Official statistics based on the Dutch Health Survey during the COVID-19 Pandemic

Jan van den Brakel[†‡], Marc Smeets[†]

†: Statistics Netherlands

‡: Maastricht University

Virtual International Total Survey Error Workshop (ITSEW) 2021

01-10-2021

Introduction

Dutch Health Survey:

- Sequential mixed-mode design, based on web interviewing (WI), face to face interviewing (CAPI)
- About 9,000 responses annually
- Annual figures on health, medical contacts, lifestyle and preventive behaviour of the Dutch population

COVID-19:

- Partial loss of CAPI during the lockdown in Q2: bias
- Annual figures:
  - Not timely (information published in Q1 2021 will be outdated)
  - No accurate information of the impact of COVID-19 on health figures

Introduction

COVID-19:

- Strong (external) demand for timely figures
- How to accommodate for loss of CAPI and suddenly changing mode effects?

Solution model-based inference method

- Based on structural time series model (STM):
- Production of official quarterly figures (time series model as a form of SAE)
- STM also correct for changing mode-effects

Dutch Health Survey

Continuously conducted survey

Sample design:

- Stratified simple random sampling of persons
- Regional stratification (COROP)
- Data collection: WI, non-response follow up with CAPI
- Response:
  - Gross sample: 15,000 persons
  - WI response: 7,300
  - Approached by CAPI: 7,900
  - CAPI response: 1,500
  - Total response 8,800

Dutch Health Survey

Estimation:

- General regression estimator (Särndal et al. 1992) using
  - Gender in 2 classes
  - Age in 16 classes
  - Marital status in 4 classes
  - Urbanisation in 5 classes
  - Regional classification based on province plus 4 major cities in 16 classes
  - Nationality in 4 classes
  - Quarters
  - Income in 5 classes
  - Equity in 5 classes
  - Crossings between gender, age and marital status

Dutch Health Survey

Official figures:

- Annuel figures on health, medical contacts, lifestyle and preventive behaviour of the Dutch population
- Figures defined as percentages

COVID-19

- Due to lockdown regulations no capi in 16-th March 2020 - July 2020
- Loss of CAPI:
  - Sudden change in mode effects (mixture of selection effects and measurement bias)
  - Mode effects will be confounded with the real effect of COVID on health figures
- Strong external demand on more timely health figures
  - Sample size is to small to produce quarterly figures
  - Annual figures
    - Obscured picture of COVID effect
    - Come with an unacceptable delay

COVID-19

- Proposed solution:
  - Quarterly figures
  - Requires a model-based inference method
  - Structural time series model:
    - Small area estimation (borrow strength over time)
    - Model is also used to compensate for changing mode effects due to the loss of CAPI

Structural time series model

Measurement error model:

$$\hat{y}_t = \theta_t + \tilde{e}_t$$

θ_t: unknown population parameter in quarter t

$\hat{y}_t$: sample estimate for θ_t

$\tilde{e}_t$: sampling error

Structural time series model (STM) for θ_t:

$$\theta_t = L_t + S_t + I_t$$

L_t: smooth trend model

S_t: trigonometric seasonal model

I_t: unexplained variation population parameter

Structural time series model

Smooth trend model

$$L_t = L_{t-1} + R_{t-1}$$

$$R_t = R_{t-1} + \eta_t^R$$

with

$$\eta_t^R \sim \mathcal{N}(0, \sigma_{\mathrm{R}}^2)$$

$$\mathrm{cov}(\eta_t^R, \eta_{t'}^R) = 0, \text{ for } t \neq t'$$

Structural time series model

Trigonometric seasonal model

$$S_t = \sum_{j=1}^{S/2} \gamma_{j,t}$$

with

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

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

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

S : number of seasons

Structural time series model

Trigonometric seasonal model for quarterly data S = 4

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

with

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

$$\gamma^*_{1,t} = -\gamma_{1,t-1} + \omega^*_{1,t}$$

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

$\omega_{1,t} \sim \mathcal{N}(0, \sigma^2_\omega)$, $\omega^*_{1,t} \sim \mathcal{N}(0, \sigma^2_\omega)$ and $\omega_{2,t} \sim \mathcal{N}(0, \sigma^2_\omega)$

Structural time series model

Measurement error model:

$$\hat{y}_t \ = \theta_t + \tilde{e}_t$$

STM:

$$\theta_t = L_t + S_t + I_t$$

Inserting measurement error model in STM:

$$\hat{y}_t \ = L_t + S_t + I_t + \tilde{e}_t \equiv L_t + S_t + e_t$$

e_t:

- Sampling error
- Unexplained population variation

Structural time series model

So far:

- STM to borrow strength over time
- Also required is a correction for the loss of CAPI response:
  - Bivariate STM
  - Input:
    $\hat{y}_t^{\mathrm{C}}$: Series of direct estimates based on complete response
    $\hat{y}_t^{\mathrm{I}}$: Series of direct estimates based on Internet response only
  - During the lockdown $\hat{y}_t^{\mathrm{C}}$ is missing
  - STM produces nowcasts based on the relation between $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{I}}$ observed in the past

Structural time series model

Bivariate model:

$$\begin{pmatrix} \hat{y}_t^{\mathrm{C}} \\ \hat{y}_t^{\mathrm{I}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (L_t + S_t) + \begin{pmatrix} \\ \lambda_t \end{pmatrix} + \begin{pmatrix} e_t^{\mathrm{C}} \\ e_t^{\mathrm{I}} \end{pmatrix}$$

λ_t:

- Systematic difference between $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{I}}$
- Modeled with a random walk (varies over time):

$$\lambda_t = \lambda_{t-1} + \eta_{\lambda,t}$$

$$\eta_{\lambda,t} \sim \mathcal{N}\left(0, \sigma_\lambda^2\right)$$

$$\mathrm{Cov}\left(\eta_{\lambda,t}, \eta_{\lambda,t'}\right) = 0, \text{ for } t \neq t'$$

Structural time series model

Measurement error e_t^{C} and e_t^{I}:

- Sampling error and unexplained population variation
- Heteroscedasticity:
  - $e_t^{\mathrm{C}} \sim \mathcal{N}\left(0, \sigma_{e^c}^2 \widehat{V(\hat{y}_t^{\mathrm{C}})}\right)$
  - $e_t^{\mathrm{I}} \sim \mathcal{N}\left(0, \sigma_{e^I}^2 \widehat{V(\hat{y}_t^{\mathrm{I}})}\right)$
  - $\widehat{V(\hat{y}_t^{\mathrm{C}})}$ and $\widehat{V(\hat{y}_t^{\mathrm{I}})}$ variances of the GREG estimates of the input series (and are derived from the micro data)

Structural time series model

Measurement error e_t^{C} and e_t^{I}:

