# *Geheimhouding bij bedrijfsstatistieken van het CBS: Geheimhoudingsmodule 2000*

Wim Koopman en Ab Mulder [1]

*Het Nederlandse Centraal Bureau voor de Statistiek (CBS) wordt al heel lang met het probleem van de vertrouwelijkheid geconfronteerd. De publicaties van veel statistieken, waaronder ook de bedrijfsstatistieken, zijn immers opgebouwd uit vertrouwelijke gegevens.*

In de (Nederlandse) Wet op de Economische Statistieken van 1936 (een opvolger van voorgaande wetten op dit punt) is een paragraaf opgenomen die de veiligheid van de gegevens van de individuele bedrijven waarborgt. Letterlijk heet het daar:
*'Gegevens, ingevolge deze wet verzameld, worden niet in zoodanigen vorm openbaar gemaakt, dat daaruit opgaven en inlichtingen over een afzonderlijken persoon, onderneming of instelling kunnen blijken, tenzij die persoon, het hoofd van de onderneming of het bestuur der instelling hiertegen geen bezwaar heeft'.*

*Geheimhouding in de statistiek*

De uitvoering van deze waarborg wordt door het CBS aan twee eenvoudige basisregels toevertrouwd:
- 1. Gegevens worden niet gepubliceerd als ze afkomstig zijn van minder dan 'N' individuele bedrijven en;
- 2. Gegevens worden niet gepubliceerd als één individueel bedrijf voor meer dan 'X' procent verantwoordelijk is voor deze gegevens.

Beide waarden, 'N' en 'X', worden om veiligheidsredenen (ook in deze publicatie) eveneens geheim gehouden. Ook dat bevordert de doelmatigheid van de verhulling. Gegevens die op basis van deze twee regels niet mogen worden gepubliceerd noemen we *primair geheim*.

In het algemeen bestaat de behoefte om zowel details als totalen te publiceren. De gebruikers van de statistieken verwachten dat ook. De primair geheime gegevens kunnen dan soms alsnog worden berekend uit het verschil tussen de totalen en de niet geheime details. Om dit te voorkomen moeten in dat geval ook andere gegevens, die op zichzelf niet geheim zijn, toch van publicatie worden uitgesloten. Dit noemen *we secundaire geheimhouding*. Zoals hierna zal blijken is dit onderscheid noodzakelijk bij het proces tot vaststelling van de geheime cellen. Bij publicatie is het onderscheid niet meer van belang. Met andere woorden: de gebruikers van een publicatie zien alleen maar of iets geheim is, niet waarom en hoe.

*Het probleem*

Waar de primaire geheimhouding relatief simpel kan worden berekend door toepassing van de genoemde twee regels zit het probleem met name bij de bepaling van de aanvullende, secundaire, geheimhouding. Omdat de secundaire geheimhouding minder eenduidig is vast te stellen dan de primaire (er zijn, zoals zal blijken, nogal wat keuzemogelijkheden), en omdat we te maken hebben met publicaties vanuit verschillende statistieken, die bovendien ook nog allerlei onderlinge relaties hebben, is dit probleem complexer dan het op het eerste gezicht lijkt. Bovendien wordt uit een dataset vaak door de verschillende onderdelen van het statistisch bureau op verschillende tijden en op heel diverse (hiërarchische) detailniveaus gepubliceerd.
Tot voor kort was er geen tijd of capaciteit om alle geheimhouding goed door te rekenen, en bovendien was het probleem er te ingewikkeld voor. Er bestond nog geen afdoend systeem voor de geheimhouding[2]. Veel regionale gegevens werden daarom vaak zekerheidshalve maar geheim verklaard en niet eens gepubliceerd. Soms ook werden er jarenlang dezelfde publicatieschema's gebruikt. Daardoor kwam het voor dat cijfers van een grote stad als Amsterdam geheim waren, terwijl van een (economisch) veel kleinere Nederlandse provincie als Drenthe veel meer gegevens beschikbaar werden gesteld.

In de jaren tachtig groeide het gebruik van de computer exponentieel. De komst hiervan in de statistische wereld maakte snelle invoer en verwerking van gegevens mogelijk. Via internet worden tegenwoordig data zo mogelijk rechtstreeks naar grote klanten als Eurostat toegestuurd. Ook de vraag naar gegevens is de laatste jaren gegroeid. Er worden steeds meer detailgegevens op steeds kortere termijn gevraagd. Kortom: er is een activiteit van jewelste die geen tijdrovende acties als geheimhoudingsbepaling oude stijl meer toelaat. Geen mens zou meer kunnen en willen begrijpen dat cijfers twee of drie maanden worden opgehouden voor publicatie omdat de geheimhouding nog moet worden bepaald.

