AUTOMATISCH CORRIGEREN van inconsistenties in de officiële statistiek

Officiële statistieken moeten consistent zijn. Dat betekent dat statistische resultaten hetzelfde moeten zijn in verschillende publicaties. Consistente uitkomsten ontstaan niet vanzelf. Vaak zijn er meerdere bronnen beschikbaar voor één statistiek. Afhankelijk van de gekozen bron kan er iets anders uitkomen. Het proces om tot consistente uitkomsten te komen heet ook wel macro-integratie. In dit artikel wordt uitgelegd dat in toenemende mate wiskundige methoden worden ingezet voor macro-integratie.

Jacco Daalmans

Stel je bezoekt met vrienden een café. De rekening wordt gesplitst. Iedereen legt een bedrag op tafel dat hij of zij denkt te hebben besteed. Na telling blijkt dat er te weinig geld ligt. Of anders gezegd: de hoeveelheid geld is niet 'consistent' met het bedrag op de rekening. Meestal wordt dan eerst gecontroleerd of er een fout is gemaakt. Heeft de ober te veel drankjes aangeslagen of zijn de bezoekers een schaal bitterballen vergeten? Als de oorzaak niet wordt gevonden dan kan een ingewikkelde discussie ontstaan, waarna alsnog één bedrag wordt afgerekend.

In de officiële statistiek speelt een vergelijkbaar probleem. Met 'officiële statistiek' worden cijfers bedoeld die door een officiële instantie, zoals het Centraal Bureau voor de Statistiek (CBS), worden gepubliceerd. Officië-

le statistieken moeten hetzelfde zijn in alle publicaties waarin deze voorkomen. Het aantal werklozen mag niet 300.000 zijn in de ene publicatie en 400.000 in een andere, uitgaande van een gelijke definitie van werkloosheid. Een verschil in uitkomsten zorgt voor verwarring en dit strookt niet met het doel van statistische bureaus om onbetwistbare statistiek te leveren.

Meer in het algemeen betekent consistentie dat statistieken aan onderlinge relaties voldoen. Een voorbeeld is dat twaalf maandcijfers optellen tot één jaarcijfer. Als hier niet aan wordt voldaan, kan een gebruiker naast het gegeven jaarcijfer, een alternatief jaarcijfer afleiden door maandcijfers op te tellen. Zo ontstaat verwarring over het 'ware' jaarcijfer.

Consistentie van statistische output ontstaat niet vanzelf. Data die een statistisch bureau verzamelt komen uit tal van bronnen, ieder met hun eigen onzuiverheden. Zo komen steekproeffouten, nonresponsfouten en meetfouten voor. Macro-integratie is een proces om de verschillen weg te werken. De uitkomsten van verschillende statistieken worden gecorrigeerd, om ze beter op elkaar af te stemmen. Men corrigeert bijvoorbeeld maandcijfers, om te zorgen dat twaalf maandcijfers optellen tot één jaarcijfer. Men zou kunnen denken dat door macro-integratie 'fouten' ontstaan, omdat de resultaten afwijken van de afzonderlijke bronnen. Dit is echter niet het geval, omdat we te maken hebben met bronnen met onzuiverheden. Door het confronteren van de verschillende bronnen, kan juist een nauwkeuriger beeld ontstaan dan door bronnen in afzondering te beschouwen.

Een toepassing van macro-integratie heeft gelijkenissen met het gesplitst betalen van een rekening. Eerst worden grote verschillen, met een aanwijsbare oorzaak, gecorrigeerd, daarna de overgebleven, kleinere verschillen. Hieronder gaat de aandacht alleen uit naar de tweede stap. We zullen zien dat wiskundige methoden hiervoor van grote waarde zijn en dat deze steeds meer worden toegepast in de officiële statistiek.

Nationale Rekeningen

Een traditionele toepassing van macro-integratie vindt plaats bij de Nationale Rekeningen. De Nationale Rekeningen bestaan uit een aantal zeer gedetailleerde tabellen, die het economisch proces van een land beschrijven. In die tabellen worden macro-economische indicatoren, zoals productie, consumptie, investeringen, import en export gepubliceerd voor een groot aantal goederen en diensten. Dit gebeurt zowel op kwartaal- als op jaarbasis. De tabellen worden samengesteld uit tal van bronnen. De integratie van de data is lastig, vanwege de grote samenhang tussen cijfers. Wanneer één cijfer moet worden aangepast, betekent dit doorgaans ook dat veel gerelateerde cijfers moeten worden gewijzigd.

Van oudsher werden informele methoden toegepast om Nationale Rekeningen consistent te maken. Dat betekent dat experts correcties aanbrengen in grote tabellen op basis van vakinhoudelijke kennis. Hoewel zo'n aanpak meestal goed werkt, is de toepassing arbeidsintensief. Bovendien zijn de uitkomsten lastig te reproduceren. Als alternatief voor informele methoden, zijn er ook formele methoden beschikbaar, gebaseerd op een wiskundige methode. Formele methoden hebben als voordeel dat deze reproduceerbare uitkomsten geven. Het mechanisme van de correcties ligt immers vast.

Wiskundige methoden

Een bekende wiskundige methode voor integratie van Nationale Rekeningen is die van Stone et al. (1942). Wiskundig gezien is dit een gewogen kwadratisch optimaliseringsprobleem met lineaire restricties. De lineaire restricties geven de eisen weer waar cijfers aan moeten voldoen. De doelfunctie minimaliseert een (gewogen) som van correcties, die men moet uitvoeren om aan de restricties te voldoen.

