

Panels

Deelthema: Bedrijvenpanels



Paul Knottnerus

Statistische Methoden (201102)



Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader voorlopig cijfer
x	=	geheim
–	=	nihil
–	=	(indien voorkomend tussen twee getallen) tot en met
o (o,o)	=	het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	=	een cijfer kan op logische gronden niet voorkomen
2010–2011	=	2010 tot en met 2011
2010/2011	=	het gemiddelde over de jaren 2010 tot en met 2011
2010/'11	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2010 en eindigend in 2011
2008/'09–		
2010/'11	=	oogstjaar, boekjaar enz., 2008/'09 tot en met 2010/'11

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

Colofon

Uitgever

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

Prepress

Centraal Bureau voor de Statistiek -
Grafimedia

Omslag

TelDesign, Rotterdam

Inlichtingen

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

Bestellingen

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

Internet

www.cbs.nl

ISSN: 1876-0333

© Centraal Bureau voor de Statistiek,
Den Haag/Heerlen, 2011.
Vereenvoudiging is toegestaan,
mits het CBS als bron wordt vermeld.

Inhoudsopgave

1.	Inleiding op het deelthema Bedrijvenpanels.....	5
1.1	Algemene beschrijving.....	5
1.2	Afbakening en relatie met andere thema's	6
1.3	Plaats in het statistische proces.....	6
1.4	Definities	6
2.	Panels uit bedrijfspopulaties zonder migratie	8
2.1	Korte beschrijving	8
2.2	Toepasbaarheid.....	8
2.3	Uitgebreide beschrijving	8
2.3.1	Inleiding	8
2.3.2	Schatter voor het omzettotaal van de populatie	9
2.3.3	Schatter voor een 12-maands groeipercentage.....	10
2.3.4	Belang van grote overlap bij panels	11
2.3.5	Belang en omvang van de jaarlijkse panelverversing	12
2.3.6	Geboorten en sterften in de populatie	13
2.4	Kwaliteitsindicatoren.....	15
2.5	Appendix. Covariantieformules bij overlappende steekproeven.....	15
3.	Panels met migratie tussen de strata	17
3.1	Korte beschrijving	17
3.2	Toepasbaarheid.....	17
3.3	Uitgebreide beschrijving	17
3.3.1	Inleiding	17
3.3.2	Jaarlijkse correctie en verversing in januari.....	17
3.3.3	De covariantieterm.....	18
3.3.4	Schatting van de covariantieterm	20
3.4	Kwaliteitsindicatoren.....	22
3.5	Appendix. Rechtvaardiging van (3.6).....	23
4.	Panels voor het schatten van indicatoren en indices	28
4.1	Korte beschrijving	28
4.2	Toepasbaarheid.....	28

4.3	Uitgebreide beschrijving	28
4.3.1	De ratioschatter voor een algemene index	28
4.3.2	De PPS-schatter voor een algemene index.....	31
4.3.3	Vervanging of aanvulling van de systematische PPS-steekproef	32
4.3.4	Fusies van bedrijven.....	32
4.3.5	Splitsingen van bedrijven.....	36
4.4	Kwaliteitsindicatoren.....	37
5.	Literatuur	38

1. Inleiding op het deelthema Bedrijvenpanels

1.1 Algemene beschrijving

Panels waarbij dezelfde eenheden van een populatie in meerdere perioden worden waargenomen, worden sinds de jaren '80 steeds vaker gebruikt. Panels zijn vooral aantrekkelijk wanneer men is geïnteresseerd in het meten van veranderingen in populatiegrootheden van bepaalde doelvariabelen.

In eerste instantie zullen we ons in dit document beperken tot panels voor het schatten van omzetveranderingen bij bedrijven en mutaties van prijsindexcijfers zoals de PPI. Maar dezelfde methoden kunnen met de nodige aanpassingen ook worden toegepast op panels bij personen/huishoudens om bijvoorbeeld de werkloosheidscijfers te bepalen. Omdat iedere statistiek zijn eigen karakteristieken heeft, zijn de aanpassingen vaak ook weer heel specifiek en moeilijk in één algemeen document samen te vatten.

Naast de desbetreffende schatters komen ook de varianties van de verschillende schatters aan de orde. Tevens gaan we nader in op de verversing van het panel, de vervanging bij uitval/sterfte van bedrijven in het panel alsook op de gevolgen voor het panel van een fusie tussen bedrijven.

Verder dient te worden opgemerkt dat in dit document veel aandacht wordt besteed aan het afleiden van de varianties van geschatte groeivoeten. Reden hiervoor is dat een regelmatig terugkerende vraag is hoe groot de onzekerheidsmarges zijn van door het CBS gepubliceerde groeicijfers; vaak is hierbij sprake van externe druk. Elders in de literatuur blijkt dit onderwerp nauwelijks aan de orde te komen. Alleen Nordberg (2000) gaat nader op dit onderwerp in, zij het helemaal toegespitst op de Zweedse situatie. Een ander nadeel van zijn aanpak is dat nog steeds simulaties nodig zijn om bepaalde componenten van diverse covarianties te schatten. Met name hoofdstuk 3 in dit document waarbij rekening wordt gehouden met migratie van bedrijven tussen strata, kan op sommige lezers wat technisch overkomen.

In hoofdstuk 2 gaan we nader in op panels onder bedrijven die zijn opgedeeld in strata. Accent ligt daarbij vooral op het schatten van de relatieve verandering van een omzettotaal in een bepaalde maand ten opzichte van 12 maanden geleden plus de bijbehorende onzekerheidsmarges. We nemen hierbij aan dat de bedrijven niet verhuizen van het éne naar het andere stratum. Wel mogen er geboorten en sterften plaatsvinden bij de bedrijven. In hoofdstuk 3 komen panels aan de orde waarbij bedrijven kunnen verhuizen van de éne grootteklasse naar de andere als gevolg van een verandering in het aantal werkzame personen in de desbetreffende periode van 12 maanden. Hoofdstuk 4 behandelt het schatten van indices en veranderingen daarin ten opzichte van een basisperiode. In dit hoofdstuk wordt ook aandacht besteed aan de gewichten van bedrijven die betrokken zijn bij een fusie of een splitsing.

1.2 Afbakening en relatie met andere thema's

Het spreekt voor zich dat er een sterk verband bestaat tussen de theorie voor panels en de deelthema's *Steekproefontwerpen* en *Ophoogmethoden*. De lezer wordt geacht vertrouwd te zijn met de bij die deelthema's behandelde stof, in het bijzonder de gestratificeerde schatter, de quotiëntschatter en de daarmee gerelateerde schatter van een ratio, inclusief de bijbehorende variantieformules. Ook is er een relatie met het thema *Wegen als correctie voor non-respons* en met het deelthema *Duurmodellen*. Hier zullen we evenwel stilzwijgend aannemen dat de non-respons niet tot systematische vertekening leidt tenzij anders vermeld. Verder bestaat er een relatie met het thema *Indexcijfers*.

1.3 Plaats in het statistische proces

Panels spelen op verschillende plekken in het statistische proces een belangrijke rol. Het ontwerpen van panels en het trekken van de bijbehorende steekproeven vergen een degelijke voorbereiding aan het begin van het proces terwijl de berekeningen voor de desbetreffende schattingen met de bijbehorende onzekerheidsmarges meer aan het einde van het statistische proces plaatsvinden.

1.4 Definities

Begrip	Omschrijving
Enkelvoudige aselechte steekproef	een steekproefontwerp waarbij de elementen met gelijke kansen uit het steekproefkader worden getrokken. Bij trekking zonder teruglegging moet elke deelverzameling ter grootte van de steekproefomvang dezelfde kans op realisatie hebben. Bij trekking met teruglegging moeten de trekkingen van de opeenvolgende elementen onafhankelijk van elkaar zijn.
Indexcijfer	cijfer dat de verhouding van een waarde tot een als referentie aangenomen grootte aanwijst; meestal als percentage
Insluitkans	de kans dat een element wordt getrokken bij een steekproef zonder teruglegging; afhankelijk van het steekproefontwerp kan deze kans per element verschillen
Panel-onderzoek	een onderzoek waarbij waarneemheden meerdere keren worden benaderd in verschillende peilingen om zo de ontwikkeling in de tijd (micro of macro) te kunnen volgen
Populatieparameter	een te schatten waarde van een kenmerk van een populatie, bijvoorbeeld de gemiddelde score in de populatie op een bepaalde variabele

Quotiëntschatte	een schatte voor een populatietotaal die gebruik maakt van een kwantitatieve hulpvariabele. De nauwkeurigheid van deze schatte is groter naarmate de quotiënten van de scores op de doel- en hulpvariabele minder variëren. Als synoniem wordt vaak ‘ratioschatte’ gebruikt, maar deze term wordt in hoofdstuk 4 gereserveerd voor een schatte van een ratio (quotiënt).
Steekproefallocatie	het verdelen van de totale steekproefomvang over de verschillende strata bij een gestratificeerde steekproef
Systematische steekproef	steekproefontwerp waarbij elementen uit het steekproefkader worden geselecteerd door er op een systematische manier doorheen te lopen. Alleen het startpunt wordt geloot. Daarna wordt een vaste staplengte (spronggrootte) gehanteerd.
Vertekening (van een schatte)	verwachting van het verschil tussen de schatte en de werkelijke waarde van een (populatie)parameter, ook wel ‘systematische fout’ genoemd. Is dat verschil gelijk aan nul, dan heet de schatte zuiver (unbiased), en anders onzuiver (biased).

2. Panels uit bedrijfspopulaties zonder migratie

2.1 Korte beschrijving

Beschouw een populatie van bedrijven die is opgesplitst in strata naar grootteklasse en SBI-code. We nemen aan dat er in de populatie geboorten en sterften van bedrijven kunnen plaatsvinden maar dat er geen verhuizingen zijn van bedrijven van het éne naar het andere stratum.

Het belangrijkste doel van een panel in deze context is om voor iedere maand t ($t = 1, 2, \dots$) zowel de totale omzet van de doelpopulatie te schatten als de relatieve verandering van het omzettotaal van die maand ten opzichte van 12 maanden geleden. Dit is inclusief de schatting van de bijbehorende onzekerheidsmarges. Uitgangspunt voor de schatting van de relatieve verandering in een omzettotaal vormen de schattingen van de omzettotaal in de desbetreffende maanden. De quotiëntschatting speelt derhalve een belangrijke rol bij de methodologie voor panels.

Het maandelijks panel dat we in dit hoofdstuk nader bekijken, is opgebouwd uit enkelvoudige, aselechte steekproeven zonder teruglegging (EAZT) uit de afzonderlijke strata. De steekproef fracties zijn vast maar kunnen wel per stratum verschillen. Hierbij geldt hoe hoger de grootteklasse, hoe hoger de steekproef fractie. De grootste grootteklassen worden overigens integraal waargenomen. De bedrijven in het panel kunnen van maand tot maand verschillen.

2.2 Toepasbaarheid

Zowel bij bedrijfseconomische als bij sociaal-economische statistieken worden panels gebruikt. Reden is dat veranderingen in allerlei populatietotaal tussen een tweetal maanden met behulp van panels vaak met grotere nauwkeurigheid kunnen worden geschat dan met volledig onafhankelijke steekproeven. Indien er nauwelijks sprake is van eenheden die verhuizen tussen verschillende strata, is dit hoofdstuk van toepassing.

2.3 Uitgebreide beschrijving

2.3.1 Inleiding

In dit hoofdstuk zullen we met name formules afleiden voor schatters van 12-maands groeicijfers van de maandomzet bij bedrijven met een bepaalde SBI-code. Eerst behandelen we populaties zonder geboorten en sterften. In paragraaf 2.3.6 komen populaties met geboorten en sterften aan de orde. Verder nemen we aan dat ieder jaar in januari in ieder stratum 10% procent van de steekproef wordt ververs.

De effecten van gaafmaken, imputeren en andere vergelijkbare rekenslagen worden bij de afleiding van de formules voor de onzekerheidsmarges voorlopig buiten beschouwing gelaten evenals de non-respons.

