Toelichting berekening kwartaalcijfers Gezondheidsenquête 2020

Marc Smeets en Jan van den Brakel

Reviewer: Rob Willems

21 augustus 2020

samenvatting In deze nota wordt een technische toelichting gegeven op de tijdreeksmethode waarmee de kwartaalcijfers van de Gezondheidsenquête over 2020 geschat zijn. Vanwege het wegvallen van de CAPI-waarneming tijdens de coronacrisis ontstaat er een breuk in de uitkomsten van de gezondheidsenquête. Via een tijdreeksmodel wordt er zo goed mogelijk voor deze breuk gecorrigeerd. Door kwartaalcijfers te berekenen, wordt een zo duidelijk mogelijk beeld gegeven van (ervaren) gezondheid, leefstijl en zorggebruik in het coronajaar 2020.

trefwoorden Coronacrisis, structureel tijdreeksmodel, nowcasten, gezondheid.

1. Inleiding

De Gezondheidsenquête (GE) is een doorlopend steekproef onderzoek dat door het CBS wordt uitgevoerd met als doel de gezondheid, de leefstijl en het zorggebruik van de Nederlandse bevolking te beschrijven. Doordat de GE herhaaldelijk wordt uitgevoerd ontstaan tijdreeksen die ontwikkeling van de variabelen met betrekking tot gezondheid, leefstijl en zorggebruik door de tijd heen zo goed mogelijk beschrijven. De dataverzameling is gebaseerd op een sequentieel mixed-mode design waarbij gebruik wordt gemaakt van internet waarneming en face-tot-face waarneming. Tot en met 2019 zijn op basis van het onderzoek alleen jaarcijfers over de gezondheid, de leefstijl en het zorggebruik van de Nederlandse bevolking gepubliceerd.

Door het tijdelijk wegvallen van de face-to-face waarneming (CAPI) als gevolg van de coronamaatregelen in 2020 ontstaat er een breuk in de uitkomsten van de Gezondheidsenquête (GE). Via een tijdreeksmethode wordt er zo goed mogelijk gecorrigeerd voor de verandering in meet- en selectiefouten t.g.v. het wegvallen van CAPI. Daarnaast wordt deze methode gebruikt om de verbanden tussen de

COVID19-pandemie en de uitkomsten van de GE zo goed mogelijk in kaart te brengen. Tot 2020 publiceerde de Gezondheidsenquête alleen jaarcijfers. Om het effect van de corona crisis op de ontwikkeling van de gezondheid, het zorggebruik en de leefstijl van de Nederlandse bevolking zo goed mogelijk in kaart te brengen, is besloten om kwartaalcijfers te publiceren. Dit heeft twee voordelen. Ten eerste kan de periode van de coronacrisis scherper worden afgebakend. Daarnaast ontstaan al gedurende 2020 snellere cijfers over de mogelijke effecten van de corona crisis op de uitkomsten van de GE. In deze nota wordt de gebruikte tijdreeksmethode beschreven.

Het netto-effect van het wegvallen van CAPI kan worden berekend op basis van responsbestanden van de voorgaande jaren. Dit kan door de CAPI-respondenten uit de respons te verwijderen en de resterende responsen opnieuw te wegen. Daarmee worden twee schattingen voor een doelvariabele verkregen: één gebaseerd op de volledige respons (CAPI plus CAWI) en één gebaseerd op alleen de internetrespons (CAWI). Voor de GE kunnen op deze manier tijdreeksen gemaakt worden vanaf het eerste kwartaal in 2014: de reguliere reeks gebaseerd op de volledige respons en de internetreeks gebaseerd op de internetrespons. De reguliere reeks is beschikbaar t/m het eerste kwartaal van 2020 en de internetreeks t/m het tweede kwartaal van 2020. In het eerste kwartaal van 2020 is CAPI slechts deels beschikbaar vanwege het wegvallen van de CAPI-waarneming vanaf halverwege maart. Voor dit kwartaal nemen we aan dat het om een normale waarneming met CAPI gaat. Op dit moment is nog niet duidelijk wanneer de CAPI-waarneming weer zal plaatsvinden. Voor het schatten van het derde en het vierde kwartaal van 2020 moet daarom nog besloten worden hoe we omgaan met het ontbreken van CAPI in deze kwartalen. Met name indien in de tweede of derde maand van een kwartaal de CAPI waarneming wordt opgestart moet een besluit worden genomen hoe hiermee in de schattingsmethodiek wordt omgegaan.

In de kwartalen waar alleen internetrespons beschikbaar is, kan het kwartaalcijfer gebaseerd op de internetrespons gecorrigeerd worden voor het systematisch verschil (kortweg de breuk) t.g.v. het wegvallen van CAPI. De breuk kan worden geschat uit de voorgaande perioden waarvoor zowel de reguliere reeks als de internetreeks beschikbaar is. De gebruikte methode houdt tevens rekening met de duur van de periode zonder CAPI. Dit kan via een bivariaat structureel tijdreeksmodel waarbij het gecorrigeerde cijfer wordt genowcast op basis van de internetreeks die beschikbaar is in de kwartalen waar CAPI ontbreekt.

Voor de Gezondheidsenquête is er dus respons beschikbaar vanaf het eerste kwartaal van 2014. De respons in de eerste kwartalen van de reeks is nodig om de onderliggende parameters van het tijdreeksmodel goed te kunnen schatten. Dit is het zogenaamde inregelen van het model. Om deze reden worden de kwartaalcijfers vanaf 2017 gepubliceerd. Er worden kwartaalcijfers gemaakt voor de volgende variabelen:

- Ervaren gezondheid (0 jaar en ouder)
- Psychisch ongezond (12 jaar en ouder)
- Huisartsbezoek in de afgelopen vier weken (0 jaar en ouder)
- Dagelijks roken (vanaf 18 jaar)
- Overgewicht (vanaf 18 jaar)
- Overmatig alcoholgebruik (vanaf 18 jaar)
- Tandartsbezoek in de afgelopen vier weken (0 jaar en ouder)
- Specialistbezoek in de afgelopen vier weken (0 jaar en ouder)

2. De tijdreeksmethode

2.1 Het model

We gaan uit van het bivariate structurele tijdreeksmodel gegeven door

$$\begin{pmatrix} \hat{y}_t^{\mathrm{R}} \\ \hat{y}_t^{\mathrm{I}} \end{pmatrix} = \begin{pmatrix} \vartheta_t \\ \vartheta_t \end{pmatrix} + \begin{pmatrix} \\ \lambda_t \end{pmatrix} + \begin{pmatrix} e_t^{\mathrm{R}} \\ e_t^{\mathrm{I}} \end{pmatrix}, \text{ met} \tag{1}$$

$\vartheta_t = L_t + S_t$ en kwartaal t.

Hierin is

- $\hat{y}_t^{\mathrm{R}}$: de directe schatting in kwartaal t gebaseerd op de volledige respons (*reguliere reeks*).
- $\hat{y}_t^{\mathrm{I}}$: de directe schatting in kwartaal t gebaseerd op de internetrespons (*internetreeks*).
- ϑ_t: de onbekende populatieparameter in kwartaal t.
- L_t: de trend, gemodelleerd als smooth trend model.
- S_t: de seizoenscomponent met kwartaal als periode, gemodelleerd met een trigonometrisch seizoen model.
- λ_t: het systematisch verschil tussen de internetreeks en de reguliere reeks, gemodelleerd als random walk.
- e_t^j: de meetfout van $\hat{y}_t^j$ voor $j \in \{\mathrm{R}, \mathrm{I}\}$, gemodelleerd als $e_t^j = \sqrt{\hat{\mathbb{V}}(\hat{y}_t^j)}\tilde{e}_t^j$ waarbij $\hat{\mathbb{V}}(\hat{y}_t^j)$ de geschatte variantie van $\hat{y}_t^j$ is en $\tilde{e}_t^j$ witte ruis.

De volledige uitwerking van dit model staat in Bijlage I.

2.2 De directe schattingen

De directe schattingen $\hat{y}_t^{\mathrm{R}}$ en $\hat{y}_t^{\mathrm{I}}$ zijn beschikbaar vanaf het eerste kwartaal van 2014. Voor de jaren 2014 t/m 2019 zijn deze schattingen gebaseerd op de gewogen jaarbestanden van de gezondheidsenquête. De weging zoals die vanaf 2014 plaatsvindt, inclusief een update met de veranderingen per 2018 is beschreven in Boonstra (2019). Voor 2020 is er een gewogen halfjaarsbestand samengesteld door de beschikbare respons in de maanden januari t/m juni via de gegeneraliseerde regressieschatter (GREG, Särndal e.a. 1992) op te hogen naar de populatie van 2020. Hierbij is uitgegaan van het weegmodel uit eerdere jaren, waarbij geen rekening gehouden wordt met de in 2018 ingevoerde doelgroepenbenadering.

