Handboek Statistische Beveiliging

Info voor externen

2e editie

Juni 2024

CBS Den Haag
Henri Faasdreef 312
2492 JP Den Haag
Postbus 24500
2490 HA Den Haag
+31 70 337 38 00

www.cbs.nl

Juni 2024

Inhoudsopgave

Lijst van Quick References

1. Inleiding

Op 19 december 2005 heeft het directieberaad van het CBS het officiële beleid op het gebied van statistische beveiliging bekrachtigd, door het accorderen van het Handboek Statistische Beveiliging. Dat handboek bevatte de belangrijkste beleidsregels voor statistische beveiliging van de verschillende soorten statistische producten van het CBS. Het huidige rapport is een herschreven versie van het oorspronkelijke Handboek Statistische Beveiliging.

De huidige versie is gemaakt om in te spelen op de vraag vanuit de werkvloer voor een beter leesbare versie. Daarmee is de huidige versie inhoudelijk gelijk aan de versie van 19 december 2005 en vertegenwoordigt daarmee nog steeds het door het directieberaad bekrachtigde beleid.

Bij de bespreking van de wettelijke kaders is in de huidige versie uiteraard wel uitgegaan van de meer recente wettelijke bepalingen.

1.1 Waar gaat dit rapport over?

Dit rapport geeft de beveiligings*regels* waaraan alle CBS-output in principe moet voldoen om gepubliceerd te kunnen worden. Die *regels* bepalen of een CBS-output, voor zover het statistische beveiliging betreft, gepubliceerd kan worden. Wanneer output niet aan de *regels* voldoet, kunnen er *methoden* gebruikt worden om te proberen de CBS-output alsnog geschikt te maken voor publicatie. De mogelijke *methoden* staan beschreven in de **Methodenreeks - Statistische beveiliging** en worden in dit rapport verder niet behandeld.

1.2 Achtergrond

Bij het publiceren van statistische informatie moet het CBS een afweging maken tussen de belangen van zijn berichtgevers en houders van registraties, de privacy van personen, bedrijven en instellingen aan de ene kant en de behoeften van zijn gebruikers aan de andere kant. Gebruikers willen van het CBS zo veel en zo gedetailleerd mogelijke informatie. De wetgever, berichtgevers, bronhouders, publiek en instituten als de Autoriteit Persoonsgegevens (AP) eisen dat de privacy wordt gewaarborgd.

Wat het CBS wèl en niet mag publiceren, moet onder andere volgen uit het statistische beveiligingsbeleid van het CBS zelf. De grondregel van statistische beveiliging zegt dat we moeten voorkómen dat er conclusies over herkenbare eenheden kunnen worden getrokken op basis van CBS-output. Uit de statistische publicaties van het CBS (StatLine-tabellen, open data, nieuwsberichten, wetenschappelijke artikelen, AI-toepassingen, algoritmes, ASD publicaties e.d.) mogen zulke conclusies niet getrokken kunnen worden. Maar ook als het CBS microdata beschikbaar stelt voor wetenschappelijke analyse, moet deze grondregel van de statistiek overeind blijven. Om bij individuele publicaties goede keuzes te kunnen maken is het belangrijk dat dit beleid duidelijk beschreven staat, en breed en gezaghebbend gedragen wordt binnen het CBS.

De technische termen in dit handboek worden niet allemaal in detail uitgelegd; soms wordt verwezen naar meer technische rapporten. Met dit rapport in de hand is het voor de meeste CBS-output in grote lijnen duidelijk hoe de statistische beveiliging gedaan moet worden. Het beleid in dit rapport laat wel enige vrijheid toe: het is wenselijk om per statistiek enige ruimte te hebben bij de statistische beveiliging, bijvoorbeeld bij het instellen van de parameters van de verschillende regels (binnen een bandbreedte).

Bovendien is het praktisch gezien onmogelijk om van te voren alle mogelijke situaties te bedenken en te beschrijven.

Voor interpretatie en uitwerking van het statistische beveiligingsbeleid kent het CBS dan ook experts en een expertgroep.

1.3 Belang

Naarmate het CBS zijn gegevens meer en meer ontleent aan de grote administraties in Nederland, wordt het issue van de statistische beveiliging nog belangrijker. Dat één individuele respondent besluit niet meer deel te nemen aan CBS onderzoeken omdat de statistische beveiliging niet op orde is, valt wellicht niet zo gauw op. Maar een administratieve bron, met gegevens over duizenden tot miljoenen statistische eenheden, mag voor het CBS nooit “droog vallen”.

Twee mechanismes die traditioneel indirect aan de bescherming van individuele gegevens bijdroegen, zijn niet meer van toepassing bij gebruik van administraties en registraties: zowel de steekproeftrekking als het interviewproces brengen marges en ruis tot stand die zich bij gebruik van integrale administraties niet voordoen. Daarmee wil overigens niet gezegd zijn dat administraties foutloos zouden zijn en altijd tot de perfecte onthulling leiden.

Statistische beveiliging is geen statisch vak. Integendeel. Nieuwe methodologieën, nieuwe databronnen en nieuwe vormen van publicaties vragen om nieuwe statistische beveiligingstechnieken. Ook de maatschappelijke perceptie van privacy bescherming en de toenemende hoeveelheid beschikbare “open data” vraagt om een continue ontwikkeling van het vakgebied van de statistische beveiliging.

De expertgroep “Statistische Beveiliging” zal in de toekomst beschikbaar blijven als klankbordgroep voor het beantwoorden van verdere vragen.

1.4 Leeswijzer

Zonder de grondbeginselen en de basisbegrippen van statistische beveiliging te kennen, is het moeilijk om de inhoud van dit handboek goed te kunnen plaatsen. In hoofdstuk 2 zullen we daarom enkele grondbeginselen van de statistische beveiliging bespreken en enkele basisbegrippen definiëren die we in de overige hoofdstukken van het handboek zullen gebruiken.
Het statistische beveiligingsbeleid van het CBS is gebaseerd op de wetgeving op dit terrein (zie hoofdstuk 3). Deze wetten moeten in verschillende situaties nader geoperationaliseerd en geïmplementeerd worden. In dit hoofdstuk staat het “waarom” van de statistische beveiliging centraal.
In de uitwerking in de praktijk heeft het CBS een eigen verantwoordelijkheid en vrijheid. Die moet ingevuld worden met regels, om willekeur te voorkomen. En die regels worden bij voorkeur gebaseerd op *best practices*. Dat de regels vervolgens softwarematig ondersteund en geïmplementeerd worden, spreekt vanzelf. De software pakketten μ- en τ-ARGUS zijn hiervoor een internationale standaard. Deze pakketten zijn in Europees verband ontwikkeld, waarbij het CBS altijd een leidende rol heeft gespeeld. De nog steeds leidende rol van het CBS in diverse Europese projecten op het gebied van de statistische beveiliging ondersteunt en onderschrijft dat.

In de hoofdstukken 4 tot en met 7 gaat het niet over het “waarom” maar over het “hoe” van de statistische beveiliging. We beschrijven dan de statistische beveiligingsregels voor de statistische aggregaten en microdata die het CBS verlaten.

De aggregaten kunnen drie vormen aannemen:

- Kwantitatieve tabellen;
- Frequentietabellen;
- Statistische analyses.

Kwantitatieve tabellen komen het meest in de economische statistieken voor. Voor een aantal achtergrondgegevens worden totale (of gemiddelde) productie-, omzet-, handel-, winst- en andere gegevens samengevat. Frequentietabellen komen vooral in de sociale statistieken voor. Voor de combinatie van een aantal (achtergrond- en doel-)kenmerken wordt bijvoorbeeld bepaald hoeveel mensen eraan voldoen. De voor de statistische beveiliging relevante verschillen tussen beide soorten tabellen vloeien mede voort uit de verschillen tussen de betreffende populaties. De bedrijvenpopulatie is bijvoorbeeld kleiner en naar belangrijke kenmerken veel schever verdeeld dan de populatie van personen. Grote bedrijven worden (daarom) integraal waargenomen. In hun economische belangen zijn zij bovendien kwetsbaarder dan de gemiddelde persoon. Bedrijfsspionage is een realiteit. Aan de andere kant kunnen groepen burgers als zodanig bijzonder kwetsbaar zijn. Begrippen als discriminatie, stigmatisering en groepsprivacy duiden daarop.

Het statistisch beveiligen van uitkomsten van algemene statistische analyses (de “output controle”) speelt vooral een rol bij de Remote Access (RA)-faciliteit van Microdataservices van het CBS en bij Aanvullende Statistische Diensten (ASD). De grote hoeveelheid aan potentieel bruikbare analyse technieken maakt de outputcontrole een ingewikkelde zaak. In het laatste hoofdstuk van dit handboek wordt kort ingegaan op de “outputcontrole” van statistische analyses bij RA en ASD.

2. Grondbeginselen en basisbegrippen

2.1 Inleiding

Om de inhoud van dit handboek goed te kunnen plaatsen, is het nodig dat de grondbeginselen en de basisbegrippen van statistische beveiliging bekend zijn bij de lezer. In dit hoofdstuk geven we een overzicht van statistische beveiliging zonder direct in de meer technische details te treden.
Als eerste gaan we in op de aard van de problematiek van statistische beveiliging. Op hoog niveau spelen bij microdata en bij geaggregeerde data dezelfde problemen een rol. De praktische invulling verschilt echter wel per type output.
Ten slotte geven we een overzicht van de belangrijkste, binnen de statistische beveiliging vaak gebruikte termen.

2.2 Aard van de problematiek van statistische beveiliging

Centraal bij statistische beveiliging staat het wel of niet kunnen afleiden van een herkenbaar gegeven over afzonderlijke personen, huishoudens, ondernemingen of instellingen. In het vervolg van dit hoofdstuk zullen we term “eenheid” gebruiken als we verwijzen naar een persoon, huishouden, onderneming of instelling.
Artikel 37 van de CBS-wet zegt onder andere dat het CBS gegevens slechts zodanig openbaar mag maken dat daaraan geen herkenbare gegevens over een afzonderlijke eenheid kunnen worden ontleend. Voor meer informatie over de wettelijke aspecten, zie hoofdstuk 3.
Binnen de statistische beveiliging wordt vaak gesproken over **onthullen**. Onder onthullen wordt dan verstaan dat iemand op basis van de door het CBS gepubliceerde gegevens in staat is een gegeven van een identificeerbare, afzonderlijke eenheid te herleiden. De term “onthullen” heeft gevoelsmatig te maken met het idee dat er informatie vrijkomt die op dat moment nog niet bekend was voor de onthuller. Voor de strikte interpretatie van de CBS-wet en de AVG is het echter niet van belang of de vrijkomende informatie nieuw is voor de onthuller of niet.

2.2.1 Soorten gegevens

Bij statistische beveiliging maken we onderscheid tussen twee soorten gegevens:

1. Gegevens die gebruikt kunnen worden om een afzonderlijke eenheid te identificeren: **identificerende** gegevens;
2. Gegevens die iets zeggen over een afzonderlijke eenheid: **overige** of **statistische** gegevens.

Bij de onder 1. genoemde gegevens horen bijvoorbeeld BSN, naam en opleiding of gemeente van vestiging en grootteklasse. Bij de tweede soort gegevens kan het gaan om zaken als politieke gezindte, inkomen en medicijngebruik of omzet en uitgaven aan investeringen. Doelbewust zijn hier voorbeelden aangehaald van zowel personen als bedrijven.
De interpretatie van wat bij anderen als bekend mag worden verondersteld is niet altijd even gemakkelijk en zal per situatie verschillen. Statistische beveiliging is dan ook niet uitsluitend een theoretische activiteit. Contact houden met de praktijk is essentieel.
Een gegeven kan zowel identificerend als statistisch zijn. Statistische gegevens worden soms ook wel suggestief “gevoelige” gegevens genoemd. Gevoelig wijst in de richting van privacy. Een gegeven als medicijngebruik past hier zeker in. Maar van het

jaarkilometrage van de gezinsauto is de mate van gevoeligheid al veel minder duidelijk. Toch moet het CBS ook met dat gegeven vertrouwelijk omgaan. Om meningsverschillen over al dan niet gevoeligheid van bepaalde gegevens te voorkomen, houdt het CBS de richtlijn aan dat uit de statistische informatieverstrekking *geen enkel* tot een specifieke eenheid herleidbaar gegeven mag blijken. Daarbij speelt het wel of niet algemeen bekend of beschikbaar zijn van de gegevens in principe geen rol.
Een belangrijke eigenschap van **identificerende** variabelen is dat de score op dergelijke variabelen op ruimere schaal bij anderen bekend is, of eenvoudig door anderen te achterhalen is. Daarnaast kunnen identificerende variabelen verschillen in de mate waarin ze eenheden uniek vastleggen. Het meest specifiek zijn variabelen als naam en adres, of een nummer waaronder een eenheid in één of meerdere registers is opgenomen. Die scores maken eenheden eenduidig herkenbaar en de variabelen worden dan ook **direct identificerend** genoemd. Voor *statistisch* onderzoek zijn deze variabelen niet nodig en worden bij gegevensverstrekking dan ook altijd weggelaten. Identificerende variabelen waarvan de scores unieke eenheden niet eenduidig herkenbaar maken, worden **indirect identificerend** genoemd. Denk daarbij aan variabelen als geslacht, leeftijd, woonregio, opleiding of beroep. Vaak zijn deze variabelen nodig voor het uitvoeren van statistisch onderzoek en kunnen ze dus niet zondermeer worden weggelaten. Hoewel de score op één indirect identificerende variabele dus meestal niet tot een herkenbare unieke eenheid leidt, kunnen combinaties van scores op verschillende indirect identificerende variabelen dat soms echter wel. Zo is de score “burgemeester” voor beroep op zichzelf niet identificerend (er zijn ruim 300 burgemeesters in Nederland) maar in combinatie met werkregio wel (er is maar één burgemeester van Amsterdam). Dat zal dus meegenomen moeten worden bij het bepalen van de benodigde statistische beveiliging.

Merk op dat het CBS nooit op naam gestelde, of gemakkelijk identificeerbare, openbaar toegankelijke gegevens aan derden zal verstrekken. Voorbeelden zijn hypotheekgegevens die door iedereen bij het Kadaster zijn op te vragen, en gegevens die in jaarverslagen van bedrijven of ondernemingen worden gepubliceerd. Het feit dat dergelijke informatie “toch al” beschikbaar is, ontslaat het CBS niet van de plicht om vertrouwelijk om te gaan met aan het CBS verstrekte gegevens, conform artikel 37 van de CBS-wet.

2.2.2 Manieren van onthullen

Onthulling kan in de praktijk op vele manieren plaatsvinden. Zowel in de literatuur als in de praktijk richt men zich meestal op **onthulling via identificatie**. Daarbij vindt onthulling plaats via het op eenduidige wijze vaststellen van de identiteit van de eenheid waarop de informatie betrekking heeft. De combinatie van de scores op (indirect) identificerende variabelen is dan uniek voor de betrokken eenheid: geen enkele andere eenheid heeft diezelfde combinatie. De bijbehorende statistische informatie is dan dus onthuld voor die herleidbare eenheid.
Naast deze manier van onthullen kan ook **groepsonthulling** leiden tot het bekend raken van statistische informatie over een individuele eenheid. Hierbij kan men denken aan de situatie waar meerdere eenheden hetzelfde scoren op een combinatie van (indirect) identificerende variabelen. De identificerende variabelen leiden in dat geval dus niet tot één unieke, herleidbare eenheid, maar tot een herleidbare groep eenheden. Als alle eenheden in die groep dezelfde score op een statistische variabele hebben, dan onthul

je die score voor iedere eenheid binnen de identificeerbare groep. Je zou dat stigmatiserend kunnen noemen, met name als de onthulling een grote impact heeft voor die groep eenheden. Denk bijvoorbeeld aan een bewering over crimineel gedrag van jongeren met een migratieachtergrond uit de Bijlmermeer of over het ontduiken van belasting- en premieafdracht in de bouw.
In zekere zin zou je het publiceren van statistische informatie kunnen zien als groepsonthulling. Statistiek gaat immers per definitie over groepen van eenheden. In hoofdstuk 5 gaan we in op de praktische invulling van groepsonthulling versus statistiek.