Eind van de jaren negentig groeide het interne geheugen en vooral daardoor de prestaties van computers zodanig, dat wat medio negentiger jaren nog voor onmogelijk werd gehouden, nu goed mogelijk is. De tijd is daarom ook technisch rijp voor een passende oplossing. Snel, betrouwbaar, en voor alle resultaten uit een willekeurige dataset.

*Geheimhoudingsmodule 2000*

In 1997 is bij de statistische afdelingen van het Nederlandse Bureau voor de Statistiek een begin gemaakt met de inventarisatie van vereisten voor een geheimhoudingsmodule voor bedrijfsstatistieken. Grofweg komt dat overeen met wat in het voorgaande beschreven is.
Sedert augustus 1999 is een programma ontwikkeld, 'Geheimhoudingsmodule 2000[3]' , dat een driedimensionale tabel van vele niveaus kan beveiligen. De drie dimensies kunnen volkomen flexibel, volgens een hiërarchische classificatie (zie verderop in de paragraaf classificaties), worden ingelezen.

De methode volgens welke het programma werkt, wordt hierna globaal uiteengezet.

Het programma verwerkt een dataset met een tabel van circa 25 miljoen cellen op een snelle machine in niet meer dan enkele minuten. De resultaten van het programma worden voor later gebruik opgeslagen in het Microlab[4]. Ze vormen daar een zogenaamde 'geheimhoudingskubus' (driedimensionaal), zodat iedere tabelpublicatie die via het Microlab[5] loopt netjes beveiligd kan worden.

Implementatie voor alle statistieken zal nog wel even tijd kosten. Voorwaarde is immers dat logische classificaties worden gebruikt en dat dit bij alle betrokken statistieken op dezelfde manier gebeurt. Bovendien zijn er meer afspraken nodig over het aanbrengen van een grote mate van samenhang tussen voordien min of meer afzonderlijke opererende statistieken. Voor een deel van de statistieken zijn dergelijke afspraken al gemaakt.

*Waarborgen van de geheimhouding*

In het onderstaande worden allereerst aan de hand van eenvoudige tweedimensionale voorbeelden een aantal noodzakelijke aanvullende regels vastgesteld, die we in het vervolg van dit betoog voor het gemak de secundaire regels zullen noemen.
We gaan uit van de volgende tabel met willekeurig gekozen getallen, waarbij P staat voor primair geheim:

*Tabel 1*

| *Regio* / *Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | 50 | 80 |
| *Handel* | 30 | 30 | 60 |
| *Dienstverlening* | 40 | 30 | 70 |
| *Overig* | 10 | 0 | 10 |
| *Totaal* | 110 | 110 | 220 |

Zowel via de kolom als via de rij is hier het cijfer van de industrie uit de regio Noord rechtstreeks afleidbaar, namelijk 30. Het blijkt dus niet voldoende om alleen de cellen waarin iets aan de hand is geheim te maken. Er is meer nodig. In vaktermen: behalve de primaire geheimhouding moet ook secundaire geheimhouding worden toegevoegd om de primaire geheimhouding te waarborgen.
Uit de tabel zien we al dat in alle dimensies secundaire geheimhouding moet worden toegevoegd. Alleen in de rij of alleen in de kolom volstaat niet.
Een eerste secundaire regel die we zo ontdekken is, dat we in *alle dimensies(1)* moeten beschermen. In dit geval zijn er twee dimensies: sectoren en regio's.
Als de indeling ook grootteklassen bevat, moet die als extra dimensie eveneens beschermd worden. Dat geldt natuurlijk voor alle andere indelingen die als extra dimensie worden gegeven.
Daarmee zijn we er echter nog niet. Want welke cel nemen we? Logisch zou zijn om de totalen te nemen, want daarmee wordt de primair geheime cel het effectiefst beschermd. Maar dan komt de tabel er zo uit te zien (P = primair, S = secundair):

*Tabel 2*

| *Regio* / *Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | 50 | S |
| *Handel* | 30 | 30 | 60 |
| *Dienstverlening* | 40 | 30 | 70 |
| *Overig* | 10 | 0 | 10 |
| *Totaal* | S | 110 | S |

