Steekproeftheorie

Deelthema: Herhaald wegen

08

José Gouweleeuw en Paul Knottnerus

Statistische Methoden (08006)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen 2008

Verklaring van tekens

.	= gegevens ontbreken
*	= voorlopig cijfer
x	= geheim
–	= nihil
–	= (indien voorkomend tussen twee getallen) tot en met
0 (0,0)	= het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	= een cijfer kan op logische gronden niet voorkomen
2005–2006	= 2005 tot en met 2006
2005/2006	= het gemiddelde over de jaren 2005 tot en met 2006
2005/'06	= oogstjaar, boekjaar, schooljaar enz., beginnend in 2005 en eindigend in 2006
2003/'04–2005/'06	= oogstjaar, boekjaar enz., 2003/'04 tot en met 2005/'06

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

Colofon

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

Prepress
Centraal Bureau voor de Statistiek - Facilitair bedrijf

Omslag
TelDesign, Rotterdam

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

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

Internet
www.cbs.nl

ISSN: 1876_0333

6016508006 X-37

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving

In de standaard manier van ophogen hoort normaliter bij iedere steekproef een vaste set van gewichten waarmee de waarnemingen van een bepaalde doelvariabele in die steekproef moeten worden opgehoogd. Dit kan er evenwel toe leiden dat er verschillende cijfers naast elkaar worden gepubliceerd over één en dezelfde variabele wanneer er twee publicaties zijn over die variabele op basis van twee verschillende onderzoeken. Een dergelijke discrepantie of inconsistentie kan zich vooral voordoen bij multidimensionale tabellen die vaak een marginaal gemeenschappelijk hebben terwijl de onderliggende tabellen zijn geschat op basis van verschillende onderzoeken.

In dit document beschrijven we de techniek van het *herhaald wegen*. Bij deze techniek wordt in voorkomende gevallen bij een bepaalde tabel de set gewichten aangepast. Deze veelal cosmetische aanpassing gebeurt op zo'n manier dat de consistentie met andere tabellen of met het register weer volledig wordt hersteld.

1.2 Afbakening en relatie met andere thema's

Het spreekt voor zich dat er relaties bestaan tussen aan de éne kant herhaald wegen en aan de andere kant de thema's *Steekproeftheorie* en *Wegen als correctie voor non-respons*. De daar besproken methoden en technieken resulteren in het algemeen in een set van gewichten die in het verdere statistische proces worden gebruikt bij de ophoging. Deze gewichten dienen weer als input voor het herhaald wegen van een bepaalde tabel wanneer de desbetreffende schattingen niet consistent blijken te zijn met reeds geschatte of bekende cijfers. Herhaald wegen resulteert dan voor die tabel in een set van nieuwe gewichten en aangepaste tabelschattingen die wel consistent zijn.

1.3 Plaats in het statistisch proces

Uit het voorafgaande moge het duidelijk zijn dat herhaald wegen zich afspeelt aan het einde van het statistische proces vlak voor het publiceren van de cijfers. Immers het doel van herhaald wegen is het voorkomen van gepubliceerde cijfers die numeriek inconsistent zijn.

2. Herhaald wegen

2.1 Korte beschrijving

In de statistische praktijk zijn verschillende schatters mogelijk om het totaal van een doelvariabele te schatten. In de praktijk moeten er meestal totalen van meerdere doelvariabelen worden geschat. Dit hoeft ook niet altijd op basis van een enkele steekproef, soms kunnen hier meerdere steekproeven voor worden gebruikt. Wanneer verschillende tabellen worden geschat op basis van meerdere steekproeven, hoeven deze niet noodzakelijke numeriek consistent te zijn. Wanneer bijvoorbeeld een tabel *leeftijd* × *opleiding* uit een of andere steekproef wordt geschat en een tabel *gezondheid* × *opleiding* uit een andere steekproef dan zullen de schattingen voor de verschillende categorieën van opleiding over het algemeen niet exact overeenkomen. Immers, op schattingen die gebaseerd zijn op steekproefdata zit een steekproeffout waardoor de schattingen kunnen afwijken van het echte populatietotaal. Uiteraard moeten ze als het goed is binnen de steekproefmarge wel overeen komen.

Met de methode van herhaald wegen worden dergelijke numerieke inconsistenties tussen schattingen uit verschillende bronnen zo veel mogelijk voorkomen (zie bijvoorbeeld Kroese en Renssen, 1999). Het voornaamste doel van herhaald wegen is dus puur cosmetisch: het voorkomen van verschillen door steekproeffouten in schattingen. In plaats van één set gewichten die wordt gebruikt om alle tabellen te schatten, krijgt bij herhaald wegen iedere tabel zijn eigen set gewichten. Feitelijk komt de methode neer op het herhaald toepassen van de regressieschatter. Om de uitvoering van herhaald wegen te vereenvoudigen is speciale software ontwikkeld: het pakket VRD (Vullen Reference Database).

2.2 Toepasbaarheid

Bij het CBS wordt het herhaald wegen vooral toegepast bij de sociaal-economische statistieken. Zo is herhaald wegen gebruikt om de tabellen voor de Volkstelling 2001 te schatten. Met deze methode was het mogelijk om 40 consistente tabellen aan Eurostat te leveren, zonder dat er een heuse volkstelling voor nodig was. Voor meer informatie over de volkstelling wordt de lezer verwezen naar Schulte Nordholt (2005) en Advokaat et al. (2004); zie ook paragraaf 2.4. Ook is herhaald wegen gebruikt bij het Loonstructuuronderzoek 2002 (zie Gouweleeuw, 2005).

Alvorens over te gaan tot herhaald wegen dient men er zich van te vergewissen dat aan de volgende vier voorwaarden is voldaan:

1. de steekproeven en registers hebben betrekking op dezelfde periode;
2. de steekproeven en registers hebben betrekking op dezelfde populatie;
3. een variabele met dezelfde naam heeft op alle plaatsen dezelfde definitie;

4. verschillende classificaties van dezelfde categoriale variabele zijn genest ofwel een klasse van een bepaalde classificatie zit altijd in precies één klasse van een minder gedetailleerde classificatie.

In de volgende paragraaf gaan we hier nader op in.

2.3 Uitgebreide beschrijving

2.3.1 Inleiding

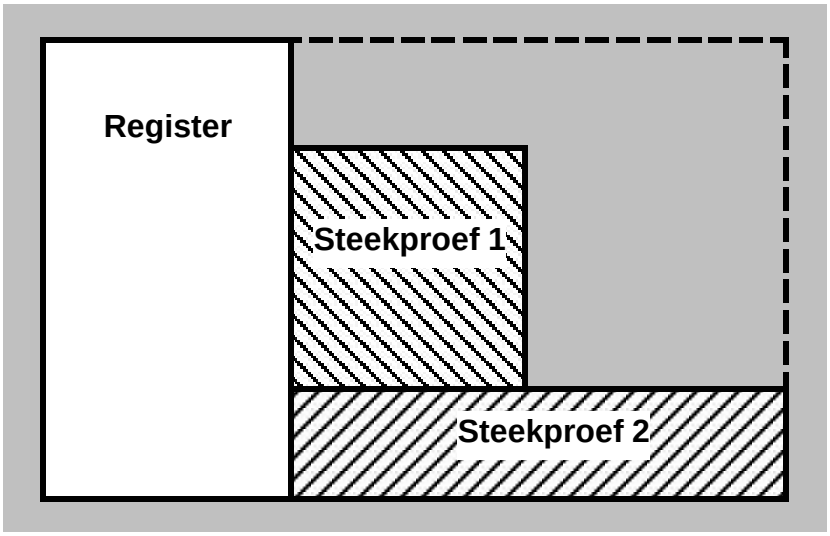
In deze paragraaf zal globaal worden beschreven welke stappen er moeten worden doorlopen bij het herhaald wegen. De precieze, wiskundige uitwerking van deze stappen zal in de volgende paragrafen worden gedaan. Herhaald wegen wordt met name gebruikt wanneer er een groot aantal tabellen moet worden geschat, en dit schatten moet numeriek consistent worden uitgevoerd. Het belangrijkste idee achter herhaald wegen is dat er meer dan één set gewichten per steekproefbestand kan worden gebruikt. Voor iedere tabel die moet worden geschat, kunnen de gewichten (als dit nodig is) worden aangepast om ervoor te zorgen dat schattingen van de nieuwe tabel consistent zijn met de schattingen van de tabellen die eerder zijn gemaakt. Dit komt in feite neer op het herhaald toepassen van de regressieschatter (zie bijvoorbeeld Särndal et al, 1992). Iedere tabel kan in principe met een andere set gewichten worden geschat, zelfs wanneer verschillende tabellen op basis van dezelfde steekproef worden geschat.

Voorbeeld. Beschouw ter illustratie het volgende (fictieve) voorbeeld. Veronderstel dat de twee tabellen *opleiding* × *leeftijd* en *opleiding* × *gezondheid* moeten worden geschat. Hiertoe zijn één register en twee steekproeven beschikbaar, zoals geschetst in Figuur 1. In het register is de variabele leeftijd (weergegeven met X) in twee categorieën (jong, oud) opgenomen. Er zijn volgens het register 30 jonge personen en 70 oude personen in de populatie. Steekproef 1 bevat naast leeftijd ook opleidingsniveau (weergegeven met Z) in twee categorieën (hoog, laag) en steekproef 2 bevat naast leeftijd en opleiding ook de variabele gezondheid (weergegeven met Y) in twee categorieën (goed, slecht). De tabel *opleiding* × *leeftijd* (= $Z \times X$) kan nu op basis van de gecombineerde steekproeven 1 en 2 worden geschat, en de tabel *opleiding* × *gezondheid* (= $Z \times Y$) kan op basis van steekproef 2 worden geschat. Dit levert de volgende tabellen op.

