Journal of Official Statistics
vol. 39, 4 (December 2023)

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 435–458, http://dx.doi.org/10.2478/JOS-2023-0021

Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach

Alejandra Arias-Salazar[1]

Obtaining reliable estimates in small areas is a challenge because of the coverage and periodicity of data collection. Several techniques of small area estimation have been proposed to produce quality measures in small areas, but few of them are focused on updating these estimates. By combining the attributes of the most recent versions of the structure-preserving estimation methods, this article proposes a new alternative to estimate and update cross-classified counts for small domains, when the variable of interest is not available in the census. The proposed methodology is used to obtain and up-date estimates of the incidence of poverty in 81 Costa Rican cantons for six postcensal years (2012–2017). As uncertainty measures, mean squared errors are estimated via parametric bootstrap, and the adequacy of the proposed method is assessed with a design-based simulation.

Key words: Extreme poverty; intercensal updating; small area estimation; log-linear models.

1. Introduction

The estimation and monitoring of socio-economic indicators is relevant for decision-making and the development of public policies aimed at improving the conditions of the citizens. Among other characteristics, high-quality statistics must be relevant, accurate, and reliable to use them in the design, development and assessment of programs of social interest (Eurostat 2017). The success of these plans depends on how they are formulated and oriented, but in many cases, the information available is not enough to achieve this objective. Traditionally, national surveys are carried out every year in many countries to produce an up-to-date status of important topics such as poverty, inequality, and unemployment. This information, which is obtained periodically, usually satisfies the quality requirements, for instance of national statistical offices only at bigger domains. In other words, due to lack of resources, the sample sizes are not large enough to study the problems of interest in detail. For example, in the case of poverty: Where is the most vulnerable population located? Which areas have been improved through the years and which areas have been stagnated? Which other conditions (e.g., sex, age, disabilities) are associated with this phenomenon, and in which local areas?

[1]University of Costa Rica, Ciudad Universitaria Rodrigo Facio, San Pedro, Montes de Oca, San José 11501-2060, Costa Rica. Email: alejandra.ariassalazar@ucr.ac.cr

Acknowledgments: The elaboration of this article was only possible thanks to the data provided by the National Statistical Institute of Statistics and Censuses from Costa Rica. The data access was provided under specific confidentiality conditions. The numerical results are not official estimates and are only produced for illustrating the methods. I would also like to deeply thank Dr. Marissa Cinco Isidro and Dr. Ángela Luna Hernández for their kind collaboration. The completion of this article could not have been accomplished without their research and the codes they shared with me.

Small area estimation (SAE) methods have the goal of producing reliable estimates in smaller domains, that is, with adequate precision. Most of these methodologies, usually classified as unit-or area-level models, provide efficiency gains if the correlation between existing auxiliary information and the survey data is sufficient. (Pfeffermann 2013; Rao and Molina 2015; Tzavidis et al. 2018). In middle-income countries, administrative records are usually not sound enough and therefore censuses are the most important auxiliary source of information for the entire population, with the limitation that it is usually collected every ten years.

The time gap between annual surveys and population censuses is usually ignored in SAE methods. The use of covariates from an earlier period may lead to less reliable indicators than what would be expected from more solid auxiliary information. Academic literature on updating estimates in small areas is limited. Post censal population estimates have been obtained, for example from traditional procedures in demography, like the component method (United Nations 1956) and vital rates (Rao 2003). Emwanu et al. (2006) use panel survey data to obtain small area estimates of welfare in post-census years by regressing recent income (or expenditure) data on household characteristics that are available in both survey and census data. The best-known tool in this field is the structure-preserving estimation (SPREE) method, which is also the focus of this article and will be described in detail in Subsection 3.1. This technique was originally introduced by Purcell and Kish (1980) to obtain post-census estimates (counts or proportions), arranged by small domains and categories of interest. SPREE have been especially applied for updating demographic information and socio-economic indicators including employment (Hidiroglou and Patak 2009; Berg and Fuller 2009; Luna-Hernández et al. 2015) and poverty (Isidro et al. 2016).

Thereafter, several versions of SPREE have been proposed with the aim of improving the method by adding flexibility and reducing bias, namely the Generalized-SPREE (GSPREE) (Zhang and Chambers 2004) and most recently, the Multivariate-SPREE (MSPREE) (Luna-Hernández 2016). These SPREE-techniques have specific assumptions and requirements for their implementation. For example, the variable of interest must be categorical, and it must be not only in the survey (most recent) but also in the census data, which for indicators like the poverty rate are usually not available for variables based on income or expenditure. An alternative version called Extended-SPREE (ESPREE) (Isidro et al. 2016) solves this problem by applying a small area estimation technique (the Elbers, Lanjouw and Lanjouw (ELL) method (Elbers et al. 2003), as a previous step to obtain the required information for the census year. Once the estimated census information is obtained, Isidro et al. (2016) perform the original SPREE (Purcell and Kish 1980) to compute post-censal estimates. Moreover, Luna-Hernández (2016) showed that MSPREE is more efficient compared to SPREE (in terms of lower mean squared errors). Therefore, the current article extends the framework of Isidro et al. (2016) by allowing for the MSPREE in the updating process. In particular, the article aims to provide a modern methodology to (a) estimate and (b) update counts or proportions of relevant indicators in small areas when the information of interest is not available in the census data.

