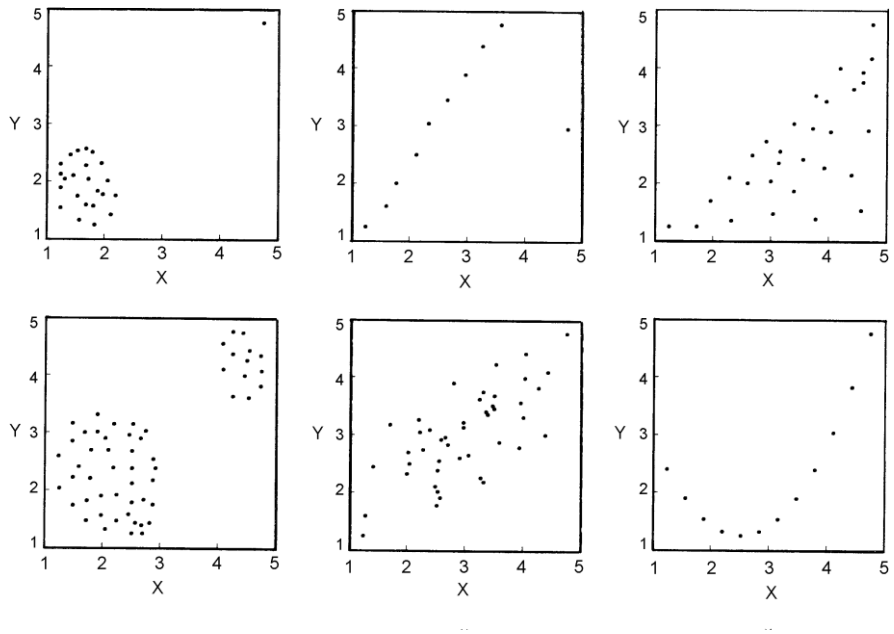
4. Op zoek naar samenhang

Een van de redenen om een spreidingsdiagram te maken is het doen van nader onderzoek naar een mogelijke samenhang tussen twee variabelen. Er bestaan ook numeriek technieken om op zoek te gaan naar samenhang. Een bekend numerieke grootheid is de correlatiecoëfficiënt. Die is echter alleen in staat om de sterkte van lineaire samenhang te meten. De correlatiecoëfficiënt is niet in staat om patronen of onverwachte problemen te signaleren. Daarom moet je je zoektocht naar bijzondere patronen en structuren nooit beginnen met het berekenen van de correlatiecoëfficiënt. Begin altijd met het bekijken van een spreidingsdiagram. Zie je daarin een mooi lineair verband, dan kun je de correlatiecoëfficiënt uitrekenen en daarmee vaststellen hoe sterk de samenhang is.

De *correlatiecoëfficiënt* neemt altijd een waarde aan tussen -1 en +1. Een waarde dichtbij -1 of +1 duidt op een bijzonder sterke lineaire samenhang. Voor een waarde in de buurt van +1 liggen de punten op een stijgende lijn. Een waarde dichtbij -1 geeft aan dat er sprake is van een dalende lijn.

De correlatiecoëfficiënt heeft dus zijn beperkingen als we hem gebruiken als instrument voor exploratief onderzoek. Figuur 4.1 geeft voorbeelden daarvan. De grafieken in deze figuur zijn ontleend aan Chambers et al. (1983). Ondanks dat de waarde van de correlatiecoëfficiënt in de zes grafieken steeds hetzelfde is (0,70), is de structuur van de samenhang steeds anders. In de grafiek linksboven is geen samenhang aanwezig. Door de aanwezigheid van een uitschieter vindt de correlatiecoëfficiënt wel samenhang. In de grafiek midden boven gebeurt het omgekeerde. Er is een perfecte samenhang. De correlatiecoëfficiënt zou dus eigenlijk 1 moeten zijn. Eén uitschieter verstoort de situatie echter. De grafiek rechtsonder vertoont ook een perfecte samenhang, maar aangezien die samenhang niet lineair is (maar kwadratisch), pakt de correlatiecoëfficiënt dat onvoldoende op. In de grafiek linksonder is er eigenlijk geen samenhang. Er is geen samenhang binnen de twee groepen, maar de relatieve ligging van de twee groepen ten opzichte van elkaar leidt tot een soort schijnbare samenhang. De grafiek midden onder is de enige met een echte lineaire samenhang. De puntenwolk heeft de vorm van een sigaar (een rechte lijn met ruis eromheen). De (lineaire) samenhang is niet zo heel sterk.

