Schatten van kenmerken van kleine deelgebieden

Dr. Hans C.J. Vrolijk
Dr. Wietse Dol
Drs. Geerte Cotteleer

Projectcode 63393

Augustus 2002

Rapport 8.02.05

LEI, Den Haag

Het LEI beweegt zich op een breed terrein van onderzoek dat in diverse domeinen kan worden opgedeeld. Dit rapport valt binnen het domein:

- ☐ Wettelijke en dienstverlenende taken
- ☐ Bedrijfsontwikkeling en concurrentiepositie
- ☐ Natuurlijke hulpbronnen en milieu
- ☐ Ruimte en Economie
- ☐ Ketens
- ☐ Beleid
- ☐ Gamma, instituties, mens en beleving
- ☑ Modellen en Data

Schatten van kenmerken van kleine deelgebieden
Vrolijk, H.C.J., W. Dol, G. Cotteleer
Den Haag, LEI, 2002
Rapport 8.02.05; ISBN 90-5242-755-0 ; Prijs € 19,- (inclusief 6% BTW)
110 p., fig., tab.

Dit rapport geeft een beschrijving en analyse van methoden voor het schatten van kenmerken van kleine deelgebieden. Deze methoden kunnen worden toegepast om betere schattingen te maken voor gebieden (bijvoorbeeld gemeentes, provincies, landbouwgebieden enzovoort) of voor groepen (landbouw sectoren).

Bestellingen:
Telefoon: 070-3358330
Telefax: 070-3615624
E-mail: publicatie@lei.dlo.nl

Informatie:
Telefoon: 070-3358330
Telefax: 070-3615624
E-mail: informatie@lei.dlo.nl

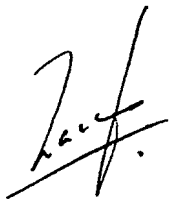
Inhoud

Woord vooraf

Binnen het LEI wordt veel statisch onderzoek verricht op basis van beschikbare gegevensverzamelingen, zoals de Landbouwtelling en het Bedrijven-Informatienet van het LEI (het Informatienet). Dit rapport beoogt de kwaliteit van deze gegevensverwerking te verhogen door na te gaan hoe onderzoekers op dit moment met de gegevens omgaan en een beschrijving en analyse te geven van methoden die de gegevensverwerking kunnen verbeteren. De methoden die hierbij centraal staan zijn de methoden voor het schatten van kenmerken van kleine deelgebieden. Door het gebruik van extra informatie bieden deze methoden de mogelijkheid betere schattingen op te leveren.

Dit rapport is opgesteld door Hans Vrolijk, Wietse Dol en Geerte Cotteleer. Daarnaast bedanken we de onderzoekers voor hun open opstelling om de gehanteerde werkwijzen in kaart te brengen. Wij hopen dat de onderzoekers in de toekomst zinvol gebruik kunnen maken van de in dit rapport beschreven methoden.

De directeur,



Prof.dr.ir. L.C. Zachariasse

Samenvatting

Regelmatig voert het LEI onderzoek uit waarbij resultaten voor een klein gebied (gemeenten, provincies, landbouwgebieden, kaartvierkanten) of kleine groep gewenst zijn. In veel gevallen worden deze resultaten geproduceerd door beschikbare of berekende bedrijfsgegevens 'op te hogen' naar het gewenste aggregatieniveau. Hierbij is het noodzakelijk dat er voldoende waarnemingen voor het gebied zijn om verantwoord te kunnen aggregeren.

Aggregatie van gegevens die betrekking hebben op Informatienet-bedrijven is voor kleine gebieden veelal niet mogelijk op basis van de gebruikelijke procedure die gebruikmaakt van wegingsfactoren. In de loop van de tijd zijn daarom op het LEI verschillende methodes toegepast/ontwikkeld die het mogelijk maken om toch bruikbare informatie op een laag ruimtelijk aggregatieniveau te genereren. Het is nuttig een vergelijking te maken van de beschikbare methoden voor het maken van schattingen voor kleine gebieden.

Conclusies uit interviews

In het vooronderzoek is geprobeerd een beeld te schetsen hoe en in welke mate op het LEI gebruik wordt gemaakt van steekproeven. De meest gebruikte steekproef is het Informatienet, daarnaast worden diverse andere steekproeven gebruikt, bijvoorbeeld in de vorm van enquêtes.

Op basis van de steekproeven worden puntschattingen voor gemiddeldes berekend voor de gehele populatie en voor delen van de populatie. In de huidige onderzoekspraktijk op het LEI wordt weinig aandacht besteed aan de *betrouwbaarheid* van uitkomsten. De consequentie hiervan is dat men geen statistisch verantwoorde uitspraken kan doen bij het vergelijken van scenario's, groepen en jaren. Het voordeel van meer geavanceerde schattingsmethodes is juist gelegen in een toename van de betrouwbaarheid en niet de zuiverheid van de schatter. Zolang er dus geen omslag plaatsvindt van een punt-denken naar een interval-denken zal het moeilijk zijn de voordelen van alternatieve methoden duidelijk te maken. Indien de onderzoeker niet de voordelen ziet zal hij of zij minder geneigd zijn om zich te verdiepen in de materie. Gezien de hogere complexiteit van alternatieve schattingswijzen is deze verdieping wel een vereiste voor een statistisch verantwoord gebruik.

Ondanks deze constatering is er een groeiend toepassingsterrein voor methoden voor het schatten van kenmerken van kleine deelgebieden. Er is een toenemende vraag naar schattingen voor regionale gebieden. Ook bestaat er behoefte aan methoden om gegevens te produceren die in modellen zoals het Ruimtelijk Economische Model (REM) kunnen worden gebruikt. Met name bestaat er een grote behoefte aan inkomensgegevens voor kleine gebieden.

Er is een duidelijke vraag naar en behoefte aan (bij)scholing op het gebied van statistiek en steekproeftechnieken. Dit wordt binnen het instituut vaak onterecht vereenzelvigd met kwantitatieve kennis. Een kwantitatieve scholing of insteek wil niet per definitie ze g-

gen dat iemand kennis heeft van statistiek en steekproeftechnieken. Gezien het belang van het gebruik van het Informatienet en andere steekproeven lijkt een uitbreiding van deze kennis onontbeerlijk voor het LEI.

Schattingsmethoden

Verschillende methoden komen in aanmerking om schattingen te maken van kenmerken van kleine deelgebieden. Er wordt een samenvatting gegeven van alle methoden die uit dit onderzoek naar voren zijn gekomen. Voor elk specifiek probleem dient een afweging gemaakt te worden, welke methode het beste kan worden gebruikt.

Directe schatter
Het is gebruikelijk om schattingen van gemiddelden en totalen te maken op basis van de waarden van de variabele zoals die in de steekproef zijn waargenomen. De totale melkproductie kan bijvoorbeeld op deze manier worden geschat door de melkproductie op de steekproefbedrijven op te hogen naar de populatie middels de in het Informatienet beschikbare gewichten.

Ratioschatter
Indien een hulpvariabele beschikbaar is die in grote mate correleert met de doelvariabele, dan kan deze hulpvariabele worden gebruikt voor het maken van betrouwbaardere schattingen. Als men bijvoorbeeld een schatting wil maken van de totale melkproductie, kan gebruik worden gemaakt van het gegeven dat de melkproductie op een bedrijf sterk zal correleren met het aantal koeien op dat bedrijf. Bij het gebruik van de ratioschatter geldt wel de voorwaarde dat het gemiddelde of totaal van deze hulpvariabele voor de hele populatie bekend moet zijn en dat deze geen stochast is. Voor het aantal koeien is dit het geval, op basis van de Landbouwtelling kan het totaal aantal koeien worden vastgesteld. De reden waarom deze indirecte schatting betrouwbaarder kan zijn dan een directe schatting is dat de verhouding tussen twee variabelen stabieler kan zijn dan de variabelen afzonderlijk. De melkproductie op verschillende boerderijen kan sterk uiteenlopen. Een directe schatting zou dan ook een hoge variantie laten zien. De melkproductie zal echter sterk afhankelijk zijn van het aantal koeien. De verhouding productie per koe zal een kleinere spreiding laten zien dan de spreiding in de melkproductie of het aantal koeien zelf over de steekproefbedrijven. Indien men op basis van andere bronnen gegevens heeft over het totaal aantal koeien dan kan een veel nauwkeurigere uitspraak over de totale melkproductie in Nederland worden gedaan.

Een bijkomend voordeel van het gebruik van ratioschatters is dat de representativiteit wordt verhoogd. Stel dat in de steekproef vooral kleine bedrijven zijn opgenomen. Doordat in de indirecte schatter van de totale melkproductie rekening wordt gehouden met het aantal koeien op de steekproefbedrijven ten opzichte van het aantal koeien in Nederland wordt automatische gecorrigeerd voor de omvang van de bedrijven. De verhouding melkproductie per koe wordt vermenigvuldigd met de uit een andere bron bekende aantal koeien. Een directe schatter zou in dit geval tot een onderschatting van de totale melkproductie leiden.

Regressieschatters
Regressieschatters maken net als de ratioschatters gebruik van extra informatie van een hulpvariabele die sterk is gecorreleerd met de doelvariabele. Wanneer er wel een verband bestaat tussen deze variabelen maar wanneer deze niet door de oorsprong gaat of wanneer er meerdere verklarende variabelen zijn, kan beter gebruik worden gemaakt van een regressieschatter dan van een ratioschatter. Bij een relatie tussen aantal koeien en de melkproductie is het aannemelijk dat de relatie door de oorsprong gaat. Een veestapel van nul koeien zal immers leiden tot een melkproductie van nul liter. In andere situaties is de aanname van een verband door de oorsprong minder waarschijnlijk. Indien bijvoorbeeld een verband wordt verondersteld tussen het nettobedrijfsresultaat en het aantal koeien dan zal dit verband niet door de oorsprong gaan. In verband met vaste lasten zal op een gespecialiseerd bedrijf een veestapel van nul koeien leiden tot een negatief bedrijfsresultaat. In dergelijke situaties is het gebruik van een regressieschatter aan te raden.

Bayesiaanse schatter
De Bayesiaanse schatter is een lineaire combinatie van de regressieschatter en de directe schatter. Als het verband tussen doel- en hulpvariabele niet alleen opgaat binnen een klein deelgebied, maar geldt voor de gehele populatie, verdient de Bayesiaanse schatter de voorkeur boven een directe of een regressieschatter. Bayesiaanse analyse maakt beter gebruik van de informatie die vooraf beschikbaar is over de te schatten grootheden dan bijvoorbeeld een directe schatter of een regressieschatter. De directe schatter gebruikt enkel de individuele eigenschappen van de doelvariabele binnen een bepaald klein deelgebied en de lineaire regressieschatter maakt enkel gebruik van de relatie tussen de doelvariabele met andere variabelen (verklarende variabelen) die sterke correlatie vertonen met de doelvariabele. De regressieschatter houdt geen rekening met het feit dat individuele eigenschappen kunnen gelden voor de doelvariabele binnen verschillende kleine deelgebieden die niet terugkomen in de verklarende variabele(n).

Voor onderzoek op het LEI zou deze techniek gebruikt kunnen worden wanneer onderzoeksvragen een bepaalde regio aangaan en verondersteld kan worden dat een bepaalde relatie tussen hulp- en doelvariabele voor het gehele land geldt.

Poststratificatie
In het geval een dataset een groot aantal toepassingen heeft, dat wil zeggen dat een groot aantal variabelen als doelvariabele gebruikt wordt, heeft poststratificatie de voorkeur boven stratificatie vooraf (Sarndal, 1992). Bij een gestratificeerde steekproefopzet worden de strata definitief ingevoerd. Dit leidt tot een reductie in de variantie voor de daarbij gehanteerde doelvariabelen, de stratificatievariabelen. Deze opzet kan echter minder efficiënt zijn voor vele andere doelvariabelen. De combinatie van een aselecte steekproef en poststratificatie kan de totale efficiency verbeteren. Bij de analyse van gegevens kan gebruik worden gemaakt van de kennis en intuïtie van de onderzoeker om bij het onderzoek passende poststratificaties vast te stellen.

Stel dat in het type glastuinbouw twee typen productiesystemen bestaan die van grote invloed zijn op het energieverbruik. Indien men kennis heeft over de verdeling van deze systemen in de populatie (bijvoorbeeld op basis van de Landbouwtelling) dan kan men deze kennis gebruiken om een betere schatting te maken. Stel dat men weet dat 30% van de

bedrijven productiesysteem A gebruikt en 70% systeem B. Omdat de steekproef niet is gestratificeerd op basis van dit kenmerk kan het voorkomen dat in de steekproef 50% van de bedrijven systeem A en 50% systeem B gebruikt. In een onderzoek naar het energieverbruik kan het zinvol zijn te corrigeren voor deze verhouding. Poststratificatie leidt er toe dat het gewicht van bedrijven met systeem A iets lager wordt (bedrijven met Systeem A zijn oververtegenwoordigd in de steekproef) en bedrijven met systeem B iets hoger wordt (bedrijven met systeem B zijn ondervertegenwoordigd) bij het maken van schattingen omtrent het energieverbruik.

Datafusie en imputatie

Datafusie is een methode om gegevens vanuit verschillende bronnen te integreren en samen te voegen. Binnen het LEI kunnen op die manier Informatienet- en Landbouwtellingsgegevens worden geïntegreerd. De kenmerken in de Landbouwtelling zijn bekend voor alle agrarische bedrijven groter dan circa 3 nge. Daarnaast is in het Informatienet een gedetailleerde administratie beschikbaar van een kleine 1.500 bedrijven. Voor het overgrote deel van de bedrijven in de Landbouwtelling is deze gedetailleerde administratie niet beschikbaar. Om toch uitspraken te kunnen doen over kenmerken die gelden voor de populatie op het kleine deelgebied, gaat men op zoek naar bedrijven (waarvan wel een administratie beschikbaar is) welke op basis van kenmerken in de Landbouwtelling sterk op het bedrijf, waarover men een uitspraak wil doen, lijken. Een bedrijf dat sterk op een ander bedrijf lijkt op basis van de beschikbare variabelen in de Landbouwtelling zal naar alle waarschijnlijkheid ook lijken op dat andere bedrijf voor variabelen die niet beschikbaar zijn, ervan uitgaande dat de beschikbare en de niet-beschikbare variabelen in grote mate met elkaar gecorreleerd zijn.

De methode kan bijvoorbeeld worden toegepast indien men een uitspraak wil doen over een regio waarvoor men over weinig directe waarnemingen beschikt. In een regio zullen bedrijven zitten van verschillende typen. Om alle typen afzonderlijk te schatten zijn veel waarnemingen nodig. Middels datafusie en imputatie gaat men op zoek naar bedrijven die een grote gelijkenis vertonen met de bedrijven in de te bestuderen regio. Men zoekt voor elk bedrijf in de regio naar een bedrijf in het Informatienet dat er sterk op lijkt gegeven de kenmerken in de Landbouwtelling. Vervolgens wordt de veronderstelling gemaakt dat de te schatten kenmerken van het bedrijf ook hetzelfde zullen zijn, ervan uitgaande dat de gebruikte kenmerken in de Landbouwtelling gecorreleerd zijn met de kenmerken uit het Informatienet. De gegevens van het gelijkende bedrijf in het Informatienet worden dus van toepassing verklaard op het bedrijf in de te bestuderen regio waar men deze gegevens niet direct heeft waargenomen. Op basis van deze (geïmputeerde) gegevens kunnen vervolgens bepaalde statistieken voor de regio worden berekend.

Evaluatie schattingstechnieken

In figuur 1 zijn de methoden op een groot aantal criteria geëvalueerd.

Methode Criteria	Directe schatter	Ratio schatter	Regressie schatter	Bayesiaanse schatter	Post-stratificatie	Regressie-model	Hot deck procedures
Betrouwbaarheid	++	++	++	-	+	-/+	-/+
Te berekenen betrouwbaarheid bij kleine aantallen	-	+	+	Nvt	-	-/+	Nvt
Zuiverheid	++	-/+	-/+	-/+	+	?	--
Indicatie GFI aannames	Nvt	+	+	+	Nvt	+	-
Validiteit bij kleine aantallen	-	+	++	++	++	+	??
Onderbouwing	Steek-proeven	Model en steekproef	Model en steekproef	Model en steekproef	Steek-proeven	Model en steekproef	Afstands-maten
Eenvoud	++	-	--	--	-	--	+
Bewerkelijkheid	++	-	--	--	-	--	+
Flexibiliteit	++	-	-	-	-	-	++
Wetenschappelijke acceptatie	++	-/+	-/+	-	+	-/+	-
Meerdere doelvariabelen	+	-	-	-	+	-	++
Gebruik extra info	Geen	Gemiddelde of totaal hulpvariabele	Gemiddelde van hulp-variabelen	Gemiddelde van hulp-variabele en directe schatter op alle deelgebieden	Verdeling in populatie	Kenmerken in de populatie	Kenmerken in populatie
Meerdere hulpvariabelen	Nvt	-	+	+	+	+	+
Nominale of ordinale hulpvariabele	Nvt	-	-	-	++	+	++
Interval of ratio hulpvariabele	Nvt	++	++	++	-	++	++
Reproduceerbaarheid	++	+	+	+	+	+	-

Figuur1 Evaluatie van de beschreven methoden

Vernieuwing Informatienet en het schatten van kenmerken van kleine deelgebieden

Het Informatienet is met ingang van 2000 sterk vernieuwd. Doel is om het Informatienet meer vraaggestuurd en flexibeler te maken. Dat gaat gepaard met harmonisatie van land- en tuinbouwbedrijven en overgang op een uniform kalenderjaar. Door die overgang komen over 2000 minder gegevens beschikbaar. In komende jaren wordt toepassing van de in dit rapport opgebouwde kennis belangrijker. Het uitwerken van een relatief zeer klein aantal boekhoudingen van het jaar 2000 zal voor veel onderzoekers tot problemen leiden. Het tekort aan data in 2000 en eventueel ook in de daaropvolgende jaren is echter één van de problemen waar onderzoekers en beleidsmakers mee te maken zullen krijgen. Een ander probleem dat zich voordoet is de herdefiniëring van bepaalde variabelen in het nieuwe systeem. Gerelateerde problemen zijn trendbreuken die ontstaan, specifieke gegevens die

binnen bepaalde modellen gebruikt worden en die niet jaarlijks, gedeeltelijk, of in het geheel niet beschikbaar zijn in het nieuwe systeem. Ook de representativiteit van de steekproef speelt een rol.

Voor het tekort aan data in 2000 en andere problemen gerelateerd aan de vernieuwing van het Informatienet is geen eenzijdige oplossing te geven. Wat een oplossing is voor het ene probleem zal voor het andere probleem geen oplossing zijn. De oplossing kan over het algemeen in twee richtingen gezocht worden. De eerste gaat uit van het gebruik van andere databronnen en de tweede maakt gebruik van schattingstechnieken voor kleine deelgebieden. Een combinatie van deze twee oplossingsrichtingen is ook mogelijk. Als ervoor gekozen wordt schattingstechnieken voor kleine deelgebieden te gebruiken, zullen met name ratio- en regressieschatters en imputatie en datafusie een rol kunnen spelen.

1. Inleiding en probleemstelling

1.1 Inleiding

Regelmatig voert het LEI onderzoek uit waarbij resultaten voor een klein gebied (gemeenten, provincies, landbouwgebieden, kaartvierkanten) of kleine groep (sectoren) gewenst zijn. In veel gevallen worden deze resultaten geproduceerd door beschikbare of berekende bedrijfsgegevens 'op te hogen' naar het gewenste aggregatieniveau. Daarvoor is het nodig dat er voldoende waarnemingen voor het gebied zijn om verantwoord te kunnen aggregeren.

Aggregatie van gegevens die betrekking hebben op Informatienet-bedrijven is voor kleine gebieden veelal niet mogelijk op basis van de gebruikelijke procedure die gebruikmaakt van wegingsfactoren. In de loop van de tijd zijn daarom binnen LEI verschillende methodes toegepast/ontwikkeld die het mogelijk maken om toch bruikbare informatie op een laag ruimtelijk aggregatieniveau te genereren. Voorbeelden daarvan zijn a) het gebruik van bedrijfsgegevens uit andere ruimtelijke eenheden die nauw verwant zijn aan de betreffende ruimtelijke eenheid (eventueel in combinatie met poststratificatie zoals bij onderzoek ten behoeve van landinrichting), b) het werken met geschatte verbanden tussen wel en niet bekende grootheden, en c) de methode die onder andere in het project grondbalansen is toegepast ('fuzzy poststratificatie').

Uitgaande van deze ontwikkeling is het nuttig een vergelijking te maken van de beschikbare methoden (en varianten daarop) voor het maken van schattingen voor kleine gebieden. Daarbij moet duidelijk worden hoe de methoden zich verhouden in kwaliteit, eenvoud, bewerkelijkheid, flexibiliteit, en dergelijke. Op basis daarvan kunnen aanbevelingen worden gedaan over eventuele verdere ontwikkeling of operationalisering van methoden.

1.2 Doelstelling

De doelstelling van dit onderzoek is: *Het verschaffen van inzicht in de gebruiksmogelijkheden en toepasbaarheid van methoden voor het schatten van kenmerken van kleine deelgebieden.*

Ten einde deze doelstelling te realiseren zullen de volgende deelvragen moeten worden beantwoord:

- welke methoden voor het schatten van kenmerken van kleine deelgebieden worden gebruikt binnen het LEI en welke zijn beschreven in de literatuur;
- wat zijn de kenmerken en voor- en nadelen van deze methoden;
- in hoeverre zijn de methoden bruikbaar binnen het LEI;
- welke criteria spelen een rol bij de keuze van een methode;
- hoe dient het LEI in de toekomst met de genoemde methoden om te gaan?

1.3 Methode van onderzoek

Voor het beantwoorden van de deelvragen is gebruikgemaakt van de volgende informatiebronnen en methoden van onderzoek:

- medewerkers binnen het LEI zijn geïnterviewd ten aanzien van het huidige gebruik van methoden en de wensen ten aanzien van het toekomstige gebruik;
- literatuur omtrent 'small area estimation', imputatie, poststratificatie en indirecte schatters;
- LEI-publicaties waarin toepassingen van in het verleden gebruikte methoden beschreven zijn.

1.4 Opzet rapport

In hoofdstuk 2 wordt een korte introductie gegeven omtrent het schatten van kenmerken van kleine deelgebieden. In hoofdstuk 3 wordt de inventarisatie van het gebruik van steekproeven en hergebruik van steekproeven binnen het LEI beschreven. Tevens wordt aandacht besteed aan de mogelijkheden tot het toepassen van methoden voor het schatten van deelgebieden in de toekomst. In hoofdstuk 4 volgt een beschrijving van de methoden voor het schatten van kenmerken van kleine deelgebieden. De methoden worden geïllustreerd aan de hand van een voorbeeld. Tevens wordt de toepasbaarheid van de methoden geëvalueerd. In hoofdstuk 5 wordt een aantal criteria voor het vergelijken van de methoden gedefinieerd. Tevens zal worden aangegeven hoe de methoden op deze criteria worden beoordeeld. In hoofdstuk 6 worden een samenvatting en de conclusies van het onderzoek gegeven. In hoofdstuk 7 worden de implicaties voor het onderzoek binnen het LEI beschreven.

2. Ervaringen omtrent het schatten van kenmerken van kleine deelgebieden binnen het LEI

In het kader van het project 'het schatten van kenmerken van kleine deelgebieden' wordt inzicht verschaft in de gebruiksmogelijkheden en toepasbaarheid van methoden voor het maken van schattingen van grootheden op kleine deelgebieden. Het schatten op kleine deelgebieden impliceert het hergebruik van de steekproef om uitspraken te doen over een deelpopulatie. Deze deelpopulatie hoeft niet per se klein te zijn. De eerste fase van het project omvat een inventarisatie van het gebruik van deze technieken binnen het LEI. Omdat wij het schatten van kleine deelgebieden definiëren als het hergebruik van steekproeven om uitspraken te doen over deelpopulaties hebben wij bij de interviews binnen het LEI een iets bredere insteek gekozen. In de interviews willen wij in kaart brengen hoe met steekproeven wordt omgegaan, zowel ten aanzien van de hele populatie als ten aanzien van deelpopulaties.

2.1 Doelstelling interviews

De doelstelling van de interviews is: *Inventariseren van het gebruik van steekproeven en het gebruik van technieken voor het maken van schattingen van kenmerken van populaties in het algemeen en deelpopulaties in het bijzonder*.

Gebruik wordt hierbij breder gedefinieerd dan alleen maar het al dan niet toepassen van de techniek. Ook ervaringen ten aanzien van het gebruik en mogelijke wensen voor de toekomst worden hierbij meegenomen.

De doelstelling is vertaald in een aantal onderzoeksvragen:

- welke steekproeven worden gebruikt en hoe worden deze gebruikt;
- worden schattingen gemaakt van deelpopulaties en zo ja op welke wijze;
- worden methoden voor het schatten van deelgebieden reeds toegepast;
- welke wensen bestaan er ten aanzien van het toekomstige gebruik van deze methoden?

2.2 Methode van onderzoek

Middels een aantal interviews binnen de verschillende afdelingen wordt antwoord gegeven op de onderzoeksvragen. Deze interviews zijn in 2000 afgenomen. Hierdoor kan de beschreven organisatiestructuur afwijken van de huidige. In de bijlage is de vragenlijst weergegeven, zoals die in de interviews is gebruikt.

2.3 Resultaten van de interviews

In de hieronder volgende paragrafen zal antwoord worden gegeven op de vier geformuleerde onderzoeksvragen, daarnaast zal nog een paragraaf worden besteed aan algemene bevindingen.

2.3.1 Gebruik van steekproeven

De meest gebruikte steekproef binnen het LEI is het Bedrijven-Informatienet van het LEI (het Informatienet) (met de daaraan gekoppelde bestanden) (zie Van Dijk et al., 2002). Naast het Informatienet wordt gebruikgemaakt van CBS-steekproeven zoals de stalsystemen en uitrijsystemen. Verder wordt gebruikgemaakt van de Landbouwtelling (al dan niet te beschouwen als een steekproef) en het Europese Farm Accountancy Data Network (FADN/RICA).

Naast de hiervoor genoemde regelmatig uitgevoerde steekproeven worden er met name bij de afdeling Structuuronderzoek (SO) diverse enquêtes uitgezet. Ook hier is nadrukkelijk sprake van steekproeven. Een deel van deze enquêtes wordt als aanvullende enquêtes uitgezet bij bestaande Informatienet-bedrijven teneinde aanvullende gegevens voor een specifiek onderzoek te verzamelen. Het ongewijzigde gebruik van de wegingsfactoren van de Informatienet-bedrijven bij de analyse van dergelijke aanvullende steekproeven kan een bijzonder verstorend effect tot gevolg hebben. Bij het trekken van een steekproef binnen een steekproef (welke weer een afzonderlijke non-response tot gevolg kan hebben) representeren de Informatienet wegingsfactoren niet meer de trekkingskans van deze bedrijven.

2.3.2 Schattingen met betrekking tot deelpopulaties

Regelmatig worden in LEI-onderzoek uitspraken gedaan over deelpopulaties. Deelpopulaties die regelmatig voorkomen zijn regio's, bedrijfstypen, inkomensklassen, leeftijdsgroepen, omvangklassen, landbouwgebieden, activiteiten en gewasgroepen.

Bij PPRF (Landbouw) wordt bij het doen van onderzoek meestal een insteek gekozen naar bedrijfstype, grootte klasse, regio en leeftijd. Bij het analyseren van de Informatienetgegevens wordt altijd gebruikgemaakt van een weging. Dit levert wel eens problemen op bij het analyseren van kleine groepen waarbij de som van de wegingen kan afwijken van het aantal bedrijven in die groep volgens de Landbouwtelling. Voor de typering van bedrijven wordt uitgegaan van de NEG-typering uit de Landbouwtelling.

Tevens is de sectie PPRF (afdeling Landbouw) betrokken bij regionaal onderzoek. Bij dit type onderzoek wordt men vaak geconfronteerd met te weinig waarnemingen. Indien het aantal waarnemingen lager is dan 20 dan is een directe schatting van een totaal of gemiddelde op een klein deelgebied niet genoeg. In het verleden werd hiertoe gebruikgemaakt van de methode Tjomme de Haan [1]. Opdrachtgevers willen graag iets weten over het financiële plaatje en deze methode biedt daartoe de mogelijkheid. Op dit moment wordt een variant van de methode gebruikt.

[1] Zie hoofdstuk 4 voor een verdere beschrijving van deze methode.

Bij de sectie AEOS (afdeling Structuuronderzoek) richt men zich met name op de 14 landbouwgebieden en op circa 25 activiteiten. Voor het maken van schattingen voor alle 14 gebieden zijn vaak niet voldoende waarnemingen beschikbaar. Om dit probleem op te lossen zijn de 14 gebieden samengevoegd tot 3 overkoepelende gebieden. Schattingen voor deze 3 gebieden worden weer gedesaggregeerd naar de 14 gebieden (het geschatte gemiddelde op hoger niveau wordt van toepassing verklaard op de gebieden op een lager niveau).

Bij de sectie MESO (afdeling Landbouw) worden op basis van gegevens in het Informatienet acceptatiegraden en kunstmestgiften per gewasgroep en regio vastgesteld. Hierbij wordt uitgegaan van de 31 mestgebieden. In een aantal gebieden is het Informatienet niet goed vertegenwoordigd. Als minimum aantal waarnemingen wordt uitgegaan van 20. Indien het aantal waarnemingen lager is worden regio's samengevoegd. De deelpopulaties die in de hiervoor beschreven aanpak zijn te onderscheiden zijn de regio's en de gewassen.

Bij het gebruik van het Financieel Economisch Simulatiemodel (FES) wordt regelmatig een indeling in typen of inkomensklassen gehanteerd. Voor het doen van uitspraken over groepen worden de gegevens opgehoogd met behulp van de Informatienet gewichten. De indeling in groepen, de typering, is gebaseerd op de Informatienet gegevens. Eventuele gevolgen van deze basis worden niet standaard beschouwd. Met FES worden er niet vaak uitspraken gedaan over regionale gebieden. FES is met name gericht op nationale problemen ten aanzien van belasting- en beleidsmaatregelen. Er wordt weinig gedaan aan regionale problemen bij bijvoorbeeld de provincie. Andere deelpopulaties die af en toe worden gebruikt zijn gebaseerd op een indeling naar inkomen.

Bij SO bestaat regelmatig de behoefte uitspraken te doen over regio's. In onderzoek wordt regelmatig een situatieschets voor een regio gemaakt. Hierbij zou men graag ook iets zeggen over financieel economische kengetallen van de landbouw. In de huidige situatie wordt vaak op basis van gegevens van het CBS een uitspraak gedaan over een regio. Deze gegevens zijn vaak op een hoger aggregatie niveau (bijvoorbeeld provincie) en moeten dus vertaald worden naar kleinere regio's.

Bij gebiedsgericht onderzoek wordt vaak geen steekproef gebruikt. In veel gevallen worden alle bedrijven zoals die in de Landbouwtelling voorkomen meegenomen. Wanneer inkomenscijfers gewenst zijn, wordt gebruikgemaakt van een steekproef in de vorm van het Informatienet (vaak in samenwerking met de sectie PPRF). Bij een verkenning probeert men voor clusters en typen op basis van gegevens uit het Informatienet een schatting te maken. Gezinsinkomen uit bedrijf per ondernemer is hierbij een belangrijke variabele. Bij een minimum van 10 à 15 per type wordt een directe schatting van het gemiddelde gemaakt. De gewichten vanuit het Informatienet worden hierbij niet gebruikt. Als het aantal bedrijven kleiner is dan 10 dan wordt bijvoorbeeld op basis van gesprekken een inschatting gemaakt. Een variant hiervan is dat men op basis van het teeltplan en schattingen van de opbrengst per hectare een schatting maakt van het inkomen.

Bij de sectie AM (Landbouw) worden tal van deelpopulaties bestudeerd. Voor kleine regio's zullen er in veel gevallen niet voldoende bedrijven beschikbaar zijn. In dergelijke gevallen wordt op basis van de beschikbare Landbouwtellingsgegevens voor elk bedrijf in die regio gezocht naar een zo sterk mogelijk gelijkend bedrijf in het Informatienet. De vergelijking wordt gemaakt op basis van 20 criteria. De exacte keuze van criteria is afhankelijk van het doel van het onderzoek. Nadat voor elk bedrijf een zo goed mogelijk

gelijkend bedrijf is gevonden wordt met deze gegevens verder gewerkt. Vervolgens kunnen bijvoorbeeld gemiddeldes voor kleine regio's worden berekend. De stellige indruk bestaat dat dit tot betere schattingen leidt dan wanneer men op basis van bijvoorbeeld slechts een of twee beschikbare bedrijven een schatting maakt voor een regio. De hier genoemde methode is niet zo zeer een statistische methode maar een methode waarin meer gebruik wordt gemaakt van expertkennis. Dit is een andere aanpak van hetzelfde probleem.

2.3.3 Gehanteerde methoden voor het schatten op kleine deelgebieden

In het hiervoor genoemde onderzoek waarin uitspraken worden gedaan over deelpopulaties zijn meer of minder expliciet methoden voor het schatten van kenmerken van kleine deelgebieden te herkennen. Daarnaast is in het verleden een aantal studies verricht die direct gericht waren op het doen van uitspraken over deelpopulaties (meestal regio's). Hieronder zullen deze methoden en studies worden besproken. Met dit overzicht pretenderen wij niet een volledig overzicht te geven.

Kunstmestonderzoek

Begin jaren tachtig is een onderzoek uitgevoerd dat specifiek gericht was op het schatten van kenmerken van kleine deelgebieden. Het doel van het onderzoek was het schatten van kunstmestgiften op gemeentelijk niveau.

In dit onderzoek zijn relaties geschat tussen de kunstmestgiften per hectare en de bedrijfskenmerken op basis van de gegevens uit de steekproef. De kunstmestgiften zijn onderverdeeld in stikstof, fosfaat en kali en ze zijn uitgesplitst naar het gebruik op grasland, bouwland, eenjarige opengrondstuinbouwgewassen, meerjarige opengrondstuinbouwgewassen en gewassen onder glas. Met de relaties en de bedrijfskenmerken die voor elk bedrijf in de Landbouwtelling te vinden zijn kan dan voor elk bedrijf een schatting worden gemaakt van de kunstmestgiften per hectare.