We hebben het totaal van de totalen ook geheim moeten maken, want anders helpt het nog niet. Dit is een tweede secundaire regel die we zien: verhullen moet dus niet alleen in alle richtingen doorwerken, *maar ook in alle (hiërarchische) tussenniveaus van die richtingen (2).*
Op deze manier is er een goed beveiligde tabel ontstaan. Toch hebben we het idee dat we ons doel wat zijn voorbij geschoten en dat we liever iets anders zouden publiceren. We zien namelijk wel allerlei details, maar missen totalen. Juist datgene wat de klant het belangrijkst vindt ontbreekt. Daarom introduceren we een derde

secundaire regel: *het geheel is meer dan de som der delen (3)*. Dat houdt in dat we totalen, indien mogelijk, niet verhullen. We kiezen in plaats daarvan een ander detail op gelijk hiërarchisch niveau voor de bescherming van de primair geheime cel. Dan zou de tabel er bijvoorbeeld zo uit kunnen zien:

*Tabel 3*

| *Regio*<br>*Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | S | 80 |
| *Handel* | 30 | 30 | 60 |
| *Dienstverlening* | 40 | 30 | 70 |
| *Overig* | S | S | 10 |
| *Totaal* | 110 | 110 | 220 |

Tabel 3 komt intussen al aardig overeen met wat we zouden willen publiceren. We hebben de cel 'Zuid / Overig' ook geheim moeten maken, anders gaat het weer mis. Nu zijn alle totalen beschikbaar en slechts een paar andere cellen verhuld. Toch doet zich nog een probleem voor. We hebben namelijk nog geen rekening gehouden met de voorkennis die de gebruiker van de publicatie al heeft. Zo is bijvoorbeeld algemeen bekend, dat er in Nederland geen winning van ijzererts is en dat er alleen in het noorden aardgaswinning plaatsvindt. Stel dat de rij 'Overig' 'Aardgaswinning' heette, dan zou iedereen direct weten dat de 'S' bij Zuid in feite een 0 is en op die manier heel eenvoudig weer alles kunnen ontrafelen. Een vierde regel wordt daarom ook nog toegevoegd: *cellen met een inhoud van nul mogen nooit worden gebruikt om andere cellen te beschermen (4).*
Cellen met een inhoud van nul kunnen in publicaties daarom ook rustig worden gegeven. Ze worden nooit gebruikt voor geheimhouding. De tabel ziet er nu bijvoorbeeld als volgt uit:

*Tabel 4*

| *Regio*<br>*Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | S | 80 |
| *Handel* | S | S | 60 |
| *Dienstverlening* | 40 | 30 | 70 |
| *Overig* | 10 | 0 | 10 |
| *Totaal* | 110 | 110 | 220 |

In dit geval is gekozen voor de 'Handel' in plaats van de 'Dienstverlening'. Het verdient meestal op logische gronden de voorkeur, waar mogelijk, *opeenvolgende waarden (5)* te nemen (in de rij of in de kolom). Impliciet wordt het totaal van die (al dan niet opeenvolgende) waarden gepubliceerd, de informatie van tabel 4 is qua inhoud namelijk hetzelfde als die van tabel 5:

*Tabel 5*

| *Regio*<br>*Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie + Handel* | 60 | 80 | 140 |
| *Dienstverlening* | 40 | 30 | 70 |
| *Overig* | 10 | 0 | 10 |
| *Totaal* | 110 | 110 | 220 |

De reden om, als de overige regels het toestaan, de opvolgende waarde te nemen is gelegen in de logica van de gebruikte indelingen (of classificaties). Branche-indelingen volgen de productiekolom van winning tot eindgebruik, en de grootteklasse-indeling is numeriek. De regio-indeling volgt een lijn van buurprovincies. Aggregaten zijn dus altijd min of meer logisch. Deze regel is overigens betrekkelijk arbitrair en ondergeschikt aan regels die eisen zijn om de geheimhouding te waarborgen. Als we nog even kijken bij tabel 4 dan zien we dat in een tweedimensionale tabel een blokje ontstaat dat geheim wordt. Bij een driedimensionale tabel wordt dat automatisch een kubus. Er komt immers een dimensie bij.

In dit betoog wordt op dit punt geen onderscheid gemaakt tussen "nodig" of "wenselijk". De regel 'Het geheel is meer dan de som der delen' (regel 3), waardoor de totalen voorrang krijgen, is weliswaar voor de secundaire geheimhouding niet persé nodig, maar wel voor de klanten. Die willen immers in ieder geval over totalen kunnen beschikken. Daarom is dat ook 'nodig'. Wel gaan secundaire eisen natuurlijk boven 'wensen' van de buitenwereld, tenminste als ze elkaar dreigen te bijten. De bovengenoemde volgorde is dan ook niet de prioriteitsvolgorde, maar slechts die waarin de regels in dit betoog aan de orde komen. Voor de duidelijkheid zal verderop een indeling van de te onderscheiden regels naar "nodig" en "wenselijk" worden gegeven.