2.2.3 Soorten onthulling

Naast verschillende manieren waarop onthulling plaats kan vinden ("hoe"), kennen we binnen de statistische beveiliging ook verschillende soorten onthulling ("wat"). De meest bekende soorten zijn **identiteit-onthulling** (identity disclosure) en **attribuut-onthulling** (attribute disclosure). Bij identiteit-onthulling wordt de identiteit van een eenheid onthult op basis van een statistische publicatie. Het is dan dus mogelijk om een herkenbare eenheid aan te wijzen in de publicatie. Zoals in de vorige subparagraaf beschreven, is dat vaak de start voor het onthullen van een statistisch gegeven (attribuut-onthulling) van een herkenbare eenheid (onthulling via identificatie).
Soms is het echter ook mogelijk om een eigenschap van een herkenbare eenheid te bepalen, zonder eerst de eenheid in de publicatie aan te kunnen wijzen. De identificatie (herkenbaarheid) ligt dan feitelijk *buiten* de publicatie. Met name bij groepsonthulling speelt deze vorm van onthulling een rol. Immers, je kunt dan een specifieke eenheid slechts identificeren tot op een herkenbare groep, maar je onthult wel een eigenschap van die eenheid. Als men weet dat een specifieke eenheid tot die groep behoort, treedt dus attribuut-onthulling op voor die herkenbare eenheid. Als voorbeeld, stel dat 95% van een herkenbare groep een bepaalde eigenschap heeft. Als men iemand kent die tot de herkenbare groep behoort (herkenbaarheid *buiten* de publicatie), dan is de kans heel groot dat die persoon ook die eigenschap heeft.

2.2.4 Onthulling bij fouten in de data

Bij statistische beveiliging gaat men er meestal vanuit dat de desbetreffende data en informatie foutloos zijn. Paass (1988) is een van de weinigen die de mogelijkheid van fouten in statistische microdata in zijn beveiligingsonderzoek heeft meegenomen. Het optreden van fouten zal de mogelijkheid tot onthulling en de mate van vertrouwen in het onthulde, in meerdere of mindere mate beïnvloeden. Volgens Paass (1988) moeten we ons echter niet te veel voorstellen van de beschermende werking van de aanwezigheid van fouten. Mede daarom laten we de aan- of afwezigheid van fouten buiten beschouwing in dit handboek.

2.2.5 Maatregelen om risico op onthulling te verkleinen

Bij de beveiliging kan men op twee manieren te werk gaan: ingrijpen in de identificerende variabelen of ingrijpen in de overige variabelen. Door het informatiegehalte van de indirect identificerende variabelen te verminderen wordt de kans op identificatie verkleind en daarmee ook de *kans* op identiteit- en/of attribuut-onthulling. Denk daarbij aan het vervangen van een exacte leeftijd door leeftijdsklassen.
Bij ingrijpen in de overige variabelen wordt de *impact* van de onthulling over het algemeen verminderd. Een voorbeeld van dergelijk ingrijpen is het toevoegen van ruis aan het inkomen van een eenheid.
Microdata bevatten vaak een groot aantal indirect identificerende variabelen. Daardoor

is de kans op identificatie vrij groot en ligt het voor de hand om vooral het informatiegehalte van die variabelen te verminderen. Additioneel kan men ook de overige gegevens (drastisch) reduceren.
Geaggregeerde data worden vaak gepubliceerd in de vorm van een tabel met een zeer beperkt aantal variabelen. Desalniettemin kan het voorkomen dat bepaalde cellen, ondanks het aggregatieproces, toch over bepaalde herkenbare eenheden te veel informatie geven. Die informatie kan verminderd worden door ofwel op een andere tabel over te stappen (via “indikken van variabelen”) ofwel door de betreffende celwaarden weg te laten.

2.3 Basisbegrippen

In deze paragraaf geven we een opsomming van de belangrijkste termen die binnen de statistische beveiliging gebruikt worden.
Statistische beveiliging is gericht op het voorkomen van **onthulling**. Met **onthulling** van individuele gegevens bedoelen we het op basis van statistische publicaties kunnen herleiden van gegevens op een manier dat deze met zekerheid of zeer hoge waarschijnlijkheid geassocieerd kunnen worden met een **identificeerbare** eenheid.
Een eenheid is **identificeerbaar** als die unieke eenheid met zekerheid of zeer hoge waarschijnlijkheid in de populatie herkenbaar is. Om een eenheid te kunnen identificeren zijn **identificerende variabelen (identificatoren)** nodig.
Identificerende variabelen zijn gegevens die aan anderen dan de desbetreffende eenheid bekend zijn en gebruikt kunnen worden om de eenheid in de populatie te herkennen. Dergelijke gegevens worden ingedeeld in **directe** en **indirecte identificatoren**. Of een variabele identificerend is kan per situatie verschillen. Zo kunnen de scores op bepaalde variabelen van een eenheid slechts bij een beperkt aantal anderen bekend zijn en dus alleen door die groep gebruikt worden om de eenheid te identificeren.
Direct identificerende gegevens zijn expliciet bedoeld om een eenheid eenduidig en snel te kunnen aanwijzen. Voorbeelden zijn namen en registratienummers (BSN, KvK-nummer, etc.).
Indirect identificerende gegevens kunnen een hulpmiddel zijn bij het identificeren. De scores op dergelijke gegevens zijn op zichzelf vaak niet voldoende om een eenheid te identificeren, maar een combinatie van indirecte identificatoren kan wel tot een kleine groep van eenheden en soms zelfs unieke eenheden leiden. Voorbeelden zijn opleiding, leeftijd(sklasse), beroep, grootteklasse, etc. Merk overigens op dat het identificerende vermogen vaak beperkt blijft tot bepaalde combinaties van scores. Zo zal de score op “beroep” in combinatie met “woonplaats” bij de meeste personen niet tot identificatie leiden. Bij specifieke scores echter wel: burgemeester wonend in Amsterdam. Een combinatie van (indirect) identificerende variabelen wordt een **sleutel** genoemd.
De scores op variabelen die niet als identificerend zijn aangemerkt moeten bij verstrekking van statistische informatie als **vertrouwelijk** behandeld worden. Let op dat het identificerend en dus ook het vertrouwelijk zijn van variabelen situatie-afhankelijk kan zijn. Bovendien kan een variabele afhankelijk van de mate van detaillering soms identificerend en soms vertrouwelijk zijn volgens de hier beschreven indeling. Bijvoorbeeld de variabele inkomen. Als die variabele in (grove) klassen is ingedeeld, is dat als indirect identificerend te beschouwen: van je buurman of je collega weet je waarschijnlijk wel in welke klasse die valt. Het exacte inkomen tot op de euro nauwkeurig weet je echter vaak niet en is daarmee dan vertrouwelijk informatie.
Merk op dat identificerende variabelen die bij personen horen, **persoonsgegevens** zijn in

de zin van de AVG en dus ook als zodanig behandeld moeten worden. Daarnaast kunnen variabelen die niet als identificerend zijn aangemerkt, ook **persoonsgegevens** zijn in de zin van de AVG. Persoonsgegevens in de zin van de AVG zijn alle gegevens over een geïdentificeerde of identificeerbare natuurlijke persoon. Voor een nadere bespreking van persoonsgegevens in de zin van de AVG verwijzen we naar hoofdstuk 3.

Er bestaan verschillende manieren om het risico op onthulling te verkleinen. We noemen hier alleen de op het CBS meest gebruikte manieren. Voor een uitgebreider overzicht verwijzen we naar de **Methodenreeks - Statistische beveiliging** en naar Hundepool et al. (2012).
Het niet vrijgeven van een specifiek gegeven binnen een publicatie heet **onderdrukken**. Bij microdata betekent dit dat de score op een bepaalde variabele in een record wordt vervangen door de score "missend". Bij tabeldata betekent dit dat een celwaarde wordt vervangen door een symbool (bijvoorbeeld een × of een ·). Een tweede veel gebruikte manier is het **indikken** of **hercoderen** van variabelen. Daarbij worden nieuwe categorieën gecreëerd die gemiddeld genomen op een grotere groep individuen betrekking hebben dan de oorspronkelijke categorieën. Daaronder valt het vervangen van de precieze waarde van een gegeven door een klasse, en het vervangen van een klasse door een bredere.
Ook kan **ruis toegevoegd** worden om het risico op onthulling te verkleinen. Dit kan zowel op microdata-niveau als op tabelniveau gedaan worden. Het desbetreffende gegeven wordt dan in een gewijzigde vorm vrijgegeven, door toevoeging van een voor de gebruiker van de informatie onbekende waarde. Bij continue variabelen kun je denken aan het toevoegen van een willekeurige waarde getrokken uit een bepaalde kansverdeling. Door ruis toe te voegen aan identificerende variabelen kunnen die variabelen niet langer de zekerheid van identificatie garanderen. Daarmee verklein je de *kans* op onthulling. Door ruis toe te voegen aan overige variabelen wordt onzekerheid over de juistheid van de informatie die onthuld zou worden geïntroduceerd. Daarmee verklein je de *impact* van een mogelijke onthulling. Over het algemeen zal aselecte ruis gebruikt worden.
Afronden kan in zekere zin ook gezien worden als een vorm van ruis toevoegen. Ook afronden kan zowel op het niveau van microdata plaatsvinden als op het niveau van geaggregeerde data. Voor afronden bestaan meerdere methoden. Zo kan er stochastisch afgerond worden (met kans p naar boven afronden en met kans $1 - p$ naar beneden afronden), maar ook deterministisch (bijvoorbeeld afronden naar het dichtstbijzijnde tiental). In beide gevallen wordt onzekerheid over de exacte waarde geïntroduceerd.
Sommige methoden om het risico op onthulling te verkleinen hebben een effect op een gering aantal onderdelen van de te verstrekken informatie. Dit is dus een lokaal effect, denk aan onderdrukken van een waarde in een record of het onderdrukken van een celwaarde in een tabel. Andere methoden hebben effect op het geheel van de te verstrekken informatie. Dit is dan een globaal effect, denk aan het indikken van variabelen.
In de praktijk zal bij het verkleinen van het onthullingsrisico gebruik gemaakt worden van een combinatie van verschillende technieken. Zo zou bij microdata bijvoorbeeld eerst ingedikt kunnen worden waarna resterende onveilige situaties onderdrukt worden.

2.4 Het proces van onthulling

Het risico op onthulling is een combinatie van de **kans** op onthulling en de **impact** van de onthulling. Hoewel juridisch gezien alleen de *kans* op onthulling van belang is, speelt de

impact van een potentiële onthulling wel degelijk een rol in het bepalen van het beleid aangaande statistische beveiliging. De impact van de onthulling is subjectief, tijdsafhankelijk en cultuurafhankelijk. Zo kunnen bepaalde variabelen in de loop der tijd meer of juist minder als stigmatiserend gezien worden, is binnen bepaalde culturen het inkomen een publiek gegeven en zal de ene berichtgever het vervelender vinden dat zijn of haar persoonsgegeven bekend wordt dan een andere berichtgever. Voor het inschatten van de impact van de onthulling is ook inhoudelijke kennis over de gegevens essentieel.
Theoretisch gezien zal het risico op onthulling nooit nul zijn: de kans op onthulling zal onder andere afhangen van de manier waarop iemand probeert informatie uit de statistische output te halen. Dit heet een **onthullingsscenario**. Daarbij spelen vele aspecten een rol, waarvan we er hier een klein aantal zullen bespreken:

- Welke middelen heeft een gebruiker om te proberen iets te onthullen;
- Welke kennis over de eenheden heeft een gebruiker zelf al;
- Is een gebruiker doelbewust op zoek naar een eenheid.

Omdat het risico op onthulling theoretisch gezien nooit nul zal zijn, zal bij statistische beveiliging aan risicomanagement gedaan moeten worden.

2.4.1 Beschikbare middelen en kennis

Behalve vaardigheid in het onthullen is extra kennis over afzonderlijke eenheden uit de desbetreffende populatie een hulpmiddel voor een (potentiële) onthuller. Meer specifiek: over welke eenheden weet de gebruiker reeds het een en ander, en hoe uitgebreid is deze kennis? Dat zal onder andere afhangen van het “type” gebruiker: een opsporingsambtenaar kan bijvoorbeeld heel andere informatie over personen beschikbaar hebben dan bijvoorbeeld een onderzoeker of een beleidsmedewerker.
Aangezien identificatie vaak voorafgaat aan onthulling, is het nodig dat niet alleen de eenheid op basis van een combinatie van (indirect) identificerende variabelen uniek is in de populatie, maar ook dat de identificatie daadwerkelijk wordt uitgevoerd. Ofwel, de onthuller moet dan zelf, onafhankelijk van de verkregen statistische informatie, al weten dat er sprake is van uniciteit en dat de betreffende eenheid ook daadwerkelijk in het bestand zit.
Dit suggereert wellicht dat alleen naar uniciteit gekeken hoeft te worden. De nadruk valt inderdaad op uniciteit. Echter, ook een klein aantal eenheden die dezelfde waarde hebben op een sleutel van (indirecte) identificatoren is riskant. Vaak zal de aandacht van een (potentiële) onthuller vooral naar zeldzame scores uitgaan. Daarna is dan slechts een kleine hoeveelheid extra informatie nodig om een unieke persoon te vinden. Die extra informatie kan soms ook komen van variabelen die in het algemeen niet als identificerend beschouwd worden, maar die wel bij de onthuller bekend zijn voor de eenheid waar hij in geïnteresseerd is.

2.4.2 Doelbewust zoeken

Wanneer een onthuller doelbewust op zoek is, zal hij zoveel mogelijk middelen inzetten om een eenheid te herkennen. Daarbij kun je ten minste twee mogelijkheden onderscheiden: een onthuller is op zoek naar informatie over een van tevoren vastgestelde eenheid, of een onthuller wil weten of het mogelijk is om informatie over een willekeurige, identificeerbare eenheid te vinden.

Het kan echter ook voorkomen dat een gebruiker van de statistische output helemaal niet van plan is om iets te onthullen, maar bij gebruik van de statistische output "toevallig" iets opmerkt dat doet denken aan een bekende eenheid.

2.5 Vaststellen van onveilige situaties

In de volgende subparagrafen zullen we in grote lijnen weergeven op welke manier onveilige situaties herkend kunnen worden bij het publiceren of beschikbaar stellen van microdata en tabeldata.

2.5.1 Microdata

Bij microdata speelt voornamelijk de identificeerbaarheid van eenheden een rol. In grove lijnen komt het er op neer dat naar combinaties van identificerende variabelen gekeken moet worden, waarbij een "te klein aantal" op een onveilige situatie duidt. Dit worden de **zeldzame combinaties** genoemd.

De internationaal bekendste risicomaat die daarbij gebruikt wordt is de ***k*-anonimiteit** (*k*-anonymity). In dat model is een eenheid *k*-anoniem, als er minimaal $k-1$ andere records in het bestand zijn die precies dezelfde score hebben op een vooraf vastgestelde sleutel van identificerende variabelen. Iedere eenheid met die score is dus, op basis van die sleutel, niet te onderscheiden van de andere $k-1$ records. Indien een onthuller dan een willekeurige eenheid kiest, is de kans dat het juiste record is gekozen gelijk aan $1/k$.

Een alternatieve maat is ***k*^{m}-anonimiteit**. Stel dat er *M* identificerende variabelen in het bestand zijn aangewezen. In dit model wordt dan niet gekeken naar *k*-anonimiteit op één sleutel van *M* variabelen, maar naar *k*-anonimiteit voor iedere sleutel van *m* variabelen. Ofwel, er wordt gekeken naar $\binom{M}{m}$ sleutels. Een eenheid is dan k^m-anoniem als hij voor *iedere* sleutel *k*-anoniem is.

Deze modellen houden geen rekening met het feit dat een steekproef-unieke eenheid niet noodzakelijk ook een populatie-unieke eenheid hoeft te zijn. Met name bij (relatief) kleine steekproeffracties zullen eenheden vaak uniek zijn op een sleutel van identificerende variabelen. De geschiktheid van het gebruik van bijvoorbeeld *k*-anonimiteit hangt daarnaast ook af van de keuze voor *k* én van de keuze van de grootte van de sleutel van identificerende variabelen. Immers, hoe meer variabelen zijn opgenomen in de sleutel, hoe sneller een eenheid uniek zal zijn.

