Statistische beveiliging

10

Anco Hundepool en Peter-Paul de Wolf

Statistische Methoden (201014)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2010

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader voorlopig cijfer
x	=	geheim
–	=	nihil
–	=	(indien voorkomend tussen twee getallen) tot en met
0 (0,0)	=	het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	=	een cijfer kan op logische gronden niet voorkomen
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

Omslag
TelDesign, Rotterdam

Inlichtingen
Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen
E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

60165201014 X-37

Inhoudsopgave

1. Statistische Beveiliging

1.1 Algemene beschrijving

Bij het publiceren van statistische informatie moet het CBS een afweging maken tussen de belangen van zijn berichtgevers en de behoeften van zijn gebruikers. Aan de ene kant willen de gebruikers van het CBS zo veel en zo gedetailleerd mogelijke informatie. Aan de andere kant eisen de berichtgevers (personen en bedrijven, en anders wel de houders van registraties en het College Bescherming Persoonsgegevens) dat hun privacy wordt gewaarborgd. *Private lives and public policies: confidentiality and accessibility of government statistics* (Duncan et al., 1993), is dan ook de treffende titel van een Amerikaans boek over deze problematiek.

Wat het CBS wèl en niet mag publiceren, volgt uit het statistische beveiligingsbeleid van het CBS, zoals vastgelegd in het Handboek Statistische Beveiliging (Hundepool et al., 2006). Onder statistische beveiliging verstaan we hier het voorkómen dat er inhoudelijke conclusies over herkenbare eenheden kunnen worden getrokken op basis van gepubliceerd of anderszins beschikbaar gesteld CBS-materiaal.

Uit de statistische publicaties van het CBS (StatLine-tabellen, web-artikelen, persberichten, wetenschappelijke artikelen) 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.

In het Handboek Statistische Beveiliging staan de (beleids)regels beschreven waaraan individuele publicaties moeten voldoen. Niet alle publicaties voldoen daar vanzelf aan. Integendeel, vaak zal een publicatie “beveiligd” moeten worden. Voor de beveiliging van microdata, van tabeldata en analyseresultaten zijn verschillende methoden beschikbaar. Het thema Statistische Beveiliging in de Methodenreeks kan dan ook opgesplitst worden in een aantal deelthema’s:

- Statistische beveiligingsmethoden voor microdata,
- Statistische beveiligingsmethoden voor kwantitatieve tabellen,
- Statistische beveiligingsmethoden voor frequentietabellen,
- Statistische beveiligingsmethoden voor analyseresultaten.

De tegengestelde belangen van privacy-bescherming en informatiebehoud speelt continu een rol bij statistische beveiliging. Bij het toepassen van de verschillende methoden voor statistische beveiliging, moeten beide aspecten dan ook worden meegenomen. Het statistische beveiligingsbeleid van het CBS schrijft een minimaal beveiligingsniveau voor. Het is de kunst van de beveiliger om op een zodanige manier verschillende beveiligingsmethoden toe te passen dat het minimaal vereiste beveiligingsniveau gehaald wordt en het informatieverlies daarbij zo klein mogelijk is. Dit zal voor iedere situatie anders zijn: het begrip “informatieverlies” kan voor verschillende gebruikers van de CBS-gegevens immers een verschillende betekenis hebben.

De in dit thema genoemde methoden worden ieder afzonderlijk beschreven. Echter, in de praktijk zullen per situatie vaak meerdere methoden tegelijk gebruikt worden om tot een “veilige” publicatie te komen. De interactie van de verschillende methoden op elkaar bij gelijktijdig gebruik wordt niet in deze Methodenreeks beschreven.

1.2 Afbakening en relatie met andere thema’s

Het beleid aangaande statistische beveiliging zal hier niet worden beschreven. Dat beleid is vastgelegd in het genoemde Handboek Statistische Beveiliging. Wel zullen verschillende methoden worden beschreven die voorhanden zijn om dat beleid op CBS-publicaties te kunnen toepassen. Sommige van de te beschrijven methoden worden actief toegepast op het CBS, andere methoden worden vooralsnog alleen op statistische bureaus in het buitenland gebruikt.

Bij het toepassen van statistische beveiligingsmethoden moet zowel het beveiligingsniveau als het informatieverlies van de publicatie bekeken worden. Aangezien het begrip “informatie” subjectief is en dus per gebruiker verschillend ingevuld kan worden (zelfs bij één publicatie), is het niet mogelijk om voor iedere specifieke situatie een bepaalde methode voor te schrijven. De methoden zullen daarom beschreven worden samen met hun voor- en nadelen en hun effecten op beveiligingsniveau en informatieverlies. Een CBS-medewerker die belast is met de statistische beveiliging van een publicatie (in welke vorm dan ook), zal vervolgens zelf de meest geschikte methode voor zijn of haar publicatie moeten kiezen.

1.3 Plaats in het statistisch proces

Statistische Beveiliging vindt traditioneel plaats aan het einde van het statistisch proces: vlak voor publicatie (in wat voor vorm dan ook) wordt de statistische beveiliging toegepast. Idealiter zou tijdens het hele statistische proces al rekening gehouden moeten worden met het feit dat de uiteindelijke publicatie aan het statistische beveiligingsbeleid zal moeten voldoen. Maar denk bijvoorbeeld ook aan maatregelen die aan het begin van het statistisch proces al genomen kunnen worden, zoals de formulering van de aanschrijfbrief voor deelname aan een enquête (“informed consent”).

Het concept “statistische beveiliging” speelt dus een rol tijdens het gehele statistische proces. De specifieke methoden zoals beschreven in het huidige document worden echter pas aan het einde van het statistisch proces toegepast, vlak voor publicatie.

1.4 Definities

Begrip	Omschrijving
μ-ARGUS	Software voor statistisch beveiligen van microdatabestanden
τ-ARGUS	Software voor statistisch beveiligen van tabellen
Identificerende variabele	Variabele waarvan de waarde kan bijdragen tot identificatie van een afzonderlijk persoon, huishouden, onderneming of instelling
Onthullen	Het uit statistische gegevens informatie achterhalen over een herkenbare afzonderlijke persoon, huishouden, onderneming of instelling
Primair onveilige cel	Cel in een tabel die niet aan de beveiligingsregels voldoet
Secundair onveilige cel	Cel in een tabel die wel aan de beveiligingsregels voldoet, maar onderdrukt moet worden om primair onveilige cellen te beveiligen
Structurele nulcel	Een cel waar van het algemeen bekend is dat die cel (op logische gronden) geen bijdragen *kan* hebben

Een uitgebreide Nederlandstalige verklarende woordenlijst (glossary) voor statistische beveiliging is te vinden op de intranetsite van het CBS: http://intranet/dmk/Methodologie/glossary.asp.

Een internationale (Engelstalige) versie is te vinden op: http://neon.vb.cbs.nl/casc/Glossary.htm.

1.5 Literatuur

Duncan, G.T., Jabine, T.B. and V.A. de Wolf (Eds.) (1993), *Private lives and public policies: confidentiality and accessibility of government statistics.* The National Academies Press, ISBN 0309086515.

Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt, E. en De Wolf, P.P. (2006), *Handboek Statistische Beveiliging*, BPA 21-06-TMO.

2. Statistische Beveiliging van Microdata

2.1 Algemene beschrijving en leeswijzer

2.1.1 Algemene beschrijving

Onder statistische beveiliging van microdata verstaan we het creëren van microdata die voldoen aan het beveiligingsbeleid van het CBS en als zodanig het CBS mogen verlaten. Het gaat dus nadrukkelijk niet om microdatabestanden die op het CBS blijven, waaronder bestanden voor on-site en remote access. Het beveiligingsbeleid voor microdata is vastgelegd in hoofdstuk 3 van het Handboek Statistische Beveiliging (Hundepool et al., 2006).

De methoden die in dit deelthema zullen worden beschreven, zijn te gebruiken om beveiligde microdatabestanden te maken. De mate waarin (of de strengheid waarmee) de methoden worden toegepast is mede afhankelijk van het type bestand dat vrijgegeven gaat worden. Dit is uitvoerig beschreven in het Handboek Statistische Beveiliging, waar onderscheid is gemaakt tussen Publicatiebestanden en Microdatabestanden onder Contract.

De in dit deelthema beschreven methoden kunnen door middel van het software pakket μ-ARGUS eenvoudig worden toegepast. Dit pakket is door DMV in Europees verband ontwikkeld.

2.1.2 Leeswijzer

Als eerste stap bij de statistische beveiliging van microdata zal bepaald moeten worden of er onthulling mogelijk is: is er informatie in de microdata aanwezig die niet over individuele respondenten onthuld mag worden? Meestal betreft dit “gevoelige” informatie over respondenten die in de microdata zijn te herkennen als uniek of zeldzaam geval. Dergelijke respondenten zullen beschermd moeten worden.

Een aantal beveiligingsmethoden die we zullen behandelen, is toepasbaar op categoriale variabelen: globaal hercoderen (paragraaf 2.3) en PRAM (paragraaf 2.7). Top (en bottom) coding is vooral bedoeld voor continue variabelen, zie paragraaf 2.5. Lokaal onderdrukken (paragraaf 2.4) kan zowel bij categoriale als continue variabelen gebruikt worden. Daarnaast is er nog de mogelijkheid om ruis toe te voegen aan (ophoog)gewichten, zie paragraaf 2.6.

Welke (combinatie van) methode(n) uiteindelijk in een specifieke situatie gebruikt wordt, is niet op voorhand vast te leggen. De afdeling die verantwoordelijk is voor de constructie van het microdatabestand, is ook verantwoordelijk voor een adequate (statistische) beveiliging. Bij de keuze voor de te gebruiken methode(n) moeten twee (concurrerende) aspecten worden meegenomen:

- onthullingrisico,
- informatieverlies.

Algemeen kan gezegd worden, dat verkleining van het onthullingrisico tot meer informatieverlies zal leiden. Ook het omgekeerde is waar: hoe kleiner het informatieverlies, hoe groter het onthullingrisico. In voorkomende gevallen zal een afweging gemaakt moeten worden, waarbij uiteraard wel altijd minimaal aan de regels in het Handboek Statistische Beveiliging (Hundepool et al., 2006) voldaan moet worden.

2.2 Afbakening en relatie met andere (deel)thema's

De methoden die in dit deelthema zullen worden beschreven, worden direct op de microdata zelf toegepast. Daarbij kunnen verschillende niveaus van beveiliging ontstaan die corresponderen met die van Publicatiebestanden of van Microdatabestanden onder Contract.

Gebruikers van onbeveiligde bestanden (waaronder bestanden voor on-site en remote access) en van Microdatabestanden onder Contract, kunnen output genereren die niet noodzakelijk aan het beveiligingsbeleid van het CBS voldoet. In die situaties zullen (ook) andere methoden gebruikt moeten worden om de output veilig te maken. Voor dergelijke methoden, zie het deelthema "Statistische Beveiliging van Analyseresultaten".

2.3 Globaal hercoderen

2.3.1 Korte beschrijving

Bij de statistische beveiliging van microdatabestanden die het CBS verlaten, wordt voornamelijk naar de variabelen gekeken aan de hand waarvan een respondent mogelijk geïdentificeerd kan worden. Dit soort variabelen worden *identificerende* variabelen genoemd. Identificerende variabelen zijn over het algemeen categoriale variabelen. Combinaties van categorieën van identificerende variabelen kunnen al snel tot unieke of zeldzame personen leiden. Denk bijvoorbeeld aan "Burgemeester in Amsterdam" (uniek) of "Vrouwelijke neurochirurg van boven de 55 jaar uit Staphorst" (zeldzaam). In de regels voor Microdatabestanden onder Contract (zie hoofdstuk 3 uit het Handboek Statistische Beveiliging (Hundepool et al., 2006)) staat dat dergelijke combinaties voldoende vaak in de (doel)populatie voor moeten komen.

Door samenvoeging van categorieën van identificerende variabelen, kunnen zeldzame combinaties minder zeldzaam gemaakt worden.

2.3.2 Toepasbaarheid

Bij de beveiliging van microdatabestanden die het CBS verlaten, moeten bepaalde combinaties van identificerende variabelen voldoende vaak in de (doel)populatie voor komen. Met name wanneer een identificerende variabele zeer gedetailleerd in

het bestand voorkomt, kan door middel van globaal hercoderen het bestand vaak voldoende beveiligd worden, terwijl het informatieverlies beperkt kan blijven.

Voor sommige onderzoekers zal globale hercodering echter te veel detail weghalen, waardoor zij hun analyses niet meer uit kunnen voeren. Het is dan ook aan de medewerker van het CBS die is belast met het statistisch beveiligen van het bestand, om in te schatten of een globale hercodering voor het onderhavige geval een geschikte beveiligingsmethode is.

Globale hercodering hoeft overigens niet beperkt te blijven tot identificerende variabelen. Ook niet-identificerende variabelen kunnen globaal gehercodeerd worden. Uiteraard moeten dit wel categoriale variabelen zijn. Bij een dergelijke toepassing zou, bij eventuele identificatie van een respondent, slechts minder gedetailleerde (en daardoor wellicht algemeen bekende) informatie onthuld worden.

2.3.3 Uitgebreide beschrijving

Bij globale herodering wordt de codelijst van een (identificerende) variabele aangepast. In het geval dat het om een hiërarchische variabele gaat (zoals bijvoorbeeld regio) is een voor de hand liggende hercodering het verwijderen van (enkele) detailniveaus. Zo zou bij een hercodering van Woonplaats alle gemeentes vervangen kunnen worden door de bijbehorende provincie.

Nadat een codelijst van een variabele is aangepast, wordt bij *ieder* record de score op die variabele aangepast aan de nieuwe codelijst. Dus niet alleen bij de onveilige records, maar ook bij de veilige records.

2.3.4 Voorbeeld

In Figuur 1 zijn een aantal records uit een fictief microdatabestand weergegeven. Om eenvoudig naar de juiste records te kunnen verwijzen zijn de records genummerd.

	Beroep	Woonplaats	Geslacht	Opleiding	…
1	Burgemeester	Amsterdam	Man	Hoog	…
2	Visser	Urk	Man	Laag	…
3	Docent	Amsterdam	Vrouw	Hoog	…
4	Loodgieter	Papendrecht	Man	Middel	
	…	…	…	…	…

Figuur 1: Enkele records uit een fictief microdatabestand

De burgemeester uit Amsterdam is uiteraard uniek. De variabele “Woonplaats” wordt nu globaal gehercodeerd door de plaatsnamen te vervangen door de bijbehorende provincie. Dit levert de records zoals gegeven in Figuur 2.

	Beroep	Woonplaats	Geslacht	Opleiding	…
1	Burgemeester	Noord-Holland	Man	Hoog	…
2	Visser	Flevoland	Man	Laag	…
3	Docent	Noord-Holland	Vrouw	Hoog	…
4	Loodgieter	Zuid-Holland	Man	Middel	
	…	…	…	…	…

Figuur 2: Records uit Figuur 1 na globale hercodering van "Woonplaats"

Nu is in record 1 de burgemeester niet langer uniek. Aangezien de hercodering *globaal* wordt toegepast, is de woonplaats in de veilige records 2 tot en met 4 ook aangepast.

2.4 Lokaal onderdrukken

2.4.1 Korte beschrijving

Bij de statistische beveiliging van microdatabestanden die het CBS verlaten, wordt voornamelijk naar de variabelen gekeken aan de hand waarvan een respondent mogelijk geïdentificeerd kan worden. Dit soort variabelen worden *identificerende* variabelen genoemd. Identificerende variabelen zijn over het algemeen categoriale variabelen. Combinaties van categorieën van identificerende variabelen kunnen al snel tot unieke of zeldzame personen leiden. Denk bijvoorbeeld aan "Burgemeester in Amsterdam" (uniek) of "Vrouwelijke neurochirurg van boven de 55 jaar uit Staphorst" (zeldzaam). In de regels voor Microdatabestanden onder Contract (zie hoofdstuk 3 uit het Handboek Statistische Beveiliging (Hundepool et al., 2006)) staat dat dergelijke combinaties voldoende vaak in de (doel)populatie voor moeten komen.

Bij lokaal onderdrukken wordt van minimaal één van de variabelen in een combinatie die onvoldoende vaak in de (doel)populatie voorkomt de score onderdrukt (of de score "Onbekend" toegekend). Daardoor beschrijft de combinatie van de overgebleven variabelen een (mogelijk) grotere groep in de (doel)populatie.

2.4.2 Toepasbaarheid

Bij de beveiliging van microdatabestanden die het CBS verlaten, moeten bepaalde combinaties van identificerende variabelen voldoende vaak in de (doel)populatie voorkomen. Met name wanneer een identificerende variabele zeer gedetailleerd in het bestand voorkomt, kan door middel van lokaal onderdrukken het bestand vaak voldoende beveiligd worden, terwijl het informatieverlies beperkt kan blijven.

