Micro-integratie

09

Bart F. M . Bakker

Statistische Methoden (09001)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2009

Verklaring van tekens

.	= gegevens ontbreken
*	= voorlopig cijfer
x	= geheim
–	= nihil
–	= (indien voorkomend tussen twee getallen) tot en met
0 (0,0)	= het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	= een cijfer kan op logische gronden niet voorkomen
2007–2008	= 2007 tot en met 2008
2007/2008	= het gemiddelde over de jaren 2007 tot en met 2008
2007/'08	= oogstjaar, boekjaar, schooljaar enz., beginnend in 2007 en eindigend in 2008
2005/'06–2007/'08	= oogstjaar, boekjaar enz., 2005/'06 tot en met 2007/'08

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

Colofon

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

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

Omslag
TelDesign, Rotterdam

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

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

Internet
www.cbs.nl

ISSN: 1876-0333

6016509001 X-37

Inhoudsopgave

1. Inleiding op het thema Micro-integratie

1.1 Algemene beschrijving

Micro-integratie houdt in dat gegevens van statistische eenheden op individueel niveau worden gekoppeld met als doel het samenstellen van betere informatie dan met de afzonderlijke bronnen mogelijk is. Micro-integratie dient te worden onderscheiden van verrijking van individuele gegevens. Bij dit laatste ontstaat geen meerwaarde maar wordt uitsluitend additionele informatie aan de individuele eenheden toegevoegd (Al & Thijssen, 2003).

Gegevens uit registers en enquêtes die voor CBS-statistieken worden gebruikt kunnen allerlei fouten bevatten, zoals representatie- en meetfouten. De doelstelling van micro-integratie is het verbeteren van de kwaliteit van de statistische uitkomsten die uit verschillende bronnen afkomstig zijn door opsporing en correctie van de representatie- en meetfouten, zodat:

- de validiteit en betrouwbaarheid van de statistische uitkomsten voldoende is,
- over een verschijnsel maar één cijfer wordt gepubliceerd (de zogenaamde ééncijfergedachte),
- gegevens uit verschillende bronnen kunnen worden gecombineerd en correcte themaoverstijgende uitkomsten worden gepubliceerd,
- correcte longitudinale uitkomsten worden gepubliceerd.

Representatiefouten bestaan eruit dat de populatie waarover uitspraken wordt gedaan niet volledig en/of selectief wordt beschreven. Daarbij kan sprake zijn van over- en onderdekking van de populatie. Door middel van het onderdeel *completeren* worden deze representatiefouten opgespoord en gecorrigeerd. Overigens kan bij het afleiden van elementen die samen de populatie vormen ook sprake zijn van meetfouten. Voor correctie van deze fouten worden naast completeren ook andere micro-integratietechnieken gebruikt.

Meetfouten bestaan eruit dat de kenmerken van de populatie niet juist worden beschreven. Dit kan allerlei verschillende oorzaken hebben. Door gebruik te maken van informatie uit verschillende bronnen kan vertekening in één of meerdere bronnen worden opgespoord en worden gecorrigeerd. Voor het corrigeren van meetfouten op conceptueel niveau wordt *harmoniseren* ingezet. Voor het corrigeren van meetfouten op dataniveau wordt *correctie voor meetfouten* toegepast.

Voor het signaleren van representatie- en meetfouten wordt gezocht naar inconsistenties. Voor het opsporen van representatiefouten wordt de data uit een bron of een combinatie van bronnen idealiter vergeleken met een referentiebestand waarin een complete opsomming van de elementen uit de doelpopulatie is opgenomen. Komen die niet overeen, dan is sprake van onder- en/of overdekking. In veel gevallen ontbreekt echter dergelijke referentiedata. Deze moeten dan tijdens het

proces van micro-integratie worden opgebouwd. Meetfouten komen aan het licht als er inconsistenties zijn tussen de scores op de variabelen uit de gecombineerde bronnen.

Consistentie is een kwaliteitsaspect dat los gezien kan worden van de vertekening of de betrouwbaarheid. Een statistische uitkomst wint aan kwaliteit puur door het feit dat zij consistent is (gemaakt) met andere statistische uitkomsten, zodat een samenhangende beschrijving van onderwerpen gegeven kan worden.

