# CPB Memorandum

Centraal Planbureau

| | | |
|---|---|---|
| Sector | : | 2 |
| Afdeling/Project | : | Deelproject Herweging |
| Samensteller(s) | : | R.J. Waaijers |
| Nummer | : | 146 |
| Datum | : | 28 februari 2006 |

## Herwegingsprocedure bij het op IPO gebaseerde microsimulatiemodel

Het herwegen van historische microdata naar toekomstige jaren waarvoor alleen (ramingen van) macro totalen van persoons- en huishoudkenmerken bekend zijn, is een beproefde en aantrekkelijke vorm van statische microsimulatie (*static aging*).
Dit memorandum beschrijft hoe de methode van *lineair wegen* is toegepast op het IPO-microgegevensbestand van het CBS, als bijdrage aan de ontwikkeling van een CPB-microsimulatiemodel voor sociale zekerheid, inkomsten- en loonheffing en koopkracht.
Daarbij is het bestand voor ieder jaar van de periode 2002-2011 herwogen naar een uitgebreide set indicatoren. Om te garanderen dat hierbij uitsluitend positieve gewichten resulteren wordt gebruik gemaakt van een iteratief procédé van Huang en Fuller (HF) voor begrensd wegen.
Voor de achtereenvolgens uit te voeren herwegingsoperaties wordt de voorkeur gegeven aan *recursieve* weging (op basis van het vorige herwegingsresultaat in plaats van het oorspronkelijke IPO-gewicht). Door een specifieke toepassing van het HF-algoritme wordt gegarandeerd dat de afwijking tussen elke afzonderlijke herwegingsset en de oorspronkelijke IPO-gewichtenset ook bij recursieve weging begrensd blijft, ten koste van een (geringe) toename van de benodigde rekentijd.

# Inhoud

# 1 Inleiding

Het project Implementatie IPO is er op gericht een microsimulatiesysteem te ontwikkelen op basis van het microbestand van het Inkomens Panel Onderzoek (IPO) voor gebruik door de afdelingen IEP en SZ. Het Inkomens Panel Onderzoek is een steekproefonderzoek van het CBS dat al dateert van 1977 met gegevens op basis van een aantal administraties waaronder de belastingaangifte. Het IPO is relatief groot opgezet: het microbestand omvat voor het jaar 2002 circa 240 000 persoonsrecords. Om tot macrototalen voor de gehele Nederlandse bevolking te komen dienen de variabelen in het bestand te worden gewogen met het in ieder record aanwezige gewicht. Deze gewichten zijn zodanig bepaald dat het aantal personen voor een combinatie van leeftijd, geslacht, burgerlijke staat, regio en soort adres aansluit op de gegevens van de Bevolkingsstatistiek en dat het aantal particuliere huishoudens naar omvang, regio en leeftijd van het hoofd overeenkomt met de uitkomsten van de Huishoudensstatistiek.

Het Deelproject Herweging heeft als doelstelling om voor een aantal persoons- en huishoudenskenmerken binnen het IPO te bewerkstelligen dat wordt aangesloten bij macrototalen die afkomstig zijn van het CBS en van andere afdelingen binnen het CPB. Daartoe dienen de in het IPO aanwezige gewichten te worden gecorrigeerd middels herweging. In dit memorandum wordt de daarbij gehanteerde methode beschreven, evenals de wijze waarop deze in SAS is geïmplementeerd. Bovendien wordt een succesvolle toepassing beschreven binnen een realistisch experimenteel kader voor de periode 2002-2011.

De herweging is gebaseerd op een procédé dat bekend staat als lineair wegen. Deze methode is eerder succesvol toegepast in Mimos2 bij de MLT voor 1999-2002. Een uitgebreide beschrijving daarvan is te vinden in Waaijers (1998).[1] De huidige procedure is een verbetering van de vroegere in drie opzichten. Ten eerste is de implementatie nu volledig ondergebracht in SAS waar dit vroeger deels in GAUSS gebeurde, omdat toen nog niet kon worden beschikt over de SAS/IML module. Hierdoor verloopt het volledig geautomatiseerde herwegingsproces veel soepeler. Een tweede en belangrijker verbetering is dat nu een algoritme bij de berekening wordt betrokken waarbij zodanige begrenzing voor de nieuwe gewichten mogelijk is, dat ze dichter bij de oude gewichten liggen en in ieder geval positief zijn. Hierdoor kunnen uitgebreidere herwegingsoperaties worden uitgevoerd (ter vergelijking: het aantal randtotalen waarnaar herwogen wordt bedraagt in dit memorandum 48; bij Mimos2 is dit tien), waarbij de verlangde aanpassing forser kan zijn. Een derde verbetering van de huidige procedure ten opzichte van de vroegere is dat ingeval een set macrototalen aanleiding geeft tot problemen nu automatisch wordt teruggevallen op een kleinere (vooraf gekozen) set waarmee een nieuwe

[1] In: R.J. Waaijers, 1998, "De ophoogprocedure bij MIMOS2", Interne notitie CPB, Den Haag. Bij een test leverden de SAS-programmatuur en de wegingsprogrammatuur van het CBS (het pc-software pakket Bascula) dezelfde resultaten op, afgezien van verschillen als gevolg van eindige rekenprecisie bij een pc en een mainframe.

poging kan worden ondernomen. Dit bevordert in hoge mate de flexibiliteit van het herwegingsproces.

# 2 Theoretische fundering

De gebruikte wegingsmethodiek is een bijproduct van een schattingsprocedure[2] om een populatietotaal Y te schatten, bijvoorbeeld de loonsom in Nederland. Daarom wordt eerst een korte uiteenzetting gegeven van deze procedure. Als uitgangspunt geldt een schattingsprocédé dat uitsluitend berust op de kans om in de steekproef voor te komen. Elk element j van een steekproef kent een trekkingskans $\pi_j$ , bepaald door de geldende steekproefopzet. Een zuivere schatter voor het populatietotaal Y, bijvoorbeeld de loonsom in Nederland, is de *Horvitz-Thompsonschatter* die als volgt is gedefinieerd:

$$\underline{Y}_{HT} = \Sigma_s \, Y_j / \pi_j \text{ , waarbij gesommeerd is over de gehele steekproef.} \qquad (1)$$

Merk op dat (1) een gewogen som voorstelt met wegingsfactoren $1/\pi_j$ , die in de praktijk worden gerepresenteerd door de in het microgegevensbestand aanwezige ophoogfactoren $d_j$.[3] Een nauwkeuriger schatter[4] voor Y kan worden verkregen door gebruik te maken van de samenhang tussen Y en mogelijk aanwezige steekproefinformatie $X_s$, zoals, ingeval van loon, leeftijd en opleiding. Deze *regressieschatter* maakt gebruik van bekende populatiegegevens omtrent X, weergegeven door $X_{pop}$ , en wordt gedefinieerd door:

$$\underline{Y}_R = \underline{Y}_{HT} + b' \, (X_{pop} - X_s'd) \qquad (2)$$

Hierbij is d de vector met ophoogfactoren en b' een rijvector van coëfficiënten die geschat is op basis van steekproefinformatie $Y_s$, $X_s$ en de wegingsmatrix D (diagonaalmatrix met op de diagonaal de elementen $d_j$ ):

$$b = (X_s' \, D \, X_s)^{-1} \; X_s' \, D \, Y_s \qquad (3)$$

Ook de regressieschatter komt neer op een vorm van wegen, waarbij in de weging informatie met betrekking tot $X_s$ wordt betrokken. Dit wordt duidelijk als de schatter (2) met behulp van (3) wordt herschreven:

$$\underline{Y}_R = Y_s'd + b' \, (X_{pop} - X_s'd) = Y_s'd + Y_s'DX_s(X_s'D \, X_s)^{-1}(X_{pop} - X_s'd) = Y_s' \, \underline{w} \text{ , met}$$

[2] Zie: J.G. Betlehem en W.J. Keller, "Linear Weighting of Sample Survey Data", Journal of Official Statistics, Vol.3, No. 2, 1987, pp.141-153.

[3] $d_j$ is de reciproque van de 'insluitkans' $\pi_j$ , eventueel herwogen naar demografische kenmerken (zie inleiding).