2.3.2 Schatter voor het omzettotaal van de populatie

Laat O^t ($t=1,2,\dots$) zijn gedefinieerd als het omzettotaal in maand t van alle bedrijven in de doelpopulatie met een bepaalde SBI-code ofwel $O^t = \sum_{i=1}^N O_i^t$ waarbij O_i^t staat voor de omzet van bedrijf i in maand t ($i=1,\dots, N^t$). Zoals is aangegeven in onder meer Banning e.a. (2010), kan O^t worden geschat met de volgende stratificatieschatter op basis van H grootteklassen

$$\hat{O}^t = \sum_{h=1}^H N_h^t \bar{o}_h^t,$$

waarbij

N_h^t : aantal bedrijven in stratum/grootteklasse h ($h=1,\dots,H$) in maand t ;

$\bar{o}_h^t$: gemiddelde omzet in steekproef s_h^t in maand t in stratum h ($\bar{o}_h^t = \frac{1}{n_h^t} \sum_{i=1}^{n_h^t} o_{hi}^t$);

o_{hi}^t : omzet van i^e bedrijf in de steekproef uit stratum h ;

n_h^t : aantal bedrijven in de steekproef uit stratum h in maand t .

Definieer verder

$n_{hh}^{t-12,t}$: aantal bedrijven in de overlap van de steekproeven uit stratum h in de maanden $t-12$ en t ;

$\bar{O}_h^t$: gemiddelde omzet per bedrijf in stratum h in maand t ($\bar{O}_h^t = \frac{1}{N_h^t} \sum_{i=1}^{N_h^t} O_{hi}^t$);

O_{hi}^t : omzet van i^e bedrijf uit stratum h ;

f_h : steekproeffractie in stratum h .

Omdat onder de aanname dat er geen geboorten en sterften zijn, N_h^t en n_h^t niet van t afhangen, kan in principe het superscript t worden weggelaten. Omdat de formules met superscript t of $t-12$ ook bruikbaar zijn voor de meer algemene situatie dat zich wel geboorten en sterften voordoen in de populatie, laten we de superscripten t in deze paragraaf toch staan. De variantie van de schatter $\hat{O}^t$ is gelijk aan

$$\begin{aligned} \text{var}(\hat{O}^t) &= \sum_{h=1}^H (N_h^t)^2 (1 - f_h) \frac{(S_h^t)^2}{n_h^t} \\ (S_h^t)^2 &= \frac{1}{N_h^t - 1} \sum_{i=1}^{N_h^t} (O_{hi}^t - \bar{O}_h^t)^2. \end{aligned} \tag{2.1}$$

De variantie van $\hat{O}^t$ kan worden geschat met

$$\hat{\text{var}}(\hat{O}^t) = \sum_{h=1}^H (N_h^t)^2 (1 - f_h) \frac{(s_h^t)^2}{n_h^t}$$

$$(s_h^t)^2 = \frac{1}{n_h^t - 1} \sum_{i=1}^{n_h^t} (o_{hi}^t - \bar{o}_h^t)^2.$$

Soortgelijke formules zijn van toepassing op $\hat{O}^{t-12}$.

2.3.3 Schatter voor een 12-maands groeipercentage

Laat $g^{t,s}$ zijn gedefinieerd als de relatieve omzetgroei tussen de maanden s en t ofwel

$$g^{t,s} = \frac{O^t}{O^s} - 1, \quad t > s.$$

De grootheid $g^{t,s}$ kan worden geschat met

$$\hat{g}^{t,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1. \quad (2.2)$$

We zullen nu stapsgewijs een uitdrukking afleiden voor $\text{var}(\hat{g}^{t,t-12})$. De variantie van het rechterlid van (2.2) kan worden benaderd door gebruik te maken van een eerste-orde Taylorreeksontwikkeling van een ratio. Dit geeft voor $s=t-12$

$$\begin{aligned} \text{var}(\hat{g}^{t,t-12}) &= \text{var}\left\{ \frac{\hat{O}^t}{\hat{O}^{t-12}} \right\} = \frac{1}{(O^{t-12})^2} \text{var}(\hat{O}^t - G^{t,t-12} \hat{O}^{t-12}) \\ &= \frac{1}{(O^{t-12})^2} \{ \text{var}(\hat{O}^t) + (G^{t,t-12})^2 \text{var}(\hat{O}^{t-12}) \\ &\quad - 2G^{t,t-12} \text{cov}(\hat{O}^{t-12}, \hat{O}^t) \} \\ G^{t,t-12} &\int \frac{O^t}{O^{t-12}} = 1 + g^{t,t-12}. \end{aligned} \quad (2.3)$$

De twee variantietermen in (2.3) kunnen op de standaard manier worden geschat zoals is vermeld in de vorige deelparagraaf. De term $G^{t,t-12}$ kan worden geschat met

$$\hat{G}^{t,t-12} \int \frac{\hat{O}^t}{\hat{O}^{t-12}}.$$

De covariantieterm $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ in (2.3) is gelijk aan

$$\begin{aligned}
\text{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \sum_{h=1}^H \text{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t) \\
&= \sum_{h=1}^H \text{cov}(N_h^{t-12} \bar{o}_h^{t-12}, N_h^t \bar{o}_h^t) \\
&= \sum_{h=1}^H N_h^{t-12} N_h^t \text{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t),
\end{aligned} \tag{2.4}$$

waarbij we hebben gebruikt dat de steekproeven per stratum onafhankelijk van elkaar worden getrokken. Voor het afleiden van een uitdrukking voor de covarianties in de laatste regel van (2.4) gebruiken we dat $n_{hh}^{t-12,t} = 0.9n_h^{t-12}$ vast is wegens de eerder genoemde aanname dat in januari in ieder stratum 10% van de steekproef wordt vervangen door andere bedrijven van buiten de steekproef. Uit (2.12) in de appendix (paragraaf 2.5) volgt dan

$$\begin{aligned}
\text{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t) &= \left(\frac{n_{hh}^{t-12,t}}{n_h^{t-12} n_h^t} - \frac{1}{N_{hh}^{t-12,t}} \right) S_{hh}^{t-12,t} \\
S_{hh}^{t-12,t} &= \frac{1}{N_{hh}^{t-12,t} - 1} \sum_{i=1}^{N_{hh}^{t-12,t}} (o_{hi}^{t-12} - \bar{o}_h^{t-12})(o_{hi}^t - \bar{o}_h^t).
\end{aligned} \tag{2.5}$$

$N_{hh}^{t-12,t}$ staat voor het aantal bedrijven die zowel in maand $t-12$ als maand t tot stratum h behoren. Onder de aanname dat er geen geboorten en sterften plaatsvinden in de populatie, geldt er $N_{hh}^{t-12,t} = N_h^{t-12} = N_h^t$. Uit (2.5) volgt dat $\text{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t)$ uit de eerste regel van (2.4) zuiver kan worden geschat met

$$\begin{aligned}
\hat{\text{cov}}(\hat{O}_h^{t-12}, \hat{O}_h^t) &= N_h^{t-12} N_h^t \left(\frac{n_{hh}^{t-12,t}}{n_h^{t-12} n_h^t} - \frac{1}{N_{hh}^{t-12,t}} \right) s_{hh}^{t-12,t} \\
s_{hh}^{t-12,t} &= \frac{1}{n_{hh}^{t-12,t} - 1} \sum_{i=1}^{n_{hh}^{t-12,t}} (o_{hi}^{t-12} - \bar{o}_h^{t-12})(o_{hi}^t - \bar{o}_h^t).
\end{aligned}$$

2.3.4 Belang van grote overlap bij panels

Om wat meer inzicht te krijgen in het belang van een grote overlap bij panels maken we de vereenvoudigende aannames dat $S_h^{t-12} = S_h^t = S_h$, $n_h^{t-12} = n_h^t = n_h$ en $G^{t,t-12} = 1$. Definieer verder de overlapfractie I_h als $I_h = n_{hh}^{t-12,t} / n_h$. Wegens de eerdere aanname $N_{hh}^{t-12,t} = N_h^{t-12} = N_h^t = N_h$ volgt uit (2.5) een bovengrens voor de covarianties in de eerste regel van (2.4)

$$\text{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t) \leq N_h^2 (I_h - f_h) \frac{S_h^2}{n_h}.$$

In combinatie met (2.1) en (2.3) geeft dit de volgende ondergrens voor $\text{var}(\hat{g}^{t,t-12})$

$$\text{var}(\hat{g}^{t,t-12}) \geq \frac{2}{(O^{t-12})^2} \sum_{h=1}^H N_h^2 (1 - f_h) \frac{S_h^2}{n_h}. \quad (2.6)$$

Deze formule geeft aan dat als de overlapfracties f_h dalen van bijvoorbeeld 0,9 naar 0,5, de ondergrens voor de variantie van schatter $\hat{g}^{t,t-12}$ met een factor 5 toeneemt. De ondergrens in bovenstaande formule wordt bereikt wanneer naast bovengenoemde aannames ook nog geldt dat $\text{corr}(O_{hi}^{t-12}, O_{hi}^t) = 1$ ofwel wanneer de correlatiecoëfficiënt tussen de O_{hi}^{t-12} en de O_{hi}^t gelijk aan 1 is. In aanvulling hierop zij nog opgemerkt dat de omvang van de overlap volstrekt irrelevant is in de extreme situatie dat $S_{hh}^{t-12,t} = 0$.

2.3.5 Belang en omvang van de jaarlijkse panelverversing

Vooraf bij omzetstatistieken is het van belang dat mutaties in de populatie snel worden verwerkt in het panel omdat de omzet zelf nu de doelvariabele is. Bij panels ten behoeve van bijvoorbeeld prijsindexcijfers is deze urgentie wat minder omdat daar de omzet alleen als gewicht dient terwijl doorgaans de correlatie tussen omzetgewicht en prijsveranderingen gering is. Behalve voor mutaties in de populatie is een verversing van het panel ofwel een roterend panel ook nodig om de vertekening van een panel beperkt te houden. De vertekening in een panel kan het gevolg zijn van paneluitval (Engels, *panel attrition*) bijvoorbeeld bij sterfte en/of weigering om nog langer aan het panel deel te nemen. Ten slotte, een derde doel van verversing is een rechtvaardiger lastendrukverdeling over de desbetreffende bedrijven.

Om te laten zien welke effect de omvang van de jaarlijkse verversing $(1 - f_h)$ heeft op de onzekerheidsmarges rondom de geschatte omzetgroei kijken we eerst naar de *relatieve* onzekerheidsmarge rondom het geschatte omzettotaal onder dezelfde aannames als in de vorige subparagraaf. Uit (2.1) volgt dat de relatieve 95%-onzekerheidsmarge rondom het geschatte omzettotaal $\hat{O}^{t-k}$ ($k = 0, 12$), zeg $RM_{95}(\hat{O}^{t-k})$, kan worden geschreven als

$$RM_{95}(\hat{O}^{t-k}) = \frac{1,96}{O^{t-k}} \sqrt{\sum_{h=1}^H N_h^2 (1 - f_h) \frac{S_h^2}{n_h}}. \quad (2.7)$$

Onder de additionele aanname dat $\text{corr}(O_{hi}^{t-12}, O_{hi}^t) = 1$ volgt verder uit (2.6) dat de *absolute* onzekerheidsmarge rondom de geschatte omzetgroei $\hat{g}^{t,t-12}$ bij benadering gelijk is aan

$$M_{95}(\hat{g}^{t,t-12}) = \frac{1,96}{O^{t-12}} \sqrt{2 \sum_{h=1}^H N_h^2 (1 - f_h) \frac{S_h^2}{n_h}}. \quad (2.8)$$

Uit deze twee formules volgt dat $RM_{95}(\hat{O}^{t-12}) = M_{95}(\hat{g}^{t,t-12})$ wanneer $f_h = (1 + f_h)/2$. Met andere woorden, wanneer de relatieve onzekerheidsmarge

bijvoorbeeld 1% is geldt onder de voorwaarde $\lambda_h = (1 + f_h)/2$ dat de absolute onzekerheidsmarge rondom de geschatte omzetgroei 1 procentpunt is. De globale formule (2.8) voor de marge maakt ook duidelijk dat naarmate de verversing toeneemt, $M_{95}(g)$ stijgt. Indien echter de hier gemaakte aannames weinig realistisch zijn, zit er weinig anders op dan één van de complexere formules voor $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ te gebruiken die elders in dit document aan bod komen. Dit is met name het geval wanneer $\text{corr}(O_{hi}^{t-12}, O_{hi}^t)$ niet in de buurt van 1 ligt.

