Centraal Bureau
voor de Statistiek

Statistische Methoden

Afronden in tabellen

2013 | 03

Leon Willenborg
Willem Sluis

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving en leeswijzer

Het afronden van getallen is een welbekend probleem. Stel men moet in de winkel contant een bedrag betalen van 1,97 € en de kleinste munt die men kan gebruiken is 5-cent. Dan betaalt men 1,95 €, door afronding naar beneden. Indien men een aantal artikelen wil kopen, zeg van de volgende bedragen: 0,78 €, 1,99 €, 3,63 € en 5,23 €, dan wordt eerst de optelling gemaakt (11,63 €) en afgerond: 11,65 €. Indien men ieder van de bedragen eerst zou afronden en daarna optellen, betaalt men contant 11,70 €. Contante betaling is een praktische reden waarom afronden in de praktijk wordt toegepast. Een andere reden om getallen af te ronden is bij hoofdrekenen wanneer men wil nagaan wat de orde van grootte is van een optelling. Ook computers ronden af wanneer met floating point getallen wordt gerekend. Dit zijn slechts enkele voorbeelden die het belang van afronden illustreren. Zie de Wikipedia-bijdrage over 'Rounding', die diverse, interessante toepassingen bespreekt, maar niet specifiek in gaat op afronding in tabellen die in dit stuk aan bod komt (http://en.Wikipedia.org/wiki/Rounding). Afronden in de officiële statistiek (en dus bij het CBS) wordt gebruikt om tabellen te vereenvoudigen (verstrekken van minder details) of om ze te beveiligen (door toevoeging van ruis).

Bij afronden in tabellen zou men in eerste instantie misschien kunnen denken dat dit ook eenvoudig is: rond iedere cel waarde onafhankelijk van de anderen af, net als bij het afronden van getallen. Dat gaat echter niet altijd goed. Problemen ontstaan wanneer men met kruistabellen werkt, waar ook marginalen bij zijn gegeven. Voor deze tabellen geldt dat de cel waarden in het binnenwerk in eenzelfde rij, kolom e.d. gelegen optellen tot het corresponderende randtotaal. Dit geldt ook voor marginalen van marginalen. We noemen de tabellen dan additief. Bij afronden eist men doorgaans ook dat de afgeronde tabellen (binnenwerk en randen) additief (optelbaar) zijn. Bovendien wil men dat de afgeronde cel waarden niet te ver afliggen van de bijbehorende originele cel waarden.

Om de gedachten te bepalen is in Tabel 1 een situatie weergegeven van een 2d tabel met twee 1d randen en een totaal generaal.

Tabel 1. 2d-tabel met randen

3	17	3	8	31
2	4	5	1	12
7	2	11	3	23
12	23	19	12	66

Stel dat de bedoeling is om de cel waarden af te ronden op 5-vouden, niet al te ver verwijderd van de bijbehorende cel waarden, en zodanig dat de afgeronde 2d tabel optelt tot de afgeronde 1d tabellen en deze op hun beurt weer optellen tot het

afgeronde totaal-generaal. In Paragraaf 2.4.1 wordt dit voorbeeld verder besproken. In Paragraaf 2.4 staan meer voorbeelden van tabelafrondproblemen.

In de boven gegeven voorbeelden wordt een getal afgerond op het dichtst bijliggende veelvoud van een bepaald getal (in de voorbeelden 5 (eurocent)). Bij het afronden van tabellen kijkt men vaak naar de twee dichtst bijliggende veelvouden van zo'n getal, afrondbasis genaamd en men heeft de keuze om naar boven of naar beneden af te ronden. Bijvoorbeeld kan men 17 dan bij een afrondbasis van 5 afronden naar 15 of naar 20. De afrondfout is dan 2 (bij afronden naar beneden) of 3 (bij afronden naar boven). Voor sommige algoritmen gebeurt de afronding stochastisch. Met een bepaalde kans wordt naar beneden of naar boven afgerond. De kans voor afronding naar beneden of naar boven wordt dan doorgaans bepaald door de afrondfout. (Zie Paragraaf 2.3.7 voor enkele voorbeelden van dergelijke methoden).

Afronden in tabellen zonder bijbehorende marginalen is eenvoudig als men de cellen afzonderlijk kan afronden, zonder verdere beperkingen. Het wordt lastiger als men meerdere tabellen heeft met gemeenschappelijke randen, bijvoorbeeld één tabel met zijn marginalen, en men wil al deze tabellen afronden zodanig dat de afgeronde tabellen ook additief zijn en bovendien niet al te veel afwijken van de originele tabellen. Of men wil, als een probabilistische afrondmethode wordt gebruikt, dat de afronding (conditioneel) zuiver is, dus dat de afgeronde tabel (opgevat als stochast) (conditioneel) zuiver is, en dus in verwachting de oorspronkelijke tabel oplevert. Dat legt beperkingen op aan de afrondkansen. (Zie Paragraaf 2.3.7)

In sommige gevallen vindt men geen oplossing voor een afrondprobleem als men iedere cel waarde slechts kan afronden naar één van de twee de dichtst bijliggende veelvouden van de gekozen afrondbasis. Men dient dan in een ruimere omgeving van cel waarden naar een geschikt veelvoud van de afrondbasis te kunnen zoeken. Uiteraard geldt dan wel: hoe verder de oplossing (afronding) verwijderd is van de oorspronkelijke cel waarde, des te groter de penalty.

De afrondmethoden die in dit stuk worden gepresenteerd leiden tot zekere optimaliseringsproblemen, die gewoonlijk alleen met een computer kunnen worden opgelost. Dat geldt eigenlijk voor alle afrondmethoden die bekend zijn in de literatuur. Deze afrondproblemen worden in de praktijk opgelost met behulp van speciale software bedoeld om lineaire optimaliseringsproblemen, of zelfs (gemengde) geheeltallige optimaliseringsproblemen, op te lossen. Voor de in dit stuk beschreven afrondmethode is software beschikbaar. (Zie Paragraaf 2.3.8). Dat geldt ook voor enkele andere methoden die in dit stuk worden genoemd maar niet verder uitgediept (zoals controlled tabular adjustment (CTA); zie Paragraaf 2.3.6).

1.1.1 Beschrijving van het thema

In dit stuk behandelen we afronden van tabellen, uitgaande van een afrondtechniek die rekening houdt met bepaalde afhankelijkheden tussen de diverse cel waarden in de tabellen die men wil afronden. Met het afronden van tabellen wordt bedoeld het

afronden van cel waarden in tabellen. Wat het onderwerp compliceert is dat aan de afgeronde tabellen voorwaarden worden gesteld. Daar gaan we nu nader op in.

Het doel van het afronden in tabellen is afgeronde tabellen te krijgen waarbij iedere cel waarde een veelvoud is van een van te voren gekozen afrondbasis β is, waarbij $\beta > 0$ doorgaans een geheel getal is. Maar dat is niet het enige. Men wil ook dat de afgeronde cel waarden niet te ver af liggen van de oorspronkelijke cel waarden. Men streeft ernaar om iedere cel waarde af te ronden naar één van de twee veelvouden van een gekozen afrondbasis β die het dichtst bij liggen. Alleen lukt dat niet in alle gevallen.

Een verfijning van het afrondprobleem krijgt men door te kijken naar de relatieve afrondfout in plaats van naar de absolute afrondfout. Bij deze laatste fout kijkt men alleen naar de afstand tussen een getal en zijn afronding. Bij de relatieve afrondfout deelt men de absolute fout door het getal zelf (als dit 0 is, hoeft men niet af te ronden en is er dus ook geen sprake van een afrondfout, absoluut of relatief). Zie Paragraaf 2.3.4 voor nadere informatie over relatief afronden.

1.1.2 Problemen en oplossingen

Men zou afronden van tabellen op de simpelste manier kunnen doen door dit per cel waarde te doen. Men hoeft dan geen rekening te houden met andere cellen en hun afrondingen. Hoewel simpel toe te passen, heeft deze methode – ongecontroleerd afronden genaamd - nadelen, namelijk dat op deze manier verkregen afgeronde tabellen niet optelbaar hoeven te zijn. Niet alleen is dit storend, het kan zelfs ineffectief zijn, in de zin dat de afronding kan worden teruggedraaid. In het geval afronding wordt toegepast om tabellen te beveiligen, is dan de afronding een schijnoperatie die niet tot veiligere tabellen leidt. In Paragraaf 2.4.3 wordt een voorbeeld gegeven van een dergelijke situatie.