[4] Deze schatter is asymptotisch zuiver, met een kleinere variantie dan de $Y_{HT}$ -schatter. Zie Betlehem en Keller, op.cit.

$$\underline{\mathbf{w}} = \mathbf{d\#g} \qquad (4)$$

waarbij # het Hadamard product (elementsgewijze vectorvermenigvuldiging) is en

$$\mathbf{g = 1 + X_s\,\varphi\,,\ met\ \ \varphi = (X_s'\,D\,X_s)^{-1}\,(X_{pop} - X_s'd)} \qquad (5)$$

Volgens (4) komt de regressieschatter neer op een weging van $Y_s$ met gewichten $\underline{w}$ die het product vormen van de oorspronkelijke ophoogfactoren en de *correctiegewichten* g. De correctiegewichten (5) zijn middels de *wegingscoëfficiënten* φ lineair afhankelijk van de per record aanwezige informatie met betrekking tot $X_s$ (vandaar de term *lineair* wegen).[5] De regressieschatter $Y_R$ met bijbehorende wegingsfactoren $\underline{w}$ behoort tot de klasse van *kalibratieschatters*.[6] Dit zijn schatters om een kengetal van een eindige populatie te schatten, zoals de loonsom in Nederland, en die daarbij gebruik maken van hulpvariabelen uit de steekproef waarvoor het populatietotaal bekend is. Een kalibratieschatter berust op gekalibreerde gewichten, die zo dicht mogelijk bij de oorspronkelijke ophoogfactoren van de steekproef liggen, volgens een gegeven afstandsmaat, en daarbij voldoen aan een aantal restricties in de vorm van *kalibratievergelijkingen*. Dit komt neer op de oplossing van het volgende optimaliseringsprobleem onder restricties (s heeft betrekking op steekproefgegevens) :

$$\text{Minimaliseer } F(w) = \Sigma_s\,(w_k/d_k - 1)^2\,d_k \text{ , onder de voorwaarden } X_s'w = X_{pop}. \qquad (6)$$

Daarmee zijn er twee wegen waarlangs de regressieschatter kan worden afgeleid: enerzijds op basis van de schatting van b (formuleringen 2 en 3), waarbij $\underline{w}$ als bijproduct geldt (formulering 4), anderzijds op basis van de oplossing van (6), waarbij $Y_R = Y_s\underline{w}$. Dit betekent dat de onder (4) gegeven gewichten zo dicht mogelijk bij de oorspronkelijke gewichten liggen en tegelijk voldoen aan de volgende zeer nuttige eigenschap[7] :

$$\mathbf{X_s'\,\underline{w}\ = X_{pop}} \qquad (7)$$

[5] Voor g = 1 levert (5) de Horvitz-Thompsonschatter uit (1) op.

[6] Zie: Jean-Claude Deville and Carl-Erik Särndal, "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, Vol. 87, No. 418, June 1992.

[7] Onder de (niet beperkende) voorwaarde dat $X_s$ een kolom ı bevat met uitsluitend enen, geldt: $\underline{w} = D\,X_s\,(\mathbf{X_s'}\,D\,X_s)^{-1}\,X_{pop}$ (formulering gebruikt in Waaijers (1998)). De voorwaarde betekent dat de som van de gewichten gelijk is aan de populatieomvang (kalibratie op populatieaantal) en dat de dummies voor 'restcategorieën' worden weggelaten in $X_s$ (bijv. de dummy voor geslacht = vrouw wordt weggelaten, alleen die voor man wordt opgenomen). Deze herformulering voor $\underline{w}$ kan als volgt worden afgeleid. De schatter b komt neer op een schatting van Y* op X* met Y* en X* transformaties van Y en X (vermenigvuldiging met $d^{½}$) . Als X de vector ı bevat dan bevat X* de vector $d^{½}$ die onafhankelijk is van het residu (Y* - X*b), zodat $Y_s'd - b'X_sd = 0$ in formulering (4), hetgeen leidt tot de genoemde herformulering voor $\underline{w}$. Deze voldoet ook aan (7), zoals direct blijkt bij substitutie.

Merk op dat dit niet afhangt van de doelvariabele Y, zodat de in (7) weergegeven formulering altijd de mogelijkheid verschaft om bestanden met microgegevens zodanig te herwegen (met minimale afwijking van de oorspronkelijke gewichten) dat voor een set variabelen $X_s$ geldt dat hun gewogen som gelijk is aan bekende ermee corresponderende *randtotalen* [8] uit de populatie $X_{pop}$.

Als de met $X_{pop}$ corresponderende categorie-indeling van kenmerken zodanig is dat de steekproef voor een aantal categorieën weinig waarnemingen oplevert, neemt de kans op negatieve waarden voor $\underline{w}$ toe, die dan moeilijk als gewicht te interpreteren zijn. Dit geldt bovendien sterker naarmate de verschillen tussen de in de steekproef aangetroffen verdelingen van $X_s$ en de populatieverdeling $X_{pop}$ groter zijn. Als remedie tegen negatieve gewichten kan het door Huang en Fuller (1978) ontwikkelde algoritme voor begrensd wegen worden ingezet. Bijgaand kader bevat een uiteenzetting van het algoritme. De eerder in (5) gegeven formulering voor de correctiegewichten is nu zodanig aangepast dat *reductiefactoren* Q (diagonaalmatrix) zijn opgenomen die een restrictieve invloed hebben op al die correctiegewichten die anders buiten een vooraf gekozen interval zouden geraken (in het bijzonder: negatief zouden worden).

Een noodzakelijke voorwaarde voor positieve gewichten is, zelfs bij het HF-algoritme, dat de randvoorwaarden zodanig zijn vastgesteld dat de gewogen steekproefcategorieën hieraan kunnen voldoen. Dit maakt controle op inconsistenties in de randvoorwaarden noodzakelijk. Zelfs consistentie in de randvoorwaarden garandeert geen positieve gewichten, omdat eisen kunnen worden opgelegd waaraan de data niet kunnen voldoen. Situaties waarin het HF-algoritme niet tot convergentie leidt, vaak resulterend in de melding van onmogelijke matrixinversie[9], zijn in de regel te herleiden tot inconsistente randvoorwaarden.

Wanneer sprake is van achtereenvolgende herwegingsexercities, waarbij telkens aan nieuwe randvoorwaarden $X_{pop}$ voldaan dient te zijn[10], kan worden gekozen tussen twee procedures: bij *recursief* wegen wordt de herweging gebaseerd op de gewichten $w_{t-1}$ uit de voorafgaande herweging, bij *direct* wegen wordt telkens uitgegaan van de oorspronkelijke bestandsgewichten d. Het voordeel van recursief wegen ten opzichte van direct wegen is dat er onderling verband blijft bestaan tussen de resulterende gewichtensets en dus in het algemeen tussen resultaten die op weging met deze sets berusten. Een neveneffect van deze interne consistentie kan zijn dat de herweging sneller verloopt.[11] Het nadeel van recursief wegen is dat de gewichten steeds verder

[8] De term 'randtotaal' is ontleend aan een representatie waarbij de steekproefuitkomsten zijn opgenomen in een matrix en de kolomrand van de matrix de somtotalen per kolom bevat.

[9] Als een reductiefactor 0 wordt kan de matrix Tα (zie kader) niet geïnverteerd worden.

[10] Zoals bij het Deelproject Herweging waarbij herweging wordt uitgevoerd voor de afzonderlijke jaren in de periode 2002-2011.

[11] Bijvoorbeeld als het eerste randtotaal sterk afwijkt van de steekproefinformatie en de achtereenvolgende randtotalen daarna weinig veranderen zijn de gewichten $w_{t-1}$ na een eerste herweging in het recursieve geval al zodanig aangepast dat slechts een beperkte correctie nodig is, terwijl bij directe weging uitgaande van de oorspronkelijke gewichten d telkens forse aanpassingen nodig zouden zijn. Bij dezelfde randtotalen geeft herweging in stappen via de recursieve methode andere

van de oorspronkelijke gewichten komen af te liggen, waardoor ze zeer klein of zeer groot kunnen worden. Dit kan voorkomen worden door de grenzen voor de correctiegewichten niet absoluut te nemen maar relatief ten opzichte van de oorspronkelijke gewichten d.[12] Dit brengt wel meer iteraties met zich mee dan wanneer direct wordt herwogen met absolute begrenzing. Bij wijze van experiment zijn drie varianten bij de herweging getest: recursief wegen met absolute begrenzing, recursief wegen met relatieve begrenzing en direct wegen met absolute begrenzing. Bij de bespreking van de resultaten (in de paragrafen 5.4 en 5.5) zullen de varianten worden beoordeeld op de mate waarin de resulterende gewichten overeenstemmen met zowel de oorspronkelijke gewichten als met de gewichten uit de voorafgaande herweging, waarbij ook de benodigde rekentijd in aanmerking zal worden genomen.

---

resultaten dan een directe herweging, hetgeen aannemelijk is op basis van formulering (6) : de te optimaliseren functie F(w) van de laatste herweging bij de recursieve methode heeft andere $d_k$-parameters dan F(w) van de directe methode.

[12] Bijvoorbeeld om te bereiken dat $0{,}1d < \underline{w}_t < 10d$ dient voor de correctiegewichten te gelden dat $0{,}1d/\underline{w}_{t-1} < g < 10d/\underline{w}_{t-1}$, zodat $L = 0{,}1d/\underline{w}_{t-1}$ en $U = 10d/\underline{w}_{t-1}$.

**Iteratief algoritme van Huang en Fuller (1978) voor begrensd wegen**

Stap 1: Kies L, U, maxit en $\zeta$ waarbij L en U de onder- en bovengrens van de correctiegewichten zijn, maxit het maximum aantal iteraties en $\zeta$ een dempingsfactor voorstelt.

Stap2: kies $\alpha = 0$ en bepaal de bijbehorende *reductiefactor* $Q_0 = I$ met I is eenheidsmatrix.

Stap3: bepaal het *correctiegewicht* $g_\alpha = 1 + X_s\, Q_\alpha T_\alpha^{-1} (X_{pop} - X_s'd)$

met $T_\alpha = X_s'D\, Q_\alpha\, X_s$.

Stap 4: als alle elementen van $g_\alpha$ binnen [L,U] zijn of als $\alpha$ = maxit: dan stop.