The motivation of the proposal is to offer updated reliable income-based poverty estimates of Costa Rican cantons in three mutually exclusive categories: “extreme poor”, “poor” (not extreme), and “not poor”. Due to its political stability and good performance in

general macroeconomic aspects, Costa Rica has been for several years an example among other economies in the region (OECD 2016). Despite this, a point that draws attention and has been the object of study in recent years is the stagnation of relative poverty that the country has had for more than two decades, unlike other Latin American countries that have achieved greater reductions in their poverty rates (CEPAL and MIDEPLAN 2016). As well as in the international agenda, previously through the Millennium Development Goals (MDGs) (United Nations 2019a) and currently, the Sustainable Development Goals (SDGs) (United Nations 2019b), one of the main concerns specifically in this country is the extreme poverty. Traditionally, and for international comparison, the National Institute of Statistic and Censuses (INEC, Instituto National de Estadıstica y Censos) of Costa Rica measures poverty based on the poverty line method. With this approach, a person or household is considered poor (or extreme poor) if its monthly per capita income is equal or below a specific poverty line. The idea of defining a threshold or line is to set the minimum amount in the per capita income that a person or household requires to satisfy food and non-food needs, included in a basket of goods and services (INEC 2015). In Costa Rica, extreme poverty had a reduction from 2004 to 2013 of only 0.8 percentage points in accordance with a diagnosis of structural gaps. Research showed that this status of poverty has three determinants: the adverse home and social environment, the insufficient scope of social programs and the exclusive labor market (CEPAL and MIDEPLAN 2016). Because of a lack of data, this kind of studies can only be conducted in census years or for larger areas, limiting the possibility of applying targeted policy interventions for specific groups or domains.

As well as other middle-income economies, Costa Rica faces several limitations to obtain small area estimates of poverty: administrative records at the unit level are not available, the census does not contain income or expenditure information to compute poverty estimates via the poverty line method, and the census is carried out only every ten years which can reasonably lead to outdated poverty estimates. The main study previously conducted in Costa Rica to obtain estimates of poverty in local areas was carried out for the same year as the census, using the ELL method. Although this work, developed by Méndez and Bravo (2011), certainly allowed to obtain more detailed information about poverty in local areas (classified as poor and not poor), two aspects can be improved with the proposal presented in this article: (a) provide poverty estimates for non-censal years, and (b) produce estimates on extreme poverty as a specific group of interest. The methodology proposed in this article can also be applied to many other countries that share similar conditions and extended to other relevant demographic and socio-economic (categorical) indicators.

This article has the following structure: Section 2 describes the data sets and explains the definition of poverty used in the application. Section 3 introduces the SPREE methods as small area estimation and updating techniques. The strategy proposed to obtain and update poverty estimates in Costa Rican cantons is also explained in this section, as well as the methodology to produce uncertainty measures. Application results are shown in Section 4. The results of a simulation study to validate the proposed method are presented in Section 5. The last section is dedicated to the conclusions and recommendations for further research.

2. Data Description and Definition of Poverty

As will be explained in detail in Subsection 3.1, the basic approach of the SPREE techniques requires one complete composition (usually a cross-classification from a census gathered in a previous time) and updated, reliable population totals (margins) for the variable of interest and for the area population sizes. Extensions to this methodology including the one implemented here require of an updated estimate of the cross-classification of interest that can be obtained from survey data. The aim of this section is to describe the data sources available and explain the definition of poverty considered in the application. Population and housing census, as well as the National Household Survey data sets, were provided by the INEC of Costa Rica, under specific confidential agreements.

2.1. Population and Housing Census 2011

In Costa Rica, the Population and Housing Censuses are carried out every ten years by the INEC. The most recent census was conducted in 2011 (data collection from 30th, May to 3rd, June) and it collected information of people, households, and dwellings necessary for the planning, execution, and evaluation of public policies (INEC 2012). With the information collected, it is possible to identify the relevant characteristics of the population such as access to education, employment, social security, technology, and health centers. Although the census 2011 includes questions to compute the unsatisfied basic needs (UBNS) index (Feres and Mancero 2001), it did not produce information about income or expenditures of the persons and households, which are necessary to calculate the incidence of poverty via the poverty line method. The sampling frame which is needed to conduct national surveys and other statistical operations is constructed based on this population and housing census. With this census, 10,461 primary sampling units (PSUs) and 1,359,168 dwellings were identified.

