Het benaderen van de onbetrouwbaarheidsmarges van OVG-cijfers

Drs. F.D. Bijleveld

R-99-21

Het benaderen van de onbetrouwbaarheidsmarges van OVG-cijfers

Toepassing van de approximatie-methode van het CBS

R-99-21
Drs. F.D. Bijleveld
Leidschendam, 1999
Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

Documentbeschrijving

Rapportnummer:	R-99-21
Titel:	Het benaderen van de onbetrouwbaarheidsmarges van OVG-cijfers
Ondertitel:	Toepassing van de approximatie-methode van het CBS
Auteur(s):	Drs. F.D. Bijleveld
Onderzoeksmanager:	Mr. P. Wesemann
Projectnummer SWOV:	53.127
Projectcode opdrachtgever:	PRDVL 98.091
Opdrachtgever:	De inhoud van dit rapport berust op gegevens verkregen in het kader van een project, dat is uitgevoerd in opdracht van de Adviesdienst Verkeer en Vervoer van Rijkswaterstaat.
Trefwoord(en):	Mobility (pers), traffic survey, mathematical model, statistics, evaluation (assessment).
Projectinhoud:	Het onderzoek verplaatsingsgedrag (OVG) wordt jaarlijks door het Centraal Bureau voor de Statistiek (CBS) uitgevoerd. Een deel van de Nederlandse bevolking wordt door middel van een enquête ondervraagd over hun mobiliteitsgedrag: hoeveel kilometer zij met welk vervoermiddel hebben afgelegd in een bepaalde periode. De resultaten van deze enquête worden daarna opgehoogd naar gegevens van 'de totale Nederlandse bevolking'. In dit rapport wordt op basis van een door het CBS gepubliceerde methode een implementatie uitgewerkt van een methode om de onzekerheid in cijfers te schatten die afgeleid zijn van het OVG.
Aantal pagina's:	17 + 6 blz.
Prijs:	f 17,50
Uitgave:	SWOV, Leidschendam, 1999

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV
Postbus 1090
2260 BB Leidschendam
Telefoon 070-3209323
Telefax 070-3201261

Samenvatting

Het onderzoek verplaatsingsgedrag (OVG) wordt jaarlijks door het Centraal Bureau voor de Statistiek (CBS) uitgevoerd. Een deel van de Nederlandse bevolking wordt door middel van een enquête ondervraagd over hun mobiliteitsgedrag: hoeveel kilometer zij met welk vervoermiddel hebben afgelegd in een bepaalde periode. De resultaten van deze enquête worden daarna opgehoogd naar gegevens van 'de totale Nederlandse bevolking'.
In dit rapport wordt op basis van een door het CBS gepubliceerde methode een implementatie uitgewerkt van een methode om de onzekerheid in cijfers te schatten die afgeleid zijn van het OVG. Het gaat hierbij uitsluitend om het schatten van de onzekerheid ten gevolge van het feit dat het OVG een steekproef is uit een populatie, en niet de gehele populatie zelf betreft. De methode geeft bijvoorbeeld geen schatting van de onzekerheid als gevolg van het feit dat de gereden afstanden door de geënquêteerden worden geschat en niet worden gemeten. De geschatte onzekerheid kan derhalve het beste gezien worden als een ondergrens.

De praktische uitwerking heeft geresulteerd in een betrekkelijk flexibel SAS-programma, waarmee zonder veel moeite verschillende meer-dimensionale tabellen kunnen worden vastgesteld. Aan de hand van een praktisch voorbeeld wordt het effect op de onzekerheid van verdere disaggregatie van OVG-gegevens aangegeven.

Summary

Estimating the confidence margins of the National Travel Survey (NTS)

The National Travel Survey (NTS), which is continuous, is carried out by Statistics Netherlands, the Dutch Central Bureau of Statistics (CBS). A sample of the population of the Netherlands is surveyed about their travel behaviour: how many kilometres, which modes of transport, during which period. The sample results are then extrapolated to represent 'the whole Dutch population'.
This report uses a method published by the CBS to estimate the uncertainties of the NTS data. Only those uncertainties resulting from the fact that NTS uses a sample, and not the whole population, are dealt with. For example, the method makes no estimation of the uncertainty which is a result of respondents estimating the distances, and their not actually being measured. The estimated uncertainty can, therefore, be best regarded as a minimum.

The practical working out resulted in a relatively flexible SAS program. Various multi-dimensional tables were established quite easily. A practical example is given to show the effect of the uncertainty of further disaggregation of NTS data.