2.5.2 Tabeldata

Ook tabellen kunnen gebaseerd zijn op steekproeven of op gehele populaties. Voor de leesbaarheid worden hieronder maten geformuleerd voor tabellen gebaseerd op de gehele populatie. Voor steekproeftabellen zijn de maten eventueel aan te passen, zie hoofdstukken 4 en 5.

De meest voor de hand liggende maat is het aantal eenheden per tabelcel. Immers, hoe minder eenheden in een cel zitten, hoe sneller informatie over een unieke eenheid afgeleid kan worden. Dit speelt met name bij frequentietabellen, waarbij de opspanvariabelen[1] identificerend kunnen zijn en/of tot de categorie "overige variabelen" kunnen behoren. Kleine aantallen die scoren op bepaalde identificerende variabelen geven dan informatie over de score van een kleine groep eenheden op de "overige" variabele(n).

Bij tabellen waarbij niet de eenheden per cel geteld worden, maar waar één van de "overige" variabelen over alle eenheden in een cel wordt opgeteld, speelt een ander

[1] Opspanvariabelen zijn variabelen die de rijen en kolommen van een tabel definiëren

aspect een rol. Denk in dit geval bijvoorbeeld aan tabellen over de omzet van bedrijven, naar regio, grootteklasse en SBI. In die situatie wordt vaak gebruik gemaakt van een **concentratiemaat**. Een concentratiemaat geeft aan hoe sterk de waarde van een tabelcel geconcentreerd is rond de bijdrage van één of enkele eenheden in die cel. Een bekend voorbeeld is de **dominantieregel of (*n*, *k*)-regel**: als *n* eenheden meer dan *k*% van de totale celwaarde bijdragen, dan is dat een onveilige situatie.
Een andere concentratiemaat, de ***p*%-regel**, controleert of een willekeurige eenheid in een tabelcel de bijdrage van een willekeurig andere eenheid in die cel, op basis van de gepubliceerde celwaarden, niet nauwkeuriger kan schatten dan +/- *p*%.
Bij bovenstaande concentratiematen wordt de aanname gemaakt dat alle bijdragen van alle eenheden niet-negatief zijn.

2.5.3 Combinaties van output

Voor zowel microdata als tabeldata geldt dat onveilige situaties vaak per bestand of per tabel bepaald worden. Echter, het is van belang om in te zien dat het combineren van verschillende outputs vaak tot een verhoogd onthullingsrisico leidt. Zo kunnen er bepaalde relaties bestaan tussen cellen uit verschillende tabellen. Denk bijvoorbeeld aan een tabel op gemeente-niveau en dezelfde tabel op provincie-niveau. Een provincie bestaat uit een aantal gemeentes, waardoor de som van de waarde van een aantal cellen uit de eerste tabel gelijk zal zijn aan de waarde in een cel uit de tweede tabel.
Bij het toepassen van beveiligingsmethoden zal rekening gehouden moeten worden met dergelijke afhankelijkheden tussen outputs. Dit kan door de verschillende gerelateerde outputs simultaan consistent te beveiligen, of door een nieuwe output te beveiligen conditioneel op eerder beveiligde output.

Referenties

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. en de Wolf, P.P. (2012). *Statistical Disclosure Control*, Wiley Series in Survey Methodology, ISBN 978-1-119-97815-2.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6**, pp. 487–500.

Quick reference van gebruikte termen

Term [pagina]	Korte beschrijving
Afronden [15]	Waarde van een gegeven vervangen door een veelvoud van een afrondbasis
Attribuut-onthulling [13]	Onthulling van een eigenschap van een identificeerbare eenheid, zonder noodzakelijkerwijs die eenheid aan te kunnen wijzen
Concentratiemaat [18]	Maat die aangeeft hoe sterk de waarde van een tabelcel geconcentreerd is rond de waarde van één of enkele eenheden
Direct identificerende variabele [12]	Variabele die op zichzelf een unieke eenheid kan identificeren
Dominantieregel of (n, k)-regel [18]	*n* eenheden in een tabelcel mogen niet meer dan *k* % bijdragen aan het celtotaal
Groepsonthulling [12]	Onthulling van een eigenschap van een identificeerbare *groep* van eenheden
Indikken [15]	Creëren van nieuwe categorieën die ieder op een (gemiddeld) grotere groep individuen betrekking hebben dan de oude categorieën
Identificerende gegevens [11]	Gegevens die gebruikt kunnen worden om een eenheid te identificeren
Identiteit-onthulling [13]	Onthulling van de identiteit van een unieke eenheid
Indirect identificerend variabele [12]	Variabele die in combinatie met andere indirect identificerende variabelen een unieke eenheid kan identificeren
k-anoniem [17]	Er zijn minimaal *k*-1 andere records in het microbestand met dezelfde score op een vooraf vastgestelde sleutel van identificerende variabelen
k^m-anoniem [17]	*k*-anoniem op iedere sleutel van *m* identificerende variabelen
Onderdrukken [15]	Het niet vrijgeven van een specifiek gegeven binnen een publicatie
Onthullen [11]	Het op basis van statistische publicaties kunnen herleiden van gegevens op een manier dat deze met zekerheid of zeer hoge waarschijnlijkheid geassocieerd kunnen worden met een identificeerbare eenheid
Onthullingsscenario [16]	Manier waarop iemand probeert informatie te onthullen
Opspanvariabelen [17]	Variabelen die de rijen en kolommen van een tabel definiëren

Overige of statistische gegevens [11]	Gegevens die niet gebruikt worden ter identificatie van unieke eenheden
***p*%-regel** [18]	Een willekeurige eenheid in een tabelcel mag een andere eenheid niet nauwkeuriger dan +/- *p*% kunnen schatten op basis van die tabelcel
Persoonsgegevens in de zin van de AVG [15]	Alle gegevens over een geïdentificeerde of identificeerbare natuurlijke persoon
Risico op onthulling [15]	Combinatie van de kans op onthulling en de impact van die onthulling
Ruis toevoegen [15]	Een onbekende waarde toevoegen aan een gegeven
Sleutel [14]	Set van (indirect) identificerende variabelen
Statistische beveiliging [14]	Toepassen van methoden op te publiceren gegevens, om het risico op onthulling te verkleinen
Zeldzame combinatie [17]	Een combinatie van identificerende variabelen waarop een “te klein” aantal eenheden in de populatie dezelfde score heeft

3. Statistische beveiliging: wet- en regelgeving

3.1 Inleiding

Statistische beveiliging is één van de fundamenten van het statistisch onderzoek van het CBS. In dit hoofdstuk worden de wet- en regelgeving en beleidsaspecten die daarbij van belang zijn, geschetst met hun achtergronden.
De taak van het CBS is vastgelegd in de CBS-wet, artikel 3 t/m 5. Artikel 3 lid 1 van de CBS-wet luidt: *Het CBS heeft tot taak het van overheidswege verrichten van statistisch onderzoek ten behoeve van praktijk, beleid en wetenschap en het openbaar maken van de op grond van zodanig onderzoek samengestelde statistieken.* Het CBS is daarnaast op nationaal niveau belast met de productie van Europese statistieken (artikel 4) en kan in incidentele gevallen statistische werkzaamheden voor derden verrichten (artikel 5 lid 1). De artikelen 3 t/m 5 geven daarmee het "wat" in het kort aan. Het "hoe" wordt o.a. in artikel 37 aangegeven.

3.2 De centrale regel van de CBS-wet

De centrale bepaling in de CBS-wet met betrekking tot statistische beveiliging is artikel 37. Daarom wordt deze hier integraal geciteerd:

Artikel 37
1. De door de directeur-generaal in het kader van de uitoefening van de taken ter uitvoering van deze wet ontvangen gegevens worden uitsluitend gebruikt voor statistische doeleinden.
2. De in het eerste lid bedoelde gegevens worden niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.
3. De in het eerste lid bedoelde gegevens worden slechts zodanig openbaar gemaakt dat daaraan geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling kunnen worden ontleend, tenzij, ingeval het gegevens met betrekking tot een onderneming of instelling betreft, er een gegronde reden is om aan te nemen dat bij de betrokken onderneming of instelling geen bedenkingen bestaan tegen de openbaarmaking.

Uit het Uitvoeringsbesluit Ambtenarenwet 2017 (bijlage), de Algemene wet bestuursrecht (artikel 2:5) en het Wetboek van Strafrecht (artikel 272)[2], vloeit al een (ambtelijke) geheimhoudingsplicht voort. Artikel 37 van de CBS-wet specificeert deze voor de door het CBS verzamelde gegevens. Alle gegevens die het CBS verzamelt voor zijn taak, het produceren en publiceren van statistische (dat wil zeggen geaggregeerde) informatie, mogen uitsluitend voor dat doel gebruikt worden.

- Het eerste en tweede lid van artikel 37 sluiten gebruik voor fiscale, administratieve, controle en gerechtelijke doeleinden uit, alsmede de verstrekking van de gegevens aan derden die niet belast zijn met de uitvoering van de taak van het CBS.
- Het derde lid van artikel 37 maakt publicatie van bij het CBS berustende statistische gegevens slechts mogelijk in beveiligde vorm. Aan de publicatie mogen geen herkenbare gegevens van afzonderlijke respondenten kunnen worden ontleend.

[2] Hierin wordt schending van een geheimhoudingsplicht die geldt uit hoofde van ambt, beroep of wettelijk voorschrift strafbaar gesteld.

In paragraaf 3 van hoofdstuk 5 in de CBS-wet staan een aantal uitzonderingen op artikel 37. Maar in het kort staat ook hier dat de geheimhoudingseisen geldig blijven behoudens voor bepaalde verstrekkingen.

Uit het geciteerde artikel 37 vloeit dan ook rechtstreeks de verplichting voort dat het CBS zijn statistische informatie beveiligt: aan tabellen en andere (statistische) publicaties mogen geen herkenbare gegevens over afzonderlijke eenheden ontleend kunnen worden. De daarvoor binnen het CBS te controleren regels worden in de andere hoofdstukken van dit handboek beschreven. Methoden voor het veilig publiceren van statistische informatie staan in de **Methodenreeks - Statistische beveiliging**.

De formulering van artikel 37 geeft aanleiding tot een aantal observaties:

- Het gaat om de bescherming van herkenbare gegevens over statistische eenheden. Schending van de statistische beveiliging (“onthulling”) komt dus neer op de combinatie van twee feiten: herkenbaarheid van een eenheid èn bekendmaking van gegevens over die eenheid.
- Er wordt een expliciet onderscheid gemaakt tussen “verstrekken” (lid 2) en “openbaar maken” (lid 3). Het openbaar maken van statistieken is de wettelijke taak van het CBS: de resultaten zijn voor iedereen toegankelijk.
- Er wordt in deze wetstekst als zodanig geen onderscheid gemaakt tussen sociale (personen, huishoudens) en economische (ondernemingen, instellingen) statistieken.3,4 Ook het verschil tussen primaire (enquêtering) en secundaire (gebruik externe registraties) waarneming is voor de statistische geheimhouding niet van belang.
- Het onthullingsverbod in publicaties is absoluut voor personen en huishoudens.
- Het onthullingsverbod in publicaties voor ondernemingen of instellingen is niet absoluut en geldt niet als er een gegronde reden bestaat om aan te nemen dat de betrokken onderneming of instelling geen bezwaar tegen de openbaarmaking heeft. Denk daarbij allereerst aan wettelijke bepalingen en machtigingen. Er kunnen wettelijke bepalingen zijn die het CBS aanzetten tot publicatie op gemeentelijk niveau. In artikel 1 van de Gemeentewet wordt bijvoorbeeld verwezen naar het aantal inwoners zoals dat door het CBS openbaar wordt gemaakt. En in de Onderwijswetgeving is voorzien dat het CBS in bepaalde gevallen bij uitzondering op het niveau van individuele scholen mag publiceren. Een betrokken onderneming of instelling kan ook zelf laten weten geen bezwaar tegen de openbaarmaking te hebben (machtiging). Onder het huidige beleid behandelt het CBS gegevens die het voor de statistiekproductie ontleent aan openbare jaarverslagen, als vertrouwelijk. Voor de intracommunautaire handelsstatistiek (“Intrastat”) geldt overigens een afwijkend regime5, het “piepsysteem”, ook wel passieve geheimhouding genoemd: ondernemingen moeten zelf aangeven dat zij CBS-informatie die hun internationale handel herkenbaar zou maken, beveiligd willen zien. In het algemeen en uitzonderingen daargelaten geldt, dat het doel van een statistiek niet is statistische

[3] In termen van mate van herkenbaarheid van de gegevens en schade voor de eenheden zijn er wèl zeer relevante verschillen tussen de sociale en economische statistieken.

[4] De functionaris voor de gegevensbescherming (FG) die het CBS heeft ingesteld in de zin van artikel 37 Algemene Verordening Gegevensbescherming (AVG) ziet dan ook zowel op de bescherming van persoonsgegevens als op die van bedrijfsgegevens toe.

[5] Zie artikel 38a van de CBS-wet, ter uitvoering van verordening (EG) nr. 638/2004 van het Europees Parlement en de Raad van de Europese Unie van 31 maart 2004 betreffende de communautaire statistieken van het goederenverkeer tussen de lidstaten en houdende intrekking van Verordening (EEG) nr. 3330/91 (PbEG L 102).

informatie te geven over één bedrijf of instelling.
- De CBS-wet gaat verder dan de AVG (zie paragraaf 3.4 hieronder). In de CBS-wet gaat het over “personen” terwijl het in de AVG gaat over “natuurlijke personen”. De CBS-wet gaat daarmee bijvoorbeeld ook over overleden personen, maar de AVG niet.
- De oorzaak van onthulling bij verstrekking of openbaarmaking doet er niet toe. Het CBS moet zich óók hoeden voor nalatigheid. Juist daarom wordt er gewerkt met formele procedures en criteria zoals in dit handboek op hoofdlijnen beschreven, en met de softwaretools μ- en τ-Argus.

Afgezien van de wettelijke verplichting liggen er twee motieven aan de verplichting tot statistische beveiliging ten grondslag. Allereerst is één van de fundamentele principes van gegevensverwerking hier van toepassing. Gegevensverwerking moet transparant, spaarzaam, en functioneel zijn.[6] Dat wil zeggen dat gegevensverwerking kenbaar moet zijn voor degene op wie de gegevens betrekking hebben. Er moeten niet meer gegevens verwerkt worden dan nodig. En de gegevensverwerking moet met een specifiek (voor betrokkenen kenbaar) doel voor ogen plaatsvinden. Als het CBS gegevens ontvangt voor het produceren van statistieken, mogen deze niet zonder medeweten en instemming van betrokkenen voor andere (niet-statistische) doeleinden gebruikt worden. Dat is overigens - voor het CBS gelukkig - niet integraal omkeerbaar: statistische doeleinden worden wèl als compatibel (verenigbaar) met andere doeleinden gezien. Mede daarom mag het CBS relatief gemakkelijk gebruik maken van allerlei gegevens uit administraties en registraties. De geheimhoudingsverplichting past naadloos in de statistische beroepsethiek.
Het tweede in dit verband relevante motief is gelegen in het eigen belang van het CBS bij continuïteit van de gegevensverzameling door (= berichtgeving aan) het CBS. Als respondenten of registratiehouders gegevens aan het CBS verstrekken voor de productie en publicatie van statistieken moeten zij erop kunnen rekenen dat die gegevens niet op een andere manier worden gebruikt. Het CBS is, zoals we hierboven zagen, wettelijk verplicht tot statistische geheimhouding maar zegt deze ook nog eens expliciet toe aan zijn berichtgevers. Houdt het CBS zich niet aan zijn wettelijke verplichtingen en eigen toezeggingen, dan moet het nadrukkelijk voor een verminderde gegevensstroom vrezen. Dat geldt zowel voor berichtgeving door respondenten als door registratiehouders.

