

DE VIRTUELE VOLKSTELLING EN HET SOCIAAL STATISTISCH BESTAND

Een verslag van de conferentie gehouden in Amsterdam op 11 november 2003

Redactie
Bart Bakker (CBS)
Lisa Putman (SISWO)

Inhoudsopgave

Voorwoord	5
Inleiding	7
Volkstellingen. De betekenis voor historisch sociaal-wetenschappelijk onderzoek J. van Maarseveen	11
Het belang van een éénprocentsteekproef uit de huishoudens van de virtuele volkstelling J. Dronkers	25
De 'virtuele volkstelling' als databron voor sociaal-wetenschappelijk onderzoek; optimaliseringsvragen Een reactie op de papers van J. van Maarseveen en J. Dronkers J. Gierveld	31
De 'virtuele volkstelling' 2001 E. Schulte Nordholt	33
De 'virtuele Volkstelling' is geen kopie van een traditionele Volkstelling Een reactie op het paper van E. Schulte Nordholt C. de Hoog	43
Het Sociaal Statistisch Bestand (SSB): de statistiek van de toekomst P. Everaers	47
SSB: wensdroom of nachtmerrie Een reactie op het paper van P. Everaers R. van der Velden	57
Dynamiek op de arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand B. Bakker en K. Arts	59
CBS wees zuinig op de enquêtes! Een reactie op het paper van B. Bakker en K. Arts P. de Beer	73

Loonstructuuronderzoek – verrijkte data door integratie E. Schulte Nordholt	75
LSO is een mooi bestand, maar... Een reactie op het paper van E. Schulte Nordholt J. Hartog	81
Ruimtelijke spreiding van mensen en hun kenmerken Zicht op autochtonen en allochtonen J. Latten	83
Nog geen zicht op allochtonen Een reactie op het paper van J. Latten P. Tesser	98
De deur van het demografisch laboratorium Perspectieven voor externe benutting van de virtuele volkstelling 2001 J. Nobel	101
Hoe ver gaat de deur van het CBS open? Een reactie op het paper van J. Nobel K. Tijdens	121
Discussie	123
Gebruikte afkortingen	127
Deelnemers	128
Auteurs	129

Voorwoord

Eind 2003 verstrekte het CBS als één van de eerste lidstaten van de Europese Unie de gegevens over de zeventiende Nederlandse Volkstelling aan het Europese statistiekbureau Eurostat. Bijzonder was dat het CBS deze gegevens kon leveren zonder dat daar ook maar één persoon speciaal voor hoefde te worden geënquêteerd. Door gebruik te maken van registers die al beschikbaar zijn voor andere dan statistische doeleinden en deze te combineren met enquêtes werd een virtuele Volkstelling voor het jaar 2001 geconstrueerd. De basis van deze Volkstelling wordt gevormd door het Sociaal Statistisch Bestand (SSB), waarin registers en enquêtegegevens aan elkaar worden gekoppeld en waarin veel meer gegevens voorkomen dan gebruikt worden voor de volkstelling.

Op 11 november 2003 organiseerden het CBS en SISWO / Instituut voor Maatschappijwetenschappen de conferentie *De virtuele Volkstelling en het Sociaal Statistisch Bestand*. Centraal stond de wijze waarop het CBS voor 2001 een volkstelling heeft samengesteld, welke mogelijkheden het SSB nog meer te bieden heeft en welke gegevens sociale wetenschappers graag aan het SSB zouden willen ontleenen.

In dit rapport zijn de bijdragen van de sprekers van het CBS, de referaten uit de sociaal-wetenschappelijke wereld en de discussie die naar aanleiding van deze presentaties plaats vond gebundeld.

De waardevolle bijdragen van sprekers, referenten en deelnemers hebben er een leerzame dag van gemaakt. Door van gedachten te wisselen hebben de wetenschappers een beter beeld gekregen van de mogelijkheden die het CBS hen te bieden heeft en met hun reacties kan het CBS weer een stap verder zetten op weg naar een kennisinstituut. Op deze plek willen wij alle aanwezigen nogmaals danken voor hun bijdrage.

P. Everaers (CBS) en E. de Gier (SISWO)

Inleiding

B. Bakker en L. Putman

De aanleiding voor de organisatie van de conferentie ligt in een aantal discussies over de volkstelling en dan met name het feit dat Nederland al vanaf 1971 geen traditionele volkstelling meer kent waarbij tellers alle huishoudens bezoeken. Veel wetenschappers zien de volkstelling als een belangrijk zo niet het belangrijkste gegevensbestand en betreuren het ontbreken van volkstellinggegevens. De reden dat na 1971 geen integrale volkstelling meer werd gehouden was de grote publieke weerstand tegen dergelijke tellingen. In 1981 gebruikte het CBS combinaties van steekproefonderzoeken om de volkstellingstabellen die internationaal werden gevraagd samen te stellen. Voor de volkstelling van 1991 werd hetzelfde principe gebruikt waarbij de gegevens wel veel diepgaander werden herwogen met data uit de geautomatiseerde bevolkingsstatistieken.

Voor de volkstelling van 2001 heeft het CBS gebruik gemaakt van het Sociaal Statistisch Bestand (SSB). Het SSB is een combinatie van gekoppelde integrale registraties en steekproefonderzoekgegevens. Met het SSB kan het CBS eigenlijk continue virtuele volkstellingen uitvoeren. Dit betekent feitelijk dat nu weer voor het eerst sinds 1971 een volkstelling beschikbaar is, met dat verschil dat de gegevens nu niet via tellers zijn verzameld maar zijn verkregen uit bestaande registraties, administratieve bronnen en steekproefenquêtes. In november 2003 zijn de laatste tabellen van de volkstelling 2001 aan Eurostat

geleverd en op korte termijn worden de uitkomsten ook door het CBS gepubliceerd. In een marginaal van Mens en Maatschappij eind jaren 80 heeft de socioloog Jaap Dronkers, door het ontbreken van volkstellingen in 1981 en 1991, Nederland gekarakteriseerd als het statistisch Jutland van de wereld. Met het ontwikkelen van het SSB en de virtuele Volkstelling hoopt het CBS niet langer aan die karikatuur te voldoen.

De informatie voor de volkstelling 2001 is dus niet verzameld op de traditionele en heel kostbare manier, maar door het benutten van de informatie die al beschikbaar was uit registers aangevuld met steekproefenquêtes. Het op grote schaal gebruik maken van registers is een belangrijk onderdeel van het bedrijfsbeleid van het CBS. In feite zijn de gegevens van de volkstelling een deelverzameling van de gegevens in het SSB. In het SSB worden de beschikbare registers en enquêtes die het CBS houdt aan elkaar gekoppeld. Waarom gebeurt dat? Daar zijn drie redenen voor:

- voldoen aan een veranderende vraag van gebruikers;
- het verlagen van de kosten voor het produceren van statistieken;
- het verlagen van de enquêtedruk.

Allereerst de gebruikerswensen. Deze blijken onder meer uit de bijdragen die gevraagd zijn uit beleid, praktijk en wetenschap voor het nieuwe Meerjarenprogramma van het CBS. In toenemende mate willen gebruikers verbanden leggen tussen kenmerken die tot een verschillend inhoudelijk terrein behoren: ze willen gegevens over de gezondheid naar arbeidsomstandigheden, of de criminaliteit naar wijk- en buurtkenmerken, of de startsalarissen

van pas afgestudeerden in het hoger onderwijs, etc. Door alle beschikbare informatie uit registers en enquêtes te koppelen kunnen die verbanden worden gelegd.

Ook neemt de belangstelling voor laagregionale informatie zoals die over postcodegebieden toe. Deze informatie speelt vooral een belangrijke rol in de discussies over de grootstedelijke problematiek. Hetzelfde geldt voor informatie over relatief kleine groepen in de samenleving. De discussies van de laatste tijd over criminele Antilliaanse en Marokkaanse jongeren zijn daar een voorbeeld van.

De belangstelling voor longitudinale informatie neemt ook hand over hand toe. Een aantal belangrijke maatschappelijk vragen kunnen alleen maar beantwoord worden met behulp van tijdreeksen. Om een voorbeeld te noemen: het beperken van de instroom in de WAO. Daarvoor is informatie nodig over wie de WAO in stromen, bij welke werkgevers zij gedurende hun loopbaan hebben gewerkt, welke arbeidsomstandigheden daar waren, hoe de gezondheid van de WAO-intreder zich heeft ontwikkeld, etc.

Dan de kosten. Het CBS wordt in de verschillende bezuinigingsrondes van de overheid helaas niet overgeslagen. Ook nu weer wordt het CBS voor de komende kabinetsperiode getroffen door een forse bezuiniging. Door registerinformatie te benutten kunnen statistieken goedkoper worden gemaakt dan met enquêtes. Het verzamelen van informatie via enquêtes is duur, omdat er veel veldwerk voor nodig is. Daarnaast neemt de bereidheid van burgers en bedrijfsleven om deel te nemen aan enquêtes steeds verder af, met alle gevolgen van dien voor de kwaliteit van de statistische uitkomsten.

Als informatie al beschikbaar is in een register hoeft deze niet meer via een enquête te worden verzameld. Daarmee bespaart het CBS de samenleving flink wat kosten: enkele honderden miljoenen.

En als laatste de enquêtedruk. Het gaat hierbij zowel om de administratieve lastendruk bij het bedrijfsleven als de druk bij individuele huishoudens. De vrij hoge non respons bij veel persoonsenquêtes van het CBS wordt mede veroorzaakt door een zekere enquêtemoeheid bij de Nederlandse huishoudens. Het gebruik van gegevens die al in registraties voorhanden zijn, wordt algemeen ook als een geaccepteerde manier om gegevens voor statistiek te verzamelen gezien. Sociale statistieken maken ook gebruik van enquêtes bij bedrijven en instellingen. Door gebruik te maken van registers en de informatie niet langer via bedrijfsenquêtes te verzamelen daalt de administratieve lastendruk.

Om wetenschappers vertrouwd te maken met de inhoud van de virtuele Volkstelling 2001 en het SSB heeft het CBS op 11 november 2003 in samenwerking met SISWO / Instituut voor maatschappijwetenschappen de conferentie De virtuele Volkstelling en het Sociaal Statistisch Bestand georganiseerd. Daarnaast werden wetenschappers in de gelegenheid gesteld aan te geven met welke gegevens zij het SSB uitgebreid zouden willen zien, zodat het meer aan de behoefte van de gebruikers uit de sociale wetenschappen voldoet. Een van de taken van het CBS is immers de levering van microdata voor sociaal-wetenschappelijk onderzoek. De relatie tussen het CBS en de wetenschappers is niet altijd even goed geweest. In de marginalen van Mens en Maatschappij

in de jaren 80, maar ook elders, is er steeds weer op gewezen dat het CBS te veel op zijn data zat. In de jaren 90 is het CBS gegevens uit surveys gaan leveren via het Wetenschappelijk Statistisch Agentschap (WSA). Daarnaast is de mogelijkheid gecreëerd om binnen de muren van het CBS analyses uit te voeren waar meer detail of waar privacygevoelige gegevens voor nodig zijn: het zogenaamde *on site* werken. Dat heeft geleid tot een bevredigender situatie, blijkend uit het flinke gebruik van WSA-bestanden en de hoeveelheid *on site* werkers op het CBS.

Met de komst van het SSB treedt een nieuwe fase aan. Omdat het CBS steeds meer van registers gebruik maakt, neemt het aantal variabelen in surveys en het aantal surveys af. De surveys bevatten daarom straks minder informatie. De vraag die dan opportuun is, is of het CBS de sociale wetenschappen nog wel voldoende van microdata kan voorzien. Een tweede belangrijk doel van de conferentie is dan ook om na te gaan welke data uit het SSB aan de sociale wetenschappers ter beschikking moet worden gesteld. En de vraag die daar direct op volgt is: op welke wijze stelt het CBS die ter beschikking en hoe organiseert hij het zodat binnen de grenzen van de wet de onderzoekers zo goed mogelijk worden bediend.

Door meer samen te werken met de academische wereld kan het CBS weer een stap zetten in de wording tot kennisinstituut. Het CBS is niet alleen een cijferfabriek, het wil de Nederlandse samenleving ook kennis bieden.

Volkstellingen. De betekenis voor historisch sociaal-wetenschappelijk onderzoek

J. van Maarseveen (1)

1. Inleiding

De volkstellingen nieuwe stijl, gebaseerd op het Sociaal Statistisch Bestand (SSB), staan niet op zichzelf maar bouwen voort op een rijke traditie van vroeger in Nederland gehouden volkstellingen. In deze bijdrage staan we kort stil bij de vraag welke tellingen er in de loop der tijd gehouden zijn en welke analyses op het verkregen materiaal zijn uitgevoerd. Allereerst wordt een korte historische terugblik gegeven, gevolgd door een uiteenzetting van het begrip bevolking dat bij deze tellingen werd gehanteerd. Hoewel ook in de negentiende eeuw volkstellingen van groot belang geacht werden voor zowel kennis van de samenleving als voor de wetenschap, bleven wetenschappelijke verhandelingen aanvankelijk beperkt. Wel werd in meer of mindere mate verslag gedaan van de methode van onderzoek. Pas in de twintigste eeuw wordt de telling in grotere mate gebruikt als databron voor allerlei onderzoek en analyses. Tenslotte wordt gewezen op de mogelijkheden van andere analyses en het belang van historische reeksen.

2. De klassieke volkstellingen

De periode dat in Nederland de klassieke volkstellingen werden gehouden loopt van 1795 tot en met 1971 (zie

tabel 1). Hierbij verzamelden ‘tellers’ via veld-onderzoek de gegevens van iedere persoon. Deze tellingen die vanaf 1829 eenmaal in de tien jaar plaats vonden, kennen een rijke historie. Ze vonden met de Volkstelling 1971 een nogal luidruchtig einde. Door de toenemende automatisering van de overheid ontstond onder de bevolking vrees voor aantasting van de privacy, wat in 1970-'71 soms tot heftige kritiek tegen de volkstelling leidde. De opzet van deze telling was inderdaad in hoge mate geautomatiseerd. Op de individuele telkaarten kwamen de antwoorden te staan, die via optische leesapparatuur in een geautomatiseerd bestand werden opgeslagen. Herinneringen aan de Tweede Wereldoorlog waarbij de Duitsers de persoonsgegevens die opgeslagen waren in de bevolkingsregisters hadden misbruikt, speelden bij het verzet een rol. Hoewel het CBS de geheimhouding garandeerde, was het wantrouwen groot. Een algemene privacywetgeving ontbrak nog. Het zou nog tot 1988 duren voordat de bescherming van persoonsgegevens via privacywetgeving was geregeld (Wet Persoonsregistraties).

Een volgende telling bleek in 1981 niet meer mogelijk. De non-response (naar schatting ca. 26%) liet geen integrale telling toe. En in 1991 werd de Volkstellingenwet tenslotte afgeschaft. Voor 1981, 1991 en 2001 zijn voor Nederland in het kader van de volkstelling in de Europese Gemeenschap respectievelijk Europese Unie vervangende gegevens geleverd, gebaseerd op de bevolkingsadministratie en steekproeven.

De volkstellingen oude stijl bevatten een schat aan statistische basisgegevens over de Nederlandse bevolking

Tabel 1. Volks-, Beroeps-, Woning- en Registertellingen

jaar	Telling
1795	Volkstellingen in de Nederlandsche Republiek (15 okt. 1795 - 31 dec. 1795)
1829	Volkstelling (16 nov.-31 dec.)
1839	Algemene Volkstelling (18 nov.-31 dec.)
1849	Derde tienjarige Volkstelling (tevens Beroepstelling)*
1859	Vierde tienjarige Volkstelling (tevens Beroepstelling)*
1869	Vijfde tienjarige Volkstelling*
1879	Zesde tienjarige Volkstelling*
1889	Zevende tienjaarlijkse Volkstelling (tevens Beroepstelling)*
1899	Achtste tienjaarlijkse Volkstelling (tevens Beroepstelling en Woningstatistiek)*
1909	Negende tienjaarlijkse Volkstelling (tevens Beroepstelling en Woningstatistiek)*
1919	Woningtelling in gemeenten met meer dan 2000 inwoners en in enkele kleinere gemeenten (laatste maanden 1919)
1920	Tiende Algemene Volkstelling (tevens Beroepstelling)*
1930	Elfde Algemene Volkstelling (tevens Woningstatistiek en Gezinsstatistiek)*
1947	Twaalfde Volkstelling, annex Algemene Woningtelling (tevens Beroepstelling) (31 mei - 1 juni 1947)
1956	Algemene Woningtelling (30 juni 1956)
1960	Dertiende Algemene Volkstelling (31 mei - 1 juni 1960)
1971	Veertiende Algemene Volkstelling (tevens Beroepstelling en Algemene B4Woningtelling) 28 febr. 1971
1981	Registertelling bevolking; Telling woningvoorraad; Steekproefonderzoeken: Arbeidskrachtentelling en Woningbehoeftenonderzoek
1991	Registertelling bevolking; Steekproefonderzoeken: Enquête Beroepsbevolking; Woningbehoeftenonderzoek
1992-2000	Telling woningvoorraad (jaarlijks geactualiseerd)
1995-2000	Jaarlijkse structuurtelling bevolking en vanaf 1998 nieuwe huishoudensstatistiek (jaarlijks)

* Tijdstip telling 31 december

en behoren in de CBS-Bibliotheek tot de meest geraadpleegde statistieken. Mede om al deze publicaties in een zo goed mogelijke staat te behouden, heeft het CBS daarom ter gelegenheid van zijn honderdjarig bestaan in 1999, in samenwerking met het Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI) al de publicaties van de volkstellingen 1795-1971 via een zogenaamde mediumversie gedigitaliseerd. De originele boeken behoeven nu minder geraadpleegd te worden en hebben derhalve minder te lijden van het gebruik. Daarnaast is van de Volkstelling 1899 een inhoudelijke versie samengesteld, waarbij voor onderzoeksdoeleinden de gepubliceerde gegevens in een gedigitaliseerd bestand via het CBS-programma StatLine zijn te raadplegen en onderzoekers zelf tabellen kunnen samenstellen. Op dit moment worden, eveneens in samenwerking met het NIWI en met subsidie van de KNAW en NWO, de gegevens van alle overige volkstellingen inhoudelijk gedigitaliseerd. Van de Volkstellingen 1960 en 1971 zijn nog ruwe databestanden beschikbaar, die thans zodanig worden bewerkt dat zij met in achtneming van privacywaarborgen eveneens door onderzoekers zijn te raadplegen. De stand van zaken van deze digitalisering van de volkstellingen is te raadplegen op internet via www.volkstellingen.nl.

3. Het begrip bevolking

Bij alle volkstellingen die vanaf 1899 zijn gehouden, stond de vaststelling van de “werkelijke bevolking” centraal, dit is de bevolking die de werkelijke woonplaats heeft in een bepaalde gemeente of daar gewoonlijk verblijf houdt

(Methorst 1902). Hierbij gaat het om personen die in een van de bevolkingsregisters “behooren te zijn opgenomen”, ongeacht of zij op het tijdstip van de telling in de gemeente van hun werkelijke woonplaats zijn, ook wel de “de jure”-bevolking genoemd. Ook tijdelijk afwezige personen (schippers e.d.) werden geteld bij de gemeente waar zij hun werkelijke woonplaats hadden en waar zij dus als zodanig ingeschreven waren in het bevolkingsregister. Buitenlanders die tijdelijk in Nederland verbleven, vielen buiten de telling. Dit gold ook voor hen die het recht van diplomatieke immuniteit bezitten. Personen die tijdelijk in het buitenland verbleven maar ingeschreven waren in een Nederlandse gemeente, vielen daarentegen wel onder de doelgroep van de telling.

Een ander, daarmee niet te verwarren, begrip is de “feitelijke bevolking”. Hiervan is sprake, wanneer personen die in een gemeente op het tijdstip van de telling toevallig tijdelijk aanwezig zijn, wel tot de bevolking van deze gemeente gerekend worden, de bevolking dus die ‘de facto’ in een gemeente aanwezig is. Hoewel dit laatste begrip bij enkele tellingen in de negentiende eeuw is toegepast, bleef dit in de tellingen van 1879 en daarna buiten beschouwing.

De commotie rond de Volkstelling 1971 leidde in de Tweede kamer op 10 februari 1971 tot een correctie op dat “de jure” - begrip: buitenlanders mochten alleen worden geteld voor zover zij in het bevolkingsregister waren opgenomen. Acht jaar later werd deze correctie in het wetsontwerp tot wijziging van de Volkstellingenwet zelfs uitgebreid tot de gehele bevolking. Alleen de personen die geregistreerd waren in het bevolkingsregister,

waren onderwerp van telling. Nadat de telling van 1981 werd uitgesteld en in 1991 de Volkstellingenwet werd ingetrokken, is deze werkwijze feitelijk ook daarna toegepast bij de levering van de alternatieve volkstellinggegevens voor 1981 en 1991. Het integrale gedeelte van de demografische basisgegevens is namelijk gebaseerd op de gemeentelijke bevolkingsregistratie, zodat alleen personen die hierin zijn opgenomen in de statistische informatie zijn betrokken.

Tabel 2. Het begrip bevolking

Tellingen	Definitie van het begrip bevolking
1829, 1839	Werkelijke bevolking
1849	Feitelijke bevolking
1859, 1869	Feitelijke en werkelijke bevolking
1879-1971	Werkelijke bevolking
1981-2001	Werkelijke bevolking

Een tweetal bevolkingscategorieën werden in de wettelijke besluitvorming speciaal vermeld: de bevolking die in inrichtingen (ook gestichten genoemd), instellingen of tehuizen verbleef en de varende en rijdende bevolking. Voor deze bevolkingsgroepen werden aan de gemeenten uitvoerige voorschriften gegeven om duidelijk vast te stellen wat hun werkelijke woonplaats was. Voor de eerstgenoemde categorie was dit het tehuis of gebouw waar zij gedurende het jaar gehuisvest waren. Personen die in deze instellingen werkzaam waren, hadden daar hun werkelijke woonplaats als zij er hun nachtverblijf hadden.

De varende en rijdende bevolking (schippers, bewoners van woonwagens e.d.) werd geteld bij de gemeente waar de werkelijke woonplaats was. Na invulling van de telkaarten ontvingen zij een bewijs van invulling. Tijdens de telling moesten zij dit bewijs kunnen tonen aan elke ambtenaar die erom vroeg. Studenten, militairen e.d. die met vakantie of verlof waren bij hun ouders of elders, werden daar als tijdelijk aanwezig aangemerkt.

4. De klassieke volkstellingen als bron voor sociaal wetenschappelijk onderzoek

Welke resultaten werden gepubliceerd? Allereerst werden binnen ongeveer een half jaar de voorlopige uitkomsten gepubliceerd van de bevolking per gemeente, onderscheiden naar geslacht. Deze uitkomsten waren gebaseerd op tellingen van de gemeenten die hierbij gebruik maakten van de bevolkingsregisters en aantekeningen van de tellers. Ongeveer een jaar na afloop van de telling volgden de definitieve uitkomsten van de zogenaamde “eigenlijke volkstelling”. Dit was de telling exclusief de beroepstelling en woningstatistiek (-telling). Resultaten van deze twee laatste kwamen een half tot heel jaar later ter beschikking. De definitieve resultaten verschenen naderhand in een reeks van publicaties met gegevens van de eigenlijke volkstelling, de beroepstelling en de woningtelling, veelal onderscheiden naar zowel gemeenten, provincies en rijk. Daarnaast verschenen er publicaties met de resultaten van afzonderlijke variabelen zoals godsdienst, onderwijs e.d. Met uitzondering van de telling van 1971 stelde het CBS alle publicaties samen.

Wanneer we voor de periode 1795-1971 het aantal publicaties en de omvang ervan overzien (zie tabel 3), dan treedt na 1899 een sterke versobering op met 1920 als dieptepunt. Gedetailleerde uitkomsten kwamen vanaf 1947 beschikbaar via zogenaamde gelichtdrukte tabellen die niet in de publicaties waren opgenomen. In 1930 werden gedetailleerde gegevens beschikbaar gesteld via zogenaamde leggers, tabellen die ter inzage lagen en konden worden overgeschreven. Nieuw was ook dat in 1947 per gemeente een standaardboekje werd samengesteld en topografische afdrukken werden gemaakt van wijk- en buurtindelingen. Deze gedragslijn is ook voor de tellingen van 1960 en 1971 gevolgd. Vermeldenswaard is nog dat bij de volkstellingen geruime tijd ook gegevens verzameld werden over de oppervlakte van de gemeenten.

5. Analyses en internationale vergelijkingen

Zoals uit voorgaand overzicht blijkt, gaat het bij de publicaties van de volkstellingen primair om de tabelresultaten, een toelichting hierop en om beschrijvende analyses. Verder wordt in het eerste deel van een reeks publicaties van een bepaalde telling (veelal summier) aandacht besteed aan de methodische aspecten. In de bijlagen hiervan werden de vragenlijsten afgedrukt met de instructies aan gemeenten en tellers, en gebruikte classificaties. Vergelijkingen met voorgaande tellingen waren beknopt en meestal beperkt tot de voorgaande telling. Hierin kwam verandering in de jaren zestig van de vorige eeuw. De mogelijkheden tot analyse namen toe door de automatisering, maar ook doordat het aantal

Tabel 3. Aantal bladzijden en delen van alle gepubliceerde Volks-, Beroeps- en Woningtellingen, 1795-1971 (incl. transparanten en lichtdrukken)

jaar volks- telling	bladzijden	delen	licht- drukken
1795	191	2	
1829	18	(*)	
1839	85	1	
1849	1165	12	
1859	1184	3	
1869	889	3	
1879	2262	12	
1889	10223	26	
1899	9925	27	
1909	4144	14	
1919**	191	1	
1920	1953	10	
1930	2353	11	
1947	1325	12	31 000
1956**	345	3	47 000
1960	1809	18	75 000
1971	4503	38	87 000
Totaal	42565	193	240 000

* Volkstelling 1829. Enkele resultaten zijn gepubliceerd in het Statistisch Jaarboekje 1830 en in deel 2 van de Volkstelling 1859.

** In 1919 en 1956 vond een afzonderlijke Woningtelling plaats; in 1919 in gemeenten met meer dan 2000 inwoners en in enkele kleinere gemeenten; in 1956 een (integrale) Algemene Woningtelling

vragen resp. variabelen en afgeleide variabelen groter was dan voorheen (2).

Vermeldenswaard is de publicatie over de Volkstelling 1960 (deel 10 C), waarin de uitkomsten van de beroepstellingen 1849-1960 met elkaar worden vergeleken. Afzonderlijke vermelding verdienen de analyses die het CBS voor de Volkstelling 1971 in samenwerking met SISWO en universiteiten heeft samengesteld in 18 censusmonografieën. Voorafgaand hieraan had SISWO wensen geformuleerd voor de indeling van variabelen en de opzet van tabellensets (Reinink, 1968). Aangezien het CBS bij de bewerking kampte met een aanzienlijke vertraging, heeft SISWO in interimrapporten eerder de resultaten van literatuurstudies gepubliceerd. De eindrapporten handelden over de volgende thema's:

1. De huisvestingssituatie
2. Geboorte-intervallen
3. Gescheidenen en verweduwen
4. Wederzijdse kenmerken van huwelijkspartners
5. Laagstgeklasseerden
6. Een sociaal-demografische analyse van de huwelijkssluiting
7. Beroepsarbeid
8. Vergrijzing
9. Ontwikkelingsfasen van het gezin
10. De positie van jongeren ten opzichte van het ouderlijk gezin
11. Huishoudenssamenstelling en samenlevingsvormen
12. Onderwijs
13. Kerkelijke gezindten
14. Groeiende en kwijnende plattelandskernen

15A. Typologieën van Nederlandse gemeenten naar bevolkingsgroei

15B. Typologieën van Nederlandse gemeenten naar stedelijkheidsgraad

16. Regionale arbeidsarbeidsmarkten

17. Opleiding, beroepen- en werkgelegenheidsstructuur (niet verschenen) (3)

18. Huishouden, huwelijk en gezin

De hier vermelde thema's geven een indruk van de mogelijkheden tot analyse bij de tellingen, die vanaf 1930 zijn gehouden. De analysemogelijkheden van de tellingen 1909-1920 zijn minder, terwijl die voor de periode 1869-1899 weer ruimer zijn.

De mogelijkheden tot analyses van de klassieke volkstellingen zijn recent verder toegenomen door de digitalisering van de volkstelling. Zoals vermeld zijn de resultaten van de Volkstelling 1899 opgenomen in een volledig te doorzoeken en analyseerbare database via het CBS-programma StatLine. Deze database is te raadplegen via <http://www.volkstellingen.nl>. Behalve een beschrijving van de methodische aspecten, het proces van digitalisering zijn in Van Maarseveen en Doorn (2001) analyses gemaakt gepresenteerd over onder meer:

- de groei van de bevolking in de negentiende eeuw
- gezin en huishouden op het breukvlak van twee eeuwen
- kerk en godsdienst rond 1900
- vreemdelingen in de volks- en beroepstellingen 1899-1971
- de sociale structuur rond 1900
- de samenstelling van de burgerij rond 1900

- werkgelegenheid 1807-1909
- de gegevens op provinciaal niveau
- regionale indelingen (een clusteranalyse)

Het bovengaande geeft slechts een indruk van de mogelijke analyses. Een uitgebreid bibliografisch onderzoek is niet beoogd. Afzonderlijke vermelding verdienen de berekeningen die met behulp van volkstellingsgegevens zijn uitgevoerd door statistici en verzekeringsdeskundigen voor de samenstelling van sterftetafels. Als eerste publiceerde de bekende statisticus Von Baumhauer, chef van het Statistisch Bureau op het departement van Binnenlandse Zaken, sterftetafels voor de jaren 1840-1851. Voor de volgende jaren (tot en met 1930) zijn behalve door Von Baumhauer sterftetafels samengesteld door Prof.dr. A.J. van Pesch en Prof. Dr. J.J.A. Mulder. Een historisch overzicht geeft Van Haften (1923).

Ook zijn internationale vergelijkingen mogelijk. Op basis van een globaal verkennend onderzoek blijkt dat in Nederland waarschijnlijk weinig internationale analyses zijn gemaakt op basis van gegevens van volkstellingen (4). Uitkomsten van volkstellingen in een groot aantal landen zijn voor de periode 1750-1993 zonder nadere analyse gepubliceerd in het standaardwerk van Mitchell (2001). Hierin zijn voor de jaren 1970, 1980 en 1991 de cijfers voor Nederland geschat. Merkwaardig want voor de volkstelling in Europees verband heeft het CBS destijds gegevens geleverd.

Overigens hebben de gegevens van de verschillende landen niet altijd op dezelfde jaren (teldata) betrekking en kunnen de gebruikte definities verschillend zijn. Zelfs in de tweede helft van de twintigste eeuw is geen sprake van

eenzelfde periodiciteit bij het houden van volkstellingen. In sommige landen vinden de volkstellingen eenmaal in de vijf jaar en in andere landen eenmaal in de tien jaar plaats, daarbij in een jaar eindigend op een '0' of op een '1'. Dit blijkt bijvoorbeeld uit de Eurostat-publicatie over de resultaten van de in 1990/1991 gehouden volkstellingen in de Europese Gemeenschap en de Europese Vrijhandels Associatie (5). Toch bieden de door Eurostat gepubliceerde gegevens beter vergelijkingsmateriaal dan voorheen, aangezien sprake is van enige coördinatie van definities, indelingen en classificaties van variabelen.

Voor deze coördinatie van de methodologie is na de Tweede Wereldoorlog het werk van het Statistical Office van de Verenigde Naties van groot belang geweest. Vanaf de jaren vijftig publiceert deze instelling met enige regelmaat een handboek voor het houden van volks- en woningtellingen (6). Dit vormt internationaal het uitgangspunt bij het uitvoeren van nieuwe tellingen. Ook de methodologie van de in Europees verband gehouden tellingen neemt de in het handboek geformuleerde aanbevelingen als leidraad.

Internationale vergelijkende analyses zijn derhalve pas goed mogelijk als de gegevens op elkaar zijn afgestemd respectievelijk vergelijkbaar zijn gemaakt. Daarvoor dienen er veelal correcties op het materiaal te worden toegepast, omdat er allerlei verschillen kunnen zijn in methodiek: zoals verschil in begrip bevolking, teldatum (telperiode), definities van de variabelen etc.

Behalve internationale vergelijkingen met behulp van demografische basisgegevens als geslacht, burgerlijke staat,

en mogelijk globaal naar categorieën van leeftijd, gezin / huishouden, godsdienst, woning, urbanisatiegraad (stad - platteland), kunnen ook de gehouden woningtellingen een bron zijn voor het maken van internationale vergelijkende analyses.

Data over onderwijs zijn, vanwege de uiteenlopende onderwijssystemen, weer minder gemakkelijk vergelijkbaar te maken. Hetzelfde geldt voor de variabele beroep. Wel vond vanaf het einde van de negentiende eeuw reeds uitwisseling van gedachten plaats over beroepsclassificaties en zeker na de Eerste Wereldoorlog met de oprichting van het International Labour Office (ILO) en na de oprichting van de Europese Gemeenschappen is in toenemende mate sprake van internationale onderlinge afstemming. Zo sluiten de beroepsclassificaties van de volkstellingen van 1960 en 1971 aan bij de International Standard Classification of Occupations (ISCO) van de ILO. De Standaard Beroepenclassificatie 1992 is een uitwerking van de ISCO 1988. Vermeldenswaard is dat historici in 2002 op basis van de ISCO 1968 een “Historical International Standard Classification of Occupations” hebben gepubliceerd. Hiermee kunnen internationale vergelijkingen worden gemaakt met gegevens van beroepstellingen die gedurende de negentiende en twintigste eeuw in een groot aantal landen zijn gehouden.

Tenslotte is vermeldenswaard het internationale historische “iPums-project” (Integrated Public Use Microdata Series International) van de University of Minnesota. Doel is met inachtneming van waarborgen voor strikte geheimhouding en na harmonisatie van de

variabelen onderzoekers in de gelegenheid te stellen om (vergelijkende) analyses uit te voeren op basis van steekproeven uit de microdata van volkstellingen van landen. De data hebben betrekking op tellingen vanaf de 19e eeuw tot heden.

6. De volkstelling als ijkpunt en ophoogkader voor sociaal-wetenschappelijk onderzoek

Vanaf het begin dat volkstellingen werden gehouden vormden zij een ijkpunt voor de demografische gegevens van de bevolking in het algemeen en specifieke bevolkingsgroepen in het bijzonder. Doel van de volkstelling was niet alleen het vaststellen van de bevolkingsomvang naar een aantal kenmerken, maar tevens het toetsen (ijken) van het bevolkingsregister. Met behulp van de via de volkstelling verkregen gegevens werden de bevolkingregistraties geactualiseerd (herijkt). Vanaf 1947 zijn door het CBS de demografische basisgegevens afkomstig van de volkstelling jaarlijks met behulp van door de gemeenten geleverde gegevens geactualiseerd. De herijkte gegevens van de bevolking zijn vastgelegd in jaarlijkse overzichten.

Na 1971 vonden partiële en vanaf 1983 ook integrale registertellingen plaats ter vervanging van de vroegere overzichten. De herijking van de demografische basisgegevens had aanvankelijk alleen betrekking op de variabelen: geslacht, geboortjaar, burgerlijke staat en land van nationaliteit. In 1990 en 1992 kwamen daar de volgende variabelen bij: geboorteland, geboorteland vader en geboorteland moeder, jaar van meest recente

burgerlijke staatswijziging, en jaar van meest recente vestiging in Nederland. Aan de herijking van de stand van de bevolking via registertellingen kwam een einde toen in 1994 de Gemeentelijke Basis Administratie persoonsgegevens (GBA) tot stand kwam. Op basis van de GBA vindt sinds 1995 vindt jaarlijks een structuurtelling plaats van de demografische basisgegevens.

Na de Tweede Wereldoorlog heeft het steekproefonderzoek een hoge vlucht genomen. Daarbij werden de uitkomsten van de Volkstelling vanwege het integrale karakter tevens veelvuldig als ophoogkader gebruikt om landelijke en regionale totaalcijfers te verkrijgen. Hoe belangrijk de volkstelling is voor het ijken van steekproeven, kwam naar voren tijdens de discussie in de Centrale Commissie voor de Statistiek (CCS) en het Parlement over de voorbereiding van een nieuwe volkstelling in 1981. Ook blijkt hieruit dat in feite toen reeds de eerste schreden waren gezet op weg naar het huidige Sociaal Statistisch Bestand, hoewel dit niet als zodanig werd geformuleerd. Het CBS had namelijk voorgesteld om een integrale telling te houden op basis van de bevolkingsadministratie en een steekproef voor aanvullende gegevens. Dit plan stuitte op krachtig verzet van de regionale, planologische en wetenschappelijke instellingen. De minister van Economische Zaken nam het advies van de CCS over en stelde daarom aan het Parlement voor een integrale telling te houden. Van een telling met verzameling van demografische gegevens was het onderzoek, aldus de minister in het Parlement, uitgegroeid tot een sociale structuurtelling waarin de demografische gegevens het bindende element vormden. De volkstelling had volgens de minister een centrale functie in het geheel van sociale

statistieken voor:

- algemene achtergrondgegevens ten behoeve van andere statistieken;
- gegevens die (vooralsnog) niet via bestaande registratiesystemen of enquêtes verkrijgbaar zijn;
- gegevens ten behoeve van het ijken van kleinschaliger steekproefonderzoek en het kiezen van de onderzoekseenheden.

De minister gaf daarbij zijn voorkeur aan een combinatie van verzameling van:

- een beperkt aantal gegevens op integrale basis (deels uit het bevolkingsregister, deels via veldonderzoek)
- alle overige gegevens op steekproefbasis.

Maar ook bij modern historisch-demografisch onderzoek keert het belang van de integrale tellingen terug. Zo vindt thans met subsidie van NWO een grootschalig onderzoek plaats onder de naam *Life courses in context*. Doel hiervan is via onderzoek in de gemeentelijke bevolkingsregistraties over de Nederlandse bevolking op individuele basis demografische basisgegevens te verzamelen die voor historisch en sociaal-wetenschappelijk onderzoek gebruikt kunnen worden (zoals op gebied van: gezondheid, sterfte, sociale bewegingen, religie, onderwijs, huishoudens, arbeidsrelaties, sociale mobiliteit, migratie, en methodologisch onderzoek naar data-integratie). Het betreft ongeveer 40.000 personen geboren tussen 1863 – 1922, waarvan de levensloop in geautomatiseerde bestanden wordt vastgelegd. Koppeling van de levenslopen met gegevens van de volkstellingen geeft informatie over de sociale, lokale en regionale context. Hoewel de beschikbare gegevens van de volkstellingen uit de periode 1795 – 1947 alleen op geaggregeerde wijze zijn gepubliceerd, bieden zij

vanwege hun grote mate van detaillering over geografische en sociale categorieën de nodige achtergrondinformatie. Een voorbeeld hiervan zijn de resultaten van de reeds inhoudelijk gedigitaliseerde Volkstelling 1899.

Tenslotte zij nog vermeld dat de adressen van de volkstellingen vaak gebruikt zijn voor andere onderzoeken. In 1930 voor de Algemene Bedrijfs-telling, voor onderzoek naar gehandicapten in 1920, terwijl de adressen van de Volkstelling 1971 geruime tijd gebruikt zijn als steekproefkader voor allerlei personenquêtes die het CBS in de jaren zeventig heeft gehouden.

7. Tot besluit

Zoals uiteengezet ging het CBS na 1971 voor de in Europees verband gehouden volkstellingen van 1981 en 1991 over op de verstrekking van alternatieve gegevens, gebaseerd op registertellingen en enquêtes. In die zin is de samenstelling van nieuwe volkstellinggegevens voor 2001 gebaseerd op het SSB, een voortzetting van een sinds 1981 ingezette aanpak. Het verschil is dat van integratie van deze twee soorten bronnen in 1981 en 1991 nog geen sprake was. Interessant is in dit verband wat Dronkers (1988) destijds over deze alternatieve gegevens heeft opgemerkt. Behalve bij de kwaliteit van de registerdata plaatst hij vraagtekens bij het intrekken van de Volkstellingenwet in 1991. Voor het alternatieve plan van volkstellinggegevens, zo merkt hij op, ontbreekt de juridische basis waardoor zo'n plan de speelbal van de politiek kan worden, afhankelijk van de omstandigheden (financiën en weerstand bij de bevolking). Hij verzucht:

“Wellicht blijft ons slechts het beroep op de Europese richtlijn over als de nationale wetgever verstek laat gaan”. De vraag is inderdaad of er in 1981, 1991 en 2001 alternatieve volkstellingsgegevens geleverd zouden zijn, als de Europese richtlijn er niet zou zijn geweest.

Tenslotte nogmaals de analyses. Op basis van de volkstellingen zijn in de loop der tijd allerlei analyses uitgevoerd. Hierboven is daarvan een beknopte schets gegeven. Een uitgebreid bibliografisch onderzoek zou deze schets moeten aanvullen. Met name op regionaal en lokaal niveau zijn allerlei deelanalyses mogelijk en ook toegepast (7). Voor analyses van kleinere categorieën van de bevolking en voor regionale analyses biedt de tabelinformatie, opgeslagen op de lichtdrukken (transparanten) voor de jaren 1947, 1956 (woningtelling), 1960 nog allerlei nieuwe mogelijkheden. Maar vooral de digitale inhoudsconversies van de Volkstellingen 1795-1971 alsmede de speciale onderzoeksbestanden die thans worden samengesteld op basis van de originele records van de Volkstellingen 1960 en 1971 bieden binnen niet al te lange tijd nieuwe mogelijkheden voor allerlei analyses.

Voor deze analyses is het van belang te beschikken over historische reeksen, zodat vergelijkingen in de tijd gemaakt kunnen worden. Bijzondere vermelding verdient in dit verband de studie van Oomens (1989), die in een diepgaande studie de resultaten van de negentiende eeuwse volkstellingen onderling vergelijkbaar heeft gemaakt. Ook Oomens en den Bakker (1994) hebben de resultaten van de beroepstellingen, gehouden vanaf 1849 – 1971 en naderhand op steekproefbasis voortgezet in de Arbeidskrachtentellingen en Enquêtes Beroepsbevolking, geanalyseerd en op een onderling vergelijkbare basis gepubliceerd.

Om in de toekomst historisch-vergelijkende studies te maken is het van groot belang dat de SSB-databestanden, waaraan de tabellen met volkstellingsgegevens worden ontleend, zorgvuldig worden bewaard en gedocumenteerd. Zo zijn de tabellen die destijds voor de in Europees verband gehouden volkstellingen van 1981 en 1991 zijn samengesteld op dit moment niet direct gemakkelijk voor gebruikers te achterhalen, terwijl ook niet duidelijk is hoe de daarbij gebruikte databestanden zijn bewaard en gedocumenteerd. Documentatie hierover is tevens van belang bij deelname van Nederland aan het eerder vermelde internationale iPums-project dat gebaseerd is op microdata.

Ook voor de Volkstellingen nieuwe stijl is het van belang dat ze kunnen worden vergeleken met die van de voorheen gehouden tellingen. Dit zal extra inspanningen vergen in vergelijking met vroeger, omdat tabellen thans worden samengesteld via koppeling van bestanden. De wijze waarop de tabellen worden vastgelegd dient dus zorgvuldig gedocumenteerd te worden, wil de toekomstige gebruiker er mee uit de voeten kunnen.