*Classificaties*

In tabel 6 is een deel van de classificatie (synoniemen: hiërarchie, dimensie of 'richting') van de bedrijven gegeven. Ter wille van de leesbaarheid is de tabel een kwart slag gedraaid, zodat het hoogste niveau links staat, in plaats van bovenaan. Het gaat in de tabel om de Nederlandse brancheclassificatie die gebaseerd is op de Europese NACE[6]. In de tabel is de volledige context van de cacaobonen-(verwerkende)-industrie weergegeven. Deze industrie, met de code 158410, is, zo blijkt uit de tabel, via een aantal tussenniveaus onderdeel van de industrie.

Het eerste niveau van de classificatie bevat het totaal van alle Nederlandse bedrijven. Op het tweede niveau zien we de belangrijkste onderverdeling daarvan gedeeltelijk terug (de codeletters daarvan lopen door tot Q!), te beginnen bij de Landbouw (A). Onze cacaobonen zijn een onderdeel van (respectievelijk) D (industrie), DA (voeding en genot), 15 (voeding), 158 (overige voeding), 1584 (cacaobonen + chocolade en suikerwerk) en 15841 (cacaobonen).

Net als deze classificatie is ook iedere andere strikt hiërarchisch. Er bestaat daarom niet een willekeurig (en, volgens de classificatie, onlogisch) aggregaat van de cacaobonen samen met de delfstoffenwinning als aparte cel. Alleen de niveaus van de hiërarchie hebben een eigen cel. Als aggregaat zitten daarbij alleen de door de classificatie bepaalde, logische, totalen.

Op het laagste, 8e, niveau in deze hiërarchie staan de cacaobonen alleen (158410). Op het niveau daarboven worden twee cellen met elkaar geconfronteerd: de cacaobonen (15841)en de fabricage van chocolade en suikerwerken (15842). Op niveau 6 zijn er zelfs negen codes op gelijk niveau. Op niveau 5 zijn er ook negen, op niveau 4 zijn er maar twee, op niveau 3 zijn er 14 en op niveau 2 zelfs 17. De hiërarchie is dus heel onevenwichtig: soms is er maar 1 cel, soms zijn er heel veel alternatieven.

Ook de bij de bedrijfsstatistieken gebruikte classificaties 'Grootteklasse' en 'Regio' kennen een vergelijkbare hiërarchische opbouw, hoewel het aantal niveaus en rijen verschilt.

*Tabel 6. Gedeeltelijke hiërarchie van de branches*[7]

| *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8e niveau* | |
|---|---|---|---|---|---|---|---|---|
| T | | | | | | | | Alle Nederlandse bedrijven |
| | A | | | | | | | Landbouw |
| | B | | | | | | | Visserij |
| | C | | | | | | | Delfstoffenwinning |
| | D | | | | | | | Industrie |
| | | DA | | | | | | Voeding en genot |
| | | | 15 | | | | | Voedingsmiddelen |
| | | | | 151 | | | | |
| | | | | 152 | | | | |
| | | | | 153 | | | | |
| | | | | 154 | | | | |
| | | | | 155 | | | | |
| | | | | 156 | | | | |
| | | | | 157 | | | | |
| | | | | 158 | | | | Overige voeding |
| | | | | | 1581 | | | |
| | | | | | 1582 | | | |
| | | | | | 1583 | | | |
| | | | | | 1584 | | | Cacaobonen, suikerwerk |
| | | | | | | 15841 | | Cacaobonen |
| | | | | | | | 158410 | Cacaobonen |
| | | | | | | 15842 | | Suikerwerk |
| | | | | | 1585 | | | |
| | | | | | 1586 | | | |
| | | | | | 1587 | | | |
| | | | | | 1588 | | | |
| | | | | | 1589 | | | |
| | | | | 159 | | | | |
| | | | 16 | | | | | Tabaksverwerking |
| | | DB | | | | | | |
| | | *tot* | | | | | | |
| | | DN | | | | | | |
| | E | | | | | | | |
| | *tot* | | | | | | | |
| | Q | | | | | | | |
| *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8e niveau* | |

*Van theorie naar methode*

De tabellen 1 tot 5 maken aanschouwelijk waarom de secundaire regels nodig zijn. In de praktijk van alledag gaat het bij bedrijfsstatistieken echter om driedimensionale tabellen. Dat zijn tabellen waar drie dimensies waarin de dataverzameling uiteen kan vallen, worden onderscheiden. Die dimensies zijn

- de branche (NACE, een codestelsel met in de Nederlandse situatie 8 niveaus met in totaal ongeveer 2600 rijen),
- de grootteklasse (een codestelsel met 3 of 4 niveaus waarin de bedrijven naar omvang zijn gerangschikt, in totaal minder dan 20 rijen) en
- de regio (in de Nederlandse situatie met 5 niveaus, en, afhankelijk van het jaartal, 600 tot 700 rijen).

