Discussion paper

Variance of the generalized regression estimator under measurement error

Jan van den Brakel and John Michiels

May 2025

Content

Summary

Official statistics published by national statistical institutes are predominantly based on sample surveys in combination with the generalized regression (GREG) estimator. The GREG estimator uses auxiliary information of which the distributions in the population are known to improve the precision of the sample estimates. This is achieved by calibrating the design weights, defined as the inverse of the inclusion probabilities of the sample design, such that the sums over the weighted auxiliary variables of the sample units are exactly equal to the distributions in the population. In most cases the population totals of these auxiliary variables are assumed to be known without error, since they are derived from registers. One exemption is two-phase sampling, where estimates for population totals based on a large first-phase sample are used in the weighting scheme of the estimates based on the second-phase sample. The variance of the GREG estimator under two-phase sampling accounts for the additional uncertainty of using sample estimates derived from the first phase in the weighting scheme of estimates obtained from the second phase.

Besides two-phase sampling there are other situations where population totals used in the weighting scheme of the GREG estimator contain uncertainty, which is ignored in the standard variance approximation of the GREG estimator. Statistics Netherlands uses a structural time series model (STM) for the production of official monthly labour force figures. Quarterly and annual figures are based on the GREG estimator. To enforce numerical consistency between monthly, quarterly and annual figures, the weighting scheme of the quarterly and annual figures contains a table that is based on the published monthly labour force figures. The additional uncertainty of using estimates for population totals, derived from a time series model, is ignored in the variances of the GREG estimator. In this paper a new variance approximation for the GREG estimator that accounts for the additional uncertainty of using population totals in the weighting scheme that are observed with measurement error is proposed. The method is illustrated with an application to the quarterly labour force figures of the Dutch Labour Force Survey.

Keywords

Labour force survey, variance estimation, weighting

This project is financed by Eurostat under grant agreement 2023-NL-LFS, nr. 101139082

1. Introduction

Based on the Dutch Labour Force Survey (DLFS) monthly, quarterly and annual figures are published on the employed and unemployed labour force in the Netherlands. Monthly publication tables are compiled with a time series model, while quarterly and annual figures are predominantly produced with the more commonly used generalized regression (GREG) estimator. To enforce consistency between monthly and quarterly figures, the weighting scheme for quarterly figures contains a component that is based on the monthly labour force estimates. The variance of the GREG estimator ignores the uncertainty of components in the weighting scheme that are observed with measurement error. In this report a variance approximation for the GREG estimator for the quarterly and annual figures is proposed that accounts for the additional uncertainty that is a result of using monthly estimates as weighting components that are observed with error.

The DLFS is based on a rotating panel design. Each month a sample of persons enters the panel. The responding persons are interviewed five times with quarterly intervals. After the fifth interview, respondents leave the panel. The inference procedure used to compile quarterly and annual figures is predominately based on the GREG estimator (Särndal et al. (1992), Ch. 6). The sample size, however, is too small to use the same design-based inference approach for calculating single month estimates . Until 2010, this problem was circumvented by publishing each month a rolling quarterly figure. In 2010 Statistics Netherlands implemented a model inference procedure for the estimation and publication of monthly labour force figures. This estimation procedure is based on a multivariate structural time series model (STM), see Van den Brakel and Krieg (2015). The model addresses five issues faced by panel surveys such as the DLFS. First, it is used as a form of small area estimation by increasing the effective sample size of the last month with sample information observed in previous months. This results in monthly estimates with a smaller standard error compared to the GREG estimator. Second, the model accounts for rotation group bias (RGB). This refers to the phenomenon that there are systematic differences in the outcomes of the subsequent waves of a rotating panel (Bailar, 1975). Under the assumption that the observations in the first wave are the most reliable, the STM benchmarks the outcomes in the follow-up waves to the level of the estimates based on the first wave. Third, the model accounts for serial correlation in the outcomes, which are a result of the partial sample overlap of the rotating panel design. Fourth, the model accounts for systematic effects that are the result of three major survey process redesigns that took place in 2010, 2012 and 2021. Systematic differences that are the result of a survey redesign are further referred to as discontinuities. Fifth and finally the model accounts for the systematic effects of the loss of face-to-face interviews during the lockdowns of the COVID-19 pandemic. For more details how the STM accounts for these problems, we further refer to Van den Brakel (2022) and Van den Brakel et al. (2022).

Quarterly figures are, as previously mentioned, predominately based on the GREG estimator. The weighting scheme of the GREG estimator is based on a linear function of social demographic auxiliary variables. The population totals of these auxiliary variables are derived from the Dutch population registers and are considered to be without error. The model-based domain estimates for the monthly employed and unemployed labour force are included as weighting terms in the GREG estimator for the quarterly and yearly releases. This enforces consistency between monthly, quarterly, and yearly labour force figures and corrects, at least partially, for the RGB, discontinuities and COVID effects in the GREG estimates of the quarterly and yearly labour force figures.

The GREG estimator treats the population totals that are used in the weighting scheme as true values that are observed without error. This assumption holds for the population totals that are derived from population registers, but not for the monthly estimates obtained with the time series model. The usual standard errors of the GREG estimator therefore ignore the additional uncertainty that is a result of including the monthly weighting terms in the GREG estimator of which the population totals are subject to error. The purpose of this project is to derive variance approximations for the GREG estimator that account for the additional uncertainty of using population totals that are estimated with the time series model for the monthly labour force figures.

The paper is organized as follows. In Section 2 the DLFS, the STM for monthly figures and the GREG estimator for the quarterly figures are described. In Section 3 an analytical expression for the variance of the GREG estimator that accounts for uncertainty in the population totals of the weighting scheme is proposed. In Section 4 the results of an application to the quarterly figures of the DLFS are presented. The paper concludes with discussion in Section 5.

2. Dutch Labour Force Survey

2.1 Survey design

The objective of the Dutch Labour Force Survey (DLFS) is to provide reliable information about the Dutch labour force. Before the redesign of 2021, each month a stratified two-stage cluster sample of addresses was drawn. Strata were formed by geographical regions. Municipalities are considered as primary and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample. All household members with an age of 15 years and older are included in the sample. Different subpopulations are oversampled to improve the accuracy of the official releases, for example, addresses where people live, who are formally registered at the employment office, and subpopulations with low response rates.

Before 2000, the DLFS was designed as a cross-sectional survey. Since October 1999, the DLFS has been conducted as a rotating panel design. Until the redesign in 2010, data in the first wave were collected by means of computer assisted personal interviewing (CAPI). Respondents were re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). During these re-interviews, a condensed questionnaire was used to establish changes in the labour market position of the respondents.

In 2010, a major redesign for the DLFS started. The main objective of this redesign was to reduce the data collection costs of this survey. This was accomplished by changing the data collection in the first wave from CAPI to a mixed data collection mode using CAPI and CATI. Households with a listed telephone number are interviewed by telephone, the remaining households are interviewed face-to-face. To make CATI data collection in the first wave feasible, the questionnaire for the first wave needed to be abridged since a telephone interview, according to data collection literature, should not take longer than 15 to 20 minutes. Therefore, parts of the questionnaire were transferred from the first to the second or the third wave.

In 2012, a second major redesign of the DLFS took place. Data collection changed to a sequential mixed-mode design that starts with Web interviewing. After three reminders the non-responding households are contacted by telephone if they have a listed telephone number. The remaining households are interviewed face-to-face. It was again necessary to change the questionnaire in all the five waves. The monthly gross sample size for the first wave averaged about 8,000 addresses commencing at the moment that the DLFS changed to a rotating panel design, i.e. October 1999, and gradually fell to about 6,500 addresses in 2012. The response rate was about 55% in the first wave and in the subsequent waves about 90% with respect to the responding households from the preceding wave. After the second redesign in 2012, the monthly sample size in the first wave increased to about 8500 households. Response rates in the first wave vary between 50% and 55%, in the second wave about 70% of the households respond with respect to the responding households from the first wave and in the third, fourth and fifth waves about 90% with respect to the responding households from the preceding wave.

The new survey design of the DLFS, that was implemented in 2021, is different in several ways. The sample design changed from a stratified two-stage cluster sample of households to a stratified two-stage cluster sample of persons. Strata are formed by the same geographical regions used under the old design. Municipalities are considered as primary and persons as secondary sampling units. The target population consists of people from 15 to 89 years old living in private households. Under the new design, samples are drawn on a weekly instead of a monthly basis. Data collection periods for several weekly samples overlap providing a continuous flow of respondents resulting in an optimized distribution of responses over the year. The rotation scheme of the panel remained unchanged. The monthly samples are still observed five times with quarterly intervals. The average monthly sample size in the first wave is about 5000 persons.

Each sampled person receives an invitation letter where she or he is asked to participate via an online questionnaire. The internet data collection runs approximately one month for every (weekly) sample with two reminders for non-respondents. After the internet data collection phase, non-respondents are approached by CATI or CAPI. In the follow-up waves, respondents are asked again to participate in the online questionnaire and non-respondents are approached by CATI if a telephone number is available (which was asked for at the end of the first wave). The same data collection process is repeated for the third, fourth and fifth wave.