- Correlated because e_t^{C} and e_t^{I} are based on the same web respondents

$$Corr\left(\hat{y}_t^{\mathrm{C}}, \hat{y}_t^{\mathrm{I}}\right) = \rho \frac{n^{ol}}{\sqrt{n^C}\sqrt{n^I}}$$ (Kish, 1965)

  - ρ: correlation of the sample overlap
  - n^{ol}: number of observations in sample overlap
  - n^C: number of observations of complete response
  - n^I: number of observations of internet response

- Because the internet response is equal to the sample overlap: $\rho = 1$ and $n^{ol} = n^I$
- $\Longrightarrow Corr\left(\hat{y}_t^{\mathrm{C}}, \hat{y}_t^{\mathrm{I}}\right) = \frac{\sqrt{n^I}}{\sqrt{n^C}}$ and $Cov(e_t^C, e_t^I) = Cov\left(\hat{y}_t^{\mathrm{C}}, \hat{y}_t^{\mathrm{I}}\right) = \frac{\sqrt{n^I}}{\sqrt{n^C}}\sqrt{\widehat{V(\hat{y}_t^{\mathrm{C}})}}\sqrt{\widehat{V(\hat{y}_t^{\mathrm{I}})}}$

Structural time series model

Additional effect of CORONA:

- Sudden increase/decrease of the population parameter
- Temporal miss-specification of the model
- Visible large values of the standardized innovations
- Trend temporarily more flexible

$$\begin{pmatrix} \hat{y}_t^{\mathrm{C}} \\ \hat{y}_t^{\mathrm{I}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (L_t + S_t) + \begin{pmatrix} \\ \lambda_t \end{pmatrix} + \begin{pmatrix} e_t^{\mathrm{C}} \\ e_t^{\mathrm{I}} \end{pmatrix}$$

$$L_t = L_{t-1} + R_{t-1}$$

$$R_t = R_{t-1} + \eta_t^R$$

with

$$\eta_t^R \sim \mathcal{N}(0, k_t \sigma_{\mathrm{R}}^2) \quad \text{instead of} \quad \eta_t^R \sim \mathcal{N}(0, \sigma_{\mathrm{R}}^2)$$

Structural time series model

Estimation of the bivariate model:

- STM written in the so-called state space form
- Filtered and smoothed estimates for trend, seasonal, bias, etc. are obtained with the Kalman filter
- Kalman filter assumes that hyperparameters (i.e. variances of the disturbance terms of the trend, seasonal , etc) are known
- Hyperparameters are estimated via maximum likelihood estimation using a numerical optimization procedure
- Uncertainty of the ML estimates is ignored in the standard errors of the Kalman filter estimates
- Software: OxMetrics in combination with Ssfpack

Structural time series model

Assumption:

- Proposed method is based on a strong assumption that cannot be verified
- Observed difference between $\hat{y}_t^{\mathrm{C}}$ and $\hat{y}_t^{\mathrm{I}}$ is not affected by the lockdown
- Internet respondents before and during the lockdown are comparable

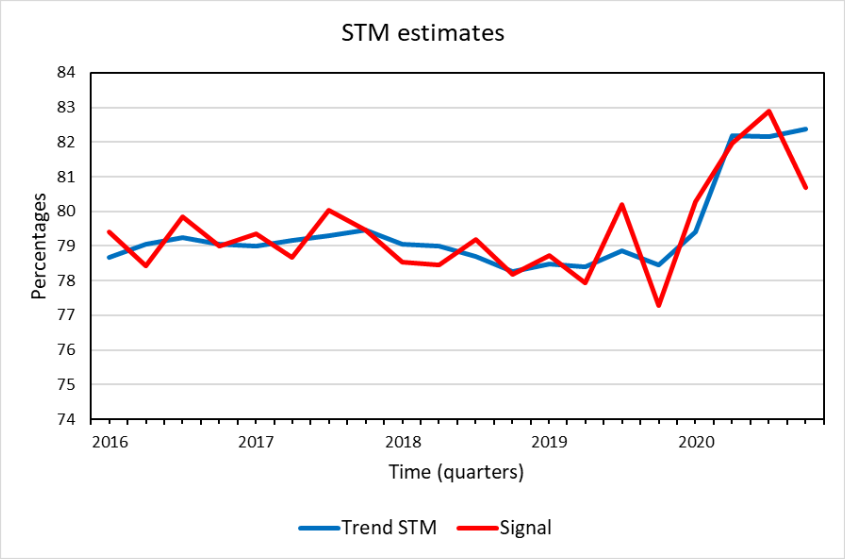
Results

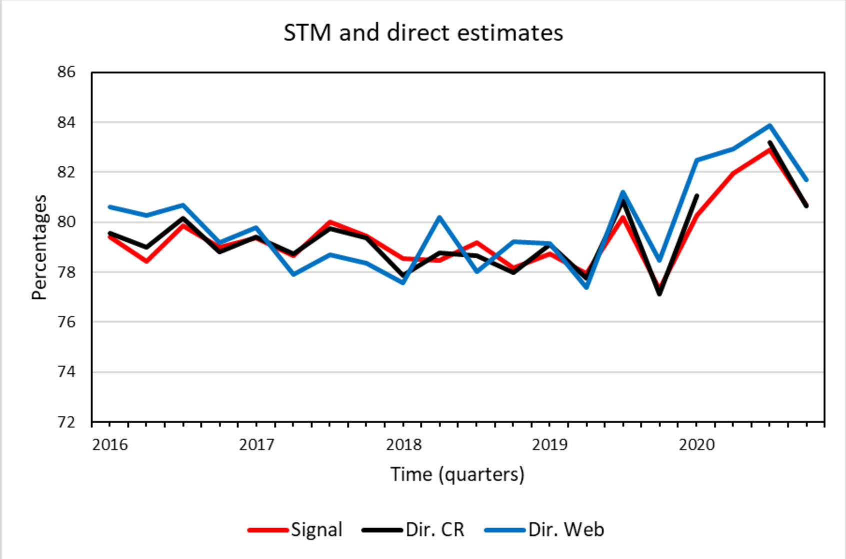
STM's are developed for 8 major indicators:

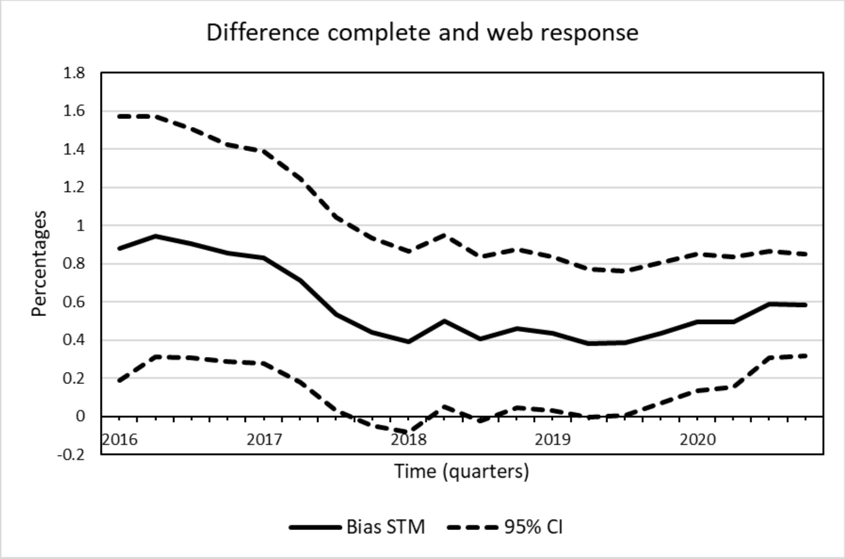
1. Perceived health (0 and older)
2. Mental unhealthy (12 and older)
3. GP visit in the past four weeks (0 and older)
4. Daily smoking (18 and older)
5. Overweight (18 and older)
6. Excessive alcohol consumption (18 and older)
7. Dental visit in the past four weeks (0 and older)
8. Specialist visit in the past four weeks (0 and older)

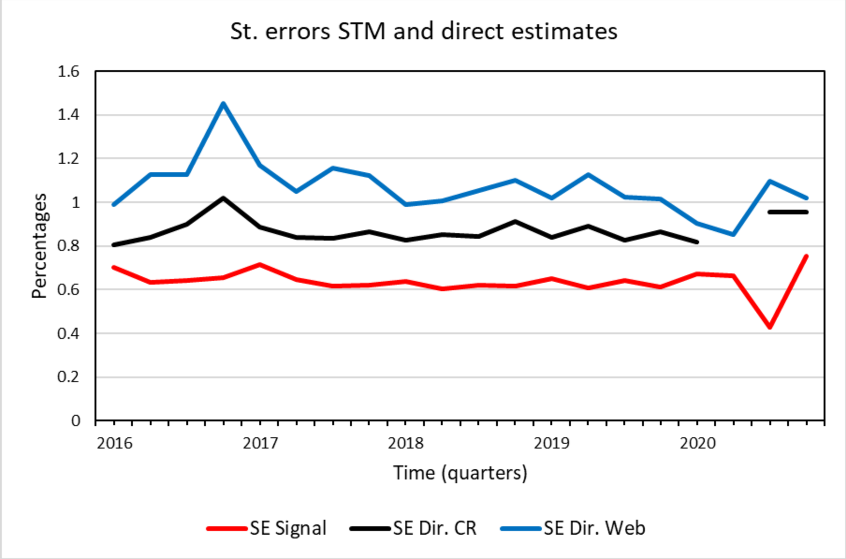
All indicators are defined as percentages

Results: Perceived health (0 and older)