Het CBS was geen voorstander van de gehanteerde aanpak omdat zij alleen van concrete waarnemingen en directe schatters wilden uitgaan. Intern was men zeer tevreden over de gehanteerde aanpak en het resultaat. Het zou een grote toegevoegde waarde hebben indien een dergelijke analyse jaarlijks zou kunnen worden uitgevoerd. Het kost echter flink wat tijd en geld om deze analyse uit te voeren. In het stofstromen model is een soortgelijke aanpak wel gedeeltelijk geïmplementeerd.

Stofstromenmodel

In het stofstromenmodel wordt een methodiek gebruikt die sterk gerelateerd is aan de problematiek van de kleine deelgebieden. Hierbij wordt een relatie gelegd tussen het Informatienet en de Landbouwtelling. Er wordt een functie geschat waarbij een variabele uit het Informatienet (bijvoorbeeld stikstofgift per hectare) wordt geschat als functie van één of meer kenmerken uit de Landbouwtelling. De functie wordt geschat op basis van de gehele populatie. De keuze van de kenmerken die in de functie worden opgenomen is een belangrijke stap. Op basis van de onderzoeksvraag en kennis van het onderwerp worden relevante variabelen geselecteerd, vervolgens wordt gekeken in hoeverre deze variabelen

echt een verklarende waarde hebben, tevens wordt op de samenhang oftewel de correlatie tussen de verschillende variabelen gelet. In het stofstromenmodel is een vergelijking voor de stikstofgift per hectare grasland en maïsland geschat op basis van een aantal jaargangen. De resulterende vergelijking is opgenomen in het stofstromenmodel.

Land en tuinbouw in Noord- en Midden-Limburg

In dit onderzoek is een schatting gemaakt van de opbrengst van de gezinsarbeid van bedrijven in Limburg. Deze wordt in dit onderzoek berekend als het verschil tussen de brutowinst en alle vaste kosten (exclusief gezinsarbeid). Voor zowel de berekening van de brutowinst als de vaste kosten zijn uit het Informatienet algemene relaties afgeleid. Deze relaties zijn vervolgens ingevuld met gegevens uit de Landbouwtelling van de bedrijven in Noord- en Midden-Limburg. Op deze manier is de arbeidsopbrengst voor elk afzonderlijk bedrijf berekend.

SIRAS

SIRAS is het simulatiemodel voor de regionale agrarische structuur. Het rekenmodel heeft tot doel inzicht te verschaffen in de toekomstige structuur van de land- en/of tuinbouw in een gebied onder invloed van onder meer economische, technische, planologische en politieke ontwikkelingen en/of ingrepen. Meer concreet betekent dit dat het model inzicht moet geven in de ontwikkeling van het aantal agrarische bedrijven; de productieomvang en de verhouding waarin de productiefactoren worden ingezet. SIRAS is voorafgegaan door het regionale model voor de prognose van de agrarische structuur. Op basis van geschatte overgangskansen en ontwikkelingen wordt een prognose van de structuur gemaakt.

Bodembalansen Zuid-Holland

Bij het opstellen van de bodembalans voor Zuid-Holland wordt een andere aanpak gehanteerd. Voor bepaalde gewassen is het aantal directe waarnemingen in deze provincie te gering. Voor dergelijke gewassen wordt de westelijke regio gehanteerd voor het maken van een schatting voor Zuid-Holland.

Grondbalansen onderzoek

Voor elk bedrijf in de Landbouwtelling wordt een steekproefbedrijf gezocht dat voor een aantal specifieke kenmerken zo goed mogelijk lijkt op het Landbouwtellingsbedrijf. Daarbij wordt een voorselectie gemaakt van steekproefbedrijven waaruit gekozen kan worden op basis van een aantal relevante variabelen. De informatie van het steekproefbedrijf wordt geacht van toepassing te zijn op het betreffende Landbouwtellingsbedrijf. Door aggregaties over Landbouwtellingsbedrijven worden vervolgens schattingen gemaakt voor kleine gebieden. Deze methode is verder ontwikkeld door Wil Hennen.

Ratioschatters

In een aantal gevallen wordt gebruikgemaakt van zogenaamde ratioschatters. Zo worden er bijvoorbeeld schattingen gemaakt van gegevens per varken. Deze schatting in combinatie met gegevens uit de Landbouwtelling worden gebruikt voor het maken van schattingen voor bepaalde gebieden. Hierbij worden alleen gemiddeldes geschat, aan de variantie van deze schatting wordt geen aandacht besteed.

Poststratificatie

Echte voorbeelden van poststratificatie zijn niet voorhanden. Wel zijn er enkele toepassingen waarbij een soort herweging plaatsvindt op basis van de verdeling in de populatie.

2.3.4 Toekomstig gebruik van methoden voor kleine deelgebieden

Ten aanzien van het toekomstig gebruik van methoden voor het schatten van kenmerken van kleine deelgebieden is het van belang om eerst stil te staan bij de vraag of er in de toekomst een toenemende vraag zal bestaan naar onderzoek op deelgebieden. Het antwoord op deze vraag loopt uiteen. Volgens sommige onderzoekers is er een duidelijke trend te constateren naar vragen omtrent kleine gebieden. Met name voor kleine sectoren is de behoefte aan betere methoden groot. Het Ministerie stelt steeds meer vragen over kleine takken. De indruk bestaat dat bij een goede marketing een aanzienlijke vraag bestaat naar onderzoek naar deelpopulaties. De indeling van deelpopulaties is in toenemende mate afhankelijk van de opdrachtgever. Ook komen er in toenemende mate vragen op provinciaal niveau. Wel geldt hierbij de kanttekening dat provinciën als moeilijke klanten worden ervaren. Sommigen hebben weinig geld over voor het doen van onderzoek of doen het onderzoek liever zelf. Bij SO wordt geen duidelijke trend geconstateerd. De vraag naar gebiedsgericht onderzoek fluctueert sterk over de afgelopen jaren. Tevens is geen duidelijke ontwikkeling te bespeuren naar nog kleinere regio's.

Als richtlijnen en leidraden zouden bestaan voor het gebruik van technieken voor het schatten op deelgebieden zouden die volgens de onderzoekers worden toegepast. Al wordt bij het mogelijke gebruik nog wel een aantal kanttekeningen geplaatst. Voor sommige regio's wordt het gebruik van dergelijke technieken als moeilijk ervaren, omdat de uitkomsten bij lange na niet stroken met de eigen expertkennis. Een combinatie van deze kennis, overige beschikbare informatie en schattingstechnieken kunnen wellicht wel leiden tot zinvolle toepassingen. In de huidige opzet wordt te veel uitsluitend gekeken naar de informatie die in het Informatienet besloten ligt.

Hiertoe zouden wel een aantal standaardtechnieken beschikbaar moeten komen. Het is te complex en het vergt te veel tijd als onderzoekers zich hierin moeten gaan verdiepen. Hoe meer gebruikgemaakt wordt van gegevens des te beter. Hierbij geldt wel de randvoorwaarde dat de kwantitatieve kennis bij een aantal afdelingen beperkt is. Men wil zich wellicht wel verdiepen in de materie, maar het belang voor het onderzoek moet heel duidelijk zijn.

Algemeen wordt genoemd dat de methoden binnen het LEI op een centrale plaats beschikbaar zouden moeten zijn; daarbij geldt de voorwaarde dat ze makkelijk toegankelijk

moeten zijn. Er bestaat dan wel degelijk belangstelling voor een intensiever gebruik van deze methoden. De behoefte aan ondersteuning in de vorm van een cursus is daarbij groot.

2.3.5 Algemene bevindingen

Wat in LEI-onderzoek meestal wordt geschat is het gemiddelde. Varianties komen nauwelijks tot niet aan de orde. Het niet specificeren van onzekerheden van de schattingen komt niet zozeer uit gemak, het zit meer besloten in de bedrijfscultuur. Daarnaast geldt dat er zelden naar de betrouwbaarheid van de uitkomsten wordt gevraagd. Naar aanleiding van de discussie omtrent de RIVM-modellen werd er recent wel eens naar gevraagd maar dan ook alleen nog door het RIVM zelf, LNV zal zeker niet vragen naar marges. De politiek kan niet leven met onzekerheden. In rapportages komen geen varianties of standaardfouten aan de orde. Dit is over het algemeen te moeilijk voor de lezers van de rapportages. De opdrachtgever wil geen moeilijke dingen; indien toch een indicatie moet worden gegeven van de spreiding is het makkelijker zoiets als min/max weer te geven. Een andere reden voor het niet aangeven van de betrouwbaarheid van schattingen is dat de modellen niet direct de mogelijkheid bieden om varianties te berekenen. In FES worden bijvoorbeeld de betrouwbaarheden niet uitgerekend.

In het verlengde van het niet berekenen van betrouwbaarheidsintervallen ligt het geringe gebruik van statistische toetsen. Als redenen voor het niet gebruiken van statistische toetsen worden genoemd: een gebrek aan tijd, gebrek aan kennis, de software is niet goed, projectleiders vragen er niet om, een gebrek aan ondersteuning en het feit dat dit niet in de opdracht wordt gevraagd. Ondanks het geringe gebruik van varianties, statistische toetsen en betrouwbaarheden wordt het wel als nuttig ervaren hier in de toekomst meer aandacht aan te besteden.

De gemiddeldes van deelpopulaties worden berekend als een gewogen gemiddelde, waarbij de weging plaatsvindt met de bedrijfsweging volgens het Informatienet. De schatting van het aantal bedrijven waarop iets van toepassing is, wordt als vaststaand beschouwd (som van de gewichten). Er wordt geen rekening gehouden met het feit dat dit een schatting is die met een bepaalde onzekerheid wordt omgeven. Net als voor schattingen voor de gehele populatie, worden er geen varianties en betrouwbaarheidsintervallen berekend voor de berekende gemiddeldes. Dit impliceert dat het moeilijk is verschillende groepen onderling of over een reeks van jaren te vergelijken.

Bij het berekenen van een directe schatter van het gemiddelde wordt gebruikgemaakt van de wegingsfactoren. Bij het gebruik van alternatieve methoden worden deze wegingsfactoren vaak weggelaten. Dit kan een verstorend effect op de resultaten hebben omdat op deze manier geen rekening wordt gehouden met de historische trekkingskansen. Niethomogene segmenten zullen hierdoor oververtegenwoordigd zijn. De gevonden resultaten zullen dus sterker worden beïnvloed door deze segmenten. Het is dan ook zeer de vraag of er nog sprake is van de nagestreefde representativiteit. Verder zijn er verschillen te constateren in de typering van bedrijven. In sommige onderzoeken gaat de onderzoeker uit van de typering volgens de Landbouwtelling. In andere wordt de typering in het Informatienet als uitgangspunt gekozen. Beide aanpakken kunnen tot fundamentele verschillen leiden.

2.4 Conclusies uit interviews

In dit vooronderzoek is geprobeerd een beeld te schetsen hoe en in welke mate op het LEI gebruik wordt gemaakt van steekproeven. De meest gebruikte steekproef is het Informatienet, daarnaast worden diverse andere steekproeven gebruikt.

Op basis van de steekproeven worden puntschattingen voor gemiddeldes berekend voor de gehele populatie en voor delen van de populatie. In de huidige onderzoekspraktijk op het LEI wordt weinig aandacht besteed aan de *betrouwbaarheid* van uitkomsten. De consequentie hiervan is dat men geen statistisch verantwoorde uitspraken kan doen bij het vergelijken van scenario's, groepen en jaren. Het voordeel van meer geavanceerde schattingsmethodes is juist gelegen in een toename van de betrouwbaarheid en niet de zuiverheid van de schatter. Zolang er dus geen omslag plaatsvindt van een punt-denken naar een interval-denken zal het moeilijk zijn de voordelen van alternatieve methoden duidelijk te maken. Indien de onderzoeker niet de voordelen ziet zal hij of zij minder geneigd zijn om zich te verdiepen in de materie. Gezien de hogere complexiteit van alternatieve schattingswijzen is deze verdieping wel een vereiste voor een statistisch verantwoord gebruik.

Ondanks deze constatering is er een groeiend toepassingsterrein voor methoden voor het schatten van kenmerken van kleine deelgebieden. Binnen AEOS is er een toenemende vraag naar schattingen voor regionale gebieden, ook binnen PPRF en MESO ziet men toepassing voor kleine deelgebieden. Binnen SO zou men graag beschikken over meer gegevens die in het REM-model kunnen worden gebruikt. Met name bestaat er bij SO een grote behoefte aan inkomensgegevens op kleine gebieden.

Er is een duidelijke vraag naar en behoefte aan (bij)scholing op het gebied van statistiek en steekproeftechnieken. Dit wordt binnen het instituut vaak onterecht vereenzelvigd met kwantitatieve kennis. Een kwantitatieve scholing of insteek wil niet per definitie zeggen dat iemand kennis heeft van statistiek en steekproeftechnieken. Gezien het belang van het gebruik van het Informatienet en andere steekproeven lijkt een uitbreiding van deze kennis onontbeerlijk voor het LEI.

3. Van concrete onderzoeksvraag tot statistische mogelijkheden

3.1 Inleiding

Een onderzoeker wordt geconfronteerd met de vraag een uitspraak te doen omtrent een bepaald kenmerk van een populatie. Een opdrachtgever wil bijvoorbeeld weten wat het gemiddelde inkomen van de Nederlandse agrariër is. Om deze vraag te beantwoorden kan een onderzoeker besluiten een steekproef uit alle agrariërs te trekken en vervolgens gegevens te verzamelen bij de steekproefelementen. Wanneer de gegevens zijn verzameld kan het gemiddelde inkomen van de agrariër in de steekproef worden berekend. Dit gemiddelde in de steekproef vormt een zo goed mogelijke schatting van het inkomen in de populatie.

Nadat de gegevens zijn verzameld ontstaan vaak aanvullende vragen, bijvoorbeeld wat het gemiddelde inkomen is in de tuinbouwsector. In plaats van gebruik te maken van een nieuwe steekproef kan de bestaande steekproef met gegevens over alle agrariërs worden hergebruikt om een uitspraak te doen over het gemiddelde inkomen van de tuinders. Indien de onderzoeker een steekproef hergebruikt om uitspraken te doen over delen van de populatie dan spreken wij over 'small area estimation'.

In het hedendaagse onderzoek wordt men meer en meer geconfronteerd met vragen die gericht zijn op deelpopulaties of kleine geografische gebieden. Gebiedsgericht onderzoek wordt alsmaar belangrijker. Het probleem dat in dergelijke gevallen vaak optreedt, is dat men slechts over een beperkt aantal directe waarnemingen beschikt. Bij het doen van uitspraken op basis van een gering aantal waarnemingen wordt men vaak geconfronteerd met zeer onbetrouwbare schattingen. De daadwerkelijke onbetrouwbaarheid zal afhankelijk zijn van de homogeniteit van de bedrijven in de te bestuderen kleine populatie. Indien alle bedrijven sterk op elkaar lijken zal het geen probleem zijn wanneer men slechts over een klein aantal waarnemingen beschikt. Bij meer heterogene onderzoekspopulaties zal het probleem wel optreden.

Er zijn in dergelijke gevallen twee oplossingsrichtingen denkbaar. Ten eerste kan getracht worden het aantal waarnemingen te vergroten door missende waarden in te vullen. Ten tweede kan men proberen de betrouwbaarheden van de schattingen te vergroten door alternatieve schattings- en stratificatiemethoden te hanteren. Men kan de waarnemingen bijvoorbeeld dusdanig poststratificeren en indelen zodat relatief homogene groepen ontstaan. Ook kan middels het gebruik van indirecte schatters de betrouwbaarheid worden vergroot doordat additionele informatie wordt gebruikt bij het maken van de schattingen.

Binnen de zojuist genoemde tweede oplossingsrichting zijn verschillende methoden mogelijk. Het vergroten van de betrouwbaarheid kan door schattingen te doen op basis van vooraf gedefinieerde aannames. Een aanname is bijvoorbeeld dat de doelvariabele verklaard wordt door een of meer andere variabelen. De betrouwbaarheid kan vergroot worden in het geval deze verklarende variabelen bekend zijn voor de gehele deelpopulatie. Een andere aanname kan worden gedaan omtrent de verhouding van de deelpopulatie ten

opzichte van de gehele populatie. Als de waarde die de variabele aanneemt op het kleine deelgebied niet in grote mate afwijkt van die van de gehele populatie, kunnen bijvoorbeeld alle steekproefelementen worden gebruikt voor het doen van schattingen op het kleine deelgebied.

Verschillende aannames worden geëxpliciteerd in een model. De kritiek op de modelgebaseerde aanpak berust op de aanname van het model. Indien het correcte model is gespecificeerd zullen betere schattingen kunnen worden gemaakt dan met de klassieke steekproef theorie. Echter, als het model niet correct is zal sprake zijn van een sterke bias en de schatting van de variantie zal te optimistisch zijn. Omdat men nooit zeker weet of het juiste model is gespecificeerd prefereren sommigen de modelvrije klassieke steekproeftheorie. Echter, bij het maken van schattingen voor kleine deelgebieden wordt men wel gedwongen aannames te maken. Gezien de redelijke werking van lineaire modellen in veel sociaal economische processen is het gebruik van modellen bij het maken van schattingen voor kleine gebieden verdedigbaar.

Kleine deelgebieden waarover onderzoeksvragen gesteld worden binnen het LEI hebben onder meer betrekking op: deelgebieden of -groepen zoals geografisch gebieden. Specialistische groepen zijn een ander voorbeeld, hierbij kan gedacht worden aan bijvoorbeeld de groep van kalvermesterijen.

3.2 Schattingsmethoden en de steekproef van het Informatienet

In het Informatienet wordt een gedetailleerde administratie bijgehouden van ruim 1.500 land- en tuinbouwbedrijven. Naast financieel-economische gegevens worden ook technisch-economische, milieueconomische en sociaal-economische gegevens van deze bedrijven vastgelegd. Het Informatienet wordt mede bijgehouden voor de Europese Unie. Daarnaast vormt het Informatienet de basis voor veel onderzoek zoals dat binnen het LEI wordt uitgevoerd. Op basis van de bedrijven in het Informatienet worden uitspraken gedaan over alle land- en tuinbouwbedrijven (of delen daarvan). Hierbij is het belangrijk dat de bedrijven die in het Informatienet zijn opgenomen, wat betreft belangrijke onderzoeksvariabelen representatief zijn voor de gehele populatie. Op deze manier kan men zelfs tot betere schattingen komen op basis van slechts een deel van de bedrijven. Bij een beperkt aantal bedrijven kan men veel nauwkeuriger en kwalitatief betere gegevens verzamelen dan wanneer men alle bedrijven zou moeten bezoeken en onderzoeken.

Een belangrijk criterium is de representativiteit van bedrijven in het Informatienet voor de bedrijven in de gehele populatie. De vraag is hoe zorg gedragen kan worden voor deze representativiteit. Hiertoe wordt gebruikgemaakt van een disproportionele gestratificeerde steekproef. Een gestratificeerde steekproef wil zeggen dat de populatie in een aantal groepen wordt opgedeeld en dat bedrijven uit elk van de afzonderlijke groepen worden geselecteerd. De kenmerken op basis waarvan de groepsindeling tot stand komt, moeten belangrijke kenmerken van de populatie zijn, zodanig dat bedrijven die in een groep terechtkomen veel op elkaar lijken wat betreft belangrijke doelvariabelen voor onderzoek. Door gebruik te maken van deze groepsindeling weet men zeker dat bedrijven uit alle groepen in de steekproef terechtkomen. Disproportioneel wil zeggen dat niet alle bedrijven een even grote kans hebben om in de steekproef terecht te komen. Groepen die heel homo-

geen zijn wat betreft belangrijke kenmerken, dat wil zeggen dat de bedrijven sterk op elkaar lijken in deze kenmerken, hebben een lagere trekkingskans. Immers, als alle bedrijven (bijna) identiek zijn, kan men op basis van een beperkt aantal waarnemingen een redelijke uitspraak doen (in het extreme geval dat alle bedrijven identiek zijn is één waarneming voldoende om een exacte uitspraak over de hele groep te doen). Bij minder homogene groepen zal men meer bedrijven moeten opnemen om betrouwbare uitspraken te doen. De variabelen op basis waarvan de groepen worden ingedeeld hebben dus een belangrijke invloed op de representativiteit van de steekproef. In het Informatienet werden de groepen t/m het jaar 2000 ingedeeld op basis van het bedrijfstype, de regio, NGE-klassen (hierbij staat NGE voor Nederlandse grootte-eenheid) en meer verfijnd naar de bedrijfsomvang in hectares, de leeftijd en een fijnmaziger regio-indeling. In 2001 is deze indeling in strata enkel nog gebaseerd op het bedrijfstype en NGE (Vrolijk en Lodder, 2002). Reden hiervoor is dat de steekproef flexibeler gebruikt kan worden. Onderzoeksvragen bij het LEI vallen binnen een breed kader, dat veel verschillende doelvariabelen betreft. Als naar bepaalde doelvariabelen gestratificeerd wordt, is er niets te zeggen over de homogeniteit ten aanzien van andere variabelen.

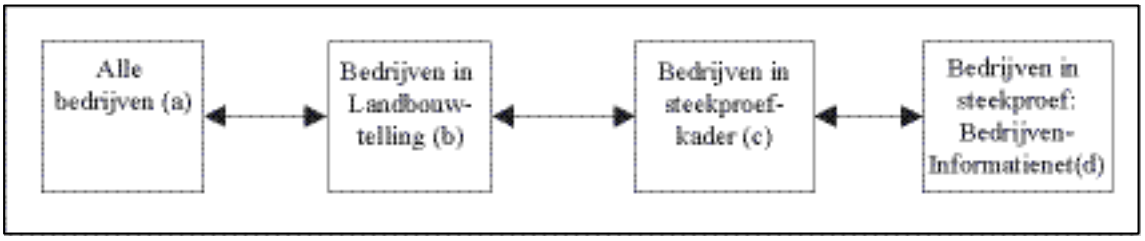
Door op deze manier de bedrijven te selecteren kunnen uitspraken worden gedaan over de hele populatie. Op basis van de bedrijven uit een groep kunnen uitspraken worden gedaan voor die groep, door de gestratificeerde steekproefopzet zijn bedrijven uit alle groepen opgenomen en kunnen uitspraken worden gedaan over alle groepen. Alle groepen tezamen vormen de gehele populatie. In het Informatienet is dit gerealiseerd door aan elk bedrijf een gewicht toe te kennen. Het gewicht wordt berekend door het aantal bedrijven in de populatie (in een bepaalde groep) te delen door het aantal bedrijven in de steekproef (in die zelfde groep).

Op deze manier wordt geprobeerd het Informatienet zo representatief mogelijk te maken voor de gehele populatie. Hierbij moeten twee kanttekeningen worden geplaatst. De eerste is dat de representativiteit is gewaarborgd ten aanzien van de kenmerken op basis waarvan de groepen zijn ingedeeld. Dit wil nog niet zeggen dat de steekproef voor elke willekeurig te bedenken variabele representatief is. Ten tweede geldt dat de populatie waarvoor het Informatienet representatief zou moeten zijn, niet betrekking heeft op alle landbouw en tuinbouwbedrijven (a in figuur 3.1). Bedrijven die te klein zijn of te laat zijn geteld maken geen deel uit van de Landbouwtelling (b). De steekproefpopulatie (of eigenlijk steekproefkader) (c) werd t/m het jaar 2000 gevormd door de bedrijven die in de Landbouwtelling zijn opgenomen en een omvang hebben van minimaal 16 NGE en maximaal 800 NGE. Vanaf 2001 is de EGE (Europese grootte-eenheid) bepalend, bedrijven met een omvang van minimaal 16 EGE en maximaal 1.200 EGE vormen het steekproefkader. Uit dit steekproefkader (zie figuur 3.1) wordt de daadwerkelijke steekproef getrokken (d).

In het huidige gebruik van het Informatienet wordt alle kennis over de populatie gestopt in de opzet van de steekproef. Kennis omtrent (onder andere) sectoren en EGE's wordt in de steekproef gebruikt om te komen tot een gestratificeerde steekproef. Op basis van deze steekproef wordt zonder gebruik te maken van aanvullende informatie een schatting gemaakt van de onderzoeksvariabele. Dit is de zogenaamde directe schatter. Het is ook mogelijk de kennis over de populatie pas te gebruiken bij het maken van de schattingen. De steekproefopzet kan dan relatief eenvoudig worden gehouden. Door het gebruik van aanvullende informatie, kan de betrouwbaarheid en validiteit van de schattingen toe-

nemen. Deze aanpak heeft als voordeel dat voor elk onderzoek aanvullende informatie kan worden gebruikt die is toegespitst op de relevante doelvariabelen in dat onderzoek. Bij het gebruik van aanvullende informatie voor het maken van schattingen wordt gesproken van een indirecte schatting.

Het voordeel van het gebruiken van de kennis in de fase van het schatten is dat het flexibeler is. Voor elk afzonderlijk onderzoek kan worden nagegaan wat de beste additionele informatie is die kan worden gebruikt om de schatting te verbeteren. Bij het gebruik van deze kennis in de steekproefopzet wordt de opzet afgestemd op een beperkt aantal doelvariabelen. Voor deze variabelen zal een dergelijke opzet voordelen bieden, voor het schatten van andere variabelen die minder samenhangen met de stratificatievariabelen kan stratificatie nadelig zijn.

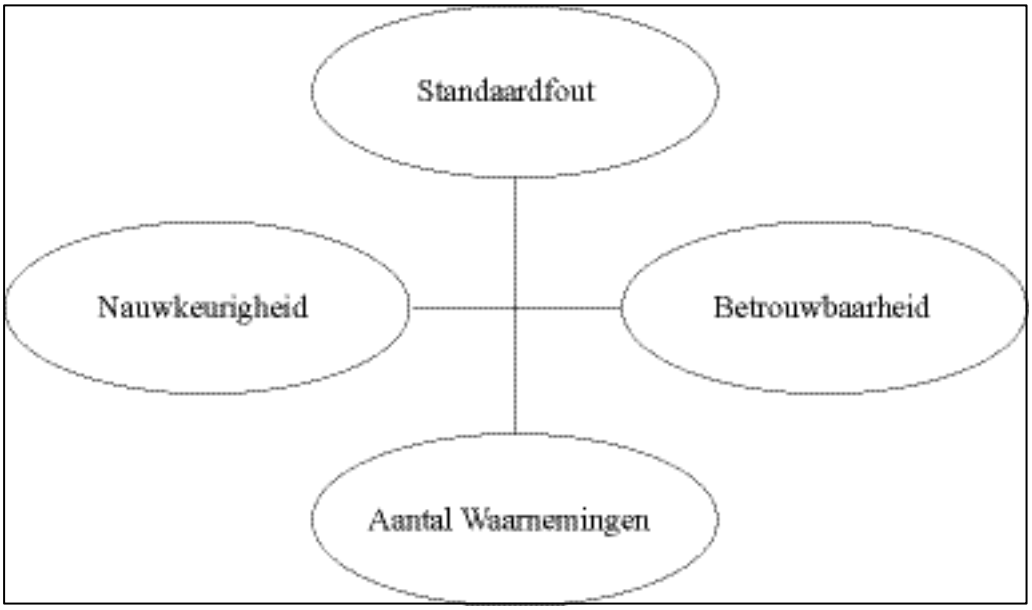


Figuur 3.1 Relatie steekproef en totale populatie

3.3 Beoordelen kwaliteit van steekproef en schattingen

Bij het beoordelen van de kwaliteit van een steekproef spelen verschillende aspecten een rol. Voor wat betreft de statistische kwaliteit zijn termen als de betrouwbaarheid, de nauwkeurigheid, het aantal steekproefbedrijven, de standaardfout en de representativiteit van groot belang. Representativiteit staat enigszins los. De eerste vier termen zijn niet los van elkaar te beoordelen. Het is onmogelijk richtlijnen ten aanzien van het beoordelen van een van de termen te geven zonder de overige in beschouwing te nemen. De samenhang is gevisualiseerd in figuur 3.2. De hierin genoemde aspecten zijn:

- de standaardfout is de standaarddeviatie van het steekproefgemiddelde;
- het aantal waarnemingen is het aantal steekproefbedrijven;
- de betrouwbaarheid wordt uitgedrukt in een percentage. Bijvoorbeeld dat men met 95% betrouwbaarheid wil stellen dat het populatiegemiddelde in het betrouwbaarheidsinterval zal vallen. De betrouwbaarheid hangt samen met de standaardfout en het aantal waarnemingen;
- de nauwkeurigheid heeft betrekking op de breedte van het betrouwbaarheidsinterval.



Figuur 3.2 Statistische kwaliteitsaspecten

Men kan van tevoren eisen stellen aan bijvoorbeeld de nauwkeurigheid of het aantal steekproefbedrijven. Als men eisen stelt aan het aantal bedrijven dan zal dit van invloed zijn op de nauwkeurigheid en de standaardfout. Als men eisen stelt aan de standaardfout dan zal dit consequenties hebben voor het minimale aantal waarnemingen. De aspecten kunnen dus niet afzonderlijk worden beoordeeld zonder naar de overige consequenties te kijken.

Representativiteit heeft betrekking op de mate waarin de steekproef een goede afspiegeling is van de onderzoekspopulatie. Dit kan getoetst worden door na te gaan in hoeverre significante verschillen bestaan tussen de steekproef en de onderzoekspopulatie. Zo kan bijvoorbeeld worden gekeken of het gemiddelde aantal hectares in de steekproef gelijk is aan die in de populatie. Ook kan getoetst worden of de verdeling in groepen gelijk is, bijvoorbeeld of het percentage akkerbouwbedrijven in de steekproef gelijk is aan die in de populatie.

De verdeling van alle mogelijke populatiegemiddeldes is normaal verdeeld (aanname gebaseerd op de centrale limietstelling) zelfs als de onderliggende x-waarden niet normaal verdeeld zijn. Op basis van het gevonden gemiddelde en de standaardfout kan een betrouwbaarheidsinterval voor het gemiddelde worden berekend. In veel gevallen wordt uitgegaan van een betrouwbaarheidsinterval van 95%. De hierbij behorende z-waarde bedraagt 1,96 [1]. Men kan vervolgens stellen dat het werkelijke populatiegemiddelde met 95% zekerheid tussen het gevonden gemiddelde plus of min 1,96 maal de standaardfout zal liggen. Bij een hogere standaardfout zal men dus geconfronteerd worden met een groter betrouwbaarheidsinterval en dus met minder nauwkeurige uitspraken.

[1] Voor steekproeven kleiner dan 50 wordt soms aanbevolen om de t-verdeling te gebruiken. Deze geeft een iets breder betrouwbaarheidsinterval dan de normale z-verdeling (Thompson, 1992).

Het rapporteren van de standaardfout is van belang om de betrouwbaarheid van de schatting weer te geven. Deze betrouwbaarheid is met name van belang indien men verschillende gemiddeldes wil vergelijken. Stel men vindt in het ene jaar een steekproefgemiddelde van 85 en het daaropvolgende jaar van 87. Het is wellicht aantrekkelijk te concluderen dat er een stijging is opgetreden. Echter, zoals in het voorgaande is geïllustreerd is het gevonden gemiddelde een van de vele mogelijke uitkomsten die tot stand komt afhankelijk van de elementen die in de steekproef zijn opgenomen. Rekeninghoudend met de gevonden standaardfouten kan de conclusie dat de werkelijke populatiegemiddeldes in beide jaren verschillend zijn, wellicht niet worden onderbouwd. Afhankelijk van de waarde van de standaardfout kan het verschil tussen beide jaren berusten op een toevalligheid of op een daadwerkelijke stijging. Bij het vergelijken van de betrouwbaarheidsintervallen voor beide jaren kan blijken dat deze een sterke overlap vertonen. Beide gevonden jaargemiddelden in de steekproef kunnen dan gebaseerd zijn op eenzelfde populatiegemiddelde. Van een significante stijging hoeft dan geen sprake te zijn.

4. Methoden voor het schatten van kenmerken van kleine deelgebieden

In de volgende paragrafen worden diverse methoden voor het schatten van deelgebieden behandeld. De methoden die aan de orde komen zijn:
- directe schatters;
- ratioschatters;
- regressieschatters;
- Bayesiaanse schatters;
- poststratificatie;
- imputatie middels modelschatters;
- hot deck-procedures.

De directe schatters maken uitsluitend gebruik van de steekproef gegevens. De overige methoden maken gebruik van aanvullende informatie uit andere bronnen of van een onderliggend model.

4.1 Directe schatters

Bij het gebruik van steekproeven wil de onderzoeker op basis van een deel van de populatie (de steekproef) uitspraken doen over een kenmerk van de hele populatie. Bij het gebruik van een directe schatter wordt het gemiddelde direct bepaald aan de hand van de beschikbare waarnemingen in de steekproef. Zonder verdere aannames [1] te hoeven maken kan op basis van de waarnemingen een schatting worden gemaakt.

4.1.1 Directe schatter op basis van een aselecte steekproef

4.1.1.1 Theorie

Het gemiddelde in de steekproef is afhankelijk van de elementen die in een steekproef zijn opgenomen. Het is dan ook eenvoudig in te zien dat er een verschil kan zitten in het gemiddelde van het kenmerk in de steekproef en het werkelijke gemiddelde in de populatie. Dit verschil wordt de steekproeffout genoemd. Door het gebruik van steekproefprocedures kan deze fout worden gekwantificeerd.

Het voorgaande zal aan de hand van een voorbeeld worden toegelicht. Stel men wil de gemiddelde doorsnede van een tomaat in een doos met 1.000 tomaten bepalen op basis van een steekproef van 20 tomaten. Het gevonden gemiddelde is afhankelijk van de toevallige selectie van tomaten. Het is theoretisch denkbaar dat men toevallig de 20 grootste tomaten selecteert en meet. Dit zal tot een groter gemiddelde leiden dan wanneer men toe-

[1] Afgezien van de aanname van normaliteit.

vallig de 20 kleinste selecteert. Afhankelijk van de tomaten die in de steekproef terechtkomen, komt men op een bepaald gemiddelde. Het gemiddelde (1) is de directe schatter van het daadwerkelijke populatiegemiddelde.

$$\bar{Y}_D = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (1)$$

Het is nu natuurlijk de vraag hoe betrouwbaar deze schatting is. Hiertoe wil men inzicht hebben in de standaardfout oftewel de standaarddeviatie van het gemiddelde. Om de standaarddeviatie van het gemiddelde vast te stellen zou men voor alle mogelijke combinaties van 20 tomaten het gemiddelde kunnen berekenen en vervolgens over al deze waarden de standaarddeviatie kunnen uitrekenen. Het moge duidelijk zijn dat we normaal gesproken gebruikmaken van een steekproef en niet alle mogelijke steekproeven trekken. Aangetoond kan worden dat de standaardfout, de wortel van de variantie (2), kan worden berekend door de variantie van een kenmerk in de populatie te delen door het aantal waarnemingen in de steekproef, n (in geval van een enkelvoudige aselecte trekking) .[1]

$$V(\bar{Y}_D) \approx \frac{s_y^2}{n} \qquad (2)$$

De populatievariantie

$$S_y^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)}{N-1} \qquad (3)$$

waarbij N het aantal populatie elementen, is echter vaak onbekend, maar kan worden geschat door de variantie van de elementen in de steekproef te berekenen. Een zuivere schatter hiervoor is:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y}_D)^2}{n-1} \qquad (4)$$

De variantie (5) van de directe schatter voor het gemiddelde kan nu worden gevonden door de variantie van de elementen in de steekproef te delen door het aantal waarnemingen in de steekproef.

$$n(\bar{Y}_D) \approx \frac{s_y^2}{n} \qquad (5)$$

[1] De correctiefactor ((N–n)/N = 1-f) voor de eindigheid van de populatie wordt hier niet meegenomen.