2.2 Time series model for monthly estimates

Since June 2010, Statistics Netherlands uses a multivariate structural time series (STS) model for the production of monthly labour force figures. The sample observed for the jth time is henceforth shortly denoted as the jth wave. As a result of the rotation scheme, each month data are collected in five independent samples, i.e. the sample of the first wave that enters the panel for the first time, the sample of the second wave that had entered the panel three months before and that is now observed for the second time, the sample of the third wave that had entered the panel six months before and is observed for the third time, etc. Let $\hat{y}_t^{[j]}$ denote the general regression (GREG) estimator for an unknown population parameter in month t, based on the sample that is observed for the jth time. As a result, each month, five GREG estimates are observed that can be collected in a five dimensional vector, say $\hat{\boldsymbol{y}}_t = (\hat{y}_t^{[1]}, \dots, \hat{y}_t^{[5]})'$. From this, a five-dimensional time series can be constructed, which is part of the input of the following STS model:

$$\hat{\boldsymbol{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\Delta}_t^1\boldsymbol{\gamma}^1 + \boldsymbol{\Delta}_t^2\boldsymbol{\gamma}^2 + \boldsymbol{\Delta}_t^3\boldsymbol{\gamma}^3 + \boldsymbol{\delta}_t^{COV}\gamma^{COV} + \boldsymbol{\varepsilon}_t. \qquad (1)$$

This is an extension of the model proposed by Pfeffermann (1991). The components in STS model (1) can be motivated as follows. The first component θ_t denotes the unknown population parameter and $\mathbf{1}_{[5]}$ is a five dimensional column vector with each element equal to one. This component states that $\hat{\boldsymbol{y}}_t$ contains five GREG estimates for the population parameter in month t. The population parameter is modelled with a so-called basic STM, i.e.

$$\theta_t = L_t + S_t + I_t, \qquad (2)$$

with L_t a time-varying or dynamic trend model for the low frequency variation in the series of the population parameter, S_t a dynamic seasonal model for the monthly effects in the series and I_t a white noise component for the unexplained variation of the population parameter. For L_t the so-called smooth trend model and for S_t the trigonometric seasonal model are used, see Durbin and Koopman 2012, Ch. 3 for details. This first component is used as a form of small area estimation (Rao and Molina, 2015), since it uses sample information observed in previous reference periods to make more stable and accurate estimates for the monthly labour force figures.

The second component in (1), i.e. $\boldsymbol{\lambda}_t$, models the rotation group bias (RGB) induced by the rotating panel design. RGB refers to the phenomenon that there are systematic differences between the outcomes of the waves of the panel

(Bailar, 1974), which are the net result of differences in questionnaire and data collection modes applied in the different waves, panel attrition and panel effects. In this application it is assumed that the first wave is free from RGB and thus gives the most reliable estimates for θ_t, see Van den Brakel and Krieg (2009) for a motivation. The other four components contain random walks, denoted $\lambda_t^{[j]}$ ($j = 2, \ldots, 5$), and model the systematic difference between the first wave and the four follow-up waves. As a result, $\boldsymbol{\lambda}_t = (0, \lambda_t^{[2]}, \lambda_t^{[3]}, \lambda_t^{[4]}, \lambda_t^{[5]})'$. Since the RGB of the first wave equals zero, the time series model estimates for θ_t are benchmarked to the level of the GREG series in the first wave. The differences between the first wave and the follow-up waves can gradually change over time since they are modelled with random walks.

The third, fourth and fifth component model the discontinuities in the input series that are the result of three major survey redesigns that took place in 2010, 2012, and 2021 respectively. The discontinuities are modelled with level interventions, i.e. $\boldsymbol{\Delta}_t^i = diag(\delta_t^{i,[1]}, \delta_t^{i,[2]}, \delta_t^{i,[3]}, \delta_t^{i,[4]}, \delta_t^{i,[5]})$, which denotes a diagonal matrix with dummy variables $\delta_t^{i,[j]}$ that change from zero to one at the moment that the survey in wave j changes from the old to the new design during redesign $i = 1$ in 2010, $i = 2$ in 2012 and $i = 3$ in 2021. Furthermore, $\boldsymbol{\gamma}^i = (\gamma^{i,[1]}, \gamma^{i,[2]}, \gamma^{i,[3]}, \gamma^{i,[4]}, \gamma^{i,[5]})'$ are five dimensional vectors that contain estimates for the discontinuities in the five waves during redesign $i = 1$ in 2010, $i = 2$ in 2012, and $i = 3$ in 2021. For all three redesigns, estimates for the discontinuities in the first wave are obtained with a parallel run, where data under the old and new design are conducted in parallel for a period of 6 months. These direct estimates are used as an estimate for $\gamma^{i,[1]}$, for $i = 1, 2, 3$, and are treated in the Kalman filter as if these values are known a priori. This also holds for the second wave of the second redesign, i.e. $\gamma^{2,[2]}$, since there was also budget for a parallel run of 6 months in the second wave in 2012. Discontinuities for all other follow up waves are estimated with the Kalman filter. See Van den Brakel and Krieg (2016) for details.

The sixth component in (1), $\boldsymbol{\delta}_t^{COV} \gamma^{COV}$, contains a correction for the loss of CAPI respondents in the first wave during the lockdown of the corona crisis in 2020 and 2021. For this component, $\boldsymbol{\delta}_t^{COV} = (\delta_t^{cov}, 0, 0, 0, 0)'$ is a five dimensional vector that contains a level intervention for the first wave only. The indicator δ_t^{cov} is equal to one during the months of the lockdown without CAPI respondents and zero otherwise. The coefficient γ^{COV} is an approximation of the systematic difference in the first wave that arises as a result of the loss of CAPI in the first wave. An estimate for γ^{COV}, is derived by modelling the time series of the first wave based on the complete response with a series that is based on the CATI and web response only and a series of people receiving claimant counts in a multivariate STS. See Van den Brakel et al. (2022) for details.

The last component in (1), i.e. $\boldsymbol{\varepsilon}_t$, is a time series model for the survey errors that accommodate heteroscedasticity due to e.g. varying sample sizes over time and serial correlation which is a result of the partial sample overlap of the rotating panel design. The sampling errors are stacked in a five dimensional vector $\boldsymbol{\varepsilon}_t = (\varepsilon_t^{[1]}, \varepsilon_t^{[2]}, \varepsilon_t^{[3]}, \varepsilon_t^{[4]}, \varepsilon_t^{[5]})'$. To account for heteroscedasticity, due to e.g. varying sample sizes over time, the sampling errors are scaled with the standard errors of the GREG estimates of the input series, i.e. $\varepsilon_t^{[j]} = \sqrt{var(\hat{y}_t^{[j]})}\tilde{\varepsilon}_t^{[j]}$. The standard errors of the GREG estimates are estimated from the survey data. The scaled sampling error for the first wave, i.e. $\tilde{\varepsilon}_t^{[1]}$, is a normally and independently distributed error term that is not correlated with past observations, since the first wave is observed for the first time. The scaled sampling errors of the follow-up waves are modeled with an AR(1) model to accommodate serial correlation with past observations. See Van den Brakel and Krieg (2015) for details.

Model (1) can be expressed in the so-called state space representation. A state space model consists of a measurement equation and a transition equation. The measurement equation defines how the observed series depends on a set of unobserved state variables. These are the variables that define the different components in the STM (1), like the level of the trend, the variables that define the seasonal effects, the RGB etc. The transition equation defines how the state variables change over time. Besides state variables, a state space model also contain hyperparameters. These are the variances and covariances of the state disturbance terms of the transition equation and the variances of the measurement equation, which defines the dynamics of the state space model.

If the STS model is expressed in state space form, the Kalman filter is applied to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). The Kalman filter is a recursive algorithm that runs from the start of the time series and provides optimal estimates for state variables for each period t that are based on the information available up to and including period t, i.e. the time series $\hat{\boldsymbol{y}}_t$ observed from $t = 1, \dots, t$. The filtered estimates of past state vectors can be updated if new data becomes available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. In this application, interest is mainly focused on the filtered estimates, since they are based on the complete set of information that is available in the regular production process to produce a model-based estimate for month t.

To start the Kalman filter, values for the hyperparameters and starting values for the state variables for the first period are required. The hyperparameters are estimated with a maximum likelihood procedure. Starting values for the state variables are obtained with a diffuse initialization, which implies that their starting values are taken equal to zero with a large filter variance. The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman et al. (2008).

