Het meten van de kwaliteit van administratieve bronnen: Recente resultaten en toekomstige ontwikkelingen

Piet J.H. Daas, Saskia J.L. Ossen en Martijn Tennekes (CBS)

1. Inleiding

Nationale Statistische Instituten (NSI's) hebben gegevens nodig om statistieken te kunnen maken. Veel van die gegevens worden met behulp van vragenlijsten verzameld. Steeds vaker maken NSI's echter ook gebruik van gegevens uit bronnen die door andere organisaties verzameld zijn. Voorbeelden van dergelijke bronnen zijn registers en administratieve bronnen (Wallgren en Wallgren, 2007). De gegevens in die bronnen worden gebruikt voor en zijn het gevolg van administratieve processen. In de praktijk blijken ze vaak ook erg interessant te zijn voor de statistiek. Dit besef is in de afgelopen 10 jaar bij steeds meer NSI's doorgedrongen (Unece 2007). De belangrijkste voordelen van het gebruik van administratieve bronnen en registers voor de statistiek zijn: i) reductie van de kosten van het verzamelen van de gegevens en ii) lastendrukvermindering voor bedrijven en personen. Omdat administratieve bronnen vaak gegevens over hele populaties bevatten, in verschillende tijdsperioden, zijn ze ook erg geschikt om te gebruiken voor (virtuele) volkstellingen (Schulte Nordholt, 2004), voor gedetailleerde longitudinale analyse van (sub)populaties en regio's (Wallgren en Wallgen, 2007) en voor cohortonderzoek van studenten (Chowdry et al., 2010).

Vanuit statistisch oogpunt bekeken kleven aan het gebruik van administratieve bronnen echter ook enkele nadelen. Deze zijn voornamelijk het gevolg van het feit dat de verzameling en verwerking van de gegevens niet door het NSI wordt uitgevoerd. Het is de beheerder van de bron (de 'bronhouder') die deze taken verricht. Een soortgelijk probleem treedt op bij de eenheden en variabelen die in een administratieve bron zijn opgeslagen. De definitie hiervan wordt uitsluitend door de administratieve regels van de bronhouder bepaald. Hierdoor kunnen de gehanteerde definities afwijken van degene die door het NSI worden gebruikt (Wallgren en Wallgen, 2007). Het is dan ook niet altijd even eenvoudig om de statistische bruikbaarheid van een administratieve bron te bepalen (Bakker et al., 2008).

Omdat de kwaliteit van statistieken sterk wordt beïnvloed door de kwaliteit van de gegevens die aan het begin van het statistisch proces liggen, is het van essentieel

belang dat NSI's de kwaliteit van administratieve bronnen eenduidig en efficiënt kunnen bepalen. Deze constatering vormde de aanleiding voor het ontwikkelen van een kwaliteitsraamwerk voor administratieve bronnen en registers op het Centraal Bureau voor de Statistiek (CBS). Dit raamwerk moet het mogelijk maken om de statistische bruikbaarheid (de kwaliteit) van extern verzamelde, secundaire, gegevens, aan het begin van het proces, op een efficiënte en transparante manier te bepalen (Daas et al., 2008).

2. *Kwaliteitsraamwerk*

Op het CBS is een uitgebreide literatuurstudie uitgevoerd om de verschillende kwaliteitsaspecten van administratieve bronnen te identificeren. Deze studie liet zien dat het perspectief op de kwaliteit van dergelijke bronnen in de diverse publicaties erg kan verschillen (Daas et al., 2008). Afhankelijk van het gehanteerde perspectief traden andere kwaliteitsaspecten op de voorgrond. Deze constatering is op zich niet nieuw. Dergelijke perspectieven worden vaak categorieën (Batini en Scannapieco. 2006) of hyperdimensies genoemd (Karr et al., 2006). De laatste term zal in de rest van dit hoofdstuk worden gebruikt.
De verschillende perspectieven die in de literatuurstudie werden geïdentificeerd bleken met drie hyperdimensies volledig beschreven te kunnen worden (Daas et al., 2008). De hyperdimensies werden Bron, Metadata en Data genoemd. De drie hyperdimensies vormen de basis van het ontwikkelde kwaliteitsraamwerk voor administratieve bronnen en registers. Elke hyperdimensie in het raamwerk is opgebouwd uit een aantal dimensies, waarbij elke dimensie een aantal kwaliteitsindicatoren bevat. Een kwaliteitsindicator wordt gemeten of geschat door één of meerdere meetmethoden die kwantitatief of kwalitatief kunnen zijn (Daas et al., 2008).
Dit hoofdstuk begint met een bespreking van de kwaliteitsaspecten en de daarbij behorende meetmethoden die voor de hyperdimensies Bron en Metadata zijn ontwikkeld. Vervolgens worden inzichten beschreven voor het bepalen van de kwaliteit van de gegevens die tot de hyperdimensie Data behoren.

2.1 Hyperdimensies Bron en Metadata

Een NSI dat van plan is een administratieve bron als inputbron voor de statistiek te gaan gebruiken dient als eerste de kwaliteitsaspecten die met de levering van de bron te maken hebben te bepalen. Deze kwaliteitsaspecten behoren tot de Bron hyperdimensie van het kwaliteitsraamwerk. Tabel 1 geeft een overzicht van de dimensies, kwaliteitsindicatoren en meetmethoden voor de Bron hyperdimensie.
De Metadata hyperdimensie richt zich op de conceptuele en proces gerelateerde kwaliteitsaspecten van de metadata van de bron. Het is uitermate belangrijk dat een NSI de metadata gerelateerde kwaliteitsaspecten volledig begrijpt. Elk misver-