Noten

- (1) Jacques van Maarseveen is historicus en oud CBS-er. Hij houdt zich onder meer bezig met de historie van de Volkstellingen en de digitalisering daarvan. smrn@cbs.nl
- (2) In de loop der tijd zijn vragen gesteld over: leeftijd, beroep, geslacht, positie in het bedrijf, burgerlijke staat, sociale groep, geboorteplaats, aard van het bedrijf, woonplaats, bedrijfsklasse, nationaliteit, onderwijsniveau, periode van vestiging, woning (eigendom/huur, bouwwijze e.d.), godsdienst/kerkelijke gezindte, lichamelijke handicaps (blindheid e.d.), verblijf/aanwezigheid, huwelijksvruchtbaarheid,

alleenstaande/hoofdhuishouden, huisvesting (samenwonen, aantal vertrekken e.d.), plaats in het gezin/huishouden, inkomens(niveau), gezin/huishouden (gezinsfase), bron van inkomsten, personen in instellingen, werkende vrouwen, varende en rijdende bevolking, woon- en werkverkeer / forensisme.

- (3) Zie voor dit onderwerp Huijgen, Riesewijk en Conen (1983). Deze studie is gebaseerd op de gegevens van de Volkstellingen 1960 en 1971 alsmede de Arbeidskrachtentelling 1977.
- (4) Een eerste globaal verkennend onderzoek in de CBS-Bibliotheek leverde slechts het volgende op:
 - Vergelijking van enkele resultaten uit de algemene volkstellingen van de Beneluxlanden 1960-1961 (Brussel 1965). De publicatie bevat alleen een cijfermatig overzicht.
 - Een publicatie over de demografische ontwikkelingen in West-Europa. Deze was echter niet zozeer gebaseerd op de volkstellingen maar op de jaarlijkse demografische gegevens en samengesteld voor onderwijskundige doeleinden (publicatie: "Over de Europese bevolking. Demografische ontwikkelingen, Geografisch Instituut Rijksuniversiteit Utrecht, 1979).
- (5) Zie hiervoor Eurostat (1996). De uitkomsten van deze 'censuses' staan ook in de database van Eurostat: New Cronos.
- (6) Handbook of population census methods (New York 1954; New York 1958-1959). Handbook of population and housing census methods (New York 1969). Handbook of population and housing censuses (New York 1992).
- (7) Zie bijv. Q.J. Munters (1970)

Referenties

Dronkers, J., 1988, Nederland het statistisch Jutland van Europa? In: *Mens en Maatschappij* jrg. 63, nr. 4, blz. 345-346

- Eurostat, 1996, *Population, Households and Dwellings in Europe: Main Results of the 1990/1991 Censuses*. Statistical Document. Theme 3: Population and social conditions, Series C: Accounts and surveys (Luxembourg: Office for Official Publications of the European Communities)
- Haaften, M. van, 1923, De sterftetafels, afgeleid uit de volkstellingen in Nederland, In: *Verzekeringsarchief* juni 1923, 128-163.
- Huijgen, F., B.J.P. Riesewijk en G.J.M. Conen, 1983, *De kwalitatieve structuur van de werkgelegenheid in Nederland: bevolking in loondienst en functieniveaustructuur in de periode 1960-1977* (Den Haag: Staatsuitgeverij)
- Maarseveen, J.G.S.J. van, en P.K. Doorn (red.), 2001, *Nederland een eeuw geleden geteld. Een terugblik op de samenleving rond 1900* (Amsterdam: NIWI).
- Methorst, H.W., 1902, Geschiedenis van de statistiek in het Koninkrijk der Nederlanden. In: *Bijdragen tot de statistiek van Nederland*. Nieuwe volgrees. (s-Gravenhage: Centraal Bureau voor de Statistiek).
- Mitchell, B.R., 2001 *International historical statistics. Europe 1750-1993* (4th ed.)
- Munters, Q.J., 1970, *Interne en externe rekrutering en kerkelijke gezindte. Sociologisch commentaar bij nog niet eerder gepubliceerde volkstellingsgegevens, met name m.b.t. enkele kleinere godsdienstige groepen in de stad Utrecht*, Mededelingen van het Sociologisch Instituut der Rijksuniversiteit Utrecht, nr. 64 (Utrecht 1970).
- Oomens, C.A. 1889, De loop der bevolking van Nederland in de negentiende eeuw. In: *Statistische onderzoekingen*, nr. M35 (Den Haag: Staatsuitgeverij).
- Oomens, C.A. en G.P. den Bakker, 1994, De beroepsbevolking in Nederland 1849-1990. In: *Supplement bij de Sociaal-economische maandstatistiek*, 1994, nr.2, blz. xxx-xxx
- Reinink, G.J., 1968, *De Volkstelling 1970. Aanpassingsdesiderata in het perspectief van het sociaal-wetenschappelijk onderzoek*. (Amsterdam: SISWO)

Bijlage. Inventarisaties van volkstellingen en bronnen

De bronnen, vindplaatsen en publicaties van de volkstellingen in de negentiende eeuw worden beschreven door:

- J.A. de Kok, Nederland op de breuklijn Rome-Reformatie. Numerieke aspecten van Protestantisering en Katholieke Herleving in de Noordelijke Nederlanden 1580-1880 (Assen 1964); betreft: de tellingen van 1795, 1809, 1829 en 1849
- J.H. Kompagnie, Uitkomsten van volkstellingen als archivalia bij de Tweede Afdeling en in gedrukte vorm (Den Haag 1992, Algemeen Rijksarchief).
- H.W. Methorst, Geschiedenis van de statistiek in het Koninkrijk der Nederlanden. In: Bijdragen tot de statistiek van Nederland. Nieuwe volgrees. Uitgegeven door het Centraal Bureau voor de Statistiek ('s-Gravenhage 1902).
- J.L. van Zanden, De Volkstelling 1897/08. In: Broncommentaren I-IV, nr. III, 41-50 (Den Haag 1987). Deze tweede nationale telling is exclusief Limburg en Zeeuws-Vlaanderen.

De volgende inventarisaties zijn door het CBS samengesteld, te weten:

- Inventarisatie van de volkstellingen, beroepstellingen, woning- en gezinstellingen 1830-1930 (CBS-publicatie; Den Haag ca. 1951).
- Bibliografie van de Nederlandse volkstellingen sinds 1795 (CBS-publicatie, ca. 1967). (betreft de periode 1795 -1960).
- Inventarislijsten gelichtdrukte tabellen.

In de CBS-Bibliotheek zijn inventarislijsten aanwezig van de gelichtdrukte tabellen van de Volkstellingen 1947, 1960 en 1971.

Vermeldenswaard is ook de inventariserende publicatie, die samengesteld is door J.G.M. Assen en M.L. van der Heijde, *Bronnen van het Centraal Bureau voor de Statistiek. Deel II Volkstellingen* (Geografisch en Planologisch Instituut, Sociologisch Instituut van de Katholieke Universiteit Nijmegen, 1985). Deze publicatie richt zich met name op vragenformulieren, uitkomsten en publicaties van de Volkstellingen van 1960 en 1971.

Voorbeeld van publicaties van de volkstelling: Volkstelling 1971

Om een indruk te geven van de reeks publicaties van de meest recente klassieke volkstelling volgt hier een summier overzicht over de publicaties van de Volkstelling 1971.

Van deze telling drie series verschenen.

Serie A betreft beschrijvende analyses van de algemene uitkomsten (8 delen)

Serie B betreft de uitkomsten per gemeente (13 delen).

Serie C betreft de systematische classificaties (3 delen; bedrijfs-, beroeps-, en onderwijsclassificaties).

Ter informatie verscheen: *Waarom een volkstelling?* Een informatiemap met toelichtingen, enz. ('s Gravenhage 1971).

Daarnaast verschenen nog enkele publicaties over de organisatie en uitvoering van de telling waaronder een technische toelichting bij de vragenlijst, een tweedelige Handleiding voor de gemeentebesturen, en het gemeentelijk tabellenprogramma.

Het belang van een éénprocentsteekproef uit de huishoudens van de virtuele volkstelling.

J. Dronkers (1)

1. Inleiding

Met de ontwikkeling van het Sociaal Statistisch Bestand lijkt er een einde te gaan komen aan een donkere episode in de Nederlandse samenleving: de periode (1972-heden) waarin geen volkstelling gehouden werd. Laten wij deze merkwaardige terugval voorlopig wegboeken als een van onbedoelde neveneffecten van de jaren zestig en er nu geen verdere woorden aan vuilmaken. Maar bij deze nieuwe start is het wel nodig een aantal zaken in de relatie tussen volkstellingorganisator (voor het gemak in het vervolg noem ik die het CBS) en de wetenschap goed te regelen. Als dat niet vanaf het begin gebeurt, kan het opnieuw goed foutlopen, net als in de jaren zestig met de 'oude' volkstelling. Bovendien kunnen wij nu profiteren van de wet van de remmende voorsprong: een nieuwe virtuele volkstelling zou weleens veel meer mogelijkheden kunnen bieden dan de oude vorm met interviews.

Centraal staat in deze bijdrage het belang van een onafhankelijke analyse van de virtuele volkstellingsgegevens door bonafide wetenschappers, die geen directe of indirecte bindingen hebben met belanghebbenden. Met belanghebbenden bedoel ik in hoofdzaak overheidsdepartementen (inclusief de quasi-onafhankelijke onderdelen als inspecties), SER, SCP,

CPB, wetenschappelijke instituten van politieke partijen en belangenorganisaties, en NWO. Zij zijn bij de huidige zwakke financiële positie van de sociale wetenschap te vaak diegenen die inhoud en richting van het sociaal-wetenschappelijk onderzoek bepalen, dankzij de 'zilveren koorden' van hun subsidies en (toekomstige) onderzoeksopdrachten (2).

Het vrij ter beschikking stellen van een éénprocentsteekproef uit die virtuele volkstelling (in het CBS jargon: het Sociaal Statistisch Bestand), zonder restricties in vraagstelling of door financiële bijdragen, zal een belangrijke bijdrage van het CBS zijn aan de versterking van de onafhankelijke wetenschap. Bovendien zal het CBS als een van de eerste de vruchten plukken van het vrij ter beschikkingstellen van een éénprocentsteekproef, omdat haar inhoudelijke positie binnen de overheidsbureaucratieën versterkt zal worden. Haar producten dragen echt bij tot meer kennis van de samenleving. Daardoor zal het publieke debat beter gevoerd kunnen worden en minder gedomineerd kunnen worden door politieke taboes, vooroordelen en belangen.

2. Wat is een éénprocentsteekproef uit een volkstelling?

Voor het antwoord op deze vraag, kunnen wij het best naar de USA gaan. Eens in de zoveel jaar wordt daar een volkstelling gehouden, die voor het interne reilen en zeilen van de Verenigde Staten erg belangrijk zijn (verdeling van zetels in huis van afgevaardigden; financiële verhoudingen tussen de staten en de federatie, etc.). Nadat de volkstelling

voltooid is en de eerste publicaties van het Censusbureau verschijnen, wordt er ook een steekproef uit die volkstelling getrokken en ter beschikking van het publiek gesteld. Deze éénprocentsteekproef is bedoeld als een representatieve steekproef van de geïnterviewde bevolking van de USA. (zie voor meer informatie over deze service van het US censusbureau: http://factfinder.census.gov/servlet/DatasetMainPageServlet?_lang=en). Aan het gebruik van deze steekproef zijn geen extra voorwaarden verbonden, buiten die welke voortvloeien uit de wet, en na enige tijd ziet men in de Amerikaanse tijdschriften dan ook geregeld analyses van deze steekproef. Ook in de meer serieuze dag- en weekbladpers wordt deze steekproef als referentiepunt gebruikt bij allerlei beschouwingen. Daardoor heeft de Amerikaanse volkstelling ook een veel breder draagvlak in de samenleving. Naast de éénprocentsteekproef, die veel inhoudelijke details geeft maar weinig regionale verschillen laat zien, bestaat er in de USA ook een vijf procentsteekproef, die weinig regionale details bevat maar veel inhoudelijke verschillen.

In Nederland hebben wij hetzelfde nodig: een representatieve steekproef uit het Sociaal Statistisch Bestand, als de voorlopig beste beschrijving van de bevolking in Nederland. Een steekproef van één procent geeft genoeg mogelijkheden om allerlei bijzondere combinaties van variabelen te maken en die valide te onderzoeken. De huidige kracht van de computer is groot genoeg om deze grote bestanden aan te kunnen. Over de mate van detail van de regionale en inhoudelijke variabelen (een belangrijk verschil tussen de twee Amerikaanse varianten) wil ik het nu niet hebben. Ik denk dat wij in Nederland een middenweg kunnen vinden tussen beiden, met een accent op de inhoudelijke informatie.

Er zou maar één volledige versie van deze éénprocentsteekproef moeten bestaan, die aan iedereen in die vorm ter beschikking wordt gesteld. Dit maakt discussie tussen de gebruikers mogelijk en verbetert de kwaliteit van de analyses, omdat ze dan onderling vergelijkbaar blijven. Het CBS moet dus niet voor elke gebruiker steeds een nieuw aangepast bestand maken: dat kost alleen maar extra geld.

Elk jaar bij het maken van de nieuwe versie van het Sociaal Statistisch Bestand wordt er ook een nieuwe éénprocentsteekproef gemaakt. Zo ontstaat er op de duur een reeks van éénprocentsteekproeven, voor elke nieuwe versie een. Deze reeks bestanden maakt het mogelijk ontwikkelingen in de tijd in kaart te brengen aan de hand van de steekproeven.

Gezien de mogelijkheid om de éénprocentsteekproef via het sofi-nummer te koppelen, is het erg wenselijk dat in de meest recente éénprocentsteekproef niet alleen de variabelen van dat jaar opgenomen zijn, maar ook individuele kerngegevens uit eerdere Sociaal Statistische Bestanden. Op deze wijze wordt de éénprocentsteekproef langzamerhand geen statisch bestand meer maar een longitudinaal bestand, waarin veranderingen op individueel niveau gevolgd en geanalyseerd kunnen worden en geconfronteerd worden met veranderingen op macroniveau. Het is erg belangrijk om in staat te zijn een onderscheid te kunnen maken tussen veranderingen op het individuele niveau (zeg het aantal verhuizingen) en het aggregaat van al deze verhuizingen (zeg een verschuiving tussen provincies en gemeenten). Uiteraard kunnen niet alle gegevens uit de eerdere Sociaal Statistische Bestanden gekoppeld worden aan de recentste éénprocentsteekproef,

want dan wordt hij onhanteerbaar. Maar de creatie op lange termijn van grote steekproeven met longitudinale gegevens die ver terugreiken, zou wel eens de beste dienst zijn die het CBS de sociale wetenschappen kan bewijzen. Het verschaft zo essentiële bouwstenen voor een goed empirisch onderzoek en gefundeerde theorieën. Daarbij is het ook van belang in deze longitudinale bestanden op te nemen in welk jaar en maand een bepaalde gebeurtenis plaatsvond (bijvoorbeeld aanvang arbeidsongeschiktheid of dienstverband, een geboorte, etc). Dat kan goed, omdat in het SSB alles al op gebeurtenisniveau en datum is vastgelegd. Toepassing van *event-history analysis* maar ook andere tijdsafhankelijke technieken laten zien hoe belangrijk tijd is voor het verloop van menselijke processen (bijvoorbeeld de lengte van een samenwonenrelatie of een huwelijksrelatie voor scheidingskansen). Zonder het betrekken van dat tijdsverloop zullen onze analyses slechter zijn dan nodig. Het systematisch inbrengen van een longitudinale dimensie in de éénprocentsteekproef (en dus ook in de virtuele volkstelling) betekent een unieke verbetering van volkstellingen, niet alleen in Nederland maar ook in Europa.

Om verdere kosten te drukken moet het CBS de bestanden ter beschikking stellen via een aparte pagina van zijn homepage. Dit is erg gebruikelijk aan het worden, ook met zeer gevoelige data zoals de PISA-data van de OECD (een crossnationale vergelijking van onderwijskwaliteit in alle OECD-landen). De data moeten dan als een kaal databestand ter beschikking gesteld worden, met daarbij de syntax om deze datafile te kunnen omzetten in een systeemfile van de belangrijkste sociaal-wetenschappelijk software (SPSS, STATA, SAS).

Ook de handleidingen en gebruikerstips kunnen via deze homepage opvraagbaar zijn. De homepage van de PISA-data (http://pisaweb.acer.edu.au/oecd/oecd_pisa_data_s1.html) is een goed voorbeeld hoe zo'n pagina ingericht moet worden.

Alle variabelen moeten zo gedetailleerd mogelijk (binnen de grenzen van de wet) beschikbaar gesteld worden. Het samenvoegen van categorieën of variabelen is niet de taak van het CBS. Omdat het om een steekproef gaat zijn ook unieke combinaties binnen de steekproef irrelevant in verband met de privacywetgeving: wat binnen de steekproef een unieke combinatie is, hoeft dat binnen het volledige bestand niet te zijn. Herkenning van individuen op basis van de steekproef is derhalve uitgesloten. Hierboven benadrukte ik al dat de grenzen van de wet gerespecteerd moeten worden bij de beschikbaarstelling van deze gedetailleerde steekproef. Maar de grenzen van de wet liggen ruimer dan de huidige cultuur van beschikbaarstelling van gegevens door het CBS. Die huidige cultuur lijkt nog steeds meer ingegeven door angst voor de belanghebbende berichtgevers en 'kritische' kamervragen dan door zijn wettelijk taak te zorgen voor goede én relevante statistieken over de Nederlandse samenleving. De mate van detaillering van geboorteland en nationaliteit in de éénprocentsteekproef lijkt mij een goede maatstaf om vast te stellen in welke mate het CBS in de toekomst zijn wettelijke taak wil uitvoeren. In elk geval moet de mate van gedetailleerdheid in de steekproef nooit minder zijn dan die van de tabellen, die het CBS zelf publiceert. Als dat laatste wel het geval zou zijn, plaatst het CBS zich buiten de vrije wetenschap, want het is oncontroleerbaar voor het wetenschappelijk

forum geworden. Het is dan een van de bovengenoemde ordinaire overheidsorganen geworden, een speelbal van belanggroepen en politieke partijen.

De enige verplichting die het CBS aan gebruikers zou mogen stellen is dat zij zo snel mogelijk laten weten welk gebruik zij hebben gemaakt van de steekproefdata en waar dat gepubliceerd is. Deze lijst van publicaties, zo mogelijk van links naar de betrokken internet-pagina's, wordt ook ondergebracht op de steekproef-homepage van het CBS. Alle geïnteresseerden kunnen dan zien welke analyses er al verricht zijn en kunnen er ook kennis van nemen. Door deze publicatielijst apart te houden van haar reguliere CBS-publicaties, is het CBS niet verantwoordelijk voor de inhoud van de publicaties maar laat het wel zien wat er met de steekproef gebeurt en legt zo verantwoording af van haar wettelijke taak.

3. Enige mogelijkheden van de steekproef voor de wetenschap.

Wat zou een sociale wetenschapper met de éénprocent-steekproef kunnen doen? De reikwijdte van mijn fantasie en kennis en de beschikbare ruimte is beperkt, dus ik zal maar twee mogelijkheden aangeven, naast de al eerder aangestipte mogelijkheid om een longitudinaal bestand op te bouwen. Al deze mogelijkheden lijken mij echter al ruim voldoende om een éénprocentsteekproef te rechtvaardigen.

De eerste mogelijkheid lijkt mij dat er voor het eerst sinds 1972 weer een referentiepunt kan ontstaan om de representativiteit van sociaal-wetenschappelijk onderzoek

aan af te meten. Ik wil nu niet het hele debat herhalen over de non-betrouwbaarheid van verkiezingsonderzoek en ander politicologisch onderzoek als gevolg van de hoge non-response van de meeste surveys. Aan de hoge non-response in Nederland kunnen wij als wetenschappers weinig doen, maar wij moeten wel beter de gevolgen voor de kwaliteit van ons werk onder ogen zien. Tot nu toe behandelen wij de extreem hoge non-response in Nederland als het zwarte schaap van de familie: je kunt er beter over zwijgen en zeker in de nabijheid van derden. De éénprocentsteekproef kan ons helpen daarmee beter om te gaan, want het biedt weer een goed en betrouwbaar referentiepunt voor non-response analyses.

De tweede mogelijkheid is de ontwikkeling van de buurt, gemeente en regio statistiek door een consequentie toepassing van een multilevel aanpak. Door het geruime tijd ontbreken van volkstellinggegevens is de statistiek van buurt, gemeente of regio weggeschrompeld: die kan namelijk alleen goed van de grond komen als men microkenmerken (zeg van individuen) kan koppelen aan macrokenmerken (zeg van regio's, gemeenten en buurten). Als men geen microgegevens aan macrokenmerken kan koppelen, zal men in de analyse van de effecten van die macrokenmerken zeker fouten maken. Over het algemeen zal men het belang van de macrokenmerken overschatten ten koste van het belang van de individuele kenmerken. Het effectieve scholenonderzoek van de laatste decennia heeft op dit punt een harde les geleerd. Schoolkenmerken zijn voor de voorspelling van schoolsucces niet onbelangrijk, maar individuele kenmerken zijn nog veel belangrijker. Ook is duidelijk dat er nog steeds verschillen naar buurt, gemeenten en regio bestaan en dat die hardnekkiger

zijn dan men jaren lang dacht. Het is daarom van groot belang aan de éénprocentsteekproef buurt, gemeente en regiokenmerken toe te voegen, zodat er een gecombineerd bestand van macro- en microgegevens ontstaat. Om binnen de grenzen van de privacy wetgeving te blijven kan dat het best door een groot aantal geaggregeerde buurt, gemeente en regio kenmerken te maken op basis van het volledige Sociaal Statistisch Bestand en vervolgens die geaggregeerde kenmerken toe te voegen aan de éénprocentsteekproef, via de postcode. Als er genoeg geaggregeerde macrokenmerken aan het bestand toegevoegd zijn, zal de noodzaak om de concrete naam van buurt, gemeente of regio in de éénprocentsteekproef op te nemen wegnemen. Welke geaggregeerde kenmerken dat precies zijn (inkomen, percentage werklozen, etc.), zal via nader overleg tussen CBS en wetenschappers vastgesteld moeten worden. Ook is het denkbaar dat externe buurt, gemeente en regiokenmerken op dezelfde manier aan de éénprocentsteekproef gekoppeld worden.

4. Is het Sociaal Statistisch Bestand wel compleet?

Het huidige Sociaal Statistisch Bestand is uitermate beperkt en weinig fantasievol samengesteld. Om de bruikbaarheid te vergroten zouden een aantal verbeteringen en aanvullingen mogelijk zijn. Ik beperk mij hierbij tot de bestaande bronnen en mogelijkheden, om luchtftietserij te voorkomen.

Merkwaardigerwijs behandelt het Sociaal Statistisch Bestand de Nederlandse bevolking alsof het een losse verzameling individuen is. Alleen de variabele gezinsrelaties doet vermoeden dat ze nog samenlevings-relaties met andere

Nederlanders zijn aangegaan en dat ze wellicht ook kinderen, partners of ouders hebben met wie zij samenwonen. De individualistische benadering van Nederlanders mag misschien wel politiek correct zijn, maar ze strookt in het geheel niet met het belang van kenmerken van partners, ouders en kinderen voor de levensloop van mensen. Het schoolsucces van kinderen wordt bijvoorbeeld door beroep, opleiding en inkomen van ouders beïnvloed; succes op de arbeidsmarkt van gehuwde vrouwen hangt mede af van de arbeidsmarktpositie van hun echtgenoten. En zo zou ik nog wel een tijdje kunnen doorgaan met sociaal-wetenschappelijk onderzoek aan te halen over het belang van de gezinsrelaties voor het functioneren van de samenleving. Zelfs economen, in stilzwijgende imitatie van sociologen, onderschrijven dat tegenwoordig blijkens publicaties van Nobelprijswinnaars als Gary S. Becker (1981). Men kan het groeiend belang van gezinsrelaties in een individualiserende samenleving het best aangeven met de titel van een boek over het gezinsleven in de USA: *A haven in a heartless world*. En naarmate die wereld individualistischer wordt, neemt het belang van die gezinsrelatie toe, juist voor de zwakkere in die samenleving. Het is daarom beter de éénprocentsteekproef van individuen uit het Sociaal Statistisch Bestand om te bouwen naar een éénprocentsteekproef uit de Nederlandse huishoudens. Omdat de gemeentelijke basisadministratie ten grondslag ligt aan dit Sociaal Statistisch Bestand moet het mogelijk zijn daarmee huishoudens te reconstrueren. Dat huishouden omvat dan alle personen die op hetzelfde adres wonen en de steekprofeenheid wordt dan ook huishouden. In dat huishouden zijn alle beschikbare kenmerken van de leden van dat huishouden opgenomen. De mogelijkheden van analyse zullen hierdoor sterk toenemen, evenals de relevantie van de uitkomsten. Er is tegenwoordig genoeg

ervaring met huishoudsurveys om daaruit bestanden te construeren met individuele personen. Maar een bestand met alleen individuele personen kan nooit meer worden omgebouwd tot een huishoudsurvey. Uitgaande van de huidige individuele volledige bestanden uit het Sociaal Statistisch Bestand betekent dit in mijn ogen dat het huishouden de steekproef eenheid moet zijn en niet de losse individuen. Alle individuen uit alle getrokken huishouden worden vervolgens in de éénprocentsteekproef opgenomen. Op deze wijze kunnen mensen in hun 'natuurlijke' setting geanalyseerd worden. Uiteraard betekent dit ook een gestratificeerde steekproef, om te zorgen dat er voldoende verschillende typen huishouden in voorkomen. Maar dat lijkt mij niet erg omdat de éénprocentsteekproef toch gestratificeerd zal moeten worden, om verschillende kleinere categorieën voldoende in te laten voorkomen. Bovendien zijn weegfactoren tegenwoordig geen probleem meer voor de meeste statistische pakketten.

Tot slot wil ik nog wijzen op de mogelijkheden die het onderwijsnummer gaat geven voor de kwaliteit van het Sociaal Statistisch Bestand. Omdat onderwijs in de huidige samenleving de ruggengraat van de maatschappelijke ongelijkheid is geworden, is het van eminent belang tijdig met de integratie van deze gegevens in het Sociaal Statistisch Bestand te beginnen. Omdat het nog geruime tijd duurt voordat het onderwijsnummer zo ver doorgevoerd is dat het bruikbaar voor het Sociaal Statistisch Bestand is, zal zo spoedig mogelijk naar andere bestanden (registers van de CWI's en de vroegere arbeidsbureaus, andere onderwijsregisters en enquêtes) omgezien moeten worden. Een Sociaal Statistisch Bestand en een daaruit afgeleide éénprocentsteekproef zonder een goede meting van

de centrale ongelijkheidsindicator zal vruchteloos zijn. Het onderwijsnummer kan echter onmiddellijk op een andere manier gebruikt worden: het vaststellen van de geaggregeerde kenmerken van scholen waarop de kinderen uit de éénprocentsteekproef zitten. Het beschikbaar komen van deze geaggregeerde schoolkenmerken (met name leerlingcompositie, type onderwijs, sector) zal aan het Sociaal Statistisch Bestand al onmiddellijk een groot wetenschappelijk en maatschappelijk gewicht geven, omdat dan de discussie over segregatie tussen scholen een steviger en onafhankelijker grondslag kan krijgen.

Noten

- (1) Jaap Dronkers is als hoogleraar verbonden aan het European University Institute, San Domenico di Fiesole (FI), Italy. E-mail: jaap.dronkers@iue.it. Personal homepage: <http://www.iue.it/Personal/Dronkers>
- (2) De inhoud van het onderwijshoofdstuk uit het rapport van het Verweij-Jonker Instituut over de integratie van migranten in de Nederlandse samenleving laat de klemmende omhelzing van de wetenschap door de belanghebbende departementen treffend zien, evenals het buitensluiten van alle meer kritische stemmen van meer onafhankelijke wetenschappers.

Referenties

- Becker, G.S., 1981, A treatise on the family (Chicago: Chicago University Press)

De virtuele volkstelling als databron voor sociaal-wetenschappelijk onderzoek; optimaliseringsvragen

Een reactie op de papers van J. van Maarseveen en J. Dronkers

J. Gierveld (1)

Met Van Maarseveen is Gierveld het eens dat we zorgvuldig met de oude Volkstellingen moeten omgaan. Ze behoren tot ons culturele erfgoed. Ze is dan ook erg blij dat de vroegere Volkstellingen gedigitaliseerd zijn. Gierveld vindt dat we dergelijke projecten moeten koesteren. Dat is erg belangrijk voor onderzoek. De digitalisering is een mooi voorbeeld van het vergroten van de toegankelijkheid van informatie. In de jaren 80 waren bestanden van het CBS nog erg duur waardoor de toegankelijkheid beperkt was. StatLine is ook een mooi voorbeeld van vergroting van de toegankelijkheid. Evenals Van Maarseveen zou Gierveld graag zien dat de censusmonografieën van 1971 een vervolg krijgen. Waarover Gierveld met Van Maarseveen van mening verschilt, is de mate van internationale vergelijkbaarheid. Zij wordt wanhopig van de slechte internationale vergelijkbaarheid van gegevens. Er is een groot verschil in definities. Zo blijkt dat er geen eenduidige definitie voor huishouden is. Wat moet er dan bijvoorbeeld gebeuren wanneer de data van Nederland worden vergeleken met die van Italië? Hoe onderscheiden we een huishouden waarin de 35-jarige zoon nog bij zijn ouders inwoont van een huishouden waarin ouderen – om welke

reden ook – zijn gaan samenwonen met een volwassen zoon of dochter? Ook de peildata variëren. Voor het doen van internationaal vergelijkend onderzoek is dat buitengewoon lastig.

Ook met Dronkers is Gierveld het in grote lijnen eens. Zij verwacht dat het beschikbaar stellen van een 1 % steekproef erg nuttig is, zeker om multivariaat en multilevelonderzoek uit te voeren. Alleen wil zij die informatie niet op huishoudensbasis hebben, zoals Dronkers nu voorstelt. Het is belangrijk om meer te weten over een persoon en zijn omgeving, maar die omgeving is breder dan het gemeenschappelijk huishouden. Bovendien is dat huishouden een moeilijk te analyseren eenheid; huishoudens wisselen steeds van vorm. Stel dat een man en vrouw trouwen, kinderen krijgen, maar na verloop van tijd op elkaar uit gekeken raken en besluiten te scheiden. De moeder krijgt de kinderen en vindt uiteindelijk weer een nieuwe partner waar ze mee trouwt, waardoor er een stiefvader in beeld komt. Wat is in dat geval dan de betekenis van het huishouden? Volgens Gierveld is het daarom beter te kijken naar familiale relaties in plaats van huishoudens. De voorkeur van Gierveld gaat uit naar een personensteekproef met daaraan gekoppeld huishoudenskenmerken behorend bij de persoon.

Noten

- (1) Jenny Gierveld is emeritus hoogleraar aan de Vrije Universiteit en Honorary Fellow van het NIDI.

De ‘virtuele volkstelling’ 2001

E. Schulte Nordholt (1)

1. Inleiding

In 2003 zijn de gegevens samengesteld voor de Volkstelling 2001 (VT 2001). Het betreft een veertigtal uitgebreide tabellen. Daarvan gaan er achtentwintig over Nederland als geheel, zijn er negen op COROP-niveau (NUTS 3) en drie op gemeenteniveau (NUTS 5). De veertig tabellen vallen in een aantal groepen uiteen. Een achttal tabellen gaat over wonen, twee tabellen gaan over woon-werkverkeer en de overige dertig tabellen zijn demografisch getint met een link naar onder andere beroep, opleidingsniveau en bedrijfstak waar mensen werken. Naast deze veertig tabellen zijn er twee tabellen samengesteld op binnengemeentelijk niveau voor de tien grote steden die aan de zogenaamde *Urban Audit* meedoen. Deze tien grote steden zijn: Amsterdam, Rotterdam, Den Haag, Utrecht, Eindhoven, Tilburg, Groningen, Enschede, Arnhem en Heerlen.

Voor Nederland is geen echte telling onder de bevolking gehouden, maar is uitsluitend van bij het CBS al beschikbare bronnen gebruik gemaakt. Hierdoor is de Nederlandse belastingbetaler aanzienlijk goedkoper uit. De kosten van een traditionele Volkstelling zouden in de orde van grootte van een paar honderd miljoen euro liggen terwijl de nu gemaakte kosten een paar miljoen euro bedragen. Bij deze laatste ruwe schatting zijn inbegrepen

alle voorbereidende werkzaamheden zoals het ontwikkelen van een nieuwe methodologie en bijbehorende software.

Behalve het financiële aspect zijn er nog andere belangrijke verschillen tussen een traditionele Volkstelling en een virtuele Volkstelling. Bij een traditionele Volkstelling zal, ondanks het verplichte karakter, altijd een deel van de bevolking niet meedoen (*unit nonrespons*) en door het deel dat wel meedoet, zullen sommige vragen niet worden beantwoord (*item nonrespons*). Met weeg- en imputatietechnieken kan worden geprobeerd hiervoor te corrigeren. Een bekend probleem bij traditionele Volkstellingen is een beperkte en selectieve participatie waardoor traditionele correctiemethoden niet meer afdoende zijn om betrouwbare uitkomsten te kunnen publiceren. In de Britse Volkstelling van 2001 is dit probleem bijvoorbeeld aan de orde. Er is in het Verenigd Koninkrijk veel maatschappelijke kritiek op de eerste uitkomsten van de in 2001 gehouden Volkstelling en inmiddels heeft het Britse Office for National Statistics (ONS) deze uitkomsten moeten herzien. Bij de laatste traditionele Volkstelling in Nederland (die van 1971) was er veel bezwaar tegen het verzamelen van integrale informatie over de in Nederland woonachtige bevolking. Hierdoor werd het non-responsprobleem groter en de verwachting is dat als nu een traditionele Volkstelling in Nederland zou worden gehouden de non-respons nog veel hoger zou zijn. Tegen een virtuele Volkstelling bestaan nauwelijks bezwaren en het non-responsprobleem speelt alleen bij de enquêtes waarvan gebruik wordt gemaakt.

In Nederland is later met de (virtuele) Volkstelling gestart dan in andere landen waar een traditionele Volkstelling

wordt gehouden. Het was pas zinvol daadwerkelijk met het Volkstellingsproject 2001 te starten toen alle bronnen gereed waren; enkele registers waren pas relatief laat beschikbaar. Toch is Nederland eerder gereed met het samenstellen van de veertig genoemde tabellen dan de meeste andere landen die een Volkstelling 2001 houden. In Nederland hoeven immers niet eerst alle binnenkomende formulieren te worden verwerkt en gecontroleerd. Wel moet worden bedacht dat voor sommige variabelen slechts steekproefinformatie beschikbaar is en dat betekent dat voor Nederland niet in alle tabellen het gewenste detailleringniveau kan worden gehaald. De voordelen van de virtuele Volkstelling - goedkoper en vrijwel ontbreken van het non-responsprobleem - wegen ruimschoots op tegen het wat beperktere detail in de uitkomsten vergeleken met een traditionele Volkstelling.

In de Noordelijke landen (Denemarken, Finland, Noorwegen, IJsland en Zweden) zijn meer variabelen in registers beschikbaar dan in Nederland en daar speelt het probleem minder dat een beperkter detail in de uitkomsten beschikbaar komt. Voor ontbrekende variabelen wordt in sommige Noordelijke landen bovendien een (beperkte) telling gehouden. De meeste andere landen bevinden zich in een vergelijkbare situatie als Nederland waarbij sommige voor de Volkstelling relevante variabelen in registers kunnen worden gevonden, terwijl andere variabelen alleen op steekproefbasis beschikbaar zijn. Daarom bestaat er veel belangstelling voor de Nederlandse aanpak om registers en enquêtes te combineren en moderne statistische technieken en bijbehorende software te gebruiken. Cruciaal is natuurlijk dat statistische bureaus gebruik mogen maken van de voor

de Volkstelling relevante registers. De mogelijkheden voor het Nederlandse CBS om gebruik te maken van allerlei registers worden versterkt in de nieuwe CBS-wet. Toch zal het CBS de komende jaren veel zorg moeten besteden aan de relaties met de registerhouders. Tijdige leveringen met de voor het CBS relevante variabelen zijn cruciaal om de Nederlandse statistiekproductie te kunnen blijven waarborgen.

De aanleiding dat het CBS is overgegaan tot het samenstellen van de tabellenset is het *gentlemen's agreement* dat is gesloten. Er bestaat dus geen (Europese) verplichting tot levering van Volkstellingsgegevens, maar het is vrijwel ondenkbaar dat Nederland niet net als alle andere Europese landen Volkstellingsgegevens samenstelt voor de internationale organisaties. Binnen de Europese Unie heeft Eurostat hierin een coördinerende rol en een taak om internationale vergelijkingen te maken van de uitkomsten. Het CBS zal de uitkomsten van 2001 vergelijken met de uitkomsten van eerdere Nederlandse Volkstellingen. In het verleden zijn dergelijke analyses ook uitgevoerd (zie hiervoor bijvoorbeeld CBS, 2002).

2. Methode van samenstellen

De huidige virtuele Volkstelling heeft als peildatum 1 januari 2001. De tabellenset bestaat geheel uit frequentietabellen; er komen dus geen kwantitatieve tabellen voor. Het SSB bestaat uit een reeks koppelbare bestanden die onderling consistent zijn gemaakt. Er worden een aantal onderdelen van het SSB gebruikt. De algemene werkwijze van het SSB en de bronnen die

zijn gebruikt zijn beschreven in Arts en Hoogteijling (2002). De ruggengraat van de Volkstelling wordt gevormd door de Gemeentelijke Basisadministratie (GBA). GBA-gegevens van 1 januari 2001 zijn gebruikt als basis voor de tabellenset. Het banenbestand 2001 (het onderdeel van het SSB met geïntegreerde gegevens over onder andere werknemers en zelfstandigen, zie Bakker en Arts, 2003) is pas eind 2003 beschikbaar gekomen en daarop kon niet worden gewacht. Daarom is gewerkt met het banenbestand van 2000 dat wel beschikbaar was en waaruit de gegevens van eind 2000 zijn afgeleid als benadering voor de situatie op 1 januari 2001. Voor verschillende variabelen, zoals beroep en hoogst behaalde opleidingsniveau, is uit de EBB geput. Om voldoende massa te verkrijgen is informatie uit de EBB 2000 en de EBB 2001 gecombineerd. Voor de woontabellen is gebruik gemaakt van GBA-gegevens van 1 januari 2001, het Woningregister 2001 en het Woningbehoeftenonderzoek (WBO) 2000. Voor de tabellen over het woon-werkverkeer is gebruik gemaakt van de GBA-gegevens van 2000 en 2001, de Enquête werkgelegenheid en lonen (EWL) 2000 en het banenbestand over 2000.

Sommige variabelen (uit de GBA en het banenbestand) zijn integraal beschikbaar zoals leeftijd, geslacht, burgerlijke staat en het al dan niet werknemer zijn. Andere variabelen (uit de EBB en het banenbestand) zijn slechts voor een deel van de personen beschikbaar. Voorbeelden hiervan zijn opleidingsniveau en of iemand voltijder is of niet. Teneinde consistentie tussen de tabellen te kunnen garanderen is gebruik gemaakt van de techniek van het herhaald wegen. De methode van het herhaald wegen is uitvoerig beschreven in Houbiers et al. (2003). Met

herhaald wegen worden de ophooggewichten van records in de microdata zodanig gekalibreerd dat een nieuwe tabelschatting consistent is met alle eerder gemaakte tabelschattingen.

Om de techniek van het herhaald wegen toe te kunnen passen is gebruik gemaakt van de nieuwste versie van het op het CBS ontwikkelde software pakket VRD. De letters VRD staan voor Vullen Reference Database en het doel van de applicatie is het vullen van en beheer over de Reference database. De belangrijkste functies van VRD zijn het schatten van tabellen via herhaald wegen, het toevoegen van deze tabellen aan de Reference database, en het opvragen van aggregaten uit de Reference database. Onder de voorwaarde van kleine, onafhankelijke steekproeven, kunnen ook de varianties van de tabelwaarden worden geschat. Het schatten van de tabellen gebeurt niet in VRD zelf, maar vindt plaats in Bascula 4.0.

Om elke tabel zo nauwkeurig mogelijk te kunnen schatten, is elke schatting gebaseerd op een zo groot mogelijk aantal waarnemingen. Tabellen die uitsluitend registervariabelen bevatten, worden geteld uit de registers. Tabellen die ten minste één variabele bevatten uit een enquête worden geschat uit de grootst mogelijke combinatie van registers en steekproeven. Door de combinatie van registers en steekproeven zijn blokken gevormd, waaruit de VT-tabellen zijn geschat. Ter illustratie is hieronder een zestal blokken weergegeven op basis waarvan de VT-tabellen voor de economisch actieven (werkenden en werklozen samen) zijn geschat:

1. Het registerblok.
2. Het SBI-blok (alle records uit het registerblok waarvoor de internationale code voor economische activiteit NACE bekend is, evenals de niet-werkzamen).
3. Het EWL-blok.
4. Het EWL-SBI-blok.
5. Het Economische Activiteitblok (in feite is dit het EBB-blok, aangevuld met informatie over de werkenden en gepensioneerden).
6. Het EBB-blok.

De blokken 2 tot en met 6 zijn op basis van steekproefdata samengesteld. Om schattingen voor de hele populatie te maken moeten ophooggewichten worden bepaald. Deze gewichten hangen af van de precieze samenstelling van het betreffende blok (één of meerdere steekproeven), het design van de steekproeven, de non-responscorrectie van de steekproeven, het eventueel trachten de variantie te reduceren met behulp van hulpinformatie en het al zo veel mogelijk bereiken van consistentie. Volledige consistentie is niet altijd mogelijk, bijvoorbeeld als er te veel restricties zijn opgelegd.

Voor de samenstelling van de VT-tabellen worden de gewichten uit de blokken telkens gekalibreerd naar alle relevante registertellingen en de eerder geschatte tabellen uit de blokken zodat alle tabellen onderling consistent zijn. Elke tabel moet worden berekend uit het grootste blok waaruit de tabel kan worden bepaald. Als alle tabellen met de juiste gewichten worden geschat, zijn de tabeluitkomsten onderling consistent. Door telkens uit te gaan van het grootst mogelijke blok, kan de grootst mogelijke detaillering in de VT-tabellen worden bereikt.

De cijfers van de VT 2001 hebben betrekking op personen woonachtig in Nederland op 1 januari 2001 (teleenheid personen). De personen die volgens de GBA aan het begin van die dag in Nederland woonachtig waren worden 'geteld' in de virtuele Volkstelling. Het grootste deel van de Nederlandse bevolking is woonachtig in particuliere huishoudens en de overigen maken deel uit van institutionele huishoudens (Instellingen, Inrichtingen en Tehuizen, IIT). De aantallen werknemers in de tabellen hebben betrekking op ultimo 2000 waarbij 22 december 2000 als peildatum is gebruikt om het aantal banen van werknemers in Nederland vast te stellen. In het gebruikte banenbestand is informatie over de banen van werknemers vastgelegd. Dat wil zeggen dat een werknemer meerdere keren kan voorkomen als hij of zij meerdere banen tegelijkertijd heeft. In de tabellenset zijn de kenmerken van de hoofdbaan gebruikt, waarbij de hoofdbaan van een werknemer is gedefinieerd als de baan met het hoogste bruto loon voor de sociale verzekeringen.

