Bijlage 1 - Algemene beschrijving dataprocessing

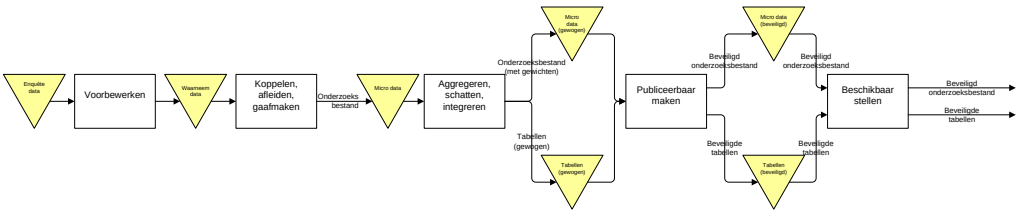
Energiemodule ENO2018

De term dataprocessing wordt gebruikt om de werkzaamheden samen te vatten die nodig zijn om aan de antwoorden die respondenten geven op de vragen in de vragenlijst, plausibele statistische informatie te kunnen ontlenen. De verwerking van de data is zoveel mogelijk geautomatiseerd. Het verwerken is uitgevoerd door de Sector Arbeid, Inkomen en Leefsituatie (SAL) van de Divisie Sociaal-economische en ruimtelijke statistieken (SER) van het CBS.

1. Inleiding

Het onderstaande figuur 1 geeft het verwerkingsproces van de onderzoeksdata in hoofdlijnen aan. De vijf deelprocessen (weergegeven in de vierkanten blokken) worden in hierna in detail beschreven.

Figuur 1: Procesmodel Verwerkingsproces



Voorbewerken: Van enquêtedata naar waarneemdata

Voordat enquêtedata geschikt is om te verwerken zijn er een aantal activiteiten nodig als voorbewerking. Kort gezegd gaat het erom de juiste data (qua cases en variabelen) in het juiste formaat beschikbaar te maken. Dit resulteert in waarneemdata. De waarneemdata is de grondstof voor het ‘echte’ verwerken.

Binnen het subproces “Voorbewerken” kunnen de volgende processtappen worden onderkend:

1.1. Controleren enquêtedata: range en routing controle

Met de range- en routingcontrole wordt gekeken of het enquêtebestand voldoet aan de eisen voor de verdere verwerking van het bestand.
Het enquêtebestand wordt hier gedefinieerd als:

- alle data uit een vragenlijst;
- van een of meerdere respondenten;
- van een specifieke mode.

De range- en routingcontrole is specifiek:

- per onderzoek;
- per mode;
- per versie van de vragenlijst.

De eisen van de controles zijn overwegend gebaseerd op de definitie van de vragenlijsten zoals vastgelegd in de gecompileerde vragenlijsten en op het ontwerp van het uniforme datamodel.
Range controles zijn o.a.:

- formaat van variabele;
- variabele heeft juiste naam;
- waardebereik van variabele conform afspraak;
- dubbele sleutels.

Routing controles zijn o.a.:

- indien een blok niet op de route ligt, mogen de variabelen ook geen waarde hebben;
- indien een blok wel op de route ligt moeten deze in principe (tenzij veld niet verplicht) ook een waarde hebben.