Population parameters estimated by the time series model are the unemployed labour force, employed labour force and the total labour force. These three parameters are estimated at the national level and a break down in six domains that is based on the cross classification of gender and age in three classes. Variables of interest are the trend (L_t) and the signal. The latter is defined as the trend plus the seasonal component ($L_t + S_t$). These estimates are corrected for discontinuities, i.e. the published trends are defined as $L_t + \gamma^{1,[1]} + \gamma^{2,[1]} + \gamma^{3,[1]}$

and the signals as $L_t + S_t + \gamma^{1,[1]} + \gamma^{2,[1]} + \gamma^{3,[1]}$. Recall that $\gamma^{i,[1]}$ is the discontinuity for the first wave of the i-th redesign. Since the LFS estimates are benchmarked to the level of the first wave, by the assumption that the RGB for the first wave is equal to zero, the trend and signal estimates are corrected for the discontinuities of the first wave. In this way the entire series for the trend and signal are at the measurement level of the most recent survey design.

The employed, unemployed and total labour force at the national level and its breakdown in six domains define a set of 21 parameters. Model (1) is applied to each of these 21 parameters separately. As a result the monthly publication tables are not numerically consistent. Therefore a Lagrange function is applied to the filtered estimates to enforce that the sum over employed and unemployed labour force is equal to the total labour force for the national level and the six domains and that the sum over the six domains equals the national total. See Van den Brakel and Krieg (2015) for details.

2.3 GREG estimator for quarterly estimates

The quarterly estimates of the DLFS are based on the GREG estimator. The purpose of the GREG estimator is to estimate population totals which are defined as the sum over the values of a variable of interest, say y of all elements of the target population U:

$$t_y = \sum_{i \in U} y_i. \tag{3}$$

To this end a probability sample s of size n is drawn from the target population where all elements in the population have non-zero first and second order inclusion probabilities π_i and π_{ij}.

The GREG estimator for t_y is derived from a linear regression model that defines the relation between the target variable and a set of auxiliary variables;

$$y_i = \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i \tag{4}$$

with y_i the response of the target variable of sampling unit i, $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,p})'$ a p dimensional vector containing the auxiliary variables of sampling unit i, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ a p dimensional vector containing the regression coefficients and ε_i a residual. It is assumed that $E_\xi(\varepsilon_i) = 0$ and $Var_\xi(\varepsilon_i) = \nu_i \sigma^2$, where $E_\xi(.)$ and $Var_\xi(.)$ denote the expectation and variance with respect to model (4) and ν_i a scaling factor that is known for each sampling unit.

In Särndal et al. (1992), Ch. 6 it is shown that the GREG estimator for t_y can be expressed as

$$\hat{t}_y^R = \hat{t}_y^\pi + \hat{\boldsymbol{\beta}}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^\pi), \tag{5}$$

with $\hat{t}_y^\pi$ the Narain-Horvitz-Thompson estimator for t_y, i.e.

$$\hat{t}_y^{\pi} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

$\boldsymbol{t}_x$ a p dimensional vector containing the population totals of the auxiliary variables, i.e.

$$\boldsymbol{t}_x = \sum_{i \in U} \boldsymbol{x}_i,$$

and $\hat{\boldsymbol{t}}_x^{\pi}$ the Narain-HorvitzThompson estimator for $\boldsymbol{t}_x$, which is defined as

$$\hat{\boldsymbol{t}}_x^{\pi} = \sum_{i \in s} \frac{x_i}{\pi_i}.$$

An estimator for the regression coefficients in (5) is defined as:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \frac{x_i x_i'}{\pi_i v_i}\right)^{-1} \sum_{i \in s} \frac{x_i y_i}{\pi_i v_i}.$$

The variance of the GREG estimator is obtained by approximating (5) with a first order Taylor linearization (Särndal et al. (1992), Ch. 6):

$$\hat{t}_y^{R} = \hat{t}_y^{\pi} + \boldsymbol{\beta}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{\pi}) + Rest \doteq \hat{t}_y^{\pi} + \boldsymbol{\beta}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{\pi}) \equiv \hat{t}_y^{R0}. \tag{6}$$

The variance of $\hat{t}_y^{R0}$ is used as an approximation of the variance of $\hat{t}_y^{R}$ and is defined as:

$$V\left(\hat{t}_y^{R0}\right) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}, \tag{7}$$

with $e_i = y_i - \boldsymbol{\beta}' \boldsymbol{x}_i$. An estimator for the variance is defined as

$$\hat{V}\left(\hat{t}_y^{R0}\right) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{\hat{e}_i}{\pi_i} \frac{\hat{e}_j}{\pi_j}, \tag{8}$$

with $\hat{e}_i = y_i - \widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i$.

The weighting model for the quarterly DLFS figures is defined as:

H_M_V*Afl_Herkomst20 + Afl_Grotegemeentes*Afl_Lft2Gesl + H_M_V*Afl_Lft43 + Afl_Lft7GeslCapti + Afl_CWI_Duur + Afl_CWI_Prov + Afl_Inkomen + Afl_LoonCat + Afl_TypeHH + Afl_X3ArbPos1 + Afl_X3ArbPos2 + Afl_X3ArbPos3 + Afl_X3ArbPos4 + Afl_X3ArbPos5 + Afl_X3ArbPos6 + Afl_X3OplNivo1 + Afl_X3OplNivo2 + Afl_X3OplNivo3 + Afl_X3OplNivo4 + Afl_X3OplNivo5 + Afl_Lft4GeslInt, (9)

with:

- H_M_V : gender in 2 classes (men, women).
- Afl_Herkomst20 : ethnicity in combination with age (15-64 years and other)
- Afl_Lft2Gesl : layout in three classes (<15 or >64, 15-64 men, 15-64 women).
- Afl_Grotegemeentes : regional classification (197 regions).

- Afl_Lft43 : Age in 43 classes (0-14, separate years in 15-30, four 5-year classes up to 49, separate years, in 50-69, and classes 70-74, 75-125).
- Afl_Lft7GeslCapti: 5 age groups (15-24, 25-34, 35-44, 45-54, 55-64) times mode (CAPI, CATI) times sex (men, women), plus the two age groups <14 and >=65.
- Afl_CWI_Duur : layout in 5 classes: not enlisted at labour office, the other four classes refer to registered labour status of enlisted persons: employed, not employed (< 1 year), not employed (>1 year, <4 years) and not employed (>= 4 years).
- Afl_CWI_Prov : layout in 13 classes (not enlisted, enlisted per province).
- Afl_Inkomen : layout in 6 net income classes (yearly in Euro's) (0-3000, 3000-9999, 10000-14999, 15000-19999, 20000-29999, >=30000).
- Afl_LoonCat : combination of type of income (benefits, wages) and region (twelve provinces plus four largest cities), plus a category unknown.
- Afl_TypeHH : layout in 3 classes (single household, household with children, other).
- Afl_ X3ArbPos : 6 continuous auxiliary variables for labour position based on a composite estimator. The six variables correspond to groups within the age group 15-64: employee with open term contract, employee without open term contract, self-employed, unemployed, not in the labor force for/without a specific reason.
- Afl_X3OplNivo : 5 continuous auxiliary variables for educational level based on a composite estimator. The five variables correspond to: 1 = Basis onderwijs, 2 = VMBO, MBO 1, AVO onderbouw, 3 = HAVO, VWO, MBO, 4 = HBO, WO Bachelor, 5 = WO Master, Doctor.
- Afl_Lft4GeslInt : crossing of age (15-24, 25-44, 45-74) x labour position (employed, unemployed not in labour force) x sex, plus one additional category (<15 or >74) derived from the monthly estimates obtained with the STS model.

Afl_ X3ArbPos and Afl_X3OplNivo are based on a composite estimator for rotating panel designs, see Fuller and Rao (2001) for details.

3. Variance of the GREG-estimator under measurement error

As explained before, the GREG estimator in (5) assumes that the population totals $\boldsymbol{t}_x$ are observed without error. For the components in (9), this doesn't hold for the last component Afl_Lft4GeslInt. The population totals of this component are the three-months averages of the monthly estimates obtained with the time series model from Section 2.2. The variance approximations (7) and (8) ignore the additional uncertainty that arises in the GREG estimates if the weighting scheme (9) contains components that are estimated with a time series model from a series of repeated monthly survey estimates. An analytic approximation for the variance of the GREG estimator that accounts for the additional uncertainty of incorporating components in the weighting scheme that are subject to measurement error is obtained as follows.

The components Afl_ X3ArbPos and Afl_X3OplNivo are also observed with error, since they are based on a composite estimator that combines data from previous periods. Since standard errors for these components are not available, they are removed from the weighting scheme for variance estimation. Once these standard errors are available, they can be incorporated into the GREG estimator's variance approximation in a manner similar to that proposed for time series estimates in this paper.

