Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Variance of the generalized regression estimator under measurement error

by Jan van den Brakel and John Michiels

Release date: June 29, 2026

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, June 2026
Vol. 52, No. 1, pp. 95-116
Statistics Canada, Catalogue No. 12-001-X

Variance of the generalized regression estimator under measurement error

Jan van den Brakel and John Michiels[1]

Abstract

With the exception of two-phase sampling, the standard variance approximation of the generalized regression (GREG) estimator assumes that the population totals in the weighting scheme are observed without error. If the weighting model of the GREG estimator contains population totals that are observed with measurement error sources other than the sampling error of first-phase estimates, then this uncertainty will be ignored by the variance approximation of the GREG estimator. This paper proposes a variance approximation for the GREG estimator that accounts for additional uncertainty arising from measurement error in one or more of the population totals used in the weighting scheme. This approach has been developed for, and is being applied to, the Dutch Labour Force Survey (DLFS). The monthly publications of the DLFS are obtained with a time series model, which corrects for rotation group bias and discontinuities caused by major redesigns and the loss of face-to-face interviews during COVID-19. The GREG estimates for the quarterly figures are benchmarked to the average of the monthly publications to enforce numerical consistency between monthly and quarterly publication tables. The standard variance approximation of the GREG estimator assumes that these population totals are observed without error. This results in an underestimation of the variance of the GREG estimator. The variance approximation proposed in this paper results in more realistic standard errors for the quarterly GREG estimates.

Key Words: Calibration; Labour Force Survey; Numerical consistency between estimated tables; Variance estimation.

1. Introduction

Official statistics published by national statistical institutes are predominantly based on sample surveys in combination with the generalized regression (GREG) estimator (Särndal, Swensson and Wretman, 1992). 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 exception 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 (Särndal and Swensson, 1987; Hidiroglou and Särndal, 1998; Singh and Kumar, 2010). A related method, proposed by Renssen and Nieuwenbroek (1997) and Berger, Muñoz and Rancourt (2009), is the variance approximation that accounts for the uncertainty of using sample estimates for shared variables in the GREG estimator that are based on two or more probability samples.

1. Jan van den Brakel, Statistics Netherlands, P.O. Box 2454, 2490 AA Den Haag, Netherlands and Maastricht University, P.O. Box 616, 6200 MD Maastricht, Netherlands. E-mail: ja.vandenbrakel@cbs.nl; John Michiels, Statistics Netherlands, P.O. Box 2454, 2490 AA Den Haag, Netherlands.

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. An example is the GREG estimator for the quarterly figures of the Dutch Labour Force Survey (DLFS).

Based on the DLFS monthly, quarterly and annual figures are published on the employed and unemployed labour force in the Netherlands. The DLFS is based on a rotating panel design. Monthly publication tables are compiled with a multivariate structural time series model (STM) to address different issues faced by panel surveys such as the DLFS. The model 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. The model also accounts for rotation group bias (RGB), see Bailar (1975). Furthermore, the model accounts for systematic effects that are the result of three major survey process redesigns and for the systematic effects of the loss of face-to-face interviews during the lockdowns of the COVID-19 pandemic (Van den Brakel and Krieg, 2015; Van den Brakel, Souren and Krieg, 2022).

Quarterly and annual figures are predominantly produced with the more commonly used GREG estimator. 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 additional uncertainty of using estimates for population totals, derived from a time series model, is ignored in the variances of the GREG estimator. This can strongly underestimate the variance of quarterly GREG estimates, in particular if they are strongly related to the weighting table that is based on the monthly time series estimates. 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.

The paper is organized as follows. In Section 2 the standard GREG estimator is briefly introduced, mainly with the purpose to define notation. In Section 3 the DLFS, the STM for monthly figures and the GREG estimator for the quarterly figures are described. Some details of the STM for monthly figures are provided as a motivation why the monthly publication table is included in the weighting scheme of the quarterly GREG estimator. In Section 4 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 5 the results of an application to the quarterly figures of the DLFS are presented. The paper concludes with a discussion in Section 6.

2. Generalized regression 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:

$$t_y = \sum_{i \in U} y_i. \tag{2.1}$$

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 probabilities π_i for population unit i and second order inclusion probabilities $\pi_{ii'}$ for population units i and i'. 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;

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i, \tag{2.2}$$

with y_i the response of the target variable of sampling unit i, $\mathbf{x}_i = (x_{i,1},\ldots,x_{i,p})'$ a p dimensional vector containing the auxiliary variables of sampling unit i, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ a p dimensional vector containing the regression coefficients and ε_i a residual. It is assumed that $E_\xi(\varepsilon_i) = 0$ and $\mathrm{Var}_\xi(\varepsilon_i) = \nu_i \sigma^2$, where $E_\xi(.)$ and $\mathrm{Var}_\xi(.)$ denote the expectation and variance with respect to model (2.2) and ν_i a scaling factor that is known for each sampling unit. In Särndal et al. (1992), Chapter 6 it is shown that the GREG estimator for t_y can be expressed as

$$\hat{t}_y^R = \hat{t}_y^\pi + \hat{\boldsymbol{\beta}}' (\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi), \tag{2.3}$$

with $\hat{t}_y^\pi$ the Narain-Horvitz-Thompson estimator (Narain, 1951 and Horvitz and Thompson, 1952) for t_y, i.e.

