Controle en correctie

19

Jeffrey Hoogland, Mark van der Loo, Jeroen Pannekoek en Sander Scholtus

Statistische Methoden (10011)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2010

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader voorlopig cijfer
x	=	geheim
–	=	nihil
–	=	(indien voorkomend tussen twee getallen) tot en met
0 (0,0)	=	het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	=	een cijfer kan op logische gronden niet voorkomen
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

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

Colofon

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

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

Omslag
TelDesign, Rotterdam

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

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

Internet
www.cbs.nl

ISSN: 1876-0333

6016510011 X-37

Inhoudsopgave

1. Inleiding op het thema

1.1 Algemene beschrijving en leeswijzer

1.1.1 Beschrijving van controle en correctie

In de databestanden die op het CBS gebruikt worden komen vrijwel altijd fouten voor. Dit geldt zowel voor de data verkregen via eigen waarneming als voor data afkomstig van registratiehouders. Voor zover deze fouten leiden tot vertekening in schattingen van publicatiecijfers is het voor het CBS van belang de fouten op te sporen en te corrigeren.

Fouten kunnen ontstaan tijdens de waarneming; er is dan een verschil tussen de gerapporteerde waarde en de werkelijke waarde. Dit kan komen omdat de respondent de werkelijke waarde niet (precies) weet of moeilijk kan achterhalen en dus een schatting maakt. Een andere mogelijke oorzaak is het verschil in definities tussen de boekhouding bij bedrijven en het CBS, omdat bijvoorbeeld het boekjaar afwijkt van het kalenderjaar. Bovendien kan het zijn dat bedrijven bepaalde informatie, die het CBS wil ontvangen, gewoonweg niet meten. In dit geval zal de respondent wederom de waarde schatten of helemaal niet invullen. Tot slot kunnen vragen ook verkeerd worden gelezen of begrepen door de respondent. Bijvoorbeeld als de respondent in euro's rapporteert terwijl gevraagd wordt om in duizenden euro's te rapporteren of als de respondent alleen voor zichzelf antwoordt en niet, zoals gevraagd, voor het hele huishouden.

Fouten kunnen ook ontstaan tijdens het proces van het verwerken van de data nadat deze zijn verzameld. Op het CBS doorlopen de verzamelde gegevens verschillende processen, zoals invoeren, coderen, controleren, corrigeren, wegen en tabelleren. Al deze processen kunnen fouten introduceren in de gegevens. Een voorbeeld hiervan is dat het handmatig invoeren van gegevens kan leiden tot misinterpretaties, bijvoorbeeld dat een 1 voor een 7 wordt aangezien of vice versa. Verder kunnen er fouten in de verwerkingssoftware zitten of goede waarden onterecht voor fout worden aangezien tijdens het controle- en correctieproces:

Controle- en correctiemethoden hebben verschillende doelstellingen

1. het identificeren van mogelijke foutenbronnen zodat in de toekomst het statistisch proces verbeterd kan worden;
2. het leveren van informatie over de kwaliteit van verzamelde en gepubliceerde gegevens;
3. het opsporen en corrigeren van invloedrijke fouten in de verzamelde gegevens;
4. het leveren van volledige en consistente gegevens.

Momenteel worden op het CBS controle- en correctiemethoden voornamelijk toegepast met als doelstellingen volledige en consistente gegevens te leveren en fouten die een grote invloed hebben op het publicatietotaal te corrigeren. Daarnaast worden naar aanleiding van de gevonden fouten ook verbeteringen doorgevoerd in de lay-out van de vragenlijst of de toelichtingen bij de vragen. Analyse van de gevonden fouten kan ook gebruikt worden om verschillen in elektronische en schriftelijke vragen vast te stellen en om inzicht te krijgen in de kwaliteit van administratieve data.

Er zijn verschillende controle- en correctiemethoden en processen ontwikkeld voor verschillende soorten van fouten. Belangrijk hierbij is het onderscheid tussen invloedrijke fouten en niet-invloedrijke fouten en het onderscheid tussen systematische en toevallige fouten.

Voornamelijk bij bedrijfsstatistieken kunnen *invloedrijke* en *niet-invloedrijke fouten* worden onderscheiden. Onder invloedrijke fouten verstaan we de fouten die een grote invloed hebben op het uiteindelijke publicatietotaal. Dit kan zijn omdat de fout is gemaakt bij een groot bedrijf dat al veel invloed heeft op het totaal of een minder groot bedrijf dat een groot gewicht heeft in de schatting voor het totaal of omdat er een grote fout is gemaakt die het totaal sterk zal beïnvloeden, bijvoorbeeld een duizendfout. Het is duidelijk dat fouten die een grote invloed hebben op een publicatietotaal tot grote vertekeningen kunnen leiden en zeer risicovol zijn voor het CBS. Daarom is het cruciaal deze fouten zo goed mogelijk op te sporen en te corrigeren. Het controle- en correctieproces zal zich dan ook vooral op deze fouten moeten richten.

Een andere onderverdeling die vaak wordt gemaakt is die in *systematische* en *toevallige fouten*. Een systematische fout is een fout die door meerdere respondenten wordt gemaakt, zoals de genoemde duizendfouten, of het feit dat verschillende respondenten bruto- in plaats van netto-inkomensgegevens opgeven of het feit dat een groep respondenten een minteken voor een bedrag zet, terwijl het minteken al op de vragenlijst staat. Omdat deze fouten door meerdere respondenten op dezelfde manier worden gemaakt kunnen zij een systematische vertekening opleveren. Als men weet welke systematische fouten worden gemaakt, dan zijn deze vaak eenvoudig op te sporen en te corrigeren. Toevallige fouten zijn fouten die per ongeluk ontstaan. De meest voorkomende oorzaak hiervan is onoplettendheid van de respondent, de interviewer of de gegevens-invoerder. Bijvoorbeeld doordat er bij het invullen twee cijfers worden verwisseld. Wanneer deze fouten niet-systematisch worden gemaakt zal de kans op het ontstaan van een systematische vertekening door dit soort fouten kleiner zijn.

Systematische fouten kunnen zowel invloedrijk (duizendfout) als niet-invloedrijk (tekenfout in een kleine waarde) zijn. Hetzelfde geldt voor toevallige fouten. Als er bijvoorbeeld per ongeluk teveel cijfers zijn ingevuld bij een groot bedrijf kan dit invloedrijk zijn, maar bij een klein bedrijf zal dit misschien niet-invloedrijk zijn.

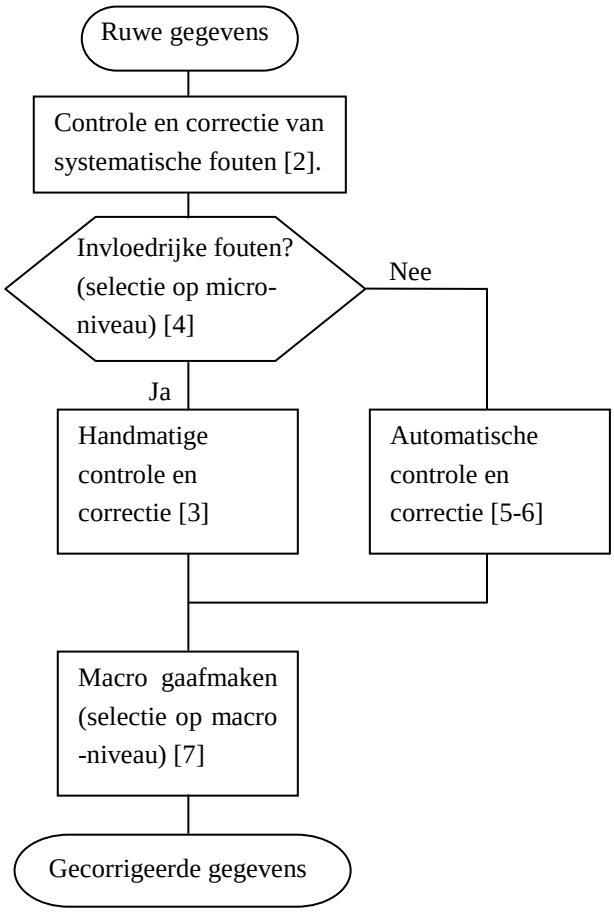
In dit thema bespreken we controle- en correctiemethoden die zijn ontwikkeld om systematische, toevallige en invloedrijke fouten op te sporen en te corrigeren. In

paragraaf 1.1.2 beschrijven we de verschillende soorten methoden aan de hand van een prototype proces waarin op globale wijze de verschillende processtappen van een mogelijk controle- en correctieproces geschetst worden.

1.1.2 Problemen en oplossingen

Nadat de gegevens zijn verzameld en ingevoerd wordt het controle- en correctieproces gestart. De specifieke manier waarop dit proces wordt toegepast zal verschillen per statistiek, maar er is wel een algemene strategie, die in veel processen in grote lijnen gevolgd wordt. Deze algemene strategie is weergegeven in figuur 1 en geeft een overzicht van het controle- en correctieproces.

Figuur 1. Overzicht van het controle- en correctieproces[1)]



[1)] *De cijfers tussen vierkante haken verwijzen naar de betreffende hoofdstukken van dit rapport.*

In de eerste fase van het controle- en correctieproces worden herkenbare systematische fouten opgespoord en gecorrigeerd. Zoals al eerder vermeld kunnen deze systematische fouten tot grote vertekening leiden. Bovendien zijn deze fouten vaak eenvoudig en zeer betrouwbaar automatisch op te sporen en te corrigeren. Het is zeer efficiënt om deze fouten in een vroeg stadium te corrigeren. Het detecteren

en corrigeren van systematische fouten wordt in hoofdstuk 2 van dit rapport behandeld.

Nadat de herkenbare systematische fouten automatisch zijn gecorrigeerd kan worden besloten om te beginnen met handmatig controle en correctie. Deze processtap wordt uitgevoerd door gaafmakers of analisten die hierbij meestal ondersteund worden door software waarmee ondermeer controleregels automatisch toegepast kunnen worden en waarden interactief veranderd kunnen worden. We spreken daarom behalve van *handmatig gaafmaken* ook van *interactief gaafmaken*. Deze vorm van gaafmaken wordt beschreven in hoofdstuk 3 van dit rapport.

Handmatig controleren en corrigeren is duur en tijdrovend. Het is daarom beter om alleen de records met invloedrijke fouten handmatig te bekijken zodat de beperkte tijd van specialisten ingezet wordt waar deze het meest effectief is. Dit betekent dat records waarvan verwacht wordt dat ze invloedrijke fouten bevatten geselecteerd worden voor interactief gaafmaken. De overige records met minder belangrijke fouten kunnen automatisch gaafgemaakt worden. Als er veel vertrouwen is in het automatisch gaafmaken dan kunnen records met invloedrijke fouten ook automatisch worden gaafgemaakt. Het beperken van interactief gaafmaken tot die records waarin waarschijnlijk invloedrijke fouten voorkomen die niet betrouwbaar automatisch op te lossen zijn wordt *selectief gaafmaken* genoemd. Dit selectieproces wordt beschreven in hoofdstuk 4 van dit rapport.

Bij selectief gaafmaken wordt er gebruik gemaakt van verwachte waarden voor de variabelen in een record om te bepalen of die waarden afwijkend zijn. Sterk afwijkende waarden kunnen veroorzaakt worden door een invloedrijke fout. Bij het bepalen van deze verwachte waarden wordt informatie uit andere bronnen gebruikt dan het actuele bestand. Vaak worden hiervoor gegevens uit een vorige periode van dezelfde statistiek gebruikt. Dit maakt het mogelijk om met het selecteren voor handmatig gaafmaken te beginnen tijdens de dataverzamelingsperiode, zodra de eerste records binnenkomen. Als alle, of een groot deel van de, data binnen zijn kunnen verdachte waarden ook opgespoord worden door te kijken naar voorlopige schattingen van totalen en naar waarnemingen die hier veel invloed op hebben. Deze vorm van selectie heet macro-gaafmaken (zie hieronder).

Het automatisch corrigeren van toevallige fouten en andere fouten waarvan de oorzaak niet te achterhalen is, gebeurt in twee stappen. In de eerste plaats wordt zo goed mogelijk bepaald welke scores op variabelen in een record fout zijn. Dit is triviaal als een waarde niet in het toegelaten waardebereik ligt, bijvoorbeeld een negatief inkomen of een ten onrechte ontbrekende waarde. De betrokken waarde is dan zeker fout. In veel gevallen treden er echter inconsistenties (schendingen van editregels) op waarbij het niet duidelijk is welke waarde(n) daarvoor verantwoordelijk zijn. Als bijvoorbeeld aan een optelbaarheidsregel (zoals: totale personeelslasten is gelijk aan de som van de salarissen, verzekeringspremies, opleidingskosten en overige personeelslasten) niet is voldaan, is niet duidelijk welke waarde(n) in die optelling voor de schending van de regel verantwoordelijk zijn. Bij het automatisch detecteren van fouten worden de foute waarden aangewezen (het

lokaliseren van fouten) volgens het principe van Fellegi en Holt, dat luidt: wijs zo min mogelijk waarden aan als foutief, maar wel zodanig dat het veranderen van deze waarden kan leiden tot een volledig consistent record dat aan alle harde editregels voldoet. Het automatisch lokaliseren van fouten op basis van het principe van Fellegi en Holt wordt behandeld in hoofdstuk 5 van dit rapport. Als de fouten opgespoord zijn worden zij vervangen door betere waarden middels imputatie. Automatische imputatie vindt plaats met behulp van modellen die de foute of ontbrekende waarden kunnen voorspellen. Hierop wordt uitgebreid ingegaan in het thema *Imputatie*.

Een alternatieve methode die automatisch fouten lokaliseert én nieuwe waarden imputeert, de zogenaamde Nearest-neighbour Imputatie Methodologie, wordt behandeld in hoofdstuk 6 van dit rapport.

Ten slotte worden in de laatste fase voorlopige publicatiecijfers berekend en geanalyseerd aan de hand van historische gegevens of externe bronnen. Deze analyse wordt ook wel fiatteren of macro gaafmaken genoemd en wordt besproken in hoofdstuk 7 van dit rapport. Als de geaggregeerde cijfers implausibel zijn, wordt er naar de individuele records gekeken door bijvoorbeeld uitbijters of invloedrijke records nader te analyseren en zonodig te corrigeren. De fouten die in deze fase worden opgespoord kunnen fouten zijn die niet in eerdere fasen van het controle- en correctieproces worden gevonden of fouten die juist zijn geïntroduceerd door het proces. Bij macro-gaafmaken begint de detectie van fouten dus op macro-niveau maar vindt de correctie altijd plaats in de individuele records, dus op micro-niveau. Als de voorlopige cijfers plausibel zijn, wordt het controle- en correctieproces afgesloten.

Het proces in figuur 1 moet gezien worden als een prototype. In de praktijk zullen voor de verschillende statistieken niet alle stappen worden doorlopen. Zo zijn er bijvoorbeeld bij statistieken over persoonsgegevens weinig controleregels aanwezig. Daardoor zal de nadruk van het controle- en correctieproces in dat geval meer liggen op de correctie voor item-nonrespons door middel van imputatie. Bij statistieken die op registers gebaseerd zijn komt (een groot deel van) de data vaak tegelijk beschikbaar. In dat geval kan meteen worden begonnen met macro gaafmaken. Ook worden er bij de selectie van records voor het handmatig gaafmaken vaak nog andere criteria gebruikt dan alleen of een record invloedrijke fouten bevat. Zo worden heel belangrijke bedrijven vaak als cruciaal aangemerkt en altijd handmatig geïnspecteerd. Dit kunnen bijvoorbeeld bedrijven zijn die in hun eentje al een groot deel van de omzet in hun branche bepalen. Ook kan er voor gekozen worden zeer goed imputeerbare variabelen automatisch gaaf te maken ook als zij mogelijk invloedrijke fouten bevatten.

1.2 Afbakening en relatie met andere thema's

In dit thema wordt ingegaan op het controleren en eventueel corrigeren van mogelijke fouten in microdata. De bedoeling van dit onderdeel van het statistisch proces is om "ruwe" microdata met fouten en inconsistenties te transformeren naar

gecorrigeerde “gave” microdata die geschikt zijn voor het schatten van publicatiecijfers en verdere analyses. Het schatten zelf, met daaraan gerelateerde problemen zoals bepalen van ophooggewichten, het corrigeren voor non-respons door weging en het behandelen van correcte uitbijters, wordt in andere thema’s van de Methodenreeks beschreven.

In het thema *Controle en Correctie* bespreken we de verschillende methoden die zijn ontwikkeld om fouten op te sporen. In dit thema worden eveneens correctietechnieken behandeld die worden gebruikt voor de correctie van fouten waarbij de ingevulde, onjuiste waarde informatie bevat over de correcte waarde, zoals bij te duiden systematische fouten. Daarnaast wordt ook ingegaan op handmatige correctie door experts (interactief gaafmaken). Correctie voor item-nonrespons en onjuiste waarden, waarbij de ingevulde waarde geen informatie bevat over de correcte waarde en daarom op missing is gezet, vindt vaak plaats door middel van imputatie. Dit onderwerp wordt besproken in het thema *Imputatie*.

1.3 Plaats in het statistisch proces

Zoals in figuur 1 is geschetst, begint het controle- en correctieproces met ruwe gegevens. Dit zijn de gegevens zoals die ontvangen zijn van de respondenten en vervolgens zijn vertoetst, bij schriftelijke enquêtes, en opgeslagen in een standaardformaat. Bij elektronische vragenlijsten en bij CATI-waarneming kunnen tijdens de waarnemingsfase al bepaalde controles uitgevoerd worden wat mogelijk leidt tot correcties door de respondent, maar ook deze gegevens worden voor het controle- correctieproces, dat na de waarneming begint, als ruwe gegevens opgevat.

Zoals in paragraaf 1.1.2 is beschreven kunnen de meeste controle-correctieprocedures uitgevoerd worden per berichtgever zonder dat kennis van de antwoorden van de overige berichtgevers nodig is (dit geldt alleen niet voor de selectie van invloedrijke fouten op macro-niveau). Het controle-correctieproces kan dan ook deels uitgevoerd worden tijdens de dataverzamelingsfase. Dit is vooral van belang voor statistische processen waarbij een substantiële hoeveelheid records interactief gaaf wordt gemaakt. Voor deze processen is het uit het oogpunt van de tijdigheid van de op te leveren cijfers van belang om zo vroeg mogelijk met gaafmaken te beginnen. Bij statistische processen waar weinig interactief wordt gaafgemaakt of de dataverzamelingsperiode van korte duur is kan de controle-correctieprocedure ook na de dataverzamelingsperiode beginnen.

Zowel de “input” als de “output” van het controle-correctieproces bestaat uit een bestand met records per berichtgever. Het controle-correctieproces heeft de ruwe micro-data met overduidelijke fouten, inconsistenties en ontbrekende waarden getransformeerd naar gave micro-data waarin deze problemen zoveel mogelijk zijn opgelost. Het gave bestand wordt in het verdere statistische proces gebruikt om te aggregeren, schattingen te maken van totalen en ontwikkelingen en voor verdere analyses. Het controle-correctieproces brengt alleen veranderingen aan in de micro-data. Correcties op geaggregeerde cijfers zoals die bij nationale rekeningen

plaatsvinden behoren niet tot de controle-correctie fase maar tot een volgende stap in het statistisch proces.

1.4 Definities

Begrip	Omschrijving
automatisch gaafmaken	een verzamelnaam voor gaafmaakmethoden waarbij een computerprogramma de data zowel controleert als corrigeert
controle en correctie	het detecteren en verbeteren van ontbrekende en onjuiste waarden in een databestand
controleregel	een restrictie op de waarden in een databestand; data die niet voldoen aan een controleregel bevatten ofwel met zekerheid (zie “harde controleregel”), ofwel met grote waarschijnlijkheid een fout (zie “zachte controleregel”)
deductieve correctie	een verzamelnaam voor gaafmaakmethoden waarbij benodigde correcties eenduidig zijn af te leiden uit de ongecorrigeerde data
gaafmaken	zie “controle en correctie”
handmatig gaafmaken	zie “interactief gaafmaken”
harde controleregel	een controleregel die aangeeft dat er met zekerheid een fout in de data zit
interactief gaafmaken	een gaafmaakmethode waarbij een computerprogramma de data controleert en een menselijke gaafmaker de data corrigeert
invloedrijke fouten	fouten die een grote invloed hebben op het te publiceren cijfer
macrogaafmaken	een verzamelnaam voor gaafmaakmethoden waarbij controle van de data plaatsvindt op een geaggregeerd niveau
microgaafmaken	een verzamelnaam voor gaafmaakmethoden waarbij controle en correctie plaatsvindt op het niveau van de individuele records
scorefunctie	een indicator van de invloed die het interactief gaafmaken van een record naar verwachting zal hebben op het te publiceren cijfer; scorefuncties worden gebruikt om records te prioriteren voor interactief gaafmaken (zie “selectief gaafmaken”)
selectief gaafmaken	een verzamelnaam voor methoden om records te selecteren voor interactief gaafmaken waarin mogelijk invloedrijke fouten voorkomen; hierbij wordt vaak een scorefunctie gebruikt
zachte controleregel	een controleregel die aangeeft dat er met grote waarschijnlijkheid een fout in de data zit; data die niet voldoen aan een zachte controleregel zijn verdacht, maar niet noodzakelijk fout

2. Methoden voor deductieve correctie

2.1 Korte beschrijving

Data die zijn verzameld voor het maken van statistieken bevatten dikwijls overduidelijke systematische fouten, d.w.z. fouten die door meerdere respondenten op dezelfde, herkenbare manier gemaakt worden. Een dergelijke systematische fout kan vaak op eenvoudige wijze automatisch worden gedetecteerd, zeker in vergelijking met de complexe algoritmen die nodig zijn voor het automatisch lokaliseren van niet-systematische fouten (zie hoofdstuk 5 en 6). Bovendien is het, nadat een systematische fout is gedetecteerd, meteen duidelijk welke correctie nodig is om hem te herstellen. We weten immers, of denken althans met voldoende zekerheid te weten, hoe de fout is ontstaan.

Voor elk type systematische fout is een apart detectie- en correctievoorschrift nodig. De precieze vorm van de correctiemethode verschilt per type fout; er is geen standaardrecept beschikbaar. Dit hoofdstuk wijkt daarom qua opbouw iets af van de rest van het rapport. De meeste ruimte wordt gebruikt voor het behandelen van praktijkvoorbeelden (paragraaf 2.4) in plaats van algemene theorie (paragraaf 2.3).

De moeilijkheid bij het toepassen van deze methode zit vooral in het bepalen *welke* systematische fouten zullen voorkomen in de data, voordat deze daadwerkelijk verzameld zijn. Dit kan onderzocht worden op basis van data uit het verleden. Wanneer bepaalde controleregels vaak op dezelfde manier geschonden zijn, is er mogelijk sprake van een systematische fout. Soms brengt een dergelijk onderzoek systematische fouten aan het licht die ontstaan wegens een tekortkoming in het vragenlijstontwerp of een bug in het verwerkingsproces. In dat geval moet de vragenlijst c.q. de procedure worden aangepast. In verband met methodebreuken kan het wenselijk zijn om aanpassingen in de vragenlijst ‘op te sparen’ tot een gepland herontwerp van de statistiek, en tot die tijd de systematische fout op te lossen met een deductieve correctiemethode.

2.2 Toepasbaarheid

Deductieve correctie is toepasbaar op zowel kwantitatieve als kwalitatieve variabelen. Deductieve methoden zijn in de eerste plaats bedoeld om systematische fouten te corrigeren. Voor het corrigeren van niet-systematische (of toevallige) fouten zijn zulke methoden in het algemeen niet geschikt. Het wordt aanbevolen om zo veel mogelijk systematische fouten te behandelen aan het begin van het controle- en correctieproces, voordat enige andere correctiemethode wordt toegepast. Van deze fouten is immers bekend hoe ze zijn ontstaan en hoe ze ongedaan gemaakt kunnen worden. De rest van het controle- en correctieproces verloopt efficiënter nadat de systematische fouten deductief zijn opgelost.

Fouten waarvan de oorzaak met voldoende zekerheid bekend is kunnen deductief worden opgelost. Bij onjuiste aannames over de foutenbron kan de methode leiden tot vertekening in de schatters. In de praktijk kunnen correctieregels (zie paragraaf 2.3.1) uit efficiencyoverwegingen ook worden toegepast op niet-systematische fouten, als de geïntroduceerde vertekening verwaarloosbaar is. Bij dit laatste kan men bijvoorbeeld denken aan het deductief oplossen van afrondfouten (cf. Scholtus, 2008a).

Systematische fouten zijn vaak te herkennen door te kijken naar veel voorkomende schendingen van controleregels. Deductieve methoden zijn daarom vooral effectief bij data waarvoor veel controleregels zijn gedefinieerd.

2.3 Uitgebreide beschrijving

2.3.1 Correctieregels

De meest eenvoudige deductieve correctiemethoden zijn weer te geven in één regel:

als (*voorwaarde*) **dan** (*aanpassing*). (2.3.1)

Hierin geeft *voorwaarde* een combinatie van waarden in een record die niet is toegestaan. Vervolgens beschrijft *aanpassing* de correctie die wordt aangebracht om de inconsistentie op te lossen. Men spreekt in dit geval wel van *correctieregels*.

Een voorbeeld van een correctieregel is:

als (*geslacht* = man **en** (*zwanger* = ja **of** *zwanger* = “leeg”))
dan *zwanger* = nee. (2.3.2)

Deze regel corrigeert records die niet voldoen aan de controleregel

als *geslacht* = man **dan** *zwanger* = nee.

Merk op dat de **als**-**dan**-constructie hier op twee verschillende manieren wordt gebruikt. In een controleregel beschrijft zij een voorwaarde waaraan de data zouden moeten voldoen, in een correctieregel beschrijft zij een actie die leidt tot aanpassingen in de data.

Een ander voorbeeld van een correctieregel is:

als (*leeftijd* < 18 **en** (*rijbewijs* = ja **of** *rijbewijs* = “leeg”))
dan *rijbewijs* = nee. (2.3.3)

Deze regel kan gebruikt worden om records te corrigeren die niet voldoen aan de controleregel

als *leeftijd* < 18 **dan** *rijbewijs* = nee.

