Centraal Bureau voor de Statistiek

Discussion Paper

Kwaliteitsmaten voor het datacorrectieproces

De weergegeven opvattingen zijn die van de auteur(s) en komen niet noodzakelijkerwijs overeen met die van het CBS

2014 | 08

Bart van den Broek, Mark van der Loo en Jeroen Pannekoek

Datacorrectie is gebaat bij indicatoren die op eenvoudige en duidelijke wijze de invloed van het correctieproces op de data aangeven. In dit rapport beschrijven we enkele indicatoren die op grafische wijze aspecten van verandering in data ten gevolge van een correctieproces uitlichten. indicatoren vallen uiteen in twee soorten, zij die betrekking hebben op waarden en zij die betrekking hebben op regelschendingen. We illustreren de indicatoren aan de hand van de statistieken van Welzijn en Kinderopvang en van Groothandel.

Inhoudsopgave

1 Inleiding

Het doel van een datacorrectieproces is om de data aan te passen zodanig dat dit de kwaliteit van de data verbetert. Inzicht in de status van de data en de invloed van het correctieproces daarop wordt verkregen door middel van indicatoren. Typisch voor een correctieproces is dat het de data verandert. Indicatoren toegespitst op datacorrectie zullen allerlei aspecten van verandering aangeven.

Een eerste bron van informatie met betrekking tot data correctie is de mate waarin waarden worden aangepast. Della Rocca et al. (2005); Brancato et al. (2009); Banning en Vink (2010) beschouwen fracties van waarden die zijn geïmputeerd, aangepast of geannuleerd. We demonstreren deze indicatoren door middel van een visuele weergave. De invloed van correctiemethoden zoals foutlokalisatie, die waarden annuleert, en deductieve, regressie en *nearest neighbour* imputatie wordt met deze indicatoren duidelijk zichtbaar.

Voor correctiemethoden die waarden aanpassen is het informatief om te kijken naar de invloed van zo'n methode op de waarden die variabelen of statistieken daarvan aannemen. Hedlin (2003) zet de waarde van een variabele voor en na toepassing van de correctiemethode tegen elkaar uit, en geeft hiermee patronen weer in de data ten gevolge van de correctie. Della Rocca et al. (2005) stelt verschillende afstandsmaten voor om de verandering in waarde van variabelen ten gevolge van correctiemethoden uit te drukken. Door waarden of statistieken van variabelen te schatten en de afwijking tot de werkelijke waarde te bepalen wordt ook een maat voor de kwaliteit van de data gegeven (Nordbotten, 1997; Hedlin, 2003). Hier demonstreren we indicatoren die voor variabelen de verandering van de gemiddelde waarde en de betrouwbaarheid daarvan weergeven over de correctiestappen. De relatieve bijdrage per correctiemethode is hieruit duidelijk af te lezen, net zoals het aandeel aan beïnvloede variabelen.

Variabelen in een statistiek kunnen op allerlei manieren en om allerlei redenen in verband staan met elkaar. Wanneer deze verbanden beschreven worden door regels waar de data aan heeft te voldoen, dan kan als onderdeel van een correctieproces op de geldigheid van regels gecontroleerd worden. Om de verandering in de data in termen van regelschendingen aan te geven beschouwen we indicatoren die veranderingen in het aantal schendingen tellen en indicatoren die de omvang van schendingen meten. De eerste is vergelijkbaar met de eerder genoemde indicatoren die gehaltes van geïmputeerde, aangepaste of geannuleerde waarden meten. Voor het softwarepakket R is het `editrules` pakket ontwikkeld waarmee ondermeer geverifieerd kan worden of data aan gegeven regels voldoet (De Jonge en Van der Loo, 2011, 2012; Van der Loo en De Jonge, 2011).

De omvang van regelschendingen drukken we uit in termen van een kortste afstand tussen het foutieve record en de records die wel aan de gegeven regel voldoen. Voor continue data, waar we ons hier met name op richten, is een veel gebruikte afstandsmaat de Minkowski-afstand (Della Rocca et al., 2005; Banning en Vink, 2010; De Waal et al., 2011). Voor deze afstand laten we zien dat de omvang van de schending van lineaire regels reduceert tot het verschil tussen een gewogen som van waarden en een gewenste waarde.

De indicatoren illustreren we aan de hand van twee statistieken, die van Welzijn en Kinderopvang en die van Groothandel. Beide statistieken zijn onderworpen aan een

correctieproces. Een korte beschrijving van deze statistieken en de stappen in het correctieproces geven we in sectie 2. Indicatoren die veranderingen in waarden uitdrukken beschouwen we in sectie **??**. Maten voor regelschendingen behandelen we in sectie 4. In sectie 5 passen we een maat voor regelschendingen toe op het vergelijken van verschillende gaafmaaksystemen voor de Welzijn en Kinderopvang statistiek.

Notatie

Een overzicht van de notaties die we gebruiken in dit rapport is gegeven in tabel 1.

symbool	definitie
N	aantal records
J	aantal variabelen
K	aantal regels
$\boldsymbol{X}$	dataset
$\boldsymbol{x}$	een data record
D	record waarde domein
NA	ontbrekende waarde
e	een regel
E	een verzameling van regels
$\boldsymbol{F}$	$N \times K$ matrix met elementen in $\{0, 1, \text{NA}\}$
$\boldsymbol{G}$	$N \times K$ matrix met elementen in $[0, \infty) \cup \{\text{NA}\}$
$i(\boldsymbol{x}, \boldsymbol{y})$	de donor-imputatie van record $\boldsymbol{x}$ met donor $\boldsymbol{y}$
$\boldsymbol{v}^\top$	de getransponeerde van een vector $\boldsymbol{v}$
$\|\boldsymbol{v}\|_p$	de L^p-norm $\left(\sum_j \|v_j\|^p\right)^{1/p}$ van een vector $\boldsymbol{v}$, $1 \leq p < \infty$
$\|\boldsymbol{v}\|_\infty$	de L^∞-norm $\max_j \|v_j\|$ van een vector $\boldsymbol{v}$
$\|\boldsymbol{v}\|_{\boldsymbol{w},p}$	de L^p-norm $\left(\sum_j \|v_j\|^p w_j\right)^{1/p}$ van een vector $\boldsymbol{v}$, $1 \leq p < \infty$, met gewichten $\boldsymbol{w} = (w_1, \ldots, w_J)$
$\|\boldsymbol{v}\|_{\boldsymbol{w},\infty}$	de L^∞-norm $\max\{\|v_j\|w_j : j = 1, \ldots, J \text{ en } w_j > 0\}$ van een vector $\boldsymbol{v}$ met gewichten $\boldsymbol{w} = (w_1, \ldots, w_J)$
$\inf S$	de grootste ondergrens van een verzameling S van getallen
$\text{sgn}(x)$	het teken van een getal x; $\text{sgn}(x)$ is -1 als $x < 0$, 0 als $x = 0$, en $+1$ als $x > 0$
$x \vee y$	het maximum van x en y

Tabel 1 Gebruikte notaties.

Een record $\boldsymbol{x}$ is een rij van variabelen $x_1, \ldots, x_J$. Iedere variabele x_j kan waarden aannemen binnen een domein D_j. Als een variabele x_j geen waarde aanneemt maar een waarde ontbreekt, dan schrijven we $x_j = \text{NA}$ (``Not Available''). Het waarde domein waarbinnen een record $\boldsymbol{x}$ waarden aan kan nemen is precies het Cartesisch product $D = \prod_{j=1}^{J} D_j$ van de waarde domeinen $D_1, \ldots, D_J$ van de variabelen $x_1, \ldots, x_J$.

2 Beschrijving data en correctiestappen

Ter illustratie van de indicatoren in dit document zullen we gebruik maken van twee statistieken, die van Welzijn en Kinderopvang en die van Groothandel. Beide statistieken

Correctie stap	aanpassen beschikbare waarden	introduceren ontbrekende waarden	invullen ontbrekende waarden
ruwe data			
afleiden var.			
mutaties	×		
duizendfouten regels	×		
duizendfouten ref. data	×		
tikfouten	×		
afrondfouten	×		
nonresponse		×	
foutlokalisatie		×	
NN-imputatie			×
deductieve imputatie	×		
aanpassen	×		

Tabel 2 Acties correctiestappen in Welzijn en Kinderopvang statistiek.

bestaan uit ruwe data die onderworpen is aan een correctieproces van meerdere stappen.

De statistiek Welzijn en Kinderopvang bestaat uit 840 records van ieder 67 variabelen. Het toegepaste correctieproces bestaat uit een aantal stappen. We geven nu een korte beschrijving van ieder van deze stappen. Een beknopt overzicht in termen van acties per correctiestap staat gegeven in tabel 2. Voor een meer uitgebreide beschrijving verwijzen we naar Van der Loo en Pannekoek (2013).

1. Afleiden var: Nieuwe variabelen worden afgeleid die verderop in het proces gebruikt worden. Dit betreft variabelen die lineaire combinaties zijn van bestaande variabelen, en een variabele (``duizendfoutenindicator'') die aangeeft of het record één of meerdere duizendfouten bevat. Deze laatste variabele wordt verderop in het proces gebruikt om duizendfouten te herstellen.
2. Mutaties: Waargenomen waarden worden vervangen op basis van eenvoudige kennisregels. Een voorbeeld hiervan is de volgende regel:

   ```
   if ( v17 == v126 & v126 != 0 ) v126 = 0
   ```

   Deze regel zegt dat als ``Bedrijfsresultaat'' (`v17`) gelijk is aan ``Saldo boekwinsten/verliezen'' (`v126`) en ``Saldo boekwinsten/verliezen'' is ongelijk aan nul, dan moet ``Saldo boekwinsten/verliezen'' gelijk worden gesteld aan nul.
3. Duizendfouten regels: Duizendfouten (waarden die een factor duizend verkeerd zijn doordat ze in de verkeerde eenheid zijn opgegeven) worden gecorrigeerd met behulp van de onder ``Afleiden var.'' geïntroduceerde ``duizendfoutenindicator''. Op basis van kennisregels worden per record variabelen aangewezen die mogelijk duizendfouten bevatten en deze worden in waarde door duizend gedeeld.
4. Duizendfouten ref. data: Op basis van gegevens uit een eerdere periode worden per record variabelen aangewezen die mogelijk duizendfouten bevatten en deze worden door duizend gedeeld.

Correctie stap	aanpassen beschikbare waarden	introduceren ontbrekende waarden	invullen ontbrekende waarden
ruwe data			
negatieve waarden	×		
duizendfouten	×		
deterministisch	×		×
schrijffouten	×		
tekenfouten	×	×	×
afrondfouten	×		
foutlokalisatie		×	
handmatig	×		×
deductieve imputatie			×
regressie imputatie			×
aanpassen	×		

Tabel 3 Acties per correctiestap in de Groothandel statistiek.

5. Tikfouten: Tikfouten worden opgespoord en gecorrigeerd met behulp van de methode zoals beschreven in de Methodenreeks, thema Controle & Correctie (versie 25 januari 2010) Hoofdstuk 2.4.6.
6. Afrondfouten: Afrondfouten worden opgelost met behulp van een algoritme van Scholtus (2008).
7. Nonresponse: In het verkrijgen van de ruwe data van Welzijn en Kinderopvang zijn alle ontbrekende waarden met nul ingevuld. Deze stap bepaalt op basis van kennisregels of waarden terecht gelijk aan nul zijn en zet als foutief aangewezen waarden op ontbrekend. Overigens wordt hierbij geen rekening gehouden met of een waarde door de respondent is ingevuld danwel om bovengenoemde reden met nul is geïmputeerd.
8. Foutlokalisatie: In records die niet aan alle regels voldoen worden een minimum aantal velden op leeg gezet zodanig dat ze met alle regels consistent kunnen worden ingevuld. Dit is het principe van Fellegi en Holt (1976). De hiervoor gehanteerde methode staat beschreven in de Methodenreeks, thema Controle & Correctie (versie 25 januari 2010), Hoofdstuk 5.
9. Deductieve imputatie: Op basis van regels waar de data aan moet voldoen, worden variabelen deductief geïmputeerd.
10. NN-imputatie: Lege waarden worden ingevuld door ze over te nemen uit records die sterk overeen komen wat ingevulde waarden betreft.
11. Aanpassen: Na het invullen van ontbrekende waarden met waarden van sterk gelijkende records voldoen de ingevulde records meestal niet aan alle controleregels. De overgenomen waarden worden hier aangepast zodat wel aan alle regels is voldaan.

De Groothandel statistiek bestaat uit 323 records van ieder 89 variabelen. De stappen die uitgevoerd worden in het correctieproces voor Groothandel zijn hieronder beschreven. Een beknopt overzicht in termen van acties per correctiestap staat gegeven in tabel 3.

1. Negatieve waarden: Variabelen die niet negatief horen te zijn, maar die toch een negatieve waarde hebben, worden op nul gezet. Een uitzondering is de variabele ``Dotaties

voorzieningen", van deze wordt de absolute waarde genomen indien de waarde negatief is.

2. Duizendfouten: In de records met duizendfouten worden financiële variabelen in waarde gedeeld door duizend en vervolgens afgerond op nul decimalen.
3. Deterministisch: De data wordt gecorrigeerd aan de hand van deterministische regels waar de data aan moet voldoen. Dit gebeurt met behulp van de functie `applyRules` uit het `deducorrect` R-package.
4. Schrijffouten: Records die niet aan gegeven lineaire gelijkheden voldoen worden indien mogelijk gecorrigeerd door eenvoudige schrijffouten op te lossen. Dit gebeurt met behulp van de functie `correctTypos` uit het `deducorrect` R-package, wat een implementatie is van een algoritme van Scholtus (2009).
5. Tekenfouten: Records die niet aan gegeven lineaire gelijkheden voldoen worden voor zover mogelijk gecorrigeerd door waarden om te wisselen of van teken te veranderen. Dit gebeurt met behulp van een algoritme zoals beschreven door Scholtus (2008) en geïmplementeerd in de functie `correctSigns` uit het `deducorrect` R-package. Het omwisselen van waarden is ook van toepassing op ontbrekende waarden, zodat er effectief ontbrekende waarden geïntroduceerd en ingevuld kunnen worden.
6. Afrondfouten: Records die gegeven lineaire (on)gelijkheden schenden worden opgespoord en indien mogelijk gecorrigeerd op afrondfouten. Dit gebeurt met behulp van de functie `correctRounding` uit het `deducorrect` R-package wat een implementatie is van een algoritme voor afronden (Scholtus, 2008).
7. Foutlokalisatie: Fouten worden gelokaliseerd op basis van het principe van Fellegi en Holt (1976).
8. Handmatig: Records worden handmatig gecorrigeerd.
9. Deductieve imputatie: Op basis van gegeven regels waar de data aan moet voldoen, worden variabelen deductief geïmputeerd. Dit gebeurt met behulp van de functie `deduImpute` uit het `deducorrect` R-package.
10. Regressie imputatie: Ontbrekende waarden worden geïmputeerd door middel van multivariate regressie.
11. Aanpassen: Records worden aangepast om aan gegeven lineaire (on)gelijkheden te voldoen. Dit gebeurt met behulp van de functie `adjustRecords` uit het `rspa` R-package (Van der Loo, 2013).

3 Verandering van waarden

3.1 Verandering op het niveau van de gehele data

totaal				
totaal beschikbaar			totaal ontbrekend	
nog steeds beschikbaar		geïmputeerd	op ontbrekend gezet	nog steeds ontbrekend
beschikbaar, onaangepast	beschikbaar, aangepast			

Tabel 4 Onderverdeling van de status van cellen.