Tabel 1
Kwaliteitsraamwerk voor registers, hyperdimensie Bron

DIMENSIES	KWALITEITSINDICATOREN	MEETMETHODEN
1. Leverancier	1.1 Contact	– Naam databron – Contactgegevens bronhouder/beheerder – Contactpersoon NSI
	1.2 Doel	– Reden gebruik bron door NSI
2. Relevantie	2.1 Nut	– Belang bron voor NSI
	2.2 Mogelijk gebruik	– Potentieel gebruik bron voor statistiek
	2.3 Informatiebehoefte	– Voldoet de bron aan informatiebehoefte v/h NSI
	2.4 Lastendruk	– Gevolgen gebruik op lastendruk van NSI
3. Privacy en beveiliging	3.1 Wettelijke basis	– Grondslag voor bestaan v/d bron
	3.2 Vertrouwelijkheid	– Is WBP van toepassing? – Is gebruik door NSI aangemeld?
	3.3 Beveiliging	– Wijze versturen gegevens naar NSI – Noodzaak beveiliging (soft- en hardware)
4. Levering	4.1 Kosten	– Kosten verbonden aan gebruik door NSI
	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 NSI wil hebben?
5. Procedures	5.1 Data verzamelen	– Bekendheid met wijze van data verzamelen
	5.2 Wijzigingsplannen	– Bekendheid met wijzigingsplannen – Wijze van communiceren met NSI
	5.3 Terugkoppeling	– Mag NSI bij problemen terugkoppelen? – Wat wel, wat niet en waarom?
	5.4 Terugvalscenario	– Afhankelijkheidsrisico v/h NSI – Maatregelen bij het niet leveren volgens afspraak

stand of fout hierin zal de kwaliteit van de geproduceerde statistieken aanzienlijk beïnvloeden. In tabel 2 zijn de dimensies, kwaliteitsindicatoren en meetmethoden van de Metadata hyperdimensie weergegeven.

2.2 Checklist voor Bron en Metadata

Voor de verschillende kwaliteitsaspecten in de hyperdimensies Bron en Metadata is een checklist ontwikkeld (Daas et al., 2009a). De Engelstalige versie van de checklist is in de CBS-publicatie van Daas et al. (2009b) opgenomen; deze publicatie is te vinden op de website van het CBS in de rubriek 'discussion papers'. De checklist leidt de gebruiker door de meetmethoden van de kwaliteitsindicatoren in de hyperdimensies Bron en Metadata. Door de vragen in de checklist, voor de betreffende verslagperiode, te beantwoorden wordt de 'waarde' voor elke meetmethode in tabel 1 en 2 bepaald.

Bij de evaluatie van het Metadata-deel in de checklist is het noodzakelijk dat de gebruiker een specifieke statistiek waarvoor de bron wordt gebruikt in gedachten heeft. Dit is nodig omdat in dit deel de definities van de eenheden, variabelen en tijdsperiode(n) van de bronhouder vergeleken worden met de definities zoals ze door de betreffende statistiek worden gebruikt.

Tabel 2
Kwaliteitsraamwerk voor registers, hyperdimensie Metadata

DIMENSIES	KWALITEITSINDICATOREN	MEETMETHODEN
1. Duidelijkheid	1.1 Populatie definitie	– Score duidelijkheid omschrijving
	1.2 Classificatievariabele definitie	– Score duidelijkheid omschrijving
	1.3 Telvariabele definitie	– Score duidelijkheid omschrijving
	1.4 Tijdsdimensie definitie	– Score duidelijkheid omschrijving
	1.5 Definitiewijzigingen	– Bekendheid met opgetreden wijzigingen
2. Vergelijkbaarheid	2.1 Populatie definitie	– Vergelijkbaarheid met NSI-definitie
	2.2 Classificatievariabele definitie	– Vergelijkbaarheid met NSI-definitie
	2.3 Telvariabele definitie	– Vergelijkbaarheid met NSI-definitie
	2.4 Tijdsverschillen	– Vergelijkbaarheid met tijdsperioden NSI
3. Unieke sleutels	3.1 Identificerende sleutels	– Aanwezigheid unieke sleutels
		– Overeenkomst met unieke sleutels van NSI
	3.2 Unieke combinaties	– Aanwezigheid bruikbare combinaties van variabelen
4. Databehandeling (door bronhouder)	4.1 Controles	– Gebruikte controles van populatie eenheden
		– Gebruikte controle van variabelen
		– Controles op combinaties van variabelen
		– Controles op extreme waarden (uitbijters)
	4.2 Aanpassingen/bewerkingen	– Bekendheid met aanpassingen/bewerkingen
		– Worden aangepaste velden gemarkeerd?
		– Bekendheid met gebruik van standaardwaarden

Om de bruikbaarheid van de checklist te testen zijn acht secundaire databronnen van het CBS geëvalueerd. Deze bronnen zijn: de Polisadministratie (PA), het bestand Wet op de Studiefinanciering (WSF), de gegevens van werkzoekenden van het Centrum voor Werk en Inkomen (CWI; tegenwoordig het UWV WERKbedrijf), het Examen Resultatenregister (ERR), het gecoördineerde eencijferregister afgeleid uit het Centraal Register Inschrijvingen Hoger Onderwijs (1CijferHO), het gecoördineerde eencijferregister afgeleid uit de onderwijsnummerbestanden voor het Voortgezet Onderwijs (1CijferVO), de Nationale Autopas gegevens (NAP) en de Gemeentelijke Basisadministratie persoonsgegevens (GBA).
Omdat het belangrijkste doel van de studie de bruikbaarheid van de resultaten van de checklist was, werden de vragenlijsten door de gebruikers ingevuld in samenwerking met één of meerdere auteurs van dit rapport. De gebruikers waren CBS-medewerkers die betrokken zijn bij: contact met de bronhouder, ontvangst van de databron, en/of verwerking en controle van de bron. Gemiddeld duurde het ongeveer 2 uur om de checklist voor een bron in te vullen.