Het gaat dan om 25 miljoen cellen.
Het blijkt trouwens heel moeilijk zich een voorstelling te maken van een driedimensionale tabel met verschillende hiërarchische niveaus erin. Het wordt nog lastiger als je dat probeert te doen terwijl er ook nog acties in plaats vinden. Het eenvoudigste is dan ook je te beperken tot een logische redenering voor één dimensie. In het volgende wordt dan ook uitgegaan van één enkele dimensie, om het probleem begrijpelijk te houden.

*Basisgedachte*

De gedachte is als volgt. Een dimensie bestaat uit een aantal verschillende niveaus Die niveaus kunnen elkaar alleen op gelijk niveau rechtstreeks beïnvloeden (zie tabel 6), want we willen als het even kan het totaal wel blijven geven. Als je alle niveaus stap voor stap doorloopt, dan komen alle zijdelingse relaties vanzelf aan de orde. Dat betekent dat het niet nodig is alle niveaus tegelijk te bekijken. Je dus kunt volstaan met het niveau dat op dat moment aan de orde is, samen met het niveau daarboven, het totaal dus, te bekijken. Je maakt als het ware een deeltabel van ieder samenhangend niveau, met zijn eigen totaal. Onze cacaobonen (zie tabel 6) komen op die manier op niveau 8 in beeld als een deeltabel van 1 cel (code 158410) met een totaal (niveau 7, code 15841). Niet erg indrukwekkend dus, maar wel overzichtelijk. Als je alle losstaande deeltabellen van een niveau hebt gehad (in tabel 6 staan alleen de cacaobonen genoemd, maar er zijn er natuurlijk veel meer, net zoveel als er totalen zijn van het niveau daarboven!), ga je het volgende niveau bekijken. Het gaat dan om niveau 7. Daar krijgen de cacaobonen (15841) in hun deeltabel gezelschap van 'suikerwerk' (code 15842), met het bijbehorende totaal (code 1584). Dit gaat zolang door tot de deeltabellen van alle niveaus afgehandeld zijn. Iedere cel komt op die manier twee keer in een deeltabel voor, eerst als detail en daarna als totaal, behalve natuurlijk de hoogste en de laagste cellen. Op die manier communiceren de verschillende niveaus met elkaar. Via het gelijke niveau communiceren ze met cellen op gelijk niveau (in tabel 6 zijn dat de codes die onder elkaar staan), en via de top is op die manier de hele hiërarchie met elkaar verbonden. Als alle deeltabellen zijn afgehandeld, is de tabel (als geheel) nog niet veilig.

Omdat niveau voor niveau wordt afgewerkt en pas in hogere niveaus kan blijken dat er zijdelings (op gelijk niveau dus) iets geheim wordt dat ook op een niveau lager tot geheimhouding moet leiden, moet het proces iteratief worden. Als je van de top uitgaat, weet je nog niet wat je beneden tegenkomt, maar ga je van beneden uit, dan heb je de lagere niveaus juist al vastgesteld voordat je weet hebt van een probleem dat hogerop bekend wordt. Je kunt het daarom niet in een keer oplossen, maar moet dat stapsgewijs (niveau voor niveau, per deeltabel) en, als je alle niveaus hebt gehad, nogmaals, iteratief doen. Na de eerste iteratie door alle niveaus begin je dus bij de volgende iteratie weer van voren af aan. De tabel met al zijn niveaus moet zo vaak worden doorlopen tot alle veranderingen in alle hoeken hebben doorgewerkt en totdat er nergens meer iets is aan te passen.

Omdat de primaire geheimhouding wordt bepaald vanuit de laagste niveaus, gaan we uit van dat niveau. Iedere iteratie begint dus vanaf de bodem van de tabel, bottom-up. Geredeneerd vanuit tabel 6: het begint op niveau 8 van die classificatie en eindigt bij niveau 1.
De kennis van lagere niveaus kan helpen om op hoger niveau de juiste cel voor secundaire geheimhouding te kiezen. Op die manier kan het aantal te verhullen cellen zo veel mogelijk worden beperkt.

