Quinten A. Meertens
Centraal Bureau voor de Statistiek
Den Haag
qa.meertens@cbs.nl

Onderzoek

Vertekening door algoritmes die fouten maken

Wat als je algoritmes, die soms fouten maken, op een hele populatie loslaat en de uitkomsten optelt? Kun je daarmee betrouwbare statistische uitspraken doen over de populatie? Quinten Meertens deed vanuit de Universiteit van Amsterdam (econometrie) en de Universiteit Leiden (informatica) promotieonderzoek bij het CBS naar deze vragen en beschrijft hier de problemen en oplossingen die hij heeft ontdekt.

Binnen de officiële statistiek is men geïnteresseerd in aantallen en hoeveelheden. Actuele voorbeelden op wereldwijde schaal zijn het aantal coronabesmettingen, de hoeveelheid bos in de Amazone en de oppervlakte ijs op de poolkappen. Op nationale schaal kun je denken aan het aantal huizen met zonnepanelen, het aandeel cybercrime binnen alle misdrijven en het aantal innovatieve bedrijven in Nederland. De gemene deler van deze zes voorbeelden is dat ze tot stand komen door een geheel (een populatie mensen, huizen, bedrijven, et cetera) in groepen te verdelen en dan de omvang van de afzonderlijke groepen te bepalen.

Het toekennen van een object aan een groep noemen we ook wel classificeren. Dat classificeren gebeurt steeds vaker met behulp van algoritmes, daarom ook wel classificatiealgoritmes genoemd. Algoritmes zijn snel, maar niet foutloos. Fouten die algoritmes maken kunnen grote gevolgen hebben op het niveau van individuen. De toeslagenaffaire is daarvan een tragisch voorbeeld. Wij zijn echter benieuwd naar de gevolgen van dezelfde fouten op het niveau van groepen. Wat gebeurt er als je de uitkomsten van het algoritme dat is toegepast op een hele populatie bij elkaar optelt? Het blijkt dat er dan vrijwel altijd statistische vertekening optreedt, ook als het algoritme relatief weinig fouten maakt. Die vertekening wordt *misclassification bias* genoemd.

Een actueel voorbeeld

We introduceren het principe van misclassification bias eerst aan de hand van een actueel voorbeeld: het aantal coronabesmettingen in Nederland. Zoals bekend kun je je bij klachten laten testen op het coronavirus. Een PCR-test geeft dan aan of je al dan niet met het coronavirus bent besmet. Alle positieve PCR-tests in Nederland worden geteld en zo komen we tot het aantal coronabesmettingen in Nederland.

Het is echter bekend dat de PCR-test niet altijd juist is. Om allerlei redenen kan iemand die wel COVID-19 heeft toch een negatieve testuitslag ontvangen. We noemen dat een fout-negatieve testuitslag. Volgens het RIVM is de kans op dit type fout tussen de 2 en de 33 procent, zie [10]. Andersom kan een fout-positieve testuitslag ook voorkomen: de kans dat je zonder COVID-19 te hebben toch een positieve testuitslag ontvangt is tussen de $0{,}5$ en 4 procent, zie wederom [10].

Met deze foutkansen kunnen we een simpel rekenvoorbeeld uitwerken. Stel dat er in een week $100\,000$ mensen met corona-gerelateerde klachten zich laten testen en stel dat hiervan $10\,000$ mensen ook daadwerkelijk COVID-19 hebben. Wat is dan (naar verwachting) het aantal coronabesmettingen op basis van de uitslagen van de afgenomen PCR-tests die week?

Het antwoord hangt af van wat de precieze kans op fout-positieven en fout-negatieven is. Laten we voor dit voorbeeld eens kijken wat er in het meest extreme geval

gebeurt. Stel dat de kans op een fout-negatieve uitslag gelijk is aan de genoemde 2 procent en de kans op een fout-positieve uitslag gelijk is aan 4 procent. Dan zullen van de $10\,000$ met COVID-19 besmette mensen naar verwachting 9800 een positieve testuitslag ontvangen. Van de $90\,000$ mensen zonder COVID-19 zullen naar verwachting 3600 een positieve testuitslag ontvangen. Het totaal aantal besmettingen komt dan uit op $13\,400$, naar verwachting. Dat ligt 34 procent hoger dan de werkelijkheid. De misclassification bias is dan $+0{,}034$.