2.2. National Household Surveys 2011–2017

The National Household Survey (Encuesta Nacional de Hogares (ENAHO)) is the primary source for poverty and inequality measures in Costa Rica. In this aspect, this survey collects information about housing characteristics, education, social security and employment of the household members. This study is carried out annually (data collection during the month of July). Surveys from 2011 to 2014 used the sampling framework from the previous census 2000, the following surveys used the sampling framework updated with the census 2011. The sampling design used in the ENAHO is a two-stage stratified random sampling where census segments are the first stage units selected with probability proportional to size, and dwellings are defined as the final stage units. Administratively, Costa Rica has four disaggregation levels: two zones, six planning regions, 81 cantons and 473 districts (municipalities). The sampling design specifies twelve strata – each planning region divided by urban and rural areas. In this case, the strata coincide with the study domains. Smaller domains are not considered to guarantee a coefficient of variation less than 15% for the main poverty measure (percentage of household under poverty) (INEC 2017). For 2011, the ENAHO selected 1,120 PSUs and 13,440 dwellings (10.7% and 9.9% of the sampling framework, respectively).

There are two main differences across the survey rounds as well as regarding the census data. On the one hand, the last surveys collected more variables than the previous ones and the census, and the definition of some of the variables also changed. On the other hand, there were land reforms in the period of study: in 2011 there were only 81 cantons but in 2017 another canton was created. To deal with these obstacles, only variables that exist with the same definition in both the census and survey data are included for the analysis. Also, for the last survey, cantons were grouped exactly the same as in the census 2011. This straightforward-aggregation is possible because the new canton was created by dividing one of the existing cantons. In this article, the objective is to obtain quality poverty information of households for the third administrative level, that is, the 81 cantons, which are defined as the target small areas.

2.3. Demographic Projections

In order to apply a SPREE technique, it is necessary to provide reliable and up-to-date area totals as it will be explained in detail in Subsection 3.1. Since survey data usually does not produce trustworthy population sizes for small areas (due to sample sizes and areas out of sample), demographic projections are used instead in this article. In Costa Rica, population projections are calculated with the cohort component method (Preston et al. 2001) which considers changes in three components: mortality, fertility and migration. The mortality projection was carried out with an autoregressive integrated moving average (ARIMA) random walk model with drift, and for the fertility and migration components, functional data analysis models were implemented. Further details can be found in INEC and CCP (2013). Because of population projections in Costa Rica consider persons at an aggregate (e.g., canton) level, but in this application the aim is to update the total of households according to their status of poverty, the headship rate (United Nations 1973) by sex and age groups is applied in order to get the household projections. A previous implementation of this methodology in this country can be found in Sáenz (2002).

2.4. Definition of Poverty

In Costa Rica, poverty is measured under different uni-and multidimensional proaches. One of the most important, and that is the focus in this article, is the (monetary) poverty rate which is based on the poverty line method (using non-equivalized household per capita income). The INEC defines two types of lines or thresholds:

1. The indigence or extreme poverty line: set by the per capita costs of a basic food basket. The composition of this basket is defined from the National Survey on Income and Expenditure of the Households (Encuesta Nacional de Ingresos y Gastos de los Hogares, ENIGH) which is carried out every five years. The value of the basket is updated every month based on the consumer price index. If the monthly per capita income of a household is below this line, it is considered under "extreme" poverty. For the census time (July 2011), the indigence line was 39,428 colones (Costa Rican currency) (INEC 2011) which was 27.9% of the median per capita income at that time.

2. The poverty line: considers non-food basic needs. A household is classified in this category if the monthly per capita income is equal to or below this value but higher than the indigence line. The poverty line in July 2011 was 84,006 colones which is 60.7% of the median per capita income (Méndez and Bravo 2011).

In this article, the number of households grouped in three categories of poverty ("extreme poor", "poor" (not extreme) and "not poor") is estimated and updated for six post-censal years (2012–2017) in 81 cantons.

3. Methodology

This section describes the methodology for estimating and updating counts and proportions for small areas. First, SPREE techniques are introduced because they are the basis of this proposal. Second, the recommended strategy for obtaining and updating estimates is explained. Finally, the steps to produce uncertainty measures are described.

3.1. Structure Preserving Estimation (SPREE) Methods

As stated above, SPREE was originally proposed by Purcell and Kish (1980) as a tool to update counts or proportions of a categorical variable of interest according to study domains in intercensal years. The target information of interest in a recent time t_1 is shown as a multi-way contingency table Y_{aj,t_1} grouped by $a = 1, ..., A$ areas or domains (rows) and $j = 1, ..., J$ categories of the variable of interest (columns) (e.g., poverty status). In other words, for a population of size N, all the units i (e.g., individuals or households) are organized according to the area and category to which they belong. The structure of a two-way contingency table can be represented as a saturated log-linear model:

$$\log Y_{aj,t_1} = \alpha^Y_{0,t_1} + \alpha^Y_{a,t_1} + \alpha^Y_{j,t_1} + \alpha^Y_{aj,t_1}. \tag{1}$$

These four terms $\alpha^Y_{0,t_1}, \alpha^Y_{a,t_1}, \alpha^Y_{j,t_1}$ and α^Y_{aj,t_1} can be defined using a centred-constraint parametrization, see, for example, Luna-Hernández (2016). Notice that data from a census (time t_0) can also be arranged as a contingency table (Z_{aj,t_0}) and represented as a saturated log-linear model ($\log Z_{aj,t_0}$) with the same constraints.