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

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

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

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

$$\hat{\mathbf{t}}_x^\pi = \sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i}.$$

An estimator for the regression coefficients in (2.3) is defined as:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \frac{\mathbf{x}_i \mathbf{x}_i'}{\pi_i \nu_i} \right)^{-1} \sum_{i \in s} \frac{\mathbf{x}_i y_i}{\pi_i \nu_i}.$$

The variance of the GREG estimator is obtained by approximating (2.3) with a first order Taylor linearization (Särndal et al. (1992), Chapter 6):

$$\hat{t}_y^R = \hat{t}_y^\pi + \boldsymbol{\beta}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi) + \text{Rest} \doteq \hat{t}_y^\pi + \boldsymbol{\beta}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^\pi) \equiv \hat{t}_y^{R0}. \tag{2.4}$$

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

$$V(\hat{t}_y^{R0}) = \sum_{i\in U}\sum_{j\in U}(\pi_{ii'} - \pi_i\pi_j)\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}, \tag{2.5}$$

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

$$\hat{V}(\hat{t}_y^{R0}) = \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ii'} - \pi_i\pi_j)}{\pi_{ii'}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j}, \tag{2.6}$$

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

3. Dutch Labour Force Survey

3.1 Survey design

The objective of the Dutch Labour Force Survey (DLFS) is to provide reliable information about the Dutch labour force. Before 2000, the DLFS was designed as a cross-sectional survey. The DLFS has been conducted as a rotating panel design since October 1999. Each month a sample enters the panel. The sample units are observed five times at quarterly intervals, which is called a 1-2-1 (5) monthly rotating panel design. Until 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. In 2021 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.

Commencing the moment that the DLFS changed to a rotating panel design about 8,000 respondents were observed in the monthly samples of the first wave. This gradually felt back to 6,500 persons in 2012. After the second redesign in 2012, the monthly sample size in the first wave increased such that about 8,500 respondents were obtained in the monthly samples of the first wave. Since 2021, the average monthly sample size in the first wave is about 5,000 persons. Response rates in the first wave are about 55%. In the follow-up waves about 90% of the respondents from the preceding wave are observed.

Since the introduction of the panel design in 2000, the survey design has been redesigned three times. The first redesign took place in 2010, where the data collection in the first wave changed from face-to-face interviewing (CAPI) only to a mixed mode design of telephone interviewing (CATI) and CAPI. This required also a fundamental change of the questionnaire. In 2012 the data collection changed to sequential mixed mode design that starts with Web interviewing (CAWI) with a non-response follow-up that is based on CATI and CAPI. In the redesign of 2021 the sample design changed from a household-based sample to a person-based sample. Data collection in the follow-up waves was based on CATI only commencing the introduction of the rotating panel design until 2021. During the redesign of 2021 a sequential mixed-mode design based on CAWI and CATI is implemented in the follow-up waves.

3.2 Time series model for monthly labour force figures

Until 2010, Statistics Netherlands published rolling quarterly labour force figures on a monthly basis using the GREG estimator. However, the monthly sample size was considered too small to produce accurate monthly labour force figures using this method. Therefore, a multivariate structural time series (STS) model is implemented in June 2010 for the production of monthly labour force figures. The model is used for small area estimation, providing the most accurate possible monthly estimates by incorporating sample information from previous reference periods. The model also corrects for RGB and the systematic effects in the outcomes resulting from survey redesigns and the loss of CAPI during the COVID-19 crisis. This subsection gives an overview of this model.

The sample observed for the j-th time is henceforth shortly denoted as the j-th wave. As a result of the rotation scheme, data are collected in five independent samples each month. Each month, a new sample enters the panel. This is the first wave, which is the sample observed for the first time. The second wave is the sample that entered the panel three months ago and is now being observed for the second time. Similarly, the samples of the third, fourth and fifth waves are those that entered the panel six, nine and twelve months ago, and are now being observed for the third, fourth and fifth times, respectively. Samples observed for the fifth time leave the panel.

The purpose is to estimate monthly population totals for labour force figures defined by (2.1). In this case, y denotes the indicators that define the labour force status of individual i in month t. Let $\hat{t}_{y_t}^{[j]}$ denote the general regression (GREG) estimator, defined by (2.3), for an unknown population total (2.1) in month t, based on the sample that is observed for the j-th time. See Appendix A.1 for the weighting scheme of this GREG estimator. The 1-2-1 (5) rotating panel design implies that each month five GREG estimates are observed that can be collected in a five dimensional vector, say $\hat{\mathbf{y}}_t = (\hat{t}_{y_t}^{[1]}, \ldots, \hat{t}_{y_t}^{[5]})'$. From this, a five-dimensional time series can be constructed, which is the input of the following STS model:

$$\hat{\mathbf{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \sum_{r=1}^{3} \boldsymbol{\Delta}_t^r \boldsymbol{\gamma}^r + \boldsymbol{\delta}_t^{\mathrm{COV}} \gamma^{\mathrm{COV}} + \boldsymbol{\varepsilon}_t. \tag{3.1}$$

This is an extension of the model proposed by Pfeffermann (1991). The first component θ_t in (3.1) 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 $\hat{\mathbf{y}}_t$ contains five GREG estimates for the population parameter in month t. The population parameter is modelled with a so-called basic STM, i.e.

$$\theta_t = L_t + S_t + I_t, \tag{3.2}$$

with L_t the so-called smooth trend model for the low frequency variation in the series of the population parameter, S_t the trigonometric seasonal model for the monthly effects in the series and I_t a white noise component for the unexplained variation of the population parameter, see Durbin and Koopman, 2012, Chapter 3 for details. This first component can be interpreted 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 (3.1), i.e. $\boldsymbol{\lambda}_t$, models the 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, 1975), 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. In this way, the labour force are comparable with the estimates obtained under the cross-sectional design used before the introduction of the rotating panel design in October 1999.