Figuur 4.1. Zes spreidingsdiagrammen met dezelfde correlatiecoëfficiënt van 0,70



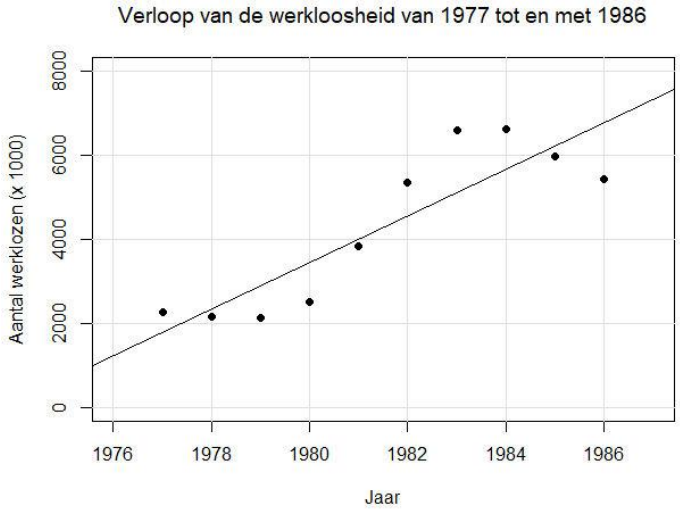
Uit deze reeks grafieken zal het duidelijk zijn dat de waarde van de correlatiecoëfficiënt weinig zegt over de structuur en de sterkte van de samenhang. Het is, zeker in een eerste verkennende fase van de analyse, beter om naar de puntenwolken te kijken.

Er zijn onderzoekers die er vanuit gaan dat de wereld lineair is, of op zijn minst bij benadering lineair. Als dit zo is, dan kun je de samenhang het beste numeriek samenvatten in de vorm van een regressielijn. Helaas is de praktijk vaak anders. De wereld is niet altijd lineair. En soms past een rechte lijn wel goed binnen bepaalde groepen, maar is de rechte lijn binnen de groepen steeds weer anders. Je krijgt dan een mix van verschillende lijnen.

Figuur 4.2. toont een puntenwolk waarin wel sprake is van een sterke samenhang, maar niet van een lineaire samenhang. Het betreft een weergave van het verloop van de werkloosheid vanaf 1977 tot en met 1986. Het is duidelijk dat de samenhang niet lineair is. Er is gepoogd een rechte lijn door de

puntenwolk te tekenen die de punten zo goed mogelijk volgt. Dit is de regressielijn. De conclusie moet dat deze lijn geen goede beschrijving van de samenhang is.

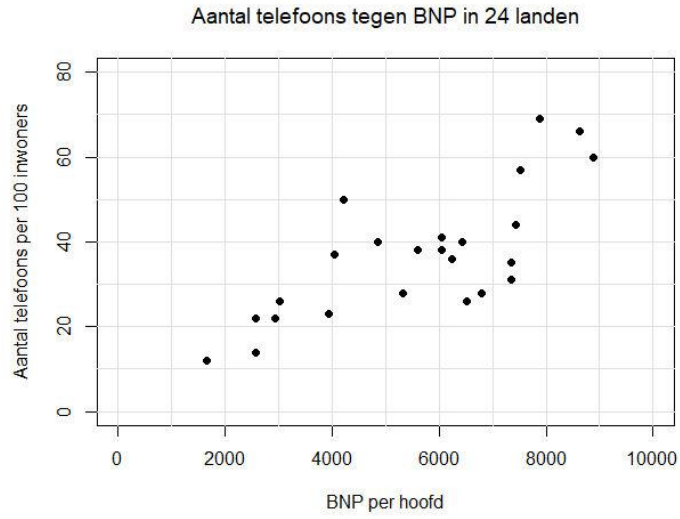
Figuur 4.2. Een voorbeeld van niet-lineaire samenhang



De grafiek in figuur 4.2 is niet vierkant maar breder-dan-hoog. De Gouden Snede is toegepast. Er is hier sprake van een verklarende variabele (jaar) en een responsvariabele (aantal werklozen). Daarvoor zou de voorkeur inderdaad moeten uitgaan naar een breder-dan-hoog grafiek.