Afronden in tabellen waar lineaire restricties tussen cel waarden bestaan is lastig als men voor de afgeronde tabellen eveneens eist dat de cel waarden in deze tabellen aan soortgelijke restricties voldoen. Ook wil men dat de afgeronde cel waarden niet te ver af liggen van hun oorspronkelijke waarden. Bij voorkeur is de afgeronde waarde van een cel waarde a één van de twee dichtst bijliggende veelvouden van de afrondbasis: het kleinste veelvoud van β groter dan of gelijk aan a of het grootste veelvoud van β kleiner of gelijk aan dan a, waarbij β de afrondbasis is.

Hoewel dat vaak wel te bewerkstelligen is, is het niet te garanderen dat het altijd mogelijk is. Er zijn voorbeelden waarbij dit niet lukt. Paragraaf 2.3.3 gaat hier op in.

Een ander fenomeen kan zich voordoen bij gekoppelde tabellen. Het is mogelijk dat de oorspronkelijke tabellen set bestaat uit een n-dimensionale tabel ($n \geq 3$) met marginalen, zodanig dat als men alleen deze marginalen afrondt er geen moedertabel is die past bij deze afgeronde tabellen. Zie Paragraaf 2.4.6 voor een voorbeeld. Indien men de moedertabel bij het afrondproces betrekt kan dit fenomeen niet optreden. Zie Paragraaf 2.4.5.

1.2 Afbakening en relatie met andere thema's

Afronden van tabellen wordt op de volgende twee plaatsen in het statistisch proces gebruikt:

- Bij statistische beveiliging: hercoderen, cel onderdrukking, random ruis toevoegen
- Bij presentatie van tabellen: afkappen van cel waarden

De beschrijving die hier gegeven wordt van het probleem van gecontroleerd afronden betreft één specifieke oplossingsmethode. Andere oplossingsmethoden voor dit probleem zijn ook mogelijk. Enkele daarvan worden besproken in Paragraaf 2.3.7.

1.3 Plaats in het statistisch proces

Afronden wordt op de volgende plaatsen in het statistische proces gebruikt:

- Output: Statistische beveiliging
- Output: Representatie van tabellen

1.4 Definities

Begrip	Omschrijving
Absolute afrondfout	Als x een waarde is en $r(x)$ is een afgeronde waarde van x bij een afronding, dan is de absolute afrondfout $\mid x - r(x) \mid$, waarbij $\mid \cdot \mid$ de absolute waarde aan duidt.
Absoluut afronden	Afronden waarbij een penalty functie wordt gebruikt die bepaald wordt door absolute afrondfouten
Afrondbasis	Een positief getal (doorgaans een geheel getal), in dit stuk aangeduid met β . Iedere cel waarde in de afgeronde tabel(len) is een veelvoud van β .
Afronden zonder restricties	Afronden per cel waarde, zonder te letten op constraints die gelden in verband met optelbaarheid. Kan daarom leiden tot afgeronde tabellen waarbij het afgeronde binnenwerk niet optelt tot de afgeronde randen. Aangeduid met 'unrestricted rounding' in Willenborg & De Waal (2001).
Binnenwerk (van een tabel met marginalen)	De cel waarden van een tabel waarbij ook (sommige van) de randen gegeven zijn. Zo'n tabel met zijn randen wordt als één geheel beschouwd. Zie ook: moedertabel.
Ceiling (bij afrondbasis β)	Zie de functie $Ce_{\beta}(x)$ in de tabel in Paragraaf 1.5.
Consistente set tabellen	Een set tabellen waarbij gemeenschappelijke randen hetzelfde zijn.
Deterministisch afronden	Afrondtechniek waarbij geen kans mechanisme wordt gebruikt om cel waarden af te ronden. 'Deterministic rounding' in Willenborg & De Waal (2001).
Floor (bij afrondbasis β)	Zie de functie $Fl_{\beta}(x)$ in de tabel in Paragraaf 1.5.
Gecontroleerd afronden	Afronden in tabellen waarbij er expliciet op wordt gelet dat de afgeronde tabellen optelbaar zijn. 'Controlled rounding' in Willenborg & De Waal (2001).
Gekoppelde tabellen	Een set tabellen die gemeenschappelijke randen hebben. 'Linked tables' in het Engels.
Hiërarchische tabel	Een bepaalde type gekoppelde tabellen, waarbij een

	hiërarchische structuur bestaat die aangeeft dat een tabel ('parent table') kan worden verkregen door aggregatie van een andere tabel ('child table'). Deze hiërarchie hoeft geen boom te zijn.
Intel variabele	Zie: Tabel.
Lattice	Zie: Rooster.
Marginaal (van een tabel)	Een tabel die ontstaat uit een andere tabel (die het binnenwerk vormt) door optelling (aggregatie) van cel waarden in dezelfde rijen, kolommen, etc.
Moedertabel (bij een gegeven set tabellen)	Een tabel van waaruit de tabellen in de gegeven set kunnen worden afgeleid door aggregatie, dat wil zeggen optelling / sommatie van cel waarden in dezelfde rij, kolom, e.d. Zo'n moedertabel hoeft niet altijd te bestaan, ook al zijn de tabellen in de tabellen set onderling consistent.
Moedertabel (bij een gegeven set tabellen)	Tabel waarvan de gegeven tabellen marginalen zijn. Anders gezegd, de moedertabel is de inwendige tabel en de gegeven tabellen zijn randen hiervan. De gegeven tabellen moeten aan bepaalde restricties voldoen, i.c. dat af te leiden gemeenschappelijke randen hetzelfde zijn. Dat is echter geen garantie voor het bestaan van een moedertabel. Zie bijvoorbeeld Paragraaf 2.4.6 voor een tegenvoorbeeld.
Ongecontroleerd afronden	Afrondmethode in tabellen waarbij men cel waarden onafhankelijk van elkaar afrondt. Zie ook: gecontroleerd afronden.
Opspanvariabelen (van een tabel)	Zie: Tabel.
Rand (van een tabel)	Zie: marginaal (van een tabel).
Relatief afronden	Afronden waarbij een penalty functie wordt gebruikt die bepaald wordt door relatieve afrondfouten.
Relatieve afrondfout	Als x een waarde is en r(x) is een afgeronde waarde van x bij een afronding en $x \neq 0$ dan is de relatieve afrondfout $\frac{\mid x - r(x) \mid}{\mid x \mid}$, waarbij $\mid \cdot \mid$ de absolute waarde aan duidt.
Rooster	Een graph waarmee kan worden aangegeven welke tabellen door aggregatie uit welke andere tabellen af te leiden zijn. Eng: Lattice.
Stochastisch afronden	Een afrondmethode waarbij men cel waarden afrondt gebruik makend van een kans mechanisme. 'Stochastic rounding' in Willenborg & De Waal (2001).
Tabel	Een functie $T : D \rightarrow B$ waarbij D een eindige verzameling is, te schrijven als cartesisch product $D = D_1 \times ... \times D_k$ voor zekere $k \in \mathrm{N}$ en waarbij B het bereik, is meestal één der verzamelingen N der natuurlijks getallen, Z der gehele getallen, Q der rationale getallen of $\Re$ der reële getallen. De variabelen geassocieerd met de domeinen $D_1,..., D_k$ worden wel opspanvariabelen genoemd van de desbetreffende tabel. De variabele geassocieerd met de verzameling B noemt men wel de Intel variabele van de desbetreffende tabel. Dit hier gegeven beschrijving slaat op een volledige tabel, zonder ontbrekende waarden. Tabellen met ontbrekende waarden kan men definiëren als een partiële functie, een functie die slechts op een deel van zijn domein is gedefinieerd.

1.5 Algemene notatie

Notatie	Omschrijving
$Fl_{\beta}(x)$	Het grootste veelvoud van β kleiner dan of gelijk aan x . 'Fl' staat voor het Engelse woord 'Floor'.
$Ce_{\beta}(x)$	Het kleinste veelvoud van β groter dan of gelijk aan x . 'Ce' staat voor het Engelse woord 'Ceiling'.
CTA	Controlled Tabular Adjustment (zie Paragraaf 2.3.7).
MIP	Mixed Integer Programming.

Typische afrondsituaties van tabellen betreffen sets gekoppelde tabellen. Deze tabellen kunnen aan elkaar gerelateerd zijn doordat ze gemeenschappelijke marginalen hebben. Deze kunnen expliciet of impliciet bestaan. Als ze expliciet bestaan is zo'n marginaal één van de tabellen in de set. In het impliciete geval is zo'n gemeenschappelijke marginaal te berekenen.