2.3 Resultaten checklist

De resultaten voor de acht databronnen zijn in tabellen 3 en 4 weergegeven. Tabel 3 bevat de resultaten voor de Bron hyperdimensie en tabel 4 die voor de Metadata hyperdimensie. Voor de PA is het Metadata-deel van de checklist ingevuld met het oog op gebruik voor de werkloosheidsstatistiek. De GBA is beoordeeld met gebruik voor de bevolkingsstatistieken in gedachten, terwijl de NAP bekeken is met beoogd gebruik voor de verkeer- en vervoersstatistieken. De andere bronnen zijn beoordeeld met het oog op gebruik voor het schatten van het opleidingsniveau van de Nederlandse bevolking (Bakker et al., 2008)

Tabel 3
Resultaten voor de hyperdimensie Bron

	Databronnen							
	PA	WSF	CWI	ERR	1FigHO	1FigVO	NAP	GBA
Dimensies								
1. Leverancier	+	+	+	+	+	+	+	+
2. Relevantie	+	+	+	o	+	+	+	+
3. Privacy en beveiliging	+	+	+	+	+	+/o	+	+
4. Levering	o	+	-	+	+	o	+	+
5. Procedures	+	+/o	+	+/o	+/o	+/o	o	+

Tabel 4
Resultaten voor de hyperdimensie Metadata

	Databronnen							
	PA	WSF	CWI	ERR	1FigHO	1FigVO	NAP	GBA
Dimensies								
1. Duidelijkheid	+	+	-	o	+	+	+	+
2. Vergelijkbaarheid	+/o	+	-	+	+	+	+	+
3. Unieke sleutels	+	+	+	+	+	+	+	+
4. Data behandeling	+/o	?(+)	?	?(o)	?(+)	?(+)	+	+

De resultaten van de evaluatie zijn in tabel 3 en 4 op het niveau van de dimensies weergegeven. De getoonde scores werden bepaald door de meest voorkomende score van de meetmethoden die tot die dimensie behoren te selecteren. De gebruikte symbolen voor de scores in tabel 3 en 4 zijn: goed (+), redelijk (o), slecht (-) en onduidelijk (?). Tussenliggende scores worden weergegeven door de symbolen te combineren met een scheidingsteken (/). Wanneer een onduidelijk score in een bepaalde dimensie voorkwam is dit resultaat voor de gehele dimensie getoond. Echter, wanneer daarnaast bleek dat alle andere meetmethoden in die dimensie wel duidelijk scoorden is de meest voorkomende score van die andere methoden tussen haakjes toegevoegd.

Hyperdimensie Bron

De resultaten in tabel 3 laten zien dat de scores van alle databronnen, voor het Bron-deel van de checklist, ietwat laag zijn voor de dimensies Levering en Procedures. Voor de dimensie Levering is dit voornamelijk het gevolg van het niet altijd tijdig leveren van de PA, CWI en 1FigHO bestanden. Dit duidt op een mogelijk risico voor CBS-gebruikers die erg afhankelijk zijn van de tijdige levering van deze bronnen. Het grootste probleem in de hyperdimensie Bron vormt de tijdige beschikbaarheid van de CWI. Deze bron wordt vrijwel nooit op tijd geleverd; een vertraging van enkele uren, dagen of zelfs weken is eerder regel dan uitzondering. Er is zelfs eens een periode van 3 maanden geweest waarin door de beheerder geen gegevens werden geleverd. De ietwat lage scores in de dimensie Levering voor de

PA en 1FigHO zijn overigens niet geheel onverwacht. Beide bronnen zijn nog niet helemaal uitontwikkeld; ze bestaan nog niet zo heel lang. Hierdoor fluctueren de leveringstijden nog enigszins.

In de dimensie Procedures zijn de scores ietwat laag door de relatief lage score op de terugvalscenario indicator (indicator nr. 5.4 in tabel 1). Niet alle CBS-gebruikers waren met de CBS-regel bekend dat een dergelijk scenario niet voor alle administratieve bronnen hoeft te worden opgesteld. Dit beïnvloedde de score voor de dimensie Procedures in negatieve zin. Wanneer hiermee rekening wordt gehouden komen er in deze dimensie, op één uitzondering na, eigenlijk nauwelijks problemen voor. De uitzondering was de NAP, hierbij verloopt het contact met de bronhouder wat moeizaam. Verzoeken om extra informatie worden niet altijd tijdig beantwoordt en de antwoorden die worden gegeven zijn niet altijd even verhelderend.

Hyperdimensie Metadata

De resultaten voor het Metadata-deel van de checklist zijn in tabel 4 weergegeven. Vergeleken met de resultaten voor Bron (tabel 3) zijn er duidelijk meer slechte (-) scores te zien. Deze scores hebben ook hier weer betrekking op het CWI. Deze databon scoort slecht in de dimensies Duidelijkheid en Vergelijkbaarheid. Dit is voor beide dimensies voornamelijk het gevolg van een verschil tussen de definitie van de CWI-variabele 'opleidingsniveau' en de corresponderende CBS-variabele. Bij het CWI is deze variabele namelijk duidelijk minder strikt gedefinieerd dan bij het CBS. Bij het CWI gaat het bij deze variabele ruwweg om het best bemiddelbare opleidingsniveau. Stel bijvoorbeeld dat iemand op een universiteit is afgestudeerd in een op dat moment lastig bemiddelbare studierichting. Dan kan het voorkomen dat deze persoon omgeschoold wordt, waarbij de omscholing gebeurt op een lager niveau dan universitair. Op het moment dat de omscholing voltooid is, zal het CWI als opleidingsniveau het niveau van de omscholing vermelden. Dit opleidingsniveau wordt dan immers beter bemiddelbaar geacht. In dat geval registreert het CWI dus een lager niveau dan het hoogstbehaalde (Bakker et al., 2008). Ook komt het voor dat het opleidingsniveau dat bij het CWI vermeld is, hoger is dan het hoogst behaalde opleidingsniveau. De geregistreerde in kwestie heeft dan slechts onderwijs genoten op een bepaald opleidingsniveau, maar de betreffende opleiding nooit met een diploma afgesloten.