The third component models 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. $\boldsymbol{\Delta}_t^r = \text{diag}\,(\delta_t^{r,[1]}, \delta_t^{r,[2]}, \delta_t^{r,[3]}, \delta_t^{r,[4]}, \delta_t^{r,[5]})$, which denotes a diagonal matrix with dummy variables $\delta_t^{r,[j]}$ 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 $r = 1$ in 2010, $r = 2$ in 2012 and $r = 3$ in 2021. Furthermore, $\boldsymbol{\gamma}^r = (\gamma^{r,[1]}, \gamma^{r,[2]}, \gamma^{r,[3]}, \gamma^{r,[4]}, \gamma^{r,[5]})'$ are five dimensional vectors that contain estimates for the discontinuities in the five waves during redesign $r = 1$ in 2010, $r = 2$ in 2012, and $r = 3$ in 2021.

The fourth component in (3.1), $\boldsymbol{\delta}_t^{\text{COV}}\gamma^{\text{COV}}$, 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, $\boldsymbol{\delta}_t^{\text{COV}} = (\delta_t^{\text{cov}}, 0, 0, 0, 0)'$ 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 (3.1), i.e. $\boldsymbol{\varepsilon}_t$, models the survey errors and 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. $\varepsilon_t^{[j]} = \sqrt{\text{var}(\hat{y}_t^{[j]})}\ \tilde{\varepsilon}_t^{[j]}$. 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(3) model to accommodate serial correlation with past observations. See Van den Brakel and Krieg (2015) for details.

Model (3.1) can be expressed in the so-called state space representation and fitted with the Kalman filter to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2012). The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman, Shephard and Doornik (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 $\gamma^{r,[1]}$ is the discontinuity for the first wave of the r-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 (3.1) 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.

3.3 GREG estimator for quarterly estimates

The quarterly estimates of the DLFS are based on the GREG estimator. Due to the use of a 1-2-1 (5) monthly design, the observations observed in each quarter are based on 15 independent samples, since none of the waves observed in one quarter are overlapping. This results in a large enough sample size to produce sufficiently precise estimates using a design-based inference approach, such as the GREG estimator. The

weighting scheme for the quarterly GREG estimator for the DLFS is based on a complex set of tables that contain social-demographic auxiliary variables for which the population totals are derived from registers.

Unlike the time series model used to produce the monthly labour force figures, the GREG estimator does not take into account RGB, the effects of the three redesigns, or the loss of CAPI respondents during the period of the pandemic. To enforce that the quarterly estimates for the labour market position are consistent with the monthly estimates, the weighting scheme for the quarterly GREG estimator also contains the three-month averages from the monthly publication tables. In this way, the quarterly estimates that correspond to the monthly figures will be numerically consistent, and the quarterly figures that are refinements or are related to the monthly figures will be corrected for these systematic effects. The additional table in the weighting scheme that contains the three-month averages from the monthly publication tables is hereafter referred to as the STM-margin (structural time series model margin). The weighting scheme for the quarterly figures is specified in Appendix A.2.

Adding a population table to the weighting scheme of the GREG estimator complicates the variance estimation. The variance of the GREG estimator assumes that the population totals of the three-month averages from the monthly publication tables is observed without error. Ignoring the variance of the time series model estimates leads to a significant underestimation of the variance of the quarterly GREG estimates. Currently the variance of the quarterly labour force figures is approximated with a GREG estimator without the three-month averages from the monthly publication tables in the weighting scheme. This is a pragmatic approach that might over-estimate the variance. In the next section a variance approximation is proposed that accommodates the uncertainty of population totals in the weighting scheme of the GREG estimator.

4. Variance of the GREG estimator under measurement error

Variance approximation (2.5) and its estimator (2.6) both ignore additional uncertainty arising in GREG estimates when the weighting scheme contains components observed with measurement error. In this application, this uncertainty arises because the weighting scheme contains population totals estimated with a time series model for series of repeated monthly survey estimates. This section develops a method to account for this uncertainty in the context of benchmarking quarterly GREG estimates against three-month averages of time series model predictions. It is understood that this method is more general and can be applied to situations where estimates are benchmarked against population totals derived from other, independent data sources.

To account for this additional uncertainty, the vector with population totals of the auxiliary variables $\mathbf{t}_x$ is split in a term for which the true population totals are known, say $\mathbf{t}_a$ (where subscript *a* stands for administration), and a term for which the true population totals are estimated, say $\tilde{\mathbf{t}}_m$ (where subscript *m* stands for model estimate), i.e. $\mathbf{t}_x = (\tilde{\mathbf{t}}_m' \, \mathbf{t}_a')'$. This results in:

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

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

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

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:

$$\mathrm{Var}(\hat{t}_y^{R0}) = E_m \mathrm{Var}_s(\hat{t}_y^{R0} \mid m) + \mathrm{Var}_m E_s(\hat{t}_y^{R0} \mid m), \tag{4.1}$$

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 (4.1) 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. As a result it follows for the first term in (4.1) that:

$$E_m \mathrm{Var}_s(\hat{t}_y^{R0} \mid m) = E_m \mathrm{Var}_s(\hat{t}_y^{R0}) = \mathrm{Var}_s(\hat{t}_y^{R0}).$$

For the second term in (4.1) it follows that:

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

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

$$\mathrm{Var}_m E_s(\hat{t}_y^{R0} \mid m) = \mathrm{Var}_m(t_y + \boldsymbol{\beta}_m'(\tilde{\mathbf{t}}_m - \mathbf{t}_m)).$$

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

$$\mathrm{Var}_m E_s(\hat{t}_y^{R0} \mid m) = \boldsymbol{\beta}_m' \mathrm{Var}_m(\tilde{\mathbf{t}}_m) \boldsymbol{\beta}_m, \tag{4.2}$$

with $\mathrm{Var}_m(\tilde{\mathbf{t}}_m)$ 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 $\mathrm{Var}_m(\tilde{\mathbf{t}}_m)$ is a diagonal matrix, since the time series estimates $\tilde{\mathbf{t}}_m$ are obtained by applying STM (3.1) to each category of $\mathbf{t}_m$ separately. Note that in the case that the target variable t_y is included in the table defined by $\mathbf{t}_m$, 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 $t_y - \boldsymbol{\beta}_m' \mathbf{t}_m = t_y - t_y = 0$. The last paragraph of this section further elaborates on this special case and a proof is provided in Appendix A.3.

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

$$\mathrm{Var}(\hat{t}_y^{R0}) = \mathrm{Var}_s(\hat{t}_y^{R0}) + \boldsymbol{\beta}_m' \mathrm{Var}_m(\tilde{\mathbf{t}}_m) \boldsymbol{\beta}_m, \tag{4.3}$$

with $\mathrm{Var}_s(\hat{t}_y^{R0})$ defined in (2.5) that can be estimated with (2.6).

To motivate the variance approximation (4.3), 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. $\mathbf{t}_m$. Let k denote the number of categories of $\mathbf{t}_m$ and l the number of categories of $\mathbf{t}_a$, i.e. the components of the weighting scheme for which the population totals are known. If the GREG estimator is applied to estimate table $\mathbf{t}_m$, then $\hat{t}_y^{R0}$ in (2.4) becomes a k dimensional vector, say $\hat{\mathbf{t}}_m^{R0}$. It is shown in Appendix A.3 that in this case the regression coefficients in (2.4) form a $(k+l)\times k$ matrix equal to

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_{[k\times k]} \\ \mathbf{O}_{[l\times k]} \end{bmatrix}, \tag{4.4}$$

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

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

This also implies for the residual in (2.5) and (2.6) that $e_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i = x_{i,k'} - \boldsymbol{\beta}_{k'}'\mathbf{x}_i = 0$ 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 (2.6) would be equal to zero, which is undesirable and incorrect, since $\hat{\mathbf{t}}_m^{R0}$ is exactly equal to $\tilde{\mathbf{t}}_m$ which is an estimate based on a time series model and subject to uncertainty. Because of (4.4), the variance of the GREG estimator based on (4.3) is exactly equal to $\mathrm{Var}_m(\tilde{\mathbf{t}}_m)$ and thus correctly corresponds to the uncertainty of the time series model estimates $\tilde{\mathbf{t}}_m$.

There are a few related methods in the literature that account for the additional uncertainty when the GREG estimator contains margins that are subject to error. As mentioned in the introduction, the classical example of two-phase sampling is where the GREG estimator for the second phase uses population totals estimated from the observations of the first phase. Särndal and Swensson (1987), Hidiroglou and Särndal (1998), and Singh and Kumar (2010) propose variance approximations that account for this additional uncertainty.

Renssen and Nieuwenbroek (1997) extended the idea of two-phase sampling to align estimates for common variables in two or more sample surveys. This includes a variance approximation that accounts for the uncertainty of using sample estimates for the shared variables in the GREG estimator for the separate sample surveys. Berger et al. (2009) further elaborate on the variance estimator of Renssen and Nieuwenbroek (1997) by providing details for partial overlapping and non-overlapping samples. This includes the variance for composite estimators in rotating panel designs. The main difference with the method proposed in this

paper and that of Renssen and Nieuwenbroek (1997) and Berger et al. (2009) is that they develop variance approximations for GREG estimators that use common variables from multiple sample surveys, following the classical design-based inference approach for probability sampling. Consequently, they can account for dependency arising from partial sample overlap between the sources. In contrast, the method proposed in this paper can also account for measurement error in the population totals of the weighting scheme that is unrelated to sampling error.

5. Results

5.1 Evaluation measures

In this subsection evaluation measures are defined to compare the point estimates and the standard errors of the GREG estimators. Let $\hat{t}_{y,t}^{R(+m)}$ denote the GREG estimate for quarter t with the STM-margin, i.e. the three-month averages from the monthly publication tables, in the weighting scheme and $\hat{t}_{y,t}^{R(-m)}$ the GREG estimate for quarter t without this table in the weighting scheme. Let period $t=1$ refer to the first quarter of the GREG estimates of the period included in this comparison. In a similar way T refers to the last quarter of the GREG estimates of the period included in this comparison. As a result T quarterly GREG estimates are included in the comparison. To evaluate the effect on the point estimates we calculate the following measures:

- Mean Estimates:

$$\text{AE} = \frac{1}{T}\sum_{t=1}^{T}\hat{t}_{y,t}^{R(x)},\ x=[+m,-m].$$

- Average Relative Difference (RD):

$$\text{RD} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{t}_{y,t}^{R(+m)}-\hat{t}_{y,t}^{R(-m)}}{\hat{t}_{y,t}^{R(-m)}}\times 100\%.$$