Om de afhankelijkheden tussen tabellen aan te geven gebruiken we roosters (lattices). Dit zijn graphen, waarbij de punten tabellen voorstellen en de kanten relaties tussen de tabellen. Iedere tabel wordt gerepresenteerd door een verzameling van variabelen die de tabel opspannen (die in het rooster als label bij de punt getoond wordt) Twee punten zijn door een kant verbonden als de tabel geassocieerd met het hoger gelegen punt kan worden verkregen door de tabel geassocieerd met het lager gelegen punt te aggregeren naar de (unieke) variabele die wel in de tabel zit geassocieerd met het lager gelegen punt maar niet in de verzameling geassocieerd met het hoger gelegen punt. Tabellen zijn vaak op meerdere manieren uit gedetailleerdere tabellen te verkrijgen door aggregatie. Alle tabellen in een rooster worden verondersteld eenzelfde intel variabele te hebben.

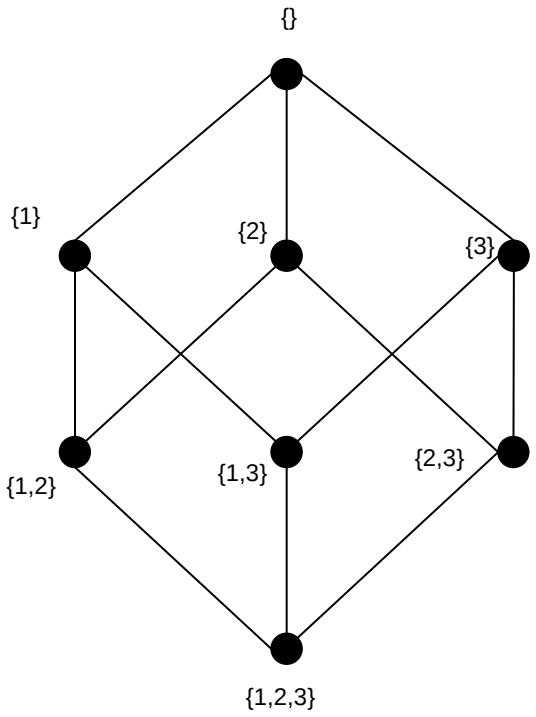
Als voorbeeld nemen we een 3d-tabel opgespannen door de variabelen 1, 2, 3. Deze heeft drie 2d-marginalen (of randen), namelijk de tabellen opgespannen door variabelen 1, 2, door variabelen 1, 3 en door variabelen 2, 3. En de 1d-marginalen opgespannen door variabele 1, door variabele 2 en door variabele 3. Tenslotte is er nog het totaal generaal. In Figuur 1 zijn deze tabellen weergeven als punten, corresponderend met verzamelingen van opspannende variabelen. Twee punten zijn met een kant verbonden als één van de verzamelingen een maximale echte deelverzameling is van een ander. Alleen voor verzamelingen A, B met $A \subset B$ is dit het geval, zó dat er geen verzameling C is met $A \subset C \subset B$. Hier duidt $\subset$ een echte deelverzameling aan (dus $\subseteq$ maar niet $=$). Hoe groter de verzamelingen zijn, des te lager in de figuur zijn ze getekend. Verzamelingen van gelijke omvang staan op hetzelfde horizontale niveau. Een figuur als weergegeven in Figuur 1 heet een rooster (Eng: lattice).

1.6 Representatie van tabellen

Er zijn verschillende manieren om tabellen te representeren. In de voorbeelden in Paragraaf 2.4 worden verschillende methoden gebruikt. De meest gebruikelijke methode voor 2d-tabellen is door middel van een matrix-vorm, waarbij de cellen een

bepaalde plaats hebben in deze matrix. Tabel 36 is een voorbeeld van een dergelijke tabelrepresentatie. Als een 2d-tabel marginalen heeft worden die vaak samen met de moedertabel afgebeeld. Tabel 4 is daar een voorbeeld van. Deze wijze van representatie is soms handig maar niet altijd. Bijvoorbeeld niet voor hoger dimensionale tabellen, met een dimensie hoger dan 2. In dat geval kan het handiger zijn om te werken met een lijstrepresentatie. Zo'n lijst bestaat uit cel coördinaten samen met de bijbehorende cel waarde. Indien men de conventie gebruikt dat alleen niet-nul cellen in de lijst hoeven te worden genomen, heeft men een representatie die efficiënt is voor grote tabellen met veel nul cellen, ook wel ijle tabellen (Engels: sparse tables) genoemd. Een voorbeeld van zo'n lijstpresentatie is niet te vinden in het onderhavige stuk. In de numerieke wiskunde wordt een dergelijke tabel (matrix) veel gebruikt. Zie ook http://en.wikipedia.org/wiki/Sparse_matrix. In het voorbeeld in Paragraaf 2.4.4 wordt een 3d-tabel met marginalen laag voor laag gerepresenteerd.

Figuur 1. Rooster van 3d-tabel met al zijn marginalen



Deze voorbeelden illustreren een diversiteit aan representatiemethoden voor tabellen die echter niet uitputtend is.

Het is, tot slot, zaak te onderscheiden tussen tabelrepresentaties voor 'menselijk gebruik' en voor computerverwerking.

Voor de eerste is een positiesysteem vaak het meest inzichtelijk, althans als de tabellen 2-dimensionaal zijn. De cel waarden worden dan in een tableau gepresenteerd waarbij de cel waarde van cel (i, j) in de i-de rij en de j-de kolom staat vermeld. Voor 3d tabellen worden dan lagen 2d tabellen getoond. Ook voor hiërarchische tabellen zijn overzichtelijke representaties bekend. Zie hiervoor recente CBS publicaties.

Voor de computerbewerking is de lijstrepresentatie juist wel handig, waarbij een matrix is weergeven als een geordende rij van paren <coördinaten, cel waarde>, waarbij een paar alleen is opgeslagen als de cel waarde ongelijk nul is. Dit levert een efficiënte wijze van opslag op voor ijle tabellen, dat wil zeggen tabellen met veel nullen. In de numerieke wiskunde wordt deze wijze van opslag veel gebruikt. Merk op dat deze representatie dezelfde vorm heeft ongeacht de dimensie van de tabel. Het enige verschil is dat de coördinaten verschillend zijn: voor n dimensionale tabellen zijn het n-tuples, ofwel geordende rijtjes van n indices.

In paragraaf 2.3 wordt een andere vorm gekozen om een tabel op te slaan die handig is voor computerverwerking. Hierbij wordt de tabel gevectoriseerd. Dit betekent dat de cellen in een tabel op een bepaalde manier geordend worden in een rij, en dat die rij alleen de cel waarden bevat. Dit betreft ook cel waarden die nul zijn. Uit het volgnummer in de rij kan men de coördinaten in een tabel berekenen, en omgekeerd. We geven een voorbeeld om dit te illustreren.

Voorbeeld. Gegeven is Tabel 2.

Tabel 2. Een voorbeeldtabelletje

43	15
0	21

In vectorvorm zou deze tabel weergegeven kunnen worden als

$$(43,0,15,21)\,. \qquad (1)$$

In Tabel 3 is de koppeling tussen cel coördinaten en vector coördinaten weergegeven.

Tabel 3. Koppeling van tabel- en vector coördinaten

Coördinaat tabelcel	**Coördinaat vector**
(1,1)	1
(2,1)	2
(1,2)	3
(2,2)	4

Tabel 3 kan gebruikt worden om coördinaten om te rekenen van de ene naar de andere representatie. ■

2. Probleemformulering en -oplossing

2.1 Korte beschrijving

Afronden van getallen is een eenvoudig probleem. Afronden in tabellen is lastiger omdat men hierbij rekening dient te houden met restricties.

In een typische afrondsituatie heeft men te maken met meerdere tabellen met gemeenschappelijke randen, dus aggregaten die hetzelfde zijn. Voor de afgeronde tabellen eist men doorgaans eenzelfde optelbaarheid als voor de oorspronkelijke tabellen. Eist men niet dat de afgeronde tabel optelbaar is, door bijvoorbeeld de afronding per cel te bekijken, dan is men soms in staat om de originele tabellen terug te rekenen. Een voorbeeld van deze situatie is te vinden in Paragraaf 2.4.

Ook bij stochastische afrondmethoden heeft men te maken met restricties bijvoorbeeld door de eis van de zuiverheid van de afrondmethode, zoals bijvoorbeeld bij de methode van Fellegi (1975) (zie ook Paragraaf 2.3.7).

In de praktijk doen zich verschillende situaties voor bij het afronden in tabellen. Een vaak voorkomend speciaal geval betreft een 2-dimensionale tabel met randtotalen en een totaal generaal. Maar men zou ook met een 3-dimensionale tabel te maken kunnen hebben met al zijn lager-dimensionale randen. Ook zijn er situaties dat men met meerdere tabellen te maken heeft waarvan er enkele een gemeenschappelijke marginaal hebben. Dit zijn zogenaamde gekoppelde tabellen. Voorbeelden van enkele in de praktijk voorkomende situaties worden gegeven in Paragraaf 2.4.