Ook de meeste algemene deductieve correctiemethoden zijn in principe uit te drukken als regels van de vorm (2.3.1). De **als**-voorwaarde bevat dan eventueel ook informatie uit andere records of zelfs van buiten het te corrigeren databestand. Het detectiecriterium kan vrij complex zijn; zie de voorbeelden in paragraaf 2.4.

2.3.2 Opstellen van deductieve correcties

Een deductieve correctiemethode is bedoeld om een inconsistentie op te lossen die op logische en/of inhoudelijke gronden slechts op één manier kan worden opgelost, onder een bepaalde aanname. Als de aanname juist is, levert de deductieve correctiemethode altijd de werkelijke waarden op. Correctieregel (2.3.2) werkt bijvoorbeeld onder de aanname dat de variabele *geslacht* nooit fouten bevat. Hetzelfde geldt voor regel (2.3.3) en de variabele *leeftijd*. Aan deze aannames zou bijvoorbeeld (bij benadering) voldaan kunnen zijn als *geslacht* en *leeftijd* afkomstig zijn uit een goed onderhouden populatieregister.

Deductieve correctiemethoden zijn aantrekkelijk vanwege hun eenvoud. Ze mogen echter alleen gebruikt worden wanneer bij een dergelijke eenvoudige aanpak geen belangrijke nuances verloren gaan. Wanneer de data niet voldoen aan de gemaakte aannames leidt deductief corrigeren tot vertekende schatters. Bijvoorbeeld: indien *geslacht* of *leeftijd* in sommige records een foutieve waarde hebben zullen we, na toepassing van correctieregels (2.3.2) en (2.3.3), het aantal zwangerschappen, respectievelijk het aantal houders van een rijbewijs in de populatie onderschatten.

Doorgaans kan een gegeven inconsistentie op veel verschillende manieren worden verklaard. Zelfs in de simpele voorbeelden uit paragraaf 2.3.1, met slechts twee variabelen, kunnen we de inconsistenties alleen deductief corrigeren door sommige verklaringen op voorhand uit te sluiten. Deductieve correctie is in het algemeen alleen toepasbaar wanneer een van de verklaringen voor de inconsistentie veel meer voor de hand ligt dan alle andere mogelijke verklaringen. Om dit te beoordelen is vaak inhoudelijke kennis van de data nodig.

Een idee dat veel wordt gebruikt bij het opstellen van deductieve correctiemethoden (soms onbewust) is het volgende: als voor een gegeven inconsistentie een correctie bestaat die zeer weinig verandert aan de huidige waarden, dan levert deze zeer waarschijnlijk de werkelijke waarden op. Hierbij kan ‘weinig veranderen’ zowel slaan op het aantal veranderingen als op de aard van de veranderingen. Dit is in feite een naïeve versie van het principe van Fellegi en Holt. (Zie hoofdstuk 5 voor het echte principe van Fellegi en Holt.)

Ter illustratie toont de eerste kolom van tabel 1 een record dat inconsistent is met betrekking tot de controleregel

$$omzet - kosten = winst. \tag{2.3.4}$$

De inconsistentie kan worden opgelost door een van de drie variabelen aan te passen. De overige kolommen van tabel 1 laten zien welke mogelijke aanpassingen dit oplevert (de aangepaste waarde is steeds vetgedrukt). Intuïtief is de oplossing waarbij *kosten* wordt aangepast de meest aantrekkelijke, omdat het veranderen van de waarde 283 in 238 minder ingrijpend is dan de andere voorgestelde correcties. Omgekeerd ligt het veel meer voor de hand dat de werkelijke waarde 238 ergens tijdens het verzamelen en verwerken van de data is veranderd in 283, dan dat 398 is veranderd in 353 of 70 in 115. We zouden daarom het volgende voorschrift voor deductieve correctie kunnen opstellen: als een record niet voldoet aan (2.3.4), maar

wél wanneer de cijfers in een van de opgegeven bedragen worden verwisseld, dan moet het record op deze manier worden gecorrigeerd. Bij het opstellen van deze correctiemethode hebben we tweemaal het ‘naïeve principe van Fellegi en Holt’ gebruikt: eerst door niet te kijken naar oplossingen waarin meer dan één variabele wordt aangepast, en vervolgens door van de overgebleven oplossingen de minst ingrijpende te kiezen.

Bij de praktijkvoorbeelden in paragraaf 2.4 komt dit principe nog een aantal keer langs.

Tabel 1. Voorbeeld van een deductief te corrigeren record

	record	correctie 1	correctie 2	correctie 3
omzet	353	**398**	353	353
kosten	283	283	**238**	283
winst	115	115	115	**70**

2.3.3 Opsporen van nog onbekende systematische fouten

Nieuwe systematische fouten kunnen worden gevonden door schendingen van controleregels te analyseren. Als een controleregel vaak is geschonden kan dit een aanwijzing zijn voor de aanwezigheid van een systematische fout in de bijbehorende variabelen. Een nadere analyse van de records die de controleregel schenden, waarbij ook de vragenlijst bekeken wordt, kan de oorzaak van de fout aan het licht brengen. Als de fout eenmaal geïdentificeerd is, is het doorgaans vrij eenvoudig om een deductieve methode op te stellen waarmee de fout automatisch wordt gedetecteerd en gecorrigeerd.

Het opsporen van nieuwe systematische fouten kan pas plaatsvinden nadat voldoende data verzameld zijn. Voor de huidige waarnemingsperiode komen de resultaten daarom meestal te laat. Als de analyse nieuwe deductieve correctiemethoden oplevert, kunnen deze ingebouwd worden in het correctieproces voor de data van de volgende waarnemingsperiode.

Voor systematische fouten geldt: voorkomen is beter dan genezen. Soms is het mogelijk het ontwerp van de vragenlijst te verbeteren zodat veel minder respondenten een bepaald type fout maken. Wanneer veel respondenten op dezelfde manier in de fout gaan kan dit een aanwijzing zijn dat de vraagstelling niet duidelijk genoeg is. Ook is het soms mogelijk het verwerkingsproces aan te passen zodat een bepaalde verwerkingsfout zich niet meer voordoet. Deze aanpak is in principe te verkiezen boven het achteraf aanbrengen van deductieve correcties. Aangezien er praktische bezwaren kleven aan het voortdurend aanpassen van de vragenlijst, kan men in eerste instantie volstaan met het inbouwen van een deductieve correctiemethode, en de opgedane kennis meenemen bij een later herontwerp.

Ter illustratie sporen we een nieuwe systematische fout op in de data van de Productiestatistiek (PS) Groothandel 2001. Een van de (vele) controleregels luidt:

LOONSOM110000 + LOONSOM121100 + LOONSOM121200 + LOONSOM122000 = LOONSOM100000.

Hierin stelt LOONSOM100000 de totale arbeidskosten voor. De andere vier variabelen zijn de deelposten van dit totaal.

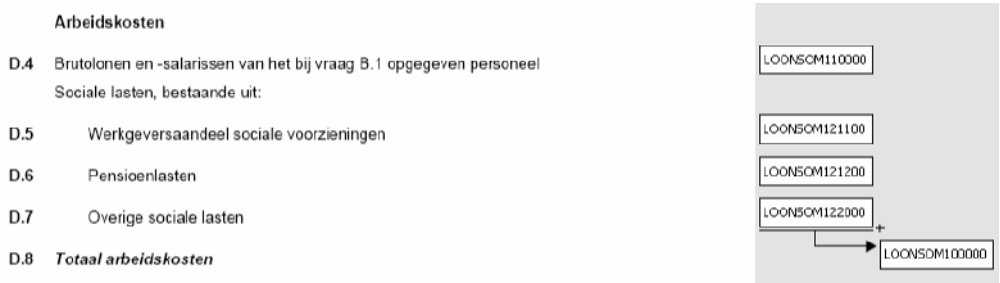
Tabel 2 toont enkele records die de controleregel schenden.

Tabel 2. Voorbeelden van inconsistente records in de PS Groothandel 2001

	record 1	record 2	record 3	record 4
LOONSOM110000	1 100	364	1 135	901
LOONSOM121100	88	46	196	134
LOONSOM121200	40	34	68	0
LOONSOM122000	42	0	42	0
LOONSOM100000	170	80	306	134

Opvallend is dat in deze records de posten LOONSOM121100, LOONSOM121200 en LOONSOM122000 optellen tot het totaal LOONSOM100000. Dat wil zeggen: het lijkt erop dat deze berichtgevers de eerste deelpost LOONSOM110000 hebben genegeerd bij het berekenen van LOONSOM100000. Een nadere blik op de vragenlijst (figuur 2) maakt duidelijk waarom dit gebeurd is: er is een gat tussen het antwoordvakje van LOONSOM110000 en de andere vakjes. Hierdoor is uit de vraagstelling niet ondubbelzinnig af te leiden of LOONSOM110000 bij de som hoort of op zichzelf staat. De meeste respondenten begrijpen uit de context wel wat de bedoeling is, maar in enkele tientallen records zien we de fout uit tabel 2 terug.

Figuur 2. Onderdeel van de vragenlijst PS Groothandel (t/m 2005)



We kunnen een correctiemethode opstellen die deze fout deductief oplost. Een meer structurele oplossing bestaat uit het wegnemen van de oorzaak van de fout door de vragenlijst aan te passen. Dit is overigens inmiddels gebeurd: de vragenlijst uit figuur 2 is voor de PS 2006 vervangen. Op de nieuwe vragenlijst staan alle antwoordvakjes even ver uit elkaar.

2.4 Voorbeelden

In deze paragraaf komen enkele praktijkvoorbeelden aan bod van reeds toegepaste of althans ontwikkelde methoden voor deductieve correctie. We behandelen eerst

voorbeelden uit de statistiek Bouwobjecten In Voorbereiding (paragraaf 2.4.1) en de Korte-termijn Statistieken (paragraaf 2.4.2). De overige voorbeelden zijn afkomstig uit de PS. We bespreken eerst de deductieve correctiemethoden in het huidige verwerkingsproces (paragraaf 2.4.3) en vervolgens drie recent ontwikkelde methoden (paragrafen 2.4.4 t/m 2.4.6).

2.4.1 Correctieregels voor de statistiek Bouwobjecten In Voorbereiding

De kwartaalstatistiek Bouwobjecten In Voorbereiding (BIV) volgt de ontwikkeling van de totale bouwwaarde van nieuwe contracten bij architectenbureaus in Nederland. In 2007 is een nieuw controle- en correctieproces ontworpen voor deze statistiek: zie Van der Loo en Pannekoek (2007), waaraan dit voorbeeld is ontleend.

Bij het invullen van de BIV-vragenlijst moet de berichtgever over elk bouwobject apart een aantal vragen beantwoorden. Zo moet hij aankruisen of het gaat om een woning (*w*), een combinatiegebouw[1] (*c*) of geen van beide (*o* van overig). Een andere vraag betreft *n*, het aantal woningen in het gebouw. Bij een combinatiegebouw wordt ook het percentage vloeroppervlak bestemd voor wonen (*p*) gevraagd.

De opgave bevat een fout wanneer van de vakjes *w*, *c* en *o* er nul, twee of drie zijn aangekruist. In dat geval ligt het type bouwobject niet eenduidig vast. In bepaalde situaties kan deze fout deductief worden gecorrigeerd op basis van *n* en *p*.

Als een waarde van *n* groter dan nul is opgegeven en als bovendien *p* gelijk is aan 100% of niet is ingevuld, ligt het voor de hand dat het bouwobject een woning is. Als *n* groter dan nul is en als bovendien een *p* ongelijk aan 0 of 100% is ingevuld, ligt het voor de hand dat het bouwobject een combinatiegebouw is. Als ten slotte zowel *n* als *p* niet is ingevuld of de waarde 0 heeft, betreft het hoogstwaarschijnlijk een bouwobject uit de categorie overig. Deze interpretaties volgen uit de aanname dat de opgave kloppend moet worden gemaakt door zo weinig mogelijk waarden te veranderen.

We schrijven $w = \mathsf{T}$ als het vakje woning is aangekruist en anders $w = \mathsf{F}$, en doen hetzelfde voor *c* en *o*. Het correctievoorschrift luidt nu als volgt:

als $(w,c,o) \in \{ (\mathsf{T},\mathsf{T},\mathsf{T}) , (\mathsf{T},\mathsf{T},\mathsf{F}) , (\mathsf{T},\mathsf{F},\mathsf{T}) , (\mathsf{F},\mathsf{T},\mathsf{T}) , (\mathsf{F},\mathsf{F},\mathsf{F}) \}$
dan
als (p = “leeg” **of** $p = 100\%$) **en** $n > 0$
dan $(w,c,o) = (\mathsf{T},\mathsf{F},\mathsf{F})$
als $0\% < p < 100\%$ **en** $n > 0$
dan $(w,c,o) = (\mathsf{F},\mathsf{T},\mathsf{F})$
als (p = “leeg” **of** $p = 0\%$) **en** (n = “leeg” **of** $n = 0$)
dan $(w,c,o) = (\mathsf{F},\mathsf{F},\mathsf{T})$.

Dit is een klein deel van het controle- en correctieproces voor de statistiek BIV.

[1] Een combinatiegebouw wordt behalve voor wonen ook nog voor andere doelen gebruikt.

Bij de implementatie van het controle- en correctieproces voor BIV is het afleiden van de correctie steeds gescheiden van het daadwerkelijk aanbrengen van de correctie. In het bovenstaande voorbeeld wordt in eerste instantie alleen een indicator aangemaakt die voor elk record aangeeft of een deductieve correctie van toepassing is, en zo ja welke. Pas in de volgende stap worden de waarden van *w*, *c* en *o* in het record aangepast. Op deze manier is sprake van een transparant controle- en correctieproces, zodat achteraf precies kan worden achterhaald welke wijzigingen elk record heeft ondergaan.

2.4.2 Correctie van duizendfouten bij de Korte-termijn Statistieken

Bedrijfsenquêtes bevatten doorgaans een instructie aan de berichtgever dat alle financiële bedragen moeten worden afgerond op veelvouden van duizend euro. Sommige respondenten negeren deze instructie en geven waarden op die een factor 1000 groter zijn dan ze eigenlijk bedoelen. Het is duidelijk dat als deze 'duizendfouten' niet worden gecorrigeerd, de resulterende schattingen voor te publiceren cijfers te hoog uitvallen.

We spreken van een *uniforme duizendfout* wanneer alle financiële bedragen in een record een factor 1000 te groot zijn. Het is bekend dat vooral bij langere vragenlijsten ook records met een niet-uniforme (of partiële) duizendfout voorkomen. Een niet-uniforme duizendfout kan ontstaan wanneer meerdere personen elk een deel van de vragenlijst voor hun rekening nemen.

Duizendfouten worden gedetecteerd door een of meer opgegeven bedragen te vergelijken met referentiewaarden. De gebruikte referentiedata en de manier waarop de vergelijking plaatsvindt verschillen per statistiek en per statistisch bureau. Voorbeelden van referentiedata zijn: de opgave van dezelfde respondent in een eerdere periode, de mediaan van een aantal soortgelijke respondenten in een eerdere periode en registerdata over de respondent. Het is belangrijk dat deze referentiedata eerder op fouten zijn gecontroleerd.

Bij de Korte-termijn Statistieken (KS) worden duizendfouten als volgt gedetecteerd (Ter Haar, 2002). De door de respondent opgegeven totale omzet wordt vergeleken met de omzet over de meest recente periode waarvoor een opgave van de respondent beschikbaar is, tot maximaal zes perioden terug. De opgegeven omzet over deze eerdere periode moet bovendien ongelijk aan nul zijn. Er is sprake van een duizendfout indien geldt (met abs(a) de absolute waarde van a):

$$\text{abs}(omzet_t) > 300 \times \text{abs}(omzet_{t-i}) > 0, \quad \text{voor zekere } i \in \{1,\ldots,6\}.$$

Als geen data van de respondent uit een eerdere periode beschikbaar zijn, wordt gekeken naar de mediaan van de omzet over de vorige periode in het stratum van de respondent. Er is sprake van een duizendfout indien geldt:

$$\text{abs}(omzet_t) > 100 \times \text{stratummediaan}(omzet_{t-1}).$$

Wanneer op deze manier een duizendfout is gedetecteerd, wordt deze opgelost door de totale omzet én al zijn deelposten te delen door 1000.

Tabel 3 toont een voorbeeld van een record met een duizendfout die op deze manier gevonden wordt.

Tabel 3. Voorbeeld van een uniforme duizendfout

	referentiedata	data voor correctie	data na correctie
eerste deelpost omzet	3 331	3 148 249	3 148
tweede deelpost omzet	709	936 142	936
totale omzet	4 040	4 084 391	4 084

2.4.3 Correctie van systematische fouten bij de PS

De PS-vragenlijst bevat een groot aantal financiële variabelen die aan allerlei controleregels moeten voldoen, zoals deelposten die optellen tot een totaal en verhoudingen die tussen bepaalde grenswaarden liggen. Deze verzameling controleregels vormt een rijke bron voor het vinden van systematische fouten in de data. Vooralsnog (PS 2007) zijn acht correctiemethoden voor systematische fouten geïmplementeerd, waarvan we er hier drie behandelen. Vanaf paragraaf 2.4.4 komen drie methoden aan de orde die wellicht in de toekomst gebruikt zullen worden.

De belangrijkste systematische fout die automatisch wordt gecorrigeerd bij de PS is de uniforme duizendfout. De aanpak lijkt op die bij de KS (zie paragraaf 2.4.2). In plaats van de omzet direct te vergelijken met een eerdere periode, kijkt men hier naar de verhouding tussen de opgegeven omzet en het opgegeven aantal werkzame personen. Er wordt een duizendfout gedetecteerd wanneer deze verhouding sterk afwijkt van de stratummediaan in de vorige periode, i.e. wanneer

$$omzet_t / wp_t > 100 \times \text{stratummediaan}(omzet_{t-1} / wp_{t-1}), \qquad (2.4.1)$$

met *wp* het aantal werkzame personen.[2] Daarnaast worden BTW-registerdata en KS-data gebruikt als referentie. Voor de respondenten waarvan een positieve BTW- of KS-jaaromzet *omzet_extern* bekend is wordt gekeken of geldt:

$$omzet_t > 100 \times omzet\_extern_t.$$

Zo ja, dan wordt een duizendfout gedetecteerd. In beide gevallen worden alle opgegeven financiële bedragen gedeeld door 1000.

---

[2] Uit PS-documentatie blijkt dat in plaats van (2.4.1) de volgende formule is gebruikt:

$$omzet_t / wp_t > 100 \times \text{stratummediaan}(omzet_{t-1}) / \text{stratummediaan}(wp_{t-1}).$$

In het algemeen is de mediaan van de verhoudingen echter niet gelijk aan de verhouding van de medianen. Een eenvoudig voorbeeld:

$$\text{mediaan}(\{ 1 , 10^6 , 10^6 \}) / \text{mediaan}(\{ 1 , 1 , 10^6 \}) = 10^6 / 1 = 10^6,$$

terwijl

$$\text{mediaan}(\{ 1 / 1 , 10^6 / 1 , 10^6 / 10^6 \}) = \text{mediaan}(\{ 1 , 10^6 , 1 \}) = 1.$$

Een tweede systematische fout bij de PS betreft ten onrechte geplaatste mintekens. Wanneer een waarde moet worden afgetrokken, geven sommige respondenten dit aan door een minteken voor het opgegeven bedrag te plaatsen. Dit gebeurt ondanks het feit dat er reeds een gedrukt minteken op het enquêteformulier staat. Na vertoetsing heeft de variabele dan ten onrechte een negatieve waarde. Deze fout wordt verholpen door de absolute waarde van het ingevulde bedrag te nemen.

Verder wordt bij de PS gekeken naar records waarin deelposten zijn ingevuld, terwijl het bijbehorende totaal leeg is. Deze fout wordt gecorrigeerd door het totaal alsnog uit te rekenen op basis van de controleregel die zegt dat de deelposten moeten optellen tot het totaal. Bij deze correctie wordt aangenomen dat eventuele lege deelposten de waarde nul hebben. Pannekoek en Tempelman (2005) laten zien dat aan deze aanname niet altijd voldaan is.

2.4.4 Correctie van tekenfouten en baten-lastenverwisselingen

De resultatenrekening is een onderdeel van de PS-vragenlijst waarin veel fouten worden gemaakt. De rekening is opgebouwd uit saldi die moeten optellen tot een eindsaldo. Verder zijn sommige saldi uitgesplitst in baten en lasten. Vrij algemeen gesteld betekent dit dat de volgende controleregels van toepassing zijn:

$$\begin{cases} x_0 = x_{0,b} - x_{0,l} \\ \quad \vdots \\ x_m = x_{m,b} - x_{m,l} \\ x_n = x_0 + x_1 + \cdots + x_{n-1} \end{cases} \tag{2.4.2}$$

Hierbij stellen $x_0, x_1, \ldots, x_{n-1}$ de saldi voor, x_n het eindsaldo, en $x_{k,b}$ en $x_{k,l}$ de baten en lasten die horen bij saldo x_k. Om de notatie simpel te houden nemen we aan dat alleen $x_0, \ldots, x_m$ zijn uitgesplitst, voor zekere $m \in \{0,1,\ldots,n-1\}$. De onderste regel uit (2.4.2) wordt wel de *externe somregel* genoemd, terwijl de andere regels *interne somregels* heten.

Tabel 4 toont de opbouw van de resultatenrekening uit de vragenlijst die tot en met 2005 werd gebruikt voor de PS. De controleregels zijn gegeven door (2.4.2) met $n = 4$ en $m = n - 1 = 3$. Tabel 4 bevat ook drie voorbeelden van records die inconsistent zijn met betrekking tot (2.4.2).

In voorbeeld (a) zijn twee controleregels geschonden: de externe somregel en de interne somregel van het financieel resultaat. Opvallend genoeg kunnen we beide schendingen opheffen door alleen de waarde van x_1 te veranderen van 10 in –10 (zie tabel 5). Het ligt voor de hand om het record op deze manier te corrigeren, want elke andere oplossing vereist dat meer dan één waarde wordt aangepast.

In voorbeeld (b) zijn twee interne somregels geschonden. De voor de hand liggende manier om dit record consistent te maken is: verwissel de waarden van $x_{1,b}$ en $x_{1,l}$ en verwissel de waarden van $x_{3,b}$ en $x_{3,l}$ (zie tabel 5). Deze correctie maakt gebruik

van de door de respondent ingevulde bedragen en heeft daarom de voorkeur boven een oplossing waarbij synthetische waarden worden geïmputeerd.

Tabel 4. Voorbeelden van tekenfouten en baten-lastenverwisselingen

variabele	naam	(a)	(b)	(c)
$x_{0,b}$	*totale bedrijfsopbrengsten*	2 100	5 100	3 250
$x_{0,l}$	*totale bedrijfslasten*	1 950	4 650	3 550
x_0	*bedrijfsresultaat*	150	450	300
$x_{1,b}$	*financiële baten*	0	0	110
$x_{1,l}$	*financiële lasten*	10	130	10
x_1	*financieel resultaat*	10	130	100
$x_{2,b}$	*onttrekkingen en vrijval van voorzieningen*	20	20	50
$x_{2,l}$	*toevoegingen aan voorzieningen*	5	0	90
x_2	*saldo voorzieningen*	15	20	40
$x_{3,b}$	*buitengewone baten*	50	15	30
$x_{3,l}$	*buitengewone lasten*	10	25	10
x_3	*buitengewoon resultaat*	40	10	20
x_4	*resultaat voor belastingen (eindsaldo)*	195	610	–140

Tabel 5. Gecorrigeerde versie van tabel 4. Aangepaste waarden zijn vetgedrukt.

variabele	naam	(a)	(b)	(c)
$x_{0,b}$	*totale bedrijfsopbrengsten*	2 100	5 100	3 250
$x_{0,l}$	*totale bedrijfslasten*	1 950	4 650	3 550
x_0	*bedrijfsresultaat*	150	450	**–300**
$x_{1,b}$	*financiële baten*	0	**130**	110
$x_{1,l}$	*financiële lasten*	10	**0**	10
x_1	*financieel resultaat*	**–10**	130	100
$x_{2,b}$	*onttrekkingen en vrijval van voorzieningen*	20	20	**90**
$x_{2,l}$	*toevoegingen aan voorzieningen*	5	0	**50**
x_2	*saldo voorzieningen*	15	20	40
$x_{3,b}$	*buitengewone baten*	50	**25**	30
$x_{3,l}$	*buitengewone lasten*	10	**15**	10
x_3	*buitengewoon resultaat*	40	10	20
x_4	*resultaat voor belastingen (eindsaldo)*	195	610	–140

De typen inconsistenties in voorbeeld (a) en (b) heten respectievelijk *tekenfouten* en *baten-lastenverwisselingen*. Kortheidshalve gebruiken we ‘tekenfout’ ook wel als overkoepelende term. Deze fouten hangen met elkaar samen en moeten daarom simultaan worden gedetecteerd.

Er is sprake van een tekenfout wanneer voldaan is aan deze twee voorwaarden:

- Het record voldoet niet aan (2.4.2).
- Het record kan worden aangepast door alleen tekens van saldi te veranderen en baten en lasten te verwisselen, zodat het wel voldoet aan (2.4.2).

Hierbij mogen de totale bedrijfsopbrengsten ($x_{0,b}$) en totale bedrijfslasten ($x_{0,l}$) niet verwisseld worden, omdat deze via andere controleregels dan (2.4.2) nog verbonden zijn met posten van buiten de resultatenrekening. Bovendien is het vanwege de opbouw van de vragenlijst zeer onwaarschijnlijk dat een respondent deze twee antwoorden door elkaar zou halen.

Een wiskundige formulering van de bovenstaande voorwaarden is dat een inconsistent record een tekenfout bevat indien het volgende stelsel vergelijkingen een oplossing $(s_0,\ldots s_n;t_1,\ldots,t_m)\in\{-1,1\}$ heeft:

$$\begin{cases} x_0 s_0 = x_{0,b} - x_{0,l} \\ x_1 s_1 = (x_{1,b} - x_{1,l})t_1 \\ \quad\vdots \\ x_m s_m = (x_{m,b} - x_{m,l})t_m \\ x_n s_n = x_0 s_0 + x_1 s_1 + \cdots + x_{n-1} s_{n-1} \end{cases} \tag{2.4.3}$$

Als een dergelijke oplossing gevonden wordt, is ook meteen duidelijk hoe de tekenfout gecorrigeerd kan worden. Voor elke $s_j = -1$ moet het teken van de bijbehorende x_j worden veranderd, en voor elke $t_k = -1$ moeten $x_{k,b}$ en $x_{k,l}$ worden verwisseld. Het is niet moeilijk om in te zien dat het resulterende record voldoet aan (2.4.2). We merken op dat een variabele t_0 ontbreekt in (2.4.3) omdat $x_{0,b}$ en $x_{0,l}$ niet in aanmerking komen voor verwisseling. Afgezien van de boven gegeven motivatie heeft dit ook een technische reden: de eerste vergelijking in (2.4.3) legt de waarde van s_0 nu uniek vast. Door een van de variabelen vast te leggen voorkomen we dat een oplossing van (2.4.3) is om te werken tot een alternatieve oplossing door alle variabelen met –1 te vermenigvuldigen.

