



## Centraal Bureau voor de Statistiek

Divisie Methodologie en Kwaliteit  
Postbus 4481  
6401 CZ Heerlen  
j.burger@cbs.nl; jgww@cbs.nl; b.buelens@cbs.nl

---

### Weging Woononderzoek Nederland 2009, basismodule Woningmarkt met oversampling

Joep M.S. Burger, José M. Gouweleeuw & Bart Buelens

*Samenvatting: Het Woononderzoek Nederland (WoON) is een statistiek over de woonsituatie en woonwensen van de Nederlandse bevolking in opdracht van het ministerie van VROM. Door selectieve non-respons en oversampling zijn niet alle respondenten even representatief voor de doelpopulatie, wat leidt tot vertekeningen in de schattingen. Een manier om hiervoor te corrigeren is door de waarnemingen te wegen, waarbij rekening wordt gehouden met ongelijke insluitkansen van personen en waarbij hulpinformatie over de doelpopulatie wordt gebruikt. In deze nota adviseren we de CBS sector Statistische analyse personen Heerlen (SAH) over de weging voor de basismodule Woningmarkt van het WoON 2009.*

*Trefwoorden: steekproeftheorie, ophogen, schatten, variantie, vertekening, selectieve non-respons, oversampling, Bascula*

#### 1. Inleiding

De donkerbruine schilderijen die Vincent van Gogh eind negentiende eeuw in Nuenen schilderde bevestigen de kwalitatief benauwde woonsituatie die blijkt uit de woningstatistieken van die tijd (Boonstra et al. 2007, p.129-152). In Nederland begon de woningstatistiek in 1899 met de oprichting van het CBS als aanvulling op

---

Projectnummer: 206499-33  
BPA-nummer: concept versie 8  
Datum: 23 december 2009

de volkstellingen die al vanaf 1795 werden gehouden (Boonstra et al. 2007, p.421-440). Sindsdien tracht ze inzicht te geven in de woonsituatie en woonwensen waarop bestuurders hun beleid kunnen baseren. De statistiek is in de loop van ruim een eeuw sterk uitgebreid en veranderd door de continue maatschappelijke ontwikkelingen. Aanvankelijk werden vragen gesteld over het samen slapen van personen boven de 12 jaar van verschillend geslacht ter behoud van een morele standaard. Dergelijke zedelijkheidsvragen zijn verdwenen en ook vragen over het aantal bedsteden of de aanwezigheid van stromend water zijn niet meer relevant. Tegenwoordig is er behoefte aan gedetailleerdere informatie en is er naast fysieke kwaliteit meer aandacht voor sociale kwaliteit.

Ook de methodologie is door de jaren heen sterk ontwikkeld. In 1964 werden de tienjaarlijkse woningstatistieken, sinds 1919 woningtellingen genoemd, vervangen door het vierjaarlijkse Woningbehoefteonderzoek (WBO). In 1975 werd de steekproef op basis van adressen vervangen door een steekproef onder 18-plussers. In 1977 werden gegevens voor het eerst met behulp van een computer verwerkt. In 1998 werd de mogelijkheid van oversampling geïntroduceerd (Faessen 1998). De enkelvoudig aselechte steekproef werd vervangen door een tweetrapssteekproef onder personen binnen gemeenten. In de twee opeenvolgende jaren is geëxperimenteerd met een kleiner maar jaarlijks WBO, zonder bevredigend resultaat. In 2002 is gestart met computerondersteunende interviews, persoonlijk (CAPI), telefonisch (CATI) of via internet (CAWI), en konden gemeenten kiezen tussen verkorte of volledige vragenlijsten (Meeuwissen et al. 2003; Rigo 2003). Groepen die bekend stonden om hun lage respons werden oververtegenwoordigd in het steekproefontwerp. Het steekproefontwerp werd ook aangepast zodat niet personen maar huishoudens zoveel mogelijk een gelijke insluitkans hadden. De tweetrapssteekproef werd vervangen door een gestratificeerde steekproef met landelijke dekking van gemeenten. In 2006 zijn het WBO en de Kwalitatieve Woningregistratie (de opvolger van het Kwalitatief Woningonderzoek) opgevolgd door het driejaarlijkse Woononderzoek Nederland (WoON) (GfK 2006; Hooft van Huijsduijnen et al. 2007; Meeuwissen & Van Til 2007; Rigo 2007). Het WoON bestaat uit modules waarvoor bij een deel van de respondenten uit de basismodule Woningmarkt vervolgonderzoek wordt gedaan naar specifieke thema's. Sinds het WoON worden ook gegevens uit registers gebruikt.

Enquêteonderzoek levert door verschillende oorzaken fouten op (Bethlehem 2005). Ten eerste is het enquêteonderzoek van het WoON gebaseerd op een steekproef omdat steekproefonderzoek goedkoper, sneller en dus actueler is dan integrale waarneming. Steekproefonderzoek brengt echter selectie- en schattingsfouten met zich mee. Selectiefouten treden op wanneer verkeerde insluitkansen worden gebruikt. Daarnaast zijn er meestal ook een aantal niet-steekproeffouten zoals dekkingsfouten, meet- en verwerkingsfouten en non-respons. Dekkingsfouten treden op wanneer het steekproefkader niet overeenkomt met de doelpopulatie. Non-respons treedt op wanneer personen in een steekproef geen of onbruikbare informatie geven. Dat kan gebeuren omdat het niet lukt contact te maken, omdat de persoon weigert of omdat de persoon niet in staat is mee te werken. Wanneer deze

personen tot een bepaalde groep horen en deze groep afwijkt van de populatie met betrekking tot de doelvariabelen, dan is de non-respons selectief.

Tot het WBO 2002 werden alle personen met gelijke insluitkans getrokken, waardoor grote huishoudens oververtegenwoordigd waren, en slecht responderende groepen ondervertegenwoordigd waren. Sinds het WBO 2002 worden personen in grote huishoudens met een kleinere insluitkans getrokken dan personen in kleine huishoudens. Groepen die bekend staan om hun lage respons krijgen ook een grotere insluitkans. Daarnaast kunnen gemeenten sinds 1998 betalen voor oversampling om uitspraken op lokaal niveau te kunnen doen. Hierdoor hebben personen in oversamplende strata een grotere insluitkans (en een kleiner insluitgewicht) dan personen in overige strata.

Door selectiefouten, dekkingsfouten, selectieve non-respons en oversampling zijn niet alle steekproefpersonen representatief voor de doelpopulatie en dat leidt tot een vertekening van de schattingen. Een effectieve en handige manier om hiervoor te corrigeren is door de waarnemingen te wegen, mits de steekproeven voldoende groot zijn. Bij wegen wordt rekening gehouden met ongelijke insluitkansen van personen en wordt hulpinformatie over de doelpopulatie gebruikt (Bethlehem 2007). Wegen verkleint naast de vertekening ook de variantie van de schattingen. In deze nota adviseren wij de CBS sector Statistische analyse personen Heerlen (SAV) over de weging van de basismodule Woningmarkt van het WoON 2009.

## 2. Methoden

### 2.1 Wegen naar personen

Een schatter voor het totaal van een doelvariabele  $Y$  kan geschreven worden als een gewogen som van de steekproefwaarnemingen  $y_i$  aan  $r$  respondenten (Särndal et al. 1992):

$$\hat{Y} = \sum_{i=1}^r w_i y_i, \quad (1)$$

waarbij  $w_i$  het aantal personen in de doelpopulatie is dat wordt vertegenwoordigd door respondent  $i$ . Dit persoonsgewicht  $w$  wordt in het woononderzoek *weegpers* genoemd. De som van de persoonsgewichten is dus gelijk aan de populatieomvang op 1 januari 2009 van  $N$  18-plussers die deel uitmaken van particuliere huishoudens:

$$\sum_{i=1}^r w_i = N = 12.762.874. \quad (2)$$

De institutionele bevolking behoort dus niet tot de doelpopulatie. Het persoonsgewicht  $w_i$  is het product van een insluitgewicht  $d_{j[i]}$  van persoon  $i$  in steekproefstratum  $j$  en een correctiegewicht  $c_i$ :

$$w_i = d_{j[i]} c_i. \quad (3)$$

Het insluitgewicht  $d_j$  is de reciproke van de insluitkans  $\pi_j$ :

$$d_j = \frac{1}{\pi_j}. \quad (4)$$

De insluitkans is het product van de steekproeffractie  $n_j/N_j$  en de responsfractie  $r_j/n_j$ , wat neerkomt op het aandeel van de doelpopulatie per stratum dat respondeert:

$$\pi_j = \frac{n_j}{N_j} \frac{r_j}{n_j} = \frac{r_j}{N_j}, \quad (5)$$

waarbij  $n_j$  de netto steekproefomvang in stratum  $j$  is (het aantal benaderde personen: de bruto steekproef exclusief kaderfouten),  $N_j$  de populatieomvang in stratum  $j$ , en  $r_j$  de netto responsomvang (het aantal bruikbare responsen) in stratum  $j$ . De insluitkansen van alle personen binnen eenzelfde stratum zijn gelijk. Bij het WoON 2009 bestaat de netto steekproefomvang uit  $n = 179.714$  personen, van wie er  $r = 78.071$  hebben gerespondeerd.