Inhoud

1. Inleiding

In veel SWOV-onderzoeken wordt ten behoeve van het verkrijgen van mobiliteitsgegevens al dan niet expliciet gebruik gemaakt van gegevens uit het onderzoek verplaatsingsgedrag (OVG) van het Centraal Bureau voor de Statistiek (CBS, 1993b). Het OVG wordt jaarlijks door het CBS uitgevoerd. Een deel van de Nederlandse bevolking wordt door middel van een enquête ondervraagd over hun mobiliteitsgedrag. De resultaten van deze enquête worden daarna opgehoogd naar gegevens van 'de totale Nederlandse bevolking'.
Aan een onderzoek, dat op een dergelijke wijze wordt uitgevoerd, kleven een aantal bezwaren. De resultaten zullen namelijk afwijken van de landelijke cijfers, omdat een toevallige selectie van geënquêteerden is gebruikt, waarvan de gegevens worden opgehoogd tot landelijke cijfers.
Dit verslag beschrijft de approximatie-methode die door het CBS is toegepast om een schatting te maken van de mate waarin de OVG-cijfers kunnen variëren, ten gevolge van het feit dat een toevallige groep geënquêteerden in een steekproef is gebruikt. Nadrukkelijk dient vermeld te worden, dat afwijkingen tussen de cijfers uit het OVG, gebaseerd op de steekproef, en gegevens van de totale Nederlandse bevolking, die het gevolg zijn van andere fouten, niet meegenomen (kunnen) worden. Met andere woorden, de nu volgende methode benadert slechts de variatie in de OVG-cijfers die zou kunnen ontstaan als in plaats van de gebruikte selectie van mensen in de steekproef, een andere groep mensen gekozen zou zijn volgens dezelfde selectiecriteria. Het feit dat juist deze mensen (toevallig) deel uitmaken van de steekproef kan een grote invloed hebben op de uitkomst, zeker omdat de cijfers sterk opgehoogd moeten worden om gegevens over de Nederlandse bevolking te verkrijgen. Deze gevolgen voor de uitkomst zijn het grootst als de steekproef bestaat uit een klein aantal mensen.
Het schatten van deze gevolgen is een belangrijke toepassing van de hierna beschreven methode. Andere fouten, zoals bij voorbeeld het (door respondenten) systematisch foutief schatten van door hen afgelegde afstanden kunnen op deze wijze natuurlijk niet worden vastgesteld.

Behalve de approximatie-methode maakt het CBS ook gebruik van andere methoden om informatie over de onzekerheid van OVG-cijfers te verkrijgen. Eén hiervan betreft de bootstrap-methode. Deze methode is (theoretisch) beter toepasbaar bij zeer kleine steekproefaantallen. Voor zowel de bootstrap-methode als de hier beschreven approximatie-methode geldt overigens dat bij extreem kleine steekproefaantallen de methode zelf ook niet meer betrouwbaar is. In die zin leveren beide methoden slechts een eerste benadering van de betrouwbaarheid van de steekproefschattingen waarmee bij nog kleinere aantallen op een verantwoorde manier gewerkt kan worden. De approximatie-methode is overigens gebaseerd op een bootstrap-methode.

Een belangrijke toepassing van deze methoden betreft het vergelijken van een reeks mobiliteitscijfers over de jaren heen. Dit is zeker van belang omdat de steekproefomvang in recente jaren aanmerkelijk groter is dan in voorgaande jaren. Met de grotere onbetrouwbaarheid van de cijfers uit de eerdere jaren - waarin met een kleinere steekproef werd gewerkt - zal ook rekening gehouden moeten worden bij de vergelijking van de gegevens over de jaren.

In deze rapportage wordt een beschrijving gegeven van een SAS-programma waarmee de betrouwbaarheidsgrenzen van OVG-cijfers kunnen worden berekend. Het statistisch pakket SAS is het standaard analyse-pakket dat bij de SWOV wordt gebruikt.

2. Werkwijze benaderingsmethode

De gebruikte werkwijze wordt in detail beschreven in 'Een vergelijking tussen de BHS-methode en analytische benaderingsformules voor het schatten van relatieve marges van cijfers uit het OVG' (CBS, 1993a). De BHS-methode staat voor Bootstrap Half Sample-methode. Een beknopte uiteenzetting van de methode die in dit onderzoek is gebruikt, is in de bijlage van de CBS-OVG publicatie 1992, (CBS, 1993b) weergegeven.

Iedere methode om de onbetrouwbaarheid van gegevens te schatten moet rekening houden met de wijze waarop de steekproef is opgezet. Dit geldt ook voor deze methode. Bij de enquête-methode die bij het OVG wordt gebruikt wordt voor iedere dag dat er geënquêteerd wordt een selectie van adressen gekozen. Vervolgens wordt ieder huishouden op een gekozen adres geënquêteerd - benaderd voor deelname aan de enquête. Deze procedure wordt maandelijks herhaald. De adressen worden zonder teruglegging gekozen. Dit levert een complex steekproefontwerp op, waarvoor de onbetrouwbaarheidsmaten in eerste instantie met behulp van de BHS-methode zijn benaderd.

Voor de approximatie-methode is het steekproefontwerp als volgt vereenvoudigd:

- er wordt gebruik gemaakt van een steekproef van huishoudens per dag in plaats van een steekproef van adressen per maand;
- de huishoudens binnen de steekproef worden onderling onafhankelijk beschouwd. Doordat het aantal huishoudens in de steekproef zeer klein is in vergelijking tot het totaal aantal huishoudens, zijn de gevolgen van het feit dat met teruglegging is getrokken in plaats van zonder teruglegging klein;
- de kans dat een huishouden tweemaal in de steekproef voorkomt is verwaarloosd . In ieder geval zijn de eventuele gevolgen hiervan verwaarloosd (idem);
- de factoren voor het ophogen naar de Nederlandse bevolking zijn bekend beschouwd in plaats van dat ze van het toeval afhankelijk zijn.