For intercensal years, the production of official statistics relies in many cases on survey data. However, due to sampling size limitations, trustworthy results are only available for big areas. The SPREE method provides a solution when updated estimates of frequency characteristics are required in smaller domains. The terms $\alpha^Y_{0,t_1}, \alpha^Y_{a,t_1}$ and α^Y_{aj,t_1} represent the *allocation* structure which are benchmarked to totals or current margins (A row and J column totals), usually provided by direct survey estimates and/or demographic projections. In this article, for simplicity, the allocation structure is also referred to as the survey margins, although it is made up of both demographic projections and survey data. It is assumed that these totals are reliable and updated for postcensal years. Furthermore, the method supposes that the interactions between rows and columns of the census Z_{aj,t_0} (inner cells in the contingency table) remain unchanged for the target years. Therefore, the structural assumption is;

$$\alpha^Y_{aj,t_1} = \alpha^Z_{aj,t_0}. \tag{2}$$

This interaction term provided by the census is usually known as the *association* structure.

Following the proposal of Purcell and Kish (1980), the updated estimates are obtained via the iterative proportional fitting (IPF) algorithm (Deming and Stephan 1940) (also found in literature as *raking* or *contingency table standardization*). A detailed description of the IPF is available for example, in Bishop et al. (2007) and Zaloznik (2011). This algorithm fits a census table by keeping reliable survey margins fixed. The process to obtain the updated SPREE of Y_{aj,t_1} is represented by Luna-Hernández (2016) as:

$$\hat{Y}^{S}_{aj,t_1} = \text{IPF}\left[\exp\left(\hat{\alpha}^{Y}_{aj,t_1}\right), Y_{a,t_1}, Y_{j,t_1}\right], \tag{3}$$

with Y_{a,t_1} and Y_{j,t_1} representing the reliable survey margins (rows and columns) and $\hat{a}^{Y}_{aj,t_1} = a^{Z}_{aj,t_0}$.

As it will be later described, new versions of SPREE estimators have been proposed and each of them define different assumptions on their association structure $\hat{\alpha}^{Y}_{aj,t_1}$. The process in Equation 3 is applied with the defined association structure of each SPREE-type estimator.

In order to apply a fitting strategy via IPF, Koebe et al. (2022) summarise some basic requirements that should be considered:

1. the data to fit must be arranged in categories (e.g., contingency tables),
2. the margins of the census and survey structures must have the same length (same number of rows and columns),
3. totals by rows and columns must be equal,
4. the census data (association structure) must contain the indicator of interest (e.g., poverty status) with the same definition or a highly correlated indicator (Green et al. 1998).

In practice, requirements two and four are not met in the Costa Rican scenario. Due to administrative reforms that occurred in postcensal years (e.g., merge or split domains), the number of local areas in the census and in the surveys differs. Another situation where requirement two may not be fulfilled is when some areas were not selected in the sample, leading to an incomplete allocation structure. For these cases, several solutions can be considered: complementary information such as administrative registries or population projections could be used as reliable survey margins, rows that are not in the survey composition can be eliminated from the census, or adding missing rows with small values (e.g., 1) to re-construct the survey compositions in the same way as defined in the census. The first alternative is implemented in the current example.

Regarding the fourth requirement, income or poverty information is not obtained directly in the Costa Rican census. A solution for this kind of situation was proposed by Isidro et al. (2016). The so-called ESPREE considers the case when the indicator of interest is not present in the census data, and another small area estimation method (e.g., the ELL method) is applied as a first step with the aim of obtaining the required information for the census year. Thereafter, the original SPREE technique is conducted as a second step to proceed with the updating process. Considering the characteristics of the

case of study exposed in this article, the methodology of ESPREE is followed to obtain updated poverty estimates.

