Journal of Official Statistics
vol. 33, i. 1 (2017)

Unit Root Properties of Seasonal Adjustment and Related Filters: Special Cases......... p. 1
Bell, William.R.

A Simple Method for Limiting Disclosure in Continuous Microdata Based on Principal Component Analysis.. p. 15
Calviño, Aida

Estimating the Count Error in the Australian Census.. p. 43
Chipperfield, James / Brown, James / Bell, Philip

Space-Time Unit-Level EBLUP for Large Data Sets... p. 61
D’Aló, Michele / Falorsi, Stefano / Solari, Fabrizio

Official Statistics and Statistics Education: Bridging the Gap.................................... p. 79
Gal, Iddo / Ograjenšek, Irena

Three Methods for Occupation Coding Based on Statistical Learning.........................p. 101
Gweon, Hyukjun / Schonlau, Matthias / Kaczmirek, Lars / Blohm, Michael / Steiner, Stefan

Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions.. p. 123
Kaminska, Olena / Lynn, Peter

Effects of Scale Direction on Response Style of Ordinal Rating Scales..................... p. 137
Liu, Mingnan / Keusch, Florian

Design of Seasonal Adjustment Filter Robust to Variations in the Seasonal Behaviour of Time Series.. p. 155
Martelotte, Marcela Cohen / Souza, Reinaldo Castro / Silva, Eduardo Antônio Barros da

Bridging a Survey Redesign Using Multiple Imputation: An Application to the 2014 CPS ASEC
.... p. 187
Rothbaum, Jonathan

Adjusting for Misclassification: A Three-Phase Sampling Approach......................... p. 207
Sang, Hailin / Lopiano, Kenneth K. / Abreu, Denise A. / Lamas, Andrea C. / Arroway, Pam / Young, Linda J.

Changing Industrial Classification to SIC (2007) at the UK Office for National Statistics .... p. 223
Smith, Paul A. / James, Gareth G.

Cost-Benefit Analysis for a Quinquennial Census: The 2016 Population Census of South Africa... p. 249
Spencer, Bruce D. / May, Julian / Kenyon, Steven / Seeskin, Zachary

Estimation when the Covariance Structure of the Variable of Interest is Positive Definite .... p. 275
Théberge, Alain

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 1–14, http://dx.doi.org/10.1515/JOS-2017-0001

Unit Root Properties of Seasonal Adjustment and Related Filters: Special Cases

William R. Bell[1]

Bell (2012) catalogued unit root factors contained in linear filters used in seasonal adjustment (model-based or from the X-11 method) but noted that, for model-based seasonal adjustment, special cases could arise where filters could contain more unit root factors than was indicated by the general results. This article reviews some special cases that occur with canonical ARIMA model based adjustment in which, with some commonly used ARIMA models, the symmetric seasonal filters contain two extra nonseasonal differences (i.e., they include an extra $(1 - B)(1 - F)$). This increases by two the degree of polynomials in time that are annihilated by the seasonal filter and reproduced by the seasonal adjustment filter. Other results for canonical ARIMA adjustment that are reported in Bell (2012), including properties of the trend and irregular filters, and properties of the asymmetric and finite filters, are unaltered in these special cases. Special cases for seasonal adjustment with structural ARIMA component models are also briefly discussed.

Key words: time series; linear filter; ARIMA model-based seasonal adjustment; canonical decomposition.

1. Introduction

Linear filters used in seasonal adjustment contain various unit root factors. Seasonal unit root factors are those of the seasonal summation operator $U_s(B) = 1 + B + \cdots + B^{s-1}$, where B is the backshift operator ($By_t = y_{t-1}$ for any time series y_t) and s is the seasonal period. A filter that contains $U_s(B)$ will annihilate fixed seasonal effects, a desirable property for seasonal adjustment, trend, and irregular filters. The other unit root factors of interest are powers of the differencing operator $1 - B$. A filter that contains $(1 - B)^d$ for $d > 0$ will annihilate polynomials in t up to degree $d - 1$. This is generally the case for seasonal and irregular filters, and it implies that the corresponding seasonal adjustment and trend filters will reproduce polynomials up to degree $d - 1$. This property has been of significant interest historically, as many empirical trend filters were explicitly designed to reproduce polynomials of a certain degree. For example, the symmetric Henderson trend filters will reproduce cubic polynomials (Kenny and Durbin 1982).

[1] U.S. Census Bureau, Washington, DC 20233, U.S.A. Email: William.R.Bell@census.gov

Acknowledgments: We thank David Findley for calling our attention to the special case of the biannual ARIMA(0,0,0)(0,1,0) model. Findley, Lytras, and Maravall (2015, sec. 8.1) explicitly consider its canonical decomposition and give formulas for the symmetric signal extraction filters, which show that the resulting seasonal filter contains an extra $(1 - B)(1 - F)$. This led to our investigations described in Sections 2 and 3.

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Bell (2012) gave general results on unit root factors contained in linear filters used in model-based and X-11 seasonal adjustment. It was noted there that special cases could arise for model-based adjustment where the filters contain more unit root factors than is obvious from the general results. The present article focuses on this point, examining some special cases for canonical ARIMA model-based adjustment (Hillmer and Tiao 1982; Burman 1980; Gomez and Maravall 1996) where the symmetric seasonal filters include two extra differencing operators, written as $(1-B)(1-F)$, where $F=B^{-1}$ is the forward shift operator ($Fy_t=y_{t+1}$). In these cases the symmetric seasonal adjustment filters will reproduce polynomials of two degrees higher than is indicated by the general results given in Bell (2012).

Section 2 defines notation and the framework used for linear model-based seasonal adjustment. Sections 3 and 4 provide results showing when the extra $(1-B)(1-F)$ factor occurs in two models considered explicitly by Hillmer and Tiao (1982), which we hereafter cite as HT: the ARIMA$(0,0,1)(0,1,1)_s$ model and the ARIMA$(0,1,1)(0,1,1)_s$ (airline) model. Values considered for the seasonal period s are 2 (biannual), 4 (quarterly), and 12 (monthly). Section 5 discusses some additional related results for canonical ARIMA model-based adjustment, while Section 6 briefly considers special cases for structural component models. Technical details of the derivations in Sections 3 and 4 are reserved to two Appendices.

2. Notation and Framework for Model-Based Seasonal Adjustment

The additive decomposition used in seasonal adjustment is:

$$y_t = S_t + T_t + I_t \tag{1}$$

where y_t is the observed time series (possibly after transformation, e.g., taking logarithms), and S_t, T_t, and I_t are the seasonal, trend, and irregular components. We also let $N_t = T_t + I_t = y_t - S_t$ denote the nonseasonal component, the estimate of which is known as the seasonally adjusted series. Many of the models proposed for model-based seasonal adjustment use component models that can be written in the following form:

$$\begin{gathered} U_s(B)S_t = u_t \\ (1-B)^d T_t = v_t \\ I_t \sim i.i.d.\,N(0,\sigma_I^2) \end{gathered} \tag{2}$$

where u_t and v_t are stationary time series that are independent of each other and of I_t. Often u_t and v_t are assumed to follow stationary autoregressive-moving average models (Box and Jenkins 1970), in which case y_t follows an ARIMA (autoregressive-integrated-moving average) model that can be written:

$$\phi(B)(1-B)^{d-1}(1-B^s)y_t = \theta(B)a_t \tag{3}$$

where $\phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$ is the AR operator, $\theta(B) = 1-\theta_1 B - \cdots - \theta_q B^q$ is the MA operator, and a_t is white noise, independent and identically distributed $N(0,\sigma_a^2)$. The operators $\phi(B)$ and $\theta(B)$, which may be products of nonseasonal and seasonal polynomials in B, are assumed to have all their zeros outside the unit circle. The expression of the model as presented in (3) requires $d \geq 1$, which is standard in seasonal adjustment practice. Note that $1 - B^s = (1 - B)U_s(B)$ so $(1 - B)^{d-1}(1 - B^s) = (1 - B)^d U_s(B)$.

This model framework covers the ARIMA model-based approach to seasonal adjustment as developed in HT and Burman (1980), and implemented in the TRAMO-SEATS software of Gomez and Maravall (1996) and in the X-13-ARIMA-SEATS program (Monsell 2007). It also covers the structural components models of Harvey (1989), Durbin and Koopman (2001), and Kitagawa and Gersch (1984). Though Harvey did not formulate all his component models in ARIMA form, they can generally be written this way – see Bell (2004).

Let $w_t = (1 - B)^d U_s(B) y_t$ be the differenced observed series. From (1) and (2),

$$w_t = (1 - B)^d u_t + U_s(B) v_t + (1 - B)^d U_s(B) I_t. \tag{4}$$

Let $\gamma_w(k) = \mathrm{Cov}(w_t, w_{t+k})$ and let $\gamma_w(B)$ be the autocovariance generating function (ACGF) of w_t, defined as $\gamma_w(B) \equiv \sum_{k=-\infty}^{\infty} \gamma_w(k) B^k$, where we treat B for this purpose as a complex variable. Given the ARMA model $\phi(B) w_t = \theta(B) a_t$, and the orthogonality of the components in (4), it follows that (Box and Jenkins 1970, 49)

$$\gamma_w(B) = \sigma_a^2 \theta(B)\theta(F)/\phi(B)\phi(F) \tag{5}$$

$$= (1-B)^d (1-F)^d \gamma_u(B) + U_s(B) U_s(F) \gamma_v(B) + (1-B)^d (1-F)^d U_s(B) U_s(F) \sigma_I^2. \tag{6}$$

Given ARMA models for u_t and v_t, analogous expressions to (5) can be given for their ACGFs, $\gamma_u(B)$ and $\gamma_v(B)$. From $w_t = (1-B)^d U_s(B) y_t$, the pseudo ACGF of y_t is defined as $\gamma_y(B) = \gamma_w(B)/[(1-B)^d (1-F)^d U_s(B) U_s(F)]$. We also define $z_t = (1-B)^d N_t = v_t + (1-B)^d I_t$ with ACGF $\gamma_z(B) = \gamma_v(B) + (1-B)^d (1-F)^d \sigma_I^2$.

Bell (1984 and 2012, 445) notes that the minimum mean squared error (MMSE) linear signal extraction estimate of S_t, given the full doubly infinite realization of the series $\{y_t\}$, is

$$\hat{S}_t = \omega_S(B) y_t \quad \text{where} \quad \omega_S(B) = \frac{\gamma_u(B)}{\gamma_w(B)} (1 - B)^d (1 - F)^d. \tag{7}$$

Analogous to (7), the linear filters for the MMSE estimates of N_t, T_t, and I_t are

$$\omega_N(B) = \frac{\gamma_z(B)}{\gamma_w(B)} U_s(B) U_s(F) \tag{8}$$

$$\omega_T(B) = \frac{\gamma_v(B)}{\gamma_w(B)} U_s(B) U_s(F) \tag{9}$$

$$\omega_I(B) = \frac{\sigma_I^2}{\gamma_w(B)} U_s(B) U_s(F) (1 - B)^d (1 - F)^d. \tag{10}$$

Note also that since $\hat{N}_t = y_t - \hat{S}_t$ and $\hat{T}_t = \hat{N}_t - \hat{I}_t$, it follows that $\omega_N(B) = 1 - \omega_S(B)$ and $\omega_T(B) = 1 - \omega_S(B) - \omega_I(B)$.

Simple inspection of (7)–(10) led to the results reported in Bell (2012) for unit root factors contained in these symmetric filters. The specific result of interest here is that $\omega_S(B)$ contains $(1 - B)^d(1 - F)^d$, implying that $\omega_S(B)$ annihilates, and $\omega_N(B)$ thus reproduces, polynomials up to degree $2d - 1$. The models most commonly used in seasonal adjustment have $d = 2$, in which case the symmetric seasonal adjustment filter must reproduce cubic polynomials in t. Less commonly used models have $d = 1$, in which case the symmetric seasonal adjustment filter must reproduce linear polynomials in t. Values of d other than 1 or 2 are uncommon in practice.

Bell (2012, 446–447) also noted that:

> Something not clear from [(7)–(10)] is whether these filters contain additional unit root factors beyond those obvious from inspection. Bell (2010) notes that $\omega_I(B)$ will not include additional unit root factors, while for $\omega_S(B)$, $\omega_N(B)$, and $\omega_T(B)$, additional unit root factors are possible if they appear in the MA polynomials of the ARIMA models for S_t, N_t, or T_t. For example, Hillmer and Tiao (1982, p. 67) examine a model for which the canonical trend component has a factor of $(1 + B)$ in its MA polynomial. While potential additional unit root factors in the filters considered can obviously be examined for any particular model, general results are difficult to give.

The polynomial factors in the MA operator of any ARMA model, such as $\theta(B)$ in (3), correspond to double factors in the numerator of the autocovariance generating function – note $\theta(B)\theta(F)$ in Equation (5). So $1 - B$ is a factor of the MA polynomial of the model for u_t if and only if the numerator of $\gamma_u(B)$ contains $(1 - B)(1 - F)$.

Sections 3 and 4 examine special cases that occur with canonical ARIMA model-based seasonal adjustment where, for two commonly used models, and depending on the seasonal period s and on the model parameter values, $\gamma_u(B)$ indeed contains a factor of $(1 - B)(1 - F)$. From (7), this implies that $\omega_S(B)$ contains an extra $(1 - B)(1 - F)$ so it will annihilate, and $\omega_N(B)$ will reproduce, polynomials in t up to degree $2d + 1$, which is two degrees higher than would otherwise be the case. For the common cases of $d = 1$ or 2, the extra $(1 - B)(1 - F)$ means that the seasonal adjustment filter will reproduce cubic and quintic polynomials, respectively, instead of just linear and cubic polynomials. This property will not be shared by the corresponding trend filter $\omega_T(B) = 1 - \omega_S(B) - \omega_I(B)$ because, as noted in the quotation above, the corresponding canonical irregular filter will not include the extra $(1 - B)(1 - F)$ factor.

3. Results for the ARIMA$(0,0,1)(0,1,1)_s$ Model

The ARIMA$(0,0,1)(0,1,1)_s$ model is

$$(1 - B^s)y_t = (1 - \theta_1 B)(1 - \theta_2 B^s)a_t. \qquad (11)$$

The nonseasonal and seasonal MA parameters θ_1 and θ_2 are both restricted to lie in the interval $(-1, 1)$, though for seasonal adjustment interest focuses on the case of $\theta_2 \geq 0$, for which the existence of the canonical decomposition is assured (HT, 68). Without loss of generality for the derivations and results presented here, we assume that $\text{Var}(a_t) = 1$.

HT's canonical decomposition starts with a partial fractions decomposition of the ACGF for y_t. For the Model (11), HT (p. 68) observe that the seasonal part of this partial fractions decomposition can be expressed as $Q_s^*(B)/U_s(B)U_s(F)$, where

$$Q_s^*(B) = \frac{(1-\theta_2)^2(1-\theta_1 B)(1-\theta_1 F)}{(1-B)(1-F)}\left\{1 - \frac{1}{s^2}U_s(B)U_s(F)\right\}. \qquad (12)$$

Appendix A observes that $1 - \frac{1}{s^2}U_s(B)U_s(F)$ contains $(1-B)(1-F)$, and so can be expressed as $(1-B)(1-F)\alpha_s(B)$, where $\alpha_s(B)$ is a symmetric polynomial in B and F. Appendix A also gives $\alpha_s(B)$ for the cases of $s = 2$, 4, and 12. Cancelling the $(1-B)(1-F)$ factors in the numerator and denominator, $Q_s^*(B)$ simplifies to $(1-\theta_2)^2(1-\theta_1 B)(1-\theta_1 F)\alpha_s(B)$. The spectrum of the canonical seasonal is then $(2\pi)^{-1}$ times $f_s(\lambda) = Q_s^*(e^{i\lambda})/|U_s(e^{i\lambda})|^2 - \epsilon_s$, where

$$\epsilon_s = \min_{\lambda\in[0,\pi]} \frac{Q_s^*(e^{i\lambda})}{|U_s(e^{i\lambda})|^2} = \min_{\lambda\in[0,\pi]} \frac{(1-\theta_2)^2[(1+\theta_1^2) - 2\theta_1\cos(\lambda)]\alpha_s(e^{i\lambda})}{|U_s(e^{i\lambda})|^2}. \qquad (13)$$

The value ϵ_s becomes part of the canonical irregular variance. If the minimum value ϵ_s occurs at $\lambda = 0$, then the resulting canonical seasonal spectrum $(2\pi)^{-1}f_s(\lambda)$ will be zero at $\lambda = 0$, and the pseudo-ACGF of S_t, which is $\gamma_u(B)/U_s(B)U_s(F)$, must include a $1-B$ factor in $\gamma_u(B)$ (so that $\gamma_u(e^{i0}) = \gamma_u(1) = 0$). By symmetry of $\gamma_u(B)$, it must then also include a $1-F$ factor, and so in such cases the canonical seasonal filter $\omega_S(B)$ given by (7) will include an extra $(1-B)(1-F)$ in its numerator. In these cases, the canonical $\omega_S(B)$ for the $(0,0,1)(0,1,1)_s$ model includes in total $(1-B)^2(1-F)^2$. Then $\omega_S(B)$ will annihilate, and $\omega_N(B)$ will reproduce, cubic polynomials in t, not just linear polynomials (the standard result for this model, which has $d = 1$).

For given values of the nonseasonal MA parameter θ_1, the value of λ that minimizes $f_s(\lambda)$ was determined through inspection by computing $f_s(\lambda)$ over a detailed grid of λ values (from 0 to π in increments of .01) and picking off the minimizing value of λ. Examining the results for a detailed set of θ_1 values revealed those values of θ_1 for which the minimum of $f_s(\lambda)$ occurs at $\lambda = 0$, so that $\omega_S(B)$ from the $(0,0,1)(0,1,1)_s$ model contains $(1-B)^2(1-F)^2$ and not just $(1-B)(1-F)$. Table 1 gives the results. Note that for $s = 2$, $\omega_S(B)$ contains $(1-B)^2(1-F)^2$ for any value of θ_1, while for $s = 4$ and $s = 12$, $\omega_S(B)$ contains $(1-B)^2(1-F)^2$ only for limited intervals of θ_1. In fact, the result for $s = 2$ can be established analytically, since it is easy to show that $f_2(\lambda)$ is increasing in λ over $[0,\pi]$ for any value of θ_1. Another point worth noting is that, for $\theta_1 > 0$, the $(1+\theta_1^2) - 2\theta_1\cos(\lambda)$ factor in (13), which does not depend on s, is an increasing function of λ on $[0,\pi]$, while $\alpha_s(e^{i\lambda})/|U_s(e^{i\lambda})|^2$, which does not depend on θ_1, has a global minimum at $\lambda = 0$. Hence, for each s and for all $\theta_1 > 0$, the minimum of $f_s(\lambda)$ occurs at $\lambda = 0$. Finally, note that the results of Table 1 are not affected by the value of θ_2.

To provide further insight into the results of Table 1, Figure 1 shows plots of $f_s(\lambda)$ (but omits the $(1-\theta_2)^2$ factor, since it does not depend on λ) for both the quarterly and

Table 1. Range of values of θ_1 for which the canonical seasonal filter $\omega_S(B)$ from (7) for the ARIMA$(0,0,1)(0,1,1)_s$ *model (11) includes* $(1-B)^2(1-F)^2$, *not just* $(1-B)(1-F)$.

Seasonal period s	2	4	12
Range of values of θ_1	all $\theta_1 \in (-1,1)$	$-.35 < \theta_1 < 1$	$-.28 < \theta_1 < 1$

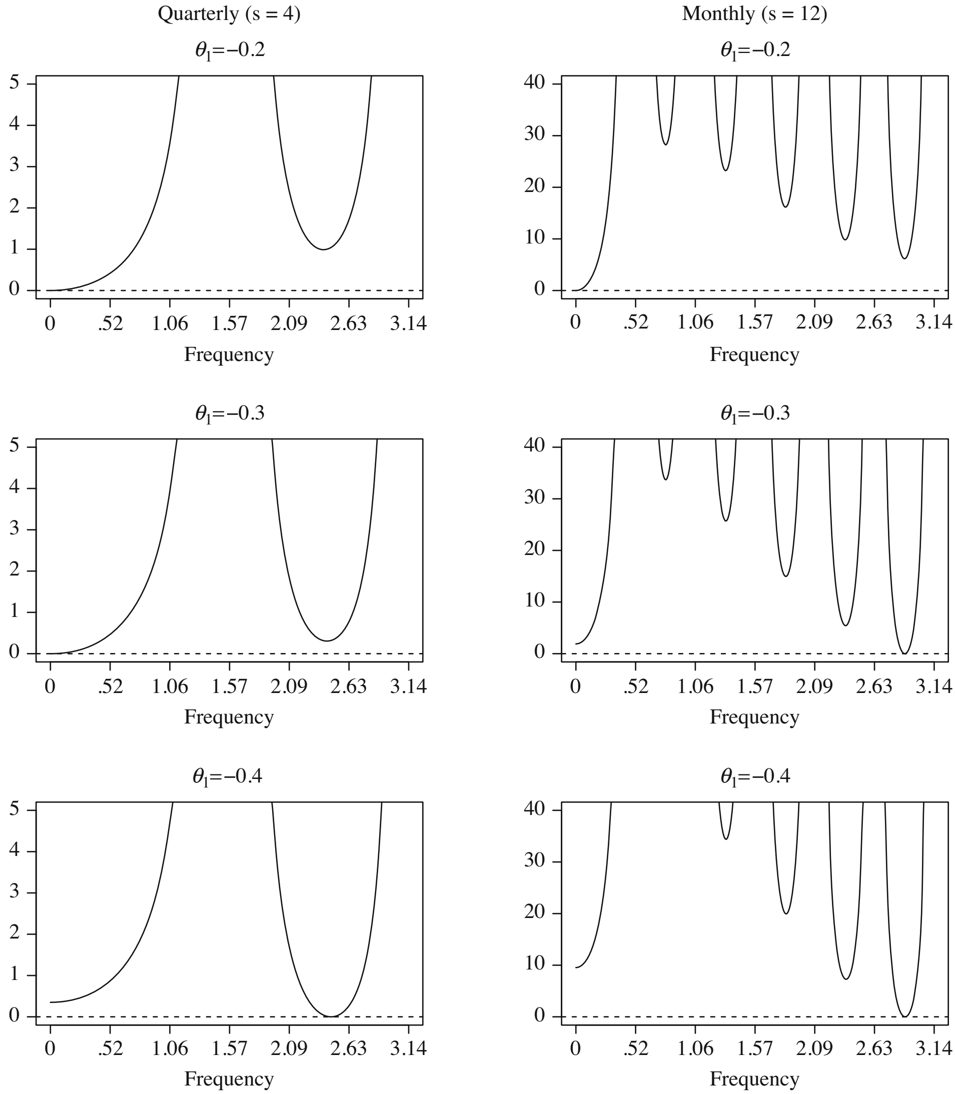


Fig. 1. Plots of the (rescaled) canonical seasonal component spectrum, $f_s(\lambda)/(1-\theta_2)^2$, for the ARIMA(0,0,1)(0,1,1)$_s$ *model. Plots are given for both the quarterly (left) and monthly (right) cases, for three values of* θ_1: $-.2$, $-.3$, *and* $-.4$. *When the minimum of* $f_s(\lambda)$ *occurs at frequency zero, the canonical symmetric seasonal filter includes* $(1-B)^2(1-F)^2$. *When the minimum occurs at a nonzero frequency, the canonical symmetric seasonal filter includes only* $(1-B)(1-F)$.

monthly cases, for three values of θ_1: $-.2$, $-.3$, and $-.4$. Features common to these plots, and to plots of $f_s(\lambda)$ for other values of θ_1, include: a local minimum at $\lambda = 0$; infinite peaks at the seasonal frequencies; and, necessarily, dips between the seasonal frequencies. The plots also show, consistent with Table 1, that (*i*) for $\theta_1 = -.2$, $f_s(\lambda)$ is minimized at $\lambda = 0$ for both the quarterly and monthly cases, (*ii*) for $\theta_1 = -.3$, this occurs for the quarterly but not the monthly cases, and (*iii*) for $\theta_1 = -.4$, this occurs for neither the quarterly nor the monthly cases. In fact, as θ_1 decreases from 1 towards -1, the dips in

$f_s(\lambda)$ between the seasonal frequencies decrease relative to the local minimum at $\lambda = 0$. Eventually, a θ_1 value is reached beyond which the global minimum of $f_s(\lambda)$ occurs at the dip between the last two seasonal frequencies, rather than at $\lambda = 0$. These θ_1 values define the lower limits of the ranges given by Table 1.

4. Results for the ARIMA(0,1,1)(0,1,1)$_s$ (Airline) Model

The ARIMA(0,1,1)(0,1,1)$_s$ (airline) model is (Box and Jenkins 1970, sec. 9.2)

$$(1 - B)(1 - B^s)y_t = (1 - \theta_1 B)(1 - \theta_2 B^s)a_t. \tag{14}$$

As with the (0,0,1)(0,1,1)$_s$ model, the nonseasonal and seasonal MA parameters θ_1 and θ_2 are restricted to lie in the interval $(-1,1)$, though again interest focuses on the case of $\theta_2 \geq 0$, for which existence of the canonical decomposition is assured. We again assume without loss of generality that $\mathrm{Var}(a_t) = 1$.

HT (p. 67) observe that, for y_t following Model (14) with $\theta_2 \geq 0$, the seasonal part of the partial fractions decomposition of $\gamma_y(B)$ can be expressed as $Q_s^*(B)/U_s(B)U_s(F)$, where now

$$\begin{aligned} Q_s^*(B) = {} & \frac{(1-\theta_2)^2}{(1-B)^2(1-F)^2} \\ & \times \left\{ \frac{(1-\theta_1)^2}{4}(1+B)(1+F)\left[1 - \frac{1}{s^2}U_s(B)U_s(F) - \frac{s^2-1}{12s^2}(1-B^s)(1-F^s)\right] \right. \\ & \left. + \frac{(1+\theta_1)^2}{4}(1-B)(1-F)\left[1 - \frac{1}{4s^2}U_s(B)U_s(F)(1+B)(1+F)\right] \right\}. \end{aligned} \tag{15}$$

Appendix B simplifies the expression in braces in (15), showing that both of its terms contain $(1 - B)^2(1 - F)^2$, so that after cancellation with the $(1 - B)^2(1 - F)^2$ of the denominator, $Q_s^*(B)$ simplifies to

$$Q_s^*(B) = (1-\theta_2)^2 \left\{ \frac{(1-\theta_1)^2}{4}(1+B)(1+F)m_{s1}(B) + \frac{(1+\theta_1)^2}{4}m_{s2}(B) \right\}$$

where $m_{s1}(B)$ and $m_{s2}(B)$ are symmetric polynomials given in Appendix B. The spectrum of the canonical seasonal is then $(2\pi)^{-1}$ times $f_s(\lambda) = Q_s^*(e^{i\lambda})/|U_s(e^{i\lambda})|^2 - \epsilon_s$, where now

$$\epsilon_s = \min_{\lambda \in [0,\pi]} \frac{(1-\theta_2)^2}{|U_s(e^{i\lambda})|^2} \left\{ \frac{(1-\theta_1)^2}{4} 2[1+\cos(\lambda)]m_{s1}(e^{i\lambda}) + \frac{(1+\theta_1)^2}{4}m_{s2}(e^{i\lambda}) \right\}.$$

For $s = 2$, 4, and 12, and for a detailed set of values of θ_1, the minima ϵ_s were again determined by inspection, noting cases when the minimum occurs at $\lambda = 0$, so $\gamma_u(B)$ contains $(1 - B)(1 - F)$, implying that $\omega_S(B)$ contains $(1 - B)^3(1 - F)^3$ and not just $(1 - B)^2(1 - F)^2$. Table 2 gives the results which, as for Table 1, are unaffected by the value of θ_2. Analogously to Table 1, we see that, for $s = 2$, $\omega_S(B)$ contains $(1 - B)^3(1 - F)^3$ for any value of θ_1, while for $s = 4$ and $s = 12$, this occurs only for limited intervals of θ_1. This is unsurprising, since plots of $f_s(\lambda)$ (not shown) reveal broadly similar patterns to the plots of Figure 1. However, the limited intervals for $s = 4$ and $s = 12$ given in Table 2 are much

Table 2. Range of values of θ_1 for which the canonical seasonal filter $\omega_S(B)$ from (7) for the ARIMA(0,1,1)(0,1,1)$_s$ *(airline) model (14) includes $(1-B)^3(1-F)^3$, not just $(1-B)^2(1-F)^2$.*

Seasonal period s	2	4	12
Range of values of θ_1	all $\theta_1 \in (-1,1)$	$.11 < \theta_1 < 1$	$.58 < \theta_1 < 1$

smaller than the corresponding intervals given in Table 1, and they exclude some positive values of θ_1.

To illustrate the results of Table 2, the symmetric seasonal filter $\omega_S(B)$ from the canonical decomposition of the quarterly airline model was applied to polynomials of the form $p_t^{(k)} = 100 \times (t-1)^k/30^k$ for $k=4$ and $k=5$. These two polynomials both take the values 0 at $t=1$ and 100 at $t=31$, while at $t=61$, the last time point used, they take the values 1,600 (for $k=4$) and 3,200 (for $k=5$). Figure 2 plots the resulting values of $\omega_S(B)p_t^{(4)}$ for $t=31$ against the value of the airline model parameter θ_1, for values of θ_1 covering the interval $-.5 \leq \theta_1 \leq .5$. The parameter θ_2 was set to zero to minimize the effective length of $\omega_S(B)$, so that its application at the mid-point of the series ($t=31$) would be negligibly affected by the absence of data prior to $t=1$ and after $t=61$. Computations were done with the X-13-ARIMA-SEATS program.

Table 2 says that the values $\omega_S(B)p_t^{(4)}$ should be zero for $\theta_1 > .11$, which is indeed the case in Figure 2. For $\theta_1 < .11$, the values are positive, and they increase as θ_1 decreases further and further below .11. However, considering that the value of $p_t^{(4)}$ is 100 at $t=31$, and increases as t increases past 31, the seasonally filtered values seem quite small. The analogous plot of $\omega_S(B)p_t^{(5)}$ (not shown) is visually identical to Figure 2, but the values of $\omega_S(B)p_t^{(5)}$ are about twice those of $\omega_S(B)p_t^{(4)}$, so they are still small. Thus, even for $\theta_1 < .11$, the symmetric quarterly canonical seasonal filter comes close to reproducing these fourth and fifth degree polynomials.

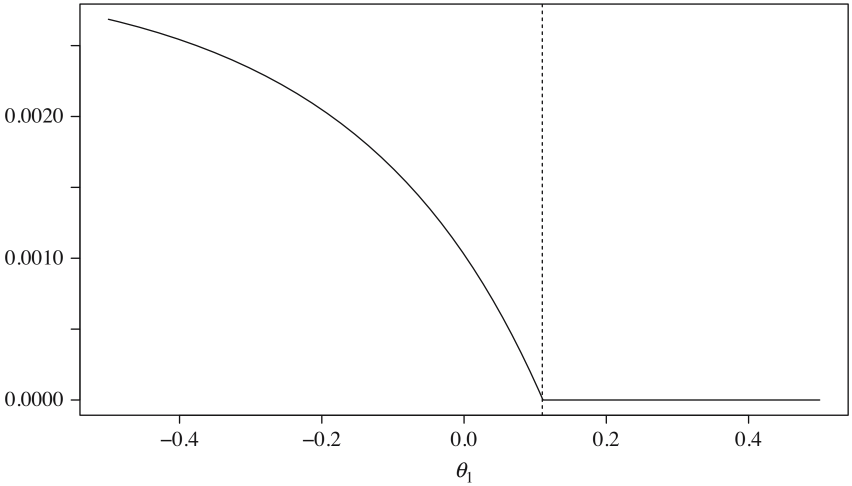


Fig. 2. Canonical decomposition of quarterly airline model for various values of θ_1: Results from applying the symmetric seasonal filter to a fourth degree polynomial, $p_t^{(4)}$, in t. The solid curve shows the values of $\omega_S(B)p_t^{(4)}$ at time point 31 (where $p_{31}^{(4)}$ =100), plotted against the value of θ_1 from the airline model. The dotted vertical line is at $\theta_1 = .11$. See text for further details.

5. Additional Results for Canonical ARIMA Model-Based Seasonal Adjustment

For any particular seasonal ARIMA model for which the canonical decomposition exists, one can obviously check for the presence of additional unit root factors in the various filters by examining the component models from the canonical decomposition. The computations can be done with the original SEATS program (Gomez and Maravall 1996) or the X-13-ARIMA-SEATS program (Monsell 2007), either of which will provide output tables giving the roots of the AR and MA polynomials of the component models. This approach was applied to the $(1,1,0)(0,1,1)_{12}$ model $(1 - \phi B)$ $(1 - B)(1 - B^{12})y_t = (1 - \theta B^{12})a_t$, for a range of values of ϕ and specific values of θ. This revealed that for $\theta = .7$, $\omega_S(B)$ contains an extra $(1 - B)(1 - F)$ factor for $\phi < -.6$, while for $\theta = .8$ this occurs for $\phi \leq -.5$. The dependence of these results on the seasonal MA parameter is in contrast to the results of Tables 1 and 2.

As noted earlier, for models of the form of (2) with $\sigma_I^2 > 0$, extra unit root factors are not present in the symmetric canonical irregular filter, and so the symmetric canonical trend filter will reproduce only polynomials up to degree $2d - 1$, not degree $2d + 1$. For models with $d = 2$ and when $\omega_S(B)$ does contain the extra $(1 - B)(1 - F)$, $\omega_S(B)$ then contains $(1 - B)^3(1 - F)^3$ while $\omega_I(B)$ contains only $(1 - B)^2(1 - F)^2$, so $\omega_N(B)$ reproduces quintic polynomials in t while $\omega_T(B)$ reproduces only cubic polynomials. This matches analogous results for X-11 symmetric filters reported in Bell (2012, 449).

The quotation in Section 2 noted that HT considered a model for which the canonical trend model had a $1 + B$ factor in its MA polynomial. This implies that $\gamma_v(B)$ contains $(1 + B)(1 + F)$, so that $\omega_T(B)$ given by (9) has this extra $(1 + B)(1 + F)$. In fact, HT's derivations for the $(0,0,1)(0,1,1)_s$ and the $(0,1,1)(0,1,1)_s$ models (the latter with $\theta_2 \geq 0$) show that the canonical trend spectrum is minimized at $\lambda = \pi$. Thus, for both these models, $\gamma_v(B)$ contains $(1 + B)(1 + F)$, so that $\omega_T(B)$ contains $U_s(B)U_s(F)(1 + B)$ $(1 + F)$, which includes $(1 + B)^2(1 + F)^2$.

Extra $1 - B$ factors will not be present in asymmetric seasonal filters because application of such filters is equivalent to application of the corresponding symmetric seasonal filter $\omega_S(B)$ after forecast and backcast extension of the time series. Since the forecast and backcast extension will reproduce polynomials only up to degree $d - 1$, this becomes the limiting factor in the degree of polynomials reproduced by the asymmetric seasonal adjustment and trend filters (Bell 2012, 447). The same argument applies to seasonal unit root factors contained in the asymmetric seasonal adjustment, trend, and irregular filters. For example, though we noted above that, for the models examined by HT, $\gamma_v(B)$ contains $(1 + B)(1 + F)$ so that $\omega_T(B)$ includes $(1 + B)^2(1 + F)^2$ instead of just $(1 + B)(1 + F)$, the asymmetric trend filters will include only the single $1 + B$ factor.

The symmetric finite filters (the filters applied at $t = m + 1$ for a time series of length $2m + 1$) provide some further exceptions to the results for both canonical ARIMA and structural component models. For the case of $d = 1$, all the finite seasonal and irregular filters will include $1 - B$, so all will annihilate constants, which are then reproduced by the corresponding finite seasonal adjustment and trend filters (Bell 2012, Table 1). However, the finite symmetric seasonal and irregular filters must, by symmetry, then include $(1 - B)(1 - F)$, so they will annihilate linear polynomials in t, which are then what is reproduced by the symmetric finite seasonal adjustment and trend filters. The symmetry argument extends to odd values of $d > 1$, though values of $d \geq 3$ are seldom used in practice. Finally, since all the finite trend filters include $U_s(B)$, which includes the factor $1 + B$, the symmetric finite trend filters must include $(1 + B)(1 + F)$ (Findley and Martin 2006, 29).

6. Special Cases for Structural Component Models

Special case results for the structural models proposed by the references cited in Section 2 differ from the special case results presented for canonical ARIMA seasonal adjustment. For the structural models, a zero in the spectrum of a component will, in most cases, arise only if model fitting estimates zero for the variance of the component's stationary part – u_t, v_t, or I_t in (2). If that happens, the component becomes deterministic, not stochastic. If $\hat{\sigma}_{\mathrm{I}}^2 = 0$, then $I_t = 0$, so it can be dropped from the model, and $N_t = T_t$. Assuming no other components have variance zero, the previous results on unit root factors in the seasonal and seasonal adjustment filters still apply.

If var(v_t) is estimated to be zero, the fitted model then has $(1 - B)^d T_t = 0$, implying that T_t is a polynomial in t of degree $d - 1$. We cannot leave the component model as $(1 - B)^d T_t = v_t$ with var$(v_t) = 0$ and apply the infinite filter signal extraction formulas (7)–(10) since, from (6), setting $\gamma_v(B) = 0$ will produce a factor of $(1 - B)^d (1 - F)^d$ in $\gamma_w(B)$, violating an assumption that underlies these formulas. Instead, we replace the stochastic component T_t in the model by a polynomial regression function $\beta_0 + \beta_1 t + \cdots + \beta_{d-1} t^{d-1}$. If this form of signal extraction estimation (including regression estimation of the β_js) is applied to a time series y_t that is exactly a polynomial in t of degree $d - 1$ or less, the polynomial will be reproduced in $\hat{T}_t$, and thus also in $\hat{N}_t = \hat{T}_t + \omega_I(B)[y_t - \hat{T}_t]$. This contrasts with the symmetric infinite filter estimates for seasonal adjustment and trend estimation that apply with var$(v_t) > 0$, which reproduce polynomials of degree $2d - 1$. For related discussion on treatment of trend constants, see Bell (2010, 5–6), including the proof given of Theorem 2.

Having var$(v_t) = 0$ is acceptable for finite sample signal extraction, but will produce the same results as modeling T_t as a $d - 1$ degree polynomial regression function. Analogous results to those just described hold if u_t is estimated to have zero variance so S_t becomes fixed seasonal effects. See Harvey (1981) and Bell (1987) for discussion related to these two points.

Special case results are more involved for the local linear trend model of Harvey (1989, 37), which is

$$(1 - B)T_t = \beta_t + \varepsilon_{1t} \quad \text{where} \quad (1 - B)\beta_t = \varepsilon_{2t}$$

with ε_{1t} and ε_{2t} independent white noise series with variances $\sigma_{\varepsilon_1}^2$ and $\sigma_{\varepsilon_2}^2$. To summarize the results, if $\sigma_{\varepsilon_2}^2 > 0$, then $\omega_N(B)$ and $\omega_T(B)$ in (8) and (9) reproduce cubics, while if $\sigma_{\varepsilon_2}^2 = 0$, then signal extraction estimation of N_t and T_t reproduces only linear functions of t. Note that estimating $\sigma_{\varepsilon_2}^2 = 0$ but $\sigma_{\varepsilon 1}^2 > 0$ occurs frequently in practice (Bell and Pugh 1990; Shephard 1993). For further discussion, see Bell (2015).

Appendix A: Derivation Details for the ARIMA(0,0,1)(0,1,1)$_s$ Model

We consider (12):

$$Q_s^*(B) = \frac{(1-\theta_2)^2(1-\theta_1 B)(1-\theta_1 F)}{(1-B)(1-F)}\left\{1 - \frac{1}{s^2}U_s(B)U_s(F)\right\}.$$

Applying $U_s(B)$ or $U_s(F)$ to a constant k yields $s \times k$. Thus, applying $1 - \frac{1}{s^2}U_s(B)U_s(F)$ to 1 yields 0, showing that $1 - \frac{1}{s^2}U_s(B)U_s(F)$ contains a factor $(1-B)$. Since $1 - \frac{1}{s^2}U_s(B)U_s(F)$ has symmetric coefficients, it must also contain $(1-F)$, and so can be expressed as $(1-B)(1-F)\alpha_s(B)$, where the polynomial $\alpha_s(B)$, which is of degree $s-2$ in B and F, also has symmetric coefficients. Cancelling the $(1-B)(1-F)$ factors in the numerator and denominator of $Q_s^*(B)$ then simplifies it to $(1-\theta_2)^2(1-\theta_1 B)(1-\theta_1 F)\alpha_s(B)$.

The coefficients of $\alpha_s(B)$ can be obtained using the following easily verified Lemma on division of polynomials in B by $1-B$ and $1-F$.

Lemma: Let $a(B) = a_0 + a_1 B + \cdots + a_k B^k$ be a polynomial in B of degree $k > 0$. Then

(i) $\frac{a(B)}{1-B} = a_0 + (a_0 + a_1)B + \cdots + (a_0 + \cdots + a_{k-1})B^{k-1} + \frac{(a_0 + \cdots + a_k)B^k}{1-B}$, and

(ii) $\frac{a(B)}{1-F} = a_k B^k + (a_k + a_{k-1})B^{k-1} + \cdots + (a_k + \cdots + a_1)B + \frac{(a_k + \cdots + a_0)}{1-F}$.

If $a_0 + \cdots + a_k = 0$, then $a(B)$ contains $1-B$ (equivalently, contains $1-F$) as a factor.

Note from the Lemma that the coefficients of the $k-1$ degree polynomial that results from dividing $a(B)$ by $1-B$ can be obtained by cumulatively summing the coefficients of $a(B)$ or, for division by $1-F$, by cumulatively summing the coefficients of $a(B)$ in reverse order. Applying this approach to $1 - \frac{1}{s^2}U_s(B)U_s(F)$ yields the following $\alpha_s(B)$ for $s = 2, 4,$ and 12:

$$s = 2: \qquad \alpha_2(B) = \frac{1}{4}$$

$$s = 4: \qquad \alpha_4(B) = \frac{1}{16}[10 + 4(B+F) + (B^2+F^2)]$$

$$\begin{aligned} s = 12: \quad \alpha_{12}(B) = \frac{1}{144}&[286 + 220(B+F) + 165(B^2+F^2) + 120(B^3+F^3) \\ &+ 84(B^4+F^4) + 56(B^5+F^5) + 35(B^6+F^6) + 20(B^7+F^7) \\ &+ 10(B^8+F^8) + 4(B^9+F^9) + (B^{10}+F^{10})]. \end{aligned}$$

Appendix B: Derivation Details for the ARIMA(0,1,1)(0,1,1)$_s$ (Airline) Model

For the airline model, we consider (15):

$$\begin{aligned} Q_s^*(B) = &\frac{(1-\theta_2)^2}{(1-B)^2(1-F)^2} \\ &\times\left\{\frac{(1-\theta_1)^2}{4}(1+B)(1+F)\left[1 - \frac{1}{s^2}U_s(B)U_s(F) - \frac{s^2-1}{12s^2}(1-B^s)(1-F^s)\right]\right. \\ &\left.+\frac{(1+\theta_1)^2}{4}(1-B)(1-F)\left[1 - \frac{1}{4s^2}U_s(B)U_s(F)(1+B)(1+F)\right]\right\}. \end{aligned}$$

We know that $1 - \frac{1}{s^2}U_s(B)U_s(F) = (1-B)(1-F)\alpha_s(B)$ and $(1-B^s)(1-F^s) = (1-B)(1-F)U_s(B)U_s(F)$. The first term in brackets on the right-hand side above is thus $(1-B)(1-F)$ times $\alpha_s(B) - \frac{s^2-1}{12s^2}U_s(B)U_s(F)$. If, for each of the cases $s = 2$, 4, and 12, we cumulatively sum and reverse sum the coefficients of $\alpha_s(B) - \frac{s^2-1}{12s^2}U_s(B)U_s(F)$, the first and last values in this twice-summed sequence are both zero. Thus, from the Lemma, $\alpha_s(B) - \frac{s^2-1}{12s^2}U_s(B)U_s(F) = (1-B)(1-F)m_{s1}(B)$, where $m_{s1}(B)$ is the symmetric polynomial whose coefficients are the nonzero terms of the sequence produced by the summing and reverse summing. For the second term in brackets on the right-hand side above, if we cumulatively sum and reverse sum the coefficients of $1 - \frac{1}{4s^2}U_s(B)U_s(F)(1+B)(1+F)$, we again get zero for the first and last coefficients, so $1 - \frac{1}{4s^2}U_s(B)U_s(F)(1+B)(1+F) = (1-B)(1-F)m_{s2}(B)$ for the symmetric polynomial $m_{s2}(B)$ whose coefficients we just produced. The terms in the second and third lines of the Expression (15) for $Q_s^*(B)$ thus both contain $(1-B)^2(1-F)^2$, and cancelling this with the $(1-B)^2(1-F)^2$ in the denominator shows that

$$Q_s^*(B) = (1-\theta_2)^2\left\{\frac{(1-\theta_1)^2}{4}(1+B)(1+F)m_{s1}(B) + \frac{(1+\theta_1)^2}{4}m_{s2}(B)\right\}.$$

The polynomials $m_{s1}(B)$ and $m_{s2}(B)$ for the cases of $s = 2$, 4, and 12 are given below.

$$s = 2: \quad m_{2,1}(B) = \frac{1}{4} \quad \text{and} \quad m_{2,2}(B) = \frac{1}{16}(6 + B + F)$$

$$s = 4: \quad m_{4,1}(B) = \frac{3}{16}[26 + 16(B+F) + 5(B^2+F^2)]$$

$$m_{4,2}(B) = \frac{1}{64}[44 + 19(B+F) + 6(B^2+F^2) + (B^3+F^3)]$$

$$\begin{aligned} s = 12: \quad m_{12,1}(B) = \frac{1}{1,728}&[16,874 + 16,016(B+F) + 14,091(B^2+F^2) \\ &+ 11,616(B^3+F^3) + 8,988(B^4+F^4) + 6,496(B^5+F^5) \\ &+ 4,333(B^6+F^6) + 2,608(B^7+F^7) + 1,358(B^8+F^8) \end{aligned}$$

$$\begin{aligned} m_{12,2}(B) = \frac{1}{576}&[1,156 + 891(B+F) + 670(B^2+F^2) + 489(B^3+F^3) \\ &+ 344(B^4+F^4) + 231(B^5+F^5) + 146(B^6+F^6) \\ &+ 85(B^7+F^7) + 44(B^8+F^8) + 19(B^9+F^9) \\ &+ 6(B^{10}+F^{10}) + (B^{11}+F^{11})]. \end{aligned}$$

7. References

Bell, W.R. 1984. "Signal Extraction for Nonstationary Time Series." *Annals of Statistics* 12: 646–664. Available at: http://www.jstor.org/stable/2241400 (accessed January 24, 2017).

Bell, W.R. 1987. "A Note on Overdifferencing and the Equivalence of Seasonal Time Series Models With Monthly Means and Models With $(0, 1, 1)_{12}$ Seasonal Parts When $\Theta = 1$." *Journal of Business and Economic Statistics* 5: 383–387. Doi: http://dx.doi.org/10.1080/07350015.1987.10509602.

Bell, W.R. 2004. "On RegComponent Time Series Models and Their Applications." In *State Space and Unobserved Component Models: Theory and Applications*, edited by A.C. Harvey, S.J. Koopman, and N. Shephard. 248–283. Cambridge, UK: Cambridge University Press.

Bell, W.R. 2010. *Unit Root Properties of Seasonal Adjustment and Related Filters (revised 8/30/2011)*. Research Report RRS2010-08. Center for Statistical Research and Methodology, U.S. Census Bureau. Available at: https://www.census.gov/srd/papers/pdf/rrs2010-08.pdf (accessed September 2016).

Bell, W.R. 2012. "Unit Root Properties of Seasonal Adjustment and Related Filters." *Journal of Official Statistics* 28: 441–461. Available at: http://www.jos.nu/Articles/abstract.asp?article=283441 (accessed January 24, 2017).

Bell, W.R. 2015. *Unit Root Properties of Seasonal Adjustment and Related Filters: Special Cases*. Research Report RRS2015-03. Center for Statistical Research and Methodology, U.S. Census Bureau. Available at: https://www.census.gov/srd/papers/pdf/RRS2015-03.pdf (accessed September 2016).

Bell, W.R. and M.G. Pugh. 1990. "Alternative Approaches to the Analysis of Time Series Components." In *Analysis of Data in Time, Proceedings of the 1989 International Symposium*, edited by A.C. Singh and P. Whitridge. 105–116. Ottawa, Ontario: Statistics Canada.

Box, G.E.P. and G.M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Burman, J.P. 1980. "Seasonal Adjustment by Signal Extraction." *Journal of the Royal Statistical Society Series A* 143: 321–337. Doi: http://dx.doi.org/10.2307/2982132.

Durbin, J. and S.J. Koopman. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Findley, D.F., D.P. Lytras, and A. Maravall. 2015. *Illuminating Model-Based Seasonal Adjustment with the First Order Seasonal Autoregressive and Airline Models*. Research Report RRS2015-02. Center for Statistical Research and Methodology, U.S. Census Bureau. Available at: http://www.census.gov/srd/papers/pdf/RRS2015-02.pdf (accessed September 2016).

Findley, D.F. and D.E. Martin. 2006. "Frequency Domain Analyses of SEATS and X-11/12-ARIMA Seasonal Adjustment Filters for Short and Moderate-Length Time Series." *Journal of Official Statistics* 22: 1–34. Available at: http://www.jos.nu/Articles/abstract.asp?article=221001 (accessed January 2017).

Gomez, V. and A. Maravall. 1996. *Programs TRAMO and SEATS: Instructions for the User (Beta Version: September 1996)*. Madrid, Spain: Banco de España. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/96/Fich/dt9628e.pdf (accessed September 2016).

Harvey, A.C. 1981. "Finite Sample Prediction and Overdifferencing." *Journal of Time Series Analysis* 2: 221–232. Doi: http://dx.doi.org/10.1111/j.1467-9892.1981.tb00323.x.

Harvey, A.C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK: Cambridge University Press.

Hillmer, S.C. and G.C. Tiao. 1982. "An ARIMA-Model-Based Approach to Seasonal Adjustment." *Journal of the American Statistical Association* 77: 63–70.

Kenny, P. and J. Durbin. 1982. "Local Trend Estimation and Seasonal Adjustment of Economic and Social Time Series." *Journal of the Royal Statistical Society Series A* 145: 1–28. Doi: http://dx.doi.org/10.2307/2981420.

Kitagawa, G. and W. Gersch. 1984. "A Smoothness Priors-State Space Modeling of Time Series With Trend and Seasonality." *Journal of the American Statistical Association* 79: 378–389. Doi: http://dx.doi.org/10.1080/01621459.1984.10478060.

Monsell, B.C. 2007. "The X-13A-S Seasonal Adjustment Program." In Proceedings of the 2007 Federal Committee On Statistical Methodology Research Conference. November 5–7, 2007. Available at: https://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Monsell-II-B.pdf (accessed September 2016).

Shephard, N. 1993. "Maximum Likelihood Estimation of Regression Models with Stochastic Trend Components." *Journal of the American Statistical Association* 88: 590–595.

Received June 2015
Revised September 2016
Accepted October 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 15–41, http://dx.doi.org/10.1515/JOS-2017-0002

A Simple Method for Limiting Disclosure in Continuous Microdata Based on Principal Component Analysis

Aida Calviño[1,2]

In this article we propose a simple and versatile method for limiting disclosure in continuous microdata based on Principal Component Analysis (PCA). Instead of perturbing the original variables, we propose to alter the principal components, as they contain the same information but are uncorrelated, which permits working on each component separately, reducing processing times. The number and weight of the perturbed components determine the level of protection and distortion of the masked data. The method provides preservation of the mean vector and the variance-covariance matrix. Furthermore, depending on the technique chosen to perturb the principal components, the proposed method can provide masked, hybrid or fully synthetic data sets. Some examples of application and comparison with other methods previously proposed in the literature (in terms of disclosure risk and data utility) are also included.

Key words: Statistical disclosure control; microdata protection; hybrid microdata; masking method; propensity score.

1. Introduction and Motivation

Limiting disclosure risk is a very important and hard task that agencies that publish collected data must deal with. The objective is to prevent users from being able to learn personal information about a certain individual (data can refer to people, enterprises, etc.) from any published data product. Statistical Disclosure Control (SDC) methods provide means of protecting the providers' confidential data, and approaches range from the simplest methods such as noise addition (see Brand (2002) for a survey and comparison of different SDC methods based on noise addition) to more complex ones such as synthetic data generation based on multiple imputation (see Rubin 1993).

Roughly speaking, SDC methods can be divided into three main categories: masking methods, (fully or partially) synthetic data generators, and hybrid data generators (for a full review on SDC methods see Hundepool et al. 2012). Masking refers to the process of producing a modified safe data set from the original, whereas synthetic data generators replace the original data for all or some variables with modeled (synthetic) data designed to

[1] Department of Computer Science and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[2] Department of Statistics and Operations Research III, Complutense University of Madrid, 28040 Madrid, Spain
Email: aida.calvino@ucm.es

Acknowledgments: The author is indebted to Prof. Josep Domingo-Ferrer for his insightful comments, and to the four referees and the Associate Editor for their interesting comments that allowed her to improve the article substantially. This work was partly supported by the Spanish Government through project TIN2014-57364-C2-R (SmartGlacis), the European Commission through project H2020 "CLARUS" and the Government of Catalonia under grant 2014 SGR 537.

preserve specified properties of the original data. Finally, hybrid data generators combine original and synthetic data in order to obtain protected data sets closer to the original one.

Regarding masking methods, Duncan and Pearson (1991) showed that many of them, such as noise addition or microaggregation, fall into the broader category of “matrix masking”, which consists of masking a matrix X as: $\tilde{X} = M_1XM_2 + M_3$, where $\tilde{X}$ is the masked matrix, M_1 is a record-transforming mask, M_2 is a variable-transforming mask, and M_3 is a displacing mask. For instance, simple noise addition is a particular case of matrix masking where M_1 and M_2 are identity matrices and M_3 is a matrix containing realizations of a specific random vector. Another example of matrix masking is microaggregation (which groups similar records together and releases the average record of each group), where M_2 and M_3 are the identity and zero matrices, respectively, and M_1 is a block diagonal matrix such that the elements in a block can be considered similar and, therefore, are aggregated together.

An interesting method that is worth mentioning, and which does not belong to the previous category, is data swapping, which consists of swapping the values of the records univariately. So as to maintain the variance-covariance matrix as close as possible to the original one, rank swapping, a particular case of data swapping proposed by Moore (1996), limits the range of swapping values to the ones whose rank does not differ more than a prespecified threshold. As noted by Muralidhar et al. (2014), data swapping has a high level of user acceptance, as the values themselves do not suffer any modification and the unidimensional distributions are preserved.

As for synthetic data generators, the first and simplest method in this category was proposed by Liew et al. (1985) and consists of releasing a random sample from the statistical distribution underlying the original data set. However, it is not always possible to find the underlying distribution. Alternatively, Rubin (1993) proposed to generate synthetic data sets by means of the multiple imputation methodology, considering the records to be protected as missing and imputing them. Although this method provides data sets with very high utility, it is quite complex (see Raghunathan et al. (2003) or Drechsler (2011) for more information on multiple imputation-based SDC methods). Therefore, simpler alternatives to generate synthetic data have been explored, such as the Information Preserving Statistical Obfuscation (IPSO) proposed by Burridge (2003). The IPSO method basically sustitutes the original confidential variables with draws from a multivariate normal distribution with parameters obtained conditional on the non-confidential variables. Another alternative includes the synthetic data by bootstrap in Fienberg (1994), which has some similarities with the methods in Liew et al. (1985) and Rubin (1993), and consists of releasing a random draw from the smoothed empirical cumulative distribution function of the original data.

When this kind of SDC methods are applied, it is possible to release only synthetic variables (fully synthetic data) or to replace only the most sensitive or identifying variables with synthetic ones and release the remaining original ones (partially synthetic data). The main advantage of synthetic data is that, at first glance, no respondent re-identification seems possible as the data are artificial. However, this is not always true, as synthetic values can be very close to original ones if overfitting takes place. Nevertheless, as stated by Hundepool et al. (2012), the intruder never knows if a unit in the released data was actually in the original data set.

Some authors, as noted in Muralidhar and Sarathy (2008), argue that the utility of synthetic data sets is limited to the properties initially selected by the generating method. For this reason, and as an attempt to obtain synthetic data sets closer to the original ones, hybrid data generators have been proposed. Hybrid data methods combine original and synthetic data and, depending on how the combination is computed, they can lead to data closer to the original data or to the synthetic data (the definition of "combination" is different depending on the author). Interesting references of this kind of methods are Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), both of whom proposed methods that exactly preserve the mean vector and the covariance matrix of the original data and use the IPSO method in Burridge (2003) as a special case.

The method in Muralidhar and Sarathy (2008) (named MS in the sequel) essentially proposes to substitute the confidential variables by a convex linear combination of the original variables and the synthetic ones given by the IPSO method. Regarding the method in Domingo-Ferrer and González-Nicolás (2010), which is known as MicroHybrid (MH), the idea is to obtain the same groups as with the classical microaggregation but, instead of releasing an aggregated value of the records in the group, they propose to apply the IPSO method independently in each group and release the resulting synthetic values. Note that, for both cases, the extreme cases imply releasing the original data set or a partially synthetic one (as only the confidential variables are protected when applying these methods).

The quality of an SDC method depends on the data utility and the disclosure risk of the output data set. However, no general measure exists to evaluate those features as, depending on the posterior use, preservation of some data characteristics may be more or less important (see, for example, Sarathy and Krishnamurty 2002). Mateo-Sanz et al. (2005) proposed the probabilistic information loss index in order to measure the preservation of certain statistics, such as the mean or the variance. Similarly, Woo et al. (2009) proposed to use propensity scores as a measure of global data utility. From the disclosure risk perspective, a very common measure is distance-based record linkage, which was first proposed in Pagliuca and Seri (1999). All these methods will be further explainedand used in Section 4.

In this article we propose a simple method for limiting disclosure in continuous microdata based on Principal Component Analysis (PCA). The use of PCA is not new in the SDC literature. Banu and Nagaveni (2009) proposed a privacy-preserving clustering method that released masked data obtained from projecting the original data onto a transformation matrix built from the principal component loading matrix (which is obtained from a subset of the original data). However, the utility of this method is limited to cluster analysis, as it permits obtaining similar clusters but does not preserve any other statistic.

Instead of perturbing the original variables directly, the proposed method alters the principal components, as they contain the same information but are uncorrelated, which permits working on each component separately. Due to this last fact, the computation cost can be reduced, as it is less computationally intensive to work univariately than multivariately. However, it is important to highlight that the method is only applicable to continuous data, as it makes use of the classic PCA, which also requires continuous variables. Extending the methodology to categorical variables is still an open question.

The main advantages of the proposed method are:

1. Along the lines of the IPSO method in Burridge (2003) and the one proposed by Muralidhar and Sarathy (2008), the proposed method aims to preserve the mean vector and covariance matrix, as they are sufficient statistics for the multivariate normal distribution. Moreover, even in the absence of normality, many parametric statistical analyses (such as linear regression) will lead to the same result if those statistics are preserved.
2. The proposed method is very flexible, as it permits choosing any (univariate or multivariate) SDC method on the principal components as long as they preserve (at least asymptotically) the mean and variance of the principal components. In this sense, the method can provide masked, hybrid, or fully synthetic data sets depending on the choice of method. This is a great advantage as it gives the user the freedom to choose the kind of data set they need.
3. The proposed method is also fast, as the protection can be applied univariately, thus reducing computation times and making the protection process easier and more effective.
4. If one of the variables is a linear combination of the others, the number of components is less than the number of the variables and this variable is not involved in the protection process. This, however, does not represent a problem, because a protected version of this variable can still be obtained by adding the corresponding protected variables.
5. Some of the methods proposed in the literature (such as the IPSO method proposed by Burridge 2003) impose the same level of perturbation in all the variables of the data set. The proposed method, on the other hand, allows a choice of different levels of perturbation by means of the weights of the original variables in the principal components.
6. The proposed method is very simple and, therefore, more accessible to the public than more complex alternatives. Nevertheless, it still leads to a good balance between data utility and disclosure risk, as will be shown in Section 4. Finally, its computational effort is linear in the number of records, making it suitable for large data sets (see Subsection 3.4).

The article is organized as follows. In Section 2, we provide background on Principal Components Analysis, including specific details on procedures. Section 3 is devoted to the proposed model and analysis of its characteristics. Some examples of application and comparison with other methods previously proposed in the literature (in terms of disclosure risk and data utility) are shown in Section 4. Section 5 includes some guidelines for the application of the proposed method. Finally, in Section 6 we give some conclusions and future research lines.

2. Principal Component Analysis

In this section, we review PCA and how it is performed. PCA is a classic statistical technique designed to identify the causes of data variability and to order them by importance. PCA builds a linear transformation that chooses a new orthogonal coordinate system and changes the data coordinates such that the largest variance is captured by the first axis, the second largest by the second, and so on. For a more comprehensive description of PCA, see Jolliffe (2002).

Mathematically speaking, the principal components are found as follows. Let $\boldsymbol{X}$ be the $n \times m$ matrix containing n observations of m random variables with column-wise zero empirical mean. PCA seeks to find a linear transformation of $\boldsymbol{X}$ as

$$\boldsymbol{Y} = \boldsymbol{XW}, \tag{1}$$

where $\boldsymbol{Y}$ is the transformed data and $\boldsymbol{W}$ is the transforming matrix, whose columns ($\boldsymbol{w}_i$) are the unit loading vectors.

The variance of the first principal component ($\boldsymbol{y}_1$) is given by:

$$\boldsymbol{y}_1 = \boldsymbol{w}'_1 \boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}_1. \tag{2}$$

As we look for the most informative component, $\boldsymbol{w}_1$ must lead to a vector $\boldsymbol{y}_1$ with maximum variance, that is,

$$\boldsymbol{w}_1 = \arg \max_{t/\|\boldsymbol{t}\|=1} \{\boldsymbol{t}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{t}\} = \arg \max_{\boldsymbol{t}} \left\{ \frac{\boldsymbol{t}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{t}}{\boldsymbol{t}'\boldsymbol{t}} \right\}. \tag{3}$$

Note that the right-most quotient corresponds to the renowned Rayleigh quotient, which reaches its maximum value λ_{max} (the highest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$) when $\boldsymbol{t}$ is the corresponding eigenvector. Therefore, $\boldsymbol{y}_1$ has maximum variance when $\boldsymbol{w}_1$ equals the eigenvector associated with the highest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$.

Following the same idea, it can be shown that the remaining loading vectors are given by the remaining eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ sorted by importance (variability) according to its corresponding eigenvalue.

To sum up, the principal components of an $n \times m$ matrix $\boldsymbol{X}$ are given by the columns of $\boldsymbol{Y} = \boldsymbol{XW}$ such that the columns of $\boldsymbol{W}$ are the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$.

It is important to highlight the fact that the variables need to be standardized prior to applying PCA, if they are not in the same unit (i.e., height and weight) or if, although being in the same unit, they have greatly varying sizes (i.e., state level population and number of employed persons), as the components are obtained as sums of the original variables.

3. Proposed Method

The proposed method consists of obtaining the principal components of the original data, to later perturb them. The reason to choose to work with the principal components is twofold: first, they are uncorrelated, which permits modifying them independently without perturbing the variance-covariance structure; and second, the components can be sorted by importance, which permits us to choose what components to alter, based on this information. Furthermore, the components are obtained by linear combinations of the original variables and, therefore, when perturbing a certain component, we are perturbing mainly the original variables with corresponding higher weights. There might be cases where the data owner is interested in perturbing a very sensitive variable more than others. In that case, we can analyze the scores (or weights) of the original variables on the components and decide to perturb only those with highest weights for the sensitive variable.

By using this method, all the variables in the data set need to be considered as, otherwise, the correlation structure with other variables could be destroyed. However, as mentioned above, it is possible to alter the sensitive variables more than the nonsensitive ones by means of a careful selection of the components to be altered. Note that it is not possible to leave any variables unchanged (unless no correlation exists between the confidential and nonconfidential variables).

3.1. *Illustrative Example*

In this section we show an example of the PCA-based method, in order to better illustrate the following sections. Consider the data set in Table 1, which consists of three continuous variables X, Y, and Z, such that X and Y are highly correlated, and Z is negatively correlated with both X and Y. Table 1 also contains the principal components (PCs) of the data set. The weights of the variables on the principal components are:

$$\begin{array}{c|rrr} & \text{PC1} & \text{PC2} & \text{PC3} \\ \hline \text{X} & -0.6183 & 0.4407 & 0.6508 \\ \text{Y} & -0.6656 & 0.1468 & -0.7317 \\ \text{Z} & 0.4180 & 0.8856 & -0.2026 \end{array} \tag{4}$$

As already stated, the basis of the method is to perturb the principal components. This perturbation can be made univariately because of the uncorrelation of the principal components (note that the correlation matrix of the PCs in Table 1 is the identity matrix). In order to illustrate the method, we apply data swapping to the first components (the one with highest variance) and undo the transformation given by the matrix in Equation (4). The data swapping process consists of randomly sorting the observations, which is done by generating a random vector whose elements range from 1 to n and by rearranging the observations according to this new order.

The results are shown in Table 2, where the perturbed components are on the left and the resulting variables are on the right.

As can be seen, compared with Table 1, the preservation of means is exact, while standard deviations and correlations are very close to the original values. Regarding protection, there are two records that remain unchanged, while the remaining ones get very different values. The reason why two records have remained unchanged is that there are only ten records altogether and, therefore, the swapping can lead to no changes with a relatively high probability taking into account that we have perturbed only the first component. If more components are perturbed, this probability decreases, as the probability of not swapping an observation equals $\frac{1}{n}$.

Table 3 shows the results obtained if PC1 and PC2 are swapped and when all the three components are swapped. Note that in both cases the level of protection and data utility is high. Moreover, no record is left unchanged and the contribution of PC3 is minimal as it contains less than four percent of the total variability (the records in the right table are not very different from those in the left one).

Table 1. Illustrative example: data and principal components.

	X	Y	Z		PC1	PC2	PC3
	26.20	26.13	85.98		2.3095	0.4611	−0.0835
	98.61	124.85	74.25		−1.1954	1.2161	0.4253
	43.97	60.01	55.26		0.4223	−1.2144	0.3255
	55.67	126.72	68.07		−0.3816	0.0625	−0.6000
	82.51	147.21	58.13		−1.6351	−0.0870	−0.0773
	33.52	87.71	52.91		0.2127	−1.4687	−0.3291
	72.80	118.89	85.34		−0.1328	1.5062	−0.3122
	40.26	32.56	80.02		1.6889	0.3184	0.2680
	38.14	43.15	70.48		1.2865	−0.3333	0.1997
	97.50	165.31	48.00		−2.5749	−0.4609	0.1835
Mean	58.92	93.25	67.84	Mean	0.0000	0.0000	0.0000
Std Dev	26.96	50.24	13.75	Std Dev	1.5261	0.9445	0.3351
Correlation				Correlation			
X	1.0000	0.8664	−0.2417	PC1	1.0000	0.0000	0.0000
Y	0.8664	1.0000	−0.4639	PC2	0.0000	1.0000	0.0000
Z	−0.2417	−0.4639	1.0000	PC3	0.0000	0.0000	1.0000

Table 2. Illustrative example: perturbed principal components (left) and masked data (right).

	PC1′	PC2	PC3		X′	Y′	Z′
	− 2.5749	0.4611	− 0.0835		103.45	181.07	59.35
	− 1.1954	1.2161	0.4253		98.61	124.85	74.25
	− 0.1328	− 1.2144	0.3255		52.75	77.62	52.23
	2.3095	0.0625	− 0.6000		13.11	41.35	82.74
	0.2127	− 0.0870	− 0.0773		53.29	88.59	68.20
	− 1.6351	− 1.4687	− 0.3291		62.74	146.33	42.84
	0.4223	1.5062	− 0.3122		64.02	101.28	88.37
	1.6889	0.3184	0.2680		40.26	32.56	80.02
	− 0.3816	− 0.3333	0.1997		64.52	96.06	61.39
	1.2865	− 0.4609	0.1835		36.43	42.82	69.05
Mean	0.0000	0.0000	0.0000	Mean	58.92	93.25	67.84
Std Dev	1.5261	0.9445	0.3351	Std Dev	27.19	48.06	14.21
Correlation				Correlation			
PC1′	1.0000	0.0583	− 0.1563	X′	1.0000	0.8664	− 0.2857
PC2	0.0583	1.0000	0.0000	Y′	0.8664	1.0000	− 0.4970
PC3	− 0.1563	0.0000	1.0000	Z′	− 0.2857	− 0.4970	1.0000

Table 3. Illustrative example: masked data when PC1 and PC2 have been swapped (left) and all the principal components have been swapped (right).

	X''	Y''	Z''		X'''	Y'''	Z'''
	93.05	174.62	48.71		97.50	165.31	48.00
	90.10	119.57	65.53		87.48	125.05	65.94
	83.42	96.65	83.66		72.52	119.48	85.38
	8.65	38.58	78.17		25.72	2.83	75.46
	57.86	91.43	72.89		57.75	91.65	72.90
	65.61	148.11	45.77		74.41	129.66	44.38
	47.75	91.18	71.69		42.96	101.22	72.45
	20.12	20.06	59.38		10.46	40.29	60.91
	67.30	97.79	64.23		62.69	107.45	64.96
	55.34	54.55	88.42		57.70	49.60	88.04
Mean	58.92	93.25	67.84	Mean	58.92	93.25	67.84
Std Dev	27.94	47.38	14.00	Std Dev	26.78	48.73	14.24
Correlation				Correlation			
X''	1.0000	0.8258	−0.2354	X'''	1.0000	0.8652	−0.2847
Y''	0.8258	1.0000	−0.5947	Y'''	0.8652	1.0000	−0.4964
Z''	−0.2354	−0.5947	1.0000	Z'''	−0.2847	−0.4964	1.0000

3.2. Mathematical Formulation

Let $\boldsymbol{X}$ be the $n \times m$ matrix containing the n observations of the m random variables such that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the $\boldsymbol{X}$'s mean and standard deviation vectors, respectively.

In order to correctly apply PCA, we first need to standardize $\boldsymbol{X}$. In matrix form this is equivalent to:

$$\boldsymbol{X}_1 = (\boldsymbol{X} - \boldsymbol{M})\boldsymbol{S}, \tag{5}$$

where $\boldsymbol{X}_1$ is the standardized data set, $\boldsymbol{M}$ is an $n \times m$ matrix with rows equal to $\boldsymbol{\mu}$ and $\boldsymbol{S}$ is a m-diagonal matrix with non-zero elements equal to $s_{jj} = 1/\sigma_j$.

Next, we obtain the scores $\boldsymbol{X}_2$ of the observations on the principal components as:

$$\boldsymbol{X}_2 = \boldsymbol{X}_1\boldsymbol{A}, \tag{6}$$

where $\boldsymbol{A}$ is a matrix whose columns are the normalized (to unit vectors) eigenvectors of $\boldsymbol{X}'_1\boldsymbol{X}_1$.

As already stated, instead of perturbing the data directly, the basis of the proposed method consists of perturbing the scores $\boldsymbol{X}_2$. Mathematically speaking, the perturbation process can be written as:

$$\boldsymbol{X}_3 = \boldsymbol{X}_2\boldsymbol{B} + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{B}$ is an $m \times m$ diagonal matrix with ones in the rows corresponding to the m_c components left unchanged and zeros otherwise and $\boldsymbol{\epsilon}$ is an $n \times m$ matrix with m_c columns equal to zero and the remaining (m_r) ones contain the variables that are replacing the original principal components.

Finally, we obtain the masked data $\tilde{\boldsymbol{X}}$ by undoing the PCA transformation and recovering the original means, variances and covariances:

$$\tilde{\boldsymbol{X}} = (\boldsymbol{X}_3\boldsymbol{A}'\boldsymbol{S}^{-1}) + \boldsymbol{M}, \tag{8}$$

where $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$ and $\boldsymbol{S}^{-1}$ is the inverse of $\boldsymbol{S}$. Recall that matrix A is composed of normalized eigenvectors and, therefore, its inverse equals its transpose.

It is important to highlight that any SDC method can be applied as long as it preserves the properties of the original principal components, that is, the mean vector is equal to zero and it has the same diagonal variance-covariance matrix (note that the preservation of higher moments in the perturbed components leads to a better data utility of the final data set). The "protected" components obtained when applying the SDC method form matrix $\boldsymbol{\epsilon}$ in Equation (7). Some examples of possible SDC methods are random scores, noise addition, and swapping.

An interesting feature of this method is that depending on the method chosen to perturb the principal components, we can get masked, hybrid or fully synthetic data. In particular, if a masking method is chosen to alter the principal components, such as data swapping (Moore 1996), the resulting data set is masked, as the original values have been modified but not substituted. Similarly, if the components are substituted by random vectors (by means of the methodology in Liew et al. (1985) or Fienberg (1994)), we get fully synthetic data sets if no component is left unchanged and hybrid data sets, otherwise.

As previously noted, the principal components can be sorted by importance based on quantity of variability and, therefore, perturbing the first component does not lead to the same level of protection and data utility of the output data as perturbing the last component. The larger the total weight of altered components, the lower the data utility and disclosure risk.

Note that one remarkable difference with regard to this method compared to others previously suggested is that the level of perturbation is fixed, in the sense that we can only modify a certain number of components between 1 and *m*, whereas in other methods, such as noise addition, any quantity of noise can be added. It is important to take into account that, although the level of perturbation of the components is not limited, the effect on the final data set is, as the weight of the components is fixed. In other words, if only the "last" component is perturbed, even if the protected component has no resemblance to the original one, the effect on the final data set will be small, as only a small portion of the original variability has been changed.

3.3. Verifying the Preservation of the Mean Vector and the Variance-Covariance Matrix

Preserving the mean vector and the variance-covariance matrix of the original data is a very important feature of a masking method which is very common in the SDC literature. For this reason, we will now show that the proposed method preserves both the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}_X$.

So as to facilitate the computations, we first show the direct relation between $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, following from (5)-(8):

$$\tilde{\boldsymbol{X}} = [(\boldsymbol{X}-\boldsymbol{M})\boldsymbol{SAB}+\boldsymbol{\epsilon}]\boldsymbol{A}'\boldsymbol{S}^{-1}+\boldsymbol{M} = (\boldsymbol{X}-\boldsymbol{M})\boldsymbol{SABA}'\boldsymbol{S}^{-1}+\boldsymbol{\epsilon A}'\boldsymbol{S}^{-1}+\boldsymbol{M}. \quad (9)$$

First, we deal with the mathematical expectation of the masked data set $\tilde{\boldsymbol{X}}$:

$$\begin{aligned} E[\tilde{\boldsymbol{X}}] &= E[(\boldsymbol{X}-\boldsymbol{M})\ \boldsymbol{SABA}'\boldsymbol{S}^{-1}+\boldsymbol{\epsilon A}'\boldsymbol{S}^{-1}+\boldsymbol{M}] \\ &= (E[\boldsymbol{X}]-E[\boldsymbol{M}])\ \boldsymbol{SABA}'\boldsymbol{S}^{-1}+E[\boldsymbol{\epsilon}]\boldsymbol{A}'\boldsymbol{S}^{-1}+E[\boldsymbol{M}] \qquad (10) \\ &= (\mu-\mu)\ \boldsymbol{SABA}'\boldsymbol{S}^{-1}+\boldsymbol{0}\boldsymbol{A}'\boldsymbol{S}^{-1}+\mu = \mu. \end{aligned}$$

Next, we focus on the variance of $\tilde{\boldsymbol{X}}$. From now on, $\boldsymbol{\Sigma}_X$ refers to the variance-covariance of data set $\boldsymbol{X}$.

$$\begin{aligned} \boldsymbol{\Sigma}_{\tilde{X}} &= (\boldsymbol{SABA}'\boldsymbol{S}^{-1})'\boldsymbol{\Sigma}_X(\boldsymbol{SABA}'\boldsymbol{S}^{-1})\ +(\boldsymbol{A}'\boldsymbol{S}^{-1})'\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{A}'\boldsymbol{S}^{-1})+\boldsymbol{0} \\ &= (\boldsymbol{S}^{-1})'\boldsymbol{A}\boldsymbol{\Sigma}_1\boldsymbol{A}'\boldsymbol{S}^{-1}+(\boldsymbol{S}^{-1})'\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{A}'\boldsymbol{S}^{-1} \qquad (11) \\ &= (\boldsymbol{S}^{-1})'\boldsymbol{A}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\boldsymbol{A}'\boldsymbol{S}^{-1}. \end{aligned}$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{S}'\boldsymbol{\Sigma}_X\boldsymbol{SAB} = \boldsymbol{B}'\boldsymbol{\Sigma}_{X_2}\boldsymbol{B}$.

Without loss of generality, we assume that the components have been sorted in such a way that the ones that remain unaltered are the first ones and the altered ones are the last ones. Note that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$have a special block structure as shown below. This is due to the fact that $\boldsymbol{\Sigma}_1$ is obtained by multiplying and premultiplying $\boldsymbol{\Sigma}_{X_2}$, which is a diagonal matrix as it refers to the variance-covariance matrix of the principal components, by matrix $\boldsymbol{B}$, which is a diagonal matrix with zeros in the rows associated to altered components. On the other hand, $\boldsymbol{\epsilon}$ has been defined to be an $n \times m$ matrix with m_c columns equal to zero and the remaining (m_r) ones replacing the original principal components. Therefore, the variance and covariance associated with the "zero" columns are also zero and we have forced the altered components to maintain the variances of the original principal components. Then,

$$\boldsymbol{\Sigma}_1 = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{X_2}[m_c] & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} \end{array} \right), \tag{12}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \left(\begin{array}{c|c} \boldsymbol{0} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{X_2}[m_r] \end{array} \right), \tag{13}$$

where, abusing notation, we have defined $\boldsymbol{\Sigma}_{X_2}[m_c]$ and $\boldsymbol{\Sigma}_{X_2}[m_c]$ to be the submatrices associated to the m_c unaltered components and m_r to the altered components, respectively.

Taking (12) and (13) into account, we have

$$\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{X_2}[m_c] & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{X_2}[m_r] \end{array} \right) = \boldsymbol{\Sigma}_{X_2}, \tag{14}$$

and, thus,

$$\boldsymbol{\Sigma}_{\tilde{X}} = (\boldsymbol{S}^{-1})' \boldsymbol{A} \boldsymbol{\Sigma}_{X_2} \boldsymbol{A}' \boldsymbol{S}^{-1} = \boldsymbol{\Sigma}_X. \tag{15}$$

The perturbing method chosen for the principal components determines whether the first and second moments are exactly or asymptotically preserved. For example, if we choose to substitute component i with a realization of a random variable with expectation and variance equal to μ_i and σ_i^2, respectively, it is very unlikely that the observed mean and variance equals μ_i and σ_i^2. In that case, the original mean vector and variance-covariance matrix are asymptotically (but not exactly) preserved. On the other hand if we choose swapping – as the values are the same – the mean is preserved exactly. However, this is not the case with the variance-covariance matrix, as the covariance of the perturbed principal components is not numerically zero.

Although, as previously stated, many perturbing methods are valid, we suggest using data swapping and random draws from the smoothed empirical cumulative distribution functions (see Fienberg 1994) for masking and hybrid/synthetic data sets, respectively, as they permit maintaining not only the first and second moments but the whole distribution on the univariate components. Note that rank swapping is not necessary here, as the components are uncorrelated and, therefore, the covariance structure does not need to be preserved. This, in turn, helps to preserve the distribution of the original data. Moreover, note that when the number of altered components is high, the resulting records do not clearly represent any of the original records. These are obtained by means of the components' scores of other records, selected randomly and, thus (although the resulting data set is not strictly synthetic), it does bear some similarity to the synthetic data generation philosophy.

3.3.1. On the Preservation of the Third Moment

In this section we deal with the preservation of the third moment (μ_3), which is related to the symmetry of the variables. We remind the reader that the third moment is given by:

$$\mu_3(\boldsymbol{X}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))^3], \tag{16}$$

and, if $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{Z}$, then it holds that:

$$\begin{aligned} \mu_3(\boldsymbol{X}) &= \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}) - 6Cov(\boldsymbol{Y},\boldsymbol{Z})(E(\boldsymbol{Y}) + E(\boldsymbol{Z})) \\ &\quad + 3(Cov(\boldsymbol{Y},\boldsymbol{Z}^2) + Cov(\boldsymbol{Y}^2,\boldsymbol{Z})). \end{aligned} \tag{17}$$

The proposed method essentially decomposes the original variables into some uncorrelated ones and then undoes the decomposition. Without loss of generality, let's assume that we want to preserve the third moment of an original variable $\boldsymbol{X}$, whose principal components are $\boldsymbol{Y}$ and $\boldsymbol{Z}$. As the principal components are uncorrelated, it holds that:

$$\mu_3(\boldsymbol{X}) = \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}) + 3\,(Cov(\boldsymbol{Y},\boldsymbol{Z}^2) + Cov(\boldsymbol{Y}^2,\boldsymbol{Z})), \tag{18}$$

and, therefore, the third moment of $\boldsymbol{X}$ is preserved as long as the addends on the right-hand side are preserved. If a perturbation method is chosen such that the third moment of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are preserved (as univariate data swapping), then the preservation of $\mu_3(\boldsymbol{X})$ depends on the preservation of $Cov(\boldsymbol{Y},\boldsymbol{Z}^2)$ and $Cov(\boldsymbol{Y}^2,\boldsymbol{Z})$, which, in turn, depends on the number and weight of the perturbed components.

It is important to highlight that, if the principal components are independent (and not only uncorrelated), it holds that:

$$\mu_3(\boldsymbol{X}) = \mu_3(\boldsymbol{Y}) + \mu_3(\boldsymbol{Z}), \tag{19}$$

and, in that case, the preservation of the third moment of the original variables can be ensured, as the perturbed principal components are also independent. If multivariate normality holds, the principal components are also normal (as a consequence of the infinite divisibility of the normal distribution) and, in that case, uncorrelation implies independence. However, although generally uncorrelation does not imply independence, multivariate normality is not the only case.

To sum up, the preservation of the third moment depends on the preservation of the third moment of the perturbed principal components, as well as on the independence of the principal components.

3.4. *Computational Effort*

As has been shown, the proposed method essentially consists of obtaining the eigenvectors of the correlation matrix (or equivalently the variance-covariance matrix of the standardized data set) and then applying products and/or sums to the original matrix data $\boldsymbol{X}$ and the transformation matrix $\boldsymbol{A}$. Finally, there might be a random number generation phase associated with the altering components phase. Therefore, the running time of the method is $O(nm^2)$, where n is the number of records and m is the number of variables.

Generally, the number of records is much larger than the number of variables and, therefore, the proposed method is suitable for large data sets.

4. Empirical Results

In this section a simulation study for the PCA-based method is shown. In particular, we evaluate its performance in terms of data utility and disclosure risk in two scenarios: a) when it is applied to get a masked data set that protects all the variables in the data set, and b) when only a subset of variables needs to be protected and the output is a hybrid data set.

The results have been obtained using R project (R Core Team 2014) and, more specifically, package sdcMicro (Templ 2008) when possible. Some ad hoc functions and programs also needed to be developed. Data sets *Tarragona* and *Census* have also been provided by this package. Regarding computation times, anonymizing a data set (of up to 55,000 records and 35 variables) by means of the proposed method takes less than one second on a Toshiba satellite L50-B-11W laptop, Intel Core i7-870 1.8GHz, 4MB, RAM: 8GB.

4.1. Fully Masked Data Set

In this case, the proposed method is applied to the *Tarragona* data set in Brand et al. (2002). It consists of 13 quantitative variables associated to 832 real companies in the province of Tarragona in 1995.

For the sake of completeness, we compare the results derived from the proposed method with two well-known masking methods that have been identified as well-performing in terms of data utility and disclosure risk: rank swapping (see Domingo-Ferrer and Torra (2001) or Jiménez et al. (2014)) and microaggregation plus noise addition (see Oganian and Karr (2006), or Woo et al. (2009)).

Following the ideas of Domingo-Ferrer and Torra (2001) or Jiménez et al. (2014), we compute a score as the mean average of disclosure risk and data utility in order to be able to compare the three methods. The disclosure risk and data utility measures also consist of a score made of the mean average of two different criteria. In particular, disclosure risk is evaluated by means of distance-based record linkage and interval disclosure (see Domingo-Ferrer and Torra 2004), while data utility is computed based on the Probabilistic Information Loss (PIL) measure proposed by Mateo-Sanz et al. (2005) and the propensity scores proposed by Woo et al. (2009). In the following we briefly described these four measures:

- ***Distance-based record linkage (DBRL)***: DBRL is one of the most common methods for evaluating the disclosure risk of a masked data set. It consists of obtaining the closest masked record (in terms of normalized euclidean distance) to all original records and determining how many of them were generated by the corresponding original record. As noted in Domingo-Ferrer and Torra (2004), variables should be standardized when using distance-based record linkage in order to avoid scaling problems. This index can take values between zero percent (no record linkage) and 100% (total record linkage).
- ***Interval disclosure (ID)***: It consists of determining the proportion of original records that lay in an interval whose center is the corresponding masked record. The extremes

of the interval are given by the two masked values whose ranks differ ± p percent of the total number of records. The measure associated with ID is obtained by averaging this proportion for p, taking values from one percent to ten percent with one percent increments. In the simulation study that follows, we have substituted the corresponding masked record by the closest record in order to be able to analyze the attribute disclosure, that is, how much the intruder can learn if re-identification takes place. Moreover, as noted previously, when the number of altered components is large, masked records have a very weak connection with the corresponding original ones (this also happens for microaggregation plus noise addition when the parameter is large). In this way, we can work with an homogeneous measure independently of the parameters of the method. This index also takes values between zero percent (no attribute disclosure) and 100% (total attribute disclosure).

- ***Probabilistic Information Loss (PIL)***: It consists of evaluating, from a probabilistically point of view, the information loss suffered from the masking processes based on the observed difference between some statistics obtained from the original and the masked data set. Given a certain parameter θ and its masked value $\hat{\theta}$, the probabilistic information loss can be measured as the standardized sample discrepancy as follows:

$$pil(\theta) = 2 \cdot P\left(0 \leq N(0,1) \leq \frac{|\theta - \hat{\theta}|}{\sqrt{Var(\hat{\theta})}}\right), \tag{20}$$

where N(0,1) is a standardized normal distribution. The variances of the considered statistics are given in Mateo-Sanz et al. (2005). The final PIL measure is given by the mean average of the *pil* associated with the means, variances, covariances, Pearson'scorrelation coefficients and quantiles. The *pil* given by Equation (20) takes values between 0 (no information loss) and 1 (total information loss) and, therefore, the total PIL also takes values in that range.

- ***Propensity scores (PS)***: Propensity scores were adopted from the statistical literature by Woo et al. (2009) in an attempt to define new global measures of data utility. In the observational study literature, propensity scores are the probabilities of being assigned to a treatment, given other variables (covariates). When two large groups have the same distributions of propensity scores, the groups should have similar distributions on the covariates. Therefore, one can consider the masked data as the treatment and estimate the probability of being assigned to the treatment (the propensity scores) for both the masked and the original data sets. If the distributions of the propensity scores of both sets are similar, we can conclude that the distributions of the original and masked data are also similar and, therefore, the data utility should be relatively high. Woo et al. (2009) propose to evaluate the similarity of the propensity scores using the following formula:

$$PS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - 0.5)^2, \tag{21}$$

where $N = 2n$ and $\hat{p}_i$ is the propensity score for unit i. Note that when the original and the masked data sets have similar distributions, it is difficult to distinguish them and, therefore, the propensity scores are close to 0.5 and *PS* equals 0. On the other hand, if the distributions are very different, they are perfectly distinguishable and the propensity scores for the original and masked data take approximately value 0 and 1, respectively. In that case, *PS* is close to 0.25. The main disadvantage of this method is that it relies on the choice of the model used to estimate the propensity scores. Nevertheless, the authors suggest to use a logistic regression using a second-order polynomial in all the variables, as well as in their interactions. In this article, we have considered the model suggested by the authors as well as a logistic regression with third-order polynomials in all the variables, and the same interactions.

It is important to highlight that we are considering all the variables at the same time in the disclosure risk phase, which is equivalent to assuming that the intruder has information about all the variables in the data set. Therefore, as this is rarely the case, the disclosure indexes should be taken as worst case ones. However, this is done with all the three methods and, thus, this fact will not affect the conclusions derived from the comparison.

On the other hand, all the indexes, except for the propensity scores, lay between 0 and 1 and, thus, when computing their mean average we get a score that also lies between 0 and 1. To overcome the inconvenience of the propensity score, we multiply it by a factor of four, thus obtaining an index that takes values between 0 and 1. Finally, note that the larger the score is, the worse the method is, as it has higher disclosure risk and less data utility.

We have decided to rely on several data utility and disclosure risk indexes, as each of them measures a different concept and permits obtaining a global score. Table 4 shows the results, including the four indexes explained previously (the propensity score is shown with its two variants) and the resulting score (two different scores arise because of the propensity score), of a simulation study performed in order to be able to evaluate the results of our method and compare it with microaggregation plus noise and rank swapping. As the three methods depend on random values in one way or another, and thus, very good or bad results can be obtained by chance and mask the real behavior of the method, we have to resort to a simulation study, which takes into account several realizations.

In order to be able to determine the better performance of a method, comparing the mean values of the score is not enough. For this reason, we have computed the confidence interval (CI) of the scores of the methods along with the mean average, based on 100 realizations. As the score does not belong to any known distribution, we resort to bootstrap techniques, in particular to the Percentile Bootstrap CI. The bootstrap method, which is one of a broader class of resampling methods, uses Monte Carlo sampling to generate an empirical sampling distribution of the estimate (see Efron and Tibshirani (1993) for more details on bootstrap methods).

For the PCA-based method, *parms* refers to the number of components that have been swapped starting by the one with more variance. The total proportion of altered variance is also shown in parentheses. With respect to microaggregation plus noise, *parms* refers to the number of records grouped together in the microaggregation phase. Finally, *parms* refers to the maximum relative rank difference allowed in rank swapping.

The skewness and kurtosis relative bias has been calculated as well, and, in addition, the mean average for all the variables and all the realizations is shown in Table 4.

Table 4. Simulation study showing the parameters used (parms), the skewness and kurtosis bias, Propensity Scores (PS_2 and PS_3 for second-degree and third-degree interactions, respectively), Probabilistic Information Loss (PIL), Distance-based Record Linkage (DBRL) and Interval Disclosure (ID) indexes, as well as the global scores for the PCA-based, "microaggregation plus noise" and rank swapping methods with different parameters. The smallest score for each of the methods has been boldfaced.

parms	Skewness Bias	Kurtosis Bias	PS_2	PS_3	PIL	DBRL	ID	$Score_2$	$Score_3$
1 (63.44%)	0.4512	0.5018	0.0020	0.0988	0.2306	0.3642	0.4567	0.2633 (0.2506.0.2791)	0.2876 (0.2757.0.2998)
2 (73.26%)	0.4863	0.5183	0.0060	0.1660	0.2436	0.1802	0.4326	0.2156 (0.2064.0.2308)	0.2556 (0.2450.0.2716)
3 (82.13%)	0.4869	0.5421	0.0104	0.2200	0.2487	0.0974	0.4269	0.1958 (0.1873.0.2078)	**0.2482 (0.2354.0.2592)**
4 (88.32%)	0.5129	0.5688	0.0148	0.2792	0.2478	0.0555	0.4221	0.1850 (0.1755.0.1957)	0.2511 (0.2372.0.2623)
5 (91.60%)	0.5192	0.5729	0.0172	0.3052	0.2494	0.0328	0.4174	0.1792 (0.1717.0.1884)	0.2512 (0.2360.0.2645)
6 (94.17%)	0.5222	0.5725	0.0204	0.3268	0.2501	0.0178	0.4120	0.1751 (0.1666.0.1856)	0.2517 (0.2342.0.2642)
7 (96.25%)	0.5232	0.5725	0.0228	0.3540	0.2549	0.0095	0.4095	0.1742 (0.1647.0.1883)	0.2569 (0.2363.0.2759)
8 (97.91%)	0.5247	0.5754	0.0240	0.3548	0.2517	0.0052	0.4061	0.1718 (0.1653.0.1889)	0.2545 (0.2293.0.2735)
9 (98.78%)	0.5234	0.5728	0.0268	0.3592	0.2547	0.0030	0.4006	0.1713 (0.1621.0.1851)	0.2544 (0.2290.0.2713)
10 (99.54%)	0.5259	0.5740	0.0284	0.3628	0.2549	0.0018	0.3987	0.1709 (0.1619.0.1811)	0.2545 (0.2373.0.2710)
11 (99.82%)	0.5232	0.5738	0.0280	0.3660	0.2550	0.0016	0.3971	0.1704 (0.1624.0.1817)	0.2549 (0.2358.0.2756)
12 (99.93%)	0.5253	0.5754	0.0284	0.3612	0.2512	0.0016	0.3963	0.1694 (0.1608.0.1808)	0.2525 (0.2367.0.2702)
13 (100.00%)	0.5268	0.5751	0.0280	0.3612	0.2501	0.0012	0.3973	**0.1691 (0.1626.0.1783)**	0.2525 (0.2275.0.2688)
3	0.3615	0.5154	0.0296	0.3092	0.2864	0.0461	0.3426	0.1762 (0.1664.0.1863)	**0.2461 (0.2332.0.2630)**
5	0.4429	0.6042	0.0320	0.4404	0.3058	0.0239	0.3283	**0.1725 (0.1614.0.1866)**	0.2746 (0.2563.0.2887)
7	0.5189	0.6781	0.0316	0.6128	0.3156	0.0166	0.3288	0.1732 (0.1597.0.1885)	0.3185 (0.2954.0.3367)
10	0.6124	0.7665	0.0324	0.6544	0.3302	0.0111	0.3329	0.1766 (0.1634.0.1947)	0.3321 (0.3170.0.3507)
25	0.7844	0.8885	0.0340	0.7788	0.3480	0.0052	0.3392	0.1816 (0.1668.0.1977)	0.3678 (0.3509.0.3928)
50	0.8864	0.9328	0.0348	0.8432	0.3625	0.0028	0.3517	0.1880 (0.1712.0.2132)	0.3901 (0.3693.0.4133)
75	0.9300	0.9449	0.0340	0.8644	0.3670	0.0023	0.3611	0.1911 (0.1734.0.2103)	0.3987 (0.3794.0.4204)
0.05	0	0	0.1676	0.1784	0.3259	0.7645	0.6972	0.4888 (0.4785.0.4992)	0.4914 (0.4817.0.5022)
0.1	0	0	0.2864	0.2984	0.3725	0.4972	0.4441	0.4000 (0.3891.0.4112)	0.4030 (0.3913.0.4140)
0.15	0	0	0.3784	0.3856	0.3829	0.2599	0.3460	0.3418 (0.3309.0.3537)	0.3436 (0.3318.0.3549)
0.2	0	0	0.4588	0.4580	0.3872	0.1210	0.3169	**0.3209 (0.3084.0.3323)**	**0.3207 (0.3037.0.3370)**
0.5	0	0	0.7396	0.6788	0.3925	0.0094	0.2645	0.3515 (0.3404.0.3606)	0.3363 (0.2972.0.3566)
1	0	0	0.8572	0.7316	0.3930	0.0015	0.2362	0.3720 (0.3647.0.3782)	0.3405 (0.2615.0.3706)

The smallest score for each of the methods has been boldfaced. It can be seen that, independently of the chosen propensity score, the smallest scores are taken by the "microaggregation plus noise" and the proposed methods. Rank swapping, which has been recognized as a very well-performing technique by Domingo-Ferrer and Torra (2004) among others, leads to worse results, as its scores are higher and its confidence intervals do not overlap with those of the other methods (this is due to the fact that the variance-covariance matrix is not very well preserved). However, rank swapping is the only method that leads to zero bias on the skewness and kurtosis of the variables.

Regarding second-degree propensity scores, we see that the best results are around 0.17 for both the "microaggregation plus noise" and the proposed methods. For the PCA-based method, the best balance between data utility and disclosure risk is taken for a high number of altered components, in which case the DBRL gets very low values because of the "synthetic" aspect of the resulting masked data set. Furthermore, the data utility from the PS perspective does not decrease dramatically with increasing numbers of swapped components. However, the CIs of Score$_2$ almost coincide for more than 94% of the variability perturbed. This is due to the fact that the utility remains almost constant and the risk, although it decreases with the number of components, is already very small.

As for microaggregation plus noise, which has been recognized as a very well-performing method in Woo et al. (2009), the lowest score takes place for $parms = 5$, although the results are not statistically different up to $parms = 10$, as can be deduced from the CIs. It can be seen that this method provides better attribute disclosure protection, but worse record linkage compared with the PCA-based method, and that it has a slightly bigger score.

As regards third-degree propensity scores, we can see that the utility is worse, as both "microaggregation plus noise" and the PCA-based methods fail at preserving third moments. Note that the relative bias is similar for both methods for the best parameters and increases with the number of perturbed components and the size of the groups. For this reason, the proposed method now reaches its best value for $parms = 3$, as the bias is smaller. Nevertheless, it can be seen that very similar results (in fact, they are not statistically different) are obtained for any number of components larger than three.

Again, the best results are reached for small parameters for the "microaggregation plus noise" method. Its CI overlaps with those of the proposed method, meaning that the balance between data utility and risk disclosure achieved is similar for both methods.

All in all, the proposed method has led to statistically similar results to "microaggregation plus noise". In spite of that, microaggregation techniques are usually slower than the proposed method and, thus, the proposed method is preferable as it can provide data sets with similar quality, but faster. For the computer specified above, the "microaggregation plus noise" method (using the "mdav" algorithm for the microaggregation phase), took between 5 and 30 times longer than the proposed method, depending on the data set.

4.2. Hybrid Data for Partial Protection

As stated in Section 3, the proposed method also permits perturbing some variables more than others by selecting the components more related to the target variables and just modifying them. This is an interesting feature, as it also allows preserving the whole original correlation structure.

We now show how to do it in the *Census* data set in Brand et al. (2002), which contains 1,080 records and 13 variables and has been previously used in Domingo-Ferrer and Torra (2004), Domingo-Ferrer and González-Nicolás (2010) or Jiménez et al. (2014). It is important to highlight that one of the 13 variables is a linear combination of other variables and, therefore, it is omitted from the analysis. However, after the masking is performed on the remaining variables, one can directly obtain its protected version by simply adding the corresponding protected variables.

In this example, we assume that only the first variable needs special protection. In order to proceed, we first need to obtain the weights of the first variable on the twelve components. The left plot in Figure 1 shows these weights. It can be seen that this variable has very little influence on many components, as their weights are close to zero, but has a great influence on the third component (represented by a gray circle). In fact, the weight of the third component on the first variable represents 59.91% of the total components' weights. Therefore, to protect the first variable, perturbing the third component is enough.

In order to check how the remaining variables can be affected with the perturbation of the third component, we can analyze the weights of the original variables on this component (shown in the middle plot of Figure 1). Note that the only significant weights are those associated with variables 1 and 8 (represented by a gray circle) and, therefore, only those variables will be significatively affected. Finally, looking at the weights of the eighth variable on all the components (see the right plot in Figure 1), we can observe that the third component is not the one with the highest weight and, thus, the effect of perturbing it will not be significant, as most of the information (around 87%) of this variable will be left unchanged.

As already stated, with this example we aim to show how to obtain hybrid data sets. For this reason, instead of using swapping to alter the components, we substitute them with a random draw from its smoothed empirical cumulative distribution function (see Fienberg 1994). Furthermore, we compare the results with the ones obtained using the methods proposed in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010) (as previously noted, we will refer to them as MS and MH, respectively). Both

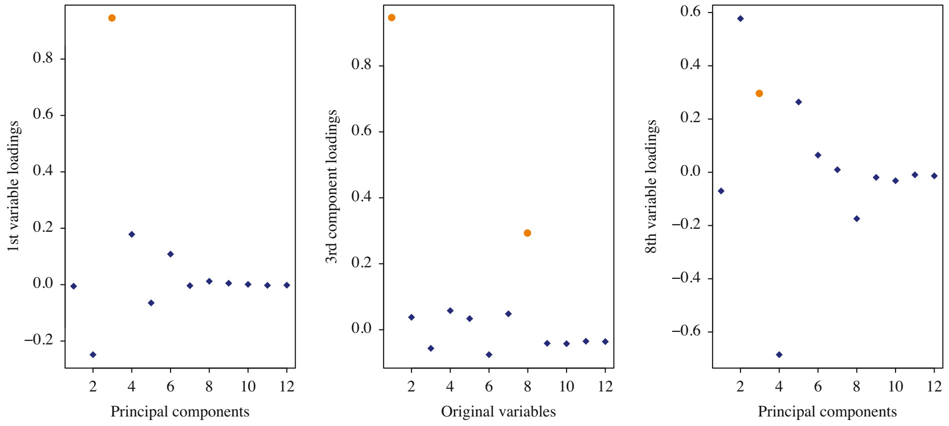


Fig. 1. Illustrative example: representation of the weights of the first and eighth variables on all the components (left and right plots, respectively) and of the third component on all the variables (middle plot).

methods provide very good results and require some sensitive and some nonsensitive variables. In our example, we assume that the first variable in the census data set is the sensitive one. Moreover, for the MS method we show the results associated with the following selected parameters:

- α equal to 0: In this case, the MS method equals the IPSO method in Burridge (2003).
- α equal to 0.4: In order to perform a fair comparison between the methods, we have selected this parameter as it leads to a result that takes 40% of the information from the original record, as it is the case with the PCA method, where we are only perturbing the third component and it represents almost 40% of the total variability of the first variable.
- α equal to 0.9: a case where the result is highly dependent on the original record.

Regarding MH, we show the results for group sizes of 54, 180, and 540, which correspond to 20, 6, and 2 microaggregated groups, respectively. Note that for the ease of brevity we only show a subset of the results obtained when comparing the methods.

After applying the method, we can analyze the characteristics of the resulting data set. In particular, Table 5 shows Pearson's correlation coefficients of the original and resulting variables after applying the PCA-based method. It can be seen that the coefficients are larger than 0.99 for all the variables except for the first one, which is low as a result of the masking process, and the eighth one, as we predicted. However, it is still higher than 0.9, so it has only suffered a mild perturbation. As for the MS and MH methods, the nonsensitive variables remain unchanged and Pearson's correlation coefficients of the original and resulting first variable are 0.0331, 0.4199, and 0.9033 for α equal to 0, 0.4, and 0.9 after applying the MS method, respectively, and 0.6528, 0.3877, and 0.0512 for k equal to 54, 180, and 540, respectively.

With regard to the preservation of the original Pearson's correlation coefficients, Table 6 contains the original and resulting Pearson's correlation coefficients of the first and third variables for the proposed method (for the ease of brevity we do not show all variables, but the results are similar to those of the third variable). It can be seen that the coefficients are similar for the masked variable (the absolute difference is around four percent) and almost coincide for the remaining ones. We did not show these coefficients for the MS and MH methods, as both exactly preserve means, variances and covariances.

Furthermore, we have analyzed how close the original records are to the masked ones by means of the classic rank interval disclosure, defined in the previous subsection. We remind that it computes the proportion of records that lie in a narrow interval around its masked value. Again, we have obtained 1,000 hybrid data sets and we have computed this proportion. Table 7 shows the mean value of this index for the sensitive variable. Moreover, the centered 95% percentile bootstrap confidence interval of the proportion has been obtained.

Table 5. Pearson's correlation coefficient between the original and the protected variables applying the PCA-based method.

Variable											
1	2	3	4	5	6	7	8	9	10	11	12
0.1736	0.9986	0.9970	0.9968	0.9989	0.9945	0.9978	0.9181	0.9984	0.9983	0.9988	0.9988

Table 6. Original and resulting Pearson's correlation coefficients of the first and third variables applying the PCA-based method.

Variable	1	2	3	4	5	6	7	8	9	10	11	12
original V_1	1.0000	0.0102	0.0490	−0.0036	0.0182	−0.0996	0.0038	−0.0236	0.0311	0.0324	0.0386	0.0362
resulting V_1	1.0000	0.0463	0.0869	0.0325	0.0572	−0.0701	0.0364	0.0040	0.0684	0.0544	0.0713	0.0810
original V_3	0.0490	0.4910	1.0000	0.3824	0.4952	0.2894	0.4359	−0.0576	0.5597	0.5469	0.5594	0.5498
resulting V_3	0.0869	0.4907	1.0000	0.3825	0.4946	0.2860	0.4361	−0.0475	0.5580	0.5457	0.5580	0.5479

Table 7. Disclosure risk DR (rank interval disclosure) and data utility DU (propensity score) of the output data sets.