Als het niet zinvol is om een regressielijn te tekenen (omdat de samenhang niet lineair lijkt te zijn), hoe moeten we dan de samenhang in beeld brengen? Neem als voorbeeld de puntenwolk in figuur 4.3. Die toont voor 24 westerse landen de samenhang tussen het aantal telefoons per 100 inwoners (*Y*-variabele) en het bruto nationaal product per hoofd van de bevolking (*X*-variabele).

Figuur 4.3. Aantal telefoons tegen BNP in 24 westerse landen



Zo op het oog lijk je het verband tussen beide variabelen in figuur 4.3 niet door een rechte lijn te kunnen beschrijven. Maar hoe moet je het dan wel doen? We werken in deze paragraaf drie aanpakken wat nader uit. Twee ervan zijn gebaseerd op zogenaamde *kruismedianen* en derde maakt gebruik van *lokaal gewogen regressieanalyse* (LOWESS).

Kruismedianen zijn geïntroduceerd door Tukey (1977). Uitgangspunt is dat de samenhang zich het beste in beeld laat brengen door het bepalen van een *functie* die zo goed mogelijk de punten in de puntenwolk volgt. Enerzijds moet de bijbehorende kromme zo dicht mogelijk door alle punten gaan en anderzijds moeten uitschieters de vorm van de kromme niet te veel beïnvloeden. *Uitschieters* zijn waarden die buiten het normale patroon van de waarnemingen liggen en daarom weinig of niets zeggen over de structuur van de samenhang.

Functies zijn een nuttig hulpmiddel, maar ze hebben wel de beperking dat er bij één *X*-waarde maar één *Y*-waarde hoort. De punten in de puntenwolk hoeven zich daar natuurlijk niet aan te houden. Het is best mogelijk dat in de puntenwolk een aantal punten boven elkaar ligt. Tukey (1977) lost dit probleem op door de puntenwolk te verdelen in een aantal verticale banden (*strippen*). In elke strip bepalen je de mediaan van de *X*-waarden en de *Y*-waarden. De combinatie van een *X*-mediaan en een *Y*-mediaan noemt Tukey (1977) een *kruismediaan*. Voor het in beeld brengen van de samenhang gebruik je vervolgens alleen de kruismedianen en niet de oorspronkelijke gegevens.

De kruismedianen zijn gebaseerd op verticale strips. Het had natuurlijk ook gekund met horizontale strips. Hier weerspiegelt zich eenzelfde vorm van asymmetrie tussen de *X*-variabele en de *Y*-variabele als in een regressieanalyse: we verklaren het gedrag van de *Y*-variabele uit dat van de *X*-variabele, en niet omgekeerd. Bij kruismedianen is het net zo. Daarvoor zijn verticale strips beter geschikt.

Een punt van aandacht is het aantal strippen dat je gebruikt in je puntenwolk. Je kunt zelf het aantal strippen bepalen en daarmee hoeveel punten je in elke strip opneemt. Kies je voor veel strippen, dan bevat elke strip maar weinig punten. De locaties van de kruismedianen zijn dan afhankelijk van maar enkele waarden. Dat zal ertoe leiden dat het gevonden verband nog veel ruis bevat. Gebruik je maar weinig strippen, dan zullen de kruismedianen veel robuuster zijn. Het is de vraag of de functie die je zo vindt nog wel veel informatie verschaft over de samenhang. Toepassing van weinig strippen zal vaak leiden tot een te ruwe beschrijving van de samenhang. Kortom, je zult een middenweg moet vinden tussen te weinig strippen en teveel strippen. Het is verstandig om verschillende aantallen strippen uit te proberen.