Uit het voorgaande verhaal blijkt dat de exacte definitie van de variabele opleidingsniveau van het CWI ondermeer afhankelijk is van de arbeidsmarkt en daarmee dus tijdsafhankelijk is. De gehanteerde relatief losse definitie van opleidingsniveau levert ook problemen op bij het communiceren van een definitiewijziging, omdat het CWI-hoofdkantoor een dergelijke wijziging vaak niet precies kan duiden. Dit maakt dat de CWI ook in de dimensie Vergelijkbaarheid slechts scoort.

Van alle dimensies die tot de Metadata hyperdimensie behoren is de dimensie 'Data behandeling' het meest onduidelijk. Dit geeft aan dat op het CBS relatief weinig bekend is over de mogelijke controles, aanpassingen en bewerkingen die

door de bronhouder worden uitgevoerd. Positieve uitzonderingen hierop zijn de PA, NAP en GBA. Bij deze bronnen is de verkregen kennis over 'data behandeling' echter voornamelijk door praktijkervaring geleerd en niet door gerichte studie. Praktijkervaring is zeer belangrijk omdat er sprake kan zijn van een verschil tussen het protocol en de daadwerkelijke uitvoering bij de bronhouder.
De scores voor de databronnen in tabel 3 en 4 laten zien dat aandacht moet worden geschonken aan het CWI en dan vooral aan de kwaliteitsaspecten die tot de dimensies Levering, Duidelijkheid en Vergelijkbaarheid van die bron behoren. De geconstateerde problemen dienen voor het CWI eerst te worden opgelost voordat enige tijd mag worden besteed aan data gerelateerde kwaliteitsstudies. Pas op het moment dat de Bron- en Metadata-gerelateerde problemen voor het CWI zijn opgelost, heeft het nut om (meer) tijd te besteden aan de bepaling van de kwaliteit van de CWI-gegevens. De positieve resultaten voor de GBA laten zien dat het mogelijk is om elk kwaliteitsaspect in de Bron en Metadata hyperdimensie onder controle te hebben. Voor de andere databronnen kan geconstateerd worden dat sommige kwaliteitsgebieden meer aandacht verdienen, maar dat hierbij in het algemeen geen grote problemen werden gevonden. Voor alle bronnen, met uitzondering van het CWI, is de volgende logische stap dan ook het bepalen van de kwaliteit van de gegevens in de bron. Dit is het onderwerp van de volgende paragraaf.

2.4 Hyperdimensie Data

Tabel 5 bevat een lijst van kwaliteitsaspecten die tot de hyperdimensie Data worden gerekend. Ze zijn onder andere het resultaat van de eerder vermelde literatuurstudie (Daas et al., 2008). Veel van de indicatoren in tabel 5 zullen statistici bekend voorkomen, maar sommige mogelijk iets minder. De minder bekende indicatoren worden dan ook kort besproken.
Een aanzienlijk deel van de indicatoren in tabel 5 is gebaseerd op de zogenaamde Representativiteitsindex (R-index). Deze indicator is op het CBS ontwikkeld (Schouten et al., 2009). R-indices meten de mate waarin de samenstelling van de eenheden in een bron, op een bepaald moment, afwijkt van de beoogde doelpopulatie. Bij steekproefonderzoeken is dit een bekend concept. Voor dergelijke onderzoeken betekent representativiteit dat alle eenheden in de doelpopulatie dezelfde kans op responderen hebben. Representativiteit is echter ook bij administratieve bronnen belangrijk, zeker wanneer de samenstelling van de populatie in een dergelijke bron tijdsafhankelijk is. Een voorbeeld hiervan is de samenstelling van de bedrijven die Belasting Toegevoegde Waarde (BTW) gegevens naar de Belastingdienst opsturen. In deze bron varieert de samenstelling van bedrijven gedurende de maandelijkse periode van verzamelen (Ouwehand et al., 2009). Dit beïnvloedt de kwaliteit van de data in die bron aanzienlijk. Omdat tijdgerelateerde (kwaliteits) aspecten in R-indices worden meegenomen, is tijdigheid niet als aparte dimensie in tabel 5 vermeld. Ook de dimensie Precisie in tabel 5 wordt door tijdsafhankelijke veranderingen op de populatiesamenstelling beïnvloedt.

Tabel 5
Kwaliteitsraamwerk voor registers, hyperdimensie Data

DIMENSIES	KWALITEITSINDICATOREN	METHODE BESCHRIJVING
1. Technische checks	1.1. Leesbaarheid	– Is alle data in de bron toegankelijk?
	1.2 Voldoen aan metadata	– Voldoet alle data aan de metadata-definitie?
		– Indien niet, meld de verschillen
2. Overdekking	2.1 Niet populatie eenheden	– Percentage eenheden dat niet tot de doelpopulatie behoort
3. Onderdekking	3.1 Ontbrekende eenheden	– Percentage t.o.v. de doelpopulatie ontbrekende eenheden
	3.2 Selectiviteit	– R-index voor de samenstelling van de eenheden
	3.3 Effect op gemiddelde	– Maximale vertekening v/h gemiddelde voor kernvariabelen
		– Maximale RMSE v/h gemiddelde voor kernvariabelen
4. Koppelbaarheid	4.1 Koppelbare eenheden	– Percentage eenduidig gekoppelde eenheden
	4.2 Miskoppelingen	– Percentage niet-correct gekoppelde eenheden
	4.3 Selectiviteit	– R-index voor gekoppelde eenheden
	4.4 Effect op gemiddelde	– Maximale vertekening v/h gemiddelde voor kernvariabelen
		– Maximale RMSE v/h gemiddelde voor kernvariabelen
5. Unit-nonrespons	5.1 Eenheden zonder data	– Percentage eenheden waar alle gegevens ontbreken
	5.2 Selectiviteit	– R-index voor samenstelling v/d eenheden
	5.3 Effect op gemiddelde	– Maximale vertekening v/h gemiddelde voor kernvariabelen
		– Maximale RMSE v/h gemiddelde voor kernvariabelen
6. Item-nonrespons	6.1 Ontbrekende warden	– Percentage velden met ontbrekende waarden
	6.2 Selectiviteit	– R-index voor variabele samenstelling
	6.3 Effect op gemiddelde	– Maximale vertekening v/h gemiddelde voor variabele
		– Maximale RMSE v/h gemiddelde voor variabele
7 Meting	7.1 Externe controle	– Is een audit of paralelle toets uitgevoerd?
		– Is de inputprocedure getest?
	7.2 Onverenigbare records	– Fractie geschonden edit/controleregels
	7.3 Meetfout	– Omvang van de relatieve meetfout
8. Verwerking	8.1 Gaafmaken	– Fractie herziene/gaafgemaakte velden
	8.2 Imputatie	– Fractie geïmputeerde velden
	8.3 Uitbijters	– Fractie velden met uitbijtercorrectie
9. Precisie	9.1 Standaardfout	– Schatter voor standaardfout v/h gemiddelde
10. Gevoeligheid	10.1 Ontbrekende warden	– Totaal percentage lege velden
	10.2 Selectiviteit	– R-index voor samenstelling van totalen
	10.3 Totaal effect	– Totale maximale vertekening v/h gemiddelde
		– Totale maximale RMSE v/h gemiddelde