		PCA-based	IPSO	MS ($\alpha = 0.4$)	MS ($\alpha = 0.9$)	MH ($k = 54$)	MH ($k = 180$)	MH ($k = 540$)
DR	Mean	0.1151	0.1088	0.1540	0.3597	0.2187	0.1592	0.1105
	CI	(0.0998, 0.1304)	(0.0946, 0.1223)	(0.1380, 0.1704)	(0.3442, 0.3771)	(0.2026, 0.2349)	(0.1441, 0.1760)	(0.0976, 0.1245)
DU	Mean	0.0035	0.0100	0.0089	0.0008	0.0004	0.0049	0.0092
	CI	(0.0019, 0.0058)	(0.0068, 0.0137)	(0.0059, 0.0125)	(0.0003, 0.0015)	(0.0002, 0.0007)	(0.0032, 0.0066)	(0.0059, 0.0127)

Based on the mean and CI disclosure risk values, it can be seen that the proposed method, the IPSO method (MS $\alpha = 0$ and MH $k = 1{,}080$) and MH ($k = 540$) lead to very similar levels of protection. However, this is not the case for large and small values of α and k, respectively, as their mean values are larger and their CIs do not overlap with the one associated with the PCA-based method.

Finally, although it is clear that the MS method preserves better the sufficient statistics (mean and variance-covariance matrix), the PCA-based method also provides good approximations. For this reason, it would be interesting to compare the results based on other indexes that take into account different features of the data sets, such as the propensity score. Table 7 shows the mean propensity score of 1,000 hybrid data sets, as well as its centered 95% bootstrap confidence interval.

As expected, for the MS method, the larger α is, the larger data utility is obtained with the MS method. It can be seen that the PCA-based method leads to more useful data sets if the parameter α is smaller than 0.4, as its mean propensity score is smaller than the others, and the CIs for α equal to 0 and 0.4 do not overlap with the PCA-based one. Similar conclusions can be drawn for the MH method: large values of k lead to data sets with lower utility. Moreover, the MH method leads to worse, similar and better utility levels than the PCA-based method for k equal to 540, 180, and 54, respectively.

However, in this case, we can conclude that the PCA-based method outperforms the MS and the MH methods in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), respectively, as it provides data sets with a better balance between data utility and disclosure risk (as neither MS nor MH provide at the same time better disclosure risk and data utility than the proposed method).

It is important to highlight that the good or bad performance of the proposed method in the case of partial protection is data-dependent in the sense that depending on the correlation structure, more or fewer variables will be affected by the protection process. For example, if the confidential variables to be protected show low-medium correlation with the non confidential ones, the principal components with high weights on the confidential ones will then tend to show low weights on the nonconfidential ones and, thus, the perturbation process will affect them mildly. On the other hand, if the confidential variables are highly correlated with the nonconfidential ones, the perturbed principal components with high weights on the confidential variables will also have high weights on the nonconfidential one, and, therefore the nonconfidential variables will be highly affected. The example showed in this subsection is of the first type. Better results than in the proposed method are expected for the MS and MH methods for the second type of situation, as the nonconfidential variables are not perturbed and that leads to better overall utility (although similar levels of protection).

5. Final Considerations

In light of the previous examples, we can give some guidelines on how to apply the proposed method:

- The proposed method is expected to lead to better results than other proposed methods where all variables need protection. If there is only a subset of confidential variables, the performance of the method depends on the correlation structure of the confidential and nonconfidential variables. If this correlation is low, then the proposed method outperforms other methods previously proposed in the literature. Otherwise, it is preferable to resort to other methods.

- If all variables need protection, selection of the components to be perturbed depends on the desired results. If third or higher moments are not critical, then it is recommended to alter all of them, as the utility achieved is similar to that obtained with fewer components, but the record linkage gets reduced.
- On the other hand, if third or higher moments are critical, we recommend starting perturbing the first components (those with higher variance) until the desired protection level has been reached.
- If all variables need some kind of protection but we wish to perturb some more than the others, then the matrix of weights needs to be analyzed for a careful selection of the components.
- Perturbing only a few of the "last" components leads to very little protection, as its corresponding variance is very small.

The proposed method cannot guarantee that the protected variables lie in a predefined interval. For those cases, we suggest applying one of the methods proposed in Kim et al. (2015).

6. Conclusions and Future Work

In this article we have presented a simple and versatile method for limiting disclosure in continuous microdata based on PCA that preserves the mean vector and the variance-covariance matrix. The versatility of the method comes from the fact that it can provide masked, hybrid or fully synthetic protected data sets and it can be used to protect all or only some of the variables in a data set.

The method is very simple and, thus, does not require complex or very powerful software. We have not compared the method with more sophisticated techniques, such as multiple imputation, as we aim to provide an easy and efficient tool, in terms of good and fast results that can be widely applicable.

Some simulation studies have been performed to compare the proposed method with other well-performing techniques in terms of data utility and disclosure risk. Regarding the application of the proposed method to protect all variables at the same time, it has been shown that the PCA-based method offers a very good balance between data utility and disclosure risk and provides much better results than rank swapping and similar ones compared to "microaggregation plus noise".

As for what we call *hybrid data for partial protection*, the PCA-based method has provided better data sets than the methods proposed in Muralidhar and Sarathy (2008) and Domingo-Ferrer and González-Nicolás (2010), in the sense that, when comparing protected data sets with these three methods with similar data utility in terms of disclosure risk, the proposed method leads to safer data sets and the same happens with data sets with similar disclosure risk. As it has been already highlighted, the PCA-based method outperforms the MS and the MH method only if the confidential variables are not highly correlated with with the nonconfidential ones.

Regarding future work, the method could be extended to categorical variables by means of Categorical Principal Components Analysis. Moreover, other types of orthogonal transformations, such as the Independent Component Analysis, are to be explored in the future to check if they can lead to better results. Finally, in order to improve the usage of the method in the case of partial protection, we plan to use Factor Analysis instead of PCA, as it is possible to rotate the factors obtained, isolating the effect of the variables on the factors.

7. References

Banu, R. and N. Nagaveni. 2009. "Preservation of Data Privacy Using PCA Based Transformation." In *International Conference on Advances in Recent Technologies in Communication and Computing*, 439–443. Doi: http://dx.doi.org/10.1109/ARTCom.2009.159.

Brand, R. 2002. "Microdata Protection through Noise Addition." In *Inference Control in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 2316: 97–116. Berlin Heidelberg: Springer. Doi: http://dx.doi.org/10.1007/3-540-47804-38.

Brand, R., J. Domingo-Ferrer, and J. Mateo-Sanz. 2002. *Reference Data Sets to Test and Compare SDC Methods for Protection of Numerical Microdata*. Deliverable of European Project IST-2000-25069 CASC. Available at: http://neon.vb.cbs.nl/casc (accessed August 2016).

Burridge, J. 2003. "Information Preserving Statistical Obfuscation." *Statistics and Computing* 13: 321–327. Doi: http://dx.doi.org/10.1023/A:1025658621216.

Domingo-Ferrer, J. and U. González-Nicolás. 2010. "Hybrid Microdata Using Microaggregation." *Information Sciences* 180: 2834–2844. Doi: http://dx.doi.org/10.1016/j.ins.2010.04.005.

Domingo-Ferrer, J. and V. Torra. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata." In *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*, edited by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz. 111–133. Elsevier. Available at: https://www.iiia.csic.es/es/publications/quantitativecomparison-disclosure-control-methods-microdata (accessed August 2016).

Domingo-Ferrer, J. and V. Torra. 2004. "Disclosure Risk Assessment in Statistical Data Protection." *Journal of Computational and Applied Mathematics* 164: 285–293. Doi: http://dx.doi.org/10.1016/S0377-0427(03)00643-5.

Drechsler, J. 2011. *Synthetic datasets for statistical disclosure control: theory and implementation*, volume 201. Springer Science & Business Media.

Duncan, G. and R. Pearson. 1991. "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future." *Statistical Science* 6: 219–239.

Efron, B. and R. Tibshirani. 1993. *An introduction to the Bootstrap*. New York: Chapman and Hall.

Fienberg, S. 1994. *A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality*. Technical Report 611, Department of Statistics, Carnegie Mellon University.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Nordholt, K. Spicer, and P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons.

Jiménez, J., G. Navarro-Arribas, and V. Torra. 2014. "JPEG-Based Microdata Protection." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 8744: 117–129. Springer International Publishing. Doi: http://dx.doi.org/10.1007/978-3-319-11257-210.

Jolliffe, I. 2002. *Principal Component Analysis*. New York, USA: Springer.

Kim, H., A. Karr, and J. Reiter. 2015. "Statistical Disclosure Limitation in the Presence of Edit Rules." *Journal of Official Statistics* 31: 121–138. Doi: http://dx.doi.org/10.1515/jos-2015-0006.

Liew, C., U. Choi, and C. Liew. 1985. "A Data Distortion by Probability Distribution." *ACM Transactions Database Systems* 10: 395–411.

Mateo-Sanz, J., J. Domingo-Ferrer, and F. Sebé. 2005. "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata." *Data Mining and Knowledge Discovery* 11: 181–193. Doi: http://dx.doi.org/10.1007/s10618-005-0011-9.

Moore, R. 1996. *Controlled Data Swapping Techniques for Masking Public use Microdata Sets*. Technical report, U.S. Bureau of the Census, Washington, D.C. Available at: https://www.census.gov/srd/papers/pdf/rr96-4.pdf (accessed August 2016).

Muralidhar, K. and R. Sarathy. 2008. "Generating Sufficiency-Based Non-Synthetic Perturbed Data." *Transactions on Data Privacy* 1: 17–33. Available: at http://www.tdp.cat/issues/tdp.a005a08.pdf (accessed August 2016).

Muralidhar, K., R. Sarathy, and J. Domingo-Ferrer. 2014. "Reverse Mapping to Preserve the Marginal Distributions of Attributes in Masked Microdata." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer. Lecture Notes in Computer Science, 8744: 105–116. Springer International Publishing. Doi: http://dx.doi.org/10.1007/978-3-319-11257-29.

Oganian, A. and A. Karr. 2006. "Combinations of SDC Methods for Microdata Protection." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and L. Franconi. Lecture Notes in Computer Science, 4302: 102–113. Berlin Heidelberg: Springer. Doi: http://dx.doi.org/10.1007/1193024210.

Pagliuca, D. and G. Seri. 1999. *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*. Esprit SDC Project, Deliverable MI-3/D2.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Website: http://www.R-project.org/.

Raghunathan, T.E., J. Reiter, and D. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16.

Rubin, D. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9: 461–468.

Sarathy, R. and M. Krishnamurty. 2002. "The Security of Confidential Numerical Data in Databases." *Information Systems Research* 13: 389–403. Doi: http://dx.doi.org/10.1287/isre.13.4.389.74.

Templ, M. 2008. "Statistical Disclosure Control for Microdata Using the Rpackage sdcMicro." *Transactions on Data Privacy* 1: 67–85. Doi: http://dx.doi.org/10.18637/jss.v067.i04.

Woo, M., J. Reiter, A. Oganian, and A. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1: 111–124.

Received September 2015
Revised August 2016
Accepted August 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 43–59, http://dx.doi.org/10.1515/JOS-2017-0003

Estimating the Count Error in the Australian Census

James Chipperfield[1], *James Brown*[2], *and Philip Bell*[3]

In many countries, counts of people are a key factor in the allocation of government resources. However, it is well known that errors arise in Census counting of people (e.g., undercoverage due to missing people). Therefore, it is common for national statistical agencies to conduct one or more "audit" surveys that are designed to estimate and remove systematic errors in Census counting. For example, the Australian Bureau of Statistics (ABS) conducts a single audit sample, called the Post Enumeration Survey (PES), shortly after each Australian Population Census. This article describes the estimator used by the ABS to estimate the count of people in Australia. Key features of this estimator are that it is unbiased when there is systematic measurement error in Census counting and when nonresponse to the PES is nonignorable.

Key words: Undercount; capture-recapture; Dual System Estimator.

1. Introduction

In many countries, counts of people are a key factor in the distribution of government resources. However, the (observed) Census counts differ from the (unobserved) true counts for an area because of overcoverage (e.g., a person is counted multiple times or counted once in the wrong location) and undercoverage (e.g., people are missed). In addition, the Census may count individuals on a 'person present' basis (de facto) while counting on a 'usual residents' basis (de jure) is typically more relevant for government policy. 'Usual residents' counts may be significantly smaller than 'persons present' counts in certain locations (e.g., tourist areas and city centres) and this will have implications for the type and size of government infrastructure projects.

Methods have been developed to correct for systematic errors in the observed Census counts (for a review see Belin and Rolph 1994). A classical approach to estimating person counts is the Dual System Estimator (DSE) as developed in Sekar and Deming (1949). In this traditional form, the DSE has been widely used by national statistical agencies, including the US Census Bureau Bureau (see Xi Chien and Tang 2011; Mule 2008; Griffin and Mule 2008; Alho et al. 1993; and Hogan 1993) and the Office for National Statistics

[1] Australian Bureau of Statistics, Methodology Division, P O Box 10 Belconnen Australian Capital Territory 2616, Australia. Email: james.chipperfield@abs.gov.au
[2] University of Technology, Sydney, School of Mathematical Sciences, Sydney, New South Wales, Australia. Email: james.brown@uts.edu.au
[3] Australian Bureau of Statistics, Methodology Division, Adelaide, South Australia, Australia. Email: philip.bell@abs.gov.au

Acknowledgments: We would like to thank Bill Gross, David Steel, Stephen Carlton and Frank Yu for pointing us in appropriate directions and linking us to relevant literature. Prof James Brown at UTS is part supported by a research grant from the ABS.

(ONS) (see Brown et al. 1999; Brown et al. 2006). However, in the Australian context, the Australian Bureau of Statistics (ABS) has developed an approach that differs fundamentally from the US and UK approach on perhaps three points.

First, in the US, the approach has developed with two independent 'audit' sample surveys of the Census (see Hogan 1993). In the UK (see Large et al. 2011), the approach has developed from a single independent coverage survey (audit sample) but with separate adjustments for erroneous enumerations (overcoverage and undercoverage). In contrast, the ABS approach integrates measurement of undercoverage and overcoverage of its Census from a single survey, called the Post Enumeration Survey (PES).

Second, while the Australian Census counts on a 'persons present' basis, the PES counts on a 'usual residents' basis, so the second count (PES) is not just a repetition of the first count (Census).

Third, as we will see, the ABS approach has also been developed to handle people who are classified differently by the PES and the Census (e.g., a person may be classified as an Aboriginal and Torres Strait Islander by the PES, but not by the Census). The assumption here is that the PES classification is correct. This is perhaps reasonable, since the PES consists of a face-to-face interview conducted by ABS's professional interviewers, while the Census typically uses self-completion, supported by a large temporary field-force.

In Section 2 we review the general DSE framework, as applied by the US Census Bureau. Section 3 introduces the standard regression estimator of person counts and motivates the more robust estimator that was used in 2011. Section 4 describes a small simulation study of the two estimators in Section 3. Section 5 describes a more realistic and involved simulation study of the two estimators in Section 3. Section 6 contains some concluding remarks.

2. Traditional Approach

In the traditional approach, a population set, U, is defined of people indexed by j. The population can be categorised into H subgroups, defined in terms of characteristics such as age, sex, Aboriginal and Torres Strait Islanders status, country of birth, and geography. Subgroups need not be mutually exclusive. The aim is to estimate the number of people in the hth subgroup $T_h = \sum_{j \in U} t_{jh}$ where $t_{jh} = 1$ if the jth person in the population belongs to subgroup h and $t_{jh} = 0$ otherwise and $h = 1, . . . H$. The population is counted by the Census. The Census count is denoted for subgroup h by X_h for $h = 1, . . . , H$. While in the traditional DSE approach T_h and X_h are conceptually the same, in the Australian situation they are not (see Section 3).

Consider the situation where following the Census, we conduct an independent PES of the population, typically by sampling dwellings. In the context of the US Census Bureau, the reference date for PES counting would be Census Night and the PES is referred to as the P-sample, as it is a sample of the population (Hogan 1993, 2003).

The PES will, of course, have nonresponse (undercoverage), but overcoverage can be assumed to be zero because it uses ABS professional interviewers who are familiar with applying rigorous procedures to correctly identify and avoid double-counting usual residents within selected dwellings. For example, these procedures ensure that overseas visitors are identified and discarded from estimation and that each person can only be selected via a single dwelling. Let the PES responding sample size be n people, and denote the sample set by s. The PES collects t_{ih} where $t_{ih} = 1$ if the ith sample person belongs to subgroup h and $t_{ih} = 0$, $h = 1, \ldots H,$ and $i = 1, \ldots n.$ After matching the PES to the Census, we can then derive an indicator, m_{ih}, for 'correct Census enumeration', where $m_{ih} = 1$ if person i is counted in subgroup h by both the PES and Census, and otherwise $m_{ih} = 0$. Let $w_i = \pi_i^{-1}$, where π_i is the probability, with respect to the PES sample design, that person i was selected.

2.1. *Estimating Undercoverage Using a P-Sample*

Now, if we impose the assumption that the PES and Census enumerate the populations independently, the classic DSE of T_h without an adjustment for overcoverage (previously used by the US Census Bureau, see Hogan 1993) is

$$\hat{T}_h = \hat{R}_h \times X_h. \qquad (1)$$

where

$$\hat{R}_h = \frac{\sum_{i \in s} w_i t_{ih}}{\sum_{i \in s} w_i m_{ih}}, \qquad (2)$$

is an estimate of the probability that an individual is missed by the Census, for all $h. = 1, \ldots H.$

This probability, $\hat{R}_h$, adjusts the total Census count, X_h, for Census undercoverage.

2.2. *Estimating Over- and Undercoverage Using E- and P-Samples*

For several practical reasons, such as enumerators following up the wrong households when the forms are posted-out, the Census count X_h will have a typically low level of overcoverage. Table 1 in Large et al. (2011) shows that historical levels of overcount have been less than one per cent in Switzerland, Canada, Australia, and the UK; the US being an

Table 1. Description of the simulation scenarios.

	Error rates*			
Scenario	Census misclassification	Census duplication	Census missing	PES missing
1	0.1	LOW	LOW	LOW
2	0.1	MEDIUM	MEDIUM	MEDIUM
3	0.2	LOW	LOW	LOW
4	0.2	MEDIUM	MEDIUM	MEDIUM
5	0.3	MEDIUM	MEDIUM	MEDIUM
6	0.3	HIGH	HIGH	HIGH

*Low scenario = 0.1 in communities and 0.05 outside communities
*Medium scenario = 0.2 in communities and 0.1 outside communities
*High scenario = 0.3 in communities and 0.15 outside communities

exception, due to a post-out Census in most areas for several decades. Recently, Census overcount has been increasing; in the 2011 Canadian Census it was 1.85%.

If there is overcoverage in the Census count, X_h, using (1) will be positively biased for T_h. Thus, an additional adjustment is required. In the US context, this involves selecting a sample of Census records and confirming whether the enumerations were correct. This sample is referred to as the E-sample (Hogan 1993). For the kth Census record sampled in the E-sample, we resolve whether the record should have been enumerated ($e_k = 1$) or should not have been enumerated ($e_k = 0$). The DSE of T_h is now

$$\hat{T}_h^{(DSE)} = \hat{R}_h \times X_h \times \hat{E}_h \tag{3}$$

where $\hat{R}_h$ still adjusts for undercoverage, while

$$\hat{E}_h = \frac{\sum_{kh} v_k e_k}{\sum_{kh} v_k} \tag{4}$$

is an estimate of the probability that an enumerated Census record is actually a correct enumeration in subgroup h, $\sum_{kh}$ is the summation of records in the E-sample classified to subgroup h, and v_k is the appropriate sampling weight for the kth record in the E-sample.

As already stated, both (1) and (3) assume that enumerations in Census and PES occur independently, conditional on subgroup. When this independence assumption is not true, the DSE will be biased for T_h (see, for example Wachter and Freedman 2000). Reducing this 'correlation bias' is possible by bringing in external information, such as a known sex ratio, as developed in Wolter (1990) and implemented in Bell (2001), or some other additional information (Brown et al. 2006).

In the 2010 Census, the US Census Bureau (see Mule 2008) extended (3) by modelling the probability of correct enumeration and the probability of incorrect enumeration at an individual level (i.e., by fitting a logistic regression to m_i in the P-sample and e_k in the E-sample).

3. The ABS Approach

3.1. Differences Between the Australian and Traditional Approaches

Now we bring in the Australian context, which has additional complications. First, there is only one additional sample - a P-sample (i.e., there is no E-sample). This must be taken into account when estimating overcoverage in the Census.

Second, there are systematic differences between Census counts and PES counts where, for the reasons outlined below, we are interested in the latter. The Census counts people on a 'person present' basis, while the PES counts people on a 'usual residents' basis. The latter is typically more useful to the government when allocating resources. This means it is quite legitimate for a person to be in one geographic area in the Census and in a different area in the PES. (To facilitate matching a person's Census and PES records, the PES asks respondents about possible locations for their Census enumeration.) The number of 'movers' is expected to be small, given that the Census and the PES are carried out only a couple of weeks apart. In the US approach, in which the time between Census and survey is longer, an adjustment is made to account for 'movers' (Griffin 2000). It is also possible for the Census and the PES to classify a person in different subgroups, even if they are enumerated at the same geographic location. This discrepancy in classification is more noticeable in some subgroups, in particular to a more significant extent in Aboriginal and Torres Strait Islanders status than in others. Since the Census uses self-enumeration, usually with one individual responding for all household members; the PES is more likely to be 'correct' in the Australian context, where the ABS utilises its professional field-force for the PES interviews. For the reasons mentioned, it is assumed here that the PES always correctly classifies a person to subgroup. No such assumption is made with regard to census classification. (It is worthwhile to note here that the traditional DSE in Section 2 does not correct for systematic differences in PES and Census classification of people to subgroup).

The observed Census counts $\boldsymbol{X^T} = (X_1, \ldots X_h, \ldots, X_H)$ are calculated by summing over all Census records in each of the H subgroups. We may consider expressing the Census counts by $\boldsymbol{X} = \sum_{j \in U} \boldsymbol{x}_j$, where $\boldsymbol{x}_j^T = (x_{j1}, \ldots x_{jh}, \ldots, x_{jH})$, $X_h = \sum_{j \in U} x_{jh}$, and x_{jh} is the number of times person j was counted by the Census in subgroup h for $j \in U$. If person j in the population is missed by the Census (i.e., not counted in any subgroup), then we can notionally set $\boldsymbol{x}_j = \boldsymbol{0}$, where $\boldsymbol{0}$ is an H column vector of zeros. While we can calculate $\boldsymbol{X}$, we do not observe $\boldsymbol{x}_j$ for all j, as this would require that each person in the population is identified and explicitly assigned a value for $\boldsymbol{x}$.

In other contexts, $\boldsymbol{x}$ can define characteristics of people in administrative data. Many countries now use administrative data as either the entire basis for their census or as a major component of their census. Valente (2010) provides a review of different approaches taken by different European countries. In the case of the Netherlands (see Nordholt 2005), $\boldsymbol{x}$ contains classification errors, and the PES functions somewhat as a quality correction for the administrative data, rather than as a coverage check (Brown and Honchar 2012).

After matching PES and Census records, we can establish x_{ih}, the number of times the ith PES respondent was counted by the Census in subgroup h. Again, $x_{ih} = 0$ if the Census did not count the ith PES respondent in subgroup h (i.e., if the ith PES record could not be matched to any Census record in that subgroup). The variable x_{ih} captures information about overcoverage (if greater than 1) *and* undercoverage (if zero). This is important since here, unlike in the DSE approach, we only have a single audit sample (i.e., there is a P-sample but no E-sample) to capture information about over- and undercoverage. Integrating the two also recognises that both errors are inherent in the Census, and our target is to recognise this in our estimation. We now have several possible types of coverage outcomes; the main ones being:

- $t_{ih} = 1$, $x_{ih} = 1$, and $x_{ig} = 0$ and $t_{ig} = 0$ for all $g \neq h$
  - the PES and Census counts a person once in the same subgroup,
- $t_{ih} = 1$ and $x_{ih} = 2$
  - the PES and Census counts a person once and twice in the same subgroup, respectively (e.g., duplication),

- $t_{ih} = 1$ and $x_{ig} = 0$ for all g
  - ○ the PES counts a person but the Census did not count them at all,
- $t_{ih} = 1$, $x_{ih} = 0$, $x_{if} = 1$, $x_{ig} = 0$ for all $g \neq h$ or $g \neq f$
  - ○ the PES and Census counts a person in different subgroups. This could be due to Census misclassification or because a person was enumerated by the PES and Census in different geographic locations, and
- $t_{ih} = 1$, $x_{ih} = 1$, $x_{if} = 1$, $x_{ig} = 0$ for all $g \neq h$ or $g \neq f$
  - ○ the PES and Census counts a person once in the same subgroup but the Census also counts the person in a different subgroup.

Third, we know that certain subgroups of the population are more likely to be missed by the PES, even after conditioning on auxiliary information available from the Census. As we see in the next section, this would mean that the standard generalised regression estimator (Subsection 3.2) for the PES in the Australian context would be biased, and so we consider an alternative (Subsection 3.3).

3.2. Generalised Regression Estimator Using a Prediction Model

For simplicity, in the rest of this article we replace t_{ih} with t_i, where $t_i = 1$ if person i is in an arbitrary subgroup of interest (i.e., we drop the h subscript). Similarly, we replace t_{jh} with t_j. Now we are interested in estimating $T = \sum_j t_j$ the usual resident population in an arbitrary subgroup.

Consider the 'working' linear prediction model

$$t_j = \boldsymbol{x}_j^T \boldsymbol{\alpha} + e_j \tag{5}$$

$$\mathrm{E}(e_j | \boldsymbol{x}_j) = 0$$

where the e_j s are independent and identically distributed and $\boldsymbol{\alpha}$ is an H column vector of coefficients that relate to membership of the H subgroups. We call (5) a 'working' model because a linear model is not ideal for a binary variable such as t. Nevertheless, we may use this model to motivate the classic generalised regression estimator,

$$\hat{T}^{(GREG)} = \sum_{i \in s} w_i \left(t_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\alpha}} \right) + \boldsymbol{X}^T \hat{\boldsymbol{\alpha}}$$

where $\hat{\boldsymbol{\alpha}} = \left(\sum_{i \in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\sum_{i \in s} w_i \boldsymbol{x}_i^T t_i \right)$ is the standard 'survey weighted' least squares estimator of $\boldsymbol{\alpha}$ (see Särndal et al. 1992).

If we now allow for nonresponse in the PES, we need to consider whether or not the condition under which $\hat{T}^{(GREG)}$ is asymptotically unbiased is reasonable. Denote the response indicator by I_j, where $I_j = 1$ if person j in the population would respond if selected in the PES and otherwise $I_j = 0$. Now consider the distribution of t given $\boldsymbol{x}$ in the population,

$$[t_j | \boldsymbol{x}_j ; j = 1, \ldots N] = [t | \boldsymbol{x}]$$

and the distribution of t given $\boldsymbol{x}$ in the population of PES respondents,

$$[t_j | \boldsymbol{x}_j, I_j = 1; j = 1, \ldots N] = [t | \boldsymbol{x}, I = 1].$$

If these two distributions are equal, we may write

$$[t|\boldsymbol{x}] = [t|\boldsymbol{x}, I = 1]. \tag{6}$$

From (5) and (6) it follows that $\hat{T}^{(GREG)}$ is asymptotically unbiased in the presence of nonresponse (see also Kott and Chang 2010). The condition in (6) means that the distribution of t given $\boldsymbol{x}$ is the same for PES respondents and PES nonrespondents and so we may say that nonresponse is ignorable given $\boldsymbol{x}$ (see Rubin and Little 2002); we may also say that the response, t, and the indictor for response, I, are independent conditional on $\boldsymbol{x}$ and so we may write $[t, I|\boldsymbol{x}\,] = [I|\boldsymbol{x}\,][t|\boldsymbol{x}\,]$.

However, there is strong evidence against (5) or (6) holding in the case of the PES. To illustrate this, consider breaking up the population U into Census respondents ($\boldsymbol{x}_j \neq \mathbf{0}$) and Census nonrespondents ($\boldsymbol{x}_j = \mathbf{0}$). For Census nonrespondents, (5) and (6) become

$$t_{\mathrm{j}} = \mathrm{e}_{\mathrm{j}} \tag{7}$$

$$\mathrm{E}(e_j|\boldsymbol{x}_j = \mathbf{0}) = 0$$

$$[t|\boldsymbol{x} = \mathbf{0}] = [t|\boldsymbol{x} = \mathbf{0}, I = 1] \tag{8}$$

for all $j \in U$. Equation (7) implies that, for Census nonrespondents, the *unconditional* mean of t in the population and in the population of PES respondents is zero. That is, if a person is missed by the Census, the model expects them to be missed by the PES. This is clearly not an appropriate assumption, since the PES is designed to capture information about Census undercoverage.

Equation (8) is equivalent to the assumption that, within the population of Census nonrespondents, *nonresponse occurs completely at random;* that is, within the population of PES nonrespondents, *t,* and the nonresponse indicator, *I*, are unconditionally independent (Rubin and Little 2002). There are at least two reasons why (8) is unlikely in the case of the PES. First, there is strong practical evidence that Aboriginal and Torres Strait Islanders people living in remote communities and people aged 20–29 have a higher rate of being missed by the PES.

Second, it is reasonable to suppose that whether a person responds to the PES may be correlated in some way to whether the person responds to the Census. For example, people may avoid the PES interviewer specifically because they do not want to own-up to being a Census nonrespondent. This would lead to the PES sample having an unrepresentatively high rate of completed Census forms. This creates the ‘correlation bias’. In dual sampling literature this correlation is sometimes assumed to be negligible after conditioning on an appropriate set of covariates. However, for Census nonrespondents, there are effectively no covariates ($\boldsymbol{x}$) on which to condition.

In order to reduce any impact of the PES on the Census response (another potential form of ‘correlation bias’), the PES is conducted four weeks after Census night. Census records matched with PES records (i.e., $\mathbf{x}$ in this article) are those that were received before the date that PES field operations began, Census forms that were returned by mail or by Internet after this date are essentially ignored. In addition, data collected by the PES and Census are processed independently.

In short, for the reasons outlined above, nonresponse is *not* ignorable given $\boldsymbol{x}$. Given $\hat{T}^{(GREG)}$ is biased in this case, next, we consider an alternative.

3.3. Using a Two Stage Prediction Model

Now consider a vector $\mathbf{z}$ that is comprised of variables on people selected in the PES. Accordingly, $\mathbf{z}$ may be a function of $\boldsymbol{x}$ or t. Now, instead of (6) consider assuming

$$[t_j|z_j, I_j = 1] = [t_j|z_j] \text{ for } j = 1, \ldots N \tag{9}$$

$$[\boldsymbol{x}_j|z_j, I_j = 1] = [\boldsymbol{x}_j|z_j] \text{ for } j = 1, \ldots N \tag{10}$$

Assumption (9) is that the distribution of $t|z$ in the population and in the population of PES respondents is the same. Assumption (10) is that the distribution of $\boldsymbol{x}|z$ in the population and in the population of PES respondents is the same. If $\mathbf{z}$ is a function of only $\mathbf{x}$, then (9) and (10) collapse to (6). In an attempt to overcome the failings of (6), we allow $\mathbf{z}$ to be a function of t. By allowing $\mathbf{z}$ to be a function t we allow the indicator for nonresponse, I, to depend upon the response value itself (see Little and Rubin 2002). In this case, nonresponse is said to be nonignorable given $\boldsymbol{x}$.

To make the ideas more concrete, consider the underlying working model

$$t_j = z_j^T \boldsymbol{\theta} + \varepsilon_{1j} \tag{11}$$

$$\boldsymbol{x}_j = z_j^T \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{2j},$$

$$E(\varepsilon_{1j}|z_j) = 0$$

$$E(\boldsymbol{\varepsilon}_{2j}|z_j) = \mathbf{0}$$

where ε_{1j} and $\boldsymbol{\varepsilon}_{2j}$ are independent over j. Using (11) in a regression of t on $\boldsymbol{x}$, it follows that $\boldsymbol{\alpha}$ and e_j from (5) can be expressed by $\boldsymbol{\alpha}^* = \boldsymbol{\gamma}^{-1}\boldsymbol{\theta}$ and $e_j^* = \varepsilon_{1j} - \boldsymbol{\varepsilon}_{2j}\boldsymbol{\gamma}^{-1}\boldsymbol{\theta}$. We may then re-consider the working model of (5) and write

$$t_j = \boldsymbol{x}_j^T \boldsymbol{\alpha}^* + e_j^* \tag{12}$$

$$E\left(e_j^*|\boldsymbol{x}_j\right) = 0$$

We know from (9), (10), and (11) that (12) holds in the population and in the population of PES responders. So while (5) and (12) are both linear regressions with the same dependent and independent variables, only (12) holds in the population.

Now we may use (12) in a standard generalised regression estimator. Accordingly, unbiased estimates of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}^*$ are

$$\tilde{\boldsymbol{\theta}} = \left(\sum_{i \in s} w_i z_i z_i^T\right)^{-1} \left(\sum_{i \in s} w_i z_i t_i\right),$$

$$\tilde{\boldsymbol{\gamma}} = \left(\sum_{i \in s} w_i z_i z_i^T\right)^{-1} \left(\sum_{i \in s} w_i z_i \boldsymbol{x}_i\right)$$

and

$$\tilde{\boldsymbol{\alpha}}^* = \tilde{\boldsymbol{\gamma}}^{-1}\tilde{\boldsymbol{\theta}} = \left(\sum_{i \in s} w_i z_i \boldsymbol{x}_i^T\right)^{-1} \left(\sum_{i \in s} w_i z_i t_i\right)$$

and the classical regression estimator of T under (12) is

$$\hat{T}^{(PREG)} = \sum_{i \in s} w_i \left(t_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\alpha}}^* \right) + \boldsymbol{X}^T \tilde{\boldsymbol{\alpha}}^* \tag{13}$$

Since under (12), we know that $E(T) = \boldsymbol{X}^T \boldsymbol{\alpha}^*$ and $E\left(\sum_{i \in s} w_i e_j^* \right) = 0$ it follows that, since $\tilde{\boldsymbol{\alpha}}^*$ is unbiased for $\boldsymbol{\alpha}^*$, $\mathrm{E}(\hat{T}^{(PREG)}) = T$. Kott and Chang (2010) show that (13) is unbiased and note that for $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\gamma}}^{-1}$ to exist the inverse of $\sum_{i \in s} w_i \boldsymbol{z}_i \boldsymbol{x}_i^T$ must exist and so constrain the dimension of $\boldsymbol{x}$ and $\mathbf{z}$ to be the same. Since (13) is essentially a function of means, the jackknife can be used to give an asymptotically unbiased estimate of the sampling variance of $\hat{T}^{(PREG)}$ in large samples.

If we let the predicted value of $\boldsymbol{x}_i$ be $\tilde{\boldsymbol{x}}_i = \boldsymbol{z}_i^T \hat{\boldsymbol{\gamma}}$ it is easy to show that an alternative expression for (13) is $\hat{T}^{(PREG)} = \sum_{i \in s} w_i g_i t_i$, where $g_i = 1 + \left(\boldsymbol{X} - \sum_{i \in s} w_i \tilde{\boldsymbol{x}}_i \right)^T \left(\sum_{i \in s} w_i \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^T \right)^{-1} \tilde{\boldsymbol{x}}_i$ - the same as $\hat{T}^{(GREG)}$ but with $\tilde{\boldsymbol{x}}$ replacing $\boldsymbol{x}$. In other words, the Predicted value of $\boldsymbol{x}$ is used in an otherwise standard REGression estimator (PREG). In other words, the weight adjustment g_i depends upon z_i (via $\tilde{\boldsymbol{x}}_i$,) instead of $\boldsymbol{x}_i$ and so the weight adjustment does not depend on whether or not person i is a Census respondent. This is analogous to classic DSE, where membership of a subgroup defines the PES nonresponse adjustment for both the Census responders and nonrespondents.

In 2011, the ABS application of $\boldsymbol{x}$ and $\mathbf{z}$ contained the number of times a person was counted on a 'persons present' basis (collected by Census) and 'usual –residents' basis (collected by the PES) in each subgroup, respectively, where subgroup was defined by cross-classifying age, sex, Aboriginal and Torres Strait Islanders status collected by the PES. The dimension of $\boldsymbol{x}$ and $\mathbf{z}$ was about $H = 450$ (i.e., 450 subgroups indicators). Defining $\mathbf{z}$ completely in terms of PES counts allowed for nonignorable nonresponse in the PES.

The proposed estimator of (13) relies on assumptions similar to the DSE:

- Assumptions (9) and (10) correspond to the assumption of *independence* between the PES and the Census. If the independence assumption was violated then the distributions in the sample and populations would not be the same, as required by (9) and (10);
- The assumption of *perfect matching* between the PES and the Census ensures the values of $\boldsymbol{x}$ that are assigned to PES respondents are correct;
- The *closed population* assumption is implicit in the definition of the population set U. We assume that everyone who responds to the Census must be in U and all people in U have a chance of being selected by the PES. The closed population assumption will be violated (and (13) will be biased) if a Census respondent *does not* correctly identify as an overseas visitor and leaves Australia before the PES; and
- The *homogeneity of response* assumption is implicit in the response model of (11).

Chang and Kott (2008) consider a response propensity model, rather than a prediction model, such as (11), to justify (13). We may suppose that $p_j = E(I_j) = 1/f\left(z_j^T \boldsymbol{\theta} \right)$, where $f()$ is some appropriate nonlinear function and $\boldsymbol{\theta}$ is a matrix of coefficients. Chang and Kott (2008) show that under this response propensity model (and certain conditions), an unbiased estimator of T is

$$\hat{T}^{(RESP)} = \sum_{i \in s} w_i \, f\left(z_i^T \boldsymbol{\omega} \right) t_i$$

where $\boldsymbol{\omega}$ is an H vector (same dimension as $\mathbf{z}$) of constants that satisfies $\boldsymbol{X} = \sum_{i \in s} w_i f\left(z_j^T \boldsymbol{\omega}\right) \boldsymbol{x}_i$. Chang and Kott (2008) show that the estimators $\hat{T}^{(RESP)}$ and $\hat{T}^{(PREG)}$ have the same form if $f\left(z_j^T \boldsymbol{\omega}\right) = 1/\left(1 + z_j^T \boldsymbol{\omega}\right)$. Here, it may be more appropriate to consider a response propensity model rather than a linear model, such as (5), for a binary variable such as t. This will be the subject of future work.

4. Simulation to Demonstrate the PREG

This section describes a simple but illustrative simulation study of the PREG and GREG. Given that we are working in a situation in which there is no E-sample, the DSE was not evaluated. In this simulation, the population is made up of four subgroups of interest (i.e., there are four different ways of defining t). The subgroup population totals of interest are 5000 (Subgroup 1: non-Aboriginal and Torres Strait Islanders males), 2000 (Subgroup 2: Aboriginal and Torres Strait Islanders males), 2000 (Subgroup 3: non-Aboriginal and Torres Strait Islanders females), and 1000 (Subgroup 4: Aboriginal and Torres Strait Islanders females). The proportion of people in subgroup 1, 2, 3, 4 living in Aboriginal communities is 0.2, 0.8, 0.3, and 0.7, respectively. Using these proportions, each person in a subgroup is randomly assigned a value for c, where $c = 1$ if a person lives in an Aboriginal community and otherwise $c = 0$. In reality, Census and PES coverage of Aboriginal communities is potentially more difficult due to their remoteness and the wide geographic area that they cover.

The population is counted by the Census. Each person in the population is assigned a value for $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$ where x_{hj} is the number of times the person was counted in subgroup h by the Census. The type of errors in the Census counting include misclassification (records in subgroup $h = 2$ are misclassified to subgroup $h = 1$), duplication (person counted twice), and missing (person not counted). Consistent with earlier notation, if a person is missed, then $\boldsymbol{x} = \mathbf{0}$, if a person in subgroup $h = 1$ is duplicated in the subgroup, then the first element of $\boldsymbol{x}$ is 2 and all other elements are zero. The probability of these Census counting errors occurring depends only on c. Table 1 gives these probabilities for a range of scenarios. For example, in Scenario 1 the Census rate of misclassification, duplication and missing for people living in Aboriginal communities was 0.1.

A simulated PES sample of size 1,000 people was selected by SRSWOR from the population. From the sample the variable c and subgroup membership variables z_{hi} were collected, where $z_{1i} = 1$ of person i belonged to population h and is otherwise zero. However, the PES did miss people. As in the Census, the probability of the PES missing a person depended upon c. Table 1 gives these probabilities for a range of scenarios. For example, in Scenario 1 the probability of missing a person living within and outside an Aboriginal community was 0.1 and 0.05, respectively. Since it is difficult, in practice, to measure the probability of missing a person when the 'missing' mechanism itself is nonignorable, the range of scenarios aims to explore a wide scope of possibilities (rather than be motivated by a particular case study).

The PREG and GREG used the same definition of $\mathbf{x}$. The following estimators were considered:

PREG1: Equation (13) with $\mathbf{z} = 1$. This assumes that the PES provides a representative sample of t and of $\mathbf{x}$ from the population.

PREG2: Equation (13) with $\mathbf{z} = (1, r)$, where $r = z_2 + z_4$ is an Aboriginal and Torres Strait Islanders status indicator. If it is suspected that PES nonresponse depends on Aboriginal and Torres Strait Islanders status alone then there would be substantive reasons for defining $\mathbf{z}$ in this way. In this situation r is a good proxy for c, which drives the PES nonresponse mechanism.

PREG3: Equation (13) with $\mathbf{z} = (1, c)$. This is the estimator that would be used if the mechanism generating the count errors in Census and PES were known.

GREG: Generalised regression estimator with auxiliary $\mathbf{x}$. This estimator assumes that the distribution of $t|\mathbf{x}$ in the PES and in the population is the same.

The results are presented in Table 2 in terms of Relative Bias of $\hat{T} = (\hat{T} - T)/T$. Since the dimension of $\mathbf{x}$ and $\mathbf{z}$ were not the same, we calculated the generalised inverse of $\left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{x}_i^T\right)$ to calculate the PREG.

Table 2 shows that the GREG is consistently biased, but that reduced the bias in the observed Census counts. The table shows that PREG1 is biased because it is not well specified. PREG2 assumes PES nonresponse depends on ‘Aboriginal and Torres Strait

Table 2. Bias (%) in population estimates for various scenarios.

	GREG				PREG1			
	Subgroup							
Scenario	1	2	3	4	1	2	3	4
1	5.6	9.8	6.0	9.0	− 0.9	2.2	− 0.8	2.5
2	11.0	20.0	12.0	17.8	− 1.8	5.5	− 2.3	4.7
3	5.3	11.0	6.0	8.8	− 0.6	1.5	− 1.9	1.2
4	10.7	21.9	11.6	17.6	− 2.5	5.0	− 2.5	4.3
5	10.6	22.1	12.0	18.5	− 2.1	4.9	− 1.8	3.5
6	15.5	34.9	17.5	26.9	− 3.8	8.2	3.9	5.8
	PREG2				PREG3			
	Subgroup							
Scenario	1	2	3	4	1	2	3	4
1	− 0.4	0.3	0.5	− 0.1	− 0.3	− 0.3	0.1	0.5
2	− 0.0	− 0.1	− 0.4	− 3.1	− 0.1	− 0.1	0.3	0.3
3	− 0.4	0.6	1.5	0.5	− 0.2	− 1.2	− 0.7	− 0.2
4	− 0.1	− 0.3	− 0.3	0.4	− 0.8	1.5	0.7	1.7
5	0.2	0.6	1.2	1.8	0.1	− 0.3	− 0.4	0.0
6	− 0.9	− 0.1	− 0.0	1.3	0.0	− 1.3	− 0.6	0.5
	Census counts							
	Subgroup							
Scenario	1	2	3	4				
1	− 10.2	26.5	0.3	2.0				
2	− 9.1	29.0	0.4	− 1.5				
3	− 20.7	53.7	− 0.2	− 0.6				
4	− 29.5	80.8	− 0.5	− 1.9				
5	− 21.3	60.7	0.85	1.4				
6	− 27.5	84.5	0.9	0.3				

Islanders' status. Interestingly, PREG 2 only has a small biase because 'Aboriginal and Torres Strait Islanders' and 'Community' are well-correlated. As expected, PREG3 is unbiased in all scenarios, as it correctly assumes that PES nonresponse is due to 'community'.

5. Set-Up of the Simulation

Subsection 5.1 creates a synthetic version of the population, U, and simulates Census counting of the population. Subsection 5.2 simulates PES sampling from the synthetic population. Subsection 5.3 evaluates the GREG and PREG. The aim was for the simulation to be realistic.

5.1. Simulating the Population

Records of the 2001 Census define the synthetic population U and subgroup t. The different subgroup totals of interest, T, are given in Tables 3 and 4 (e.g., in the last row of Table 3, t is defined as the membership indicator for the Australian Capital City and T is the population total of the Australian Capital City).

Define $\mathbf{z}$ to be a $H = 449$ vector of 2001 Census variables given by Aboriginal and Torres Strait Islanders status and the cross-classification of region, sex, and age. This defines $\mathbf{z}_j$ for all $j \in U$. Given that t and $\mathbf{z}$ are defined in terms of the same source of data (i.e., the 2001 Census), their values are consistent.

In the simulation, $\mathbf{x}$ has the same categories as $\mathbf{z}$. However, to allow for errors in the Census counting, $\mathbf{x}$ and $\mathbf{z}$ can be different. To do this, we used the 2001 PES to model the following three probabilities, conditional on a range of dwelling and person-level covariates (including covariates in $\mathbf{z}$) from the 2001 Census:

$p^{(1)}$ = the probability that a person was counted correctly by the Census (i.e., Census and PES classification is the same),

$p^{(2)}$ = the probability that a person is misclassified by the Census (i.e., Census and PES classification is *not* the same), and

$p^{(3)}$ = the probability that a person is missed (i.e., not counted) by the Census.

Table 3. Relative Bias (RB), Relative Root Mean Squared Error (RRMSE) and Relative Standard Error at State and National Level.

	RB (%)		RSE (%)		RRMSE (%)	
Subgroup	PREG	GREG	PREG	GREG	PREG	GREG
Australia	−0.03	−0.02	0.10	0.10	0.11	0.10
New South Wales	−0.03	−0.01	0.21	0.21	0.22	0.21
Victoria	−0.03	−0.02	0.21	0.20	0.21	0.20
Queensland	−0.03	−0.01	0.25	0.25	0.25	0.25
South Australia	−0.04	−0.04	0.28	0.27	0.28	0.28
Western Australia	−0.06	−0.05	0.30	0.29	0.30	0.30
Tasmania	−0.07	−0.07	0.40	0.39	0.41	0.40
Northern Territory	−0.11	−0.08	1.00	0.97	1.01	0.98
Australian Capital Territory	−0.04	−0.05	0.53	0.51	0.53	0.52

Table 4. Mean squared error and bias for other subgroups.

	MSE (relative to PR)		Bias (%)	
	PREG	GREG	PREG	GREG
Male, age 0–19	100	94	−0.04	0.00
Male, age 20–29	100	92	−0.03	0.24
Male, age 30–59	100	89	−0.03	0.05
Male, age 60+	100	93	−0.03	0.27
Female, age 0–19	100	97	−0.02	−0.10
Female, age 20–29	100	91	0.06	0.02
Female, age 30–59	100	104	−0.01	−0.17
Female, age 60+	100	91	0.04	0.06
Not Aboriginal and Torres Strait Islanders	100	111	−0.02	−0.07
Aboriginal and Torres Strait Islanders	100	139	0.08	2.80
Born in Australia	100	100	−0.01	0.00
Born overseas	100	99	0.00	0.02
Not married	100	100	−0.02	0.00
Married	100	97	−0.02	−0.01

Each person in the simulated population was assigned these three probabilities under the model, giving $\left(p_j^{(1)}, p_j^{(2)}, p_j^{(3)}\right)$ for all $j \in U$. Then the value of $\mathbf{x}_j$ for population record j was equal to:

- $\mathbf{z}_j$ with probability $p_j^{(1)}$.
- $\mathbf{z}_j^*$ with probability $p_j^{(2)}$, where $\mathbf{z}_j^*$ is same as $\mathbf{z}_j$ but changed to an 'adjacent' category in one dimension (e.g., if $\mathbf{z}_j$ indicates a male aged 20–25, then $\mathbf{z}_j^*$ may indicate a male aged 26–30, where the dimension that is changed is age and the 'adjacent' categories are 20–25 and 26–30).
- $\mathbf{0}$ with probability $p_j^{(3)}$
- $2\mathbf{z}_j$ with probability $1 - p_j^{(1)} - p_j^{(2)} - p_j^{(3)}$ (i.e., person j is counted twice).

Finally, all people in the simulated population were assigned a value for η_j, the probability that person j would respond to the PES. The important point here is that η_j was allowed to be a function of $\mathbf{z}$. It is worth noting here that Aboriginal and Torres Strait Islanders status had a strong influence on the propensity to respond. (The coefficients in the propensity model were obtained from the logistic regression of 2001 Census response propensity using the 2001 PES. One difference between the variables in the models used to obtain $p_j^{(3)}$ and η_j is that the latter included additional dwelling and person-level covariates, such as whether born outside of Australia, dwelling type, and marital status).

5.2. Simulating the PES Samples

Repeated PES samples of size 90,000 people were drawn from the synthetic population. The simulated PES sampling scheme was designed to mimic the actual PES sampling scheme. The first stage of this sampling scheme divides the Census Collector's Districts

(CDs) into strata, and chooses a sample of these CDs with probability proportional to the number of dwellings in the CDs. The second stage divides the CDs into blocks, and selects a block at random. Finally, a cluster of dwellings is selected within each selected block by skipping through the list of dwellings. The skip lengths are such that each dwelling has an equal probability of selection in a state. Once the PES sample was selected from the synthetic population, each selected person was randomly assigned to be a PES nonrespondent with the probability 1-η_j. The variable t was collected from PES respondents.

Given that the propensity to respond to the PES depends upon on **z** (and not **x**), nonresponse is nonignorable given **x**. If the propensity to respond was based only on **x**, nonresponse would be ignorable given **x**. The nonresponse rate was simulated to be about 94% (of those who were contacted by the 2011 PES, 94% responded).

5.3. *Evaluation of Alternative Estimators Using Simulation*

The estimators in the simulation are:

GREG: Generalised regression estimator, $\hat{\boldsymbol{T}}^{(\boldsymbol{GREG})}$ where **x** is defined in Subsection 5.1.

PREG: The proposed regression estimator, $\hat{\boldsymbol{T}}^{(\boldsymbol{PREG})}$ where **x** and **z** are defined in Subsection 5.1 This definition of **x** and **z** was used in estimation for the actual 2011 PES.

For each of 1,000 simulated PES samples, we calculated the

- Relative Bias (RB) $(\hat{\mathrm{T}} - \mathrm{T})/\mathrm{T}$
- Relative Standard Error (RSE) of $\hat{\mathrm{T}} = \sqrt{\widehat{\mathrm{Var}(\hat{T})}/T}$
- Relative Root Mean Squared Error (RRMSE) of $\hat{\mathrm{T}} = \sqrt{\mathrm{Var}(\hat{T} - T)/T}$

for the GREG and PREG estimators, where $\widehat{\mathrm{Var}(\hat{T})}$ is calculated using the group jackknife. Table 3 gives the average RB, RSE and RRMSE over these 1,000 simulations. Table 3 shows that GREG and PREG perform equally well. However, Table 4 shows that PREG outperforms GREG for estimates of the Aboriginal and Torres Strait Islanders population. This is driven by the fact that the Aboriginal and Torres Strait Islanders population was simulated to have significant and nonignorable influence on PES nonresponse status. While the results are not shown here, the coverage rates of PREG were close to their nominal level of 95%.

6. Discussion

The development of the ABS's Census coverage estimation strategy for Australia has been driven by Census counting on a 'persons present' basis and PES counting on a 'usual residents' basis, Census misclassification, and by nonignorable nonresponse in the PES.

The estimator proposed here also has the potential to aid population estimation when using an imperfect administrative source as the basis, rather than a traditional Census. In such a situation, the estimation strategy has to deal with individuals having some characteristics poorly defined and being in the wrong locations in the administrative data, as well as having people completely missing from the administrative data.

This concept of a survey for quality assessment of administrative data has been proposed in Brown and Honchar (2012), where they suggest the ABS PREG estimator as an approach.

Completely erroneous returns was part of the reason for the E-sample, as adopted by the US Census. In part because there is no E-sample, the proposed estimator is biased if there are completely erroneous returns. More generally, it is biased if people who cannot be counted by the PES can be counted by the Census. Although there is no evidence of an erroneous returns problem in the Australian context, it does mean that temporary residents should be excluded (i.e., excluded from **X**). In the context of estimating people counts from an administrative list (instead of a Census), this means that individuals who have emigrated, but erroneously remain on the list, need to be removed using other approaches, such as evidence of activity within the system prior to estimation.

As discussed, the PREG relies on the same assumptions as classic DSE, including independence between the Census and PES. Gerritse et al. (2015) explicitly explored issues of dependence with two lists. More generally, within the classic capture-recapture framework work has been done exploring the use of multiple systems to tackle the issue of dependence. A recent review is given by Baffour et al. (2013); while Zhang (2015) makes an interesting contribution to using multiple administrative sources with a survey to deal with completely erroneous returns. Clearly, looking at how the ABS PREG can fit into this multiple system situation is important future work, as many countries now using a traditional Census are looking to use multiple administrative sources as an alternative.

7. References

Alho, J.M., M.H. Mulry, K. Wurdeman, and J. Kim. 1993. “Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation.” *Journal of the American Statistical Association* 88: 1130–1136.

Baffour, B., J.J. Brown, and P.W.F. Smith. 2013. “An Investigation of Triple System Estimators in Censuses.” *Statistical Journal of the IAOS* 1: 53–68. Doi: http://dx.doi.org/10.3233/SJI-130760.

Belin, T.R. and J.E. Rolph. 1994. “Can We Reach Consensus on Census Adjustment?” *Statistical Science* 9: 486–508. Doi: http://dx.doi.org/10.1214/ss/1177010261.

Bell, W. 2001. *ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Estimates Using Revised Demographic Analysis Results*. Report No. 10 to the Executive Steering Committee for A.C.E. Policy II, U.S. Bureau of the Census.

Brown, J., O. Abbott, and I. Diamond. 2006. “Dependence in the 2001 One-Number Census Project.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169: 883–902. Available at: URL: http://www.jstor.org/stable/3877405.

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. “A Methodological Strategy for a One-Number Census in the UK.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162: 247–267. Available at: URL: http://www.jstor.org/stable/2680581.

Brown, J.J. and O. Honchar. 2012. "Design and Estimation of Surveys to Measure Data Quality Aspects of Administrative Data." *Lithuanian Journal of Statistics* 51: 5–16.

Chang, T. and P.S. Kott. 2008. "Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model." *Biometrika* 95: 557–571. Available at: http://www.jstor.org/stable/20441486.

Gerritse, S.C., P.G.M. van der Heijden, and B.F.M. Bakker. 2015. "Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models." *Journal of Official Statistics* 31: 357–379. Doi: http://dx.doi.org/10.1515/jos-2015-0022.

Griffin, R.A. 2000. "Accuracy and Coverage Evaluation: Dual System Estimation." DSSD Census 2000 Procedures and Operations Memorandum Series, Q-20, US Census Bureau.

Griffin, R.A. and T. Mule. 2008. DSSD 2010 Census Coverage Measurement Memorandum #2010-E-20. Available at: http://www.census.gov/coverage_measurement/post_enumeration_surveys/2010_results.html (accessed December 2015).

Hogan, H. 1993. "The 1990 Post Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060.

Hogan, H. 2003. "The Accuracy and Coverage Evaluation: Theory and Design." *Survey Methodology* 29: 129–138.

Kott, P.S. and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105: 1265–1275. Doi: http://dx.doi.org/10.1198/jasa.2010.tm09016.

Large, A., J.J. Brown, O. Abbott, and A. Taylor. 2011. "Estimating and Correcting for Over-count in the 2011 Census." *Survey Methodology Bulletin* 69: 35–48.

Mule, T. 2008. DSSD 2010 Census Coverage Measurement Memorandum #2010-E-18. Available at: https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf (accessed January 2017).

Nordholt, E.S. 2005. "The Dutch Virtual Census 2001: A New Approach by Combining Different Sources." *Statistical Journal of the United Nations Economic Commission for Europe* 22: 25–37.

Rubin, D.B. and R.J.A. Little. 2002. *Statistical Analysis of Missing Data (2nd Edition)*. New Jersey: John Wiley and Sons.

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

Sekar, C.C. and W.E. Deming. 1949. "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44: 101–115. Doi: http://dx.doi.org/10.2307/2280353.

Valente, P. 2010. "Census Taking in Europe: How Are Populations Counted in 2010?" *Population & Societies* 467: 1–4.

Wachter, K.W. and D.A. Freedman. 2000. "The Fifth Cell: Correlation Bias in U.S. Census Adjustment." *Evaluation Review* 24: 191–211. Doi: Http://dx.doi.org/10.1177/0193841X0002400202.

Wolter, K.M. 1990. "Capture–Recapture Estimation in the Presence of a Known Sex Ratio." *Biometrics* 46: 157–162. Doi: http://dx.doi.org/10.2307/2531638.

Xi Chien, S. and C.Y. Tang. 2011. "Properties of Census Dual System Population Size Estimators." *International Statistical Review* 79: 336–361. Doi: http://dx.doi.org/10.1111/j.1751-5823.2011.00150.x.

Zhang, Li-Chun. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://dx.doi.org/10.1515/jos-2015-0023.

Received February 2015
Revised September 2016
Accepted October 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 61–77, http://dx.doi.org/10.1515/JOS-2017-0004

Space-Time Unit-Level EBLUP for Large Data Sets

Michele D'Alò[1], *Stefano Falorsi*[1], *and Fabrizio Solari*[1]

Most important large-scale surveys carried out by national statistical institutes are the repeated survey type, typically intended to produce estimates for several parameters of the whole population, as well as parameters related to some subpopulations. Small area estimation techniques are becoming more and more important for the production of official statistics where direct estimators are not able to produce reliable estimates. In order to exploit data from different survey cycles, unit-level linear mixed models with area and time random effects can be considered. However, the large amount of data to be processed may cause computational problems. To overcome the computational issues, a reformulation of predictors and the correspondent mean cross product estimator is given. The R code based on the new formulation enables the elaboration of about 7.2 millions of data records in a matter of minutes.

Key words: Small area estimation; time series; linear mixed model; small area estimation software.

1. Introduction

Large-scale surveys are usually aimed at providing estimates of target parameters for the whole population, as well as for relevant subpopulations defined at the sampling stage. Design-consistent and design-unbiased direct estimates are produced for the parameters of interest. However, in most surveys, the sample size is not large enough to guarantee reliable estimates for all the target subpopulations. When direct estimates cannot be provided, small area estimation (SAE) methods should be used to overcome the problem (see Rao 2003; Pfeffermann 2002, 2013). SAE methods, usually referred to as indirect estimators, cope with the lack of information from each domain by borrowing strength from samples that belong to other domains, with the result that it increases the effective sample size for each small area.

The most important surveys carried out by national statistical institutes are repeated surveys (see Duncan and Kalton 1987, and Kish 1987). The repeated nature of these surveys allows them to borrow strength not only from other areas but also from other survey cycles.

[1] Italian National Statistical Institute, via Cesare Balbo 16, 00184 Rome, Italy. Emails: dalo@istat.it, stfalors@istat.it, and solari@istat.it.

Acknowledgments: We would like to express our deepest appreciation to the reviewers and the associated editor for providing insightful comments and providing directions for improving the quality of this article. A special thank is also extended to Professor Maria Giovanna Ranalli for her support and for carefully aimed comments and suggestions.

In this context, Saei and Chambers (2003) proposed the use of unit-level linear mixed models (LMMs) with area and time random effects. However, this presents a computational challenge, since large amounts of data from different survey cycles have to be processed. The aim of this article is propose a method to overcome computational problems that may arise from using the predictors and correspondent errors given by Saei and Chambers (2003). For this reason, a reformulation of these expressions will be presented. Furthermore, these more efficient expressions will be applied to the estimation of the unemployment rate at Labour Market Area (LMA) level, using data from the Italian Labour Force Survey (LFS). The case study aims to show the potential gains in efficiency as a result of SAE methods borrowing strength from space and time. It does not aim to suggest a ready solution for official LFS statistics, which necessarily involves many other issues and considerations that are outside the scope of this article.

The LFS is a quarterly survey based on a two-stage stratified cluster design. Municipalities are the primary sampling units, and households are the secondary sampling units. The survey follows a rotating panel sample design, according to the rotation design 2-(2)-2. Households are interviewed in two consecutive quarters. After a two-quarter break, they are interviewed for an additional two consecutive quarters. The sample is uniformly spread across all the weeks, such that all territorial domains are represented in each month and in each of the four waves. The LFS is the main source of information on the Italian labour market and aims to produce monthly, quarterly, and yearly estimates of employment, unemployment, and inactivity rates for different planned territorial domains. Each sample contains information about approximately 170,000 respondents. LMAs, on the other hand, are unplanned areas that are defined every ten years based on daily commuting flows detected by the Population Census. At present, there are 611 LMAs, of which about 450 are included in at least one of the LFS samples in the years 2004–2014. The most unstable estimates refer to the estimation of the unemployment rate. In this case, the Coefficient of Variations (CVs) of the direct estimates are very large, and about three out of four CVs are larger than 30%. Therefore, SAE methods are needed in order to obtain more precise estimates of the unemployment rate that are suitable for dissemination. However, the areas are sampled with unequal selection probabilities in relation to the values of the target variable values. In such situations, standard SAE methods are biased; the magnitude of the bias depends on the sampling fraction and the covariance between the sampling weights and the target variable. However, in the LFS, bias resulting from informative sampling is considered to be small. Treatment of informative sampling in SAE is not considered in this article.

As mentioned above, when LMMs with area and time random effects are assumed, computational problems may result from the large amounts of data to be used in the estimation process. For instance, the data used in this article comes from the 44 LFS quarterly samples in 2004 to 2014, and the overall data size processed comprises about 7,200,000 records.

Usually, in order to overcome the computational problems deriving from large data sets, area-level models are applied. For instance, Rao and Yu (1994) proposed an extension of the basic Fay-Herriot (Fay and Herriot 1979) model to handle time series and cross-sectional data by means of an AR(1) model specification. Datta et al. (2002) and You (1999) used the Rao-Yu model but replace the AR(1) model specification with a random walk model. Pfeffermann and Burck (1990) proposed a general model involving area-by-time specific random effects. Hidiroglou and You (2016) compared the performances of unit- and area-level models, showing that the former outperforms the latter in terms of bias and mean squared error. Furthermore, Gershunskaya (2015) showed that due to errors associated with the variance of direct estimates, in terms of mean squared error, there is no benefit to introducing temporal correlations between small areas over using the regular Fay-Herriot model. The benefits only become apparent when theoretical variances of direct estimates are used in Rao-Yu model specification.

To avoid the computational issues related to unit-level LMMs, formulas given in Saei and Chambers (2003) have been rewritten in order to involve only small dimensional matrices. The revised expressions, implemented in the ad hoc R function, enable the processing of millions of survey records from different survey cycles in a matter of minutes.

The two-way unit-level linear mixed model with area and time random effects is described in Section 2, while Section 3 is devoted to the reformulation of the expressions needed to compute small area estimates and errors. Section 4 describes some particular SAE methods obtained from the general model. Section 5 includes a case study based on LFS data that aimed to compare the empirical performances of alternative model specifications. Section 6 compares the computational performances of the available SAE software tools with the R function implementing the new expression presented in Section 3. In conclusion, Section 7 presents the most important conclusions of the work.

2. Two-Way Linear Mixed Model

Let d ($d = 1, \ldots, D$) and t ($t = 1, \ldots, T$) denote the generic domain and time indices respectively. For domain d and time t, let N_{dt} and n_{dt} denote population and sample sizes, respectively, and let y_{dti} be the observed value of the target variable for the generic unit i. The parameter of interest is the vector $\boldsymbol{\theta}$ including the population means $\bar{y}_{dt} = (1/N_{dt}) \sum_i y_{dti}$, for all domains and times ($d = 1, \ldots, D$, $t = 1, \ldots, T$). Other relevant parameters for large-scale repeated surveys, such as totals, or net changes between two survey cycles, can be expressed as a linear combination of $\boldsymbol{\theta}$. For this reason, the results in this article can be easily extended to the other types of parameters.

Let us suppose that the data follows the two-way unit-level additive LMM (see Searle et al. 1992; Saei and Chambers 2003)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \tag{1}$$

where $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$ are known full rank matrices, and $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{e}$ are random vectors, independently distributed from each other. The random effect vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, modeling between area and time variations not explained by fixed effects, include D and T levels respectively. Furthermore, we assume for $\alpha = 1, 2$, $\mathbf{u}_\alpha \sim N(\mathbf{0}, \mathbf{G}_\alpha)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, where the covariance matrices $\mathbf{G}_\alpha = \sigma_\alpha^2 \boldsymbol{\Omega}_\alpha(\rho_\alpha)$ and $\mathbf{R} = \sigma^2 \mathbf{W}^{-1}$, with $\mathbf{W}$ as a known diagonal matrix. In particular, for $\alpha = 1, 2$, σ_α^2 and ρ_α denote, respectively, the variance and a measure of correlation for the elements of $\mathbf{u}_\alpha$, while σ^2 is the variance of the generic element of $\mathbf{e}$. For notational simplicity, it will be useful to introduce the parametrisation $\phi_\alpha = \sigma_\alpha^2/\sigma^2$. Hence, $\mathbf{y}$ is $N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\omega})$ given by

$$\boldsymbol{\Sigma}(\boldsymbol{\omega}) = \sigma^2(\mathbf{W}^{-1} + \mathbf{Z}\boldsymbol{\Omega}\boldsymbol{\Omega}'),$$

where $\boldsymbol{\omega} = (\sigma^2, \phi_1, \rho_1, \phi_2, \rho_2)$ is the overall variance component vector, $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2]$, and $\boldsymbol{\Omega} = diag_\alpha\{\phi_\alpha \boldsymbol{\Omega}_\alpha(\rho_\alpha)\}$. The uncorrelated random effect case is obtained by setting $\boldsymbol{\Omega}_1(0) = \mathbf{I}_D$ and $\boldsymbol{\Omega}_2(0) = \mathbf{I}_T$. For models with correlated area effects and correlated time effects, different structures of covariance matrices of random effects can be assumed. For example, $\boldsymbol{\Omega}_1(\rho_1)$ may depend on the distances among the areas, while $\boldsymbol{\Omega}_2(\rho_2)$ may follow an auto-regressive model.

Once the sample is collected, it is useful to partition Model (1) into two parts, depending on whether or not units are observed. In the following, we use the subscripts s and r to refer to sampled and nonsampled population units, respectively. The predicted values for nonsampled population units of $\boldsymbol{\eta}_r = E[\mathbf{y}_r | \mathbf{X}_r, \boldsymbol{\beta}, \mathbf{u}] = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u}$ are (see Royall 1976)

$$\tilde{\boldsymbol{\eta}}_r(\boldsymbol{\omega}) = \mathbf{X}_r\tilde{\boldsymbol{\beta}} + \mathbf{Z}_r\tilde{\mathbf{u}}, \tag{2}$$

where $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})$, the Best Linear Unbiased Estimator (BLUE) of β, is given by

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}_s^{\prime}\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s\right]^{-1}\mathbf{X}_s^{\prime}\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{y}_s,$$

and $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\omega)$, the Best Linear Unbiased Predictor (BLUP) of $\mathbf{u} = \left[\mathbf{u}'_1\mathbf{u}'_2\right]'$, is

$$\tilde{\mathbf{u}} = \boldsymbol{\Omega}\mathbf{Z}_s^{\prime}\boldsymbol{\Sigma}_{ss}^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\tilde{\boldsymbol{\beta}}\right).$$

Then, the BLUP of the target parameter $\boldsymbol{\theta}$ is

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega}) = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\tilde{\boldsymbol{\eta}}_r(\boldsymbol{\omega}), \tag{3}$$

where matrices $\mathbf{L}_s$ and $\mathbf{L}_r$ have the block-wise structure $diag_d\left\{diag_t\left\{\mathbf{l}_{dt}^{\prime}\right\}\right\}$, being $\mathbf{l}_{dt}^{\prime} = N_{dt}^{-1}\mathbf{1}_{n_{dt}}^{\prime}$ and $\mathbf{l}_{dt}^{\prime} = N_{dt}^{-1}\mathbf{1}_{N_{r,dt}}^{\prime}$ for $\mathbf{L}_s$ and $\mathbf{L}_r$, respectively, and $N_{r,dt} = N_{dt} - n_{dt}$ is the number of nonsampled units in area d at time t.

The BLUP estimator $\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega})$, given in (3), depends on the variance component vector $\boldsymbol{\omega}$, which is unknown in practical applications. By replacing $\boldsymbol{\omega}$ by an estimator, $\hat{\boldsymbol{\omega}}$, a two stage estimator called the Empirical Best Linear Unbiased Predictor (EBLUP) is obtained. Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and the method of fitting constants can be applied to the estimation of fixed effects and variance components (for details see Harville 1977; Searle et al. 1992; Cressie 1992; Rao 2003; Saei and Chambers 2003). Then, the EBLUP of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\omega}}) = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\hat{\boldsymbol{\eta}}_r(\hat{\boldsymbol{\omega}}),$$

where $\hat{\boldsymbol{\eta}}_r(\hat{\boldsymbol{\omega}})$ is the EBLUP correspondent to (2).

3. Reformulation

In this section, computationally more efficient expressions for predicted area means and the mean cross product error are derived. Results 1 to 5 consist in rewriting the expressions given in Saei and Chambers (2003) as a function of terms dependent on area and time level matrices instead of unit-level matrices. In particular, Result 1 gives the expression of the predicted value for $\mathbf{u}_1$ and $\mathbf{u}_2$, while Result 2 provides the estimate of the regression coefficient $\boldsymbol{\beta}$. Result 3 gives the mean cross-product error (MCPE) for the BLUP of $\boldsymbol{\theta}$. Result 4 computes the expression for updating the variance component estimates when

EBLUP is performed, and finally Result 5 provides the MCPE of the EBLUP of $\boldsymbol{\theta}$. For the sake of simplicity and with obvious notation, we will use matrix operators col{ · }, row{ · }, diag{ · }, and matr{ · }. Different types of correlation matrices $\boldsymbol{\Omega}(\rho)$ can be used for both area and time effects, provided that they depend on a one-dimensional correlation parameter ρ. For instance, the spatial correlation can be specified either as a SAR model based on an adjacency matrix (Cressie 1993), or as exponential or gaussian correlation structures, while the time correlation can follow an AR(1) process.

Two alternative cases for fixed effects are considered. In the first case (Case A), a different regression coefficient vector $\boldsymbol{\beta}_t$, of dimension K, is defined for each time t, determining $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T'\right)$ to be a $(T \times K)$-dimensional vector. In the second case (Case B), a common regression coefficient vector $\boldsymbol{\beta}$, of dimension K, is considered for all times t. The block-wise structure of matrix $\mathbf{X}$ under the two cases is given by

$$\mathbf{X} = \begin{cases} \mathrm{col}_d\{\mathrm{diag}_t\{\mathbf{X}_{dt}\}\}, & \text{for case } A \\ \mathrm{col}_d\{\mathrm{col}_t\{\{\mathbf{X}_{dt}\}\}, & \text{for case } B \end{cases},$$

where $\mathbf{X}_{dt}$ is the $N_{dt} \times K$ design matrix for area d and time t. The ith row of $\mathbf{X}_{dt}$ is $\mathbf{x}_{dti} = (x_{dti,1}, \ldots, x_{dti,K})'$.

For the random effect part of the model, $\mathbf{u} = col_\alpha\{\mathbf{u}_\alpha\}$, and $\mathbf{Z} = row_\alpha\{\mathbf{Z}_\alpha\}$, where

$$\mathbf{Z}_\alpha = \begin{cases} \mathrm{diag}_d\{\mathrm{col}_t\{\mathbf{1}_{N_{dt}}\}\}, & \text{for } \alpha = 1 \\ \mathrm{col}_d\{\mathrm{diag}_t\{\mathbf{1}_{N_{dt}}\}\}, & \text{for } \alpha = 2 \end{cases}.$$

Finally, $\mathbf{W} = \mathrm{diag}_d\{\mathrm{diag}_t\{\mathbf{W}_{dt}\}\}$ in which $\mathbf{W}_{dt}$ is a diagonal N_{dt} − dimensional matrix, whose generic element, $w_{dti}(i = 1, \ldots, N_{dt})$, is a known constant expressing the heteroscedasticity weight for the unit i in area d at time t.

It is worthwhile to note that matrices and vectors partitioned into sampled and nonsampled units have the same block-wise matrix structure of the corresponding nonpartitioned matrices and vectors, but matrices or vectors referred to area d and time t are, respectively, of size N_{dt} and $N_{r,dt}$ instead of N_{dt}.

Let us define the following quantities referred to as area d and time t:

$$f_{dt} = n_{dt}/N_{dt},$$

$$\bar{y}_{s,dt} = n_{dt}^{-1} \sum_i y_{s,dti},$$

$$\bar{y}_{w,dt} = w_{dt}^{-1} \sum_i w_{dti} y_{dti},$$

$$\bar{\mathbf{x}}_{w,dt} = w_{dt}^{-1} \sum_i w_{dti} \mathbf{x}_{dti},$$

$$\bar{\mathbf{x}}_{r,dt} = N_{r,dt}^{-1} \sum_{\mathbf{i}} \mathbf{x}_{r,dti}.$$

Then, the general aggregated expression of $\tilde{\boldsymbol{\theta}}(\boldsymbol{\omega})$ is

$$\tilde{\boldsymbol{\theta}} = \mathrm{col}_d\left\{\mathrm{col}_t\{\tilde{\bar{y}}_{dt}\}\right\},$$

where $\tilde{\bar{y}}_{dt} = \tilde{\bar{y}}_{dt}(\boldsymbol{\omega})$ is

$$\tilde{\bar{y}}_{dt} = f_{dt}\bar{y}_{s,dt} + (1 - f_{dt})\left(\bar{\mathbf{x}}'_{r,dt}\tilde{\boldsymbol{\beta}} + \tilde{u}_{1,d} + \tilde{u}_{2,t}\right), \tag{4}$$

in which $\tilde{u}_{1,d}$ and $\tilde{u}_{2,t}$ are the dth and tth element of $\tilde{\mathbf{u}}_{\alpha}$, $\alpha = 1, 2$. Let us define $\mathbf{T}^* = \mathbf{T}^*(\boldsymbol{\omega})$ as

$$\mathbf{T}^* = \left[\mathbf{Z}'_s\mathbf{W}_s\mathbf{Z}_s + \boldsymbol{\Omega}^{-1}\right]^{-1}$$

$$= \begin{bmatrix} \mathrm{diag}_d\{w_d\} + \phi_1^{-1}\boldsymbol{\Omega}_1^{-1}(\rho_1) & \mathrm{matr}_{dt}\{w_{dt}\} \\ \mathrm{matr}_{td}\{w_{dt}\} & \mathrm{diag}_t\{w_t\} + \phi_2^{-1}\boldsymbol{\Omega}_2^{-1}(\rho_2) \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{T}^*_{11} & \mathbf{T}^*_{12} \\ \mathbf{T}^*_{21} & \mathbf{T}^*_{22} \end{bmatrix},$$

being $w_d = \sum_t w_{dt}$ and $w_t = \sum_d w_{dt}$, in which $w_{dt} = \sum_i w_{dti}$. Note that $\mathrm{matr}_{td}\{w_{dt}\}$ is the transpose of $\mathrm{matr}_{dt}\{w_{dt}\}$.

Result 1. The predicted values $\tilde{\mathbf{u}}_{\alpha} = \tilde{\mathbf{u}}_{\alpha}(\boldsymbol{\omega})$, $\alpha =$ 1,2, are obtained as

$$\tilde{\mathbf{u}}_{\alpha} = \mathbf{T}^*_{\alpha 1} \cdot \mathrm{col}_d\left\{w_d\tilde{\bar{e}}_{w,d}\right\} + \mathbf{T}^*_{\alpha 2} \cdot \mathrm{col}_t\left\{w_t\tilde{\bar{e}}_{w,t}\right\}, \tag{5}$$

for $w_d\tilde{\bar{e}}_{w,d} = \sum_t w_{dt}\tilde{\bar{e}}_{w,dt}$ and $w_t\tilde{\bar{e}}_{w,t} = \sum_d w_{dt}\tilde{\bar{e}}_{w,dt}$, being $\tilde{\bar{e}}_{w,dt} = \tilde{\bar{e}}_{w,dt}(\boldsymbol{\omega})$ given by $\tilde{\bar{e}}_{w,dt} = \bar{y}_{w,dt} - \bar{\mathbf{x}}'_{w,dt}\tilde{\boldsymbol{\beta}}$.

Result 2. When case A is considered, the aggregated expression of $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})$ is

$$\tilde{\boldsymbol{\beta}} = [\mathbf{B}_{s,11} - \tilde{\mathbf{B}}_{s,12}]^{-1}[\mathbf{b}_{s,21} - \tilde{\mathbf{b}}_{s,22}], \tag{6}$$

being

$$\mathbf{B}_{s,11} = \mathrm{diag}_t\left\{\sum_d\sum_i w_{dti}\mathbf{x}_{dti}\mathbf{x}'_{dti}\right\}, \tag{7}$$

$$\mathbf{b}_{s,21} = \mathrm{col}_t\left\{\sum_d\sum_i w_{dti}\mathbf{x}'_{dti}\mathbf{y}_{dti}\right\}, \tag{8}$$

$$\tilde{\mathbf{B}}_{s,12} = \mathbf{B}_{\bar{x}_w}\mathbf{T}^*\mathbf{B}'_{\bar{x}_w},$$

$$\tilde{\mathbf{b}}_{s,22} = \mathbf{B}_{\bar{x}_w}\mathbf{T}^*\mathbf{b}_{\bar{y}_w},$$

where

$$\mathbf{B}_{\bar{x}_w} = [\mathrm{matr}_{td}\{w_{dt}\bar{\mathbf{x}}_{w,dt}\}, \mathrm{diag}_t\{w_t\bar{\mathbf{x}}_{w,t}\}],$$

$$\mathbf{b}_{\bar{y}_w} = [\mathrm{row}_d\{w_d\bar{y}_{w,d}\}, \mathrm{row}_t\{w_t\bar{y}_{w,t}\}]',$$

Under Case B, the external block-wise matrix operators in (7) and (8), $\mathrm{diag}_t\{\cdot\}$ and $\mathrm{col}_t\{\cdot\}$, are substituted by $\sum_t\{\cdot\}$, $\mathbf{B}_{\bar{x}_w}= [\mathrm{row}_d\{w_d\bar{\mathbf{x}}_{w,d}\}, \mathrm{row}_t\{w_t\bar{\mathbf{x}}_{w,t}\}]$, and $\mathbf{b}_{\bar{y}_w}$ does not change.

Result 3. Following Saei and Chambers (2003), the MCPE matrix of the BLUP $\tilde{\boldsymbol{\theta}}$ is given by

$$\mathrm{MCPE}(\tilde{\theta}) = \mathrm{E}[(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{'}] = \mathbf{G}_1(\boldsymbol{\omega}) + \mathbf{G}_2(\boldsymbol{\omega}) + \mathbf{G}_4(\boldsymbol{\omega}), \tag{9}$$

where the aggregated expressions of $\mathbf{G}_1(\boldsymbol{\omega})$, $\mathbf{G}_2(\boldsymbol{\omega})$ and $\mathbf{G}_4(\boldsymbol{\omega})$ are:

$$\mathbf{G}_1(\boldsymbol{\omega}) = \sigma^2\mathbf{Z}_r^*\mathbf{T}^*\mathbf{Z}_r^{*'} = \sum_{\alpha}\sum_{\alpha'}\mathbf{a}_\alpha\mathbf{T}^*_{\alpha,\alpha'}\mathbf{a}_{\alpha'}, \tag{10}$$

$$\begin{aligned}\mathbf{G}_2(\boldsymbol{\omega}) =&\sigma^2\left(\mathbf{X}_r^* - \mathbf{Z}_r^*\mathbf{T}^*\mathbf{Z}_s^{'}\mathbf{W}_s^{-1}\mathbf{X}_s\right)\left(\mathbf{B}_{s,11} - \tilde{\mathbf{B}}_{s,12}\right)^{-1}\\ &\times\left(\mathbf{X}_r^{*'} - \mathbf{X}_s^{'}\mathbf{W}_s^{-1}\mathbf{Z}_s\mathbf{T}^*\mathbf{Z}_r^{*'}\right),\end{aligned} \tag{11}$$

$$\mathbf{G}_4(\boldsymbol{\omega}) = \sigma^2\mathbf{L}_r\mathbf{W}_r^{-1}\mathbf{L}_r^{'} = \sigma^2(\mathrm{diag}_d\{\mathrm{diag}_t\{\mathbf{W}_{r,dt}\}\}), \tag{12}$$

being

$$\mathbf{X}_r^* = \mathbf{L}_r\mathbf{X}_r = \mathrm{col}_d\left\{\mathrm{diag}_t\left\{N_{r,dt}\bar{\mathbf{x}}_{r,dt}^{'}\right\}\right\},$$

when case A is considered, while the internal operator $\mathrm{diag}_t\{\cdot\}$ is substituted by $\mathrm{col}_t\{\cdot\}$ under case B. In addition,

$$\mathbf{Z}_r^* = \mathbf{L}_r\mathbf{Z}_r = [\mathrm{row}_d\{\mathrm{diag}_t\{N_{r,dt}\}\}, \mathrm{diag}_d\{\mathrm{row}_t\{N_{r,dt}\}\}], \tag{13}$$

in which $\mathbf{a}_1 = \mathrm{col}_d\{\mathrm{diag}_t\{N_{r,dt}\}\}$, $\mathbf{a}_2 = \mathrm{diag}_d\{\mathrm{col}_t\{N_{r,dt}\}\}$, $\mathbf{a}_3 = \mathbf{a}_1^{'} = \mathrm{row}_d\{\mathrm{diag}_t\{N_{r,dt}\}$, $\mathbf{a}_4 = \mathbf{a}_\mathbf{2}^{'} = \mathrm{diag}_d\{\mathrm{row}_t\{N_{r,dt}\}\}$.

Hence, the BLUP estimator, $\tilde{\boldsymbol{\theta}}$, given in Results 1 and 2, depends on the variance components vector $\boldsymbol{\omega}$, which is unknown in practical applications. Replacing $\boldsymbol{\omega}$ by an estimator, $\hat{\boldsymbol{\omega}}$, the correspondent EBLUP is obtained.

Result 4. The EBLUP $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\omega}})$ of $\boldsymbol{\theta}$ corresponding to (4) is given by

$$\hat{\boldsymbol{\theta}} = \mathrm{col}_d\{\mathrm{col}_t\{\hat{\bar{y}}_{dt}\}\}, \tag{14}$$

where $\hat{\bar{y}}_{dt}$ is the EBLUP of $\bar{y}_{dt}$. The explicit expression of $\hat{\bar{y}}_{dt} = \hat{\bar{y}}_{dt}(\hat{\boldsymbol{\omega}})$ is obtained by substituting the estimate $\hat{\boldsymbol{\omega}}$ of the variance component vector $\boldsymbol{\omega}$ into (6) and (5), namely $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$, $\hat{\mathbf{u}}_1 = \hat{\mathbf{u}}_1(\hat{\boldsymbol{\omega}})$ and $\hat{\mathbf{u}}_2 = \hat{\mathbf{u}}_2(\hat{\boldsymbol{\omega}})$.

REML estimates of model parameters are obtained following the iterative algorithm given in Saei and Chambers (2003). Compact expressions for updating the variance components from iteration k to iteration $k+1$ are:

$$\hat{\sigma}^2 = (n - Q)^{-1}\left(\sum_d \sum_t \sum_i w_{dti} y_{dti}\left(y_{dti} - \mathbf{x}_{dti}^{'}\hat{\boldsymbol{\beta}}\right) + \hat{\mathbf{u}}_1 \mathbf{1}_D + \hat{\mathbf{u}}_2 \mathbf{1}_T\right),$$

$$\hat{\varphi}_1 = \frac{1}{T}\left(\mathrm{tr}\left\{\hat{\mathbf{T}}_{s,11} + \hat{\mathbf{P}}_1\left(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21}\right)^{-1}\hat{\mathbf{P}}_1^{'}\right\} + \hat{\sigma}^{-2}\hat{\mathbf{u}}_1^{'}\boldsymbol{\Omega}_1^{-1}\hat{\mathbf{u}}_1\right),$$

$$\hat{\varphi}_2 = \frac{1}{D}\left(\mathrm{tr}\left\{\hat{\mathbf{T}}_{s,22} + \hat{\mathbf{P}}_2\left(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21}\right)^{-1}\hat{\mathbf{P}}_2^{'}\right\} + \hat{\sigma}^{-2}\hat{\mathbf{u}}_2^{'}\boldsymbol{\Omega}_2^{-1}\hat{\mathbf{u}}_2\right),$$

where Q denotes the number of columns of $\mathbf{X}$, $\hat{\mathbf{T}}_s = \hat{\mathbf{T}}^* + \hat{\mathbf{P}}(\hat{\mathbf{B}}_{11} - \hat{\mathbf{B}}_{21})^{-1}\hat{\mathbf{P}}'$, with $\hat{\mathbf{T}}^* = \mathbf{T}^*(\hat{\omega})$ and

$$\hat{\mathbf{P}} = \begin{bmatrix} \hat{\mathbf{P}}_1 \\ \hat{\mathbf{P}}_2 \end{bmatrix} = \begin{bmatrix} \mathrm{matr}_{dt}\{N_{dt}\bar{\mathbf{x}}_{dt}^{'}\}\hat{T}_{11}^{*} + \mathrm{diag}_t\{N_t\bar{\mathbf{x}}_t^{'}\}\hat{T}_{12}^{*} \\ \mathrm{matr}_{dt}\{N_{dt}\bar{\mathbf{x}}_{dt}^{'}\}\hat{T}_{21}^{*} + \mathrm{diag}_t\{N_t\bar{\mathbf{x}}_t^{'}\}\hat{T}_{22}^{*} \end{bmatrix},$$

$$\hat{\rho}_1(k+1) = \hat{\rho}_1(k) + I(\hat{\rho}_1) + \Delta(l_{\mathrm{REML}}(\hat{\rho}_1)), \tag{15}$$

$$\hat{\rho}_2(k+1) = \hat{\rho}_2(k) + I(\hat{\rho}_2) + \Delta(l_{\mathrm{REML}})(\hat{\rho}_2)), \tag{16}$$

where $\Delta(l_{\mathrm{REML}})(\cdot)$ is the derivative of the likelihood function with respect to the parameter of interest, whereas $I(\cdot)$ is the relevant element of the inverse of the information matrix. The expressions given above are updated iteratively together with the expression (6) for $\tilde{\boldsymbol{\beta}}$ given in Result 2 until convergence is attained.

Result 5. The MCPE of the EBLUP $\hat{\boldsymbol{\theta}}$ is given by the diagonal elements of the following matrix

$$\mathrm{MCPE}(\hat{\boldsymbol{\theta}}) = \mathrm{MCPE}(\tilde{\boldsymbol{\theta}}) + 2\mathbf{G}_3(\hat{\boldsymbol{\omega}}) = \mathbf{G}_1(\hat{\boldsymbol{\omega}}) + \mathbf{G}_2(\hat{\boldsymbol{\omega}}) + 2\mathbf{G}_3(\hat{\boldsymbol{\omega}}) + \mathbf{G}_4(\hat{\boldsymbol{\omega}}),$$

where $\mathbf{G}_1(\hat{\boldsymbol{\omega}})$, $\mathbf{G}_2(\hat{\boldsymbol{\omega}})$, $\mathbf{G}_4(\hat{\boldsymbol{\omega}})$ are computed, respectively, plugging into (9), (10), (11), and (12) the estimated values of the variance components. Matrix $\mathbf{G}_3(\hat{\boldsymbol{\omega}})$ takes into account the uncertainty of the estimation of the variance components. The explicit expression of $\mathbf{G}_3(\hat{\boldsymbol{\omega}})$ is

$$\mathbf{G}_3(\hat{\boldsymbol{\omega}}) = \hat{\sigma}^2\left[tr\left(\nabla_\alpha \hat{\boldsymbol{\Sigma}}_s^{*} \nabla_{\alpha'}^{'} \hat{\mathbf{B}}\right)\right],$$

where $\hat{\mathbf{B}}$ is the asymptotic covariance matrix of the REML estimates of the variance component vector $\boldsymbol{\omega}$. It depends on the diagonal elements of the inverse of the Fisher information matrix of REML estimators $\hat{\boldsymbol{\omega}}$. For more details, see Saei and Chambers (2003). Furthermore, ∇_α and $\boldsymbol{\Sigma}_s^*$ have the following expression

$$\nabla_\alpha = -\left(\mathbf{Z}_\alpha^* \hat{\mathbf{T}}^* \otimes \mathbf{I}_H\right)\left(\frac{\delta \boldsymbol{\Omega}^{-1}}{\delta \boldsymbol{\omega}}\right)_{\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}} \hat{\mathbf{T}}^*,$$

$$\boldsymbol{\Sigma}_s^* = \mathbf{A} + \mathbf{A}\boldsymbol{\Omega}\boldsymbol{\Omega},$$

where, denoting with $\otimes$ the Kronecker product, $\mathbf{Z}^*_\alpha$ is the αth row of matrix $\mathbf{Z}^*_r$ given in (13), $\mathbf{I}_H$ is the identity matrix of dimension $H_{=4}$, and $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \text{diag}_d\{w_{s,d}\} & \text{matr}_{dt}\{w_{s,dt}\} \\ \text{matr}_{td}\{w_{s,dt}\} & \text{diag}_t\{w_{s,t}\} \end{bmatrix},$$

for $w_{s,d} = \sum_t w_{s,dt}$, $w_{s,t} = \sum_d w_{s,dt}$ and $w_{s,dt} = \sum_i^{n_{dt}} w_{dti}$.

4. Particular Cases

Starting from the general LMM specification in Saei and Chambers (2003), we describe the more relevant model, and two random effect model specifications presented in the literature. The general model (1) will be denoted by $\text{M}^{\text{ST}}_{\text{CC}}$, where the superscript ST stands for model with spatial and temporal random effects, and subscript CC stands for using a correlation structure for both the random effects.

When N_d is large, $f_{dt} \cong 0$ and $\bar{\mathbf{x}}_{r,dt} \cong \bar{\mathbf{x}}_{dt}$, and the general formula (4) of the unit-level EBLUP with space and time correlation, $\hat{\bar{y}}_{dt}$, can be approximated by

$$\text{M}^{ST}_{CC} : \hat{\bar{y}}_{dt} = \bar{\mathbf{x}}'_{dt}\hat{\boldsymbol{\beta}} + \sum_{d'}\sum_{t'} \hat{\gamma}_{d't'}\hat{\bar{e}}_{w,d't'}, \tag{17}$$

where $\hat{\gamma}_{d't'} = w_{d't'}\hat{\Gamma}_{dt}(d',t')$, being $\hat{\Gamma}_{dt}(d',t') = \hat{T}^*_{11,dd'} + \hat{T}^*_{12,dt'} + \hat{T}^*_{21,td'} + \hat{T}^*_{22,tt'}$. The corresponding estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by means of (14).

Special cases of (17) are obtained through particular settings for $\hat{\Gamma}_{dt}(d',t')$. Using analogous notation, $\text{M}^{\text{ST}}_{\text{II}}$ is the two-way model with independent and identically distributed area and time effects, while $\text{M}^{\text{ST}}_{\text{IC}}$ and $\text{M}^{\text{ST}}_{\text{CI}}$ denote, respectively, the two-way linear mixed model with independent area effects and correlated time effects, and spatially correlated area effects and independent time effects.

The case of two independent random effects, $\text{M}^{\text{ST}}_{\text{II}}$, is obtained when $\hat{\Gamma}_{dt}(d',t') = \hat{T}^*_{11,dd} + \hat{T}^*_{22,tt}$. Therefore, the expression for the estimator is given by

$$\text{M}^{\text{ST}}_{\text{II}} : \hat{\bar{y}}_{dt} = \bar{\mathbf{x}}'_{dt}\hat{\boldsymbol{\beta}} + \hat{\gamma}_d\hat{\bar{e}}_{w,d} + \hat{\gamma}_t\hat{\bar{e}}_{w,t},$$

where $\hat{\gamma}_d = w_d\hat{T}^*_{11,dd} = \hat{\sigma}^2_1/\left(\hat{\sigma}^2_1 + \hat{\sigma}^2/w_d\right)$ and $\hat{\gamma}_t = w_t\hat{T}^*_{22,tt} = \hat{\sigma}^2_2/\left(\hat{\sigma}^2_2 + \hat{\sigma}^2/w_t\right)$. This estimator may be applied in many real situations, for example, when the spatial and temporal correlation between area and time effects is lower than a given threshold. Furthermore, it may be useful for cross-sectional surveys in which index t, instead of representing time, represents a set of T domains which form a different partition of the population than the D areas.

In many practical situations, it may be useful to consider the two estimators $\text{M}^{\text{ST}}_{\text{CI}}$ and $\text{M}^{\text{ST}}_{\text{IC}}$. Model $\text{M}^{\text{ST}}_{\text{IC}}$ can be used for repeated business surveys, in which the small areas of interest are small domains different from territorial subpopulations (e.g., industry segments) and it is not possible, or straightforward, to define spatial correlation among domains.

One-way models $\text{M}^{\text{S}}_{\text{C}}$ and $\text{M}^{\text{S}}_{\text{I}}$, respectively, with spatially correlated area effects and independent and identically distributed area effects, allow traditional cross-sectional small area estimation to borrow strength from other domains, but not from other survey cycles. Specifically, M_I^S corresponds to the standard model defined by Battese et al. (1988), while examples for M_C^S are given in Saei and Chambers (2003), and Petrucci and Salvati (2004). To borrow strength from other survey cycles but not from other domains, alternative modelisations for usual time series models are M_C^T and M_I^T, that is, linear mixed models with correlated time effects and independent and identically distributed time effects.

5. Application to Real Data

In this section we present a case study aimed at comparing several alternative SAE models and at testing different SAE software estimation tools. To this end, LFS data from 2004 to 2014 has been used to produce estimates of the unemployment rate at LMA level. The overall amount of data is about 7,200,000 records and about 25% of LMAs are not covered by the samples.

LMMs with both area and time random effects are considered, and their estimation is made possible by means of the expressions described in Section 3. The corresponding estimator has been applied to compute quarterly LMA unemployment rates and compared with other standard SAE methods.

The binary nature of the target variable should suggest the use of non-normal mixed models, for instance a binomial with a logistic link function. However, D'Aló et al. (2012) showed that the use of logistic models does not improve substantially the quality of the estimates with respect to normal model. Furthermore, Boonstra et al. (2007) do not find evidence for the superiority of logistic mixed models over their normal counterparts in the estimation of unemployment counts in Dutch municipalities. In addition, we are not usually interested in individual predictions, but rather in predicting area and time aggregates. Besides, for non-normal mixed models, easy interpretable closed-form expressions for predictors are not available. Linear mixed models only need area and time population totals for prediction, while non-normal models require cross-classified population totals for the fixed effects, even though only marginal effects are included in the model specification.

The LMMs and the correspondent estimators considered in the experimental study are reported in Table 1.

In addition to the direct estimator, EBLUPs arising from one-way and two-way unit-level LMMs are considered. Therefore, the estimator with area- and time-correlated

Table 1. List of models and estimators considered.

Model	Estimator
–	Direct
M_I^S	$EBLUP_I^S$
M_C^S	$EBLUP_C^S$
M_{CC}^{ST}	$EBLUP_{CC}^{ST}$
$M_I^{S(**)}$	$EBLUP\_ALL_I^S$
$M_C^{S(**)}$	$EBLUP\_ALL_C^S$

(**)Model parameters are estimated using all LFS data from 2004 to 2014.

random effects, $\mathrm{EBLUP}_{\mathrm{CC}}^{\mathrm{ST}}$, is compared with two SAE cross-sectional methods, specifically with the EBLUP with uncorrelated area random effects, $\mathrm{EBLUP}_{\mathrm{I}}^{\mathrm{S}}$, and with spatially correlated area random effects, $\mathrm{EBLUP}_{\mathrm{C}}^{\mathrm{S}}$. Furthermore, $\mathrm{EBLUP}_{\mathrm{CC}}^{\mathrm{ST}}$ is computed using the whole set of available time series data, while the cross-sectional methods exploit only the last quarter data set. Then, in order to be able to set aside the effect of the amount of data, when comparing $\mathrm{EBLUP}_{\mathrm{CC}}^{\mathrm{ST}}$ with its competitors, the one-way model parameters have also been estimated using the overall set of data. These last two estimators are denoted by $\mathrm{EBLUP\_ALL}_{\mathrm{I}}^{\mathrm{S}}$ and $\mathrm{EBLUP\_ALL}_{\mathrm{C}}^{\mathrm{S}}$, respectively.

In particular, for $\mathrm{EBLUP}_{\mathrm{CC}}^{\mathrm{ST}}$, the between-area correlation matrix proposed by Saei and Chambers (2003) has been considered. This matrix is dependent on the distances among the areas and on a scale parameter ρ_1 connected to the spatial structure of the areas, and is given by

$$\mathbf{\Omega}_1(\rho_1) = \left[1 + \delta_{d,d'}\exp\left(\frac{\mathrm{dist}(d,d')}{\rho_1}\right)\right]^{-1},$$

with $\delta_{d,d'} = 0$ if $d = d'$ and $\delta_{d,d'} = 1$ otherwise and $dist(d,d')$ denoting the Euclidean distance between area d and d'. Instead, the between-time correlation matrix arises from an autoregressive AR(1) process whose expression is

$$\mathbf{\Omega}_2(\rho_2) = \frac{1}{1-\rho_2^2}\begin{bmatrix} 1 & \rho_2 & \cdots & \rho_2^{T-1}\rho_2 \\ 1 & \cdots & \rho_2^{T-2}\vdots & \vdots \\ \ddots & \vdots\rho_2^{T-1} & \rho_2^{T-2} & \cdots \\ 1 & & & \end{bmatrix}.$$

The scope of the empirical study is to assess the statistical properties of the estimators. To this aim, the estimates computed for the last quarter of 2011 (October 2011–December 2011) are compared with the correspondent 2011 Census values, which are referred to on 9 October 2011.

The auxiliary information used in the experimental study, that is, the cross-classification of 14 age groups by sex, is similar to what is used in the LFS calibration process. A common regression coefficient vector is defined for all the quarters. This is the hypothesis defined in Section 2 as Case B. We note that the assumption of fixed effects over time is not very realistic, but the correlated random effects are expected to smooth the estimates. A first comparison among the estimators has been carried out by means of Average Absolute Relative Error (AARE) and Average Squared Error (ASE), defined as

$$\mathrm{AARE}(\hat{\boldsymbol{\theta}}) = \frac{1}{D}\sum_{d=1}^{D}\mathrm{ARE}_d = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{\hat{\theta}_d}{\theta_d} - 1\right|,$$

$$\mathrm{ASE}(\hat{\boldsymbol{\theta}}) = \frac{1}{D}\sum_{d=1}^{D}\mathrm{SE}_d = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d - \theta_d\right)^2,$$

where for domain d, $d = 1, \ldots, D$, $\hat{\theta}_d$ and θ_d are, respectively, the estimate computed with a given estimator and the true parameter of interest.

Table 2. AARE and ASE with respect to 2011 Census data.

Estimator	AARE	ASE$^{(*)}$
Direct	0.65	18.86
$\text{EBLUP}_{\text{I}}^{\text{S}}$	0.34	2.67
$\text{EBLUP}_{\text{C}}^{\text{S}}$	0.33	2.59
$\text{EBLUP}_{\text{CC}}^{\text{ST}}$	0.26	2.07
$\text{EBLUP\_ALL}_{\text{I}}^{\text{S}(**)}$	0.36	3.56
$\text{EBLUP\_ALL}_{\text{C}}^{\text{S}(**)}$	0.36	3.56

(*)ASE is multiplied by 1,000.
(**)Model parameters are estimated using all LFS data from 2004 to 2014.

Table 2 displays the values of AARE and ASE evaluated over the 611 LMAs. The $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ outperforms the others estimators both in terms of AARE and ASE. It shows better performances than $\text{EBLUP}_{\text{I}}^{\text{S}}$ and $\text{EBLUP}_{\text{C}}^{\text{S}}$. $\text{EBLUP}_{\text{I}}^{\text{S}}$ and $\text{EBLUP}_{\text{C}}^{\text{S}}$ performed similarly, with a slight preference for the $\text{EBLUP}_{\text{C}}^{\text{S}}$. This implies there is no strong evidence for a significant spatial correlation. Therefore, the introduction of the time random effect substantially increases the efficiency of the estimates. In fact, the estimated value of the time correlation coefficient ρ_2, computed with (16), is equal to 0.73, while the estimate of the spatial parameter, obtained by means of (15), is 0.29. The spatial correlation defined for the area random effects allows us to obtain more accurate estimates for out-of-sample areas than the corresponding estimates computed only by synthetic prediction. Furthermore, the better performance of $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ is not only due to the larger set of data involved in the estimation process. In fact, $\text{EBLUP\_ALL}_{\text{I}}^{\text{S}}$ and $\text{EBLUP}_{\text{C}}^{\text{S}}$, which use the same data as $\text{EBLUP}_{\text{CC}}^{\text{ST}}$, perform poorly because they do not capture the true time pattern of data.

Table 3 reports the value of the coefficients of variation for all the estimates, with the exception of $\text{EBLUP\_ALL}_{\text{I}}^{\text{S}}$ and $\text{EBLUP\_ALL}_{\text{C}}^{\text{S}}$. It shows that $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ outperforms the other methods, aside from minimum and maximum values. The direct estimator shows a better coefficient of variation value only for the minimum value.

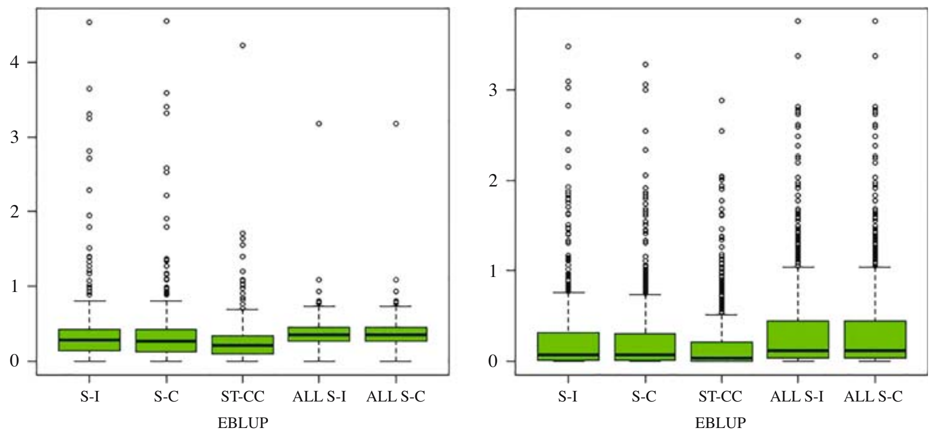
Figures 1a and 1b show the distribution of ARE and SE, respectively. The error distribution of the direct estimator is not included due to its poor performance. In accordance with Table 2, in both cases the distribution of the errors for $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ is more concentrated around zero than the other distributions, with the exception of $\text{EBLUP\_ALL}_{\text{I}}^{\text{S}}$ and $\text{EBLUP\_ALL}_{\text{C}}^{\text{S}}$ for the ARE.

Figure 2 displays the spatial distribution of the estimates for direct estimator (a), $\text{EBLUP}_{\text{I}}^{\text{S}}$ (b), $\text{EBLUP}_{\text{C}}^{\text{S}}$ (c) and $\text{EBLUP}_{\text{CC}}^{\text{ST}}$ (d). The direct estimates are plotted for

Table 3. CV% distribution.

Estimator	Min.	1st Q	Median	Mean	3rd Q	Max.
Direct$^{(*)}$	0.72	31.57	52.01	54.56	77.34	119.80
$\text{EBLUP}_{\text{I}}^{\text{S}}$	4.83	19.83	26.99	26.42	33.95	45.46
$\text{EBLUP}_{\text{C}}^{\text{S}}$	4.83	19.71	27.18	26.35	33.79	45.60
$\text{EBLUP}_{\text{CC}}^{\text{ST}}$	1.19	4.23	6.16	7.80	9.44	27.97

(*)There are 158 empty LMAs.



(a) *ARE*. (b) *SE*.

Fig. 1. ARE and SE distributions. SE is multiplied by 1,000.

provinces, while the estimates for the EBLUPs are plotted on LMAs. This is because the 110 provinces are planned domains for which the direct estimator produce reliable estimates. The spatial distribution of the direct estimates can be considered as a good picture of the spatial distribution of the true unemployment rates, and for that it can be used to benchmark SAE estimates. As showed in Figure 1, all the EBLUPs have analogous spatial patterns to the distribution of the direct estimates.

