GEDETAILLEERDE EN TIJDIGE STATISTIEKEN OVER DE NEDERLANDSE SAMENLEVING

Officiële statistieken gepubliceerd door statistiekbureaus zijn traditioneel gebaseerd op schattingsmethoden uit de steekproeftheorie die parametervrij zijn. Het nadeel van deze technieken is dat alleen met relatief grote steekproefomvangen voldoend nauwkeurige cijfers kunnen worden gemaakt. Dit leidt ertoe dat met deze methoden alleen cijfers op een betrekkelijk hoog aggregatieniveau voor een vrij lange referentieperiode kunnen worden gepubliceerd. Recente ontwikkelingen maken modelgebaseerde schattingsmethoden steeds aantrekkelijker voor statistiekbureaus om meer gedetailleerde schattingen te maken die sneller beschikbaar komen omdat ze over een kortere verslagperiode gaan. Hierdoor wordt statistische informatie voor veel gebruikers en beleidsmakers een stuk relevanter. In dit artikel beschrijven we het gebruik en de ontwikkelingen van multilevel- en tijdreeksmodellen om officiële statistieken in meer detail en sneller te kunnen produceren.

Jan van den Brakel, Harm Jan Boonstra, Sabine Krieg & Marc Smeets

De taak van statistiekbureaus is het verzamelen en publiceren van betrouwbare statistische informatie over sociaaleconomische ontwikkelingen in een samenleving. Hierbij gaat het om variabelen die gedefinieerd zijn als totalen of gemiddelden. Bijvoorbeeld het percentage werklozen uitgesplitst naar regio of demografische achtergrondkenmerken. Veel statistiekbureaus, waaronder het Centraal Bureau voor de Statistiek (CBS), verzamelen de benodigde data (voor zover niet beschikbaar via registers) via kanssteekproeven en maken gebruik van traditionele 'design'-gebaseerde schattingsmethoden voor het maken van statistische informatie. Deze methoden zijn

hoofdzakelijk gebaseerd op het kansmechanisme van het steekproefontwerp. Statistische modellen spelen hierbij geen of slechts een ondergeschikte rol. Dergelijke schatters corrigeren voor de ongelijke insluitkansen uit het steekproefontwerp en voor selectieve non-respons.

De precisie van deze schatters kan verbeterd worden door gebruik te maken van gerelateerde hulpinformatie waarvan de populatietotalen bekend zijn uit bijvoorbeeld registers. Deze schatters worden gemotiveerd met een lineair regressiemodel dat het verband tussen de doel- en hulpvariabelen beschrijft en zijn in de literatuur bekend als gegeneraliseerde regressie (GREG) schatters. Hoewel GREG-schatters worden gemotiveerd door een lineair regressiemodel is de methodiek design-gebaseerd omdat eigenschappen van de schatters, zoals verwachting en variantie, worden geëvalueerd met betrekking tot het kansmechanisme van het steekproefontwerp in plaats van het veronderstelde lineaire model.

In de literatuur zijn ook veel schatters voor eindige populatieparameters bekend die gebaseerd zijn op een statistisch model. Over het algemeen zijn statistiekbureaus, waaronder het CBS, terughoudend in het gebruik van deze modelgebaseerde schattingsmethoden. Bij veel statistiekbureaus is het uitgangspunt dat officiële statistieken, die worden gebruikt voor het nemen van beleidsmatige beslissingen, bij voorkeur zijn gebaseerd op empirie en data en zo min mogelijk op modelaannames waarvan de validiteit moeilijk verifieerbaar is.

Design-gebaseerde schatters zijn met name bij grote steekproefomvangen geschikt voor het maken van statistische informatie omdat in deze situatie design-varianties klein zijn. Bij kleine steekproefomvangen zijn deze schatters vanwege grote design-varianties vaak onbruikbaar. De relevantie van statistische informatie neemt echter toe met de mate van detail, frequentie en tijdigheid van de cijfers. Cijfers op een gedetailleerd regionaal niveau op maandbasis zijn doorgaans relevanter dan cijfers op nationaal niveau op jaarbasis. Gedetailleerde uitsplitsingen in combinatie met een korte referentieperiode leiden snel tot situaties waarbij slechts weinig waarnemingen beschikbaar zijn. Dan heeft het de voorkeur om informatie te gebruiken uit andere bronnen, deelpopulaties of perioden, die via een statistisch model aan elkaar gerelateerd worden. Deelpopulaties of perioden waarvoor onvoldoende waarnemingen beschikbaar zijn om betrouwbare design-gebaseerde schattingen te maken worden kleine domeinen genoemd. Rao en Molina (2015) geven een uitgebreid overzicht van methoden voor het schatten over kleine domeinen.