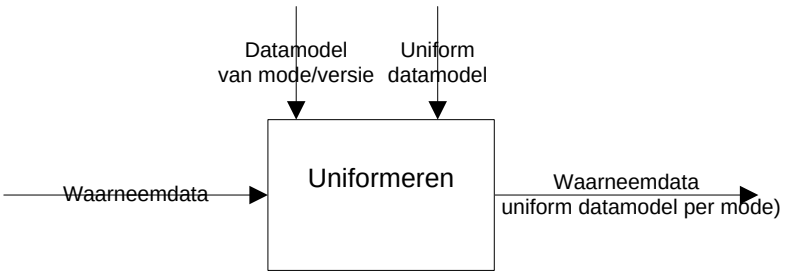
1.2 Uniformeren

De recordstructuur van enquêtedata kan mode specifiek zijn. Dit komt omdat de vragenlijst mode specifiek is. Daarmee is ook de metadata mode specifiek.
Het kan ook voorkomen dat gedurende het onderzoek er voor een mode meerdere versies van een vragenlijst worden gebruikt (eventueel ook met een verschil in variabelen). Dus ook per versie kunnen er verschillen qua recordstructuur zijn.
Doel bij uniformeren is om tot één uniforme recordstructuur te komen. De uniforme recordstructuur is onderzoekspecifiek en geldt in principe voor de duur van het hele onderzoek. Het bepalen van de uniforme recordstructuur[1] en het beschrijven van de bijbehorende metadata is een ontwerpactiviteit. Dit betekent dat de metadata van het uniforme datamodel vooraf gedefinieerd is; dus voordat het uniformeren daadwerkelijk plaats vindt. Bij het uniformeren van de data hoeft er in principe dus geen metadata meer te worden aangepast; als je de data in het uniforme model zet is de metadata automatisch correct.
Pre-conditie:

- het mode-specifieke (en vragenlijstversie specifieke) datamodel moet bekend zijn;
- het uniforme datamodel moet bekend zijn.

Figuur 2: Procesmodel Uniformeren

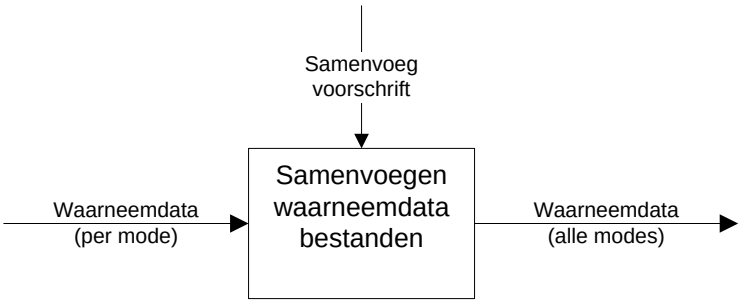
[1] en trouwens ook de mode-specifieke datamodellen



1.3 Controleren enquêtedata: range en routing controle

De waarneemdatabestanden van de diverse modes worden samengevoegd tot één fysiek bestand. Per case moet wel duidelijk blijven uit welke mode het record komt.

Figuur 3: Procesmodel Samenvoegen waarneemdata



2. Koppelen, afleiden en gaafmaken

Bij het koppelen, afleiden en gaafmaken wordt de data verrijkt met o.a. data uit de steekproef, registerdata en andere bronnen. Tevens wordt de respons afgebakend, vindt imputeren en gaafmaken plaats en worden variabelen afgeleid (inclusief coderen). Dit resulteert in een onderzoeksbestand.

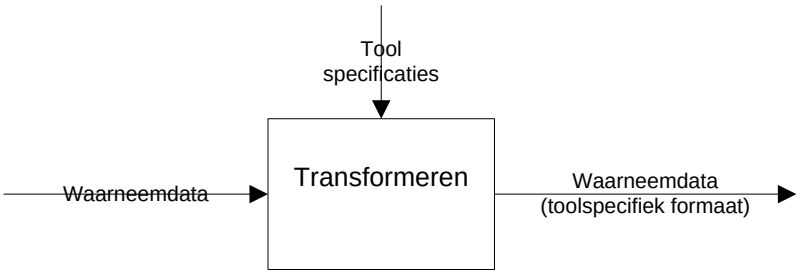
2.1 Transformeren

Doel van deze procesactiviteit is de waarneemdata transformeren naar een vorm die gebruikt kan worden bij de vervolg activiteiten. De waarneemdata is in een bepaald technisch formaat, in dit geval Blaise. Bij de vervolgstappen van het verwerkingsproces wordt SPSS als verwerkingstool gebruikt. Daartoe moet het technische formaat van de data worden aangepast (van Blaise via ASCII naar SPSS).

T.b.v. SPSS is transformeren bijvoorbeeld:

- Dichitomiseren (men werkt in het onderzoeksbestand niet met meervoudige antwoorden).
- Labels aanbrengen: Op basis van de metadata uit de vragenlijst worden variabelen en value labels gegenereerd, die worden gecombineerd met dit SPSS-systeembestand. Deze labels vormen de beschrijving van de data.
- Imputatie routing: Respondenten hoeven in de vragenlijst alleen die vragen te beantwoorden die op hun situatie van toepassing zijn. Vragen die door de respondent niet beantwoord hoefden te worden, gaan als blanco naar ASCII en krijgen vervolgens in SPSS de waarde SYSMIS. Dit betekent dat het SPSS-systeem de waarden als “n.v.t.” beschouwd bij de uitvoering van statistische analyses.