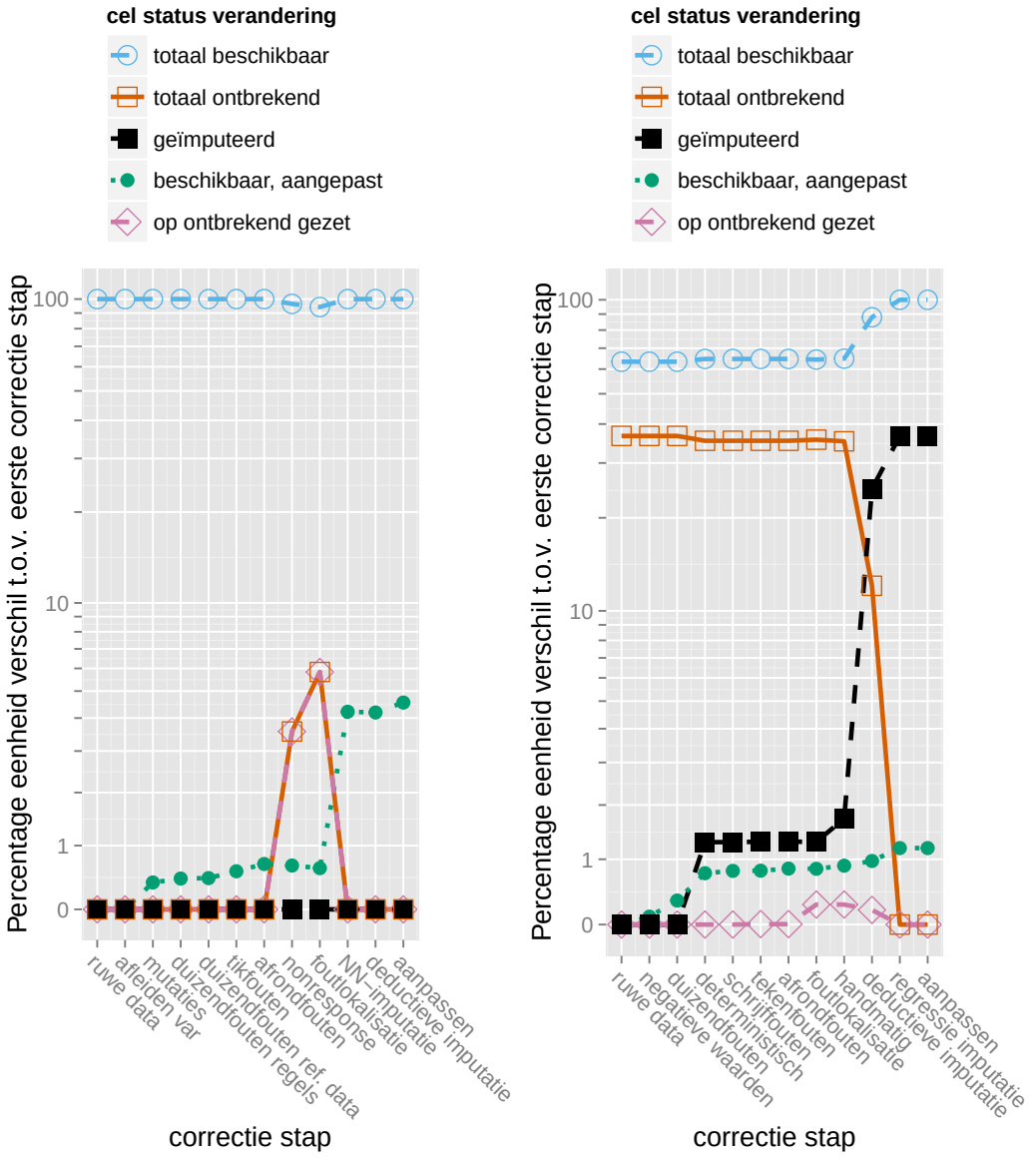
Gedurende een correctieproces worden, indien nodig, de waarden in cellen in de data aangepast. Op ieder moment in het correctieproces kunnen we aan iedere cel een status toekennen, op basis van de geschiedenis van die cel. Een onderverdeling van de status van cellen is gegeven in tabel 4. Het totaal aantal cellen is op te splitsen in die cellen met beschikbare waarden en die waarvoor de waarden ontbreken. Cellen waarvoor de waarden beschikbaar zijn, zijn weer verder onder te verdelen in cellen met geïmputeerde danwel nog steeds beschikbare waarden, de laatste is weer onder te verdelen in cellen met aangepaste danwel onaangepaste waarden. Cellen waarvoor de waarden ontbreken zijn verder onder te verdelen in cellen waarvoor de waarden nog steeds ontbreken danwel waarvoor de waarden op ontbrekend zijn gezet.

De verandering van de status van cellen over de stappen in het correctieproces is gegeven in de tabellen 5 en 6, dit zijn aantallen en de verandering is ten opzichte van de ruwe data. Een grafische weergave van deze verandering van de status van cellen wordt gegeven in figuur 1(a) voor de Welzijn en Kinderopvang statistiek en in figuur 1(b) voor de Groothandel statistiek. De weergave is nu in procenten en op een pseudologaritmische schaal. Dit is een schaal die bij benadering lineair is rond nul en bij benadering logaritmisch daarbuiten. Ze wordt verkregen uit een lineaire schaal door middel van de volgende transformatie:

$$\text{pseudolog}(x) = \frac{\text{arsinh}(x/2)}{\log(10)} = \frac{\log(x + \sqrt{x^2 + 1})}{\log(10)} \tag{1}$$

waarbij log de natuurlijke logaritme is.

In figuur 1(a) zie we dat in de ruwe data in de Welzijn en Kinderopvang statistiek geen waarden ontbreken: de curven voor percentages totaal ontbrekende waarden en totaal beschikbare waarden beginnen op 0 respectievelijk 100 procent. Alleen in de stappen ``nonresponse'' en ``foutenlokalisatie'' tonen deze curves aan dat er ontbrekende waarden zijn. De stappen ``nonresponse'' en ``foutlokalisatie'' introduceren ontbrekende waarden, en ``NN-imputatie'' brengt het percentage ontbrekende waarden weer terug tot nul. Het percentage op ontbrekend gezette waarden is hier gelijk aan het percentage totaal ontbrekende waarden, aangezien de hoeveelheid op ontbrekend gezette waarden relatief ten opzichte van de ruwe data is, en de bijbehorende curves vallen derhalve samen. Om soortgelijke redenen is de curve van het percentage geïmputeerde waarden overal nul. De curve van het percentage beschikbare, aangepaste waarden laat zien dat de stappen tot aan ``foutlokalisatie'' minder dan één procent van de cellen in waarde aanpassen. De stappen ``nonresponse'' en ``foutlokalisatie'' zorgen voor een lichte daling in het aantal beschikbare, aangepaste waarden, zie tabel 5. Dit betekent dat een klein aantal waarden terug is gezet op hun oorspronkelijke waarden zoals deze waren in



(a) Welzijn en Kinderopvang **(b) Groothandel**

Figuur 1 Statusverdeling van cellen over correctiestappen met betrekking tot (a) de Welzijn en Kinderopvang statistiek en (b) de Groothandel statistiek. De hoeveelheden zijn weergegeven in percentages op een pseudologaritmische schaal. Percentages geïmputeerde, aangepaste en op ontbrekend gezette cellen zijn relatief ten opzichte van de ruwe data.

Correctie stap	totaal	totaal beschikbaar	beschikbaar, aangepast	geïmputeerd	op ontbrekend gezet	totaal ontbrekend
ruwe data	56280	56280	0	0	0	0
afleiden var	56280	56280	0	0	0	0
mutaties	56280	56280	230	0	0	0
duizendfouten regels	56280	56280	264	0	0	0
duizendfouten ref. data	56280	56280	268	0	0	0
tikfouten	56280	56280	328	0	0	0
afrondfouten	56280	56280	393	0	0	0
nonresponse	56280	54273	379	0	2007	2007
foutlokalisatie	56280	52999	356	0	3281	3281
NN-imputatie	56280	56280	2374	0	0	0
deductieve imputatie	56280	56280	2359	0	0	0
aanpassen	56280	56280	2562	0	0	0

Tabel 5 Veranderingen in de status van cellen over correctiestappen voor de Welzijn en Kinderopvang statistiek. Veranderingen zijn ten opzichte van de ruwe data.

de ruwe data. ``NN-imputatie'' is vervolgens verantwoordelijk voor een verdere toename tot vier procent aan cellen die in waarde zijn aangepast ten opzichte van de ruwe data.

In figuur 1(b) zien we dat de ruwe data in de Groothandel statistiek wel ontbrekende waarden bevat: de curven voor percentages totaal ontbrekende waarden en totaal beschikbare waarden beginnen op ongeveer 36 respectievelijk 64 procent. Uit deze curves lezen we ook een daling af in het aantal ontbrekende waarden tot nul procent ten gevolge van ``deductieve imputatie'' en ``regressie imputatie''. Een introductie van ontbrekende waarden vindt plaats als het resultaat van de stappen ``tekenfouten'' en ``foutlokalisatie''. In de stap ``tekenfouten'' worden twee ontbrekende waarden geïntroduceerd, dit lezen we af uit tabel 6. Het is het gevolg van het omwisselen van waarden, wat ook tot de mogelijkheden behoort in de correctiestap ``tekenfouten''. In twee records wordt de waarde van variabele ``Vrijval voorz.'', die in deze records ontbreekt, omgewisseld met die van variabele ``Dotaties voorz.'', die in deze records wel een waarde heeft. Hoewel de omwisseling geen invloed heeft op het totaal aantal ontbrekende waarden worden er wel nieuwe ontbrekende waarden geïntroduceerd en net zo veel lege velden ingevuld. Een groter aantal ontbrekende waarden wordt geïntroduceerd ten gevolge van ``foutlokalisatie'', de curve voor het aantal op ontbrekend gezette waarden laat dit zien, als ook dat alle op ontbrekend gezette waarden weer geïmputeerd worden door ``deductieve imputatie'' en ``regressie imputatie''. De correctiestappen met het grootste aandeel in imputatie van ontbrekende waarden zijn duidelijk af te lezen uit de curve die het percentage geïmputeerde waarden weer geeft; deterministische en handmatige correctie leveren een kleine bijdrage, de grootste bijdragen worden geleverd door ``deductieve imputatie'' en ``regressie imputatie''. Uit de curve van het percentage beschikbare, aangepaste waarden is af te lezen dat iets meer dan één procent van de beschikbare waarden in waarde wordt aangepast, en dat dit met name gebeurt door correctie op duizendfouten, deterministische correctie en ``regressie imputatie''.

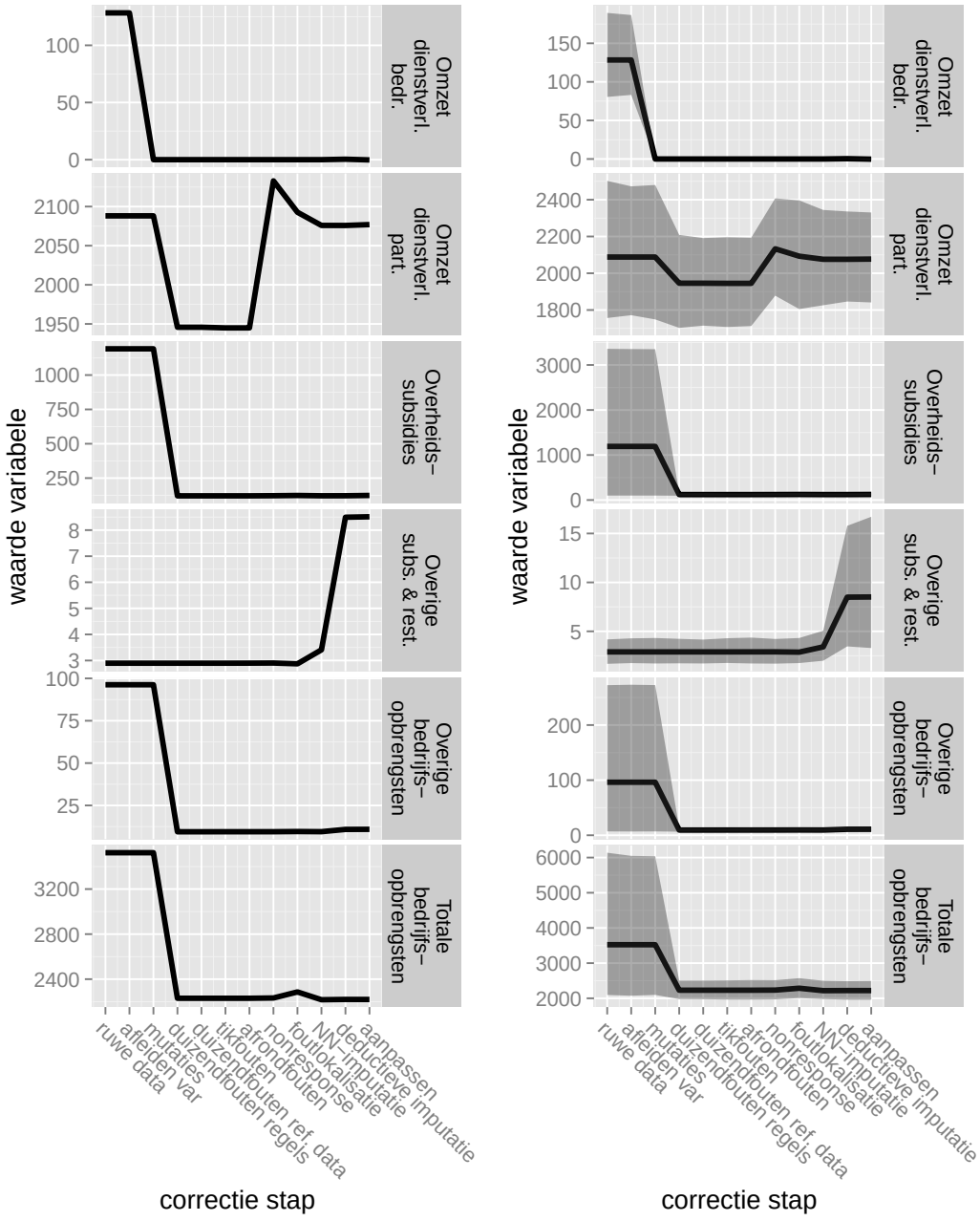
Correctie stap	totaal	totaal beschikbaar	beschikbaar, aangepast	geïmputeerd	op ontbrekend gezet	totaal ontbrekend
ruwe data	30039	19033	0	0	0	11006
negatieve waarden	30039	19033	35	0	0	11006
duizendfouten	30039	19033	107	0	0	11006
deterministisch	30039	19421	233	388	0	10618
schrijffouten	30039	19421	244	388	0	10618
tekenfouten	30039	19421	245	390	2	10618
afrondfouten	30039	19421	255	390	2	10618
foutlokalisatie	30039	19334	254	390	89	10705
handmatig	30039	19462	269	518	89	10577
deductieve imputatie	30039	26414	292	7446	65	3625
regressie imputatie	30039	30039	357	11006	0	0
aanpassen	30039	30039	357	11006	0	0

Tabel 6 Veranderingen in de status van cellen over correctiestappen voor de Groothandel statistiek. Veranderingen zijn ten opzichte van de ruwe data.