Another line of research within the SPREE framework is the bias reduction of the estimator. The GSPREE introduced by Zhang and Chambers (2004) makes the structural assumption of SPREE more flexible by adding a *proportionality coefficient* (constant obtained through direct estimates $\hat{Y}^{Dir}_{aj,t_1}$) adjusting Equation 2 to:

$$\alpha^{Y}_{aj,t_1} = \beta \alpha^{Z}_{aj,t_0}. \tag{4}$$

Note that the SPREE assumption defined previously in Equation 2 is the case of Equation 4 when $\beta = 1$. Moreover, Luna-Hernández (2016) proposes a version that aims to relax two restrictions of the former methods (SPREE and GSPREE): the relationship between the association structures (interaction terms) of the census and target compositions are controlled only by one parameter, and in the case of the GSPREE the proportionality parameter β, is assumed to be the same for all the categories. GSPREE and MSPREE mainly differ from the original SPREE because a multi-way contingency table of direct estimates is also necessary in order to update the association structure, meanwhile, the former version only requires the availability of suitable (total) margins. The novel method is called *multivariate* SPREE (MSPREE) because in this case, the target compositions are the interactions within each area. Then, the coefficient $\boldsymbol{\beta}$, similar to the proportionality coefficient specified by Zhang and Chambers (2004), now is represented by a $J \times J$ matrix (with $(J-1) \times (J-1)$ free parameters), and varies inside each area, from one category to another. The main benefit of this proposal is to be able to capture better relationships between categories, instead of assuming that these interactions remain identical over time. With this, the bias that can occur through changes in the association structure (which is not accounted in SPREE and GSPREE), is also reduced. Similarly as in Equations 2 and 4, the MSPREE structural assumption is expressed as:
which is equivalent to:

$$\alpha^{Y}_{a,t_1} = \boldsymbol{\beta} \alpha^{Z}_{a,t_0} \tag{5}$$

where $a = 1, ..., A$ areas, and for each area, the interaction terms are: $\alpha^{Y}_{a,t_1} = (\alpha^{Y}_{a1,t_1}, ..., \alpha^{Y}_{aJ,t_1})$ and $\alpha^{Z}_{a,t_0} = (\alpha^{Z}_{a1,t_0}, ..., \alpha^{Z}_{aJ,t_0})$. The target MSPREE composition Y^{M}_{aj,t_1} can also be obtained via IPF as in Equation 3. In the same way as in the original SPREE and GSPREE, the reliable total margins are preserved, but the same list of requirements aforementioned must be fulfilled.

Estimates of β can been obtained via Maximum Likelihood (ML) or Iterative Weighted Least Squares (IWLS), in both cases by using a log or a logit link. Because a target contingency table can be represented as a log-linear model, SPREE fits within the framework of generalized linear models (Marker 1999; Noble et al. 2002). Moreover, as Agresti (2002) indicates, a log link with a Poisson response is commonly applied to model cell counts in contingency tables, but he also shows that the Poisson expected frequencies μ_{aj,t_1} are equal to $n\pi_{aj,t_1}$, where π_{aj,t_1} are the cell probabilities used under multinomial sampling and n the sample size. For this reason, when working with log-linear models (as SPREE), the estimation of coefficients β can be obtained with a Poisson $\hat{Y}^{Dir}_{aj,t_1} | \alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim}$ Poisson(μ_{aj,t_1}) or Multinomial $\hat{Y}^{Dir}_{aj,t_1} | \alpha^{Z}_{a,t_0} \overset{\text{ind}}{\sim}$ Multinomial $(\sum_{j=1}^{J} \hat{Y}^{Dir}_{aj,t_1}, \pi_a, t_1)$

leading to similar results. Notice that, similarly to the GSPREE, the computation of β requires direct estimates (i.e., not only survey margins but a survey composition $\hat{Y}^{Dir}_{aj,t_1}$).

The second alternative to obtain estimates of β is using IWLS. This algorithm is implemented in Zhang and Chambers (2004) and Luna-Hernández 2016 suggests it when the sample is drawn using a complex sampling design because using ML for the parameter estimation can lead to misspecification if sampling design information is ignored. In the case of the survey data used in this article, the sample was gathered in a two-stage selection process considering unequal selection probabilities, and for this reason, a fully distributional approach should not be assumed. Consequently, in this article, the parameters of interest are estimated via IWLS. This method requires an estimate of the variance-covariance matrix of the target composition $\hat{Y}^{M}_{aj,t_1}$ which is usually not available. Luna-Hernández 2016 solves this issue by multiplying a design effect with the variance that corresponds to a simple random sample without replacement design. This proposal assumes that samples are independently selected in each area and there are no existing correlations among estimates from different areas. The estimate of the variance-covariance matrix is represented as:

$$\hat{V}_{aj,t_1} = \frac{\text{deff}_{j,t_1}\ \hat{\pi}^{M}_{aj,t_1}\left(1-\hat{\pi}^{M}_{aj,t_1}\right)}{n_{a,t_1}}, \tag{6}$$

with $\hat{\pi}^{M}_{aj,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{\hat{Y}_{a,t_1}}$ and n_{a,t_1} the area sample sizes. Further details about the IWLS algorithm can be consulted in Jiang (2007) and Luna-Hernández (2016).