De directe schatting $\hat{y}_t^{\mathrm{R}}$ is berekend als gewogen gemiddelde van de volledige respons in kwartaal t en de directe schatting $\hat{y}_t^{\mathrm{I}}$ als gewogen gemiddelde van de internetrespons in kwartaal t. De standaardfouten van de directe schattingen, d.w.z. $\sqrt{\hat{V}(\hat{y}_t^R)}$ en $\sqrt{\hat{V}(\hat{y}_t^I)}$, zijn berekend in R (R Core Team, 2015) met het package 'survey' (Lumley, 2014), waarbij rekening gehouden wordt met het steekproefontwerp van de gezondheidsenquête. Er wordt maandelijks een zelfwegende gestratificeerde tweetrapssteekproef getrokken met gemeente als cluster en corop als stratum in de eerste trap met een minimale clusteromvang gelijk aan één. In de tweede trap worden personen per gemeente ge-

trokken. In eerdere jaren werd deelgemeente als cluster gebruikt. Bij het schatten van de standaardfouten gaan we uit van een gestratificeerd steekproefontwerp met de combinatie maand × provincie als stratum, aangezien de combinatie maand × corop tot strata met te weinig respons leidt. Op deze manier zijn de twee inputreeksen voor het bivariate model geconstrueerd. De reguliere reeks bestaat uit de schattingen $\hat{y}_t^{\mathrm{R}}$ en loopt van het eerste kwartaal 2014 t/m het eerste kwartaal van 2020. De internetreeks bestaat uit de schattingen $\hat{y}_t^{\mathrm{I}}$ en loopt van het eerste kwartaal 2014 t/m het tweede kwartaal van 2020.

2.3 De modelgebaseerde schattingen

Aan de hand van model (1) kunnen schattingen gemaakt worden voor de populatieparameter ϑ_t vanaf het eerste kwartaal van 2014 t/m het tweede kwartaal van 2020. Hiertoe berekenen we via het Kalman filter (Durbin en Koopman, 2012) de gefilterde schattingen van $\hat{L}_t + \hat{S}_t$. Gefilterd wil zeggen dat de schatting $\hat{L}_t + \hat{S}_t$ in kwartaal t gebaseerd is op alle beschikbare respons vanaf het eerste kwartaal van 2014 t/m kwartaal t. De nowcasts voor de reguliere reeks in de kwartalen zonder CAPI volgen hier automatisch uit. Hierbij worden twee sterke aannames gemaakt. Ten eerste wordt verondersteld dat de samenstelling van de internetrespons niet verandert tijdens de coronacrisis. Deze aanname is geëvalueerd via een responsanalyse en dit lijkt inderdaad het geval te zijn. De tweede veronderstelling is dat het verschil tussen de CAPI- en CAWI-respons niet verandert door de coronacrisis. Deze modelveronderstelling kan niet worden geëvalueerd.

Om het Kalman filter toe te kunnen passen wordt model (1) eerst als toestandsruimtemodel geschreven:

$$\begin{aligned} &\boldsymbol{y}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t \\ &\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \text{ voor kwartaal } t, \text{ waarbij} \\ &\boldsymbol{\eta}_t \sim \mathcal{N}(0, \mathbf{H}). \end{aligned} \tag{2}$$

In het toestandsruimtemodel wordt de vector $\boldsymbol{y}_t = (\hat{y}_t^{\mathrm{R}}, \hat{y}_t^{\mathrm{I}})'$ van de directe schattingen uitgedrukt in een vector van de niet-waargenomen toestandsvariabelen $\boldsymbol{\alpha}_t$, zoals de trend en de seizoencomponenten. Verder wordt in het toestandsruimtemodel het verloop van deze toestandsvariabelen over de tijd beschreven. De covariantiematrix $\mathbf{H}$ van de vector $\boldsymbol{\eta}_t$ met storingstermen bevat de hyperparameters van het model, die geschat worden met maximum likelihood (ML). De berekeningen zijn uitgevoerd in Ssfpack 3.0 (Koopman et al., 2008) in combinatie met Ox (Doornik, 2009).

Voor iedere doelvariabele zijn de modelaannames van model (1) geëvalueerd door voor zowel de reguliere reeks als de internetreeks de gestandaardiseerde innovaties te analyseren. Innovaties zijn de fouten van de voorspellingen van de waarnemingen $\boldsymbol{y}_t = (\hat{y}_t^{\mathrm{R}}, \hat{y}_t^{\mathrm{I}})'$ gegeven door het Kalman filter gebaseerd op gegevens die tot en met een kwartaal eerder beschikbaar zijn. De door het Kalman filter voorspelde waarden noteren we met $\boldsymbol{y}_{t|t-1} = (\hat{y}_{t|t-1}^{\mathrm{R}}, \hat{y}_{t|t-1}^{\mathrm{I}})'$. De innovaties voor de reguliere reeks zijn dan gegeven door $\hat{y}_t^{\mathrm{R}} - \hat{y}_{t|t-1}^{\mathrm{R}}$ en voor de internetreeks door $\hat{y}_t^{\mathrm{I}} - \hat{y}_{t|t-1}^{\mathrm{I}}$. De gestandaardiseerde innovaties zijn gelijk aan de innovaties gedeeld door de bijbehorende standaardfouten. Onder het model zijn de gestandaardiseerde innovaties standaard normaal verdeeld. Via deze gestandaardiseerde innovaties is het model onderzocht op normaliteit, heteroscedasticiteit en autocorrelatie. Hiervoor zijn een aantal toetsen uitgevoerd, zie Durbin en Koopman, (2012) Hoofdstuk 2 voor details.

Toepassing van model (1) op de doelvariabelen leidt tot verwerping van de normaliteit bij de variabelen huisartsbezoek en dagelijks roken. Bij ervaren gezondheid wordt heteroscedasticiteit gemeten.

Verder worden er geen modelveronderstellingen verworpen. Bij huisartsbezoek en dagelijks roken blijken enkele uitbijters voor te komen in eerdere jaren: bij huisartsbezoek in kwartaal 3 van 2016 en bij dagelijks roken in kwartaal 2 van 2019. De gestandaardiseerde innovaties vallen in deze kwartalen net buiten het interval $(-2,2)$. Bij ervaren gezondheid en de drie variabelen over het zorggebruik worden in de eerste kwartalen van 2020 gestandaardiseerde innovaties gemeten die in absolute waarde variëren van 3 tot 6. Dat is een teken dat voor deze variabelen model (1) de effecten van corona in 2020 niet goed beschrijft en dus aangepast moet worden.

2.4 Aanpassing model vanwege corona-effecten

Schattingen gebaseerd op model (1) lenen informatie uit het verleden om zo de nauwkeurigheid van de schattingen te verbeteren. Er wordt daarbij aangenomen dat de cijfers uit het verleden samenhangen met de huidige cijfers. In het tweede kwartaal van 2020 wordt bij een viertal variabelen een sterke afwijking gemeten in de internetreeks. We zien het zorggebruik sterk afnemen. Het gaat hier om de variabelen huisartsbezoek, tandartsbezoek en specialistbezoek. Bij de variabele ervaren gezondheid zien we juist een sterke toename. Voor deze variabelen geldt de aanname van de samenhang met het verleden dus niet meer en passen we het model aan. Dit wordt gesignaleerd doordat de gestandaardiseerde innovaties waardes aannemen die (absoluut) veel groter zijn dan 2 (zie vorige paragraaf).