Bij de beoordeling van een deeltabel kunnen zich in principe drie situaties voordoen:

I Er wordt een cel op gelijk niveau geheim gemaakt om een broedercel te beschermen (of meer dan 1);
II Er wordt een totaal geheim gemaakt, omdat de details (broedercellen) onvoldoende mogelijkheden tot bescherming bieden (nulcellen bijvoorbeeld!) en
III Er is een geheim totaal, zonder dat (minstens) één van de details geheim is. Deze situatie ontstaat na I, wanneer een geheimgemaakt detail, een niveau lager!, het totaal geworden is. Dan moeten er, naar behoefte, details geheim worden.

Het toevoegen van een tweede, een derde of zelfs van volgende dimensies is niets anders dan het herhalen van hetzelfde procédé. Het principe wordt er niet ingewikkelder of gecompliceerder van. In de tabellen 1 tot en met 5 is dat voor twee dimensies gedeeltelijk uitgewerkt. Door extra dimensies komen er wel (voorrangs)regels bij.
*Verdere complicaties en regels*

Bij het toepassen van de tot nu toe gevonden 5 secundaire regels ontstaan nog een paar andere problemen. Dit geldt vooral voor meer ingewikkelde gevallen, wanneer er meer keus is, en toegepast op een meerdimensionale tabel. De volgende punten kunnen daarom worden beschouwd als even zovele extra regels:
6. In een tabel met meer dan 1 dimensie moet voor iedere richting een keus worden gemaakt. Vanzelfsprekend is de dimensie waarvoor het eerst de keus wordt gemaakt in het voordeel, want daar wordt de keus nog niet beperkt

of beïnvloed door andere gemaakte keuzen. *Afspraak is dat de eerste dimensie (de branche) voor de andere dimensies gaat. De tweede dimensie (grootteklasse) gaat weer voor derde (regio) (6).* De eerste dimensie is daarmee bevoordeeld op de tweede en die weer op de derde. Een voorbeeld:

Bij het beveiligen van tabel 7 kun je uitgaan van de verticale dimensie of de horizontale dimensie:

*Tabel 7*

| *Regio*<br>*Sector* | *Noord* | *Oost* | *Zuid* | *West* | *Totaal* |
|---|---|---|---|---|---|
| *Industrie* | P | 20 | 20 | 20 | 70 |
| *Handel* | 30 | 0 | 20 | 20 | 70 |
| *Diensten* | 30 | 0 | 30 | 30 | 90 |
| *Overig* | 30 | 80 | 30 | 30 | 170 |
| *Totaal* | 100 | 100 | 100 | 100 | 400 |

In tabel 8 staat het resultaat uitgaande van de horizontale dimensie:

*Tabel 8*

| *Regio*<br>*Sector* | *Noord* | *Oost* | *Zuid* | *West* | *Totaal* |
|---|---|---|---|---|---|
| *Industrie* | P | S | 20 | 20 | 70 |
| *Handel* | 30 | 0 | 20 | 20 | 70 |
| *Diensten* | 30 | 0 | 30 | 30 | 90 |
| *Overig* | S | S | 30 | 30 | 170 |
| *Totaal* | 100 | 100 | 100 | 100 | 400 |

In tabel 9 staat echter het resultaat uitgaande van de verticale dimensie:

*Tabel 9*

| *Regio*<br>*Sector* | *Noord* | *Oost* | *Zuid* | *West* | *Totaal* |
|---|---|---|---|---|---|
| *Industrie* | P | 20 | S | 20 | 70 |
| *Handel* | S | 0 | S | 20 | 70 |
| *Diensten* | 30 | 0 | 30 | 30 | 90 |
| *Overig* | 30 | 80 | 30 | 30 | 170 |
| *Totaal* | 100 | 100 | 100 | 100 | 400 |

Hieruit blijkt duidelijk dat de resultaten kunnen verschillen naargelang het probleem in een andere volgorde wordt opgelost.

7. In tabellen met veel niveaus is het handig de keus te laten beïnvloeden door de al aanwezige problemen op lagere niveaus. Daardoor wordt er minder verhuld. *De problemen moeten zoveel mogelijk worden geconcentreerd, door rekening te houden met problemen in lagere niveaus (7).*

8. Als ergens een cel geheim wordt gemaakt, moet mogelijk op het niveau daaronder ook nog een cel geheim worden. Het is noodzakelijk op dat lagere niveau te weten hoeveel er moet worden gecompenseerd om niet in een situatie terecht te komen dat er te veel of te weinig wordt verhuld. *Te weinig verhullen mag niet en te veel verhullen is ongewenst. Een zo goed mogelijke benadering van de te verhullen omvang is dus noodzakelijk (8).*