Merk op dat we met andere combinaties uit de twee genoemde intervallen van foutkansen ook heel andere uitkomsten kunnen aantreffen. Als we uitgaan van het andere extremum, 33 procent kans op fout-negatieven en $0{,}5$ procent kans op fout-positieven, dan vinden we 7150 positieve tests. De misclassification bias is nu $-0{,}0295$. Een mooie laatste combinatie is 9 procent kans op fout-negatieven en 1 procent kans op fout-positieven, wat ook binnen de genoemde intervallen past: dan komen we precies op $10\,000$ uit. De misclassification bias is dan 0.

Formele definitie

Voordat we oplossingen beschrijven om misclassification bias te verkleinen, introduceren we eerst de formele definitie van misclassification bias. Daartoe is de volgende notatie benodigd. We beschouwen een populatie I van omvang N. Voor nu beperken we ons tot binaire classificatie. Dat wil zeggen dat elk object $i \in I$ tot een van twee klassen behoort. We noteren de klasse van een object $i \in I$ met $s_i \in \{0,1\}$ (s staat voor stratum). We zijn geïnteresseerd in de zogenaamde *base rate* α, die gelijk is aan

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} s_i.$$

In het voorbeeld over coronabesmettingen geldt dat $N = 100\,000$, dat $s_i = 1$ dan en slechts dan als persoon i besmet is met COVID-19 en dat $\alpha = 0{,}1$.

Zoals in het voorbeeld is de waarde van s_i niet bekend. We kunnen alleen een schatting of, in het geval van classificatiealgoritmen, een voorspelling maken van de waarde van s_i. Die schatting noteren we met $\hat{s}_i$. De geschatte base rate die we verkrijgen door de schattingen $\hat{s}_i$ te middelen noteren we met $\hat{\alpha}$. We beschouwen $\hat{s}_i$ als een stochastische variabele waarvan de verdeling afhangt van de werkelijke waarde s_i. De *(mis)classificatiekansen* p_{ab}, met $a,b \in \{0,1\}$, zijn gedefinieerd als

$$p_{ab} = \mathbb{P}(\hat{s}_i = b \mid s_i = a).$$

De kans op een fout-negatieve uitkomst wordt dus met p_{10} genoteerd. De kans op een fout-positieve uitkomst wordt met p_{01} genoteerd. De matrix $P = (p_{ab})$ noemen we de *misclassificatiematrix*. In deze matrix staat p_{11} linksboven, dus P is gelijk aan

$$P = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix}.$$

Merk op dat waarden in elke rij van P optellen tot 1, waarmee de getransponeerde matrix P^T een stochastische matrix is.

Tot slot introduceren we nog de *base rate vector* $\boldsymbol{\alpha} = (\alpha, 1-\alpha)^T \in \mathbb{R}^2$. De vector $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}, 1-\hat{\alpha})^T$ is de geschatte base rate vector. Nu geldt de gelijkheid

$$\mathbb{E}[\hat{\boldsymbol{\alpha}}] = P^T\boldsymbol{\alpha}.$$

Met dit resultaat en de bovenstaande notatie kunnen we misclassification bias formeel definiëren en uitdrukken.

Definitie. *Misclassification bias* is de vertekening van $\hat{\alpha}$ als schatter voor de base rate α en is gelijk aan