$Z \times X$	jong	oud	totaal
hoog	20	30	50
laag	9	41	50
totaal	29	71	100

$Z \times Y$	goed	slecht	totaal
hoog	32	20	52
laag	28	20	48
totaal	60	40	100

Figuur 1. Schematische weergave van de beschikbare data



De geschatte tabellen leveren een tweetal inconsistenties op. Allereerst komt de geschatte verdeling over de categorieën van leeftijd niet overeen met de bekende verdeling in de populatie. Daarnaast leveren beide tabellen inconsistente schattingen voor het opleidingsniveau op. Met herhaald wegen worden de schattingen als volgt aangepast. Allereerst worden de gewichten van de tabel $Z \times X$ aangepast, zodat de marginale tabel van leeftijd consistent is met de bekende verdeling in de populatie. Hiervoor wordt de regressieschatter gebruikt. Vervolgens worden de gewichten van de tabel $Z \times Y$ aangepast, zodat de marginale tabel van opleiding consistent is met de marginaal uit $Z \times X$. Hiervoor wordt wederom de regressieschatter gebruikt, waarbij de geschatte tabel van opleiding als bekend populatietotaal wordt beschouwd. Dit levert de volgende tabellen op, die nu wel numeriek consistent zijn.

$Z \times X$	jong	oud	totaal
hoog	20	30	50
laag	10	40	50
totaal	30	70	100

$Z \times Y$	goed	slecht	totaal
hoog	31	19	50
laag	29	21	50
totaal	60	40	100

Zoals eerder is opgemerkt, is herhaald wegen alleen bruikbaar wanneer aan de volgende voorwaarden is voldaan. De verschillende databronnen die worden gebruikt, moeten betrekking hebben op dezelfde periode en populatie. Verder moet een variabele met dezelfde naam op alle plaatsen dezelfde definitie hebben. Om herhaald wegen te kunnen toepassen moet een aantal stappen worden doorlopen:

1. Specificatie van de tabellen.
2. Afbakenen van datablokken.
3. Bepalen van startgewichten.
4. Uitvoeren van herhaald wegen.

Deze stappen zullen hier kort achtereenvolgens worden beschreven.

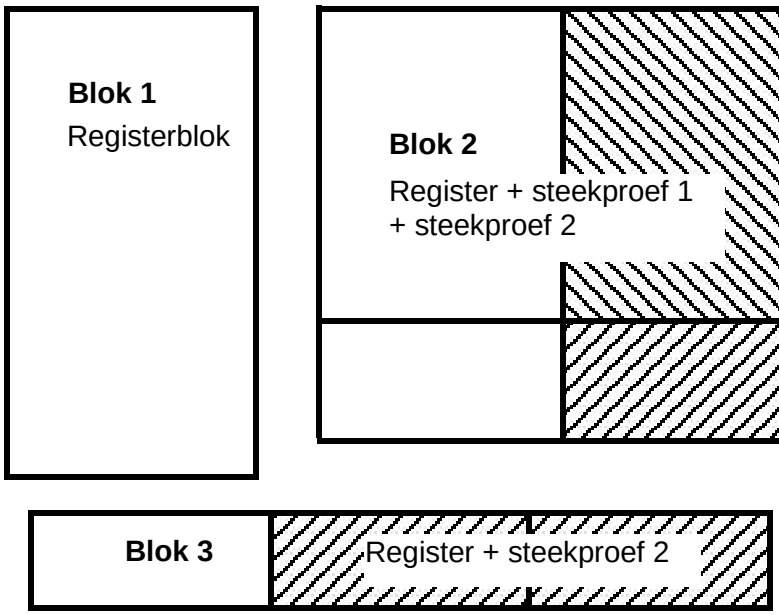
Allereerst moeten de te schatten tabellen worden gespecificeerd. Iedere tabel wordt gekarakteriseerd door één of meerdere classificatievariabelen in combinatie met hetzij een telvariabele hetzij een kwantitatieve variabele. De telvariabele geeft aan

dat er populatie-elementen worden geteld in de tabel. Dit kan bijvoorbeeld het aantal personen of huishoudens zijn. Bij een kwantitatieve variabele kan men denken aan bijvoorbeeld een tabel met per cel het (geschatte) inkomenstotaal van alle personen in die cel. De classificatievariabelen beschrijven welke groepen men binnen de populatie beschouwt, bijvoorbeeld geslacht of leeftijd. Een classificatievariabele verdeelt de populatie in een aantal disjuncte groepen. Deze variabelen kunnen op verschillende niveaus voorkomen, bijvoorbeeld leeftijd per jaar en leeftijd in 5-jaarsklassen. Het classificatieniveau wordt altijd aangegeven door een getal tussen haakjes achter de naam van de variabele waarbij het hoogste cijfer slaat op de meest gedetailleerde classificatie, bijvoorbeeld $L^{(1)}$ en $L^{(2)}$ voor leeftijd in 5-jaarsklassen en leeftijd per jaar.

Voor herhaald wegen is het van belang dat de gebruikte classificaties hiërarchisch zijn. Dit betekent dat iedere klasse op een hoger niveau kan worden gevormd door hele klassen van een lager niveau samen te voegen. Op deze manier is een tabel met een variabele op een lager niveau altijd een marginale tabel van dezelfde variabele op een hoger niveau. Dergelijke hiërarchische classificaties voorkomen dat herhaald wegen nodeloos ingewikkeld wordt.

In stap 2 moet eerst worden onderzocht welke bronnen (registraties en/of enquêtes) er beschikbaar zijn voor het schatten van de tabellen. Deze bronnen moeten vervolgens op microniveau aan elkaar worden gekoppeld. Hierdoor ontstaat een microdatabase die voor ieder element in de populatie een record bevat. De variabelen in de registers zijn beschikbaar voor ieder element. Voor sommige elementen is er extra informatie beschikbaar uit één of meerdere steekproeven. Uit deze database moeten rechthoekige, volledig gevulde datablokken worden afgeleid. Ieder datablok bevat alle records die een bepaalde maximale verzameling variabelen gemeenschappelijk hebben. Er is bijvoorbeeld een registerblok dat records bevat voor alle elementen in de populatie en alle variabelen uit het register.

Figuur 2. Datablokken bij Figuur 1



Een ander voorbeeld is een blok gebaseerd op één steekproef. Dit blok bevat records voor alle elementen die in de steekproef voorkomen en alle variabelen uit zowel het register als de steekproef. Gecompliceerdere blokken die uit meer dan een steekproef bestaan zijn ook mogelijk. Uit de datasets die in figuur 1 zijn weergegeven, kan in totaal een drietal relevante blokken worden afgeleid. Deze blokken zijn in figuur 2 weergegeven. Voordat de gewenste tabellen uit de rechthoekige datablokken kunnen worden geschat of geteld, moet ieder blok worden voorzien van een startgewicht voor herhaald wegen. Uit het registerblok moeten alleen tellingen worden gemaakt. Om dit te kunnen uitvoeren, wordt aan ieder record in het registerblok een startgewicht dat gelijk is aan 1 toegekend. Voor een blok dat ook steekproefinformatie bevat, worden de startgewichten veelal gebaseerd op de publicatiegewichten uit het betreffende onderzoek. Wanneer een blok uit meerdere steekproeven bestaat, moeten de gewichten van de verschillende steekproefonderzoeken gecombineerd worden. In de volgende subparagraaf zal worden beschreven hoe dit combineren moet worden uitgevoerd.

Wanneer de startgewichten zijn bepaald, kunnen de tabellen in stap 4 worden geschat. Hierbij moet eerst worden vastgesteld in welke volgorde de tabellen worden geschat. De tabellen die uit het grootste datablok (met de meeste records) kunnen worden geschat staan in principe vooraan. De reden hiervoor is dat een groter datablok een kleinere steekproeffout heeft althans wanneer alle steekproeven aselect zijn getrokken. Daarna komen de tabellen die uit het op één na grootste datablok kunnen worden geschat, etc. Wanneer van een bepaalde tabel de marginalen uit een groter blok kunnen worden geschat, worden deze marginalen vóór de tabel zelf geschat. Per blok komen eerst de tabellen die consistent kunnen worden geschat op basis van de startgewichten, daarna komen de overige tabellen in een min of meer willekeurige volgorde. Wanneer herhaald wegen wordt uitgevoerd, moet per tabel eerst worden onderzocht welke marginalen deze gemeenschappelijk heeft met reeds geschatte tabellen. De (geschatte of getelde) populatietotalen van deze gemeenschappelijke marginalen worden gebruikt als bekende populatietotalen waarop de regressieschatter wordt gecalibreerd. Merk op dat de resultaten van herhaald wegen op deze manier afhankelijk zijn van de volgorde waarin de tabellen worden geschat. Immers, iedere tabel moet consistent worden geschat met de eerder geschatte tabellen, en dit is duidelijk afhankelijk van de volgorde. Deze afhankelijkheid kan worden omzeild door gebruik te maken van de splitting-up procedure. Deze houdt in dat bij iedere tabel die wordt geschat ook alle marginale tabellen afzonderlijk (en eerder) worden geschat.

2.3.2 Methodologie

In deze subparagraaf geven we een wat technische beschrijving van herhaald wegen. Voor een goed begrip van deze subparagraaf is het nodig dat de lezer vertrouwd is met de basisprincipes van matrixalgebra.

Bij herhaald wegen kunnen twee belangrijke varianten worden onderscheiden: de splitting-up procedure en het zogenaamde minimaal (her)wegen. In deze subparagraaf zullen we eerst aandacht besteden aan de splitting-up procedure die bij

een eerste kennismaking in het algemeen als eenvoudiger wordt ervaren. Daarna gaan we in op de zogenaamde minimale weegmodellen. De laatste methode wordt in de praktijk veel vaker gebruikt omdat de splitting-up procedure bij een beperkte dataset leidt tot een te groot aantal te schatten tabellen. Bij de tabellen die dan onderaan de lijst te schatten tabellen staan, zijn inmiddels zoveel randvoorwaarden dat het niet mogelijk is om de tabel consistent te schatten.