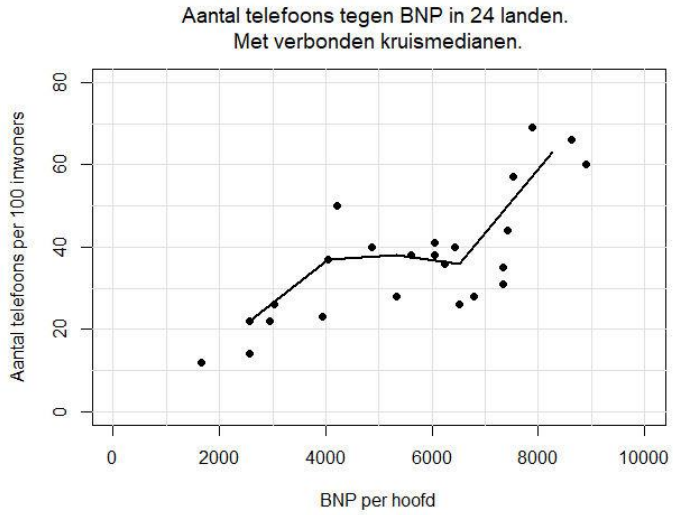
Hebben we eenmaal de kruismedianen berekend, dan moeten we gaan bedenken wat voor soort functie door deze punten moet lopen. We bespreken in het kort twee verschillende aanpakken. Het eerste voorstel is om de kruismedianen te verbinden door rechte lijnen. Het zo ontstane reeks verbonden lijnstukken geeft een simpel beeld van de samenhang. Door medianen te gebruiken in plaats van gemiddelden, is het gevonden patroon niet gevoelig voor uitschieters. Wel vertoont de reeks verbonden lijnstukken een hoekig patroon en het daarom zal het in veel gevallen niet een realistisch beeld van de werkelijke samenhang geven. Niettemin zou het als een soort eerste orde benadering zijn diensten kunnen bewijzen. Het kan bijvoorbeeld heel goed knikken detecteren, waar een gewone regressieanalyse geen raad mee weet.

Figuur 4.4 bevat een voorbeeld van verbonden kruismedianen. Uitgangspunt was de grafiek in figuur 4.3. De puntenwolk is hier verdeeld in vijf strips. Duidelijk is te zien dat de lijnen in de strippen allemaal een andere richting hebben. Opvallend is ook dat één van de lijnen horizontaal loopt Het is zeker niet verantwoord de samenhang in de grafiek samen te vatten in de vorm van een regressielijn.

De verbonden kruismedianen geven enig inzicht in de vorm van de samenhang. Het is echter onwaarschijnlijk dat deze hoekige reeks van lijnstukken de werkelijke samenhang weergeeft. Meestal zitten er niet van die scherpe knikken in de samenhang. Vaak is het niet onredelijk te veronderstellen dat de functie die de samenhang weergeeft een continu verloop heeft. In dat geval ligt het voor de hand om door de kruismedianen een functie met een continue verloop te trekken. De vraag is alleen: welke functie? Als we *n* punten tot onze beschikking hebben, dan kunnen we daardoor altijd een *n*-de graads

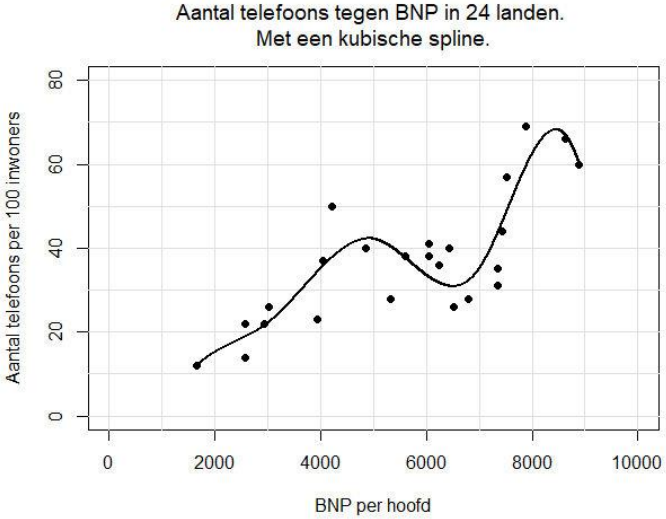
polynoom trekken. Voor puntenwolken met veel punten kan dit wel eens onrealistische beeld opleveren. Bovendien is de vorm van de kromme dan erg gevoelig voor kleine veranderingen in de *Y*-waarden.