Voor sommige uitkomsten van de VT 2001 moet rekening worden gehouden met onnauwkeurigheidsmarges, omdat de VT 2001 gedeeltelijk op basis van steekproefgegevens is samengesteld. In verband met de betrouwbaarheid van de uitkomsten worden vuistregels gehanteerd voor celwaarden die zijn gebaseerd op een steekproef uit de VT-populatie. De exacte onnauwkeurigheidsmarges zijn niet te geven vanwege het samenstellen van de blokken uit de steekproeven en het complexe steekproefontwerp van deze steekproeven. De vuistregels zijn afgeleid op basis van de veronderstellingen dat de twee EBB-bestanden (voor 2000 en 2001) één steekproef vormen en dat de 'insluitkansen' voor deze steekproef worden gegeven door

de blokgewichten van het EBB-blok. De vuistregels voor waarnemingen van personen uit de EBB luiden als volgt:

- Cellen met minder dan 10 waarnemingen worden altijd onderdrukt.
- Cellen met 25 of meer waarnemingen worden altijd gepubliceerd.
- Cellen met 10–24 waarnemingen worden alleen gepubliceerd als ze deel uitmaken van een uitsplitsing (naar leeftijd of geslacht), waarin geen cellen met minder dan 10 waarnemingen voorkomen, en ten minste 50 procent van de cellen in de uitsplitsing meer dan 25 waarnemingen heeft. De grens van 25 waarnemingen komt overeen met een geschatte relatieve onnauwkeurigheid van ten hoogste 20 procent (i.e. de geschatte marges bedragen maximaal 40 procent).

De vuistregels voor waarnemingen uit het WBO zijn van dezelfde vorm. Er worden echter wat hogere drempelwaarden gehanteerd in verband met het feit dat de steekproefomvang van het WBO wat beperkter is dan die van de EBB. Voor cellen met als teleenheid huishoudens of woningen in Nederland worden analoge vuistregels gebruikt.

3. Uitkomsten Volkstelling 2001

3.1 Bevolking naar geslacht, leeftijd en type huishouden

Aan het begin van 2001 waren er in Nederland in totaal 16,0 miljoen mensen woonachtig, 7,9 miljoen mannen en 8,1 miljoen vrouwen. In de leeftijdscategorieën 0-14 en 15-74 jaar waren er iets meer mannen dan vrouwen, maar

in de categorie 75 jaar of ouder waren er bijna tweemaal zoveel vrouwen als mannen. Meer dan 200 duizend mensen woonden in institutionele huishoudens zoals verpleeg- en verzorgingshuizen. Van deze groep was ongeveer 36 procent man en 64 procent vrouw. Van de mensen in institutionele huishoudens was 57 procent 75 jaar of ouder. Meer informatie over de bevolking naar geslacht, leeftijd en type huishouden is te vinden in tabel 1.

3.2. Bevolking naar economische activiteit

Van de begin 2001 in Nederland woonachtigen behoorde iets minder dan de helft tot de economisch actieve bevolking (beroepsbevolking). De werkende beroepsbevolking omvatte 7,4 miljoen mensen: 6,8 miljoen werknemers en 0,6 miljoen zelfstandigen. De werkloze beroepsbevolking was een kleine 200 duizend mensen groot. Bij deze aantallen moet bedacht worden dat in de opzet van de VT de werknemers, zelfstandigen en werklozen elkaar uitsluitende categorieën zijn. Dit impliceert dat zelfstandigen die ook voor een aantal uren per week in loondienst werkzaam zijn, worden gerekend tot de werknemers. Het aantal werklozen is geschat op basis van steekproefinformatie, waarbij iemand die tot de werkende beroepsbevolking behoort niet tegelijkertijd tot de werkloze beroepsbevolking wordt gerekend. Van de economisch actieve bevolking was 58 procent man, terwijl van de economisch inactieve bevolking 58 procent vrouw was. Tot de economisch inactieven worden onder meer gerekend onderwijsvolgenden, gepensioneerden en huisvrouwen en huismannen. Het aantal huisvrouwen is meer dan 18 maal zo groot als het aantal huismannen.

Tabel 1. Bevolking naar geslacht, leeftijd en type huishouden

	Totaal	0-14	15-74	75+
Totale bevolking	15.985.538	2.977.283	12.036.171	972.084
Man	7.909.052	1.522.811	6.047.425	338.816
Vrouw	8.076.486	1.454.472	5.988.746	633.268
Bevolking in particuliere huishoudens	15.766.606	2.970.545	11.947.996	848.065
Man	7.829.914	1.518.611	5.998.189	313.114
Vrouw	7.936.692	1.451.934	5.949.807	534.951
		0-14	15-74	75+
Bevolking in IIT	218.932	6.738	88.175	124.019
Man	79.138	4.200	49.236	25.702
Vrouw	139.794	2.538	38.939	98.317

Meer informatie over de bevolking naar geslacht, leeftijd en type huishouden is te vinden in tabel 2.

Economische activiteit naar bedrijfssector

De 7,4 miljoen leden tellende werkende beroepsbevolking kunnen we onderverdelen naar bedrijfssector met behulp van de NACE-code. Voor een werknemer die meer dan één betrekking heeft wordt gekeken naar de kenmerken van zijn hoofdbaan. De hoofdbaan van een persoon is in het kader van de VT gedefinieerd als de baan waarmee in 2000 het hoogste loon vóór de sociale verzekeringen werd behaald. Op deze wijze bleken er (begin 2001) 0,2

miljoen werkenden te zijn in de bedrijfssector landbouw en visserij, 1,5 miljoen in de industrie en bouwnijverheid en 5,7 miljoen in de dienstverlening. Van de werkenden in de dienstverlening waren er 3,5 miljoen werkzaam in de commerciële dienstverlening en ruim 2,1 miljoen in de niet-commerciële dienstverlening.

Arbeidsduur van werknemers

Van de 6,8 miljoen werknemers kan worden bekeken hoeveel uur zij per week werken in hun hoofdbaan. Fulltime (≥ 35 uur per week) werken 4,2 miljoen werknemers, 1,8 miljoen werknemers hebben een grote

Tabel 2. Bevolking naar economische activiteit

	Totaal	Man	Vrouw
Economisch actieve bevolking	7.586.914	4.388.239	3.198.675
Werkende	7.394.777	4.287.967	3.106.810
Werknemer	6.786.511	3.883.813	2.902.698
Zelfstandige	608.266	404.154	204.112
Werkloze	192.137	100.272	91.865
Economisch inactieve bevolking			
Alle leeftijden	8.398.624	3.520.813	4.877.811
15-74	4.449.257	1.659.186	2.790.071
Onderwijsvolgend	640.446	342.934	297.512
Gepensioneerd	1.355.940	620.493	735.447
Huisvrouw /-man	1.270.420	65.821	1.204.599
Anders economisch inactief	1.182.451	629.938	552.513

parttime baan (minder dan 35 uur, maar ten minste 15 uur per week) en 0,8 miljoen hebben een kleine parttime baan (minder dan 15 uur per week). Van de fulltime werknemers is 77 procent man, terwijl van de parttime werknemers 75 procent vrouw is. Meer informatie over de arbeidsduur van werknemers is te vinden in tabel 3.

3.3. Werkenden naar beroep

Met behulp van de *International Standard Classification of Occupations* (ISCO) kunnen werkenden worden ingedeeld naar beroep. Bij de mannen zijn de beroepsgroepen wetenschappelijke beroepen, parlementariërs, landsbe-

stuurders, leidinggevende beroepen en handwerks- en ambachtsslieden de meest voorkomende. Bij de vrouwen zijn dat de beroepsgroepen hogere en middelbare technici en vakspecialisten, administratieve beroepen en lagere dienstverlenende en commerciële beroepen. Meer informatie over de werkenden naar beroep is te vinden in tabel 4.

3.4. Bevolking naar opleidingsniveau

Met behulp van de *International Standard Classification of Education* (ISCED) kan de in Nederland woonachtige bevolking worden ingedeeld naar opleidingsniveau. Bij de 12 miljoen mensen in de leeftijdscategorie 15 – 74 jaar

Tabel 3. Werknemers naar arbeidsduur

	Totaal	Man	Vrouw
Werknemers	6.786.511	3.883.813	2.902.698
Fulltime (≥ 35 uur per week)	4.222.228	3.236.504	985.724
Parttime totaal	2.564.283	647.309	1.916.974
Lang parttime (15- <35 uur per week)	1.793.656	419.071	1.374.585
Kort parttime (<15 uur per week)	770.627	228.238	542.389

Tabel 4. Werkenden naar beroep

		Totaal	Man	Vrouw
Werkenden		7.394.777	4.287.967	3.106.810
1	Parlementariërs, landsbestuurders, leidinggevende beroepen	926.631	695.563	231.068
2	Wetenschappelijke beroepen	1.205.163	705.357	499.805
3	hogere en middelbare technici en vakspecialisten	1.248.759	607.819	640.939
4	Administratieve beroepen	841.219	271.862	569.358
5	lagere dienstverlenende en commerciële beroepen	800.629	259.173	541.456
6	vakkrachten in landbouw en visserij	105.256	78.280	26.976
7	handwerks- en ambachtsslieden	712.093	677.256	34.837
8	bedieners fabrieksinstallaties en machines en montagemedewerkers	446.722	398.845	47.877
9	elementaire beroepen	522.901	272.435	250.467
0	Strijdkrachten	37.032	34.227	2.805
99	Onbekend	548.374	287.151	261.223

komt het secundaire niveau het vaakst voor en zijn er meer mensen met een hoogst behaalde opleiding op het tertiaire niveau dan op het primaire niveau. Voor de groep van 75 jaar en ouder is het secundaire niveau ook het meest voorkomende niveau, maar in deze groep hebben aanzienlijk meer mensen een hoogst behaalde opleiding op primair dan op tertiair niveau. Meer informatie over de bevolking naar opleidingsniveau is te vinden in tabel 5.

4. Slot

In 2003 zijn gegevens samengesteld voor de Volkstelling 2001 (VT 2001). Voor Nederland is geen echte telling onder de bevolking gehouden, maar is uitsluitend van, bij het CBS, al beschikbare bronnen gebruik gemaakt. Hierdoor is de Nederlandse belastingbetaler aanzienlijk goedkoper uit geweest en zijn toch gegevens over Nederland beschikbaar gekomen die worden vergeleken met die van andere landen.

De Volkstelling heeft als peilmoment 1 januari 2001. De tabellen zijn voornamelijk samengesteld uit het SSB. Dit bestaat uit een reeks koppelbare bestanden die

Tabel 5. Bevolking naar opleidingsniveau

	Totaal	0-14	15-74	75+
Totale bevolking	15.985.538	2.977.283	12.036.171	972.084
Geen opleiding	1.244.031	1.244.031	0	0
Pre-primair onderwijs (ISCED 0)	1.370.511	1.198.580	154.832	17.098
Primair onderwijs (ISCED 1)	2.787.104	534.672	1.825.655	426.778
Lager secundair onderwijs (ISCED 2)	3.145.529		2.924.405	221.125
Hoger secundair onderwijs (ISCED3c)	2.711.384		2.566.372	145.012
Hoger secundair onderwijs (ISCED3b)				
Hoger secundair onderwijs (ISCED3a)	1.873.656		1.828.072	45.584
Post secundair onderwijs (ISCED4)	483.684		468.699	14.985
Tertiair onderwijs (ISCED5b)	247.194		238.029	9.165
Tertiair onderwijs (ISCED5a)	2.081.590		1.992.670	88.920
Onderwijs aan promovendi (ISCED6)	32.760		31.082	1.678
Onbekend	8.094	0	6.356	1.738

onderling consistent zijn gemaakt. Er worden een aantal onderdelen van het SSB gebruikt. GBA-gegevens van 1 januari 2001 zijn gebruikt als basis voor de tabellenset. Het banenbestand 2000 met geïntegreerde gegevens over onder andere werknemers en zelfstandigen van ultimo 2000. Voor verschillende persoonskenmerken, zoals het hoogst behaalde opleidingsniveau, is uit de EBB van 2000 en 2001 geput.

Sommige variabelen zijn integraal beschikbaar zoals leeftijd, geslacht, burgerlijke staat en het al dan niet werknemer zijn. Andere variabelen, voornamelijk uit de EBB en het Woningbehoeftenonderzoek (WBO), zijn slechts voor een deel van de personen beschikbaar. Voorbeelden hiervan zijn hoogst behaalde opleidingsniveau (EBB) en of iemand in een huur- of koopwoning woont (WBO). Teneinde consistentie tussen de tabellen te garanderen is gebruik gemaakt van de techniek van het herhaald wegen. Elke tabel moet berekend worden uit het grootste blok waaruit de tabel bepaald kan worden. Als alle tabellen met de juiste gewichten worden geschat, zijn de tabeluitkomsten onderling consistent. Onder bepaalde voorwaarden kunnen ook de varianties van de tabelwaarden worden geschat.

Noten

- (1) Eric Schulte Nordholt is projectleider van de Volkstelling 2001 op het CBS. esle@cbs.nl

Referenties

- Arts, C.H. en E.M.J. Hoogteijling, 2002, Het Sociaal Statistisch Bestand 1998 en 1999. In: *Sociaal-economische maandstatistiek*, Jaargang 19, december 2002, blz. 13-21.
- Bakker, B.F.M., en C.H. Arts, 2003, Dynamiek op de Arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand, In: deze bundel.
- CBS, 2002, Algemene tellingen in de twintigste eeuw: de methode van onderzoek bij personen en bedrijven. J.G.S.J. van Maarseveen (red.). Centraal Bureau voor de Statistiek, Voorburg / Heerlen.
- Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen en V. Snijders, 2003, Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Centraal Bureau voor de Statistiek, Voorburg / Heerlen. <http://www.cbs.nl/en/publications/articles/general/discussion-papers/discussion-paper-03005.pdf>

De ‘virtuele Volkstelling’ is geen kopie van een traditionele Volkstelling

Een reactie op het paper van E. Schulte Nordholt

C. de Hoog (1)

1. Inleiding

In 1971 werd in Nederland de laatste Volkstelling gehouden. Een telling die in dat jaar op drie pijlers berustte. Gewoontegetrouw kon een gedeelte van de verzamelde gegevens worden gebruikt om de gemeentelijke bevolkingsregisters te controleren. Deze controle had niets te maken met het opsporen van illegalen of met andere activiteiten die betrokkenen niet zichtbaar wilden maken. Het ging om, tamelijk globale, administratieve correcties, zoals bijvoorbeeld een verbetering van het aantal leden in een huishouden.

De tweede pijler waarop de Volkstelling 1971 was gebaseerd, bestond, traditioneel, uit het verkrijgen van data die een overzicht gaven van basisstructuren. Basisstructuren die demografisch, economisch en sociaal-cultureel van aard waren. Een bijkomend voordeel van deze datacollectie was dat de transversale gegevens enigszins een longitudinaal karakter kregen, omdat op basis van eerdere Volkstellingen ontwikkelingen in de tijd getraceerd konden worden. Typisch voor de Nederlandse situatie was dat er aan de drie laatste Volkstellingen (1947, 1960, 1971) een annex was verbonden. Deze

Volkstellingen waren ook woningtellingen. Dit behoeft geen verwondering te wekken omdat de woningnood door politici en beleidsvoerders, zeker in de eerste twee decennia na de Tweede Wereldoorlog, als volksvijand nummer één werd beschouwd.

De Volkstelling 1971 was in wetenschappelijk opzicht een eigenaardige telling, omdat voor de eerste en voor de laatste keer in de lange geschiedenis van de Volkstellingen de wetenschap, via SISWO en universitaire onderzoekinstellingen, direct betrokken was bij analyse van de data (censusmonografieën) en in iets mindere mate bij de constructie en de opzet van de vragenlijsten (Reinink, 1968). Deze derde pijler van de Volkstelling 1971 bestond derhalve uit de productie van de wetenschappelijke censusmonografieën.

SISWO had de projectleiding van de in totaal 18 monografieën. De Ministeries van Cultuur, Recreatie en Maatschappelijk Werk, van Sociale Zaken en van Volkshuisvesting en Ruimtelijke Ordening verstrekten in eerste instantie een subsidie voor de gezinssociologische censusmonografieën (twaalf). De auteur van deze bijdrage was toentertijd lid van het gezinssociologische coördinatieteam (2). In 1971 werd financiering van de gehele reeks verkregen (ruim vier miljoen gulden) uit de “beleidsruimte”, een overheidsfonds voor onderzoek onder verantwoordelijkheid van het Interdepartementaal Overleg Wetenschapsbeleid (IOW). Andere financiers waren het CBS en de deelnemende universitaire onderzoeksinstituten die faciliteiten en vooral menskracht beschikbaar stelden (Thoenes e.a., 1979).

Met de Volkstelling 1971 als voorbeeld, leek de census een zonnige toekomst tegemoet te gaan. Het beleid en de wetenschap hadden elkaar gevonden. De analyses van de komende Volkstellingen zouden moderne analyses worden. Niets bleek echter minder waar te zijn. Elders in deze bundel wordt nader ingegaan op de problemen waarmee de Volkstelling 1971 te kampen heeft gehad. Naast de problemen over de vermeende schending van de privacy en de onbereikbaarheid van bepaalde groeperingen (studenten, tweeverdieners die overdag niet thuis waren et cetera), dient ook vermeld te worden dat er moeilijkheden waren met de technische uitvoering. Een gedeelte van de ponskaarten was met drukinkt vervuild. Hierdoor werden foutieve gegevens afgelezen. Bovendien toonde de (nieuwe) computerapparatuur gebreken. Het gevolg was dat de eerste censusmonografie pas in 1978 is verschenen en de laatste in 1980. Wel moet worden gezegd dat deze tegenslagen die wetenschappelijke bedrijvigheid niet hebben beïnvloed. Door SISWO en de onderzoeksinstellingen werden interim-rapporten uitgebracht. Alle achttien definitieve censusmonografieën zijn in druk verschenen.

2. Het einde van de traditionele Volkstelling in Nederland

Het is juist dat het in de 21ste eeuw in Nederland niet meer mogelijk is om een traditionele Volkstelling te houden. Het sterkste argument tegen een dergelijke telling vormt niet de kosten. De berekeningen zijn gebaseerd op ruwe schattingen die variëren van 50 miljoen euro tot een paar honderd miljoen euro. Er heeft geen grondige

kosten-baten analyse plaatsgevonden. Bovendien is bij de berekeningen geen rekening gehouden met de kosten van onderzoeken die worden gehouden, omdat gegevens via een recente Volkstelling niet meer te verkrijgen zijn. Er is ook niet ingegaan op de impuls die de sociale wetenschappen (wetenschappelijke innovatie) via een moderne census zouden kunnen krijgen.

Een sterker argument om geen Volkstelling meer te houden is de te verwachten non-response (item en unit). In zijn bijdrage beschrijft Schulte Nordholt terecht dat de leemten die hierdoor zullen ontstaan integrale informatie over de bevolking vrijwel onmogelijk maakt. We hebben hier niet alleen te maken met onderzoekstechnisch probleem, maar ook met een administratief probleem. De bevolkingsgegevens in de grote steden, in het bijzonder in sommige wijken en buurten, zijn in belangrijke mate gebaseerd op een virtuele werkelijkheid. Een werkelijkheid die tijdens de Bijlmerramp pijnlijk duidelijk is geworden. Een werkelijkheid waarmee woningbouwverenigingen en woningstichtingen door het verschijnsel onderhuur dagelijks mee te maken hebben. Door de leemten in de gemeentelijke administratie kan zelfs de vraag worden gesteld in hoeverre de bevolkingsboekhouding betrouwbaar is.

Toch is het de vraag of 'een virtuele Volkstelling' een traditionele telling kan vervangen. Aan internationale verplichtingen kan weliswaar voldaan worden, maar een virtuele telling lijkt niet geschikt te zijn om, in navolging van de vroegere censusmonografieën min of meer gecompliceerde kruisverbanden op micro- en macroniveau te construeren en te analyseren. Dergelijke verbanden

kunnen zowel voor het beleid als voor de wetenschap van groot belang zijn. Een voorbeeld, via een Volkstelling is het in principe mogelijk om een tabel samen te stellen waarin gegevens zijn opgenomen over de samenstelling van het huishouden, het opleidingsniveau van de leden van het huishouden en het huishoudinkomen. Een tabel die een overzicht geeft van de sociale strata binnen de samenleving. Een virtuele telling lijkt deze mogelijkheid niet of nauwelijks te kunnen bieden.

Er is nog een ander probleem aanwezig bij een virtuele telling. De gegevens van deze tellingen zijn afkomstig uit verschillende bronnen en statistieken. Dit kan betekenen dat we te maken krijgen met verschillende definities. Ook daarvan een voorbeeld: in een traditionele Volkstelling wordt een nooit gehuwde die bij zijn ouder(s) woont in statistische zin als een kind beschouwd (positie in het huishouden). Hierdoor was het mogelijk om op betrekkelijk eenvoudige wijze het verlaten van het ouderlijk huis te analyseren. Tegenwoordig zien we steeds meer statistieken waarin niet meer wordt uitgegaan van één positie binnen het huishouden, maar bijvoorbeeld van de combinatie kinderen en volwassenen. Dit kan betekenen dat een huishouden met vier volwassenen kan bestaan uit een gezin met oudere kinderen, maar ook dat het huishouden eventueel een woongroep kan zijn of een ander samengesteld huishouden dat uit twee (echt)paren kan bestaan. Kortom een koppeling van verschillende bestanden kan tot verwarring leiden.

3. Een andere route?

We hebben aandacht gevraagd voor de leemten die aanwezig kunnen zijn in de gemeentelijke administratie. Een 'virtuele Volkstelling' lost deze moeilijkheden niet op. Het is daarom van belang dat de rijksoverheid met enige nadruk de gemeenten opdracht geeft om hun bevolkingsregisters, wellicht via steekproeven te controleren. De jaren vijftig, toen de afwijkingen gering waren, zullen hierdoor niet terugkeren. Wel worden de leemten en de problemen zichtbaar en wordt duidelijkheid verkregen welke groeperingen in de bevolkingsbestanden ondervertegenwoordigd zijn.

De Volkstelling 1971 was vooral uniek, omdat er door de censusmonografieën een samenwerking aanwezig was tussen het CBS, ministeries, SISWO en universitaire onderzoeksinstellingen. Juist hierdoor had de laatste telling een meerwaarde. Daarom verdient het aanbeveling om te pogen om een dergelijk samenwerkingsverband in ere te herstellen.

Een andere route dan de 'virtuele telling' zou een minicensus (steekproef van 30.000 huishoudens) kunnen zijn. Deze minicensus zou elke vijf jaar herhaald moeten worden. Nog beter zou het zijn als deze minicensus het karakter krijgt van een longitudinaal onderzoek (dataverzameling om de twee jaar). Bij het entameren van een minicensus zouden SISWO en het CBS, in samenwerking met andere organisaties (bijvoorbeeld KNAW, NWO, SCP, NIDI en de Nederlandse Gezinsraad) en uiteraard de sociaal-wetenschappelijke faculteiten en departementen van de universiteiten, het voortouw kunnen nemen.

De virtuele Volkstelling is geen slecht alternatief om aan internationale statistische verplichtingen te voldoen. Een minicensus is echter een beter alternatief voor een Volkstelling, in het bijzonder als het gaat over beleidsanalyses en wetenschappelijke analyses op nationaal niveau.

Noten

- (1) Cees de Hoog is hoogleraar gezinssociologie en gezinsbeleid aan de Wageningen Universiteit, leerstoelgroep Sociologie van Consumenten en Huishoudens Betrokkenheid. Lid van het coördinatieteam gezinssociologische censusmonografieën 1970-1979, gebaseerd op de Volkstelling 1971. SISWO/Landbouwniversiteit
- (2) Het gezinssociologische coördinatieteam bestond uit C.J.M. Corver (SISWO), A.M. van der Heiden (Universiteit Utrecht), C de Hoog (v/h Landbouwhogeschool Wageningen) en L. Th. van Leeuwen (v/h Landbouwhogeschool Wageningen).

Referenties

Reinink, G.J., 1968, De Volkstelling 1970, Aanpassings-desiderata, SISWO, Amsterdam.

Thoenes, P., J.G.M. Sterk en W. Begeer, 1979, Ten geleide, In: C.J.M. Corver, A.M. van der Heiden, C. de Hoog en L.Th. van Leeuwen, Huishouden, huwelijk en gezin, SISWO / CBS, Amsterdam / Voorburg.

Het Sociaal Statistisch Bestand (SSB): de statistiek van de toekomst

P. Everaers (1)

1. Inleiding

In het Sociaal Statistisch Bestand (SSB) van het CBS worden gegevens uit registers en gegevens uit surveys op microniveau, dat wil zeggen op het niveau van individuele personen, gecombineerd tot een geïntegreerd databestand. Dit geïntegreerde bestand vormt in de nabije toekomst de basis voor de statistieken over personen en huishoudens van het CBS. De afkorting SSB wordt niet alleen gebruikt voor het aanduiden van dit stelsel van gekoppelde gegevens maar ook voor de werkwijze voor het verzorgen en beschikbaar stellen van gegevens over sociale en maatschappelijke gebeurtenissen uit dit stelsel. In deze bijdrage worden de achtergronden van het opzetten van het SSB geschetst en wordt aangegeven hoe het SSB zich onderscheidt van de traditionele wijze waarop het CBS informatie over personen en huishoudens samenstelde.

In paragraaf 2 worden de achtergronden geschetst van het SSB als een actueel, flexibel en geïntegreerd systeem voor de sociale statistiek. In de derde paragraaf worden de soorten gegevensbronnen voor de sociale statistieken beschreven en wordt kort getoond hoe deze bronnen in het SSB worden gebruikt. Daarbij wordt aangegeven hoe de werkwijze van het SSB verschilt van de traditionele wijze van statistiek

maken. In de vierde paragraaf wordt betoogd dat de moderne statistiekbeoefening een combinatie vereist van inhoudelijke kennis op specifieke sociaal wetenschappelijke terreinen en hoogwaardige IT kennis. Het SSB als een exponent van deze moderne manier van statistiek maken is in hoge mate op een moderne IT infrastructuur gebaseerd.

2. De achtergronden achter het SSB als werkwijze en geïntegreerd databestand

2.1 De eisen te stellen aan moderne sociale statistieken

Aan de sociale statistieken worden een aantal eisen gesteld: actualiteit, relevantie en flexibiliteit. Deze eisen zijn niet uitsluitend van toepassing voor het SSB, maar gelden ook voor de meer traditionele wijze van statistiek maken. De ontwikkeling van het SSB maakt het echter mogelijk in de nabije toekomst beter aan deze eisen te voldoen.

Statistische informatie moet in de eerste plaats actueel zijn, gaan over zaken die op dit moment in de belangstelling staan en daarmee aansluiten op de vigerende vraag naar informatie. Dit betekent dat informatie die wordt verzameld niet pas na drie of vier jaar beschikbaar moet zijn. De tijd tussen verzamelen van de gegevens en beschikbaar komen van de hierop gebaseerde statistische informatie moet zo kort mogelijk zijn. Het CBS hanteert voor het criterium van tijdigheid voor de meeste statistieken de 1 op 1 norm: bijvoorbeeld een maandstatistiek wordt binnen een maand na de maand van verzamelen van de gegevens gepubliceerd. In Europees verband worden bij tal van indicatoren

vergelijkbare en soms zelfs nog strengere eisen gesteld. Statistische informatie moet ook relevant zijn. Ze moet informatie geven over vraagstukken die voor beleid en maatschappij belangrijk zijn. De relevantie kan worden afgemeten aan de mate waarin de statistieken informatie geven over categorieën van de bevolking, beleidsmaatregelen of regio's die sterk in de belangstelling staan. Sterk beleidsgericht is de informatie die inzicht geeft in de kwaliteit van het gevoerde beleid.

Een volgende eis aan moderne sociale statistieken is flexibiliteit. Snel en eenvoudig moeten aanvullende gegevens kunnen worden geproduceerd. Vanzelfsprekend is dat sociale statistieken ook betrouwbaar moeten zijn, dat wil zeggen dat deze bij herhaalde meting tot dezelfde conclusie leidt. Hiervoor dienen de meetmethoden en de basisgegevens aan een aantal belangrijke voorwaarden te voldoen.

2.2 De redenen voor de ontwikkeling van het SSB

De ontwikkeling van het SSB is ingegeven door de vanaf het begin van de jaren negentig van de vorige eeuw sterk toegenomen vraag naar informatie van verschillende domeinen van de leefsituatie in relatie tot elkaar: arbeid en gezondheid, onderwijs en beroepsloopbaan, inkomen en transitie op de arbeidsmarkt, etc. Deze samenhangen worden onderzocht om in meer detail ontwikkelingen in de maatschappij te kunnen beschrijven en te verklaren. Voor een dergelijke beschrijving van samenhangen zijn kenmerken van personen en huishoudens nodig afkomstig uit verschillende domeinen die in samenhang met elkaar beschikbaar zijn, dus tegelijkertijd in hetzelfde databestand opgenomen. Dit vraagt om een statistische

output die meer multivariaat is en dus meer domeinen van de leefsituatie omvat, maar ook om informatie die sneller en eenvoudiger met elkaar kan worden gecombineerd.

Een tweede reden voor de ontwikkeling van het SSB zijn de toegenomen mogelijkheden van de Informatie Technologie (ICT). In het begin van de jaren negentig was het mogelijk een bestand van 100.000 cases op een pc te behandelen. In het midden van de jaren negentig zijn de mogelijkheden zo toegenomen dat meerdere bestanden met miljoenen cases tegelijk kunnen worden geanalyseerd. Heden ten dage is het koppelen van bestanden met miljoenen cases een zaak van minuten. Daarnaast zijn ook de mogelijkheden van de software voor de analyse van grote databestanden sterk uitgebreid. Een statisticus is daarmee steeds meer een medewerker die van achter zijn computerscherm (eventueel zelfs op grote afstand van de fysieke locatie waar de gegevens zijn opgeslagen) zeer grote databestanden manipuleert en analyseert.

Een derde reden voor de ontwikkeling van het SSB, met name voor Nederland gezien de nadruk op onderzoek naar mogelijkheden van registraties voor de statistiek, maar zeker niet uniek voor Nederland, is de hoge non-respons bij persoonsenquêtes. Onder non-respons wordt hier verstaan het ontbreken van de bereidheid bij groepen van de bevolking om medewerking te verlenen aan surveys en de daardoor slechts gedeeltelijke representativiteit voor de gehele populatie van steekproefonderzoek. Slechts met grote inspanningen kunnen bepaalde categorieën van de bevolking worden overgehaald via een telefonisch of een mondeling interview informatie te verstrekken. Dit

is een van de redenen waarom de gegevensverzameling via enquêtering zeer duur is. De hoge non-respons bracht het CBS al in het begin van de jaren negentig er toe te onderzoeken welke mogelijkheden registraties de statistiek kunnen bieden. Hierdoor en door de informatisering van de maatschappij als geheel, zijn veel nieuwe mogelijkheden voor het maken van statistieken ontstaan.

De vierde kracht achter de ontwikkeling van het SSB is de noodzaak de kosten van statistiekproductie te verlagen. Dit kan in de eerste plaats door meer gebruik te maken van bestaande registraties zoals boven beschreven. Daarnaast kan door een verdergaande mate van standaardisatie voorkomen worden dat veel tijd wordt besteed aan het afstemmen van uitkomsten of basisgegevens van afzonderlijke bronnen. Door vergelijkbare begrippen en classificaties te gebruiken wordt efficiency bereikt in de samenstelling van gegevens. Binnen het CBS en in internationaal verband samen met andere statistische bureaus wordt standaardisatie nagestreefd. Van groot belang is ook de harmonisatie van begrippen en classificaties met instanties die de registraties beheren die door het CBS voor het maken van statistieken in het kader van het SSB worden gebruikt. Deze harmonisatie wordt bereikt door met deze instanties permanent te overleggen welke begrippen en classificaties het beste voor enerzijds de registratie zelf en anderzijds voor de statistiek kunnen worden gebruikt.

Ten slotte is de ontwikkeling van het SSB als flexibel systeem ingegeven door de gedachte dat met het beschikbaar zijn van een SSB de druk op rapportages,

maatwerk en specifieke vragenlijsten kan worden verminderd. Min of meer op aanvraag zijn alle mogelijke combinaties van informatie over de verschillende domeinen van de sociale statistieken beschikbaar. De informatie die gevraagd wordt is namelijk al in het SSB opgenomen en er kan snel en zonder een grote extra inspanning specifieke statistische informatie uit worden samengesteld.

2.3 Het levensloopmodel als conceptueel model achter het SSB

Voor een geïntegreerd systeem van sociale statistieken is het nodig dat een invalshoek voor de beschrijving van de sociale werkelijkheid wordt gekozen. Voor de bedrijfsstatistieken fungeert het concept van de nationale rekeningen al vanaf de jaren zestig als een ordeningskader. Definities en classificaties zijn in het systeem van nationale rekeningen onderling op elkaar afgestemd en bieden meta-informatie die bijvoorbeeld het invoegen van nieuwe statistieken of het belichten van een specifiek aspect in relatie met andere aspecten aanzienlijk vergemakkelijkt.

In de sociale statistieken is er van een dergelijk algemeen geaccepteerd conceptueel model echter geen sprake. Wel zijn er in de loop van de afgelopen decennia diverse pogingen gedaan om een dergelijk model te ontwikkelen aan de hand waarvan de sociale werkelijkheid in de statistiek kan worden gemodelleerd. Zo is er vanuit het conceptueel model van de nationale rekeningen gezien of een systeem van sociale statistieken gebaseerd op toegevoegde waarde of kosten van voorzieningen

voor huishoudens en personen kon worden opgezet. Door onderzoekers op het gebied van tijd, ruimte en geografie is gezocht naar de mogelijkheden om de sociale statistieken te ordenen met als bindend element de tijd dat handelingen plaatsvinden. Omdat onderdelen van de sociale werkelijkheid niet of moeilijk in de eenheden tijd of geld zijn uit te drukken zijn deze modellen echter minder geschikt voor de ordening van *alle* fenomenen die zich in de maatschappij of in het leven van individuen voordoen. Voor het opzetten van een SSB waarin juist alle aspecten van de sociale statistieken een plaats moeten krijgen ten opzichte van de andere aspecten is een model nodig dat in principe juist alle aspecten *kan* omvatten.

2.4 Het 'levensloopmodel'

Door onderzoekers op het gebied van armoede en sociale deprivatie wordt voor het beschrijven van de ernst van de sociale uitsluiting vaak gebruik gemaakt van een door Bergman (1997) beschreven conceptueel model. Dit model maakt gebruik van de relaties van (variabelen in) de verschillende domeinen van de sociale statistieken vanuit een hiërarchisch perspectief waarbij de sterk "objectief" waarneembare en genetisch of door afkomst bepaalde kenmerken vooraf gaan aan meer subjectieve en op keuzes van individuen gebaseerde kenmerken. Er wordt gebruik gemaakt van de meest waarschijnlijke tijdsvolgorde en causaliteit van de variabelen in en tussen de verschillende domeinen van de maatschappelijke werkelijkheid. Dit conceptueel model toont sterke gelijkenis met keuzen die in de loop van een leven worden gemaakt. Bij de geboorte krijgt men bepaalde kenmerken mee, nationaliteit, andere gezondheidskenmerken; de gezondheid bepaalt in

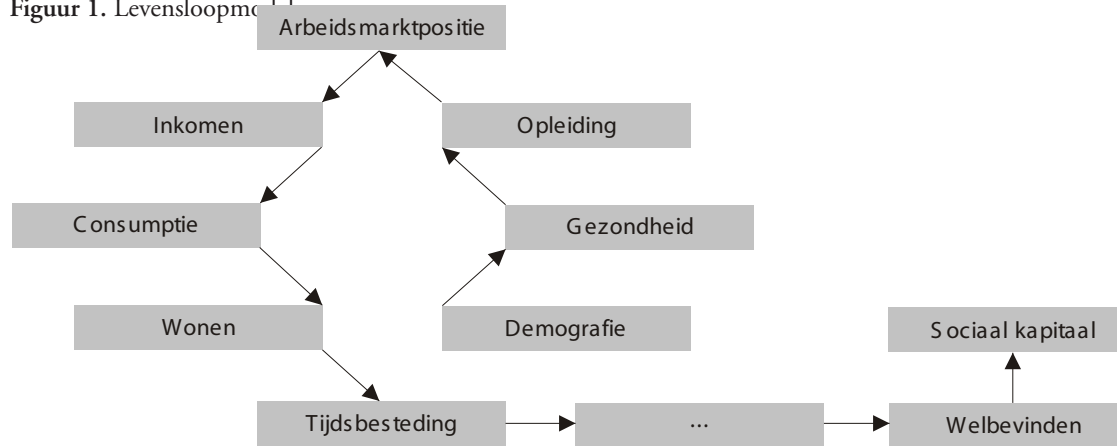
belangrijke mate of een bepaalde opleiding kan worden gevolgd die weer aanleiding kan zijn voor het beoefenen van een bepaald beroep. Hiermee wordt inkomen verworven dat kan worden geconsumeerd, bijvoorbeeld via de woonsituatie maar ook door andere aspecten van de tijdsbesteding.

Het model biedt de mogelijkheid de verschillende onderdelen van de sociale statistieken in hun onderlinge relatie te beschrijven en vertoont grote overeenkomsten met zogenaamde levensloopmodellen. Levensloop is als uitgangspunt voor de benadering van de maatschappij en ontwikkelingen een veel gebruikte methode voor sociaal wetenschappelijk onderzoek. Kern van levenslooponderzoek is de studie van onderling samenhangende individuele processen die zich over het leven ontvouwen gedurende een bepaalde periode. De meest gebruikte omschrijvingen van levensloop (bijvoorbeeld Runyan 1984 en Mayer en Tuma 1990) verwijzen respectievelijk naar de samenhang tussen eerdere en latere gebeurtenissen in een levensloop en naar de zich in de tijd wijzigende samenhang van levensdomeinen en naar de

In de verschillende domeinen bevinden zich de bijbehorende relevante variabelen. Bijvoorbeeld in het domein demografie de positie van een persoon in de demografische context. In het domein arbeidsmarktpositie variabelen die relevant zijn voor de beschrijving van de positie van personen en huishoudens op de arbeidsmarkt maar ook in relatie tot de variabelen in de andere domeinen.

individuele en historische tijdsdimensie.

Figuur 1. Levensloopperspectief

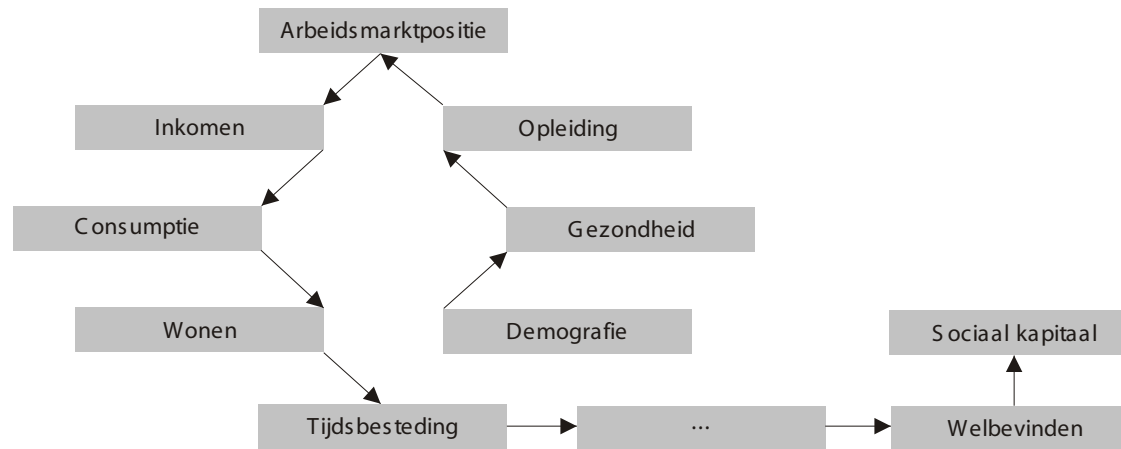


Het geschetste conceptueel model biedt goede aanknopingspunten voor het beschrijven van de relevante statistieken. De meeste beleidsmaatregelen en beleidsvoornemens waarvoor statistische informatie in onze *evidence based* maatschappij onmisbaar is, kunnen in dit model worden geplaatst. In verschillende beleidsonderzoeken wordt het accent gelegd op de samenhang van eerdere ervaringen en keuzes met de latere levensloop en de posities die een persoon of huishouden dan op bepaalde variabelen inneemt. Dit is bijvoorbeeld het geval op het gebied van armoede en sociale uitsluiting. Bergman stelt bij de beschrijving van multiple deprivatie en de daarmee samenhangende sociale uitsluiting dat in de moderne westerse maatschappij een achterstand op

een van de variabelen in een van de domeinen in ernst of gevolg vermindert door middel subsidiesystemen en ondersteuning- en hulpprogramma's van de overheid. Een achterstand in twee opeenvolgende domeinen, door hen multiple deprivatie genoemd, leidt tot sociale uitsluiting.

Het model zoals in figuur 1 weergegeven biedt een integratiekader voor de sociale en ruimtelijke statistieken die het CBS maakt. Uit dit model blijkt immers de relevantie van het levensloopperspectief voor de statistieken op het terrein van demografie, gezondheid, onderwijs, arbeidsmarkt, leefsituatie, inkomen, sociale zekerheid etc.

Figuur 2. Levensloopmodel in de historische dimensie en de effecten van overheidsmaatregelen



Voor het goede gebruik kunnen maken van de kracht van het model is informatie per domein en per variabele nodig voor elk individu op verschillende momenten in de tijd (longitudinaal). Indien gebruik wordt gemaakt van panelinformatie – bij dezelfde personen en huishoudens op verschillende momenten in de tijd verzamelde gegevens over dezelfde kenmerken – wordt optimaal aan deze behoefte voldaan. Beschikt men over dergelijke informatie op meerdere tijdstipmomenten dan biedt deze informatie ruime analysemogelijkheden.

De analysemogelijkheden die ook aansluiten bij de hedendaagse vraag naar statistische informatie kunnen als volgt worden getypeerd:

- *informatie over toestanden* (hoeveel mensen zijn er werk-

loos, bevinden zich in een bepaalde gezondheids-situatie of woonomstandigheden),

- *informatie over trends* (wat zijn de veranderingen in de omvang van de werkloosheid)
- *informatie over transities tussen toestanden* (welke mensen zijn nu werkzaam en waren een jaar geleden werkloos) en
- *de duur in bepaalde toestanden* (hoe lang is iemand werkloos geweest alvorens weer werkzaam te zijn?).

De combinatie van de informatie en de analyse mogelijkheden door deze vier aspecten van meerdere (variabelen van) domeinen in hun onderlinge samenhang te beschrijven biedt uitstekende mogelijkheden voor het monitoren en evalueren van beleid en beschrijven van effecten van beleidsmaatregelen.