2.3.6 Geboorten en sterften in de populatie

Indien er geboorten en sterften plaatsvinden bij de bedrijven in de doelpopulatie, kunnen vrijwel dezelfde formules worden gebruikt als in de paragrafen 2.3.2 en 2.3.3. Alleen de formule voor $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ in (2.3) wordt wat ingewikkelder. Dit hangt samen met de maandelijkse update procedure van de steekproef waarbij rekening moet worden gehouden met de geboorten en de sterften in de populatie iedere maand. Hieronder beschrijven we de maandelijkse update procedure.

Definieer $U_{0h}^{t-1,t}$ als de verzameling geboorten in stratum h in maand $t-1$ en definieer $N_{0h}^{t-1,t}$ als haar omvang. Het aantal bedrijven dat wordt getrokken uit $U_{0h}^{t-1,t}$ in maand t is $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. Vervolgens wordt de rest van de steekproef uit stratum h hieraan aangepast op de volgende manier. Definieer $n_{h,RST}^{t-1,t}$ als het benodigde aantal bedrijven voor de rest van de steekproef, dat wil zeggen $n_{h,RST}^{t-1,t} = n_h^t - n_{0h}^{t-1,t}$. Merk op dat $n_h^t = f_h N_h^t$ vast is. Definieer verder de zogenaamde pre-steekproef $s_{h,PRE}^t$ in maand $t-1$ als $s_{h,PRE}^t = s_h^{t-1} \cap U_h^t$ oftewel de verzameling bedrijven in s_h^{t-1} die nog steeds bestaan in maand t . Noteer de omvang van $s_{h,PRE}^t$ met $n_{h,PRE}^t$. Indien $n_{h,PRE}^t > n_{h,RST}^{t-1,t}$, verwijder dit verschil dan aselekt uit de steekproef en indien $n_{h,PRE}^t \leq n_{h,RST}^{t-1,t}$, trek dan het verschil aselekt bij uit $(U_h^t \setminus U_{0h}^{t-1,t}) \setminus s_{h,PRE}^t$. Merk op dat bedrijven die aldus bij een update uit de steekproef worden gehaald, later weer in de steekproef kunnen belanden.

Op grond van de maandelijkse update procedure kan stratum h in maand $t-12$ in twee substrata worden verdeeld: (i) substratum $U_{h,H+1}^{t-12,t}$ van bedrijven in stratum h die sterven in de maanden $t-12, \dots, t-1$ en (ii) substratum $U_{hh}^{t-12,t}$ van de overige bedrijven in stratum h in maand $t-12$. Geheel analoog kan stratum h in maand t ook in twee substrata worden verdeeld: (i) substratum $U_{0h}^{t-12,t}$ van alle bedrijven die zijn geboren in stratum h in de maanden $t-12, \dots, t-1$ en (ii) stratum $U_{hh}^{t-12,t}$ van de ove-

rige bedrijven in maand t . De omvangen van de steekproeven uit $U_{h,H+1}^{t-12,t}$ en $U_{hh}^{t-12,t}$ in maand $t-12$ geven we aan met $n_{h,H+1}^{t-12}$ en n_{hh}^{t-12} . De desbetreffende steekproefgemiddelden van de omzet geven we aan met $\bar{o}_{h,H+1}^{t-12}$ en $\bar{o}_{hh}^{t-12}$. Bij de steekproeven uit $U_{0h}^{t-12,t}$ en $U_{hh}^{t-12,t}$ in maand t geven we deze vier grootheden aan met n_{0h}^t , n_{hh}^t , $\bar{o}_{0h}^t$ en $\bar{o}_{hh}^t$. Verder veronderstellen we dat sterften zich niet voordoen bij geboorten en dat geboorten na hun eerste maand in de populatie voor de rest van de studieperiode niet meer zijn betrokken bij de maandelijkse updates. Dit heeft het voordeel dat n_{0h}^t en n_{hh}^t vast zijn. In paragraaf 3.5 zullen we zien dat een aldus geïntroduceerde fout doorgaans gering is.

We kunnen $\bar{o}_h^{t-12}$ en $\bar{o}_h^t$ nu schrijven als

$$\begin{aligned}\bar{o}_h^{t-12} &= \frac{n_{hh}^{t-12}}{n_h^{t-12}} \bar{o}_{hh}^{t-12} + \frac{n_{h,H+1}^{t-12}}{n_h^{t-12}} \bar{o}_{h,H+1}^{t-12} \\ \bar{o}_h^t &= \frac{n_{0h}^t}{n_h^t} \bar{o}_{0h}^t + \frac{n_{hh}^t}{n_h^t} \bar{o}_{hh}^t.\end{aligned}$$

Vervolgens kan de covariantie in de laatste regel van (2.4) worden herschreven als

$$\text{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t) = \text{cov}\left(\sum_{g \in \{h, H+1\}} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}, \sum_{k \in \{0, h\}} \frac{n_{kh}^t}{n_h^t} \bar{o}_{kh}^t\right) \quad (2.9)$$

$$= \frac{1}{n_h^{t-12} n_h^t} \text{cov}(n_{hh}^{t-12} \bar{o}_{hh}^{t-12}, n_{hh}^t \bar{o}_{hh}^t). \quad (2.10)$$

In (2.10) hebben we gebruikt dat n_{kh}^t vast is en dat $\text{cov}(\bar{o}_{hg}^{t-12}, \bar{o}_{kh}^t) = 0$ voor $g \neq h$ of $k \neq h$ omdat de desbetreffende steekproeven afkomstig zijn uit verschillende substrata.

Om een uitdrukking te vinden voor de covariantieterm in (2.10) gebruiken we de formule voor conditionele covarianties waarbij we conditioneren op $\nu_h = (n_{hh}^{t-12}, n_{hh}^{t-12,t})$

$$\begin{aligned}\text{cov}(n_{hh}^{t-12} \bar{o}_{hh}^{t-12}, n_{hh}^t \bar{o}_{hh}^t) &= \text{cov}\{E(n_{hh}^{t-12} \bar{o}_{hh}^{t-12} \mid \nu_h), E(n_{hh}^t \bar{o}_{hh}^t \mid \nu_h)\} \\ &\quad + E\{\text{cov}(n_{hh}^{t-12} \bar{o}_{hh}^{t-12}, n_{hh}^t \bar{o}_{hh}^t \mid \nu_h)\} \\ &= \bar{o}_{hh}^{t-12} \bar{o}_{hh}^t \text{cov}(n_{hh}^{t-12}, n_{hh}^t) + n_{hh}^t E\{n_{hh}^{t-12} \text{cov}(\bar{o}_{hh}^{t-12}, \bar{o}_{hh}^t \mid \nu_h)\} \\ &= 0 + E(n_{hh}^{t-12,t} - \frac{n_{hh}^{t-12} n_{hh}^t}{N_{hh}^{t-12,t}}) S_{hh}^{t-12,t}.\end{aligned} \quad (2.11)$$

In de laatste regel hebben we opnieuw gebruik gemaakt van (2.12) in de appendix en dat n_{hh}^t vast is. Uit (2.10) volgt dat de covariantie in (2.4) zuiver kan worden geschat met

$$\hat{\text{cov}}(\hat{O}_h^{t-12}, \hat{O}_h^t) = \frac{N_h^{t-12} N_h^t}{n_h^{t-12} n_h^t} n_{hh}^{t-12,t} \left(1 - \frac{n_{hh}^{t-12} n_{hh}^t}{n_{hh}^{t-12,t} N_{hh}^{t-12,t}}\right) s_{hh}^{t-12,t}.$$

Merk op dat de schatter zuiver is omdat $E(s_{hh}^{t-12,t} | \nu_h) = s_{hh}^{t-12,t}$.

2.4 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft panels lijken veel op die van de directe schatter bij een EAZT-steekproef, dat wil zeggen

- de onzekerheidsmarges van de desbetreffende schatters bij een panel;
- de omvang van de non-respons.
- de omvang van de overlap van de panels in de maanden t en $t-12$.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele. Bij een panel kan het ook voorkomen dat iemand na enige tijd niet meer bereid is aan het panel deel te nemen. Dit type uitval bij een panel wordt ook wel aangeduid met de Engelse term *sample attrition* of *panel attrition*.

Een ander belangrijk aspect voor de kwaliteit van een panel is de omvang van de overlap. Indien door non-respons of door veel migratie tussen de strata de overlap tussen de panels in de maanden t en $t-12$ een beperkte omvang heeft, kan de onzekerheidsmarge van een geschatte omzetgroei aanzienlijk oplopen.

2.5 Appendix. Covariantieformules bij overlappende steekproeven

Laat s_{123} een moedersteekproef zijn die bestaat uit drie (disjuncte) EAZT-substeekproeven s_1, s_2 en s_3 (EAZT: *enkelvoudig aselect zonder teruglegging*). Laat de variabele x waargenomen zijn in s_{12} en de variabele y in s_{23} . De bijbehorende steekproefgemiddelden worden genoteerd als respectievelijk $\bar{x}_{12}$ en $\bar{y}_{23}$. De omvang van s_k wordt aangegeven met n_k ($k=1,2,3,12,23$). Definieer verder $\lambda = n_2 / n_{12}$, $\mu = n_2 / n_{23}$ en $f_k = n_k / N$. Er geldt dan dat de covariantie van $\bar{x}_{12}$ en $\bar{y}_{23}$ gelijk is aan

$$\begin{aligned} \text{cov}(\bar{x}_{12}, \bar{y}_{23}) &= \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right) S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right) S_{xy} \\ &= \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right) S_{xy} = \left(\frac{n_2}{n_{12}n_{23}} - \frac{1}{N}\right) S_{xy} \end{aligned} \quad (2.12)$$

$$S_{xy} = \frac{1}{N-1} \sum_{j=1}^N (X_j - \bar{X}_p)(Y_j - \bar{Y}_p).$$

In de context van (2.5) heeft s_{12} betrekking op de steekproef in maand $t-12$ en s_{23} op die in maand t , terwijl s_2 betrekking heeft op de overlap van beide steekproeven. N is het aantal bedrijven die zowel in maand $t-12$ als maand t tot de doelpopulatie behoorden. Het bewijs van (2.12) is als volgt

$$\begin{aligned}
\text{cov}(\bar{x}_{12}, \bar{y}_{23}) &= \text{cov}\{(1-\lambda)\bar{x}_1 + \lambda\bar{x}_2, \mu\bar{y}_2 + (1-\mu)\bar{y}_3\} \\
&= (1-\lambda)\text{cov}(\bar{x}_1, \bar{y}_{23}) + \lambda\mu\text{cov}(\bar{x}_2, \bar{y}_2) + \lambda(1-\mu)\text{cov}(\bar{x}_2, \bar{y}_3) \\
&= -(1-\lambda)\frac{S_{xy}}{N} + \lambda\mu\left(\frac{1}{n_2} - \frac{1}{N}\right)S_{xy} - \lambda(1-\mu)\frac{S_{xy}}{N} \\
&= \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}.
\end{aligned}$$

In de derde regel hebben we gebruikt dat $\text{cov}(\bar{x}_2, \bar{y}_3) = -S_{xy}/N$ voor twee steekproefgemiddelden uit twee onderling *disjuncte* EAZT-steekproeven ongeacht hun omvang. Dit volgt uit

$$\begin{aligned}
\text{cov}(\bar{x}_2, \bar{y}_3) &= E\{\text{cov}(\bar{x}_2, \bar{y}_3 | s_2)\} + \text{cov}\{E(\bar{x}_2 | s_2), E(\bar{y}_3 | s_2)\} \\
&= 0 + \text{cov}\left\{\bar{x}_2, \frac{\bar{Y}_p - f_2\bar{y}_2}{1-f_2}\right\} \\
&= -\frac{f_2}{1-f_2}\text{cov}(\bar{x}_2, \bar{y}_2) = -\frac{f_2}{1-f_2}(1-f_2)\frac{S_{xy}}{n_2} = -\frac{S_{xy}}{N}.
\end{aligned}$$

3. Panels met migratie tussen de strata

3.1 Korte beschrijving

Evenals in het vorige hoofdstuk gaan we in dit hoofdstuk uit van een populatie van bedrijven die is opgedeeld in strata naar SBI-code en grootteklasse. Het belangrijkste verschil met het vorige hoofdstuk is dat we nu rekening gaan houden met het verschijnsel dat bedrijven kunnen verhuizen van de éne grootteklasse naar de andere als gevolg van een verandering in het aantal werkzame personen.