Toegepast op voorbeeld (c) uit tabel 4 wordt (2.4.3):

$$\begin{cases} 300 s_0 = -300 \\ 100 s_1 = 100 t_1 \\ 40 s_2 = -40 t_2 \\ 20 s_3 = 20 t_3 \\ -140 s_4 = 300 s_0 + 100 s_1 + 40 s_2 + 20 s_3 \end{cases}$$

Dit stelsel heeft de volgende oplossing:

$$(s_0 = -1, s_1 = 1, s_2 = 1, s_3 = 1, s_4 = 1; t_1 = 1, t_2 = -1, t_3 = 1).$$

Dus voorbeeld (c) bevat een tekenfout. Ter correctie moeten we de waarde van x_0 veranderen van 300 in –300, en de waarden van $x_{2,b}$ en $x_{2,l}$ verwisselen. In tabel 5 is te zien dat deze aanpassingen inderdaad een consistent record opleveren.

Samengevat luidt de methode voor het detecteren en corrigeren van tekenfouten en baten-lastenverwisselingen in de PS-resultatenrekening als volgt:

1. Gegeven een record dat niet voldoet aan (2.4.2), bepaal stelsel (2.4.3).
2. Vind de[3] oplossing $(s_0,\ldots s_n;t_1,\ldots,t_m)\in\{-1,1\}$ van (2.4.3), als zij bestaat. Stop als (2.4.3) geen oplossing heeft, ga anders verder met stap 3.
3. Voor $j=0,\ldots,n$ en voor $k=1,\ldots,m$: verander het teken van x_j indien $s_j=-1$, en verwissel de waarden van $x_{k,b}$ en $x_{k,l}$ indien $t_k=-1$.

De enige niet-triviale stap in dit schema is stap 2, het oplossen van stelsel (2.4.3). Aangezien n en m klein zijn kan de oplossing in principe worden gevonden door systematisch alle $2^{n+m+1}-1$ combinaties van $s_0,\ldots,s_n$ en $t_1,\ldots,t_m$ te proberen. In Scholtus (2007) wordt het oplossen van (2.4.3) herschreven als een binair lineair programmingsprobleem dat met standaardsoftware kan worden opgelost.

2.4.5 Correctie van doortelfouten

Een andere fout die regelmatig optreedt in de PS-resultatenrekening is de zogeheten *doortelfout*. Tabel 6 toont drie voorbeelden van records met een doortelfout. De fout ontstaat omdat de respondent de resultatenrekening als het ware 'cumulatief' invult. In voorbeeld (a) en (b) gebeurt dit consequent, in voorbeeld (c) niet. Bovendien zijn in voorbeeld (c) ook nog de financiële baten en lasten verwisseld.

Stel dat een gegeven record niet voldoet aan de externe somregel en ook niet aan de k^{e} interne somregel, waarbij $k\in\{1,\ldots,n-1\}$, maar dat wel geldt:

$$x_k = x_{k-1} + x_{k,b} - x_{k,l}\,. \tag{2.4.4}$$

In dat geval is sprake van een doortelfout, die gecorrigeerd kan worden door de waarden van x_k, $x_{k,b}$ en $x_{k,l}$ te vervangen door

$$x'_k = x_k - x_{k-1}\,, \qquad x'_{k,b} = x_{k,b}\,, \qquad x'_{k,l} = x_{k,l}\,.$$

Uit (2.4.4) volgt onmiddellijk dat $x'_k = x'_{k,b} - x'_{k,l}$. Door deze stap achtereenvolgens uit te voeren voor elke $k\in\{1,\ldots,n-1\}$ kunnen voorbeeld (a) en (b) volledig consistent gemaakt worden. Hierbij moet overigens steeds de *oorspronkelijke* waarde van x_{k-1} in (2.4.4) worden ingevuld en niet x'_{k-1}.

Om ook rekening te houden met mogelijke tekenfouten moet in plaats van (2.4.4) gekeken worden of er $(\lambda,\mu)\in\{-1,1\}$ bestaan zodat geldt:

$$\lambda x_k = x_{k-1} + \mu(x_{k,b} - x_{k,l})\,. \tag{2.4.5}$$

---

[3] In appendix A van Scholtus (2008a) wordt bewezen dat (2.4.3) onder zeer milde voorwaarden hooguit één oplossing heeft.

Tabel 6. Voorbeelden van doortelfouten

variabele	naam	(a)	(b)	(c)
$x_{0,b}$	*totale bedrijfsopbrengsten*	6 700	8 300	6 900
$x_{0,l}$	*totale bedrijfslasten*	5 650	5 400	6 150
x_0	*bedrijfsresultaat*	1 050	2 900	750
$x_{1,b}$	*financiële baten*	0	0	0
$x_{1,l}$	*financiële lasten*	0	150	40
x_1	*financieel resultaat*	1 050	2 750	790
$x_{2,b}$	*onttrekkingen en vrijval van voorzieningen*	0	0	0
$x_{2,l}$	*toevoegingen aan voorzieningen*	0	30	0
x_2	*saldo voorzieningen*	1 050	2 720	0
$x_{3,b}$	*buitengewone baten*	0	0	0
$x_{3,l}$	*buitengewone lasten*	0	110	0
x_3	*buitengewoon resultaat*	1 050	2 610	0
x_4	*resultaat voor belastingen (eindsaldo)*	1 050	2 610	790

Zo ja, dan bevat het record een doortelfout. Als $\lambda = -1$ heeft bovendien x_k een verkeerd teken en als $\mu = -1$ zijn $x_{k,b}$ en $x_{k,l}$ verwisseld. De doortelfout én de tekenfout worden gecorrigeerd door x_k, $x_{k,b}$ en $x_{k,l}$ te vervangen door:

$$\begin{cases} x'_k = \lambda x_k - x_{k-1} \\ x'_{k,b} = \dfrac{1+\mu}{2} x_{k,b} + \dfrac{1-\mu}{2} x_{k,l} \\ x'_{k,l} = \dfrac{1-\mu}{2} x_{k,b} + \dfrac{1+\mu}{2} x_{k,l} \end{cases} \tag{2.4.6}$$

Merk op dat $x'_{k,b} = x_{k,b}$ als $\mu = 1$ en $x'_{k,b} = x_{k,l}$ als $\mu = -1$, en iets soortgelijks voor $x'_{k,l}$. Uit (2.4.5) volgt dat $x'_k = x'_{k,b} - x'_{k,l}$.

Bijvoorbeeld: in voorbeeld (c) is te zien dat

$$1 \cdot x_1 = 790 = 750 - (-40) = x_0 + (-1) \cdot (x_{1,b} - x_{1,l}),$$

d.w.z. (2.4.5) geldt voor $(\lambda = 1, \mu = -1)$. Volgens (2.4.6) is de fout te corrigeren door te kiezen: $x'_1 = x_1 - x_0 = 40$, $x'_{1,b} = x_{1,l} = 40$ en $x'_{1,l} = x_{1,b} = 0$. Nu geldt inderdaad dat $x'_1 = x'_{1,b} - x'_{1,l}$. Omdat in dit record geen andere fouten voorkomen, is met deze correctie ook meteen voldaan aan de externe somregel.

Een iets gedetailleerdere uitwerking van deze methode staat in Scholtus (2008a).

2.4.6 Correctie van eenvoudige schrijffouten

In paragraaf 2.3.2 zagen we een voorbeeld waarbij een inconsistentie deductief kon worden opgelost door aan te nemen dat de respondent per ongeluk twee cijfers had

verwisseld (door ‘283’ te schrijven in plaats van ‘238’). Het verwisselen van twee opeenvolgende cijfers is een voorbeeld van een eenvoudige schrijffout. Andere voorbeelden zijn:

- het toevoegen van een cijfer (bijvoorbeeld: ‘46297’ in plaats van ‘4627’);
- het weglaten van een cijfer (bijvoorbeeld: ‘427’ in plaats van ‘4627’);
- het vervangen van een cijfer (bijvoorbeeld: ‘4687’ in plaats van ‘4627’).

Eenvoudige schrijffouten zijn gemakkelijk te maken en komen in de praktijk dan ook vaak voor. Uit een inventarisatie bij de PS Groothandel 2007 bleek bijvoorbeeld dat bijna 10% van alle inconsistenties in lineaire gelijkheden te verklaren was als één van de vier zojuist genoemde fouten (Scholtus, 2009).

In het geval dat de data moeten voldoen aan één lineaire gelijkheid kunnen eenvoudige schrijffouten gemakkelijk worden gedetecteerd, zoals in het voorbeeld uit paragraaf 2.3.2. Van de Pol e.a. (1997) behandelen deze situatie in detail. Bij de controleregels van de PS is echter sprake van een systeem van lineaire gelijkheden die met elkaar samenhangen. Naast controleregel (2.3.4) moeten *omzet* en *kosten* bijvoorbeeld gelijk zijn aan de som van enkele deelposten. Een deductieve methode voor het corrigeren van eenvoudige schrijffouten in deze complexere situatie is beschreven door Scholtus (2009). Hier behandelen we de methode alleen aan de hand van een voorbeeld.

Stel, een record bestaat uit elf variabelen die moeten voldoen aan vijf controleregels:

$$\begin{cases} x_1 + x_2 = x_3 \\ x_2 = x_4 \\ x_5 + x_6 + x_7 = x_8 \\ x_3 + x_8 = x_9 \\ x_9 - x_{10} = x_{11} \end{cases}$$

Het volgende record schendt de tweede, vierde en vijfde controleregel:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1452	116	1568	161	323	76	12	411	19979	1842	137

Om te zien of één of meer inconsistenties kunnen worden verklaard als eenvoudige schrijffout, bepalen we eerst welke variabelen alleen voorkomen in geschonden lineaire gelijkheden. Dit zijn namelijk de enige variabelen die we deductief kunnen veranderen zonder nieuwe inconsistenties te introduceren. In dit voorbeeld gaat het om de variabelen die alleen voorkomen in de tweede, vierde en vijfde controleregel, en dit zijn x_4, x_9, x_{10} en x_{11}.

Voor elke genoemde variabele lopen we de lineaire gelijkheden af waarin hij voorkomt. Bij elke controleregel bepalen we welke waarde de variabele in kwestie zou moeten krijgen om de inconsistentie op te heffen. Merk op: een dergelijke waarde bestaat altijd als de controleregel de vorm heeft van een lineaire gelijkheid. Vervolgens vergelijken we de nieuwe waarde met de ingevulde waarde. Als de verandering kan worden verklaard door een eenvoudige schrijffout, dan bewaren we

de voorgestelde verandering, en anders niet. Nadat alle controleregels zijn behandeld, kijken we hoe vaak elke voorgestelde waarde genoemd is.

In dit voorbeeld komt x_4 alleen voor in de tweede controleregel. Om te voldoen aan deze regel zouden we de waarde $\tilde{x}_4 = 116$ moeten invullen. De huidige waarde is 161, en de nieuwe waarde is hieruit te verklaren door een eenvoudige schrijffout, namelijk het verwisselen van twee opeenvolgende cijfers. Dat wil zeggen: het is voorstelbaar dat de werkelijke waarde 116 door een schrijffout is veranderd in de waargenomen waarde 161.

Variabele x_9 komt voor in zowel de vierde als de vijfde controleregel. Het blijkt dat aan beide regels voldaan kan worden door de waarde $\tilde{x}_9 = 1979$ in te vullen. Ook deze waarde kan verklaard worden door een eenvoudige schrijffout: de werkelijke waarde 1979 is wellicht veranderd in de waargenomen waarde 19979 omdat per ongeluk een cijfer is toegevoegd.

Voor x_{10} vinden we $\tilde{x}_{10} = 19842$, en uit deze waarde kan de waargenomen waarde 1842 worden gevonden door het weglaten van een cijfer. Ook dit zou dus een eenvoudige schrijffout kunnen zijn.

De benodigde waarde van x_{11} ten slotte, $\tilde{x}_{11} = 18137$, kan niet worden verklaard door een van de bovengenoemde eenvoudige schrijffouten. Dit betekent dat we x_{11} verder buiten beschouwing laten.

Vervolgens moet uit de gevonden mogelijke schrijffouten een keuze worden gemaakt. Het is duidelijk dat voor elke variabele hooguit één nieuwe waarde kan worden gekozen. Bovendien heeft het geen zin om twee variabelen te veranderen die voorkomen in dezelfde lineaire gelijkheid: per saldo blijft de controleregel dan geschonden. Gegeven deze twee beperkingen kiezen we de combinatie van voorgestelde veranderingen die leidt tot een maximaal aantal kloppend gemaakte inconsistenties.

In het voorbeeld hebben we bij elke variabele hooguit één nieuwe waarde gevonden, dus de eerste beperking speelt geen rol. Nadere beschouwing van de controleregels laat zien dat x_4 niet in dezelfde regel voorkomt als x_9 en x_{10}, maar dat x_9 en x_{10} wel samen in een gelijkheid voorkomen. Er zijn daarom twee mogelijke keuzes: ofwel x_4 en x_9 veranderen, ofwel x_4 en x_{10}. Het aantal kloppend gemaakte inconsistenties bij deze keuzes is respectievelijk drie en twee. We kiezen daarom voor de eerste combinatie van deductieve correcties. Het resulterende record is:

x_1	x_2	x_3	$\tilde{x}_4$	x_5	x_6	x_7	x_8	$\tilde{x}_9$	x_{10}	x_{11}
1452	116	1568	116	323	76	12	411	1979	1842	137

Dit record voldoet aan alle controleregels.

Een uitgebreide beschrijving van dit algoritme voor het deductief corrigeren van eenvoudige schrijffouten is te vinden in Scholtus (2009).

2.5 Kwaliteitsindicatoren

Zoals gezegd doet men bij het opstellen van een deductieve correctiemethode altijd aannames over de data. Als deze aannames geldig zijn, levert de methode de best mogelijke correcties op. Bij onrealistische aannames kan de methode echter vertekening introduceren. Het is daarom belangrijk om te onderzoeken of de data voldoen aan de gemaakte aannames.

Een indicator voor het nut van het toepassen van een deductieve correctiemethode is het aantal fouten dat hij oplost in een realistisch databestand. Een ander aspect is de winst in efficiëntie die behaald wordt doordat een aantal records na implementatie van de deductieve methode minder controle en correctie nodig heeft, of een lichtere vorm. Een voorbeeld van dit laatste vinden we bij het controle- en correctieproces van de PS, waar de keuze bestaat tussen handmatig (‘duur’) en automatisch gaafmaken (‘goedkoop’). De in paragraaf 2.4.3 beschreven deductieve correcties zorgen ervoor dat meer records geschikt zijn voor de geautomatiseerde variant.

3. Interactief gaafmaken

3.1 Korte beschrijving

Bij interactief of handmatig gaafmaken voert een gaafmaker correcties uit op een record, gebruikmakende van een programma dat controles uitvoert en toont welke variabelen zijn betrokken bij een geschonden controleregel. Het is dan aan de gaafmaker om te besluiten welke variabele moet worden gecorrigeerd en wat de correcte waarde van deze variabele zou kunnen zijn. Op het CBS en vele andere statistische bureaus wordt hiervoor BLAISE gebruikt. Bij telefonische of persoonlijke interviews kan het gaafmaken al tijdens het invoeren gebeuren. Om de kwaliteit van het gaafmaken te waarborgen is het raadzaam om eerst een gaafmaakinstructie op te stellen.

3.2 Toepasbaarheid

Om het doel van interactief gaafmaken duidelijk te maken wordt er eerst een aantal soorten waarden onderscheiden.

De *ware waarde* wordt verkregen bij een ideaal meet- en verwerkingsproces. Oftewel de waarde die wij verkrijgen als de berichtgever volgens de juiste definitie en met behulp van de juiste boekhouding cijfers levert die niet tijdens het verwerkingsproces zijn gewijzigd. De ware waarde kennen we echter niet, en dus weten we ook niet of we deze hebben waargenomen.

Wat we kunnen bereiken is een *correcte waarde*, namelijk als een branche expert deze correct vindt op basis van de beschikbare informatie, oftewel de bij het CBS beschikbare informatie over waarden van gerelateerde variabelen en records en een uitgebreide set van gaafmaakregels. Of deze ‘correcte’ waarde de ware waarde benadert hangt af van de beschikbare informatie, de kunde van de gaafmaker, de kwaliteit van de gaafmaakinstructie en de mate waarin deze wordt opgevolgd.

In veel gevallen kunnen we echter niets beters krijgen dan een *acceptabele waarde*, d.w.z. dat deze aan harde gaafmaakregels voldoet. Als een gaafmaker alleen harde fouten oplost dan levert dit een acceptabele waarde. Dit geldt ook als een record automatisch wordt gaafgemaakt. In dat geval worden namelijk alleen harde fouten opgelost, zie hoofdstuk 5.

Het doel van interactief gaafmaken is om waarden in een record correct te maken. In het geval van hoge tijdsdruk kan het voorkomen dat een beperkt aantal variabelen uitvoerig wordt gecontroleerd en dat de waarden van de overige variabelen hooguit acceptabel worden gemaakt tijdens de microgaafmaakfase. Dit moet een keuze zijn van de projectleider, niet van de gaafmaker zelf. Invloedrijke fouten die blijven zitten bij het microgaafmaken moeten tijdens het macrogaafmaken alsnog interactief worden gaafgemaakt.

Interactief gaafmaken is vooral interessant als de data maar deels automatisch kunnen worden gaafgemaakt of als de kwaliteit van interactief gaafgemaakte data aanzienlijk beter is. Een ander voordeel is dat bij het interactief gaafmaken informatie kan worden opgezocht op o.a. internet of een schriftelijk ingevulde vragenlijst. Bij primaire waarneming kunnen berichtgevers worden gebeld. Het is raadzaam om dit alleen te doen als dit cruciaal is voor inzicht in het statistisch proces of de kwaliteit van een publicatiecijfer. Interactief gaafmaken geeft tevens de mogelijkheid om foutpatronen te herkennen die regelmatig voorkomen. Er kan dan gekeken worden of deze foutpatronen in het vervolg automatisch kunnen worden (voor)gaafgemaakt.

Een belangrijke voorwaarde is dat er een set van controleregels beschikbaar is waarmee onderlinge relaties en het waardebereik van (ratio's van) variabelen kan worden gecontroleerd. Tevens dient er een programma beschikbaar te zijn dat controleregels kan nalopen en onderscheid kan maken tussen harde en zachte fouten. Dit programma moet ook in staat zijn om referentiewaarden te tonen voor een record, zie paragraaf 3.3.

3.3 Uitgebreide beschrijving

3.3.1 Inleiding

Bij het interviewen van personen kan een formulier direct interactief worden gaafgemaakt als er gebruik wordt gemaakt van een computer-assisted interviewing (CAI) systeem, zoals BLAISE. Inconsistenties kunnen dan worden gedetecteerd door het CAI-systeem en gecorrigeerd door de interviewer in samenspraak met de geïnterviewde. Bij bedrijven is dit alleen mogelijk als deze door de buitendienst worden bezocht. Als een bedrijf een formulier opstuurt dan is het raadzaam om alleen interactief gaaf te maken als deze potentiële invloedrijke waarden bevat. Dit kan worden bepaald aan de hand van scorefuncties, zie hoofdstuk 4. De gaafmaker kan deze scores bekijken om te bepalen welke variabelen potentiële invloedrijke fouten bevatten.

Als een door een berichtgever ingevulde enquête binnenkomt op het CBS vindt er eerst een automatische correctieslag plaats, waarbij overduidelijke fouten worden gecorrigeerd. Bij het interactief corrigeren van *voorgaafgemaakte* data worden resterende foute waarden verbeterd door contact op te nemen met de berichtgever of door gebruik te maken van expertkennis in combinatie met referentiegegevens zoals andere gegevens van dezelfde berichtgever (uit een vorige periode, een andere enquête, een administratie), de oorspronkelijke opgave van de berichtgever of representatieve waarden of trends voor soortgelijke berichtgevers.

Bij interactief corrigeren van een enquête met gerelateerde variabelen kan het wijzigen van een waarde van een variabele leiden tot schending van andere controleregels. Mogelijk moeten er dan ook andere variabelen worden gecorrigeerd. Een gaafmaker zal er in ieder geval voor moeten zorgen dat de data voldoen aan alle

harde controleregels. Hij/zij moet bepalen welke variabele in een geschonden controleregel dient te worden gecorrigeerd en wat de correcte waarde is.

Bij interactief corrigeren van een korte-termijnstatistiek kan bijvoorbeeld bekeken worden of een invloedrijke verdachte waarde past in het seizoenpatroon voor vergelijkbare eenheden. Bij economische statistieken kan rekening worden gehouden met een algemeen beeld van de economische ontwikkeling in de afgelopen perioden.

3.3.2 Opstellen van gaafmaakinstructie

Het is niet voldoende om een programma te hebben dat de variabelen van te controleren records toont, aangeeft waarom een record is geselecteerd, welke controleregels zijn geschonden, en gerelateerde variabelen uit andere bronnen en eerdere perioden of processtappen zichtbaar maakt. Om te voorkomen dat een gaafmaker een verkeerde gaafmaakstrategie hanteert is het belangrijk om een gaafmaakinstructie op te stellen.

Een gaafmaakinstructie moet in ieder geval de volgende onderdelen bevatten:

- Uitleg over het waarneem- en verwerkingsproces dat heeft plaatsgevonden.
- Instructie over de volgorde waarin de geselecteerde records moeten worden afgehandeld. Als het interactief gaafmaken onderdeel uitmaakt van macrogaafmaken dan is er ook een analyseinstructie nodig. Deze geeft aan hoe (extra) records geselecteerd kunnen worden.
- Bij selectief gaafmaken is er uitleg nodig over het selectiecriterium en hoe dit gebruikt kan worden bij het opsporen van fouten in een record.
- Een overzicht van het soort fouten dat kan optreden in de data, zoals SBI-fouten, grootteklasse-fouten, meetfouten en verwerkingsfouten.
- Tips over het opsporen van een bepaalde fout. Bij statistieken over meerdere gerelateerde variabelen kan gekeken worden naar scatterplots en kengetallen. Dit is voor andere statistieken ook mogelijk als er gerelateerde variabelen uit andere bronnen gekoppeld kunnen worden. Bij korte-termijnstatistieken kan het seizoenpatroon voor een record afgezet worden tegen het het seizoenpatroon voor de branche.
- Suggesties over aanvullende informatie die kan worden opgezocht, bijvoorbeeld via raadpleeg-ABR, branche organisaties, internet of Cdfoon. Googlen van een bedrijfsnaam helpt bijvoorbeeld bij het bepalen of er sprake is van een SBI-fout.
- Per type fout een indicatie hoe de fout kan worden gecorrigeerd. Bij systematische fouten kan er mogelijk een correctieregel worden gespecificeerd.
- Een instructie over het vastleggen van de gaafmaakacties die zijn ondernomen, bijvoorbeeld via een opmerkingenveld in de gaafmaaktool. Als er een SBI-fout of grootteklasse-fout is geconstateerd dan moet duidelijk zijn of deze door moet worden gegeven aan de mensen die het populatiekader beheren.

3.4 Kwaliteitsindicatoren

Om te bepalen of interactief gaafmaken leidt tot verbetering van microdata en publicatiecijfers kan gekeken worden naar een aantal aspecten:

1. percentage en aantal records dat niet aan een controleregel voldoet vóór interactief gaafmaken;
2. percentage en aantal records dat niet aan een controleregel voldoet na interactief gaafmaken;
3. publicatiecijfer berekend op basis van voorgaafgemaakte data voor interactief gaafgemaakte records en acceptabele data voor automatisch gaafgemaakte records;
4. publicatiecijfer berekend op basis van interactief gaafgemaakte data.

Het verschil tussen indicator 1 en 2 geeft inzicht in de mate waarin schendingen van controleregels worden opgelost bij het interactief corrigeren. Het verschil tussen indicator 3 en 4 geeft aan wat het effect van interactief gaafmaken is op het publicatiecijfer.

4. Selectief gaafmaken

4.1 Korte beschrijving

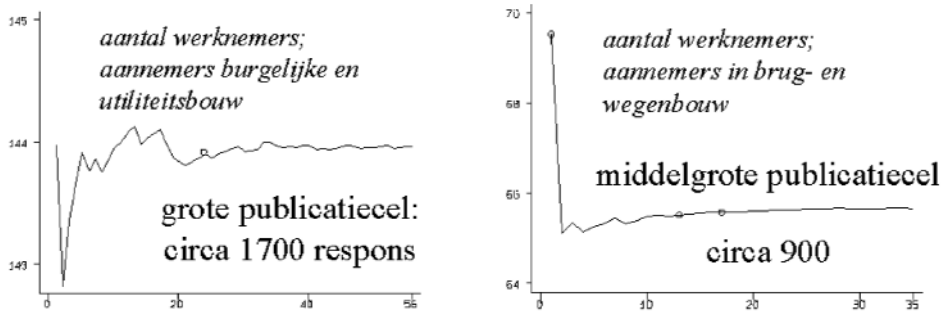
Handmatig of interactief gaafmaken is een van de meest tijdrovende en kostbare onderdelen van het statistische verwerkingsproces voor bedrijfsstatistieken. In het verleden werden vaak alle records handmatig gaafgemaakt wat gepaard ging met hoge kosten en lange doorlooptijden. Dit is een stimulans geweest voor onderzoek naar mogelijkheden om het handmatige werk te beperken. De laatste twee decennia heeft dit onderzoek laten zien dat het niet nodig en niet wenselijk is om *alle* records handmatig gaaf te maken. Voor veel records hebben de handmatige correcties namelijk nauwelijks invloed op de uiteindelijke publicatiecijfers zoals schattingen van (deel)populatietotalen of ontwikkelingen. Het handmatig gaafmaken kan daarom beperkt worden tot de records waarbij correcties wél van invloed zijn op de publicatiecijfers, deze gerichte inperking wordt selectief gaafmaken genoemd. De bedoeling van selectief gaafmaken is het beperken van het handmatig gaafmaken om daarmee kosten te besparen, de doorlooptijd te verkorten, en de enquêtedruk te verminderen met slechts een minimaal verlies in de kwaliteit van de publicatiecijfers.