- Average Relative Absolute Difference (RAD):

$$\text{RAD} = \frac{1}{T}\sum_{t=1}^{T}\frac{\left|\hat{t}_{y,t}^{R(+m)}-\hat{t}_{y,t}^{R(-m)}\right|}{\hat{t}_{y,t}^{R(-m)}}\times 100\%.$$

To evaluate the effect on the standard errors, let $\text{se}(\hat{t}_{y,t}^{R(+m)})$ denotes the standard error for the GREG estimator with the STM-margin using the variance approximation that ignores the uncertainty of the STM-margin in the weighting scheme. In a similar way $\text{se}(\hat{t}_{y,t}^{R(-m)})$ denotes the standard error for the GREG estimator without the STM-margin using the standard variance approximation that doesn't account for uncertainty in the population totals of the weighting scheme. Finally $\text{se}(\hat{t}_{y,t}^{R(me)})$ refers to the standard errors of the GREG estimator with the STM-margin using the variance approximation that accounts for the uncertainty of the STM-margin in the weighting scheme. The following measures are calculated:

- Average Standard Errors (ASE):

$$\text{ASE} = \frac{1}{T}\sum_{t=1}^{T} \text{se}(\hat{t}_{y,t}^{R(x)}),\ x = [+m, -m, me].$$

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 and always refers to the GREG estimator that includes the STM-margin in the weighting scheme.

- Average Relative Difference between Standard Errors (RDSE):

$$\text{RDSE}(x1, x2) = \frac{1}{T}\sum_{t=1}^{T} \frac{\text{se}(\hat{t}_{y,t}^{R(x1)}) - \text{se}(\hat{t}_{y,t}^{R(x2)})}{\text{se}(\hat{t}_{y,t}^{R(x2)})} \times 100\%,\ x1, x2 = [+m, -m, me].$$

5.2 The effect of weighting quarterly figures to monthly publication tables

In this subsection, point estimates based on the GREG estimator with a weighting with and without the STM-margin are compared. In this way it is investigated to what extent the extension of the weighting model with the monthly publications affects the point estimates of the quarterly figures. This will be done for the 12 quarters of 2017, 2018 and 2019.

Point estimates for the variable “relation to the labour market” for the GREG estimator with and without the monthly publication tables in the weighting model are compared in Table 5.1 and also presented in Figure 5.1. In this example, the monthly time series predictions correspond to the measurement level of the survey design employed between 2012 and 2021. Consequently, the correction to the quarterly GREG estimates, using the STM margin, is the average RGB across the four follow-up waves during this period. For the employed labour force, estimates based on the first wave are systematically smaller than those based on the follow-up waves. This results in a negative RGB for the employed labour force. Consequently, the quarterly GREG estimates with the STM margin result in a downward correction. For the unemployed labour force, the RGB was small at this time. Therefore, the differences between the GREG estimates with and without the STM margin are small. The other groups comprise different categories of the non-labour force population. For these groups, the RGB is the complement of the RGB for the employed labour force, since the RGB for the unemployed is almost zero. Consequently, the quarterly GREG estimates with the STM margin result in an upward bias correction.

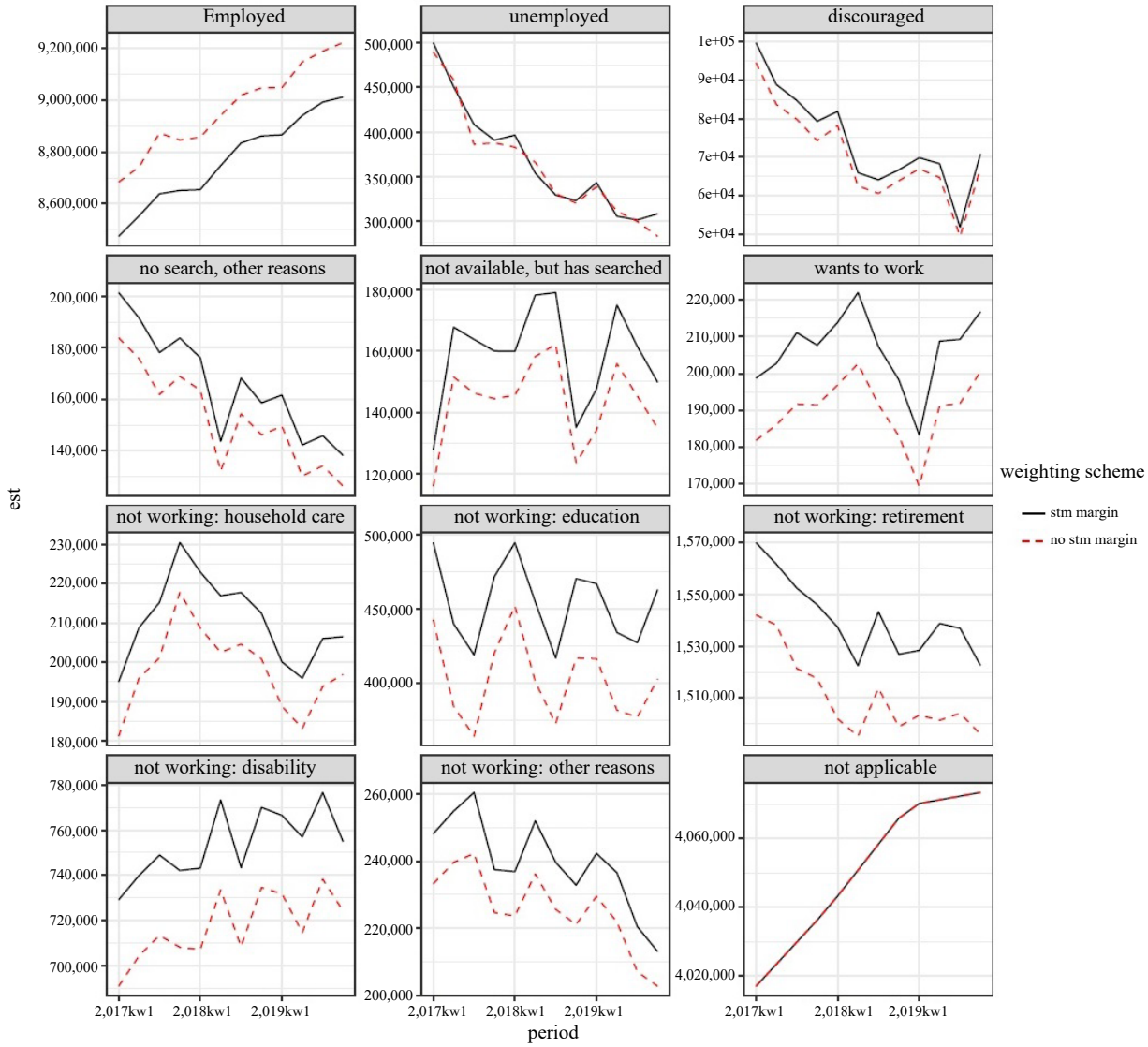
The effect of including the STM-margin in the weighting scheme has pronounced impact for some of the target variables. The relative (absolute) differences are largest for the point estimates of “not working: education” and “not available, but has searched”. They are respectively 9.1 and 7.3 percent. 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. The absolute difference for “employed” is about 130,000 which is nevertheless substantial. This illustrates the need for numerical consistency between the monthly and quarterly estimates, and the extent to which the quarterly estimates benefit from the RGB and discontinuity correction applied to the monthly figures using the time series model.