6. SAE Software for Unit-Level Linear Mixed Models

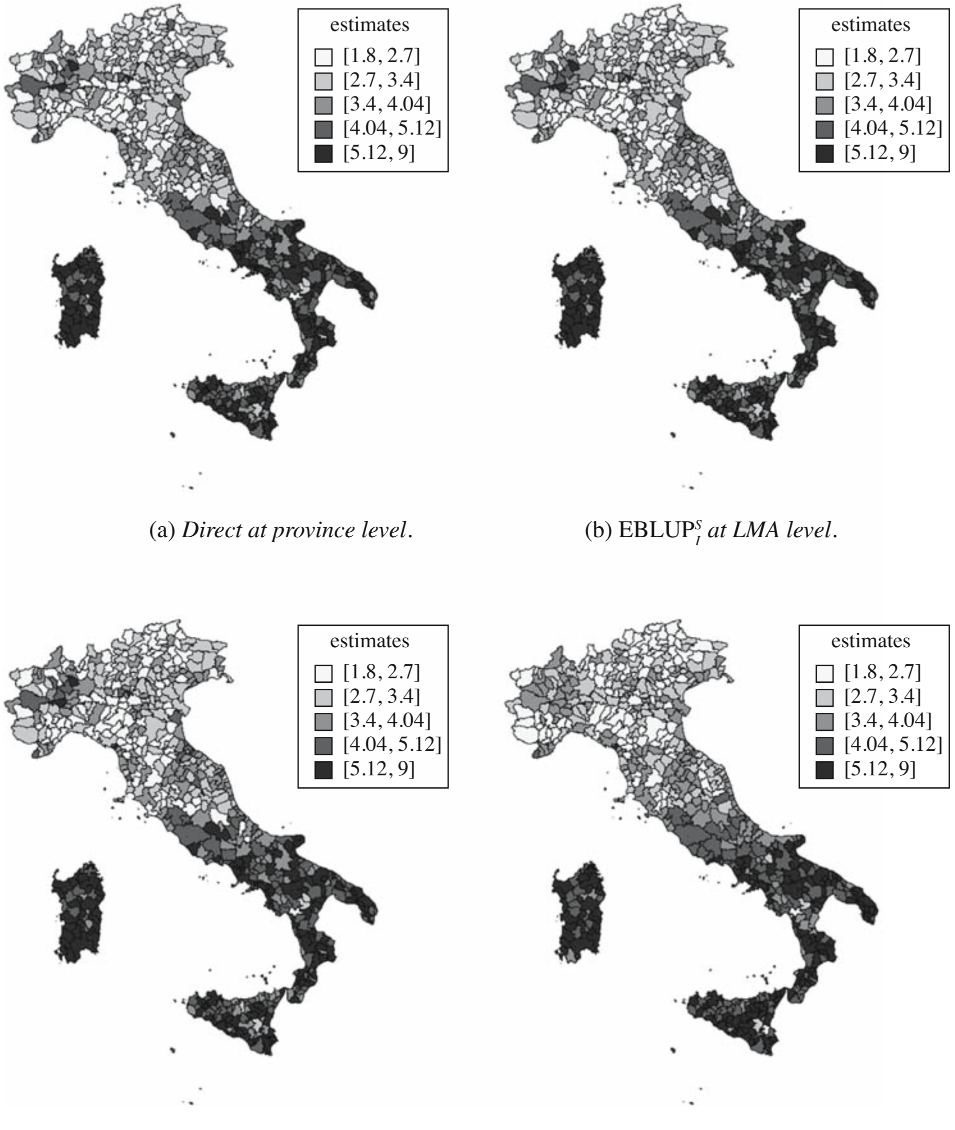
We implemented the new formulation given in Section 3 in an R function named space.time.eblup, which allows the computation of (a) estimates of the model parameters; (b) SAE estimates and their MSEs for sampled areas; (c) SAE estimates and their MSEs also for out-of-sample areas. In this section, the performance of this function is compared with the most used software tools, available for SAE or for LMMs fitting. An exhaustive review of available SAE software tools is provided by the Essnet SAE project.

The available SAE software packages carry out a complete estimation process with the computation of (a) and (b), but, usually, do not allow (c). LMMs can be estimated using general software tools for model fitting. In this case, they allow only (a), and extra work is needed to complete the estimation process, that is, (b) and (c).

The result of the comparative analysis of space.time.eblup compared with the other available functions and SAE packages shows evidence that space.time.eblup, in addition to performing a more complete estimation process, is more efficient in terms of runtime.

Table 4 reports SAE software tools developed recently by national or international projects dealing with small area estimation. All SAE software provides a complete tool for treating SAE problems, but only the R functions produced by SAMPLE are able to deal with LMMs that include area and time random effects. Specifically, time random effects are nested within area random effects instead of including additive random effects as in (1). Furthermore, no correlation structure can be specified for the area random effects.

Besides the software tools described in Table 4, R packages specifically dedicated to SAE are available for download at the CRAN, https://cran.r-project.org/. The SAE



(a) *Direct at province level.*

(b) EBLUP^S_I *at LMA level.*

(c) EBLUP^S_C *at LMA level.*

(d) EBLUP$^{ST}_{CC}$ *at LMA level.*

Fig. 2. Unemployment rate estimates for direct (a), EBLUP^S_I (b), EBLUP^S_C (c), EBLUP$^{ST}_{CC}$ (d). Legends display the estimated unemployment rate classes.

packages implementing unit-level LMMs are hbsae, JoSAE, rsae, sae, but do not include time random effects in the model. Furthermore, as far as the software toos describe in Table 4 are concerned, it is worthwhile to underline that they can only handle sets of data much smaller than the 7,200,000 records processed for the case study.

Besides SAE packages, there are many R packages that provide functions for fitting LMMs. A general package for LMMs is lme4. It can fit linear mixed models by means of the function lmer. These models can also be fitted using the function lme from the package

Table 4. Description of SAE software based on unit-level LMMs produced by projects on small area estimation.

Project	Enviroment	Area random effects	Time random effects
EURAREA	SAS	Correlated	No
BIAS	R	Uncorrelated	No
SAMPLE	R	Uncorrelated	Nested, Correlated
AMELI	R	Uncorrelated	No
ESSnet SAE	R	Correlated	No

nlme. This package supports various correlation and heteroscedasticity structures for the variance within.

Concerning the statistical software SAS, (see https://www.sas.com/), apart from the macro program codes developed by the EURAREA project, no ad hoc SAE software is available. The SAS procedure MIXED can fit a variety of LMMs. It performs model estimation that provides both fixed and random effects estimates, and variance components estimates.

Table 5 compares, in terms of computation times, the performances of the most used SAS and R functions to fit LMMs with the space.time.eblup function. Only the procedure MIXED in SAS allows us to handle the whole set of data used in the case study of the Italian LFS. However, when spatial and temporal correlation is introduced, it can only process much smaller data sets. The R functions tested to fit LMMs were not able to process the whole set of data, but only a subset including about 3,000,000 records related to the first 18 survey occasions. Furthermore, similarly to the SAS procedure MIXED, lme and lmer can fit models with correlated random effects only for very small sets of data. For this reason, the only comparison framework that can be set up is restricted to the 18 survey occasions sets of data, and without taking into account any type of correlation structure. Moreover, the space.time.eblup function is a complete SAE tool providing computation of (a), (b), and (c).

All the performances of R and SAS codes were run on an Intel CoreTM i7-3770K 3.50 GHz processor with 8 GB RAM on a 64 bit Windows 7 personal computer.

7. Conclusions

Since the most important surveys carried by national statistical institutes are repeated surveys, it is important to carefully consider SAE problems within this broad and relevant survey framework. Standard small area models usually take into account cross-sectional

Table 5. Comparison of performances, in terms of computer time, for R and SAS functions fitting unit-level LMMs and space.time.eblup R function for Italian LFS data, complete and reduced.

Package	Complete data set	Restricted data set
PROC MIXED$^{(*)}$	30 sec	12 sec
lme$^{(*)}$	–	3 min 00 sec
lmer$^{(*)}$	–	2 min 21 sec
space.time.eblup	4 min 54 sec	2 min 18 sec

$^{(*)}$Elaboration times are related to independent area and time random effects.

estimation. Nonetheless, in the context of repeated surveys, more realistic and efficient models can be considered by adding a temporal random effect for exploiting previous survey occasions data. It potentially allows us to increase the efficiency of results by using more realistic SAE models that can better capture the real variability of the phenomena under study. Furthermore, unit-level models have potentially more predictive power than area-level models, and they are able to exploit the individual correlations between target variable and fixed effects covariates.

As a consequence, large amount of data have to be processed and computational problems may occur. The empirical test, conducted on Italian LFS quarterly data, displayed good statistical performance, outperforming the other estimators. Furthermore, the new formulation was shown to be effective when dealing with extremely large amounts of data. As a matter of fact, the function space.time.eblup, implementing the new expressions was able to process 7,200,000 survey records from the 44 LFS quarterly samples from 2004 to 2014 in about five minutes. Therefore, the new formulation allows us to manage very large amounts of data, overcoming the computational limits underlying the software currently available. Moreover, it can provide a valuable starting point for building more sophisticated models.

Currently, only the R function is available for use. However, an R package will be produced and made available as soon as possible.

8. References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. “An Error Components Model for Prediction of County Crop Areas Using Survey and Satellite Data.” *Journal of American Statistical Association* 83: 28–36. Doi: http://dx.doi.org/10.1080/01621459.1988.10478561.

Boonstra, H., B. Buelens, and M. Smeets. 2007. “Estimation of Municipal Unemployment Fractions - A Simulation Study Comparing Different Small Area Estimators.” Internal report, BPA-no. DMK-DMH-2007-04-20-HBTA, Herleen: Statistics Netherlands.

Cressie, N. 1992. “REML Estimation in Empirical Bayes Smoothing of Census Undercount.” *Survey Methodology* 18: 75–94.

Cressie, N. 1993. *Statistics for Spatial Data*. New York: Wiley.

D’Aló, M., L. Di Consiglio, S. Falorsi, M.G. Ranalli, and F. Solari. 2012. “Use of Spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-Provincial Areas in Italy.” *Journal of the Indian Society of Agricultural Statistics* 66: 43–54.

Datta, G.S. P. Lahiri, and T. Maiti. 2002. “Empirical Bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-sectional Data.” *Journal of Statistical Planning and Inference* 102: 83–97. Doi: http://dx.doi.org/10.1016/S0378-3758(01)00173-2.

Duncan, G.J. and G. Kalton. 1987. “Issues of Design and Analysis of Surveys across Time.” *International Statistical Review* 55: 97–117. Doi: http://dx.doi.org/10.2307/1403273.

Fay, R. and R. Herriot. 1979. “Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data.” *Journal of the American Statistical Association* 74: 269–277. Doi: http://dx.doi.org/10.1080/01621459.1979.10482505.

Gershunskaya, J. 2015. “Combining Time Series and Cross-Sectional Data for the Current Employment Statistics Estimates.” In Proceedings of the Section on Statistical Computing: American Statistical Association, August 9, 2015. 1085–1096. Alexandria, VAL: American Statistical Association. Available at: http://www.amstat.org/sections/srms/proceedings/y2015/files/233962.pdf (January 16, 2017).

Harville, D.A. 1977. “Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems.” *Journal of the American Statistical Association* 72: 320–338. Doi: http://dx.doi.org/10.2307/2286797.

Hidiroglou, M.A. and Y. You. 2016. “Comparison of Unit Level and Area Level Small Area Estimators.” *Survey Methodology* 42: 41–61. Available at: http://www.statcan.gc.ca/pub/12-001-x/2016001/article/14540-eng.pdf (January 16, 2017).

Kish, L. 1987. *Statistical Designs for Research*. New York: Wiley.

Petrucci, A. and N. Salvati. 2004. “Small Area Estimation Considering Spatially Correlated Errors: the Unit Level Rrandom Effects Model.” Working Paper 2004/10, Department of Statistics, Florence University.

Pfeffermann, D. 2002. “Small Area Estimation: New Developments and Directions.” *International Statistical Review* 70: 125–143. Doi: http://dx.doi.org/10.2307/1403729.

Pfeffermann, D. 2013. “New Important Developments in Small Area Estimation.” *Statistical Science* 28: 40–68. Doi: http://dx.doi.org/10.1214/12-sts395.

Pfeffermann, D. and L. Burck. 1990. “Robust Small Area Estimation Combining Time Series and Cross-Sectional Data.” *Survey Methodology* 16: 217–237.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: Wiley.

Rao, J.N.K. and M. Yu. 1994. “Small Area Estimation by Combining Time Series and Cross-Sectional Data.” *Canadian Journal of Statistics* 22: 511–528. Doi: http://dx.doi.org/10.2307/3315407.

Royall, R.M. 1976. “The Linear Least-Squares Prediction Approach to Two-Stage Sampling.” *Journal of the American Statistical Association* 71: 657–664. Doi: http://dx.doi.org/10.1080/01621459.1976.10481542.

Saei, A. and R. Chambers. 2003. “Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects”. *Methodology Working Paper M03/15*. University of Southampton: Southampton Statistical Sciences Research Institute. Available at: http://eprints.soton.ac.uk/8165/1/8165-01.pdf (January 16, 2017).

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. *Variance Components*. New York: Wiley.

You, Y. 1999. “Hierarchical Bayes and Related Methods for Model Based Small Area Estimation.” Ph.D. Thesis, School of Mathematics and Statistics, Carleton University.

Received September 2015
Revised September 2016
Accepted October 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 79–100, http://dx.doi.org/10.1515/JOS-2017-0005

Official Statistics and Statistics Education: Bridging the Gap

Iddo Gal[1] *and Irena Ograjenšek*[2]

This article aims to challenge official statistics providers and statistics educators to ponder on how to help non-specialist adult users of statistics develop those aspects of statistical literacy that pertain to official statistics. We first document the gap in the literature in terms of the conceptual basis and educational materials needed for such an undertaking. We then review skills and competencies that may help adults to make sense of statistical information in areas of importance to society. Based on this review, we identify six elements related to official statistics about which non-specialist adult users should possess knowledge in order to be considered literate in official statistics: (1) the system of official statistics and its work principles; (2) the nature of statistics about society; (3) indicators; (4) statistical techniques and big ideas; (5) research methods and data sources; and (6) awareness and skills for citizens' access to statistical reports. Based on this ad hoc typology, we discuss directions that official statistics providers, in cooperation with statistics educators, could take in order to (1) advance the conceptualization of skills needed to understand official statistics, and (2) expand educational activities and services, specifically by developing a collaborative digital textbook and a modular online course, to improve public capacity for understanding of official statistics.

Key words: Statistical literacy; skills and competencies; official statistics literacy; dissemination; adult education.

1. Background and Motivation

In recent years, both national and international statistical offices as well as other producers of official statistics (hereafter: official statistics providers) have been paying increasing attention to the formal training of professional statisticians who work in national and international statistical systems, and sometimes to the training of other user groups. Programs awarding either a diploma or a full academic degree related to official statistics are offered by several intergovernmental institutions or networks, such as the European Master in Official Statistics (EMOS; Zwick 2016), the Statistical Institute for Asia and the Pacific (SIAP), and the University of the South Pacific. Several national statistical offices (some via institutional collaboration) are very active in this regard as well. For example, in New Zealand, a postgraduate course in official statistics is offered that covers areas such as data visualization, confidentiality, geographic information system, demography, health

[1] University of Haifa, Department of Human Services, Abba Houshi Road 199, Haifa, 3498838, Israel. Email: iddo@research.haifa.ac.il
[2] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, Ljubljana, 1000, Slovenia. Email: irena.ograjensek@ef.uni-lj.si (corresponding author)

Acknowledgments: The first author was partially supported by a grant from the ERASMUS+ program of the European Commission for the ProCivicStat project. However, the opinions expressed in this article reflect the author's own views and not necessarily those of the sponsoring agency.

statistics, and economic statistics (Harraway and Forbes 2013), in collaboration with New Zealand's National Certificate of Official Statistics (Forbes and Keegan 2016). The Central Statistics Office Ireland (MacCuirc 2015) or Statistics Finland (Helenius and Mikkelä 2011) have also developed training modules, or full diploma programs, for specific target groups of users such as government employees and analysts, business managers, or journalists, who are usually not statisticians but who work with official statistics in various ways.

This article focuses on a gap in the world of formal training in official statistics, pertaining to wider, non-professional audiences. These include, among other groups within the adult public at large, the many educators who may teach non-specialists about statistics (for example, lecturers in introductory statistics at the undergraduate level in many different disciplines and departments, mathematics teachers who also teach statistics at the high school level), their many students (who would soon be adults and enter the workforce), or various administrators, and managers in diverse sectors.

On the one hand, official statistics providers are interested in increasing the use of their information products through multiple user groups that include the general public. They are taking many steps to improve the quality of their information services: they have been opening up free access to their information products through digital portals, and have been continuously seeking ways to improve levels of public trust and confidence in official statistics, as well as the level of satisfaction with their information products (Biemer et al. 2014; Steenvoorden et al. 2015).

On the other hand, the provision of training or resources related to official statistics for wider, non-professional audiences, has been largely left aside. Even if official statistics websites are being made more user-friendly, comprehension of the statistical information in them is far from optimal (Schield 2011). Very few official statistics providers offer structured materials designed to enable the public, or stakeholders from the education sector (i.e., teachers and students), to better understand official statistics on their websites. Even leading national statistical offices such as Statistics Canada or the Australian Bureau of Statistics have cut down on their support to statistics education at schools over the last few years.

The gaps noted above also exist within the professional field of statistics. Official statistics providers have been operating for decades around the world, and represent an indispensable element in the information system of a democratic society (United Nations 2014). However, a dire and surprising lack of solid educational materials designed for professionals (i.e., statistics or economics majors entering careers in official statistics) has been noted by numerous scholars involved in the training of statisticians (Murphy 2002; Nathan 2007). Pfeffermann (2015) has recently reviewed curricula of statistics departments at over 20 leading universities and concluded that most departments pay little attention to formal instruction in key aspects of official statistics (such as survey sampling, seasonal adjustment, or national accounts). Given that official statistics is a prime employment area for statistics graduates, this is a very surprising finding.

Furthermore, a literature search we conducted did not find a single current textbook that describes key knowledge bases that have to be emphasized in detail when educating statistics majors about official statistics. Over 20 years ago the modular online Course on European Economic Statistics (CEES) was developed with the support of Eurostat (in cooperation with the Institute for Training of European Statisticians – TES) within the Phare Multi-Country Co-Operation for Distance Education Programme (Bregar et al. 2000). Unfortunately, it seems to have been published too early to be adopted by traditional universities, most of which at that point in time had not begun to recognize digitalization as the future of educational systems. Lacking solid marketing support at its launch, the course therefore remained a short-lived attempt to fill in the gap that was identified decades ago and still exists today.

The only book currently available that appears to be dedicated to the role of an official statistics provider is Citro and Straf (2013), which is also in use by the EMOS programme. This US-based text focuses on key aspirations or expectations from an official statistics provider (for example, relevance to policy issues, credibility among data users, trust among data providers, independence from political and external influences), and on numerous important administrative and organizational practices and roles (such as mission clarity, confidentiality, continuous development of useful data, openness about sources, data limitations transparency, and more). These are core issues for all official statistics providers around the world, yet they are not related to a comprehension of the actual products from the contents' point of view. Consequently, the text should be regarded as a very incomplete basis from which to define what non-specialists need to know to understand official statistics products.

The situation described above implies that educators who wish to introduce non-majors, high school students, business graduates or adults in general to the fundamentals of official statistics do not have a set of suitable resources geared for their needs, even at the beginning of the twenty-first century. If one accepts the tenet that citizens should know something useful about official statistics, many questions arise: first of all, the question of "what are the basics that citizens (or non-specialists) should know about official statistics?".

While this question seems simple, the answer is not straightforward. It has not been discussed in detail in the professional literature on official statistics; and certainly not with regard to non-majors and adults at large. Other related questions are "whose responsibility is it to develop materials on official statistics for non-specialists?", and "to what extent (if at all) should official statistics providers divest resources in order to increase public knowledge of official statistics?".

Our goal in this article is to assist, but also to challenge official statistics providers to ponder the questions raised above. We focus our contribution on specific issues that official statistics providers may face if they want to help non-specialist users develop the aspects of statistical literacy (Gal 2002) that pertain to knowledge of, and engagement with, official statistics (for brevity we refer to this desired knowledge base as *official statistics literacy* or OSL). To this end, in our article we first briefly review the general ecology of skills and competencies that adults may need in order to make sense of statistical information regarding societal matters. We then examine possible building blocks of the desired knowledge base that is specific to OSL in more detail. Based on this conceptualization, we then discuss some directions for future developments that official statistics providers could make in order to contribute to educational efforts aimed at increasing official statistics literacy, thereby enriching the course for the development of statistics education in general.

2. On Quantitative Competencies and Literacies

A discussion of the statistical capacity needed to understand and engage with official statistics requires that we first describe the larger environment within which understanding of official statistics (by non-specialists) is situated.

Over the last few decades, the academic and professional literature has identified several related but separate constructs that describe general competencies that adults should possess in order to effectively cope with the quantitative demands of the adult world, including those related to statistics and probability. Key constructs that have so far been defined and discussed at some length are *(adult) numeracy and mathematical literacy* (Gal et al. 2005; PIAAC Numeracy Expert Group 2009; Geiger et al. 2015; Stacey 2015; Tout and Gal 2015), *quantitative literacy and quantitative reasoning* (Steen 2001; Madison 2014; Karaali et al. 2016), as well as *statistical literacy and probability literacy* (Gal 2002, 2005; Watson 2016). Separate constructs such as *health numeracy* (Ancker and Kaufman 2007), *scientific literacy* (Rutherford 1997), *financial literacy* (Lusardi and Mitchell 2014; Xu and Zia 2012) or *media literacy* (Coddington 2015) also encompass, among other components, diverse quantitative skills which incorporate understanding of specific types of statistics and data collection methods. Examples include an understanding of long-range trends in the economy or in ageing which affect pensions or poverty levels; risk estimates associated with health conditions, pollution levels, and mortality rates; or notions of (the strength of) evidence.

The usage of 'literacy' when coupled with a term denoting an area of human activity (e.g., 'statistical literacy') may conjure an image of a minimal subset of basic skills expected of all citizens in this area, as opposed to a more advanced set of skills and knowledge that only specialists may achieve. Yet, many scholars warn against such a restrictive interpretation, and argue that "literacy", when used to describe people's capacity for goal-oriented behavior in a specific domain, suggests a complex cluster of skills that may range on a continuum from very low to very high; and furthermore, that such skills involve not only certain formal and informal knowledge, but also desired beliefs and attitudes, habits of mind, and a critical perspective (Gal 2002; Geiger et al. 2015). This has already been recognized in the area of mathematics education, where conceptions of mathematical literacy (Kilpatrick 2001) or quantitative literacy (Steen 2001) have extended the definitions of the mathematical knowledge desired of school graduates, in light of the complex nature of everyday situations adults have to understand and manage.

The literacies pertaining to the area of statistics usually fall under the umbrella term *statistical literacy*, though there are several related constructs, such as *probability literacy* (Gal 2005), *data literacy*, or *risk literacy*. According to Gal (2002), statistical literacy refers to people's ability to interpret, critically evaluate, and (when relevant) express their opinions regarding statistical information, data-related arguments, or stochastic phenomena. He further argues that statistically literate behavior requires the joint activation of dispositions (supporting motivation, positive attitudes, and a critical stance), coupled with five cognitive knowledge bases: literacy skills, statistical knowledge (also including some knowledge of probability, albeit informal), mathematical knowledge, contextual or world knowledge, and knowledge of critical questions that have to be asked. Watson (2002), as well as Watson and Callingham (2003), have described three levels that reflect increasing degrees of sophistication in statistical literacy and can be viewed as a developmental trajectory through which learners may progress: (1) basic understanding of probabilistic and statistical terminology; (2) understanding of statistical language and concepts when they are embedded in the context of wider social discussion; (3) and the ability to apply a questioning attitude to statistical claims and arguments.

Several of the constructs described above have also been defined and evaluated as part of large-scale international comparative assessments and, in some countries, as part of national assessments. Consequently, we do have some information about proficiency distributions. For instance, results from the Programme for International Assessment of Adult Competencies (PIAAC, also referred to as the OECD Survey of Adult Skills) for 33 countries that participated in the first two waves of this comparative assessment show that in the area of adult *numeracy*, a large percentage of adults in most countries, usually between 20–40%, has low or very low numeracy skills. In most countries few adults (less than eight percent) reach the highest proficiency levels possible in the assessment, though there is considerable variation around these general patterns at the country level (OECD 2013b).

International comparative data that shed some light on knowledge and skills of adults in the specific area of statistical knowledge comes mainly from the OECD's Programme for International Student Assessment (PISA). While PISA assesses proficiencies of students aged 15–16 years, it shares many similarities with the PIAAC assessment of adult proficiencies (Tout and Gal 2015), both in terms of its conceptual framework and its use of assessment items that purport to simulate real-world demands facing future adults. Specifically, the PISA 2003 and PISA 2012 assessment cycles have reported separate findings in four subareas of mathematical literacy, one of which is 'uncertainty & data' (i.e., statistics & probability). In PISA 2003 (OECD 2004), whose test-takers now approach 30 years of age and thus classify as adults, results were reported for six levels of proficiency, from 1 (lowest) to 6 (highest), including a seventh group of 'below level 1'. At the risk of oversimplifying the complex pattern of reported results, the findings suggest that, across all 25 participating countries, on average, 46% of the respondents did not reach level 3, showing poor ability to read and interpret statistical displays and statistical messages that involve more than a few straightforward data elements. A similar pattern was reported in PISA 2012 (OECD 2013a), whose participants are now aged around 18 years.

Results concerning numeracy (in PIAAC) and mathematical literacy (in PISA) proficiencies thus suggest that in many countries the adult population is very diverse in terms of its ability to comprehend quantitative and statistical messages. Further, PIAAC also shows similar patterns regarding other skills that are involved in finding, understanding and engaging official statistics, in particular *reading literacy* and the *ability to solve problems in [information] technology-rich environments*. It is, of course, possible that quantitative and statistical competencies at the individual level change (even evolve positively) over time. Nevertheless, when viewed together, findings and gaps documented by PISA and PIAAC, motivate and inform further dialogue about ways to conceptualize, and in turn improve, official statistics literacy.

3. Towards a Definition of Official Statistics Literacy

3.1. Making Sense of Official Statistics: An Overview of Sources

In this section we reflect on what are the unique or specific knowledge bases and skills that citizens at large and non-specialists need in order to make sense of official statistics in addition to having the knowledge bases and skills subsumed under the more generalized constructs reviewed in the previous section. A specific point of comparison pertains to the knowledge expected of students who have taken an introductory statistics course at the undergraduate level, which may be the last, and for some students the only, structured exposure to statistics (Moore 1998).

The traditional content of introductory courses for non-specialists is reflected in the table of contents of basic statistics textbooks. There is no single structure for an introductory course across textbooks and disciplines, and even well-established series (such as Freedman et al. 2007; or Moore 2012) change some of their internal elements over time. That said, a typical introductory course for non-majors may cover a mix of ideas and techniques related to topics such as:

- the purpose of statistics,
- descriptive statistics (e.g., measures of center and spread), normal curve and distributions such as z and t,
- some graphing,
- notions of association and correlation as well as some regression,
- sampling and sampling error,
- basic ideas concerning probability and binomial distribution,
- basics of statistical inference (including expected values, confidence intervals and simple statistical significance tests),
- and possibly other subtopics such as data collection methods (surveys and experiments), measurement and questionnaire design.

Not surprisingly, the contents of an introductory course and related teaching approaches, have been the subject of expert analysis over several decades, in the United States in particular. Numerous scholars have debated the sequencing as well as relative importance and weight of some components (Moore and Cobb 2000; Chance and Rossman 2001; Cobb 2007; Malone et al. 2012). There are calls to change the balance between conceptual understanding and computations or the use of technology, along with the need to deepen understanding of big ideas in statistics via the use of randomizations or simulations (e.g., Tintle et al. 2015), for examining alternative approaches to teaching (Vehkalahti 2016), or for expanding the attention to qualitative ideas in statistics (Ograjenšek and Gal 2016).

There is a plethora of introductory textbooks and scholarly interest in, and debates on, the content of introductory statistics courses for university and high-school students. However, there are virtually no scholarly debates or sources that provide an integrative view of basic knowledge elements regarding *official statistics* expected of the same students, and adults at large. In this context, we note the work by the United Nations Economic Commission to Europe (UNECE), which, as part of its efforts to improve good practices for communicating and using official statistics, has also aimed to define general knowledge elements in statistics required of decision-makers and citizens. The UNECE (2012) proposed four primary areas:

(1) data awareness,
(2) ability to understand statistical concepts,
(3) ability to analyse, interpret and evaluate statistical information, and
(4) ability to communicate statistical information and understandings.

We believe that these areas are generally important, but not sufficiently specific to the area of official statistics.

To contribute to further thinking in this regard, we have reviewed and integrated information from the references mentioned in this section so far, along with references from the following three types of sources:

- Syllabi of established programs that impart either graduate degrees or diplomas related to official statistics (e.g., by Central Statistics Office Ireland or Statistics Finland); selected key publications on the websites of national official statistics providers active in statistics education (e.g., Australian Bureau of Statistics, Statistics New Zealand, Statistics Canada); and texts from Eurostat's Statistics Explained website.
- Preliminary insights from ProCivicStat, a new collaboration effort by six universities in five countries (Germany, Hungary, Israel, Portugal, and the United Kingdom) funded by the European Commission's ERASMUS+ program. The project (see http://community.dur.ac.uk/procivic.stat) aims to promote civic engagement and understanding among young adults regarding 'civic statistics' about key societal phenomena (Engel et al. 2016). Among other things, the consortium of partners has analyzed the cognitive demands of texts and displays in publications of official statistics providers, news media, and other stakeholders, and is developing a new framework regarding skills and attitudes needed to understand civic statistics and related teaching resources.

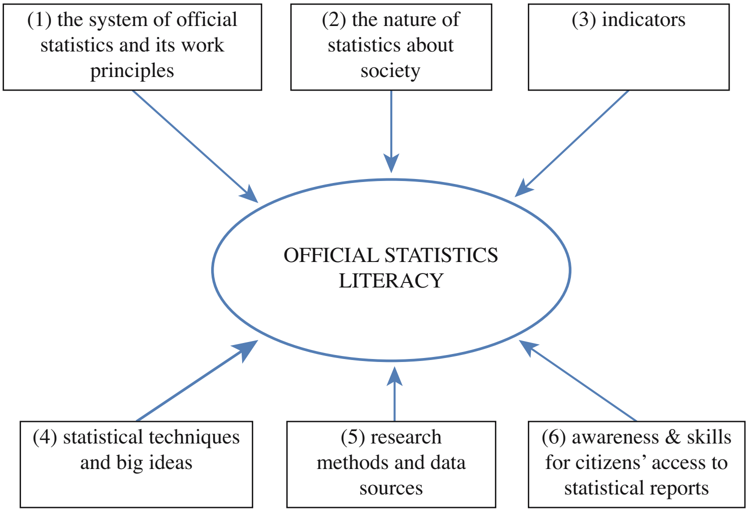

(1) the system of official statistics and its work principles
(2) the nature of statistics about society
(3) indicators
OFFICIAL STATISTICS LITERACY
(4) statistical techniques and big ideas
(5) research methods and data sources
(6) awareness & skills for citizens' access to statistical reports

Fig. 1. Proposed model of six building blocks (areas) of official statistics literacy.

- Analyses of products, users of official statistics providers, and discussions of official statistics aspects of media literacy and science literacy (e.g., Bregar et al. 2000; Gal 2003a, 2003b; Gal and Bosley 2005; Gal and Murray 2011; Lancaster 2011; Ograjenšek et al. 2013; von Roten and de Roten 2013; Poljičak Sušec et al. 2014; Coddington 2015).

While a full analysis of information from these diverse sources is still in progress, at this interim stage we can propose a new model, depicted in Figure 1 and explained later in this section in more detail.

The model encompasses six elements about which non-specialists and adults in general should possess knowledge to be considered literate in official statistics:

(1) the system of official statistics and its work principles,
(2) the nature of statistics about society,
(3) indicators,
(4) statistical techniques and big ideas,
(5) research methods and data sources, and
(6) awareness and skills for citizens' access to statistical reports.

3.2. *The System of Official Statistics and Its Work Principles*

Adults can be expected to know that their country has a system of statistics producers or official statistics providers that work cooperatively on the basis of fundamental principles (United Nations 2014). These official statistics providers aim to make data and diverse information products available to keep policy-makers, various user groups, and the general public apprized of the current economic and social situation. Their aim is also to facilitate description of changes over time (historical analysis) and to create predictions (e.g., population projections) in order to anticipate future trends for a wide range of topics relevant to society.

Towards these goals, official statistics providers employ scientific principles, accepted procedures and standards, as well as quality criteria for data collection, analysis, reporting, release, and dissemination (Biemer et al. 2014). Official statistics providers aim to collect, analyze data and report findings in an impartial and ethically sound way, and work in ways that create and retain public trust and confidence in the national statistical system (Holt 2008).

We argue that citizens may also need to know about seemingly more technical aspects of the broad statistical system that affect how, and what types of, statistics are reported to the public. For example, the fact that official statistics providers release certain statistics (e.g., regarding economic indicators such as the CPI, the GDP, or population statistics) using prescribed release schedules; that they may revise and correct already published findings due to methodological or other considerations; or that they have to use international standards for collecting and reporting key statistics in order to enable comparability across societies. These, and related details about the statistical system, are normally not included in introductory statistics courses, yet are essential for adults to understand, in general terms, where official statistics come from, how they are produced and reported, and why they are produced and reported in specific ways.

3.3. The Nature of Statistics about Society

Based on work by the ProCivicStat project noted earlier, Engel et al. (2016) claim that to be fully engaged, citizens need to understand 'civic statistics' with regard to past trends, present situations, and possible future changes in diverse areas of importance to society such as demographics, employment, wages, migration, health, crime, poverty, access to services, energy, education, human rights, and other domains**.** The *ProCivicStat* analysis points to five general characteristics of civic statistics and the ways in which they are reported to the public:

- *Multivariate phenomena*. Data about social variables of interest usually do not stand in isolation; their description and understanding involves other variables that are correlated, interact with each other, or have non-linear relationships.
- *Aggregated data*. Statistics about society are often reported not with regard to continuous raw variables per se, but involve data that are grouped in diverse ways, sometimes using qualitative variables (Ograjenšek and Gal 2016). Thus, data may be combined into indicators, or reported for multiple subgroups.
- *Dynamic data*. Civic statistics are often not the result of a one-time data collection effort (e.g., unlike a single survey discussed in an introductory statistics course) but based on data collected periodically (e.g., each month, quarter, year) or on a comparative basis (e.g., in multiple countries). Consequently, data are often reported as a trend over time, and may be updated when new data become available or old data are re-evaluated, leading to the creation of an information space and displays that are more complex and dense compared to the simplified data used in teaching introductory statistics.
- *The use of rich texts*. Statistical information about society is brought to the public mainly via texts published by statistics producers (e.g., press releases or brief reports) or via articles in the media. Thus, text is a primary medium for communicating statistics (Gal 2002), and the public needs to be capable of comprehending and critically interpreting different genres of writing, such as formal language used in official reports, journalistic writing, and more.
- *Diverse visualizations*. Since data and findings about social phenomena are multivariate, dynamic, and aggregated, their description across time or comparison units requires the use of diverse types of representations. Hence, today users encounter a range of static, dynamic, and interactive visualizations (Ridgway 2016) that are much broader and more sophisticated compared with the limited range of graphs and histograms used in introductory classes.

The five broad characteristics of 'civic statistics' outlined by the ProCivicStat project influence the nature of the data and statistical messages from statistics providers that reach the general public and non-specialist user groups, albeit in different ways. Information products describing statistics with the above characteristics are made available to the public via multiple channels, including traditional (printed and visual) media, social media, private entities such as NGOs, advocacy groups, independent research centers, and other information or data intermediaries (e.g., bloggers). These 'secondary players' usually present only selected aspects of the original publications or findings, and may sometimes re-analyze or present them in ways that aim to explore specific topics of social or political significance or advance specific points of view. Some findings may also be redistributed via social or digital networks and discussed by private citizens, NGOs, or academic instructors, outside the purview of the original producers.

3.4. Indicators

What kind of official statistics are conveyed to the public (and to policy-makers) via media channels? The answer is complex, of course, as many types of findings and insights are shared, and their flavor may change across topics or countries. Yet all official statistics providers create messages regarding levels or changes in dozens of indicators, such as unemployment level, child mortality, gross domestic product, or income inequality (e.g., Gini coefficient), that reflect the state of some aspect of our society, economy, or well-being.

These and many other indicators in use by official statistics providers are often not raw variables, such as those encountered in introductory statistics, but rather combinations of data elements that may be expressed as percentages, ratios, or numbers on arbitrary scales. They may be computed or derived, from simple rates to complex aggregates of weighted elements. They may be based either on objective (e.g., consumer spending) or subjective data (e.g., consumer confidence), and their definitions may develop and change over time to reflect society's needs for information about itself. However they are defined, indicators are widely used by official statistics providers to report on a wide range of issues, and their understanding is essential for all citizens.

Although they are seemingly included in the broad description of the prior aspect regarding the nature of statistics about society, we highlight indicators as a separate aspect of official statistics because of their privileged role in public discourse and as a key product category that may influence policy-makers. Yet, surprisingly, despite their centrality in society and their prevalence in public and political discourse, indicators are hardly ever described or analyzed in textbooks and statistics curricula for non-specialists. (That said, see Haack's 1979 textbook for non-statisticians for an early, yet quite comprehensive, treatment of indicators.)

3.5. Statistical Techniques and Big Ideas

There is a vast range of techniques used by official statistics providers. The basics of descriptive statistics and statistical inference may be encountered by the subgroup of those who learn statistics at an introductory level at the high school or college level.

In this section, however, we refer to an array of additional techniques and ideas that are frequently used in official statistics, such as moving averages, seasonal adjustment, data smoothing, case weighting, and the like. Specific areas of official statistics may have additional important approaches, such as the use of models and assumptions for population projections, or national accounts and purchasing power parities in economic statistics (Pfeffermann 2015).

Understanding of these and related techniques may not be essential for the understanding of statistics reported in the media, as technical terminology related to the methods listed above is quite often not used in the regular media, except in the business section of newspapers. However, knowing about their existence, even if they are treated as a 'black box' and their actual computational nature is not learned, may be important if an adult wants to adopt a questioning stance or desires to understand more deeply how certain conclusions are derived, or how credible the underlying data are. For instance, how is it possible to conduct comparisons across different economic, financial and social systems that have monetary systems with different characteristics, or if social or economic conditions (e.g., inflation) have changed the base against which comparisons are being made?

Furthermore, critical interpretation of the statistical findings released by official statistics providers also requires an understanding of notions pertaining to confounding variables or conditioning of probabilities (Schield 2011) and related statistical ideas and techniques that are usually not afforded much attention in introductory-level classes.

3.6. Research Methods and Data Sources

Knowledge bases related to methodological issues are often spread between the discipline of statistics and the domain loosely called 'research methods' (Murtonen 2015). There is an overlap between them (Gal 2007; Meng 2009), and consequently there are long-standing debates as to where statistics ends and research methods begin. What statisticians view as fitting under 'methodology and enquiry processes' may only cover some elements of what experts from other disciplines may have in mind (Gal and Ograjenšek 2010; Ograjenšek and Gal 2016).

At university, the learning of research methods is spread over multiple degree levels (e.g., undergraduate, graduate, doctoral), and is organized in diverse ways across different academic institutions and departments (Deem and Lucas 2006). Regardless of the existing diversity, however, the logic of the statistical enquiry process (Wild and Pfannkuch 1999) or the PPDAC (problem, plan, data, analysis, conclusion) cycle (MacKay and Oldford 2000) is likely to be encountered.

Consequently, some students may learn about surveys vs. experiments, sampling and randomization, some aspects of measurement or questionnaire design, or sources affecting internal and external validity of different research designs. Official statistics providers, however, make use of a wider range of data sources and methods for data collection. Examples include the use of a national census, the increasing role of administrative records or public registers, and the many potential types of 'big data' (Daas et al. 2015) that accumulate from sources that fall outside the traditional distinction between surveys and experiments. Further, even when samples are used by official statistics providers, they are usually utilized on a large scale or a cycling basis (e.g., social surveys, employment surveys, employer-based or enterprise surveys) and involve weighting issues if a whole country or sector is to be represented. Given the repeated nature of many official surveys or data-collection efforts and the high-stakes nature of the findings derived from them, issues related to various error sources such as sample design, nonresponse, or respondent bias that determine data quality or credibility receive much attention in official statistics.

3.7. Awareness and Skills for Citizen Access to Statistical Reports

As already explained, citizens need to know that much of the statistical information or statistics-based messages that appear in the media, in fact, originate in a release or report prepared by a statistics provider (Gal 2003b). This is a source differentiated from reports generated by journalists based on ‘open data’ sources (Coddington 2015), even though such open data sources themselves may, in fact, have been created by an official statistics provider.

Hence, as UNECE (2012) also recognizes, engaged citizens need to be aware of the fact that they can access the website of an official statistics provider and often get free and easy access to the same data products or publications used by the media (i.e., a press release or a technical report). This means that adults can verify or cross-check claims they have encountered in the media, and learn about a certain topic beyond the selective information in a media article.

However, the website of a typical official statistics provider presents a complex environment, certainly to newcomers and often even to more experienced users (Gal 2003a, 2005; Bregar et al. 2006; Ridgway 2015). Citizens have to search for information without necessarily knowing how to search for it, or how to use glossaries or help systems that are often written for professionals and not for the general public (Gal and Bosely 2005). Further, they need to be aware of the fact that on the provider’s website, they may find prior versions of the same information products (e.g., press releases and reports from the same survey which was conducted earlier). In addition, official statistics providers also publish technical information, ‘metadata’, about how the data were gathered or a survey was implemented, how variables were defined and measurements performed, including access to the actual phrasing of survey questions. Finally, some official statistics providers enable citizens to use data visualizations in order to view certain data from multiple viewpoints, and in some cases even provide online analytic tools that enable citizens to conduct their own analysis on aggregated data.

The upshot is that the scope of the information presented on an official statistics provider’s website about a topic is much broader and deeper compared with the simplified information or data that students encounter in a statistics class, and may require more sophistication and mental flexibility on the part of the users.

4. Discussion: Achieving Official Statistics Literacy

4.1. Critical Examination of Past and Present Efforts to Promote Official Statistics Literacy

To date, discussions of the connections between official statistics providers and statistics educators have focused in large part on how official statistics providers can facilitate improvement of generic statistics education at the school or university level. Within this framework, official statistics providers have been contributing to teachers’ professional development by offering datasets, lesson plans, ideas for projects and poster competitions, and other resources that can inform class activities or highlight the importance of official statistics. Some official statistics providers have also developed specialized sections on their websites that are geared towards teachers and students, or support the international CensusatSchool project and its various derivatives (Davies 2011). The richness and importance of such and related activities have been noted and appreciated around the world (see e.g., Sanchez 2008; Townsend 2011; Helenius and Mikkelä 2011; or MacCuirc 2015).

As valuable as these efforts to increase general statistical literacy are, it needs to be pointed out that they did little to systematically promote understanding of issues pertaining specifically to official statistics.

In this article we sketched a new model of six interconnected knowledge elements of the world of official statistics, about which non-specialists and adults at large should possess knowledge to be considered literate in official statistics. In the prior sections, we analyzed how such knowledge elements go above and beyond what is usually associated with learning introductory statistics, or how statistical literacy related to official statistics is understood by bodies such as the UNECE (2012). All our findings imply that unique efforts are needed to promote official statistics literacy.

We believe that improvements that may affect knowledge levels of current (primary and secondary) school pupils or tertiary students, as valuable as they are, do not directly impact the skill set of the current adult population, which is outside of formal education systems' range, yet comprises the main audience that statistics providers try to reach. Given the relatively slow rate at which the adult population is replaced by younger cohorts, even if knowledge among school and university graduates about official statistics vastly improved overnight, it would still take two to three decades for new knowledge gained at school level to be shared among (the younger) half of the adult population. Consequently, many adults will still lack such knowledge for decades to come.

For these reasons, it is important to continue existing specially targeted collaborations between official statistics providers and school-level educators, as noted by sources discussing the development of statistical literacy at school level (e.g., Gal 2002; Sanchez 2008; Watson 2013). Townsend (2011), Helenius and Mikkelä (2011), UNECE (2012), MacCuirc (2015), de Smedt (2016), and others, describe numerous relevant initiatives and services aiming to promote official statistics literacy that have been implemented over the years by statistics providers at national and sometimes international level.

Examples include:

- the provision of workshops, brief online courses and supportive training materials about official statistics designed for specific non-specialist user groups with known characteristics such as journalists, business leaders, or government workers,
- the provision of short leaflets about key indicators that affect the general public, such as the consumer price index,
- the preparation and posting of answers to frequently asked questions about finding or interpreting selected key official statistics on the provider's website,
- the provision of simplified explanations about official statistics in selected key areas (e.g., Eurostat's Statistics Explained mini-website on migration statistics),
- the preparation and posting of answers and non-technical explanations about selected basic statistical terms, statistical glossaries, and more.

Such initiatives and activities are essential and have the potential to contribute to the mission of official statistics providers and to the ability of users to comprehend specific information products in several important ways. Yet, we believe the vision of systematically promoting official statistics literacy within the general adult population (including actions in countries with characteristics that differ from the few that have spearheaded educational services and activities) requires an examination of additional directions –from a long-range future collaborative perspective.

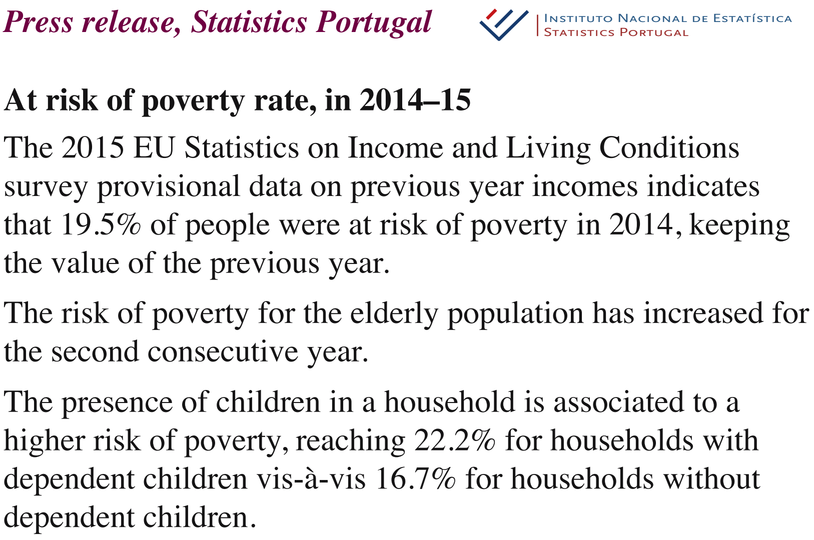
4.2. Proposed Directions for Future Collaborative Actions to Promote Official Statistics Literacy

Taken together, the six elements of official statistics proposed in this article and depicted in Figure 1 imply that if citizens aim to understand official statistics about society (i.e., civic statistics) to which they are exposed through the media, or if citizens attempt to find, read, and critically comprehend actual products (e.g., press releases, highlights, annotated visualizations) on the website of an official statistics provider, they need a knowledge base that is above and beyond what is taught in regular introductory statistics classes for non-specialists.

Figure 2 presents an illustrative example for how several of the six elements or knowledge areas in our proposed model co-exist in a seemingly simple product from an official statistics provider. The text in Figure 2 is an excerpt from a one-page regular press release by a national statistics provider (Statistics Portugal) that the public may hear about via a news website or a newspaper item. The example is taken (with permission) from Gal et al. (2016) who developed it for a workshop on understanding 'civic statistics' that is part of ongoing work by the aforementioned ProCivicStat project.

Despite its brevity, this excerpt can be used to show how multiple areas in our proposed model are all called upon to comprehend the given text. The text refers to:

- the production of statistics as part of a system of official statistics that relies on general international standards, and generates modifiable or provisional data (area 1),
- the nature of statistics about society, that is, use of rich text to convey a statistical finding, or the dynamic and aggregated nature of statistics (area 2),



Press release, Statistics Portugal

Instituto Nacional de Estatística
Statistics Portugal

At risk of poverty rate, in 2014–15

The 2015 EU Statistics on Income and Living Conditions survey provisional data on previous year incomes indicates that 19.5% of people were at risk of poverty in 2014, keeping the value of the previous year.

The risk of poverty for the elderly population has increased for the second consecutive year.

The presence of children in a household is associated to a higher risk of poverty, reaching 22.2% for households with dependent children vis-à-vis 16.7% for households without dependent children.

Fig. 2. News about poverty – press release from an official statistics provider.

- the use of an indicator, that is, risk of poverty (area 3),
- big ideas in statistics, for example, risk (area 4),
- the use of specific research methods (area 5).

As Gal et al. (2016) explain, the example in Figure 2 also illustrates the need for adults to be able to critically reflect on the origin and quality of data, and on how variables or social phenomena are defined and measured.

With the above in mind, we outline two possible initiatives at the international level, and some additional ideas that specific official statistics providers can implement at a local level.

Firstly, we propose the development of a textbook on official statistics geared towards statistics majors as well as non-majors who may study selected topics in statistics. We note that there are many more non-majors than majors who take only introductory statistics, and the provision of an accessible textbook may be the first step to helping educational institutions develop new modules or whole courses related to official statistics that are currently lacking.

Secondly, we propose the development of an MOOC or a collection of digital (video and audio) teaching modules for entry-level majors, non-majors, and other groups of interest among the general public.

It is hard to expect a single official statistics provider to shoulder responsibility and allocate resources related to both initiatives outlined above, although it would be technically possible. Both initiatives thus call for an international collaborative effort of official statistics providers, statistics educators, specialists in applied fields that rely on official statistics when discussing major concepts inherent to their disciplines, and other stakeholders. Such an effort can, of course, benefit from existing materials and frameworks developed in the context of existing diploma and degree programs listed in the previous sections of this article. Textbook developers participating in this collaborative effort could build on experiences gained within the framework of the already mentioned Phare project, which resulted in the modular online Course on the European Economic Statistics (Bregar et al. 2000).

Several organizations, of which some have been referred to earlier in this article, appear to have both the infrastructure, resources, and interest necessary to promote both a textbook and a MOOC as outlined above. These include, among others, the EMOS community, which presently includes over 20 universities and cooperating national statistics offices, with support from Eurostat or UNECE, as well as SIAM and networks of official statistics providers in Asia and Oceania. Furthermore, PARIS21 and the UNESCO Institute of Statistics, and other organizations involved in statistical capacity-building in developing countries are well positioned to further clarify the knowledge needs of non-specialists who engage official statistics in such countries.

In addition, large professional associations with an international outreach and long-standing interest and activities in statistics education can also facilitate collaborations and the long-term development of a textbook and a MOOC. Key actors may be the International Statistical Institute (ISI) and its relevant divisions (the International Association for Statistics Education – IASE and the International Association for Official

Statistics – IAOS) as well as the Royal Statistical Society (RSS), the American Statistical Association (ASA), and others.

The need for a comprehensive knowledge base related to official statistics may raise questions about the relative importance of the six knowledge elements outlined in our proposed model, as well as about preferred learning trajectories. Such questions may seem useful, given the need to prioritize development efforts when writing a new textbook or developing a new MOOC. We believe that all six areas are important in the long term, and there is no known consensus yet as to what may be considered 'basic' or more 'advanced' levels of knowledge in this regard, or a best learning sequence. One needs to take into account possible learning trajectories for learners with different starting points (e.g., in terms of basic knowledge in statistics or other parameters), and the need to motivate learners diverse in their background, learning styles, and so on, along the way. Arguably, at an initial stage of development, it may be advisable to select a few high-visibility, some simple and some more advanced indicators or findings of interest to the general public, and discuss some basic methodologies and working principles related to them. At a later stage, it is possible to expand the coverage of these and all other areas in our model.

The above preliminary ideas notwithstanding, we believe that the design of a textbook and a MOOC can, and should, benefit from current technological flexibilities, and be conceived from the outset as an integrated collection of digital learning resources that will be developed in parallel by multiple partners. This may reduce the need for topic prioritization. Many potential development partners that were mentioned above can build on the already existing partial resources (shareable materials from existing diploma and degree programs aimed at non-specialists) as well as ongoing work by other stakeholders (e.g., the already mentioned ProCivicStat, or individual instructors around the world) who can be called upon to share their teaching materials.

The envisioned collaborative digital resource enables the development of multiple variants of textbook chapters or MOOC units, distributed across multiple partners who work in parallel; with common as well as nation-specific modules. Subsequent review and revision processes can also move in parallel, with new resources added and hyperlinked in iterative stages. Such an approach can help to shorten the development timeline to a degree that enables the coverage of all six areas in our proposed model, initially in English, given its position in the international statistical system, with translation to other languages and localized adaptations moving ahead as materials in English become available.

Finally, apart from the two initiatives envisioned above, at the local level official statistics providers can take additional steps in order to help educate providers who work with adult learners and college or school-level populations. Educators can be equipped with collections of examples of how the media reports about press releases or other official publications, since virtually all official statistics providers nowadays use clipping services or media analysis companies that monitor all media channels. Hence, official statistics providers could develop focused packages, organized around specific issues of social significance, including the original press release and several real-life examples of how data and findings were reported in diverse media channels, selected to illustrate proper, as well as distorted, or one-sided use of statistics. Such packages could be accompanied by suggestions for in-class discussion and take-home assignments.

In summary, it is important to state that the conceptualization of the building blocks of official statistics literacy presented in this article is preliminary and open to debate, since we live in a dynamic world. Discussions emerge in professional channels on the era of ‘open data’ and its implications for both producers and users of statistics, and on the need for public understanding of statistics (von Roten 2006; von Roten and de Roten 2013). Ridgway (2016) points out that significant developments such as open data, big data, data visualisation and the rise of data-driven journalism, provoke new sorts of questions, make possible new sorts of answers and are changing the nature of available evidence, the ways in which it is presented and used to influence policy, public opinion and business practices, and the skills needed to interpret it.

The six elements we identified combine both abstract ideas and a general understanding of a complex working system in its social ecology, as well as knowledge bases of a more technical nature. Details of these elements and their operationalization have to be further examined and developed in more detail, both because official statistics itself is practiced in somewhat different ways in different contexts or by official statistics providers with different missions, and because it is evolving over time, as outlined above. We hope that the ideas proposed in this article will initiate a productive dialogue and ultimately lead to further pragmatic development-friendly decisions among statistics providers and other stakeholders interested in active promotion of official statistics literacy.

5. References

Ancker, J.S. and D. Kaufman. 2007. “Rethinking Health Numeracy: A Multidisciplinary Literature Review.” *Journal of the American Medical Informatics Association* 14: 713–721.

Biemer, P., D. Trewin, H. Bergdahl, and L. Japec. 2014. “A System for Managing the Quality of Official Statistics.” *Journal of Official Statistics* 30: 381–415. Doi: http://dx.doi.org/10.2478/jos-2014-0022.

Bregar, L., I. Ograjenšek, and M. Bavdaž. 2000. “Teaching Economic Statistics in a Digital Environment.” In *New approaches in applied statistics*, edited by A. Ferligoj and A. Mrvar, 237–249. Available at: http://mrvar.fdv.uni-lj.si/pub/mz/mz16/abst/bregar.htm (accessed 30 August 2016).

Bregar, L., I. Ograjenšek, and M. Bavdaž. 2006. “Use of Web-Based Public Databases in Statistics Courses: Experiences and Challenges. *ICOTS-7*. Available at: http://iase-web.org/documents/papers/icots7/7A4_BREG.pdf (accessed 30 August 2016).

Chance, B.L. and A.J. Rossman. 2001. “Sequencing Topics in Introductory Statistics: A Debate on What to Teach When.” *The American Statistician* 55: 140–144.

Citro, C.F. and Straf, M.E. (Eds.). 2013. *Principles and Practices for a Federal Statistical Agency*. National Research Council, Committee on National Statistics, 5th ed. Washington: National Academy Press.

Cobb, G.W. 2007. “The Introductory Statistics Course: A Ptolemaic Curriculum?” *Technology Innovations in Statistics Education* 1(1).

Coddington, M. 2015. “Clarifying Journalism’s Quantitative Turn: A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-Assisted Reporting.” *Digital Journalism* 3: 331–348.

Daas, P.J., M.J. Puts, B. Buelens, and P.A. van den Hurk. 2015. “Big Data as a Source for Official Statistics.” *Journal of Official Statistics* 31: 249–262.

Davies, N. 2011. “Developments of AtSchool Projects for Improving Collaborative Teaching and Learning in Statistics.” *Statistical Journal of the IAOS* 27(3, 4): 205–227.

de Smedt, M. 2016. “European Statistics and Eurostat’s Contribution to Improving Statistical Literacy.” In J. Engel, Ed., Proceedings, IASE Roundtable on Promoting Understanding of Statistics About Society, Berlin. Available at: http://iase-web.org/Conference_Proceedings.php?p=Promoting_Understanding_of_Statistics_about_Society_2016. (accessed 2 January, 2017).

Deem, R. and L. Lucas. 2006. “Learning About Research: Exploring the Learning and Teaching/Research Relationship Amongst Educational Practitioners Studying in Higher Education.” *Teaching in Higher Education* 11(1) : 1–18.

Engel, J., I. Gal, and J. Ridgway. 2016. “Mathematical Literacy and Citizen Engagement: The Role of Civic Statistics.” Paper presented at the 13th International Congress on Mathematics Education (ICME13).

Forbes, S. and A. Keegan. 2016. “Helping Raise the Official Statistics Capability of Government Employees.” *Journal of Official Statistics* 32: 811–826.

Freedman, D., R. Pisani, and R. Purves. 2007. *Statistics* (4th Ed). Norton.

Gal, I. 2002. “Adult Statistical Literacy: Meanings, Components, Responsibilities.” *International Statistical Review* 70: 1–25.

Gal, I. 2003a. “Teaching for Statistical Literacy and Services of Statistics Agencies.” *The American Statistician* 57: 80–84.

Gal, I. 2003b. “Expanding Conceptions of Statistical Literacy: An Analysis of Products from Statistics Agencies.” *Statistics Education Research Journal* 2: 3–21. Available at: http://iase-web.org/documents/serj/serj2(1).pdf (accessed 30 August, 2016).

Gal, I. 2005. “Towards ‘Probability Literacy’ for All Citizens.” In *Exploring probability in school: Challenges for teaching and learning*, edited by G. Jones, 43–71. London: Kluwer Academic Publishers.

Gal, I. 2007. “Research Methods: Reflections on Teaching Frameworks and Research.” In *Learning and teaching of Research methods at University*, edited by M. Murtonen, J. Rautopuro, and P. Väisänen. Turku, Finland: Finnish Educational Research Association.

Gal, I. and S. Murray. 2011. “Users’ Statistical Literacy and Information Needs: Institutional and Educational Implications.” *Statistical Journal of the IAOS* 27(3–4): 185–195.

Gal, I. and I. Ograjenšek. 2010. “Qualitative Research in the Service of Understanding Learners and Users of Statistics.” *International Statistical Review* 78: 287–298.

Gal, I., J. Ridgway, and J. Nicholson. 2016. “Exploration of Skills and Conceptual Knowledge Needed for Understanding Statistics About Society.” In J. Engel, Ed., Proceedings, IASE Roundtable on Promoting Understanding of Statistics About Society, Berlin. Available at: http://iase-web.org/Conference_Proceedings.php?p= Promoting_Understanding_of_Statistics_about_Society_2016 (accessed 2 January, 2017).

Gal, I. and J. Bosley. 2005. “Non-Specialist Users and Their Information Needs: An Exploratory Study at the US Bureau of Labor Statistics.” *Proceedings, 55th World*

Statistics Congress, Sydney. Voorburg: International Statistical Institute. Available at: http://www.stat.auckland.ac.nz/~ iase/publications/13/Gal-Bosley.pdf (accessed 30 July, 2016).

Gal, I., M. van Groenestijn, M. Manly, M.J. Schmitt, and D. Tout. 2005. "Adult Numeracy and Its Assessment in the ALL Survey: A Conceptual Framework and Pilot Results." In *Measuring Adult Literacy and Life Skills: New Frameworks for Assessment*, edited by S.T. Murray, Y. Clermont, and M. Binkley, 137–191. Ottawa, Canada: Statistics Canada.

Geiger, V., M. Goos, and H. Forgasz. 2015. "A Rich Interpretation of Numeracy for the 21st Century: A Survey of the State of the Field." *ZDM* 47(4): 531–548.

Haack, D.G. 1979. *Statistical Literacy: A Guide to Interpretation*. Duxbury Press/ Wadsworth.

Harraway, J.A. and S.D. Forbes. 2013. "Partnership Between National Statistics Offices and Academics to Increase Official Statistical Literacy." *Statistical Journal of the IAOS* 29: 31–40.

Helenius, R. and H. Mikkelä. 2011. "Statistical Literacy and Awareness as Strategic Success Factors of a National Statistical Office: The Case of Statistics Finland." *Statistical Journal of the IAOS* 27(3, 4): 137–144.

Holt, D.T. 2008. "Official Statistics, Public Policy and Public Trust." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2): 323–346. Doi: http://dx.doi.org/10.1257/jel.52.1.5.

Karaali, G., E.H. Villafane Hernandez, and J.A. Taylor. 2016. "What's in a Name? A Critical Review of Definitions of Quantitative Literacy, Numeracy, and Quantitative Reasoning." *Numeracy* 9(1): Article 2. Doi: http://dx.doi.org/10.5038/1936-4660.9.1.2.

Kilpatrick, J. 2001. "Understanding Mathematical Literacy: The Contribution of Research." *Educational Studies in Mathematics* 47(1): 101–116.

Lancaster, G.A. 2011. "How Statistical Literacy, Official Statistics and Self-Directed Learning Shaped Social Enquiry in the 19th and Early 20th Centuries." *Statistical Journal of the IAOS* 27(3, 4): 99–111.

Lusardi, A. and O.S. Mitchell. 2014. "The Economic Importance of Financial Literacy: Theory and Evidence." *Journal of Economic Literature* 52(1): 5–44. Doi: http://dx.doi.org/10.1257/jel.52.1.5.

MacCuirc, E. 2015. "You Don't Teach, Students Learn: Lessons Learned in Statistical Literacy and Statistical Education in Ireland." *Austrian Journal of Statistics* 44: 73–83. Available at: http://www.ajs.or.at/index.php/ajs/article/view/62 (accessed 30 July, 2016).

MacKay, R.J. and R.W. Oldford. 2000. "Scientific Method, Statistical Method and the Speed of Light." *Statistical Science* 15: 254–278.

Madison, B.L. 2014. "How Does One Design or Evaluate a Course in Quantitative Reasoning?" *Numeracy* 7(2): 3. Doi: http://dx.doi.org/10.5038/1936-4660.7.2.3.

Malone, C.J., J. Gabrosek, P. Curtiss, and M. Race. 2012. "Resequencing Topics in An Introductory Applied Statistics Course." *The American Statistician* 64: 52–58.

Meng, X. 2009. "Desired and Feared—What Do We Do Now and Over the Next 50 Years?" *The American Statistician* 63: 202–210.

Moore, D.S. 2012. *Statistics: Concepts & Controversies* (8th Ed). Freeman.

Moore, D.S. 1998. “Statistics Among the Liberal Arts.” *Journal of the American Statistical Association* 93: 1253–1259.

Moore, D.S. and G.W. Cobb. 2000. “Statistics and Mathematics: Tension and Cooperation.” *American Mathematical Monthly* 107: 615–630.

Murphy, P. 2002. “Teaching Official Statistics in an Irish University Statistics Department.” Proceedings, 6th International Conference on Teaching Statistics (ICOTS6), Pretoria, South Africa. Available at: http://iase-web.org/documents/papers/icots6/4e5_murp.pdf (accessed 30 July, 2016).

Murtonen, M. 2015. “University Students’ Understanding of the Concepts Empirical, Theoretical, Qualitative and Quantitative Research.” *Teaching in Higher Education* 20(7): 684–698.

Nathan, G. 2007. “Cooperation Between a Statistical Bureau and an Academic Department of Statistics as a Basis for Teaching Official Statistics.” Proceedings of the 56th Session, of the International Statistical Institute, Lisbon. Available at: http://iase-web.org/documents/papers/isi56/IPM43_Nathan.pdf (accessed 30 July, 2016).

OECD. 2013a. *PISA 2012 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Vol. I)*. Paris: OECD Publishing.

OECD. 2013b. *Skilled for life? Key Findings from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. 2004. *Learning for Tomorrow’s World: First Results from PISA 2003*. Paris: OECD Publishing.

Ograjenšek, I. and I. Gal. 2016. “Enhancing Statistics Education by Including Qualitative Research.” *International Statistical Review* 84(2): 165–178. Doi: http://dx.doi.org/10.1111/insr.12158.

Ograjenšek, I., M. Bavdaž, and L. Perviz. 2013. “Factors Influencing Integration of Official Statistics into Business Study Programmes: In Search of Evidence.” In proceedings of the 58th World Statistics Congress, Hong Kong. (STS40). Available at: www.statistics.gov.hk/wsc/STS087-P5-S.pdf (accessed 30 July, 2016).

Pfeffermann, D. 2015. “Methodological Issues and Challenges in the Production of Official Statistics.” (24th Annual Morris Hansen Lecture). *Journal of Survey Statistics and Methodology* 3: 425–483.

PIAAC Numeracy Expert Group. 2009. *PIAAC Numeracy: A Conceptual Framework*. OECD Education Working Papers, No. 35. Paris: OECD Publishing. Available at: www.oecd-ilibrary.org/education/piaac-numeracy-a-conceptual-framework_220337421165. Doi: http://dx.doi.org/10.1787/220337421165.

Poljièak Sušec, M., N. Jerak Muravec, and H. Stanèiæ. 2014. “Statistical Literacy as an Aspect of Media Literacy.” *Medijska istraživanja* 20(2): 131–155. Available at: http://hrcak.srce.hr/133811 (accessed 30 July, 2016).

Ridgway, J. 2016. “Implications of the Data Revolution for Statistics Education.” *International Statistical Review*. Doi: http://dx.doi.org/10.1111/insr.12110.

Ridgway, J., J. Nicholson, S. Sutherland, and S. Hedger. 2015. “Strategies for Public Engagement with Official Statistics.” In *Proceedings, Advances in Statistics Education*. IASE Satellite conference, Rio de Janeiro, Brazil. Available at: http://iase-web.org/documents/papers/sat2015/IASE2015%20Satellite%2038_RIDGWAY.pdf (accessed 30 July, 2016).

Rutherford, J.F. 1997. "Thinking Quantitatively About Science." In *Why Numbers Count: Quantitative Literacy for Tomorrow's America*, edited by L.A. Steen, 69–74. New York: The College Board.

Sanchez, J. 2008. *Government Statistical Offices and Statistical Literacy*. International Statistical Literacy Project and ISI. Available at: http://iase-web.org/islp/Publications.php?p=Books (accessed 30 July, 2016).

Schield, M. 2011. "Statistical Literacy: A New Mission for Data Producers." *Statistical Journal of the IAOS* 27: 173–183.

Stacey, K. 2015. "The Real World and the Mathematical World." In *Assessing Mathematical Literacy: The PISA Experience*, edited by K. Stacey and R. Turner, 57–84. Springer.

Steen, L.A. (Ed.). 2001. *Mathematics and Democracy: The Case for Quantitative Literacy*. Washington, D.C.: National Council on Education and the Disciplines. Available at: http://www.maa.org/sites/default/files/pdf/QL/MathAndDemocracy.pdf (accessed 30 July, 2016).

Steenvoorden, T., T. Řvigelj, and M. Bavdaž. 2015. "Satisfaction with Official Statistics Producers." *Statistical Journal of the IAOS* 31(4): 645–654.

Tintle, N., B. Chance, G. Cobb, S. Roy, T. Swanson, and J. VanderStoep. 2015. "Combating Anti-Statistical Thinking Using Simulation-Based Methods Throughout the Undergraduate Curriculum." *The American Statistician* 69(4): 362–370.

Tout, D. and I. Gal. 2015. "Perspectives on Numeracy: Reflections from International Assessments." *ZDM–The International Journal of Mathematics Education* 47(4): 691–706.

Townsend, Mary. 2011. "The National Statistical Agency as Educator." *Statistical Journal of the IAOS* 27(3, 4): 129–136.

United Nations. 2014. *Fundamental Principles of Official Statistics*. Available at: http://unstats.un.org/unsd/dnss/gp/FP-NEW-e.pdf (accessed 30 July, 2016).

UNECE. 2012. *Making Data Meaningful: A Guide to Improving Statistical Literacy*. United Nations Economic Commission for Europe. Available at: http://www.unece.org/stats/documents/writing (accessed 30 July, 2016).

Vehkalahti, K. 2016. "The Relationship Between Learning Approaches and Students' Achievements in an Introductory Statistics Course in Finland." In *Proceedings, 60th ISI World Statistics Congress, Rio de Janeiro, Brazil.*

von Roten, F.C. 2006. "Do We Need a Public Understanding of Statistics?" *Public Understanding of Science* 15(2): 243–249.

von Roten, F.C. and Y. de Roten. 2013. "Statistics in Science and in Society: From a State-of-the-Art to a New Research Agenda." *Public Understanding of Science* 22(7): 768–784.

Watson, J.M. 2002. "Discussion: Statistical Literacy Before Adulthood." *International Statistical Review* 70(1): 26–30.

Watson, J. and R. Callingham. 2003. "Statistical Literacy: A Complex Hierarchical Construct." *Statistics Education Research Journal* 2(2): 3–46. Available at: http://iase-web.org/documents/SERJ/SERJ2(2)_Watson_Callingham.pdf (accessed 30 July, 2016).

Watson, J.M. 2013. *Statistical Literacy at School: Growth and Goals*. Routledge.

Wild, C.J. and M. Pfannkuch. 1999. "Statistical Thinking in Empirical Enquiry." *International Statistical review* 67: 223–248.

Xu, L. and B. Zia. 2012. "Financial Literacy Around the World: An Overview of the Evidence with Practical Suggestions for the Way Forward." *World Bank Policy Research Working Paper* #6107. Available at: http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-6107 (accessed 30 July, 2016).

Zwick, M. 2016. "EMOS–Der European Master in Official Statistics." In *Human Resources*, 127–141. Wiesbaden: Springer Fachmedien.

Received October 2016
Accepted January 2017

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 101–122, http://dx.doi.org/10.1515/JOS-2017-0006

Three Methods for Occupation Coding Based on Statistical Learning

Hyukjun Gweon[1], *Matthias Schonlau*[1], *Lars Kaczmirek*[2], *Michael Blohm*[2], *and Stefan Steiner*[1]

Occupation coding, an important task in official statistics, refers to coding a respondent's text answer into one of many hundreds of occupation codes. To date, occupation coding is still at least partially conducted manually, at great expense. We propose three methods for automatic coding: combining separate models for the detailed occupation codes and for aggregate occupation codes, a hybrid method that combines a duplicate-based approach with a statistical learning algorithm, and a modified nearest neighbor approach. Using data from the German General Social Survey (ALLBUS), we show that the proposed methods improve on both the coding accuracy of the underlying statistical learning algorithm and the coding accuracy of duplicates where duplicates exist. Further, we find defining duplicates based on ngram variables (a concept from text mining) is preferable to one based on exact string matches.

Key words: Automated coding; Machine learning; ISCO-88; ALLBUS.

1. Introduction

Classifying a respondent's occupation is essential in official statistics and social science research. It enables the international comparison of the official statistics on occupation and work and is the starting point for numerous status scales or prestige measures. It is a "foundation of much, if not most research on social stratification" (Ganzeboom and Treiman 2003, 159) and social inequality. Because occupation is a risk factor in many diseases, classifying occupations is an important first step for epidemiological analyses, industrial hygiene, and other biomedical sciences.

There are quite a few different classification schemes, but all have hundreds of occupation codes and the codes are always nested in hierarchies. For example, the International Standard Classification of Occupations 1988 (ISCO-88) (Elias 1997) is a classification of four nested levels characterized by four digits. The first digit distinguishes nine major groups, and an undifferentiated tenth major group for the Armed Forces. There are 28 sub-major groups (two-digit combinations), 116 minor groups (three-digit

[1] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1 Canada. Emails: hgweon@uwaterloo.ca, schonlau@uwaterloo.ca, and shsteiner@uwaterloo.ca.
[2] GESIS – Leibniz-Institute for the Social Sciences, PO Box 12 21 55, D-68072 Mannheim, Germany. Emails: lars.kaczmirek@gesis.org and michael.blohm@gesis.org

Acknowledgments: This research was supported in part by the Social Sciences and Humanities Research Council of Canada (SSHRC # 435-2013-0128) (PI: Schonlau) and we are grateful for the support.

combinations) and 390 unit groups (four-digit combinations). Table 1 gives coding for sub-major group 71, extraction and building trades workers.

To ascertain a survey respondent's occupation, typically an open-ended question is asked (Belloni et al. 2014). Alternative ways to find a respondent's occupation include the use of search trees in web surveys (Tijdens 2014, 2015), but open-end questions are most common. The main example in this article is the biannual ALLBUS survey (ALLBUS 2015) conducted by GESIS – Leibniz Institute for the Social Sciences. The ALLBUS survey uses open-ended questions to ask about occupation (Scholz and Wasmer 2009). Using multiple choice questions to elicit four-digit occupation codes is not sensible because there are too many codes, and more importantly, respondents often would not know how to classify themselves because occupation coding rules are complex (International Labour Office 1990; Geis 2011; Elias 1997; Belloni et al. 2014).

Traditionally, assigning an occupation code to each answer text has been conducted manually by human coders. Manual coding is time-consuming and expensive, requiring professional knowledge. Occupation coding is also difficult: there are hundreds of predefined occupation codes and even more occupation titles. For example, the ISCO-88 classification contains 390 four-digit occupation codes. Another difficulty is that coding even by professional coders may be inconsistent. The coding quality of a record depends on the length of the occupation description as well as the difficulty of the words in the record (Conrad et al. 2016).

Table 1. ISCO-88 Sub-Major Group 71: extraction and building trades workers.

71	Extraction and building trades workers		
	711	Miners, shotfirers, stone cutters and carvers	
		7111	Miners and quarry workers
		7112	Shotfirers and blasters
		7113	Stone splitters, cutters and carvers
	712	Building frame and related trades workers	
		7121	Builders
		7122	Bricklayers and stonemasons
		7123	Concrete placers, concrete finishers and related workers
		7124	Carpenters and joiners
		7129	Building frame and related trades workers not elsewhere classified
	713	Building finishers and related trades workers	
		7131	Roofers
		7132	Floor layers and tile setters
		7133	Plasterers
		7134	Insulation workers
		7135	Glaziers
		7136	Plumbers and pipe fitters
		7137	Building and related electricians
		7139	Building finishers and related trade workers not elsewhere classified
	714	Painters, building structure cleaners and related trades workers	
		7141	Painters and related workers
		7143	Building structure cleaners

In an attempt to partially automate coding, researchers have implemented various rule-based coding schemes. For example, if the text answer contained a word matching an entry in a predefined dictionary, then the corresponding code in the dictionary was assigned. More recently, statistical learning or machine learning approaches have been employed: a model is trained on manually coded training data and is then used to predict the most probable code for new data (Statistical learning and machine learning are synonymous for the purpose of this article. For brevity we just use the phrase "statistical learning" for the remainder of the article). This approach is favored, for example, by the Australian Bureau of Statistics (Clarke and Brooker 2011). Autocoders based on statistical learning have also been developed in the United States (Day 2014) and in Germany (Bethmann et al. 2014).

Although the automated methods reduce costs for occupation coding, fully automated coding remains challenging. With partial automatic coding, easy-to-code answers are coded automatically, and-hard-to-code answers are coded manually. A measure of confidence – a numerical score – is used to distinguish between easy-to-code and hard-to-code text answers (Scholtus et al. 2014). For example, the CASCOT system proposes manual coding when a score for the coding quality drops below a modifiable threshold (Jones and Elias 2004).

In this article we consider three new techniques for improving automated coding:

(a) a combination of two statistical learning models for different levels of aggregation,
(b) a combination of a duplicate-based approach with a statistical learning one, and
(c) a modified nearest neighbor approach.

The remainder of this article is organized as follows: In Section 2 we give background on approaches to automated occupation coding. In Section 3, we introduce the three techniques for improving automated coding. In Section 4, we evaluate the proposed approaches with data from the 2006 German ALLBUS survey coded by GESIS based on ISCO-88 codes. In Section 5, we conclude with a discussion.

2. Automated Occupation Coding

This section gives an overview of how to evaluate the performance in automated occupation coding, as well as two types of commonly used approaches: rule-based approaches and approaches based on statistical learning. The new approaches we introduce in this article are mostly based on statistical learning.

2.1. Production Rate and Accuracy

When some answer texts are coded automatically and some are coded manually, a score or a probability is needed to distinguish between hard-to-code and easy-to-code answers. All new records with scores above a threshold are coded automatically; all others are coded manually. The threshold is set according to the desired combination of accuracy and production rate. The production rate is the proportion of observations that can be coded automatically. For a given production rate, accuracy is the proportion of codes that are coded correctly. Note that there is a tradeoff between accuracy and production rate. High accuracy can be achieved for a small number of easy-to-code records. However, as the production rate increases and more difficult answers are included, accuracy tends to decrease. The tradeoff relationship was illustrated in Chen et al. (1993).

2.2. Preprocessing

Before automated coding begins, text is often preprocessed. There is no standardized way of preprocessing, but there are a range of options, such as lower or upper casing all letters, removing duplicate blank spaces, automatically correcting spelling errors, removing very common words (so-called stopwords), and, less common in occupation coding but common in text mining, reducing words to their grammatical root (stemming). Preprocessing is an attempt to reduce the noise in the data.

2.3. Rule-Based Occupation Coding

If the text answer meets a prespecified logical condition (e.g., presence of a certain word) a specific code is assigned. Such "if-then" statements are called rules. Rules are written by experts or can be based on previous data analysis. Rules can be combined using boolean logic. Any one rule-based coding scheme consists of hundreds of rules leading to large dictionaries or look-up tables. Schierholz (2014) reports that this approach rarely codes more than 50% of records accurately. A variation on rule-based methods is to assign a score in favor of a category. If a text answer matches a rule, evidence can accumulate for multiple codes. In the end, the text answer is classified into the occupation code with the highest score. One of the earliest references to rule-based coding is O'Reagan (1972).

Rule-based systems are implemented in many institutions: the Washington State Department of Health (Ossiander and Milham 2006), the 1970 U.S. Population and Housing Census (Knaus 1987), the 1991 census data for Croatia and Bosnia-Herzegovina (Kalpic 1994), and the AIOCS system at the U.S. Census Bureau (Appel and Hellerman 1983; Chen et al. 1993). Statistics Canada further developed the AIOCS system and created the G-Code (formerly ACTR) software (Wenzowski 1988; Tourigny and Moloney 1995), which was also used for Italian census data (Ferrillo et al. 2008). The University of Warwick has a popular tool for automatic categorization called CASCOT (Jones and Elias 2004; see also Elias and Birch 2010 for performance of CASCOT), which has also been adapted to the Dutch language (Belloni et al. 2014).

2.4. Occupation Coding Based on Statistical Learning

Statistical models learn from already classified training data. Such methods can be used not only for occupation coding but also for general classification problems. Once the model has been trained, other observations can be classified automatically.

To build a model, text is first converted to numerical data. The standard text mining approach is to create a variable for each word that occurs in any of the answer texts. These unigram variables or one-grams either record the frequency of the word occurring in an answer text or simply the presence or absence of the word from the given answer text (Weiss et al. 2010; Joachims 1998). There are many different variations of this text mining approach, adding variables for the presence or absence of multi-word sequences (ngram variables), removing highly used words (stopwords) because they are probably not useful, and stemming words to their grammatical root. The large number of variables are modeled with black-box statistical learning algorithms, such as support vector machines (*SVM*) (Vapnik 2000). The model may incorporate additional variables if available.

Different learning algorithms have been used for occupation coding. The Australian Bureau of Statistics (ABS) employed fully automatic categorization using support vector machines to code data from the 2006 Australian Census (Clarke and Brooker 2011). The ABS uses the Australian and New Zealand Standard Classification of Occupation (ANZSCO) scheme. To our knowledge this system is still in use by the ABS.

The American Community Survey (ACS) uses a variation on text mining (Thompson et al. 2012). Variables created from the text include one-word and two-word sequences (called "wordbits") as well as the full text. To limit the number of variables for analysis, a rareness threshold of 30 is used (i.e., the text has to occur at least 30 times before it is used as a variable). To further limit the number of variables for analysis, the corresponding text has to be "associated with a single industry/occupation code at least 50% of the time". The remaining variables, as well as variables like age and gender, are fed into a logistic regression. The code with the highest probability obtained by the logistic regression is assigned to a new record.

Some authors have investigated a nearest neighbor strategy, which assigns the code of the answer in the training data most closely resembling the answer in question. Different similarity metrics have been employed to measure nearness or resemblance between two answers. The PACE system employed the *k* nearest neighbor method with weighted feature metrics and reported accuracy 0.86 at production rate 0.57 for the U.S. Census Bureau data (Creecy et al. 1992). Jung et al. (2008) used cosine similarity but found this did not work well, possibly because they were working in Korean, a language quite different from languages with roots in Latin. Russ et al. (2014) used the nearest neighbor approach with a Jaccard similarity measure for classifying text answers into the Standard Occupational Classification (SOC) scheme. Coding by the nearest neighbour approach was considered correct if it agreed with one or both of the codes provided by the two human coders. The accuracy, that is, the proportion of correctly classified observations, for fully automated coding was 0.51 at the six-digit level and 0.64 at the three-digit level.

The ALWA survey at the German Institute for Employment Research (IAB) used the five-digit German national classification KldB 2010 (Schierholz 2014). The approach presented in Schierholz (2014) used the full preprocessed verbatim answer text rather than the text mining approach using ngram variables. Preprocessing included converting special German characters into regular ones, stripping leading and trailing spaces. Using verbatim answers (rather than ngrams) drastically reduced the number of variables for learning. Schierholz (2014) then experimented with various methods including Naive Bayes and a gradient boosting model (Friedman 2001). The experiment concluded that boosting and the Bayesian approaches performed similarly when high accuracy was desired.

3. Three Methods for Automated Occupation Coding

We first explain the duplicate method, a simple automated coding approach based on duplicate training observations. Next, we propose three new methods for automated occupation coding. The first of these methods, combining statistical learning models at different levels of aggregation, is later also incorporated with the second method, resulting in two versions of the second method. For statistical learning models, any method that outputs probabilities can be used. In Section 4, we choose Support Vector Machines (Vapnik 2000) for our application.

For each method, the predicted occupation code is the code that has the highest score.

3.1. The Duplicate Method With the Ngram-Based Definition of Duplicates

An exact-string duplicate refers to two strings that are identical. Simple string preprocessing could improve performance and leads to what we call a preprocessed-string duplicate. Preprocessing the string might consist, for example, of lower-casing allletters and removing leading and trailing blanks. For example "Apotheker" (pharmacist), "apotheker" and " apotheker" would be considered duplicates after preprocessing.

We introduce a different definition of duplicates based on ngram variables: an ngram duplicate refers to a training observation with a text answer that has the same ngram representation (i.e., the same values for the variables created from the text). This is slightly different than an observation with the identical text answer. For example, the answer "Verwaltungsangestellte im Krankenhaus" (administrator in the hospital) and "Verwaltungsangestellte in einem Krankenhaus" (administrator in a hospital) are not identical texts. However, since "in", "im" and "einem" are stopwords and stopwords are removed, these two strings contain the same unigrams ("Verwaltungsangestellte", "Krankenhaus").

Suppose that there exist some duplicates of a new input record $\mathbf{x}$. Let $m_i(\mathbf{x})$ be the number of training duplicates having code c_i $(i = 1,2,\ldots,L)$. We estimate the probability $p_d(c_i|\mathbf{x})$ based on the relative frequency of the training duplicates having code c_i:

$$\hat{p}_d(c_i|\mathbf{x}) = \begin{cases} \dfrac{m_i(\mathrm{x})}{M(\mathrm{x})} & \text{if } \mathrm{M(x)} > 0 \\ \dfrac{1}{L} & \text{otherwise} \end{cases},$$

where $M(\mathrm{x}) = \sum_{i=1}^{L} m_i(\mathbf{x})$ is the number of duplicates of $\mathbf{x}$ found in the training date. If no duplicate is found, the method assigns equal probability to each class. The code with the highest probability is chosen as the predicted code. The duplicate method leads to high accuracy for duplicates, although not to 100% accuracy, since coders try to resolve ambiguous situations with additional undocumented information or due to human error.

3.2. Combining Models from Different Levels of Aggregation

As seen in Table 1, occupation codes have a hierarchical structure. The ISCO-88 occupation codes consist of four-digit numbers. For example, the code 7131 (roofers) is part of the minor group 713 (Building finishers and related trades workers). Three-digit group codes aggregate related occupations. We propose to apply statistical learning separately to the four-digit unit occupation codes and to the three-digit group codes, and to combine probabilities as explained in the next paragraph. The motivation is as follows: Given the large number of occupation codes, the number of observations at the four-digit level can be sparse. The number of observations will be relatively less sparse at the three-digit level. If classification from a four-digit classifier results in a near tie of occupation codes with different minor groups (different third digit), the evidence from the three-digit classifier may sway the classification to the correct four-digit code.

Suppose that code c_i $(i = 1, \ldots, L)$ belongs to a three-digit minor group m_j $(j = 1, \ldots, l)$ where L and l are the numbers of the four-digit and three-digit group codes respectively. Denote the probabilities from the statistical learning model for three-digits and four-digits as $\hat{p}_{3digit}(m_j|\mathbf{x})$ and $\hat{p}_{4digit}(c_i|\mathbf{x})$ for a record $\mathbf{x}$, respectively. We average the two probabilities:

$$\hat{p}_{3/4digit}(c_i|\mathbf{x}) = \frac{\hat{p}_{3digit}(m_j|\mathbf{x}) + \hat{p}_{4digit}(c_i|\mathbf{x})}{2}. \tag{1}$$

This averaging approach will also break ties at the four-digit level, unless the tied codes have the same three-digit code. A recent review of hierarchical classification methods in general (Silla and Freitas 2011), does not contain the proposed method. However, the proposed method may be viewed as a member of the local-classifier-per-level approaches as it fits a classifier for each three-digit and four-digit level independently.

3.3. A Hybrid Approach: Combining Duplicate and Statistical Learning Approaches

The proposed hybrid approach combines the approach based on duplicates in the training data with a statistical learning approach.

Let $\hat{p}_s(c_i|x)$ be the estimated probability obtained by a statistical learning approach. For the hybrid approach we define a combined score $\theta(c_i|\mathbf{x})$ as

$$\theta(c_i|\mathbf{x}) = \frac{M(\mathbf{x})}{M(\mathbf{x}) + 1} \cdot \hat{p}_d(c_i|\mathbf{x}) + \frac{1}{M(\mathbf{x}) + 1} \cdot \hat{p}_s(c_i|\mathbf{x}) \tag{2}$$

If there are no duplicates, the score equals the probability from the statistical learning approach $\hat{p}_s(c_i|\mathbf{x})$. When there are duplicates, coding by the duplicate method is desirable, as it leads to high accuracy. Hence, in the hybrid approach the statistical learning algorithm only influences the prediction when there is a tie among different duplicate codes. Equation (2) assigns the statistical learner a weight equivalent to that of a single duplicate, and the single duplicate is downweighted by the probability $\hat{p}_s(c_i|\mathbf{x}) < 1$.

When the production rate is less than 100%, the easier-to-learn new records are categorized automatically. The statistical learning algorithms also influence this prioritization of new records. When two new records each have the same number of duplicates and if $\hat{p}_d(c_i|\mathbf{x})$ is the same in each case, the record with the larger $\hat{p}_s(c_i|\mathbf{x})$ is assigned a greater $\theta(c_i|\mathbf{x})$ and therefore is prioritized for lower production rates.

We call this approach "hybrid-4digit" when $p_s(c_i|\mathbf{x})$ in Equation (2) is estimated using the statistical learning model for four-digit occupation codes, $\hat{p}_{4digit}(c_i|\mathbf{x})$. Subsection 3.2 defined $\hat{p}_{3/4digit}(c_i|x)$ in Equation (1), which combined two statistical learning models from different levels of aggregation. This idea can also be applied here. We call this approach "hybrid-3/4digit" when $p_s(c_i|\mathbf{x})$ in Equation (2) is estimated using $\hat{p}_{3/4digit}(c_i|\mathbf{x})$.

3.4. A Modified Nearest Neighbor Approach

The nearest neighbour approach (*NN*) (Fix and Hodges 1951) is another method employed in the occupation coding. *NN* classification finds a new record's nearest neighbor in the training data and also assigns the occupation code of that nearest neighbor to the new record. There can be multiple nearest neighbors (Yu 2002). *NN* can be viewed as a generalization of the duplicate approach: duplicates are nearest neighbors with a distance of zero. To define "near", a measure of distance, or, equivalently, a measure of similarity is needed. For text classification, cosine similarity is widely used (Knaus 1987; Iezzi et al. 2014; Maitra and Ramler 2010). Cosine similarity between two vectors **u** and **v** is defined as

$$\text{cosine}(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{\sum u_i v_i}{\sqrt{\sum u_i^2}\sqrt{\sum v_i^2}}. \tag{3}$$

where **u** and **v** are vector representations of presence or absence of ngrams in the text. Similarity ranges from 0 to 1 depending on the degree of the similarity between two records. Similarity is 0 if two records have no common words and 1 if the two records are identical (in the sense of having the same ngram representation). When duplicates exist, the *NN* method predicts the code of records with similarity 1, which is equivalent to the duplicate method.

As before, we may want to only code easy-to-code text answers and leave difficult ones for manual coding. Hence, we propose to use a score that assigns a higher value to *NN* predictions that are believed to be more accurate. Given a new text input $\mathbf{x}$, denote $K(\mathbf{x})$ the number of nearest neighbors in the training data and $s(\mathbf{x})$ the similarity of the nearest neighbors. (Often $K(\mathbf{x}) > 1$ when multiple observations are the nearest neighbors.) Suppose that $k_i(\mathbf{x})$ out of the $K(\mathbf{x})$ records have the code c_i $(i = 1,\ldots,L)$. As in the duplicate method, we estimate the probability for code c_i in the *NN* approach by $\hat{p}_{nn}(c_i|\mathbf{x}) = k_i(\mathbf{x})/K(\mathbf{x})$. We define the score for the text answer as

$$\gamma(c_i|\mathbf{x}) = \hat{p}_{nn}(c_i|\mathbf{x})s(\mathbf{x})\left(\frac{K(\mathbf{x})}{K(\mathbf{x})+0.1}\right). \tag{4}$$

The predicted code depends only on $\hat{p}_{nn}(c_i|\mathbf{x})$ because $K(\mathbf{x})$ and $s(\mathbf{x})$ are constant for any given answer text. The role of $s(\mathbf{x})$ and $K(\mathbf{x})/(K(\mathbf{x}) + 0.1)$ is to order observations such that easier-to-classify-answers have a higher score.

The multiplier $s(\mathbf{x})$ makes sense: greater similarity of a new text and its nearest neighbor leads to more accurate classifications. The last term in Equation (4) can be motivated as follows: all else being equal, classification based on a larger number of nearest neighbors will likely be more accurate than that based on fewer nearest neighbors. The multiplier $K(\mathbf{x})/(K(\mathbf{x}) + 0.1)$ equals 0.91 when $K(\mathbf{x}) = 1$ and converges to 1 as $K(\mathbf{x})$ increases. Reflecting lesser importance, this multiplier can, at most, reduce the score by about ten percent, whereas both $\hat{p}_{nn}(c_i|\mathbf{x})$ and s can drive the score to zero. Below, we will show that this works empirically. However, we readily admit this is not the only multiplier that achieves this goal, and that the choice of 0.1 is arbitrary. Using a larger constant extends the range of the multiplier component and thus makes the score more sensitive to $K(\mathbf{x})$. (This is not desirable, as the other two multipliers are more important.)

Table 2. Illustration of calculating $\gamma(c_i|\mathbf{x})$. The unigram variables contain 1 if the word is present in the record and 0 otherwise.

	(Nonzero) ngram variables							
Record	heizung	lüftungsbauer	druck	Occ. Code	$\hat{p}_{nn}(c_i\|\mathbf{x})$	$s(\mathbf{x})$	$\frac{K(\mathbf{x})}{K(\mathbf{x})+0.1}$	$\gamma(c_i\|\mathbf{x})$
Training 1	0	0	1	8251	0.75	0.5774	0.9756	0.4225
Training 2	0	0	1					
Training 3	0	0	1					
Training 4	0	1	0	7136	0.25	0.5774	0.9756	0.1408
Test answer	1	1	1	$\hat{c}_i = 8251$				

For example, the text answer of a new record was "Heizungs und Lüftungsbauer, Drucker". The text consisted of three (stemmed) unigram variables: "heizung" (heating), "lüftungsbau" (ventilation construction) and "druck" (printer). No duplicates existed, but four records in the training data contained one of the three words. Table 2 shows that three out of the four training records had the answer "Drucker" ("druck" in the stemmed ngram representation) with code 8251 and the other had "Lüftungsbauer" ("lüftungsbau" in the stemmed ngram representation) with code 7136. Based on Equation (3), the similarity between the test answer and any of the training records in Table 2 was $\frac{1}{\sqrt{3}\sqrt{1}} = 0.5774$. So the multiplier in Equation (4) is $K(\mathbf{x})/(K(\mathbf{x}) + 0.1) = 4/4.1 = 0.9756$. However, $\hat{p}_{nn}(c_i = 8251|\mathbf{x}) = 3/4$ and $\hat{p}_{nn}(c_i = 7136|\mathbf{x}) = 1/4$. The difference of the γ scores of the two codes was dueto the different probability estimates. In this example, the test answer was assigned code 8251 because it had the largest score ($\gamma = 0.4225$).

4. Occupation Coding for the ALLBUS Survey

We first describe the ALLBUS data (Subsection 4.1) and then show the importance of our definition of duplicates (Subsection 4.2). Next, we compare the proposed automatic coding methods using the ALLBUS data (Subsections 4.3 and 4.4). We conclude with a simulation to explore the influence of duplicates and noise variables in Subsection 4.5.

4.1. Problem and Data

The German General Social Survey (ALLBUS) conducts repeated cross-sectional surveys of the adult German population living in private households, with an oversampling of the residents of East Germany. ALLBUS has been conducted every two years since 1980; initially covering West Germany and expanding to former East Germany after German reunification in 1990 (ALLBUS 2015; Koch and Wasmer 2004). The main topics concern attitudes, behavior, and social structure.

The targeted net sample size is usually 3,500. Since 1994, the samples have been drawn in two stages. In the first stage, about 160 communities (primary sampling units) are selected. In the second stage, addresses of individuals are randomly selected from thelists of residents in every community. Every two years, a fresh probability sample is drawn from the German register. ALLBUS surveys are conducted face-to-face.

ALLBUS interviewers asked about occupation multiple times: current occupation of respondent, last occupation of respondent (if not employed), occupation of spouse (if married), occupation of partner (if not married but with partner), occupation of father, and occupation of mother. In the ALLBUS survey, the interviewer asks the following questions which are recommended by official statistics in Germany (Statistisches Bundesamt 2010): “What work do you do in your main job? Please describe your work precisely. Does this job, this work have a special name?” (Scholz and Wasmer 2009). Interviewers were free to combine the answers, and were not asked to write one answer after another. The occupation questions for partners/spouses/parents are analogous, using the same format. The answers were pooled to form a single data set. Prior to the open-ended questions about all occupations, respondents were also asked: “Please classify your occupational status according to this list.” The list contains 32 occupation statuses in twelve categories. We refer to this below as (self-recorded) occupation status.

The ISCO-88 coding of the text answers was done by GESIS in a two-step procedure. First, automatic coding was attempted using the in-house software, *textpack* (Geis and Hoffmeyer-Zlotnik 2000; Züll 2014). Then, such automatically coded answers were verified by a professional coder. All remaining responses were manually coded in a second step according to an extensive coding manual (Geis 2011). The in-house software used a dictionary with about 4,500 predefined combinations of ISCO codes. Because the dictionary mostly contains duplicates from previous surveys, *textpack* implements the duplicate approach, with additional hand-crafted rules (however, the coder may also override some codes in light of occupational status, education, or other information).

For each word or phrase listed in the dictionary, *textpack* searches for exact matches in the data and outputs the associated code. Such rules were applied one at a time (and the rule order may affect the result). If a rule was matched exactly, a response was coded. If none of the rules applied, it was manually coded by professional coders. Typically, *textpack* coded about 50% of the responses. GESIS used self-reported occupation status only if text was unclear or ambiguous. In the 2006 survey, 9,137 observations were coded into 399 distinct unit occupation codes and 140 minor group codes (see appendix A).

To apply the proposed methods, we encoded text answers into unigram variables (Schonlau and Guenther 2016). All such variables were indicator variables specifying the presence or absence of the corresponding word. We applied stemming, using a German Porter stemmer (Snowball 2015) and removed German “stopwords” as well as punctuation marks. The removal of stopwords and the use of stemming reduced the number of ngram variables. As is standard practice, we also created a variable that counted the number of words contained in the answer. All in all, 4,232 indicator variables were created in addition to the number-of-words variable. In addition to the text response, the survey also contains self-reported occupation status, which was also included among the independent variables.

For a statistical learning approach, we use support vector machines (*SVM*) (Vapnik 2000) with a linear kernel, which has been shown to work well in text categorization (Joachims 1998). The linear kernel requires only a single tuning parameter, C, that controls the trade-off between the training error and model complexity. In this data set, the choice of C had little influence on prediction accuracy and we used $C = 1$ throughout the study. As is common, the *SVM* scores were converted into probabilities using Platt’s method (Platt 1999), which performs a regularized logistic regression of class membership on the *SVM* score.

We evaluate the approaches using ten-fold cross validation (*CV*). This means we randomly divide the data into ten equal-sized parts. We use the first nine parts to train the model, and the last part to test the model. Accuracy is only evaluated on the test data. In turn, we use each of the ten parts as test data and average the results. As a consequence, the size of the training data is therefore 90% of the data, or 8,223 observations. For the purpose of evaluating prediction accuracy we assume that the original codes assigned by GESIS and the professional coders are correct.

The analysis was carried out in *R* (R Core Team 2014), and package *e*1071 (Meyer et al. 2014) is used for the construction of the *SVM* models.

Most open-ended answers were short; 66.5% of the answers consisted of a single word. The median length was one word; the average length was 1.8 words and the maximum length was 17 words. About 60% of the data consisted of (ngram-based) duplicate observations. Among duplicate observations, the median number of duplicates was three, with a higher average (6.8) due to some very frequent duplicates (maximum = 221 duplicates). The text with the most duplicates was “Landwirt” (farmer).

4.2. Ngram Vs. String-Based Definition of Duplicates

The purpose of this section is to demonstrate that the ngram-based method of duplicate is preferable to the string-based methods. Here we explore how much the definition of duplicate mattered for the two best performing methods, NN-3 and hybrid-3/4digit, which are explained later. We compared the ngram-based method with original string (without any processing) and preprocessed string methods. Preprocessed strings refer to lower casing and stripping off leading and trailing spaces in the original strings. As described in Subsection 4.1, ngram variables were obtained after stemming, and removing stopwords and punctuation marks.

The percentage of duplicates is 52.6% for the identical-string-duplicates, 56.7% for the preprocessed-string-duplicates, and 60.0% for the ngram-duplicates. However, the quality of the duplicates did not degrade: identical-string-duplicates (preprocessed-string-duplicates, ngram-duplicates) had identical occupation codes 91.9% (91.6%, 92.0%) of the time. The remaining eight percent represent coders’ attempt to recode otherwise unambiguous text in light of occupational status or education. For example, a pharmacist with lower occupational status might be reclassified as pharmaceutical assistant. Of course, misclassification errors are also possible.

Figure 1 shows the trade-off between accuracy and production rate for the three definitions of duplicates for hybrid-3/4digit (left panel) and NN-3 (right panel). The use of the ngram definition of duplicates improved accuracy in both methods for moderate and high production rates. With full automation, accuracy increased from 0.54 (without preprocessed) to 0.65 for the hybrid-3/4digit method, and from 0.47 (without preprocessed) to 0.65 for the NN-3 method. Preprocessed-string-duplicates fare somewhat better than unprocessed strings, but the success of the ngram-based definition clearly goes far beyond string preprocessing.

4.3. Accuracy of the Nearest Neighbor Method

We first investigated the coding performance of the modified *NN* method. The score in Equation (4) has three components. To demonstrate that all three components are helpful, we evaluate both the proposed overall score (NN-3) as well as a reduced score missing one

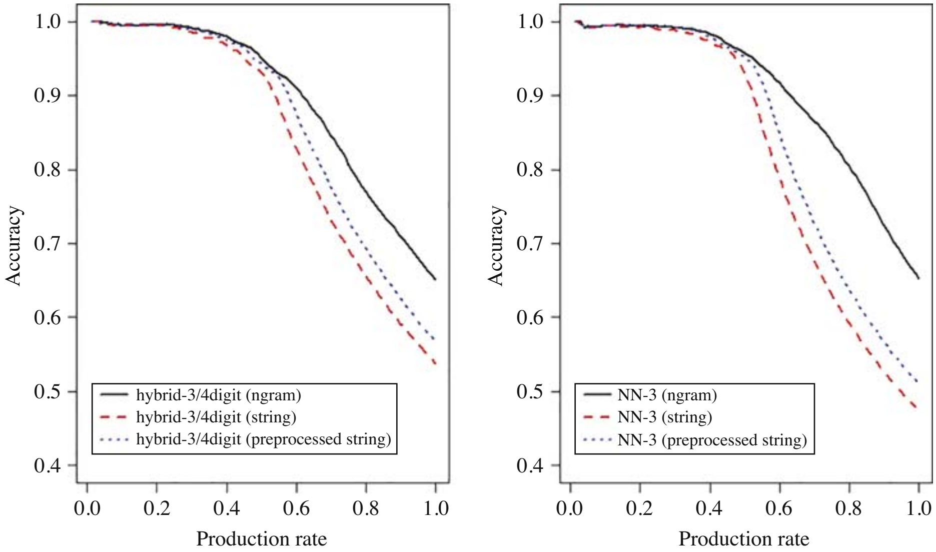


Fig. 1. Accuracy for a given production rate for two approaches based on three different definitions of duplicates "ngram", "string" and "preprocessed string". The left panel shows the results of hybrid-3/4digit and the right panel shows those of NN-3. The "ngram" definition of duplicates is far superior.

(NN-2) or two components (NN-1) with corresponding scores γ_1,γ_2 and γ_3:

$$\text{(NN-1)} \quad \gamma_1 \quad = \max_i \hat{p}_{nn}(c_i|\mathbf{x})$$

$$\text{(NN-2)} \quad \gamma_2 \quad = \max_i \hat{p}_{nn}(c_i|\mathbf{x})\, s(\mathbf{x})$$

$$\text{(NN-3)} \quad \gamma_3 \quad = \max_i \hat{p}_{nn}(c_i|\mathbf{x})\, s(\mathbf{x}) \left(\frac{K(\mathbf{x})}{K(\mathbf{x}) + 0.1}\right)$$

Figure 2 shows the accuracies of each approach as a function of the production rate. (These were average accuracies from the ten-fold cross validation mentioned earlier). Answer texts with higher scores were coded first; a production rate of, say, ten percent refers to coding ten percent of the answer texts with the highest scores automatically. When the production rate equals 100%, the accuracy is the same for all the approaches because the second and third terms in Equation (4) do not affect which code is assigned, but rather are used to prioritize more similar observations and observations with multiple nearest neighbors by assigning them a higher score. Prioritizing affects the accuracy at production rates of less than 100% (because observations with the highest score are chosen first). The improvement from NN-1 to NN-2 showed that similarity s was helpful for finding easier-to-classify-answers. Likewise, the accuracy differences between NN-2 and NN-3 showed that the term $\frac{K(\mathbf{x})}{K(\mathbf{x})+0.1}$ improved the performance at low to medium production rates.

Having established that NN-3 is preferable to NN-1 and NN-2, we next compare NN-3 with all other approaches.

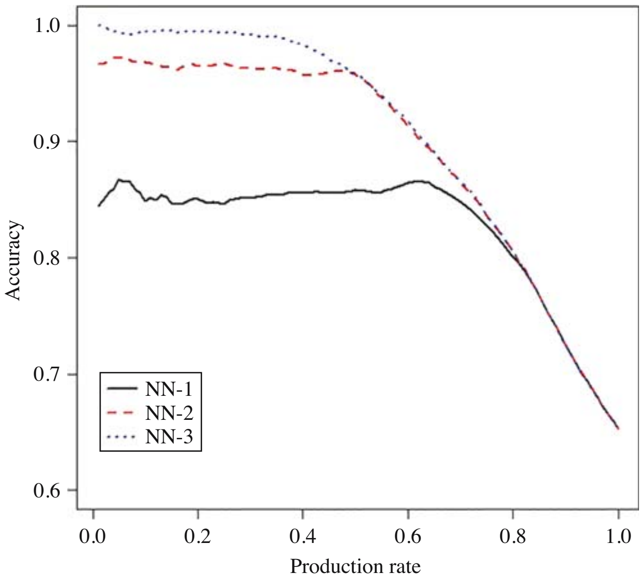


Fig. 2. Accuracy of three variations on the nearest neighbor approach as a function of production rates. NN-1, NN-2, and NN-3 refer to scores using $\gamma_1 = \hat{p}_{nn}(c_i|\mathbf{x})$, $\gamma_2 = \hat{p}_{nn}(c_i|\mathbf{x})s$ and $\gamma_3 = \hat{p}_{nn}(c_i|\mathbf{x})s\left(\frac{K(\mathbf{x})}{K(\mathbf{x})+0.1}\right)$, respectively.