De onbetrouwbaarheidsmarges van dit ontwerp blijken redelijk eenvoudig te berekenen.

Daar men voor verkeersveiligheidstoepassingen meestal geïnteresseerd is in totale aantallen reizigerskilometers, is de aandacht in dit verslag vooral daarop gericht. De CBS-rapportages zijn, mede in verband met het bredere toepassingsgebied, op dit gebied uitvoeriger.

3. Berekening van de marges van een totaal aantal reizigerskilometers uit OVG

CBS (1993b) levert de volgende twee formules (nummering CBS, 1993b):

$$(1.1) \qquad \hat{y} = \sum_{c=1}^{C} \sum_{h=1}^{n_c} y_{ch}$$

Hiermee wordt het totaal aantal reizigerskilometers $\hat{y}$ berekend. De variantie hiervan staat in (1.6):

$$(1.6) \qquad \mathrm{var}(\hat{y}) = \sum_{c=1}^{C} \frac{n_c}{n_c - 1} \sum_{h=1}^{n_c} (y_{ch} - \bar{y}_c)^2$$

Hierbij is n_c het aantal huishoudens op dag c, C is het totaal aantal dagen. y_{ch} is de totale (opgehoogde) afstand die door mensen in een huishouden tezamen is afgelegd. Een bepaald type mobiliteit, bijvoorbeeld de verplaatsingen van oudere fietsers, wordt eerst per huishouden opgehoogd gesommeerd. Technisch betekent dit dat iedere verplaatsing die aan de criteria voldoet (hier fietsverplaatsingen door ouderen) eerst wordt opgehoogd naar een landelijk cijfer en vervolgens worden deze cijfers per huishouden opgeteld. y_{ch} staat dus voor de gewogen som van de lengten van alle fietsritten uitgevoerd door ouderen, maal hun ophoogfactor naar de landelijke cijfers van huishouden h op dag c. $\bar{y}_c$, het gemiddelde van deze cijfers voor dag c, en is dus gelijk aan:

$$\bar{y}_c = \frac{1}{n_c} \sum_{h=1}^{n_c} y_{ch}$$

met (steekproef)variantie:

$$s^2(c) = \frac{1}{n_c - 1} \sum_{h=1}^{n_c} (y_{ch} - \bar{y}_c)^2$$

Dit is dus de variantie van de afstanden op een dag tussen de huishoudens. Hiermee wordt (1.6):

$$\mathrm{var}(\hat{y}) = \sum_{c=1}^{C} n_c s^2(c) = \sum_{c=1}^{C} S(c)$$

Waarmee *S(c)* gedefinieerd is als som van varianties per dag.
Analoog kan (1. 1) herschreven worden als:

$$\hat{y} = \sum_{c=1}^{C} M(c)$$

Nota bene: het feit dat de totale variantie wordt berekend met behulp van varianties per dag, betekent dat voor gegevens, die in principe betrekking kunnen hebben op eenzelfde dag, de varianties van apart berekende cijfers *niet* zomaar bij elkaar opgeteld kunnen worden!

Het zou nu handig zijn om bij een analyse van OVG-cijfers als eerste stap voor iedere dag de getallen *M(c)* en *S(c)* te berekenen. En vervolgens deze cijfers voor de totalen op te tellen.

Helaas is *S(c)* niet direct met behulp van SAS uit de OVG-gegevens te berekenen. Bij het berekenen van de variantie moet namelijk rekening worden gehouden met de 'nul'-afstanden. In de huidige datastructuur staan alleen afstanden als die werkelijk vermeld zijn. Zo is het mogelijk dat er op een dag 60 huishoudens in de steekproef zitten, en er op een bepaald tijdstip slechts in één huishouden een bromfietsrit wordt gemaakt. Dit betekent dat de variantie berekend moet worden van 59 keer nul kilometer en één keer een positief aantal kilometers. Indien de variantie wordt berekend op basis van de gerapporteerde afstanden, dan krijgt men niet het gewenste cijfer. Dit probleem kan het beste omzeild worden door niet de variantie per dag te berekenen maar in plaats daarvan de kwadraten-som (USS (Uncorrected Sum of Squares) in de SAS procedure 'PROC SUMMARY') en de som (SUM in PROC SUMMARY) in het kwadraat.