Model (1) leent informatie uit het verleden via zowel de trend L_t als het seizoen S_t. De verspreiding van corona kan zowel invloed hebben op de trend als het seizoenspatroon. Aangezien het niet mogelijk is het effect op het seizoenspatroon apart te schatten nemen we in het model aan dat corona alleen effect heeft op de dynamiek van de trend. In model (1) is de trend gemodelleerd als een smooth trend model:

$$L_t = L_{t-1} + R_{t-1}$$
$$R_t = R_{t-1} + \eta_t^R, \text{ waarbij}$$

$$\eta_t^R \sim \mathcal{N}(0, \sigma_{\mathrm{R}}^2)$$
$$\mathrm{cov}(\eta_t^R, \eta_{t'}^R) = 0, \text{ voor } t \neq t'.$$

De trend is opgebouwd uit een niveau L_t en een hellingsparameter R_t (de slope), waarbij de variantie σ_{R}^2 van de storingsterm η_t^R van de slope constant is over de tijd. De variantie bepaald de flexibiliteit van de trend. Voor de variabelen over het zorggebruik en de ervaren gezondheid maken we de trend in model (1) flexibeler door de variantie σ_{R}^2 met een tijdsafhankelijke factor $f_t \geq 1$ te vermenigvuldigen:

$$\eta_t^R \sim \mathcal{N}(0, f_t\sigma_{\mathrm{R}}^2)$$
$$\mathrm{cov}(\eta_t^R, \eta_{t'}^R) = 0, \text{ voor } t \neq t'.$$

De kwartalen waar $f_t > 1$ gekozen wordt en de waarden van f_t bij die kwartalen worden bepaald aan de hand van de gestandaardiseerde innovaties. Hierbij zorgen we ervoor dat de gestandaardiseerde innovaties in alle kwartalen niet te ver buiten het interval $(-2,2)$ vallen, zodat de normaliteit van de innovaties gegarandeerd blijft. Ook proberen we de variantie van de slope storingstermen $f_t\sigma_{\mathrm{R}}^2$ zo min mogelijk aan te passen (dat wil zeggen, waarden voor f_t zo klein mogelijk houden) , zodat het model nog informatie kan lenen uit het verleden. Merk op dat het aanpassen van σ_{R}^2 in kwartaal t invloed heeft op de storingsterm in kwartaal t en daarom op de slope pas in kwartaal $t + 1$ en op de

trend pas in kwartaal $t + 2$. Er is dus een vertraging van twee kwartalen in het effect op de uitkomsten na aanpassing van σ_{R}^2.

Na analyse van de gestandaardiseerde innovaties passen we bij de variabelen over het zorggebruik de variantie σ_{R}^2 aan van kwartaal 3 in 2019 t/m kwartaal 2 in 2020 en bij ervaren gezondheid al vanaf een kwartaal eerder, dus vanaf kwartaal 2 in 2019 t/m kwartaal 2 in 2020. Bij de andere variabelen is het niet nodig om de slope flexibeler te maken, omdat de gestandaardiseerde innovaties daar geen aanleiding toe geven. Tabel 1 geeft een overzicht van de modelaanpassingen per doelvariabele.

Tabel 1. Aanpassingen model (1) per doelvariabele

doelvariabele	aanpassing trend ($f_t > 1$)	aanpassing f_t in
ervaren gezondheid	ja	2019 kw2 t/m 2020 kw2
psychisch ongezond	nee	
huisartsbezoek	ja	2019 kw3 t/m 2020 kw2
dagelijks roken	nee	
overgewicht	nee	
overmatig alcoholgebruik	nee	
tandartsbezoek	ja	2019 kw3 t/m 2020 kw2
specialistbezoek	ja	2019 kw3 t/m 2020 kw2

Na deze aanpassingen worden er bij de meeste doelvariabelen geen modelveronderstellingen meer verworpen. Alleen bij huisartsbezoek en dagelijks roken is er nog een verwerping van de veronderstelde normaliteit. Bij deze variabelen blijken in enkele kwartalen nog uitbijters voor te komen, maar de gestandaardiseerde innovaties in deze kwartalen vallen net buiten het interval $(-2{,}2)$. Bij huisartsbezoek gaat het om de kwartaal 3 in 2016 en kwartaal 1 in 2020. Bij dagelijks roken gaat het om kwartaal 2 in 2019.

In Bijlage II zijn voor alle doelvariabelen de resultaten van de modelschattingen weergegeven. Tabel 2 toont de ML-schattingen van de hyperparameters van model (1). De figuren 1 t/m 16 laten de modelgebaseerde schattingen (STM) zien met de bijbehorende standaardfouten vanaf 2017 en vergelijken deze met de directe schattingen, gebaseerd op de volledige respons (internet waarneming en face-to-face waarneming) en de internetrespons (volledig).

Merk op dat de kwartaalcijfers over de jaren 2017 t/m 2020 gebaseerd zijn op het structureel tijdreeksmodel terwijl de eerder gepubliceerde jaarcijfers directe schattingen zijn, gebaseerd op de weging zoals beschreven in Boonstra (2019). Hierdoor wijkt het gemiddelde van de kwartaalcijfers soms iets af van de jaarcijfers.

Op basis van de kwartaalcijfers kunnen, net als bij de jaarcijfers, jaarontwikkelingen berekend worden door het verschil te berekenen van hetzelfde kwartaal in twee opeenvolgende jaren. Aangezien de kwartaalcijfers op het structureel tijdreeksmodel gebaseerd zijn, wijken ook de standaardfouten van de jaarontwikkelingen gebaseerd op de kwartaalcijfers af van de standaardfouten van de jaarontwikkelingen gebaseerd op de jaarcijfers. Over het algemeen zijn de jaarontwikkelingen van de modelgebaseerde kwartaalcijfers nauwkeuriger dan die van de directe jaarcijfers. Hierdoor zal er bij de kwartaalcijfers eerder een significante jaarontwikkeling gemeten worden dan bij de jaarcijfers.

3. Discussie

Tot op heden werden op basis van de Gezondheidsenquête (GE) op jaarbasis cijfers over de gezondheid, de leefstijl en het zorggebruik van de Nederlandse bevolking gepubliceerd. Naar aanleiding van de corona crisis en de daarmee gepaard gaande lockdown is besloten om aan de hand van een structureel tijdreeksmodel cijfers op kwartaalbasis te publiceren. Dit dient meerdere doelen. Ten eerste kan met kwartaalcijfers de coronaperiode scherper afgebakend worden, waardoor mogelijke effecten van de crisis op cijfers met trekking tot gezondheid, zorggebruik en leefstijl duidelijker in beeld worden gebracht. Bovendien komen kwartaalcijfers sneller beschikbaar, namelijk al gedurende 2020 in plaats van begin 2021. Hierdoor neemt de relevantie van de GE cijfers duidelijk toe. Ook is het mogelijk om met het tijdreeksmodel te corrigeren voor het wegvallen van de face-to-face waarneming gedurende de lockdown.

De correctie voor het wegvallen van de face-to-face waarneming via het tijdreeksmodel is gebaseerd op twee sterke aannames. Ten eerste wordt verondersteld dat de samenstelling van de internetrespons niet verandert tijdens de coronacrisis. Deze aanname is geëvalueerd via een responsanalyse en dit lijkt inderdaad het geval te zijn. De tweede veronderstelling is dat het verschil tussen de CAPI- en CAWI-respons niet verandert door de coronacrisis. Deze modelveronderstelling kan niet worden geevalueerd. De correcties voor het wegvallen van de face-to-face waarneming kan in de jaarcijfers worden verwerkt door in de weging van het jaarbestand een tabel op te nemen met de gecorrigeerde kwartaalcijfers van de belangrijkste doelvariabelen, geschat via het tijdreeksmodel.

Een bijkomend voordeel van het tijdreeksmodel is dat de schattingen op basis van het model nauwkeuriger zijn dan de directe schattingen. Met name periode-op-periode veranderingen kunnen veel nauwkeuriger worden geschat dankzij de positieve correlatie tussen de trendschattingen van de opeenvolgende perioden.

Voor een aantal variabelen lijkt de coronacrisis een sterk effect te hebben op de ontwikkeling. Om in het tijdreeksmodel rekening te houden met de plotselinge toename in de dynamiek van deze cijfers, is het noodzakelijk de trendcomponent van het model flexibeler te maken. Dit is gedaan door de variantie van de storingstermen van de trendcomponent gedurende de corona crisis groter te maken. Het gevolg is dat de standaardfouten van de tijdreeksschattingen voor deze perioden groter worden.

De cijfers over het eerste kwartaal van 2020 zijn behandeld alsof de face-to-face waarneming volledig beschikbaar is. Daarbij is het wegvallen van deze waarnemingen in de laatste twee weken van maart dus genegeerd, hetgeen een redelijke veronderstelling is. Een openstaande vraag is hoe gecorrigeerd moet worden voor het wegvallen van de face-to-face waarneming in het kwartaal waarin de face-to-face waarneming weer wordt opgestart. De meest geschikte methodiek hangt af van de maand binnen het betreffende kwartaal waar de face-to-face waarneming weer beschikbaar komt. Daarnaast moet in de komende periode worden bezien of het noodzakelijk is om de cijfers voor de tot nu toe gepubliceerde kwartalen over 2020 te reviseren op basis van de uitkomsten die na afloop van volgende kwartalen beschikbaar komen.

Referenties

Binder, D.A., en J.P. Dick (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, pp. 239-253.