$$(p_{11} - 1)\,\alpha + (1 - p_{00})\,(1 - \alpha).$$

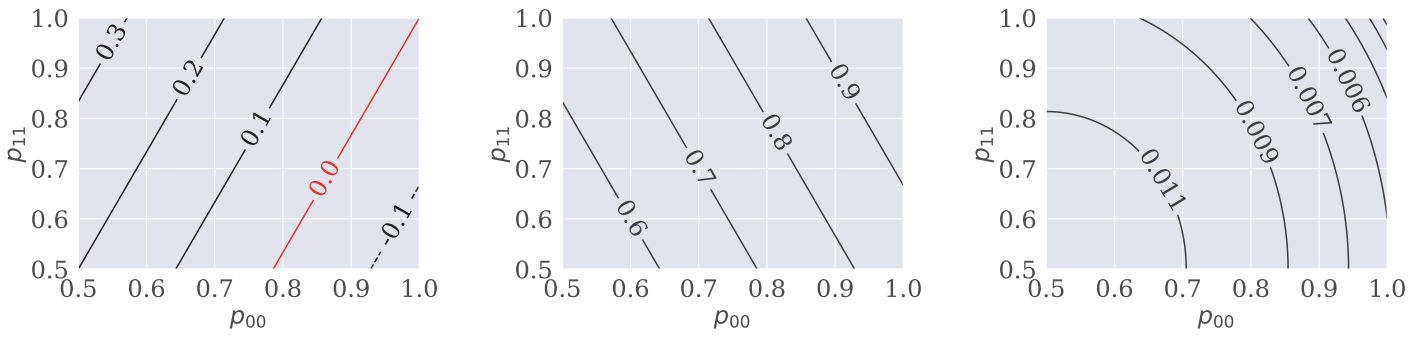
Het is nu eenvoudig in te zien dat de misclassification bias gelijk is aan 0 dan en slechts dan als (p_{00}, p_{11}) op de lijn tussen $(1-\alpha, \alpha)$ en $(1,1)$ ligt, zie de rode lijn in paneel (a) van Figuur 1.

Oplossingsrichtingen

Een classificatiealgoritme is, nadat de parameters zijn vastgesteld op basis van een zogenaamde 'training dataset', niets meer dan een deterministische functie van de kenmerkenruimte (feature space) naar, in ons geval, de verzameling $\{0,1\}$. We kunnen nu kiezen uit twee modellen om de fouten van dit algoritme te beschrijven.

In het eerste model beschouwen we de werkelijke klasse s_i als stochastisch, gegeven de uitkomst $\hat{s}_i$ van het algoritme. Dat wordt ook wel het *Berkson foutmodel* genoemd, geïntroduceerd in [1]. Dit is in de econometrie ook gebruikelijk: de onafhankelijke variabele wordt geschreven als een uitdrukking in de afhankelijke variabele plus een ruisterm. We zijn dan niet benieuwd naar de misclassificatiekansen p_{ab} van hierboven, maar naar de *calibratiekansen* c_{ab}, die gedefinieerd zijn als $c_{ab} = \mathbb{P}(s_i = b \mid \hat{s}_i = a)$. Samen vormen deze kansen de calibratiematrix $C = (c_{ab})$, met c_{11} linksboven.

In het tweede model beschouwen we de voorspelde klasse $\hat{s}_i$ gegeven de echte klasse s_i als stochastisch. Het idee daarachter komt uit de epidemiologie: de symptomen (als afhankelijke variabelen) worden bepaald door het al dan niet heb-



(a) Misclassification bias (b) Classification accuracy (c) Standaarddeviatie ($N = 2000$)

Figuur 1 Deze contourlijnen laten het verschil zien tussen (a) misclassification bias en (b) classification accuracy (dat wil zeggen: het percentage correct geclassificeerde objecten) voor base rate $\alpha = 0{,}3$ en variabele classificatiekansen p_{00} en p_{11}. Op de rode lijn in paneel (a) is de misclassification bias precies 0. De essentie van deze figuur is dat de richtingen van de contourlijnen in (a) en (b) tegengesteld zijn (en zelfs loodrecht op elkaar staan als $\alpha = 0{,}5$). Paneel (c) toont de standaarddeviatie van de schatter $\hat{\alpha}$ bij populatiegrootte $N = 2000$. Panelen (a) en (c) zijn overgenomen van [8].

ben van de ziekte (de binaire onafhankelijke variabelen), niet andersom. Dit tweede model wordt het *klassieke meetfoutmodel* genoemd, zie [2, p. 6].

Wij gaan uit van het laatste foutmodel, omdat het in de toepassingen waar we interesse in hebben het meest intuïtief is. In aanvulling op dat model nemen we aan dat de misclassificatiekansen voor elk object $i \in I$ gelijk zijn, gegeven de werkelijke klasse s_i van dat object. Bovendien nemen we aan dat de uitkomsten van het algoritme tussen de verschillende objecten stochastisch onafhankelijk van elkaar zijn.