Uitgaande van de volgende relaties, per dag:

$$\sum_{h=1}^{n_c} (y_{ch} - \bar{y}_c)^2 =$$

$$\sum_{k=1}^{n_c} \left({y_{ch}}^2 - 2\bar{y}_c y_{ch} + \bar{y}^2 \right)^2 =$$

$$\sum_{h=1}^{n_c} {y_{ch}}^2 - 2\bar{y}_c \sum_{h=1}^{n_c} y_{ch} + n_c {\bar{y}_c}^2 .$$

Daar

$$n_c \bar{y}_c = \sum_{h=1}^{n_c} y_{ch}$$

volgt:

$$\sum_{h=1}^{n_c} (y_{ch} - \bar{y}_c)^2 = \sum_{h=1}^{n_c} {y_{ch}}^2 - \bar{y}_c \sum_{h=1}^{n_c} y_{ch} .$$

Als nu

(a) $$\mathrm{afst}(c) = \sum_{h=1}^{n_c} y_{ch}$$

(dit is gelijk aan SUM in SAS PROC SUMMARY) en

(b) $$\mathrm{afst2}(c) = \sum_{h=1}^{n_c} y_{ch}^{\,2}$$

(dit is gelijk aan USS (Uncorrected Sum of Squares) in SAS PROC SUMMARY) dan valt te berekenen:

(c) $$S(c) = \frac{n_c}{n_c - 1} \times \left(\mathrm{afst2}(c) - \mathrm{afst}(c) \times \mathrm{afst}(c) / n_c\right)$$

waarmee $\mathrm{var}(\hat{y})$ te berekenen is door te sommeren.

Deze procedure is in de Bijlage *SAS-programma* gebruikt. Hierbij dient rekening te worden gehouden met het feit dat het mogelijk is dat er veel afrondingsfouten onstaan bij de berekening van de USS als met grote (opgehoogde) afstanden wordt gerekend. Deze dienen door een (grote) constante gedeeld te worden, bijvoorbeeld tot miljoenen kilometers. De gegevens over het aantal huishoudens per dag in de steekproef zijn apart verzameld. Deze zouden ten minste aan het huishoudrecord toegevoegd kunnen worden. Om efficiency-redenen (in de zin van rekentijd) zijn deze gegevens aan alle records toegevoegd.

Een praktisch probleem blijkt de 'datum' van een huishouden te zijn. In tegenstelling tot de instructies vullen niet alle gezinsleden dezelfde datum in voor hun ritten. Dit probleem is op de 'officiële' wijze opgelost: in principe is gekozen voor de datum die door de meest door gezinsleden ingevuld is. Dit is dus per gezinslid, niet per rit, één datum. In geval dat dit geen uitsluitsel geeft, wordt de oudste datum gekozen.

4. Eerste controle op resultaten

Omdat het werken met OVG-cijfers op zich niet altijd eenvoudig is en er veel keuzes gemaakt moeten worden, is begonnen met het reproduceren van gegevens die afkomstig zijn van het CBS. OVG-gegevens worden binnen de SWOV op een aantal punten aangepast. Bovendien blijkt dat de toewijzing van huishoudens aan een bepaalde datum niet in alle gevallen eenduidig is. Deze twee problemen mogen geen substantiële verschillen veroorzaken, toch betekenen ze wel dat een referentietabel niet exact te reproduceren is.

De hier gebruikte referentietabel geeft de totale afgelegde afstand in 1986 weer exclusief de afstand afgelegd tijdens veelvuldige verplaatsingen. Tevens is een opdeling naar geslacht gemaakt. De heer Konen van het CBS heeft de referentietabel geleverd.

	Totaal	Man	Vrouw
CBS afstand	14013602,1	8563261,66	5450340,44
SWOV afstand	14013602,0999	8563261,663316	5450340,4365943
CBS absolute marge	402833,95	283251,89	214161,17
SWOV absolute marge	402847,01	283200,37	214139,64
CBS relatieve marge (%)	2,87	3,31	3,93
SWOV relatieve marge (%)	2,87469	3,30716	3,92892

Tabel 4.1. *Referentietabel van totaal afgelegde afstanden in 1986*

Zoals uit *Tabel 4.1.* blijkt zijn de verschillen marginaal. De afstanden zijn tot ver achter de komma vergelijkbaar, de absolute marges (en waarschijnlijk dus ook de relatieve marges) wijken pas in de vijfde decimaal af. Dit geringe verschil kan heel goed het gevolg zijn van de hier niet optimale berekening van de varianties. In de *Bijlage* wordt de bijbehorende code weergegeven.

5. Voorbeeld

In dit hoofdstuk wordt aan de hand van een voorbeeld nagegaan in welke mate de OVG-cijfers de werkelijkheid weergeven. Er is gekozen voor een voorbeeld waarin de reizigerskilometers van bromfietsers (en snorfietsers) en bijbehorende marges daarvan per jaar worden berekend. Ook wordt naar twee leeftijdscategorieën gedisaggregeerd. Hierbij worden twee gevolgen duidelijk zichtbaar: ten eerste het effect van het feit dat bepaalde combinaties in mindere mate voorkomen in de steekproef, en ten tweede het effect van de (plotselinge) vergroting van de steekproef vanaf 1994 en de daardoor ontstane grotere betrouwbaarheid. De SAS-code is in de *Bijlage* opgenomen.