R-index: Representativiteits index, een indicator die de selectiviteit van de ontbrekende gegevens schat door gebruik te maken van informatie afkomstig uit andere bronnen (Schouten et al., 2009).

RMSE: root mean square error; een veel gebruikte maat in de statistiek om de kwaliteut van een schatter te bepalen. De RMSE is gelijk aan de wortel van de som van de vertekening en en de variantie van de schatter.

Een ander belangrijk aandachtspunt voor de Data hyperdimensie is het verschil tussen de input- en outputkwaliteit van de gegevens in secundaire bronnen. Dit verschil treedt uitsluitend op wanneer er sprake is van secundair gebruik van de gegevens in een bron. In een dergelijk geval worden de gegevens van een bron namelijk voor een heel ander doel gebruikt dan waar ze oorspronkelijk voor verzameld zijn. Hierdoor is de oorspronkelijke kwaliteit van de gegevens in de bron (= de kwaliteit van de output volgens de bronhouder) immers, per definitie, niet gelijk aan die van het nieuwe beoogde gebruik (= de kwaliteit van de input volgens het CBS). Wanneer het oorspronkelijke doel en het beoogde doel van gebruik gelijk zijn valt dit verschil weg. Daarnaast is er bij het secundair gebruik van administratieve brongegevens ook nog sprake van kwaliteit van de output. Dit is namelijk de

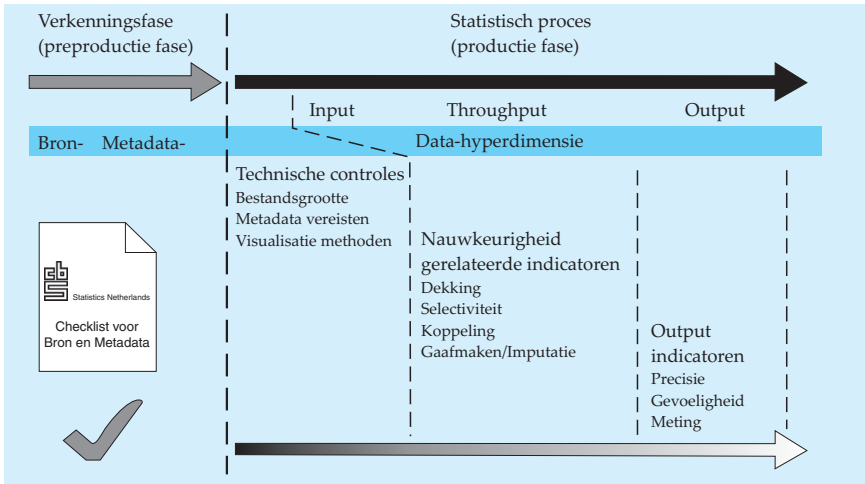
kwaliteit van de statistiek die m.b.v. de secundaire gegevens is gemaakt. Het moge hierbij duidelijk zijn dat de kwaliteit van de statistische output door de kwaliteit van de input (van de secundaire bronnen) wordt beïnvloed.

Gestructureerde studie van datakwaliteit

Bij het bepalen van de kwaliteit van de gegevens in administratieve bronnen dienen veel kwaliteitsaspecten gemeten te worden. Tabel 5 geeft hiervoor in totaal 33 meetmethoden weer. Het lijkt dan ook niet erg efficiënt voor een NSI om de waarde van elk van die meetmethoden bij elke levering van een bron te bepalen. Zeker niet wanneer bronnen 'stukje bij beetje' geleverd worden, zoals bijvoorbeeld bij de levering van BTW-gegevens door de Belastingdienst aan het CBS het geval is. Het is dan ook aan te raden om bij het ontvangen van secundaire gegevens te beginnen met het bepalen van een beperkte set van essentiële, absoluut noodzakelijke, indicatoren. Wanneer daarbij geen problemen worden gevonden dient de data specifieker, meer gedetailleerd, bestudeerd te worden.

Een overzicht van de aanpak die uit de hierboven beschreven pragmatische aanpak volgt is in figuur 1 weergegeven. Het figuur bevat, naast een voorstel voor een stapsgewijze aanpak van het meten van de kwaliteit van de data, ook de eerder besproken checklist voor Bron en Metadata. Dit is gedaan om het overzicht compleet te maken. Daarnaast is ook het onderscheidt in input- en outputkwaliteit in de figuur opgenomen. De drie stappen die bij het bepalen van de datakwaliteit onderscheiden worden zijn: i) Technische controles, ii) Nauwkeurigheid gerelateerde indicatoren, en iii) Output gerelateerde indicatoren. Elke stap wordt in de volgende paragrafen nader besproken.