Lokaal onderdrukken wordt vaak als laatste beveiligingsmethode gebruikt. De meeste beveiliging is dan al aangebracht door andere methoden en lokaal onderdrukken wordt gebruikt om de laatste onveilige records te beveiligen.

Door lokale onderdrukking ontstaan er ontbrekende waarden in het bestand. De manier waarop die ontbrekende waarden worden gekozen, is echter absoluut niet aselect: het is immers bedoeld om records die tot kleine, identificeerbare groepen beho-

ren te beschermen. Het effect van deze ontbrekende waarden op uit te voeren analyses is dan ook anders bij dan ontbrekende waarden ten gevolge van non-response.

Lokaal onderdrukken hoeft overigens niet beperkt te blijven tot identificerende variabelen. Ook niet-identificerende variabelen kunnen lokaal onderdrukt worden. Daardoor zou, bij eventuele identificatie van een respondent, geen gevoelige informatie onthuld kunnen worden.

2.4.3 Uitgebreide beschrijving

Bij lokaal onderdrukken wordt de waarde van een (identificerende) variabele op "Onbekend" gezet. Volgens de beveiligingsregels uit het Handboek Statistische Beveiliging (Hundepool et al., 2006), moeten combinaties van identificerende variabelen voldoende vaak in de (doel)populatie voorkomen. Door bij minimaal één variabele uit zo'n combinatie de score te onderdrukken, ontstaat in feite een lager dimensionale combinatie. Het gevolg daarvan is, dat de combinatie (waarschijnlijk) een grotere groep respondenten in de (doel)populatie zal beschrijven.

Lokaal onderdrukken wordt alleen op onveilige records toegepast. Het is mogelijk dat in één record meerdere onveilige combinaties van identificerende variabelen voorkomen. Door op een slimme manier de juiste variabele(n) te onderdrukken, kunnen soms meerdere onveilige combinaties tegelijk beveiligd worden.

Indien een microdatabestand meerdere records bevat van personen uit hetzelfde huishouden, moet daar bij lokaal onderdrukken rekening mee gehouden worden. Dergelijke records kunnen zogenaamde huishoudvariabelen bevatten. Dit zijn variabelen waarbij ieder lid van het huishouden dezelfde score zal hebben. Denk daarbij bijvoorbeeld aan huishoudinkomen, huishoudgrootte en woonplaats. Wanneer nu bij (minimaal) één persoon van een huishouden een onveilige combinatie voorkomt met daarin een huishoudvariabele en die huishoudvariabele wordt lokaal onderdrukt, dan moet die variabele voor alle personen in dat huishouden onderdrukt worden. In dat geval kan het dus zijn dat ook in veilige records waarden worden onderdrukt.

De keuze uit een zeldzame combinatie van scores op identificerende variabelen van de variabele die onderdrukt gaat worden, is in principe vrij. Binnen μ-ARGUS zijn echter twee opties mogelijk.

Ten eerste kan de gebruiker door middel van het toekennen van gewichten aan variabelen aangeven in welke mate het onderdrukken van de score op die variabele al dan niet gewenst is. μ-ARGUS kiest dan vervolgens die variabelen te onderdrukken, waarvan de som van de gewichten zo klein mogelijk is. Daardoor is het mogelijk om bijvoorbeeld variabelen die al door andere beveiligingsmethoden zijn aangepast enigszins te ontzien bij het lokaal onderdrukken.

Bij de tweede optie gebruik μ-ARGUS een soort entropieargument om de te onderdrukken variabele(n) te kiezen. Iedere variabele krijgt dan het volgende gewicht toegewezen:

$$w_X = -\sum_{i=1}^{K_X} \frac{f_X(i)}{n} \log \frac{f_X(i)}{n}, \qquad (2.4.1)$$

waarbij K_X het aantal categorieën van variabele X is, n het aantal records in het microdatabestand en $f_X(i)$ het aantal records met score i op variabele X. Daardoor worden variabelen met grotere aantallen categorieën minder snel onderdrukt dan variabelen met slechts een paar categorieën.

2.4.4 Voorbeeld

In Figuur 3 zijn een aantal records uit een fictief microdatabestand weergegeven. Om eenvoudig naar de juiste records te kunnen verwijzen zijn de records genummerd.

	Beroep	Woonplaats	Geslacht	Opleiding	…
1	Burgemeester	Amsterdam	Man	Hoog	…
2	Visser	Urk	Man	Laag	…
3	Docent	Amsterdam	Vrouw	Hoog	…
4	Loodgieter	Papendrecht	Man	Middel	
	…	…	…	…	…

Figuur 3: Enkele records uit een fictief microdatabestand

De burgemeester uit Amsterdam is uiteraard uniek. De variabele “Woonplaats” wordt nu lokaal onderdrukt door de plaatsnaam te vervangen door de score “Onbekend” in de onveilige records. Dit levert de records zoals gegeven in Figuur 4.

	Beroep	Woonplaats	Geslacht	Opleiding	…
1	Burgemeester	Onbekend	Man	Hoog	…
2	Visser	Urk	Man	Laag	…
3	Docent	Amsterdam	Vrouw	Hoog	…
4	Loodgieter	Papendrecht	Man	Middel	
	…	…	…	…	…

Figuur 4: Records uit Figuur 3 na lokale onderdrukking van Woonplaats

Nu is in record 1 de burgemeester niet langer uniek. Aangezien de onderdrukking *lokaal* wordt toegepast, is de woonplaats in de veilige records 2 tot en met 4 niet onderdrukt.

2.5 Top-coding

2.5.1 Korte beschrijving

De meeste aandacht bij het beveiligen van microdata gaat uit naar de behandeling van de identificerende variabelen. Zij spelen een belangrijke rol bij de beveiliging. De numerieke variabelen zijn vaak de variabelen waar de uiteindelijke interesse van een gebruiker van de data (en ook van een mogelijke onthuller) naar uitgaat, zoals inkomen etc. Het feitelijk inkomen van een gemiddelde Nederlander zal niet erg identificerend zijn, echter wel het inkomen van de extreme veelverdienders. Opeens

is de variabele inkomen identificerend geworden en derhalve moet extra beveiliging worden overwogen.

Top-coding is in die situatie een geschikte methode. Het is een eenvoudige methode, waarbij alle waarden boven een bepaalde drempelwaarde door eenzelfde standaard waarde worden vervangen. Dit kan een indicatie zijn als (‘veel’) of (‘> drempel waarde’). Maar ook het gemiddelde van alle records met een waarde boven die drempelwaarde kan gebruikt worden. Dit laatste heeft als voordeel dat het gemiddelde voor de top-coded variabele over alle records hetzelfde blijft.

Op equivalente wijze kan ook bottom-coding worden toegepast.

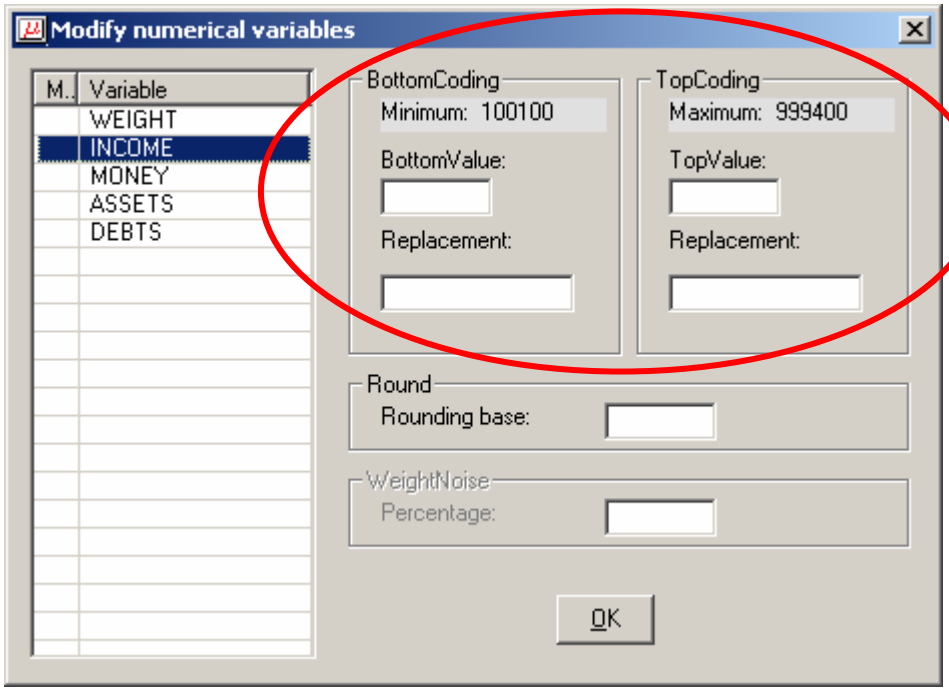
Het is duidelijk dat top-coding alleen zinvol is bij numerieke variabelen. Voor kwalitatieve variabelen kan globaal hercoderen (zie paragraaf 2.3) gebruikt worden om ook een soort top-coding te verkrijgen.

2.5.2 Toepasbaarheid

Deze methode kan toegepast worden als een additionele beveiliging voor die situaties waarin sommige extreme waarden van numerieke variabelen als identificerend moeten worden beschouwd.

2.5.3 Uitgebreide beschrijving

In µ-ARGUS is een implementatie van deze methode aanwezig. Via de menu-optie Modify|ModifyNumericalVariables kan in een dialoogvenster (zie Figuur 5) eenvoudig worden aangegeven op welke variabele top- of bottom-coding moet worden toegepast en welke vervangingswaarde in het bestand moet worden opgenomen.



Figuur 5: Dialoogvenster van µ-ARGUS voor top/bottom-coding

Als de informatie is opgegeven, zal μ-ARGUS in deze fase alleen de specificatie opslaan. Pas als echt een beveiligd bestand wordt weggeschreven, zal de top- of bottom-coding feitelijk worden uitgevoerd.

2.6 Ruis toevoegen aan gewichten

2.6.1 Korte beschrijving

Indien het bestand ophooggewichten bevat (ter correctie voor de steekproef en/of non-respons) moet de beveiliger zich afvragen of met behulp van informatie over het steekproefontwerp, uit die ophooggewichten bepaalde informatie is af te leiden die tot onthulling zou kunnen leiden. Een bekend voorbeeld is dat vaak regio als een stratificatievariabele is gebruikt. Indien bij de beveiliging via globaal hercoderen (zie paragraaf 2.3) de regio-informatie wordt beperkt of eventueel geheel verwijderd, dient men zich af te vragen of uit de waarde van de ophoogvariabele toch informatie is af te leiden over de regio. Als bijvoorbeeld gemeente wordt vervangen door provincie, kan de mogelijkheid ontstaan dat uit het ophooggewicht kan blijken dat het om een grote gemeente gaat. En dus is duidelijk om welke (onderdrukte) gemeente-informatie het gaat.

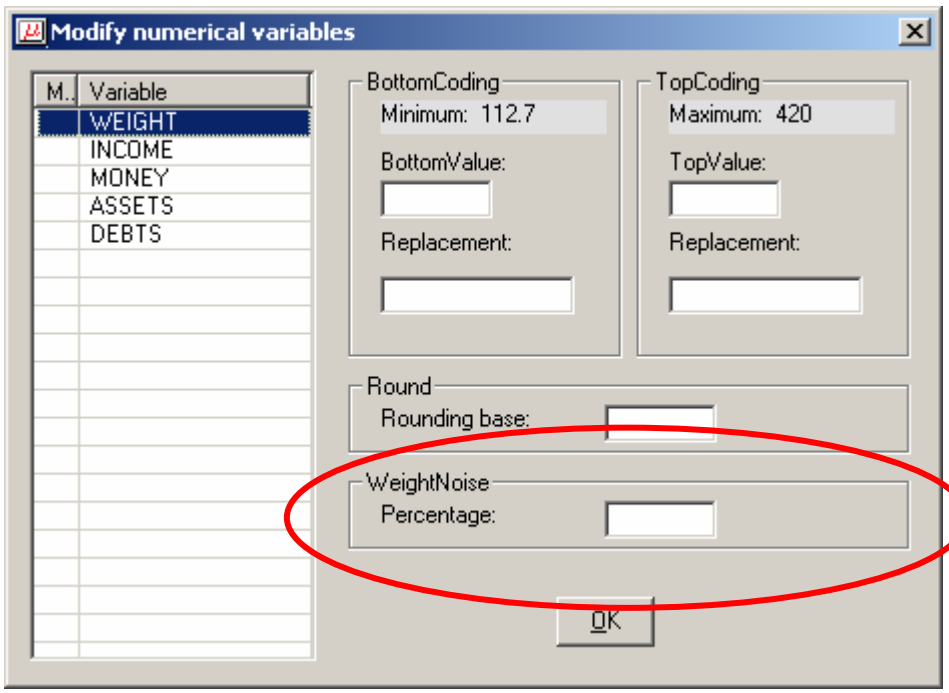
Dit soort onthulling kan worden vermeden door voldoende ruis toe te voegen aan het ophooggewicht. Door het toevoegen van aselecte ruis zal in het algemeen het ophooggewicht nog steeds goed bruikbaar zijn bij analyses.

2.6.2 Toepasbaarheid

In die gevallen zoals hierboven aangegeven, zal kennis over het steekproefdesign informatie kunnen vrijgeven, die kan helpen bij het onthullen van gegevens. Deze methode kan helpen te voorkomen dat uit ophooggewichten informatie onthuld kan worden over individuele respondenten.

2.6.3 Uitgebreide beschrijving

In μ-ARGUS is een implementatie van deze methode aanwezig. Via de menu-optie Modify|ModifyNumericalVariables kan in een dialoogvenster (zie Figuur 6) eenvoudig worden aangegeven hoeveel ruis aan het ophooggewicht moet worden toegevoegd.



Figuur 6: Dialoogvenster van μ-ARGUS voor het toevoegen van ruis aan gewichten

Er kan een percentage p opgegeven worden, zodat μ-ARGUS het gewicht w_i zal vervangen door een willekeurige waarde uit het interval

$$\left[\frac{(100-p)}{100}w_i, \frac{(100+p)}{100}w_i\right]. \qquad (2.6.1)$$

Als de informatie is opgegeven, zal μ-ARGUS in deze fase alleen de specificatie opslaan. Pas als echt een beveiligd bestand wordt weggeschreven, zal de ruis aan de ophoogvariabele worden toegevoegd.

2.7 PRAM

2.7.1 Korte beschrijving

De Post Randomisatie Methode (PRAM) is een methode voor de statistische beveiliging van categoriale variabelen. PRAM is te beschouwen als een bewuste misclassificatie, waarbij de misclassificatiekansen door de beveiliger zijn vastgelegd. Daarnaast is PRAM ook gerelateerd aan de zogenaamde Randomised Response (RR) techniek. RR wordt echter bij de vraagstelling uitgevoerd, terwijl PRAM pas wordt toegepast nadat het antwoord al is gegeven.

Bij het toepassen van PRAM wordt bij ieder record in een microdatabestand de score op één of meer categoriale variabelen met een bepaalde kans wel of niet veranderd. Dit gebeurt onafhankelijk over alle records. Het kansmechanisme dat de overgang van scores bepaald, is vooraf vastgelegd in een zogenaamde Markov-matrix.

Aangezien PRAM een stochastische methode is, wordt het onthullingrisico direct beïnvloed: wanneer een onthuller denkt een record te herkennen, is er een bepaalde

kans dat dit record *niet* overeenkomt met de persoon die de onthuller in gedachte heeft. Immers, met een bepaalde kans zijn enkele scores op identificerende variabelen veranderd.

Het feit dat het gebruikte kansmechanisme bekend is wanneer PRAM wordt toegepast, heeft tot gevolg dat het mogelijk is om met behulp van de beveiligde microdata en de Markov-matrix zuivere schatters te construeren voor bepaalde statistische eigenschappen van de originele data. Daarbij kan ook gebruik gemaakt worden van technieken uit de misclassificatie en de Randomised Response.

Voor een gedetailleerde beschrijving van PRAM, verwijzen we naar Gouweleeuw et al. (1998a en 1998b).

2.7.2 Toepasbaarheid

Volgens het CBS-beleid voor de statistische beveiliging van microdata onder contract, moet voorkomen worden (of in ieder geval bemoeilijkt) dat individuele personen herkend kunnen worden. Om een individuele persoon te kunnen herkennen, zal een onthuller gebruik maken van identificerende variabelen, zoals geslacht, burgerlijke staat, leeftijd en opleidingsniveau. Dit werkt uiteraard alleen in het geval de onthuller er zeker van kan zijn dat de scores op die variabelen in het geleverde bestand ook daadwerkelijk de echte scores zijn. Door PRAM op identificerende variabelen toe te passen, wordt die zekerheid weggehaald: er is immers een positieve kans dat de score niet de originele score meer is.