3.2 Verandering op het niveau van variabelen

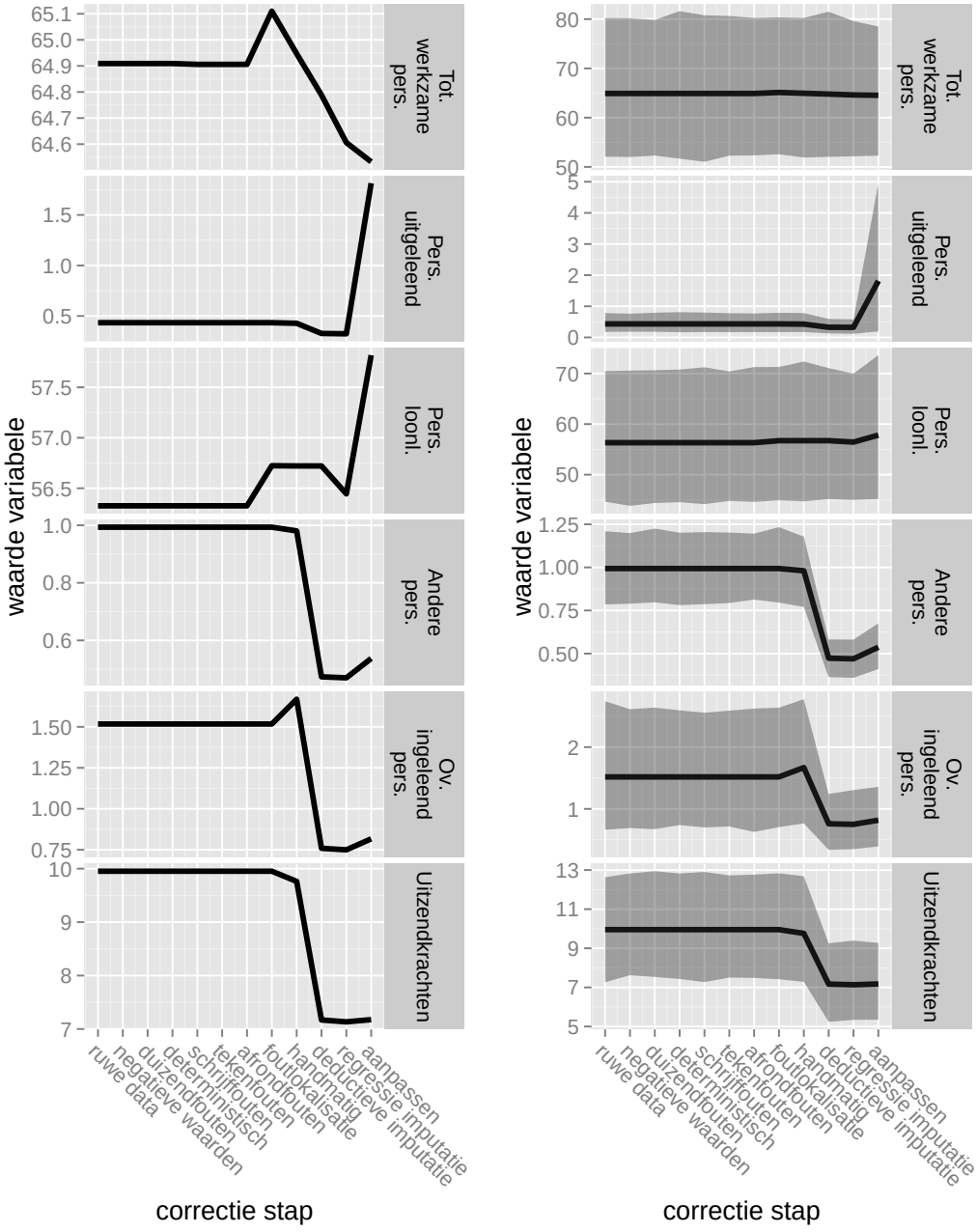
De bijdrage van iedere afzonderlijke stap in een correctieproces is af te lezen aan hoe het de waarden van variabelen in de data beïnvloed. Het verloop van enkele variabelen gedurende een correctieproces is gegeven in figuren 2 en 3 voor de statistieken Welzijn en Kinderopvang respectievelijk Groothandel. De waarden zijn gemiddeld over de records in de data.

Kijken we enkel naar de gemiddelde waarden, figuren 2(a) en 3(a), dan zien we allerlei veranderingen ten gevolge van de correctiestappen. Deze gemiddelden zijn echter steekproefgrootheden die behept zijn met steekproefonnauwkeurigheid. Deze onnauwkeurigheid kunnen we weergeven door middel van betrouwbaarheidsintervallen rond die gemiddelden, zoals weergegeven in Figuur 2(b) en 3(b). De achtereenvolgende betrouwbaarheidsintervallen kunnen als volgt worden geïnterpreteerd. Als we het gemiddelde uitrekenen direct na waarneming, dus vóór de eerste correctiestap, dan is dat gemiddelde een zuivere schatter voor de gemiddelde direct geobserveerde waarden in de populatie. Dat wil zeggen, een zuivere schatter voor de werkelijke populatiewaarde die wordt verkregen door de waarden voor de hele populatie uit te vragen en alle fouten te laten zitten. Als we het gemiddelde berekenen na één correctiestap, krijgen we een zuivere schatter voor het populatiegemiddelde wat wordt verkregen door alle waarden uit te vragen en deze ene correctiestap uit te voeren. Door de betrouwbaarheidsintervallen op na elke stap uit te rekenen kan inzicht worden verkregen in het effect van correctiestappen op de precisie van de uiteindelijke schatting. De betrouwbaarheidsintervallen zijn berekend op basis van de boostrap methode. Bij het berekenen van de gemiddelden en betrouwbaarheidsintervallen is geen rekening gehouden met eigenschappen van het gebruikte steekproefontwerp. Zaken zoals ongelijke insluitkansen of stratificatie zijn genegeerd. Omdat het hier vooral gaat om methoden waarmee effecten van processtappen in kaart kunnen worden gebracht en onderling kunnen worden vergeleken volstaat deze benadering.



(a) Zonder betrouwbaarheidsinterval **(b) Met betrouwbaarheidsinterval**

Figuur 2 Het verloop gedurende correctie van de Welzijn en Kinderopvang statistiek van de gemiddelde waarde van variabelen ``Omzet dienstverl. bedr.'', ``Omzet dienstverl. part.'', ``Overheidssubsidies'', ``Overige subs. & rest.'', ``Overige bedrijfsopbrengsten'' en ``Totale bedrijfsopbrengsten''.



(a) Zonder betrouwbaarheidsinterval **(b) Met betrouwbaarheidsinterval**

Figuur 3 Het verloop gedurende correctie van de Groothandel statistiek van de gemiddelde waarde van variabelen ``Tot. werkzame pers.'', ``Pers. uitgeleend'', ``Pers. loonl.'', ``Andere pers.'', ``Ov. ingeleend pers.'' en ``Uitzendkrachten''.

Met behulp van de betrouwbaarheidsintervallen kunnen we grote veranderingen in de gemiddelde waarden opsporen, met inachtneming van de steekproefonnauwkeurigheid van de gemiddelden. Hiertoe definieren we een ``substantiële veranderingäls een verandering waarbij de betrouwbaarheidsintervallen voor en na de verandering geen overlap vertonen. De intervalschattingen bij de gemiddelden voor en na de verandering laten dan zien dat het zeer onwaarschijnlijk is dat de verandering aan steekproeffluctuaties geweten kan worden. Zo zien we bijvoorbeeld dat de variabelen ``Omzet dienstverl. bedr.'' aan bedrijven'' en ``Overheidssubsidies'' in figuur 2 en ``Andere pers.'' en ``Uitzendkrachten'' in figuur 3 wel in waarde veranderen, maar dat deze veranderingen niet substantieel zijn, in de hierboven gedefiniëerde zin, omdat de betrouwbaarheidsintervallen voor en na verandering elkaar overlappen. In het algemeen is het dus belangrijk ook te letten op de betrouwbaarheid van de gemiddelde waarden.

Naast de gemiddelde waarde wordt ook de breedte van het betrouwbaarheidsinterval beïnvloed door het correctieproces. Zo zien we in figuur 2(b) dat deze breedte voor de variabelen ``Overheidssubsidies'', ``Overige bedrijfsopbrengsten'' en ``Totale bedrijfsopbrengsten'' aanzienlijk daalt ten gevolge van correctie voor duizendfouten, wat duidt op de aanwezigheid van uitschieters waar door de correctie voor duizendfouten voor is gecorrigeerd.

Als we kunnen bepalen welke variabelen het meest beïnvloed worden door het correctieproces, en welke stappen in het proces daarbij de grootste bijdrage leveren, dan geeft dat de mogelijkheid om aan die variabelen of correctiestappen meer aandacht te schenken. We maken daartoe een overzicht van de verandering van de waarden voor alle variabelen en correctiestappen. Om een beeld te krijgen hoe zowel de gemiddelde waarde van variabelen als ook de betrouwbaarheid daarvan verandert, maken we een opsplitsing in de verandering van de gemiddelde waarde en de verandering van de breedte van het betrouwbaarheidsinterval. In het algemeen zullen de waarden van de variabelen op verschillende schalen liggen en daardoor niet direct met elkaar te vergelijken zijn. Door de verandering van de gemiddelde waarde en de breedte van het betrouwbaarheidsinterval daaromheen in procenten te beschouwen, kunnen we deze toch met elkaar vergelijken.

Figuren 4(a) en 5(a) geven de verandering in betrouwbaarheid van de variabelen in de statistieken Welzijn en Kinderopvang respectievelijk Groothandel: per variabele en per correctiestap is door middel van een kleurschaal weergegeven hoeveel het betrouwbaarheidsinterval om de gemiddelde waarde in percentage is toe- of afgenomen relatief ten opzichte van de ruwe data. De variabelen zijn gesorteerd op volgorde van de procentuele verandering van de breedte van het betrouwbaarheidsinterval ter hoogte van de laatste stap in het correctieproces (``aanpassen''). Figuren 4(b) en 5(b) geven substantiële veranderingen in de gemiddelde waarde van de variabelen weer: per variabele en per correctiestap is door middel van een kleurschaal weergegeven hoe groot de substantiële verandering van de gemiddelde waarde is. Hiervoor is, voor substantiële veranderingen, de mate van verandering in een percentage uitgedrukt. Dit ``percentage substantiële verandering''definiëren we aan de hand van een voorbeeld. In figuur 3(b) zien we de betrouwbaarheidsintervallen rond de gemiddelde waarde van variabele ``Andere pers.'' voor iedere stap in het correctieproces. Tot en met correctiestap ``handmatig'' zien we dat de betrouwbaarheidsintervallen overlap vertonen met die ter hoogte van de ruwe data, en om die reden is het percentage substantiële verandering gelijk aan nul. In de volgende stap, ``deductieve imputatie'', is er geen overlap meer: het betrouwbaarheidsinterval ligt veel lager, met een bovengrens op 0.58 wat 24 procent lager is dan de ondergrens van het

(a) Betrouwbaarheid **(b) Significante verandering**

Figuur 4 Variabelwaarden gedurende het correctieproces in de Welzijn en Kinderopvang statistiek. Zowel (a) de verandering in breedte van het betrouwbaarheidsinterval en (b) de significante verandering in gemiddelde waarde zijn weergegeven. De verandering is relatief ten opzichte van de ruwe data. Variabelen liggen op volgorde van verandering in breedte van het betrouwbaarheidsinterval.

betrouwbaarheidsinterval ter hoogte van de ruwe data, en we zeggen dat het percentage substantiële verandering of verschil 24 procent is.

Voor de Welzijn en Kinderopvang statistiek, figuur 4, zien we een substantiële verandering in de waarde bij de variabelen ``Welzijnswerk oud.'' en ``Omzet dienstverl. bedr.''. Voor ongeveer 20% van de variabelen verbetert de betrouwbaarheid van de gemiddelde waarde aanzienlijk, deze verbetering is in de meeste gevallen toe te schrijven aan correctie op duizendfouten. Voor ongeveer 8% verslechtert de betrouwbaarheid. Het is opmerkelijk dat deze verslechtering in de meeste gevallen toe te schrijven is aan de laatste stap in het correctieproces.

Voor de Groothandel statistiek, figuur 5, zien we dat verschillende variabelen een substantiële verandering in waarde vertonen, vrijwel allemaal ten gevolge van de stap ``deductieve imputatie''. Deze stap en de stap ``duizendfouten'' zijn ook verantwoordelijk voor een aanzienlijke toename in de betrouwbaarheid voor het meerendeel van de variabelen. De betrouwbaarheid daalt voor de variabelen ``Omzet rep. en onderh.'', ``Pers. uitgeleend'' en ``Vrijval voorz.'', dit gebeurt in de laatste stappen ``regressie imputatie'' en ``aanpassen''.

(a) Betrouwbaarheid **(b) Significante verandering**

Figuur 5 Variabelwaarden gedurende het correctieproces in de Groothandel statistiek. Zowel (a) de verandering in breedte van het betrouwbaarheidsinterval en (b) de substantiële verandering in gemiddelde waarde zijn weergegeven. De verandering is relatief ten opzichte van de ruwe data. Variabelen liggen op volgorde van verandering in breedte van het betrouwbaarheidsinterval.

4 Regelschending

In het algemeen heeft een verzameling data aan allerlei regels te voldoen voordat de data als acceptabel is voor statistische analyse. Het is dan informatief om te weten in welke mate de data aan opgelegde regels voldoet. We beschouwen daartoe een harde en een zachte maat voor regelschending. De harde maat controleert of aan regels strict is voldaan, en dit noemen we regelverificatie. De zachte maat bekijkt of aan regels is voldaan, en zo niet in welke mate deze regels zijn geschonden. Dit introduceert een tolerantie waaronder aan regels is voldaan.

4.1 Regelverificatie

In regelverificatie wordt nagegaan of records aan bepaalde regels voldoen. We gaan er in het vervolg van uit dat in records waarden kunnen ontbreken. Een regel kan geïnterpreteerd worden als een drie-waardige functie, met de volgende definitie:

$$e(\boldsymbol{x}) = \begin{cases} 0 & \text{als } \boldsymbol{x} \text{ voldoet aan } e \\ 1 & \text{als } \boldsymbol{x} \text{ niet voldoet aan } e \\ \texttt{NA} & \text{als niet bepaald kan worden of } \boldsymbol{x} \text{ aan } e \text{ voldoet.} \end{cases}$$

De waarden 0 en 1 kunnen geïnterpreteerd worden als ``waar'' en ``niet waar'', respectievelijk. De waarde NA wordt teruggegeven indien in het record x minstens één waarde ontbreekt die nodig is voor de bepaling ofdat record $\boldsymbol{x}$ aan regel e voldoet.

Wanneer regelverificatie gedaan wordt voor een verzameling van K regels en data bestaande uit N records, dan levert dit $N \times K$ waarden op gelijk aan 0, 1 of NA. Deze waarden kunnen overzichtelijk in een matrix geplaatst worden, wat een $N \times K$ matrix $\boldsymbol{F}$ definieert met cellen

$$\boldsymbol{F}_{nk} = e_k(\boldsymbol{x}_n), \text{ met } \boldsymbol{x}_n \text{ het } n\text{de record en } e_k \text{ de } k\text{de regel.} \tag{2}$$

4.2 Tolerantie voor regels

Het gegeven dat een record $\boldsymbol{x}$ een regel e schendt, indien dit het geval is, zegt ons iets over één aspect van de kwaliteit van dat record: het voldoet niet aan de in e gestelde voorwaarde. Het zegt echter niets over de mate waarin deze voorwaarde geschonden wordt. De *mate* van schending van e is echter ook een aspect van de kwaliteit van het record $\boldsymbol{x}$. Een mogelijke maat hiervoor is de ``afstand'' tussen $\boldsymbol{x}$ en dat deel van het waardedomein $D$ waarbinnen aan regel $e$ voldaan is, waarbij de ``afstand'' nog nader te definiëren is. Het kan geïnterpreteerd worden als de tolerantie waaronder $\boldsymbol{x}$ aan regel e zou voldoen. De mate van schending wordt hier per regel gedefinieerd, dat is eenvoudig voor de interpretatie van de schendingen. Een alternatief is om alle regels simultaan te beschouwen en de mate van schending te definieren als de kortste afstand tussen $\boldsymbol{x}$ en de records die aan *alle* regels voldoen, zie hoofdstuk 6 voor een discussie van de mogelijkheden hiervoor.

Definitie: Gegeven een afstandsmaat d, die de afstand bepaalt tussen records, definiëren we de tolerantie $t(\boldsymbol{x}, e)$ waaronder een record $\boldsymbol{x}$ aan een regel e voldoet als volgt. Indien $\boldsymbol{x}$ geen ontbrekende waarden bevat dan is de tolerantie de grootste ondergrens van de verzameling van afstanden $d(\boldsymbol{x}, \boldsymbol{y})$ tussen record $\boldsymbol{x}$ en die records $\boldsymbol{y}$ die aan e voldoen, dat wil zeggen,

$$t(\boldsymbol{x}, e) = \inf\{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} \in D \text{ voldoet aan } e\}, \tag{3}$$

waarbij inf de operatie is van het nemen van de grootste ondergrens. Indien $\boldsymbol{x}$ ontbrekende waarden bevat maar wel bepaald kan worden of $\boldsymbol{x}$ aan e voldoet, dan is de tolerantie gedefinieerd als in (3) maar met de afstand $d(\boldsymbol{x},\boldsymbol{y})$ vervangen door $d(i(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{y})$ waarbij $i(\boldsymbol{x},\boldsymbol{y})$ de donor-imputatie is van $\boldsymbol{x}$ met als donor $\boldsymbol{y}$:

$$i(\boldsymbol{x},\boldsymbol{y})_j = \begin{cases} x_j & \text{als } x_j \neq \text{NA} \\ y_j & \text{als } x_j = \text{NA} \end{cases} \tag{4}$$

voor $j = 1, \ldots, J$ en

$$t(\boldsymbol{x}, e) = \inf\{d(i(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{y}) : \boldsymbol{y} \in D \text{ voldoet aan } e\}. \tag{5}$$

Indien niet bepaald kan worden of $\boldsymbol{x}$ aan e voldoet, dan is de tolerantie onbepaald:

$$t(\boldsymbol{x}, e) = \text{NA}. \tag{6}$$

Merk op dat formule (3) een speciaal geval is van formule (5) aangezien $i(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}$ indien $\boldsymbol{x}$ geen ontbrekende waarden heeft.