2.3.2.1 De splitting-up procedure

Bij de splitting-up procedure zijn drie stappen van belang. Eerst moet de set van doeltabellen worden gespecificeerd. Na een eventuele aanvulling moeten de tabellen in de goede volgorde worden gezet. Ten tweede dienen alle tabellen vervolgens in die volgorde te worden geschat met de gebruikelijke regressieschatter. Ten derde, wanneer in stap 2 de schattingen van een bepaalde tabel strijdig zijn met die van eerder geschatte tabellen of registers, moeten de gewichten van die tabel zodanig worden aangepast dat hij in overeenstemming is met de eerder geschatte cijfers en met de cijfers van de beschikbare registers. Hieronder zullen we de drie stappen wat uitgebreider beschrijven.

Stap 1. Specificatie en ordenen van de tabellen

Na een eerste specificatie van de tabellenset moet de set worden uitgebreid met de tabellen van alle splitting-up marginalen van de tabellen uit de oorspronkelijke tabellenset. Om uit te leggen wat een splitting-up marginaal is, beschouwen we een willekeurige driedimensionale tabel $T = A^{(r_1)} \times B^{(r_2)} \times C^{(r_3)}$. Tussen haakjes staat een getal r_k (k=1,2,3) dat de mate van detaillering aangeeft van de desbetreffende categoriale variabele. Hoe groter dat getal is, des te fijner is de classificatie van deze variabele in de tabel. In het extreme geval dat $r_k = 0$ hebben we te maken met de grofst mogelijke classificatie ofwel met een categoriale variabele die voor alle elementen in de populatie de waarde 1 aanneemt. De drie splitting-up marginalen van tabel T zijn nu gedefinieerd als de volgende drie (multiple categoriale) variabelen

$$A^{(r_1-1)} \times B^{(r_2)} \times C^{(r_3)}$$
$$A^{(r_1)} \times B^{(r_2-1)} \times C^{(r_3)}$$
$$A^{(r_1)} \times B^{(r_2)} \times C^{(r_3-1)}.$$

Dit zijn de meest gedetailleerde marginalen van tabel T die er zijn. Vervolgens moeten ook de splitting-up marginalen van deze marginalen worden toegevoegd, enzovoorts. Nadat alle tabellen van de splitting-up marginalen zijn toegevoegd aan de set, moeten de tabellen zo worden geordend dat de marginalen van een tabel altijd eerder worden geschat dan de tabel zelf.

Merk ook op dat gegeven het feit dat deze tabel gaat over de categoriale variabelen A, B en C de tabel volledig wordt gekarakteriseerd door de rijvector $r = (r_1, r_2, r_3)$.

Stap 2. Schatting van de tabellen met behulp van de regressieschatter

Om de formules bij het schatten van de tabellen goed te kunnen weergeven, moet eerst enige aandacht worden besteed aan het vectoriseren van de tabellen. Indien *A* een categoriale variabele is met *P* categorieën, is de vector a_k (k=1,…,N) voor element *k* van de populatie gedefinieerd als een vector met (*P*-1) nullen en één 1 die de correcte categorie van element *k* aangeeft. Hetzelfde kunnen we doen voor de categoriale variabelen *B, C,* enz. De rechthoekige frequentietabel *A×B* in gevectoriseerde vorm waarbij de kolommen van de tabel op elkaar worden gestapeld, kan dan worden geschreven als

$$t_{A\times B} = \sum_{k=1}^{N} b_k \otimes a_k \ .$$

Hierbij staat $\otimes$ voor het Kroneckerproduct. Voor twee willekeurige matrices *M* en *N* betekent dit symbool

$$M \otimes N = \begin{pmatrix} m_{11}N & \ldots & m_{1q}N \\ : & & \\ m_{p1}N & \ldots & m_{pq}N \end{pmatrix},$$

waarbij *p* en *q* de dimensies van *M* aangeven. Voor een meerdimensionale tabel met een multiple categoriale variabele *Y* van de vorm *Y=A×B×C×…* kan dit vectoriseren eenvoudig worden gegeneraliseerd. De resulterende vector duiden we kortweg aan met t_Y.

Verder definiëren we x_k als de vector met de waarden van de *J* hulpvariabelen voor element *k* in de populatie. Definieer t_x als de vector van de bijbehorende populatietotalen uit het register.

Voorbeeld 1. Indien *leeftijd* met bijvoorbeeld vijf 10-jaarsklassen als hulpvariabele wordt gebruikt, bestaat x_k in feite uit vijf dummyvariabelen waarvan één de waarde 1 aanneemt corresponderend met de leeftijdsklasse van persoon *k*. De andere dummy's krijgen de waarde 0. In dit geval zeggen we ook wel dat we wegen naar *leeftijd*.

Voorbeeld 2. Indien naast *leeftijd* ook de variabele *regio* met vier categorieën als hulpvariabele wordt gebruikt komen er in de vector x_k nog vier dummy's bij om de regio van persoon *k* aan te geven. Nu wordt er gewogen naar *leeftijd* en *regio* ofwel kortweg *L+R*. Dit laatste wordt ook wel het weegschema of weegmodel genoemd. Wanneer we de vector met de dummy's voor *leeftijd* aanduiden met l_k en die voor *regio* met r_k, geldt er

$$x_k = \begin{pmatrix} l_k \\ r_k \end{pmatrix}.$$

Voorbeeld 3. Vaak wordt ook een kruisvariabele zoals bijvoorbeeld *leeftijd×regio* gebruikt als hulpvariabele. Dit geeft zoals we hierboven al hebben gezien $x_k = r_k \times l_k$.

Met S_Y duiden we het grootste blok aan waarmee t_Y kan worden geschat. Eenvoudigheidshalve nemen we aan dat bij het schatten van de tabellen steeds dezelfde hulpvariabelen worden gebruikt. De regressieschatter van een willekeurige, gevectoriseerde tabel t_Y kan dan worden geschreven als

$$\begin{aligned} \hat{t}_Y^{REG(S_Y)} &= \hat{t}_Y^{HT(S_Y)} + \hat{B}_{d,x}'(t_x - \hat{t}_x^{HT(S_Y)}) \\ &= \sum_{k \in S_Y} d_k^{(S_Y)} y_k + \hat{B}_{d,x}'(t_x - \sum_{k \in S_Y} d_k^{(S_Y)} x_k) \\ \hat{B}_{d,x} &= (\sum_{k \in S_Y} d_k^{(S_Y)} x_k x_k')^{-1} \sum_{k \in S_Y} d_k^{(S_Y)} x_k y_k', \end{aligned} \tag{2.1}$$

waarbij $\hat{t}_Y^{HT(S_Y)}$ en $\hat{t}_x^{HT(S_Y)}$ Horvitz-Thompson (HT) schatters zijn uit blok S_Y, terwijl y_k', x_k' en B' de getransponeerde vormen zijn van respectievelijk y_k, x_k en B. De $d_k^{(S_Y)}$ staan voor de startgewichten in blok S_Y. Voor een enkelvoudige steekproef zijn deze gelijk aan $d_k^{(S_Y)} = 1/\pi_{S_Y,k}$. Hierbij staat $\pi_{S_Y,k}$ voor de eerste orde insluitkans die hoort bij het blok of in dit geval de steekproef S_Y. Voor een blok dat bestaat uit de vereniging van twee steekproeven zoals bijvoorbeeld EBB_{2006} en EBB_{2007}, zeg S_1 and S_2, kunnen we de desbetreffende schatter definiëren als

$$\hat{t}_Y^{HT(S_Y)} \circ \lambda_1 \hat{t}_Y^{HT(S_1)} + (1 - \lambda_1)\hat{t}_Y^{HT(S_2)} \circ \sum_{k \in S_Y} d_k^{(S_Y)} y_k,$$

waar λ_1 het relatieve belang aangeeft van S_1 in blok S_Y. Dit houdt in dat

$$d_k^{(S_Y)} = \begin{cases} \lambda_1 / \pi_{1k} & \text{als } k \in S_1 \text{ en } k \notin S_2 \\ (1 - \lambda_1)/\pi_{2k} & \text{als } k \in S_2 \text{ en } k \notin S_1 \\ \lambda_1/\pi_{1k} + (1 - \lambda_1)/\pi_{2k} & \text{als } k \in S_1 \cap S_2. \end{cases} \tag{2.2}$$

Nu moet alleen de waarde van λ_1 nog worden bepaald. Een eenvoudige manier is om λ_1 evenredig te nemen aan de omvang van S_1. Ook kan men de waarde van λ_1 laten afhangen van de steekproeffout en het non-respons percentage in S_1 in vergelijking met die van S_2.

Voor een blok van de vorm $S_Y = S_1 \cap S_2$ zijn de startgewichten gelijk aan

$$d_k^{(S_Y)} = 1/\pi_{1k}\pi_{2k|1k} \qquad (k \in S_1 \cap S_2),$$

waarbij $\pi_{2k|1k}$ staat voor de voorwaardelijke kans dat element *k* wordt geselecteerd voor S_2, gegeven het feit dat element *k* reeds is geselecteerd voor S_1. Wanneer beide steekproeven onafhankelijk zijn, geldt er $d_k^{(S_Y)} = 1/\pi_{1k}\pi_{2k}$.