Figuur 1. Overzicht van het voorgestelde proces van kwaliteitscontrole voor secundaire bronnen op het CBS

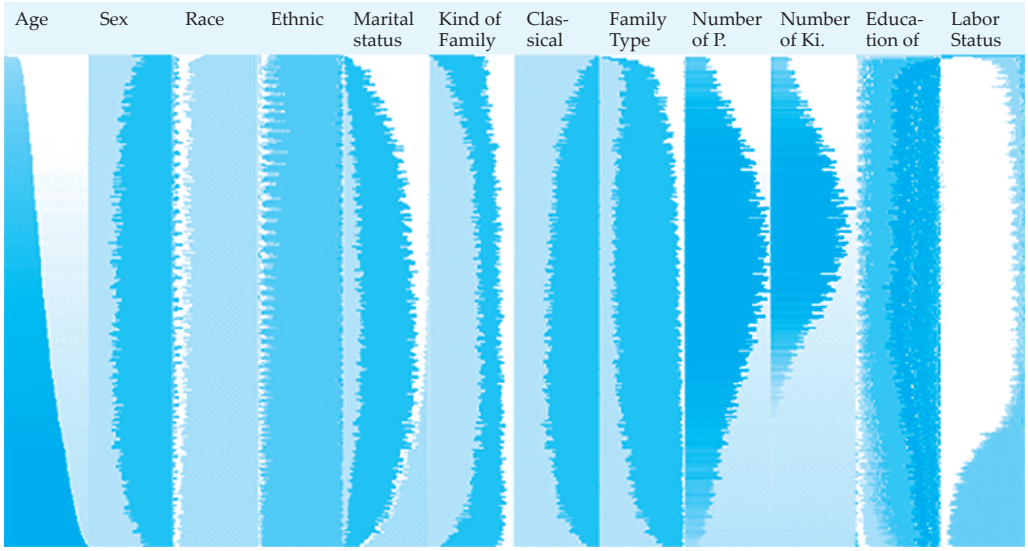


Stap 1: Technische controles

In de eerste stap, die van de technische controles, worden snelle controles op (net) ontvangen gegevens uitgevoerd. Deze controles zijn op de input gericht, erg basaal, dienen snel te kunnen worden uitgevoerd en dienen uitsluitend te worden gebruikt om serieuze problemen te identificeren. Voorbeelden van dergelijke controles zijn: het vergelijken van de omvang van een bestand en/of het aantal (unieke) eenheden in een levering ten opzichte van die van eerdere leveringen en het technisch valideren van een leveringsbestand. In dit laatste geval wordt gekeken of de gegevens en opmaak van de gegevens in een levering aan de metadata-definitie voldoen. Dit is met name bij XML-bestanden een zeer gebruikelijke controle (Van der Vlist, 2002).

Een interessante toevoeging aan de controles in deze stap van het proces zijn controlemethoden waarin gebruik wordt gemaakt van grafische visualisaties. Doordat hedendaagse computers steeds sneller worden, meer grafische mogelijkheden bieden en steeds beter met (steeds) grote(re) databestanden om kunnen gaan wordt het gebruik van dergelijke methoden steeds aantrekkelijker. Figuur 2 laat een voorbeeld zien van het gebruik van een gevisualiseerde controle van een groot databestand. De figuur is afkomstig uit Theus (2006) en toont een 'table plot' van geaggregeerde gegevens van 12 variabelen. De gegevens zijn afkomstig uit een bestand met censusgegevens van de Verenigde Staten. Elke kolom geeft één variabele weer en elk 'meetpunt' is een samenvoeging van 250 waarnemingen. De gegevens zijn gesorteerd naar leeftijd; de variabele in de eerste kolom. Bij categoriale variabelen is voor elke antwoordcategorie een andere kleur gebruikt. Een 'table plot' wordt gebruikt om een totaal overzicht te geven van alle gegevens in een dataset. Wat in figuur 2 opvalt, is dat de verhouding mannen en vrouwen (in de 2e kolom) afwijkt voor de lage en hoge leeftijdsgroepen in het bestand. Voor de lage leeftijdsgroepen is dat opvallend te noemen omdat te verwachten is dat deze verhouding dicht bij die van de geboorte zal liggen (1,05:1; CIA, 2010).

Figuur 2. Een 'table plot' van geaggregeerde gegevens van de eerste 12 variabelen in Amerikaanse censusgegevens (uit Theus, 2006)



Een ander voorbeeld waarbij het gebruik van visualisatie methoden -aan de inputkant van het proces- nuttig kan zijn is het weergeven van ontbrekende gegevens. Vaak worden deze voor de gehele bron of per variabele getoond. Met behulp van visualisatie technieken kan de relatie tussen het ontbreken van gegevens voor meerdere variabelen bekeken worden. Een voorbeeld hiervan is te vinden in het artikel van Templ en Filmozer (2008). Daarnaast is het ook denkbaar de patronen van de ontbrekende gegevens te visualiseren. Dit soort patronen kan bijvoorbeeld anders zijn voor twee bronnen waarin dezelfde hoeveelheid gegevens ontbreken maar voor een ander aantal variabelen. Bijvoorbeeld wanneer in de eerste bron 10 000 gegevens van variabelen (min of meer) willekeurig ontbreken en in de tweede bron de gegevens van 100 variabelen voor exact 100 eenheden ontbreken. Het eerste geval is waarschijnlijk geen reden om contact op te nemen met de bronhouder en het tweede geval wel!

Mede omdat visualisatiemethoden ook bij datamining gebruikt worden (Pyle, 1999) is het zeer waarschijnlijk dat deze aanpak ook op de kwaliteitscontrole van secondaire databronnen toepasbaar is. Een mogelijk beperking zou kunnen zijn dat de visualisatie techniek (d.w.z. de weergave) die wordt gebruikt specifiek voor elke databron moet worden aangepast. Het gebruik van visualisatiemethoden voor de statistiek is iets dat serieus onderzocht moet worden. Indien dergelijke methoden inderdaad toepasbaar blijken dan is het belangrijk dat standaard methodieken ontwikkeld moeten worden om het gebruik hiervan te stimuleren. Groot voordeel van visualisatiemethoden is de mogelijkheid om alle gegevens in een bron snel te kunnen controleren. Dit is iets dat niet alleen voor de statistiek maar ook voor andere wetenschapsgebieden nuttig is.