Hieruit volgt dat de standaardfout, die berekend kan worden door de wortel te trekken uit de variantie van de directe schatter, af zal nemen indien de standaarddeviatie afneemt of indien het aantal waarnemingen toeneemt. De eerste gevolgtrekking is aannemelijk te maken door het extreme voorbeeld waarin alle tomaten (bijna) dezelfde doorsnede hebben. In dit extreme geval zal de variantie (bijna) nul zijn. Het meten van één tomaat volstaat voor het maken van een goede schatting van het populatiegemiddelde. Indien de tomaten sterk uiteenlopen zal men meer waarnemingen nodig hebben om zinvolle uitspraken te kunnen doen. Naarmate het aantal waarnemingen toeneemt zal het waarschijnlijker zijn dat de steekproef een goede afspiegeling vormt van de populatie.

Als rekening gehouden wordt met de correctiefactor voor de eindigheid van de populatie ziet de variantie er als volgt uit:

$$n(\bar{Y}_D) = \frac{(1-f)}{n} s_y^2 \qquad (6)$$

De correctiefactor $(1-f) = \frac{N-n}{N}$ krijgt een waarde 1 als N naar oneindig gaat bij relatief kleine steekproeven.

4.1.1.2 Toepassing van de directe schatter

Stel men wil een schatting maken van de arbeidsopbrengst ondernemer van de mestkalverijen. In het Informatienet zijn slechts een beperkt aantal mestkalverijen opgenomen.

De waarden voor de doelvariabele Y en de afwijkingen van het gemiddelde in het kwadraat zijn hieronder weergegeven. Uit tabel 4.1 zijn de directe schatter voor het gemiddelde en de variantie voor deze schatter af te leiden .[1]

$$\bar{Y}_D = \frac{1}{n}\sum_{i=1}^{n} y_i = 73.086$$

$$n(\bar{Y}_D) = \frac{(1-f)}{n} s_y^2 = 254.772.886$$

De bijbehorende standaardfout heeft een waarde 15.962.

[1] De weergegeven waarden van verschillende schattingen in het rapport zijn in Excel berekend en afgerond weergegeven in dit rapport; om deze reden kunnen bepaalde waarden in kleine mate afwijken als deze nagerekend worden met behulp van andere weergegeven waarden.

Tabel 4.1 Gegevensvoorbeeld directe schatter

Arbeidsopbrengst ondernemer (y_i)	Afwijking van het gemiddelde in het kwadraat $(y_i - \bar{Y}_D)^2$
179.716	11.369.946.745
9.867	3.996.647.982
71.109	3.908.717
94.430	455.564.303
118.874	2.096.536.583
9.483	4.045.347.666
116.438	1.879.391.775
122.384	2.430.288.109
39.082	1.156.275.254
109.502	1.326.121.588
16.858	3.161.593.339
238.166	27.251.390.678
14.012	3.489.743.102
28.054	2.027.885.313
1.491	5.125.850.844
203	5.311.938.630
77.246	17.305.204
231.612	25.130.477.578
6.782	4.396.226.731
19.043	2.920.650.996
30.455	1.817.406.221

4.1.1.3 Evaluatie directe schatter

De waarde van de directe schatter kan gemakkelijk berekend worden. Een nadeel van de directe schatter op een klein deelgebied is de onbetrouwbaarheid van de schatting van het gemiddelde of het totaal wanneer doelvariabele een grote variantie heeft. Als de doelvariabele in de populatie een variantie 0 heeft, zal een schatting op basis van een klein aantal waarnemingen nog steeds betrouwbaar zijn. Als echter de variantie van de doelvariabele groot is en het aantal waarnemingen klein, is de directe schatter zeer onbetrouwbaar.

4.1.2 Directe schatter op basis van een gestratificeerde steekproef

4.1.2.1 Theorie

In het Informatienet wordt gebruikgemaakt van een gestratificeerde steekproef. Dit betekent dat er groepen worden samengesteld die relatief homogeen zijn. Stratificatie heeft als voordeel dat de betrouwbaarheid van de schattingen toeneemt. Dit komt doordat de variantie van de schatting wordt bepaald door de variantie binnen de groepen en niet door de variantie tussen de groepen.

Daarnaast heeft stratificatie het voordeel dat de representativiteit van de steekproef verbetert. Door een aantal strata te definiëren en vervolgens een steekproef uit deze strata te trekken, weet men zeker dat bedrijven uit alle strata in de steekproef terecht zullen ko-

men. Een bijkomend voordeel ten aanzien van de representativiteit treedt op bij het voorkomen van non response. Een niet responderend bedrijf uit een bepaald stratum kan worden vervangen door een ander bedrijf in hetzelfde stratum. Indien de strata redelijk homogeen zijn, kan men er van uitgaan dat het vervangende bedrijf lijkt op het niet responderende bedrijf, wat betreft de stratificatievariabelen.

In geval van stratificatie zal het gemiddelde en de standaardfout worden berekend op basis van de gemiddeldes en standaardfouten van de afzonderlijke strata.

Een zuivere schatter voor het steekproefgemiddelde in stratum h, het stratumgemiddelde, is de directe schatter op dit deelgebied:

$$\bar{Y}_{Dh} = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \tag{7}$$

Waarbij n_h het aantal steekproefelementen in stratum h weergeeft. Het gemiddelde van de populatie is een gewogen gemiddelde van de individuele stratumgemiddelden. De gewichten zijn gelijk aan het aantal populatie-elementen in het stratum gedeeld door het aantal elementen in de gehele populatie. De schatter op basis van een gestratificeerde steekproef is:

$$\bar{Y}_S = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{Y}_{Dh} \tag{8}$$

N_h is het aantal populatiebedrijven in stratum h. Doordat de schatters voor de stratumgemiddelden zuiver zijn, is ook de schatter voor het gemiddelde van de populatie, als een gewogen som van de geschatte stratumgemiddelden, een zuivere schatter voor het werkelijke populatiegemiddelde.

De formule voor het populatiegemiddelde is te herschrijven als:

$$\bar{Y}_S = \frac{1}{N}\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi} \tag{9}$$

$\frac{N_h}{n_h}$ wordt gedefinieerd als de ophoogfactor. Dit is het gewicht van een steekproefbedrijf zoals gedefinieerd in de databank van het Informatienet. Als de steekproeven onafhankelijk in de strata getrokken worden dan is de variantie van de stratificatieschatter voor het gemiddelde gelijk aan de som van de gewogen varianties van de stratumgemiddelden. De variantie hangt dus alleen af van de varianties binnen de strata:

$$V(\bar{Y}_S) = \sum_{h=1}^{H} W_h^2 V(\bar{Y}_{Dh}) \tag{10}$$

Waarbij W_h wordt gedefinieerd als de fractie populatiebedrijven in stratum h.

$$W_h = \frac{N_h}{N} \tag{11}$$

Invullen van de volgende formule [1]

$$n(\bar{Y}_{Dh}) = \frac{N_h - n_h}{N_h}\frac{s_{yh}^2}{n_h} \tag{12}$$

in (10) levert een zuivere schatter voor de variantie van het populatiegemiddelde:

$$n(\bar{Y}_S) = \frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_{yh}^2}{n_h} = \frac{1}{N^2}\left[\sum_{h=1}^{H}\frac{1}{n_h}N_h^2 s_{yh}^2 - \sum_{h=1}^{H} N_h s_{yh}^2\right] \tag{13}$$

Om de populatievariantie in stratum h te kunnen berekenen moeten er ten minste twee waarnemingen in dat stratum zijn. Een complicerende factor kan optreden indien de strata niet samenvallen met de doelgroepen. Dit betekent dat van de n_h elementen in een stratum h, n_{hj} elementen tot de doelgroep horen. Een schatter voor de variantie van het gemiddelde in groep j wordt dan:

$$V(\bar{Y}_{Sj}) = \frac{1}{N_j}\sum_{h=1}^{H}\frac{N_h^2(1-f_h)}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h}\left(y_{hij} - \bar{Y}_{Dhj}\right)^2 + n_{hj}\left(1 - \frac{n_{hj}}{n_h}\right)\left(\bar{Y}_{Dhj} - \bar{Y}_j\right)^2\right] \tag{14}$$

Hierbij is $f_h = \frac{n_h}{N_h}$ en $\bar{Y}_{Dhj} = \sum_{i=1}^{n_{hj}}\frac{y_{hij}}{n_{hj}}$.

De tussenvariantie veroorzaakt door verschillen in de gemiddelden van strata wordt slechts ten dele verwijderd. Deze toevoeging van de tussenvariantie is gering indien n_h niet veel verschilt van n_{hj}, dus wanneer het aantal waarnemingen behorende tot de groep bijna gelijk is aan het totaal aantal waarnemingen in stratum h.

4.1.2.2 Toepassing van de directe schatter in een gestratificeerde steekproef

Ter illustratie zal het voorbeeld besproken in paragraaf 4.1.1.2 gebruikt worden om een directe schatter te berekenen in een gestratificeerde steekproef. Een voorbeeld van stratificatie in het Informatienet is de stratificatie naar NGE-klassen. Tot het jaar 2001

[1] De variantie binnen een bepaald stratum is gelijk aan de variantie van de directe schatter van het gemiddelde binnen dat stratum. In deze formule wordt in tegenstelling tot formule 2, wel rekening gehouden met de eindigheid van de populatie, door de factor (N-n)/N mee te nemen. In het vervolg zal deze factor altijd meegenomen worden.

werden vier NGE-klassen onderscheiden. Uitgegaan wordt van de verdeling van de bedrijven over de vier strata zoals weergegeven in tabel 4.2.

Tabel 4.2 Gegevensvoorbeeld stratificatieschatter

NGE-Klasse	16-37	37-66	66-123	123-800
Verdeling waarnemingen over de verschillende strata (arbeidsopbrengst ondernemer)	203 1.491 6.782 9.483 9.867	14.012 16.858 19.043 28.054 30.455 39.082	71.109 77.246 94.430 109.502 116.438 118.874 122.384	179.716 231.612 238.166
$\overline{Y}_{Dh}$	5.565	24.584	101.426	216.498
S_h^2	20.171.884	91.772.497	431.164.882	1.025.425.372
S_h	4.491	9.579	20.764	32.022
N_h	5	6	7	3
N_h	289	278	261	127
W_h	0,24	0,29	0,33	0,14
$n(\overline{Y}_{Dh})$	3.964.578	14.965.299,3	59.943.010,4	333.734.242
$W_h^2 n(\overline{Y}_{Dh})$	1.847.291	7.776.701	32.204.581	18.134.465

Uit tabel 4.2 zijn ook de stratificatieschatter en de variantie van deze schatter af te leiden.

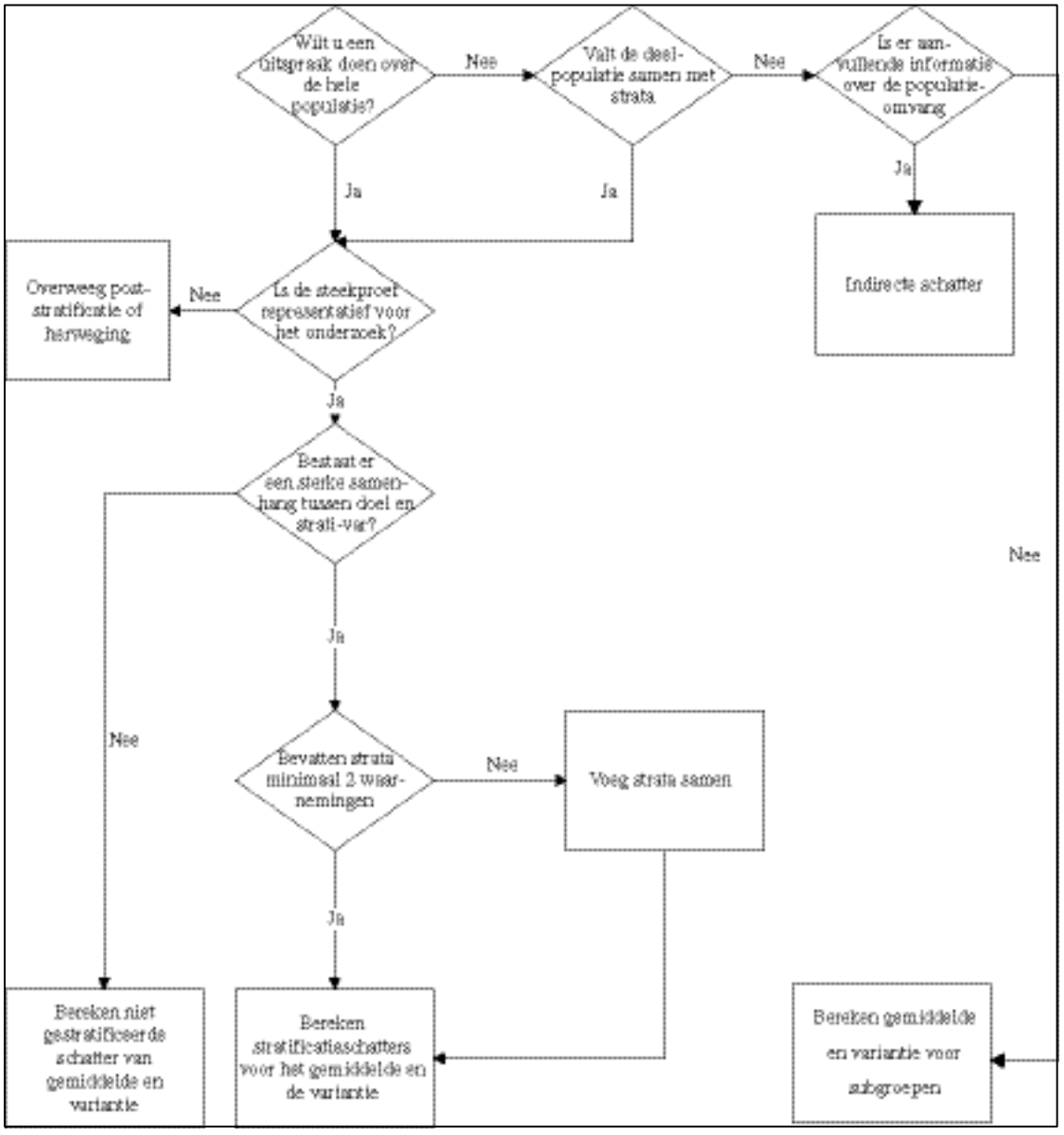
$$\overline{Y}_S = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{Y}_{Dh} = 65.351$$

$$n(\overline{Y}_S) = \sum_{h=1}^{H} W_h^2 n(\overline{Y}_{Dh}) = 59.963.039$$

De bijbehorende standaardfout is 7.744. Uit dit voorbeeld blijkt dat de variantie veel kleiner is dan de variantie van de directe schatter in de aselecte steekproef. In het geval de stratificatievariabele gelijk is aan of in hoge mate gecorreleerd met de doelvariabele zal de variantie afnemen als gewerkt wordt met een gestratificeerde steekproef. Is er echter geen duidelijk verband tussen de stratificatievariabele en de doelvariabele dan kan de variantie van de stratificatieschatter zelfs groter zijn dan die van de directe schatter in de aselecte steekproef.

4.1.2.3 Evaluatie directe schatter in een gestratificeerde steekproef

Als de steekproef vooraf gestratificeerd is, kan een directe schatter die hier rekening mee houdt een kleinere variantie en dus een grotere betrouwbaarheid opleveren. Voorwaarde hierbij is dat de stratificatie heeft plaatsgevonden naar de doelvariabele of een variabele die in hoge mate correleert met de doelvariabele. Het aantal muizen dat rondloopt op een boerderij zal naar alle waarschijnlijkheid niet samenhangen met het aantal NGE's van dat bedrijf. Het heeft dan ook geen zin gebruik te maken van de directe schatter die rekening houdt met de stratificatie vooraf, aangezien de verschillende strata niet homogeen zijn.



Figuur 4.3 Besluitvormingsschema voor het gebruik van directe schatters

4.2 Ratioschatters

4.2.1 Theorie

Het is gebruikelijk om schattingen van gemiddelden en totalen te maken op basis van de waarden van de variabele zoals die in de steekproef zijn waargenomen. De totale melkproductie kan op deze manier worden geschat door de melkproductie op de steekproefbedrijven op te hogen naar de populatie middels de in het Informatienet beschikbare gewichten.

Indien een hulpvariabele beschikbaar is die in hoge mate correleert met de doelvariabele, dan kan deze hulpvariabele worden gebruikt voor het maken van betrouwbaardere schattingen. Als men bijvoorbeeld een schatting wil maken van de totale melkproductie dan kan gebruik worden gemaakt van het gegeven dat de melkproductie op een bedrijf sterk zal correleren met het aantal koeien op dat bedrijf. Bij het gebruik van de ratioschatter geldt de voorwaarde dat het gemiddelde of totaal van deze hulpvariabele voor de hele populatie bekend moet zijn. Voor het aantal koeien is dit het geval, op basis van de Landbouwtelling kan het totaal aantal koeien worden vastgesteld. De reden waarom deze indirecte schatting betrouwbaarder kan zijn dan een directe schatting is dat de verhouding tussen twee variabelen stabieler kan zijn dan de variabelen afzonderlijk. De melkproductie op verschillende boerderijen kan sterk uiteenlopen. Een directe schatting zou dan ook een hoge variantie laten zien. De melkproductie zal echter sterk afhankelijk zijn van het aantal koeien. De verhouding productie per koe zal een kleinere spreiding laten zien dan de spreiding in de melkproductie of het aantal koeien over de steekproefbedrijven. Indien men op basis van andere bronnen gegevens heeft over het totaal aantal koeien dan kan een veel nauwkeuriger uitspraak over de totale melkproductie in Nederland worden gedaan.

Een bijkomend voordeel van het gebruik van ratioschatters is dat de representativiteit wordt verhoogd. Stel dat in de steekproef vooral kleine bedrijven zijn opgenomen. Doordat in de indirecte schatter van de totale melkproductie rekening wordt gehouden met het aantal koeien op de steekproefbedrijven ten opzichte van het aantal koeien in Nederland wordt automatische gecorrigeerd voor de omvang van de bedrijven. De verhouding melkproductie per koe wordt vermenigvuldigd met de uit een andere bron bekende aantal koeien. Een directe schatter zou in dit geval tot een onderschatting van de totale melkproductie leiden.

Bij de ratioschatter wordt dus gebruikgemaakt van een hulpvariabele die een sterke correlatie vertoont met de doelvariabele. Van deze hulpvariabele moet het populatietotaal of -gemiddelde bekend zijn. Het doel van deze schatter is een verhoogde precisie door gebruik te maken van de correlatie tussen de doel- en hulpvariabele.

$$\bar{Y}_R = R\bar{X} = \frac{Y}{X}\bar{X} = \frac{\bar{Y}_D}{\bar{X}_D}\bar{X} \qquad (15)$$

Y en X zijn de totalen in de steekproef. Uitgaande van het hiervoor beschreven voorbeeld is Y gelijk aan de totale melkproductie op de steekproefbedrijven en X is het totaal aantal melkkoeien in de steekproef. $\bar{X}$ is het gemiddelde aantal koeien op een bedrijf in de populatie en $\bar{Y}$ is de resulterende schatting van de gemiddelde melkproductie op een be-

drijf. De ratio R kan berekend worden aan de hand van directe schatters van gemiddelden van de variabelen of aan de hand van de totalen van deze variabelen. Als de ratio $\frac{y_i}{x_i}$ redelijk constant is voor alle eenheden dan heeft de ratioschatter een hogere precisie dan een directe schatter.

Het gebruik van de ratioschatter levert geen problemen op indien de steekproef voldoende groot is. Als vuistregel geldt dat steekproefomvang groter dan 30 is en de coëfficiënt van variantie voor zowel x als y kleiner dan 10% zijn.

Het berekenen van de variantie van deze schatting is complexer. Een deel van de variantie komt voort uit het gebruik van de steekproef een ander deel van de variantie komt voort uit de modelspecificatie. Voor het doen van een uitspraak over de variantie van het gemiddelde moet een combinatie van beide bronnen van variantie worden genomen.

De variantie voortvloeiende uit de steekproef heeft tot nu toe centraal gestaan. Als echter met aannames en modellen gewerkt wordt, gaat de modelvariantie ook een rol spelen. De voorspelde waarde $\overline{Y}_R$ is gegeven het model de best mogelijke voorspelling voor bedrijf i. Het is echter onwaarschijnlijk dat de samenhang tussen de waarden van de doel- en de verklarende variabele exact voldoet aan $y_i = Rx_i$. Zowel de waargenomen waarden als de niet waargenomen waarden zullen deels groter zijn dan Rx_i en deels kleiner. Een maatstaf voor deze spreiding is de standaardfout van de schatting (standard error of the estimate).

Deze standaardfout wordt voor de ratioschatter op de volgende wijze berekend:

$$\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - Rx_i)^2}{n-1}} \qquad (16)$$

Waarbij n het aantal waarnemingen weergeeft en e_i het verschil tussen y_i en Rx_i. Van de totale variantie in y_i wordt een deel verklaard door het model en een deel niet. Het deel dat niet verklaard wordt door het model, is de standaardfout van de schatting.

De totale variantie van de schatter, waarbij ook rekening gehouden wordt met de steekproefvariantie is:

$$V(\overline{Y}_R) = \frac{1-f}{n}\left[\frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1}\right] \qquad (17)$$

Waarbij f wordt gedefinieerd als de fractie steekproefbedrijven, populatiebedrijven (f = n/N). Omdat echter geen waarden van y_i bekend zijn voor de hele populatie, zal de variantie benaderd moeten worden. Een zuivere schatter voor de variantie ziet er als volgt uit:

$$n(\bar{Y}_R) = \frac{1-f}{n}\left[\frac{\sum_{i=1}^{n}(y_i - Rx_i)^2}{n-1}\right] = \frac{1-f}{n}\left[s_y^2 + R^2 s_x^2 - 2Rrs_x s_y\right] \quad (18)$$

Hierbij is r een schatting van de correlatiecoëfficiënt

$$r = \frac{s_{xy}}{s_x s_y} \quad (19)$$

tussen de doel- en de hulpvariabele. Een schatting voor de co-variantie tussen x en y, s_{xy}, is s_{xy}:

$$s_{xy} = \frac{\sum_{i=1}^{n}(y_i - \bar{Y}_D)(x_i - \bar{X}_D)}{n-1} \quad (20)$$

Een belangrijke eis voor het toepassen van de ratioschatter is dat de hulpvariabele gecorreleerd moet zijn met de doelvariabele. Des te hoger deze correlatie des te voordeliger is het gebruik van de ratioschatter. Door de variantie van de ratioschatter te vergelijken met de variantie van de directe schatter

$$n(\bar{Y}_D) = \frac{1-f}{n}s_y^2 \quad (21)$$

kan worden afgeleid wanneer de ratioschatter resulteert in een lagere variantie. Hiervoor geldt dat $s_y^2 + R^2 s_x^2 - 2Rrs_x s_y$ kleiner moet zijn dan s_y^2. Uitwerken van deze expressie toont dat de ratioschatter resulteert in een lagere variantie dan de gewone schatter indien aan de volgende voorwaarde is voldaan:

$$r > \frac{1}{2}\frac{CV_x}{CV_y} = \frac{1}{2}\frac{\frac{s_x}{\bar{X}_D}}{\frac{s_y}{\bar{Y}_D}} \quad (22)$$

4.2.2 Voorbeeld van het gebruik van ratioschatters

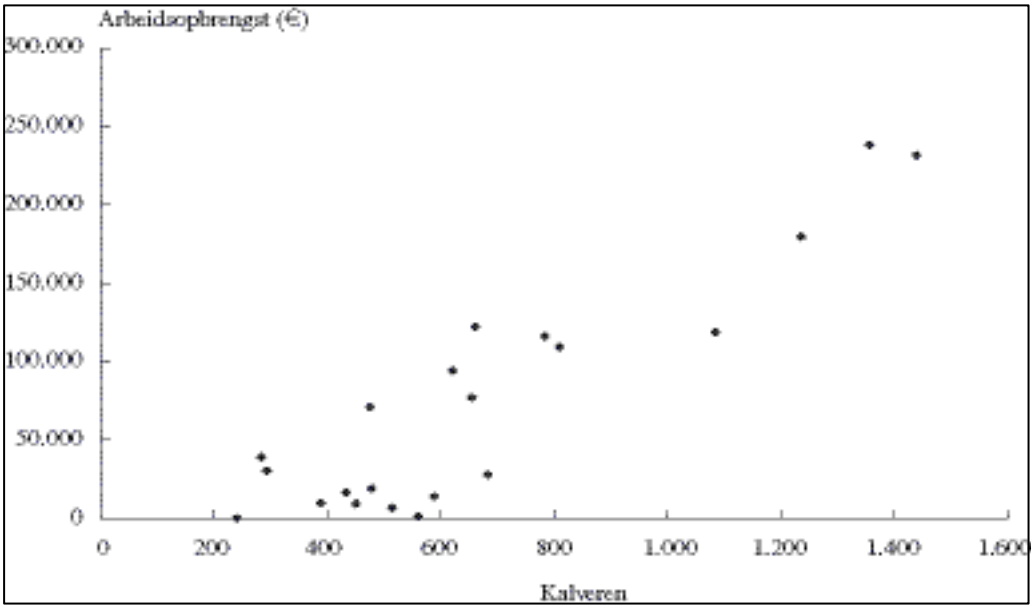
Evenals in paragraaf 4.1 wil men een schatting maken van de arbeidsopbrengst ondernemer van de mestkalverijen. In het Informatienet is slechts een beperkt aantal mestkalverijen opgenomen. De arbeidsopbrengst ondernemer op deze bedrijven laat een grote variantie zijn. In dit voorbeeld zal worden nagegaan in hoeverre de variantie van de schatter kan worden gereduceerd door het gebruik van indirecte schatters. Voor het toepassen van een indirecte schatter is het nodig een hulpvariabele te selecteren die sterk gecorreleerd is met de doelvariabele en waarvan bovendien het gemiddelde of het populatie totaal bekend is. Variabelen die in de Landbouwtelling voor handen zijn voldoen aan deze voorwaarde. De populatie totalen of gemiddelden kunnen worden berekend op basis van de gegevens uit de Landbouwtelling. In dit voorbeeld lijkt het aantal mestkalveren een goede hulpvariabele te zijn .[1] De waarden voor de doel- en hulpvariabele zijn hieronder weergegeven.

Tabel 4.3 Gegevensvoorbeeld ratioschatter

Aantal kalveren (x_i)	Arbeidsopbrengst ondernemer in euro (y_i)
1.236	179.716
389	9.867
475	71.109
621	94.430
1.085	118.874
450	9.483
784	116.438
661	122.384
283	39.082
810	109.502
433	16.858
1.357	238.166
589	14.012
683	28.054
560	1.491
240	203
655	77.246
1.440	231.612
514	6.782
478	19.043
293	30.455

Visualisatie van de hierboven genoemde gegevens laat duidelijk zien dat er een samenhang bestaat tussen het aantal kalveren en de arbeidsopbrengst ondernemer.

[1] In deze voorbeelden is afgezien van het gebruik van de wegingscoëfficiënten.



Figuur 4.4 Samenhang tussen aantal kalveren en arbeidsopbrengst (in €)

Om na te gaan of de hulpvariabele daadwerkelijk geschikt is, is de correlatiecoëfficient tussen de hulp- en doelvariabele berekend. Deze blijkt 0,91 te zijn. De ratioschatter resulteert in een lagere variantie dan de gewone schatter indien aan de volgende voorwaarde is voldaan:

$$r > \frac{1}{2}\frac{CV_x}{CV_y} = \frac{1}{2}\frac{\frac{s_x}{\overline{\overline{X}}_D}}{\frac{s_y}{\overline{\overline{Y}}_D}}$$

Waarbij x de hulpvariabele is en y de doelvariabele. Na invullen van de rechterexpressie blijkt deze gelijk te zijn aan 0,27. De gevonden correlatiecoëfficiënt is hoger dan deze waarde, er mag dus geconstateerd worden dat aan de voorwaarden is voldaan waaronder een ratioschatter tot een betere schatter leidt.

Tabel 4.4 Extra gegevensvoorbeeld ratioschatter

Aantal kalveren (x)	Arbeidsopbrengst ondernemer (€)	$y_i - Rx_i$	$(y_i - Rx_i)^2$
1.236	179.716	66.199	4.382.328.740
389	9.867	-25.859	668.717.794
475	71.109	27.484	755.370.293
621	94.430	37.396	1.398.464.818
1.085	118.874	19.225	369.612.826
450	9.483	-31.845	1.014.164.322
784	116.438	44.433	1.974.361.746
661	122.384	61.676	3.803.974.538
283	39.082	13.090	171.366.023
810	109.502	35.109	1.232.704.789
433	16.858	-22.909	524.851.190
1.357	238.166	113.536	12.890.483.492
589	14.012	-40.083	1.606.646.821
683	28.054	-34.674	1.202.297.157
560	1.491	-49.940	2.494.061.345
240	203	-21.839	476.946.503
655	77.246	17.089	292.048.343
1.440	231.612	99.359	9.872.284.502
514	6.782	-40.424	1.634.167.799
478	19.043	-24.857	617.896.580
293	30.455	3.545	12.568.893

$\overline{X}$ = 510 (gemiddeld aantal vleeskalveren in de populatie, volgens de Landbouwtelling)

$R = \frac{Y}{X} = \frac{\overline{Y}_D}{\overline{X}_D}$ = 109 (gemiddelde arbeidsopbrengst per vleeskalf)

$\overline{Y}_R = R\overline{X}$ = 55.767 (arbeidsopbrengst per vleeskalf * aantal kalveren)

N = 928 (aantal bedrijven in de populatie (Landbouwtelling))

n = 21 (aantal bedrijven in de steekproef (Informatienet))

Een schatting voor de variantie is:

$$n(\overline{Y}_R) = \frac{1-f}{n}\left[S_y^2 + R^2 S_x^2 - 2RrS_y S_x\right] = 85.437.591$$

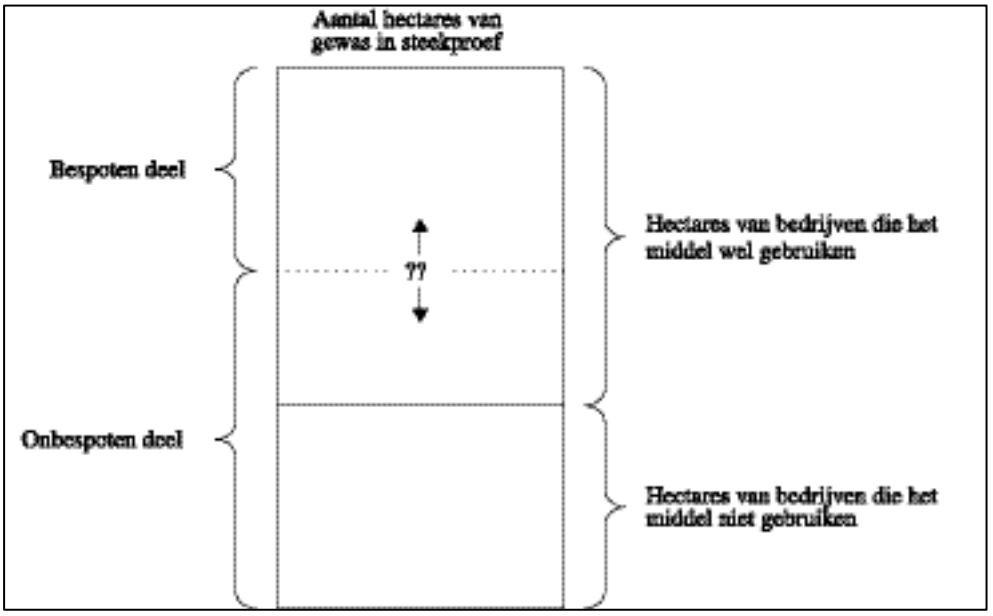
Hieruit volgt dat de standaardfout van de schatter voor het gemiddelde 9.243 bedraagt.

Een directe schatting op basis van de hierboven vermelde gegevens resulteerde in een waarde van ongeveer 73.000 met een standaardfout van bijna 16.000.

Hieruit blijkt dat de directe schatting hoger ligt dan de indirecte schatter (73.000 versus 58.000). Dit verschil is te verklaren uit het feit dat het gemiddelde aantal kalveren van Informatienet-bedrijven hoger ligt dan het gemiddelde aantal kalveren van bedrijven in de Landbouwtelling (668 versus 510). Doordat de indirecte schatter meer informatie in de schatting betrekt is de stelling te verdedigen dat de indirecte schatter een waarheidsgetrouwere schatting oplevert.

4.2.3 Toepassing van de ratioschatter: gebruik gewasbeschermingsmiddelen

Het LEI verzamelt in het Informatienet gebruiksgegevens van gewasbeschermingsmiddelen. Op basis van deze steekproef kan een schatting worden gemaakt van het totale verbruik.



Figuur 4.5 Opbouw gewashectares en bespoten hectares in de steekproef

De bedrijven in de steekproef hebben een bepaald aantal hectares van een gewas (figuur 4.5). Een gewas kan al dan niet bespoten worden met een middel. Een deel van de bedrijven gebruikt het middel niet. Voor deze bedrijven kan dus worden uitgerekend wat het totaal aantal hectares is waar dit middel niet wordt gebruikt. Een ander deel van de bedrijven maakt wel gebruik van het middel. In het Informatienet is voor deze bedrijven echter niet vastgelegd of het middel op een deel van de hectares wordt toegepast of dat alle hectares van het gewas zijn bespoten. De grens tussen het bespoten deel en het onbespoten deel is voor deze bedrijven die het middel voor dit gewas hebben ingezet dus niet volledig duidelijk.