De gegevens voor het totaal (alle leeftijden) zijn:

Jaar	Aantal	Afstand	Variantie	Absolute marge	Relatieve marge (%)
1985	234	167,1469	135,0127	22,77421	13,62527
1986	257	171,7975	136,9915	22,9405	13,35322
1987	246	157,6366	105,2072	20,10383	12,75327
1988	241	167,8931	194,0654	27,30424	16,26288
1989	234	138,8729	115,9387	21,10426	15,19681
1990	232	150,5526	110,6038	20,61299	13,69155
1991	198	122,2442	105,764	20,15696	16,48909
1992	203	118,97	82,0093	17,74956	14,91936
1993	197	130,0894	126,2885	22,02612	16,93153
1994	336	122,3683	39,0443	12,24715	10,00843
1995	355	117,9889	12,5473	6,94276	5,88425
1996	351	118,6675	16,1693	7,88137	6,64155
1997	357	115,8513	15,7978	7,79031	6,72440

Tabel 5.1. *Reizigerskilometers van brom- en snorfietsers*

Opvallend is ten eerste dat onder 'aantal' (dat is het aantal dagen dat er bromfietsverkeer in de steekproef voorkomt) nooit het maximale aantal van 365/366 wordt bereikt. Dit aantal dagen is duidelijk toegenomen sinds 1994 terwijl de de marges duidelijk kleiner zijn geworden. Dit betekent dat de betrouwbaarheid ongeveer is verdubbeld sinds 1993.

Dezelfde tabel, maar nu voor de categorie van 30 jaar en ouder levert het volgende beeld op:

Jaar	Aantal	Afstand	Variantie	Absolute marge	Relatieve marge (%)
1985	67	31,06054	28,76516	10,5121	33,84391
1986	92	36,00248	34,12774	11,45011	31,80368
1987	99	40,14197	27,86992	10,34723	25,77658
1988	87	43,01984	93,66264	18,96877	44,09308
1989	85	31,35601	22,424	9,28138	29,60001
1990	83	29,09307	17,73837	8,25492	28,37418
1991	78	29,16435	17,40456	8,17688	28,03724
1992	79	28,12173	16,51357	7,96483	28,3227
1993	65	28,04006	20,63644	8,90376	31,75371
1994	232	37,16119	9,54574	6,05565	16,29563
1995	291	37,39044	3,70594	3,77316	10,09124
1996	272	36,14268	4,01818	3,9289	10,87052
1997	280	36,46773	3,99352	3,91682	10,74052

Tabel 5.2. *Reizigerskilometers van brom- en snorfietsers in de leeftijdsgroep 30 jaar en ouder*

Voor deze categorie is de verbetering door de uitbreiding van de steekproef werkelijk substantieel. Waar het cijfer in het begin makkelijk 30% af kon wijken van de werkelijkheid is dit nu aanzienlijk teruggebracht. Een deel hiervan kan het gevolg zijn van de toename in het gebruik (in de steekproef). Risicodalingen in het algemeen van ongeveer 10 procent zullen echter moeilijk aantoonbaar blijven.
Voor de groep 'jongeren', jonger dan 30 jaar in dit geval, geldt:

Jaar	Aantal	Afstand	Variantie	Absolute marge	Relatieve marge (%)
1985	205	136,0863	105,6744	20,14842	14,80562
1986	225	135,795	100,976	19,69542	14,50379
1987	207	117,4946	75,54134	17,03525	14,49874
1988	208	124,8732	96,20584	19,22458	15,39528
1989	197	107,5169	86,60995	18,24064	16,96536
1990	196	121,4595	90,67518	18,66381	15,36628
1991	160	93,0798	85,90375	18,16612	19,51671
1992	157	90,8483	65,39243	15,84965	17,44629
1993	169	102,0493	105,5259	20,13426	19,72993
1994	306	85,2071	29,223	10,59543	12,43491
1995	343	80,5985	8,77656	5,80655	7,2043
1996	339	82,5249	11,78242	6,7278	8,15246
1997	342	79,3836	11,71437	6,70835	8,45054

Tabel 5.3. *Reizigerskilometers van brom- en snorfietsers in de leeftijdsgroep jonger dan 30 jaar*

Deze groep domineert het totaal, de resultaten zijn dan ook vergelijkbaar met het totaal, behalve dat de onzekerheid enigszins groter is.

6. Conclusies en aanbevelingen

In dit verslag is een methode aangegeven die kan worden gebruikt om met behulp van het computerprogramma SAS (relatieve) betrouwbaarheden te berekenen voor OVG-cijfers.

Een vergelijking met door het CBS gepubliceerde schattingen laat zien dat de methode gelijkwaardige uitkomsten geeft.

Het is gebleken dat de marges voor sommige cellen soms meer dan twee keer de geregistreerde afstand in reizigerskilometers kunnen bedragen. Dergelijke afwijkingen zijn zo groot ten opzichte van de geregistreerde afstanden dat voor dergelijke cellen getwijfeld moet worden of de benaderingsmethode voldoende nauwkeurig is om de onzekerheidsmarges zelf goed te berekenen. Verder geldt dat in zo'n geval de marges in ieder geval niet meer symmetrisch om de geregistreerde afstand zullen liggen, hetgeen één van de aannames van de methode is. Waarschijnlijk is het beter in die gevallen een bootstrap-methode volgens CBS(1995) toe te passen en aan de hand daarvan de onder- en bovengrens van de afstanden vast te stellen. In vergelijking met de hier beschreven benaderingsmethode betekent dit dat er meer een onderzoek op maat en een grote hoeveelheid rekenwerk nodig is, waardoor het alternatief voor de meeste toepassingen (voorlopig) niet praktisch is.

