Automatisch opsporen van fouten in data voor officiële statistiek

Sander Scholtus

Meetfouten komen voor in praktisch alle gegevens die gebruikt worden voor statistisch onderzoek. In de officiële statistiek wordt veel aandacht besteed aan het opsporen en verbeteren van meetfouten, om te voorkomen dat deze de kwaliteit van te publiceren statistieken aantasten. Traditioneel gebeurde dit handmatig. Er bestaan ook methoden om fouten automatisch op te sporen, wat in potentie veel tijd en kosten zou kunnen besparen. Echter, de kwaliteit van automatische foutlocalisatie valt tot nu toe vaak tegen, zodat in de praktijk veel handwerk nodig blijft. In dit artikel ga ik in op een bestaande methode voor automatische foutlocalisatie, en op manieren om deze methode te verbeteren.

Gegevens die zijn verzameld voor statistisch onderzoek bevatten in de praktijk altijd meetfouten. Bij vragenlijstonderzoek kunnen fouten bijvoorbeeld optreden wanneer een respondent een vraag verkeerd begrijpt, of het antwoord niet precies weet, of zelfs bewust een verkeerd antwoord geeft. Gegevens die afkomstig zijn uit een register kunnen ook meetfouten bevatten. In dit geval is een extra bron van fouten dat personen of bedrijven er belang bij kunnen hebben om (ten onrechte) op een bepaalde manier geregistreerd te staan.

Meetfouten werken door in statistieken die uit de data worden afgeleid. In het gunstigste geval zorgen meetfouten alleen voor extra onzekerheid (variantie), maar ze kunnen ook leiden tot vertekening. Het is daarom belangrijk om bij statistisch onderzoek alert te zijn op de invloed van meetfouten en deze eventueel te corrigeren.

In de officiële statistiek is van oudsher veel aandacht besteed aan gestandaardiseerde methoden om gegevens te controleren en waar nodig te corrigeren voor meetfouten. Traditioneel werden alle data 'met de hand' gecontroleerd door inhoudelijk deskundigen. Dit was tijdrovend en kostbaar; naar schatting besteedden statistische bureaus tot 40% van hun budget aan dit controlewerk.

Vanaf de jaren tachtig van de vorige eeuw groeide het besef dat deze aanpak verbeterd kon worden. Veel individuele meetfouten zijn nauwelijks zichtbaar in een geaggregeerde statistiek. Door het handmatige controlewerk toe te spitsen op verdachte waarden die naar verwachting de grootste invloed hebben op te publiceren statistieken, kan veel tijd en geld worden bespaard terwijl de nauwkeurigheid van de statistieken nauwelijks verandert. De kwaliteit van de statistieken kan zelfs worden verhoogd door een deel van het vrijgekomen budget te besteden aan andere verbeteringen, zoals het bestrijden van non-respons. De afgelopen decennia is een standaard-methodologie ontwikkeld voor het op een effectieve manier selecteren van invloedrijke verdachte waarden (De Waal et al., 2011).

Automatische foutlocalisatie

Tegelijk met de ontwikkeling van selectiemethoden om het handmatige controlewerk terug te dringen, is er ook onderzoek gedaan naar methoden om meetfouten automatisch te verbeteren. Het idee is dat men dergelijke methoden toepast op het deel van de data dat niet voor handmatige controle wordt geselecteerd. Als meetfouten redelijk betrouwbaar worden gecorrigeerd door een automatische methode, kan de selectie voor handmatige controle verder worden ingeperkt, zodat statistieken nog efficiënter en sneller worden geproduceerd. Een ander voordeel van automatische methoden is dat ze transparanter en beter reproduceerbaar zijn dan handmatige correctie.

Om fouten in data automatisch te herkennen kan men controleregels opstellen waaraan foutloze data zouden moeten voldoen. Deze regels bevatten inhoudelijke kennis die ook (maar soms impliciet) wordt gebruikt bij handmatige controles. Met name bij bedrijfsstatistieken kunnen vaak sterke controleregels worden opgesteld, op basis van boekhoudkundige definities. Stel bijvoorbeeld dat bij een bedrijf de drie variabelen *omzet*, *kosten* en *resultaat (winst/verlies)* worden uitgevraagd. In een bepaalde bedrijfstak moeten deze variabelen voldoen aan de volgende regels:

omzet – kosten = resultaat,
omzet ≥ 0,
kosten ≥ 0,
kosten ≥ 60% × omzet.