Stap 2: Nauwkeurigheid gerelateerde indicatoren

Wanneer voor een NSI duidelijk is voor welke publicatie een bron gebruikt gaat worden, kunnen meer specifieke kwaliteitscontroles worden toegepast. Indicatoren die hiervoor worden gebruikt zijn nog steeds input georiënteerd omdat ze naar de bruikbaarheid van de gegevens aan het begin van het proces kijken. Ze worden 'nauwkeurigheid gerelateerde' indicatoren genoemd omdat deze indicatoren, direct of indirect, aan de nauwkeurigheid van de gegevens gerelateerd zijn. Veel van de indicatoren die in tabel 5 staan vermeld behoren tot deze groep. Voorbeelden van 'nauwkeurigheid gerelateerde' indicatoren voor eenheden zijn: over- en onderdekking, selectiviteit en koppelbaarheid. Voorbeelden van dergelijke indicatoren voor variabelen zijn: selectiviteit, het percentage gaafgemaakte en geïmputeerde waarden en externe controle.

Veel van de indicatoren die in tabel 5 zijn weergegeven worden op het CBS al gebruikt voor het bepalen van de kwaliteit van enquêtegegevens. Dit vormt een goed startpunt voor het ontwikkelen van indicatoren die zowel voor primaire (enquête) als voor secondaire (administratieve) gegevens te gebruiken zijn. Dit uitgangspunt is een nobel streven maar op dit moment is nog onduidelijk in hoeverre dit in de praktijk mogelijk is.

Stap 3: Output gerelateerde indicatoren

De kwaliteitsindicatoren die tot de derde stap van het bepalen van de datakwaliteit behoren, duiden de kwaliteit van administratiegegevens aan op geaggregeerd niveau. De indicatoren in deze groep zijn allen duidelijk output georiënteerd. Ze rapporteren op een niveau van kwaliteit dat met de inhoud van de geproduceerde statistieken samenhangt. Met de indicatoren in de 'aan output gerelateerde' groep wordt de vraag "hoe goed is de kwaliteit van de publicatie die is gebaseerd op deze set van gegevens?" beantwoord. Voorbeelden van indicatoren die tot de groep van 'aan output gerelateerde' indicatoren behoren zijn indicatoren die proberen de precisie van kernvariabelen te bepalen en indicatoren die proberen de selectiviteit van samengestelde totalen te bepalen. Onderzoek naar precisie-indicatoren heeft recentelijk aanzienlijk voortgang geboekt wanneer het schattingen betreft die op de combinatie van register- en enquêtegegevens zijn gebaseerd (Harmsen et al., 2009). De bepaling van de precisie van schattingen die volledig op administratieve gegevens zijn gebaseerd vormt echter nog steeds een grote uitdaging (Zhang, 2009). Ook de studie naar de selectiviteit van eenheden, het tweede voorbeeld, heeft de afgelopen jaren door het 'Representative Indicators for Survey Quality (RISQ)' project (Schouten et al., 2008) aanzienlijk voortgang geboekt. Hoewel de representativiteitindicatoren oorspronkelijk voor enquêtegegevens ontwikkeld zijn blijken ze ook toepasbaar op de eenheden in administratieve bronnen (Ouwenhand et al., 2009).

Er is echter ook een belangrijke algemene beperking in de toepasbaarheid van indicatoren die tot de derde stap behoren; mogelijk is dit ook bij -een deel van- de

indicatoren in de 2e stap het geval. Het is belangrijk dat de kwaliteitsindicatoren in de 3e stap algemeen toepasbaar zijn. Zeer specifieke indicatoren kunnen niet worden opgenomen omdat het eenvoudigweg niet mogelijk is alle denkbare indicatoren op te nemen (Daas et al., 2008). Dit belangrijk punt wordt door het volgende voorbeeld verduidelijkt. Een zeer specifieke indicator is bijvoorbeeld een indicator die de resultaten voor de schatting van 'het percentage werkloze personen in Nederland' (voor een bepaalde maand) afkomstig uit een administratieve bron en uit enquêtegegevens vergelijkt. Het moge duidelijk zijn dat een dergelijke indicator nuttig is maar niet zo gedetailleerd kan en mag worden opgenomen in de 3e stap van de Data hyperdimensie omdat deze indicator niet algemeen toepasbaar is. Om die reden is dan ook besloten de algemene indicator 'externe controle' (nr. 7.1) in tabel 5 te vermelden. Daarnaast is het ook nog eens zo dat verschillende gebruikers van een bron deze voor verschillende doeleinden willen gaan gebruiken waardoor ze elk weer andere eisen stellen aan de kwaliteit van de gegevens in de bron. Dit maakt het noodzakelijk de set van indicatoren die in de 3e stap van de hyperdimensie Data kan worden opgenomen enigszins te beperken.

Een andere belangrijke vraag voor de indicatoren in de 3e stap is de vraag of ze informatie over de bruikbaarheid van de bron leveren voor- of nadat de gegevens in de bron bewerkt zijn? Indien het indicatoren betreft die alleen na het bewerken van de gegevens kwaliteitinformatie leveren dan kun je je afvragen of deze indicatoren wel zo geschikt zijn om als indicatoren voor de kwaliteit van een bron te fungeren. Procesmatig gezien is een indicator die zo vroeg mogelijk in een proces een indicatie over de kwaliteit geeft te prefereren. Wat dat betreft lijken de controles en indicatoren in de eerste 2 stappen van het bepalen van de datakwaliteit veel beter geschikt. Hierbij is zeker sprake van het meten van kwaliteit aan het begin van het proces.