Boonstra, H.J. (2019). *Weging Gezondheidsenquête 2014*. Discussion paper CBS, Centraal Bureau voor de Statistiek, Heerlen. URL: https://www.cbs.nl/nl-nl/achtergrond/2019/18/weging-gezondheidsenquete-2014

Doornik, J.A. (2009). *An Object-oriented Matrix Programming Language Ox 6*. London: Timberlake Consultants Press.

Durbin, J. en S. Koopman (2012). *Time Series Analysis by State Space Methods (second edition)*. Oxford University Press, Oxford.

Koopman, S.J., Shephard, N. en Doornik, J.A. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form.* London: Timberlake Consultants Press.

Lumley, T. (2014). '*survey': analysis of complex survey samples.* R package version 3.30.

R Core Team (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Särndal, C-E, B. Swensson en J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer Verlag.

Bijlage I. Uitwerking model (1)

In model (1) is de trend L_t als een smooth trend gemodelleeerd, gegeven door

$$L_t = L_{t-1} + R_{t-1} \qquad \text{(I.1)}$$
$$R_t = R_{t-1} + \eta_t^R, \text{ waarbij}$$

$$\eta_t^R \sim \mathcal{N}(0, \sigma_R^2)$$
$$\mathrm{Cov}(\eta_t^R, \eta_{t'}^R) = 0, \text{ voor } t \neq t'.$$

De seizoenscomponent S_t is met een trigonometrisch model gemodelleerd en gegeven door

$$S_t = \gamma_{1,t} + \cdots + \gamma_{J/2,t}, \text{ waarbij} \qquad \text{(I.2)}$$

$$\gamma_{j,t} = \gamma_{j,t-1}\cos\left(\tfrac{\pi j}{J/2}\right) + \gamma_{j,t-1}^*\sin\left(\tfrac{\pi j}{J/2}\right) + \omega_{j,t}$$
$$\gamma_{j,t}^* = \gamma_{j,t-1}^*\cos\left(\tfrac{\pi j}{J/2}\right) - \gamma_{j,t-1}\sin\left(\tfrac{\pi j}{J/2}\right) + \omega_{j,t}^* \text{ voor } j = 1, \ldots, J/2.$$

Voor kwartalen $J = 4$ geldt $\gamma_{j,t}^* = 0$ omdat $\sin(\pi) = 0$. Daarom vereenvoudigt het trigonometrisch model tot

$$S_t = \gamma_{1,t} + \gamma_{2,t}, \text{ waarbij}$$

$$\gamma_{1,t} = \gamma_{1,t-1}^* + \omega_{1,t}$$
$$\gamma_{1,t}^* = -\gamma_{1,t-1} + \omega_{1,t}^*$$

$\gamma_{2,t} = -\gamma_{2,t-1} + \omega_{2,t}$, met

$\omega_{1,t} \sim \mathcal{N}(0, \sigma_\omega^2)$
$\omega_{1,t}^* \sim \mathcal{N}(0, \sigma_\omega^2)$
$\omega_{2,t} \sim \mathcal{N}(0, \sigma_\omega^2)$, en

$\mathrm{Cov}(\omega_{j,t}, \omega_{j,t'}) = 0$, voor $t \neq t'$ en $j = 1, 2$,
$\mathrm{Cov}(\omega_{1,t}^*, \omega_{1,t'}^*) = 0$, voor $t \neq t'$,
$\mathrm{Cov}(\omega_{j,t}, \omega_{1,t}^*) = 0$, voor alle kwartalen t en $j = 1, 2$,
$\mathrm{Cov}(\omega_{1,t}, \omega_{2,t}) = 0$, voor alle kwartalen t.

Het systematisch verschil λ_t tussen de reguliere reeks en de internetreeks is gemodelleerd als een random walk, gegeven door

$$\lambda_t = \lambda_{t-1} + \eta_{\lambda,t}, \text{ waarbij} \qquad \text{(I.3)}$$

$\eta_{\lambda,t} \sim \mathcal{N}(0, \sigma_\lambda^2)$
$\mathrm{Cov}(\eta_{\lambda,t}, \eta_{\lambda,t'}) = 0$, voor $t \neq t'$.

De meetfout e_t^{j} is een combinatie van steekproefruis en ruis in de populatieparameter. Omdat de gezondheidsenquête gebaseerd is op een cross-sectioneel onderzoek is het niet mogelijk deze twee termen met een structureel tijdreeksmodel te scheiden. In het model houden we wel rekening met veranderingen in de variantie van de directe schattingen, veroorzaakt door veranderingen in de responsomvang en het steekproefontwerp. Ook wordt er rekening gehouden met de overlap van de respons waarop de directe schattingen $\hat{y}_t^{\mathrm{R}}$ en $\hat{y}_t^{\mathrm{I}}$ gebaseerd zijn.

De meetfout e_t^{j} is gemodelleerd met het volgende meetfoutmodel (Binder en Dick, 1990):

$$e_t^j = \sqrt{\widehat{\mathbb{V}}(\hat{y}_t^j)}\,\tilde{e}_t^j \text{ voor } j \in \{\mathrm{R}, \mathrm{I}\}, \text{ waarbij} \qquad \text{(I.4)}$$

$\tilde{e}_t^j \sim \mathcal{N}(0, \sigma_{\mathrm{e},j}^2)$
$\mathrm{Cov}(\tilde{e}_t^j, \tilde{e}_{t'}^j) = 0$, voor $t \neq t'$ en
$\mathrm{Cov}(\tilde{e}_t^{\mathrm{R}}, \tilde{e}_t^{\mathrm{I}}) = \frac{\sqrt{n_t^{\mathrm{I}}}}{\sqrt{n_t^{\mathrm{R}}}}$.

De laatste formule impliceert

$$\mathrm{Cov}(e_t^{\mathrm{R}}, e_t^{\mathrm{I}}) = \frac{\sqrt{n_t^{\mathrm{I}}}}{\sqrt{n_t^{\mathrm{R}}}} \sqrt{\widehat{\mathbb{V}}(\hat{y}_t^R)} \sqrt{\widehat{\mathbb{V}}(\hat{y}_t^I)}.$$

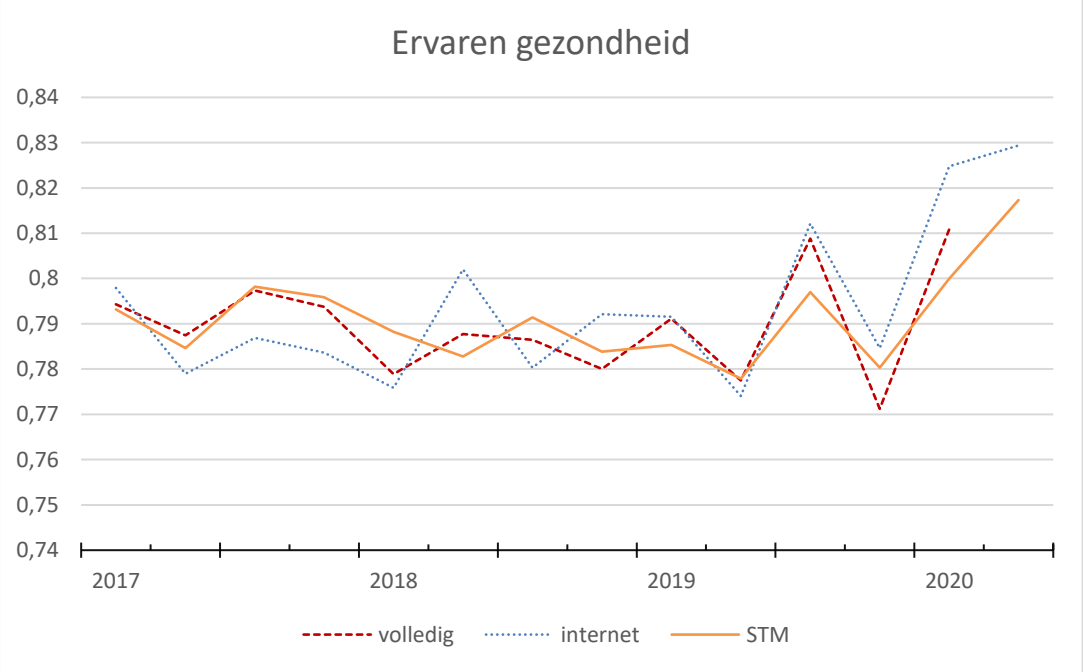
Hierin is n_t^{I} de omvang van de internetrespons in kwartaal t en n_t^{R} de omvang van de volledige respons in kwartaal t.