Stap 5: neem $\alpha = \alpha + 1$ en bepaal per element de relatieve afstand tot de grenzen:

$f_\alpha = (g_{\alpha-1} - 1)/(L - 1)$ als $g_{\alpha-1} \leq 1$

$f_\alpha = (g_{\alpha-1} - 1)/(U - 1)$ als $g_{\alpha-1} > 1$.

Stap 6: bepaal per element de nieuwe reductiefactoren:

$q_\alpha = q_{\alpha-1}$ als $0 \leq f_\alpha < 0{,}5$

$q_\alpha = q_{\alpha-1} (1 - \zeta (f_\alpha - 0{,}5)^2)$ als $0{,}5 \leq f_\alpha < 1$

$q_\alpha = q_{\alpha-1} (1 - \zeta /4)\, f_\alpha$ als $f_\alpha \geq 1$.

Bepaal $Q_\alpha$ (diagonaalmatrix met elementen $q_\alpha$ op de diagonaal) en herhaal vanaf stap 3.

Ontleend aan N.J. Nieuwenbroek en H.J. Boonstra, 2002, Bascula 4.0, Reference Manual, Statistics Netherlands, Heerlen. Voor de betekenis van $X_{pop}$ , $X_s$, d en D wordt verwezen naar de tekst. De dempingsfactor $\zeta$ heeft veelal 0,8 als optimale waarde. Voor de bovengrens U geldt in het algemeen U > 1 ( in Bascula wordt U default op 1000 gezet). Ter vermijding van negatieve gewichten wordt 0 < L <1 genomen. Als zowel L als U dichtbij 1 worden genomen kan hierdoor het aantal iteraties en daardoor de benodigde rekentijd toenemen. Anderzijds kan een te groot interval [L,U] leiden tot zeer kleine of zeer grote gewichten waardoor wegingsresultaten aanzienlijk vertekend kunnen worden. Bij navraag (met dank aan Harm Jan Boonstra) blijkt het CBS het begrenzingsalgoritme te gebruiken bij de herwegingen van de Enquete Beroepsbevolking en van de productiestatistieken.

# 3 Implementatie

De herweging is ondergebracht in een drietal SAS-programma's. Hiervan is één programma gericht op controle op consistentie in de randvoorwaarden, omdat dit een noodzakelijke voorwaarde is voor succesvolle herweging. Bijlage 1 beschrijft hoe deze controle verloopt. Een aantal van de geformuleerde condities is niet zozeer opgenomen vanwege hun dwingende voorschrift als wel om de gegevens ook op plausibiliteit te controleren. De gehele set randvoorwaarden, met in dit geval een tiental deelsets die betrekking hebben op de periode 2002-2011, wordt aldus doorgelicht voordat aan de herweging wordt begonnen.

De herwegingsprocedure is zodanig geïmplementeerd dat voor het geval er binnen het maximaal aantal toegestane iteraties geen oplossing wordt gevonden met uitsluitend gewichten binnen de voorafgestelde grenzen, automatisch wordt teruggevallen op een gekozen beperktere set randvoorwaarden, waardoor de kans op succes groter wordt.[13] Daarnaast is het altijd mogelijk per herweging de set randvoorwaarden anders in te stellen, waardoor problematische randtotalen kunnen worden vermeden.

In de toepassing voor het project Implementatie IPO vallen de herwegingswerkzaamheden in twee delen uiteen. Eén deel betreft, in principe, eenmalige herwegingsprocedures die betrekking hebben op gerealiseerde en vaststaande randvoorwaarden. De overige herwegingsprocedures worden vaker uitgevoerd en hebben betrekking op randvoorwaarden die de status hebben van een voorlopige raming. Beide categorieën werkzaamheden zijn ondergebracht in aparte programma's. Tot de eenmalige procedures behoort ook dat uitgaande van een basisbestand met persoonsgegevens een huishoudensbestand wordt aangemaakt met daarin de totaalscores van de leden van het huishouden met betrekking tot de randvoorwaarden. Deze randvoorwaarden hebben betrekking op zowel originele IPO-variabelen als daarvan afgeleide indicatoren. Het huishoudensbestand dient uiteindelijk als het bestand waarop de herweging wordt uitgevoerd. De samenstelling van het daaraan ten grondslag liggende basisbestand is het onderwerp van de volgende paragraaf.

# 4 Basisbestand[14]

Het basisbestand is een uit het derde IPO-testbestand afgeleid persoonsbestand met naast het huishoudnummer alle variabelen die van belang zijn bij de herweging.[15] Daarbij gaat het om persoonskenmerken als leeftijd, geslacht, inkomensbron, hoofdbron van inkomen, combinaties

[13] In de beperkte set randtotalen vervallen de volgende kenmerken: leeftijdsklasse 25-29 (wordt samengevoegd met klasse 30-59), tweeverdiener, hoofdinkomensbron, eigen woningbezit, samenloop inkomensbronnen.

[14] Met dank aan Joke Goes voor deze bijdrage.

[15] Zie h:\p_ipo\microdata\randtotalen\randbestandrob_def.sas.

van een aantal inkomensbronnen en huishoudkenmerken als type huishouden, eigen woningbezit of huurwoning, tweeverdiener of alleenverdiener (bij paren). Van al deze kenmerken is van belang dat, na correctie van de per record aanwezige ophoogfactoren, hun macrototaal (i.c. hun gewogen steekproeftotaal) correspondeert met de in de randvoorwaarden opgelegde totalen. In deze paragraaf wordt besproken hoe de variabelen in het basisbestand zijn aangemaakt uit de originele IPO gegevens.

De verschillende inkomenbronnen worden ingedeeld, zie tabel 4.1, en aan de hand daarvan wordt de (hoofd)inkomensbron vastgesteld. Het loon wordt bijvoorbeeld als inkomensbron gezien als de variabele bl_m_incl_en groter is dan nul. Dit geldt voor alle bronnen, behalve voor winst, daar mag t2_win_en ook negatief zijn. Een bron is hoofdinkomensbron als het het grootste is van alle bronnen. Ook hier is de winst weer een uitzondering: als er winst is, is dat de hoofdbron. Er wordt ook rekening gehouden met samenloop. Bij twee of meer inkomensbronnen wordt gekeken naar de twee grootste bronnen, de overige bronnen tellen voor de indeling niet mee.

Met behulp van de (geïmputeerde[16]) deeltijdfactoren worden voor de markt, zorg en overheid deeltijdfactoren berekend op basis van gewerkt aantal weken per jaar: djaarw_* = dtf_*_dn / dtfsim_dn, waarbij * = m,z,o, dtf_*_dn = deeltijdfactor per jaar in de markt/zorg/overheid en dtfsim_dn = deeltijdfactor per week. In de aantallen werknemers met loon markt, zorg en overheid (de variabelen mktdlt, zrgdlt en ovhdlt) is deze deeltijdfactor inbegrepen. Daarnaast wordt deze factor ook gebruikt als een van de drie bronnen het hoofdinkomen is (hmktdlt, hzrgdlt en hovhdlt). Bij samenloop van deze inkomensbronnen wordt de djaarw_* van de hoogste inkomensbron toegekend.

[16] Deze variabelen zijn aan het IPO toegevoegd in het kader van het Deelproject Imputatie. Zie: Startdocument IPO Deelproject 01: Imputatiemodel, dd. 24 mei 2004.

**Tabel 4.1 Definitie inkomen**

| inkomensbron | variabelen uit IPO |
|---|---|
| markt | bl_m_incl_en |
| zorg | bl_z_incl_en |
| overheid | bl_o_incl_en |
| WW | t5_ww_en + t5_wa_en |
| WAO | t5_ao_en + ao_waj_en |
| bijstand | t6_b_en |
| VUT | t5_pen_en, mits lftkl_en <= 14 |
| AOW | t5_aow_en |
| winst | t2_win_en, als ongelijk aan nul, dan winst hoofdbron |
| overig arbeidsink. | t1_dgn_en + ovink_en + t9_onb_en + buiink_en |
| student | t6_stu_en |
| geen ink | sec_kn = 14 |
| overig ink. (geen arbeidsink.) | de rest |

Er worden twee indelingen naar soort huishouden gemaakt. De eerste indeling is te vinden in tabel 4.2. Hiervan wordt voor de herweging alleen het kenmerk ‘tweeverdiener’gebruikt. Op basis van de eerste indeling wordt voor de tweede indeling aangesloten bij een CBS-indeling waarbij gekeken wordt hoeveel personen er in een huishouden zitten en hoeveel daarvan kinderen zijn. De personen met pos_kn =7 in het IPO worden gezien als “anderen” en zijn niet van invloed op de indeling. De gemaakte indeling is te vinden in tabel 4.3. Hierbij wordt per IPO-huishouden het aantal pos_kn =5 en pos_kn=6 geteld; dit is het aantal kinderen per huishouden. Alle categorieën van de tweede indeling worden gebruikt bij de herweging.