3.2 Toepasbaarheid

Panels uit dit hoofdstuk worden in de praktijk vaak gebruikt voor het berekenen van een bepaald groeipercentage van de omzet voor bedrijven uit een bepaalde SBI-klasse terwijl bedrijven in de beschouwde periode zijn verwisseld van grootteklasse. Hierbij is het vaak ook relevant om een indicatie te hebben voor de onzekerheidsmarges rondom de geschatte groeipercentages.

3.3 Uitgebreide beschrijving

3.3.1 Inleiding

Zoals we al in het vorige hoofdstuk hebben gezien, wordt het panel op verschillende manieren verversd. De maandelijkse update voor geboorten en sterften is hetzelfde als in paragraaf 2.3.6. Daarnaast vinden doorgaans jaarlijks in januari correcties plaats voor bedrijven die qua grootteklasse niet meer in het correcte stratum zitten. Verder wordt ieder jaar in januari de gebruikelijke 10% procent van de steekproef verversd.

3.3.2 Jaarlijkse correctie en verversing in januari

In januari wordt de steekproef gecorrigeerd waarbij rekening wordt gehouden met de nieuwe stratumindeling van januari en de panelerversing van 10%.

Alle bedrijven die in de steekproef van december zaten en in januari nog steeds bestaan worden opnieuw in strata ingedeeld overeenkomstig hun actuele aantal werkzame personen en SBI-code in januari. Dit heeft tot gevolg dat de aldus ontstane steekproef uit een bepaald stratum kan bestaan uit bedrijven met verschillende insluitkansen omdat bedrijven kunnen zijn overgegaan naar een ander stratum waar de steekproeffractie anders is dan die in het stratum waar ze vandaan zijn gekomen.

Om de strata te corrigeren voor de aanwezigheid van bedrijven met verschillende insluitkansen, introduceren we het substratum $U_{h\ell}^{dec,jan}$ dat bestaat uit bedrijven die in december behoorden tot stratum h en in januari tot stratum ℓ . De omvang van dit stratum wordt aangegeven met $N_{h\ell}^{dec,jan}$ ($h, \ell = 1, \dots, H$). In analogie met de maandelijkse update procedure in deelparagraaf 2.3.6 wordt $s_{h\ell,PRE}^{jan}$ gedefinieerd als $s_{h\ell,PRE}^{jan} = s_h^{dec} \cap U_{h\ell}^{dec,jan}$. Zij $n_{h\ell,PRE}^{jan}$ de omvang van $s_{h\ell,PRE}^{jan}$. Aangezien de vereiste omvang van $s_{h\ell,RST}^{dec,jan}$ uit $U_{h\ell}^{dec,jan}$ in januari zou moeten zijn $n_{h\ell,RST}^{dec,jan} = f_\ell N_{h\ell}^{dec,jan}$, wordt de jaarlijkse verversing van de steekproef $s_{h\ell,PRE}^{jan}$ in januari als volgt uitgevoerd.

Indien $n_{h\ell,PRE}^{jan} > n_{h\ell,RST}^{dec,jan}$, verwijder dan op aselechte wijze dit verschil uit $s_{h\ell,PRE}^{jan}$. Vervolgens wordt 10% van de overgebleven bedrijven in de resterende steekproef vervangen door bedrijven uit $U_{h\ell}^{dec,jan} \setminus s_{h\ell,PRE}^{jan}$. Hierbij is aangenomen dat deze verzameling voldoende bedrijven bevat. Mocht dat niet zo zijn, dan worden er slechts $N_{h\ell}^{dec,jan} - n_{h\ell,PRE}^{jan}$ bedrijven vervangen in de steekproef. Verder wordt er in geval dat $n_{h\ell,PRE}^{jan} \leq n_{h\ell,RST}^{dec,jan}$ het desbetreffende verschil bijgetrokken uit $U_{h\ell}^{dec,jan} \setminus s_{h\ell,PRE}^{jan}$. Vervolgens vindt er in dit geval nog een additionele vervanging plaats van $n_{h\ell,PRE}^{jan} - .9n_{h\ell,RST}^{dec,jan}$ bedrijven in $s_{h\ell,RST}^{dec,jan}$ wanneer (i) dit verschil positief is en (ii) er nog voldoende nieuwe bedrijven beschikbaar zijn. Deze procedure wordt uitgevoerd voor alle substrata $h\ell$, inclusief $h = \ell$. Ten slotte is evenals bij de maandelijkse update procedure het aantal bedrijven dat in januari moet worden getrokken uit het substratum $U_{0\ell}^{dec,jan}$ van nieuwe geboorten in stratum ℓ , gelijk aan $n_{0\ell}^{dec,jan} = f_\ell N_{0\ell}^{dec,jan}$.

Opmerking. Bovenstaande procedure kan ook worden gevolgd wanneer de stratumgrenzen van de grootteklassen in januari worden gewijzigd of wanneer de steekproeffracties in strata veranderen.

3.3.3 De covariantieterm

Ook al veranderen bedrijven van grootteklasse in januari, de maandelijkse omzettalen kunnen op dezelfde manier worden geschat als is beschreven in hoofdstuk 2. Dit geldt ook voor de desbetreffende varianties en onzekerheidsmarges. Daarom zullen we in dit hoofdstuk alleen ingaan op de covariantieterm $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ en de schatting daarvan.

Om rekening te houden met de migratie van bedrijven tussen grootteklassen schrijven we de covariantie nu op een iets andere manier dan (2.4) en wel als volgt

$$\begin{aligned}\text{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \text{cov}\left(\sum_{h=1}^H N_h^{t-12} \bar{o}_h^{t-12}, \sum_{\ell=1}^H N_\ell^t \bar{o}_\ell^t\right) \\ &= \sum_{h=1}^H \sum_{\ell=1}^H N_h^{t-12} N_\ell^t \text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t).\end{aligned}\quad (3.1)$$

Om rekening te houden met de migratie van bedrijven tussen strata, definiëren we de volgende grootheden:

$N_{h\ell}^{t-12,t}$: omvang van substratum $U_{h\ell}^{t-12,t}$ ofwel het aantal bedrijven dat in maand $t-12$ behoorde tot stratum h en in maand t tot stratum ℓ ($h, \ell = 0, 1, \dots, H+1$);

$\bar{O}_{h\ell}^{t-m}$: het substratumgemiddelde van de omzet in $U_{h\ell}^{t-12,t}$ in maand $t-m$ ($m = 0, 12$);

$n_{h\ell}^{t-m}$: omvang van steekproef $s_{h\ell}^{t-m}$, i.e., de feitelijke steekproef uit $U_{h\ell}^{t-12,t}$ in maand $t-m$ ($0 \leq m \leq 12$);

$\bar{o}_{h\ell}^{t-m}$: het steekproefgemiddelde van de omzet in $s_{h\ell}^{t-m}$ ($m = 0, 12$);

$n_{h\ell}^{t-12,t}$: aantal bedrijven in de overlap $s_{h\ell}^{t-12,t} \cap s_{h\ell}^{t-12} \ll s_{h\ell}^t$;

$\bar{o}_{h\ell,OLP}^{t-m}$: het steekproefgemiddelde van de omzet in de overlap $s_{h\ell}^{t-12,t}$ in maand $t-m$ ($m = 0, 12$).

Evenals in het vorige hoofdstuk wordt stratum 0 gereserveerd voor de geboorten in de maanden $t-12, \dots, t-1$ en stratum $H+1$ voor de sterften in die periode. Nu kunnen we $\bar{o}_h^{t-12}$ en $\bar{o}_\ell^t$ schrijven als

$$\bar{o}_h^{t-12} = \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}$$

$$\bar{o}_\ell^t = \sum_{k=0}^H \frac{n_{k\ell}^t}{n_\ell^t} \bar{o}_{k\ell}^t.$$

Vervolgens kunnen de covarianties in (3.1) worden herschreven als

$$\text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) = \text{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}, \sum_{k=0}^H \frac{n_{k\ell}^t}{n_\ell^t} \bar{o}_{k\ell}^t\right) \quad (3.2)$$

$$= \frac{1}{n_h^{t-12} n_\ell^t} \text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) \quad (1 \leq h, \ell \leq H). \quad (3.3)$$

In de tweede regel hebben we gebruikt dat

$$\text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{k\ell}^t \bar{o}_{k\ell}^t) = 0 \quad (3.4)$$

voor $k \neq h$. Verder geldt er $\text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) = 0$ voor $g \neq \ell$ ($g=1, \dots, H+1$) omdat

$$\begin{aligned} \text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) &= E \text{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | n_{hg}^{t-12}, n_{h\ell}^t) \\ &\quad + \text{cov}\{E(n_{hg}^{t-12} \bar{o}_{hg}^{t-12} | n_{hg}^{t-12}, n_{h\ell}^t), E(n_{h\ell}^t \bar{o}_{h\ell}^t | n_{hg}^{t-12}, n_{h\ell}^t)\} \\ &= 0 + \bar{O}_{hg}^{t-12} \bar{O}_{h\ell}^t \text{cov}(n_{hg}^{t-12}, n_{h\ell}^t) = 0. \end{aligned} \quad (3.5)$$

In de laatste regel is gebruik gemaakt van

$$\text{cov}(n_{hg}^{t-12}, n_{h\ell}^t) = 0 \quad (g=1, \dots, H+1). \quad (3.6)$$

Voor een rechtvaardiging van (3.6) en de daarbij gemaakte aannames zie de appendix (paragraaf 3.5). In de appendix leiden we bovendien een alternatieve schattingsmethode af voor het geval dat dit type covariantie niet verwaarloosbaar is. Op dezelfde manier als (3.5) kan de covariantie in (3.3) worden herschreven als

$$\begin{aligned} \text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t) &= E\{\text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | \nu_{h\ell})\} \\ &\quad + \text{cov}\{E(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12} | \nu_{h\ell}), E(n_{h\ell}^t \bar{o}_{h\ell}^t | \nu_{h\ell})\}. \end{aligned} \quad (3.7)$$

waar $\nu_{h\ell} = (n_{h\ell}^{t-12}, n_{h\ell}^{t-12,t}, n_{h\ell}^t)$. Voor de eerste component in het rechterlid geldt er

$$\begin{aligned} E\{\text{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^t \bar{o}_{h\ell}^t | \nu_{h\ell})\} &= E\{n_{h\ell}^{t-12} n_{h\ell}^t \text{cov}(\bar{o}_{h\ell}^{t-12}, \bar{o}_{h\ell}^t | \nu_{h\ell})\} \\ &= E\{n_{h\ell}^{t-12} n_{h\ell}^t \left(\frac{n_{h\ell}^{t-12,t} / n_{h\ell}^{t-12}}{n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t}\}. \end{aligned} \quad (3.8)$$

In de laatste regel hebben we weer (2.12) gebruikt. Verder is $S_{h\ell}^{t-12,t}$ gedefinieerd als

$$S_{h\ell}^{t-12,t} = \frac{1}{N_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{N_{h\ell}^{t-12,t}} (O_{h\ell i}^{t-12} - \bar{O}_{h\ell}^{t-12})(O_{h\ell i}^t - \bar{O}_{h\ell}^t).$$

De tweede component in het rechterlid van (3.7) is wegens (3.6) gelijk aan

$$\bar{O}_{h\ell}^{t-12} \bar{O}_{h\ell}^t \text{cov}(n_{h\ell}^{t-12}, n_{h\ell}^t) = 0.$$

Uit (3.3), (3.7) en (3.8) volgt

$$\text{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) = E\left\{ \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_h^{t-12} n_\ell^t} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (3.9)$$

3.3.4 Schatting van de covariantieterm

Uitdrukking (3.9) kan worden geschat uit de overlap $s_{h\ell}^{t-12,t}$ met

$$\begin{aligned}\widehat{\text{cov}}(\bar{o}_h^{t-12}, \bar{o}_\ell^t) &= \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_h^{t-12} n_\ell^t} \left(\frac{n_{h\ell}^{t-12, t}}{n_{h\ell}^{t-12} n_{h\ell}^t} - \frac{1}{N_{h\ell}^{t-12, t}} \right) \hat{S}_{h\ell, OLP}^{t-12, t} \\ \hat{S}_{h\ell, OLP}^{t-12, t} &= \frac{1}{n_{h\ell}^{t-12, t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12, t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell, OLP}^{t-12}) (o_{h\ell i}^t - \bar{o}_{h\ell, OLP}^t).\end{aligned}\tag{3.10}$$