Binnengekomen data die niet voldoen aan alle controleregels bevatten blijkbaar minimaal één fout. De eerste rij van tabel 1 toont een voorbeeld van een ingevulde vragenlijst die de eerste controleregel schendt.

In het algemeen is niet onmiddellijk duidelijk *welke* waarden fout zijn, omdat controleregels meerdere variabelen kunnen bevatten. Om fouten automatisch aan te wijzen is naast de controleregels een selectiecriterium nodig. Fellegi en Holt (1976) gaven de eerste logisch sluitende aanpak voor automatische foutlocalisatie. Hun voorstel is uitgegroeid tot de standaardaanpak in de officiële statistiek. Volgens het principe van Fellegi en Holt moet de kleinst mogelijke deelverzameling van de variabelen worden aangewezen als fout waarbij het mogelijk is om (alleen) deze variabelen aan te passen zodat voldaan wordt aan alle controleregels.

In het voorbeeld uit tabel 1 kan aan alle controleregels

	OMZET	KOSTEN	RESULTAAT
oorspronkelijke data	200	125	755
data na handmatige correctie	200	125	75
data na automatische correctie (Fellegi-Holt)	200	125	75

Tabel 1. Eerste voorbeeld van foutlocalisatie in een dataset met drie variabelen

	OMZET	KOSTEN	RESULTAAT
oorspronkelijke data	100	50	50
data na handmatige correctie	100	50	50
data na automatische correctie (Fellegi-Holt; alle controleregels meegenomen)	100	60	40

Tabel 2. Tweede voorbeeld van foutlocalisatie in een dataset met drie variabelen

worden voldaan door alleen de waarde van *resultaat* aan te passen, maar niet door alleen de waarde van *omzet* of *kosten* aan te passen. Volgens het principe van Fellegi en Holt moet daarom de oorspronkelijke waarde van *resultaat* als fout worden aangewezen (derde rij). In dit voorbeeld is dit ook de oplossing die een inhoudelijk deskundige zou kiezen (tweede rij).

Uitgaande van het principe van Fellegi en Holt komt automatische foutlocalisatie neer op een minimalisatieprobleem onder restricties. Voor een grote klasse van controleregels die in de praktijk voorkomen kan dit worden geschreven als een gemengd geheeltallig lineair programmeringsprobleem (Engelse afkorting: MILP) – een optimalisatieprobleem waarvan de restricties en de doelfunctie lineair zijn en waarbij sommige beslisvariabelen geheeltallig zijn (De Jonge & Van der Loo, 2014). In de praktijk moeten foutlocalisatieproblemen worden opgelost die maximaal enkele honderden variabelen en enkele honderden controleregels bevatten; dit lukt meestal goed met standaard-softwarepakketten voor MILP-problemen. Daarnaast zijn ook specifieke algoritmen ontwikkeld om dit foutlocalisatieprobleem op te lossen (De Waal et al., 2011).

In het voorbeeld uit tabel 1 leidde het principe van Fellegi en Holt tot dezelfde oplossing die een inhoudelijk deskundige zou kiezen. In toepassingen blijkt echter vaak dat data die automatisch zijn gecorrigeerd volgens dit principe systematisch verschillen van data die handmatig zijn gecorrigeerd. Dit beperkt de toepasbaarheid van automatische controle en correctie in de praktijk. In mijn promotieonderzoek (Scholtus, 2018) heb ik gewerkt aan twee uitbreidingen van het principe van Fellegi en Holt die deze aanpak flexibeler maken, zodat de uitkomsten beter aansluiten bij die van inhoudelijk deskundigen.