Figuur 4: Procesmodel Transformeren



2.2 Verrinnen

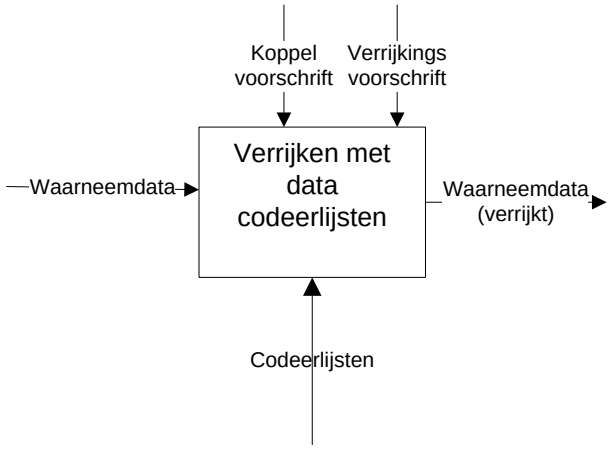
Binnen de verwerkingsprocessen gebeurt “het verrinnen” alleen t.b.v. het koppelen met registerdata en dus niet om met geanonimiseerde data in het verwerkingsproces te werken. Om waarneemdata te kunnen verrijken met registerdata dient de waarneemdata eerst verrind te worden. Dit betekent dat voor iedere persoon in de waarneemdata een betekenisloze identificerende variabele wordt bepaald (genaamd “RINPersoon”). Dit nummer is gebaseerd op data uit de GBA.

De te koppelen waarneemdata wordt daartoe geleverd aan CBK. Deze koppelen de data aan het Centrale Koppelbestand Personen (CKP). Sector CBK levert de verrinde data vervolgens terug aan het verwerkingsproces. Voor het leggen van een koppeling zijn het Burger Service Nummer (BSN) en/of de combinatie van geboortedatum, geslacht en adres nodig. Een geslaagde koppeling betekent in concreto dat aan het originele record RINPersoon en RINPersoonVolgNr uit het CKP worden toegevoegd. Hiermee is de desbetreffende persoon in het CKP identificeerbaar. Naast de CKP persoonidentifiCATIe wordt nog informatie over de koppeling aan het record toegevoegd.

2.3 Verrijken met data codeerlijsten

De waarneemdata wordt hier gekoppeld aan verschillende codeerlijsten. Hier wordt bijvoorbeeld op basis van de viercijferige postcode de gemeentecode bepaald. Op basis van de gemeentecode worden de bovengemeentelijke regionale indelingen bepaald.

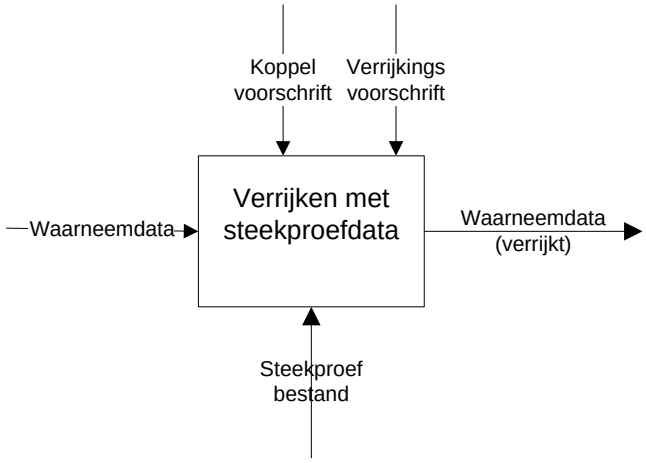
Figuur 5 : Procesmodel Verrijken met data codeerlijsten



2.4 Verrijken met steekproefdata

T.b.v. non-respons analyses kan de waarneemdata worden verrijkt met de complete steekproef met voor elk element de voor uitdunning van de adressensteekproef gebruikte variabelen (alle adresgegevens bijvoorbeeld), het startgewicht en een eindresultaat (bijvoorbeeld: uitgedunde GBA65plus, niet uitgezet door regiomanager, geen woonadres, leegstand, niemand thuis, taalbarrière, weigering, enzovoorts). Op deze manier kan een betere (uitgebreidere) non-respons analyse naar allerlei achtergrondkenmerken gemaakt worden. Bij het koppelen kan het voorkomen dat er een steekproefeenheid is waarvoor (nog) geen waarneemdata is. En tevens waarneemdata waarvoor geen steekproefeenheid is. In het laatste geval is blijkbaar een verkeerde respondent bevraagd.