Bij de statistische beveiliging van een microdatabestand onder contract, is het over het algemeen niet mogelijk om zeer gedetailleerde regionale variabelen op te nemen. Met name in het geval ook andere (gedetailleerde) identificerende variabelen in het bestand aanwezig zijn. De traditionele statistische beveiligingsmethoden als hercoderen, top-coding en locaal onderdrukken zouden dan een vrijwel onbruikbaar bestand opleveren voor analyses waarbij het regionale detail van belang is. PRAM zou dan een mogelijk alternatief zijn: het detailniveau blijft gehandhaafd, alleen is niet meer met zekerheid de juiste score op een identificerende variabele te zien.

Een gebruiker van een bestand dat met PRAM is beveiligd, moet echter over voldoende statistische kennis beschikken, om de analysemethode die hij wil toe passen te kunnen corrigeren voor de aangebrachte veranderingen in de records. Van een aantal analysemethoden is bekend op welke manier de methoden aangepast moeten worden. Zie bijvoorbeeld Gouweleeuw et al. (1998a en 1998b), Van den Hout (1999), Van den Hout en van der Heijden (2002) en Ronning et al. (2004).

Bestanden die met PRAM zijn beveiligd, zijn dan ook vooral bedoeld voor (theoretisch) ervaren statistici. Daarnaast kunnen microdatabestanden waarop PRAM is toegepast ook gebruikt worden als “testbestand”. Bijvoorbeeld om scripts uit te testen of om onderzoeksrichtingen te bepalen. De uiteindelijke definitieve analyse zou dan vervolgens uitgevoerd kunnen worden op het originele (onbeveiligde) bestand via een remote execution of een on-site sessie.

2.7.3 Uitgebreide beschrijving

Voor een uitgebreide theoretische beschrijving van de methode verwijzen we naar Gouweleeuw et al. (1998a en 1998b).

Bij het toepassen van PRAM speelt de Markov-matrix met overgangskansen een belangrijke rol. De overgangskansen bepalen het beveiligingsniveau en hebben invloed op het informatieverlies. Het is dus van belang om die kansen op een juiste manier te kiezen. Iedere gebruiker zal informatieverlies op een verschillende manier ervaren. Het verdient dan ook de voorkeur om, bij het bepalen van de overgangskansen, de wensen van de gebruiker in het achterhoofd te houden. In De Wolf (2006) zijn verschillende maten voor informatieverlies gegeven.

Aangezien PRAM een stochastische beveiligingsmethode is, zijn de standaard regels zoals beschreven in het Handboek Statistische Beveiliging (Hundepool et al., 2006) niet direct toepasbaar. Wel zijn er alternatieve regels gegeven in bijvoorbeeld De Wolf (2006), die gerelateerd zijn aan de standaardregels uit het Handboek Statistische Beveiliging.

Het mag duidelijk zijn dat de keuze van de overgangskansen geen eenvoudige zaak is. Er is geen universele manier om in iedere situatie de juiste beslissing te nemen. De volgende vragen spelen een rol bij het bepalen van de overgangskansen:

- Op welke variabelen wordt PRAM toegepast?
- Wordt PRAM op (een deel van) de variabelen onafhankelijk toegepast?
- Zijn er onmogelijke combinaties die voorkomen moeten worden door de bijbehorende overgangskansen op nul te zetten?
- Wat is het effect op het informatieverlies?
- Wat is het effect op het onthullingrisico?

Per geval zullen antwoorden op deze vragen de keuze van de specifieke overgangskansen bepalen. Er is dus geen universele manier voorhanden om de ideale overgangskansen te bepalen.

Voor een empirische studie naar de gevolgen van verschillende mogelijkheden voor de overgangskansen op zowel het onthullingrisico als het informatieverlies, verwijzen we naar De Wolf (2006).

Bij de keuze van een matrix van overgangskansen, zijn een aantal typische structuren mogelijk. Zo kan een bandmatrix met bandbreedte *b* goed bruikbaar zijn bij ordinale variabelen als Leeftijd. In dat geval wordt een leeftijd met een bepaalde kans vervangen door een leeftijd binnen plus of min *b* jaar. Volledig gevulde matrices zijn vooral goed bruikbaar bij nominale variabelen met een beperkt aantal categorieen, zoals de variabele Burgerlijke Staat. Zie Figuur 7 voor een paar voorbeelden.

$$\begin{pmatrix} 0.70 & 0.10 & 0.10 & 0.10 \\ 0.02 & 0.94 & 0.02 & 0.02 \\ 0.09 & 0.09 & 0.73 & 0.09 \\ 0.11 & 0.11 & 0.11 & 0.67 \end{pmatrix} \qquad \begin{pmatrix} 0.80 & 0.20 & 0 & 0 \\ 0.10 & 0.80 & 0.10 & 0 \\ 0 & 0.20 & 0.60 & 0.20 \\ 0 & 0 & 0.10 & 0.90 \end{pmatrix}$$

(a) (b)

Figuur 7: Voorbeelden van matrices met overgangskansen: (a) Volledig gevulde matrix, (b) Bandmatrix met bandbreedte 1

Voor andere variabelen ligt een blokmatrix meer voor de hand. Bijvoorbeeld voor een variabele als Regio (op gemeenteniveau) kan gedacht worden aan een blokmatrix waarbij de blokken corresponderen met de Provincies. Op die manier kunnen gemeentes alleen vervangen worden door andere gemeentes uit dezelfde provincie. Zie Figuur 8 voor een voorbeeld van een blokmatrix met overgangskansen.

$$\begin{pmatrix} 0.90 & 0.10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.20 & 0.80 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.70 & 0.20 & 0.10 & 0 & 0 & 0 \\ 0 & 0 & 0.10 & 0.80 & 0.10 & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0.15 & 0.70 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.75 & 0.15 & 0.10 \\ 0 & 0 & 0 & 0 & 0 & 0.09 & 0.82 & 0.09 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0.05 & 0.90 \end{pmatrix}$$

Figuur 8: Voorbeeld blokmatrix met drie blokken met overgangskansen.

2.7.4 Voorbeeld

Vooralsnog wordt in mARGUS slechts een beperkte faciliteit geboden om bestanden statistisch te beveiligen met behulp van PRAM. Binnen dat pakket is het mogelijk om PRAM per variabele toe te passen, waarbij de Markov-matrix ofwel een bandmatrix, ofwel een volledig gevulde matrix kan zijn. De bandbreedte van een bandmatrix is instelbaar, evenals de diagonaalkansen (de kansen dat bepaalde categorieën *niet* veranderen).

Aangezien PRAM een stochastische beveiligingsmethode is (alleen de overgangs*kansen* zijn vastgelegd), zal een beveiligd bestand na iedere toepassing van PRAM er anders uit kunnen zien: een dergelijk beveiligd bestand is immers de uitkomst van een kansexperiment. Analyses kunnen dan ook alleen *in verwachting* gecorrigeerd worden voor het feit dat ze zijn toegepast op een bestand dat met PRAM is beveiligd. Dat wil zeggen dat bijvoorbeeld de verwachting van parameters die gecorrigeerd geschat zijn, gelijk zal zijn aan de parameterschattingen op basis van het originele bestand.

Beschouw, om een indruk te geven van mogelijke aanpassingen van analyses, het eenvoudige geval van PRAM toegepast op de variabele Geslacht (twee categorieën) waarbij we de frequentietabel van het aantal mannen en het aantal vrouwen willen schatten. De variabele Geslacht noteren we met ξ waarbij ξ = 1 = Man en ξ = 2 = Vrouw. De bijbehorende frequentietabel noteren we met $\mathbf{T}_\xi$. Stel dat het oorspronkelijke bestand 100 mannen en 100 vrouwen bevat, dus $\mathbf{T}_\xi = (100, 100)^t$. PRAM wordt met de volgende matrix met overgangskansen toegepast op de variabele Geslacht:

$$\mathbf{P} = \begin{pmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{pmatrix}. \tag{2.7.1}$$

In woorden: de kans dat het geslacht Man in Vrouw wordt veranderd is 10%, de kans dat het geslacht Vrouw in Man wordt veranderd is 20%. De variabele ξ wordt met X genoteerd na toepassing van PRAM. De frequentietabel van Geslacht op basis van het beveiligde bestand wordt dan genoteerd met $\mathbf{T}_X$. Eenvoudig valt dan af te leiden dat

$$E\left(\mathbf{T}_X \middle| \xi\right) = \mathbf{P}^t \mathbf{T}_\xi, \tag{2.7.2}$$

waarbij de verwachting voorwaardelijk op het originele bestand is. In het voorbeeld betekent dit dat, in verwachting, 110 mannen en 90 vrouwen in het beveiligde bestand zullen voorkomen. Uit vergelijking (2.7.2) volgt direct een zuivere schatter voor de oorspronkelijke frequentietabel, namelijk

$$\hat{\mathbf{T}}_\xi = \left(\mathbf{P}^t\right)^{-1} \mathbf{T}_X . \tag{2.7.3}$$

De originele frequentietabel zal door deze gecorrigeerde schatter slechts *in verwachting* gereproduceerd worden. Dat wil zeggen,

$$E\left(\hat{\mathbf{T}}_\xi \middle| \xi\right) = \mathbf{T}_\xi . \tag{2.7.4}$$

Stel dat het beveiligde bestand 112 mannen en 88 vrouwen bevat (NB: dit is een voorbeeld, want dit kan per realisatie van het kansexperiment verschillen), dan zou (afgerond op gehele getallen) de zuivere schatting dus gegeven zijn door

$$\hat{\mathbf{T}}_\xi = \left(\mathbf{P}^t\right)^{-1} \mathbf{T}_X = \begin{pmatrix} 0.90 & 0.20 \\ 0.10 & 0.80 \end{pmatrix}^{-1} \begin{pmatrix} 112 \\ 88 \end{pmatrix} = \begin{pmatrix} 103 \\ 97 \end{pmatrix}. \tag{2.7.5}$$

Merk op dat deze gecorrigeerde schatting van de frequentietabel veel dichter bij de originele frequentietabel ligt dan de niet gecorrigeerde schatting (de directe telling uit het beveiligde bestand), maar dat de exacte originele waarden niet zijn verkregen.

2.8 Afsluiting

Voor de beveiliging van microdatabestanden is op het CBS het pakket μ-ARGUS beschikbaar. Voor een uitgebreide beschrijving van dat pakket verwijzen we naar de handleiding ervan (Hundepool et al., 2007).

Wanneer μ-ARGUS wordt gebruikt, wordt na iedere sessie waarin één of meerdere bestanden zijn beveiligd een rapport opgemaakt. In dat rapport is opgenomen welke methoden en parameters gebruikt zijn.

Met μ-ARGUS is het eenvoudig om de effecten van verschillende statistische beveiligingsmethoden te zien. Methoden kunnen worden toegepast en ook weer ongedaan gemaakt worden (binnen één sessie). Met behulp van het automatisch gegenereerde rapport, kan voor een volgende versie van hetzelfde bestand, eenvoudig worden afgeleid welke (combinatie van) methode(n) uiteindelijk gebruikt is.

Aangezien μ-ARGUS in Europees verband is (wordt) ontwikkeld, zijn daar ook enkele methoden opgenomen, die niet in deze Methodenreeks zijn beschreven. Dat zijn dan methoden die wel door enkele andere EU-landen worden gebruikt, maar niet door het CBS.

2.9 Literatuur

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. en De Wolf, P.P. (1998a), *Post Randomisation for Statistical Disclosure Control: Theory and Implementation*, Journal of Official Statistics, vol. 14, 4, pp. 463 – 478.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. en De Wolf, P.P. (1998b), *The post randomisation method for protecting microdata*, Qüestiió, Quaderns d'Estadística i Investigació Operativa, vol. 22, 1, pp. 145 – 156.

Van den Hout, A. (1999), *The analysis of data perturbed by pram*, Delft University Press, Delft.

Van den Hout, A. en Van der Heijden, P.G.M. (2002), *Randomized response, statistical disclosure control and misclassification: a review*, International Statistical Review 70(2), pp. 269 – 288.

Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt E. en De Wolf, P.P. (2006), *Handboek Statistische Beveiliging*, BPA 21-06-TMO.

Hundepool, A., v. d. Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra,V., Brand, R. en Giessing, S. (2007), *μ-ARGUS user manual 4.1*, Voorburg.

Ronning, G., Rosemann, M. en Strotmann, H. (2004), *Estimation of the probit model using anonymized micro data*, Paper prepared for the 'European Conference on Quality and Methodology in Official Statistics (Q2004)', Mainz, 24–26 May 2004.

De Wolf, P.P. (2006), *Risk, Utility and PRAM*, in 'Privacy in Statistical Databases 2006', Domingo-Ferrer, J. en Franconi, L. (Eds.), LNCS 4302, Springer-Verlag, Berlin Heidelberg, pp. 189 – 204.

3. Statistische Beveiliging van Kwantitatieve Tabellen

3.1 Algemene beschrijving en leeswijzer

3.1.1 Algemene beschrijving

Onder statistische beveiliging van kwantitatieve tabellen verstaan we het produceren van kwantitatieve tabellen die voldoen aan het CBS-beleid aangaande statistische beveiliging en die als zodanig gepubliceerd kunnen worden. Het beveiligingsbeleid voor kwantitatieve tabellen is vastgelegd in hoofdstuk 4 van het Handboek Statistische Beveiliging (Hundepool et al., 2006). Kwantitatieve tabellen zijn tabellen waarbij de celwaardes zijn ontstaan door sommering van een continue variabele over alle bijdragers tot een cel. Dit in tegenstelling tot frequentietabellen waarbij alleen het *aantal* bijdragers per cel wordt gegeven. Voor frequentietabellen gelden andere regels en kunnen andere beveiligingsmethoden beter geschikt zijn dan die voor kwantitatieve tabellen. Beveiligingsmethoden voor frequentietabellen worden in het deelthema “Statistische Beveiliging van Frequentietabellen” behandeld.

Wanneer precies één of twee bijdragers een celtotaal leveren, is het duidelijk dat die cel niet te publiceren is. In het geval van één bijdrager wordt direct individuele informatie vrijgegeven en in het geval van twee bijdragers kan de ene bijdrager de andere bijdrage exact uitrekenen door zijn eigen bijdrage van het celtotaal af te trekken.

Maar ook in het geval dat er meer dan twee bijdragers in een cel zitten, kunnen ongewenste situaties ontstaan. In principe moeten we bij de statistische beveiliging van kwantitatieve tabellen voorkomen (of op zijn minst bemoeilijken) dat een willekeurige bijdrage te nauwkeurig kan worden geschat. Dit kan bijvoorbeeld ook in het geval dat een zeer grote bijdrager in één cel zit, samen met een aantal relatief kleine bijdragers. De op één na grootste bijdrager kan dan uitrekenen dat de grootste bijdrage niet méér bijdraagt dan het celtotaal minus zijn eigen bijdrage. Daarmee is, in strijd met de beveiligingsregels van het CBS, een vrij goede schatting van de bijdrage van de grootste bijdrager verkregen.

Ook het vóórkomen van lege cellen vraagt extra aandacht. In een aantal gevallen zal een lege cel een zogenaamde *structurele nulcel* zijn. Dat wil zeggen dat het algemeen bekend is dat die cel (op logische gronden) geen bijdragen *kan* hebben. Dergelijke cellen kunnen dan ook niet gebruikt worden bij de beveiliging: wat je ook doet, iedereen weet dat het een lege cel moet zijn.

Tegelijkertijd kan met *niet-structurele nulcellen* soms wel degelijk informatie onthuld worden. In het geval dat er wel bijdragers in zo’n cel zitten, is er immers in feite een soort groepsonthulling: het is direct duidelijk dat alle bijdragers een bijdrage nul hebben geleverd (onder de aanname dat bijdragen niet-negatief zijn). In het

geval er geen bijdragers in de cel zitten, maar het niet op voorhand onmogelijk is om een bijdrager in die cel te hebben, geeft dat op zich ook direct informatie.

De in dit deelthema beschreven methoden kunnen door middel van het software pakket τ-ARGUS eenvoudig worden toegepast. Dit pakket is door DMV in Europees verband ontwikkeld.

3.1.2 Leeswijzer

Als eerste stap bij het bepalen van de juiste statistische beveiliging voor een kwantitatieve tabel, zal eerst bepaald moeten worden of er onthulling mogelijk is. Als basis geldt daarbij in eerste instantie het “gezonde verstand”: is er informatie in de tabel aanwezig die niet over individuele respondenten onthuld mag worden? Bij kwantitatieve tabellen is dergelijke informatie over het algemeen de individuele bijdrage van een respondent aan het totaal van een specifieke cel in de tabel.