3.3 Wettelijke uitzonderingen op de centrale regel

In de CBS-wet heeft de wetgever een aantal uitzonderingen op de statistische beveiligingsverplichting in strikte zin gesteld. Eigen aan ieder van deze uitzonderingen is dat ze, exclusief, het doel van statistiek en onderzoek voor de samenleving dienen. Zonder ze in het geheel te citeren, laat staan te bespreken, gaat het om de volgende uitzonderingen:

- Artikel 39 maakt het mogelijk om gegevens te verstrekken aan de communautaire en nationale instanties voor de statistiek van de lidstaten van de Europese Unie of leden van het Europees Stelsel van Centrale Banken, voor zover deze verstrekking noodzakelijk is voor de productie van specifieke communautaire statistieken. Ook als het niet juridisch verplicht is mag het CBS gegevens verstrekken aan deze instanties maar dan onder strakkere voorwaarden die het CBS zelf moet monitoren. In de

[6] De AVG (zie paragraaf 3.4) kent als belangrijkste algemene beginselen:
- doelbinding en rechtmatigheid;
- transparantie, gegevenskwaliteit, bewaartermijnen en rechten van de burger;
- techniek en beveiliging.

praktijk hanteert het CBS dan dezelfde voorwaarden als bij andere verstrekkingen aan derden.

- Artikel 40 heeft betrekking op gegevensuitwisseling met De Nederlandsche Bank, voor statistische doeleinden.
- Artikelen 41 en 42 leggen de grondslag voor het microdatabeleid van het CBS. Het CBS mag microbestanden uitleveren, die dan beveiligd worden volgens de regels van hoofdstuk 5 van dit handboek. Het mag ook onderzoekers on site, via remote execution of via remote access toegang verlenen tot zijn microbestanden. In die situatie is vooral cruciaal dat de onderzoekers in (de tabellen en tekst van) hun publicaties geen herkenbare gegevens onthullen. De methoden en criteria voor beveiliging van kwantitatieve tabellen, frequentietabellen en statistische analyse zijn dan van toepassing. Welke onderzoekers aldus voor statistische en wetenschappelijke doeleinden de microbestanden van het CBS kunnen gebruiken is geregeld in het tweede lid van artikel 41. Voor zover zij niet expliciet in de wet genoemd worden, mag de DG van het CBS op basis van een aantal expliciete criteria de betreffende instelling een machtiging verlenen. De criteria voor het verlenen van toegang tot microdata van het CBS zijn vastgelegd in de “Beleidsregel toegang instellingen tot microdata CBS”.7
- Artikel 42a ten slotte maakt het mogelijk dat het CBS onder bepaalde, strikte voorwaarden koppelbare doodsoorzakengegevens uit kan leveren.

3.4 De algemene privacywetgeving in Nederland

De algemene privacywetgeving is in Nederland vastgelegd in de Algemene Verordening Gegevensbescherming (AVG) en de Uitvoeringswet AVG (UAVG).[8] Samen vaak aangeduid als AVG of (U)AVG.

- De AVG verstaat onder een persoonsgegeven: elk gegeven betreffende een geïdentificeerde of identificeerbare natuurlijke persoon9. Als er aan één van beide elementen (natuurlijk persoon, identificeerbaarheid) niet is voldaan, dan is er geen sprake van persoonsgegevens en is de AVG niet van toepassing.
- Een persoon is identificeerbaar indien zijn identiteit redelijkerwijs, zonder onevenredige inspanning, vastgesteld kan worden. Twee factoren spelen hierbij een rol: de aard van de gegevens en de mogelijkheden van de verantwoordelijke om identificatie tot stand te brengen.
- Ten aanzien van de identificeerbaarheid van een persoon kunnen in de zin van de AVG drie categorieën van gevallen worden onderscheiden.
    - Verwerking van persoonsgegevens op naam: de heer X heeft een krediet tot een bedrag Y van bank Z. In die situatie is vanzelfsprekend sprake van een persoonsgegeven.
    - Gevallen waarbij gegevens niet direct op naam zijn terug te vinden, maar de betrokken persoon met aanwending van beschikbare middelen alsnog kan worden achterhaald, bijvoorbeeld aan de hand van een (direct identificerend) nummer, of met behulp van indirect identificerende gegevens.
    - Gevallen waarbij de identiteit van de betrokken personen niet of slechts

[7] Beleidsregel van 1 augustus 2021, nr. CSB-2021-072.

[8] Daarnaast zijn van belang de Wet basisregistratie personen (2013) en de Wet op de Geneeskundige Behandelingsovereenkomst (Afdeling 5 Boek 7 Burgerlijk Wetboek). Deze zijn echter zo (domein)specifiek dat er hier niet verder op wordt ingegaan.

[9] Dat moet niet restrictief worden uitgelegd: gegevens met betrekking tot kentekens of eenpersoonsbedrijven worden ook als persoonsgegevens aangemerkt.

met een disproportionele aanwending van geld, menskracht of middelen kan worden achterhaald. Deze laatste gegevens worden niet als persoonsgegevens aangemerkt.

- Bij het voortschrijden van de informatietechnologie moet rekening worden gehouden met het feit dat waar voorheen wellicht nog sprake was van een onevenredige inspanning (en dus niet van een persoonsgegeven), deze inspanning geringer wordt met het beschikbaar komen van nieuwe technieken. Dit betekent dat gegevens die voorheen niet als persoonsgegeven beschouwd werden, dat later wel kunnen worden en dan wel onder de AVG zouden vallen. Nieuwe publicaties hebben daar rekening mee te houden. Het mag duidelijk zijn dat het geen zin heeft om oudere publicaties hierop aan te passen: die zijn immers al publiek gemaakt.
- Bij de verwerking van gegevens voor wetenschappelijk onderzoek en statistiek (o&s) ondervindt betrokkene daar in beginsel geen gevolgen van. De AVG bevat dan ook specifieke uitzonderingsbepalingen voor wetenschappelijk onderzoek en statistiek:
    - Verenigbaar gebruik: verwerking voor o&s wordt verenigbaar geacht met de oorspronkelijke doelen waarvoor gegevens verzameld zijn.
    - Bewaartermijnen: langer bewaren van persoonsgegevens gaat tegen de oorspronkelijke doelbinding in maar kan onder voorwaarden voor o&s toelaatbaar zijn.
    - Bijzondere (gevoelige) en strafrechtelijke gegevens: voor het verwerken van bijzondere (gevoelige) en strafrechtelijke persoonsgegevens ten behoeve van o&s zijn de voorwaarden iets minder strikt dan bij andere verwerkingen van deze gegevens.
    - Rechten van de betrokkenen: betrokkenen hebben bij o&s geen recht op gegevenswissing/vergetelheid indien dit het maken van statistiek onmogelijk maakt of in het gedrang brengt. Daarnaast kan het recht op inzage, rectificatie en beperking van de verwerking eveneens buiten beschouwing worden gelaten.
- De AVG voorziet in zelfregulering en in een aantal instrumenten om het toezicht op de uitvoering en handhaving van haar bepalingen te vergemakkelijken. Daarbij behoren de functionarissen gegevensbescherming (FG's) die binnen/door instellingen aangewezen kunnen worden als toezichthouder. Daarnaast kunnen sectoren gezamenlijke gedragscodes vastleggen en bij de AP melden.10

3.5 Communautaire wetgeving[11]

Naast de nationale wetgeving is er communautaire wetgeving van toepassing. In deze paragraaf noemen we de wetgeving die specifiek voor de officiële statistiek van belang is:

- De algemene statistische wet van de EU, ook wel de Europese statistiekverordening genoemd12. In artikel 2 daarvan wordt de statistische geheimhouding genoemd als één van de zes statistische beginselen en omschreven als: 'vertrouwelijke gegevens betreffende individuele statistische eenheden die direct voor statistische doeleinden

[10] Zo heeft de AP een instemmende verklaring afgelegd bij de gedragscode voor de brancheverenigingNLdigital.
[11] Internationaal recht (Europees recht) - wet- en regelgeving die betrekking heeft op de Europese Gemeenschap.
[12] Verordening (EG) Nr. 223/2009 van het Europees Parlement en de Raad van 11 maart 2009 betreffende de Europese statistiek en tot intrekking van Verordening (EG, Euratom) nr. 1101/2008 betreffende de toezending van onder de statistische geheimhoudingsplicht vallende gegevens aan het Bureau voor de Statistiek van de Europese Gemeenschappen, Verordening (EG) nr. 322/97 van de Raad betreffende de communautaire statistiek en Besluit 89/382/EEG, Euratom van de Raad tot oprichting van een Comité statistisch programma van de Europese Gemeenschappen.

of indirect uit administratieve of andere bronnen zijn verkregen, moeten worden beschermd, wat betekent dat het verboden is de verkregen gegevens voor niet-statistische doeleinden te gebruiken of ze op onrechtmatige wijze openbaar te maken;' Daarnaast kent hoofdstuk 5 (artikel 20 t/m 26) van de Europese statistiekverordening bepalingen over statistische geheimhouding.

- Het beginsel van statistische geheimhouding uit artikel 2 van de Europese statistiek verordening is conform artikel 11 van de verordening verder uitgewerkt in de Praktijkcode voor Europese statistieken van 16 november 2017. Beginsel 5 van de Praktijkcode gaat over statistische geheimhouding en gegevensbescherming.
- Om een Eurostat-microdatabeleid ten behoeve van wetenschappelijk onderzoekers mogelijk te maken is er ten slotte de Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002. Deze verordening betreft een groot aantal bestanden waarvan de meeste kunnen worden geclassificeerd als social sample surveys. Toelating van nieuwe gebruikers vergt een expliciet besluit. Discussies over de (statistische) beveiliging van deze bestanden komen aan bod in de specifieke working groups, de expert group on statistical disclosure control (EGSDC) en de microdata access network group (MANG).

Quick reference wet en regelgeving

Wetgeving [pagina]	Korte beschrijving
Artikel 37 CBS wet [21]	Centrale bepaling in de CBS-wet met betrekking tot statistische beveiliging: *Artikel 37* *Lid 1. De door de directeur-generaal in het kader van de uitoefening van de taken ter uitvoering van deze wet ontvangen gegevens worden uitsluitend gebruikt voor statistische doeleinden.* *Lid 2. De in het eerste lid bedoelde gegevens worden niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.* *Lid 3. De in het eerste lid bedoelde gegevens worden slechts zodanig openbaar gemaakt dat daaraan geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling kunnen worden ontleend, tenzij, ingeval het gegevens met betrekking tot een onderneming of instelling betreft, er een gegronde reden is om aan te nemen dat bij de betrokken onderneming of instelling geen bedenkingen bestaan tegen de openbaarmaking.*
Artikel 39 CBS wet [23]	Uitzondering voor communautaire statistieken
Artikel 40 CBS wet [24]	Uitzondering voor De Nederlandse Bank
Artikel 41 en 42 CBS wet [24]	Grondslag voor toegang tot microdata
Artikel 42a CBS wet [24]	Grondslag voor toegang tot doodsoorzakengegevens
AVG [24]	Algemene Verordening Gegevensbescherming
Verordening 223 [25]	Algemene statistische wet van de EU
Verordening 557 [26]	Grondslag voor toegang tot microdata bij Eurostat

4. Beveiliging van kwantitatieve tabellen

4.1 Inleiding

In dit hoofdstuk beperken we ons tot **kwantitatieve tabellen**; tabellen waar de cel--inhoud de som van de scores van een numerieke variabele is. Dit in tegenstelling tot frequentietabellen, waar de cel-inhoud tot stand komt door de aantallen eenheden te tellen. Deze tabellen dienen volgens een ander stramien te worden beveiligd en zullen in hoofdstuk 5 worden behandeld.

Kwantitatieve tabellen komen vaak voor bij economische statistieken. Denk daarbij aan tabellen over de totale omzet naar bedrijfstak, de (internationale) handelsomzet naar land van bestemming of het aantal werknemers naar bedrijfstak en regio. Maar ook bij sociale statistieken kunnen kwantitatieve tabellen geproduceerd worden, zoals tabellen over inkomen naar geslacht of beroepsgroep.

De indruk zou kunnen bestaan dat tabellen die slechts geaggregeerde gegevens bevatten, geen enkel beveiligingsrisico hebben. Niets is minder waar. Ook uit geaggregeerde tabellen kunnen gegevens van individuele eenheden worden achterhaald. Denk bijvoorbeeld aan een cel met slechts één enkele bijdrager. Dan is de celwaarde de score van die ene bijdrager die daarmee dus wordt onthuld. Maar dit is niet het enige gevaar; ook cellen met twee bijdragers vormen een risico. Immers de ene bijdrager kan met behulp van de tabel gegevens achterhalen van de andere door zijn eigen bijdrage van het celtotaal af te trekken. Dit is een ernstig risico bij met name economische statistieken, mede omdat het niet onwaarschijnlijk is dat de twee bijdragers elkaar kennen, elkaars concurrenten zijn en dus zeer geïnteresseerd zijn in de gevoelige bedrijfsgegevens van elkaar. Zelfs cellen met meer dan twee bijdragers kunnen een probleem vormen. Als de cel is samengesteld uit één grote bijdrager en een paar kleine bijdragers, dan is het celtotaal een heel goede schatting van de waarde van die grote bijdrager.

Dit speelt uiteraard niet alleen bij tabellen gebaseerd op enquêtes, maar ook bij tabellen gebaseerd op administratieve data. Bij economische statistieken worden bij enquêtes overigens vaak de grootste bedrijven integraal waargenomen, waardoor het verschil tussen administratieve data en op enquêtes gebaseerde data bij dergelijke statistieken niet echt van betekenis is wat betreft statistische beveiliging.

Voor mogelijke methoden om tabellen met risicovolle cellen te beveiligen, verwijzen we naar de **Methodenreeks - Statistische beveiliging**. Daarnaast willen we aangeven dat op het CBS de software τ-ARGUS beschikbaar is voor zowel het controleren van de *regels* als voor het toepassen van verschillende beveiligings*methoden*.

4.2 Te beveiligen cellen

In deze paragraaf wordt beschreven hoe kan worden vastgesteld welke informatie in tabellen al dan niet kan worden gepubliceerd. Zoals al is aangegeven zal een cel met één bijdrager vrijwel altijd als onveilig moeten worden beschouwd. Het betreft immers eigenlijk de individuele score van een bijdrager. Bij cellen met meerdere bijdragers moeten we de individuele scores van de bijdragers apart bekijken om te kunnen beslissen of een cel veilig is of niet. Bij twee bijdragers kunnen ze elkaars waarde exact bepalen; buitenstaanders zijn daar in het algemeen minder goed toe in staat.

Een “geheimhoudingsregel” moet uitkomst bieden om te bepalen of een cel veilig is of niet. Omdat we individuele bijdragen moeten beveiligen, gebruiken we de *p*%-regel om

te bepalen of een tabelcel veilig is of niet. De *p*%-regel zegt dat een bijdrager in een cel niet nauwkeuriger dan *p*% geschat mag kunnen worden. De “worst-case” situatie is dat de een-na-grootste bijdrager (x_2) de grootste bijdrager (x_1) probeert te schatten met $T_C - x_2$, waarbij T_C het celtotaal is. Om te bepalen of een cel onveilig is, is het dan ook noodzakelijk en voldoende om te toetsen of

$$\frac{(T_C - x_2) - x_1}{x_1} < \frac{p}{100}$$

Deze regel is nog uit te breiden met de aanname dat de individuele score van elke bijdrage tot de cel tot op *q*% nauwkeurig bekend is bij een potentiële onthuller. Het is niet geheel onrealistisch te veronderstellen dat dergelijke a-priori kennis aanwezig is, met name als het gaat om concurrerende bedrijven. De CBS standaard is om de *p*%-regel te gebruiken bij kwantitatieve tabellen
Experimenteel is vastgesteld dat voor *p* waarden uit het interval [5, 15] gewenst zijn. Dit wordt ook internationaal gezien als *best practice*.
Per definitie heeft de *p*%-regel tot gevolg dat er minimaal 3 bijdragers in een cel moeten zitten. Eventueel kan als extra regel toegevoegd worden dat er minimaal k_0 bijdragers in een cel moeten zitten: de zogenaamde minimale frequentieregel. In de praktijk wordt, in combinatie met de *p*%-regel, een drempelwaarde k_0 van 4 of 5 genomen.
Tevens kan de *p*%-regel uitgebreid worden naar een situatie waar niet één bijdrager een ander probeert te schatten, maar waar meerdere bijdragers samenwerken om een andere bijdrager te proberen te schatten. In de praktijk wordt dit echter als een minder waarschijnlijk scenario gezien. Daarom wordt, ook internationaal, meestal met de oorspronkelijke *p*%-regel gewerkt.