In a first step the correction term of the regression estimator is split in a term for which the true population totals are known, say $\boldsymbol{t}_a$ (where subscript a stands for administration), and a term for which the true population totals are estimated, say $\tilde{\boldsymbol{t}}_m$ (where subscript m stands for model estimate). This results in:

$$\hat{t}_y^R = \hat{t}_y^\pi + \widehat{\boldsymbol{\beta}}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^\pi) = \hat{t}_y^\pi + \widehat{\boldsymbol{\beta}}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \widehat{\boldsymbol{\beta}}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi).$$

Here $\hat{\boldsymbol{t}}_a^\pi$ and $\hat{\boldsymbol{t}}_m^\pi$ are the Narain-Horvitz-Thompson estimators for $\boldsymbol{t}_a$ and $\boldsymbol{t}_m$ and $\widehat{\boldsymbol{\beta}}_a$ and $\widehat{\boldsymbol{\beta}}_m$ the corresponding estimates for the regression coefficients. With a first order Taylor approximation it follows that:

$$\hat{t}_y^R \doteq \hat{t}_y^\pi + \boldsymbol{\beta}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi) \equiv \hat{t}_y^{R0}.$$

An expression for the variance of $\hat{t}_y^{R0}$ must account for two sources of variation; sampling error of the sample design of the LFS and the measurement error of the time series model. This is achieved by conditioning on the measurement error of the time series models using the following decomposition:

$$Var(\hat{t}_y^{R0}) = E_m Var_s(\hat{t}_y^{R0}|m) + Var_m E_s(\hat{t}_y^{R0}|m), \qquad (10)$$

where E_m and Var_m denote the expectation and variance with respect to the time series model and E_s and Var_s the expectation and variance with respect to the sample design. For the first term in (10) it follows that the variance of the regression estimator, conditionally on the time series model is equal to the variance of the regression estimator treating the population totals obtained with the time series model as fixed known values, i.e.:

$$E_m Var_s(\hat{t}_y^{R0}|m) = E_m Var_s(\hat{t}_y^{R0}) = Var_s(\hat{t}_y^{R0}) \, .$$

For the second term in (10) it follows that:

$$Var_m E_s(\hat{t}_y^{R0}|m) = Var_m E_s(\hat{t}_y^\pi + \boldsymbol{\beta}'_a(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^\pi) + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^\pi)|m).$$

Taking the expectation with respect to the sample design and using $E_s \hat{t}_q^\pi = t_q$, (for $q = y, a, m$) implies that

$$Var_m E_s(\hat{t}_y^{R0}|m) = Var_m(t_y + \boldsymbol{\beta}'_m(\tilde{\boldsymbol{t}}_m - \boldsymbol{t}_m)).$$

Since t_y and $\boldsymbol{t}_m$ are finite population totals, which are constants with respect to the time series model it follows that

$$Var_m E_s(\hat{t}_y^{R0}|m) = \boldsymbol{\beta}'_m Var_m(\tilde{\boldsymbol{t}}_m)\boldsymbol{\beta}_m, \tag{11}$$

with $Var_m(\tilde{\boldsymbol{t}}_m)$ a covariance matrix containing the (co)variances of the time series model estimates on the diagonal, which are available from the software used to produce the monthly labour force figures. In this application $Var_m(\tilde{\boldsymbol{t}}_m)$ is a diagonal matrix, since the time series estimates $\tilde{\boldsymbol{t}}_m$ are obtained by applying STM (1) to each category of $\boldsymbol{t}_m$ separately. Note that in the case that the target variable t_y is included in the table defined by $\boldsymbol{t}_m$, the target variable y will be regressed on itself. In this case the corresponding regression coefficient equals one and all other regression coefficients are all equal to zero. In that case $t_y - \boldsymbol{\beta}'_m\boldsymbol{t}_m = t_y - t_y = 0$. The last paragraph of this section further elaborates on this special case and a proof is provided in Appendix B.

As a result we have the following variance approximation for the GREG estimator

$$Var(\hat{t}_y^{R0}) = Var_s(\hat{t}_y^{R0}) + \boldsymbol{\beta}'_m Var_m(\tilde{\boldsymbol{t}}_m)\boldsymbol{\beta}_m\ , \tag{12}$$

with $Var_s(\hat{t}_y^{R0})$ defined in (7) that can be estimated with (8).

To motivate the variance approximation (12), consider the situation were the GREG estimator is used to estimate the table from the weighting scheme that corresponds to the components that are estimated with the time series model, i.e. $\boldsymbol{t}_m$. Let k denote the number of categories of $\boldsymbol{t}_m$ and l the number of categories of $\boldsymbol{t}_a$, i.e. the components of the weighting scheme for which the population totals are known. If the GREG estimator is applied to estimate table $\boldsymbol{t}_m$, then $\hat{t}_y^{R0}$ in (6) becomes a k dimensional vector, say $\hat{\boldsymbol{t}}_m^{R0}$. It is shown in the Appendix that in this case the regression coefficients in (4) form a $(k+l) \times k$ matrix equal to

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{I}_{[k\times k]} \\ \boldsymbol{O}_{[l\times k]} \end{bmatrix}, \tag{13}$$

where it is assumed, without loss of generality, that the weighting variables are ordered such that the ones corresponding to Afl_Lft4Geslint come first. Furthermore, $\boldsymbol{I}_{[k\times k]}$ denotes the k dimensional identity matrix and $\boldsymbol{O}_{[l\times k]}$ an $l \times k$ matrix with each element equal to zero. This implies that each regression estimator for the totals in $\hat{\boldsymbol{t}}_m^{R0}$ has one regression coefficient that is equal to one for the auxiliary variable in the weighting scheme that corresponds to the target variable and all other $(k + l - 1)$ regression coefficients are equal to zero. This implies that

$$\hat{\boldsymbol{t}}_m^{R0} = \hat{\boldsymbol{t}}_m^{\pi} + \boldsymbol{O}_{[k\times l]}(\boldsymbol{t}_a - \hat{\boldsymbol{t}}_a^{\pi}) + \boldsymbol{I}_{[k\times k]}(\tilde{\boldsymbol{t}}_m - \hat{\boldsymbol{t}}_m^{\pi}) = \tilde{\boldsymbol{t}}_m.$$

This also implies for the residual in (6) and (8) that $e_i = y_i - \boldsymbol{\beta}'\boldsymbol{x}_i = x_{i,k'} - \boldsymbol{\beta}'_{k'}\boldsymbol{x}_i = 0$ for all $i \in s$, where $\boldsymbol{\beta}_{k'}$ denotes the k' -th column of $\mathbf{B}$ with the k' -th element equal to one and the remaining $(k + l - 1)$ elements equal to zero. As a result it follows that the variance of the GREG estimator based on (8) would be equal to zero, which is undesirable and incorrect, since $\hat{\boldsymbol{t}}_m^{R0}$ is exactly equal to $\tilde{\boldsymbol{t}}_m$ which is an estimate based on a time series model and subject to uncertainty. Because of (13) the variance of the GREG estimator based on (12) is exactly equal to $Var_m(\tilde{\boldsymbol{t}}_m)$ and thus correctly corresponds to the uncertainty of the time series model estimates $\tilde{\boldsymbol{t}}_m$.

4. Results

4.1 Evaluation measures

In this subsection evaluation measures are defined to compare the point estimates and the standard errors of the GREG estimators. Let $\hat{t}_y^{R(+m)}$ denote the GREG estimate with the monthly publication table Afl_Lft4GeslInt in the weighting scheme and $\hat{t}_y^{R(-m)}$ the GREG estimate without the monthly publication table Afl_Lft4GeslInt in the weighting scheme. To evaluate the effect on the point estimates we calculate the following measures:

- Mean point estimates:

$$AE = \frac{1}{T}\sum_t^T \hat{t}_y^{R(x)}, x = [+m, -m].$$

- Relative Difference (RD):

$$RD = \frac{1}{T}\sum_t^T \frac{\hat{t}_y^{R(+m)} - \hat{t}_y^{R(-m)}}{\hat{t}_y^{R(-m)}} \times 100\%.$$

- Relative Absolute Difference (RAD):

$$RAD = \frac{1}{T}\sum_t^T \frac{\left|\hat{t}_y^{R(+m)} - \hat{t}_y^{R(-m)}\right|}{\hat{t}_y^{R(-m)}} \times 100\%.$$

To evaluate the effect on the standard errors we calculate the following measures
- Average standard errors (ASE):

$$ASE = \frac{1}{T}\sum_t^T se\left(\hat{t}_y^{R(x)}\right), x = [+m, -m, me].$$

It is understood that me refers to the standard errors of the GREG estimator that accounts for the uncertainty of the monthly publication table in the weighting scheme.

- Relative Difference standard errors (RDSE):

$$RDSE(x1, x2) = \frac{1}{T}\sum_t^T \frac{se\left(\hat{t}_y^{R(x1)}\right) - se\left(\hat{t}_y^{R(x2)}\right)}{se\left(\hat{t}_y^{R(x2)}\right)} \times 100\%,\ x1, x2 = [+m, -m, me]\ .$$

4.2 The effect of weighting quarterly figures to monthly publication tables

As a first step the GREG estimates and the standard errors based on weighting scheme (9) with and without the component that contains the monthly labour force figures Afl_Lft4GeslInt are compared using variance approximation (8). In this way it is investigated to what extent the extension of the weighting model with the monthly publications affects the point estimates and decreases or even increases the variance of the quarterly figures. If the variance of the GREG estimator is larger without Afl_Lft4GeslInt in the weighting scheme, then this variance could serve as a pragmatic and conservative approximation for the variance of the quarterly figures. This will be done for the 12 quarters of 2017, 2018 and 2019.