Daarnaast is een objectieve methode nodig om te bepalen welke cellen van de tabel respondenten bevatten die mogelijk gevaar lopen op onthulling van hun individuele bijdrage. De p%-regel (zie paragraaf 3.3) is bedoeld om dergelijke primair onveilige cellen aan te kunnen wijzen. Deze methode is alleen toepasbaar in het geval van kwantitatieve tabellen en niet in het geval van frequentietabellen.

Nadat de onveilige cellen zijn aangewezen, zal over het algemeen de tabel nog verder beveiligd moeten worden. Daarvoor zijn grofweg drie methoden beschikbaar: herstructureren van de tabel (zie paragraaf 3.4), onderdrukken van cellen (zie paragraaf 3.5) en afronden (zie paragraaf 3.6).

Welke (combinatie van) methode(n) uiteindelijk in een specifieke situatie gebruikt wordt, is niet op voorhand vast te leggen. Dit hangt sterk af van de beoogde gebruikers. Zo zal het bij Eurostat-verordeningen niet altijd mogelijk zijn om de tabel te herstructureren en zal vaak voor celonderdrukking gekozen (moeten) worden. De afdeling die verantwoordelijk is voor de betreffende kwantitatieve tabel, is ook verantwoordelijk voor een adequate (statistische) beveiliging daarvan. Bij de keuze voor de te gebruiken methode(n) moeten twee (concurrerende) aspecten worden meegenomen:

- onthullingrisico;
- informatieverlies.

Algemeen kan gezegd worden, dat verkleining van het onthullingrisico tot meer informatieverlies zal leiden. Ook het omgekeerde is waar: hoe kleiner het informatieverlies, hoe groter het onthullingrisico. In voorkomende gevallen zal een afweging gemaakt moeten worden, waarbij uiteraard wel altijd minimaal aan de regels in het Handboek Statistische Beveiliging (Hundepool et al., 2006) voldaan moet worden.

3.2 Afbakening en relatie met andere (deel)thema's

In dit deelthema worden methoden behandeld die te gebruiken zijn voor de statistische beveiliging van kwantitatieve tabellen. In dit hoofdstuk worden *geen* methoden beschreven die alleen voor frequentietabellen zijn te gebruiken. Voor dergelijke methoden, zie het deelthema "Statistische Beveiliging van Frequentietabellen".

Enkele in dit hoofdstuk beschreven methoden kunnen in principe zowel bij kwantitatieve tabellen als bij frequentietabellen gebruikt worden. Dergelijke methoden zullen in het deelthema "Statistische Beveiliging van Frequentietabellen" herhaald worden.

De methoden in dit hoofdstuk zijn in te delen in twee varianten: methoden voor het bepalen van de (primair) onveilige cellen van een kwantitatieve tabel en methoden voor het alsnog publiceerbaar maken van tabellen met onveilige cellen.

3.3 *P* % regel

3.3.1 Korte beschrijving

Statistische beveiliging heeft tot doel de onthulling van informatie over individuele bijdragers aan een tabel te voorkomen of op z'n minst te bemoeilijken. Om dit te kunnen bereiken, zullen eerst de cellen moeten worden aangewezen waarin risico is op een mogelijk onthulling. Daarvoor is een objectieve maat nodig, die aangeeft hoe goed een individuele bijdrage aan een cel te schatten is op basis van de gepubliceerde tabel. De *p* % regel voorziet hierin. Tevens is de methode te gebruiken om aan te geven in welke mate de beveiligingsregel wordt geschonden en hoe groot de te treffen maatregelen zouden moeten zijn.

3.3.2 Toepasbaarheid

Voordat een kwantitatieve tabel statistisch beveiligd kan worden, zal eerst aangegeven moeten worden waar in die tabel zich mogelijk problemen voordoen. De *p* % regel geeft aan hoe goed een bijdrager in een cel een andere bijdrager in diezelfde cel zou kunnen schatten. Daarmee zijn de primair onveilige cellen bepaald en is tevens aangegeven hoeveel bescherming geleverd moet worden om aan het CBS-beleid te kunnen voldoen bij het publiceren van kwantitatieve tabellen.

Met deze methode kan ook rekening gehouden worden met eventuele machtigingen: bijdragers die hebben aangegeven dat ze geen bezwaar hebben tegen publicaties waaruit hun bijdrage te halen is. Dergelijke bijdragers worden dan eenvoudigweg uitgesloten bij de toepassing van de *p* % regel.

De *p* % regel is alleen toe te passen

- in geval van kwantitatieve tabellen;
- met niet-negatieve bijdragers;
- waarvan de grootste bijdragers identificeerbaar zijn voor de onthuller;
- op niet-lege cellen met een positief celtotaal.

3.3.3 Uitgebreide beschrijving

Zij T_A de celwaarde van cel A in de betreffende tabel. Schrijven we de n_A individuele bijdragers (zonder machtigingen) aan die cel van groot naar klein als $X_1 \geq X_2 \geq \cdots \geq X_{n_A} \geq 0$, dan is cel T_A onveilig indien

$$\frac{(T_A - X_2) - X_1}{X_1} < \frac{p}{100}. \tag{3.3.1}$$

Ofwel, een cel is onveilig als de één na grootste bijdrager de grootste bijdrager met een grotere nauwkeurigheid dan p % kan schatten.

Eenvoudig is in te zien dat dit het slechtste scenario is: indien de één na grootste bijdrager de grootste bijdrager *niet* nauwkeuriger dan p % kan schatten, dan kan geen enkele andere bijdrager een willekeurig andere bijdrager nauwkeuriger dan die p % schatten en dus is de cel dan veilig. Met andere woorden: de meest nauwkeurige schatting kan worden gemaakt door de één na grootste bijdrager, wanneer die de grootste bijdrage schat.

De waarde van het verschil tussen de linkerkant en de rechterkant van de ongelijkheid in formule (3.3.1) geeft ook aan hoeveel bescherming een onveilige cel nodig heeft. Voor meer detail verwijzen we naar Loeve (2001).

Met de standaardsoftware op het CBS voor de beveiliging van tabellen, τ-ARGUS, is het eenvoudig om de p % regel toe te passen. Deze methode is één van de standaard ingebouwde regels die gebruikt kunnen worden om de primair onveilige cellen aan te wijzen. Bovendien berekent τ-ARGUS automatisch hoeveel bescherming een onveilige cel nodig heeft en gebruikt dat bij de verdere beveiliging van de betreffende tabel. Om het τ-ARGUS mogelijk te maken de primair onveilige cellen aan te wijzen met behulp van de p % regel, is het wel noodzakelijk dat de invoer voor τ-ARGUS bestaat uit de microdata waaruit de betreffende tabel wordt opgebouwd. Om de p % regel toe te kunnen passen is immers informatie nodig over de individuele bijdragers. Voor meer informatie over het gebruik van τ-ARGUS verwijzen we naar de handleiding daarvan (Hundepool et al., 2003).

De waarde die voor p wordt gekozen, is bepaald door het CBS-beleid. In het Handboek Statistische Beveiliging van het CBS (Hundepool et al., 2006) wordt een interval gegeven waarbinnen p gekozen dient te worden ($5 \leq p \leq 15$). De exacte waarde voor p wordt door de statistische divisie bepaald en mag nooit aan externen worden verteld, omdat dit zou kunnen helpen bij het terugrekenen van onderdrukte cellen.

Een grote waarde voor p resulteert in een strenge beveiliging, aangezien in dat geval bij het schatten van een willekeurige bijdrage zelfs geen relatief “grote” fout gemaakt mag worden. Een kleine waarde voor p resulteert in een minder strenge beveiliging, aangezien een cel dan pas onveilig is als een bijdrage zeer nauwkeurig kan worden geschat.

3.3.4 Voorbeeld

In dit fictieve voorbeeld kijken we naar een cel in een tabel met omzet naar SBI en Regio. Stel dat de cel met SBI = 32 en Regio = Noord Brabant bestaat uit 4 bijdragers met de waarden 324, 4, 2 en 10. Stel dat we de p % regel toe willen passen met $p = 5$, dan moeten we de bijdragers eerst sorteren: $X_1 = 324$, $X_2 = 10$, $X_3 = 4$ en $X_4 = 2$. Het celtotaal T_A is dan 340. Berekenen we vervolgens het quotiënt uit formule (3.3.1), dan krijgen we de waarde 0,0185. Dit is duidelijk kleiner dan 5 % en dus is de cel onveilig.

3.4 Tabel herstructureren

3.4.1 Korte beschrijving

In paragraaf 3.3 is beschreven wanneer een cel in een tabel als onveilig moet worden beschouwd. In het algemeen zijn cellen met een beperkt aantal bijdragers danwel een cel met één of twee grote bijdragers de voor de hand liggende kandidaten om als gevoelig te worden aangemerkt. Alle onveilige cellen moeten beveiligd worden. Alvorens op ruime schaal te gaan onderdrukken, kan ook overwogen worden de tabel te herstructureren. Door het samenvoegen van rijen en/of kolommen worden cellen samengevoegd en wordt de vulling per cel vergroot. Dit leidt ertoe dat minder cellen als onveilig worden aangemerkt door de p % regel, zoals beschreven in paragraaf 3.3.

3.4.2 Toepasbaarheid

Deze methode zal er in het algemeen toe leiden dat er minder onveilige cellen in de tabel zullen voorkomen. Door cellen samen te voegen ontstaan cellen die veiliger zijn dan de cellen die waren samengevoegd.

Er zijn geen methodologische voorwaarden voor het toepassen van deze methode. Echter, door extern opgelegde leveringsverplichtingen wordt soms voorgeschreven in welke mate van detail een tabel dient te worden gepubliceerd. Dit kan een Eurostat-verplichting zijn, maar ook het CBS-beleid kan ertoe leiden dat een bepaald detailniveau van een tabel moet worden gepubliceerd. In die gevallen kan de methode technisch gezien dus wel toegepast worden, maar wordt dit door (externe) beleidsbeslissingen verhinderd.

Verder moet een afweging gemaakt worden tussen het informatieverlies ten gevolge van het grotere aantal kruisjes (onderdrukte cellen) dat nodig is om de tabel te beveiligen en het informatieverlies ten gevolge van het samenvoegen van kolommen/rijen, waarbij dan (veel) minder kruisjes nodig zijn.

3.4.3 Uitgebreide beschrijving

Binnen het sofware pakket τ-ARGUS zijn voorzieningen aanwezig om rijen en/of kolommen in tabellen te hercoderen. Daarbij wordt onderscheid gemaakt tussen twee situaties:

- In het geval van een hiërarchische opspanvariabele bestaat de hercodering eruit dat bepaalde uitsplitsingen op het laagste niveau worden weggelaten.
- In het geval van een ongestructureerde opspanvariabele is de gebruiker vrij om naar eigen inzicht kolommen of rijen van een tabel samen te voegen.

3.4.4 Voorbeeld

In Figuur 9 is een (fictieve) tabel gegeven van omzet naar Regio (hiërarchisch) en GrootteKlasse. In Figuur 10 zijn twee mogelijke herstructureringen van deze tabel gegeven. De hercodering van de variabele Grootteklasse is dat de categorieën 2 tot en met 6 zijn samengevoegd tot de categorie MiddelKlein en dat de categorieën 7, 8 en 9 zijn samen genomen tot de categorie Groot. Merk op dat op deze manier alle primair onveilige cellen zijn samengevoegd tot veilige cellen. De hercodering van de variabele Regio is zodanig dat het kleinste detailniveau is verwijderd. Deze herstructurering lost niet alle problemen op: de primair onveilige cellen op landsdeel-niveau (bij Noord en Oost) blijven in de tabel aanwezig.

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
Noord	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
1	1,986,129.00	×	×	398,062.00	348,039.00	354,711.00	418,778.00	466,529.00	-
2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
Oost	3,703,896.00	15.00	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
4	124,336.00	×	-	36,311.00	32,132.00	25,770.00	18,150.00	×	-
5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	81,799.00	145,004.00	-
6	2,234,995.00	×	×	345,803.00	251,358.00	251,188.00	303,377.00	1,083,254.00	-
7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
Zuid	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
11	2,752,743.00	-	15.00	488,613.00	392,395.00	363,490.00	402,925.00	1,105,305.00	-
12	1,441,228.00	-	-	212,936.00	209,886.00	254,547.00	244,096.00	519,763.00	-
99	-	-	-	-	-	-	-	-	-

Figuur 9: Kwantitatieve tabel omzet naar regio en grootteklasse.

	tot	Groot	MiddelKlein	99
tot	16,847,646.84	11,814,874.84	5,032,387.00	385.00
Noord	4,373,664.00	2,994,540.00	1,378,739.00	385.00
1	1,986,129.00	1,240,018.00	746,111.00	.
2	1,809,246.00	1,363,539.00	445,322.00	385.00
3	578,289.00	390,983.00	187,306.00	.
Oost	3,703,896.00	2,546,635.00	1,157,261.00	.
4	124,336.00	55,888.00	68,448.00	.
5	526,279.00	337,733.00	188,546.00	.
6	2,234,995.00	1,637,819.00	597,176.00	.
7	818,286.00	515,195.00	303,091.00	.
West	4,576,115.84	3,383,573.84	1,192,542.00	.
8	485,326.00	346,117.00	139,209.00	.
9	3,664,559.84	2,695,797.84	968,762.00	.
10	426,230.00	341,659.00	84,571.00	.
Zuid	4,193,971.00	2,890,126.00	1,303,845.00	.
11	2,752,743.00	1,871,720.00	881,023.00	.
12	1,441,228.00	1,018,406.00	422,822.00	.
99	.	.	.	.

(a) Hercodering Grootteklasse (alle primair onveilige cellen zijn beveiligd)

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
Noord	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
Oost	3,703,896.00	15.00	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	.
West	4,576,115.84	.	.	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	.
Zuid	4,193,971.00	.	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	.
99	.	.	.	.	.	.	.	.	.

(b) Hercodering Regio (niet alle primair onveilige cellen zijn beveiligd)

Figuur 10: Twee mogelijke herstructureringen toegepast op de tabel uit Figuur 9

3.5 Cellen onderdrukken

3.5.1 Korte beschrijving

Een veel gebruikte methode om primair onveilige cellen te beveiligen, is het onderdrukken (niet publiceren) van bepaalde cellen. De celwaarde wordt dan eenvoudig vervangen door een kruisje (×).

In een kwantitatieve tabel waarbij de marginalen ook gegeven zijn, is het echter vaak niet voldoende om alleen de primair onveilige cellen te onderdrukken. Wanneer een onderdrukte cel de enige onderdrukte cel in een rij is, is de onderdrukte waarde immers eenvoudig uit te rekenen door de overige celwaarden in die rij van de bijbehorende marginaal af te trekken.

Om primair onveilige cellen toch voldoende te kunnen beschermen is het dan ook noodzakelijk om ook andere, van zichzelf veilige, cellen te onderdrukken. Dit heet *secundair onderdrukken*. Het is niet eenvoudig om dit op een zodanige manier te doen dat de primair onveilige cellen voldoende worden beveiligd, terwijl er ook niet te veel informatie uit de tabel wordt weggehaald. Bovendien moet ook rekening worden gehouden met het feit dat structurele nulcellen niet gebruikt kunnen worden als secundaire onderdrukkingen: iedereen weet immers dat die cellen per definitie leeg zijn.

Om te voorkomen dat onderdrukte, primair onveilige cellen exact kunnen worden teruggerekend, zijn dus secundaire onderdrukkingen nodig. Echter, ook nu speelt weer een rol dat een “te nauwkeurige” schatting voor een onderdrukte cel niet gewenst is. Wat is immers het verschil tussen de uitspraak “Deze onderdrukte cel heeft

eigenlijk een waarde van 10000" en "Deze onderdrukte cel heeft eigenlijk een waarde tussen 9998 en 10002". Gegeven een onderdrukkingspatroon is het altijd[1] mogelijk om een interval te berekenen waarbinnen een onderdrukte cel moet liggen. De methode "Cellen Onderdrukken" moet dan ook een onderdrukkingspatroon opleveren, waarbij de te berekenen intervallen groot genoeg zijn. De grootte van die intervallen wordt bepaald door de regel die is gebruikt om de primair onveilige cellen te bepalen.