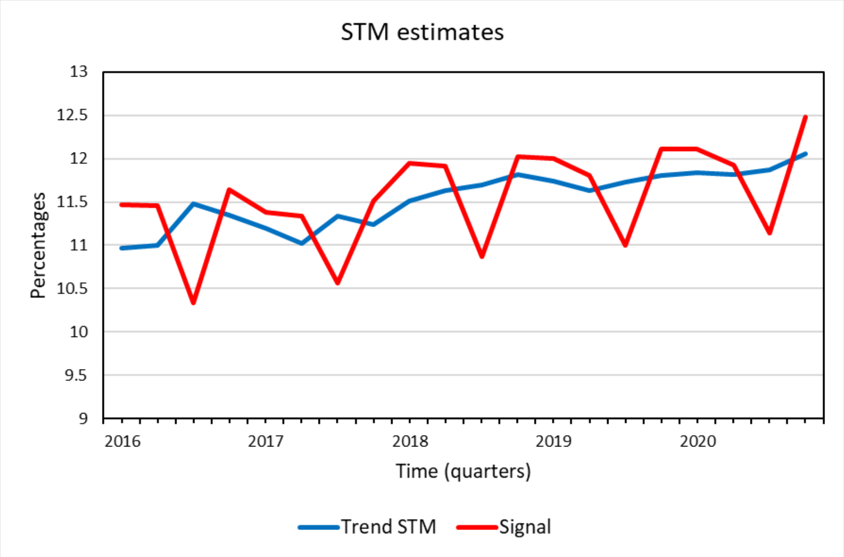


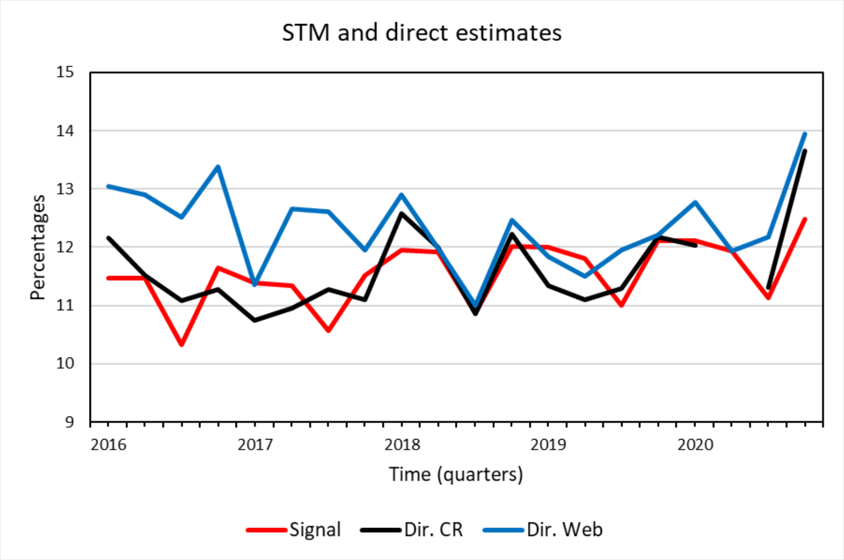


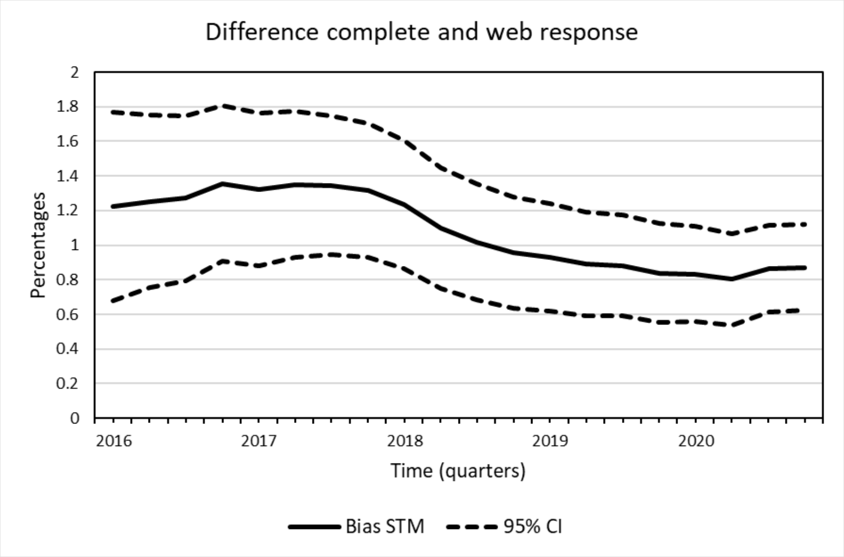


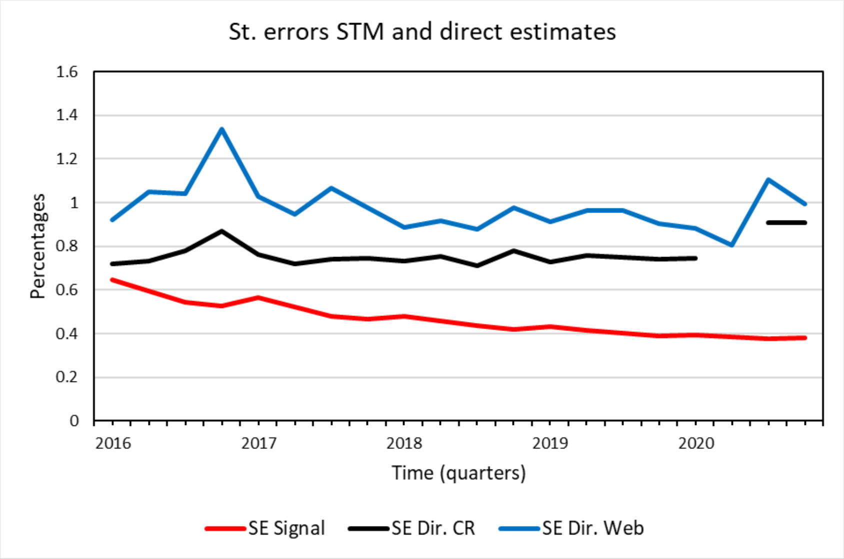


Results: Mental unhealthy (12 and older)