Merk op dat (3.10) een zuivere schatter is voor (3.9) wegens

$$E(\hat{S}_{h\ell, OLP}^{t-12, t} | \nu_{h\ell}) = S_{h\ell}^{t-12, t}.$$

Een nadeel van schatter $\hat{S}_{h\ell, OLP}^{t-12, t}$ in (3.10) is evenwel dat het kan leiden tot een negatieve uitkomst van de variantieschatter

$$\begin{aligned}\widehat{\text{var}}(\hat{O}^t - G^{t, t-12} \hat{O}^{t-12}) &= \widehat{\text{var}}(\hat{O}^t) + (\hat{G}^{t, t-12})^2 \widehat{\text{var}}(\hat{O}^{t-12}) \\ &\quad - 2\hat{G}^{t, t-12} \widehat{\text{cov}}(\hat{O}^t, \hat{O}^{t-12}).\end{aligned}\tag{3.11}$$

Knottnerus en Van Delden (2006) stellen een verfijning van (3.10) voor. Definieer de standaarddeviaties

$$\hat{S}_{h\ell}^{t-m} = \sqrt{\frac{1}{n_{h\ell}^{t-m} - 1} \sum_{i=1}^{n_{h\ell}^{t-m}} (o_{h\ell i}^{t-m} - \bar{o}_{h\ell}^{t-m})^2} \quad (m = 0, 12).$$

Dan wordt de nieuwe, gemodificeerde schatter voor $S_{h\ell}^{t-12, t}$

$$\hat{S}_{h\ell}^{t-12, t} = \hat{\rho}_{h\ell, OLP}^{t-12, t} \hat{S}_{h\ell}^{t-12} \hat{S}_{h\ell}^t, \tag{3.12}$$

waar $\hat{\rho}_{h\ell}^{t-12, t}$ de correlatie aangeeft tussen de variabelen o^t en o^{t-12} in $U_{h\ell}^{t-12, t}$ en $\hat{\rho}_{h\ell, OLP}^{t-12, t}$ haar schatting op basis van $s_{h\ell}^{t-12, t}$. Overeenkomstig (3.10) en (3.12) kan de covariantie in (3.1) worden geschat met

$$\widehat{\text{cov}}(\hat{O}^{t-12}, \hat{O}^t) = \sum_{h=1}^H \sum_{\ell=1}^H \frac{N_h^{t-12} N_\ell^t}{n_h^{t-12} n_\ell^t} n_{h\ell}^{t-12, t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12, t} N_{h\ell}^{t-12, t}} \right) \hat{S}_{h\ell}^{t-12, t}. \tag{3.13}$$

Voor de zo gedefinieerde schatting $\hat{\rho}_{h\ell, OLP}^{t-12, t}$ geldt met kans 1 $|\hat{\rho}_{h\ell, OLP}^{t-12, t}| \leq 1$ terwijl gebruik van (3.10) impliciet kan leiden tot een geschatte correlatie die groter dan 1 is en aldus tot een mogelijk negatieve uitkomst van (3.11). Voor de volledigheid zij overigens vermeld dat in bijzondere omstandigheden ook gebruik van (3.12) tot een negatieve uitkomst van (3.11) kan leiden.

Bij het gebruik van (3.12) kan zich een speciaal probleem voordoen wanneer $n_{h\ell}^t = 1$ of $n_{h\ell}^{t-12} = 1$. Om de benodigde varianties te berekenen is er allereerst de optie de steekproefvariantie te lenen van een verwant substratum of van hetzelfde substratum in een eerdere maand. Ook kan men een variantie imputeren wanneer uit de data blijkt dat er een relatie is van de vorm $S_{h\ell}^2 \approx \sigma^2 \bar{O}_{h\ell}^\beta$; zie Särndal et al. (1992, blz.

461). Bovendien kan de desbetreffende covariantieterm worden verwaarloosd wanneer zijn verwachte bijdrage aan de totale variantie gering is. Dit is vaak het geval wanneer het gaat om strata h en ℓ met relatief kleine steekproef fracties en derhalve ook met relatief kleine varianties in vergelijking met de grotere bedrijven uit de strata met de grotere steekproef fracties. Soortgelijke opmerkingen zijn van toepassing op de geïmputeerde waarden voor $\rho_{h\ell}^{t-12,t}$ wanneer $n_{h\ell}^{t-12,t} \leq 2$ en $n_{h\ell}^{t-m} \geq 2$ ($m=0,12$). Aangezien de $\rho_{h\ell}^{t-12,t}$ vaak hoog zijn, lijkt dit een effectieve aanpak. Verder zij opgemerkt dat wanneer $n_{h\ell}^{t-m} = 0$ ($m=0,12$), de desbetreffende covariantieterm in (3.13) kan worden verwaarloosd zonder de zuiverheid van (3.13) aan te tasten mits de overige $S_{h\ell}^{t-12,t}$ zuiver kunnen worden geschat. Onder deze aanname kan een dergelijke term met $n_{h\ell}^{t-m} = 0$ ($m=0,12$) worden verwaarloosd omdat de verwachting

$$n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right) \hat{S}_{h\ell}^{t-12,t} \quad (3.14)$$

uit (3.10) gelijk is aan

$$E_{\nu} E_{\hat{S}} \left\{ n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right) \hat{S}_{h\ell}^{t-12,t} \middle| \nu_{h\ell} \right\} = E \left\{ n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^t}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right) S_{h\ell}^{t-12,t} \right\}$$

en de verwachting van het rechterlid is de parameter die moet worden geschat. Bovendien wanneer $n_{h\ell}^{t-m} = 0$ ($m=0,12$), geldt er ook $n_{h\ell}^{t-12,t} = 0$ en derhalve is de uitkomst van (3.14) nul zodat de schatter $\hat{S}_{h\ell}^{t-12,t}$ van $S_{h\ell}^{t-12,t}$ irrelevant is. Daarom tast verwaarlozing van dergelijke termen met $n_{h\ell}^{t-m} = 0$ ($m=0,12$) de zuiverheid van de schatters (3.13) en (3.14) niet aan, mits de overige $S_{h\ell}^{t-12,t}$ zuiver zijn geschat.

3.4 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft panels met migratie van bedrijven tussen strata/grootteklassen zijn:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons;
- de omvang van de overlap van de panels in de maanden t en $t-12$.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele.

Een ander belangrijk aspect voor de kwaliteit van een panel is de omvang van de overlap. Indien door non-respons of door veel migratie tussen de strata de overlap tussen de panels in de maanden t en $t-12$ een beperkte omvang heeft, kan de onzekerheidsmarge van een geschatte omzetgroei aanzienlijk oplopen.

3.5 Appendix. Rechtvaardiging van (3.6)

Eerst beschouwen we het eenvoudige geval met strata zonder geboorten en sterften. Afgezien van de jaarlijkse update in januari zijn er nu geen maandelijkse updates in de overige maanden. Derhalve geldt er dat $n_{h\ell}^t = n_{h\ell, RST}^{dec, jan}$ vast is waaruit (3.6) volgt. Deze situatie is bijvoorbeeld van toepassing op de supermarkten die door de jaren heen een redelijk stabiele populatie vormen.

Wanneer er wel geboorten en sterften voorkomen in een populatie, schrijf $n_{h\ell}^t$ dan als

$$n_{h\ell}^t = n_{\ell}^t - n_{0\ell}^{t-12,t} - \sum_{k \neq h}^H n_{k\ell}^t, \quad (3.15)$$

waarbij $n_{0\ell}^{t-12,t}$ of kortweg $n_{0\ell}^t$ staat voor het aantal geboorten uit de maanden $t-12$, ..., $t-1$ die vallen in steekproef s_{ℓ}^t van maand t . Omdat de steekproeven onder de nieuwe geboorten na maand $t-12$ onafhankelijk zijn van de n_{hg}^{t-12} , zijn de random variabelen $n_{0\ell}^{t-12,t}$ en n_{hg}^{t-12} ongecorreleerd. Omdat tevens geldt dat $\text{cov}(n_{hg}^{t-12}, n_{k\ell}^t) = 0$ voor $k \neq h$, volgt uit (3.15) dat $\text{cov}(n_{hg}^{t-12}, n_{h\ell}^t) = 0$ ($h = 1, \dots, H$).

In feite is tot nu toe aangenomen dat de verdeling van $n_{k\ell}^t$ ($k \neq h$) goed kan worden beschreven met een hypergeometrische verdeling met parameters $(N_{\ell}^t, N_{k\ell}^{t-12,t}, n_{\ell}^t)$ ongeacht de waarden van de n_{hg}^{t-12} . Eenzelfde opmerking is van toepassing op $n_{0\ell}^{t-12,t}$. Er kan echter worden aangetoond dat in de praktijk deze aannames kunnen leiden tot een geringe tweede-orde fout in de variantieformules. Om wat meer inzicht in deze fout te krijgen maken we de volgende vier vereenvoudigende aannames: (i) geboorten en sterften migreren niet tussen strata, (ii) er zijn geen sterften onder de geboorten, (iii) na hun eerste maand in de populatie doen geboorten niet meer mee bij de updates gedurende de rest van de studieperiode, inclusief de verversing in januari en (iv) bij de maandelijkse updates worden geen sterften geselecteerd in of verwijderd uit de steekproef; aldus wordt een derde-orde fout nog steeds veronachtzaamd. Onder deze aannames kijken we nu wat uitgebreider naar de covarian-

tie in (3.2). Analoog aan (3.7) kan deze voor $\ell = h$ worden opgesplitst in twee componenten. De tweede component, zeg $C_{hh,sec}$, kan worden geschreven als

$$\begin{aligned} C_{hh,sec} &\equiv \frac{1}{n_h^{t-12} n_h^t} \text{cov}\{E(\sum_{g=1}^{H+1} n_{hg}^{t-12} \bar{o}_{hg}^{t-12} | \nu_h), E(\sum_{k=0}^H n_{kh}^t \bar{o}_{kh}^t | \nu_h)\} \\ &= \frac{1}{n_h^{t-12} n_h^t} \sum_{g=1}^{H+1} \sum_{k=1}^H \bar{o}_{hg}^{t-12} \bar{o}_{kh}^t \text{cov}(n_{hg}^{t-12}, n_{kh}^t) \\ \nu_h &= (n_{h,1}^{t-12}, \dots, n_{h,H+1}^{t-12}, n_{1h}^t, \dots, n_{Hh}^t). \end{aligned} \quad (3.16)$$

Merk op dat onder de bovengenoemde veronderstellingen $C_{h\ell,sec} = 0$ voor $\ell \neq h$. Alvorens de covarianties in (3.16) te schatten bekijken we eerst de formule voor de conditionele verwachting van y gegeven $x = x_0$ wanneer y en x een bivariate normale verdeling volgen. Dat wil zeggen, in standaard notatie,

$$E(y|x_0) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2} (x_0 - \mu_x).$$

Voor een gegeven verandering Δx_0 in x is de conditionele verwachting van de verandering in y gelijk aan $E(\Delta y | \Delta x_0) = \sigma_{yx} \Delta x_0 / \sigma_x^2$ ofwel

$$\sigma_{yx} = \frac{E(\Delta y | \Delta x_0)}{\Delta x_0} \sigma_x^2. \quad (3.17)$$

Dus om onder normaliteit bijvoorbeeld $\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t)$ in (3.16) te schatten is het voldoende na te gaan wat het verwachte effect is op $y = n_{kh}^t$ dat wordt veroorzaakt door een verandering in de toekomstige sterften $x = n_{h,H+1}^{t-12}$ in s_h^{t-12} . Zij $\Delta n_{h,H+1}^{t-12}$ een additionele (positieve) verandering in het aantal doden in s_h^{t-12} . Definieer $p_{h,H+1}^{jan,t}$ als $p_{h,H+1}^{jan,t} = N_{h,H+1}^{jan,t} / N_{h,H+1}^{t-12}$ waar $N_{h,H+1}^{jan,t}$ staat voor het aantal sterften in stratum h tussen januari en maand t . Evenzo $p_{hg}^{t-12} = N_{hg}^{t-12,t} / N_h^{t-12}$ ($g = 1, \dots, H+1$). Op basis van aanname (iv) kan nu het verwachte aantal additionele sterften in de steekproef van januari voor de verversing worden geschat op $p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12}$. Vervolgens kan het verwachte aantal additionele sterften in de steekproef na de verversing worden geschat met

$$\begin{aligned} \gamma_{red}^{jan} p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12} \\ \gamma_{red}^{jan} = (0.9 - f_h) / (1 - f_h), \end{aligned} \quad (3.18)$$

waar γ_{red}^{jan} staat voor de reductiefactor ten gevolge van de verversing in januari. Voor de afleiding van (3.18) zie het slot van deze appendix.