Fischetti en Salazar (2000) hebben een methode bedacht om op een optimale manier bovenstaand probleem op te lossen. Hun methode is in theorie toepasbaar op willekeurige, additieve tabellen met niet-negatieve bijdragers. In de praktijk blijkt hun oplossing echter te veel rekentijd te kosten wanneer de tabellen te groot worden in omvang dan wel complexiteit. Vandaar dat er een aantal sub-optimale methoden zijn ontwikkeld voor het vinden van geschikte onderdrukkingspatronen voor grotere en/of complexere tabellen.

Zo splitst de "modulaire aanpak" (HiTaS) een hiërarchische tabel op in een groot aantal niet-hiërarchische deeltabellen en past de optimale methode toe op iedere afzonderlijke deeltabel. Door de resultaten op een juiste manier te combineren is een suboptimale oplossing voor de gehele tabel te krijgen, in een aanzienlijk kortere rekentijd.

De "hypercube aanpak" kan ook grote tabellen beveiligen door op een bepaalde iteratieve manier deeltabellen te beveiligen. De beveiliging van iedere deeltabel vindt ook op een suboptimale manier plaats. Daardoor is de aanpak relatief snel, maar worden over het algemeen meer cellen onderdrukt dan strikt noodzakelijk is om een beveiligde tabel te krijgen.

3.5.2 Toepasbaarheid

Deze methode kan gebruikt worden om kwantitatieve tabellen met cellen die niet aan de eisen van het statistische beveiligingsbeleid van het CBS voldoen, op een adequate manier te beveiligen. Met name in het geval dat een herstructurering van de tabel niet (verder) mogelijk is, is de methode van celonderdrukking goed bruikbaar.

De bijdragen tot de te beveiligen tabel moeten niet negatief zijn en de tabel moet additief zijn, waarbij de marginalen ook zijn gegeven.

Bij de modulaire aanpak, mag de tabel maximaal driedimensionaal zijn. Iedere dimensie mag hiërarchisch zijn. Gekoppelde tabellen zijn eventueel te beveiligen door de onderdrukkingen van één tabel over te nemen in de andere en die vervolgens te beveiligen. Dit zou dan mogelijk iteratief uitgevoerd moeten worden.

---

[1] In het geval dat de tabel is opgebouwd uit niet-negatieve bijdragers en de marginalen gegeven zijn.

Bij de hypercube aanpak zoals geïmplementeerd in τ-ARGUS, mag de tabel maximaal zevendimensionaal zijn. De tabel mag in iedere dimensie hiërarchisch zijn. Gekoppelde tabellen zijn eventueel mogelijk, maar dit is (nog) niet volledig geïmplementeerd in τ-ARGUS.

Voor beide aanpakken geldt overigens dat het uit oogpunt van performance wordt aangeraden om geen lange, ongestructureerde (niet-hiërarchische) codelijsten te gebruiken.

3.5.3 Uitgebreide beschrijving

In het softwarepakket τ-ARGUS zit een voorziening waarmee celonderdrukking op kwantitatieve tabellen kan worden toegepast. τ-ARGUS zal, wanneer de originele microdata als invoer gebruikt zijn, zelf de primair onveilige cellen bepalen met de bijbehorende beveiligingsintervallen (zie ook paragraaf 3.3).

Vervolgens zal τ-ARGUS een onderdrukkingspatroon moeten bepalen dat de benodigde beveiligingsintervallen garandeert. Daarvoor zijn verschillende opties, waarvan we de twee voor het CBS meest interessante aanpakken zullen bespreken.

3.5.3.1 Modulaire aanpak

Voor een gedetailleerdere beschrijving en een uitgewerkt voorbeeld van de modulaire aanpak, zie De Wolf (2002).

Globaal is de modulaire aanpak als volgt te beschrijven:

1. Splits de hiërarchische tabel op in alle logische niet-hiërarchische deeltabellen.
2. Groepeer de deeltabellen op een zodanige manier in klassen, dat alle tabellen binnen één klasse onafhankelijk van elkaar beveiligd kunnen worden. Voor een geschikte indeling, zie De Wolf (2002).
3. Beveilig alle tabellen binnen klasse *K*.
4. Wanneer er geen secundaire onderdrukkingen in de marginalen van de deeltabellen van klasse *K* worden gezet, ga dan verder met klasse $K + 1$, waarbij eventuele secundaire onderdrukkingen in het binnenwerk van een tabel worden meegenomen als primaire onderdrukkingen voor klasse $K + 1$.
5. Wanneer er wel secundaire onderdrukkingen in een marginaal van minimaal één deeltabel gezet moeten worden, ga dan terug naar klasse $K - 1$ waarbij alleen de secundaire onderdrukkingen in de marginalen als primaire onderdrukkingen worden meegenomen.
6. Herhaal stappen 4 en/of 5 totdat alle deeltabellen op het laagste (meest gedetailleerde) hiërarchische niveau zijn beveiligd.

Iedere niet-hiërarchische deeltabel wordt beveiligd met de mixed integer aanpak van Fischetti en Salazar (2000). Bij die aanpak worden de vereiste beveiligingsintervallen gegarandeerd, terwijl een bepaalde kostenfunctie wordt geminimaliseerd. Die

kostenfunctie kan op verschillende manieren gekozen worden, waardoor er verschillende vormen van informatieverlies geminimaliseerd kunnen worden. Die minimalisatie vindt *lokaal* plaats, zodat de uiteindelijke oplossing voor de gehele (hiërarchische) tabel niet noodzakelijk ook optimaal hoeft te zijn.

Voor de kostenfunctie kan in τ-ARGUS o.a. gekozen worden voor:

- een variabele uit de dataset (bijvoorbeeld de kwantitatieve waarde waarover getabelleerd wordt),
- een constante (zodat het aantal onderdrukkingen geminimaliseerd wordt),
- het aantal bijdragers per cel (zodat het totale aantal onderdrukte bijdragen wordt geminimaliseerd).

Bij de beveiliging van een deeltabel, wordt ook rekening gehouden met zogenoemde singletons: cellen met slechts één bijdrage. Wanneer dergelijke cellen in een onderdrukkingspatroon zitten, kunnen de bewuste bijdragers het onderdrukkingspatroon geheel of gedeeltelijk ongedaan maken. Zij weten immers hun eigen bijdrage en kunnen dus die onderdrukte waarde invullen, waardoor mogelijk ook andere onderdrukte cellen teruggerekend kunnen worden. Binnen de huidige implementatie in τ-ARGUS van de mixed integer aanpak, is het niet mogelijk om iedere denkbare combinatie van een singleton met een andere onderdrukte cel onder controle te houden bij het zoeken naar een onderdrukkingspatroon. Wel is het mogelijk om rekening te houden met de combinaties binnen één rij, kolom of laag[2] van de tabel. De combinaties waarmee rekening gehouden moet worden bestaan uit precies twee primair onveilige cellen in één rij, kolom of laag, waarvan minimaal één cel een singleton is. Door de grootste van die twee primair onveilige cellen een beveiligingsinterval te geven dat net niet door de andere primair onveilige cel geleverd kan worden, zal er altijd minimaal één (extra) secundaire onderdrukking in de betreffende rij, kolom of laag geplaatst worden.

Op een vergelijkbare manier wordt ervoor gezorgd, dat binnen één rij, kolom of laag alle onderdrukte cellen samen, meer dan het minimaal vereiste aantal bijdragers voor een veilige cel bevatten.

3.5.3.2 Hypercube aanpak

Voor een gedetailleerdere beschrijving van de hypercube aanpak, zie Giessing en Repsilber (2002).

Ook bij deze aanpak wordt een hiërarchische tabel opgesplitst in niet-hiërarchische deeltabellen. De niet-hiërarchische deeltabellen worden vervolgens in een bepaalde volgorde beveiligd, waarbij de deeltabellen op het hoogste niveau als eerste aan de beurt zijn.

---

[2] Een rij bestaat uit de cellen met de coördinaten (*r*, *k*, *l*) waarbij *k* en *l* vast zijn. Een kolom bestaat uit de cellen met de coördinaten (*r*, *k*, *l*) waarbij *r* en *l* vast zijn. Een laag bestaat uit de cellen met coördinaten (*r*, *k*, *l*) waarbij *r* en *k* vast zijn.

Per deeltabel worden voor iedere primair onveilige cel alle mogelijke hyperkubussen geconstrueerd, waarbij die primair onveilige cel één van de hoekpunten is. Voor iedere hyperkubus wordt het interval berekend waarbinnen de primair onveilige cel nog kan liggen wanneer alle overige hoekpunten van de hyperkubus ook worden onderdrukt. Indien dat interval groot genoeg is (afhankelijk van de gebruikte beveiligingsregel) wordt de bijbehorende hyperkubus toelaatbaar genoemd. Vervolgens wordt voor iedere toelaatbare hyperkubus het informatieverlies berekend. Ten slotte wordt de toelaatbare hyperkubus met het kleinste informatieverlies gekozen om de betreffende primair onveilige cel te beveiligen.

Voor het berekenen van de beveiligingsintervallen ten gevolge van een hyperkubus, hoeft geen lineair programmeringprobleem opgelost te worden. Dit versnelt de procedure aanzienlijk. De hypercube aanpak is dan ook over het algemeen sneller dan de modulaire aanpak, waarbij een mixed integer programmeringprobleem opgelost moet worden.

Nadat alle deeltabellen op deze manier zijn beveiligd, wordt de hele procedure herhaald. Secundair onderdrukte cellen van een bepaalde deeltabel die ook in andere deeltabellen voorkomen, worden in die andere deeltabellen als primair onveilige cellen beschouwd en als zodanig behandeld. Dit proces wordt herhaald, totdat er geen veranderingen meer plaatsvinden.

Het gebruik van hyperkubussen voor de beveiliging van primair onveilige cellen is overigens een voldoende maar niet een noodzakelijke voorwaarde voor een veilig onderdrukkingspatroon. Met andere woorden, in sommige gevallen zal de combinatie van de verschillende hyperkubussen niet tot een optimaal onderdrukkingspatroon leiden, maar wel altijd een veilig onderdrukkingpatroon opleveren. Daardoor heeft deze aanpak de neiging om meer cellen te onderdrukken dan noodzakelijk is voor een veilig onderdrukkingspatroon.

Bij deze aanpak wordt ook rekening gehouden met de zogenaamde singletons. Een cel met slechts één bijdrager zou immers alle onderdrukte hoekpunten van een hyperkubus terug kunnen rekenen. De extra eis bij singletons is dan ook dat zo'n cel een hoekpunt moet zijn van minimaal twee verschillende hyperkubussen.

3.5.4 Voorbeeld

Met behulp van τ-ARGUS kan op een eenvoudige manier celonderdrukking worden toegepast op een kwantitatieve tabel. Zowel de modulaire aanpak als de hypercube aanpak zijn in τ-ARGUS geïmplementeerd. Ook is het mogelijk om meerdere informatieverliesmaten te kiezen voor de kostenfunctie die geminimaliseerd moet worden. Voor het gebruik van τ-ARGUS verwijzen we naar de handleiding (Hundepool et al., 2003).

In Figuur 11 is een voorbeeldtabel opgenomen waarin alleen de primair onveilige cellen zijn onderdrukt.

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
Noord	4,373,664.00	x	x	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
1	1,986,129.00	x	x	398,062.00	348,039.00	354,711.00	418,778.00	466,529.00	-
2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
Oost	3,703,896.00	15.00	x	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
4	124,336.00	x	-	36,311.00	32,132.00	25,770.00	18,150.00	x	-
5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	81,799.00	145,004.00	-
6	2,234,995.00	x	x	345,803.00	251,358.00	251,188.00	303,377.00	1,083,254.00	-
7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
Zuid	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
11	2,752,743.00	-	15.00	488,613.00	392,395.00	363,490.00	402,925.00	1,105,305.00	-
12	1,441,228.00	-	-	212,936.00	209,886.00	254,547.00	244,096.00	519,763.00	-
99	-	-	-	-	-	-	-	-	-

Figuur 11: Kwantitatieve tabel omzet naar regio en grootteklasse

Duidelijk is te zien dat dit niet voldoende is: zowel de cel (Oost, 4) als de cel (4, 9) kan direct uitgerekend worden: (Oost, 4) = 3 703 896 – 15 – 642 238 – 515 003 – 534 147 – 620 392 – 1 392 096 = 5 en (4, 9) = 1 392 096 – 145 004 – 1 083 254 – 151 870 = 11 968.

In Figuur 12 is het onderdrukkingspatroon weergegeven dat met τ-ARGUS is bepaald met behulp van de hypercube aanpak. Figuur 13 geeft hetzelfde weer op basis van de modulaire aanpak. Uiteraard zou in een publicatie geen onderscheid meer gemaakt moeten kunnen worden tussen primaire en secundaire onderdrukkingen.

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
Noord	4,373,664.00	x	x	719,049.00	x	688,962.00	756,529.00	1,549,049.00	385.00
1	1,986,129.00	x	x	398,062.00	x	354,711.00	418,778.00	466,529.00	-
2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
Oost	3,703,896.00	x	x	x	x	534,147.00	620,392.00	1,392,096.00	-
4	124,336.00	x	-	36,311.00	x	25,770.00	18,150.00	x	-
5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	81,799.00	145,004.00	-
6	2,234,995.00	x	x	x	x	251,188.00	303,377.00	x	-
7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
Zuid	4,193,971.00	-	x	x	x	618,037.00	647,021.00	1,625,068.00	-
11	2,752,743.00	-	x	488,613.00	x	363,490.00	402,925.00	1,105,305.00	-
12	1,441,228.00	-	-	x	x	254,547.00	244,096.00	519,763.00	-
99	-	-	-	-	-	-	-	-	-

Figuur 12: Onderdrukkingspatroon voor de tabel uit Figuur 11, m.b.v. de hypercube aanpak

	tot	2	4	5	6	7	8	9	99
tot	16,847,646.84	20.00	25.00	2,711,808.00	2,320,534.00	2,505,042.58	2,799,074.26	6,510,758.00	385.00
Noord	4,373,664.00	×	×	719,049.00	659,680.00	688,962.00	756,529.00	1,549,049.00	385.00
1	1,986,129.00	×	×	398,062.00	348,039.00	354,711.00	418,778.00	466,529.00	-
2	1,809,246.00	0.00	-	223,990.00	221,332.00	241,913.00	258,233.00	863,393.00	385.00
3	578,289.00	-	-	96,997.00	90,309.00	92,338.00	79,518.00	219,127.00	-
Oost	3,703,896.00	×	×	642,238.00	515,003.00	534,147.00	620,392.00	1,392,096.00	-
4	124,336.00	×	-	36,311.00	32,132.00	×	×	×	-
5	526,279.00	-	-	93,589.00	94,957.00	110,930.00	×	×	-
6	2,234,995.00	×	×	345,803.00	251,358.00	×	303,377.00	1,083,254.00	-
7	818,286.00	-	-	166,535.00	136,556.00	146,259.00	217,066.00	151,870.00	-
West	4,576,115.84	-	-	648,972.00	543,570.00	663,896.58	775,132.26	1,944,545.00	-
8	485,326.00	-	-	63,767.00	75,442.00	87,305.00	59,953.00	198,859.00	-
9	3,664,559.84	-	-	537,911.00	430,851.00	515,019.58	643,762.26	1,537,016.00	-
10	426,230.00	-	-	47,294.00	37,277.00	61,572.00	71,417.00	208,670.00	-
Zuid	4,193,971.00	-	15.00	701,549.00	602,281.00	618,037.00	647,021.00	1,625,068.00	-
11	2,752,743.00	-	15.00	488,613.00	392,395.00	363,490.00	402,925.00	1,105,305.00	-
12	1,441,228.00	-	-	212,936.00	209,886.00	254,547.00	244,096.00	519,763.00	-
99	-	-	-	-	-	-	-	-	-

Figuur 13: Onderdrukkingspatroon voor de tabel uit Figuur 11, m.b.v. de modulaire aanpak

3.5.5 Kwaliteitsindicatoren

Wanneer een tabel is beveiligd met behulp van celonderdrukking, is het mogelijk om per onderdrukte cel het gerealiseerde beveiligingsinterval te berekenen. Gegeven het onderdrukkingspatroon en de structuur van de tabel moeten dan per onderdrukte cel twee LP-problemen worden opgelost (minimaliseren en maximaliseren van de waarde voor de onderdrukte cel).

Indien τ-ARGUS wordt gebruikt voor de beveiliging van een kwantitatieve tabel, wordt aan het einde van de sessie een rapport opgemaakt, waarin de genomen stappen en de resultaten daarvan staan vermeld. Ook is het mogelijk om tijdens de sessie al informatie te verkrijgen over de al dan niet beveiligde tabel (bijvoorbeeld: aantal primair onveilige cellen, aantal secundaire onderdrukkingen, informatieverlies).