Voorbeelden van afstandsmaten zijn de Euclidische afstand

$$d(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{j=1}^{J}(x_j - y_j)^2\right)^{1/2},$$

die typisch gebruikt wordt voor continue data, en de Hamming afstand, die het aantal variabelen telt waarin twee records van elkaar verschillen,

$$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^{J}\delta(x_j, y_j) \qquad \delta(x_j, y_j) = \begin{cases} 1 & \text{als } x_j = y_j \\ 0 & \text{als } x_j \neq y_j \end{cases}$$

en die een meer voordehandliggende keuze is voor categoriale data.

Als een voorbeeld van tolerantiebepaling in het geval van continue data nemen we het volgende.

Voorbeeld: Gegeven is een record $\boldsymbol{x}$ met de variabelen ``kosten'', ``omzet'' en ``winst'',

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}[\text{kosten}] \\ \boldsymbol{x}[\text{omzet}] \\ \boldsymbol{x}[\text{winst}] \end{pmatrix} = \begin{pmatrix} -3 \\ 8 \\ 5 \end{pmatrix},$$

een regel

$$e : \text{kosten} > 0,$$

en een afstandsmaat

$$d(\boldsymbol{x},\boldsymbol{y}) = \Big(\big(\boldsymbol{x}[\text{kosten}] - \boldsymbol{y}[\text{kosten}]\big)^2 + \big(\boldsymbol{x}[\text{omzet}] - \boldsymbol{y}[\text{omzet}]\big)^2 + \big(\boldsymbol{x}[\text{winst}] - \boldsymbol{y}[\text{winst}]\big)^2\Big)^{1/2}.$$

Record $\boldsymbol{x}$ voldoet niet aan regel e want $\boldsymbol{x}[\text{kosten}] = -3$ is niet groter dan nul, en de tolerantie waaronder $\boldsymbol{x}$ aan e voldoet vinden we met formule (3):

$$t(\boldsymbol{x}, e) = \inf\{|-3 - \boldsymbol{y}[\text{kosten}]| : \boldsymbol{y}[\text{kosten}] > 0\} = |-3 - 0| = 3.$$

Stel dat de waarde voor de variabele ``omzet'' in $\boldsymbol{x}$ ontbreekt: $\boldsymbol{x}[\text{omzet}] = \text{NA}$. Dit verandert niets aan het feit dat $\boldsymbol{x}$ niet aan regel e voldoet, en de tolerantie is nu gegeven door (5): voor iedere $\boldsymbol{y}$ die aan regel e voldoet is de donor-imputatie van $\boldsymbol{x}$ met donor $\boldsymbol{y}$ gegeven door

$$i(\boldsymbol{x},\boldsymbol{y}) = \begin{pmatrix} -3 \\ \boldsymbol{y}[\text{omzet}] \\ 5 \end{pmatrix},$$

de afstand tussen $i(\boldsymbol{x}, \boldsymbol{y})$ en $\boldsymbol{y}$ is

$$\begin{aligned} d(i(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{y}) &= \Big(\big(-3 - \boldsymbol{y}[\text{kosten}] \big)^2 \\ &\qquad + \big(\boldsymbol{y}[\text{omzet}] - \boldsymbol{y}[\text{omzet}] \big)^2 + \big(5 - \boldsymbol{y}[\text{winst}] \big)^2 \Big)^{1/2} \\ &= \Big(\big(-3 - \boldsymbol{y}[\text{kosten}] \big)^2 + \big(5 - \boldsymbol{y}[\text{winst}] \big)^2 \Big)^{1/2}, \end{aligned}$$

en voor de tolerantie vinden we

$$\begin{aligned} t(\boldsymbol{x}, e) &= \inf\Big\{ \Big(\big(-3 - \boldsymbol{y}[\text{kosten}] \big)^2 + \big(5 - \boldsymbol{y}[\text{winst}] \big)^2 \Big)^{1/2} : \boldsymbol{y}[\text{kosten}] > 0 \Big\} \\ &= \inf\{ | -3 - \boldsymbol{y}[\text{kosten}] | : \boldsymbol{y}[\text{kosten}] > 0 \} = | -3 - 0 | = 3. \end{aligned}$$

In het geval dat de waarde voor de variabele ``kosten'' in $\boldsymbol{x}$ zou ontbreken kan niet bepaald worden of $\boldsymbol{x}$ aan regel e voldoet, en de tolerantie is dan niet bepaald: $t(\boldsymbol{x}, e) = \text{NA}$.

Het dient opgemerkt te worden dat de gegeven tolerantie slechts voor individuele regels gedefinieert is. Voldoet een record onder kleine tolerantie aan een specifieke regel, dan in het algemeen sluit dit niet uit dat er andere regels zijn waar slechts onder grote tolerantie aan voldaan is.

4.3 Tolerantie voor lineaire regels

In de praktijk hebben regels voor continue data vaak de vorm van een lineaire (on)gelijkheid. Voorbeelden zijn

$$\begin{aligned} e_1 &: && \text{kosten} > 0 \\ e_2 &: && \text{winst} = \text{omzet} - \text{kosten}. \end{aligned}$$

Zulke lineaire regels kunnen wiskundig weergegeven worden als

$$e(\boldsymbol{x}) = \begin{cases} 0 & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \in B \\ 1 & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \notin B \\ \text{NA} & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \text{ niet bepaald kan worden,} \end{cases} \tag{7}$$

waarbij $\boldsymbol{a} \neq 0$ en record $\boldsymbol{x}$ vectoren zijn in de Euclidische ruimte $\mathbb{R}^J$, J is het aantal variabelen, en B is een interval of een verzameling van één element in $\mathbb{R}$. De regels e_1 en e_2 in bovenstaand voorbeeld zijn van de vorm (7) met

$$\boldsymbol{x} = \begin{pmatrix} \text{kosten} \\ \text{omzet} \\ \text{winst} \end{pmatrix} \qquad \boldsymbol{a} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad B = (0, \infty)$$

voor regel e_1 en

$$\boldsymbol{x} = \begin{pmatrix} \text{kosten} \\ \text{omzet} \\ \text{winst} \end{pmatrix} \qquad \boldsymbol{a} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \qquad B = \{0\}$$

voor regel e_2.

Voor lineaire regels kunnen we toleranties definiëren volgens formules (3), (5) en (6). We zullen nu laten zien dat de tolerantie een gereduceerde vorm heeft in het bijzondere geval waarin de afstandsmaat van de vorm

$$d(\boldsymbol{x}, \boldsymbol{y}) = \| \boldsymbol{x} - \boldsymbol{y} \|_p$$

is, waarbij voor iedere $\boldsymbol{v} \in \mathbb{R}^J$

$$\| \boldsymbol{v} \|_p = \left(\sum_{j=1}^{J} |v_j|^p \right)^{1/p}$$

voor $1 \leq p < \infty$ en

$$\|\boldsymbol{v}\|_p = \max\{|v_j| : j = 1, \ldots, J\}$$

voor $p = \infty$. Daarbij nemen we de conventie aan dat

$$0 \cdot \texttt{NA} = \texttt{NA} \cdot 0 = 0. \tag{8}$$

Stelling: *Indien de afstandsmaat gegeven is door*

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p$$

waarbij $1 \leq p \leq \infty$*, dan heeft de tolerantie waaronder een record* $\boldsymbol{x}$ *aan een lineaire regel*

$$e(\boldsymbol{x}) = \begin{cases} 0 & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \in B \\ 1 & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \notin B \\ \texttt{NA} & \text{als } \boldsymbol{a}^\top \boldsymbol{x} \text{ niet bepaald kan worden} \end{cases}$$

voldoet de gereduceerde vorm

$$t(\boldsymbol{x}, e) = \inf\{\|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| : b \in B\}, \tag{9}$$

waarbij $\frac{1}{p} + \frac{1}{q} = 1$.

In het bewijs zullen we gebruik maken van het puntsgewijs product van vectoren, wat voor twee vectoren $\boldsymbol{x}$ en $\boldsymbol{y}$ in $\mathbb{R}^J$ gedefinieerd is als de vector

$$\boldsymbol{xy} = \begin{pmatrix} x_1 y_1 \\ \vdots \\ x_J y_J \end{pmatrix}.$$

Bewijs: De reductie is triviaal in het geval waarin niet bepaald kan worden of een record $\boldsymbol{x}$ aan de lineaire regel e voldoet: het inproduct $\boldsymbol{a}^\top \boldsymbol{x}$ is onbepaald, en daarmee de tolerantie volgens formule (9), wat in overeenstemming is met de algemene definitie volgens formule (6).

Stel nu dat wel bepaald kan worden of $\boldsymbol{x}$ aan regel e voldoet. Dit betekent dat het inproduct $\boldsymbol{a}^\top \boldsymbol{x}$ een bepaalde waarde heeft, dat $a_j = 0$ als $x_j = \texttt{NA}$ voor $j = 1, \ldots, J$ wegens conventie (8), en dat

$$\boldsymbol{a}^\top i(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{a}^\top \boldsymbol{x} \text{ voor iedere } \boldsymbol{y} \in \mathbb{R}^J. \tag{10}$$

Verder geldt voor iedere $\boldsymbol{y} \in \mathbb{R}^J$ dat

$$\begin{aligned} d(i(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}) &= \|i(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{y}\|_p \\ &\geq \|\boldsymbol{a}\|_q^{-1} \|\boldsymbol{a}(i(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{y})\|_1 \\ &= \|\boldsymbol{a}\|_q^{-1} \sum_{j=1}^{J} |a_j (i(\boldsymbol{x}, \boldsymbol{y})_j - y_j)| \\ &\geq \|\boldsymbol{a}\|_q^{-1} \left| \sum_{j=1}^{J} a_j (i(\boldsymbol{x}, \boldsymbol{y})_j - y_j) \right| \\ &= \|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top (i(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{y})| \\ &= \|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{y})| \end{aligned} \tag{11}$$

waar de eerste ongelijkheid een toepassing is van de ongelijkheid van Hölder (Steele, 2004).

Enerzijds volgt uit (11) dat

$$\begin{aligned} d(i(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}) &\geq \inf\{\|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{y})| : \boldsymbol{y} \in \mathbb{R}^J \text{ voldoet aan } e\} \\ &= \inf\{\|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| : b \in B\} \end{aligned}$$

voor iedere $\boldsymbol{y} \in \mathbb{R}^J$ die voldoet aan e, en dus

$$t(\boldsymbol{x}, e) \geq \inf\{\|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| : b \in B\}$$

wegens de definitie van $t(\boldsymbol{x}, e)$.

Anderzijds, als voor iedere $b \in B$ er een $\boldsymbol{y} \in \mathbb{R}^J$ is waarvoor $\boldsymbol{a}^\top \boldsymbol{y} = b$ en gelijkheid geldt in (11), dan wegens de definitie van $t(\boldsymbol{x}, e)$ is

$$\begin{aligned} t(\boldsymbol{x}, e) \leq d(i(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}) &= \|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{y})| \\ &= \|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| \end{aligned}$$

en dus

$$t(\boldsymbol{x}, e) \leq \inf\{\|\boldsymbol{a}\|_q^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| : b \in B\}.$$

Formule (9) is daarmee afgeleid als we kunnen aantonen dat zo'n $\boldsymbol{y}$ bestaat.

Voor de rest van het bewijs laat $b \in B$ en $\boldsymbol{y} \in \mathbb{R}^J$ waarvoor gelijkheid geldt in (11) en $\boldsymbol{a}^\top \boldsymbol{y} = b$. We merken eerst op dat de tweede ongelijkheid in (11) een gelijkheid is dan en slechts dan als de termen $a_j(i(\boldsymbol{x}, \boldsymbol{y})_j - y_j)$ elkaar niet annuleren, oftewel als er een $s \in \{-1, +1\}$ is zodanig dat

$$\text{sgn}(a_j(i(\boldsymbol{x}, \boldsymbol{y})_j - y_j)) \in \{0, s\} \tag{12}$$

voor iedere j.

In het geval $p = 1$ en $q = \infty$ geldt gelijkheid in (11) dan en slechts dan als

$$i(\boldsymbol{x}, \boldsymbol{y})_j - y_j = 0 \text{ of } |a_j| = \|\boldsymbol{a}\|_\infty$$

en (12) voor iedere j. Hieruit volgt dat

$$b - \boldsymbol{a}^\top \boldsymbol{x} = \sum_{\substack{j=1,\ldots,J \\ |a_j| = \|\boldsymbol{a}\|_\infty}} a_j(y_j - x_j).$$

Een $\boldsymbol{y} \in \mathbb{R}^J$ die hier aan voldoet is gegeven door

$$y_j = \begin{cases} 0 & \text{als } x_j = \text{NA} \\ x_j & \text{als } x_j \neq \text{NA en } |a_j| \neq \|\boldsymbol{a}\|_\infty \\ x_j + (b - \boldsymbol{a}^\top \boldsymbol{x}) N^{-1} a_j^{-1} & \text{als } x_j \neq \text{NA en } |a_j| = \|\boldsymbol{a}\|_\infty \end{cases}$$

waarbij N het aantal componenten in $\boldsymbol{a}$ waarvoor $|a_j| = \|\boldsymbol{a}\|_\infty$.

In het geval $p = \infty$ en $q = 1$ geldt gelijkheid in (11) dan en slechts dan als

$$|i(\boldsymbol{x}, \boldsymbol{y})_j - y_j| = \|i(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{y}\|_\infty \text{ of } a_j = 0$$

en (12) voor iedere j. Hieruit volgt voor iedere j waarvoor $a_j \neq 0$ en $y_j - i(\boldsymbol{x}, \boldsymbol{y})_j \neq 0$ dat

$$\begin{aligned} y_j - i(\boldsymbol{x}, \boldsymbol{y})_j &= \text{sgn}(y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) |y_j - i(\boldsymbol{x}, \boldsymbol{y})_j| \\ &= \text{sgn}(a_j) s \|\boldsymbol{y} - i(\boldsymbol{x}, \boldsymbol{y})\|_\infty \end{aligned}$$

en

$$b - \boldsymbol{a}^\top \boldsymbol{x} = \sum_{j=1}^{J} a_j(y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) = \|\boldsymbol{a}\|_1 s \|\boldsymbol{y} - i(\boldsymbol{x}, \boldsymbol{y})\|_\infty$$

wat s definieert. Een $\boldsymbol{y} \in \mathbb{R}^J$ die hier aan voldoet is gegeven door

$$y_j = \begin{cases} 0 & \text{als } x_j = \text{NA} \\ x_j + (b - \boldsymbol{a}^\top \boldsymbol{x}) \|\boldsymbol{a}\|_1^{-1} \text{sgn}(a_j) & \text{als } x_j \neq \text{NA}. \end{cases}$$

In het geval $1 < p < \infty$ en $1 < q < \infty$ geldt gelijkheid in (11) voor $\boldsymbol{y} \in \mathbb{R}^J$ dan en slechts dan als er een $c \geq 0$ is zodanig dat

$$|i(\boldsymbol{x}, \boldsymbol{y})_j - y_j|^p = c|a_j|^q$$

en (12) voor iedere j. Als $c = 0$ dan is $\boldsymbol{y} - i(\boldsymbol{x}, \boldsymbol{y}) = 0$, anders geldt voor iedere j waarvoor $a_j \neq 0$ dat $y_j - i(\boldsymbol{x}, \boldsymbol{y})_j \neq 0$ en $\text{sgn}(a_j)\text{sgn}(y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) = s \in \{-1, +1\}$ wegens (12), en dus

$$\begin{aligned} y_j - i(\boldsymbol{x}, \boldsymbol{y})_j &= \text{sgn}(y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) |y_j - i(\boldsymbol{x}, \boldsymbol{y})_j| \\ &= \text{sgn}(y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) c^{1/p} |a_j|^{q/p} \\ &= \text{sgn}(a_j) s c^{1/p} |a_j|^{q/p} \end{aligned}$$

en

$$b - \boldsymbol{a}^\top \boldsymbol{x} = \sum_{j=1}^{J} a_j (y_j - i(\boldsymbol{x}, \boldsymbol{y})_j) = sc^{1/p} \sum_{j=1}^{J} |a_j| |a_j|^{q/p}$$
$$= sc^{1/p} \sum_{j=1}^{J} |a_j|^q$$
$$= sc^{1/p} \|\boldsymbol{a}\|_q^q$$

wat $sc^{1/p}$ definieert. Een $\boldsymbol{y} \in \mathbb{R}^J$ die hier aan voldoet is gegeven door

$$y_j = \begin{cases} 0 & \text{als } x_j = \text{NA} \\ x_j + (b - \boldsymbol{a}^\top \boldsymbol{x}) \|\boldsymbol{a}\|_q^{-q} \text{sgn}(a_j) |a_j|^{q/p} & \text{als } x_j \neq \text{NA}. \end{cases}$$

QED

De afstandsmaat

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p = \left(\sum_{j=1}^{J} |x_j - y_j|^p \right)^{1/p}$$

weegt iedere component *j* even zwaar. Dit is niet altijd wenselijk, en de ongewogen som in bovenstaande afstand kan dan vervangen worden door een gewogen som met gewichten $w_j > 0, j = 1, \ldots, J$. Dit levert een afstand

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{w},p}$$

op, waarbij

$$\|\boldsymbol{v}\|_{\boldsymbol{w},p} = \left(\sum_{j=1}^{J} |v_j|^p w_j \right)^{1/p}$$

voor iedere vector $\boldsymbol{v} \in \mathbb{R}^J$ en vector $\boldsymbol{w} = (w_1, \ldots, w_J)$ van gewichten w_j en $1 \leq p < \infty$. We definiëren ook

$$\|\boldsymbol{v}\|_{\boldsymbol{w},p} = \max\{|v_j| : w_j > 0 \text{ en } j = 1, \ldots, J\}$$

voor $p = \infty$. Voor zulke afstanden is er een vergelijkbare reductie van de formules (3), (5) en (6) voor de tolerantie.