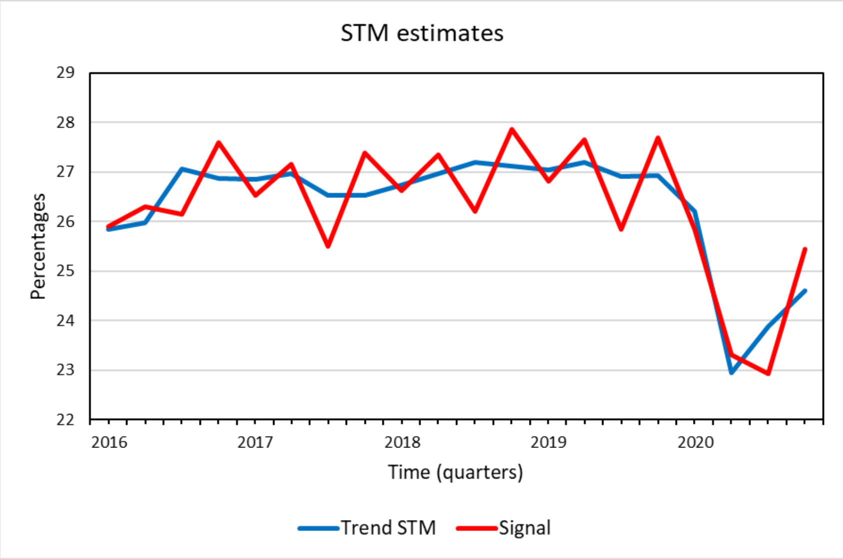


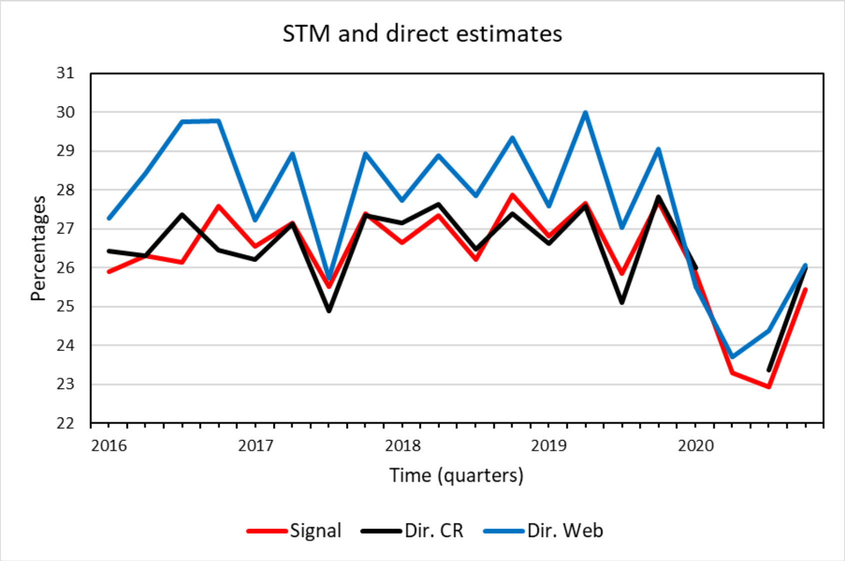


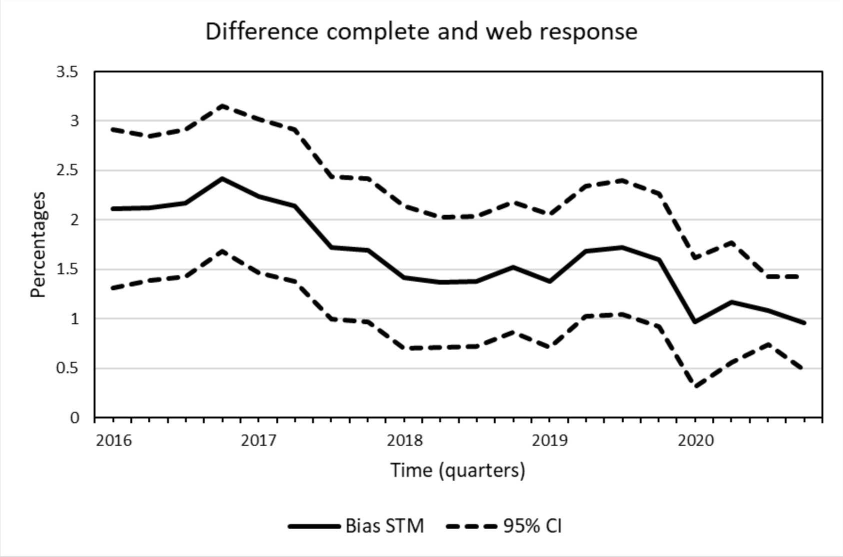


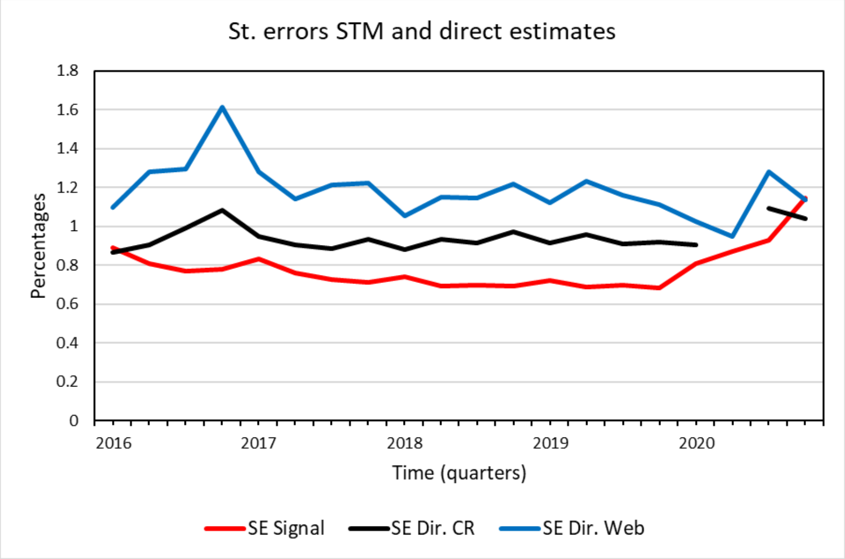


Results: GP visit in the past four weeks (0 and older)