**Tabel 4.2 Indeling 1 soort huishouden**

| soort huishouden | conditie |
|---|---|
| alleenstaande | h_sam_kn = 1 |
| alleenstaande ouder | 6 <= h_sam_kn <= 8 OR 13 <= h_sam_kn <= 15 |
| alleenverdiener | (pos_kn = 3 OR pos_kn = 4) AND sec_kn = 14 |
| tweeverdiener | (pos_kn = 3 OR pos_kn = 4) AND sec_kn ne 14 |
| overig | de rest |

**Tabel 4.3 Indeling 2 soort huishouden**

| soort huishouden | waarde h_sam_kn |
|---|---|
| Institutionele huishoudens | 17 |
| Eenpersoonshuishoudens | 1 |
| Meerpersoonshuishoudens zonder kinderen | 2,9 |
| Paren met 1 kind | 3,4,5,10,11,12, en aantal kinderen = 1 |
| Paren met 2 of meer kinderen | 3,4,5,10,11,12, en aantal kinderen > 1 |
| Eenoudergezinnen met 1 kind | 6,7,8,13,14,15, en aantal kinderen = 1 |
| Eenoudergezinnen met 2 of meer kinderen | 6,7,8,13,14,15, en aantal kinderen > 1 |
| Overige meerpersoonshuishoudens | 16 |

Verder wordt het geslacht rechtstreeks uit het IPO overgenomen (ges_kn), de IPO-variabele h_woon_kn geeft aan of iemand huurt of een eigen huis heeft en er worden leeftijdsklassen gemaakt aan de hand van de IPO-variabele lftkl_kn.

# 5 Herweging voor de periode 2002-2011

De in hoofdstuk 2 beschreven herwegingsprocedure is toegepast op het basisbestand 2002 waarbij herwogen is naar randtotalen voor de periode 2002-2011. De berekening van de nieuwe gewichten is gebaseerd op een uit het basisbestand aangemaakt huishoudensbestand, waarop vervolgens de herwegingsprocedure uit hoofdstuk 2 is toegepast. Na bepaling van de nieuwe gewichten $\underline{w}$ zijn deze vervolgens overgebracht naar het basisbestand met behulp van de koppelingsvariabele 'huishoudnummer'. Weging met $\underline{w}$ toont aan dat in het basisbestand inderdaad wordt voldaan aan de eis dat de gewogen steekproeftotalen corresponderen met de vooraf vastgestelde randtotalen voor de periode 2002-2011. In de volgende paragrafen bespreken we de daartoe ondernomen stappen.

## 5.1 De representativiteit van het basisbestand

In tabel 5.1 is weergegeven aan welke kenmerken het basisbestand na herweging dient te voldoen in 2002. In de tweede kolom van de tabel staan de aantallen zoals deze worden gevonden bij gebruikmaking van de in het IPO aanwezige originele ophoogfactor, corresponderend met de vector $X_s$'d uit hoofdstuk 2, vergelijking (5). De derde kolom geeft een representatief beeld voor 2002 op basis van diverse bronnen. Deze bronnen betreffen het CBS en een aantal afdelingen binnen het CPB, maar ook het IPO zelf ingeval geen andere betere

bron voorhanden is. De derde kolom geeft daarmee het randtotaal (corresponderend met $X_{pop}$[17] in hoofdstuk 2, vergelijking (5)) waaraan de gewogen kenmerken na herweging met de nieuwe gewichten w dienen te voldoen. Het opnemen van randtotalen waarvoor het IPO zelf de bron is garandeert dat na herweging de oorspronkelijke totalen uit kolom 2 bewaard blijven. Voor de categorieën waarvoor SZ als bron geldt zijn de vereiste randtotalen gelijk aan de IPO-aantallen. Het betreft hier een aantal inkomensbronnen waarvoor SZ een relatie heeft gelegd tussen het IPO en cijfers uit de Nationale Rekeningen, Sociale Verzekeringsbank en UWV. Daarnaast heeft de categorie IPO2002/SZ betrekking op IPO-cijfers waarop SZ een factor heeft gezet.[18] Een uitgebreide beschrijving van de in dit memorandum gebruikte randtotalen voor de periode 2002-2011 en een vergelijking met het IPO2002 is te vinden in Stegeman (2005).[19]
De tabel laat zien dat waar een vergelijking met andere betrouwbare bronnen mogelijk was het IPO redelijk representatieve uitkomsten geeft.

**Tabel 5.1 Representativiteit IPO2002**

| | 2002 IPO macro | 2002 Representatief beeld | bron |
|---|---|---|---|
| **Aantal personen naar leeftijd** | | | |
| totaal | 16193 | 16193 | Statline CBS |
| 0-5 jaar | 1220 | 1219 | Statline CBS |
| 6-11 jaar | 1207 | 1195 | Statline CBS |
| 12-17 jaar | 1169 | 1180 | Statline CBS |
| 18-24 jaar | 1346 | 1348 | Statline CBS |
| 25-29 jaar | 1031 | 1031 | Statline CBS |
| 30-59 jaar | 7226 | 7226 | Statline CBS |
| 60-64 jaar | 773 | 773 | Statline CBS |
| 65 jaar en ouder | 2220 | 2220 | Statline CBS |
| **Aantal personen naar geslacht** | | | |
| mannen | 8015 | 8015 | Statline CBS |
| vrouwen | 8177 | 8177 | Statline CBS |
| **Aantal personen naar inkomensbron** | | | |
| aantal personen met loon markt | 5564 | 5701 | CON/ARB |
| aantal personen met loon zorg | 706 | 723 | CON/IEP |
| aantal personen met loon overheid | 664 | 680 | CON/ARB |

[17] Voor de herweging is redundante informatie in $X_{pop}$ en in $X_s$ weggelaten. Het betreft de categorieën leeftijdsklasse 0-5 jr, man, overige meerpersoonshuishoudens en huurder, omdat de totalen voor 'aantal personen' en 'aantal huishoudens' al zijn opgenomen. Zie ook voetnoot 7.

[18] Deze factor is gebaseerd op de verhouding tussen totaal aantal werknemers NR en totaal aantal werknemers IPO (beide gecorrigeerd voor de periode waarover het inkomen is genoten).

[19] Zie het CPB-memo "Randtotalen II.2" van Hans Stegeman dd. 20 oktober 2005.

| | | | |
|---|---|---|---|
| aantal personen met overig arbeidsinkomen | 483 | 483 | IPO2002 |
| aantal personen met winst uit onderneming | 831 | 831 | ARB |
| | | | |
| aantal personen met WW-uitkering | 443 | 443 | SZ |
| aantal personen met WAO-uitkering | 796 | 796 | SZ |
| aantal personen met bijstand e.d. | 469 | 469 | SZ |
| aantal personen met VUT/prepensioen-uitkering | 710 | 710 | IPO2002 |
| aantal personen met AOW-uitkering | 2197 | 2197 | SZ |
| aantal personen met studiebeurs | 650 | 650 | IPO2002 |
| overige personen met inkomen | 57 | 57 | IPO2002 |
| aantal personen zonder inkomen | 4180 | 4180 | IPO2002 |
| | | | |
| arbeidsjaren markt (exclusief zelfstandigen!) | 4357 | 4633 | CON/ARB |
| arbeidsjaren zorg | 477 | 588 | CON/IEP |
| arbeidsjaren overheid | 599 | 553 | CON/ARB |
| | | | |
| **Aantal personen naar hoofdinkomensbron** | | | |
| aantal personen met loon markt | 5230 | 5359 | IPO2002/SZ |
| aantal personen met loon zorg | 659 | 675 | IPO2002/SZ |
| aantal personen met loon overheid | 631 | 646 | IPO2002/SZ |
| aantal personen met WW-uitkering | 136 | 136 | IPO2002 |
| aantal personen met WAO-uitkering | 576 | 576 | IPO2002 |
| aantal personen met bijstand | 358 | 358 | IPO2002 |
| | | | |
| **Samenloopgevallen>0,5%** | | | |
| markt/ww | 239 | 239 | IPO2002 |
| markt/vut | 161 | 161 | IPO2002 |
| markt/zelfstandigen | 161 | 161 | IPO2002 |
| markt/overig arbeidsinkomen | 106 | 106 | IPO2002 |
| | | | |
| markt/bijstand | 98 | 98 | IPO2002 |
| markt/wao | 149 | 149 | IPO2002 |
| wao/vut | 86 | 86 | IPO2002 |
| | | | |
| **Huishoudens** | | | |
| aantal huishoudens totaal | 7201 | 7211 | Statline CBS |
| Institutionele huishoudens | 206 | 215 | Statline CBS |
| eenpersoonshuishoudens | 2383 | 2384 | Statline CBS |
| meerpersoonshuishoudens zonder kinderen | 2056 | 2047 | Statline CBS |
| paren met 1 kind | 742 | 760 | Statline CBS |
| paren met 2 of meer kinderen | 1306 | 1333 | Statline CBS |
| eenoudergezinnen met 1 kind | 266 | 251 | Statline CBS |

| | | | |
|---|---|---|---|
| eenoudergezinnen met 2 of meer kinderen | 175 | 173 | Statline CBS |
| overige meerpersoonshuishoudens | 67 | 48 | Statline CBS |
| aantal huishoudens eigen woning | 3394 | 3685 | Statline |
| aantal huishoudens huurders | 3807 | 3311 | Statline |
| aantal huishoudens met tweeverdieners | 3228 | 3228 | IPO2002 |

De vermelde aantallen betreffen dzd personen respectievelijk dzd huishoudens.