Een gebruikelijke doelstelling binnen statistical learning is om classificatiealgoritmes te verbeteren. De kwaliteit van een algoritme wordt dan gemeten met bepaalde kwaliteitsmaten, zoals *precision*, *recall*, *accuracy* en vele andere. Minder fouten, oftewel het verhogen van p_{11} of p_{00} (of beide) zorgt ervoor dat elk van de gebruikelijke maten het algoritme beter beoordeelt. Het is dan ook tegenintuïtief dat dit niet automatisch betekent dat de misclassification bias kleiner wordt. Figuur 1 laat dit mooi zien. Stel bijvoorbeeld dat we met een classificatiealgoritme starten dat fouten maakt, maar α wel zuiver schat. Dat wil zeggen, we starten op de rode lijn in paneel (a). Als we nu het algoritme op zo'n manier verbeteren dat p_{00} gelijk blijft en p_{11} toeneemt, dan stijgt de accuracy (zie paneel (b)) en daalt de standaardafwijking (zie paneel (c)), maar *stijgt* de misclassification bias (zie daarvoor weer paneel (a)).

De uitdrukking voor misclassification bias zoals hierboven opgenomen kan gebruikt worden om aan te tonen dat een classificatiealgoritme de base rate α zuiver schat dan en slechts dan als precision en recall gelijk zijn. Het verwachte aantal fout-positieven is dan namelijk gelijk aan het verwachte aantal fout-negatieven. We zouden deze restrictie dus kunnen meenemen in het trainen van het algoritme, bijvoorbeeld door een maat te bedenken die naast het belonen van een hogere classification accuracy ook het verschil tussen precision en recall bestraft.

De oplossingsrichting die in mijn proefschrift *Misclassification Bias in Statistical Learning* wordt onderzocht is een heel andere, namelijk die uit de epidemiologie. Het idee is om het algoritme niet verder aan te passen, maar om na afloop voor vertekening te corrigeren. Simpel gezegd: we schatten α door $(P^T)^{-1}\hat{\alpha}$.

Dat lijkt een prachtige oplossing voor het probleem, maar er zit een addertje onder het gras. De misclassificatiematrix P is namelijk niet bekend en moet ook geschat worden. Helaas is de inverse van een geschatte matrix geen zuivere schatter van de inverse matrix. Bovendien kan de variantie zeer groot worden als de schatting van P op een klein aantal datapunten is gebaseerd. Om deze twee redenen bestuderen we alternatieve methoden uit de literatuur die na afloop voor vertekening corrigeren. We vergelijken de mean squared error (MSE) van de schatters die op deze correctiemethoden zijn gebaseerd en kunnen in verschillende situaties theoretisch bewijzen welke correctiemethode het beste is.

Correctiemethoden

De correctiemethoden zijn allemaal gebaseerd op schattingen van het aantal fouten dat het classificatiealgoritme maakt. We schatten dat aantal fouten met behulp van een zogenaamde testset, van grootte n. Van de objecten in deze (kleine) aselecte steekproef van de populatie achterhalen we de werkelijke klasse s_i, vaak door experts met domeinkennis in te zetten. We kunnen het classificatiealgoritme ook op deze testset loslaten om zo s_i en $\hat{s}_i$ te vergelijken. We komen zo tot tellingen n_{ab} ($a,b \in \{0,1\}$), het aantal objecten in de testset waarvoor geldt dat $s_i = a$ en $\hat{s}_i = b$. Zo komen we tot schattingen voor p_{ab}, namelijk $\hat{p}_{ab} = n_{ab}/(n_{a0} + n_{a1})$.

We bespreken hier twee correctiemethoden. De eerste methode is gebaseerd op het inverteren van de geschatte misclassificatiematrix. We noemen de verkregen schatter voor α de *misclassificatieschatter* en noteren deze schatter als $\hat{\alpha}_p$. Er geldt dat

$$\hat{\alpha}_p = \frac{\hat{\alpha} + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1}.$$

De tweede methode is gebaseerd op het idee van de calibratiekansen c_{ab}, die we ook met behulp van de testset schatten, namelijk als $\hat{c}_{ab} = n_{ab}/(n_{0b} + n_{1b})$. We nemen het eerste element van de 2-vector $\hat{C}\hat{\alpha}$ als schatter voor α, oftewel

$$\hat{\alpha}_c = \hat{c}_{11}\alpha + \hat{c}_{10}(1-\alpha).$$

In [5] worden uitdrukkingen afgeleid voor de vertekening en variantie van de misclassificatieschatter en voor de calibratieschatter. Deze uitdrukkingen zijn benaderingen, nauwkeurig tot op orde $1/n^2$.