Het insluitgewicht is het aantal personen in de populatie dat wordt vertegenwoordigd door een respondent, aangenomen dat de non-respons niet selectief is. Het correctiegewicht  $c_i$  corrigeert daarnaast voor selectieve non-respons door middel van hulpvariabelen die verondersteld worden te correleren met de doelvariabelen en waarvoor de gegevens op populatieniveau bekend zijn. Om een beeld te krijgen van de selectiviteit in de respons, hebben we de verdeling van verschillende hulpvariabelen in het responsbestand na ophoging vergeleken met de verdeling in de doelpopulatie. De natuurlijke logaritme van de “odds ratio” is een maat voor de relatieve vertegenwoordiging van een klasse (Neter et al. 1996):

$$V = \ln \left( \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \right), \quad (6)$$

waarbij  $V$  de relatieve vertegenwoordiging van een klasse is,  $p_1$  de proportie van respondenten die tot de klasse behoort in het opgehoogde responsbestand (dus al rekening houdend met ongelijke insluitkansen), en  $p_2$  de proportie van personen die tot de klasse behoort in de populatie. Een waarde groter dan 0 betekent dat de klasse is oververtegenwoordigd, en een waarde kleiner dan 0 betekent dat de klasse is ondervertegenwoordigd. De numerieke waarde zelf heeft geen directe betekenis. Stel dat na ophoging 55% van de respondenten vrouw is, terwijl dat in de populatie bijvoorbeeld 49% is. Vrouwen zijn dan oververtegenwoordigd in de steekproef ( $V = 0.24$ ) en krijgen daarom een correctiegewicht kleiner dan 1 ( $49/55 = 0.89$ ), terwijl mannen in dit voorbeeld ondervertegenwoordigd zijn ( $V = -0.24$ ) en daarom een correctiegewicht groter dan 1 krijgen ( $51/45 = 1.13$ ).

De keuze voor het uiteindelijke weegmodel wordt door een aantal factoren bepaald. Ten eerste moeten populatietotalen voor de hulpvariabelen bekend zijn. Ten tweede hebben we ons gericht op de hulpvariabelen waarvoor de non-respons het meest selectief was. Ten derde hebben we gekeken naar de verdeling van de correctiegewichten, waarbij negatieve gewichten en gewichten groter dan 3 ongewenst zijn. Ten vierde hebben we gekeken naar de variatiecoëfficiënt (de standaarddeviatie gedeeld door het gemiddelde) van schatters voor een aantal variabelen waarvan plausible waarden bekend zijn. Meer termen in het weegmodel betekent over het algemeen een kleinere variantie maar ook een onstabielere schatting door minder respondenten per klasse of combinatie van klassen (de zogenaamde celvulling).

De meeste hulpvariabelen en hun populatietotalen komen uit de Gemeentelijke Basisadministratie. Besteedbaar inkomen van het huishouden is afkomstig van de belastingdienst, WOZ-waarde van het Kadaster, en eigendom (koop- of huurwoning) is afgeleid van belastinggegevens en WOZ-waarde. Het CBS weegt met registervariabelen, omdat deze overeenkomen met de stand en definitie van het populatieregister. Op verzoek van VROM hebben we ook gewogen met leeftijd en positie in het huishouden uit de vragenlijst, omdat deze methode in vorige woononderzoeken is toegepast. Verder hebben we de herkomstklasse “onbekend” (109 personen in het register, van wie 4 respondenten) samengevoegd met de grootste herkomstklasse, dat wil zeggen “Nederlands”. De 48 respondenten met onbekende positie in het huishouden hebben we beschouwd als “met partner”. De oude gemeentecodes hebben we vervangen door nieuwe volgens de gemeentelijke indeling 2009 (65 door 1891, 372 door 377, 483 door 1884 en 645 door 1884). De meest recente gegevens over het besteedbaar inkomen van huishoudens betroffen 2008. Omdat we geen populatieaantallen hadden van inkomen, WOZ-waarde en geregistreerd eigendom hebben we deze geschat op basis van de verdeling in de netto steekproef:

$$\hat{N}(x) = \sum_{j=1}^J \frac{n_j(x)}{n_j} N_j, \quad (7)$$

waarbij  $\hat{N}(x)$  het geschatte populatieaantal is voor klasse  $x$  van hulpvariabele  $X$  (inkomen, WOZ-waarde of geregistreerd eigendom). Hierbij nemen we dus aan dat per stratum de verdeling van de hulpvariabele in de netto steekproef representatief is voor de verdeling in de populatie. Dit is aannemelijk als personen binnen strata aselekt zijn getrokken (zie discussie).

De correctiegewichten hebben we bepaald met behulp van lineair wegen en het softwarepakket Bascula (Nieuwenbroek & Boonstra 2001). Varianties van schatters zijn geschat op basis van Taylor linearisatie.

## 2.2 Wegen naar huishoudens

Een schatter voor het totaal van een doelvariabele  $Y$  kan ook geschreven worden als een gewogen som van de steekproefwaarnemingen  $y_i$  aan  $h$  huishoudens:

$$\hat{Y} = \sum_{i=1}^r \frac{w_i}{hhfac_i} y_i, \quad (8)$$

waarbij  $hhfac_i$  het aantal 18-plussers in de huishoudkern is waartoe respondent  $i$  behoort. Respondenten die lid van het huishouden/inwonend zijn (huisvestingssituatie  $hvs = 6$ ) worden hierbij niet meegeteld, omdat zij geen vragen over het huishouden krijgen. In plaats van  $hhfac_i$  kan in Vgl. 7 ook  $ghhfac_i$  worden ingevuld: het aantal 18-plussers in het potentiële huishouden van een verhuisgeneigde respondent ( $verh \neq 4$ ). Hierbij worden leden van huishoudens na verhuizing (gewenste huisvestingssituatie  $ghvs = 4$ ) niet meegeteld. Het huishoudensgewicht  $w/hhfac$  wordt in het woononderzoek *hweegwon* genoemd, en het huishoudensgewicht na verhuizing  $w/ghhfac$  komt overeen met *gweegwon*.

Het aantal 18-plussers in het huishouden ( $hhfac$ ) is als volgt geschat:

$$hhfac = \begin{cases} 1 & \text{als } hhkern < 5 \text{ en } VRlftpa < 18 \\ 2 & \text{als } hhkern < 5 \text{ en } VRlftpa \geq 18 \\ \min(aantalpp, 10) & \text{als } hhkern \geq 5 \end{cases} \quad (9)$$

waarbij  $hhkern$  de samenstelling van het huishouden in zeven klassen is (1 t/m 4 = echtpaar/vaste partner met of zonder kinderen en/of anderen, 5 t/m 6 = één ouder met kinderen met of zonder anderen, 7 = andere samenstelling),  $VRlftpa$  de leeftijd van de partner uit de vragenlijst, en  $aantalpp$  het aantal personen in het huishouden.

Het aantal 18-plussers in het huishouden na verhuizing ( $ghhfac$ ) is als volgt geschat:

$$ghhfac = \begin{cases} hhfac & \text{als } samhhnv = 1 \\ 1 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 1 \\ 2 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 2 \\ 2 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 3 \\ 1 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 4, \\ 0 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 5 \\ 0 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 6 \\ 0 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 7 \\ 2 & \text{als } samhhnv \neq 1 \text{ en } tposhh = 8 \end{cases} \quad (10)$$

waarbij  $samhhnv$  codeert of de samenstelling van het huishouden na verhuizing hetzelfde is (1 = ja, 2 = nee, 8 = weigert, 9 = weet niet), en  $tposhh$  de positie in het huishouden na verhuizing in acht klassen (1 = alleen, 2 = hoofd/partner zonder kinderen, 3 = hoofd/partner met kinderen, 4 = hoofd eenoudergezin, 5 = kind tweeoudergezin, 6 = kind eenoudergezin, 7 = overig lid, 8 = niet-gezinshuishouden).