## 5.2 De ophoging naar randtotalen

Zoals in hoofdstuk 3 is besproken wordt bij de implementatie van de herweging onderscheid gemaakt tussen eenmalige en herhaaldelijke werkzaamheden. In principe eenmalig van karakter is de herweging naar randtotalen, die niet meer zullen wijzigen omdat ze inmiddels als gerealiseerd worden beschouwd. Daarnaast is er de herhaaldelijk uitgevoerde herweging naar randtotalen die het karakter van een raming dragen. Hierbij aansluitend wordt de gehele herwegingsoperatie hieronder in twee afzonderlijke paragrafen besproken.[20]

### 5.2.1 De ophoging in het basisjaar en in de gerealiseerde jaren

Het IPO2002 spoort goed met de gekozen set randtotalen voor 2002, zo kon in paragraaf 5.1 worden geconcludeerd. Voor de eerste herwegingexercitie, die in het algemeen lastig kan zijn omdat verschillen tussen IPO en andere informatiebronnen dienen te worden rechtgetrokken, resulteerde dit in een succesvolle afronding, waarin het niet nodig was terug te vallen op de beperkte set randtotalen. Wel waren drie iteratieslagen nodig voor het HF-algoritme om negatieve gewichten kwijt te raken. Voor de jaren 2003-2004 kon worden volstaan met hooguit twee iteraties omdat na de herweging 2002 de aanpassing aan de randtotalen in deze jaren minder extreem was.[21] De totale ophogingexercitie nam voor de periode 2002-2004 iets minder dan vier minuten in beslag.

### 5.2.2 De ophoging in de ramingjaren

Hetgeen in de vorige paragraaf is opgemerkt over de relatief beperkte verschillen tussen aanwezige gewogen randtotalen en gewenste randtotalen nadat de herweging 2002 heeft plaatsgevonden, geldt ook voor de periode 2005-2011. Zoals tabel 5.2 laat zien verloopt de ontwikkeling voor de afzonderlijke randtotalen in het algemeen zeer gelijkmatig. De grote

[20] Zie bijlage 2 voor enige verdelingskenmerken van de IPO-gewichten en van de na herweging resulterende gewichten.

[21] De grote verschillen tussen 'IPO na (her)weging' en 'gewenst' randtotaal, zoals in 2002 voor de categorieën 'arbeidsjaren zorg', 'hoofdinkomensbron loon markt' en 'bezit van eigen woning', zijn dan verdwenen. Wel is het aantal randtotalen groter waarbij er een verschil is tussen 'IPO na (her)weging' en 'gewenst'.

verschillen tussen aanwezig en gewenst randtotaal, zoals in 2002 voor de categorieën ‘arbeidsjaren zorg’, ‘hoofdbroninkomen loon markt’ en ‘bezit van eigen woning’, zijn dan verdwenen (wel is het aantal randtotalen groter waarbij er een verschil is tussen ‘aanwezig’ en ‘gewenst’).

Dit resulteerde in een soepel verloop van de herwegingsprocedure voor deze periode: voor alle zeven uitgevoerde herwegingsoperaties waren twee iteratieslagen voldoende om aan alle randvoorwaarden te voldoen. De herwegingsprocedure voor de gehele periode 2005-2011 duurde ook in dit geval vier minuten. Mogelijk dat dit in de praktijk nog wat sneller verloopt als de aanpassingen bij achtereenvolgende ramingen kleiner worden.

**Tabel 5.2 Randtotalen: raming tot en met 2011**

| | 2005 | 2008 | 2011 |
|---|---|---|---|
| **Aantal personen naar leeftijd** | | | |
| totaal | 16339 | 16439 | 16545 |
| 0-5 jaar | 1201 | 1143 | 1100 |
| 6-11 jaar | 1186 | 1209 | 1196 |
| 12-17 jaar | 1203 | 1194 | 1189 |
| 18-24 jaar | 1360 | 1399 | 1438 |
| 25-29 jaar | 992 | 983 | 991 |
| 30-59 jaar | 7235 | 7039 | 6933 |
| 60-64 jaar | 838 | 1038 | 1064 |
| 65 jaar en ouder | 2324 | 2435 | 2635 |
| | | | |
| **Aantal personen naar geslacht** | | | |
| mannen | 8079 | 8122 | 8171 |
| vrouwen | 8260 | 8316 | 8375 |
| | | | |
| **Aantal personen naar inkomensbron** | | | |
| aantal personen met loon markt | 5413 | 5592 | 5574 |
| aantal personen met loon zorg | 788 | 837 | 835 |
| aantal personen met loon overheid | 687 | 708 | 706 |
| aantal personen met overig arbeidsinkomen | 472 | 487 | 486 |
| aantal personen met winst uit onderneming | 779 | 785 | 782 |
| | | | |
| aantal personen met WW-uitkering | 766 | 435 | 433 |
| aantal personen met WAO-uitkering | 748 | 645 | 626 |
| aantal personen met bijstand e.d. | 500 | 498 | 496 |
| aantal personen met VUT/prépensioenuitkering | 770 | 898 | 866 |
| aantal personen met AOW-uitkering | 2299 | 2419 | 2619 |
| aantal personen met studiebeurs | 656 | 674 | 693 |
| overige personen met inkomen | 57 | 57 | 57 |
| aantal personen zonder inkomen | 4199 | 4212 | 4199 |

| | | | |
|---|---|---|---|
| arbeidsjaren markt (exclusief zelfstandigen!) | 4389 | 4533 | 4519 |
| arbeidsjaren zorg | 639 | 679 | 677 |
| arbeidsjaren overheid | 557 | 574 | 572 |
| | | | |
| **Aantal personen naar hoofdinkomensbron** | | | |
| aantal personen met loon markt | 5088 | 5256 | 5239 |
| aantal personen met loon zorg | 736 | 782 | 779 |
| aantal personen met loon overheid | 652 | 673 | 671 |
| aantal personen met WW-uitkering | 234 | 133 | 133 |
| aantal personen met WAO-uitkering | 541 | 467 | 453 |
| aantal personen met bijstand | 381 | 380 | 379 |
| | | | |
| **Samenloopgevallen>0,5%** | | | |
| markt/ww | 240 | 234 | 234 |
| markt/vut | 155 | 163 | 162 |
| markt/zelfstandigen | 153 | 157 | 157 |
| markt/overig arb.inkomen | 101 | 104 | 104 |
| markt/bijstand | 94 | 97 | 96 |
| markt/wao | 141 | 143 | 142 |
| wao/vut | 87 | 88 | 85 |
| | | | |
| **Huishoudens** | | | |
| aantal huishoudens totaal | 7356 | 7506 | 7663 |
| Institutionele huishoudens | 215 | 215 | 215 |
| eenpersoonshuishoudens | 2497 | 2635 | 2769 |
| meerpersoonshuishoudens zonder kinderen | 2055 | 2047 | 2060 |
| paren met 1 kind | 749 | 739 | 734 |
| paren met 2 of meer kinderen | 1337 | 1341 | 1331 |
| eenoudergezinnen met 1 kind | 269 | 289 | 308 |
| eenoudergezinnen met 2 of meer kinderen | 185 | 192 | 196 |
| overige meerpersoonshuishoudens | 48 | 49 | 50 |
| | | | |
| aantal huishoudens eigen woning | 3882 | 4090 | 4308 |
| aantal huishoudens huurders | 3258 | 3201 | 3140 |
| | | | |
| aantal huishoudens met tweeverdieners | 3250 | 3259 | 3275 |

De vermelde aantallen betreffen dzd personen respectievelijk dzd huishoudens.

## 5.3 De herwogen verdeling van deeltijd en sociale verzekeringsdagen

Na de boven beschreven herwegingexercitie is nagegaan wat de invloed van de nieuwe gewichten is op een tweetal variabelen die niet zijn opgenomen in de set randtotalen maar wel belangrijk zijn binnen het project Implementatie IPO. Het gaat dan om de variabelen

'deeltijdfactor per week' en 'aantal sociale verzekeringsdagen per week'. Deze variabelen zijn in het IPO geïmputeerd in het kader van Deelproject 01.[22]

Vergelijking van de voorlaatste regel met de daaraan voorafgaande twee regels in tabel 5.3 laat zien dat de herweging een gunstige invloed heeft op de aansluiting van het IPO bij het LoonStructuurOnderzoek (LSO) dat als een betrouwbare bron mag worden gezien. De verdeling ondergaat niet veel invloed van de herweging in andere jaren. Ter illustratie is op de onderste regel de gewogen verdeling weergegeven met gewichten voor het jaar 2011.

**Tabel 5.3 Deeltijdfactor per week bij IPO en LSO**

| | | **Deeltijdfactor per week** (IPO-variabele: dtfsim) | | | |
|---|---|---|---|---|---|
| | | Gemiddelde | 1e deciel | 1e kwartiel | mediaan |
| **2002 ongewogen** | | | | | |
| | IPO | 0,7592 | 0,2215 | 0,5000 | 1 |
| | LSO | 0,7999 | 0,3542 | 0,6000 | 1 |
| **2002 gewogen** | | | | | |
| | IPO | 0,7665 | 0,2187 | 0,5252 | 1 |
| | LSO | 0,8053 | 0,3095 | 0,6122 | 1 |
| **herwogen** | | | | | |
| 2002 | IPO | 0,7949 | 0,2603 | 0,6000 | 1 |
| 2011 | IPO | 0,7966 | 0,2685 | 0,6003 | 1 |

Dat herweging ook voor de verdeling van het aantal SV-dagen per week goed uitpakt toont tabel 5.4. Vergelijking van de voorlaatste kolom met de daaraan voorafgaande twee kolommen in deze tabel laat zien hoe na weging met de nieuwe gewichten voor 2002 de herwogen frequentieverdeling van het IPO beter aansluit bij de gewogen verdeling van het LSO (vijfde kolom).[23] Ook hier geldt dat de herweging in latere jaren niet veel verandert aan de verdeling binnen het IPO, zoals de onderlinge vergelijking tussen de laatste twee kolommen laat zien.