Corollarium: *Indien de afstandsmaat gegeven is door*

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{w},p}$$

waarbij

$$\boldsymbol{w} = (w_1, \ldots, w_J) \qquad\qquad w_1, \ldots, w_J > 0$$

en $1 \leq p \leq \infty$*, dan heeft de tolerantie waaronder een record* $\boldsymbol{x}$ *aan een lineaire regel*

$$e(\boldsymbol{x}) = \begin{cases} 0 & \textit{als } \boldsymbol{a}^\top \boldsymbol{x} \in B \\ 1 & \textit{als } \boldsymbol{a}^\top \boldsymbol{x} \notin B \\ \textit{NA} & \textit{als } \boldsymbol{a}^\top \boldsymbol{x} \textit{ niet bepaald kan worden} \end{cases}$$

voldoet de gereduceerde vorm

$$t(\boldsymbol{x}, e) = \inf\{\|\boldsymbol{w}^{-1} \boldsymbol{a}\|_{\boldsymbol{w},q}^{-1} |\boldsymbol{a}^\top \boldsymbol{x} - b| : b \in B\}, \tag{13}$$

waarbij $\frac{1}{p} + \frac{1}{q} = 1$ *en*

$$\boldsymbol{w}^{-1} = \begin{pmatrix} w_1^{-1} \\ \vdots \\ w_J^{-1} \end{pmatrix}.$$

Bewijs: Het bewijs is equivalent aan het eerder gegeven bewijs voor de ongewogen som, maar met ongelijkheid (11) aangepast tot

$$\begin{aligned} d(i(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{y}) &= \left\|i(\boldsymbol{x},\boldsymbol{y})-\boldsymbol{y}\right\|_{\boldsymbol{w},p} \\ &\geq \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}\left\|\boldsymbol{w}^{-1}\boldsymbol{a}(i(\boldsymbol{x},\boldsymbol{y})-\boldsymbol{y})\right\|_{\boldsymbol{w},1} \\ &= \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}\sum_{j=1}^{J}|a_j(i(\boldsymbol{x},\boldsymbol{y})_j-y_j)| \\ &\geq \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}\left|\sum_{j=1}^{J}a_j(i(\boldsymbol{x},\boldsymbol{y})_j-y_j)\right| \\ &= \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}|\boldsymbol{a}^{\top}(i(\boldsymbol{x},\boldsymbol{y})-\boldsymbol{y})| \\ &= \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}|\boldsymbol{a}^{\top}(\boldsymbol{x}-\boldsymbol{y})| \end{aligned}$$

waar de eerste ongelijkheid een toepassing is van de ongelijkheid

$$\left\|\boldsymbol{f}\boldsymbol{g}\right\|_{\boldsymbol{w},1} \leq \left\|\boldsymbol{f}\right\|_{\boldsymbol{w},p}\left\|\boldsymbol{g}\right\|_{\boldsymbol{w},q}$$

van Hölder (Steele, 2004), met $\boldsymbol{f} = i(\boldsymbol{x},\boldsymbol{y})-\boldsymbol{y}$ en $\boldsymbol{g} = \boldsymbol{w}^{-1}\boldsymbol{a}$, en vervolgens iedere instantie van $\|\boldsymbol{a}\|_q$, $|a_j|^q = \left\|\boldsymbol{a}\right\|_{\infty}$ en $\|\boldsymbol{y}-i(\boldsymbol{x},\boldsymbol{y})\|_{\infty}$ vervangen door respectievelijk $\left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}$, $|w_j^{-1}a_j|^q = \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},\infty}$ en $\left\|\boldsymbol{y}-i(\boldsymbol{x},\boldsymbol{y})\right\|_{\boldsymbol{w},\infty}$. QED

Formules (9) en (13) kunnen nog verder gereduceerd worden indien de lineaire regels de vorm hebben van (on)gelijkheden. Dit betekent dat we ze kunnen schrijven als

$$e(\boldsymbol{x}) = \begin{cases} 0 & \text{als } \boldsymbol{a}^{\top}\boldsymbol{x} \odot b \\ 1 & \text{als niet } \boldsymbol{a}^{\top}\boldsymbol{x} \odot b \\ \text{NA} & \text{als } \boldsymbol{a}^{\top}\boldsymbol{x} \text{ niet bepaald kan worden,} \end{cases}$$

waarbij $b \in \mathbb{R}$ en $\odot$ is één van de relaties $<$, $\leq$, $=$, $\geq$ of $>$. We onderscheiden de volgende gevallen.

1. De relatie $\odot$ is een ongelijkheid $<$ of $\leq$, en de tolerantie is gegeven door
   $$t(\boldsymbol{x},e) = \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}(\boldsymbol{a}^{\top}\boldsymbol{x}-b) \vee 0.$$
2. De relatie $\odot$ is een gelijkheid $=$, en de tolerantie is gegeven door
   $$t(\boldsymbol{x},e) = \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}|\boldsymbol{a}^{\top}\boldsymbol{x}-b|.$$
3. De relatie $\odot$ is een ongelijkheid $>$ of $\geq$, en de tolerantie is gegeven door
   $$t(\boldsymbol{x},e) = \left\|\boldsymbol{w}^{-1}\boldsymbol{a}\right\|_{\boldsymbol{w},q}^{-1}(b-\boldsymbol{a}^{\top}\boldsymbol{x}) \vee 0.$$

Bepaling van de toleranties voor een verzameling van data bestaande uit N records die hebben te voldoen aan K regels levert een totaal aantal van $N \times K$ tolerantiewaarden op. Deze waarden kunnen overzichtelijk in een matrix geplaatst worden, wat een $N \times K$ matrix $\boldsymbol{G}$ definieert met cellen

$$\boldsymbol{G}_{nk} = t(\boldsymbol{x}_n, e_k), \text{ met } \boldsymbol{x}_n \text{ het } n\text{de record en } e_k \text{ de } k\text{de regel.} \tag{14}$$

4.4 Regelschendingen op het niveau van de gehele data

We hebben gekeken naar harde en zachte maten voor schendingen van regels door respectievelijk regels te verifiëren en een maat van tolerantie te introduceren. Met de harde maat van schendingen tellen we het aantal gevallen waarin een record aan een regel voldoet, het aantal gevallen waarin een record een regel schendt, en het aantal gevallen waarin

totaal					
totaal verifieerbaar				totaal niet verifieerbaar	
totaal geschonden		totaal ongeschonden			
nog steeds geschonden	geschonden, extra	nog steeds onge-schonden	onge-schonden, extra	nog steeds niet verifieer-baar	niet verifieer-baar, extra

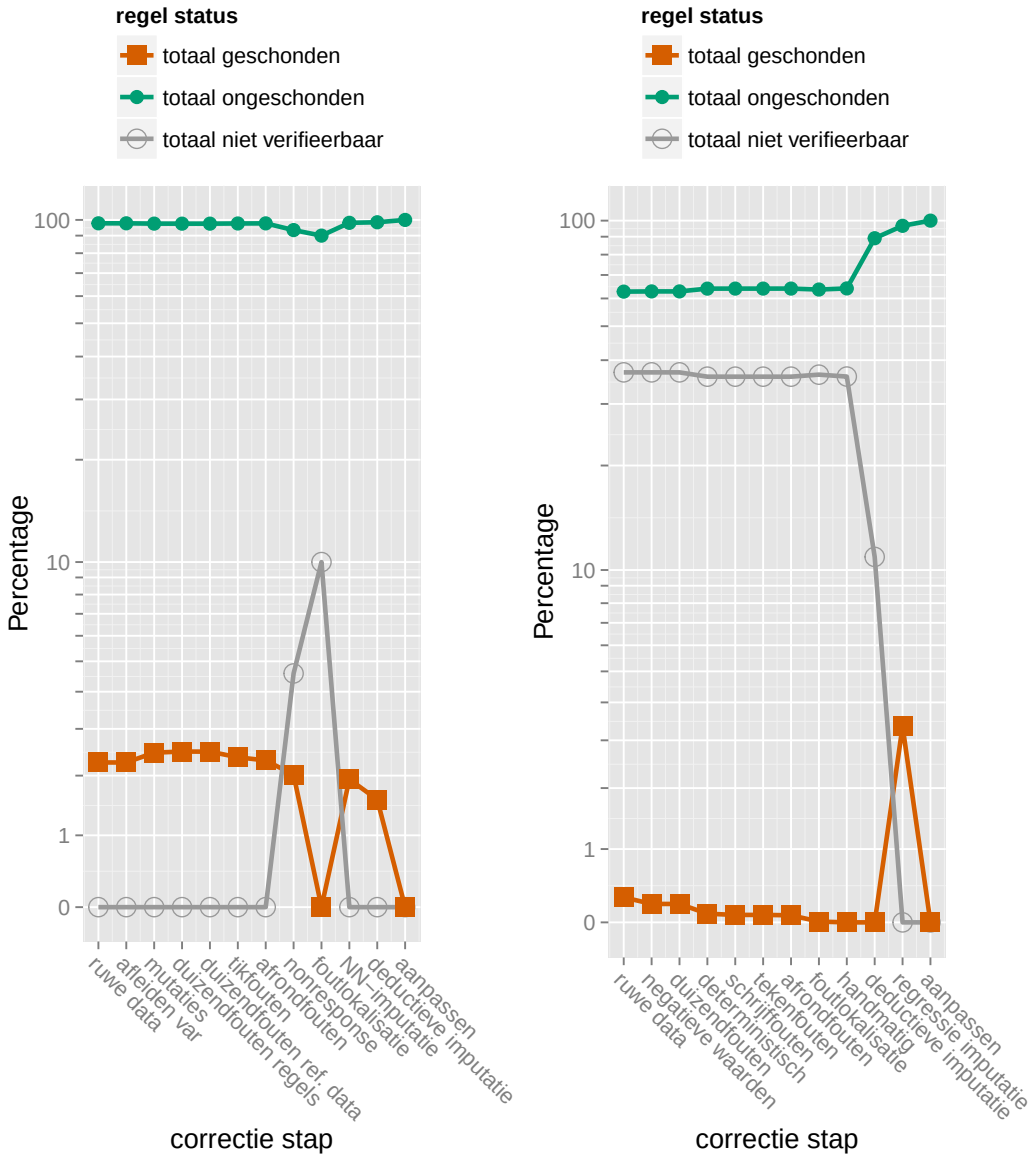
Tabel 7 Onderverdeling van de status van regelschendingen.

regelschending niet verifieerbaar is. Doen we dit voor iedere stap in een correctieproces, dan kunnen we aan iedere combinatie van een record en een regel een status toekennen, net zoals we in sectie 3.1 gedaan hebben voor cellen in data. Een onderverdeling hiervan is weergegeven in tabel 7. Een eerste onderverdeling wordt gemaakt op basis van verifieerbaarheid, verifieerbare gevallen worden verder onderverdeeld in gevallen waarbij een regel geschonden danwel niet geschonden wordt. De resulterende statussen ``totaal geschonden'', ``totaal ongeschonden'' en ``totaal niet verifieerbaar'' worden verder onderverdeeld op basis van enige verandering ten opzichte van een eerder referentiepunt, zoals de voorgaande of de eerste stap in het correctieproces.

Tabellen 8 en 9 geven de statusverdelingen van regelschendingen in absolute aantallen voor de statistieken Welzijn en Kinderopvang respectievelijk Groothandel. Figuur 6 geeft de status voor beide statistieken in percentages. De percentages zijn weergegeven op een pseudologaritmische schaal, zie formule (1) in sectie 3.1. Voor de Welzijn en Kinderopvang statistiek zien we dat het aantal regelschendingen in het begin enigzins toeneemt, om uiteindelijk tot nul te reduceren. De stappen ``nonresponse'' en ``foutlokalisatie'' laten het aantal regelschendingen dalen maar maken tevens een groot aantal regels niet verifieerbaar, dit is te begrijpen doordat deze stappen ontbrekende waarden introduceren zoals we zagen in tabel 5 en figuur 1(a). Voor de Groothandel statistiek zien we dat het aantal regelschendingen langzaam tot nul reduceert, met uitzondering van een piek rond ``regressie imputatie''. Het aantal niet verifieerbare regels is hoog en daalt pas aanzienlijk ten gevolge van ``deductieve imputatie'' en ``regressie imputatie'', dit is te begrijpen doordat deze stappen ontbrekende waarden invullen, zoals we zagen in tabel 6 en figuur 1(b).

Een nader inzicht krijgen we door naar de tolerantiewaarden te kijken. De tolerantiewaarden zijn een maat voor de omvang van de regelschendingen, een hogere tolerantiewaarde betekent een grotere mate van regelschending, en hun verdeling verschaft ons derhalve informatie over de omvang van de regelschendingen. Er zijn verschillende manieren om verdelingen visueel weer te geven, hier kiezen we voor de boxplot. Een alternatief dat we ook overwogen hebben is weergave van de kansdichtheidsfunctie, dit bleek echter minder duidelijk te zijn.

Figuren 7 en 8 laten de verdeling van de positieve tolerantiewaarden zien voor iedere stap in het correctieproces van de data en door middel van boxplots. De waarden liggen op een pseudologaritmische schaal. Een boxplot bestaat onder meer uit een doos die over het interkwartielbereik ligt, dat wil zeggen, over het gebied tussen de 25% laagste tolerantiewaarden en de 25% hoogste tolerantiewaarden. De hoogte van de boxplot kiezen we proportioneel aan de vierkantswortel van het aantal positieve tolerantiewaarden. De mediaan wordt weergegeven door een lijn die de doos in twee niet-noodzakelijkerwijs gelijke delen deelt. Van de doos uit lopen twee lijnen, een naar de kleinste tolerantiewaarde die groter is dan of gelijk aan de 25% laagste tolerantiewaarden min anderhalf keer de lengte van het



(a) Welzijn en Kinderopvang **(b) Groothandel**

Figuur 6 Statusverdeling van regels over correctiestappen met betrekking tot de statistieken (a) Welzijn en Kinderopvang en (b) Groothandel. De hoeveelheden zijn weergegeven in percentages op een pseudologaritmische schaal. Percentages pas geschonden, pas ongeschonden en niet meer verifieerbare regels zijn relatief ten opzichte van de ruwe data.