### 3. Resultaten

#### 3.1 Insluitgewichten

Insluitgewichten  $d_j$  worden bepaald door het steekproefontwerp. Het steekproefontwerp inclusief oversampling bestaat uit  $J = 655$  strata (Fig. 1; Vosmer & Van Berkel 2008). Nederland is opgesplitst in de 31 grootste gemeenten (G31) en de rest. Beide gebieden zijn opgesplitst in een regulier gedeelte en een oversampled gedeelte. Het reguliere gebied dat niet tot de G31 behoort (Fig. 1A) is opgesplitst in COROP-gebieden. Binnen elk COROP-gebied is onderscheid gemaakt tussen personen met een partner en personen zonder een partner. Personen met een partner worden met een twee keer zo kleine kans getrokken als de overige personen, zodat meerpersoonshuishoudens een minder grote insluitkans hebben dan eenpersoonshuishoudens. Alle gemeenten in COROP 13, 14, 16, 21, 26, 27, 31 en 32 behoren tot de G31 en/of zijn oversampled. Dit deel van de steekproef (Fig. 1A) bestaat dus uit  $(40 - 8 \text{ COROP gebieden}) \times 2 \text{ partnerklassen} = 64$  strata.

Het oversampled gebied dat niet tot de G31 behoort (Fig. 1B) bestaat uit 91 gemeenten. Binnen deze gemeenten is onderscheid gemaakt tussen personen met partner en personen zonder partner. De gemeente Maassluis (0556) is aanvullend opgedeeld in drie leeftijdsklassen (18-30 jaar, 31-64 jaar en 65-plussers), en de gemeente Vlaardingen (0622) in elf postcodegebieden. De overige postcodes van Vlaardingen zijn bij postcodegebied 04 gevoegd. De gemeente Rozendaal (niet te verwarren met de gemeente Roosendaal), is zo klein dat het is samengevoegd met de gemeente Rheden (0275). De gemeente Bennebroek (0372) is samengevoegd met de gemeente Bloemendaal (0377) conform de nieuwe gemeentelijke indeling op 1 januari 2009. Dit deel van de steekproef (Fig. 1B) bestaat dus uit  $(89 \text{ standaard gemeenten} + 1 \text{ gemeente Maassluis} \times 3 \text{ leeftijdsklassen} + 1 \text{ gemeente Vlaardingen} \times 11 \text{ postcodegebieden}) \times 2 \text{ partnerklassen} = 206$  strata.

Het reguliere gebied dat wel tot de G31 behoort (Fig. 1C) bestaat uit 18 gemeenten. Deze zijn elk opgesplitst in Vogelaarwijk en/of ISV-gebied enerzijds, en de rest anderzijds. Daarbinnen is onderscheid gemaakt tussen personen met partner en personen zonder partner, en daarbinnen tussen westerse en niet-westerse personen. In Emmen (0114), Hengelo (0164) en Sittard-Geleen (1883) is er binnen het stratum Vogelaarwijk en/of ISV-gebied geen onderscheid gemaakt naar herkomst. In Heerlen (0917) is er bij personen met partner geen onderscheid gemaakt naar herkomst. Dit deel van de steekproef (Fig. 1C) bestaat dus uit  $18 \text{ gemeenten} \times 2 \text{ Vogelaarwijk/ISV-klassen} \times 2 \text{ partnerklassen} \times 2 \text{ herkomstklassen} - 8 \text{ samenvoegingen} = 136$  strata.

Het oversampled gebied dat tot de G31 behoort (Fig. 1D) bestaat uit 14 gemeenten. Almelo (0141), Enschede (0153), Zaanstad (0479), Rotterdam (0599), Maastricht (0935) en Lelystad (0995) zijn net als de reguliere G31 elk opgesplitst naar wel/ geen Vogelaarwijk en/of ISV-gebied, met/zonder partner, en wel/niet westers. Zwolle (0193) en Nijmegen (0268) zijn buiten de zone Vogelaarwijk en/of ISV-gebied

aanvullend opgesplitst in twee postcodegebieden. In Zwolle zijn deze postcodegebieden gebaseerd op 4-cijferige postcodes.

Arnhem (0202) en Utrecht (0344) hebben oversampled in twee postcodegebieden (Fig. 1D). De rest van Arnhem is buiten de zone Vogelaarwijk en/of ISV-gebied bij postcodegebied 02 gevoegd en daarbinnen bij postcodegebied 01. De rest van Utrecht behoort tot de reguliere steekproef (Fig. 1C). Buiten de zone Vogelaarwijk en/of ISV-gebied is in Utrecht ook het postcodegebied 01 bij de reguliere steekproef gevoegd.

Haarlem (0392) is opgesplitst in 14 buurtgebieden, deels gebaseerd op postcodes en deels op buurtcodes (Fig. 1D). Binnen deze buurtgebieden is onderscheid gemaakt naar wel/geen partner en wel/niet westers. Binnen buurtgebied 13 is aanvullend onderscheid gemaakt naar wel/geen Vogelaarwijk en/of ISV-gebied. In buurtgebieden 03, 09 en 14 is geen onderscheid gemaakt naar herkomst. De rest van Haarlem is buiten de zone Vogelaarwijk en/of ISV-gebied bij buurtgebied 02 gevoegd en daarbinnen bij buurtgebied 08. Buiten de zone Vogelaarwijk en/of ISV-gebied zijn buurtgebieden 06 en 08 bij buurtgebied 07 gevoegd. Binnen de zone is buurtgebied 02 bij buurtgebied 01 gevoegd.

Den Haag (0518) is opgesplitst in 11 buurtgebieden, gebaseerd op wijkcodes en enkele buurtcodes (Fig. 1D). Binnen deze buurtgebieden is onderscheid gemaakt naar wel/geen partner en wel/niet westers. In buurtgebieden 03 en 11 is aanvullend onderscheid gemaakt naar wel/geen Vogelaarwijk en/of ISV-gebied. In buurtgebied 11 (overige wijk- en buurtcodes) is ook nog onderscheid gemaakt naar wel/geen particuliere huur en leeftijd (wel/niet jonger dan 25 jaar). Er is geen onderscheid naar herkomst gemaakt bij personen zonder particuliere huur, jonger dan 25 jaar, met partner; bij personen met particulier huur, van 25 jaar of ouder, in een Vogelaarwijk en/of ISV-gebied, met partner; en bij personen met particulier huur, jonger dan 25 jaar, met partner. Buiten de zone Vogelaarwijk en/of ISV-gebied is buurtgebied 02 bij buurtgebied 03 gevoegd, buurtgebied 05 bij 06, en 07 bij 09. Binnen de zone is buurtgebied 06 bij buurtgebied 05 gevoegd, en 09 bij 07.

Leiden (0546) is opgesplitst in drie postcodegebieden, waarbinnen onderscheid is gemaakt naar wel/geen partner en wel/niet westers (Fig. 1D). Binnen postcodegebied 01 is aanvullend onderscheid gemaakt naar wel/geen Vogelaarwijk en/of ISV-gebied. Buiten de zone Vogelaarwijk en/of ISV-gebied is postcodegebied 02 bij postcodegebied 03 gevoegd. Binnen de zone is postcodegebied 03 bij postcodegebied 02 gevoegd.

Schiedam (0606) is oversampled in twee postcodegebieden, gebaseerd op 4-cijferige postcodes (Fig. 1D). Binnen deze postcodegebieden is onderscheid gemaakt naar wel/geen partner en wel/niet westers. Binnen postcodegebied 01 is aanvullend onderscheid gemaakt naar wel/geen Vogelaarwijk en/of ISV-gebied. De rest van Schiedam is bij postcodegebied 01 gevoegd.

Dit deel van de steekproef (Fig. 1D) bestaat dus uit  $65 \text{ strata} \times 2 \text{ partnerklassen} \times 2 \text{ herkomstklassen} - 11 \text{ samenvoelingen} = 249 \text{ strata}$ . Daarmee komt het totaal aantal