Schatting van het totale verbruik

Voor elke waarneming (elk bedrijf) is bekend of het middel op een bepaald gewas is gebruikt en wat het aantal hectares van dit gewas is. Dit leidt tot het schatten van het aantal kilogram per gewas hectare (voor een specifiek gewas/middel combinatie).

$$(\hat{kg}/ha)_{m,g} = \frac{\sum_{i=1}^{n} kg_{i,m,g}}{\sum_{i=1}^{n} ha_{i,g}}$$

Waarbij geldt dat $ha_{i,g}$ het aantal hectares van gewas g op bedrijf i is, $(\hat{kg}/ha)_{m,g}$ is het gemiddeld aantal kilogram van middel m gespoten op een hectare gewas g, n is het aantal steekproefbedrijven en $\sum_{i=1}^{n} kg_{i,m,g}$ is het aantal kilogram van middel m op bedrijf i gespoten op gewas g. Hierbij is het aantal kilogram voor bedrijven die het middel niet toepassen op het specifieke gewas gelijk aan nul.

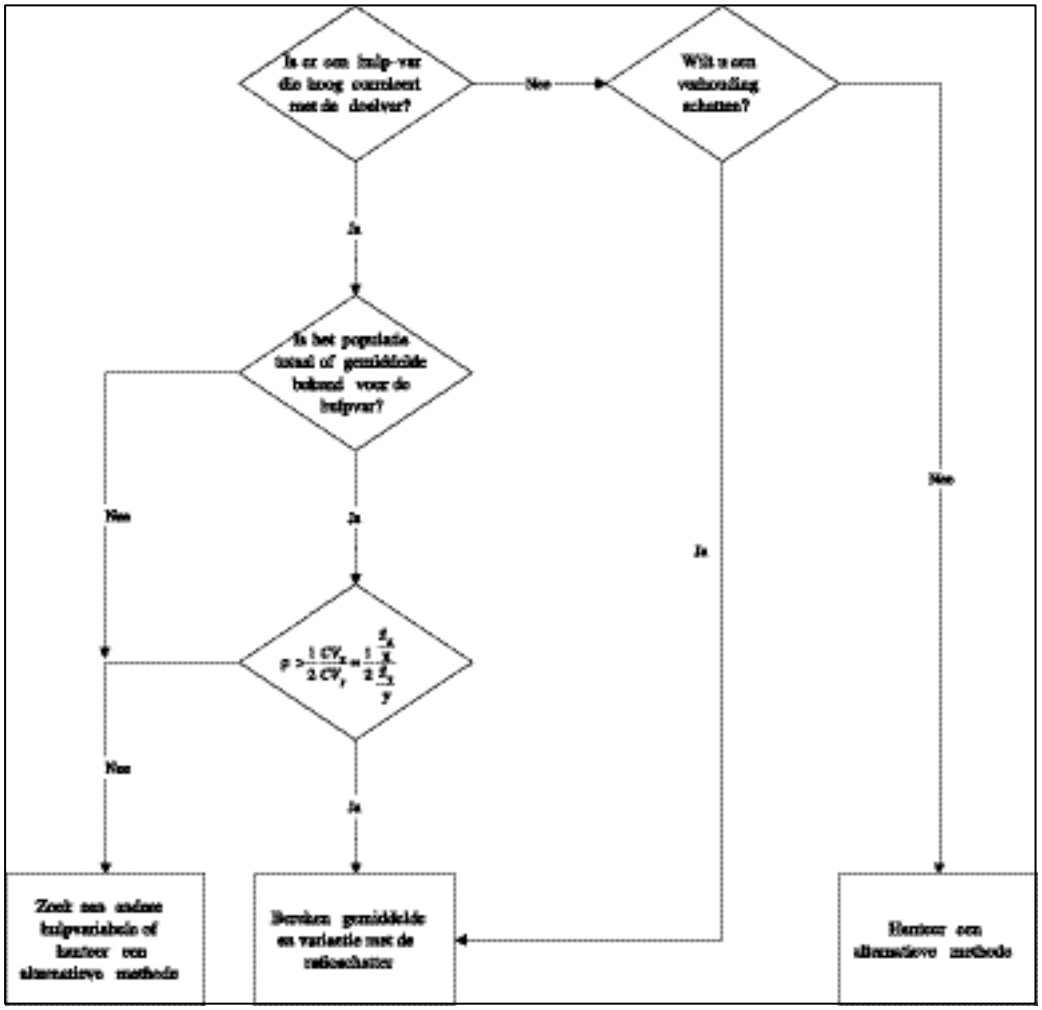
Het totale verbruik van een bepaald middel op een bepaald gewas kan vervolgens worden berekend als kilogram per gewashectare vermenigvuldigd met het aantal hectare van dit gewas volgens de Landbouwtelling.

$$\hat{kg}_{m,g,tot} = (\hat{kg}/ha)_{m,g} * ha_{g,lt}$$

Waarbij geldt dat $\hat{kg}_{m,g,tot}$ het geschatte totale verbruik is van middel m op gewas g en $ha_{g,lt}$ het aantal hectare van gewas g is volgens de Landbouwtelling.

4.2.4 Evaluatie ratioschatters

De voordelen van ratioschatters zijn evident. Door gebruik te maken van additionele informatie kan een betrouwbaardere schatting worden gemaakt. Daarnaast kan dit gebruik van additionele informatie een representativiteitsverhogend effect hebben. Het nadeel van het gebruik van ratioschatters is dat de toepassing veel bewerkelijker is dan een directe schatting. Bij ratioschatters moet een bewuste keuze worden gemaakt welke hulpvariabele kan worden gebruikt voor het maken van een schatting. Tevens geldt dat getoetst moet worden of aan de voorwaarden voor het toepassen van de ratioschatter wordt voldaan. Dit betekent dat het gebruik van indirecte schatters moeilijker in standaard programmatuur is op te nemen. De onderzoeker moet zeer bewust met de problematiek omgaan.



Figuur 4.6 Beslissingsboom voor het gebruik van ratioschatters

4.3 Regressieschatters

Regressieschatters kunnen gebruikt worden om een schatting van de doelvariabele te verkrijgen aan de hand van één verklarende variabele. Het is echter ook mogelijk grotere modellen te definiëren, die gebruikmaken van meerdere verklarende variabelen. In paragraaf 4.3.1 zal het simpele regressiemodel beschreven worden, uitgaande van één verklarende variabele. In paragraaf 4.3.2 zal een uitbreiding op dit simpele model beschreven worden, waarbij meerdere variabelen in de schatting betrokken worden.

4.3.1 Theorie simpel regressiemodel

De regressieschatters maken net als de ratioschatters gebruik van extra informatie van een hulpvariabele die sterk is gecorreleerd met de doelvariabele. Wanneer er wel een verband bestaat tussen deze variabelen maar dit niet door de oorsprong gaat, kan beter gebruik

worden gemaakt van een regressieschatter dan van een ratioschatter. De ratioschatter is in feite een gerestricteerde versie van de regressieschatter, omdat de constante bij de ratioschatter een veronderstelde waarde 0 heeft. Bij een relatie tussen aantal koeien en de melkproductie is het aannemelijk dat de relatie door de oorsprong gaat. Een veestapel van nul koeien zal immers leiden tot een melkproductie van nul liter. In andere situaties is de aanname van een verband door de oorsprong minder waarschijnlijk. Indien bijvoorbeeld een verband wordt verondersteld tussen het nettobedrijfsresultaat en het aantal koeien dan zal dit verband niet door de oorsprong gaan. In verband met vaste lasten zal op een gespecialiseerd bedrijf een veestapel van nul koeien leiden tot een negatief bedrijfsresultaat. In dergelijke situaties is het gebruik van een regressieschatter aan te raden.

De regressieschatter voor het populatiegemiddelde, uitgaande van één verklarende variabele (plus de constante), is te berekenen uit:

$$\overline{Y}_L = \overline{Y}_D + b_1(\overline{X} - \overline{X}_D) = b_0 + b_1\overline{X} \tag{23}$$

De regressieschatting wordt gevormd door het gemiddelde van de doelvariabele in de steekproef, en een correctie voor het verschil tussen het gemiddelde van de hulpvariabele in de populatie en in de steekproef.

De constante, b_0, is gelijk aan:

$$b_0 = \overline{Y}_D - b_1\overline{X}_D \tag{24}$$

Als b wordt geschat op basis van de steekproef dan is b te schatten met behulp van de bekende kleinste kwadraten methode.

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{Y}_D)(x_i - \overline{X}_D)}{\sum_{i=1}^{n}(x_i - \overline{X}_D)^2} \tag{25}$$

De variantie van de regressieschatter is net als bij de ratioschatter opgebouwd uit de steekproefvariantie en de modelvariantie en bedraagt:

$$V(\overline{Y}_L) = \frac{1-f}{n}\frac{\sum_{i=1}^{N}\left[(y_i - \overline{Y}) - b_1(x_i - \overline{X})\right]^2}{N-1} = \frac{1-f}{n}\frac{\sum_{i=1}^{N}(y_i - b_0 - b_1x_i)^2}{N-1} \tag{26}$$

Een schatter voor de variantie op basis van de steekproefgegevens is:

$$n(\overline{Y}_L) = \frac{(1-f)}{n}\frac{\sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2}{n-1} = \frac{(1-f)}{n}s_y^2(1-r^2) \tag{27}$$

Waarbij *r* een schatting van de correlatie tussen de hulp- en doelvariabele weergeeft. Vergelijking van deze variantie met de variantie van een directe schatter laat zien dat de variantie van de regressieschatter lager is, tenzij de correlatie 0 bedraagt. De variantie van de regressieschatter is lager dan de variantie van de ratioschatter behalve wanneer het verband tussen de hulp en doelvariabele wordt gevormd door een rechte lijn die door de oorsprong gaat.

4.3.2 Uitbreiding van het lineaire regressiemodel

4.3.2.1 Theorie

Op basis van theoretische gronden moet een conceptueel model worden geconstrueerd. In deze modelspecificatie moet worden aangegeven welke verklarende variabelen op welke wijze de te verklaren variabele beïnvloeden.

Als meerdere verklarende worden opgenomen, ziet het model er als volgt uit:

$$\overline{Y}_L = b_0 + b_1\overline{X}_1 + b_2\overline{X}_2 + ... + b_k\overline{X}_k \qquad (28)$$

Geschreven in matrixvorm is het te schatten model:

$$y = Xb \qquad (29)$$

Y is een nx1 vector, waarbij n het aantal waarnemingen. X is een nxk matrix, waarbij k het aantal verklarende variabelen. b is een kx1 vector met te schatten parameters. De eerste kolom van X bestaat uit enen en de bijbehorende eerste parameter b_0 is de constante.

Vervolgens moeten de parameters uit dit conceptuele model op basis van empirische gegevens worden geschat. Een schatting voor de vector b wordt verkregen uit:

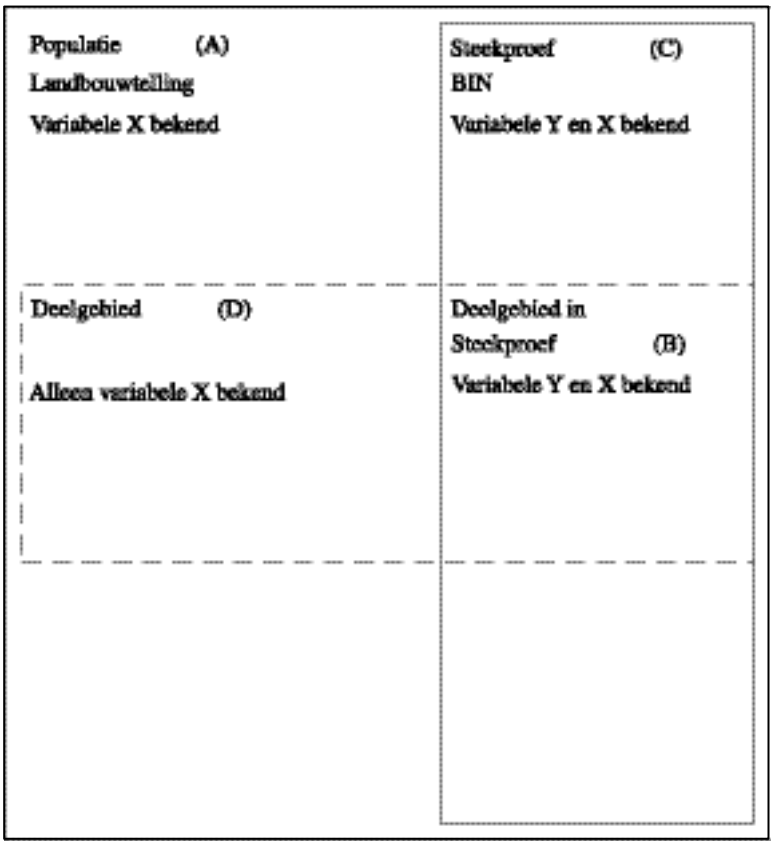
$$b = (X'X)^{-1}(X'y) \qquad (30)$$

De resulterende parameters geven aan hoe een verandering in een verklarende variabele doorwerkt op de te verklaren variabele. Na deze fase moeten diverse controles worden uitgevoerd. Inspectie van de Durbin-Watson statistic kan wijzen op het bestaan van autocorrelatie. Van autocorrelatie is sprake als een verband tussen de verschillende storingstermen bestaat. Inspectie van het patroon van verstoringen kan wijzen op heteroscedasticiteit of andere afwijkingen. Heteroskedasticiteit bestaat als de variantie van de storingstermen niet constant is. Tevens moet worden nagegaan of de variabelen een significante bijdrage leveren aan het verklaren van de totale variantie in de doelvariabele. De omvang van de residuen geeft een indicatie van de mate waarin het model een voorspellend vermogen heeft en of er wellicht belangrijke variabelen buiten beschouwing zijn gelaten. Hierbij moet een afweging worden gemaakt tussen het verklarende vermogen van het model en de eenvoud van het model. Indien het verklarende vermogen sterk toeneemt door een variabele toe te voegen dan is het raadzaam deze variabele daadwerkelijk in het model op te nemen. Als de verklaringsgraad slechts in geringe mate toeneemt is een goede

regel om een dergelijke variabele niet mee te nemen (Sarndal, 1992). Het schatten van een simpel model is beter dan het schatten van een gecompliceerd model, tenzij de fit van het model in belangrijke mate toeneemt. Na het succesvol uitvoeren van toetsen en inspecties op het model, kan het model worden toegepast.

4.3.2.2 Methoden om een schatting voor β te verkrijgen

De informatie die over de doelvariabele (Y) en de hulpvariabele (X) beschikbaar is kan op verschillende manieren worden aangewend om een schatting voor β te krijgen. Figuur 4.7 laat de samenhang zien tussen de beschikbare informatie over verschillende variabelen in de steekproef en de populatie. Het doel is een schatting te maken voor de doelvariabele voor de gehele populatie op het kleine deelgebied (gebied B). Op dit gebied is echter alleen de waarde van de hulpvariabele X bekend. De situatie waar men zich in bevindt is bepalend voor de manier waarop de extra informatie aangewend wordt.



Figuur 4.7 Samenhang bekende gegevens populatie en steekproef

Methode 1: berekenen van b aan de hand van informatie op gebied D

In het voorbeeld over ratioschatters werden enkel de gebieden (B) en (D), respectievelijk de steekproef- en de populatie-elementen op het kleine deelgebied, gebruikt bij het schatten van $\overline{Y}_R$. De ratio R werd bepaald aan de hand van informatie die over Y en X op het kleine deelgebied beschikbaar was in de steekproef (gebied D). Vervolgens werd de ratioschatter berekend aan de hand van informatie over X die beschikbaar was in de hele populatie op het kleine deelgebied. Om een regressieschatting te maken kan iets soortgelijks gedaan worden. De informatie over X en Y in de steekproef op het kleine deelgebied kan gebruikt worden om een schatting voor b (b) te vinden. Vervolgens kan de informatie op het kleine deelgebied in de populatie gebruikt worden om een schatting voor $\overline{Y}_L$ te maken. Deze methode zal uitgewerkt worden in het voorbeeld in paragraaf 4.4.2. De matrix X en de vector y die gebruikt worden voor de schatting van b zijn:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n_H 1} & \cdots & x_{n_H k} \end{bmatrix} \quad en \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_H} \end{bmatrix}$$

Methode 2: berekenen van b aan de hand van informatie op gebied C

Men kan echter ook uitgaan van de veronderstelling dat het verband tussen X en Y buiten het kleine deelgebied hetzelfde is als daarbinnen. Dan kan de informatie over X en Y in de gehele steekproef (C), worden gebruikt om een schatting voor b te vinden. Omdat meer waarnemingen beschikbaar zullen zijn, zal de lineaire regressieschatting betrouwbaarder worden. In dit geval zien X en y er als volgt uit:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad en \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Methode 3: berekenen van b aan de hand van informatie op gebied A

Een andere mogelijkheid is gebied (A) te gebruiken om een schatting voor b te vinden als verondersteld wordt dat de relatie tussen de doel- en de hulpvariabele voor de gehele populatie geldt. Gebied (C), de steekproef voor alle verschillende deelgebieden, kan gebruikt worden om de directe schatters voor de gemiddelden van de doelvariabelen over de verschillende deelgebieden te berekenen. Gebied (A), de populatie, kan gebruikt worden om waarden van alle bekende gemiddelden van de hulpvariabele te verkrijgen over alle verschillende kleine deelgebieden. X en y worden gegeven door:

$$X = \begin{bmatrix} \overline{X}_{11} & \cdots & \overline{X}_{1k} \\ \vdots & \ddots & \vdots \\ \overline{X}_{H1} & \cdots & \overline{X}_{Hk} \end{bmatrix} \quad en \quad y = \begin{bmatrix} \overline{Y}_{D1} \\ \vdots \\ \overline{Y}_{DH} \end{bmatrix}$$

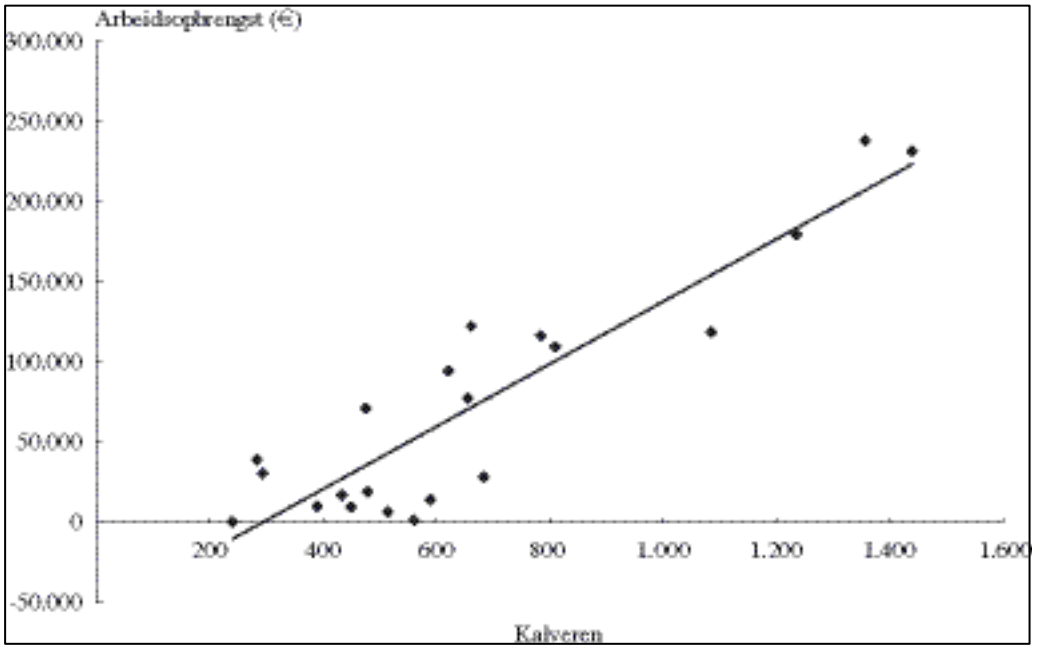
4.3.3 Voorbeeld van het gebruik van een regressieschatter

Uitgaande van het eerdere voorbeeld kan met behulp van de kleinste kwadraten functie de regressiefunctie worden geschat op basis van de steekproefwaarnemingen. Hieruit volgt dat de helling of richtingscoëfficiënt

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{Y}_D)(x_i - \overline{X}_D)}{\sum_{i=1}^{n}(x_i - \overline{X}_D)^2} = 195$$

bedraagt. De intercept of constante is gelijk aan:

$$b_0 = \overline{Y}_D - b_1 \overline{X}_D = -57.190$$



Figuur 4.8 Relatie tussen aantal kalveren en de arbeidsopbrengst (in €)

Invullen van het voorgaande met als steekproefgemiddelde van het aantal kalveren:

$\overline{X}_D = 668$

en het steekproefgemiddelde voor de opbrengst:

$\overline{Y}_D = 73.086$

en het populatiegemiddelde van het aantal kalveren per bedrijf:

$\overline{X} = 510$

levert als regressieschatting voor het gemiddelde:

$\overline{Y}_L = \overline{Y}_D + b(\overline{X} - \overline{X}_D) = 42.216$

De variantie bedraagt:

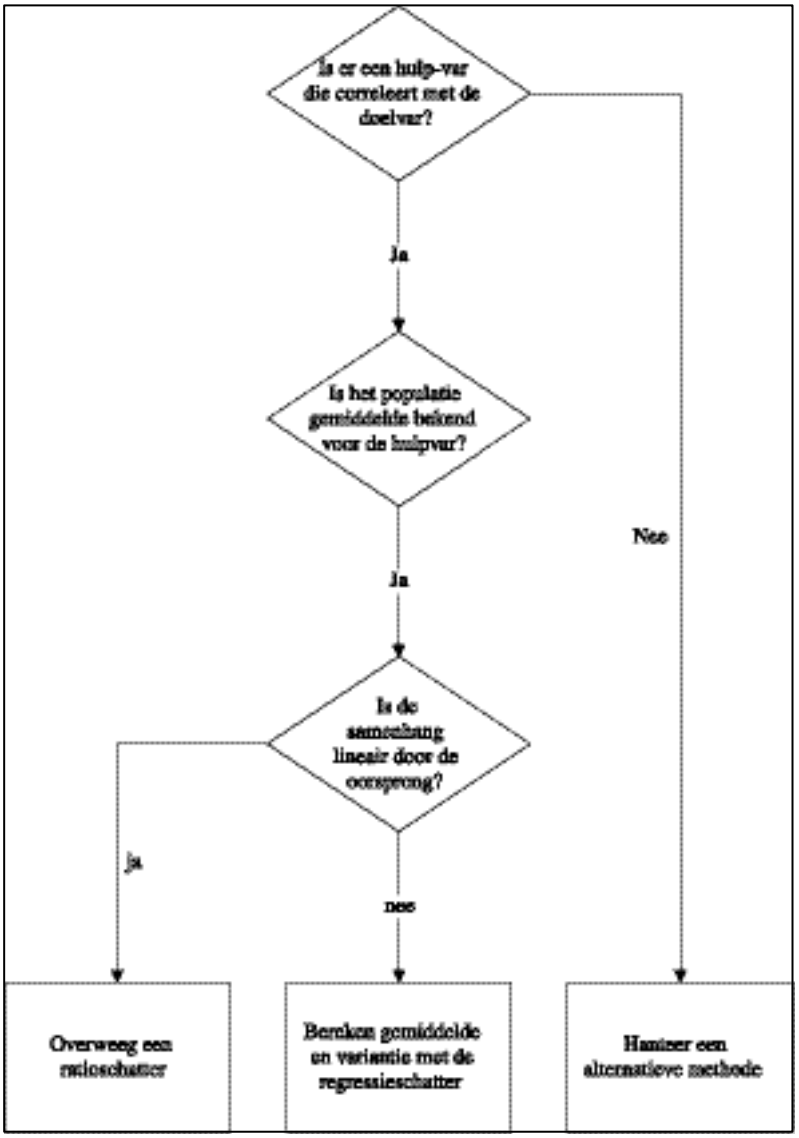
$$V(\overline{Y}_L) = \frac{(1-f)}{n} s_y^2 (1-r^2) = 45.014.842$$

De standaardfout van de regressieschatting van het gemiddelde bedraagt hiermee 6.709.

Hieruit blijkt dat de regressieschatting lager ligt dan de ratioschatting (58 versus 42 duizend). Dit verschil is te verklaren uit het feit dat het verband tussen het aantal kalveren en de opbrengst niet lineair door de oorsprong gaat. Op basis van figuur 4.8 is het aannemelijk te maken dat een regressielijn die gedwongen wordt door de oorsprong te gaan bij het gemiddelde aantal kalveren van 510 boven de afgebeelde regressielijn zal liggen. De ratioschatter die een verband door de oorsprong verondersteld zal dus tot een hogere schatting leiden.

4.3.4 Evaluatie regressieschatter

De voordelen van regressieschatters komen sterk overeen met die van ratioschatters. Door gebruik te maken van additionele informatie kan een betrouwbaardere schatting worden gemaakt. Net als voor ratioschatters geldt het nadeel dat het gebruik van regressieschatters veel bewerkelijker is. Bij regressieschatters moet een bewuste keuze worden gemaakt welke hulpvariabele wordt gebruikt voor het specificeren van een relatie met de doelvariabele. De onderzoeker moet op basis van zijn of haar kennis deze keuzes maken.



Figuur 4.9 Beslissingsboom voor het gebruik van regressieschatters

4.4 Bayesiaanse schatter

4.4.1 Theorie

De Bayesiaanse schatter is een lineaire combinatie van de regressieschatter en de directe schatter. Als het verband tussen doel- en hulpvariabele niet alleen opgaat binnen een klein deelgebied, maar geldt voor de gehele populatie, verdient de Bayesiaanse schatter de voorkeur boven een directe of een regressieschatter. Bayesiaanse analyse maakt beter gebruik van de informatie die vooraf beschikbaar is over de te schatten grootheden dan de directe schatter of regressieschatter. De directe schatter gebruikt enkel de individuele eigenschappen van de doelvariabele binnen een bepaald klein deelgebied en de lineaire

regressieschatter maakt enkel gebruik van de relatie tussen de doelvariabele met andere variabelen (verklarende variabelen) die sterke correlatie vertonen met de doelvariabele. De regressieschatter houdt geen rekening met het feit dat individuele eigenschappen kunnen gelden voor de doelvariabele binnen verschillende kleine deelgebieden die niet terugkomen in de verklarende variabele(n). De Bayesiaanse schatter gebruikt beide schattingstechnieken om een uitspraak te doen over een bepaalde eigenschap van de doelvariabele.

De Bayesiaanse schatter voor $\bar{Y}$ op deelgebied h ziet er als volgt uit:

$$\bar{Y}_{Bh} = B(\bar{Y}_{Lh}) + (1-B)\bar{Y}_{Dh} \qquad (31)\ ^{1}$$

Hierbij is $\bar{Y}_{Bh}$ de Bayesiaanse schatter op deelgebied h voor $\bar{Y}_h$, $\bar{Y}_{Dh}$ is de directe schatter en $\bar{Y}_{Lh}$ is de lineaire regressieschatter op deelgebied h.

De lineaire regressieschatter komt volgens methode 3 beschreven in paragraaf 4.3.2.2 tot stand. Er vanuitgaande dat er één verklarende variabele en een constante worden gebruikt, wordt de regressieschatter weergegeven door:

$$\bar{Y}_{Lh} = b_0 + b_1\bar{X}_h \qquad (32)$$

b_0 en b_1 worden op de volgende manier afgeleid:

$$b_0 = \bar{Y}_D - b_1\bar{X} \qquad (33)$$

$$b_1 = \frac{\sum_{h=1}^{H}(\bar{Y}_{Dh} - \bar{Y}_D)(\bar{X}_h - \bar{X})}{\sum_{h=1}^{H}(\bar{X}_h - \bar{X})^2} \qquad (34)$$

Hierbij is $\bar{X}$ de gemiddelde waarde van de hulpvariabele over de gehele populatie en alle deelgebieden. $\bar{X}_h$ is de gemiddelde waarde van de hulpvariabele binnen populatie h over de gehele populatie. Deze variabele wordt in de andere paragrafen $\bar{X}$ genoemd. $\bar{Y}_D$ is de directe schatter over de gehele steekproef en $\bar{Y}_{Dh}$ is de directe steekproef schatter voor het kleine deelgebied. $\bar{Y}_{Dh}$ werd in voorgaande paragrafen $\bar{Y}_D$ genoemd.

B is de wegingsfactor [2]:

---

[1] Het subscript h wordt gebruikt om aan te geven dat het om een klein deelgebied h gaat. In eerdere hoofdstukken is bij het schatten geen gebruikgemaakt van informatie in andere deelgebieden, dus werden geen subscripten gebruikt.

[2] Voor het gemak wordt ervan uitgegaan dat de variantie van de steekproefschatter gelijk is voor de verschillende kleine deelgebieden. Als dit niet het geval is, wordt de te hanteren methode lastiger. Hiervoor wordt verwezen naar Dol (1991, paragraaf 2.3.8).

$$B = \frac{s^2}{s^2 + t^2} \tag{35}$$

De wegingsfactor geeft aan welk gewicht aan de lineaire schatter toegekend wordt en welk gewicht aan de directe schatter. B is opgebouwd uit s^2 en t^2, s^2 is de variantie van de steekproefschatter op basis van de steekproef. t^2 is de variantie van de veronderstelde verdelingsfunctie van $\overline{Y}$. De weging B heeft gevoelsmatig de juiste eigenschappen. Als de variantie van de steekproefschatter s^2 klein is, zouden we enkel de directe schatter willen gebruiken om tot een schatting voor $\overline{Y}$ te komen. In het extreme geval $s^2 = 0$ zijn alle waarnemingen y_i hetzelfde en is $\overline{Y}_{Dh}$ de beste schatter. Als echter de steekproef erg klein is en s^2 erg groot ten opzichte van $s^2 + t^2$, dan geeft de directe schatter geen betrouwbare schatting en wordt een groter gewicht toegekend aan de lineaire schatter. Een schatter voor B is:

$$\widehat{B} = \frac{(H - k - 2)s^2}{w^2} \tag{36}$$

Waarbij k het aantal te schatten parameters weergeeft. Een schatting voor s^2 is, zoals eerder besproken in paragraaf 4.1 over directe schatters:

$$s^2 = \frac{\sum_{i=1}^{n_h} (y_{ih} - \overline{Y}_{Dh})^2}{n_h - 1} \tag{37}$$

en w^2 wordt berekend uit de som van de kwadratische afwijkingen:

$$w^2 = \sum_{h=1}^{H} (\overline{Y}_{Dh} - \overline{Y}_{Lh})^2 \tag{38}$$

De Bayesiaanse schatter zou binnen het LEI voor onderzoeksvragen over regio's gebruikt kunnen worden. Als het verband tussen de doelvariabele en de hulpvariabele(n) gelijk is voor verschillende regio's. Voor onderzoeksvragen over bepaalde deelpopulaties (zoals de kalveren uit het voorbeeld beschreven in voorgaande paragrafen) is de veronderstelling dat een bepaalde relatie tussen doel- en hulpvariabele ook buiten deze deelpopulatie geldt onaannemelijk. In het voorbeeld over de kalveren kunnen bijvoorbeeld geen berekeningen buiten het kleine deelgebied gemaakt worden, omdat de relatie tussen het aantal kalveren en de arbeidsopbrengst niet bestaat voor bijvoorbeeld een bedrijf waar geen kalveren gehouden worden. De Bayesiaanse schattingstechniek is dan ook niet toepasbaar voor dit soort vragen. De betrouwbaarheid van de regressieschatter neemt toe naarmate er meer kleine deelgebieden zijn, aangezien de schatting voor β gebaseerd is op meer waarnemingen. De directe schatter wordt betrouwbaarder naarmate meer waarnemingen beschikbaar zijn binnen het kleine deelgebied.

4.4.2 Toepassing van de Bayesiaanse schatter

Met behulp van de Bayesiaanse methode zal een schatting van de opbrengst uit melk worden gemaakt voor melkveebedrijven in de provincie Overijssel. De geselecteerde bedrijven zijn in het bezit van ten minste één koe. De hulpvariabele die hierbij gebruikt wordt is het aantal melk- en kalfkoeien dat een bedrijf in bezit heeft. De gemiddelde waarden van de doel- en hulpvariabelen zijn per deelgebied weergegeven in tabel 4.5. Het aantal melkkoeien per bedrijf is een variabele uit de Landbouwtelling. Deze variabele is bekend voor elk deelgebied in de hele populatie. De opbrengst uit melk is een Informatienet-variabele. Van deze variabele zijn dan ook alleen de steekproefwaarden bekend.