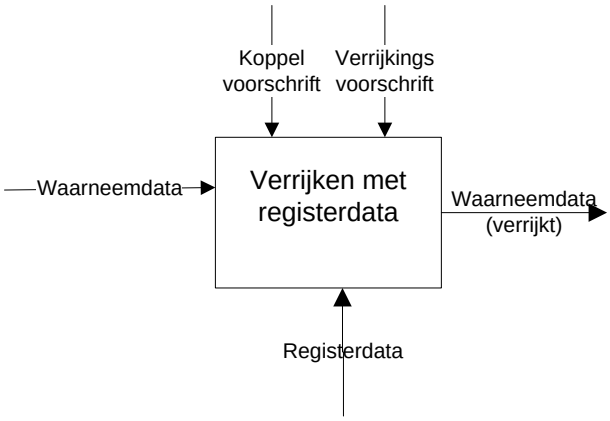
Figuur 6: Procesmodel Verrijken met steekproefdata



2.5 Verrijken met registerdata

Het koppelen met de registerdata gebeurt op basis van het RIN-nummer. Veel gebruikte registers zijn de GBA, de Polisadministratie en het UWV-Werkbedrijf.
De waarneemdata wordt uit de registers verrijkt met o.a. type huishouden, geboorteland persoon en van diens vader en moeder en afleidingen daarop (GBA), bron inkomen en hoogte inkomen (Polisadministratie), provincie, inschrijfduur (is een afleiding), werkend (UWV-Werkbedrijf).
Het verrijken gebeurt niet alleen voor de OP maar ook voor alle vastgestelde personen in het huishouden. Als er een koppeling is, wordt de waarneemdata vervolgens verrijkt met een aantal variabelen uit de registers. Als er geen koppeling is blijven de betreffende register-variabelen leeg.

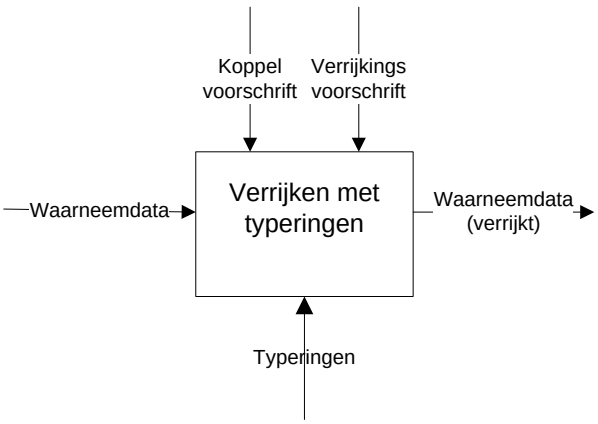
Figuur 7: Procesmodel Verrijken met registerdata



2.6 Verrijken met typeringen

Deze stap kan pas worden uitgevoerd als de typeringen beschikbaar zijn. Afhankelijk van het specifieke onderzoek heeft het typeren een bepaalde doorlooptijd, waardoor deze data (meestal) niet meteen beschikbaar is.

Figuur 8: Procesmodel Verrijken met typeringen

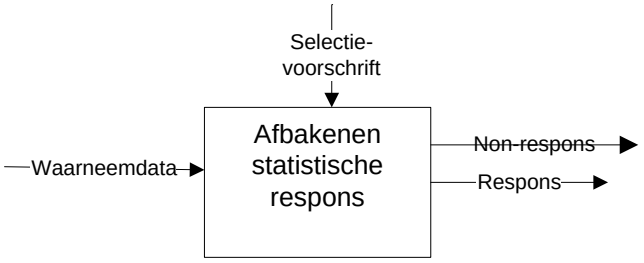


2.7 Afbakenen statistische respons

Alleen statistische respons wordt meegenomen in de verdere verwerking. Wat wel/niet tot respons behoort, staat in een voorschrift. In deze activiteit wordt o.b.v. het voorschrift de respons bepaald.