The effect on point estimates and standard errors of including the monthly publication table in the weighting scheme depends on the type of target variable. For the variable 'educational attainment' (table 4.2.1) the impact is much smaller than for variable 'relation to the labour market' (table 4.2.2). The relative absolute differences (RAD) for all of the categories of 'educational level' are below 1 percent, and the impact on the standard errors seems negligible. For most of the educational levels the effect of including the monthly publication table is a slight increase (less than 0.7 percent) in standard error.

In contrast to the results in table 4.2.1. the variable 'relation to the labour market' shows a more pronounced impact of including the monthly publication table (see table 4.2.2). The relative (absolute) differences are largest for the point estimates of 'not working: education'[1] and 'not available, but has searched'[2]: they are respectively 9.1 and 7.3 percent. For the two categories 'employed' and 'unemployed', that are also part of the monthly publication table, the impact on the point estimates is relatively small. However, the impact on the standard errors is very large: the RDSE values for these categories amount to 100%. Because the categories are part of the weighting scheme the point estimates correspond to the weighting margins of the monthly publication table, whose values are assumed to be known without error from the perspective of the GREG-weighting approach. Therefore the standard errors are zero when the monthly publication table is implemented. Other categories with a large reduction in standard error are 'not working: education' and 'not working: retirement'. For some of the other

[1] Respondents are not available for work and have not recently searched fopr a job because they receive education.

[2] Respondents are not available for work but have recently searched for a job.

categories the extra weighting term leads to an increase in the estimated standard errors; the most notable category in this respect is ‘not available, but has searched’ where the relative increase in standard error is 4.5 percent.

Table 4.2.1 effect on point measures and standard errors for educational attainment

Category	AE(+m)	AE(-m)	RD	RAD	ASE(+m)	ASE(-m)	RDSE (+m,-m)
	Counts	counts	%	%	counts	counts	%
Primary education	1445919	1440859	0,4	0,4	12834	12749	0,7
Pre-secondary vocational education, secondary vocational education level 1	1758372	1754767	0,2	0,2	18401	18343	0,3
General education undergrad	1222520	1222129	0	0,1	16105	16062	0,3
Secondary vocational education level 2 and 3	2118997	2120048	-0,1	0,1	20905	20892	0,1
Secondary vocational education level 4	1889155	1893420	-0,2	0,2	19028	19036	0
General education	1306684	1304098	0,2	0,2	16377	16348	0,2
Higher Vocational Education, University Bachelor	2751881	2755440	-0,1	0,1	14359	14360	0
University Master, PhD	1532420	1535789	-0,2	0,2	10249	10298	-0,5
Unknown	2965746	2965145	0	0	7778	7748	0,4

Table 4.2.2 effect on point measures and standard errors for relation to the labour market

Category	AE(+m)	AE(-m)	RD	RAD	ASE(+m)	ASE(-m)	RDSE (+m,-m)
	counts	counts	%	%	counts	counts	%
Employed	8768846	8897074	-1,4	1,4	0	16568	-100
Unemployed	367392	361486	1,7	2,5	0	8480	-100
Discouraged	73290	70964	3,2	3,2	4706	4586	2,6
No search, other reasons	164637	155059	6,2	6,2	6623	6407	3,4
Not available, but has searched	152885	142442	7,3	7,3	5850	5597	4,5
Wants to work	205787	195523	5,2	5,2	7338	7155	2,6
Not working: household care	217006	211615	2,6	2,6	6875	6969	-1,3
Not working: education	454345	416703	9,1	9,1	7522	9255	-18,7
Not working: retirement	1538607	1514953	1,6	1,6	12343	13980	-11,7
Not working: disability	765337	750500	2	2	12173	12477	-2,4
Not working: other reasons	232533	224346	3,6	3,6	7829	7680	1,9

Point estimates for the target variables ‘educational level’ and ‘relation to the labour market’ for the GREG estimator with and without the monthly publication tables in the weighting model are presented in Figures 4.2.1 and 4.2.2 respectively. The quarterly estimates relate to the period 2017 – 2019.

Figure 4.2.1: Quarterly GREG estimates with and without the STM-margin in the weighting scheme for the target variable ‘educational level’. The following abbreviations are used: sve: secondary vocational training, ge: general education, hve: higher vocational training.

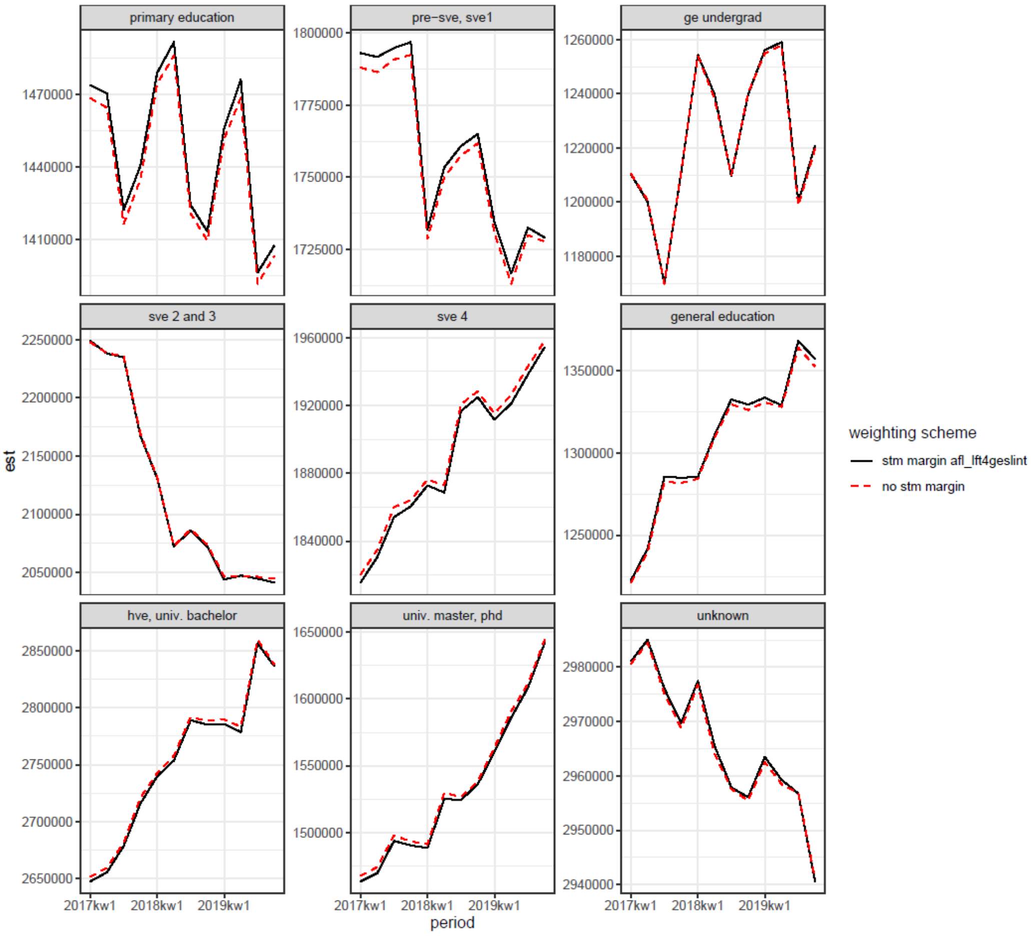
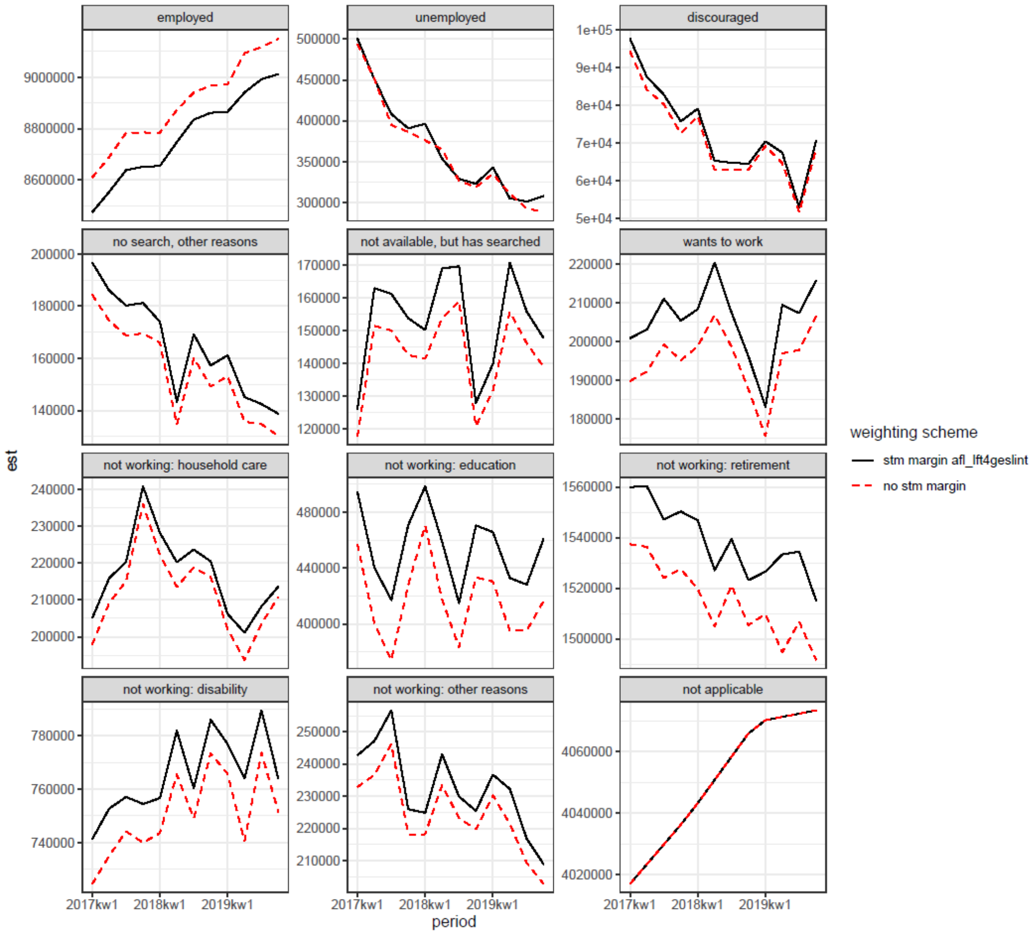