[22] Zie Startdocument IPO Deelproject 01: Imputatiemodel dd. 24 mei 2004.

[23] Een plausibele verklaring voor de betere aansluiting tussen IPO en LSO voor deeltijdfactor en SV-dagen is het impliciet meenemen van de deeltijdfactor per week in de randtotalen. Dit gebeurt door zowel het opnemen van het aantal arbeidsjaren als de quote arbeidsweken/52 (in 'loon markt', 'loon zorg' en 'loon overheid'; zie de beschrijving van het basisbestand in hoofdstuk 4). In de verhouding tussen deze twee factoren is de deeltijdfactor per week verdisconteerd.

**Tabel 5.4 Cumulatieve frequentieverdeling SV-dagen per week bij IPO en LSO**

**Verdeling: cumulatieve percentages**

| SV-dagen per week | 2002 ongewogen | | 2002 gewogen | | 2002 | 2011 |
|---|---|---|---|---|---|---|
| | IPO | LSO | IPO | LSO | IPO herwogen | |
| 0,2 | 0,16 | 0,10 | 0,14 | 0,06 | 0,10 | 0,09 |
| 0,5 | 0,78 | 0,55 | 0,74 | 0,75 | 0,52 | 0,49 |
| 1,0 | 5,35 | 3,38 | 5,52 | 4,27 | 4,36 | 4,28 |
| 1,5 | 7,48 | 4,26 | 7,66 | 5,54 | 6,15 | 5,85 |
| 2,0 | 12,0 | 7,61 | 12,17 | 9,19 | 10,04 | 9,73 |
| 2,5 | 14,82 | 9,94 | 14,62 | 11,15 | 12,17 | 11,80 |
| 3,0 | 22,14 | 18,48 | 21,35 | 17,96 | 18,52 | 18,35 |
| 3,5 | 24,14 | 20,57 | 23,29 | 19,57 | 20,33 | 20,17 |
| 4,0 | 31,22 | 30,10 | 30,42 | 27,70 | 27,71 | 27,83 |
| 4,5 | 33,33 | 31,72 | 32,50 | 29,28 | 29,66 | 29,77 |
| 5,0 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |

## 5.4 Resulterende correctiegewichten

Voor de achtereenvolgende herwegingen in de periode 2002-2011 is geëxperimenteerd met drie varianten (zie hoofdstuk 2): recursieve herweging, met absolute of relatieve (ten opzichte van de oorspronkelijke gewichten) begrenzing, en directe herweging. Recursieve weging met absolute begrenzing (RA) kan leiden tot gewichten die uiteindelijk ver van de oorspronkelijke gewichten af liggen. Bij toepassing van deze variant ligt in de vijfde herwegingsronde (jaar 2006) de verhouding ten opzichte van de oorspronkelijke IPO-gewichten tussen 0,01 en 7,89, in de tiende herwegingsronde (jaar 2011) ligt dit tussen 0,0003 en 6,57. De uitersten van deze intervallen hebben betrekking op relatief weinig records (90% ligt tussen 0,3 en 1,9). De toename van zeer kleine gewichten blijkt kenmerkend voor deze variant. De overige kenmerken komen nagenoeg overeen met het beeld dat bij recursieve weging met relatieve begrenzing (RR) ontstaat en dat is weergegeven in tabel 5.5. Voor wat betreft de verhouding van de correctiegewichten ten opzichte van de oorspronkelijke IPO-gewichten $W_0$ geldt nu dat deze minimaal 0,1 bedraagt (op basis van $L = 0{,}1W_0/W_{t-1}$). Het maximale correctiegewicht blijft ruimschoots onder 10 (bij $U = 10W_0/W_{t-1}$). De gewichten liggen redelijk rond de waarde 1 te oordelen naar het gemiddelde, het 10e en 90e percentiel en de standaardafwijking.[24] In de

[24] Hoewel de populatieomvang toeneemt (van 16193 in 2002 naar 16545 in 2011, zie de paragrafen 5.1 en 5.2) daalt het gemiddelde correctiegewicht $\bar{g}$ . Dit berust op de positieve relatie (covariantie) tussen correctiegewicht g en oorspronkelijk gewicht $W_0$. Er geldt: $\bar{g} = (\Sigma W - cov(g,W_0)) / \Sigma W_0$. Als cov = 0 dan is $\bar{g} = \Sigma W / \Sigma W_0 = \bar{W}/\bar{W}0$ (ofwel gemiddelde verhouding

laatste kolom is F(w) gegeven als mate van overeenstemming tussen $W_t$ en $W_0$.[25] Hieruit blijkt dat de afwijking ten opzichte van de oorspronkelijke gewichten groter wordt bij elke herwegingsronde (idealiter geldt F(w) = 0). Dit is het gevolg van een steeds grotere afwijking tussen gewenste randtotalen en de met de oorspronkelijke gewichten gewogen totalen (dit verschijnsel doet zich dus voor bij alle wegingsvarianten). De laatste twee regels van de tabel geven een beeld van de samenhang tussen de gewichten uit de opeenvolgende herwegingen. De spreiding van de correctiegewichten (i.c. de verhouding $W_t / W_{t-1}$) om 1 is nu veel kleiner dan bij de zojuist gegeven vergelijking ten opzichte van $W_0$ terwijl ook F(w) veel kleiner is in vergelijking met de twee erboven vermelde waarden.[26]

**Tabel 5.5 Verdeling van correctiegewichten bij recursieve weging met relatieve gewichtsbegrenzing (RR)**

| Jaar | gemiddelde | minimum | maximum | 10e Pctl | 90e Pctl | St.afw. | F(w) |
|---|---|---|---|---|---|---|---|
| 2002 | 0,9933 | 0,1606 | 4,7868 | 0,5705 | 1,3303 | 0,3567 | 1824 |
| tov $W_0$ : | | | | | | | |
| 2006 | 0,9880 | 0,1002 | 7,6444 | 0,5114 | 1,4026 | 0,4514 | 3057 |
| 2011 | 0,9741 | 0,1000 | 6,6467 | 0,3901 | 1,5791 | 0,5123 | 4156 |
| tov $W_{t-1}$ : | | | | | | | |
| 2006 | 1,0036 | 0,3183 | 1,6501 | 0,9054 | 1,0899 | 0,0902 | 167 |
| 2011 | 0,9933 | 0,6922 | 1,0615 | 0,9417 | 1,0405 | 0,0419 | 27 |

In tabel 5.6 zijn de kenmerken van de correctiegewichten bij directe weging (D) weergegeven (met absolute begrenzing). Zoals te verwachten was, is de overeenstemming met de oorspronkelijke gewichten groter (resp. F(w) is kleiner) en met de voorafgaande wegingsresultaten kleiner (resp. F(w) is groter) dan bij RR en RA. Overigens verschillen de resultaten niet veel.[27]

---

is gelijk aan verhouding der gemiddelden). Bij de geldende sets randvoorwaarden is cov > 0 en stijgt deze van 0,0004 in 2002 naar 0,0015 in 2006 en vervolgens naar 0,0032 in 2011.

[25] $F(w) = \Sigma_s(w_k/d_k - 1)^2 d_k = \Sigma_s(g - 1)^2 d_k$ , zie hoofdstuk 2. Wordt $W_0$ genormeerd zodat de som van de gewichten 1 is, dan komt F(w) zeer goed overeen met de standaardafwijking van g. Omdat F(w) berust op de afwijking van g ten opzichte van 1 en de standaardafwijking berust op de afwijking ten opzichte van het gemiddelde van g, is F(w) een beter criterium voor de mate van overeenstemming tussen gewichtensets.

[26] F(w) is in dit geval: $F(w) = \Sigma_s(w_k/w_{-1,k} - 1)^2 w_{-1,k} = \Sigma_s(g^* - 1)^2 w_{-1,k}$.

[27] Er is een sterke relatie met de gewichten uit de RR-variant : de correlatie tussen de RR- en de D-gewichten voor het jaar 2011 bedraagt 0,99. Verder is het D-gewicht gemiddeld 0,4% lager dan het RR-gewicht en geldt voor de verhouding D-gewicht/ RR-gewicht dat 10% groter dan 1,10 is (maximaal 3,42) en 10% kleiner dan 0,78 is (minimaal 0,24).