3.2. The Empirical Best Predictor (EBP) Method

The EBP is applied in order to obtain the information of poverty status in the census structure. The EBP methodology proposed by Rao and Molina (2015) implements a unit-level nested error regression model to get estimates of a specific variable of interest in the census, using (usually) survey data that contains this variable. This method has been extensively implemented in SAE problems and also specifically for poverty estimation (see e.g., Pratesi 2016 or Das and Haslett 2019). The process assumes a random effects model for a finite population of size *N:*

$$y_{ai} = \mathbf{x}^{T}_{ai}\beta + u_a + e_{ai},$$

where y_{ai} represents the target variable and $\mathbf{x}_{ai}$ the set of covariates for the ith individual or household in ath area, u_a indicates the area-specific random effects and e_{ai}, the unit-level error. The two last terms are assumed to be normal, independent and identically distributed. By using survey data, the following estimates are obtained: $\hat{\beta}, \hat{\sigma}^2_u, \hat{\sigma}^2_e$ and the weighting factors $\hat{\gamma}_a = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u + \frac{\hat{\sigma}^2_e}{n_a}}$, where n_a denotes the sample size in area a, and σ^2_u and σ^2_e indicates the between and within group variance respectively.

Then, $l = 1, ..., L$ Monte Carlo simulations to generate a pseudo population are conducted:

$$y_{ai} = \mathbf{x}^{T}_{ai}\hat{\beta} + \hat{u}_a + v^{(l)}_a + e^{(l)}_{ai}, \tag{7}$$

where $v_a^{(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_a))$ and $e_{ai}^{(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and the predicted random effect $\hat{u}_a$ is defined as $\hat{u}_a = E(u_a|\mathbf{y}_{as})$. The final indicator of interest is obtained taking the mean over the L iterations.

Molina and Rao (2010) explain that the EBP estimator can be biased when model error terms depart from normality. When working with income data, a common practice to achieve Gaussian assumptions is by using a logarithmic transformation (Elbers et al. 2003; Molina and Rao 2010) which is a special case of the Box-Cox transformation (Box and Cox 1964). In this case study, as will be explained in detail in Subsection 5.1, the incidence of poverty is approximated by modeling income, for this reason, departures from normality are reduced with a Box-Cox transformation. Further details on the performance of the EBP under data-driven transformations can be found in Rojas-Perilla et al. (2020). The EBP was conducted using R (R Core Team 2018), specifically with the Package emdi (Kreutzmann et al. 2019).

3.3. *Strategy to Estimate and Update Poverty Estimates in Costa Rican Cantons*

The goal of this article is to obtain and update poverty estimates in three categories: "extreme poor", "poor" (not extreme), and "not poor". However, as noted previously, no poverty information is collected directly from the census, nor income or expenditure data. Thus, the applied methodology considers characteristics of some of the SPREE methods, specifically the ESPREE and the MSPREE. Since the census data does not contain poverty information, this article adjusts the ESPREE framework. Instead of applying an ELL model to estimate poverty in the census data followed by the original SPREE to update the counts as in Isidro et al. (2016), the empirical best predictor (EBP) (Rao and Molina 2015) is implemented followed by the MSPREE in this work (for simplicity, also referred as EBP-MSPREE). To the best of my knowledge, MSPREE is the most recent and complete technique mentioned in SPREE literature. This version provides more flexibility and bias reduction compared with the previous versions, therefore it is implemented as the main tool in the updating part of the process.

The estimation and updating strategy can be summarized in the following steps:

1. **Estimating the proxy association structure via EBP** Considering that census and survey data, both for the same year at individual level are available, the EBP explained in Subsection 3.2 is applied in order to obtain the information of poverty status in the census structure.
2. **Obtaining the allocation structure.** Household projections as mentioned in Subsection 2.3 provide the total of households in each canton (row margins) and survey data described in Subsection 2.2, the total of households by poverty status. Both sources are available for postcensal years (2012–2017).
3. **Updating estimates via MSPREE.** Intercensal EBP-MSPREE compositions $\hat{Y}_{aj,t_1}^{EM}$ are estimated considering the outputs from the two previous steps (association structure α_{aj,t_0}^{Z} from Step 1, and row Y_{a,t_1} and column Y_{j,t_1} margins from Step 2, that represent the allocation structure). Direct estimates from survey data and design effects are also required as explained at the end of Subsection 3.1. With these inputs, the procedure can be described as follows:

1. the matrix of coefficients β is estimated with an IWLS algorithm that requires a variance-covariance matrix. This matrix is approximated using design effects as showed in Equation 6 with EBP composition estimated in t_0 and the direct estimate obtained from the survey in t_1,
2. $\hat{\alpha}^{Y}_{aj,t_1}$ is estimated with Equation 5,
3. taking into account all these elements, the target estimate $\hat{Y}^{EM}_{aj,t_1}$ is finally obtained with Equation 3.