Statistische informatie kan op basis van de analyse-mogelijkheden worden verdeeld in:

1. domeinspecifieke informatie over toestanden en trends
2. transitie en duren binnen een domein
3. relaties tussen domeinen
4. relaties tussen transitie en duren tussen domeinen en
5. monitorinformatie

Onder monitorinformatie wordt die informatie verstaan waarbij effecten van maatregelen over een veel langere periode een voortdurende bron van onderzoek zijn. Dit in tegenstelling tot de vier eerdere analyse-mogelijkheden waarbij de analyses zich tot een periode van maximaal 10 jaar beperken.

3. Bronnen voor de sociale statistieken

Het aantal beschikbare domeinen van de leefsituatie (zie figuur 1) en daarmee ook het aantal gewenste kenmerken van personen en huishoudens is groot. Gezien de verschillen in aard (subjectief, objectief, eenvoudig meetbaar, complexere concepten) zijn ze niet allemaal via dezelfde waarnemingsmethode te verzamelen. Sommige gegevens laten zich goed via mondelinge enquêtering verzamelen, andere bijvoorbeeld alleen via observatie of fysiek onderzoek. Weer andere gegevens zijn in Nederland voor de hele populatie beschikbaar via registraties (bevolkingsadministratie, inkomensgegevens via de belastinginformatie) andere zijn alleen voor specifieke populaties of voor bepaalde doelen beschikbaar (ziekenhuisregistratie). Uit het voorgaande is duidelijk dat het van belang is dat kenmerken van verschillende

domeinen bij individuen en huishoudens in hun onderlinge samenhang verzameld worden en dat deze gegevens ook longitudinaal beschikbaar zijn.

Gebaseerd op het aantal domeinen dat een bron dekt en of de gegevens in de tijd gezien op meerdere momenten op dezelfde manier worden verzameld kunnen de bronnen langs de volgende lijnen worden onderscheiden.

- Eendimensionale survey: voorbeeld is het Consumenten Conjunctuuronderzoek (CCO). Dit onderzoek richt zich op een aspect (consumentenvertrouwen en consumentengedrag). De enquêtering vindt plaats bij een telkens wisselende steekproef van de Nederlandse bevolking.
- Multidimensionale survey (eenmalig): een voorbeeld van deze gegevensbron is het Permanent Onderzoek Leefsituatie (POLS) van het CBS. Gegevens van verschillende domeinen, demografie, gezondheid, recht en veiligheid en wonen worden in een survey aan een (telkens wisselende) steekproef van de Nederlandse bevolking gevraagd.
- Multidimensionale survey: repeterend. Het Sociaal Economische Panelonderzoek (SEP) is hier een voorbeeld van. Gegevens van verschillende domeinen, demografie, gezondheid, recht en veiligheid en wonen worden in een survey aan steeds dezelfde steekproef van de Nederlandse bevolking gevraagd.
- Eendimensionaal register: een voorbeeld hiervan is de GBA-bevolkingsadministratie. Hierin worden van elke persoon in Nederland woonachtig uitsluitend demografische gegevens opgeslagen. Dit register wordt door de gemeenten onderhouden.
- Meerdimensionaal register: een voorbeeld hiervan is de inkomensstatistiek gebaseerd op de gegevens van

de belastingdienst gecombineerd met gegevens van de huursubsidie en studiefinanciering.

Het principe van het SSB is dat de bovenbeschreven longitudinaal beschikbare informatie en de informatie die maar op een moment beschikbaar is uit de verschillende bronnen wordt gecombineerd tot een bestand. Dit is eenvoudig weer te geven als een matrix met in de kolommen alle inwoners van Nederland en in de rijen alle gewenste variabelen. De stap van afzonderlijke bronnen naar een geïntegreerd stelsel wordt gevormd door de bronnen te koppelen op basis van individuele identificatoren (in Nederland het A-nummer en het sofi nummer en in mindere mate geboortedatum, geslacht en adres). De methode om de gegevens van de individuen

Figuur 3. De combinatie van register en survey informatie in het SSB (gesimplificeerd)



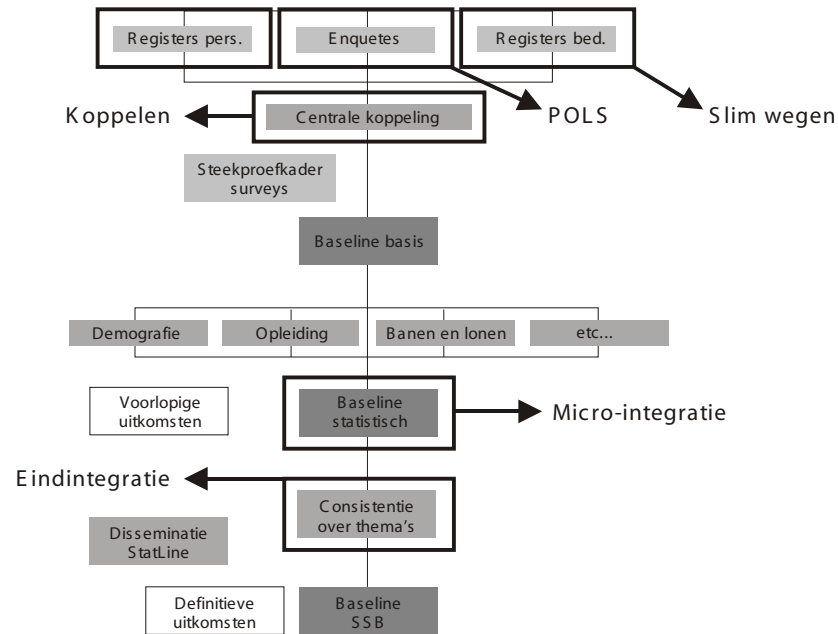
uit verschillende bronnen te koppelen is elders uitvoerig beschreven (Arts, Bakker en Van Lith, 2000). In figuur 3 is schematisch de combinatie van register en survey informatie in dit geïntegreerde bestand weergegeven. In Bakker en Arts (2003) is een overzicht opgenomen van de bronbestanden die op dit moment in het SSB zijn opgenomen.

4. Het SSB is in hoge mate gebaseerd op een moderne IT infrastructuur

De moderne statistiekbeoefening is een combinatie van inhoudelijke en hoogwaardige IT kennis. Het SSB als een exponent van deze moderne manier van statistiek maken is in hoge mate op een moderne IT infrastructuur gebaseerd. In figuur 4 staat het systeem van het SSB in grote lijnen beschreven. In diverse andere publicaties wordt op dit model in meer detail ingegaan (bijvoorbeeld Arts en Hoogteijling, 2002, Bakker, 2002, Van der Laan, 2000, Bakker en Arts, 2003), een korte beschrijving volstaat dan ook hier.

Gegevens uit verschillende soorten bronnen zoals huishoudenenquêtes, bedrijfsregisters en op persoonsinformatie gebaseerde registers worden in een database gecombineerd. Het centraal koppelen van de verschillende bestanden is daarbij een belangrijke handeling waar de identificatoren een belangrijke rol spelen in het succes van de koppelingen. De verschillende onderwerpspecialisten gaan vervolgens aan de slag om de uit verschillende bronnen afkomstige gegevens per domein consistent te maken. Programma's zoals demografie, gezondheid, arbeid etc gebruiken daarvoor elk specifieke methoden gebaseerd op hun kennis over de verschillende bronnen

Figuur 4. Het systeem van baselines



en de maatschappelijke werkelijkheid. Hierover worden snelle cijfers gepubliceerd. De dan geschoonde informatie per domein wordt aan de database toegevoegd. Hierna worden de verschillende bronnen onderling met elkaar geconfronteerd om inconsistenties en fouten op te sporen. Dit wordt eindintegratie genoemd. Met eindintegratie wordt een bestand gecreëerd waarin in principe een variabele per fenomeen beschikbaar is, consistent over tijd en plaats. Ten slotte wordt dan het SSB als outputdatabase samengesteld op basis waarvan analyses over meerdere domeinen kunnen worden uitgevoerd.

Gezien de omvang van de bestanden, het terugkerende karakter van de bewerkingen en de noodzaak optimale kwaliteit in de verschillende stappen te behalen en te behouden is het gewenst al deze stappen via geteste en gedocumenteerde IT processen uit te voeren. De Business architectuur van het SSB omvat ongeveer 100 verschillende deelstappen waarvan de meeste geautomatiseerd worden uitgevoerd. Voor het opzetten en ontwikkelen van een dergelijk systeem is een goede samenwerking nodig tussen de inhoudelijk specialisten, onderzoeksmethodologen en de IT ontwikkelaars en automatiseringsmedewerkers.

Het SSB is dan ook behalve een geïntegreerd systeem en een werkwijze voor statistiekproductie een IT product. Moderne statistiek kan niet meer zonder hoogwaardige Informatie Technologie.

Noten

- (1) Pieter Everaers is directeur Sociale en Ruimtelijke Statistiek op het CBS en als zodanig verantwoordelijk voor zowel de Volkstelling als het Sociaal Statistisch Bestand. pevs@cbs.nl

Referenties

- Arts, K., B.F.M. Bakker & E. van Lith, 2000, Linking administrative registers and household surveys, In: Pieter Al en Bart F.M. Bakker (eds.), *Re-engineering Social Statistics by micro-integration of different sources*. Themanummer Netherlands Official Statistics, jrg. 15, nr. summer, p. 16-22
- Arts, C.H. & E.M.J. Hoogteijling, 2002, Het Sociaal Statistisch Bestand 1998 en 1999, In: *Sociaal-economische maandstatistiek*, december 2002, p. 13-21
- Bakker, B.F.M., 2002. Statistics Netherlands' Approach to Social Statistics: The Social Statistical Dataset, In: *OECD Statistics Newsletter*, vol. 2002, nr. 11, p. 4-6
- Bakker, B. & K. Arts, 2003, Dynamiek op de arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand (in deze bundel)

- Berghman, J. (1997) The resurgence of poverty and the struggle against exclusion: A new challenge for social security in Europe? In: *International Social Security Review*, p. 3-21.

- Laan, P. van der, 2000, Integrating administrative registers and household surveys, In: P. Al en B.F.M. Bakker (red.), *Re-engineering social statistics by micro-integration of different sources*. Themanummer Netherlands Official Statistics, jrg. 15, nr. summer, p. 7-15

- Mayer, K.U. en N.B. Tuma, (1990). Life course research and event history analysis: an overview, in Mayer, K.U. en N.B. Tuma (eds). *Event history analysis in Life course Research*. The University of Wisconsin Press, p. 2-30.

- Runyan, W.M. (1984). The life course as a theoretical orientation. In: W.M. Runyan, *Life histories and psychobiography. Explorations in theory and methods*, New York, Oxford University, p. 81-99.

SSB: wensdroom of nachtmerrie

Een reactie op het paper van P. Everaers

R. van der Velden (1)

Met het SSB lijkt het CBS in staat geweest de dromen van wetenschappers in vervulling te laten gaan. Maar als we niet waakzaam blijven zou het SSB zou ook wel eens als een nachtmerrie voor de gebruikers kunnen uitpakken, aldus Van der Velden. Voordat hij aangeeft waarom dat zo zou kunnen zijn, zet hij uiteen waarom het CBS dromen laat uitkomen.

Een belangrijk voordeel van het SSB is dat het in staat is dure informatie op goedkope en makkelijke wijze te leveren. Er is erg veel informatie beschikbaar, over de ontwikkeling van de lonen bijvoorbeeld, maar er zijn ook gedetailleerde gegevens per regio en informatie over specifieke categorieën. Verder is met het SSB de betrouwbaarheid van de gegevens verbeterd; er worden minder klassieke meetfouten gemaakt. Ook de selectiviteit in de gegevens is verminderd, hoewel registers hun eigen selectiviteitsprobleem hebben. Door in het SSB registergegevens aan elkaar te koppelen kan daar een virtuele Volkstelling uit samengesteld worden en hoeft die informatie niet meer via enquêtes verzameld te worden. Daardoor speelt in de virtuele Volkstelling het probleem van de non-respons geen rol meer, daar hebben alleen enquêtes last van. Verder neemt door het SSB ook de enquêtedruk af. Hoewel Van der Velden zich afvraagt of de enquêtedruk wel ooit een probleem is geweest. Uit

een klein onderzoekje dat hij heeft gehouden onder zijn collega's blijkt dat niemand van hen of van hun directe familieleden ooit benaderd is door een enquêteur van het CBS. Mensen worden wel veel lastig gevallen door marketing bureaus. Tot slot is een belangrijk voordeel van het SSB dat de kosten van productie van statistieken verlaagd kunnen worden. Met hetzelfde geld is nu meer te doen.

Over de mogelijkheden die het SSB biedt is Van der Velden eveneens zeer te spreken. Het SSB zou als steekproefkader gebruikt kunnen worden voor aanvullende surveys; er zijn nu gerichte steekproeven mogelijk. Daarnaast zijn de data in het SSB verrijkt, doordat een koppeling met andere data heeft plaatsgevonden. Tot slot is door SSB ook de ontwikkeling van multileveldata (personen, huishoudens, regio, bedrijven) mogelijk.

Maar zoals gezegd ziet Van der Velden niet alleen kansen, maar ook bedreigingen. Een van de grootste nadelen van het SSB, aldus Van der Velden, is het gebrek aan diepgang van de data; de dataset is beperkt, met name de vulling van opleidingsgegevens zal lang duren. Het zal erg moeilijk worden om aan een groot aantal gegevens voor een grote groep mensen te komen. Een ander minpunt ziet Van der Velden in de validiteit van de gegevens. De registratie van gegevens is afhankelijk van de gehanteerde definitie en die varieert in de loop van de tijd - dat geldt bijvoorbeeld voor het begrip werkloosheid. Wijzigingen in de definities zorgen voor trendbreuken in de gegevens, waardoor de geldigheid wordt aangetast. Bovendien worden vaak definities gehanteerd die niet uit de wetenschap afkomstig zijn. Van der Velden vindt dat bij het opzetten van een enquête eigenlijk gekeken zou moeten worden naar de

wetenschap, maar de houder van de registratie is geen wetenschapper: ambtenaren zijn geen wetenschappers. Wat Van der Velden ook jammer vindt, is dat met het SSB uiteindelijk weinig flexibiliteit geboden kan worden. Registerinformatie kan niet op verzoek aangepast worden, zodat er aanvullende informatie vrij komt. Dat kan wel bij enquêtes. Problemen zijn er verder met de meetmethode. Er zijn verschillen tussen uiteenlopende registers en wat is dan een betrouwbare meting van inkomen en / of opleiding. Verder vreest hij dat het CBS de gegevens 'oppoetst'. Koppeling van bestanden leidt er soms toe dat iemand in een zelfde periode geregistreerd staat als werkend en werkloos. Het CBS heeft een systematiek uitgedacht waarmee deze inconsistentie verholpen kan worden. Daar is Van der Velden echter niet zo blij mee. Deze geforceerde-één-cijfer-gedachte, zoals hij dat noemt, kan misschien wel nuttig zijn voor beleidsmakers, maar niet voor wetenschappers. Zij hebben geen behoefte aan één officieel cijfer. Zij willen juist de gegevens hebben op basis waarvan het CBS tot dat ene cijfer komt. Wat Van der Velden verder mist in het SSB zijn goede metadata. Wetenschappers willen niet alleen de precieze inhoud van de registers weten, maar ook die van de enquêtes. Welke vragen zijn gesteld om tot de informatie te komen en welke definities zijn er gehanteerd? Zolang goede metadata ontbreken is het beoordelen van de kwaliteit een moeizaam karwei. Tot slot vreest Van der Velden dat het SSB zich beperkt tot hetgeen wat beschikbaar is via registraties en dat de toegankelijkheid als steekproefkader voor derden nihil is. Aangezien enquêtes een snellere oplevering kennen dan registers, zou het wel eens kunnen gebeuren dat als derden geen toegang krijgen tot de gegevens er een autonome druk ontstaat om buiten het

SSB om enquêtes te houden. Die enquêtes zullen niet noodzakelijkerwijs door het CBS gehouden worden.

Wat zou er moeten gebeuren om die bedreigingen in kansen om te zetten? Van der Velden zou graag zien dat het CBS meer openheid van zaken geeft en dan niet alleen over de mogelijkheden, maar ook over de onmogelijkheden. Hij wil dat het CBS kritisch is, de interne rapporten openbaar maakt en op de website zet, zodat de huidige geslotenheid doorbroken wordt. Het geeft meer vertrouwen als het CBS ook meldt wat er niet kan of nog niet goed gaat, dan wanneer er alleen maar positieve berichten naar buiten komen. Daarnaast zou Van der Velden graag zien dat het CBS voor goede metadata zorgt. Dus waar komt de informatie vandaan, wat zijn de mogelijke problemen? Maar ook: hoe zijn de registers opgezet en hoe verhoudt de informatie zich tot elkaar? Tot slot zou hij willen dat er een optimale toegang tot de databronnen voor derden wordt gecreëerd.

Noten

- (1) Rolf van der Velden is hoofd van de afdeling Onderwijs en Beroepsloopbanen bij ROA.

Dynamiek op de arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand

B. Bakker en K. Arts (1)

1. Inleiding

Statistische informatie over stromen op de arbeidsmarkt is onmisbaar voor een goed inzicht in het functioneren van deze markt en het nagaan van de effectiviteit van werkgelegenheidsbeleid. De statistische informatiebehoefte is de afgelopen jaren veranderd door de toegenomen differentiatie gedurende het leven van mensen. Voorbeelden hiervan zijn de nog steeds in beweging zijnde rolverdeling tussen mannen en vrouwen, de flexibilisering en toegenomen dynamiek van de arbeidsmarkt, en de veranderende rol van privé- en werkomstandigheden.

De groeiende behoefte aan samenhangende longitudinale informatie kan voor een deel worden gedekt met gegevens uit het Sociaal Statistisch Bestand (SSB). De in het SSB opgenomen informatie is nog beperkt tot demografische en sociaal-economische gegevens. Op dit moment zijn voor de verslagjaren 1998-2001 Sociaal Statistische Bestanden beschikbaar. Het verslagjaar 2002 volgt volgens planning in juli 2004. Omdat de behoefte aan longitudinale informatie groeit is, door het CBS het programma Sociale Dynamiek en Arbeidsmarkt gestart. De hoofddoelstelling van het strategisch programma Sociale dynamiek en arbeidsmarkt van het CBS is het ontwikkelen van nieuwe, geïntegreerde, samenhangende

statistische persoonsinformatie over de arbeidsmarkt-, huishouden-, inkomen- en vermogensdynamiek (2). In deze projecten wordt dikwijls SSB-informatie gebruikt. In dit artikel wordt een overzicht gegeven van de mogelijkheden die het SSB biedt voor longitudinale analyses, waarbij in sommige gevallen gebruik gemaakt van de eerste resultaten uit dit programma. Eerst wordt echter de veranderde informatiebehoefte in beeld gebracht.

2. Veranderde informatiebehoefte

De behoefte aan stroominformatie komt mede voort uit het besef dat het beleid zich meer moet richten op het faciliteren van in de levensloop veranderende combinaties van privé- en werkomstandigheden. Het standaardpatroon waarin de mannen de kost verdienen en vrouwen de zorgtaken voor kinderen en ouders op zich nemen, neemt een steeds minder dominante positie in. De veranderde rolverdeling tussen mannen en vrouwen en de toegenomen arbeidsparticipatie van vrouwen leidt ertoe dat de combinatie van werk- en zorgtaken voor huishoudens problematischer wordt. Er is meer behoefte ontstaan aan mogelijkheden om tijdelijk meer of juist minder te werken om zorgtaken voor kinderen en ouders te kunnen uitvoeren. Dit is natuurlijk afhankelijk van de specifieke situatie waarin het huishouden zich bevindt. Om dit statistisch te kunnen beschrijven is stroominformatie noodzakelijk.

Verder is er meer behoefte aan stroominformatie ter evaluatie van het beleid gericht op bevordering

van instroom in en tegengaan van uitstroom uit de arbeidsmarkt. Zo is het verminderen van de instroom in de arbeidsongeschiktheid een belangrijke prioriteit van het Ministerie van Sociale Zaken en Werkgelegenheid. Daarnaast staat vergroting van de arbeidsmarktparticipatie van ouderen, vrouwen en allochtonen hoog op de politieke agenda, onder meer vanwege de gewenste emancipatie van deze categorieën en de vergrijzing van de beroepsbevolking. Om de processen van in- en uitstroom en de factoren die daarop van invloed zijn goed te beschrijven is stroominformatie onontbeerlijk.

De toegenomen werkloosheid van het afgelopen jaar heeft de belangstelling voor de aansluiting van het onderwijs en de arbeidsmarkt doen oplaaien. In het bijzonder is de toename van de jeugdwerkloosheid zorgelijk. Inzicht in de schoolloopbanen en de start op de arbeidsmarkt van jongeren is belangrijk voor de onderbouwing van het beleid op dit terrein.

Ook neemt de behoefte aan samenhangende en met elkaar in verband te brengen informatie toe. Het is onvoldoende om de stromen als zodanig te beschrijven. Ook de factoren die van invloed zijn op de omvang van de stromen zijn relevant. In de discussie over de gevolgen van de vergrijzing speelt het proces van pensionering bijvoorbeeld een grote rol. Bij het krapper worden van de arbeidsmarkt door de vergrijzing zal een beroep gedaan worden op de ouderen om langer aan het arbeidsproces deel te nemen. Inzicht in het proces van pensionering, de factoren die daarop van invloed zijn en de veranderingen in de tijd zijn van groot belang om na te gaan op welke wijze ouderen actief kunnen blijven in het arbeidsproces.

Onder de factoren die van belang zijn voor dit proces, zijn de individuele pensioenaanspraken (inclusief AOW-rechten), de vermogenspositie, de huishoudensituatie, de gezondheidstoestand, arbeidsgerelateerde ziekte, ziekteverzuim, de arbeidsomstandigheden, en de actualiteit van de kennis en vaardigheden van de betrokken werknemers.

De bovenstaande opsomming van vragen is zeker niet uitputtend. Het zijn slechts enkele voorbeelden van maatschappelijke vraagstukken die alleen met behulp van longitudinale gegevens statistisch beschreven kunnen worden. In het verleden werden deze en soortgelijke vragen beantwoord door longitudinale surveys te analyseren. Goede longitudinale surveys zijn echter schaars en worden onder druk van de bezuinigingen steeds schaarser. Bovendien hebben longitudinale surveys te kampen met een hoge en selectieve initiële non-respons en selectieve panelsterfte, waardoor de uitkomsten worden vertekend. In het SSB is een aantal registers over meerdere jaargangen gekoppeld, zodat hieruit longitudinale informatie beschikbaar komt.

3. De bronnen van het SSB

De ruggengraat van het SSB bestaat uit de *Gemeentelijke Basisadministratie* (GBA). Dit bestand bevat demografische gegevens zoals geslacht, geboortedatum, geboorteland en woonplaats van alle personen die vanaf 1995 enig moment tot de bevolking van Nederland behoorden. Alle wijzigingen hierin, zoals verhuizingen en echtscheidingen zijn in het bestand opgenomen. Alle overige bronnen van

het SSB die weergegeven zijn in tabel 1 zijn gekoppeld aan de GBA. Eerst worden zoveel mogelijk records gekoppeld op basis van het sofi-nummer. Voor records die geen sofi-nummer bevatten, wordt gekoppeld op de kenmerken geslacht, geboortedatum en adres. Deze laatste wijze van koppelen wordt in ieder geval toegepast voor de enquêtes.

Op basis van de gekoppelde bronnen is een groot aantal gegevens over personen, banen en uitkeringen integraal beschikbaar voor statistische doeleinden. De groeiende behoefte aan informatie waarin de samenhang van een groot aantal verschillende verschijnselen wordt beschreven wordt gerealiseerd doordat alle beschikbare bronnen onderling gekoppeld zijn. Doordat dat voor meerdere jaren het geval is, kunnen ook longitudinale analyses worden uitgevoerd. Bovendien hebben veel bestanden een volumekarakter, zodat ook duuranalyses tot de mogelijkheden behoren. Omdat meer gebruik wordt gemaakt van integrale registerinformatie, kan meer en betrouwbaarder informatie worden samengesteld over kleine regionale eenheden en kleine groepen in de samenleving. Voorbeelden hiervan zijn in- en uitstroom van personen die in de gezondheid- en welzijnszorg werkzaam zijn, maar ook het percentage van de bevolking van de gemeente Pekela met een werkloosheidsuitkering.

Het is echter niet zo dat alle bronnen direct bruikbaar zijn voor de statistiek. Ten eerste zijn in administratieve bronnen soms gegevens opgenomen die niet relevant zijn voor statistische doeleinden of een andere definitie kennen dan in de statistiek. Ten tweede bestrijken niet alle bronnen de gehele populatie van eenheden, die

onderwerp zijn van de statistiek. Ten derde bevatten de registers en enquêtes fouten van allerlei aard. Het is daarom noodzakelijk de verschillende bronnen te integreren.

4. De productie van het SSB: integratie van de bronnen

4.1. Micro-integratie

Het proces waarin de registers en enquêtes worden verwerkt tot statistische gegevens wordt micro-integratie genoemd (zie ook Van der Laan, 2000; Bakker, 2002; Arts en Hoogteijling, 2002). Bij micro-integratie gaat het erom dat de fouten in de bronnen worden opgespoord en worden gecorrigeerd. Als twee bronnen dezelfde variabele bevatten, dan horen die voor dezelfde persoon ook dezelfde waarde op te leveren. Het kan echter ook zijn dat er een logische relatie is tussen variabelen, bijvoorbeeld het uurloon vermenigvuldigd met het aantal gewerkte uren per jaar levert het jaarloon op. Door gebruik te maken van deze systematische relaties, worden fouten in de bronnen opgespoord. Alleen als de fouten zo omvangrijk zijn dat ze van invloed zijn op de te publiceren uitkomsten, wordt naar een manier gezocht om deze te corrigeren.

Om dit proces van micro-integratie te beschrijven worden voorbeelden ontleend aan de productie van het banenbestand. Een banenbestand van een bepaald verslagjaar is opgebouwd uit drie bronnen: de Fibase, VZA en EWL. De FiBase en de VZA bevatten baaninformatie

Tabel 1. Gegevensbronnen van het SSB naar herkomst en inhoud

<i>Bron</i>	<i>Afkomstig van</i>	<i>Belangrijke variabelen</i>
Gemeentelijke basisadministratie vanaf 1995 (GBA)	Gemeenten	geslacht, geboortedatum, adres, burgerlijke staat, nationaliteit, geboorteland, gezinsrelaties
Voorheffing loonbelasting (Fibase)	Belastingdienst	soort inkomen, fiscaal loon, aanvang en einde, loonbelastingnummer bedrijf/instelling
Verzekerdenadministratie werknemers (VZA)	Uitvoeringsinstellingen werknemersverzekeringen	Bruto loon sociale verzekering, sociale verzekeringsdagen, soort dienstverband, aanvang en einde, aansluitingsnummer bedrijf/instelling
Enquête werkgelegenheid en lonen (EWL)	CBS, enquête onder bedrijven	soort dienstverband, looncomponenten, arbeidsduur, aanvang en einde, economische activiteit
Administratie arbeidsongeschiktheidswet (AO)	Uitvoeringsinstellingen werknemersverzekeringen	soort uitkering, mate en aard van arbeidsongeschiktheid, aanvang en einde, reden beëindiging, WSW-indicatie, bedragen
Definitieve aanslaggegevens fiscale ondernemingswinst natuurlijke personen (Winst)	Belastingdienst	winstbedrag
Administratie werkloosheidswet (WW)	Uitvoeringsinstellingen werknemersverzekeringen	aanvang en einde, reden beëindiging, bedragen
Administratie studiefinanciering (WSF)	Informatie Beheer Groep	soort studiefinanciering, bedrag
Administratie algemene bijstandswet (ABW)	Gemeenten	soort uitkering, leefvorm, aanvang en einde, bedragen
Administratie wet inkomensvoorziening oudere en gedeeltelijk arbeidsongeschikte werkloze werknemers (IOAW) / arbeidsongeschikte gewezen zelfstandigen (IOAZ)	Gemeenten	soort uitkering, leefvorm, aanvang en einde, bedragen
Centraal register ingeschrevenen hoger onderwijs (CRIHO)	Informatie Beheergroep	ingeschreven bij HBO of WO, niveau en richting opleiding
Enquête Beroepsbevolking (EBB)	CBS, enquête onder personen	opleiding, beroep, werkloosheid, arbeidspositie, zoekgedrag op arbeidsmarkt

die het gehele jaar beslaat. De EWL bevat de banen aan het einde van een jaargang. Voor het bepalen van de economische activiteit van bedrijven wordt gebruik gemaakt van het Algemeen Bedrijfsregister (ABR) van het CBS. De GBA wordt gebruikt om correcties in het banenbestand aan te brengen. De bedoeling van de integratie van de bronnen is dat (banen)records uit de verschillende bronnen die betrekking hebben op dezelfde baan, aan elkaar gekoppeld worden. Tegelijkertijd wordt aan iedere baan de economische activiteit van het bedrijf toegekend. Figuur 1 laat zien dat vrijwel alle records uit de EWL aan VZA of Fibase-records gekoppeld kunnen worden. Ruim een kwart van alle banenrecords (26%) in het banenbestand vinden hun oorsprong in alle drie de

bronnen. Ruim 60% is niet in de EWL maar wel in de doorsnede van de VZA en de Fibase aanwezig.

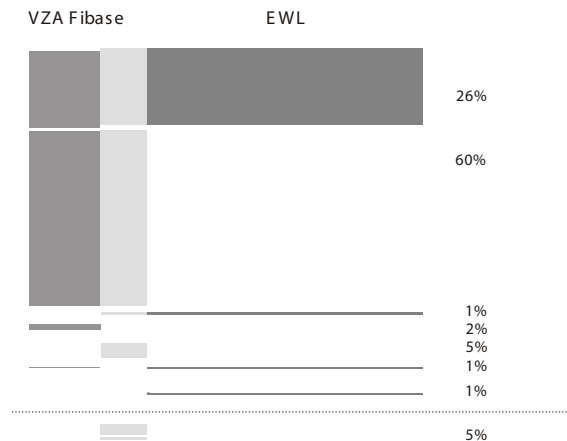
4.2. Definitieverschillen

Bij de start van de micro-integratie wordt eerst beoordeeld welke definities worden gebruikt in de verschillende bronnen. Daarbij wordt eerst de gewenste inhoud van de variabele gedefinieerd. Bij deze definities wordt wel met een schuin oog gekeken naar de mogelijkheden om deze in de beschikbare bronnen te operationaliseren. Vervolgens wordt per bron beoordeeld of de daarin gebruikte definitie afwijkt van de gewenste. Als de definities van elkaar verschillen wordt gezocht naar andere variabelen waarmee het verschil kan worden gecorrigeerd. De ene definitie wordt als het ware kwantitatief vertaald in de andere.

In het banenbestand is een *baan* van een werknemer gedefinieerd als een relatie tussen een persoon en een bedrijf of instelling in statistische zin, de zogenaamde *bedrijfsseenheid*. Bij het bepalen of sprake is van een baan en voor de bepaling van aanvangs- en beëindigings-datums van banen is binnen het SSB gekozen voor het *betalingsconcept*. Een baan of uitkering telt in deze optiek mee als er daadwerkelijk betalingen hebben plaatsgevonden. Wijziging van aanvangs- en beëindigingsdatums en het selecteren van bepaalde banenrecords hebben tot doel beter aan te sluiten op het betalingsconcept. Er wordt daarbij aangenomen dat het betaalde loon in de vastgestelde periode is verdiend.

In administraties komen records voor die niet onder de bovenstaande definitie van 'baan' vallen. Records uit de

Figuur 1. Het banenbestand SSB: relaties tussen de bronnen.



Fibase worden niet als werknemersbanen beschouwd als het vermelde fiscale loon nul of lager dan nul is, of de loonbelastingtabel aangeeft dat het niet om inkomen uit tegenwoordige arbeid gaat of het soort inkomen aangeeft dat het een uitkering is. Voor het VZA-bestand gelden soortgelijke afbakeningen: een record wordt niet als een werknemersbaan beschouwd als het bruto loon voor de sociale verzekering nul of lager dan nul is, en de baan niet als werknemersbaan voorkomt in de Fibase. Als een record van een persoon in het VZA-bestand uitsluitend als uitkeringsrecord voorkomt in het fiscale bestand, waarvan het fiscale loon overeenstemt met het bruto loon voor de sociale verzekering uit het VZA-bestand, dan komt het VZA-record als baan te vervallen.

4.3. Administratieve fouten

Als er daarna nog verschillen overblijven, wordt beoordeeld wat voor ieder van de variabelen de beste bron is. Bij de bepaling van wat de beste bron voor een variabele is, wordt rekening gehouden met de administratieve praktijk. Vooral als een variabele voor een berichtgever niet erg belangrijk is (bijvoorbeeld begin- en einddatum van een baan in fiscale gegevens) is de kwaliteit dikwijls twijfelachtig. Daarnaast is vaak sprake van administratieve vertraging. Er wordt geprobeerd om de gevolgen van administratieve fouten op de uitkomsten zoveel mogelijk te beperken.

Wanneer een persoon met een werknemersbaan volgens het GBA-bestand is gestorven, dan geldt de sterftedatum als beëindigingsdatum van de baan. Wanneer de sterftedatum vóór het verslagjaar ligt, dan is geen sprake van een baan

en wordt het record niet in het banenbestand opgenomen. Het gaat hierbij om enkele honderden gevallen per jaar. Als beste bron is hier dus voor de GBA gekozen.

Indien het kiezen van een beste bron niet mogelijk is, bijvoorbeeld omdat de kwaliteit van de bronnen onvoldoende bekend is, dan kan ook een nieuwe variabele bepaald worden op grond van twee of meer bronnen. Een voorbeeld daarvan is het bepalen van de begindatum van een baan. Veel banen hebben in de Fibase en de VZA als aanvangsdatum 1 januari en als beëindigingsdatum 31 december van een verslagjaar. De reden hiervoor is dat voor veel registers 1 januari en 31 december een soort ‘administratieve’ standaarddata zijn. Wanneer verschillende bronnen verschillende aanvangs- en beëindigingsdata aangeven, dan worden keuzen gemaakt. Wanneer een baan in één bron een aanvangsdatum van 1 januari en een andere bron een andere datum heeft, dan wordt de datum uit de laatste bron overgenomen. Het omgekeerde geldt ook voor banen met een beëindigingsdatum van 31 december. Voor banen die mede gebaseerd zijn op informatie uit de VZA en de EWL geldt bovendien als voorwaarde dat het aantal sociale verzekeringsdagen past binnen een eventueel nieuw vastgestelde periode tussen aanvang en einde van de baan.

5. Resultaten

5.1. Mogelijkheden voor longitudinale analyses

In deze paragraaf worden enkele voorbeelden gegeven van de mogelijkheden die het SSB biedt tot samenstellen van stroominformatie. Deze worden uitsluitend ontleend aan de informatie die afkomstig is uit de registers. Een voorbeeld heeft betrekking op de stromen tussen werkenden en zelfstandigen, gepensioneerden, uitkeringstrekkers, en scholieren en studenten. Een ander voorbeeld betreft de positie op de arbeidsmarkt van afgestudeerden uit het hoger onderwijs.

5.2. Dynamiek op de arbeidsmarkt

Hoewel in beginsel iedere gebeurtenis (bijvoorbeeld de overgang van een baan naar een uitkering) ook van een datum voorzien kan worden, is ervoor gekozen om de stromen te laten zien tussen twee peilmomenten: de stromen tussen de laatste werkdag van september van 1999 en 2000. De reden dat de laatste werkdag van een kwartaal gekozen is, is omdat in het weekend allerlei gebeurtenissen plaatsvinden zoals beëindiging van banen. De reden dat de laatste werkdag van het derde kwartaal gekozen is, is omdat dan sprake is van een relatief stabiele situatie: er zijn dan geen sterke seizoensinvloeden en ook de wetgeving die van invloed is op de inhoud van de gebruikte registers wijzigt meestal niet in die periode.

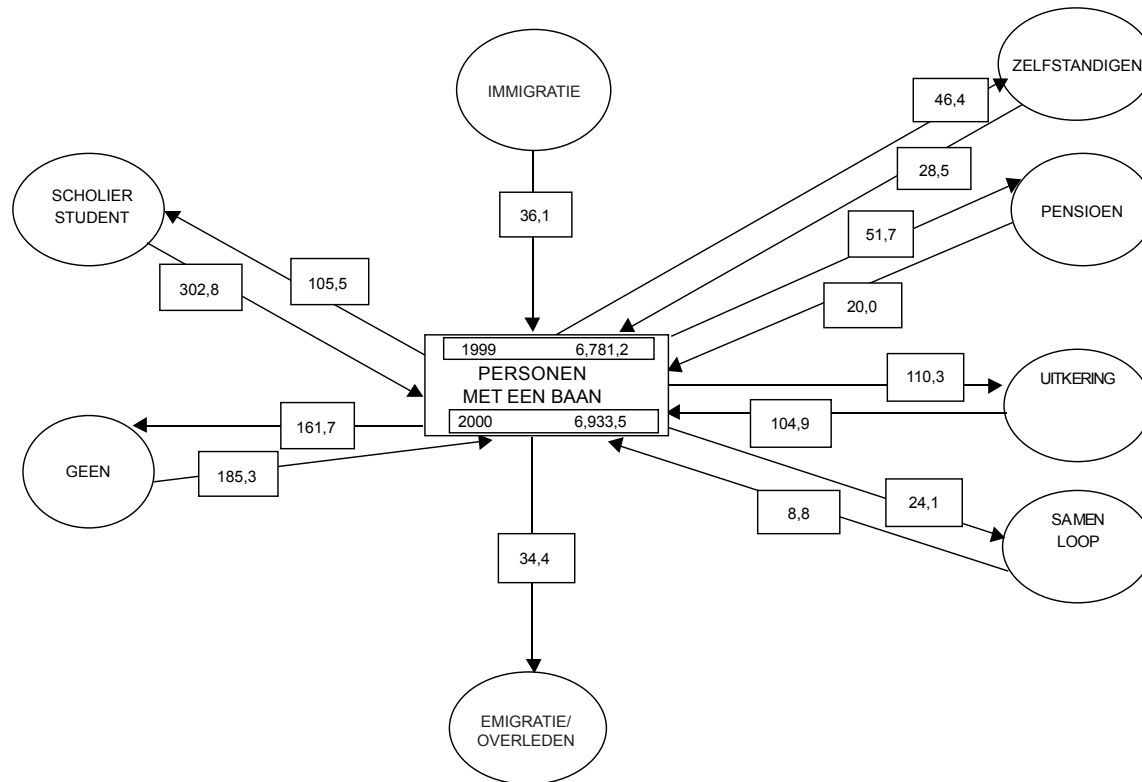
Eind september 1999 waren er van de Nederlandse bevolking 6,8 miljoen personen met een baan als werknemer en een jaar later ruim 6,9 miljoen. Per saldo

nam het aantal werknemers toe met 152 duizend. Dit saldo bestaat uit meerdere stromen tussen een baan als werknemer en andere posities op de arbeidsmarkt (figuur 2). De uitgaande pijlen geven de aantallen mensen weer die eind september 1999 als werknemer werkten en eind september 2000 niet meer, maar toen wel een andere positie innamen. De ingaande pijlen geven het omgekeerde weer: personen die eind september 1999 geen werknemer waren en een jaar later wel.

In totaal waren er ruim 534 duizend personen die in september 1999 een baan als werknemer hadden, maar in september 2000 niet meer (uitstroom). Daar stonden bijna 686 duizend personen tegenover die in september 1999 geen baan hadden, maar wel in september 2000 (instroom). Bijna 92 procent van de werknemers van eind september 1999 was ook een jaar later werkzaam.

De grootste instroom op de arbeidsmarkt tussen september 1999 en september 2000 vond plaats vanuit de positie scholier/student (303 duizend). Enerzijds gaat het hierbij om scholieren en studenten die naast school of studie een baan als werknemer hebben, anderzijds om scholieren en studenten die na hun studie zijn gaan werken. Per saldo kwamen er uit de positie scholier/student tussen september 1999 en september 2000 ongeveer 197 duizend werknemers bij. Ook waren er grote stromen van en naar de positie “geen”. Het gaat daarbij om personen die niet werkzaam zijn, geen uitkering of pensioen hebben en ook niet als scholier/student getypeerd kunnen worden. Vanuit deze positie stroomden ruim 185 duizend personen tussen september 1999 en september 2000 in op de arbeidsmarkt,

Figuur 2. Stromen op de arbeidsmarkt, september 1999 – september 2000 (×1000)



terwijl bijna 162 duizend personen vanuit een baan als werknemer naar “geen” uitstroomden.

Er waren eind september 1999 110 duizend personen met een baan als werknemer, die eind september 2000 geen baan meer hadden, maar een uitkering. Dat wil overigens

niet zeggen dat deze personen in september 1999 naast hun baan geen andere inkomstenbron hadden. Ruim 40 procent hiervan had eind september 1999 naast de baan nog een andere inkomensbron. Omgekeerd waren er in september 1999 bijna 105 duizend mensen met een uitkering die in september 2000 werkten als werknemer.

Verder zijn er nog enkele relatief kleine stromen van en naar pensioen, zelfstandigen en *samenloop*. Van samenloop is bijvoorbeeld sprake wanneer iemand naast pensioen nog een andere uitkering ontvangt of daarnaast als zelfstandige werkzaam is. Van de mensen die in september 1999 pensioen ontvingen, hadden er 20 duizend in september 2000 een baan als werknemer. In bijna alle gevallen was dat een baan naast de pensioenuitkering.

Tot slot komen er door immigratie ruim 36 duizend werknemers bij en stromen er door sterfte en emigratie 34 duizend personen uit.

Interessant is om een van deze stromen onder de loep te nemen en uit te splitsen naar achtergrondvariabelen. Dat wordt gedaan door de uitstroom van werknemers naar arbeidsongeschiktheid uit te splitsen naar de herkomstgroepering. Daarmee kan de vraag worden beantwoord of de kans dat (verschillende categorieën) allochtonen uitstromen naar arbeidsongeschiktheid verschilt van die van autochtonen. Voor de totale bevolking is de kans om arbeidsongeschikt te worden 0,8% (zie figuur 3). Voor Marokkanen ligt die kans echter op 1,7% en voor Turken zelfs op 2,0%. De overige niet-westerse allochtonen lijken veel meer op autochtonen.

5.3. Kans op werk bij hbo-ers groter dan bij wo-ers

Voor de studenten uit het hoger onderwijs die in het studiejaar 1997-1998 hun diploma hebben gehaald is de arbeidsmarktpositie nagegaan binnen één, twee en drie jaar na hun studie (figuur 4). Tevens is nagegaan wat de startsalarissen zijn voor deze gediplomeerden. Uiteraard

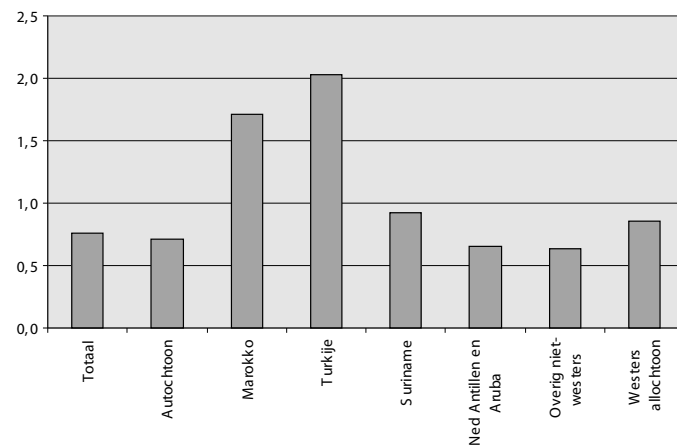
kunnen studenten het gehele jaar hun studie afronden. Voor het startsalaris is de peildatum de laatste werkdag van september 1998 genomen. We hebben ons beperkt tot de voltijd werknemers (figuur 5).