Figure 4.2.2: Quarterly GREG estimates with and without the STM-margin in the weighting scheme for the target variable ‘relation to the labour market’.



4.3 Standard errors of the GREG estimator under measurement error

In this subsection the proposed analytical variance approximation that accounts for the measurement error in the weighting scheme, i.e. expression (12) in Section 3, is evaluated and compared with two different versions of standard variance approximation of the GREG estimator. The first version relates to the standard errors based on weighting scheme (9), which includes the STM-margin Afl_Lft4GeslInt which is based on the monthly STS model. This approximation will underestimate the standard errors if the target variable is highly correlated or coincides with the estimates in Afl_Lft4GeslInt. The second version has standard errors based on weighting scheme (9), where table Afl_Lft4GeslInt is left out. The purpose of this comparison is to investigate to what extent the latter approach can be used as a pragmatic approximation for the standard errors of the quarterly figures.

The contribution of measurement error to the standard variance approximation of the GREG estimator depends on the type of target variable. Regarding the two target variables considered in this article the impact is quite different. Figure 4.3.1 presents the results for educational attainment. The ASE's and RDSE's are summarized in Table 4.3.1. The standard errors under the three approaches are almost similar. The relation between Afl_Lft4GeslInt and educational attainment is weak, resulting in very small values for $\boldsymbol{\beta}_m$ in (11). As a result, the measurement error correction leads to changes in the standard errors that are relatively small compared to the size of the standard errors. Leaving out the STM-margin Afl_Lft4GeslInt from the weighting scheme also hardly effects the standard errors of the standard GREG approximations. This is an indication that adding the STM-margin to the weighting scheme does not severely inflate the dispersion of the GREG weights and therefore the standard errors for a target variable like educational attainment, which is weakly related to Afl_Lft4GeslInt.

Results for relation to the labour market are presented in Figure 4.3.2. The ASE's and RDSE's are summarized in Table 4.3.2. It can be seen that the impact of accounting for the uncertainty of the STM-margin through the proposed measurement error correction is more pronounced, although there are differences among the categories of this target variable. Categories employed and unemployed show the greatest changes. Without the measurement error correction term, the GREG-estimates produce zero standard errors, as the weighting procedure includes the STM-margins. With the measurement error correction term the standard estimates rebound to the level of the standard errors of the three months averages of the STM-margins. The standard errors of the GREG estimator that accounts for the uncertainty of the STM-margins are still smaller than the standard errors of the GREG-estimates under the weighting scheme without the STM-margins. For the employed the standard errors are roughly 20 percent smaller over the period 2017-2019, for the unemployed 30 percent. For target variables that coincide with the STM-margin, the over-estimation of the standard error with a GREG estimator that leaves out the STM-margin from the weighting scheme can be substantial.

Turning to the other categories of relation to the labour market it can be seen that accounting for the uncertainty of the STM-margin leads to relatively smaller changes in standard errors. The impact is most clear for the categories not in the labour force due to education and not in the labour force due to retirement or high age.

Figure 4.3.1 GREG standard errors without (black, red) and with measurement error correction (green) for the target variable ‘educational attainment’. The following abbreviations are used: sve: secondary vocational training, ge: general education, hve: higher vocational training.

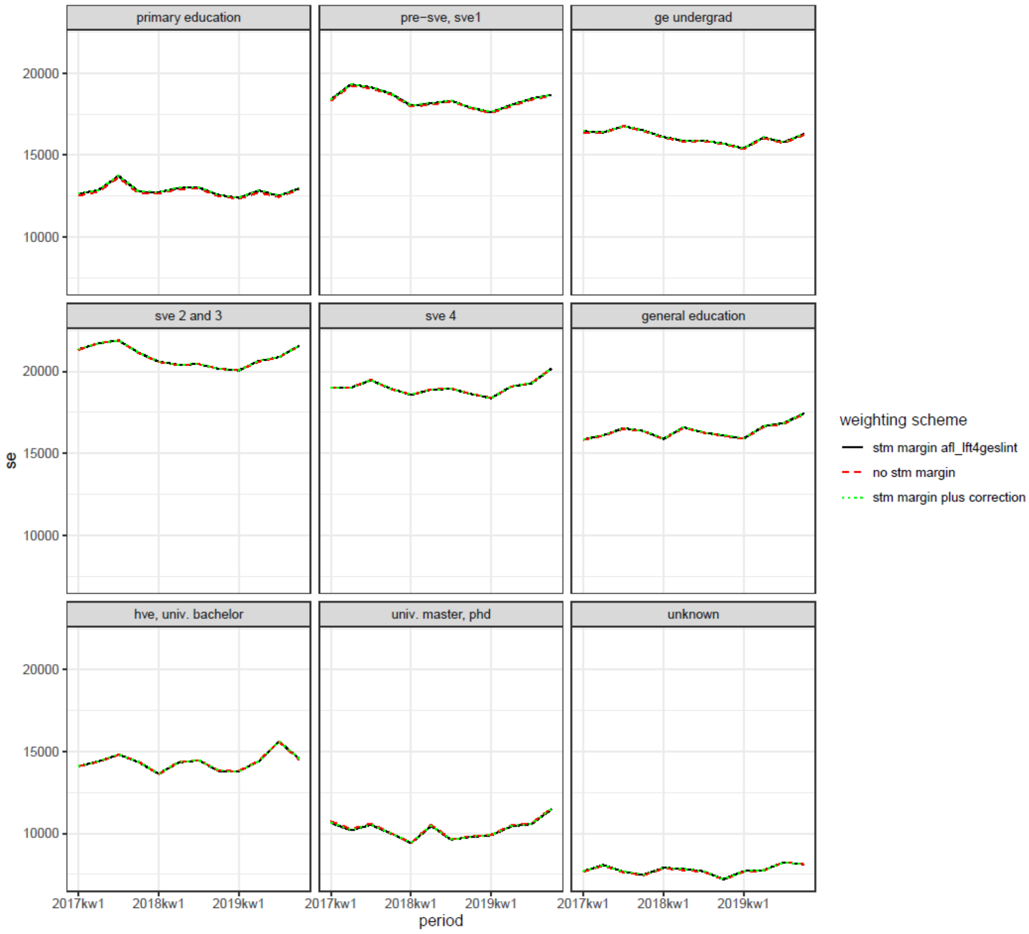


Figure 4.3.2 GREG standard errors without (black, red) and with measurement error correction (green) for the target variable ‘relation to the labour market’