Table 5.1
Effect on point measures and standard errors for relation to the labour market

Category	AE(+m) counts	AE(-m) counts	RD %	RAD %
Employed	8,768,846	8,897,074	-1.4	1.4
Unemployed	367,392	361,486	1.7	2.5
Discouraged	73,290	70,964	3.2	3.2
No search, other reasons	164,637	155,059	6.2	6.2
Not available, but has searched	152,885	142,442	7.3	7.3
Wants to work	205,787	195,523	5.2	5.2
Not working: household care	217,006	211,615	2.6	2.6
Not working: education	454,345	416,703	9.1	9.1
Not working: retirement	1,538,607	1,514,953	1.6	1.6
Not working: disability	765,337	750,500	2.0	2.0
Not working: other reasons	232,533	224,346	3.6	3.6

Note: Average estimates (AE); average Relative Absolute Difference (RAD); avergage Relative Difference (RD).

Figure 5.1 Quarterly GREG estimates with and without the STM-margin in the weighting scheme for the target variable "relation to the labour market" for the period 2017-2019



Note : Generalized regression (GREG); structural time series model (STM).

5.3 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 (4.3), is evaluated and compared with two different versions of the standard variance approximation of the GREG estimator. The first version relates to the standard errors based on the weighting scheme that includes the STM-margin 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 the table of the STM-margin. The second version are the standard errors based on the weighting scheme where the STM-margin 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 for relation to the labour market are presented in Figure 5.2. The ASE's and RDSE's are summarized in Table 5.2.

It can be seen that the impact of accounting for the uncertainty of the STM-margin through the proposed measurement error correction is pronounced for some variables (compare the green dotted lines with the black solid lines). Categories employed and unemployed show the greatest changes but also for variables not working – education and not working – retirement, the standard errors are underestimated if the uncertainty of the STM-margin is ignored. For the categories employed and unemployed variables the regression coefficients are equal to one for the category that corresponds with the target variable while the remaining coefficients are zero. 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-month averages from the monthly publication tables. See also the discussion around equation (4.4) in Section 4.

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 (compare the green dotted lines with the red broken lines). For the employed the standard errors are roughly 25 percent smaller over the period 2017-2019, for the unemployed 40 percent, as follows from Table 5.2. 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.