Hóe de non-respons gescheiden wordt van de respons is vanuit logisch oogpunt niet relevant; dit kan bijvoorbeeld door de records fysiek van elkaar te scheiden, maar kan bv ook door met indicatoren te werken. Voorwaarde is dat de non-respons op een gegeven moment beschikbaar is ten bate van non-respons analyses.

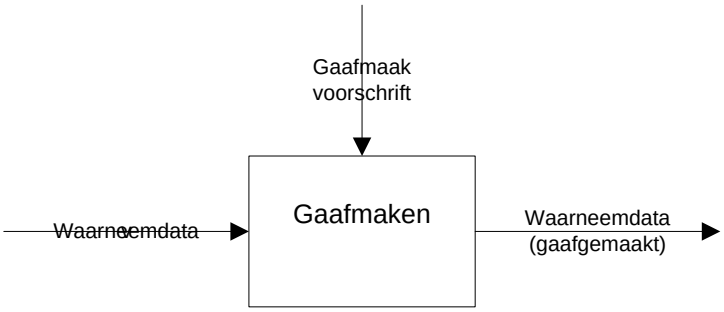
Figuur 9: Procesmodel Afbakenen statistische respons



2.8 Gaafmaken (micro)

Gaafmaken is het opsporen en corrigeren van foutieve gegevens in de waarneemdata.
Bij micro gaafmaken vinden zowel de controles als de correcties plaats op microniveau.
Voorbeelden van voorkomende fouten zijn: het geboortejaar klopt niet of is onwaarschijnlijk, een respondent rapporteert in euro's in plaats van in duizenden euro's (of omgekeerd), of de winst van een bedrijf is niet gelijk aan het verschil tussen baten en lasten. Gaafmaken gebeurt op basis van, bij het ontwerp bepaalde, voorschriften.

Figuur 10: Procesmodel Gaafmaken (micro)



2.9 Imputeren

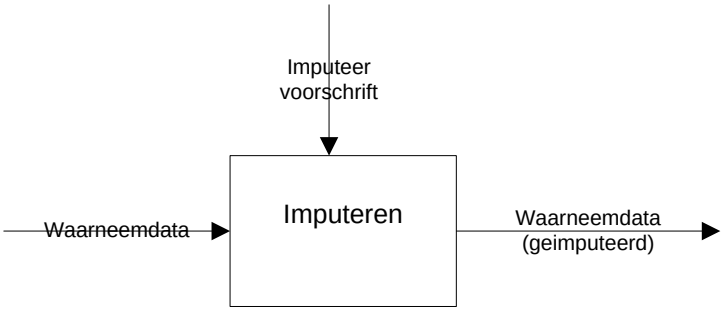
Imputeren is het bepalen en introduceren van een (nieuwe) waarde op een plaats waar een waarde ontbreekt of op 'onbekend' (ontbrekend) is gezet.
Bij enquêtes komt het voor dat respondenten op één of meer vragen geen antwoord geven, terwijl dit wel van ze wordt verlangd. Men spreekt dan van item-nonrespons (of partiële nonrespons) en van (ten onrechte) ontbrekende waarden (missing values). Redenen om een vraag niet te beantwoorden zijn het niet kunnen of willen geven van het antwoord. Op ingewikkelde of moeilijk te begrijpen vragen kan men vaak geen antwoord geven, op gevoelige vragen wil men het dikwijls niet. Ook bij registers kunnen gegevens ontbreken die het CBS wel had willen hebben.

Er zijn een aantal manieren om met ontbrekende waarden om te gaan. Eén daarvan is imputeren van een geldige waarde voor de ontbrekende waarde in het databestand.
Een alternatief voor imputeren is om het achterwege te laten. De ontbrekende waarden blijven dan onbekend. Redenen om te imputeren, in plaats van het veld leeg te laten, zijn:

1. het verkrijgen van een 'volledig' (geheel gevuld) databestand;
2. verhoging van de kwaliteit van het micro-bestand en/of van de parameterschattingen.

We maken verder onderscheid tussen imputeren en afleiden. Bij het afleiden van variabelen worden nieuwe variabelen gecreëerd als functie van in het bestand reeds bestaande variabelen. Bij imputeren worden ontbrekende waarden op een bestaande variabele gecreëerd. Imputeren gebeurt op basis van, bij het ontwerp bepaalde, voorschriften. Steeds dient te worden vastgelegd dat een waarde is geïmputeerd.