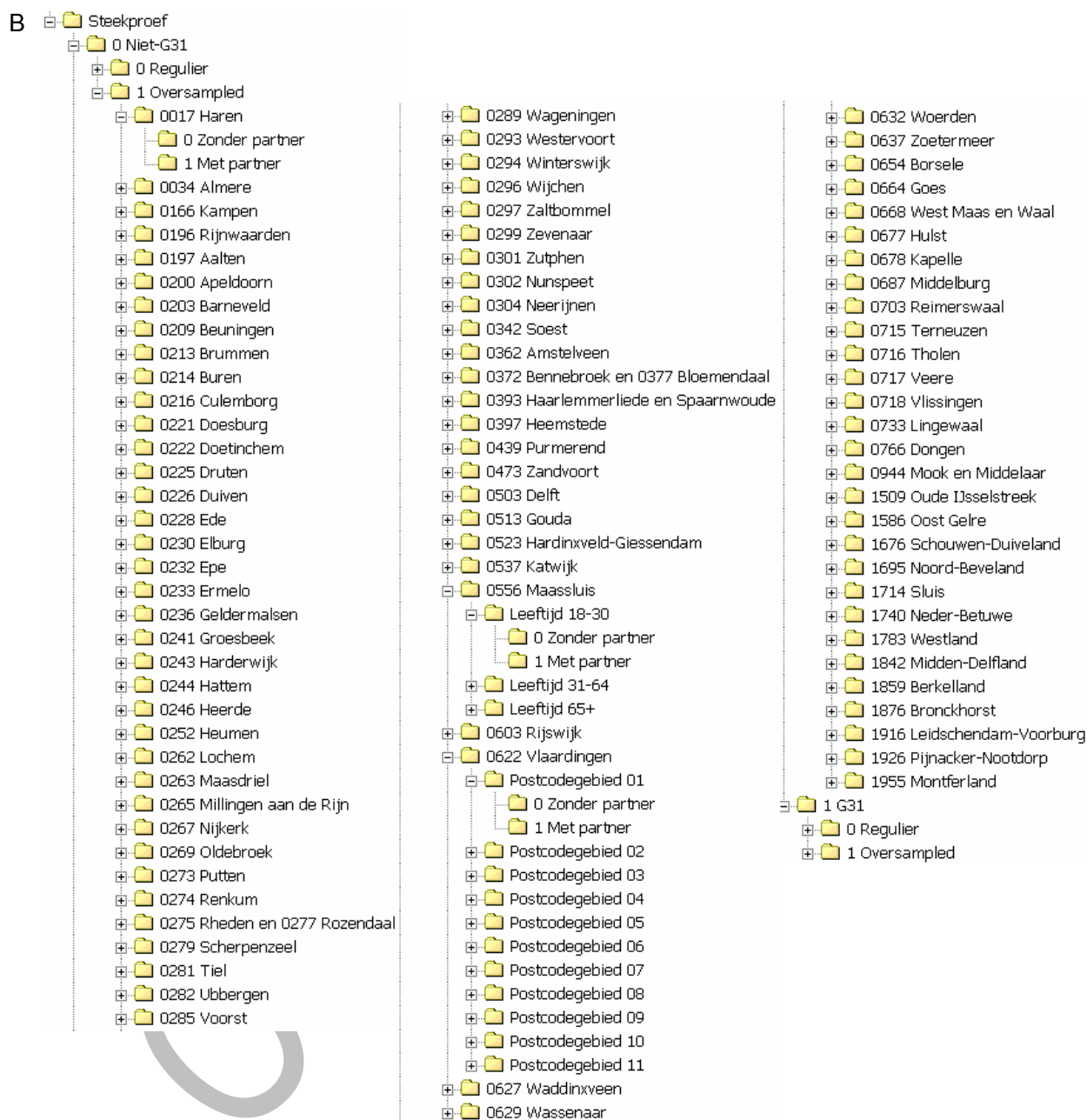


strata in de steekproef op  $J = 64 + 206 + 136 + 249 = 655$ , waarvan  $64 + 136 = 200$  uit de reguliere steekproef en  $206 + 249 = 455$  uit de oversampled steekproef.

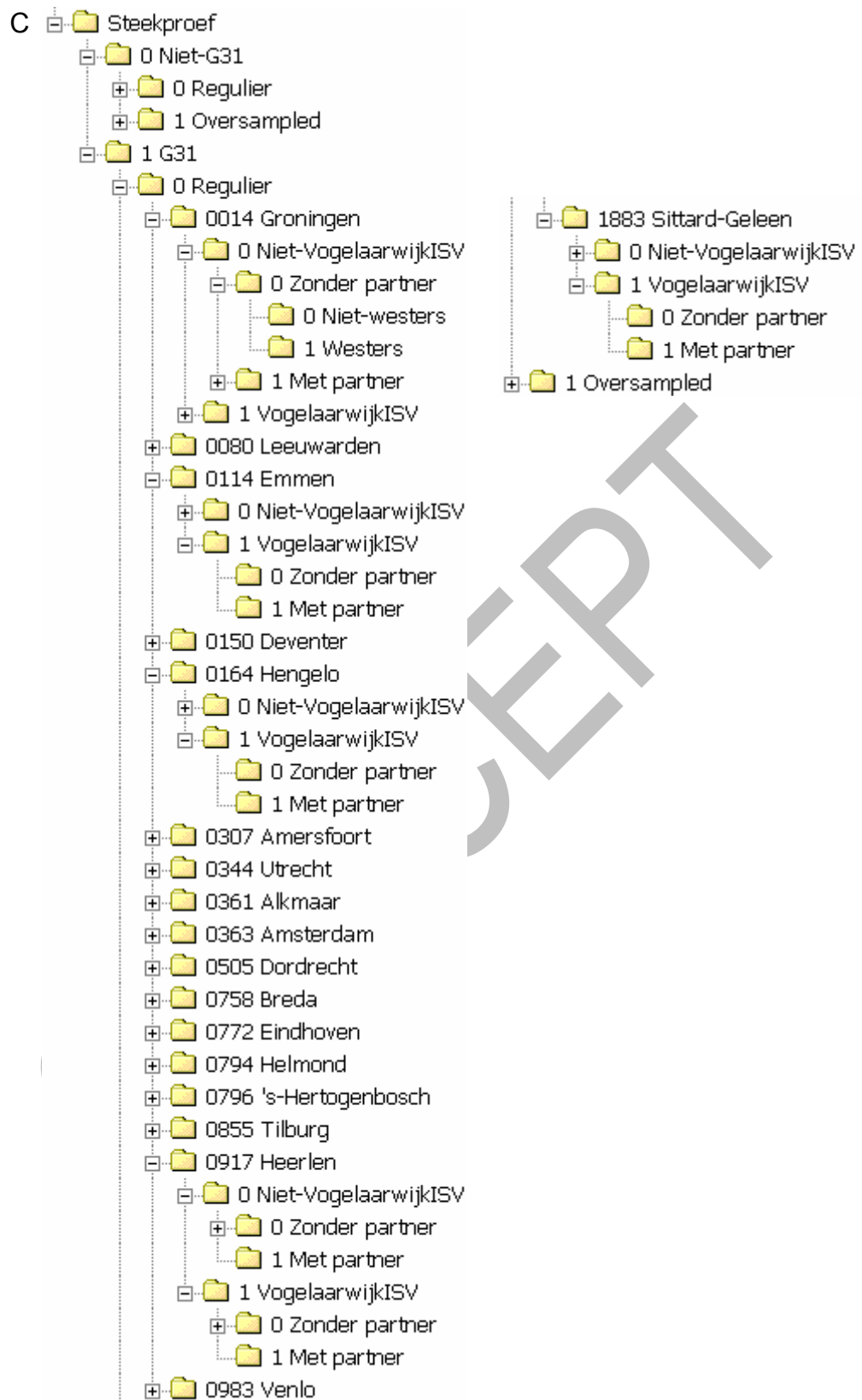


**Figuur 1.** Steekproefontwerp inclusief oversampling. (A) Reguliere steekproef buiten de G31, (B) Oversampled steekproef buiten de G31, (C) Reguliere steekproef binnen de G31, en (D) Oversampled steekproef binnen de G31. Een ‘+’ voor een knoop in de hiërarchie betekent dat de onderliggende structuur verborgen is want gelijk aan

de onderliggende structuur van de eerste knoop op hetzelfde hiërarchische niveau.



**Figuur 1. Vervolgd.**

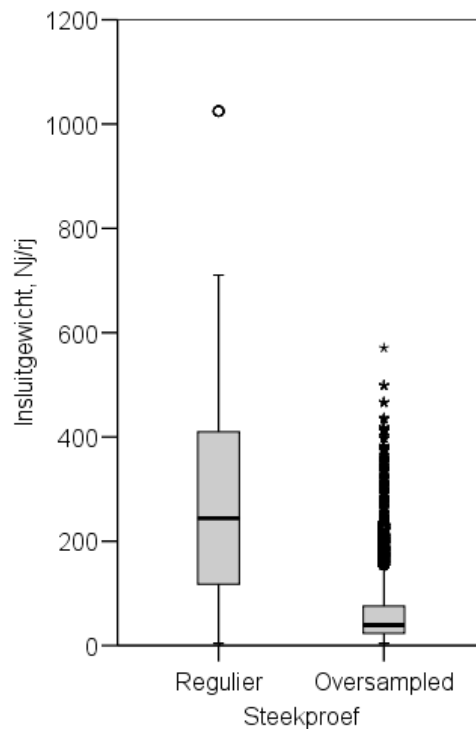


**Figuur 1.** Vervolgd.



Figuur 1. Vervolgd.

Het reguliere deel van de steekproef bevat 40.752 respondenten. Eén respondent vertegenwoordigt vóór correctie tussen de 4 en 1025 personen, met een mediaan van 244 (Fig. 2). In het oversampled deel, met 37.319 respondenten, vertegenwoordigt een respondent vóór correctie tussen de 4 en 571 personen met een mediaan van 39. De grote spreiding wordt niet alleen veroorzaakt door spreiding in responsfractie  $r_j/n_j$  maar ook door spreiding in steekproeffractie  $n_j/N_j$  omdat de gewenste responsaantallen naar regionale kenmerken niet constant zijn. De oversampled steekproef bevat per definitie een grotere proportie van de populatie, waardoor de insluitgewichten lager zijn.