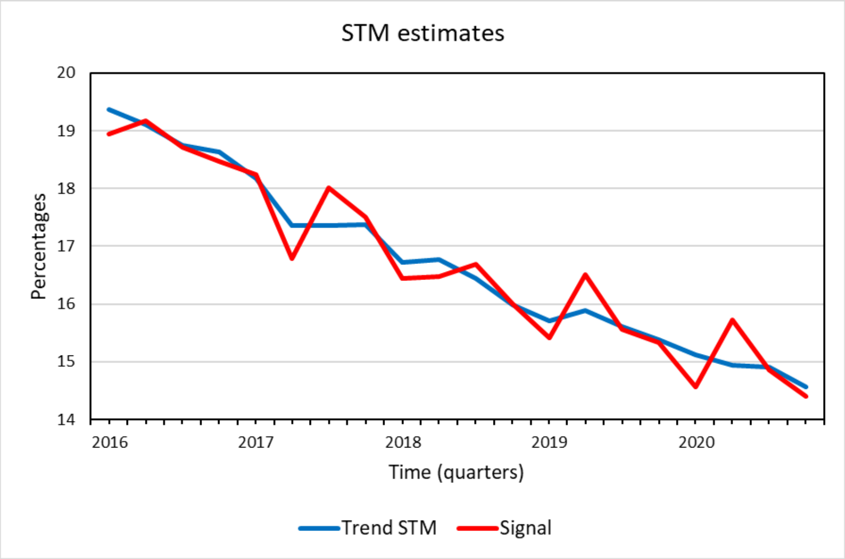


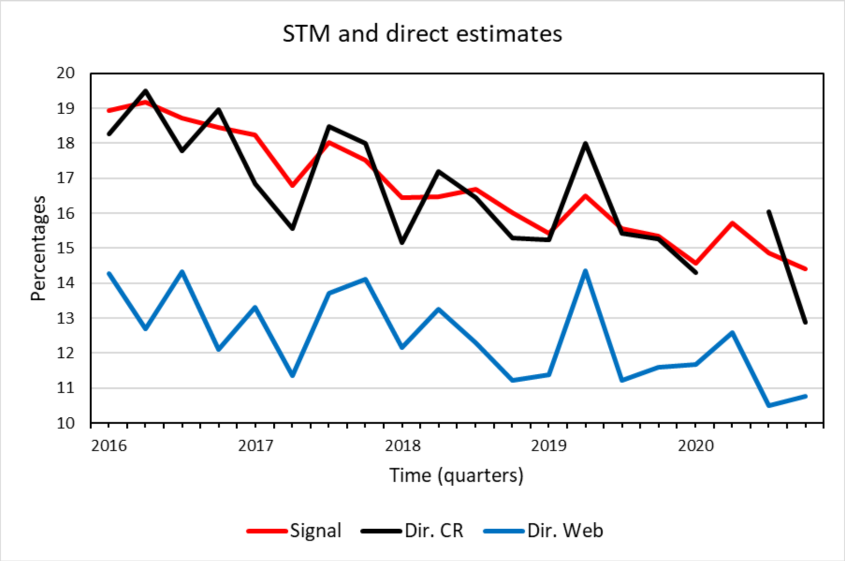


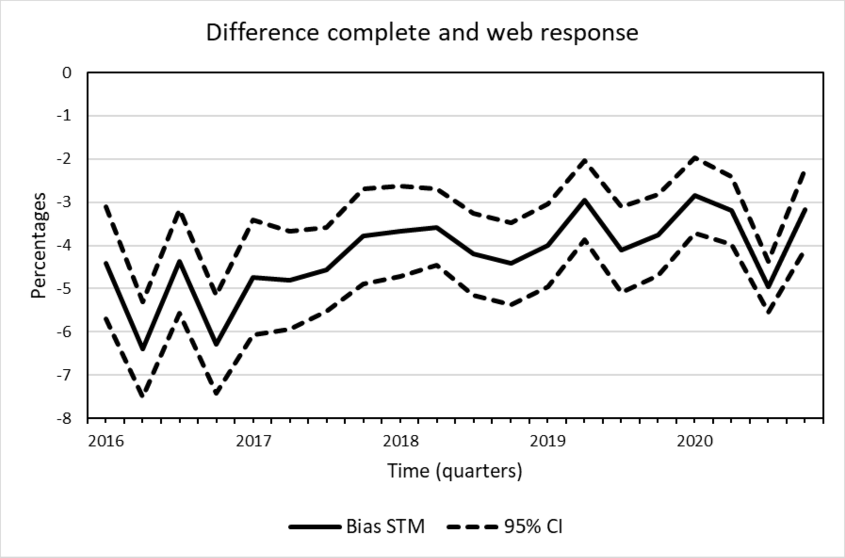


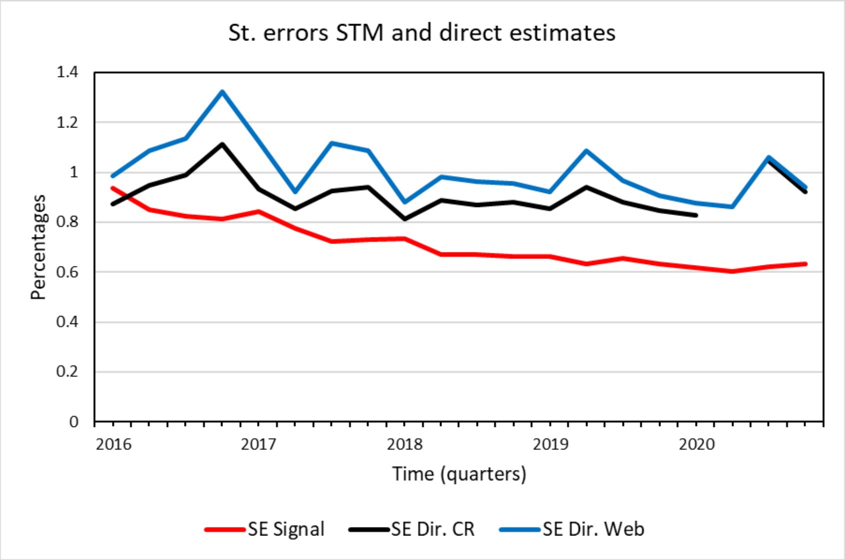


Results: Daily smoking (18 and older)