4.4. Comparison of Methods

Here we compare the accuracy as a function of production rate for the proposed methods (hybrid-4digit, hybrid-3/4digit, and NN-3) as well as some default methods (duplicate method, svm-4digit, svm-3/4digit). The duplicate method refers to assigning the code of ngram duplicates (or a random code if no duplicates exist), svm-4digit refers to an *SVM* model based on four-digit occupation codes. The svm-3/4 digit refers to an *SVM* model based on averaged probability from separate models for three-digit and four-digit occupation codes as described in Equation (1). For all methods, a production rate of x% refers to the x% of the data that have the highest score (or probability).

Figure 3 shows the accuracy as a function of the production rate for the different methods. For all methods, there were trade-offs between the accuracy and the production rate. The modified nearest neighbor method, NN-3, performs equal to or slightly better than the next best method, hybrid-3/4digit. NN-3, hybrid-4digit, and hybrid-3/4digit uniformly beat the duplicate method and both svm methods.

A production rate of 100% corresponds to classifying all answers automatically. At full automation, NN-3 and hybrid-3/4digit perform equally well. At full automation, svm-3/4digit has an accuracy of 59%, the duplicate method has an accuracy of 53%, and the hybrid-3/4digit method increases the accuracy to 65%.

Figure 3 also shows the duplicate accuracy remained at around 95% up to a production rate of about 0.55. About 55% of the test data in any given cross-validation were duplicates and thus duplicates were used for coding. However, when no duplicates exist in the training data, the duplicate approach assigned equal probabilities to all codes, resulting in the random code assignment and accuracy near zero. The accuracy started decreasing at a production rate of around 0.55, from which no additional records of some *CV* test samples

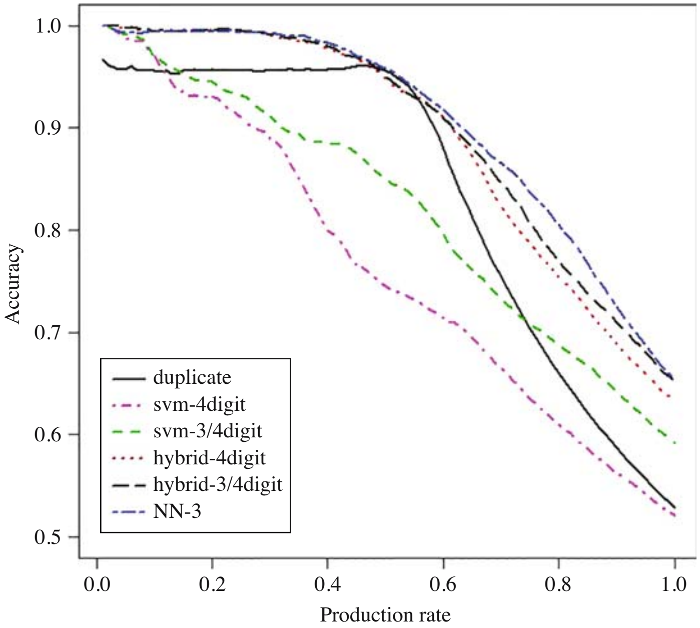


Fig. 3. Comparison of different methods for occupation coding. Methods include statistical learning (svm-4digit), statistical learning from two models at different levels of aggregation (svm-3/4digit), and two hybrid methods combining duplicate-predictions with svm-4digit and svm-3/4digit, respectively.

could be classified by the method. From a production rate of 0.60, all of the *CV* test data sets had no duplicates and the method performed poorly. NN-3, hybrid-4digit, and hybrid-3/4digit beat the duplicate method even for production ranges where duplicates are available.

Combining the four-digit unit and three-digit minor code methods (svm-3/4digit) was uniformly superior to using the unit code method only (svm-4digit). For example, for fully automated coding, the accuracy for svm-3/4digit was 0.59, as compared with 0.52 for svm-4digit. The hybrid approaches performed very similarly up to a production rate of about 60%. After that, the hybrid-3/4digit performs a little better than hybrid-4digit. When duplicates were available for hybrid-3/4digit, the predicted codes mostly agreed (83%) with those predicted by the duplicate method.

The performances of hybrid-3/4digit and the NN-3 were similar for fully-automated coding as well as at low-medium production rates. NN-3 appeared to slightly outperform hybrid-3/4digit at medium-high production rates.

The curves in Figure 3 help us decide which texts should be classified automatically and which should be classified manually. For example, if the client decides that 80% accuracy is required, then Figure 3 suggests that 76% of the data can be classified automatically with the hybrid method and 81% with the NN-3 method. Relative to applying the duplicate-based approach, this increases production from about 58% to 76% or 81%.

4.5. Simulation

The purpose of this section is to explore to what extent the methods are robust to possible idiosyncrasies of the data. We considered two possible concerns with our example data:

1) The data contain a large percentage (50%) of duplicates. 2) The text answers are unusually clean and contain fewer superfluous words than usual.

In the first case, in the context of occupation coding a large number of duplicates is very common. (Duplicates here refers to ngram duplicates). To simulate a data set with fewer duplicates, a random subset of duplicate records was removed so that in the reduced data only about ten percent duplicates of the test records had duplicates. The reduced data set contained 4,722 observations.

As expected, Figure 4 shows that the accuracy (for a given production rate) for all methods decreased for this much more difficult problem. The relative performance of the methods is very similar with one notable exception: previously, both NN-3 and hybrid3/4-digit performed similarly. Now, NN-3 clearly outperforms the hybrid-3/4digit method. The NN-3 method remains superior to NN-1 and NN-2 analogous to Figure 2 (The analogous figure is not shown).

In the second case, less clean text answers would have resulted in additional words that are not related to the occupation code. Such additional words translate into indicator variables (presence or absence of the word) in the data. There are typically many such variables, each with a low probability. We added 100 independent "noise" indicator variables to the data. Each variable followed a Bernoulli distribution with an 0.01 probability of success.

The results are shown in Figure 5. Adding the noise variables decreased the number of duplicates. Hence the accuracy of the duplicate method started decreasing at a production rate of around 0.2 instead of around 0.55. The results lead to roughly the same conclusions as we obtained from Figures 3 and 4. NN-3 and hybrid-3/4digit were comparable, with NN-3 having a slight edge at lower production rates.

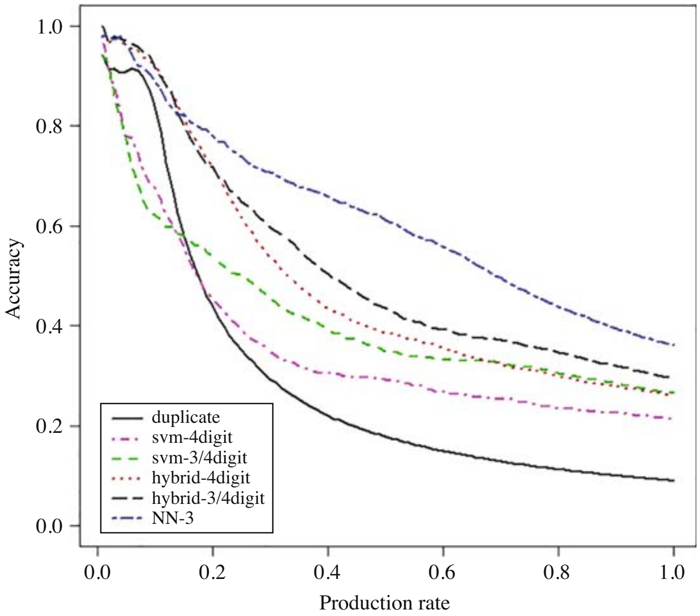


Fig. 4. Comparison of the same methods as in Figure 3 on a reduced data set containing only ten percent duplicates.

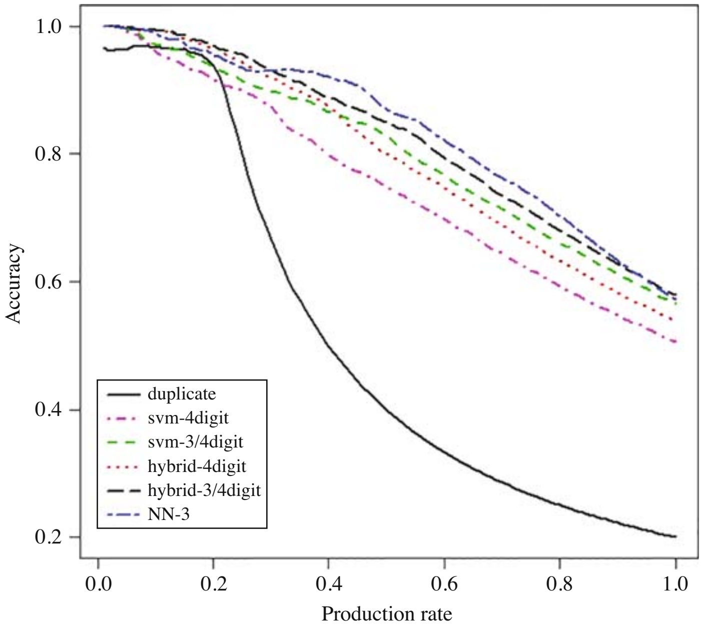


Fig. 5. Comparison of the same methods as in Figure 3 with 100 noise variables added to the data.

5. Discussion

We have investigated several novel approaches for automated occupation coding for any desired production rate. The two best-performing methods, the modified nearest neighbor method (NN-3) and a hybrid method (hybrid-3/4digit) substantially improve the accuracy compared with both statistical learning (*SVM* in the example) by itself and the duplicate method at any production rate in the ALLBUS data. As the percentage of duplicates decreases, a simulation shows that NN-3 gains a relative advantage over the hybrid method.

Either accuracy or production rate can be set at a target rate which determines the second measure. For example, targeting 80% accuracy for the automated coding, the hybrid-3/4digit and NN-3 approaches could categorize 76% and 81% of the data automatically, while the numbers obtained by the *SVM* and duplicate methods individually were 60% and 66%, respectively. If production rate is fixed at 80%, the hybrid-3/4digit and NN-3 could achieve an accuracy of 77% and 81%, while the *SVM* and duplicate approaches reported accuracy of 69% and 66%. Note that accuracy for each category may differ from the overall accuracy. Categories that contain more hard-to-code answers than others achieve lower accuracies.

In addition, we have learned:

(1) Even at low production rates when duplicates exist, NN-3 and hybrid achieve a higher accuracy than the duplicate method.

(2) Using the duplicate method where duplicates exist and using statistical learning otherwise is not the best strategy (Figure 3 shows the proposed methods beat the duplicate method where duplicates exist.). We instead recommend the hybrid method that integrates the two approaches.

(3) Combining aggregate and detailed learners improves accuracy for some learning algorithms. For example, where svm-4digit and svm-3/4digit disagree in the ALLBUS data, svm-3/4digit is correct 87% of the time.

Why do the NN-3 and hybrid methods beat *SVM* and the duplicate approach? Because a duplicate is also a nearest neighbor, both methods rely on nearest neighbors. Nearest neighbor algorithms are effective when prediction is highly local and little can be gained from observations further away. This may explain why NN-3 and hybrid methods beat *SVM*, one of best statistical learning algorithms in existence. Both proposed methods beat the duplicate approach because a) they both can distinguish between easier-to-code and harder-to-code duplicates leading to higher accuracies at lower production rates, b) the hybrid- 3/4 method can break ties among duplicates, and c) the duplicate approach performs poorly when no duplicates exist.

The NN-3 approach can be computationally expensive when the training data set is very large. The hybrid method requires finding duplicates, but on the other hand, finding duplicates is much less expensive because it does not require a sorting step.

We have combined the aggregate method with the hybrid method, leading to better results. The modified nearest neighbor method could also be combined with the idea of aggregating different level scores. However, the resulting method showed almost the same performance as NN-3.

We now comment on the importance of some data analysis choices. First, duplicates were defined as having the same ngram representation rather than being identical strings. This increased the number of duplicates and substantially improved accuracy at moderate and high production levels. Second, self-reported occupation status (STIB) was used as a covariate for statistical learning. We found that including STIB made little difference. Third, we supported German language stemming, but it turned out this had almost no effect. Because the text was written by interviewers (rather than respondents) our data were relatively clean with many one-word answers. Stemming is likely more important with messier data.

We next comment on possible limitations arising from idiosyncrasies of the ALLBUS data set. The proposed methods are not limited to the ISCO-88 coding scheme. One of the methods relies on a hierarchical coding scheme, but all occupation codes are hierarchical. We have analysed 9,137 observations. While this data set is probably larger than most data sets analysed in statistical journals, at national statistics agencies far larger data sets arise sometimes with millions of observations. The proposed methodology is not limited to a specific data size, but it is unclear whether the performance of the proposed methodology relative to the alternative algorithms would be equally impressive with millions of observations. We have pooled self-recorded occupations and occupations from partners, spouses, and parents. We investigated whether this distorted results somehow. Specifically, we reduced the data set to one occupation question per respondent. We found this did not meaningfully affect the results.

For the hybrid method, we used *SVM* as the statistical learning method of choice. While *SVM* is one of best performing methods available, other statistical learning methods could be chosen, provided that they output a probability (or a score that can be transformed into a pseudo-probability) rather than just a classification. Naturally, better predictions from the statistical learning method will tend to improve the hybrid method also, particularly when there are no duplicates.

All proposed approaches rely on training data. For statistical learning, the size of the training data needs to be large relative to the number of occupation codes. In the ALLBUS data, the size of the training data (implied by cross-validation) was 8,226. Relative to the 399 occupation codes, this is an average of 20.6 observations per code. More training data will tend to increase the number of duplicates.

Cross-validation deals with unseen data, but does not take into account time trends. To the extent that language use changes from year to year, any classifier would slowly degrade over time.

In summary, we proposed new approaches to automated occupation coding that lead to vastly improved coding accuracy at both high and low production rates in our example data. While not conclusive, this bodes well for other occupation data sets.

Appendix A

There are more distinct codes in the GESIS data than the 390 ISCO-88 unit codes for several reasons: 1) When there is sufficient information to identify a minor group, but not sufficient information to identify a unit code, the minor code is used and a zero is appended (e.g., minor group 112 would turn into 1120). 2) Sometimes a minor group can be identified and the text is specific enough to identify the exact occupation, but that occupation is not listed. In that case a separate code is used ending in a nine (e.g., 1129 in the previous example) 3) ISCO-88 allows users to define additional codes for occupations that are not explicitly mentioned. GESIS has defined 10 such codes (e.g., housewife, not codable, don't know). The total of possible GESIS codes is 641 (390 unit codes ± 116 minor groups ± 28 sub-major groups ± 10 major groups ± 10 GESIS specific codes ± 87 codes for occupations not elsewhere classified). In the ALLBUS 2006 survey 399 of the 641 distinct codes were observed.

6. References

ALLBUS. 2015. Available at: http://www.gesis.org/allbus (accessed October 10, 2016).

Appel, M.V. and E. Hellerman. 1983. "Census Bureau Experiments with Automated Industry and Occupation Coding." In Proceedings of the American Statistical Association, Section on Survey Research Methods. August 15–18, 1983, Toronto, Canada. 32–40.

Belloni, M., A. Brugiavini, E. Meschi, and K. Tijdens. 2014. *Measurement Error in Occupational Coding: an Analysis on SHARE Data*. Ca' Foscari University of Venice, Department of Economics, Working Paper 24. Doi: http://dx.doi.org/10.2139/ssrn.2539080.

Bethmann, A., M. Schierholz, K. Wenzig, and M. Zielonka. 2014. "Automatic Coding of Occupations." In Proceedings of Statistics Canada Symposium. August 29–31, 2014, Québec, Canada. Available at: http://www.statcan.gc.ca/sites/default/files/media/14291-eng.pdf (accessed October 10, 2016).

Chen, B.-C., R.H. Creecy, and M.V. Appel. 1993. "Error Control of Automated Industry and Occupation Coding." *Journal of Official Statistics* 9: 729–745. http://www.jos.nu/Articles/abstract.asp?article=94729 (accessed October 10, 2016).

Clarke, F.R. and S.J. Brooker. 2011. Use of Machine Learning for Automated Survey Coding. In Proceedings of the 58th ISI World Statistics Congress. August 21–26, 2011, Dublin, Ireland.

Conrad, F.G., M.P. Couper, and J.W. Sakshaug. 2016. "Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes." *Journal of Official Statistics* 32: 75–92. Doi: http://dx.doi.org/10.1515/JOS-2016-0003.

Creecy, R.H., B.M. Masand, S.J. Smith, and D.L. Waltz. 1992. "Trading MIPS and Memory for Knowledge Engineering." *Communications of the ACM* 35: 48–64. Doi: http://dx.doi.org/10.1145/135226.135228.

Day, J. 2014. *Using an Autocoder to Code Industry and Occupation in the American Community Survey*. Presentation for the Federal Economic Statistics Advisory Committee Meeting. Available at: http://www2.census.gov/adrm/fesac/2014-06-13_day.pdf (accessed October 10, 2016).

Elias, P. 1997. "Occupational Classification (ISCO-88): Concepts, Methods, Reliability, Validity and Cross-National Comparability." OECD Labour Market and Social Policy Occasional Papers 20, OECD Publishing. Available at: https://ideas.repec.org/p/oec/elsaaa/20-en.html (accessed October 10, 2016).

Elias, P. and M. Birch. 2010. *Tuning CASCOT for Industry and Occupation Coding in the Scottish Census of Population 2011*. Technical Report, Institute for Employment Research. Coventry: University of Warwick.

Ferrillo, A., S. Macchia, and P. Vicari. 2008. "Different Quality Tests on the Automatic Coding Procedure for the Economic Activities Descriptions." In Proceedings of the European Conference on Quality in Official Statistics – Q2008. July 8–11, 2008, Rome, Italy. Available at: http://q2008.istat.it/sessions/paper/15Ferrillo.pdf (accessed January 2017).

Fix, E. and J.L. Hodges. 1951. *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. Technical Report, USAF School of Aviation Medivine, Randolph Field, Texas. Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951.

Friedman, J.H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29: 1189–1232. Available at: http://www.jstor.org/stable/2699986 (accessed October 10, 2016).

Ganzeboom, Harry B.G. and Donald J. Treiman. 2003. "Three Internationally Standardised Measures for Comparative Research on Occupational Status." In *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, edited by J.H.P. Hoffmeyer-Zlotnik and C. Wolf, pp. 159–193. Doi: http://dx.doi.org/10.1007/978-1-4419-9186-7_9.

Geis, A. 2011. *Handbuch für die Berufsvercodung*. Technical Report, GESIS, Mannheim, Germany. Available at: http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/handbuch_der_berufscodierung_110304.pdf (accessed October 10, 2016).

Geis, A.J. and J.H.P. Hoffmeyer-Zlotnik. 2000. “Stand der Berufsvercodung.” *ZUMA Nachrichten* 24: 103–128.

Iezzi, D.F., M. Lori, F. Lorenzini, M. Nicosia, and S. Stoppiello. 2014. “An Application of Text Mining Technique for the Census of Nonprofit Institutions.” In *Statistical Methods and Applications from a Historical Perspective*, edited by F. Crescenzi and S. Mignani, pp. 143–152. Springer. Doi: http://dx.doi.org/10.1007/978-3-319-05552-7_13.

International Labour Office. 1990. International Standard Classification of Occupations, ISCO-88. International Labour Office. Available at: http://www.ilo.org/public/libdoc/ilo/1990/90B09_411_engl.pdf (accessed October 10, 2016).

Joachims, T. 1998. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” In Proceedings of the 10th European Conference on Machine Learning, Volume 1398. April 21–23, 1998, Chemnitz, Germany, 137–142. Doi: http://dx.doi.org/10.1007/BFb0026683.

Jones, R. and P. Elias. 2004. *CASCOT: Computer-Assisted Structured Coding Tool*. Technical Report, Institute for Employment Research. Coventry: University of Warwick. Available at: http://www2.warwick.ac.uk/fac/soc/ier/publications/software/cascot/ (accessed October 10, 2016).

Jung, Y., J. Yoo, S.-H. Myaeng, and D.-C. Han. 2008. “A Web-Based Automated System for Industry and Occupation Coding.” In *Web Information Systems Engineering - WISE 2008*, edited by J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. Wang. Volume 5175, 443–457. Springer. Doi: http://dx.doi.org/10.1007/978-3-540-85481-4_33.

Kalpic, D. 1994. “Automated Coding of Census Data.” *Journal of Official Statistics* 10: 449–463.

Knaus, R. 1987. “Methods and Problems in Coding Natural Language Survey Data.” *Journal of Official Statistics* 3: 45–67.

Koch, A. and M. Wasmer. 2004. “Der ALLBUS als Instrument zur Untersuchung sozialen Wandels: Eine Zwischenbilanz nach 20 Jahren.” In *Sozialer und Politischer Wandel in Deutschland*, edited by R. Schmitt-Beck, M. Wasmer, and A. Koch, 13–41. VS Verlag für Sozialwissenschaften.

Maitra, R. and I.P. Ramler. 2010. “A k-mean-directions Algorithm for Fast Clustering of Data on the Sphere.” *Journal of Computational and Graphical Statistics* 19: 377–396. Doi: http://dx.doi.org/10.1198/jcgs.2009.08155.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2014. *e1071: Misc Functions of the Department of Statistics, TU Wien*. Available at: http://CRAN.R-project.org/package=e1071 (accessed October 10, 2016).

O’Reagan, R.T. 1972. “Computer-Assigned Codes from Verbal Responses.” *Communications of the ACM* 15: 455–459. Doi: http://dx.doi.org/10.1145/361405.361419.

Ossiander, E.M. and S. Milham. 2006. “A Computer System for Coding Occupation.” *American Journal of Industrial Medicine* 49: 854–857. Doi: http://dx.doi.org/10.1002/ajim.20355.

Platt, J. 1999. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.” In *Advances in Large Margin Classifiers*, edited by A.J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, 61–74. Cambridge, Massachusetts: MIT Press.

R Core Team. 2014. "R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing." Available at: http://www.R-project.org/ (accessed October 10, 2016).

Russ, D.E., K.-Y. Ho, C.A. Johnson, and M.C. Friesen. 2014. "Computer-Based Coding of Occupation Codes for Epidemiological Analyses." In Proceedings of the 27th IEEE International Symposium on Computer-Based Medical Systems. May 27–29, 2014, New York, USA, 347–350. Doi: http://dx.doi.org/10.1109/CBMS.2014.79.

Schierholz, M. 2014. "Automating Survey Coding for Occupation." Master's thesis, Ludwig-Maximilians-Universität Munich. Available at: https://epub.ub.uni-muenchen.de/21444/index.html (accessed October 10, 2016).

Scholtus, S., R. van de Laar, and L. Willenborg. 2014. *The Memobust Handbook on Methodology for Modern Business Statistics*. Available at: https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_246.pdf (accessed January 2017).

Scholz, E., and M. Wasmer. 2009. *German General Social Survey 2006. English Translation of the German "ALLBUS"- Questionnaire*. Technical Report, GESIS, Mannheim, Germany. Available at: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-207035 (accessed October 10, 2016).

Schonlau, M., and N. Guenther. 2016. Text Mining Using N-Grams. *Social Science Research Network*. Doi: http://dx.doi.org/10.2139/ssrn.2759033.

Silla, C.N., and A.A. Freitas. 2011. "A Survey of Hierarchical Classification across Different Application Domains." *Data Mining and Knowledge Discovery* 22: 31–72. Doi: http://dx.doi.org/10.1007/s10618-010-0175-9.

Snowball. 2015. Available at: http://snowball.tartarus.org/algorithms/german/stemmer.html (accessed October 10, 2016).

Statistisches Bundesamt. 2010. *Demographische Standards*. Technical Report, Wiesbaden, Germany. Available at: https://www.destatis.de/DE/Methoden/StatistikWissenschaft-Band17.html (accessed October 10, 2016).

Thompson, M., M.E. Kornbau, and J. Vesely. 2012. "Creating an Automated Industry and Occupation Coding Process for the American Community Survey." Available at: http://ftp.census.gov/adrm/fesac/2014-06-13_thompson_kornbau_vesely.pdf (accessed October 10, 2016).

Tijdens, K. 2014. "Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey." *Journal of Official Statistics* 30: 23–43. Doi: http://dx.doi.org/10.2478/jos-2014-0002.

Tijdens, K. 2015. "Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-Up Tables." *Survey Methods: Insights from the Field (SMIF)*. Doi: http://dx.doi.org/10.13094/SMIF-2015-00008.

Tourigny, J.Y., and J. Moloney. 1995. "The 1991 Canadian Census of Population Experience with Automated Coding." In United Nations Statistical Commission on Statistical Data Editing.

Vapnik, V.N. 2000. *The Nature of Statistical Learning Theory*. 2nd edition. New York: Springer.

Weiss, S.M., N. Indurkhya, T. Zhang, and F. Damerau. 2010. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Wenzowski, M.J. 1988. “ACTR – A Generalised Automated Coding System.” *Survey Methodology* 14: 299–308.

Yu, C. 2002. *High-Dimensional Indexing: Transformational Approaches to High-Dimensional Range and Similarity Searches*. Volume 2341. Berlin: Springer. Doi: http://dx.doi.org/10.1007/3-540-45770-4.

Züll, C. 2014. *Berufscodierung*. Technical Report, GESIS – Leibniz Institut für Sozialwissenschaften (SDM Survey Guidelines). Mannheim. Doi: http://dx.doi.org/10.15465/sdm-sg_019.

Received March 2016
Revised October 2016
Accepted October 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 123–136, http://dx.doi.org/10.1515/JOS-2017-0007

Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions

Olena Kaminska[1] *and Peter Lynn*[1]

In multi-national surveys, different countries usually implement different sample designs. The sample designs affect the variance of estimates of differences between countries. When making such estimates, analysts often fail to take sufficient account of sample design. This failure occurs sometimes because variables indicating stratification, clustering, or weighting are unavailable, partially available, or in a form that is unsuitable for cross-national analysis. In this article, we demonstrate how complex sample design should be taken into account when estimating differences between countries, and we provide practical guidance to analysts and to data producers on how to deal with partial or inappropriately-coded sample design indicator variables. Using EU-SILC as a case study, we evaluate the inverse misspecification effect (*imeff*) that results from ignoring clustering or stratification, or both in a between-country comparison where countries' sample designs differ. We present *imeff* for estimates of between-country differences in a number of demographic and economic variables for 19 European Union Member States. We assess the magnitude of *imeff* and the associated impact on standard error estimates. Our empirical findings illustrate that it is important for data producers to supply appropriate sample design indicators and for analysts to use them.

Key words: Cross-national studies; imeff; multiple frame design; complex sample estimation.

1. Introduction

There are many examples of multi-country surveys that are designed specifically for the purpose of cross-national comparisons (Lynn et al. 2006; Smith 2010), though the challenges that must be met in order to provide useful comparability are considerable (Kish 1994, 1999). In order to provide a basis for unbiased estimation of between-country differences, such surveys apply a standard definition of the target population (Heeringa and O'Muircheartaigh 2010) and select a probability sample from that population (Häder and Gabler 2003; Lynn et al. 2007). As well as enabling unbiased estimation, cross-national surveys sometimes also aim to standardize the precision of estimates within each

[1] Institute for Social and Economic Research, University of Essex, Colchester, Essex CO4 3SQ, United Kingdom of Great Britain and Northern Ireland. Email: olena@essex.ac.uk and plynn@essex.ac.uk

Acknowledgments: This research was carried out under the award Analysis of Life Chances in Europe (ALICE), funded by the UK Economic and Social Research Council. The ALICE Principal Investigators were Richard Berthoud and Maria Iacovou. EU-SILC data was supplied by Eurostat. We are grateful to Vijay Verma for advice, to Alexandra Skew and Francesco Figari for helping to document and interpret the data, and to Alita Nandi and Steve Pudney for advice on Stata routines. The empirical study of mis-specification effects (sections 3 and 4 of this article) was presented at the Joint Statistical Meetings in Vancouver, Canada, in August 2010 and at the World Statistical Congress in Dublin, Ireland, in August 2011. Additionally, our thoughts on preparing sample design indicators for cross-national analysis (section 2) were presented at the Comparative Survey Design and Implementation (CSDI) workshop in London, UK, in March 2011.

country (European Commission 2013). One way to achieve this is to select one specific important statistic and develop a sample in each country such that it leads to a defined precision for the estimate of that statistic (European Commission 2013). This defined precision has to be common across countries. For a multi-purpose survey, a more appropriate method is to set a common effective sample size (Lynn et al. 2007; Gabler et al. 2006). Effective sample size indicates how many cases a simple random sample would need in order to have the same precision as a particular (complex) sample design.

These requirements for a standard population definition, a probability sample, and a required precision leave scope for sample designs to vary across countries. Kish (1989, 41) mentions ". . . the selection methods and the sample designs of the surveys whose results are compared need not be at all similar. If they are based on good probability methods, the sampling method for each can be entirely distinct. Actually, for each sample we should utilize whatever selection method is most appropriate, feasible, and efficient . . .". If countries implement the most efficient and appropriate sample design, then differences in geography, population distribution, available sampling frames, and survey systems make it inevitable that countries will vary in whether and how they use stratified sampling, clustering, and unequal selection probabilities.

Differences in sample designs need not be a problem for estimation, but appropriate estimation requires the existence of appropriate indicators of components of the sample design. Specifically, indicators are needed of the strata used in a stratified sampling design, of the primary sampling units (PSUs) used in a multi-stage design, and of the design weights used in a design with variable selection probabilities. Furthermore, these indicators must be in a form that reflects the sample design when viewed as a single multinational sample. If sample design indicators are either not available or not in an appropriate form, this can cause problems for analysis. Alternatively, the data producer could supply analysts with replicate weights (Dippo et al. 1984) that have been produced in a way that appropriately takes into account all features of the sample design. However, it can be argued that using replicate weights places a slightly higher burden on the analyst. Cross-national survey data sets often have one or more of the following problematic features:

- the indicator of sampling stratum is set to 'missing' for countries that don't implement stratification,
- the PSU indicator is left with missing values for countries where a single-stage design is implemented,
- the weight variable is set to 'missing' in countries where the sample is selected with equal selection probabilities, and
- for either the stratum or PSU indicator, the same range of values may have been (partially) used in different countries.

The consequences can be either that the analyst fails to notice the problematic features, leading to incorrect results, or that the analyst chooses to carry out analysis that ignores one or more components of the sample design (for example, clustering may be ignored if the PSU indicator has missing values), leading at least to biased estimates of standard errors.

This article has two aims. We first explain how missing information from countries that omitted a particular sample design feature can be 'filled', and how variables can be recoded if national data sets have been prepared without regard for the requirements for a cross-national data set. This should be useful for users who encounter these problems, but more importantly for data release organizations that, by following these steps, can make it easier for users to account for complex sample design. We then apply the method developed in the first section to create the best possible information on stratification, clustering, and weighting for a large cross-national survey data set. Estimates that use our filled and edited sampling information are compared with those that ignore one or more of the sample design indicators. Specifically, we study misspecification effects if all or part of a complex sample design is ignored in the situation where countries have different sample designs. We examine country comparisons of means and their standard errors for a number of demographic and economic variables.

While in this article we refer to comparisons between countries, the methodology presented has broader application. It applies to any situation where sample designs differ between domains, and these domains are either combined or compared in analysis. Such domains might include regions of a country or strata in a multi-stratum sample.

2. Preparing Sample Design Indicators for Cross-National Analysis

A cross-national sample can be viewed as a special case of a multiple-frame sample. Multiple-frame samples use more than one sampling frame to represent a population (Hartley 1962). Most literature on multiple frames discuss cases where one frame covers all units and another frame is cheap but covers only a subset, or where two frames overlap (Hartley 1962; Cochran 1965; Lohr, 2007; Lepkowski and Groves 1986). A cross-country survey represents a different situation, specifically where none of the frames overlap. According to Hartley (1962), a multiple-frame sample should meet the following requirements:

1) each unit in the population of interest should belong to at least one of the frames, and
2) for each sampled unit, it should be possible to record whether or not it belongs to the other frame(s).

In the cross-national survey context, these requirements are clearly met if we can assume the frames to be non-overlapping. Furthermore, cross-national surveys can be viewed as Hartley's case number 1, where all domain sizes are known (i.e., country totals). According to Hartley (Hartley 1962, 204), in this situation the frames (countries) should be treated as strata. He then notes "In case 1 the estimation problem is reduced to the standard methodology for stratified sampling." Thus each frame (country) should be viewed as a top-level explicit stratum, between which sample designs can vary.

2.1. *Cross-National Stratum Indicator*

For cross-national analysis, a single stratum indicator is required that reflects the complete multi-frame design. This indicator should reflect the sampling strata within each country, and treat countries as the top level strata (as samples were selected independently in each country). It is important that each stratum from the cross-national perspective should take a unique value, and therefore if one country supplies a stratum indicator taking values of 1 to 5 and another country uses 1–7 to indicate strata, the values should be recoded (for example the second country's strata should be coded as 6–12). Any country that does not use stratified sampling should be treated as a single stratum. Thus, in countries with stratification the cross-national stratum indicator should take a different value for each national stratum, while for countries with no stratification, the cross-national stratum indicator should take the same value for each sample element. This is analogous to the situation in national surveys where some regions are treated as a single stratum, while others are subdivided into more detailed strata. If none of the countries has a stratified design, each country should be treated as a separate stratum and the stratum indicator for cross-national analysis should simply take a different value for each country.

2.2. *Cross-National PSU Indicator*

Analogously to the stratum indicator, the cross-national PSU indicator should indicate the units selected from each frame at the first stage of selection when the survey is viewed as a single cross-national sample. If none of the countries has a multi-stage sample design, the PSU indicator can be omitted with caution. Caution is needed in case there are multiple possible levels of analysis relating to hierarchically-associated units such as households and individuals. In this case, a single-stage sample of households would produce a multi-stage sample of individuals, where households are the PSUs within which individuals are clustered. In this situation, we suggest that the PSU indicator should be equivalent to a household indicator. Again, a different range of values should be used in each country so that each household has a unique value in the cross-national data set. Defined thus, the PSU indicator is important for individual-level analysis, while for analysis at household level it will, correctly, have no effect, as it will indicate the absence of clustering.

If all countries have a multi-stage design, then the cross-national PSU indicator should reflect this with a unique value for each PSU when the sample is viewed from a cross-national perspective. Attention is again needed to avoid the same value being used in more than one country.

In a situation where some, but not all countries use a multi-stage design, the indicator should take a unique value for each PSU in each multi-stage country, while it should take a unique value for each sample element in countries with a single-stage design. In this way, using the indicator will provide correct complex sample estimation in an analysis of multiple countries with and without multi-stage designs.

2.3. *Cross-National Weights*

For comparison of estimates between countries it is only necessary that the weight variable reflects the relative inclusion probabilities within each country; between-country differences in the mean weight will not affect comparisons for any type of ratio estimate such as means, proportions or model coefficients (Dorofeev and Grant, 2006, 82–84; see also Brewer 1963). However, we suggest routinely applying what we will call 'population scaling' to the weights. This will render them suitable for any kind of analysis, including that which combines countries, such as estimation for the total cross-national sample or comparison of groups of countries. For unit *i* in country *j* the population-scaled weight for cross-national analysis should take the form:

$$w_{ij}^{s} = \frac{w_{ij}^{u} N_j}{\sum_{i=1}^{n_j} w_{ij}^{u}} \tag{1}$$

where w_{ij}^{u} is the national (unscaled) weight for the unit, and

n_j is the sample size in country *j*, and
N_j is the (assumed known) population size of country *j*.

Using this population-scaled weight, the weighted sample size for each country equals the population size of the country, that is $\sum_{i=1}^{n_j} w_{ij}^{s} = N_j$. An equivalent approach is used by the European Social Survey – see the description of "population size weight" in European Social Survey (2014).

In the special case where a country has a sample design with equal selection probabilities, the national weight may be missing. In this case, it should first be set to a constant value such as 1 for all sample elements in the country, that is $w_{ij}^{u} = 1 \, \forall \, i$. Then, Expression (1) can be applied though for such countries it can be simplified to:

$$w_{ij}^{s} = \frac{N_j}{n_j} \tag{2}$$

2.4. *Cross-National Data Set*

Once the steps outlined above have been followed, the three sample design indicator variables (stratum, PSU, and weight) are ready for use in any kind of cross-national analysis and can be incorporated into standard procedures for complex sample design estimation. Ideally, these steps should be carried out by the data production organization, so that data released to analysts is already in a suitable form for analysis. In that way, the analyst needs only to know how to carry out standard survey analysis, and does not additionally need to perform the data preparation relating to sample design.

3. Empirical Study of Misspecification Effects: Methods

Next, we study how important it is for an analyst of cross-national survey data to have full information on the complex sample design, and whether conclusions about differences between countries can be influenced by ignoring all or part of the sample design information. We concentrate on studying the effect of ignoring stratified and/or multi-stage (clustered) sampling where countries differ in their sample design, compared to estimation using stratum and PSU indicators that have been completed and edited following the procedures outlined in the previous section.

3.1. *Data: EU-SILC*

For our study we use data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. The EU-SILC has been carried out in all 27 EU Member States since 2007 (some started earlier) plus four non-Member States (Wolff et al. 2010). Both cross-sectional and longitudinal data are collected on income, poverty, social exclusion, and other living conditions. Most items are collected through individual interviews with each adult in a household though some items are collected through a household interview. In most countries the data is collected by means of a survey with a rotating panel design (Iacovou and Lynn 2017). Though the details of the design vary, a typical design involves a four-wave rotation with annual interviews. Some countries select a sample of households via addresses, while others first select a sample of individuals and then identify the household of each selected individual. The latter group further subdivides into countries where all adult household members are interviewed and countries where only the selected individual is interviewed, as information on the other household members can be collected from population registers. Furthermore, some countries use a multi-stage clustered design, while others use a single-stage design. Key sample design parameters for each country are summarized in the supplemental data, Appendix 1 (available online at http://dx.doi.org/ 10.1515/jos-2017-0007).

We use data related to 2007, extracted from the longitudinal EU-SILC data set (EUSILC LONGITUDINAL UDB 2007 – version-1 of August 2009 [EOM]). The cross-sectional data set could not be used as it did not include a PSU indicator. We drop a number of countries from our analysis that either had not yet provided this data at the time of analysis, or for whom the indicators of sample design parameters – which are crucial to our analysis – were either missing completely or did not correspond to the description of the design (and where these discrepancies could not be resolved). This leaves 19 countries for analysis. The details can be found in the supplemental data, Appendix 1 (available online at http://dx.doi.org/10.1515/jos-2017-0007).

3.2. *Data Editing: Complex Sample Design Variables*

We apply the procedures outlined in Section 2 to the EU-SILC data. Although a majority of countries used stratified sampling, no stratum indicator exists in the data files, so we treat countries as strata and create a stratum indicator that takes a unique value for each country. For countries with single-stage sample designs we create a PSU indicator that is coterminous with household; for countries with multi-stage designs we use the existent PSU indicator, but recode to avoid between-country overlap in the ranges of values. We do not utilize weights provided by Eurostat, as these incorporate nonresponse adjustments for some countries, but not all, and do not always appear to reflect the described sample design. Instead, we derive our own design weights based on the documented description of the sample design in each country and, where relevant, the data item indicating the number of adults in the household. No attempt is made to develop nonresponse adjustments to these weights, as our focus in this article is on the effects of sample design on precision of estimates. For some countries, the weight for individual-level analysis is different from that for household-level analysis. Specifically, if a country implemented a sample of households, but only one individual was selected, we corrected for within-household selection for individual-level analysis (no such correction was needed for household-level analysis). If a country selected a sample of individuals and then included the household of each individual, we corrected for the fact that households are sampled with probability proportional to the size of the household (while no correction is needed for individual-level analysis). For further details please see the supplemental data, Appendix 1 (available online at http://dx.doi.org/10.1515/jos-2017-0007).

3.3. Estimation

We use the svy commands in Stata 11.0 to provide estimates that take into account aspects of the sample design. Similar approaches can be used in other software packages. Our Stata syntax for estimating a difference between two countries in mean value of the variable var1 is as follows, where the variables strata1, psu, and weight1 are the three sample design indicator variables derived as described in the previous paragraph:

```
svyset psu [pw=weight1], strata(strata1)

svy: mean var1 if cntry1==1 | cntry1==2, over(cntry1)

lincom [var1]1 - [var1]2
```

It can be seen that this form of estimation is very simple to implement once the design variables have been correctly derived. We estimate differences between pairs of countries in a number of descriptive parameters (means and proportions, including some subgroup means). We note in passing that Stata estimation routines will incorrectly estimate the degrees of freedom used to construct the design-based confidence interval for the difference between countries whenever the true degrees of freedom differ between the countries. The effect is likely to be negligible when the degrees of freedom are large, but may not be negligible if the design in at least one of the countries has a small number of degrees of freedom. This problem exists independently of whether or not the design is correctly specified and is therefore not the focus of this article. The interested reader is referred to Valliant and Rust (2010) for discussion of this issue.

Our objective is to estimate what we call the inverse misspecification effect, *imeff,* in a range of scenarios. The misspecification effect, *meff* (Skinner 1989), is the ratio of the true variance of a sample statistic under the complex sample design to the estimated variance, when ignoring all or part of the sample design. The *imeff* (which equals 1/*meff*) is useful because it indicates the factor by which the variance of the estimate is under- or overestimated. If *imeff* is over 1 the variance is overestimated, but usually *imeff* is under 1, which means that the variance is underestimated by a factor of *imeff.*

In all cases, we assume that weights are correctly specified in the analysis. We consider three likely forms of misspecification when using the EU-SILC data:

- failing to take into account that samples are selected independently in each country (i.e., failing to treat countries as strata),
- failing to take into account that the sample is clustered (i.e., treating the sample as if it were a single-stage design), and
- only partially taking into account that the sample is clustered (suboptimal specification of clusters), specifically, recognizing that individuals are clustered within households, but not that households may be clustered within larger PSUs.

In combination, this leads to five possible types of misspecification (Table 1). For each type of misspecification, we estimate *imeff* for each of 90 pairs of countries, specifically all

Table 1. Design misspecification scenarios.

Five types of misspecification:	
Type 1	Ignore independence of samples and ignore clustering
Type 2	Ignore independence of samples
Type 3	Ignore clustering
Type 4	Ignore independence of samples and only partially consider clustering
Type 5	Only partially consider clustering

the pairs that consist of one country with a multi-stage (clustered) design and one with a single-stage design. (Of the 19 countries available for analysis, ten had multi-stage design and nine had single-stage design.) For household-level analysis, only misspecification Types 1, 2, and 3 are possible, as clustering of individuals within households is not relevant to household-level estimation. We estimate differences between countries for five household-level variables (listed in Table 2) and fifteen individual-level variables (listed in Tables 3, 4, and 5), leading to 8,100 estimates of *imeff*.

4. Results

As described above, we carry out analysis for five household-level estimates for each of three types of misspecification and for fifteen individual-level estimates for each of five types of misspecification. This is done for all 90 country pairs. Overall, we find that the *imeff* is, in general, considerable when the clustering is not specified, whereas the effect of ignoring the stratification is negligible for most estimates. Thus, results for Type 1 and Type 3 misspecification (see Table 1) are very similar, as are results for Types 4 and 5, while all 1,530 estimates of *imeff* for Type 2 are in the range 0.98–1.00. Therefore, we present here only the results from misspecification Type 1 and Type 4, as these capture all of the important findings.

4.1. Household-Level Questions

Starting with Type 1 results for each of the five household variables, in Table 2 we present the mean *imeff* (across the 90 country pairs). These are in the range 0.70–0.90. However, we also present the minimum and maximum estimated *imeff* for each variable, and this shows that in specific pairwise comparisons, *imeff* can be as low as 0.07. This means that the true variance could be 14 times the size of the estimated one if the design is misspecified in this way, and standard errors could be nearly four times the size of the estimated ones.

Table 2. Results for five household-level variables: misspecification Type 1 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Income	19160.32	0.80	0.25	0.07	1.00
Capacity to afford holidays	0.25	0.71	0.20	0.33	0.96
Capacity to afford meals	0.12	0.81	0.14	0.54	0.99
Ability to make ends meet	0.06	0.83	0.15	0.43	0.99
Number of household members	0.28	0.87	0.11	0.55	1.00

Table 3. Results for twelve individual-level variables: misspecification Type 1 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Gender	0.024	2.31	0.34	1.73	3.08
Age	2.06	0.64	0.09	0.39	0.78
Equivalized disposable income	11,737	0.38	0.14	0.03	0.63
Education (ISCED)	0.099	0.55	0.19	0.06	0.81
Economic activity	0.070	0.76	0.16	0.27	0.97
Employment	0.044	0.73	0.15	0.41	1.01
Education (males)	0.097	0.74	0.22	0.09	0.97
Economic activity (males)	0.066	0.99	0.14	0.57	1.22
Employment (males)	0.039	0.81	0.11	0.62	1.00
Education (females)	0.112	0.77	0.23	0.13	1.00
Economic activity (females)	0.075	0.94	0.18	0.37	1.14
Employment (females)	0.052	0.86	0.13	0.51	1.06

4.2. Individual-Level Questions Available for All Household Members

Table 3 summarizes results for those individual-level estimates that are based on observations of all individuals in each sample household, either because all individuals were interviewed or because only one person was interviewed, but information for other individuals was obtained from a population register. Twelve of the 15 individual-level estimates are of this type, of which six are whole-sample, three are based on males only and three on females only. Among these estimates, the largest mean *meff* (across the 90 country pairs) is 2.31 for gender which is, unusually, (well) above the value of 1.00. This is a unique situation, which suggests that failing to take into account clustering results in an overestimate of the standard error of the difference. This reflects that PSUs (which, for several countries, consist of households) in the population are more heterogeneous with respect to gender than random samples of the same size from the whole population would be. As a consequence, sample-clustering reduces the standard error of the estimated gender distribution.

Apart from gender, the mean *imeff* (across the 90 country pairs) ranges from 0.38 for mean equivalized disposable income to 0.99 for the proportion of males who are economically active. This is a much greater range than observed above for household-level estimates, reflecting the larger intra-cluster correlation for individual variables due to the additional level of clustering (individuals within households) and the larger sample size per PSU. Failing to correctly take clustering into account is therefore particularly problematic for individual-level estimation. Some values of *imeff* for differences between two countries are very low indeed, with the smallest being 0.03 for a difference in mean equivalized disposable income, implying that standard errors could be underestimated by a factor of six. As an indicator of the extent to which this underestimation may affect analytical conclusions, we would note that, excluding gender, 27 of the 990 comparisons (2.7%) appear significant ($P < 0.05$) if the design is misspecified in this way, but not significant if correctly specified.

Unlike Type 1 misspecification (Table 2), which completely ignores clustering, Type 4 misspecification partially accounts for clustering. Specifically, household IDs are treated as clusters in both countries and this is compared to correctly specifying PSUs in countries

Table 4. Results for twelve individual-level variables: misspecification Type 4 over 90 country-pairs.

	$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
Gender	0.024	0.95	0.06	0.78	1.02
Age	2.06	0.93	0.12	0.59	1.07
Equivalized disposable income	11,737	0.77	0.26	0.07	1.00
Education (ISCED)	0.099	0.72	0.24	0.08	0.97
Economic activity	0.070	0.90	0.19	0.33	1.10
Employment	0.044	0.82	0.16	0.45	1.03
Education (males)	0.097	0.78	0.23	0.10	0.97
Economic activity (males)	0.066	0.94	0.13	0.54	1.13
Employment (males)	0.039	0.86	0.10	0.65	1.00
Education (females)	0.112	0.79	0.23	0.14	1.00
Economic activity (females)	0.075	0.90	0.17	0.35	1.00
Employment (females)	0.052	0.88	0.13	0.53	1.06

where such are present. The same variables are used and the same comparisons are implemented as in Table 3.

As expected, the estimate of the difference itself is not influenced (Table 4). Overall, the mean *imeff* for Type 4 misspecification is much less pronounced than the mean *imeff* for Type 1 misspecification. It comes closer to 1.0 for all estimates except for two (economic activity for males and females), which were close to 1.0 already in Table 3 (the change for these two estimates is minor). For example, the mean *imeff* changes from 0.38 to 0.77 for equalized disposable income. The minimum and maximum *imeff* are also much closer to 1.0. Overall, taking into account clustering of individuals within households improves the estimates considerably, even when ignoring prior stages in a multi-stage sampling design.

4.3. Individual-Level Questions Available for All Household Members in Some Countries and for One Household Member in Other Countries

Thus far, we have discussed the situation in which information is available for all household members, obtained either through an interview or from a register. However, in countries where only one person was interviewed in each household, some variables were not available from a register, leading to a situation in which some variables (for example health evaluation) are only available for one household member. When using such variables to construct estimates of differences between countries, the effect of misspecification can be different from that of variables available for all household members, even though correct specification takes the same form. When comparing two countries, one with a multi-stage sample of households and one with a single-stage sample of households, we distinguish between four situations:

a) both countries may have one individual observed per household,
b) both have all individuals observed per household,
c) only the clustered country has all observed, or
d) only the unclustered country has all observed.

These four scenarios have potentially different implications for misspecification, so in Table 5 we present results separately for each scenario.

Table 5. Results for self-assessed general health (individual-level): misspecification Type 1.

		$\overline{y_1 - y_2}$	$\overline{imeff}$	*s.d. (imeff)*	*Min. (imeff)*	*Max. (imeff)*
All individuals in both countries (48 comparisons)	All	0.071	0.72	0.06	0.61	0.85
	Men	0.060	0.91	0.07	0.69	1.06
	Women	0.080	0.89	0.06	0.77	0.98
One per household in both countries (six comparisons)	All	0.057	0.95	0.01	0.93	0.97
	Men	0.050	0.98	0.01	0.97	0.99
	Women	0.063	0.97	0.03	0.94	1.00
All individuals in PSU country; one per household in non-PSU country (24 comparisons)	All	0.087	0.83	0.08	0.68	0.97
	Men	0.076	0.93	0.07	0.74	1.05
	Women	0.095	0.92	0.06	0.81	0.99
One per household in PSU country; all individuals in non-PSU country (twelve comparisons)	All	0.071	0.83	0.03	0.77	0.89
	Men	0.063	0.96	0.02	0.93	0.98
	Women	0.079	0.95	0.03	0.92	1.00

It can be seen that values of *imeff* are modest when both countries interview only one person per household, but a little more substantial when one of the countries interviews all persons. The largest values of *imeff* arise when both countries interview all persons, as in this case an entire level of clustering is being ignored in both countries.

5. Conclusions

Our findings show that misspecification effects in cross-national comparisons can be considerable and can result in serious bias in standard errors of estimates of between-country differences. This would result in biased hypothesis testing (Type 1 errors). Bias is greatest when multi-stage sample selection is ignored completely in estimation. Bias is smaller, but still substantial (for individual-level estimates) when the first stage is ignored and only the clustering of individuals within households is acknowledged. Furthermore, misspecification effects have been shown to depend on the nature of the difference in sample design between the two countries being compared. The corollary of this is that in multi-country comparisons, if designs are misspecified in estimation, the chances of a country being identified as an outlier depends on the sample design adopted in that country. This is clearly undesirable.

To avoid misspecification effects in cross-national comparisons, it is necessary not only for sample design indicators (PSU, stratum, and design weight) to be present on the data set, but also for these indicators to be in a form that is suitable for cross-national analysis. Indicators that are suitable for national analysis of each country do not necessarily meet this requirement, but in Section 2 above we have set out the steps necessary to convert these indicators into a suitable form. These steps are not particularly demanding and we propose that they should be carried out by a relevant central agency before data is released to analysts. This is efficient, as it avoids duplication of effort, and mistakes by analysts who may not be experts in sample design. Once suitable indicators for cross-national analysis have been produced, correct specification can easily be achieved with standard software, leading to unbiased estimation of standard errors.

However, we are aware that the EU-SILC is certainly not the only cross-national survey data set in which the sample design indicators are not in suitable form. An analyst of any such data would be well-advised to follow the data preparation steps that we propose here. Furthermore, there are some cross-national survey data sets that do not release indicators of sampling strata or primary sampling units to secondary analysts at all. The European Social Survey is one prominent example (see http://www.europeansocialsurvey.org/data/). The producers of such data sets should be encouraged to release these indicators so that analysts can appropriately estimate standard errors and test hypotheses.

While we have focused here on how best to estimate the impact of sampling error on cross-country comparisons, the impact of other components of statistical error may be equally important. It is not our intention to imply otherwise. In addition, estimating the magnitude of error *post-hoc* is no substitute for controlling the error at the design and data collection stages. All sources of error (coverage, sampling nonresponse, measurement, editing, and so forth) should be given due attention within a total survey error framework (Biemer 2010; Groves and Lyberg 2010) that recognizes interactions and dependencies between the error sources. Our comments on sampling error should be considered within that context, although further discussion of the broader context is outside the scope of this article.

Finally, we should note some limitations of our research. We have not examined all possible variants of misspecification. In particular, we have not assessed the effects of ignoring variation in design weights. Nor have we assessed the effects of ignoring stratified sampling within countries. The first of these is, in general, likely to lead to even greater underestimation of standard errors. The second is likely to have a rather more modest effect in the opposite direction. Furthermore, we have examined a limited number of estimates for one survey, albeit important ones. Effects might be different in magnitude for estimates of substantially different parameters and for substantially different sample designs (e.g., those with much larger, or smaller, cluster sample sizes). However, we do not feel that any of these limitations invalidate our main conclusion, which is that misspecification can have a serious effect and can (and should) be avoided. Though the effect may be different in magnitude in other circumstances, the data preparation steps outlined here guarantee that the effects can be completely avoided. As implementing the steps has very modest resource implications, we think that this should always be done.

6. References

Biemer, P.P. 2010. “Total Survey Error: Design, Implementation, and Evaluation.” *Public Opinion Quarterly* 74: 817–848. Doi: http://dx.doi.org/10.1093/poq/nfq058.

Brewer, K.R.W. 1963. “Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process.” *Australian Journal of Statistics* 5: 93–105.

Cochran, R.S. 1965. “Theory and Application of Multiple Frame Surveys.” *Retrospective Theses and Dissertations*. Paper No. 4080.

Dippo, C.S., R.E. Fay, and D.H. Morgenstein. 1984. “Computing Variances from Complex Samples with Replicate Weights.” In Proceedings of the Survey Research Methods Section of the American Statistical Association. 489–494.

Dorofeev, S. and Grant, P. 2006. *Statistics for Real-Life Sample Surveys: Non-Simple-Random Samples and Weighted Data*. Cambridge: Cambridge University Press.

European Commission. 2013. *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys*. Luxemburg: Publications Office of the European Union. Available at: http://ec.europa.eu/eurostat/documents/3859598/5927001/KS-RA-13-029-EN.PDF (accessed January 2017).

European Social Survey. 2014. *Weighting European Social Survey Data*. Available at: www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf (accessed 28 October 2015).

Gabler, S., S. Häder, and P. Lynn. 2006. “Design Effects for Multiple Design Samples.” *Survey Methodology* 32: 115–120.

Groves, R.M. and L. Lyberg. 2010. “Total Survey Error: Past, Present, and Future.” *Public Opinion Quarterly* 74: 849–879. Doi: http://dx.doi.org/10.1093/poq/nfq065.

Häder, S. and S. Gabler. 2003. “Sampling and Estimation.” In *Cross-Cultural Survey Methods*, edited by J.A. Harkness, F.J.R. Van de Vijver, and P.Ph. Mohler, 117–134. Hoboken, New Jersey: Wiley.

Hartley, H.O. 1962. “*Multiple Frame Surveys*.” In Proceedings of Social Science Section of American Statistical Association meetings. Minneapolis, Minnesota. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y1962/Multiple Frame Surveys.pdf (accessed January 2017).

Heeringa, S.G. and C. O'Muircheartaigh. 2010. "Sampling Designs for Cross-Cultural and Cross-National Survey Programs." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, and T. Smith, 251–268. New Jersey: Wiley.

Iacovou, M. and P. Lynn. 2017. *Design and Implementation Issues to Improve the Research Value of the Longitudinal Component of EU-SILC*. Monitoring Social Inclusion in Europe, edited by A.B. Atkinson, A.-C. Guio and E. Marlier. Chapter 27. EU Publications.

Kish, L. 1989. Q/A 21.1 Comparisons of Surveys. *Questions/Answers. From the Survey Statistician*, edited by A.M. Vespa-Leyder, 40–41.

Kish, L. 1994. "Multipopulation Survey Designs." *International Statistical Review* 62: 167–186.

Kish, L. 1999. "Cumulating/Combining Population Surveys." *Survey Methodology* 25: 129–138.

Lepkowski, J. and R.M. Groves. 1986. "A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design." *Journal of the American Statistical Association* 81: 930–937.

Lohr, S. 2007. *Recent Developments in Multiple Frame Surveys*. In Proceedings of Survey Research Methods Section of American Statistical Association meetings, Salt Lake City, Utah. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000580.pdf (accessed January 2017).

Lynn, P., L. Japec, and L. Lyberg. 2006. "What's So Special about Cross-National Surveys?" *Conducting Cross-National and Cross-Cultural Surveys: Papers from the 2005 Meeting of the International Workshop on Comparative Survey Design and Implementation (CSDI)*, edited by J. Harkness. ZUMA, Mannheim.

Lynn, P., S. Häder, S. Gabler, and S. Laaksonen. 2007. "Methods for Achieving Equivalence of Samples in Cross-National Surveys: the European Social Survey Experience." *Journal of Official Statistics* 23: 107–124.

Skinner, C.J. 1989. "Introduction to Part A." In *Analysis of Complex Surveys*, edited by C.J. Skinner, D. Holt, and T.M.F. Smith, 23–58. Chichester: Wiley.

Smith, T.W. 2010. "The Globalization of Survey Research." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.Ph. Mohler, B.-E. Pennell and T.W. Smith, 477–484. Hoboken, New Jersey: Wiley.

Valliant, R. and K.F. Rust. 2010. "Degrees of Freedom Approximations and Rules of Thumb." *Journal of Official Statistics* 26: 585–602.

Wolff, P., F. Montaigne, and G.R. González. 2010. "Investing in Statistics: EU-SILC." In *Income and Living Conditions in Europe*, edited by A.B. Atkinson and E. Marlier, 37–55. Luxembourg: Publications Office of the European Union.

Received December 2015
Revised June 2016
Accepted June 2016

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 1, http://dx.doi.org//10.1515/JOS-2017-0007

Appendix 1. Summary of Sample Design Data.

Country**	Type of units drawn	Personal level data collection	Stratification	PSUs above household level	Weights for ind-level*	Weights for HH-level and for income estimates
Austria (AT)	Households	Interview	1 var	No	1	1
Belgium (BE)	Households	Interview	1 var (11 strata)	Yes	1	1
Cyprus (CY)	Households	Interview	1 var (9 strata)	No	1	1
Czech Republic (CZ)	Households	Interview	2 vars	Yes	1	1
Estonia (EE)	Households	Interview	1 var (3 strata)	No	1	1
Finland (FI)	Persons	Interview and register	1 var	No	1	$1/n_i$
France (FR)	Households	Interview	2 vars	Yes	1	1
Hungary (HU)	Households	Interview	2 vars	Yes, for one part	1	1
Iceland (IS)	Persons	Interview and register		No	1	$1/n_i$
Italy (IT)	Households	Interview and register	2 vars	Yes	n_i	1
Lithuania (LT)	Households	Interview	1 var (7 strata)	No	1	1
Luxembourg (LU)	Households	Interview	yes, no info on var #	No	1	1
Latvia (LV)	Households	Interview	1 var	Yes	1	1
Netherlands (NL)	Households	Interview and register	2 vars	Yes	n_i	1
Poland (PL)	Households	Interview	2 vars	Yes	1	1
Sweden (SE)	Persons	Interview and register		No	1	$1/n_i$
Slovenia (SI)	Persons	Interview and register	2 vars	Yes	1	$1/n_i$
Slovakia (SK)	Households	Interview	2 vars	No	1	1
United Kingdom (UK)	Households	Interview	yes, no info on var #	Yes	1	1

*Weights for individual level indicates weights for the questions (e.g., on health) that are available only from an interview and not from register; weights for HH-level indicate weights for a household-level analysis; weights for income estimates indicate weights for the questions that are available from both interviews and registers.

**As of 1st August 2009 release of EU-SILC data set used in this article (EUSILC LONGITUDINAL UDB 2007 – version-1 of August 2009 [EOM]) the data from the following countries were not available: Germany, Denmark, Greece, and Ireland. The following countries are present on the data set but are excluded from the analysis: Spain because it used substitutes for nonrespondents (and no indicator for this is present on the data set); Portugal because no PSU information is provided while the sampling description suggests a multistage design; Norway because all persons have PSU information while the sample design description suggests a two-domain design with only one part being multistage. Some rotational groups where the sample design description does not match the data set are also excluded: two of eight groups in France (missing information on PSUs); two groups in Hungary (where all members have PSU information while it should be two-domain design with a multi-stage design only in one domain); and one group in Poland (missing PSU information). Additionally, because the cross-sectional data set has missing information on PSUs from at least seven countries, we use the longitudinal dataset. This means that new rotational groups that entered EU-SILC study in 2007 are excluded from the analysis.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 137–154, http://dx.doi.org/10.1515/JOS-2017-0008

Effects of Scale Direction on Response Style of Ordinal Rating Scales

Mingnan Liu[1] *and Florian Keusch*[2]

Although ordinal rating scales have received much research attention in survey methodology literature, the direction of the rating scales has not been as extensively studied as other design features. Research on scale direction effect has mainly focused on the influence on response distribution, while largely overlooking its impact on latent constructs. This study examines the scale direction effect on extreme and acquiescent response style latent class variables in an experiment embedded in a national probability sample. We found a higher level of acquiescent response style from scales starting with positive adjective words using a web survey. No significant effect of scale direction was detected on extreme response style or in a face-to-face survey (with show cards). This study also demonstrates that scale direction does not affect the substance latent class variables, once the response style latent class variables are included in the model. Implications of these findings and future research directions are discussed.

Key words: Rating scale; scale direction; response style; latent class analysis; data collection mode.

1. Background and Literature Review

Attitudes, as different from behaviors, are not directly observable. In surveys, attitudes are typically measured by a series of items with ordinal rating scales. Given the popularity of attitudinal questions, many research efforts have been devoted to examining measurement error associated with this type of question, which leads to a very rich body of literature on this topic (for a review, see Krosnick and Presser 2010). When designing a rating scale, many decisions have to be made by researchers, one of which is the scale direction. In rating scales, the two endpoints are usually defined by two adjective words. One way to design the rating scale is to put the positive/high adjective on the left side/top of the scale and the negative/low adjective on the right side/bottom of the scale. Another way is to reverse the order of the response options by putting the negative/low adjective first and the positive/high adjective last. This seemingly trivial design decision could affect response behavior and result in a shift in the response distribution – the 'scale direction effect' (Belson 1966).

Research on the scale direction effect dates back more than half a century (Belson 1966) and several studies focused on the impact on the univariate distribution. For example,

[1] SurveyMonkey, 101 Lytton Avenue, Palo Alto, CA 94301, U.S.A. Email: mingnanliu@gmail.com
[2] University of Mannheim, 68131 Mannheim, Germany. Email: f.keusch@uni-mannheim.de

He et al. (2014) report in their summary of empirical studies involving an experiment on scale direction that twelve out of 27 experiments found a shift in responses based on the starting point. That is, the high/positive end of the scale was endorsed more often when the scale started with the high/positive end, and the low/negative end of the scale attracted more endorsement when the scale began with the low/negative end. Dillman et al. (1995) examined scale direction effects in 82 experiments in twelve surveys using ordinal scales over the telephone and in mail questionnaires. Out of 21 experiments involving mail surveys, only one showed a significant scale direction effect. Three out of 22 telephone experiments showed a significant effect. Furthermore, of the fifteen experiments using both mail and telephone, only two showed a significant effect across both modes. The authors concluded that the scale direction effect "has been overestimated by past research" (p. 674). Another set of experiments by Christian et al. (2008) tested presenting the positive option versus the negative option first using five-point ordinal scales in both telephone and web surveys. None of the comparisons in each mode yielded significant differences. A recent experiment by Stapleton (2013) on scale direction effects among respondents who could choose from one of two modes (PC and mobile device) showed that the univariate distribution shifted by the scale direction in both modes, although mobile respondents were more prone to the scale direction effect than PC respondents.

In addition to the distribution of response options, researchers have also examined the impact of scale direction on the latent structure of estimates. For example, Chan (1991) administered five items with five response options to a group of high school students and in a factor analysis, the results showed that the positive-negative scale had a better model fit than the negative-positive scale. However, item discrimination is higher in the latter than the former scale. These findings mean that the two forms of scale direction produce different estimations of the latent trait in the same group of individuals. A study by Krebs and Hoffmeyer-Zlotnik (2010) showed no substantial difference in terms of the dimensional structure, factor loadings, and the latent means between the two scale directions. A recent study adopted two analytical approaches – the Rasch model and confirmatory factor analysis – to examine the interaction effect between scale direction and response speed (Salzberger and Koller 2013). The two modeling approaches revealed different findings, including differential interaction effects with response speed. Saris and Gallhofer (2007) conducted a meta-analysis of multitrait–multimethod experiments on question reliability and validity. Scale direction is one of the factors they examined, and the findings showed that providing the negative option first reduced the reliability but improved the validity of the responses.

Taken together, empirical studies report mixed findings with regard to the influence of the direction of rating scales on response. Some studies found no significant difference in terms of univariate distribution (Dillman et al. 1995; Christian et al. 2008), while others did find a difference (Krebs and Hoffmeyer-Zlotnik 2010; Belson 1966; Stapleton 2013). Also, some studies found no latent structure/underlying structure difference between the two scale directions (Krebs and Hoffmeyer-Zlotnik 2010), while others concluded that the latent variables differed by the scale direction (Chan 1991; Saris and Gallhofer 2007) and different modeling approaches reached different conclusions (Salzberger and Koller 2013). There was no obvious difference between interviewer-administered and self-administered modes (Dillman et al. 1995; Christian et al. 2008).

Response style is another form of measurement bias that is particularly associated with ordinal rating scales. Response style refers to the tendency of choosing certain response options from an ordinal rating scale based on some question content-irrelevant information rather than the question itself (Paulhus 1991). Two of the most frequently studied response styles are acquiescent response style (ARS) and extreme response style (ERS). ARS refers to the tendency to agree with a statement regardless of the question content (Baumgartner and Steenkamp 2001). ERS refers to the tendency to select the endpoints of a rating scale more frequently than other response options (Paulhus 1991). Previous research has shown that several design features of ordinal rating scales (e.g., the presence of a midpoint, scale length, labeling) can impact response styles (Moors 2008; Moors et al. 2014; Kieruj and Moors 2010). For example, Moors (2008) found that extreme response style existed regardless of the presence or absence of a middle response category. Kieruj and Moors (2013) examined the impact of scale length on both ARS and ERS. Specifically, they studied Likert scales that ranged from 5–11 points and found that both types of response styles existed regardless of scale length. More recently, Moors et al. (2014) examined the labeling of a Likert scale and its impact on response styles. The authors demonstrated that both ARS and ERS existed, regardless of whether numeric or verbal labels were used, and whether the scales were fully or endpoint labeled. As will be discussed in detail below, survey satisficing and anchoring-and-adjustment are potential causes of scale direction effects. Reinforced by acquiescence, these mechanisms can manifest themselves as response styles.

2. Conceptual Framework and Expectations

Different theories have been brought forward when trying to explain the influence of scale direction on response behavior to rating scales. One explanation sees the scale direction effect as a special case of the primacy effect observed in survey modes where lists of categorical response options are presented visually. Respondents are more likely to choose options offered first than later in such lists of unordered response options, due to response satisficing (Krosnick 1991). This is because, unlike optimizers, satisficers choose a response option that is good enough without going through the whole response option list and selecting the best answer. This approach is cognitively less burdensome than providing an optimized response option. As a result, satisficing has also been used as a potential explanation for scale direction effects in ordinal rating scales (e.g., Krebs and Hoffmeyer-Zlotnik 2010).

However, several studies also show that the scale direction effect is not stronger in situations conducive to satisficing (Carp 1974; Mingay and Greenwell 1989), and scale direction effects are also observed in surveys employing aural administration (Kalton et al. 1978; Mingay and Greenwell 1989; Yan and Keusch 2015). Yan and Keusch (2015) propose that the scale direction effect may also be caused by respondents' use of anchoring-and-adjustment (Tversky and Kahneman 1974) when constructing and mapping their answers to one of the scale points. The basic idea is that people make numerical estimates when under uncertainty, by anchoring on an initial value (e.g., the start of a rating scale), and then adjusting to that anchor until a plausible estimate is reached. Because the adjustments made to the anchor are more often incomplete and insufficient, the final estimate is usually biased toward the anchor. Although anchoring-and-adjustment was initially proposed as a heuristic approach for numerical estimates, Yan and Keusch (2015) showed that it can also be used to explain scale direction effects in rating scales. Whether anchoring-and-adjustment also applies to Likert-type agree-disagree questions is yet unknown. Salzberger and Koller (2013) attributed the scale direction effect in multi-item grid questions to the 'near means related' heuristic that respondents use when answering rating scale questions (Tourangeau et al. 2004; Tourangeau et al. 2007). That is, the spatial proximity of the survey items and the starting point of a response scale lead to a higher endorsement of this side of the scale in self-administered modes.

The direction of a response order is an important design feature of rating scales. However, the relationship between scale direction and response styles, including ARS and ERS, is yet unknown. The goal of this study is to evaluate and compare response styles across scale directions through experimental data. Survey satisficing, anchoring-and-adjustment, and acquiescence are all potential causes of measurement bias for ordinal rating scales. Satisficing can lead to a primacy effect, that is, higher endorsement of the response option that is presented first, whether positive or negative. Similarly, the anchoring-and-adjustment process used for responding to ordinal rating scales also predicts a higher level of agreement in the positive-negative scale than the negative-positive scale, since respondents are more likely to use the first presented option as the anchoring point, and it is most likely to be selected if the adjustment process is insufficient. Acquiescence, different from the other two mechanisms, manifests itself in higher endorsement of the positive option, regardless of the order of presentation. Therefore, when measuring ARS, we would expect a higher level of ARS when the positive option is presented first and the negative option is presented last. This is due to the effect being reinforced by scale direction effect, which is likely to be caused by satisficing and the anchoring-and-adjustment process.

Given that more national surveys are moving toward web data collection, this study also compares the impact of scale direction on response styles between web and face-to-face surveys. Liu et al. (2016) showed that face-to-face survey respondents reveal more ARS and ERS than web survey respondents. They attributes the mode effect on response styles to the impression management concerns of respondents in the face-to-face survey. Another goal of this study is to evaluate and compare the ARS and ERS bias between two scale directions in the two modes of data collection separately. Since the primacy effect is typically stronger in a self-administered survey than in an interviewer-administered survey, it is possible that we will observe a larger scale direction effect on ARS in a web survey than in a face-to-face survey. Also, web surveys use only visual display, hence respondents are more likely to use the first visually displayed option as the anchoring point when making the judgment than face-to-face respondents. Interviewers in a face-to-face survey utilize both visual (show cards) and aural channels of communication and may facilitate a more full interpretation of the rating scales. This consideration also leads to the prediction of a larger scale direction effect on ARS in web surveys than in face-to-face surveys.

To sum up, this study utilizes an experiment embedded in a national probability survey to explore the impact of scale direction on response styles in face-to-face and web survey data collection. Specifically, this study will answer three research questions. First, do response styles, including ARS and ERS, differ by scale direction in agree-disagree Likert scales? Second, does the impact of scale direction on response style, if any, differ between face-to-face surveys and web surveys? Third, after adjusting for response style, does the scale direction still have an impact on the substantive content latent class variables? Given the previous research in this area, we expect to find more endorsement of positive options when they are presented first than when they are presented last. Also, we expect to see a larger effect of this in web surveys than in face-to-face surveys. As for the scale direction effect on ERS, our study is largely exploratory, and we do not have a clear expectation. However, given that ERS is an important response style and a previous study shows that the rating scales used in this analysis suffer from ERS (Liu et al. 2015), we also examine whether or not the scale direction has any impact on ERS.

3. Data and Measures

The data analyzed in this study come from the 2012 American National Election Studies (ANES). The ANES is a national time series survey on political candidates, parties, politics in general, and other related topics conducted in election years in the United States. In 2012, the ANES, for the first time, conducted surveys using two modes of data collection, namely face-to-face and web. The survey has two independent national representative samples: one for each mode, and one identical questionnaire. The target population for both samples is US citizens aged 18 or older by the 2012 Election Day. The web survey sample came from the GfK KnowledgePanel, a probability online panel of US adults. The online panel members were recruited using two sampling methods: address-based sampling and random-digit dialing. After a household was selected, all members in the household were enumerated, and panel members were selected. Households without Internet access or necessary equipment for participating in web surveys were provided with such equipment. The face-to-face survey involves an address-based, stratified, multi-stage cluster sample. The sample includes a nationally representative main sample and two oversamples for African Americans and Hispanic Americans. (For more information about the ANES sampling design, see http://www.electionstudies.org/studypages/anes_timeseries_2012/anes_timeseries_2012_userguidecodebook.pdf.)

The ANES contains two stages of data collection, with one wave prior to the presidential election and one wave after the election. In 2012, 5,914 pre-election interviews (2,014 of the face-to-face surveys) and 5,510 post-election interviews (1,929 of the face-to-face surveys) were completed. Response rates for the pre-election study were 38% and two percent for face-to-face and web, respectively (AAPOR RR1). The re-interview rates (conditional on the pre-election study response rates) in the post-election stage for the two modes were 94% and 93%, respectively. According to the survey organization, the response rate for the web survey is a function of the recruitment of panel members, the retention of panelists from the time of recruitment to the point at which they were invited to take the ANES survey, and the response to those survey invitations.

The 2012 ANES pre- and post-election studies included an experiment that randomly assigned respondents to one of two scale direction conditions. Of the 190 ordinal items included in the experiment, three sets of balanced multi-item Likert scales can fully serve the purpose of examining the effect of scale direction on ARS and ERS. Randomization is performed at the respondent level, not at the question level. That is, each respondent received all ordinal rating items labeled in one direction or the other.

For the multi-scale items analyzed in this study, the first scale contains four items about attitudes toward traditionalism. The second scale contains four items about the position of Blacks in society. The third scale contains six items about the attitudes toward egalitarianism (see Appendix for wording of items). All items were measured on the same five-point rating scale. The scales were labeled ‘disagree strongly,’ ‘disagree somewhat,’ ‘neither agree nor disagree,’ ‘agree somewhat’ or ‘agree strongly’ (forward condition) or in reversed order (reversed condition) without numeric labels. In the face-to-face survey, show cards were used to visually present the response options to the respondents. (For show cards in the face-to-face survey, see http://www.electionstudies.org/studypages/anes_timeseries_2012/anes_timeseries_2012_respbooklet_post.pdf.) The response options were displayed vertically in both modes. A “don’t know” option was not explicitly provided in either mode but it was accepted in the face-to-face survey. In the web survey, respondents could skip the question if they chose not to answer. Both “don’t knows” and “skips” are coded as missing in the analysis. Overall, the design of the questionnaire was kept as unified as possible between these two modes. Also, the fully labeled five-point agree-disagree scale conforms to the best practice in the literature (Revilla et al. 2013; Krosnick and Presser 2010).

In this study, latent class analysis (LCA) was used to examine the experimental data on scale direction. Several analytical models exist in the literature on how to measure response styles. We choose this particular analytical approach because it can simultaneously estimate ARS, ERS, and the substantive content of the rating scales as different latent class variables. Moors (2003) was among the first to adopt the LCA model to examine response styles. Specifically, he treated the rating scales as nominal variables and the latent class variables as equidistant ordered variables in order to estimate ERS. The reason for treating the rating scales as nominal rather than as ordinal is because a U-shape is expected for the ERS. In other words, the coefficients for the two endpoints are positive, and the coefficients for the middle options are negative, which is an indication of the endpoint preference. Morren et al. (2011) simplified the model by treating the rating scales as ordinal variables when measuring the content latent class variables and maintaining the rating scales as nominal variables when measuring ERS latent class variables in order to capture the nonmonotonic (U-shape) relationship. This is a more parsimonious model in that, for each response item, only one coefficient is estimated for the content latent class variables. This modeling approach was further simplified by forcing the style latent class variable coefficients to be equal across all items (Kieruj and Moors 2013; Moors et al. 2014). This model only outputs one set of coefficients for all items when estimating the response style latent class variables. This constraint is theoretically meaningful because, by definition, a response style is content-irrelevant, and its impact on all items should be equally likely. For the data in this study, the model can be written as:

$$log\frac{P(Y_{ij}=c+1|F1_i,F2_i,F3_i,E_i,A_i)}{P(Y_{ij}=c|F1_i,F2_i,F3_i,E_i,A_i)}$$
$$=\left(\beta_{0jc+1}-\beta_{0jc}\right)+\beta_{1j_1}F1_i+\beta_{2j_2}F2_i+\beta_{3j_3}F3_i+\left(\beta_{4jc+1}-\beta_{4jc}\right)E_i+\beta_{5j}A_i$$

Where Y_{ij} denotes the response of respondent i to Likert-type item j, $i = 1,\ldots,I$, $j = 1,\ldots,14$;
$F1_i$ denotes the “moral traditionalism” latent class variable;
$F2_i$ denotes the “position of Blacks in society” latent class variable;
$F3_i$ denotes the “egalitarianism” latent class variable;
E_i denotes the ERS latent class variable;
A_i denotes the ARS latent class variable;
β_{1j_1} denotes the effects on the adjacent category logit for the “moral traditionalism” latent class variable, $j_1 = 1, 2, 3,$ or 4;
β_{2j_2} denotes the effects on the adjacent category logit for the “position of Blacks in society” latent class variable, $j_2 = 5, 6, 7,$ or 8;
β_{3j_3} denotes the effects on the adjacent category logit for the “egalitarianism” latent class variable, $j_3 = 9, 10, 11, 12, 13,$ or 14;
$\beta_{4jc+1} - \beta_{4jc}$ denotes the nonmonotonic (U-shape) relationship between the ERS latent class variable and the Likert-type items;
β_{5j} denotes the effects on the adjacent category logit for the ARS latent class variable; and
c denotes the response options, $c = 1, 2, 3$ or 4.

This model is illustrated in Figure 1, showing that the individual items only load on their corresponding content latent class variables (F1, F2, F3) with no cross-loadings specified. The three content latent class variables are allowed to be correlated with each other. For ARS and ERS latent class variables, all items are loaded on these two style latent class variables. This is because response styles should affect all items regardless of the specific question content. The two style latent class variables are not correlated with each other, nor do they correlate with the content latent class variables. The rating items are estimated as nominal variables for measuring ERS. Effect coding is used and, thus, the model outputs five coefficients – one for each response option. For the other four latent class variables, the rating items are estimated as ordinal variables and, thus, one coefficient is estimated for each item. The scale direction is introduced into the model as a covariate. This is the key variable of interest in this model. Previous research has shown that, given the data structure, this model should fit the data well (Liu et al. 2015). However, we also test a few other alternative models in order to find the best fitting empirical model. We use the Bayesian information criterion (BIC) to guide our choice of models. A smaller BIC indicates a better model fit. However, although a more complex model tends to have a better model fit based on BIC, we try to present a conceptually meaningful and parsimonious model rather than a purely data-driven model. The model is applied to face-to-face and web survey data separately because a previous study has shown that

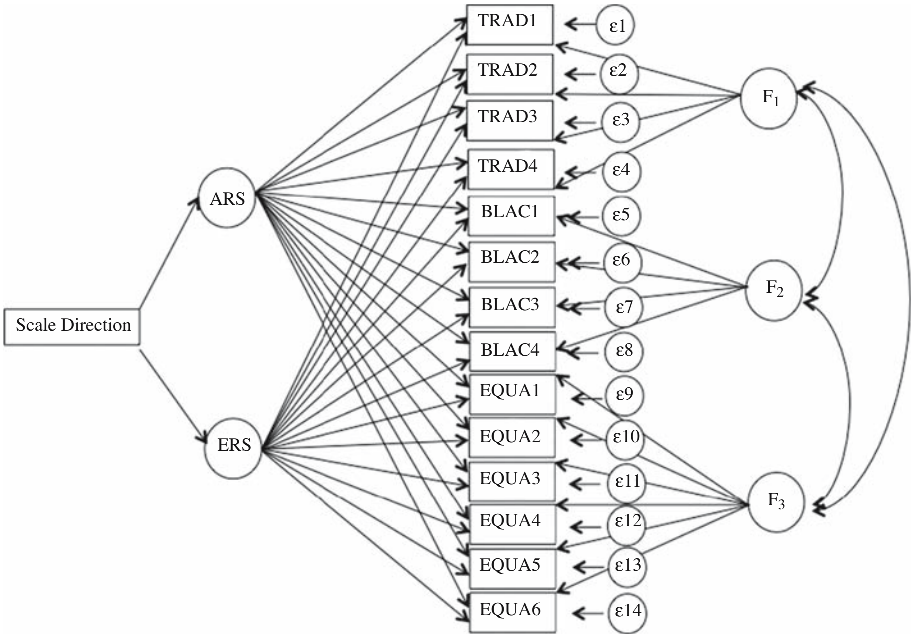


Fig. 1. Latent class analysis model of acquiescent response style (ARS), extreme response style (ERS), and content latent class variables (F_1, F_2, F_3), with covariates (scale direction).

response styles differ in these two modes of data collection (Liu et al. 2015). All analysis is weighted using the weight variable provided by the survey organization, which adjusts for the probability of household selection, the probability of respondent selection within the household, nonresponse, and random sampling error. Based on the survey documentation, the weights are poststratified to produce estimates that match known population proportions for selected characteristics. The weights were created separately for face-to-face and web survey so that the weighted analysis of each survey should, in theory, produce unbiased estimates of the same target population. Missing values are deleted listwise. All analyses are performed in Latent Gold 5.0 (Vermunt and Magidson 2013).

4. Results

Before fitting the LCA models, we first estimated the demographic distributions between the two experimental conditions (forward vs. reversed) in the two modes separately. As Appendix B shows, none of the weighted demographic variables, including gender, age, race/ethnicity, education, marital status, or household income, differed in the two conditions in the two response order conditions and the two modes of data collection. Therefore, it is not likely that the weighted differences observed between the two conditions and two modes are due to the demographic composition differences.

Although the LCA model introduced above is theoretically meaningful, we fit several alternative models, and compare them using BIC in order to find the best fitting empirical model. The first step was to determine whether the data actually reflected response styles. That is, whether adding latent class variables for ARS, ERS, or both to the model in addition to the content latent class variables could improve the model's fit. According to Table 1, this was the case for both the face-to-face and the web survey. In comparison with the content-only model (Model 1), when ARS (Model 2), ERS (Model 3), or both ARS and ERS (Model 4) were included in the model, the BIC dropped; Model 4 has the smallest BIC among the four models. This indicated that both ARS and ERS were critical latent class variables to be added to the model. In other words, respondents' answers to the rating scales not only reflected their opinions toward the substantive content of the questions, but also their response styles.

The next step was to determine the number of levels for all five latent class variables. As we mentioned above, the latent class variables are equidistant ordered latent class variables with at least two levels. Each of the latent class variables in the aforementioned four models contains two levels. Next, we increased the number of levels to three (Model 5) and four (Model 6). However, the four-level latent class model has small and not meaningful class sizes for several latent class variables for both fact-to-face and web surveys. (For the face-to-face surveys, the class sizes for ARS are 0.19, 0.50, 0.24, and 0.07, and for ERS they are 0.11, 0.63, 0.20, and 0.06. For the web surveys, the class sizes for ARS are 0.20, 0.66, 0.09, and 0.05, and for ERS they are 0.28, 0.56, 0.10, and 0.06.) Furthermore, the model becomes more difficult to interpret and replicate. We hope to identify a simple, easy-to-interpret, and theoretically meaningful model. Considering this, we choose to proceed with the three-level model with ordinal latent classes (Model 5).

The last step in model-building is to examine whether or not adding equality constraints can improve the model fit. In Model 5a, for each latent class variable, the coefficients are set to be equal. The BIC of Model 5a is substantially larger than in Model 5 for data from both the face-to-face and the web survey, indicating such constraints deteriorate the model fit. In Model 5b, the coefficients are set to be equal for ARS and ERS, but the coefficients for the three content latent class variables are free to vary. The BIC increases slightly compared with Model 5, but the model has 65 (158-93) more free parameters, which makes it a much more parsimonious model. In addition, this model is conceptually more meaningful, since response styles are content-free, and they should equally influence responses to items. As a result, we conclude that Model 5b is the best-fit model.

Given this information, we deduced that two response style latent class variables exist. An important subsequent step is to determine whether these two style latent class variables

Table 1. Model Fit Statistics, 2012 American National Election Studies.

	BIC		
	Face-to-face	Web	Npar
Model 1: Content only (two-level)	72909	137782	79
Model 2: Content + ARS (two-level)	70639	131711	137
Model 3: Content + ERS (two-level)	71610	133396	95
Model 4: Content + ARS + ERS (two-level)	69581	128620	153
Model 5: Content + ARS + ERS (three-level)	68907	126078	158
Model 5a: Equality on all latent class variables	73643	141588	82
Model 5b: Equality on style latent class variables	68928	126467	93
Model 6: Content + ARS + ERS (four-level)	68791	125355	163

Table 2. Estimated Regression Coefficients (Log Odds) and Standard Errors of ERS and ARS on the Likert Scale Items by Mode of Data Collection, 2012 American National Election Studies.

		Face-to-face		Web	
Response style	Response option	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
ERS	Disagree strongly	2.45	0.16***	1.70	0.16***
	Disagree somewhat	−0.72	0.13***	−1.11	0.15***
	Neither agree nor disagree	−3.96	0.30***	−1.88	0.25***
	Agree somewhat	−0.63	0.12***	−0.92	0.10***
	Agree strongly	2.86	0.14***	2.20	0.16***
ARS		1.11	0.07***	0.98	0.05***

$^{***}p < .0001$.

actually represent ARS and ERS. Table 2 shows the estimated regression coefficients for these two latent class variables in the face-to-face and web survey separately. As described above, there are five coefficients for ERS: one for each response option. For ERS, the coefficients for the two endpoints need to be in the opposite direction from the three middle options. This is what we find from both modes. The coefficients for the two endpoints are positive, and the coefficients for the middle three options are negative. This confirms that this latent class variable is indeed ERS. Recall that ERS is a three-level latent class variable. The coefficients in Table 2 suggest that respondents at the higher level of the latent class variable tend to select the endpoint of the scales more frequently than respondents at the lower level of the latent class variable. That is, a lower level of the latent class variable indicates "avoid-ERS" and a higher level of the latent class variable indicates "pro-ERS". For ARS, because the rating items are treated as ordinal variables, there is only one coefficient output from the model. In the analysis, the negatively worded items are recoded so that they go from negative (disagree strongly = 1) to positive (agree strongly = 5), the positive coefficients in both modes mean that a higher level of the ARS variable indicates pro-ARS and a lower level of the ARS variable indicates avoid-ARS.

The scale direction effect on response styles is presented in Table 3. In this case, the latent class variables are dependent variables and scale direction is the predictor.

Table 3. Estimated Scale Direction Effect (Log Odds) on Response Style Latent Class Variables by Mode of Data Collection, 2012 American National Election Studies.

	Face-to-face					
	ERS			ARS		
	$\hat{\beta}$	S.E.	p-value	$\hat{\beta}$	S.E.	p-value
Scale direction[a]	−0.17	0.23	0.44	−0.28	0.27	0.31
	Web					
	ERS			ARS		
	$\hat{\beta}$	S.E.	p-value	$\hat{\beta}$	S.E.	p-value
Scale direction	−0.10	0.18	0.60	1.73	0.35	<.0001

[a] Scale direction (dependent variable): strongly agree first vs. strongly disagree first (reference group).

Table 4. Estimated Scale Direction Effect (Log Odds) on Content Latent Class Variables by Mode of Data Collection, 2012 American National Election Studies.

	Face-to-face			Web		
	$\hat{\beta}$	S.E.	p-value	$\hat{\beta}$	S.E.	p-value
F1	−0.07	0.33	0.83	0.04	0.19	0.81
F2	−0.2	0.33	0.53	−0.01	0.22	0.96
F3	−0.06	0.4	0.87	−0.06	0.24	0.81

The coefficients for both ERS ($\hat{\beta} = -0.17$, $p = 0.44$) and ARS ($\hat{\beta} = -0.28$, $p = 0.31$) are not significant in the face-to-face survey. This means that, although both types of response styles exist in the face-to-face survey, they do not differ by the direction of the scale. This is possibly due to the presentation channels (both visual and aural) in the face-to-face survey reduced the impact of design features such as scale direction. Also, the presence of interviewers might have also increased the respondents' motivation so that they were more likely to provide a careful answer and less likely to be influenced by the presentation of the scales. In the web survey, ERS also does not show a significant scale direction effect ($\hat{\beta} = -.1$, $p = .6$), while ARS does show a significant scale direction effect ($\hat{\beta} = 1.73$, $p < .0001$). Morren et al. (2011) found in their study that the three levels of ERS variable were not ordinally related to covariate variables. Therefore, we also analyzed the data by treating the ERS latent variable as a nominal variable. The result showed that scale direction was not significantly related to ERS.

When the positive option is presented first on the scale (i.e., agree-disagree scale), the odds of having ARS change are by a factor of 5.6 (=exp (1.73)) compared to the scale where the negative option is presented first (i.e., disagree-agree scale). As was described in the introduction, the examination of the ERS is exploratory and we did not have any expectation on how the scale direction could influence ERS. The significant scale direction effect reflect on ARS is possibly due to a combination of satisficing and anchoring-and-adjustment.

The literature shows that scale direction can affect content latent variables. However, previous studies do not control for response styles when testing scale direction effects on the substantive content latent class variables. Table 4 shows scale direction effects on the three content latent class variables after explicitly controlling for ARS and ERS in the LCA model. The effects on all three variables and in both models are small, and none are significant. Since earlier literature that examined the relationship between scale direction and substantive content variables did not control for response style, we also analyzed the data without adjustment for response styles. We reached the same conclusion: that scale direction was not significantly related to the content variables, without adjusting response styles (results not shown).

5. Discussion and Conclusion

The direction of rating scales and their influence on responses to attitudinal questions has attracted researchers' attention for more than half a century (Belson 1966). Previous research has primarily focused on the effect of scale direction on the response distribution of individual items, with occasional focus on latent variables. Response styles are one form of measurement bias that is frequently examined for multi-item rating scales. Some recent studies have examined the different scale design formats and their impacts on response styles (Kieruj and Moors 2013; Moors et al. 2014). However, no such effort has been devoted to testing the relationship between scale direction and response styles.

This study set out to test whether response styles, including ARS and ERS, differ by the direction of scales using an experiment embedded in a national probability survey conducted both face-to-face and on the web. Using latent class analysis, we reached the following conclusions:

1) ARS and ERS exist in both scale directions in both survey modes;
2) ERS is similar for scales presented in both directions in both face-to-face and web surveys;
3) ARS is significantly different between the two scale directions in the web surveys but not in the face-to-face surveys;
4) the scale direction does not impose a significant influence on the substantive content latent class variables, with or without controlling for response styles.

Several factors can explain the higher endorsement of the positive options in the agree-disagree scale than the disagree-agree scale in the web surveys. First, this could be interpreted as the result of respondents satisficing (Krosnick 1991) that leads to a primacy effect due to the more frequent selection of the option presented first, regardless of whether the option is positive or negative. Respondents might not read and process all scale points once they find a good enough response option. Assuming that satisficing might play a bigger role in self-administered survey modes than interviewer administration, this would also explain the nonfindings for the face-to-face survey.

Second, it is possible that respondents use the top-most response option as the anchoring point and then adjust their responses. Since the adjustment process is often insufficient, the anchoring point and the adjacent options are most likely to be endorsed. Both satisficing and anchoring-and-adjustment predict more selection of the position option when it is presented first than when it is presented last. Third, acquiescence predicts more selection of the positive option regardless of the scale direction. As a result, the overall shift of the distribution toward the positive end of the scale is most likely to be the consequences of a combination of the three mechanisms. Future research needs to be done to determine which of these explanations reflects the true cause of the scale direction effect.

The lack of scale direction effect in the face-to-face survey may be attributable to the channel of communication. In the face-to-face survey, as opposed to the web survey, both visual and aural communication is adopted during the interview. Such a design feature can alleviate the impact of scale direction and, hence, no such effect on response style latent class variables is detected from face-to-face surveys. In the web survey, questions and response options are only present visually. Thus, there is a higher likelihood that the respondent's anchoring point changes from one scale direction to another. Consequently, we observe a significant effect on ARS in the web survey. Note that this finding is the opposite of recency effects found in telephone interviews (Krosnick 1991). Furthermore, interviewers in face-to-face surveys can explain the purpose of the survey, answer questions, address concerns, and motivate respondents. Hence, respondent commitment is likely to be higher among face-to-face respondents than among web respondents. This can also be a reason for the lack of response style difference between the two scale directions, as respondents might just have been more committed to the task and less likely to be impacted by the design difference.

This study also finds that substantive latent class variables do not show a reliable scale direction effect, whether the response style latent class variables were controlled for or not. Previous research reported mixed results on the impact of scale direction on latent traits (Krebs and Hoffmeyer-Zlotnik 2010; Saris and Gallhofer 2007; Chan 1991), and this study provides one more piece of evidence that supports the lack of impact of the scale direction. Although the univariate response distribution shifted by the change of scale direction, it does not necessarily reflect a change in the substantial latent construct. Rather, it is likely to be a reflection of the change of response style latent variables.

This study probably raises more questions than it can answer. First, is it possible to generalize the findings in this study to other scale types? In this study, we only examine the scale direction effect using five-point agree-disagree Likert scales. A previous study has shown that other question types, such as item-specific scales, produce different response style patterns (Liu et al. 2015). Whether the scale direction influences other rating scales similarly should be examined in the future. Second, do other data collection modes exhibit a different scale direction effect? This study demonstrates that the scale direction effects are not identical between face-to-face and web surveys. Future studies should also explore and compare the scale direction effect on response styles among other survey modes (e.g., telephone, mobile web). Third, are the results replicable for other survey questions? Although response style is content-irrelevant, the scale direction effect on response style may differ regarding the question topic (e.g., sensitive vs. nonsensitive topics) or type (e.g., attitudinal vs. behavioral questions). We also find no scale direction effect on content latent class variables after controlling for the response styles. Future studies should attempt to replicate this finding and test whether it is possible to generalize. Last but not least, the scale direction may interact with the wording of the item. A positively worded item may show a different effect from a negatively worded item. We encourage future research to explore this possibility.

Regardless of the unsolved questions, this study demonstrates that the scale direction, a seemingly trivial survey design feature, can influence response style in web surveys. Researchers need to take the scale direction into serious consideration when designing rating scales. As many flagship national and international surveys move toward using the web, this is becoming a particularly relevant issue. The good news is that after controlling for response styles in the analysis model, the scale direction does not exert significant impact on the substantial latent class variables. Since in most cases the substantial, rather than the response style latent variables, are of interest, it is important to control for the response styles in the analysis model.

Appendix A
Question Wordings

The world is always changing and we should adjust our view of moral behavior to those changes. (TRAD1)
The newer lifestyles are contributing to the breakdown of our society. (TRAD2)
We should be more tolerant of people who choose to live according to their own moral standards, even if they are very different from our own. (TRAD3)
This country would have many fewer problems if there were more emphasis on traditional family ties. (TRAD4)

Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors. (BLAC1)
Generations of slavery and discrimination have created conditions that make it difficult for Blacks to work their way out of the lower class. (BLAC2)
Over the past few years, Blacks have gotten less than they deserve. (BLAC3)
It's really a matter of some people not trying hard enough; if Blacks would only try harder they could be just as well off as whites. (BLAC4)

Our society should do whatever is necessary to make sure that everyone has an equal opportunity to succeed. (EQUA1)
We have gone too far in pushing equal rights in this country. (EQUA2)
One of the big problems in this country is that we don't give everyone an equal chance. (EQUA3)
This country would be better off if we worried less about how equal people are. (EQUA4)
It is not really that big a problem if some people have more of a chance in life than others. (EQUA5)
If people were treated more equally in this country we would have many fewer problems. (EQUA6)

Appendix B

Demographic Distributions (means and standard errors [S.E.]) by Scale Direction in Face-to-face and Web Surveys, 2012 American National Election Studies.

	Face-to-face						Web					
	Forward		Reversed				Forward		Reversed			
	Mean	S.E.	Mean	S.E.	χ^2	p	Mean	S.E.	Mean	S.E.	χ^2	p
Male	47.98	2.18	48.00	2.14	0.00	0.99	47.29	1.52	48.48	1.49	0.31	0.58
Female	52.02	2.18	52.00	2.14			52.71	1.52	51.52	1.49		
<30	20.16	1.81	22.14	1.82	7.49	0.19	20.32	1.42	21.69	1.44	3.32	0.65
30–39	16.47	1.50	14.76	1.45			15.31	1.13	14.66	1.08		
40–49	15.59	1.52	18.80	1.76			16.01	1.08	17.37	1.13		
50–59	19.64	1.70	18.20	1.64			20.13	1.14	20.07	1.10		
60–69	16.59	1.77	12.21	1.38			16.30	1.01	16.24	0.94		
70<	11.56	1.48	13.89	1.58			11.93	0.93	9.96	0.79		
Non-Hispanic white	70.78	1.71	70.96	1.75	0.35	0.95	70.11	1.44	71.73	1.40	4.92	0.18
Non-Hispanic black	12.20	1.15	11.69	1.13			11.21	0.97	12.59	1.11		
Hispanic	11.04	1.00	10.75	1.06			11.76	1.04	10.76	0.92		
Other	5.98	0.88	6.60	1.02			6.92	0.83	4.92	0.64		
Less than high school	9.95	1.15	10.43	1.14	1.01	0.91	11.30	1.08	9.26	0.94	2.70	0.61
High school	29.23	2.05	31.46	2.08			29.33	1.47	30.35	1.43		
Some post-high school, no Bachelor's degree	31.09	1.99	29.25	1.89			30.29	1.34	30.51	1.38		
Bachelor's degree	19.83	1.79	18.74	1.72			18.21	1.13	19.10	1.11		
Graduate degree	9.91	1.29	10.13	1.31			10.88	0.82	10.78	0.83		
Married	53.76	2.15	52.78	2.12	1.02	0.91	53.91	1.52	52.62	1.50	4.51	0.34
Widowed	5.86	0.97	6.27	0.92			5.94	0.69	5.42	0.60		
Divorced	12.19	1.19	12.98	1.32			13.49	1.02	12.69	0.95		
Separated	2.94	0.71	2.22	0.40			2.67	0.51	1.96	0.37		
Never married	25.26	1.83	25.75	1.80			23.98	1.36	27.31	1.42		
0-49999	51.71	2.25	48.60	2.19	3.16	0.37	49.30	1.55	49.66	1.51	0.56	0.90
50000-99999	27.71	2.09	31.96	2.06			31.62	1.44	32.25	1.44		
100000-149999	11.08	1.58	12.19	1.67			11.83	0.98	11.57	0.90		
150000+	9.50	1.49	7.25	1.37			7.25	0.72	6.53	0.66		

7. References

Baumgartner, H. and J.E.M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38: 143–156. Doi: http://dx.doi.org/10.1509/jmkr.38.2.143.18840.

Belson, W.A. 1966. "Effects of Reversing Presentation Order of Verbal Rating Scales." *Journal of Advertising Research* 6: 30–37.

Carp, F.M. 1974. "Position Effects on Interview Responses." *Journal of Gerontology* 29: 581–587. Doi: http://dx.doi.org/10.1093/geronj/29.5.581.

Chan, J.C. 1991. "Response-Order Effect in Likert-Type Scales." *Educational and Psychological Measurement* 51: 531–540. Doi: http://dx.doi.org/10.1177/0013164491513002.

Christian, L.M., D.A. Dillman, and J.D. Smyth. 2008. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. De Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, R.L. Sangster, 250–275. New Jersey: John Wiley & Sons, Inc.

Dillman, D.A., T.L. Brown, J.E. Carlson, E.H. Carpenter, F.O. Lorenz, R. Mason, J. Saltiel, and R.L. Songster. 1995. "Effects of Category Order on Answers in Mail and Telephone Surveys." *Rural Sociology* 60: 674–687. Doi: http://dx.doi.org/10.1111/j.1549-0831.1995.tb00600.x.

He, L., T. Yan, F. Keusch and S. Han. 2014. "The impact of question and scale characteristics on scale direction effect." In Proccedings of the AAPOR 69th Annual Conference, Anaheim, California, May 15–18. Available at: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2014/Session-J-6-2-He-Y.pdf (accessed January 2017).

Kalton, G., M. Collins, and L. Brook. 1978. "Experiments in Wording Opinion Questions." *Applied Statistics*, 149–161. Doi: http://dx.doi.org/10.2307/2346942.

Kieruj, N.D. and G. Moors. 2010. "Variations in Response Style Behavior by Response Scale Format in Attitude Research." *International Journal of Public Opinion Research* 22: 320–342. Doi: http://dx.doi.org/10.1093/ijpor/edq001.

Kieruj, N.D. and G. Moors. 2013. "Response Style Behavior: Question Format Dependent or Personal Style?" *Quality & Quantity* 47: 193–211. Doi: http://dx.doi.org/10.1007/s11135-011-9511-4.

Krebs, D. and J.H.P. Hoffmeyer-Zlotnik. 2010. "Positive First or Negative First?: Effects of the Order of Answering Categories on Response Behavior." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6: 118–127. Doi: http://dx.doi.org/10.1027/1614-2241/a000013.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213–236. Doi: http://dx.doi.org/10.1002/acp. 2350050305.

Krosnick, J.A. and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research, Second Edition*, edited by P.V. Marsden and J.D. Wright, 263–314. United Kingdom: Emerald Group Publishing Limited.

Liu, M., F.G. Conrad, and S. Lee. 2016. "Comparing Acquiescent and Extreme Response Styles in Face-to-Face and Web Surveys." *Quality & Quantity*. Doi: http://dx.doi.org/10.1007/s11135-016-0320-7.

Liu, M., S. Lee, and F.G. Conrad. 2015. "Comparing Extreme Response Styles between Agree-Disagree and Item-Specific Scales." *Public Opinion Quarterly* 79: 952–975. Doi: http://dx.doi.org/10.1093/poq/nfv034.

Mingay, D.J. and M.T. Greenwell. 1989. "Memory Bias and Response-Order Effects." *Journal of Official Statistics* 5: 253–263.

Moors, G. 2003. "Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Reexamined." *Quality and Quantity* 37: 277–302. Doi: http://dx.doi.org/10.1023/A:1024472110002.

Moors, G. 2008. "Exploring the Effect of a Middle Response Category on Response Style in Attitude Measurement." *Quality & Quantity* 42: 779–794. Doi: http://dx.doi.org/10.1007/s11135-006-9067-x.

Moors, G., N.D. Kieruj, and J.K. Vermunt. 2014. "The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias." *Sociological Methodology* 44: 369–399. Doi: http://dx.doi.org/10.1177/0081175013516114.

Morren, M., J.P.T.M. Gelissen, and J.K. Vermunt. 2011. "Dealing With Extreme Response Style In Cross-Cultural Research: A Restricted Latent Class Factor Analysis Approach: Extreme Response Style In Cross-Cultural Research." *Sociological Methodology* 41: 13–47. Doi: http://dx.doi.org/10.1111/j.1467-9531.2011.01238.x.

Paulhus, D.L. 1991. "Measurement and Control of Response Bias." *Measures of Personality and Social Psychological Attitudes. Volume 1 in Measures of Social Psychological Attitudes Series*. Available at: http://doi.apa.org/psycinfo/1991-97206-001.

Revilla, M.A., W.E. Saris, and J.A. Krosnick. 2013. "Choosing the Number of Categories in Agree-Disagree Scales." *Sociological Methods & Research* 43: 73–97. Doi: http://dx.doi.org/10.1177/0049124113509605.

Salzberger, T. and M. Koller. 2013. "Towards a New Paradigm of Measurement in Marketing." *Journal of Business Research* 66: 1307–1317.

Saris, W.E. and I.N. Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." In *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, edited by W.E. Saris and I.N. Gallhofer, 237–253. New Jersey: John Wiley & Sons, Inc.

Stapleton, C. 2013. "The Smart (phone) Way to Collect Survey Data." *Survey Practice* 6(2).

Tourangeau, R., M.P. Couper, and F.G. Conrad. 2004. "Spacing, Position, and Order Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68: 368–393. Doi: http://dx.doi.org/10.1093/poq/nfh035.

Tourangeau, R., M.P. Couper, and F.G. Conrad. 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71: 91–112. Doi: http://dx.doi.org/10.1093/poq/nfl046.

Tversky, A. and D. Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185: 1124–1131. Doi: http://dx.doi.org/10.1126/science.185.4157.1124.

Vermunt, J.K. and J. Magidson. 2013. Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. Belmont, MA: Statistical Innovations Inc.

Yan, T. and F. Keusch. 2015. “The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey.” *Public Opinion Quarterly* 79: 145–165. Doi: http://dx.doi.org/10.1093/poq/nfu062.