De kans om een betaalde baan te vinden binnen een jaar na afronding van de studie is voor hbo-ers groter dan voor universitair opgeleiden. Het verschil wordt echter wel steeds kleiner naarmate de tijd vordert.

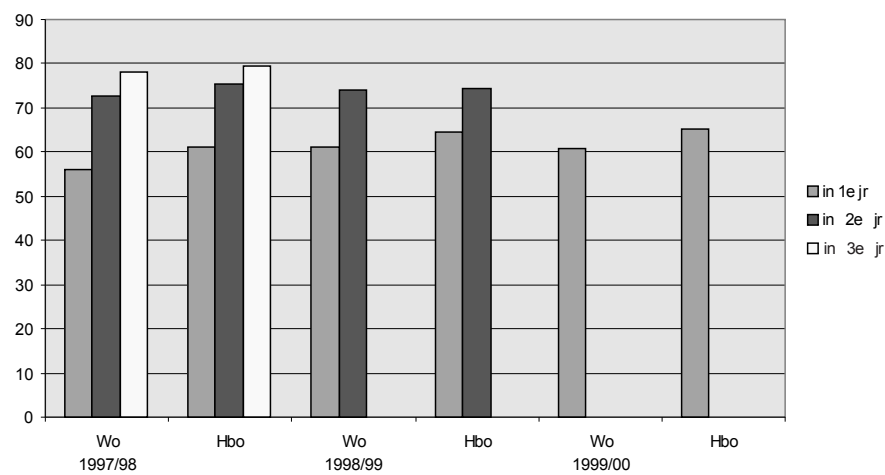
Gediplomeerden van het wetenschappelijk onderwijs hebben dan ook vaker een uitkering dan gediplomeerden van het hoger beroepsonderwijs. Ook dit verschil verdwijnt echter na enkele jaren. Er zijn ook kleine verschillen tussen de afstudeercohorten. Deze verschillen zijn echter zo klein dat daaraan weinig betekenis moet worden toegekend.

Wetenschappelijk opgeleiden hebben gemiddeld wel een hoger loon. Met uitzondering van de landbouwkundige opleidingen geldt dat voor alle studierichtingen. Het verschil is het grootst bij de medische opleidingen. Daar verdienen wetenschappelijk opgeleiden ongeveer 30% meer dan hun collega's die een hbo-opleiding hebben afgerond. Dit is uiteraard geen verrassing, omdat dit de tegenstelling weerspiegelt tussen (tand)artsen en verpleegkundigen.

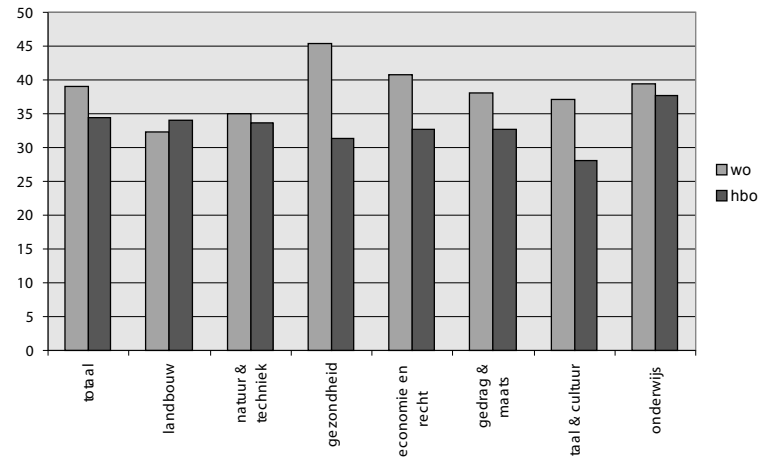
Figuur 3. Stroom van werknemers naar arbeidsongeschiktheid tussen september 1999 en september 2000 naar herkomst-groepering



Figuur 4. Het percentage werkenden na diplomering in het hoger onderwijs naar opleidingsniveau en afstudeercohort



Figuur 5. Gemiddeld jaarloon ($\times 1000$ Euro) voor voltijd werknemers binnen een jaar na diplomering naar opleidingsniveau en –richting, 1998



6. Slot

In dit paper zijn de mogelijkheden geschetst om longitudinaal onderzoek te verrichten waarbij gebruik gemaakt wordt van de gegevens uit het SSB. Daarin is een groot aantal registers onderling koppelbaar opgeslagen en geïntegreerd. Door gebruik te maken van het feit dat deze registers zijn gekoppeld voor meerdere jaargangen ontstaat een groot integraal longitudinaal bestand. Er zijn drie mogelijkheden om longitudinale analyses te verrichten op de gegevens uit het SSB.

In de eerste plaats is dat de analyse van de gegevens in het SSB zelf. Daarbij gaat het uitsluitend om de registratieve gegevens die in het SSB zijn opgenomen. Omdat die integraal zijn en grotendeels op volumebasis, kunnen deze

longitudinaal worden gebruikt. Enige voorzichtigheid dient hierbij wel betracht te worden, omdat de gegevens niet volledig longitudinaal gecorrigeerd zijn. Zo is niet voor alle variabelen gecorrigeerd voor administratieve veranderingen. Dit kan betekenen dat de stromen enigszins overschat worden. Ook is nog niet beoordeeld, wat voor gevolgen veranderingen van de toekenning van de economische activiteit aan bedrijven voor de uitkomsten hebben. Hetzelfde bedrijf kan in twee opeenvolgende jaren een andere economische activiteit hebben toegewezen gekregen. Als dit zou worden gerekend tot baanmobiliteit, dan zou de totale baanmobiliteit worden overschat.

Op de langere termijn worden ook andere registers in het SSB opgenomen. Daarbij gaat het om registers waarin opleidingsinformatie is opgenomen (bestanden

van de vroegere Arbeidsbureaus en de huidige Centra voor Werk en Inkomen, het Examen Resultaten Register en de onderwijsnummerbestanden van de Informatie Beheer Groep), en de registers op het terrein van inkomen (voornamelijk fiscale gegevens). Dit vergroot de analysemogelijkheden nog aanzienlijk.

In de tweede plaats kunnen oudere surveys en registers aan het SSB gekoppeld worden. Dit is als het ware een virtuele *follow up* van deze surveys. Zo kan men de respondenten uit het Sociaal-economisch panelonderzoek aan het SSB koppelen, zodat voor een aantal cruciale variabelen de actuele situatie in beeld wordt gebracht. Ook hier is enige voorzichtigheid op zijn plaats. Uit eerste confrontaties van registers en enquêtes blijkt dat deze elkaar nog wel eens tegenspreken. Bijvoorbeeld omdat een persoon in een enquête zegt een baan te hebben, maar deze baan niet in registers wordt teruggevonden. Als dit een flinke omvang zou hebben, dan is de schatting van de omvang van stromen vertekend.

In de derde plaats kunnen oudere jaargangen van het SSB worden gekoppeld aan recentere surveys. Zo kan bijvoorbeeld worden nagegaan wat de gevolgen voor de leefsituatie zijn van mensen die arbeidsongeschikt zijn geworden. Ook hier geldt hetzelfde voorbehoud als gemaakt is bij het koppelen van oudere surveys aan recentere data uit het SSB.

Met het SSB worden nieuwe wegen geopend voor longitudinale analyses waarmee belangrijke maatschappelijke vragen beantwoord kunnen worden. Door het SSB uit te breiden met registers over het opleidingsniveau, de

op dit moment gevolgde opleiding en inkomens, worden de mogelijkheden nog verder vergroot.

Noten

- (1) Bart Bakker is programmamanager Eindintegratie en Sociaal Statistisch Bestand op het CBS. bbkr@cbs.nl. Koos Arts is projectleider bij hetzelfde programma. kars@cbs.nl
- (2) Voor meer informatie over het strategisch programma zie <http://www.cbs.nl/nl/service/onderzoek/strat-ondz/SDA-projecten.htm>

Referenties

- Arts, C.H. & E.M.J. Hoogteijling, 2002, Het Sociaal Statistisch Bestand 1998 en 1999, In: *Sociaal-economische maandstatistiek*, december 2002, pp. 13-21
- Bakker, B.F.M., 2002, Statistics Netherlands' Approach to Social Statistics: The Social Statistical Dataset, In: *OECD Statistics Newsletter*, vol. 2002, nr. 11, blz. 4-6
- Laan, P. van der, 2000, Integrating administrative registers and household surveys, In: P. Al en B.F.M. Bakker (red.), *Re-engineering social statistics by micro-integration of different sources*. Themanummer Netherlands Official Statistics, jrg. 15, nr. summer blz. 7-15

CBS wees zuinig op de enquêtes!

Een reactie op het paper van B. Bakker en K. Arts

P. de Beer (1)

Zo'n twintig jaar geleden waren de bestanden die inzicht moesten geven in de werkloosheid vervuild. De betrouwbaarheid van de statistiek van de geregistreerde werkloosheid nam dan ook af. Om weer met een betrouwbaar cijfer te komen startte het CBS met de Enquête BeroepsBevolking (EBB). Daarin werden mensen vragen voorgelegd en op basis daarvan werd bepaald of iemand werkloos was of niet. Met de EBB kwam er een nieuwe definitie voor werkloosheid.

De eerste gedachte die De Beer dan ook bekreep toen hij hoorde van het SSB was dat het CBS de klok wilde terug draaien. Dat komt doordat de enquêtes die vroeger nodig waren om tot een Volkstelling te komen nu niet meer nodig zijn, want deze kan samengesteld worden uit het SSB. Dat registers niet altijd even betrouwbaar zijn gaf de registratie van de werkloosheid in de jaren 80 al aan en sinds de Bijlmerramp weten we dat ook de Gemeentelijke BasisAdministratie – de basis van het SSB - niet betrouwbaar is. Dus de betrouwbaarheid van registers blijft ter discussie staan. Daarom vindt De Beer, dat het nodig blijft dat er naast registers ook enquêtes gehouden worden. Hij ziet echter dat het zwaartepunt steeds meer verschuift richting registers.

Registers zijn niet alleen onderhevig aan vervuiling, het aantal categorieën waarover informatie verzameld wordt is ook beperkt. Zo is het maar zeer de vraag of de registers informatie bevatten over bijvoorbeeld ZZP'ers (zelfstandigen zonder personeel). Bovendien zijn er zeer vage overgangscategorieën. De overgangen worden steeds diffuser. Volgens De Beer brengt het CBS een kunstmatige scheiding aan, daardoor ontstaat een schijnwereld. En die virtuele schijnwereld moet niet gecreëerd worden. Als dat wel gebeurt, dan is het contact met de echte wereld verloren. Dat is volgens De Beer ook de grootste zorg die wetenschappers hebben voor het SSB en de virtuele Volkstelling die daaruit samengesteld wordt.

Door de koppeling van bestanden ontstaan soms, zoals Bakker heeft laten zien, inconsistente gegevens. Het CBS probeert zorgvuldig om te gaan met de problemen die verbonden zijn aan het koppelen van bestanden. De Beer laat echter weten dat we erg voorzichtig moeten zijn met de methode die het CBS hanteert. Mensen moeten niet dwangmatig aan één bepaalde categorie toegewezen worden. Volgens De Beer kunnen mensen soms gewoon in twee of meer categorieën vallen. Iemand kan aangeven student, werkend en werkloos te zijn. Dit hoeft niet per se een inconsistente combinatie te zijn. De door het CBS gehanteerde methode zorgt al snel dat we iets als inconsistent beschouwen, en dan wordt het misschien ten onrechte geschrapt. We moeten oppassen dat statistici de bevolking te makkelijk eenduidig indelen. De werkelijkheid is diffuser. Dat betekent ook dat we duidelijk moeten kunnen vaststellen wat de oorspronkelijke en wat de geschatte gegevens zijn.

Het is zeer toe te juichen dat met het SSB stromen op de arbeidsmarkt in kaart kunnen worden gebracht. Het is namelijk erg belangrijk om goed zicht te hebben op de dynamiek van de arbeidsmarkt. Alle lof voor de bereidheid dat in kaart te brengen, aldus De Beer.

Maar waarop is de dynamiek in het SSB gebaseerd, op register- of ook op enquêtegegevens?

De dynamiek voor ieder individu afzonderlijk is alleen in kaart te brengen via registers. Met de diffuse grenzen tussen de categorieën in de registers is het echter makkelijk om de dynamiek te overschatten. Stel dat iemand een uur in de week meer gaat werken, bijvoorbeeld van elf uur in de week naar twaalf uur in de week. In feite is de toename in werkgelegenheid miniem. Uit de statistiek blijkt echter iets anders: die geeft namelijk aan dat iemand van werkloos werkend is geworden.

Het afschaffen van het Sociaal- Economische Panel (SEP) is een zorgelijke ontwikkeling, want de informatie die daarmee verzameld wordt, is niet uit de registers te halen.

Een ander probleem dat De Beer signaleert is de sterke focus op personen - er zou ook meer naar dynamiek van de statistische eenheid baan gekeken moeten worden. De definitie die het CBS hanteert voor een baan is een relatie tussen een persoon en de werkgever. Dus als de persoon een andere baan vindt en de relatie verbreekt, dan verdwijnt de baan. Dat is raar, want in feite blijft de functie gewoon bestaan. Het verbreken van de relatie verandert niets aan de baan.

De Beer sluit zijn betoog af met de hoop dat onderzoekers toegang blijven houden tot de oorspronkelijke gegevens

en niet dat zij niet alleen maar toegang krijgen tot de door het CBS bewerkte gegevens.

Noten

- (1) Paul de Beer is bijzonder hoogleraar arbeidsverhoudingen (Henri Polak-leerstoel) aan de Universiteit van Amsterdam en verbonden aan ALAS en De Burcht. Hij heeft in het verleden veel kwantitatief arbeidsmarktonderzoek gedaan op basis van microdata van het CBS (m.n. de EBB) en is lid geweest van de Commissie van Advies voor Arbeidsstatistiek die onder andere heeft geadviseerd over het SSB.

Loonstructuuronderzoek – verrijkte data door integratie

E. Schulte Nordholt (1)

1. Inleiding

Het CBS stelt regelmatig gegevens samen over de verdiende lonen van werknemers in relatie met hun opleidings- en beroepsniveau op basis van het Loonstructuuronderzoek (LSO). In de jaren zeventig werd nog een zelfstandig grootschalig Loonstructuuronderzoek gehouden. In de jaren tachtig werden zogenaamde na-enquêtes gehouden onder een steekproef van bedrijven die in het kader van het Jaarlijks Loononderzoek van het CBS al looninformatie over hun werknemers hadden verstrekt.

De onderzoeken zijn sinds de jaren negentig samengesteld uit al beschikbare gegevens. Informatie over LSO95 en LSO97 is gepubliceerd in respectievelijk Boerdam, Loeve en Ruijs (1998) en Schulte Nordholt en Ruijs (2000). Het onderzoek over het verslagjaar 1998 is in het kader van het Sociaal Statistisch Bestand (SSB) op een andere manier uitgevoerd, waardoor de uitkomsten met de voorgaande LSO-en niet zonder meer te vergelijken zijn. Een verschil tussen LSO98 en de voorgaande LSO-en is veroorzaakt door het gebruik maken van meer bronnen. Daardoor worden meer, vooral kleinere banen meegenomen in het LSO98. Kleinere banen hebben een lage wekelijkse arbeidsduur of beslaan een beperkte

periode van het jaar. De populatie voor het LSO98 is beperkt tot werknemers van 15 tot en met 64 jaar die woonachtig zijn in Nederland en deel uitmaken van de bevolking in particuliere huishoudens. De kosten van een Loonstructuuronderzoek zijn tegenwoordig relatief beperkt en er is geen sprake meer van enquêtedruk omdat louter gebruik wordt gemaakt van al bij het CBS aanwezige bronnen.

In dit artikel wordt beknopt ingegaan op de gehanteerde methode van onderzoek en worden enige belangrijke uitkomsten van het LSO98 gepresenteerd. In paragraaf 2 wordt de methode van samenstellen van Loonstructuuronderzoekgegevens toegelicht. Ook worden enige aandachtspunten bij de uitkomsten weergegeven. In paragraaf 3 worden uitkomsten gepresenteerd over banen, gemiddelde uurlonen en maandlonen naar opleidingsniveau en beroepsniveau voor een aantal kenmerken.

Voor een beschrijving van de gebruikte methode in het LSO98 wordt verwezen naar Arts en Hoogteijling (2002). Dit artikel kan worden beschouwd als een verkorte en geactualiseerde versie van Van Cruchten, Hartgers en Schulte Nordholt (2002).

2. Methode van samenstellen

De benodigde informatie voor de LSO-en over de verslagjaren 1995, 1996 en 1997 zijn verkregen door combinatie van gegevens op persoonsniveau uit drie bronnen: de Enquête werkgelegenheid en lonen (EWL, een bedrijfsenquête), de Enquête beroepsbevolking (EBB,

een persoonsenquête) en de Verzekerdenadministratie (VZA, een register). Het onderzoek over het verslagjaar 1998 is op een andere manier uitgevoerd, waardoor de uitkomsten met de voorgaande LSO-en niet zonder meer te vergelijken zijn.

In de eerste plaats is de methode om enquêtegegevens te koppelen aan registergegevens veranderd. Voor het LSO97 zijn de VZA en de EWL afzonderlijk gekoppeld aan de EBB. In een volgende stap zijn voor de ontbrekende EWL-variabelen in het deelbestand VZA-EBB de gegevens geïmputeerd en zijn de deelbestanden samengevoegd. Het LSO98 is samengesteld met behulp van de methode van herhaald wegen (Houbiers et al., 2003). De bestanden die worden gebruikt zijn: de Gemeentelijke Basisadministratie (GBA), de EWL, het SSB-Banenregister en de EBB. Door de combinatie van registers en steekproeven worden vier blokken gevormd, waaruit de LSO-tabellen geschat worden (Arts en Hoogteijling, 2002, paragraaf 3.7.3, Schema 1). Voor de samenstelling van het LSO98 worden de gewichten uit de steekproeven telkens gekalibreerd naar alle relevante registertellingen en de eerder geschatte tabellen uit de steekproeven zodat alle tabellen onderling consistent zijn. Elke tabel moet berekend worden uit het grootste blok waaruit de tabel bepaald kan worden. Als alle tabellen met de juiste gewichten worden geschat, zijn de tabeluitkomsten onderling consistent en zijn de uitkomsten onafhankelijk van de gebruikte bron. Door telkens uit te gaan van het grootst mogelijke blok, kan de structuur van de lonen van werknemers nauwkeurig worden beschreven.

In de tweede plaats is de populatie waarop de cijfers betrekking hebben in het LSO98 anders dan in de voorgaande LSO-en. De populatie is beperkt tot werknemers van 15 tot en met 64 jaar die woonachtig zijn in Nederland en deel uitmaken van de bevolking in particuliere huishoudens. In de voorgaande LSO-en zijn ook werknemers van 65 jaar en ouder opgenomen, evenals (een gering aantal) werknemers die wel in Nederland werken, maar niet in Nederland wonen.

In de derde plaats zijn bij het LSO97 de uitkomsten herwogen naar de uitkomsten van de EWL 1997, waardoor de totalen exact hetzelfde zijn voor de kenmerken die bij de herweging zijn gebruikt. De totalen van het LSO98 zijn echter niet gelijk aan die van de EWL 1998 zoals gepubliceerd in de Sociaal-economische maandstatistiek (CBS, 2000, paragraaf 2.2). Het is de bedoeling dat in de toekomst bij publicaties over banen en lonen ook de kleinere banen die niet in bedrijfsadministraties voorkomen, worden meegenomen.

De gepubliceerde aantallen werknemers in de tabellen hebben betrekking op ultimo 1998 waarbij de periode 25-31 december 1998 als peilweek is gebruikt om het aantal banen vast te stellen. In de gebruikte bestanden is informatie over banen van werknemers vastgelegd. Dat wil zeggen dat een werknemer meerdere keren kan voorkomen als hij of zij meerdere banen tegelijkertijd heeft. De gemiddelde maandlonen zijn berekend door het maandloon van alle banen in een tabelcel te sommeren en te delen door het aantal banen in die cel. Voor de berekening van de gemiddelde uurlonen wordt voor elke tabelcel het maandloon van alle banen vermenigvuldigd

met 12 en vervolgens gedeeld door de som van de arbeidsduren van banen in die betreffende cel.

In de uitkomsten van het LSO moet rekening worden gehouden met onnauwkeurigheidsmarges, omdat het LSO gedeeltelijk op basis van steekproefgegevens is samengesteld. In verband met de betrouwbaarheid van de uitkomsten wordt de vuistregel gehanteerd dat uitsluitend gegevens worden verstrekt die betrekking hebben op meer dan 2 500 werknemers in de populatie. De exacte onnauwkeurigheidsmarges zijn niet te geven vanwege het complexe steekproefontwerp van zowel de EWL als de EBB. Naast de nauwkeurigheid worden de publicatiemogelijkheden beperkt door de toegepaste geheimhouding op de gegevens. Deze heeft tot doel het voorkomen dat gegevens herleidbaar zijn tot individuele bedrijven of personen.

3. Uitkomsten Loonstructuuronderzoek 1998

3.1. Banen naar opleidings- en beroepsniveau

Ultimo 1998 hadden werknemers in Nederland in totaal 6,4 miljoen banen. Gemiddeld verdienden zij ruim 14 euro per uur. Het gemiddelde maandloon bedroeg 1631 euro. Werknemerskenmerken die van belang zijn voor het niveau van het feitelijk verdiende loon zijn onder meer het opleidingsniveau, het beroepsniveau en de arbeidservaring.

Voor de indeling van het opleidingsniveau zijn de eerste drie cijfers van de Standaard onderwijsindeling

gebruikt. Het opleidingsniveau wordt voor het LSO als volgt ingedeeld: basisonderwijs, mavo, vbo, havo/vwo, mbo, hbo en wo. De kleine aantallen waarvan het opleidingsniveau niet bekend is, zijn in het LSO98 toegevoegd aan de categorie basisonderwijs.

De Standaard beroepenclassificatie wordt gebruikt voor de indeling in beroepsniveaus. Voor het LSO worden vijf verschillende niveaus onderscheiden: elementair, lager, middelbaar, hoger en wetenschappelijk. In het LSO98 is er voor gekozen om de personen van wie geen beroep bekend is, alleen op te nemen in het totaal.

Over de arbeidservaring van werknemers zijn niet voldoende gegevens beschikbaar om betrouwbare uitspraken te doen. De arbeidservaring kan worden benaderd door uit te gaan van andere gegevens, zoals de leeftijd van de werknemers. De meeste banen behoren toe aan werknemers in de leeftijdsklasse van 25 tot 35 jaar, namelijk 29 procent.

Ultimo 1998 werd van alle banen van werknemers 15 procent vervuld door allochtonen; bijna 7 procent door niet-westerse en ruim 8 procent door westerse allochtonen.

Het aantal banen van werknemers met een hbo- of wo-opleiding bedroeg bij de autochtonen 25 procent van het totale aantal banen van autochtonen. Het aandeel banen van werknemers met hbo of wo bij de niet-westerse allochtonen was beduidend lager: 15 procent. Bij de westerse allochtonen lag het aandeel iets hoger dan bij de autochtonen: 29 procent.

Ruim 71 procent van alle banen bevindt zich bij een bedrijf of instelling in de cao-sector particuliere bedrijven, 15 procent bij de gesubsidieerde sector en 14 procent bij de overheid. Tussen de cao-sectoren was er een verschil in banen naar opleidingsniveau. Bijna 46 procent van de banen in de gesubsidieerde sector betrof banen van werknemers met mbo. Bij de particuliere bedrijven was dat 37 procent en bij de overheid 28 procent. Bij de banen van werknemers met een hogere beroepsopleiding (hbo) of een wetenschappelijke opleiding (wo) was de hiërarchie in de cao-sectoren duidelijk anders. De overheid was hierin koploper: 55 procent van de banen in deze sector betrof banen van werknemers op hbo- of wo-niveau. Bij de gesubsidieerde sector was het aandeel 27 procent en bij de particuliere bedrijven 19 procent.

Onder werknemers kwam ultimo 1998 het middelbaar beroepsonderwijs (mbo) het meest voor: 37 procent. Het beroepsniveau dat onder de werknemers het meeste voorkwam is het middelbare beroepsniveau. Op dit niveau bevond 35 procent van de banen van werknemers zich. Arbeidservaring en aanvullende opleidingen kunnen een reden zijn dat het uitgeoefende beroep van een ander niveau is dan het hoogst behaalde opleidingsniveau zou doen verwachten.

3.2. Beloningsverschillen naar opleidingsniveau

Werknemers met een hogere beroepsopleiding (hbo) verdienden eind 1998 gemiddeld 18 euro per uur en werknemers met een wetenschappelijke opleiding (wo) 22 euro. Een hoger opleidingsniveau leidt voor een werknemer niet altijd tot een hoger uurloon. Naast

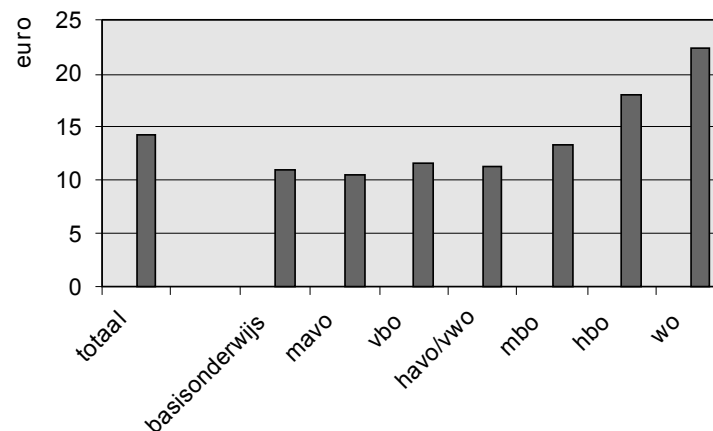
het opleidingsniveau kan een aantal andere kenmerken van invloed zijn op het uurloon, zoals geslacht, leeftijd, beroep, arbeidservaring, dienstverband en de bedrijfstak waar men werkt. De opleidingsniveaus zijn heterogeen van samenstelling naar deze kenmerken. Werknemers met alleen basisonderwijs zijn bijvoorbeeld gemiddeld ouder dan de werknemers met een mavo-diploma. Dit heeft een verhogend effect op het gemiddelde uurloon van werknemers met alleen basisonderwijs. Alle kenmerken van de groepssamenstelling samen hebben tot gevolg dat de gemiddelde werknemer met alleen basisonderwijs een iets hoger uurloon heeft dan de gemiddelde werknemer met een mavo-diploma.

Uit Figuur 1 blijkt dat de gemiddelde uurlonen toenemen met het opleidingsniveau. Het laagste gemiddelde uurloon werd verdiend door werknemers met basisonderwijs of een mavo-diploma (11 euro) en het hoogste gemiddelde uurloon door werknemers met een wetenschappelijke opleiding (22 euro).

Geslacht en leeftijd

Het gemiddelde uurloon van mannen was eind 1998 ongeveer 27 procent hoger dan dat van vrouwen. Dit percentage verschilde per opleidingsniveau. Het kleinst waren de verschillen tussen mannen en vrouwen met een havo/vwo-diploma (7 procent) en het grootst bij de werknemers met een vbo en met een hbo-diploma (34 en 35 procent). De verschillen tussen de maandlonen van mannen en vrouwen zijn voor alle opleidingsniveaus groter dan die tussen de uurlonen. Dit komt doordat vrouwen vaker in deeltijd werken.

Figuur 1. Uurloon van alle werknemers naar opleidingsniveau, ultimo 1998



Het verschil in uurloon tussen jongere en oudere werknemers was niet voor alle opleidingsniveaus gelijk. Gemiddeld verdiende een 55-plusser per uur 2,4 keer zoveel als een werknemer die jonger is dan 25 jaar. Voor werknemers met mbo was dit verhoudingsgetal het laagst (1,9) en voor werknemers met havo/vwo het hoogst (3,1).

Herkomstgroepering en cao-sector

Westerse allochtonen lijken wat betreft gemiddeld uurloon meer op autochtonen dan op niet-westerse allochtonen. Het verschil in gemiddeld uurloon tussen autochtonen en westerse allochtonen was ultimo 1998 minder dan 0,5 euro. Het verschil met de niet-westerse allochtonen bedroeg meer dan 3 euro.

Werknemers met een baan bij de overheid verdienden gemiddeld per uur 28 procent meer dan werknemers bij particuliere bedrijven en zelfs 31 procent meer dan de werknemers in de gesubsidieerde sector. Dat werknemers bij de overheid gemiddeld meer verdienen dan in de andere cao-sectoren, komt niet zozeer door de verschillen in uurloon per opleidingsniveau, maar vooral door het grotere aandeel werknemers met een hbo- of wo-opleiding bij de overheid. Ook moet rekening worden gehouden met het gegeven dat de werknemers bij de overheid gemiddeld genomen ouder zijn dan de werknemers in de andere cao-sectoren.

Werknemers bij de overheid verdienden ultimo 1998 in bijna alle opleidingsniveaus meer dan de werknemers in de andere cao-sectoren. Werknemers met een wo-opleiding

bij de overheid verdienden per uur gemiddeld meer dan de werknemers met een soortgelijk opleidingsniveau bij de particuliere bedrijven en de gesubsidieerde sector; respectievelijk 4 en 5 procent. Het leeftijdseffect kan hier debet aan zijn. Het gemiddelde maandloon van werknemers met een wo-opleiding bij de overheid was lager dan het gemiddelde maandloon van werknemers bij de particuliere bedrijven. Dit komt doordat werknemers bij de overheid een kortere gemiddelde wekelijkse arbeidsduur hebben.

3.3. Beloningsverschillen naar beroepsniveau

De Standaard beroepenclassificatie deelt beroepen in de eerste plaats in naar niveau, op basis van de voor een bepaald beroep benodigde kennis en ervaring. Als startpunt voor de bepaling van het niveau, wordt het niveau van de opleiding gebruikt die het beste voorbereidt op de werkzaamheden. Het opleidingsniveau en het beroepsniveau zullen dan ook bij veel werknemers samenhangen.

Uit Figuur 2 blijkt dat de gemiddelde uurlonen toenemen met het beroepsniveau. In elementaire beroepen werd ultimo 1998 gemiddeld 9 euro per uur verdiend en in wetenschappelijke beroepen 23 euro.

Geslacht en leeftijd

Ultimo 1998 verdienden mannen op alle beroepsniveaus een hoger uurloon dan vrouwen. Daarbij gold: hoe hoger het beroepsniveau, hoe groter het verschil. Bij de elementaire beroepen verdienden mannen gemiddeld bijna één euro

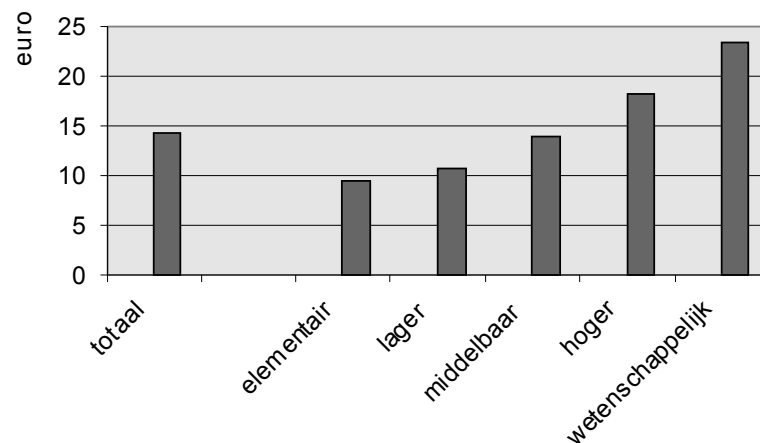
per uur meer. Bij de lagere beroepen en de middelbare beroepen ging het om een verschil van ongeveer 2,5 euro per uur. Bij de hogere en wetenschappelijke beroepen was het verschil meer dan 4,5 euro. Relatief gezien varieerde het verschil van 10 procent op het elementaire niveau tot 30 procent op het hogere beroepsniveau.

In het algemeen gaat een hogere leeftijd samen met een hoger beloningsniveau. Gemiddeld verdiende een 55-plusser per uur 2,4 keer zo veel als een werknemer die jonger is dan 25 jaar. Voor werknemers met een lager beroep was dit verhoudingsgetal met 1,7 het laagst. Bij middelbare en hogere beroepen was het verschil in beloning tussen ouderen en jongeren veel groter. De 55-plussers in de middelbare en hogere beroepen verdienden het dubbele van hun jongste collegae en de 55-plussers in de wetenschappelijke beroepen het drievoudige.

4. Slot

Het CBS stelt regelmatig gegevens samen over de verdiende lonen van werknemers in relatie met hun opleidings- en beroepsniveau op basis van het Loonstructuuronderzoek (LSO). Vroeger werd hiervoor een speciale enquête georganiseerd. Later werd door incidentele koppeling van verschillende bestanden een LSO-bestand samengesteld. Nu worden de gegevens samengesteld in het kader van het Sociaal Statistisch Bestand (SSB). Door gebruik te maken van bij het CBS al beschikbare gegevens zijn de kosten relatief beperkt en is er geen sprake meer van enquêtedruk.

Figuur 2. Uurloon van alle werknemers naar beroepsniveau, ultimo 1998



Door de koppeling van registers en steekproeven worden vier blokken gevormd, waaruit de LSO-tabellen worden geschat. Elke tabel moet berekend worden uit het grootste blok waaruit de tabel bepaald kan worden. Voor de samenstelling van het LSO98 worden de gewichten uit de steekproeven telkens gekalibreerd naar alle relevante registertellingen en de eerder geschatte tabellen uit de steekproeven zodat alle tabellen onderling consistent zijn. Als alle tabellen met de juiste gewichten worden geschat, zijn de tabeluitkomsten onderling consistent en zijn de uitkomsten onafhankelijk van de gebruikte bron. Door telkens uit te gaan van het grootst mogelijke blok kan de structuur van de lonen van werknemers nauwkeurig worden beschreven.

Noten

- (1) Eric Schulte Nordholt is behalve projectleider van de Volkstellingen ook nauw betrokken geweest bij het produceren van het Loonstructuuronderzoek vanuit het Sociaal Statistisch Bestand. esle@cbs.nl

Referenties

- Arts, C.H. en E.M.J. Hoogteijling, 2002, Het Sociaal Statistisch Bestand 1998 en 1999. In: *Sociaal-economische maandstatistiek*, Jaargang 19, december 2002, blz. 13-21

Boerdam, A.A., J.A. Loeve en G.P.C.M. Ruijs, 1998, Loon naar opleiding en beroep: het loonstructuuronderzoek 1995. In: *Sociaal-economische maandstatistiek*, Jaargang 15, maart 1998, blz. 31-57

Cruchten, J.M.J. van, M.I. Hartgers en E. Schulte Nordholt, 2002, Loon naar opleidings- en beroepsniveau: het Loonstructuuronderzoek 1998. In: *Sociaal-economische maandstatistiek*, jrg. 19, december 2002, blz. 72-81

CBS, 2000, Werkgelegenheid en vacatures. In: *Sociaal-economische maandstatistiek*, Jaargang 17, december 2000, blz. 37-94.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen en V. Snijders, 2003. Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Centraal Bureau voor de Statistiek, Voorburg / Heerlen. <http://www.cbs.nl/en/publications/articles/general/discussion-papers/discussion-paper-03005.pdf>

Schulte Nordholt, E. en G.P.C.M. Ruijs, 2000, Loon naar opleidingsniveau en beroep: het Loonstructuuronderzoek 1997. In: *Sociaal-economische maandstatistiek*, Jaargang 17, april 2000, blz. 19-49.

LSO is een mooi bestand, maar...

Een reactie op het paper van E. Schulte Nordholt

J. Hartog (1)

Hartog vond het erg lastig om een reactie te geven op het paper van Eric Schulte Nordholt. Het was alsof hij gevraagd was commentaar op het weer te geven. Het is er! En wat dan? Vandaar dat hij voor zijn reactie de insteek gekozen heeft, hoe hij als gebruiker tegen het Loon Structuur Onderzoek (LSO) aan kijkt.

Hij vindt het LSO een interessante bron voor onderzoekers. Pluspunten zijn dat het een groot bestand is met veel betrouwbare waarnemingen over inkomen en lonen. Toch constateert Hartog ook een aantal minpunten. Inkomen

en lonen zijn dan wel betrouwbaar weergegeven in het LSO, maar geldt dat ook voor de variabelen opleiding en etniciteit? Het mag dan wel een groot bestand zijn, het LSO bevat niet zo heel erg veel variabelen. Verder zit er geen systematiek in de cao-variabele. Er kan niet bepaald worden of iemand bij een monopolist meer verdient dan in een situatie van volkomen concurrentie. Verder vraagt Hartog zich af welke consequenties het herhaald wegeen heeft voor de consistentie van de gegevens of met andere woorden hoe zit het met de vergelijkbaarheid van LSO-en uit voorgaande jaren; zijn er als gevolg van de koppeling van bestanden nog wel vergelijkingen te maken?

Wat is er mogelijk met het LSO? Om dat te kunnen beoordelen heeft Hartog voor het LSO '98 onderzocht in hoeverre loonverschillen veroorzaakt worden door de

Tabel 1. In hoeverre reageren lonen op verschillen in risico?

	Man	Vrouw	Autochtoon	Allochtoon
Duur opleiding	0,041	0,056	0,048	0,043
Leeftijd	0,066	0,034	0,051	0,026
Leeftijdkwadr.	-0,0006	-0,0003	-0,0004	-0,0004
Etniciteit	-0,089	-0,035	-	-
Geslacht	-	-	-0,215	-0,172
Variantie	2,374	1,474	2,045	2,101
Scheefheid	-0,752	-0,495	-0,632	-0,662
R ²	0,35	0,23	0,36	0,31

mate van risico die verbonden is aan het rendement van human capital: in hoeverre reageren lonen op verschillen in risico? Uit zijn onderzoek blijkt dat als iemand een opleiding volgt waar een hoger risico aan kleeft, dan is er kans op een hoger loon. Daarbij is risico gemeten als de residuele variantie van de inkomens voor individuen met dezelfde opleiding; bovendien is de scheefheid toegevoegd (het derde moment), omdat de theorie voorspelt dat individuen waardering hebben voor positieve scheefheid (kleine kans op hoog inkomen) en derhalve een lager verwacht inkomen accepteren. De residuele variantie en scheefheid komen uit een regressie van inkomen op opleidingsduur, leeftijd en leeftijdkwadraat. De regressie levert in alle gevallen het verwachte teken: positief voor risico, negatief voor scheefheid (zie tabel 1).

Noten

- (1) Joop Hartog is hoogleraar economie aan de UvA (FEE) en verricht onderzoek op het terrein van arbeidsmarkt en onderwijs. De regressieresultaten zijn ontleend aan gezamenlijk werk met Dinand Webbink (CPB), die ook de berekeningen heeft verricht. Momenteel houdt Hartog zich ook bezig met onderzoek naar de economische aspecten van migratie.

Ruimtelijke spreiding van mensen en hun kenmerken

Zicht op autochtonen en allochtonen

J. Latten (1)

1. Inleiding

Ruimtelijke concentratie of spreiding blijkt een algemeen verschijnsel. Personen met bepaalde posities in de maatschappelijke structuur kunnen relatief vaker voorkomen in het ene gebied dan het andere. Sociale ongelijkheid vindt zijn weerslag in ruimtelijke ongelijkheid. In andere woorden: ruimtelijke structuur en sociale structuur hangen met elkaar samen.

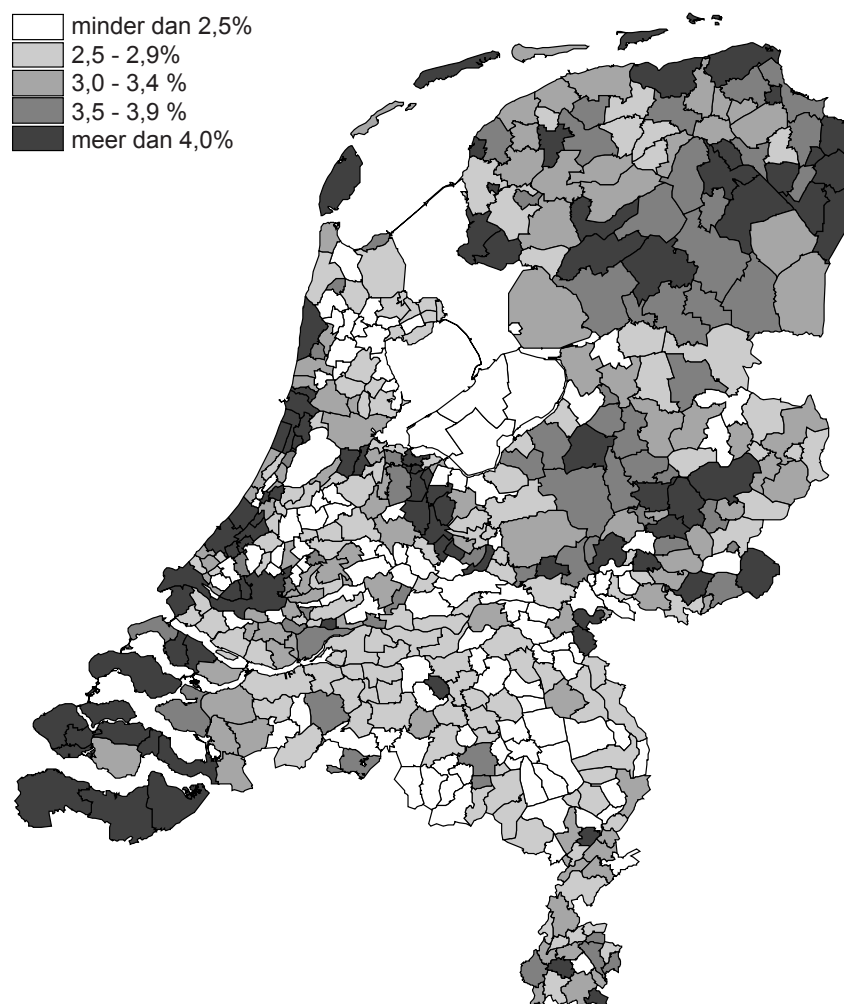
De actuele discussie over integratie gaat voor een deel over de betekenis van ruimtelijke spreiding voor sociale integratie, zo men wil over maatschappelijke uitsluiting. Via het Sociaal Statistisch Bestand (SSB) kan op zijn minst een antwoord worden gegeven op de vraag in welke mate ruimtelijke spreiding van mensen en hun kenmerken zich tot op laag regionaal niveau (gemeente, wijk, buurt, postcode) voordoet. Het SSB biedt immers een combinatieperspectief voor demografische en een groot aantal sociaal-economische persoonskenmerken. Tot die persoonskenmerken behoren ook ruimtelijke locatiekenmerken tot op zeer laag niveau. Zo kan bij voorbeeld voor allochtone bevolkingsgroepen op gemeente- of buurniveau de statistische samenhang tussen ruimtelijke spreiding en sociaal economische kenmerken worden weergegeven. Voor een daadwerkelijke inhoudelijke verklaring van

samenhang tussen ruimtelijke en sociale structuur is echter het levensloopperspectief van belang, in het bijzonder de ruimtelijke levensloop in de zin van verhuizingen. Dit introduceert het dynamische aspect in de samenhang tussen ruimtelijke en sociale structuur. Ook daarvoor zal het SSB mogelijkheden bieden. Het SSB kan op den duur zelfs de mogelijkheid bieden dit dynamisch perspectief uit te breiden naar intergenerationele ruimtelijke en sociale mobiliteit.