Door de uitdrukking voor $\hat{B}_{d,x}$ te substitueren in (2.1) kan de regressieschatter ook worden geschreven in termen van gewichten

$$\begin{aligned} \hat{t}_Y^{REG(S_Y)} &= \sum_{k\in S_Y} w_k^{(S_Y)} y_k \\ w_k^{(S_Y)} &= d_k^{(S_Y)}\{1 + x_k' (\sum_{k\in S_Y} d_k^{(S_Y)} x_k x_k')^{-1}(t_x - \hat{t}_x^{HT(S_Y)})\}. \end{aligned} \tag{2.3}$$

De aldus verkregen gewichten voldoen aan de zogenaamde calibratievergelijkingen $\sum_{k\in S_Y} w_k^{S_y} x_k = t_x$. Dit volgt uit

$$\begin{aligned} \sum_{k\in S_Y} w_k^{(S_Y)} x_k &= \sum_{k\in S_Y} d_k^{(S_Y)}\{x_k + x_k x_k' (\sum_{k\in S_Y} d_k^{(S_Y)} x_k x_k')^{-1}(t_x - \hat{t}_x^{HT(S_Y)})\} \\ &= \hat{t}_x^{HT(S_Y)} + t_x - \hat{t}_x^{HT(S_Y)} = t_x. \end{aligned}$$

Verder dient te worden opgemerkt dat wanneer de te inverteren matrices in (2.1) en (2.3) singulier blijken te zijn, de inversen moeten worden vervangen door de gegeneraliseerde inversen. Ook de zo verkregen gewichten voldoen dan aan de zojuist genoemde calibratievergelijkingen; zie Renssen en Martinus (2002). Een alternatieve methode om singuliere matrices te voorkomen is om overtollige hulpvariabelen te verwijderen.

Stap 3. Herhaald wegen

Indien bij het schatten van een bepaalde tabel t_Y in stap 2 een marginaal is geschat die al eerder is geschat met andere uitkomsten, moet deze tabel worden herwogen. Een andere mogelijkheid waarbij moet worden herwogen is dat één van de marginalen van de tabel al bekend is uit het register maar bij het wegen in stap 2 is daar geen rekening mee gehouden of anders gezegd, de desbetreffende marginaal maakte geen deel uit van de hulpvariabelen in x_k.

Zij m_k de vector met de waarden voor element *k* van alle variabelen uit de splitting-up marginalen van de desbetreffende tabel. Laat $\hat{t}_m^{HW}$ de vector zijn van de geschatte marginalen op basis van een eerdere tabel of een integrale telling uit het register. Wanneer een marginaal in eerdere tabellen niet voorkomt en ook niet bekend is uit het register, is het desbetreffende element in $\hat{t}_m^{HW}$ simpelweg gelijk aan de regressieschatter uit stap 2. Herhaald wegen van tabel t_Y is dan nodig wanneer de regressiegewichten $w_k^{(S_Y)}$ uit stap 2 *niet* voldoen aan de consistentie-eisen $\sum_{k\in S_Y} w_k^{(S_Y)} m_k = \hat{t}_m^{HW}$. De HW-schatter van tabel t_Y is nu analoog aan de regressieschatter gedefinieerd als

$$\hat{t}_Y^{HW} = \hat{t}_Y^{REG(S_Y)} + \hat{B}_{w,m}(\hat{t}_m^{HW} - \hat{t}_m^{REG(S_Y)})$$
$$\hat{B}_{w,m} = (\sum_{k \in S_Y} w_k^{(S_Y)} m_k m_k')^{-1} \sum_{k \in S_Y} w_k^{(S_Y)} m_k y_k'. \qquad (2.4)$$

Analoog aan (2.3) kan de HW-schatter ook weer worden geformuleerd in termen van gewichten

$$\hat{t}_Y^{HW} = \sum_{k \in S} r_k^{(Y)} y_k$$
$$r_k^{(Y)} = w_k^{(S_Y)} \{1 + m_k' (\sum_{k \in S_Y} w_k^{(S_Y)} m_k m_k')^{-1} (\hat{t}_m^{HW} - \hat{t}_m^{REG(S_Y)})\}. \qquad (2.5)$$

De gewichten $r_k^{(Y)}$ voldoen wel aan de consistentie-eisen $\sum_{k \in S_Y} r_k^{(Y)} m_k = \hat{t}_m^{HW}$. Merk op dat in feite de regressieschatter weer wordt gebruikt, waarbij wordt gecalibreerd op de *geschatte* totalen van m_k.

Voorbeeld 3. Het voorbeeld dat we nu geven, is al eerder aan de orde geweest in subparagraaf 2.3.1. We nemen aan dat beide steekproeven S_1 en S_2 in dat voorbeeld enkelvoudig aselect zonder teruglegging (EAZT) zijn getrokken. Aan de hand van de te schatten doeltabel *opleiding ´ gezondheid* ofwel $Z´\ Y$ zullen we de stappen 2 en 3 hierboven nader toelichten. Verder nemen we nu aan dat alle tabellen standaard worden gewogen met leeftijd ofwel worden geschat met behulp van de regressieschatter waarbij *leeftijd* als hulpvariabele wordt gebruikt of eigenlijk de dummy l_k met $l_k = 1$ als persoon *k* jong is en $l_k = 0$ als dat niet zo is. Alvorens de doeltabel $Z´\ Y$ te schatten op basis van S_2 moet eerst de fractie hoogopgeleiden worden geschat op basis van het blok dat bestaat uit de vereniging van beide steekproeven, hier kortweg aangeduid met S_{12}. We nemen aan dat S_{12} eveneens kan worden gezien als een EAZT-steekproef. Uit de linkertabel op pagina 7 blijkt dat de schatting van de fractie hoogopgeleiden op basis van de desbetreffende regressieschatter gelijk is aan $\hat{\bar{z}}_{12,reg} = 0{,}50$. Volgens de splitting-up procedure zou formeel ook eerst de fractie gezonde personen moeten worden geschat op basis van S_2. Ter wille van de eenvoud laten we dat hier achterwege. Om de notatie eenvoudig te houden beperken we ons in dit voorbeeld ook alleen tot de linkerboven cel van de doeltabel $Z´\ Y$ ofwel de fractie personen met een hoge opleiding en een goede gezondheid. Definieer de dummy a_k

$$a_k = \begin{cases} 1 & \text{als persoon } k \text{ hoogopgeleid én gezond is} \\ 0 & \text{anders.} \end{cases}$$

Voor steekproef *s* (*s*=1, 2, 12) geven we de omvang aan met n_s en het steekproef-gemiddelde met $\bar{a}_s$. Het populatiegemiddelde duiden we aan met $\bar{a}_{pop}$. Omdat de populatie bestaat uit *N*=100 personen, kan $N\bar{a}_{pop}$ worden gezien als het percentage personen met een hoge opleiding en een goede gezondheid. Omdat iedere variabele slechts twee klassen heeft, neemt de regressieschatter van de fractie $\bar{a}_{pop}$ op basis van S_2 overeenkomstig stap 2 in dit geval de volgende vorm aan

$$\hat{\bar{a}}_{2,reg} = \bar{a}_2 + \hat{b}_{d,l}(\bar{l}_{pop} - \bar{l}_2)\,.$$

Uit het register is bekend dat $\bar{l}_{pop} = 0{,}3$. Uit de rechtertabel op pagina 6 blijkt dat $\bar{a}_2 = 0{,}32$. Verder is $\hat{b}_{d,l}$ de geschatte regressiecoëfficiënt in een regressie van a_k op l_k, inclusief de constante term, op basis van de data in steekproef 2 ofwel

$$\hat{b}_{d,l} = \frac{\sum_{k\in S_2}(l_k - \bar{l}_2)a_k}{\sum_{k\in S_2}(l_k - \bar{l}_2)^2}\,.$$

Merk op dat alle startgewichten $d_k^{(S_2)}$ gelijk zijn. Onder de aanname dat $\bar{l}_2 = 0{,}28$ en $\hat{b}_{d,l} = 0{,}5$ is $\hat{\bar{a}}_{2,reg}$ gelijk aan 0,33. Substitutie van de laatste uitdrukking voor $\hat{b}_{d,l}$ in de regressieschatter levert op dat de regressieschatter kan worden geschreven als een gewogen gemiddelde

$$\hat{\bar{a}}_{2,reg} = \sum_{k\in S_2} w_k a_k$$
$$w_k = 1/n_2 + (\bar{l}_{pop} - \bar{l}_2)(l_k - \bar{l}_2)/\Sigma_{k\in S_2}(l_k - \bar{l}_2)^2.$$

Merk op dat deze w_k ook van toepassing zijn op de andere cellen van de doeltabel *Z*×*Y*. Stel nu dat de data in S_2 zodanig zijn dat $\sum_{k\in S_2} w_k z_k = 0{,}52$. Om dan consistentie te bereiken met de eerder geschatte fractie hoogopgeleiden $\hat{\bar{z}}_{12,reg} = 0{,}50$ moet overeenkomstig stap 3 de geschatte tabel *Z*×*Y* worden herwogen met *opleiding*. In dit voorbeeld wordt de HW-schatter van $\bar{a}_{pop}$ dan gedefinieerd als

$$\hat{\bar{a}}_{2,HW} = \hat{\bar{a}}_{2,reg} + \hat{b}_{w,z}(0{,}50 - \hat{\bar{z}}_{2,reg})$$
$$\hat{b}_{w,z} = \frac{\sum_{k\in S_2} w_k(z_k - \hat{\bar{z}}_{2,reg})a_k}{\sum_{k\in S_2} w_k(z_k - \hat{\bar{z}}_{2,reg})^2}\,.$$

Stap 2 leverde op dat $\hat{\bar{a}}_{2,reg} = 0{,}33$. Onder de aanname dat $\hat{b}_{w,z} = 1$ en $\hat{\bar{z}}_{2,reg} = 0{,}52$ is $\hat{\bar{a}}_{2,HW}$ dan gelijk aan 0,31. Net als de regressieschatter hierboven kan de HW-schatter worden geschreven als een gewogen gemiddelde

$$\hat{\bar{a}}_{2,HW} = \sum_{k\in s_2} r_k a_k$$
$$r_k = w_k\{1 + (0{,}50 - \hat{\bar{z}}_{2,reg})(z_k - \hat{\bar{z}}_{2,reg})/\Sigma_{k\in S_2} w_k(z_k - \hat{\bar{z}}_{2,reg})^2\}.$$

Het is niet moeilijk na te gaan dat de zo verkregen gewichten r_k voldoen aan $\sum_{k\in S_2} r_k z_k = 0{,}50$ precies zoals de bedoeling was.