Received June 2015
Revised August 2016
Accepted August 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 155–186, http://dx.doi.org/10.1515/JOS-2017-0009

Design of Seasonal Adjustment Filter Robust to Variations in the Seasonal Behaviour of Time Series

Marcela Cohen Martelotte[1], *Reinaldo Castro Souza*[2], *and*
Eduardo Antônio Barros da Silva[3]

Considering that many macroeconomic time series present changing seasonal behaviour, there is a need for filters that are robust to such changes. This article proposes a method to design seasonal filters that address this problem. The design was made in the frequency domain to estimate seasonal fluctuations that are spread around specific bands of frequencies. We assessed the generated filters by applying them to artificial data with known seasonal behaviour based on the ones of the real macroeconomic series, and we compared their performance with the one of X-13A-S. The results have shown that the designed filters have superior performance for series with pronounced moving seasonality, being a good alternative in these cases.

Key words: Moving seasonality; Filter design; Frequency domain; Time series decomposition; X-13A-S.

1. Introduction

Changing seasonality of time series was first noted in the nineteenth century (Gilbart 1852, quoted in Bell and Hillmer 1984), and is common in macroeconomic data (Canova and Ghysels 1994; Wells 1997; Franses and Koehler 1998; Van Dijk et al. 2003). Such changes can be due to variations in seasonal amplitude from year to year or in the proportionality relationship between the seasonal at each month and the seasonal at each other month (i.e., the seasonal pattern) (Godfrey and Karreman 1964). We refer to them as 'moving seasonality'.

Kuznets (1932) was among the first authors to highlight the importance of moving seasonality. Since then, statistical tests have been created to evaluate the presence of changing seasonal behaviour (Higginson 1975; Canova and Hansen 1995; Sutradhar and Dagum 1998) and several seasonal adjustment methods have been suggested to tackle it, many of them developed in the frequency domain. Among the frequency domain approaches, we highlight the pioneering work of Hannan (1964), Nerlove (1964), and Nettheim (1964).

X-13ARIMA-SEATS (X-13A-S) is the most recent enhanced version of the 'X-11 family' (U.S. Census Bureau 2013). This program contains two seasonal adjustment

[1] Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro. Rua Marquês de São Vicente, 225 - Gávea, Rio de Janeiro, Brazil. 22451-900. Email: marcela.cohen@prof.iag.puc-rio.br
[2] Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro. Rua Marquês de São Vicente, 225 - Gávea, Rio de Janeiro, Brazil. 22451-900. Email: reinaldo@ele.puc-rio.br
[3] Electrical Engineering Program, Federal University of Rio de Janeiro. C.P. 68504, Rio de Janeiro, Brazil. 21941-972. Email: eduardo@smt.ufrj.br

modules: the X-11 method and the SEATS. The latter is a seasonal adjustment procedure that follows the ARIMA model-based signal extraction technique (Gómez and Maravall 1996). The former module is the X-11, or Census X-11, one of the most commonly used methods for seasonal adjustment of economic time series used by government agencies and statistical bureaus. This method, based on moving averages, was introduced in 1965 by the U.S. Census Bureau (Shiskin et al. 1967) and further contributions have been added to the basic version (Dagum 1980; Dagum 1988; Findley et al. 1998). It is important to mention that these methods are also implemented in JDEMETRA+, which is the software officially recommended by Eurostat and the European Central Bank for the seasonal and calendar adjustment of official statistics in the European Union. The ESS guidelines on seasonal adjustment (Eurostat 2015) highlight the unstable seasonality problem, warning that the standard seasonal adjustment cannot be used in this case.

In the literature, there are several works comparing X-11 with other methods of seasonal adjustment, especially with SEATS (Hood et al. 2000; Findley 2005; Tiller et al. 2007). The results point to similar performance when the time series presents common seasonal behaviour. However, in cases of data with moving seasonality, the X-11 method has some drawbacks (Planas 1998; Kaiser and Maravall 2000; Maravall and Pérez 2011).

Nettheim (1965) listed strategies for dealing with moving seasonality. One of them is a filter designed to have unit gain around each seasonal frequency and very small gain elsewhere. As pointed out by the author, a drawback of such a method is that one should determine in advance how wide the unit gain region should be. A way of circumventing this problem is to use spectral estimation methods. Examples are the non-ad-hoc methods in Melnick and Moussourakis (1974) and Geweke (1978). The ARIMA model-based approach of SEATS (Gómez and Maravall 1996) is another attempt to treat moving seasonality, but in some cases it is not trivial to find a good-fitting model with a valid decomposition into components (Tiller et al. 2007). The Structural Models (STM) (Koopman et al. 2000) can deal with moving seasonality via a sophisticated model-based approach that requires expert operators. However, the simplicity of the seasonal adjustment programs is sometimes preferred to seasonally adjusting a large number of series.

In this context, considering that the seasonal adjustment is largely used in the production of official statistics, we propose a methodology to design seasonal filters to deal with moving seasonality. The design of such filters, which we refer to as Seasonal-WLS (S-WLS), is based on least squares criteria in the frequency domain. The design is inspired by the requirements set forth in Nettheim (1965). We assess the performance of the proposed S-WLS filters by running them on artificial data derived from the behaviour of the real macroeconomic time series. Then, we compared its adjustment with the adjustment of the X-11 method. We make this comparison because, besides the fact that X-11 tends to misadjust series with highly moving seasonality, it is ad-hoc and has been one of the most widely used seasonal adjustment methods.

This article is organised as follows. Section 2 briefly describes the theoretical framework of the X-11 method, as well as the frequency domain representation of its filters. Section 3 presents the proposed method to design the S-WLS filters for seasonal adjustment, describing its structure and the parameters' choice. Section 4 shows the results of the application of the proposed S-WLS filters, comparing their performance with that of X-11. Finally, Section 5 summarises and discusses the main findings. Appendix A presents the X-11 algorithm in the frequency domain, and Appendix B and C present, respectively, the details about the selection of the filter parameters and about the signal-to-noise ratio (SNR) computation.

2. The X-11 Method in the Frequency Domain

The X-13ARIMA-SEATS (X-13A-S) program contains the implemented X-11 seasonal adjustment method. This method, as well as other programs of the ‘X-11 family’, consists of a moving average procedure for seasonally adjusting series (it is fully explained in Findley et al. 1998). The frequency domain properties of the X-11 method have been discussed in the works of Wallis (1982), Bell and Monsell (1992), Dagum et al. (1996), Gómez and Maravall (2001), Findley and Martin (2006) and others. Here, the X-11 procedure will be briefly discussed for the purpose of introducing its transfer function, which will be instrumental in analysing the X-11 behaviour in the presence of moving seasonality.

The seasonal adjustment filters of X-11 are available in X-13A-S (U.S. Census Bureau 2013), X-12-ARIMA (Findley et al. 1998) and X-11-ARIMA (Dagum 1980). In the literature, the hybrid name ‘X-11/12-ARIMA filters’ was adopted to designate these filters (Findley and Martin 2006). In this work, they will be referred to just as ‘X-11 filters’.

The step-by-step application of the X-11 method (default setting) can be summarised in two stages, for seasonal factor and seasonal adjustment (Findley et al. 1998). In the default procedure, it specifies a 3 × 3 seasonal moving average (usually called ‘seasonal filter’), $M_{3\times3}$, for the initial seasonal factor estimates, and the 3 × 5 seasonal filter ($M_{3\times5}$) thereafter. It also prefilters the input series with a 2 × 12 moving average, ($M_{2\times12}$). The whole procedure is depicted in Figure 1, considering the additive decomposition of monthly series, where: Y is the original time series; T is the trend estimate; SI is the estimate of the seasonal-irregular; $\hat{S}$ is the preliminary seasonal factor; S is the seasonal factor; A is the seasonally adjusted time series. The superscript$^{(1)}$ means the initial estimate and the superscript$^{(2)}$ refers to the final one. H_{13} is the 13-term Henderson trend filter.

The coefficients of the seasonal filters present in X-11, as well as the 2 × 12 moving average, are listed in Tables 1 and 2.

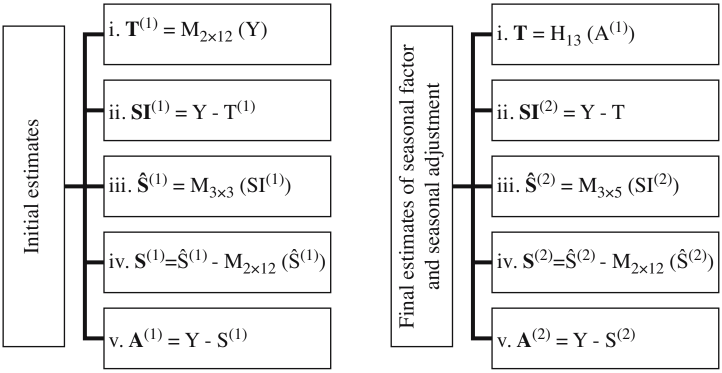

Initial estimates
i. $\mathbf{T}^{(1)} = M_{2\times12}(Y)$
ii. $\mathbf{SI}^{(1)} = Y - T^{(1)}$
iii. $\hat{\mathbf{S}}^{(1)} = M_{3\times3}(SI^{(1)})$
iv. $\mathbf{S}^{(1)}=\hat{S}^{(1)} - M_{2\times12}(\hat{S}^{(1)})$
v. $\mathbf{A}^{(1)} = Y - S^{(1)}$
Final estimates of seasonal factor and seasonal adjustment
i. $\mathbf{T} = H_{13}(A^{(1)})$
ii. $\mathbf{SI}^{(2)} = Y - T$
iii. $\hat{\mathbf{S}}^{(2)} = M_{3\times5}(SI^{(2)})$
iv. $\mathbf{S}^{(2)}=\hat{S}^{(2)} - M_{2\times12}(\hat{S}^{(2)})$
v. $\mathbf{A}^{(2)} = Y - S^{(2)}$

Fig. 1. X-11 default procedure for seasonal adjustment considering the additive decomposition of monthly series.

Table 1. Coefficients of the X-11 Seasonal Filters ($m(n) = m(-n)$): $m(n)$ are the coefficients of filter $M_{PXQ}(z)$ (see Equation (A.1) in Appendix A).

Seasonal filters	m(5)	m(4)	m(3)	m(2)	m(1)	m(0)
3 × 9	1/27	2/27	3/27	3/27	3/27	3/27
3 × 5			1/15	2/15	3/15	3/15
3 × 3				1/9	2/9	3/9

In the automatic selection procedure, the program may replace the 3 × 5 seasonal moving average filter in step (iii) of the 'final estimates of seasonal factor' in Figure 1, by either a 3 × 3 or a 3 × 9 seasonal filter (Findley et al. 1998; U.S. Census Bureau 2013). Regarding the Henderson trend filter, the program selects a trend moving average based on statistical characteristics of the data. For monthly data, either a 9-, 13-, or 23-term Henderson trend filter can be selected by the automatic procedure.

The procedure used to compute the seasonally adjusted series ($A^{(2)}$) described in Figure 1 can be expressed in the frequency domain by the following expression:

$$\begin{aligned} A^{(2)} = Y(z)\{1 - \mathrm{M}_{3\times5}(z^{12})[1 - \mathrm{M}_{2\times12}(z)][1 - H_{13}(z) \\ \times \{1 - [1 - \mathrm{M}_{2\times12}(z)]^2 \mathrm{M}_{3\times3}(z^{12})\}]\} \end{aligned} \tag{1}$$

where the functions of z are the z-transforms of the corresponding filters. The coefficients $m(n)$ for each filter are listed in Tables 1 and 2. Detailed explanation about this expression is given in Appendix A.

The expression in Equation (A.12) provides a useful way to evaluate the transfer function of the various X-11 filters, both for the default and the optional choices in the automatic procedure. Figure 2 shows the magnitude of the transfer functions of the X-11 for the three types of seasonal moving average, considering a 13-term Henderson filter and monthly data.

Figure 2 illustrates the fact that the smaller the size of the seasonal filter, the wider its passband width is, and the more suitable it is for treating moving seasonality data (for more details, see Subsection 3.2). However, even the smallest seasonal filter in the automatic option of the seasonal adjustment (3 × 3) does not have a large enough passband in order to deal with moving seasonality. The X-11 method also provides the possibility of using a three-term moving average filter, although it is not available in the automatic procedure; in addition, it produces a transfer function with a poor attenuation in the stopband.

Table 2. Coefficients of the X-11 Moving Average Filter ($m(n) = m(-n)$): $m(n)$ are the coefficients of filter $M_{PXQ}(z)$ (see Equation (A.1) in Appendix A).

Moving average	m(6)	m(5)	m(4)	m(3)	m(2)	m(1)	m(0)
2 × 12	1/24	1/12	1/12	1/12	1/12	1/12	1/12

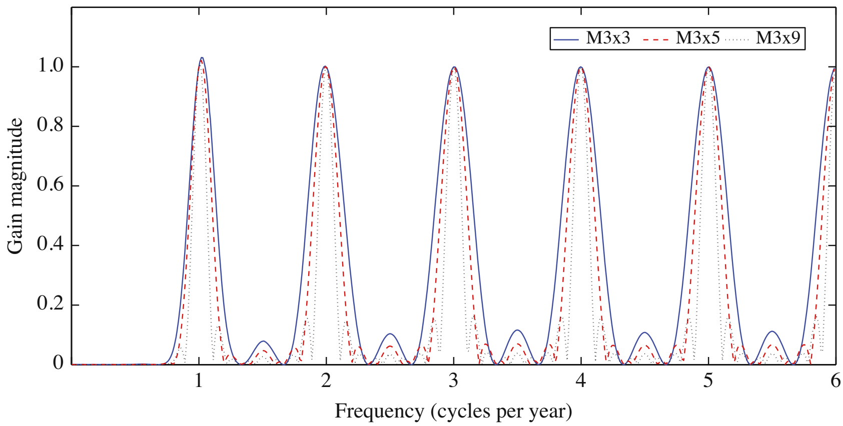


Fig. 2. Magnitude of the transfer function of the X-11 for different seasonal filters: monthly series.

In this article we propose a design method that generate filters with bandwidths that are large enough so that they can deal with the most common seasonality variations, without compromising the filter attenuation outside the seasonal frequencies. The construction of this filter will be presented in the next section.

3. The Proposed Seasonal Filter: a Frequency Domain Moving Seasonal Filter to Deal With Moving Seasonality

To introduce the proposed seasonal filters, we assume that a monthly observable time series at time t, $Y(t)$, can be represented as follows:

$$Y(t) = T(t) + S(t) + I(t), (t = 1, 2, \ldots) \tag{2}$$

where $Y(t)$ is the original time series, $T(t)$, $S(t)$, and $I(t)$ are unobservable trend-cycle (treated here as 'trend'), seasonal and irregular components.

From a frequency domain point of view, a filter designed to extract the seasonality should be able to isolate the movements in the series which occur in the seasonal frequency and in its harmonics, usually called 'seasonal frequencies': ($2\pi/12$, $4\pi/12, \ldots, 12\pi/12$). However, when the series has moving seasonality, its spectral mass is not restricted to the seasonal frequencies, but is spread around their neighborhoods. Considering this, we want a filter with frequency response equal to one in the bands around the seasonal frequencies (passbands) and zero in the remaining frequencies (stopbands). This is one of the filters mentioned by (Nettheim 1965), illustrated in Figure 3.

This filter has the objective of not disturbing the frequency components around the harmonics of the seasonal frequency, and to this end its transfer function has a flat shape in a neighbourhood of width Δ around these frequencies, as shown in Figure 3. This is important in the case of moving seasonality. An example that illustrates this situation is given by the following time series, composed by an irregular component and a seasonal

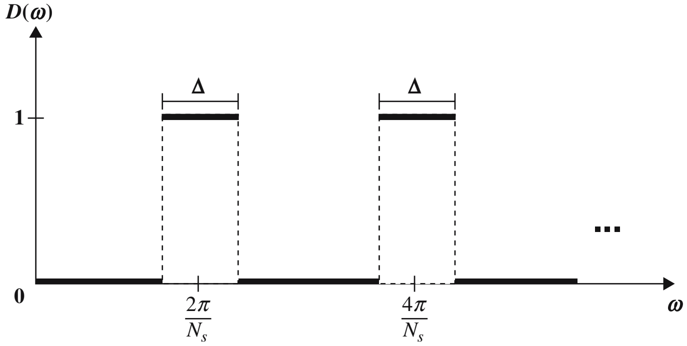


Fig. 3. Magnitude of the transfer function of the ideal filter.

component with nonstationary changes:

$$Y(t) = \sum_{i=1}^{P} \left[1 + b\sin\left(2\pi\frac{(t-iQ)}{k_i}\right)\right]\left[\sin 2\pi\frac{t}{2}\right] + I(t) \tag{3}$$

where $b = 0.9$, $Q = 120$, k_i are samples from a random variable uniformly distributed in the interval [70, 240] and $I(t)$ is the irregular component, an independent zero-mean Gaussian process.

The spectrum of this time series is shown in Figure 4. It is possible to note that the seasonal component has significant energy over a bandwidth of approximately 0.03 around the frequency 1/12 cycles/month. In order not to attenuate the frequency components that deviate from the harmonic of the seasonal frequencies, the gain of the filter should be constant for all the frequencies in the neighbourhood of the seasonal frequencies. This block shape has been suggested by Nettheim (1965).

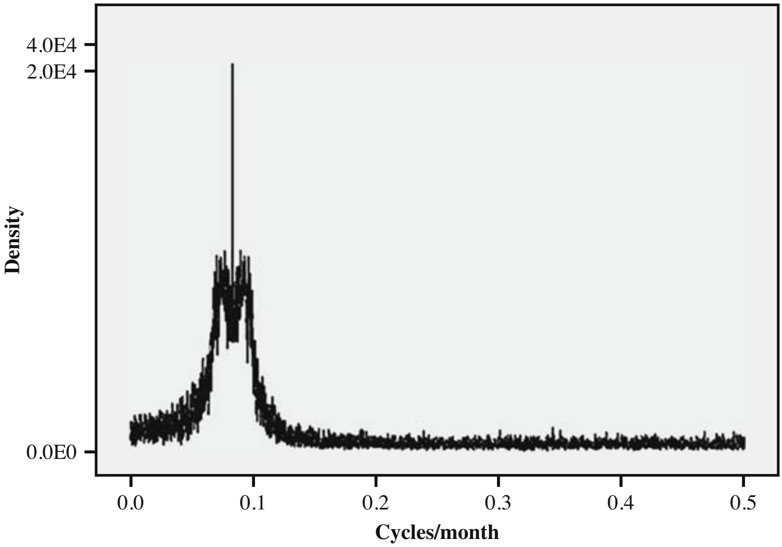


Fig. 4. Spectral density of Y(t) by frequency.

Such a filter can be designed to accommodate the different kinds of seasonality variations. These can be seen as combinations of variations in the seasonal periods as well as variations in the seasonal amplitude. We should take these into account when computing the filter parameters. To this end, one can express a seasonal component with frequency ω_s and moving seasonality as

$$s(t) = [1 + a(t)]h(t), \tag{4}$$

where h is a periodic function with period $(\omega_s + \Delta\omega)$ and $\Delta\omega$ may vary with time t; $a(t)$ represents the seasonal amplitude variation.

Considering that the rate of variation of $\Delta\omega$ is much smaller than the seasonal period, $h(t)$ can be expressed by a Fourier series as:

$$h(t) = \sum_n c_n e^{jn[\omega_s + \Delta\omega(t)]t} \tag{5}$$

From Equation (4), the seasonal signal in the frequency domain can be written as:

$$S(\omega) = \sum_n 2\pi c_n \delta(\omega - n\omega_s - n\Delta\omega) + \sum_n 2\pi c_n A(\omega - n\omega_s + n\Delta\omega) \tag{6}$$

where $A(\omega)$ is the Fourier transform of $a(t)$ and $\delta(\omega)$ is the Dirac delta function.

Equation (6) is depicted in Figure 5, where the arrows indicate the Dirac delta functions and the bell-shaped functions are repetitions of $A(\omega)$ centred at the frequencies $n\omega_s + n\Delta\omega$. From this, since the bandwidth of $a(t)$ is given by B, and considering that one has to account for up to the nth harmonic component, the width Δ from each of the filter's passbands has to be larger than:

$$\Delta = 2\max\{n\Delta\omega + B, -n\Delta\omega + B\} = 2(n|\Delta\omega| + B) \tag{7}$$

Therefore, to perform the seasonal adjustment it is necessary to determine the filter design parameter Δ, that is, a function of the seasonal behaviour of the series. In practice, this can be done, for example, by computing the series spectrogram in order to determine ω_s, $\Delta\omega$, and B.

Regarding what was discussed in this subsection, we propose a methodology to design seasonal filters that are able to deal with moving seasonality, that is, with transfer functions

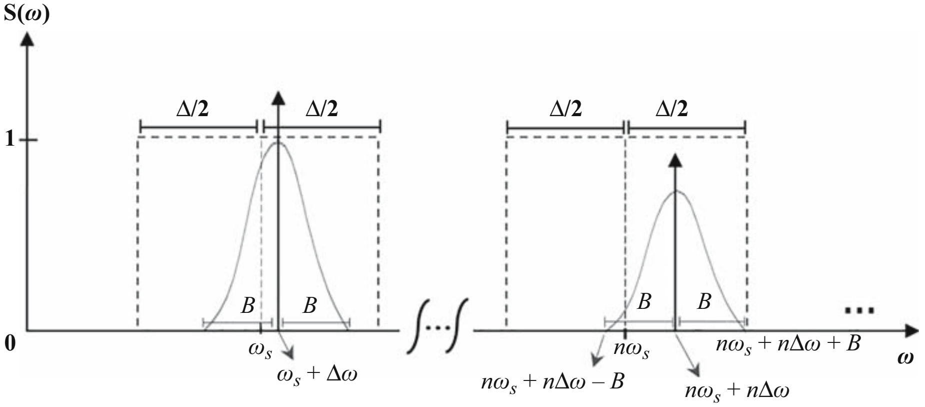


Fig. 5. Magnitude of the ideal filter transfer function and spectrum of the seasonal signal from Equation (6).

that approximates the one in Figure 3. We refer to them as S-WLS (Seasonal Weighted Least Squares) filters. These filters are finite and symmetric, designed in the frequency domain, and can be applied to a monthly seasonal time series, independent of its distribution.

Applying the proposed S-WLS filters, the data will be seasonally adjusted by eliminating the trend component and performing the seasonal extraction using a single filtering operation. The following subsection presents the design of the S-WLS filters. The theory used in this section is based on Diniz et al. (2010).

3.1. The Structure of the S-WLS Filters

First of all, to extract the seasonality, it is necessary to also eliminate the trend component (Hassani 2007; Cleveland et al. 1990; Burman 1980).

With this purpose in mind, the z-transform of the filter frequency response should have a term of the form $(1 - z^{-1})^{j+1}$, that accounts for eliminating a trend polynomial up to order j.

Therefore, we can represent the S-WLS filters, to extract the seasonal component, by the following z-transform:

$$P(z) = (1 - z^{-1})^{j+1} G(z) \tag{8}$$

where $G(z)$ is defined as

$$G(z) = \sum_{t=-p}^{L-p-1} g(t) z^{-t}. \tag{9}$$

In Equation (9), L is the number of degrees of freedom of the filter, given by the coefficients $g(t) \in \mathbb{R}$; the filter length is $(L + j + 1)$; p gives a shift in the filter output, and for a filter with zero delay, it should be equal to $(L + j + 1)/2$. The index t represents the time period $(t = 1, 2, \ldots)$.

The coefficients $g(t)$ of $G(z)$ must be optimised so that the frequency response of the filter can approximate the desired frequency response $D(\omega)$. In other words, $G(z)$ is adjusted so that the resulting filter can approximate the one from Nettheim (1965) (illustrated in Figure 3) with bandwidths around the harmonics of the seasonal frequency as flat as possible. Besides, to make the filter robust to variations in seasonality, these coefficients must be optimised to consider the seasonal variation around the harmonics in a frequency range corresponding to a percentage of the seasonal frequency. Moreover, it must suppress as much of the irregular component as possible. Thus, in the S-WLS design, the passbands have a desired frequency response $(D(\omega))$ equal to one and the stopbands have a desired frequency response equal to zero.

In addition, in order to help in the optimisation process, we introduce 'don't care' bands, where the desired response is not specified, between each adjacent passband and stopband. Their width is adjusted experimentally so that the obtained frequency response is as close as possible to the desired one. These design parameters are illustrated in Figure 6, where N_s indicates the assumed seasonal period, which is twelve for monthly data.

As can be observed in Figure 4, the filter is robust to seasonality variations up to a fraction $(\alpha/2)$ of the assumed seasonal frequency.

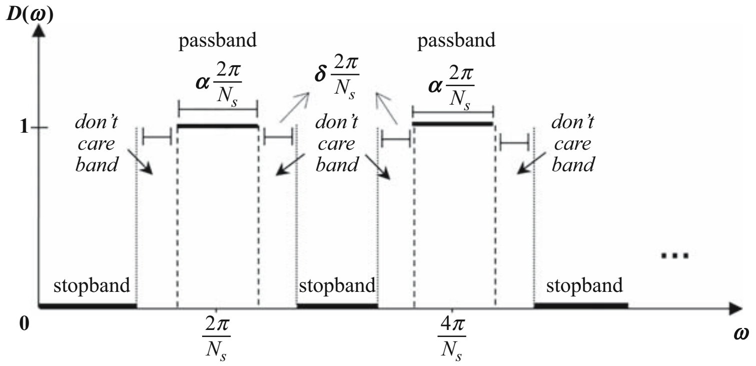


Fig. 6. Magnitude of the transfer function of the proposed S-WLS filter.

The filter coefficients are obtained by an optimisation process, which minimises the Euclidean distance between the desired frequency response $D(\omega)$ and the filter frequency response $P(e^{i\omega})$.

Since the frequency response of the filter can be computed from its z-transform by making $z = e^{i\omega}$ (Diniz et al. 2010), we have that, from Equation (8), it becomes

$$P(e^{i\omega}) = (1 - e^{-i\omega})^{j+1} G(e^{i\omega}) \tag{10}$$

$$= e^{-i\left(\frac{\omega}{2}\right)(j+1)}\left(2i\sin\frac{\omega}{2}\right)^{j+1} G(e^{i\omega})\,. \tag{11}$$

Since $G(z)$ can be written as

$$G(e^{i\omega}) = e^{-i\omega(p)}\mathbf{E}^{\mathrm{T}}(\omega)\mathbf{g} \tag{12}$$

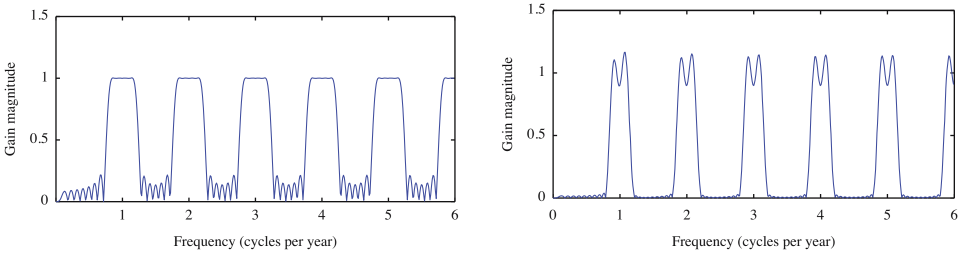
where $\mathbf{E}(\omega) = \left[1 \;\; e^{-i\omega} \;\; \ldots \;\; e^{-i\omega(L-1)}\right]^{\mathrm{T}}$ and $\mathbf{g} = \left[g(-p) \;\; g(-p+1) \;\; \ldots \;\; g(L-p-1)\right]^{\mathrm{T}}$, the filter frequency response becomes

$$P(e^{i\omega}) = e^{-i\left(\frac{\omega}{2}\right)(j+1)-i\omega(p)}\left(2i\sin\frac{\omega}{2}\right)^{j+1}\mathbf{E}^{\mathrm{T}}(\omega)\mathbf{g} \tag{13}$$

$$= s(\omega, j, p)\mathbf{E}^{\mathrm{T}}(\omega)\mathbf{g}. \tag{14}$$

where $s(\omega, j, p) = e^{-i\omega\left[\left(\frac{j+1}{2}\right)+p\right]}\left(2i\sin\frac{\omega}{2}\right)^{j+1}$.

In the optimisation process, we will discretise the frequency variable ω in the passband and in the stopband. Thus, it is relevant to consider the possibility that the errors in the passbands and in the stopbands have different importance. To allow this in the optimisation process, we assign a weight $W(\omega)$ to each frequency. It establishes the relative importance of the frequency response at each frequency ω during the optimisation. For example, if we assign a higher importance to the error in the passband, the transfer function would tend to be like the one in Figure 7a. In contrast, if the importance of attenuation in the stopband is much higher than the one in the passband, we would tend to have the transfer function like the one in Figure 7b. As can be observed, the transfer function may change considerably.



(a) (b)

Fig. 7. Magnitude of the transfer function of the S-WLS filter when $N = 145$, $\alpha = 1/3$, $\delta = 1/30$ and (a) $w_0 = 30$. (b) $w_0 = 0.05$.

Formally, such frequency response weighting is equivalent to minimising the average of the weighted squared error below:

$$|\mathrm{e}_{\mathrm{r}}(\omega)|^2 = |[P(\omega) - D(\omega)]W(\omega)|^2 \tag{15}$$

where $D(\omega)$ is the desired frequency response (see Figure 6).

In order to perform this minimisation we discretise ω at the set of frequencies $(\omega_1, \omega_2, \ldots, \omega_n)$. The number n of frequency samples is equal to $401N$, where N is the filter order. Therefore, each of the functions of ω can be represented as a column vector consisting of the samples of the function at the discrete set of frequencies. For example, we represent $P(\omega)$ as

$$\mathbf{P} = \begin{bmatrix} P(\omega_1) & P(\omega_2). \; . \; . & P(\omega_n) \end{bmatrix}^{\mathrm{T}}. \tag{16}$$

Using this notation, Equation (14) is equivalent to:

$$\mathbf{P} = \mathbf{Ug} \tag{17}$$

where $\mathbf{P}$ is defined in Equation (16) and the matrix $\mathbf{U}$ of dimensions $n \times L$ is defined as

$$\mathbf{U} = \begin{bmatrix} \mathbf{E}^{\mathrm{T}}(\omega_1)s(\omega_1, j, p) \\ \mathbf{E}^{\mathrm{T}}(\omega_2)s(\omega_2, j, p) \\ \cdots \\ \mathbf{E}^{\mathrm{T}}(\omega_n)s(\omega_n, j, p) \end{bmatrix}. \tag{18}$$

If the samples of error (ω) and the desired frequency response $D(\omega)$ are represented analogously as column vectors, and we define

$$\mathbf{W} = \begin{bmatrix} \mathrm{w}_1(\omega_1) & 0 & 0 & 0 \\ 0 & \mathrm{w}_2(\omega_2) & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathrm{w}_n(\omega_n) \end{bmatrix}, \tag{19}$$

then, from Equation (17) to (19), Equation (15) can be expressed in matrix form as

$$\mathbf{e_r} = \mathbf{W}[\mathbf{P} - \mathbf{D}] = \mathbf{W}[\mathbf{U}\mathbf{g} - \mathbf{D}]. \tag{20}$$

The sum of squared errors can be written as

$$\|\mathbf{e_r}\|_2^2 = \|\mathbf{e}_\mathbf{r}^\mathrm{T}\mathbf{e_r}\| = (\mathbf{U}\mathbf{g} - \mathbf{D})^{*\mathrm{T}}\mathrm{W}^{*\mathrm{T}}\mathrm{W}(\mathbf{U}\mathbf{g} - \mathbf{D}) \tag{21}$$

which is minimized by the vector

$$\mathbf{g} = \left(\mathbf{U}^{*\mathrm{T}}\mathbf{W}_s^2\mathbf{U} + \mathbf{U}^\mathrm{T}\mathbf{W}_s^2\mathbf{U}^*\right)^{-1}(\mathbf{U}^\mathrm{T} + \mathbf{U}^{*\mathrm{T}})\mathbf{W}_s^2\mathbf{D}. \tag{22}$$

Convolving the vector **g** with the coefficients of the polynomial $(1 - z^{-1})^{j+1}$ (Equation (8)), we obtain the vector with the S-WLS filter coefficients. The filtering operation is accomplished by convolving the vector of the S-WLS coefficients with the time series. The output of this operation is the extracted seasonal component. The adjusted series is obtained by subtracting this result from the original series.

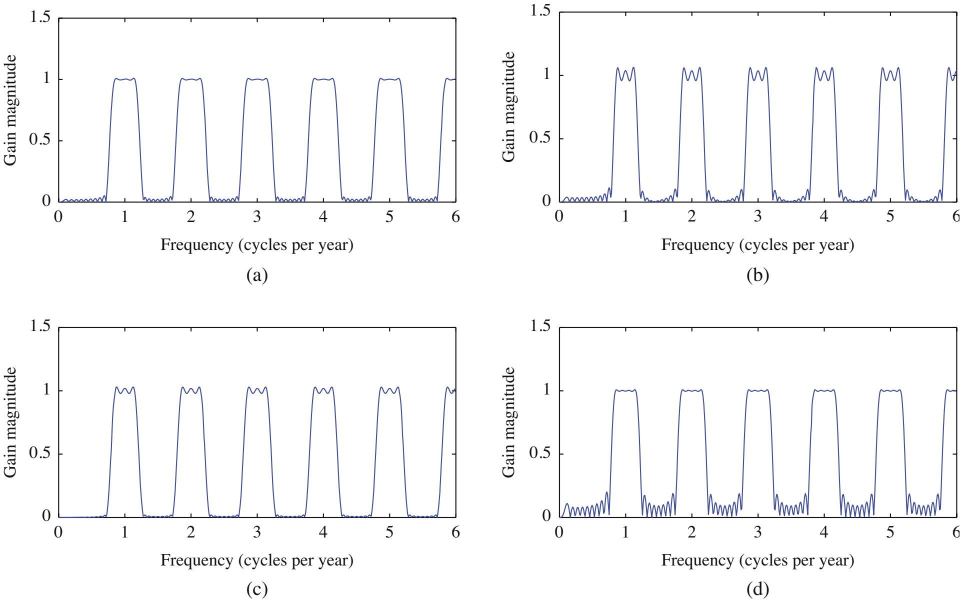
The S-WLS filters have five design parameters (see Figure 6):

(i) the parameter α is equivalent to the bandwidth around the seasonal frequencies, being related to the seasonal stability (it depends on the data characteristics);
(ii) the parameter δ is related to the width of the 'don't care' band, helping in the optimisation process;
(iii) the weight (w_0) indicates the importance given to the error minimisation in the passbands compared with the one in the stopbands – large values of weight (w_0) result in gain close to 1 around the seasonal frequencies, but the attenuation outside the seasonal frequencies decreases;
(iv) the filter size N, representing the number of coefficients of the filter;
(v) the number of frequency samples used during the optimisation. In the filter experiments we used $401N$ because it was shown to be enough to provide a good approximation.

Considering a fixed value for the parameter N, and to a given α, different δ and w_0 lead to considerable changes in the filter transfer function. Figures 8a to 8d show the transfer function for some values of the parameters δ and w_0 for $\alpha = 1/3$ and $N = 169$.

The designed filters should have as much attenuation as possible in the stopband and as little ripple as possible in the passband. Analysing the transfer functions in Figures 8a and 8b, we can see that, for a given w_0, a larger δ (that is, a larger transition, or 'don't care', band), allows a smaller ripple in the passband, as well as a larger attenuation in the stopband. On the other hand, more of the irregular component can leak through a larger transition band, yielding a filter that tends to overadjust the seasonality. Consider now a given δ, the bigger the w_0, the smaller the ripple in the passband (see Figures 8a, 8c, Figure 8b and 8d), but the attenuation in the stopband gets worse.

Being aware that different values of the filter parameters result in distinct transfer functions, it is important to have a methodology for choosing their values. In this work, we adopted the strategy of analysing the performance of the filter when applied to artificial series with behaviour similar to the one of real macroeconomic series. This issue will be dealt with in the following subsection.



(a)

(b)

(c)

(d)

Fig. 8. Magnitude of the transfer function of the S-WLS filter when N = 169, α = 1/3 and (a) δ = 1/10 and w_0 = 1. (b) δ = 1/100 and w_0 = 1. (c) δ = 1/10 and w_0 = 0.3. (d) δ = 1/100 and w_0 = 10.

3.2. The Choice of the S-WLS Filter Parameters

As mentioned at the end of last subsection, we choose the filter parameters by analysing their performance when filtering simulated series. These artificial series should have features similar to the ones of real macroeconomic series. Therefore, our first step was to analyse real macroeconomic data, for the purpose of identifying the behaviour of their moving seasonality. From 144 monthly macroeconomic time series analysed, 53% showed changing seasonal behaviour, according to the F-test for moving seasonality implemented in X-13A-S program, considering a p-value $<5\%$. Those series are listed in SMT-UFRJ (2014), and were obtained from the OECD (2014), the U.S. Census Bureau (2014), the U.S. Bureau of Labour Statistics (2014), the IPEA (2014), and the IBGE (2014).

An example of a time series with moving seasonality is the ‘USA Employment Level’ (p-value $<0.1\%$), from the U.S. Bureau of Labour Statistics (jan/93 to sep/2013). Its seasonal component, adjusted using X-13A-S program, is shown in the plot in Figure 9.

As can be observed from Figure 9, this seasonal component changes its amplitude and shape over the months, confirming the changing seasonality. In order to generate monthly artificial seasonal components with similar behaviour, we used a sinusoidal series whose amplitude is modulated by another sine wave, as follows:

$$S(t) = A\left[1 + b\sin\left(2\pi\frac{t}{k}\right)\right]\left[\cos\left(2\pi\frac{t}{12}\right)\right] \tag{23}$$

where: A is the seasonal amplitude ($A \in \mathbb{R}$); b is related to the rate of change in the signal amplitude ($b \in (0,1)$), k is related to the change in the seasonal pattern, and t is the time index ($t = 1, 2, \ldots$).

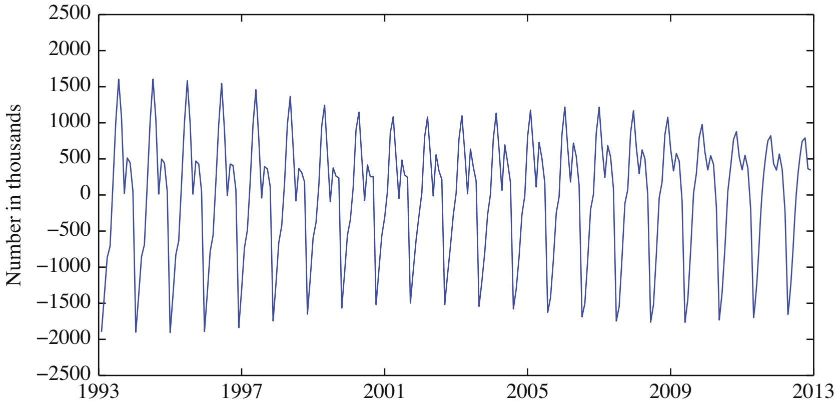


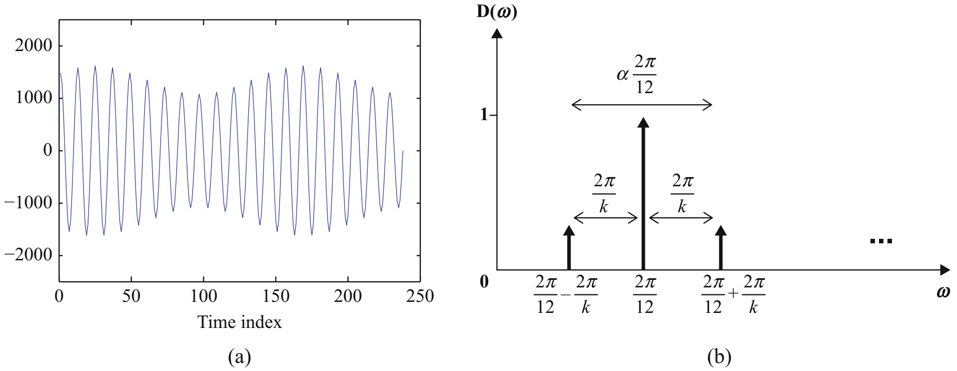
Fig. 9. Seasonal component of the 'USA Employment Level' (jan/93 to sep/2013).

A seasonal component represented by Equation (23) is illustrated in the time and frequency domains in Figures 10a and 10b. Its parameters are: $A = 1350$, $b = 20\%$ and $k = 144$. In Figure 10a it is possible to identify the amplitude change of the seasonal component, and in Figure 10b we notice that the variations in seasonality appear as two sinusoidal components distant $\pm 2\pi/k$ rad/month from the seasonal frequency, that is $2\pi/12$ rad/month. Note that the α parameter of the filter (see Figure 6) must be such that it can properly deal with seasonal frequency variations, that is

$$\alpha \frac{2\pi}{N_s} \geq \frac{4\pi}{k} \tag{24}$$

where N_s is the the seasonal period, that is twelve for monthly data and four for quarterly data.

It is important to note that the parameter α of the filter gives an upper bound to the maximum variation of the seasonal frequency that the filter is able to handle. However, since the filter design method is deterministic, in a practical application one could perform a spectral analysis of the time series prior to the seasonal adjustment in order to estimate



(a)

(b)

Fig. 10. Seasonal component (Equation (23)) (a) in time domain, (b) in frequency domain.

the bandwidth of the seasonal variation. With this estimate, one could design a filter that would have the α parameter large enough to handle the amount of seasonal variation present in the time series.

An artificial seasonal signal such as the one in Equation (23) can be used to evaluate the response of the filter for different levels of variation in seasonality, either in amplitude or frequency.

To determine the appropriate α for the filter, we analysed the seasonal component of a wide range of macroeconomic time series with moving seasonality. This analysis led us to $\alpha = 1/3$ as a good compromise, that is appropriate for most of the analysed series (in time domain, $\alpha = 1/3$ corresponds approximately to a range between ten and 14 months). Note that, in order to accommodate as much variation on the seasonal component as possible, α should be as large as possible. However, we cannot increase α too much, because the larger the α, the larger the leak of the irregular component through the passbands around the seasonal components, which increases the error in seasonality estimation.

After this, we found a good combination of the parameters δ and w_0, based on the seasonal adjustment of artificial data containing moving seasonal behaviour (Equation (23)), trend and irregular components with the same features of the real time series listed in the webpage in SMT-UFRJ (2014). As observed in Figure 8, for a given α, we have to vary δ and w_0 to find a good compromise between the attenuation in the stopband and the ripple in the passband. In other words, these parameters are responsible, respectively, for the error in the seasonality estimation for the noiseless case (no irregular component) and for the error due to the irregular component, as discussed in Subsection 3.1.

The results showed that a good compromise for the parameters α, δ and w_0 is given by $\alpha = 1/3$, $\delta = 1/30$ and $w_0 = 1$. The complete methodology used to find the combination of the filter parameters is exposed in Appendix B.

Regarding the choice of filter length, it is important to note that one of the aims of this article is to compare the S-WLS filter performance with the one of X-11, considering the filters in the automatic option. Therefore, we only compared the performances of S-WLS and X-11 for the same filter lengths.

Figure 11 shows the transfer function of the S-WLS filter together with the one of the X-11 filter of the same length ($N = 145$). The dashed line represents the transfer function

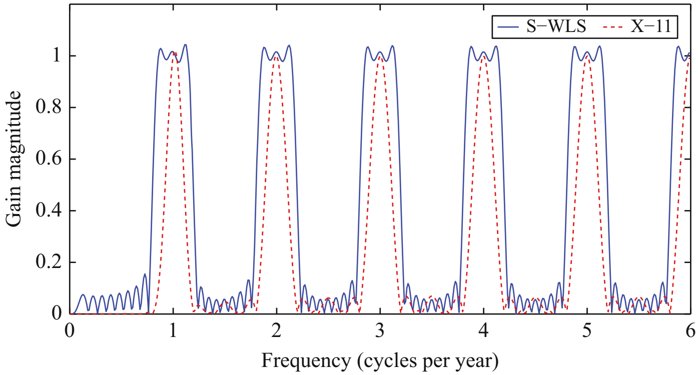


Fig. 11. Magnitude of the transfer function of the S-WLS filter and X-11 filter.

of the X-11 filter, and the continuous line represents the transfer function of the proposed S-WLS filter. As can be seen, the bandwidth of the proposed filter around the seasonal frequencies is larger than the one of X-11, with just a moderate amount of ripple. This allows estimating the seasonal component more accurately in the presence of instability in the seasonal frequency. Yet, the attenuation at the stopband is equivalent to the one of X-11, thus keeping the leaks of the irregular component at a level similar to the one of X-11.

It is important to highlight that, although in this article we determined the design parameters based on the behaviour of a large amount of time series, this filter can be designed according to the characteristics of a specific time series.

The MATLAB program used to implement the S-WLS filter is provided in SMT-UFRJ (2014). In the next section we will present a summary of the experimental results obtained with the S-WLS filter.

4. Results: the S-WLS Filter Performance

Since real time series have unobserved components, we decided to use artificial ones in order to better assess the filter performance. This is so because all the parameters of an artificial series are known and the estimation errors of the seasonal adjustment method can be precisely computed.

The artificial time series used were generated with several degrees of moving seasonality, considering some seasonal behaviours that, in aggregate, characterise the variety of monthly macroeconomic series (see the data in SMT-UFRJ (2014)). Their generation procedure is fully described in the next subsection.

To identify in which conditions of moving seasonality the proposed S-WLS filter performs better than the X-11 method, we applied both S-WLS and X-11 filters to seasonally adjust the mentioned series.

4.1. Data: Application on Artificial Time Series

To assess the ability of S-WLS filter to provide a satisfactory seasonal adjustment for series with moving seasonality, as well as to determine the conditions in which this filter performs better that X-11, we used monthly artificial time series with additive decomposition. These series were divided into two sets: in the first set the series were composed of a seasonal component with moving seasonality added to an irregular component; in the second set of data, a trend component following a cubic polynomial was added to these series, so that the performance of the proposed filters in the presence of a trend component could be assessed. We have chosen an order three polynomial to allow a fair comparison with the X-11 method. This is so because X-11 uses the Henderson filters, which can handle polynomial trend up to order three.

The seasonal component was generated with three parameters (A, b, k) defined in Equation (23). The choice of the parameters A, b, k, as well as the standard deviation of the irregular (s), was based on the characteristics of real monthly series, as mentioned before. As an example, in Figure 12a we show an artificial series with moving seasonality, where we have a cubic polynomial trend component added to an irregular component with parameters $b = 40\%$, $k = 120$, and $A/s = 6$ (Equation (23)). In Figure 12b we show this series without the trend component, and Figure 12c shows just the seasonal component.

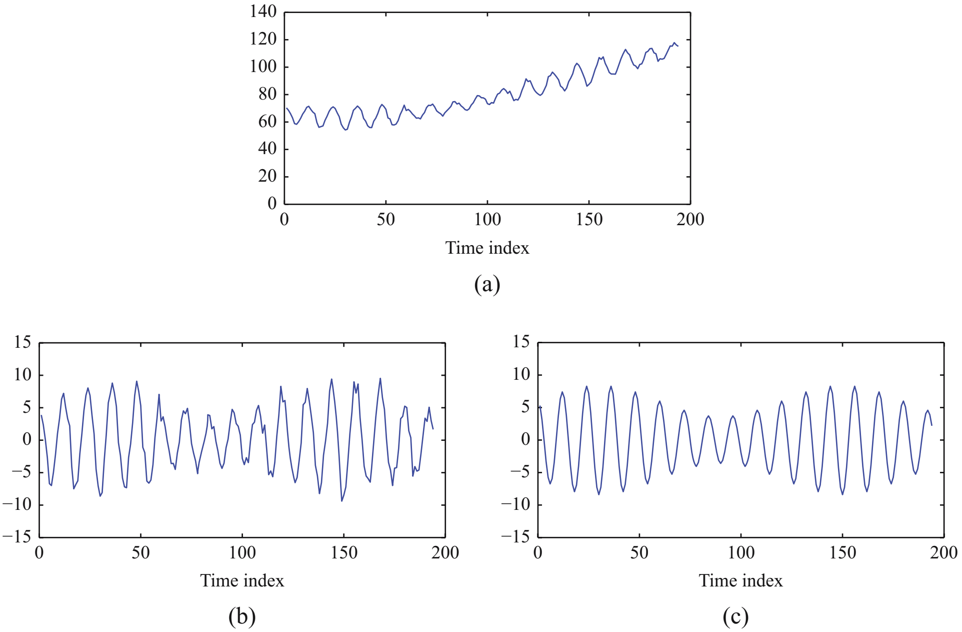


Fig. 12. Equation (23) with b = 40%, k = 120, and A/s = 6 (a) trend plus seasonal component and irregular, (b) seasonal component plus irregular, (c) seasonal component.

In this analysis we used time series of size 400. It is important to note that it lies outside the scope of this work to extend the series using forecast and backcast. We have done so to avoid masking the differences among the analysed filters. Therefore, since the considered filters (S-WLS and X-11) are symmetric, observations at both ends of the series had to be discarded. It is also important to note that these series were generated without outliers or missing values, so we could focus just on the filters, suitability to extract seasonality.

We generated 1,200 time series. Each simulation was replicated 100 times, randomising the irregular component.

4.2. *Criteria for Comparison*

To assess the ability of S-WLS filter to provide a satisfactory seasonal adjustment for series with moving seasonality, as well as to determine the conditions in which this filter performs better that X-11, we compared the accuracy in seasonality estimation of both filters when applied to the artificial time series. The procedure used in this performance comparison is described by the following steps:

(1) initially, the artificial series were seasonally adjusted by X-11 method considering the seasonal moving averages and Henderson trend filters in the automatic procedure of X-13A-S program;
(2) the X-11 filter that showed the best SNR for the analysed series was chosen and the filter length was determined;
(3) after determining the filter length, the proposed S-WLS filter with the same length was applied to the data.

The accuracy of each set of estimates (S-WLS and X-11) was measured by comparing them to the known seasonal component underlying that series. For this we used the signal-to-noise ratio (SNR – see details in Appendix C), the Mean Squared Error (MSE) and the Mean Absolute Deviation (MAD). We define the mean of the MSE and the mean of the MAD as the average of these statistics over 100 replications of the irregular. A one-sided t-test was applied to the pairs of means of the MSE statistics obtained for S-WLS and X-11 filters, with the alternative hypothesis $\mu_{S-WLS} < \mu_{X-11}$ (negative difference). The same was done considering the MAD statistic.

4.3. Simulation Results

In order to evaluate the conditions under which the S-WLS filters have a better performance than X-11, we considered different possibilities for the seasonal component. These characteristics refer to the parameters b, k, and A (Equation (23)), which were taken from macroeconomic series.

The parameter b is related to the rate of change in the seasonal amplitude, taking values in the interval (0,1). In real data the maximum value found for b was 52%.

Table 3 shows a performance comparison for values of b from 10% to 80% considering $A/s = 6$ and $k = 120$. Figure 13a illustrates the relation between the MSE of X-11 and of the S-WLS filter.

As can be seen, the higher the b value, the better the MSE of the S-WLS filter is compared to the one of X-11. Note that the MSE of the S-WLS filter does not change substantially with the variation of b, while the MSE of X-11 significantly changes with b. The same occurs with the MAD and SNR statistics. Table 3 shows that for smaller values of b, the performance of X-11 tends to improve relative to the one of S-WLS. It is important to note that the value of b from which S-WLS starts to perform better than X-11 depends on the values of k and A/s.

In order to evaluate the performance of the filters S-WLS and X-11 based on the variation of k, we set $b = 40\%$ and $A/s = 6$. As the parameter k is related to the change in the seasonal pattern, the smaller the k, the more unstable the seasonality. In these cases, S-WLS tends to perform better than X-11. Table 4 shows numerical figures illustrating this behaviour.

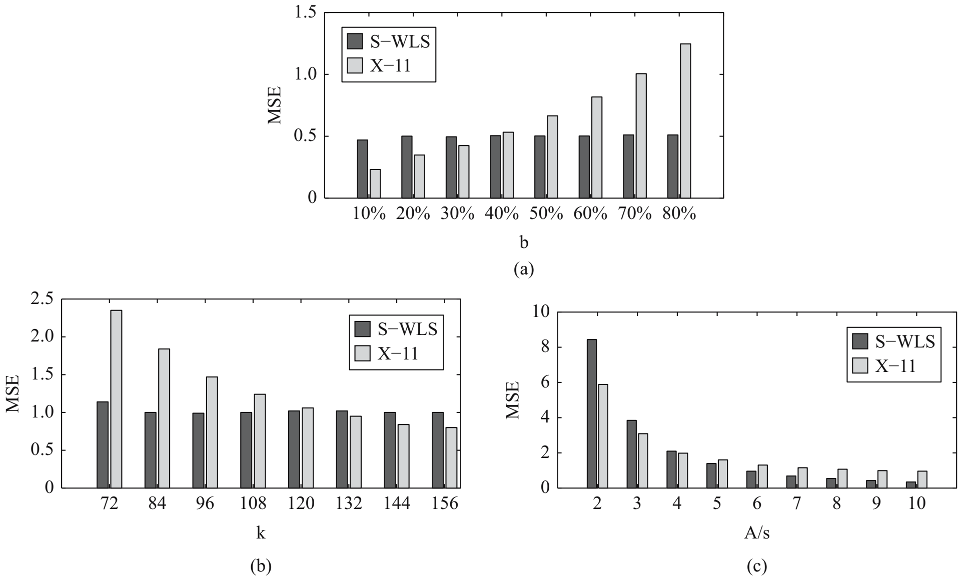
Figure 13b shows that when k decreases, the MSE of the S-WLS filter remains at the same level, indicating robustness of this filter, while the MSE of X-11 increases. Considering that the S-WLS filter uses the parameter $\alpha = 1/3$ (Subsection 3.2), the minimum value for k that it is able to deal with is 72.

The ratio between the amplitude of the signal (A) and the standard deviation of the irregular component (s) has a substantial influence on the MSE of the filters, as can be seen in Figure 13c. In both filters (X-11 and S-WLS), the MSE drops as the ratio A/s increases, but in the S-WLS this drop is more pronounced. Figure 13c and Table 5 also show that for large values of A/s, the S-WLS outperforms the X-11 filter. In typical series, the minimum value observed of A/s was 2.2 and the maximum was 11.7, and 50% of the monthly series with additive decomposition showed $A/s \geq 6$.

It is important to mention that for different values of k and b, the ratio A/s in which the proposed filter outperforms X-11 changes. For $k = 120$, the minimum A/s for which

Table 3. MSE, MAD, and SNR for b values, with A/s = 6, k = 120: monthly data with trend component (Equation (23)).

	MSE			MAD			SNR		
b	S-WLS	X-11	*p*-value	S-WLS	X-11	*p*-value	S-WLS	X-11	S-WLS/X-11
10%	0.940	0.463	1	0.773	0.547	1	53.7	110.1	0.5
15%	0.990	0.648	1	0.792	0.643	1	50.5	77.1	0.7
20%	1.002	0.697	1	0.798	0.665	1	50.9	71.8	0.7
25%	0.988	0.769	1	0.794	0.702	1	51.4	66.0	0.8
30%	0.992	0.850	1	0.792	0.740	1	52.0	60.3	0.9
40%	1.010	1.065	0.001	0.801	0.833	0.000	53.5	49.7	1.1
50%	1.006	1.331	0.000	0.800	0.940	0.000	55.5	41.1	1.3
60%	1.005	1.636	0.000	0.800	1.050	0.000	57.8	34.5	1.7
70%	1.022	2.012	0.000	0.808	1.173	0.000	60.5	29.4	2.1
80%	1.022	2.494	0.000	0.805	1.312	0.000	63.6	25.5	2.5



(a)

(b)

(c)

Fig. 13. Average of the MSE statistic in the simulations: X-11 and S-WLS filter (a) values of b, with A/s = 6, k = 120. (b) values of k, with A/s = 6, b = 40%. (c) values of A/s, with b = 40%, k = 120 (Equation (23)).

S-WLS overperforms X-11 is five; however, if $b = 40\%$ and $k = 72$, then $A/s \geq 3$ is enough for the S-WLS filter to perform better that X-11. The complete table with all the possibilities is available in SMT-UFRJ (2014).

Another way to verify the adequacy of the seasonal adjustment filter is to analyse the spectrum of the deseasonalised series. In Figures 14a and 14b we show the spectrum of the series deseasonalised by S-WLS filter and by X-11, respectively. The parameters of the series were the same as those used in the examples above: $b = 40\%$, $k = 120$ and $A/s = 6$. In Figure 14a it is possible to note that there is a peak in the frequency 1/12 cycles/month, indicating that some seasonality remains in the series after being deseasonalised by X-11, while in Figure 14b there is no peak, meaning that the seasonality was removed after applying the S-WLS filter to the data.

The performances of the S-WLS filter and the X-11 are analysed and compared for artificial time series based on Equation (3), that simulates a seasonal component with nonstationary changes. The results are presented in Table 6, showing the MSE and MAD for all combinations of Henderson filter and seasonal moving average filters.

Analysing the results we note that the proposed method (S-WLS) is able to estimate this kind of non-stationary seasonality better than X-11.

4.4. Data: Application on Real Time Series

In order to illustrate the S-WLS filter on a real-life time series, we applied it to the Austrian Consumer Price Index (all items non-food non-energy). This monthly time series was obtained from the OECD (http://stats.oecd.org/index.aspx?DatasetCode=MEI, extracted on April 2016), with a time span of 41 years (from jan/1975 to jan/2016). Besides this,

Table 4. MSE, MAD, and SNR for k values, with A/s = 6, b = 40%: monthly data with trend component (Equation (23)).

	MSE			MAD			SNR		
k	S-WLS	X-11	*p*-value	S-WLS	X-11	*p*-value	S-WLS	X-11	S-WLS/X-11
72	1.14	2.35	0.000	0.85	1.25	0.000	48.6	23.3	2.1
84	1.00	1.84	0.000	0.80	1.11	0.000	53.5	30.2	1.8
96	0.99	1.47	0.000	0.79	0.99	0.000	53.5	35.9	1.5
108	1.00	1.24	0.000	0.80	0.90	0.000	53.5	42.7	1.3
120	1.02	1.06	0.001	0.80	0.83	0.000	53.5	49.7	1.1
132	1.02	0.95	1	0.81	0.79	0.999	53.6	55.8	1.0
144	1.00	0.84	1	0.80	0.74	1	53.7	64.5	0.8
156	1.00	0.80	1	0.80	0.71	1	53.8	65.8	0.8
180	1.01	0.73	1	0.80	0.68	1	53.8	73.4	0.7

Table 5. MSE, MAD, and SNR for A/s values, with b = 40% and k = 120: monthly data with trend component (Equation (23)).

	MSE			MAD			SNR		
A/*s*	S-WLS	X-11	*p*-value	S-WLS	X-11	*p*-value	S-WLS	X-11	S-WLS/X-11
2	8.438	5.884	1	2.320	1.939	1	6.41	9.17	0.70
3	3.845	3.089	1	1.565	1.404	1	14.38	17.76	0.81
4	2.096	1.981	0.999	1.153	1.130	0.977	25.47	26.42	0.96
5	1.394	1.603	0.000	0.945	1.025	0.000	39.62	34.13	1.16
6	0.957	1.308	0.000	0.780	0.932	0.000	56.74	40.56	1.40
7	0.686	1.156	0.000	0.661	0.884	0.000	76.59	45.72	1.68
8	0.540	1.070	0.000	0.587	0.855	0.000	99.50	49.90	1.99
9	0.427	0.991	0.000	0.520	0.827	0.000	125.14	53.23	2.35
10	0.347	0.960	0.000	0.468	0.817	0.000	152.80	55.86	2.74

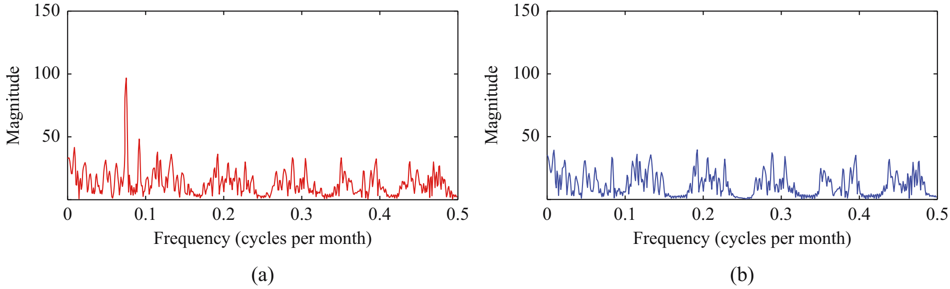


Fig. 14. Spectrum of the deseasonalised series (a) by X-11. (b) by S-WLS.

we used X-11 (X-13A-S program) to seasonally adjust this time series. Then, we compared the results.

Note that we do not perform series extrapolation at its extremes because the effects of the asymmetrical weights could mask the differences that we want to analyse.

The seasonal component extracted by X-11 and S-WLS filters is shown in Figures 15a and 15b, respectively.

We can see that by 1994 the seasonal component of S-WLS is higher than the one of X-11, showing that it can capture more seasonality than X-11.

5. Concluding Remarks

In this article we have proposed a seasonal adjustment filter design methodology, in which the main feature is to be robust to changes in the seasonal behaviour. This filter, named S-WLS, was designed in the frequency domain, based on least squares criteria, allowing the specification of an adequate passband width for filtering series with moving seasonality.

Several seasonal adjustment filters have been proposed in the literature. Our contribution is in the fact that this robustness is achieved while preserving its automatic character (so, it can be used to adjust a large amount of series). In addition, its parameters were determined based on the seasonal behaviour of typical macroeconomic series.

Table 6. Comparison results of S-WLS and X-11 for different Henderson filters and seasonal moving average filters.

		MSE		MAD	
Henderson filter	Seasonal MA	S-WLS	X-11	S-WLS	X-11
9	3 × 3	0.021	0.066	0.097	0.198
9	3 × 5	0.019	0.090	0.091	0.230
9	3 × 9	0.021	0.148	0.097	0.298
13	3 × 3	0.019	0.063	0.092	0.193
13	3 × 5	0.020	0.088	0.094	0.228
13	3 × 9	0.020	0.149	0.096	0.299
23	3 × 3	0.019	0.049	0.092	0.167
23	3 × 5	0.020	0.080	0.094	0.213
23	3 × 9	0.020	0.151	0.096	0.299

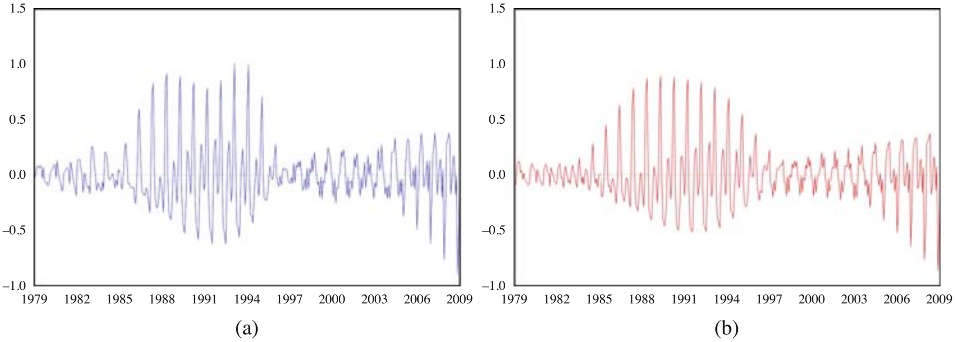


Fig. 15. Seasonal component of Austrian Consumer Price Index (all items non-food non-energy) (a) extracted with X-11. (b) extracted with S-WLS.

With the aim of assessing the performance of our filter, we compared it to the X-11 method, since this is an ad-hoc filter and has been widely used. In the comparisons we took care to use the S-WLS filters with the same lengths of X-11 (considering the automatic procedure in X-13A-S). These comparisons were performed both using time domain characteristics, based on the statistics MSE and MAD, and frequency domain characteristics, using the SNR and the inspection of the spectrum of the seasonally adjusted series. In these comparisons, we used simulated monthly data with changing seasonal patterns based on the ones of the typical macroeconomic series.

Due to space constraints, this study was limited to monthly additive series. Yet, we have verified that our filter can be easily extended to quarterly data and other periodicities, and it also performs well in multiplicative series.

The simulation results show that for series with a very slowly changing seasonal pattern, our filter tends to overadjust the data in comparison to X-11. On the other hand, as the degree of moving seasonality increases, X-11 tends to underadjust the series (i.e., not to remove all the seasonality), while our filter shows a good performance. This occurs thanks to the larger passband width of the S-WLS filter that allows robustness in cases of moving seasonality, providing a better quality of adjustment than X-11.

Detailing these findings regarding the characteristics of the seasonal component, we have that:

(i) if the amplitude of the seasonal component is large when compared to the standard deviation of the irregular component, the proposed S-WLS filter performs better than X-11;
(ii) the same occurs when the rate of change in the seasonality amplitude is large enough;
(iii) regarding the period of change in the seasonal pattern, the faster the changes in seasonal pattern, the better is the performance of S-WLS.

In brief, we recommend the X-11 method – in X-13A-S – if the variations in seasonality are sufficiently small. In cases of stronger variations, the proposed S-WLS filter performs better. Interesting further investigation would be to extend the current work for the case of multivariate series, such as is done in Infante et al. (2015) in the context of testing common seasonal patterns.

Appendix A. The X-11 Algorithm in the Frequency Domain

First of all, we define the $P \times Q$ moving average in the frequency domain, using the Z-transform, as follows:

$$M_{PXQ}(z) = \frac{z^{\left(\frac{p+Q-2}{2}\right)}}{PQ}\left[\sum_{n=0}^{P-1}(n+1)z^{-n} + \sum_{n=P}^{Q-2}Pz^{-n} + \sum_{n=Q-1}^{P+Q-1}(P+Q-1-n)z^{-n}\right] \tag{A.1}$$

The standard X-11 algorithm as shown in (Dagum 1988) and in (Findley et al. 1998) is presented below. For this, we consider a monthly time series and additive decomposition: $y_t = t_t + s_t + i_t$, where y_t is the original series, and t_t, s_t, i_t are the non-observable components of trend, seasonality and irregular, respectively. The filtering operations are presented in the frequency domain, using the Z-transform.

Stage 1 Preliminary Estimates

The first estimate of the trend component is obtained by applying a 'centered 12-term' moving average $M_{2X12}(z)$, that is

$$T^{(1)}(z) = Y(z)M_{2X12}(z). \tag{A.2}$$

The first estimate of Seasonal and Irregular components (SI) is given by

$$SI^{(1)}(z) = Y(z) - T^{(1)}(z). \tag{A.3}$$

One obtains the preliminary estimate of the Seasonal Factor $\hat{S}^{(1)}(z)$ by applying a weighted five-term moving average ($M_{3X3}(z^{12})$) to the SI component,

$$\hat{S}^{(1)}(z) = SI^{(1)}(z)M_{3X3}(z^{12}). \tag{A.4}$$

Then, the initial Seasonal Factor ($S^{(1)}(z)$) and the preliminary Seasonal Adjustment ($A^{(1)}(z)$) are obtained as

$$S^{(1)}(z) = \hat{S}^{(1)}(z) - \hat{S}^{(1)}(z)M_{2X12}(z) \tag{A.5}$$

$$A^{(1)}(z) = Y(z) - S^{(1)}(z). \tag{A.6}$$

Stage 2 Seasonal Factors and Seasonal Adjustment

We then perform the intermediate trend estimate by applying the '13'-term Henderson filter ($H_{13}(z)$) to the seasonally adjusted series, from Equation (A.6),

$$T^{(2)}(z) = A^{(1)}(z)H_{13}(z). \tag{A.7}$$

The second estimate of the SI component is then given by

$$SI^{(2)}(z) = Y(z) - T^{(2)}(z). \tag{A.8}$$

We then obtain the second estimate of the Seasonal Factor via a seven-term moving average (‘3 × 5’ seasonal moving average):

$$\hat{S}(2)(z) = SI^{(2)}(z)M_{3X5}(z^{12}) \tag{A.9}$$

from which the Seasonal Factor ($S^{(2)}(z)$) and the Seasonally Adjusted series ($A^{(2)}(z)$) are obtained as

$$S^{(2)}(z) = \hat{S}^{(2)}(z) - \hat{S}^{(2)}(z)M_{2X12}(z) \tag{A.10}$$

$$A^{(2)}(z) = Y(z) - S^{(2)}(z). \tag{A.11}$$

As the operations presented in Stages 1 and 2 are linear, it is possible to represent them as an equivalent filter of X-11 method for the seasonal adjustment. This filter, from Equations (A.2) to (A.11), is

$$\begin{aligned} A^{(2)} = Y(z)\{1 - \mathrm{M}_{3\times5}(z^{12})[1 - \mathrm{M}_{2\times12}(z)][1 - H_{13}(z) \\ \times\{1 - [1 - \mathrm{M}_{2\times12}(z)]^2\mathrm{M}_{3\times3}(z^{12})\}]\}. \end{aligned} \tag{A.12}$$

Appendix B. The Procedure for Selecting the Values of the S-WLS Filter Parameters

The procedure for selecting the values of the S-WLS filter parameters is summarised in the following steps below. For further details, the reader is referred to SMT-UFRJ (2014):

B.1. Experimental Determination of the Values For A, b, k (see Equation (23)), and the Standard Deviation (s) of the Irregular Component

For each real time series, we used the X-13A-S program (considering the X-11 adjustment mode) to estimate the seasonal component. The behaviour of the seasonal components of these real time series with moving seasonality was individually analysed. The specification of the value of α was based on the values of k, using the relation in Equation (24) exemplified by Figure (10).

B.2. Choosing the Filter Length

The X-11 filter that best fitted the data was chosen based on the theoretical evaluation of the SNR, for all the combinations of seasonal moving average filters and Henderson filters in the automatic mode of X-11. For this we used the values of A, b, k and s from step A.1 to build artificial seasonal signals and irregular components. Then we searched for the best X-11 filter for a given combination of these parameters, based on the X-11 SNR (Equation (C.9)).

B.3. Searching for the S-WLS Parameters

Based on the filter length defined in Subsection 2, we searched for the combination of δ and w_0 with largest ratio between the SNRs of S-WLS and X-11 (see Equations (C.8) and (C.9)). At this stage, we restricted the choice of the lengths of the S-WLS filter to the possible lengths of the X-11 filter. The parameter α was fixed at 1/3, while δ and w_0 values were varied over a wide range. Then, for several combinations of A, k, b (Equation (23)) and s, we chose the S-WLS filters that were, in general, based on SNR better than X-11. Since there is no single set of parameters that best fits the data, we worked with the top eight combinations.

B.4. Determination of the Best Parameters for a Wide Range of the Parameter k

At this stage we performed a simulation by filtering time series with moving seasonality, cubic trend component, and irregular component following a $N(0,\sigma^2)$. In it, 100 replications of the irregular component were generated for each series. The artificial seasonal components were created considering 100 values of k (Equation (23)), drawn randomly from a set of possible values based on the seasonal behaviour of macroeconomic data.

To search for the S-WLS parameters, we first chose the X-11 filter that had the lowest MSE for the estimation of the seasonal component of each replication of the series. Setting the same length of the S-WLS filter as the one of X-11, we searched for the parameters δ and w_0 that provided good MSE figures.

The S-WLS parameter combination that obtained, in general, the lowest MSE compared to the MSE of the X-11, was identified as $\alpha = 1/3$, $\delta = 1/30$ and $w_0 = 1$.

Appendix C. X-11 and S-WLS SNRs

When we filter a time series to extract its irregular component, according to the model in Equation (2), the errors at its output may have three main causes:

(i) residuals of the trend component;
(ii) irregular component at the output of the filter;
(iii) errors caused by the seasonal component.

In our case, the errors in (i) are automatically eliminated by the filter's structure (Equation (8)). In Sections B.1 and B.2, we deal with the errors in (ii) and (iii), respectively.

C.1. Variance of the Irregular at the Output of the Filter (Noise Power)

When a stochastic process $x(t)$ with power spectral density $S_X(e^{i\omega})$ is input to a filter with transfer function $\mathcal{H}(e^{i\omega})$, the power spectral density of its output $y(t)$ is $S_X(\omega)|\mathcal{H}(e^{i\omega})|^2$ (Diniz et al. 2010). Since an uncorrelated irregular component with variance, or noise power, σ_X^2 has a power spectral density equal to σ_X^2, then the power spectral density of its output $y(t)$ is

$$S_Y(e^{i\omega}) = \sigma_X^2 |\mathcal{H}(e^{i\omega})|^2. \tag{C.1}$$

Therefore, the variance of the irregular component, or noise power, of the output of the filter is

$$\sigma_Y^2 = \int_{-\pi}^{\pi} S_Y(e^{i\omega})d\omega = \int_{-\pi}^{\pi} \sigma_X^2 |\mathcal{H}(e^{i\omega})|^2 d\omega = \sigma_X^2 \sum_{t=-\infty}^{\infty} |h(t)|^2. \tag{C.2}$$

where the rightmost equality comes from Parseval's theorem (Diniz et al. 2010), with $h(t)$ being the filter coefficients. In other words, in the case of an uncorrelated irregular component, the noise power of the irregular at the output of a filter is proportional to the sum of the squares of its coefficients.

C.2. Errors Caused by the Seasonal Component

(a) The case of the S-WLS filter

Consider Figure C.16, which illustrates a typical frequency response of the S-WLS filter at the passband (see also Figures 8a to 8d in Subsection 3.1). There, we highlight two important deviations from the desired unit gain in the passband. The first one is given by γ_0, which is the gain at the fundamental frequency. The second one gives the maximum deviation from the desired response at the passband. Since the desired response is 1, and the corresponding gain is γ_1, the maximum deviation is given by $|1 - \gamma_1|$.

For an input time series according to Equation (23), since it is composed of sinusoidal components, the contribution of the above deviations to the noise power is,

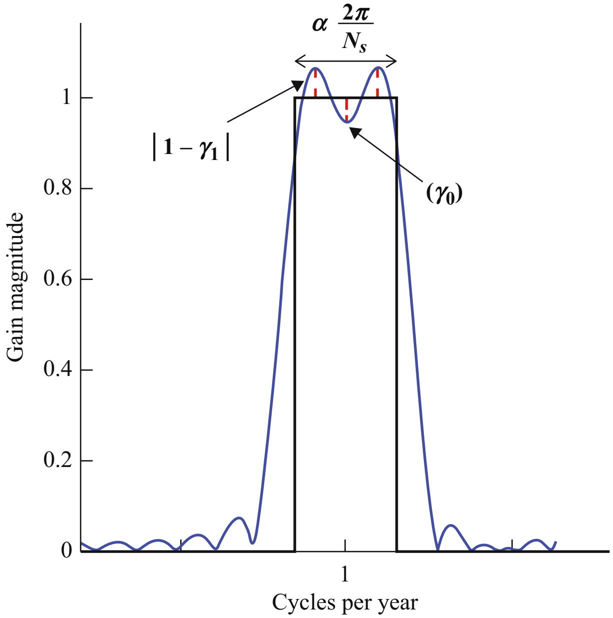


Fig. C.16. Deviations from the desired passband gain for the S-WLS filter.

in the worst case,

$$e_1 = (1-\gamma_0)^2\frac{A^2}{2} + (1-\gamma_1)^2\frac{A^2b^2}{4}. \tag{C.3}$$

where A is the seasonal amplitude ($A \in \mathbb{R}$), b is related to the rate of change in the seasonal amplitude ($b \in (0,1)$), k is related to the change in the seasonal pattern, and t is the time index ($t = 1, 2, \ldots$).

(b) The case of the X-11 filter

As can be seen from the frequency responses of the X-11 filter in Figure 11, Subsection 3.2, the typical frequency response of the X-11 filter always has a peak at the seasonal frequency (1/12 cycles per month, for monthly series, and 1/4 cycles per quarter, for quarterly series). As you move away from the peak, the response decreases monotonically. Therefore, the two largest deviations from the ideal unit gain in the passband are given by the two frequencies at the edges of the passband, as illustrated in Figure C.17, in which the response function differs from the 'ideal' by $(1-\beta_1)$ and $(1-\beta_2)$. There, the dashed line represents the spectrum of the X-11 equivalent filter, and the continuous line represents the magnitude of the ideal frequency response for an allowed seasonal frequency variation of $\alpha(2\pi/N_s)$ around the nominal frequency (in this case, $2\pi/N_s$ or its harmonics).

Using an argument equivalent to the one that led to Equation (C.3), the contribution of the above deviations to the noise power is, for the X-11 filter,

$$e_2 = \frac{A^2b^2}{4}\frac{[(1-\beta_1)^2 + (1-\beta_2)^2]}{2} \tag{C.4}$$

(c) Computation of the SNR of the S-WLS and X-11 filters

Therefore, if we refer to the sum of the squares of the coefficients of the S-WLS filter as SQ and to the one of the X-11 as S, then we have, from Equations (C.3) and (C.4),

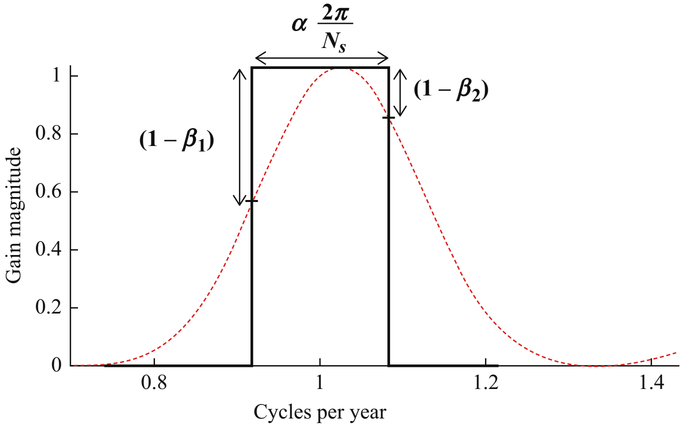


Fig. C.17. Deviations from the desired passband gain for the X-11 filter.

that the total noise power at the output of S-WLS and X-11 filters is given by:

$$e_{S-WLS} = (1-\gamma_0)^2 \frac{A^2}{2} + (1-\gamma_1)^2 \frac{A^2 b^2}{4} + SQ\,\sigma^2 \tag{C.5}$$

$$e_{X-11} = \frac{A^2 b^2}{8}\left[(1-\beta_1)^2 + (1-\beta_2)^2\right] + S\,\sigma^2. \tag{C.6}$$

Since the seasonal signal in Equation (23) is composed of three sinusoids, its average squared value is given by

$$E_s = A^2 \frac{1}{2} + \frac{A^2 b^2}{4}\frac{1}{2} + \frac{A^2 b^2}{4}\frac{1}{2} = \frac{A^2}{2}\left(1 + \frac{b^2}{2}\right). \tag{C.7}$$

Therefore, from Equation (C.5) to (C.7), we have that the SNRs of the S-WLS and X-11 filters are:

$$SNR_{S-WLS} = \frac{\frac{A^2}{2}\left(1+\frac{b^2}{2}\right)}{(1-\gamma_0)^2 \frac{A^2}{2} + (1-\gamma_1)^2 \frac{A^2 b^2}{4} + SQ\,\sigma^2} \tag{C.8}$$

$$SNR_{X-11} = \frac{\frac{A^2}{2}\left(1+\frac{b^2}{2}\right)}{\frac{A^2 b^2}{8}\left[(1-\beta_1)^2 + (1-\beta_2)^2\right] + S\,\sigma^2}. \tag{C.9}$$

6. References

Bell, W.R. and S.C. Hillmer. 1984. "Issues Involved with the Seasonal Adjustment of Economic Time Series." *Journal of Business & Economic Statistics* 2: 291–320. Doi: http://dx.doi.org/10.2307/1391275.

Bell, W.R. and B.C. Monsell. 1992. *X-11 Symmetric Linear Filters and Their Transfer Functions*. Bureau of the Census, Research Report n. RR 92: 15. Available at: https://www.census.gov/srd/papers/pdf/rr92-15.pdf (accessed September 2015).

Burman, J.P. 1980. "Seasonal Adjustment by Signal Extraction." *Journal of the Royal Statistical Society: Series A (General)* 143: 321–337. Doi: http://dx.doi.org/10.2307/2982132.

Canova, F. and E. Ghysels. 1994. "Changes in Seasonal Patterns: Are They Cyclical?" *Journal of Economic Dynamics and Control* 18: 1143–1171. Available at: http://apps.eui.eu/Personal/Canova/Articles/chanseapat.pdf (accessed September 2015).

Canova, F. and B.E. Hansen. 1995. "Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability." *Journal of Business & Economic Statistics* 13: 237–252. Doi: http://dx.doi.org/10.2307/1392184.

Cleveland, R.B., W.S. Cleveland, J.E. McRae, and I. Terpenning. 1990. "STL: A Seasonal-Trend Decomposition Procedure Based on Loess." *Journal of Official Statistics* 6: 3–73.

Dagum Bee, E. 1980. *The X-11-ARIMA Seasonal Adjustment Method. Statistics Canada - Seasonal Adjustment and Time Series Staff*. Available at: https://www.census.gov/ts/papers/1980X11ARIMAManual.pdf (accessed September 2015).

Dagum Bee, E. 1988. *X-11-ARIMA/88 Seasonal Adjustment Method - Foundations and Users' Manual*. Technical Report, Statistics Canada.

Dagum Bee, E., N. Chhab, and K. Chiu. 1996. "Derivation and Properties of the X11ARIMA and Census X11 linear filters." *Journal of Official Statistics* 12: 329–348.

Diniz, P., E.A. da Silva, and S.L. Netto. 2010. *Digital Signal Processing: System Analysis and Design*. Cambridge University Press.

Eurostat. 2015. *ESS Guidelines on Seasonal Adjustment*. Luxembourg: Publications Office of the European Union. Available at: http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf (accessed April 2016). Doi: http://dx.doi.org/10.2785/317290.

Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto, and B.-C. Chen. 1998. "New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program." *Journal of Business & Economic Statistics* 16: 127–152. Doi: http://dx.doi.org/10.2307/1392565.

Findley, D.F. 2005. "Some Recent Developments and Directions in Seasonal Adjustment." *Journal of Official Statistics* 21: 343–365. Available at: https://www.census.gov/ts/papers/recentdevelopmentsjos.pdf (accessed September 2015).

Findley, D.F. and D.E.K. Martin. 2006. "Frequency Domain Analyses of SEATS and X-11/12-ARIMA Seasonal Adjustment Filters for Short and Moderate Length Time Series." *Journal of Official Statistics* 22: 1–34. Available at: http://www.jos.nu/Articles/abstract.asp?article=221001 (accessed September 2015).

Franses, P.H. and A.B. Koehler. 1998. "A Model Selection Strategy for Time Series with Increasing Seasonal Variation." *International Journal of Forecasting* 14: 405–414. Doi: http://dx.doi.org/10.1016/S0169-2070(98)00041-7.

Geweke, J. 1978. "The Temporal and Sectoral Aggregation of Seasonally Adjusted Time Series." *Seasonal Analysis of Economic Time Series*, 411–432. NBER.

Gilbart, J. 1852. "On the Laws of the Currency in Ireland, as Exemplified in the Changes that Have Taken Place in the Amount of Bank Notes in Circulation in Ireland, Since the Passing of the Act of 1845." *Journal of the Statistical Society of London* 15: 307–326.

Godfrey, M.D. and H.F. Karreman. 1964. *A Spectrum Analysis of Seasonal Adjustment*. Technical Report 64, Econometric Research Program - Princeton University. Available at: https://www.princeton.edu/~erp/ERParchives/archivepdfs/M64.pdf (accessed September 2015).

Gómez, V. and A. Maravall. 1996. "Programs TRAMO (Time Series Regression with Arima Noise, Missing Observations, and Outliers) and SEATS (Signal Extraction in Arima Time Series). Instructions for the User." *Documento de Trabajo*, vol. 9628.

Gómez, V. and A. Maravall. 2001. "Seasonal Adjustment and Signal Extraction in Economic Time Series." Peña, D., et al. 202–246. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.8918rep=rep1type=pdf (accessed September 2015).

Hannan, E.J. 1964. “The Estimation of a Changing Seasonal Pattern.” *Journal of the American Statistical Association* 59: 1063–1077. Doi: http://dx.doi.org/10.2307/2282624.

Hassani, H. 2007. “Singular Spectrum Analysis: Methodology and Comparison.” *Journal of Data Science* 5: 239–257.

Higginson, J. 1975. *An F Test for the Presence of Moving Seasonality when Using Census Method II-X-11 Variant*. Statistics Canada.

Hood, C.C., J.D. Ashley, and D.F. Findley. 2000. “An Empirical Evaluation of the Performance of TRAMO/SEATS on Simulated Series.” In of the American Statistical Association, Business and Economic Statistics Section, American Statistical Association, Alexandria, VA., 2000. Available at: https://www.census.gov/ts/papers/asa00_ts.pdf (accessed September 2015).

IBGE - Instituto Brasileiro de Geografia & Estatística. 2014. Available at: http://www2.sidra.ibge.gov.br/bda/ (accessed June 2014).

Infante, E., D. Buono, and A. Buono. 2015. “A 3-Way ANOVA a Priori Test for Common Seasonal Patterns and Its Application to Direct Versus Indirect Methods.” *Eurostat Review on National Accounts and Macroeconomic Indicators* (1/2015): 67–77.

IPEA - Instituto de Pesquisa Econômica Aplicada. 2014. Available at: http://www.ipeadata.gov.br (accessed June 2014).

Kaiser, R. and A. Maravall. 2000. *An Application of Tramo-Seats: Changes in Seasonality and Current Trend-Cycle Assessment: the German Retail Trade Turnover Series*. UC3M Working papers. Statistics and Econometrics 00-63, no. 29. Available at: https://core.ac.uk/download/pdf/30043292.pdf (accessed September 2015).

Koopman, S.J., A.C. Harvey, J.A. Doornik, and N. Shephard. 2000. *STAMP 6.0: Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants.

Kuznets, S. 1932. “Seasonal Pattern and Seasonal Amplitude: Measurement of Their Short-Time Variations.” *Journal of the American Statistical Association* 27: 9–20. Doi: http://dx.doi.org/10.2307/2277876.

Maravall, A. and D. Pérez. 2011. *Applying and Interpreting Model-Based Seasonal Adjustment; The Euro-Area Industrial Production Series*. Technical Report 1116, Banco de España. Available at: http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/11/Fich/dt1116e.pdf (accessed September 2015).

Melnick, E.L. and J. Moussourakis. 1974. “Filter Design for the Seasonal Adjustment of a Time Series.” *Communications in Statistics-Theory and Methods* 3: 1171–1186. Doi: http://dx.doi.org/10.1080/03610927408827219.

Nerlove, M. 1964. “Spectral Analysis of Seasonal Adjustment Procedures.” *Econometrica: Journal of the Econometric Society* 32: 241–286. Doi: http://dx.doi.org/10.2307/1913037.

Nettheim, N.F. 1964. *A Spectral Study of Overadjustment for Seasonality*. Technical Report, DTIC Document. Available at: https://statistics.stanford.edu/research/spectral-study-overadjustment-seasonality (accessed September 2015).

Nettheim, N.F. 1965. “Fourier Methods for Evolving Seasonal Patterns.” *Journal of the American Statistical Association* 60: 492–502. Doi: http://dx.doi.org/10.1080/01621459.1965.10480805.

OECD - Organisation for Economic Co-operation and Development. 2014. Available at: http://stats.oecd.org (accessed July 2014).

Planas, C. 1998. “The Analysis of Seasonality in Economic Statistics: A Survey of Recent Developments.” *Questiió: Quaderns d’Estadística, Sistemes, Informatica i Investigació Operativa* 22: 157–171. Available at: http://eudml.org/doc/40241 (accessed September 2015).

Shiskin, J., A.H. Young, and J.C. Musgrave. 1967. *The X-11 Variant of the Census Method II Seasonal Adjustment Program*. Technical Report, Economic Research and Analysis Division, US Department of Commerce, Bureau of the Census. Available at: https://www.census.gov/ts/papers/ShiskinYoungMusgrave1967.pdf (accessed September 2015).

SMT-UFRJ. 2014. Available at: http://www02.smt.ufrj.br/~eduardo/moving_seasonality/ (accessed December 2014).

Sutradhar, B.C. and E. Bee Dagum. 1998. “Bartlett-Type Modified Test for Moving Seasonality with Applications.” *Journal of the Royal Statistical Society: Series D (The Statistician)* 47: 191–206. Doi: http://dx.doi.org/10.1111/1467-9884.00123.

Tiller, R.T., D. Chow, and S. Scott. 2007. *Empirical Evaluation of X-11 and Model-Based Seasonal Adjustment Method*. Technical Report, Working Paper. Washington, DC: Bureau of Labor Statistics.

U.S. Bureau of Labor Statistics. 2014. Available at: http://www.bls.gov/data (accessed January 2014).

U.S. Census Bureau. 2013. *X-13ARIMA-SEATS Reference Manual version 1.1*. Time Series Research Staff, US Census Bureau. Available at: http://www.census.gov/ts/x13as/docX13AS.pdf (accessed September 2015).

U.S. Census Bureau. 2014. Available at: http://www.census.gov (accessed January 2014).

Van Dijk, D., B. Strikholm, and T. Teräsvirta. 2003. “The Effects of Institutional and Technological Change and Business Cycle Fluctuations on Seasonal Patterns in Quarterly Industrial Production Series.” *The Econometrics Journal* 6: 79–98. Doi: http://dx.doi.org/10.1111/1368-423X.00103.

Wallis, K.F. 1982. “Seasonal Adjustment and Revision of Current Data: Linear Filters for the X-11 Method.” *Journal of the Royal Statistical Society Series A (General)* 145: 74–85. Doi: http://dx.doi.org/10.2307/2981422.

Wells, J.M. 1997. “Modelling Seasonal Patterns and Long-Run Trends in US Time Series.” *International Journal of Forecasting* 13: 407–420. Doi: http://dx.doi.org/10.1016/S0169-2070(97)00027-7.

Received September 2015
Revised May 2016
Accepted September 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 187–206, http://dx.doi.org/10.1515/JOS-2017-0010

Bridging a Survey Redesign Using Multiple Imputation: An Application to the 2014 CPS ASEC

Jonathan Rothbaum[1]

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) serves as the data source for official income, poverty, and inequality statistics in the United States. In 2014, the CPS ASEC questionnaire was redesigned to improve data quality and to reduce misreporting, item nonresponse, and errors resulting from respondent fatigue. The sample was split into two groups, with nearly 70% receiving the traditional instrument and 30% receiving the redesigned instrument. Due to the relatively small redesign sample, analyses of changes in income and poverty between this and future years may lack sufficient power, especially for subgroups. The traditional sample is treated as if the responses were missing for income sources targeted by the redesign, and multiple imputation is used to generate plausible responses. A flexible imputation technique is used to place individuals into strata along two dimensions: 1) their probability of income recipiency and 2) their expected income conditional on recipiency for each income source. By matching on these two dimensions, this approach combines the ideas of propensity score matching and predictive means matching. In this article, this approach is implemented, the matching models are evaluated using diagnostics, and the results are analyzed.

Key words: Multiple imputation; survey redesign; bridge; CPS ASEC.

1. Introduction

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) is among the most widely used surveys conducted by the U.S. Census Bureau. CPS ASEC data are used to calculate measurements of national income and the official poverty rate. Rothbaum (2015) shows that the CPS ASEC suffers from underreporting of certain income types, including property income (especially interest and dividends), retirement income, and income from means-tested government transfer programs. Meyer et al. (2009) also show underreporting of participation in means-tested government programs.

To address this underreporting, the U.S. Census Bureau, in consultation with the private sector, implemented a redesign of the survey (see Czajka and Denmead (2008) and Hicks and Kerwin (2011) for results of that consultation). In 2014, approximately 30% of the CPS ASEC sample received the redesigned survey instrument, and approximately 70%

[1] Social and Economic Housing Statistics Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, U.S.A. Email: jonathan.l.rothbaum@census.gov

Acknowledgments: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological or operational issues are those of the author and are not necessarily those of the U.S. Census Bureau. Any error or omissions are the sole responsibility of the author. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau. All data used in this article are confidential.

received the unchanged traditional instrument (in use since 1994). Assignment into the two groups was random at the household level. For more details about the redesign and the content tests, see Semega and Welniak (2013, 2015).

A major focus of the redesign was to improve reporting of property income, especially income earned from assets in the form of interest or dividends. In addition, since 1980, the nature of retirement savings has shifted from defined benefit to defined contribution plans. From 1980 to 2008, the share of private wage and salary workers with defined benefit plans fell from 38% to 20%. The share of private workers with defined contribution plans grew from eight percent to 31% over the same period (Butrica et al. 2009). Therefore, the survey was redesigned to improve reporting of retirement income, which has also historically been underreported (Czajka and Denmead 2008).

The redesigned instrument is being used for the full sample, starting with the 2015 CPS ASEC. However, in order to make apples-to-apples comparisons between the results in 2014 and 2015 and beyond, only 30% of the 2014 sample can be used. This significantly reduces the power of the comparisons that can be made, for example of median income or poverty rates, and is especially relevant for subgroups.

While the survey redesign significantly increased recipiency and aggregates for many income types, the majority of income (by US dollars) was not affected. For example, earnings comprised 75.9% of all income in the redesign sample, and there were no statistically significant differences in the number of earners or mean earnings across the two instruments. Although we do not observe what respondents to the traditional instrument would have said to the redesigned questions, we do have a considerable amount of information about them that is unaffected by the redesign.

This suggests treating the problem as one of missing data – as if the recipients of the traditional instrument did not respond to the redesigned income questions. The "missing" responses to the redesigned questions are multiply imputed for individuals in the traditional sample. This article adds to the literature using multiple imputation to bridge a survey or data classification change. Clogg et al. (1991) used multiple imputation to impute industry and occupation codes across a change in the coding scheme between 1970 and 1980 census data. Schenker and Parker (2003) imputed single-race reporting for multiple-race respondents after a change from single- to multi-race reporting in government survey data.

An approach developed by Bondarenko and Raghunathan (2007) is applied to impute these missing responses in the traditional sample. This approach combines the ideas of propensity score matching and predictive means matching. By matching donors to recipients within propensity score/predictive mean cells, this approach is similar to the hot deck procedure used in normal CPS ASEC processing. That makes it appealing for use in this case, as the completed data from the 2014 CPS ASEC can be used to make comparisons with data in subsequent years where all imputation of missing values is done using the hot deck.

The Bondarenko and Raghunathan technique is used to create an "Income-Consistent" full file that uses all of the CPS ASEC sample with imputed income in the affected categories for respondents to the traditional instrument. It is called the Income-Consistent file, as the responses for all individuals are consistent with the questions in the redesign survey instrument.

The article is organized as follows. In Section 2, the CPS ASEC and the survey redesign are described. Section 3 discusses the imputation methodology. Section 4 discusses diagnostic results to evaluate the models used. Section 5 contains results relating to income and poverty, measurement using the imputed data. Section 6 contains a conclusion.

2. Data and Survey Redesign

The CPS ASEC is among the most widely used surveys conducted by the U.S. Census Bureau. The CPS ASEC uses a stratified random sample to survey about 100,000 households each year and includes questions on income and health insurance coverage.

The 2014 survey redesign included a number of changes. First, the survey was redesigned to specifically ask if anyone in the household has a pension, and separately if anyone has a retirement account (401(k), 403(b), IRA, or other account designed specifically for retirement savings). The traditional instrument includes one broad question on the receipt of pension and retirement income. The redesigned instrument also asks individuals over 70 years old about required distributions from retirement accounts. To ensure that the distribution is correctly identified as income, a follow-up question asks if the required distribution was "rolled over" or reinvested in another account. The traditional ASEC instrument makes no distinction between investment income received in retirement accounts or separately from them. This more detailed set of questions can improve misreporting of income and cue respondents to decrease underreporting.

Several additional changes were also made to the survey. Prior to the redesign, only households that reported less than USD 75,000 in combined family income were asked questions about means-tested transfer programs such as Temporary Assistance to Needy Families (TANF). Semega and Welniak (2015) cite evidence from the American Community Survey (ACS) that some screened households were likely to be recipients of these transfers making it inappropriate to remove them using the income screener. To prevent respondent fatigue from affecting answers to the income recipiency questions, the recipiency questions were separated from the amount questions as part of a "dual-pass" approach. Respondents were asked first about all sources of income received and then later were asked about amounts for only the received sources. In addition, the order of the income questions was changed based on respondent characteristics to match those sources most likely to be received. If a respondent was unsure of the income generated from assets, the value of the assets was collected. The questions on disability were clarified to eliminate confusion between disability income from Social Security and Supplemental Security Income (SSI).

2.1. Results of the Redesign

In 2014, the CPS ASEC sample was randomly divided into two groups at the household level, with 31% (30,000 housing units) receiving the redesigned instrument and about 69% (68,000 housing units) receiving the traditional instrument. Within each sample, individual observations were weighted to national population controls, as is standard with the CPS ASEC. Both surveys were conducted primarily by home visit (with some by phone) by trained field representatives. Even the interviewer was not aware of the selection of a given household into the traditional or redesign sample until they began the survey.

Semega and Welniak (2015) compared income aggregates between the two samples. Table 1 shows a subset of their results for median income, updated to reflect recent edits of the redesign sample file. Household median income was USD 51,939 in the traditional sample and USD 53,585 in the redesign, a difference of 3.2%. When decomposed by race, the only statistically significant differences are for whites (and non-Hispanic whites).

Table 2 shows income statistics for total income and various income sources collected in the CPS ASEC. For each source, Semega and Welniak report the number of recipients in the population, the mean income earned by those recipients, and the aggregate value of that income estimated, using the traditional and redesign samples separately. For example, for total income, the number of income recipients estimated using the traditional sample is 218.7 million compared with 222.0 from the redesigned sample, a statistically significant difference of 1.5%. The estimated difference in mean total income is 2.6% (USD 41,319 in the traditional vs. USD 42,394 in the redesign), and the estimated difference in aggregate total income is 4.2% (USD 9.04 trillion in the traditional compared with USD 9.41 trillion in the redesign), both statistically significant. At the 90% confidence level, there are a number of income sources that have statistically significant differences in the number of recipients, mean income, or aggregate income. The sources with statistically significant differences in aggregate income include farm self-employment income (−42.1%), public assistance (28.8%), veterans' benefits (−23.1%), disability benefits (36.4%), retirement income (21.9%), interest (113.0%), and dividends (−20.1%).

Mitchell and Renwick (2015) study the effects of the redesign on poverty rates. While they find no statistically significant difference in the overall poverty rate, they do find differences for child and elderly poverty in the redesigned sample. In both cases, they suggest that differences in the sample populations may explain the increase in poverty in the redesigned sample. For child poverty, they show that the redesigned sample has a higher share of children living with female householders than the traditional sample (single-mother families). They also find that means-tested program recipiency was higher in the redesigned sample.

These potential differences in sample characteristics support the approach taken in this article. Because the changes in the questionnaire are treated as a problem of missing information, any differences in the samples can be controlled for as a part of the imputation modeling, and the combined sample should better reflect the intended full CPS ASEC sample.

2.2. *Selection of Income Sources to be Imputed*

Taking these analyses together, the redesign increased aggregate income, increased income recipiency and reporting in a number of income categories. However, some of the differences, especially in income types with no or little change in the questionnaire, may be due to random variation or differences in the samples. This is supported by differences in poverty that Mitchell and Renwick attribute to sample differences.

Because of this evidence of sample differences, the analysis focuses on those income types which were targeted by the questionnaire redesign. This eliminates farm self-employment, and veterans' benefits.

The income types that are sufficiently different between the two surveys and were specifically targeted by the questionnaire redesign include: 1) retirement income,

Table 1. Comparison of traditional and redesign samples: median income.

	Traditional				Redesign					
	Number (thousands)		Median income (USD)		Number (thousands)		Median income (USD)		Percentage change $\left(\frac{R-T}{T}\right)$	
Characteristic	Estimate	90% CI	Estimate	90% CI	Estimate	90% CI	Estimate	90% CI	Estimate	90% CI
All households	122,952	723	51,939	455	123,931	942	53,585	1,076	3.2*	2.08
Race[1] and Hispanic origin of householder										
White	97,774	605	55,257	699	98,807	756	56,745	850	2.7*	1.81
White, not Hispanic	83,641	544	58,270	1,066	84,432	732	60,329	876	3.5*	2.04
Black	16,108	262	34,598	1,198	16,009	355	35,324	1,410	2.1	5.13
Asian	5,759	151	67,065	2,830	5,818	215	72,383	5,531	7.9	7.92
Hispanic (any race)	15,811	210	40,963	908	16,088	354	39,687	1,953	−3.1	5.00
Earnings of full-time year-round workers										
Men with earnings	60,769	600	50,033	404	61,240	787	50,015	935	−0.4	1.23
Women with earnings	45,068	510	39,414	596	44,629	659	38,792	1,145	−0.9	3.19

Notes: Income in 2013 dollars. Households and people as of March of the following year. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see www.census.gov/prod/techdoc/cps/cpsmar14.pdf. * indicates statistically different from zero at the 90-percent confidence level. A 90-percent confidence interval (CI) is a measure of an estimate's variability. The larger the CI is in relation to the size of the estimate, the less reliable the estimate. CIs shown in this table are based on standard errors calculated using replicate weights. For more information, see "Standard Errors and Their Use" at <www.census.gov/hhes/www/p60_245sa.pdf>.

[1] Federal surveys now give respondents the option of reporting more than one race. Therefore, two basic ways of defining a race group are possible. A group such as Asian may be defined as those who reported Asian and no other race (the race-alone or single-race concept) or as those who reported Asian regardless of whether they also reported another race (the race-alone-or-in-combination concept). This table shows data using the first approach (race alone). The use of the single-race population does not imply that it is the preferred method of presenting or analyzing data. The U.S. Census Bureau uses a variety of approaches. Information on people who reported more than one race, such as White and American Indian and Alaska Native or Asian and Black or African American, is available from Census 2010 through American FactFinder. About 2.9 percent of people reported more than one race in Census 2010. Data for American Indians and Alaska Natives, Native Hawaiians and Other Pacific Islanders, and those reporting two or more races are not shown separately in this table.

Source: U.S. Census Bureau, Current Population Survey, 2014 Annual Social and Economic Supplement.

Table 2. Comparison of traditional and redesign samples: income recipiency, mean income, and aggregate income by source.

	Traditional						Redesign						Percent change $\left(\frac{K-T}{T}\right)$					
	Number (thousands)		Mean income (USD)		Aggregate income (thousands)		Number (thousands)		Mean income (USD)		Aggregate income (thousands)		Number		Mean income		Aggregate income	
Type of income	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Total income	218,662	311	41,319	279	9,035,004	60,841	222,003	418	42,394	345	9,411,655	79,158	1.5*	0.2	2.6*	1.1	4.2*	1.1
Earnings	158,081	489	44,416	334	7,021,280	58,106	158,655	638	44,999	438	7,139,254	74,929	0.4	0.5	1.3	1.2	1.7	1.3
Wages and salary	148,752	492	44,931	336	6,683,647	55,005	149,546	684	45,695	457	6,833,462	75,382	0.5	0.5	1.7	1.2	2.2	1.4
Nonfarm self-emp	8,702	190	35,145	1,215	305,825	12,033	8,508	298	33,777	1,411	287,376	15,300	− 2.2	4.1	− 3.9	5.6	− 6.0	6.2
Farm self-emp	627	59	50,728	6,870	31,808	5,359	601	73	30,662	4,448	18,416	3,278	− 4.2	13.7	− 39.6*	12.4	− 42.1*	14.3
Unemployment	6,818	165	5,841	151	39,825	1,362	6,435	233	5,870	167	37,768	1,806	− 5.6	4.1	0.5	4.1	− 5.2	5.6
Workers' comp	1,186	60	9,224	566	10,940	930	952	77	10,156	851	9,671	1,135	− 19.7*	7.3	10.1	10.5	− 11.6	11.4
Social security	48,370	332	13,979	55	676,178	5,142	49,055	418	14,052	84	689,325	6,782	1.4	1.0	0.5	0.7	1.9	1.2
SSI	6,053	176	7,782	105	47,104	1,459	6,642	230	7,728	148	51,333	2,021	9.7*	5.0	− 0.7	2.3	9.0	5.5
Public assistance	1,775	81	3,195	149	5,671	340	2,189	124	3,337	160	7,305	508	23.3*	8.8	4.4	6.5	28.8*	10.9
Veterans' benefits	3,517	127	14,640	424	51,493	2,503	3,296	164	12,021	584	39,619	2,757	− 6.3	5.1	− 17.9*	4.8	− 23.1	6.1
Survivors' benefits	3,033	110	12,972	559	39,340	2,260	2,970	156	14,526	995	43,139	3,502	− 2.1	6.5	12.0	8.6	9.7	11.0
Disability benefits	1,771	76	15,543	736	27,524	1,850	3,099	163	12,110	635	37,535	2,654	75.0*	11.8	− 22.1*	5.6	36.4*	12.8
Retirement income	18,871	251	20,034	307	378,054	7,865	20,698	372	22,262	449	460,784	12,668	9.7*	2.4	11.1*	3.0	21.9*	4.2
Interest	86,142	588	2,120	68	182,619	5,963	123,772	887	3,142	107	388,943	13,403	43.7*	1.4	48.2*	6.8	113.0*	9.7
Dividends	33,243	432	4,424	170	147,050	6,225	31,804	568	3,693	211	117,454	6,954	− 4.3*	1.9	− 16.5*	5.8	− 20.1*	5.6

*Statistically different from zero at the 90-percent confidence level. Standard errors are calculated using replicate weights. For more information, see "Standard Errors and Their Use" at <www.census.gov/hhes/www/p60_245sa.pdf>.