Het blijkt verder dat de marges in OVG-cijfers sterk kunnen verschillen indien verder in de steekproef wordt gedisaggregeerd. Hierdoor blijft het relevant om rekening te houden met de onbetrouwbaarheid indien verschillende groepen met elkaar worden vergeleken. Mede om deze redenen valt het sterk aan te bevelen dat een methode voor de benadering van de betrouwbaarheidsmaten van de OVG-cijfers standaard wordt gekoppeld aan programmatuur welke OVG-cijfers produceren. Het feit dat de onzekerheid in samengevoegde cellen niet gelijk is aan de som van de onzekerheden van de individuele cellen, waar dat voor de afstanden wel het geval is, zal tot complicaties leiden bij een algehele implementatie binnen bijvoorbeeld BIS-V. Gezien het belang van BIS-V en de invloed die de onzekerheid op uitspaken kan hebben lijkt het verstandig te onderzoeken of en hoe BIS-V met een OVG-onzekerheidsmodule kan worden uitgerust.

Literatuur

CBS (1993a). *Een vergelijking tussen de BHS-methode en analytische benaderingsformules voor het schatten van relatieve marges van cijfers uit het OVG'*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen. BPA no.: 2646-93-M1/INTERN.

CBS (1993b). *De mobiliteit van de Nederlandse bevolking in 1992*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen.

CBS (1995). *De bootstrap methode toegepast op het ondersoek veplaatsingsgedrag 1994*. Centraal Bureau voor de Statistiek, Voorburg/Heerlen. BPA no.: 10515-95-RSM.

Bijlage Sas-programma

Controle 1986 (hoofdstuk 4)

Beginfase programma:

Initialisatie code, opties en dergelijke....
belangrijk is dat aan data, of een andere logical, de directory van de OVG-data wordt verbonden. Ook is het handig aan 'library' de formats te verbinden.

```
LIBNAME data 'disk3:[ovg.data]' ;
LIBNAME library     'sido_OVG_form';
```

Nu twee relevante instructies, die in principe als argument aan een macro moeten worden doorgegeven. Aan het macro symbool vars ('&vars' als daar in SAS aan gerefereerd wordt) moeten de variabelen waarvoor gedisaggregeerd moet worden worden opgegeven. Bij meerdere variabelen moeten er spaties of tabs tussen de variabelen staan, geen komma's o.i.d. Met jaar wordt het jaar van de steekproef opgegeven. Hier wordt dus een opdeling naar geslacht gemaakt voor het jaar (19)86. De bestandsstructuur is dus, strikt genomen, nog niet 'millenium-proof'.

```
%let vars = sexe;
%let jaar = 86 ;
```

De volgende stap levert een subset op van de (vrij grote) OVG-bestanden. In principe moeten hier de ritten worden geselecteerd. Als categorieën moeten worden samengenomen, dan kan dat over het algemeen het beste hier. Zoals in *Hoofdstuk 3 en 6* is beschreven, kunnen onbetrouwbaarheidsgegevens die (gedeeltelijk) uit dezelfde huishoudens kunnen komen later niet meer samengenomen worden. Dit geldt natuurlijk niet voor de afgelegde afstanden, maar wel voor de varianties daarvan, dit in verband met het bestaan van covarianties.

```
data   temp
  (Keep=
       VolgJr       /* jaar steekproef */
       VolgMnd      /* maand steekproef */
       VolgNr       /* adres in steekproef (nummer) */
       Hhnr         /* huishouden nummer op adres */
       veelvvpl     /* selectie variabele aan de hand waarvan
                       veelvuldige verplaatsingen uit de
                       selectie worden gehaald */
       datum        /* 'officieele' datum van de rit.
                         Zie vorig hoofdstuk */
       nhuish       /* aantal huishoudens op deze officieele
                       datum */
       Afstand      /* afgelegde afstand van deze rit */
       faktorV      /* Factor (CBS) om deze rit naar een
                       landelijk cijfer op te hogen */

       &vars
);

       set    data.ovg&jaar ;
```

veelvuldige verplaatsingen worden weggelaten

```
If     veelvvpl = 0 ;
```

Afstand in de gewenste eenheid:

```
Afstand = Afstand / 10000.000;
```

Op dit moment zijn de basisbestanden van het OVG bij de SWOV nog niet gesorteerd naar volgjaar, volgmaand, volgnummer en huishoudnummer (dit zijn de unieke huishoudens). Tot dat wel het geval is, moet het bestand op die wijze worden gesorteerd om daarna verder verwerkt te kunnen worden.

```
Proc   sort ;
by     volgjr volgmnd VolgNr Hhnr ;
```

Vervolgens wordt, voor ieder huishouden (combinatie van volgjr volgmnd VolgNr Hhnr) per klasse van de variabelen in &vars (hier: oudere) plus het totaal de (totale, opgehoogde) afstand opgeteld.

```
Proc   summary ;
by     volgjr volgmnd VolgNr Hhnr ;
Class  &vars ;
id     datum nhuish;
var    Afstand ;
WEIGHT faktorv ;
output out = hhdata (drop=_type_ _freq_ )
       Sum(Afstand) = Afstand ;
```

Het werkbestand 'hhdata' bevat nu:

- de huishouden indentificatie variabelen: volgjr volgmnd VolgNr Hhnr,
- de variabelen in '&vars',
- de totale afstand (som) in 'Afstand' ,
- de 'ID' variabelen 'datum' en 'nhuish', respectievelijk de 'officiele huishouden datum (dus niet noodzakelijk de toegewezen datum) en het aantal huishoudens dat op die datum in de steekproef bleek te zitten. Het programma gaat er vanuit dat de combinaties datum en nhuish consistent zijn in de dataset. Deze worden niet meer gecontroleerd.