**Tabel 5.6 Verdeling van correctiegewichten bij directe weging (D)**

| Jaar | gemiddeld | minimum | maximum | 10e Pctl | 90e Pctl | St.afw. | F(w) |
|---|---|---|---|---|---|---|---|
| 2002 | 0,9933 | 0,1606 | 4,7868 | 0,5705 | 1,3303 | 0,3567 | 1824 |
| tov $W_0$ : | | | | | | | |
| 2006 | 0,9880 | 0,1558 | 5,2954 | 0,4470 | 1,4326 | 0,4472 | 2959 |
| 2011 | 0,9742 | 0,1426 | 4,7383 | 0,2890 | 1,5983 | 0,5127 | 4037 |
| tov $W_{t-1}$ : | | | | | | | |
| 2006 | 1,0064 | 0,2687 | 2,3947 | 0,8840 | 1,1027 | 0,1089 | 184 |
| 2011 | 0,9919 | 0,4620 | 1,3750 | 0,9231 | 1,0468 | 0,0544 | 34 |

In tabel 5.7 is een beeld gegeven van de correctiegewichten die resulteren bij recursieve herweging met relatieve begrenzing voor de beperkte set randvoorwaarden (RRK).[28]
De spreiding in de correctiegewichten is kleiner in vergelijking met die van de uitgebreide set randvoorwaarden in tabel 5.5. Omdat het aantal randvoorwaarden waaraan voldaan dient te worden kleiner is, heeft dit als gevolg dat de overeenstemming met zowel $W_0$ als met $W_{t-1}$ groter is: de waarden voor F(w) en de spreidingsmaatstaven zijn nu kleiner dan bij de uitgebreide set.

**Tabel 5.7 Verdeling van correctiegewichten bij recursieve weging: kleine set randvoorwaarden (RRK)**

| Jaar | gemiddeld | minimum | maximum | 10e Pctl | 90e Pctl | St.afw. | F(w) |
|---|---|---|---|---|---|---|---|
| 2002 | 0,9946 | 0,1724 | 4,7289 | 0,6307 | 1,3032 | 0,3339 | 1593 |
| tov $W_0$ : | | | | | | | |
| 2006 | 0,9897 | 0,1004 | 5,7117 | 0,5821 | 1,3994 | 0,3879 | 2211 |
| 2011 | 0,9764 | 0,1000 | 5,3879 | 0,4735 | 1,4324 | 0,4327 | 2823 |
| tov $W_{t-1}$ : | | | | | | | |
| 2006 | 1,0014 | 0,2804 | 1,4506 | 0,8969 | 1,0779 | 0,0822 | 121 |
| 2011 | 0,9969 | 0,7564 | 1,1079 | 0,9590 | 1,0385 | 0,0320 | 16 |

Daar staat tegenover dat niet meer voldaan is aan alle randvoorwaarden. De grootste afwijkingen betreffen de samenloopcategorieën markt/ww en markt/vut, zoals tabel 5.8 laat zien. Het alsnog meenemen van deze randtotalen verandert weinig aan de overige afwijkingen,

[28] Zie voetnoot 13.

terwijl de rekentijd zowel in de basis- als in de ramingsperiode met circa één minuut toeneemt. De benodigde rekentijd bij de diverse varianten vormt het onderwerp van de volgende paragraaf.

**Tabel 5.8 Verhouding niet meegenomen kenmerken ten opzichte van randvoorwaarden bij beperkte set**

| Jaar: | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|
| Leeftijd | | | | | | | | | | |
| 18-24 jaar | 1,03 | 1,05 | 1,06 | 1,06 | 1,04 | 1,04 | 1,03 | 1,02 | 1,02 | 1,01 |
| 25-29 jaar | 0,99 | 0,99 | 0,99 | 0,99 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| Hoofdinkomensbron | | | | | | | | | | |
| markt | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| zorg | 1,01 | 1,01 | 1,02 | 1,02 | 1,02 | 1,02 | 1,01 | 1,01 | 1,01 | 1,01 |
| overheid | 0,98 | 0,98 | 0,99 | 0,99 | 0,99 | 0,98 | 0,98 | 0,98 | 0,98 | 0,98 |
| WW | 0,89 | 0,91 | 0,95 | 0,96 | 0,93 | 0,88 | 0,87 | 0,86 | 0,86 | 0,86 |
| WAO | 0,96 | 0,95 | 0,96 | 0,96 | 0,96 | 0,95 | 0,94 | 0,94 | 0,93 | 0,94 |
| bijstand | 0,96 | 0,96 | 0,98 | 0,98 | 0,97 | 0,96 | 0,96 | 0,95 | 0,94 | 0,95 |
| Samenloop | | | | | | | | | | |
| markt/ww | 1,02 | 1,37 | 1,61 | 1,72 | 1,48 | 1,00 | 1,00 | 0,98 | 0,96 | 1,00 |
| markt/vut | 1,03 | 1,05 | 1,08 | 1,09 | 1,19 | 1,28 | 1,31 | 1,32 | 1,33 | 1,27 |
| markt/zelfst | 1,11 | 1,06 | 0,98 | 0,96 | 1,01 | 1,06 | 1,09 | 1,10 | 1,12 | 1,10 |
| markt/over | 1,07 | 1,05 | 0,99 | 0,98 | 1,02 | 1,08 | 1,09 | 1,10 | 1,11 | 1,10 |
| markt/bijst | 1,01 | 1,03 | 1,05 | 1,06 | 1,09 | 1,08 | 1,09 | 1,09 | 1,10 | 1,10 |
| markt/wao | 1,06 | 1,06 | 1,00 | 0,96 | 0,90 | 0,87 | 0,86 | 0,86 | 0,85 | 0,84 |
| wao/vut | 1,04 | 1,01 | 0,98 | 0,95 | 0,96 | 1,00 | 1,00 | 1,00 | 1,00 | 0,98 |
| Eigen woning | 0,93 | 0,92 | 0,90 | 0,89 | 0,88 | 0,88 | 0,87 | 0,85 | 0,84 | 0,83 |
| Aantal tweeverdieners | 1,02 | 1,02 | 1,01 | 1,00 | 1,00 | 0,99 | 0,98 | 0,98 | 0,97 | 0,97 |

## 5.5 Benodigde rekentijd

In tabel 5.9 is een overzicht gegeven van de benodigde rekentijd en het aantal uitgevoerde iteraties. Voor de zojuist besproken varianten is er een nauw verband tussen rekentijd en aantal iteraties: per iteratie is ongeveer ¼ minuut nodig bij de uitgebreide set randtotalen en bijna de helft hiervan bij de beperkte set randtotalen (het aantal kolommen van de datamatrix is dan 31 in plaats van 48). De grootste verschillen doen zich voor in de ramingsperiode (2005-2011) omdat het aantal herwegingen dan het grootst is (zeven maal). De onderdelen I tot en met III hebben betrekking op de uitgebreide set en IV heeft betrekking op de beperkte set. De bovenste twee categorieën hebben betrekking op de zojuist besproken drie varianten. De directe methode

(D) heeft de meeste rekentijd nodig (zes minuten voor de ramingsperiode, met 21 iteraties). Een verklaring hiervoor is dat bij deze variant bij iedere herweging wordt uitgegaan van de oorspronkelijke IPO-gewichten, zodat telkens forse aanpassingen nodig zijn om aan de randtotalen te voldoen. De recursieve methode duurt met relatieve begrenzing (RR) iets langer dan met absolute begrenzing (RA). Bij toepassing van de RR-methode op de set beperkte randvoorwaarden (RRK) halveert in de ramingsperiode de rekentijd, hoewel het aantal iteraties hetzelfde blijft als bij de uitgebreide set (de rekentijd per iteratie ligt lager als gevolg van een kleinere datamatrix).

Ten slotte is onder III het effect weergegeven van een vernauwing van het [L,U]-interval. Bij een relatieve begrenzing van [0,3 , 3] verdrievoudigt de rekentijd. Daarbij kost verlaging van de bovengrens het meest. De in de tabel weergegeven rekentijden zijn redelijk optimaal omdat alle berekeningen in de IML-module van SAS worden uitgevoerd, waarbij de benodigde IPO-dataset (240 920 records) slechts éénmaal in een matrix wordt ingelezen, zodat tijdrovende datasteps worden vermeden. De met * aangegeven methodes zijn als standaardmethodes gekozen en geïmplementeerd. Hiervan is de uitgebreide RR-methode onder I gebruikt voor de uiteindelijke herwegingsoperatie binnen het Deelproject Herweging.

**Tabel 5.9 Benodigde rekentijd en aantal iteraties bij grote en kleine set randvoorwaarden**

| | | | Benodigde Rekentijd (minuten) | | Aantal iteraties |
|---|---|---|---|---|---|
| | Ondergrens L | Bovengrens U | Periode: basis | raming | Periode: raming |
| I Recursieve weging: | | | | | |
| absolute begrenzing (RA) | 0,1 | 10 | 4 | 3 | 8 |
| relatieve begrenzing (RR)* | 0,1 | 10 | 4 | 4 | 14 |
| II Directe methode: | | | | | |
| absolute begrenzing (D) | 0,1 | 10 | 5 | 6 | 21 |
| III Recursieve weging: | | | | | |
| relatieve begrenzing (RR) | 0,3 | 3 | 5 | 12 | 45 |
| relatieve begrenzing (RR) | 0,1 | 3 | 5 | 9 | 36 |
| relatieve begrenzing (RR) | 0,3 | 10 | 5 | 7 | 26 |
| Kleine set randvoorwaarden | | | | | |
| IV Recursieve weging: | | | | | |
| relatieve begrenzing (RRK)* | 0,1 | 10 | 3 | 2 | 14 |

# 6 Conclusie

Het Deelproject Herweging IPO is succesvol afgerond met de oplevering van een drietal SAS-modules waarmee een uitgebreide herwegingsprocedure op het IPO is uitgevoerd voor de periode 2002-2011. De herwegingsprocedure heeft als doel om een aantal demografische en sociaal-economische persoons- en huishoudenskenmerken van het IPO-microdatabestand (240 920 records) te laten sporen met cijfers van het CBS en van afdelingen binnen het CPB. Daartoe is gebruik gemaakt van de methode van *lineair wegen* zoals deze eerder zijn nut heeft bewezen bij de MLT 1999-2002 (op. cit. Waaijers, 1998).