2.3.2.2 Minimale weegmodellen

Het voordeel van de hierboven besproken splitting-up procedure is dat de volgorde waarin de tabellen worden geschat min of meer vastligt. Dat wil zeggen, eerst moeten de (grofste) marginalen van een tabel worden geschat, maar welke van twee zulke marginalen van een tabel het eerst moet worden geschat doet er niet toe. Een alternatieve aanpak is om uit te gaan van een bepaalde vaste volgorde. De set van tabellen hoeft dan in principe ook niet meer te worden aangevuld. Wel is het verstandig om tabellen die op basis van de grote blokken kunnen worden geschat bovenaan te zetten. Dit verhoogt de nauwkeurigheid van de schattingsresultaten. Aan de hand van een voorbeeld zullen we nu uitleggen wat wordt bedoeld met minimaal wegen.

Laten *leeftijd* (*L*), *geslacht* (*G*) en *beroepsniveau* (*B*) bekend zijn uit het register. Laat het *aantal gewerkte uren* (*W*) worden waargenomen in steekproef 1 ofwel blok B_1. Laat *opleiding* ($O^{(2)}$) zijn waargenomen in steekproef 2 ofwel blok B_2. Verder nemen we aan dat het gebruikte weegmodel bij de regressieschatter voor alle tabellen gelijk is aan $X = L \times G$. In Tabel 1 staat een overzicht van vijf tabellen die achtereenvolgens moeten worden geschat. Ook is aangegeven welke tabellen opnieuw (herhaald) moeten worden gewogen en wat hun herweegschema (-model) is. Dit herhaald wegen gebeurt weer precies zoals is beschreven in de vorige subparagraaf. Het zal duidelijk zijn dat tabellen over *gewerkte uren* × *opleiding* alleen kunnen worden geschat op basis van de overlap van de steekproeven 1 en 2 ofwel blok B_3.

Tabel 1. Doeltabellen met hun herweegschema's

tabellen	blok	minimaal herweegschema $M\{T_k\}$
$T_1 = L \times W$	B_1	nvt
$T_2 = L \times G \times O^{(2)}$	B_2	nvt
$T_3 = B \times O^{(1)}$	B_2	$B + O^{(1)}$
$T_4 = W \times O^{(1)}$	B_3	$W + O^{(1)}$
$T_5 = L \times W \times O^{(1)}$	B_3	$L \times W + L \times O^{(1)} + W \times O^{(1)}$

In principe zijn alle doeltabellen van de vorm

$$T_k = L^{(r_{k1})} \times G^{(r_{k2})} \times B^{(r_{k3})} \times W^{(r_{k4})} \times O^{(r_{k5})} = T(r_k) \qquad (k=1,\ldots,5).$$

Zoals al eerder is opgemerkt wordt iedere tabel volledig gekarakteriseerd door de bijbehorende classificatievector r_k. Voor een willekeurige tabel T_k (k=1,…,5) uit de geordende tabellenset is het minimale herweegschema gelijk aan

$$M\{T_k\} = T_{k0} + T_{k1} + \ldots + T_{k,k-1}.$$

Hierbij is T_{kh} (h=0,…,k-1) gedefinieerd als

$$T_{kh} = T[\min(r_k, r_h)]$$
$$[\min(r_k, r_h)]_j = \min(r_{kj}, r_{hj}) \qquad (j = 1,...,5).$$

Verder staat T_0 voor het register dat in deze context moet worden gezien als een grote tabel. De eerste twee tabellen hoeven niet te worden herwogen want ze zijn al consistent met het register omdat beide tabellen reeds zijn gewogen met leeftijd maal geslacht. Anders gezegd, uit de categoriale variabele *leeftijd× geslacht* zijn de hulpvariabelen voor de gebruikte regressieschatter afgeleid. Ook wordt het *aantal gewerkte uren* niet waargenomen buiten blok 1 en *opleiding* niet buiten blok 2. Verder is bij T_5 in het herweegschema T_{50} =*L* weggelaten omdat *L* al is opgenomen in het herweegschema via *L*×*W*. Hetzelfde geldt voor $T_{53} = O^{(1)}$.

Merk ook op dat wanneer B_3 de enige steekproef was, herhaald wegen niet nodig zou zijn geweest behalve voor T_3. Immers wanneer alle tabellen worden geschat op basis van B_3 met behulp van de regressieschatter met de hulpvariabelen uit het weegmodel $X = L \times G,$ zijn de vijf aldus geschatte tabellen reeds onderling consistent. Qua leeftijd en geslacht zijn ze ook consistent met het register. Alleen T_3 moet worden herwogen omdat *B* niet in het overall weegmodel zit.

Soms is het aan te bevelen om de gegeven set van te schatten tabellen toch nog uit te breiden met andere tabellen. Dit doet zich voor wanneer een marginaal, zeg T_m, van de kruistabel T_k kan worden geschat op basis van een groter blok dan T_k. In dat geval is het verstandig T_m toe te voegen aan de tabellenset zodanig dat hij voor T_k wordt geschat. Hetzelfde geldt wanneer T_k niet consistent kan worden geschat met de gewone regressieschatter terwijl een bepaalde marginaal van die tabel op basis van *hetzelfde* blok wel consistent kan worden geschat. Ook in die situatie dient een dergelijke marginaal aan de set te worden toegevoegd wanneer dat nog niet is gebeurd en te worden geschat voor T_k.

Als bezwaar tegen minimaal (her)wegen wordt wel naar voren gebracht dat de uitkomsten van herhaald wegen op deze manier van de volgorde van de doeltabellen afhangen. Op zich is deze opmerking correct. Zolang de verschillen tussen de verschillende uitkomsten blijven binnen de onzekerheidsmarges van de oorspronkelijke schattingen is er weinig aan de hand. Een andere steekproef had ook tot andere uitkomsten geleid. Dergelijke statistische verschillen zijn inherent aan het schattingsproces.

2.3.2.3 Tabellen met kwantitatieve variabelen

In de vorige paragrafen hebben we het steeds gehad over het herwegen van frequentietabellen. Enigszins formeel kunnen frequentietabellen worden geschreven als *Y*=[*F*]×1. Hierbij staat *F* voor een (multiple) categoriale variabele van de vorm *A*×*B*×…, terwijl de "1" aangeeft dat het gaat om het tellen van de populatie-eenheden die horen bij de desbetreffende cellen van tabel *Y*. Soms wordt de "1" ook wel weggelaten en schrijven we kortweg *Y*=*F*. In deze subparagraaf zullen we in het kort ingaan op het herwegen van een tabel met de kwantitatieve inkomensvariabele,

zeg Z_{kw}. Een dergelijke tabel duiden we aan met Y=[F]×Z_{kw}. Met andere woorden, in een cel van een dergelijke tabel staat het (geschatte) inkomenstotaal van alle populatie-eenheden in die cel.

Een belangrijk begrip bij het herwegen van tabellen met de kwantitatieve inkomensvariabele is de zogenaamde gekwantiseerde versie van een frequentietabel. Dit is vooral van belang wanneer er doeltabellen zijn met categoriale variabelen met inkomensklassen, zeg $Z_{cat}^{(1)},...,Z_{cat}^{(\max)}$. Aan de hand van een voorbeeld leggen we uit wat we verstaan onder de gekwantiseerde versie van een frequentietabel. Bekijk de volgende (fictieve) 2×2 frequentietabel van *inkomen* × *leeftijd* ofwel $Z_{cat}^{(1)} \times L^{(1)}$

Tabel 2. Frequentietabel van *inkomen* × *leeftijd* (in mln)

	jong	oud
laag inkomen	5	1,5
hoog inkomen	2	1,5

Stel nu dat het gemiddelde inkomen in de groep met lage inkomens grofweg is geschat op 20.000 Euro en voor de groep met hoge inkomens op 50.000 Euro. Het totale inkomen van de jongeren en de ouderen kan dan worden geschat op basis van de cijfers in Tabel 2. De aldus geschatte inkomenstotalen voor *jong* en *oud* staan in Tabel 3.

Tabel 3. Gekwantiseerde versie van $Z_{cat}^{(1)} \times L^{(1)}$ (in mld Euro's)

	jong	oud
inkomen	200	105

Dit wordt de gekwantiseerde versie van Tabel 2 genoemd en wordt ook wel aangeduid met $K(Z_{cat}^{(1)} \times L^{(1)})$. Het voordeel van de gekwantiseerde Tabel 3 ten opzichte van de oorspronkelijke frequentietabel is dat het aantal weegvariabelen op deze manier kan worden gereduceerd van 4 tot 2. Dit is vooral van belang wanneer er niet voldoende waarnemingen zijn bij veel multidimensionale tabellen. Bij meer dan twee inkomensklassen wordt ook wel simpelweg het klassemidden gebruikt als schatting voor het gemiddelde inkomen in de desbetreffende inkomensklasse. Het gaat er om dat er een representatieve waarde wordt gekozen.