	OMZET	KOSTEN	RESULTAAT
oorspronkelijke data	100	60.000	40.000
data na handmatige correctie	100	60	40
data na automatische correctie (Fellegi-Holt; alleen harde controleregels meegenomen)	100	60.000	–59.000

Tabel 3. Derde voorbeeld van foutlocalisatie in een dataset met drie variabelen

	OMZET	KOSTEN	RESULTAAT
oorspronkelijke data	90	130	40
data na handmatige correctie	130	90	40
data na automatische correctie (Fellegi-Holt)	170	130	40

Tabel 4. Vierde voorbeeld van foutlocalisatie in een dataset met drie variabelen

Zachte controleregels

In het bovenstaande voorbeeld komen de eerste drie controleregels direct voort uit de definities van de variabelen. Dit zijn 'harde' restricties: elke vragenlijst die een van deze regels schendt bevat gegarandeerd een fout. De vierde regel is 'zacht': een vragenlijst waarin de opgegeven kosten minder dan 60% van de opgegeven omzet bedragen is verdacht, maar bevat niet per se een fout. De grens van 60% is gekozen omdat verwacht wordt dat deze restrictie geldt voor de meeste bedrijven in deze branche, maar er zijn uitzonderingen.

Zachte regels worden veel gebruikt tijdens handmatige controles. Echter, het principe van Fellegi en Holt ziet alle regels als hard. Bij automatische foutlocalisatie moeten zachte regels dus ofwel worden opgevat als harde regels, of worden weggelaten. Beide opties zijn niet ideaal, zoals blijkt uit de voorbeelden in tabel 2 en 3. De vragenlijst uit tabel 2 schendt de zachte controleregel. Bij handmatige controle is geconcludeerd dat deze (kleine) schending voor dit bedrijf terecht is. Bij automatische foutlocalisatie worden de data wel aangepast. In dit voorbeeld zou het beter zijn geweest om de zachte regel weg te laten. Het weglaten van de zachte regel kan echter ook leiden tot ongewenste aanpassingen, zoals blijkt uit tabel 3.

Een betere oplossing is mogelijk door onderscheid te maken tussen harde en zachte controleregels. In de besliskunde is een bekende aanpak om zachte restricties mee te nemen in een minimalisatieprobleem, een extra term aan de doelfunctie toe te voegen die kosten verbindt aan schendingen van deze restricties. Deze aanpak is direct toepasbaar op de MILP-formulering van het principe van Fellegi en Holt (Scholtus, 2013). De oplossing van het foutlocalisatieprobleem hoeft dan niet altijd aan alle zachte controleregels te voldoen; een eventuele schending van een zachte regel wordt afgewogen tegen het aanwijzen van een extra fout om deze schending op te heffen.

Algemenere aanpassingen

Een tweede belangrijke bron van systematische verschillen tussen handmatige en automatische foutlocalisatie is dat het principe van Fellegi en Holt ervan uitgaat dat elke variabele onafhankelijk wel of geen fout bevat. Inhoudelijk deskundigen zien echter dat respondenten ook fouten maken waarbij meerdere variabelen tegelijk betrokken zijn. Ze passen dan correcties toe die vanuit het oogpunt van Fellegi en Holt suboptimaal zijn. Bekijk bijvoorbeeld tabel 4. De inhoudelijke deskundige concludeert hier dat de respondent de bedragen bij *omzet* en *kosten* heeft verwisseld; hij/zij ziet dit als één fout. De automatische methode ziet dit als twee aanpassingen en kiest daarom een andere oplossing waarbij maar één waarde wordt aangepast.

In Scholtus (2016) is een uitbreiding van het principe van Fellegi en Holt voorgesteld die ruimte biedt aan dit soort correcties. Definieer per toepassing een verzameling toegelaten aanpassingen. Inhoudelijk komen deze aanpassingen overeen met correcties voor fouttypen waarvan men verwacht dat ze regelmatig voorkomen in de data. Dit kunnen fouten in individuele waarden zijn, zoals in de oorspronkelijke formulering van Fellegi en Holt, maar ook ingewikkeldere correcties zoals omwisselingen en overhevelingen van bedragen. Technisch is elke aanpassing een affiene transformatie van de data, waarbij eventueel vrije parameters worden ingevoerd. (In het oorspronkelijke probleem van Fellegi en Holt is de nieuwe waarde voor een als fout aangewezen variabele zo'n vrije parameter.) Het criterium voor foutlocalisatie is nu om te zoeken naar de combinatie van het kleinste aantal toegelaten aanpassingen die ervoor zorgt dat vol-

daan wordt aan alle controleregels. Het oorspronkelijke principe van Fellegi en Holt is een speciaal geval van deze formulering, voor een specifieke verzameling toegelaten aanpassingen.