Een belangrijke eigenschap van Stone's methode is dat betrouwbaarheidsgewichten mee kunnen worden genomen. Dit maakt het mogelijk om de meest nauwkeurige cijfers het minst aan te passen en vice versa. Niet alleen sluit zo'n aanpak het best aan bij onze intuïtie, maar ook wiskundig is te bewijzen dat op die manier de meest nauwkeurige resultaten worden verkregen. Dit betekent echter wel dat een inschatting moet worden gemaakt over de nauwkeurigheid van de broncijfers.

Naast de methode van Stone bestaan er nog veel andere methoden in de literatuur, met verschillende toepassingsmogelijkheden. Voor het consistent maken van reeksen kwartaal- en jaarcijfers is bijvoorbeeld de methode van Denton (1971) beschikbaar.

De methoden, die in de vorige eeuw zijn ontwikkeld, zijn niet meteen grootschalig toegepast op statistische bureaus. Eén van de redenen hiervoor is de stand van de IT. Toepassing van een automatische data-integratie methode op de aanbod- en gebruik tabellen, zoals die momenteel op het CBS worden samengesteld, betekent bijvoorbeeld dat een optimaliseringsprobleem met meer dan 500.000 variabelen moet worden opgelost. Dit stelt eisen aan de software. Tegenwoordig is efficiënte software beschikbaar voor het oplossen van grote kwadratische optimaliseringsproblemen. Deze software kan relatief eenvoudig op een standaard PC worden toegepast. Een optimaliseringsprobleem met 500.000 variabelen vormt geen enkel probleem. Voorbeelden van commerciële software voor kwadratische optimalisering zijn: CPLEX, XPRESS, GARUBI en AIMMS.

Een tweede beperkende factor voor toepassing van automatische methoden bestaat uit de begrensde mogelijkheden van de methodologie. Zo kunnen de meest eenvoudige methoden alleen lineaire gelijkheidsrestricties aan. Voor de toepassing bij de Nationale Rekeningen is het nodig om een breder scala van relaties te kunnen modelleren. Een analyse van het integratieproces wees uit dat in ieder geval de volgende drie soorten van functionaliteiten toepasbaar moeten zijn.

Ten eerste is het nodig om ongelijkheidsrestricties mee te nemen, zoals de niet-negativiteitsrestrictie. Veel economische cijfers kunnen geen negatieve waarde aannemen. Dit moet men dan ook kunnen afdwingen in een automatische methode.

Ten tweede moet het mogelijk zijn om zogenaamde ratio-restricties op te stellen. In economische rekeningenstelsels zijn verhoudingen tussen twee cijfers vaak erg belangrijk. Een voorbeeld is de verhouding tussen productie van een bepaald goed en het verbruik van de benodigde grondstoffen. Voor de productie van 1 kg kaas is bijvoorbeeld 10 liter melk nodig. Het is nodig dat men restricties op kan leggen aan de uitkomsten van een ratio.

Ten derde moet de mogelijkheid bestaan om zogenaamde zachte relaties mee te nemen. Het gaat namelijk om relaties die sturend zijn, dus niet bindend. Of in andere woorden, die bij benadering moeten gelden. Een voorbeeld is dat verhouding tussen lonen en productie ongeveer hetzelfde moet zijn als vorig jaar.

Nieuwe methode

In een hoofdstuk van mijn proefschrift (Daalmans 2019) beschrijf ik, met Reinier Bikker en Nino Mushkudiani, een uitbreiding van de oorspronkelijke Denton methode (Bikker et al. 2013). De nieuwe methode combineert verschillende functionaliteiten van bestaande methoden en is specifiek gericht op de behoeften van de Nationale Rekeningen. De drie soorten van relaties, die hierboven zijn genoemd, kunnen bijvoorbeeld worden meegenomen.

De voorgestelde methode wordt al enige tijd daadwerkelijk toegepast voor de integratie van Nationale Rekeningen op het CBS. Daarmee is een groter deel van het productieproces geautomatiseerd en is de reproduceerbaarheid van de officiële statistiek vergroot. Hoewel de uitgebreide Denton specifiek is ontworpen voor de Nationale Rekeningen, is de methode ook toepasbaar voor andere statistieken waarin consistente tabellen moeten worden samengesteld. Het CBS past een vergelijkbare methode bijvoorbeeld ook toe voor energiestatistieken. De methodologie voor macro-integratie wordt steeds verder uitgebreid. Hierdoor ontstaan ook steeds nieuwe toepassingsmogelijkheden. Het splitsen van een caférekening zullen we echter zelf moeten blijven doen.

Literatuur

Bikker, R. P., Daalmans, J. A., & Mushkudiani, N. (2013). Benchmarking large accounting frameworks: a generalized multivariate model. *Economic Systems Research*, *25*, 390–408.

Daalmans, J. A. (2019). *Pushing the boundaries for automated data reconciliation in official statistics* (PhD Thesis). Tilburg University, Tilburg.

Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: An Approach based on quadratic minimization. *Journal of the American Statistical Association, 66*(333), 99–102.

Stone, J. R. N., Champerowne, D. G., & Meade, J. E. (1942). The Precision of National Income Estimates. *Reviews of Economic Studies, 9*, 111–135.

Jacco Daalmans is methodoloog bij het Centraal Bureau voor de Statistiek. In maart 2019 is hij gepromoveerd aan de Universiteit van Tilburg, op een proefschrift dat onder andere gaat over het onderwerp van dit artikel.
E-mail: j.daalmans@cbs.nl