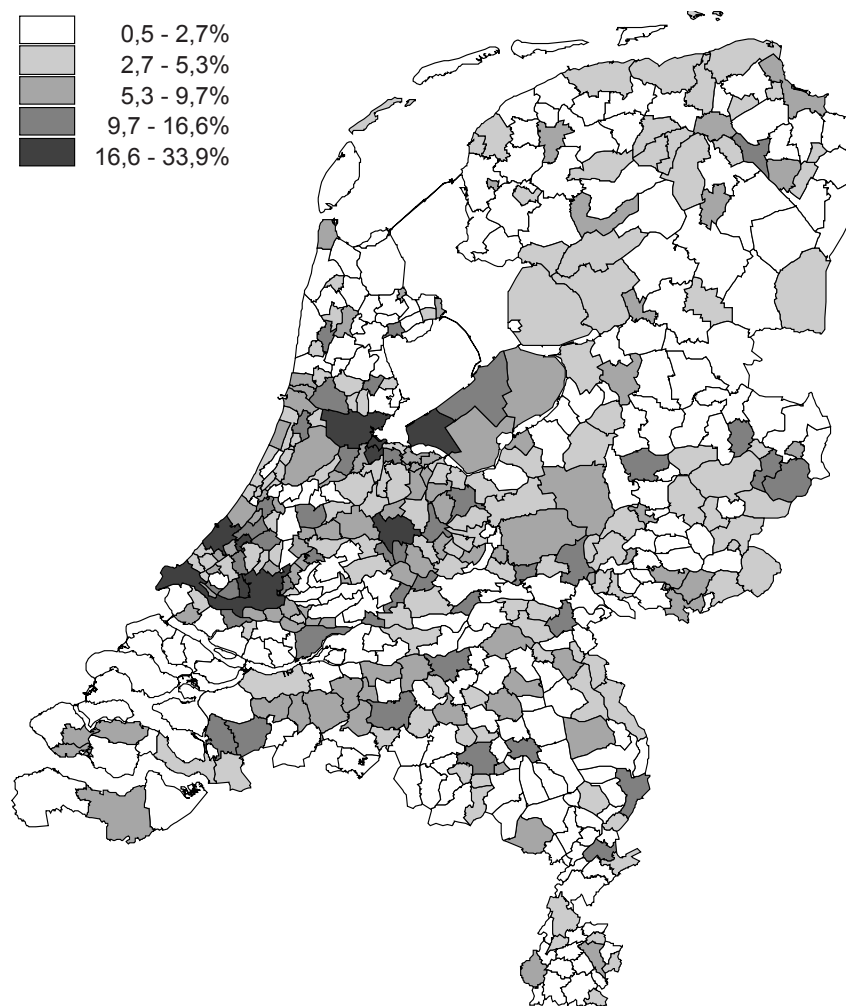
2. Ruimtelijke spreiding als empirische realiteit

Nederland staat bekend om een relatief hoge bevolkingsdichtheid (in 2003: 479 inwoners per vierkante kilometer land). In feite is zo'n gemiddelde een formele weergave van een realiteit die op lager ruimtelijk niveau nauwelijks herkenbaar is. Gedifferentieerd naar gemeenten is de realiteit heel anders. De gemeente Den Haag heeft bijvoorbeeld een bevolkingsdichtheid van 5610 inwoners per vierkante kilometer. De gemeente Rozendaal in Gelderland komt op 55 inwoners per vierkante kilometer. In andere woorden: de dagelijkse realiteit laat ongelijke ruimtelijke spreiding zien van inwoners van Nederland. Dat geldt niet alleen voor het totale inwonertal maar bijvoorbeeld ook voor subcategorieën als 80-plussers (figuur 1), voor de spreiding van niet-westerse allochtonen (figuur 2) of alleenstaande 30-34-jarigen (figuur 3), de leeftijd waarop meestal gezinsvorming plaatsvindt (Latten, 2002).

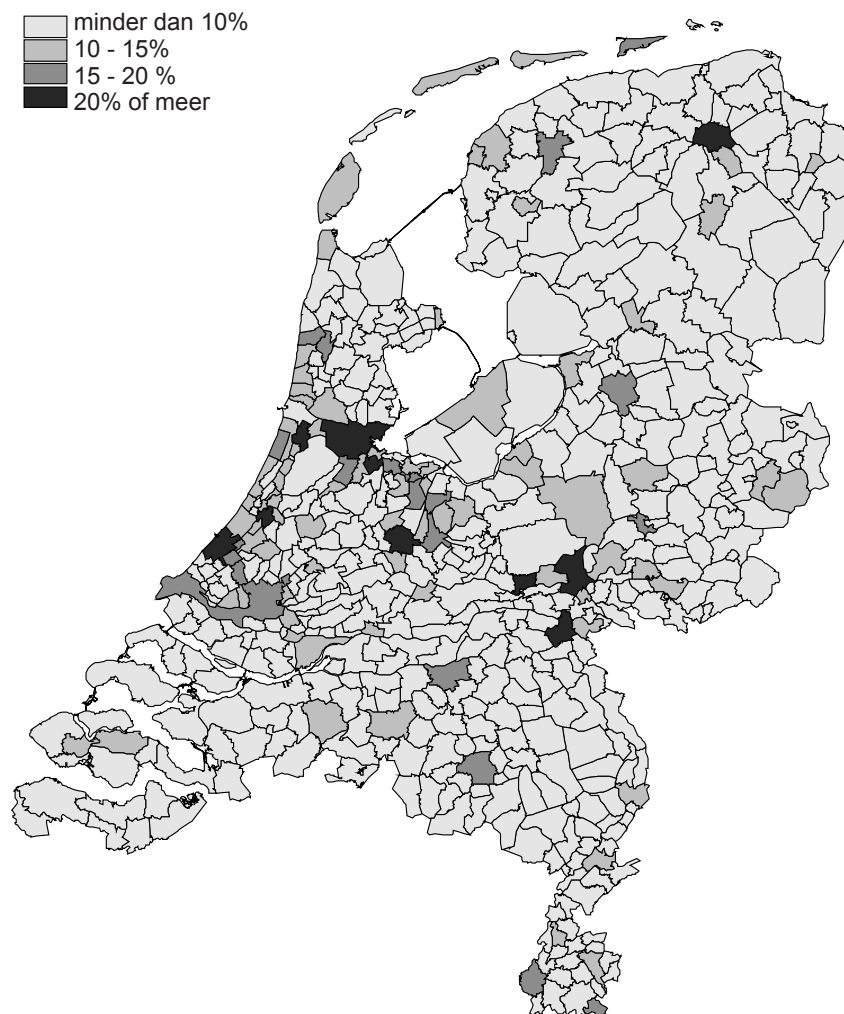
Figuur 1. Gemeenten naar percentage 80-plussers, 1 januari 2003



Figuur 2. Gemeenten naar percentage niet-westerse allochtonen, 1 januari 2003



Figuur 3. Gemeenten naar percentage alleenstaande vrouwen, 30-34-jaar, 1 januari 2000



Wat het laatstgenoemde voorbeeld betreft: gemiddeld is in Nederland twaalf procent van de 30-34-jarige vrouwen alleenstaand. Maar in de grotere steden is het een veelvoud. Zo woont in Amsterdam bijna een op de drie vrouwen van 30-34-jaar alleen. Ook de stad Groningen, Utrecht en Wageningen scoren hoog met meer dan een kwart. En er zijn gebieden, gemeenten of wijken waar relatief rijkere mensen wonen, gebieden waar relatief veel grote gezinnen zijn of waar men gemiddeld langer leeft. Zelfs informele concentraties zijn algemeen voorkomend. Zo kent bijvoorbeeld het Scheveningse strand een eigen informele gebruikersdifferentiatie. Het Noorderstrand is er overwegend voor hippe loungers en de havenhoofden fungeren als flaneergebied voor allochtonen.

Blijkbaar bestaat de demografische en sociale opbouw van Nederland uit een stapeling van ongelijke ruimtelijke concentraties van mensen en hun specifieke kenmerken. Hoewel een ruimtelijk ongelijke spreiding zo algemeen voorkomt, leidt het zelden tot maatschappelijk debat. Anders is het recent bij de discussie rond spreiding over het land van niet-westerse allochtonen. Zwarte scholen en verkleurende steden dragen het karakter van segregatie. Met name omdat ongelijke ruimtelijke spreiding dan samenhangt met cumulatie van achterstanden. Ruimtelijk ongelijke spreiding van 80-plussers of van 30-34-jarige alleenstaanden zou evenzo goed tot een fenomeen van segregatie kunnen worden bestempeld zodra de ruimtelijke spreiding een pendant heeft in andere aspecten van sociale ongelijkheid. Informatie over de mate van spreiding tot op laag regionaal niveau wat betreft diverse aspecten zoals wonen, inkomen of arbeidsparticipatie is in dergelijke situaties urgent. Om

aan die maatschappelijke behoefte aan informatie te voldoen beschikt het CBS thans in het SSB over een databron die voor alle inwoners van Nederland diverse demografische en sociaal-economische kenmerken – waaronder herkomst - combineert. Zo kan via het SSB bijvoorbeeld voor diverse bevolkingscategorieën informatie over inkomensbronnen (werk of uitkering) tot op wijk- en buurtniveau worden gegeven.

Als voorbereiding op de ontwikkeling van het SSB heeft het CBS eerder demografische gegevens uit de Gemeentelijke BasisAdministratie (GBA) gekoppeld aan een zeer grote steekproef van belastinggegevens. Daarmee kon het CBS voor het eerst integraal zicht geven op inkomens in wijken en buurten. Wat leverde die koppeling inhoudelijk op aan informatie over grote steden en wijken?

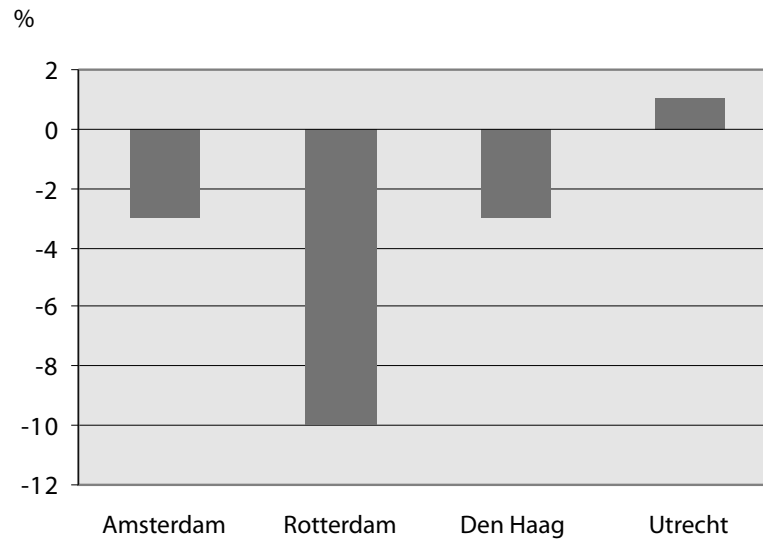
3. Inkomenshoogte en ruimtelijke spreiding

Los van de landelijke spreiding ligt het gestandaardiseerd huishoudinkomen in de drie grootste steden onder het landelijk gemiddelde van 18 duizend euro (Kasperski, 2003).

Het gestandaardiseerd huishoudinkomen in Rotterdam ligt zelfs 10 procent onder het landelijk gemiddelde. In de gemeente Den Haag is het gemiddelde minder ongunstig (zie figuur 4).

Maar door te differentiëren naar wijkniveau wordt een sterke differentiatie binnen de stad zelf zichtbaar. De Schilderswijk scoorde het laagst. Het gestandaardiseerde

Figuur 4. Afwijking gestandaardiseerd inkomen t.o.v. landelijk gemiddelde, 2000



Bron: CBS

inkomen lag daar 28 procent onder het landelijk gemiddelde. Het Benoordenhout daarentegen scoort opvallend hoog: 63 procent boven het landelijk gemiddelde. De verschillen in de stad tussen de wijken, wat de spreiding van inkomen betreft, zijn blijkbaar enorm. Ze hebben inzake ruimtelijke spreiding van ongelijkheid meer empirische realiteit dan een stedelijk gemiddelde. De laagregionale differentiatie helpt in het geval van de gemeente Den Haag ruimtelijk zichtbaar te maken dat de stad, net als in de 19e eeuw, nog steeds een stad van 'pruiken en petten' is.

Dat ruimtelijke spreiding van ongelijkheid actuele relevantie heeft en geen gepasseerd maatschappelijk fenomeen is, blijkt ook uit een ander CBS-onderzoek op basis van 'kleinschalige' ruimtelijke grids van 500 bij 500 meter (Trimp, 2001). Dit onderzoek liet zien dat ruimtelijke inkomenssegregatie tussen 1994 en 1998 is toegenomen. Lage inkomens zijn dichter bij elkaar komen te wonen (2). In 1998 zouden bijna drie van de tien huishoudens met een laag inkomen moeten verhuizen om een gelijkmatige spreiding van lage inkomens over het land te verkrijgen. In 1994 was dat nog een kwart.

4. Ruimtelijke spreiding en valkuilen

Statistische informatie op laagregionaal niveau als wijk, buurt, postcode of grid, biedt veel perspectief maar ook nieuwe vragen. Statistische samenhang op laagregionaal niveau is niet zonder meer te duiden. Bijvoorbeeld als het gaat om de samenhang tussen inkomens en inkomensbronnen. Door combinatie van bestanden, zoals die in het SSB is gerealiseerd, is namelijk van iedere inwoner van Nederland de inkomensbron vast te leggen. Daarom is het mogelijk om wijken te beschrijven naar zowel gemiddeld inkomen als naar omvang van bepaalde inkomensbronnen. Daaruit blijkt bijvoorbeeld dat de Schilderswijk in Den Haag niet alleen het laagste gestandaardiseerde inkomen in de stad heeft maar met 35 procent ook het hoogste percentage uitkeringsafhankelijken. Dit in scherp contrast met de wijk Benoordenhout, waar 3 procent een uitkering ontving en het gemiddeld gestandaardiseerd inkomen het hoogst was in de stad. Statistisch is er samenhang. Hoe is het beeld landelijk? En wat kan dat betekenen?

Voorbeeld: wat betreft de inkomensbronnen volgt uit de analyse van Linder (2002) op SSB data dat gemiddeld 46,6% van alle inwoners van Nederland (alle leeftijden) werkzaam is. Zonder te standaardiseren naar bevolkingsopbouw varieert op provincieniveau het percentage werkzamen van 43,6% in Groningen en Friesland tot 49,3% in Utrecht. In de provincie Utrecht waren in 1999 relatief de meeste inwoners werkzaam. Bijna de helft had een baan als werknemer of zelfstandige.

Op gemeenteniveau is de differentiatie groter, is de polarisatie sterker (tabel 1). Bovenaan staat Nieuwegein. In Nieuwegein werkt meer dan de helft van alle inwoners. In Vaals (3), maar ook in Kerkrade, Rozendaal, Laren en Pekela bijvoorbeeld, zijn percentages werkenden te zien die zich het verst onder het landelijk gemiddelde bevinden.

Pekela en Kerkrade bevinden zich anderzijds in de top tien van gemeenten met hoogste percentages personen met een uitkering (anders dan pensioen). In Kerkrade en Pekela gaan zeer lage percentages werkenden en zeer hoge percentages uitkeringen statistisch gezien samen. In dergelijke situaties lijkt een inhoudelijke interpretatie, de oorzaak van statistische samenhang voor de hand liggend, maar is toch niet terecht, omdat het in feite een correlatie op mesoniveau betreft en niet op persoonsniveau.

Er kan namelijk ook sprake zijn van andere inhoudelijke samenhang. Het geringe aantal werkzamen kan bijvoorbeeld duiden op vergrijzing of welvaart. Dat zou kunnen worden geïndiceerd door het feit dat Laren en Rozendaal weliswaar ook weinig werkenden hebben maar zich niet bevinden in de groep gemeenten met hoge percentages uitkeringen!

Om foutieve interpretaties te voorkomen moet in feite de inkomenshoogte en inkomensbron op persoonsniveau worden geanalyseerd. Dat is hier niet gebeurd, maar het SSB biedt daartoe juist de mogelijkheden. Ik kom daar later op terug middels het concept verhuizingen, maar los daarvan wordt hierna ingegaan op het beeld van regionale samenhang tussen inkomensbron van niet-westerse allochtonen en wijkniveau.

Tabel 1. Rangorde van gemeenten naar werkzame personen en personen met uitkering

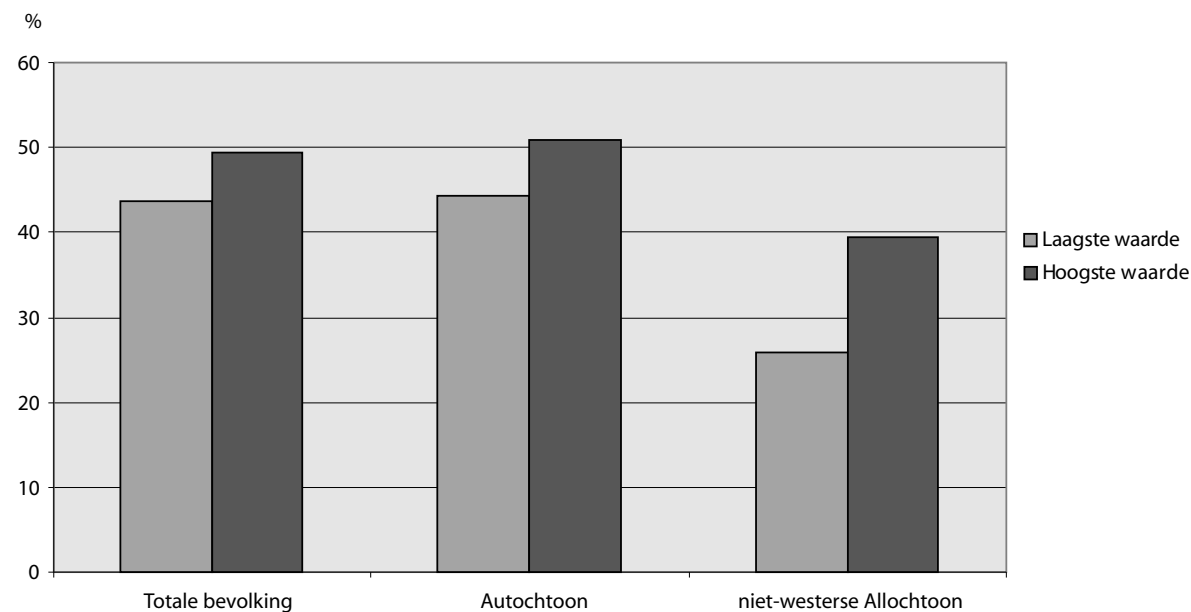
Gemeenten in rangorde	Werkzame personen			Personen met uitkering	
1	Nieuwegein	55,4	1	Amsterdam	17,7
2			2	Heerlen	17,0
3			3	Rotterdam	16,7
4			4	Reiderland	16,3
5			5	Gennep	16,2
6			6	Pekela	16,1
7			7	Kerkrade	15,8
8			8	Sittard	15,7
9			9	Arnhem	15,5
10			10	Hoogezand-Sappemeer	15,4
.	NEDERLAND	46,6	.	.	.
529	Dantumadeel	39,3			
530	Pekela	39,1			
531	Laren	38,6			
532	Bellingwedde	38,1			
533	Wassenaar	38,0			
534	Winschoten	37,9			
535	Reiderland	37,2			
536	Rozendaal	36,2			
537	Kerkrade	34,5			
538	Vaals	26,0			

5. Inkomensbron, etniciteit en ruimtelijke spreiding

Zoals hiervoor vermeld was het aandeel werkzame personen in de bevolking in 1999 met 49,3 procent het hoogst in de provincie Utrecht en met 43,6 procent het laagst in de noordelijke provincies Groningen en Friesland. Uitsplitsing naar etniciteit laat de volgende differentiatie zien.

Voor autochtonen is er een variatie van 44,3 tot 50,6 procent. Voor niet-westerse allochtonen variëren de percentages van 25,8 tot 39,4. Niet alleen is de variatie groter, ook is het niveau van het percentage werkenden onder niet-westers allochtonen in alle provincies lager. Verder blijkt dat in die provincies waar het aandeel werkenden onder de autochtonen hoger is, veelal ook het percentage werkzame niet-westerse allochtonen hoger

Figuur 5. Limietwaarden werkenden per provincie

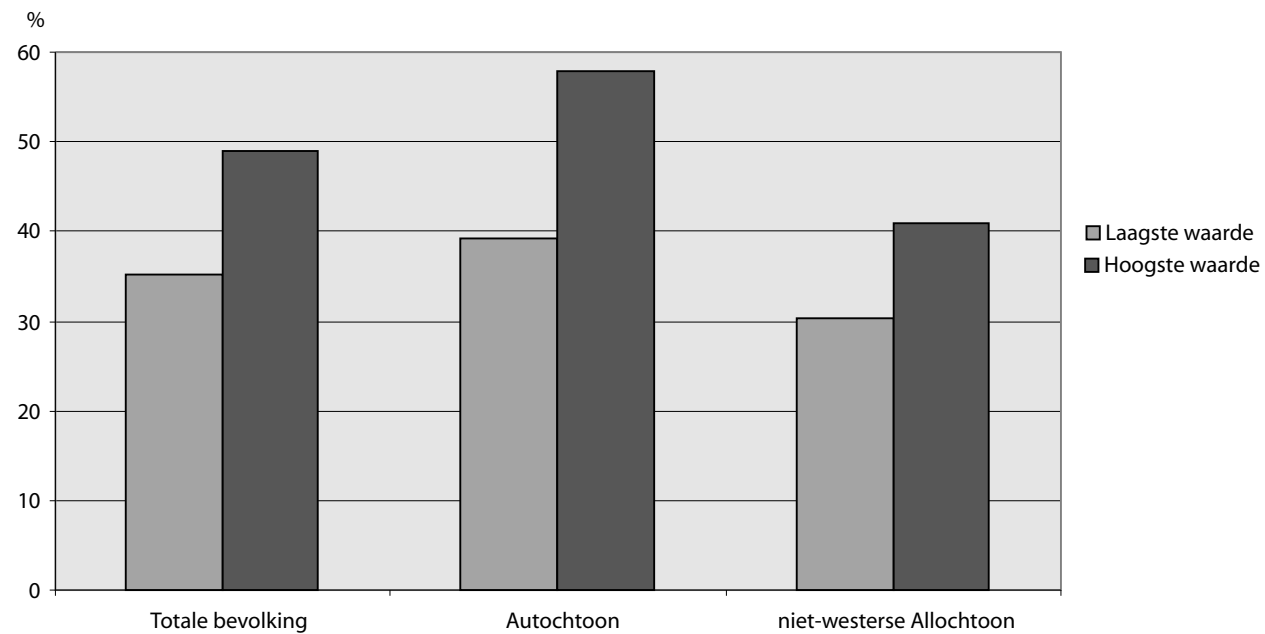


is (Linder, p. 44). De suggestie dringt zich dan op of er wellicht, statistisch, een regio-effect is?

Laag-regionaal gezien, op wijkniveau, is de differentiatie sterker. Binnen Rotterdam komt op wijkniveau namelijk

veel spreiding in het percentage werkenden voor (zie figuur 6). Differentiatie naar herkomst laat zien dat in alle wijken niet-westerse allochtonen vaker niet werkzaam zijn en het percentage werkende niet-westerse allochtonen in één wijk (Prins Alexander) dat voor de autochtonen benadert (40%).

Figuur 6. Limietwaarden werkenden per wijk Rotterdam



Zowel de geconstateerde samenhang op provincieniveau als die op wijkniveau suggereert een voordeel bij fysieke spreiding. Maar, voor de juiste interpretatie zal een analyse van diverse kenmerken op persoonsniveau noodzakelijk zijn. Dat kan nader worden toegelicht met een ander voorbeeld, namelijk over het regio-effect dat wordt waargenomen bij zuigelingensterfte.

6. Zuigelingensterfte onder allochtonen. Buurt of selectieve migratie?

Vooruitlopend op beschikbaarheid van inkomensgegevens op persoonsniveau is in een multilevel-analyse geëxperimenteerd met toevoeging van een buurtkenmerk aan de persoonskenmerken in een analyse omtrent zuigelingensterfte (van Duin, 2002).

Per jaar overlijden in Nederland gemiddeld 4,9 op de duizend kinderen vóór hun eerste verjaardag. Dit komt neer op ongeveer duizend zuigelingen per jaar. Onder niet-westerse allochtonen sterven 6,4 op de duizend levendgeboren kinderen binnen het eerste levensjaar. Onder personen met een westerse (inclusief autochtone) herkomst bedraagt deze sterfte 4,9 per duizend. De zuigelingensterfte onder niet-westerse allochtonen is dus zo'n 30 procent hoger. Door dit verschil overlijden jaarlijks veertig zuigelingen meer onder niet-westerse allochtonen.

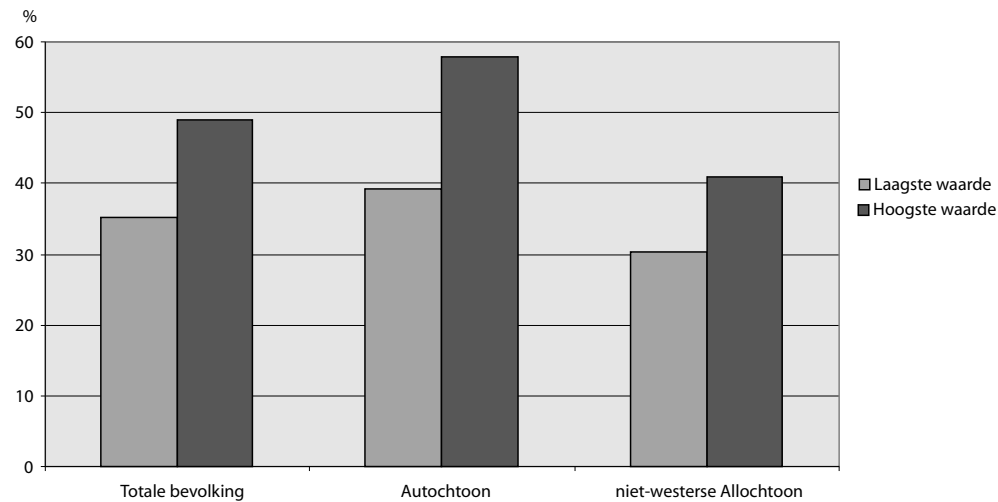
In zijn analyse naar zuigelingensterfte is van Duin nagegaan in welke mate hogere zuigelingensterfte onder allochtonen te maken heeft met het land van herkomst

van de moeder en in welke mate met sociaal-economische kenmerken van de buurt waarin de moeder woont. Dat laatste is bepaald op basis van de gemiddelde huiswaarde en het gemiddelde huishoudensinkomen. Daarmee werd de variabele Sociaal-economische Status (SES) als indicator voor de status van de woonbuurt geconstrueerd. Voor de variabele SES is een indeling naar twee groepen gebruikt, lage en hoge status. Uit analyse kwam naar voren dat zowel de SES (als indicator voor de status van de woonbuurt) en het kenmerk etniciteit van de moeder een rol spelen in de verklaring van hogere zuigelingensterfte.

Bij de invloed van etniciteit kan men denken aan taalachterstanden die de communicatie met de gezondheidsdiensten of de huisarts kunnen belemmeren, de andere gezondheidssituatie van de populatie niet-westerse allochtonen of misschien vatbaarheid voor bepaalde ziekten. Maar ook trouwpatronen waarbij men trouwt met een ver familielid kunnen een rol spelen.

Zowel bij de lage als de hoge welvaartsbuurt wordt een significant hogere zuigelingensterfte gevonden onder de niet-westerse dan onder de westerse kinderen. Bij de niet-westerse moeders in de lage welvaartsbuurten blijkt de zuigelingensterfte het hoogst. Bij de autochtone en westerse moeders in de hoge welvaartsbuurten het laagst (zie figuur 7). Wanneer de allochtone moeder in een sociaal-economisch betere buurt woont is de zuigelingensterfte dus kleiner dan wanneer de allochtone moeder in een armere buurt woont. Een betere buurt hangt dus statistisch gezien samen met een lager sterfterisico voor kinderen van niet-westerse moeders. Overigens werd een overeenkomstig effect van buurt en sociaal-economische status geconstateerd voor de jaarlijkse

Figuur 7. Zuigelingensterfte per 1000 levendgeborenen in de periode 1995-2000 naar herkomstgroep per welvaartsgroep



sterftekans in het algemeen (Smits, Keij, & Westert, 2001). Er zijn in Nederland aanzienlijke verschillen in sterfte tussen personen die wonen in gebieden met een hoge sociaal-economische status en personen in gebieden met een lage sociaal-economische status. Een rol kan spelen dat risicofactoren als slechte behuizing, vervuiling en (verkeers)onveiligheid minder aanwezig zijn in buurten met een hoge sociaal-economische status. Maar ook hier komt de vraag op: vormen buurtkenmerken ook echt een verklaring? Een positief antwoord zou immers betekenen dat ruimtelijke spreiding van herkomstgroepen zeker is aan te bevelen, maar het zou best wel eens zo kunnen zijn dat het slechts een schijnverband is. Misschien is

het wel zo dat niet-westerse moeders die zich sterk op de Nederlandse samenleving oriënteren, maatschappelijk succesvoller zijn en ook andere trouwpatronen hebben. De overheersende neiging onder Turken en Marokkanen om een partner uit herkomstland te halen zou mede tot gevolg hebben dat er huwelijken worden gesloten tussen neven en nichten, niet zelden iemand uit hetzelfde dorp, waar al eeuwenlang onderling wordt getrouwd. Mede daardoor zou er via trouwpatronen een groter risico op zuigelingensterfte in Nederland kunnen ontstaan.

Als dat zo is kan dat betekenen dat juist allochtonen die zich op een ruimere huwelijksmarkt oriënteren minder

snel een kind zouden kunnen verliezen vanwege minder erfelijk belaste ziekten. Misschien valt een dergelijke oriëntatie op de Nederlandse huwelijksmarkt ook wel samen met gunstige sociaal-economische kenmerken. Gunstige sociaal-economische kenmerken op hun beurt leiden wellicht weer tot bijvoorbeeld een grotere kans te gaan wonen in een sociaal-economisch gunstige buurt. En zo ontkomt beter begrip van de werkelijkheid niet aan de noodzaak analyses uit te voeren op microniveau naar een combinatie van dimensies zoals sociaal-economische status, inkomensbron of woonsituatie maar zeker ook naar de individuele verhuisgeschiedenis als aspect van de levensloop (4). Daarmee zou zelfs vaker duidelijk worden dat ruimtelijke spreiding of segregatie mede gevolg is van selectieve migratie. Zo gezien zijn ruimtelijke sociale structuren mede het resultaat van individuele levenslopen.

7. Nieuwe data, nieuwe vragen

Het voorgaande verwijst naar het belangrijk aandachtspunt: de mogelijkheid die het SSB biedt om levensloopbenaderingen te kiezen waarbij diverse thema's worden geïntegreerd en longitudinaal worden geanalyseerd. Dat geldt voor de sociaal-economische posities die worden ingenomen alsook voor de posities in de ruimtelijke structuur. Nieuwe bestanden van het CBS en nieuwe koppelingsmogelijkheden zullen de keuze voor zo'n analyseperspectief vereenvoudigen. Op dit punt heeft de ontwikkeling van bestanden de aandacht.

Los van de exacte inhoud van het actuele SSB en los van gewenste ontwikkelingen zijn er voorbeelden te geven van onderzoeksdesigns die rekening houden met de dynamische aspecten van de levensloop zoals de individuele verhuisgeschiedenis of de sociaal-economische levensloop. De twee eerstgenoemde onderzoeken worden reeds door het CBS in samenwerking met derden uitgevoerd.

7.1 Sociaal-economische dynamiek van vier grote steden.

In een samenwerking met de afdeling Geografie en Planologie van de Universiteit van Amsterdam worden de demografische en sociaal-economische kenmerken van de in- en uitstromers van grote steden tussen twee meetmomenten in kaart gebracht.

Daarmee is er ook een beeld te schetsen van stedelijke verloop en de duur van verblijf in de stad: welke groepen nieuwkomers zijn in bepaalde onderzoeksperiode in de stad gebleven en wie zijn vertrokken, en hoe lang is men gebleven? Wat zijn de achtergrondkenmerken van 'blijvers' en '*movers*'? In welke mate is het bijvoorbeeld zo dat bepaalde migratietrajecten worden gekenmerkt door bepaalde inkomenstrajecten. Ofwel: zullen degenen die sneller in inkomen stijgen bijvoorbeeld ook sneller een wijk of de stad weer verlaten en zo bijdragen aan inkomenssegregatie?

En hoe kan de actuele sociaal-economische en demografische structuur (verarmd, verrijkt, verjongd etc.) van een gemeente worden verklaard met behulp van ruimtelijke verplaatsingen van subcategorieën in de bevolking, waaronder ook categorieën als allochtonen en autochtonen?

7.2 Sociaal-economische dynamiek van immigranten

In het kader van het strategisch programma 'sociale dynamiek en arbeidsmarkt' wordt door het CBS in samenwerking met de Afdeling Algemene Economie van de Universiteit van Amsterdam onderzoek gedaan naar de sociaal-economische dynamiek van immigranten. Door koppeling van bestanden met achtergronden over migratiemotieven en inkomensbestanden over een langere periode kan een analyse worden uitgevoerd naar veranderingen in economische status (werkzaam, uitkering enz.) en naar inkomen van verschillende categorieën internationale immigranten (zoals volgmigranten, arbeidsmigranten en asielmigranten).

7.3 Overdrachten tussen generaties

In theorie kunnen met het SSB ook sommige vormen van intergenerationele overdracht in kaart worden gebracht. Als er immers voor alle personen in Nederland die in de GBA zijn opgenomen overeenkomstige informatie beschikbaar is, dan zijn in het bestand ook zonen en vaders, moeders en dochters te identificeren. Via de GBA zijn in principe familierechtelijke banden te herkennen. Op die wijze kan het SSB voor onderzoek tussen generaties dienen. Een dergelijke benadering wordt thans uitgewerkt in een analyse naar intergenerationele sterfte waarmee achterhaald kan worden in welke mate een beperkte levensverwachting van ouders zich herhaalt bij kinderen. Intergenerationeel gedrag zou ook kunnen worden onderzocht in termen van ruimtelijke spreiding van generaties.

7.4 Tenslotte: zicht op allochtonen

Sociaal-economische dynamiek van vier grote steden, sociaal-economische dynamiek van immigranten en overdrachten tussen generaties vormen onderzoeksonderwerpen die een belangrijke samenhang zullen hebben met subcategorieën in de bevolking als allochtonen. Dergelijke analyses zullen mede het zicht op de allochtone bevolking verhelderen. Daarnaast zal in de toekomst blijken dat de allochtone bevolkingsgroep divers is wat betreft sociaal-economische en demografische kenmerken. Ook de levenslopen van allochtonen zullen diverser worden. De omvang van de categorieën in het SSB zal het mogelijk maken dat met deze diversiteit rekening kan worden gehouden. Analyses zullen zich kunnen richten op subgroepen binnen de allochtone bevolking. Het zal niet alleen maar gaan om Turken, maar om Turken met succesvolle maatschappelijke en ruimtelijke carrières versus Turken met minder succesvolle carrières, over hun verhuisgeschiedenis en hun sociale mobiliteit, over kansen en het gebruik maken van kansen kortom.

Noten

- (1) Jan Latten is demograaf en de woordvoerder voor de pers van het CBS. jltm@cbs.nl
- (2) De ruimtelijke afzondering van huishoudens met een laag inkomen is gemeten met de segregatie-index. Voor het bepalen daarvan is Nederland verdeeld in vierkanten van 500 bij 500 meter. De index wordt berekend door de verdeling van de huishoudens met een laag inkomen over de vierkanten te vergelijken met die van de overige huishoudens. Als de huishoudens met een laag inkomen en de overige huishoudens op dezelfde manier verdeeld zijn over de vierkanten, is de segregatie-index gelijk aan 0. Zijn alle huishoudens met een laag inkomen geconcentreerd in vierkanten waar geen andere huishoudens voorkomen, dan is de segregatie-index gelijk aan 100.
- (3) In de Limburgse grensgemeente Vaals was het aandeel werkzame personen in de bevolking extreem laag: 26% tegenover 46,6% landelijk. In deze grensgemeente wonen veel Duitsers. In 1999 was het percentage westerse allochtonen in Vaals dan ook 50 procent. Gegevens over de inkomensbronnen van deze personen ontbreken meestal in de Nederlandse administraties. Een deel van de Duitse populatie in Vaals werkte vermoedelijk over de grens in Duitsland en wordt daarmee aan het zicht onttrokken van de Nederlandse werknemersadministraties.
- (4) Uiteraard is de aandacht voor levensloopbenadering, zoals hiervoor beschreven, in een veranderende empirie geworteld. Maar om demografisch gedrag te begrijpen, en op juiste wijze te beschrijven, is het noodzakelijk om de individuele demografische transitie in hun onderlinge chronologische samenhang te beschouwen. Daarmee is de aandacht theorie-gestuurd, maar niet minder is de aandacht ook data-gestuurd. Door verruiming van technologische mogelijkheden voor de bouw van databestanden alsook analysemethoden kunnen opeenvolgende gebeurtenissen van individuele levenslopen immers steeds beter worden opgeslagen en verwerkt. Ruimere databestanden (geïntegreerde, gekoppelde databestanden, levensloopdata) in combinatie met nieuwe analysetechnieken vormen een uitdaging in de sociaal-demografische analyse. Het Sociaal

Statistisch Bestand zal nieuwe geïntegreerde informatie bieden waarmee de kennis over de empirische realiteit in de pas blijft met de toenemende dynamiek daarin.

Referenties

- Duin, C. van, 2002, Hogere zuigelingensterfte in minder welvarende gebieden en onder niet-westerse allochtonen in Nederland. CBS, Maandstatistiek van de Bevolking, nr. 3, p. 4-6.
- Kasperski, J.M.J., 2003, Inkomens in de wijken van de grote steden, CBS Webmagazine, 3 februari.
- Latten, J. J., 2002, Op zoek naar alleenstaande dertigers, CBS Webmagazine, 8 april.
- Linder F.S., 2002, Bronnen van inkomen in de regio, in : CBS, Sociaal-Economische Maandstatistiek, nr. 12, p. 42-52.
- Smits J., Keij I., Westert G., 2001, Effecten van sociaal-economische status van kleine, middelgrote en grote geografische eenheden op de sterfte. CBS, Maandstatistiek van de Bevolking, nr. 11, p. 4-10.
- Trimp R., 2001, Lage inkomens dicht bij elkaar, CBS Webmagazine, 15 januari.

Nog geen zicht op allochtonen

Een reactie op het paper van J. Latten

P. Tesser (1)

In het najaar van 1993, toen het SCP besloot de Rapportage minderheden 1994 te wijden aan de ruimtelijke concentratie van allochtonen, werd al gesproken over het SSB. Voor Tesser had het SSB toen iets van een fata morgana. “Aanlokkelijk, maar steeds als je er bijna bij dacht te zijn, lag het toch weer een eind verderop. Bijna zoiets als het onderwijsnummer.”

Ondanks het feit dat het SSB destijds nog niet beschikbaar was, kon het SCP in de Rapportage wel uitspraken doen over de ruimtelijke concentratie van allochtonen, omdat het de beschikking had over een survey met gegevens over allochtone huishoudens: de SPVA. Van de SPVA-respondenten was de postcode bekend en met die postcodes kon het SCP via bestanden als de woonmilieudatabank en het Geo-marktprofiel weer aan aanvullende informatie komen. Er kon nog geen variantiecomponentenanalyse op buurt- en individueel niveau mee uitgevoerd worden, maar er konden wel samenhangen tussen bijvoorbeeld het aandeel allochtonen in een buurt en het werkloosheidsrisico van allochtone en autochtone buurtbewoners berekend worden. Met de Politie-monitor bevolking was het overigens wel mogelijk om een multilevelanalyse uit te voeren. En daarmee wil Tesser maar aangeven dat het SCP ook zonder het SSB

op verschillende niveaus al aardig wat zicht had op allochtonen. Hij vindt het dan ook een beetje wonderlijk dat Latten in zijn paper suggereert dat met het SSB voor het eerst mogelijk is te analyseren wat de concentratie van allochtonen in grote stadswijken is: wat kan het SSB beter dan het SCP tien jaar geleden, toen het de segregatie van minderheden analyseerde?

Latten licht in zijn presentatie een tipje van de sluier op. Wat daaronder te zien valt, is vanuit het gezichtspunt van de onderzoeker die tuk is op uitbreiding van de toepassingsmogelijkheden die de hedendaagse analytische trucendoos biedt niet te versmaden. Tesser zou graag zien dat de (wetenschappelijke) buitenwereld op grote schaal gebruik kan maken van het SSB. Hij ondersteunt ook het pleidooi van Dronkers voor een optimale toegankelijkheid van een steekproef uit het SSB voor de hele onderzoekswereld. Maar het SSB moet natuurlijk meer zijn dan een speeltuin voor multileveljunks, aldus Tesser.

Een jaar geleden is hij van het onderzoek bij het SCP overgestapt naar het beleid van het ministerie van Justitie. Het onderwerp - de integratie van etnische minderheden - is hetzelfde gebleven. Binnen het ministerie functioneert hij een beetje als de go-between tussen onderzoek en beleid. Het is zijn taak om bevindingen en inzichten uit onderzoek om te zetten in een effectieve beleidsstrategie. Vandaar dat de titel van het paper van Latten zijn niet geringe nieuwsgierigheid had opgewekt: Zicht op allochtonen. “Als strateeg van het integratiebeleid kun je daar niet gauw te veel van hebben.”

Na lezing van Lattens bijdrage is Tesser toch een beetje teleurgesteld over het geboden zicht op de spreiding van allochtonen. Latten begint met iets te vertellen over de relevantie van ruimtelijke spreiding voor de beschrijving van de sociale werkelijkheid. Dat is voor Tesser niet echt een eye-opener meer. Vervolgens laat Latten wat geaggregeerde inkomensgegevens zien. Dat is mooi, vooral omdat op die manier de verschijnselen geconcretiseerd kunnen worden, een naam krijgen. De excentrische positie van Rotterdam wordt duidelijk. Amsterdam is met Rotterdam te vergelijken. Het gemiddelde inkomen in de Schilderswijk is af te zetten tegen het landelijk gemiddelde. Maar zicht op allochtonen geeft het niet. Lattens paper wordt pas interessant voor een strateeg van het integratiebeleid bij de analyse van de zuigelingensterfte. Als blijkt dat onder allochtone vrouwen in hogere inkomenswijken de zuigelingensterfte lager is dan onder allochtone vrouwen in lagere inkomenswijken, dat roept volgens Tesser meteen de vraag op of een verschil in inkomen daar de oorzaak van is. Hij zou dan ook graag vernemen wat de resultaten zijn als gecorrigeerd wordt voor het individuele of huishoudensinkomen.