9. In gevallen waarbij cellen met nul waarnemingen zijn betrokken kan de situatie ontstaan dat *het probleem niet op gelijk niveau kan worden opgelost (9)* (zie tabel 10, 11 en 12). In zo'n geval kan overigens slechts bottom-up (iteratief) worden gewerkt:

*Tabel 10*

| *Regio*<br>*Sector* | Noord | Zuid | Totaal |
|---|---|---|---|
| *Industrie* | P | 110 | 120 |
| *Overig* | 100 | 0 | 100 |
| *Totaal* | 110 | 110 | 220 |

Om een of andere reden is in tabel 10 de cel 'Noord / Industrie' geheim. Gegeven secundaire regel 4 (cellen met nul niet gebruiken voor verhulling), is het niet voldoende om de details geheim te maken, want de geheime cellen zijn dan eenvoudig af te leiden:

*Tabel 11*

| *Regio* / *Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | S | 120 |
| *Overig* | S | 0 | 100 |
| *Totaal* | 110 | 110 | 220 |

Daarom is het nodig om ook de totalen te verhullen en daarmee een niveau hoger te gaan:

*Tabel 12*

| *Regio* / *Sector* | *Noord* | *Zuid* | *Totaal* |
|---|---|---|---|
| *Industrie* | P | S | S |
| *Overig* | S | 0 | S |
| *Totaal* | S | S | 220 |

10. Elkaar compenserende cellen moeten *zo weinig mogelijk van 'gewicht' verschillen (10)*, omdat anders het risico van te veel of te weinig geheimhouden groot wordt.
11. Als meer onderdelen van een bedrijf (bijvoorbeeld regionale vestigingen) in 1 cel zijn opgenomen, moet het totaal van de onderdelen worden afgezet tegen het celtotaal om te bepalen of de cel primair geheim is. Het gaat dan immers niet om ieder individueel onderdeel. Omhooggaand in de dimensie moeten voor de (nieuwe) totaalcel opnieuw alle onderdelen gezamenlijk worden afgezet tegen het nieuwe celtotaal. Dat geldt natuurlijk voor alle niveaus van een dimensie. Bij het bepalen van de geheimhouding in een deeltabel waar sprake is van dit verschijnsel ('verdeelde bedrijven' of vestigingen), is, om programmatechnische redenen, niet altijd het celtotaal beschikbaar van willekeurige aggregaten die kandidaat geheimhoudingscel zijn. Om tot een 100% veilige uitkomst te komen, wordt in voorkomende probleemgevallen overdreven, waardoor *het probleem bij meerdere vestigingen net als bij 0-cellen een niveau kan stijgen (11)* (zie regel 9).

*Samenvatting van de regels*

In het voorgaande hebben we aan de hand van voorbeelden laten zien dat er regels nodig zijn, en welke dat zijn. Een aantal daarvan is absoluut nodig voor de beveiliging zelf, terwijl de overige alleen nodig zijn om zoveel mogelijk nuttige, en door de klant gevraagde, informatie te kunnen leveren. De eerste categorie kunnen we de harde beveiligingsregels, en de tweede de optimalisatieregels noemen. Beide soorten regels kunnen we beschouwen vanuit het perspectief van de tabel (A) en vanuit het perspectief van de verzameling van tabellen (B).
De tussen haakjes geplaatste nummering verwijst naar de regels zoals die eerder werden toegelicht. De Romeinse cijfers I, II en III corresponderen met de drie situaties die zich kunnen voordoen en zoals vermeld in het hoofdstukje "Basisgedachte".

<u>De harde beveiligingsregels</u>
*A. Vanuit het perspectief van de tabel.*
1. Om een primair geheime cel te beveiligen heeft de secundaire beveiliging zoveel mogelijk plaats op hetzelfde niveau. Dit betekent het volgende:
   - Een cel (of meer dan 1) wordt op gelijk niveau geheim gemaakt om een broedercel te beschermen (I); behalve
   - In het geval dat de details (broedercellen) onvoldoende mogelijkheden tot bescherming bieden, bijvoorbeeld bij 0-cellen (9) of bij meerdere vestigingen (11) wordt een totaal geheim gemaakt (II)
   - In een tabel waar alleen het totaal geheim is, moet de secundaire geheimhouding op het onderliggend niveau gezocht worden (III).
2. De elkaar compenserende cellen moeten zoveel mogelijk van een gelijk gewicht zijn
   - Op hetzelfde niveau (10) of indien nodig
   - Op een onderliggend niveau (8) en
   - Mogen dus zeker niet de inhoud '0' hebben (4)

*B. Vanuit het perspectief van de verzameling van tabellen.*
1. Beveiliging moet plaatsvinden
   - In alle dimensies (1) en
   - Op alle tussenniveaus van die dimensies (2).