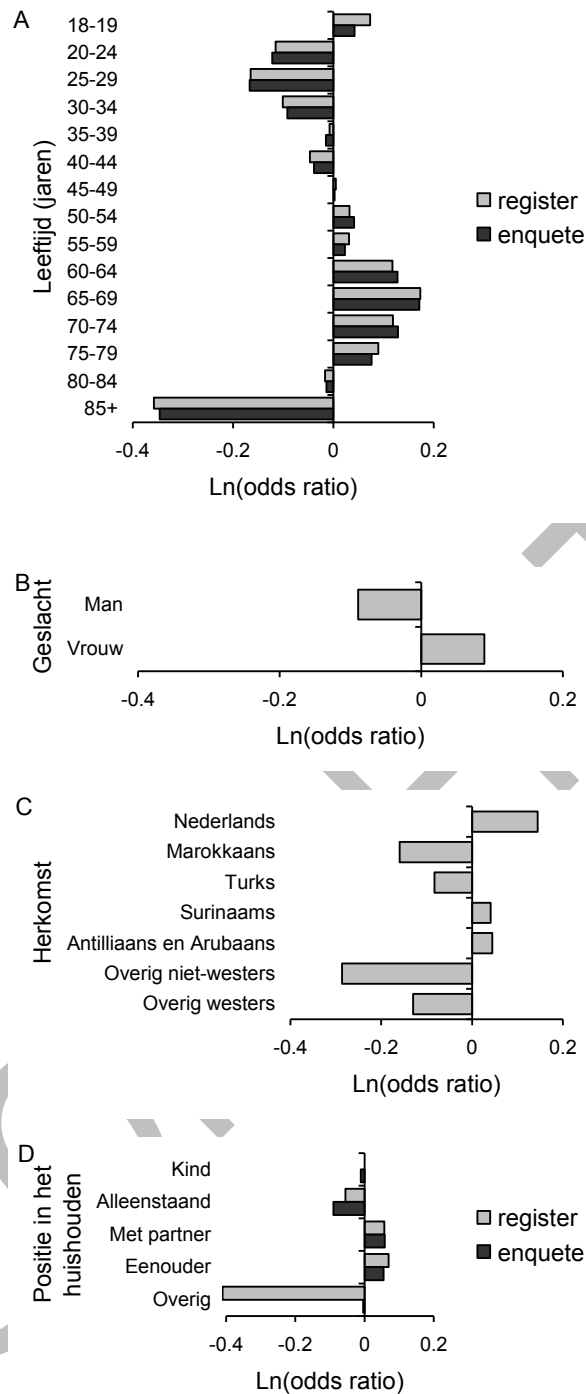
**Figuur 2.** Verdeling van insluitgewichten  $d_j = N_j/r_j$  in reguliere steekproef (40.752 respondenten) en oversampled steekproef (37.319 respondenten).

### 3.2 Selectiviteit in respons

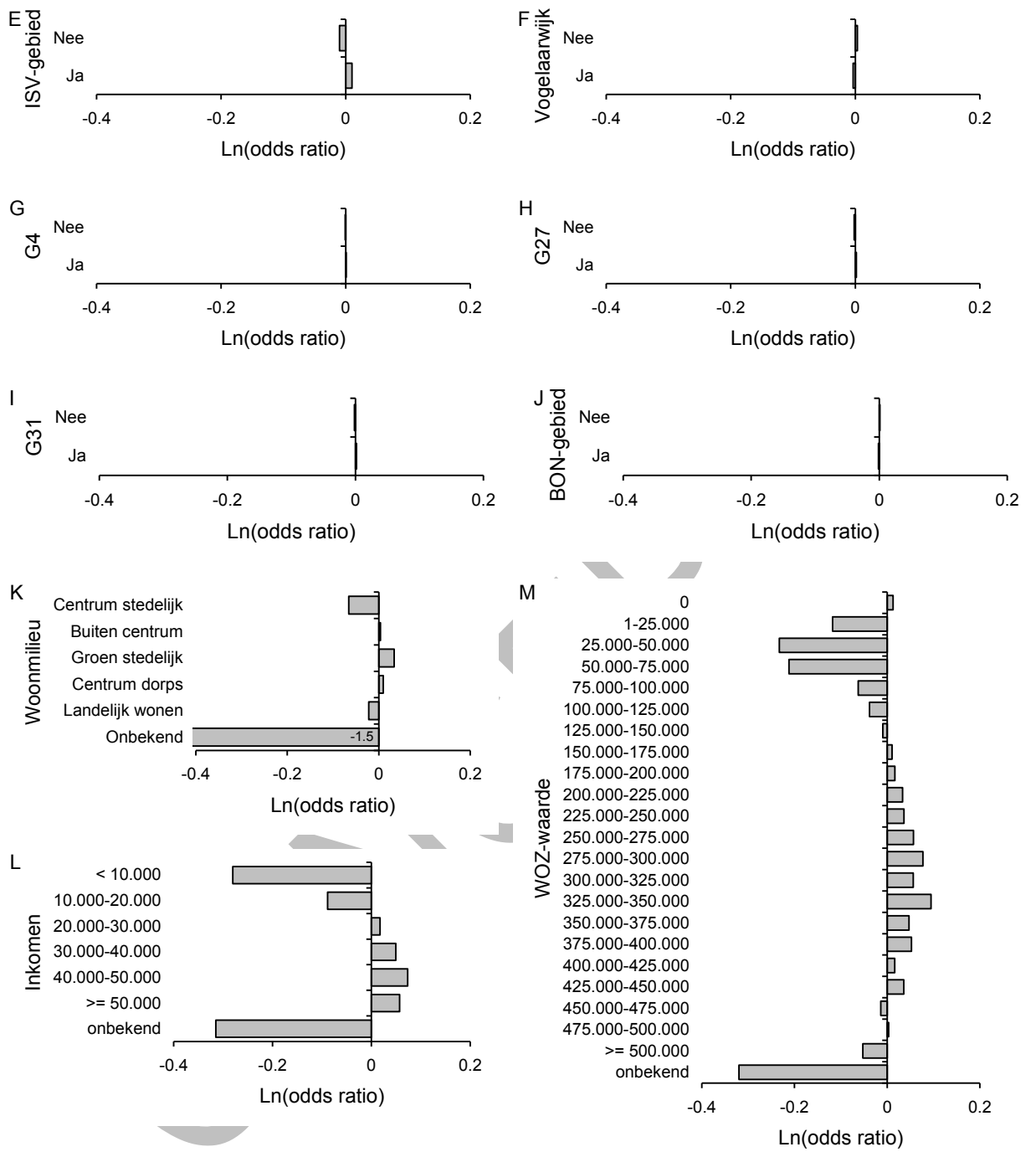
Om een beeld te krijgen van de selectiviteit in de respons, hebben we de verdeling van een hulpvariabele in het opgehoogde responsbestand vergeleken met de verdeling in het populatieregister (Fig. 3; Vgl. 6). Hierbij is dus al rekening gehouden met de verschillende insluitkansen. Voor de eenvoud hierbij hebben we ons in Fig. 3 beperkt tot de hoofdeffecten, hoewel daartussen veel interacties mogelijk zijn. De selectiviteit kan bijvoorbeeld regionaal anders zijn dan het hier gepresenteerde landelijke niveau, en de selectiviteit naar leeftijd kan bijvoorbeeld verschillen tussen mannen en vrouwen. Daarnaast zijn deze effecten correlaties en geen oorzakelijke verbanden.