Het zijn met name de indicatoren voor integratie en de factoren die integratie beïnvloeden, waar de makers van het integratiebeleid in geïnteresseerd zijn. Direct daarna komt de vraag naar beleidseffecten. Wat is de invloed van de beleidsingrepen op de integratie van allochtonen of minderheden?

Met het SSB had het ministerie van Justitie antwoord op de volgende vragen gehoopt te krijgen:

1. Hoe staat het met de integratie van onderscheiden groepen minderheden?

2. Welke ontwikkelingen hebben zich de afgelopen jaren hierin voorgedaan?

3. Wat is de betekenis van de ruimtelijke segregatie voor de integratie van minderheden?

4. Helpt een beleid dat gericht is op vermindering van de segregatie?

5. Wat is het effect van het integratiebeleid (of delen daarvan) op de integratie van de minderheden?

In het paper van Latten heeft Tesser de antwoorden op zijn vragen jammer genoeg niet gezien. Hij hoopt echter dat in de toekomst het SSB daar wel in slaagt.

Noten

- (1) Paul Tesser is hoofd van de afdeling Strategie, Onderzoek en Communicatie (SOC) van de Directie Coördinatie Integratiebeleid etnische Minderheden (DCIM) van het Ministerie van Justitie. Tot voor kort was hij als medewerker van het SCP belast met de jaarlijkse Rapportage Minderheden.

De deur van het demografisch laboratorium Perspectieven voor externe benutting van de virtuele volkstelling 2001

J. Nobel (1)

1. Inleiding

In 1968 publiceerde SISWO een boekje over De Volkstelling 1970. Het bevatte “Aanpassingsdesiderata in het perspectief van het sociaal-wetenschappelijk onderzoek”. De wensen hadden betrekking op de te gebruiken indelingen en de te produceren tabellen. De Volkstelling kwam er, zij het iets later dan CBS en SISWO voorzagen, en leidde onder meer tot een groot aantal Census monografieën van SISWO. Het ontbreken van een Volkstelling sinds 1971 is veelvuldig betreurd, onder meer in de redactionele kolommen van Mens en Maatschappij. Zonder enige enquête durf ik wel de stelling aan dat de sociaal-wetenschappelijke belangstelling anno 2003 voor een Volkstelling allereerst zal uitgaan naar de microdata. Het bewerken en tabuleren dat dertig jaar terug aan het CBS met zijn mainframe computer was voorbehouden, doet elke zichzelf respecterend socioloog nu immers thuis op zijn eigen PC.

In mijn artikel beschrijf ik eerst de voordelen van *data sharing* en de daaraan te stellen randvoorwaarden vanuit het perspectief van de gegevensbescherming. Vervolgens zet ik uiteen met welke methoden microdata nuttig en tegelijk veilig gebruikt kunnen worden. Die methoden vallen grofweg uiteen in

safe data en *safe settings*. Dan geef ik kort de Nederlandse ontwikkeling weer met betrekking tot de beschikbaarheid van officiële statistische microdata. De huidige situatie in Nederland kunnen we vergelijken met beleid en praktijk in andere westerse landen. Daarvan beschouw ik de Noord-Amerikaanse en Scandinavische nader. De eerste omdat de VS wel als een walhalla voor onderzoekers worden neergezet, de laatste omdat de Nederlandse officiële statistiek vanwege het toenemende gebruik van gegevens uit registraties in plaats van enquêtes steeds meer op de Scandinavische statistiek gaat lijken. Ter afsluiting formuleer ik een aantal conclusies en aanbevelingen.

2. *Data sharing*: voordelen en nadelen

Aan het meervoudig gebruik van microdata, de bouwstenen van het sociaal-wetenschappelijk onderzoek, zijn tal van voordelen verbonden, vanuit elk betrokken perspectief.

- De voordelen voor de gebruiker, de *secundaire onderzoeker*, vaak de academische wetenschapper, zijn evident. Hij hoeft minder geld en tijd te besteden aan dure en tijdrovende gegevensverzameling. De microdata zijn al *gecleaned*, *ge-edit* en gedocumenteerd. Er is consultatie mogelijk van primaire onderzoekers die expertise met de data hebben. Als de data voor collega-secundaire onderzoekers onder dezelfde voorwaarden toegankelijk zijn, is intercollegiale toetsing door het wetenschappelijk forum goed mogelijk. Een cumulatie van onderzoeksresultaten wordt zo bevorderd. En replicerbaarheid wordt wel als een van de basiskennmerken van echt wetenschappelijk onderzoek gezien.

- De voordelen voor de *primaire onderzoeker*, vaak de officiële statisticus, zijn even helder. Zijn werk wordt breder benut, wat tot meer erkenning en meer feedback van gebruikers leidt. Die erkenning slaat zowel op zijn instituut als zodanig als op de gebruikte standaarden en classificaties.
- De *belastingbetaler* krijgt meer waar voor zijn geld als dure, op zijn kosten verzamelde gegevens intensiever gebruikt worden.
- De *respondent* (of registratiehouder, als alternatieve bron van gegevens) hoeft minder vaak lastig gevallen te worden door de onderzoeker.
- De *opdrachtgever* heeft baat bij resultaten die snel en relatief goedkoop tot stand gebracht worden. De kans dat de resultaten consistent zijn met andere onderzoeksresultaten, wordt bevorderd als onderzoekers van dezelfde (standaard) datasets gebruik maken.

Kortom, het mes snijdt aan meer kanten. Dat neemt niet weg dat er ook wel nadelen aan *data sharing* vast zitten. Zo heeft de secundaire onderzoeker minder eigen invloed op de keuze, definitie en meting van variabelen, of op het moment van beschikbaarheid van zijn onderzoeksgegevens. En de primaire onderzoeker zal juist het gevoel kunnen hebben dat anderen aan de haal gaan met onderzoeksresultaten waarmee hij zelf graag had willen komen. Op zijn minst kan er een tegenstrijdig belang in de *timing* van de beschikbaarstelling van de microdata zitten. De primaire onderzoeker heeft ze immers met een bepaald doel voor ogen verzameld. En het bereiken van dat doel heeft prioriteit boven de beschikbaarstelling van microdata aan secundaire onderzoekers. Daarbij merk ik nog op dat de voor- en

nadelen zich ook voordoen als we de begrippen primair en secundair een schakel naar voren halen in de hele keten van bewerking van primaire gegevens tot publicatie van geaggregeerde informatie. Nu het CBS in toenemende mate en om verschillende redenen gebruik maakt van gegevens uit registraties, kan het CBS immers zelf ook als secundair beschouwd worden in zijn relatie tot de primaire registratiehouders. De voordelen van lagere kosten, snellere beschikbaarheid, gemeenschappelijke standaarden, minder enquêtedruk en dergelijke, doen evenzeer opgeld. Maar ook de daartegenover staande nadelen gaan tot op zekere hoogte op.

3. Statistische beveiliging: motieven en bedreigingen

Als de voordelen zwaarder wegen dan de nadelen, komt de vraag aan de orde onder welke voorwaarden *data sharing* mogelijk is. We hebben het immers over microdata, gegevens van individuele personen. Die personen hebben een grondwettelijk recht op bescherming van hun privacy, hun persoonlijke levenssfeer. En *data sharing* kan daar op gespannen voet mee staan. Dat grondwettelijk recht is in verschillende wetten nader uitgewerkt. Drie daarvan zijn er in het bijzonder van belang voor het CBS: de Wet Bescherming Persoonsgegevens (WBP), de recent in werking getreden Wet op het Centraal Bureau voor de Statistiek (Eerste Kamer, 2003, zie bijlage 1 voor de meest relevante artikelen), en de Wet op de Geneeskundige Behandelovereenkomst (WGBO). Voor een uitvoerige beschouwing over de relevante wetgeving verwijs ik naar Hooghiemstra (2002), Hij beschrijft onder meer

uitgebreid de verschillende belangrijke uitzonderingen die voor wetenschappelijk onderzoek en statistiek in de WBP opgenomen zijn. In dit artikel concentreer ik mij op de bepalingen van de CBS-wet. Die is geschreven met het oog op de verzelfstandiging van het CBS tot Zelfstandig Bestuursorgaan en is per 3 januari 2004 in werking getreden. De bepalingen daarvan verschillen in essentie niet van die van de sinds 1 juli 1996 geldende Wet op het Centraal bureau en de Centrale commissie voor de statistiek (2).

In artikel 3 van de CBS-wet wordt de taak van het CBS omschreven als het van overheidswege verrichten van statistisch onderzoek ten behoeve van praktijk, beleid en wetenschap en het openbaar maken van de op grond van zodanig onderzoek samengestelde statistieken. Het CBS maakt en publiceert derhalve statistieken, dat wil zeggen geaggregeerde informatie. Voor dit collectieve goed wordt het vanuit de collectieve middelen gefinancierd. En de wetenschap is één van de daarmee beoogde doelgroepen. Het is dus letterlijk genomen niet een taak van het CBS om microdata, in feite een tussen- en geen eindproduct van het CBS, voor wetenschappelijk onderzoek beschikbaar te stellen. Integendeel, geheimhouding van de individuele gegevens is een basisregel in de CBS-wet. Dat blijkt nadrukkelijk uit artikel 37: de door het CBS ontvangen gegevens worden uitsluitend gebruikt voor statistische doeleinden, en worden in het bijzonder niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.

Volgens de Memorie van Toelichting worden de individuele gegevens van het CBS daarmee ook vrij-waard van de fiscus, het openbaar ministerie, enzovoorts. Die vrijwaring is de afgelopen decennia herhaaldelijk van belang gebleken, maar beperkt zich tot het CBS met zijn wettelijke positie. Ze strekt zich niet uit tot de individuele CBS gegevens waarover academische onderzoekers beschikken.

Met die strikte verplichting tot statistische geheimhouding zouden de voordelen van *data sharing* zoals hierboven geschetst ten enenmale onbereikbaar zijn. En dat is nu ook weer niet de bedoeling. Artikel 41 maakt het dan ook mogelijk dat het CBS in expliciete afwijking van de basisregel van geheimhouding voor statistisch of wetenschappelijk onderzoek microdata verstrekt dan wel er toegang toe verleent. De beschikbaarstelling van microdata voor het CBS is zo een expliciete uitzondering voor één bepaald doel op de algemene regel van de statistische geheimhouding.

Er is dus een uitdrukkelijke wettelijke verplichting om voorzichtig om te springen met de beschikbaarstelling van microdata. Maar daarvoor bestaan ook andere redenen: de onderzoeksethiek en het eigenbelang. De ethiek van het onderzoek brengt met zich dat gegevens die voor een bepaald doel verzameld worden, alleen voor dat doel gebruikt worden en niet voor een ander doel, en al helemaal niet als dat andere doel niet in het verlengde van het oorspronkelijke doel ligt. Dit principe van de doelbinding is zowel in verschillende wetten zoals de WBP, als in verschillende gedragscodes (zie

bijvoorbeeld Sociaal-Wetenschappelijke Raad (2003)). De onderzoeker heeft voorts een eigen belang bij statistische geheimhouding omdat daarmee de respons op enquêtes bevorderd wordt. Respondenten hechten belang aan een zorgvuldige omgang met hun data. Dan is het in het belang van de respons en de kwaliteit van de resulterende data om de privacy van de respondenten te respecteren. Zoals gezegd, maakt het CBS in toenemende mate van gegevens uit administraties in plaats van uit eigen enquêtes gebruik. Dan blijft het argument van het welbegrepen eigenbelang intact. De registratiehouder zal immers niet graag zijn individuele data zien weglekken via het CBS. Het CBS moet er dan veel, zo niet alles, aan doen om de geheimhouding van die data en daarmee de privacy van de geregistreerden te garanderen.

Dat dit geen loos motief is, bleek het afgelopen voorjaar toen in de Tweede Kamer het VVD-lid Blok een amendement op de CBS-wet indiende. Het amendement was in eerste instantie beslist onhandig geformuleerd, zodat alle data die het CBS voor zijn statistiekproductie ontvangt, via de achterdeur het CBS dreigden te kunnen verlaten. Verschillende registratiehouders lieten al gauw doorschemeren daar absoluut niet van gediend te zijn en hun berichtgeving aan het CBS te willen heroverwegen als het amendement aangenomen zou worden.

Waartegen moeten nu de individuele gegevens beschermd worden? Door welke slordigheid kunnen gegevens lekken? Hoe kan een *intruder* gegevens onthullen? Er zijn globaal

drie onthullingsscenario's te onderscheiden: responskennis, systematische koppeling en spontane onthulling.

- Als een onderzoeker van een hem bekende derde weet dat deze in een bestand opgenomen is, wordt het relatief gemakkelijk deze persoon en daarmee eventueel onbekende scores van hem in het bestand op te sporen. Wel is het zo dat als bestanden verouderd zijn of op andere landen betrekking hebben, de kans op correcte responsekennis al snel gering gaat worden.
- Onthulling door systematische koppeling wordt mogelijk als de microdata via een of andere sleutel gekoppeld worden aan een ander bestand. Die sleutel kan variëren van een uniek identificatienummer tot een combinatie van naam, adres en woonplaats (NAW). De mogelijkheid van zo'n systematische koppeling wordt groter naarmate er meer databases met persoonsgegevens te koop of te vinden zijn. Wat dat betreft liggen de Verenigde Staten ver vóór op Europa. Bovendien worden de koppelsoftware en de hardware steeds krachtiger en sneller.
- Van spontane onthulling spreken we als een individu door een relatief zeldzame combinatie van scores op achtergrondvariabelen herkenbaar wordt, ook als zijn NAW-gegevens niet in het bestand zijn opgenomen. Die zeldzame combinatie kan op allerlei variabelen betrekking hebben, zoals beroep en woonplaats, of aantal kinderen en woonplaats. Naarmate de variabelen gedetailleerder zijn en betrouwbaarder gemeten worden, wordt de kans op zo'n zeldzame combinatie groter.

4. Beveiligingsmethoden: *safe data* versus *safe settings*

Om te voorkomen dat records geïdentificeerd worden en vervolgens gegevens van geïdentificeerde personen onthuld, zijn verschillende soorten methoden gangbaar. Deze worden wel gerubriceerd als *safe data* versus *safe settings*.

Data kunnen worden veilig gemaakt door met name de indirect identificerende variabelen in te dikken. We doen dat door afronden, afkappen, aftoppen, en dergelijke. De nieuwe getransformeerde waarden zijn grover, en maken daardoor herkenning van de oorspronkelijke eenheden moeilijker. Maar het is niet alleen de herkenning die moeilijker wordt, er zijn ook vanuit het oogpunt van de onderzoeker nadelen aan verbonden. Het wordt voor hem lastiger om tot eigen categorisering in bijvoorbeeld een beroepenclassificatie te komen, en om contextuele kenmerken van bijvoorbeeld 4 cijfer postcoderegio's, zoals het gemiddelde inkomen, het percentage allochtonen of de nabijheid van bepaalde voorzieningen, toe te voegen.

Twee soorten variabelen zijn in het bijzonder van belang omdat ze niet alleen sterk bijdragen aan de identificatie van respondenten maar ook ingedikt als achtergrondvariabele in veel modellen een belangrijke rol spelen: postcode en geboortedatum. Die ken je wel van je broer, je buurman, etc., maar niet zijn sofi- of A-nummer. Het zijn bovendien variabelen die het koppelen aan integrale registraties aanzienlijk vergemakkelijken. Het is dus de vraag of je zes- of vijf- of vier- of drie- of twee- of één-positie postcode als regiovariabele moet opnemen, en geboortedag of -maand of -jaar of -jaarklasse.

Meer geavanceerde statistische beveiligingsmethoden kijken niet naar de records separaat. Specifieke waarden kunnen tussen twee records omgewisseld worden om kwetsbare cellen te verhullen: *data swapping*. Het is zelfs mogelijk om vanuit de onderliggende multivariate verdeling synthetische records te genereren die voor onderzoek beschikbaar worden gesteld. Zie bijvoorbeeld het werk van Woodcock en Abowd van Cornell University op gekoppelde werkgevers-werknemers bestanden uit Frankrijk. Zonder nadere studie zijn dit soort meer geavanceerde methoden voor veel secundaire onderzoekers overigens niet altijd even gemakkelijk toepasbaar.

Aan de andere kant kan ook het onderzoeksproces zelf beveiligd worden. Dat kan allereerst door slim gebruik te maken van automatiseringshulpmiddelen, hard- en software. Het afschermen van de microdata ten opzichte van dreigingen uit de rest van de wereld behoort daarbij. Daarvoor kan men denken aan allerlei fysieke en logische vormen van toegangsbeveiliging. Maar bij sommige werkwijzen is het ook goed mogelijk te loggen welke analyses de onderzoeker met de data uitvoert. Een alternatieve mogelijkheid is juridisch van aard. In een contract kunnen beveiligingsbepalingen worden opgenomen. Tot die bepalingen kunnen behoren: een koppelverbod, een verplichting om de data na gebruik te vernietigen, een verplichting om concept resultaten voor publicatie door het CBS op beveiligingsaspecten te laten screenen, enzovoorts. Ook een juridisch gefundeerde selectie van toegelaten onderzoekers hoort hier bij.

In de literatuur zijn er zo drie soorten modellen ontstaan. Bij *on-desk* gebruik van microdata hebben onderzoekers de

microdata op hun eigen PC voor analyse klaarstaan. Bij *on site* analyse gaat de onderzoeker voor de data naar het CBS toe. Bij *remote access* of *execution* blijven de onderzoeker en de data elk op hun eigen plek: de onderzoeker biedt analyseprogramma's aan, bijvoorbeeld in SPSS, aan ter verwerking op de microdata binnen het CBS achter een veilige *firewall*. Alleen de (veilige) geaggregeerde output gaat terug naar de onderzoeker.

5. De CBS praktijk: van *safe data* naar *safe settings*

In de loop van de jaren tachtig raakten de verhoudingen tussen het CBS en de academische wetenschap sterk bekoeld, nadat in de jaren vijftig en zestig tal van sociale wetenschappers het onderzoeksvak nog bij het CBS geleerd hadden. Het CBS begon de nieuwe analytische mogelijkheden van het koppelen en analyseren van grote bestanden in de vingers te krijgen: de dynamische inkomensstatistiek kwam tot stand door een panel van belastingplichtigen te construeren. En de werkloosheidsstatistiek ging gebaseerd worden op een combinatie van de nieuwe Enquête Beroepsbevolking met de administratieve bestanden van de arbeidsbureaus. Daarbij werd echter duidelijk dat het weglaten van NAW gegevens niet volstaat om respondenten *de facto* anoniem te houden. En de recente geschiedenis van de Volkstelling had het CBS voldoende duidelijk gemaakt hoe belangrijk deze privacy in het maatschappelijk *discours* is. Bij de privacy kwestie voegde zich een budgettair probleem. De kabinetten Lubbers legden het CBS de taakstelling op een miljoen gulden aan ontvangsten uit de beschikbaarstelling van microbestanden voor wetenschappelijk onderzoek te genereren. Zowel de bezuinigingsproblematiek als – meer ideologisch

– het profijtbeginsel vormde een achtergrond hiervoor. Uit antwoorden op Kamervragen van het toenmalige lid Wallace bleek dat de Minister van Onderwijs deze insteek van de voor het CBS verantwoordelijke Minister van Economische Zaken volledig deelde. De protesten waren niet van de lucht. De Sociaal-Wetenschappelijke Raad van de KNAW bracht met de Vereniging van Onderzoeksinstituten VOI een pamflet Wetenschap en Statistiek uit. Het hoofd van het Steinmetz archief De Guchteneire en het CCS lid De Kok vanuit VWS ontketenden campagnes tegen de toenmalige DG van het CBS, die onder meer de kolommen van de NRC haalden. En er werden onvermijdelijk Kamervragen gesteld.

Op initiatief van de toenmalige DG Wetenschapsbeleid Van Spiegel werd begin jaren negentig de weg geplaveid voor een verzoening tussen CBS en Academia. Op een gezamenlijk memorandum van CBS DG Begeer en demografische *éminence grise* Van de Kaa volgde een pilot-project van NWO, grotendeels uitgevoerd onder supervisie van de Tilburgse methodoloog Segers. Van de kant van het CBS was er een serieuze bereidheid om afspraken te maken over bestandsleveringen. Die afspraken betroffen de persoonsenquêtes, waarvan de bestanden partieel beveiligd zouden worden, waarna het CBS ze onder overeenkomst aan de onderzoekers zou leveren. NWO toonde zich bereid om het CBS in staat te stellen aan zijn financiële taakstelling van een miljoen gulden te voldoen. In het voorjaar van 1994 richtte NWO het Wetenschappelijk Statistisch Agentschap (WSA) op en sloot het een overeenkomst voor vier jaar met het CBS. Het WSA stond aanvankelijk onder leiding van ex-CBS methodoloog Stronkhorst, die over

een goede academische reputatie en ingangen beschikte. De inwerkingtreding van de CBS/CCS-wet per 1 juli 1996 verschaftte het juridisch fundament voor deze beschikbaarstelling door het CBS van microdata voor wetenschappelijk onderzoek.

De jaren sinds 1994 bleken vervolgens een relatieve oase van rust en productiviteit. De grote inhaalvraag werd gestild, en de voortdurende datastroom hield de statistici en de onderzoekers aan het werk. Tweemaal, telkens na vier jaar, vond er een onafhankelijke evaluatie plaats, die tot een positieve uitkomst leidde. De conclusie dat het WSA op basis van de bestandsovereenkomst met het CBS goed werk levert, werd in 2003 onderschreven door zowel NWO voorzitter Nijkamp als een SWR/KNAW commissie onder voorzitterschap van professor Van der Knaap. WSA hoofd Dekker heeft het WSA herhaaldelijk buiten de landsgrenzen, onder meer in Duitsland, als een geslaagd model voorgehouden. Dat neemt niet weg dat er ook wel contrapunten zijn aan te geven zoals Borghans, op wiens precieze argumentatie overigens nog wel wat af te dingen valt, dit voorjaar in ESB deed. Nederlandse onderzoekers wenden zich naar zijn idee meer tot niet-Nederlandse microdata dan dat niet-Nederlandse onderzoekers Nederlandse data analyseren. En daardoor weten we minder van de Nederlandse samenleving dan wenselijk. Uit verschillende initiatieven en rapporten blijkt een zekere onrust op dit ogenblik in sociaal-wetenschappelijk Nederland ten aanzien van de data-infrastructuur.

In de jaren sinds de oprichting van het WSA is de CBS facilitering van wetenschappelijk onderzoek, vaak met steun van NWO en KNAW, gaandeweg uitgebreid. Tot

de mijlpalen behoren in chronologische volgorde:

- De opening van *Cerem* in 1998 als laboratorium voor onderzoek op bedrijfseconomische microdata binnen de muren van het CBS. Op deze microdata is een zwaarder geheimhoudingsregime van toepassing zodat er intensief overleg met het Ministerie van EZ, VNO-NCW en juristen van een aantal multinationals nodig was om het - getuige de onder voorzitterschap van Lubbers halverwege de jaren negentig opgestelde Onderzoeks-verkenning economische wetenschappen ook door KNAW en NWO - gewenste resultaat te bereiken. De opbouw van de databestanden werd financieel gesteund door NWO. In 2002 verschenen alleen al drie dissertaties mede gebaseerd op onderzoek in Cerem. Dezelfde technische faciliteiten worden overigens ook regelmatig op andere bestanden gebruikt. Zo start er elke maand wel een onderzoek *on site* op de doodsoorzaken-gegevens.
- De digitalisering van de Volkstellinguitkomsten in vroeger jaren, mede ter gelegenheid van het CBS eeuwfeest in 1999. De digitale publicatie had als zodanig niet betrekking op microdata, hoewel het een uitdaging zou zijn om uit de tabellen van 1899 het onderliggende microbestand voor zover mogelijk te reconstrueren. Ze leidde echter ook tot hernieuwde belangstelling voor de reconstructie van de microdata van de Volkstellingen van 1960 en 1971, die nog grotendeels maar niet geheel ongeschonden in de digitale archieven aanwezig waren.
- De eerste stappen in 2003 op de weg naar *remote execution* faciliteiten voor departementale onderzoekers in het kader van het Centrum voor Beleidsstatistiek. Niet alleen universitaire onderzoekers maar ook depar-

tementale beleidsonderzoekers hebben grote behoefte aan toegang tot CBS microdata. Een overeenkomst met het Ministerie van Sociale Zaken en Werkgelegenheid geeft onderzoekers van dat departement de mogelijkheid om *on site* bij het CBS te werken op microdata. In het verlengde hiervan wordt tevens onderzocht of *remote execution* soelaas biedt voor de SZW analyses. Vooralsnog gebeurt dit door menselijke tussenkomst, maar de ambitie is er om dit verregaand te automatiseren en te generaliseren naar andere onderzoekers respectievelijk andere bestanden dan voor SZW relevant.

- De uitwerking en instrumentering in 2003 van het amendement Blok voor epidemiologisch onderzoek met behulp van CBS doodsoorzakengegevens die momenteel onderhanden is. De doodsoorzakengegevens vormen de enige registratie die het CBS zelf integraal bijhoudt voor statistisch onderzoek. De gegevens vormen een voor statistisch onderzoek vaak welkome afsluiting van de onderzoeksrecords en patiëntenbestanden die epidemiologen en andere medisch onderzoekers bijhouden. Onder nader te bepalen voorwaarden, waarbij het *informed consent* van betrokkenen een belangrijke rol speelt, mogen deze gegevens het CBS verlaten zodat ze aan deze records en bestanden gekoppeld kunnen worden voor met het oog op verdere analyse. Het simpele feit dat deze wettelijke bepaling noodzakelijk is om het CBS in staat te stellen aan de epidemiologische wens te voldoen, houdt overigens in dat andere identificeerbare gegevens het CBS zeker niet zullen mogen verlaten.

Daarnaast gaat het CBS in toenemende mate strategische samenwerkingsverbanden aan met universitaire

onderzoeksgroepen en planbureaus. Het resultaat daarvan geldt als officiële CBS output, en moet dus mede in het CBS belang zijn en aan CBS kwaliteitsstandaarden voldoen. Het analysewerk dient uiteraard binnen de CBS standaarden te vallen. Maar voor het overige geeft deze samenwerkingsvariant externe onderzoekers wellicht nog meer mogelijkheden dan het *on site* werken.

6. De laatste Nederlandse Volkstelling, privacy en statistiek

Er valt veel te zeggen over de Volkstelling 1971, die voor 1970 gepland was en de laatste klassieke Volkstelling in Nederland zou blijken te zijn. Een paar krenten uit de pap.

- De klassieke Volkstelling heeft nooit een exclusieve onderzoeksdoelstelling gehad. Integendeel, nevendoel was steeds waar nodig correcties op de bevolkingsregisters aan te brengen. Voor de Volkstelling van 1971 werd uiteindelijk besloten dat correcties van de bevolkingsregisters geen nadelige gevolgen voor betrokkenen, zoals illegaal in Nederland verblijvende buitenlanders zouden hebben. Voor de voor 1980 geplande Volkstelling is deze administratieve functie expliciet geschrapt. Maar juist deze telling is er zoals bekend nooit gekomen.
- Het verzet tegen de Volkstelling 1971, geleid door een Comité Waakzaamheid Volkstelling, was omvangrijk en leidde uiteindelijk, met een zeer aanzienlijke vertraging, tot de Nederlandse privacywetgeving, de WPR van 1989. In die privacywetgeving zijn tamelijk liberale en werkbare uitzonderingen voor statistiek en wetenschappelijk onderzoek op de standaardregels

geïntroduceerd (zie ook Hooghiemstra, 2003). Een tussenstap in het proces van VT 1971 naar WPR 1989 is overigens de motie Brinkhorst (de huidige D66 Minister van EZ) die de Tweede Kamer op 28 november 1978 aanvaardde, waarin de regering verzocht werd geen maatregelen te treffen tot het houden van een Volkstelling alvorens overleg met de Kamer was gepleegd. Een van de overwegingen van de motie Brinkhorst was dat er geen telling gehouden moest worden als er geen voldoende wettelijke waarborgen bestonden ter bescherming van persoonlijke gegevens.

- De gegevensverzameling voor de Volkstelling 1971 vond plaats met ponskaartenboekjes. Dat alleen al

voedde het *Big Brother* syndroom. Maar in de politieke discussie werd eveneens veelvuldig herinnerd aan het misbruik van de bevolkingsadministratie door de Duitse bezetter in de oorlogsjaren en de mogelijkheid van een coup leidende tot een kolonelsregime.

- Naam- en adresgegevens werden op een aparte omslagkaart genoteerd. Deze werden na controle op volledigheid, los van de overige gegevens naar het CBS gezonden. De Jonge, hoofd Algemene Tellingen van het CBS, stelde dan ook in NRC/Handelsblad in het najaar van 1970: “Wij zijn het enige land ter wereld waar absolute anonimiteit bestaat, want voor verwerking van de gegevens zijn naam en adres al verdwenen.”

De toenmalige Directeur-generaal Algemene Beleidsaan-gelegenheden van het Ministerie van Sociale Zaken en Volksgezondheid had in 1968 al een methodiek voorgesteld om de Volkstellingsgegevens te gebruiken voor longitudinaal vervolgonderzoek. Daartoe zou het CBS-nummer van de omslagkaart van de Volkstelling 1970 overgenomen moeten worden op de persoonskaart in het gemeentelijk bevolkingsregister en tien jaar later weer op de omslagkaart van de Volkstelling van dat jaar. Daarmee zou tegemoet gekomen kunnen worden aan de grote behoefte aan gegevens omtrent mobiliteit per beroep, per geografisch gebied en per bedrijfstak. De Directeur-generaal van de Volkshuisvesting sloot zich nadrukkelijk bij deze wens aan.

De bedoeling van CBS, CCS en Ministerraad was bescheidener: het bewaren van een tien-procentssteekproef op naam en adres. De Minister van Economische Zaken verdedigde dit bewaren in de Tweede Kamer met een naar eigen zeggen “wat bizar voorbeeld. Laten wij nu eens zeggen dat de statistici of de beleidsinstanties geïnteresseerd zijn in de vraag, in hoeverre mensen met een drankneus in 1970 ook in 1980 nog een drankneus hebben, dat wil zeggen, hoeveel daarvan nog dezelfden zijn. Dit betekent, dat men dan nominatief moet nagaan, hoeveel onder de pakweg 12 000 Nederlanders met een drankneus in 1970 er in 1980 nog aanwezig zijn met dezelfde, ietwat bizarre afwijking.”

Een amendement om alle op naam en adres gestelde gegevens binnen drie jaar te vernietigen haalde het in 1970 niet in de Kamer. Bij de bespreking van de uitvoeringsmaatregelen een jaar later ging de Minister met dat voorstel echter alsnog akkoord “teneinde i.c. een belangrijke extra waarborg te scheppen voor de bescherming van de persoonlijke levenssfeer.” Onder meer het hoofd van het Amsterdamse Bureau voor Onderzoek en Statistiek was hier buitengewoon ontevreden over.

- Van de aanvankelijke plannen om uitgebreid wetenschappelijk gebruik te maken van de microdata van de Volkstelling, met name voor longitudinaal onderzoek, is niets terechtgekomen.
- De discussies over de Volkstellingen 1971 en 1980 hebben bijgedragen aan de instelling door het CBS van de Arbeidskrachtentelling (AKT, later Enquête Beroepsbevolking EBB) en het Sociaal-economisch Panelonderzoek SEP.
- Achteraf is Jan Holvast, toentertijd één van de voormannen van het Comité Waakzaamheid Volkstelling, nu partner in een privacy consultancy bedrijf, overigens aanzienlijk milder over het bedreigende karakter van de Volkstelling 1971. De toenmalige commotie over de bedreiging van de privacy noemde hij in 1997 nogal “overdreven” en eigenlijk “stelde die volkstelling zelf niet zoveel voor.” Over de huidige aanpak van het CBS maakt hij zich ook geen zorgen in Computable in 1999. “De belangrijkste reden daarvoor is dat het CBS als geen ander geleerd heeft van de lessen uit 1970.” (Holvast, 1999).

7. Het gras aan de andere kant van de oceaan

In zekere zin zijn de Verenigde Staten de bakermat van de microdata. Al in de jaren zestig verschenen er bij wijze van *pilot tapes* met *public use microdata samples* uit de constitutioneel verplichte tienjaarlijkse Volkstelling. Voor de 1970 Census waren deze PUMs al een standaard product. Deze data zijn voor iedereen vrijwel om niet beschikbaar, voor academici, maar ook voor lagere

overheden, journalisten en commerciële ondernemingen. Omdat de data voor iedereen laagdrempelig beschikbaar zijn, en het Census Bureau tegelijk onder strikte geheimhoudingsbepalingen (Title 13, U.S. Code) moet werken, wordt een scala aan beveiligingsmethoden gebruikt: “*data-swapping, topcoding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables*”. De verspreiding van de gegevens vond aanvankelijk per tape plaats, later per CD-ROM, en tegenwoordig kunnen gebruikers *on line* microdata downloaden c.q. zelf maatwerk tabellen vervaardigen.

Recent is een zekere verstrakking van het beveiligingsregime te zien, die voortvloeit uit wat hierboven gezegd is over de immer groeiende koppelmogelijkheden, met massale *databases* met persoonsgegevens en *high capacity hard- en software*, die in de VS bij uitstek voor het oprapen liggen. “*Because of the rapid advances in computer technology and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through disclosure-limitation techniques*.” Hoe belangrijk deze beveiliging voor het Census Bureau is blijkt in de inleiding van de meer dan 700 pagina's tellende technische documentatie. De eerste twee zinnen daarvan luiden: “*Public use microdata samples are ASCII files which contain individual records for a sample of people and housing units. Information which could identify a household or an individual is excluded in order to protect the confidentiality of respondents*.” (zie bijlage 2 voor de complete tekst van deze pagina, inclusief een aantal door de Amerikanen toegepaste beveiligingsmethoden).

Voor de Volkstelling 2000 levert het *Census Bureau* - vanaf mei 2003, dat wil zeggen met een vertraging van drie jaar - twee soorten PUMs:

- De een procent bestanden geven de gebruikers het maximum aan inhoudelijke informatie. Om dat mogelijk te maken is de geografische drempel opgetrokken naar wat de super-PUMA (*Public Use Microdata Area*) genoemd wordt: Super-PUMA's hebben ten minste 400.000 inwoners. Elk van de 50 staten is herkenbaar maar een staat wordt pas opgesplitst in super-PUMA's bij een bevolking van 800.000.
- De 5 procent bestanden bevatten meer regionaal maar navenant minder inhoudelijk detail. De PUMA's hebben minimaal 100.000 inwoners. Op onderzoeksvariabelen worden de genoemde *disclosure limitation* technieken strenger toegepast.

Om de gedachten te bepalen: Nederland telt drie steden met meer dan 400.000 inwoners, en 25 met meer dan 100.000 inwoners.

Om ook geavanceerd analytisch onderzoek op de vele bedrijfseconomische gegevens die het *Census Bureau* verzamelt, mogelijk te maken, is bij het *Census Bureau* in Washington in de jaren tachtig het *Center for Economic Studies* opgericht, dat als model voor het Nederlandse *Cerem* gefungeerd heeft. Het succes van dit *Center* en de honger naar data bij prestigieuze en kapitaalkrachtige universiteiten buiten Washington heeft geleid tot een aantal dependances. Met financiële steun van de *National Science Foundation* heeft het *Census Bureau* bij circa tien prestigieuze en kapitaalkrachtige universiteiten *Research Data Centers* opgericht. Om van de microdata bij deze RDC's gebruik te kunnen maken moeten onderzoekers

een onderzoeksopzet indienen die door het *Census Bureau* getoetst wordt op wetenschappelijke merites en op het nut ervan voor het *Census Bureau* zelf. Onderzoek dat niet in het belang van het *Census Bureau* zelf is, behoort niet tot de mogelijkheden. Met het beoordelingsproces kan bovendien behoorlijk wat tijd gemoeid zijn. En de gemiddelde onderzoeker die niet in de buurt van de topuniversiteiten van Boston, Chicago, en dergelijke, zijn onderzoek verricht, heeft niet het gevoel hier enig profijt van te hebben. In deze RDC's zijn ook wel meer gedetailleerde *Census* microdata toegankelijk, onder dezelfde beperkende voorwaarden.

Het *Census Bureau* is maar één van de 72 federale statistische bureaus in de VS, zij het wel het grootste. Voor de andere statistische bureaus variëren de geheimhoudingsbepalingen en de wettelijke mogelijkheden om toegang tot microdata te geven aanzienlijk. De fiscale inkomensgegevens vormen het ene uiterste: deze zijn vrijwel niet toegankelijk voor externe onderzoekers. Maar microdata met betrekking tot de beleidsterreinen gezondheid, justitie en onderwijs worden op ruime schaal onder licentie verspreid door de verantwoordelijke bureaus. Deze microdata worden alleen beschikbaar gesteld voor onderzoeksdoeleinden. Vanwege de contractuele afdekking kennen deze bestanden veelal een rijkere inhoud dan wat voor de *Census* PUMs gebruikelijk resp. toegestaan is. Ook hier staat overigens de statistische geheimhouding permanent onder druk. In de nasleep van 11 september heeft de federale regering, in strijd met de geldende relevante statistiekwetgeving en in strijd met alle toezeggingen aan de berichtgevende onderwijsinstanties, met terugwerkende kracht de bevoegdheid gecreëerd

om van het NCES, het nationaal bureau voor de onderwijsstatistiek, individuele gegevens over studenten op te eisen. Men was in het bijzonder geïnteresseerd in studenten van Arabische origine aan luchtvaartsscholen (<http://www.uwm.edu/~margo/govstat/jsm.pdf>).

In het noordelijke buurland Canada heeft het zogeheten *Data Liberation Initiative* tot een sterke verruiming van toegankelijkheid en gebruik van microdata geleid. *Statistics Canada* en de *Social Science Research Council* hebben in de tweede helft van de jaren negentig daartoe de handen ineengeslagen. Onder dit initiatief zijn inmiddels zo'n 400 PUMs van *Statistics Canada*, meestal van cross-sectionele persoonssurveys, toegankelijk voor onderzoekers van universiteiten en ministeries. Voor een aantal bestanden kunnen onderzoekers via *remote access* toegang krijgen tot meer detail in deze PUMs. Negen Canadese universiteiten hebben bovendien een eigen *Research Data Centre* waar meer gedetailleerde microdata van *Statistics Canada* (bijvoorbeeld *National Population Health Survey* en *Survey of Labour & Income Dynamics*) geanalyseerd kunnen worden. Per mei 2003 liepen er 244 projecten in deze RDCs waarbij bijna 400 onderzoekers betrokken waren.

8. Het gras aan de andere kant van Jutland

Bij afwezigheid van een bevolkingsregistratie hebben de Angelsaksische landen een periodieke Volktelling nodig om elementaire statistische informatie als het aantal inwonenden per staat of per gemeente te kunnen bepalen. In de Angelsaksische landen is het concept van een PUM

dan ook vrij concreet: de vragen op de vragenlijst leiden tot variabelen en een steekproef van de respondenten bepaalt het aantal cases. In de Scandinavische landen daarentegen worden de officiële statistieken grotendeels gebaseerd op gekoppelde gegevens uit administraties en registraties, die over het algemeen van een zeer goede kwaliteit zijn. Met deze werkwijze kan de Volktelling gesimuleerd worden. De burgers hoeven er niet opnieuw voor benaderd te worden, de statistische bureaus maken slechts efficiënt gebruik van gegevens waarover de overheid voor administratieve doeleinden toch al beschikt. De koppelbaarheid van de gegevens brengt in potentie een veel rijker arsenaal aan cross-sectionele en longitudinale analyses met zich dan bij een traditionele ('echte') Volktelling mogelijk is. Het concept van een PUM verliest daarmee tegelijk voor de meest geavanceerde gebruikers aan zeggingskracht.

Sinds 1987 kent Statistics Denmark een *on site* of *in-house researcher arrangement*. Buiten Kopenhagen is er een kleine dependance in Århus. In de loop van 2001 hebben Statistics Denmark, het Ministerie van Onderzoek en de onderzoekswereld een contract getekend dat leidde tot de oprichting van een speciale *Research Service Unit* binnen *Statistics Denmark* met de speciale taak om de toegang van onderzoekers tot microdata te verbeteren (<http://www.micro2122.scb.se/The%20Danish%20System%20for%20Access%20to%20Microdata.doc>) Het budget voor deze unit bedraagt 6 miljoen Deense Kronen of 800 000 Euro's. Een deel daarvan wordt besteed aan *hardware*, voor *remote access*. Het grootste deel gaat op aan de personele bezetting, met zes statistici, zes automatiseerders en een administratieve kracht. De unit

zet de data voor een onderzoeksproject volgens het *Need to know* principe geanonimiseerd klaar voor de onderzoekers. Voor elk project wordt 8 uur kosteloos beschikbaar gesteld, daarna moet er een prijs per uur betaald worden. Per 1 augustus 2003 hebben 55 Deense onderzoeksgroepen zich voor dit arrangement aangemeld. De toegang moet altijd lopen via een Deense onderzoeksgroep en binnen het Deense recht. Het gebruik van de microdata verloopt via een veilig en *dedicated* netwerk. De analyse opdrachten op de bestanden worden gelogd. De toegang tot bedrijfsmicrodata is beperkt. Data moeten ten minste een jaar oud zijn en onderzoeksafdelingen van particuliere ondernemingen krijgen er geen toegang toe. Voor het overige vindt er op de microdata bewust geen toepassing van (geautomatiseerde) *disclosure avoidance* technieken plaats. De Deense statistici vertrouwen geheel op de overeenkomst die de onderzoekers met hen sluiten en de automatisering. De *remote access* faciliteit is zo succesvol dat de *on site* mogelijkheid gesloten wordt, behalve een paar werkplekken voor buitenlandse onderzoekers. Het Deense model is succesvol: het aantal onderzoekers dat zo van Deense microdata gebruik gemaakt is tussen 1997 en 2003 gestegen van 70 tot 200. Het is ook zo succesvol dat het grotere buurland Zweden overweegt het te introduceren.

9. Het gras aan de andere kant van de landsgrenzen

Een aantal grote internationale projecten maken eveneens gebruik van *remote access* maar op een andere manier dan in het boven geschetst Scandinavische model.