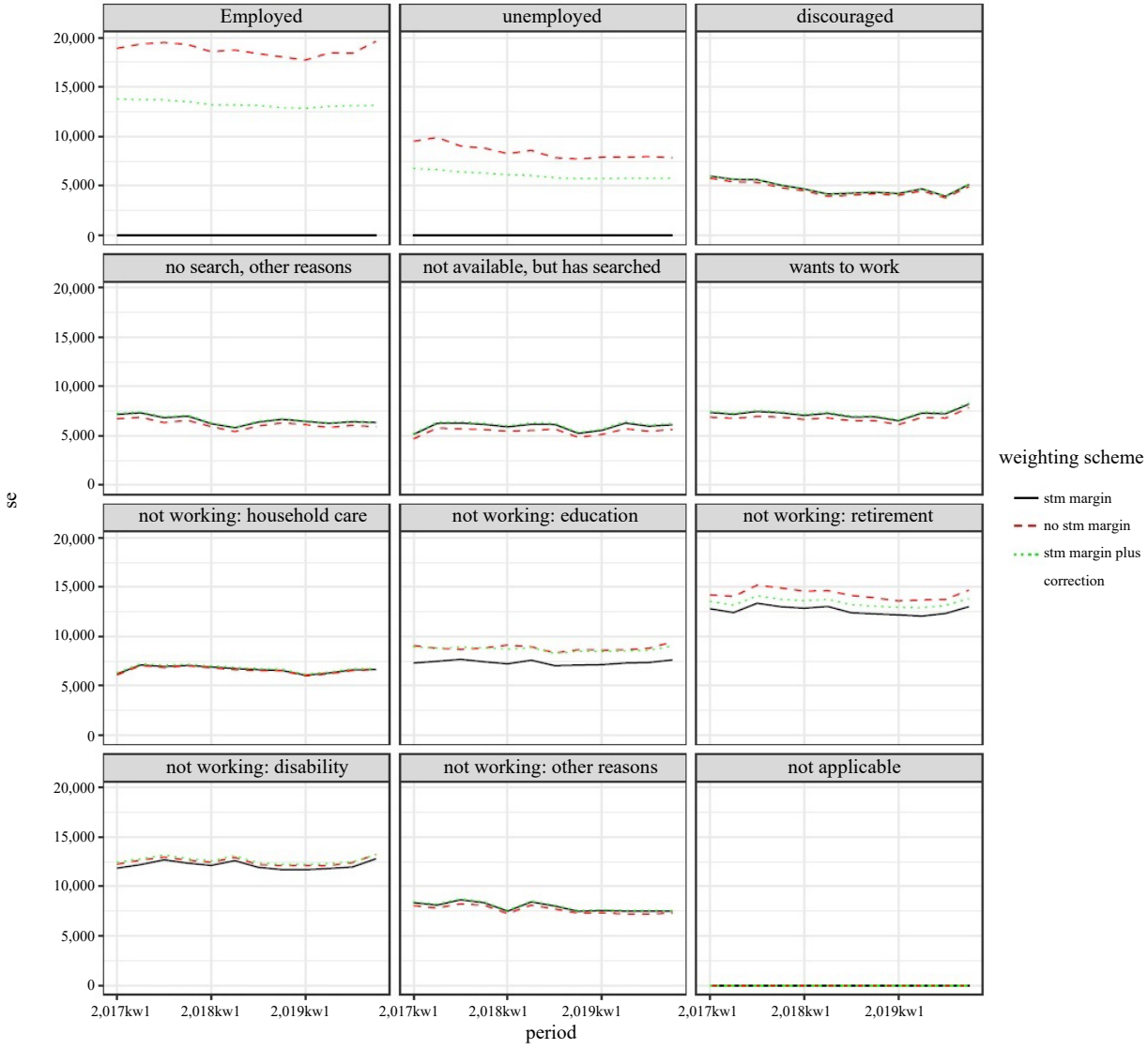
Other categories with a large reduction in standard error are "not working: education" and "not working: retirement". For some of the other categories the extra weighting term leads to an increase in the estimated standard errors; the most notable category in this respect is "not available, but has searched" where the relative increase in standard error is 4.5 percent, as follows from Table 5.2. Turning to the other categories of relation to the labour market, like discouraged or not working: other reasons it can be seen that accounting for the uncertainty of the STM-margin leads to relatively smaller changes in standard errors. This is probably because the relation between these target variables and the population totals in the STM-margin is less strong.

Table 5.2
Effect on point measures and standard errors for relation to the labour market

Category	ASE(+m) counts	ASE(-m) counts	ASE(me) counts	RDSE (+m,-m) %	RDSE (+m,me) %	RDSE (-m,me) %
Employed	0	16,568	13,280	-100.0	-100.0	24.8
Unemployed	0	8,480	6,064	-100.0	-100.0	39.7
Discouraged	4,706	4,586	4,720	2.6	-0.3	-2.8
No search, other reasons	6,623	6,407	6,710	3.4	-1.3	-4.5
Not available, but has searched	5,850	5,597	6,003	4.5	-2.5	-6.7
Wants to work	7,338	7,155	7,447	2.6	-1.5	-3.9
Not working: household care	6,875	6,969	6,967	-1.3	-1.3	0.0
Not working: education	7,522	9,255	8,820	-18.7	-14.7	4.9
Not working: retirement	12,343	13,980	13,321	-11.7	-7.3	4.9
Not working: disability	12,173	12,477	12,496	-2.4	-2.6	-0.2
Not working: other reasons	7,829	7,680	7,914	1.9	-1.1	-2.9

Note: Average Standard Errors (ASE); average Relative Difference between Standard Errors (RDSE).

Figure 5.2 GREG standard errors without (black, red) and with measurement error correction (green) for the target variable "relation to the labour market"



Note : Generalized regression (GREG); structural time series model (STM); standard errors (SE).

6. Conclusion

In this paper a variance approximation is proposed for generalized regression (GREG) estimators that contain population totals 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 totals 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. A related method is the variance approximation that accounts for the uncertainty of using sample estimates for shared variables in the GREG estimator that are based on two or more probability samples, proposed by Renssen and Nieuwenbroek (1997) and Berger et al. (2009).

The situation considered in this paper is different. We consider GREG estimators that contain population totals in the weighting scheme that are included to enforce numerical consistency between the GREG estimates and these population totals, even if the population totals are observed with error. This uncertainty may be due to design-based sampling errors, as in Renssen and Nieuwenbroek (1997) and Berger et al. (2009). However, it may also account for other sources of uncertainty. For example the uncertainty of model-based estimates from time-series or small-area estimation models. One practical situation in which this occurs is when GREG estimates are benchmarked against output levels corrected for rotation group bias (RGB) or discontinuities. If these population totals are based on estimates and are therefore subject to measurement error, then it is important that the variance of the GREG estimates also accounts for this additional uncertainty. Unlike two-phase sampling, the variance of the GREG estimator cannot account for this additional uncertainty because this source of measurement error is unrelated to the sample design. Therefore a variance approximation for the GREG estimator is proposed that contains an additional variance component. It is naturally assumed that the level of uncertainty of the population totals is known a priori.