Source: U.S. Census Bureau, Current Population Survey, 2014 Annual Social and Economic Supplement.

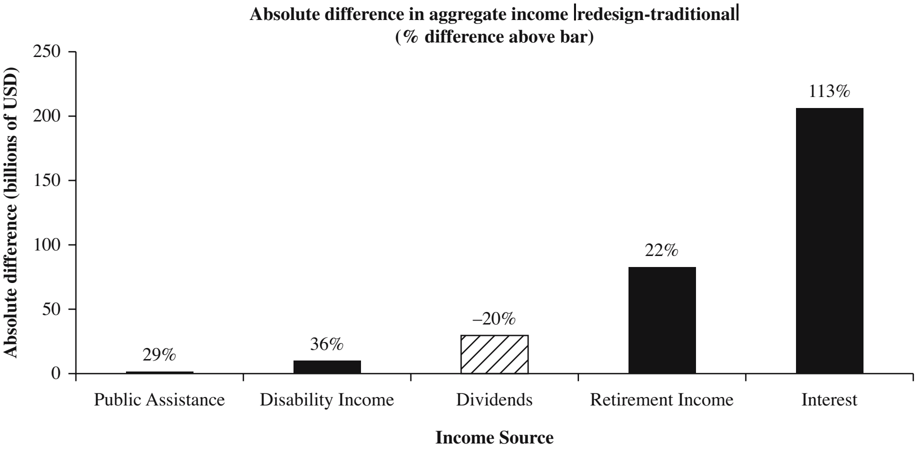


Fig. 1. Aggregate income differences between the traditional and redesign samples. Source: U.S. Census Bureau, Current Population Survey, 2014 Annual Social and Economic Supplement.

2) interest, and 3) dividends. These three sources had the largest difference in estimated aggregate income of the types affected by the redesign. Fig. 1 shows changes in aggregate income for all income sources with a statistically significant difference in aggregate income between the traditional and redesign samples. For interest income, the number of recipients increased by 37.6 million and aggregate income increased by USD 206.3 billion. For retirement income, the number of recipients increased by 1.8 million and aggregate income increased by USD 82.7 billion. For dividend income, the number of recipients decreased by 1.4 million and aggregate income decreased by USD 29.6 billion.

3. Imputation Methodology

3.1. Hot Deck Imputation

As a part of the standard processing of the CPS ASEC, when an individual does not respond to a particular question, missing values are imputed using a hot deck procedure. In the hot deck, individuals are divided into cells based on the characteristics specified in the hot deck model. Within each cell, individuals without missing information (donors) are randomly selected and their income is assigned to the individuals with missing information (recipients). Donors and recipients in each cell must match on every variable in the hot deck model. If there are no donors in a given recipient's cell, the hot deck model is amended to reduce the number of categories for some variables (for example from nine age groupings to six) and to reduce the number of variables in the model.

The different hot deck models used in the CPS ASEC are called match levels. The first match level includes the largest number of variables and categories within each variable. If no matches are found at the first level, an attempt to match recipients and donors is made using the model at the second match level. This continues until a match level is reached for a given recipient in which at least one donor is present in the same cell. For missing earnings in the longest job, at the first match level there are 16 variables in the model and 621 billion possible cells; at the second match level there are 14 variables and 17 billion possible cells; at the third match level there are eleven variables and 3.8 million possible cells, and by the sixth match level there are four variables and 96 possible cells. In the traditional sample for those observations missing earnings from the longest job only, 4.4% matched at the first level, 13.0% matched at the second level, 51.5% matched at the third level, and 6.4% matched at the sixth level. The variables and number of categories at each match level are available in Supplemental data online (URL: http://dx.doi.org/10.1515/JOS-2017-0010), Table 1.

As these numbers make clear, the number of variables that can be included in a hot deck model is limited by the size of the sample. While this is clearly a constraint even in the full CPS ASEC sample of about 200,000 individuals, the constraint is even more binding when imputing income from the redesign sample of about 60,000 individuals. If retirement, interest, and dividend income in the traditional sample were imputed using the hot deck model, it would not be possible to incorporate many variables in the model that are potentially correlated with each income type. This would limit the ability of the imputation to accurately match similar individuals as donors and recipients and reduce the quality of the matches.

3.2. *Model-Based Matching Imputation*

Instead, a more flexible technique is implemented to impute the missing responses to the redesigned questions in the traditional sample for the research file. The approach, developed by Bondarenko and Raghunathan (2007), hereafter BR, matches donors and recipients using summaries of covariates estimated by logistic and ordinary least squares regression modelling.

The primary reason the BR approach was chosen in this research is its similarity to the hot deck model. As in the hot deck model, individuals are matched based on similarities in observable characteristics. In the hot deck model, matching is directly based on the characteristics. In the BR approach, the matching is based on the predicted probability of recipiency and expected income conditional on recipiency, both of which can be estimated from observable characteristics. This is advantageous, as the imputed data must be comparable to data from subsequent years where all missing data are imputed using the hot deck model. However, by efficiently summarizing the model covariates in two statistics, recipiency and expected income, the BR approach allows for the inclusion of many more variables in the imputation model.

Next, the BR method is described, with slight modification for this application. Suppose that the dataset has P variables of observable characteristics, X_p, $p = 1, 2, \ldots, P$ and $X = (X_1, \ldots, X_P)$ and Q income types where Y_q, $q = 1, 2, \ldots, Q$, represents the income value and R_q represents recipiency status ($R_q \in \{0, 1\}$). There are two groups in the sample, one for which the income types q are observed (group O) and one for which income types q are unobserved (group M) so that each vector can be partitioned among O and M as $X_p = \left(X_p^O, X_p^M\right)$, $Y_q = \left(Y_q^O, Y_q^M\right)$, and $R_q = \left(R_q^O, R_q^M\right)$. Because missingness is complete for all Y_q^M, income can be imputed sequentially without iteration. Therefore, $<q$ is defined as the set of incomes with indices less than q so that $Y_{<q} = \left(Y_1, \ldots, Y_{q-1}\right)$ and $R_{<q} = \left(R_1, \ldots, R_{q-1}\right)$ and $Y_{<1}$ and $R_{<1}$ are empty sets. Two efficient summaries are constructed of the income variables through two regression predictions:

1. Probability of recipiency: $\hat{R}_q = \Pr\left(R_q = 1 | Y_{<q}, R_{<q}, X\right)$ estimated using a logistic regression model. This is an efficient summary of R_q that can be used to balance income recipients and nonrecipients.
2. $\hat{R}_q$ is stratified into K equal size strata, where $k = 1, \ldots, K$.
3. Predicted value of income conditional on recipiency within each stratum k: $\hat{Y}_q = E\left(Y_q | R_q = 1, Y_{<q}, R_{<q}, X\right)$ is estimated using an OLS regression model on all individuals in stratum k. Then, individuals are subdivided in stratum k into J equal sized substrata, where $j = 1, \ldots, J$. This creates $K \times J$ equal size strata.

Within each stratum k, j there are n individuals with observed income and recipiency, and m individuals with missing income and recipiency for income type q. A sample size m is drawn from the observed set of n individuals as the imputed values by Approximate Bayesian Bootstrap (ABB). This step is repeated for each stratum k, j and income type q and then sequentially for all $q = 1, \ldots, Q$. This entire process is repeated independently to obtain multiple imputations.

This approach relies on the same assumption that underlies matching models (see, for example, Rosenbaum and Rubin 1983). The traditional sample is M in this exercise, as the responses to the redesigned questions are missing for all individuals in the traditional sample. The redesign sample is O as the responses are observed. For this approach to be valid, it must be assumed that inclusion in the traditional sample can be controlled for by the variables in the imputation model (unconfoundedness). Specifically, given the probability of missingness $P(M)$, it is assumed that $P\left(M | X, Y_q^M, Y_q^O\right) = P(M|X)$ where X can be summarized by $\hat{R}_q$ and $\hat{Y}_q$. Although the random selection into the redesign sample implies missingness should be completely at random (MCAR), Mitchell and Renwick (2015) suggest that sample differences do exist on observable characteristics. Therefore, it is conservatively only assumed that the responses are missing at random (MAR).

There are a number of challenges to implementing BR method in the CPS ASEC. First, many income types do not follow a normal distribution or any simple transformation of a normal distribution. Because the missing income sources are modeled with continuous covariates, some distributional assumptions must be made about the relationships between them. Second, predictors (X) must be selected for the modelling of each income variable from a very large set of possible covariates ($>$1,200) in the CPS ASEC.

As shown in Hokayem et al. (2015), the distribution of income is rarely normally distributed. Simple transformation (such as log) and more flexible ones such as Tukey's gh distribution (He and Raghunathan 2006) also can fail to convert the distribution to normal. Therefore, an empirical normal transformation proposed by Woodcock and Benedetto (2009) is used to convert all income values to normal distributions (this includes income and other continuous variables in X as well) prior to imputation.

The most significant challenge to applying the BR method to the CPS ASEC was to select the models for each imputed variable. In order to avoid omitted variable bias in the imputation model, as many potential predictors as possible should be included. However, if too many variables are included, overfitting the model is a risk. The list of potential predictors used includes all unchanged income information (imputation flag, recipiency, value), spouse/partner earnings, race (separate dummy for each), gender, age (including dummies for each age between 62 and 70), weeks worked last year, hours worked per week, as well as the hot deck categories for relationship to householder, education level, marital status, presence of children, occupation (22 categories), type of residence, Census region, recipiency of means-tested government transfers. A large set of interaction terms are included in the list of predictors, including for major income types (earnings, spouse earnings, etc.), education, weeks and hours worked, race and age, and means-tested transfers. In all, over 1,200 potential predictors and interaction terms can be included in the BR models. A list of the modelling variables is available in Supplemental data, Table 2 (URL: http://dx.doi.org/10.1515/JOS-2017-0010).

The parameters of two models are estimated: (1) $R_q = F(X_q\beta_R + v_q)$ using logistic regression and (2) $Y_q = X_q\beta_Y + e_q$ by OLS. However, with more than 1,000 possible covariates, all possible covariates cannot be included in X and some values in β must be set to 0 in each regression.

Stepwise model selection is used to determine which values in β in each regression to set to 0. It was chosen as a pragmatic tool to efficiently capture the correlations between covariates X_q and dependent variables R_q and Y_q. However, there is uncertainty about which are the correct items in each β that should be set to zero which must be accounted for in order for the imputation to be proper. If the model variables were known with certainty (known nonzero items in β), after regressing Y on $X\beta$ parameter uncertainty could be accounted for using the variance-covariance matrix of β. However, in this case, both parameter and model uncertainty are present. In order to approximate both sources of uncertainty and have proper imputation variance estimates, for each income type, all regressions for each income type q are run on an approximate Bayes Bootstrap (ABB) sample.

In summary, the imputation steps to create the Income-Consistent file are:

1. Normal transformation – transform all income value variables to normal distribution with empirical normal transformation.
2. BR Imputation – sequentially impute interest, dividends, and retirement income from the redesign (donors/observed) to the traditional sample (recipients/missing). For each income type:
   a. Select a random sample by ABB.
   b. Predict probability of income recipiency using logistic regression on the redesign ABB sample with stepwise model selection to choose list of predictors. Only those individuals with non-imputed values of recipiency are included in the regression.
   c. Stratify the sample into K equal-sized groups based on probability of income recipiency in the original sample.
   d. Within each stratum k, predict expected income conditional on recipiency using OLS regression on the redesign ABB sample that is within the probability of recipiency bounds of that stratum.
   e. Stratify subsample k into J equal sized substrata based on the expected income of the original sample.
   f. Within each substratum j, select a random sample of m donors from the redesign sample (where m is the number of recipients with missing responses in stratum k,j) using ABB. Each donor receives all income, source, and value variables from the recipient. By donating source information (i.e., whether the retirement income is

from a 401k, IRA, or others), this implies the additional assumption that for each income type q and source S_q, $P\left(M|X, Y_q^M, Y_q^O, S_q^M, S_q^O\right) = P(M|X)$, where X can be summarized by $\hat{R}_q$ and $\hat{Y}_q$.

   g. Repeat for each stratum k, j until all missing observations for income type q have been imputed.

3. Transform to original scale – return all variables to their original scales.
4. Repeat the entire process to create ten implicates.

These steps are done after processing and allocation of the survey data. This means that hot deck imputed values in the redesign file can be used as part of the imputation process. However, all modelling and prediction is done only on actual responses with allocated values excluded from the modelling step.

Since all of the interest, dividends, and retirement income are missing for all observations in the traditional sample, the order of imputation should not matter. Consider, for example, the case where interest is imputed first and dividends second. In that case, the imputation for interest should capture the relationship between interest and all other variables in $X_{<q}$. In the imputation for dividends, information on interest is included in $X_{<q}$, which should capture the relationship between dividends and interest as well as dividends and all other variables in X. For both missing income types, the imputation captures the conditional relationship between the other type and the variables in X. The same is true if dividends are imputed first and interest second. As a result, the variables are imputed by frequency of recipiency, from most common to least common: 1) interest, 2) dividends, and 3) retirement income. Note that this invariance to imputation order is only true if missingness for each imputed variable implies missingness for all others.

4. Diagnostic Results

One way of evaluating the imputation model is to construct an R^2 from the set of regressions on the ABB sample. For the logistic regressions, the Tjur-R^2 (Tjur 2009) is used, which is calculated by comparing the average predicted probability of recipiency for those who did and did not receive income of that type, or

$$R^2_{\text{Tjur}} = E\left(\hat{R}_q|R = 1,\ Y_{<q}, R_{<q}, X\right) - E\left(\hat{R}_q|R = 0,\ Y_{<q}, R_{<q}, X\right).$$

The Tjur-R^2 is bounded between 0 and 1.

For the OLS regressions, the R^2 used is the squared correlation between the transformed income and the predicted income from the strata regressions, shown in Table 3. The average Tjur-R^2 for interest, dividends, and retirement are 0.35, 0.30, and 0.39 respectively. The OLS R^2 values for interest, dividends, and retirement income are 0.12, 0.10, and 0.15 respectively.

The relatively low R^2 are in part due to the fact that predictions are made on ABB samples, not the original one. The regression R^2 are much higher, but they reflect the match between the predictions and the bootstrapped sample, which will by definition be higher than for the original sample, which was not used for the prediction.

After imputing interest, dividend, and retirement income responses for the traditional sample, the two samples are combined to create the Income-Consistent file, as all responses

Table 3. Model diagnostics – effective R^2 of recipiency and value regressions.

	Recipiency			Value		
Variable	Average	Min	Max	Average	Min	Max
Interest	0.35 (0.01)	0.33	0.36	0.12 (0.03)	0.05	0.15
Dividends	0.30 (0.01)	0.28	0.32	0.10 (0.03)	0.03	0.13
Retirement	0.39 (0.02)	0.36	0.43	0.15 (0.05)	0.03	0.19

The R^2 are calculated by taking the predicted recipiency and values conditional on recipiency from the prediction models used to define the donor/recipient cells and calculating the Tjur R^2 for recipiency and squared correlation for the value. The average, minimum, and maximum effective R^2 across the 10 implicates are reported with standard deviations in parentheses.

are now consistent with the redesigned income questionnaire. Estimates are calculated for the number of recipients and mean income in the traditional, redesign, each of the Income-Consistent implicates as well as the multiple imputation estimates, shown in Table 4. Standard errors in each file or implicate are calculated using replicate weights. Throughout the article, multiple imputation standard errors for the Income-Consistent file are calculated using the multiple imputation variance formula in Rubin and Schenker (1986). There are no statistically significant differences when comparing recipiency or mean income between the redesign and Income-Consistent files for any of the three income sources.

Another statistic that can be used to evaluate the value of applying the imputation to create the Income-Consistent file is the estimated rate of missing information, denoted by γ (Rubin 1987). Very high values of γ (for example, 0.7) would imply that there is little additional benefit to using the traditional sample with imputed interest, dividend, and retirement income. As the relevance of the missing interest, dividend, and retirement income may differ for different statistics, for each parameter of interest, a separate γ can be computed.

I estimated γ for the recipiency and mean income statistics in Table 4. Recall that approximately 30% of the Income-Consistent sample comes from the redesign sample and is the same across all ten implicates. The rate of missing information varies across the income types from 0.09 (interest recipiency) to 0.53 (mean retirement income).

I also calculated γ for household median income and poverty of 0.15 and 0.08 respectively. Both of these are low values, which indicates that a considerable amount of information in estimating median income and poverty is contributed by the other variables in the traditional sample that were not imputed, as they were not affected by the redesign. For family and household income statistics, these low γ values also validate the general approach of combining the samples to take full advantage of the information available in the questions unaffected by the redesign.

5. Income and Poverty Statistics

To further assess the Income-Consistent file, median income and poverty statistics are calculated. These statistics are in the annual Income and Poverty reports published by the

Table 4. Recipiency and mean Income in traditional, redesign, and income-consistent files.

	Retirement		Interest		Dividends	
	Recipiency (Thousands)	Mean (USD)	Recipiency (Thousands)	Mean (USD)	Recipiency (Thousands)	Mean (USD)
Traditional	18,871	20,034	86,142	2,120	33,243	4,424
	(251)	(307)	(588)	(68)	(432)	(170)
Redesign	20,698	22,262	1,23,772	3,142	31,804	3,693
	(372)	(449)	(887)	(107)	(568)	(211)
Income-Consistent Implicate #						
1	20,709	22,406	1,25,594	3,033	32,095	3,584
	(265)	(271)	(1,866)	(75)	(587)	(172)
2	20,777	22,195	1,24,925	3,160	32,045	3,669
	(292)	(377)	(1,404)	(75)	(411)	(141)
3	20,511	22,054	1,25,339	3,168	31,236	3,706
	(241)	(251)	(1,834)	(87)	(386)	(109)
4	20,395	22,383	1,25,813	3,083	31,657	3,734
	(225)	(257)	(2,152)	(54)	(368)	(96)
5	20,826	22,667	1,24,748	3,115	32,167	3,633
	(455)	(291)	(1,228)	(55)	(616)	(122)
6	20,945	22,371	1,24,858	3,087	31,357	3,923
	(428)	(287)	(1,282)	(49)	(366)	(133)
7	20,891	22,826	1,24,804	3,082	31,660	3,706
	(520)	(377)	(1,191)	(51)	(295)	(135)
8	20,659	22,252	1,25,513	3,195	31,827	3,629
	(248)	(267)	(2,129)	(75)	(404)	(129)
9	20,788	22,019	1,25,805	3,037	31,741	3,597
	(327)	(353)	(2,026)	(74)	(302)	(165)
10	20,793	21,867	1,25,996	3,134	31,733	3,711
	(418)	(393)	(2,189)	(58)	(312)	(116)
Income-Consistent Multiple Imputation	20,729	22,304	1,25,340	3,109	31,752	3,689
	(398)	(442)	(1,841)	(88)	(525)	(168)
Rate of Missing Information (γ)	0.23	0.53	0.09	0.47	0.41	0.41

Source: U.S. Census Bureau, Current Population Survey, 2014 Annual Social and Economic Supplement. Standard errors (in parenthesis) are calculated using replicate weights. For more information, see "Standard Errors and Their Use" at <www.census.gov/hhes/www/p60_245sa.pdf>. Standard errors for the Income-Consistent file are calculated using the multiple imputation formula in Rubin and Schenker (1986).

Census Bureau from the CPS ASEC (available at http://www.census.gov/topics/income-poverty/income.html). Table 5 shows the median income statistics (Table 1 from the annual report) comparison between the Income-Consistent full sample and the traditional and redesign sample. Compared with the redesign sample, the only statistically significant differences are for median income of nonfamily households with a female householder (3.6% greater) and households headed by individuals without a disability (2.3% greater). At the 90% confidence interval, fewer than ten percent of the tested statistics are significantly different. For the comparison with the traditional sample, nearly all of the median income comparisons are statistically significant.

Table 6 compares poverty estimates in each of the traditional and redesign sample to the Income-Consistent file. The headline poverty number for all individuals is not statistically significantly different between the Income-Consistent file and either sample. However, for the traditional sample, poverty is lower in the Income-Consistent file for blacks (1.0%),

Table 5. Comparison of traditional, redesign, and income-consistent files: household median income by selected characteristics.

	Traditional (T)			Redesign (R)		Income-Consistent Multiple Imputation (IC)		(IC-T)/T	(IC-R)/R
Characteristic	Number (thousands)	Median (USD)	SE	Median (USD)	SE	Median (USD)	SE	% Diff	% Diff
All households	1,22,952	51,939	455	53,585	1,076	53,499	687	* 3.00	−0.16
Family households	81,192	65,587	643	66,923	872	67,301	673	* 2.61	0.56
Married-couple families	59,669	76,509	674	78,897	1,359	79,275	946	* 3.61	0.48
Female householder, no husband present	15,193	35,154	832	35,412	1,512	35,896	793	* 2.11	1.37
Male householder, no wife present	6,330	50,625	1,503	52,480	2,730	52,248	1,528	* 3.21	−0.44
Nonfamily households	41,760	31,178	518	31,480	951	31,977	539	* 2.56	1.58
Female householder	22,266	26,425	795	26,238	1,019	27,186	766	* 2.88	* 3.61
Male householder	19,494	36,876	937	39,379	1,674	38,242	1,094	* 3.70	−2.89
White	97,774	55,257	699	56,745	850	56,708	606	* 2.63	−0.06
White, not Hispanic	83,641	58,270	1,006	60,329	876	60,225	657	* 3.36	−0.17
Black	16,108	34,598	1,198	35,324	1,410	35,429	948	2.40	0.29
Asian	5,759	67,065	2,830	72,383	5,531	71,743	2,564	* 6.98	−0.88
Hispanic (any race)	15,811	40,963	908	39,687	1,954	41,341	894	0.92	4.17
Under 65 years	94,223	58,448	958	60,265	771	60,528	499	* 3.56	0.44
15 to 24 years	6,323	34,311	1,808	33,791	3,156	34,845	1,610	1.55	3.12
25 to 34 years	20,008	52,702	1,489	52,416	2,098	53,592	1,366	1.69	2.24
35 to 44 years	21,046	64,973	1,620	67,594	1,976	66,985	1,198	* 3.10	−0.90
45 to 54 years	23,809	67,141	1,265	70,598	2,114	70,671	1,214	* 5.26	0.10
55 to 64 years	23,036	57,538	1,662	60,481	1,835	60,735	1,503	* 5.56	0.42
65 years and older	28,729	35,611	722	37,297	1,283	36,352	808	2.08	−2.53
Native born	1,05,328	52,779	754	55,087	940	54,615	737	* 3.48	−0.86

Table 5. Continued.

	Traditional (T)			Redesign (R)		Income-Consistent Multiple Imputation (IC)		(IC-T)/T	(IC-R)/R
Characteristic	Number (thousands)	Median (USD)	SE	Median (USD)	SE	Median (USD)	SE	% Diff	% Diff
Foreign born	17,624	46,939	1,037	46,795	1,563	48,156	1,259	* 2.59	2.91
Naturalized citizen	9,491	54,974	2,898	56,354	3,098	57,406	1,947	* 4.42	1.87
Not a citizen	8,133	40,578	1,113	40,185	1,944	41,020	950	1.09	2.08
Households with householders aged 18 to 64	94,024	58,492	955	60,310	742	60,566	501	* 3.55	0.42
With disability	8,794	25,421	1,260	25,337	1,746	26,476	1,152	* 4.15	4.50
Without disability	84,784	61,979	564	62,487	1,021	63,924	892	* 3.14	* 2.30
Northeast	22,053	56,775	1,426	56,868	2,563	58,121	1,852	2.37	2.20
Midwest	27,214	52,082	1,160	53,426	2,102	53,207	1,427	* 2.16	−0.41
South	46,499	48,128	1,104	49,854	1,335	50,136	809	* 4.17	0.57
West	27,186	56,181	1,190	59,525	2,067	57,775	1,336	* 2.84	−2.94

The 2014 CPS ASEC included redesigned questions for income and health insurance coverage. All of the approximately 98,000 addresses were eligible to receive the redesigned set of health insurance coverage questions. The redesigned income questions were implemented to a subsample of these 98,000 addresses using a probability split panel design. Approximately 68,000 addresses were eligible to receive a set of income questions similar to those used in the 2013 CPS ASEC and the remaining 30,000 addresses were eligible to receive the redesigned income questions. * if statistically significant difference at the 90% confidence level. Income in 2013 US dollars. Households and people as of March of the following year. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see http://www2.census.gov/programs-surveys/cps/techdocs/cpsmar14.pdf. Standard errors calculated using replicate weights for the traditional and redesigned samples, and using replicate weights for each Income-Consistent implicate and the multiple imputation variance formula from Rubin and Schenker (1986) to combine estimates across implicates for the Income-Consistent estimates.

Table 6. Comparison of traditional, redesign, and income-consistent files: poverty by selected characteristics.

		Traditional				Redesign				Income-Consistent Multiple Imputation				IC-T	IC-R
Characteristic	Total Population	Number (thousands)	SE	% in Pov	SE	Number (thousands)	SE	% in Pov	SE	Number (thousands)	SE	% in Pov	SE	% in Pov	% in Pov
PEOPLE															
Total	3,12,970	45,318	616	14.5	0.2	46,269	896	14.8	0.3	45,257	549	14.5	0.2	0.0	− 0.3
Family Status															
In families	2,54,990	31,530	514	12.4	0.2	32,786	833	12.8	0.3	31,668	480	12.4	0.2	0.0	− 0.4
Householder	81,217	9,130	150	11.2	0.2	9,645	256	11.7	0.3	9,171	145	11.3	0.2	0.0	* − 0.5
Related children under 18	72,573	14,142	271	19.5	0.4	15,116	440	20.9	0.6	14,492	256	20.0	0.4	* 0.5	* − 0.9
Related children under 6	23,585	5,231	137	22.2	0.6	5,590	207	23.7	0.9	5,339	122	22.6	0.5	0.5	− 1.0
In unrelated subfamilies	1,413	608	69	43.0	3.8	776	134	47.7	5.1	622	62	42.5	3.2	− 0.5	− 5.2
Reference person	595	246	29	41.3	3.9	291	52	44.0	5.0	243	24	40.3	3.0	− 1.0	− 3.6
Children under 18	714	340	42	47.7	4.1	448	79	53.1	5.7	358	40	47.5	3.5	− 0.2	− 5.6
Unrelated individual	56,564	13,181	252	23.3	0.4	12,707	352	22.9	0.5	12,967	225	23.0	0.3	− 0.3	0.0
Race and Hispanic Origin															
White alone	2,43,080	29,936	496	12.3	0.2	31,287	652	12.9	0.3	30,235	415	12.4	0.2	0.1	* − 0.4
White alone, not Hispanic	1,95,170	18,796	439	9.6	0.2	19,552	495	10.0	0.3	19,102	349	9.8	0.2	0.1	− 0.3
Black alone	40,615	11,041	308	27.2	0.8	10,186	384	25.2	0.9	10,615	270	26.2	0.7	* − 1.0	1.0
Asian alone	17,063	1,785	107	10.5	0.6	2,255	201	13.1	1.2	1,881	100	11.1	0.6	0.6	* − 2.0
Hispanic (of any race)	54,145	12,744	312	23.5	0.6	13,356	487	24.7	0.9	12,750	299	23.5	0.6	0.0	− 1.1
Sex															
Male	1,53,360	20,119	345	13.1	0.2	20,294	468	13.2	0.3	20,101	291	13.1	0.2	0.0	− 0.1
Female	1,59,600	25,199	348	15.8	0.2	25,975	548	16.3	0.3	25,156	327	15.8	0.2	0.0	* − 0.5
Age															
Under 18 years	73,625	14,659	277	19.9	0.4	15,801	441	21.5	0.6	15,039	261	20.5	0.4	* 0.6	* − 1.1
18 to 64 years	1,94,830	26,429	394	13.6	0.2	25,899	533	13.3	0.3	26,000	358	13.3	0.2	− 0.2	0.0
65 years and over	44,508	4,231	138	9.5	0.3	4,569	174	10.2	0.4	4,218	135	9.5	0.3	0.0	* − 0.7

Table 6. Continued.

		Traditional				Redesign				Income-Consistent Multiple Imputation				IC-T	IC-R
Characteristic	Total Population	Number (thousands)	SE	% in Pov	SE	Number (thousands)	SE	% in Pov	SE	Number (thousands)	SE	% in Pov	SE	% in Pov	% in Pov
Nativity															
Native	2,71,970	37,921	573	13.9	0.2	38,831	790	14.3	0.3	38,064	492	14.0	0.2	0.0	−0.3
Foreign born	40,997	7,397	227	18.0	0.5	7,438	338	18.3	0.7	7,193	201	17.7	0.4	−0.4	−0.6
Naturalized citizen	19,147	2,425	105	12.7	0.5	2,132	151	11.1	0.8	2,245	89	11.7	0.4	* −0.9	0.7
Not a citizen	21,850	4,972	189	22.8	0.7	5,306	303	24.8	1.1	4,948	174	22.9	0.7	0.1	* −1.9

The 2014 CPS ASEC included redesigned questions for income and health insurance coverage. All of the approximately 98,000 addresses were eligible to receive the redesigned set of health insurance coverage questions. The redesigned income questions were implemented to a subsample of these 98,000 addresses using a probability split panel design. Approximately 68,000 addresses were eligible to receive a set of income questions similar to those used in the 2013 CPS ASEC and the remaining 30,000 addresses were eligible to receive the redesigned income questions. * if statistically significant difference at the 90% confidence level. Income in 2013 US dollars. Households and people as of March of the following year. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see ftp://ftp2.census.gov/programs-surveys/cps/techdocs/cps-mar13.pdf. Standard errors calculated using replicate weights for the traditional and redesigned samples, and using replicate weights for each Income-Consistent implicate and the multiple imputation variance formula from Rubin and Schenker (1986) to combine estimates across implicates for the Income-Consistent estimates.

naturalized citizens (−0.9%), residents of principal cities (0.5%), and workers (0.2%), and is higher for children (0.6%). For the redesign sample, unlike median income, there are significant differences: lower in poverty in the Income-Consistent file for children (1.1%) and those aged 65 and older (0.7%).

To summarize the results, the Income-Consistent file household median income estimates are more like the redesign file, but the poverty estimates lie between the two files. While the point estimate for poverty of 14.5% is not statistically significantly different from the point estimate for either file, it is much closer to the 14.5% estimate from the traditional file than the 14.8% estimate of the redesign file.

6. Conclusion

In this article, multiple imputation is applied to the problem of a split sample receiving different survey instruments in a bridge year. One possible way to use data from all survey respondents is shown, even though distinct sets of respondents answered different questions. This idea has an important potential benefit – by making use of all of the data during a bridge year, it potentially lowers the cost in terms of decreased statistical power of survey redesigns and bridges.

To address this problem of missing information during a survey bridge year, a semiparametric multiple imputation technique proposed by Bondarenko and Raghunathan is applied to the CPS ASEC 2014 redesign. The technique performs reasonably well and analysis of basic summary statistics shows how this technique affects important economic statistics that are widely reported on from the CPS ASEC.

For the 2014 CPS ASEC, this technique increases the potential sample that can be used to make comparisons to data from subsequent years, which uses the redesigned questionnaire for the entire sample. The larger sample facilitates analyses on subgroups, such as by state, where the redesign sample may lack the statistical power needed for comparisons. By combining the two samples, this technique may also address concerns about differences in sample composition raised in previous research by Mitchell and Renwick (2015).

7. References

Bondarenko, I. and T.E. Raghunathan. 2007. “Multiple Imputations Using Sequential Semi and Nonparametric Regressions.” In Proceedings Section on Survey Research Methods: American Statistical Association, July 29, 2007. 3293–3300. Salt Lake City, UT. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000624.pdf (accessed November 2016).

Clogg, C.C., D.B. Rubin, N. Schenker, B. Schultz, and L. Weidman. 1991. “Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression.” *Journal of the American Statistical Association* 86: 68–78. Doi: http://dx.doi.org/10.1080/01621459.1991.10475005.

Czajka, J.L. and G. Denmead. 2008. “Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys.” Mathematica Reference No.: 6302-601. Available at:

https://www.mathematica-mpr.com/~/media/publications/PDFs/incomedata.pdf (accessed November 2016).

Hokayem, C., T.E. Raghunathan, and J. Rothbaum. 2015. "SRMI in the CPS ASEC." *Unpublished Manuscript*.

He, Y. and T.E. Raghunathan. 2006. "Tukey's gh Distribution for Multiple Imputation." *The American Statistician* 60: 251–256. Doi: http://dx.doi.org/10.1198/000313006X126819.

Hicks, W. and J. Kerwin. 2011. "Cognitive Testing of Potential Changes to the Annual Social and Economic Supplement of the Current Population Survey." Unpublished Westat report to the U.S. Census Bureau. July 25, 2011.

Butrica, B.A., I. Howard, K.E. Smith, and E.J. Toder. 2009. "The Disappearing Defined Benefit Pensions and Its Potential Impact on the Retirement Income of Baby Boomers." *Social Security Bulletin* 69: 1–28.

Meyer, B.D., W.K.C. Mok, and J.X. Sullivan. 2009. "The Under-Reporting of Transfers in Household Surveys: its Nature and Consequences." *National Bureau of Economic Research Working Paper #15181*.

Mitchell, J. and T. Renwick. 2015. "A Comparison of Official Poverty Estimates in the Redesigned Current Population Survey Annual Social and Economic Supplement." *U.S. Census Bureau SEHSD Working Paper #2014-35*.

Rosenbaum, P.R. and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. Doi: http://dx.doi.org/10.1093/biomet/70.1.41.

Rothbaum, J. 2015. "Comparing Income Aggregates: How Do the CPS and ACS Match the National Income and Product Accounts, 2007–2012." *U.S. Census Bureau SEHSD Working Paper #2015-01*.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons. Doi: http://dx.doi.org/10.1002/9780470316696.

Rubin, D.B. and N. Schenker. 1986. "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81: 366–374. Doi: http://dx.doi.org/10.1080/01621459.1986.10478280.

Schenker, N. and J.D. Parker. 2003. "From Single-Race Reporting to Multiple-Race Reporting: Using Imputation Methods to Bridge the Transition." *Statistics in Medicine* 22: 1571–1587. Doi: http://dx.doi.org/10.1002/sim.1512.

Semega, J. and E. Welniak Jr. 2013. "Evaluating the 2013 CPS ASEC Income Redesign Content Test." Presented at the proceeding of the 2013 FCSM Conference. U.S. Census Bureau Income Statistics Working Paper. Available at: http://census.gov/library/working-papers/2013/demo/semega-01.html (accessed November 2016).

Semega, J. and E. Welniak Jr. 2015. "The Effects of the Changes to the Current Population Survey Annual Social and Economic Supplement on Estimates of Income." Presented at the Proceedings of the 2015 Allied Social Science Association (ASSA) Research Conference. U.S. Census Bureau Income Statistics Working Paper. Available at: http://www.census.gov/library/working-papers/2015/demo/cpsasec-red-income.html (accessed November 2016).

Tjur, T. 2009. "Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination." *The American Statistician* 63: 366–372. Doi: http://dx.doi.org/10.1198/tast.2009.08210.

Woodcock, S.D. and G. Benedetto. 2009. "Distribution-Preserving Statistical Disclosure Limitation." *Computational Statistics & Data Analysis* 53: 4228–4242. Doi: http://dx.doi.org/10.1016/j.csda.2009.05.020.

Received November 2015
Revised December 2016
Accepted December 2016

Supplemental data, Table 1. CPS hot deck imputation for missing earnings from longest job.

	Match level					
Match variable	1	2	3	4	5	6
Sex	2	2	2	2	2	2
Race	3	2	2			
Age	9	6	3	3		
Relationship	7	7	4	4	4	
Years of school completed	6	5	5	4	4	4
Marital status	4	4				
Presence of children	3					
Labor force status of spouse	3					
Weeks worked	5	5	4	4	4	4
Hours worked	3	3	3	3	2	
Occupation	528	528	66	66	66	
Class of worker	5	5	5	3	3	3
Other earnings	8	8				
Type of residence	3	2	2			
Region	4	4				
Transfers payments receipt	2	2	2	2		
Number of cells	620,786,073,600	17,031,168,000	3,801,600	456,192	50,688	96

Supplemental data, Table 2. Potential predictor variables in imputation model.

Recipiency/binary variables	Values	Other variables
Earnings	Earnings	Race/Ethnicity - White
Other Job Wage	Other Job Wage	Race/Ethnicity - Black
Other Job Self-Employment	Other Job Self-Employment	Race/Ethnicity - Native American
Other Job Farm Self-Employment	Other Job Farm Self-Employment	Race/Ethnicity - Asian
Unemployment Compensation	Unemployment Compensation	Race/Ethnicity - Pacific Islander
Veterans' Benefits	Veterans' Benefits	Race/Ethnicity - Hispanic
Survivors' Benefits	Survivors' Benefits (Source 1)	Age
Rental Income	Survivors' Benefits (Source 2)	Weeks Worked in Last Year
Educational Assistance	Rental Income	Usual Hours Worked
Child Support Payment	Educational Assistance	Property Value (Non-imputed)
Financial Assistance	Child Support Payment	Gender
Spouse/Partner Present	Financial Assistance	Supplement Weight (Full Sample)
Spouse/Partner Earnings	Spouse/Partner Earnings	Relationship to Household Head
Interest Income	Interest Income	Education
Dividend Income	Dividend Income	Marital Status
		Children in Family
		Occupation
		City Type of Residence (Urban/CBSA)
		Census Region
		Transfer Payments/ Program Participation

This table lists the potential predictor variables in the imputation model. The more detailed list, with interaction terms and CPS ASEC variable codes, is available from the author upon request. The large number of variables in the model come from the conversion of categorical variables to sets of dummy variables, the inclusion of multiple recodes of particular (for example age, age squared, dummies for ages relevant to retirement such as 62, 65, and 70), and the inclusion of a large number of possible interactions.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 207–222, http://dx.doi.org/10.1515/JOS-2017-0011

Adjusting for Misclassification: A Three-Phase Sampling Approach

Hailin Sang[1], *Kenneth K. Lopiano*[2], *Denise A. Abreu*[3], *Andrea C. Lamas*[3], *Pam Arroway*[4], *and Linda J. Young*[3]

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) conducts the June Agricultural Survey (JAS) annually. Substantial misclassification occurs during the prescreening process and from field-estimating farm status for nonresponse and inaccessible records, resulting in a biased estimate of the number of US farms from the JAS. Here, the Annual Land Utilization Survey (ALUS) is proposed as a follow-on survey to the JAS to adjust the estimates of the number of US farms and other important variables. A three-phase survey design-based estimator is developed for the JAS-ALUS with nonresponse adjustment for the second phase (ALUS). A design-unbiased estimator of the variance is provided in explicit form.

Key words: Estimation under the three-phase sampling design; nonresponse; unbiased estimator; variance estimation.

1. Introduction

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) conducts numerous statistical surveys to provide information about current and future supplies of agricultural commodities. See Fecso et al. (1986), Vogel (1995), and Nusser and House (2009) for the evolution and development of agricultural statistics and the surveys conducted at the United States Department of Agriculture. The June Agricultural Survey (JAS) is conducted annually. For the JAS, a stratified random sample is drawn using an area frame, which ensures complete coverage. Information about US crops, livestock, grain storage capacity, type and size of farms are collected from agricultural operations in the sample. NASS uses the JAS to estimate numerous items relating to US agriculture, including the number of farms.

[1] Department of Mathematics, University of Mississippi, University, MS 38677, USA. Email: sang@olemiss.edu
[2] Roundtable Analytics, Research Triangle Park, NC 27709, USA. Email: klopiano@roundtableanalytics.com
[3] National Agricultural Statistics Service Research and Development, USDA, 1400 Independence Ave. SW, Room 6035, USA. Emails: Denise.Abreu@nass.usda.gov, Andrea.Lamas@nass.usda.gov, and Linda.Young@nass.usda.gov
[4] EDUCAUSE, 282 Century Place, Ste 5000 Louisville, CO 80027, USA. Email: parroway@educause.edu

Acknowledgments: The authors would like to thank the Editor, the Associate Editor, and the referees for numerous suggestions that improved the presentation of this article.

This work was supported by the National Agricultural Statistics Service, the United States Department of Agriculture, and the National Institute of Statistical Sciences. Hailin Sang was at the National Institute of Statistical Sciences, Pam Arroway was on faculty at North Carolina State University, and Kenneth K. Lopiano and Linda J. Young were at the University of Florida when most of this research was conducted.

Every five years, the annual number of farms estimate is compared to the one obtained from the quinquennial Census of Agriculture, which is a dual-frame survey conducted during years ending in 2 and 7. See Kott and Vogel (1995) for details on the dual-frame survey. In 2007, the difference between the estimated number of farms from the JAS and the 2007 Census of Agriculture could not be attributed to sampling error alone. A preliminary study showed that the JAS estimate was biased because some farms were incorrectly classified as non-farms. In addition, some non-farms were misclassified as farms, but at a lower rate. Prior to this study, NASS had assumed that no misclassification was present in the JAS or any other survey that it conducted.

Bross (1954) first showed that, when misclassification is present, conventional methods can be seriously biased. Tenenbein (1970, 1972) proposed a double-sampling scheme for inference from categorical data subject to misclassification. The double-sampling schemes utilize a sample of n_1 units classified by both a fallible and true device, and another sample of n_2 units classified only by a fallible device. The double-sampling scheme and its variants are popular approaches to estimation when misclassification is present (see Thall et al. 1996, Stewart et al. 1998, and the references therein). Bayesian methods are also popular for inference from categorical data subject to misclassification (see Swartz et al. 2004, the book by Gustafson 2003, and the references therein).

In this article, a design-based approach that addresses misclassification and leads to improved estimates of the number of farms is suggested. First, the JAS sampling design is discussed, with an emphasis on the factors leading to the misclassification of farms and non-farms. Then, a proposed revision to the JAS sampling design is presented, and the properties of the resulting farm number estimates from this revised design are explored. Finally, the implications of the work on the JAS are considered.

2. The June Agricultural Survey (JAS)

The JAS is conducted annually utilizing an area frame, ensuring complete coverage of the population. Land within the JAS area frame is divided into homogeneous land-use strata. Although minor definitional adjustments may be made depending on the specific needs of the state, land-use strata with more than 50% cultivated land are generally labeled with a value in the 10s, agri-urban and commercial land-use strata are typically given a label in the 30s, and so on (see Table 1). The general land-use strata definitions are similar from state to state; however, minor definitional adjustments may be made depending on the specific needs of a state. Each land-use stratum is further divided into substrata (called "design strata") by grouping areas that are agriculturally similar, providing greater

Table 1. Land-use strata.

Land-use strata	
≥50% cultivated land	10s
15–49% cultivated land	20s
<15% cultivated land	40s
Agri-urban/Commercial areas	30s
Non-agricultural land	50s

precision for state-level estimates of individual commodities. Within each design stratum, the land is divided into primary sampling units (PSUs). A sample of PSUs is selected and smaller, similar-sized segments (each of about a square mile (640 acres)) of land are delineated within these selected PSUs. Finally, one segmentis randomly selected from each selected PSU to be fully enumerated.

Once selected for inclusion in the JAS, a segment stays in the sample for five years. Thus, each year the sample has about 20% new segments, and the 20% of the segments that have been in the sample for five years rotate out. Segments rotating in during the same year are called replicates; thus, each JAS sample consists of five replicates (see Cotter et al. 2010, Benedetti et al. 2015, for further details on JAS).

Through 2010, the JAS prescreening was conducted in the two weeks prior to data collection. During prescreening, field enumerators (data collectors) divide each segment into tracts of land. Each tract represents a unique land operating arrangement. Field enumerators do not interview tract operators during prescreening. Instead, they complete an area screening form which provides an inventory of all tracts within a sampled segment, and contains screening questions that determine whether or not each tract has agricultural activity. Using this form, each tract within the segment is screened for agricultural activity, and the screening applies to all land in the identified operating arrangement. Each screened tract is classified as agricultural or non-agricultural. Non-agricultural tracts are assigned to one of three categories: (1) non-agricultural with potential, (2) non-agricultural with unknown potential, or (3) non-agricultural with no potential.

The JAS is conducted during the first two weeks of June. During the sampling period, field enumerators return to only those tracts classified as agricultural during the earlier screening period. Data collection continues until some type of response is obtained for every sampled tract. If a respondent cannot be reached, the information may be obtained from administrative data, data collected for other surveys, or estimates made by field enumerators. Regardless of the information source, these tracts are identified as being field estimated. Based on the JAS, an agricultural tract is classified as a farm if its entire operation, which could include land outside the sampled tract, qualifies with at least USD 1,000 in agricultural sales or potential sales. All non-agricultural tracts and agricultural tracts with less than USD 1,000 in sales are classified as non-farms.

In 2009, NASS conducted a one-time follow-on survey to the JAS segments, the Farm Numbers Research Project (FNRP) (Abreu et al. 2010). The sampling design of the FNRP targeted the 20% of JAS segments that were newly rotated in for 2009 (2009 segments). All tracts in the 2009 segments that were non-agricultural or field estimated in JAS were selected for FNRP. During the FNRP, all places of interest within a selected tract were considered subtracts.

A shortened form based on the JAS questionnaire was used to classify each subtract as a farm or as a non-farm.

A major finding in FNRP was that, assuming misclassification rates are the same for all rotations (did not differ from that observed for the 2009 segments), the JAS estimate of the number of farms would increase by approximately 580,000 (see Table 2). The bulk of these farms were found in tracts that had been identified as non-agricultural with no potential in the JAS.

Table 2. FNRP results by type of tract.

Type of tract	FNRP sample size (subtracts)	Number of FNRP farms	Net expanded number of farms
Field estimated as farm	1,591	1,466	(7,822)
Field estimated as non-farm	121	37	13,032
Non-agricultural with potential	487	95	38,346
Non-agricultural with unknown potential	364	56	37,479
Non-agricultural with no potential	14,628	905	500,338
FNRP total	17,191	2,559	581,373

Several factors could lead to the misclassification of farms as non-farms and of non-farms as farms. During prescreening, the agricultural activity may not have been evident when the field enumerator observed the tract from a distance (tract operators are not interviewed during this process), or the primary agricultural activity could have been outside the sampled tract (the response for a tract includes agriculture associated with all of the operation, not just that within the tract). In FNRP, 86.1% (500,338) of the field-estimated number of farms misclassified as non-farms were found in tracts prescreened to be non-agricultural with no potential. Small farms are more likely to be misclassified. In FNRP, 58.3% (335,902) of the field-estimated number of farms misclassified as non-farms had less than 25 acres. Operations that recently went out of business or small farms whose production fell below the USD 1000 threshold in sales could be misclassified as farms when field estimated.

To obtain a more accurate estimate of the number of US farms from the JAS, the current estimation approach must be revised to account for misclassification. The Annual Land Utilization Survey (ALUS), a follow-on survey to the JAS, has been proposed for this purpose. FNRP results are used as guidelines for the ALUS design, but ALUS will be able to detect different types of trends as well.

3. The Annual Land Utilization Survey (ALUS): Design

The ALUS focuses on those JAS tracts that were potentially misclassified as farm or non-farms either during the prescreening process or during field estimation of farm status for nonresponding or inaccessible operations. These tracts are treated as nonresponders, and data collection is focused on obtaining accurate information on them. ALUS represents the second phase of a two-phase sample, with the first phase being the traditional JAS. As in the JAS, the proposed ALUS is a stratified sample of segments, using JAS land-use strata and sampling across rotations. Segments that are eligible for inclusion in ALUS must have at least one tract that was prescreened as non-agricultural (regardless of potential) or that was field estimated in JAS (as either a farm or non-farm); that is, only JAS segments that had completed interviews for all tracts are not eligible for possible inclusion in the ALUS sample. For a selected segment, all tracts are to be reevaluated using a modified combined JAS-ALUS questionnaire. The collection of eligible segments in a particular year will be called the ALUS population.

Table 3. Guidelines for ALUS allocation scheme.

Land-use strata	Proportion of FNRP adjustment from non-agricultural tracts (%)	Proportion of ALUS-eligible segments in 2009 JAS (%)	Proportion of ALUS-eligible segments in 2010 JAS (%)	Suggested Proportion of ALUS sample (%)
10s	16	53	52	27
20s	34	26	27	30
30s	<1	3	3	3
40s	50	17	17	39
50s	<1	<1	<1	1
Total	576,000 farms	10,168 segments	10,121 segments	

For ALUS, the sample allocation of segments to each state-stratum combination considers two factors: the proportion of the ALUS population in the land-use stratum and the proportion of the FNRP adjustment from non-agricultural tracts in the land-use stratum (see Table 3). The latter simultaneously accounts for the number of converted non-agricultural tracts and the expansion factors associated with them, allowing states and land-use strata that contributed most to the FNRP adjustment to be targeted. In the JAS, the sampling scheme favors cultivated areas. For ALUS, the sampling will lean more heavily on moderately and less cultivated land-use strata where the largest portion of the FNRP adjustment originates. For example, although the exact land-use stratum definition varies from state to state, land-use strata 10s $(10, 11, \cdots)$ are highly cultivated areas, with generally at least 50% cultivated land. In the JAS, over half of the selected segments are from these land-use strata. However, 10s made uponly 16% of the FNRP adjustment arising from non-agricultural tracts, so only about 27% of the ALUS sample will come from these strata. The sample will be evenly distributed over the five rotations, with approximately 20% of the ALUS sample selected from each.

Within each land-use stratum of the ALUS population, segments will be selected with probability proportional to size (pps) sampling where the size measure of a segment is defined as the sum of the number of tracts either prescreened as non-agricultural or field estimated to be non-farms, and one-tenth of the number of tracts field estimated to be a farm. Because most tracts (92%) field estimated as farms in the JAS were confirmed as farms in FNRP, ALUS only takes a tenth of the number of these tracts within a segment when determining size. If a segment is selected, all ALUS-eligible tracts withinthat segment will be in the sample, including those field estimated as farms.

Precise estimates of uncertainty can be obtained by viewing the combination of JAS and ALUS as a two-phase sample, with JAS being the first phase and ALUS being the second. Given that each phase makes use of a probability sampling design with known inclusion probabilities, standard results can be used to construct a design-based estimator (Särndal and Swensson 1987). However, nonresponse is also expected to occur in ALUS.

Instead of using the estimated tract values to account for this nonresponse, the two-phase design estimator of Särndal and Swensson (1987) has been extended to a third phase (see Section 4). The resulting estimator is used for the two-phase JAS-ALUS, with the self-selection of response treated as a third phase of random sampling. This methodology can be applied not only to estimates of the number of farms but to all variables collected in the ALUS.

4. Estimation

In this section we first extend the two-phase π^* estimator (Särndal and Swensson 1987) to a three-phase survey sampling estimator. Legg and Fuller (2009), Särndal et al. (1992) and Singh (2003) provide a review of the two-phase sampling estimator. Jeyaratnam et al. (1984) studied a multiphase design in a forest study. Fuller (2003) studied a three-phase regression estimator for the mean of a vector population. Magnussen (2003) studied estimators for three-phase sampling of categorical variables. Then in the second subsection, we study the application to the ALUS estimator with nonresponse adjustment.

4.1. Estimation Under a Three-Phase Sampling Design

To be consistent and complete, the notation used by Särndal and Swensson (1987) for the two-phase design is extended for the third phase.

Let y_k be the response of interest for the kth unit in a finite population U. The population total is $T = \sum_U y_k$. A general sampling design is allowed in each phase.

(a) The first-phase sample $S(S \subset U)$ is drawn according to a sampling design $P_a(\cdot)$, such that $P_a(S)$ is the probability of choosing S. The inclusion probabilities are defined by

$$\pi_{ak} = \sum_{k \in S} P_a(S),\ \pi_{akp} = \sum_{k,p \in S} P_a(S)$$

with $\pi_{akk} = \pi_{ak}$. Set $\Delta_{akp} = \pi_{akp} - \pi_{ak}\pi_{ap}$. It is assumed that $\pi_{ak} > 0$ for all k, and $\pi_{akp} > 0$ for all $k \neq p$ in variance estimation. π_{ak} is the probability of selection of the kth unit in the first-phase sampling. π_{akp} is the probability of selection both the kth unit and the pth unit in the first-phase sampling.

(b) Given S, the second-phase sample $R(R \subset S)$ is drawn according to a sampling design $P(\cdot|S)$, such that $P(R|S)$, is the conditional probability of choosing R. The inclusion probabilities given S are defined by

$$\pi_{k|S} = \sum_{k \in R} P(R|S),\ \pi_{kp|S} = \sum_{k,p \in R} P(R|S).$$

$\pi_{kk|S} = \pi_{k|S}$. Set $\Delta_{kp|S} = \pi_{kp|S} - \pi_{k|S}\pi_{p|S}$. It is assumed that for any S, $\pi_{k|S} > 0$ for all $k \in S$, and $\pi_{kp|S} > 0$ for all $k \neq p \in S$ in variance estimation. $\pi_{k|S}$ is the probability of selection of the kth unit in the second-phase sampling given the result of the first-phase sampling. $\pi_{kp|S}$ is the probability of selecting both the kth unit and the pth unit in the second-phase sampling given the result of the first-phase sampling.

(c) Given R, the third-phase sample $F(F \subset R)$ is drawn according to a sampling design $P(\cdot|R)$, such that $P(F|R)$ is the conditional probability of choosing F. F is the set of selected units in a three-phase sampling design or the set of responses for the second phase in a two-phase sampling design. The inclusion probabilities given R are defined by

$$\pi_{k|R} = \sum_{k \in F} P(F|R),\ \pi_{kp|R} = \sum_{k,p \in F} P(F|R).$$

$\pi_{kk|R} = \pi_{k|R}$. Set $\Delta_{kp|R} = \pi_{kp|R} - \pi_{k|R}\pi_{p|R}$. In a three-phase sampling design, $\pi_{k|R}$ is the probability of selection of the kth unit in the third phase of sampling given the result of the first two phases of sampling. $\pi_{kp|R}$ is the probability of selecting both the kth unit and the pth unit in the third phase of sampling given the result of the first two phases of sampling. In a two-phase sampling design, $\pi_{k|R}$ is the probability when the kth unit has response for the second phase. $\pi_{kp|R}$ is the probability that both the kth unit and the pth unit have a response for the second phase.

Now, for any S and for all $k, p \in S$, define $\pi^{*}_{k} = \pi_{ak}\pi_{k|S}$, $\pi^{*}_{kp} = \pi_{akp}\pi_{kp|S}$. $\pi^{*}_{kk} = \pi^{*}_{k}$. Next, define $\pi^{\#}_{k} = \pi^{*}_{k}\pi_{k|R} = \pi_{ak}\pi_{k|S}\pi_{k|R}$ for all $k \in R$ and any R. Then the first-phase expanded y-value is $\breve{y}_k = y_k/\pi_{ak}$. The second-phase expanded y-value is $\breve{y}^{*}_{k} = \breve{y}_k/\pi_{k|S} = y_k/\pi^{*}_{k}$. The third-phase expanded y-value is $\breve{y}^{\#}_{k} = \breve{y}^{*}_{k}/\pi_{k|R} = \breve{y}_k/(\pi_{k|S}\pi_{k|R}) = y_k/(\pi_{ak}\pi_{k|S}\pi_{k|R}) = y_k/\pi^{\#}_{k}$. The expanded Δ values are $\breve{\Delta}_{akp} = \Delta_{akp}/\pi_{akp}$, $\breve{\Delta}^{*}_{kp|S} = \Delta_{akp}/\left(\pi^{*}_{kp}\right) = \Delta_{akp}/\ (\pi_{akp}\pi_{kp|S})$. $\breve{\Delta}_{kp|S} = \Delta_{kp|S}/\pi_{kp|S}$. Now, the expansion estimator in three-phase sampling is defined as

$$\hat{t}_{\#} = \sum_{k\in F}\breve{y}^{\#}_{k} = \sum_{k\in F} y_k/\pi^{\#}_{k}. \tag{1}$$

The following theorem gives an unbiased estimator of the variance of the triple expansion estimator $\hat{t}_{\#}$.

Theorem 1. *The estimator in (1) is design unbiased, and a design-unbiased estimator of* $Var(\hat{t}_{\#})$ *is given by*

$$\begin{aligned}\widehat{Var}\,(\hat{t}_{\#}) = &\sum\sum_{F}\breve{\Delta}^{*}_{kp|S}\breve{y}_k\breve{y}_p/\pi_{kp|R} + \sum\sum_{F}\breve{\Delta}_{kp|S}\breve{y}^{*}_{k}\breve{y}^{*}_{p}/\pi_{kp|R}\\ &+\sum\sum_{F}\Delta_{kp|R}\breve{y}^{\#}_{k}\breve{y}^{\#}_{p}/\pi_{kp|R}.\end{aligned} \tag{2}$$

The proof of Theorem 1 is deferred to the Appendix.

4.2. The ALUS Estimator

Let T be the number of US farms in a specific year. First, consider the JAS estimate of the number of farms. Then the estimator incorporating the information obtained during the ALUS (second-phase sample) and the nonresponse adjustment in ALUS will be developed.

Under stratified simple random sampling, the JAS estimator of T is

$$\hat{T} = \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}\sum_{k=1}^{n_{ij}}\sum_{m=1}^{x_{ijk}} t_{ijkm} \tag{3}$$

where

- i is the index of land-use stratum, l is the number of land-use strata;
- j is the index of design stratum, s_i is the number of design strata in land-use stratum i;
- k is the index of segment, n_{ij} is the number of segments in design stratum j within land-use stratum i;
- d_{ij} is the expansion factor or the inverse of the probability of selection for each segment in design stratum j in land-use stratum i;

- m is the index of tract, x_{ijk} is the number of farm tracts in the segment; and
- t_{ijkm} is the tract-to-farm ratio, which is $\frac{\textit{tract acres for the } m^{th} \textit{ tract}}{\textit{farm acres for the } m^{th} \textit{ tract}}$.

Under the assumption that the JAS provides accurate information for all tracts, $\hat{T}$ is unbiased. The variance is

$$Var(\hat{T}) = \sum_{i=1}^{l}\sum_{j=1}^{s_i}\frac{1-1/d_{ij}}{1-1/n_{ij}}\sum_{k=1}^{n_{ij}}(c_{ijk}-c_{ij.})^2 \tag{4}$$

where $c_{ijk} = d_{ij}\sum_{m=1}^{x_{ijk}} t_{ijkm}$, $c_{ij.} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} c_{ijk}$. This formula is given by Kott (1990).

However, the JAS estimate is biased because some tracts are misclassified either during prescreening when agricultural tracts may be identified as non-agricultural, or during the JAS when tracts are incorrectly field estimated to be farms or non-farms.

Now consider the JAS-ALUS two-phase estimator with nonresponse adjustment for the second phase. The estimator is

$$\hat{\hat{T}} = \hat{T}_1 + \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}a_{ij}\sum_{k=1}^{n'_{ij}} r_{ijk}\sum_{m=1}^{z_{ijk}} t_{ijkm} := \hat{T}_1 + \hat{T}_2. \tag{5}$$

Here, the first term $\hat{T}_1$ has the same form as $\hat{T}$ in (3). However, it only includes the JAS segments comprised of all farm tracts confirmed through an interview with the operator (not estimated) in the first phase. In the second phase, the ALUS sample only includes the JAS tracts that were either prescreened as non-agricultural or field estimated as either a farm or a non-farm. Thus each tract in the ALUS sample has been potentially misclassified and is treated as a nonrespondent from the first phase. n'_{ij} is the number of ALUS segments in design stratum j within land-use stratum i. a_{ij} is the expansion factor or the inverse of the probability of selection in the second phase for each segment in design stratum j in land-use stratum i. z_{ijk} is the number of farm tracts in the given ALUS-selected segment. r_{ijk} is the expansion factor or the inverse of the response probability of each tract in segment k, design stratum j, land-use stratum i.

Here we assume that all tracts in the same segment have the same response probability and this probability r_{ijk} is known. If r_{ijk} is unknown, it can be estimated by modeling under the assumption of stratified Bernoulli subsampling for nonresponse, that is, a response is assumed to have the Bernoulli distribution. In this case, we would have another variance component. This is a complex case and is not considered here. A referee suggested that, instead of assuming r_{ijk} known, the last phase could be treated conditionally (on the number of good responses) as a simple random sample within each segment. The assumption needed for this approach is for at least two responses to be obtained within each segment. Readers are referred to Särndal et al. (1992) for the modeling on nonresponse in a quasi-design-based framework ("quasi" because response if modeled). Hidiroglou and Estevao (2013) used a follow-up sample of the nonrespondents to deal with nonresponse.

Now we apply (2) in Theorem 1 to obtain a design-unbiased estimator of $Var(\hat{T}_2)$. For convenience, we use (i,j) to denote design stratum j within land-use stratum i. We also use k or p to be the index of segment. In the JAS-ALUS sampling design, the unit is a segment. One unit is one segment in (i,j). It includes all tracts in that segment. Recall that all segments within the same design stratum have the same expansion factor. The first phase expansion factor is $d_{ijk} = d_{ij}$ and the second phase expansion factor is $a_{ijk} = a_{ij}$ for all segments k in (i,j). Therefore, $\pi_{ak} = d_{ijk}^{-1} = d_{ij}^{-1}$, and

$$\breve{y}_k = y_k/\pi_{ak} = d_{ij}\sum_{m=1}^{z_{ijk}} t_{ijkm}. \tag{6}$$

$\pi_{k|S} = a_{ij}^{-1}$. There are $n'_{ij}a_{ij}d_{ij}$ segments in (i,j). If $k \neq p$ and these segments are in a same design stratum (i,j),

$$\pi_{akp} = (n'_{ij}a_{ij} - 1)/[d_{ij}(n'_{ij}a_{ij}d_{ij} - 1)],$$

$$\Delta_{akp} = \pi_{akp} - \pi_{ak}\pi_{ap} = (1 - d_{ij})/\left[d_{ij}^2(n'_{ij}a_{ij}d_{ij} - 1)\right],$$

$$\breve{\Delta}_{akp} = \Delta_{akp}/\pi_{akp} = (1 - d_{ij})/[d_{ij}(n'_{ij}a_{ij} - 1)],$$

$$\pi_{kp|S} = (n'_{ij} - 1)/[a_{ij}(n'_{ij}a_{ij} - 1)].$$

If k, p are from different design strata $(i, j), (i', j')$, $\Delta_{akp} = 0$. $\breve{\Delta}_{akp} = 0$. $\pi_{kp|S} = 1/(a_{ij}a_{i'j'})$. If $k = p$,

$$\pi_{akk} = \pi_{ak} = d_{ij}^{-1},$$

$$\Delta_{akk} = d_{ij}^{-1} - d_{ij}^{-2},$$

$$\breve{\Delta}_{akk} = \Delta_{akk}/\pi_{akk} = 1 - d_{ij}^{-1},$$

$$\pi_{kk|S} = \pi_{k|S} = a_{ij}^{-1}.$$

Therefore,

$$\breve{\Delta}^*_{kp|S} = \breve{\Delta}_{akp}/\pi_{kp|S} = a_{ij}(1 - d_{ij})/[d_{ij}(n'_{ij} - 1)]$$

if $k \neq p$ are in the same design stratum. $\breve{\Delta}^*_{kp|S} = 0$ if k, p are from different design strata.

$$\breve{\Delta}^*_{kp|S} = \breve{\Delta}_{akp}/\pi_{kp|S} = [a_{ij}(d_{ij} - 1)]/d_{ij} \tag{7}$$

if $k = p$. In the second phase of ALUS, recall that $\pi_{k|S} = \pi_{kk|S} = 1/a_{ij}$, $\pi_{kp|S} = (n'_{ij} - 1)/[a_{ij}(n'_{ij}a_{ij} - 1)]$ if the two different segments are in the same design stratum. Otherwise, $\pi_{kp|S} = 1/(a_{ij}a_{i'j'})$. Therefore,

$$\breve{y}^*_k = \breve{y}_k/\pi_{k|S} = d_{ij}a_{ij}\sum_{m=1}^{z_{ijk}} t_{ijkm}. \tag{8}$$

$$\Delta_{kp|S} = \pi_{kp|S} - \pi_{k|S}\pi_{p|S} = (1 - a_{ij})/\left[a_{ij}^2(n'_{ij}a_{ij} - 1)\right]$$

and

$$\breve{\Delta}_{kp|S} = \Delta_{kp|S}/\pi_{kp|S} = (1 - a_{ij})/[a_{ij}(n'_{ij} - 1)]$$

if the two different segments are in the same design stratum. $\Delta_{kp|S} = 0 = \check{\Delta}_{kp|S}$ if the two segments are in different design strata. $\Delta_{kp|S} = \pi_{kp|S} - \pi_{k|S}\pi_{p|S} = (a_{ij} - 1)/\left(a_{ij}^2\right)$ and $\check{\Delta}_{kp|S} = \Delta_{kp|S}/\pi_{kp|S} = (a_{ij} - 1)/a_{ij}$ if $k = p$. $\pi_{k|R}$ is the probability of response of the tracts in segment k. $\pi_{kp|R}$ is the probability that two tracts have response in segments k, p. $\pi_{k|R} = \pi_{kk|R} = 1/r_{ijk}$ and $\pi_{kp|R} = 1/(r_{ijk}r_{ijp})$ if $k \neq p$. Then $\Delta_{kp|R} = \pi_{kp|R} - \pi_{k|R}\pi_{p|R} = 0$ if $k \neq p$ and $\Delta_{kk|R} = \pi_{k|R} - \pi_{k|R}^2 = (r_{ijk} - 1)/r_{ijk}^2$. By (8), the third-phase expanded y-value

$$\check{y}_k^{\#} = \check{y}_k^{*}/\pi_{k|R} = d_{ij}a_{ij}r_{ijk}\sum_{m=1}^{z_{ijk}} t_{ijkm}.$$

Together with all the analysis, the design-unbiased estimator (2) of $Var(\hat{T}_2)$ is

$$\begin{aligned}\widehat{Var}(\hat{T}_2) =& \sum_{i=1}^{l}\sum_{j=1}^{s_i} a_{ij}d_{ij}(d_{ij} - 1)\sum_{k=1}^{n'_{ij}} r_{ijk}\left(\sum_{m=1}^{z_{ijk}} t_{ijkm}\right)^2 \\ &+ \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}a_{ij}(1 - d_{ij})(n'_{ij} - 1)^{-1}\sum_{1\le k<p\le n'_{ij}}\left(\sum_{m=1}^{z_{ijk}} r_{ijk}t_{ijkm}\sum_{m=1}^{z_{ijp}} r_{ijp}t_{ijpm}\right) \\ &+ \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}^2 a_{ij}(a_{ij} - 1)\sum_{k=1}^{n'_{ij}} r_{ijk}\left(\sum_{m=1}^{z_{ijk}} t_{ijkm}\right)^2 \\ &+ \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}^2 a_{ij}(1 - a_{ij})(n'_{ij} - 1)^{-1}\sum_{1\le k<p\le n'_{ij}}\left(\sum_{m=1}^{z_{ijk}} r_{ijk}t_{ijkm}\sum_{m=1}^{z_{ijp}} r_{ijp}t_{ijpm}\right) \\ &+ \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}^2 a_{ij}^2\sum_{k=1}^{n'_{ij}} r_{ijk}(r_{ijk} - 1)\left(\sum_{m=1}^{z_{ijk}} t_{ijkm}\right)^2.\end{aligned} \tag{9}$$

In (9), the first two summands give the first quantity in (2); summand 3 and 4 give the second quantity in (2); and the last summand gives the third quantity in (2). $\widehat{Var}(\hat{T}_2)$ can be further simplified to

$$\begin{aligned}\widehat{Var}(\hat{T}_2) =& \sum_{i=1}^{l}\sum_{j=1}^{s_i}\sum_{k=1}^{n'_{ij}} a_{ij}d_{ij}r_{ijk}(a_{ij}d_{ij}r_{ijk} - 1)\left(\sum_{m=1}^{z_{ijk}} t_{ijkm}\right)^2 \\ &+ \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}a_{ij}(1 - d_{ij}a_{ij})(n'_{ij} - 1)^{-1}\sum_{1\le k<p\le n'_{ij}}\left(\sum_{m=1}^{z_{ijk}} r_{ijk}t_{ijkm}\sum_{m=1}^{z_{ijp}} r_{ijp}t_{ijpm}\right).\end{aligned} \tag{10}$$

We denote $\widehat{Var}(\hat{T}_2) = \sum_{i=1}^{l}\sum_{j=1}^{s_i} V_{ij}$ where V_{ij} is the contribution to the variance from the segments in design stratum j in land-use stratum i. In the special case that $r_{ijk} = r_{ijp} = r_{ij}$

and $\sum_{m=1}^{z_{ijk}} t_{ijkm} = \sum_{m=1}^{z_{ijp}} t_{ijpm} = c_{ij}$, $1 \leq k < p \leq n'_{ij}$, for some i,j, the V_{ij} is

$$\begin{aligned} V_{ij} &= n'_{ij}a_{ij}d_{ij}r_{ij}(a_{ij}d_{ij}r_{ij} - 1)c_{ij}^2 \\ &\quad + d_{ij}a_{ij}(1 - d_{ij}a_{ij})(n'_{ij} - 1)^{-1}r_{ij}^2\frac{n'_{ij}(n'_{ij} - 1)}{2}c_{ij}^2 \qquad (11) \\ &= \frac{1}{2}d_{ij}a_{ij}r_{ij}n'_{ij}[r_{ij}(d_{ij}a_{ij} + 1) - 2]c_{ij}^2. \end{aligned}$$

$V_{ij} \geq 0$ as expected since the expansion factors $d_{ij}, a_{ij}, r_{ij} \geq 1$. The contribution $V_{ij} = 0$ if $d_{ij} = a_{ij} = r_{ij} = 1$. $\widehat{Var}(\hat{T}_2) = 0$ if $d_{ij} = a_{ij} = r_{ij} = 1$ for all i,j. This is the case of complete census without nonresponse.

To derive the variance of $\hat{\hat{T}}$, let $E(\cdot|JAS)$ and $Var(\cdot|JAS)$ refer, respectively, to the conditional expectation and conditional variance given the outcome of the JAS. We use the formula

$$\begin{aligned} Var(\hat{\hat{T}}) &= Var(\hat{T}_1 + \hat{T}_2) \\ &= E[Var(\hat{T}_1 + \hat{T}_2|JAS)] + Var[E(\hat{T}_1 + \hat{T}_2|JAS)] \qquad (12) \\ &= E[Var(\hat{T}_2|JAS)] + Var[\hat{T}_1 + E(\hat{T}_2|JAS)]. \end{aligned}$$

By the proof of Theorem 1, the first term of (12) is estimated by the second and third quantities in Theorem 1, which are the summands 3, 4, and 5 in (9). By (16) in the Appendix and (6),

$$E(\hat{T}_2|JAS) = \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}\sum_{k=1}^{a_{ij}n'_{ij}}\sum_{m=1}^{z_{ijk}} t_{ijkm}.$$

Here $a_{ij}n'_{ij}$ is the number of segments in the ALUS population in design stratum j within land-use stratum i, since a_{ij} is the expansion factor and n'_{ij} is the number of ALUS segments in (i,j). Together with (3), we have

$$\hat{T}_1 + E(\hat{T}_2|JAS) = \sum_{i=1}^{l}\sum_{j=1}^{s_i} d_{ij}\left(\sum_{k=1}^{n_{ij}}\sum_{m=1}^{x_{ijk}} t_{ijkm} + \sum_{k=1}^{a_{ij}n'_{ij}}\sum_{m=1}^{z_{ijk}} t_{ijkm}\right).$$

By (4),

$$Var[\hat{T}_1 + E(\hat{T}_2|JAS)] = \sum_{i=1}^{l}\sum_{j=1}^{s_i} \frac{1 - 1/d_{ij}}{1 - 1/(n_{ij} + a_{ij}n'_{ij})} \sum_{k=1}^{n_{ij}+a_{ij}n'_{ij}} (c_{ijk} - c_{ij.})^2 \qquad (13)$$

where

$$c_{ijk} = d_{ij}\sum_{m=1}^{x_{ijk}} t_{ijkm},$$

if

$$1 \le k \le n_{ij},$$

$$c_{ijk} = d_{ij} \sum_{m=1}^{z_{ijk}} t_{ijkm},$$

if

$$n_{ij} + 1 \le k \le n_{ij} + a_{ij} n'_{ij},$$

$$c_{ij} = \frac{1}{n_{ij} + a_{ij} n'_{ij}} \sum_{k=1}^{n_{ij}+a_{ij}n'_{ij}} c_{ijk}.$$

Nevertheless, we cannot calculate (13), since only the ALUS sample information, which includes n'_{ij} segments in (i,j), is known. A design-unbiased estimator of (13) is given by

$$\widehat{Var}\,[\hat{T}_1 + E(\hat{T}_2|JAS)]$$

$$= \sum_{i=1}^{l} \sum_{j=1}^{s_i} \frac{1 - 1/d_{ij}}{1 - 1/(n_{ij} + a_{ij} n'_{ij})} \left(\sum_{k=1}^{n_{ij}} \left(c_{ijk} - \widehat{c_{ij.}}\right)^2 + a_{ij} \sum_{p=1}^{n'_{ij}} \left(\widehat{c_{ijp}} - \widehat{c_{ij}}\right)^2 \right), \tag{14}$$

where $\widehat{c_{ijp}} = d_{ij} r_{ijp} \sum_{m=1}^{z_{ijp}} t_{ijpm}$, $1 \le p \le n'_{ij}$, and

$$\widehat{c_{ij}} = \frac{1}{n_{ij} + a_{ij} n'_{ij}} \left(\sum_{k=1}^{n_{ij}} c_{ijk} + a_{ij} \sum_{p=1}^{n'_{ij}} \widehat{c_{ijp}} \right).$$

Hence, we have the design-unbiased estimator of $Var(\hat{\hat{T}})$,

$$\begin{aligned}
\widehat{Var}\,(\hat{\hat{T}}) = & \sum_{i=1}^{l} \sum_{j=1}^{s_i} d_{ij}^2 a_{ij} (a_{ij} - 1) \sum_{k=1}^{n'_{ij}} r_{ijk} \left(\sum_{m=1}^{z_{ijk}} t_{ijkm} \right)^2 \\
& + \sum_{i=1}^{l} \sum_{j=1}^{s_i} d_{ij}^2 a_{ij} (1 - a_{ij}) (n'_{ij} - 1)^{-1} \sum_{1 \le k < p \le n'_{ij}} \left(\sum_{m=1}^{z_{ijk}} r_{ijk} t_{ijkm} \sum_{m=1}^{z_{ijp}} r_{ijp} t_{ijpm} \right) \\
& + \sum_{i=1}^{l} \sum_{j=1}^{s_i} d_{ij}^2 a_{ij}^2 \sum_{k=1}^{n'_{ij}} r_{ijk} (r_{ijk} - 1) \left(\sum_{m=1}^{z_{ijk}} t_{ijkm} \right)^2 \\
& + \sum_{i=1}^{l} \sum_{j=1}^{s_i} \frac{1 - 1/d_{ij}}{1 - 1/(n_{ij} + a_{ij} n'_{ij})} \left(\sum_{k=1}^{n_{ij}} \left(c_{ijk} - \widehat{c_{ij.}}\right)^2 + a_{ij} \sum_{p=1}^{n'_{ij}} \left(\widehat{c_{ijp}} - \widehat{c_{ij.}}\right)^2 \right).
\end{aligned}$$

5. Conclusions

The JAS is the largest annual survey conducted by NASS. Its results are used to develop a number of official estimates. Here, the focus has been on estimating the total number of

US farms. The substantial misclassification of farms and non-farms has led to a biased estimate of the number of farms. The two-phase JAS-ALUS has been suggested as an improvement that would produce a (quasi-)unbiased estimation of farm numbers. The proposed three-phase survey design-based estimator (1) is an extension of the two-phase sampling estimator in Särndal and Swensson (1987), which allows for a general sampling design in each phase. For the JAS-ALUS application considered here, the JAS is the first phase; ALUS is the second phase; and modeling response/nonresponse in the second phase is the final phase. More importantly, a design-unbiased variance estimator for estimator (1) is given in Theorem 1. The estimator (10) of $Var(\hat{T}_2)$ was developed by applying our three-phase variance estimator (2).

Although the focus here has been on estimating the number of US farms, the same ALUS follow-on and adjustment for nonresponse in the second phase allow unbiased estimates of other variables to also be obtained. The experience gained from the FNRP described in Section 2, the change in JAS protocols following the FNRP, and the fact that the FNRP included only 2009 segments could lead to the ALUS results being different from those anticipated here. ALUS has been proposed during a time of declining budgets, and its additional expense is the primary reason NASS has yet to implement ALUS.

Following the FNRP, additional training on JAS prescreening was conducted, and the time that field enumerators were given to complete prescreening was extended from two to four weeks. This resulted in an initial increase in the estimated number of farms, using Equation 1, and then the estimates began to decrease. Some of the decrease may be due to a decline in the number of farms; however, misclassification may again be increasing. Currently, NASS is using modeling approaches to adjust for this misclassification in JAS. It is hoped that ALUS can be conducted at least once, allowing the estimates based on the methods presented here to be compared to the modeled results.

Appendix

Proof of Theorem 1

The proof is an application of the variance formula $Var(X) = Var[E(X|Y)] + E[Var(X|Y)]$. We sketch the necessary steps for readers' convenience.

Recall that $T = \sum_U y_k$ is the population total. From the design, it is easy to see that $\hat{t}_\#$ is unbiased for T. To provide the variance formula for this estimator, first decompose $\hat{t}_\# - T$ as

$$\begin{aligned}\hat{t}_\# - T &= \left(\sum_S \breve{y}_k - \sum_U y_k\right) + \left(\sum_R \breve{y}_k^* - \sum_S \breve{y}_k\right) + \left(\sum_F \breve{y}_k^\# - \sum_R \breve{y}_k^*\right)\\ &= A_S + B_R + C_F.\end{aligned}$$

Now let $E_S(\cdot) = E(\cdot|S)$ and $Var_S(\cdot) = Var(\cdot|S)$ refer, respectively, to the conditional expectation and variance in phase two, given the outcome S of phase one. We also define $E_R(\cdot) = E(\cdot|R)$ and $Var_R(\cdot) = Var(\cdot|R)$ similarly. Then, the variance of the three-phase estimator is

$$Var(\hat{t}_\#) = Var(\hat{t}_\# - T) = Var[E(\hat{t}_\# - T|S)] + E[Var(\hat{t}_\# - T|S)]. \qquad (15)$$

Given the first phase sample, A_S is constant, and the second and third phase estimators are unbiased. Therefore,

$$E(\hat{t}_{\#} - T|S) = E(A_S + B_R + C_F|S) = A_S + 0 + 0 = A_S. \tag{16}$$

Since

$$Var(\hat{t}_{\#} - T|S) = Var_S[E(\hat{t}_{\#} - T|R)] + E_S[Var(\hat{t}_{\#} - T|R)], \tag{17}$$

by a similar argument as in (16), one can easily have

$$Var(\hat{t}_{\#} - T|S) = Var_S(B_R) + E_S[Var(C_F|R)]. \tag{18}$$

From (15), (16), and (18),

$$\begin{aligned} Var(\hat{t}_{\#}) &= Var(A_S) + E\{Var_S(B_R) + E_S[Var(C_F|R)]\} \\ &= Var(A_S) + E[Var_S(B_R)] + E\{E_S[Var_R(C_F)]\}. \end{aligned} \tag{19}$$

Here,

$$Var(A_S) = \sum\sum_U \Delta_{akp}\breve{y}_k\breve{y}_p, \tag{20}$$

$$Var_S(B_R) = \sum\sum_S \Delta_{kp|S}\breve{y}^*_k\breve{y}^*_p, \tag{21}$$

$$Var(C_F|R) = Var_R(C_F) = \sum\sum_R \Delta_{kp|R}\breve{y}^{\#}_k\breve{y}^{\#}_p. \tag{22}$$

However, this variance formula (19) cannot be applied directly. Therefore, a design-unbiased estimator of the variance is needed. For arbitrary constant c_{kp},

$$\begin{aligned} E\left\{E_S\left[E\left(\sum\sum_F c_{kp}/\pi_{kp|R}|R\right)\right]\right\} &= E\left[E_S\left(\sum\sum_R c_{kp}\right)\right] \\ = E\left(\sum\sum_S \pi_{kp|S}c_{kp}\right) &= \sum\sum_U \pi_{akp}\pi_{kp|S}c_{kp} = \sum\sum_U \pi^*_{kp}c_{kp}. \end{aligned} \tag{23}$$

Let $c_{kp} = \breve{\Delta}^*_{kp|S}\breve{y}_k\breve{y}_p$ in the above argument (23). A design-unbiased estimator of the first term of (19) is

$$\sum\sum_F \breve{\Delta}^*_{kp|S}\breve{y}_k\breve{y}_p/\pi_{kp|R}. \tag{24}$$

Let $c_{kp} = \check{\Delta}_{kp|S}\check{y}^*_k\check{y}^*_p$. By using the first two equalities of (23), a design-unbiased estimator of $E[Var_S(B_R)]$ (the second term of (19)) is

$$\sum\sum_F \check{\Delta}_{kp|S}\check{y}^*_k\check{y}^*_p/\pi_{kp|R}. \tag{25}$$

Let $c_{kp} = \Delta_{kp|R}\check{y}^\#_k\check{y}^\#_p$. By using the first equality of (23), a design-unbiased estimator of the first term of $E\{E_S[Var_R(C_F)]\}$ (the third term of (19)) is

$$\sum\sum_F \Delta_{kp|R}\check{y}^\#_k\check{y}^\#_p/\pi_{kp|R}. \tag{26}$$

Putting (24), (25), and (26) together, we have (2), a design-unbiased estimator of (19).

6. References

Abreu, D.A., J.S. McCarthy, and L.A. Colburn. 2010. *Impact of the Screening Procedures of the June Area Survey on the Number of Farms Estimates*. Research and Development Division. RDD Research Report #RDD-10-03. Washington, DC: USDA, National Agricultural Statistics Service.

Benedetti, R., F. Piersimoni, and P. Postiglione. 2015. *Sampling Spatial Units for Agricultural Surveys, Advances in Spatial Science*. Berlin Heidelberg: Springer-Verlag.

Bross, I. 1954. "Misclassification in 2 × 2 Tables." *Biometrics* 10: 478–486. Doi: http://dx.doi.org/10.2307/3001619.

Cotter, J., C. Davies, J. Nealon, and R. Roberts. 2010. "Area Frame Design for Agricultural Surveys." In *Agricultural survey methods*, edited by R. Benedetti, M. Bee, G. Espa, and F. Piersimoni. Chichester: Wiley.

Fecso, R., R.D. Tortora, and F.A. Vogel. 1986. "Sampling Frames for Agriculture in the United States." *Journal official Statistics* 2: 279–292.

Fuller, W.A. 2003. "Estimation for Multiple Phase Samples." In *Analysis of Survey Data*, edited by R.L. Chambers and J. Skinner. 307–322. Chichester: John Wiley and Sons.

Gustafson, P. 2003. *Measurement Error and Misclassifcation in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Ration, FL: CRC Press.

Hidiroglou, M.A. and V. Estevao. 2013. "Dealing with Nonresponse Using Follow Up." In Proceedings of the Joint Statistical Meeting, Survey Research Methods Section, August 6, 2016, Montreal, Canada, 1478–1489.

Jeyaratnam, S., D.C. Bowden, F.A. Graybill, and W.E. Frayer. 1984. "Estimation in Multiphase Designs for Stratification." *Forest Science* 30: 484–491.

Kott, P.S. 1990. *Mathematical formulae for the 1989 Survey Processing System (SPS) summary*. NASS Staff Report, SRB-90-08, National Agricultural Statistics Service, USDA.

Kott, P.S. and F.A. Vogel. 1995. "Multiple-Frame Business Surveys." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, C.B. Nanjamma, A. Christianson, M.J. Colledge, and P.S. Kott. New York: Wiley.

Legg, J.C. and W.A. Fuller. 2009. "Two-Phase Sampling." In *Handbook of statistics 29A, sample surveys: design, methods and applications*, edited by D. Pfeffermann and C.R. Rao. The Netherlands: Elsevier.

Magnussen, S. 2003. "Stepwise Estimators for Three-Phase Sampling of Categorical Variables." *Journal of Applied Statistics* 30: 461–475. Doi: http://dx.doi.org/10.1080/0266476032000053628.

Nusser, S.M. and C.C. House. 2009. "Sampling, Data Collection, and Estimation in Agricultural Surveys." In *Handbook of statistics 29A, sample surveys: design, methods and applications*, edited by D. Pfeffermann and C.R. Rao. The Netherlands: Elsevier.

Särndal, C.-E. and B. Swensson. 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. Doi: http://dx.doi.org/10.2307/1403406.

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

Singh, S. 2003. *Advanced Sampling Theory with Applications: How Michael Selected Amy, 1 & 2*. The Netherlands: Kluwer Academic Publishers.

Stewart, S.L., K.C. Swallen, S.L. Glaser, P.L. Horn-Ross, and D.W. West. 1998. "Adjustment of Cancer Incidence Rates for Ethnic Misclassification." *Biometrics* 54: 774–781. Doi: http://dx.doi.org/10.2307/3109783.

Swartz, T., Y. Haitovsky, A. Vexler, and T. Yang. 2004. "Bayesian Identifiability and Misclassification in Multinomial Data." *Canadian Journal of Statistics* 32: 285–302. Doi: http://dx.doi.org/10.2307/3315930.

Tenenbein, A. 1970. "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications." *Journal of the American Statistical Association* 65: 1350–1361.

Tenenbein, A. 1972. "A Double Sampling Scheme for Estimating from Multinomial Data with Application to Sampling Inspection." *Technometrics* 14: 187–202.

Thall, P.F., D. Jacoby, and S.O. Zimmerman. 1996. "Estimating Genomic Category Probabilities from Fluorescent in Situ Hybridization Counts with Misclassification." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 4: 431–436. Doi: http://dx.doi.org/10.2307/2986066.

Vogel, F.A. 1995. "The Evolution and Development of Agricultural Statistics at the United States Department of Agriculture." *Journal of Official Statistics* 11: 161–180.

Received May 2014
Revised September 2016
Accepted October 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 223–247, http://dx.doi.org/10.1515/JOS-2017-0012

Changing Industrial Classification to SIC (2007) at the UK Office for National Statistics

Paul A. Smith[1] *and Gareth G. James*[2]

As part of the changes to industrial classifications following the United Nations' revision to the International Standard Industrial Classification, ISIC Rev. 4, the UK moved to its version of a new classification between 2007 and 2011. We describe the processes involved in changing an industrial classification, including model-based adjustment methods and changes to survey designs and operations. We discuss the quality of the approaches used for different time periods in the same series, and the ways in which consistent time series are produced for users of economic statistics. We provide some general evaluation of the changeover, and guidance on the best approaches to follow when updating classifications.

Key words: Conversion matrix; overlap control; quality assessment; classification change.

1. Introduction

Changes in standard classifications, such as that used for economic activity, usually occur at reasonably regular, though infrequent intervals. They are required to ensure classifications remain up-to-date and relevant, but when they occur, they consume large amounts of resources and can create discontinuities in time series. Therefore, they need to be carefully managed (MacDonald 1995). A change in the classification system has effects in many of the stages of the statistical production process (as codified by, for example, the Generic Statistical Business Process Model, UNECE 2013). In this article, we consider the most recent classification change implemented by the UK Office for National Statistics (ONS) as a model to demonstrate some techniques and derive some general guidance.

Industrial classifications are hierarchical. They are harmonised internationally down to a particular level of detail, including:

- an international framework: the International Standard Industrial Classification (ISIC, the current version is Rev. 4), produced by the United Nations,
- a regional implementation: in the European Union (EU), Eurostat (the statistical office of the EU) currently uses NACE (Nomenclature statistique des Activités

[1] Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk
[2] Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK. Email: gareth.james@ons.gov.uk

Acknowledgments: Many people in ONS were involved in implementing SIC (2007), and it would not have been possible to write this article without their efforts. We are very grateful to the Associate Editor and three anonymous referees for their comments, which have led to a much improved article. Some of the final considerations derive from a visit to ONS of the late David Binder. The views expressed in this article are those of the authors and do not necessarily represent those of the ONS.

économiques dans la Communauté Européenne) Rev. 2 (a four-digit hierarchical classification), which is consistent with ISIC Rev. 4 at the two-digit level, and

- national implementations, such as the UK's Standard Industrial Classification (SIC). Within the EU, Member States have the option to enhance NACE for their own purposes, as long as the national classification corresponds with the harmonised classification to at least a given level of detail (usually four-digit industry for NACE). National implementations are also available in an appropriate native language with support for dictionaries and look-ups to make the classification usable (see, for example Beekman 1992).

Harmonised classifications within the UK have been an important way to make statistics from different government departments comparable since the first UK SIC in 1947, and (with a few exceptions) have been used consistently throughout the UK's statistical system. The UK's current classification is known as SIC (2007) (the '2007' denoting the year of publication). The classification forms the basis of statistical outputs, and provides a framework for combining estimates from different surveys in derived statistics such as the National Accounts. The classification codes also provide one dimension for stratification in business survey designs.

Changes in the classification in the UK have occurred about every 10–15 years (see Smith and Penneck 2009 for an overview of industrial classifications used in UK statistics since 1907). The most recent large changes in the UK's classification were from SIC (80) to SIC (92) – the implementation occurring in the mid-1990s – and from SIC (2003) to SIC (2007). (The change from SIC (92) to SIC (2003) was not large, and involved only minor changes at the most detailed level.)

The most recent change proved particularly challenging, as the new classification (SIC (2007)) contained more detail than its predecessors, and included a number of industry restructures. Examples include the separate identification of Repair and Maintenance, a new section on Water Supply and related activities, a new section on Information and Communication, the move of Retail Sale of Automotive Fuel from Motor Trades to Retail, and the move of some publishing and printing activities from Manufacturing to Services. Table 1 shows the changing numbers of categories at different levels of the classification. The detail of the five-digit level was agreed in the UK following a series of user consultations called 'Operation 2007' (Hughes 2008) run by a cross-government group (including the ONS). Subdivisions of the four-digit codes were agreed only where there was user demand, and where it would be practical and meaningful to distinguish between the proposed subcategories. It can be seen that fewer

Table 1. Comparison of detail (number of categories) between SIC (2003) and SIC (2007).

	SIC (2003)	SIC (2007)
Section (letter)	17	21
Division (two-digit)	62	88
Group (three-digit)	225	272
Class (four-digit)	514	615
Subclass (five-digit)	699	728

disaggregations of four-digit codes to five-digit codes were accepted under SIC (2007) than under SIC (2003), probably reflecting the greater detail already present in the new four-digit codes, but also reflecting a reluctance to create five-digit codes unless there was a strong case that they were both necessary and practicable. That cross-government group also coordinated the implementation of SIC (2007) across the Government Statistical Service in the UK, following the timetable set by Eurostat for all EU Member States.

The change in industrial classification was also aligned with a change to the European product classification to CPA 2008 (Classification of Products by Activity), and this particularly affected the National Accounts, which use both product and industry estimates in producing balanced measures of the national economy. The scale of these changes called for review of most aspects of the methodology used in business surveys in the ONS to allow the production of estimates on the new classification.

Planning for the project and work to get agreement on the classification breakdowns using the fifth digits started in the early 2000s, so that the new classification, with its index and accompanying notes for coding could be published in January 2007. Thereafter, the work to implement the new classification in surveys and outputs really began; the timing of changes is outlined in Figure 1. All organisations on the Inter-Departmental Business Register (IDBR) – the ONS's sampling frame for most business surveys – were dual-coded during 2007, so that by January 2008 all had a SIC (2003) and a SIC (2007) code. The annual, structural surveys collecting information about activity in 2008 were the first to be sampled on the new classification, with selections taking place from the IDBR towards the end of 2008. The first outputs on the new classification, from short-period (monthly and quarterly) surveys, were reported to Eurostat from the start of 2009. The old classification was still used for sample selection in monthly and quarterly surveys until the start of 2010, however, and inputs to National Accounts were retained on the SIC (2003) basis until the publication of the Blue Book (the UK's main National Accounts publication) in September 2011, after which time all reporting of economic output from the ONS has used the new classification.

Timetable of implementation of SIC (2007) at ONS:

Pre-2007	Agreement of fifth digits in SIC codes
2007	Addition of SIC (2007) codes to the IDBR, available from January 2008
Late 2008	Selection of first annual survey samples using SIC (2007) stratification
Jan/Q1 2009	Delivery of short-period statistics to Eurostat using SIC (2007)
Jan/Q1 2010	Short-period surveys stratified by SIC (2007); outputs still required in National Accounts on SIC (2003)
Sep 2011	National Accounts move to SIC (2007) with publication of the Blue Book

The prolonged timetable has created some interesting challenges, particularly around the publication of two simultaneous sets of outputs. In some cases, publication was retrospective, with backcasting (see Subsection 4.1) used to estimate how historical outputs would have looked on the new classification. In other cases, publication was (almost) concurrent, with more recent periods being reported on both classifications.

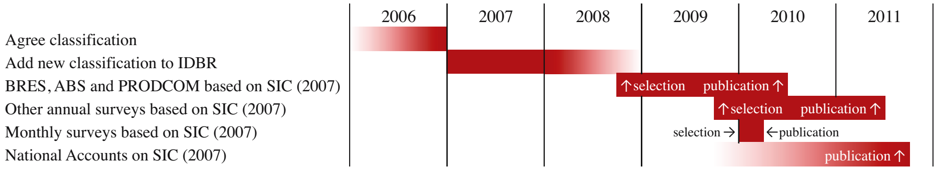

2006
2007
2008
2009
2010
2011
Agree classification
Add new classification to IDBR
BRES, ABS and PRODCOM based on SIC (2007)
Other annual surveys based on SIC (2007)
Monthly surveys based on SIC (2007)
National Accounts on SIC (2007)
↑selection
publication ↑
↑selection
publication ↑
selection →
←publication
publication ↑

Fig. 1. Diagrammatic representation of transition to SIC (2007) in ONS. Only the first instance of each survey with design and sample selection updated to the new classification is represented. Fading shows continuing work leading up to or following an implementation. IDBR = Inter-Departmental Business Register, BRES = Business Register and Employment Survey, ABS = Annual Business Survey and PRODCOM = Products of the European Community survey.

In this article we review the implementation of SIC (2007), and the developments and changes required in the methodology of surveys, particularly with respect to consistency of estimates across time. Section 2 covers the changes to the IDBR, and Section 3 describes changes to the sample designs of business surveys largely (but not wholly) to accommodate the new classification. Section 4 covers different methods of estimating using the new classification and includes comments on some of the effects on the National Accounts. Section 5 discusses the impacts on quality of estimates using the standard European dimensions of output quality (Eurostat 2015a), and the article concludes with some general comments about the implementation, and includes some lessons learned and suggestions for future implementations.

2. The Business Register

The IDBR is maintained by the ONS, and is the main sampling frame used for official business surveys in the UK. It contains information on about 2.7 million sites (known as local units (LUs) – shops, factories, offices, and so on), which are grouped into about 2.1 million reporting units (RUs), which are the sampling units used in most ONS business surveys (in this article we will use ‘business’ synonymously with RU for this reason). One or more RUs form an enterprise, which is the smallest autonomous business structure in the IDBR. In the majority of cases, an enterprise has one RU and one LU, and the three units are indistinguishable; even large enterprises normally consist of one RU, although a relatively small number are split into multiple RUs to facilitate data collection (for more information see Smith et al. 2003). The distribution of enterprise sizes is very skewed, and large proportions of activity and employment take place in enterprises composed of multiple LUs.

The IDBR contains a range of variables for each type of unit, the values of which are obtained from various data sources, with a hierarchy specifying which sources are preferred. One of the most important variables ascribed to the unit is the SIC code, which represents its principal economic activity, and rules exist to determine the SIC code for enterprises and RUs based on the codes of the LUs (for a simple summary see Smith 2013, Subsection 5.2.1.2 and Box 5.1). A stable SIC code is stored for sampling purposes, and this code is used throughout a calendar year for monthly surveys, for example. Then, reclassification effects can be saved up and handled together at a fixed point, usually the year-end. A ‘live’ classification variable stores the most up-to-date SIC code. Updating the IDBR includes copying the current live codes to the stable codes once a year. Additional variables were added to the IDBR in preparation for the change to SIC (2007), so that (stable and live) codes for both SIC (2003) and SIC (2007) could be stored for each unit.

One source of information about new businesses in the IDBR is Value Added Tax (VAT) records from HM Revenue and Customs (HMRC). When a business is formed, it must describe its activity when it registers for VAT, and these descriptions are coded by HMRC using automated coding software (called ACTR – Automatic Coding by Text Recognition). The software and coding dictionary are harmonised across government departments, so that there is consistency in the way in which classification codes are assigned. Both the original descriptions and the assigned codes are passed to the IDBR and stored, so the descriptions are available for recoding to deal with future classification updates. For businesses already in the IDBR, another source of information is the ONS's annual Business Register and Employment Survey (BRES), which updates register information and collects employment variables; the register-updating part has previously been administered separately. Not all businesses are surveyed in BRES, but all large businesses are included every year, and medium-sized ones once every third or fourth year. The smallest businesses are included in BRES only with small sampling fractions, and primarily for estimation of employment, so classification information for such businesses in the IDBR is derived mainly from the administrative data sources. Businesses selected for BRES are asked to supply a written description of the principal economic activity at each site (LU), and these are then coded to the SIC using the ACTR automated coder (Williams 2006). (Previous descriptions of economic activity are prefilled on the questionnaire for existing sites to make the respondents' task less onerous – only changes need be notified.)

To facilitate coding to the new classification, a knowledge base for the new classification was developed by constructing a list of full descriptions of economic activity. These full descriptions were edited by a classification specialist for consistency, and used as the basis for the automated coder to code business descriptions to SIC (2007). This enabled dual coding of most businesses in the IDBR (1.7 million of the 2.7 million local units, for which original business descriptions were available and usable). For the remaining units – those without a description, or where the description was insufficient – SIC (2007) codes were assigned probabilistically, based on the distributions observed in those units that had been dual-coded. Of course, any SIC (2003) codes that mapped entirely to just one SIC (2007) did not need to go through this process. Based on reclassifications of businesses, 462 of the five-digit SIC (2003) codes matched exactly to a five-digit SIC (2007) code; 229 matched to multiple codes, and the eight remaining codes did not have any businesses classified to them. Of SIC (2007) five-digit codes, 503 matched one code, 205 matched multiple codes, and 20 had no businesses. Some checking and manual intervention was necessary to avoid problems arising from the descriptions (and to feed these back as improvements to the automated coder), but the IDBR was fully dual-coded before January 2008. However, feedback on codes from businesses meant that further cleaning continued, and some further changes in the register were quite evident, particularly in the early months of 2008. There were some noticeable changes in the proportion of SIC (2003) industries contributing to SIC (2007) industries early in the year (Figure 2), but by mid-2008 these proportions had largely stabilised. The same sort of pattern is shown for local units in aggregate in Scottish Government (2012).

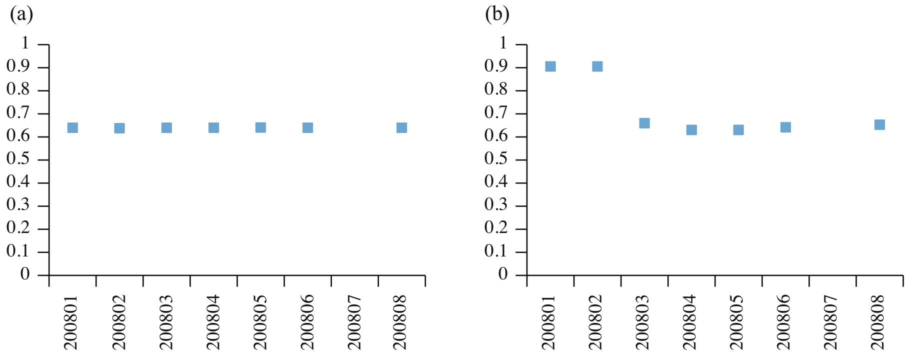


Fig. 2. Proportions of undeflated turnover from an original SIC (2003)-based classification code contributing to a new SIC (2007) code in the early months of 2008, based on dual-coded microdata from the IDBR. Here, the classifications are aggregates of industries used for UK Supply and Use Tables (SUT); the identities of the classifications are withheld for disclosure control reasons. Note that there is no data point for July 2008. (a) A typical example with little variation; (b) some industries had noticeable changes in early 2008 before stabilising.

The new SIC (2007) codes were used for sample selection purposes (as stratification variables) for the first time in the autumn of 2008, when samples for three annual surveys were selected: BRES, the Annual Business Survey (ABS) and PRODCOM (PRODucts of the European COMmunity; the UK implementation is headed 'UK manufacturers sales by product').

3. Survey Redesign

The changes in the SIC meant that the business surveys needed to be redesigned. Almost all of the ONS's business surveys are stratified by a cross-classification of industry (SIC) and size, with size usually defined as employment-based sizebands. Therefore changing the surveys' sample designs to the new codes, together with reallocating the (fixed) sample size was the minimum requirement, but in many cases the opportunity was taken to introduce other improvements at the same time.

An example of wider survey redesign was the introduction of the Monthly Business Survey (MBS) at the ONS, of which an overview is given here, and a more detailed account can be found in Taylor et al. (2011). The MBS was introduced as a replacement for a number of short-period surveys, each covering different parts of the economy (namely production, and parts of the services sector). Improvements to the surveys had been planned for some time, but the need to introduce the new classification made this work a higher priority. (Here, we see one of the incidental effects of changing a classification, that it often prompts implementation of further changes.) The new design covered much more of the economy as a single survey (construction and agriculture were not included initially, although construction surveys have since been added to the MBS family). A combined survey has advantages when businesses change classification, particularly for moves between the production and services sectors. Such a difference does not now result in a change to the title of the questionnaires, which previously might have seen the same business receive a 'Manufacturing' questionnaire on one survey and a 'Services' questionnaire on another. Thus, a combined approach removes the source of some queries from businesses, and it also facilitates easier transfer of data between processing systems. In addition to changes in stratification and sample allocation (detailed further, below), improvements were made to the coverage, questionnaire design, and editing and imputation methodology, with the aim of introducing greater standardisation and therefore more coherence. This approach followed general practice, which often sees several updates to a survey implemented simultaneously. This approach can be convenient for users, as they need to deal with only one discontinuity. On the other hand, some users want to know how much of any discontinuity is attributable to each development, and these may not be separately estimable when changes are introduced together.

The first task in designing the new MBS was to decide on appropriate SIC (2007) codes to define the industry strata. This was not a straightforward task, as different levels of the hierarchy and different groupings of codes can be used, as was the case across its predecessor surveys. Consultations were carried out with the main customer of the surveys, National Accounts, so that the strata would be consistent with their requirements, and based around the industrial groupings used in the supply and use tables (SUTs). In 2009, those were as yet undefined under SIC (2007), which required some work to be conducted quickly to establish the likely groupings to be used in National Accounts from 2011 onwards when SIC (2007) would be introduced. The results of this exercise were used as the basis for the MBS stratification. (Further details of the 114 input-output industry groups, which would be defined for use in the SUTs under the new classification can be found in Drew and Dunn 2011.) Many of the SIC (2007) SUT industry groups were formed at the two- or three-digit SIC code level, and meant that SIC (2007) stratification of the MBS could be carried out at a more aggregated level than in the previous designs based on SIC (2003) – only around 150 industries instead of 300 (Taylor et al. 2011). A similar approach, with the stratification detail being driven by the principal output requirements, was taken when redesigning other surveys. Use of a broader stratification gave more flexibility to produce an efficient sample allocation without increasing the fixed sample size. However, it should be noted that this also increased some estimator variances (particularly for detailed industry classifications that no longer formed strata), as domain estimation was needed to get the full SIC (2007) industry detail, and also for the SIC (2003)-based estimates that were still required by National Accounts until their move to SIC (2007).

With industry stratum boundaries defined, a review was conducted as to whether the sizebands, which separate the industry strata into sampling strata, should be retained or updated. Some of this work was necessary, as various combinations of sizebands were used across the predecessor surveys, but the opportunity was taken to conduct a wider review. In many cases, updated sizeband boundaries were introduced to improve the design, even in industries which mapped one-to-one from SIC (2003) to SIC (2007). However, the number of sizebands was kept the same as changing this would have necessitated changes to the processing systems that could not have been easily accommodated at the same time.