- In het international PUMs (iPUMS) project van Ruggles en McCaa worden Volkstellingsmicrodata (steekproeven) uit zoveel mogelijk landen uit zoveel mogelijk jaren ontsloten. Momenteel zijn er microdata voor 48 miljoen personen tussen 1960 en 1999 uit Colombia, Frankrijk, Kenia, Mexico, de VS, en Vietnam in ondergebracht. Deze dataset wordt snel uitgebreid. De (meta)data worden zoveel mogelijk geharmoniseerd voordat ze voor door een licentie toegelaten onderzoekers beschikbaar komen. Deze kunnen de data vervolgens downloaden. (<http://www.ipums.org>)
- In de *Luxembourg Income Study* worden microdata met betrekking tot persoons- en huishoudensinkomens bijeengebracht en geharmoniseerd voor internationaal vergelijkend onderzoek. Vergelijkbaar is de *Luxembourg Employment Study*. In deze beide projecten kunnen toegelaten onderzoekers analyseopdrachten in SAS, SPSS of Stata aanbieden die in Luxemburg achter een *firewall* verwerkt worden. Zowel de analyseopdrachten als het analyseresultaat worden gescreend op beveiligingsrisico's, geautomatiseerd en zo nodig ook door de LIS/LES staf zelf. (<http://www.lisproject.org>)

Zowel iPUMS als LIS/LES worden gesubsidieerd door de Amerikaanse *National Science Foundation* en andere fondsen zodat er een infrastructuur gecreëerd wordt met voor de individuele onderzoeker relatief beperkte kosten.

De metadata moeten vanwege de internationale groep aan gebruikers goed vastgelegd en gedocumenteerd zijn. Dat is een relatief eenvoudige taak omdat het om een overzichtelijke set aan begrippen en variabelen gaat. In tegenstelling daarmee is het Scandinavische model veel minder gestandaardiseerd. De meeste onderzoeksgroepen hebben specifieke eigen wensen, en de data binnen de Scandinavische statistische bureaus zijn aanmerkelijk complexer en flexibeler dan de eenvoudige bestanden die bij iPUMS en LIS/LES horen.

De in de vorige alinea genoemde projecten maken gebruik van microdata vanuit de officiële statistiek maar zijn gelanceerd door academische onderzoekers. In aanvulling daarop zijn er ook initiatieven vanuit de aanbodkant van de officiële statistiek. Door inspanningen van de noordelijke statistische bureaus in de Europese Unie, waaronder het CBS, is er een commissieverordening die het in principe mogelijk maakt dat voor een viertal verzamelingen van microdata, de ontsluiting door Eurostat geregeld wordt. De context van comitologie en nationale veto's gevoegd bij de huidige Eurostat crisis lijken echter een snelle feitelijke implementatie in de weg te staan.

Buiten deze praktische vooruitgang is de internationale statistische context ook van belang voor de meningsvorming en de uitwisseling van ervaringen. In Noord-Amerika komen conferenties en publicaties over de toegang tot microdata met grote regelmaat voor, en is er een relatief intensieve interactie tussen de officiële statistiek en de academische gemeenschap. Zo kwam in 2001 het boek CDDA uit met steun van het Amerikaanse *Census Bureau* maar volgeschreven door

anderen dan officiële statistici. Uit de Lage Landen was Jules Theeuwes één van de redacteurs. De conferentie waar dit boek gepresenteerd werd, was zo gewild dat er een aantal inschrijvingen niet gehonoreerd kon worden (zie verder voor de VS <http://www.amstat.org/comm/cmtepc/index.cfm>) Sinds 1994 organiseert Eurostat met een zekere regelmaat internationale conferenties over statistische geheimhouding: van Dublin (1994) tot en met Lissabon (2002). Zie bijvoorbeeld het meest recent: http://europa.eu.int/comm/eurostat/Public/datashop/print-catalogue/EN?catalogue=Eurostat&product=KS-PB-03-001-__-C-EN. Hetzelfde geldt voor de *Conference of European Statisticians*, die onder de VN *Economic Commission for Europe* ressorteert en behalve de West-Europese ook de Oost-Europese en Noord-Amerikaanse statistici bijeen brengt (zie <http://www.unece.org/stats/archive/02.01.e.htm>). Tijdens de laatste *Conference of European Statisticians* presenteerde Trewin uit Australië een paper over *Access to microdata – issues, organisation and approaches*. Ook gegeven Trewins andere functies, als voorzitter van het *International Statistical Institute* en organisator van het ISI congres in Australië over twee jaar, kunnen we er zeker van zijn dat de verbetering van de toegang tot microdata hoog op de statistische agenda zal blijven staan de komende jaren.

10. Conclusies en aanbevelingen

Welke conclusies trek ik uit het bovenstaande?

- Er is sprake van een internationale trend van facilitering van wetenschappelijk onderzoek met microdata vanuit de officiële statistiek. Onderzoekers accepteren

in toenemende mate de statistische beveiliging en de juridische en technologische voorzieningen. De statistici zien het grote maatschappelijk en wetenschappelijk nut van een veilig gebruik van 'hun' officiële microdata. De officiële statistiek ziet de veilige beschikbaarstelling van microdata in toenemende mate als haar eigen, legitieme taak.

- Het gebruik van die microdata is niet symmetrisch. Nederlandse onderzoekers zullen vaker naar Amerikaanse of internationale microdata grijpen dan buitenlandse onderzoekers naar Nederlandse microdata. De indruk bestaat dat dat ten koste van de nationale beleidsrelevantie van het onderzoek gaat.
- Goede data zijn duur. De laatste Amerikaanse Volkstelling kostte 65 dollar per huishouden, dat is meer dan 6 miljard dollar in totaal. Terugvertaald naar Nederlandse verhoudingen zou een Volkstelling uitgaande van die bedragen twee extra CBS jaarbudgetten vergen. Mede dankzij goede en voor de statistiek toegankelijke nationale registraties zijn we in Nederland dus relatief goedkoop uit.
- Ook de beschikbaarstelling van microdata is duur. Die kosten moeten gezamenlijk worden opgebracht door de overheid, de onderzoeksorganisaties, het statistisch bureau en de gebruikers. Veel van de goede ontwikkelingen over de grens zijn gestimuleerd door een Ministerie van Onderzoek (Denemarken), een *Social Science Research Council* (Canada) en een *National Science Foundation* (VS), waarbij de bijdrage van de eindgebruikers variabel is. In Nederland heeft NWO daar in en na 1994 een zeer welkome bijdrage aan geleverd.
- De kosten van beschikbaarstelling van microdata zijn

in kleine landen relatief hoog vanwege allerlei *economies of scale*. Zo is het aantal variabelen belangrijker voor de kosten van de documentatie dan het aantal records. *Ceteris paribus* kan er in kleinere landen minder dan in grote landen.

- Een land als Nederland dat internationaal wil meetellen, moet dan ook investeren in zijn data-infrastructuur. Uit verschillende rapporten (<http://www.knaw.nl/publicaties/pdf/20031077.pdf> van de SWR bijvoorbeeld) en het concept Meerjarenprogramma van het CBS (zie in het bijzonder de pagina's 24 en 25 van <http://www.cbs.nl/nl/organisatie/corporate-informatie/MJP-2004-2008-def-concept-06-08-03.pdf>) blijkt daartoe de nodige bereidheid. Die zal echter ook gehonoreerd moeten worden door onze budget autoriteiten, de Ministers van OCW en EZ.

Mijn toekomstbeeld voor de sociaal-wetenschappelijke (VT-)microdata is meerledig:

- Standaardmicrodata sets moeten via *remote access* voor alle groepen onderzoekers toegankelijk zijn, à la iPUMS en LIS: gestandaardiseerde *remote access*. Deze sets zullen selecties uit het SSB zijn, door het CBS te standaardiseren in overleg met de gebruikers. Gemakshalve, maar ook met het oog op longitudinaal onderzoek, denk ik dan aan uit het registratieve SSB afgeleide microdata samples die sterk lijken op de huidige microdata uit de sociale sample surveys, maar ook aan de historische VT data. Een voordeel voor het CBS is dat het versiebeheer naar de gebruikers toe eenvoudiger wordt en dat uit de analyses en *loggings* veel geleerd kan worden over de interesses van het onderzoeksveld en zwakke plekken in de beveiliging.

- Waar deze standaarddata niet toereikend zijn, is voor veel data de Deense optie het eerste alternatief: maatwerk *remote execution*. Er wordt voor de onderzoekers een dataset klaargemaakt en klaargezet. Het voordeel voor de onderzoekers bestaat eruit dat zij hun werk van huis uit kunnen doen en dat zij altijd over de meest actuele versie van de data zullen kunnen beschikken.
- Voor de meest gedetailleerde data moeten de *on site* laboratoria gehandhaafd en uitgebouwd worden. Daarbij ware te denken aan veel van de meest gedetailleerde regionale specificaties maar ook aan panelbestanden, bedrijfsgegevens en andere gegevens. Het werken met deze data kan echter vergemakkelijkt worden door kleine, eventueel synthetische samples beschikbaar te stellen zodat onderzoekers thuis hun onderzoeksanpak alvast kunnen uittesten.

Zo'n meerledig beeld is standaard aan het worden in de officiële statistiek: zie het overzicht in bijlage 3 van de producten van *Statistics Canada* door Chuck Humphrey van de Universiteit van Alberta.

Bestaat voor zo'n toekomstbeeld *commitment* aan de kant van het CBS? Mijn eerste en laatste antwoord is beslist positief: kijk naar onze performance over de afgelopen periode, kijk ook naar het concept meerjarenprogramma voor de komende jaren.

Over het gebied tussen droom en daad moet ik intussen echter wel enkele opmerkingen maken:

- Voor zijn eigen taak en dus ook voor de daarvan afgeleide microdata blijft het CBS buitengewoon afhankelijk van zijn berichtgevers, of het nu om respondenten of om registratiehouders gaat. Alleen al

daarom zal het CBS altijd enigszins aan de voorzichtige kant moeten blijven.

- Voor zijn budget is en blijft het CBS afhankelijk van de rijksoverheid. Alleen al om die reden is brede steun voor het CBS noodzakelijk. Tegenover een afnemend budget staat een toename aan taken voor het CBS, die voor een groot deel een wettelijk verplicht karakter heeft. U moet daarbij denken aan een versnelling van de conjunctuurinformatie en een verbreding van de macro- en micro-economische statistieken tot de dienstensector.
- De opeenstapeling van bezuinigingen (het CBS is in twee decennia terug van 3.500 naar 2.200 plaatsen) heeft per saldo een enigszins verlamdend effect op organisatie, expertise en cultuur binnen het CBS. De wil tot innovatie is soms groter dan het vermogen, zeker gezien bovengenoemde druk op het budget en het programma.

Mijn laatste conclusie is dan ook dat het CBS en de academische wereld alles te winnen hebben bij een intensivering van de betrekkingen.

Noten

- (1) Joris Nobel is adviseur van de Raad van Bestuur van het CBS en speciaal belast met de wetgeving op statistisch terrein en het beveiligingsbeleid. jnbl@cbs.nl
- (2) De voor het sociaal-wetenschappelijk onderzoek relevante verschillen, zijn de volgende:
 - De herformulering dat passende (beveiligings)maatregelen moeten worden genomen met betrekking tot het gebruik van de data;

- de toevoeging dat naast de verstrekking van data er nu ook toegang toe verleend kan worden;
- het geheel nieuwe artikel 42a, dat zeer specifieke bepalingen bevat met het oog op de toegankelijkheid van de doodsoorzakengegevens van het CBS voor epidemiologisch onderzoek vanwege het amendement Blok.

Referenties

Doyle, P., J. Lane, J. Theeuwes, and L. Zayatz, red., 2002, Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies. Elsevier Science (te vinden op: http://www.elsevier.com/wps/find/bookdescription.cws_home/622129/description#description)

Eerste Kamer, 2003, De Wet op het Centraal Bureau voor de Statistiek <http://eerstekamer.cust.pdc.nl/9324000/1/j9vvgh5ihkk7kof/vgh3jl2jwib0/f=y.pdf>

Holvast, J., 1999, (<http://www.computable.nl/artikels/archief9/d44ag9gx.htm>).

Hooghiemstra, Th., 2002, Privacy bij wetenschappelijk onderzoek en statistiek. Kader voor een gedragscode (Den Haag: CBP) (http://www.cbweb.nl/documenten/rap_2002_Privacy_en_statistiek.htm)

Sociaal-Wetenschappelijke Raad (2003), Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek. (<http://www.knaw.nl/publicaties/pdf/20031019.pdf>).

Bijlage 1. De relevante artikelen van de CBS-wet

Artikel 37

1. De door de directeur-generaal in het kader van de uitoefening van de taken ter uitvoering van deze wet ontvangen gegevens worden uitsluitend gebruikt voor statistische doeleinden.
2. De in het eerste lid bedoelde gegevens worden niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.
3. De in het eerste lid bedoelde gegevens worden slechts zodanig openbaar gemaakt dat daaraan geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling kunnen worden ontleend, tenzij, ingeval het gegevens met betrekking tot een onderneming of instelling betreft, er een gegronde reden is om aan te nemen dat bij de betrokken onderneming of instelling geen bedenkingen bestaan tegen de openbaarmaking.

Artikel 41

1. In afwijking van artikel 37 kan de directeur-generaal op verzoek, ten behoeve van statistisch of wetenschappelijk onderzoek, een verzameling van gegevens met betrekking tot het gebruik waarvan passende maatregelen zijn genomen om herkenning van afzonderlijke personen, huishoudens, ondernemingen of instellingen te voorkomen, verstrekken aan een dienst, organisatie of instelling als bedoeld in het tweede lid, dan wel daartoe toegang verlenen.
2. Een verzameling van gegevens als bedoeld in het eerste lid kan worden verstrekt, dan wel daartoe kan toegang worden verleend aan:
 - a. een universiteit in de zin van de Wet op het hoger onderwijs en wetenschappelijk onderzoek;

- b. een bij wet ingestelde organisatie of instelling voor wetenschappelijk onderzoek;
- c. bij of krachtens de wet ingestelde planbureaus;
- d. de communautaire en nationale instanties voor de statistiek van de lidstaten van de Europese Unie;
- e. onderzoeksafdelingen van ministeries en andere diensten, organisaties en instellingen, voor zover daartoe instemming van de CCS is verkregen.

Artikel 42

De directeur-generaal willigt een verzoek als bedoeld in artikel 41 slechts in, indien de verzoeker naar het oordeel van de directeur-generaal voldoende maatregelen heeft getroffen om te voorkomen dat de verzameling van gegevens voor andere doeleinden dan statistisch of wetenschappelijk onderzoek wordt gebruikt.

Artikel 42a

1. In afwijking van artikel 37 kan de directeur-generaal op verzoek, ten behoeve van wetenschappelijk onderzoek op het terrein van de volksgezondheid, gegevens verstrekken die het CBS ten behoeve van statistisch onderzoek op grond van artikel 12a van de Wet op de lijkbezorging heeft verzameld, van personen die in een wetenschappelijk onderzoek waren betrokken.
2. De directeur-generaal willigt een verzoek als bedoeld in het eerste lid slechts in, indien de betrokkene tot een zodanige verstrekking zijn uitdrukkelijke toestemming heeft gegeven dan wel, indien de betrokkene tot een zodanige verstrekking geen uitdrukkelijke toestemming heeft kunnen geven, voor zover verzoeker voldoende aantoont dat:
 - a. het vragen van toestemming bij leven van de betrokkene in redelijkheid niet mogelijk was of kon worden gevergd;

- b. niet gebleken is dat de betrokkene bij leven bezwaar heeft gemaakt tegen het verwerken van diens persoonsgegevens ten behoeve van wetenschappelijke onderzoek;
 - c. het onderzoek een algemeen belang dient;
 - d. het onderzoek niet zonder de desbetreffende gegevens kan worden uitgevoerd;
 - e. het onderzoek overigens voldoet aan daaraan redelijkerwijs te stellen eisen.
3. De directeur-generaal kan nadere voorwaarden verbinden aan een verstrekking krachtens het eerste lid.
 4. De artikelen 41, tweede lid, en 42 zijn van overeenkomstige toepassing.
 5. Onze Minister stelt, in overeenstemming met Onze Minister van Volksgezondheid, Welzijn en Sport, gehoord de CCS en het College bescherming persoonsgegevens, nadere regels met betrekking tot de in het eerste lid bedoelde bevoegdheid.

Bijlage 2. De inleiding tot de US Bureau of the Census PUMs publicatie

Introduction

OVERVIEW

Public use microdata sample files are ASCII files which contain individual records of the characteristics for a sample of people and housing units. Information which could identify a household or an individual is excluded in order to protect the confidentiality of respondents. Within the limits of the sample size, the geographic detail, and the confidentiality protection, these files allow users to prepare virtually any tabulation they require.

WHAT ARE MICRODATA?

Microdata are the individual records which contain information collected about each person and housing unit. They include the census basic record types, computerized versions of the questionnaires collected from households, as coded and edited during census processing. The Census Bureau uses these confidential microdata in order to produce the summary data that go into the various reports, summary files, and special tabulations. Public use microdata samples are extracts from the confidential microdata taken in a manner that avoids disclosure of information about households or individuals. For Census 2000, the microdata are only available to the public through the Public Use Microdata Sample (PUMS) products.

PROTECTING CONFIDENTIAL INFORMATION

All data released (in print or electronic media) by the Census Bureau are subject to strict confidentiality measures imposed by the legislation under which our data are collected: Title 13, U.S. Code. Responses to the questionnaire can be used only for statistical purposes, and Census Bureau employees are sworn to protect respondents' identities.

Because of the rapid advances in computer technology since 1990 and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through enhanced disclosure limitation techniques. At the same time, the Census Bureau recognizes the data user's need for characteristic detail and geographic specificity. Hence, there are two sets of files: one that provides a fuller range of detailed characteristics (the 1-percent files) and one that provides greater geographic detail but less characteristic detail (the 5-percent files).

Confidentiality is protected, in part, by the use of the following processes: data-swapping, topcoding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables.

Dataswapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. Swapping is applied to individual records and, therefore, also protects microdata. Top-coding is a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category.

Geographic population thresholds prohibit the disclosure of data for individuals or housing units for geographic units with population counts below a specified level.

Age perturbation, that is, modifying the age of household members, is required for large households (households containing ten people or more) due to concerns about confidentiality.

Detail for categorical variables is collapsed if the number of occurrences in each category does not meet a specified national minimum threshold.

U.S. Census Bureau, Census 2000

Bron: <http://www.census.gov/prod/cen2000/doc/pums.pdf>

Bijlage 3. A Continuum of Access to Statistics Canada Products

Service:	Statistics Canada Website	Depository Service Program	Data Liberation Initiative	Product Sales & Customized Tabulations	Remote Job Submission	Research Data Centres
Who is eligible:	General public	General public through DSP libraries	Post-secondary academic staff & students	Individuals or organizations	Researchers with accepted proposals	Approved researchers
Under what conditions:	Available on the Internet at www.statcan.ca	Available on site	Restricted to teaching and research purposes	Contract between STC and individual or organization	Arrangement between STC & researcher	SSHRC peer review & deemed STC employee status
Products:	- The Daily - Canadian Statistics - Census - Statistical profiles of Canadian communities - Downloadable publications	- Paper publications - Electronic publications, including priced down-loadable publications & select CDs	Standard data products: aggregate databases, public use microdata files, and geography files	Tables from confidential files that are specially processed by STC for a fee; also access to specialized data-bases	"Dummy" or synthetic files to build analysis setups that must then be submitted to STC for processing	Confidential data files from the longitudinal surveys begun in the 1990's, such as the NPHS, SLID, NLSCY
Notes:	Warning: some services on the web site are fee-based	Some DSP libraries provide off-site access to authenticated users	Interface to CANSIM I and Trade Analyzer available through CHASS (University of Toronto)	Specialized databases include CANSIM II and the Trade Analyzer	Available for only select surveys; this service most developed for the NPHS	Applications can be submitted through the SSHRC web site

Chuck Humphrey, University of Alberta, Continuum of Access
Bron: <http://www.statcan.ca/english/Dli/continuumofaccess.htm>

Hoe ver gaat de deur van het CBS open?

Een reactie op het paper van J. Nobel

K. Tijdens (1)

De deur van het CBS is de laatste jaren steeds verder open gegaan. Een voorbeeld daarvan is de mogelijkheid op de website van het CBS tabellen met behulp van StatLine samen te stellen. CEREM (microdata van bedrijven) is ook een mooi voorbeeld van het verruimde data-aanbod. Maar ook met de oprichting van het WSA zijn meer databestanden aan derden ter beschikking gesteld. Het openen van de deuren komt niet alleen tot uiting in een groter aanbod van data, maar ook in de informatie die het CBS verstrekt om de data en de databestanden te gebruiken, de zogenaamde gebruikersdagen. Tot slot noemt Tijdens nog de cd-rom waarop de gegevens van 200 jaar Volkstellingen staan.

Hoewel een steeds grotere stroom gegevens van het CBS zijn weg naar buiten vindt, zouden de wetenschappelijk onderzoekers de stroom graag nog groter zien worden. Dat komt doordat de gewenste schaal van dataverzameling steeds groter is geworden, maar ook omdat het budget dat de wetenschappelijke onderzoekers bij de universiteit voor dataverzameling kunnen aanspreken bij lange na niet toereikend is. Daarnaast is het ook erg lastig om elders financiering voor dataverzameling te regelen. NWO heeft bij reguliere projecten slechts een klein deel van de begroting gereserveerd voor de financiering

van gegevensverzameling. Op Europees niveau, bij de EU, is wel geld beschikbaar. Het probleem daar is dat als voorwaarde voor financiering wordt gesteld dat ook andere landen in de EU soortgelijke data moeten verzamelen, zodat er internationaal vergelijkbare analyses uitgevoerd kunnen worden. Op commercieel vlak is de interesse om geld te steken in dataverzameling niet erg groot. Niet alleen de beperkingen in eigen middelen vormen een reden waarom volgens Tijdens onderzoekers graag toegang tot de gegevens van het CBS willen hebben. Het CBS heeft al de beschikking over erg veel (mooie) gegevens. Probleem is alleen of wetenschappers ook gemakkelijk toegang tot de gegevens kunnen krijgen. Bij *data sharing* spelen zoals Nobel al in zijn paper heeft duidelijk gemaakt een aantal problemen. Een bijkomend probleem volgens Tijdens is dat de gegevens die het CBS verzamelt niet primair voor wetenschappelijk onderzoek worden verzameld, waardoor het aantal variabelen vaak beperkt is. Tijdens denkt dat voorgestelde *data sharing* ondersteund zou kunnen worden als het CBS meer voorlichting en informatie biedt via gebruikersdagen en de eigen website. Informatie die Tijdens ook toejuicht betreft die informatie waarmee landenvergelijkingen gemaakt kunnen worden, de *international classifications*, historische reeksen en het variabelen query systeem. Voor wat betreft het SSB hoopt Tijdens dat de gestandaardiseerde en maatwerk *remote access* en *on site access* worden voortgezet. Op deze manier zouden ook meer tabellen beschikbaar komen met beschrijvende data van context variabelen (zoals postcode, regio, beroep, sector, cohorten). Dergelijke geaggregeerde data kan goed gebruikt worden in analyses van eigen dataverzamelingen van de onderzoeker.

Noten

- (1) Kea Tijdens is socioloog en promoveerde in 1989. Zij is sinds 1982 werkzaam bij de Universiteit van Amsterdam, op dit moment als universitair hoofddocent en onderzoekscoördinator van AIAS. Zij heeft onder meer onderzoek verricht naar flexibilisering bij bedrijven, gelijke beloning tussen mannen en vrouwen, afspraken in CAO's, internationale vergelijkingen van deeltijdarbeid in Europa, uitbesteding van huishoudelijke arbeid en kenmerken van herintreedsters op de arbeidsmarkt. Zij neemt deel aan de dataverzameling van lonen en arbeidsuren van werknemers via de Loonwijzer-website.

Discussie

Uit de papers is gebleken dat het CBS beschikt over erg veel gegevens, dat het allerlei initiatieven neemt om de buitenwereld ook toegang tot die informatie te geven en dat de gegevensstroom de laatste jaren steeds beter zijn weg naar buiten heeft gevonden. Maar wat vinden de gebruikers van de gegevens die ter beschikking worden gesteld: zijn het er voldoende en zijn ze gedetailleerd genoeg? En als dat niet zo is wat zou er dan moeten gebeuren? Allereerst krijgen de mensen in de zaal de gelegenheid te reageren op wat zij gehoord hebben.

De deelnemers onderschrijven het commentaar van de referenten. Zij vinden dat wetenschappers toegang tot 'alle' informatie van het CBS moeten hebben. Dat betekent niet alleen dat zij Dronkers' voorstel voor het ter beschikking stellen van een 1 % steekproef een goed idee vinden – hoewel sommigen vinden dat dat ten koste van regionale detaillering gaat -, maar ook dat zij in staat moeten worden gesteld te achterhalen hoe de verschillende bestanden gekoppeld en bewerkt worden, zodat zij het 'ene cijfer' dat het CBS samenstelt, kunnen reconstrueren. Zo geeft Schulte Nordholt in zijn paper aan dat er 40 tabellen voor de Volkstelling samengesteld zijn. De deelnemers zijn benieuwd of het ook mogelijk is om nog andere dan deze 40 tabellen op te vragen. Schulte Nordholt laat weten dat dat in principe niet kan, omdat daar een weegtechniek – het herhaald wegen – voor nodig is, die nog al tijdsintensief is. Wat wel zou kunnen is dat wetenschappers *on site* bij het CBS komen werken om zo de gewenste tabellen samen te stellen.

De CBS'ers maken ook van de discussiegelegenheid gebruik om in te gaan op een aantal vragen en opmerkingen die de referenten ter sprake brachten. Op de vraag of het wel zo verstandig is om naar één cijfer te streven, antwoordt Schulte Nordholt dat het erg van de gebruikers afhangt. Wetenschappers willen toegang tot alle gegevens hebben om daar zelf combinaties uit samen te stellen en analyses op uit te voeren. Maar beleidsmakers willen dat helemaal niet, die willen geen inzicht in de achtergrond of bewerkingen op de data uitvoeren; ze willen hupklare brokken. En het CBS moet met beide gebruikers rekening houden.

In aansluiting op hetgeen Gierveld meldde meent een deelnemer dat Dronkers' voorstel om niet langer individuen maar huishoudens in de analyse centraal te stellen niet haalbaar is, omdat de statistische eenheid huishouden immers erg veranderlijk is. Dat is Everaers met haar eens, maar hij stelt dat een dataverzamelmethode voor, die een oplossing biedt voor dit probleem. Dan kan ook aan de wens van Dronkers voldaan worden. Het CBS gebruikt deze methode al bij het Inkomenspanelonderzoek (IPO). Het IPO bevat gegevens over 75.000 kernpersonen en 150.000 overige huishoudensleden. Daarnaast kunnen via het GBA ook familiale relaties achterhaald worden. Daarin is een verwijzing naar de vader en de moeder opgenomen, waardoor het mogelijk is om iets over het herkomstmilieu van iemand te zeggen. Verder wordt nog even stil gestaan bij de reactie van Hartog. De variabele opleiding is redelijk betrouwbaar geschat. Daarvoor is de Enquête Beroepsbevolking (EBB) gebruikt, waarin de gehele onderwijsloopbaan wordt gereconstrueerd. Op grond daarvan is de hoogst afgeronde

opleiding bepaald. Om te kunnen beoordelen of het loon in een situatie van monopolie hoger is dan in een situatie van volledige mededinging moeten bedrijven op dat kenmerk geclassificeerd worden en die informatie is helaas (nog) niet beschikbaar.

Wat de vergelijkbaarheid van de LSO-en betreft: onderzoekers kunnen het vierde blok gebruiken, dat is goed vergelijkbaar met vorige jaren. Er zijn alleen kleine populatie verschillen.

Tot slot antwoordt Bakker in reactie op het referaat van Tesser dat met het SSB wel degelijk meer kan dan wat het SCP tien jaar geleden al deed. De meerwaarde van het SSB ten opzichte van SPVA (een steekproef) is dat met het SSB een zeer gedetailleerde indeling naar herkomst beschikbaar is, maar ook dat de samenhang met andere variabelen, zoals demografische gegevens en data uit voorgaande jaren, mogelijk is. Het SPVA is alleen rijker aan variabelen. Volgens Tesser gaat het beleidsmakers daar wel om. Zij hebben veel variabelen / gegevens nodig om beleid te kunnen maken.

Ook in de forumdiscussie staat de confrontatie tussen de wensen en de mogelijkheden centraal. Onder leiding van Erik de Gier discussiëren Bart Bakker (CBS), Paul de Beer (UvA), Jaap Dronkers (UIE), Joris Nobel (CBS) en Kea Tijdens (UvA) met elkaar. In zijn referaat pleit Kees de Hoog voor een minicensus. Dat lijkt een beetje op het voorstel van Dronkers om een 1 % steekproef ter beschikking te stellen. De Gier is benieuwd wat de panelleden van een minicensus vinden.

Tijdens denkt niet dat een minicensus erg nuttig is. Ze ziet wel wat in het ter beschikking stellen van een 1 % of 5 % steekproef, maar daar is geen minicensus voor

nodig. Bakker vraagt zich ook af of een minicensus veel nut op zal leveren. In een minicensus gaat het om 30.000 waarnemingen en dat aantal is veel te weinig om detaillering naar regionaal niveau mogelijk te maken. Bovendien is Bakker van mening dat het SSB / de virtuele Volkstelling eigenlijk al een soort minicensus is, omdat naast de registergegevens ook de EBB gekoppeld is. De minicensus is dan een met registergegevens verrijkte EBB. Volgens Nobel zou een minicensus wel zijn nut kunnen hebben. Bijvoorbeeld als deze gehouden wordt voor een bepaalde gemeente en wanneer de gegevens die met de census verzameld zijn vergeleken worden met het SSB om deze te valideren. In een minicensus ziet De Beer niets, maar ook niet in een 1 % of 5 % steekproef. Hij wil dat gebruikers toegang hebben tot het complete SSB om multivariate analyses uit te kunnen voeren. Daar is een steekproef niet geschikt voor, want die is onnauwkeurig. Om die multivariate analyses te kunnen maken zou De Beer graag meer combinatiemogelijkheden in Statline willen hebben. Ook Dronkers zou willen dat Statline uitgebreid wordt en wel zo dat er meer variabelen in opgenomen worden.

Maar is een 1 % steekproef wel mogelijk? Ja, de met registergegevens verrijkte EBB is zo'n steekproef, aldus Bakker. Deze steekproef voldoet volgens Dronkers niet aan zijn eisen. De EBB is alleen een arbeidsmarktbestand. Dat is ook het voornaamste kritiekpunt van Dronkers op het SSB. Het is veel te veel gericht op werk en arbeid. Dronkers wil ook informatie hebben over de andere domeinen van de levensloop (zorg, leren en rusten). Daarop antwoordt Bakker dat de EBB wel een huishoudenssteekproef is, die te verrijken is met gegevens uit registerbestanden.

Er worden veel wensen geuit, maar wie moet die in 'vervulling' laten gaan. Is dat een taak van het WSA of van de data intermediairs? De Beer vindt dat het CBS daarin het voortouw moet nemen. Het beschikt immers al over de data. De Beer koppelt echter wel een voorwaarde aan de uitvoering door het CBS: het moet in dialoog blijven met de wetenschappers, zodat regelmatig geïnventariseerd wordt wat de wensen van de gebruikers zijn. Hij vindt dat de tabellen in huidige publicaties niet getuigen van inzicht in de interesses van onderzoekers.

Het CBS bij monde van Nobel, geeft te kennen dat het die taak ook graag zou uitvoeren, en dat het ook geen bezwaar is dat de gebruikers vanaf het begin meekijken met wat het CBS doet, maar er moet wel geld voor beschikbaar zijn. Dronkers vraagt zich of het meedenken van de gebruikers wel iets uithaalt, want de voor wetenschappers meest interessante bestanden kennen de grootste onthullingsrisico's. Bakker meldt dat de data in principe toegankelijk zijn, maar de modaliteit waarmee verschilt. Als alle onderzoekers en erkende instellingen *on site* en via *remote execution* toegang zouden krijgen tot de complete bestanden van het CBS, dan zou Dronkers zijn wens voor de 1 % steekproef laten vallen. Dat vindt Tijdens niet verstandig. Er kunnen gevallen zijn waarbij een 1 % steekproef toch nuttig kan zijn. Nobel wil nog graag aan de discussie toevoegen dat het belangrijk is rekening te houden met de wensen van alle gebruikers, dus niet alleen de wetenschappers, maar ook de beleidsmakers. Ook Bakker beaamt dat. Het CBS heeft een grote hoeveelheid data. Bij het CBS bestaat dan ook de behoefte om met het beleid en de wetenschap samen vast te stellen hoe de toegankelijkheid van de data op een wijze kan worden

ingericht die nuttig is voor alle gebruikers.

Tot slot biedt De Gier de wetenschappelijke panelleden de gelegenheid aan te geven wat zij van het SSB en de virtuele Volkstelling willen. Dronkers is erg blij met de mogelijkheid via *remote execution* toegang tot de data te hebben, maar wat hem via die weg ter beschikking komt, moet, vindt hij, ook herhaald kunnen worden en een relatie met het verleden hebben. De Beer bepleit nogmaals dat hij vindt dat het onderhoud van enquêtes even veel aandacht dient te krijgen als onderhoud van het SSB. Tijdens ziet, naast het creëren van de mogelijkheid van *remote execution*, graag grote tabellen op StatLine verschijnen.

Ter afsluiting vat Bakker de wensen nog eens samen:

1. Wens: uitbreiding van StatLine met *giant tables*.
Mogelijk: ja, en het CBS is bovendien al op weg in de goede richting. In het komende jaar worden de uitkomsten van de verslagjaren 1999-2002 via StatLine ontsloten in twaalf *giant tables*.
2. Wens: bestanden die internationaal te vergelijken zijn.
Mogelijk: ja, in het kader van iPUMS kunnen die bestanden beschikbaar worden gesteld; omdat er weinig veiligheidsrisico's aan verbonden zijn. Overigens zal de internationale vergelijkbaarheid beperkt zijn vanwege de verschillende peilmomenten en definities van variabelen in verschillende landen.
3. Wens: enquêtegegevens verrijkt met register-informatie
beschikbaar stellen via *remote execution* voor sociaal-wetenschappelijke onderzoekers.
Mogelijk: ja, mits er voldoende financiën ter beschik-

king worden gesteld om daarvoor een productielijn op te zetten. Daarnaast is te overwegen ook standaardbestanden beschikbaar te stellen voor allerlei gebruikers, waaronder ook de pers.

4. Wens: meer gebruikersdagen organiseren in de toekomst.

Mogelijk: ja. Het CBS zou dat dan graag naar thema organiseren.

5. Wens: gebruik van oude gegevens.

Mogelijk: Het CBS zou beleid moeten formuleren voor het gebruik van deze oude gegevens. Wat kan ter beschikking worden gesteld en wat niet? Wanneer zijn de veiligheidsrisico's zo verkleind dat een bestand volledig ter beschikking kan worden gesteld?

6. Wens: duidelijkheid over wie welke versie van de databestanden beheert.

Mogelijk: ja. In principe is het te regelen dat de bestanden die ter beschikking worden gesteld bevroren worden per jaargang. Dat betekent dat er geen data aan toegevoegd, verwijderd of veranderd worden. Dat maakt de analyses van verschillende sociale wetenschappers controleerbaar, omdat de analyses gerepliceerd kunnen worden op exact dezelfde data.

Gebruikte afkortingen

ABR	Algemeen Bedrijfsregister	NIWI	Nederlands Instituut voor Wetenschappelijke Informatiediensten
ABW	Administratie algemene bijstandswet	ONS	Office for National Statistics
AKT	Arbeidskrachtentelling	POLS	Permanent Onderzoek Leefsituatie
CCO	Consumenten ConjunctuurOnderzoek	RDC	Research Data Center
CCS	Centrale Commissie voor de Statistiek	SEP	Sociaal Economische Panelonderzoek
CRIHO	Centraal register ingeschrevenen hoger onderwijs	SES	Sociaal-economische Status
CWI	Centrum voor Werk & Inkomen	SSB	Sociaal Statistisch Bestand
EBB	Enquête Beroepsbevolking	SWR	Sociaal-Wetenschappelijke Raad
EWL	Enquête werkgelegenheid en lonen	VT	volkstelling
GBA	Gemeentelijke BasisAdministratie	VZA	Verzekerdenadministratie werknemers
IOAW	Wet Inkomensvoorziening Oudere en gedeeltelijk Arbeidsongeschikte werkloze Werknemers	VRD	Vullen Reference Database (software)
IOAZ	Wet Inkomensvoorziening Oudere en gedeeltelijk Arbeidsongeschikt gewezen Zelfstandigen	WAO	Wet op de arbeidsongeschiktheidsverzekering
IIT	Instellingen, Inrichtingen en Tehuizen	WBO	Woningbehoeftenonderzoek
ILO	International Labour Office	WBP	Wet Bescherming Persoonsgegevens
IOW	Interdepartementaal Overleg Wetenschapsbeleid	WGBO	Wet op de Geneeskundige Behandelovereenkomst
IPO	Inkomenspanelonderzoek	WSA	Wetenschappelijk Statistisch Agentschap
iPums	Integrated Public Use Microdata Series International	WSF	Wet studiefinanciering
ISCO	International Standard Classification of Occupations	WSW	Wet Sociale Werkvoorziening
ISCED	International Standard Classification of Education	WW	Werkloosheidswet
IT	Informatie Technologie	ZZP	zelfstandigen zonder personeel
LES	Luxembourg Employment Study		
LIS	Luxembourg Income Study		
LSO	Loonstructuuronderzoek		
NAW	naam, adres en woonplaats		

Deelnemers

Dr. B. Bakker	CBS	F. Linder	CBS
Prof. dr. E.J. Bartelsman	Vrije Universiteit	Dr. R. Luijkx	Universiteit van Tilburg
Prof dr. P. de Beer	AIAS	J. van Maarsseveen	(ex-)CBS
J. Besseling	TNO	P. van der Meer	RUG
Prof. dr. R.J. Bosker	RUG	Dr. C.H. Mulder	Universiteit van Amsterdam
Dr. J.M. Dagevos	SCP		ITS, KUN
Dr. J. van Dijk	Rijksuniversiteit Groningen	L. Mulder	Universiteit Maastricht
		Drs G. Notten	CBS
Prof. dr. P.A. Dykstra	NIDI	Drs. J. Nobel	SISWO
Dr. P.K. Doorn	NIWI	Drs. L. Putman	NIWI
Prof dr. J. Dronkers	IUE, Italië	Drs. L. Schreven	CBS
Dr. P. Everaers	CBS	Drs. E. Schulte Nordholt	CBS
Drs. L. Faase	SER	Prof dr. J. van Sinderen	Katholieke Universiteit Nijmegen
Dr. D.J. Fouarge	OSA	Dr. J.P.J.M. Smits	TNO
J. Geerits	ITS, KUN		BiZa
Prof dr. E. de Gier	SISWO	Dr. P.G.W. Smulders	Universiteit van Amsterdam
Prof dr. J. Gierveld	NIDI	P. Tesser	ROA
Prof. dr. N.D. de Graaf	Katholieke Universiteit Nijmegen	Prof dr. K. Tijdens	EIM
	Erasmus Universiteit, ISEO		
Dr. S.M. Groeneveld	UvA	Dr. R. van der Velden	
	Universiteit van Amsterdam, SEO	P. Vroonhof	
Prof dr. J. Hartog	Universiteit Wageningen		
Dr. A. Heyma	CBS		
	Katholieke Universiteit Nijmegen		
Dr. C. de Hoog	OSA		
J. Jonker	CBS		
Dr. G.L.M. Kraaykamp	CBS		
Dr. M.J.M. Kerkhofs			
Dr. J. Latten			

Auteurs

Koos Arts is projectleider Eindintegratie en Sociaal Statistisch Bestand op het CBS. kars@cbs.nl

Bart Bakker is programmamanager Eindintegratie en Sociaal Statistisch Bestand op het CBS. bbkr@cbs.nl

Paul de Beer is bijzonder hoogleraar arbeidsverhoudingen (Henri Polak-leerstoel) aan de Universiteit van Amsterdam en verbonden aan AIAS en De Burcht. Hij heeft in het verleden veel kwantitatief arbeidsmarktonderzoek gedaan op basis van microdata van het CBS (m.n. de EBB) en is lid geweest van de Commissie van Advies voor Arbeidsstatistiek die onder andere heeft geadviseerd over het SSB.

Jaap Dronkers is als hoogleraar verbonden aan het European University Institute, San Domenico di Fiesole (FI), Italy. jaap.dronkers@iue.it. personal homepage: <http://www.iue.it/Personal/Dronkers>

Pieter Everaers is directeur Sociale en Ruimtelijke Statistiek op het CBS en als zodanig verantwoordelijk voor zowel de Volkstelling als het Sociaal Statistisch Bestand. pevs@cbs.nl

Jenny Gierveld is hoogleraar aan de Vrije Universiteit en Honorary Fellow van het NIDI.

Joop Hartog is hoogleraar economie aan de UvA (FEE) en heeft onderzoek verricht op het terrein van arbeidsmarkt en onderwijs. Momenteel houdt hij zich bezig met onderzoek naar de economische aspecten van migratie.

Cees de Hoog is hoogleraar gezinssociologie en gezinsbeleid aan de Wageningen Universiteit, leerstoelgroep Sociologie van Consumenten en Huishoudens Betrokkenheid. Lid van het coördinatieteam gezinssociologische censusmonografieën 1970-1979, gebaseerd op de Volkstelling 1971. SISWO / Landbouwniversiteit

Jan Latten is demograaf en de woordvoerder voor de pers van het CBS. jltm@cbs.nl

Joris Nobel is adviseur van de Raad van Bestuur van het CBS en speciaal belast met de wetgeving op statistisch terrein en het beveiligingsbeleid. jnbl@cbs.nl

Jacques van Maarseveen is historicus en oud CBS-er. Hij houdt zich onder meer bezig met de historie van de Volkstellingen en de digitalisering daarvan. smrn@cbs.nl

Eric Schulte Nordholt is projectleider van de Volkstelling 2001 op het CBS. Daarnaast is nauw betrokken geweest bij het produceren van het Loonstructuuronderzoek vanuit het Sociaal Statistisch Bestand. esle@cbs.nl

Paul Tesser is hoofd van de afdeling Strategie, Onderzoek en Communicatie (SOC) van de Directie Coördinatie Integratiebeleid etnische Minderheden (DCIM) van het Ministerie van Justitie. Van hieruit is hij in hoge mate geïnteresseerd in onderzoeksgegevens die zicht geven op de mate van integratie van minderheden. Tot voor kort was hij medewerker van het SCP belast met de

jaarlijkse Rapportage Minderheden. De SCP-rapportage beschrijft en analyseert de positie van de minderheden in de samenleving en brengt deze in verband met het gevoerde beleid. Van daaruit nauw betrokken bij de data-infrastructuur over minderheden.

Kea Tijdens is socioloog en promoveerde in 1989. Zij is sinds 1982 werkzaam bij de Universiteit van Amsterdam, op dit moment als universitair hoofddocent en onderzoekscoördinator van AIAS. Zij heeft onder meer onderzoek verricht naar flexibilisering bij bedrijven, gelijke beloning tussen mannen en vrouwen, afspraken in CAO's, internationale vergelijkingen van deeltijdarbeid in Europa, uitbesteding van huishoudelijke arbeid en kenmerken van herintreedsters op de arbeidsmarkt. Zij heeft een eigen dataverzameling van lonen en arbeidsuren van werknemers via de Loonwijzer-website.

Rolf van der Velden is hoofd van de afdeling Onderwijs en Beroepsloopbanen bij ROA.