Figuur 11: Procesmodel Imputeren



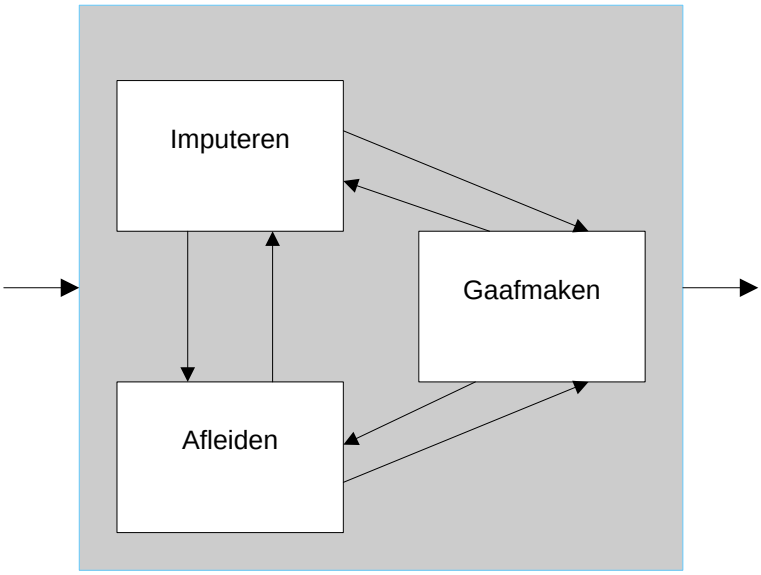
2.10 Afleiden

Met afleiden wordt hier bedoeld het creëren van nieuwe variabelen als functie van in het bestand reeds bestaande variabelen.
Coderen is ook een vorm van afleiden. Het coderen van een vraag is het (keuze)proces waarbij een beslissing wordt genomen om een antwoord te interpreteren in termen van een voor gedefinieerde verzameling mogelijke antwoorden. Een dergelijke keuze wordt, tijdens een interview of bij het invullen van een vragenformulier, vaak gedaan door respondenten, al of niet met hulp van een interviewer. Soms echter wordt deze keuze achteraf gedaan, op het CBS en zonder de aanwezigheid van een respondent, door codeurs of typeurs.

2.11 Gaafmaken, imputeren en afleiden: samenhang

De logische volgorde is eerst gaafmaken, dan imputeren en dan variabelen afleiden.
Echter, gaafmaken, imputeren en afleiden zijn geen activiteiten die voor een dataset en individuele case sequentieel verlopen. De activiteiten worden per variabele of set van variabele doorlopen, waarna de activiteiten voor andere variabelen worden doorlopen.
Eventueel kan de volgorde van de activiteiten ook anders zijn: bv eerst imputeren dan gaafmaken. Dit is afhankelijk van de specifieke regels die van toepassing zijn binnen het onderzoek.

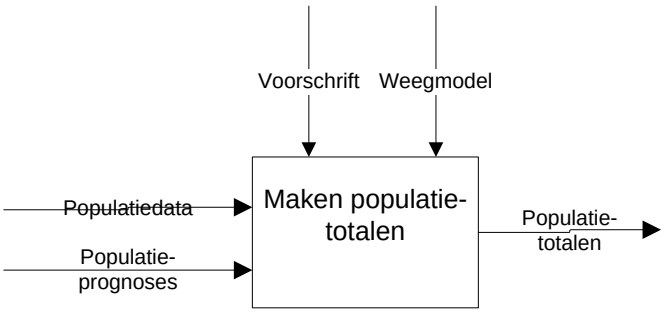
Figuur 12: Procesmodel Gaafmaken, imputeren en afleiden: Samenhang



2.12 Maken Populatietotalen

Voor het wegen zijn populatietotalen nodig. Het kan zijn dat de populatietotalen zelf geschat moeten worden aangezien de data bij de taakgroep Demografie niet altijd voldoende gedetailleerd en soms onvoldoende actueel zijn. Populatietotalen worden bepaald op het totale register, niet op de met register verrijkte waarneemdata.