<u>De optimalisatieregels</u>
*A. Vanuit het perspectief van de tabel.*
1. Het geheel is meer dan de som der delen, geheimhouding moet zoveel mogelijk op details (3).
2. Compenserende cellen worden zoveel mogelijk gezocht op opeenvolgende rijen binnen 1 niveau (5) en

3. De volgorde van de beveiliging binnen de tabellen is: branche, grootteklasse, regio (6)

*B. Vanuit het perspectief van de verzameling van tabellen.*

1. Beveiligingsproblemen worden zoveel mogelijk geconcentreerd (7).

De harde beveiligingsregels zijn altijd geldig. De optimalisatieregels zijn mede afhankelijk van omstandigheden, die per situatie (bijvoorbeeld per land) kunnen verschillen. Hier kan in de toekomst uitbreiding of verfijning nodig zijn, zeker in het geval dat statistische bureaus uit andere landen het programma willen gebruiken.

*Slot*

De 'Geheimhoudingsmodule 2000' (versie 2) komt tegemoet aan de hoofdzaken van de geschetste problematiek. Dit is mede te danken aan de inzet van vele deskundigen bij verschillende onderdelen van het statistische bureau tijdens de implementatie. Daardoor zijn ook nog een aantal aanvullingen gerealiseerd ten opzichte van versie 1 van november 1999.

Uitgegaan wordt van ASCII-bestanden. De classificaties en de noodzakelijke informatie over de bedrijven worden op aparte bestanden aangeleverd. Ook kan geschiedenis waarmee rekening moet worden gehouden meegeleverd worden. Na het opstarten wordt door de module korte tijd later een ASCII-bestand opgeleverd dat het Microlab, als publicatiebasis van de bedrijfsstatistieken, voedt met de voor publicaties noodzakelijke geheimhoudingsinformatie. Iedere tabel uit het Microlab is op die manier helemaal te beveiligen, op alle gevraagde niveaus. Daarnaast kan ook veel meer detailinformatie gepubliceerd worden dan vroeger, omdat we regionale informatie niet meer bij voorbaat geheim verklaren, maar eveneens netjes hebben doorgerekend.

In versie 2 van het programma is nog geen rekening gehouden met door bedrijven eventueel afgegeven publicatiemachtigingen. Ook wordt de keuze voor secundaire geheimhouding nog niet in alle gevallen even optimaal gedaan. Met name de secundaire regels 7 en 8 verdienen nog betere implementatie. Er wordt intussen gewerkt aan versie 3, die naar verwachting in de loop van 2000 beschikbaar komt. Versie 3 zal zowel Engelstalig als Nederlandstalig zijn.

---

[1] Wim Koopman (email: wkpn@cbs.nl) en Ab Mulder (email: amlr@cbs.nl) zijn werkzaam bij de divisie Landbouw, Nijverheid en Milieu van het Centraal Bureau voor de Statistiek.

[2] Zie ook: Accessibility of business microdata, Andrea Groot and Cor A.W. Citteur, Netherlands Official Statistics, winter 1997. Alle onderzochte landen hebben hiermee meer of minder problemen.

[3] Eurostat is het statistische bureau van de Europese Unie.

[4] De Geheimhoudingsmodule 2000 is ontworpen in Delphi door Ab Mulder. Tevens is het grootste deel van de methode door hem ontwikkeld. Het programma kan worden uitgevoerd onder Windows 95 op een computer met een processor uit de Pentium 100 serie of beter, en een intern geheugen van 128 Mb of meer. Hoe sneller de computer, des te sneller de resultaten. Hoe meer geheugen, des te groter de tabel kan zijn.

[5] Het Microlab is een gegevensbestand met microdata van bedrijfsstatistieken. De betreffende publicaties worden in toenemende mate gegenereerd vanuit het Microlab. Op termijn zal dit alle bedrijfsstatistieken betreffen. Daarna is dan sprake van een Economisch Statistisch Basis Bestand. Wim Koopman is vanaf 1995 statistisch coördinator ten behoeve van de bedrijfsstatistieken.

[6] NACE: Nomenclature générale des Activités économiques dans les Communautés Européennes. Dit is de classificatie van economische activiteiten van de Europese Unie, waarop ook de Nederlandse classificatie is gebaseerd.

[7] Alleen de regels die voor het voorbeeld Cacaobonen relevant zijn, worden in de tabel getoond.