Bijlage II. Resultaten

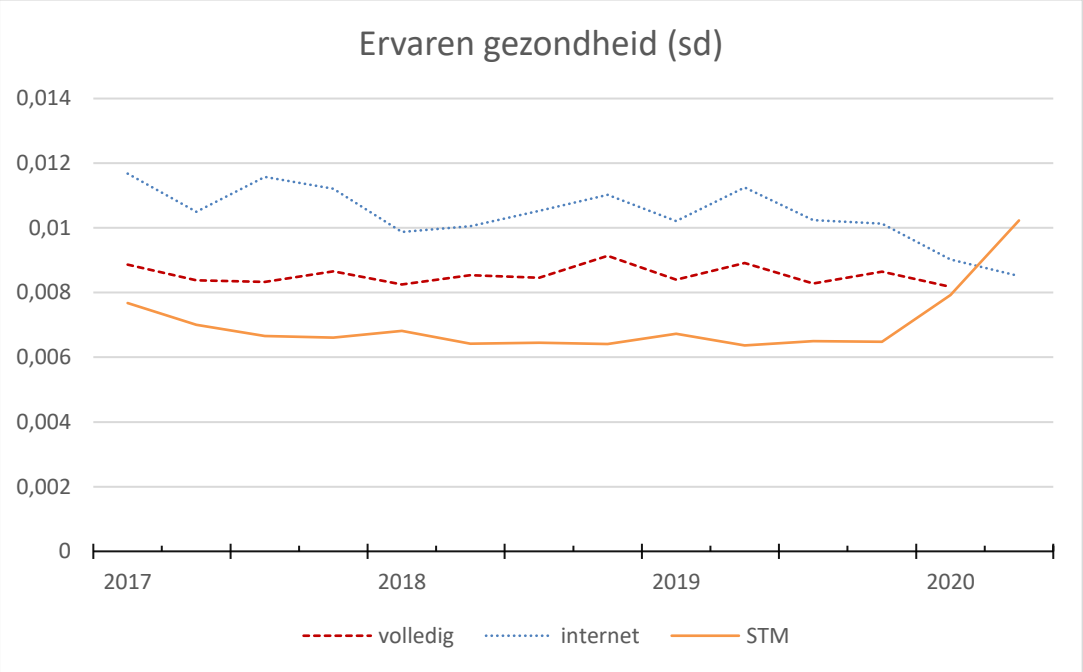
Tabel 2. ML-schattingen hyperparameters model (1)

doelvariabele	$\hat{\sigma}_{\eta}$	$\hat{\sigma}_{\omega}$	$\hat{\sigma}_{\lambda}$	$\hat{\sigma}_{e,R}$	$\hat{\sigma}_{e,I}$
ervaren gezondheid	0,001	<0,001	<0,001	1,120	1,340
psychisch ongezond	<0,001	<0,001	0,003	0,964	0,783
huisartsbezoek	0,002	<0,001	<0,001	1,030	1,210
dagelijks roken	<0,001	<0,001	0,003	1,220	1,210
overgewicht	<0,001	<0,001	<0,001	1,080	1,030
overmatig alcoholgebruik	<0,001	0,004	0,003	0,933	0,005
tandartsbezoek	0,003	<0,001	<0,001	1,130	1,080
specialistbezoek	0,002	0,003	0,007	0,619	<0,001

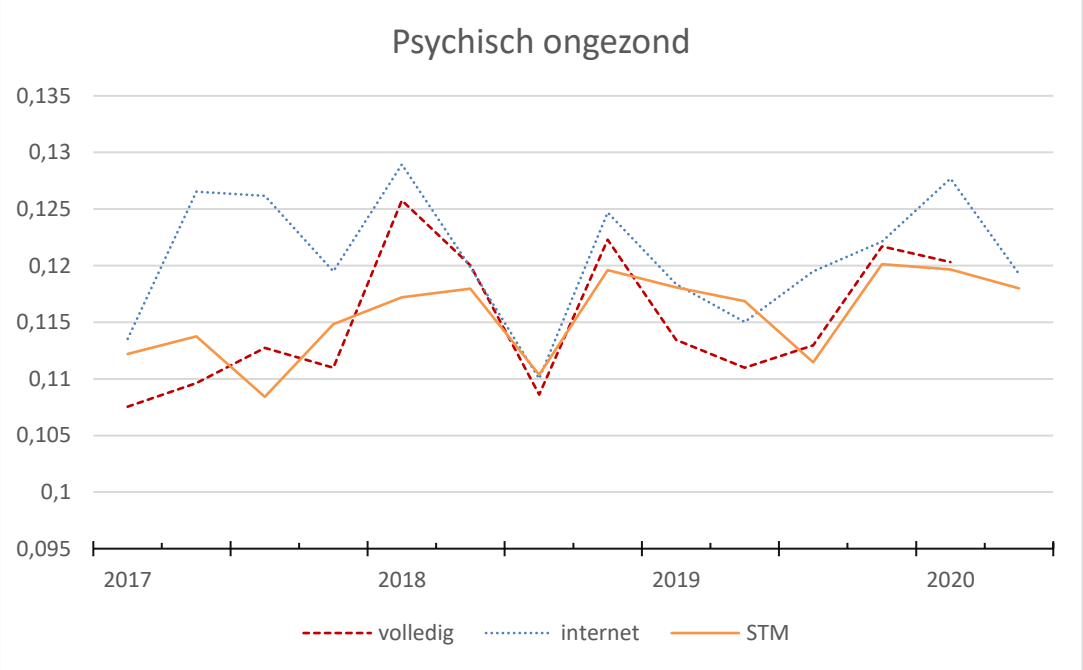
Figuur 1. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor ervaren gezondheid



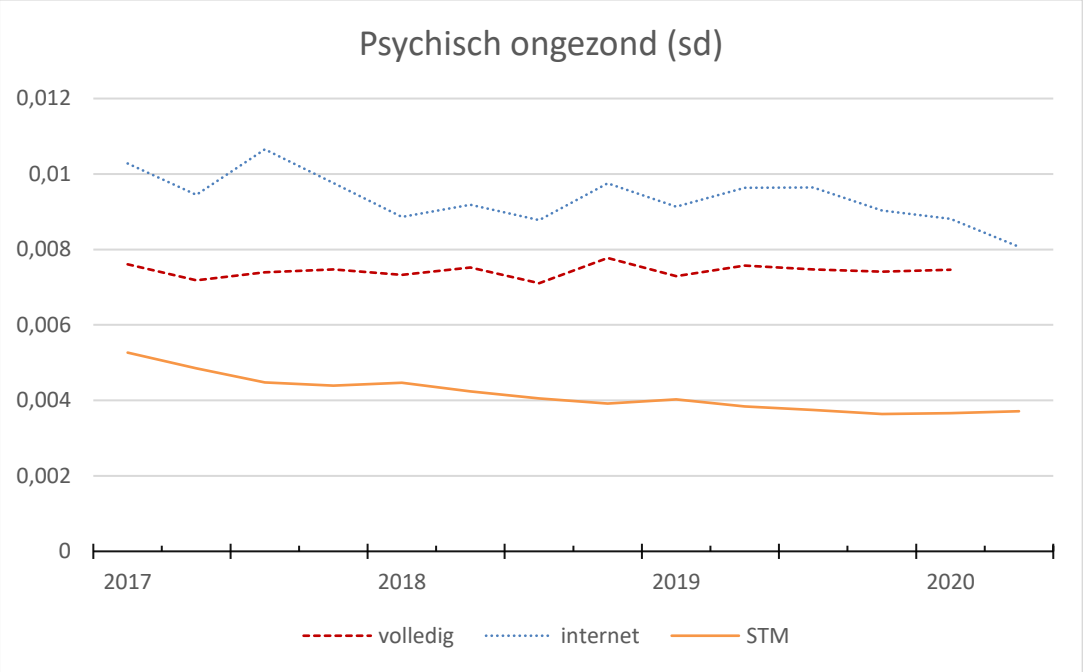
Figuur 2. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor ervaren gezondheid