Figuur 13: Procesmodel maken populatietotalen

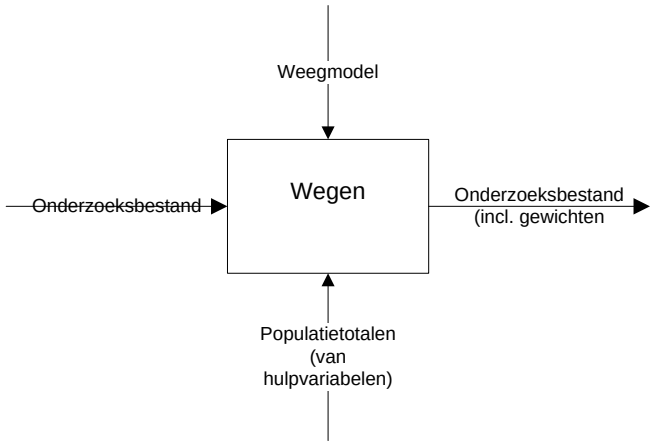


2.13 Wegen

Het wegen is de activiteit waarbij weegfactoren worden bepaald. In het weegmodel staat beschreven hoe het wegen moet plaatsvinden.
Bij het wegen wordt de verdeling van variabelen in de steekproef in overeenstemming gebracht met de verdeling daarvan in de populatie. Daartoe wordt aan iedere case een gewicht toegekend. Ten bate van het wegen zijn populatietotalen (hulpvariabelen) nodig op persoonsniveau. Tevens kan

gewogen worden naar verdelingen waaraan het responsproces idealiter moet voldoen, bijvoorbeeld: elke dag evenveel respons of een gelijk responspercentage per mode. De weging resulteert in één of meerdere ophoogfactoren (afhankelijk van het aantal entiteiten waarvoor gewogen wordt).

Figuur 14: Procesmodel Wegen



3. Publiceerbaar maken

Bij het publiceerbaar maken worden de tabellen en het gewogen onderzoeksbestand statistisch beveiligd. Binnen het subproces “publiceerbaar maken” kunnen de hieronder genoemde processtappen plaatsvinden.

3.1 Maken micro output

Op basis van het onderzoeksbestand wordt de output (microbestanden) voor bijvoorbeeld de externe klant, SAL-SAD/SIL, CvB, DANS en/of Eurostat gemaakt.
De output voor de diverse afnemers bevatten meestal slechts een deelverzameling van de variabelen uit het onderzoeksbestand. Deze deelverzamelingen voor de verschillende afnemers worden in deze stap gemaakt.
SAL-SAD/SIL krijgen meestal het gehele onderzoeksbestand ten behoeve van het maken van publiCATIes. Hiervoor hoeven geen extra stappen gezet te worden.

3.2 Statistisch beveiligen microdata

De microbestanden voor de verschillende afnemers worden vaak nog statistisch beveiligd. De wijze van beveiliging kan wel verschillen.
Onder statistische beveiliging verstaan we hier het voorkómen dat er inhoudelijke conclusies over herkenbare eenheden kunnen worden getrokken op basis van gepubliceerd of anderszins beschikbaar gesteld CBS-materiaal. Uit de statistische publiCATIes van het CBS (StatLine-tabellen, web-artikelen, persberichten, wetenschappelijke artikelen) mogen zulke conclusies niet getrokken kunnen worden. Maar ook als het CBS microdata beschikbaar stelt voor wetenschappelijke analyse, moet deze grondregel van de statistiek overeind blijven.

3.3 Statistisch beveiligen standaardtabellen

De tabellen (t.b.v. de Externe klant, Ministeries,Statline, CvB, DANS en/of Eurostat) dienen ook statistisch beveiligd te worden.

4. Beschikbaar stellen

De statistisch beveiligde producten worden vervolgens als output geleverd (beschikbaar gesteld) aan diverse belanghebbenden, waaronder externe opdrachtgevers, verschillende ministeries, Eurostat, DANS, Centrum voor Beleidsstatistiek. Binnen het subproces “Beschikbaar stellen” kunnen de hieronder genoemde processtappen plaatsvinden.

4.1 Leveren microbestanden

Dit is de activiteit waarbij microdata daadwerkelijk wordt geleverd. Aan welke partijen geleverd wordt is afhankelijk van het betreffende onderzoek.
Er worden voor het onderzoek CV twee bestanden opgeleverd: een responsbestand met data voor de ene vragenlijstvariant en een responsbestand met data voor de andere vragenlijstvariant.

4.2 Leveren standaardtabellen

Dit is de activiteit waarbij de tabellen daadwerkelijk worden geleverd. Aan welke partijen geleverd wordt is afhankelijk van het betreffende onderzoek.