Alleen als het oplossen van inconsistenties van belang is voor de te publiceren gegevens, wordt naar een manier gezocht om deze op te lossen. Micro-integratie wordt meestal toegepast in een situatie waarin er verschillende administratieve bronnen beschikbaar zijn voor de beschrijving van hetzelfde thema of dezelfde bron voor verschillende verslagperiodes. Deze zijn meestal min of meer integraal, dat wil zeggen dat de populatie van de administratie er in zijn geheel in is opgenomen. Het is echter niet noodzakelijk om micro-integratie te beperken tot integrale bronnen. Ook enquêtedata en registers kunnen worden geïntegreerd op microniveau en heel soms zelfs data van twee enquêtes met een voldoende overlap van respondenten. Er is echter maar weinig ervaring met de correctie van representatie- en meetfouten bij de integratie van enquêtedata en registers. We ontlenen dan ook uitsluitend voorbeelden aan de registerpraktijk.

1.2 Afbakening en relatie met andere thema's

Micro-integratie onderscheidt zich van gaafmaken (het thema "Controle en correctie/imputatie"), omdat de beslisregels die gedefinieerd worden voor het corrigeren van fouten van velerlei soort over verschillende bronnen heen gaan. Bij gaafmaken gaat het om beslisregels die binnen één bron worden toegepast.

Micro-integratie onderscheidt zich van het thema 'Macro-integratie', doordat bij micro-integratie de gegevens op microniveau consistent worden gemaakt, terwijl dat bij macro-integratie niet het geval is.

Er is lang gedebatteerd over de vraag of micro-integratie nu een methode is of niet. Micro-integratie is voor een belangrijk deel niet generaliseerbaar, in de zin dat hiervoor kennis van de bronnen en de maatschappelijke werkelijkheid nodig is. Deze zorgen ervoor dat steeds specifieke micro-integratieregels worden opgesteld en toegepast. Dit kun je zien als kleine stapjes in een totaal verwerkingsproces van data. De integratieregels kunnen wel in een aantal onderscheiden categorieën worden ondergebracht, zoals hierboven is beschreven. Er is besloten om deze categorieën te beschrijven en in de Methodenreeks op te nemen.

1.3 Plaats in het statistisch proces

Uitgangspunt voor de plaats in het statistisch proces zijn de processen van de CBS-brede waarnemingstrategie en het Sociaal Statistisch Bestand (SSB). Globaal worden de volgende processtappen onderscheiden:

- Specificatie van de output op basis van de informatiebehoefte.

- Verzameling registergegevens waarmee de informatiebehoefte wordt gedekt.
- Opbouw steekproefkader en steekproeftrekking.
- Aanvullend enquêteren.
- Koppelbaar maken van de registers en de enquêtes.
- Completeren: corrigeren voor onder- of overdekking van de doelpopulatie.
- Harmoniseren en minimalisatie van meetfouten.
- Wegen, consistent herhaald wegen.
- Tabelleren, aggregeren, publiceren

De eerste twee punten behoeven geen statistische methoden maar goed accountmanagement. De opbouw van het steekproefkader en de steekproeftrekking vallen onder ‘Steekproeftheorie’. Aanvullend enquêteren valt onder het thema ‘Verzamelen van gegevens’. Koppelbaar maken valt onder het thema ‘Koppelen’. ‘Wegen, consistent herhaald wegen’ valt onder ‘Steekproeftheorie’. Voor het laatste punt is het thema ‘Grafieken’ beschikbaar.

Dan blijven er twee onderdelen over:

- Completeren: corrigeren van onder- of overdekking van de doelpopulatie.
- Harmoniseren en minimalisatie van meetfouten.

Deze twee onderdelen komen in dit thema ‘Micro-integratie’ aan de orde.

1.4 Definities

Begrip	Omschrijving
SSB	Sociaal Statistisch Bestand
Micro-integratie	Een methode waarin gegevens van statistische eenheden op individueel niveau worden gekoppeld met als doel het samenstellen van betere informatie dan met de afzonderlijke bronnen mogelijk is. Bij de kwaliteit gaat het om de validiteit, betrouwbaarheid en de (themaoverstijgende en longitudinale) consistentie. Micro-integratietechnieken zijn completeren, harmoniseren en corrigeren voor meetfouten.
Completeren	Met behulp van een stelsel van regels corrigeren voor de onder- en overdekking van de doelpopulatie van een onderzoek
Harmoniseren	Met behulp van een stelsel van regels gegevens onder één noemer brengen. De noemer is het concept dat men wil meten.
Corrigeren voor meetfouten	Met behulp van een stelsel van regels inconsistenties in data oplossen zodat de kwaliteit wordt verbeterd.