Het resultaat van het meten van datakwaliteit

Het werk dat is uitgevoerd naar de kwaliteitsindicatoren in de hyperdimensie Bron en Metadata laat tevens zien dat het voor de toepasbaarheid van dat werk belangrijk is ook een gestructureerde manier van meten te ontwikkelen. Voor de Bron en Metadata hyperdimensie is daarvoor een checklist ontwikkeld (Daas et al., 2009b). Mede doordat de meetmethoden in beide hyperdimensies uit vrijwel uitsluitend kwalitatieve vragen bestaan was de keuze voor een checklist als hulpmiddel ook relatief eenvoudig.

Voor het gestructureerd meten van de kwaliteitsaspecten die tot Data hyperdimensie behoren, lijkt een checklist niet de meest handige oplossing. Zo zullen in de eerste, technische controle, stap visuele inspecties worden uitgevoerd terwijl daarnaast in de tweede en eventuele derde stap een aanzienlijk hoeveelheid kwantitatieve indicatoren moeten worden gemeten. De ontwikkeling van standaards scripts en/of een software programma lijken voor de Data hyperdimensie dan ook veel betere manieren om de gebruiker te ondersteunen.

Het is belangrijk dat de meetresultaten voor de indicatoren en controles in de hyperdimensie Data in een enkele rapportage worden samengebracht. Deze rapportage, die eerder Kwaliteitskaart of kwaliteitsinstrument is genoemd (Daas et al., 2008), dient algemeen toepasbaar te zijn en een duidelijk gestructureerd overzicht te geven van alle kwaliteitsaspecten die essentieel zijn voor de gegevens in de betreffende databron. Een andere uitdaging is de normering van de resultaten. Het vaststellen van normen voor de kwantitatieve resultaten van de indicatoren in Data zal een stuk lastiger zijn dan die voor de meer kwalitatieve kwaliteitsgegevens van de Bron en Metadata hyperdimensie (Daas et al., 2009a).

3. *Conclusie*

Dit document geeft een overzicht van de huidige stand van zaken en de ideeën voor de bepaling van de statistische bruikbaarheid van registers en administratieve bronnen op het CBS. Het eerste deel van het onderzoek naar het bepalen van de kwaliteit van dergelijke bronnen is recentelijk op het CBS afgerond. Dit werk heeft geresulteerd in de ontwikkeling van een raamwerk, dat uit drie hyperdimensies bestaat, en een checklist om de kwaliteitsindicatoren in de eerste twee hyperdimensies, Bron en Metadata, te bepalen. Indicatoren voor de bepaling van de kwaliteit van de gegevens van administratieve bronnen en registers treffen we in de Data hyperdimensie aan. Deze hyperdimensie wordt momenteel onderzocht. Dit document bevat een voorstel voor het op een gestructureerde manier bepalen van de kwaliteit van deze gegevens. Dit voorstel zal in het in april 2010 gestarte onderzoekproject BLUE Enterprise and Trade Statistics (BLUE-ETS), een project gefinancierd door het zevende kaderprogramma voor onderzoek van de Europese Unie, nader worden uitgewerkt. Hierbij zijn, naast de auteurs van dit document, ook statistici van de NSI's van Italië, Noorwegen, Slowakije en Zweden betrokken. Uiteindelijk doel van dit onderzoek is het opleveren van een instrument dat het mogelijk maakt de statistische bruikbaarheid van administratieve bronnen en registers op een efficiënt en eenduidige wijze te bepalen.

Referenties

Bakker, B.F.M., Linder, F., and Van Roon, D. (2008). *Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys.* Proceedings of IAOS Conference on Reshaping Official Statistics, Shanghai, China.

Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques.* Berlin: Springer.

Chowdry, H., Crawford, C., Dearden, ,L., Goodman, A., and Vignoles, A. (2010) *Widening Participation in Higher Education: Analysis Using Linked Administrative Data.* DoQSS Working Papers 1008, Department of Quantitative Social Science – Institute of Education, University of London.

CIA (2010) *The World Factbook, Unites States, People, Sex-ratio,* Online at https://www.cia.gov/library/publications/the-world-factbook/geos/us.html.

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

Daas, P.J.H., Ossen, S.J.L., and Arends-Tóth, J. (2009a). *Framework of Quality Assurance for Administrative Data Sources.* Paper for the 57th session of the International Statistical Institute, Durban, South Africa.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., and Arends-Toth, J. (2009b). *Checklist for the Quality evaluation of Administrative Data Sources.* Discussion paper 09042, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Harmsen, C., Van Der Laan, J., and Kuijvenhoven, L. (2009). *Deriving longitudinal consistent household statistics from register information.* Paper for the 57th session of the International Statistical Institute, Durban, South Africa.

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

Ouwehand, P., Schouten, B., and De Heij, V. (2009). *Representativity indicators for business surveys based on population totals.* Paper for the European Establishment Statistics Workshop, Stockholm, Sweden.

Pyle, D. (1999). *Data preparation for data mining.* San Francisco: Morgan Kaufmann.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology,* 35, 101–113.

Schulte Nordholt, E. (2004). Introduction to the Dutch virtual census of 2001. In *The Dutch Virtual Census of 2001, analysis and methodology,* eds. E. Schulte Nordholt, M. Hartgers, R. Gircour, Voorburg: Statistics Netherlands.

Templ, M. and Filzmoser, P. (2008). *Visualization of missing values using the R-package VIM.* Forschungsbericht CS-2008–1, Institut f. Statistik u. Wahrscheinlichkeitstheorie, Wien, Austria.

Theus, M. (2006). Statistical Graphics. In *Graphics of Large Datasets: Visualizing a Million,* eds. A. Unwin, M. Theus, H. Hofmann, Singapore: Spinger, pp. 31–54.

Unece (2007). *Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics.* Geneva: United Nations Publication.

Van der Vlist, E. (2002). *XML Schema.* Sebastopal: O'Reilly & Associates.

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

Zhang, L-C. (2009). *Unit errors in statistical registers and their effects.* Paper for the 57th session of the International Statistical Institute, Durban, South Africa.