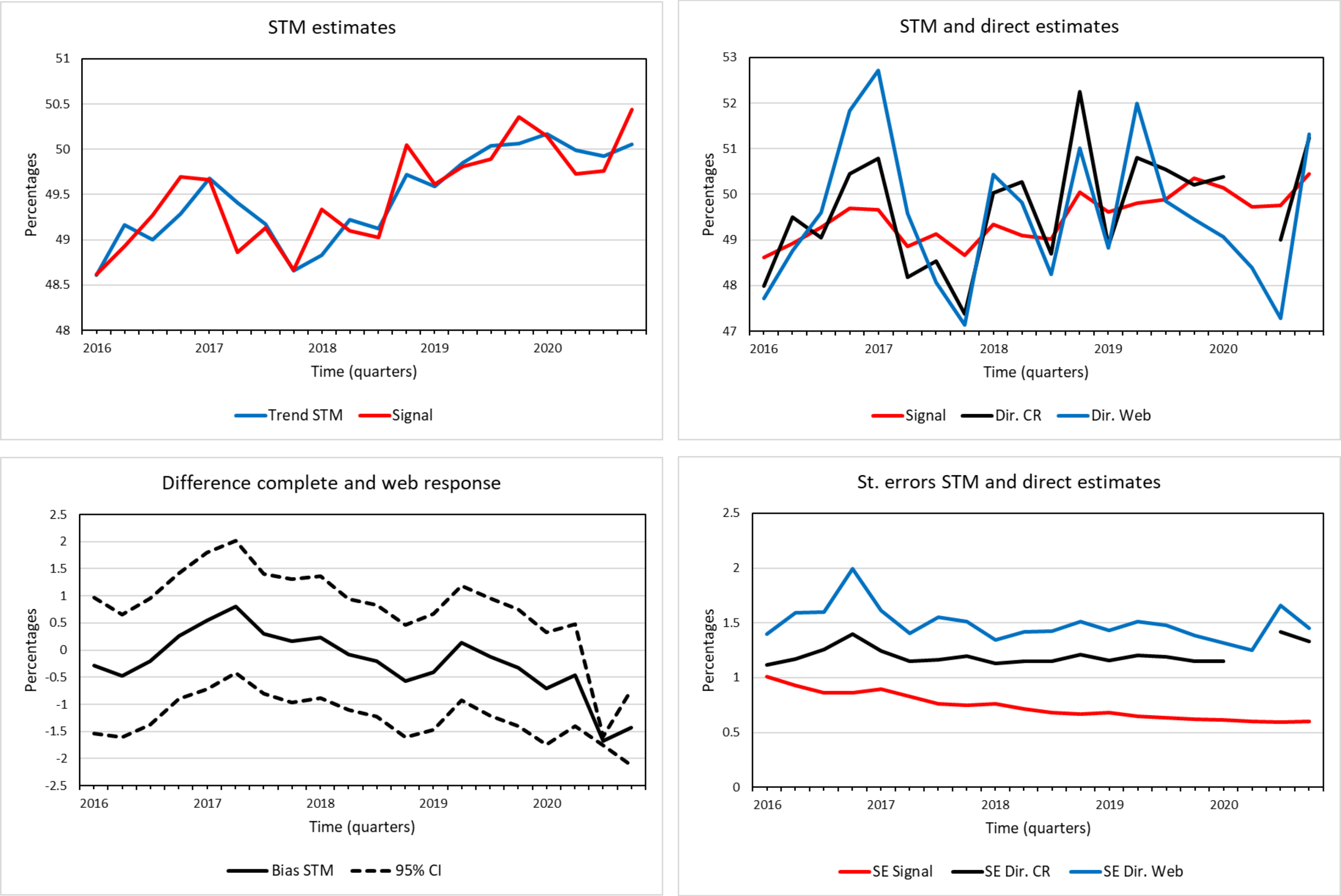




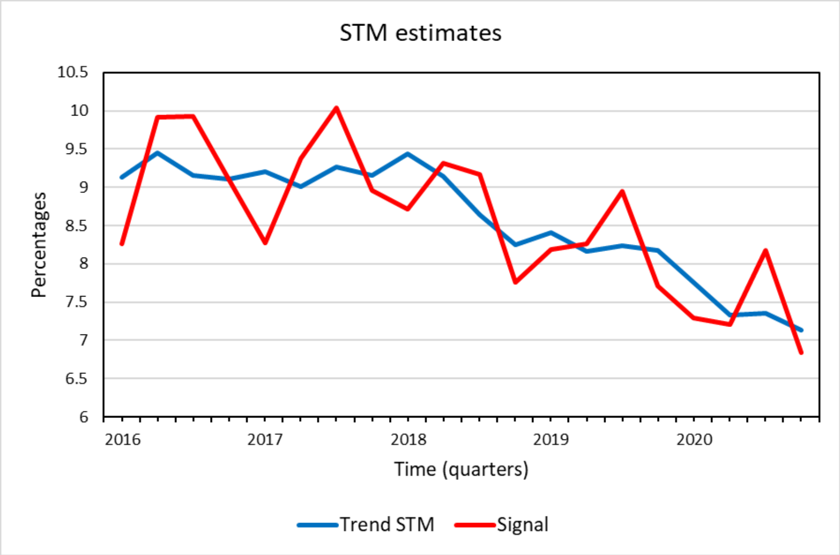


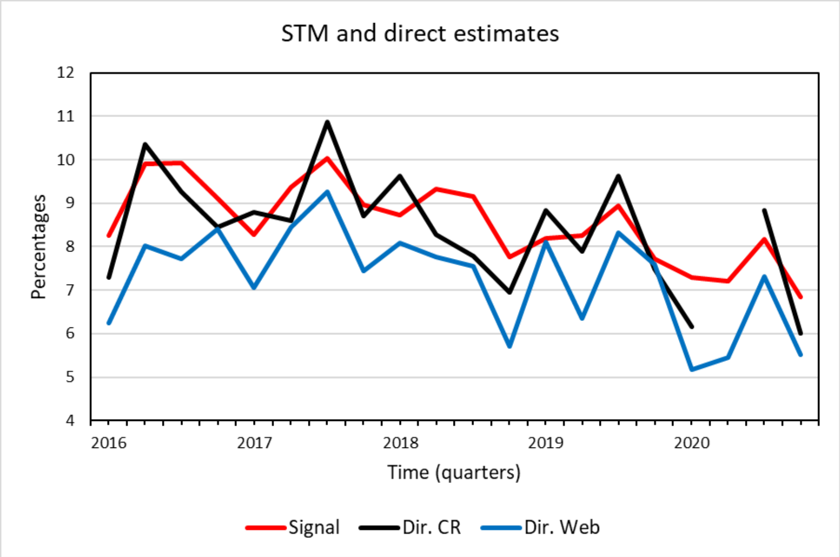


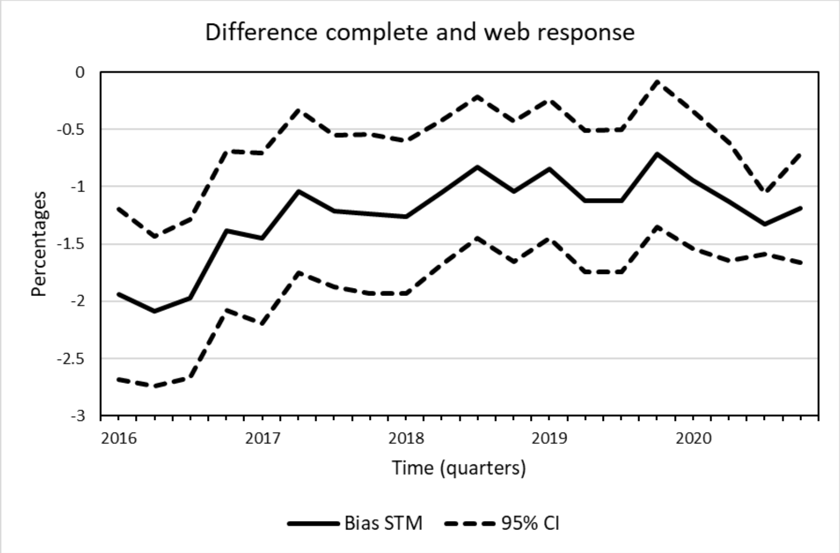
Results: Overweight (18 and older)

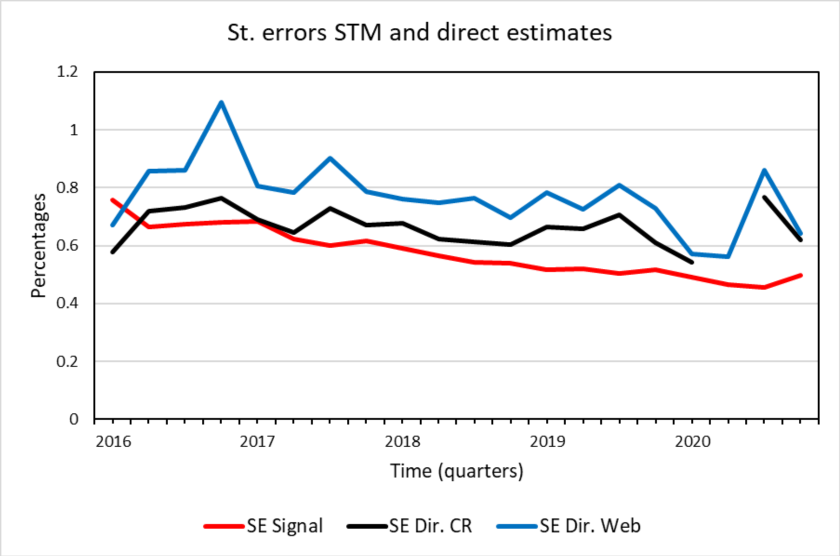


Results: Excessive alcohol consumption (18+)