Om selectief gaaf te maken kan aan ieder record een score toegekend worden die aangeeft hoe groot de verwachte invloed is op publicatiecijfers als het record handmatig gaafgemaakt zou worden. De records met hoge scores (veel invloed) hebben de hoogste prioriteit voor het handmatig gaafmaken. Voor lage scores, onder een bepaalde grenswaarde, hoeft niet meer handmatig gaafgemaakt te worden. Methoden waarmee de scores bepaald kunnen worden en methoden waarmee de grenswaarde vastgesteld kan worden behoren tot de methodologie van het selectief gaafmaken die in dit hoofdstuk wordt besproken.

De onderstaande figuren illustreren de afnemende invloed van steeds minder belangrijke correcties op de schattingen van totalen (overgenomen uit Hoogland e.a. 2002). De geschatte totalen zijn weergegeven als functie van het aantal gaafgemaakte records waarbij de records zijn gaafgemaakt in volgorde van afnemende invloed op deze schatting. De linkergrafiek in figuur 3 betreft de schatting van het aantal werknemers in de burgerlijke- en utiliteitsbouw in een publicatiecel van 1700 respondenten. Uit deze figuur blijkt dat het corrigeren van meer dan 40 records geen invloed meer heeft op de schatting van het populatietotaal. De rechtergrafiek in figuur 3 laat zien dat de schatting van het aantal werknemers in de brug- en wegenbouw in een publicatiecel van 900 werknemers nauwelijks meer veranderd nadat de 20 records met de grootste fouten zijn gecorrigeerd. Deze figuren laten voorbeelden zien waarbij handmatig gaafmaken beperkt kan blijven tot een klein deel van de records. Niet in alle situaties zal de mate van handmatig gaafmaken zo sterkt beperkt kunnen blijven als in deze voorbeelden, maar uit het oogpunt van efficiëntie is het vrijwel altijd van belang om handmatig gaafmaken

selectief toe te passen. Methoden om te bepalen welke records in aanmerking komen voor handmatig gaafmaken en voor welke records dat niet nodig is worden in dit hoofdstuk besproken.

Figuur 3. Geschatte totalen als functie van het aantal gaafgemaakte records, waarbij is gaafgemaakt in volgorde van invloed op het totaal



4.2 Toepasbaarheid

Selectief gaafmaken wordt vrijwel uitsluitend toegepast bij bedrijfsstatistieken en numerieke variabelen. De invloed van het gaafmaken op publicatiecijfers verschilt sterk tussen bedrijven eenvoudigweg omdat ze (soms veel) in omvang verschillen, en daardoor een sterk verschillend aandeel hebben in de schatting van een totaal. Voor persoonsstatistieken geldt dit veel minder. Ieder individu is ongeveer even belangrijk voor de schatting van een totaal, dit belang wordt uitgedrukt in de ophooggewichten die niet heel sterk verschillen tussen respondenten. Toch kunnen ook bij persoonsstatistieken records met sterk afwijkende waarden, zoals extreem hoge inkomens, opgespoord en handmatig gecontroleerd worden.

Hoewel het bij vrijwel alle economische statistieken de moeite loont om selectief gaaf te maken hoeft dit niet altijd op micro-niveau te gebeuren. Een alternatief is macro-gaafmaken (zie hoofdstuk 7). Een voordeel van selectief gaafmaken op micro-niveau boven macro-gaafmaken is dat al tijdens de dataverzamelingsperiode met gaafmaken kan worden begonnen. Een voordeel van gaafmaken op macro-niveau is dat de belangrijke fouten beter op te sporen zijn als alle data binnen zijn.

In het verwerkingsproces vindt de selectie van records voor het handmatig gaafmaken plaats nadat de systematische fouten zijn gecorrigeerd (zie hoofdstuk 2). Dit is een automatische correctieslag waarbij vaak belangrijke fouten (zoals duizendfouten) gecorrigeerd worden. Als dergelijke fouten niet eerst gecorrigeerd zouden worden, zullen ze tijdens het selectief gaafmaken herkend worden als invloedrijke fouten en zullen de betrokken records aan de gaafmakers worden aangeboden. Het is uiteraard inefficiënt om gaafmakers te belasten met deze automatisch oplosbare fouten.

4.3 Uitgebreide beschrijving

4.3.1 *Inleiding*

Het belangrijkste instrument in een selectief gaafmaakproces is de scorefunctie. Deze functie kent scores toe aan de individuele records op basis waarvan zij een prioriteit krijgen voor het handmatig gaafmaken. De records met de hoogste scores komen hiervoor als eerste in aanmerking. Dergelijke scores zijn in principe gelijk aan wat op het CBS ook wel de plausibiliteitsindex genoemd wordt. Het enige verschil is dat bij de plausibiliteitsindex lage waarden corresponderen met een hoge prioriteit voor handmatig gaafmaken terwijl dit bij de gebruikelijke scorefuncties andersom is.

Een score voor een record kan opgebouwd zijn uit een aantal verschillende deelscores of lokale scores. Vaak zijn dit afzonderlijke scores voor ieder van de belangrijkste variabelen. Deze scores geven een indicatie van het verwachte effect van het gaafmaken van die variabelen voor de schattingen van belangrijke doelparameters zoals de totalen van die variabelen en ontwikkelingen in die totalen. Een lokale score voor een variabele *j* in een record *i* heeft in het algemeen de vorm,

$$s_{ij} = belang_{ij} \times risico_{ij}$$

De risicofactor wordt bepaald door de ruwe waarde van de variabele te vergelijken met een zogenaamde referentiewaarde. De referentiewaarde geeft een indicatie van de waarde die men zou kunnen verwachten en wordt bepaald op grond van informatie uit andere bronnen dan het huidige onderzoek, zoals een eerdere versie van dat onderzoek, of andere onderzoeken of registraties met soortgelijke variabelen. De mate waarin een waarde afwijkt van de referentiewaarde bepaalt het risico. Het risico is groot als de afwijking groot is, de ruwe waarde zou dan wel eens fout kunnen zijn en in dat geval bovendien tot een grote correctie kunnen leiden. Als de afwijking klein is, is er geen reden om aan te nemen dat de waarde fout zou kunnen zijn en bovendien, mocht dat toch het geval zijn dan is de correctie waarschijnlijk klein. De belangfactor geeft aan hoe zeer het record bijdraagt aan de schatting van het publicatiecijfer. Deze factor heeft vooral te maken met de omvang van een bedrijf, een procentueel kleine correctie in de waarde van een groot bedrijf kan toch een substantiële invloed hebben op een publicatiecijfer.

De hierboven beschreven lokale scores zijn gerelateerd aan de schatting voor een doelparameter. Deze scores worden daarom ook wel schatter-gerelateerde scores genoemd. Een ander type scores is gebaseerd op schendingen van controleregels (edits) zoals het aantal harde fouten of het aantal ten onrechte lege velden (partiële non-respons) die ook iets zeggen over de kwaliteit van een record. Dit laatste type scores wordt edit-gerelateerde scores genoemd. In paragraaf 4.3.3. wordt aangegeven dat edit-gerelateerde en schatter-gerelateerde scores gecombineerd kunnen worden tot één record score.

Om een selectieve gaafmaak strategie te implementeren moeten de volgende stappen genomen worden:

- Definieren van lokale scores op basis van beschikbare referentiewaarden die de verwachte waarden zo goed mogelijke benaderen. In paragraaf 4.3.2 worden enkele veel gebruikte lokale scorefuncties besproken.
- Het combineren van de lokale scores tot een record score of globale score. Dit wordt besproken in paragraaf 4.3.3.
- Het vaststellen van een grenswaarde voor de record scores waarmee de handmatig te verwerken records geselecteerd kunnen worden. Het bepalen van de grenswaarde wordt besproken in paragraaf 4.3.4. De overige records zullen automatisch worden gaaf gemaakt. Automatisch gaafmaken wordt behandeld in hoofdstuk 5 en 6.

4.3.2 Lokale scorefuncties voor totalen en ontwikkelingen

De twee belangrijkste doelparameters bij bedrijfsstatistieken zijn totalen van (deel)populaties en ontwikkelingen binnen (deel)populaties. In deze paragraaf worden voor ieder van deze doelparameters veel gebruikte scorefuncties besproken. De meeste in de praktijk toegepaste lokale scorefuncties kunnen opgevat worden als varianten van de hier besproken functies, waarbij soms de belang of risico-component aan een specifieke situatie wordt aangepast.

Om een scorefunctie op te stellen die gericht is op de effecten van het gaafmaken op de schatting van totalen, beschouwen we eerst de gebruikelijke schatter voor het populatietotaal van een doelvariabele y_j. Deze schatter kan geschreven worden als

$$\hat{Y}_j = \sum_{i \in s} w_i \hat{y}_{ij} \ , \tag{4.3.1}$$

met s de verzameling responderende eenheden en w_i gewichten die corrigeren voor ongelijke insluitkansen en non-respons. De $\hat{y}_{ij}$ in (4.3.1) zijn gaafgemaakte waarden, d.w.z. ze hebben een gaafmaakproces doorlopen waarbij sommige van de ruwe waarden, zeg y_{ij}, hetzij door gaafmakers of door een geautomatiseerd proces zijn vervangen betere waarden $\hat{y}_{ij}$. Het effect van het gaafmaken van een enkel record op de uiteindelijke schatting kan uitgedrukt worden als het verschil

$$\delta_{ij} = w_i (y_{ij} - \hat{y}_{ij}) . \tag{4.3.2}$$

De grootheid δ_{ij} bevat de onbekende gecorrigeerde waarde van $\hat{y}_{ij}$ en kan dus niet uitgerekend worden. Daarom wordt $\hat{y}_{ij}$ benaderd door een referentiewaarde. De referentiewaarde dient als een beoordelingsmaatstaf voor de kwaliteit van de ruwe waarde. In de Engelstalige literatuur spreekt men van “anticipated value”. Veel gebruikte bronnen voor referentiewaarden zijn de volgende:

- Gaafgemaakte gegevens van hetzelfde bedrijf uit een eerdere versie van hetzelfde onderzoek eventueel vermenigvuldigd met een schatting van de ontwikkeling tussen de huidige en de vorige waarneming. Bij korte-termijn-statistieken is deze bron veel belangrijker dan bij jaarstatistieken omdat de

overlap tussen de steekproeven voor opvolgende perioden daar veel groter is.

- Gegevens van hetzelfde bedrijf uit een ander onderzoek of een registratie. Bij de productiestatistieken kunnen bijvoorbeeld gegevens uit de Korte-termijn Statistiek gebruikt worden en bij beide kunnen belastinggegevens worden gebruikt.
- Gegevens over een homogene subgroep van gelijkende bedrijven. Bij de productiestatistieken bijvoorbeeld wordt de mediaan van de gaafgemaakte data van een vorige periode in dezelfde bewerkingscel gebruikt. Bewerkingscellen worden gevormd door (samenvoegingen van) SBI-categorieën en grootteklassen.

Behalve de onbekende gecorrigeerde waarde bevat (4.3.2) ook een nog onbekend gewicht w_i. Omdat de gewichten w_i niet alleen corrigeren voor ongelijke insluitkansen maar ook voor non-respons, kunnen ze pas berekend worden wanneer de non-respons bekend is, dus na de periode waarin de data worden verzameld. Het gaafmaken begint echter al tijdens deze periode. Een scorefunctie voor selectief gaafmaken kan daarom geen gebruik maken van deze gewichten. Als oplossing is het gebruikelijk om de gewichten w_i te benaderen door de "begingewichten", zeg v_i, die alleen compenseren voor de ongelijke insluitkansen en al berekend kunnen worden zodra het steekproefontwerp bekend is, namelijk als de inverse van deze insluitkansen. Als vooraf al een schatting gemaakt kan worden van de verwachte non-repons, kan hiervan gebruikgemaakt worden bij het bepalen van de gewichten. Met behulp van de referentiewaarde en de begingewichten kan het effect van het gaafmaken op de schatting van het totaal gekwantificeerd worden met de scorefunctie

$$s_{ij} = v_i \left| y_{ij} - \tilde{y}_{ij} \right| = v_i \tilde{y}_{ij} \times \left| y_{ij} - \tilde{y}_{ij} \right| / \tilde{y}_{ij} = b_{ij} \times r_{ij}, \qquad (4.3.3)$$

met $\tilde{y}_{ij}$ de referentiewaarde. Zoals (4.3.3) laat zien kan deze scorefunctie geschreven worden als het product van een "belangfactor" b_{ij} en een "risicofactor" r_{ij}. De belangfactor is het aandeel van het record i in de totaalschatting gebaseerd op de referentiewaarden en de risicofactor is de absolute waarde van de relatieve afwijking van de geobserveerde waarde ten opzichte van de referentiewaarde. Het risico r_{ij} geeft de verwachte mate van aanpassing door het gaafmaken weer. In plaats van het belang b_{ij} wordt vaak het relatieve belang $b_{ij} / \sum_i b_{ij}$ genomen. Omdat $\sum_i b_{ij} = \sum_i v_i \tilde{y}_{ij} = \tilde{Y}_j$ is de resulterende score $s'_{ij} = s_{ij} / \tilde{Y}_j$ op te vatten als een geschaalde versie van s_{ij}, door te delen door een schatting van het totaal (op basis van de referentiewaarden) wordt de score onafhankelijk van de meeteenheid. Deze schaling maakt de scores voor verschillende variabelen beter vergelijkbaar wat voordelen biedt wanneer lokale scores gecombineerd worden tot een record score (zie paragraaf 4.3.3).

Merk op dat de op deze wijze gedefinieerde score hogere waarden aanneemt naarmate het risico groter is én naarmate het belang groter is, dus naarmate de waarschijnlijkheid van een invloedrijke fout toeneemt en het record eerder in aanmerking komt voor handmatig gaafmaken. Bij de productiestatistieken worden de scores getransformeerd naar scores op een schaal van 1 to 10 waarbij 10 staat voor een zeer plausibele waarde van y_{ij} en 1 voor een zeer implausibele waarde. Na deze transformatie worden de scores plausibiliteitsindicator genoemd.

Een andere bekende scorefunctie wordt verkregen door de risicofactor te baseren op de verhouding tussen de ruwe waarde en de referentiewaarde in plaats van het absolute verschil zoals in (4.3.3). Deze risicofactor, voorgesteld door Hidiroglou en Berthelot (1986) is als volgt gedefinieerd:

$$r_{ij} = \max\left(\frac{\tilde{y}_{ij}}{y_{ij}}, \frac{y_{ij}}{\tilde{y}_{ij}}\right) - 1. \qquad (4.3.4)$$

Door deze definitie worden multiplicatieve afwijkingen van de referentiewaarde naar boven even zwaar meegeteld als multiplicatieve afwijkingen naar beneden en is de minimum waarde 0, voor $y_{ij} = \tilde{y}_{ij}$.

Soms zijn verhoudingen tussen variabelen in een record geschikter om afwijkende waarden op te sporen dan de afzonderlijke variabelen zelf. Voorbeelden hiervan zijn de verhouding van de omzet van een bedrijf en het aantal werknemers of de verhouding van de prijs van een huis en het aantal vierkante meters. De omzet per werknemer en de prijs per vierkante meter vertonen veel minder fluctuatie binnen een bewerkingscel dan de omzet en de prijs. Afwijkende waarden zijn daarom beter te onderscheiden, behalve als de teller en de noemer in dezelfde richting afwijken. Scorefuncties op basis van verhoudingen kunnen verkregen worden door y_{ij} en $\tilde{y}_{ij}$ in de risicofactor in (4.3.3) of (4.3.4) te vervangen door respectievelijk de ruwe waarde en de referentiewaarde van de verhouding.

Voor sommige statistieken zoals de korte-termijnstatistieken zijn de voornaamste doelparameters ontwikkelingen voor (deel)populaties. In deze gevallen is het gebruikelijk om een scorefunctie te kiezen die gericht is op het detecteren van bedrijven met afwijkende ontwikkelingen. De ontwikkeling in een doelvariabele y_j tussen het huidige tijdstip t en een vorig tijdstip t-1, voor een bedrijf i is $\hat{o}_{ij} = \hat{y}_{ij,t} / \hat{y}_{ij,t-1}$. We gaan ervan uit dat de t-1 gegevens al eerder gaafgemaakt zijn en dat het de bedoeling van de te construeren scorefunctie is om alleen de huidige data selectief gaaf te maken. De ruwe waarde van de ontwikkeling is dus $o_{ij} = y_{ij,t} / \hat{y}_{ij,t-1}$. Een risicofactor in een scorefunctie zal afwijkende waarden van de individuele ontwikkelingen o_{ij} proberen op te sporen door deze te vergelijken met een referentiewaarde $\tilde{o}_{ij}$. Hidiroglou en Berthelot (1986) kiezen als referentiewaarde de mediaan van de o_{ij} in een bewerkingscel, wat het nadeel heeft dat pas met gaafmaken begonnen kan worden als er voldoende respons binnen is om die mediaan te kunnen bepalen. Als alternatief kiezen Latouche en Berthelot (1992)

daarom de mediaan van de individuele ontwikkelingen tussen *t*-2 en *t*-1, wat zinvol is als de ontwikkeling tussen *t* en *t-1* lijkt op die tussen *t-1* en *t-2*. Bij de korte-termijnstatistieken op het CBS wordt de referentiewaarde verkregen door eerst een referentiewaarde voor $\hat{y}_{ij,t}$ te bepalen en vervolgens de referentiewaarde voor de ontwikkeling te berekenen als $\tilde{o}_{ij} = \tilde{y}_{ij,t} / \hat{y}_{ij,t-1}$. De referentiewaarde $\tilde{y}_{ij,t}$ wordt bepaald door extrapolatie van $\hat{y}_{ij,t-1}$ met behulp van een op eerdere data geschat seizoenpatroon. Met behulp van een referentiewaarde kan een risicofactor bepaald worden, waarvoor in het geval van ontwikkelingen meestal een multiplicatieve vorm wordt gekozen, zoals (4.3.4).

Een scorefunctie kan nu gevormd worden door r_{ij} te vermenigvuldigen met een belangfactor, Hidiroglou en Berthelot gebruiken hiervoor (de ongewogen versie van)

$$b_{ij} = \left[\max\left(v_{i,t} y_{ij,t}, w_{i,t-1} \hat{y}_{ij,t-1}\right)\right]^c, \qquad (4.3.5)$$

met $0 \leq c \leq 1$. Met de parameter *c* kan de invloed van het belang bepaald worden; de invloed van het belang neemt af bij lagere waarden voor *c*. Op basis van empirisch onderzoek bij Statistics Canada suggereren Latouche en Berthelot om voor *c* de waarde 0,5 te kiezen. De maximum-functie in (4.3.5) heeft tot gevolg dat een fout in de opgave van $y_{ij,t}$ eerder tot een over- dan tot een onderschatting van het belang zal leiden. Een te lage gerapporteerde waarde van $y_{ij,t}$ kan immers nooit resulteren in een belang kleiner dan $\hat{y}_{ij,t-1}$, terwijl een te hoge waarde van $y_{ij,t}$ het belang in principe ongelimiteerd kan verhogen. Een geschaalde versie van een score met belangfactor volgens (4.3.5) kan verkregen worden door $v_{i,t} y_{ij,t}$ en $w_{i,t-1} \hat{y}_{ij,t-1}$ in (4.3.5) te delen door schattingen voor hun totaal, respectievelijk $\tilde{Y}_{j,t}$ en $\hat{Y}_{j,t-1}$. Het geschatte *t*-1 totaal is eenvoudig $\hat{Y}_{j,t-1} = \sum_i w_{i,t-1} \hat{y}_{ij,t-1}$. Omdat er van uitgegaan wordt dat alle data nog niet beschikbaar zijn moet het actuele totaal benaderd worden, bijvoorbeeld door een schatting op basis van referentiewaarden en begingewichten: $\tilde{\tilde{Y}}_{j,t} = \sum_i v_{i,t} \tilde{y}_{ij,t}$. De geschaalde versie van de belangfactor kan geschreven worden als

$$b'_{ij} = \left[\max\left(\frac{v_{i,t} y_{ij,t}}{\tilde{\tilde{Y}}_{j,t}}, \frac{w_{i,t-1} \hat{y}_{ij,t-1}}{\hat{Y}_{j,t-1}} \right) \right]^c . \qquad (4.3.6)$$

4.3.3 Combineren van lokale scores tot een globale score

Om een record al of niet te selecteren voor het handmatig gaafmaken is een score op record niveau nodig. Deze globale score combineert de informatie van de lokale scores over de verwachte invloed van het gaafmaken van verschillende variabelen in het record tot één score die het belang van het handmatig gaafmaken voor het hele record aangeeft.

Bij het combineren van de scores is het van belang dat de orde van grootte van de scores vergelijkbaar is omdat anders de verschillende scores onbedoeld een verschillend gewicht krijgen in de globale score. Het is daarom gebruikelijk om de scores te schalen zodat ze beter vergelijkbaar worden. Eén methode hiervoor is in de vorige paragraaf beschreven. Een andere methode is het delen van de score door de standaardafwijking van de referentiewaarden, $(s_{ij} / \sigma(\tilde{y}_j))$, zie Lawrence en McKenzie (2000). Dit laatste heeft het voordeel dat afwijkingen in variabelen met een grote spreiding minder hoge scores krijgen, en dus minder snel als verdacht worden aangemerkt, dan afwijkingen in variabelen met een kleinere spreiding.

Er zijn verschillende methoden voorgesteld om de gestandaardiseerde scores te combineren tot een globale score. Vaak wordt de som van de lokale scores genomen (Latouche en Berthelot, 1992). Records met veel afwijkende waarden krijgen hierdoor hoge scores en dus hoge prioriteit voor het handmatig gaafmaken. Dit is een voordeel omdat het gaafmaken van meerdere variabelen in hetzelfde record relatief minder werk is dan het gaafmaken van een enkele variabele in een record, zeker als daarbij opnieuw contact met de berichtgever wordt opgenomen. De methode heeft tot gevolg dat records met veel, maar minder sterk afwijkende waarden eerder handmatig gaaf gemaakt zullen worden dan records met slechts enkele maar wel sterk afwijkende waarden. Als het gewenst is dat een sterk afwijkende waarde voor een enkele variabele in een overigens niet-verdacht record toch handmatig gaafgemaakt wordt is de som van de lokale scores geen goed criterium.

Als alternatief voor de som van de lokale scores stellen Lawrence en McKenzie, (2000) voor om het maximum te nemen van de geschaalde scores. Het voordeel hiervan is dat de methode voor iedere variabele garandeert dat afwijkende waarden boven een bepaalde grens handmatig geïnspecteerd zullen worden. Het nadeel van deze veilige strategie is dat er geen onderscheid meer gemaakt wordt tussen records met een enkele ernstige afwijking en records met veel even ernstige afwijkingen. Als compromis tussen de som en het maximum stelt Farwell (2005) voor om de Euclidische metriek te gebruiken. Deze drie voorstellen zijn te generaliseren tot het combineren van de lokale scores tot een globale score volgens de zogenaamde Minkowski metriek (zie Hedlin, 2008) gegeven door

$$S_i^{(\alpha)} = \left(\sum_{j=1}^{J} s_{ij}^{\alpha} \right)^{1/\alpha}, \qquad (4.3.7)$$

met J het aantal lokale scores. De parameter α in (4.3.7) bepaalt de invloed van grote waarden van de lokale scores op de globale score, deze invloed neemt toe met α. Voor $\alpha=1$ is (4.3.7) de som van de lokale scores en voor $\alpha=\infty$ is (4.3.7) gelijk aan het maximum van de lokale scores, alleen de grootste waarde telt dan nog mee. Voor $\alpha=2$, is (4.3.7) de Euclidische metriek.

Bij uitgebreide vragenlijsten zoals bij de productiestatistieken zijn niet alle variabelen even belangrijk. Totalen van omzet en aantallen werknemers zijn veel belangrijker dan uitsplitsingen van onderdelen van de bedrijfslasten. In zulke gevallen worden aan de lokale scores binnen de sommatie in (4.3.7) nog gewichten

toegekend die verschillen in belang uitdrukken. Bij de productiestatistieken bijvoorbeeld worden door inhoudelijk deskundigen gewichten toegekend waarbij de keuze is tussen 0,1,10 of 100.

In formule (4.3.7) worden schatter-gerelateerde scorefuncties gecombineerd. Daarnaast is er soms echter ook nog sprake van edit-gerelateerde scorefuncties, zoals het aantal harde fouten of het aantal ten onrechte lege velden (partiële non-respons) die ook iets zeggen over de kwaliteit van een record. De edit-gerelateerde scores moeten nog toegevoegd worden aan de globale score. Dit kan door ze op dezelfde manier te behandelen als de schatter-gerelateerde scores en ze, na een geschikte schaling en eventueel met hun eigen gewichten, toe te voegen aan de sommatie in (4.3.7). Een andere mogelijkheid is om de edit-grelateerde scores eerst te combineren met hun eigen metriek en deze gecombineerde scores op te tellen bij de gecombineerde schatter-gerelateerde scores (zie paragraaf 4.4.2 voor een voorbeeld).

4.3.4 Bepalen van grenswaarde voor de globale score en pseudo-bias

Het uiteindelijke doel van een globale scorefunctie is het selecteren van de records voor het handmatig gaafmaken. Als het gaafmaken kan wachten tot na de waarnemingsperiode dan kan handmatig gaafgemaakt worden volgens de prioritering van de globale score totdat schattingen van de belangrijkste doelparameters hierdoor niet meer substantieel veranderen (vergelijk figuur 3). Aangezien handmatig gaafmaken tijdrovend is leidt deze aanpak, vooral bij statistieken met grotere hoeveelheden data en variabelen, tot onacceptabel lange doorlooptijden. Om met handmatig gaafmaken te kunnen starten tijdens de dataverzamelingsfase is het nodig om op basis van de score per record, zonder vergelijking met de scores van de andere records, een beslissing te nemen om het record al dan niet handmatig gaaf te maken. Met dit doel wordt er een grenswaarde voor de record score vastgesteld zodanig dat records met een score hoger dan de grenswaarde handmatig gaafgemaakt worden en records met een score lager dan de grenswaarde automatisch of helemaal niet worden gaafgemaakt.