Tabel 4.5 Gegevensvoorbeeld Bayesiaanse schatter

Provincie	Melkkoeien $\overline{X}_h$	Arbeidsopbrengst in € $\overline{Y}_{Dh}$
Groningen	57,4	319.312
Friesland	64,2	464.322
Drenthe	52,4	339.020
Overijssel	43,6	290.865
Flevoland	77,7	397.357
Gelderland	42,4	298.426
Utrecht	45,9	276.211
Noord-Holland	42,7	274.834
Zuid-Holland	46,6	303.269
Zeeland	40,5	358.816
Noord-Brabant	54,8	380.714
Limburg	54,3	341.470

Verder zijn de gemiddelden over de gehele populatie en de directe steekproefschatter gegeven door:

$\overline{Y}_D = 336.830$

$\overline{X} = 50$

Een berekening van de coëfficiënten b_0 en b_1 is als volgt:

$$b_1 = \frac{\sum_{h=1}^{H}(\overline{Y}_{Dh} - \overline{Y}_D)(\overline{X}_h - \overline{X})}{\sum_{h=1}^{H}(\overline{X}_h - \overline{X})^2} = 3.579$$

$$b_0 = \overline{Y}_D - b_1\overline{X} = 157.594$$

De lineaire regressieschatter die hier voor Overijssel uit volgt is:

$$\overline{Y}_{Lh} = b_0 + b_1\overline{X}_h = 313.674$$

De directe schatter heeft een waarde:

$$\overline{Y}_{Dh} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_i = 290.866$$

Om tot een schatting van B, de wegingsfactor te komen, worden s^2 en w^2 berekend:

$$s^2 = \frac{\sum_{i=1}^{n_h}(y_{ih} - \overline{Y}_{Dh})^2}{n_h - 1} = 181.919$$

$$w^2 = \sum_{h=1}^{H}(\overline{Y}_{Dh} - \overline{Y}_{Lh})^2 = 17.781.061.540$$

Een schatting voor B is dan:

$$\widehat{B} = \frac{(H - k - 2)s^2}{w^2} = 0{,}000092$$

Uit voorgaande kan de Bayesiaanse schatter worden afgeleidt:

$$\overline{Y}_{Bh} = B(\overline{Y}_{Lh}) + (1 - B)\overline{Y}_{Dh} = 290.868$$

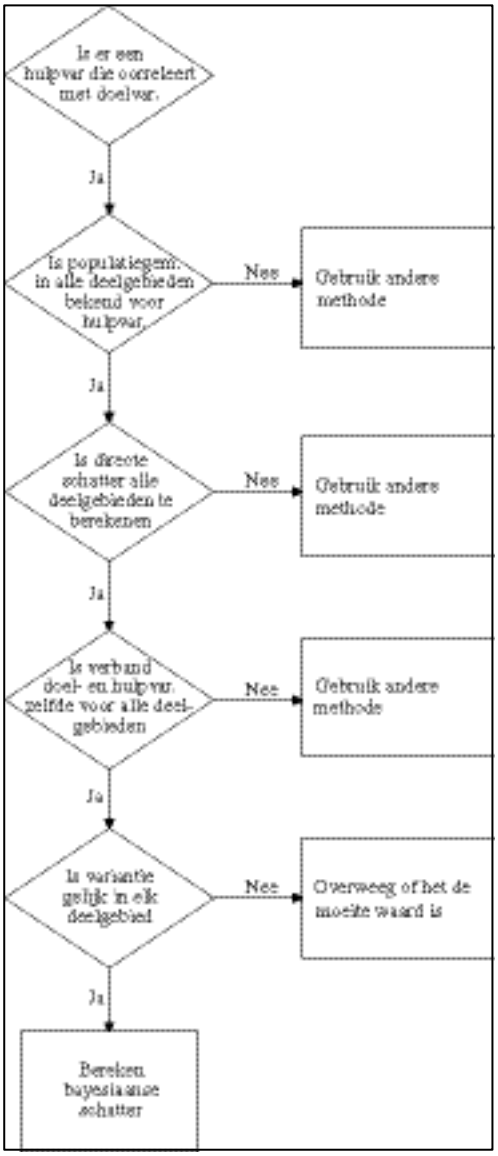
Opvallend is dat de Bayesiaanse schatter een waarde heeft die dicht bij de waarde van de directe schatter ligt. Reden hiervoor is dat s^2 relatief een kleine waarde heeft ten opzichte van $s^2 + t^2$. De schatting is uitgevoerd met een klein aantal deelgebieden waar-

door de lineaire schatter een grote standaardfout en een kleine betrouwbaarheid heeft. De directe schatter levert in dit geval een betere schatting op en krijgt dan ook relatief een zeer grote wegingsfactor. Een reden hiervoor kan zijn dat het aantal kleine deelgebieden te klein is, waardoor de lineaire schatter een zeer kleine betrouwbaarheid heeft, waardoor de bijbehorende wegingsfactor voor de lineaire schatter klein is. Een andere reden kan zijn dat de bedrijven in het gehele land geen goede afspiegeling zijn van de bedrijven die in Overijssel gevestigd zijn. Om deze redenen zou de lineaire schatter relatief slechte resultaten kunnen geven.

Verder is opvallend dat de waarnemingen geclusterd zijn doordat uitgegaan wordt van gemiddelden per provincie. Hierdoor loopt de geschatte regressielijn niet door de oorsprong, terwijl het verband tussen de opbrengst uit melk en het aantal koeien volgens de theorie wel door de oorsprong zou moeten gaan. Deze restrictie zou vooraf opgelegd kunnen worden, dan krijgt b_0 de waarde 0 en wordt b_1 volgens het principe van de ratioschatter berekend.

4.4.3 Evaluatie van de Bayesiaanse schatter

Het voordeel van de Bayesiaanse schatter is dat deze de voordelen van de lineaire regressieschatter en die van de directe schatter combineert. Er wordt gebruikgemaakt van additionele informatie van één of meerdere verklarende variabele(n) en ook wordt rekening gehouden met specifieke kenmerken van de doelvariabele op het kleine deelgebied. De Bayesiaanse schatter is dan ook een gewogen gemiddelde tussen de lineaire regressieschatter en de directe schatter. Nadelen van deze methode zijn echter dat deze vrij bewerkelijk is en dat er vooraf een aantal eisen aan de data gesteld worden. Het aantal deelgebieden dient vrij groot te zijn. Een schatting op basis van 12 deelgebieden levert geen goede resultaten op voor de regressieschatter. Verder dient goed nagedacht te worden over de specificatie van het lineaire regressiemodel en eventueel op te leggen restricties.



Figuur 4.10 Beslissingsboom voor het gebruik van Bayesiaanse schatter

4.5 Poststratificatieschatter

4.5.1 Theorie

In paragraaf 4.1.2 is het gebruik van strata voor het maken van een directe schatting beschreven. Deze strata worden voorafgaand aan het nemen van een steekproef gedefinieerd. In sommige gevallen is het niet mogelijk of niet wenselijk om de strata vooraf te definiëren, bijvoorbeeld omdat men niet weet tot welk stratum elk element behoort, of omdat het aantal variabelen waarnaar gestratificeerd kan worden erg groot is. Poststratificatie of stra-

tificatie achteraf houdt in dat men achteraf strata definieert. Op basis van de frequenties in de populaties kan een betrouwbaardere schatting worden gemaakt.

In het geval een dataset een groot aantal toepassingen heeft, dat wil zeggen dat een groot aantal variabelen als doelvariabele gebruikt wordt, heeft poststratificatie de voorkeur boven stratificatie vooraf (Sarndal, 1992). Bij een gestratificeerde steekproefopzet worden deze strata definitief ingevoerd. Dit leidt tot een reductie in de variantie voor de daarbij gehanteerde doelvariabelen, de stratificatievariabelen. Deze opzet kan echter minder efficiënt zijn voor vele andere doelvariabelen. De combinatie van een aselecte steekproef en poststratificatie kan de totale efficiency verbeteren. Bij de analyse van gegevens kan gebruik worden gemaakt van de kennis en intuïtie van de onderzoeker om bij het onderzoek passende poststratificaties vast te stellen.

Stel dat in het type glastuinbouw twee typen productiesystemen bestaan die van grote invloed zijn op het energieverbruik. Indien men kennis heeft over de verdeling van deze systemen in de populatie (bijvoorbeeld op basis van de Landbouwtelling) dan kan men deze kennis gebruiken om een betere schatting te maken. Stel dat men weet dat 30% van de bedrijven productiesysteem A gebruikt en 70% systeem B. Omdat de steekproef niet is gestratificeerd op basis van dit kenmerk kan het voorkomen dat in de steekproef 50% van de bedrijven systeem A en 50% systeem B gebruikt. In een onderzoek naar het energieverbruik kan het zinvol zijn te corrigeren voor deze verhouding. Poststratificatie leidt er toe dat het gewicht van bedrijven met systeem A iets lager wordt (bedrijven met Systeem A zijn oververtegenwoordigd in de steekproef) en bedrijven met systeem B iets hoger wordt (bedrijven met systeem B zijn ondervertegenwoordigd) bij het maken van schattingen omtrent het energieverbruik.

De poststratificatieschatter is:

$$\bar{Y}_P = \sum_{h=1}^{H} W_h \bar{Y}_{Dh} \tag{39}$$

Poststratificatie resulteert in een (iets) hogere variantie dan stratificatie vooraf. Dit wordt veroorzaakt doordat de variantie van de schatter bij poststratificatie toeneemt doordat de n_h stochastisch zijn. Dat de n_h stochastisch zijn wil zeggen dat men van tevoren niet precies weet hoeveel steekproefeenheden in een bepaald stratum vallen. In het hiervoor genoemde voorbeeld: men trekt de bedrijven random en het is dus niet van tevoren aan te geven hoeveel bedrijven met systeem A en hoeveel bedrijven met systeem B in de steekproef terecht zullen komen. Als men opnieuw zou trekken zou deze verdeling anders kunnen liggen. Deze onzekerheid omtrent het exacte aantal waarnemingen van een bepaald type leidt tot een hogere variantie van de schatting voor het gemiddelde. De formule voor de variantie van poststratificatie is:

$$n(\bar{Y}_P) = \frac{1-f}{n} \sum_{h=1}^{H} W_h s_h^2 + \frac{1}{n^2} \sum_{h=1}^{H} (1-W_h) s_h^2 \tag{40}$$

De eerste term is gelijk aan de variantie van de schatter voor het populatiegemiddelde bij een gestratificeerde aselecte steekproef en de tweede term geeft de variantie ten

gevolge van de aselecte steekproefomvang in elk stratum. In tegenstelling tot stratificatie vooraf is het aantal waarnemingen per stratum bij poststratificatie stochastisch. Door de aanwezigheid van de factor $\frac{1}{n^2}$ zal de tweede term vrij klein zijn waardoor de variantie (en dus ook de standaardfouten) van de variabelen bij grote aantallen bij poststratificatie niet veel hoger zijn dan de variantie van deze variabelen bij stratificatie.

Ten einde na te gaan of er verschil bestaat in de gemiddelden van de doelvariabele tussen de verschillende strata kan een variantieanalyse uitgevoerd worden. De nulhypothese van de ANOVA-test, die voor dit doeleinde gebruikt kan worden, is dat er geen verschil bestaat tussen de gemiddelden voor de verschillende strata. Onder deze nulhypothese geldt:

$$\frac{\sum_{h=1}^{H} n_h(\overline{Y}_{Dh} - \overline{Y}_D)^2 \Big/ (H-1)}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{ih} - \overline{Y}_{Dh})^2 \Big/ (n_1 + n_2 + ... + n_h - H)} \approx F(H-1, n_1 + n_2 + ...n_h - H) \qquad (41)$$

Hierbij geeft $\sum_{h=1}^{H} n_h(\overline{Y}_{Dh} - \overline{Y}_D)^2$ de som van de kwadratische afwijkingen van het gemiddelde tussen de strata weer en $\sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{ih} - \overline{Y}_{Dh})^2$ de som van de kwadratische afwijkingen van de gemiddelden binnen de strata. Als deze twee kwadratische afwijkingen gedeeld door het bijbehorende aantal vrijheidsgraden veel van elkaar afwijken, kan de nulhypothese verworpen worden. De nulhypothese verwerpen dan wel aannemen is afhankelijk van de gekozen betrouwbaarheid. Het verwerpen van de nulhypothese betekent dat de gemiddelden in de verschillende strata significant van elkaar verschillen. Poststratificatie is in dit geval aan te raden.

4.5.2 Toepassing van poststratificatie

Het aantal kalveren hangt sterk samen met de arbeidsopbrengst ondernemer. Door een klasse-indeling te maken naar het aantal kalveren en vervolgens de verdeling over deze klassen in de steekproef en de populatie te bekijken en de juiste gewichten toe te kennen, kan men zich verzekeren van het representatief zijn van de steekproef ten aanzien van het aspect omvangsklasse. Het aantal klassen dat wordt gespecificeerd is in dit voorbeeld 3. De klassengrenzen zijn vastgesteld aan de hand van het aantal kalveren dat een bepaald bedrijf in bezit heeft en zijn te vinden in tabel 4.6, evenals de aantallen steekproefbedrijven en populatiebedrijven in de verschillende klassen.

Tabel 4.6 Klassengrenzen

	Klassengrenzen	n_h	N_h
Klasse 1	0 < X > 500	8	552
Klasse 2	500 < X > 1.000	9	351
Klasse 3	1.000 < X	4	52

Voordat het gemiddelde en de variantie worden berekend, zal eerst de ANOVA-analyse uitgevoerd worden om zodoende na te gaan of de gemiddelden in de verschillende klassen significant van elkaar verschillen. Is dit het geval dan is er een basis voor de post-stratificatieschatter.

Tabel 4.7 Output van SPSS ANOVA-analyse

	Sum of squares	Df	Mean square	F	Sig.
Between groups	76.374.222.386	2	38.187.111.193	20,8	,000
Within groups	33.036.274.972	18	1.835.348.609		
Total	109.410.497.358	20			

Hierbij is:

Sum of squares between groups: $\sum_{h=1}^{H} n_h(\overline{Y}_{Dh} - \overline{Y}_D)^2$

Sum of squares within groups: $\sum_{h=1}^{H}\sum_{i=1}^{n_h}(y_{ih} - \overline{Y}_{Dh})^2$

Df is het aantal vrijheidsgraden (Degrees of freedom) en de Mean square is de sum of squares gedeeld door het aantal vrijheidsgraden. De F-waarde is dan de mean square between groups gedeeld door de mean sqaure within groups.

Uit de variantieanalyse blijkt een duidelijk verschil in de arbeidsopbrengst van de ondernemer tussen de verschillende groepen, dit kan op verschillende manieren afgeleid worden. De tussenvariantie is veel groter dan de binnenvariantie. De gevonden F-waarde geeft aan dat de nulhypothese verworpen wordt. De kritieke waarde bij een betrouwbaarheid van 95% is 3,55 en daar zit 20,8 ver boven. Als naar de laatste kolom van tabel 4.7 gekeken wordt, kan de significantie afgelezen worden. Als deze waarde kleiner is dan 0,05 dan wordt de nulhypothese verworpen. In dit voorbeeld is de waarde veel kleiner dan 0,05. Er is dus een basis voor poststratificatie.

Het gemiddelde bedraagt:

$$\bar{Y}_P = \sum_{h=1}^{H} W_h \bar{Y}_{Dh} = 47.919$$

Indien men de aanname durft te maken dat de oorspronkelijke steekproef aselect getrokken is, kan de variantie van de schatter op de volgende wijze worden berekend:

$$n(\bar{Y}_P) = \frac{1-f}{n} \sum_{h=1}^{H} W_h s_h^2 + \frac{1}{n^2} \sum_{h=1}^{H} (1-W_h) s_h^2 = 75.512.833$$

De benodigde gegevens worden uit tabel 4.8 afgeleid.

Tabel 4.8 Gegevensvoorbeeld poststratificatieschatter

Klasse	s_h	n_h	N_h	W_h	s_h^2	$(1-W_h)s_h^2$	$W_h s_h^2$
1	22.483	8	552	0,57801	505.521.797	213.324.869	292.196.843
2	50.372	9	351	0,367539	2.537.367.202	1.604.785.372	932.582.228
3	55.373	4	52	0,05445	3.066.228.259	2.899.270.917	166.956.908

De afgeleide standaardfout bedraagt 8.690. Uitgaande van de formule voor stratificatie (vooraf), kan de variantie als volgt berekend worden:

$$n(\bar{Y}_S) = \sum_{h=1}^{H} W_h^2 n(\bar{Y}_{Dh}) = 60.011.509$$

De bijbehorende standaardfout heeft een waarde 7.747.

Tabel 4.9 Extra gegevensvoorbeeld poststratificatieschatter

Klasse	s_h	n_h	N_h	W_h	s_h^2	$n(\bar{Y}_{Dh})$	$W_h^2 n(\bar{Y}_{Dh})$
1	22.483,81	8	552	0,57801	505.521.797	62.274.413	20.805.639
2	50.372,29	9	351	0,367539	2.537.367.202	274.700.766	37.107.983
3	55.373,53	4	52	0,05445	3.066.228.259	707.591.036	2.097.887

Hierbij wordt echter geen rekening gehouden met de variantie als gevolg van de onzekerheid van de verdeling van n over de verschillende klassen. Uit deze vergelijking blijkt dus dat poststratificatie tot een iets hogere variantie leidt dan stratificatie vooraf.

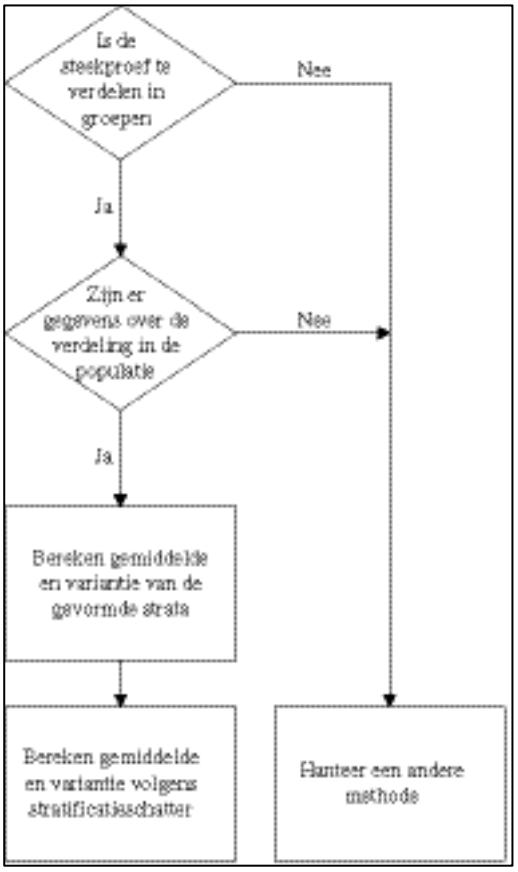
4.5.3 Evaluatie poststratificatie

Poststratificatie heeft een variantiereducerend effect doordat achteraf groepen worden gevormd die homogeen zijn. Hiermee wordt direct een voorwaarde voor het zinvol gebruik van poststratificatie aangegeven. Poststratificatie is alleen zinvol indien de gevormde strata relatief homogeen zijn. Poststratificatie kan een representativiteitsverhogend effect hebben doordat met de verdeling over de groepen in de steekproef ten opzichte van de verdeling in de populatie rekening kan worden gehouden. Het gebruik van poststratificatie bij de problematiek van kleine deelgebieden wordt bemoeilijkt door de eis van minimaal 20 waarnemingen per groep. Het karakter van de problematiek van de kleine deelgebieden is hier vaak mee in strijd. Tevens geldt dat de opzet van het Informatienet poststratificatie bemoeilijkt. Poststratificatie kan eenvoudig worden toegepast op een aselecte steekproef. Doordat bij de opzet van het Informatienet wordt uitgegaan van een disproportioneel gestratificeerde steekproef wordt de berekening van de variantie van de schatters bijzonder complex.

Poststratificatie maakt de schatting van de varianties veel complexer. Bij het gebruik van stratificatie kan de variantie van de schatter binnen de afzonderlijke cellen worden gesommeerd (gewogen). De neiging bestaat een dergelijk simpele procedure toe te passen bij poststratificatie. Door de steekproefeenheden toe te wijzen aan de nieuw te vormen strata en vervolgens de varianties binnen de strata te berekenen en te sommeren kan de variantie van de schatter worden vastgesteld. Dit leidt echter tot een onderschatting van de variantie. Bij stratificatie weet men van tevoren hoeveel steekproefeenheden in elk van de strata zullen vallen. Bij poststratificatie weet men dit niet, de eenheden zijn niet gekozen op basis van deze variabele en het zal dus toevallig zijn welke en hoeveel eenheden in elk van de strata zullen vallen (tenzij er een sterke samenhang bestaat tussen de stratificatievariabelen en de variabelen waarop de poststratificatie wordt toegepast). Het onbekend zijn van het aantal eenheden in een stratum zal leiden tot een hogere variantie van de schatter.

De gedachte leeft bij velen dat het aantrekkelijk zou zijn meerdere sets van gewichten te definiëren en een van deze te selecteren afhankelijk van de te bestuderen doelvariabele. Hiermee zouden een reeks van poststratificatiemogelijkheden van tevoren worden gedefinieerd en worden aangeboden aan de onderzoekers. Echter, het definiëren van een tweede set gewichten maakt geen onderscheid tussen stratificatie of poststratificatie. Het gebruik van een dergelijke set, alsof deze door stratificatie tot stand is gekomen, zal dus resulteren in een onderschatting van de variantie. Ook houdt men in deze procedure onvoldoende rekening met de oorspronkelijke trekkingskansen. In de nieuw te vormen strata (te ontstaan in de poststratificatie) kunnen bedrijven terechtkomen met uiteenlopende trekkingskansen. Door al deze bedrijven eenzelfde gewicht toe te kennen kan er sprake zijn van een systematische afwijking oftewel een bias. De gehanteerde procedure kan dus resulteren in zowel fouten in de schatting van het gemiddelde als in de schatting van de variantie. Bij het op een zorgvuldige manier toepassen van poststratificatie moet het dus

mogelijk zijn dat eenheden binnen een stratum ongelijke gewichten hebben die voldoende rekening houden met de oorspronkelijke trekkingskansen.



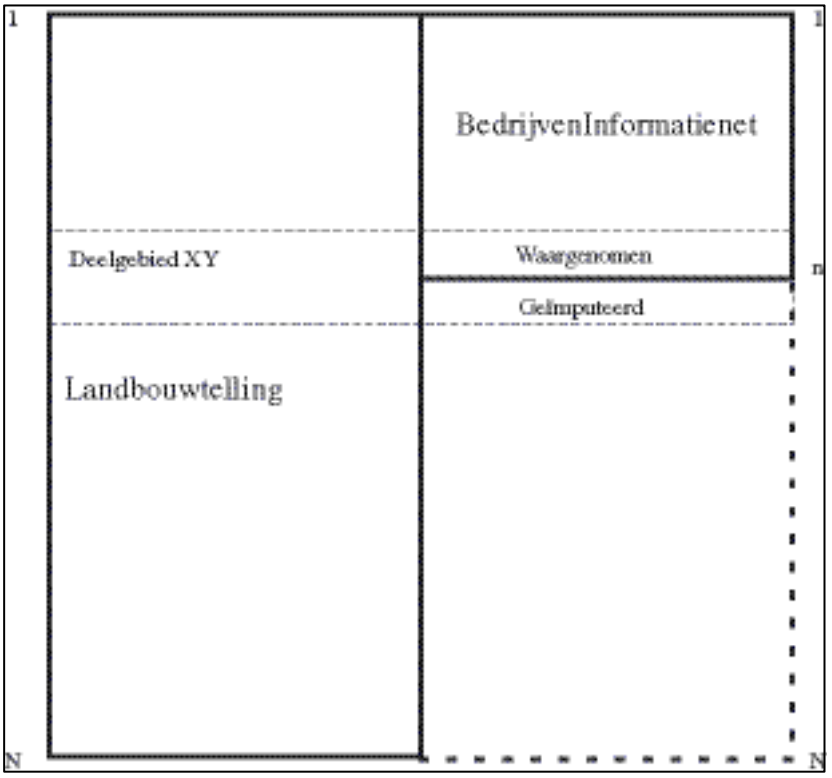
Figuur 4.11 Beslissingsboom voor het gebruik van poststratificatie

4.6 Datafusie en imputatie

4.6.1 Theorie

Datafusie en imputatie komen met name voort uit het marktonderzoek (Vrolijk en Wedel, 1996). In het uitvoeren van marktonderzoek moeten vaak gegevens vanuit verschillende bronnen met elkaar in verband worden gebracht. Het verzamelen van alle relevante gegevens in één onderzoek (single source) is vaak niet mogelijk. Hiervoor zijn enkele redenen. Ten eerste zijn kosten en tijd beperkende factoren die verhinderen alle variabelen in één onderzoek te meten. Een (te) grote hoeveelheid vragen leidt tot een zware belasting van de respondent. Dit kan ten koste gaan van de kwaliteit en daarmee de validiteit van de gegevens. Door vermoeidheid en verveling kan er partiële non-respons optreden. Ten tweede geldt dat men in veel gevallen geïnteresseerd is in een koppeling van generieke onderzoeksresultaten zoals (markt)segmenten aan specifieke gegevens van respondenten die niet

in het betreffende onderzoek zijn gemeten. Het is weinig efficiënt bij elk onderzoek dezelfde segmenten opnieuw te identificeren. Deze twee situaties geven aan dat het koppelen van onderzoeksgegevens wenselijk of noodzakelijk kan zijn. Het samenvoegen van gegevens uit verschillende bestanden wordt datafusie genoemd.



Figuur 4.12 Samenhang Informatienet- en Landbouwtellingsgegevens

In de voorgaande figuur is het principe van datafusie en imputatie geïllustreerd aan de hand van de Landbouwtelling en het Bedrijven-Informatienet van het LEI (het Informatienet). De kenmerken in de Landbouwtelling zijn bekend voor alle agrarische bedrijven groter dan circa 3 nge. Daarnaast is in het Informatienet een gedetailleerde administratie beschikbaar van een kleine 1.500 bedrijven. Voor het overgrote deel van de bedrijven in de Landbouwtelling is deze gedetailleerde administratie niet beschikbaar. Om toch uitspraken te kunnen doen over kenmerken die gelden voor de populatie op het kleine deelgebied, gaat men op zoek naar bedrijven (waarvan wel een administratie beschikbaar is) welke op basis van kenmerken in de Landbouwtelling sterk op het bedrijf, waarover men een uitspraak wil doen, lijken. Een bedrijf dat sterk op een ander bedrijf lijkt op basis van de beschikbare variabelen in de Landbouwtelling zal naar alle waarschijnlijkheid ook lijken op dat andere bedrijf voor variabelen die niet beschikbaar zijn, ervan uitgaande dat de beschikbare en de niet-beschikbare variabelen in grote mate met elkaar gecorreleerd zijn.

De methode kan bijvoorbeeld worden toegepast indien men een uitspraak wil doen over een regio waarvoor men over weinig directe waarnemingen beschikt. In een regio zullen bedrijven zitten van verschillende typen. Om alle typen afzonderlijk te schatten zijn veel waarnemingen nodig. Middels datafusie en imputatie gaat men op zoek naar bedrijven die een grote gelijkenis vertonen met de bedrijven in de te bestuderen regio. Men zoekt voor elk bedrijf in de regio naar een bedrijf in het Informatienet dat er sterk op lijkt gegeven de kenmerken in de Landbouwtelling. Vervolgens wordt de veronderstelling gemaakt dat de te schatten kenmerken van het bedrijf ook hetzelfde zullen zijn, ervan uitgaande dat de gebruikte kenmerken in de Landbouwtelling gecorreleerd zijn met de kenmerken uit het Informatienet. De gegevens van het gelijkende bedrijf in het Informatienet worden dus van toepassing verklaard op het bedrijf in de te bestuderen regio waar men deze gegevens niet direct heeft waargenomen. Op basis van deze (geïmputeerde) gegevens kunnen vervolgens bepaalde statistieken voor de regio worden berekend.

4.6.2 Methoden datafusie en imputatie

Datafusie wordt nu vaker dan voorheen als legitieme en bruikbare methode voor het samenvoegen van gegevens beschouwd (Ford, 1983; Antoine en Santini, 1987; Buck, 1989; O'Brien, 1991; Roberts, 1994; Baker et al., 1995). Voorwaarde hierbij is dat de fusie op een zorgvuldige wijze plaatsvindt. Voor het uitvoeren van datafusie en imputatie bestaat een aantal methoden:

Een eerste mogelijkheid om ontbrekende waarden in te vullen is gebruik te maken van *random selectie*. Hiertoe worden de objecten/subjecten in een aantal voor het onderzoek relevante strata opgedeeld, bijvoorbeeld op basis van leeftijd en regiocode. Vervolgens wordt binnen een dergelijk stratum volgens toeval een waarde getrokken om de ontbrekende waarde in te vullen. Ondanks het feit dat men op deze manier kan waarborgen dat de frequentieverdeling van de waargenomen waarden gelijk is aan die van de ingevulde ontbrekende waarden, gaat de samenhang tussen de gegevens verloren.

Een tweede methode is het gebruik van *regressiemodellen*. Hierbij wordt een regressiefunctie geschat tussen de variabelen. Op basis van deze functie wordt de waarde van ontbrekende variabelen geschat en ingevuld.

Een derde methode voor het invullen van ontbrekende waarden is de *'hot deck' procedure*. Hierbij wordt een subject gezocht die zoveel mogelijk lijkt op het subject met de ontbrekende waarden.

De tweede en derde methode zullen in de onderstaande paragrafen uitvoeriger worden besproken. Ook zal een imputatiemethode die binnen het LEI ontwikkeld is beschreven worden. Deze methode wordt intern binnen het LEI ook wel de 'Tjomme de Haan-methode' genoemd, naar haar bedenker.

4.6.2.1 Regressiemodellen

Eerder in dit hoofdstuk is het gebruik van de regressieschatter beschreven. Door gebruik te maken van aanvullende informatie in de vorm van een hulpvariabele die is gecorreleerd met de doelvariabele kan een betrouwbare schatting worden gemaakt. Met een regressieschatter is het niet direct de bedoeling uitspraken te doen over individuele bedrijven. Het

verband tussen doel en hulpvariabele zal vaak onvoldoende zijn om een groot deel van de variantie van de doelvariabele te verklaren. Met behulp van regressiemodellen met een groter aantal verklarende variabelen kan wel geprobeerd worden de doelvariabele te verklaren uit een set van verklarende variabelen. Indien een model met een redelijke verklaring kan worden geconstrueerd kunnen voorspellingen worden gedaan voor bedrijven waarvan de waarden van de doelvariabelen onbekend zijn.

Meer concreet betekent dit dat geprobeerd kan worden om een doelvariabele uit het Informatienet te verklaren met waarden van variabelen uit de Landbouwtelling. Op basis van de gegevens van de Informatienet-bedrijven kan het model worden geschat. Als een redelijk voorspellend model kan worden geproduceerd, kan dit model worden toegepast op elk bedrijf in de Landbouwtelling. Voor elk bedrijf in de Landbouwtelling kan een voorspelling worden gedaan over de doelvariabele op basis van de beschikbare gegevens in de Landbouwtelling.

Op basis van de waargenomen waarden van de doelvariabele (y_i) van bedrijven in het Informatienet i = 1 tot n_w

$$y_i = b_0 + b_1 x_{1i} + ... + b_n x_{ki} \qquad (42)$$

kunnen de parameters (de regressiecoëfficiënten en de constante) van het model worden geschat. y representeert de doelvariabele (een variabele uit het Informatienet), de x'n zijn de verklarende variabelen (variabelen uit de Landbouwtelling). Voor het construeren van dit model moeten de normale stappen voor modelbouw zoals besproken in paragraaf 4.3.2 worden doorlopen. Als de schattingen voor de bèta's zijn vastgesteld kan voor alle niet waargenomen bedrijven i = n_w+1 tot N de geschatte waarde voor y_i worden berekend:

$$\hat{y}_i = b_0 + b_1 x_{1i} + ... + b_n x_{ki} \qquad (43)$$

De manier waarop de verkregen waarden vervolgens gebruikt dienen te worden, is voor de verschillende imputatiemethoden gelijk en zal worden besproken in paragraaf 4.6.2. De modelvariantie voor lineaire regressieschatters is:

$$s^2 = \frac{\sum_{i=1}^{N} (y_i - \sum_{j=0}^{k} b x_{ji})}{n-k} \qquad (44)$$

Het voordeel van deze aanpak is dat men een indicatie krijgt van de juistheid van de aannames. Indien een model wordt gespecificeerd waarbij de verklarende variabelen weinig tot niets van de variantie in de doelvariabele verklaren dan zal dit tot uitdrukking komen in een lage R^2 en F-waarde van het model. Dit kan aanleiding zijn tot heroverwegen van het model in de vorm van bijvoorbeeld het toevoegen van extra verklarende variabelen. Bij het koppelen van objecten volgens de hot deck-procedure krijgt men niet een dergelijke indicatie.

In de hiervoor beschreven aanpak wordt een model geschat op basis van cross-sectiedata. Een andere mogelijkheid is gebruik te maken van paneldata. Bij paneldata

wordt een groep van bedrijven of personen gedurende meerdere tijdsperioden gevolgd. Dit heeft als voordeel dat men veranderingen in de tijd op individueel niveau beter kan volgen. Deze samenhang tussen de waarnemingen heeft direct tot gevolg dat bij het gebruik van panel data de storingstermen gecorreleerd zullen zijn. Het schatten van de parameters kan dan niet langer plaatsvinden met behulp van OLS.

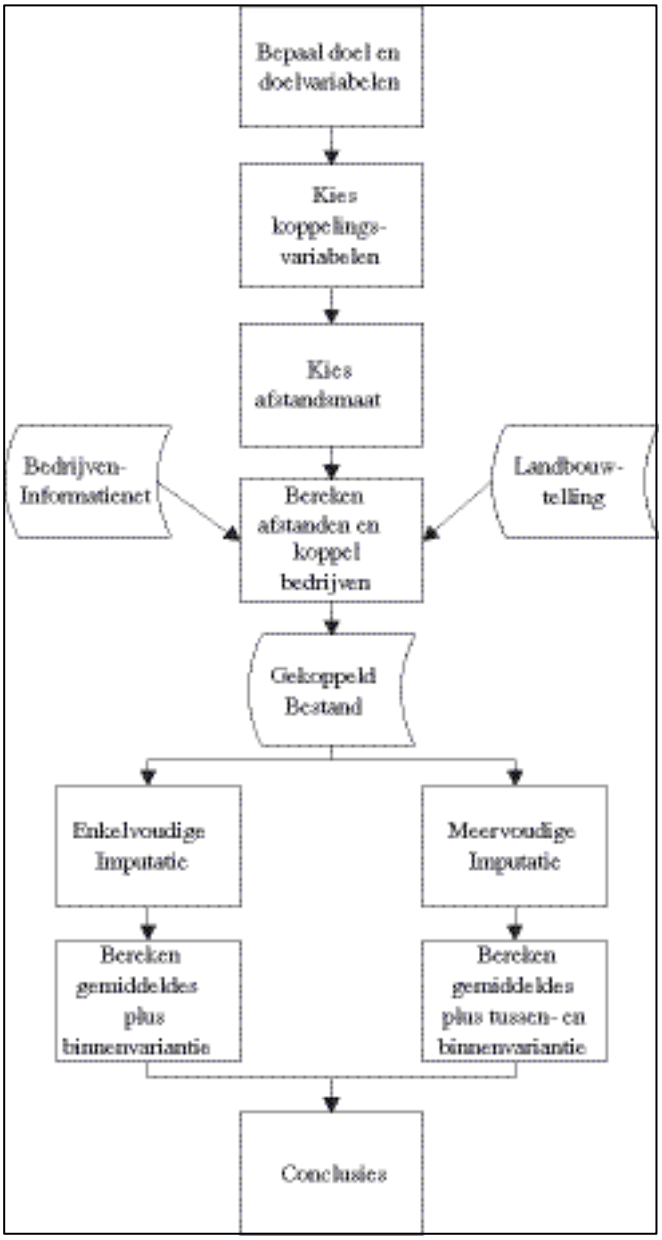
Voor een uitgebreider onderzoek over panel data verwijzen we naar het seo onderzoek naar het gebruik van paneldata (Reinhard en Van Staalduinen).