De desbetreffende maandelijkse updates in stratum h tussen januari en maand t dankzij deze additionele sterften in de steekproef suggereren de volgende schatting van de verwachte toename van instromende n_{kh}^t uit stratum k ($k \neq h$) in de steekproef van maand t

$$\begin{aligned} E(\Delta n_{kh}^t | \Delta n_{h,H+1}^{t-12}) &= \gamma_{red}^{jan} p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12} p_{kh}^t \\ p_{kh}^t &= N_{kh}^{t-12,t} / (N_h^t - N_{0h}^t). \end{aligned} \quad (3.19)$$

Zoals we hebben gezien in deelparagraaf 2.3.6, is er alleen een update in maand s wanneer $d_h^{s-1} \neq f_h D_h^{s-1}$, waarbij D_h^s (d_h^s) staat voor het aantal sterften in U_h^s (s_h^s) en is $n_{kh}^t = f_h N_{kh}^{t-12,t}$ vast wanneer $N_{h,H+1}^{jan,t} = 0$ ($k \neq h$). Merk verder op dat geboorten zijn uitgesloten in de definitie in (3.19) wegens aanname (iii).

Definieer vervolgens voor $m = 0, 12$

$$\begin{aligned} \bar{O}_h^{t-m} &= \frac{1}{N_h^{t-m}} \sum_{i=1}^{N_h^{t-m}} O_{hi}^{t-m}; \quad (S_h^{t-m})^2 = \frac{1}{N_h^{t-m} - 1} \sum_{i=1}^{N_h^{t-m}} (O_{hi}^{t-m} - \bar{O}_h^{t-m})^2 \\ p_{h,\leq H}^{t-12} &= 1 - p_{h,H+1}^{t-12}; \quad p_{in,h}^t = 1 - p_{hh}^t \\ \bar{O}_{h,\leq H}^{t-12} &= \sum_{g=1}^H \frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \bar{O}_{hg}^{t-12}; \quad \bar{O}_{in,h}^t = \sum_{\substack{k=1 \\ k \neq h}}^H \frac{p_{kh}^t}{p_{in,h}^t} \bar{O}_{kh}^t. \end{aligned}$$

Met behulp van (3.17) en (3.19) krijgen we voor $k \neq h$ de volgende benadering voor de covariantie

$$\begin{aligned} \text{acov}(n_{h,H+1}^{t-12}, n_{kh}^t) &= \frac{E(\Delta n_{kh}^t | \Delta n_{h,H+1}^{t-12})}{\Delta n_{h,H+1}^{t-12}} \text{var}(n_{h,H+1}^{t-12}) \\ &\approx \gamma_{red}^{jan} p_{h,H+1}^{jan,t} p_{kh}^t n_h^{t-12} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h) \\ &= n_h^{t-12} p_{kh}^t A_h / p_{in,h}^t \\ A_h &= \gamma_{red}^{jan} p_{in,h}^t p_{h,H+1}^{jan,t} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h), \end{aligned} \quad (3.20)$$

waarbij ter vereenvoudiging in de tweede regel de term $N_h^{t-12} / (N_h^{t-12} - 1)$ hebben weggelaten. Omdat n_h^{t-12} vast is, geldt er

$$\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) = -\text{cov}(n_{h1}^{t-12} + \dots + n_{hH}^{t-12}, n_{kh}^t).$$

In analogie met de multihypergeometrische verdeling kunnen we voor $1 \leq g \leq H$ en $k \neq h$ de volgende relatie gebruiken voor een benadering van $\text{cov}(n_{hg}^{t-12}, n_{kh}^t)$

$$\text{acov}(n_{hg}^{t-12}, n_{kh}^t) = -\frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \text{acov}(n_{h,H+1}^{t-12}, n_{kh}^t) = -n_h^{t-12} \frac{p_{hg}^{t-12}}{p_{h,\leq H}^{t-12}} \frac{p_{kh}^t}{p_{in,h}^t} A_h, \quad (3.21)$$

waarbij ook gebruik is gemaakt van (3.20). Merk in aanvulling hierop op dat

$$\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) = - \text{cov}(n_{h,\mathcal{E}H}^{t-12}, n_{kh}^t) = - \sum_{g \in \mathcal{E}H} \sum_{i \in \mathcal{U}_{hg}^{t-12,t}} \text{cov}(d_{hgi}^{t-12}, n_{kh}^t)$$

$$d_{hgi}^{t-12} = \begin{cases} 1 & \text{indien bedrijf } i \text{ in } U_{hg}^{t-12,t} \text{ is geselecteerd in steekproef } s_h^{t-12} \\ 0 & \text{anders.} \end{cases}$$

Om redenen van symmetrie geldt er derhalve

$$\text{cov}(d_{hgi}^{t-12}, n_{kh}^t) = - \text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t) / N_{h,\mathcal{E}H}^{t-12,t},$$

waaruit (3.21) volgt ($1 \notin g \in \mathcal{E}H$). Evenzo krijgen we voor $k = h$ op basis van (3.20) en (3.21)

$$\begin{aligned} \text{acov}(n_{h,H+1}^{t-12}, n_{hh}^t) &= - n_h^{t-12} A_h \\ \text{acov}(n_{hg}^{t-12}, n_{hh}^t) &= n_h^{t-12} p_{hg}^{t-12} A_h / p_{h,\mathcal{E}H}^t \quad (1 \notin g \in \mathcal{E}H). \end{aligned} \quad (3.22)$$

Substitutie van (3.20)-(3.22) in (3.16) geeft vervolgens

$$C_{hh,\text{sec}} = A_h (\bar{O}_{h,H+1}^{t-12} - \bar{O}_{h,\mathcal{E}H}^{t-12}) (\bar{O}_{in,h}^t - \bar{O}_{hh}^t) / n_h^t. \quad (3.23)$$

Onder de aanname dat de twee termen tussen haakjes in (3.23) absoluut kleiner zijn dan S_h^t , volgt uit (3.23) dat

$$|C_{hh,\text{sec}}| \leq \frac{g_{\text{red}}^{jan} p_{h,H+1}^{jan,t} p_{in,h}^t p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12}) (1 - f_h)}{n_h^t} (S_h^t)^2.$$

Hieruit volgt dat wanneer $p_{in,h}^t, p_{h,H+1}^{t-12} \leq 0.1$, we mogen concluderen dat onder de eerder gemaakte veronderstellingen de bijdrage van de tweede covariantie component minder is dan 1% van $\text{var}(\bar{o}_h^t)$. Met andere woorden, (3.6) kan worden gebruikt zonder al teveel nadelige gevolgen voor uitkomsten. Wanneer in de praktijk $C_{hh,\text{sec}}$ niet verwaarloosbaar is, kan die uit de data worden geschat met

$$\hat{C}_{hh,\text{sec}} = A_h (\bar{o}_{h,H+1}^{t-12} - \bar{o}_{h,\mathcal{E}H}^{t-12}) (\bar{o}_{in,h}^t - \bar{o}_{hh}^t) / n_h^t. \quad (3.24)$$

We sluiten deze appendix af met de afleiding van (3.18). In januari is het verwachte aantal additionele sterften dat gedurende de verversing in de steekproef blijft $0.9 p_{h,H+1}^{jan,t} D n_{h,H+1}^{t-12}$. Een schatting voor het aantal sterften buiten de steekproef net voor de verversing is $N_{h,H+1}^{jan,t} - n_{h,H+1}^{jan} - p_{h,H+1}^{jan,t} D n_{h,H+1}^{t-12}$. Derhalve kan het aantal nieuwe sterften in de steekproef ten gevolge van de verversingen in alle substrata $U_{hg}^{t-12,t}$ ($g = 1, \dots, H$) in januari worden geschat op

$$0.1 (n_h^{jan} - n_{0h}^{jan}) \frac{N_{h,H+1}^{jan,t} - n_{h,H+1}^{jan} - p_{h,H+1}^{jan,t} D n_{h,H+1}^{t-12}}{N_h^{jan} - N_{0h}^{t-12,jan} - (n_h^{jan} - n_{0h}^{jan})}.$$

Omdat $n_{0h}^{jan} = f_h N_{0h}^{t-12,jan}$ op basis van de eerdere aannames, krijgen we

$$\gamma_{red}^{jan} = 0.9 - \frac{0.1(n_h^{jan} - n_{0h}^{jan})}{N_h^{jan} - N_{0h}^{t-12,jan} - (n_h^{jan} - n_{0h}^{jan})} = \frac{0.9 - f_h}{1 - f_h}.$$

4. Panels voor het schatten van indicatoren en indices

4.1 Korte beschrijving

Een index kan in feite worden gezien als een ratio van twee populatietotalen die kan worden geschat met de ratioschatter¹ op basis van een EAZT-steekproef uit de populatie of uit de desbetreffende strata. Een andere mogelijkheid om een index te schatten is de Horvitz-Thompson schatter op basis van een probability proportional to size (PPS-) steekproef. Hierbij is de insluitkans van een bedrijf evenredig met het desbetreffende omzetcijfer van dat bedrijf. In dit hoofdstuk werken we de desbetreffende formules voor indices nader uit. De gevolgen van een splitsing of een fusie van bedrijven worden eveneens besproken. Formules uit dit hoofdstuk zijn ook van toepassing op prijsmutaties en conjunctuurindicatoren die zijn gebaseerd op panels onder bedrijven. Tenzij anders vermeld, zullen we in dit hoofdstuk steeds aannemen dat de doelparameter een gewogen gemiddelde is van prijsmutaties bij een doelgroep van bedrijven waarbij de wegenen gelijk zijn genomen aan de omzetaandelen van de desbetreffende bedrijven. Voor indexmutaties, updates en verversing van panels wordt de lezer verwezen naar hoofdstuk 3.

4.2 Toepasbaarheid

Bij CBS worden panels ook gebruikt om speciaal gewogen populatiegemiddelden te berekenen. Men kan in dit verband denken aan bijvoorbeeld de PPI, DPI, COEN en de Conjunctuurtest. De eerste twee zijn prijsindices en de laatste twee zijn conjunctuurindicatoren.

4.3 Uitgebreide beschrijving

4.3.1 De ratioschatter voor een algemene index

Een samengestelde prijsindex I van prijsmutaties bij N verschillende bedrijven kan voor onze doeleinden ook worden geschreven in de volgende vorm

$$I = \frac{\sum_{h=1}^H \sum_{i=1}^{N_h} O_{hi} I_{hi}}{\sum_{h=1}^H \sum_{i=1}^{N_h} O_{hi}} = \sum_{h=1}^H \sum_{i=1}^{N_h} O_{hi} I_{hi} / O, \quad (O = \sum_{h=1}^H \sum_{i=1}^{N_h} O_{hi}) \quad (4.1)$$

waarbij

¹ Merk op dat met *ratioschatter* in deze context van indexcijfers letterlijk wordt bedoeld de schatter van een ratio en niet de in de literatuur meer gebruikelijke schatter van een populatietotaal op basis van een geschatte ratio, die ook wel met ‘quotiëntschatte’ wordt aangeduid..

N_h : aantal bedrijven in stratum h ($h = 1, \dots, H$; $\sum_{h=1}^H N_h = N$)

I_{hi} : prijsmutatie van bedrijf i in stratum h

O_{hi} : omzet van bedrijf i in stratum h in een bepaalde periode.

Voor schattingsdoeleinden is het handig (4.1) te herschrijven als

$$\begin{aligned} I &= \left[\sum_{h=1}^H O_h \right]^{-1} \sum_{h=1}^H O_h I_h = \sum_{h=1}^H \frac{O_h}{O} I_h & (O &= \sum_{h=1}^H O_h) \\ I_h &= \sum_{i=1}^{N_h} \frac{O_{hi}}{O_h} I_{hi} & (O_h &= \sum_{i=1}^{N_h} O_{hi}). \end{aligned} \quad (4.2)$$

Omdat de omzetcijfers in het algemeen bekend mogen worden verondersteld, is een voor de hand liggende schatter van (4.1) of (4.2)

$$\hat{I} = \sum_{h=1}^H \frac{O_h}{O} \hat{I}_h. \quad (4.3)$$

$\hat{I}_h$ staat voor de standaard ratioschatter van de gemiddelde prijsmutatie in stratum h

$$\hat{I}_h = \frac{\sum_{i=1}^{n_h} o_{hi} I_{hi}}{\sum_{i=1}^{n_h} o_{hi}}, \quad (4.4)$$

waarbij o_{hi} en I_{hi} hier respectievelijk staan voor de omzet en de prijsmutatie van bedrijf i in de EAZT-steekproef van omvang n_h uit stratum h ($i=1, \dots, n_h$). Merk op dat $\hat{I}_h$ kan worden gezien als een standaard ratioschatter voor de ratio I_h in (4.2).

