Geheimhoudingsmodule

Veel statistiekpublicaties zijn opgebouwd uit vertrouwelijke gegevens. In de Wet op de Economische Statistieken (1936) is de veiligheid van de gegevens van de individuele bedrijven expliciet gewaarborgd. Hoe gaat het Centraal Bureau voor de Statistiek nu om met vertrouwelijke gegevens?

De uitvoering van deze waarborg wordt door het CBS aan twee eenvoudige basisregels toevertrouwd. De eerste is dat gegevens niet gepubliceerd worden als ze afkomstig zijn van minder dan 'N' individuele bedrijven. Gegevens worden in de tweede plaats niet gepubliceerd als één individueel bedrijf voor meer dan 'X' procent verantwoordelijk is voor deze gegevens. Beide waarden, 'N' en 'X', worden om veiligheidsredenen 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*.

Secundaire geheimhouding

In het algemeen bestaat de behoefte om details én 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*. Dit onderscheid is noodzakelijk bij het proces tot vaststelling van de geheime gegevens. 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.

Terwijl de primaire geheimhouding 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 en omdat we te maken hebben met publicaties vanuit verschillende statistieken, die bovendien ook nog allerlei onderlinge relaties hebben. Het probleem is dan complexer dan het op het eerste gezicht lijkt. Bovendien wordt uit een dataset vaak op verschillende tijden en op diverse hiërarchische niveaus gepubliceerd.

Tot voor kort was er geen tijd of capaciteit om alle geheimhouding goed door te rekenen, bovendien was het probleem er te ingewikkeld voor. Veel regionale gegevens werden daarom vaak zekerheidshalve maar geheim verklaard en niet eens gepubliceerd.

Beveiliging

In augustus 1999 heeft het CBS de *Geheimhoudingsmodule 2000* ontwikkeld, dat een driedimensionale tabel van vele niveaus kan beveiligen. De drie dimensies kunnen volkomen flexibel, volgens een hiërarchische classificatie worden ingelezen. Hierna worden aan de hand van eenvoudige tweedimensionale voorbeelden een aantal noodzakelijke aanvullende regels vastgesteld, de *secundaire regels*. De uitgangssituatie is tabel 1 met willekeurig gekozen getallen, waarbij P staat voor primair geheim.

Zowel via de kolom als via de rij is hier het cijfer van de industrie uit de regio Noord rechtstreeks af te leiden, 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.

Maar: welke cel nemen we? Logisch zou zijn om de totalen te nemen, want daarmee wordt de primair geheime cel het meest effectief beschermd. Zo ontstaat tabel 2 (P = primair, S = secundair).

Informatieverlies

We hebben het totaal van de totalen ook geheim moeten maken, want anders helpt het nog niet. Dit is een tweede secundaire regel: verhullen moet 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 dat wat de klant het belangrijkste vindt, ontbreekt. Er treedt dus informatieverlies op. 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. (tabel 3) Deze tabel komt 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 zijn slechts een paar andere cellen verhuld.

Voorkennis

Toch doet zich nog een probleem voor. We hebben namelijk nog geen rekening gehouden met de voorkennis die de gebruiker van de publicatie bezit. Zo is algemeen bekend dat er in Nederland geen ijzererts wordt gewonnen en dat er alleen in het noorden aardgaswinning plaatsvindt. Stel dat de rij 'Overig' nu 'Aardgaswinning' heette, dan zou iedereen direct weten dat de 'S' bij Zuid in feite een 0 is. Op die manier zou iemand alles heel eenvoudig kunnen ontrafelen. Daarom wordt ook een vierde regel 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. (tabel 4)

In dit geval is gekozen voor de 'Handel' in plaats van de 'Dienstverlening'. (tabel 5) 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 van tabel 5, behalve voor het totaal van Industrie en Handel.

Logica

De presentatie van tabel 4 verdient de voorkeur, omdat daarbij meer informatie wordt gegeven dan in tabel 5. De reden om 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. De grootteklasse-indeling is numeriek. De regio-indeling volgt een lijn van buurprovincies. Aggregaten, mits opeenvolgend, zijn dus altijd min of meer logisch. Deze regel is overigens arbitrair en ondergeschikt aan regels die *eisen* zijn om de geheimhouding te waarborgen. Om de gehanteerde methode gemakkelijk te kunnen illustreren is uitgegaan van een tweedimensionale tabel. In werkelijkheid moet echter een stelsel van tabellen worden beveiligd met drie dimensies: branche, grootteklasse en regio.

Het toevoegen van meer dimensies brengt met zich mee dat de hiervoor beschreven procedure moet worden herhaald. Het wordt er in principe niet anders van. Er komen door de extra dimensies wel (voorrangs)regels bij, die voor het begrip van de beschreven methodiek niet essentieel zijn, en dus in het kader van deze beknopte uitleg ook niet verder zijn genoemd. ◀

Wim Koopman en Ab Mulder

W. Koopman en A. Mulder, 'Geheimhouding bij bedrijfsstatistieken van het CBS: Geheimhoudingsmodule 2000', *Industriemonitor*, juli 2000.

Tabel 1

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

Tabel 2

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

Tabel 3

	Regio		
	Noord	Zuid	Totaal
Sector			
Industrie	P	S	80
Handel	30	30	60
Dienstverlening	40	30	70
Overig	S	S	10
Totaal	110	110	220

Tabel 4

	Regio		
	Noord	Zuid	Totaal
Sector			
Industrie	P	S	80
Handel	S	S	60
Dienstverlening	40	30	70
Overig	10	0	10
Totaal	110	110	220

Tabel 5

	Regio		
	Noord	Zuid	Totaal
Sector			
Industrie + Handel	60	80	140
Dienstverlening	40	30	70
Overig	10	0	10
Totaal	110	110	220