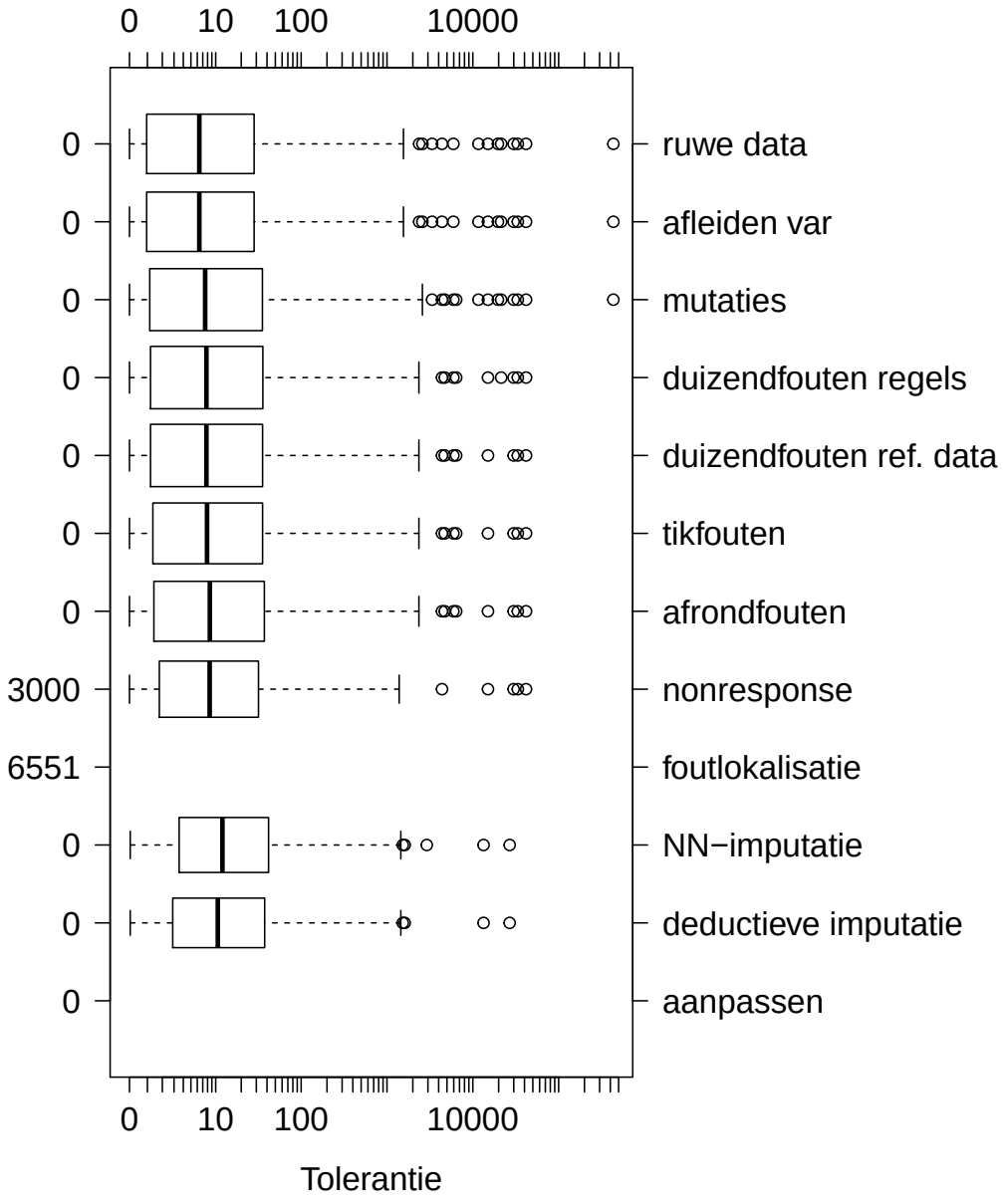
Correctie stap	totaal	totaal geschonden	geschonden, extra t.o.v. ruwe data	totaal ongeschonden	ongeschonden, extra t.o.v. ruwe data	totaal niet verifieerbaar	niet verifieerbaar, extra t.o.v. ruwe data
ruwe data	65520	1479	0	64041	0	0	0
afleiden var	65520	1479	0	64041	0	0	0
mutaties	65520	1612	138	63908	5	0	0
duizendfouten regels	65520	1628	158	63892	9	0	0
duizendfouten ref. data	65520	1626	158	63894	11	0	0
tikfouten	65520	1549	153	63971	83	0	0
afrondfouten	65520	1506	148	64014	121	0	0
nonresponse	65520	1321	81	61199	107	3000	3000
foutlokalisatie	65520	0	0	58969	104	6551	6551
NN-imputatie	65520	1265	682	64255	896	0	0
deductieve imputatie	65520	1029	536	64491	986	0	0
aanpassen	65520	0	0	65520	1479	0	0

Tabel 8 Veranderingen in de status van regels over correctiestappen voor de Welzijn en Kinderopvang statistiek. Veranderingen zijn ten opzichte van de ruwe data.

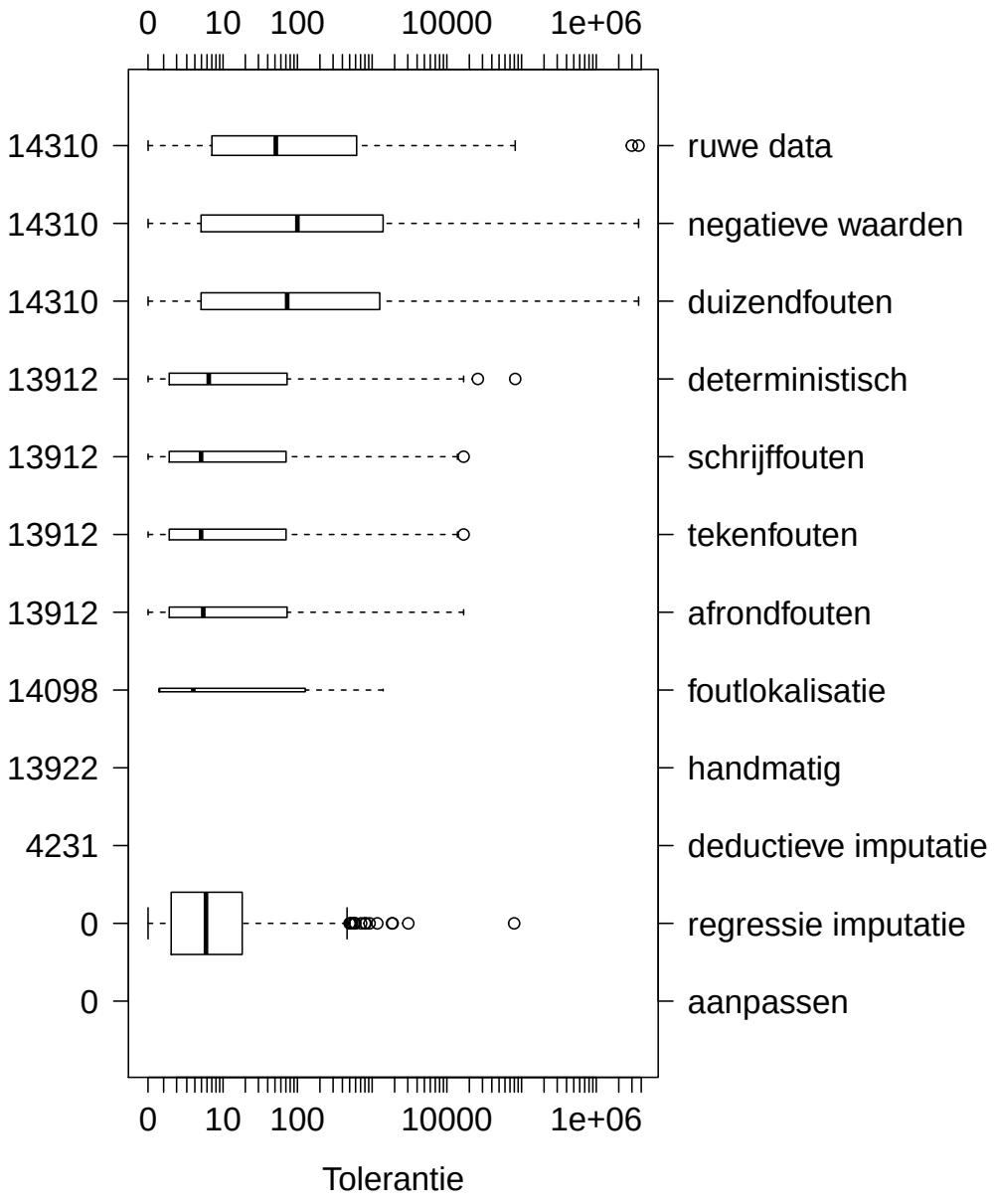
interkwartielbereik, en een naar de grootste tolerantiewaarde die kleiner is dan of gelijk aan de 25% hoogste tolerantiewaarden plus anderhalf keer de lengte van het interkwartielbereik. Deze twee lijnen markeren een bereik, en tolerantiewaarden die hier buiten vallen worden als uitbijters opgevat en weergegeven als individuele punten. Per correctiestap hebben we links van de boxplot het aantal onbepaalde tolerantiewaarden weergegeven.

In figuur 7 zien we dat de correctie op duizendfouten enkele uitbijters verwijdert. ``Nonresponse'' en ``foutlokalisatie'' maken cellen in de data leeg, daarmee reduceert het aantal positieve tolerantiewaarden tot nul ten koste van het aantal onbepaalde tolerantiewaarden wat stijgt tot 6551. De leeg gemaakte cellen worden weer ingevuld door ``NN-imputatie'', daarmee reduceert het aantal onbepaalde tolerantiewaarden weer tot nul ten koste van het aantal positieve tolerantiewaarden wat stijgt. ``Deductieve imputatie'' zorgt voor enige reductie in de positieve tolerantiewaarden, en een totale reductie wordt bewerkstelligd door de laatste stap ``aanpassen''.

In figuur 8 zien we dat in de stappen tot en met ``handmatig'' het aantal positieve tolerantiewaarden geleidelijk daalt tot nul. Correctie voor negatieve waarden doet het relatieve aandeel buiten het interkwartielbereik dalen, deterministische correctie zorgt voor zowel een daling in het aantal onbepaalde tolerantiewaarden als een halvering in het aantal positieve tolerantiewaarden. ``Deductieve imputatie'' zorgt voor een aanzienlijke daling in het aantal onbepaalde tolerantiewaarden. ``Regressie imputatie'' reduceert het aantal onbepaalde tolerantiewaarden tot nul, echter ten koste van een stijging in het aantal positieve tolerantiewaarden. De laatste stap ``aanpassen'' reduceert ook het aantal positieve tolerantiewaarden tot nul.



Figuur 7 Grafische weergave van de distributie van positieve tolerantiewaarden per individuele correctieregel over de stappen van het correctieproces voor de Welzijn en Kinderopvang statistiek. De waarden zijn weergegeven door middel van boxplots op een pseudologaritmische schaal, de hoogte van de boxplots is evenredig met de vierkantswortel van het aantal positieve tolerantiewaarden. De getallen links geven het aantal onbepaalde tolerantiewaarden per stap weer.



Figuur 8 Grafische weergave van de distributie van positieve tolerantiewaarden per individuele correctieregel over de stappen van het correctieproces voor de Groothandel statistiek. De waarden zijn weergegeven door middel van boxplots op een pseudologaritmische schaal, de hoogte van de boxplots is evenredig met de vierkantswortel van het aantal positieve tolerantiewaarden. De getallen links geven het aantal onbepaalde tolerantiewaarden per stap weer.

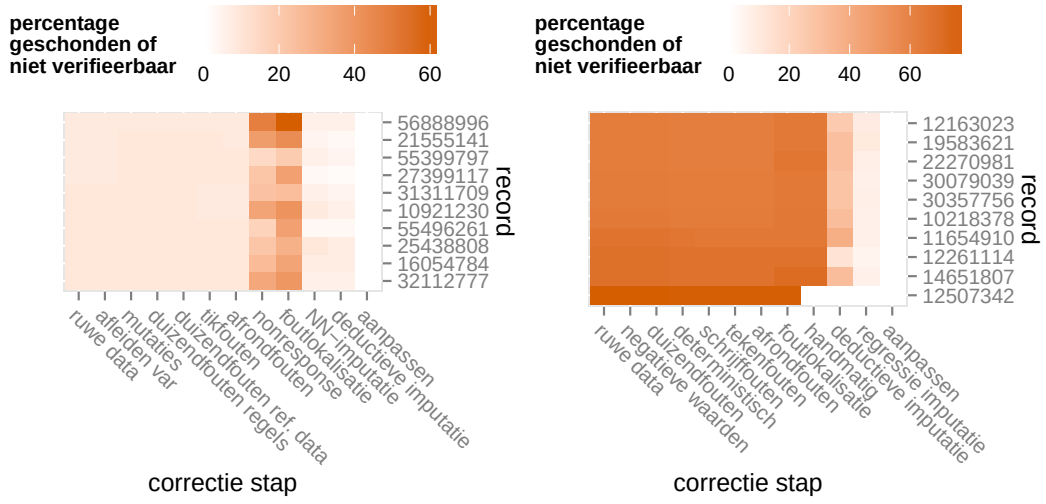
Correctie stap	totaal	totaal geschonden	geschonden, extra t.o.v. ruwe data	totaal ongeschonden	ongeschonden, extra t.o.v. ruwe data	totaal niet verifieerbaar	niet verifieerbaar, extra t.o.v. ruwe data
ruwe data	38760	128	0	24322	0	14310	0
negatieve waarden	38760	94	2	24356	36	14310	0
duizendfouten	38760	94	2	24356	36	14310	0
deterministisch	38760	44	3	24804	485	13912	0
schrijffouten	38760	38	3	24810	491	13912	0
tekenfouten	38760	38	3	24810	491	13912	2
afrondfouten	38760	36	3	24812	493	13912	2
foutlokalisatie	38760	4	1	24658	492	14098	188
handmatig	38760	0	0	24838	671	13922	188
deductieve imputatie	38760	0	0	34529	10338	4231	154
regressie imputatie	38760	1297	1288	37463	13146	0	0
aanpassen	38760	0	0	38760	14438	0	0

Tabel 9 Veranderingen in de status van regels over correctiestappen voor de Groothandel statistiek. Veranderingen zijn ten opzichte van de ruwe data.

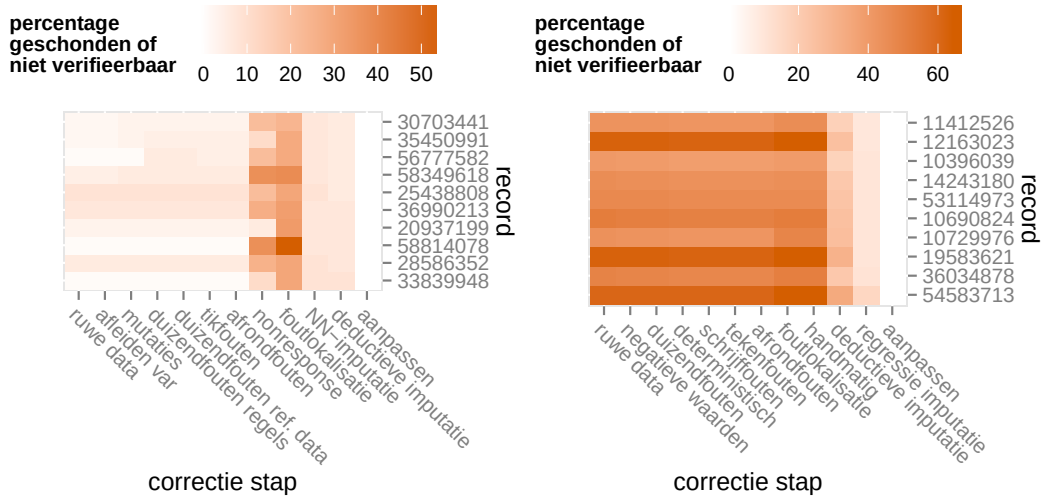
4.5 Regelschendingen op het niveau van records

In plaats van naar regelschendingen te kijken op het niveau van de gehele data doen we dit nu op het niveau van records. Daartoe tellen we per record het aantal geschonden en niet te verifiëren regels. Figuur 9 geeft het percentage geschonden en niet te verifiëren regels per record over de stappen in het correctieproces, voor zowel de Welzijn en Kinderopvang statistiek als de Groothandel statistiek. De records zijn gesorteerd op percentage hoogte: in figuren 9(a) en 9(b) ligt het record met het hoogste percentage ter hoogte van de eerste correctiestap (``ruwe data'') onderaan, gevolgd door het record met het op-een-na hoogste percentage, enzovoorts, en in figuren 9(c) en 9(d) ligt het record met het hoogste percentage ter hoogte van de laatste correctiestap (``aanpassen'') onderaan; in het geval van gelijke percentages zijn percentages ter hoogte van latere respectievelijk eerdere correctiestappen in beschouwing genomen om de records te sorteren. Enkel de tien records met de hoogste percentages zijn weergegeven.