De gebruikelijke methode om een grenswaarde te bepalen is via een simulatiestudie waarin het effect van verschillende grenswaarden, en dus verschillende mate van handmatig gaafmaken, op de vertekening in de belangrijkste doelparameters onderzocht wordt. Zo'n simulatie is gebaseerd op een verzameling ruwe data en de daarbij behorende volledig handmatig gaafgemaakte data. Deze data moeten vergelijkbaar zijn met de data waarop de grenswaarden toegepast gaan worden. De gebruikelijke keuze hiervoor zijn de data van een eerdere versie van het onderzoek.

Voor de simulatiestudie worden nu eerst volgens de gekozen methoden globale scores berekend voor de records met ruwe data en vervolgens worden deze records geordend volgens deze scores. Daarna wordt gesimuleerd dat alleen de eerste p% van de records geselecteerd zijn voor handmatig gaafmaken. Dit wordt gedaan door voor die eerste p% van het bestand de ruwe waarden te vervangen door de

gaafgemaakte waarden. Het deelbestand met de gaafgemaakte records geven we aan met H_p.

Vervolgens wordt het verschil bepaald tussen de schatting van het totaal van een variabele op basis van het *p%*-gaafgemaakte bestand en op basis van het volledig gaafgemaakte bestand. De absolute waarde van het relatieve verschil tussen deze schattingen wordt de absolute pseudo-bias genoemd (Latouche en Berthelot, 1992), gegeven door

$$D_j(p) = \frac{1}{\hat{Y}_j}\left|\sum_{i \in H_p} w_i(\hat{y}_{ij} - y_{ij})\right| \quad . \tag{4.3.8}$$

Zoals (4.3.8) laat zien wordt de absolute pseudo-bias bepaald door het verschil in de totalen van de gaafgemaakte waarden en de niet gaafgemaakte waarden voor het niet voor handmatig gaafmaken geselecteerde deel van de records. Als het gaafmaken tot gevolg heeft dat alle fouten (en alléén fouten) gecorrigeerd worden dan is (4.3.8) de relatieve vertekening (bias) die ontstaat doordat niet alle records gaafgemaakt worden. Omdat het niet zeker is dat het gaafmaken de werkelijke waarden reproduceert, is (4.3.8) een benadering van deze vertekening en wordt daarom dan ook pseudo-bias genoemd.

De pseudo-bias bij *p%*-gaafmaken kan ook gezien worden als een schatting van de winst in nauwkeurigheid die te bereiken is door de overige (1-*p*%) van de records ook gaaf te maken. Door de pseudo-bias te berekenen voor een groot aantal verschillende waarden van *p* wordt een beeld verkregen van de winst in nauwkeurigheid als functie van *p*. Als de sortering van de records op grond van de score het gewenste effect heeft, zal deze winst afnemen naarmate *p* toeneemt. Bij een zekere waarde voor *p* zal men besluiten dat de overgebleven psuedo-bias klein genoeg is en dat het niet loont om meer records gaaf te maken. De met deze waarde voor *p* corresponderende record score wordt dan de grenswaarde.

De pseudo-bias zoals hierboven beschreven, is gebaseerd op een vergelijking tussen handmatig gaafgemaakte data en ruwe data en gaat er dus van uit dat er of handmatig wordt gaafgemaakt of helemaal niet wordt gaafgemaakt. In veel gevallen wordt echter automatisch gaafgemaakt in plaats van helemaal niet gaafgemaakt. Als we er van uitgaan dat het automatisch gaafmaken in ieder geval niet tot meer vertekening leidt dan helemaal niet gaafmaken, kan de bovenstaande pseudo-bias opgevat worden als een bovengrens voor de pseudo-bias in situaties waarin automatisch gaafmaken wordt toegepast.

4.4 Voorbeelden

In deze paragraaf komen enkele praktijkvoorbeelden aan bod van (onderdelen van) het selectief gaafmaken. We behandelen eerst de constructie van een plausibiliteitsindex voor de korte-termijnstatistieken (§ 4.4.1) en geven vervolgens een korte beschrijving van de plausibiliteitsindex van de statistiek Bouwobjecten In Voorbereiding (§ 4.4.2). Deze praktijkvoorbeelden dienen slechts ter illustratie van de technieken en daarom wordt niet op alle details van de implementatie ingegaan.

Uitgebreidere beschrijvingen zijn te vinden in Van Duin, 2003 (plausibiliteitsindex voor de korte-termijnstatistieken), en Van der Loo en Pannekoek, 2007 (plausibiliteitsindex van de statistiek Bouwobjecten In Voorbereiding).

4.4.1 Scorefunctie voor de Korte-termijn Statistieken

De belangrijkste variabele bij de Korte-termijn Statistieken (KS'en) is de omzet en de belangrijkste doelparameter is de ontwikkeling in die omzet tussen opeenvolgende perioden (maanden of kwartalen). In het standaard productieproces voor deze statistieken (IMPECT 2) wordt selectief gaafgemaakt waarbij de selectie uitsluitend bepaald wordt op basis van de variabele omzet.

Het selectieproces maakt gebruik van varianten van de belang- en risicofactoren die in paragraaf (4.3.2) zijn besproken. De risicofactor is de verhouding van de waargenomen omzet ontwikkeling tussen tijdstippen *t* en *t-1* en een referentiewaarde voor deze ontwikkeling:

$$r_{i,t} = \frac{o_{i,t}}{\tilde{o}_t}, \text{ met } o_{i,t} = y_{i,t} / \hat{y}_{i,t}.$$

Merk op dat bij deze definitie van een risicofactor, in tegenstelling tot die volgens (4.3.4), zowel waarden veel groter dan 1 als waarden veel kleiner dan 1, "verdacht" zijn.

Om de referentiewaarde te bepalen wordt voor een aantal jaren uit het verleden de mediaan berekend van de omzet ontwikkeling tussen *t-1* en *t*. Voor een maandstatistiek zijn *t* en *t-1* steeds dezelfde maanden maar uit verschillende jaren en voor een kwartaal statistiek slaan *t* en *t-1* steeds op dezelfde kwartalen, uit verschillende jaren. Vervolgens wordt het meetkundig gemiddelde genomen, over de jaren, van deze ontwikkelingen uit het verleden. De referentiewaarde wordt bepaald door dit meetkundig gemiddelde te vermenigvuldigen met een correctiefactor voor het verschil in het aantal werkdagen op *t-1* en *t*.

De belangfactor is het geschaalde belang volgens (4.3.6), dus het maximum van de bijdrage aan het *t-1*-totaal en een schatting voor de bijdrage aan het actuele totaal. In plaats van *t-1* wordt hier echter geleken naar *t*-2, dus

$$b_{i,t} = \max\left(\frac{v_{i,t}, y_{i,t}}{\tilde{Y}_t}, \frac{w_{i,t-2}, \hat{y}_{i,t-2}}{\hat{Y}_{t-2}} \right).$$

Er wordt gekeken naar *t*-2 omdat er voor de periode *t*-1 mogelijk nog geen gefiatteerd totaal beschikbaar is. De benadering voor het actuele totaal $\tilde{Y}_t$ wordt verkregen door het geschatte totaal op *t-2* te vermenigvuldigen met de geschatte omzetontwikkeling volgens de referentiewaarde zoals hierboven bepaald.

Met behulp van de bovenstaande risico- en belangfactoren $r_{i,t}$ en $b_{i,t}$ worden records geselecteerd voor interactief gaafmaken. De strategie die hierbij gevolgd wordt is anders dan de in §4.3.1 beschreven strategie. In plaats van het combineren van de risico en belangfactoren tot een score en vervolgens de selectie toepassen aan

de hand van een grenswaarde voor deze score, worden hier aparte grenswaarden voor de belang en de risicofactoren gehanteerd. Dit selectieproces laat zich in de volgende twee stappen samenvatten:

1. Als $b_{i,t} > b_{\min 1}$ dan wordt interactief gaafgemaakt, onafhankelijk van de waarde van $r_{i,t}$. De waarde van $b_{\min 1}$ is zodanig gekozen dat hiermee een klein aantal zeer belangrijke bedrijven wordt geselecteerd waarvoor het de moeite loont om ze altijd door een gaafmaker te laten controleren. Bij de KS'en gaat het maar om een paar variabelen, dus een record is snel te controleren.

2. Voor de overige records geldt:

Alleen als $b_{i,t} > b_{\min 2}$ en ($r_{i,t} < r_{\min}$ of $r_{i,t} > r_{\max}$) dan wordt interactief gaafgemaakt. Merk op dat zowel $r_{i,t} < r_{\min}$ als $r_{i,t} > r_{\max}$ duidt op een groot risico.

In de tweede stap geldt dus net als bij de eerder besproken aanpak dat als het risico én het belang groot is dan wordt er interactief gaafgemaakt. In tegenstelling tot de eerder besproken aanpak zijn er hier echter minder "compensatiemogelijkheden". Bij de scorefunctie benadering kan een record met een klein risico toch een hoge score krijgen en daarmee geselecteerd worden voor interactief gaafmaken, als het belang maar groot genoeg is. Dat is met de hier toegepaste benadering niet mogelijk; als het risico in het interval [$r_{\min}$, $r_{\max}$] ligt is het record plausibel en wordt het niet interactief gaafgemaakt. Eveneens geldt dat als $b_{i,t} < b_{\min 2}$ dan is het record onbelangrijk en wordt het niet interactief gaafgemaakt, hoe groot het risico ook is.

4.4.2 Plausibiliteitsindex voor statistiek Bouwobjecten In Voorbereiding

De kwartaalstatistiek Bouwobjecten In Voorbereiding (BIV) volgt de ontwikkeling van de totale bouwwaarde van nieuwe contracten bij architectenbureaus in Nederland en wordt gebruikt als een snelle indicator voor ontwikkelingen in de bouw (zie ook paragraaf 2.4.1 waar de deductieve correcties binnen het gaafmaakproces van deze statistiek zijn besproken). Het budget van zo'n contract is de centrale variabele voor deze statistiek. Hiermee worden schattingen gemaakt van het totale budget voor bouwobjecten in de klassen gedefinieerd door de combinaties van type bouwwerk (woning, niet-woning, combinatiegebouw) en soort werk (nieuwbouw, renovatie). Het is daarom van primair belang om fouten in het opgegeven budget te corrigeren. De budgetten van de bouwobjecten vertonen een zeer grote spreiding waardoor het moeilijk is om afwijkende waarden te detecteren. Het budget per vierkante meter vertoont echter een veel geringere spreiding. Sterk afwijkende waarden voor het budget per vierkante meter zijn daarom een indicatie voor mogelijk foutieve opgaven van het budget (of het aantal vierkante meters). Daarom is gekozen voor een risicofactor die gebaseerd is op de afwijking van het budget per vierkante meter van een bouwobject ten opzichte van de mediaan voor de betreffende klasse. Als we voor een bouwwerk *i* in klasse *k* het budget weergeven met b_{ki} en de oppervlakte in vierkante meters met o_{ki} dan kan de vierkante meter

prijs gedefinieerd worden als $x_{ki} = b_{ki}/o_{ki}$. De risicofactor kan dan geschreven worden als

$$r_{ik} = \max\left(\frac{\tilde{x}_k}{x_{ki}}, \frac{x_{ki}}{\tilde{x}_k}\right),$$

met $\tilde{x}_k$ de mediaan van de vierkante meter prijs van de bouwobjecten in klasse *k*. Dit is een risicofactor van de vorm (4.3.4) (alleen de constante -1 is hier weggelaten wat geen gevolgen heeft voor de ordening van de records naar hun risico).

Een belangfactor moet het belang van respondent *i* voor de schatter van het totaal van de doelvariabele representeren. Het totale budget per klasse, Y_k, wordt geschat met

$$\hat{Y}_k = \sum_i w_{ki} y_{ki} = \sum_i w_{ki} o_{ki} x_{ki},$$

met w_{ki} het ophooggewicht. Het belang van een record voor de schatter van het totale budget kan daarom uitgedrukt worden in de belangfactor $b_{ki} = w_{ki} o_{ki}$ en de risico- en belangfactoren kunnen gecombineerd worden tot de scorefunctie

$$s_{ki}^{(Y)} = b_{ki} r_{ki}.$$

Deze scorefunctie per klasse is geschaald door te delen door het maximum per klasse.

Behalve naar invloedrijke verdachte waarden van het budget per vierkante meter, wordt bij het selectief gaafmaken van de BIV ook gekeken naar harde fouten. Dit kunnen waarden zijn die buiten het toegelaten waardebereik liggen (zoals percentages woonoppervlak die niet tussen 0 en 100 liggen) of ontbrekende waarden op de variabelen Budget of Aantal woningen. Het kunnen ook conflicterende waarden zijn, zoals het opgeven van meerdere objecttypen voor één bouwobject of het invullen van een percentage (<100%) woonoppervlak voor een bouwobject dat een woning is. In totaal zijn er 13 typen harde fouten gedefinieerd. Het aantal harde fouten is een kenmerk van de kwaliteit van het record en hiervoor is ook een scorefunctie gedefinieerd:

$$s_{ki}^{(E)} = \sum_{j=1}^{J} g_j^{(E)} E_j,$$

waarin E_j =1 als er een harde fout van type E_j is opgetreden en 0 anderszins. Met de gewichten $g_j^{(E)}$ kan het relatieve belang van de verschillende harde fouten ingesteld worden. De gewichten zijn zodanig geschaald dat $\sum_j g_j^{(E)} = 1$. Hierdoor ligt de score voor harde fouten tussen 0 en 1. Door de twee scorefuncties gewogen te combineren ontstaat de uiteindelijk toegepaste scorefunctie

$$s_{ki} = g_1 s_{ki}^{(E)} + g_2 s_{ki}^{(Y)}.$$

met $g_1 + g_2 = 1$.

Deze scorefunctie is een schatter-gerelateerde score $s_{ki}^{(Y)}$ gecombineerd met een edit-gerelateerde score $s_{ki}^{(E)}$, waarbij het relatieve belang van deze twee componenten wordt weergegeven door, respectievelijk, de gewichten g_2 en g_1.

5. Foutlocalisatie op basis van het principe van Fellegi en Holt

5.1 Korte beschrijving

Met deze methode wordt een gegevensbestand record voor record gecontroleerd aan de hand van voorgedefiniëerde controleregels. Wanneer een record één of meer controleregels schendt, levert de methode een aantal velden op die zodanig kunnen worden geïmputeerd, dat geen enkele regel meer geschonden wordt. Het imputeren zelf is geen onderdeel van de methode.

Bij het selecteren van de velden wordt uitgegaan van het (gegeneraliseerde) principe van Fellegi en Holt. Dat wil zeggen, dat het kleinste (gewogen) aantal velden wordt gekozen waarmee het record consistent kan worden geïmputeerd. Het aanwijzen van de te imputeren velden wordt foutlocalisatie genoemd, en kan geautomatiseerd worden uitgevoerd. De controleregels kunnen zowel rekenkundig (zoals het controleren van optellingen) als logisch van aard zijn (bijvoorbeeld: als *geslacht*=man dan *zwanger*=nee). Combinaties zijn ook mogelijk.

Op het CBS is software ontwikkeld waarmee deze geautomatiseerde foutlocalisatie kan worden uitgevoerd, in de vorm van SLICE/CherryPie.

5.2 Toepasbaarheid

Deze methode is bedoeld om in een record de fout ingevulde velden op te sporen. De methode kan worden toegepast op gegevensbestanden die numerieke, categoriale of beide gegevenstypen bevatten. Voor numerieke gegevens geldt dat controleregels uit lineaire relaties tussen de variabelen moeten bestaan (zie 5.3.1). Voor categoriale gegevens kunnen willekeurige relaties tussen variabelen worden vastgelegd. Controleregels moeten per record gecontroleerd kunnen worden. Bijvoorbeeld, een controleregel waarbij de waarde in een veld wordt vergeleken met de gemiddelde waarde voor dat veld over het hele bestand is geen geldige controleregel. Dit betekent wel dat deze foutlocalisatiemethode al kan worden toegepast voordat alle data binnen zijn.

Het (gegeneraliseerde) principe van Fellegi en Holt, kan voor elke enquête worden ingezet, hoewel het niet voor alle typen fouten geschikt is. Bij sommige inconsistenties, zoals eenheidsfouten (bijvoorbeeld duizendfouten), tekenverwisselingen en kolomverwisselingen kan beter deductieve correctie worden toegepast, zoals beschreven in hoofdstuk 2. Het belangrijkste verschil tussen deductieve correctie en de methode die hier beschreven wordt, is dat bij deductieve correctie wel gebruik wordt gemaakt van de opgegeven waarden in velden om fouten te lokaliseren, en bij de huidige methode niet. Wanneer de waarde in een veld een aanwijzing kan geven over de opgetreden fout (en daarmee de oplossing) kan beter deductieve correctie worden toegepast. Naast de hiervoor gegeven

voorbeelden vallen daar bijvoorbeeld nog cijferverwisselingen en kommafouten onder.

Bij foutlocalisatie volgens het (gegeneraliseerde) principe van Fellegi en Holt wordt geen onderscheid gemaakt tussen zogenaamde harde en zachte controleregels: alle regels worden als harde controleregel behandeld. Met harde controleregels worden regels bedoeld die door rekenkundige of logische verbanden worden vastgelegd, zoals $omzet = winst + kosten$. Harde controleregels definiëren waardecombinaties die *zeker* fout zijn. Zachte controleregels geven aan of een waarde, of waardecombinatie *onwaarschijnlijk* is, zoals $kosten / omzet \geq 0.6$. Bij de foutlocalisatie worden alle records die één of meer controleregels schenden als zeker fout gezien. Wanneer teveel zachte controleregels worden gedefinieerd, bestaat het gevaar op *over-editing*: het onterecht aanpassen van juist ingevulde gegevens. Zie bijvoorbeeld Di Zio e.a. (2005).

5.3 Uitgebreide beschrijving

De hier gegeven beschrijving is voornamelijk gericht op geautomatiseerde foutlocalisatie, waarbij voor elk record *on-the-fly* een veldkeuze wordt bepaald. Bij enquêtes met weinig vragen kan op basis van het (gegeneraliseerde) principe van Fellegi en Holt met de hand, per combinatie van schendingen, een veldkeuze worden vastgelegd. Dit is bijvoorbeeld gebeurd bij het gaafmaken van de statistiek Bouwobjecten in Voorbereiding (Van der Loo en Pannekoek, 2007), waarin 5 variabelen een rol spelen. Wanneer het aantal variabelen en relaties daartussen toeneemt, neemt de omvang van het foutlocalisatieprobleem snel toe. Daarom is op het CBS software voor foutlocalisatie ontwikkeld in de vorm van SLICE (De Waal, 2005a). Met SLICE kunnen grote en complexe enquêtes worden verwerkt. SLICE wordt bijvoorbeeld toegepast bij het verwerken van gegevens in de ProductieStatistiek (De Jong, 2002).

Dit hoofdstuk is als volgt ingedeeld. In paragraaf 5.3.1 beschrijven we de formulering van records en controleregels. In paragraaf 5.3.2 geven we een overzicht van de eigenschappen van het foutlocalisatieprobleem en de oplossing ervan. Om het foutlocalisatieprobleem op te lossen is het nodig om uit expliciet gedefinieerde controleregels de regels die daar logisch uit volgen (automatisch) af te kunnen leiden. De technieken hiervoor worden beschreven in paragraaf 5.3.3. Deze (vrij technische) paragraaf kan eventueel worden overgeslagen bij eerste lezing. De volgende paragraaf (5.3.4) is gewijd aan de oplossing zoals die bij het CBS is geïmplementeerd: het *branch-and-bound* algoritme. Deze paragraaf is weer vrij technisch en kan eventueel worden overgeslagen. Tot slot wordt ingegaan op de door het CBS ontwikkelde software SLICE/CherryPie, waarmee foutlocalisatieproblemen kunnen worden opgelost.

5.3.1 Records en controleregels

Een record is een rijtje van velden of variabelen uit een vragenlijst. Een record x kan worden weergeven als $x = (x_1, x_2, \ldots, x_n)$. De waarden die door variabele x_i

aangenomen kunnen worden wordt het *domein* D_i genoemd. Voorbeelden zijn x_i =geslacht met $D_i = \{man, vrouw\}$, x_i =aantal woningen met $D_i = \mathrm{N}$ of x_i =winst met D_i = R. Het totale domein D, waarbinnen alle mogelijke records liggen kan geschreven worden als $D = (D_1, D_2, \ldots, D_n)$.

Controleregels geven aan waaraan variabelen of variabelecombinaties in een gegevensbestand per regel moeten voldoen. Controleregels worden vaak gaafmaakregels of (naar het Engels) *edits* of *edit checks* genoemd. Alle controleregels moeten per record controleerbaar zijn, en mogen dus niet afhangen van waarden in velden van andere records.

De typen controleregels die kunnen worden gebruikt, kunnen worden onderscheiden op grond van de typen gegevens waarop zij betrekking hebben:

- **Numerieke gegevens.** Kunnen worden gecontroleerd op basis van lineaire relaties zoals $omzet \geq 0$, of $winst + kosten = omzet$. De algemene vorm van een lineaire controleregel luidt:

$$\sum_{i=1}^{n} a_{ji} x_i \geq b_j \quad \text{of} \quad \sum_{i=1}^{n} a_{ji} x_i = b_j \ ,$$

  waarbij j de controleregels nummert, a_{ji} lineaire coëfficiënten zijn en b_j constanten. Merk op dat regels zoals $kosten / omzet \geq 0.6$ ook lineaire controleregels zijn, omdat ze in de vorm $kosten - 0.6 \cdot omzet \geq 0$ kunnen worden geschreven.

- **Categoriale gegevens.** Willekeurige combinaties van categoriale gegevens kunnen worden uitgesloten. De regels worden vaak geschreven in als-dan vorm, bijvoorbeeld: **als** *geslacht* = man **dan** *zwanger* = nee.

- **Combinaties van beide.** Deze worden ook in als-dan vorm geschreven, bijvoorbeeld: **als** *burg. staat.*=getrouwd **dan** $leeftijd \geq 16$.

Wanneer een controleregel e expliciet betrekking heeft op variabele x_i, zeggen we dat x_i *voorkomt* in e. Omgekeerd zeggen we e *bevat* x_i. Merk op dat een controleregel een deelverzameling van alle mogelijke records in D vastlegt, waarvoor geldt dat alle records in die deelverzameling minimaal één fout bevatten (zie ook paragraaf 5.3.3).

Om het foutlocalisatieprobleem op te lossen, is het van belang dat niet alleen met de voorgedefinieerde regels, maar ook met de regels die daar logisch uit volgen rekening wordt gehouden. Regels die door de gebruiker zijn gedefinieerd worden *expliciete controleregels* genoemd, regels die daaruit worden afgeleid heten *impliciete controleregels.* Bijvoorbeeld, gegeven de controleregels $x_1 > x_2$ en $x_2 > x_3$, dan volgt daar de impliciete regel $x_1 > x_3$ uit. Het is niet nodig (en voor lineaire regels zelfs onmogelijk) om alle impliciete controleregels te genereren. Fellegi en Holt (1976) bewezen dat het voldoende is om de zogenaamde *essentieel nieuwe* controleregels af te leiden (zie paragraaf 5.3.3).

5.3.2 Foutlocalisatie

Foutlocalisatie is het aanwijzen van één of meerdere velden in een record, zodanig dat door aanpassen van de inhoud van die velden het record geen controleregel meer schendt. Het is van belang te beseffen dat het niet zeker is dat de werkelijk (door de respondent) gemaakte fout wordt gevonden. Voor het aanwijzen van de "foute" waarden wordt altijd een aanname gemaakt. Het gegeneraliseerde principe van Fellegi en Holt is gebaseerd op zo'n aanname, en kan als volgt worden samengevat: wanneer een record x één of meer controleregels schendt, wordt gezocht naar de verzamelingen van velden G die voldoen aan:

(G1) De inhoud van de velden $g \in G$ kan worden aangepast, zodanig dat record x geen enkele expliciete of essentieel nieuwe controleregel meer schendt.

(G2) De waarde van $\sum_{g \in G} w(g)$ wordt geminimaliseerd.

Hierbij zijn $w(g)$ betrouwbaarheidsgewichten voor de velden g in G. Een hogere waarde voor $w(g)$ betekent dat veld g geacht wordt beter te zijn ingevuld. Het testen van de betrouwbaarheidsgewichten is niet mogelijk binnen de foutlocalisatiemethode. De geldigheid van de gekozen betrouwbaarheidsgewichten moet dus apart worden onderzocht. Zie bijvoorbeeld Hoogland en Smit (2008).

Een speciaal geval is wanneer voor alle gewichten $w(g)$ dezelfde waarde wordt gekozen, bijvoorbeeld $w(g) = 1$ voor alle g. In dat geval houden we foutlocalisatie op basis van het (oorspronkelijke) principe van Fellegi en Holt over. De verzameling G bestaat dan uit de kleinste verzameling velden waarmee het record consistent geïmputeerd kan worden. In dat geval kan worden bewezen (Fellegi en Holt, 1976) dat G gegeven wordt door de kleinst mogelijke verzameling van variabelen die alle geschonden expliciete en essentieel nieuwe controleregels overdekt. Dat wil zeggen, de kleinste verzameling van velden waarvoor geldt dat elk veld in minstens één van de geschonden expliciete en essentieel nieuwe regels voorkomt. De aanname hierachter is dat fouten toevallig worden gemaakt, en dat de grootste verzameling consistent ingevulde velden naar waarheid zijn ingevuld.

De laatste mogelijkheid is om geen extra aanname te doen over de beste oplossing voor het foutlocalisatieprobleem, en een willekeurige oplossing te kiezen uit alle verzamelingen van velden die aan eis (G1) voldoen.