Figuur 4.4. Aantal telefoons tegen BNP in 24 westerse landen



Om dit probleem te vermijden, kunnen we gebruik maken van *splines*. De naam spline komt van een instrument dat tekenaars vroeger gebruikten om vloeiende krommen te tekenen. Het is een lang, dun en buigzaam latje. Door het plaatsen van gewichten op bepaalde plaatsen op het latje, kun het latje de gewenste kromming geven. Bij toepassing van splines in puntenwolken verbindt je elke tweetal opeenvolgende punten door een derdegraads polynoom. Vanwege het derdegraads karakter spreken we ook wel van een *kubische spline*. We plakken al die stukjes derdegraads polynoom aan elkaar tot één vloeiende kromme. Je kunt door twee punten een heleboel verschillende derdegraads polynomen trekken. De vrijheid die we hebben bij de keuze van deze polynomen gebruiken we om restricties op te leggen die er voor zorgen dat de overgang van het ene stukje polynoom naar het volgende 'gladjes' verloopt. Dat doen we door te eisen dat de eerste orde en tweede orde afgeleiden van de polynomen in het ‘knooppunt’ aan elkaar gelijk moeten zijn. Aan de uiteinden zijn geen knooppunten. Daar eisen we dat de tweede orde afgeleiden 0 moeten zijn. Aldus ontstaat een vloeiende curve die nergens sprongen of knikken vertoont.

Figuur 4.5. Aantal telefoons tegen BNP in 24 westerse landen



Figuur 4.5 bevat een voorbeeld van het gebruik van splines. Eerst is de puntenwolk in vijf strippen verdeeld. In die strippen zijn de kruismedianen berekend. Vervolgens is een kubische spline getrokken door de vijf kruismedianen. Ook hier is duidelijk te zien dat we de samenhang niet kunnen beschrijven door een rechte lijn. Er is sprake van een flinke 'dip' in de buurt van een BNP van 6500. Ook aan het rechter uiteinde zit er een bocht in de trend. Deze 'bocht' was niet te zien bij de verbonden kruismedianen in figuur 4.4.

Er zijn nog andere manieren om de samenhang in een puntenwolk te onderzoeken. We verwijzen daarvoor naar Tukey (1977). Een laatste techniek die we hier nog behandeld, is gebaseerd op standaard regressieanalyse, maar niet een regressieanalyse die noodzakelijkerwijs een rechte lijn op moet leveren. De techniek heet *locally weighted regression scatterplot smoothing* (LOWESS). Hij is ontwikkeld door Cleveland (1979). Korte beschrijvingen van de techniek zijn te vinden in Chambers et. al. (1983) en Cleveland en McGlll (1984).

Bij lokale gewogen regressie bepalen we voor elk punt (X_i, Y_i) in de puntenwolk een regressielijn bepaald. We gaan er vanuit dat de X-waarden $X_1, X_2, \ldots, X_n$ zijn geordend in oplopende grootte. Voor elke waarde X_i bepalen we een omgeving van m waarden X_j (inclusief X_i) die het dichtst bij X_i liggen. Meestal is voor m een bepaalde fractie f van het totaal aantal punten n gekozen ($m = fn$, afgerond). Op de m waarnemingsparen (X_j , Y_j) voeren we een *gewogen regressieanalyse* uit. Het gewicht van een punt (X_j ,Y_j) is gebaseerd op de afstand D_j van X_j tot X_i . Laat D_{max} de grootste afstand zijn die voorkomt onder de m punten. Dan is het gewicht van (X_j ,Y_j) gelijk aan

$$W_j = \left(1 - \left|\frac{D_j}{D_{\max}}\right|^3\right)^3 .$$

Met de gewichten W_j voeren we een gewogen regressieanalyse uit. Vervolgens rekenen we de residuen R_j uit. Om de invloed van uitschieters zo klein mogelijk te houden, maken we nieuwe gewichten. Zij R_{med} de mediaan van de absolute waarden van de residuen. We bepalen vervolgens factoren V_j met de formule

$$V_j = \left(1 - \left|\frac{R_j}{R_{\max}}\right|^2\right)^2 .$$

We maken nu gewichten W_j' door de oude gewichten W_j te vermenigvuldigen met de factoren Vj:

$W_j' = W_j V_j$.