Al de genoemde situaties hebben gemeen dat er lineaire relaties bestaan tussen de celinhouden van sommige cellen in de set tabellen die men consistent wil afronden. Voor de corresponderende afgeronde cel waarden eist men dat aan dezelfde lineaire relaties wordt voldaan. Bij het afronden zoekt men waarden die een veelvoud zijn van een getal, de afrondbasis genoemd. Bij voorkeur zoekt men voor iedere cel waarde een afgeronde waarde die dicht ligt bij de af te ronden waarde. Bij voorkeur één van de twee: het corresponderende afgeronde getal naar boven (ceiling - plafond) of naar beneden (floor - vloer). Als β de afrondbasis is dan is $Fl_\beta(a)$ gelijk aan het grootste β-voud kleiner dan of gelijk aan a, en $Ce_\beta(a)$ is het kleinste veelvoud van β groter dan of gelijk aan a. Dus, bijvoorbeeld, $Fl_5(37) = 35$ en $Ce_5(37) = 40$.

2.2 Toepasbaarheid

Afronden voor datarepresentatie is interessant in die gevallen waarin het niet nodig is om detailinformatie te tonen, wanneer men schijnnauwkeurigheid wil vermijden of als de orde van grootte van getallen goed genoeg is om een betoog te ondersteunen. Het is daarbij gewenst dat de afgeronde tabellen optelbaar zijn, en bovendien, voor ieder van de cellen, niet te ver van de oorspronkelijke cel waarden verwijderd zijn. In dit geval mogen cel waarden positief, negatief of nul zijn.

Afronden gebruikt als beveiligingsmethode, moet van een onveilige tabel, of set gekoppelde tabellen, een veilige versie maken. In dit geval zijn de tabellen doorgaans niet-negatief, dat wil zeggen dat ieder van de cel waarden groter dan of gelijk aan nul is. Dit levert restricties op voor de cel waarden in de tabel die men kan gebruiken bij het bepalen van de waardenbereiken voor de originele cel waarden. De wijze van afronding dient gecontroleerd te zijn, om te voorkomen dat cel waarden terug te rekenen zijn, exact of met grote nauwkeurigheid. In feite houdt het afrondalgoritme rekening met het interval waarmee men gevoelige cellen kan terugrekenen. Zie ook Paragraaf 2.4.3.

2.3 Uitgebreide beschrijving

2.3.1 Het afrondprobleem

Het afrondprobleem kan in algemene vorm als volgt worden geformuleerd. Initieel gegeven zijn reële getallen $a = (a_1,..., a_n)$ die aan een aantal lineaire restricties voldoen $Ma \equiv b$ waarbij de elementen van M gehele getallen zijn en $b = (b_1,..., b_m)$ gehele veelvouden zijn van een afrondbasis β. Vind een vector y die aan dezelfde lineaire restricties $My \equiv b$ voldoet, die alleen veelvouden van β mag bevatten en die zo dicht mogelijk bij de initiële waarde a ligt. Het symbool $\equiv$ staat hier voor componentsgewijs toepassen van de symbolen $\leq, \geq$ en $=$.

Deze formulering is nog niet compleet zonder een uitleg over wat “dicht bij” precies betekent. Hier wordt deze notie expliciet gemaakt door de keuze van een optimalisatie expressie die geminimaliseerd dient te worden. Er zijn verschillende mogelijkheden voor de optimalisatie expressie, elk met voor- en nadelen.

Bovenstaand geval kan nog iets algemener geformuleerd worden door ook toe te staan dat de vector b uit gebroken veelvouden van de afrondbasis bestaat. Echter, door introductie van een extra inputvariabele is dit te reduceren tot bovenstaand geval.

2.3.2 Absoluut afronden ‘binnen de grenzen’

In navolging van het gebruikelijke afronden van een enkel reëel getal eisen we dat

$$y_i = Fl_\beta(a_i) \; of \; y_i = Ce_\beta(a_i) \tag{2}$$

Merk op dat we toestaan dat getallen als 4.12 afgerond worden naar 5.0 (met afrondbasis 1.0). Dit effect moeten we in het algemeen toestaan om oplossingen te vinden die aan de lineaire restricties voldoen.

Door introductie van beslisvariabelen $x_i = 0 \; of \; x_i = 1$ herformuleren we het probleem tot

$$y_i = Fl_\beta(a_i) + \beta x_i \tag{3}$$

$$x_i = 0 \quad als \; Fl_\beta(a_i) = Ce_\beta(a_i) \tag{4}$$

dan moet (4) gelden en dat

$$Mx \equiv \frac{1}{\beta}(b - M\ Fl_{\beta}(a)) \tag{5}$$

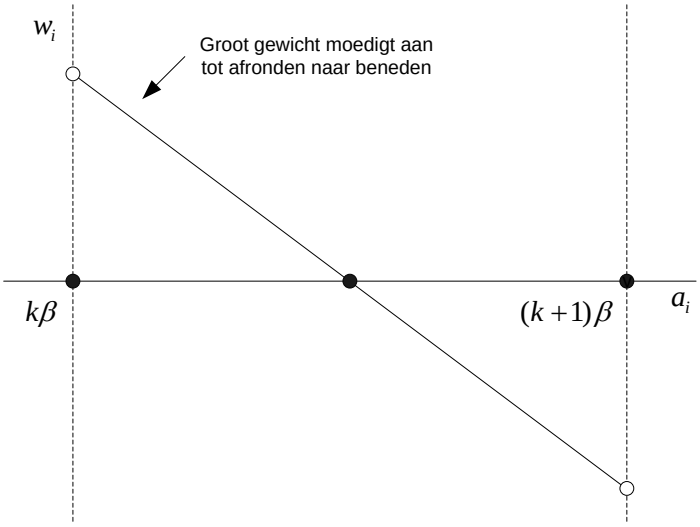
Dit vormt een stelsel lineaire restricties in de beslisvariabelen x. In het algemeen zal dit meerdere mogelijke waarden voor x toestaan. Door een extra criterium in te stellen kunnen we een voorkeur voor bepaalde x kenbaar maken. Voor de doelfunctie en de gewichten daarin maken we de volgende keuzes:

$$\min \sum_i w_i x_i \tag{6}$$

$$w_i = \frac{Fl_{\beta}(a_i) + Ce_{\beta}(a_i)}{2} - a_i \tag{7}$$

de gewichten w_i worden bepaald uit de initiële waarden en drukken uit dat getallen a_i die dicht bij een veelvoud van de afrondbasis liggen bij voorkeur afgerond worden tot dat veelvoud, zie Figuur 2. Merk op dat de gewichten negatief kunnen zijn. Zowel de keuze van de doelfunctie als van de gewichten, gegeven de keuze van een lineaire doelfunctie, kan ook anders uitvallen dan hierboven is gedaan.

Figuur 2. Gewichten voor de afrondmethode



Merk ook op dat als de initiële waarde al een veelvoud van de afrondbasis is, dan blijft deze in de oplossing ongewijzigd, vanwege (4). Merk verder op dat het gewicht $w_i = 0$ in het geval dat a_i precies midden tussen twee grenzen ligt. In dat geval is de penalty voor het omlaag afronden net zo groot als voor omhoog afronden (nl. 0).

Er zijn twee mogelijke problemen met deze aanpak.

Ten eerste is de oplossing mogelijk niet uniek. Het is denkbaar om in dat geval extra criteria te bedenken die een keuze forceert. Zo kun je bij gelijkwaardige oplossingen bijvoorbeeld ervoor kiezen om de randen zoveel mogelijk traditioneel af te ronden.

Daarnaast kun je de voorkeur geven om de kleinste variabelen zoveel mogelijk traditioneel af te ronden omdat daarvoor afrondingen relatief gezien de meeste impact hebben. In deze nota gaan we op deze criteria niet verder in[1], ook al omdat het niet duidelijk is hoe deze extra criteria in de context van een professionele optimizer geformuleerd kunnen worden. Daarnaast is het in veel gevallen acceptabel om uit de niet-unieke oplossingen een willekeurige keuze te maken.

Ten tweede is het mogelijk dat er geen oplossingen zijn. Dit betekent dat de beperking dat er ofwel omlaag ofwel omhoog afgerond moet worden te strikt is.