In het opgehoogde responsbestand zijn vooral de 20- tot en met 34-jarigen de 85-plussers ondervertegenwoordigd, terwijl vooral de 60- tot en met 79-jarigen juist

oververtegenwoordigd zijn (Fig. 3A). Ook de 18- en 19-jarigen zijn iets oververtegenwoordigd. Voor de selectiviteit naar leeftijd maakt het (gelukkig) niet veel uit of de registervariabele op peildatum 1 januari 2009 wordt gebruikt of het antwoord van de respondent op de enquêtedatum. Verder zijn vrouwen iets beter vertegenwoordigd dan mannen (Fig. 3B). Nederlanders, Surinamers, en Antillianen en Arubanen zijn ook oververtegenwoordigd, terwijl andere herkomstklassen ondervertegenwoordigd zijn (Fig. 3C). Voor de selectiviteit naar positie in het huishouden maakt het wel uit of de registervariabele of de enquêtevariabele wordt vergeleken met het register (Fig. 3D). Beide variabelen geven aan dat personen met partner en in eenoudergezinnen oververtegenwoordigd zijn, en dat alleenstaanden ondervertegenwoordigd zijn. Volgens de registervariabele is de restklasse echter sterk ondervertegenwoordigd, terwijl deze klasse volgens de enquêtevariabele juist evenwichtig vertegenwoordigd is. De ISV-gebieden zijn op zichzelf beschouwd evenwichtig vertegenwoordigd (Fig. 3E), net als de Vogelaarwijken (Fig. 3F), de 4 grootste gemeenten (Fig. 3G), de 31 grootste gemeenten zonder de 4 grootste (Fig. 3H), de 31 grootste gemeenten (Fig. 3I), en de BON-gebieden (Fig. 3J). Sterk ondervertegenwoordigd zijn personen waarvan het woonmilieu nog niet bekend is ( $\ln(\text{OR}) = -1,5$ ) en in mindere mate ook personen in een centrum stedelijk woonmilieu (Fig. 3K). De laagste inkomenklassen en personen van wie het besteedbaar huishoudinkomen onbekend is zijn ondervertegenwoordigd, terwijl het opgehoogde responsbestand te veel hoge inkomenklassen bevat (Fig. 3L). Personen in woningen met een lage, zeer hoge of onbekende WOZ-waarde zijn ondervertegenwoordigd, terwijl personen in woningen van 150 tot 450 duizend euro oververtegenwoordigd zijn (Fig. 3M). De COROP-gebieden zijn evenwichtig vertegenwoordigd (Fig. 3N).



**Figuur 3.** Selectiviteit van de respons. Relatieve vertegenwoordiging (Vgl. 6) van een respondent in het opgehoogde responsbestand ten opzichte van het populatiebestand per klasse van potentiële hulpvariabele. (A) Leeftijd, (B) Geslacht, (C) Herkomst, (D) Positie in het huishouden, (E) ISV-gebied, (F) Vogelaarwijk, (G) G4, (H) G31 exclusief G4, (I) G31, (J) BON-gebied, (K) Woonmilieu, (L) Besteedbaar inkomen van het huishouden 2008 (euro per jaar), (M) WOZ-waarde (euro), en (N) COROP-gebied. Voor leeftijd en positie in het huishouden zijn naast registervariabelen op verzoek van de opdrachtgever ook enquêtevariabelen vergeleken.



**Figuur 3. Vervolgd.**





**Figuur 3. Vervolgd.**

**Tabel 1** Onderzochte weegmodellen.

Model	Weegtermen (aantal klassen)	Celvulling	Persoonsgewicht $w_{ij}$
m0	Constante: $c_i = 1$	-	$N/r = 12.762.874/78.071 = 163$
m1	Constante: $c_i = 1$	-	$N_j/r_j$ ( $J = 655$ strata)
m2	Oversampled ( $2^A$ ) × Leeftijd (15) × Geslacht (2) + Oversampled (2) × Leeftijd ( $7^B$ ) × Geslacht (2) + Oversampled (2) × Leeftijd (7) × Herkomst (7) + Oversampled (2) × Geslacht (2) × Herkomst (7) + Oversampled (2) × Geslacht (2) × COROP (40) + Leeftijd (7) × Geslacht (2) × Herkomst (7) + Leeftijd (7) × Geslacht (2) × COROP (40) + Herkomst (7) × Landsdeel (4) + Inkomen (7) × Provincie (12) × Eigendom ( $3^C$ ) + WOZ (23) × Provincie (12) + Gemeente ( $123^D$ ) + Leeftijd ( $11^E$ ) × Positie (5) + Woonmilieu (6)	$\geq 1$	$(N_j/r_j)c_i$
m3	Oversampled (2) × Leeftijd (15) × Geslacht (2) + Oversampled (2) × Geslacht (2) × Herkomst (7) + Leeftijd (7) × Herkomst (7) + Leeftijd (7) × COROP (40) + Geslacht (2) × COROP (40) + Herkomst (7) × Landsdeel (4) + Inkomen (7) × Provincie (12) + WOZ (23) + Gemeente (123) + Leeftijd (11) × Positie (5) + Woonmilieu (6)	$\geq 4$	$(N_j/r_j)c_i$
m4	Oversampled (2) × Leeftijd (15) × Geslacht (2) + Oversampled (2) × Geslacht (2) × Herkomst (7) + Leeftijd (7) × Herkomst (7) + Leeftijd (7) × COROP (40) + Geslacht (2) × COROP (40) + Herkomst (7) × Landsdeel (4) + Inkomen (7) × Provincie (12) × <b>Eigendom (3)</b> + WOZ (23) + Gemeente (123) + Leeftijd (11) × Positie (5) + Woonmilieu (6)	$\geq 4$	$(N_j/r_j)c_i$
m5	Als m4 met Leeftijd en Positie uit enquête in plaats van uit register; begrenzingsalgoritme toegepast tegen negatieve gewichten	$\geq 5$	$(N_j/r_j)c_i$

<sup>A</sup>Reguliere of oversamplende gemeente.

<sup>B</sup>18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+ jaar.

<sup>C</sup>Koopwoning, huurwoning of onbekend.

<sup>D</sup>31 grootste gemeenten (G31) + 91 oversamplende gemeenten buiten de G31 + 1 restklasse = 123 klassen.

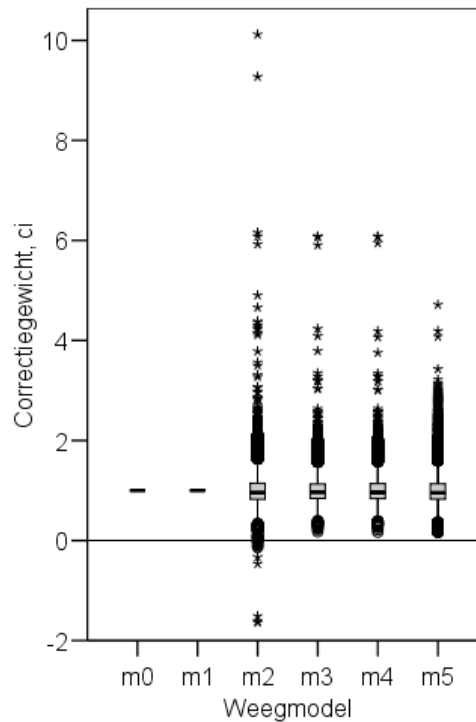
<sup>E</sup>18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+ jaar.

### 3.3 Weegmodellen vergeleken

In deze paragraaf bespreken we het effect van verschillende weegmodellen op de spreiding van de correctiegewichten, op de schattingen van een aantal variabelen en op de variatiecoëfficiënten van de schattingen. Uit de vorige paragraaf is gebleken dat door het complexe steekproefontwerp het opgehoogde responsbestand in een aantal opzichten al vrij goed overeen kwam met het populatiebestand. De variabelen waar de non-respons het meest selectief was zaten (vrijwel) niet in het steekproefontwerp (leeftijd, geslacht, inkomen, WOZ) of te eenvoudig (herkomst: westers/niet-westers, en positie in het huishouden: met/zonder partner). Op basis hiervan hebben we vijf weegmodellen vergeleken (Tabel 1).