4.6.2.2 'Hot deck'-procedures

De 'hot deck'-procedure wordt ook veel gebruikt voor datafusie (Bronner, 1988). Hierbij wordt een ontbrekende waarde in bestand A ingevuld met een waarde uit bestand B. Hiertoe wordt in bestand B een subject (de donor) gezocht die zoveel mogelijk overeenkomsten vertoont (op de gemeenschappelijke variabelen) met het subject in bestand A (de ontvanger). De achterliggende gedachte is dat naarmate de ontvanger met betrekking tot de gemeenschappelijke kenmerken meer lijkt op de donor het waarschijnlijk is dat deze respondenten op de unieke variabelen ook gelijke scores hebben. Om de overeenkomsten vast te kunnen stellen geldt de voorwaarde dat bepaalde variabelen zich in beide bestanden bevinden. De mate van overeenkomst tussen personen wordt vastgesteld op basis van variabelen. Deze zogenaamde gezamenlijke variabelen hebben vaak betrekking op demografische kenmerken, zoals de leeftijd, het geslacht, gezinsomvang, regio, inkomensklasse enzovoort, omdat deze standaard in marktonderzoeken worden meegenomen.

De manier waarop de subjecten in de twee bestanden worden gekoppeld, is onder meer afhankelijk van het type variabele. Als er sprake is van kwantitatieve variabelen wordt een afstandsmaat op basis van de gemeenschappelijke variabelen gebruikt om de mate van overeenkomst uit te drukken. Een voorbeeld van een dergelijke afstandsmaat is de Euclidische afstand. De donor uit bestand A wordt vervolgens gekoppeld aan een ontvanger uit bestand B die de kleinste afstand tot de donor heeft.

Bij gebruik van kwalitatieve variabelen, die in categorieën zijn ingedeeld, wordt naar subjecten gezocht waarvan de kenmerken in precies dezelfde categorieën vallen. Indien geen perfecte koppeling tot stand kan worden gebracht worden de categorieën ruimer gedefinieerd of worden bepaalde variabelen buiten beschouwing gelaten. Er wordt een onderscheid gemaakt tussen kritische variabelen en koppelingsvariabelen. De waarden van de kritische variabelen dienen exact overeen te komen. Een koppeling is bijvoorbeeld alleen mogelijk als het geslacht van de donor en de ontvanger gelijk is. Een dergelijke categorie gedefinieerd op basis van de kritische variabelen wordt een imputatiegroep genoemd. Voor de koppelingsvariabelen is deze eis minder stringent. Deze variabelen worden gebruikt om een zo goed mogelijke matching van donor en ontvanger te bewerkstelligen gegeven de overeenstemming van de kritische variabelen.



Figuur 4.13 Aanbevolen stappen in het proces van datafusie en -imputatie

Bij ieder bedrijf uit de CBS-populatie wordt een kopie gezocht uit het LEI-bestand dat zoveel mogelijk overeenkomt op basis van een aantal van tevoren opgegeven gelijkheidsvariabelen. Voor iedere combinatie 'CBS' en 'LEI' wordt de afstand berekend. De totale afstand voor een bepaalde combinatie wordt verkregen door de afstand per gelijkheidsvariabele te berekenen en deze over de variabelen te sommeren:

$$\text{Afstand} = \sum_{i=1}^{\#var} \left(\frac{CBS_i - LEI_i}{Range_i}\right)^{EXP_i} \qquad (45)$$

Het is mogelijk om per variabele verschillende EXP's voor positieve en negatieve afwijkingen op te geven. Dit kan handig zijn in het geval de populatie in bestand 'CBS' sterk verschilt van de verzameling potentiële kopieën in het bestand 'LEI'. Middels een kleinere EXP+ kunnen bijvoorbeeld positieve afwijkingen sterker worden afgestraft dan negatieve. Via de range wordt ongeveer het verschil aangegeven tussen de maximum- en de minimumwaarde. Verkleining van de range geeft een hogere weging aan de betreffende variabele.

Het bedrijf met de kleinste afstand wordt beschouwd als het best vergelijkbare bedrijf. Per bedrijf uit de populatie 'CBS' worden de tien beste potentiële kopieën uit 'LEI' onthouden (top 10), met bijbehorende afstand. In principe wordt het bedrijf uit 'LEI' met de kleinste afstand (nummer 1 uit de top 10) gekozen. Het kan echter voorkomen dat een 'LEI'-bedrijf bij te veel 'CBS'-bedrijven als nummer 1 voorkomt, daarmee oververtegenwoordigd is in de verzameling gebruikte kopieën. Het percentage van de populatie uit 'CBS' dat maximaal gerepresenteerd mag worden door een bedrijf uit 'LEI' kan worden opgegeven.

Nadat alle bedrijven uit de populatie 'CBS' zijn voorzien van een top 10 bedrijven uit 'LEI', wordt per 'LEI'-bedrijf gekeken hoe vaak dit bedrijf als nummer 1 voorkomt in de diverse top 10's van de 'CBS'-bedrijven. Indien een bepaald 'LEI'-bedrijf te vaak als nummer 1 voorkomt, dan wordt nagegaan bij welke 'CBS'-bedrijven het 'LEI'-bedrijf op 1 staat. Per 'CBS'-bedrijf uit dit lijstje wordt vervolgens nagegaan hoeveel de afstand groter zou worden, indien gekozen zou worden voor de tweede keus uit de persoonlijke top 10. Bij het 'CBS'-bedrijf met de kleinste toename in afstand schuift nummer 2 naar nummer 1 in de persoonlijke top 10 (3 naar 2, 4 naar 3, enzovoort) van dit bedrijf. Het aantal maal dat het 'LEI'-bedrijf als nummer 1 voorkomt is nu met één verminderd. Dit proces herhaalt zich tot alle 'LEI'-bedrijven niet vaker dan het maximum aantal malen gebruikt worden.

4.6.3 Verwerking van geïmputeerde waarden

Op basis van de geïmputeerde waarden kan vervolgens een schatting van het gemiddelde worden gemaakt. Na imputatie zal voor alle populatie elementen een waarde zijn waargenomen of geïmputeerd (niet waargenomen):

$$\overline{Y}_I = \frac{\sum_{i=1}^{n_w} y_i + \sum_{i=n_w+1}^{N} y_i}{N} \qquad (46)$$

Het gemiddelde kan met volledige nauwkeurigheid worden bepaald. Hiermee worden de onzekerheden omtrent de geïmputeerde waarden echter volledig buiten beschouwing gelaten.

Imputaties worden vaak als daadwerkelijke waarden gebruikt. Hierdoor wordt de onzekerheid omtrent de geïmputeerde waarden niet meegenomen. Meervoudige imputaties

maken het expliciet omgaan met onzekerheden mogelijk. Bij meervoudige imputaties wordt een aantal data sets gemaakt. Elke dataset bevat alternatieve geïmputeerde waarden. Door de analyses uit te voeren op de afzonderlijke datasets en de resultaten van de verschillende sets te vergelijken wordt de onzekerheid omtrent de geïmputeerde waarden expliciet gemaakt.

Bij het gebruik van meerdere imputatiesets kan de totale variantie worden opgedeeld in de variantie van het gemiddelde gegeven een bepaalde imputatie set (within variance) en de variantie van het gemiddelde tussen de imputatiesets (between variance) (Levy en Lemeshow, 1991). De binnenvariantie (within) bedraagt:

$$s_w^2 = \frac{\sum_{i=1}^{k} \left(\frac{\sum_{j=1}^{n} (y_{ij} - \bar{Y}_{Ii})^2}{n-1} \right)}{k} \qquad (47)$$

Waarbij k het aantal replicaties van het imputatieproces weergeeft. De tussenvariantie (between) bedraagt:

$$s_b^2 = \frac{\sum_{i=1}^{k} (\bar{Y}_{Ii} - \bar{\bar{Y}}_I)^2}{k-1} \qquad (48)$$

Dit leidt tot de volgende schatting van de totale variantie van het gemiddelde:

$$V(\bar{Y}_I) = s_w^2 + (1 + \frac{1}{k} s_b^2) \qquad (49)$$

Indien de resultaten op basis van de verschillende sets niet sterk verschillen (between variance verschilt niet veel van nul) dan is de onzekerheid die volgt uit de imputatie gering. Indien de resultaten sterk verschillen dan is deze onzekerheid groot. Een belangrijk nadeel van deze meervoudige imputaties is natuurlijk dat het zeer bewerkelijk is om dezelfde analyses op alle datasets toe te passen en vervolgens de schattingen te vergelijken.

4.6.4 Richtlijnen voor gebruik

Bij het gebruik van datafusie bestaat het gevaar van 'garbage in garbage out'. De methode resulteert altijd in een dataset, ongeacht de kwaliteit van de match van de bedrijven. Het probleem is dat er geen formele statistiek bestaat ter indicatie van de kwaliteit van de data dit in tegenstelling tot een methode zoals regressieanalyse waarbij de R^2 een indicatie geeft van het percentage van de variantie dat door het model wordt verklaard. Zonder kwaliteitscontrole loopt men het risico dat het toepassen van de methode is te vergelijken met een random generator.

De keuze van de koppelingsvariabelen is hierdoor des te belangrijker. De keuze van variabelen dient afhankelijk te zijn van het probleem. Het moet meer dan aannemelijk worden gemaakt dat de koppelingsvariabelen van invloed zijn op de doelvariabelen. Naast een intuïtieve beoordeling van deze verbanden verdient het aanbeveling het verband tussen de koppelingsvariabelen en de doelvariabele nader te toetsen. Met behulp van bijvoorbeeld variatieanalyse en regressieanalyse kan worden nagegaan of de koppelingsvariabelen daadwerkelijk de variantie in de doelvariabele(n) verklaren. Een meer geavanceerde methode die zou kunnen worden toegepast is datamining. Datamining zoekt naar verborgen verbanden in een grote hoeveelheid gegevens. Datamining zou bij datafusie kunnen worden toegepast om in een grote verzameling variabelen, de variabelen te identificeren die de waarde van de doelvariabele in sterke mate bepalen. Als dergelijke variabelen geïdentificeerd kunnen worden, dan zijn deze geschikt om toegepast te worden als koppelingsvariabelen.

Een verdere verbetering kan worden gerealiseerd door onderscheid te maken tussen kritische en niet-kritische koppelingsvariabelen. Door uitsluitend te kijken naar de euclidische afstanden tussen bedrijven op basis van een verzameling koppelingsvariabelen wordt uitgegaan van een soort compensatorisch model. Een verschil in leeftijd wordt even zwaar gewogen als een bepaald verschil in hectares (het exacte aantal hectares dat overeenkomt met 1 jaar leeftijdsverschil is afhankelijk van de gehanteerde afstandsmaat). In een onderzoek waar leeftijd een belangrijke rol speelt is deze afweging misschien niet realistisch. Door onderscheid te maken tussen kritische en niet-kritische variabelen zou men kunnen afdwingen dat er een goede match op basis van leeftijd plaatsvindt. Een koppeling wordt uitsluitend tot stand gebracht indien de kritische koppelingsvariabelen overeenstemmen.

Het al dan niet koppelen van twee bedrijven wordt bepaald door de variabelen die in de analyse worden meegenomen en de afstandsmaat die wordt gebruikt om de afstand tussen twee bedrijven te definiëren. De gevonden oplossing van het koppelen van bestanden wordt in grote mate bepaald door de variabelen die worden gekozen. De koppeling van bedrijven is afhankelijk van deze keuze. De keuze van de variabelen moet dan ook plaatsvinden op basis van theoretische en praktische gronden. De keuze moet variabelen omvatten die een goede karakterisering van de bedrijven geven en die sterk zijn gerelateerd aan het doel van de analyse. De uitkomst van de fusie kan sterk worden beïnvloed door de keuze van niet-relevante variabelen.

De euclidische afstand is een veel gebruikte afstandsmaat. De euclidische afstand is bij twee variabelen de lengte van de schuine zijde van een driehoek. Een veel gebruikt alternatief is de 'city block' afstand waarbij de totale afstand wordt bepaald door de som van de afstanden op de afzonderlijke variabelen.

De afstandsmaat is gevoelig voor de gehanteerde schaal. Dit kan worden geïllustreerd aan de hand van een eenvoudig voorbeeld waarbij op basis van leeftijd en oppervlakte van het bedrijf de koppeling tot stand wordt gebracht. Bij het niet standaardiseren van de scores zal een eenheid van oppervlakte opwegen tegen een eenheid van leeftijd. Het is eenvoudig te zien dat oppervlakte een veel sterker effect heeft indien deze gemeten wordt aan de hand van vierkante meters dan bij meting in hectares. In de meeste gevallen verdient het dus aanbeveling om de scores voorafgaand aan het berekenen van de afstandsmaat te standaardiseren. Standaardiseren impliceert het converteren van elke vari-

abele naar standaardscores door het gemiddelde van de waarde af te trekken en te delen door de standaarddeviatie. Standaardisatie heeft twee belangrijke voordelen:

1. het wordt makkelijker variabelen te vergelijken omdat ze op dezelfde schaal zijn gemeten. Positieve waarden vallen boven en negatieve onder het gemiddelde. De score geeft aan hoeveel standaarddeviaties de originele score van het gemiddelde valt;
2. het veranderen van de schaal leidt niet tot andere gestandaardiseerde scores. De uitkomsten van het koppelingsproces zijn dus onafhankelijk van de gehanteerde schaal.

Een ander aspect waarop gelet moet worden is multi-collineariteit (samenhang tussen variabelen). Indien multi-collineariteit optreedt, betekent het dat deze variabelen impliciet een groter gewicht krijgen. In het extreme geval zijn twee variabelen identiek, in dit geval telt de afstand voor deze variabele dubbel mee (een keer via de eerste variabele en de tweede keer door de tweede variabele die identiek is aan de eerste). Op deze manier werkt multi-collineariteit als een wegingprocedure die niet direct duidelijk is voor de onderzoeker.

4.6.5 Toepassing

Stel dat men een uitspraak wil doen over melkveebedrijven in een gemeente in het Noordelijk Weidegebied. Op basis van het aantal waarnemingen in het Informatienet is het lastig directe uitspraken te doen over een gemeente. Middels datafusie/imputatie gaat men op zoek naar Informatienet-bedrijven in een groter gebied die qua Landbouwtellingskenmerken sterk lijken op de bedrijven in de gemeente. Om de basis voor de vergelijkbaarheid te vergroten wordt uitgegaan van andere bedrijven in het totale Noordelijke Weidegebied. Van dit gebied zijn 70 melkveebedrijven in het Informatienet opgenomen. De koppelingsvariabelen die zijn gebruikt zijn (zie volgende paragraaf voor een onderbouwing van de keuze van de variabelen):

- leeftijd;
- hectares gras;
- hectares voedergewas;
- melkkoeien; en
- NGE.

In het voorbeeld wordt geprobeerd een schatting te maken van de variabelen opbrengsten, kosten, nettobedrijfsresultaat, arbeidsopbrengst ondernemer en aantal ondernemers.

In de onderstaande tabel zijn de resultaten van het fusieproces beschreven. Hierbij is uitgegaan van een enkelvoudige imputatie. Voor elk bedrijf in de gemeente van onderzoek is in het Informatienet gekeken welk bedrijf in het Noordelijk Weidegebied daar het sterkst op lijkt gegeven de 5 koppelingsvariabelen (deze variabelen zijn eerst gestandaardiseerd). Vervolgens is het gemiddelde voor de 5 doelvariabelen voor de gemeente van onderzoek berekend uitgaande van de veronderstelling dat de waarden voor de variabelen van het meest gelijkende bedrijf in het Noordelijk Weidegebied de beste benadering voor het bedrijf in de gemeente is.

Tabel 4.10 Resultaten fusie (enkelvoudige imputatie)

	Resultaat fusie	Standaardfout
Opbrengsten	415.020	15.028
Kosten	506.479	15.103
Nettobedrijfsresultaat	-80.069	4.581
Arbeidsopbrengst ondernemer	58.066	5.010
Aantal ondernemers	1,47	0,05

Zoals in de theoretische evaluatie van de methode is beschreven, heeft het gebruik van de enkelvoudige imputatie als nadeel dat de variantie wordt onderschat. De geïmputeerde waarde wordt als waarheid opgevat terwijl er toch een bepaalde onzekerheid omtrent deze waarde hangt. In de volgende tabel is de situatie doorgerekend op basis van een meervoudige imputatie. Hierbij is gebruikgemaakt van de 3 beste koppelingen en een random trekken van 1 van de 3 bedrijven voor de daadwerkelijke koppeling. Middels 100 replicaties is een schatting van de uitkomsten gemaakt.

Tabel 4.11 Resultaten fusie (meervoudige imputatie)

	Resultaat fusie	Standaardfout	Min.	Max.
Opbrengsten	417.203	16.723	405.002	431.081
Kosten	505.405	16.354	492.738	521.129
Nettobedrijfsresultaat	-76.984	5.502	-85.138	-69.606
Arbeidsopbrengst ondernemer	63.899	6.459	56.126	75.055
Aantal ondernemers	1,49	0,05	1,4	1,6

Uit een vergelijking van de voorgaande 2 tabellen blijkt dat de resultaten dicht bij elkaar liggen. Wel komt duidelijk naar voren dat het doorrekenen van 100 replicaties tot gevolg heeft dat de variantie van de schatter toeneemt. Deze verhoging is het gevolg van het toevoegen van een stuk variantie als gevolg van de verschillen in het gemiddelde tussen replicaties. Uit de kolommen min en max blijkt bijvoorbeeld dat de schatting van de opbrengsten kan uiteenlopen van 405.000 tot 431.000. Deze variantie komt boven op de variantie als gevolg van de verschillen in waarden tussen bedrijven binnen een replicatie. Voor de verschillende doelvariabelen gaat de standaardfout met circa 10% omhoog.

4.6.6 Validatie

In de huidig gehanteerde werkwijze is het niet goed mogelijk een uitspraak te doen over de kwaliteit van de fusie/imputatie. In deze paragraaf wordt een validatie beschreven waarbij de beschreven fusie procedure wordt toegepast. Door het bestand te koppelen aan het eigen

bestand met als voorwaarde dat een bedrijf niet aan zichzelf gekoppeld kan worden, kan de geschiktheid van de methode en de keuze van de koppelingsvariabelen worden beoordeeld.

In dit voorbeeld worden melkveebedrijven (NEG-type 4410) uit het Noordelijk Weidegebied als uitgangspunt genomen. Als afhankelijke variabelen worden opbrengsten, kosten, nettobedrijfsresultaat, arbeidsopbrengst ondernemer en aantal ondernemers gehanteerd. Een mogelijke lijst van verklarende variabelen (koppelingsvariabelen) is hieronder weergegeven:

Leeftijd	Percentage overige weide vee
Hectare cultuurgrond	Percentage zeugen
Hectare grasland	Percentage vleesvarkens
Hectare voederoppervlak	Percentage pluimvee
Percentage overig oppervlak	Percentage voeder oppervlak
Melkkoeien	Percentage granen
Melkkoeien per hectare	Percentage knolgewas
SBE	Percentage overige akkerbouw
SBE varkens	Percentage tuinbouw open
Percentage melkvee	Percentage tuinbouw glas

Kader 4.1 Mogelijke koppelingsvariabelen

In tabel 4.12 zijn de resultaten weergegeven:

Tabel 4.12 Imputatie op basis van volledige variabelenlijst

	Werkelijke waarden	Resultaat fusie	Standaardfout
Opbrengsten	476.902	493.360	32.869
Kosten	569.488	573.109	33.472
Nettobedrijfsresultaat	-79.303	-66.473	9.536
Arbeidsopbrengst ondernemer	67.817	80.157	11.858
Aantal ondernemers	1,53	1,49	0,09

Zoals uit de tabel is af te lezen valt de voorspelde waarde middels fusie dicht in de buurt van de werkelijke waarde. Indien de standaardfout wordt bekeken is geen sprake van een significant verschil tussen de voorspelde en werkelijke waarde.

Het is echter de vraag in hoeverre de hele lijst van variabelen in de fusie betrokken moet worden. In principe moet een afweging worden gemaakt tussen de juistheid van de uitkomsten en een eenvoud van het datafusie model. In het onderstaande tabel is een extreem voorbeeld weergegeven waarbij de datafusie alleen op basis van de leeftijd en het aantal hectares grasland is uitgevoerd.

Tabel 4.13 Fusie op basis van leeftijd en aantal hectares

	Werkelijke waarden	Resultaat fusie	Standaardfout
Opbrengsten	476.902	355.033	21.028
Kosten	569.488	459.701	14.797
Nettobedrijfsresultaat	-79.303	-91.233	9.601
Arbeidsopbrengst ondernemer	67.817	12.530	10.507
Aantal ondernemers	1,53	1	0

Uit de voorgaande tabel is af te lezen dat er grote verschillen zijn tussen de voorspelde en werkelijke waarden. Deze verschillen zijn significant. Op basis van deze analyse mag geconcludeerd worden dat het uitvoeren van een imputatie op basis van slechts 2 variabelen resulteert in onvoldoende kwaliteit.

In tabel 4.14 is een inhoudelijke keuze gemaakt voor vijf koppelingsvariabelen. De resultaten van de fusie zijn weergegeven. Uit de tabel blijkt dat een fusie op basis van een beperkte set van variabelen leidt tot even goede of zelfs betere resultaten dan een fusie op basis van de uitgebreide set van variabelen.

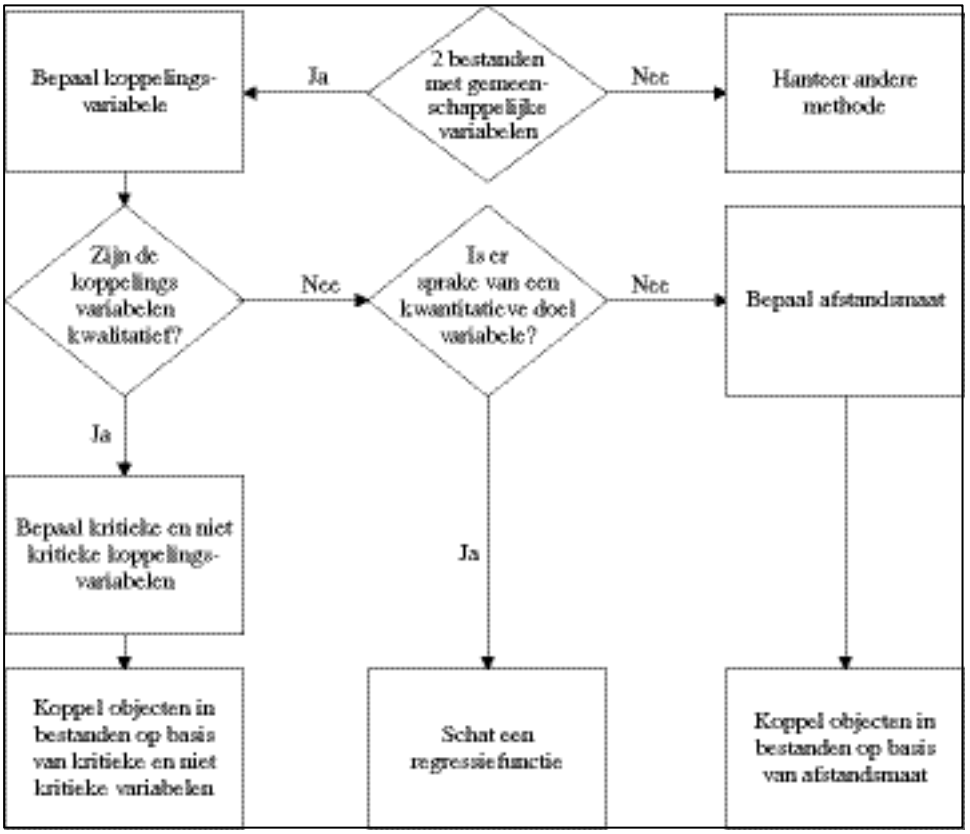
Het voordeel van deze aanpak is dat men inzicht krijgt in de kwaliteit van het fusieproces en dat men een betere onderbouwing kan geven van de variabelen die in het proces worden opgenomen.

Tabel 4.14 Fusie op basis van leeftijd, ha gras, ha voeder gewas, melkkoeien en NGE

	Werkelijke waarden	Resultaat fusie	Standaardfout
Opbrengsten	476.902	470.917	34.330
Kosten	569.488	560.114	33.836
Nettobedrijfsresultaat	-79.303	-76.492	9.182
Arbeidsopbrengst ondernemer	67.817	68.500	11.297
Aantal ondernemers	1,53	1,53	0,09

4.6.7 Evaluatie van datafusie en imputatie

Alhoewel de hot deck-procedure in de praktijk goed kan werken, heeft hij belangrijke nadelen. Op de eerste plaats moeten een aantal subjectieve keuzen worden gemaakt, met betrekking tot wat de kritische variabelen zijn en wat koppelingsvariabelen, en welke afstandsmaat gebruikt wordt om donoren en ontvangers te koppelen. Deze keuzen kunnen een belangrijke invloed hebben op het resultaat. Ten tweede zijn de statistische eigenschappen van het gefuseerde databestand onbekend. Dit komt omdat in het gefuseerde bestand de gegevens van twee objecten worden gekoppeld en in vervolganalyses als één object worden beschouwd. De onzekerheid in de gegevens, die ontstaat door de koppeling wordt in vervolganalyses niet meer meegenomen.



Figuur 4.14 Beslissingsboom voor het gebruik van datafusie en imputatie

Het zal in alle gevallen in meer of mindere mate statistisch incorrect zijn de geïmputeerde waarden op dezelfde manier te beschouwen als de daadwerkelijke geobserveerde waarden (Sarndal et al., 1992). De mate van verstoring is afhankelijk van de kwaliteit van de geïmputeerde waarden. De imputaties kunnen slechte substituten zijn van de daadwerkelijke waarden. Wanneer ze bijvoorbeeld tot stand komen op basis van voorspellingen vanuit de steekproef zal de voorspelde waarde bepaalde statistische eigenschappen hebben zoals een verwachte variantie. Door uitsluitend naar de verwachte waarde te kijken worden de varianties onderschat.

Bij het matchen van bedrijven op basis van de Landbouwtelling zal de kwaliteit afhankelijk zijn van de variabelen die worden gebruikt bij het matchen en de mate waarin de afhankelijke variabele kan worden verklaard door deze matchvariabelen. Het matchen veronderstelt impliciet een model waarbij de doelvariabele kan worden verklaard uit de matchvariabelen. Als dit niet het geval is, kunnen bedrijven die sterk overeenkomen op de match variabelen zeer sterk verschillen op de doelvariabelen. In dat geval is matching niet zinvol.

Imputaties worden vaak als daadwerkelijke waarden gebruikt. Hierdoor wordt de onzekerheid omtrent de geïmputeerde waarden niet meegenomen. Meervoudige imputaties maken het expliciet omgaan met onzekerheden mogelijk.

5. Evaluatie methoden

Voor de evaluatie en vergelijking van de methoden is gebruikgemaakt van een lijst van criteria. Deze criteria hebben betrekking op diverse aspecten van de methoden, zoals het gemak waarmee de methode kan worden toegepast, de concepten die ten grondslag liggen aan de methode en vele andere. Hieronder volgt de lijst van criteria en de definitie van deze criteria.

Criterium	Omschrijving
Betrouwbaarheid te berekenen	Kan met de methode aangegeven worden hoe betrouwbaar of hoe nauwkeurig een schatting is?
Betrouwbaarheid bij kleine aantallen	Hoe betrouwbaar zijn de schattingen bij het gebruik van kleine aantallen steekproefbedrijven?
Zuiverheid	Is de schatting zuiver?
Indicatie goodness of fit van de aannames	In hoeverre krijgt de onderzoeker een indicatie van de juistheid van de aannames die aan de methode ten grondslag liggen?
Validiteit	Wordt daadwerkelijk gemeten wat men wil meten, in hoeverre kan men ervan uitgaan dat de schatting juist (bias vrij) is voor de populatie?
Onderbouwing	Wat voor ideeën of concepten liggen aan de methode ten grondslag?
Eenvoud	In hoeverre is het principe van de methode eenvoudig in te zien?
Bewerkelijkheid	Hoeveel moeite en tijd kost het om de methode in een praktijksituatie toe te passen?
Flexibiliteit	In hoeverre is de methode flexibel in de zin dat andere aannames makkelijk kunnen worden doorgerekend?
Wetenschappelijke acceptatie	In hoeverre is er uitvoerig onderzoek verricht naar de toepasbaarheid en bruikbaarheid van een methode?
Meerdere doelvariabelen	Wil men uitspraken doen over een of meerdere doelvariabelen?
Gebruik extra info	Wat is de aard van de extra informatie die wordt gebruikt?
Meerdere hulpvariabelen	Is het gebruik van meerdere hulpvariabelen mogelijk?
Nominale of ordinale hulpvariabele	Is de extra informatie in de vorm van de hulpvariabele nominaal of ordinaal geschaald?
Interval of ratio hulpvar	Is de extra informatie in de vorm van de hulpvariabele interval of ratio geschaald?
Reproduceerbaarheid	In hoeverre kunnen de resultaten op een later tijdstip of door een andere onderzoeker worden gereproduceerd?

Kader 5.1 Criteria voor de evaluatie van de methoden

In de volgende tabel zijn de in dit rapport beschreven methoden uitgezet tegen de hiervoor genoemde criteria. In de tabel is aangegeven hoe een methode scoort op een bepaald criterium. Na de tabel zullen de scores worden toegelicht per criterium.

Tabel 5.1 Evaluatie van de beschreven methoden

Methode Criteria	Directe schatter	Ratio-schatter	Regressie-schatter	Bayesiaanse schatter	Post-stratificatie	Regressie-model	Hot deck-procedures
Betrouwbaarheid	++	++	++	-	+	-/+	-/+
Te berekenen betrouwbaarheid bij kleine aantallen	-	+	+	Nvt	-	-/+	Nvt
Zuiverheid	++	-/+	-/+	-/+	+	?	--
Indicatie GFI aannames	Nvt	+	+	+	Nvt	+	-
Validiteit bij kleine aantallen	-	+	++	++	++	+	??
Onderbouwing	Steek-proeven	Model en steekproef	Model en steekproef	Model en steekproef	Steek-proeven	Model en steekproef	Afstands-maten
Eenvoud	++	-	--	--	-	--	+
Bewerkelijkheid	++	-	--	--	-	--	+
Flexibiliteit	++	-	-	-	-	-	++
Wetenschappelijke acceptatie	++	-/+	-/+	-	+	-/+	-
Meerdere doel-variabelen	+	-	-	-	+	-	++
Gebruik extra info	Geen	Gemiddelde of totaal van hulp-variabele	Gemiddelde van hulp-variabele	Gemiddelde van hulp-variabele en directe schatter op deelgebieden	Verdeling in de populatie	Kenmerken in populatie	Kenmerken in populatie
Meerdere hulp-variabelen	Nvt	-	+	+	+	+	+
Nominale of ordinale hulpvariabelen	Nvt	-	-	-	++	+	++
Interval of ratio hulpvariabelen	Nvt	++	++	++	-	++	++
Reproduceerbaarheid	++	+	+	+	+	+	-

5.1 Berekenen betrouwbaarheid

Directe schatters:	op basis van de (gestratificeerde) random steekproef kan de steekproeffout worden bekend.
Ratioschatters:	de variantie kan op basis van de genoemde formules worden berekend. Door gebruik te maken van additionele informatie kan de betrouwbaarheid toenemen.
Regressieschatters:	de variantie kan op basis van de genoemde formules worden berekend. Door gebruik te maken van additionele informatie kan de betrouwbaarheid toenemen.
Bayesiaanse schatter:	vereist expert kennis om de variantie van de Bayesiaanse schatter te bepalen
Poststratificatie:	de variantie kan worden berekend. Praktische problemen in het Informatienet zijn ongelijke trekkingskansen zodat aan de veronderstelling van een enkelvoudige aselecte steekproef niet is voldaan.
Modelschatters:	afhankelijk van de aannames die gemaakt worden ten aanzien van de status van de geschatte waarden (worden ze beschouwd als werkelijke waarden) kunnen al dan niet uitspraken worden gedaan over de betrouwbaarheid.
Hot deck-procedures:	de huidige gebruikte methode van Tjomme de Haan biedt geen enkel aanknopingspunt om uitspraken te doen over de betrouwbaarheid. De voorgestelde uitbreiding middels meervoudige imputaties maakt het wel mogelijk enige indicatie te krijgen van de te realiseren betrouwbaarheid.

5.2 Betrouwbaarheid bij kleine aantallen

Directe schatters:	bij kleine aantallen zal de standaardfout groot zijn.
Ratioschatters:	door de ratio op basis van een grotere regio te schatten en deze toe te passen op de regio van onderzoek kan een redelijke betrouwbaarheid worden gerealiseerd.
Regressieschatters:	door de ratio op basis van een grotere regio te schatten en deze toe te passen op de regio van onderzoek kan een redelijke betrouwbaarheid worden gerealiseerd.
Bayesiaanse schatter:	de huidige methode biedt geen mogelijkheden tot het vaststellen van betrouwbaarheden.
Poststratificatie:	bij kleine aantallen is poststratificatie moeilijk toe te passen.
Modelschatters:	afhankelijk van de aannames die gemaakt worden ten aanzien van de status van de geschatte waarden kan al dan niet een redelijke betrouwbaarheid worden gerealiseerd.
Hot deck-procedures:	de huidige methode biedt geen mogelijkheden tot het vaststellen van betrouwbaarheden. Een uitbreiding van de methode maakt dit wel mogelijk, ook bij kleine aantallen.

5.3 Zuiverheid

Directe schatters:	afgezien van niet steekproeffouten is de schatter zuiver.
Ratioschatters:	zuiver als aannames kloppen.
Regressieschatters:	zuiver als aannames kloppen.
Bayesiaanse schatter:	zuiver als aannames kloppen.
Poststratificatie:	zuiver.
Modelschatters:	zuiverheid staat of valt met de juistheid van het model.
Hot deck-procedures:	de huidige methode biedt geen mogelijkheden tot het beoordelen van de zuiverheid. Een uitbreiding maakt een soort validatie op basis van een testset mogelijk.