2.3.3 Absoluut afronden 'over de grenzen heen'

Het is ook mogelijk om toe te staan dat de afgeronde waarden over de grenzen van aanliggende gehele waarden gaan. Dus bijvoorbeeld dat 4.12 afgerond wordt naar 3.0 (uitgaande van een afrondbasis van 1.0). We hebben dan

$$
\begin{aligned}
& y_i = Fl_\beta(a_i) + \beta x_i + \beta x_i^+ - \beta x_i^- \\
& x_i = 0 \ of \ \ x_i = 1 \\
& x_i = 0 \ als \ \ Fl_\beta(a_i) = Ce_\beta(a_i) \\
& x_i^+ \geq 0, x_i^- \geq 0 \\
& \\
& x_i^+, x_i^- \text{ geheeltallig}
\end{aligned}
\tag{8}
$$

Dergelijk grensoverschrijdend gedrag wordt mogelijk gemaakt door strikt positieve waarden voor x_i^+ of x_i^-. Natuurlijk willen we dit zoveel mogelijk voorkomen, en dat komt dan ook tot uiting in de aangepaste optimalisatie-expressie.

$$
\min \sum_i w_i x_i + \frac{\beta}{2} x_i^+ + \frac{\beta}{2} x_i^- \tag{9}
$$

$$
w_i = \frac{Fl_\beta(a_i) + Ce_\beta(a_i)}{2} - a_i \tag{10}
$$

Samen vormt dit wederom een optimalisatieprobleem, dit keer met een mix van beslisvariabelen en integer variabelen.

Merk op dat een optimale oplossing bij probleem (9) altijd ofwel $x_i^+ = 0$ of $x_i^- = 0$ heeft. Merk verder op dat het algemenere probleem veel meer variabelen gebruikt dan het beperkte probleem. In de praktijk kan daarom gekozen worden om het uitgebreidere probleem pas op te lossen als het beperkte probleem geen oplossing heeft.

2.3.4 Relatief afronden

Bovenstaande optimalisatiefuncties kennen een absolute penalty toe in de zin dat het afronden van bijvoorbeeld 4.12 naar 4.0 even zwaar telt als het afronden van

[1] In sommige speciale gevallen zoals één dimensionaal afronden (totaal = som van termen) is dit goed te doen.

2004.12 naar 2004.0. Voor sommige doeleinden is het beter om te kiezen voor een relatieve penalty. De optimalisatiefunctie is

$$\min \sum_i \frac{w_i x_i + \frac{\beta}{2} x_i^+ + \frac{\beta}{2} x_i^-}{\varepsilon + |a_i|} \tag{11}$$

$$w_i = \frac{Fl_\beta(a_i) + Ce_\beta(a_i)}{2} - a_i \tag{12}$$

Hierbij is ε een kleine constante die voorkomt dat we in numerieke problemen komen in het geval van (hele) kleine waarden a_i. Dit laatste is niet onwaarschijnlijk in het geval een grote inputvector a.

2.3.5 Verantwoording

De afrondmethode die in deze paragraaf is beschreven vindt zijn oorsprong in Fischetti & Salazar-González (1998). In vervolgartikelen van Salazar-González is de methode verbeterd. Zie bijvoorbeeld Salazar-González (2005, 2006). De methode van Salazar-González maakt gebruikt van een mixed integer programming (MIP) formulering. De hier gepresenteerde methode is op het CBS geïmplementeerd en is beschikbaar voor toepassingen (zie Paragraaf 2.3.8).

2.3.6 Afronden en tabelbeveiliging

Zoals eerder opgemerkt wordt afronding in tabellen ook gebruikt om tabellen te beveiligen voordat ze gepubliceerd worden. Het is een speciale manier van ruis toevoegen aan tabellen, met het doel om gevoelige gegevens te verhullen. Het idee is om gevoelige cellen in tabellen te beschermen. Voor ieder gevoelige cel wordt een beveiligingsinterval gedefinieerd (of een lengte van een dergelijk interval). Dit is een interval waarbinnen de desbetreffende te beschermen cel waarde niet mag worden teruggerekend. Een geschikte keuze van een afrondbasis is gebaseerd op al deze intervallen (of hun lengtes). Zie verder Willenborg & De Waal (2001).

Een techniek die als een veelbelovende generalisatie van afronden kan worden gezien is Controlled Tabular Adjustment (CTA). Hierbij wordt aan de oorspronkelijke cel waarden ruis toegevoegd, zodanig dat een veilige set (synthetische) tabellen wordt gegenereerd, die bovendien additief zijn. Bij afronden voegt men ook ruis toe, maar men is veel meer beperkt in de keuze van die ruis voor iedere cel waarde. Het voordeel van CTA is bovendien dat het tot minder complexe optimaliseringsmodellen leidt dan in Paragraaf 2.3. Voor details zie Dandekar & Cox (2002), Castro (2006, 2011).

2.3.7 Alternatieve afrondmethoden

Naast de hierboven gepresenteerde afrondmethode van tabellen zijn in de literatuur nog diverse andere methoden bekend om tabellen af te ronden. Een aantal van deze methoden willen we hier de revue laten passeren. De bedoeling van deze paragraaf is om de geïnteresseerde lezer te wijzen op enkele andere methoden van afronden,

zonder echter uitvoerig op deze methoden in te gaan. Daarom zijn literatuurverwijzingen opgenomen voor ieder van de methoden die hier kort worden besproken. Daar zijn de details te vinden die in deze paragraaf ontbreken. Het overzicht in deze paragraaf is bedoeld om illustratief te zijn en niet limitatief.

De methode van Fellegi (1975) is probabilistisch. Dit betekent dat het afronden van de cel waarden gebeurt met behulp van een kans mechanisme. Iedere cel waarde van het binnenwerk van een tabel wordt naar boven of naar beneden afgerond met bepaalde kansen, die afhangen van de afstand van de oorspronkelijke cel waarde tot de boven- respectievelijk ondergrens (of eigenlijk, de afstand van het fractionele deel tot 0 respectievelijk 1 in een genormaliseerd afrondprobleem; zie hiervoor Paragraaf 2.4.7). De afrondprocedure is zodanig dat in verwachting de afgeronde tabel gelijk is aan de oorspronkelijke niet-afgeronde tabel. Ook behoudt de procedure de optelbaarheid. De methode van Fellegi is alleen te gebruiken voor 1d tabellen. Bovendien heeft deze methode tegenwoordig alleen historische waarde. Men zou deze methode kunnen beschouwen als een voorloper van de beneden te bespreken randomisatiemethoden, toegepast bij het specifieke probleem van gecontroleerd afronden in tabellen.

In Cox & Ernst (1982), Causey, Cox & Ernst (1985), Cox (1987), Cox & George (1989) wordt (onder andere) het gecontroleerd afronden in tabellen besproken, in het bijzonder voor 2- en 3-dimensionale tabellen met hun marginalen. Optimaliserings-modellen voor dergelijke problemen worden besproken evenals (benaderende) oplossingen. Dit werk is tegenwoordig vooral historisch van belang, aangezien het de aanzet vormt tot modernere probleemformuleringen voor gecontroleerd afronden in tabellen.

Een andere methode van afronden die gebruikt maakt van kans mechanismen om een oplossing voor een afrondprobleem te vinden is die van Kelly, Golden & Assad (1990, 1993), Kelly, Golden, Assad en Baker (1990) en is gebaseerd op een zogenaamde lokale zoekmethode, zoals simulated annealing of tabu search. Zie hiervoor Aarts en Lenstra (1997). Het gaat hierbij om heuristische methoden die gebruikt worden in de combinatorische optimalisering. Hierbij probeert men, beginnend vanuit een startpositie, uit te komen bij een voor de praktijk acceptabele oplossing, door van toestand naar toestand te springen. De oplossing die zo'n methode oplevert hoeft echter niet per se een optimum te zijn voor het probleem. Het zou wel een goede benadering moeten zijn.

Een andere methode voor het afronden van tabellen maakt gebruik van zogenaamde gerandomiseerde algoritmen, afkomstig van Raghavan & Thompson (1987); zie ook Motwani & Raghavan (1995). Dergelijke algoritmen zijn niet alleen geschikt om efficiënt benaderingen te vinden van combinatorische optimaliseringsproblemen (zoals gecontroleerd afronden) maar ook om afschattingen te geven over de kwaliteit van de gevonden oplossing, in de zin van de (verwachte) afwijking tot het optimum. Zie Doerr e.a. (2006) en Doerr and Klein (z.j).

2.3.8 Software

De afrondmethode beschreven in de Paragrafen 2.3.1 tot en met Paragraaf 2.3.4 is geïmplementeerd op het CBS (door Willem Sluis), die de software ook onderhoudt. In feite gaat het hier niet om een tool met een grafisch interface maar om een API in de vorm van een .Net assembly. Het eigenlijke optimalisatieprobleem wordt opgelost met behulp van het optimaliseringstool Xpress.