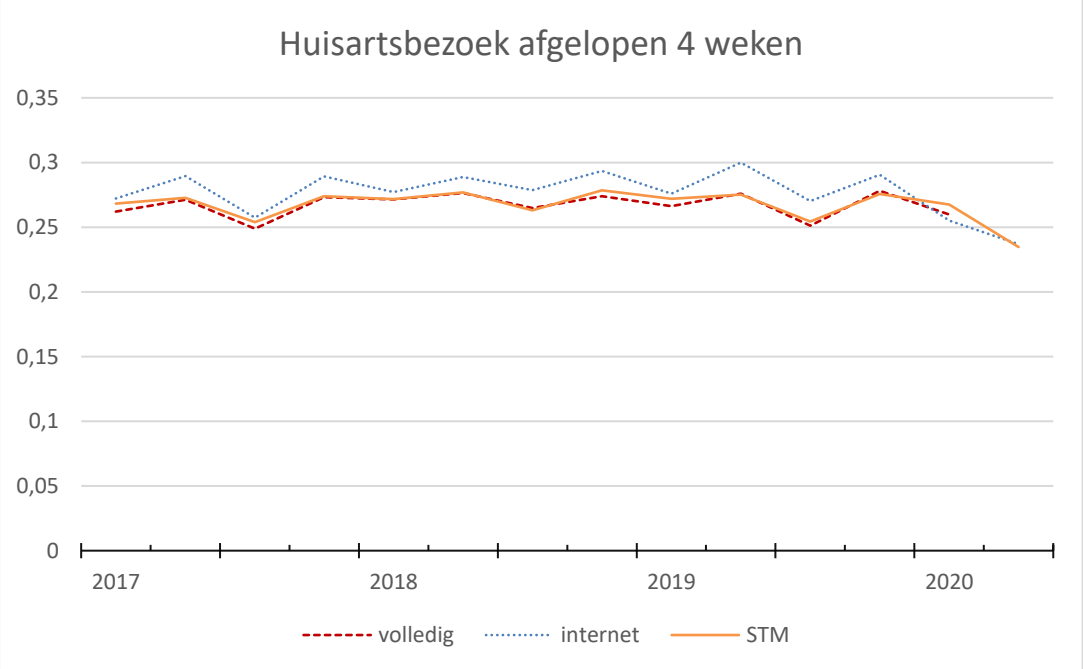
Figuur 3. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor variabele psychisch ongezond



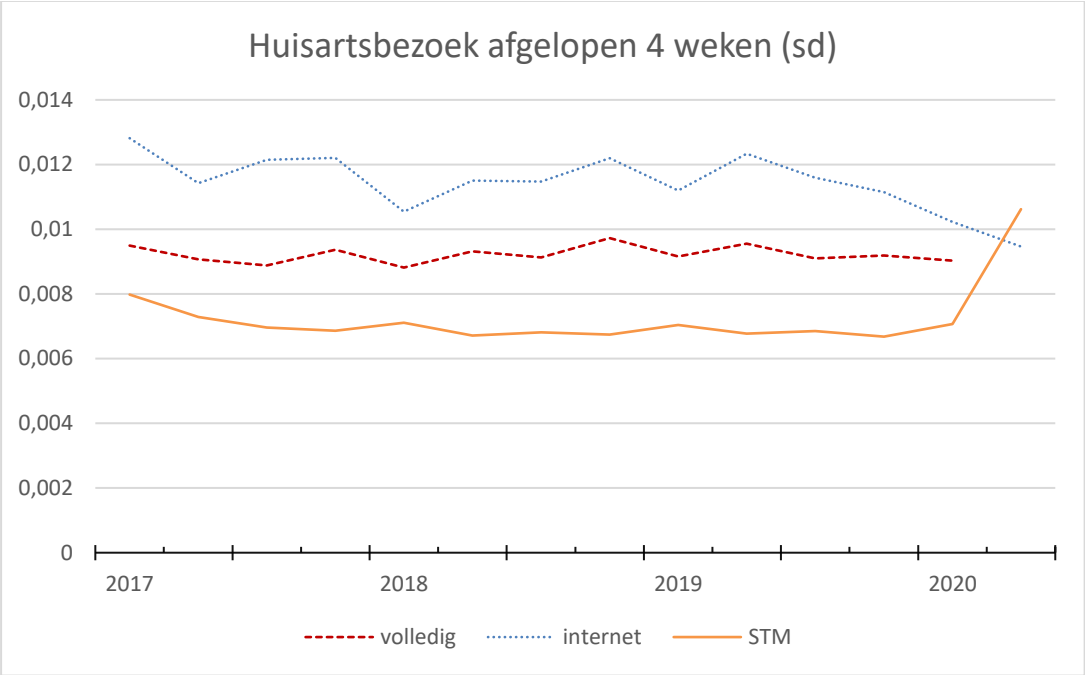
Figuur 4. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor psychisch ongezond



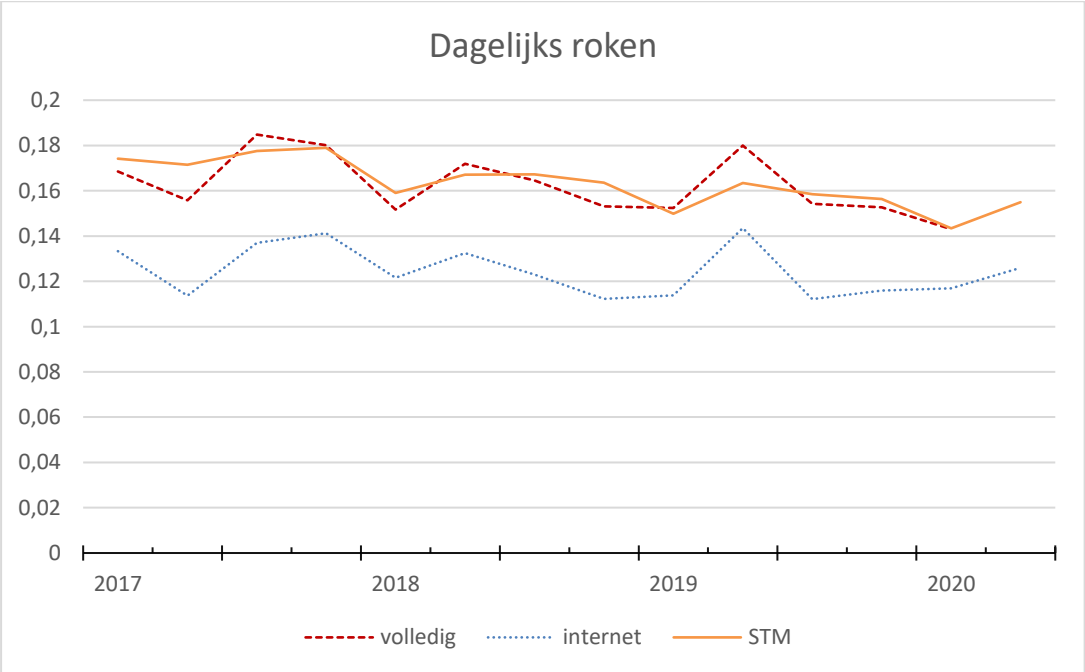
Figuur 5. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor huisartsbezoek afgelopen 4 weken



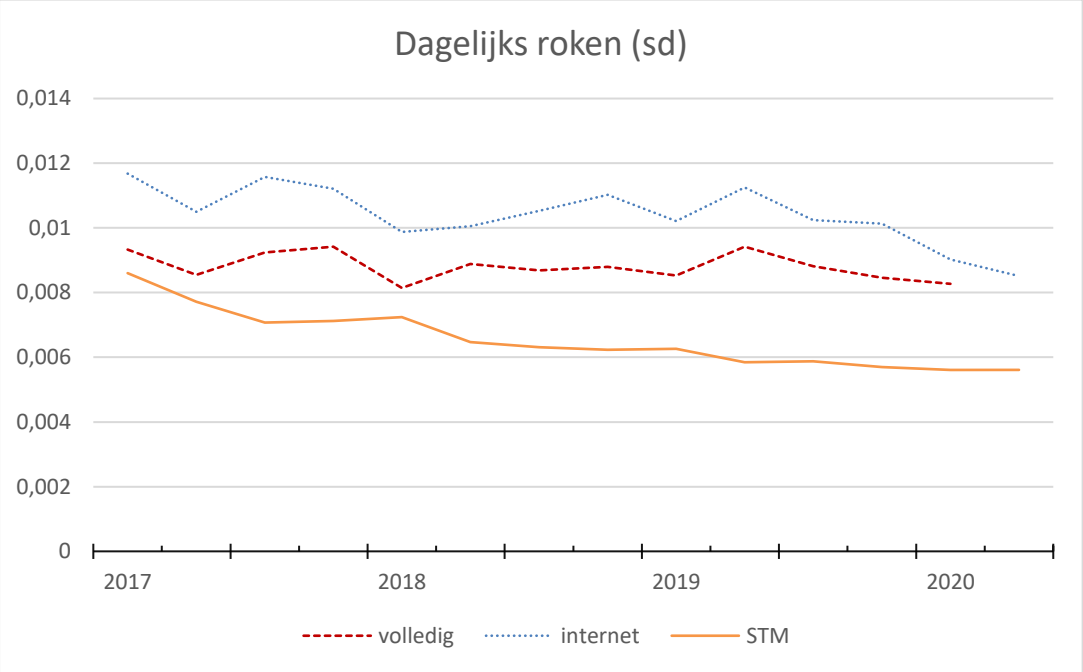
Figuur 6. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor huisartsbezoek afgelopen 4 weken