Recent is gebleken dat ook dit uitgebreide foutlocalisatieprobleem is te schrijven als MILP-probleem (Daalmans en Scholtus, 2018). Het kan daarom op dezelfde manier worden opgelost als het oorspronkelijke probleem van Fellegi en Holt. Resultaten op data met gesimuleerde fouten suggereren dat hiermee een verbetering in de kwaliteit van automatisch gecorrigeerde data kan worden bereikt, mits men in staat is om relevante toegelaten aanpassingen te vinden.

Slot

De twee besproken uitbreidingen maken de automatische foutlocalisatie flexibeler, maar ze introduceren ook extra keuzes die per toepassing moeten worden gemaakt, zoals de precieze kostenterm voor geschonden zachte regels en de keuze van de toegelaten aanpassingen. Vervolgonderzoek zal zich richten op het uitwerken en testen van deze methoden voor concrete toepassingen bij het Centraal Bureau voor de Statistiek.

Literatuur

Daalmans, J., & Scholtus, S. (2018). *A MIP Approach for a Generalised Data Editing Problem*. Discussion paper, Den Haag: Centraal Bureau voor de Statistiek.

Fellegi, I.P., & Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, *71*, 17–35.

Jonge, E. de, & Loo, M. van der (2014). *Error Localization as a Mixed Integer Problem with the Editrules Package*. Discussion paper 2014-07. Den Haag: Centraal Bureau voor de Statistiek.

Scholtus, S. (2013). Automatic Editing with Hard and Soft Edits. *Survey Methodology*, *39*, 59–89.

Scholtus, S. (2016). A Generalized Fellegi-Holt Paradigm for Automatic Error Localization. *Survey Methodology*, *42*, 1–18.

Scholtus, S. (2018). *Editing and Estimation of Measurement Errors in Administrative and Survey Data*. Proefschrift, Amsterdam: Vrije Universiteit. http://dare.ubvu.vu.nl/handle/1871/55568.

Waal, T. de, Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

Sander Scholtus is methodoloog bij het Centraal Bureau voor de Statistiek in Den Haag. In maart 2018 is hij cum laude gepromoveerd aan de Vrije Universiteit op een proefschrift dat onder andere gaat over het onderwerp van dit artikel.
E-mail: s.scholtus@cbs.nl.

ONBEKEND MAAKT ONBEMIND

Het is inmiddels alweer een jaar geleden dat aan onze verliefdheid op 'het algoritme' een eind kwam.[1] We hielden het niet langer vol. Het algoritme zorgde voor filterbubbels, beïnvloedde verkiezingen, propageerde vooroordelen en elimineerde bovenal onze privacy. Cathy O'Neil heeft ons de ogen geopend, het 'algoritme' is manipulatief en heeft niet het beste met ons voor.[2] Nee, het kon niet langer zo, het moest stoppen. Nu we een jaar verder zijn, komt het besef dat de breuk wellicht wat impulsief was. Zijn we te generalistisch geweest door alle algoritmes in de ban te doen? Maar hoe kunnen we elkaar dan weer vertrouwen?

In mijn werk als *analytics consultant* kom ik, jammer genoeg, steeds meer organisaties tegen die het vertrouwen in het gebruik van data en algoritmes aan het verliezen zijn. In de afgelopen jaren hebben ze, veelal op aangeven van strategieconsultants en technologieleveranciers, geïnvesteerd in opslagcapaciteit en analysesoftware en zijn ze als een dolle data gaan verzamelen en analyseren. Dit heeft geleid tot een overspannen vraag naar statistici en econometristen terwijl al deze inspanningen tot niet meer dan een open-deurenshow[3] van inzichten hebben geleid. Er zijn ook maar weinig echt impactvolle toepassingen van data science en kunstmatige intelligentie in de praktijk te vinden. Andrew Ng geeft aan dat bijna alle economische waarde die gegenereerd wordt met machine-leren op het conto komt van *supervised learning*[5] toegepast op online *display ad*.[4] Een kleine verbetering van de *click trough rates* kan in deze industrie veel geld opleveren.