5.4 Indicatie Goodness of Fit aannames

Directe schatters:	niet van toepassing. Er worden geen aannames gemaakt.
Ratioschatters:	middels de correlatie coëfficiënt en de coëfficiënt van variatie kan worden gekeken of het zinvol is een ratioschatter toe te passen.
Regressieschatters:	de kwaliteit van de regressielijn in termen van het deel van de variantie dat wordt verklaard, kan worden bepaald.
Bayesiaanse schatter:	de R^2 van de regressieschatter geeft een indicatie voor de goodness of fit.
Poststratificatie:	niet van toepassing. Wat betreft het schatten van de variantie is het gebaseerd op de directe schatting. Hierbij worden geen aannames gemaakt.
Modelschatters:	op basis van de R^2 en de t- en F-waarden kan de juistheid van het model worden beoordeeld.
Hot deck-procedures:	de huidige methode biedt geen mogelijkheden tot het vaststellen van de goodness of fit. De voorgestelde uitbreiding biedt wel aanknopingspunten.

5.5 Validiteit bij kleine steekproeven

Directe schatters:	bij kleine aantallen kent de schatting een grote variantie. De validiteit kan dus beperkt zijn.
Ratioschatters:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.
Regressieschatters:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.
Bayesiaanse schatter:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.
Poststratificatie:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.

Modelschatters:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.
Hot deck-procedures:	door het gebruik van aanvullende informatie kan de validiteit sterk verhoogd worden.

5.6 Onderbouwing

Directe schatters:	steekproeftheorie.
Ratioschatters:	model en steekproef.
Regressieschatters:	model en steekproef.
Bayesiaanse schatter:	model en steekproef.
Poststratificatie:	steekproeven.
Modelschatters:	model en steekproef.
Hot deck-procedures:	afstandsmaten. Er bestaat een grote samenhang met methoden zoals cluster analyse waarbij de afstanden tussen objecten centraal staan.

5.7 Eenvoud

Directe schatters:	methode is zeer eenvoudig. Gewoon een kwestie van een (gewogen) gemiddelde.
Ratioschatters:	gebruik van aanvullende informatie is intuïtief goed te volgen.
Regressieschatters:	concept van regressieschatters is iets complexer dan die van ratioschatters.
Bayesiaanse schatter:	modelmatig ingewikkeld. Praktijkuitleg intuïtief redelijk te begrijpen.
Poststratificatie:	opdelen in groepen is een logische activiteit.
Modelschatters:	maken van schattingen van doelvariabelen.
Hot deck-procedures:	

5.8 Bewerkelijkheid

Directe schatters:	absoluut niet bewerkelijk. Gewoon optellen en delen door het aantal waarnemingen.
Ratioschatters:	het gebruik van ratioschatters is veel bewerkelijker. Er moet een bewuste keuze worden gemaakt welke hulpvariabele wordt gebruikt. Tevens moet getoetst worden of aan de voorwaarden voor het toepassen van de ratioschatter wordt voldaan.
Regressieschatters:	het gebruik van regressieschatters is bewerkelijk. Er moet een bewuste keuze worden gemaakt welke hulpvariabelen worden gebruikt.

Bayesiaanse schatter:	het gebruik van de Bayesiaanse schatter is bewerkelijk. Er moet een bewuste keuze worden gemaakt welke hulpvariabelen worden gebruikt. Verder worden aanvullende eisen aan de data en de modelkeuze gesteld.
Poststratificatie:	indien frequentieverdelingen in de populatie beschikbaar zijn voor variabelen in de populatie die men voor de poststratificatie wil gebruiken dan is de methode redelijk simpel toe te passen.
Modelschatters:	het toepassen van modelschatters is zeer bewerkelijk. Het gebruik vergt het volledig doorlopen van het modelspecificatie en schattingsproces, inclusief beoordeling van de kwaliteit van het model.
Hot deck-procedures:	makkelijk toe te passen (indien geschikte programmatuur beschikbaar is). Men moet echter waken voor een garbage-in-garbage-out situatie. Een gedegen toepassing is complex.

5.9 Flexibiliteit

Directe schatters:	zeer gemakkelijk.
Ratioschatters:	de te volgen procedure is duidelijk, maar tijdrovend indien deze opnieuw moet worden uitgevoerd. Afhankelijk van de te veranderen aannames. Is aanname fundamenteel, of niet.
Regressieschatters:	de te volgen procedure is duidelijk, maar tijdrovend indien deze opnieuw uitgevoerd moet worden. Afhankelijk van de te veranderen aannames. Is aanname fundamenteel, of niet.
Bayesiaanse schatter:	de te volgen procedure is duidelijk, maar tijdrovend indien deze opnieuw uitgevoerd moet worden. Afhankelijk van de te veranderen aannames. Is aanname fundamenteel, of niet.
Poststratificatie:	de te volgen procedure is duidelijk, maar tijdrovend indien deze opnieuw uitgevoerd moet worden. Afhankelijk van de te veranderen aannames. Is aanname fundamenteel, of niet.
Modelschatters:	de te volgen procedure is duidelijk, maar tijdrovend indien deze opnieuw uitgevoerd moet worden. Afhankelijk van de te veranderen aannames. Is aanname fundamenteel, of niet.
Hot deck-procedures:	het berekenen van de gemiddelde waarde is zeer gemakkelijk. Als echter de methode van Tjomme de Haan opnieuw uitgevoerd wordt kan dit tijdrovend zijn.

5.10 Wetenschappelijke acceptatie

Directe schatters:	groot, het gebruik van directe schatters is de meest objectieve methode.
Ratioschatters:	afgezien van een meer fundamentele discussie omtrent de waarde van modelgebaseerde schattingsmethoden is de acceptatie

	van de ratioschatter als indirecte schattingsmethode relatief hoog.
Regressieschatters:	afgezien van een meer fundamentele discussie omtrent de waarde van modelgebaseerde schattingsmethoden is de acceptatie van de regressieschatter als indirecte schattingsmethode relatief hoog.
Bayesiaanse schatter:	de Bayesiaanse schattingstechniek ik relatief onbekend. In de statistiek wordt echter op verschillende niveaus veel gebruikgemaakt van Bayesiaanse technieken.
Poststratificatie:	de acceptatie van poststratificatie is hoog.
Modelschatters:	de kritiek op de model gebaseerde aanpak berust op de aanname van het model. Omdat men nooit zeker weet of het juiste model is gespecificeerd prefereren sommigen de model vrije klassieke steekproef theorie. Echter, bij het maken van schattingen voor kleine deelgebieden wordt men wel gedwongen aannames te maken. Gezien de redelijke werking van lineaire modellen in veel sociaal economische processen is het gebruik van modellen bij het maken van schattingen voor kleine gebieden verdedigbaar.
Hot deck-procedures:	er bestaat enige scepsis, maar de acceptatie is de afgelopen jaren groter geworden. Tal van praktijkvoorbeelden zijn bekend.

5.11 Meerdere doelvariabelen

Directe schatters:	elke doelvariabele wordt afzonderlijk geschat. Geen schaalvoordelen bij het schatten van meerdere variabelen.
Ratioschatters:	voor elke doelvariabele moet een afweging worden gemaakt welke hulpvariabele wordt gebruikt. Meerdere doelvariabelen kunnen dus niet tegelijkertijd worden geschat.
Regressieschatters:	voor elke doelvariabele moet een afweging worden gemaakt welke hulpvariabele wordt gebruikt. Meerdere doelvariabelen kunnen dus niet tegelijkertijd worden geschat.
Bayesiaanse schatter:	voor elke doelvariabele moet een afweging worden gemaakt welke hulpvariabele wordt gebruikt. Meerdere doelvariabelen kunnen dus niet tegelijkertijd worden geschat.
Poststratificatie:	voor een specifiek onderzoek met meerdere doelvariabelen kan eenmalig een poststratificatie worden opgesteld die voor dat specifieke onderzoek relevant is. Het schatten van volgende doelvariabelen zal dan ook relatief eenvoudiger zijn.
Modelschatters:	in principe wordt een model gespecificeerd om een doelvariabele te verklaren. Het opnemen van meerdere doelvariabelen vereist het opnieuw specificeren van een model. Het schatten van meerdere doelvariabelen is dan ook zeer bewerkelijk.

Hot deck-procedures:	voor een specifiek onderzoek met meerdere doelvariabelen kan eenmalig een datafusie en imputatie worden uitgevoerd. Voor doelvariabelen die voor wat betreft de keuze van de relevantie van de koppelingsvariabelen samenhangen biedt dit schaalvoordelen. Dezelfde geïmputeerde objecten kunnen voor verschillende (samenhangende) doelvariabelen worden gebruikt.

5.12 Gebruik extra informatie

Directe schatters:	niet van toepassing. Bij directe schatters wordt geen aanvullende informatie gebruikt.
Ratioschatters:	gemiddelde of totaal van een hulpvariabele in de populatie die sterk correleert met de doelvariabele moet bekend zijn.
Regressieschatters:	gemiddelde of totaal van een hulpvariabele in de populatie die sterk correleert met de doelvariabele moet bekend zijn.
Bayesiaanse schatter:	gemiddelde of totaal van de hulpvariabele in alle kleine deelgebieden moet bekend zijn. Verder wordt gebruikgemaakt van de directe schatters van de doelvariabele op alle kleine deelgebieden.
Poststratificatie:	verdeling in de populatie over groepen moet bekend zijn.
Modelschatters:	aanvullende kenmerken worden gebruikt om een verband tussen een doelvariabele en een of meer hulpvariabelen te schatten.
Hot deck-procedures:	aanvullende kenmerken worden gebruikt om de afstand tussen bekende en (gedeeltelijk) onbekende objecten vast te stellen.

5.13 Meerdere hulpvariabelen

Directe schatters:	niet van toepassing. Bij directe schatters wordt geen aanvullende informatie gebruikt.
Ratioschatters:	bij het toepassen van een ratioschatter wordt doorgaans een hulpvariabele gebruikt die sterk correleert met de doelvariabele. In Krishnaiah en Rao (1988) wordt een aanpak beschreven met een multivariate ratioschatter.
Regressieschatters:	bij het toepassen van de simpele vorm van de regressieschatter wordt één hulpvariabele gebruikt die sterk correleert met de doelvariabele. De uitgebreidere vorm houdt ook rekening met meerdere hulpvariabelen.
Bayesiaanse schatter:	het gebruik van meerdere hulpvariabelen is mogelijk.
Poststratificatie:	bij poststratificatie kunnen in principe strata worden gevormd op basis van meerdere stratificatievariabelen.
Modelschatters:	in het regressiemodel om een doelvariabele te verklaren worden waarschijnlijk meerdere hulpvariabelen opgenomen.

Hot deck-procedures:	bij het definiëren van de afstand tussen objecten kunnen meerdere variabelen worden gebruikt.

5.14 Nominale of ordinale hulpvariabele

Directe schatters:	niet van toepassing. Bij directe schatters wordt geen aanvullende informatie gebruikt.
Ratioschatters:	ratioschatters vereisen een interval- of ratiogeschaalde hulpvariabele. Bij nominaal of ordinaal is het begrip correlatie lastig te operationaliseren.
Regressieschatters:	regressieschatters vereisen een interval- of ratiogeschaalde hulpvariabele. Bij nominaal of ordinaal is het verband lastig te operationaliseren.
Bayesiaanse Schatter:	de Bayesiaanse schatter vereist een interval- of ratiogeschaalde hulpvariabele. Bij nominaal of ordinaal is het verband lastig te operationaliseren.
Poststratificatie:	indien nominaal of ordinaal wordt gebruikt voor het definiëren van groepen dan is het gebruik van dergelijke variabelen uitermate geschikt voor poststratificatie.
Modelschatters:	nominaal of ordinale variabelen kunnen meegenomen worden in een regressiemodel. Bijvoorbeeld in de vorm van dummies.
Hot deck-procedures:	bij de koppeling van variabelen kunnen alle typen van variabelen worden gehanteerd. Nominaal of ordinaal zijn met name geschikt voor het definiëren van kritische koppelingsvariabelen.

5.15 Interval- of ratiogeschaalde hulpvariabele

Directe schatters:	niet van toepassing. Bij directe schatters wordt geen aanvullende informatie gebruikt.
Ratioschatters:	interval- of ratiogeschaalde hulpvariabelen zijn bij uitstek geschikt voor een ratioschatter.
Regressieschatters:	interval- of ratiogeschaalde hulpvariabelen zijn bij uitstek geschikt voor een regressieschatter.
Bayesiaanse schatter:	interval- of ratiogeschaalde hulpvariabelen zijn bij uitstek geschikt voor een Bayesiaanse schatter.
Poststratificatie:	interval- of ratiogeschaalde hulpvariabelen zijn lastig bruikbaar. Vergt de definiëring van een aantal klassen. De grenzen tussen de klassen zijn redelijk arbitrair.
Modelschatters:	interval- of ratiogeschaalde hulpvariabelen kunnen worden gebruikt om zoveel mogelijk van de variantie in de doelvariabelen te verklaren.
Hot deck-procedures:	interval- of ratiogeschaalde hulpvariabelen kunnen worden gebruikt om de afstand tussen twee objecten te operationaliseren.

5.16 Reproduceerbaarheid

Directe schatters:	eenvoudig te reproduceren. Zelfde dataset leidt altijd tot dezelfde uitkomsten.
Ratioschatters:	bij het goed documenteren van de keuzen en waarden van de hulpvariabelen zijn de uitkomsten goed te reproduceren.
Regressieschatters:	bij het goed documenteren van de keuzen en waarden van de hulpvariabelen zijn de uitkomsten goed te reproduceren.
Bayesiaanse schatter:	bij het goed documenteren van de keuzen en waarden van de hulpvariabelen zijn de uitkomsten goed te reproduceren.
Poststratificatie:	bij het goed documenteren van de keuze van de strata en de frequenties van deze in de populatie zijn de uitkomsten goed te reproduceren.
Modelschatters:	het model dat gebruikt wordt voor het schatten van waarden van doelvariabelen voor individuele bedrijven moet goed worden gedocumenteerd. Bij beschikbaarheid van het model zijn de uitkomsten reproduceerbaar.
Hot deck-procedures:	de oplossing is afhankelijk van een groot aantal keuzes (variabelen, afstandsmaat enzovoort). Deze keuzes moeten in de programmatuur worden ingevoerd. Reproduceerbaarheid vereist een goede documentatie van deze uitgangspunten.

6. Vernieuwing Informatienet en het schatten van kenmerken van kleine deelgebieden

6.1 Inleiding

Naar aanleiding van de vernieuwing van het Bedrijven-Informatienet van het LEI (het Informatienet) en de overgang van boekjaar naar kalenderjaar is het aantal bedrijven dat in het Informatienet verwerkt wordt voor het jaar 2000 veel kleiner dan het aantal van 1.500 waarvan in andere jaren de gegevens uitgewerkt werd. Het aantal uit te werken boekhoudingen zal voor het jaar 2000 tussen de 200 en 500 uitkomen. Op het moment dat dit rapport werd geschreven, bestond onduidelijkheid over het aantal dat precies uitgewerkt zou gaan worden. Ook van het traject dat gevolgd zou worden na 2000 en de bijbehorende aantallen uitgewerkte boekhoudingen, was nog geen duidelijk beeld ten tijde van uitwerking van dit rapport.

Het uitwerken van een relatief zeer klein aantal boekhoudingen van het jaar 2000 zal voor veel onderzoekers tot problemen leiden. Het tekort aan data in 2000 en eventueel ook in de daaropvolgende jaren is slechts één van de problemen waar onderzoekers en beleidsmakers mee te maken zullen krijgen. Een ander probleem dat zich voordoet is de herdefiniëring van bepaalde variabelen in het nieuwe systeem. Gerelateerde problemen zijn trendbreuken die ontstaan, specifieke gegevens die binnen bepaalde modellen gebruikt worden en die niet jaarlijks, gedeeltelijk, of in het geheel niet beschikbaar zijn in het nieuwe systeem. Ook de representativiteit van de steekproef speelt een rol.

6.2 Aanpak tekort aan data

Voor het tekort aan data in 2000 en andere problemen gerelateerd aan de vernieuwing van het Informatienet is geen eenduidige oplossing te geven. Wat een oplossing is voor het ene probleem zal voor het andere probleem geen oplossing zijn. Voor de bovengenoemde problemen zal een oplossing of oplossingsrichting beschreven worden. De oplossing kan over het algemeen in twee richtingen gezocht worden. De eerste gaat uit van het gebruik van andere databronnen en de tweede maakt gebruik van schattingstechnieken voor kleine deelgebieden. Een combinatie van deze twee oplossingsrichtingen is ook mogelijk. Het vergelijken van verschillende methoden kan inzicht bieden in de gebruikte methoden en de uitkomsten.

6.2.1 Geen gebruikmaken van de gegevens van 2000

Indien gekozen wordt geen gebruik te maken van de gegevens van het Informatienet van 2000, zal gebruikgemaakt moeten worden van andere databronnen. Hierbij kan gedacht worden aan beschikbare gegevens van 1999, maar ook het gebruik van gegevens van 2001, die naar alle waarschijnlijkheid binnenkort beschikbaar zullen zijn. Een andere mogelijk-

heid is gebruik te maken van periodieke voorspellingen zoals gepubliceerd in het *Landbouw-Economisch Bericht (LEB)* (Silvis et al., diverse jaren). Ook gebruikmaken van RICA-gegevens is een optie.

De gegevens van 1999 kunnen uitkomst bieden in het geval geen grote veranderingen hebben plaatsgevonden in een bepaalde sector in de periode 1999/2000. In dat geval zijn de gegevens van 1999 ook voor 2000 representatief voor de sector die weergegeven wordt. Daarnaast is het voor sommige beleidsonderzoeken het te analyseren jaar van minder belang omdat een uitspraak over de normale of toekomstige situatie wordt gevraagd.

Voor het geven van jaarlijkse prognoses is het niet mogelijk de gegevens van 1999 te gebruiken. De voorspelling van een jaar eerder is reeds op deze gegevens gebaseerd. Ook kan het zijn dat veranderingen hebben plaatsgevonden zodat gegevens van 1999 niet meer bruikbaar zijn bij het maken van een analyse. Een mogelijkheid is in dit geval gebruik te maken van gegevens van het 1e kwartaal van 2001. Naar alle waarschijnlijkheid zullen van het jaar 2001 gegevens van 800 tot 1.100 bedrijven uitgewerkt worden. Dit is een aanzienlijke verbetering ten opzichte van het jaar 2000. Voordeel van de invoering van ARTIS is dat gegevens niet over een geheel jaar verwerkt worden en ruim een jaar na beëindiging van een bepaald jaar beschikbaar komen. Ingevoerde gegevens zijn direct na invoering beschikbaar en kunnen per kwartaal verwerkt worden.

In het *LEB* worden ramingen gemaakt van financieel economische gegevens voor een aantal sectoren. Deels worden ramingen gebaseerd op gegevens van het CBS en deels op gegevens uit het Informatienet. Eind 2000 en juni 2001 werden in het *LEB* ramingen gepubliceerd voor het jaar 2000. De ramingen zijn gebaseerd op (voorlopige) gegevens van 1998 en 1999. Ook in 2001 zullen prognoses van financieel economische gegevens gepubliceerd worden in het *LEB*. Informatienet-gegevens van 2000 zijn echter niet of nauwelijks aanwezig, dus ramingen voor 2001 die uitgaan van Informatienet-gegevens kunnen niet op beschikbare gegevens van 2000 gebaseerd worden. De gekozen oplossing gaat uit van de ramingen die gemaakt zijn voor 2000 en de beschikbare (voorlopige) gegevens van 1999. In dit voorbeeld worden de gepubliceerde ramingen voor het jaar 2000 als beschikbare gegevens gebruikt. Ook in andere onderzoeksprojecten die financieel-economische gegevens voor 2000 betreffen zou overwogen kunnen worden de ramingen uit het *LEB* te gebruiken.

Het Informatienet is uniek in Nederland wat betreft de gegevensverzameling, zowel in omvang als in detail. Door verschillende onderzoeksbureaus en organisaties wordt echter 'aanvullend' onderzoek verricht op agrarisch gebied. Ook enquêtering en gegevensverzameling spelen hierbij een rol. In het geval onderzoekers naar zeer specifieke informatie op zoek zijn, zoals informatie over een bepaalde regio of voor een bepaalde deelsector, zouden zij bepaalde organisaties kunnen benaderen die op dit gebied een gegevensverzameling beheren. Zodoende kan verzamelde informatie eventueel beschikbaar worden gesteld voor onderzoek door LEI-medewerkers.

Als laatste komt een mogelijke oplossing voor het trendbreukprobleem aan de orde. Door onder meer herdefiniëring van variabelen in het nieuwe systeem ontstaan trendbreuken. Een variabele die in het oude systeem een bepaalde betekenis had, heeft in het nieuwe systeem een andere betekenis gekregen. Hierbij zou het kunnen gaan om herschaling, maar ook om hergroeperingen. De betekenis van variabelen in het nieuwe systeem dient men in de gaten te houden. Een WOT (Wettelijke Onderzoekstaak) van het LEI is het leveren van

gegevens aan de Europese Unie. De gegevens die in het RICA terechtkomen zijn echter uniform gedefinieerd. Dit betekent dat de vastgelegde gegevens elk jaar dezelfde betekenis hebben. Als benodigde gegevens voor onderzoek terug te vinden zijn in het RICA, dan kan gebruik worden gemaakt van deze gegevens en zijn herdefiniëring en trendbreuk niet langer een probleem.

6.2.2 Gebruikmaken van de gegevens van 2000

Als ervoor wordt gekozen gebruik te maken van de beschikbare gegevens van 2000, is er hoogstwaarschijnlijk sprake van een zeer klein databestand. Een aantal technieken besproken in dit rapport kunnen in dat geval een bijdrage leveren aan de berekening van kengetallen, gemiddelden en totalen. Voor een groot deel zal hierbij gebruikgemaakt worden van databestanden als de Landbouwtelling en eventueel gegevens uit het Informatienet van andere jaren.

Van de besproken modellen kunnen voornamelijk ratio- en regressieschatters, datafusie en imputatie een rol spelen bij het maken van schattingen voor 2000. De methoden zijn ook van toepassing op het maken van prognoses voor 2001 en eventueel 2002 of 2003.

De methode die gebruikmaakt van directe schatters wordt niet aangeraden aangezien de steekproef voor het jaar 2000 niet representatief zal zijn voor de gehele populatie. Van de ongeveer 1.000 bedrijven die geselecteerd zijn om in 2000 deel te nemen aan het Informatienet, zullen in het ergste geval 250 bedrijven uitgewerkt worden. Deze 250 bedrijven vormen ten eerste geen aselecte steekproef binnen de verschillende gedefinieerde strata en ten tweede is het totaal aantal bedrijven in de steekproef dusdanig klein dat extrapolatie van schattingen naar de gehele populatie niet aan te raden is.

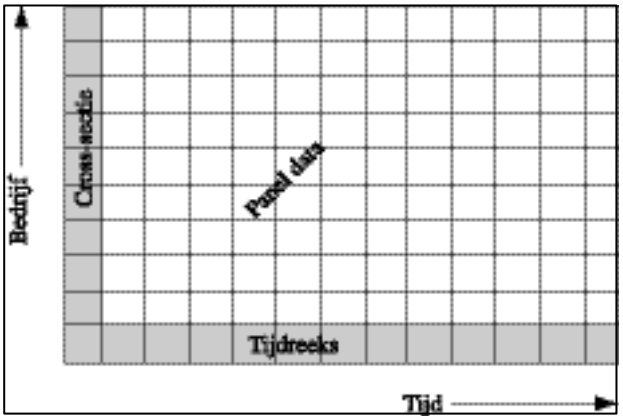
Ratio- en regressieschatters maken gebruik van gegevens uit de Landbouwtelling en kunnen dan ook gebruikt worden in het geval zeer weinig gegevens beschikbaar zijn. Regressie- en ratioschatters zullen in dit hoofdstuk in één adem genoemd worden aangezien de ratioschatter een gerestricteerde versie van de regressieschatter is en kunnen dan ook volgens dezelfde procedure gebruikt worden.

Poststratificatie is een methode om de betrouwbaarheid van de schattingen te vergroten. Op de data van 2000 is deze methode echter niet van toepassing, aangezien de data niet representatief zijn. De betrouwbaarheid kan wel vergroot worden, maar in het geval schattingen absoluut niet nauwkeurig zijn heeft dit relatief weinig zin.

Het gebruik van datafusie en imputatie is zinvol indien regressiemodellen worden gebruikt om data te imputeren. Hot deck-procedures zijn minder zinvol voor het imputeren van gegevens in de dataset van 2000, omdat deze geen gebruikmaken van extra gegevens. Hot deck-procedures maken gebruik van ongebruikte gegevens uit de eigen dataset. Indien schattingen van een grootheid gemaakt worden voor een bepaalde regio biedt de hot deck-procedure uitkomst. Informatie van bedrijven uit andere regio's wordt gebruikt voor het maken van een schatting. In 2000 is echter sprake van een dusdanig kleine dataset, dat problemen ontstaan bij schattingen voor het gehele land en in dat geval is geen extra informatie beschikbaar. Imputatie middels regressiemodellen is zinvol aangezien gebruikgemaakt kan worden van extra gegevens uit de Landbouwtelling.

6.2.3 Procedure

De methoden die een mogelijke oplossing bieden voor dataproblemen in 2000 gaan uit van hetzelfde principe. Regressiemodellen worden gebruikt eventueel in combinatie met gegevens uit de Landbouwtelling om aanvullende informatie te verkrijgen. Indien gebruikgemaakt wordt van gegevens uit de Landbouwtelling, wordt gezocht naar verbanden tussen variabelen uit het Informatienet en variabelen uit de Landbouwtelling.



Figuur 6.1 Samenhang data verschillende modellen

Verklarende en voorspellende modellen kunnen op verschillende manieren geformuleerd worden om een betere schatting te maken van grootheden in 2000. De eerste mogelijkheid, besproken in eerdere paragrafen, gaat uit van cross-sectiedata. Dat wil zeggen gegevens van meerdere bedrijven over één bepaald jaar. De tweede mogelijkheid (evenals de derde, niet eerder besproken) gaat uit van gegevens van één bepaalde grootheid of één bepaald bedrijf genomen over meerdere jaren. Modellen die hiervan uitgaan worden ook wel tijdreeksmodellen genoemd. De derde mogelijkheid, die uitgaat van paneldata, benut de mogelijkheden van het Informatienet optimaal. Paneldata zijn gegevens van verschillende bedrijven beschikbaar voor meerdere jaren. Aangezien het Informatienet de vorm van een roterend panel aanneemt, zijn gegevens uit het Informatienet geschikt voor paneldata modellen. De verschillende modellen en bijbehorende mogelijkheden zullen uitgebreider besproken worden.

Cross-sectiemodellen zijn besproken in de paragrafen over regressiemodellen, en imputatie middels regressiemodellen. Een probleem dat speelt bij deze modellen is dat de relatie tussen de te verklaren variabele uit het Informatienet en de verklarende variabele uit de Landbouwtelling geschat dient te worden aan de hand van de beschikbare gegevens in het Informatienet. Als het gaat om bijvoorbeeld 10 beschikbare bedrijven in het Informatienet, dan is de betrouwbaarheid van de geschatte coëfficiënt(en) in het model zeer klein. In dit geval kunnen mogelijkheden omtrent tijdreeksmodellen of paneldatamodellen bekeken worden. Een andere mogelijkheid is het schatten van de coëfficiënt(en) in het model aan de hand van gegevens van het voorgaande jaar. Voorwaarde hierbij is de juistheid van

de veronderstelling dat de relatie tussen afhankelijke en onafhankelijke variabelen hetzelfde is gebleven ten op zichte van het voorgaande jaar.

Tijdreeksmodellen worden niet gebruikt om relaties mee te schatten, maar voornamelijk voor het vastleggen van trends en patronen in variabelen waarvan waarden op verschillende tijdstippen vastgelegd zijn. Vaak worden tijdreeksmodellen gebruikt voor voorspellingen van waarden in de toekomst. Aan de hand van gegevens van een bepaalde variabele in het verleden wordt een relatie geschat waarin de geschatte waarde van de variabele op een bepaald tijdstip afhankelijk wordt gesteld van de waarde van die variabele een periode eerder. De geschatte relatie wordt vervolgens verondersteld ook in de toekomst te bestaan. Zodoende kunnen toekomstige waarden voorspeld worden. De voorspelde waarden zouden in de dataset van 2000 geïmputeerd kunnen worden.

In paneldata zijn voor verschillende bedrijven tijdreeksen van een variabele vastgelegd. Dit houdt in dat de voordelen van tijdreeksen en cross-secties gecombineerd zijn in paneldatamodellen. Het aantal vrijheidsgraden neemt toe met het aantal waarnemingen dat beschikbaar is. Voor paneldatamodellen zijn in tegenstelling tot cross-sectiemodellen zowel gegevens van een variabele voor verschillende bedrijven als op verschillende momenten in de tijd vastgelegd. Het genoemde probleem van een tekort aan data waardoor de relatie tussen verschillende variabelen niet geschat kan worden zou opgelost kunnen worden door gebruik te maken van paneldatamodellen. Waarden van coëfficiënten in de tijd kunnen hierbij verschillen. Bij eventuele voorspellingen zou dan ook rekening gehouden kunnen worden met patronen en trends die zich voordoen in de tijd. Een probleem is echter dat voor het Informatienet een roterende panel gehanteerd wordt. Elk jaar wordt een deel van de steekproefbedrijven vervangen. Niet elk bedrijf zal de gehele te onderzoeken periode in de dataset aanwezig zijn geweest. Statistische pakketten kunnen veelal zonder problemen omgaan met ontbrekende gegevens.

7. Samenvatting en conclusies

De verschillende methoden die het mogelijk maken op een klein deelgebied te schatten, hebben ieder hun eigen sterke en zwakke punten. Niet elke methode is geschikt om in elke situatie toe te passen. In veel gevallen heeft dit betrekking op de data, maar ook op de voorkeuren van de onderzoekers. De afweging tussen de tijd die aan een berekening wordt besteed en de kwaliteit van de resultaten moet elke keer worden gemaakt. Het berekenen van een directe schatter is bijvoorbeeld vele malen eenvoudiger en minder bewerkelijk dan het berekenen van een ratioschatter of een regressieschatter. Dit gaat echter wel ten koste van de betrouwbaarheid. De betrouwbaarheid van schattingen kan vergroot worden door uit te gaan van vooraf gedefinieerde aannames die geëxpliciteerd worden in een model. Indien een correct model wordt gespecificeerd, zal een betere schatting gemaakt kunnen worden dan wanneer directe schatters gebruikt worden. Wanneer de aannames echter onterecht zijn, is er sprake van een bias in de resultaten en zal de schatting van de standaardfout te optimistisch zijn.Voor elke schatting dient de onderzoeker zijn eigen criteria vast te stellen en aan de hand van deze criteria en de beschikbare data een keuze moeten maken tussen de verschillende methoden.

Onderstaande tabel geeft een vergelijking van de resultaten die de verschillende schattingsmethoden opleveren met betrekking tot het voorbeeld over de arbeidsopbrengsten ondernemers in de kalvermesterij. Dit overzicht is niet volledig, aangezien de illustratie voor de Bayesiaanse schatter en de verschillende imputatiemethoden een ander voorbeeld betrof.

Tabel 7.1 Vergelijking resultaten schattingen in voorbeelden

	Schatting voor $\overline{Y}$	Variantie schatting	Standaardfout
Directe schatter aselecte steekproef	73.086	254.772.886	15.962
Directe schatter met stratificatie vooraf	65.351	59.963.039	7.744
Ratioschatter	55.767	85.437.591	9.243
Regressieschatter	42.216	45.014.842	6.709
Poststratificatieschatter	47.919	75.512.833	8.690

De betrouwbaarheid van de directe schatter op basis van de aselecte steekproef is zeer klein in vergelijking met die van andere schatters. De directe schatter op basis van de gestratificeerde steekproef heeft een grotere betrouwbaarheid dan de poststratificatieschatter. Wanneer de steekproef echter voor vele verschillende doeleinden gebruikt wordt, is stratificatie vooraf niet aan te raden.

De schatter die gebruikmaakt van stratificatie vooraf heeft een grotere betrouwbaarheid dan de directe schatter in het geval de stratificatievariabele gelijk is aan of in hoge mate correleert met de doelvariabele. Is er echter geen duidelijk verband tussen de stratificatievariabele en de doelvariabele dan kan de variantie van de stratificatieschatter zelfs groter zijn dan de directe schatter in de aselecte steekproef.

Als de directe schatter vergeleken wordt met de ratioschatter blijkt dat de directe schatting een grotere waarde heeft dan de indirecte schatter. Dit verschil is te verklaren uit het feit dat het gemiddelde aantal kalveren van Informatienet-bedrijven hoger ligt dan het gemiddelde aantal kalveren van bedrijven in de Landbouwtelling. Doordat de indirecte schatter meer informatie in de schatting betrekt is de stelling te verdedigen dat de indirecte schatter een waarheidsgetrouwere schatting oplevert.

Als de ratioschatter vergeleken wordt met de regressieschatter blijkt dat de regressieschatting lager ligt dan de ratioschatting. Dit verschil is te verklaren uit het feit dat het verband tussen het aantal kalveren en de opbrengst niet lineair door de oorsprong gaat. De ratioschatter legt in dit geval een onterechte restrictie op aan het model. De ratioschatter die een verband door de oorsprong verondersteld leidt dan ook tot een hogere schatting en een kleinere betrouwbaarheid.

Verder is bekend dat de Bayesiaanse schatter een waarde heeft die tussen de waarde van de regressieschatter en die van de directe schatter in ligt.