The method proposed in this paper for accounting for additional uncertainty when benchmarking quarterly GREG estimates against three-month averages of time series model predictions for the Dutch labour force can be applied more generally in situations where estimates are benchmarked against population totals derived from other independent data sources. Wulansari (2026), applied the method as proposed in this paper in the context of small area estimation to approximate the variance of the Fay-Herriot model under complex sampling and measurement error in the covariates.

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-month averages from the monthly publication 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 40%.

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 variance component. 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 three-month averages from the time series model 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. Depending on their stability over time, these standard errors will need to be updated periodically, but not necessarily on a quarterly frequency.

Acknowledgements

The authors would like to thank the anonymous associate editor and the two reviewers for carefully reading our manuscript and providing constructive feedback that proved to be helpful to further improve our paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. This research project is financed by Eurostat under grant agreement 2023-NL-LFS, nr. 101139082.

Appendix

A.1 Weighting model for the monthly labour force figures

In this appendix the weighting models for the GREG estimator used to produce the input series for the time series model in Subsection 3.2 is specified. The numbers between brackets refer to the number of categories that are specified by the auxiliary variable. The monthly GREG estimates per wave that compose the input series for the time series model (3.1) is based on the following weighting scheme:

Nationality(3) + Gender(2) + Age(21) + Householdtype(3) + Region(44) + Age*Gender(7) + UnemplBenefit(5) + Income(3).

Most of the variable names are evident, with the exception of *UnemplBenefit(5)* , which stands for the duration time that a respondent is receiving unemployment benefits in 5 different categories.

A.2 Weighting model for the quarterly labour force figures

In this appendix the weighting models for the GREG estimator used for the quarterly figures in Subsection 3.3 is specified. This model is defined as:

Gender(2)*Nationallity(20) + BigMunicipalities(197)*AgeGender(3) + Gender(2)*Age(43) + Age(5)*Mode(3)*Gender(2) + UnemplBenefit(5) + UnemplBenefit(2)* Province(12) + Income(6) + IncomeType(3)*Province(12) + Householdtype(3) + LMPos(3)*Gender(2)*Age(3).

The numbers between brackets refer to the number of categories that are specified by the auxiliary variable. Most of the variable names are evident, with the exception of:

- AgeGender(3): layout in three classes (<15 or >64, 15-64 men, 15-64 women).
- *LMPos(3)*Gender(2)*Age(3)*: STM-margin or the three-month average from the monthly publication tables Labour market position in three categories for gender and age.

A.3 Proof of formula (4.4)

In this appendix it is shown that the matrix of regression coefficients of the GREG estimator equals (4.4) 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:

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

with $\hat{\mathbf{t}}_y^{R0}$ and $\hat{\mathbf{t}}_y^{\pi}$ 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

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

with $\mathbf{Y}$ an $n \times q$ matrix where each row corresponds to the q target values $(y_{i,1}, \ldots, y_{i,q})$ of sampling unit $i = 1,\ldots,n$, $\mathbf{X}$ an $n \times p$ matrix where each row corresponds to the p auxiliary variables from the weighting scheme $(x_{i,1}, \ldots, x_{i,p})$ of sampling unit $i = 1,\ldots,n$, and $\mathbf{E}$ 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:

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

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

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

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

Proof:

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

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

Elaborating on the matrix operations gives:

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

with

$$\mathbf{X}_{11} = \mathbf{X}_m'\boldsymbol{\Sigma}^{-1}\mathbf{X}_m,$$

$$\mathbf{X}_{21} = \mathbf{X}_a'\boldsymbol{\Sigma}^{-1}\mathbf{X}_m,$$

$$\mathbf{X}_{12} = \mathbf{X}_m'\boldsymbol{\Sigma}^{-1}\mathbf{X}_a,$$

$$\mathbf{X}_{22} = \mathbf{X}_a'\boldsymbol{\Sigma}^{-1}\mathbf{X}_a.$$

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

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

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

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

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

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

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

From this it follows for the first component that

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

and for the second component

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

Collecting both results gives

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

which proves (4.4).

References

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

Berger, Y.G., Muñoz, J.F. and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals – An application to the extended regression estimator and the regression composite estimator. *Computational Statistics and Data Analysis*, 53, 2596-2604.

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.

Hidiroglou, M.A. and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24(1), 11-20. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998001/article/3905-eng.pdf.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Ass*ociation, 47, 663-685.

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

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

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

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

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, 2nd edition*. New York: John Wiley & Sons, Inc.

Renssen, R.H. and Nieuwenbroek, N. (1997). Alligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

Särndal, C.-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

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

Singh, H.P. and Kumar, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers*, 51, 559-582.

van den Brakel, J.A. and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35(2), 177-190. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11040-eng.pdf.

van den Brakel, J.A. and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41(2), 267-296. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14231-eng.pdf.

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

Wulansari, I.Y. (2026). *Small Area Estimation with Measurement Error in Covariates: Simulation Studies and Applications in Official Statistics*. PhD thesis (forthcoming), School of Mathematical and Physical Sciences, University of Technology Sydney, Australia.