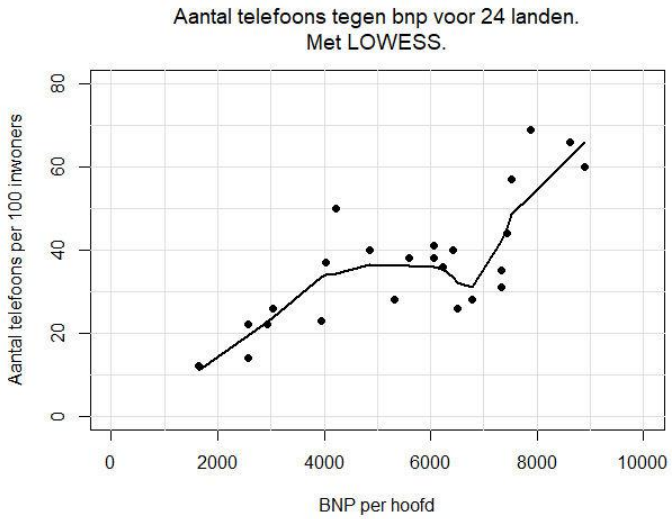
Opnieuw voeren we een gewogen regressieanalyse uit. Weer bepalen we de residuen en weer passen we de gewichten aan op basis van deze residuen. En voor de derde maal voeren we een regressieanalyse uit. Zo krijgen we uiteindelijk een regressielijn die is gebaseerd is op punten in de omgeving van het punt (X_i, Y_i) en waarop veraf gelegen punten en uitschieters weinig invloed hebben gehad. Deze regressie is dus een beschrijving van de samenhang in de buurt van het punt (X_i, Y_i). De lijn die de samenhang beschrijft in de puntenwolk moet nu in de waarde X_i gaan door de door de regressielijn voorspelde waarde $\hat{Y}_i$.

Hierboven is beschreven hoe je voor één punt (X_i, Y_i) de waarde $\hat{Y}_i$ van de lijn bepaald. Deze procedure moet je nu herhalen voor elk van de n punten in de puntenwolk. Zo ontstaat n punten (X_i, $\hat{Y}_i$) en die punten verbind je door lijnstukken.

Figuur 4.6 toont een voorbeeld van een grafiek die met LOWESS is gemaakt. Aan de linkerkant zien je een groep punten die je wel redelijk met een simpele regressielijn kunt beschrijven, maar in het

rechterdeel lukt dit niet. Voor een deel van de puntenwolk zou je misschien een spline kunnen gebruiken, maar voor de knik zal dit waarschijnlijk niet werken. In dit voorbeeld moeten we concluderen dat LOWESS hier het beste werkt. Deze techniek is goed in staat de knik mee te nemen.

Figuur 4.6. Aantal telefoons tegen BNP in 24 westerse landen



Chambers et al (1983) raden aan de fractie punten *f* voor de bepaling van de lokale regressielijn te nemen tussen de 1/3 en 2/3 te nemen. In het voorbeeld was *f* gelijk aan 0,40.

Uit bovenstaande beschrijving blijkt wel dat de LOWESS-procedure zeer rekenintensief is. Voor elk punt in de puntenwolk moet je drie regressie-berekeningen uitvoeren. Dit wordt erger naarmate *f* groter is. Je zou eventueel kunnen overwegen de procedure iets minder robuust te maken door de derde lokale regressie-berekening per punt achterwege te laten, maar dan moet je wel rekening houden met een iets grotere invloed van uitbijters.

Referenties

Bergami, D, (1969), Mathematics. *Time-Life Science Library*, Time-Life Book, Netherlands.

Bethlehem, J.G. (1987), *Het spreidingsdiagram opnieuw bekeken*. Rapport 4436-87-M1, Centraal Bureau voor de Statistiek, Voorburg.

Bethlehem, J.G. (2018), *Understanding Public Opinion Polls*. CRC Press, Boca Raton, FL.

Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Wadsworth, Belmont, CA.

Cleveland, W.S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74, blz. 829-836.

Cleveland, W.S. & McGlll, R. (1984), The Many Faces of a Scatterplot. *Journal of the American Statistical Association* 79, blz. 807-822.

Kendall, M.G. & Buckland, W.R. (1960), *A Dictionary of Statistical Terms*. Oliver and Boyd, London.

Playfair, J. (1786), The Commercial and Political Atlas. London.

Schmid, C.F. (1983), Statistical Graphics, Design Principles and Practices. Wiley and Sons, New York.

Tufte, E. (1983), The Visual Display of Quantitative Information. Graphics Press, Chesire, CT.

Tukey, J.W. (1977), Exploratory Data Analysis. Addison-Wesley, Reading, MA.