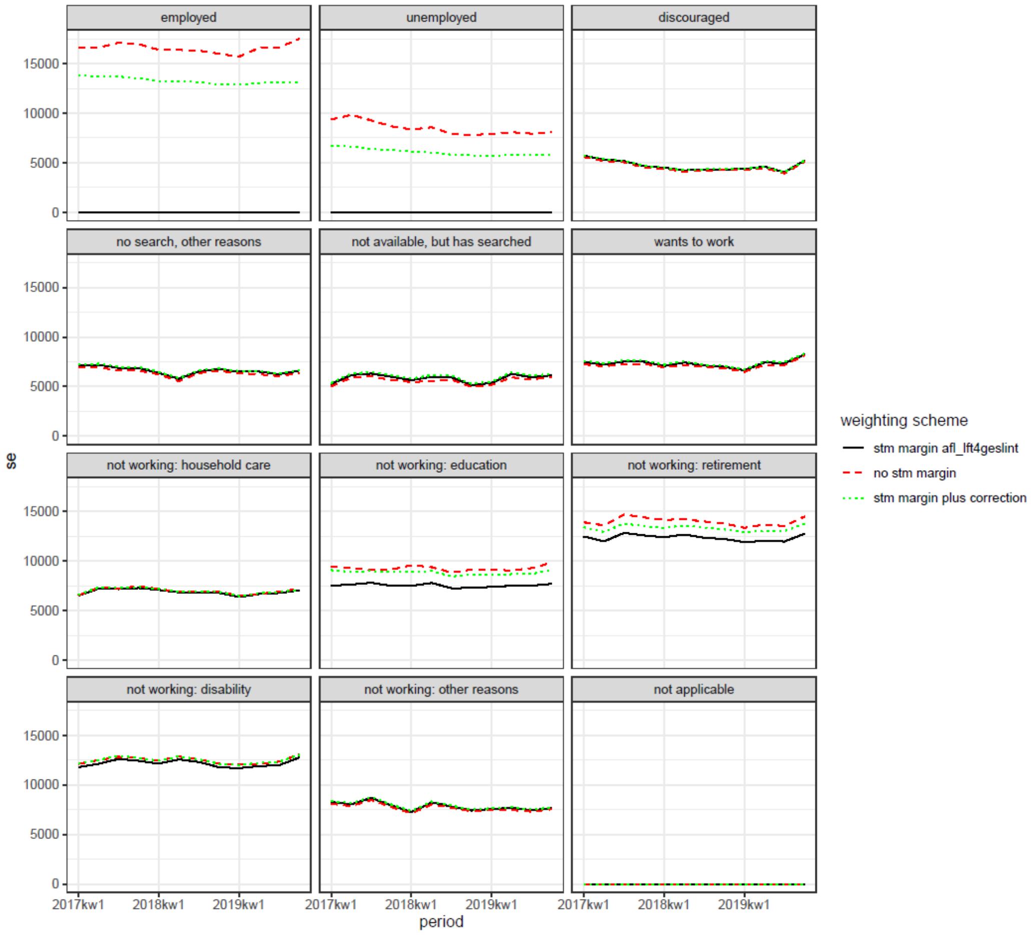


Table 4.3.1 effect on point measures and standard errors for educational attainment

Category	ASE(+m)	ASE(-m)	ASE(me)	RDSE (+m,-m)	RDSE (+m,me)	RDSE (-m,me)
	counts	counts	counts	%	%	%
Primary education	12834	12749	12849	0,7	-0,1	-0,8
Pre-secondary vocational education, secondary vocational education level 1	18401	18343	18411	0,3	-0,1	-0,4
General education undergrad	16105	16062	16108	0,3	0	-0,3
Secondary vocational education level 2 and 3	20905	20892	20912	0,1	0	-0,1
Secondary vocational education level 4	19028	19036	19035	0	0	0
General education	16377	16348	16388	0,2	-0,1	-0,2
Higher Vocational Education, University Bachelor	14359	14360	14370	0	-0,1	-0,1
University Master, PhD	10249	10298	10264	-0,5	-0,1	0,3
Unknown	7778	7748	7780	0,4	0	-0,4

Table 4.3.2 effect on point measures and standard errors for relation to the labour market

Category	ASE(+m)	ASE(-m)	ASE(me)	RDSE (+m,-m)	RDSE (+m,me)	RDSE (-m,me)
	counts	counts	counts	%	%	%
Employed	0	16568	13280	-100	-100	24,8
Unemployed	0	8480	6064	-100	-100	39,7
Discouraged	4706	4586	4720	2,6	-0,3	-2,8
No search, other reasons	6623	6407	6710	3,4	-1,3	-4,5
Not available, but has searched	5850	5597	6003	4,5	-2,5	-6,7
Wants to work	7338	7155	7447	2,6	-1,5	-3,9
Not working: household care	6875	6969	6967	-1,3	-1,3	0
Not working: education	7522	9255	8820	-18,7	-14,7	4,9
Not working: retirement	12343	13980	13321	-11,7	-7,3	4,9
Not working: disability	12173	12477	12496	-2,4	-2,6	-0,2
Not working: other reasons	7829	7680	7914	1,9	-1,1	-2,9

The impact of accounting for the measurement error of the STM-margin Afl_Lft4GeslInt in the variance and standard error estimates depends on the values of the regression coefficients $\boldsymbol{\beta}_m$. As can be seen from equation (11), the measurement error contribution has a variance-covariance matrix with non-zero diagonal terms (and zero off-diagonal entries) that does not vary across the quarterly target variables. Differences in the impact from this contribution come from the regression coefficients which reflect the (partial) correlation between the target variables and the STM-margin. See table 4.3.3 for values of the regression coefficients for different relations to the labour market categories for the first quarter of 2017. Target variables such as the employed and unemployed are also included in the categories of the STM-margins. For these variables the regression coefficient is equal to one for the category that corresponds with the target variable while the remaining coefficients are zero. For such target variables, the variance of the quarterly estimates equals the variance of the STM-margin, which is the variance of the three monthly estimates from the STS-model, see also the discussion around equation (13) in Section 3. For target variables not included in the STM-margin, the regression coefficients are likely to be smaller (in absolute values) and depend on the relation between the target variable and the variables included in the STM-margin. The larger the (absolute) values of the regression coefficients, the larger the contribution of the variance of the STM-margins to the variance of the GREG estimator for the quarterly target variable.

Table 4.3.3 Parameter coefficients of STM-weighting terms for different relation to the labour market categories (2017 Q1)

	Employed	Unemployed	Not working (education)	Not working (retirement)	Discouraged
Coefficient					
Unemployed, 15-24 years, men	-5,22E-14	1,00E+00	-0,635	-0,006	-0,025
Unemployed, 15-24 years, women	-2,78E-13	1,00E+00	-0,661	-0,011	-0,017
Unemployed, 25-44 years, men	3,74E-12	1,00E+00	-0,138	-0,051	-0,059
Unemployed, 25-44 years, women	4,87E-16	1,00E+00	-0,071	-0,034	-0,042
Unemployed, 45-64 years, men	-4,66E-13	1,00E+00	0,015	-0,416	-0,048
Unemployed, 45-64 years, women	1,91E-13	1,00E+00	0,015	-0,260	-0,055
Employed, 15-24 years, men	1,00E+00	-6,07E-15	-0,648	-0,043	-0,021
Employed, 15-24 years, women	1,00E+00	-5,63E-15	-0,651	-0,042	-0,013
Employed, 25-44 years, men	1,00E+00	-9,86E-14	-0,148	-0,153	-0,045
Employed, 25-44 years, women	1,00E+00	-7,99E-15	-0,080	-0,135	-0,024
Employed, 45-64 years, men	1,00E+00	4,09E-15	-0,006	-0,565	-0,033
Employed, 45-64 years, women	1,00E+00	-1,30E-14	-0,002	-0,374	-0,034

5. Discussion

In this paper a variance approximation is proposed for generalized regression (GREG) estimators that contain population tables in the weighting scheme that are observed with measurement error. Existing methods in the literature that account for uncertainty in the weighting scheme of the GREG estimator are based on two-phase sampling where estimates of population tables from the first phase are used as auxiliary information for the GREG estimator in the second phase. In this situation the sample design can be used to derive a variance approximation for the GREG estimator that accounts for this additional uncertainty. See e.g. Särndal et al (1992), Chapter 9 for details.

The situation considered in this paper is different in a sense that the GREG estimator accounts for uncertainty in certain population tables included in the weighting scheme, which are subject to uncertainty. It is naturally assumed that the level of uncertainty of the population tables is known a priori.

The proposed method is applied to the quarterly estimates of the Dutch Labour Force Survey (DLFS). The DLFS applies a multivariate structural time series (STS) model for the production of official monthly labour force figures. The quarterly three months averages of the monthly output tables are included in the weighting scheme of the GREG estimator for the quarterly figures with the purpose to enforce numerical consistency between monthly and quarterly publication tables. The standard variance approximation of the GREG estimator will treat the monthly publication tables as if they are observed without error. As a result, the variance of quarterly target variables that are strongly related to this monthly publication table will be severely underestimated. In the most extreme case, the variance will tend to zero for target variables that are also included in this monthly publication table. In the proposed method, the variance approximation of the GREG estimator is increased with the sum over the variances of the STS model estimates of the monthly publication table multiplied with the squared values of the regression coefficients of the monthly output table of the GREG estimator.

A more pragmatic approach is to use the standard variance approximation of the GREG estimator with a weighting scheme without the monthly publication table. Compared to the proposed variance approximation, this alternative does not or only slightly overestimate the standard errors for quarterly target variables that are not or only weakly related to the variables in the monthly output table. For quarterly target variables that are highly correlated with the variables in the monthly output table, this approach may overestimate the standard errors by up to 20%.

Statistics Netherlands will use the variance approximation developed in this paper to publish standard errors of the quarterly DLFS figures. One disadvantage of the proposed variance approximation is that it requires the evaluation of an additional component to account for the uncertainty of the monthly publication table. This involves a linear combination of the regression coefficients of the monthly publication table from the weighting scheme and the standard errors of the monthly estimates of the time series model. For target variables that are not or only weakly related to the monthly publication table from the weighting scheme, the standard errors of the quarterly figures will be approximated using the variance of the GREG estimator without the monthly publication table.