3.6 Additief afronden

3.6.1 Korte beschrijving

Door celwaarden in een kwantitatieve tabel af te ronden, zijn de exacte celwaarden slechts binnen een bepaald interval bekend. Op deze manier kan een tabel met primair onveilige cellen ook beveiligd worden. De mate waarin wordt afgerond zal uiteraard invloed hebben op de grootte van de intervallen. Wanneer iedere cel onafhankelijk afgerond zou worden, zou de optelbaarheid van de tabel niet noodzakelijk gehandhaafd blijven.

Uiteraard kan op een simpele manier de optelbaarheid gegarandeerd worden: de cellen in het binnenwerk worden onafhankelijk van elkaar afgerond en de marginalen worden opnieuw berekend. Daardoor zullen de marginalen over het algemeen echter vrij veel van de (afgeronde) originele waarden af komen te liggen.

Bij additief afronden wordt de tabel zodanig afgerond, dat de optelbaarheid gehandhaafd blijft en de afgeronde tabel zo min mogelijk van de originele tabel afwijkt. Bovendien is het mogelijk om op een zodanige manier additief af te ronden, dat ook

vooraf gespecificeerde beveiligingsintervallen gegarandeerd kunnen worden. De haalbaarheid van dit laatste, hangt echter af van de grootte van de gekozen afrondbasis in relatie tot de beveiligingsintervallen.

3.6.2 Toepasbaarheid

Additief afronden kan gebruikt worden voor de statistische beveiliging van zowel kwantitatieve tabellen als frequentietabellen. Vaak zal een presentatieargument ook een rol spelen: een groot aantal significante cijfers suggereert een hoge nauwkeurigheid die niet altijd terecht is ten gevolge van steekproeffouten en meetfouten. Door de tabelwaarden af te ronden wordt ook die schijnnauwkeurigheid enigszins ingeperkt.

3.6.3 Uitgebreide beschrijving

Bij additief afronden worden celwaarden in een tabel afgerond op veelvouden van een afrondbasis *b*, waarbij de totalen en subtotalen in de tabel gelijk blijven aan de som van de corresponderende delen.

Vaak wordt additief afronden “zero restricted” uitgevoerd. Dat wil zeggen, celwaarden die al een veelvoud zijn van de afrondbasis worden niet veranderd, terwijl de overige celwaarden worden afgerond op één van de naastliggende veelvouden van die afrondbasis. De afgeronde waarden worden zodanig gekozen, dat de som van de absolute afwijkingen van de celwaarden in de afgeronde tabel ten opzichte van de celwaarden in de originele tabel geminimaliseerd wordt, onder de restrictie dat de afgeronde tabel optelbaar blijft. Hierdoor is het mogelijk dat celwaarden niet worden afgerond op het dichtstbijzijnde veelvoud van de afrondbasis.

In bepaalde omstandigheden is het niet mogelijk om onder het zojuist beschreven scenario een afgeronde tabel te construeren. In dat geval kan de restrictie dat wordt afgerond op één van de naastliggende veelvouden van de afrondbasis worden afgezwakt door toe te laten dat een celwaarde ook mag worden afgerond op niet-naastliggende veelvouden van afrondbasis. Deze afzwakking kan nog enigszins beperkt worden door een maximum te stellen aan het aantal stappen dat de afgeronde waarde van de originele waarde af mag liggen.

In het geval van “zero restricted” additief afronden op afrondbasis $b > 0$ van het niet-negatieve getal $z = ub + r$, met $0 \le r < b$, wordt afgerond op het getal a, waarbij

$$a \in \{ub, (u + 1_{(0,b)}(r))b\} \tag{3.6.1}$$

met $1_{(0,b)}(r)$ gelijk aan 1 als $r \in (0, b)$ en gelijk aan 0 als $r = 0$.

Dus in het geval dat $r = 0$ wordt altijd afgerond op ub en in het geval dat $r \in (0, b)$ wordt afgerond op ub of op $(u + 1)b$.

Wanneer echter de restrictie wordt afgezwakt met maximaal $K > 0$ stappen verder dan de naastliggende veelvouden van de afrondbasis, dan wordt afgerond op het getal a, waarbij

$$a \in \left\{(0 \vee (u+j))b \mid j = -K, \ldots, \left(K + 1_{(0,b)}(r)\right)\right\} \tag{3.6.2}$$

met $x \vee y = \max(x, y)$.

Voor een gegeven tabel kunnen meerdere additief afgeronde versies bestaan. Dit zijn allemaal *toelaatbare* tabellen. Van de toelaatbare tabellen kan dan vervolgens die tabel genomen worden, die het dichtst bij de originele tabel ligt. In τ-ARGUS wordt de afstand die geminimaliseerd wordt gegeven door

$$\sum_{i=1}^{N} \left| z_i - a_i \right| \tag{3.6.3}$$

met N het aantal cellen in de tabel (inclusief alle (sub)-totalen), z_i de celwaarden in de oorspronkelijke tabel en a_i de corresponderende afgeronde celwaarden.

Het vinden van de optimale oplossing, is een rekenintensief probleem (NP-volledig). Voor grote tabellen kan dit tot onacceptabel lange rekentijden leiden. Daarom is in τ-ARGUS een partitionering ingebouwd: een grote tabel kan dan opgesplitst worden in een aantal deeltabellen die afzonderlijk afgerond worden. Nadat die deeltabellen zijn afgerond, worden ze weer gecombineerd waarbij de betreffende (sub)totalen (indien nodig) worden berekend uit de afgeronde delen.

3.6.4 Voorbeeld

Met τ-ARGUS kunnen (kwantitatieve) tabellen eenvoudig additief afgerond worden, waarbij de vereiste beveiligingsmarges gegarandeerd worden.

In Figuur 14 is een voorbeeld tabel opgenomen met daarin een aantal primair onveilige cellen. Figuur 15 bevat de bijbehorende additief afgeronde tabel, bij een afrondbasis van 2000. Uiteraard zouden bij een publicatie de primair onveilige cellen niet meer te herkennen mogen zijn.

	tot	2	4	5	6	7	8	9	99
tot	16,847,647	20	25	2,711,808	2,320,534	2,505,043	2,799,074	6,510,758	385
Noord	4,373,664	5	5	719,049	659,680	688,962	756,529	1,549,049	385
1	1,986,129	5	5	398,062	348,039	354,711	418,778	466,529	-
2	1,809,246	0	-	223,990	221,332	241,913	258,233	863,393	385
3	578,289	-	-	96,997	90,309	92,338	79,518	219,127	-
Oost	3,703,896	15	5	642,238	515,003	534,147	620,392	1,392,096	-
4	124,336	5	-	36,311	32,132	25,770	18,150	11,968	-
5	526,279	-	-	93,589	94,957	110,930	81,799	145,004	-
6	2,234,995	10	5	345,803	251,358	251,188	303,377	1,083,254	-
7	818,286	-	-	166,535	136,556	146,259	217,066	151,870	-
West	4,576,116	-	-	648,972	543,570	663,897	775,132	1,944,545	-
8	485,326	-	-	63,767	75,442	87,305	59,953	198,859	-
9	3,664,560	-	-	537,911	430,851	515,020	643,762	1,537,016	-
10	426,230	-	-	47,294	37,277	61,572	71,417	208,670	-
Zuid	4,193,971	-	15	701,549	602,281	618,037	647,021	1,625,068	-
11	2,752,743	-	15	488,613	392,395	363,490	402,925	1,105,305	-
12	1,441,228	-	-	212,936	209,886	254,547	244,096	519,763	-
99	-	-	-	-	-	-	-	-	-

Figuur 14: Kwantitatieve tabel omzet naar Regio en GrootteKlasse

	tot	2	4	5	6	7	8	9	99
tot	16,848,000	0	0	2,712,000	2,320,000	2,506,000	2,800,000	6,510,000	0
Noord	4,374,000	0	0	720,000	660,000	690,000	756,000	1,548,000	0
1	1,986,000	0	0	398,000	348,000	356,000	418,000	466,000	-
2	1,810,000	0	-	224,000	222,000	242,000	258,000	864,000	0
3	578,000	-	-	98,000	90,000	92,000	80,000	218,000	-
Oost	3,704,000	0	0	642,000	514,000	534,000	622,000	1,392,000	-
4	124,000	0	-	36,000	32,000	26,000	18,000	12,000	-
5	526,000	-	-	94,000	94,000	110,000	82,000	146,000	-
6	2,236,000	0	0	346,000	252,000	252,000	304,000	1,082,000	-
7	818,000	-	-	166,000	136,000	146,000	218,000	152,000	-
West	4,576,000	-	-	648,000	544,000	664,000	776,000	1,944,000	-
8	486,000	-	-	64,000	76,000	88,000	60,000	198,000	-
9	3,664,000	-	-	538,000	430,000	514,000	644,000	1,538,000	-
10	426,000	-	-	46,000	38,000	62,000	72,000	208,000	-
Zuid	4,194,000	-	0	702,000	602,000	618,000	646,000	1,626,000	-
11	2,752,000	-	0	488,000	392,000	364,000	402,000	1,106,000	-
12	1,442,000	-	-	214,000	210,000	254,000	244,000	520,000	-
99	-	-	-	-	-	-	-	-	-

Figuur 15: Tabel uit Figuur 14 additief, beveiligend afgerond met afrondbasis 2000

3.7 Afsluiting

Voor de beveiliging van kwantitatieve tabellen is op het CBS het pakket τ-ARGUS beschikbaar. Voor een uitgebreide beschrijving van dat pakket verwijzen we naar de handleiding ervan (Hundepool et al., 2007).

Wanneer τ-ARGUS wordt gebruikt, wordt na iedere sessie waarin één of meerdere tabellen zijn beveiligd een rapport opgemaakt. In dat rapport is opgenomen welke methoden en parameters gebruikt zijn.

Met τ-ARGUS kunnen de effecten van verschillende statistische beveiligingsmethoden op de tabellen eenvoudig zichtbaar gemaakt worden. De verschillende methoden kunnen worden toegepast, maar binnen dezelfde sessie ook weer ongedaan gemaakt worden.

3.8 Literatuur

Fischetti, M. en Salazar Gonzales, J.J. (2000), *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints*, Journal of the American Statistical Association, vol. 95, pp. 916 – 928.

Giessing, S. en Repsilber, D. (2002), *Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine*, In: 'Inference Control in Statistical Databases' Domingo-Ferrer (Ed.), LNCS 2316, Springer-Verlag Berlin Heidelberg, pp. 181 – 192.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Wolf, P.P. de, Giessing, S., Fischetti, M., Salazar, J.J., Castro, J. en Lowthian, P. (2007), *τ-ARGUS user manual 3.2*, Voorburg.

Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt E. en De Wolf, P.P. (2006), *Handboek Statistische Beveiliging*, BPA 21-06-TMO.

Loeve, A. (2001), Notes on sensitivity measures and protection levels, BPA 01892-01-S-TMO.

De Wolf, P.P. (2002), *HiTaS: A Heuristic Approach to Cell Suppression in Hierarchical Tables*, In: 'Inference Control in Statistical Databases' Domingo-Ferrer (Ed.), LNCS 2316, Springer-Verlag Berlin Heidelberg, pp. 74 – 82.

4. Statistische Beveiliging van Frequentietabellen

4.1 Algemene beschrijving en leeswijzer

4.1.1 Algemene beschrijving

Onder statistische beveiliging van frequentietabellen verstaan we het produceren van frequentietabellen die voldoen aan het CBS-beleid aangaande statistische beveiliging en die als zodanig gepubliceerd kunnen worden. Het beveiligingsbeleid voor frequentietabellen is vastgelegd in hoofdstuk 5 van het Handboek Statistische Beveiliging (Hundepool et al., 2006). Frequentietabellen zijn tabellen waarbij het aantal bijdragers per cel wordt gegeven. Dit in tegenstelling tot kwantitatieve tabellen waarbij de celwaardes zijn ontstaan door sommering van een continue variabele over alle bijdragers tot een cel. Voor kwantitatieve tabellen gelden andere regels en kunnen andere beveiligingsmethoden beter geschikt zijn dan die voor frequentietabellen. Beveiligingsmethoden voor kwantitatieve tabellen worden in het deelthema "Statistische Beveiliging van Kwantitatieve Tabellen" behandeld.

Artikel 37 van de CBS-wet (2004) verplicht tot bescherming van herkenbare gegevens over statistische eenheden. Schending van de statistische geheimhouding ("onthulling") komt dus op de combinatie van twee feiten neer: herkenning van een eenheid èn bekendmaking van nadere gegevens over die eenheid.

Voor frequentietabellen kan dat als volgt worden geformuleerd. De gebruiker moet eerst een bijdrager of groep bijdragers herkennen in de tabel. Daarna volgt een uitspraak over deze bijdrager(s) door de frequentie-verdeling over de cellen. De uitspraak die de tabel over die groep mogelijk maakt, moet meer informatie geven over de leden van de groep dan alleen de groepsgrootte. Kennis die nodig is om de leden van de groep te herkennen vormt in die zin geen informatie over de leden van de groep.

Aan de wettelijke verplichting wordt voldaan indien de tabel geen informatie oplevert over een individuele statistische eenheid als zodanig. De statistische beroepsethiek en het eigen belang van het CBS bij continuïteit van de berichtgeving aan het CBS leiden echter in bepaalde gevallen tot de eis dat de tabel geen informatie oplevert over groepen van statistische eenheden (personen of huishoudens, enz.). Dat is met name het geval indien de tabel variabelen omvat die kwetsende of kwetsbaar makende informatie over die groepen kan opleveren. Zulke gegevens worden hierna "gevoelige gegevens" genoemd.

De in dit deelthema beschreven methoden kunnen door middel van het software pakket τ-ARGUS eenvoudig worden toegepast. Dit pakket is door de sector DMV (en haar voorgangers) in Europees verband ontwikkeld.

4.1.2 Leeswijzer

Als eerste stap bij het bepalen van de juiste statistische beveiliging voor een frequentietabel, zal bepaald moeten worden of er onthulling mogelijk is. Als basis geldt daarbij in eerste instantie het “gezonde verstand”: is er informatie in de tabel aanwezig die niet over individuele respondenten onthuld mag worden? Bij frequentietabellen kan dergelijke informatie verborgen zijn in de opspanvariabelen. Een deel van de opspanvariabelen kan worden beschouwd als identificerend en de rest als gevoelig. Wat betreft gevoeligheid wordt nog onderscheid gemaakt naar mate van gevoeligheid. Zie Hundepool et al. (2006) voor meer informatie.

Nadat onveilige situaties zijn aangewezen, zal over het algemeen de tabel nog verder beveiligd moeten worden. Daarvoor zijn grofweg drie methoden beschikbaar: herstructureren van de tabel (zie paragraaf 4.4), onderdrukken (zie paragraaf 4.5) en afronden (zie paragraaf 4.6).

Welke (combinatie van) methode(n) uiteindelijk in een specifieke situatie gebruikt wordt, is niet op voorhand vast te leggen. Dit hangt sterk af van de beoogde gebruikers. Zo zal het bij Eurostat-verordeningen niet altijd mogelijk zijn om de tabel te herstructureren en zal vaak voor onderdrukking of afronding gekozen (moeten) worden. De afdeling die verantwoordelijk is voor de betreffende frequentietabel, is ook verantwoordelijk voor een adequate (statistische) beveiliging daarvan. Bij de keuze voor de te gebruiken methode(n) moeten twee (concurrerende) aspecten worden meegenomen:

- onthullingrisico;
- informatieverlies.

Algemeen kan gezegd worden, dat verkleining van het onthullingrisico tot meer informatieverlies zal leiden. Ook het omgekeerde is waar: hoe kleiner het informatieverlies, hoe groter het onthullingrisico. In voorkomende gevallen zal een afweging gemaakt moeten worden, waarbij uiteraard wel altijd minimaal aan de regels in het Handboek Statistische Beveiliging (Hundepool et al., 2006) voldaan moet worden.

4.2 Afbakening en relatie met andere (deel)thema’s

In dit deelthema worden methoden behandeld die te gebruiken zijn voor de statistische beveiliging van frequentietabellen. In dit hoofdstuk worden *geen* methoden beschreven die alleen voor kwantitatieve tabellen zijn te gebruiken. Voor dergelijke methoden, zie het deelthema “Statistische Beveiliging van Kwantitatieve Tabellen”.

Enkele in dit hoofdstuk beschreven methoden kunnen in principe zowel bij kwantitatieve tabellen als bij frequentietabellen gebruikt worden. Dergelijke methoden zullen in het deelthema “Statistische Beveiliging van Kwantitatieve Tabellen” herhaald worden.

De methoden in dit hoofdstuk zijn in te delen in twee varianten: methoden voor het bepalen van de (primair) onveilige cellen van een frequentietabel en methoden voor het alsnog publiceerbaar maken van tabellen met onveilige cellen.