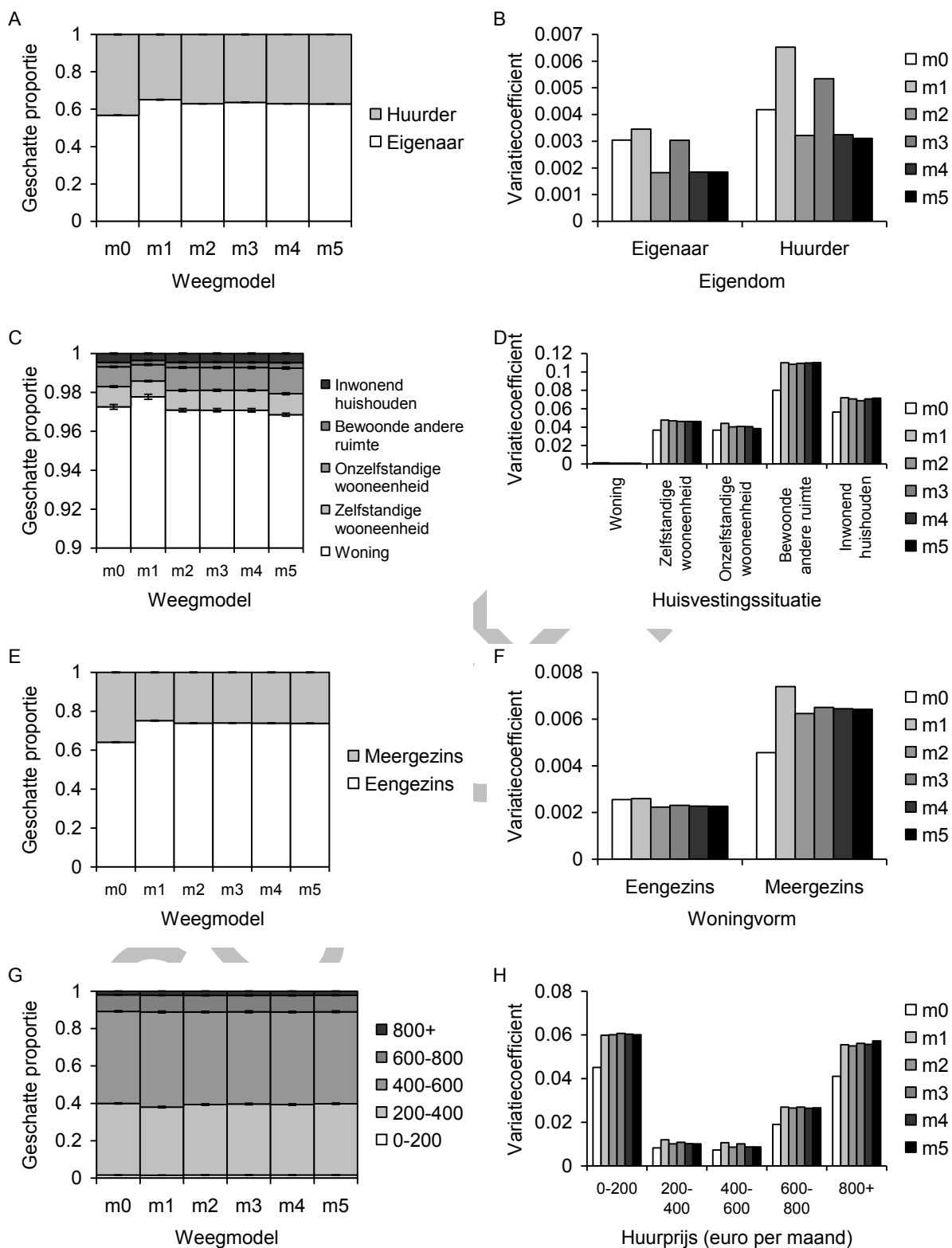
Het eenvoudigste model (m0) is een ophoging zonder rekening te houden met het steekproefontwerp. Het tweede model (m1) is een ophoging per steekproefstratum. De andere modellen corrigeren daarnaast voor selectieve non-respons. Model m2 is het maximale weegmodel op voorwaarde van minimaal 1 waarneming per combinatie van klassen. Model m3 is een vereenvoudiging van model m2 door het aantal waarnemingen per weegterm te vergroten om negatieve gewichten te voorkomen. Model m4 bevat op verzoek van de opdrachtgever aanvullend eigendomsgegevens gebaseerd op de belastingsaangifte en WOZ-waarde. Model m5 gebruikt op verzoek van de opdrachtgever leeftijd en positie in het huishouden uit de enquête als hulpvariabelen in plaats van uit het register, omdat dit overeenkomt met de methode van 1998, 2002 en 2006. Er is dus één weegmodel voor het gehele bestand inclusief de oversampled steekproef. Het voordeel hiervan is dat lokale cijfers optellen tot regionale en landelijke cijfers. Er is wel rekening gehouden met het verschil tussen de reguliere en oversampled steekproef door zoveel mogelijk variabelen te kruisen met de variabele *Oversampled*.

Het correctiegewicht van model m0 en m1 is per definitie gelijk aan 1 (Fig. 4). Correctie voor selectieve non-respons met het maximale weegmodel (m2) levert een grote spreiding van correctiegewichten, waarbij ook negatieve gewichten ontstaan. Dit kan worden voorkomen door meer waarnemingen per combinatie van klassen te eisen (m3). Het toevoegen van eigendomsgegevens (m4) heeft weinig effect op de spreiding van de correctiegewichten. Het gebruik van leeftijd en positie in het huishouden uit de enquête in plaats van het register (m5) resulteert zonder begrenzingsalgoritme in 10 negatieve gewichten. Na begrenzing is de verdeling iets minder leptokurtisch dan model m4 (minder hoge piek, minder dunne staarten).

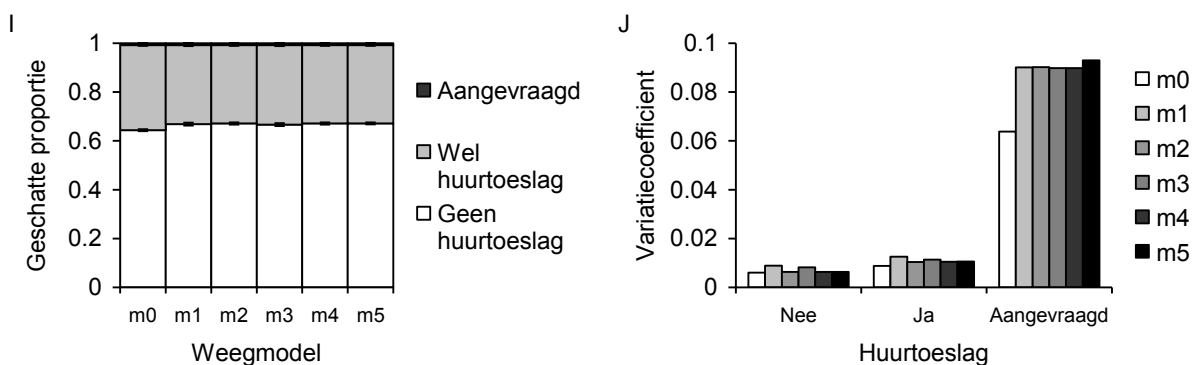


**Figuur 4.** Effect van weegmodel op de verdeling van correctiegewichten.

Het gebruik van insluitgewichten per stratum (model m1) levert aanzienlijk andere landelijke schattingen op dan het gebruik van een enkel insluitgewicht (model m0) (Fig. 5), met name in de geschatte proportie woningeigenaren (Fig. 5A) en eengezinswoningen (Fig. 5E). Hoewel schattingen op basis van model m1 per definitie realistischer zijn, nemen de variatiecoëfficiënten meestal echter toe. Vergeleken met het toevoegen van de steekproefstrata, zijn landelijke schattingen relatief ongevoelig voor het gebruik van correctiegewichten (m2 t/m m5). Gezien de kleine standaardfouten en grote steekproefomvang kunnen kleine verschillen in uitkomsten wel significant zijn. Het toevoegen van het maximale weegmodel verkleint in een aantal gevallen de variatiecoëfficiënten (m2 ten opzichte van m1). De vereenvoudiging van dit model om negatieve gewichten te voorkomen heeft weinig effect op de schattingen (m4 ten opzichte van m2). Uiteraard nemen de variatiecoëfficiënten van de eigendomsvariabele toe met het verwijderden van registergegevens die nauw met deze variabele samenhangen (Fig. 5B, m3 ten opzichte van m4). Het gebruik van enquêtevariabelen in plaats van registervariabelen heeft ook weinig effect op de schattingen (m5 ten opzichte van m4). Merk overigens op dat de hier gepresenteerde schattingen op persoonsgewichten zijn gebaseerd, en niet op huishoudensgewichten. Het is namelijk niet triviaal hoe standaardfouten berekend moeten worden van schattingen gebaseerd op huishoudensgewichten. Het gaat hier echter om het relatieve effect van het weegmodel en niet om de absolute waarden van de schattingen. Verder kunnen de verschillen natuurlijk groter zijn op lokaal niveau of voor andere variabelen. Een uitgebreide plausibiliteitsanalyse is door de sector SAH (2009) beschreven.



**Figuur 5.** Effect van weegmodel op schattingen van enkele doelvariabelen (linker panelen; gemiddelde  $\pm$  SD) en variatiecoëfficiënten (rechter panelen). (A,B) eigendom, (C,D) huisvestingssituatie (let op afwijkende y-as in C), (E,F) woningvorm, (G,H) huurprijs, (I,J) Huurtoeslag.