4.3 Regels voor kwantitatieve tabellen

4.3.1 De *p*%-regel

Voor het bepalen van de primair onveilige cellen in een kwantitatieve tabel gebruikt het CBS de *p*%-regel:

Een cel met de waarde T_C is **onveilig** als

$$\frac{(T_C - x_2) - x_1}{x_1} < \frac{p}{100}$$

waarbij x_1 de grootste en x_2 de op een na grootste bijdrage aan die cel is.

De parameter *p* moet voldoen aan $5 \leq p \leq 15$. Er is hier vanuit gegaan dat alle bijdragers niet-negatief zijn en dat minimaal één bijdrager groter dan nul is.
Een grote waarde van *p* is strenger in die zin dat een cel sneller als onveilig wordt aangewezen dan bij een kleine waarde van *p*.
De keuze voor een specifieke waarde van *p* kan bijvoorbeeld afhangen van de impact die onthulling van een (nauwkeurige) schatting van de grootste bijdrager zou hebben.

4.3.2 De minimale frequentie regel

Aanvullend op de *p*%-regel kan de minimale frequentieregel gebruikt worden: een cel is **onveilig** als er minder dan k_0 bijdragers aan het celtotaal zijn.

Let op:

Statistieken die over dezelfde populatie en dezelfde variabelen gaan, moeten dezelfde waarde van *p* en k_0 gebruiken.

4.3.3 Aanvullingen bij het gebruik van de *p*%-regel

Steekproef

Indien de tabel gemaakt wordt op basis van een steekproefbestand, kan de *p*%-regel worden uitgebreid. Elke bijdrage x_i aan een cel heeft zijn eigen ophooggewicht w_i. Voor het celtotaal wordt uiteraard elke bijdrage opgehoogd met dat gewicht. Voor de beveiliging wordt een virtuele cel gemaakt waarin van elke bijdrage w_i replica's voorkomen. Op de virtuele cel wordt dan de *p*%-regel toegepast. Bij gebroken gewichten is dit eenvoudig uit te breiden. In het algemeen leidt dit tot minder gevoelige cellen dan indien men deze ophooggewichten negeert.

Holding

Van belang is ook om te bepalen wat een eenheid is die men wil beveiligen. Als meerdere records die tot een cel bijdragen tot eenzelfde holding behoren, zou ten onterechte de indruk kunnen ontstaan dat de cel veilig is. In werkelijkheid bestaat de cel slechts uit een geringer aantal grotere bijdragers. Het is dan ook verstandig de *p*%-regel op het niveau van de holding toe te passen.

4.4 Bijzondere situaties

4.4.1 Buitenlandse handel

Voor de statistiek van de buitenlandse handel zijn alternatieve beveiligingsregels van kracht. Op basis van een wettelijk kader wordt hier afgeweken van de actieve toepassing van de bovengenoemde beveiligingsregels en wordt gesproken over **passieve beveiliging**. Slechts in die situatie dat een bedrijf/bijdrager zelf aan het CBS verzoekt zijn gegevens niet herkenbaar in de tabellen te publiceren, wordt geheimhouding toegepast. In dat geval wordt, als dat bedrijf meer dan een bepaald percentage bijdraagt aan de cel of de cel uit slechts een gering aantal bijdragers bestaat, de cel als onveilig beschouwd. De wettelijke basis voor deze speciale aanpak bij de buitenlandse handel is te vinden in artikel 38a van de CBS-wet.

4.4.2 Machtigingen

Soms hebben de bijdragers van een paar zeer grote bedrijven een zeer drastische invloed op het beveiligingspatroon. Om de paar grote bijdragers afdoende te kunnen beveiligen zijn vaak vergaande maatregelen nodig. Omdat de gegevens van deze grote bedrijven soms toch al op andere wijze bekend zijn, bijvoorbeeld uit jaarverslagen, kan het helpen deze bedrijven expliciet toestemming te vragen om hun gegevens herkenbaar te mogen publiceren. Deze toestemming is nodig, omdat het CBS nooit uit eigen

beweging deze gegevens zal publiceren. Indien op deze wijze een paar zeer grote cellen vrijgegeven kunnen worden, kan dit leiden tot aanzienlijk minder informatieverlies ten gevolge van de toegepaste beveiligingsmethoden.
Machtigingen dienen altijd voor een beperkte periode (max. 5 jaar) geldig te zijn.
De *p*%-regel is aan te passen zodanig dat er rekening wordt gehouden met machtigingen. In feite komt het er op neer dat gecontroleerd moet worden dat geen enkele bijdrager een schatting nauwkeuriger dan *p*% kan maken van een andere bijdrager *zonder machtiging*.

4.4.3 Negatieve bijdragers

De *p*%-regel zoals hier gegeven, gaat ervan uit dat alle bijdragen tot een cel niet-negatief zijn. Wanneer er zowel positieve als negatieve bijdragen zijn, zal er een aangepaste versie van de *p*%-regel gebruikt moeten worden. Afhankelijk van de situatie zijn er meerdere mogelijkheden. Het belangrijkste is dat de essentie van de *p*%-regel behouden blijft. Voor deze situatie wordt geadviseerd om contact op te nemen met een van de statistische beveiligingsexperts.

4.4.4 Nul-cellen

Cellen die alleen bijdragers kennen met bijdrage nul zijn een bijzonder probleem. Hier wordt uitdrukkelijk niet een lege cel bedoeld. Een lege cel is een cel zonder bijdragers. Een cel met de *waarde* nul is onveilig (bij uitsluitend niet-negatieve bijdragers): hieruit is immers af te leiden dat alle bijdrages nul zijn en in dat geval is er sprake van onthulling. De *p*%-regel is in dit geval niet toepasbaar. Daarom zullen dergelijke cellen handmatig als onveilig bestempeld moeten worden.

4.5 Aandachtspunten

4.5.1 Detaillering

Alvorens de beveiligingsregels toe te passen dient men te bedenken in welke mate van detail het wenselijk is de tabel te publiceren. Veel detaillering leidt tot veel cellen met slechts een klein aantal bijdragers. Na toepassing van de beveiligingsregels worden die onveilig en het toepassen van een beveiligingsmethode zal ook effect hebben op vele andere (op zichzelf veilige) cellen. Het enigszins indikken van een tabel kan zeer zinnig zijn, omdat door het effectief samenvoegen van cellen vaak veel onveilige cellen zullen verdwijnen.

4.5.2 “Missing” code bij opspanvariabelen

Het kan voorkomen dat de variabelen die de kolommen en rijen van een tabel definiëren (de opspanvariabelen) een code voor “missing” hebben. Dat wil zeggen dat voor iedere cel in de rij of de kolom met de code “missing” de categorie van die opspanvariabele niet bekend is. Hoewel een degelijke cel onveilig kan zijn volgens de regels, is het de vraag of die cel daadwerkelijk als onveilig gezien moet worden.
In het algemeen zal het immers lastig zijn de individuele bijdragers van die cel te identificeren. Afhankelijk van hoe identificeerbaar bijdragers in dergelijke cellen zijn kan gekozen worden om zulke cellen per definitie als veilig te beschouwen.

4.5.3 Gekoppelde tabellen

Vaak worden tabellen gemaakt die gemeenschappelijke opspanvariabelen hebben. Denk bijvoorbeeld aan omzet naar SBI x Regio en omzet naar SBI x Grootteklasse. Beide tabellen bevatten dan de randtotalen naar SBI.
Het is evident dat bij dergelijke tabellen altijd dezelfde parameterwaarden voor de regels gebruikt moeten worden. Echter, ook bij toepassing van een methode om de tabellen te beveiligen zal rekening gehouden moeten worden met de koppeling tussen dergelijke tabellen.

4.5.4 Meerdere hiërarchieën

Een vaak geuite wens is de behoefte om eenzelfde tabel te publiceren naar verschillende hiërarchische indelingen van de opspanvariabelen. Bijvoorbeeld regio opgesplitst naar provincies en naar gemeenten. Ook hier is het evident dat in al die tabellen dezelfde parameterwaarden voor de regels gebruikt moeten worden. En ook nu zal bij het toepassen van een methode om de tabellen te beveiligen rekening gehouden moeten worden met de koppeling tussen de tabellen. Dat is relatief eenvoudig als de gebruikte hiërarchieën genest zijn. Genest wil zeggen dat de ene hiërarchie een nette deelhiërarchie van de andere is. Denk bijvoorbeeld aan een tabel naar 2-digit SBI en een tabel naar 3-digit SBI. Het komt echter ook voor dat de gebruikte hiërarchieën *niet* genest zijn. Denk daarbij aan een tabel naar Provincies en een tabel naar Zorgregio's: een Provincie is niet altijd precies gelijk aan een groep Zorgregio's. Het verschil kan soms resulteren in (zeer) kleine gebieden. In het geval van niet-geneste hiërarchieën is het toepassen van beveiligingsmethoden een stuk ingewikkelder. Het wordt aangeraden om in die gevallen contact op te nemen met een van de statistische beveiligingsexperts.

Quick reference voor beveiliging van kwantitatieve tabellen

Regel [pagina]	**Korte beschrijving**	**Toegestane parameter waarden**
***p*%-regel** [30]	Onveilig als $\frac{(T_C - x_2) - x_1}{x_1} < \frac{p}{100}$ waar T_C de celwaarde is, x_1 de grootste en x_2 de op een na grootste bijdrage aan de cel.	$5 \leq p \leq 15$
Minimale frequentie [31]	Optioneel; aanvullend op *p*%-regel. Onveilig als aantal eenheden in cel kleiner is dan k_0.	

Kleine *p* is minder streng, grote *p* is strenger
Kleine k_0 is minder streng, grote k_0 is strenger

Aandachtspunten	
	Statistieken die over dezelfde populatie en dezelfde variabelen gaan, moeten dezelfde waarde van *p* en k_0 gebruiken.

Speciale situaties [pagina]	
Nul-cel [32]	Als *alle* bijdragers aan een cel de *waarde* nul bijdragen, kan er sprake zijn van groepsonthulling.
Steekproef [31]	Indien de tabel gemaakt wordt op basis van een steekproefbestand, kan de *p*%-regel worden uitgebreid zodat rekening wordt gehouden met de ophooggewichten.
Holding [31]	Als meerdere records die aan een celtotaal bijdragen tot eenzelfde eenheid horen (meerdere vestigingen van één holding bijvoorbeeld), moeten die records als één bijdrager beschouwd worden bij het controleren van de regel(s).
Passieve beveiliging [31]	Alleen als een berichtgever zelf aan het CBS verzoekt zijn gegevens niet herkenbaar in de tabellen te publiceren wordt geheimhouding toegepast. Is alleen van toepassing op enkele statistieken over de Internationale Handel.
Machtiging [31]	Expliciete toestemming van een bedrijf om zijn gegevens herkenbaar te mogen publiceren.

5. Beveiliging van frequentietabellen

5.1 Inleiding

In dit hoofdstuk beperken we ons tot **frequentietabellen**: tabellen waarbij de celwaarde gelijk is aan het aantal eenheden dat aan een combinatie van een aantal kenmerken voldoet. Dit in tegenstelling tot kwantitatieve tabellen, waar de celwaarde tot stand komt door de scores op een numerieke variabele op te tellen van alle eenheden die in die cel zitten.
Frequentietabellen komen vooral in de sociale statistieken voor. Denk daarbij aan tabellen met het aantal personen naar regio, geslacht en opleiding of het aantal huishoudens naar type huishouden. Maar ook bij economische statistieken komen frequentietabellen voor, zoals bijvoorbeeld tabellen met aantallen startende bedrijven naar regio en SBI.
In hoofdstuk 2 staat beschreven dat onthulling vaak neerkomt op een combinatie van twee feiten: herkenning van een eenheid en het daardoor bekend worden van "overige" gegevens over die eenheid.
Voor frequentietabellen kunnen we dat als volgt formuleren. Een gebruiker moet eerst een bijdrager of een groep bijdragers herkennen in de tabel aan de hand van de score op een aantal identificerende opspanvariabelen[13]. Bijvoorbeeld de groep mannen in Son en Breugel van 100 jaar of ouder met een universitaire opleiding zou herkenbaar kunnen zijn. Daarna volgt een uitspraak over deze bijdrager(s) door de verdeling van de eenheden over de cellen van de andere opspanvariabelen. De uitspraak die dan mogelijk is, moet meer informatie bevatten dan alleen de groepsgrootte van die identificeerbare groep. Het is dan immers niet mogelijk om meer te weten te komen dan nodig was om de identificeerbare groep in de tabel te herkennen. In een frequentietabel van het aantal personen naar regio, leeftijd, opleiding en geslacht zou bijvoorbeeld de groep mannen in Son en Breugel van 100 jaar of ouder met een universitaire opleiding wellicht te herkennen zijn (de identificeerbare groep), maar deze tabel bevat geen andere informatie dan de groepsgrootte. Zou de tabel daarnaast ook informatie bevatten over de migratieachtergrond, dan zou je die extra informatie over de identificeerbare groep kunnen herleiden.

Een frequentietabel voldoet aan de wettelijke eisen waar we ons als CBS aan moeten houden als de tabel geen informatie oplevert over een herkenbare *individuele* statistische eenheid. De statistische beroepsethiek en het eigenbelang van het CBS in continuïteit van de berichtgeving aan het CBS leiden in bepaalde gevallen tot de eis dat een tabel geen informatie op mag leveren over *groepen* van statistische eenheden. Dat is met name het geval als de tabel variabelen bevat die kwetsend of kwetsbaar makende informatie over die groepen kan opleveren.

Anders dan bij kwantitatieve tabellen, zijn er voor frequentietabellen geen eenvoudige "geheimhoudingsregels". Ook internationaal zijn er geen algemene regels bekend.
Ter verkenning zijn een aantal statistische bureaus van andere landen benaderd (Australië, Canada, Duitsland, Groot-Brittannië en de Verenigde Staten) en is gevraagd naar de regels die zij hanteren voor frequentietabellen. Samengevat komt het er op neer dat deze bureaus een ad-hocbeleid voeren als het gaat om de beveiliging van

[13] Opspanvariabelen zijn variabelen die de rijen en kolommen van een tabel definiëren.

frequentietabellen. Een aantal bureaus noemt groepsonthulling (ook wel het 100%-cellen probleem genoemd) expliciet als een (potentieel) onveilige situatie.
In dit hoofdstuk formuleren we een aantal regels die voor de Nederlandse situatie in de meeste voorkomende gevallen van frequentietabellen gebruikt kunnen worden. De regels worden besproken met het uitgangspunt dat de frequentietabellen over de hele populatie gaan. De regels zijn echter ook toepasbaar op tabellen gebaseerd op steekproeven en die vervolgens zijn opgehoogd naar de hele populatie.

5.2 Onthullingsrisico bij frequentietabellen

Het onthullingsrisico bij frequentietabellen hangt af van **alle** variabelen die bij de definitie van de tabel een rol spelen. Dat zijn dus niet alleen de variabelen die de rijen en kolommen van een tabel opspannen, maar ook variabelen die gebruikt zijn voor het bepalen van de (deel)populatie waar de tabel betrekking op heeft. Dus een tabel over de mannelijke bevolking, met als celwaarde het aantal personen van een bepaald beroep in een bepaalde regio heeft drie variabelen die een rol spelen: Beroep, Regio en Geslacht. Beroep en Regio spannen de rijen en kolommen op en Geslacht bepaalt de doelpopulatie waar de tabel betrekking op heeft.

Om algemene regels te kunnen formuleren, stellen we ons voor dat alle variabelen die bijdragen tot de herkenbaarheid van groepen van eenheden (de identificerende variabelen) zijn samengenomen in de voorkolom van die tabel. De kolommen van de tabel bevatten dan de overige variabelen die informatie geven over die groepen van eenheden. In de praktijk zijn tabellen meestal niet zo ingericht, maar in gedachte kan een willekeurige tabel op deze manier worden herordend.

Neem als voorbeeld Tabel 1 van aantal personen die een niet-natuurlijke dood zijn gestorven, uitgesplitst naar type doodsoorzaak, geslacht en leeftijd (NB: de getallen zijn uiteraard fictief). Daarbij zijn Geslacht en Leeftijd de variabelen die de herkenbare groepen definiëren en is Niet-natuurlijke doodsoorzaak een “overige” variabele.