4.3 Tijdelijk standaardiseren frequentietabel

4.3.1 Korte beschrijving

Statistische beveiliging heeft tot doel de onthulling van informatie over individuele bijdragers aan een tabel te voorkomen of op z'n minst te bemoeilijken. Om dit te kunnen bereiken, zullen eerst de cellen moeten worden aangewezen waarin risico aanwezig is op een mogelijke onthulling. Bij frequentietabellen spelen (minimaal) twee aspecten daarbij een rol: herkenbare groepen en gevoelige variabelen. Kort gezegd komt het er op neer dat zich een onveilige situatie in een frequentietabel voordoet wanneer ofwel een cel die correspondeert met een herkenbare groep respondenten te weinig respondenten bevat ofwel wanneer de verdeling van de respondenten uit een herkenbare groep te veel geconcentreerd zijn binnen één of twee categorieën. In het Handboek Statistische Beveiliging (Hundepool et al., 2006) is vastgelegd wat het CBS-beleid verstaat onder "te klein" en "te veel geconcentreerd".

Om dit te kunnen doen, is het handig om de frequentietabel in een standaard formaat te bekijken. In sommige gevallen zal de frequentietabel al in dat formaat beschikbaar zijn. In andere gevallen is het nodig om tijdelijk een vertaalslag te maken. Nadat de frequentietabel publicabel gemaakt is, kan die weer in het oorspronkelijke formaat worden omgezet.

4.3.2 Toepasbaarheid

Voordat een frequentietabel statistisch beveiligd kan worden, zal eerst aangegeven moeten worden waar in die tabel zich mogelijk problemen voordoen. Daarvoor is het handig om de frequentietabel (tijdelijk) op een standaard manier te bekijken, waarbij een duidelijk onderscheid zichtbaar is tussen identificerende en gevoelige variabelen.

4.3.3 Uitgebreide beschrijving

Het is voor het detecteren van onveilige situaties in frequentietabellen nodig om de opspanvariabelen in te delen in identificerende variabelen en gevoelige variabelen. De kwalificatie bij ieder variabele is in principe door de betreffende afdeling te bepalen. Om coördinatie tussen de verschillende te publiceren frequentietabellen te bevorderen, is het verstandig om dit centraal bij te houden.

Vervolgens kan de frequentietabel (tijdelijk) zodanig worden ingericht dat de identificerende variabelen in de voorkolom zijn samengebracht en de categorieën van de gevoelige variabelen zijn samengebracht in de overige kolommen. Daarbij moeten ook "verborgen" variabelen worden betrokken die de (deel)populatie definiëren waarover de frequentietabel gaat.

Op de aldus gestandaardiseerde frequentietabel kunnen vervolgens de regels zoals genoemd in Hundepool et al. (2006) worden toegepast.

4.3.4 Voorbeeld

Stel dat Tabel 1 een frequentietabel is van het aantal personen in een bepaald jaar, die een niet-natuurlijke dood zijn gestorven, zoals gegeven in een publicatie[3].

Type		Leeftijd						
niet-natuurlijke dood	Geslacht	Totaal	<15	15-<20	20-<40	40-<60	60-<80	>=80
Zelfdoding	Totaal	1530	8	43	418	674	298	89
	Man	1027	8	34	297	453	181	54
	Vrouw	503	-	9	121	**221**	117	35
Moord en doodslag	Totaal	141	11	13	74	34	8	1
	Man	96	9	5	47	32	2	1
	Vrouw	45	2	8	27	2	6	-
Wegverkeersongeval	Totaal	880	47	120	380	67	179	87
	Man	636	23	87	315	52	98	61
	Vrouw	244	24	33	65	15	81	26
Bedrijfsongeval	Totaal	81	-	3	30	42	6	-
	Man	79	-	3	28	42	6	-
	Vrouw	2	-	-	2	-	-	-
Privé-ongeval	Totaal	2013	64	6	120	60	481	1282
	Man	834	32	2	100	56	223	421
	Vrouw	1179	32	4	20	4	258	**861**
Overig/onbekend	Totaal	110	2	6	24	8	37	33
	Man	63	1	4	18	7	20	13
	Vrouw	47	1	2	6	1	17	20
Totaal	Totaal	4755	132	191	1046	885	1009	1492
	Man	2735	73	135	805	642	530	550
	Vrouw	2020	59	56	241	243	479	942

Tabel 1: *Aantal personen dat in jaar J een niet-natuurlijke dood is gestorven*

De voor statistische beveiliging gestandaardiseerde vorm van deze frequentietabel ontstaat door de identificerende variabelen naar de voorkolom te halen en de gevoelige variabelen in de overige kolommen te zetten. In Tabel 1 zijn de identificerende variabelen “Geslacht” en “Leeftijd”. De gevoelige variabele is de variabele “Type niet-natuurlijke dood”. De gestandaardiseerde versie van Tabel 1 is gegeven in Tabel 2.

[3] De in de tabel genoemde aantallen zijn fictief.

		Type niet-natuurlijke dood						
Geslacht	Leeftijd	Zelf-doding	Moord en doodslag	Wegverkeers-ongeval	Bedrijfs-ongeval	Privé-ongeval	Overig / Onbekend	Totaal
Man	<15	8	9	23	-	32	1	73
	15-<20	34	5	87	3	2	4	135
	20-<40	297	47	315	28	100	18	805
	40-<60	453	32	52	42	56	7	642
	60-<80	181	2	98	6	223	20	530
	>=80	54	1	61	-	421	13	550
	Totaal	1027	96	636	79	834	63	2735
Vrouw	<15	-	2	24	-	32	1	59
	15-<20	9	8	33	-	4	2	56
	20-<40	121	27	65	2	20	6	241
	40-<60	**221**	2	15	-	4	1	243
	60-<80	117	6	81	-	258	17	479
	>=80	35	-	26	-	**861**	20	942
	Totaal	503	43	220	2	1147	46	1961
Totaal	<15	8	11	47	-	64	2	132
	15-<20	43	13	120	3	6	6	191
	20-<40	418	74	380	30	120	24	1046
	40-<60	674	34	67	42	60	8	885
	60-<80	298	8	179	6	481	37	1009
	>=80	89	1	87	-	1282	33	1492
	Totaal	1530	139	856	81	1981	109	4696

Tabel 2: *Gestandaardiseerde versie van Tabel 1*

4.4 Tabel herstructureren

4.4.1 Korte beschrijving

In paragraaf 4.3 is beschreven hoe onveilige situaties bij frequentietabellen ontdekt kunnen worden door de tabel tijdelijk op een gestandaardiseerde manier te bekijken. Indien vervolgens onveilige cellen worden gevonden, zal de tabel beveiligd moeten worden voordat tot publicatie wordt overgegaan. Een eerste mogelijkheid om een frequentietabel met onveilige cellen alsnog publicabel te maken, is door de tabel te herstructureren. Door het samenvoegen van categorieën wordt de vulling per cel vergroot. Daardoor wordt de verdeling over de verschillende categorieën van de gevoelige opspanvariabele(n) beïnvloed. Daarnaast kan met deze methode de vulling per herkenbare groep worden vergroot.

4.4.2 Toepasbaarheid

Deze methode zal er in het algemeen toe leiden dat er minder onveilige cellen in de tabel zullen voorkomen. Door het samenvoegen van rijen en/of kolommen worden cellen samengevoegd en wordt de vulling per cel vergroot. Ook de verdeling over de verschillende categorieën van de gevoelige opspanvariabele(n) wordt daardoor beïnvloed.

Er zijn geen methodologische voorwaarden voor het toepassen van deze methode. Echter, door extern opgelegde leveringsverplichtingen wordt soms voorgeschreven in welke mate van detail een tabel dient te worden gepubliceerd. Dit kan een Eurostat-verplichting zijn, maar ook het CBS-beleid kan ertoe leiden dat een bepaald detailniveau van een tabel moet worden gepubliceerd. In die gevallen kan de metho-

de technisch gezien dus wel toegepast worden, maar wordt dit door (externe) beleidsbeslissingen verhinderd.

4.4.3 Uitgebreide beschrijving

In de gestandaardiseerde versie van de frequentietabel wordt bepaald of zich een onveilige situatie voordoet. De herstructurering kan op twee manieren worden toegepast:

a. Herstructureren van de oorspronkelijke tabel

b. Herstructureren van de gestandaardiseerde versie van de tabel

Wanneer voor a wordt gekozen, zal na herstructurering de tabel opnieuw in de gestandaardiseerde vorm bekeken moeten worden om na te gaan of dan wel aan de beveiligingsregels wordt voldaan. Ook zal op basis van de gestandaardiseerde vorm bepaald moeten worden om welke onveilige cellen het gaat. In geval van optie b is direct duidelijk om welke cellen het gaat, maar zal de herstructurering nog vertaald moeten worden naar de oorspronkelijke tabel.

4.4.4 Voorbeeld

In Tabel 2 blijkt dat de verdeling van de respondenten over de verschillende categorieën van de gevoelige variabele in twee rijen niet voldoet aan de beveiligingsregels. Deze regels schrijven voor dat niet bijna alle respondenten in een cel zijn geconcentreerd. De cel (Vrouw, 80+, Privé-ongeval) heeft 91% van de totale groep vrouwen van 80+ die een niet-natuurlijke dood zijn gestorven en de cel (Vrouw, 40-60, Zelfdoding) heeft ook 91% van de totale groep van vrouwen tussen de 40 en 60 jaar die een niet-natuurlijke dood zijn gestorven.

Uitgaande van de originele tabel, zou gekozen kunnen worden voor het niet naar geslacht uitsplitsen van doodsoorzaken “Zelfdoding” en “Privé-ongeval”.

Uitgaande van de gestandaardiseerde vorm, zou gekozen kunnen worden voor het indikken van de leeftijdscategorieën tot “<15”, “15-<20”, “20-<60” en “>=60”. Dit resulteert in een tabel waarin geen sterk in één cel geconcentreerde verdeling van herkenbare groepen meer voorkomt. Zie Tabel 3 voor de bijbehorende tabel in originele vorm.

Type niet-natuurlijke dood	Geslacht	Leeftijd Totaal	<15	15-<20	20-<60	>=60
Zelfdoding	Totaal	1530	8	43	1092	387
	Man	1027	8	34	750	235
	Vrouw	503	-	9	342	152
Moord en doodslag	Totaal	141	11	13	108	9
	Man	96	9	5	79	3
	Vrouw	45	2	8	29	6
Wegverkeersongeval	Totaal	880	47	120	447	266
	Man	636	23	87	367	159
	Vrouw	244	24	33	80	107
Bedrijfsongeval	Totaal	81	-	3	72	6
	Man	79	-	3	70	6
	Vrouw	2	-	-	2	-
Privé-ongeval	Totaal	2013	64	6	180	1763
	Man	834	32	2	156	644
	Vrouw	1179	32	4	24	1119
Overig/onbekend	Totaal	110	2	6	32	70
	Man	63	1	4	25	33
	Vrouw	47	1	2	7	37
Totaal	Totaal	4755	132	191	1931	2501
	Man	2735	73	135	1447	1080
	Vrouw	2020	59	56	484	1421

Tabel 3: Beveiligde versie van Tabel 1

4.5 Onderdrukken

4.5.1 Korte beschrijving

Een veel gebruikte methode om primair onveilige cellen te beveiligen, is het onderdrukken (niet publiceren) van bepaalde cellen. De celwaarde wordt dan eenvoudig vervangen door een kruisje (×).

In een frequentietabel waarbij (sub)totalen ook gegeven zijn, is het echter vaak niet voldoende om alleen de primair onveilige cellen te onderdrukken. Wanneer een onderdrukte cel de enige onderdrukte cel in een rij is, is de onderdrukte waarde immers eenvoudig uit te rekenen door de overige celwaarden in die rij van de bijbehorende marginaal af te trekken.

Om primair onveilige cellen toch voldoende te kunnen beschermen is het dan ook noodzakelijk om ook andere, zelf veilige, cellen te onderdrukken. Dit heet *secundair onderdrukken*. Het is niet eenvoudig om dit op een zodanige manier te doen dat de primair onveilige cellen voldoende worden beveiligd, terwijl er ook niet te veel informatie uit de tabel wordt weggehaald. Bovendien moet ook rekening worden gehouden met het feit dat structurele nulcellen niet gebruikt kunnen worden als secundaire onderdrukkingen: iedereen weet immers dat die cellen per definitie leeg zijn.

Om te voorkomen dat onderdrukte, primair onveilige cellen exact kunnen worden teruggerekend, zijn dus secundaire onderdrukkingen nodig. Echter, ook nu speelt weer een rol dat een “te nauwkeurige” schatting voor een onderdrukte cel niet gewenst is. Wat is immers het verschil tussen de uitspraak “Deze onderdrukte cel heeft eigenlijk een waarde van 10000” en “Deze onderdrukte cel heeft eigenlijk een waar-

de tussen 9998 en 10002". Gegeven een onderdrukkingspatroon is het altijd[4] mogelijk om een interval te berekenen waarbinnen een onderdrukte cel moet liggen. De methode "Onderdrukken" moet dan ook een onderdrukkingspatroon opleveren, waarbij de te berekenen intervallen groot genoeg zijn.

Fischetti en Salazar (2000) hebben een methode bedacht om op een optimale manier bovenstaand probleem op te lossen. Hun methode is in theorie toepasbaar op willekeurige, additieve tabellen met niet-negatieve bijdragers. In de praktijk blijkt hun oplossing echter te veel rekentijd te kosten wanneer de tabellen groot worden in omvang of complexiteit. Vandaar dat er een aantal sub-optimale methoden zijn ontwikkeld voor het vinden van geschikte onderdrukkingspatronen voor grotere en/of complexere tabellen.

Zo splitst de "modulaire aanpak" (HiTaS) een hiërarchische tabel op in een groot aantal niet-hiërarchische deeltabellen en past de optimale methode toe op iedere afzonderlijke deeltabel. Door de resultaten op een juiste manier te combineren is een suboptimale oplossing voor de gehele tabel te krijgen, in een aanzienlijk kortere rekentijd.

De "hypercube aanpak" kan ook grote tabellen beveiligen door op een bepaalde iteratieve manier deeltabellen te beveiligen. De beveiliging van iedere deeltabel vindt ook op een suboptimale manier plaats. Daardoor is de aanpak relatief snel, maar worden over het algemeen meer cellen onderdrukt dan strikt noodzakelijk is om een beveiligde tabel te krijgen.

4.5.2 Toepasbaarheid

Onveilige situaties in frequentietabellen zijn onder te verdelen in twee gevallen:

a. De herkenbare groep is te klein;

b. De verdeling van de herkenbare groep over de gevoelige variabele(n) is te veel geconcentreerd in één (gevoelige) cel.

Voor het bepalen van een geschikt onderdrukkingspatroon is het nodig om te weten op welke manier voldaan kan worden aan de gestelde beveiligingsregels. In veel algoritmes worden daar zogenaamde veiligheidsintervallen (safety ranges) voor gebruikt. Dit zijn de minimale intervallen voor primair onderdrukte cellen, die zouden moeten volgen uit het onderdrukkingspatroon. Vooralsnog is, in tegenstelling tot het geval van kwantitatieve tabellen, nog geen methodiek beschikbaar om minimale intervallen te berekenen voor primair onveilige cellen in frequentietabellen. De methodiek zoals beschreven in Fischetti en Salazar (2000) is dan ook vooralsnog niet direct toepasbaar.

[4] In het geval dat de tabel is opgebouwd uit begrensde (b.v. niet-negatieve) bijdragers en de marginalen gegeven zijn.

4.5.3 Uitgebreide beschrijving

Wanneer een rijtotaal in de gestandaardiseerde vorm van de tabel te klein is (de herkenbare groep is te klein) zal die cel onderdrukt moeten worden. Uiteraard zullen meerdere cellen onderdrukt moeten worden om te voorkomen dat het rijtotaal weer kan worden teruggerekend. Over het algemeen zal dit betekenen dat de totale rij onderdrukt zal moeten worden, inclusief een tweede mogelijk “veilige” rij.

Een tweede situatie die zich voor kan doen bestaat uit een voldoend groot rijtotaal dat echter te veel geconcentreerd is in één gevoelige categorie van de variabele. In dat geval is het rijtotaal in principe publiceerbaar. De cel behorende bij de categorie van de gevoelige variabele waarin de respondenten zijn geconcentreerd, kan dan gezien worden als primair te onderdrukken cel. In een tabel met (sub)totalen moeten dan ook secundair te onderdrukken cellen gezocht worden. In veel algoritmes worden daar veiligheidsintervallen voor gebruikt. Dit zijn de minimale intervallen voor primair onderdrukte cellen, die zouden moeten volgen uit het onderdrukkingspatroon. Vooralsnog is, in tegenstelling tot het geval van kwantitatieve tabellen, nog geen methodiek beschikbaar om minimale intervallen te berekenen voor primair onveilige cellen in frequentietabellen. De methodiek zoals beschreven in Fischetti en Salazar (2000) is dan ook vooralsnog niet direct toepasbaar.