Recente ontwikkelingen maken modelgebaseerde schattingsmethoden steeds aantrekkelijker voor statistiekbureaus. Naast de toenemende vraag naar gedetailleerde en tijdige cijfers is er een constante druk op statistiekbureaus om de kosten en lastendruk voor de berichtgevers te verlagen door meer gebruik te maken van informatie uit registers en bigdatabronnen.

Kleindomeinschatters voor de beroepsbevolking

Literatuur over kleindomeinschatters is hoofdzakelijk gebaseerd op multilevelmodellen die gebruik maken van cross-sectionele correlaties tussen domeinen met als doel de precisie van een domeinschatter te verbeteren met informatie uit andere domeinen. De meeste surveys op statistiekbureaus worden bovendien herhaaldelijk uitgevoerd. Tijdreeksmodellen zijn dan relevant omdat hiermee informatie uit voorgaande perioden gebruikt kan worden om de precisie van de schatting voor de laatste periode te verbeteren. Door tijdreeksen voor meerdere domeinen te modelleren in één multivariaat tijdreeksmodel ontstaat de mogelijkheid om zowel cross-sectionele als temporele correlaties te gebruiken om de precisie van domeinschatters te verbeteren.

Onderzoek naar het gebruik van dit soort modelgebaseerde schatters op het CBS volgde twee aanpakken. De eerste aanpak was het ontwikkelen van een cross-sectioneel multilevel model voor het maken van jaarcijfers over de beroepsbevolking op gemeenteniveau, gebaseerd op de Enquête Beroepsbevolking (EBB). Hiervoor wordt een model op persoonsniveau toegepast dat bestaat uit drie componenten. De eerste component is een regressiecomponent zoals bij de GREG-schatters waarin veel achtergrondkenmerken worden meegenomen. De tweede

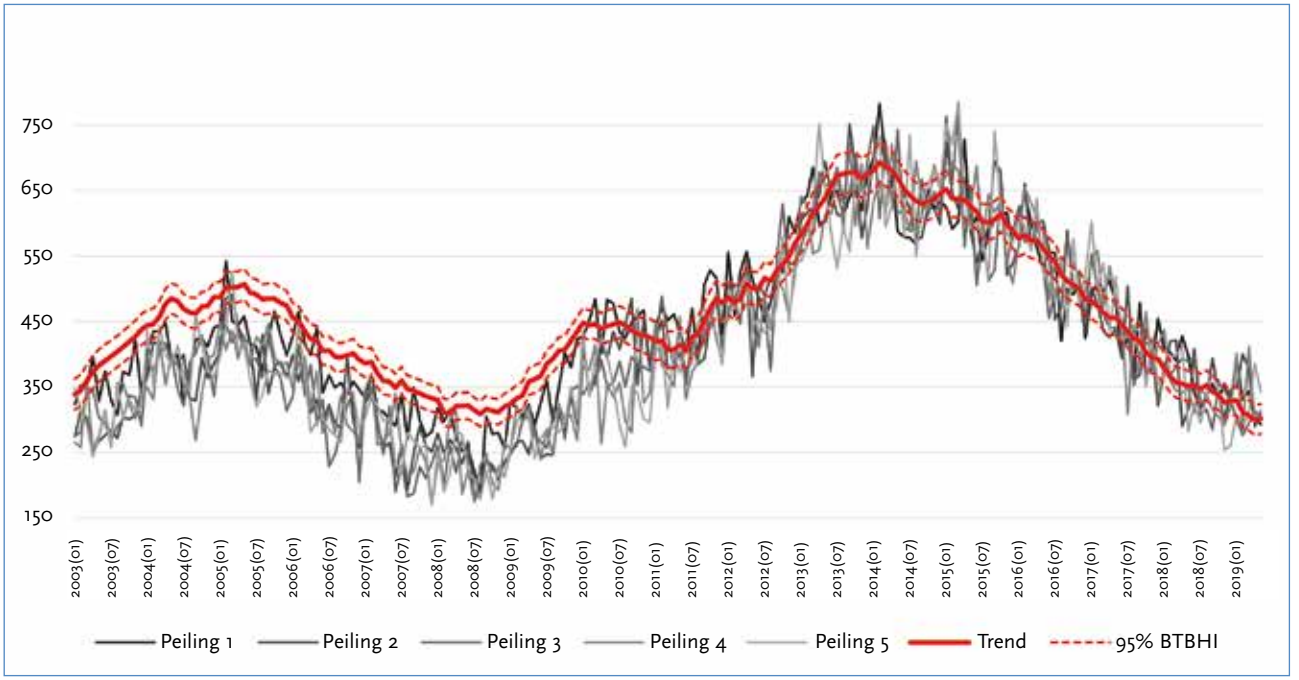
component bestaat uit de gemeente-effecten. Vanwege de geringe hoeveelheid data in veel gemeenten worden de gemeente-effecten gemodelleerd in een tweede 'level' van het model, waardoor effectief informatie uit andere gemeenten wordt gebruikt om de schattingen voor een specifieke gemeente te verbeteren. De derde component is een ruisterm om de resterende variatie te verklaren. Het multilevel model wordt jaarlijks geschat en sinds 2015 worden hiermee de officiële gemeentelijke cijfers over de beroepsbevolking gemaakt.