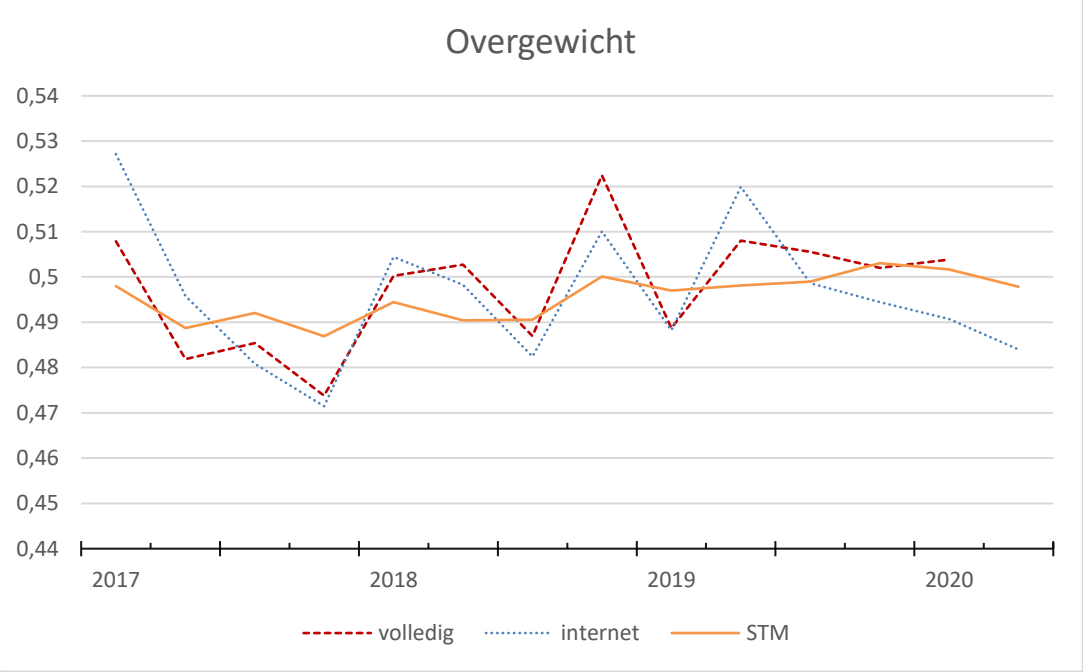
Figuur 7. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor dagelijks roken



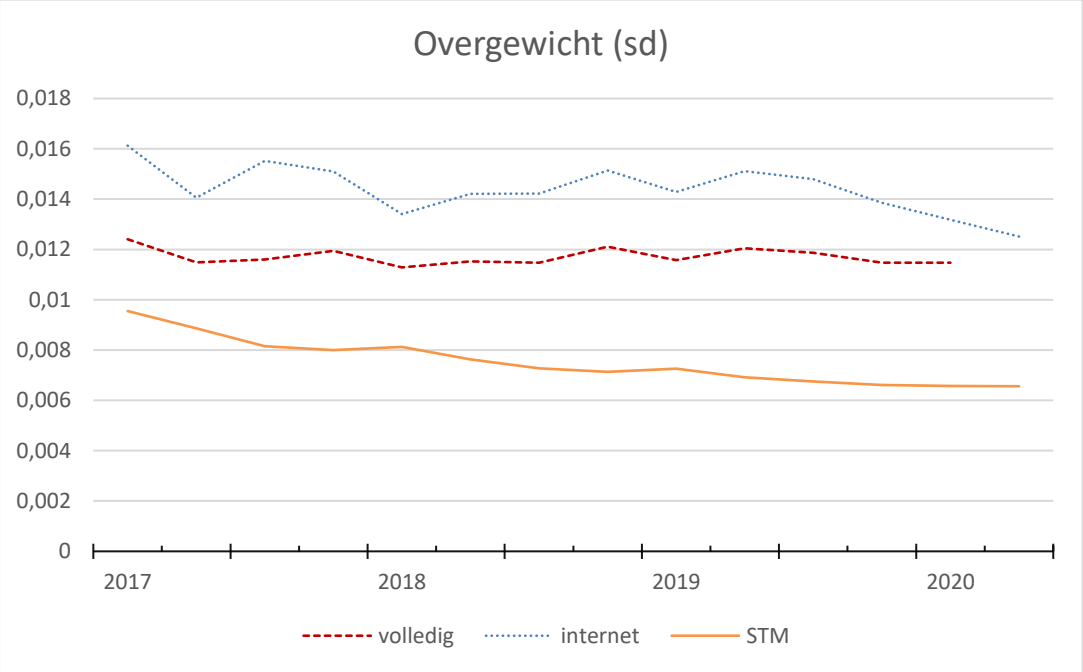
Figuur 8. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor dagelijks roken



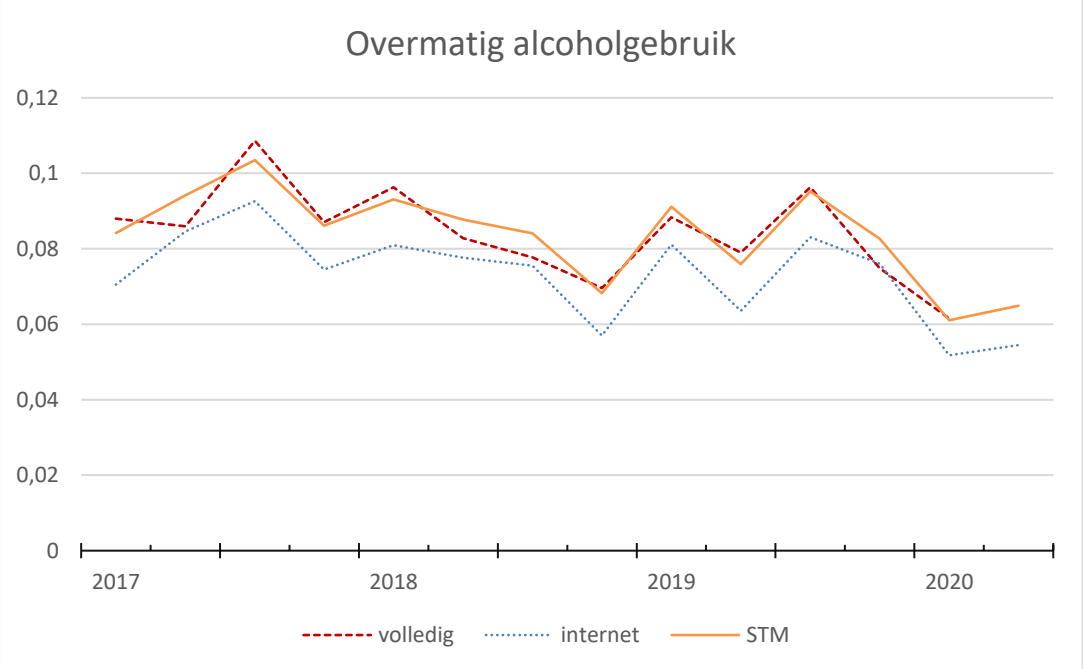
Figuur 9. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor overgewicht



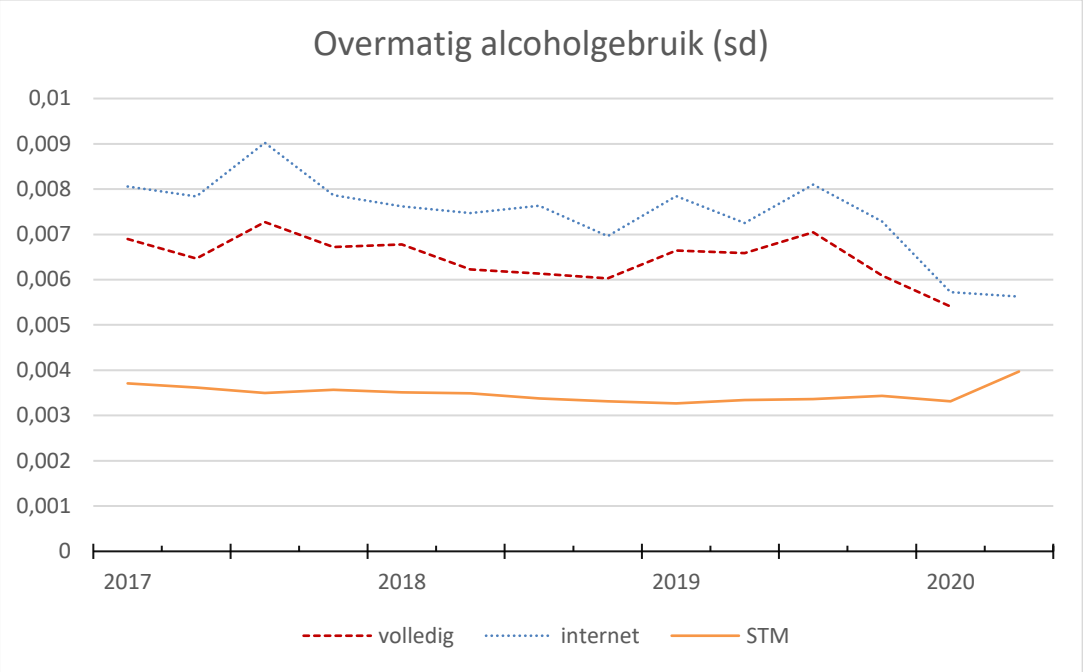
Figuur 10. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor overgewicht