De data worden op volgorde van datum gezet (volgjr datum is qua programmering 'veiliger', maar niet noodzakelijk. Het is meer de bedoeling om op een handige wijze 'volgjr' mee te kunnen nemen in de proc summary stappen die volgen).

```
Proc   sort ;
By     volgjr datum ;
```

De volgende stap berekent uit Hoofdstuk 3 de formules van (a) en (b). De optie MISSING zorgt er voor dat de marginalen van de tabel ook berekend worden. Er wordt een eenvoudige controle op het gecodeerde aantal huishoudens per datum uitgevoerd. Indien er iets niet klopt worden de records in de dataset 'ERGFOUT' weggeschreven.

```
PROC   Summary nway missing;
Class  &vars ;
BY     volgjr datum ;
VAR    afstand nhuish;

OUTPUT OUT = dag
       Min( nhuish )  = nhuish
       Max( nhuish )  = nhuisht
       SUM( afstand ) = afstand   /* formule (a) */
       USS( afstand ) = afst2 ;   /* formule (b) */
```

De volgende stap bestaat uit het uitvoeren van een controle op aantallen huishoudens. Indien het maximum van de gecodeerde aantallen huishoudens niet gelijk is aan het minimum van de gecodeerde aantallen huishoudens (dit betekent dat er dus verschillen op een datum voorkomen) wordt er een alarm gegeven. In principe zou dit niet mogen voorkomen en in de toekomst zal dit stukje code verdwijnen.

```
Data   fout   dag ;
       set    dag ;
       If     nhuish ne nhuisht then output fout ;
       output dag ;

proc   append base=ergfout new = fout force;
```

In de nu volgende stap wordt er per datum (en disaggregatie) de variantie van de afstanden berekend. Dit is stap (c) in Hoofdstuk 3.

Om onnodige waarschuwingen te voorkomen worden onbruikbare records eerst weggelaten.

```
data   dag ;
       set    dag    ;

       If Not(nhuish eq . OR nhuish LE 1 or afstand eq . or
              afst2 eq .);

       vari = (nhuish/(nhuish-1)) *
              (afst2-afstand*afstand/nhuish) ; /* stap (c) */
```

Er bestaat nu een bestand met voor iedere dag (en voor ieder disaggregatie-niveau) een afstand en een variantie. Deze worden nu opgeteld.

```
PROC   Summary nway Missing;
Class  volgjr &vars;
VAR    afstand vari;

OUTPUT OUT = out
       SUM( afstand vari) = afstand vari ;
```

Vervolgens worden de 95% marges berekend. In 'abs' de absolute marge, in 'rel' de relatieve marge.

```
data   out    ;
       set    out    ;

         rel = 100 * abs/afstand ;
         abs = 1.96 * Sqrt(vari) ;
```

De gegevens worden aan een bestand toegevoegd, zodat de macro voor meerdere jaren 'apart' kan worden uitgevoerd.

```
proc   append base=total new = out force;
run;
```

Vervolgens worden de resulaten geprint. Ook worden eventuele fouten geprint.

```
proc   print data = total ;

Proc   print data = ergfout ;
```

SAS programma van het voorbeeld (hoofdstuk 5)

```
options nofmterr;
libname      current "[]";
libname data 'disk3:[ovg.data]' ;
LIBNAME library        base    'sido_OVG_form';

%let vars = oudere;
%let jaar = 86 ;

%macro dojaar(jaar) ;

data   temp
  (Keep=
       VolgJr       /* jaar steekproef */
       VolgMnd      /* maand steekproef */
       VolgNr       /* adres in steekproef (nummer) */
       Hhnr         /* huishouden nummer op adres */
       veelvvpl     /* selectie variabele aan de hand waarvan
                       veelvuldige verplaatsingen uit de
                       selectie worden gehaald */
       datum        /* 'officieele' datum van de rit.
                         Zie vorig hoofdstuk */
       nhuish       /* aantal huishoudens op deze officieele
datum */
       Afstand      /* afgelegde afstand van deze rit */
       faktorV      /* Factor (CBS) om deze rit naar een
landelijk
                       cijfer op te hogen */
       vvmid        /* vervoersmiddel */
       kleeft       /* code leeftijd */

       &vars
);

       set    data.ovg&jaar ;
```

veelvuldige verplaatsingen worden weggelaten

```
If     veelvvpl = 0 ;
```

alleen ritten met brom- of snorfiets worden geselecteerd

```
If vvmid eq 30       /* snorfiets */
  or vvmid eq 40 ;   /* bromfiets */
```

onder de dertig jaar wordt als jongere gecodeerd

```
If kleeft gt 2529 then oudere = 1 ; else oudere = 0 ;


       Afstand = Afstand / 10000000; /* miljoenen kms's */
```