In een tweede aanpak is een tijdreeksmodel ontwikkeld voor het schatten van maandcijfers over de beroepsbevolking op basis van de EBB waarbij gebruik gemaakt wordt van temporele correlaties. De EBB is een doorlopend onderzoek gebaseerd op een roterend panelontwerp. De respondenten worden in vijf peilingen waargenomen, steeds met een interval van één kwartaal. Dit survey ontwerp heeft drie problemen:

- De steekproefomvang is te klein om met de GREG-schatter voldoend precieze schattingen te maken voor maandcijfers over de beroepsbevolking.
- Er zijn systematische verschillen tussen de opeenvolgende peilingen. De werkloze beroepsbevolking wordt bijvoorbeeld systematisch hoger geschat met de data van de eerste peiling ten opzichte van de daaropvolgende vier peilingen. Dit fenomeen wordt in de literatuur 'rotation group bias' (RGB) genoemd (Bailar, 1975).
- Door veranderingen in de manier waarop data worden verzameld en het veldwerk wordt uitgevoerd ontstaan er breuken in de waargenomen tijdreeksen.

Als oplossing voor deze drie problemen is een vijfdimensionaal structureel tijdreeksmodel ontwikkeld. Het rotatieschema van het panelontwerp impliceert dat iedere maand data worden verzameld in vijf onafhankelijke steekproeven: de steekproef die voor het eerst wordt waargenomen, de steekproef die drie maanden geleden is getrokken en voor de tweede keer wordt waargenomen, et cetera. Op basis hiervan kunnen iedere maand vijf onafhankelijke schattingen worden gemaakt voor de onbekende populatieparameter. De vijf tijdreeksen die op deze manier worden waargenomen vormen de input voor het tijdreeksmodel.

Het vijfdimensionale tijdreeksmodel bestaat uit vier componenten. De eerste component modelleert de onbekende populatieparameter als de som van een stochastische trend voor de laagfrequente variatie, een seizoen-



Figuur 1. Maandcijfers totale werkloze beroepsbevolking (in duizend) op nationaal niveau. De reeksen van de peilingen zijn beïnvloed door de breuken van 2010 en 2012. Voor de eerste peiling treden deze breuken op in januari 2010 en in april 2012. Voor de vervolgpeilingen is dat 3 tot 12 maanden later

component voor de cyclische variatie en een witte ruis voor de onverklaarde hoogfrequente variatie. Via deze component wordt steekproefinformatie uit het verleden gebruikt om de precisie van de schattingen te verbeteren. De tweede component modelleert de RGB. Onder de aanname dat de waarnemingen verkregen bij de eerste peiling onvertekend zijn, wordt het niveauverschil tussen de tijdreeksen van de tweede, derde, vierde en vijfde peiling ten opzichte van de eerste peiling gemodelleerd met vier verschillende random walk processen. Deze component zorgt ervoor dat de schattingen voor de populatieparameter op het niveau van de tijdreeks van de eerste peiling liggen. De derde component modelleert de breuken ten gevolge van veranderingen in het surveyproces. De vierde component modelleert de steekproefruis waarbij rekening wordt gehouden met de autocorrelatie die ontstaat door de paneloverlap. Zie Van den Brakel en Krieg (2015) voor een technische verantwoording van deze schattingsmethodiek.