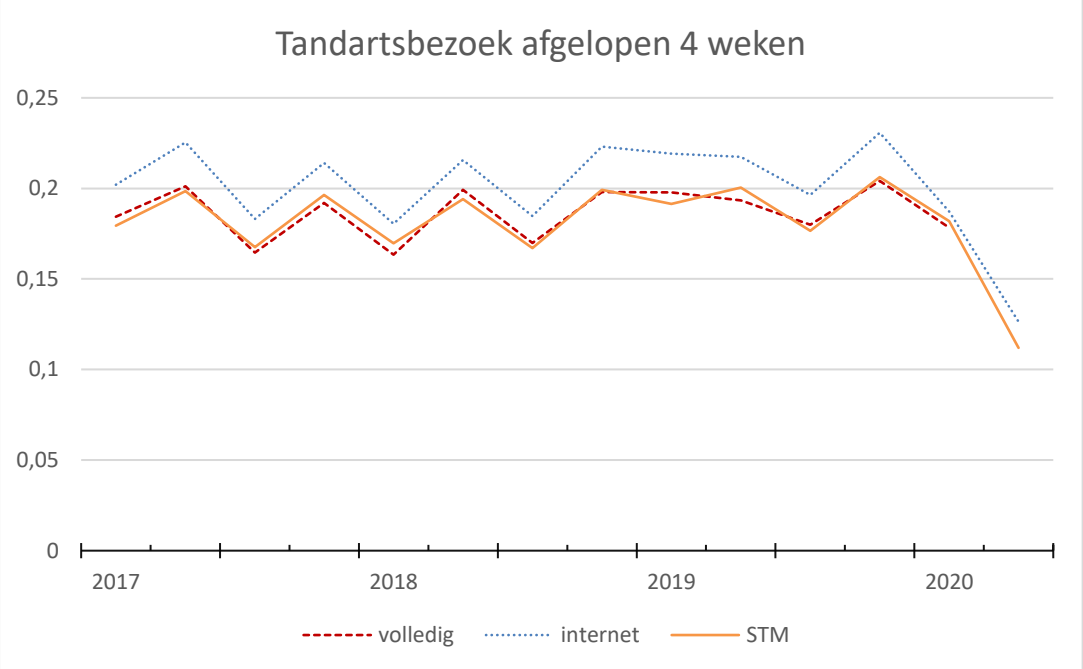
Figuur 11. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor overmatig alcoholgebruik



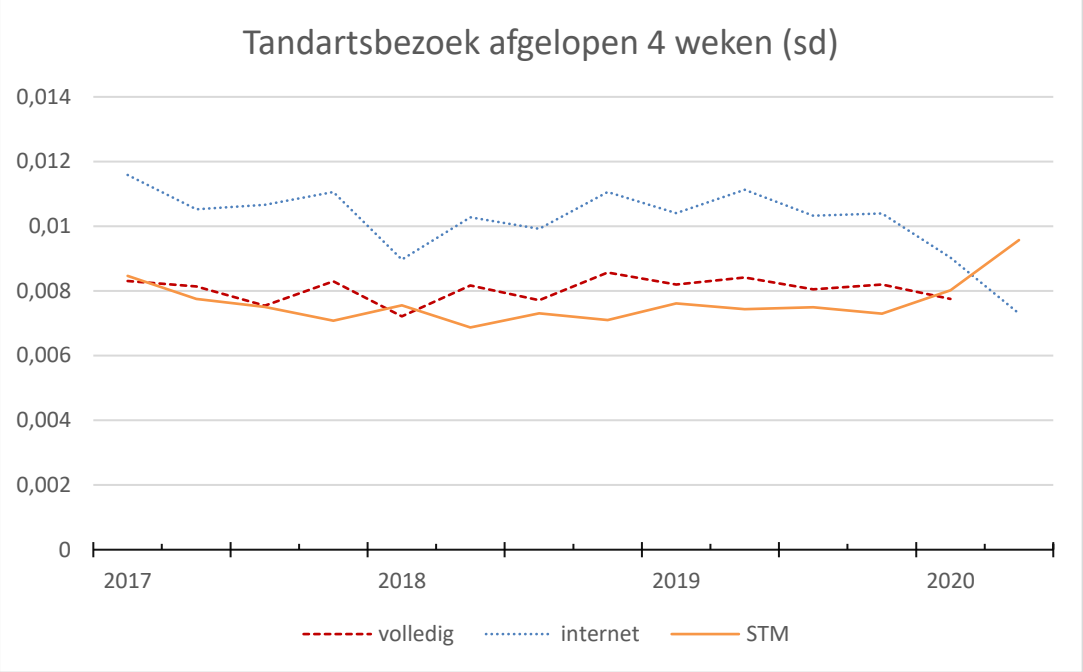
Figuur 12. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor overmatig alcoholgebruik



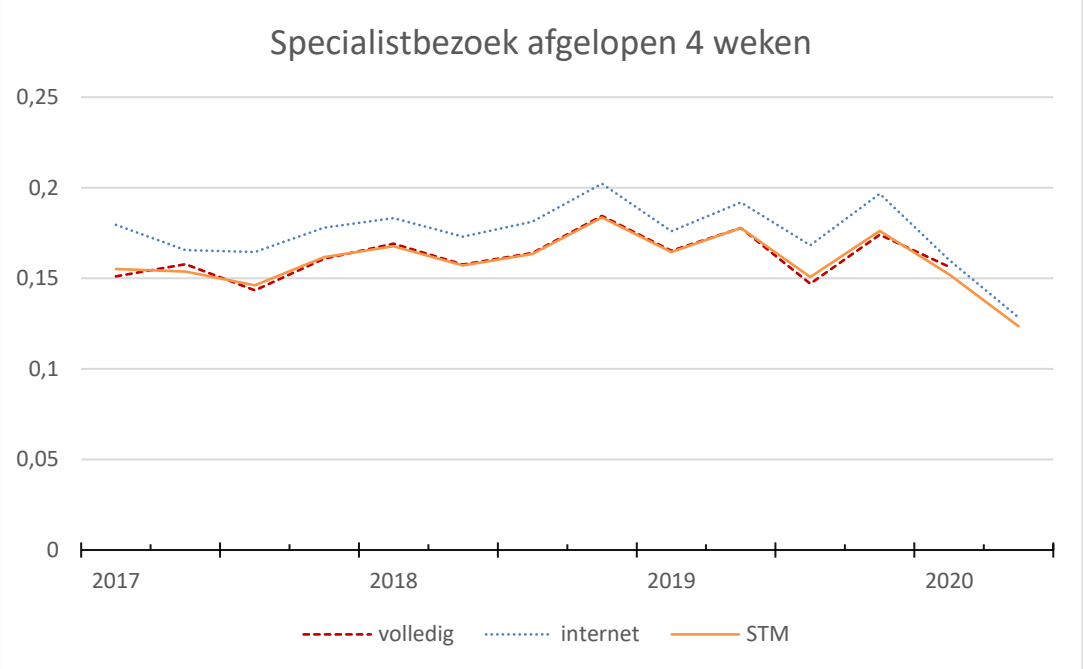
Figuur 13. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor tandartsbezoek afgelopen 4 weken



Figuur 14. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor tandartsbezoek afgelopen 4 weken



Figuur 15. Modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor specialistbezoek afgelopen 4 weken



Figuur 16. Standaardfouten van modelgebaseerde schattingen (STM) en directe schattingen gebaseerd op volledige respons (volledig) en internetrespons (internet) voor specialistbezoek afgelopen 4 weken