Een eerste verschil met eerdere programmatuur is dat alle code nu is ondergebracht in SAS met gebruikmaking van SAS/IML (eerder was een deel in GAUSS geprogrammeerd). Een tweede verschil is dat de set van randtotalen waarnaar herwogen wordt nu veel uitgebreider is waardoor de kans toeneemt dat de resulterende gewichten niet alle positief zijn. Om dit ongewenste resultaat tegen te gaan is daarom een iteratief algoritme van Huang en Fuller (HF) geïmplementeerd, dat normaliter uitsluitend positieve gewichten oplevert. Het algoritme is ook nuttig om de resulterende gewichten verder te begrenzen: omdat de herweging voor tien achtereenvolgende jaren wordt uitgevoerd en *recursief* wordt herwogen (weging gebaseerd op voorafgaande wegingsresultaten, zodat een relatie wordt gelegd tussen de gewichten van opeenvolgende jaren) zouden zonder beperking de gewichten in later jaren zeer groot of zeer klein kunnen worden. Dit zou tot aanzienlijke vertekening kunnen leiden van resultaten die op dergelijke gewichten berusten (bij kleine subsets uit het IPO). Door de begrenzing te relateren aan de oorspronkelijke IPO-gewichten wordt bereikt dat de resulterende gewichtensets enerzijds relatief dicht bij de oorspronkelijke IPO-gewichten blijven en er anderzijds, door het recursieve karakter van de herweging, zoveel mogelijk verband blijft bestaan tussen de gewichten van opeenvolgende jaren. Hoewel een nauwere begrenzing op zich aantrekkelijk is, heeft dit een prijs in de vorm van een langere rekentijd. De rekentijd is zoveel mogelijk teruggebracht door de feitelijke iteratieve herweging geheel uit te voeren in de IML-module van SAS, waardoor tijdrovende datasteps worden vermeden. Ingeval het algoritme geen convergentie bereikt, wordt door de programmatuur automatisch teruggevallen op een beperktere (vooraf gekozen) set randtotalen waarmee hopelijk wel succesvol kan worden herwogen. Eventueel laat de programmatuur zich eenvoudig wijzigen om per herweging de set randtotalen aan te passen.

Een noodzakelijke voorwaarde in alle gevallen is dat de opgelegde set randtotalen, waaraan de herweging dient te voldoen, intern consistent is. De controle hierop wordt uitgevoerd door een aparte SAS-module die voorafgaand aan de herweging wordt gerund.

Een tweede SAS-module heeft betrekking op werkzaamheden die in principe eenmalig worden uitgevoerd. Hieronder valt de aanmaak van een huishoudensbestand uit het basisbestand met tellingen van de betreffende kenmerken op huishoudniveau. Op het huishoudensbestand worden daarna alle herwegingen uitgevoerd, waarna de resulterende gewichten met het IPO-

huishoudnummer worden weggeschreven naar een apart bestand, dat daarmee gekoppeld kan worden aan andere IPO-bestanden. Alle herwegingen waarbij de randtotalen vast staan omdat ze als gerealiseerd worden beschouwd, worden met deze module uitgevoerd (periode 2002-2004).

De derde SAS-module heeft betrekking op de herweging naar randtotalen die nog het karakter van een raming hebben. Deze module zal in een ramingsperiode regelmatig worden gebruikt (periode 2005-2011).

De modules zijn succesvol getest op uitgebreide sets randtotalen (met 48 categorieën) voor de periode 2002-2011. Het HF-algoritme heeft hier al in de eerste herweging zijn nut bewezen (jaar 2002), waarbij het IPO in lijn wordt gebracht met andere gegevens van CPB en CBS. Slechts 3 iteraties waren voldoende hierbij. Tenslotte zij opgemerkt dat de feitelijke herweging zowel voor de periode 2002-2004 als voor de periode 2005-2011 vier minuten bedroeg (voor de variant met een beperkte set randvoorwaarden met 31 categorieën bedroeg dit respectievelijk drie en twee minuten).

# Bijlage 1

**Controle op randtotalen: melding foute en implausibele waarden**

AANTAL AOW GROTER DAN AANTAL 65+

AANTAL PERSONEN TE KLEIN IN RELATIE TOT AANTALLEN HH-STRUCTUUR

AANTAL STUDENTEN IMPLAUSIBEL TOV AANTALLEN LEEFTIJDSKLASSEN

AANTAL KINDEREN IMPLAUSIBEL TOV AANTALLEN HH-TYPOLOGIE

AANTAL HOOFDBRONINKOMEN WAO GROTER DAN AANTAL WAO

AANTAL HOOFDBRONINKOMEN WW GROTER DAN AANTAL WW

AANTAL HOOFDBRONINKOMEN BIJSTAND GROTER DAN AANTAL BIJSTAND

AANTAL TWEEVERDIENERS GROTER DAN AANTAL PAREN

SAMENLOOP WW GROTER DAN AANTAL WW

SAMENLOOP BIJSTAND GROTER DAN AANTAL BIJSTAND

SAMENLOOP WAO GROTER DAN AANTAL WAO

SAMENLOOP ZELFSTANDIGEN GROTER DAN AANTAL ZELFSTANDIGEN

SAMENLOOP OVERIG ARBEIDSINKOMEN GROTER DAN AANTAL OVERIG ARBEIDSINKOMEN

SAMENLOOP VUT GROTER DAN AANTAL VUT

ARBEIDSJAREN MARKT GROTER DAN AANTAL ARBEIDSWEKEN MARKT

ARBEIDSJAREN ZORG GROTER DAN AANTAL ARBEIDSWEKEN ZORG

ARBEIDSJAREN OVERHEID GROTER DAN AANTAL ARBEIDSWEKEN OVERHEID

SOM WEKENQUOTE HOOFDBRONINKOMEN MARKT GROTER DAN SOM WEKENQUOTE BRONINKOMEN MARKT

SOM WEKENQUOTE HOOFDBRONINKOMEN ZORG GROTER DAN SOM WEKENQUOTE BRONINKOMEN ZORG

SOM WEKENQUOTE HOOFDBRONINKOMEN OVERHEID GROTER DAN SOM WEKENQUOTE BRONINKOMEN OVH

# Bijlage 2

**Verdelingskenmerken van ophooggewichten en correlatie met oorspronkelijke ophooggewichten**

**N = 240 920**

| Jaar | gem | st.afw. | min. | max | 5e Pctl | 25e Pctl | 75e Pctl | 95e Pctl | som | correl |
|---|---|---|---|---|---|---|---|---|---|---|
| IPO-2002 | 0,0672 | 0,0361 | 0,0120 | 0,2699 | 0,0291 | 0,0406 | 0,0832 | 0,1591 | 16193 | 1 |
| na herweging: | | | | | | | | | | |
| 2002 | 0,0672 | 0,0435 | 0,0030 | 0,5652 | 0,0167 | 0,0370 | 0,0868 | 0,1598 | 16193 | 0,8390 |
| 2003 | 0,0675 | 0,0458 | 0,0016 | 0,8735 | 0,0165 | 0,0363 | 0,0866 | 0,1624 | 16258 | 0,8113 |
| 2004 | 0,0677 | 0,0487 | 0,0013 | 1,0956 | 0,0161 | 0,0354 | 0,0860 | 0,1649 | 16306 | 0,7728 |
| 2005 | 0,0678 | 0,0513 | 0,0012 | 1,1367 | 0,0154 | 0,0347 | 0,0855 | 0,1670 | 16339 | 0,7498 |
| 2006 | 0,0680 | 0,0502 | 0,0012 | 0,7074 | 0,0149 | 0,0345 | 0,0861 | 0,1682 | 16372 | 0,7824 |
| 2007 | 0,0681 | 0,0504 | 0,0012 | 0,6092 | 0,0142 | 0,0344 | 0,0873 | 0,1689 | 16405 | 0,7942 |
| 2008 | 0,0682 | 0,0523 | 0,0012 | 0,6390 | 0,0133 | 0,0329 | 0,0875 | 0,1711 | 16439 | 0,7804 |
| 2009 | 0,0684 | 0,0541 | 0,0012 | 0,6822 | 0,0125 | 0,0320 | 0,0873 | 0,1730 | 16474 | 0,7691 |
| 2010 | 0,0685 | 0,0560 | 0,0012 | 0,7309 | 0,0117 | 0,0310 | 0,0873 | 0,1752 | 16509 | 0,7567 |
| 2011 | 0,0687 | 0,0569 | 0,0012 | 0,7504 | 0,0113 | 0,0304 | 0,0871 | 0,1767 | 16545 | 0,7590 |