Zelfs wanneer het gegeneraliseerde principe van Fellegi en Holt wordt toegepast, kan het foutlocalisatieprobleem meerdere oplossingen hebben. Om in dat geval een unieke oplossing te genereren kan daarom gebruik worden gemaakt van een (hiërarchische) combinatie van selectieprincipes. Bijvoorbeeld: (1) genereer de oplossingen $G_1, G_2, \ldots, G_m$ volgens het gegeneraliseerde principe van Fellegi en Holt. (2) wanneer de oplossing niet uniek is $(m > 1)$, kies dan een willekeurige oplossing uit de m mogelijkheden. Een andere methode kan zijn: (1) genereer de oplossingen $G_1, G_2, \ldots, G_m$ volgens het gegeneraliseerde principe van Fellegi en Holt. (2) wanneer de oplossing niet uniek is $(m > 1)$, kies dan de oplossing met het

kleinste aantal velden. (3) wanneer de oplossing dan nog niet uniek is, kies er dan willekeurig één uit de overgebleven oplossingen. Zie ook Stoop (2003) voor meer selectiemechanismen.

Er bestaan verschillende algoritmen om, uitgaande van het gegeneraliseerde principe van Fellegi en Holt, de mogelijke oplossingen $G_1, G_2, \ldots, G_m$ te vinden. De methode die op het CBS is geïmplementeerd in CherryPie (een onderdeel van SLICE), is gebaseerd op het zogenaamde *branch-and-bound* algoritme (De Waal, 2003; 2008). In het kort komt dit algoritme er op neer dat voor elke relevante combinatie van velden wordt getest of het aan eis (G1) kan voldoen. De relevante combinaties worden afgelopen met behulp van een zogenaamde binaire boom. Vervolgens kan op basis van (een combinatie van) de selectieprincipes een keuze uit de mogelijke oplossingen worden gemaakt. In paragraaf 5.3.4 wordt dit algoritme beschreven. Het algoritme neemt als invoer een record en een verzameling controleregels. De uitvoer bestaat uit een verzameling velden die consistent kunnen worden geïmputeerd. De complexiteit (de mate waarin de looptijd van het algoritme toeneemt met de invoer) van het *branch-and-bound* algoritme is vrij hoog. Ten eerste heeft het construeren van de boom een asymptotische (maximale) complexiteit van $O(2^n)$ in het aantal variabelen. Dat wil zeggen: elke extra variabele die in een controleregel voorkomt kan de looptijd verdubbelen. Tijdens het construeren van de boom moeten in elke stap variabelen uit controleregels worden geëlimineerd. Dit is voor categoriale variabelen een probleem met complexiteit $O(2^{k_s})$, met k_s het aantal controleregels dat x_s bevat. De precieze loopduur neemt dus snel toe met het aantal variabelen en het aantal controleregels. Omdat de looptijd zo snel toeneemt, is bij de implementatie van deze methode voor de productiestatistiek gekozen voor een tijdslimiet van enkele minuten. Records waarvoor na die tijd nog geen oplossing is gevonden, worden dan handmatig gaafgemaakt.

Er kunnen wel maatregelen getroffen worden om de looptijd te bekorten, namelijk door n en/of k zo klein mogelijk te houden. Ten eerste kan een gegevensbestand worden voorbewerkt, zodanig dat zoveel mogelijk deductieve correcties al zijn toegepast. Het *branch-and-bound* algoritme loopt sneller af wanneer minder regels zijn geschonden. Ten tweede kunnen de kolommen van bestanden die niet aan elkaar gerelateerd zijn door correctieregels als aparte blokken worden behandeld (verkleinen van n). Ten derde wordt in SLICE de binaire boom op een slimme manier opgebouwd: de volgorde van opbouwen is zo gekozen dat oplossingen (meestal) snel gevonden worden en takken die geen, of geen betere dan eerder gevonden oplossingen kunnen opleveren worden afgebroken (De Waal, 2005b; Daalmans, 2000). Tot slot kan bij elektronische waarneming eventueel rekening gehouden worden met het gaafmaakproces door het inbouwen van gaafmaakregels in de vragenlijst. Door harde controleregels in te bouwen in bijvoorbeeld webformulieren, wordt het aantal geschonden controleregels voor het latere gaafmaakproces verminderd. Door een slimme keuze van de in te bouwen controleregels kan het aantal variabelen waarmee het *branch-and-bound* algoritme rekening moet houden worden verminderd. (zie ook Van der Loo, 2008).

5.3.3 Elimineren van variabelen

Het is mogelijk om door logische of rekenkundige bewerkingen impliciete controleregels te genereren uit een gegeven aantal expliciete regels. Bekijk bijvoorbeeld de volgende twee lineaire controleregels:

$e_1 : kosten + winst - omzet = 0$
$e_2 : kosten - 0.6 \cdot omzet \geq 0.$

Door kosten op te lossen uit e_1, en in te vullen in e_2, krijgen we

$e_3 : 0.4 \cdot omzet - winst \geq 0$.

De nieuwe regel e_3 bevat de variabele *kosten* niet, terwijl e_1 en e_2 dat wel doen. We zeggen dat de variabele *kosten* is geëlimineerd. De algemene procedure om lineaire controleregels af te leiden heet Fourier-Motzkin eliminatie (zie bijvoorbeeld De Waal, 2003, blz. 46). De methode bestaat uit het oplossen van een variabele uit één van de lineaire (on)gelijkheden, waarna de oplossing in de andere vergelijkingen wordt ingevuld, rekening houdend met de onderlinge tekens van de ongelijkheden.

Voor categoriale (logische) controleregels bestaat een andere procedure. Daarvoor definiëren we eerst de *normaalvorm* voor categoriale controleregels. Elke controleregel voor categoriale variabelen kan worden geschreven als een combinatie van deelverzamelingen van de domeinen $D_i, 1 \leq i \leq n$, namelijk:

$e_j = (F_1^j, F_2^j, \ldots, F_n^j)$, waarbij elke $F_i^j \subseteq D_i$.

De deelverzamelingen zijn zo gedefinieerd dat als een record $x \in e_j$, dan schendt x controleregel e_j.

Neem als voorbeeld een bestand met de velden $x_1 = burgerlijke\,staat$, $x_2 = leeftijd$ en $x_3 = relatie\,tot\,hoofd\,huishouden$. De domeinen, behorend bij deze variabelen zijn gegeven door:

$D_1 = \{getrouwd, ongetrouwd, verweduwd, gescheiden\}$,

$D_2 = \{<16, \geq 16\}$,

$D_3 = \{huwelijkspartner, kind, anders\}$.

De controleregel die zegt dat iemand die jonger is dan 16 jaar niet getrouwd kan zijn, ziet er in normaalvorm uit als

$e_1 = (\{getrouwd\}, \{<16\}, \{huwelijkspartner, kind, anders\})$.

De controleregel die zegt dat iemand die niet getrouwd is, geen huwelijkspartner kan zijn, wordt in deze notatie weergegeven als

$e_2 = (\{ongetrouwd, verweduwd, gescheiden\}, \{<16, \geq 16\}, \{huwelijkspartner\})$.

Met andere woorden, een controleregel legt een deelverzameling van het totale domein D vast waarvoor geldt dat alle records in die deelverzameling minstens één

fout bevatten. Een controleregel e_j bevat precies die variabelen x_i waarvoor geldt dat $F_i^j \subset D_i$ ($F_i^j \neq D_i$). Dus, regel e_1 in het voorbeeld bevat de variabelen *burgerlijke staat* en *leeftijd*, en regel e_2 bevat *burgerlijke staat* en *relatie tot hoofd huishouden*.

Gezien de twee controleregels uit het voorbeeld is het intuïtief duidelijk dat iemand die jonger is dan 16 jaar, geen huwelijkspartner van het hoofd van het huishouden kan zijn. Deze regel kan inderdaad formeel worden afgeleid uit e_1 en e_2. De algemene procedure verloopt als volgt. Gegeven twee controleregels e_j en e_k, kan een nieuwe geïmpliceerde controleregel $F_s(j,k)$ worden gevormd door middel van de operatie

$$F_s(j,k) = (F_1^j \cap F_1^k, F_2^j \cap F_2^k, \ldots, F_{s-1}^j \cap F_{s-1}^k, F_s^j \cup F_s^k, F_{s+1}^j \cap F_{s+1}^k, \ldots, F_n^j \cap F_n^k),$$

waarbij geldt dat $F_s(j,k) = \varnothing$ wanneer $F_i^j \cap F_i^k = \varnothing$ voor een of meer van de velden x_i, met $i \neq s$. De regels e_j en e_k heten *genererende controleregels* en x_s heet het *genererend veld*. Het is gemakkelijk in te zien dat $F_s(j,k)$ inderdaad een controleregel is. Namelijk, als $x \in F_s(j,k)$, dan volgt uit de definitie dat $x \in e_j$, en/of $x \in e_k$. Omdat $F_s(j,k)$ een controleregel is, volgt meteen dat geïmpliceerde controleregels weer gebruikt kunnen worden om nieuwe geïmpliceerde regels te produceren. De formule voor $F_s(j,k)$ kan daarom eenvoudig worden gegeneraliseerd tot $F_s(E)$ met E een verzameling controleregels. Het kan voorkomen dat de resulterende controleregel de variabele x_s niet meer bevat. Dit gebeurt bijvoorbeeld wanneer $F_s^j \cup F_s^k = D_s$, en wordt het *elimineren* van x_s genoemd. Het gevolg is dat wanneer een record de regel $F_s(j,k)$ schendt, dit record met geen enkele aanpassing van x_s zo kan worden gecorrigeerd dat aan beide regels e_j en e_k wordt voldaan.

In het voorbeeld kan de variabele *burgerlijke staat* geëlimineerd worden door de regel $F_1(1,2)$ te vormen, namelijk

$$F_1(1,2) = (\{getrouwd, ongetrouwd, verweduwd, gescheiden\}, \{<16\}, \{huwlijkspartner\}).$$

Deze regel kan inderdaad worden geïnterpreteerd als: iemand jonger dan 16 jaar kan niet huwelijkspartner zijn van het hoofd van het huishouden. Een record $x \in F_1(1,2)$ kan niet gecorrigeerd worden voor e_1 en e_2 door de variabele *burgerlijke staat* aan te passen. Namelijk, waneer in x voor *burgerlijke staat* de waarde *getrouwd* wordt ingevuld, schendt x zowel e_1 als e_2, wanneer een andere waarde wordt ingevuld schendt x regel e_2.

Controleregels die zowel categoriale als numerieke data bevatten, kunnen in een algemene vorm geschreven worden. Daartoe schrijven we een record eerst als $x = (v, y) = (v_1, v_2, \ldots, v_n, y_{n+1}, y_{n+2}, \ldots, y_{n+m})$, met categoriale variabelen v_i en numerieke variabelen y_i. De algemene vorm voor controleregels wordt dan gegeven door de normaalvorm van categoriale regels te combineren met lineaire controleregels middels een als-dan statement:

$$e_j : \textbf{IF } v \in F^j \textbf{ THEN } y \in \{y : a_{j\bullet} \cdot y \geq b_j\},$$

waarbij F^j, een deelverzameling is van alle mogelijke combinaties van categoriale variabelen. De THEN-conditie is een lineaire voorwaarde voor y met $a_{j\bullet} = (a_{j1}, a_{j2}, \ldots, a_{jm})$. Merk op dat in deze notatie ook lineaire gelijkheden weergegeven kunnen worden, omdat elke lineaire gelijkheid als twee lineaire ongelijkheden kan worden geschreven.

Voor een numerieke variabele y_s zijn deze regels als volgt te combineren tot een geïmpliceerde controleregel:

$$F_s(j,k) = \textbf{IF } v \in F^j \cap F^k \textbf{ THEN } y \in \{y : \tilde{a} \cdot y \geq \tilde{b}\},$$

met $\tilde{a}$ en $\tilde{b}$ de lineaire coëfficiënten en constante die worden verkregen door y_s uit de THEN-condities van e_j en e_k te elimineren met behulp van Fourier-Motzkin eliminatie. Voor het oplossen van het foutlocalisatieprobleem is het niet nodig om impliciete regels te genereren uit dergelijke algemene controleregels waarbij het genererend veld een categoriale variabele is. Het *branch-and-bound* algoritme is zo opgesteld, dat categoriale variabelen pas behandeld worden wanneer alle numerieke variabelen al zijn verwerkt (zie ook volgende paragraaf).

Zoals hiervoor genoemd, is het voor het foutlocalisatieprobleem van belang niet alleen de expliciete, maar ook de impliciete controleregels te kennen. In het algemeen kan het aantal impliciete controleregels erg groot of zelfs oneindig zijn. Echter, Fellegi en Holt (1976) bewezen dat het voldoende is om een eindig aantal zogenaamde *essentieel nieuwe* controleregels te kennen. Een impliciete controleregel $F_s(j,k)$ is een essentieel nieuwe controleregel wanneer geldt dat

(E1) x_s komt niet voor in $F_s(j,k)$ en,

(E2) er is geen controleregel waarvan $F_s(j,k)$ een deelverzameling is.

De eerste eis zegt dat controleregels essentieel nieuw zijn wanneer een variabele uit de genererende regels wordt geëlimineerd. De tweede eis legt vast dat redundante controleregels niet essentieel nieuw zijn. Merk (nogmaals) op dat e_j en e_k ook impliciete regels kunnen zijn.

5.3.4 Het branch-and-bound algoritme

Wanneer een record $x = (x_1, x_2, \ldots, x_n)$ één of meer controleregels schendt, wordt met behulp van een binaire boom de verzameling van mogelijke foutpatronen afgezocht. Een binaire boom is een veelgebruikte gegevensstructuur uit de informatica en is opgebouwd uit *knopen* die zijn verbonden via *gerichte zijden*, of *pijlen*. Er is een unieke beginknoop, die de *stam* wordt genoemd. Vanuit de stam lopen twee zijden, die de stamknoop met twee knopen verbindt, die zijn *kinderen* worden genoemd. Elke knoop in de boom heeft maximaal twee kinderen: het linkerkind en het rechterkind. Elk kind heeft precies één ouder. Een knoop die geen kinderen heeft heet een *blad*, en bevindt zich aan het einde van de boom.

Met elke knoop, behalve de stam, wordt een verzameling controleregels en één variabele geassocieerd. De stam bevat alle expliciete controleregels, en geen variabele. De boom wordt vanuit de stam opgebouwd door één voor één de kandidaat variabelen $x_1, x_2, \ldots, x_n$ te behandelen, als volgt. Kies variabele x_1. In het linkerkind van de stam wordt aangenomen dat x_1 correct is ingevuld en in het rechterkind wordt aangenomen dat x_1 fout is ingevuld. Vervolgens wordt voor het linkerkind en het rechterkind een verzameling correctieregels gegenereerd. Voor het linkerkind worden de correctieregels uit de ouder gekopieerd en wordt de waarde voor x_1 uit het record in die regels ingevuld. De regels die overblijven moeten gelden voor de niet-geselecteerde variabelen $x_2, x_3, \ldots, x_n$ wanneer x_1 niet wordt aangepast. Na invullen van de waarde in x_1 kunnen sommige gaafmaakregels interne tegenspraak vertonen (bijvoorbeeld "$0 \geq 1$"). In dat geval kunnen de kinderen van deze knoop niet leiden tot een oplossing van het localisatieprobleem en wordt deze tak afgebroken. Als er geen interne tegenspraak is kan de tak worden voortgezet. Daarnaast kan het voorkomen dat de verzameling controleregels tautologieën bevat, zoals "1=1". Deze regels kunnen worden verwijderd omdat ze geen enkele variabele bevatten. Voor het rechterkind wordt de variabele x_1 geëlimineerd uit de controleregels van de ouder, met de methoden uit de voorgaande paragraaf. De resulterende verzameling controleregels in het rechterkind zijn de controleregels waaraan de variabelen $x_2, x_3, \ldots, x_n$ moeten voldoen, welke waarde er ook voor x_1 wordt ingevuld. Vervolgens wordt de boom voortgezet door x_2 te selecteren en voor elk kind een linkerkind en een rechterkind te genereren zoals hiervoor. Dit gaat door totdat alle variabelen $x_1, x_2, \ldots, x_n$ geselecteerd zijn. De bladeren die uiteindelijk als variabele x_n hebben en waarbij de bijbehorende verzameling controleregels geen interne tegenspraak bevat komen overeen met een oplossing G voor het localisatieprobleem die voldoet aan eis (G1). Ook kan worden bewezen dat met deze procedure precies alle oplossingen worden gevonden, zie stellingen 1 en 2 uit De Waal en Quere (2003). Elke oplossing wordt gegeven door het unieke pad van de stam naar het blad af te lopen en bij te houden welke variabelen vastgezet zijn en welke geëlimineerd. Hierna moet met een van de eerder genoemde selectieprincipes nog een van de mogelijke oplossingen gekozen worden.

Het is niet nodig om alle mogelijke oplossingen die aan (G1) voldoen te vinden. Namelijk, wanneer in een van de takken blijkt dat de som van de betrouwbaarheidsgewichten van geëlimineerde variabelen groter is dan die van een eerder gevonden oplossing, hoeft deze tak niet te worden voortgezet. Door steeds de oplossing met de kleinste som aan betrouwbaarheidsgewichten vast te houden, wordt efficiënter gezocht naar de oplossingen die aan eis (G1) én (G2) voldoen.

Er dient nog te worden opgemerkt, dat er bij het voorgaande vanuit is gegaan dat voor alle velden $x_1, x_2, \ldots, x_n$ voor het betreffende record een waarde was ingevuld. Wanneer er sprake is van item nonrespons kunnen de leeggelaten velden worden geëlimineerd in de originele verzameling controleregels, aangezien zij toch geïmputeerd moeten worden. Voor de overige variabelen en regels kan de boom geconstrueerd worden zoals hiervoor beschreven.

Het hierboven beschreven algoritme is een basisprocedure. In de praktijk (SLICE) zijn nog aanpassingen aangebracht, waarvan we er hier enkele bespreken. Ten eerste worden de numerieke variabelen eerder dan de categoriale variabelen behandeld om enkele technische moeilijkheden bij het elimineren van variabelen te voorkomen (De Waal, 2005b). Ten tweede kan per record geprobeerd worden om de variabelen in een zo gunstig mogelijke volgorde te doorlopen, zodat oplossingen zo snel mogelijk gevonden worden (zie Daalmans, 2000; De Waal, 2005b). Ten derde kan aan variabelen, naast betrouwbaarheidsgewichten, ook de status "*locked*" worden toegewezen. Het algoritme zoekt dan naar oplossingen waarbij de betreffende variabele niet wordt aangepast. Zie ook De Jong (2002).

5.3.5 Programmatuur op het CBS: SLICE/CherryPie

Sinds 2007 is bij het CBS SLICE 1.6 beschikbaar. Voor een uitgebreide beschrijving verwijzen we naar De Waal (2005a,b), hier wordt slechts een kort overzicht gegeven van de mogelijkheden die SLICE biedt.

SLICE is de aan het CBS ontwikkelde software bibliotheek voor automatisch gaafmaken. De verschillende functies van SLICE zijn ondergebracht in modules. De module CherryPie is in staat om foutlocalisatieproblemen op basis van het gegeneraliseerde principe van Fellegi en Holt op te lossen. CherryPie kan werken met zowel numerieke als categoriale gegevens, en kan lineaire, categoriale en gecombineerde controleregels aan, zoals beschreven in paragraaf 5.3.1. De regels kunnen worden opgesteld in de daarvoor ontwikkelde scripttaal van CherryPie, maar er is ook een module waarmee regels uit Blaise kunnen worden geïmporteerd. Daarnaast is er een imputatiemodule voor numerieke gegevens, waarmee eenvoudige imputatiemethoden, gebaseerd op ratioschatters kunnen worden geïmplementeerd. Na imputatie voldoen de records niet altijd aan alle controleregels, omdat daar in de imputatiemethode niet expliciet rekening mee wordt gehouden. Vandaar dat er nog een extra module (AdaptValues) is die de geïmputeerde waarden weer aanpast. Het is mogelijk om SLICE te gebruiken in combinatie met andere software voor imputatie. De Jong (2002) geeft een uitgebreid overzicht van de mogelijkheden van SLICE.

SLICE zelf heeft geen (grafische) gebruikersinterface, maar bestaat uit een bibliotheek van routines, in de vorm van een .dll (*dynamically linked library*) bestand die vanuit andere programmas kunnen worden aangeroepen. Door deze opzet is SLICE zeer flexibel in te zetten. Er is ook een demonstratieprogramma beschikbaar (SLICEDemo, zie Sluis, 2004) waarmee de functionaliteit van SLICE kan worden uitgetest.

5.4 Voorbeeld

Als voorbeeld werken we een klein stukje van het gaafmaakproces van de statistiek bouwobjecten in voorbereiding (BIV) uit (zie Van der Loo en Pannekoek, 2007). Bij BIV worden architectenbureaus ondervraagd over nieuwe opdrachten. Hierbij wordt onder andere gevraagd naar het type bouwobject $t \in \{w, c, o\}$, waarbij w staat voor woning(en), c voor combinatiegebouwen (deels woning deels andere bestemming) en o voor overige bouwwerken. Daarnaast wordt gevraagd naar het percentage woonoppervlak $p \in [0,100]$ (indien combinatiegebouw) en het aantal woningen $n \in \mathrm{N}$. Het spreekt voor zich dat voor de categorie overig zowel het percentage woonoppervlak als het aantal woningen gelijk moet zijn aan 0. In de notatie van de vorige paragrafen leidt dit tot de volgende controleregels:

$$e_1 = (o, (0,100], \mathrm{N})$$
$$e_2 = (o, [0,100], \mathrm{N}^+).$$

Hierbij zegt e_1 dat voor overige bouwwerken het percentage niet groter kan zijn dan 0, en zegt e_2 dat voor overige bouwwerken het aantal woningen niet groter dan 0 kan zijn. We kiezen alle betrouwbaarheidsgewichten gelijk aan 1. Er zijn in dit geval geen essentieel nieuwe geïmpliceerde controleregels. Controleer namelijk dat $F_1(1,2) \subset e_1$, zodat deze aan geen van de eisen (E1) en (E2) voldoet. Ga verder na dat $F_2(1,2) = e_2$ en $F_3(1,2) = e_1$, zodat ook deze regels niet aan (E2) voldoen. Bekijk nu een record $r = (o, 10\%, 0)$. Dit record schendt alleen expliciete regel e_1. Regel e_1 bevat de velden t en p zodat er twee minimale overdekkende verzamelingen mogelijk zijn, namelijk $G_1 = \{t\}$ en $G_2 = \{p\}$. Bekijk tot slot een record $r' = (o, 10\%, 3)$. Dit record schend zowel e_1 als e_2. De enige variabele die in beide geschonden controleregels voorkomt is het type bouwwerk t, zodat er slechts één optimale oplossing is, namelijk $G = \{t\}$.

5.5 Eigenschappen

We merken nog op dat de foutopsporingsmethode op basis van het principe van Fellegi en Holt, geschikt is voor parallelle verwerking over meerdere servers. Omdat alle records onafhankelijk verwerkt worden schaalt de verwerkingstijd vrijwel lineair met het aantal servers wat kan worden ingezet.

5.6 Kwaliteitsindicatoren

De methode werkt beter wanneer inderdaad de werkelijk gemaakte fouten worden opgespoord. Met behulp van simulaties kan een idee worden gekregen of dit inderdaad zo is. Men kan bijvoorbeeld realistische fouten in een gaaf databestand aanbrengen om te kijken onder welke omstandigheden ze worden teruggevonden met behulp van SLICE.

Een tweede aspect kan de efficiëntie zijn waarmee het *branch-and-bound* algoritme oplossingen voor het foutlocalisatieprobleem vindt. Dit kan in SLICE min of meer worden gecontroleerd door betrouwbaarheidsgewichten aan te passen, of door variabelen op “locked” te zetten.

6. Foutlocalisatie met de Nearest-neighbour Imputatie Methodologie

6.1 Korte beschrijving

De Nearest-neighbour Imputatie Methodologie (NIM) is een alternatieve methode voor automatische foutlocalisatie op recordniveau. In tegenstelling tot de methode uit hoofdstuk 5 is de NIM niet gebaseerd op het principe van Fellegi en Holt, maar op een daarvan afgeleid principe. De NIM bepaalt niet alleen een oplossing van het foutlocalisatieprobleem – d.w.z. een verzameling velden die kan worden geïmputeerd zodat aan alle controleregels is voldaan – maar ook de te imputeren waarden. Wat dat betreft kan men deze methode ook zien als een imputatiemethode. In feite vormt de NIM een uitbreiding van hot-deck donorimputatie op basis van een afstandsfunctie (zie hoofdstuk 6 in het thema *Imputatie*), bedoeld voor de situatie dat de data nog fouten kunnen bevatten.

Voor elk record dat niet aan alle controleregels voldoet, maakt de NIM een lijst met donorrecords die (volgens een zekere afstandsfunctie) goed lijken op het te imputeren record. Aan de hand van de donorrecords bepaalt de NIM manieren om fouten aan te wijzen in het record, zodat de foute velden kunnen worden geïmputeerd met de bijbehorende waarden uit een donorrecord, op zodanige wijze dat aan alle controleregels is voldaan. Ten slotte kiest de NIM de beste van alle voorgestelde geïmputeerde versies van het record, volgens een criterium dat wordt toegelicht in paragraaf 6.3.

Voor het toepassen van de NIM is de door Statistics Canada ontwikkelde software CANCEIS (CANadian Census Edit & Imputation System) op het CBS beschikbaar.

6.2 Toepasbaarheid

De NIM is bij Statistics Canada ontwikkeld voor één doel: controle en correctie van de vijfjaarlijkse volkstelling (zie bijvoorbeeld Bankier e.a., 1994). Dit komt tot uitdrukking in een aantal eigenschappen van de methode:

- De NIM is in staat om zeer grote datasets snel te verwerken. Een belangrijke voorwaarde is wel dat er voldoende foutloze donorrecords beschikbaar zijn. Dit is precies de situatie die zich voordoet bij een volkstelling: miljoenen records waarvan de meeste geen fouten bevatten. De NIM is geen geschikte methode in de situatie dat bijna alle records fouten bevatten. In dat geval zouden namelijk steeds dezelfde records als donor worden gebruikt.
- De NIM kan zowel numerieke als categoriale gegevens verwerken, alsook een combinatie van beide. De methode is echter vooral geschikt voor datasets met voornamelijk categoriale variabelen (en eventueel een paar numerieke variabelen), zoals bij een volkstelling. De methode is niet geschikt voor volledig numerieke datasets die moeten voldoen aan lineaire gelijkheden, zoals bij de Productiestatistieken. In dat geval is het namelijk

bijna onmogelijk om een geschikte donor te vinden waaruit een record kan worden geïmputeerd zodat aan de lineaire gelijkheden is voldaan.