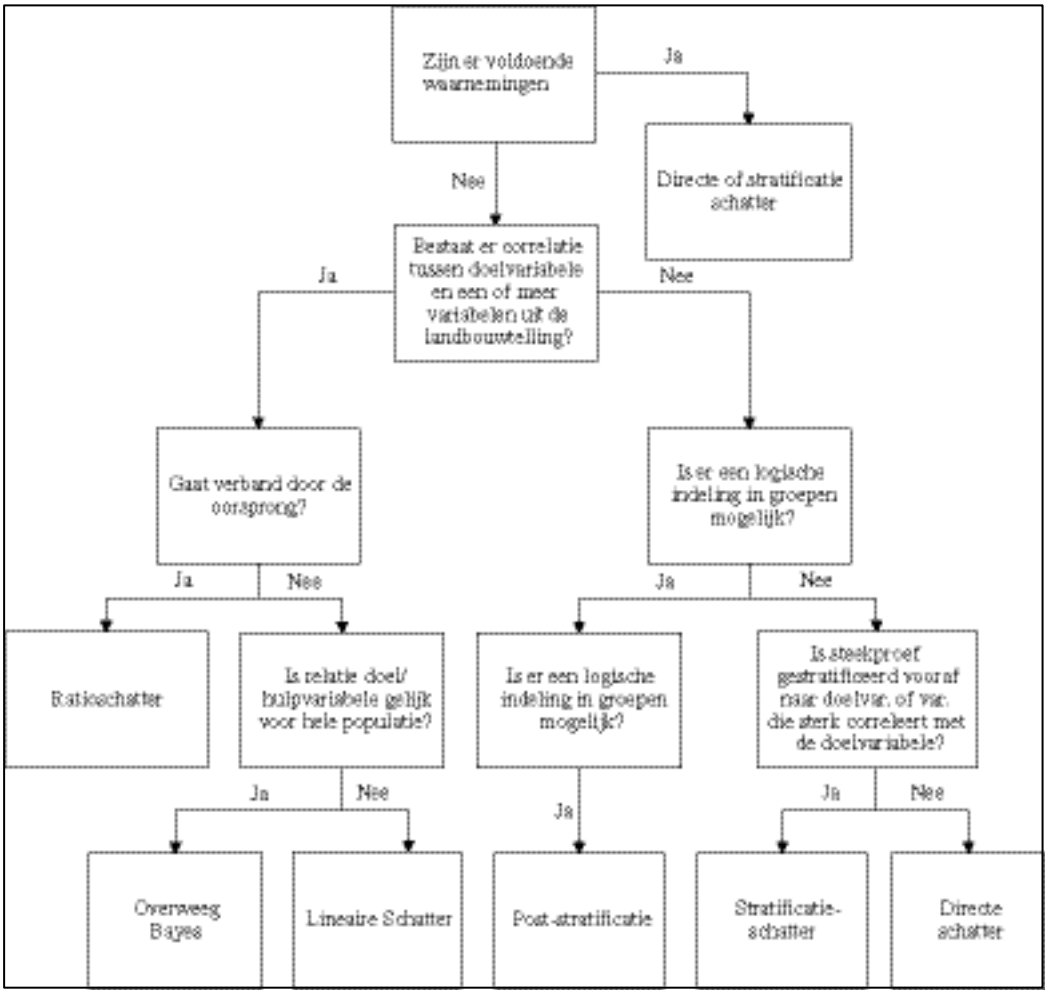
8. Implicaties voor het onderzoek

Verschillende methoden om te schatten brengen verschillende mogelijkheden met zich mee. Elke methode heeft zijn eigen voor- en nadelen. Er bestaat dan ook geen methode die in elke situatie de beste resultaten geeft. De keuze tussen de verschillende methoden is onder meer afhankelijk van de beschikbare data, de mogelijkheid tot het combineren van data uit verschillende bronnen, de deelpopulatie en de aannames die eventueel gemaakt kunnen worden. Voor het oplossen van elk probleem moet opnieuw de afweging tussen de verschillende methoden worden gemaakt. Dit rapport geeft de richting aan waarin de oplossing gezocht kan worden. In figuur 8.1 is geprobeerd een algemeen overzicht te geven van de te volgen procedure wanneer een schatting gemaakt dient te worden van een kenmerk, zoals een gemiddelde of een totaal, van een populatie of een deel van een populatie.

Als eerste dient nagegaan te worden of correlatie bestaat tussen de doelvariabele en één of meer hulpvariabelen. Voorwaarde hierbij is wel dat meer waarnemingen beschikbaar zijn voor de hulpvariabele dan voor de doelvariabele. Voor onderzoek binnen het LEI zou bijvoorbeeld een vraag over een bepaalde variabele in het Informatienet kunnen ontstaan. Mogelijke hulpvariabelen zouden beschikbare variabelen uit de Landbouwtelling zijn.

Wanneer sprake is van een relatie tussen bepaalde doel- en hulpvariabelen, ontstaat de vraag of de relatie door de oorsprong gaat. Als dit het geval is, kan het beste gebruikgemaakt worden van de ratioschatter. Als dit niet het geval is, dient de afweging gemaakt te worden of de relatie tussen doel- en hulpvariabele niet alleen op het kleine deelgebied maar voor de hele populatie geldt. Als deze aanname niet gemaakt kan worden dan kan het beste gebruikgemaakt worden van de lineaire regressieschatter. Kan deze aanname wel gemaakt worden dan kan een Bayesiaanse schatter overwogen worden. Als de variantie voor alle deelpopulaties echter verschilt, wordt het uitvoeren van de Bayesiaanse analyse zeer bewerkelijk. Als dit niet gewenst is, kan ook in dit geval de lineaire regressieschatter overwogen worden.

Als geen verband bestaat tussen de doelvariabele en eventuele hulpvariabelen dan komt men in de hoek van de directe schatters terecht. Het probleem in deze hoek is dat vaak te weinig waarnemingen beschikbaar zijn, waardoor de betrouwbaarheid van deze schatters zeer klein wordt. Een vuistregel is dat directe schattingen op basis van minder dan 20 waarnemingen geen betrouwbare resultaten geven. Is dit aantal waarnemingen niet beschikbaar, dan moet een datafusie of imputatie overwogen worden. Zijn er wel genoeg waarnemingen dan kan de procedure van de meest eenvoudige directe schattingsmethode gevolgd worden.



Figuur 8.1 Gebruik van schatters

Als de steekproef vooraf gestratificeerd is naar de doelvariabele of een variabele die in grote mate gecorreleerd is met de doelvariabele kan gebruikgemaakt worden van de stratificatieschatter. Als dit niet het geval is dient nagegaan te worden of vooraf een verdeling bekend is van de grootte van verschillende groepen in de populatie, of als een logische groepenindeling gemaakt kan worden. Als dit het geval is kan gebruik worden gemaakt van de poststratificatieschatter, als dit niet het geval is kan de directe schatter gebruikt worden.

Literatuur

Baker, K., P. Harris en J. O'Brien, 'Data Fusion: An Appraisal and Experimental Evaluation'. In: *Journal of The Market Research Society* 31 (2), 152-212. 1994.

Buck, S., 'Single Source Data -The Theory and the Practice'. In: *Journal of the Market Research Society* 31 (4), 489-500. 1989.

Bronner, A.E., 'Einde fusiefobie in Nederland?'. In: *Jaarboek van de Nederlandse Vereniging van Marktonderzoekers* 1988/1989, 9-18. 1988.

Cochran, W.G.C., *Sampling Techniques*. Wiley, New York, 1977.

Dijk, van J.P.M., K. Lodder en H.C.J. Vrolijk, *De steekproef voor het Bedrijven-Informatienet van het LEI.* Rapport 1.01.02. Den Haag, LEI, 2002.

Dol, W., *Small area astimation, a synthesis between sampling theory and econometrics*. Wolters Noordhoff, Groningen, 1991.

Dudewicz, E.J. en S.N. Mishra, *Modern mathematical statistics*. Wiley, Singapore, 1988.

Ford, B., 'An Overview of Hot-Deck Procedures'. In: *Incomplete Data In Sample Surveys*. Academic Press, Volume II, Part 2, 185-207. 1983.

Gelman, A., G. King en C. Liu, 'Not asked and not answered: Multiple imputation for Multiple Surveys'. In: *Journal of the American Statistical Association.* September 1998, Vol. 93, No. 443, pp. 846-857. 1998.

Greene, W.H., *Econometric analysis.* Fourth edition. Upper Saddle River, Prentice Hall, 2000.

Krishnaiah, P.R. en C.R. Rao, *Handbook of statistics 6: Sampling, North-Holland, Amsterdam-NewYork-Oxford.* 1988.

Levy, P.S. en S. Lemeshow, *Sampling of populations, Methods and Applications*. Wiley, New York, 1991.

Muilwijk, J., T.A.B. Snijders en J.J.A. Moors, *Kanssteekproeven*. Stenfert Kroese, Leiden, 1992.

O'Brien, S., 'The Role of Data Fusion in Actionable Media Targeting in the 1990s'. In: *Marketing & Research Today* 19 (February), 15-22. 1991.

McNally, J.W., *Generating Hot-Deck imputation estimates: Using SAS for simple and multiple imputation allocation routines* Working paper 97-12. Population Studies and Training Center, Brown University, 1997.

McNally, J.W., S. Sassler en R. Schoen, *Misplaced affection, the use of multiple imputation to reconstruct missing cohabiting partner information in the NSFH.* Working paper 97-09. Population Studies and Training Center, Brown University, 1997.

Montalto, C.P. en J. Sung, 'Multiple Imputation in the 1992 Survey of Consumer Finances'. In: *Financial Counseling and Planning* 7, 133-146. 1996.

Roberts, A., 'Media Exposure and Consumer Purchasing: An Improved Data Fusion Technique'. In: *Marketing and Research Today* 22 (August) 159-172. 1994.

Sarndal, C.E., B. Swensson en J. Wretman, *Model Assisted Survey Sampling*. Springer Verlag, New York, 1992.

Santini A., J. en G., 'Fusion Techniques: Alternative to Single Source Methods?'. In: *European Research* 15 (August), 178-187. 1987.

Silvis, H.J. en C. van Bruchem, *Landbouw Economisch Bericht 2000*. LEI, Den Haag, 2000.

Silvis, H.J. en C. van Bruchem, *Landbouw Economisch Bericht 2001*. LEI, Den Haag, 2001.

Thompson, S.K., *Sampling*. Wiley, New York,1992.

Vrolijk, H.C.J. en K.Lodder, *Voorstel tot vernieuwing van het steekproefplan voor het Bedrijven-Informatienet.* LEI, Den Haag, 2002.

Vrolijk, H.C.J. and M.Wedel, *Een Datafusie-procedure voor het maken van kruistabellen, in: Jaarboek van de Vereniging voor Marktonderzoek en Informatiemanagement*. pp. 95-106. Uitgeverij de Vrieseborch, Haarlem, 1996.

Bijlage 1 Checklist interviews

- Gebruikt de onderzoeker steekproeven en zo ja welke steekproeven worden gebruikt?

- Op welke manier maakt de onderzoeker op basis van deze steekproeven schattingen van populatie parameters?

- In hoeverre worden uitspraken gedaan over delen van de populatie (voorbeelden)?

- Hoe worden deze uitspraken gedaan en in hoeverre is de onderzoeker op de hoogte van technieken voor kleine deelgebieden?

- In welke mate en in welke situaties worden deze technieken toegepast (voorbeelden van LEI-onderzoek)?

- Wat zijn de ervaringen met het gebruik van deze technieken (welke problemen, wat ging goed, voldoende kennis, tevredenheid enzovoort)?

- Welke wensen bestaan er ten aanzien van toekomstig gebruik (toepassingsgebieden, methoden, ondersteuning …?

- Overige punten:
  - gebruik van varianties/standaardfouten;
  - gebruik statistische toetsen;
  - weging en stratificatie in BDL.

Bijlage 2 Interviewverslagen

De hier vermelde interviewverslagen zijn voorgelegd aan en gecorrigeerd door de betrokkenen.

John Helming

Bij AEOS wordt gebruikgemaakt van een steekproef in de vorm van het Bedrijven-Informatienet van het LEI (het Informatienet). In het onderzoek richt men zich op de 14 landbouwgebieden en op circa 25 activiteiten. Voor het maken van schattingen voor alle 14 gebieden zijn vaak niet voldoende waarnemingen beschikbaar. Daarom zijn de 14 gebieden samengevoegd tot 3 overkoepelende gebieden. Schattingen voor deze 3 gebieden worden weer gedesaggregeerd naar de 14 gebieden (het geschatte gemiddelde op hoger niveau wordt van toepassing verklaard op de gebieden op een lager niveau). Alternatieve indelingen die wellicht resulteren in andere schattingen zijn niet geëvalueerd.

Verder wordt er gebruikgemaakt van ratio's. Zo worden bijvoorbeeld schattingen gemaakt van gegevens per varken. Deze schatting in combinatie met gegevens uit de Landbouwtelling worden gebruikt voor het maken van schattingen voor bepaalde gebieden. Hierbij worden alleen gemiddeldes geschat, aan de variantie van deze schatting wordt geen aandacht besteed. De schatting van het gemiddelde wordt vervolgens in simulaties gebruikt.

Binnen het onderzoek wordt geen gebruikgemaakt van statistische toetsen. Men werkt uitsluitend met gemiddeldes, trends bestaan uit een reeks van gemiddeldes. De redenen voor het niet gebruiken van statistische toetsen zijn een gebrek aan tijd, gebrek aan kennis en het feit dat dit niet in de opdracht wordt gevraagd.

Men is wel geïnteresseerd in het gebruik van varianties, statistische toetsen en andere schatters. Men is altijd geïnteresseerd om de kwaliteit te verbeteren. De kennis moet wel gratis ter beschikking zijn. Het zou handig zijn om de medewerkers te scholen.

Wil Hennen

De sectie AM maakt gebruik van het Informatienet als steekproef. In het Approxi-model worden bedrijven ingewogen middels de weging. Het bedrijf representeert zoveel bedrijven als de weging groot is.

Als er een uitspraak wordt gedaan over een bedrijfstype worden de beschikbare bedrijven opgehoogd naar de deelpopulatie middels de beschikbare gewichten.

Voor kleine regio's zullen er in veel gevallen niet voldoende bedrijven beschikbaar zijn. In dergelijke gevallen wordt op basis van de beschikbare Landbouwtellingsgegevens voor elk bedrijf in die regio gezocht naar een zo sterk mogelijk gelijkend bedrijf in het Informatienet. De vergelijking wordt gemaakt op basis van 20 criteria. De exacte keuze van criteria is afhankelijk van het doel van het onderzoek. Is men bijvoorbeeld geïnteresseerd

in de vergrijzing dan licht een keuze van de variabelen zoals leeftijd en opvolgingssituatie voor de hand, is men geïnteresseerd in milieu effecten dan zal men eerder kiezen voor variabelen zoals de intensiteit. De vergelijking resulteert in een soort van best en second best keuze van vergelijkbare bedrijven. Op basis van de fit tussen de oplossing en de nagestreefde situatie wordt de keuze geoptimaliseerd. De fit wordt uitgedrukt in termen van het gemiddelde. De variantie wordt hoogstens meegenomen in termen van een minimalisatie van de variantie en niet zozeer door een vergelijking van de daadwerkelijk variantie en de variantie van de oplossing. Nadat voor elk bedrijf een zo goed mogelijk gelijkend bedrijf is gevonden wordt met deze deterministische oplossing gewerkt. Vervolgens kunnen bijvoorbeeld gemiddeldes voor kleine regio's worden berekend. De stellige indruk bestaat dat dit tot betere schattingen leidt dan wanneer men op basis van bijvoorbeeld slechts een of twee beschikbare bedrijven een schatting maakt voor een regio. Varianties worden niet berekend omdat nog geen enkele opdrachtgever daar naar heeft gevraagd. Verder geldt dat de hier genoemde methode niet zo zeer een statistische methode is maar een methode waarin meer gebruik wordt gemaakt van expertkennis. Dit is een andere aanpak van hetzelfde probleem.

Een eerdere variant van de hier beschreven procedure is bijvoorbeeld toegepast in het grondbalansen onderzoek van Tjomme de Haan.

Machiel Mulder

De afdeling SO werkt in de huidige praktijk weinig met steekproeven. Steekproeven komen wel voor bij het houden van enquêtes.

Bij het doen van uitspraken over regio's komt wel een soortgelijke problematiek als die bij het schatten van kenmerken van kleine deelgebieden aan de orde. Omdat het vaak niet mogelijk is om het gehele onderzoek voor elke regio uit te voeren wordt geprobeerd uitkomsten naar andere gebieden te transfereren (middels value transfer). Op basis van specifieke kenmerken van het gebied wordt een waarde getransfereerd (een probleem hierbij is dat de selectie van de oorspronkelijke gebieden eigenlijk niet a select kan worden genoemd).

FES maakt gebruik van Informatienet-gegevens. Voor het doen van uitspraken over groepen worden de gegevens opgehoogd met behulp van de Informatienet gewichten. De indeling in groepen, de typering, is gebaseerd op de Informatienet gegevens. Eventuele gevolgen van deze basis worden niet standaard beschouwd. In FES worden de betrouwbaarheden niet uitgerekend. In het verleden is dit door de afdelingsleiding afgeraden. De reden hiervoor is onduidelijk, wellicht was men bang voor de uitkomsten en daarnaast geldt dat opdrachtgevers er meestal niet om vragen.

Met FES worden er niet vaak uitspraken gedaan over regionale gebieden. FES is vaak gericht op nationale problemen. Er wordt weinig gedaan aan regioproblemen bij bijvoorbeeld de provincie. Verder geldt het probleem dat het aantal bedrijven in een regio vaak laag is. Men gaat hoogstens uit van de reeds gehanteerde indeling (bijvoorbeeld Westland versus rest van Nederland).

Bij SO bestaat regelmatig de behoeft uitspraken te doen over regio's. In onderzoek wordt regelmatig een situatieschets voor een regio gemaakt. Hierbij zou men graag ook iets zeggen over financieel economische kengetallen van de landbouw. In de huidige situatie wordt vaak op basis van gegevens van het CBS een uitspraak gedaan over een regio. Deze

gegevens zijn vaak van een hoger aggregatie niveau (bijvoorbeeld provincie) en moeten dus vertaald worden naar kleinere regio's.

De markt voor REM groeit. Gegevens is vaak een probleem. Gegevens verkregen middels technieken voor kleine deelgebieden zouden hier deels een oplossing voor kunnen bieden. Hiertoe zouden wel een aantal standaardtechnieken beschikbaar moeten komen. Het is te complex en het vergt te veel tijd als je je hier als onderzoeker in moet gaan verdiepen. Er bestaat wel enthousiasme voor het idee. Des te meer gebruik je kunt maken van gegevens des te beter. Hierbij geldt wel de randvoorwaarde dat de kwantitatieve kennis bij SO beperkt is. Men wil zich wellicht wel verdiepen in de materie, maar het belang voor het onderzoek moet heel duidelijk zijn.

Piet Rijk

Gebiedsgericht onderzoek betreft vaak een regio bestaande uit een paar gemeentes. Het aantal landbouwbedrijven in een dergelijke regio bedraagt doorgaans tussen de 300 en 600 (in enkele gevallen zijn de regio's kleiner, bijvoorbeeld 100 bedrijven). Er is niet echt een ontwikkeling te bespeuren naar nog kleinere regio's. De vraag naar gebiedsgericht onderzoek fluctueert sterk over de afgelopen jaren. Het aantal opdrachten gaat op en neer.

Bij gebiedsgericht onderzoek wordt vaak geen steekproef gebruikt. In veel gevallen worden alle bedrijven zoals die in de Landbouwtelling voorkomen meegenomen. Wanneer men wil komen tot een inkomensplaatje wordt gebruikgemaakt van een steekproef in de vorm van het Informatienet. Bij een verkenning probeert men voor clusters en typen op basis van gegevens uit het boekhoudnet een schatting te maken. Gezinsinkomen uit bedrijf per ondernemer is hierbij een belangrijke variabele. Bij een minimum van 10 à 15 per type wordt een directe schatting van het gemiddelde gemaakt. De gewichten vanuit het Informatienet worden hierbij niet gebruikt.

Als het aantal bedrijven kleiner is dan 10 dan wordt bijvoorbeeld op basis van gesprekken een inschatting gemaakt. Een variant hiervan is dat men op basis van het teeltplan en schattingen van de opbrengst per hectare een schatting maakt van het inkomen.

Vergelijking van bijvoorbeeld de relatieve omvang van de bedrijven in de regio ten opzichte van die in het Informatienet biedt aanknopingspunten voor het maken van een schatting. Als de bedrijven 10% groter zijn dan de bedrijven in het Informatienet dan wordt bijvoorbeeld verondersteld dat het inkomen ook 10% groter is dan die van de bedrijven in het Informatienet. Ook komt het voor dat men voor het maken van een schatting naar bedrijven in een iets ruimere regio kijkt.

Onafhankelijke van de gehanteerde methodiek, in de rapportages komen geen varianties of standaardfouten aan de orde. Dit is over het algemeen te moeilijk voor de lezers van de rapportages. Wel maakt men regelmatig vergelijkingen tussen regio's en tussen jaren. Voor deze vergelijkingen wordt geen gebruikgemaakt van betrouwbaarheidsintervallen van schattingen.

Tevens komt het voor dat men een steekproef trekt voor een nader onderzoek. Bij 25 à 30 bedrijven wordt middels een enquête aanvullende gegevens verzameld.

Er bestaat tevredenheid over de op dit moment gehanteerde werkwijze. De opdrachtgever vraagt niet zo diep. Het berekenen van standaardfouten enzovoort zou te breed en te uitvoerig zijn. Er bestaat belangstelling voor technieken voor het schatten van kenmerken

van kleine deelgebieden. Als ze er zijn en als ze makkelijk toegankelijk zijn dan wil men ze graag gebruiken. Een standaardpakket waarbij de berekeningen om inkomens te schatten zijn geautomatiseerd is wenselijk (bijvoorbeeld het vroegere GPS).

Met name bestaat een zeer concrete behoefte aan betere inkomensgegevens voor kleine gebieden. In het verleden was er een programma beschikbaar waarmee op basis van een schatting van het inkomen per sbe een schatting voor een klein gebied kon worden gemaakt. Ook de methode waarin Landbouwtellingsbedrijven worden vervangen door soortgelijke Informatienet-bedrijven vindt men een aantrekkelijke optie.

Harry Luesink

Er wordt in ruime mate gebruik van steekproeven. Naast het Informatienet wordt gebruikgemaakt van CBS steekproeven zoals het stalsystemen en uitrijsystemen. Bewerkte Informatienet-data is een van de inputs voor de mest en ammoniakmodellen. Op basis van het Informatienet worden acceptatiegraden en kunstmestgiften per gewasgroep en regio vastgesteld. Hierbij wordt uitgegaan van de 31 mestgebieden. In een aantal gebieden is het Informatienet niet goed vertegenwoordigd. Als minimumaantal waarnemingen wordt uitgegaan van 20. Indien het aantal waarnemingen lager is worden regio's samengevoegd. Deze samenvoeging kan anders zijn voor elk van de gewassen. Bij het berekenen van de acceptatiegraden bestaan 2 vormen van onzekerheid, ten eerste de onzekerheid als gevolg van de Informatienet steekproeffout en ten tweede als gevolg van de onzekerheid omtrent het uitrijden. De tweede wordt inzichtelijk gemaakt door gevoeligheidsanalyses uit te voeren, aan de eerste wordt tot nu toe weinig aandacht aan besteed. Het is ook niet overwogen deze onzekerheden uit te rekenen. De schattingen vormen de invoer voor het mest en ammoniakmodel. De Informatienet uitvoer is te laag ten aan zien van de acceptatiegraden. Indien het Mest- en Ammoniakmodel niet tot een oplossing komt worden de acceptatiegraden iteratief aangepast tot er wel een oplossing mogelijk is.

De deelpopulaties die in de hiervoor beschreven aanpak zijn te onderscheiden zijn de regio's en de gewassen.

Er wordt zelden naar de betrouwbaarheden van de uitkomsten gevraagd. Na aanleiding van de RIVM-toestanden wordt er nu wel eens naar gevraagd maar dan ook alleen nog door het RIVM zelf. LNV zal zeker niet vragen naar marges. De politiek kan niet leven met onzekerheden. Het wordt wel als nuttig ervaren in de toekomst meer aandacht te besteden aan betrouwbaarheden.

Als er richtlijnen en leidraden zouden bestaan voor het gebruik van technieken voor het schatten van deelgebieden zouden die zeker worden toegepast. Voor sommige regio's wordt het gebruik van dergelijke technieken als moeilijk ervaren omdat de uitkomsten bij verre na niet stroken met de eigen expert kennis. Een combinatie van deze kennis, overige beschikbare informatie en schattingstechnieken kunnen wellicht wel leiden tot zinvolle toepassingen. In de huidige opzet wordt te veel uitsluitend gekeken naar het Informatienet.

Er is een duidelijke trend te constateren naar vragen omtrent kleine gebieden. Gebiedsgericht beleid omvat echter niet alleen milieu aspecten maar ook tal van andere. Deze vragen komen meestal voor op provinciaal niveau. Deze worden als moeilijke klanten ervaren. Sommige hebben weinig geld over voor het doen van onderzoek of doen het onderzoek liever zelf.

In begin jaren tachtig is een onderzoek uitgevoerd wat specifiek gericht was op het schatten van kenmerken van kleine deelgebieden. Het doel van het onderzoek was het schatten van kunstmestgiften op gemeentelijk niveau.

In dit onderzoek zijn relaties geschat tussen de kunstmestgiften per hectare en de bedrijfskenmerken op basis van de gegevens uit de steekproef. De kunstmestgiften zijn onderverdeeld in stikstof, fosfaat en kali en ze zijn uitgesplitst naar het gebruik op grasland, bouwland, een jarige opengrondstuinbouwgewassen, meerjarige -opengrondstuinbouwgewassen en gewassen onder glas. Met de relaties en de bedrijfskenmerken die voor elk bedrijf in de Landbouwtelling te vinden zijn kan dan voor elk bedrijf een schatting worden gemaakt van de kunstmestgiften per hectare.

Het CBS was geen voorstander van de gehanteerde aanpak omdat zij alleen van concrete waarnemingen en directe schatters wilden uitgaan. Intern was men zeer tevreden over de gehanteerde aanpak en het resultaat. Het zou een grote toegevoegde waarde hebben indien een dergelijke analyse jaarlijks zou kunnen worden uitgevoerd. Het kost echter flink wat tijd en geld om deze analyse uit te voeren. In het stofstromen model is een soortgelijke aanpak wel gedeeltelijk geïmplementeerd.

Ook voor het schatten van de stikstof giften in Zuid-Holland is een relatie geschat. Op basis van onder andere de grondsoort, de veebezetting en de dierlijke mestgift wordt een schatting gemaakt voor elk bedrijf van de N-gift op grasland.

Bij het opstellen van de bodembalans voor Zuid-Holland wordt een andere aanpak gehanteerd. Voor bepaalde gewassen is het aantal directe waarnemingen in deze provincie te gering. Voor dergelijke gewassen wordt de westelijke regio gehanteerd voor het maken van een schatting voor Zuid-Holland.

Gabe Venema

Binnen de sectie PPRF wordt veel van steekproeven gebruikgemaakt. Gebruikte steekproeven zijn onder andere het Informatienet, de Landbouwtelling en RICA. Het onderzoek heeft met name betrekking op de rentabiliteit en financiering en de inkomens- en financieringsontwikkeling. Bij het doen van onderzoek wordt meestal een insteek gekozen naar bedrijfstype, grote klasse, regio en leeftijd. Bij het analyseren van de Informatienet gegevens wordt altijd gebruikgemaakt van fip of bul weging. Dit levert wel eens problemen op bij het analyseren van kleine groepen waarbij de som van de wegingen kan afwijken van het aantal bedrijven in die groep volgens de Landbouwtelling. In de voorkomende gevallen dat de standaardfouten worden berekend worden die uit BDL gebruikt. Voor de typering van bedrijven wordt uitgegaan van de negtypering uit de Landbouwtelling.

Ondanks het feit dat men meestal geen betrouwbaarheidsintervallen berekent, maakt men toch regelmatig vergelijkingen tussen groepen. Officieel kun je hiermee de fout in gaan, maar de indruk bestaat dat dit bij grote groepen wel mee zal vallen.

In de sectie PPRF wordt weinig gebruikgemaakt standaardfouten en van statistische toetsen. Een discussie binnen de sectie leverde de volgende redenen: de software is niet goed, projectleiders vragen er niet om en er is een gebrek aan ondersteuning. Methoden zouden binnen het LEI op een centrale plaats beschikbaar moeten zijn, daarbij geldt de voorwaarde dat ze makkelijk toegankelijk moeten zijn. Er bestaat wel degelijk belangstel-

ling voor een intensiever gebruik van methoden. De behoefte aan een cursus is daarbij groot.

Met name voor kleine sectoren is de behoefte aan betere methoden groot. Het ministerie stelt steeds meer vragen over kleine takken. De indruk bestaat dat bij een goede marketing een aanzienlijke vraag bestaat naar onderzoek naar deelpopulaties. De indeling van deelpopulaties is in toenemende mate afhankelijk van de opdrachtgever.

Voor regionaal onderzoek wordt men vaak geconfronteerd met te weinig waarnemingen. Indien het aantal waarnemingen lager is dan 20 dan moet er iets gebeuren om iets te kunnen zeggen. In het verleden werd hiertoe gebruikgemaakt van de methode Tjomme de Haan. Opdrachtgevers willen graag iets weten over het financiële plaatje en deze methode biedt daartoe de mogelijkheid. Na een stroeve samenwerking met AM wordt een eigen variant van de methode gebruikt.

Hennie van der Veen

Informatienet en de daaraan gekoppelde bestanden vormen de steekproeven die worden gebruikt. Meestal worden schattingen voor het hele land gemaakt. Bij het doen van uitspraken over deelpopulaties wordt soms de indeling in BUT-types gehanteerd. Gemiddeldes worden uitgerekend als een gewogen gemiddelde, waarbij de weging in het Informatienet wordt gebruikt. In de meeste gevallen worden geen varianties berekend. Wanneer dit in specifieke gevallen wel wordt uitgerekend wordt niet de bdl software gebruikt maar wordt dit berekend in een statistisch pakket (spss). Het niet uitrekenen van varianties komt niet zo zeer uit gemak, het zit meer besloten in de bedrijfscultuur. Het wordt door bijna niemand gedaan, en op den duur denk je er zelf ook niet meer aan. Verder geldt dat de opdrachtgever geen moeilijke dingen wil, indien toch een indicatie moet worden gegeven van de spreiding is het makkelijker zoiets als min/max weer te geven.

Andere deelpopulaties die af en toe worden gebruikt zijn gebaseerd op een indeling naar inkomen. Op basis van de Informatienet gewichten kan een uitspraak worden gedaan voor een inkomensgroep.

In FES worden ook met name berekeningen uitgevoerd voor de totale populatie. De vraag naar opdrachten op regionaal gebied is nihil. FES wordt met name gebruikt bij belasting- en beleidsmaatregelen. Een evaluatie op landelijk niveau ligt hierbij voor de hand. Er is nog niet zozeer over nagedacht in hoeverre FES ook voor regionale vraagstukken kan worden ingezet. Bij kleine regio's ontstaat wel het probleem dat het aantal waarnemingen erg klein wordt. FES biedt geen directe mogelijkheden om varianties uit te rekenen.

Bij het simuleren van een langere periode ontstaat vaak het probleem van een gering aantal waarnemingen. Door het uitvallen van bedrijven blijven er soms weinig over. Verder zou men graag inzicht willen hebben hoe met de wegingen om te gaan. In FES zijn deze gewichten constant. Bij een evaluatie van meerdere jaren zouden de gewichten kunnen of moeten veranderen door bedrijfsuitval en veranderingen in de omvang van de bedrijven.

Hierbij wordt een relatie gelegd tussen het Informatienet en de Landbouwtelling. Er wordt een functie geschat waarbij een variabele uit het Informatienet (bijvoorbeeld stikstofgift per hectare) wordt geschat als functie van een of meer kenmerken uit de Landbouwtelling. De functie wordt geschat op basis van de gehele populatie. De keuze van de kenmerken die in de functie worden opgenomen is een belangrijke stap. Op basis van de onderzoeksvraag en verstand van het onderwerp worden relevante variabelen geselecteerd, vervolgens wordt gekeken in hoeverre deze variabelen echt een verklarende waarde hebben, tevens wordt op de samenhang oftewel de correlatie tussen de verschillende variabelen gelet.

Een voorbeeld van een dergelijk onderzoek is de Lopiker Waard. In dit gebied zijn ongeveer 450 bedrijven actief. Voor een soort toets op de bruikbaarheid van de geschatte functie (geschat voor het hele land) wordt gekeken in hoeverre de functie redelijk voorspelt voor de Informatienet-bedrijven in die regio. Indien te onderzoek regio klein is en weinig Informatienet-bedrijven bevat, wordt een toetsing op een iets ruimer gebied uitgevoerd. Hierbij wordt met name gekeken naar het gemiddelde. Een vraag die hierbij nog niet beantwoord is hoe je kunt en of je mag corrigeren voor een eventueel verschil. Een door een opdrachtgever gesuggereerde correctie is het domweg toevoegen van een constante. In principe worden bij het schatten van een functie alle bedrijfstypes meegenomen.

In het stofstromenmodel is een vergelijking geschat op basis van een aantal jaargangen. De resulterende vergelijking wordt toegepast maar niet jaarlijks geëvalueerd en of bijgesteld. In het stofstromenmodel worden de uitkomsten van de functie als vaststaand beschouwd, er wordt geen rekening gehouden met het betrouwbaarheidsinterval behorende bij de regressieschatting voor een bepaald bedrijf.

Men is tevreden over de uitkomsten van deze werkwijze. Wel heeft het vrij veel tijd gekost om het te implementeren. Met name het samenstellen van de dataset is bewerkelijk. Een validatie van de uitkomsten op nationaal niveau gaf goede resultaten en dus geen aanleiding om de functie aan te passen. Bij kleine gebieden speelt wel eens het probleem dat er veel specifieke kenmerken een rol spelen. Deze kenmerken zijn niet in de functie opgenomen.

Een probleem dat wordt ervaren is dat je minder variatie overhoudt. Bij een lage R^2 geeft de functie slechts een deel van de variantie weer. Een voorstel om een random factor op te nemen om deze variatie weer op het gewenste niveau te brengen is als niet zinvol afgewezen.

In Wageningen heeft men kritiek op de hier beschreven werkwijze. Dit kan worden verklaard uit een gebrek aan inzicht in de problematiek en dus de reden waarom je het doet en uit het feit dat men in Wageningen geen langdurige relatie tussen bemesting en andere variabelen ziet.

Een regressiefunctie wordt geschat indien de variabelen het gewenste meetniveau hebben. Bij andere variabelen (bijvoorbeeld drie typen beweidingsystemen) wordt op basis van bijvoorbeeld het staltype en de intensiteit gekeken hoe de frequentieverdelingen zijn. Vervolgens wordt op basis van deze kennis een beweidingsysteem toegekend aan een bedrijf.