De winnaar

Het verschil tussen de MSE van de misclassificatieschatter en die van de calibratieschatter kan, tot op orde $1/n$ nauwkeurig, uitgedrukt worden als

$$\mathrm{MSE}[\hat{\alpha}_p] - \mathrm{MSE}[\hat{\alpha}_c] = \frac{T^2}{n(p_{00}+p_{11}-1)^2\beta(1-\beta)},$$

waarbij

$$T = (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$$

en

$$\beta = (1-\alpha)(1-p_{00}) + \alpha p_{11}.$$

Aangezien de bovenstaande uitdrukking altijd strikt positief is (want alleen gelijk aan 0 als $p_{00} = p_{11} = 1$, dat wil zeggen, als het algoritme nooit fouten maakt), komt de calibratieschatter $\hat{\alpha}_c$ altijd als winnaar uit de bus.

Een relevante uitzondering is wanneer een van de cruciale aannames wordt geschonden, namelijk de aanname dat de testset een aselecte steekproef is van de populatie. Vaak wordt een classificatiealgoritme namelijk eenmaal getraind en worden de foutkansen geschat op een aselecte steekproef als testset zoals eerder beschreven. Vervolgens wordt het algoritme voor een langere periode gebruikt zonder op een later tijdstip opnieuw de kwaliteit van het algoritme te bekijken. Wat hier essentieel is, is dat de base rate α over de tijd kan veranderen. Maar dan kan de eerder gebruikte testset niet meer als aselecte steekproef gezien worden van de populatie op dit latere tijdstip. En juist dit scenario is waarom het nauwkeurig schatten van α relevant is, denk wederom aan het voorbeeld over het aantal coronabesmettingen in Nederland. Het is interessant om dat aantal te schatten, juist omdat het over de tijd verandert.

Als we te maken hebben met een veranderende base rate α, zonder dat de foutkansen van het algoritme veranderen, dan spreken we van *prior probability shift*. Deze shift noteren we met δ. De uitdrukkingen van de MSE van $\hat{\alpha}_p$ en $\hat{\alpha}_c$ gaan dan afhangen van δ. In [7] laten we voor elk algoritme zien bij welke waarde van δ de eerdere conclusie verandert: voor δ voldoende groot zal de misclassificatieschatter een kleinere MSE hebben dan de calibratieschatter.

Een nog mooiere oplossing is om te allen tijde de twee correctiemethoden te combineren. Het idee is om op tijdstip $t = 0$

gebruik te maken van $\hat{\alpha}_c = \hat{\alpha}_c(0)$ en om op tijdstip $t > 0$ hier het verschil $\hat{\alpha}_p(t) - \hat{\alpha}_p(0)$ bij op te tellen. De eerste simulaties laten zien dat deze correctiemethode een kleinere MSE heeft dan zowel $\hat{\alpha}_c(t)$ als $\hat{\alpha}_p(t)$, voor alle $t > 0$ en elke waarde van δ, zie [4].

Toegepast op webwinkels

Binnen het CBS hebben we de misclassificatieschatter toegepast om een schatting te maken van de aankopen van Nederlandse consumenten bij buitenlandse webwinkels die binnen de EU zijn gevestigd, zie [6]. Dergelijke winkels zijn verplicht om in Nederland belasting af te dragen. Via de Belastingdienst zijn de aangiftes van dergelijke winkels beschikbaar binnen het CBS. (Het CBS hanteert strenge eisen omtrent privacy en statistische beveiliging opdat gepubliceerde cijfers nooit herleid kunnen worden tot individuele personen of bedrijven, zie ook [9].) Het probleem is dat veel andere buitenlandse bedrijven ook belastingaangifte doen en dat het onderscheid tussen webwinkels en andere bedrijven op basis van de beschikbare gegevens niet te maken is.

We hebben daarom een algoritme ontwikkeld dat automatisch de website van een bedrijf opzoekt met de naam van een bedrijf als uitgangspunt en vervolgens bepaalt of de gevonden website een webwinkel is. Voor een kleine aselecte steekproef hebben experts bepaald of een bedrijf al dan niet een webwinkel is. Dat is met name voor buitenlandse bedrijven met weinig omzet soms een behoorlijke klus. Met die informatie maken we een schatting van de misclassificatiekansen van het algoritme.