- De NIM gebruikt de statistische eigenschappen van de verzameling donorrecords als benadering voor de statistische eigenschappen van de hele populatie. De methode is daarom in de eerste plaats bedoeld voor statistieken die zijn gebaseerd op een integrale telling, zoals de volkstelling. Gegevens verkregen uit een steekproef geven doorgaans alleen een correcte afspiegeling van de hele populatie indien gebruik wordt gemaakt van ophooggewichten. Dit is niet mogelijk met de huidige vorm van de NIM.

6.3 Uitgebreide beschrijving

6.3.1 Records en controleregels

Evenals in hoofdstuk 5 gaan we uit van een databestand met records van n velden, die we noteren als $x = (x_1, \ldots, x_n)$. Zoals gezegd in paragraaf 6.2 mag een record voor de NIM zowel categoriale als numerieke velden bevatten.

Voor het lokaliseren van fouten maken we gebruik van controleregels die aangeven welke waarden en combinaties van waarden niet zijn toegestaan. De implementatie van de NIM in de huidige versie van CANCEIS gaat ervan uit dat alle controleregels de volgende algemene vorm hebben:

als (Δ_1 **en** Δ_2 **en** $(\cdots)$ **en** Δ_S) **dan** $\varnothing$. (6.3.1)

Bij numerieke velden staat elke Δ_s voor een lineaire propositie van de vorm

$$\Delta_s: \quad a_{s1}x_1 + \cdots + a_{sn}x_n \lhd b_s,$$

waarbij op de plaats van het symbool $\lhd$ een van de operatoren $<, >, \leq, \geq, =, \neq$ moet worden ingevuld. Bij categoriale velden heeft elke Δ_s de vorm

$$\Delta_s: \quad x_i \in F_i^s, \text{ voor zekere } i \in \{1, \ldots, n\},$$

met F_i^s een deelverzameling van het domein van x_i, analoog aan paragraaf 5.3.3. Een record voldoet niet aan controleregel (6.3.1), en bevat dus een fout, indien alle proposities $\Delta_1, \ldots, \Delta_S$ evalueren tot *waar* wanneer de waarden uit het record worden ingevuld.

Ter illustratie verwijzen we in deze paragraaf steeds naar een klein voorbeeld met vier velden: $x = (x_1, x_2, x_3, x_4) = (\textit{Leeftijd, Inkomen, Burgerlijke Staat, Relatie tot Hoofd Huishouden})$. De eerste twee velden zijn numeriek met de niet-negatieve gehele getallen als domein. De laatste twee velden zijn categoriaal. Variabele x_3 heeft als mogelijke waarden *Getrouwd*, *Ongetrouwd*, *Verweduwd* en *Gescheiden*. Variabele x_4 heeft als mogelijke waarden *Partner*, *Kind* en *Overig*.

Er zijn in dit voorbeeld drie controleregels waaraan de records moeten voldoen. In woorden zeggen deze regels het volgende:

1. Personen jonger dan 18 jaar kunnen niet getrouwd zijn (geweest).
2. Personen jonger dan 12 jaar hebben geen inkomen boven de 0 Euro.
3. Personen die niet getrouwd zijn kunnen geen partner zijn van het hoofd van het huishouden.

We schrijven de drie controleregels als volgt in de algemene vorm (6.3.1):

1. **als** ($x_1 < 18$ **en** $x_3 \in \{Getrouwd, Verweduwd, Gescheiden\}$) **dan** $\varnothing$.
2. **als** ($x_1 < 12$ **en** $x_2 > 0$) **dan** $\varnothing$.
3. **als** ($x_3 \in \{Ongetrouwd, Verweduwd, Gescheiden\}$ **en** $x_4 \in \{Partner\}$) **dan** $\varnothing$.

De lezer kan zelf nagaan dat deze twee formuleringen dezelfde regels beschrijven.

6.3.2 Donorselectie

Om te bepalen in welke mate twee records op elkaar lijken, definiëren we een globale afstandsfunctie. De afstand tussen de records $x^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$ en $x^{(2)} = (x_1^{(2)}, \ldots, x_n^{(2)})$ is

$$D(x^{(1)}, x^{(2)}) = \sum_{i=1}^{n} w_i D_i(x_i^{(1)}, x_i^{(2)}), \tag{6.3.2}$$

met $w_i \geq 0$ het gewicht van variabele x_i en $D_i(x_i^{(1)}, x_i^{(2)})$ de afstand tussen de waarden $x_i^{(1)}$ en $x_i^{(2)}$. Voor elke variabele kiest men een lokale afstandsfunctie, met als enige voorwaarden dat $0 \leq D_i(x_i^{(1)}, x_i^{(2)}) \leq 1$ en dat $D_i(x_i^{(1)}, x_i^{(2)}) = 0$ als $x_i^{(1)} = x_i^{(2)}$, en een gewicht dat het belang van de variabele uitdrukt. Een hogere waarde van w_i betekent dat variabele x_i meer invloed heeft op de afstandsfunctie. Om een variabele weg te laten uit (6.3.2) kiezen we $w_i = 0$.

In de eerste stap van de NIM worden alle records in het databestand gecontroleerd aan de hand van de door de gebruiker gekozen controleregels. Records die ten minste één controleregel schenden bevatten blijkbaar fouten, en worden in de tweede stap onderworpen aan de automatische foutlocalisatie. Alle andere records worden geplaatst in de zogenaamde *donorpool,* d.w.z. de verzameling potentiële donorrecords. Om de NIM met succes te kunnen gebruiken moet de donorpool de grote meerderheid van de records uit het databestand bevatten (zie paragraaf 6.2).

De records die in de eerste stap van de NIM zijn aangewezen voor foutlocalisatie, worden in de tweede stap één voor één behandeld. Gegeven een record met fouten $x^{(F)}$ zoekt de NIM in de donorpool naar records $x^{(D)}$ met een zo klein mogelijke afstand $D(x^{(F)}, x^{(D)})$. Om de rekentijd laag te houden worden niet alle records uit de donorpool bekeken, maar alleen de records die in het oorspronkelijke databestand in de buurt van $x^{(F)}$ liggen. De achterliggende aanname is dat het databestand zo gesorteerd is dat records die dicht bij elkaar liggen meer op elkaar lijken dan records die ver uit elkaar liggen. In het geval van de Canadese volkstelling is het

bijvoorbeeld gebruikelijk om het databestand te sorteren op geografische kenmerken. De N_D records met de kleinste afstand $D(x^{(F)}, x^{(D)})$ worden bewaard als potentiële donoren, met N_D een instelbare parameter.

In het voorbeeld met vier variabelen uit paragraaf 6.3.1 is

$$x^{(F)} = (x_1^{(F)} = 9, x_2^{(F)} = 25000, x_3^{(F)} = Ongetrouwd, x_4^{(F)} = Partner)$$

een record dat niet aan alle controleregels voldoet. Dit record schendt namelijk zowel de tweede als de derde regel. Een voorbeeld van een record dat wel aan alle controleregels voldoet, en dus geschikt is als donorrecord, is

$$x^{(D)} = (x_1^{(D)} = 8, x_2^{(D)} = 0, x_3^{(D)} = Ongetrouwd, x_4^{(D)} = Kind) .$$

6.3.3 Genereren van imputatie-acties

Nadat de potentiële donoren voor record $x^{(F)}$ uit de donorpool zijn geselecteerd, probeert de NIM de fouten in $x^{(F)}$ op te lossen door sommige waarden te vervangen door de bijbehorende waarden uit een donorrecord $x^{(D)}$. Het overnemen van waarden uit een donor in een ander record heet een *imputatie-actie*. We noteren een imputatie-actie formeel als $I = (x^{(F)}, x^{(D)}, \delta)$, waarbij δ een rij binaire variabelen voorstelt, $\delta = (\delta_1, \ldots, \delta_n)$, met $\delta_i = 1$ als de waarde $x_i^{(F)}$ wordt vervangen door $x_i^{(D)}$, en anders $\delta_i = 0$.

Het resultaat van de imputatie-actie $I = (x^{(F)}, x^{(D)}, \delta)$ is een aangepast record $x^{(A)} = (x_1^{(A)}, \ldots, x_n^{(A)})$, met

$$x_i^{(A)} = \delta_i x_i^{(D)} + (1 - \delta_i) x_i^{(F)}, \qquad i = 1, \ldots, n .$$

Het is duidelijk dat we variabelen waarvoor $x_i^{(F)} = x_i^{(D)}$ buiten beschouwing kunnen laten bij het genereren van imputatie-acties.

Een imputatie-actie heet *toegelaten* wanneer zij een aangepast record $x^{(A)}$ oplevert dat aan alle controleregels voldoet. De NIM bepaalt alle toegelaten imputatie-acties die worden gegenereerd door de N_D potentiële donoren. Vaak genereert een donorrecord meerdere toegelaten imputatie-acties.

In het eerder gegeven voorbeeld verkrijgen we een toegelaten imputatie-actie voor $x^{(F)}$ door de waarden van x_2 en x_4 uit $x^{(D)}$ over te nemen. Het aangepaste record is in dat geval namelijk

$$x^{(A)} = (x_1^{(A)} = 9, x_2^{(A)} = 0, x_3^{(A)} = Ongetrouwd, x_4^{(A)} = Kind)$$

en het is eenvoudig na te gaan dat dit record voldoet aan de drie controleregels. In formele notatie schrijven we deze imputatie-actie als $I = (x^{(F)}, x^{(D)}, \delta = (0,1,0,1))$.

Omdat $x_3^{(F)} = x_3^{(D)} = Ongetrouwd$, kunnen in dit voorbeeld alleen de drie variabelen x_1, x_2 en x_4 worden gebruikt voor het genereren van (zinvolle) imputatie-acties. In

totaal zijn er daarom $2^3 - 1 = 7$ mogelijke imputatie-acties. Het blijkt dat slechts twee van deze acht imputatie-acties toegelaten zijn; de enige toegelaten imputatie-actie naast de reeds genoemde imputeert de variabelen x_1, x_2 en x_4, met als resultaat:

$$x^{(A)} = (x_1^{(A)} = 8, x_2^{(A)} = 0, x_3^{(A)} = Ongetrouwd, x_4^{(A)} = Kind).$$

6.3.4 Selecteren uit toegelaten imputatie-acties

In het voorbeeld uit paragraaf 6.3.3 geldt voor de laatste imputatie-actie: $x^{(A)} = x^{(D)}$. In het algemeen bestaat er altijd een imputatie-actie met deze eigenschap (kies $\delta_i = 1$ voor alle $i = 1, \ldots, n$), en zij is per definitie toegelaten. Het vinden van een toegelaten imputatie-actie is dus zeer eenvoudig. Het doel van de NIM is echter ambitieuzer, namelijk het vinden van de *best mogelijke* toegelaten imputatie-actie. Onder 'best mogelijk' wordt bij de NIM verstaan de toegelaten imputatie-actie $I = (x^{(F)}, x^{(D)}, \delta)$ met de volgende twee eigenschappen:

(B1) $x^{(A)}$ lijkt zo veel mogelijk op $x^{(F)}$.

(B2) $x^{(A)}$ lijkt zo veel mogelijk op $x^{(D)}$.

Enerzijds is het wenselijk dat een toegelaten imputatie-actie zo weinig mogelijk verandert aan het oorspronkelijke record; dit is de redenering achter eigenschap (B1), die doet denken aan het principe van Fellegi en Holt uit hoofdstuk 5. Anderzijds geldt dat het aangepaste record een kunstmatig record is, samengesteld uit twee verschillende records. We weten dat de combinatie van waarden in het aangepaste record niet in strijd is met de controleregels, maar het zou kunnen dat deze combinatie van waarden in de populatie zeer zeldzaam is[4]. Dergelijke imputatie-acties zijn weinig plausibel. Naarmate het aangepaste record beter lijkt op het donorrecord, is de plausibiliteit van het aangepaste record groter, omdat het lijkt op een foutloos record dat op natuurlijke wijze is verkregen; dit is de redenering achter eigenschap (B2).

Als voorbeeld bekijken we het volgende record, dat niet voldoet aan controleregel 3 uit paragraaf 6.3.1:

$$x^{(F)} = (x_1^{(F)} = 56, x_2^{(F)} = 30000, x_3^{(F)} = Ongetrouwd, x_4^{(F)} = Partner),$$

en twee potentiële donorrecords:

$$x^{(D,1)} = (x_1^{(D,1)} = 59, x_2^{(D,1)} = 28000, x_3^{(D,1)} = Getrouwd, x_4^{(D,1)} = Partner),$$

$$x^{(D,2)} = (x_1^{(D,2)} = 21, x_2^{(D,2)} = 30000, x_3^{(D,2)} = Ongetrouwd, x_4^{(D,2)} = Kind).$$

---

[4] Cf. het onderscheid tussen harde en zachte controleregels (paragraaf 5.2). Bij automatische foutlocalisatie op basis van de NIM worden alle controleregels als harde regels opgevat, net als bij foutlocalisatie op basis van het principe van Fellegi en Holt. Men kan daarom bij de NIM geen zachte controleregels gebruiken om ongebruikelijke waardecombinaties te identificeren, zoals bij interactief gaafmaken wel kan.

Twee toegelaten imputatie-acties zijn: imputeer $x_3^{(D,1)}$ of imputeer $x_4^{(D,2)}$. De bijbehorende aangepaste records zijn:

$$x^{(A,1)} = (x_1^{(A,1)} = 56, x_2^{(A,1)} = 30000, x_3^{(A,1)} = Getrouwd, x_4^{(A,1)} = Partner),$$

$$x^{(A,2)} = (x_1^{(A,2)} = 56, x_2^{(A,2)} = 30000, x_3^{(A,2)} = Ongetrouwd, x_4^{(A,2)} = Kind).$$

De records $x^{(A,1)}$ en $x^{(A,2)}$ voldoen beide aan de controleregels. In beide gevallen is slechts één variabele uit het oorspronkelijke record veranderd. In de populatie zal men echter veel meer 56-jarigen aantreffen die getrouwd zijn met het hoofd van het huishouden waartoe zij behoren, dan 56-jarigen die een kind zijn van het hoofd van het huishouden. Merk op: donorrecord $x^{(D,2)}$ hoort zelf niet bij een 56-jarige die een kind is van het hoofd van het huishouden, maar een dergelijk record ontstaat wanneer waarden uit $x^{(F)}$ en $x^{(D,2)}$ worden gecombineerd.

Voor elke toegelaten imputatie-actie $I = (x^{(F)}, x^{(D)}, \delta)$ bepaalt de NIM de volgende maat:

$$\mu(I) = \alpha D(x^{(F)}, x^{(A)}) + (1-\alpha) D(x^{(A)}, x^{(D)}).$$

Hierbij zijn $D(x^{(F)}, x^{(A)})$ en $D(x^{(A)}, x^{(D)})$ gedefinieerd via (6.3.2) en is α een door de gebruiker te kiezen parameter met $1/2 < \alpha \leq 1$. De best mogelijke imputatie-actie is nu gedefinieerd als de imputatie-actie met de kleinste waarde van $\mu(I)$. De keuze van α bepaalt of, en zo ja in welke mate, gekeken wordt naar de plausibiliteit van het aangepaste record: met $\alpha = 1$ kijkt de NIM alleen naar eigenschap (B1), met $\alpha < 1$ doet eigenschap (B2) ook mee. Bij de controle en correctie van de Canadese volkstelling zijn de waarden $\alpha = 0{,}75$ en $\alpha = 0{,}9$ gebruikt.

Door bij het beoordelen van toegelaten imputatie-acties te letten op $D(x^{(A)}, x^{(D)})$, hoopt men dat de NIM in staat is om de univariate en multivariate verdelingen in de populatie te behouden. Een aangepast record $x^{(A)}$ met een bepaalde combinatie van waarden die in, zeg, 5% van alle donorrecords voorkomt, zal naar verwachting ook in ongeveer 5% van de gevallen een kleine $D(x^{(A)}, x^{(D)})$ hebben. In de andere 95% van de gevallen is $D(x^{(A)}, x^{(D)})$ groot en zal de bijbehorende toegelaten imputatie-actie waarschijnlijk niet worden gekozen. Hierbij wordt aangenomen dat de donorpool een goede afspiegeling vormt van de populatie als geheel.

Stel, voor de toegelaten imputatie-acties bij een record is de kleinste waarde van $\mu(I)$ gelijk aan $\mu_{\min}$. Een toegelaten imputatie-actie heet in de terminologie van de NIM een *near minimum change imputation action* (NMCIA) als zij voldoet aan

$$\mu(I) \leq \gamma \mu_{\min},$$

met $\gamma \geq 1$ een door de gebruiker te kiezen parameter. Voor elk record worden alleen de NMCIA's bewaard. Men kan γ iets groter dan 1 kiezen, omdat toegelaten imputatie-acties met $\mu(I)$ in de buurt van $\mu_{\min}$ nauwelijks slechter zijn dan de best

mogelijke imputatie-actie. Het bewaren van dergelijke imputatie-acties helpt te voorkomen dat steeds dezelfde donorrecords worden gebruikt bij het imputeren. De waarde $\gamma = 1{,}1$ is in Canada gebruikt bij de volkstelling.

Ten slotte maakt de NIM een willekeurige keuze uit de lijst met NMCIA's bij een record. Het bijbehorende aangepaste record $x^{(A)}$ vervangt $x^{(F)}$ in de uitvoer. Op deze manier worden alle records die niet aan de controleregels voldoen één voor één behandeld.

6.3.5 Programmatuur: CANCEIS

Voor toepassingen van de NIM op statistische bureaus is de software CANCEIS gratis beschikbaar gesteld door Statistics Canada. CANCEIS wordt nog steeds doorontwikkeld aan de hand van ervaringen bij de Canadese volkstelling. De onderstaande beschrijving gaat uit van CANCEIS versie 4.5 uit 2006, die op het CBS beschikbaar is. Zie ook CANCEIS (2006).

CANCEIS bestaat uit drie modules. De eerste module analyseert controleregels en werkt alleen als ondersteuning voor de andere modules. De *Derive Module* wordt gebruikt voor het uitvoeren van afleidingen en deductieve correcties middels correctieregels (zie hoofdstuk 2). De *Hotdeck Module* ten slotte bevat een implementatie van de NIM. Bij deze implementatie is een efficiënt algoritme gebruikt voor het zoeken naar toegelaten imputatie-acties; zie Bankier (2006) voor een uitgebreide behandeling van dit algoritme.

De invoer van CANCEIS bestaat uit een aantal ASCII-bestanden, waaronder het ruwe databestand en een bestand met controleregels van het type (6.3.1). De controleregels moeten worden geformuleerd in de vorm van zogenaamde *Decision Logic Tables* (DLT's). Een DLT bestaat uit rijen en kolommen. Elke kolom behalve de eerste correspondeert met een controleregel. De eerste kolom bevat alle proposities Δ_s die voorkomen in de controleregels, en elke rij heeft betrekking op de propositie in zijn eerste kolom. Het binnenwerk van een DLT is opgebouwd uit de elementen 'Y', 'N' en '–', die aangeven of en zo ja hoe een propositie voorkomt in een bepaalde controleregel: 'Y' betekent dat de propositie zelf voorkomt, 'N' betekent dat de ontkenning van de propositie voorkomt en '–' betekent dat de propositie niet voorkomt.

Ter illustratie schrijven we de drie controleregels uit paragraaf 6.3.1 in een DLT:

```
                      1   2   3
x1 < 18               Y   -   -
x1 < 12               -   Y   -
x2 > 0                -   Y   -
x3 = Ongetrouwd       N   -   -
x3 = Getrouwd         -   -   N
x4 = Partner          -   -   Y
```

Voor meer voorbeelden van DLT's: zie CANCEIS (2006) en Scholtus (2008b).

6.4 Voorbeeld

Op het CBS is CANCEIS getest voor gebruik bij de productie van demografische statistieken op basis van de Gemeentelijke BasisAdministratie (GBA). De situatie is hier enigszins vergelijkbaar met de Canadese volkstelling: er is sprake van een min of meer integraal populatiebestand met een groot aantal records, waarin sporadisch nog fouten voorkomen. De NIM lijkt hier een geschikte keuze, omdat voldoende donorrecords beschikbaar zijn. Zie Pannekoek e.a. (2008) en Scholtus (2008b) voor meer informatie over deze toepassing van CANCEIS.

6.5 Kwaliteitsindicatoren

De kwaliteit van een methode voor automatische foutlocalisatie wordt in de eerste plaats bepaald door de mate waarin foute velden correct worden geïdentificeerd. Omdat de NIM tevens een imputatiemethode omvat, is ook van belang in hoeverre de imputaties overeenkomen met de werkelijke waarden. Hierbij kan men zowel geïnteresseerd zijn in de kwaliteit van de individuele imputaties als in de mate waarin de geïmputeerde data bepaalde populatieverdelingen correct weergeven. Al deze eigenschappen zijn in de praktijk alleen meetbaar door middel van simulaties, waarbij bekende fouten en ontbrekende waarden worden aangebracht in een gaaf databestand.

Een ander aspect van de kwaliteit van de NIM is de efficiëntie van het zoekalgoritme. De gebruiker kan de benodigde rekentijd tot op zekere hoogte zelf beïnvloeden, door zijn keuzes voor de parameter N_D en voor het aantal records in de omgeving van $x^{(F)}$ dat bekeken wordt tijdens het zoeken naar potentiële donoren (zie paragraaf 6.3.2).

7. Macrogaafmaken

7.1 Aggregaatmethode

7.1.1 Korte beschrijving

Bij de *aggregaatmethode* worden eerst aggregaten berekend, meestal de publicatiecijfers. Indien de berekende aggregaten duidelijk afwijken van wat men zou verwachten, bijvoorbeeld op grond van vroegere gegevens, wordt gekeken naar een lager aggregatieniveau en worden onderliggende records gecontroleerd en eventueel gecorrigeerd. Voor aggregaten die weinig afwijken van wat men zou verwachten kunnen verdere controles worden uitgevoerd om te bepalen of een aggregaat correct is. Aggregaten kunnen incorrect zijn door invloedrijke fouten of foutieve ophoogfactoren.

7.1.2 Toepasbaarheid

Het doel van de aggregaatmethode is het fiatteren van publicatiecijfers. Daarbij kan op een lager aggregatieniveau gekeken worden om de stabiliteit van de cijfers te bepalen, bijvoorbeeld per grootteklasse. Als niveaus of ontwikkelingen afwijken van de verwachting dan kunnen potentiële invloedrijke fouten worden gedetecteerd met behulp van scorefuncties.

Het is essentieel om naar aggregaten te kijken, vooral als de waargenomen data incompleet zijn en er daarom is geïmputeerd of opgehoogd. Er kunnen naast problemen met de microdata namelijk ook problemen met de imputatie- of ophoogmethode zijn. Invloedrijke fouten kunnen bij het microgaafmaken over het hoofd zijn gezien of daar zijn geïntroduceerd. De aggregaatmethode is vooral nuttig als invloedrijke fouten structureel voorkomen in microgaafgemaakte data, of als er structurele problemen zijn met ophoogfactoren of imputaties.

De voorwaarden voor de aggregaatmethode zijn:

- Systematische fouten (overduidelijke én minder duidelijke fouten) zijn verwijderd bij het microgaafmaken.
- Er is voor ieder record een waarneming of imputatie beschikbaar of er is een ophoogfactor beschikbaar voor ieder waargenomen record;
- Er zitten niet al te veel invloedrijke fouten in de microgaafgemaakte data. Oftewel, er kunnen zinvolle aggregaten worden bepaald;
- Er is een referentiekader. Oftewel, er zijn referentiedata of een gaafmaker heeft voldoende branchekennis om aggregaten te kunnen beoordelen.

Dat een publicatiecijfer plausibel is wil niet zeggen dat deze correct is. Er kunnen nog steeds invloedrijke fouten in de data zitten. Het is daarom aan te raden om de

aggregaatmethode te combineren met de verdelingsmethode, zie paragraaf 7.2. Aan de hand van de verdelingsmethode kan ook worden bepaald of er nog systematische of invloedrijke fouten in de data zitten. Bovendien kunnen uitbijters worden gedetecteerd met de verdelingsmethode.

7.1.3 Uitgebreide beschrijving

Er zijn diverse redenen waarom publicatiecijfers afwijken van de verwachting.

- Er kunnen invloedrijke meet- of verwerkingsfouten in de data zitten;
- Er kunnen problemen zijn met het ophoogkader of de ophoogmethodiek;
- Er kunnen onverwachte ontwikkelingen zijn, die wel degelijk reëel zijn.

Om te bepalen of er nader naar de microdata moet worden gekeken kan de relatieve afwijking van een aggregaat voor variabele y_j in periode t t.o.v. een referentieaggregaat $\hat{Y}_j^s$ worden berekend:

$$\frac{\hat{Y}_j^t - \hat{Y}_j^s}{\hat{Y}_j^s}, \tag{7.1.1}$$

waarbij

$$\hat{Y}_j^t = \sum_{i=1}^{n} w_i^t y_{ij}^t \ . \tag{7.1.2}$$

Het referentieaggregaat kan zijn bepaald op basis van een andere bron of dezelfde bron voor een eerdere periode s.

Er kan ook worden gekeken naar een kengetal. Dit is een verhouding van twee gerelateerde variabelen y_j en y_k. Een kengetal is in principe stabieler en beter interpreteerbaar dan de variabelen afzonderlijk. De relatieve afwijking van een kengetal kan als volgt worden bepaald:

$$\left(\frac{\hat{Y}_j^t}{\hat{Y}_k^t} - \frac{\hat{Y}_j^s}{\hat{Y}_k^s}\right) \Big/ \frac{\hat{Y}_j^s}{\hat{Y}_k^s}. \tag{7.1.3}$$

Als blijkt dat aggregaten of kengetallen teveel afwijken van de verwachting dan is het raadzaam om de onderliggende data nog eens te bekijken. Dit kan aan de hand van de verdelingsmethode, zie paragraaf 7.2. Er kan ook gebruik worden gemaakt van scorefuncties om mogelijke invloedrijke fouten te detecteren, zie paragraaf 4.3.2. Groot voordeel is dat de ophoogfactoren en aggregaten voor de verslagperiode dan beschikbaar zijn. Met behulp van scorefuncties kan de invloed van een record op een adequate wijze worden meegenomen in de foutdetectie.