For quarterly figures that are strongly related to the monthly publication table used in the weighting scheme, it is necessary to account for its additional uncertainty. This is relatively straightforward for target variables that coincide with one of the monthly publication table categories, since the regression coefficients are equal to one in this case. Consequently, the standard error of the quarterly figure is equal to the standard error of the average of the three monthly time series estimates. Since the standard errors of the monthly time series estimates are very stable over time, there is no need to recalculate them for each quarter.

For quarterly figures that are detailed cross-classifications of the categories in the monthly publication table from the weighting scheme, it is also necessary to account for the additional uncertainty in the monthly time-series estimates. In this case, the regression coefficients will not be exactly equal to one, so an estimate from the GREG estimator will be required. A follow-up project will investigate the stability of the variances of the GREG estimates of the quarterly DLFS figures over time. If the standard errors are relatively time invariant, the standard errors of the quarterly DLFS figures will be published in a separate document. Depending on their stability over time, these standard errors will need to be updated periodically.

Acknowledgement

The views expressed in this paper are those of the authors and not reflect the policy of Statistics Netherlands. Harm Jan Boonstra reviewed the paper. The authors are grateful for his constructive comments.

References

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

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

Durbin, J., and Koopman, S.J. (2012). Time series analysis by state space methods. Second edition. Oxford: Oxford University Press.

Fuller, W.A. en Rao, J.N.K. (2001), A Regression Composite Estimator with Application to the Canadian Labour Force Survey, Survey Methodology 27, No. 1, p. 45-51.

Koopman, S.J., N. Shephard, and J. Doornik (2008). Ssfpack 3.0: Statistical algorithms for models in state-space form. Timberlake Consultants, Press London.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979). Multivariate Analysis. London: Academic Press

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. Journal of Business & Economic Statistics, 9, pp. 163-175.

Rao, J. and I. Molina (2015) Small Area Estimation, 2nd edition, Wiley & Sons, New York.

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

Van den Brakel, J.A. (2022). Monthly Labour Force Figures during the 2021 redesign of the Dutch Labour Force Survey. Discussion paper, January 2022. Statistics Netherlands, Heerlen.

Van den Brakel, J.A., S. Krieg, (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Survey Methodology. Vol. 35, pp. 177-190.

Van den Brakel, J.A., S. Krieg, (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. Survey Methodology. Vol. 41, pp. 267-296.

Van den Brakel, J.A. and J. Michiels (2024). Variance of the generalized regression estimator under measurement error - Interim report. Statistics Netherlands.

Van den Brakel, J.A., M. Souren and S. Krieg (2022). Estimating monthly Labour Force Figures during the COVID-19 pandemic in the Netherlands. Journal of the Royal Statistical Society, Series A. Vol 185, pp 1560-1583.

Appendix: Proof formula (13)

In this appendix it is shown that the matrix of regression coefficients of the GREG estimator equals (13) if the GREG estimator is applied to estimate the table from the weighting scheme that corresponds to the components that are estimated with the time series model. First the GREG estimator is defined for a vector of population totals:

$$\hat{\boldsymbol{t}}_y^{R0} = \hat{\boldsymbol{t}}_y^{\pi} + \mathbf{B}'(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x^{\pi}),$$

with $\hat{\boldsymbol{t}}_y^{R0}$ and $\hat{\boldsymbol{t}}_y^{\pi}$ two q dimensional column vectors containing the GREG and Horvitz-Thompson estimators for the q population totals of interest and $\mathbf{B}$ a $p \times q$ dimensional matrix containing the regression coefficients of the multivariate regression model that motivates the GREG estimator. The multivariate regression model is defined as

$$\boldsymbol{Y} = \boldsymbol{X}\,\mathbf{B} + \boldsymbol{E},$$

with $\boldsymbol{Y}$ an $n \times q$ matrix where each row corresponds to the q target values $(y_{i,1}, \dots, y_{i,q})$ of sampling unit $i = 1, \dots, n$, $\boldsymbol{X}$ an $n \times p$ matrix where each row corresponds to the p auxiliary variables from the weighting scheme $(x_{i,1}, \dots, x_{i,p})$ of sampling unit $i = 1, \dots, n$, and $\boldsymbol{E}$ an $n \times q$ matrix with the corresponding residuals of the regression model. The generalized regression estimator for $\mathbf{B}$ is now defined in matrix notation as:

$$\widehat{\mathbf{B}} = [\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

with $\boldsymbol{\Sigma}$ an $n \times n$ diagonal matrix with diagonal elements $\pi_i v_i$, i.e. the product of the inclusion probabilities π_i and the scaling factor of the variance of the linear regression model v_i of sampling unit $i = 1, \dots, n$. Recall from Section 4 that the p dimensional vector with population totals for the auxiliary variables $\boldsymbol{t}_x$ can be split in a vector for which the true population totals are known $\boldsymbol{t}_a$ of length l and a vector for which the population totals for the monthly publication tables are estimated with the time series model $\tilde{\boldsymbol{t}}_m$ of length k. Thus $\boldsymbol{t}_x = (\tilde{\boldsymbol{t}}_m' \, \boldsymbol{t}_a')'$, $\boldsymbol{X} = (\boldsymbol{X}_m \, \boldsymbol{X}_a)$ with $\boldsymbol{X}_m$ an $n \times k$ matrix where each row corresponds to the k auxiliary variables of $\tilde{\boldsymbol{t}}_m$ and with $\boldsymbol{X}_a$ an $n \times l$ matrix where each row corresponds to the l auxiliary variables of $\boldsymbol{t}_a$.

Consider the situation that $\hat{\boldsymbol{t}}_y^{R0}$ corresponds to the component of the weighting model that is estimated with the time series model $\boldsymbol{t}_m$. In this case $\hat{\boldsymbol{t}}_y^{R0} = \hat{\boldsymbol{t}}_m^{R0}$. In this case it can be shown that

$$\widehat{\mathbf{B}} = \begin{bmatrix} \boldsymbol{I}_{[k \times k]} \\ \boldsymbol{O}_{[l \times k]} \end{bmatrix}.$$

Proof:
In this case $\boldsymbol{Y} = \boldsymbol{X}_m$ and $\boldsymbol{X} = (\boldsymbol{X}_m \, \boldsymbol{X}_a)$, leading to the following expression for $\widehat{\mathbf{B}}$:

$$\widehat{\mathbf{B}} = [(\boldsymbol{X}_m \, \boldsymbol{X}_a)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_m \, \boldsymbol{X}_a)]^{-1}(\boldsymbol{X}_m \, \boldsymbol{X}_a)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_m.$$

Elaborating on the matrix operations gives:

$$\widehat{\mathbf{B}} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{21} \end{bmatrix},$$

with

$\boldsymbol{X}_{11} = \boldsymbol{X}_m'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_m,$
$\boldsymbol{X}_{21} = \boldsymbol{X}_a'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_m,$
$\boldsymbol{X}_{12} = \boldsymbol{X}_m'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_a,$
$\boldsymbol{X}_{22} = \boldsymbol{X}_a'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_a.$

Taking the inverse of the partitioned matrix, see e.g. Mardia, Kent and Bibby (1979), Section A.2.4, gives:

$$\widehat{\mathbf{B}} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1} \\ -\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A} & (\boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{21} \end{bmatrix}, \tag{A.1}$$

with $\boldsymbol{A} = (\boldsymbol{X}_{11} - \boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21})^{-1}$.

Using formula (A.2.4f) from Mardia, Kent and Bibby (1979), we have:

$$(\boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12})^{-1} = \boldsymbol{X}_{22}^{-1} + \boldsymbol{X}_{22}^{-1} \, \boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}. \tag{A.2}$$

Inserting (A.2) in (A.1) and further elaborating on the matrix operations gives:

$$\widehat{\mathbf{B}} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{X}_{11} - \boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} \\ -\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{X}_{11} + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} \end{bmatrix}.$$

From this it follows for the first component that

$$\boldsymbol{A}\boldsymbol{X}_{11} - \boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} = \boldsymbol{A}\boldsymbol{A}^{-\mathbf{1}} = \boldsymbol{I},$$

and for the second component

$$\begin{aligned} &-\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{X}_{11} + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} \\ &= -\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}(\boldsymbol{X}_{11} - \boldsymbol{X}_{12}\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}) + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} \\ &= -\boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21}\boldsymbol{A}\boldsymbol{A}^{-\mathbf{1}} + \boldsymbol{X}_{22}^{-1}\boldsymbol{X}_{21} = \boldsymbol{O}. \end{aligned}$$

Collecting both results gives

$$\widehat{\mathbf{B}} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{O} \end{bmatrix},$$

which proves (13).

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/’18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/’14–2017/’18	Crop year, financial year, etc., 2015/’16 to 2017/’18 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information