In het vervolg noteren we het weegschema van een doeltabel Y in deze subparagraaf als $W\{Y\}$. Het minimale weegmodel uit de vorige subparagraaf geven we aan met $M\{.\}$. In feite kan een tabel van de vorm Y=[F]×Z_{kw} ook weer worden gekarakteriseerd door een vector r waarvan het laatste element de waarde 1 aanneemt als de inkomensvariabele Z_{kw} in de tabel voorkomt en 0 als dat niet zo is. Merk op dat wanneer het inkomenstotaal kan worden geschat op basis van een groter blok, de tabel [1]× Z_{kw} aan de tabellenset moet worden toegevoegd en moet worden geschat voor Y; zie vorige subparagraaf. Hierbij is [1]× Z_{kw} gebruikt als een als een wat formele aanduiding voor de nuldimensionale tabel *totaal inkomen*.

Hieronder geven we enkele resultaten met betrekking tot het herwegen van kwantitatieve tabellen:

1. Indien er geen categoriale inkomensvariabelen in de doeltabellen voorkomen, is het weegschema voor $Y=[F]\times Z_{kw}$ gelijk aan $W\{Y\} = M\{Y\} + F$. De uitbreiding van het weegmodel met F heeft het voordeel dat men later ook consistent het gemiddelde inkomen kan bepalen waarbij men deelt door het bijbehorende aantal personen per categorie.

2. Indien er van de kwantitatieve inkomensvariabele Z_{kw} ook categoriale versies $Z_{cat}^{(1)},...,Z_{cat}^{(\max)}$ voorkomen in de set te schatten doeltabellen, wordt de zaak wat ingewikkelder. Het herweegschema van de frequentietabel $Y=F\times Z_{cat}^{(k)}$ (k=1,…,max) is bijvoorbeeld gelijk aan

$$W\{Y\}=M\{Y\}+M\{K(F\times Z_{cat}^{(k)})\}.$$

Nadat de frequentietabel $F\times Z_{cat}^{(k)}$ is geschat met herhaald wegen, moet met dezelfde gewichten ook de gekwantiseerde versie $K(F\times Z_{cat}^{(k)})$ worden berekend, alsmede alle 'lagere' gekwantiseerde versies $K(F\times Z_{cat}^{(l)})$ met l =1 .... k-1. De aldus verkregen gekwantiseerde versies zijn allemaal onderling consistent.

3. Opnieuw veronderstellen we dat er van de kwantitatieve inkomensvariabele Z_{kw} categoriale versies $Z_{cat}^{(1)},...,Z_{cat}^{(\max)}$ bestaan. Bij herweging van de tabel $Y=[F]\times Z_{kw}$ is het herweegschema gelijk aan

$$W\{Y\}= F +M\{Y\} + K(M\{F\times Z_{cat}^{(\max)}\}).$$

Met andere woorden, als het minimale weegmodel van $F\times Z_{cat}^{(\max)}$ gelijk is aan $A\times Z_{cat}^{(1)} + B\times Z_{cat}^{(2)}$ dan moet nog de volgende component aan het weegschema moet worden toegevoegd

$$K(A\times Z_{cat}^{(1)}) + K(B\times Z_{cat}^{(2)}).$$

Voorbeeld. Stel dat de volgende doeltabellen moeten worden geschat

$$T_1=[L\times Z_{cat}^{(1)}]\times 1=[leeftijd\times inkomen^{(1)}]\times 1$$
$$T_2=[L\times R]\times Z_{kw}=[leeftijd\times regio]\times inkomen.$$

De tabel $L\times R$ is bekend uit het register (T_0). Bovendien is gegeven dat het totale inkomen 300 mld bedraagt. Nadat T_1 is geschat eventueel met behulp van herwegen, moet ook de gekwantiseerde versie $K(L\times Z_{cat}^{(1)})$ hiervan worden geschat met dezelfde gewichten als T_1. Dan is tabel T_2 aan de beurt. Volgens punt 3 hierboven bestaat het weegschema van T_2 uit drie componenten. De eerste component bestaat uit de bijbehorende frequentietabel $L\times R$. De tweede component bestaat uit het minimale

weegschema van T_2 zoals hierboven besproken. Dit geeft als resultaat het reeds bekende inkomenstotaal ofwel $[1]\times Z_{kw}$ (NB: L en $L\times R$ zitten reeds in de eerste component). Ten slotte is de derde component gelijk aan de gekwantiseerde versie van het minimale weegmodel van $L\times R\times Z_{cat}^{(1)}$. Het minimale weegmodel van de laatste tabel levert als extra bijdrage alleen nog op $L\times Z_{cat}^{(1)}$. De gekwantiseerde versie van de laatste tabel wordt gegeven door Tabel 3 aan het begin van deze subparagraaf en is berekend nadat T_1 was geschat. Tabel T_2 moet derhalve herwogen of gekalibreerd worden op *leeftijd*×*regio* uit het register, het inkomenstotaal van 300 mld en op Tabel 3.

Voor meer details zie Renssen et al. (2001) en Houbiers en Renssen (2001). Schatten van dergelijke tabellen met herhaald wegen is op grote schaal uitgevoerd voor het Loonstructuuronderzoek 2002, zie Gouweleeuw (2005).

2.3.3 Slotopmerkingen

Tot nu toe is gesproken over tabellen waarvan de cijfers geen onderlinge verschillen mogen laten zien als het over dezelfde variabele gaat. Een andere vorm van consistentie heeft te maken met het bestaan van zogeheten editregels. Bijvoorbeeld het aantal personen met een rijbewijs kan niet meer zijn dan het aantal personen boven de 18 jaar. Naast deze if-then relaties (op microniveau) kan het ook nog voorkomen dat kwantitatieve variabelen aan een bepaalde relatie moeten voldoen zoals winst is gelijk aan omzet minus kosten. Een dergelijke relatie is van toepassing op ieder bedrijf afzonderlijk maar moet ook gelden voor de geschatte aggregaten. Daas en Renssen (2001) en Van de Laar (2004) besteden uitgebreid aandacht aan het schatten van tabellen met dit soort editregels.

Wanneer er onderlinge relaties tussen de (geschatte) variabelen bestaan, kunnen ook macro-integratietechnieken worden gebruikt. Voor een uitgebreide beschrijving van deze aanpak en de relatie met Kalmanvergelijkingen zie Knottnerus (2003, Hfdst 12) en Sefton en Weale (1995). Boonstra (2004b) werkt een en ander verder uit voor het calibreren van tabelschattingen.

In dit document wordt niet ingegaan op de variantie van de HW-schatter. In principe kan onder bepaalde voorwaarden de variantie van de HW-schatter worden geschat. Er bestaat een nauwe relatie tussen de variantie van de HW-schatter en die van de regressieschatter. Uit herhaald toepassen van (2.4) volgt ook dat de HW-schatter altijd is te schrijven als een lineaire combinatie van regressieschatters. Hierbij is wel aangenomen dat er voldoende waarnemingen (n>>1) zijn voor iedere cel van de tabel zodat de stochastiek in de geschatte regressiecoëfficiënten kan worden verwaarloosd. Wanneer dit niet het geval is, neemt de variantie van de HW-schatter toe net als bij de regressieschatter. Dit verschijnsel doet zich vooral voor wanneer de tabellen te gedetailleerd worden met teveel categoriale variabelen in verhouding tot de omvang van de blokken (steekproeven). Voor een uitgebreide beschrijving van het schatten van de variantie van de HW-schatter zie Boonstra et al. (2003) en Knottnerus en van Duin (2006). Voor verschillende simulatiestudies naar de

performance van de HW-schatter zie Boonstra (2004a) en Van Duin en Snijders (2003).

2.4 Toepassen van herhaald wegen bij de Volkstelling 2001

Herhaald wegen is toegepast in de Virtuele volkstelling van 2001 (VT2001). Deze volkstelling omvatte 40 tabellen die numeriek consistent aan Eurostat moesten worden geleverd. Het CBS heeft deze tabellen volledig gevuld op basis van register en enquêtedata die reeds beschikbaar was. Dit maakte de VT relatief goedkoop. Daarnaast was de doorlooptijd relatief kort: hoewel Nederland pas laat is gestart, heeft het toch als een van de eerste landen de tabellen aan Eurostat kunnen leveren. Doordat er gebruik is gemaakt van herhaald wegen, zijn de tabellen bovendien numeriek consistent geschat.[1] Er is een softwarepakket ontwikkeld (VRD = Vullen Reference Database) om het herhaald wegen te automatiseren. In deze subparagraaf zal nader worden ingegaan op de verschillende stappen (zoals beschreven in paragraaf 2.3.1) die zijn doorlopen bij het uitvoeren van herhaald wegen voor de VT2001.

2.4.1 Tabellen, variabelen en datablokken

Allereerst moeten de te schatten tabellen worden gespecificeerd. In het geval van de volkstelling was deze set door Eurostat voorgeschreven. De telvariabele was hier in alle gevallen het aantal personen of het aantal huishoudens.

Wanneer de variabelen volledig gespecificeerd zijn, moet er worden onderzocht uit welke databron (register of enquête) deze kunnen worden afgeleid. Merk op dat verschillende niveaus van één variabele niet noodzakelijk uit dezelfde databron hoeven worden afgeleid. De databronnen die bij de VT zijn gebruikt zijn: de gemeentelijke basisadministratie (GBA, register) voor demografische variabelen zoals geslacht, leeftijd, geboorteland, nationaliteit; het Sociaal Statistisch bestand (SSB, register) om vast te stellen of iemand al dan niet werkzaam is; de Enquête Werkgelegenheid en Lonen (EWL, steekproef) voor de arbeidsduur; de Enquête Beroepsbevolking (EBB, steekproef) voor de variabelen opleiding, beroep en economische status.