Sinds 2010 wordt het hierboven beschreven tijdreeksmodel gebruikt voor het schatten van maandcijfers over de beroepsbevolking. Hierbij gaat het om de totale werkloze en werkzame beroepsbevolking op nationaal niveau en een uitsplitsing naar de kruising van geslacht en leeftijd in drie categorieën. Figuur 1 illustreert de vijf inputreeksen van het tijdreeksmodel en de outputreeks voor de totale werkloze beroepsbevolking op nationaal niveau. De outputreeks is gecorrigeerd voor de breuken in 2010 en 2012 ten gevolge van veranderingen in de dataverzameling. Vóór 2012 schat het model de aantallen zoals deze geweest zouden zijn met de nieuwe waarneemmethode. Dit model reduceert de standaardfouten met circa 20% op nationaal niveau en circa 50% op het niveau van de zes domeinen.

Om naast temporele informatie ook gebruik te maken van cross-sectionele informatie kunnen de vijfdimensionale tijdreeksen voor de zes domeinen worden gecombineerd in één 30-dimensionaal structureel tijdreeksmodel. Op een vergelijkbare manier kan het model worden uitgebreid met hulpreeksen, bijvoorbeeld het aantal geregistreerde uitkeringsgerechtigden. Hiermee kan de precisie van de modelschattingen verder worden verbeterd. Zie Van den Brakel en Krieg (2016).

In Boonstra en Van den Brakel (2019) wordt een alternatieve aanpak gevolgd. In plaats van een state-space model dat wordt gefit met een Kalman filter, wordt een structureel tijdreeksmodel voor de vijf peilingen in de twaalf provincies geschat met een Bayesiaans multilevel tijdreeksmodel via een MCMC-simulatie.

Kleindomeinschatters voor andere onderzoeken

Behalve bij de EBB worden kleindomeinschatters ook bij een aantal andere statistieken toegepast, of wordt daar op dit moment onderzoek naar gedaan.

Het Consumenten Conjunctuur Onderzoek (CCO) meet maandelijks het consumentenvertrouwen van de Nederlandse bevolking. Het consumentenvertrouwen wordt berekend op basis van vijf vragen over de financiële situatie van het eigen huishouden en de economische situatie van ons land. Naast het consumentenvertrouwen zelf worden ook tijdreeksen over deze vijf onderliggende vragen gepubliceerd. Sinds 2017 worden het consumentenvertrouwen en de onderliggende variabelen geschat met een structureel tijdreeksmodel. Dit model lijkt in grote lijnen op het model dat voor de maandcijfers over de beroepsbevolking toegepast wordt. Natuurlijk is het model aangepast om met de design-aspecten van het CCO rekening te houden.

Het Verplaatsingsonderzoek is een jaarlijks steekproefonderzoek waarin personen wordt gevraagd naar hun verplaatsingen op een bepaalde dag. Ook hier is behoefte aan gedetailleerde schattingen naar onder andere motief en vervoerwijze van de verplaatsing en naar persoonskenmerken zoals geslacht en leeftijd. Daarnaast worden trends over de tijd geschat. Hiervoor is recentelijk een multilevel tijdreeksmodel ontwikkeld waarmee zowel cross-sectionele als temporele verbanden worden gemodelleerd. Het model corrigeert bovendien voor breuken die zijn ontstaan door veranderingen in het surveyproces. Innovatief aan deze toepassing is dat ruim 500 tijdreeksen in één Bayesiaans multilevel tijdreeksmodel worden gecombineerd. Deze methodiek is in 2019 in gebruik genomen.

Met het Schoolverlatersonderzoek (SVO) worden verschillende aspecten van de aansluiting van werk en opleiding gemeten. Hierover worden gedetailleerde cijfers naar onderwijsinstelling of arbeidsmarktregio en opleidingsrichting gemaakt. Het SVO wordt de afgelopen jaren onder alle mbo-schoolverlaters uitgevoerd. Omdat slechts een deel van hen uiteindelijk respondeert, moeten

de ontbrekende gegevens worden bijgeschat. Hiervoor wordt, net zoals voor de gemeentelijke jaarcijfers over de beroepsbevolking, een multilevel model op persoonsniveau toegepast. In dit geval bevat het model naast een uitgebreide set van registervariabelen meerdere gemodelleerde effecten voor indelingen naar school, regio en opleiding.