In figuren 9(a) en 9(c) zien we dat voor records met het hoogste percentage geschonden of niet te verifiëren regels bij aanvang respectievelijk op het einde van het correctieproces in de Welzijn en Kinderopvang statistiek er weinig verandering is in het percentage gedurende de eerste correctiestappen. De correctiestappen ``nonresponse'' en ``foutlokalisatie'' zorgen voor een aanzienlijke stijging in het percentage. Dit komt doordat deze stappen een groot aantal waarden in de data als foutief aanwijzen en op ontbrekend zetten, zoals we gezien hebben in sectie 3.1. Het gevolg is een aanzienlijke toename in het aantal niet te verifiëren regels. De volgende stap, ``NN-imputatie'', vult alle ontbrekende waarden weer in en zorgt daarmee voor een aanzienlijke daling in de percentages geschonden of niet-verifieerbare regels. Ook in de Groothandel statistiek, figuren 9(b) en 9(d), zien we dat de percentages weinig veranderen gedurende de eerste correctiestappen. Vanaf correctiestap ``handmatig'' reduceren de percentages geleidelijk tot nul. In tegenstelling tot de Welzijn en Kinderopvang statistiek zien we hier geen grote stijgingen in de percentages. Dit heeft te maken met het feit dat het aandeel



(a) Welzijn en Kinderopvang **(b) Groothandel**



(c) Welzijn en Kinderopvang **(d) Groothandel**

Figuur 9 Het percentage geschonden of niet-verifieerbare regels per record over de stappen in het correctieproces met betrekking tot (a, c) de Welzijn en Kinderopvang statistiek en (b, d) de Groothandel statistiek. Records zijn gesorteerd op het hoogste percentage aan het begin (a, b) danwel einde (c, d) van het correctieproces; alleen de tien records met de hoogste percentages zijn weergegeven.

ontbrekende waarden bij aanvang veel hoger is, zoals we gezien hebben in sectie 3.1.

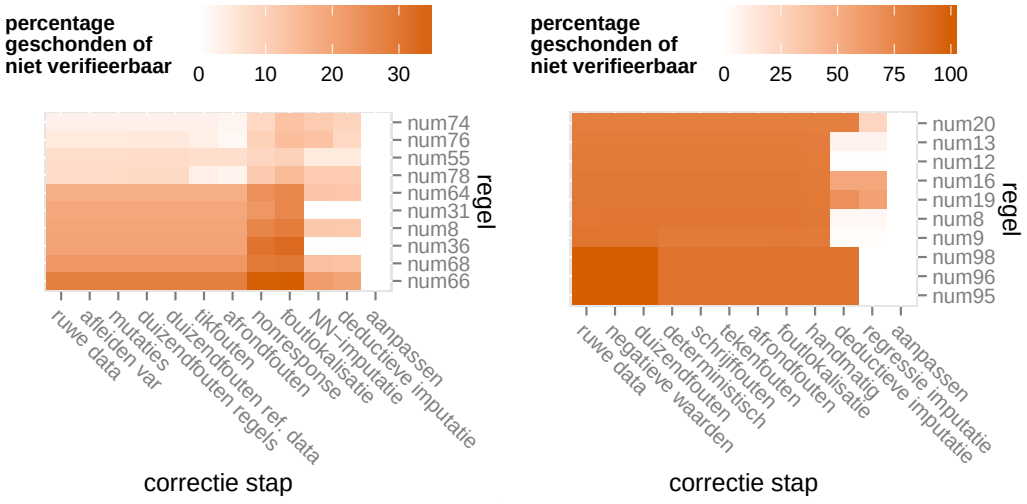
Behalve het aantal geschonden en niet-verifieerbare regels per record te tellen kunnen we ook per record naar de tolerantiewaarden kijken. Van deze tolerantiewaarden kunnen we dan bijvoorbeeld het gemiddelde bepalen, of de mediaan of het maximum, afhankelijk van de informatie die we willen verkrijgen. Net als bij het aantal regelschendingen kunnen we zo de records vinden met de hoogste gemiddelde tolerantie, etc. De informatie die dit verschaft over de bijdragen van de afzonderlijke stappen lijkt hierbij vergelijkbaar te zijn met wat we zagen bij het tellen van het aantal geschonden en niet te verifiëren regels in figuur 9. Een weergave door middel van een figuur laten we achterwege.

4.6 Regelschendingen op het niveau van regels

Net zoals we in sectie 4.5 naar regelschendingen op het niveau van records hebben gekeken, kijken we nu naar regelschendingen op het niveau van regels door per regel het aantal records te tellen dat deze regel schendt of waarvoor deze regel niet verifieerbaar is. Figuur 10 geeft het percentage geschonden of niet-verifieerbare regels voor de Welzijn en Kinderopvang statistiek en de Groothandel statistiek. In figuren 10(a) en 10(b) zijn de regels gegeven waarvoor de percentages aan het begin van het correctieproces het hoogst zijn, en figuren 10(c) en 10(d) de regels waarvoor de percentages aan het eind van het correctieproces het hoogst zijn. Wat de invloed van de stappen in het correctieproces betreft zien we patronen vergelijkbaar met die in de percentages geschonden of niet-verifieerbare regels op het niveau van records, figuur 9. In het bijzonder valt op dat in de Groothandel statistiek bij aanvang van het correctieproces er regels zijn waar door geen enkel record aan wordt voldaan.

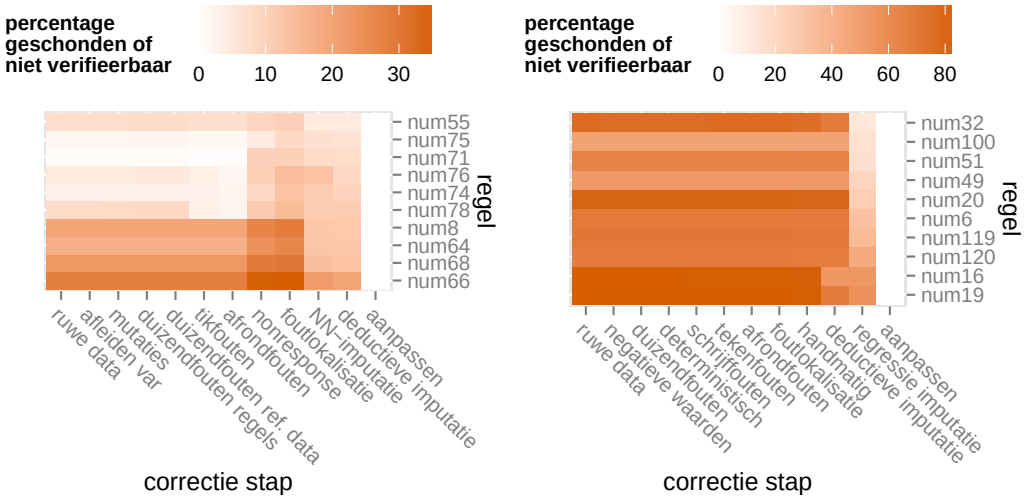
4.7 Regelschendingen op het niveau van variabelen

Tot slot kijken we naar regelschendingen op het niveau van variabelen. Daartoe bepalen we per variabele de regels die op deze variabele betrekking hebben, voor ieder van deze regels tellen we het aantal records dat deze regel schendt of waarvoor deze regel niet verifieerbaar is, en deze aantallen tellen we bij elkaar op; we refereren hiernaar met het aantal geschonden of niet-verifieerbare regels per variabele. Figuur 11 geeft het percentage geschonden of niet-verifieerbare regels per variabele voor de Welzijn en Kinderopvang statistiek en de Groothandel statistiek. Figuren 11(a) en 11(b) geven de tien variabelen die aan het begin van het correctieproces de meeste hoogste percentages hebben, en figuren 11(c) en 11(d) geven de tien variabelen met aan het eind van het correctieproces de hoogste percentages. Wat de invloed van de stappen in het correctieproces betreft zien we weer een patroon vergelijkbaar met het de percentages geschonden of niet-verifieerbare regels op het niveau van records en van regels, figuren 9 en 10.



(a) Welzijn en Kinderopvang

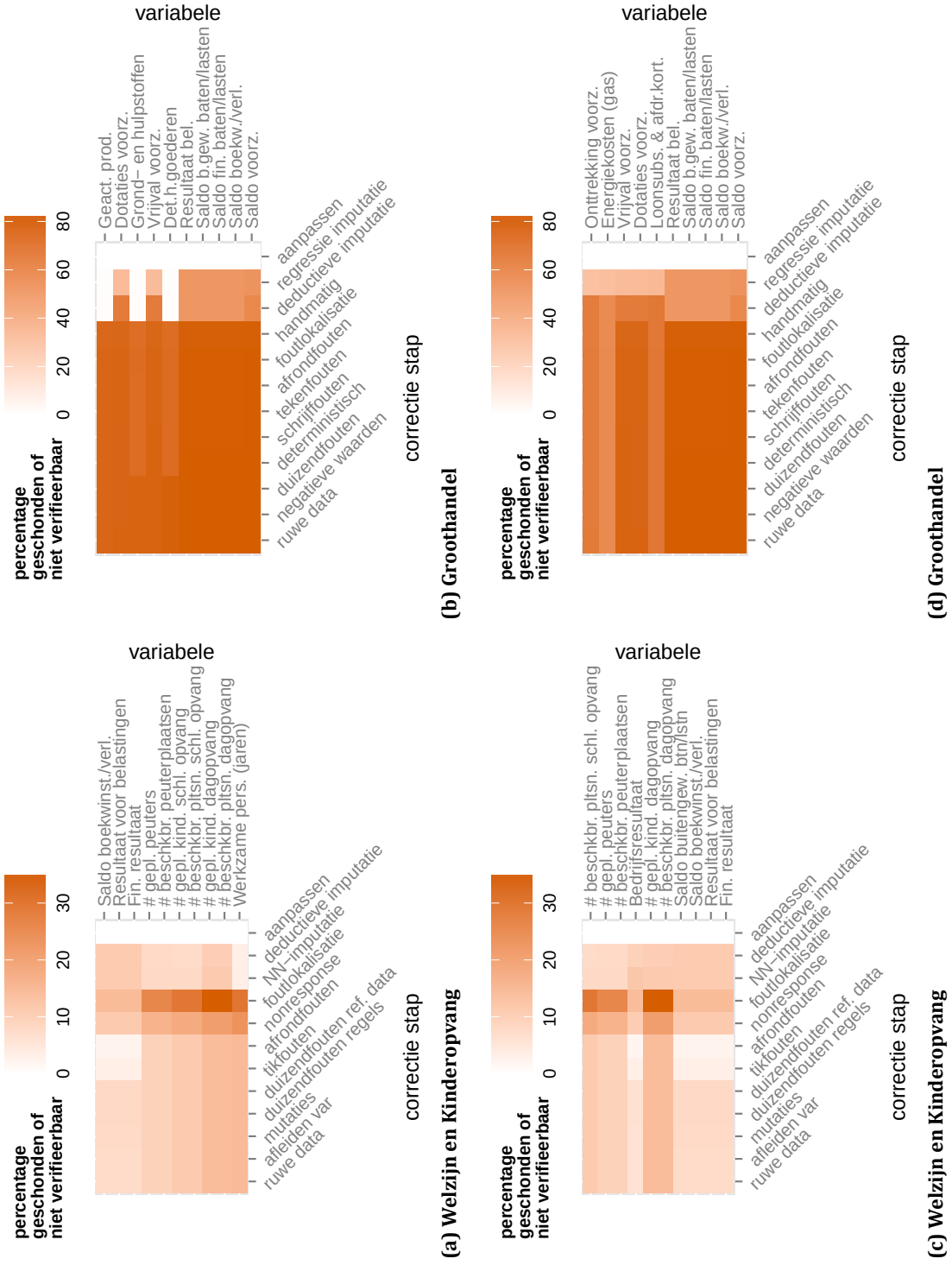
(b) Groothandel



(c) Welzijn en Kinderopvang

(d) Groothandel

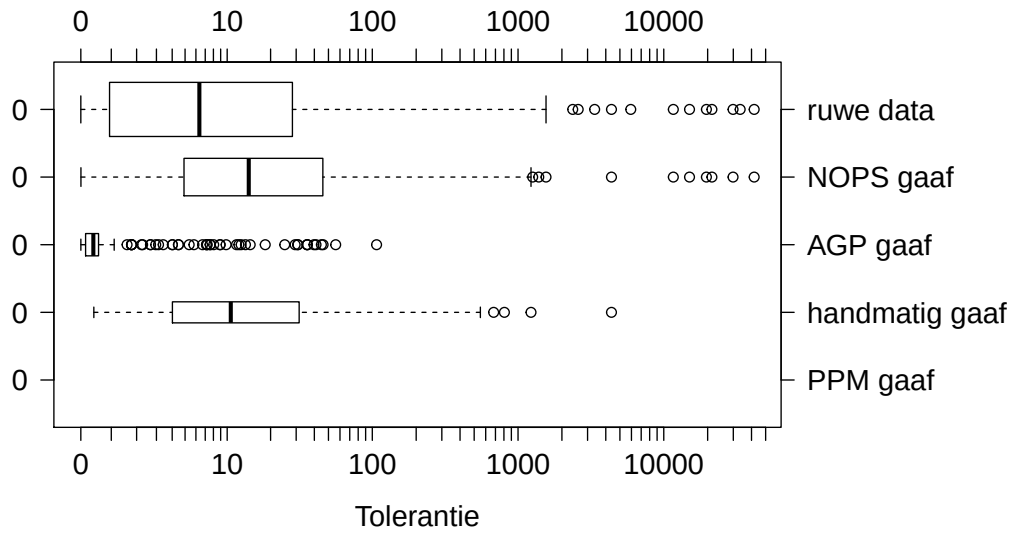
Figuur 10 Het percentage geschonden of niet-verifieerbare regels per regel over de stappen in het correctieproces met betrekking tot (a, c) de Welzijn en Kinderopvang statistiek en (b, d) de Groothandel statistiek. Regels zijn gesorteerd op het hoogste percentage aan het begin (a, b) danwel einde (c, d) van het correctieproces; alleen de tien regels met de hoogste percentages zijn weergegeven.



Figuur 11 Het percentage geschonden of niet-verifieerbare regels per variabele over de stappen in het correctieproces met betrekking tot (a, c) de Welzijn en Kinderopvang statistiek en (b, d) de Groothandel statistiek. Variabelen zijn gesorteerd op het hoogste percentage aan het begin (a, b) danwel einde (c, d) van het correctieproces; alleen de tien variabelen met de hoogste percentages zijn weergegeven.

5 Vergelijking van controle- en correctiesystemen

De hiervoor beschreven indicatoren bieden de mogelijkheid om het effect van verschillende (geautomatiseerde) controle-correctiesystemen met elkaar te vergelijken. Hier vergelijken we het effect van vier verschillende controle-correctiesystemen op de gegevens voor de statistiek Welzijn en Kinderopvang, te weten het NOPS systeem (wat wordt ingezet als onderdeel bij het controleren van gegevens voor de productiestatistiek), het AGP systeem (eerder ingezet als onderdeel bij het controleren van gegevens van de statistiek Welzijn en Kinderopvang), het effect van controle door specialisten, en het PPM systeem wat al beschreven werd in Hoofdstuk 2.



Figuur 12 data. Grafische weergave van de verdeling van positieve tolerantiewaarden voor de ruwe data en de verschillende opgeleverde gaafgemaakte data. De waarden zijn weergegeven door middel van boxplots op een pseudologaritmische schaal, de hoogte van de boxplots is evenredig met de vierkantswortel van het aantal positieve tolerantiewaarden. De getallen links geven het aantal niet te verifiëren regelschendingen weer.