Samples for business surveys at the ONS are usually allocated using principles of Neyman allocation (Cochran 1977, 98-99), although adapted to account for various practical considerations and precision requirements on some lower-level outputs. The process requires information on the number of businesses in each sampling stratum (obtained from the IDBR) and a measure of the variability in the responses, given the estimator used within each sampling stratum. The standard approach for estimating the population variance is to use information from survey responses in a particular stratum in a previous period:

$$\hat{v}_h = \frac{1}{n_h - 1} \sum_{i \in h} (y_i - \hat{y}_i)^2 \tag{1}$$

where y_i is the previous period response for business i, n_h is the number of responses in stratum h in that period, and $\hat{y}_i$ is an estimate of the mean for unit i appropriate for the estimation model in use in the survey (e.g., the overall mean $\bar{y}$ for expansion estimation, or βx_i for ratio estimation with auxiliary variable x). The same basic approach applies when h represents old strata or when it represents the new strata, though with the change in classifications, past responses for use in the new SIC (2007) strata sometimes come from businesses in several different SIC (2003)-based strata, and these needed to be weighted appropriately. Standard options for a weighted variance include:

$$\hat{v}_g = \frac{1}{\left[\sum_{i \in g} d_i \right] - 1} \sum_{i \in g} d_i (y_i - \tilde{y})^2 \tag{2}$$

and

$$\hat{v}_g = \frac{1}{(n_g - 1)\bar{d}} \sum_{i \in g} d_i (y_i - \tilde{y})^2 \tag{3}$$

where d_i is the inverse of the sampling probability for unit i, $\bar{d} = \frac{1}{n_g} \sum_{i \in g} d_i$, g indexes the new (SIC (2007) × sizeband) strata, and $\tilde{y} = \sum_{i \in g} d_i \hat{y}_i / \sum_{i \in g} d_i$ is a weighted estimate of the mean of the y appropriate for the estimation model used in the survey. Only (3) collapses to (1) when $d_i = k \, \forall \, i$. Even what may appear to be a relatively simple case, for example of two old (SIC (2003)) industry strata being mapped entirely into one new (SIC (2007)) industry stratum may require this treatment if different employment sizebands or sampling fractions were used in the old industries or between the old and new industries.

A similar pooling, but based on an assumption that variances within a group of strata are equal, is presented in Van den Brakel (2010, Equation (30)); this might be a better approach where sample sizes in group g are small, which may make variance estimates unstable. In ONS business surveys, groups g based on SIC (2007) were of sufficient size to use the approach in (2). For monthly surveys, variances were calculated separately for each month over a year, and the average was used in allocation (Taylor et al. 2011). In some cases, a new SIC (2007) sampling stratum had no previous information to use, since the SIC (2003) equivalent industry was out of scope. In these cases, an allocation was made based on variances derived from IDBR turnover data; the resulting allocation was checked and adjusted in cases where it looked implausible. Thus, the final allocations were based on a number of procedures and assumptions, and were reviewed when real data had been collected, resulting in some minor adjustments.

The ONS uses a Permanent Random Number (PRN) system to coordinate its sampling from the IDBR (Ohlsson 1995, Smith et al. 2003). Each business in the IDBR is allocated a random number in the range (0,1] which it retains thereafter. When a sample is drawn, the businesses within a stratum are ordered by PRN (along the 'PRN line') and the sample consists of a consecutive group of businesses in this ordering, with rotation achieved by adding businesses with the next largest PRNs and dropping businesses with the smallest PRNs. The PRN line is considered to 'wrap around' from 1 back to 0.

The sample selection procedures employed specify how long a business should expect to be retained in a survey, how many periods should elapse before it is selected in the same survey again, and for the smallest businesses, a maximum of one survey at any given time. The change in industry stratification using the new classifications meant that currently sampled businesses from several different strata in the old design formed the initial sample in a new stratum. If there was complete freedom to choose which businesses to include in the sample, then it would be possible to maximise the overlap appropriately using optimisation methods (see Mach et al. 2006, Johnson et al. 2012 and Schiopu-Kratina et al. 2014 for some examples of suitable methods); similarly, if particular surveys were restricted to parts of the PRN line (as in the Swedish system SAMU, see Lindblom 2003), the problem would be simplified. But in the ONS there are no restrictions to parts of the line for particular surveys, and the sample selection is managed by the PRN rotation system in the IDBR, so it was important to retain the existing system.

By using the PRNs for businesses in each stratum in the design using the new classification, we automatically achieve the required selection probabilities (each PRN sample is a simple random sample of the stratum population, using the properties in Ohlsson 1995). We have a choice of which PRN to use as the starting value for the new rotation, and the best start is the one which maximises the overlap, but also obeys rules on time in survey and survey holidays. The best starting PRN was found by evaluating a penalty function for each possible PRN start (because the PRNs in any particular register are fixed, there are N_h possible starts in stratum h). The PRN start corresponding with the minimum value of this penalty was chosen as the start for the sample on the new classification. The penalty is

$$\sum_{t=1}^{r}\sum_{i\in h}(L_{it}+B_{it}+D_{it})$$

where r is the rotation period for the survey, t sums over a full rotation of the survey, the three components of the penalty function were defined as:

$$L_{it}=\begin{cases} r-u_{it} & \text{if } \ i\in S_{ht} \ \text{ and } \ 0<u_{it}<r \\ 0 & \text{otherwise} \end{cases}$$

$$B_{it}=\begin{cases} r & \text{if } \ i\in S_{ht} \ \text{ and } \ s_{it}>r \\ 0 & \text{otherwise} \end{cases} \qquad ,$$

$$D_{it}=\begin{cases} r-s_{it} & \text{if } \ i\notin S_{ht} \ \text{ and } \ 0<s_{it}<r \\ 0 & \text{otherwise} \end{cases}$$

S_{ht} is the set of sample units defined by the PRNs at period t, s_{it} is the number of consecutive periods before period t for which unit i was selected in this survey ($= 0$ if it was not selected in period $t-1$), and u_{it} is the number of consecutive periods before period t for which unit i was not selected in this survey ($= 0$ if it was selected in period $t-1$). L penalises units which have an insufficient holiday between periods in the survey (taking a holiday of length at least r as the target), B penalises units which spend more than r consecutive periods in the sample (and more heavily than the expected $s_{it} - r$ in order to penalise breaches to published expectations for the length of time in particular surveys), and D penalises units which are dropped from the survey early, that is, before they have spent r periods in the sample. In practice, D tends to dominate the penalty function, but this is acceptable because the minimum of $\sum\sum D_{it}$ occurs where most previously sampled units (which have not reached r periods in sample) are being reincluded in the sample.

This algorithm was applied stratum-by-stratum and survey-by-survey, so it did not coordinate samples across surveys. Nevertheless, it provided a way to maintain a large overlap within the existing system, although the resulting overlap was often smaller than would have been achieved had the design remained unchanged. This resulted in larger variances in estimates of period to period change across the transition. A number of further initiatives were introduced to try to maximise response and realise the largest possible overlap, including giving advance notice to businesses of any changes in the questionnaire, and increasing resources for response-chasing in these periods.

4. Creation of New Outputs

For the most part, survey outputs at ONS have been produced on the same classification as the stratification. However, with a change in classification, outputs have been required on both classifications. Careful consideration has been needed to ensure that this is achieved sensibly.

Two broad approaches have been used: backcasting and dual-running. Backcasting is a *macro*-method, because it uses only aggregate statistics as the basis for a model, which is used to produce estimates on a different classification. Dual-running is a *micro*-method, because it uses the microdata, usually dual-coded, as the basis for estimation on both classifications.

The principal use of backcasting has been in the production of historical time series on SIC (2007), formerly available only under SIC (2003). For many series, backcasting has been used to present a reclassified series starting in the 1990s. This corresponds with survey redesigns in the UK (for example, the Annual Business Inquiry, now the Annual Business Survey), which was introduced in 1997 (Smith et al. 2003), so many annual series start from this date. This provides a sufficiently long run of reclassified estimates for many uses. Econometric models, however, may require very long spans of data, so in some cases longer series have been developed, though the method of backcasting should be taken into account when developing and using such models.

Dual-running, the micro-method, may be thought of as the production of two sets of estimates (on SIC (2003) and SIC (2007)) with respect to the same reference period, at approximately the same time. It has been used for current or recent periods where dual-coded register data are available. This method was used both to produce reclassified estimates for SIC (2003)-stratified surveys (for example, the short-period surveys in 2009 were sampled using SIC (2003), but SIC (2007) outputs were required by Eurostat), and for SIC (2007)-stratified surveys (for example the short-period surveys in 2010, which were stratified and processed under SIC (2007), but where SIC (2003) outputs were also required for National Accounts until the publication of the 2011 Blue Book). See Van den Brakel (2010) for a discussion of stratification and estimation methods for these two micro-method scenarios, and Subsection 4.2 below.

The two methods (the macro-method of backcasting, and the micro-method of dual-running) were used at the ONS to produce different parts of the same time series. Backcasting is used for estimates and outputs referring to periods further in the past, and dual-running is used for more recent or current periods. Therefore, the time series produced with SIC (2007) have distinct sections to them:

- the oldest periods estimated directly using the old classification and adjusted by backcasting,
- more recent periods with data produced by dual-running from surveys designed on SIC (2003), and
- the most recent periods, based on surveys redesigned on SIC (2007).

These sections are linked together where appropriate to avoid discontinuities (see De la Fuente Moreno 2014 for information on linking and splicing methods). Part of such a series for the Retail Sales Index is shown in Figure 3. We now examine more closely the macro- and micro-methods used.

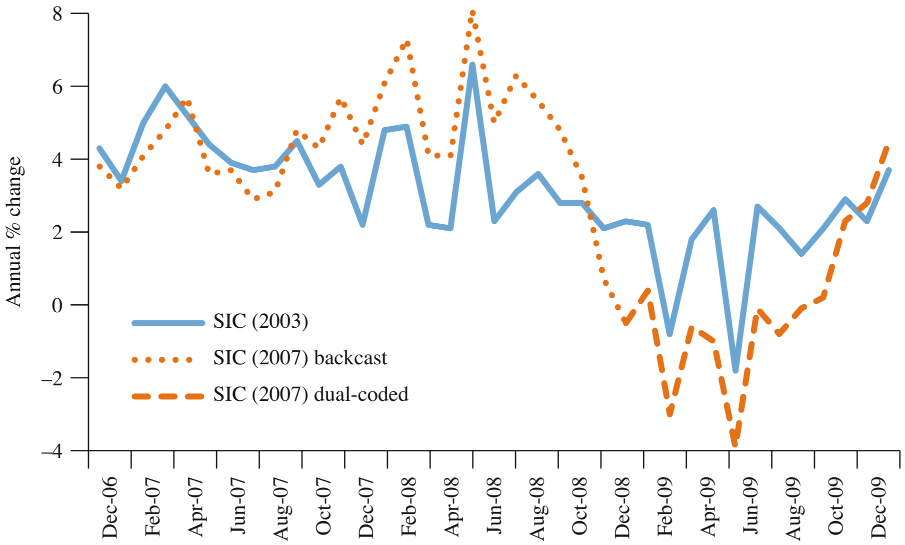


Fig. 3. An example reclassified series composed of linked parts – yearly movements in Retail Sales value (i.e., not deflated), seasonally adjusted, December 2006-December 2009 (redrawn from McLaren 2010, Figure 5). The definition of retailing is expanded in SIC (2007) to include automotive fuels (previously collected in another survey, as it was part of Motor Trades), and the pattern in the fuel series accounts for the very different patterns of movement in the last six months of 2008. The backcast portion of the SIC (2007) series contains an estimate for automotive fuels, derived from the survey which previously included it, converted to the new classification. The whole period shown is based on a stratification using SIC (2003) – the change to SIC (2007) stratification came in January 2010.

4.1. Backcasting

Backcasting, the process used to derive most historical estimates using SIC (2007) is a macro-method, which is basically dependent on fitting a suitable model to the available aggregate data. The model is assumed to hold over a long time period. Such methods are applied only to the industry-level (that is, aggregate) estimates, and do not involve the microdata in any direct way other than (sometimes) to fit the model. Conversion matrices (or concordances) are the most frequently used method for codifying these models for reclassifications; such methods are commonly used by national statistical institutes (e.g., Bayard and Klimek 2004, Russell et al. 2004, Yuskavage 2007) when handling a change in classifications. Indeed, this approach was used by the ONS at the last major classification change, to SIC (92) in the mid-1990s. The major advantage of such methods lies in their efficiency: there is no need to retrieve and recode historical microdata (which may no longer exist, or may be impossible to recode), and the quality of the resulting estimates is usually acceptable. However, they do rely on assumptions about the stability and appropriateness of the model, particularly where they are used for conversion over a long period. Where stability cannot be assumed, it may be possible to use historical data to develop a sequence of time-varying conversion matrices (Yuskavage 2007). In the change in the ONS, conversion matrices from 2008 were used consistently for years before 2008, using the assumption of stability. This was largely because of a lack of resources for recoding earlier versions of the IDBR.

Conversion matrices do not deal with classification changes which incorporate or remove whole areas of activity (structural zeroes), as there is then no information to which to apply conversion factors to obtain an estimate. A further disadvantage of the application of fixed conversion matrices is that historical reclassifications of individual large businesses will be missed (since the same, recent proportions are used to split and reaggregate estimates for all historical periods), which may lead to unrealistic results. This emphasises the need for manual examination of individual cases and converted series to identify when the assumptions about model stability have broken down. In these cases, a suitable adjustment can be made to the converted series.

The foundation of conversion matrices is a dual-coded business register or census, from which cross-tabulations show the proportions of businesses that map from a SIC (2003) code (or group of codes) to a SIC (2007) code (or group of codes), and vice versa. In the ONS, such matrices were produced from extracts of the IDBR, based on the most detailed (five-digit) level of classification, with proportions reflecting business sizes (in terms of turnover and employment). These matrices were then used to apportion SIC (2003) aggregate estimates to SIC (2007) codes, which were then resummed to derive SIC (2007) aggregate estimates for historical periods.

Conversion matrix options: The choice of which conversion matrices to use was decided on a case-by-case basis for the various ONS outputs, with compromises being inevitable; consistency was kept wherever possible:

- Timing: For most of the short-period surveys, conversion matrices were used to create SIC (2007)-based estimates for reference periods up to the end of 2008, with micro-methods (see below) being used from the start of 2009. In order to estimate and link out any discontinuities, conversion matrix-based estimates were also produced for 2009, and the matrices themselves were produced with IDBR extracts taken at that time. Matrices were somewhat volatile through early-2008, caused by the new SIC (2007) codes settling (Figure 2), but became much more stable later in the year. Early conversions were therefore of lower quality, but were redone when the stabilised matrices were available. Further register updates were applied from the 2008 structural surveys, further increasing quality and stability.

- Base variable: Most ONS business survey estimates are based on ratio estimation (Smith et al. 2003), and outputs that use register employment as an auxiliary variable in estimation (generally those with labour-market related outputs) used conversion matrices based on employment size, whereas those that use turnover as an auxiliary variable (generally related to business output) used conversion matrices based on turnover size. Note that this approach can reduce the consistency in estimates of variables derived from both types of data, such as productivity. The inconsistency was accepted in ONS because of the benefits of using conversion matrices appropriate to the variables, but other decisions would be possible. Matrices based on counts of businesses were generally not used, but would be appropriate if trying to backcast series of estimates of numbers of businesses.
- Statistical units: The matrices could be calculated separately using different types of units as their basis; the usual survey practice guided which one to use. Most ONS business surveys are based on responses from RUs (see Section 2), and these used conversion matrices calculated from RU information. Labour market estimates are based on additional information from LUs, and these series used conversion matrices calculated from LU information.
- Level of conversion: Although the matrices are usually calculated at the five-digit classification level, alternative levels of aggregation can be used. Naturally, the level of the series (estimates) to be converted will largely inform the decision, but coding error would be more likely at lower levels (the detail may be incorrect within the appropriate broad industry group). Examination of the conversions at detailed level in the ONS suggested that the timing of extracts (Figure 2) from the IDBR had a larger effect on coding error than the choice of the level at which the conversion matrix was applied.
- In all cases, each derived series was carefully checked, and this allowed manual interventions to be made where appropriate.

The settling of the SIC (2007) codes on the IDBR also had some notable effects on the conversion matrices. Many of the initial codes were imputed from old SIC (2003) codes, as there was no business description available, and these cases could only follow the table of ‘official’ correspondences. Since then, many of the imputed codes have been replaced with directly-coded SIC (2007) codes, which has seen the number of nonzero correspondences in the cross-classified tables increase (sometimes codes derived from business descriptions fall outside the ‘official’ correspondences). As an example, in the matrices for conversion from four-digit to three-digit codes (with a few exceptions), the number of nonempty cells approximately doubled from January to May 2008. (Of course, some of the changes in SIC (2007) codes may reflect actual changes in activity after the SIC (2003) code was assigned, or errors in the original assignment of SIC (2003) codes, as well as errors in the SIC (2007) code.)

There are also choices to be made regarding the stage of processing at which conversion should occur. ONS's Index of Production (IoP), for example, is largely based around estimates of turnover, which are then deflated using price indices before being seasonally adjusted. Thus, for the IoP, there were three reasonable choices regarding conversion: (1) conversion of all inputs (turnover estimates and price indices); (2) conversion after deflation; or (3) conversion after seasonal adjustment. Investigations revealed that the differences between deflating before conversion and conversion followed by deflation seem to be generally small. The largest effects were present where there were changes over time in the weights of components with different deflators. Consultation with experts suggests that converting after deflation might make more sense economically. However, there are practical considerations, such as coordination of the timing of the reclassification of the deflator (a price index) and the data to which it is applied, which may be more important than the choice of stage to convert. One reviewer suggested that another reason for converting the values before deflation is that the conversion matrix was developed using these values. Relationships among the variables in these matrices are more likely to remain stable than the relationships between variables before and after deflation. The need for a history for deflators on the new classification, too, suggested that conversion of deflators first would give the best coherence between the various outputs of the reclassification. A considerable amount of checking was required to ensure that the historical price indices were coherent and credible.

Applying the conversion matrix to seasonally adjusted series would give linear combinations of the seasonally adjusted input series, and these outputs should, in general, appear seasonally adjusted. This seems to work in practice, although there remains a concern that a linear combination of the errors in the seasonal adjustment decomposition may show some residual seasonality. In line with the principles in the European Statistical System Guidelines on Seasonal Adjustment (Eurostat 2015b Sec. 3.4) we therefore prefer direct seasonal adjustment of the new series. This allows for different seasonal patterns in the new component series. Therefore we prefer conversion of all inputs (including the deflators) to the new classification, and for seasonal adjustment to take place after deflation.

Conversion matrices were made available for users to do their own conversions. As well as the turnover- and employment-based versions, a conversion matrix using the number of units was produced and published (ONS 2010); each published matrix had entries rounded, where appropriate, to reduce the risk of disclosure.

A further use of conversion matrices, noted here for completeness, can lie in the dual-coding of microdata. The ONS found a demand for this on historical social survey datasets, which are made available for use by approved researchers. The actual application was on Occupational Classification codes, but the principle for Industrial Classification codes would be the same. The datasets contained an old classification code for each case (row), for which users wanted to assign a new code. Naturally, without a frame (or the ability to link to one) and no means to recontact the respondent, there is no way of gathering the required information precisely.

However, a new classification code can be assigned probabilistically using proportions from a conversion matrix. As an example, if an old code maps to three new codes in the ratio 60%: 30%: 10%, then a Uniform (0,1) random number can be generated for the case, and a new code generated according to which of the intervals (0.0,0.6], (0.6, 0.9] or (0.9, 1.0] the random number lies in. Although this does not guarantee that the code for any particular case is correct, at an aggregate level distributions should be reasonable. The method can be refined further, as required, for example for longitudinal or panel data, and judicious use of case IDs as random number generator seeds can be used to ensure consistency of coding over time.

To this end, the ONS made conversion matrices and program code in the more popular statistical programming languages available to researchers to dual-code their own datasets in a controlled way.

4.2. *Microdata Methods*

The foundation of micro-methods lies in the survey responses themselves. Each sampled unit has two codes, SIC (2003) and SIC (2007), one of which will have been used for stratification, and the returns are reaggregated to form estimates on the other classification. There are different methods that may be used for this (domain estimation, poststratification, and others), each with its own advantages and disadvantages. Four alternative approaches are considered by Van den Brakel (2010). The method most commonly used for ONS outputs, a domain estimation approach, is described here.

We first note that estimates for most ONS business surveys are derived using the ratio estimator. For an estimated total of a variable y, $\hat{t}_y$, this may be written as follows (Särndal et al. 1992, Eq. 6.5.9):

$$\hat{t}_y = \sum_{i \in s} d_i g_i y_i$$

where:

d_i is the design weight (the inverse of the selection probability), and reflects the stratification of the survey.

g_i is a calibration factor, in general determined by defining calibration groups and an appropriate estimator (from a very wide class). The specific case of the (within-stratum) ratio estimator gives $g_i = X_h / \sum_{i \in s_h} d_i x_i$ for i in stratum h where X_h is the known total of x.

y_i is the (sometimes Winsorised) survey response, or imputed value in the case of nonresponse, for the variable y. i indexes the businesses, and s is the selected sample. It was decided to maintain the Winsorisation from the original classification for resource and consistency reasons. However, another option would be to reassess outliers on the poststratification on the new classification.

The choice of the most appropriate calibration groups for outputs was discussed at length, and the Government Statistical Service Methodology Advisory Committee was consulted (GSS MAC 2008). Under stratification by only one SIC, the calibration groups are usually just the sampling strata, or groups of sampling strata. However, with two classifications being used for simultaneous outputs, the choice is not clear. Any change in the groups (for example, to cross-classification of the SICs) would lead to a break in the time series, and would also present a risk of small sample sizes in each group. Therefore, the decision was made to calibrate the SIC (2003) estimates within SIC (2003)-defined groups, and SIC (2007) estimates within SIC (2007)-defined groups, with the compromise of the two sets of outputs being calibrated differently. Thus, for example, during periods in which the surveys were stratified by SIC (2003), the two sets of estimates would be compiled as follows:

$$\hat{t}_y^{03} = \sum_{i \in s} d_i^{03} g_i^{03} y_i$$

$$\hat{t}_y^{07} = \sum_{i \in s} d_i^{03} g_i^{07} y_i$$

(with natural notation) and when stratified by SIC (2007), the estimates would be given by

$$\hat{t}_y^{03} = \sum_{i \in s} d_i^{07} g_i^{03} y_i$$

$$\hat{t}_y^{07} = \sum_{i \in s} d_i^{07} g_i^{07} y_i$$

It was felt that having consistency within time periods was better than having consistency between the SIC (2003) and SIC (2007) outputs. This is a composition of the methods in Sections 3 and 5 of Van den Brakel (2010), with a need for a way to link the estimates from the two approaches at the transition point.

4.3. National Accounts

The main conversions for National Accounts purposes were based on SUTs. These provide the basis for calculation of Gross Value Added (GVA) weights to allow other elements of the National Accounts to be weighted together appropriately for the new classification. The industrial groupings to be used for SUTs based on the new classification were agreed quite early, so that they could be used as a basis for the revised stratifications in surveys (see Section 3). The change to a new product classification, CPA 2008, occurred at the same time, which had an effect on the way groups were put together. A consultation provided evidence which was used to form a 114-group classification consistent with SIC (2007) and CPA 2008, which encompassed international reporting requirements and provided some additional detail for users within the UK. These groups form the main processing level in the construction of the National Accounts, and component processes have also moved to this structure, introducing greater harmonisation into National Accounts compilation. We focus on the industrial classification changes rather than the product classification ones in this section.

Supply-side estimates in the SUTs up to 2006 were converted by the application of the conversion matrices used for the surveys. The later conversion of National Accounts allowed a stable set of matrices to be used throughout the accounts, mostly derived from the final IDBR reclassification, but with specific information for industries not covered by the IDBR derived from a range of different sources and used to complete the conversion. Since the basis of National Accounts is monetary, the turnover-based versions of the conversion matrices were used. Such a conversion is closest to the concepts used in measuring output and intermediate consumption. The IDBR does not adequately cover financial industries and nonmarket producers, and in these cases turnover is often not a suitable target for measuring output. Thus, other sources were used to derive conversion factors here (many of which were largely unaffected by the change in classification, and so could be dealt with by one-to-one conversions), so that all the elements of the economy were covered. Once the full matrix was available, it was used to convert the industry totals (on one margin of the SUTs) to give totals on the SIC (2007) classification.

SUTs contain considerable detail on products, some of which is derived from the EU's harmonised product survey, PRODCOM. The parallel update to the EU's product classification system (to CPA 2008) allowed product patterns in SIC (2003) industries to be converted to CPA 2008. Then, the product distributions were applied to the converted industry totals by associating SIC (2007) industries with one or more SIC (2003) industries (Drew and Dunn 2011). This led to some inconsistencies between the industry totals and the product totals, which required some further balancing interventions. The SUT is a matrix, and its major entries are generally on the diagonal (showing the principal products of the various industries). In cases where manual interventions were required, there was therefore a working assumption that the diagonal entries should be maintained as far as possible. While aggregate values should conceptually remain unchanged through such a conversion process, there were some minor revisions to totals in the SUTs up to 2006 in moving to the SIC (2007) classification. The largest of these was 0.05% of the total, so the old and new series are indistinguishable graphically. The overall effect of the classification change on GDP is included in the differences shown in Figure 4; it is confounded with some other changes introduced at the same time, so it is not as clear as having the classification change separately identified, but it gives an impression of the reclassification effect.

The SUTs for 2007 to 2009 could be based on data classified according to SIC (2007). The 2007 estimates from the ABS, the main survey source for SUTs, were reworked using the SIC (2007) classification, and from 2008 onwards the ABS was designed on SIC

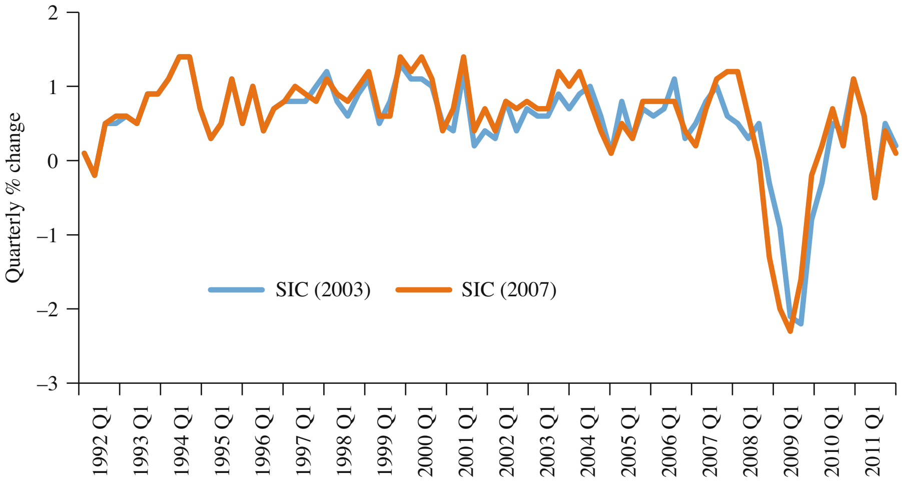


Fig. 4. Time series of quarterly changes (%) in UK GDP before and after the implementation of the reclassification to SIC (2007). The differences are due to reclassification and some additional changes made at the same time, so only part of the difference is the reclassification effect.

(2007) and could be used directly. The 2008 and 2009 ABS estimates were initially converted, and used with existing monthly movements to provide an interpolated series of within-year changes. Such changes are important and for interpretation they need to be seasonally adjusted. In order for seasonal adjustment of this derived monthly series on the new classification to be sufficiently stable, there was a need for a longer time series, so additional years of ABS data were also converted and monthly paths interpolated.

To ensure that estimates were consistent over time, any statistically important differences between the three parts of the series – estimates processed by conversion matrices, the reworked data in 2007 and the newly reclassified data for 2008 – were taken into account. For the ABS estimates, this was "done through a linking process where statistically [important] differences are blended into the historical estimates over time. This ensures that revisions to the levels of historical published data are minimised and the economic context of historical data is preserved as closely as possible" (Drew and Dunn 2011, 3). Yuskavage (2007) describes in detail a similar approach with series composed of three parts when the US Bureau of Economic Analysis introduced the North American Industry Classification System (NAICS); in their case the middle period was based on a succession of conversion matrices calculated over ten years by recoding the underlying industry data.

Seasonal adjustment of all time series forming the SUTs was also needed, and following the same principles as discussed above (Subsection 4.1), was undertaken after the conversions and any adjustments had been made.

5. Considerations in Choosing a Conversion Method

5.1. Disclosure Issues Associated with Changing Classifications

Reclassification potentially causes difficulties for statistical disclosure control, because small differences in classifications can result in 'slivers', domains which contain very few observations but which can be estimated quite accurately by differencing estimates on the two classifications (Hundepool et al. 2012, sec. 5.2.3). This can apply to the results of the reclassification, or to the conversion matrix itself. For example, ONS (no date) were able to calculate only a restricted range of ABS estimates for 2008 on the old SIC (2003) classification because of disclosure issues, which are described as 'especially complex' and not detailed (a common practice to increase disclosure protection). Similarly, Bayard and Klimek (2004) could not use some cells of the conversion matrix from the US Census Bureau because they failed disclosure tests. In many cases, the effect of converting value series and deflators separately will result in added protection from disclosure for the final outputs, as long as care is taken with any release of the conversion matrix itself.

5.2. Quality of Conversions

5.2.1. Consistency

The application of the model-based conversion matrices to part of a time series, with other parts of the time series covered by direct estimation from surveys designed on different classification systems generates some fairly standard problems in consistency of surveys, and general guidance on producing consistent time series of estimates in this case is given by Van den Brakel et al. (2008). A model-based approach to this involves the Kalman filter, and Bollineni-Balabay et al. (2016) give a complex example covering multiple changes, which demonstrates the efficacy of the procedure in more complex cases than reclassification.

The use of conversion matrices generally builds in some consistency of estimates. As long as the matrix covers all of the measured activity, and the new classification does not measure any activity not previously covered, the conversion approach guarantees that the total in the economy is unaltered. Reprocessing the microdata does not have the same property, and this results in changes to estimated totals.

In the UK, almost all of the National Accounts were converted by application of the conversion matrices based on turnover, which were derived from the business register. Conversion of other variables (such as fixed capital formation, FCF) using this matrix gives consistency, but is not necessarily the best for FCF. A conversion matrix could be constructed for FCF, but since this is not a variable in the IDBR, it would need to be constructed from a dual-classified survey source, and would therefore be subject to sampling error. It is an interesting question whether a matrix tailored for a specific variable but with sampling error is better than a nontailored matrix (which relies on the relationship between the target variable and the variable in the conversion matrix) based on the whole population.

5.2.2. Historical Consistency

It is interesting to consider what the effect of successive reclassifications is on historical data. The UK has had six national classifications since the first harmonised classification was introduced in 1948. It is not known for certain whether adjustments were made in the early transitions to new classifications, but certainly the updates to SIC (80), SIC (92) and SIC (2007) have involved the application of conversion matrices, and the SIC (92) and SIC (2007) conversions required adjustments to the National Accounts. This means that the oldest information has had a series of model-based adjustments applied to keep it consistent with modern classifications, so that users can use long time series in their economic models. Soroka et al. (2006) consider the effects of such changes, and conclude that the final adjusted data no longer bears a close resemblance to the original estimates (both in the level and in some cases the pattern) – historical accuracy is downplayed in favour of long-term historical consistency.

5.2.3. Timing Consistency

Because of the different periodicities of structural and short period surveys, samples are often selected for different periods at approximately the same time. When classifications change, this can mean that the timing of the change in classification is different for different types of surveys (see, for example Walker 1993). The same effect was felt for the change to SIC (2007), when the annual surveys changed to sampling on the new classification in 2008, for which samples were selected at the end of 2008, and the monthly and quarterly surveys changed from January/Q1 2010. This leads to consistency issues between annual and short period surveys for estimates for 2008 and 2009. These are ameliorated through conversions to and from the new classification.

5.2.4. Accuracy

Ultimately the decision to change classifications is taken to improve the accuracy of the outputs based on the classification – making it relevant to current industrial organisation and emerging activities. Making classifications current necessarily makes them less relevant for historical statistics, and care must be taken in long-term backcasting not to introduce clearly nonsensical patterns (such as production of high-definition televisions in the 1980s).

At the time a classification is introduced, there are some more-or-less short-term effects on the accuracy of the resulting series, which are a consequence of the process of changing. There is the time taken for the new classification to stabilise (as businesses are progressively contacted through surveys and administrative processes and errors in assigning them to new classification codes are gradually corrected); the effect of the errors in the classification codes is generally to increase the variance and not to institute a bias (imputed classifications must be done in a stochastic manner where there is no other information in order to ensure this property).

The sample design also undergoes a transition, and has several possible effects. Variability in new classifications will introduce lack of homogeneity into new stratifications until the classification matures, which will increase the variance of estimates. The effect of changing to rotational sampling in new strata with a maximised, but smaller than steady state, overlap is used to reduce the beneficial effect from covariances between periods, and therefore to increase the variance of estimates of change.

Thus, during the transition period there are several effects that all act to increase the variability of survey estimates. Therefore, one strategy for maintaining quality during reclassification might be to temporarily increase sample sizes to compensate for the increase in variability; this approach was taken in a limited number of surveys in the Netherlands in 1993 (Beekman 1992). In the medium term, the move to a more relevant classification should marginally increase homogeneity in strata and therefore reduce variability.

5.2.5. Coherence and Comparability

The debate over the use of single, harmonised classification systems in official statistics has long been won, but there are still issues of historical coherence, particularly when different parts of series are converted by different methods. This requires attention to the joins between methods, which need to be adjusted so as to give a coherent time series with the best estimates on the new classification.

5.3. Communication

The implementation of a new classification system is an important event for users of statistics because of the impacts on their own inferences, and it is therefore important that the plans and procedures being used to update the classifications are communicated to them. There is a case for providing some preliminary information on the effects of reclassification (as was done for the Retail Sales Index in the UK (McLaren 2010), as a way of helping users by giving them early warning of the impacts of the change in classification.

Once the material changes have been made, it is important for the metadata which describes these changes to be recorded and readily available to users to enable them to use this information to support their own analysis. Some of the metadata is directly published, such as the description of the new classification, its codebook and any indices. However, other information, such as how particular portions of a time series have been converted and the methods by which these pieces have been spliced together coherently, can be less easily discovered, and should also be made explicitly available.

The effect of classification changes is not generally well presented in graphical presentations. It has become more commonplace (at least in the UK) to label graphs with significant events that help users to understand the evolution of time series. At least at the time of issuing reclassified estimates, it would be beneficial to have visualisations which highlight where the changes in methods have taken place within a time series, in a similar way to Figure 3.

6. Conclusions

It is necessary to update classification systems occasionally so that they remain relevant, by accounting for changes and developments in the goods and services which form the economy. This is probably even more important now, as innovative products are frequently introduced to the market, and as the modern economy starts to take on new forms. As a consequence, there will be changes to statistical series, and these will reduce the quality of estimates for a period, while the new classification is introduced and settles down, in the sense that statistical units move from temporary codes at the instant of transition to ‘correct’ codes with time.

This process affects many stages of the statistical production process (Generic Statistical Business Process Model, UNECE 2013), but will most usually be felt as a period where the variability of estimates is increased somewhat. It is necessary to work through the detail of a reclassification procedure – there are many steps in processing, and all of them need to be consistent with the new classification, otherwise there is a large risk of having estimates that are not coherent. Similarly, there is a large job of quality assurance of the outputs from the various methods for converting series onto a new classification – a need to check the credibility of, for example, macro-methods applied across historical classifications, when the model assumption of consistency over time breaks down. In general, conversions should be checked to identify anomalies such as missing series, sizeable changes in turnover ratios, spikes that are attributable to a single business rather than an industry class as a whole and others. Time series consistency is usually a strong requirement from users, though it can have the effect of making estimates very different (in level and sometimes in evolution) on new classifications from their original estimates on the original classification (Soroka et al. 2006).

The temporary effects on the quality of statistical outputs at the time of the classification update are necessary in order to avoid more widespread effects on quality as a result of the classification itself becoming out of date. In the medium term, there are benefits. The new industrial structure should better measure new industries, and enable rapidly growing sectors to be separately identified so that they can be monitored.

We conclude by offering some considerations on the process of implementing a new industrial classification in a national statistical institute, which may be of use to other organisations undertaking such a change:

- With limited resources, it will be impossible to do everything. A better approach is to decide on areas which are most important, and to focus on these, spending most time ensuring the quality of backcasts of these estimates. Even so, it will be unlikely that all user requirements can be met, and it must be accepted that the quality of some converted series will be lower than that of the original estimates, possibly substantially so.
- The various options available in conversion, with no unique, obviously 'right' approach, can lead to difficult decisions. Where the quality resulting from any one approach cannot be demonstrated as being appreciably better than others, it may be prudent to choose approaches that can be easily explained and justified.
- Discontinuities will be introduced in some time series. These, and differences in treatment of different periods in series should be clearly identified for users (for example Figure 3). Explanations of these differences should be presented to users. (Obviously when fitting seasonal models, discontinuities should be estimated from the evolution of the series, and then removed.)
- Good and clear communication is essential. Document the decisions made and the methods employed; these will be invaluable next time.
- Retain information on the effects of changes in classifications (separately from other changes, if possible), and on experimental work to investigate alternative approaches and their effect on outputs. We had hoped to include more examples in this article to illustrate different effects, but the detailed comparisons were not available. Gathering these details will provide extra evidence with which to evaluate approaches and confirm (or otherwise) the validity of our recommendations.
- The period of implementation is likely to be long. Plan ahead, and get input from those areas moving last (for example, National Accounts) at the start, to inform redesign work at the beginning.
- Take the opportunity to review and improve other aspects of data collection and the production of statistics when implementing a change in classification; the additional cost of doing this at the same time may be quite low.

7. References

Bayard, K.N. and S.D. Klimek. 2004. "Creating a Historical Bridge for Manufacturing between the Standard Industrial Classification System and the North American Industry Classification System". In Proceedings of the Business and Economic Statistics Section: American Statistical Association, August 2003: 478–84.

Beekman, M.M. 1992. "Development and Implementation of a New Standard Industrial Classification." *Netherlands Official Statistics* 7: 18–26.

Bollineni-Balabay, O., J. van den Brakel, and F. Palm. 2016. "Multivariate State-Space Approach to Variance Reduction in Series with Level and Variance Breaks Due to Survey Redesigns." *Journal of the Royal Statistical Society, Series A* 179: 377–402. Doi: http://dx.doi.org/10.1111/rssa.12117.

Cochran, W.G. 1977. *Sampling Techniques*. New York: Wiley.

De la Fuente Moreno, Á. 2014. "A "Mixed" Splicing Procedure for Economic Time Series." *Estadística Española* 56: 107–121.

Drew, S. and M. Dunn. 2011. *Blue Book 2011: Reclassification of the UK Supply and Use Tables. Office for National Statistics*. Newport: ONS. Available at: http://www.ons.gov.uk/ons/rel/input-output/input-output-supply-and-use-tables/reclassification-of-the-uk-supply-and-use-tables/reclassification-of-the-uk-supply-and-use-tables-pdf.pdf?format=hi-vis (accessed 9 January 2017).

Eurostat. 2015a. *ESS Handbook for Quality Reports, 2014 Edition*. Luxembourg: Publications Office of the European Union.

Eurostat. 2015b. *ESS Guidelines on Seasonal Adjustment, 2015 Edition*. Luxembourg: Publications Office of the European Union.

GSS MAC (2008) GSS MAC 15 – minutes. Available from http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/2008-2011/15th-meeting/gss-mac-fifteenth-meeting-minutes.pdf (accessed 15 January 2017).

Hughes, J.C. 2008. "SIC 2007: Implementation in ONS." *Economic and Labour Market Review* 2: 41–44.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.-P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester: Wiley. Doi: http://dx.doi.org/10.1002/9781118348239.

Johnson, W., S. Paben, and J. Schilp. 2012. The Use of Sample Overlap Methods in the Consumer Price Index Area Redesign. In Proceedings of the fourth International Conference on Establishment Surveys (ICES-IV), 12–14 June 2012. Available at: www.amstat.org/meetings/ices/2012/papers/301813.pdf (accessed 27 October 2016).

Lindblom, A. 2003. *SAMU 4 The System for Coordination of Frame Populations and Samples from the Business Register at Statistics Sweden. Background Facts on Economic Statistics* 2003: 3. Stockholm: Statistics Sweden.

MacDonald, B. 1995. "Implementing a Standard Industrial Classification (SIC) System Revision." In *Business Survey Methods*, edited by B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott. 115–129. New York: Wiley. Doi: http://dx.doi.org/10.1002/9781118150504.ch7.

Mach, L., P.T. Reiss, and I. Şchiopu-Kratina. 2006. "Optimizing the Expected Overlap of Survey Samples via the Northwest Corner Rule." *Journal of the American Statistical Association* 101: 1671–1679. Doi: http://dx.doi.org/10.1198/016214506000000320.

McLaren, C. 2010. *Classification Changes in Retail Sales*. Newport: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/retail-sales/classification-changes-in-retail-sales.pdf?format=hi-vis (accessed 9 January 2017).

Ohlsson, E. 1995. Co-ordination of Samples Using Permanent Random Numbers. In *Business Survey Methods*, edited by B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott. 153–169. New York: Wiley. Doi: http://dx.doi.org/10.1002/9781118150504.ch9.

ONS no date. *SIC 2003 Comparison of Main Industrial Groupings in 2007 and 2008 by Key Variables*. Newport: ONS. Available at: https://www.ons.gov.uk/file?uri=/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007/abiindustrygroupings_tcm77-247614.pdf (accessed 9 January 2017).

ONS. 2009. *UK Standard Industrial Classification of Economic Activities 2007 (SIC 2007) – Structure and Explanatory Notes*. Basingstoke: Palgrave MacMillan. Available at: http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/standard-industrial-classification/sic2007—explanatory-notes.pdf (accessed 9 January 2017).

ONS. 2010. *Weighted Tables with Percentages UK SIC 03 - UK SIC 07 and UK SIC 07 - UK SIC 03*. Newport: ONS. Available at: https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007 (accessed 9 January 2017).

Russell, M., P. Takac, and L. Usher. 2004. "Industry Productivity Trends under the North American Industry Classification System." *Monthly Labor Review* November 2004: 31–42.

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

Schiopu-Kratina, I., J.M. Fillion, L. Mach, and P.T. Reiss. 2014. "Maximizing the Conditional Overlap in Business Surveys." *Journal of Statistical Planning and Inference* 149: 98–115. Doi: http://dx.doi.org/10.1016/j.jspi.2014.02.002.

Scottish Government 2012. *Supply-Use Tables and Standard Industrial Classification (2007) Conversion*. The Scottish Government. Available at: www.gov.scot/Resource/0042/00421501.doc (accessed 28 October 2015).

Smith, P. 2013. "Sampling and Estimation for Business Surveys." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, G. Haraldsen, J. Jones, and D.K. Willimack. 165–218. Hoboken, New Jersey: Wiley. Doi: http://dx.doi.org/10.1002/9781118447895.ch05.

Smith, P. and S. Penneck. 2009. *100 Years of the Census of Production in the UK*. GSS Methodology Series No 38. Newport: ONS. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series–38–100-years-of-the-census-of-production-in-the-uk.pdf (accessed 9 January 2017).

Smith, P., M. Pont, and T. Jones. 2003. "Developments in Business Survey Methodology in the Office for National Statistics, 1994–2000" (with discussion). *Journal of the Royal Statistical Society, Series D* 52: 257–295. Doi: http://dx.doi.org/10.1111/1467-9884.03571.

Soroka, S.N., C. Wlezien, and I. McLean. 2006. "Public Expenditure in the UK: How Measures Matter." *Journal of the Royal Statistical Society: Series A* 169: 255–271. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00397.x.

Taylor, C., G. James, and P. Pring. 2011. "The Development of the Monthly Business Survey." *Economic and Labour Market Review* 5: 95–103.

UNECE. 2013. *Generic Statistical Business Process Model GSBPM (Version 5.0)*. Geneva: UNECE. Available at: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0 (accessed 10 November 2015).

Van den Brakel, J. 2010. "Sampling and Estimation Techniques for the Implementation of New Classification Systems: the Change-Over from NACE Rev. 1.1 to NACE Rev. 2 in Business Surveys." *Survey Research Methods* 4: 103–119. Doi: http://dx.doi.org/10.18148/srm/2010.v4i2.2354.

Van den Brakel, J.A., P.A. Smith, and S. Compton. 2008. “Quality Procedures for Survey Transitions –Experiments, Time series and Discontinuities.” *Survey Research Methods* 2: 123–141. Doi: http://dx.doi.org/10.18148/srm/2008.v2i3.68.

Walker, C. 1993. “Transition to the New Standard Industrial Classification (SIC(92)).” *Economic Trends* 472: 88–94.

Williams, N. 2006. “ACTR/IDBR Test Evaluation Report.” *Survey Methodology Bulletin* 57: 33–38.

Yuskavage, R.E. 2007. “Converting Historical Industry Time Series Data from SIC to NAICS.” The Federal Committee on Statistical Methodology 2007 Research Conference, November 5–7 2007. Available at: https://www.bea.gov/papers/pdf/SIC_NAICS.pdf (accessed 9 January 2017).

Received December 2015
Revised November 2016
Accepted November 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 249–274, http://dx.doi.org/10.1515/JOS-2017-0013

Cost-Benefit Analysis for a Quinquennial Census: The 2016 Population Census of South Africa

Bruce D. Spencer[1], *Julian May*[2], *Steven Kenyon*[3], *and Zachary Seeskin*[4]

The question of whether to carry out a quinquennial Census is faced by national statistical offices in increasingly many countries, including Canada, Nigeria, Ireland, Australia, and South Africa. We describe uses and limitations of cost-benefit analysis in this decision problem in the case of the 2016 Census of South Africa. The government of South Africa needed to decide whether to conduct a 2016 Census or to rely on increasingly inaccurate post-censal estimates accounting for births, deaths, and migration since the previous (2011) Census. The cost-benefit analysis compared predicted costs of the 2016 Census to the benefits of improved allocation of intergovernmental revenue, which was considered by the government to be a critical use of the 2016 Census, although not the only important benefit. Without the 2016 Census, allocations would be based on population estimates. Accuracy of the postcensal estimates was estimated from the performance of past estimates, and the hypothetical expected reduction in errors in allocation due to the 2016 Census was estimated. A loss function was introduced to quantify the improvement in allocation. With this evidence, the government was able to decide not to conduct the 2016 Census, but instead to improve data and capacity for producing post-censal estimates.

Key words: Demographic statistics; fiscal allocations; loss function; population estimates; post-censal estimates.

1. Introduction

1.1. *Background on Costs and Benefits of Mid-Decade Censuses*

At all times, but especially in challenging economic times, governments considering investment in an accurate census or other social information face simultaneous decision problems of how much to invest and how much accuracy to seek. In the United States, the constitutional requirement of a census every ten years has been met, at increasing cost, and with varying degrees of accuracy. On the other hand, the US Congress has never provided

[1] Department of Statistics and Institute for Policy Research, Northwestern University, Evanston, IL 60208-4070, U.S.A. Email: bspencer@northwestern.edu
[2] DST-NRF Centre of Excellence in Food Security, University of the Western Cape, Bellville, South Africa. Email: jmay@uwc.ac.za
[3] National Treasury, Pretoria, South Africa. Email: steven.kenyon@treasury.gov.za
[4] NORC at the University of Chicago, Chicago, IL, 60603, U.S.A. Email: Seeskin-Zachary@norc.org

Acknowledgments: This research was supported a grant from the US National Science Foundation (SES-1129475). We thank Thamsanqa Mohale of the South African National Treasury and Diego Iturralde and Louis van Tonder of Statistics South Africa, for the financial and demographic statistics inputs that they compiled for this report, and Ben Mphahlele of the Statistics Council for recognizing and convincing others of the need for a cost-benefit analysis. Comments received from the Statistics Council, the Statistician General, Mr. Pali Lehohla, and the Associate Editor are also gratefully acknowledged.

funds for a mid-decade census, despite the legal requirement that a mid-decade census be carried out “in the year 1985 and every ten years thereafter” (Census Act of 1976, PL 94-521; 13 USC §141(d)). Since the 2010/2011 round of census-taking, media reports suggest that the timing and format of national censuses is being debated in several countries including Australia (The Guardian 2015), Canada (The Globe and Mail 2011), Ireland (The Journal 2012), and Nigeria (Nigerian Tribune 2016).

In South Africa, the Statistics Act of 1999 requires “a population census to be taken in the year 2001 and every five years thereafter . . . unless the Minister, on the advice of the Statistician-General . . . determines otherwise.” The Act further provides for an independent Statistics Council to advise both the Minister and the Statistician-General on a wide range of matters pertaining to official statistics, with the taking of a population census specifically identified. In accordance with the law, censuses were taken in 2001 and 2011, but not in 2006, under the advice of the Statistics Council. The analysis described in this article was prepared as part of the evaluation of the 2011 Census to help the Statistics Council advise the Statistician-General and the Minister responsible for official statistics on whether a 2016 Census should be carried out by the government statistical agency, Statistics South Africa (Stats SA).

Considering the costs and benefits of government data programs, such as the 2016 Census, is essential for making informed decisions on how much to invest in such data programs. In November 2009, representatives of national statistical agencies and UN agencies met in Dakar to discuss improving the provision of statistics in the context of the United Nations Millennium Development Goals. The Dakar Declaration on the Development of Statistics that followed from this meeting proposed that official statistics are a public good, and that their production and dissemination is a core responsibility of all governments. Considering the costs and benefits of data programs is necessary because the market does not lead to socially optimal investment in public goods (Sims 1984). The cost of a 2016 Census is estimated to be at least ZAR 3 billion, which was the cost of the 2011 Census. (All amounts are in 2011 prices and at the time of the census, the South African Rand was equivalent to USD 0.14.) Note that the value of the census really refers to the added value of the census, compared with the value of alternatives, in particular a large sample survey to provide data on inter-provincial migration since the 2011 Census. The more accurately population change can be measured without a census, the less is the 2016 Census’s value, ceteris paribus.

Benefits of data programs arise largely from their use, and understanding the causal pathways by which outputs from the data program affect outcomes is enormously complex. In particular, we would want to predict the outcomes if the 2016 Census were to be conducted and the outcomes if it were not conducted. The benefit of the 2016 Census reflects the difference in the value of the outcomes in the two scenarios, and therefore outcomes that would be the same in both scenarios can be ignored in the analysis. Even so, to consider all actions or outcomes by carrying out or not carrying out the 2016 Census is not feasible. Furthermore, assigning values (e.g., monetary values) to outcomes is challenging with regard to many uses of statistics (Spencer 1982a).

The impossibility of studying all the benefits of a major data program, such as a census, implies that cost-benefit analysis of the program must, necessarily, be incomplete in that some benefits – perhaps even the majority of the benefits – will be unmeasured. Our analysis is a partial cost-benefit analysis, in that not all benefits are considered. As discussed below, we focus on just a single use of the census data: allocation of national funds to subnational jurisdictions by formulas. There are many other uses of census statistics which may be important. For consideration of other benefits from the South African Census, see May et al. (2013). The Office for National Statistics in the United Kingdom explicitly considered costs and benefits of the 2011 Census after receiving the recommendation of the "House of Commons Treasury Select Committee . . . that: "any future Census should also be justified in cost-benefit terms" (Cope 2015, 2). However, the detailed "business case" that was developed to "make the case" for the 2011 Census is not publicly available, only a high-level summary (Parliament of the United Kingdom 2009) and links thereto discussing some identified uses. The business case analysis for the 2011 Scotland Census is available, and it contains an analysis of shifts in fund allocations to Health Board Areas that would have occurred with a 2001 Census and without a 2001 Census (in which case post-censal estimates would have been used). (General Register Office for Scotland 2006, 27–34). Bakker (2014) analyzes costs and benefits of the New Zealand census. However, in all of these studies, the quantification of benefit of nonallocative uses of census statistics typically is highly uncertain.

The earliest identified cost-benefit analysis of a quinquennial census is that of Redfern (1974), who focused on benefits of more accurate fund allocations arising from a mid-decade census in England and Wales; the analysis did not appear to support carrying out a mid-decade census there (Spencer 1980a, 13–17; Alho and Spencer 2005, 368). Spencer (1980a) conducted a cost-benefit analysis comparing two alternative versions of the 1970 US Census. Seeskin and Spencer (2015) analyze benefits of improved allocations of funds and political representation under alternative accuracy profiles of the 2020 US Census. May and Lehohla (2005) discuss reasons for cost increases in South Africa's 2001 Census, but only describe some of the benefits.

Assigning values to alternative outcomes is a challenge for cost-benefit analysis of data programs. To compare costs and benefits most directly, it is convenient for benefits to be quantified in the same units as costs. However, when such a comparison is not feasible, the issue should not be forced. Instead, summaries can be prepared showing what benefits are attainable at what costs. Savage (1985) and Sims (1984) offer cautionary critiques of misguided attempts to force benefits of data programs to be measured in units comparable to those used by costs.

A partial cost-benefit analysis of a data program should not be narrowly interpreted as a formal set of calculations that will point to the "correct" or "optimal" decision (Savage 1985, 4). Cost-benefit analysis in the narrow sense can be misleading when applied to data programs, as pointed out by the National Research Council (1985).

> Cost-benefit analysis, as we understand and use the term, means describing a program as a set of commodities produced (benefits) and a set of commodities consumed (costs) and aggregating those using prices, market prices when possible, otherwise "shadow prices" that emerge from calculations based on assumptions of optimization, either by individuals or by components of a market economy.

> With information dissemination programs, this analytical framework is not helpful. Technical analysts can determine some of the political and economic decisions to which the information is relevant, and they can look for alternative pathways through which the information might flow, if the program were reduced or eliminated. But these efforts will involve tracing out the operation of incomplete and imperfect markets and of nonmarket information transfer mechanisms; the usual practices of relying on market prices and on the uniqueness of the values of traded goods will not be available. Trying to proceed nonetheless to attach dollar values to the effects of the information will nearly always lead to guesswork and arbitrary assumptions that obscure, rather than clarify, the analysis. (54–55)

We use the term cost-benefit analysis in the broad sense of providing a way of thinking about, and a way of organizing information on, some of the benefits and costs of a data program. There should be no automatic presumption that the measured benefits will outweigh the measured costs, even in a data program that is implemented in full, in the sense that the difference between its actual benefits and its actual costs is greater than for other programs. Failure to demonstrate that measured benefits exceed costs does not mean that the data program is unjustified or should not be carried out. The value of a cost-benefit analysis is a reduction in the uncertainty concerning the benefits and costs, and in an ideal world this would improve decisions concerning statistical programs. However, there is a risk in this approach that decision-makers may conclude that a data program is not worth funding if the partial cost-benefit analysis does not show benefits exceeding costs.

Although additional practical constraints on statistical agencies could, in principle, be incorporated into the cost function, factors other than cost may influence whether a data program is carried out. These potential factors include the capacity of the responsible institution to undertake data collection, competing demands by other data collection programs, and anticipated technology or methodology changes that improve the accuracy of estimating the population. In the case of capacity constraints, the institution may opt to reprioritize its work program, delaying or suspending other data collection activities in order to undertake the activity which it deems a priority. In the case of technology or methodology changes, improvements in the capacity to sample, such as satellite imagery, may permit the institution to opt for a large survey rather than a full census, thereby affecting the cost function of an alternative to a census.

There are major limitations in scope to partial cost-benefit analyses that must be communicated by researchers. If incorrectly interpreted, a partial cost-benefit analysis could do more harm than good. Key assumptions must be presented in a transparent way. Decision-makers within the statistical agency should be aware of all the limitations. In their communication with decision-makers and the general public, the researchers should explain the limitations in an understandable, albeit abbreviated form.

1.2. Legal Context for the Census in South Africa

In South Africa, census-taking has a longstanding and sometimes controversial history dating back to the 18th century. However, most Statistics Acts (1976, 1978, and 1980) and censuses were designed during the apartheid regime, and therefore considered to be too narrow and insufficient to protect and promote the rights of all citizens of South Africa. To address the limitations of the previous Acts, the current democratic South African Government designed the 1999 Statistics Act (Act No. 6 of 1999). The Act provides for "a Statistician-General as head of Statistics South Africa, who is responsible for the collection, production and dissemination of official and other statistics, including the conducting of a census of the population, and for coordination among producers of statistics; to establish a Statistics Council and provide for its functions; to repeal certain legislation; and to provide for connected matters." The first responsibility of the Statistician-General specified in the Act is to "cause a population census to be taken in the year 2001 and every five years thereafter . . . unless the Minister [of Finance, or other Minister as chosen by the President], on the advice of the Statistician-General . . . determines otherwise."

1.3. Uses of Census Data

The additional information that a 2016 Census would provide about the population would lead to changes of various kinds, including, but not limited to the following.

1. Under South Africa's system of multi-tier government, funds are allocated by the national government to provinces and municipalities on the basis of population and other data. Fund allocations will differ depending on whether a 2016 Census is carried out or not.
2. Additional social information about population sizes (for groups classified by geography, ethnicity, and other criteria) would be provided, along with information about internal migration and migration between South Africa and other countries. Such information is important for understanding, and may or may not lead to identifiable changes in actions or outcomes. May et al. (2013) discuss a survey conducted to yield some limited insight into this.
3. Surveys carried out by Stats SA and by other survey organizations can be designed more efficiently (using updated sampling frames) based on information that the 2016 Census will provide. The survey analysis is also improved by the availability of more accurate population totals for various and diverse subgroups, which can be used to calibrate the survey data.
4. Policy analyses in all spheres of government will change to some degree as a result of having the 2016 Census data available.
5. Social planning and allocation of funding for electricity, water, sanitation, education facilities, and telecommunications can be based on more accurate data about population distribution.
6. Businesses may make different decisions about where to locate, about product design, or about risk assessment.

In addition, a census can have an important ceremonial aspect and be taken as a symbol of government efficiency (or inefficiency, depending on point of view), as observed by Kruskal (1984) and confirmed in the survey of data users as discussed by May et al. (2013, viii).

Uses of census data for formula-based allocation of funds are perceived as important in the context of a multi-tier government system such as the one adopted by South Africa, and are the focus of the benefit analysis in this article. Subsection 1.4 provides further context.

Uses of census data for policy analysis (item 4 in the list) appear to be important as well. McCaa et al. (2006) discuss the strategic importance of the census in providing demographic, economic, and social data pertaining, at a specified time, to all persons in a country or a well-defined part of the country. They further note that a census helps in undertaking efficient management of economic and social policies or programmes, and one infers that census information is a key element in evidence-based policymaking. Indeed, concern for effects of population change and numbers is reflected by the South African Government's *White Paper on Population Policy*, which emphasizes "the need for reliable and up-to-date information on the population and human development situation in the country to inform policy making and programme design, implementation, monitoring and evaluation" (Ministry for Welfare and Population Development 1998, 16).

Understanding how data affect policy development and analysis is a challenge, and may require careful case studies of policy processes. Although we did not attempt this in full, we considered how changes in population numbers would affect outputs from the kind of microsimulation analyses that would be produced in the policy context, and found moderate impact (May et al. 2013, 32–35). The findings were communicated to the Statistics Council, the Statistician-General and the Minister responsible for Stats SA, but will not be further discussed in this article.

1.4. Formula-Based Allocations of Funds

The South African Constitution considers various aspects of intergovernmental fiscal relations, including the devolution of certain revenue and expenditure assignments to subnational governments. Responsibility for revenue generation is unequally distributed between the national, provincial and local spheres of government. The national government has a wide variety of tax instruments available for raising revenue. In contrast, the provinces have limited options for taxation, and the municipalities largely rely on property taxes and service charges. Although the revenue-generating power of municipal governments was strengthened following the Municipal Property Rates Act (2004), the bulk of national revenue accrues to national government (Yemek 2005, 9). To address this, the Constitution also provides for a nonpartisan Financial and Fiscal Commission (FFC) that advises parliament and subnational governments on a variety of issues concerning intergovernmental fiscal relations, including the allocation of revenue among the three spheres of government, that is, national, provinces, and municipalities. According to Section 214 of the Constitution, one of the two main instruments for transferring revenue from the national sphere to the other two spheres is the "equitable shares" program. The provincial equitable share accounts for around 80% of transfers to provinces and the local government equitable share accounts for over half of the transfers to municipalities (National Treasury 2015).

The provincial shares and local government shares are divided between the provinces and the municipalities according to revenue-sharing formulae that are revised periodically. The Provincial Equitable Share (PES) and Local Government Equitable Share (LGES) formulas are based on the demographic and economic profiles of the subnational jurisdictions, as revealed by population sizes and other statistics. To align with the mandated responsibilities of these jurisdictions, the PES has included the following components: an education share based on the average size of the school-age population (ages 5 to 17) and the number of learners enrolled in public ordinary schools; a health share based on the use of the public health system and the number of people without medical aid or health insurance; a general component based on population size. The LGES formula depends mainly on population numbers from the latest census, since updated population statistics are not available at municipal level in non-census years. This article is based on the LGES formula that was used prior to 2013, as this was the formula in use at the time the research was conducted. The new formula, introduced in 2013, is still driven mainly by the number of poor households in each municipality (National Treasury 2013, 34–43).

1.5. Refining the Set of Choices

In a cost-benefit or other decision analysis, it is important to specify the alternative choices and underlying assumptions. We assume that a census will be taken in 2021 irrespective of whether or not a census is taken in 2016. Further, we assume that if the 2016 Census is not taken, Stats SA will conduct a large sample survey in 2016 similar to the Community Survey undertaken in 2007, which sampled 300,000 households. This will provide data on inter-provincial migration since the 2011 Census. Uses of population numbers in 2016 will be unaffected by the 2016 Census, since the census results would not yet be available. Users of population numbers for 2022 and beyond will rely on the 2021 Census numbers. Although post-2021 analysis of population dynamics would still be improved by the availability of 2016 Census data, we assess the benefits of the improvement to be relatively small in comparison to other benefits of the 2016 Census. These considerations lead us to focus on benefits arising from uses of population numbers for the five-year period, 2017–2021.

If a 2016 population census is not carried out, province-level population numbers for 2017–2021 will be available from the mid-year population estimates, which are derived by allowing for births, deaths, and net movements into and out of each province since the time of the 2011 Census (Stats SA 2011). The first two are derived from civil registration of vital statistics, but the last item can only be estimated, as internal migration is not recorded and there is a potentially substantial unrecorded international immigration. Thus, in the absence of a 2016 Census, the mid-year estimates for provinces will need to account for 6–10 years of population change since the 2011 Census; the Community Survey will be useful for this. If the 2016 Census is conducted, the population numbers for provinces in 2017–2021 will again be provided by the mid-year population estimates. However, these need only account for 1–5 years of population change since the 2016 Census, and official population numbers below the province level will be 1–5 years out of date instead of 6–10 years. Mid-year estimates are not available below the province level. Thus, municipal population numbers for 2017–2021 will be based either on the 2016 census, if it is conducted, or on the 2011 Census, if no 2016 Census is carried out.

1.6. Organization of Article

As noted, we focus on the benefits of the 2016 Census that arise from improved allocations from the LGES and PES over the period 2017–2021. For this analysis, we treat the PES allocations as correct if the input data for the allocation calculations were entirely correct. A loss function for measuring the aggregate discrepancy between the calculated allocations, $\hat{\theta}$, and the correct (or "true") allocations, θ, is developed (Section 2). We consider two alternative ways in which $\hat{\theta}$ can be developed, according to the construction of mid-year population estimates for 2017–2021: the "*cen16*" alternative uses the 2016 Census results either as the estimates (LGES) or as the base for mid-year estimates (PES), whereas the "*nocen16*" alternative relies on the 2011 Census for municipal estimates (LGES) and as the base for the mid-year estimates for provinces (PES), supplemented by a 2016 Community Survey. To model the accuracy of the two alternative sets of mid-year population estimates, we assess the past performance of mid-year population estimates by comparing them to the 2011 Census results (Section 3), then we model their accuracy for 2017–2021 (Section 4). The distributions of PES and LGES allocations are then derived under the "*cen16*" and "*nocen16*" alternatives (Section 5), leading to estimates of improvement in allocation as a result of the 2016 Census. Limitations of the analysis are described (Section 7). After discussing census cost (Section 8), we discuss the benefits in light of the costs (Section 9). The article concludes with a brief discussion of the decisions made concerning the 2016 Census and alternatives (Section 10).

2. Use of Loss Functions to Measure Improvement in the Allocations of Funds

2.1. Loss Functions for Errors in Allocations

An important identified use of population census data in South Africa is the allocation of funds using a formula with inputs from statistics of various kinds and with an output that specifies the share that each province should receive. As already noted, the formula is called the Provincial Equitable Share (PES). A similar important use is the allocation of funds to municipalities using the Local Government Equitable Share (LGES) formula. The design and weighting of the formulas are agreed by intergovernmental forums that include provincial and municipal representatives. The formulas are also reviewed by an independent constitutional advisory institution, the Fiscal and Financial Commission (FFC). These formulas are used annually by the National Treasury to allocate shares of a total that is not affected by population statistics.

Distortions in the allocations arise from error in the data used to compute the allocations. We will use a loss function, as applied in statistical decision theory, to accomplish two purposes. First, the loss function will reflect rankings over alternative patterns of errors in allocation, with smaller loss corresponding to higher ranking and greater preference (National Research Council 1980, 84ff; Spencer 1980c). The loss functions considered here all take the value zero when there is no error in allocations arising from statistical error. The loss function is thus the negative of a utility function and satisfies the properties of a regret function (Berger 1985, 46ff, 376ff). The scale of the utility function is chosen (at least in theory), so that preferences under uncertainty, including risk aversion, are automatically taken into account when expected utility (or expected loss) is considered. Alternative axioms for preferences under uncertainty lead to focus on minimizing the maximum regret rather than expected regret or loss (Manski 2011). More generally, providing the probability distribution of loss – either the full multivariate distribution or the marginal distributions for each of the recipients (e.g., local governments) can be informative. Second, we will use the loss function to compare costs of improving data to the benefits in terms of improved allocations (Spencer 1980a, 31–33).

Different perspectives have been taken in the literature on the effects of the distortions in allocations. One perspective addresses inequities that arise because the allocations differ from those that would arise if the legislated formulas were applied to error-free data. A second perspective looks at inefficiencies and reductions in social welfare that are believed to arise when the allocations are based on data with error instead of error-free data. Our analysis will focus on inequities because we believe that measuring changes in social welfare caused by distortions in allocations arising from data error is simply too difficult.

2.2. *Loss in Social Welfare from Errors in Allocations*

Analyses of benefits of censuses arising from increased "utility" or social welfare have been conducted recently for England and Wales (Cope 2015) and New Zealand (Bakker 2014). Although the details of the analysis for England and Wales could not be discovered by the authors, Cope mentions differences in utility from overallocations and underallocations and refers to the sum of net differences as "efficiency loss." More details are available for the analysis of the value of the New Zealand (NZ) census and associated population statistics. Bakker (2014, 50–53) considered distortions in allocations with the NZ health funding formula. The analysis assumed that the allocations based on error-free population data maximized the welfare of NZ residents. In particular, let H_a and $\hat{H}_a$ denote the health expenditure allocations to area a with error-free data and actual data, respectively, and let X_a denote other final consumption expenditure to area a, with $a = 1, \ldots, A$. The analysis specified that the social welfare W from health formula allocations $\hat{H}_a$ and other final consumption expenditures X_a has the form $W(\hat{\mathbf{H}}, \mathbf{X}) = \sum_a X_a + u_a(\hat{H}_a)$ with $\hat{\mathbf{H}} = (\hat{H}_1, \ldots, \hat{H}_A)$, $\mathbf{X} = (X_1, \ldots, X_A)$, and $u_a(\hat{H}_a) = H_a \log(\hat{H}_a)$. This social welfare specification implies that the optimal distribution of a fixed sum equal to $\sum_a H_a$ occurs when the allocation to area a is indeed equal to H_a. The total loss from distortions in health expenditure allocations was taken to be $W(\mathbf{H}, \mathbf{X}) - W(\hat{\mathbf{H}}, \mathbf{X})$. This is non-negative and is equal to $\sum_a u_a(H_a) - u_a(\hat{H}_a)$ or $= \sum_a H_a[\log(H_a) - \log(\hat{H}_a)]$. It is important to note that, other than the assumptions of optimality and decreasing marginal utility from health-funding allocations as reflected by $u_a(\cdot)$, the analysis made no attempt to justify the specifications involving W and $u_a(\cdot)$. Different specifications would lead to different assessments of loss from distortions in allocations due to data error.

The assumption that an allocation formula is optimal should not be made casually. The United States' experience indicates diverse ways that formulas fail to be optimal (Buehler and Holtgrave 2007). The National Research Council (2003) report, *Statistical Issues in Allocating Funds by Formula*, commissioned several papers examining the design, development, structure, and inherent compromises in intergovernmental aid formulas. Downes and Pogue (2002) discuss the "often contradictory aid objectives . . . [and] assess the extent to which, in practice, formulas deviate from the ideal" (National Research Council 2003, 97). Zaslavsky and Schirm (2002) describe formula complexities such as hold-harmless provisions, floors, ceilings, and inconsistent data sources; they describe how their effects can be difficult to predict and can "produce allocations that don't line up with original intentions" (National Research Council 2003, 97). Similar critiques appear in Spencer (1982b). Melnick (2002) describes the legislative process by which allocation formulas "pass the test for face validity while generating the necessary political support" (National Research Council 2003, 97). Possibly, legislators are motivated to secure the most funding for their constituents, but by including factors representing need, capability, and effort, the formulas appear as if they are addressing program goals. A legislator who participated in the development of a complex formula for General Revenue Sharing, a program that would distribute more than USD 55 billion in the United States between 1972 and 1980, described the process this way: "We finally quit, not because we hit on a rational formula, but because we were exhausted. And finally we got one that almost none of us could understand at the moment. We were told that the statistics were not available to run the [computer] print on it. So we adopted it, and it is here for you today" (quoted in Spencer 1980a, 152). Furthermore, even if the formula could be regarded as optimal when the input data were error-free, the formula allocations may not be optimal, for example, if the allocations also depended on other data series that contained error. For example, Schirm et al. (1999) discuss estimation error for local governments.

In conclusion, analysis of benefits of improved data in terms of increased social welfare arising from more accurate formula-based allocation of funds should be used with caution, unless the formula can be demonstrated to be optimal and the form of the social welfare function can be justified.

2.3. *Loss from Inequity in Allocations Due to Data Error*

The very names of the PES and LGES, Provincial Equitable Share and Local Government Equitable Share, indicate the importance of equitable allocations in South Africa. Therefore, we did not attempt a social welfare analysis based on assumptions of formula optimality. Instead, we considered which patterns of distortions of allocations would lead to larger increases in inequity for the local governments and their people.

For the purposes of the analysis, the allocations will be considered to be correct if there is no error in the statistics used as inputs to the allocation formulas. We will index the n units (provinces or municipalities) receiving allocations by $i = 1, \ldots, n$. The correct allocation to recipient unit i will be denoted by θ_i and the allocation based on statistics will be denoted by $\hat{\theta}_i$. The arrays of allocations are respectively denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$. The component loss function for misallocation to unit i is denoted by $\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ and the aggregate loss equals the sum of the component losses,

$$\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}). \tag{1}$$

Summing the component losses to the recipients, as in (1), is consistent with a utilitarian view of social welfare measurement (Spencer 1985, 816–817). In addition to considering aggregate loss, it is important to also ensure that the expected component loss $E\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is not excessive for any recipient i. This principle could be extended to see that the upper quantiles of the component loss functions are not excessive for any recipient.

To motivate the form of the component loss functions $\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ consider the asymmetry of the recipients' views regarding positive and negative errors in allocation. If the error in the allocation, $\hat{\theta}_i - \theta_i$, is negative (an underpayment), the recipient unit suffers a shortfall equal to that amount. A simple measure of loss in this case is $a(\theta_i - \hat{\theta}_i)$ with $a > 0$. If the error in the allocation, $\hat{\theta}_i - \theta_i$, is positive (an overpayment), the recipient is receiving a positive benefit. In this case, a simple measure of loss is $-b(\hat{\theta}_i - \theta_i)$ with $b > 0$. A simple component loss function for recipient unit i that takes this perspective into account is

$$\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = a(\theta_i - \hat{\theta}_i)^+ - b(\hat{\theta}_i - \theta_i)^+, \tag{2}$$

where $(x)^+ = \max\{x, 0\}$. Perceiving an underpayment to be somewhat more consequential than an overpayment of the same magnitude, we have $a > b \geq 0$, but the ratio b/a will not be too small. For the PES and LGES, the fact that the total amount allocated is fixed implies that the sum of the overallocations must equal the sum of the underallocations, and hence

$$\sum_{i=1}^{n} a(\theta_i - \hat{\theta}_i)^+ - b(\hat{\theta}_i - \theta_i)^+ = c\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|, \tag{3}$$

with $c = (a - b)/2$. The non-negativity of b implies $c \leq a/2$. The value of c is considered further in Section 9.

The loss function (3) refers to one year's allocation at a time. To account for multiple years of allocation, we sum the loss functions for the individual years from 2017 through 2021 to obtain the aggregate loss function

$$\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = c\sum_{y=2017}^{2021}\sum_{i=1}^{n} |\hat{\theta}_{iy} - \theta_{iy}|. \tag{4}$$

In effect, this treats the years independently and does not allow for cancellation of a recipient unit's underpayment one year by equivalent overpayment the following year. However, the factor c does account for offsetting of underpayments and overpayments to different units in the same year. The benefit of reducing errors in allocations is measured by the reduction in the expected value of the aggregate loss when $\hat{\boldsymbol{\theta}}$ is developed with the availability of the 2016 Census data, versus when 2016 Census data are not available.

2.4. Additional Rationale for the Loss Function

In applying statistical decision theory, the optimality criterion should lead to the desired choices. The loss function (3) satisfies the criterion of Fisher-consistency, in that minimization of loss occurs precisely when the allocations are correct, that is, when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ (Spencer 1980a, 36). If Fisher-consistency is violated, then minimization of expected loss would lead to statistical inaccuracy being optimal, which is contrary to the principles of statistical agencies. A generalization of (2) is given by

$$\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = w_i\left[aH\left((\theta_i - \hat{\theta}_i)^+\right) - bH\left((\hat{\theta}_i - \theta_i)^+\right)\right], \tag{5}$$

with $w_i > 0$, $H(0) = 0$, and H strictly increasing on $[0, \infty)$. The criterion of Fisher-consistency imposes strong restrictions on the weights w_i and the shape of H in (5). Requiring that (1) remain Fisher-consistent for an arbitrary number n of recipients and any size errors in allocation leads to the conditions that

$$\frac{\max w_i}{\min w_i} < \frac{a}{b} \tag{6}$$

and

$$A \leq \frac{H(x)}{x} \leq B \tag{7}$$

for $x \geq 1$ and for positive constants A, B not depending on n (Spencer 1980a, 41–46). Condition (6) implies that the weights cannot be inversely proportional to θ_i, for example, because the values of θ_i vary widely. Provided that condition (6) holds, it is possible that the weights might be inversely proportional to per capita income in the areas. On the other hand, distribution of income (or wealth) within the recipient units (provinces or municipalities) could also be important, motivating alternative weights. Condition (7) rules out choice of a nonlinear power function for H. Thus, choosing more complicated component loss functions either leads to violation of Fisher-consistency or to component loss functions similar to (2).

3. Accuracy of Mid-Year Population Estimates for Provinces, 2002–2011

3.1. Overview and Motivation

The performance of mid-year estimates based on the 2001 Census and accounting for ten years of change can be assessed by comparing with the 2011 Census results. The error structure observed for the 2001–2011 period will be extrapolated to the 2011–2021 period (Section 4). As in other evaluations of population estimates to account for post-censal change, we find that the estimates under-predict growth or decline in shares of the population (Subsection 3.2). To estimate the variances of mid-year estimates that account for ten years of change, we analyze deviations in the average errors for provinces in which the relative share of the population was growing or shrinking. To model the variances for time spans less than ten years, we consider two models of year-to-year correlation between estimates of yearly population change, independence or correlation equal to 1.

Thus, in the absence of a 2016 Census, mid-year estimates will need to account for 6–10 years of population change since the 2011 Census. Evidence of their accuracy is derived from the analysis of accuracy of the mid-year population estimates produced using the 2001 Census as a base, as discussed in Subsection 3.2.

3.2. Biases of Estimates of Population

Denote the (mid-year) estimate of a province's population size t years after the census by $\hat{P}_t$, and denote the actual population size by P_t. Thus, P_0 denotes the population size at the time of the last census. Numerical values for P_0 and P_{10} are taken from prior census results (Stats SA 2012a, Table 2.1 for 2001 population and Table 2.9 for 2011 population), with undercount adjustments for both censuses (Stats SA 2012b, section 5). All censuses in South Africa (at least since the 1996) have had undercount adjustments based on data from a post-enumeration survey. Following a matching procedure to identify persons who should have been enumerated (and those that should not have been), the adjustments are predicted using Chi-square Automatic Interaction Detection, CHAID, technique using race, geographic category, sex, and age. These are applied to produce adjustment classes. Summing up the adjusted population across adjustment classes produces a separate ratio estimate of a total, from which the national adjusted population could be calculated. At the municipal level, the effect of adjustment will vary according to the share of different adjustment classes present in that municipality; see Stats SA (2012b, Section 5).

Figure 1 plots estimated percent change based on the mid-year population estimates for 2011, $(\hat{P}_t - P_0)/P_0$, against the observed percent change based on the 2011 Census adjusted for undercount, $(P_t - P_0)/P_0$. Note that small changes are overestimated and larger changes are underestimated.

Let ε_t denote the relative error in the estimate of population change t years since the last census,

$$\varepsilon_t = \frac{(\hat{P}_t - P_0) - (P_t - P_0)}{P_t - P_0} = \frac{\hat{P}_t - P_t}{P_t - P_0}.$$

Figure 2 plots ε_t versus the observed relative change in population. The relative errors are positive for relative changes below 15% and are negative for changes above 15%. Calculations based on the 2011 mid-year population estimates ($\hat{P}_t$), adjusted census counts

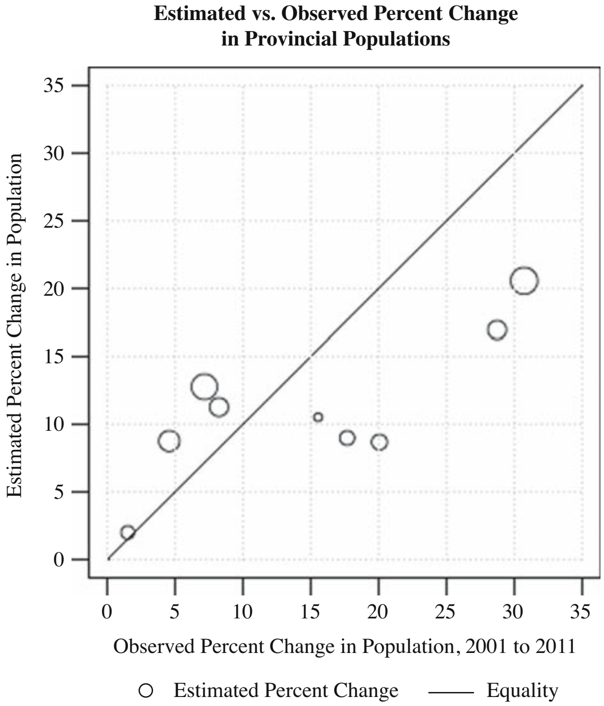


Fig. 1. Estimated versus observed change in province population size ten years after 2001 Census. Area of circle is proportional to average of 2001 and 2011 population sizes.

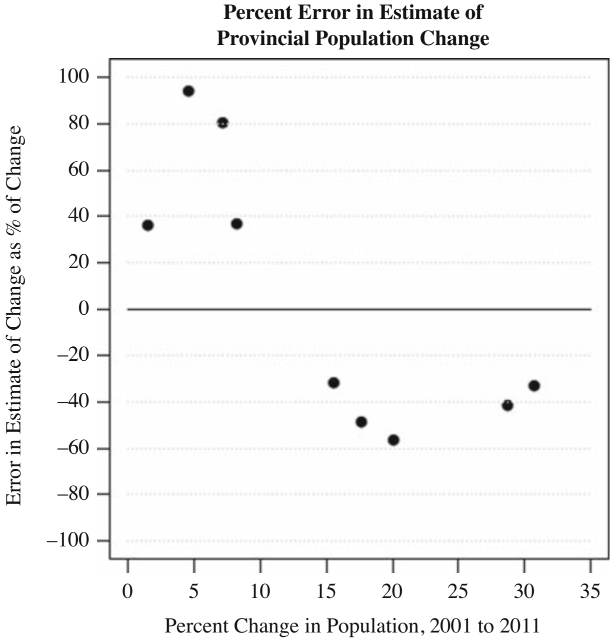


Fig. 2. Relative error of estimate of change versus observed percent change in province population size ten years after 2001 Census.

for 2011 (P_t), and the 2001 Census counts (P_0) for provinces show that the average value of ε_{10} is $+0.42$ for the five provinces that grew by more than 12% in size between 2001 and 2011 and is -0.62 for the four provinces that grew by less than 12% over that period. Given the small number of observations and the similarity in magnitudes, we decided to specify the same magnitude for provinces growing faster and slower than average, leading to the model that the expected value of ε_{10} is

$$E(\varepsilon_{10}) \approx 0.52 \times \operatorname{sgn}(dP_0 - P_{10}), \tag{8}$$

with $d = 1.12$ and $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = -1$ if $x < 0$, and $\operatorname{sgn}(x) = 0$ if $x = 0$.

Specifying the mean of ε_t for intermediate times $1 \leq t < 10$ requires some assumptions, since we have direct information only about ε_{10}. Denote the incremental error in the estimate of annual change by $\delta_t = (\hat{P}_t - \hat{P}_{t-1}) - (P_t - P_{t-1})$, with $\hat{P}_0 = P_0$ by assumption. It follows that

$$\hat{P}_t - P_t = (P_t - P_0)\varepsilon_t = \sum_{s=1}^{t} \delta_s.$$

Given the short time span, it is reasonable to use the simple approximation that the expected incremental error for a province is the same for each of the ten years, that is,

$$E(\delta_t) = (P_{10} - P_0)E(\varepsilon_{10})/10, \quad 1 \leq t \leq 10. \tag{9}$$

3.3. Variances of Estimates of Population

The average squared deviation regarding the mean for the relative errors observed for 2011 was 0.01067, leading to the model that the variance of ε_{10} is

$$V(\varepsilon_{10}) = 0.01067.$$

As in the case of the mean, specifying the variance of ε_t for $1 \leq t < 10$ requires assumptions, since we have direct information only about ε_{10}. A simple model for the variances of the incremental errors is that $V(\delta_s)$ does not change with s. If the incremental errors in a province are independent over time, then $V\left(\sum_{s=1}^{t} \delta_s\right)$ grows linearly with t, and hence $V(\delta_s) = 0.001067(P_{10} - P_0)^2$. On the other hand, if the incremental errors in a province are perfectly correlated over time, then $V\left(\sum_{s=1}^{t} \delta_s\right)$ is quadratic in t, and $V(\delta_s) = 0.0001067(P_{10} - P_0)^2$. We are assuming that the incremental errors in different provinces are mutually independent. To summarize, we have two alternative models for the variances of sums of incremental errors within provinces, the independent increments model

$$V\left(\sum_{s=1}^{t} \delta_s\right) = 0.001067t\,(P_{10} - P_0)^2, \tag{10}$$

and the dependent increments model,

$$V\left(\sum_{s=1}^{t} \delta_s\right) = 0.0001067t^2\,(P_{10} - P_0)^2. \tag{11}$$

3.4. Accuracy of Estimates of School-age Population of Provinces

The observed errors in mid-year estimates of ten-year change in the school-age population (i.e., persons aged 5–17) from 2001 to 2011 were all positive. The magnitudes were proportional to the error in the estimated ten-year change in the total province population, with different constants of proportionality for overestimates and underestimates of total population change. Estimates of those proportionality constants are 0.80 and −0.26, respectively. This means that the prediction of error in the school-age population estimate is 0.80 times the predicted error in the estimate of the total province population if the predicted error is positive. The prediction of error in the school-age population estimate is −0.26 times the predicted error in the estimate of total province population if the predicted error is negative. In both cases, the predicted error in the estimate of school-age population is positive.

4. Distributions of Mid-Year Population Estimates, 2017–2021

4.1. Overview

If no 2016 Census is carried out, population estimates for 2017–2021 must account for 6–10 years of change since the 2011 Census. If the 2016 Census is carried out, mid-year estimates for 2017–2021 will need to account for only 1–5 years of population change since the 2016 Census. Error distributions for the two sets of estimates are based on the analysis of Section 3. To specify the distributions of the estimates for 2017–2021, we add the errors to the specified true values of the population. The true values of the future population are developed in Subsection 4.2. Then, the distributions of estimates without a 2016 Census (Subsection 4.3) and with a 2016 Census (Subsection 4.4) are developed.

4.2. Specifications of True Population of Provinces, 2017—2021

To specify true values of total population and school-age population (ages 5–17) in provinces, we utilized projections of future population prepared in 2003 by the Actuarial Society of South Africa (ASSA). These projections are referred to as the ASSA2003 population projections, and they are prepared using the 2001 Census as a base (after adjustment for undercount). We assume that the true population sizes are unaffected by whether or not a 2016 Census is carried out. This is a nontrivial assumption since more accurate population data may lead to better provision of services, which can in turn influence fertility and mortality rates, as well as migration flows, as migrants seek access to better resourced areas that can provide better services. For example, in the case of the former, HIV/AIDS, low birth weight, and diarrheal diseases accounted for more than 60% of under age five deaths in South Africa at the time of the 2001 Census (Bradshaw et al. 2003). A range of primary health and basic service interventions has been found to have a direct impact on these causes (Bhutta et al. 2013). Many of these would be affected by inequalities arising from inaccurate population data and inadequate resource allocation to the authority responsible for their implementation (Say and Raine 2007). This includes vitamin A supplementation, the provision of Antiretroviral Therapy, the availability of healthcare workers, and the provision of adequate sanitation and protected water.

We use the ASSA projections in two alternative ways to specify true future population values. One specification is simply the total population as projected by the ASSA, and the other specification multiplies ASSA forecasts by the ratio of the undercount-adjusted 2011 Census figure for the province to the ASSA forecast for the 2011 population of the province. The latter "calibrated" population thus coincides with the undercount-adjusted census number for 2011. For school-age population (ages 5–17), one specification was derived from the ASSA2003 projections for five-year age groups, with population numbers disaggregated by single age based on the Sprague multiplier software on the Stats SA website. As with total population, a second specification was developed by ratio-adjusting (calibrating) the school-age population forecasts to agree with undercount-adjusted 2011 Census school-age population numbers. The two alternative sets of true values are denoted by the indicator k taking values 1 (uncalibrated) and 2 (calibrated).

4.3. Specifications of True Population of Municipalities, 2017–2021

The true values of total population for municipalities as used in the LGES can be taken to be the values for 2016, because no updating for post-censal population change is used in the LGES. Lacking ASSA projections of 2016 values for municipalities, we carried out a simple modeling of future values by extrapolating the 2001–2011 trends in the statistical inputs to the formula to 2016. This was subject to the constraint that the change from 2011 to 2016 could not exceed 50% of the 2011 total population size of the municipality.

4.4. Distribution in the Absence of a 2016 Census

For province estimates in the no-2016-census scenario, the variances of sums of incremental errors in mid-year population estimates are given by (10) or alternatively by (11). Using the independent increments assumption, we model the ten values $\delta_1, \ldots, \delta_{10}$ as independently normally distributed with means given by (8) and (9) with $d = 1$ and variances given by (10). Expression (10) can be evaluated because the modeling described in Subsection 4.2 specifies P_0 and P_{10}. Alternatively, using the dependent increments assumption, we model δ_1 as normally distributed, with means given by (8) and (9) and variance given by (11), and $\delta_{10} = \cdots = \delta_1$. The two alternative independence assumptions are denoted by the indicator l taking values 1 (independence) and 2 (perfect dependence).

The population estimate for province i, in year y, for dependence model l, corresponding to true value specification k (indicating uncalibrated or calibrated forecast), is denoted by $\hat{P}_{iykl}^{nocen16}$ when no 2016 Census is conducted and by $\hat{P}_{iykl}^{cen16}$ when a 2016 Census is conducted.

For municipalities, the error in the population estimate for municipality m in year $y, 2017 \leq y \leq 2021$, is equal to $P_{m2011} - P_{m2016}$ in the no-census scenario, since errors in census numbers are ignored.

4.5. Accuracy in the Presence of a 2016 Census

In a scenario with a 2016 Census, mid-year estimates for provinces for $2011 + t$, $6 \leq t \leq 10$ are based on the 2016 Census, and therefore account for only $t - 5$ years of population change. This is in contrast to the estimates in the no-2016-census scenario, which must account for the full t years of population change. Therefore, in a 2016 Census scenario, for each province the joint distribution of δ_t, $6 \leq t \leq 10$ is equal to the joint distribution of the corresponding values of δ_{t-5} in the no-2016 census scenario.

For municipalities, the error in the population estimate for municipality m in year y, $2017 \leq y \leq 2021$, is identically zero under the 2016 Census scenario, since errors in census numbers are ignored.

5. Distributions of PES and LGES Allocations

5.1. Hypothetical True Values

In the analysis, the true values of the allocations are allowed to change over time as the true population changes (Subsection 4.2). LGES allocations depend only on the population numbers for municipalities according to the latest census. PES allocations depend not only on population statistics, but on other statistics as well. To fully model the joint distribution of the various statistics and their underlying true values would have involved substantial additional work and would have added to the complexity of the analysis. Instead, our analysis conditions on (i.e., takes as fixed) the values of the nonpopulation statistics which served as inputs to the 2011 PES allocations.

The true allocation for province i, in year y, for true value specification k (indicating uncalibrated or calibrated forecast) is denoted by θ_{iyk}.

5.2. Specifying and Simulating Errors in PES Allocations

The errors in PES allocations are functions of population numbers only, because any nonpopulation statistics are held fixed. The joint distributions of the true and estimated allocations are determined by the joint distributions of the true and estimated populations. Recall that for the population of a province in a given year, in both the 2016 Census and no-census scenario, there are four alternative specifications, depending on whether or not the forecasts specifying the true values were calibrated and whether the estimates of year-to-year change are independent or perfectly dependent over time. For each of the eight specifications, we randomly generated four independent replications, which we denote by $r = 1, \ldots, 4$. These yielded four replications of population estimates $\hat{P}_{iyklr}^{cen16}$ and $\hat{P}_{iyklr}^{nocen16}$, respectively, in the case when a 2016 Census is and is not taken. (To increase the precision of estimated reduction in expected loss due to the 2016 Census, we set not only the distributions, but the realizations of δ_t, $6 \leq t \leq 10$, equal to realizations of δ_{t-5} in the no-2016 census scenario.) Each replication of population estimates leads to a replication of the allocation, $\hat{\theta}_{iyklr}^{cen16}$ and $\hat{\theta}_{iyklr}^{nocen16}$, respectively. The corresponding errors in allocation are $\hat{\theta}_{iyklr}^{cen16} - \theta_{iyk}$ and $\hat{\theta}_{iyklr}^{nocen16} - \theta_{iyk}$.

5.3. Specifying and Simulating Errors in LGES Allocations

As with the PES, the errors in LGES allocations are functions of population numbers only, because any nonpopulation statistics are held fixed. The joint distributions of the true and estimated allocations are determined by the joint distributions of the true and estimated populations. Recall that there are only two possible alternative estimates for the population of a municipality m in year y, P_{m2016} and P_{m2011}, corresponding to the 2016 Census scenario and the no-census scenario. The corresponding allocations to municipality m in year y are denoted by $\hat{\theta}_{my}^{cen16}$ and $\hat{\theta}_{my}^{nocen16}$.

6. Estimating Improvement in the Allocations As a Result of the 2016 Census

6.1. Estimating Reduction in Expected Loss from Errors in PES Allocations

The reduction in expected loss from errors in PES allocations when the 2016 Census is conducted is $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$, where the loss function is specified by (4). To estimate this reduction in expected loss, we use the scaling constant c times

$$\frac{1}{16}\sum_{y=2017}^{2021}\sum_{i=1}^{9}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{r=1}^{4}\left|\hat{\theta}_{pyklr}^{nocen16} - \theta_{pykl}\right| - \frac{1}{16}\sum_{y=2017}^{2021}\sum_{i=1}^{9}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{r=1}^{4}\left|\hat{\theta}_{pyklr}^{cen16} - \theta_{pykl}\right|. \quad (12)$$

Expression (12) shows the model averaging approach used to manage the different options for calculating the true population (calibrated or not) and the two variance options.

For practical considerations arising from tight decision deadlines, instead of computing the allocations for each year from 2017 to 2021, we computed the allocations just for 2021 for both $\hat{\theta}$ and θ, and we used those values for each year. This likely led to a modest overstatement of the reduction in expected loss due to the 2016 Census, since the accuracy of the mid-year population estimates is at its lowest in 2021. The calculated value of (12) is ZAR 4.8 billion.

One technical point is worth noting. By ignoring error in any nonpopulation statistics in the allocation formulas, we are, in effect, approximating $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$ by $E[\ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{cen16})]$, where $E[\cdot]$ denotes expectation and $\boldsymbol{\theta}'$ denotes the array of allocations when the population statistics have no error, but the other statistics are observed with possible error. Research in progress suggests that the approximation either overstates or only modestly understates the reduction in expected loss.

6.2. *Estimating Reduction in Expected Loss from Errors in LGES Allocations*

Recall that the LGES allocations for 2017–2021 will be based on the 2011 Census, if the 2016 Census is not conducted, and on the 2016 Census if it is conducted. As was the case for the PES, we approximate $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$ by $E[\ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{cen16})]$, where $E[\cdot]$ denotes expectation and $\boldsymbol{\theta}'$ denotes the array of allocations to municipalities when the population statistics have no error, but the other statistics are observed with possible error. By construction, $\boldsymbol{\theta}' = \hat{\boldsymbol{\theta}}^{cen16}$ and so we estimate the reduction in expected loss by scaling constant c times

$$5\sum_{m=1}^{278}\left|\hat{\theta}^{nocen}_{m2016} - \hat{\theta}^{cen}_{m2016}\right|, \tag{13}$$

where m indexes the 278 municipalities and the allocations are calculated for 2016. The calculated value of (13) is ZAR 32.1 billion. As the LGES is assumed to allocate only 1/15 as much money as the PES program over the five-year period, ZAR 38.9 billion for the LGES versus ZAR 600 billion for the PES, it is surprising that (13) is more than six times as large as (12). The explanation is much larger differences in LGES allocations in the presence or absence of the 2016 Census. Even though mid-year estimates do not estimate population change accurately, PES allocations are based on the total population level. Mid-year estimates predict population levels much more accurately than population change, whereas in the LGES, the municipal estimates of population levels are not updated at all in the absence of a 2016 Census.

7. Limitations

Several limitations to the analysis of reduction in PES and LGES misallocations arising from the 2016 Census may be noted.

1. The specifications for true values for province populations, which depend on the ASSA projections, are inaccurate to an unknown degree.
2. The true values of population for the LGES allocation are taken to be for 2016 rather than the true population sizes for 2017–2021.
3. We are ignoring the effects of errors in nonpopulation statistics that are used to calculate PES and LGES allocations. This may well increase the estimated magnitude of improvement in allocations in conducting the 2016 Census.

4. Distribution of error in mid-year population estimates 2017–2021 could be different than in last decade, due either to changes in patterns of population growth or decline or to differences in quality of data used to estimate births, deaths, and net migration among provinces.
5. Errors in 2011 Census numbers used in the analysis can cause errors in estimates of error in mid-year population estimates for 2011 (Spencer 1980b). Our analysis ignores possible error in the 2011 Census numbers.
6. Hold-harmless provisions in the allocation formulas were not taken into account.

The effects of the limitations noted in points 1, 2, and 5 might be slightly reduced by use of a prior distribution to specify uncertainty about true values, as part of a full Bayesian decision theoretic analysis. However, it is unlikely that this will greatly change the estimates of expected loss.

8. 2016 Census Cost

The costs of census-taking include the investment cost, the amount spent on the collection, capture, cleaning, and data assurance (quality control). Other costs that are often forgotten include data curatorship, which refers to looking after, updating and maintaining the data and the ongoing assistance provided to the users of this data. Finally, dissemination and publicity also carry costs. Nonetheless, for a standard cost-benefit analysis, estimating the direct costs of a Census is relatively straight-forward, to the extent that reliable and up-to-date expenditure data are available from the appropriate government departments. Some indirect costs, such as calculating cost of the time taken by respondents to complete a census questionnaire, may be more complex, but can be estimated using an appropriate shadow wage rate. This has not been undertaken for this study.

In the absence of a 2016 Census, it is assumed that some variation of the 2007 Community Survey would be conducted, and it is assumed that the cost of the mid-year estimates program is essentially unchanged regardless of whether the census or the Community Survey is taken in 2016. The net additional cost of the 2016 Census (over and above the 2016 Community Survey) was predicted to be on the order of ZAR 3 billion.

9. Measuring Benefit from Improvement in Allocations

The measures of reduction in absolute values of misallocations, such as (12) and (13) should not be interpreted directly as measures of benefit. In monetary terms, the sum of the overallocations equals the sum of the underallocations, or equivalently, one area's loss is another area's gain. As discussed in Subsection 2.3, the benefit arises from reduction of inequity of the allocations. The translation from (12) and (13) to benefit, or reduction of expected loss, is achieved through the scaling constant c in the loss function (4). The scaling constant c should reflect the sensitivity of society or the decision-makers to misallocations. Logically, the value of c should not be as large as 1, as in the cautionary example of *Jarndyce v. Jarndyce* (Dickens 1985). If overallocations are viewed as beneficial or benign for the local governments that receive them, then $c \leq 0.5$, as noted in Subsection 2.1. Ultimately, however, the magnitude of c depends on the decision-maker's preferences regarding tradeoffs for equitable allocations versus spending money to achieve the equitable allocations. If it is just worth spending ZAR ten million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.01$. If it is just worth spending ZAR 100 million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.10$, and if it is just worth spending ZAR 500 million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.50$.

We believe that the specification of c is inherently subjective and should be openly addressed. People's values are not objectively determined, and the choice of the scaling constant c involves a question of values – how much is it worth spending to achieve more equitable allocations. Our analysis has drawn on technical analyses to compute the expected loss as parameterized by the scaling constant c. However, the specific choice for c reflects the willingness of the decision-makers to use tax dollars to reduce inequity in fund allocations. Having a single, easily interpretable parameter for social values conveys the additional advantage of providing transparency to the analysis.

Spencer (1980a) suggested that, $c = 0.01$ in the 1970s context of General Revenue Sharing in the United States The rationale, as discussed in Subsection 2.3 above, was that if $x = \hat{\theta}_i - \theta_i$ and $x < 0$, then local government i incurs a deficit of $|x|$, and if $x > 0$ then it incurs a surplus of $|x|$. If a deficit of $|x|$ is incurred, local government is assumed to borrow an amount equal to the shortfall, to be repaid in the next fiscal period. If the interest rate for the period was $a - 1$, the monetary loss to the local government would be $a|x|$. We neglected long-term effects because they are hard to trace and because the local government cannot make adjustments for the deficit before the end of the current period, but it can make adjustments after the period. Conversely, if $x > 0$, so that a surplus is produced, the local government invests $|x|$ for the period at interest rate $b - 1$. The local government's monetary loss incurred is $-b|x|$, a negative loss (i.e., a gain). From (3), $c = (a - b)/2$, and so we may interpret c as half of the difference between the local governments' interest rates for investing versus borrowing for the period. Spencer (1980a) took the period to be one year and the difference between interest rates to be 0.02, leading to a specification that $c = 0.01$. In this scenario, the choice of c would reflect economic conditions and the length of the period that the local government would need to adjust for the shortfall.

Table 1 shows the expected improvement in allocation when a 2016 Census is conducted, for various values of c.

To illustrate, if $c = 0.06$ is a reflection of the preference tradeoff between PES and LGES equity on the one hand and expenditure on the other, the benefit, in terms or more equitable allocation of funds, is ZAR 2.2 billion. If a 2016 Census will cost an additional ZAR three billion (beyond the cost of a 2016 Community Survey), then the improvement in allocation of funds justifies about three quarters of the census cost. Other uses of the data would need to justify the remaining ZAR 800 million of the census cost. If $c > 0.08$, then the benefit of improvement in allocation of funds equals or exceeds the census cost, in which case the analysis would provide strong support for a 2016 Census.

10. Discussion

As mentioned at the outset, the decision to fund a Census in 2016 is not only dependent upon the costs and anticipated financial benefits involved, and the South African

Table 1. Effects of 2016 Census on Improvement of Allocation of Funds to Provinces and Municipalities, 2017–2021, with Alternative Levels of Scaling.

Expected reduction from misallocations when a 2016 Census is carried out (ZAR millions)											
	Scaling constant, c										
	1.00	0.50	0.30	0.20	0.15	0.10	0.08	0.06	0.04	0.02	0.01
PES	4,800	2,400	1,440	960	720	480	384	288	192	96	48
LGES	32,121	16,061	9,636	6,424	4,818	3,212	2,570	1,927	1,285	642	321
Total	36,921	18,461	11,076	7,384	5,538	3,692	2,954	2,215	1,477	738	369

Government made the decision not to undertake a census in 2016 (Stats SA 2014a, 21). Instead, the Government decided to improve its data collection program. An enlarged Community Survey, with a sample size increased from 300,000 households to one million households, is being undertaken in 2016 and is projected to cover all enumerator areas in the country (Parliament of the Republic of South Africa 2014, 3424). Furthermore, the agency has focused on improving civil registration of vital statistics to be able to better estimate the mid-year population (Stats SA 2014a, 59). Further considerations include a long-term strategy to introduce a continuous population survey that will collect population and other social statistics on an ongoing basis. The methodology described above permitted this decision to be evidence-based, up to the subjective specification of the parameter *c*, and to confront the possible effects of error (Stats SA, 2014b). Indeed, the impact of prior error resulting from the ten-year gap between 2001 and 2011 has been taken into account in South Africa's most recent government budget. The Annex to the Budget notes that by not properly accounting for migration, the division of revenue between provinces has become inequitable, with receiving provinces such as Gauteng and the Western Cape being allocated less resources than would have been provided with accurate data. However, as the National Treasury (2015:17–18) acknowledges, provinces which have been receiving more resources need time to adjust to revised allocations, and a total ZAR 4.2 billion has had to be added to the PES over the three years from 2013 to 2015 to cushion the impact of the census data. The results of this partial cost-benefit analysis of South African census-taking contributed to greater awareness of the role played by official statistics in the allocation of resources, greater awareness of the wider costs of error, and of assessing the 'value for money' of official statistics. The decision to triple the size of the Community Survey in 2016 and the introduction of methodological improvements by Stats SA to improve cost effectiveness are examples of ongoing reflection concerning official statistics in South Africa and elsewhere (Stats SA 2016). The cost-benefit approach used in this article is applicable to other data programs as well, such as improvements in sample surveys and vital registration statistics, provided uses of the statistics are sufficiently understood.

11. References

Alho, J.M. and B.D. Spencer. 2005. *Statistical Demography and Forecasting*. New York: Springer.

Bakker, C. 2014. *Valuing the Census*. Wellington: Statistics New Zealand. Available at: www.stats.govt.nz.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. *2nd ed.* New York: Springer.

Bhutta, Z.A., J.K. Das, A. Rizvi, M.F. Gaff, N. Walker, S. Horton, P. Webb, A. Lartey, and R.E. Black. 2013. "Evidence-based Interventions for Improvement of Maternal and Child Nutrition: What Can Be Done and at What Cost?" *Lancet* 382: 452–477.

Bradshaw, D., D. Bourne, and N. Nannan. 2003. "What Are the Leading Causes of Death among South African Children?" MRC Policy Brief, No 3., Medical Research Council, Bellville.

Buehler, J.W. and D.R. Holtgrave. 2007. “Challenges in Defining an Optimal Approach to Formula-Based Allocations of Public Health Funds in the United States.” *BMC Public Health* 7: 44. Doi: http://dx.doi.org/10.1186/1471-2458-7-44.

Cope, I. 2015. The Value of Census Statistics in England and Wales. Note by the Office for National Statistics, United Kingdom. United Nations Economic and Social Council, 4 September 2015. Report ECE/CES/GE.41/2015/16. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1/CES_GE.41_2015_16_-_UK.pdf (accessed 25 March 2016).

Dickens, C. 1985. *Bleak House.* 1852–53. New York: Penguin.

Downes, T.A. and T.E. Pogue. 2002. “How Best to Hand Out Money: Issues in the Design and Structure of Intergovernmental Aid Formulas.” *Journal of Official Statistics* 18: 329–352.

General Register Office for Scotland. 2006. 2011 Census Business Case. Prepared by John Aldridge, Consultant, July 2006. Available at: https://www.whatdotheyknow.com/request/8345/response/20302/attach/3/business%20case.pdf (accessed 8 March 2016).

Kruskal, W.H. 1984. “The Census as a National Ceremony.” In *Federal Statistics and National Needs* prepared for the Subcommittee on Energy, Nuclear Proliferation and Government Processes, an arm of the Committee on Government Affairs of the United State Senate, by the Congressional Research Service of the Library of Congress, 177–180. Washington DC: U.S. Government Printing Office.

Manski, C.F. 2011. “Actualist Rationality.” *Theory and Decision* 71: 195–210.

May, J., M. Dimbabo, J. Tamri, G. Wright, Z. Seeskin, and B.D. Spencer. 2013. *Cost Benefit Analysis of South Africa’s Population Census, Final Report*, 21 May, 2013. Bellville, South Africa: Institute for Social Development, University of the Western Cape.