Vervolgens nemen we aan dat de misclassificatiekans onafhankelijk is van de hoogte van de omzet van een bedrijf. De schatting van de aankopen van Nederlandse consumenten bij buitenlandse webwinkels binnen de EU berekenen we dan door de geschatte 2-vector van omzetten op dezelfde manier te corrigeren voor misclassification bias als dat voor de geschatte base rate vector is gedaan.

Het resultaat van deze methodologie is, voor het jaar 2016, een uitkomst (met een geschatte standaardafwijking van 8 procent) die zes keer zo hoog is als eerdere schattingen op basis van consumentenenquêtes. De reden voor dit grote verschil is dat consumenten zich vaak niet realiseren dat ze iets bij een buitenlandse webwinkel bestellen, bijvoorbeeld omdat de website in het Nederlands is. Het grote effect van deze zogenaamde language bias kon nu met onze resultaten voor het eerst goed aangetoond worden.

Deze toepassing laat zien dat classificatiealgoritmes die niet perfect zijn toch goed binnen de officiële statistiek gebruikt kunnen worden, met uitstekende resultaten tot gevolg.

Vervolgonderzoek

Het gebruik van algoritmes om aantallen en hoeveelheden te schatten is een deelgebied van statistical learning dat *quantification learning* wordt genoemd. Er is weinig (maar toenemende) aandacht voor dit deelgebied dat verrassend veel toepassingen heeft in tal van uiteenlopende vakgebieden, van epidemiologie tot actuariaat en van remote sensing tot mariene biologie, zie [3]. En dus binnen de officiële statistiek. Binnen quantification learning worden drie typen correctiemethoden onderscheiden. Het eerste type, eerst tellen en dan corrigeren, hebben we hier besproken. Mijn proefschrift bevat het eerste theoretische bewijs over welke methode (binnen dit type) in welke situatie het beste is. Correctiemethoden van het tweede type maken gebruik van classificatiekansen, dat wil zeggen de continue uitkomsten van een algoritme die binnen het interval $[0,1]$ liggen. Methoden van het derde type slaan het classificeren helemaal over. Bij dit type wordt α direct geschat door de verdeling van de data binnen de kenmerkenruimte (feature space) tussen de training set en de rest van de populatie te vergelijken. Voor de methoden van het tweede en derde type zijn tot dusver al wat empirische resultaten bekend, maar nog geen theoretische resultaten. Daar zijn de komende jaren nog mooie stappen in te zetten.

Referenties

1 J. Berkson, Are there two regressions?, *Journal of the American Statistical Association,* 45(250) (1950), 164–180.

2 J.P. Buonaccorsi, *Measurement Error: Models, Methods, and Applications*, Chapman & Hall/CRC, 2010.

3 P. Gonzalez, A. Castaño, N.V. Chawla en J.J. del Coz, A review on quantification learning, *ACM Computing Surveys* 50(5) (2017), 74:1–74:40.

4 K. Kloos, A new generic method to improve machine learning applications in official statistics, *Statistical Journal of the IAOS* (2021), in pre-press, DOI: 10.3233/SJI-210885.

5 K. Kloos, Q.A. Meertens, S. Scholtus en J.D. Karch, Comparing correction methods to reduce misclassification bias, in L. Cao, W.A. Kosters en J. Lijffijt (eds.), *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC)*, Leiden, 2020, pp. 103–129.

6 Q.A. Meertens, C.G.H. Diks, H.J. van den Herik en F.W. Takes, A datadriven supply-side approach for estimating cross-border Internet purchases Within the European Union, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(1) (2020), 61–90.

7 Q.A. Meertens, C.G.H. Diks, H.J. van den Herik en F.W. Takes, Improving the output quality of official statistics based on machine learning algorithms, *Journal of Official Statistics* (2021), geaccepteerd voor publicatie.

8 S. Scholtus en A. van Delden, The accuracy of estimators based on a binary classier, Discussion Paper Nr. 202007, CBS, 2020.

9 https://www.cbs.nl/nl-nl/over-ons/organisatie/privacy.

10 https://www.rivm.nl/sites/default/files/2020-12/Toelichting%20betrouwbaarheid%20PCR.pdf.