We vergelijken de resulterende gaafgemaakte data met elkaar en met de ruwe data. Dit doen we door te kijken naar de verdelingen van de positieve tolerantiewaarden die geassocieerd zijn met regelschendingen, weergegeven door middel van boxplots, en het aantal niet te verifiëren regelschendingen. Het resultaat is gegeven in figuur 12. Een toelichting op de boxplot is te vinden in sectie 4.4, waar we vergelijkbare figuren hebben gemaakt voor de onderlinge stappen in een correctieproces. We zien dat ieder gaafmaakproces het aantal regelschendingen heeft gereduceert, aangezien de bijbehorende boxplots smaller zijn dan die horende bij de ruwe data, en het aantal niet-verifieerbare regelschendingen op nul laat. De meeste regelschendingen zijn opgelost in ``PPM gaaf'', de minste door ``NOPS gaaf''. Regelschendingen met relatief hoge tolerantie zijn in grotere mate opgelost in ``AGP gaaf'' dan in ``NOPS gaaf'' en ``handmatig gaaf''.

6 Discussie

In dit rapport hebben we indicatoren behandeld die de status van een dataset binnen een controle-correctieproces aangeven op basis van de waarden van de variabelen en de mate waarin aan regels is voldaan.

De indicatoren meten onder meer de mate van verandering van waarden en gemiddelden en het geschatte betrouwbaarheidsinterval rond het gemiddelde. Ook geven de indicatoren het aantal ontbrekende waarden, het aantal en de omvang van regelschendingen, en het aantal niet te verifiëren regels weer. Sommige acties in het controle-correctieproces hebben effect op verschillende indicatoren. Zo heeft de actie ``foutlokalisatie'' zowel in de statistiek Welzijn en Kinderopvang als in de statistiek van de Groothandel invloed op de geschatte betrouwbaarheid van bepaalde gemiddelden alsook op het aantal verifieerbare regels. Andere acties, zoals corectie voor duizendfouten, hebben een duidelijke invloed op de geschatte betrouwbaarheid van gemiddelden maar worden niet of nauwelijks opgemerkt in termen van het aantal veranderingen in waarden of de verandering in het aantal regelschendingen.

Gerelateerd aan het vergelijken van de invloed van verschillende acties op dezelfde dataset is de vergelijking van datasets met een verschillende voorgeschiedenis. Zo zijn ontbrekende waarden in de ruwe data van de Welzijn en Kinderopvang statistiek allemaal met nul ingevuld terwijl dit bij de Groothandel statistiek niet is gebeurd. Logischerwijs resulteert dit in verschillende weergaves bij indicatoren die het aantal ontbrekende waarden of niet te verifiëren regels meten. Maar het heeft ook invloed op andere indicatoren, zoals die die het aantal geschonden regels meten: in de Welzijn en Kinderopvang statistiek is het aantal geschonden regels bij aanvang relatief hoog terwijl dit in de Groothandel statistiek niet het geval is, en de invloed van verschillende correctiestappen in termen van regelschendingen is als gevolg duidelijker merkbaar in de Groothandel statistiek dan in de Welzijn en Kinderopvang statistiek.

6.1 Toepassingen

De indicatoren lenen zich voor het beoordelen van gegevenskwaliteit of de veranderingen daarin en het controle-correctieproces met de daarbij gehanteerde regels. We geven een viertal voorbeelden waarbij deze indicatoren van nut kunnen zijn.

Ten eerste kan men verschillende gegevensversies die door hetzelfde systeem worden behandeld met elkaar vergelijken, bijvoorbeeld voor een statistiek die periodiek wordt geproduceerd. Door indicatoren over de tijd te volgen kan een verloop in gegevenskwaliteit zoals een systematische toename in het aantal regelschendingen ontdekt worden.

Ten tweede kunnen verschillende processen die op dezelfde gegevensset worden toegepast vergeleken worden. Bijvoorbeeld, bij de statistiek Welzijn en Kinderopvang zijn grote verschillen in aantallen en grootte van regelschendingen te zien in de gegevens die door vier verschillende controle-correctie processen zijn behandeld. Het meten van verschillen tussen processen of processtappen biedt aanknopingspunten voor het optimaliseren van het process.

Ten derde kan het effect van achtereenvolgende processtappen op de gegevenskwaliteit worden beoordeeld. Hiervoor zijn op zich weer twee toepassingen aan te wijzen. Ten eerste

zegt de mate waarin een processtap een record aanpast ook iets over de kwaliteit van het record. Een record dat slechts gecorrigeerd kan worden door ingrijpende aanpassingen te doen kan misschien beter door een specialist worden behandeld, of helemaal worden weggelaten bij verdere analyse van het gegevensbestand. De indicatoren fungeren dan als "scores"voor selectief gaafmaken; ze geven aan of een record automatisch of handmatig moet worden gaafgemaakt. Ten tweede kan men een kosten-baten afweging maken voor een processtap wanneer het effect op de gegevenskwaliteit gemeten kan worden. Merk op dat ook als een processtap niets veranderd aan een gegevensbestand, er wel sprake is van een kwaliteitsmeting. Neem als voorbeeld de processtap `duizendfouten oplossen'. Als deze stap geen effect heeft op een bepaalde gegevensset, dan weten we vóór deze processtap niet of de gegevens vrij zijn van duizendfouten, en na de processtap wel. Althans binnen de aannames van het gebruikte algoritme. Processtappen kunnen dus intrinsiek de waarde van een gegevensset verhogen ook zonder een merkbaar effect te hebben omdat zij ook een vorm van validatie uitvoeren.

Ten slotte kan men het effect bepalen van de parametrisering van een proces. In het algemeen zal de uitvoer van een processtap afhangen van parameters zoals de gebruikte regelset of grenswaarden voor het detecteren van uitschieters. Het effect van verschillende parametrisaties kan direct vergeleken worden door deze indicatoren per processtap of in samenhang te bekijken. Een duidelijk voorbeeld wordt gegeven bij de statistiek van de Groothandel, waarbij in de laatste processtap (een aanpassignsalgoritme) het betrouwbaarheidsinterval rondom het gemiddelde van de variabele `Pers. uitgeleend' aanmerkelijk toeneemt. Men kan zich daarom afvragen of de regelset die wordt gebruikt tijdens en/of voor deze processtap wel afdoende is.

6.2 Open vragen

De tolerantie zoals we deze beschreven hebben in sectie 4 is een maat die aangeeft hoe goed een record aan een regel voldoet. Het is de kortste afstand van het record tot die records die aan de regel voldoen. In het geval dat er meerdere regels zijn, zoals bij de statistiek toor Welzijn en Kinderopvang en de statistiek voor Groothandel die we als voorbeelden hebben gebruikt, hebben we voor iedere regel afzonderlijk tolerantiewaarden bepaald. De tolerantiewaarden voor individuele regels geven samen een indicatie hoe goed een record aan de regels voldoet, het houdt echter geen rekening met enige afhankelijkheden tussen regels. Zo is het bijvoorbeeld niet mogelijk uit de tolerantiewaarden af te lezen of een regel een andere regel wellicht impliceert of juist tegenspreekt. Voor de hand liggend is de generalisatie van tolerantie naar meerdere regels gedefinieerd als de kortste afstand van het record naar die records die aan alle regels voldoen. Dit heeft echter als complicatie dat de afstandsmaat juist gekozen moet worden zodat rekening gehouden wordt met verschillen in de schalen en eenheden van de variabelen. Ter vergelijking, de gereduceerde vorm voor tolerantie van één regel uit sectie 4.3 kent dit probleem niet aangezien verwacht mag worden dat de variabelen al juist worden gewogen door middel van de coëfficienten die de lineaire regel definiëren. Een alternatief is om een gewogen combinatie te maken van de toleranties voor individuele regels met behulp van de Mahalanobis afstand (Mahalanobis, 1936; Aggarwal, 2013): deze afstandsmaat houdt zowel met onderlinge correlaties als verschillende schalen rekening door paarsgewijze covarianties van toleranties als gewichten te gebruiken. Een vereiste is daarbij dat de covariantiematrix van toleranties inverteerbaar is, afhankelijkheden tussen de regels mogen dus niet te sterk zijn. Duidelijke afhankelijkheden zoals duplicaten van regels kunnen in R opgelost worden met

behulp van het `editrules` package (De Jonge en Van der Loo, 2012). Hedlin (2003) gebruikt de Mahalanobis afstand om het verschil in omvang van regelschendingen voor en na een correctiestap te meten, en noemt naast het voordeel van een enkele maat voor meerdere regels als nadeel dat het een complexere uitdrukking is.

De op toleranties gebaseerde indicatoren geven onder meer een weergave van het aantal regelschendingen of het aantal niet te verifiëren regels. Het spreekt voor zich dat deze aantallen in het algemeen lager zullen zijn in het geval waarbij het aantal regels minder is. Een lager aantal regelschendingen of niet te verifiëren regels impliceert automatisch een beter resultaat. Men dient hiermee rekening te houden, in het bijzonder bij vergelijkingen waarin het aantal regels niet constant is. Een mogelijkheid is om in plaats van absolute aantallen percentages te gebruiken.

De reductie van de formule voor de tolerantie, zoals beschreven in sectie 4.3, maakt het berekenen van de tolerantie eenvoudiger voor regels die de vorm hebben van lineaire (on)gelijkheden. In het geval dat de afstandsmaat Euclidisch is, kan de reductie geïnterpreteerd worden als het gevolg van een projectie. Deze interpretatie generaliseert naar de situatie van meerdere regels, en een algoritme voor het vinden van het record dat op kortste afstand ligt van de records die aan alle regels voldoen is beschreven door Pannekoek en Zhang (2012) en geïmplementeerd in R door Van der Loo (2012, 2013). Naast lineaire (on)gelijkheden zijn logische formules een veel voorkomende vorm voor regels. Het berekenen van een tolerantie voor zulke regels is computationeel tijdrovend, en het zou winstgevend zijn als hier ook een reductie voor mogelijk is.

Referenties

Aggarwal, C. (2013). *Outlier Analysis*. New York: Springer.

Banning, R. en G. Vink (2010). Methoden en indicatoren voor het evalueren en beschrijven van controle- en correctieprocessen. Technical report, Centraal Bureau voor de Statistiek, The Hague.

Brancato, G., R. Carbini, en G. Simeoni (2009). Metadata and quality indicators to report on editing and imputation to different users. In *Conference of European statisticians*.

De Jonge, E. en M. Van der Loo (2011). Manipulation of linear edits and error localization with the editrules package. Technical Report 201120, Statistics Netherlands, The Hague.

De Jonge, E. and M. Van der Loo (2012). *editrules: R package for parsing and manipulating of edit rules and error localization.* R package version 2.5.

De Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Wiley handbooks in survey methodology. John Wiley & Sons.

Della Rocca, G., O. Luzi, M. Signore, en G. Simeoni (2005). Quality indicators for evaluating and documenting editing and imputation. In *Conference of European Statisticians*.

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

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. office for national statistics. *Journal of Official Statistics 19*(2), 177--199.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India 2*(1), 49--55.

Nordbotten, S. (1997). Metrics for the quality of editing, imputation and prediction. In *UN/ECE Work Session on Statistical Data Editing, Prague*.

Pannekoek, J. and L.-C. Zhang (2012). Optimal adjustment for inconsistency in imputed data. Technical Report 201219, Statistics Netherlands.

Scholtus, S. (2008). Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Technical Report 08015, Statistics Netherlands, Den Haag.

Scholtus, S. (2009). Automatic correction of simple typing errors in numerical data with balance edits. Technical Report 09046, Statistics Netherlands, Den Haag.

Steele, J. M. (2004). *The Cauchy--Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. New York: The Mathematical Association of America.

Van der Loo, M. (2012). *rspa: Adapt numerical records to fit (in)equality restrictions with the Successive Projection Algorithm*. R package version 0.1-1.

Van der Loo, M. (2012). *The rspa package for minimal record adjustment*. Interne nota, BPA-nummer PPM-2012-09-03-01-MPLO

Van der Loo, M. and E. De Jonge (2011). Manipulation of categorical data edits and error localization with the editrules package. Technical Report 201129, Statistics Netherlands.

Van der Loo, M. en J. Pannekoek (2013). Een automatisch controle-correctie systeem voor de zorgstatistieken. Technical report, Centraal Bureau voor de Statistiek.

Appendices

I Meest geschonden regels

I.1 Welzijn en Kinderopvang statistiek

We geven hier de meest geschonden regels uit de Welzijn en Kinderopvang statistiek, zoals vermeld in sectie 4.7.

num8 : Totaal werkn. loonlst. (jaren) $\leq$ Werkzame pers. (jaren)

num31 : $0.1 <$ Werkzame pers. (jaren)

num36 : Werkzame pers. (pers.) $\leq 20 \times$ Werkzame pers. (jaren)

num55 : Totale bedrijfsopbrengsten $< 2 \times$ Totaal bedrijfslasten

num64 : # beschkbr. peuterplaatsen $\leq$ # gepl. peuters

num66 : # beschkbr. pltsn. dagopvang $\leq$ # gepl. kind. dagopvang

num68 : # beschkbr. pltsn. schl. opvang $\leq$ # gepl. kind. schl. opvang

num71 : Werkzame pers. (pers.) = Werknemers niet op loonlijst
+ Totaal werkn. niet uitgeleend

num74 : Totaal pers. kosten = Pensioenlasten + Kosten uitzendkr.
+ Sociale voorzieningen + Overige soc. lasten
+ Opleidingskosten + Overige pers.kosten + Lonen & salarissen

num75 : Totaal bedrijfslasten = Afschr. vaste activa + Totaal pers. kosten
+ Totaal overige bedrijfslasten

num76 : Totaal bedrijfslasten + Bedrijfsresultaat
= Totale bedrijfsopbrengsten

num78 : Resultaat voor belastingen = Bedrijfsresultaat+Saldo voorzieningen
+ Saldo boekwinst./verl. + Saldo buitengew. btn/lstn + Fin. resultaat

I.2 Groothandel statistiek

We geven hier de meest geschonden regels uit de Groothandel statistiek, zoals vermeld in sectie 4.7. Met uitzondering van de variabelen IMPORTS113000, OMZETPS213300, OMZETPS213400 en OMZETPS213700 hebben alle variabelen een naam die aangeeft waar de variabele voor staat.

num6 : $0 \leq$ Onttrekking voorz.

num8 : Tot. ov. bedr.lasten = Energiekosten + Houderschapsbel.
+ Ov. kosten vervoerm. + Huisv. huur/lease + Milieuheff. & zuiv.lasten
+ Onr.z.belasting + Ov. huisvesting + Apparatuur / invent.
+ Agentenprovisie + Ov. verkoopkosten + Communicatie
+ Automatisering + Vrachtkosten + Research & devel.
+ Andere diensten + Licenties e.d. + Beheerskosten
+ Ov. kostpr.verh. bel. + Ov. alg. kosten

num9 : Tot. inkoopw. = Gr.h.goederen + Det.h.goederen
+ Grond- en hulpstoffen + Uitbesteed werk + Ov. inkoopw.

num12 : Tot. ov. bedr.opbr. = Geact. prod. + Subs. & rest.
+ Ontvangen schade-uitk. + Beh.vergoedingen + Vergoedingen uitleen
+ Ov. bedr.opbr.

num13 : Tot. pers. kosten = Kosten uitzendkr. + Kosten ov. inleen
+ Opleidingskosten + Comm.beloning + Ov. pers.kosten
+ Brutolonen / -sal. + Soc. voorz. + Pensioenlasten
+ Ov. soc. lasten

num16 : Resultaat bel. = Bedr.resultaat + Saldo boekw./verl.
+ Saldo b.gew. baten/lasten + Saldo fin. baten/lasten + Saldo voorz.

num19 : Saldo voorz. + Dotaties voorz. = Vrijval voorz.

num20 : $0 \leq$ IMPORTS113000

num32 : $0 \leq$ Dotaties voorz.

num49 : $0 \leq$ Energiekosten (gas)

num51 : $0 \leq$ Energiekosten (net)

num95 : $0 \leq$ OMZETPS213300

num96 : $0 \leq$ OMZETPS213400

num98 : $0 \leq$ OMZETPS213700

num100 : $0 \leq$ Doorber. vr.kosten

num119 : $0 \leq$ Tot. pers. kosten

num120 : Energiekosten (gas) + Energiekosten (elec.)
+ Energiekosten (net) $\leq$ Energiekosten