May, J. and P. Lehohla. 2005. “Counting the Costs of a 21st Century Census: South Africa’s Census 2001.” *Development Southern Africa* 22: 215–232.

McCaa, R., A. Esteve, S. Ruggles, and M. Sobek. 2006. “Using Integrated Census Microdata for Evidence-Based Policy Making: The IPUMS-International Global Initiative.” *The African Statistical Journal* 2: 83–100.

Melnick, D. 2002. “The Legislative Process and the use of Indicators in Formula Allocations.” *Journal of Official Statistics* 18: 353–370.

Ministry for Welfare and Population Development. 1998. *White Paper on Population Policy*. 7 September 1998. Pretoria: *Government Gazette*.

National Research Council. 1980. *Estimating Population and Income of Small Areas*. Panel on Small-Area Estimates of Population and Income, Committee on National Statistics. Washington DC: The National Academies Press.

National Research Council. 1985. *Natural Gas Data Needs in a Changing Regulatory Environment*. Panel on Statistics on Natural Gas, Committee on National Statistics. Washington DC: The National Academies Press.

National Research Council. 2003. *Statistical Issues in Allocating Funds by Formula.* Panel on Formula Allocations, edited by T.A. Louis, T.B. Jabine, and M.A. Gerstein. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Treasury. 2013. Annexure W1 to the Budget Review: Explanatory Memorandum to the Division of Revenue, National Treasury, Pretoria. Available at: http://www.treasury.gov.za/documents/national%20budget/2013/review/Annexure%20W1.pdf (accessed 1 August 2016).

National Treasury. 2015. Website Annexure to the 2015 Budget Review: Explanatory Memorandum to the Division of Revenue, National Treasury, Pretoria. Available at: http://www.treasury.gov.za/documents/national%20budget/2015/review/Annexure%20W1.pdf (accessed 1 March 2015).

Nigerian Tribune. 2016. "Towards a Credible Census." Available at: http://tribuneonlineng.com/towards-credible-census/ (accessed 16 January 2017).

Parliament of the Republic of South Africa. 2014. Announcements, Tablings and Committee Reports No. 95–2014 [First Session, Fifth Parliament, 19 November 2014], 3402, Cape Town. Available at: http://www.parliament.gov.za/live/commonrepository/Processed/20141124/593735_1.pdf (accessed 2 March 2015).

Parliament of the United Kingdom. 2009. Draft Census (England and Wales) Order 2009 etc – Merits of Statutory Instruments Committee. Available at: http://www.publications.parliament.uk/pa/ld200809/ldselect/ldmerit/176/17606.htm (accessed 8 March 2016).

Redfern, P. 1974. "The Different Roles of Population Censuses and Interview Surveys, Particularly in the U.K. Context." *International Statistics Review* 42: 131–146.

Savage, I.R. 1985. "Hard-Soft Problems." *Journal of the American Statistical Association* 80: 1–7.

Say, L. and R. Raine. 2007. "A Systematic Review of Inequalities in the Use of Maternal Health Care in Developing Countries: Examining the Scale of the Problem and the Importance of Context." *Bulletin of the World Health Organization* 85: 812–817.

Schirm, A.L., A.M. Zaslavsky, and J.L. Czajka. 1999. "Large Numbers of Estimates for Small Areas." *Proceedings of the 1999 FCSM Research Conference*. Available at: https://fcsm.sites.usa.gov/files/2014/05/IV-A_Schirm_FCSM1999.pdf (accessed 7 March 2016).

Seeskin, Z.H. and B.D. Spencer. 2015. "Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds." Institute for Policy Research Working Paper WP- 15-05. Evanston, IL: Northwestern University. Available at: http://www.ipr.northwestern.edu/publications/papers/2015/ipr-wp-15-05.html (accessed 8 March 2016).

Sims, C.A. 1984. "Can We Measure the Benefits of Data Programs?" In *Proceedings of the Social Statistics Section: American Statistical Association*, 60–67. Washington, D.C.: American Statistical Association.

Spencer, B.D. 1980a. *Benefit-Cost Analysis of Data Used to Allocate Funds*. New York: Springer.

Spencer, B.D. 1980b. "Effects of Biases in Census Estimates on Evaluation of Postcensal Estimates." In National Research Council, 1980, *Estimating Population and Income of Small Areas*. Panel on Small-Area Estimates of Population and Income: 232–6. Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D.C.: The National Academy Press.

Spencer, B.D. 1980c. "Implications of Equity and Accuracy for Undercount Adjustment: A Decision- Theoretic Approach." In U.S. Bureau of the Census, *Conference on Census*

Undercount: Proceedings of the 1980 Conference: 204–216. Washington, D.C. U.S. Department of Commerce.

Spencer, B.D. 1982a. "Feasibility of Benefit-Cost Analysis of Data Programs." *Evaluation Review* 6: 649–672.

Spencer, B.D. 1982b. "Technical Issues in Allocation Formula Design." *Public Administration Review* 4: 524–529.

Spencer, B.D. 1985. "Statistical Aspects of Equitable Apportionment." *Journal of the American Statistical Association* 80: 815–822.

Stats SA. 2011. *Mid-Year Population Estimates, 2011*. Report P0302. Pretoria: Statistics South Africa.

Stats SA. 2012a. *Census 2011, Census in Brief*. Report 03-01-41. Pretoria: Statistics South Africa.

Stats SA. 2012b. CENSUS 2011: Post Enumeration Survey, Report 03-01-46. Pretoria: Statistics South Africa.

Stats SA. 2014a. Annual Report 2013/2014 (Book 1). Pretoria: Statistics South Africa. Available at: http://www.gov.za/sites/www.gov.za/files/STATSAnnual_Report_2013-2014.pdf (accessed 2 March 2015).

Stats SA. 2014b. "Population Household Surveys in the Case of South Africa." Presentation at the Workshop on Strengthening the Collection and Use of International Migration Data for Development, 18–21 November 2014, Addis Ababa, Ethiopia. Available at: http://www.un.org/en/development/desa/population/migration/events/other/workshop/docs/Session%20VI%20South%20Africa.pdf (accessed 3 March 2015).

Stats SA. 2016. Community Survey 2016: Technical Report 03-01-01, Statistics South Africa, Pretoria. Available at: http://cs2016.statssa.gov.za/wp-content/uploads/2016/06/CS-2016-Technical-report_Web.pdf (accessed 1 August 2016).

The Globe and Mail. 2011. "Traditional Census the Only Option for 2016, Statistics Canada Says." Available at: http://www.theglobeandmail.com/news/politics/traditional-census-the-only-option-for-2016-statistics-canada-says/article556243/ (accessed 4 March 2015).

The Guardian. 2015. "Census in Doubt as 10-year Data Collection Is Considered." Available at: http://www.theguardian.com/world/2015/feb/19/census-in-2016-in-doubt-as-10-year-data-collection-considered (accessed 4 March 2015).

The Journal. 2012. "2016 Census May Be Delayed in Government Spending Review." Available at: http://www.thejournal.ie/2016-census-may-be-delayed-in-government-spending-review-455663-May2012/ (accessed 4 March 2015).

Yemek, E. 2005. *Understanding Fiscal Decentralisation in South Africa*. IDASA Budget Information Service Occasional Paper. Cape Town: Institute for Democratic Alternatives in South Africa. Available at: http://www.gsdrc.org/docs/open/CC107.pdf (accessed 1 March 2015).

Zaslavsky, A.M. and A.L. Schirm. 2002. "Interactions between Survey Estimates and Federal Funding Formulas." *Journal of Official Statistics* 18: 371–391.

Received March 2015
Revised August 2016
Accepted September 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 275–299, http://dx.doi.org/10.1515/JOS-2017-0014

Estimation when the Covariance Structure of the Variable of Interest is Positive Definite

Alain Théberge[1]

Generalized regression (GREG) estimation uses a model that assumes that the values of the variable of interest are not correlated. An extension of the GREG estimator to the case where the vector of interest has a positive definite covariance structure is presented in this article. This extension can be translated to the calibration estimators. The key to this extension lies in a generalization of the Horvitz-Thompson estimator which, in some sense, also assumes that the values of the variable of interest are not correlated. The Godambe-Joshi lower bound is another result which assumes a model with no correlation. This is also generalized to a vector of interest with a positive definite covariance structure, and it is shown that the generalized calibration estimator asymptotically attains this generalized lower bound. Properties of the new estimators are given, and they are compared with the Horvitz-Thompson estimator and the usual calibration estimator. The new estimators are applied to the Canadian Reverse Record Check survey and to the problem of variance estimation.

Key words: Asymptotic setup; calibration estimators; Godambe-Joshi lower bound; Horvitz-Thompson estimator; Moore-Penrose inverse.

1. Introduction

Let s be a sample drawn from a population of size N according to a sampling plan p, let $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ be a vector of interest, and let $\mathbf{c} = (c_1, c_2, \ldots, c_N)'$ be a vector of known constants. The parameter to estimate is $\theta = \mathbf{y}'\mathbf{c}$. A commonly used estimator is that of Horvitz and Thompson (1952). This estimator can be written $\hat{\theta}_{HT} = \mathbf{y}'\mathbf{W}_{s\,HT}\mathbf{c}$, where $\mathbf{W}_{s\,HT} = \mathbf{\Delta}_s(E(\mathbf{\Delta}_s))^{-1}$ with $\mathbf{\Delta}_s \in \mathbb{R}^{N \times N}$ the diagonal matrix of the δ_k, $k = 1, 2, \ldots, N$, with δ_k equal to 1 if unit $k \in s$ and 0 otherwise (it is assumed that $E(\delta_k) = \pi_k > 0,\ k = 1, 2, \ldots, N$). The weight matrix $\mathbf{W}_{s\,HT}$ is diagonal. Even in the absence of auxiliary data, other useful estimators exist. An estimator of the form $\mathbf{y}'\mathbf{W}_s\mathbf{c}$, where $\mathbf{W}_s$ is not necessarily diagonal, will be proposed. An unbiased estimator is wanted, thus $E(\mathbf{W}_s) = \mathbf{I}_N$ will be required, where $\mathbf{I}_N$ is the identity matrix of order N. Not requiring the weight matrix $\mathbf{W}_s$ to be diagonal could prove useful if the variance matrix of $\mathbf{y}$ is not diagonal. For example, from the frame, it could be known that units 1 and 2 are twins, that $y_1 = y_2 = y_{twin}$, without knowing the value y_{twin}. Noting $\pi_{kl} = E(\delta_k\delta_l)$, an alternative to the Horvitz-Thompson estimator $\sum_{k=1}^{N} \frac{\delta_k y_k}{\pi_k}$ for the population total, is the unbiased estimator $\frac{2y_{twin}(\delta_1+\delta_2-\delta_1\delta_2)}{\pi_1+\pi_2-\pi_{12}} + \sum_{k=3}^{N} \frac{\delta_k y_k}{\pi_k}$. That is, if either unit 1 or unit 2 are sampled, the

[1] 1221 du Gouverneur, Ottawa, Ontario K4C 1E2, Canada. Email: alain.theberge1@gmail.com

Acknowledgments: The author wishes to thank the Associate Editor and four referees for their valuable comments and suggestions that helped to improve the quality of the article significantly.

value $2y_{twin}$ is given a weight equal to the inverse of the probability of selecting either of the two units. The number of twins in the sample being random, this will add to the variance of this estimator. However, if both this new estimator and the Horvitz-Thompson estimator are calibrated so that the sum of their weights equals the population size, then the calibrated new estimator is superior to the similarly calibrated Horvitz-Thompson estimator, because it makes use of the information that units 1 and 2 are twins by acknowledging that observing one of the two units is equivalent to observing both. This article will suggest estimators that can improve on the Horvitz-Thompson estimator if some of the ys are simply correlated, without necessarily being equal. For example, because of the increased risk of transmission, the incidence of the flu in two individuals from the same household are two correlated events. Depending on the variable of interest, other examples may occur for workers clustered by establishment.

In the next section, what is meant by "the variance matrix of $\mathbf{y}$" is made more precise through the asymptotic setup. With the help of the Moore-Penrose inverse, a generalization of the Horvitz-Thompson estimator, $\hat{\theta}_{GHT}$, is presented in Section 3. The generalized Horvitz-Thompson estimator will depend on an estimate of the variance matrix of $\mathbf{y}$ and will reduce to the usual Horvitz-Thompson estimator when that variance matrix estimate is diagonal. Godambe and Joshi (1965) gave a lower bound applicable to unbiased estimators under the assumption that the variance matrix of $\mathbf{y}$ is diagonal. In Section 4, their result is also generalized to a positive definite variance matrix. In Section 5, the calibration problem, as stated in Deville and Särndal (1992) and in Théberge (1999), is generalized; the desired weights should be close to those of $\hat{\theta}_{GHT}$ rather than those of $\hat{\theta}_{HT}$. The solution to that problem will lead to generalized calibration estimators. Generalized calibration estimators are shown to be optimal in the sense that they asymptotically attain the generalized Godambe-Joshi lower bound. In Section 6, the problem of computing the weights of the generalized estimators is examined with an example where the variance matrix of $\mathbf{y}$ is block diagonal. Modified versions of the generalized estimators are described in Section 7. The new estimators are compared to that of Horvitz-Thompson and to the calibration estimator in Section 8. Applications to the Canadian Reverse Record Check Survey and to the problem of variance estimation are given in Sections 9 and 10 respectively. Finally, concluding remarks are found in Section 11.

2. Asymptotic Setup

In order to discuss the large sample properties of an estimator $\hat{\theta}$, an asymptotic setup will be needed. Such setups have been described by Brewer (1979) and by Isaki and Fuller (1982). The setup shall serve two main purposes: (1) to establish a link between the setup and the variance matrix of the variable of interest, (2) to establish three results that will be useful for deriving the asymptotic properties of $\hat{\theta}_{GHT}$ and calibrated estimators. The setup described here is one that serves those purposes.

Given an auxiliary information matrix $\mathbf{X} \in \mathbb{R}^{N \times q}$ assumed to be of full rank, a sequence of increasingly large populations and samples is generated with the help of an N-dimensional distribution ξ of mean $\mathbf{X\beta}$, $\boldsymbol{\beta} \in \mathbb{R}^q$, and variance $\boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma}$ positive definite, and the sampling plan p. The sequence starts with the original population and the original sample. The tth population, of size tN, is obtained by adding N units to the $(t-1)$th population. With respect to the auxiliary information, those added units are identical to the original population. The vector of interest of the added units is generated with the distribution ξ. From the added units, a sample of units is selected using the plan p, and together with the units of the $(t-1)$th sample, they will form the tth sample of expected size $n_t = tn$, where n is the expected size of the original sample.

More precisely, if $\mathbf{1}_{a\times b} \in \mathbb{R}^{a\times b}$ is a matrix of ones and $\mathbf{I}_a \in \mathbb{R}^{a\times a}$ is the identity matrix of dimension a, then define $\mathbf{X}_t \in \mathbb{R}^{tN\times q}$ equal to $\mathbf{1}_{t\times 1} \otimes \mathbf{X}$, the auxiliary information matrix of the tth population. Set $\mathbf{c}_t = t^{-1}(\mathbf{1}_{t\times 1} \otimes \mathbf{c})$ for estimating a mean, that is, if $\mathbf{c} = N^{-1}\mathbf{1}_{N\times 1}$, but set $\mathbf{c}_t = (\mathbf{1}_{t\times 1} \otimes \mathbf{c})$ for estimating a total, that is, if $\mathbf{c} = \mathbf{1}_{N\times 1}$. More generally, set $\mathbf{c}_t = t^{\gamma-1}(\mathbf{1}_{t\times 1} \otimes \mathbf{c})$ if $\theta = \mathbf{y}'\mathbf{c} = O_p(N^{\gamma})$. Define $\mathbf{y}_t = \left(\mathbf{y}'_{[1]}\ \mathbf{y}'_{[2]} . . .\mathbf{y}'_{[t]}\right)'$ and $\theta_t = \mathbf{y}'_t\mathbf{c}_t$, where the subscript $[i]$ is used to denote the N units that belong to population i, but not to population $(i-1)$, $\boldsymbol{\Delta}_t = \begin{pmatrix} \boldsymbol{\Delta}_{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\Delta}_{[t]} \end{pmatrix}$ is the diagonal matrix of the δ_k, where δ_k is equal to 1 if unit k is sampled and 0 if not (to ease the notation, in this section, the subscript s denoting the sample will be omitted from $\boldsymbol{\Delta}$). With this setup, $E_\xi(\mathbf{y}_t) = \mathbf{X}_t\boldsymbol{\beta}$ and $V_\xi(\mathbf{y}_t) = \mathbf{I}_t \otimes \boldsymbol{\Sigma}$.

Before presenting the asymptotic results of this section, for any matrix $\mathbf{F}$, let $\mathbf{F}^{\dagger}$ denote the Moore-Penrose inverse, and if $\mathbf{F}$ is positive definite, then define $\mathbf{Q}_{\mathbf{F}} = (E_p((\boldsymbol{\Delta}\mathbf{F}\boldsymbol{\Delta})^{\dagger}))^{-1}$, where $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{[1]}$. It will be shown in the following section that $\mathbf{Q}_{\mathbf{F}}$ is well defined if and only if $\pi_k > 0,\ \ k = 1, 2, . . . , N$.

Let $\mathbf{T} \in \mathbb{R}^{q\times q}$ and $\mathbf{U} \in \mathbb{R}^{N\times N}$ be symmetric positive definite matrices, $\mathbf{U}_t = \mathbf{I}_t \otimes \mathbf{U}$ and

$$\hat{\boldsymbol{\beta}}_t = \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'_t(\boldsymbol{\Delta}_t\mathbf{U}_t\boldsymbol{\Delta}_t)^{\dagger}\mathbf{X}_t\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'_t(\boldsymbol{\Delta}_t\mathbf{U}_t\boldsymbol{\Delta}_t)^{\dagger}\mathbf{y}_t. \tag{1}$$

With this asymptotic setup, the following three results hold.

RESULT 1. If the sampling plan is noninformative (see, for example Cassel et al. 1977), then $\hat{\boldsymbol{\beta}}_t \to \boldsymbol{\beta}$ in probability.

The next two results apply to a positive definite estimator, $\hat{\boldsymbol{\Sigma}}$, of $V_\xi(\mathbf{y}) = \boldsymbol{\Sigma}$. One must first define a block diagonal matrix $\hat{\boldsymbol{\Sigma}}_t = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\boldsymbol{\Sigma}}_{[t]} \end{pmatrix}$ where $\hat{\boldsymbol{\Sigma}}_{[i]}$ is the estimator of $\boldsymbol{\Sigma}$ based on the sample represented by $\boldsymbol{\Delta}_{[i]}$ and then define $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t} = \left(E_p\left(\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\right)\right)^{-1}$.

RESULT 2. For a positive definite estimator $\hat{\boldsymbol{\Sigma}}$, the difference $\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t\mathbf{c}_t$ is $O_p(t^{\gamma-1/2})$.

RESULT 3. For a positive definite estimator $\hat{\boldsymbol{\Sigma}}$ such that $\hat{\boldsymbol{\Sigma}} \to \boldsymbol{\Sigma}$ in probability, the difference $\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t(\boldsymbol{\Delta}_t\boldsymbol{\Sigma}_t\boldsymbol{\Delta}_t)^{\dagger}\mathbf{Q}_{\boldsymbol{\Sigma}_t}\mathbf{c}_t$ is $o_p(t^{\gamma-1/2})$.

Since $n_t = tn$ and $N_t = tN$, what, for example, is $O_p(t^{\gamma-1/2})$, is also $O_p\left(n_t^{\gamma-1/2}\right)$ and $O_p\left(N_t^{\gamma-1/2}\right)$. The proofs of these three results can be found in Appendix A. This asymptotic setup incorporates the superpopulation model ξ; a separate superpopulation model is not needed. This avoids possible inconsistencies between the asymptotic setup's model and that of a superpopulation.

3. A Generalization of the Horvitz-Thompson Estimator

In this section and the next section, an auxiliary data matrix is not needed, or at least, it need not be known. Only a positive definite estimate, $\hat{\mathbf{\Sigma}}$, of the variance matrix implied by the setup, $V_\xi(\mathbf{y}) = \mathbf{\Sigma}$, will be needed. In the absence of auxiliary data, the Horvitz-Thompson estimator, $\hat{\theta}_{HT} = \mathbf{y}'\mathbf{\Delta}_s(E_p(\mathbf{\Delta}_s))^{-1}\mathbf{c}$, is an estimator that is often used. In order to generalize the Horvitz-Thompson estimator, it will first be proven that $E_p\left(\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\right)$ is nonsingular if and only if $\pi_k > 0,\ \ k = 1, 2, \ldots, N$, where $\mathbf{F}^\dagger$ denotes the Moore-Penrose inverse of the matrix $\mathbf{F}$.

LEMMA 1. *If $\mathbf{F}_i \in \mathbb{R}^{N\times N}$ $i = 1, 2, \ldots, K$ are symmetric positive semi-definite matrices, then the null space of $\mathbf{F} = \sum_{i=1}^K \mathbf{F}_i$, noted $\mathcal{N}(\mathbf{F})$, equals $\cap_{i=1}^K \mathcal{N}(\mathbf{F}_i)$.*

The proof of Lemma 1 is given in Appendix A. From Ben-Israel and Greville (2002, Exercise 2.38), $\mathcal{N}\left(\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\right) = \mathcal{N}\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right) = \mathcal{N}(\mathbf{\Delta}_s)$. Applying Lemma 1, the matrix sum $E_p\left(\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\right)$ is invertible if and only if $\bigcap_{\text{all samples } s} \mathcal{N}(\mathbf{\Delta}_s) = \mathbf{0}$, that is, $\pi_k > 0,\ \ k = 1, 2, \ldots, N$.

Note $\mathbf{Q}_{\hat{\mathbf{\Sigma}}} = \left(E_p\left(\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\right)\right)^{-1}$ and define

$$\begin{aligned}\hat{\theta}_{GHT} &= \mathbf{y}'\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\mathbf{Q}_{\hat{\mathbf{\Sigma}}}\mathbf{c} \\ &= \mathbf{y}'\mathbf{W}_{s\,GHT}\mathbf{c}.\end{aligned} \tag{2}$$

It is readily seen that regardless of the choice of $\hat{\mathbf{\Sigma}}$, $\hat{\theta}_{GHT}$ is unbiased for estimating $\theta = \mathbf{y}'\mathbf{c}$. Also, although $\hat{\theta}_{GHT}$ depends on $\hat{\mathbf{\Sigma}}$, it does not depend on auxiliary data. As required of an estimator, the rows of $\mathbf{W}_{s\,GHT}$ corresponding to nonsampled units are all $\mathbf{0}$, that is, $\mathbf{W}_{s\,GHT} = \mathbf{\Delta}_s\mathbf{W}_{s\,GHT}$. This follows from the following lemma, also proven in Appendix A, and the fact that $\mathbf{\Delta}_s$ is an orthogonal projection, that is, $\mathbf{\Delta}_s$ is symmetric and $\mathbf{\Delta}_s^2 = \mathbf{\Delta}_s$.

LEMMA 2. *Let $\mathbf{P} \in \mathbb{R}^{N\times N}$ be an orthogonal projection;*

(a) If $\mathbf{F} \in \mathbb{R}^{q\times N}$, then $(\mathbf{FP})^\dagger = \mathbf{P}(\mathbf{FP})^\dagger$.
(b) If $\mathbf{F} \in \mathbb{R}^{N\times q}$, then $(\mathbf{PF})^\dagger = (\mathbf{PF})^\dagger\mathbf{P}$.
(c) If $\mathbf{F} \in \mathbb{R}^{N\times N}$, then $(\mathbf{PFP})^\dagger = \mathbf{P}(\mathbf{PFP})^\dagger\mathbf{P}$.

Thus, if $\mathbf{W}_{[s]\,GHT} \in \mathbb{R}^{n \times N}$ and $\mathbf{y}_{[s]} \in \mathbb{R}^{n}$ are the submatrices of $\mathbf{W}_{s\,GHT}$ and $\mathbf{y}$ respectively, with rows corresponding to the sampled units, then

$$\hat{\theta}_{GHT} = \mathbf{y}'_{[s]}\mathbf{W}_{[s]\,GHT}\mathbf{c}. \tag{3}$$

It will be shown in Section 4 that among linear unbiased estimators $\hat{\theta}$, $\hat{\theta}_{GHT}$ minimizes $E_p V_\xi\left(\hat{\theta}\right)$.

If $\hat{\mathbf{\Sigma}}$ is a diagonal matrix, then $\hat{\theta}_{GHT}$ reduces to the Horvitz-Thompson estimator. Because, for diagonal matrices $\mathbf{F}_1$, $\mathbf{F}_2 \in \mathbb{R}^{N \times N}$, $(\mathbf{F}_1\mathbf{F}_2)^\dagger = \mathbf{F}_1^\dagger\mathbf{F}_2^\dagger$, diagonal matrices permute, $\mathbf{\Delta}_s^\dagger = \mathbf{\Delta}_s$, $\mathbf{\Delta}_s^2 = \mathbf{\Delta}_s$, and because $\hat{\mathbf{\Sigma}}^\dagger = \hat{\mathbf{\Sigma}}^{-1}$, it follows that

$$\begin{aligned} \hat{\theta}_{GHT} &= \mathbf{y}'\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\left(E_p\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\right)^{-1}\mathbf{c} \\ &= \mathbf{y}'\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}(E_p(\mathbf{\Delta}_s))^{-1}\mathbf{c} \\ &= \mathbf{y}'\mathbf{\Delta}_s(E_p(\mathbf{\Delta}_s))^{-1}\mathbf{c} \\ &= \hat{\theta}_{HT}. \end{aligned} \tag{4}$$

Somewhat more generally, if for every possible sample s, $\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s$ is diagonal, then $\hat{\theta}_{GHT}$ will also reduce to the Horvitz-Thompson estimator.

Note that the Horvitz-Thompson estimator, which uses a diagonal $\hat{\mathbf{\Sigma}}$, is unbiased, even if a more appropriate estimate of $\mathbf{\Sigma}$ would have a more complex structure; $\hat{\theta}_{GHT}$ is just as forgiving.

When $\hat{\mathbf{\Sigma}}$ is diagonal, so is $\mathbf{W}_{s\,GHT}$. The weights on the diagonal, the Horvitz-Thompson weights, are often referred to as the design weights. If $\hat{\mathbf{\Sigma}}$ is not diagonal, then the weight matrix $\mathbf{W}_{s\,GHT}$ and the vector $\mathbf{W}_{s\,GHT}\mathbf{c}$ depend on both the sampling design and on $\hat{\mathbf{\Sigma}}$. It may not be appropriate to refer to $\mathbf{W}_{s\,GHT}$ or $\mathbf{W}_{s\,GHT}\mathbf{c}$ as design weights.

The following is a simple consequence of Result 3 and will be needed before discussing the variance of $\hat{\theta}_{GHT}$.

RESULT 4. If $\hat{\mathbf{\Sigma}}$ is a positive definite estimator and $\hat{\mathbf{\Sigma}} \to \mathbf{\Sigma}$ in probability, then $\hat{\theta}_{GHT} = \mathbf{y}'\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^\dagger\mathbf{Q}_{\hat{\mathbf{\Sigma}}}\mathbf{c}$ and $\overset{*}{\theta}_{GHT} = \mathbf{y}'\left(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s\right)^\dagger\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c}$ are asymptotically equivalent.

Under the conditions of the preceding result, one has

$$\begin{aligned} V_p\left(\hat{\theta}_{GHT}\right) &\doteq V_p(\mathbf{y}'(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^\dagger\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c}) \\ &= V_p(\mathrm{vec}(\mathbf{y}'(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^\dagger\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c})) \\ &= (\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c} \otimes \mathbf{y})' V_p(\mathrm{vec}((\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^\dagger))(\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c} \otimes \mathbf{y}) \\ &= \|\mathbf{Q}_{\mathbf{\Sigma}}\mathbf{c} \otimes \mathbf{y}\|^2_{V_p(\mathrm{vec}((\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^\dagger))}, \end{aligned} \tag{5}$$

where vec($\mathbf{F}$) denotes the vector obtained by stacking the successive columns of the matrix $\mathbf{F}$.

In practice, $\mathbf{\Sigma}$ is unknown. The statistician will simply assume that a matrix $\mathbf{\Sigma}$ of a certain structure reflects the correlations among the population units. The matrix may, or may not, depend on certain parameters that need to be estimated. For example, in the concluding section, $\mathbf{\Sigma}$ is a block diagonal matrix where each block equals $\mathbf{1}_{2\times 2}$; there are no parameters to estimate. In Section 9, the correlation between persons of a same household is estimated; however, in that particular example, the estimated correlation is not used directly; a compromise that works for two important variables of interest is chosen. The computation of $\hat{\theta}_{GHT}$ also requires the computation of $\mathbf{Q}_{\hat{\mathbf{\Sigma}}}$. Although it can be difficult to find a closed form expression for $\mathbf{Q}_{\hat{\mathbf{\Sigma}}} = \left(E_p\left(\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^{\dagger}\right)\right)^{-1}$, its value can be approximated by repeatedly sampling the population using the sampling plan p, computing the average over the samples of $\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^{\dagger}$, and inverting that average. If $\hat{\mathbf{\Sigma}}$ varies with the sample s, then it would not be possible to compute $\left(\mathbf{\Delta}_s\hat{\mathbf{\Sigma}}\mathbf{\Delta}_s\right)^{\dagger}$ for all the samples. The alternative is to fix $\hat{\mathbf{\Sigma}}$ to the estimate obtained for the sample effectively drawn, then $\hat{\theta}_{GHT}$ will be biased, but still asymptotically unbiased. In the case of a two-stage sampling plan, the Horvitz-Thompson weights would likely be applied to the primary sampling units and the methods of this article, including the method just described to compute $\mathbf{Q}_{\hat{\mathbf{\Sigma}}}$, would apply to the secondary sampling units. For that purpose, the population consists of the secondary sampling units that belong to the primary sampling units selected in the first stage. For that "population", $\mathbf{\Sigma}$ would typically be block diagonal with each block corresponding to a selected primary sampling unit.

4. A Generalization of the Godambe and Joshi Lower Bound

Although it wasn't in the context of an asymptotic setup and although it was assumed that $V_{\xi}(\mathbf{y}) = \mathbf{\Sigma}$ was diagonal, for any unbiased estimator $\hat{\theta}$ of θ, Godambe and Joshi (1965) have given a lower bound for the value of $E_{\xi}V_p\left(\hat{\theta}\right)$. The derivation of that lower bound used the following identity:

$$E_{\xi}V_p\left(\hat{\theta}\right) = E_pV_{\xi}\left(\hat{\theta}\right) + E_p\left[E_{\xi}\left(\hat{\theta} - \theta\right)\right]^2 - V_{\xi}(\theta). \tag{6}$$

Also, for any linear unbiased estimator $\hat{\theta}$ of θ

$$\begin{aligned} E_pV_{\xi}\left(\hat{\theta}\right) &= E_pV_{\xi}\left(\theta^*_{GHT} + \left(\hat{\theta} - \theta^*_{GHT}\right)\right) \\ &= E_pV_{\xi}\left(\theta^*_{GHT}\right) + E_pV_{\xi}\left(\hat{\theta} - \theta^*_{GHT}\right) + 2\,E_pCov_{\xi}\left(\theta^*_{GHT},\ \left(\hat{\theta} - \theta^*_{GHT}\right)\right) \\ &\geq E_pV_{\xi}\left(\theta^*_{GHT}\right) + 2\,E_pCov_{\xi}\left(\theta^*_{GHT},\ \left(\hat{\theta} - \theta^*_{GHT}\right)\right) \\ &= E_pV_{\xi}\left(\theta^*_{GHT}\right), \end{aligned} \tag{7}$$

because for any linear unbiased estimator $\hat{0}$ of 0, $E_pCov_{\xi}\left(\theta^*_{GHT},\ \hat{0}\right) = 0$. To show this, let $\hat{0}$ be written $\mathbf{y}'\mathbf{\Delta}_s\boldsymbol{\lambda}_{0s} + \kappa_s$, with $\boldsymbol{\lambda}_{0s} \in \mathbb{R}^N$ and κ_s independent of $\mathbf{y}$, may depend on the sample s. Setting $\mathbf{y} = \mathbf{0}$ yields $E_p(\kappa_s) = 0$. The following derivation uses Lemma 2 as well as Ben-Israel and Greville (2002, Exercise 2.21):

$$\begin{aligned} E_p Cov_\xi\left(\overset{*}{\theta}_{GHT}, \hat{0}\right) &= E_p Cov_\xi(\mathbf{y}'(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger \mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}, \mathbf{y}'\boldsymbol{\Delta}_s\boldsymbol{\lambda}_{0s}) \\ &= E_p(\mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger Cov_\xi(\mathbf{y},\mathbf{y})\boldsymbol{\Delta}_s\boldsymbol{\lambda}_{0s}) \\ &= E_p(\mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_s(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger \boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s\boldsymbol{\lambda}_{0s}) \\ &= E_p(\mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_s\boldsymbol{\lambda}_{0s} + \kappa_s) \\ &= 0, \end{aligned} \tag{8}$$

because $\mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_s\boldsymbol{\lambda}_{0s} + \kappa_s$ is the unbiased estimator $\hat{0}$ with $\mathbf{y} = \mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}$ as the vector of interest.

The inequality (7) is what makes $\overset{*}{\theta}_{GHT}$, and $\hat{\theta}_{GHT}$, special. Taken together with (6) it shows that for any linear unbiased estimator, $\hat{\theta}$,

$$E_\xi V_p\left(\hat{\theta}\right) \geq E_p V_\xi\left(\overset{*}{\theta}_{GHT}\right) - V_\xi(\theta). \tag{9}$$

Knowing that $V_\xi(\theta) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} = ||\mathbf{c}||^2_{\boldsymbol{\Sigma}}$ and that

$$\begin{aligned} E_p V_\xi\left(\overset{*}{\theta}_{GHT}\right) &= E_p\left(\mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger\boldsymbol{\Sigma}(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\right) \\ &= \mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}E_p((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger)\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c} \\ &= \mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}E_p(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c} \\ &= \mathbf{c}'\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c} \\ &= ||\mathbf{c}||^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}}, \end{aligned} \tag{10}$$

allows the following generalization to a positive definite matrix, $\boldsymbol{\Sigma}$, of a lower bound given in Godambe and Joshi (1965).

RESULT 5. For any linear unbiased estimator, $\hat{\theta}$ of $\theta = \mathbf{y}'\mathbf{c}$, if $V_\xi(\mathbf{y}) = \boldsymbol{\Sigma}$ is positive definite, then

$$E_\xi V_p\left(\hat{\theta}\right) \geq ||\mathbf{c}||^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}}. \tag{11}$$

If $\boldsymbol{\Sigma}$ is a diagonal matrix equal to $\text{diag}\left(\sigma_k^2\right)_{k=1,2,\ldots,N}$, then (11) reduces to $E_\xi V_p(\hat{\theta}) \geq ||\mathbf{c}||^2_{(E_p(\boldsymbol{\Delta}_s)^{-1}-\mathbf{I}_N)\boldsymbol{\Sigma}} = \sum_k\left(\frac{1}{\pi_k} - 1\right)\sigma_k^2 c_k^2$, which is the Godambe and Joshi lower bound, usually given for a total, that is, for $\mathbf{c} = \mathbf{1}_{N\times1}$. If $\mathbf{c} = \mathbf{1}_{N\times1}$, then the generalized lower bound equals the sum of all entries in the matrix $\mathbf{Q}_{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$. It should be noted that Godambe and Joshi (1965) had proven that, if $\boldsymbol{\Sigma}$ is diagonal, the lower bound holds for the class of all unbiased estimators, not only for the class of linear unbiased estimators. What if $\boldsymbol{\Sigma}$ is allowed not to be diagonal? Does the generalized Godambe-Joshi lower bound apply to all unbiased estimators? The answer is no, and a nonlinear counter-example is given in Appendix B.

When the variance matrix $\mathbf{\Sigma}$ is diagonal, it is known that the calibration estimator introduced by Deville and Särndal (1992), or the equivalent generalized regression estimator, asymptotically attains the Godambe and Joshi lower bound (see, for example Särndal et al. 1992). In the next section, the calibration estimator will be generalized to the case of a positive definite variance matrix $\mathbf{\Sigma}$. It will then be shown that this generalized calibration estimator asymptotically attains the lower bound given in (11).

5. A Generalization of the Calibration Estimator

Define $\mathbf{1}'_{N\times 1}\mathbf{W}_{s\,GHT}\mathbf{c}$ to be the effective sample weight for estimating $\theta = \mathbf{y}'\mathbf{c}$. The variance of the effective sample weight will often be larger if $\hat{\mathbf{\Sigma}}$ is nondiagonal, and this variance will negatively affect the estimator $\hat{\theta}_{GHT}$. This source of variance can be eliminated with the use of a weight vector $\mathbf{w}_s$ that satisfies the calibration equation $\mathbf{1}'_{N\times 1}\mathbf{w}_s = \mathbf{1}'_{N\times 1}\mathbf{c}$. In this section, to estimate $\theta = \mathbf{y}'\mathbf{c}$, an estimator $\hat{\theta}_{GCAL} = \mathbf{y}'\mathbf{w}_{s\,GCAL}$ will be derived through calibration using an auxiliary variable matrix $\mathbf{X}$ assumed to be of full rank. More precisely, noting $\mathbf{w}_{s\,GHT} = \mathbf{W}_{s\,GHT}\mathbf{c}$, the following problem is addressed:

Calibration Problem: Among the weight vectors $\mathbf{w}_s$ in the range of $\mathbf{\Delta}_s$, $\mathcal{R}(\mathbf{\Delta}_s)$, (nonsampled units should have a weight of 0) which minimize $||\mathbf{X}'\mathbf{w}_s - \mathbf{X}'\mathbf{c}||_{\mathbf{T}}$, that is, which best satisfy the calibration equations, seek one that minimizes $||\mathbf{w}_s - \mathbf{w}_{s\,GHT}||_{\mathbf{U}}$, that is, a weight vector as close as possible to the generalized Horvitz-Thompson weights, where $\mathbf{T} \in \mathbb{R}^{q\times q}$ and $\mathbf{U} \in \mathbb{R}^{N\times N}$ are positive semi-definite matrices.

Weights, $\mathbf{w}_s$, that satisfy the calibration equations, $\mathbf{X}'\mathbf{w}_s = \mathbf{X}'\mathbf{c}$, do not always exist, especially if the number of equations, q, is high relative to the sample size. To prepare for this eventuality, the matrix $\mathbf{T}$ is at the statistician's disposal for specifying the relative importance of the q calibration equations.

This formulation of the calibration problem generalizes that of Théberge (1999), where $\mathbf{T}$ and $\mathbf{U}$ were diagonal matrices and the Horvitz-Thompson weights were used instead of the generalized Horvitz-Thompson weights.

Setting $\mathbf{v} = \mathbf{w}_s - \mathbf{w}_{s\,GHT}$, a minimum-norm least-squares solution is sought. A helpful theorem is given in Rao and Mitra (1971).

THEOREM 1. *Let* $\mathbf{T} \in \mathbb{R}^{q\times q}$ *and* $\mathbf{U} \in \mathbb{R}^{N\times N}$ *be symmetric positive semi-definite matrices, also let* $\mathbf{A} \in \mathbb{R}^{q\times N}$ *and* $\mathbf{b} \in \mathbb{R}^{q}$. *There is a matrix* $\mathbf{G} \in \mathbb{R}^{N\times q}$ *such that* $\mathbf{Gb}$ *minimizes* $||\mathbf{v}||_{\mathbf{U}}$ *among the vectors* $\mathbf{v} \in \mathbb{R}^{N}$ *which minimize* $||\mathbf{Av} - \mathbf{b}||_{\mathbf{T}}$, *if and only if*

$$\mathbf{TAGA} = \mathbf{TA} \quad \mathbf{UGAG} = \mathbf{UG} \quad \mathbf{TAG} = (\mathbf{TAG})' \quad \mathbf{UGA} = (\mathbf{UGA})'. \tag{12}$$

Choices for $\mathbf{G}$ *are*

$$\begin{aligned} \mathbf{G} &= (\mathbf{I}_N - (\mathbf{P}_{\mathcal{N}(\mathbf{A}'\mathbf{TA})}\mathbf{U}\mathbf{P}_{\mathcal{N}(\mathbf{A}'\mathbf{TA})})^{\dagger}\mathbf{U})(\mathbf{A}'\mathbf{TA})^{\dagger}\mathbf{A}'\mathbf{T}, \\ &= \mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{A}\mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2} \quad \textit{if } \mathbf{U} \textit{ is positive definite}, \end{aligned} \tag{13}$$

where $\mathbf{T}^{1/2}$ *is the symmetric positive semi-definite square root of* $\boldsymbol{T}$ *and* $\mathbf{P}_S$ *is the orthogonal projection on* S*, a subspace of* $\mathbb{R}^N$.

The first part of the theorem is proven in Rao and Mitra (1971), where other choices for $\mathbf{G}$ are given. It is shown in Appendix C that (13) does satisfy (12). To compute $\mathbf{P}_{\mathcal{N}(\mathbf{A}'\mathbf{TA})}$, the identity $\mathbf{P}_{\mathcal{N}(\mathbf{F})} = \mathbf{I} - \mathbf{F}^{\dagger}\mathbf{F}$ can be used.

The first choice of $\mathbf{G}$ given in (13) is derived from Ben-Israel and Greville (2002, Corollary 8.2) which is itself a consequence of the generalized Gauss-Markov theorem, see Zyskind and Martin (1969) and Albert (1973). The second choice of $\mathbf{G}$ given in (13) is derived from Théberge (1999).

If $\mathbf{U}$ is positive definite, there are two other possible forms for $\mathbf{G}$, namely

$$\mathbf{G} = \mathbf{U}^{-1/2}(\mathbf{T}^{1/2}\mathbf{A}\mathbf{U}^{-1/2})^{\dagger}\mathbf{T}^{1/2} \tag{14}$$

and

$$\mathbf{G} = \mathbf{U}^{-1/2}(\mathbf{U}^{-1/2}\mathbf{A}'\mathbf{T}\mathbf{A}\mathbf{U}^{-1/2})^{\dagger}\mathbf{U}^{-1/2}\mathbf{A}'\mathbf{T}, \tag{15}$$

where $\mathbf{U}^{1/2}$ is the symmetric positive semi-definite square root of $\mathbf{U}$. Applying the identity $\mathbf{F}^{\dagger} = \mathbf{F}'(\mathbf{F}\mathbf{F}')^{\dagger}$ to the Moore-Penrose inverse on the right-hand side of (14) yields the second part of (13); applying the identity $\mathbf{F}^{\dagger} = (\mathbf{F}'\mathbf{F})^{\dagger}\mathbf{F}'$ to the Moore-Penrose inverse on the right-hand side of (14) yields (15). Both of those identities are found in Ben-Israel and Greville (2002, Exercise 1.18).

The theorem must be modified to take into account that $\mathbf{v}$ is constrained to $\mathcal{R}(\mathbf{\Delta}_s)$, a subspace S of $\mathbb{R}^N$. This is done by applying the method of Ben-Israel and Greville (2002, sec. 2.9) and minimizing $\|\mathbf{v}\|_{\mathbf{U}} = \|\mathbf{P}_S\mathbf{z}\|_{\mathbf{U}} = \|\mathbf{z}\|_{\mathbf{P}_S\mathbf{U}\mathbf{P}_S}$ among the vectors which minimize $\|\mathbf{A}\mathbf{P}_S\mathbf{z} - \mathbf{b}\|_{\mathbf{T}}$. Using the preceding theorem to find the optimal $\mathbf{z}$, gives the constrained analog:

THEOREM 2. *Let* $\mathbf{T} \in \mathbb{R}^{q\times q}$ *and* $\mathbf{U} \in \mathbb{R}^{N\times N}$ *be symmetric positive semi-definite matrices,* $\mathbf{A} \in \mathbb{R}^{q\times N}$, $\mathbf{b} \in \mathbb{R}^q$ *and* $\mathbf{P}_S$ *be the orthogonal projection on S a subspace of* $\mathbb{R}^N$. *There is a matrix* $\mathbf{G} \in \mathbb{R}^{N\times q}$ *such that* $\mathbf{Gb}$ *minimizes* $\|\mathbf{v}\|_{\mathbf{U}}$ *among the vectors* $\mathbf{v} \in S$ *which minimize* $\|\mathbf{A}\mathbf{v} - \mathbf{b}\|_{\mathbf{T}}$, *if and only if*

$$\begin{aligned}&\mathbf{T}\mathbf{A}\mathbf{P}_S\mathbf{G}\mathbf{A}\mathbf{P}_S = \mathbf{T}\mathbf{A}\mathbf{P}_S \quad \mathbf{P}_S\mathbf{U}\mathbf{P}_S\mathbf{G}\mathbf{A}\mathbf{P}_S\mathbf{G} = \mathbf{P}_S\mathbf{U}\mathbf{P}_S\mathbf{G}\\ &\mathbf{T}\mathbf{A}\mathbf{P}_S\mathbf{G} = (\mathbf{T}\mathbf{A}\mathbf{P}_S\mathbf{G})' \quad \mathbf{P}_S\mathbf{U}\mathbf{P}_S\mathbf{G}\mathbf{A}\mathbf{P}_S = (\mathbf{P}_S\mathbf{U}\mathbf{P}_S\mathbf{G}\mathbf{A}\mathbf{P}_S)'.\end{aligned} \tag{16}$$

Choices for $\mathbf{G}$ *are*

$$\begin{aligned}\mathbf{G} &= (\mathbf{I}_N - (\mathbf{P}_{\mathcal{N}(\mathbf{P}_S\mathbf{A}'\mathbf{T}\mathbf{A}\mathbf{P}_S)}\mathbf{P}_S\mathbf{U}\mathbf{P}_S\mathbf{P}_{\mathcal{N}(\mathbf{P}_S\mathbf{A}'\mathbf{T}\mathbf{A}\mathbf{P}_S)})^{\dagger}\mathbf{P}_S\mathbf{U}\mathbf{P}_S)(\mathbf{P}_S\mathbf{A}'\mathbf{T}\mathbf{A}\mathbf{P}_S)^{\dagger}\mathbf{A}'\mathbf{T},\\ &= (\mathbf{P}_S\mathbf{U}\mathbf{P}_S)^{\dagger}\mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{A}(\mathbf{P}_S\mathbf{U}\mathbf{P}_S)^{\dagger}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2} \quad \textit{if } \mathbf{U} \textit{ is positive definite.}\end{aligned} \tag{17}$$

Applying this to our calibration problem stated at the beginning of this section means setting $\mathbf{P}_S = \mathbf{\Delta}_s$, $\mathbf{b} = \mathbf{X}'\mathbf{c} - \mathbf{X}'\mathbf{w}_{s\,GHT}$ and $\mathbf{A} = \mathbf{X}'$. This gives the generalized calibration weight vector

$$\mathbf{w}_{s\,GCAL} = \mathbf{w}_{s\,GHT} + \mathbf{G}(\mathbf{X}'\mathbf{c} - \mathbf{X}'\mathbf{w}_{s\,GHT}), \tag{18}$$

and the calibration estimator becomes

$$\begin{aligned}\hat{\theta}_{GCAL} &= \mathbf{y}'\mathbf{w}_{s\,GCAL}\\ &= \hat{\mathbf{y}}'\mathbf{c} + (\mathbf{y} - \hat{\mathbf{y}})' \ \mathbf{w}_{s\,GHT},\end{aligned} \tag{19}$$

where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = \mathbf{G}'\mathbf{y}$ and $\mathbf{G}$ is given by (17) with $\mathbf{P}_S = \mathbf{\Delta}_s$ and $\mathbf{A} = \mathbf{X}'$.

It should be noted that the weight vector $\mathbf{w}_{sGCAL}$ can be written in the form $\mathbf{W}_{s\,GCAL}\mathbf{c}$ with $\mathbf{W}_{s\,GCAL} \in \mathbb{R}^{N\times N}$. For (18) to hold true for any vector $\mathbf{c}$, one must have $\mathbf{W}_{sGCAL} = \mathbf{W}_{sGHT} + \mathbf{G}(\mathbf{X}' - \mathbf{X}'\mathbf{W}_{sGHT})$.

In the remainder of this section, it will be assumed that $\mathbf{U}$ is positive definite and the second choice for $\mathbf{G}$ given in (17) will be used. Then, one can write

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{X}'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\mathbf{X}\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\mathbf{y}\end{aligned} \tag{20}$$

and

$$\begin{aligned}\mathbf{w}_{s\,GCAL} = \mathbf{w}_{s\,GHT} &+ (\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{X}'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\mathbf{X}\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2} \\ &\times(\mathbf{X}'\mathbf{c} - \mathbf{X}'\mathbf{w}_{s\,GHT}).\end{aligned} \tag{21}$$

If one notes $\mathbf{w}_{[s]\,GCAL} \in \mathbb{R}^n$, $\mathbf{w}_{[s]\,GHT} \in \mathbb{R}^n$, $\mathbf{U}_{[ss]} \in \mathbb{R}^{n\times n}$ and $\mathbf{X}_{[s]} \in \mathbb{R}^{n\times q}$, the subvectors and submatrices with lines corresponding to the sampled units, then using Lemma 2,

$$\begin{aligned}\mathbf{w}_{[s]\,GCAL} = \mathbf{w}_{[s]\,GHT} &+ \mathbf{U}_{[ss]}^{-1}\mathbf{X}_{[s]}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'_{[s]}\mathbf{U}_{[ss]}^{-1}\mathbf{X}_{[s]}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2} \\ &\times(\mathbf{X}'\mathbf{c} - \mathbf{X}'_{[s]}\mathbf{w}_{[s]\,GHT}),\end{aligned} \tag{22}$$

because the weights of nonsampled units are zero. Thus, $\hat{\theta}_{GCAL} = \mathbf{y}'_{[s]}\mathbf{w}_{[s]\,GCAL}$. This shows that for computing $\hat{\theta}_{GCAL}$ the population parameter $\mathbf{X}'\mathbf{c}$ must be known, but the individual rows of $\mathbf{X}$ need only be known for those corresponding to sampled units. It is seen that the weights given by (22) could be interpreted as those from a GREG estimator, see Cassel et al. (1977), except that the Horvitz-Thompson weights are replaced with those of the generalized Horvitz-Thompson estimator, a matrix $\mathbf{T}$ has been introduced in case $\mathbf{X}_{[s]}$ is not of full rank and the matrix $\mathbf{U}$ would be set equal to $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}$ in a GREG estimator. For $\hat{\boldsymbol{\Sigma}}$ diagonal, $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}$ reduces to $\hat{\boldsymbol{\Sigma}}(E_p(\boldsymbol{\Delta}_s))^{-1}$. Equation (20) with $\mathbf{U} = \mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}$ is thus a generalization of the value of $\hat{\mathbf{y}}$ for a GREG estimator when $\hat{\boldsymbol{\Sigma}}$, the estimated variance matrix of $\mathbf{y}$ under the model, is not necessarily diagonal.

If $\mathbf{U}^{1/2}$ is the unique positive definite square root of $\mathbf{U}$, then defining $\mathbf{Z}_{[s]} = \mathbf{U}_{[ss]}^{-1/2}\mathbf{X}_{[s]}$ yields

$$\begin{aligned}\hat{\theta}_{GCAL} &= \mathbf{y}'_{[s]}\mathbf{w}_{[s]\,GCAL} \\ &= \mathbf{y}'_{[s]}\mathbf{w}_{[s]\,GHT} + \left(\mathbf{U}_{[ss]}^{-1/2}\mathbf{y}_{[s]}\right)'\mathbf{Z}_{[s]}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{Z}'_{[s]}\mathbf{Z}_{[s]}\mathbf{T}^{1/2}\right)^{\dagger} \\ &\quad\times\mathbf{T}^{1/2}\left(\mathbf{X}'\mathbf{c} - \mathbf{X}'_{[s]}\mathbf{w}_{[s]\,GHT}\right).\end{aligned} \tag{23}$$

If $\mathbf{T}$ is also positive definite and if $\mathbf{X}_{[s]}$ is of full rank, then (23) simplifies to

$$\hat{\theta}_{GCAL} = \mathbf{y}'_{[s]}\mathbf{w}_{[s]\,GHT} + \left(\mathbf{U}_{[ss]}^{-1/2}\mathbf{y}_{[s]}\right)'\mathbf{Z}_{[s]}\left(\mathbf{Z}'_{[s]}\mathbf{Z}_{[s]}\right)^{-1}\left(\mathbf{X}'\mathbf{c} - \mathbf{X}'_{[s]}\mathbf{w}_{[s]\,GHT}\right). \tag{24}$$

In this form, a parallel can be drawn with the use of "instrumental variables", as for example, in Estevao and Särndal (2003).

Replacing $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}$ in $\hat{\theta}_{GCAL}$ and noting $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta}$ gives the random variable $\theta^*_{GCAL} = (\mathbf{y}^*)'\mathbf{c} + (\mathbf{y} - \mathbf{y}^*)'\mathbf{w}_{s\,GHT}$. The bias of θ^*_{GCAL} is zero. Also, $\hat{\theta}_{GCAL}$ and θ^*_{GCAL} are asymptotically equivalent. Indeed,

$$\begin{aligned} t^{1/2-\gamma}\left(\hat{\theta}_{GCALt} - \theta^*_{GCALt}\right) &= t^{1/2-\gamma}\left(\mathbf{y}^*_t - \hat{\mathbf{y}}_t\right)'\left(\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{c}_t\right) \\ &= t^{1/2-\gamma}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_t)'\left[\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t\mathbf{c}_t\right] \end{aligned} \tag{25}$$

tends to 0 in probability since, from the results of Section 2, $\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t\mathbf{c}_t$ is $O_p(t^{\gamma-1/2})$ and $\hat{\boldsymbol{\beta}}_t \rightarrow \boldsymbol{\beta}$ in probability. This leads to the following result.

RESULT 6. For any positive definite matrix $\mathbf{U}$, if the calibration equations can be satisfied, $E_{\xi}V_p\left(\theta^*_{GCAL}\right)$ attains the lower bound given in (11).

To prove this, use (6) with $\hat{\theta} = \theta^*_{GCAL}$ while noting that $E_{\xi}\left(\theta^*_{GCAL} - \theta\right) = 0$, that $V_{\xi}(\theta) = \|\mathbf{c}\|^2_{\boldsymbol{\Sigma}}$, and that $E_pV_{\xi}\left(\theta^*_{GCAL}\right) = E_pV_{\xi}\left(\hat{\theta}_{GHT}\right) \doteq E_pV_{\xi}\left(\theta^*_{GHT}\right) = \|\mathbf{c}\|^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}}$. It should be noted that generally, $\hat{\theta}_{GHT}$ does not attain the lower bound; calibration is required.

For the generalized calibration estimator to asymptotically attain the lower bound, it is important for the generalized Horvitz-Thompson weights, $\mathbf{w}_{s\,GHT}$, to be calculated with a matrix $\hat{\boldsymbol{\Sigma}}$ that satisfies the conditions of Result 3. Also, the same auxiliary variables as appear in the model are to be used for the calibration, so that $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$. Note that $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$ whatever the choice of the positive definite matrix **U,** and the choice has no impact on whether or not the generalized calibration estimator asymptotically attains the lower bound. The case of a variance matrix $\boldsymbol{\Sigma}$ which is not diagonal has been examined before, see, for example, Montanari and Ranalli (2002). The focus is usually on the choice of the estimator $\hat{\boldsymbol{\beta}}$, or more precisely on the choice of the matrix **U**. Result 6 puts the importance of **U** in perspective.

It was seen in Section 3 that if $\hat{\boldsymbol{\Sigma}}$ is diagonal, then $\mathbf{w}_{s\,GHT}$ reduces to the usual Horvitz-Thompson weights. If the matrices **U** and **T** are also chosen to be diagonal, then the generalized calibration estimator reduces to the usual calibration estimator as given in Théberge (1999).

Assuming that $\hat{\boldsymbol{\Sigma}}$ satisfies the conditions of Result 3, the variance of the generalized calibration estimator is

$$\begin{aligned} V_p\left(\hat{\theta}_{GCAL}\right) &\doteq V_p\left(\theta^*_{GCAL}\right) \\ &= V_p((\mathbf{y} - \mathbf{y}^*)'\mathbf{w}_{s\,GHT}) \\ &\doteq \|\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c} \otimes (\mathbf{y} - \mathbf{y}^*)\|^2_{V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}))}. \end{aligned} \tag{26}$$

Theorem 1 may also be used to find an optimal vector $\boldsymbol{\beta}$; one which minimizes the variance. Among the vectors $\boldsymbol{\beta}$ which minimize

$$\begin{aligned} V_p\left(\hat{\theta}_{GCAL}\right) &\doteq \|\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes(\mathbf{y}-\mathbf{y}^*)\|^2_{V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger))} \\ &= \|(\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{X})\boldsymbol{\beta}-\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{y}\|^2_{V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger))}, \end{aligned} \tag{27}$$

the one that minimizes $\|\boldsymbol{\beta}\|^2_{\mathbf{U}}$ can be found by applying Theorem 1 with $\mathbf{A}=\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{X}$, $\mathbf{b}=\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{y}$, $\mathbf{T}=V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger))$ and by using $\mathbf{G}$ given by (15). This gives

$$\begin{aligned} \boldsymbol{\beta}_{opt} = {}& \mathbf{U}^{-1/2}[\mathbf{U}^{-1/2}(\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{X})'V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger))(\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{X})\mathbf{U}^{-1/2}]^\dagger\mathbf{U}^{-1/2} \\ & \times(\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{X})'V_p(\mathrm{vec}((\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^\dagger))(\mathbf{Q}_{\boldsymbol{\Sigma}}\mathbf{c}\otimes\mathbf{y}). \end{aligned} \tag{28}$$

If $\boldsymbol{\Sigma}$ is diagonal, then $V_p\left(\hat{\theta}_{GCAL}\right)\doteq\|\mathrm{diag}(\mathbf{c})(\mathbf{y}-\mathbf{y}^*)\|^2_{\mathbf{A\Pi A}-\mathbf{1}_{N\times N}}=\|\mathrm{diag}(\mathbf{c})\mathbf{X}\boldsymbol{\beta}-\mathrm{diag}(\mathbf{c})\mathbf{y}\|^2_{\mathbf{A\Pi A}-\mathbf{1}_{N\times N}}$, where diag $(\mathbf{v})$ denotes the diagonal matrix formed from the vector $\mathbf{v}$, $\boldsymbol{\Pi}$ is the matrix of the second-order inclusion probabilities and $\mathbf{A}=(E(\boldsymbol{\Delta}_s))^{-1}$. Applying Theorem 1, again with $\mathbf{G}$ given by (15), to find the optimal $\boldsymbol{\beta}$ will give

$$\begin{aligned} \boldsymbol{\beta}_{opt} = {}& \mathbf{U}^{-1/2}[\mathbf{U}^{-1/2}\mathbf{X}'\mathrm{diag}(\mathbf{c})(\mathbf{A\Pi A}-\mathbf{1}_{N\times N})\mathrm{diag}(\mathbf{c})\mathbf{X}\mathbf{U}^{-1/2}]^\dagger\mathbf{U}^{-1/2} \\ & \times\mathbf{X}'\mathrm{diag}(\mathbf{c})(\mathbf{A\Pi A}-\mathbf{1}_{N\times N})\mathrm{diag}(\mathbf{c})\mathbf{y}, \end{aligned} \tag{29}$$

a result similar to that found in Montanari (1998), if one sets $\mathbf{U}=\mathbf{I}$ and $\mathbf{c}=\mathbf{1}_{N\times 1}$.

The variance $V_p\left(\hat{\theta}_{GCAL}\right)$ is an unbiased estimator, under ξ, of $E_\xi V_p\left(\hat{\theta}_{GCAL}\right)\doteq\|\mathbf{c}\|^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}}$. It is then possible for $V_p\left(\hat{\theta}_{GCAL}\right)$ to be smaller than the lower bound $E_\xi V_p\left(\hat{\theta}_{GCAL}\right)\doteq\|\mathbf{c}\|^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}}$. Also, for any other linear unbiased estimator, $\hat{\theta}$, the variance $V_p\left(\hat{\theta}\right)$ is an unbiased estimator, under ξ, of $E_\xi V_p\left(\hat{\theta}\right)\geq\|\mathbf{c}\|^2_{\mathbf{Q}_{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}}$. The fact that $V_p\left(\hat{\theta}_{GCAL}\right)$ is an unbiased estimator of a parameter not greater than the parameter estimated by $V_p\left(\hat{\theta}\right)$ is not a guarantee that $V_p\left(\hat{\theta}_{GCAL}\right)$ is not greater than $V_p\left(\hat{\theta}\right)$, but it is a point in favor of $\hat{\theta}_{GCAL}$.

For a population parameter, $\boldsymbol{\Omega}$, one could use the asymptotic setup to define $\lim\limits_{U\to U_\infty}\boldsymbol{\Omega}=E_\xi(\boldsymbol{\Omega})$. For example, with $\mathbf{U}\in\mathbb{R}^{N\times N}$ a symmetric positive definite matrix, a population regression parameter $\mathbf{B}=\left(\mathbf{X}'\mathbf{Q}_{\mathbf{U}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Q}_{\mathbf{U}}^{-1}\mathbf{y}$ can be defined; by definition, one has $\lim\limits_{U\to U_\infty}\mathbf{B}=E_\xi(\mathbf{B})$, which in this case is $\boldsymbol{\beta}$. In that sense, $\lim\limits_{U\to U_\infty}\left(V_p\left(\hat{\theta}_{GCAL}\right)\right)\leq\lim\limits_{U\to U_\infty}\left(V_p\left(\hat{\theta}\right)\right)$ for any linear unbiased estimator $\hat{\theta}$.

6. Example

The computation of $\hat{\theta}_{GHT}$ and of $\hat{\theta}_{GCAL}$ requires the computation of $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}$. An iterative method of computation was described at the end of Section 3. In this section, an example is examined where it is possible to obtain a closed form expression for $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}$. It will be assumed that $\hat{\boldsymbol{\Sigma}}$ is a block diagonal matrix. Units of a same block could be persons of a same household, workers of a same establishment, children of a same school, or similar groupings. Let us say that there are F blocks, then

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{\boldsymbol{\Sigma}}_F \end{pmatrix}. \tag{30}$$

In this example, it will be assumed that units belonging to the same block have the same variance and the same covariance. Note that multiplying $\hat{\boldsymbol{\Sigma}}_f \; f = 1, \ldots, F$ by a scalar leaves $\mathbf{W}_{s\ GHT}$ unchanged, even if the scalar varies with f. More precisely, if N_f is the size of block f, it will be assumed that $\hat{\boldsymbol{\Sigma}}_f = \mathbf{I}_{N_f} + \rho_f(\mathbf{1}_{N_f \times N_f} - \mathbf{I}_{N_f}) \; f = 1, \ldots, F$ with $\frac{-1}{N_f - 1} < \rho_f < 1$.

With $\hat{\boldsymbol{\Sigma}}$ of this form, it is possible to find a closed form expression for the block diagonal matrix

$$\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}} = \begin{pmatrix} \mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_F} \end{pmatrix}. \tag{31}$$

For block f, $\mathbf{Q}^{-1}_{\hat{\boldsymbol{\Sigma}}_f} = E_p\left(\left(\boldsymbol{\Delta}_{s_f} \hat{\boldsymbol{\Sigma}}_f \boldsymbol{\Delta}_{s_f}\right)^{\dagger}\right)$, where $\boldsymbol{\Delta}_{s_f}$ is the $N_f \times N_f$ submatrix of $\boldsymbol{\Delta}_s$ that corresponds to block f. Conditioning on the number of units of the block that are sampled, S_f, one obtains

$$\mathbf{Q}^{-1}_{\hat{\boldsymbol{\Sigma}}_f} = \sum_{n_f=1}^{N_f} \mathrm{P}(S_f = n_f) E_p\left(\left(\boldsymbol{\Delta}_{s_f} \hat{\boldsymbol{\Sigma}}_f \boldsymbol{\Delta}_{s_f}\right)^{\dagger} | S_f = n_f\right). \tag{32}$$

The probabilities, $\mathrm{P}(S_f = n_f)$, can be expressed as a function of the inclusion probabilities. If one writes

$$\mathrm{P}(S_f = n_f) = \sum_{i=n_f}^{N_f} \left(k(i) \sum_{\text{block}f} \pi^{[i]} \right) \quad 1 \leq n_f \leq N_f \tag{33}$$

with $\sum\limits_{\text{block}f} \pi^{[i]}$ being the sum of all probabilities of inclusion of order i, where all i units are in block f, then the application of the inclusion-exclusion principle will yield the recurrence relation

$$k(i) = -\sum_{j=1}^{i-n_f} \binom{i}{j} k(i-j) \quad n_f + 1 \leq i \leq N_f \tag{34}$$

with $k(n_f) = 1$. An example with $N_f = 5$ is given in Appendix D.

Note that $E_p\left(\left(\boldsymbol{\Delta}_{s_f} \hat{\boldsymbol{\Sigma}}_f \boldsymbol{\Delta}_{s_f}\right)^{\dagger} | S_f = 1\right) = E_p(\boldsymbol{\Delta}_{s_f} | S_f = 1)$, which is the diagonal matrix of the first order conditional (on $S_f = 1$) inclusion probabilities. Finally, for $n_f \geq 2$,

$E_p\left(\left(\mathbf{\Delta}_{s_f}\hat{\mathbf{\Sigma}}_f\mathbf{\Delta}_{s_f}\right)^{\dagger}|S_f = n_f\right) = E_p(\mathbf{\Delta}_{s_f}\mathbf{M}_f\mathbf{\Delta}_{s_f}|S_f = n_f) = \mathbf{\Pi}_{n_f} \circ \mathbf{M}_f$, where the diagonal elements of $\mathbf{M}_f \in \mathbb{R}^{N_f \times N_f}$ are equal to $\frac{-(n_f-2)\rho_f-1}{(n_f-1)\rho_f^2-(n_f-2)\rho_f-1}$, whereas the offdiagonal elements are equal to $\frac{\rho_f}{(n_f-1)\rho_f^2-(n_f-2)\rho_f-1}$, $\mathbf{\Pi}_{n_f}$ is the matrix of second order conditional (on $S_f = n_f$) probabilities of inclusion, and $\circ$ denotes the Hadamard product, that is, element-wise multiplication. The value of the elements of $\mathbf{M}_f$ come from inverting an $n_f \times n_f$ submatrix of $\hat{\mathbf{\Sigma}}_f$.

7. A Modification to the Generalized Estimators

Note that $(\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})$ is positive definite. Indeed, for any non-zero vector $\mathbf{z} \in \mathbb{R}^N$, $E_p(V_\xi(\mathbf{z}'\mathbf{\Delta}_s\mathbf{\Sigma}^{-1}\mathbf{y})) = \mathbf{z}'(\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})\mathbf{z} > 0$. If $\mathbf{\Sigma}$ is known, instead of $\hat{\theta}_{GHT} = \theta^{*}_{GHT} = \mathbf{y}'(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}(E_p((\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}))^{-1}\mathbf{c}$, one could use

$$\begin{aligned}\hat{\theta}_{MGHT} &= \mathbf{y}'\mathbf{\Delta}_s\mathbf{\Sigma}^{-1}\mathbf{\Delta}_s(E_p(\mathbf{\Delta}_s\mathbf{\Sigma}^{-1}\mathbf{\Delta}_s))^{-1}\mathbf{c} \\ &= \mathbf{y}'\mathbf{\Delta}_s\mathbf{\Sigma}^{-1}\mathbf{\Delta}_s(\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})^{-1}\mathbf{c}.\end{aligned} \tag{35}$$

If $\mathbf{\Sigma}$ must be estimated, it can be shown that an asymptotically equivalent estimator could be obtained by replacing $\mathbf{\Sigma}$ with $\hat{\mathbf{\Sigma}}$ in (35), if $\hat{\mathbf{\Sigma}}$ satisfies the conditions of Result 3. Contrary to $\hat{\theta}_{GHT}$, which relies on the computation of $\mathbf{Q}_{\mathbf{\Sigma}}$, $\hat{\theta}_{MGHT}$ is readily given by a closed-form formula. It is seen that $\hat{\theta}_{MGHT}$ is an estimator; it does not depend on unobserved values of $\mathbf{y}$. Knowledge of $\mathbf{\Pi}$ is required, thus two-phase sampling for example, may be problematic. If $\mathbf{\Sigma}$ is diagonal, then $\hat{\theta}_{MGHT} = \hat{\theta}_{GHT} = \hat{\theta}_{HT}$. Like $\hat{\theta}_{GHT}$, $\hat{\theta}_{MGHT}$ is unbiased. Also, a closed-form formula can be given for its variance:

$$\begin{aligned}V_p\left(\hat{\theta}_{MGHT}\right) &= ||(\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})^{-1}\mathbf{c} \otimes \mathbf{y}||^2_{V_p(\mathrm{vec}(\mathbf{\Delta}_s\mathbf{\Sigma}^{-1}\mathbf{\Delta}_s))} \\ &= ||(\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})^{-1}\mathbf{c} \otimes \mathbf{y}||^2_{V_p\left(\mathrm{diag}(\mathrm{vec}(\mathbf{\Sigma}^{-1}))\mathbf{\Delta}_s^{(2)}\mathbf{1}_{N^2\times 1}\right)} \\ &= ||\mathrm{diag}(\mathrm{vec}(\mathbf{\Sigma}^{-1}))((\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})^{-1}\mathbf{c} \otimes \mathbf{y})||^2_{\mathbf{\Pi}^{(4)}-\mathrm{vec}(\mathbf{\Pi})(\mathrm{vec}(\mathbf{\Pi}))'},\end{aligned} \tag{36}$$

where $\mathbf{\Delta}_s^{(2)} = \mathrm{diag}(\mathrm{vec}(\mathbf{\Delta}_s\mathbf{1}_{N\times N}\mathbf{\Delta}_s))$, and $\mathbf{\Pi}^{(4)} = E\left(\mathbf{\Delta}_s^{(2)}\mathbf{1}_{N^2\times N^2}\mathbf{\Delta}_s^{(2)}\right)$ is a matrix of fourth-order inclusion probabilities.

Noting $\hat{\theta}_{MGHT} = \mathbf{y}'\mathbf{w}_{s\,MGHT}$, the calibration problem could now be changed in order to find weights as close as possible to $\mathbf{w}_{s\,MGHT}$, instead of $\mathbf{w}_{s\,GHT}$. The resulting estimator would be

$$\begin{aligned}\hat{\theta}_{MGCAL} &= \mathbf{y}'\mathbf{w}_{s\,MGCAL} \\ &= \hat{\mathbf{y}}'\mathbf{c} + (\mathbf{y} - \hat{\mathbf{y}})'\mathbf{w}_{s\,MGHT}.\end{aligned} \tag{37}$$

The estimator $\hat{\theta}_{MGCAL}$ is asymptotically unbiased and

$$V_p\left(\hat{\theta}_{MGCAL}\right) \doteq ||\mathrm{diag}(\mathrm{vec}(\mathbf{\Sigma}^{-1}))((\mathbf{\Sigma}^{-1} \circ \mathbf{\Pi})^{-1}\mathbf{c} \otimes (\mathbf{y} - \mathbf{y}^*))||^2_{\mathbf{\Pi}^{(4)}-\mathrm{vec}(\mathbf{\Pi})(\mathrm{vec}(\mathbf{\Pi}))'}. \tag{38}$$

Of course, it is not expected that $\hat{\theta}_{MGCAL}$ will attain the lower bound given in (11). However, it does not rely on generalized inverses, and does not require the computation of $\mathbf{Q}_{\mathbf{\Sigma}}$.

8. Estimator Comparison

In this section, six estimators are compared: the Horvitz-Thompson estimator, the calibration estimator, the generalized Horvitz-Thompson estimator, the generalized calibration estimator, and the modified versions of the latter two as described in Section 7. All estimators are, at least asymptotically, unbiased. For comparing their variance, a population of 1,000 units, 200 clusters of five units each, was used. The variable of interest was generated from a normal distribution with mean 10 and variance 2. This was done in such a way that units from the same cluster have a covariance of one, whereas units from different clusters are independent. The parameter to be estimated is the population mean.

The variance, or asymptotic variance, of the generalized estimators and of the modified generalized estimators was computed with a block diagonal matrix $\mathbf{\Sigma}$, with each 5×5 block having the value 2 on the diagonal and 1 offdiagonal, thus reflecting the distribution used to generate the population. The variances were computed under two sampling plans: simple random sampling, and Poisson sampling with the five units from a block being selected with probability (0.15, 0.15, 0.2, 0.2, 0.3). It was assumed that $\mathbf{X} = \mathbf{1}_{N\times 1}$ for computing the asymptotic variances of the calibration estimator, the generalized calibration estimator, and the modified generalized calibration estimator. That is, the only calibration equation is the one specifying that the sum of the weights should equal the population size – the trivial calibration equation. The value of $V_p(\mathrm{vec}((\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}))$, needed to compute (5) and (26), was approximated by computing $(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}$ for 10,000 different samples, all drawn according to the appropriate sampling plan: simple random sampling or Poisson sampling. With the covariance matrix used for generating the population, the lower bound given in (11) equals 0.0070 under simple random sampling, and 0.0075 under Poisson sampling.

Table 1 gives the variances of the estimators, or their asymptotic variances in the case of calibrated estimators, under the sampling plan. The table shows that, for simple random sampling, the generalized Horvitz-Thompson estimator is much less precise than the regular Horvitz-Thompson estimator. The explanation for this was given at the beginning of Section 5. The generalized Horvitz-Thompson estimator was not meant to be optimal; its interest lies in relation (7). In contrast, the generalized calibration estimator outperforms the Horvitz-Thompson estimator. Note that under simple random sampling with $\mathbf{X} = \mathbf{1}_{N\times 1}$, the calibration estimator is equal to the Horvitz-Thompson estimator. The asymptotic variance under the sampling plan, $V_p\left(\hat{\theta}_{GCAL}\right)$, is very close to the generalized Godambe-Joshi lower bound for $E_{\xi}V_p\left(\hat{\theta}\right)$, which for this $\mathbf{\Sigma}$ and sampling plan is equal to 0.0070. The performance of the modified generalized Horvitz-Thompson estimator can be significantly different from that of the generalized Horvitz-Thompson estimator. It cannot be seen as a good approximation of the generalized Horvitz-Thompson estimator. Nevertheless, the modified generalized calibration estimator performs better than the Horvitz-Thompson estimator and practically as well as the generalized calibration estimator for both simple random sampling and the Poisson sampling plan.

Table 1. Variances of six estimators of the mean.

Estimator	Simple random sampling	Poisson sampling
Horvitz-Thompson	0.0077	0.4458
Calibration	–	0.0084
Generalized Horvitz-Thompson	0.0477	0.2848
Generalized calibration	0.0069	0.0073
Modified generalized Horvitz-Thompson	0.0361	0.3237
Modified generalized calibration	0.0070	0.0076
Generalized Godambe-Joshi lower bound	0.0070	0.0075

With the Poisson sample being of random size, it is not surprising that the noncalibrated estimators (Horvitz-Thompson, generalized Horvitz-Thompson, and modified generalized Horvitz-Thompson) are performing poorly with this sampling plan. The generalized calibration estimator outperforms the calibration estimator. Its asymptotic variance is comparable to the generalized Godambe-Joshi lower bound for $E_{\xi}V_{p}\left(\hat{\theta}\right)$, which for this $\boldsymbol{\Sigma}$ and sampling plan is equal to 0.0075.

Since the calibration estimator is the generalized calibration estimator computed with a diagonal matrix $\boldsymbol{\Sigma}$, the asymptotic variance for the calibration estimator in Table 1 shows what can happen if the generalized version is used with a matrix $\boldsymbol{\Sigma}$ different from the true variance matrix, $V_{\xi}(\mathbf{y})$, . . . the generalized calibration estimator could become the ordinary calibration estimator.

9. Application to the Canadian Reverse Record Check Survey

The Reverse Record Check (RRC) is a Canadian postcensal undercoverage survey used in conjunction with the Census of Population and a postcensal overcoverage study to arrive at population estimates; see Statistics Canada (2015). In this section, the estimates and the methodology used for the Canadian Territory of Yukon for the 2011 RRC are examined, and the generalized estimates are compared to the current one. A list frame of persons is sampled with stratified random sampling. There is one large take-all stratum, where all units are enumerated by the Census, and take-some strata comprising units that are either enumerated by the Census, missed by the Census, or out of scope for the Census. The main objective of the RRC is to estimate the number of missed persons. Information on the frame is available to group persons by household. Because of the Census methodology, the variable “missed”, which takes the value 1 if the person is missed and 0 if not, is highly correlated for persons belonging to the same household. If the Census enumerated (or missed) someone, it likely enumerated (or missed) the other members of that household. With the current RRC methodology, the Horvitz-Thompson weight of responding units is multiplied by a factor to account for nonresponse and a factor to account for frame undercoverage. The latter factor is such that the estimate of persons enumerated coincides with the equivalent Census number. In this application, the Horvitz-Thompson weights are replaced with the generalized weights. For the generalized calibrated weights, the calibration equation simply ensured that the sum of the calibrated weights equalled the

Table 2. RRC estimates of missed persons.

Estimator	Estimate	Variance estimate
Horvitz-Thompson	5,272	91,727
Generalized Horvitz-Thompson	5,150	82,920
Generalized calibration	5,137	82,505
Modified generalized Horvitz-Thompson	5,194	86,945
Modified generalized calibration	5,173	85,999

number of units in the stratum. The same nonresponse adjustment factors were used, and the frame undercoverage adjustment factors were all computed so that the estimates of persons enumerated coincides with the equivalent Census number: 29,982. The generalized weights were computed assuming that the correlation structure is block diagonal, each block representing the persons of a same household, according to the frame information. All offdiagonal elements of each block are set to 0.95. This is because the estimates of the correlations within households are 0.956 for the variable "missed" and 0.947 for the variable "enumerated". The correlations are less than one because the Census sometimes partly enumerates or partly misses a household, and because the frame household may differ from the census household.

The estimates obtained along with the corresponding variance estimates are given in Table 2. The Horvitz-Thompson estimates are those currently used by the survey. The variance estimate of the generalized Horvitz-Thompson estimator is lower than that of the Horvitz-Thompson estimator, in spite of having an additional source of variance, as discussed at the beginning of Section 5. This is because, for all estimators, the last step in computing the estimates is a calibration on the number of persons enumerated. The variance estimates of the generalized calibration estimator and the modified generalized calibration estimator, those that would be used in practice, are lower than that of the Horvitz-Thompson estimator. The advantage of the modified generalized estimator is that its computation did not require calculating the matrix $\mathbf{Q}_{\boldsymbol{\Sigma}}$, although it was easy to approximate this matrix by repeatedly sampling the frame one million times, and using the method described at the beginning of Section 6. The estimates of missed persons are not significantly different from one another.

10. Variance Estimation

Statisticians are better at estimating totals or weighted totals than they are at estimating variances. Why not write variances in the form of weighted totals? A variance can be written in the form $\theta_{\text{var}} = \mathbf{y}'\mathbf{V}\mathbf{y} = (\mathbf{y} \otimes \mathbf{y})'\text{vec}(\mathbf{V})$, with $\mathbf{V} \in \mathbb{R}^{N \times N}$. Such parameters have been estimated in Sections 3, 5, and 7. There are N^2 units, each corresponding to a pair of units of the original population, with a vector of interest equal to $\mathbf{y} \otimes \mathbf{y}$ and $\mathbf{c} = \text{vec}(\mathbf{V})$. The methods of this article apply here, because in general, $Cov(y_i y_j,\ y_k y_l) \neq 0$ for $(i,\ j) \neq (k,\ l)$. Whatever the asymptotic setup, $V_\xi(\mathbf{y} \otimes \mathbf{y}) = \boldsymbol{\Sigma}_2$ will not be a positive definite matrix. For example, the row of $\boldsymbol{\Sigma}_2$ which corresponds to unit (i, j) $i \neq j$ is equal to the row which corresponds to unit (j, i). In fact, the event $(i, j) \in s$ is identical to the event $(j,\ i) \in s$. Thus, from the N^2 units, only the $N(N+1)/2$ with $j \geq i$ need to be kept. The estimators suggested in this article would require the inversion of a matrix of order $N(N+1)/2$.

The variance matrix will be diagonal, $Cov(y_iy_j, y_ky_l) = 0$ for $(i, j) \neq (k, l)$, if $V_\xi(\mathbf{y})$ is diagonal and if $E_\xi(\mathbf{y}) = \mathbf{0}$. The last assumption is reasonable if $\mathbf{y}$ is a vector of residues, as would be the case if one is estimating the variance of a calibration estimator (regular, generalized, or modified). With $Cov(y_iy_j, y_ky_l) = 0$ for $(i, j) \neq (k, l)$, the regular calibration estimator will suffice to estimate the variance of a calibrated estimator. A choice of calibration equations must still be made. The findings made in Théberge (1999) remain valid; namely, to use an auxiliary variables matrix in a block diagonal form with two blocks: a trivial model for the cross products terms of $\theta_{\text{var}} = \mathbf{y}'\mathbf{V}\mathbf{y}$, that is, $\mathbf{y}'(\mathbf{V} - (\mathbf{V} \circ \mathbf{I}))\mathbf{y}$ (this will yield a Horvitz-Thompson estimator in the case of fixed-size sampling plans) and a nontrivial model to estimate the squared terms of $\theta_{\text{var}} = \mathbf{y}'\mathbf{V}\mathbf{y}$, that is, $\mathbf{y}'(\mathbf{V} \circ \mathbf{I})\mathbf{y}$. In the examples examined in Théberge (1999), the nontrivial model used for estimating the squared terms of the variance was the one corresponding to a ratio estimator.

Whatever the models are, it is important to use an auxiliary variables matrix in a block diagonal form with two blocks. For example, the intercept used to estimate the cross-product terms has nothing to do with a possible intercept to estimate the squared terms. Therefore there should not be an auxiliary variable taking the value one for all $(i,j)\, j \geq i$. It is preferable to have an auxiliary variable taking the value one for all $(i,j)\, j > i$ and zero otherwise, and another auxiliary variable taking the value one for $(i, i)\; i = 1, 2, \ldots, N$ and zero otherwise.

11. Conclusion

An asymptotic setup is necessary to discuss the asymptotic properties of the estimators. The setup used here integrates a superpopulation model. There is no need for a superpopulation model separate from the asymptotic setup. The setup's model does not assume that the units are uncorrelated.

Even the Horvitz-Thompson estimator can be viewed as relying on a model. The generalized Horvitz-Thompson estimator, like the Horvitz-Thompson estimator, is unbiased. Both estimators, but especially the former, can be affected by the variance in the effective sample weight. Even without auxiliary data, it is possible to calibrate the weights so their total equals the population size. If this is done, then the generalized estimator will have a lower asymptotic variance than the ordinary estimator.

The calibration estimator was generalized in two ways: firstly, one is seeking weights close to the generalized Horvitz-Thompson weights; secondly, the matrices $\mathbf{T}$ and $\mathbf{U}$, used in measuring distances, need no longer be diagonal.

A somewhat easier way to compute the modified generalized calibration estimator was shown to perform practically as well as the generalized calibration estimator in the examples given in this article.

The Godambe-Joshi lower bound can be generalized to the case where the units are correlated. The asymptotic variance of the generalized calibration estimators attains the generalized Godambe-Joshi lower bound, if the model is correct, that is, if $\hat{\mathbf{\Sigma}} \rightarrow \mathbf{\Sigma}$ in probability and the matrix $\mathbf{X}$ used by the generalized calibration estimator agrees with that of the asymptotic setup. This is regardless of the choice for the matrix $\mathbf{U}$, used in measuring the distance between the calibrated weights and the generalized Horvitz-Thompson weights.

By viewing variances as weighted totals, the theory developed here provides a framework for variance estimation. The general case would require the inversion of very large matrices, but there are simplifications to be made if one is estimating the variance of a calibration estimator. Those simplifications will often result in what was called the "hybrid estimator" in Théberge (1999).

Even though there are workarounds, such as dropping variables or using the limit of a positive definite matrix, it would be interesting to generalize the results of this article to the case of $\mathbf{\Sigma}$ positive semi-definite. This strategy of using the methods of this article with a positive definite covariance matrix that differs only infinitesimally from a block diagonal matrix where each block equals $\mathbf{1}_{2\times2}$ will allow concluding by revisiting the example in the introduction. The assumption that the correlation between y_{2i-1} and y_{2i} $i = 1, 2, \ldots, N/2$ is 1, is weaker than the assumption that the two units are equal, as was done in the introduction. The resulting generalized Horvitz-Thompson estimator of the total is $\sum_{i=1}^{N/2} 2y_{2i-1}\delta_{2i-1} + 2y_{2i}\delta_{2i} - (y_{2i-1} + y_{2i})\,\delta_{2i-1}\delta_{2i} + (y_{2i-1} - y_{2i})\,\delta_{2i-1}\delta_{2i}(\pi_{2i} - \pi_{2i-1})/ \pi_{2i-1} + \pi_{2i} - \pi_{2i-1\ \ 2i}$. Setting $y_1 = y_2$ will yield the term given in the introduction. For example, for *N* spouses grouped into *N*/2 couples, the variable of interest may be the place of residence (very high correlation), or education level (significant correlation). Using a calibrated version of this estimator to ensure that the sum of the weights equals the population size, will be preferable to using a similarly calibrated version of the Horvitz-Thompson estimator, if the correlation is somewhat close to 1. Using an estimator optimized for a correlation of 1 will often be preferable to using an estimator optimized for a correlation of 0. Note that two samples drawn with the same sampling plan made up of individuals from the same couples will have the same effective sample weight, regardless of how many of the spouses, one or two, are sampled from each observed couple.

Appendix A: Proofs of Results of Section 2 and Lemmas of Section 3

Proof of Result 1:

Note that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_t &= \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'_t(\mathbf{\Delta}_t\mathbf{U}_t\mathbf{\Delta}_t)^{\dagger}\mathbf{X}_t\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'_t(\mathbf{\Delta}_t\mathbf{U}_t\mathbf{\Delta}_t)^{\dagger}\mathbf{y}_t \\
&= \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\sum_{i=1}^{t}\mathbf{X}'\left(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]}\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\sum_{i=1}^{t}\mathbf{X}'\left(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]}\right)^{\dagger}\mathbf{y}_{[i]} \\
&= \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'t^{-1}\sum_{i=1}^{t}(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]})^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'t^{-1}\sum_{i=1}^{t}(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]})^{\dagger}\mathbf{y}_{[i]}.
\end{aligned}$$

The $\mathbf{\Delta}_{[i]}$ being independent and identically distributed, from the weak law of large numbers, $t^{-1}\sum_{i=1}^{t}(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]})^{\dagger}$ tends in probability to $\mathbf{Q}_{\mathbf{U}}^{-1} = E_p((\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]})^{\dagger})$. Since the sampling plan is noninformative, $t^{-1}\sum_{i=1}^{t}(\mathbf{\Delta}_{[i]}\mathbf{U}\mathbf{\Delta}_{[i]})^{\dagger}\mathbf{y}_{[i]}$ tends in probability to $\mathbf{Q}_{\mathbf{U}}^{-1}\mathbf{X}\boldsymbol{\beta}$.

Thus,

$$\hat{\boldsymbol{\beta}}_t \to \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\mathbf{Q}_{\mathrm{U}}^{-1}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'\mathbf{Q}_{\mathrm{U}}^{-1}\mathbf{X}\boldsymbol{\beta}$$

$$= \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\mathbf{Q}_{\mathrm{U}}^{-1}\mathbf{X}\mathbf{T}^{1/2}\right)^{-1}\mathbf{T}^{1/2}\mathbf{X}'\mathbf{Q}_{\mathrm{U}}^{-1}\mathbf{X}\mathbf{T}^{1/2}\mathbf{T}^{-1/2}\boldsymbol{\beta}$$

$$= \boldsymbol{\beta}$$

in probability.

Proof of Result 2:

Note that

$$\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t\mathbf{c}_t = t^{\gamma}\mathbf{X}'\left(\left(t^{-1}\sum_{i=1}^{t}\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}\right)\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}} - \mathbf{I}_N\right)\mathbf{c}. \qquad (A.39)$$

The expectation, under the plan p, of $\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}$ is equal to $\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}^{-1}$. Also, the variance of the vector $\mathbf{X}'\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}\mathbf{c}$ is finite. Indeed, there is a finite number of possible values for $\boldsymbol{\Delta}_{[i]}$, and for any $\boldsymbol{\Delta}_{[i]}$, $\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}$ exists. According to the central limit theorem, $\left(\mathbf{X}'\left(t^{-1}\sum_{i=1}^{t}\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}\right)\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}\mathbf{c} - \mathbf{X}'\mathbf{c}\right)t^{1/2}$ converges to a normal distribution with mean $\mathbf{0}_{q\times 1}$ and finite variance. Since $\left(\mathbf{X}'\left(t^{-1}\sum_{i=1}^{t}\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}\right)\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}}\mathbf{c} - \mathbf{X}'\mathbf{c}\right)t^{1/2}$ is $O_p(1)$, from (A.39) it follows that $\mathbf{X}'_t\left(\boldsymbol{\Delta}_t\hat{\boldsymbol{\Sigma}}_t\boldsymbol{\Delta}_t\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}_t}\mathbf{c}_t - \mathbf{X}'_t\mathbf{c}_t$ is $O_p(t^{\gamma-1/2})$.

Proof of Result 3:

The difference can be written as a sum of differences $t^{\gamma}\mathbf{X}'\left(t^{-1}\sum_{i=1}^{t}\left(\left(\boldsymbol{\Delta}_{[i]}\hat{\boldsymbol{\Sigma}}_{[i]}\boldsymbol{\Delta}_{[i]}\right)^{\dagger}\mathbf{Q}_{\hat{\boldsymbol{\Sigma}}} - (\boldsymbol{\Delta}_{[i]}\boldsymbol{\Sigma}_{[i]}\boldsymbol{\Delta}_{[i]})^{\dagger}\mathbf{Q}_{\boldsymbol{\Sigma}}\right)\right)\mathbf{c}$. The differences under the summation sign tend to $\mathbf{0}$ in probability and the central limit theorem yields the result. Note that the condition of $\hat{\boldsymbol{\Sigma}}_t$ being positive definite is needed to ensure the Moore-Penrose inverse is continuous at $\boldsymbol{\Sigma}_t$ (see Ben-Israel and Greville 2002, 212).

Proof of Lemma 1:

It is obvious that $\mathcal{N}(\mathbf{F}) \supseteq \cap_{i=1}^{K}\mathcal{N}(\mathbf{F}_i)$. To show that $\mathcal{N}(\mathbf{F}) \subseteq \cap_{i=1}^{K}\mathcal{N}(\mathbf{F}_i)$, let $\mathbf{v} \in \mathrm{N}(\mathbf{F})$, then $\mathbf{v}'\mathbf{F}\mathbf{v} = \sum_{i=1}^{K}\mathbf{v}'\mathbf{F}_i\mathbf{v} = 0$. The matrices $\mathbf{F}_i$ being positive semi-definite, one must have $\mathbf{v}'\mathbf{F}_i\mathbf{v} = 0 \;\; i = 1, 2, \ldots, K$. With $\mathbf{F}_i$ also being symmetric, there exists a symmetric positive semi-definite matrix $\mathbf{K}_i$ such that $\mathbf{F}_i = \mathbf{K}_i^2$. Therefore $\mathbf{v}'\mathbf{F}_i\mathbf{v} = 0$ implies $\mathbf{v}'\mathbf{K}'_i\mathbf{K}_i\mathbf{v} = 0$ and one must have $\mathbf{v} \in \mathcal{N}(\mathbf{F}_i)$.

Proof of Lemma 2:

Writing $\mathcal{R}(\mathbf{F})$ for the range of a matrix $\mathbf{F}$, from Ben-Israel and Greville (2002, Exercise 2.38) it is known that $\mathcal{R}((\mathbf{FP})^\dagger) = \mathcal{R}(\mathbf{PF}') \subseteq \mathcal{R}(\mathbf{P})$, which proves a). The proof of b) is obtained by taking the transpose of each side of the identity in a). Finally, from using a) and b) in succession it follows that $(\mathbf{PFP})^\dagger = \mathbf{P}(\mathbf{PFP})^\dagger = \mathbf{P}(\mathbf{PFP})^\dagger\mathbf{P}$.

Appendix B: Counter-example

To estimate a population total ($\mathbf{c} = \mathbf{1}$), an estimator $\hat{\theta} = \hat{\theta}_{GCAL} + \hat{0}$, where $\hat{0}$ is a nonlinear unbiased estimator of 0, will be used. The computation of $\hat{\theta}_{GCAL}$ will use $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}$. It will be shown that for the asymptotically equivalent random variable $\theta^* = \overset{*}{\theta}_{GCAL} + \hat{0}$, one has $E_\xi V_p(\theta^*) < E_\xi V_p(\overset{*}{\theta}_{GCAL}) = \text{GGJLB}$, where GGJLB is the generalized Godambe-Joshi lower bound.

Starting with Equation (9)

$$\begin{aligned} E_\xi V_p(\theta^*) &= E_p V_\xi(\theta^*) + E_p(E_\xi(\theta^* - \theta))^2 - V_\xi(\theta) \\ &= E_p V_\xi\left(\overset{*}{\theta}_{GCAL}\right) + E_p V_\xi(\hat{0}) + 2E_p Cov_\xi\left(\overset{*}{\theta}_{GCAL}, \hat{0}\right) + E_p(E_\xi(\theta^* - \theta))^2 - V_\xi(\theta) \\ &= E_p V_\xi(\hat{\theta}_{GHT}) - V_\xi(\theta) + E_p V_\xi(\hat{0}) + 2E_p Cov_\xi\left(\overset{*}{\theta}_{GCAL}, \hat{0}\right) + E_p\left(E_\xi(\hat{0})\right)^2 \\ &= \text{GGJLB} + E_p V_\xi(\hat{0}) + 2E_p Cov_\xi\left(\overset{*}{\theta}_{GCAL}, \hat{0}\right) + E_p\left(E_\xi(\hat{0})\right)^2. \end{aligned}$$

In this example, the population $U = \{1, 2, 3\}$. Under the sampling plan, the samples $s_1 = \{1, 2\}$ and $s_2 = \{2, 3\}$ can each be selected with probability 0.5. The vector $\mathbf{u} \in \mathbb{R}^3$ is composed of independent and identically distributed variables taking the values 1 or -1 each with probability 0.5. Under the model ξ, the vector of interest is $\mathbf{y} = \mathbf{\Sigma}^{1/2}\mathbf{u}$, with $\mathbf{\Sigma}^{1/2} = \begin{pmatrix} 2 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Thus, with this model, $E_\xi(\mathbf{y}) = \mathbf{0}$, $V_\xi(\mathbf{y}) = \mathbf{\Sigma} = \begin{pmatrix} 4.25 & 1.5 & 0 \\ 1.5 & 1.25 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{Q}_\mathbf{\Sigma} = \begin{pmatrix} 6.7 & 1.5 & 0 \\ 1.5 & 1.25 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

A nonlinear unbiased, under the sampling plan, estimator of 0 is $\hat{0} = -[y_2]$ if s_1 is selected and $\hat{0} = [y_2]$ if s_2 is selected, where $[y_2]$ represents the integer part of y_2.

Under those conditions, $\text{GGJLB} = \mathbf{c}'(\mathbf{Q}_\mathbf{\Sigma} - \mathbf{\Sigma})\mathbf{c} = 3.45$, $E_p V_\xi(\hat{0}) = 0.5$, $E_\xi(\hat{0}) = 0$, $Cov_\xi(\mathbf{y}, [y_2]) = \begin{pmatrix} 1.25 \\ 0.75 \\ 0 \end{pmatrix}$,

$E_p Cov_\xi\left(\overset{*}{\theta}_{GCAL}, \hat{0}\right) = 0.5 \times \mathbf{c}'\mathbf{Q}_\mathbf{\Sigma}((\mathbf{\Delta}_{s_2}\mathbf{\Sigma}\mathbf{\Delta}_{s_2})^\dagger - (\mathbf{\Delta}_{s_1}\mathbf{\Sigma}\mathbf{\Delta}_{s_1})^\dagger)Cov_\xi(\mathbf{y}, [y_2]) = -0.35$ and

$$\begin{aligned} E_\xi V_p(\theta^*) &= \text{GGJLB} + E_p V_\xi\big(\hat{0}\big) + 2E_p Cov_\xi\Big(\theta^*_{GCAL}, \hat{0}\Big) + E_p\big(E_\xi\big(\hat{0}\big)\big)^2 \\ &= 3.45 + 0.5 - 0.7 + 0 \\ &= 3.25 \\ &< \text{GGJLB}. \end{aligned}$$

Thus, asymptotically, $E_\xi V_p\big(\hat{\theta}_{GCAL} + \hat{0}\big) < E_\xi V_p\big(\hat{\theta}_{GCAL}\big) = \text{GGJLB}$. With $N = 3$, the asymptotic properties are not very meaningful. However the example could be expanded to include a large number of strata of size 3, each with the model and sampling plan described above.

Appendix C: Proof that (13) Satisfies (12)

To simplify, $\mathbf{P}_{\mathcal{N}(\mathbf{A}'\mathbf{TA})}$ will be denoted by $\mathbf{P}$. First, using the first part of (13), set $\mathbf{G} = (\mathbf{I}_N - (\mathbf{PUP})^\dagger\mathbf{U})(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{T}$.

Because for an arbitrary real square matrix $\mathbf{M}$, $\mathcal{N}(\mathbf{MM}') = \mathcal{N}(\mathbf{M}) = \mathcal{N}(\mathbf{M}')$, it follows that

$$\begin{aligned} \mathbf{UGA} &= \mathbf{U}(\mathbf{I}_N - (\mathbf{PUP})^\dagger\mathbf{U})\,(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{TA} \\ &= \mathbf{U}(\mathbf{I}_N - (\mathbf{PUP})^\dagger\mathbf{U})(\mathbf{I}_N - \mathbf{P}) \\ &= \mathbf{U} - \mathbf{U}(\mathbf{PUP})^\dagger\mathbf{U} - \mathbf{UP} + \mathbf{UP}(\mathbf{PUP})^\dagger(\mathbf{PUP}) \\ &= \mathbf{U} - \mathbf{U}(\mathbf{PUP})^\dagger\mathbf{U} - \mathbf{UP} + \mathbf{UP}(\mathbf{I}_N - \mathbf{P}_{\mathcal{N}(\mathbf{PUP})}) \\ &= \mathbf{U} - \mathbf{U}(\mathbf{PUP})^\dagger\mathbf{U} - \mathbf{UPP}_{\mathcal{N}(\mathbf{PUP})} \\ &= \mathbf{U} - \mathbf{U}(\mathbf{PUP})^\dagger\mathbf{U}, \end{aligned} \tag{C.1}$$

which is symmetrical. Also symmetrical, is

$$\begin{aligned} \mathbf{TAG} &= \mathbf{TA}(\mathbf{I}_N - (\mathbf{PUP})^\dagger\mathbf{U})(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{T} \\ &= (\mathbf{TA} - \mathbf{TAP}(\mathbf{PUP})^\dagger\mathbf{U})(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{T} \\ &= \mathbf{TA}(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{T}, \end{aligned} \tag{C.2}$$

because $\mathbf{TAP} = \mathbf{0}$. From (C.2),

$$\begin{aligned} \mathbf{TAGA} &= \mathbf{TA}(\mathbf{A}'\mathbf{TA})^\dagger\mathbf{A}'\mathbf{TA} \\ &= \mathbf{TA}(\mathbf{I}_N - \mathbf{P}) \\ &= \mathbf{TA}. \end{aligned} \tag{C.3}$$

Finally, from (C.1),

$$\begin{aligned}\mathbf{UGAG} &= (\mathbf{U} - \mathbf{U}(\mathbf{PUP})^{\dagger}\mathbf{U})\mathbf{G}\\ &= \mathbf{UG} - \mathbf{U}(\mathbf{PUP})^{\dagger}\mathbf{U}(\mathbf{I}_N - (\mathbf{PUP})^{\dagger}\mathbf{U})(\mathbf{A}'\mathbf{TA})^{\dagger}\mathbf{A}'\mathbf{T}\\ &= \mathbf{UG} - [\mathbf{U}(\mathbf{PUP})^{\dagger}\mathbf{U} - \mathbf{U}(\mathbf{PUP})^{\dagger}\mathbf{PUP}(\mathbf{PUP})^{\dagger}\mathbf{U}](\mathbf{A}'\mathbf{TA})^{\dagger}\mathbf{A}'\mathbf{T}\\ &= \mathbf{UG}.\end{aligned} \tag{C.4}$$

If $\mathbf{G} = \mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}$ with $\mathbf{U}$ positive definite, then

$$\mathbf{UGA} = \mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\mathbf{A} \tag{C.5}$$

is a symmetrical matrix. Also,

$$\begin{aligned}\mathbf{TAG} &= \mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\\ &= \mathbf{T}^{1/2}\mathbf{P}_{\mathcal{R}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})}\mathbf{T}^{1/2}\end{aligned} \tag{C.6}$$

is symmetrical, since an orthogonal projection matrix is symmetrical. From the properties of the Moore-Penrose inverse,

$$\begin{aligned}\mathbf{UGAG} &= \mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\\ &= \mathbf{A}'\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\\ &= \mathbf{UG}.\end{aligned} \tag{C.7}$$

Finally, because for an arbitrary real matrix $\mathbf{M}$, $\mathcal{R}(\mathbf{MM}') = \mathcal{R}(\mathbf{M})$ and because $\mathbf{U}$ is positive definite, it follows that

$$\begin{aligned}\mathbf{TAGA} &= \mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}\mathbf{A}\\ &= \mathbf{T}^{1/2}\mathbf{P}_{\mathcal{R}(\mathbf{T}^{1/2}\mathbf{AU}^{-1}\mathbf{A}'\mathbf{T}^{1/2})}\mathbf{T}^{1/2}\mathbf{A}\\ &= \mathbf{T}^{1/2}\mathbf{P}_{\mathcal{R}(\mathbf{T}^{1/2}\mathbf{A})}\mathbf{T}^{1/2}\mathbf{A}\\ &= \mathbf{TA}.\end{aligned} \tag{C.8}$$

Appendix D: Example of Computing $\mathrm{P}(S_f = n_f)$ with a Block of Size 5

$$\mathrm{P}(S_f = 5) = \pi^{[5]}$$

$$\mathrm{P}(S_f = 4) = \sum_{\text{block}} \pi^{[4]} - 5\pi^{[5]}$$

$$\mathrm{P}(S_f = 3) = \sum_{\text{block}} \pi^{[3]} - 4 \sum_{\text{block}} \pi^{[4]} + 10\pi^{[5]}$$

$$P(S_f = 2) = \sum_{\text{block}} \pi^{[2]} - 3 \sum_{\text{block}} \pi^{[3]} + 6 \sum_{\text{block}} \pi^{[4]} - 10\pi^{[5]}$$

$$P(S_f = 1) = \sum_{\text{block}} \pi^{[1]} - 2 \sum_{\text{block}} \pi^{[2]} + 3 \sum_{\text{block}} \pi^{[3]} - 4 \sum_{\text{block}} \pi^{[4]} + 5\pi^{[5]}$$

$$P(S_f = 0) = 1 - \sum_{n_f=1}^{5} P(S_f = n_f)$$

12. References

Albert, A. 1973. "The Gauss–Markov Theorem for Regression Models with Possibly Singular Covariances." *SIAM Journal on Applied Mathematics* 24: 182–187.

Ben-Israel, A. and T.N.E. Greville. 2002. *Generalized Inverses: Theory and Applications (second ed.)*. New York: Springer-Verlag.

Brewer, K.R.W. 1979. "A Class of Robust Sampling Designs for Large Scale Surveys." *Journal of the American Statistical Association* 74: 911–915.

Cassel, C.-M., C.E. Särndal, and J.H. Wretman. 1977. *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons.

Deville, J.-C. and C.E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382.

Estevao, V.M. and C.E. Särndal. 2003. *A New Perspective on Calibration Estimators*. In Proceedings of the Section on Survey Research Methods, American Statistical Association, 1346–1356.

Godambe, V.P. and V.M. Joshi. 1965. "Admissibility and Bayes Estimation in Sampling Finite Populations, 1." *Annals of Mathematical Statistics* 36: 1707–1722.

Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Isaki, C.T. and W.A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77: 89–96.

Montanari, G.E. 1998. "On Regression Estimation of Finite Population Means." *Survey Methodology* 24: 69–77.

Montanari, G.E. and M.G. Ranalli. 2002. "Asymptotically Efficient Generalised Regression Estimators." *Journal of Official Statistics* 18: 577–590.

Rao, C.R. and S.K. Mitra. 1971. *Generalized Inverse of Matrices and its Applications*. New York: John Wiley & Sons.

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

Statistics Canada. 2015. *Census Technical Report: Coverage*. Statistics Canada Catalogue no. 98-303-X, Ottawa, Ontario. Available at: https://www12.statcan.gc.ca/census-recensement/2011/ref/guides/98-303-x/index-eng.cfm (accessed January 2017).

Théberge, A. 1999. “Extensions of Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 94: 635–644.

Zyskind, G. and F.B. Martin. 1969. “On Best Linear Estimation and a General Gauss–Markov Theorem in Linear Models with Arbitrary Non-negative Covariance Structure.” *SIAM Journal on Applied Mathematics* 17: 1190–1202.

Received July 2015
Revised April 2016
Accepted July 2016