2. Completeren

2.1 Representatiefouten

Bij de start van een onderzoek wordt een doelpopulatie bepaald. Deze doelpopulatie is de populatie waarover in het onderzoek gegevens worden verzameld en uitkomsten worden gepresenteerd. De mate waarin de onderzochte populatie afwijkt van de doelpopulatie wordt de totale representatiefout genoemd. In het geval van onderzoek op basis van administratieve databronnen kan de totale representatiefout bestaan uit de volgende onderdelen:

- onder- en overdekking van de doelpopulatie doordat de doelpopulatie van de administratieve gegevens afwijkt van de doelpopulatie van het statistisch onderzoek.
- onderdekking doordat niet alle elementen uit de doelpopulatie van de administratieve gegevens in de administratie zijn opgenomen.
- overdekking doordat elementen die niet tot de doelpopulatie van de administratieve gegevens behoort er (nog) in voorkomen.
- onderdekking veroorzaakt door het niet kunnen koppelen van elementen uit de administratieve gegevens die wel behoren tot de doelpopulatie.
- overdekking veroorzaakt door de (mis)koppeling van elementen uit de administratieve gegevens die niet tot de doelpopulatie behoren.

Enkele voorbeelden illustreren dit.

Voorbeeld 1. De doelpopulatie van het onderzoek betreft de studenten in het hoger onderwijs die deel uitmaken van de Nederlandse bevolking op 1 oktober 2008. De meest geschikte bron om de populatie af te bakenen betreft het zogenaamde Eéncijferbestand Hoger Onderwijs. Daarbij gaat het echter om het in Nederland bekostigde onderwijs. Daardoor worden de studenten die in het buitenland studeren (voornamelijk in België of Duitsland) en studenten die particulier hoger onderwijs volgen gemist. Dan is er dus sprake van onderdekking van de doelpopulatie. Ook zijn er studenten in het bestand opgenomen die in het buitenland wonen en in Nederland studeren. Die zorgen voor overdekking van de doelpopulatie.

Voorbeeld 2. De doelpopulatie van het onderzoek betreft de werkzoekenden die geen baan hebben ultimo september 2008. De meest geschikte administratieve bron daarvoor is de administratie van de CWI. Omdat er sprake is van administratieve vertraging worden mensen die inmiddels een baan hebben gevonden niet direct als werkzoekende uitgeschreven. In de praktijk blijkt tevens dat naarmate de werkloosheid hoger is, de administratieve vertraging ook groter is: bij meer cliënten is er domweg minder tijd om die vertraging in te lopen. Deze administratieve vertraging leidt in dit geval tot overdekking van de doelpopulatie.

Voorbeeld 3. De doelpopulatie van het onderzoek bestaat uit de personen die behoren tot de Nederlandse bevolking op 1 januari 2008 en die in 2007 verdacht zijn van een misdrijf. De meest geschikte administratieve bron hiervoor is de combinatie van de structuurtelling van de Gemeentelijke Basis Administratie op 1 januari 2008 en het HerKenningsdienstSysteem (HKS) van de politie. Vanwege koppelingsproblemen en fouten is er echter sprake van zowel over- als onderdekking. Er zijn koppelingen gemist omdat de identificerende informatie te beperkt was om een koppeling te leggen. Daarnaast zijn vanwege verschrijvingen of andere redenen koppelingen gelegd tussen gegevens van twee verschillende personen.

De methode van het completeren corrigeert voor de totale representatiefout. Daarnaast worden de meetfouten die ontstaan door miskoppelingen gecorrigeerd bij correctie voor meetfouten. Het is voldoende dat wordt gecorrigeerd voor over- en onderdekking, ongeacht de achterliggende oorzaak daarvan.

2.2 Correctie voor onderdekking

Er zijn verschillende methodes om voor onderdekking te corrigeren. We onderscheiden:

- het combineren van verschillende bronnen waardoor een complete opsomming van alle elementen uit de populatie wordt gerealiseerd.
- de wel waargenomen elementen van een gewicht te voorzien waardoor in de uitkomsten de totale doelpopulatie wordt gerepresenteerd.
- een vorm van unit-imputatie waardoor in de uitkomsten de totale doelpopulatie wordt gerepresenteerd.

De tweede en derde manier om te corrigeren worden uitgebreid besproken in andere delen van de Methodenreeks, te weten respectievelijk in Bethlehem (2007) en Israëls, Pannekoek en Schulte Nordholt (2007). Daarom bespreken we hier alleen de vorm van completeren waarin verschillende bronnen worden gecombineerd.

Bij de correctie voor onderdekking wordt gestart met het zo precies mogelijk definiëren van de doelpopulatie. Zo kan in het eerste voorbeeld van onderdekking de doelpopulatie worden gedefinieerd als: “de in Nederland wonende studenten die hoger onderwijs volgen”. De onderdekking die met het Eéncijferbestand Hoger Onderwijs ontstaat kan gedeeltelijk worden gecompleteerd door gebruik te maken van de informatie uit een andere bron: de registratie van de studenten die studiefinanciering krijgen voor een studie in het hoger onderwijs. Ook voor het particulier onderwijs en onderwijs dat in België of Duitsland wordt genoten kan studiefinanciering toegekend zijn. Weliswaar zullen niet alle studenten die ontbreken in het Eéncijferbestand Hoger Onderwijs hiermee worden opgespoord, maar wel een groot deel.

In het geval van de verdachten van misdrijven wordt een andere manier toegepast om de onderdekking te verkleinen. Het is bekend dat er gemiste koppelingen zijn door het verhuizen van mensen. Daar komt nog eens bovenop dat de verdachten van

misdrijven er belang bij hebben om niet met hun juiste personalia te worden geregistreerd. Dit leidt vaak (tijdelijk) tot het niet koppelbaar zijn van hun gegevens. Ook hier is namelijk sprake van administratieve vertraging: de juiste personalia worden in veel gevallen toch na verloop van tijd in de registratie opgenomen. Met terugwerkende kracht worden dan de oudere gegevens van dezelfde verdachte alsnog hieraan gekoppeld. Door gebruik te maken van de meest recente jaargang van het HKS van de politie kunnen de oudere jaargangen worden gecompleteerd.

Het is echter niet altijd mogelijk om voor onderdekking te corrigeren. Er moeten immers (administratieve) gegevens aanwezig zijn om een dergelijke correctie uit te voeren. Die gegevens zijn er niet altijd. Het is echter wel mogelijk om de omvang van de onderdekking te benaderen door de totale populatie te schatten met behulp van een enquête. Daarmee wordt de fout die wordt veroorzaakt door de onderdekking in beeld gebracht. Deze kan als kwaliteitskenmerk aan de gegevens worden toegevoegd.

2.3 Correctie voor overdekking

Overdekking van de statistische doelpopulatie wordt gecorrigeerd door de elementen die niet tot de populatie behoren te verwijderen. Om dit te kunnen doen dienen deze elementen eerst te worden geïdentificeerd. Een exacte operationele definitie van de doelpopulatie is daarvoor noodzakelijk. Een voorbeeld daarvan is: de personen die in het CWI-bestand 2008 (versie 1 april 2010 die gecorrigeerd is voor de administratieve vertraging tot en met het jaar 2009) zijn opgenomen en op 1 oktober 2008 op de variabele WERKZOEK de waarde “ja” hebben. Daarbij kunnen de volgende situaties worden onderscheiden:

- De definitie van de doelpopulatie is te operationaliseren binnen de jaargangen van één bron. In dat geval is deze handeling relatief eenvoudig. In het bovenstaande geval is de administratieve vertraging die voor overdekking zorgde verwijderd in de versie van 1 april 2010.
- De definitie van de doelpopulatie kan niet binnen één bron worden geoperationaliseerd, maar er is een andere bron nodig om de elementen te verwijderen die voor overdekking zorgen. De overdekking die veroorzaakt wordt door vertraagde uitschrijving bij de CWI van mensen die al een baan hebben gevonden, wordt gecorrigeerd door het CWI-bestand op individueel niveau te koppelen met een bestand met gegevens over werknemers. De elementen die zowel in het werknemersbestand zijn opgenomen als in het CWI-bestand hebben reeds werk gevonden en dienen te worden verwijderd. De administratieve vertraging in het CWI-bestand wordt daarmee gecorrigeerd.

Het is echter niet altijd mogelijk om voor overdekking te corrigeren. Dit is bijvoorbeeld het geval voor overdekking die veroorzaakt wordt door miskoppelingen. Het aantal miskoppelingen kan weliswaar worden geschat (zie daarvoor bijvoorbeeld Arts, Bakker en Van Lith, 2000), maar tot hoeveel overdekking dat precies leidt is daarmee nog niet vastgesteld. Een deel van de miskoppelingen kan immers wel degelijk tot de doelpopulatie behoren. Omdat er in

een dergelijk geval gegevens van twee verschillende personen aan elkaar zijn gekoppeld leidt dat tot fouten die vergelijkbaar zijn met meetfouten. Bijkomend nadeel is dat de elementen die de overdekking veroorzaken niet kunnen worden aangewezen.

2.4 Toepasbaarheid

Voor het correct toepassen van de methode is kennis nodig van het te onderzoeken terrein en de bronnen die gebruikt worden in het onderzoek. De methode is toe te passen door beslisregels te formuleren en te programmeren.

Het signaleren van onderdekking is geen eenvoudige zaak. Alleen als er een volledig kader is dat als referentiebestand kan dienen, is het mogelijk om vast te stellen of er sprake is van totale dekking van de doelpopulatie. Meestal ontbreekt een dergelijk referentiebestand echter. In veel gevallen wordt dit bestand opgebouwd terwijl wordt gecompleteerd. Voor de controle of de methode van completeren ook daadwerkelijk geleid heeft tot dekking van de doelpopulatie is aanvullend onderzoek nodig. Zo kan bijvoorbeeld daar gericht een enquête voor worden gebruikt.

Ook overdekking is niet altijd goed te traceren. Overdekking door administratieve vertraging is dikwijls na verloop van tijd of door middel van een andere bron te corrigeren. Voor overdekking veroorzaakt door miskoppelingen is dat echter niet het geval.

Als aangetoond is of een sterk vermoeden bestaat dat er sprake is van ontoelaatbare hoeveelheden onder- of overdekking, kan worden besloten om deze gegevens niet voor de desbetreffende doelpopulatie te gebruiken. Normen hiervoor zijn moeilijk te geven. Wat voor het ene onderzoek nog acceptabel is, kan voor een ander onderzoek niet acceptabel zijn.

3. Harmoniseren en corrigeren voor meetfouten

3.1 Meetfouten

Meetfouten bestaan eruit dat de kenmerken van de populatie niet juist worden beschreven. We spreken dan van vertekening, *bias* in het Engels. In registers en enquêtes komen allerlei verschillende meetfouten voor. Door Groves et al. (2004) worden de meetfouten in enquêtes in drie groepen ondergebracht: fouten die gemaakt worden in de operationalisering van concepten, fouten die gemaakt worden in de waarneming in de enquête en fouten die gemaakt worden in de verwerking van de enquête. De fouten in administratieve gegevens komen grotendeels overeen met die in enquêtes, omdat vaak een of meerdere enquêtetechnieken worden gebruikt om administratieve databases te vullen. Zo kunnen (elektronische) belastingformulieren beschouwd worden als een vorm van Computer Assisted Web Interviewing (CAWI) of Paper and Pencil Interviewing (PAPI), is het intakegesprek bij de CWI een vorm van Computer Assisted Personal Interviewing (CAPI), etc. (Bakker, Linder en Van Roon, 2008).

Er zijn echter wel enkele belangrijke verschillen te noemen. Het eerste verschil is dat de omvang van de fouten onder meer afhangt van het belang dat de registerhouder hecht aan de kwaliteit van de informatie. Als de informatie noodzakelijk is voor het goed uitvoeren van de kerntaken van de registerhouder dan zal deze de informatie beter controleren dan de informatie die er voor hem weinig toe doet. Bekend is bijvoorbeeld dat de begin- en einddatums van banen die ontleend worden aan de belastingaangiftes van personen niet goed van kwaliteit zijn. De reden daarvoor is dat de hoogte van het te betalen belastingbedrag daarvan niet afhankelijk is en er wordt door de Belastingdienst dan ook niet op gecontroleerd. Een voorbeeld waarin wel goed wordt gecontroleerd is de informatie uit de Onderwijsnummerbestanden die gebruikt worden voor de bekostiging van het onderwijs. Daarop wordt een accountantscontrole uitgevoerd. We kunnen gevoeglijk aannemen dat de kwaliteit na accountantscontrole beter is dan ervoor: scholen hebben belang bij een groot aantal leerlingen, terwijl de accountants van het Ministerie van Onderwijs, Cultuur en Wetenschap belang hebben bij een kleiner aantal leerlingen. Door de gezonde spanning tussen beide partijen worden de grenzen van de definities en de toepassing daarvan goed bewaakt (Bakker, 2003, blz. 129-130).

Een tweede verschil is dat er een onderscheid gemaakt moet worden tussen de meetfouten die gemaakt worden bij de operationalisering en meting van het administratieve concept en de mate waarin deze administratieve begrippen te vertalen zijn in statistische begrippen. In de conceptualisering en operationalisering van de statistische begrippen moet vaak een vertaalslag worden gemaakt van deze administratieve informatie. Ook daarbij kunnen fouten worden gemaakt.

Een derde verschil is de administratieve vertraging. Dit is een meetfout die specifiek is voor administratieve gegevens: een gebeurtenis wordt (veel) later geregistreerd dan hij plaatsvond. Daardoor worden deze gebeurtenissen op een peilmoment dat ligt voor het moment van registratie niet meegeteld. Dat hoeft overigens niet altijd problematisch te zijn voor de statistische uitkomsten. Wanneer de omvang van de gebeurtenissen en de mate van administratieve vertraging in een bron constant is in de tijd, dan zijn de standgegevens die daaruit worden gepubliceerd correct.

Meetfouten komen aan het licht als er inconsistenties zijn tussen de variabelen uit de gecombineerde bronnen. Er is onder meer sprake van inconsistenties in de kenmerken van de eenheden als:

- Variabelen uit twee bronnen hetzelfde verschijnsel beschrijven, maar op individueel niveau een verschillende uitkomst hebben. Bijvoorbeeld als in de ene bron staat dat iemand geslacht man heeft, terwijl die in de andere bron geslacht vrouw heeft.
- Er een logische relatie is tussen variabelen die bij toepassing een onjuiste uitkomst oplevert. Bijvoorbeeld dat een jaarloon ongelijk is aan de lonen die in de 12 maanden van dat desbetreffende jaar zijn verdiend.
- Standen en stromen niet op elkaar aansluiten: de stand op peilmoment t plus de mutaties tussen t en t+1 moet de stand op t+1 opleveren. Er is bijvoorbeeld sprake van een inconsistentie als de stand van de bevolking op 1-1-2007 plus alle geboortes en immigraties min alle sterftes en emigraties niet de stand van de bevolking op 1-1-2008 oplevert.
- Er een onmogelijke overgang van de ene naar de andere situatie is. Bijvoorbeeld een overgang van gehuwd naar ongehuwd. Dit is onmogelijk: er kan slechts een overgang plaatsvinden naar gescheiden of verweduwd.
- Er een onwaarschijnlijke samenloop van situaties is. Bijvoorbeeld dat iemand op hetzelfde tijdstip twee fulltime banen heeft, of op hetzelfde tijdstip een fulltime baan en een volledige WW-uitkering.

Consistentie is een kwaliteitsaspect dat los gezien kan worden van de vertekening of de betrouwbaarheid. Een statistische uitkomst wint aan kwaliteit wanneer die consistent is (gemaakt) met andere statistische uitkomsten, zodat een samenhangende beschrijving van onderwerpen gegeven kan worden.

Een bijzonder geval van consistentie is de longitudinale consistentie. Daarmee bedoelen we dat de gegevens over een tijdsperiode consistent zijn (gemaakt) zodat de omvang van de mutaties of veranderingen voor de (sub)populaties goed worden geschat. Longitudinale consistentie wordt onder meer aangetast door administratieve vertraging. Een voorbeeld daarvan is dat de huwelijken van allochtonen die in het land van herkomst zijn gesloten soms pas twee jaar na afsluiting ervan worden geregistreerd. Als deze administratieve vertraging niet wordt verwerkt worden de mutaties in burgerlijke staat voor deze categorie onderschat.

Voor het corrigeren van meetfouten op conceptueel niveau wordt *harmoniseren* ingezet. Voor het corrigeren van meetfouten op dataniveau wordt *correctie voor meetfouten* toegepast.

3.2 Harmoniseren

Onderzoek start met de vraag welke begrippen gemeten dienen te worden. Deze worden in eerste instantie conceptueel gedefinieerd. Voorbeelden kunnen worden ontleend aan de begrippen uit het zogenaamde Bureau of Standards (CBS, 2009). Daarna moeten ze worden geoperationaliseerd. De operationalisering bestaat uit het opschrijven van de exacte afleiding van het begrip uit de informatie die in de databron aanwezig is. In feite komen de afleidingsregels erop neer dat precieze criteria worden aangegeven om het begrip te meten. Bij enquêtes gaat het om een operationalisering in vraagstellingen, bij administraties is daarin veel minder vrijheid van definitie: de kwaliteit van de operationalisering hangt af van de mogelijkheden die de informatie in het register biedt. Harmoniseren houdt in dat gegevens uit één of uit verschillende bronnen onder één noemer worden gebracht.

Voorbeeld 1. Het begrip “Baan” is gedefinieerd als “Een door een werkzame persoon bezette arbeidsplaats. Een werkzame persoon kan meerdere banen naast elkaar hebben. In dat geval wordt van een hoofd- en een bijbaan gesproken” (StatLine, 2009). In deze definitie gaat het om een relatie tussen het bedrijf dat de arbeidsplaats biedt en de persoon die de arbeidsplaats bezet. In de tweede helft van de definitie wordt ingegaan op het mogelijk samengaan van meerdere banen, maar dat doet hier niet ter zake. De definitie wordt geoperationaliseerd door in de daarvoor geschikte administratieve gegevens combinaties te zoeken van personen (geïdentificeerd met een geanonimiseerd persoonsnummer, de RIN-persoon) en bedrijven die de arbeidsplaatsen bieden (geïdentificeerd met een Bedrijfsidentificatienummer, de zogenaamde BE_ID). Ieder record waar een dergelijke combinatie in voorkomt staat voor een baan met een begin- en een einddatum (Arts en Hoogteijling, 2002; Bakker & Arts, 2003).

Bij de Verzekerdenadministratie van het UWV staan de bedrijven geregistreerd onder het zogenaamde Bedrijfsvereniging & Aansluitnummer (BVA) en de werknemers daarbinnen met hun Burgerservicenummer. Nadat de BSN’s vanwege beveiligingsredenen zijn omgezet in een RIN-persoon, worden de BVA-codes omgezet in BE_ID-codes. In de FiBase, gegevens van de bedrijven voor de loonheffing, zijn de bedrijven geregistreerd onder een Loonbelastingnummer en de werknemers met een BSN. Door ook de loonbelastingnummers om te zetten in BE_ID-codes worden de gegevens geharmoniseerd en hebben de banen in deze twee bronnen dezelfde betekenis. Hetzelfde wordt gedaan met de informatie uit de Enquête Werkgelegenheid en Lonen, met dien verstande dat deze enquête voor een deel van de bedrijven integraal is en voor een ander deel op steekproefbasis. Bij het steekproefdeel wordt de steekproef getrokken uit het ABR waarin de bedrijfseenheden zijn geïdentificeerd als BE_ID.

Sinds 1 januari 2006 zijn de bovenstaande registraties grotendeels opgegaan in de zogenaamde Polisadministratie en is het voldoende dat de bovenstaande regels worden toegepast op de records uit deze administratie.

Voorbeeld 2. Het “Behaald opleidingsniveau” is gedefinieerd als “Het niveau volgens de Standaard onderwijsindeling (SOI) van de hoogste met succes gevolgde opleiding”. Voor de toepassing van de SOI is een lijst van opleidingen ontwikkeld waarbij iedere onderscheiden opleiding voorzien is van een zogenaamd opleidingsnummer. Van deze opleidingsnummers kan vervolgens de SOI worden afgeleid. Voor de operationalisering van het opleidingsniveau met behulp van administratieve gegevens worden verschillende bronnen met elkaar gecombineerd, zowel registers als enquêtes. Ieder van die bronnen bevat grotendeels unieke informatie over de opleidingen die mensen gedurende hun leven hebben gevolgd. Ieder van die bronnen heeft weer een eigen codesysteem waarin is vastgelegd welke opleiding is gevolgd en eventueel is afgerond. Door voor ieder van deze codesystemen een conversietabel te maken naar de opleidingsnummers uit de SOI wordt deze informatie geharmoniseerd.

Harmoniseren bestaat voor het grootste deel uit het formuleren van een aantal beslisregels waarin zo precies mogelijk wordt aangegeven hoe een bepaald begrip wordt gemeten, gegeven de beschikbare informatie in de data. Daarvoor is zowel kennis noodzakelijk over de wetenschappelijke en maatschappelijke betekenis van het begrip als kennis van de informatie die in de bron(nen) aanwezig is.