Ook het tabelbeveiligingspakket τ-ARGUS kan tabellen (gecontroleerd) afronden. Zie Hundepool e.a. (2011). τ-ARGUS kent wel een grafisch user interface.

2.4 Voorbeelden

2.4.1 Voorbeeld: 2d-tabel met marginalen

Een 2d-tabel met marginalen die moet worden afgerond komt vaak voor in de praktijk. Het voorbeeld dat we hier bekijken kan als typisch worden beschouwd, zij het dat de in de praktijk voorkomende tabellen vaak (veel) groter zijn, dat wil zeggen, meer cellen bevatten.

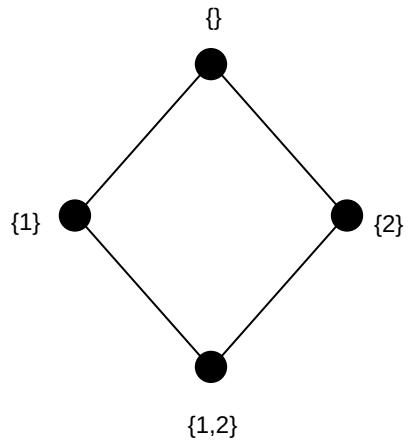
Beschouw Tabel 4 waarin een 2d-tabel te zien is met 2 1d-marginalen en een totaal generaal.

Tabel 4. 2d-tabel met randen

3	17	3	8	31
2	4	5	1	12
7	2	11	3	23
12	23	19	12	66

Figuur 3 geeft het bijbehorende rooster weer.

Figuur 3. Rooster van een 2d-tabel met al zijn marginalen



Tabel 4 is in Paragraaf 1 al geïntroduceerd (zie Tabel 1) om te lezer een concreet voorbeeld voor te schotelen van een tabelafrondprobleem. De bedoeling is om deze Tabel 4 af te ronden met afrondbasis 5. Een oplossing staat in Tabel 5.

Tabel 5. Gecontroleerd afgeronde Tabel 4 met afrondbasis 5

5	15	5	5	30
0	5	5	0	10
5	5	10	5	25
10	25	20	10	65

In een aantal gevallen is een cel waarde niet afgerond naar het dichtstbij gelegen veelvoud van 5. Deze cel waarden zijn grijs gekleurd.

2.4.2 Voorbeeld: ongecontroleerd afronden

Uitgaande van de vier tabellen weergegeven in Tabel 4 kunnen we laten zien wat er gebeurt als ongecontroleerd wordt afgerond naar het dichtstbijzijnde veelvoud van 5, dus iedere cel op zichzelf. Men krijgt dan het resultaat weergegeven in Tabel 6.

Tabel 6. Ongecontroleerd afgeronde Tabel 4 met afrondbasis 5

5	15	5	10	30
0	5	5	0	10
5	0	10	5	25
10	25	20	10	65

Het resultaat is echter niet additief, en dus inconsistent. Rijen 1 en 3 tellen niet op tot de bijbehorende rijtotalen. Dit geldt ook voor de kolommen 2 en 4. In Tabel 6 zijn de bijbehorende marginale cel waarden grijs gekleurd om aan te geven dat de bijbehorende rijen en kolommen niet additief zijn.

Opmerking. Ook stochastisch afronden, waarbij celsgewijs wordt afgerond en geen rekening wordt gehouden met de cumulatieve effecten, kan leiden tot niet-additieve tabellen. De reden is in feite dezelfde als in het in deze paragraaf besproken voorbeeld: er wordt lokaal afgerond en de overall structuur in de tabellen speelt geen rol. ■

2.4.3 Voorbeeld: ongecontroleerd afronden en terugrekenbaarheid

In dit voorbeeld laten we zien dat het ongecontroleerd afronden tot een situatie kan leiden waarbij de oorspronkelijke tabel met zijn marginalen kan worden teruggerekend. Als het doel was om de tabellen te beschermen door toepassing van afronding, is dit dus een totaal mislukte operatie. Het voorbeeld toont aan dat het gecontroleerd afronden niet alleen een esthetisch doel dient (in de zin dat afgeronde optelbare tabellen er beter uitzien, of handiger te gebruiken zijn) maar dat dit ook inhoudelijk iets te betekenen heeft. Dit is een klein voorbeeld, en ook atypisch in de dagelijkse afrondpraktijk, maar het is wel leerzaam.

We starten met Tabel 7.

Tabel 7. Oorspronkelijke tabel met marginalen.

6	2	8
3	1	4
9	3	12

Als we deze tabel ongecontroleerd afronden naar dichtstbijzijnde veelvouden van 5, krijgen we Tabel 8.

Tabel 8. Tabel 7, ongecontroleerd afgerond.

5	0	10
5	0	5
10	5	10

Iedere cel waarde in Tabel 8 staat in feite voor een verzameling mogelijke waarden. Deze zijn weergegeven in Tabel 9.

Tabel 9. Mogelijke waarden voor iedere cel

{3,4,5,6,7}	{0,1,2}	{8,9,10,11,12}
{3,4,5,6,7}	{0,1,2}	{3,4,5,6,7}
{8,9,10,11,12}	{3,4,5,6,7}	{8,9,10,11,12}

In Tabel 9 zijn de mogelijke waarden berekend, door louter per cel te kijken. Als we ook rekening houden met optelbaarheidseisen kunnen we de mogelijke waarden weer verder inperken. We kijken eerst naar de optelbaarheidseisen horizontaal. Het resultaat staat in Tabel 10.

Tabel 10. Mogelijkheden geëlimineerd, eerst rijsgewijs.

{6,7}	{0,1,2}	{8,9}
{3,4,5,6,7}	{0,1,2}	{3,4,5,6,7}
{8,9}	{3,4}	{11,12}

Indien we nu kolomsgewijs kijken kunnen we de mogelijkheden verder beperken. Het resultaat staat in Tabel 11. Het blijkt dat de cel waarden in de eerste kolom nu allemaal bekend zijn.

Tabel 11. Mogelijkheden geëlimineerd, nu kolomsgewijs

{6}	{1,2}	{8,9}
{3}	{1,2}	{3,4}
{9}	{3,4}	{11,12}

Door nu wederom rijsgewijs te kijken vinden we een verdere reductie van de mogelijkheden. Van drie nieuwe cellen zijn nu ook de cel waarden bekend, naast de drie uit de vorige stap. Het resultaat is te zien in Tabel 12.

Tabel 12. Mogelijkheden geëlimineerd, weer rijsgewijs

{6}	{2}	{8}
{3}	{1}	{4}
{9}	{3,4}	{11,12}

Door nog een keer kolomsgewijs te rekenen vinden we de mogelijke waarden van alle cellen. Het resultaat staat in Tabel 13.

Tabel 13. Oplossing, door weer kolomsgewijs te kijken

{6}	{2}	{8}
{3}	{1}	{4}
{9}	{3}	{12}

De oorspronkelijke tabel is nu feitelijk gereconstrueerd, omdat Tabel 13 maar één oplossing voorstelt.

Merk op dat bij deze herleiding van de oorspronkelijke tabel gebruik is gemaakt van bepaalde kennis over het afrondproces. Zo dient bekend te zijn dat de afrondbasis 5 is (dat is wel uit het resultaat af te lezen), maar ook dat ongecontroleerd is afgerond.

2.4.4 Voorbeeld: 3d-tabel met marginalen

We beschouwen de $3 \times 3 \times 2$ tabel in Tabel 14, kolommen 1 en 2. De bijbehorende marginalen zijn weergeven in Tabel 15 tot en met Tabel 21. Het bijbehorende rooster is eerder al getoond, namelijk in Figuur 1, waar het als voorbeeld van zo'n diagram is gepresenteerd voor precies dezelfde situatie als we hier hebben.

De bedoeling is om deze tabellen gecontroleerd af te ronden, met afrondbasis 1.

Tabel 14. Moedertabel opgespannen door variabelen x, y en z.

z=0	**y**		
x	0.498	0.854	0.047
	0.618	0.911	0.084
	0.711	0.961	0.119

z=1	**y**		
x	0.282	0.364	0.434
	0.304	0.383	0.449
	0.325	0.400	0.465

Tabel 15. x,y-rand

som(z)	**y**		
x	0.780	1.218	0.481
	0.922	1.294	0.533
	1.036	1.361	0.584

Tabel 16. x,z-rand

som(y)	**z**	
x	1.399	1.080
	1.613	1.136
	1.791	1.190

Tabel 17. y,z-rand

som(x)	**z**	
y	1.827	0.911
	2.726	1.147
	0.250	1.348

Tabel 18. x-rand

	som(y,z)
x	2.479
	2.749
	2.981

Tabel 19. y-rand

	som(x,z)
y	2.738
	3.873
	1.598

Tabel 20. z-rand

	som(x,y)
z	4.803
	3.406

Tabel 21. Totaal generaal

som(x,y,z)
8.209

De afgeronde tabellen zijn te vinden in Tabel 22 tot en met Tabel 29. De tabellen staan in dezelfde volgorde als hun originelen. Het bijschrift van de originele tabellen is aangevuld door er afgerond(e) aan toe te voegen.