Het toepassen van kleindomeinschatters bij bedrijfsstatistieken is vaak complexer. Hierbij moet rekening gehouden worden met grote schaalverschillen tussen kleine en grote bedrijven. In dit geval is de doelvariabele scheef verdeeld. Daarnaast kan de doelvariabele in sommige surveys vaak 0 zijn. Dit is het geval bij de investeringsen R&D-statistieken. In de afgelopen jaren is onderzocht hoe modelgebaseerde domeinschatters gebruikt kunnen worden voor het regionaliseren van deze bedrijfsstatistieken. Hierbij worden schattingen gemaakt per COROP-gebied en gemeente. De resultaten zijn veelbelovend en op dit moment onderzoeken we samen met de betrokken afdelingen de mogelijkheden om deze schatters in productie te nemen.

Toekomstige ontwikkelingen

Zoals aangegeven in de inleiding, neemt de relevantie van statistische informatie toe met de mate van detail, frequentie en snelheid waarmee cijfers beschikbaar komen. Modelgebaseerde domeinschatters bieden de mogelijkheid om meer relevante statistische output te produceren zonder de dataverzamelingskosten extravagant te laten stijgen. Het ligt dan ook voor de hand dat het CBS door zal gaan met het verder ontwikkelen van dergelijke schattingsmethodieken.

Deze schattingsmethodieken bieden ook de mogelijkheid om efficiënt gebruik te maken van gerelateerde informatie uit zogenaamde bigdatabronnen. Een belangrijk aspect van dit soort nieuwe databronnen is dat ze vaak sneller en met een veel hogere frequentie beschikbaar komen dan herhaald waargenomen steekproefonderzoeken. Google trends kunnen bijvoorbeeld op dag- of weekbasis worden afgeleid. Een potentiële toekomstige toepassing is om reeksen gebaseerd op surveys te combineren met snel beschikbare hulpreeksen uit alternatieve databronnen. Hierdoor ontstaat de mogelijkheid om in real time nauwkeurige voorlopige schattingen te maken voor de doelvariabelen van het steekproefonderzoek. Eerste resultaten daartoe zijn beschreven in Van den Brakel et al. (2017) voor het consumentenvertrouwen en Schiavoni et al. (2019) voor de maandcijfers van de beroepsbevolking.

Literatuur

Bailar, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23–30.

Boonstra, H. J., & Brakel, J. A. van den. (in press). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*.

Brakel, J. A. van den, & Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, 267–296.

Brakel, J. A. van den & S. Krieg (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statistical Society Series A*, 179, 763–791.

Brakel, J. A. van den, Söhler, E., Daas P., & Buelens B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, 43, 183–210.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation*. Wiley, New York.

Schiavoni, C., Palm, F., Smeekes, S., & Brakel, J. A. van den. (2019). *A dynamic factor model approach to incorporate Big Data in state space models for official statistics*. Discussion paper January, 2019. Heerlen: Statistics Netherlands.

Jan van den Brakel heeft biometrie gestudeerd aan de Universiteit van Wageningen en werkt momenteel als methodoloog bij het CBS en als bijzonder hoogleraar Survey Methodology bij de vakgroep Kwantitatieve Economie van de Universiteit van Maastricht. Zijn belangrijkste onderzoeksgebieden zijn steekproeftheorie, ontwerp en analyse van proeven, tijdreeksanalyse, kleindomeinschatters en het gebruik van big data in de productie van officiële statistieken.
Email: jbrl@cbs.nl

Harm Jan Boonstra heeft natuurkunde gestudeerd aan de Rijksuniversiteit Groningen en heeft daar ook zijn promotieonderzoek gedaan. Hij werkt sinds 1999 als onderzoeker bij het Centraal Bureau voor de Statistiek en heeft zicht gespecialiseerd in Bayesiaanse schattingsmethoden toegepast op de officiële statistiek.
E-mail: hbta@cbs.nl

Sabine Krieg heeft wiskunde gestudeerd in Jena (Duitsland). Nadat ze een aantal jaren onderzoek gedaan heeft op wiskundig gebied op de universiteiten van Essen (Duitsland) en Groningen, werkt ze sinds 1997 als methodoloog op het CBS. Ze is gespecialiseerd in steekproeftheorie, modelgebaseerd schatten en seizoenscorrectie.
E-mail: skrg@cbs.nl

Marc Smeets heeft wiskunde gestudeerd aan de Technische Universiteit Eindhoven en heeft daar ook promotieonderzoek gedaan. Hij werkt sinds 2001 als methodoloog bij het Centraal Bureau voor de Statistiek en is gespecialiseerd in modelgebaseerde schattingsmethoden, steekproeftheorie en steekproefcoördinatie.
E-mail: mset@cbs.nl