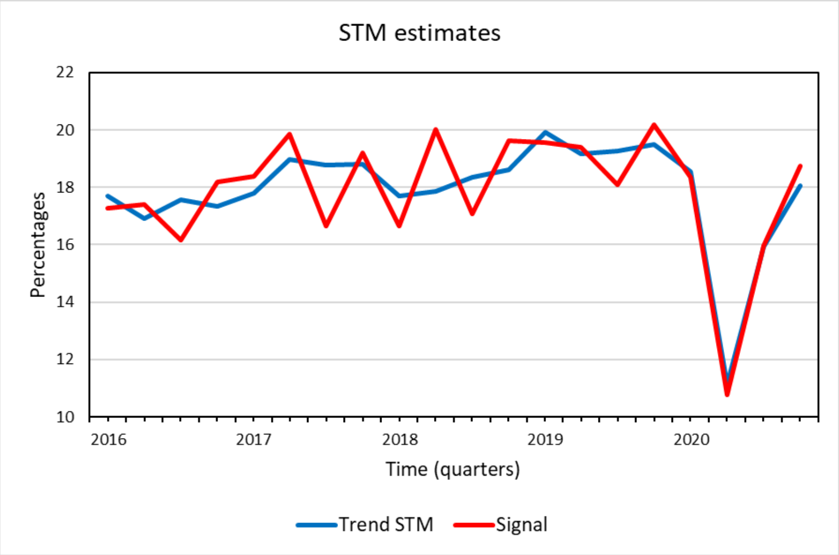


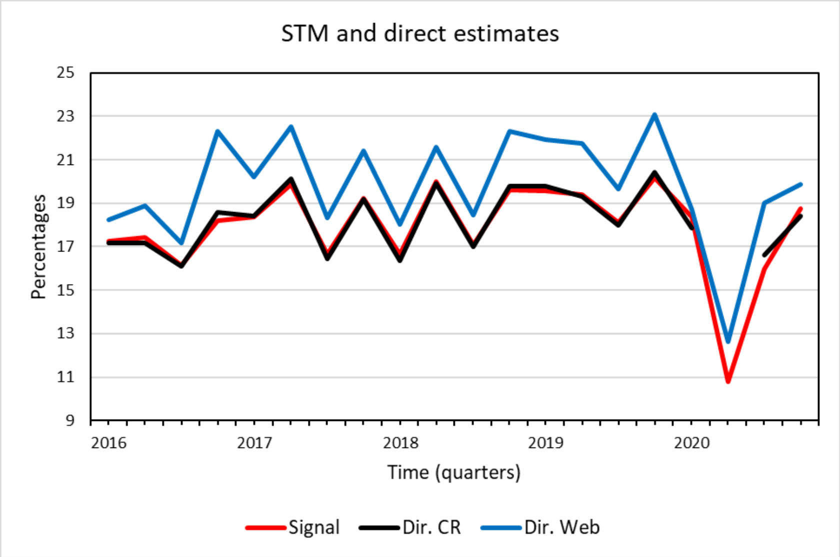


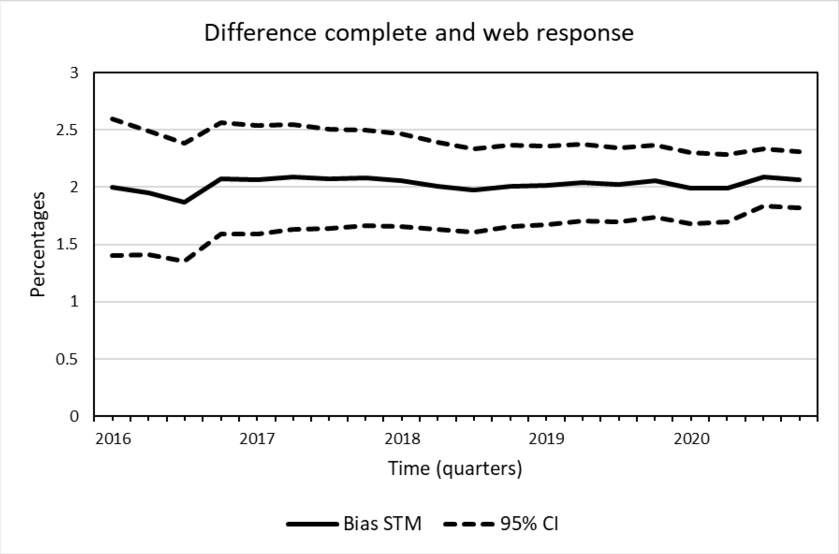


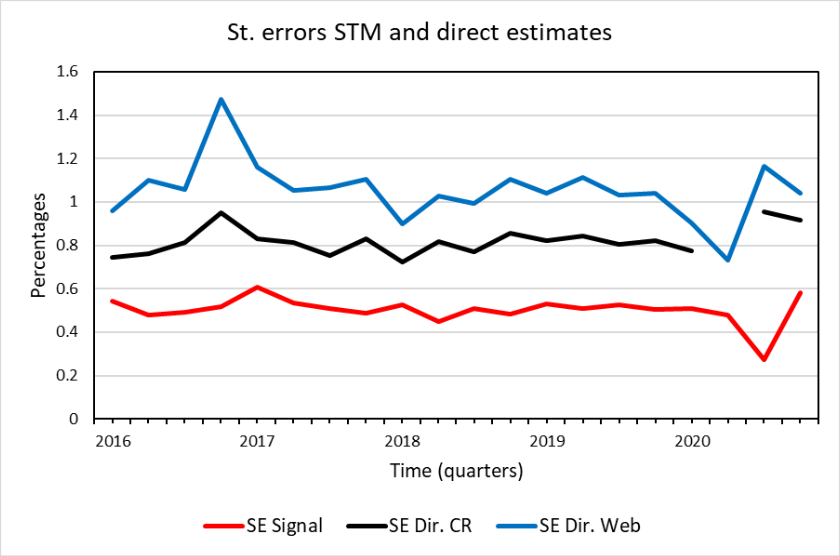


Results: Dental visit in the past four weeks (0 and older)









Results: Specialist visit in the past four weeks (0+)

Results: k-values trend

Increased flexibility trend for:

1. Perceived health (0 and older)
2. GP visit in the past four weeks (0 and older)
3. Dental visit in the past four weeks (0 and older)
4. Specialist visit in the past four weeks (0 and older)

Values for k_t are tuned such that standardized innovations have absolute values around 2.0-2.5

Results: k-values trend

	2019(1)	2019(2)	2019(3)	2019(4)	2020(1)	2020(2)	2020(3)	2020(4)
Perceived Health	1	10	100	100	100	10	1	1
GP visit	1	10	100	100	1000	1000	100	1
Dentist visit	1	1	10	100	100	10	1	1
Specialist visit	1	1	10	100	100	10	1	1

Results: annual figures

Annual figures

- Based on GREG estimator
- Weighting model is extended with quarterly STM estimates for the 8 variables
  - Numerical consistency between annual and quarterly publications
  - Correction for the loss of CAPI for more detailed breakdowns of the 8 variables
  - A good as possible correction for the loss of CAPI for other related variables for which no STMs are developed

Results: annual figures

Summary statistics correction weights

	Regular weighting	Extended weighting
Minimum weight	0.230	0.030
First quantile	0.759	0.707
Median	0.954	0.945
Third quantile	1.184	1.237
Maximum weight	2.799	3.067
Standard deviation	0.345	0.414
Numb. of negative weights	0	0

Discussion

Impact of Covid-19 on the Health Survey:

- Introduction of mode effects due to temporal loss of CAPI
- Standard annual figures lose their relevance:
  - Outdated since first figures available in Q1 2021
  - Effect of COVID-19 is less clear in annual figures
- Resulted in a strong demand for quarterly figures
- Sample size is to small to produce official quarterly figures with standard direct inference methods (GREG estimator)

Discussion

Solution:

- Model-based inference method based on bivariate STM
  - Form of SAE by borrowing strength over time
  - Nowcast for the complete response during the lockdown
- Resulted in timely quarterly official statistics
  - Trivial from academic point of view
  - Non-trivial for a national statistical institute to develop and implement a model-based inference method for the production official statistics within just one quarter

Discussion

Correction for the loss of CAPI:

- Strong relative mode-effects for almost all variables
  - Observed differences are the result of selection effects and measurement bias
- Correction based on strong assumptions
- Better than doing nothing

Other drawbacks

- Not appropriate for multipurpose surveys; requires a separate model for each variable
- Method is only possible for the most important variables