De variantie van $\hat{I}$ in (4.3) is gelijk aan

$$\text{var}(\hat{I}) = \sum_{h=1}^H W_h^2 \text{var}(\hat{I}_h), \quad (W_h = O_h/O) \quad (4.5)$$

waarbij

$$\begin{aligned} \text{var}(\hat{I}_h) &= \frac{1-f_h}{n_h \bar{O}_h^2} S_{he}^2 \\ f_h &= n_h / N_h \\ \bar{O}_h &= O_h / N_h \\ S_{he}^2 &= \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} O_{hi}^2 (I_{hi} - I_h)^2 \end{aligned} \quad (4.6)$$

zie Banning e.a. (2010, p. 76). Combinatie van (4.5) en (4.6) geeft

$$\text{var}(\hat{I}) = \sum_{h=1}^H \frac{W_h^2}{\bar{O}_h^2} (1 - f_h) \frac{S_{he}^2}{n_h}. \quad (4.7)$$

De variantie in (4.7) kan worden geschat met

$$\begin{aligned} \hat{\text{var}}(\hat{I}) &= \sum_{h=1}^H \frac{W_h^2}{\bar{O}_h^2} (1 - f_h) \frac{S_{he}^2}{n_h} \\ s_{he}^2 &= \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} o_{hi}^2 (I_{hi} - \hat{I}_h)^2; \end{aligned}$$

Analoog aan de Neyman-allocatie bij de standaard stratificatieschatter kan worden bewezen dat de variantie in (4.5) minimaal is bij de volgende allocatie van de n_h

$$n_h = \frac{W_h S_{he} / \bar{O}_h}{\sum_{h=1}^H W_h S_{he} / \bar{O}_h} n_h = \frac{N_h S_{he}}{\sum_{h=1}^H N_h S_{he}} n_h. \quad (4.8)$$

Indien kan worden verondersteld dat de I_{hi} goed met het volgende model kunnen worden beschreven $I_{hi} = I_h + e_{hi}$ met $E(e_{hi}) = 0$, $E(e_{hi}^2) = g^2$ en $E(e_{hi} e_{hj}) = 0$ ($i \neq j$), kan $S_{he}^2 / \bar{O}_h^2$ worden vereenvoudigd tot

$$\begin{aligned} \frac{S_{he}^2}{\bar{O}_h^2} &\approx \frac{g^2 \sum_{i=1}^{N_h} O_{hi}^2 / N_h}{\bar{O}_h^2} \approx \frac{g^2 (S_{ho}^2 + \bar{O}_h^2)}{\bar{O}_h^2} = g^2 (1 + CV_{ho}^2) \\ S_{ho}^2 &= \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} (O_{hi} - \bar{O}_h)^2 \\ CV_{ho} &= S_{ho} / \bar{O}_h. \end{aligned}$$

waarbij is aangenomen dat $N_h \approx N_h - 1$. De allocatie in (4.8) gaat aldus over in

$$n_h = \frac{W_h \sqrt{1 + CV_{ho}^2}}{\sum_{h=1}^H W_h \sqrt{1 + CV_{ho}^2}} n_h = \frac{O_h \sqrt{1 + CV_{ho}^2}}{\sum_{h=1}^H O_h \sqrt{1 + CV_{ho}^2}} n_h.$$

Ten slotte, zij nog opgemerkt dat I_h in (4.2) kan worden geschreven als een normaal stratumtotaal $I_h = \sum_{i=1}^{N_h} Y_{hi}$ met $Y_{hi} = O_{hi} I_{hi} / O_h$. Derhalve kan I_h ook worden geschat met de directe schatter

$$\hat{I}_{h,DIR} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{o_{hi} I_{hi}}{O_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{o_{hi} I_{hi}}{O_h}.$$

Onder bovengenoemde modelaannames geldt evenwel dat in een regressie van Y_{hi} op O_{hi} en een constante ($i = 1, \dots, N_h$) de regressiecoëfficiënt b_h bij benadering gelijk is aan $b_h = I_h$. Wegens $b_h > I_h / 2$ verdient de ratioschatter de voorkeur boven de directe schatter; zie Banning e.a. (2010, p. 77). Merk op dat b_h kan worden geschre-

ven als $b_h = \rho_{hyo} S_{hy} / S_{ho}$ waarbij ρ_{hyo} staat voor de correlatiecoëfficiënt tussen de Y_{hi} en de O_{hi} .

4.3.2 De PPS-schatter voor een algemene index

Zoals we hebben gezien in hoofdstuk 5 van Banning e.a. (2010) kan een (samen-
gestelde) prijsindex ook worden geschat op basis van een systematische PPS-
steekproef waarbij de insluitkansen $\pi_i = nW_i$ ($W_i = O_i / O$) evenredig zijn met de
desbetreffende omzetcijfers. Verder worden de bedrijven geacht in een random
volgorde op de desbetreffende lijst te staan. De PPS-schatter van de parameter
 $I = \sum_{i=1}^N W_i I_i$ neemt nu overeenkomstig vergelijking (5.5) van Banning e.a. (2010)
de volgende gedaante aan

$$\begin{aligned} \hat{I}_{PPS} &= \sum_{i=1}^N a_i \frac{W_i I_i}{\pi_i} = \frac{1}{n} \sum_{i=1}^n I_i = \bar{I}_s \\ a_i &= \begin{cases} 1 & \text{als bedrijf } i \text{ in de steekproef zit} \\ 0 & \text{anders,} \end{cases} \end{aligned} \quad (4.9)$$

waarbij $\bar{I}_s$ staat voor het gemiddelde van de waargenomen prijsmutaties in de
steekproef. Om de notatie eenvoudig te houden nemen we in deze deelparagraaf aan
dat de populatie bestaat uit 1 stratum. Mocht de populatie uit meerdere strata be-
staan, dan kan voor de afzonderlijke strata dezelfde aanpak worden gevolgd bij het
schatten van de desbetreffende index I_h .

Onder de aannames dat de volgorde van de populatie-elementen geen rol speelt bij
de PPS-steekproeftrekking en dat de I_i goed kunnen worden beschreven met het
model $I_i = I + \varepsilon_i$ met $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \gamma^2 O_i^\beta$ en $E(\varepsilon_i \varepsilon_j) = 0$ ($i \neq j$), kan als
variantiebenadering van $\hat{I}_{PPS}$ in (4.9) worden gebruikt

$$\begin{aligned} \text{var}(\hat{I}_{PPS}) &= \frac{1}{n^2} \sum_{i=1}^N \pi_i (1 - \pi_i) (I_i - I)^2 \\ &= \frac{1}{n} \sum_{i=1}^N W_i (1 - nW_i) (I_i - I)^2. \end{aligned} \quad (4.10)$$

Onder bovengenoemde modelaannames met $\beta > -1$ kan worden aangetoond dat
voor $N \rightarrow \infty$ $\hat{I}_{PPS}$ een kleinere variantie heeft dan de ratioschatter $\hat{I}_{EATZ}$ op basis
van een EATZ-steekproef ofwel

$$\hat{I}_{EATZ} = \frac{\sum_{i=1}^n o_i I_i}{\sum_{i=1}^n o_i};$$

zie Knottnerus (2011). De variantie in (4.10) kan onder eerder genoemde aannames
asymptotisch zuiver worden geschat met

$$\text{var}(\hat{I}_{PPS}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(1 - \frac{no_i}{O}\right) (I_i - \bar{I}_s)^2. \quad (4.11)$$

4.3.3 Vervanging of aanvulling van de systematische PPS-steekproef

Tot nu toe gingen we uit van een stabiele steekproef die niet verandert in de tijd. Helaas komt het regelmatig voor dat eenheden uit de steekproef verdwijnen als gevolg van non-respons of sterfte bijvoorbeeld. Een manier om een PPS-steekproef aan te vullen is als volgt. Veronderstel dat de (vermoede) uitval in een periode 10% van de steekproef bedraagt. In dat geval zou men vooraf al een bruto steekproef s_b kunnen trekken met een omvang van $n_b = 1,1n$. Door vervolgens uit s_b een EAZT-steekproef van omvang n te trekken heeft men per saldo een PPS-steekproef van de gewenste omvang n . Wanneer er later bedrijven uit het panel wegvallen, kunnen die aldus aselekt uit het restant van s_b worden aangevuld. Deze aanpak kan in de praktijk ook worden toegepast op een EAZT-steekproef.

Een andere aanpak is die van ‘circular systematic PPS sampling’ waarbij de bedrijven achter elkaar liggen op een cirkel in plaats van op een rechte lijn zodat bedrijf N ligt tussen de bedrijven $N-1$ en 1 (de random volgorde hierbij even negerend). Indien een bedrijf uitvalt, kan op eenvoudige wijze het procédé om systematisch te trekken worden voortgezet met dezelfde staplengte L . Wanneer dat nodig is, kan een nieuw random startgetal worden gekozen wanneer de cirkel is doorlopen. Zo’n nieuw random startgetal is bijvoorbeeld nodig wanneer de populatie stabiel is gebleven zonder geboorten en sterften. Na de eerste ronde door de gehele cirkel komt men dan automatisch weer uit bij het bedrijf dat hoort bij het eerste random gekozen getal. Nadeel van deze circulaire aanpak is dat niet helemaal duidelijk is wat de insluitkansen per saldo precies worden wanneer de insluitkansen ongelijk zijn en de cirkel voor de tweede keer doorlopen gaat worden.

Bij geboorten verdient het aanbeveling om die als aparte groep te zien waaruit separaat een kanssteekproef moet worden getrokken. Indien het gaat om een systematische steekproef dient de staplengte L hetzelfde te zijn als in de bijbehorende populatie van reeds bestaande eenheden. Ook kan men op dezelfde manier rekening houden met latere uitvallers in de steekproef uit deze groep.

4.3.4 Fusies van bedrijven

Beschouw een populatie $\{1, 2, \dots, N\}$ van N bedrijven. Definieer π_i als de eerste orde insluitkans van bedrijf i . We zullen allereerst de situatie bekijken waarbij Y_i de waarde van een willekeurige variabele y voor bedrijf i aangeeft ($i=1, \dots, N$). De gemodificeerde Horvitz-Thompson (HT) schatter voor het populatietotaal Y in geval van een fusie van de bedrijven 1 en 2 ziet er dan als volgt uit

$$\hat{Y}_{HT}^{\text{mod}} = (a_1 + a_2) \frac{Y_1 + Y_2}{\pi_1 + \pi_2} + \sum_{i=3}^N a_i \frac{Y_i}{\pi_i}. \quad (4.12)$$

Wegens $E(a_1 + a_2) = \pi_1 + \pi_2$ is (4.12) een zuivere schatter voor Y evenals de standaard HT-schatter in de situatie zonder fusies. We kunnen (4.12) overigens ook schrijven als

$$\begin{aligned}\hat{Y}_{HT}^{\text{mod}} &= \sum_{i=1}^N a_i \frac{Y_i^*}{\pi_i} \\ Y_i^* &= \begin{cases} \pi_i(Y_1 + Y_2)/(\pi_1 + \pi_2) & \text{als } i \leq 2 \\ Y_i & \text{als } i \geq 3. \end{cases}\end{aligned}\quad (4.13)$$

Met andere woorden, $Y_1 + Y_2$ wordt verdeeld over de bedrijven 1 en 2 met een gewichtsverhouding π_1 / π_2 . Formule (4.13) is weer de standaard gedaante van een HT-schatter en kan derhalve als basis dienen voor verdere variantieberekeningen. Bijvoorbeeld bij een enkelvoudige aselechte steekproef zonder teruglegging (EAZT) met $\pi_i = n / N$ ($i=1, \dots, N$) geldt er

$$\text{var}(\hat{Y}_{HT}^{\text{mod}}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \frac{1}{N-1} \sum_{i=1}^N (Y_i^* - \bar{Y})^2.$$

Merk op dat in dit geval $Y_1^* = Y_2^* = (Y_1 + Y_2) / 2$ en dat de populatiegemiddelden gelijk zijn $\bar{Y}^* = \bar{Y}$. Ook wanneer een systematische PPS-steekproef is getrokken, kan (4.13) worden gebruikt voor de variantiebenadering en -schatting zoals besproken in deelparagraaf 4.3.2, mits de elementen in de populatie min of meer in een random volgorde staan.