Een bijkomend probleem wordt gevormd door de in de beveiligingsregels genoemde “zinvolle aggregaten”. Wanneer meerdere cellen in een rij worden onderdrukt, wordt eigenlijk het totaal van die onderdrukte cellen gepubliceerd. Als de onderdrukte cellen een zinvol aggregaat vormen, dan mogen de respondenten ook niet te veel geconcentreerd zijn in die gecombineerde cel. Bij het bepalen van secundaire onderdrukkingen zou daar dus rekening mee gehouden moeten worden. Het is nog niet duidelijk of het model van Fischetti en Salazar (2000) algemeen genoeg is om daar rekening mee te kunnen houden.

4.5.4 Voorbeeld

In Tabel 4 is een onderdrukkingspatroon opgenomen, waarbij is aangenomen dat het aggregaat “Zelfdoding” + “Privé-ongeval” geen “zinvol aggregaat”is. Beide problematische cellen worden onderdrukt door het plaatsen van kruisjes.

Type niet-natuurlijke dood	Geslacht	Leeftijd Totaal	<15	15-<20	20-<40	40-<60	60-<80	>=80
Zelfdoding	Totaal	1530	8	43	418	674	298	89
	Man	1027	8	34	297	×	181	×
	Vrouw	503	-	9	121	×	117	×
Moord en doodslag	Totaal	141	11	13	74	34	8	1
	Man	96	9	5	47	32	2	1
	Vrouw	45	2	8	27	2	6	-
Wegverkeersongeval	Totaal	880	47	120	380	67	179	87
	Man	636	23	87	315	52	98	61
	Vrouw	244	24	33	65	15	81	26
Bedrijfsongeval	Totaal	81	-	3	30	42	6	-
	Man	79	-	3	28	42	6	-
	Vrouw	2	-	-	2	-	-	-
Privé-ongeval	Totaal	2013	64	6	120	60	481	1282
	Man	834	32	2	100	×	223	×
	Vrouw	1179	32	4	20	×	258	×
Overig/onbekend	Totaal	110	2	6	24	8	37	33
	Man	63	1	4	18	7	20	13
	Vrouw	47	1	2	6	1	17	20
Totaal	Totaal	4755	132	191	1046	885	1009	1492
	Man	2735	73	135	805	642	530	550
	Vrouw	2020	59	56	241	243	479	942

Tabel 4: Onderdrukkingspatroon voor beveiliging van Tabel 1

4.6 Additief afronden

4.6.1 Korte beschrijving

Bij frequentietabellen is afronding een vrij natuurlijke beveiligingsmethode. In de eerste plaats zijn de exacte celwaarden slechts binnen een bepaald interval bekend wanneer is afgerond. De mate waarin wordt afgerond zal uiteraard invloed hebben op de grootte van de intervallen. Ten tweede geeft een niet afgeronde frequentietabel de indruk van een grote precisie: er zou dan immers geteld zijn tot op de individuele eenheden toe. In geval van geschatte frequenties is dat een schijnnauwkeurigheid. Die schijnnauwkeurigheid wordt door af te ronden ook beperkt.

Wanneer iedere cel onafhankelijk afgerond zou worden, zou de optelbaarheid van de tabel niet noodzakelijk gehandhaafd blijven. Uiteraard kan op een simpele manier de optelbaarheid gegarandeerd worden: de cellen in het binnenwerk worden onafhankelijk van elkaar afgerond en de marginalen worden opnieuw berekend. Daardoor kunnen de marginalen echter vrij veel van de (afgeronde) originele waarden af komen te liggen.

Bij additief afronden wordt de tabel zodanig afgerond, dat de optelbaarheid gehandhaafd blijft en de afgeronde tabel zo min mogelijk van de originele tabel afwijkt. De grootte van de afrondbasis bepaalt in welke mate de frequentietabel wordt beveiligd: hoe groter de afrondbasis hoe groter de beveiliging in het algemeen zal zijn. Voor het bepalen van de juiste afrondbasis is vooralsnog geen methodiek beschikbaar.

4.6.2 Toepasbaarheid

Additief afronden kan gebruikt worden voor de statistische beveiliging van zowel kwantitatieve tabellen als frequentietabellen. Vaak zal een presentatieargument ook een rol spelen: een groot aantal significante cijfers suggereert een hoge nauwkeurigheid die niet altijd terecht is ten gevolge van steekproeffouten en meetfouten. Door de tabelwaarden af te ronden wordt ook die schijnnauwkeurigheid enigszins ingeperkt.

4.6.3 Uitgebreide beschrijving

Bij additief afronden worden celwaarden in een tabel afgerond op veelvouden van een afrondbasis *b*, waarbij de totalen en subtotalen in de tabel gelijk blijven aan de som van de corresponderende delen.

Vaak wordt additief afronden “zero restricted” uitgevoerd. Dat wil zeggen, celwaarden die al een veelvoud zijn van de afrondbasis worden niet veranderd, terwijl de overige celwaarden worden afgerond op één van de naastliggende veelvouden van die afrondbasis. De afgeronde waarden worden zodanig gekozen, dat de som van de absolute afwijkingen van de celwaarden in de afgeronde tabel ten opzichte van de celwaarden in de originele tabel geminimaliseerd wordt, onder de restrictie dat de afgeronde tabel optelbaar blijft. Hierdoor is het mogelijk dat celwaarden niet worden afgerond op het dichtstbijzijnde veelvoud van de afrondbasis.

In bepaalde omstandigheden is het niet mogelijk om onder het zojuist beschreven scenario een afgeronde tabel te construeren. In dat geval kan de restrictie dat wordt afgerond op één van de naastliggende veelvouden van de afrondbasis worden afgezwakt door toe te laten dat een celwaarde ook mag worden afgerond op niet-naastliggende veelvouden van de afrondbasis. Deze afzwakking kan nog enigszins beperkt worden door een maximum te stellen aan het aantal stappen dat de afgeronde waarde van de originele waarde af mag liggen.

In het geval van “zero restricted” additief afronden op afrondbasis $b > 0$ van het niet-negatieve getal $z = ub + r$, met $0 \le r < b$, wordt afgerond op het getal a, waarbij

$$a \in \{ub, (u + 1_{(0,b)}(r))b\} \tag{4.6.1}$$

met $1_{(0,b)}(r)$ gelijk aan 1 als $r \in (0, b)$ en gelijk aan 0 als $r = 0$.

Dus in het geval dat $r = 0$ wordt altijd afgerond op ub en in het geval dat $r \in (0, b)$ wordt afgerond op ub of op $(u + 1)b$.

Wanneer echter de restrictie wordt afgezwakt met maximaal $K > 0$ stappen verder dan de naastliggende veelvouden van de afrondbasis, dan wordt afgerond op het getal a, waarbij

$$a \in \{(0 \vee (u + j))b \mid j = -K, \ldots, (K + 1_{(0,b)}(r))\} \tag{4.6.2}$$

met $x \vee y = \max(x, y)$.

Voor een gegeven tabel kunnen meerdere additief afgeronde versies bestaan. Dit zijn allemaal *toelaatbare* tabellen. Van de toelaatbare tabellen kan dan vervolgens die tabel genomen worden, die het dichtst bij de originele tabel ligt. In τ-ARGUS (zie Hundepool et al. 2007) wordt de afstand die geminimaliseerd wordt gegeven door

$$\sum_{i=1}^{N} |z_i - a_i| \tag{4.6.3}$$

met N het aantal cellen in de tabel (inclusief alle (sub)-totalen), z_i de celwaarden in de oorspronkelijke tabel en a_i de corresponderende afgeronde celwaarden.

Het vinden van de optimale oplossing, is een rekenintensief probleem (NP-volledig). Voor grote tabellen kan dit tot onacceptabel lange rekentijden leiden. Daarom is in τ-ARGUS een partitionering ingebouwd: een grote tabel kan dan opgesplitst worden in een aantal deeltabellen die afzonderlijk afgerond worden. Nadat die deeltabellen zijn afgerond, worden ze weer gecombineerd waarbij de betreffende (sub)totalen (indien nodig) worden berekend uit de afgeronde delen.

4.6.4 Voorbeeld

In Tabel 5 is een afgeronde versie van Tabel 1 opgenomen, waarbij additief is afgerond met afrondbasis 50.

Type		Leeftijd						
niet-natuurlijke dood	Geslacht	Totaal	<15	15-<20	20-<40	40-<60	60-<80	>=80
Zelfdoding	Totaal	1550	0	50	400	700	300	100
	Man	1050	0	50	300	450	200	50
	Vrouw	500	-	0	100	250	100	50
Moord en doodslag	Totaal	150	0	0	100	50	0	0
	Man	100	0	0	50	50	0	0
	Vrouw	50	0	0	50	0	0	-
Wegverkeersongeval	Totaal	850	50	150	350	50	200	50
	Man	600	0	100	300	50	100	50
	Vrouw	250	50	50	50	0	100	0
Bedrijfsongeval	Totaal	100	-	0	50	50	0	-
	Man	100	-	0	50	50	0	-
	Vrouw	0	-	-	0	-	-	-
Privé-ongeval	Totaal	2000	50	0	150	50	450	1300
	Man	850	50	0	100	50	200	450
	Vrouw	1150	0	0	50	0	250	850
Overig/onbekend	Totaal	100	0	0	0	0	50	50
	Man	50	0	0	0	0	50	0
	Vrouw	50	0	0	0	0	0	50
Totaal	Totaal	4750	100	200	1050	900	1000	1500
	Man	2750	50	150	800	650	550	550
	Vrouw	2000	50	50	250	250	450	950

Tabel 5: Afgeronde versie van Tabel 1, afrondbasis 50. Een "0" is een afgeronde 0, een "-" is een lege cel

4.7 Literatuur

CBS (2004) *Wet op het Centraal Bureau voor de Statistiek*, Staatsblad 2004, 695, Zie ook: http://www.cbs.nl/nl-NL/menu/organisatie/corporate-informatie/default.htm

Fischetti, M. en Salazar Gonzales, J.J. (2000), *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints*, Journal of the American Statistical Association, vol. 95, pp. 916 – 928.

Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt E. en De Wolf, P.P. (2006), *Handboek Statistische Beveiliging*, BPA 21-06-TMO.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J. en Lowthian, P. (2007), *τ-ARGUS user manual 3.2*, Voorburg.

5. Statistische Beveiliging van Analyseresultaten

5.1 Algemene beschrijving en leeswijzer

5.1.1 Algemene beschrijving

Naast de in de eerdere paragrafen beschreven problemen en methoden voor het beveiligen van microdata, kwantitatieve tabellen en frequentietabellen is er nog een zeer ruime, diverse groep van statistisch output. Dit betreft de resultaten van allerlei statistische analyses en modelschattingen. Ook deze resultaten hebben in principe een risico op de onthulling van de gegevens van individuele respondenten en moeten dus voorzichtig behandeld worden. Vooral bij uitschieters is de kans op onthulling aanwezig. Bij het bepalen of deze resultaten voldoende veilig zijn, wordt vaak gekeken naar de onderliggende frequentietabellen. Vaak is er een sterk verband tussen het model van de analyse en een onderliggende frequentietabel. In het handboek Statistische Beveiliging (Hundepool et al., 2006) is ook een eerste aanzet voor de beveiliging van analyse-resultaten te vinden.

5.1.2 Leeswijzer

Het probleem van de bepaling of de uitkomsten van statistische analyses voldoende veilig zijn komt vooral voor bij het controleren van de output van het OnSite werken en bij Remote Access. Dat is waar veel statistische analyses op onbeveiligde data worden verricht, terwijl de gebruikers de resultaten van hun onderzoek graag willen meenemen buiten het CBS en publiceren. Het controleren van de output is een noodzakelijk onderdeel van deze gewaardeerde service van het CBS en statistische bureaus in het algemeen. Qua controle van de output maakt het in het geheel niet uit of de output via OnSite dan wel via Remote Access is verkregen. In beide gevallen wordt op dezelfde databestanden met dezelfde hulpmiddelen (SPSS, SAS etc.) dezelfde analyses uitgevoerd.

Omdat dit probleem niet alleen bij het CBS voorkomt, maar eigenlijk bij elk statistisch bureau in Europa, is besloten dit een onderwerp te maken van het ESSNet project Statistische Beveiliging (2008-2009). Het ESSNet is gesubsidieerd door Eurostat. Het CBS leverde de projectleider van dit project. Een van de taken in het ESSNet-project was het opstellen van richtlijnen voor het controleren van output. Voor dit onderwerp in de Methodenreeks wordt dan ook gebruik gemaakt van deze “Guidelines for Output Checking”, die te vinden zijn op de ESSnet website (http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf).

Het onderwerp is nog in ontwikkeling en de projectgroep ziet de huidige versie dan ook niet als de laatste wijsheid, maar wel als een zeer goed bruikbare eerste versie. De projectgroep hoopt dan ook dat zij in een volgend project in staat gesteld zal worden haar werk op dit gebied voort te zetten.

Door de diversiteit van het probleem, zowel qua aantal mogelijke analyse-methoden als wel het diverse aantal statistische pakketten met ieder hun eigen output-vormen is het ook ondoenlijk hiervoor kant en klare software te ontwikkelen.

5.2 Afbakening en relatie met andere (deel)thema's

In dit deelthema worden methoden behandeld die te gebruiken zijn om te bepalen of de resultaten van statistische analyses voldoende veilig zijn. Er wordt in grote mate gebruik gemaakt van de resultaten van een Europese projectgroep die richtlijnen (guidelines) heeft opgesteld. In die richtlijnen wordt ook gesproken over tabellen. Maar aangezien die onderwerpen al afgedekt zijn in de voorafgaande hoofdstukken, zijn die onderwerpen van de guidelines hier minder relevant.

5.3 Beveiliging van analyseresultaten

De methoden van de beveiliging van analyse-resultaten sluiten aan bij deze Europese richtlijnen.

Bij het opstellen van de "guidelines for output checking" heeft een aantal overwegingen een rol gespeeld. Uiteraard is het ondoenlijk alle mogelijke vormen van output in extenso te behandelen. Het aantal verschillende methodes, beschikbaar in SAS en SPSS is dermate groot, dat het onmogelijk is alle methodes op zijn mogelijke onthullingsrisico's te beoordelen. Men denke slechts aan de omvang van de SPSS- of SAS-documentatie.

Een ander aspect dat een belangrijke rol speelt in de richtlijnen is de praktische uitvoerbaarheid. Bij het beoordelen van output moeten we rekening houden met twee mogelijke fouten: ten eerste het ten onrechte goedkeuren van onveilige resultaten en ten tweede het ten onrechte tegenhouden van veilige resultaten.

In de richtlijnen worden dan ook voor elk onderwerp twee methoden aangegeven. Een "Rule-of-Thumb", die vooral de eerste fout minimaliseert en een "principles-based" rule die beide fouten probeert te minimaliseren.

De gedachte achter deze tweedeling is dat veel output van onderzoek met geringe inspanning door de eenvoudige regel afgedaan kan worden. Indien de output niet door de "Rule-of-Thumb" wordt doorgelaten en de onderzoeker toch prijs stelt op de goedkeuring, moet er extra werk (ook door de onderzoeker) verricht worden om aan te tonen dat de resultaten toch veilig zijn.

Typen uitvoer die op dit moment in de richtlijnen behandeld worden zijn:

Descriptive statistics	Frequency tables
	Magnitude tables
	Maxima, minima and percentiles (incl. median)
	Mode
	Means, indices, ratios, indicators
	Concentration ratios
	Higher moments of distributions (incl. variance, co-variance, kurtosis, skewness)
	Graphs: pictorial representations of actual data
Correlation and	Linear regression coefficients

Regression Analysis	Non-linear regression coefficients
	Estimation residuals
	Summary and test statistics from estimates (R^2, χ^2 etc.)
	Correlation coefficients

Voor de rest wordt hier verwezen naar de Europese Guidelines, i.p.v. een vertaling te maken en een complete kopie in te voegen.

5.4 Literatuur

Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt, E. en De Wolf, P.P. (2006), *Handboek Statistische Beveiliging*, BPA 21-06-TMO.

Ritchie, F., Welpton, R., Franconi, L., Lucarelli, M., Seri, G., Brandt, M., Guerke, C., Hundepool, A.J. and Mol, J. (2010), *Guidelines for the checking of output based on microdata research*, ESSNet-SDC-project. http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf