Secundaire waarneming

09

Piet Daas en Judith Arends-Tóth

Statistische methoden (09002)

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 - Graimedia

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

© Centraal Bureau voor de Statistiek, Voorburg/Heerlen, 2009.
Verveelvoudiging is toegestaan, mits het CBS als bron wordt vermeld.

6016509002 X-37

Inhoudsopgave

1. Inleiding op het thema 4
2. Aanwezigheid van en toegang tot secundaire bronnen 9
3. Geschiktheid van secundaire bronnen 17
4. Terugvalscenario's 27
5. Afsluiting 34
6. Literatuur 35
Bijlage A. Overzicht basisregistraties 38
Bijlage B. Overzicht kwaliteitsindicatoren voor secundaire bronnen 41
Bijlage C. Lijst van imagobepalende statistieken (2008) 44

1. Inleiding op het thema

1.1 Algemene beschrijving en leeswijzer

Het Centraal Bureau voor de Statistiek (CBS) wil onbetwiste en actuele statistieken over de Nederlandse samenleving samenstellen. Om dit doel te bereiken dient het CBS over voldoende actuele en betrouwbare gegevens te beschikken. Dit kunnen gegevens zijn die door het bureau zelf zijn verzameld of gegevens die door anderen worden verzameld en bijgehouden. Uit kostenoverweging en vanuit het oogpunt van de administratieve lastendruk, streeft het CBS ernaar om bij het maken van statistieken zoveel mogelijk gebruik te maken van bronnen met gegevens die door anderen (niet door het CBS) verzameld zijn. Dergelijke bronnen worden secundaire bronnen genoemd en de gegevens in die bronnen noemen we secundaire gegevens. Het proces van het verzamelen van secundaire gegevens door het CBS is het terrein van de secundaire waarneming.

1.1.1 Beschrijving van het thema

Het via secundaire waarneming verkrijgen van gegevens is een aanpak die niet uniek is voor de statistiek. Het is een methode die duidelijk multidisciplinair is; zeer verschillende wetenschapsgebieden maken van secundaire bronnen gebruik (Daas en Beukenhorst, 2008). De gehanteerde methoden behoren alle tot het wetenschapsgebied dat secundair onderzoek wordt genoemd (Golden, 1976). Bij secundair onderzoek worden reeds bestaande gegevens voor een ander doel gebruikt dan waarvoor ze oorspronkelijk verzameld zijn. Binnen het secundair onderzoek zijn drie onderzoeksstrategieën te onderscheiden (’t Hart et al., 2005; Golden, 1976), content analyse, secundaire analyse en systematische review (meta-analyse). Bij *content analyse* wordt voornamelijk de inhoud van verschillende vormen van menselijke communicatie bestudeerd. Bronnen die veel gebruikt worden, zijn o.a.: kranten, boeken, tv-beelden, websites en schilderijen. Probleem bij content analyse is het goed categoriseren en coderen van de, vaak grote hoeveelheden en ongestructureerde, gegevens. *Secundaire analyse* betreft het gebruiken van reeds - door anderen - verzamelde kwantitatieve gegevens voor een ander doel dan waarvoor ze oorspronkelijk verzameld zijn. In het algemeen zullen de analysemethoden die bij secundaire analyse gebruikt worden niet erg afwijken van de methoden die voor primaire bronnen (zoals enquêtegegevens van het CBS) worden gebruikt (Golden, 1976; Wallgren and Wallgren, 2007). Bij *systematische review* (meta-analyse) worden de resultaten van andere onderzoeken over hetzelfde of een soortgelijk fenomeen gecombineerd en bestudeerd. Een voorbeeld van een systematisch review is een onderzoek naar de relatie tussen kanker en voeding door het combineren van alle hierover eerder gepubliceerde onderzoeken in de wetenschappelijke literatuur in de afgelopen 15 jaar.

De hierboven beschreven drie methoden van secundair onderzoek worden alle ook door het CBS gebruikt. De meest gebruikelijke methode is zonder twijfel secundaire analyse. De via secundaire waarneming verkregen gegevens worden voornamelijk als input voor statistieken gebruikt. CBS-voorbeelden van secundaire analyse zijn het gebruik van de Belasting toegevoegde waarde (BTW) gegevens van ondernemingen (Belastingdienst, 2008) voor de Korte termijn statistiek (CBS, 2008b) en het gebruik van registers met bevolkingsgegevens voor de Virtuele Volkstelling (Schulte Nordholt et al., 2004). De andere twee methoden van secundair onderzoek, content analyse en systematische review, worden minder vaak door het CBS gebruikt. Een CBS-gerelateerd voorbeeld van content analyse is het historische overzicht getiteld 'Een eeuw statistieken' van Erwich en Van Maarseveen (1999). Een voorbeeld van een systematische review waar het CBS aan meegewerkt heeft, is de publicatie '200 jaar statistiek in tijdreeksen' (Van der Bie en Smits, 2001).

1.1.2 Problemen en oplossingen

Het CBS wil steeds meer gegevens uit externe, secundaire databronnen gebruiken voor de statistiekproductie. Het voordeel hiervan is dat de administratieve lasten en de kosten van dataverzameling afnemen. Dit laatste is uiteraard afhankelijk van de kosten die het CBS moet maken voor het verkrijgen van secundaire gegevens en de hoeveelheid werk die nodig is om deze gegevens geschikt te maken voor gebruik door het CBS. Verder bevatten bepaalde secundaire bronnen, zoals registers, gegevens over een (meestal) volledige populatie. Dit maakt het publiceren van zeer gedetailleerde statistieken mogelijk. Daarnaast kunnen, zeker wanneer meerdere bronnen gecombineerd worden, zonder aanvullende lastendruk, nieuwe statistieken worden gepubliceerd.

Deze voordelen maken secundaire bronnen zeer geschikt voor de statistiekproductie. Op het CBS worden secundaire bronnen dan ook voor de volgende statistische toepassingen gebruikt:

1) voor de statistiekproductie, ter vervanging van primaire gegevens;
2) als steekproefkader en bron van hulpinformatie bij het steekproefontwerp;
3) als bron van hulpvariabelen bij schattingen;
4) als input voor statistische registers (zoals het Algemeen Bedrijfsregister en het Sociaal Statistisch Bestand).

Daarnaast zijn er specifieke statistische toepassingen waarvoor de gegevens in secundaire bronnen bij uitstek geschikt zijn. Dit komt vooral omdat bepaalde bronnen gegevens over een vrijwel volledige populatie bevatten. Deze specifieke toepassingen zijn:

5) voor gedetailleerde publicaties (zoals regionale statistieken);
6) voor publicaties over bijzondere (niet vaak voorkomende) gebeurtenissen.

Wanneer er gegevens over meerdere tijdsperioden in een bron aanwezig zijn of de bron, vrijwel ongewijzigd, gedurende een langere periode door het CBS gebruikt wordt zijn secundaire bronnen ook zeer geschikt:

7) voor gedetailleerde longitudinale studies.

Doordat het CBS steeds vaker gebruik maakt van secundaire gegevens, wordt het CBS meer afhankelijk van:

1) de aanwezigheid van en toegang tot secundaire bronnen;
2) de geschiktheid (kwaliteit) van beschikbare secundaire bronnen;
3) de tijdige levering van secundaire bronnen.

Problemen in één of meerdere afhankelijkheden kunnen ernstige gevolgen hebben voor de statistische output van het CBS. In het meest extreme geval is het hierdoor mogelijk dat het CBS niet meer aan zijn publicatieplicht kan voldoen. De drie bovengenoemde afhankelijkheden van het CBS (aanwezigheid, geschiktheid en tijdige levering) met hun oplossingen worden in dit onderdeel van de Methodenreeks besproken. Samen vormen zij het onderdeel Secundaire waarneming.

1.1.3 Leeswijzer

In hoofdstuk 2 van dit document wordt het onderwerp ‘aanwezigheid van en toegang tot secundaire bronnen’ toegelicht. Om gegevens uit secundaire bronnen te kunnen gebruiken moet het CBS immers weten welke secundaire bronnen aanwezig en beschikbaar zijn en moet het bureau bekend zijn met de inhoud van die bronnen. Ook het maken van leveringsafspraken en het contact opnemen bij vragen en problemen met de bronhouder worden in dit hoofdstuk besproken.

De geschiktheid van een secundaire bron voor statistisch gebruik is het onderwerp van het derde hoofdstuk. Door het CBS zullen de gegevens in een secundaire bron (vaak) voor een ander doel worden gebruikt dan waarvoor ze oorspronkelijk zijn verzameld. Dit kan problemen veroorzaken. Zo kan in een bron de definitie van een belangrijke variabele niet exact gelijk zijn aan die van het CBS. Het is belangrijk dat het CBS snel inzicht krijgt in de mate waarin de gegevens in een secundaire bron geschikt zijn voor statistisch gebruik en waar zich de problemen bevinden. Dit zijn allemaal kwaliteitsgerelateerde aspecten.

In hoofdstuk 4 worden maatregelen besproken om zo goed mogelijk met de tijdelijke uitval van de levering van secundaire bronnen om te gaan. Door risicoanalyse worden de problemen in kaart gebracht. De combinatie van maatregelen wordt een terugvalscenario genoemd wat ervoor moet zorgen dat de statistische outputverplichtingen van het CBS zo min mogelijk hinder ondervinden van storingen in de levering van bronnen. Hoofdstuk 5 bevat afsluitende opmerkingen.

1.2 Afbakening en relatie met andere thema's

Bij secundaire waarneming verzamelt het CBS voor zijn statistische behoefte bronnen met gegevens die door anderen vergaard zijn. Het CBS heeft hierbij geen of zeer weinig zeggenschap over de eenheden waarover gegevens zijn verkregen, het proces van het verzamelen van de gegevens en de metadata definitie van de variabelen (Daas en Beukenhorst, 2008). Als het CBS zelf zijn gegevens verzamelt of als er een samenwerkingsverband met een externe partner gesloten is waarbij het CBS intensief betrokken is bij het verzamelen van gegevens, is er geen sprake van secundaire waarneming maar van primaire waarneming. Meer informatie over de door het CBS gevolgde aanpak bij het zelf verzamelen van gegevens (primaire waarneming) is te vinden in de thema's 'Benaderingsstrategieën', 'Vragenlijstontwerp' en 'Organisatie veldwerk' van de Methodenreeks. Meer informatie over de selectie van de eenheden waarover gegevens door het CBS verzameld worden, is te vinden in de thema's 'Steekproeftheorie' en 'Panels' van de Methodenreeks. De verdere bewerking van de secundaire gegevens behoort nadrukkelijk niet tot het onderwerp dat in dit document beschreven is. Hiervoor wordt de lezer verwezen naar de thema's 'Controle en correctie/imputatie' en 'Micro-integratie' van de Methodenreeks en de nota over Registermethodologie (Daas et al., 2005).

1.3 Plaats in het statistisch proces

De gegevens die door secundaire waarneming verkregen zijn vervangen (deels) de oorspronkelijk door het CBS zelf verzamelde (primaire) gegevens. Secundaire waarneming is als zodanig dus een onderdeel van de inputfase van het statistisch proces. Primaire en secundaire waarneming leveren samen alle gegevens die het CBS nodig heeft om aan zijn outputverplichtingen te voldoen. Het feit dat het CBS niet of slechts zeer beperkt betrokken is bij het verzamelen van secundaire gegevens (Daas en Beukenhorst, 2008) heeft overigens ook consequenties voor de processtappen daarna. Met name bij het bewerken van de gegevens, het integreren en het aggregeren dient er rekening mee gehouden te worden dat secundaire gegevens worden gebruikt. Deze vervolgstappen behoren echter niet meer tot het terrein van de secundaire waarneming.

1.4 Definities

Begrip	Omschrijving
Administratieve bron	Bron met gegevens die het gevolg is van het beheer van openbare of particuliere zaken
Basisregistratie	Een registratie die voldoet aan dan wel er naar toe werkt om te voldoen aan alle eisen die aan de authentieke status worden gesteld. Basisregistraties vormen het fundament van het stelsel van overheidsregistraties
Bron	Algemene aanduiding voor een verzameling van gegevens
Primaire bron	Een bron met primaire gegevens
Primaire gegevens (primaire data)	Gegevens die in opdracht van het CBS zijn verzameld, voor de statistiekproductie worden gebruikt en waarvan de conceptuele en procesmetadata door het CBS is vastgesteld.

Primair onderzoek	Onderzoek waarbij van primaire gegevens gebruik wordt gemaakt
Primaire waarneming	Het verkrijgen van primaire gegevens door het CBS
Register / registratie	Een verzameling gegevens die op gestructureerde wijze zijn vastgelegd en onderhouden wordt
Secundaire bron	Een bron met secundaire gegevens
Secundaire gegevens (secundaire data)	Gegevens die door anderen (niet door het CBS) zijn verzameld, voor de statistiekproductie van het CBS worden gebruikt en waarvan de conceptuele en/of procesmetadata niet door het CBS is vastgesteld.
Secundair onderzoek	Onderzoek waarbij van secundaire bronnen gebruik wordt gemaakt
Secundaire waarneming	Het verkrijgen van secundaire gegevens door het CBS

1.5 Algemene informatie

Door de businessarchitectuur van het CBS is een voorkeursvolgorde voorgesteld voor het statistisch gebruik van bronnen (Huigen, 2006). Efficiency, kostenreductie en het verminderen van de administratieve lastendruk voor bedrijven en personen vormen hierbij de uitgangspunten. In afnemende voorkeur is de volgorde voor het gebruik van gegevens uit bronnen de volgende:

i) aanwezige interne bronnen;

ii) reeds gebruikte secundaire bronnen;

iii) nieuwe (nog niet gebruikte) secundaire bronnen.

Indien (een deel van) de absoluut noodzakelijke gegevens in geen enkele bestaande bron te vinden zijn, de gegevens niet voldoende bruikbaar zijn, de levering erg instabiel is of de kosten voor de verwerving of verwerking van de bron te hoog zijn (CBS, 2003a-b, 2008a) kan besloten worden om deze gegevens te verzamelen door:

iv) één of meerdere vragen aan een reeds bestaande enquête toe te voegen;

v) een nieuwe enquête te houden.

Uit het bovenstaande lijstje volgt dat het CBS bij het verzamelen van nieuwe gegevens de voorkeur geeft aan gegevens afkomstig uit secundaire bronnen. Alleen wanneer informatie aantoonbaar niet aanwezig is, niet geschikt blijkt voor statistisch gebruik (met andere woorden wanneer de kwaliteit van de aanwezige informatie tekort schiet), er niet tijdig geleverd kan worden of de kosten voor gebruik te hoog zijn, wordt door het CBS zelf waargenomen.

2. Aanwezigheid van en toegang tot secundaire bronnen

2.1 Korte beschrijving

Om gegevens in secundaire bronnen te gebruiken moet het CBS bekend zijn met de aanwezigheid van geschikte bronnen en moet toestemming geregeld worden om de gegevens te mogen gebruiken. Ook voor wat betreft de leveringen en het terugkoppelen van vragen en advies bij problemen moeten afspraken worden gemaakt met de bronhouder.

2.2 Toepasbaarheid

Om het voor een statistisch bureau mogelijk te maken om op een grootschalige wijze gegevens uit secundaire bronnen te gebruiken wordt door het Nationaal statistisch bureau van Finland aanbevolen aan zoveel mogelijk van de volgende voorwaarden te voldoen (Statistics Finland, 2004):

1) Wettelijke basis
   Het dient wettelijk geregeld te zijn dat het statistische instituut gegevens uit secundaire bronnen mag gebruiken. Hierbij dient de privacy gewaarborgd te zijn.

2) Publieke instemming
   De bevolking (personen en bedrijven) dient geen bezwaar te hebben tegen het gebruik van 'hun' gegevens door het statistische bureau. Het imago van het statistische instituut als betrouwbare en optimale gebruiker van secundaire bronnen is een belangrijke factor bij het verkrijgen en behouden van publieke instemming.

3) Unieke identificatoren
   Voor het gebruik van gegevens uit verschillende bronnen is het beschikbaar zijn van algemeen in gebruik zijnde unieke identificatoren (voor de verschillende objecttypen) erg belangrijk. Door dergelijke identificatoren wordt de verwerking van de gegevens versneld en ontstaan er minder koppelingsfouten. Bronnen waarin dergelijke identificatoren niet voorkomen kunnen wel worden gebruikt maar tegen hogere kosten en met een grotere kans op fouten (door verkeerde en gemiste koppelingen).

4) Betrouwbare gegevens
   De gebruikte secundaire bronnen dienen betrouwbare gegevens van een zo groot mogelijk deel van de doelpopulatie te bevatten. Gebruik van deze bronnen door meerdere officiële instanties en door de populatie zelf verhoogt de betrouwbaarheid van de gegevens en de dekkingsgraad.

5) Samenwerking tussen autoriteiten
   Een goede samenwerking tussen de autoriteiten die betrokken zijn bij het gebruik en beheer van de bronnen verbetert de ontwikkeling van een op

secundaire bronnen gebaseerd statistisch systeem. Hoe hoger het niveau waarop de afspraken worden gemaakt hoe beter.

De introductie van het stelsel van basisregistraties in Nederland (paragraaf 2.3.2) en de CBS-wet (2003b), waarin de wettelijke toestemming is geregeld voor het kosteloze gebruik van de gegevens uit publieke bronnen (paragraaf 2.3.1), zijn een zeer goede stap in de richting om aan het merendeel van de hierboven genoemde voorwaarden te voldoen. Daarnaast is het belangrijk dat het CBS zeer zorgvuldig met de verzamelde gegevens omgaat en zich ook als zodanig naar de buitenwereld profileert. Het imago van het CBS als een betrouwbaar en zorgvuldig instituut mag absoluut geen 'deuk' oplopen.

2.3 Uitgebreide beschrijving

2.3.1 Mogelijkheden gebruik secundaire bronnen

In hoofdstuk 5 van de CBS-wet (2003b) is de wettelijke toestemming geregeld voor het gebruik van gegevens uit publieke bronnen. Het CBS mag voor statistische doeleinden kosteloos gebruik maken van bronnen die worden bijgehouden door:

i) instellingen en diensten van het Rijk, provincies, gemeenten en waterschappen;

ii) openbare lichamen (zoals beschreven in de Wet gemeenschappelijke regelingen en artikel 134 van de Grondwet);

iii) zelfstandige bestuursorganen op het niveau van de centrale overheid.

Daarnaast kunnen, na ruggespraak met de Centrale Commissie voor de Statistiek, bronnen worden gebruikt die worden bijgehouden door:

iv) aangewezen rechtspersonen die een taak uitoefenen die geheel of gedeeltelijk wordt bekostigd door de Staat of uit de opbrengst van bij wet ingestelde heffingen;

v) aangewezen categorieën van ondernemingen, vrije beroepsbeoefenaren, instellingen en rechtspersonen (uitsluitend wanneer de bronnen genoemd onder i t/m iv niet de benodigde gegevens opleveren).

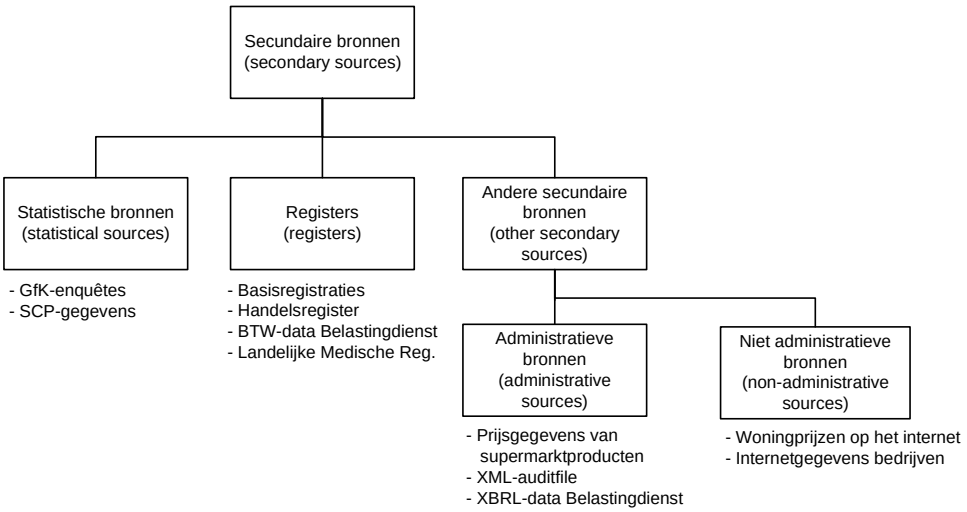
De gegevens dienen kosteloos, binnen een nader gespecificeerde termijn, door de hierboven genoemde leveranciers aan het CBS te worden versterkt. Een uitzondering vormen de gegevens die liggen op het terrein van het Nederlandse bankwezen. Deze worden of door de Nederlandsche Bank verstrekt of in overleg met de Nederlandsche Bank verzamelt. In artikel 36 van de CBS-wet (2003b) is vermeld dat de administratieve lasten voor ondernemingen, vrije beroepsbeoefenaren, instellingen en rechtspersonen zo laag mogelijk dienen te zijn. In het Besluit gegevensverwerving CBS (2003a) zijn de leveranciers, de te leveren gegevens en de leveringstermijn nader gespecificeerd.

2.3.2 Soorten secundaire bronnen

De secundaire bronnen die het CBS gebruikt of mogelijk kan gaan gebruiken omvat een heel diverse verzameling bronnen. Voorbeelden van secundaire bronnen zijn basisregistraties[1], omzetgegevens van de Belastingdienst, enquêtegegevens van het Sociaal Cultureel Planbureau, prijsgegevens van supermarktproducten en prijsgegevens van woningen op het internet. Een aantal van deze bronnen kan als een register of een administratieve bron worden aangemerkt. Dit is echter niet in alle gevallen zo eenvoudig. De prijsgegevens op een website zijn bijvoorbeeld duidelijk geen register en worden ook niet voor administratieve doeleinden bijgehouden.

Vanuit de informatiebehoefte van het CBS bekeken, zijn de secundaire bronnen van het CBS in drie hoofdgroepen in te delen, namelijk: statistische bronnen, registers (meer specifiek administratieve registers) en 'andere' bronnen. Deze indeling is gebaseerd op het vergelijken van de verschillende kenmerken van deze bronnen; zie Daas en Beukenhorst (2008) voor meer details. In figuur 1 is de indeling van secundaire bronnen getoond. In deze figuur is tevens te zien dat binnen de groep van de 'andere' secundaire bronnen er een nader onderscheid wordt gemaakt tussen administratieve en niet-administratieve 'andere' secundaire bronnen.

Figuur 1. Indeling van secundaire bronnen inclusief voorbeelden



Enkele voorbeelden ter verduidelijking. Een statistische secundaire bron die het CBS gebruikt zijn de enquêtegegevens die door anderen, zoals het Sociaal en Cultureel Planbureau, verzameld zijn. Voorbeelden van registers zijn het Handelsregister van de Kamers van Koophandel, de Landelijke Medische Registratie en de Gemeentelijke Basisadministratie (GBA) persoonsgegevens van de Nederlandse Gemeenten. De groep van registers kan ook als administratieve

[1] Overheidsbreed gebruikte registers die een volledig overzicht van een bepaalde populatie van eenheden geven waarvan de gegevens bij wettelijk voorschrift als authentiek zijn aangemerkt (E-overheid, 2009). In Bijlage A is de lijst met basisregistraties opgenomen.

registers worden aangeduid (Daas en Beukenhorst, 2008). Ook de basisregistraties behoren tot deze groep.

Basisregistraties zijn speciale bronnen omdat ze een fundament gaan vormen voor de uitvoeringstaken van de Nederlandse overheid. Basisregistraties zijn namelijk registers met gegevens die veel gebruikt worden door de overheid bij beleid, uitvoering en handhaving (E-overheid, 2009). Door gegevens vast te leggen in een stelsel van registraties is het de verwachting dat de kwaliteit zal toenemen. Daarbij komt dat bestuursorganen verplicht zijn, gebruik te maken van de gegevens in basisregistraties en verplicht zijn om, indien zij vermoeden dat gegevens onjuist zijn, dit aan de registerhouder terug te melden. Ook hierdoor zal de kwaliteit van de gegevens toenemen (zie ook paragraaf 2.3.5). Momenteel zijn er elf aangewezen basisregistraties in Nederland (E-overheid, 2009). Deze zijn in Bijlage A beschreven.

Naast de basisregistraties zijn er nog meer registraties die gegevens bevatten die veel door de overheid gebruikt worden. Wanneer uit een verkenning of haalbaarheidsstudie blijkt dat het register bijdraagt aan administratieve lastenverlichting, er sprake is van een groot reductiepotentieel en een veelvoudig gebruik van de gegevens, kan een dergelijk register als toekomstige basisregistratie worden aangemerkt (E-overheid, 2009). Momenteel zijn er twee toekomstige basisregistraties voorzien (Bijlage A). Het is de verwachting dat het CBS in de nabije toekomst steeds vaker gebruik zal gaan maken van (achtergrond)gegevens uit basisregistraties. Dit is uiteraard afhankelijk van de snelheid waarmee deze registraties in Nederland beschikbaar komen (Daas en Prins, 2007).

De laatste groep van secundaire bronnen omvat de ‘andere’ bronnen. Deze kunnen administratief of niet-administratief van aard zijn. Een voorbeeld van een administratieve ‘andere’ bron is een bestand met prijsgegevens van supermarktproducten, op het CBS meestal scannerdata genoemd. Een voorbeeld van een niet-administratieve ‘andere’ secundaire bron is een internetbestand dat prijsgegevens van koopwoningen bevat. Naast de hierboven genoemde voorbeelden zijn er ook ‘andere’ secundaire bronnen met gegevens die het CBS (nog) niet gebruikt. Enkele voorbeelden van dergelijke bronnen zijn: Radio Frequency Identification tags (RFID’s), satelliet gebaseerde navigatiesystemen (GPS), mobieletelefoniegegevens (GSM) en slimme elektriciteits- en gasmeters. Meer informatie over de mogelijkheden van het statistisch gebruik van deze bronnen is te vinden in Daas et al. (2008). De verwachting is dat het CBS in de nabije toekomst steeds vaker gebruik gaat maken van gegevens in ‘andere’ secundaire bronnen.

Statistische registers, zoals het Sociaal Statistisch Bestand (SSB) en het Algemeen Bedrijfsregister (ABR), behoren nadrukkelijk *niet* tot de secundaire bronnen van het CBS. Dergelijke registers vormen immers interne producten van het CBS, ze worden uit primaire en secundaire bronnen samengesteld en moeten als zodanig dan ook niet als een externe inputbron worden beschouwd.

2.3.3 Beschikbare bronnen

Omdat te verwachten is dat het CBS steeds vaker van gegevens in secundaire bronnen gebruik zal gaan maken, is het belangrijk om te weten welke informatie er in principe in Nederland beschikbaar is. Het CBS moet er naar streven een zo actueel mogelijk overzicht te bezitten van alle mogelijk bruikbare secundaire bronnen in Nederland. Op dit moment is een dergelijk overzicht niet op het CBS beschikbaar. Wel is er een overzicht van de secundaire bronnen die het CBS reeds gebruikt. Om te weten te komen welke ongebruikte bronnen potentieel interessant zijn is het College Bescherming Persoonsgegevens (CBP, 2009) een goed beginpunt. Deze instantie bezit namelijk een overzicht van alle instanties en bronnen die persoonsgegevens bevatten omdat iedereen die dergelijke gegevens verwerkt dit verplicht aan het college dient te melden. Het register met deze informatie wordt het openbare meldingenregister Wet bescherming persoonsgegevens genoemd. De enige bronnen die hier niet in zijn opgenomen zijn: i) bronnen die zijn vrijgesteld, zoals leden- en salarisadministraties en ii) bestanden die door politie en justitie worden gebruikt.

2.3.4 Relatiebeheer

Het toenemende gebruik van secundaire bronnen door het CBS maakt het noodzakelijk dat het CBS goed contact onderhoudt met de leverancier van de bron, de bronhouder. Dit behoort tot het terrein van het relatiemanagement. Hiertoe zijn, voor de belangrijkste bronhouders, op het CBS accountmanagers ingesteld. Dit is bijvoorbeeld voor de Polisadministratie en de GBA het geval. De accountmanagers dienen, voor de betreffende bron(nen) die ze onder beheer hebben, zowel informatie te leveren als te vergaren. Hierbij dient gedacht te worden aan het maken en bewaken van afspraken, het managen van verwachtingen en het tijdig signaleren van nieuwe ontwikkelingen (Huigen, 2006). Zo zal, om vast te stellen of een potentiële bron daadwerkelijk bruikbaar is, een afspraak over een voorverkenning moeten worden gemaakt. Voor secundaire bronnen waarvan besloten is dat ze door het CBS zullen worden gebruikt, dienen met de bronhouder duidelijke afspraken te worden opgesteld over de levering van de bron (inclusief metadata), het gebruik van de brongegevens en de wederzijdse verplichtingen die daarbij horen. De afspraken moeten worden vastgelegd in een formele overeenkomst.

De accountmanager vormt tevens het interne aanspreekpunt van het CBS bij vragen en problemen over (de leveringen van) de bron en bronhouder. Het contact met de bronhouder verloopt dus altijd via of na overleg met de accountmanager. Naast het *ad hoc* contact via de accountmanager is er vaak ook sprake van één of meerdere structurele overleggen van het CBS met de bronhouder. Denk hierbij aan bijvoorbeeld jaarlijkse vergaderingen op hoog bestuurlijk niveau, driemaandelijkse gebruikersoverleggen en tweemaandelijkse bilaterale bijeenkomsten van inhoudelijk deskundigen. Niet in alle gevallen zal de accountmanager namens het CBS aan dergelijke vergaderingen deelnemen. Het is uiteraard wel zo dat de accountmanager van deze overleggen op de hoogte dient te zijn en geïnformeerd moet worden over de daarin gemaakte afspraken.

2.3.5 Terugmelden

Onder terugmelden wordt verstaan het door het CBS terugkoppelen van informatie met betrekking tot de gegevens in de bron. Terugmelden zal vaak het gevolg zijn van problemen met de data in de bron en kan betrekking hebben op de individuele gegevens van personen of bedrijven. Dit laatste conflicteert met de in de CBS-wet vastgelegde bescherming van vertrouwelijke gegevens (CBS, 2003b). Dit vormt een dilemma voor het CBS. Aan de ene kant is het belangrijk dat de bronhouder zo exact mogelijk op de hoogte wordt gesteld van problemen met gegevens in de bron; dit om problemen in de toekomst te voorkomen en de kwaliteit van de data te verbeteren. Terwijl het CBS aan de andere kant wettelijk verplicht is de geheimhouding van individuele gegevens te garanderen. De aanpak die bij basisregistraties is gevolgd, illustreert de gemaakte keuze door het CBS.

Voor basisregistraties is er een terugmeld*plicht* die bestuursorganen - voor wie het gebruik van authentieke gegevens in basisregistraties verplicht is - verplicht om, bij het vermoeden van fouten in die gegevens, deze (op individueel niveau) terug te melden aan de beheerder van de data (Daas en Prins, 2008). De terugmeldplicht wordt voor elke basisregistratie afzonderlijk ingevuld. Momenteel is dit voor één basisregistratie, de GBA, geregeld. Voor de GBA is bepaald dat bestuursorganen verplicht zijn vermeende fouten terug te melden. Het CBS is op grond van het gestelde in de CBS-wet (2003b) van terugmelding vrijgesteld. Over de terugmeldplicht voor de GBA is een besluit genomen door het directieberaad van het CBS. Op 3 juli 2006 is namelijk besloten dat:

> “het CBS geen gevolg geeft aan de terugmeldingsplicht, zoals deze in de basisregistratie-wetgevingen i.o. zijn opgenomen, ook niet als die registratie niet over personen of bedrijven/instellingen gaat. Het CBS mag immers niet betrokken raken bij eventuele rechtsgevolgen van terugmeldingen of inbreuken op privacy, zowel vanuit strategisch oogpunt als op grond van het ‘geheimhoudingsartikel’ (artikel 37) van de CBS-wet” (2003b).

Dit besluit wordt gesteund door de argumentatie van het agentschap Basisadministratie Persoonsgegevens en Reisdocumenten (BPR). Het PBR is de beheerder van de infrastructuur van de GBA. De steun van het BPR heeft te maken met het feit dat, volgens hen, het CBS geen bestuursorgaan is in de zin van de Algemene Wet Bestuursrecht. Het CBS begeeft zich namelijk niet op het terrein van het bestuurlijk verkeer; d.w.z. op het nemen van besluiten en het verrichten van andere handelingen die individuele burgers raken. Hierdoor heeft de terugmeldplicht voor het CBS geen betekenis. Het is dan ook te verwachten dat het CBS bij geen enkele basisregistratie vermeende fouten verplicht zal terugmelden.

Het vorenstaande moet echter niet worden verward met de van oudsher bestaande praktijk dat het CBS contact kan opnemen met de bronhouder als de aangeleverde gegevens niet deugen. Bij het constateren van fouten of onduidelijkheden in de gegevens van een eenheid afkomstig uit één en dezelfde bron *kan* het CBS besluiten hierover bij de berichtgever navraag te doen. Dit gebeurt bijvoorbeeld als van een gemeente informatie over een bouwvergunning voor een aantal woningen wordt

ontvangen waarbij een onwaarschijnlijk hoge of lage bouwsom wordt vermeld en het absoluut noodzakelijk is duidelijkheid te verkrijgen over de plausibiliteit van die gegevens. Het betreft hier dus een controle van de interne consistentie van een aangeleverd record. Ook bij problemen met de levering van de bron mag worden teruggekoppeld. Zo zou de keten verantwoordelijk voor de levering van de Polisadministratiegegevens (Belastingdienst en Uitvoeringsinstituut Werknemers-verzekeringen) nooit snel verbeterd zijn zonder de hulp van het CBS. Belangrijk is dat het CBS hierbij nadrukkelijk niet op individueel (record) niveau terugkoppelt maar op globaal niveau en, in uitzonderlijke gevallen, geanonimiseerd. De privacy van personen en bedrijven moet immers gerespecteerd blijven. Ook mogen *nooit* feiten worden teruggekoppeld die het gevolg zijn van het koppelen van verschillende bronnen van verschillende leveranciers. Wanneer deze uitgangspunten worden gehanteerd moet terugkoppelen door het CBS tot de mogelijkheden behoren, zeker bij bronnen waar nog veel (opstart)problemen zijn.

2.3.6 Invloed

Voor het CBS zal een toename in het gebruik van secundaire bronnen tot gevolg hebben dat het steeds minder invloed zal hebben op datgene wat wordt verzameld. Zo kunnen allerlei beslissingen die buiten het CBS worden genomen, bijvoorbeeld door de bronhouder of door de politiek, er toe leiden dat er minder gegevens of gegevens van mindere kwaliteit beschikbaar zijn. Het CBS dient hier zoveel mogelijk op te anticiperen (zie ook hoofdstuk 4). Daarnaast dient het CBS te bekijken wat de mogelijkheden zijn om de invloed te vergroten op de gegevens die in Nederlandse registers en andere openbare bronnen, zoals de basisregistraties, worden verzameld. Door het Deense statistisch bureau is dit probleem opgelost door in de wet te regelen dat het geraadpleegd dient te worden bij het opzetten van nieuwe registers (Statistics Denmark, 1995).

2.4 Voorbeelden

2.4.1 BTW-gegevens voor Korte termijn statistiek

In 2006 heeft de Tweede Kamer middels het steunen van de motie Aptroot het CBS opgeroepen om te stoppen met de enquêtering van bedrijven in het midden- en kleinbedrijf. Om hieraan tegemoet te komen is door het CBS o.a. besloten de eigen waarneming in het midden- en kleinbedrijf (<50 werknemers) voor de Korte termijn statistieken zoveel mogelijk te vervangen door gegevens afkomstig uit secundaire bronnen. Bedrijven die tot de steekproef van de Korte termijn statistiek behoren, dienen de door hun behaalde omzet over een bepaalde periode (meestal maand- of kwartaal) aan het CBS te melden. Een secundaire bron met soortgelijke gegevens is het bestand met omzetbelastinggegevens van de Belastingdienst (2008). Elk bedrijf is in principe verplicht de Belasting Toegevoegde Waarde (BTW) periodiek aan de Belastingdienst te melden. Uitzonderingen zijn bedrijven met uitsluitend activiteiten in de landbouw en gezondheidszorg en kleine bedrijven met een BTW-afdracht van minder dan €5380 op jaarbasis (Belastingdienst, 2008). Er is door het CBS dan ook

een vooronderzoek naar de statistische bruikbaarheid van de BTW-gegevens van de Belastingdienst uitgevoerd. Hierbij bleek o.a. dat het aanpassen van de leveringsafspraken met de Belastingdienst noodzakelijk was. Zo werd afgesproken om i.p.v. eens per twee weken, om de twee dagen gegevens te leveren; dit om ervoor te zorgen dat het CBS sneller over zoveel mogelijk brongegevens beschikte. Het is de verwachting dat eind 2009-begin 2010 voor het midden- en kleinbedrijf de primaire gegevens van de Korte termijn statistiek door de BTW-gegevens van de Belastingdienst zullen worden vervangen.

2.4.2 Opleidingsgegevens

Om te voorzien in de groeiende informatiebehoefte over het onderwijs maakt het CBS in toenemende mate gebruik van secundaire bronnen die onderwijsgerelateerde gegevens bevatten. Door bestanden met onderwijsgegevens van verschillende jaren te koppelen is het mogelijk de doorstroming binnen het onderwijs in kaart te brengen. Dit biedt de mogelijkheid om de overgang van leerlingen in het voortgezet- en het middelbaar beroepsonderwijs naar andere onderwijssoorten te bestuderen en daardoor ook het (voortijdig) schoolverlaten in kaart te brengen. De bestanden die hierbij gebruikt worden zijn o.a. onderwijsnummerbestanden van de Informatie Beheergroep en de daaruit afgeleide 1-cijferbestanden per opleidingsniveau (zoals: voortgezet- en hoger-onderwijs). De 1-cijferbestanden worden door de Centrale Financiën Instellingen, een uitvoeringsorganisatie van het Ministerie van Onderwijs, Cultuur en Wetenschappen, in nauwe samenwerking met de afnemers (o.a. het CBS) opgesteld. Het doel hiervan is ervoor te zorgen dat elke betrokken instelling dezelfde set gegevens hanteert. De overleggen die in het kader van het samenstellen van de verschillende 1-cijferbestanden zijn gehouden boden het CBS de mogelijkheid om: i) de inhoud van de bestanden (mede) te bepalen en ii) de gehanteerde afleidingsregels, die voor het samenstellen van de variabelen in de bestanden worden gebruikt, (mede) te bepalen. Dit levert naast bestanden die zo goed mogelijk aan de eisen van het CBS voldoen ook een goed contact op met de bronhouder en de andere afnemers. In het geval van problemen of vragen is het dan ook geen enkel probleem contact op te nemen en de problemen in goede verstandhouding op te lossen.

2.5 Kwaliteitsindicatoren

De toenemende behoefte aan gegevens uit secundaire bronnen zorgt ervoor dat het CBS steeds meer moet gaan zoeken naar geschikte en beschikbare bronnen. Ook het opstellen en beheren van leveringsafspraken met bronhouders zal een steeds belangrijker onderdeel gaan vormen van het statistisch proces. Voor het beoordelen van de kwaliteit van secundaire bronnen is een kwaliteitskader opgesteld (Daas et al., 2008). Een onderdeel van dit kader richt zich specifiek op de kwaliteitsaspecten die met de leverancier, de afspraken en de levering van de bronnen te maken hebben. Het kwaliteitskader wordt in zijn geheel besproken in paragraaf 3.5. In tabel B1 van Bijlage B zijn indicatoren te vinden die voor het bewaken van de kwaliteit van het onderdeel bronnen zijn opgesteld.

3. Geschiktheid van secundaire bronnen

3.1 Korte beschrijving

Een secundaire bron bevat gegevens die door anderen verzameld zijn. In het algemeen zal de reden (het doel) waarvoor de gegevens oorspronkelijk verzameld zijn niet gelijk zijn aan de reden waarvoor het CBS ze wil gebruiken. De wijze waarop en de mate waarin de uitgangspunten van de bronhouder en die van het CBS verschillen, bepaalt voor een groot deel de geschiktheid (de kwaliteit) voor het statistisch gebruik van een bron. De aspecten die de mate van geschiktheid van een dergelijke bron voor het merendeel bepalen, worden in dit hoofdstuk besproken.

3.2 Toepasbaarheid

Voor het CBS vormen de gegevens uit secundaire bronnen een zeer belangrijke informatiebron. Op het CBS worden de gegevens in één secundaire bron dan ook meestal door meerdere, verschillende, statistieken gebruikt. Als gevolg hiervan zullen de eisen die aan de bron en de gegevens worden gesteld per statistiek variëren. Zo kan het zijn dat bepaalde variabelen wel door de ene en niet door de andere statistiek worden gebruikt. Maar ook de eisen die aan één en dezelfde variabele worden gesteld, voor wat betreft de actualiteit of nauwkeurigheid, kunnen per statistiek verschillen. Een voorbeeld maakt dit duidelijk.

De gegevens over de bevolking in de GBA worden op verschillende plaatsen binnen het CBS gebruikt. Zowel de Bevolkingsstatistieken als de Enquête Beroepsbevolking (EBB) gebruiken deze bron. De eisen die deze statistieken aan de actualiteit en volledigheid van de gegevens stellen variëren echter nogal. De Bevolkingsstatistieken willen een zo betrouwbaar en actueel mogelijk beeld geven van de demografische ontwikkelingen van de bevolking in Nederland (Prins en Kuijper, 2007). Het op het CBS zo snel mogelijk ontvangen en verwerken van allerlei veranderingen in de populatie en wijzigingen in de basisgegevens van personen is hierbij van essentieel belang. Voor de EBB worden de GBA-gegevens voornamelijk als steekproefkader en voor het leveren van achtergrondvariabelen gebruikt. Dit maakt dat de eisen die de EBB stelt aan de snelheid waarmee wijzigingen worden doorgegeven in de GBA-gegevens anders zijn. Dit heeft te maken met de verschillen in de wijze waarop bepaalde variabelen de statistieken beïnvloeden. Zo zal het te laat of het gedurende een bepaalde periode niet doorgeven van geboorten de EBB niet beïnvloeden maar de Bevolkingsstatistieken uiteraard wel. Het te laat doorgegeven van verhuisberichten - binnen dezelfde regio - zullen echter wel een effect hebben op de (op een steekproef gebaseerde) EBB-enquête en niet op de Bevolkingsstatistieken. Een te laat doorgegeven verhuizing zal het een CBS-enquêtrice immers (vrijwel) onmogelijk maken de geselecteerde persoon te interviewen.

3.3 Uitgebreide beschrijving

Bij het statistisch gebruik van gegevens uit secundaire bronnen doen zich een aantal methodologische aspecten voor die kenmerkend zijn voor (het gebruik van) dergelijke bronnen. Deze zaken vormen een onderdeel van het onderzoeksterrein dat op het CBS 'registermethodologie' wordt genoemd (Daas et al., 2005); dit ter onderscheid van de steekproeftheorie die bij het primair verzamelen van gegevens (b.v. met enquêtes) wordt gebruikt. In deze paragraaf worden de waarnemingsgerelateerde onderdelen van registermethodologie behandeld. De eigenschappen van secundaire bronnen en de gegevens van de bronnen die daarbij besproken worden, vormen een belangrijk onderdeel van de kwaliteit van een dergelijke bron.

3.3.1 Technische aspecten

Secundaire bronnen kunnen op verschillende manieren aan het CBS worden geleverd. Vaak betreft het de levering van een groot tot zeer groot bestand, maar dat hoeft niet. Zo wordt de GBA aan het begin van het jaar één maal volledig aan het CBS geleverd, waarna - gedurende de rest van het jaar - uitsluitend de mutaties van een beperkte set van variabelen worden doorgegeven (Prins en Kuijper, 2007). De meest gebruikte methode voor het leveren van secundaire bronnen is op elektronische wijze. Hierbij wordt het bestand (of de bestanden) via een beveiligde verbinding en een overdrachtsprotocol, zoals het File Transfer Protocol, een webservice of een overheidsbrede dienst, naar het CBS verzonden. Daarnaast is er in enkele gevallen ook sprake van leveringen op een fysieke drager per post of m.b.v. een koerier. Hierbij worden één of meerdere CD's, DVD's of harde schijven afgeleverd. Dit zijn echter vaak tijdelijke oplossingen die vaak in het begin van het leveringstraject worden gebruikt.

Het bronbestand of het deelbestand van een bron dient bij ontvangst op het CBS geopend en volledig gelezen te kunnen worden. Leesproblemen kunnen veroorzaakt worden door beschadiging van het bestand of de drager, decodeerproblemen en/of het gebruik van een onbekend (niet-standaard CBS) bestandstype. In het geval van een gecodeerd bestand, dient het voor het decoderen benodigde wachtwoord separaat verstuurd en tijdig ontvangen te zijn. Een volledig leesbaar bestand dient ten slotte aan de afgesproken metadata-definitie te voldoen. Dit is helaas niet altijd het geval.

3.3.2 Eenheden

Secundaire bronnen, met name registers, leveren naast gegevens over objecten ook een overzicht van de populatie voor een bepaald objecttype. Dit is ook de reden waarom secundaire bronnen van oudsher binnen de statistiek gebruikt werden, namelijk als steekproefkader (Erwich en Van Maarseveen, 1999). Voor statistisch gebruik is het belangrijk dat de eenheden in een secundaire bron van een duidelijk te identificeren objecttype zijn. In de ideale situatie is het objecttype in de bron identiek aan dat van het CBS. Wanneer dit echter niet het geval is, zal het gebruik

van een dergelijke bron door het CBS problematischer zijn. We zullen twee voorbeelden bespreken, één waarbij sprake is van een identiek objecttype en één waarbij een niet-identiek objecttype wordt gebruikt.

Secundaire bronnen met het objecttype natuurlijk persoon zijn goed door het CBS te gebruiken. Dit objecttype wordt ook door het CBS gebruikt en valt samen met een uniek te identificeren natuurlijke voorkomende eenheid. Deze eenheid wordt in veel bronnen, zoals de GBA, met een Burger Servicenummer (BSN) geïdentificeerd. Dit hoeft echter niet altijd zo te zijn. In dat geval wordt het gebruik van een dergelijke bron wat lastiger maar niet onmogelijk (Arts et al., 2000; zie paragraaf 3.3.3).

Moeilijker wordt het wanneer we het over het objecttype 'bedrijf' hebben. Wat is immers een bedrijf? Is dat de eenheid die aangifte doet bij de Belastingdienst of is dat de eenheid die de vragenlijst voor het CBS invult? Het probleem met het objecttype 'bedrijf' is dat veel bronhouders hun eigen definitie van dit objecttype hanteren. Zo gebruikt de Belastingdienst fiscale eenheden, beschouwt het CBS bedrijfseenheden en registreren de Kamers van Koophandel (KvK's) rechtspersonen en ondernemingen. Voor kleine bedrijven maakt dit niet veel uit; voor dit soort bedrijven zijn de eenheden van de Belastingdienst, het CBS en de KvK vaak aan elkaar gelijk (Aelen, 2004). Het probleem speelt met name bij de grote en middelgrote bedrijven. Hierbij zijn de eenheden die de verschillende overheidsinstituten hanteren niet altijd aan elkaar gelijk. Het kan zelfs zo zijn dat de fiscale eenheden die een 'bedrijf' voor de verschillende (verplichte) rapportages aan de Belastingdienst gebruikt niet allemaal aan elkaar gelijk zijn (Aelen, 2008). Dit maakt het erg lastig (de onderdelen van) grote en middelgrote bedrijven uniek te identificeren in de verschillende registers en eenduidig te combineren tot de eenheden die het CBS hanteert. Dit probleem wordt het 'eenhedenprobleem' genoemd en heeft een grote invloed op de statistieken van het CBS (Aelen, 2004, 2008). Het zijn namelijk vooral de grote bedrijven die de economische ontwikkeling in Nederland bepalen. Het CBS probeert dit lastige probleem op te lossen door zo vroeg mogelijk de informatie over grote, cruciale en/of complexe bedrijven in het verwerkingsproces te integreren. Deze aanpak dient te zorgen voor consistentie tussen en binnen statistieken (Verlinden, 2008). Een koppelkader waarin de onderlinge relatie tussen de verschillende registereenheden voor deze zo belangrijke groep van bedrijven/ondernemingen is vastgelegd moet hier zorg voor dragen. Hiervoor wordt in 2009 een eenhedenbase opgesteld (Aelen, 2008).

Naast dit probleem bevat een secundaire bron niet altijd de gehele populatie van eenheden die onderwerp zijn van de statistiek. Door alle bronnen die betrekking hebben op soortgelijke eenheden samen te voegen, wordt geprobeerd dit probleem zo goed mogelijk op te lossen. Hierdoor wordt een completer beeld van de populatie van eenheden verkregen. Meer over dit onderwerp is te vinden in het onderdeel 'Micro-integratie' van de Methodenreeks.

3.3.3 Identificerende gegevens

Een basisvoorwaarde voor het efficiënt gebruiken van secundaire bronnen is de aanwezigheid van uniforme identificatiegegevens voor elk object. Dit vereist dat de objecttypen duidelijk en consistent gedefinieerd zijn (zie vorige paragraaf) en dat de identificatiecodes precies en doelmatig zijn en bestand zijn tegen veel voorkomende fouten. De precisie waarmee eenheden gedefinieerd zijn, hangt af van het objecttype. Voor sommige objecten, zoals personen, auto's en schepen is dit geen probleem. Deze objecten vallen namelijk samen met een natuurlijk voorkomende eenheid. In andere gevallen (b.v. bij percelen en gebouwen) moeten operationele of juridische definities gezocht worden die aan de ene kant objectief zijn waar te nemen en aan de andere kant aansluiten bij de doelen van een bepaalde bron. Mogelijk biedt het stelsel van basisregistraties (E-overheid, 2009) een praktisch bruikbaar referentiekader voor dergelijke objecttypen. Voor complexe en dynamische objecten, zoals ondernemingen, bedrijven en huishoudens, is nog andere informatie noodzakelijk. Naast duidelijke definities dienen hiervoor tevens bronnen beschikbaar te zijn die zowel identificerende als relatieleggende gegevens bevatten. Daarnaast is het voor dergelijke objecten belangrijk dat één consciëntieus beheerd centraal beschikbaar koppelkader (zoals de eerder vermelde eenhedenbase voor bedrijfsobjecten) wordt gebruikt om aan dit soort objecten een praktische bruikbare invulling te geven.

De eisen aan de te gebruiken identificatiecodes zijn meer technisch van aard. Hierbij dient rekening gehouden te worden met het voorkomen van dubbeltellingen en telfouten. Een identificatiecode met een interne controlemogelijkheid, zoals de 11-proef [2] voor het BSN-nummer, verdient de voorkeur omdat deze op (type)fouten te controleren is. Daarnaast speelt de vraag of een identificatiecode informatieloos moet zijn of niet. Zo bevat het belangrijkste identificatienummer in Denemarken en Finland o.a. informatie over geboortedatum en geslacht (Statistics Denmark, 1995; Statistics Finland, 2004). Omdat deze kenmerken vrijwel onveranderbaar zijn levert dit geen grote problemen op, maar het kan wel (b.v. bij geslachtsverandering). Nadeel van een identificatiecode die een eigenschap van de te identificeren eenheid bevat (b.v. een geboortedatum of oprichtingsdatum) is de kans op herkenning en het feit dat een dergelijke code niet voor (andere type) eenheden gebruikt kan worden die niet een dergelijk kenmerk bevatten (b.v. percelen).

Wanneer de objecten in een bron geen unieke identificerende nummers bevatten dient een (unieke) combinatie van variabelen gebruikt te worden om elk van de objecten in de bron uniek te identificeren. Voor personen kan dit b.v. de combinatie van naam-, adres- en woonplaatsgegevens zijn (Arts et al., 2000).

---

[2] Het 9-cijferige BSN-nummer moet aan de volgende rekenregel voldoen: (1e cijfer BSN x 9 + 2e cijfer BSN x 8 + 3e cijfer BSN x 7 + 4e cijfer BSN x 6 + 5e cijfer BSN x 5 + 6e cijfer BSN x 4 + 7e cijfer BSN x 3 + 8e cijfer BSN x 2 - 9e cijfer BSN x 1) / 11 = een geheel getal.

3.3.4 Tijdsaanduidingen

Tijdsaanduidingen vormen een essentieel onderdeel van een databron. Zo moeten de tijdsaanduidingen die de existentie van een eenheid aangeven in een bron er ook in zijn opgeslagen. Tijdsaanduidingen hebben betrekking op periode- en standvariabelen. Zo is de variabele 'inkomen' een voorbeeld van een periodevariabele; het gaat om het inkomen gedurende een bepaalde tijdsperiode. Periodevariabelen worden ook wel volume- of inhoudvariabelen genoemd. Een voorbeeld van een standvariabele is 'leeftijd'. Het beschrijft de situatie op één bepaald tijdstip. Standvariabelen in een bron kunnen continu of discreet gevolgd worden. Continu houdt in dat de periode van geldigheid vermeld is, discreet dat alleen de waarde op een bepaald tijdstip of op bepaalde tijdstippen bekend is.

Van de gegevens in secundaire bronnen dient duidelijk vermeld te zijn op welke periode of perioden ze betrekking hebben. Problemen die hierbij kunnen optreden zijn: onvolledige vermeldingen van begin- en/of eindperioden, verschillen tussen de geregistreerde en feitelijke data van een gebeurtenis (m.n. bij retrospectieve wijzigingen lastig) en verschillen tussen de tijdsperiode van de bronhouder en die van het CBS.

Een andere tijdsgerelateerde oorzaak van problemen is administratieve vertraging bij de bronhouder. Indien de bronhouder de ontvangen gegevens niet direct verwerkt en opslaat zal de informatie in de bron achterlopen t.o.v. de actualiteit. Dit probleem komt op het CBS bijvoorbeeld naar voren wanneer aan een dergelijke bron (zeer recente) enquêtegegevens worden toegevoegd. Als duidelijk is dat het niet gelijk zijn van de waarde van één of meer variabelen in twee bronnen wordt veroorzaakt door administratieve vertraging in één van de bronnen, is dit probleem eenvoudig op te lossen. Dit is echter niet altijd direct duidelijk.

3.3.5 Variabele-definities

Het kan voorkomen dat de definitie van een variabele in de bron niet (volledig) overeenkomt met die van de variabele waarover het CBS wil publiceren. Dit is bijvoorbeeld het geval bij het begrip 'werkloosheid' van het CBS en van de arbeidsbureaus (Van den Elshout et al., 2007). Er is een aantal manieren om de publicatie van een statistiek gebaseerd op dergelijke gegevens door het CBS mogelijk te maken. De waarde van een variabele in de bron kan door *correctie* aangepast worden aan de (te verwachten) waarde van de variabele volgens de CBS-definitie of kan door het *combineren van gegevens* afgeleid worden. Het is echter ook mogelijk niet meer van de oude (CBS-)definitie van een variabele gebruik te maken maar een nieuwe definitie te introduceren. Een voorbeeld hiervan is de introductie van het begrip 'betalingsconcept' in het Sociaal Statistisch Bestand (Arts en Hoogteijling, 2002). Mocht één van de hierboven genoemde opties, om wat voor reden dan ook, niet mogelijk zijn dan kan overwogen worden om andere bronnen te zoeken die wel bruikbare informatie over de variabele bevatten of te besluiten de noodzakelijke gegevens m.b.v. een enquête te verzamelen.

3.4 Voorbeelden

3.4.1 Inkomens Informatie Systeem

Het CBS verzamelt samen met het ministerie van Financiën informatie over de samenstelling van de inkomens en vermogens van natuurlijke personen in Nederland. Hiertoe is het Inkomens Informatie Systeem (IIS) opgezet. Dit systeem wordt door het CBS beheerd. De gegevens in het IIS zijn afkomstig uit registraties van de Belastingdienst, de dienst Toeslagen en de Informatie Beheer Groep. Hiertoe worden de volgende secundaire bronnen gecombineerd:

- Aangifte- en aanslaggegevens Inkomstenbelasting
- Aangifte- en aanslaggegevens Zorgverzekeringswet
- Aangiftegegevens Winstaangifte Inkomstenbelasting en Vennootschapsbelasting
- Toeslaggegevens (gegevens over aanvragen en toekenningen van Huurtoeslag, Zorgtoeslag en Kinderopvangtoeslag)
- Loonbelasting gegevens (Fibase)
- Rentebetalingen (Rentebase)
- Dividendbetalingen (Dividendbase)
- Administratie Wet Studiefinanciering (betalingen studiefinanciering)
- Waardering Onroerende Zaken
- Beheer van Relaties voor personen (kenmerkende informatie over natuurlijke personen).

De bronnen worden gekoppeld op persoonsniveau m.b.v. het BSN-nummer. Het ministerie van Financiën gebruikt de gegevens voor het doorrekenen van beleidsvoornemens en de evaluatie van de fiscale regelgeving en belastingramingen. Het CBS gebruikt de gegevens voornamelijk voor het samenstellen van de personele inkomens- en vermogensstatistieken. Zo worden voor het Inkomens Panel Onderzoek (IPO) gedetailleerde inkomens- en vermogensinformatie per sociaal-economisch kenmerk en regio geleverd. De uitkomsten van het IPO worden gepubliceerd op StatLine (de outputdatabase van het CBS) en gebruikt in het Jaarboek Welvaartsverdeling, de Armoedemonitor en diverse andere artikelen en persberichten.

3.4.2 Sociaal Statistisch bestand

Het Sociaal Statistisch Bestand (SSB) wordt gemaakt door gegevens over personen uit secundaire bronnen en CBS-enquêtes op microniveau te koppelen (Arts en Hoogteijling, 2002). De kern van het SSB bestaat uit een aantal onderling gekoppelde registers, waarin demografische en sociaal-economische gegevens zijn opgenomen. De SSB-kern is niet één fysiek bestand al doet de naam dat wel vermoeden. Dat is ook niet mogelijk, omdat er verschillende eenheden worden onderscheiden (banen, uitkeringen, personen). Uit de gekoppelde bronnen wordt per eenheid een bestand gemaakt: een banenbestand van werknemers in Nederland, een bestand met zelfstandigen, een bestand met werknemers die in het buitenland werken, zes uitkeringenbestanden en een personenbestand waarin demografische informatie is opgenomen. Daarnaast bestaan er SSB-satellieten waarmee een

bepaald onderwerp nader wordt beschreven, zoals de maatschappelijke positie van allochtonen of de criminaliteit in Nederland. Zowel de kern als de satellieten die samen het SSB vormen zijn nog volop in ontwikkeling.

De secundaire bronnen die voor de SSB-kern worden gebruikt, zijn:

- Gemeentelijke basisadministratie persoonsgegevens
- Loonbelasting gegevens (Fibase)
- Verzekerdenadministratie werknemers
- Aangiftegegevens Inkomstenbelasting
- Landbouwtelling
- Administratie arbeidsongeschiktheidswetten
- Administratie werkloosheidswet
- Administratie algemene bijstandswet
- Administratie wet inkomensvoorziening oudere en gedeeltelijk arbeidsongeschikte werkloze werknemers / arbeidsongeschikte gewezen zelfstandigen
- Waardering Onroerende Zaken
- Beschikkingen ziekenfondsverzekeringsplicht zelfstandigen
- Beheer van relaties
- Woningstatistiek
- Inschrijvingen Centrum Werk en Inkomen
- Centraal Register Ingeschrevenen Hoger Onderwijs
- Les en cursusgeldregister
- Examen Resultaten Register
- Administratie Wet Studiefinanciering.

De gegevens van de Enquête Beroepsbevolking en Statistiek werkgelegenheid en lonen van het CBS worden hieraan toegevoegd. De combinatie van bronnen levert een groot aantal gegevens over personen, banen en uitkeringen voor het samenstellen van statistische overzichten. Een belangrijk doel van het SSB is om samenhangende en consistente informatie te publiceren. Het SSB vormt daarom ook de basis van veel sociale statistieken van het CBS.

Doordat de informatie in het SSB voor meerdere jaren beschikbaar is, kunnen objecten gevolgd worden in de tijd. Zo kan worden nagegaan wie er uitstromen uit de WAO en wie daarvan een baan vindt. Bovendien behoren ook duuranalyses tot de mogelijkheden. Daarmee kunnen vragen beantwoord worden over hoe lang iemand een werkloosheidsuitkering heeft en of dat verschilt tussen verschillende bevolkingsgroepen. Omdat vrijwel alle bronnen in het SSB integraal zijn, kan meer en betrouwbaarder informatie worden samengesteld over kleine regionale eenheden en kleine groepen in de samenleving. Voorbeelden hiervan zijn in- en uitstroom van werkzame personen in de gezondheids- en welzijnszorg en het percentage van de bevolking van de gemeente Pekela met een werkloosheidsuitkering.

Omdat het SSB individuele en gevoelige informatie bevat, worden de gegevens niet zonder meer ter beschikking gesteld. Gegevens zijn in te zien bij publicatie op geaggregeerd niveau in StatLine of tegen betaling voor een beperkt aantal instituten na overleg met en beveiliging van de gegevens door het Centrum voor Beleidsstudies van het CBS.

3.5 Kwaliteitsindicatoren

3.5.1 Kwaliteitskader voor secundaire bronnen

Op het CBS is een kwaliteitskader ontwikkeld voor secundaire bronnen (Daas et al., 2008). Dit wordt gebruikt om de *algemene* statistische bruikbaarheid van een secundaire bron te bepalen. Het kader is niet opgezet om te bepalen of een bron voor een bepaald specifiek gebruik geschikt is. Bij dit laatste dient gedacht te worden aan bijvoorbeeld het bepalen van de bruikbaarheid van het werkloosheidpercentage afkomstig van het Centrum voor Werk en Inkomen (CWI) ter vervanging van dat bepaald op basis van de Enquête Beroepsbevolking. Dit is een zeer specifieke vraag en is derhalve niet in het algemene raamwerk opgenomen. Daarnaast zullen de verschillende gebruikers elk weer andere specifieke kwaliteitsmetingen in gedachten hebben, wat weer tot andere geschiktheidsresultaten kan leiden. Het is eenvoudigweg onmogelijk om een algemeen kwaliteitskader op te stellen dat aan alle specifieke controles van alle potentiële gebruikers voldoet. Dientengevolge is het te verwachten dat na het bepalen van de algemene geschiktheid voor statistisch gebruik van een bron, zoals het CWI-bestand, er - na een positieve uitkomst - nog één of meerdere aanvullende, zeer specifieke, geschiktheidcontroles zullen worden uitgevoerd (Daas et al., 2008).

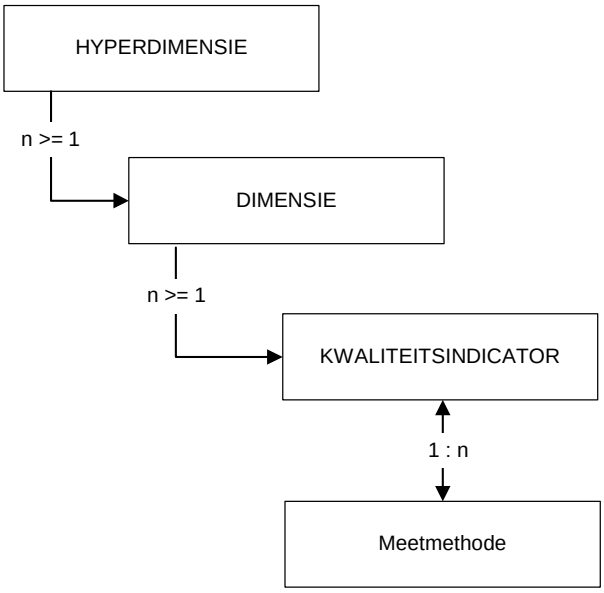
3.5.2 Opbouw kwaliteitskader

Bij het kwaliteitskader voor secundaire bronnen worden vier duidelijk verschillende manieren van kijken onderscheiden. Deze worden ‘hyperdimensies’ genoemd (Karr et al., 2006). Het kwaliteitskader omvat de volgende vier hyperdimensies: Bron, Metadata, Data en Proces (Daas et al., 2008).

De vier ‘manieren van kijken’ naar de kwaliteit van administratieve bronnen belichten ieder andere kwaliteitsaspecten. De eerste drie hyperdimensies hebben geen overlappende aspecten en zijn tevens geordend naar toenemend detailniveau. Ze zijn alle product gebaseerd. De Proces-hyperdimensie is duidelijk anders omdat hier specifiek naar de verwerking van een secundaire bron wordt gekeken. Voor het beoordelen van de kwaliteit van een secundaire bron, d.w.z. de statistische bruikbaarheid voor het CBS, is het noodzakelijk de hyperdimensies Bron, Metadata en Data te evalueren. Omdat Proces hiervoor niet noodzakelijk is, wordt deze niet verder in dit document behandeld.

Voor elk van de Bron-, Metadata- en Data-hyperdimensies zijn de bijbehorende kwaliteitsaspecten in dimensies vastgelegd. Een dimensie dient te worden opgevat als een onderdeel van kwaliteit dat zich beperkt tot één specifiek kenmerk. Elke dimensie bevat één of meerdere kwaliteitsindicatoren, die elk verwijzen naar een bepaald meetbaar aspect van de kwaliteit van dat kenmerk. Een kwaliteitsindicator wordt altijd door minimaal één methode gemeten; maar dit kunnen er meer zijn. Een meetmethode is een procedure om de waarde van een kwaliteitsindicator te bepalen of te schatten. Het verband tussen de verschillende kwaliteitsaspecten in het ontwikkelde kwaliteitskader is in figuur 2 weergegeven.

Figuur 2. De verschillende aspecten van kwaliteit en hun onderlinge relatie



Hyperdimensie Bron

De Bron-hyperdimensie bekijkt de secundaire bron als één bestand dat aan het CBS wordt geleverd waarbij de aandacht voornamelijk is gericht op de leverancier en de levering van het bestand. Voorbeelden van kwaliteitsaspecten die tot de Bron-hyperdimensie behoren zijn: afspraken over de levering van de bron, gegevens van de leverancier en contactpersonen en beveiligings- en privacyafspraken. De kwaliteitsaspecten van Bron zijn in tabel B1 van Bijlage B weergegeven. De meetmethoden in de Bron-hyperdimensie zijn voornamelijk kwalitatief maar er zijn ook enkele kwantitatieve methoden aanwezig, zoals het effect van het gebruik van de bron op de administratieve lastendruk van het CBS. Voor het beoordelen van de kwaliteitsaspecten in de Bron-hyperdimensie is een checklist opgesteld (Daas en Arends-Tóth, 2007). De evaluatie van de Bron-hyperdimensie geeft aan of een bron wel of niet potentieel geschikt is voor statistisch gebruik. Na een positieve evaluatie dienen de kwaliteitsaspecten in de Metadata-hyperdimensie geëvalueerd te worden.

Hyperdimensie Metadata

De Metadata-hyperdimensie richt zich op de conceptuele metadata-aspecten van een secundaire bron en bevat ook enkele procesgerelateerde meta-aspecten. Voorbeelden van kwaliteitsaspecten van de Metadata-hyperdimensie zijn duidelijkheid en vergelijkbaarheid van de definities van variabelen en de populatie, gegevens over de tijdsperiode waarop de data betrekking hebben en de (eventuele) controles die door de bronhouder op de gegevens zijn uitgevoerd. De kwaliteitsaspecten van Metadata zijn in tabel B2 van Bijlage B weergegeven. In de Metadata-hyperdimensie komen uitsluitend kwalitatieve meetmethoden voor. Het beoordelen van de kwaliteitsaspecten van deze hyperdimensie vindt eveneens m.b.v. een checklist plaats (Arends-Tóth en Daas, 2008). De evaluatie van de kwaliteitsindicatoren van

Metadata geeft aan of de bron wel, niet of deels geschikt is voor statistisch gebruik. Indien een bron geschikt of deels geschikt is dient de Data- hyperdimensie te worden geëvalueerd.

Hyperdimensie Data

De Data-hyperdimensie omvat de kwaliteitsaspecten die op de gegevens in de secundaire bron betrekking hebben. De kwaliteitsaspecten die tot de Data-hyperdimensie behoren komen in het algemeen overeen met de nauwkeurigheid van schatters of statistieken (Eurostat, 2003, 2005; Kuijvenhoven en Schouten, 2008). Uitzondering zijn de aspecten die tot de technische datacontroles behoren. De kwaliteitsaspecten van Data zijn in tabel B3 van Bijlage B weergegeven. De meetmethoden in de Data-hyperdimensie zijn vooral kwantitatief (Daas et al., 2008). Veel van de meetmethoden in de Data-hyperdimensie zijn gebaseerd op de zogenaamde Representativiteitsindex (R-index). Deze index is ontwikkeld door het CBS (Schouten and Cobben, 2007; Cobben and Schouten, 2008). Met een R-index wordt bepaald in hoeverre de mate waarin de samenstelling van de eenheden in een bron, op een bepaald tijdstip, afwijkt van de populatie van die eenheden.

3.5.3 Gebruik van het kwaliteitskader

Bij het beoordelen van de kwaliteit van een databron dient de gebruiker eerst de Bron-, daarna de Metadata- en vervolgens de Data-hyperdimensie te beoordelen. Voor het beoordelen van de Bron- en de Metadata-hyperdimensie zijn checklists ontwikkeld (Daas en Arends-Tóth, 2007; Arends-Tóth en Daas, 2008). Voor Data wordt nog bestudeerd wat de meest geschikte aanpak is. Wanneer er problemen geconstateerd worden in een bepaalde hyperdimensie dienen deze eerst opgelost te worden voordat de volgende hyperdimensie wordt bestudeerd. Indien in een bepaalde hyperdimensie geconstateerd wordt dat een bron niet bruikbaar is voor de statistiek dan dient de evaluatie gestopt te worden en hoeven de daarop volgende hyperdimensie(s) niet verder beoordeeld te worden. Na evaluatie van de Data-hyperdimensie kan een gebruiker definitief weten of een databron voor statistisch gebruik geschikt is. Het is mogelijk (en zelfs zeer waarschijnlijk) dat daarna nog een eventuele specifieke meting dient te worden uitgevoerd om de mate van geschiktheid exact te bepalen (Kuijvenhoven en Schouten, 2008).

4. Terugvalscenario's

4.1 Korte beschrijving

Een terugvalscenario is een combinatie van maatregelen om de nadelige gevolgen van het tijdelijk uitvallen van een deel of de gehele secundaire bron op te vangen voor de statistiekproductie van het CBS. Door risicoanalyse worden de problemen in kaart gebracht. Omdat dit terrein erg in ontwikkeling is, kan het zijn dat de precieze invulling op sommige details enigszins afwijkt van de situatie die in dit document is beschreven.

4.2 Toepasbaarheid

De toename in het gebruik van secundaire bronnen voor de statistiekproductie maakt het CBS meer afhankelijk van de tijdige beschikbaarheid van gegevens in secundaire bronnen. Hierdoor ontstaan risico's die de statistiekproductie negatief kunnen beïnvloeden. Het (tijdelijk) wegvallen van een secundaire bron is hier een voorbeeld van. Indien een secundaire bron tijdelijk wegvalt, moet het CBS - vaak snel - beslissen hoe met het ontbreken van gegevens wordt omgegaan (Frenken et al., 2008). Een dergelijke situatie kan er immers toe leiden dat bepaalde statistieken (tijdelijk) niet gepubliceerd kunnen worden. Een oplossing hiervoor is het opstellen van een terugvalscenario; een combinatie van maatregelen om de nadelige gevolgen voor de gebruikers te beperken. Een terugvalscenario dient uitsluitend te worden toegepast om het tijdelijk wegvallen van een deel of de gehele secundaire bron op te vangen.

Door het directieberaad van het CBS is besloten dat voor de imagobepalende statistieken van het CBS (zie Bijlage C) die direct of indirect van secundaire bronnen afhankelijk zijn een risicoanalyse dient te worden uitgevoerd en een terugvalscenario moet worden opgesteld (verslag CBS-directieberaad, 26 januari 2009). Tot de imagobepalende statistieken van het CBS behoren die statistieken waarvan het niet-tijdig publiceren aanzienlijke risico's heeft voor de gebruikers én het imago van het CBS. Van de 16 imagobepalende statistieken zijn er 15 van secundaire bronnen afhankelijk (zie Bijlage C). Voor alle andere CBS-statistieken die van gegevens in secundaire bronnen gebruik maken, is het opstellen van terugvalscenario's aan te bevelen.

In het algemeen wordt de volgende aanpak aanbevolen bij het ontwikkelen van een scenario voor het omgaan met het tijdelijk wegvallen van een belangrijke bron:

1. ga na of eigen waarneming, in eerste instantie via alternatieve externe bronnen, qua kosten en tijdsbeslag haalbaar is, zo ja, dan verdient dit de voorkeur;
2. pas een modelbenadering toe als eigen waarneming niet haalbaar is en een deel van de data over de verslagperiode nog aanwezig is, mits aannemelijk

is dat het toegepaste model tot een resultaat van aanvaardbare kwaliteit leidt;

3. meld de belangrijke gebruikers de mogelijke gevolgen van de uitval van de bron;
4. besluit tot uitstel als beide voorgaande mogelijkheden vervallen.

4.3 Uitgebreide beschrijving

Het CBS wil zich manifesteren als een betrouwbare en stabiele leverancier van maatschappelijk relevante informatie. Dientengevolge moet de tijdige publicatie en de kwaliteit van de statistische informatie van het CBS onomstreden zijn. De toename in het gebruik van secundaire bronnen maakt het CBS wat dat betreft meer kwetsbaar. Om dit probleem zoveel mogelijk te voorkomen kunnen voor statistieken die van secundaire bronnen gebruik maken terugvalscenario's opgesteld worden. Door risicoanalyse worden kwetsbaarheden opgespoord. Hierbij dient de methodiek van de standaardprocesbeschrijvingen en het voor het voorschrift informatiebeveiliging rijksdienst 2007 (VIR, 2007) geadopteerde ontwikkeltempo te worden gevolgd.

4.3.1 Afbakening bruikbaarheid terugvalscenario's

Het is echter niet realistisch om ervan uit te gaan dat voor alle denkbare situaties terugvalscenario's moeten klaarliggen. Terugvalscenario's zijn in de praktijk vaak maatwerk. Wat in een bepaalde situatie de beste oplossing is, hangt onder meer af van wat er precies ontbreekt en de kwaliteit van de resterende informatie. De gekozen oplossing dient tevens rekening te houden met de kosten en de doorgaans korte tijd die beschikbaar is. Daarom is het raadzaam alleen voor de statistieken van het CBS waarvan de uitval van een secundaire bron serieuze gevolgen zal hebben voor de kwaliteit van de te publiceren statistiek terugvalscenario's op te stellen. Vroegtijdig signaleren van mogelijke aanstaande problemen vergroot de kansen op adequaat reageren. Om die reden is actief relatiebeheer - in dit geval vooral het contact met de bronhouder - ook van belang.

Voor een bron die permanent uitvalt hoeft er, in principe, geen terugvalscenario te worden opgesteld. Om in een dergelijk geval alsnog aan de statistische output verplichting te voldoen dient er een nieuw statistisch dataverzamelingsproces opgestart te worden. Het is te verwachten dat een dergelijke overgang enig tijd in beslag zal nemen. Externe druk en verplichtingen kunnen het noodzakelijk maken om in de tussenliggende tijd, in ieder geval tijdelijk, andere "creatieve" oplossingen te gebruiken; denk aan een volledig modelgebaseerd cijfer, een genowcast[3] cijfer,

[3] Een nowcast is een voorspelling van het nu (of beter: van het recente verleden). Een nowcast maakt een raming voor de periode die achter ons ligt maar die nog niet door directe statistische waarneming beschreven kan worden. Bij het maken van een nowcast bedient men zich van hetzelfde instrumentarium als bij het maken van een voorspelling (een forecast).

een expert ‘guess’ en/of gebruik van de Delphi-methode[4]. Het spreekt vanzelf dat hierover helder naar de buitenwereld dient te worden gecommuniceerd. De noodoplossingen die in de overgangsperiode gebruikt worden zijn in principe als (tijdelijke) terugvalscenario op te vatten.

4.3.2 Ontwikkelen terugvalscenario

Bij het tijdelijk wegvallen van een secundaire bron zal allereerst moeten worden nagegaan of het mogelijk is om langs andere weg tijdig tot kwalitatief acceptabele uitkomsten te komen. Als het mogelijk is de ontbrekende gegevens uit alternatieve secundaire bronnen of door eigen waarneming te verkrijgen, zonder dat dit aanzienlijke kosten met zich meebrengt, verdient dit in principe de voorkeur. Het CBS zal zijn cijfers namelijk altijd zoveel mogelijk op directe statistische waarneming willen baseren. Wanneer blijkt dat een dergelijke aanpak op korte termijn niet mogelijk is, kan een modelmatige aanpak gebruikt worden. Gezien de beperkte tijd die beschikbaar is zal een dergelijk model reeds ontwikkeld (en getest) moeten zijn om ingezet te kunnen worden. Indien geen van de voorgestelde oplossingen gebruikt kan worden, moet tot uitstel van de publicatie worden besloten. Overigens bestaat altijd de mogelijkheid dat de verantwoordelijke directeur in specifieke gevallen van deze voorkeursvolgorde afwijkt. In de praktijk heeft men doorgaans slechts enkele dagen de tijd en zal er in veel gevallen dan ook niet veel anders opzitten dan naar bevind van zaken te handelen en te proberen de op dat moment verstandigste beslissing te nemen.

Uit het bovenstaande blijkt dat bij het tijdelijk of gedeeltelijk wegvallen van een secundaire bron er in principe twee manieren zijn om alsnog tot een kwalitatief acceptabele uitkomst te komen:

1) andere bronnen gebruiken (primair of secundair)
2) modelmatig schatten.

Uiteraard is het denkbaar deze mogelijkheden te combineren. Indien dit niet mogelijk is, is de laatste optie:

3) de publicatie uitstellen.

Hoe met deze mogelijkheden moet worden omgegaan, wordt hieronder nader besproken. Risico van het inzetten van een terugvalscenario is een mogelijke trendbreuk in de uitkomsten. Repareren van een trendbreuk is geen eenvoudige zaak en doorgaans ook niet goed mogelijk binnen de korte termijn waarop een terugvalscenario wordt ingezet. Tijdens het ontwikkelen van een terugvalscenario moet getracht worden ook hier zo goed mogelijk mee om te gaan.

---

[4] De Delphi-methode is een onderzoeksmethode waarbij de meningen van een groot aantal experts over een onderwerp wordt gevraagd waar geen consensus over bestaat. Door de antwoorden van de andere experts (anoniem) terug te koppelen wordt geprobeerd in een aantal rondes tot consensus te komen. De methode is genoemd naar het orakel van Delphi.

4.3.2.1 Andere bronnen gebruiken

Als het mogelijk is om de weggevallen broninformatie snel en exact, zonder aanzienlijke kosten, langs andere weg te verkrijgen - door eigen waarneming of uit alternatieve secundaire bronnen die dezelfde gegevens bevatten - wordt aan deze optie de voorkeur gegeven boven modelmatig schatten. Het CBS zal zijn cijfers altijd zoveel mogelijk op directe statistische waarneming willen baseren. Het direct (telefonisch) benaderen van de contactpersoon van de (sub)leverancier van de bron is een voorbeeld van een methode om de ontbrekende gegevens alsnog, mogelijkerwijs op een minder detailniveau, te verkrijgen (Frenken et al., 2008). Gebruik maken van een andere (secundaire) bron met identieke of soortgelijke gegevens is een andere optie. Indien er een andere leverancier kan worden gevonden, dient met het verkrijgen van de gegevens niet veel tijd gemoeid te zijn. Bij het gebruiken van een nieuwe bron is het zeer waarschijnlijk dat een trendbreuk zal optreden.

4.3.2.2 Modelmatig schatten

Een modelmatige aanpak kan uitkomst bieden wanneer slechts een deel van de gegevens is ontvangen. Dit kan het ontbreken van een deel van de data van alle berichtgevers zijn, zoals het ontbreken van de gegevens over de derde periode van een kwartaalstatistiek. Maar ook het ontbreken van alle gegevens van een deel van de populatie, bijvoorbeeld als gevolg van een (niet goed werkende) foutendetectieroutine van de bronhouder of onherstelbare ICT-gerelateerde opslagfouten of verzendproblemen. Essentiële voorwaarden voor het gebruiken van een modelmatige aanpak zijn dat:

i) een deel van de gegevens beschikbaar is

ii) reeds een model is ontwikkeld.

De benodigde onderzoekstijd die voor het ontwikkelen en toetsen van een model nodig is en het vaak ontbreken van voldoende lange tijdreeksen vormen een obstakel voor de inzet van modellen als terugvalscenario. Daarnaast is het ook de vraag of een volgens de regels ontwikkeld en getest model überhaupt kan worden toegepast op het specifieke probleem dat zich op dat moment voordoet. Dit vormt een verdere beperking voor het toepassen van de modelmatige aanpak. Indien bij het ontwikkelen van het model niet de nodige flexibiliteit wordt betracht - deze zal moeten bestaan uit het snel kunnen bedenken en toepassen van een ad hoc terugvalscenario - zal de modelmatige benadering niet vaak (goed) kunnen worden toegepast. Dat is op zich jammer, omdat het in een flink aantal gevallen mogelijk moet zijn om op basis van kennis en ervaring een modelschatting te maken die een acceptabele kwaliteit bezit. De creatieve noodoplossingen die aan het eind van paragraaf 4.3.1 als voorbeelden genoemd zijn om de publicatie van cijfers in de overgangsfase tussen twee bronnen te overbruggen, behoren in principe ook tot het domein van de ‘modelmatige’ aanpak. Dit zijn echter zwakke modellen en verdienen niet de voorkeur. Indien deze moeten worden toegepast, dient hierover duidelijk naar de gebruikers gecommuniceerd te worden.

4.3.2.3 Publicatie uitstellen

Tot uitstel van de publicatie wordt besloten indien de mogelijkheden ontbreken om, via andere manieren van waarneming of modelmatig schatten, relatief snel tot een kwalitatief acceptabele uitkomst te komen. Uitstellen is uitsluitend een optie wanneer te verwachten valt dat:

i) de problemen met betrekking tot de bron binnen een bepaalde (bij voorkeur korte) termijn kunnen worden opgelost

ii) het imago van het CBS hierdoor niet heel erg wordt beschadigd (maatschappelijk belang).

Indien bij één of beide punten problemen zijn te verwachten is uitstellen geen optie meer. Voor de imagobepalende statistieken van het CBS is uitstel van publicatie nooit een toe te passen oplossing. Hiervoor dienen dan ook andere bronnen of een model beschikbaar te zijn.

4.4 Voorbeelden

In eerste instantie lijken de mogelijkheden van het gebruik van vervangende bronnen of modelaannames beperkt. Vaak wordt gedacht dat het snel opstarten van het gebruik van nieuwe secundaire bronnen, het (op)nieuw zelf gaan waarnemen of het ontwikkelen van een model de nodige tijd in beslag zal nemen. Dit kan tevens met aanzienlijke kosten gepaard gaan. In de CBS-praktijk heeft zich een aantal situaties voorgedaan dat het belang van terugvalscenario's illustreert. Belangrijk is dat de voorgestelde maatregelen reëel zijn; soms zijn het zelfs voor de hand liggende opties. Enkele voorbeelden van terugvalscenario's uit de CBS-praktijk illustreren dit.

1. Een voorbeeld van het gebruik van primaire waarneming ter vervanging:
   - Probleem: Een centraal orgaan in de gezondheidszorg blijft in gebreke bij het tijdig verstrekken van de verpleegdagtarieven.
   - Oplossing: De ziekenhuizen werden telefonisch benaderd om de benodigde tarieven te verkrijgen.
2. Een voorbeeld van het gebruik van een secundaire bron ter vervanging:
   - Probleem: De scannerdata met daarin de prijsgegevens van producten van een grote supermarktorganisatie laten op zich wachten.
   - Oplossing: De prijsgegevens werden van de internetsite van de betreffende supermarktorganisatie gehaald. Deze oplossing verdiende de voorkeur boven imputeren op basis van het prijsverloop van andere supermarkten.
3. Een voorbeeld van het gebruik van een modelaanname:
   - Probleem: Een brancheorganisatie in de energiemarkt blijft in gebreke bij het tijdig verstrekken van informatie over de ontwikkeling van de elektriciteitstarieven.

- Oplossing: Met behulp van informatie die het CBS wel ter beschikking stond werd verwacht dat de elektriciteitstarieven (met terugwerkende kracht) fors zouden worden verhoogd. Hierdoor werd de kans reëel geacht dat de buitenwereld verkeerd zou worden geïnformeerd met een veel te laag inflatiecijfer. Er werd besloten om een raming te maken van de te verwachten stijging. Die raming is vervolgens ingezet. Dit bleek achteraf een verstandige beslissing te zijn.

4.5 Kwaliteitsindicatoren

In het VIR-besluit (VIR, 2007) zijn de beschikbaarheid van informatiesystemen (waarvan gegevensverzamelingen deel uitmaken), risicoafwegingen en te nemen maatregelen verwoord. In dit kader is de problematiek van terugvalscenario's ingepast (verslag CBS-directieberaad, 26 januari 2009). Het VIR beschikt over standaardprocesmaatregelen ('Afhankelijkheids- en kwetsbaarheidsanalyses'), die periodiek geëvalueerd en gemonitord worden. Als onderdeel van de periodieke evaluatie in het kader van het VIR worden van een proces de discontinuïteitrisico's als gevolg van niet tijdige levering of onvoldoende kwaliteit meegenomen.

Voor antwoord op de vraag of het raadzaam is voor een bepaalde statistiek een terugvalscenario te ontwikkelen is een standaardsjabloon ontwikkeld. Met behulp van het onderdeel risicobepaling in de sjabloon wordt ingeschat of het nodig wordt geacht een terugvalscenario op te stellen. Belangrijk hiervoor zijn de inschatting van problemen met de levering van de bron, de stabiliteit van de levering en de gevolgen voor het CBS. Indien de kans op uitval groot wordt geacht én de gevolgen voor het CBS groot zijn, wordt aanbevolen een terugvalscenario op te stellen. Omdat dit laatste voor de imagobepalende statistieken van het CBS altijd het geval is, is besloten voor deze groep van statistieken, indien ze direct of indirect van secundaire bronnen afhankelijk zijn, altijd een terugvalscenario op te stellen (verslag CBS-directieberaad, 26 januari 2009). In alle andere gevallen is het aan de proceseigenaar om te besluiten of een terugvalscenario ontwikkeld dient te worden. De volledige inhoud van de sjabloon voor terugvalscenario's is in tabel 1 weergegeven.

Tabel 1. Sjabloon evaluatie voor terugvalscenario

Om welke statistiek gaat het?

- Naam statistiek
- Divisie, sector, taakgroep
- Maakt gebruik van de volgende secundaire bronnen: …

Algemene informatie per secundaire bron

- Naam register
- Naam registerhouder
- Contactpersoon bij register
- CBS-accountmanager
- Overige CBS-contactpersonen
- Zijn er reguliere contacten tussen de bronhouder en het CBS?

Risicobepaling

- Hoe groot wordt het risico ingeschat dat de bronhouder de bron niet kan leveren?
- Wat zijn de gevolgen voor het CBS?
- Hoe stabiel is de bron?

Procesinformatie van de statistiek

- Zijn er alternatieve bronnen of is er bestaand onderzoek waaruit blijkt dat de gegevens modelmatig kunnen worden afgeleid indien de bron of de benodigde variabelen niet beschikbaar zijn?
- Terugvalscenario: 1. wachten; 2. modelschatting; 3. alternatieve bron gebruiken

Samenvatting

- Risico op niet (tijdig) kunnen publiceren van de statistiek
- Gevolgen voor het CBS
- Beschikbare alternatieven

Meta-informatie checklist

- Frequentie waarmee de checklist wordt geactualiseerd
- Datum laatste update
- Opgesteld door:
- Ondertekend door (naam en functie)

5. Afsluiting

De toename van het gebruik van gegevens in secundaire bronnen door het CBS leidt ertoe dat het bureau steeds meer afhankelijk wordt van de verzameling van gegevens door anderen. Hier zal bij de (her)inrichting van de statistische processen op het CBS in toenemende mate rekening mee moeten worden gehouden. Dit maakt dat het CBS zich meer extern moet gaan focussen. Er dient actief naar secundaire bronnen gezocht te worden en het contact met de bronhouders dient, op management-, proces- en expert-niveau, actief onderhouden te worden. Daarnaast is het essentieel dat het CBS de invloed op de inhoud van secundaire bronnen vergroot. Dit om ervoor te zorgen dat bronnen de voor het CBS relevante gegevens verzamelen en blijven verzamelen. Om dit geheel te bewaken is implementatie van het in dit document beschreven kwaliteitskader noodzakelijk.

6. Literatuur

Aelen, F.W.L. (2004), *Startcursus statistiekboom – Bedrijvensregister*. Versie, 28 september, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Aelen, F. (2008), *HEcS+ Eenhedenproblematiek, Eindrapportage*. Versie 1.0p2, 22 september, Project-nota, Centraal Bureau voor de Statistiek, Heerlen/Voorburg.

Arends-Tóth J. en Daas, P. J. H. (2008), *Kwaliteitsindicatoren van registers: de Metadata hyperdimensie*. CBS-nota, Sector Methodologie, Centraal Bureau voor de Statistiek, Heerlen.

Arts, C.H., Bakker, B.F.M. and Lith, F.J. van (2000), Linking administrative registers and household surveys. *Netherlands Official Statistics* 15, pp.16-21.

Arts, C.H. en Hoogteijling, E.M.J. (2002), Het Sociaal Statistisch Bestand 1998 en 1999. *Sociaal Economische Maandstatistiek* 12, blz. 13-21.

Belastingdienst (2008), *Btw algemeen*. Informatie op de website van de Belastingdienst (http://www.belastingdienst.nl/zakelijk/omzetbelasting/ob01/ob01.html).

CBP (College Bescherming Persoonsgegevens) (2009), Website van het College Bescherming Persoonsgegevens, (http://www.cbpweb.nl/).

CBS (Centraal Bureau voor de Statistiek) (2003a), Besluit gegevensverwerving CBS. *Staatsblad van het Koninkrijk der Nederlanden*, 552.

CBS (Centraal Bureau voor de Statistiek) (2003b), Wet op het Centraal Bureau voor de Statistiek. *Staatsblad van het Koninkrijk der Nederlanden*, 516.

CBS (Centraal Bureau voor de Statistiek) (2008a), *CBS Kwaliteitsverklaring*, 28 april, Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

CBS (Centraal Bureau voor de Statistiek) (2008b), *Dataverzameling*. Informatie over CBS-onderzoeken op de website van het Centraal Bureau voor de Statistiek (http://www.cbs.nl/nl-NL/menu/methoden/dataverzameling/default.htm).

Cobben, F. and Schouten, B. (2008), *An empirical validation of R-indicators*. Discussion paper 08006, Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Daas, P. J. H. en Arends-Tóth, J. (2007), *Kwaliteitsindicatoren voor registers: evaluatie van de Bron hyperdimensie*. CBS-nota, Sector Methodologie, Centraal Bureau voor de Statistiek, Heerlen.

Daas, P.J.H., Arends-Tóth, J., Schouten, B. and Kuijvenhoven, L. (2008), Quality Framework for the Evaluation of Administrative Data. In: *Proceedings of Q2008 European Conference on Quality in Official Statistics*. Statistics Italy and Eurostat, Rome.

Daas, P. en Beukenhorst, D. (2008), *Databronnen van het CBS: Primaire en secundaire bronnen*. CBS-nota, Sector Methodologie, Centraal Bureau voor de Statistiek, Heerlen.

Daas, P., Jeurissen, E., Boonstra, H.J. en Nieuwenbroek, N. (2005), *Registertheorie: Registers en het CBS*. CBS-nota, Sector Methoden en Ontwikkeling, Centraal Bureau voor de Statistiek, Heerlen.

Daas, P.J.H. en Prins, C.J.M. (2007), *Programma van eisen voor basisregistraties en voor het statistisch gebruik van deze registraties*. CBS-nota, Sector Methodologie en Sector Ontwikkeling en Ondersteuning, Centraal Bureau voor de Statistiek, Heerlen/Voorburg.

Daas, P., Roos, M. en Puts, M. (2008), *Waarnemingsinnovatie: nieuwe bronnen en mogelijkheden*. Sector Methodologie, Centraal Bureau voor de Statistiek, Heerlen.

Erwich, B. en Van Maarseveen, J.G.S.J. (1999), *Een eeuw statistieken, historisch-methodologische schetsen van de Nederlandse officiële statistieken in de twintigste eeuw*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Eurostat (2003), *Definition of quality in statistics*. Item 4.2: Methodological documents, Working group 'Assessment of the quality in statistics', Sixth meeting, Luxembourg, 2-3 October.

Eurostat (2005), *Standard quality indicators*. Quality in statistics, Seventh meeting, Luxembourg, 23-24 May.

E-overheid (2009), *E-overheid planningsoverzicht / voortgangsrapportages*. Webpagina met toegang tot de voortgangsrapportages van de basisvoorzieningen van de Elektronische overheid (http://www.e-overheid.nl/atlas/planning/planning.html).

Frenken, J., Hoven, L., Prins, K. en Slootbeek, G. (2008), *CBS-beleid terugvalscenario's*. Versie 5 september CBS-nota, Centraal Bureau voor de Statistiek, Heerlen/Voorburg.

Golden, M.P. (1976), *The research experience*. F.E. Peacock Publishers Inc., Itasca, Illinois, USA.

Huigen, R. (2006), *ICT Masterplan: CBS-architectuur, Business- en informatiemodel*. Versie 1.0, 3 april, Centraal Bureau voor de Statistiek, Heerlen/Voorburg.

Karr, A. F., A. P. Sanil, and D. L. Banks (2006), Data quality: A statistical perspective. *Statistical Methodology*, 3, pp. 137-173.

Kuijvenhoven, L. en Schouten, B. (2008), *Kwaliteitsaspecten voor de hyperdimensie data*. CBS-nota, Sector Methodologie, Centraal Bureau voor de Statistiek, Voorburg.

Prins, K. en Kuijper, H. (2007), Bevolkingsstatistieken onder het persoonskaartenstelsel en het GBA-stelsel: overeenkomsten en verschillen. *Bevolkingstrends* 55(1), blz.14-33.

Schouten, B. and Cobben, F. (2007), *R-indexes for the comparison of different fieldwork strategies and data colection modes*. Discussion paper 07002, Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Schulte Nordholt, E., Hartgers, N. and Gircour, R. (2004), *The Dutch virtual census of 2001, Analysis and methodology*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Statistics Denmark (1995), *Statistics on persons in Denmark, a register-based statistical system*. Office for Official Publications of the European Communities, Luxembourg.

Statistics Finland (2004), *Use of registers and administrative data sources for statistical purposes*. Best practices of Statistics Finland. Handbook 45.

‘t Hart, H., Boeije, H. en Hox, J. (2005), *Onderzoeksmethoden*. 7e druk, Uitgeverij Boom, Amsterdam.

VIR (Voorschrift Informatiebeveiliging Rijksdienst) (2007) Besluit voorschrift informatiebeveiliging rijksdienst 2007. *Staatscourant* 122, blz. 1-7.

Van den Elshout, B., Jacobi, C. en Van de Valk, J. (2007), Werklozen versus niet-werkende werkzoekenden. *Sociaaleconomische trends*, 3, p. 46-53.

Van der Bie, R.J. en Smits, J-P. (2001), *Tweehonderd jaar statistiek in tijdreeksen 1800–1999*. Centraal Bureau voor de Statistiek en Rijksuniversiteit Groningen, Voorburg/Heerlen, Groningen.

Verlinden, B. (2008), *Procesgang ConGO in relatie tot gebruik registers*. Versie 27 maart. CBS-nota, Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology, John Wiley & Sons, Ltd, Chichester, England.

Bijlage A. Overzicht basisregistraties

Momenteel zijn er elf aangewezen basisregistraties in Nederland (E-overheid, 2009). De aangewezen basisregistraties zijn:

1) Gemeentelijke Basisadministratie persoonsgegevens (GBA)
   Bevat de persoonsgegevens van de ingezetenen van de Nederlandse gemeenten. Op 1 april 2007 is de GBA als basisregistratie in werking getreden met een overgangsperiode tot 1 januari 2010. De GBA gaat samen met de Basisregistratie Niet-ingezetenen, de basisregistratie personen vormen.

2) Nieuwe Handelsregister (NHR)
   Gaat de gegevens van alle bedrijven, ondernemingen, rechtspersonen, eenmanszaken in de landbouw, maatschappen, vrije beroepsbeoefenaars, verenigingen van eigenaars en de gehele publieke sector, zoals ministeries, gemeenten en zelfstandige bestuursorganen, in Nederland bevatten. De wet NHR is op 1 juli 2008 in werking getreden met een overgangsperiode tot 1 januari 2014.

3) Basis Gebouwenregistratie (BGR)
   Register van alle panden, verblijfsobjecten, standplaatsen en ligplaatsen, inclusief de identificerende en beschrijvende kenmerken, zoals bouwjaar en geometrie. De BGR werkt samen met de Basis Registratie Adressen onder de naam Basisregistratie Adressen en Gebouwen (BAG). Verwacht wordt dat de wet waarin het BGR is opgenomen medio 2009 in werking treedt met een overgangsperiode van 2 jaar.

4) Basisregistratie Adressen (BRA)
   Bevat alle adressen van adresseerbare objecten (verblijfsobjecten, ligplaatsen van onder meer woonboten en standplaatsen van onder meer woonwagens) in Nederland. De BRA en de Basis Gebouwenregistratie werken samen onder de naam Basisregistratie Adressen en Gebouwen (BAG). Verwacht wordt dat de wet waarin het BRA is opgenomen medio 2009 in werking treedt met een overgangsperiode van 2 jaar.

5) Basisregistratie Topografie (BRT)
   Register dat de gedigitaliseerde gegevens van geografische objecten (bijvoorbeeld wegdeel, waterdeel, spoorbaandeel, terrein) in een schaal van 1:10.000 bevat. De BRT werkt nauw samen met de basisregistratie Kadaster. De wet waartoe de BRT behoort, is op 1 januari 2008 in werking getreden met een overgangsperiode tot 1 januari 2009. Landsdekkende topografische bestanden op een schaalniveau kleiner dan 1:10 000 zullen op 1 januari 2010 beschikbaar zijn.

6) Basisregistratie Kadaster (BRK)
Registratie van elk perceel in Nederland in een schaal van 1:10.000. Een perceel kan een gebouw zijn maar ook een maïsveld. De BRK werkt nauw samen met de basisregistratie Topografie. De wet waartoe de BRK behoort, is op 1 januari 2008 in werking getreden met een overgangsperiode tot 1 januari 2009. Landsdekkende bestanden op een schaalniveau kleiner dan 1:10 000 zullen op 1 januari 2010 beschikbaar zijn.

7) Kentekenregistratie (KR)
Bevat gegevens over het kenteken, het voertuig en de persoonsgegevens van de houder van het voertuig. Wordt soms ook wel Basisregistratie Voertuigen genoemd. De wet waarin de KR is vastgelegd is op 1 juli 2008 in werking getreden maar het gebruik van de authentieke gegevens uit KR is nog niet verplicht. De datum van de inwerkingtreding voor het verplicht gebruik, moet nog nader bepaald worden.

8) Basisregistratie Lonen, Arbeidsverhoudingen en Uitkeringsverhoudingen (BLAU)
Bevat gegevens over lonen, uitkeringen en dienstverbanden. Zal gebaseerd zijn op een deel van de gegevens in de huidige Polisadministratie. De aanvraag wordt ambtelijk voorbereid. De inwerkingtreding is op zijn vroegst voorzien in 2009 met een overgangsperiode.

9) Basisregistratie Waardering Onroerende Zaken (WOZ)
Bevat de waarde en het belang van zogenaamde 'Waardering Onroerende Zaken' (WOZ) objecten. Dit zijn bewoonde of onbewoonde eigendommen of gedeelten daarvan. Het wetsvoorstel voor de basisregistratie WOZ is, samen met dat voor de Basisregistratie Inkomen, in 2008 parlementair aanvaard. De inwerkingtreding is voorzien voor 2009 met een overgangsperiode.

10) Basisregistratie Inkomen (BRI)
Bevat inkomensgegevens van personen. Het wetsvoorstel voor de BRI is, samen met dat voor de basisregistratie Waardering Onroerende Zaken, in 2008 parlementair aanvaard. De inwerkingtreding is op zijn vroegst voorzien in 2009 met een overgangsperiode.

11) Registratie van Niet-ingezetenen (RNI)
Bevat de gegevens van personen die niet in Nederland wonen maar er wel een relatie mee hebben; bijvoorbeeld omdat ze hier (tijdelijk) werken. Bij inschrijving wordt aan de niet-ingezetene (mits goed identificeerbaar) een BSN-nummer toegekend. Het RNI gaat samen met het GBA, de basisregistratie personen vormen. De voorbereiding van de totstandkoming van de RNI is in de afrondende fase.

Kandidaat basisregistraties zijn:

i) Grootschalige Basiskaart Nederland (GBKN)

   Zal topografische gegevens (schaal circa 1:100 tot 1:5000) bevatten van belangrijke topografische objecten (o.a. gebouwen, wegen, bruggen, spoorlijnen, waterlopen, dijken en hoogspanningsmasten) in Nederland. Er is tevens een basiskaart voorzien die als ondergrond dient. De verkenningsfase is in 2008 gestart. Verwacht wordt dat in 2009 het besluit tot voorbereiding zal worden genomen.

ii) Basisregistratie Ondergrond (BRO)

   Deze registratie zal gegevens over verkenningen en gebruiksrechten van en ruimtelijke (geologische) en infrastructurele objecten in de ondergrond (bodem) bevatten. Voor heel Nederland zal een geologisch model van de bodem en een bodemkaart worden opgesteld. Een haalbaarheidsonderzoek is in 2008 afgerond met positief resultaat. Verwacht wordt dat in 2009 definitief besloten wordt tot oprichting van dit register. De realisatiefase zal in 2011 worden gestart.

Bijlage B. Overzicht kwaliteitsindicatoren voor secundaire bronnen

Tabel B1. Dimensies, kwaliteitsindicatoren en meetmethoden van Bron

DIMENSIES	KWALITEITSINDICATOREN	MEETMETHODEN
1. Leverancier	1.1. Contact	- Naam bron - Contactgegevens bronhouder/beheerder - Contactpersoon CBS
	1.2. Doel	- Reden gebruik bron door CBS
2. Relevantie	2.1. Nut	- Belang bron voor het CBS
	2.2. Mogelijk gebruik	- Potentieel gebruik bron voor statistieken; ter vervanging, aanvulling (welke) of nieuw
	2.3. Informatiebehoefte	- Voldoet aan informatiebehoefte v/h CBS
	2.4. Lastendruk	- Gevolgen op lastendruk (incl. berekening)
3. Privacy en beveiliging		
	3.1. Wettelijke basis	- Grondslag van bijhouden bron
	3.2. Vertrouwelijkheid	- Is de Wbp[1] van toepassing? - Is gebruik door het CBS aangemeld?
	3.3. Beveiliging	- Wijze versturen gegevens naar CBS - Beveiliging noodzakelijk (zowel soft- als hardware aspecten)
4. Levering	4.1. Kosten	- Kosten verbonden aan gebruik
	4.2. Afspraken	- Is er een leveringsovereenkomst? - Frequentie leveringen
	4.3. Stiptheid	- Hoe stipt kan geleverd worden? - Snelheid doorgifte afwijkingen - Snelheid opslag gegevens bij bronhouder
	4.4. Opmaak	- Format(s) waarin data geleverd kan worden
	4.5. Selectie	- Welke gegevens kunnen geleverd worden? - Is dit wat het CBS wil hebben?
5. Procedures	5.1. Dataverzameling	- Bekendheid met wijze van data verzamelen
	5.2. Wijzigingsplannen	- Bekendheid met wijzigingsplannen - Wijze van communiceren naar het CBS
	5.3. Terugkoppeling	- Mag het CBS problemen terugkoppelen? - Wat wel, wat niet en waarom
	5.4. Terugvalscenario	- Afhankelijkheidsrisico voor het CBS - Maatregelen bij het niet leveren volgens afspraak (zie: sjabloon terugvalscenario's)

[1] Wbp: Wet bescherming persoonsgegevens.

Tabel B2. Dimensies, kwaliteitsindicatoren en meetmethoden van Metadata

DIMENSIES	KWALITEITSINDICATOREN	MEETMETHODEN
1. Duidelijkheid	1.1 Populatie-eenheden definitie	- Score voor de duidelijkheid van de omschrijving
	1.2 Kwalitatieve variabelen definitie	- Score voor de duidelijkheid van de omschrijving
	1.3 Kwantitatieve variabelen definitie	- Score voor de duidelijkheid van de omschrijving
	1.4 Tijdsdimensie definitie	- Score voor de duidelijkheid van de omschrijving
	1.5 Definitiewijzigingen	- Bekendheid en score voor de duidelijkheid van de omschrijving van wijzigingen
2. Vergelijkbaarheid		
	2.1 Populatie-eenheden	- Overeenkomst met CBS-definitie
	2.2 Kwalitatieve variabelen	- Overeenkomst met CBS-definitie
	2.3 Kwantitatieve variabelen	- Overeenkomst met CBS-definitie
	2.4 Tijdsdimensie	- Overeenkomst met CBS-definitie
3. Unieke sleutels	3.1 Identificerende sleutel	- Aanwezigheid unieke sleutel - Relatie met CBS-unieke sleutel
	3.2 Unieke combinaties van variabelen	- Aanwezigheid bruikbare combinaties (als identificerende sleutel)
4. Databehandeling door registerhouder		
	4.1 Controles	- Bekendheid met controles van populatie-eenheden - Bekendheid met controles van variabelen - Bekendheid met controles van combinaties van variabelen - Bekendheid met controles van extreme waarden (uitbijters)
	4.2 Aanpassingen / bewerkingen	- Bekendheid met aanpassingen - Markering bewerkte velden - Bekendheid met gebruik van standaardwaarden

Tabel B3. Dimensies, kwaliteitsindicatoren en meetmethoden van Data

DIMENSIES	KWALITEITSINDICATOREN	MEETMETHODEN
1. Technische controles		
	1.1 Leesbaarheid	- Zijn alle gegevens toegankelijk?
	1.2 Voldoen aan metadata	- Voldoet de data aan de metadata definitie?
		- Indien niet, rapporteer de afwijkingen
2. Overdekking	2.1 Niet tot de doelpopulatie behorende eenheden	
		- Percentage records dat niet tot doelpopulatie behoort
3. Onderdekking	3.1 Ontbrekende eenheden	- Percentage ontbrekende eenheden dat tot de doelpopulatie behoort
	3.2 Selectiviteit	- R-index[1] voor samenstelling eenheden
	3.3 Effect op gemiddelde	- Maximale vertekening van gemiddelde
		- Maximale RMSE[2] van gemiddelde
4. Koppelbaarheid	4.1 Koppelbare eenheden	- Percentage eenheden dat (eenduidig) gekoppeld kan worden
	4.2 Miskoppelingen	- Percentage eenheden dat niet correct is gekoppeld
	4.3 Selectiviteit	- R-index voor samenstelling van gekoppelde eenheden
	4.4 Effect op gemiddelde	- Maximale vertekening van gemiddelde
		- Maximale RMSE van gemiddelde
5. Unit-nonrespons	5.1 Eenheden zonder gegevens	- Percentage eenheden waarvan alle gegevens ontbreken
	5.2 Selectiviteit	- R-index voor samenstelling eenheden
	5.3 Effect op gemiddelde	- Maximale vertekening van gemiddelde
		- Maximale RMSE van gemiddelde
6. Item-nonrespons	6.1 Ontbrekende waarden	- Percentage velden dat ontbreekt
	6.2 Selectiviteit	- R-index voor samenstelling
	6.3 Effect op gemiddelde	- Maximale vertekening van gemiddelde
		- Maximale RMSE van gemiddelde van variabele
7. Meting	7.1 Externe controle	- Is een audit of parallelle toets gedaan?
		- Is de vragenlijst getest in vragenlab?
	7.2 Onverenigbare records	- Fractie geschonden editregels
	7.3 Vertekening	- Omvang van de vertekening (relatieve meetfout)
8. Verwerking	8.1 Gaafmaken	- Fractie herziene/gaafgemaakte velden
	8.2 Imputatie	- Fractie geïmputeerde velden
	8.3 Uitbijterdetectie	- Fractie velden met een uitbijtercorrectie
9. Precisie	9.1 Standaardfout	- Schatter voor standaardfout van gemiddelde
10. Gevoeligheid	10.1 Ontbrekende gegevens	- Totaal percentage lege velden
	10.2 Selectiviteit	- R-index voor samenstelling
	10.3 Totaal effect op gemiddelde	- Totaal van maximale vertekening
		- Totaal van maximale RMSE

[1] R-index: een indicator die de mate van selectiviteit van ontbrekende gegevens met beschikbare informatie uit andere bronnen schat (Schouten en Cobben 2007, Cobben en Schouten 2008).
[2] RMSE: root mean square error; een gangbare statistische maat voor de kwaliteit van een schatter. De RMSE is gelijk aan de wortel van de som van het kwadraat van de vertekening (systematische fout) en de variantie van de schatter.

Bijlage C. Lijst van imagobepalende statistieken (2008)

Naam statistiek (afkorting)	*Gebruikte secundaire bronnen*
Bevolkingsstatistieken	- Gemeentelijke Basisadministratie (GBA)
Bijstand en Re-integratie door gemeenten	- Bijstanduitkeringen
Consumentenprijsindex (CPI)	- Scannerdata supermarkten
Economische groei per kwartaal	- Indirect afhankelijk (via KS, IH, EBB, Werkgelegenheid per kwartaal en CPI)
Enquête Beroepsbevolking (EBB)	- GBA, Centrum Werk en Inkomen, Bronnen Belastingdienst, Polisadministratie
Internationale Handel in Goederen (IH)	- Belasting Toegevoegde Waarde (BTW)
Inkomens Panel Onderzoek en Regionale Inkomensonderzoek	- Bronnen Belastingdienst, Studiefinanciering
Korte termijn statistieken (KS)	- BTW-gegevens
Overheidstekort en -schuld, EMU-criteria	- Polisadministratie
Producentenprijzen	- Niet afhankelijk van secundaire bronnen
Productiestatistieken (PS)	- BTW-gegevens, Vennootschapsbelasting
Sociaal Statistisch Bestand (SSB), satelliet Integratie	- Indirect afhankelijk (via SSB-kern)
Veiligheidsmonitor	- GBA, Bronnen Belastingdienst
Verdeelmaatstaven SZW/BZK	- GBA, Adres coördinaten Nederland, Top10Vector bestand, Geografisch Basis register
Werkgelegenheid per kwartaal	- Polisadministratie
Werkloosheid per maand	- Indirect afhankelijk (via EBB).