De hier beschreven aanpak kan worden gevolgd wanneer men bedrijven zoveel mogelijk wil behouden voor het panel. Men kan dan denken aan bijvoorbeeld PPI en DPI. Een alternatieve benadering is om de afzonderlijke gefuseerde bedrijven te zien als sterften en het nieuw ontstane, gefuseerde bedrijf als een geboorte. Deze alternatieve benadering geeft in principe geen systematische vertekening, maar kan wel leiden tot grotere onzekerheidsmarges rondom de geschatte mutaties.

4.3.4.1 Fusies bij een PPS-steekproef

Beschouw heel algemeen een prijsindex $I = \sum_{i=1}^N W_i I_i$ ($\sum_{i=1}^N W_i = 1$) van N bedrijven waarbij W_i staat voor het gewicht/relevantie van bedrijf i . Bedrijven in de integrale waarneming laten we buiten beschouwing. Beschouw nu een PPS-steekproef van omvang n waarbij de insluitkansen π_i evenredig zijn met de gewichten W_i ofwel $\pi_i = nW_i$. Indien er geen fusies zijn, hebben we al eerder gezien dat de standaard HT-schatter van I gelijk is aan

$$\hat{I}_{PPS} = \bar{I}_s = \frac{1}{n} \sum_{i=1}^n I_i.$$

Indien de bedrijven 1 en 2 zijn gefuseerd, is het handiger de bedrijven in de steekproef voortaan te noteren met hun rangnummers $i_1, \dots, i_n$. Voor de drie verschillende situaties die zich kunnen voordoen, ziet de gemodificeerde HT-schatter er als volgt uit. Ten eerste, wanneer $a_1 = a_2 = 0$, is (4.12) gelijk aan de standaard HT-schatter

$$\hat{I}_{PPS}^{\text{mod}} = \frac{1}{n} \sum_{b=1}^n I_{i_b}.$$

Ten tweede, wanneer precies één van de twee gefuseerde bedrijven in de oorspronkelijke steekproef zit ofwel $a_1 + a_2 = 1$, neemt (4.12) de volgende vorm aan

$$\begin{aligned} \hat{I}_{PPS}^{\text{mod}} &= \frac{W_1 I_1 + W_2 I_2}{n(W_1 + W_2)} + \frac{I_{i_2} + \dots + I_{i_n}}{n} \\ &= \frac{I_{12} + I_{i_2} + \dots + I_{i_n}}{n} \quad (I_{12} = \frac{W_1 I_1 + W_2 I_2}{W_1 + W_2}). \end{aligned} \quad (4.14)$$

In de praktijk staat I_{12} voor de prijswaarneming van het nieuwe bedrijf 12 dat is ontstaan uit de fusie van de bedrijven 1 en 2. Ten slotte, wanneer bedrijven 1 en 2 beide in de oorspronkelijke steekproef zitten, gaat (4.12) analoog aan (4.14) over in

$$\hat{I}_{PPS}^{\text{mod}} = (2I_{12} + I_{i_3} + \dots + I_{i_n}) / n.$$

Samengevat is de gemodificeerde HT-schatter van de PPI in geval van een PPS-steekproef gelijk aan

$$\hat{I}_{PPS}^{\text{mod}} = \begin{cases} (I_{i_1} + \dots + I_{i_n}) / n & \text{als } a_1 = a_2 = 0 \\ (I_{12} + I_{i_2} + \dots + I_{i_n}) / n & \text{als } a_1 + a_2 = 1 \\ (2I_{12} + I_{i_3} + \dots + I_{i_n}) / n & \text{als } a_1 = a_2 = 1. \end{cases}$$

De laatste formule laat zien dat uitgaande van een zuivere prijsindex van Laspeyres er sprake kan zijn van een zichtbare trendbreuk in het niveau van de prijsindex ten opzichte van het basisjaar wanneer I_1 en I_2 veel verschillen en n niet erg groot is. In de hypothetische situatie dat bijvoorbeeld $a_1 = 1$ en $a_2 = 0$, kan de nieuwe I_{12} in principe aanzienlijk afwijken van de oude I_1 die in de steekproef werd geobserveerd. Omdat de index wordt berekend op basis van een kettingindex, is van een dergelijke zichtbare trendbreuk in de praktijk geen sprake.

Samengevat kan men stellen dat toepassing van de gemodificeerde HT-schatter op een PPS-steekproef voor de PPI met achteraf een fusie van de bedrijven 1 en 2 inhoudt dat wanneer bedrijf i ($i=1,2$) reeds voor fusie in de steekproef zat, zijn prijsindex I_i na fusie wordt vervangen door de prijsindex I_{12} van het nieuwe, gefuseerde bedrijf. $\hat{I}_{PPS}^{\text{mod}}$ blijft ook na fusie gelijk aan het ongewogen gemiddelde van de aldus waargenomen prijsmutaties ten opzichte van het basisjaar.

4.3.4.2 Fusies bij een EAZT-steekproef

In deze subparagraaf zullen we in het kort nog iets zeggen over een enkelvoudige aselechte steekproef van omvang n waarbij alle bedrijven een gelijke insluitkans ($\pi_i = n/N$) hebben. Net als in de vorige paragraaf nemen we aan dat de bedrijven 1 en 2 (van de populatie) zijn gefuseerd tot het bedrijf 12. Zolang de bedrijven 1 en 2 geen van beide in de steekproef zitten ($a_1 = a_2 = 0$), is de gebruikelijke ratioschatter van de index I gelijk aan

$$\hat{I}_{EAZT} = \frac{\sum_{i=1}^n W_i I_i}{\sum_{i=1}^n W_i}.$$

Wanneer zich de situatie voordoet dat bedrijf 1 wel in de oorspronkelijke steekproef zit en bedrijf 2 niet of omgekeerd ($a_1 + a_2 = 1$), wordt overeenkomstig (4.12) de ratioschatter gelijk aan

$$\begin{aligned} \hat{I}_{EAZT}^{\text{mod}} &= \frac{\frac{W_1 I_1 + W_2 I_2}{2n/N} + \frac{W_{i_2} I_{i_2}}{n/N} + \dots + \frac{W_{i_n} I_{i_n}}{n/N}}{\frac{W_{12}}{2n/N} + \frac{W_{i_2}}{n/N} + \dots + \frac{W_{i_n}}{n/N}} \\ &= \frac{\bar{W}_{12} I_{12} + W_{i_2} I_{i_2} + \dots + W_{i_n} I_{i_n}}{\bar{W}_{12} + W_{i_2} + \dots + W_{i_n}} \quad (\bar{W}_{12} = \frac{W_{12}}{2} = \frac{W_1 + W_2}{2}). \end{aligned}$$

Dit houdt in dat het totale gewicht van het gefuseerde bedrijf moet worden gehalveerd bij de verdere raming van de prijsindex na de fusie. Om dit te voorkomen stellen we de volgende modificatie voor

$$\hat{I}_{EAZT}^{\text{mod}} = \frac{W_i I_{12} + W_{i_2} I_{i_2} + \dots + W_{i_n} I_{i_n}}{W_i + W_{i_2} + \dots + W_{i_n}} \quad \text{met } i = \begin{cases} 1 & \text{als } a_1 = 1 \\ 2 & \text{als } a_2 = 1. \end{cases}$$

Het voordeel is dat de gewichten van de bedrijven in de steekproef na de fusie niet hoeven te worden aangepast, evenmin als bij de PPS-steekproef. Verder blijven de teller en de noemer als schatter zuiver wegens $E(a_1 W_1 + a_2 W_2 | a_1 + a_2 = 1) = \bar{W}_{12}$.

Wanneer de bedrijven 1 en 2 beide in de steekproef zitten ($a_1 = a_2 = 1$), leidt (4.12) tot de volgende ratioschatter

$$\begin{aligned} \hat{I}_{EAZT}^{\text{mod}} &= \frac{(W_1 I_1 + W_2 I_2 + W_{i_3} I_{i_3} + \dots + W_{i_n} I_{i_n}) / (n/N)}{(W_1 + W_2 + W_{i_3} + \dots + W_{i_n}) / (n/N)} \\ &= \frac{(W_1 + W_2) I_{12} + W_{i_3} I_{i_3} + \dots + W_{i_n} I_{i_n}}{W_1 + W_2 + W_{i_3} + \dots + W_{i_n}}. \end{aligned}$$

Met andere woorden, het gefuseerde bedrijf 12 krijgt net als de andere bedrijven zijn bijbehorende gewicht ($W_1 + W_2$) overeenkomstig het omzetaandeel van het zojuist gefuseerde bedrijf.

Samengevat, betekent dit dat in geval van een steekproef voor de PPI met gelijke insluitkansen het gebruik van de gemodificeerde HT-schatter bij een fusie van de bedrijven 1 en 2 inhoudt dat wanneer één van de twee bedrijven, zeg bedrijf i ($i=1,2$), reeds voor fusie in de steekproef zat, de prijsmutatie I_i na fusie dient te worden vervangen door de prijsmutatie I_{12} van het nieuwe, gefuseerde bedrijf, terwijl zijn gewicht w_i hetzelfde blijft. Indien beide bedrijven reeds voor de fusie in de steekproef zaten, komt dit er op neer dat I_{12} in de steekproef een gewicht $w_1 + w_2$ krijgt.

4.3.5 Splitsingen van bedrijven

In deze subparagraaf gaan we kort in op de gevolgen voor de gewichten van de bedrijfswaarnemingen wanneer een bedrijf in de steekproef wordt opgesplitst in twee bedrijven. Allereerst bekijken we de situatie waarin een indexcijfer wordt geschat. We nemen aan dat een geschatte index voor een bepaalde populatie of stratum kan worden geschreven als

$$\hat{I}_s = \sum_{i=1}^n w_i I_i = w_1 I_1 + \sum_{i=2}^n w_i I_i. \quad (4.15)$$

Stel nu dat bedrijf 1 in de steekproef wordt opgesplitst in twee bedrijven 1a en 1b. Verder nemen we aan dat I_1 kan worden geschreven als een gewogen gemiddelde van I_{1a} en I_{1b} oftewel

$$I_1 = w_{1a} I_{1a} + w_{1b} I_{1b}. \quad (4.16)$$

Substitutie van (4.16) in (4.15) geeft

$$\begin{aligned} \hat{I}_s &= \sum_{i=1}^n w_i I_i = w_1 (w_{1a} I_{1a} + w_{1b} I_{1b}) + \sum_{i=2}^n w_i I_i \\ &= w_1 w_{1a} I_{1a} + w_1 w_{1b} I_{1b} + \sum_{i=2}^n w_i I_i. \end{aligned} \quad (4.17)$$

Met andere woorden, de gewichten van de gesplitste bedrijven 1a en 1b worden respectievelijk $w_1 w_{1a}$ en $w_1 w_{1b}$. Hierbij is aangenomen dat beide bedrijven na de splitsing in de steekproef blijven.

De situatie wordt iets anders wanneer bijvoorbeeld de totale omzet O moet worden geschat van een populatie/stratum van bedrijven. Onder de aanname dat het geschatte omzettotaal kan worden geschreven als

$$\hat{O}_s = w_1 O_1 + \sum_{i=2}^n w_i O_i, \quad (4.18)$$

gaat de schatter na splitsing van bedrijf 1 over in

$$\hat{O}_s = w_1 O_{1a} + w_1 O_{1b} + \sum_{i=2}^n w_i O_i. \quad (4.19)$$

Met andere woorden, de gesplitste bedrijven behouden hetzelfde gewicht als het moederbedrijf voor de splitsing.

4.4 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor panels om indices te schatten zijn:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele.

5. Literatuur

- Banning, R., Camstra, A. en Knottnerus, P. (2010). *Methodenreeks: Thema: Steekproeftheorie, Deelthema's: Steekproefontwerpen en Ophoogmethoden*, CBS, Den Haag.
- Knottnerus, P. and Delden, A. van (2006). Estimation of changes in repeated surveys and their significance. <http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/paper/Knottnerus.doc>
- Knottnerus, P. (2011). On the efficiency of randomized PPS sampling, *Survey Methodology* (forthcoming).
- Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers, *Journal of Official Statistics* 16, 363–378.