In (7.1.1) en (7.1.3) wordt geen rekening gehouden met de steekproefvariantie van aggregaten. Als een schatting voor de standaarddeviatie van het verschil in respectievelijk aggregaten en kengetallen beschikbaar is dan kunnen onderstaande relatieve afwijkingen bepaald worden:

$$\frac{\hat{Y}_j^t - \hat{Y}_j^s}{s.d.(\hat{Y}_j^t - \hat{Y}_j^s)}, \tag{7.1.4}$$

$$(\frac{\hat{Y}_j^t}{\hat{Y}_k^t} - \frac{\hat{Y}_j^s}{\hat{Y}_k^s}) \Big/ s.d.(\frac{\hat{Y}_j^t}{\hat{Y}_k^t} - \frac{\hat{Y}_j^s}{\hat{Y}_k^s}). \tag{7.1.5}$$

7.1.4 Voorbeeld

Met een methode die ontwikkeld is op het CBS kunnen op basis van BTW-omzetten niveaus en ontwikkelingen worden geschat van de omzet van het midden- en kleinbedrijf in de Detailhandel. In dit voorbeeld richten we ons op de kwartaalomzet van winkels in herenkleding gedurende vijf kwartalen, zie tabel 7. De aggregaten zijn bepaald voordat invloedrijke verdachte waarden zijn gecontroleerd die met behulp van scorefuncties zijn gedetecteerd. BTW-omzetten worden verwijderd als de gaafmaker denkt dat deze incorrect zijn. Het is moeilijk om BTW-omzetten te corrigeren, omdat deze niet door het CBS en volgens CBS-definities zijn waargenomen en berichtgevers niet mogen worden benaderd.

Tabel 7. Geschatte totale kwartaalomzet en omzetontwikkeling van winkels in herenkleding in gk 10-40 voor 1e kwartaal 2008 t/m 1e kwartaal 2009

Periode	Totale omzet (in mln euro's)	Kw. op kw.-ontwikkeling	Ontwikkeling t.o.v. 1e kwartaal 2008
1e kwartaal 2008	120	-	-
2e kwartaal 2008	154	29,1%	29,1%
3e kwartaal 2008	136	-12,3%	13,2%
4e kwartaal 2008	174	28,3%	45,3%
1e kwartaal 2009	115	-33,6%	-3,6%

De kwartaalomzetten lijken op het eerste gezicht plausibel. Door de uitverkoop in juni en december is er relatief veel omzet in het tweede en vierde kwartaal. Een kwartaalomzet van 120 miljoen wordt gehaald als een volwassen man gemiddeld 20 euro per kwartaal in een kleine tot middelgrote herenkledingwinkel spendeert.

De kwartaal-op-kwartaal ontwikkelingen in tabel 7 passen dus in het verwachte seizoenspatroon voor winkels in herenkleding. Per grootteklasse is dezelfde ontwikkeling zichtbaar. Een negatieve jaar-op-jaar ontwikkeling voor het 1e kwartaal in 2009 past ook in het plaatje van de kredietcrisis. De ontwikkelingen lijken dus plausibel. Er dient echter toch naar invloedrijke verdachte waarden te worden gekeken, omdat de marge waarbinnen een ontwikkeling plausibel is behoorlijk groot is. Het doel is om de ware ontwikkeling te benaderen.

7.1.5 Kwaliteitsindicatoren

We kunnen (7.1.1) en (7.1.4) berekenen vóór en na macrogaafmaken en bepalen in hoeverre de relatieve afwijking van een aggregaat is afgenomen. Hetzelfde geldt

voor (7.1.3) en (7.1.5) als er naar kengetallen wordt gekeken. Als de relatieve afwijking vergelijkbaar blijft dan heeft het macrogaafmaken weinig opgeleverd. Het is evenwel mogelijk dat een aggregaat/kengetal de werkelijkheid benadert, omdat een afwijking t.o.v een referentie aggregaat/kengetal terecht kan zijn. Het macrogaafmaken kan er dan wel voor gezorgd hebben dat

- Er meer vertrouwen is in de geconstateerde afwijking op macroniveau;
- Er verbeteringen hebben plaats hebben gevonden op microniveau. Dit leidt tot betere referentiewaarden voor het gaafmaken van de volgende periode.

Voor één of meerdere variabelen in de gecontroleerde records kan een viertal percentages worden bepaald:

- percentage records met een gevonden fout;
- percentage records met een gevonden invloedrijke fout;
- percentage records met een gevonden niet-representatieve (correcte) uitbijter;
- percentage records met een gevonden representatieve (correcte) uitbijter.

7.2 Verdelingsmethode

7.2.1 Korte beschrijving

Bij de verdelingsmethode worden waarden van variabelen in een groep van records met elkaar vergeleken aan de hand van de univariate en multivariate verdeling van deze variabelen. Dit kan met grafische hulpmiddelen of met behulp van statistische maten. De uitbijters, de meest verdachte records, worden vervolgens gecontroleerd. Als blijkt dat uitbijters invloedrijke foutieve waarden zijn, dan worden deze gecorrigeerd. Als een uitbijter correct wordt geacht, dan is de vraag of deze representatief is voor de populatie. Bij een steekproef van bijvoorbeeld 1 op 10 is een uitbijter representatief als er buiten de steekproef negen vergelijkbare uitbijters voorkomen in de populatie. Als een uitbijter niet representatief wordt geacht, dan wordt de ophoogfactor aangepast. Methoden voor het opsporen van uitbijters worden besproken in Krieg en Smeets (2009).

7.2.2 Toepasbaarheid

Het voornaamste doel is het detecteren van invloedrijke fouten, die bij het microgaafmaken over het hoofd zijn gezien of zijn geïntroduceerd. De verdelingsmethode is bruikbaar voor kwantitatieve variabelen. Als de verdeling van de variabelen niet symmetrisch is dan is het beter om de data eerst te transformeren, zodat deze meer lijkt op een normale verdeling. Diverse verdelingsmethoden kunnen anders een vertekend beeld geven.

7.2.3 Uitgebreide beschrijving

Voor uitbijterdetectie kan gebruik worden gemaakt van diverse statistische maten (Projectgroep Mesoanalyse, 2009). Deze geven inzicht in de verdeling van de microdata en kunnen gebruikt worden om opvallende en/of structurele wijzigingen in de microdata vast te stellen.

Op zichzelf doen statistische maten geen uitspraak over de kwaliteit van de data. Als er bijvoorbeeld een grote spreiding is dan hoeft dit niet te zeggen dat de kwaliteit slecht is. Als de spreiding in een publicatiecel aanzienlijk groter is ten opzichte van eerdere perioden, dan kan dit wel aangeven dat de kwaliteit minder is.

Onderstaande maten hebben betrekking op een publicatiecel of een deel hiervan.

<u>Representatieve waarde (van een variabele) in de respons:</u>

- *Gemiddelde*. Deze wordt vaak gebruikt, maar is gevoelig voor uitbijters.
- *Voortschrijdend gemiddelde*. Deze wordt bijvoorbeeld gebruikt bij de berekening van de meetlat bij Internationale Handel. Hierbij worden de waarden voor eerdere (gaafgemaakte) perioden meegenomen.

<u>Robuuste representatieve waarde in de respons:</u>

- *Getrunceerd gemiddelde*. Als de waarde van een variabele kleiner is dan *c* of groter dan *d* dan waarde verwijderen. Bereken vervolgens het gemiddelde. Dit kan eenzijdig (alleen een onder- of bovengrens) of tweezijdig.
- *Gecensureerd gemiddelde*. Als de waarde van een variabele kleiner is dan *c* dan waarde=*c*. Als de waarde van een variabele groter is dan *d* dan waarde=*d*. Bereken vervolgens het gemiddelde. Dit kan eenzijdig (alleen een onder- of bovengrens) of tweezijdig.
- *Mediaan*. Dit is de middelste waarde van op de variabele gesorteerde data. Uitermate robuust tegen uitbijters, vooral als er evenveel te kleine waarden als te grote waarden zijn.

<u>Maat voor spreiding (van een variabele) in de respons:</u>

- *Variantie*. Deze maat wordt zeer vaak gebruikt, maar is gevoelig voor uitbijters en heeft niet dezelfde schaal als het gemiddelde.
- *Standaard deviatie (s.d.)*. Ook gevoelig voor uitbijters, maar heeft dezelfde schaal als het gemiddelde. Deze maat is gelijk aan de wortel van de variantie.
- *Bereik*. Het verschil tussen de minimum en maximum waarde.

<u>Robuuste maat voor spreiding in de respons:</u>

- *Interkwartielafstand*. Het verschil tussen het eerste kwartiel en derde kwartiel van de cumulatieve verdeling van een variabele. Deze wordt gebruikt bij boxplots. Deze en onderstaande maat zijn handig bij symmetrische verdelingen.
- *(100-α)% percentiel minus het α% percentiel*. α=25 geeft de interkwartielafstand. Als er veel waarnemingen zijn dan is het handiger om bijvoorbeeld α=5 te nemen.

- *Derde kwartiel minus tweede kwartiel*. Dit is een spreidingsmaat voor de waarden die groter zijn dan de mediaan. Net als onderstaande maat goed bruikbaar bij een scheve verdeling.
- *Tweede kwartiel minus eerste* kwartiel. Dit is een spreidingsmaat voor de waarden die kleiner zijn dan de mediaan (tweede kwartiel).

Maat voor spreiding van de schatting van een populatiekenmerk:

- *Standaardfout van de schatter*. Als deze niet analytisch kan worden bepaald dan kan deze empirisch worden bepaald met bootstrap of jackknife technieken.

Overige verdelingskenmerken:

- *Minimum*. De minimale waarde van een variabele in de respons. Als deze bijvoorbeeld negatief is en de variabele kan alleen positief zijn dan betekent dit dat een deel van de data inconsistent is.
- *Maximum*. De maximale waarde van een variabele in de respons. Als deze bijvoorbeeld extreem groot is dan betekent dit dat er zeker één waarde verdacht is.
- *Scheefheid*. Deze geeft aan in hoeverre de verdeling van een variabele asymmetrisch is. Als de rechterstaart van de verdeling langer is dan de linkerstaart dan is de scheefheid positief. Als het omgekeerde geldt dan is de scheefheid negatief. Scheefheid kan worden veroorzaakt door uitbijters.
- *Kurtosis*. Deze is hoog als de staarten van de verdeling relatief lang zijn. Uitbijters leiden tot een hoge kurtosis.

Voor het betrouwbaar bepalen van een verdelingskenmerk dienen er voldoende records beschikbaar te zijn. Voor het nauwkeurig bepalen van een interkwartielafstand zijn bijvoorbeeld meer dan twintig waarden nodig. Een verdelingskenmerk wordt vooral interessant als we deze combineren met andere verdelingskenmerken of hetzelfde kenmerk vergelijken voor verschillende perioden. Het doel is om meer inzicht te krijgen in de ontwikkeling in publicatiecijfers.

Voor ieder verdelingskenmerk geldt dat het interessant is om deze te vergelijken met hetzelfde kenmerk voor een eerdere periode. Als het kenmerk sterk is gewijzigd dan is dat verdacht. Interessante combinaties van verdelingskenmerken zijn:

- bereik gedeeld door interkwartielafstand
- gemiddelde gedeeld door mediaan
- gemiddelde gedeeld door getrunceerd/gecensureerd gemiddelde
- variantie *t* gedeeld door variantie *t-1*

Als een van deze groot is dan is er minimaal één uitbijter.

Om uitbijters te bepalen worden ook vaak grafische hulpmiddelen, zoals scatter- en boxplots, gebruikt. Deze Exploratieve Data Analyse (EDA) technieken zijn breed toepasbaar (Tukey, 1977) en beschikbaar via bijvoorbeeld Excel en SPSS. DesJardins (1997) illustreert toepassingen van en nuttige aanvullingen op traditionele EDA technieken.

Er zijn ook wiskundige technieken om uitbijters te detecteren zoals de Mahalanobis afstand, zie Hoogland, Houbiers en De Waal (2002). Om een beeld te krijgen van de

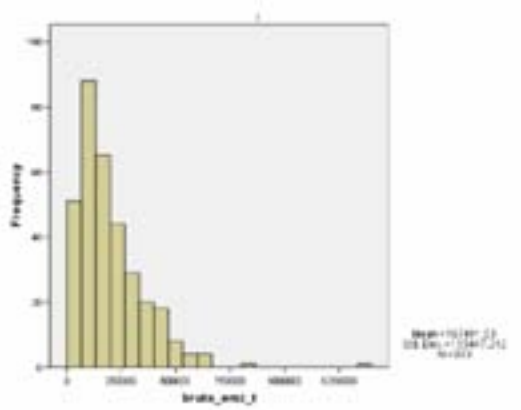
samenhang tussen twee variabelen kunnen regressietechnieken gebruikt worden. Standaard regressietechnieken kunnen een vertekend beeld geven van de samenhang als er uitbijters zijn. In dat geval kan er beter gebruik worden gemaakt van robuuste regressietechnieken. Er zijn diverse robuuste regressietechnieken, zoals M-schatters (Huber, 1981), de kleinste mediaan van kwadraten methode (Rousseeuw, 1984), de herwogen kleinste kwadraten methode (Rousseeuw en Leroy, 1987) en gegeneraliseerde S-schatters (Croux, Rousseeuw en Hössjer, 1994). Een aantal van deze technieken is beschikbaar in R, S-Plus, STATA en Matlab. M-schatters verkleinen de invloed van uitbijters, maar één uitbijter kan nog steeds voldoende zijn om deze schatters te verstoren. Een aantal technieken zijn robuust tegen uitbijters in de afhankelijke variabele, maar niet robuust tegen uitbijters in de predictoren.

7.2.4 Voorbeeld

Op het CBS worden steeds vaker belastinggegevens ingezet voor het maken van bedrijfsstatistieken. We zijn bijvoorbeeld geïnteresseerd in de jaar-op-jaar ontwikkeling van BTW-kwartaalomzetten. Op recordniveau wordt voor het gaafmaken gekeken naar de BTW-omzet en de groeivoet, d.w.z. de BTW-omzet in de verslagperiode gedeeld door de BTW-omzet in de referentieperiode.

Om een beeld te krijgen van de verdeling van de BTW-data kan er een histogram worden gemaakt, zie figuur 4. Er kunnen ook allerlei kenmerken van de verdeling worden bepaald, zie tabel 8. De interkwartielafstand is een robuuste maat voor de spreiding. Dat wil zeggen dat deze ongevoelig is voor uitbijters. Uit tabel 8 blijkt dat er minimaal één negatieve waarde voorkomt in gk 21 en 30 en dat de maximale omzet in gk 10 opvallend hoog is. Tevens blijkt dat de BTW-omzet scheef verdeeld is voor gk 10-30. Dit kan een gevolg zijn van uitbijters, zeker bij een scheefheid groter dan 2.5 of kleiner dan -2.5. Een kurtosis groter dan 10 duidt ook op uitbijters.

Figuur 4. Histogram van BTW-kwartaalomzetten voor 4e kwartaal 2008 voor kerncel 47110; gk 22

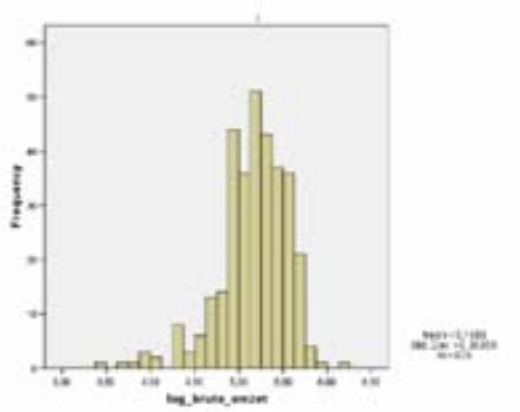


Tabel 8. Minimum, maximum, 1e, 2e en 3e kwartiel, interkwartielafstand (IKA), scheefheid en kurtosis van BTW-omzet (in duizend euro's) van kwartaalomzetten supermarkten in 4e kwartaal 2008

GK	Min.	Max.	Kwart 1	Kwart 2	Kwart 3	IKA	Scheefh.	Kurt.
10	0	2569	14	34	62	48	16,5	320,1
21	-177	1663	33	69	117	84	7,8	85,3
22	0	1349	91	163	273	182	2,0	9,1
30	-74	1873	212	331	492	280	2,1	6,9
40	0	2609	478	778	1343	865	0,6	-0,3
50	0	5245	1651	1960	2715	1064	0,5	1,3

Als de data zeer scheef verdeeld zijn dan is het is raadzaam om eerst een logaritmische transformatie toe te passen voordat er grafische analyses plaatsvinden. Dit zal vooral gelden voor de groeivoet. Na deze transformatie kunnen scheef verdeelde data meer symmetrisch verdeeld zijn, afgezien van uitbijters. In figuur 5 is dit te zien voor de BTW-kwartaalomzet van een aantal supermarkten. Het valt nu ook op dat een tiental omzetten per grootteklasse relatief klein is.

Figuur 5. Histogram van logaritme van BTW-kwartaalomzetten voor 4e kwartaal 2008 voor kerncel 47110; gk 22

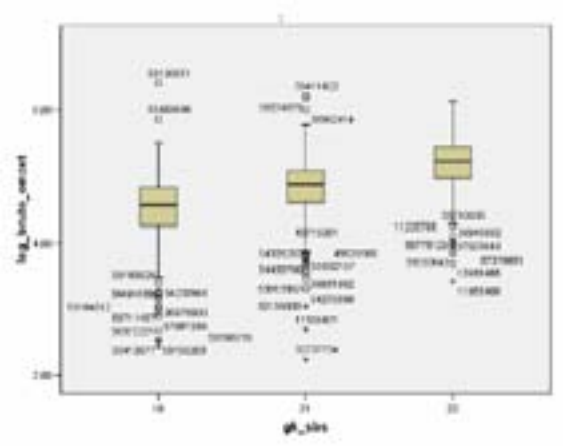


Met ondergenoemde grafieken kunnen verdachte waarden worden opgespoord.

- histogram van de waarden in een publicatiecel of stratum voor jaar *t*
- scatterplot (2-dimensionale puntenwolk) van de waarde of groeivoet voor jaar *t* versus de waarde of groeivoet voor jaar *t-1*. In dit geval zien we alleen eenheden waarvoor in de jaren *t* en *t-1* een waarde of groeivoet beschikbaar is voor de verslagperiode.
- twee boxplots (1-dimensionale plot om uitbijters te detecteren) naast elkaar: één met de waarde of groeivoet voor jaar *t* en één met de waarde of groeivoet voor jaar *t-1*. Eventueel aangevuld met boxplots voor eerdere jaren.

In figuur 6 staat een voorbeeld gemaakt met SPSS met BTW-omzet na een logtransformatie. Dit betreft boxplots voor supermarkten in een drietal grootteklassen. De ‘*’ zijn extreme uitbijters, de ‘o’ minder extreme uitbijters. Iedere uitbijter kun je een waarde meegeven (bijvoorbeeld be_id) om het record terug te kunnen vinden in de microdata. Het grote aantal uitbijters aan de onderkant van een boxplot geeft aan dat de data na logtransformatie ook niet symmetrisch verdeeld zijn en dat we nu juist een verdeling hebben met een “lange linkerstaart”.

Figuur 6. Boxplots per gk van log10(bruto omzet) van supermarkten



8. Literatuur

Bankier, M., J.-M. Fillion, M. Luc en C. Nadeau, 1994, Imputing numeric and qualitative variables simultaneously. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 242-247.

Bankier, M., 2006, *Imputing numeric and qualitative variables simultaneously*. Memo, Statistics Canada, Social Survey Methods Division.

CANCEIS, 2006, *CANCEIS version 4.5 user's guide*. Statistics Canada, Social Survey Methods Division.

Croux C., P. Rousseeuw en O. Hössjer, 1994, Generalized S-estimators. *Journal of the American Statistical Association* 89, 1271-1281.

Daalmans, J., 2000, *Automatic error localisation of categorical data*. Research paper 0024, Centraal Bureau voor de Statistiek, Voorburg.

DesJardins, D., 1997, *Experiences with introducing new graphical techniques for the analysis of census data*. UNECE Work Session on Statistical Data Editing, Praag.

Di Zio, M., U. Guarnera en O. Luzi, 2005, *Improving the effectiveness of a probabilistic editing strategy for business data*. ISTAT, Rome

Duin, C. van, 2003, *Plausibiliteitsindicator voor Impect 2*. Interne nota (BPA-nummer 2243-03-TMO), Centraal Bureau voor de Statistiek, Voorburg.

Fellegi, I.P. en D. Holt, 1976, A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* 71, 17-35.

Granquist, L. en J. G. Kovar, 1997, Editing of Survey Data: How Much Is Enough? In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz en D. Trewin (eds), *Survey Measurement and Process Quality*. New York: Wiley, pp. 415-435.

Haar, M. ter, 2002, *IMPECT 2: automatische correctie, versie 0.2 (rev. 1)*. Intern document, Centraal Bureau voor de Statistiek, Heerlen.

Hedlin, D., 2003, Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics* 19, 177-199.

Hidiroglou, M. A. en J.-M. Berthelot, 1986, Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* 12 (1), 73-83.

Hoogland, J., 2002, *Selective editing by means of Plausibility Indicators*. UNECE Work Session on Statistical Data Editing, Helsinki, working paper no. 33.

Hoogland, J., M. Houbiers en T. de Waal, 2002, *Syllabus bij de cursus gaafmaakmethoden en software voor bedrijfseconomische statistieken. Versie 2*. Interne nota (BPA-nummer 531-02-TMO), Centraal Bureau voor de Statistiek, Voorburg.

Hoogland, J. en R. Smit, 2008, *Selective automatic editing of mixed mode questionaires for structural business statistics*. UNECE Work Session on Statistical Data Editing, Wenen, working paper no. 2.

Huber, P., 1981, *Robust statistics*. Wiley, New York.

ISTAT, CBS en SFSO, 2007, *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys* (http://edimbus.istat.it/EDIMBUS1/document/RPM_EDIMBUS).

Jong, A. de 2002, *UNI-EDIT: Standardized processing of structural business statistics in the Netherlands*. UNECE Work Session on Statistical Data Editing, Helsinki, working paper no. 27.

Krieg, S. en M. Smeets, 2009, *Representatieve uitbijters*. Rapport Methodenreeks, Centraal Bureau voor de Statistiek, Heerlen.

Latouche, M. en J.-M. Berthelot, 1992, Use of a score function to prioritise and limit recontacts in editing business surveys. *Journal of Official Statistics* 8, 389-400.

Loo, M. van der en J. Pannekoek, 2007, *Advies gaafmaken en imputeren van de statistiek Bouwobjecten in Voorbereiding*. Interne nota (BPA-nummer DMK-2007-07-10-JPNK), Centraal Bureau voor de Statistiek, Voorburg.

Loo, M. P. J. van der, 2008, *An analysis of editing strategies for mixed-mode establishment surveys*. Discussion paper (08004), Centraal Bureau voor de Statistiek, Voorburg.

Pannekoek, J. en C. Tempelman, 2005, *Evaluatie van imputatiemethoden voor IMPECT: deductieve imputatie en correctie voor overduidelijke fouten*. Interne nota (BPA-nummer TMO-20050707-JPNK), Centraal Bureau voor de Statistiek, Voorburg.

Pannekoek, J., C. Harmsen, M. van Huis en K. Prins, 2008, *Automatisch gaafmaken van GBA-gegevens met de "Nearest-neighbor Imputation Methodology"*. Interne nota (BPA-nummer DMV-2008-10-2-JPNK), Centraal Bureau voor de Statistiek, Den Haag.

Pol, F. van de, F. Bakker en T. de Waal, 1997, *On principles for automatic editing of numerical data with equality checks*. Rapport (BPA-nummer 7141-97-RSM), Centraal Bureau voor de Statistiek, Voorburg.

Projectgroep Mesoanalyse, 2009, *HEcS+ Mesoanalyse: Analyse en interactief gaafmaken in de HEcS-keten Versie 0.1p3*. Interne nota, CBS, 1 mei 2009.

Rousseeuw, P., 1984, Least median of squares regression. *Journal of the American Statistical Association* 79, 871-880.

Rousseeuw, P. en A. Leroy, 1987, *Robust regression and outlier detection*. Wiley series in probability and mathematical statistics.

Scholtus, S., 2007, *Automatische correctie van tekenfouten en verwisselingen van baten en lasten*. Interne nota (BPA-nummer DMV-2007-12-14-SSHS), Centraal Bureau voor de Statistiek, Voorburg.

Scholtus, S., 2008a, *Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data*. Discussion paper (08015), Centraal Bureau voor de Statistiek, Den Haag.

Scholtus, S., 2008b, *Automatisch gaafmaken van GBA-gegevens met CANCEIS*. Interne nota (BPA-nummer DMV-2008-11-03-SSHS), Centraal Bureau voor de Statistiek, Den Haag.

Scholtus, S., 2009, *Automatic correction of simple typing errors in numerical data with balance edits*. Discussion paper (09046), Centraal Bureau voor de Statistiek, Den Haag.

Sluis, W., 2004, *SLICEDemo*. Interne nota, Centraal Bureau voor de Statistiek, Voorburg.

Stoop, J.-R., 2003, *Het lekkerste stuk uit CherryPie. Selectie van een optimale oplossing bij automatisch gaafmaken*. Interne nota (BPA-nummer 2098-03-TMO), Centraal Bureau voor de Statistiek, Voorburg.

Tukey, J., 1977, *EDA: Exploratory Data Analysis*. Addison-Wesley, Massachusetts.

Waal, T. de en R. Quere, 2003, A fast and simple algorithm for automatic editing for mixed data. *Journal of Official Statistics* 19, 383-402.

Waal, T. de, 2003, *Processing of erroneous and unsafe data*. Proefschrift, Erasmus Universiteit Rotterdam.

Waal, T. de, 2005a, *SLICE 1.5: Software voor automatisch gaafmaken en imputeren*. Interne nota (BPA nummer TMO-R&D-20050707-TWAL-4), Centraal Bureau voor de Statistiek, Voorburg.

Waal, T. de, 2005b, *De methodologie van SLICE 1.5: Het algoritme van Cherry Pie*. Interne nota (BPA nummer TMO-R&D-20050707-TWAL-2), Centraal Bureau voor de Statistiek, Voorburg.

Waal, T. de, 2008, *An overview of statistical data editing*. Discussion paper (08018), Centraal Bureau voor de Statistiek, Den Haag.