Ondanks dat deze toepassing van machine-leren kennelijk veel oplevert, moet ik bekennen dat ik nog steeds niet erg onder de indruk ben van de relevantie van de advertenties die ik online te zien krijg. Ik blijf me verbazen dat er in de media zoveel aandacht geschonken wordt aan de beloftes van data science en kunstmatige intelligentie, terwijl tastbare resultaten schaars zijn. De voorbeelden zijn beperkt en hebben niet de potentie ook in de praktijk veel impact te hebben. Zeg nu zelf, zou je een algoritme[6] dat GO kan spelen en winnen de opdracht geven je auto te besturen of je agenda te beheren? Vragen naar bewijs voor de gouden bergen die worden beloofd[7] worden bij voorkeur genegeerd, het lijkt wel een religie.

Dat de inzet van statistiek en operations research juist veel impact heeft, komt minder vaak over het voetlicht. Onze overheid bespaart miljarden euro's doordat we met statistiek en operations research kunnen vaststellen welke dijkhoogte[8] optimaal is en de hoge benutting van ons spoornet kan alleen bereikt worden omdat de NS operations research gebruikt om treinen te *schedulen*. Begin juli was ik als jurylid van de EURO Excellence in Practice Award9 (EEPA) op de EURO-conferentie in Valencia. Ik vond het fantastisch om te zien hoeveel praktische toepassingen van operations research gepresenteerd werden en te horen welke impact die hebben gehad. Toepassingen als het opstellen van een eerlijk spelschema voor de kwalificatie van het WK-voetbal, het bepalen van de beste manier om UNESCO-erfgoed te conserveren, het optimaal aansluiten van offshore windmolens op het energienet, het dynamisch plannen van melkcollectieritten, het verbeteren van het kinderbeschermingsbeleid en het bepalen van een marktmechanisme voor de verkoop van visserijrechten. Stuk voor stuk praktische toepassingen met veel impact, zowel maatschappelijk als economisch. Ik heb er alleen niets over gelezen in de media.

Ondanks dat ik al lang in ons vakgebied werkzaam ben blijf ik het lastig vinden aan iemand van buiten ons vakgebied uit te leggen wat operations research nu precies is. Voor hen is er geen verschil tussen operations research, data science, analytics, machine learning of kunstmatige intelligentie. In elk van deze disciplines gebruik je vanuit hun optiek data en wiskunde om inzichten te genereren die besluitvorming ondersteunen. Als gevolg projecteren ze hun teleurstellende ervaringen op allemaal en dat is niet terecht. Om hun vertrouwen te herwinnen is het belangrijk ze te laten zien wat de impact van operations research werkelijk is. Dat kan in mijn optiek het beste aan de hand van voorbeelden zoals de inzendingen voor de EEPA, de Franz Edelman Award[10] of natuurlijk je eigen ervaringen. Het is belangrijk dat duidelijk wordt wat operations research te bieden heeft. Zeker nu de wereld steeds complexer wordt en veranderingen zich sneller aandienen is de behoefte aan inzicht en op feiten gebaseerde, gestructureerde besluitvorming groter dan ooit. Wij moeten de wereld laten weten dat operations research daarbij een instrumentele rol kan spelen. Het verleden heeft dat al vele malen bewezen en – je kunt me vertrouwen – in dit geval zijn resultaten uit het verleden een garantie voor de toekomst.

Noten

1. Https://www.wired.com/story/2017-was-the-year-we-fell-out-of-love-with-algorithms/
2. https://www.wired.com/2016/10/big-data-algorithms-manipulating-us/
3. Een van mijn klanten vatte de resultaten van het data science initiatief binnen zijn organisatie zo samen
4. https://www.mckinsey.com/featured-insights/artificial-intelligence/how-artificial-intelligence-and-data-add-value-to-businesses
5. https://en.wikipedia.org/wiki/Supervised_learning
6. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/
7. https://www.gartner.com/newsroom/id/3872933
8. https://www.cpb.nl/persbericht/3213331/franz-edelman-award-voor-project-optimale-dijkhoogtes-nederland
9. https://www.euro-online.org/web/pages/209/excellence-in-practice-award-eepa
10. https://www.informs.org/Recognizing-Excellence/INFORMS-Prizes/Franz-Edelman-Award

John Poppelaars is Practice leader of the Advanced Analytics and Business Intelligence unit of BearingPoint.
E-mail: john.poppelaars@bearingpoint.com