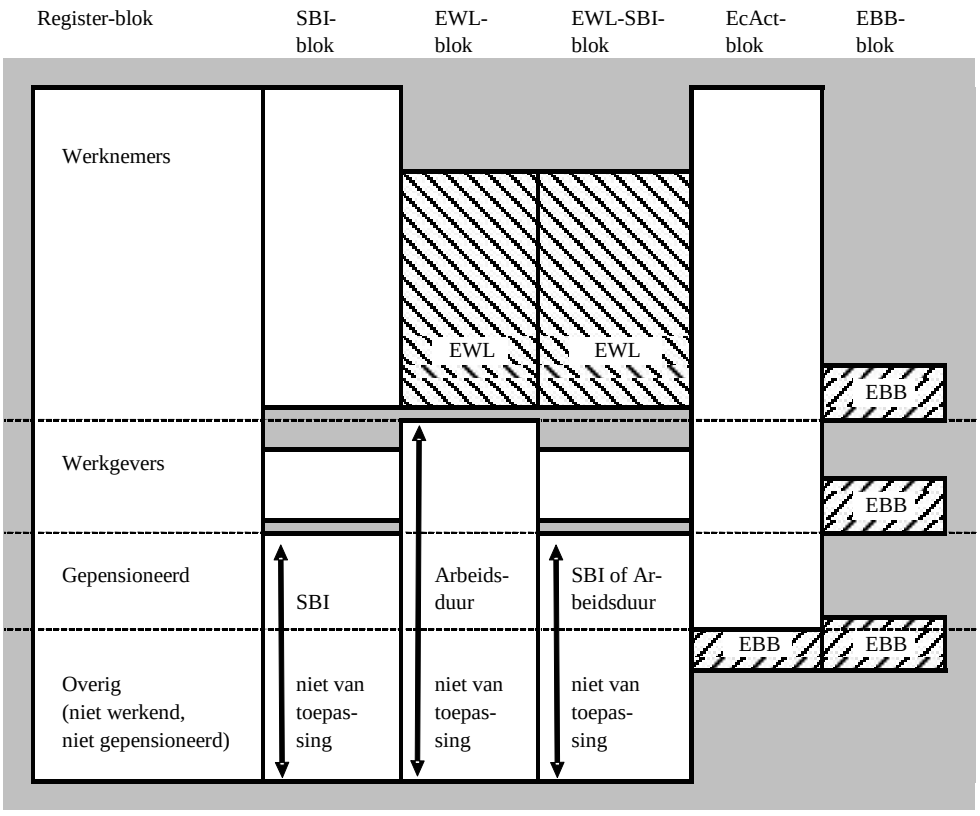
Wanneer al deze databronnen aan elkaar zijn gekoppeld, kunnen de rechthoekige datablokken worden afgeleid. Deze zijn in Figuur 3 weergegeven. Het eerste blok is het registerblok. De variabelen in dit blok (zoals geslacht en leeftijd op alle niveaus) zijn voor alle personen in de populatie bekend. Het tweede blok is het SBI-blok. Dit bestaat uit de variabelen in het registerblok, aangevuld met de variabele SBI. (Het betreft dan de SBI van de hoofdbaan van de betreffende persoon.) Het blok bevat records voor bijna alle personen in de populatie (met uitzondering van degenen van wie de SBI niet bekend is). Dit blok wordt gebruikt voor de schatting van de tabellen waarin de variabele SBI uitsluitend wordt gecombineerd met registervariabelen. Het

[1] Dit betreft overigens niet de tabellen die over wonen handelen. Deze zijn in een apart project geschat.

volgend blok is het EWL-blok. Dit blok bevat alleen records voor personen die in de EWL (een grote steekproef) voorkomen en bestaat uit de variabelen van het registerblok aangevuld met de variabele arbeidsduur. Het vierde blok is het EWL-SBI-blok. Dit bevat records voor alle personen die in het SBI-blok en het EWL-blok voorkomen (de doorsnede van beide blokken). Dit blok is nodig om de tabel te schatten waarin de variabele SBI wordt gekruist met arbeidsduur. Het vijfde blok is het EcActblok. Voor personen die tot de werknemers, werkgevers en gepensioneerden behoren komt dit overeen met het register, voor de overige zijn alleen personen die in de EBB voorkomen opgenomen. Dit blok kan worden gebruikt om de vele tabellen waarin register variabelen worden gekruist met al dan niet economisch actief zijn (d.w.z. werkend of werkloos) zo nauwkeurig mogelijk te schatten. Het laatste blok is het EBB-blok. Dit bevat records voor alle personen die in de EBB voorkomen, en wordt gebruikt voor tabellen waarin opleiding, beroep of economische status wordt gekruist met registervariabelen.

Figuur 3.

Schema datablokken VT2001; objecttype personen, 15-74 jaar



2.4.2 Startgewichten

Wanneer de datablokken zijn gedefinieerd en afgeleid, moeten er voor ieder datablok startgewichten voor het herhaald wegen worden bepaald. Allereerst worden er per datablok initiële blokgewichten bepaald, en uit deze worden vervolgens de definitieve blokgewichten, of startgewichten bepaald. Voor het registerblok is dit eenvoudig. Dit blok bevat records voor alle personen uit de populatie. Iedere persoon krijgt hierin dus een startgewicht gelijk aan 1.

De overige blokken zijn opgebouwd uit registers in combinatie met één of meer steekproeven. Iedere steekproef is op zichzelf al voorzien van een (of meer) sets gewichten. De startgewichten voor herhaald wegen kunnen nu per blok in drie stappen worden bepaald:

1. Het bepalen van de gewichten per steekproef (binnen een blok)
2. Het combineren van de gewichtensets uit stap 1 tot blokgewichten. (Wanneer een blok slechts één enquête omvat, kan stap 2 achterwege blijven.)
3. Het herwegen van de gewichten uit stap 2 naar bekende populatietotalen.

Binnen stap 1 is het meest eenvoudige om voor ieder steekproefonderzoek de kant-en-klare publicatiegewichten van het betreffende onderzoek te nemen. Deze gewichten zijn meteen beschikbaar. Bovendien zijn zij gebaseerd op uitgebreide analyses en veel inhoudelijke kennis van zaken. Stap 2 is van belang voor het EcAct-blok en het EBB-blok. Deze blokken bestaan namelijk uit twee steekproeven (namelijk twee jaargangen van de EBB). Laat $w_k^{(1)}$, $k = 1,\ldots,n_1$ de gewichten uit stap 1 voor het eerste onderzoek zijn, en evenzo, laat $w_k^{(2)}$, $k = 1,\ldots,n_2$ de gewichten uit stap 1 voor het tweede onderzoek zijn.

De gewichten die in stap 1 zijn bepaald, tellen ongeveer op tot het populatietotaal. Wanneer het blok uit twee steekproeven bestaat, zullen alle gewichten bij elkaar optellen tot twee maal de populatieomvang. Het is natuurlijk mogelijk om de gewichten uit het blok te vermenigvuldigen met 1/2, om op deze manier gewichten te krijgen die tot de juiste populatieomvang optellen. Het is echter beter om de gewichten te vermenigvuldigen met een factor die rekening houdt met de relatieve omvang van de steekproeven. Dit zorgt ervoor dat de gewichten in het uiteindelijke steekproefblok van vergelijkbare grootte zijn. De gewichten na stap 1 worden dan vermenigvuldigd met:

$$\frac{N}{\sum_{k=1}^{n_1} w_k^{(1)}} \frac{n_1}{n_1+n_2}, \text{ resp. } \frac{N}{\sum_{k=1}^{n_2} w_k^{(2)}} \frac{n_2}{n_1+n_2}.$$

Ten slotte kan het handig zijn om de blokgewichten te herwegen naar een aantal populatietotalen. Dit kan ervoor zorgen dat, wanneer er veel tabellen moeten worden geschat, er een aantal met de startgewichten consistent wordt geschat. Hoe dit moet worden gekozen hangt van het onderzoek en de te schatten tabellen af.

2.4.3 Het schatten van de tabellen

Wanneer de tabellen en variabelen gedefinieerd zijn, de steekproefblokken bepaald en afgebakend en voorzien van startgewichten kan het schatten van de tabellen beginnen. Zoals al gezegd kan dit worden uitgevoerd met behulp van het software-pakket VRD. Allereerst moet er worden besloten of er wordt gekozen voor de minimale weegprocedure of voor de splitting-up procedure. Dit is afhankelijk van het onderzoek. Als het aantal tabellen dat moet worden geschat erg groot wordt, zal de splitting-up procedure tot nog meer tabellen leiden, en dat kan op een bepaald moment tot schattingsproblemen leiden. Splitting up is dan niet aan te raden, zeker

niet als van tevoren exact vaststaat welke tabellen er moeten worden geschat. Dit was bijvoorbeeld bij de volkstelling het geval.

Vervolgens moet de volgorde worden gekozen waarin de tabellen worden geschat. Uiteraard worden eerst de tabellen bepaald, die volledig uit het registerblok kunnen worden afgeleid. Daarna verdient het de aanbeveling om te beginnen met de tabellen die uit het grootste blok kunnen worden geschat, daarna de tabellen uit het op één na grootste blok, enzovoorts. Per blok moeten de tabellen die met behulp van de startgewichten consistent kunnen worden geschat het eerst worden geschat.

2.4.4 Praktijkproblemen en mogelijke oplossingen

In de praktijk kan men bij het toepassen van herhaald wegen tegen een aantal problemen aanlopen. Allereerst is daar het probleem van steekproefnullen. Dit betekent dat een bepaalde categorie in de populatie wel voorkomt, terwijl deze in de steekproef ontbreekt. Een tabel die last heeft van steekproefnullen kan onmogelijk consistent worden geschat. Alle problemen die te maken hebben met steekproefnullen kunnen echter voordat er wordt geschat al worden geïdentificeerd. Voor iedere tabel kan worden nagegaan of alle categorieën die in de populatie voorkomen ook in de steekproef zijn vertegenwoordigd. Als dit niet het geval is, zullen er cellen moeten worden samengevoegd. Dit kan betekenen dat er voor bepaalde variabelen een extra niveau moet worden aangemaakt. Daarom is het handig om dit van tevoren te doen. Een andere mogelijke oplossing is om een dergelijke tabel niet in volledig detail, maar alleen marginalen te schatten. In het algemeen kan worden gezegd dat eigenschappen die zeldzaam zijn in de populatie dit ook zullen zijn in de steekproef en dus gemakkelijk kunnen leiden tot steekproefnullen. Denk hierbij bijvoorbeeld aan werkzame personen ouder dan 60 jaar (uitgesplitst naar leeftijd en nationaliteit).