Tabel 22. Afgeronde moedertabel opgespannen door variabelen x, y en z.

z=0	**y**		
x	0	1	0
	1	1	0
	1	1	0

z=1	**y**		
x	1	0	0
	0	0	1
	0	1	0

Tabel 23. Afgeronde x,y-rand

som(z)	**y**		
x	1	1	0
	1	1	1
	1	2	0

Tabel 24. Afgeronde x,z-rand

som(y)	**z**	
x	1	1
	2	1
	2	1

Tabel 25. Afgeronde y,z-rand

som(x)	**z**	
y	2	1
	3	1
	0	1

Tabel 26. Afgeronde x-rand

	som(y,z)
x	2
	3
	3

Tabel 27. Afgeronde y-rand

	som(x,z)
y	3
	4
	1

Tabel 28. Afgeronde z-rand

	som(x,y)
z	5
	3

Tabel 29. Afgerond totaal generaal

som(x,y,z)
8

2.4.5 Voorbeeld: afronden met moedertabel

Dit voorbeeld dient te worden gezien in relatie tot dat in Paragraaf 2.4.6. Op zichzelf genomen illustreert het voorbeeld in de onderhavige paragraaf hetzelfde als dat in Paragraaf 2.4.4.

We gaan uit van een $2 \times 2 \times 2$ tabel die we nu echter laagsgewijs beschrijven als 3 2×2 tabellen met hun 1d randen en het bijbehorende totaal generaal. We beschouwen feitelijk de drie 2×2 tabellen en hun marginalen, als weergegeven in Tabel 30, Tabel 31 en Tabel 32. Het bijbehorende rooster is als in Figuur 1.

Tabel 30. Laag 1

0.3	0.1	0.4
0.3	0.2	0.5
0.6	0.3	0.9

Tabel 31. Laag 2

0.3	0.3	0.6
0.1	0.3	0.4
0.4	0.6	1

Tabel 32. Laag 3 = som van laag 1 en laag 2

0.6	0.4	1
0.4	0.5	0.9
1	0.9	1.9

De bedoeling is om deze tabellen af te ronden met afrondbasis 1. Het resultaat is weergegeven in Tabel 33, Tabel 34 en Tabel 35.

Tabel 33. Laag 1 afgerond

0	0	0
1	0	1
1	0	1

Tabel 34. Laag 2 afgerond

0	1	1
0	0	0
0	1	1

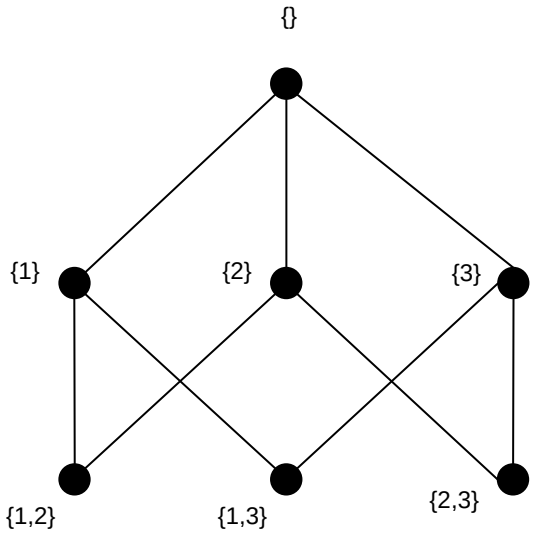
Tabel 35. Laag 3 afgerond (= som van laag 1 afgerond en laag 2 afgerond)

0	1	1
1	0	1
1	1	2

2.4.6 Voorbeeld: afronden zonder moedertabel

In dit voorbeeld gaan we uit van dezelfde tabellen als in het vorige voorbeeld, behalve het binnenwerk (de ‘moedertabel’). Met andere woorden we hebben drie 2d-tabellen en drie 1d-tabellen en een totaal generaal, als in Tabel 36 tot en met Tabel 39 weergegeven. Het bijbehorende rooster is afgebeeld in Figuur 4.

Figuur 4. Rooster van 3 gekoppelde tabellen



De randtabellen zijn als volgt.

Tabel 36. 1, 2-rand = 2,3-rand

0.4	0.6
0.5	0.4

Tabel 37. 1,3-rand

0.6	0.4
0.4	0.5

Tabel 38. 1-rand = 2-rand = 3-rand

1
0.9

Tabel 39. Totaal generaal

1.9

Na gecontroleerde afronding levert dat Tabel 40 tot en met Tabel 43 op.

Tabel 40. Afgeronde 1,2-rand = afgeronde 2,3 rand

0	1
1	0

Tabel 41.Afgeronde 1,3-rand

1	0
0	1

Tabel 42. Afgeronde 1-rand = afgeronde 2-rand = afgeronde 3-rand

1
1

Tabel 43. Afgerond totaal generaal

2

Met deze afgeronde tabellen, Tabel 40 tot en met Tabel 43 doet zich de situatie voor dat er geen moedertabel is voor deze tabellen, zoals eenvoudig te verifiëren is. Zie ook Cox (2003), waar de afgeronde tabellen set gegeven wordt als voorbeeld van een consistente tabellen set zonder moedertabel. In het onderhavige stuk is dit voorbeeld gebruikt als illustratie van een paradoxale situatie bij het afronden van tabellen. Overigens dient men zich te realiseren dat dergelijke consistente tabellen sets zonder moedertabel niet een grote uitzondering vormen. Wil men vermijden dat de afgeronde tabellen sets geen moedertabel hebben, dan dient men bij het afronden de oorspronkelijke moedertabel te betrekken.

Opmerking. Dit resultaat is een paradox, dus een schijnbare tegenstrijdigheid. Hij ontstaat als men de afgeronde cel waarden als 'echte' waarden neemt. Als men zich echter realiseert dat het om afgeronde waarden gaat, en dat ieder van deze waarden in feite staat voor een interval, is er niets aan de hand. Er is dan wel degelijk een moedertabel bij de afgeronde tabellen. Dit is een feite een niet lege verzameling van dergelijke tabellen. De oorspronkelijke moedertabel behoort tot deze verzameling. ■

2.4.7 Voorbeeld: normaliseren

We kunnen een tabelafrondingsprobleem normaliseren (dat wil zeggen in een normaal- of standaardvorm brengen) door bepaalde voorbewerkingen op de tabel toe te passen. Deze voorbewerkte tabel wordt dan afgerond en de uiteindelijke afgeronde tabel wordt verkregen door de afgeronde voorbewerkte tabel na te bewerken. Deze nabewerkingen zijn te zien als de inversen van de voorbewerkingen.

Stel $\beta > 0$ is de gekozen afrondbasis. We kunnen een cel waarde c in het binnenwerk van een tabel schrijven als $c = n_c\beta + r_c$, waarbij n_c een geheel getal is en voor r_c geldt dat $0 \le r_c < \beta$. We noemen r_c het residu, en als de afrondbasis 1 is ook wel het fractionele deel. Het afronden heeft alleen betrekking op het residu van de cel waarde. Dat geldt daarmee ook voor de hele tabel (binnenwerk). We kunnen overal de veelvouden van de afrondbasis aftrekken van de cel waarden, en dienovereenkomstig de marginalen aanpassen. Een voorbeeld illustreert de bedoeling. Stel dat de af te ronden tabellen gegeven zijn in Tabel 4 en dat de afrondbasis 5 is.

De bedoelde reductie leidt tot Tabel 44.

Tabel 44. Binnenwerk van Tabel 4 (modulo 5) en bijbehorende randen

3	2	3	3	11
2	4	0	1	7
2	2	1	3	8
7	8	4	7	26

Tabel 45 bevat de verschiltabellen van Tabel 4 en Tabel 44. Tabel 44 dient te worden afgerond en Tabel 45 dient dan bij het resultaat te worden opgeteld om een afgeronde Tabel 4 te krijgen.

Tabel 45. Verschiltabellen van Tabel 4 en Tabel 44

0	15	0	5	20
0	0	5	0	5
5	0	10	0	15
5	15	15	5	40

Op dit afronden van Tabel 44 gaan we hier verder niet in. Dat is niet anders dan bij andere tabellen. Het voordeel van de hier toegepaste reductie echter is dat het bij grote waarden in de begintabel tot een tabel leidt met alle cel waarden in een beperkte range, van 0 tot β . Het nadeel is dat het een extra rekenslag vergt.