De identificerende variabelen zijn hier Geslacht en Leeftijd. Herordenen we de tabel zodat de identificerende variabelen in de voorkolom staan, dan krijgen we Tabel 2.
In het vervolg van dit hoofdstuk gaan we er vanuit dat de frequentietabellen in de “standaard” vorm staan: de rijen geven de herkenbare groepen weer en de kolommen de categorieën van de overige (“gevoelige”) variabele(n).

Niet-natuurlijke doodsoorzaak	Geslacht	Leeftijd				
		Totaal	[0, 25)	[25, 50)	[50, 75)	[75, ∞)
Zelfmoord	Totaal	2047	51	1656	298	42
	Man	1544	42	1314	181	7
	Vrouw	503	9	342	117	35
Moord	Totaal	69	24	36	8	1
	Man	24	14	7	2	1
	Vrouw	45	10	29	6	-
Verkeersongeval	Totaal	559	167	173	179	40
	Man	315	110	93	98	14
	Vrouw	244	57	80	81	26
Werkplaats ongeval	Totaal	35	3	26	6	-
	Man	33	3	24	6	-
	Vrouw	2	-	2	-	-
Persoonlijk ongeval	Totaal	1518	70	30	481	937
	Man	339	34	6	223	76
	Vrouw	1179	36	24	258	861
Overig/onbekend	Totaal	76	8	10	37	21
	Man	29	5	3	20	1
	Vrouw	47	3	7	17	20
Totaal	Totaal	4304	323	1931	1009	1041
	Man	2284	208	1447	530	99
	Vrouw	2020	115	484	479	942

Tabel 1: Aantal personen die een niet-natuurlijke dood zijn gestorven, naar geslacht en leeftijd

		Niet-natuurlijke doodsoorzaak						
Geslacht	Leeftijd	Zelfmoord	Moord	Verkeers-ongeval	Werkplaats ongeval	Persoonlijk ongeval	Overig / onbekend	Totaal
Man	[0, 25)	42	14	110	3	34	5	208
	[25, 50)	1314	7	93	24	6	3	1447
	[50, 75)	181	2	98	6	223	20	530
	[75, ∞)	7	1	14	-	76	1	99
	Totaal	1544	24	315	33	339	29	2284
Vrouw	[0, 25)	9	10	57	-	36	3	115
	[25, 50)	342	29	80	2	24	7	484
	[50, 75)	117	6	81	-	258	17	479
	[75, ∞)	35	-	26	-	861	20	942
	Totaal	503	45	244	2	1179	47	2020
Totaal	[0, 25)	51	24	167	3	70	8	323
	[25, 50)	1656	36	173	26	30	10	1931
	[50, 75)	298	8	179	6	481	37	1009
	[75, ∞)	42	1	40	-	937	21	1041
	Totaal	2047	69	559	35	1518	76	4304

Tabel 2: Geherordende versie van Tabel 1

5.2.1 Herkenbaarheid

In de inleiding van dit hoofdstuk hebben we aangegeven dat het onthullingsproces bij frequentietabellen uit twee elementen bestaat: de definitie van een herkenbare groep bijdragers en de verdeling daarvan over de categorieën van de andere opspanvariabelen. Als eerste moeten we ons dus afvragen wat we als herkenbare groepen moeten beschouwen. Zoals we in hoofdstuk 2 hebben aangegeven, speelt daarbij de voorkennis van de gebruiker en de mate van identificeerbaarheid van de variabelen een rol.
Merk op dat herkenbaarheid een relatief begrip is. Soms kunnen gebruikers in bepaalde omstandigheden specifieke voorkennis hebben waardoor een groep voor hen herkenbaar wordt, terwijl voor gebruikers zonder die voorkennis de groep niet herkenbaar is. Bij het controleren van herkenbaarheid zal dus rekening gehouden moeten worden met de gebruiker van de frequentietabel.
De groep van identificerende variabelen is niet beperkt tot de bekende identificatoren zoals adres, geslacht, geboortedatum of leeftijd, opleiding en beroep. Daarnaast zijn er nog tal van kenmerken die kunnen bijdragen tot de herkenbaarheid van afzonderlijke eenheden, zoals: het krijgen van een bepaalde uitkering, het bezitten van een bepaald soort goed of het meedoen aan bepaalde activiteiten. Het is daarom op dit moment niet mogelijk om een operationele receptuur te geven voor het systematisch vaststellen van herkenbaarheid.
Vooralsnog gaat de controle op herkenbaarheid dus op zicht. Wel wordt gewerkt aan een samenvattende rapportage van in de praktijk aangetroffen en beoordeelde combinaties van variabelen. Met zo'n rapportage en een regelmatige actualisering daarvan dient de beoordeling op herkenbaarheid efficiënt en consequent plaats te vinden.
Als de herkenbare groepen eenmaal gedefinieerd zijn, is de vervolgvraag of er dan informatie wordt gegeven over individuele leden van die groepen. De mededeling bijvoorbeeld dat er twee mensen van het mannelijk geslacht in Nederland zijn met een lengte van meer dan 230 cm, vormt een mededeling over een herkenbare groep (de groep Nederlandse mannen van meer dan 230 cm). Maar er is geen sprake van het geven van informatie over de individuele leden van de groep: de genoemde kenmerken van lengte en geslacht zijn nodig om ze te herkennen. Pas als daarnaast ook gemeld wordt dat deze twee personen beide in de WAO zitten, wordt informatie gedeeld over de leden van de groep.

5.2.2 Overige of "gevoelige" gegevens

Zoals in hoofdstuk 2 is aangegeven, houdt het CBS de richtlijn aan dat uit statistische informatieverstrekking *geen enkel* tot een specifieke eenheid herleidbaar gegeven mag blijken. Dit om de discussie over al dan niet gevoeligheid van bepaalde gegevens te voorkomen. In de praktijk zal soms toch over gevoelige gegevens gesproken worden. Daarbij wordt in eerste instantie gedacht aan de bijzondere categorieën van persoonsgegevens uit artikel 9 van de AVG en aan de strafrechtelijke persoonsgegevens uit artikel 10 van de AVG:

> "… persoonsgegevens waaruit ras of etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen, of het lidmaatschap van een vakbond blijken, […] genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een persoon, of gegevens over gezondheid, of gegevens met betrekking tot iemands seksueel gedrag of seksuele gerichtheid …"

> "Persoonsgegevens betreffende strafrechtelijke veroordelingen en strafbare feiten of daarmee verband houdende veiligheidsmaatregelen …"

Bij herkenbare groepen van eenheden speelt het probleem dat één of meerdere categorieën van een overige ("gevoelige") variabele informatie over die groep oplevert die in het bijzonder kwetsend is voor die groep of die de groep in het bijzonder maatschappelijk kwetsbaar maakt.

In dat geval is de (moeilijke) afweging aan de orde tussen het openbaar maken van statistisch relevante informatie en de ethische bezwaren tegen het bekend maken van gevoelige gegevens over herkenbare groepen in de samenleving.

5.3 Te beveiligen situaties

In de volgende paragrafen gaan we er vanuit dat er minimaal één overige ("gevoelige") variabele is. Met andere woorden, de tabel bevat meer dan alleen de groepsomvang van de herkenbare groepen.

Merk op dat, zoals ook in hoofdstuk 2 is uitgelegd, sommige variabelen tegelijkertijd als identificerend en als "gevoelig" beschouwd kunnen worden. Dat maakt het operationaliseren van het controleren op onveilige cellen van frequentietabellen extra lastig.

5.3.1 Kleine aantallen eenheden

Simpel gezegd moeten we tabellen toetsen op de aanwezigheid van informatie over afzonderlijke, herkenbare eenheden. Als een herkenbare groep dus uit één eenheid bestaat kan dit niet gepubliceerd worden. Echter, aangezien de gebruiker van de tabel mogelijk zelf deel uitmaakt van een in de tabel herkenbare groep, mag een herkenbare groep ook niet over twee eenheden gaan. Voor die gebruiker zou de tabel dan immers informatie geven voor één andere eenheid in zijn herkenbare groep.

5.3.2 Groepsonthulling

Naast het rijtotaal voor een herkenbare groep eenheden, is de verdeling van die eenheden over de kolommen van de tabel van belang. Als 100% van de eenheden uit de herkenbare groep in dezelfde kolom zit (alle eenheden hebben dus dezelfde score op de overige ("gevoelige") variabele), wordt over al die eenheden informatie gegeven. De betreffende cel wordt dan ook wel een "100%-cel" genoemd. We spreken dan van "groepsonthulling". Dit geldt ook als bijna alle eenheden dezelfde waarde hebben: de "afwijkende" eenheden binnen de herkenbare groep kunnen dan met grote mate van zekerheid informatie afleiden over de rest van de groep.

Merk op dat groepsonthulling ook vanuit de andere kant gezien kan worden: als op een categorie niet gescoord wordt, wordt 100% op de andere categorieën gescoord. Dus lege cellen kunnen ook een indicator voor groepsonthulling zijn.

5.4 Regels voor frequentietabellen

Op basis van de redenatie onder paragraaf 5.3.1 is de eerste regel dat een herkenbare groep eenheden uit minimaal 3 eenheden moet bestaan.

Als een herkenbare groep eenheden uit 3 of meer eenheden bestaat, gelden er aanvullende regels die rekening houden met groepsonthulling. Merk op dat, wanneer de

groepsomvang relatief klein is het vaak niet om statistisch relevante informatie gaat. Bovendien is de statistische betrouwbaarheid van uitspraken over relatief kleine groepen ook vaak beperkt. Aan de andere kant, als de groepsomvang niet heel klein is, weegt de plicht tot het openbaar maken van statistische relevante informatie zwaarder dan het aspect groepsonthulling. Deze twee aspecten maken een keuze nodig voor de minimale omvang van herkenbare groepen waarvoor groepsonthulling wel is toegestaan.

Dit wordt geoperationaliseerd met behulp van een grenswaarde *k* voor het rijtotaal (de grootte van de herkenbare groep). Indien het rijtotaal kleiner is dan *k* en ten minste 90% van de eenheden van die groep dezelfde waarde voor een kolomvariabele hebben, mag de informatie over de herkenbare groep niet als zodanig gepubliceerd worden. De waarde van *k* kan verschillend gekozen worden afhankelijk van de impact die het bekend worden van die informatie zou hebben. Als de impact verwaarloosbaar is, wordt gekozen voor $k = k_0$ (< 30). In alle andere gevallen wordt een hogere waarde voor *k* gekozen. Afhankelijk van de aard van de overige ("gevoelige") gegevens wordt *k* gekozen tussen 30 en 100. De exacte waarde voor *k* die wordt gekozen, zal bepaald moeten worden door de betrokken statistische sector en vervolgens consequent toegepast moeten worden.

5.4.1 Aggregatieprobleem

Samenvoegen van kolommen of rijen (aggregeren) levert een tabel op met hooguit evenveel informatiewaarde en vaak zelfs met minder informatiewaarde. Uit het oogpunt van statistische beveiliging moet daarbij wel worden voorkomen dat een geaggregeerde tabel onthullende informatie bevat, terwijl de onderliggende gedetailleerdere tabel publicabel geacht wordt. Dat risico is er met name in verband met groepsonthulling wanneer dat puur op de verdeling over de categorieën van de betreffende tabel wordt gebaseerd.

Kijken we bijvoorbeeld naar de rij "Man", "[75, ∞)" in Tabel 2 van paragraaf 5.2. In deze rij is geen groepsonthulling: de grootste cel ("Persoonlijk ongeval") is 77% van het rijtotaal.

Als we de kolommen "Verkeersongeval", "Werkplaats ongeval" en "Persoonlijk ongeval" samenvoegen tot de categorie "Ongeval", krijgen we de volgende tabel:

Geslacht	Leeftijd	Zelfmoord	Moord	Ongeval	Overig / onbekend	Totaal
Man	[0, 25)	42	14	147	5	208
	[25, 50)	1314	7	123	3	1447
	[50, 75)	181	2	327	20	530
	[75, ∞)	7	1	90	1	99
	Totaal	1544	24	687	29	2284
Vrouw	[0, 25)	9	10	93	3	115
	[25, 50)	342	29	106	7	484
	[50, 75)	117	6	339	17	479
	[75, ∞)	35	-	887	20	942
	Totaal	503	45	1425	47	2020
Totaal	[0, 25)	51	24	240	8	323
	[25, 50)	1656	36	229	10	1931
	[50, 75)	298	8	666	37	1009
	[75, ∞)	42	1	977	21	1041
	Totaal	2047	69	2112	76	4304

Tabel 3: Geaggregeerde versie van Tabel 2

In deze tabel is het aandeel “Ongeval” in de rij “Man”, “[75, ∞)” 91%, wat dus als groepsonthulling gezien kan worden (bij $k = 100$).

Door de operationalisering van groepsonthulling puur te baseren op de verdeling over de categorieën van Tabel 2 zou daar dus geen groepsonthulling zijn in de rij “Man”, “[75, ∞)”, maar gebaseerd op de verdeling over de categorieën in de minder gedetailleerde Tabel 3 wel.

Het probleem doet zich met name voor bij “zinvolle” aggregaties van een indeling. Met “zinvolle” aggregaties bedoelen we dan aggregaten die een statistisch informatieve betekenis hebben. In bovenstaande voorbeeld zou het samenvoegen van “Zelfmoord” met “Werkplaats ongeval” geen “zinvolle” aggregaat zijn. Welke aggregaten “zinvol” zijn, is in feite een beleidsbeslissing. Als een gebruiker zelf een nieuwe indeling maakt van de kolommen, zal gekeken moeten worden in hoeverre daaruit door het CBS als zinvol bestempelde aggregaten bepaald kunnen worden. Alleen in dat geval zal de regel voor groepsonthulling ook op de tabel met zinvolle aggregaten getoetst moeten worden.

Merk op dat de aanpak van “zinvolle” aggregaten ook aansluit bij het probleem van categorieën waar helemaal niet op gescoord wordt. Als het gaat om één categorie zonder eenheden, dan kan de combinatie van alle overige categorieën ook een “zinvol” aggregaat zijn waar dan 100% op scoort.

5.4.2 Samenvattende regels

Samenvattend komen we dan tot de volgende drie regels voor frequentietabellen:

F1 Als een herkenbare groep uit minder dan 3 eenheden bestaat, is de informatie over die groep niet als zodanig te publiceren.

F2 Als een herkenbare groep uit minder dan k eenheden bestaat en ten minste 90% van de eenheden (met een minimum van het rijtotaal min 1) zijn geconcentreerd in één categorie, is de informatie over die groep niet als zodanig te publiceren. De waarde voor k is afhankelijk van de impact die het bekend worden van de informatie zou hebben. Voor verwaarloosbare impact geldt $k = k_0$ (< 30). Voor niet-verwaarloosbare impact ("gevoelige" informatie) kan een waarde voor k gekozen worden tussen 30 en 100. Hoe groter de impact, hoe hoger de waarde voor k.

F3 Regel F2 moet voor alle "zinvolle" aggregaten van de categorieën van de overige variabelen gecontroleerd worden.

Bij het controleren van een frequentietabel op onveilige situaties, moeten alle drie de regels gecontroleerd worden.

De regels moeten worden toegepast op de onbeveiligde tabellen, dus voordat eventuele beveiligingsmethoden (zoals afronden) zijn toegepast.

Quick reference voor beveiliging van frequentietabellen

Regel [pagina]	Korte beschrijving	Toegestane parameter waarden
F1 [42]	Onveilig als herkenbare groep kleiner is dan 3	
F2 [42]	Onveilig als herkenbare groep kleiner is dan k en tenminste 90% van de eenheden in die herkenbare groep scoort op dezelfde categorie van de "overige variabele(n)"	"Niet gevoelige" overige variabele(n): $k = k_0$ (< 30) "Gevoelige" overige variabele(n): $30 \leq k \leq 100$
F3 [42]	Regel F2 moet ook gecontroleerd worden voor alle "zinvolle aggregaten" van de overige variabele(n)	

Kleine k is minder streng, grote k is strenger

Regels F1, F2 en F3 worden toegepast op de originele, onbeveiligde tabel. Dus vóór eventuele afronding van aantallen of andere beveiligingsmethoden.