Een tweede probleem dat kan voorkomen, wordt veroorzaakt door editregels tussen verschillende variabelen. Het proces van herhaald wegen houdt hier niet automatisch rekening mee. Dit kan tot inconsistenties tussen verschillende tabellen leiden, hetgeen niet wenselijk is. Denk bijvoorbeeld aan de variabele economische status (die aangeeft of een persoon werkzaam, werkloos, gepensioneerd of iets anders is) en de variabele beroep. Alleen een werkzame persoon heeft een beroep, maar hier wordt niet automatisch rekening mee gehouden. Dit kan worden opgelost door tabellen naar beroep te kruisen met economische status. Een betere oplossing is in dit geval om een extra niveau in de variabele beroep op te nemen, waarin wordt uitgesplitst naar werkzaam en niet werkzaam.

2.5 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft de HW-schatter van een bepaalde tabel zijn:

- de omvang van de non-respons;
- de onzekerheidsmarges van de HW-schatters van de totalen in de tabel;
- de celvulling ofwel het aantal waarnemingen per cel;

- het verschil tussen de startgewichten en de uiteindelijke gewichten na het herhaald wegen van de tabel.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Vaak kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele. Zie ook het thema *Wegen als correctie voor non-respons*.

Zoals al eerder is opgemerkt, kunnen onder bepaalde voorwaarden de variantie en de onzekerheidsmarge van de HW-schatter worden berekend, zie bijvoorbeeld Knottnerus (2001) en Snijders en Houbiers (2002). Net als bij de regressieschatter zijn de voorwaarden hiervoor dat er voldoende celvulling moet zijn. De variantieformules zijn echter wel complex en het kost vrij veel rekentijd om met VRD varianties te schatten.

Om zonder varianties desondanks een uitspraak over de nauwkeurigheid van tabelschattingen te kunnen doen, kan worden gekeken naar het aantal waarnemingen dat bijdraagt aan een cel in een tabel. De relatieve onnauwkeurigheid van de schatting van een tabelcel hangt samen met het aantal waarnemingen in die cel, ook bij herhaald wegen, zie Van Duin en Snijders (2003). De relatieve onnauwkeurigheid van een cel in de tabel wordt gedefinieerd als de (geschatte) standaardfout gedeeld door het (geschatte) aantal elementen in die cel, dus als:

$$\text{Relatieve onnauwkeurigheid}_{\text{cel}} = \frac{\hat{\sigma}_{\text{cel}}}{\hat{N}_{\text{cel}}}.$$

Hoe kleiner de relatieve onnauwkeurigheid, hoe betrouwbaarder een cel is geschat. Cellen met een te grote onnauwkeurigheid mogen niet gepubliceerd worden. In plaats van de echte herhaald-wegenvarianties te schatten, wordt de standaardfout (grof) benaderd met de formule

$$\hat{\sigma}_{\text{cel}} \approx N\sqrt{\hat{p}_{\text{cel}}(1-\hat{p}_{\text{cel}})/n}\sqrt{1-f}\ ,$$

waarin N de populatieomvang, n de steekproefomvang en $f = n / N$ de steekproeffractie is, en $\hat{p}_{\text{cel}} = \hat{N}_{\text{cel}} / N$ de geschatte relatieve celfrequentie. Deze formule vormt een redelijke benadering van de herhaald-wegenstandaardfout onder de aanname dat het effect van de ongelijke insluitkansen in het steekproefblok op de schattingen verwaarloosbaar is. Bovendien is in bovenstaande formule het variantiereducerende effect van de calibraties in het herhaald toepassen van de regressieschatter buiten beschouwing gelaten. Het effect hiervan op de standaardfout bedraagt een factor

$$\sqrt{1-R^2} \le 1,$$

waarin R^2 de R-kwadraat van de regressie is, zie Knottnerus (2003). Omdat er bij het herhaald wegen sprake is van regressie op regressie op regressie, is deze factor niet

goed te kwantificeren. Over het algemeen geldt dat herhaald wegen leidt tot lagere varianties dan gewoon wegen, zie Van Duin en Snijders (2003) en Boonstra (2004a). Onder de aanname dat in de onderhavige toepassing het variantieverhogende effect van de ongelijke insluitkansen wordt gecompenseerd door het variantiereducerende effect van het herhaald wegen vormt de eerder genoemde formule een redelijke benadering voor de echte herhaald-wegenstandaardfout. Bij een bovengrens A voor de relatieve onnauwkeurigheid volgt dat het geschatte celtotaal moet voldoen aan

$$\hat{N}_{\text{cel}} \geq \frac{N}{1+[nA^2/(1-f)]}.$$

Uitgaand van een gemiddeld ophooggewicht van N / n volgt dat het aantal waarnemingen in de cel moet voldoen aan

$$n_{\text{cel}} \cong n\frac{\hat{N}_{\text{cel}}}{N} \geq \frac{n}{1+[nA^2/(1-f)]}.$$

In het algemeen zal de steekproeffractie f klein zijn en worden verwaarloosd. Bovendien geldt $nA^2 >> 1$ voor redelijke waardes van A, zodat het aantal waarnemingen in de cel ten minste gelijk moet zijn aan

$$n_{\text{cel}} \approx \frac{1}{A^2}$$

voor een schatting met een maximale relatieve onnauwkeurigheid A.

Zo is bijvoorbeeld bij de eerder beschreven Volkstelling uitgegaan van een maximale relatieve onnauwkeurigheid van 20 procent (A = 0.2). Hiervoor zijn ten minste 25 steekproefwaarnemingen nodig. Cellen met minder waarnemingen mogen dan in principe niet worden gepubliceerd.

Een laatste kwaliteitsindicator en een aanwijzing voor mogelijke instabiliteit van de toegepaste HW-procedure is vaak al dat de gewichten veel gaan afwijken van de startgewichten. Dit houdt in dat de HW-procedure niet langer leidt tot alleen wat cosmetische tabelaanpassingen, maar dat hij ook meer substantiële wijzigingen tot gevolg kan hebben.

3. Literatuur

Advokaat, W., Cruchten, J. van, Gouweleeuw, J., Harmsen, C., Hartgers, M., Houbiers, M., Linder, F., Oroh, H. en Schulte Nordholt, E. (2004), *VT – 2001, documentatie van het proces*. CBS, Voorburg.

Boonstra H.J.H. (2004a), *DACSEIS deliverable 7.3: A simulation study of repeated weighting estimation*. CBS, Heerlen.

Boonstra H.J.H. (2004b), *Calibration of tables of estimates*. Research Paper, CBS, Heerlen.

Boonstra, H.J.H., Brakel, J.A. van den, Knottnerus, P., Nieuwenbroek, N.J. en Renssen, R.H. (2003), *DACSEIS deliverable 7.2: A strategy to obtain consistency among tables of survey estimates*. CBS, Heerlen.

Daas, P.J.H. en Renssen, R.H. (2001), *On the use of prior knowledge in estimates: if-then edit rules*. Research Paper, CBS, Heerlen.

Deville, J.C. en Särndal, C.E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Gouweleeuw, J.M. (2005), *Loonstructuuronderzoek 2002 nationaal: Schattingsproces, marges en samenstellen van tabellen*. CBS, Voorburg.

Houbiers, M. (2004), Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics*, 20, 55–75.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H. en Snijders, V. (2003), *Estimating consistent table sets: position paper on repeated weighting*. Discussion Paper, CBS, Voorburg.

Houbiers, M. en Renssen, R.H. (2001), *Kwantitatieve variabelen en herhaald wegen*. CBS, Voorburg.

Knottnerus, P. (2001), *Varianties bij herhaald wegen*. CBS, Voorburg.

Knottnerus, P. (2003), *Sample survey theory: some Pythagorean perspectives*. Springer-Verlag, New York.

Knottnerus, P. en Duin, C. van (2006), Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565–584.

Kroese, A.H. en Renssen, R.H. (1999), Weighting and imputation at Statistics Netherlands. In: *Proceedings of the IASS Conference on Small Area Estimation*, Riga, 109–120.

Renssen, R.H., Kroese, A.H. en Willeboordse, A. (2001), *Aligning estimates by repeated weighting*. CBS, Heerlen.

Renssen, R.H. en Martinus, G.H. (2002), On the use of the generalized inverse in sampling theory. *Survey Methodology*, 28, 209–212.

Särndal, C.E, Swensson, B. en Wretman, J.H. (1992), *Model assisted survey sampling*. Springer-Verlag, New York.

Schulte Nordholt, E. (2005), The Dutch virtual Census 2001: A new approach by combining different sources. *Statistical Journal United Nations ECE*, 22, 25–37.

Snijders, V. en Houbiers, M. (2002), *Variantieschatting bij herhaald wegen voor VRD 1.4*. CBS, Voorburg.

Sefton, J. en Weale, M. (1995), *Reconciliation of national income and expenditure*. Cambridge University Press, UK.

Van Duin, C. en Snijders, V. (2003), *Simulation studies of repeated weighting*. Discussion paper, CBS, Voorburg.

Van de Laar, R.W.A. (2004), *Edit rules and the strategy of consistent table estimation.* Discussion paper, CBS, Voorburg.