De volgende reductie is door de waarden in Tabel 44 door 5 te delen. Dit levert Tabel 46 op.

Tabel 46. Tabel 44 gedeeld door 5

0,6	0,4	0,6	0,6	2,2
0,4	0,8	0	0,2	1,4
0,4	0,4	0,2	0,6	1,6
1,4	1,6	0,8	1,4	5,2

Tabel 46 kan nu worden afgerond met afrondbasis gelijk 1. Het resultaat van de afronding staat in Tabel 47.

Tabel 47. Tabel 46 gecontroleerd afgerond met afrondbasis 1

1	0	1	0	2
0	1	0	0	1
0	1	0	1	2
1	2	1	1	5

Een afronding van de oorspronkelijke tabel krijgt men nu door 5 maal Tabel 47 te nemen en Tabel 45 daarbij op te tellen. Het resultaat is gelijk aan Tabel 5.

Het normaliseren is in de praktijk meestal niet nodig. Niettemin is het aardig te weten dat afrondproblemen op de beschreven manier gestandaardiseerd kunnen worden.

2.5 Kwaliteitsindicatoren

Een kwalitatieve kwaliteitsindicator voor een set afgeronde tabellen is of zij additief, dus optelbaar, zijn of niet, als de oorspronkelijke set tabellen dat ook was. Zoals Voorbeeld 2.4.3 laat zien kan niet-additiviteit van de afgeronde tabellen, bijvoorbeeld verkregen door niet-gecontroleerd afronden, impliceren dat men de oorspronkelijke tabellen exact kan terugrekenen.

Een kwaliteitsindicator die men kan zien als een maat voor informatieverlies is de mate waarin iedere afgeronde tabel afwijkt van zijn origineel. Indien men een tabel T als vector weergeeft en $\overline{T}$ een afronding is van T (en ook als vector weergegeven) dan is $d(T,\overline{T})$, met d een metriek, een maat voor het informatieverlies. Een voorbeeld van zo'n metriek is $d_1(T,\overline{T}) = \sum_i |T_i - \overline{T_i}|$.

Een afgeronde tabel wordt doorgaans gevonden na oplossing van een geschikt optimaliseringsprobleem, waarbij een dergelijke afstand de doelfunctie is, en optelbaarheidscondities onder andere als constraints worden gebruikt. Hoe dichter de procedure een oplossing vindt in de buurt van het optimum, hoe minder informatieverlies er is en hoe beter de kwaliteit van de afronding is.

Als de afgeronde tabel bedoeld is om de corresponderende niet-afgeronde te beschermen, dan spelen ook overwegingen uit de statistische beveiliging mee om de kwaliteit van de afronding te beoordelen. Algemeen kan men stellen dat de mate waarin een afgeronde tabel de oorspronkelijke tabel beveiligt een kwalitatieve kwaliteitsindicator is. Ervan uitgaande dat de regels die het CBS hanteert om tabellen te beveiligen deugen, dan kan men stellen dat als de juiste procedure is gevolgd voor het afronden van een tabel, en men de juiste parameters heeft gebruikt, de afgeronde tabellen veilig zijn. De keuze van de afrondbasis kan men relateren aan de beveiligingsintervallen voor gevoelige cellen in de tabellen. (Zie Willenborg & De Waal (2001, Paragraaf 6.5) voor een bespreking van dit onderwerp.

Of een afgeronde tabel veilig is zou geverifieerd moeten worden door een onafhankelijke deskundige die de gevolge afrondprocedure heeft beoordeeld en goed heeft bevonden. Uiteraard is dit alles mensenwerk en ook de controleur kan fouten maken. Maar in de praktijk lijkt dit een geschikte werkwijze om er zeker van te zijn dat een toegepaste afrondingsprocedure veilige tabellen heeft opgeleverd.

3. Conclusie

Afronden van tabellen wordt op het CBS op diverse plaatsen gebruikt en in diverse processtappen. Het formaat van de af te ronden tabellen (in termen van het aantal cellen betrokken bij de afronding) kan ook zeer variëren, van enkele tientallen tot vele miljoenen.

In het bovenstaande hebben we een aantal oplossingen beschreven voor het afronden van tabellen. Dit is geen uitputtende lijst. De vraag dringt zich op wat de beste methode is, zo die er al is. Deze vraag kan eigenlijk niet beantwoord worden, omdat dit enerzijds van de specifieke doelstellingen van het afrondprobleem in kwestie afhangt, en anderzijds van de grootte van het afrondprobleem. Het resultaat van een afronding moet door een gebruiker bekeken worden, die vervolgens moet beslissen of hij daarmee kan leven, of dat het afrondprobleem iets veranderd moet worden, of dat zelfs een heel ander afrondalgoritme moet worden gebruikt.

Er zijn een aantal tools op het CBS beschikbaar waarmee tabellen kunnen worden afgerond (τ-ARGUS, WinAdjust). Deze tools zijn niet geschikt voor zeer grote afrondproblemen. Voor dergelijke problemen is een maatwerkoplossing nodig (zie Paragraaf 2.3.8).

4. Literatuur

Aarts, E. & Lenstra, J.K. (1997), *Local Search in Combinatorial Optimization*, Wiley.

Castro, J. (2006), Minimum-distance Controlled Perturbation Methods for Large-scale Tabular Data Protection. *European Journal of Operational Research* **171**, 39-52.

Castro, J. (2011), Extending Controlled Tabular Adjustment for Non-additive Tabular Data with Negative Protection Levels. *Statistics and Operations Research Transactions* **35**, 3-20.

Causey, B.D., Cox, L.H. & Ernst, L.R. (1985), Applications of Transportation Theory to Statistical Problems. *Journal of the American Statistical Association* **80**, 903-909.

Cox, L.H. (1987), A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association* **82**, 520-524.

Cox, L.H. (2003), On Properties of Multidimensional Tables. *Journal of Statistical Planning and Inference* **117**, 251-273.

Cox, L.H. & Ernst, L.R. (1982), Controlled Rounding. *INFOR* **20**, 423-432.

Cox, L.H. & George, J.A. (1989), Controlled rounding for Tables with Subtotals. *Annals of Operations Research* **20**, 141-157.

Dandekar, R.A. & Cox, L.H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. Manuscript, Energy Information Administration, U.S. Department of Energy.

Doerr, B. & Klein, C. (z.j.), Unbiased Rounding of Rational Matrices. Report, Max-Planck-Institut für Informatik, Saarbrücken, BRD.

Doerr, B., Friedrich, T., Klein, C., Osbild, R. (2006), Unbiased Matrix Rounding. In: *10-th Scandinavian Workshop on Algorithm Theory, Riga, Latvia*, Lecture Notes in Computer Science, Springer-Verlag, Vol. 4059, 102-112.

Fellegi, I.P. (1975), Controlled Random Rounding. *Survey Methodology* **1**, 123-133.

Fischetti, M & Salazar-González, J.-J. (1998), Computational Experience with the Controlled Rounding Problem in Statistical Disclosure Control. *Journal of Official Statistics* **14**, 553-565.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P.-P., Giessing, S., Fischetti, M., Salazar-González, J.-J, Castro, J. & Lowthian, Ph. (2011), *τ-ARGUS (version 3.5) User's Manual*. CBS, The Hague.

Kelly, J.P., Golden, B.L. & Assad, A.A. (1990), Using Simulated Annealing to Solve Controlled Rounding Problems. *ORSA Journal on Computing* **2**, 174-185.

Kelly, J.P., Golden, B.L. & Assad, A.A. (1993), Large-scale Controlled Rounding Using TABU Search with Strategic Oscillation. *Annals of Operations Research* **41**, 69-84.

Kelly, J.P., Golden, B.L. & Assad, A.A. & Baker, E.K. (1990), Controlled Rounding of Tabular Data. *Operations Research* **38**, 760-772.

Motwani, R. & Raghavan, P. (1995), *Randomized Algorithms*, Cambridge University Press.

Raghavan, P. & Thompson, C.D. (1987), Randomized Rounding: A Technique for Provably Good Algorithms and Algorithmic Proofs. *Combinatorica* **7**, 365-374.

Rounding in Wikipedia: http://en.Wikipedia.org/wiki/Rounding.

Salazar-González, J.-J. (2005), A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods. *Operations Research* **53**, 819-829.

Salazar-González, J.-J. (2006), Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. *Mathematical Programming, Series B* **105**, 583-603.

Willenborg, L. & De Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics 155, Springer-Verlag.