3.3 Corrigeren voor overige meetfouten

Eerst wordt voor de oplossing van inconsistenties harmoniseren toegepast. Als er daarna nog verschillen overblijven, wordt beoordeeld wat voor ieder van de variabelen de beste bron is. Bij de bepaling van de beste bron dient rekening gehouden te worden 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. Een variabele in een bron kan op sommige punten sterk en op andere punten zwak zijn. Bijvoorbeeld: het jaarloon van werkenden bij de Rijksoverheid in bron A is slechter dan hetzelfde gegeven in bron B, maar voor de overige sectoren zijn de jaarlonen in bron A beter van kwaliteit. Door aan een bepaalde bron de voorkeur te geven (eventueel onder een of meerdere condities) kan het jaarloon worden bepaald (Bakker, 2003).

Als 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 nemen van een gemiddelde waarde van twee bronnen. Eveneens kunnen beslisregels worden geformuleerd om te forceren dat een relatie tussen verschillende variabelen klopt. Daarbij hangt het van de kwaliteit van de bronnen af welke informatie wordt aangepast.

3.4 Toepasbaarheid

Zowel harmoniseren als corrigeren van meetfouten worden met behulp van beslisregels uitgevoerd. Voor het juist formuleren van deze beslisregels is kennis nodig van de maatschappelijke werkelijkheid en de inhoud en kwaliteit van de informatie uit de beschikbare bronnen.

Het signaleren van meetfouten is geen eenvoudige zaak. Hetzelfde geldt voor het formuleren van beslisregels. In een aantal gevallen zal het gaan om regels waarmee foute records worden gecorrigeerd, maar waarmee een klein aantal juiste records wordt omgezet in een onjuist record. Om de omvang van deze twee te bepalen is aanvullend onderzoek nodig. Daar kan bijvoorbeeld gericht een enquête voor worden gebruikt.

4. Afsluiting

Het is belangrijk om te documenteren welke waarden zijn gecorrigeerd en welke beslisregels daarvoor zijn gebruikt. Dit is nodig om het proces reproduceerbaar te maken. Dergelijke documentatie is ook voor gebruikers van de resulterende microbestanden van belang. Bij harmoniseren is het eveneens van belang om de gebruikte begrippen goed te definiëren, de beslisregels te bepalen en deze in de documentatie op te nemen.

5. Literatuur

Al, P. en Thijssen, J. (2003), Bespiegelingen over het waarom, de mogelijkheden en beperkingen van micro-integratie in de sociale statistieken, In: J. Nobel, S. Algera, M. Biemans en P. van der Laan (red.), *Gedacht en gemeten*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen, blz. 112-122.

Arts, K., Bakker, B.F.M. en Lith, E. van (2000), Linking administrative registers and household surveys, In: P. Al en B.F.M. Bakker (red.) *Re-engeneering Social Statistics by micro-integration of different sources*. Themanummer Netherlands Official Statistics, jrg. 15, Summer, blz. 16-22.

Arts, C.H. en Hoogteijling, E.M.J. (2002), Het Sociaal Statistisch Bestand 1998 en 1999. *Sociaal Economische Maandstatistiek* 2002 (12), 66-71.

Bakker, B.F.M. (2003) Hoe nieuw zijn nieuwe ideeën?, In: J. Nobel, S. Algera, M. Biemans & P. van der Laan (red.), *Gedacht en gemeten*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen, blz. 123-132.

Bakker, B., & Arts, K. (2003), Dynamiek op de arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand. In: B.F.M. Bakker & L. Putman (red.), *De virtuele volkstelling en SSB.* SISWO/CBS, Amsterdam, blz. 59-70.

Bakker, B.F.M., Linder, F. en Roon, D. van (2008), Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. In: *IAOS Conference on Reshaping Official Statistics.* International Association of Official Statistics and the National Bureau of Statistics of China, Shanghai, 14-16 Oktober.

Bethlehem, J., (2007), *Methodenreeks: Wegen als correctie voor non-respons* Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

CBS (2009), http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm Centraal Bureau voor de Statistiek, Den Haag/Heerlen.

Groves, R.M., Fowler F.J. jr., Couper, M.P. Lepkowski, J.M., Singer, E. en Tourangeau, R. (2004), *Survey Methodology*. Wiley Interscience, New York.

Israëls, A., Pannekoek, P. en Schulte Nordholt, E. (2007), *Methodenreeks: Thema: Controle en correctie/imputatie. Deelthema: Imputatie*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

StatLine (2009), http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm?ConceptID=92. Centraal Bureau voor de Statistiek, Den Haag/Heerlen.