6. Beveiliging van microdata

6.1 Inleiding

In dit hoofdstuk behandelen we de beveiliging van microdata die het CBS verlaten. Microdata die op het CBS blijven, zoals microdata die door CBS medewerkers worden gebruikt voor de productie van CBS-statistieken, vallen hier nadrukkelijk buiten. Ook microdata die on-site of via remote acces kunnen worden geanalyseerd (en dus ook het CBS niet verlaten) vallen niet onder dit hoofdstuk. In het laatste geval vindt de beveiliging plaats op de statistische output van de analyses.
Bij microdata die het CBS verlaten onderscheiden we twee soorten bestanden: bestanden met microdata die uitsluitend met een begeleidend contract worden geleverd (microdatabestanden onder contract) en bestanden met microdata die zonder contract worden geleverd (publicatiebestanden). In paragraaf 2 gaan we in op de achtergronden van beide soorten bestanden en lichten we met name toe wat onder ieder type bestand wordt verstaan. De regels voor publicatiebestanden en die voor microdatabestanden onder contract worden in paragraaf 6.3 respectievelijk paragraaf 6.4 besproken.
Voor mogelijke methoden om microdata waarin onveilige situaties voorkomen te beveiligen, verwijzen we naar de **Methodenreeks - Statistische beveiliging**. Daarnaast willen we aangeven dat op het CBS de software μ-ARGUS beschikbaar is voor zowel het controleren van de *regels* als voor het toepassen van verschillende beveiligings*methoden*.

6.2 Typen microdatabestanden

Internationaal worden drie typen bestanden met microdata onderscheiden: Public Use Files, Scientific Use Files en Secure Use Files. Public Use Files en Scientific Use Files zijn bestanden met microdata die het CBS verlaten en komen overeen met onze publicatiebestanden respectievelijk microdatabestanden onder contract. Secure Use Files zijn bestanden die het CBS *niet* verlaten en vallen zoals gezegd buiten de scope van dit hoofdstuk.

6.2.1 Publicatiebestanden

Publicatiebestanden zijn bestemd voor willekeurige externe gebruikers, zonder aanvullende eisen aan het gebruik of de gebruikers. Bij het CBS zijn alleen publicatiebestanden mogelijk waarbij de eenheden natuurlijke personen zijn. Het is dus niet mogelijk om een publicatiebestand te maken over bijvoorbeeld bedrijven, instellingen, gemeentes of overheidsdiensten.
Aangezien publicatiebestanden door willekeurig wie gebruikt kunnen worden, moeten deze bestanden zeer goed statistisch beveiligd worden. De regels om te bepalen of een bestand als publicatiebestand gepubliceerd kan worden, zijn dan ook zeer strikt. Dit leidt over het algemeen tot een grote reductie van het informatiegehalte. Het nut van publicatiebestanden ligt dan ook niet zozeer op het terrein van wetenschap of beleidsvorming. Publicatiebestanden zijn voornamelijk bedoeld voor gebruik in het onderwijs of als illustratie bij de ontwikkeling van statistische methoden. Een groot voordeel van publicatiebestanden is dat ze gratis zijn en gedeeld mogen worden.
Op het CBS noemen we een bestand een publicatiebestand als het aan alle drie de volgende eisen voldoet:

- het bestand heeft betrekking op gedesaggregeerde gegevens;
- het bestand gaat over natuurlijke personen;
- het bestand wordt zonder contract geleverd.

Het eerste punt geeft aan dat het om een bestand met microdata gaat en dus *niet* over geaggregeerde gegevens zoals een tabel. Het tweede punt geeft aan dat het *niet* over bedrijven, instellingen, gemeentes en overheidsdiensten kan gaan. En het laatste punt geeft aan dat het vrij beschikbaar is en dus qua toegang afwijkt van de microdatabestanden onder contract.
Merk overigens op dat het maken en leveren van een publicatiebestand wel met een contract kan zijn vastgelegd, maar dat er niet noodzakelijk contractueel is vastgelegd wat wel en niet gedaan mag worden met het bestand. Dit nadrukkelijk in tegenstelling tot de situatie bij microdatabestanden onder contract, zoals we in de volgende paragraaf zullen zien.

6.2.2 Microdatabestanden onder contract

Microdatabestanden onder contract zijn bedoeld voor statistisch en wetenschappelijk gebruik en worden met een begeleidend contract geleverd aan bonafide onderzoeksinstellingen. In een dergelijk begeleidend contract staat o.a. wat er wel en niet met het bestand gedaan mag worden. Denk daarbij aan het niet delen met derde partijen, wel/niet vernietigen na afloop van het project en wel/niet koppelen met andere bestanden. Ook wordt in het contract vastgelegd wat de consequenties zijn als niet aan de vastgelegde afspraken wordt voldaan. Zie hoofdstuk 3, paragraaf 3.2 voor meer informatie over welke onderzoeksinstellingen deze microdatabestanden onder contract mogen ontvangen.
Vanwege deze contractuele en juridische beperkingen, kan de statistische beveiliging van microdatabestanden onder contract minder streng zijn dan in het geval van publicatiebestanden. Daardoor kan het informatiegehalte ook hoger blijven en zijn de microdatabestanden onder contract wel geschikt voor wetenschap of beleidsvorming.
Een microdatabestand onder contract bestaat, net als een publicatiebestand, uit records op individueel niveau. In tegenstelling tot publicatiebestanden, kunnen microdatabestanden onder contract niet alleen over natuurlijke personen gaan, maar in principe ook over huishoudens, bedrijven, instellingen, gemeentes of overheidsdiensten. Echter, bedrijfsgegevens zijn over het algemeen veel schever verdeeld dan persoonsgegevens. Dit impliceert dat bedrijven veel herkenbaarder zijn dan personen. Het gevolg is dat er voor microdatabestanden onder contract over bedrijven alsnog vrij zware statistische beveiligingsmethoden toegepast moeten worden, waardoor vaak een minder bruikbaar bestand overblijft. In de praktijk verstrekt het CBS dan ook geen microdatabestanden onder contract over bedrijven.
De in dit hoofdstuk geformuleerde regels zijn opgesteld vanuit het oogpunt van microdatabestanden onder contract over natuurlijke personen en huishoudens. Voor microdatabestanden onder contract over gemeentes, instellingen of overheidsdiensten, en eventueel bedrijven, zullen de regels naar die specifieke situaties vertaald moeten worden.

Controlled Circulation Files
Een speciaal geval van microdatabestanden onder contract zijn de zogenaamde Controlled Circulation Files (CCFs). Deze vorm van microdatabestanden is in 2020 door het directieberaad bekrachtigd. CCFs zijn specifiek toegesneden op wensen van het SCP.

Voor CCFs zijn geen specifieke *regels*, maar zijn *criteria* opgesteld die ervoor moeten zorgen dat CCFs met grote waarschijnlijkheid aan de regels voor microdatabestanden onder contract voldoen.
Deze criteria voor CCFs zijn beschreven in een intern document. De criteria zijn echter gevoelig voor veranderingen in de definities van (de categorieën van) variabelen en zullen dus van tijd tot tijd moeten worden bijgesteld. De criteria worden dan ook niet in dit handboek genoemd, maar we verwijzen naar de meest recente versie van het document op intranet.

6.3 Regels voor publicatiebestanden

Voor alle type bestanden met microdata die het CBS verlaten, geldt dat er geen directe identificatoren in het bestand mogen voorkomen. Dit is dus ook van toepassing op publicatiebestanden.

P0	Directe identificatoren mogen niet in het bestand voorkomen.

Daarnaast gelden voor publicatiebestanden nog zes regels: de Ouderdomsregel, de Selectieregel, de Ophoogregel, de Toelatingsregel, de Huishoudensregel en de Volgorderegel. Publicatiebestanden moeten aan alle regels voldoen.
In volgende paragrafen gaan we in op iedere specifieke regel.

6.3.1 P1: de Ouderdomsregel

P1	Er moet ten minste één jaar verstreken zijn tussen de beëindiging van het veldwerk en de levering van het bestand.

Bij steekproefbestanden neemt door deze regel de bruikbaarheid van responskennis om personen te identificeren af: responskennis verwatert vaak naarmate de tijd verstrijkt. Bovendien, en dat geldt ook voor bestanden gebaseerd op administratieve data, neemt de kennis over de scores op een aantal identificerende variabelen af. Daardoor wordt de kans ook groter dat het publicatiebestand niet meer aansluit op recentere (eventueel extern beschikbare) bestanden. De onzekerheid over de juistheid van een veronderstelde onthulling neemt daarmee toe.

6.3.2 P2: de Selectieregel

De Selectieregel is opgesplitst in vier sub-regels:

P2.1	Directe regioaanduidingen mogen niet in het bestand voorkomen.
P2.2	In het bestand is maximaal één type regio toegestaan. De betreffende regiovariabele moet voldoende gespreid zijn.
P2.3	Er mogen maximaal 15 identificerende variabelen in het bestand voorkomen.
P2.4	Er mogen geen “gevoelige” gegevens in het bestand voorkomen.

Deze regel sluit de aanwezigheid van bepaalde variabelen in een publicatiebestand uit. In het bijzonder mogen in een publicatiebestand geen directe regioaanduidingen en geen “gevoelige” variabelen voorkomen. Publicatiebestanden zijn eigenlijk landelijke bestanden met hooguit “grove” regionale informatie. Directe regioaanduidingen als Provincie, Gemeente en dergelijke zijn dus *niet* toegestaan. Indirecte regioaanduidingen

als Stedelijkheidsgraad zijn wel mogelijk, mits voldoende gespreid. Voldoende gespreid betekent hier:

- Geografische spreiding: Iedere categorie van de indirecte regioaanduiding komt in minimaal 6 provincies voor;

èn

- Demografische spreiding: Geen enkele gemeente in een categorie van de indirecte regioaanduiding mag meer dan 50% van het totaal aantal inwoners binnen die categorie van de regioaanduiding bevatten.

Bij toepassen van de sub-regels moeten ook de variabelen meegenomen worden die een eventuele doelpopulatie van het hele bestand definiëren. Zo verbiedt P2.1 een publicatiebestand over Drentenaren, aangezien Drenthe een directe regioaanduiding is. Ook voor de berekening van het aantal identificerende variabelen in sub-regel P2.3 moeten de variabelen meegenomen worden die een eventuele doelpopulatie van het hele bestand definiëren. Als bijvoorbeeld de doelpopulatie van het publicatiebestand "mannen van 21 jaar of ouder" is, dan mogen er nog maximaal 13 identificerende variabelen in het bestand voorkomen. De doelpopulatie wordt hier immers gedefinieerd door twee identificerende variabelen: geslacht en leeftijd.
Ten slotte mogen er geen "gevoelige" gegevens in een publicatiebestand voorkomen, om onthulling van dergelijke gegevens met 100% zekerheid uit te sluiten. Voor discussie over "gevoelige" gegevens, zie paragraaf 2.2.1.

6.3.3 P3: de Ophoogregel

P3	Het moet uitgesloten zijn dat ophooggewichten gebruikt kunnen worden om additionele informatie met betrekking tot identificerende variabelen af te leiden.

Ophooggewichten kunnen extra informatie bevatten over identificerende variabelen. Denk bijvoorbeeld aan gewichten bij een gestratificeerde steekproef, waarbij de strata de provincies zijn. Volgens regel P2.1 mag er geen variabele Provincie in het bestand zitten. Maar in zo'n gestratificeerde steekproef zijn de gewichten constant per provincie, en wordt er dus impliciet toch informatie over de provincies gegeven. Deze regel is bedoeld om dit te voorkomen.

6.3.4 P4: de Toelatingsregel

De Toelatingsregel is opgesplitst in twee sub-regels:

P4.1	Er mogen geen categorieën van identificerende variabelen voorkomen die minder dan *K* personen in de populatie betreffen.
P4.2	Er mogen geen waardencombinaties van twee gekruiste identificerende variabelen voorkomen die minder dan *L* personen in de populatie betreffen.

Het uitgangspunt hier is dat bepaalde combinaties van identificerende variabelen in het *bestand* voldoende vaak in de *populatie* voorkomen. Voor steekproefbestanden moeten de grenswaarden op populatieniveau omgerekend worden naar aantallen in de steekproef.
Wanneer het publicatiebestand over een deelpopulatie van de Nederlandse bevolking gaat, geldt de Toelatingsregel ook voor de identificerende variabelen die de

deelpopulatie definiëren. Een bestand over leerkrachten in het basisonderwijs in Nederland is dan bij bijvoorbeeld *K* = 200 000 nooit toegestaan, aangezien er in Nederland minder dan 200 000 leerkrachten in het basisonderwijs werkzaam zijn, maar kan bij *K* = 100 000 mogelijk wel gepubliceerd worden omdat er meer dan 100 000 leerkrachten in het basisonderwijs werkzaam zijn.

6.3.5 P5: de Huishoudensregel

P5	Ieder record van een meerpersoonshuishouden dat met ten minste twee records in het bestand zit, moet op de kruising van alle huishoudensvariabelen een waardencombinatie bevatten die bij ten minste *H* huishoudens in het bestand voorkomt.

Deze regel is bedoeld om te vermijden dat huishoudensrecords kunnen worden geformeerd op basis van persoonsrecords. Kennis van de samenstelling van een huishouden leidt immers veel sneller tot identificatie dan kennis van afzonderlijke leden van dit huishouden.

6.3.6 P6: de Volgorderegel

P6	De volgorde van de records in het bestand moet willekeurig zijn.

Deze regel is bedoeld om het hergroeperen van persoonsrecords uit eenzelfde huishouden of regio moeilijker te maken.

6.4 Regels voor microdatabestanden onder contract

Voor alle type bestanden met microdata die het CBS verlaten geldt dat er geen directe identificatoren in het bestand mogen voorkomen. Dit is dus ook van toepassing op microdatabestanden onder contract.

M0	Directe identificatoren mogen niet in het bestand voorkomen.

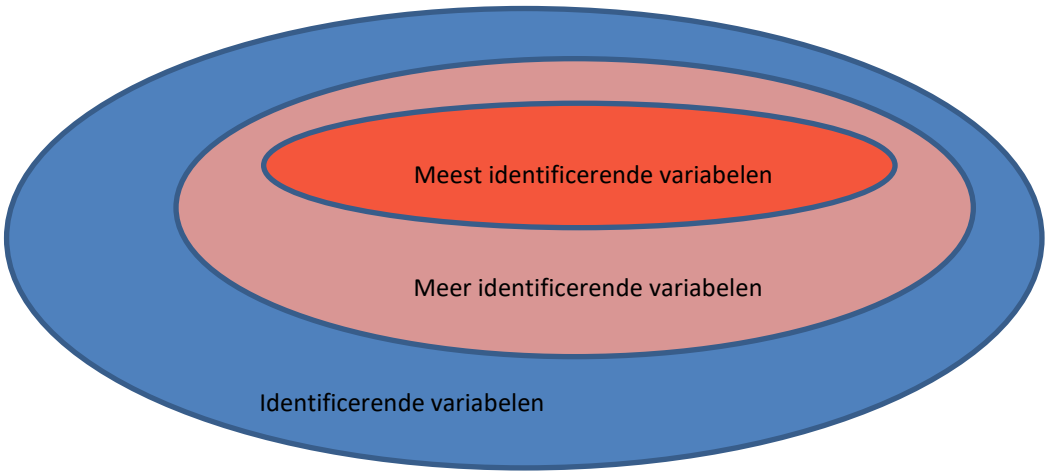
Daarnaast gelden voor microdatabestanden onder contract nog vier regels: de Zeldzaamheidsregel, de Digitsregel, de Regioregel en de Panelregel. Microdatabestanden onder contract moeten aan alle regels voldoen.
In de volgende paragrafen gaan we in op iedere specifieke regel.

6.4.1 M1: de Zeldzaamheidsregel

M1	In een record mag geen waardencombinatie voorkomen die minder dan *M* keer in de *populatie* voorkomt, waarbij waardencombinaties worden bekeken op kruisingen van variabelen van het type meest identificerend × meer identificerend × identificerend

Identificerende variabelen worden ingedeeld in drie geneste klassen: meest identificerend, meer identificerend en identificerend. Meest identificerende variabelen

zijn dus ook meer identificerend en meer identificerende variabelen zijn ook identificerend. Onderstaand figuur geeft dit schematisch weer.