**Figuur 5.** Vervolgd.

### 3.4 Eindgewichten

Model m4 bevat de meeste hulpinformatie zonder negatieve correctiegewichten en is vanuit methodologisch oogpunt logischer dan model m5, omdat de hulpvariabelen in het weegmodel dezelfde variabelen zijn als waarmee de populatieverdeling is bepaald. Toch is op verzoek van de opdrachtgever model m5 geleverd omdat dit beter aansluit bij de methode zoals die in eerdere woononderzoeken is gebruikt. Het

totaal aantal huishoudens ( $\sum_{i=1}^r w_i / hhfac_i$ ) komt op 7,32 miljoen. Om dit consistent

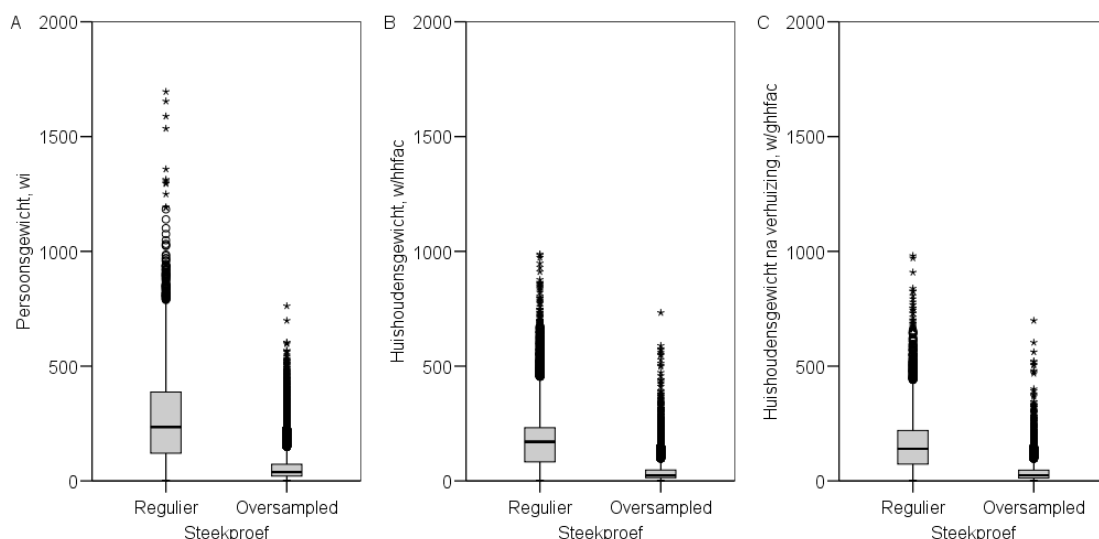
te maken met de huishoudensstatistiek (7,31 miljoen huishoudens), zijn de huishoudensgewichten herwogen naar het aantal huishoudens per gemeente volgens Statline, waarbij reguliere gemeenten buiten de G31 zijn samengevoegd. Deze correctiegewichten variëren van 0,88 tot 1,14 met een mediaan van 0,99. Het totaal aantal huishoudens van verhuiscandidate respondenten na verhuizing

( $\sum_{i=1}^r w_i / ghhfac_i$ ) komt op 2.142.147.

Na correctie van de insluitgewichten met model m5 vertegenwoordigt een respondent in de reguliere steekproef (40.752 respondenten) tussen de 3 en 1696 personen met een mediaan van 235, en in de oversampled steekproef (37.319 respondenten) tussen de 2 en 762 personen met een mediaan van 38 (Fig. 6A).

Een respondent die vragen over het huishouden heeft gekregen vertegenwoordigt in de reguliere steekproef (35.626 respondenten) tussen de 2 en 988 huishoudens met een mediaan van 171, en in de oversampled steekproef (33.523 respondenten) tussen de 1 en 733 huishoudens met een mediaan van 24 (Fig. 6B).

Een verhuiscandidate respondent ( $verh \neq 4$ ) die geen lid van het huishouden wordt ( $ghvs \neq 4$ ) vertegenwoordigt in de reguliere steekproef (10.606 respondenten) tussen de 2 en 981 potentiële huishoudens met een mediaan van 140, en in de oversampled steekproef (11.954 respondenten) tussen de 2 en 698 potentiële huishoudens met een mediaan van 25 (Fig. 6C).



**Figuur 6.** Verdeling van eindgewichten in de reguliere en oversampled steekproef. (A) Persoonsgewicht  $w_i$ , (B) Huishoudensgewicht  $w_i/hhfac_i$ , en (C) Gewicht potentiële huishoudens  $w_i/ghhfac_i$ , waarbij  $hhfac_i$  het aantal 18-plussers is in de huishoudkern van respondent  $i$ , en  $ghhfac_i$  het aantal 18-plussers in het potentiële huishouden van verhuisgeneigde respondent  $i$ .

#### 4. Discussie

In deze nota hebben we beschreven hoe we zijn gekomen tot de geleverde insluitgewichten op basis van het steekproefontwerp; de correctiegewichten op basis van hun spreiding en hun effect op schattingen en variatiecoëfficiënten; en de eindgewichten voor personen, huishoudens en potentiële huishoudens. Een aantal punten dient bij gebruik van het geleverde weegmodel in acht te worden genomen. Ten eerste is wegen gebaseerd op de aanname dat doelvariabelen met hulpvariabelen correleren. Niet alle doelvariabelen zullen echter even sterk correleren met alle hulpvariabelen, waardoor het beste weegmodel afhankelijk kan zijn van de doelvariabele. Ten tweede is het beste weegmodel niet te bepalen, omdat er geen objectief criterium voor is en omdat het aantal theoretisch mogelijke weegmodellen exponentieel toeneemt met het aantal hulpvariabelen. Ten derde is de steekproef binnen strata niet helemaal aselect. De respons viel lager uit dan verwacht, waardoor gedurende het veldwerk steekproefpersonen selectief zijn bijgetrokken. Daarnaast hebben niet alle respondenten via dezelfde waarneemmethode gerepondeerd, terwijl waarneemmethode (CAWI/CATI/CAPI) wel een effect kan hebben op de selectiviteit van de respons en op de antwoorden. Tot slot hebben we gekozen om één weegmodel te ontwikkelen voor het gehele responsbestand inclusief oversampling. Dit heeft als voordeel dat lokale cijfers optellen tot landelijke cijfers. Nadeel is dat het weegmodel suboptimaal kan zijn voor de afzonderlijke oversampled gebieden.

## 5. Dankwoord

Met dank aan Saskia Janssen-Jansen, Coen van Heukelingen, Jos Beckers, Harm-Jan Boonstra, Anton Kantebeen, Roelf-Jan van Til, Remi de Cock, Floor Langendijk, Peter Meuwissen, Frank Meessen, Koos Arts, en Frans Brepoels voor logistieke ondersteuning en/of inhoudelijke discussie.

## 6. Referenties

- Bethlehem, J., 2005. Internet survey developments at Statistics Netherlands. CBS, Voorburg.
- Bethlehem, J., 2007. Methodenreeks: wegen als correctie voor non-respons. CBS, Voorburg.
- Boonstra, O.W.A., P.K. Doorn, M.P.M. van Horik, J.G.S.J. van Maarseveen & J. Oudhof (red.), 2007. Twee eeuwen Nederland geteld. Onderzoek met de digitale Volks-, Beroeps- en Woningtellingen 1795-2001. DANS, Den Haag.
- Buelens, B & J. van den Brakel, 2009. Weging Integrale Veiligheidsmonitor 2008. BPA-nummer DMH-2009-04-15-BBUS. CBS, Heerlen.
- Faessen, W., 1998. WoningBehoeftte Onderzoek 1998. Onderzoeksdocumentatie. CBS, Voorburg.
- GfK Panel Services Benelux, 2006. Onderzoeksverantwoording Woon Onderzoek Nederland 2006 Module Woningmarkt. GfK, Dongen.
- Hooft van Huijsduijnen, J., R.-J. van Til, E. Verhoog, K. Gopal, B. Ferment & J. van Galen, 2007. WoON 2006. Onderzoeksdocumentatie module Woningmarkt. ABF Research, Delft.
- Meeuwissen, B., B. Ferment, J. Hooft van Huijsduijnen & R.J. van Til, 2003. WoningBehoeftte Onderzoek 2002. Onderzoeksdocumentatie. ABF Research, Delft.
- Meeuwissen, B. & R.-J. van Til, 2007. Steekproefstructuur Woononderzoek Nederland 2006. ABF Research, Delft.
- Neter, J., M.H. Kutner, C.J. Nachtsheim & W. Wasserman, 1996. Applied Linear Statistical Models. Fourth edition. McGraw-Hill, Boston.
- Nieuwenbroek, N. & H.J. Boonstra, 2001. Bascula 4.0 reference manual. BPA 3554-99-RSM. CBS, Heerlen.
- Rigo Research en Advies, 2003. Beter thuis in wonen. Kernpublicatie WoningBehoeftte Onderzoek 2002. Ministerie van VROM, Den Haag.
- Rigo Research en Advies, 2007. Wonen op een rijtje. De resultaten van het Woononderzoek Nederland 2006. Ministerie van VROM, Den Haag.



SAH, 2009. Dataprocessing WoON 2009. Plausibiliteitcontroles versie 1.4, 17 december 2009. CBS, Heerlen.

Särndal, C.-E., B. Swensson & J. Wretman, 1992. Model Assisted Survey Sampling. Springer, New York.

Vosmer, M. & K. van Berkel, 2008. Onderzoeksdesign dataverzameling Woningmarktmodule WoON 2009. CBS memo, versie 1.1, 24 juni 2008, BPA-nummer SDV-2008-H68. CBS, Heerlen.

CONCEPT