Op dit moment zijn de basisbestanden van het OVG bij de SWOV nog niet gesorteerd naar volgjaar, volgmaand, volgnummer en huishoudnummer (dit zijn de unieke huishoudens) tot die tijd moet het bestand op die wijze worden gesorteerd om daarna verder verwerkt te kunnen worden

```
Proc   sort ;
by     volgjr volgmnd VolgNr Hhnr ;
```

vervolgens wordt, voor ieder huishouden (combinatie van volgjr volgmnd VolgNr Hhnr) per klasse van de variabelen in &vars (hier: oudere) plus het totaal de (totale, opgehoogde) afstand opgeteld.

```
Proc     summary ;
by       volgjr volgmnd VolgNr Hhnr ;
Class    &vars ;
```

```
id        datum nhuish;
var       Afstand ;
WEIGHTfaktorv ;
output    out = hhdata (drop=_type_ _freq_ )
          Sum(Afstand) = Afstand ;
```

het werkbestand 'hhdata' bevat nu:

- de huishouden indentificatie variabelen: volgjr volgmnd VolgNr Hhnr
- de variabelen in '&vars',
- de totale afstand (som) in 'Afstand'
- de 'ID' variabelen 'datum' en 'nhuish', respectievelijk de 'officiele huishouden datum (dus niet noodzakelijk de toegewezen datum) en het aantal huishoudens dat op die datum in de steekproef bleek te zitten.

Het programma gaat er van uit de combinaties datum en nhuish consistent zijn in de dataset. Deze worden niet meer gecontroleerd.

De data worden op volgorde van datum gezet (volgjr datum is qua programmering 'veiliger', niet noodzakelijk. Het is meer de bedoeling om op een handige wijze volgjr mee te kunnen nemen in de proc summary stappen die volgen).

```
Proc   sort ;
By     volgjr datum ;
```

De volgende stap berekent de formules van (a) en (b) uit. De optie MISSING zorgt er voor dat de marginalen van de tabel ook berekend worden. Er wordt een eenvoudige controle op het gecodeerde aantal huishoudens per datum uitgevoerd. Indien er iets niet klopt worden de records in de dataset 'ERGFOUT' weggeschreven.

```
PROC   Summary nway missing;
Class  &vars ;
BY     volgjr datum ;
VAR    afstand nhuish;

OUTPUT OUT = dag
       Min( nhuish )   = nhuish
       Max( nhuish )   = nhuisht
       SUM( afstand ) = afstand   /* formule (a) */
       USS( afstand ) = afst2 ;   /* formule (b) */
```

uitvoeren controle op aantallen huishoudens. Indien het maximum van de gecodeerde aantallen huishoudens niet gelijk is aan het minimum van de gecodeerde aantallen huishoudens (er dus verschillen op een datum voorkomen) wordt er een alarm gegeven. In principe zou dit niet mogen voorkomen en in de toekomst zal dit stukje code verdwijnen

```
Data   fout   dag ;
       set    dag ;
       If     nhuish ne nhuisht then output fout ;
       output dag ;

proc   append base=ergfout new = fout force;
```

in de nu volgende stap wordt er per datum (en disaggregatie) de variantie van de afstanden berekend. Dit is stap (c) in Hoofdstuk 3.
Om onnodige waarschuwingen te voorkomen worden onbruikbare records eerst weggelaten

```
data   dag ;
```

```
      set    dag    ;

      If Not(nhuish eq . OR nhuish LE 1 or afstand eq . or
afst2 eq .);

      vari = (nhuish/(nhuish-1)) *
             (afst2-afstand*afstand/nhuish) ; /* stap (c) */
```

Er bestaat nu een bestand met voor iedere dag (en voor ieder disaggregatie niveau) een afstand en een variantie. Deze worden nu opgeteld.

```
PROC   Summary nway Missing;
Class  volgjr &vars;
VAR    afstand vari;

OUTPUT OUT = out
       SUM( afstand vari) = afstand vari ;
```

Vervolgens worden de 95% marges berekend. In 'abs' de absolute marge, in rel derelatieve marge

```
data   out    ;
       set    out    ;

         abs = 1.96 * Sqrt(vari) ;
         rel = 100 * abs/afstand ;
```

de gegevens worden aan een bestand toegevoegd, zodat de macro voor meerdere jaren 'apart' kan worden uitgevoerd

```
proc   append base=total new = out force;
run;

%mend;
```

de volgende macro voert de vorige macro uit voor de jaren (19)85 tot en met (19)97.

```
%macro jaren;
%do jaar=85 %to 97 ;
%dojaar(&jaar);
%end;
%mend;
```

roep de laatste macro aan

```
%jaren;
```

voer resultaten uit

```
proc   print data = total (drop=_type_ vari) ;
Format afstand abs rel 10.3;

Proc   print  data = ergfout ;

data   current.voorb;
       set    total ;
       file   voorbeeld ;
       Put    VOLGJR &vars afstand abs rel ;
```