De mate waarin een variabele identificerend is, wordt bepaald door de zeldzaamheid, de zichtbaarheid en de zoekbaarheid van de variabele; de zogenaamde 3Z-criteria.
Naarmate een variabele meer aan de 3Z-criteria voldoet, wordt die variabele een grotere identificerende kracht toegekend.
Een waarde van een variabele die weinig in de populatie voorkomt, wordt **zeldzaam** genoemd. Een variabele met een zeldzame categorie wordt gemakshalve een zeldzame variabele genoemd. Een voorbeeld is nationaliteit: de waarde “Samoaans” is zeldzaam in Nederland. Maar de waarde “Nederlands” is dat uiteraard niet. Toch wordt nationaliteit dan een zeldzame variabele genoemd.
Een waarde van een variabele wordt **zichtbaar** genoemd als van een persoon bekend is, of makkelijk kan worden waargenomen, dat hij die waarde bezit. Wederom wordt een variabele met minimaal één zichtbare categorie ook een zichtbare variabele genoemd. Een voorbeeld van een zichtbare variabele is geslacht.
Een waarde van een variabele wordt **zoekbaar** genoemd wanneer de personen die deze waarde hebben eenvoudig te traceren zijn. Ook de desbetreffende variabele noemen we dan weer zoekbaar. Een voorbeeld van een zoekbare variabele is woonregio. Als we weten dat een persoon in Overijssel woont, dan wordt het gebied waarin we deze persoon moeten zoeken meteen sterk verkleind, wat het zoeken een stuk makkelijker maakt.
In de volgende tabel geven we voor een aantal variabelen een mogelijke score op de 3Z-criteria:

Variabele	Zeldzaam	Zichtbaar	Zoekbaar
Regio	+	++	++
Geslacht		++	
Herkomst	++	++	
Nationaliteit	++	++	
Geboorteland	++	++	
Soort bedrijf	++	+	+
Beroep	++	+	+
Opleiding	++		+
Leeftijd	+	+	
Burgerlijke staat		+	
Huishoudenssamenstelling	+	+	

+ = hoog scorend, ++ = zeer hoog scorend

Op basis van deze tabel wordt bijvoorbeeld Regio als Meest identificerend, Nationaliteit als Meer identificerend en Burgerlijke staat als Identificerend geclassificeerd.
Bij het bepalen van de identificerende variabelen voor het toepassen van de Zeldzaamheidsregel, moeten ook variabelen meegenomen worden die de doelpopulatie van het bestand definiëren. Zo zal voor een bestand met alleen inwoners van Drenthe ook de variabele Regio meegenomen moeten worden en moeten in feite alle vierdimensionale kruisingen “Drenthe × meest identificerend × meer identificerend × identificerend” gecontroleerd worden.
De zeldzaamheidsregel is geformuleerd op het niveau van de populatie: een waardencombinatie moet minimaal *M* keer in de *populatie* voorkomen. Die populatie-informatie is niet altijd voor alle identificerende variabelen bekend. In dat geval zal de Zeldzaamheidsregel getoetst moeten worden met behulp van het steekproefbestand en zal het minimale aantal *M* in de populatie vertaald moeten worden naar een aantal voor het betreffende bestand. De te gebruiken grenswaarde hangt dan af van de bestandsgrootte relatief ten opzichte van de grootte van de (doel)populatie.
NB: de doelpopulatie hoeft niet altijd de totale Nederlandse bevolking te zijn.

6.4.2 M2: de Digitsregel

M2	Het maximale detailleringsniveau voor beroep, bedrijf en opleiding wordt bepaald door de meest gedetailleerde aanduiding voor woon-, werk- of opleidingsregio.

Deze regel is in feite ondersteunend aan de zeldzaamheidsregel. Als aan de Digitsregel wordt voldaan, zal het makkelijker zijn om aan de Zeldzaamheidsregel te kunnen voldoen.
Wil men bijvoorbeeld een microdatabestand maken met een (directe) woonregio-indeling met per woonregio ten minste *N* inwoners, dan mag men de gegevens over bedrijf, beroep en opleiding op de twee minst gedetailleerde niveaus geven. Daarnaast moet nog per woonregio met behulp van de Zeldzaamheidsregel worden onderzocht of er nog zeldzame waarden voor deze variabelen voorkomen.

6.4.3 M3: de Regioregel

M3	Elke categorie van een regio-aanduiding moet ten minste *X* inwoners in de *doelpopulatie* bevatten.

Ook deze regel is in feite ondersteunend aan de zeldzaamheidsregel. Als aan de Regioregel wordt voldaan, zal het makkelijker zijn om aan de Zeldzaamheidsregel te kunnen voldoen.

6.4.4 M4: de Panelregel

M4	In een panelbestand mag geen woon-, werk- en opleidingsregio voorkomen.

Deze regel geldt uiteraard alleen voor een beoogd microdatabestand onder contract met panelinformatie over de eenheden. In een panelbestand staat informatie van een herhaald onderzoek onder eenzelfde groep eenheden. Dit vergroot de kans op identificatie aanzienlijk omdat bijzondere combinaties van scores op variabelen op verschillende tijdstippen door dezelfde eenheid zeldzamer zullen zijn dan één enkele score. Zo zal iemand die op tijdstip t in Groningen woont en op tijdstip $t+1$ in Middelburg, makkelijker te identificeren zijn dan iemand waarvan alleen bekend is dat die op tijdstip t in Groningen woont.
Merk op dat de Panelregel impliceert dat er geen informatie over verhuizingen uit de paneldata af te leiden mag zijn.

Quick reference voor beveiliging van publicatiebestanden

Regel [pagina]	Korte beschrijving
P0 [47]	Directe identificatoren mogen **niet** in het bestand voorkomen.
P1 [47]	Er moet ten minste **één** jaar verstreken zijn tussen de beëindiging van het veldwerk en de levering van het bestand.
P2 [47]	P2.1 Directe regioaanduidingen mogen **niet** in het bestand voorkomen. P2.2 In het bestand is maximaal **één** type regio toegestaan. De betreffende regiovariabele moet voldoende gespreid zijn. P2.3 Er mogen maximaal **15** identificerende variabelen in het bestand voorkomen. P2.4 Er mogen **geen** "gevoelige" gegevens in het bestand voorkomen.
P3 [48]	Het moet uitgesloten zijn dat ophooggewichten gebruikt kunnen worden om additionele informatie met betrekking tot identificerende variabelen af te leiden.
P4 [48]	P4.1 Er mogen geen categorieën van identificerende variabelen voorkomen die minder dan ***K*** personen in de populatie betreffen. P4.2 Er mogen geen waardencombinaties van twee gekruiste identificerende variabelen voorkomen die minder dan ***L*** personen in de populatie betreffen.
P5 [49]	Ieder record van een meerpersoonshuishouden dat met ten minste twee records in het bestand zit, moet op de kruising van alle huishoudensvariabelen een waardencombinatie bevatten die bij ten minste ***H*** huishoudens in het bestand voorkomt.
P6 [49]	De volgorde van de records in het bestand moet willekeurig zijn.

	Variabelen die de deelpopulatie van het bestand definiëren moeten in bovenstaande regels als identificerende variabelen worden meegenomen.

Voldoende spreiding bij sub-regel P2.2 van regel P2 betekent:

- Iedere categorie van de indirecte regioaanduiding komt in minimaal **6** provincies voor;

èn

- Geen enkele gemeente in een categorie van de indirecte regioaanduiding mag meer dan **50%** van het totaal aantal inwoners binnen die categorie van de regioaanduiding bevatten.

Quick reference voor beveiliging van microdatabestanden onder contract

Regel [pagina]	Korte beschrijving
M0 [49]	Directe identificatoren mogen **niet** in het bestand voorkomen.
M1 [49]	In een record mag geen waardencombinatie voorkomen die minder dan ***M*** keer in de *populatie* voorkomt, waarbij waardencombinaties worden bekeken op kruisingen van variabelen van het type meest identificerend × meer identificerend × identificerend
M2 [51]	Het maximale detailleringsniveau voor beroep, bedrijf en opleiding wordt bepaald door de meest gedetailleerde aanduiding voor woon-, werk- of opleidingsregio
M3 [52]	Elke categorie van een regio-aanduiding moet ten minste ***X*** inwoners in de *doelpopulatie* bevatten.
M4 [52]	In een panelbestand mag **geen** woon-, werk- en opleidingsregio voorkomen.

	Variabelen die de deelpopulatie van het bestand definiëren moeten in bovenstaande regels als identificerende variabelen worden meegenomen.

7. Beveiliging van analyseresultaten

7.1 Inleiding

In hoofdstukken 4 en 5 zijn enkele regels geformuleerd voor het beoordelen van kwantitatieve tabellen en frequentietabellen op onthulling van gegevens over individuele eenheden of herkenbare groepen van eenheden. Bij het beoordelen van resultaten van analyses spelen vergelijkbare problemen een rol. Het aantal mogelijke analyses is echter oneindig groot, waardoor het onmogelijk is om alle analyseresultaten te behandelen. In dit hoofdstuk zullen we enkele richtlijnen geven voor de beveiliging van bepaalde analyseresultaten.
Ook internationaal zijn er op het gebied van de beveiliging van analyseresultaten geen expliciete regels. Aan de andere kant wordt remote access, on-site en ASD steeds populairder. Bij dergelijke faciliteiten moeten analyseresultaten gecontroleerd worden op onthulling, voordat ze de beveiligde omgeving mogen verlaten. Die controle wordt vaak nog op ad-hoc basis door (inhoudelijke) experts uitgevoerd.
De informatie in het huidige hoofdstuk is beperkt. Voor meer informatie verwijzen we naar richtlijnen zoals genoemd in de guidelines for outputchecking (Bond et al., 2013) en naar de richtlijnen voor remote access output zoals opgesteld door Microdata services (zie **https://www.cbs.nl/-/media/cbs/onze-diensten/maatwerk/zelf-onderzoek-doen/richtlijnen-voor-ra-output.pdf**).

7.2 Onthullingsrisico bij analyseresultaten

Analyse resultaten kunnen (niet uitputtend) ingedeeld worden in

- Samenvattende statistieken, zoals maximum, minimum, gemiddelde, standaard deviatie, etc.
- Parametrische modelschattingen, zoals lineaire regressie, log-lineaire modellen, correlaties, ANOVA, t-toets, etc.
- Niet-parametrische modelschattingen, zoals cluster analyse, niet-parametrische regressie, principale componentanalyse, $\chi 2$-toets, etc.
- Grafische uitvoer, zoals boxplots, scatterplots, heatmaps, etc.

Soms is het overduidelijk dat het gaat om informatie over individuele eenheden of (kleine) groepen van eenheden. Zo is het maximum vaak gerelateerd aan één individu. Maar denk ook aan uitbijters of aan een gemiddelde met een zeer kleine (bijna nul) variantie.
Analyseresultaten kunnen ook gebaseerd zijn op identificeerbare (groepen) van eenheden: in veel gevallen zal de uitkomst van een analyse een uitspraak doen over een doelvariabele op basis van een aantal verklarende variabelen. Zo is bijvoorbeeld een log-lineaire analyse in feite een analyse van de structuur van een frequentietabel.
Bij grafische uitvoer ligt het wat genuanceerder. Dergelijke uitvoer is vaak een bijproduct van een statistische analyse. De manier waarop de grafische uitvoer gepresenteerd wordt speelt ook een belangrijke rol. Denk aan de mate van gedetailleerdheid van de gebruikte schaal. Dit kan, in combinatie met “opvallende” eenheden (uitbijters, maximum, …) een redelijk accurate schatting opleveren van een gegeven over een individuele eenheid.
Daarnaast speelt nog een ander aspect een rol bij grafische uitvoer, met name bij het elektronisch beschikbaar stellen van die uitvoer. Zo kan bij een aantal statistische pakketten grafische uitvoer niet alleen uit een puur grafische weergave bestaan (een

plaatje zoals een bitmap), maar kan het ook de onderliggende data bevatten. Denk aan een staafdiagram waarbij je de exacte waarde van de hoogte van de staaf te zien krijgt als je er met de muis overheen gaat. In dat geval zou de grafische uitvoer directe informatie over individuele (groepen van) eenheden kunnen bevatten. De richtlijn is dan ook om grafische uitvoer alleen als puur grafische weergave ter beschikking te stellen en niet als een uitvoerbestand van de gebruikte (statistische) software.

7.3 Richtlijnen

De eerste vraag die gesteld zou moeten worden, is in hoeverre de analyseresultaten "bruikbare" informatie zouden vrijgeven over herkenbare (groepen van) individuele eenheden. Dat is niet eenvoudig. Zo zijn regressiecoëfficiënten op zichzelf vaak geen informatie over herkenbare (groepen van) individuele eenheden. De coëfficiënten kunnen via het regressiemodel uiteraard wel een mogelijk zeer goede schatting geven van de score van een herkenbare individuele eenheid op de afhankelijke variabele. In speciale gevallen kunnen geschatte coëfficiënten van twee regressieanalyses informatie onthullen over een individuele eenheid. Denk daarbij een regressie A op alle waarnemingen en een regressie B op alle waarnemingen minus één (bijvoorbeeld minus een uitbijter). Zie Ritchie (2011) voor meer informatie over mogelijke onthulling bij regressieanalyses.

Indien analyseresultaten geïnterpreteerd kunnen worden als informatie over herkenbare (groepen van) individuele eenheden uit een frequentietabel, dan kunnen de regels uit hoofdstuk 5 gebruikt worden. Dit is bijvoorbeeld het geval wanneer een frequentietabel als input gebruikt kan worden (denk aan log-lineaire regressie), maar ook wanneer de verklarende variabelen van een analyse herkenbare (groepen van) individuele eenheden definiëren. De verklarende variabelen zijn dan in feite de opspanvariabele van de frequentietabel en het analyseresultaat geeft dan informatie over die herkenbare (groepen) van individuen.

Bij samenvattende statistieken als uitbijters, maxima, minima en dergelijke, mogen de analyseresultaten ook niet over herkenbare (groepen van) individuele eenheden gaan.

Grafische uitvoer mag alleen in de vorm van pure plaatjes ter beschikking gesteld worden. De onderliggende data mag op geen enkele manier met de grafische uitvoer meekomen. De grafische uitvoer moet voldoende onzekerheid geven over de herkenbare (groepen van) individuele eenheden. Dat kan bijvoorbeeld door een niet te nauwkeurige schaalverdeling te gebruiken.

In de gevallen die niet "eenvoudig" onder de eerdere hoofdstukken vallen zal een beveiligingsexpert, samen met een inhoudelijke expert, een advies moeten geven over mogelijk onthullingsrisico. Daarvoor kan bijvoorbeeld contact opgenomen worden met (leden van) de expertgroep statistische beveiliging van het CBS (zie **Expert groep**).

Referenties

Bond, S., Brandt, M. and de Wolf, P.P. (2013). *Guidelines for the checking of output based on microdata research*, resultaat van een project binnen het zevende kader programma van de Europese Unie (FP7/2007-2013), onder grant agreement nummer 262608 (DwB, Data without Boundaries), **https://cros-legacy.ec.europa.eu/system/files/dwb_standalone-document_output-checking-guidelines.pdf**.

Ritchie, F. (2011). *Disclosure control for regression outputs.* WISERD Data Resources, WISERD/WDR/005, **https://wiserd.ac.uk/wp-content/uploads/WISERD_WDR_005.pdf**.

Quick reference voor beveiliging van analyseresultaten

Richtlijnen [pagina]	**Toelichting**
Guidelines for output checking [55]	Richtlijnen voor beveiliging van analyseresultaten, resultaat van Europese projecten
Richtlijnen RA output [55]	Richtlijnen voor beveiliging van remote access output, opgesteld door Microdata services
	Belangrijkste **vuistregels** uit de RA richtlijnen: • Geen microdata • Tabellen en soortgelijke output minimaal 10 eenheden (ongewogen) als grondslag voor elke cel of datapunt • Geen maximum of minimum • Modellen hebben minimaal 10 vrijheidsgraden • Geen groepsonthulling in frequentie tabellen • Geen dominantie in kwantitatieve tabellen