3.4. Uncertainty of the Updated Estimates

The benefits of two SPREE versions are used in this article. To motivate the proposed $\widehat{\text{MSE}}(\hat{Y}^{EM}_{aj,t_1})$, the procedures via bootstrap of the predecessors (MSE of ESPREE and MSPREE) are briefly described in this Section. Details about other approaches, for example, via linearization methods can be found in Isidro (2010), Isidro et al. (2016) and Luna-Hernández (2016).

Two sources of variation are considered when obtaining estimates via the ESPREE method: survey data (allocation structure) and pseudo-populations (association structure). Being $\hat{Y}^{E}_{aj,t_1}$ the ESPREE estimates, the uncertainty estimate is the sum of two variances: $\text{Var}(\hat{Y}^{E}_{aj,t_1}) = \text{Var}^{\text{survey}}(\hat{Y}^{E}_{aj,t_1}) + \text{Var}^{\text{census}}(\hat{Y}^{E}_{aj,t_1})$. The first variance term $\text{Var}^{\text{survey}}$ is obtained by generating $b = 1, ..., B$ independent bootstrap samples from the original survey data and computing $\hat{Y}^{E,b}_{aj,t_1}$ by keeping fixed the census data (i.e., association component) in every replication. The second term $\text{Var}^{\text{census}}$, is obtained in a similar way. In this case, B-ESPREE estimates $\hat{Y}^{E,b}_{aj,t_1}$ are computed based on $b = 1, ..., B$ bootstrap populations or pseudo-census to account for the uncertainty provided by the association structure, and the allocation structure (survey margins) will be held fixed over the replications. Because in ESPREE, the census values are estimated via ELL (or EBP), the $b = 1, ..., B$ pseudo-census generated from the ELL process can be also be used here. Both, $\widehat{\text{Var}}^{\text{survey}}$ and $\widehat{\text{Var}}^{\text{census}}$ are unconditional variances. As aforementioned, the final uncertainty estimation is only the sum of these two terms, meaning that the authors assume that there is no covariance between both estimators.

Regarding MSPREE uncertainty measures of $\hat{Y}^{M}_{aj,t_1}$, Luna-Hernández (2016) proposed three alternatives: an analytical approximation for the variance of the estimator, finite population MSE (FP-MSE) and an unconditional MSE (U-MSE). For simplicity and consistency, the last one is here described. From the point estimate $\hat{Y}^{M}_{aj,t_1}$, calculate the within-area proportions $\hat{\pi}_{a,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{\hat{Y}_{a,t_1}}$ for each area $a = 1, ..., A$. Generate B-bootstrap populations Y^{*b}_{aj,t_1} under the assumption that the target estimate across areas has the following distribution: $\mathbf{Y}_a | \alpha^{Z}_{a,t_0} \sim \text{Multinomial}(Y_{a,t_1}, \hat{\pi}_{a,t_1})$. From each bootstrap population draw a sample and follow the steps of Subsection 3.2 to obtain $\hat{Y}^{M,b}_{a,j,t_1}$. Finally, compute $\widehat{\text{U-MSE}}(\hat{Y}^{M}_{aj,t_1}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Y}^{M,b}_{aj,t_1} - \hat{Y}^{*b}_{aj,t_1})^2$.

The two procedures to obtain the MSE of $\hat{Y}^{E}_{aj,t_1}$ and $\hat{Y}^{M}_{aj,t_1}$, were briefly described to better contextualize the MSE proposed in this article and to point out the differences and similarities between them. Thus, results from these two previously described MSEs are not produced in this article.

Because the estimation and updating process in this application makes use of the EBP followed by the MSPREE, (EBP-MSPREE) the sources of uncertainty in all of the steps

should be considered. The idea of including the variation of each element is taken from Isidro et al. (2016). The objective is to contemplate that the MSE of the EBP-MSPREE entails three sources of uncertainty: allocation structure, association structure, and the yearly association updating). The difference in this proposal is that instead of calculating each variability term and adding them ($\text{Var}^{\text{survey}} + \text{Var}^{\text{census}} + \text{Var}^{\boldsymbol{\beta}}$), a single bootstrap procedure will be performed, varying each of the required elements in each replication. This decision is made since it cannot be denied that covariances between the estimators exist. This aspect requires special attention and it should be studied in further investigations. The steps to obtain the MSE for the EBP-MSPREE are as follows:

1. From Equation 7, L Monte Carlo pseudo populations in t_0 were generated. Based on defined thresholds (poverty lines), L cross-classified population tables are created with dimension: A areas and J categories of poverty. Notice that the average across them was used to defined α^{Z}_{aj,t_0} and finally compute the point estimate $\hat{Y}^{EM}_{aj,t_1}$. Now, these L pseudo populations are used to create $\alpha^{z,b}_{aj,t_0}$ (with $L = B$), to account for the uncertainty that this structure provides, similarly as in Isidro (2010) and Isidro et al. (2016).
2. To take into consideration the uncertainty from the allocation structure, generate B pairs of margins $Y^{b}_{a,t_1}, Y^{b}_{j,t_1}$ from the point estimate $\hat{Y}^{EM}_{aj,t_1}$ assuming a multinomial distribution, e.g., $Y^{b}_{a,t_1} \sim \text{Multinomial}(\tilde{Y}, \tilde{\pi}_a)$, where $\tilde{\pi}_a = \frac{\hat{Y}_{a,t_1}}{\tilde{Y}}$ and $\tilde{Y} = \sum_{a=1}^{A} \hat{Y}_{a,t_1}$.
3. The uncertainty due to the estimation of β required in the MSPREE is obtained following the procedure of the U-MSE previously described: From the point estimate $\hat{Y}_{aj,t_1}$ calculate the within-area proportions $\tilde{\pi}_{a,t_1} = \frac{\hat{Y}^{M}_{aj,t_1}}{Y_{a,t_1}}$ for each area $a = 1, ..., A$. Generate B-bootstrap populations Y^{*b}_{aj,t_1} under the assumption that the target estimate across areas has the following distribution: $\text{Y}_a | \alpha^{Z}_{a,t_0} \sim \text{Multinomial}(Y_{a,t_1}, \hat{\pi}_{a,t_1})$. From each B-bootstrap population draw a sample to get $y^{EM,b}_{aj,t_1}$.
4. With the output of Step 1 and 3, specify the MSPREE structural assumption $\hat{\alpha}^{Y,b}_{a,t_1}$ as in Equation 5.
5. With $L = B$ and the output of Step 2 and 4, compute B EBP-MSPREE estimates

$$\hat{Y}^{EM,b}_{aj,t_1} = \text{IPF}\left[\exp\left(\hat{\alpha}^{Y,b}_{a,t_1}\right), Y^{b}_{a,t_1}, Y^{b}_{j,t_1}\right].$$

6. Finally, estimate the MSE:

$$\widehat{\text{MSE}}\left(\hat{Y}^{EM}_{aj,t_1}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{Y}^{EM,b}_{aj,t_1} - \hat{Y}^{EM}_{aj,t_1}\right)^2. \tag{8}$$

4. Results of the Application

In this section, the results of the estimation and the updating process are explained. First, the EBP model for obtaining the association structure is described, (i.e., poverty information in the census composition) with its corresponding model evaluation and descriptive statistics. Second, some relevant results in the practical sense are shown, for example, the evolution of (updated) poverty indicators from 2012 to 2017 for selected cantons, and the cantons with the highest poverty rates in the last year of study. Finally, uncertainty measures and validation results are presented.

4.1. Poverty Estimates for the Census

An EBP model to obtain poverty incidence by domain was conducted by setting income per capita of the household as the dependent variable and the socio-economic covariates described in Table 1 as predictors. In Costa Rica, urbanity plays an important role in socioeconomic topics. For this reason, the variable zone (urban/rural) is added to the model. The other groups of variables contain information on the head of the household, the household members, and the housing. Most of them have been used in previous studies as predictors of poverty or are part of indexes such as the UBNS and the multidimensional poverty index (MPI) (Alkire and Foster 2007). Under the model here specified, the sample design is assumed non-informative.

To select the model, several transformations were considered to reduce normality departures of the error terms, but the final version applies a Box-Cox transformation (Box and Cox 1964) with an optimal lambda (0.1585) using the Restricted maximum likelihood (REML) approach. Regarding the normality assumptions of this linear mixed model, graphical diagnostic of the residuals are used. Figure 1 shows that normality assumptions are rejected for the unit level (Skewness -0.0491 and Kurtosis 6.4179), but not for the random effects (canton-level). The latter is also confirmed with the Shapiro Wilk test ($W = 0.9860$ and $p = 0.5286$), and Skewness (-0.3584) and Kurtosis (3.0649) measures. For this example, normality for both, the unit level and the random effects, is assumed. Also, the marginal $R^2 = 0.5003$ and the conditional $R^2 = 0.5081$ were observed.

Table 2 shows the summary statistics of the population and sample domains. For 2011, all cantons are in-sample although it is not the case for some post-censal years. Domain sizes from the census data vary from 1705 (minimum) to 84,066 (maximum) households, and in the case of the ENAHO, it varies from 12 to 877 households.

Table 1. Covariables included in the EBP model to obtain poverty estimates in census.

Category	Variable
Geographical	1. Zone
head of the	2. Age
household	3. Highest degree of education completed
	4. Sex
	5. Labor condition
Household	6. Proportion of employees in the household
conditions	7. Equivalized size of the household
	8. Overcrowding
	9. At least one member without health insurance
	10. Quantity of economically dependent members
	11. At least one member not attending to formal education
	12. At least one member not with a educational lag
Housing	13. Poor condition of the floor or ceiling
conditions	14. Any member not used internet last 3 months
	15. No garbage disposal system
	16. No exclusive toilet for the household

Notes: Labor condition has three categories: employed, unemployed, out of Labor Force.
Variable 9 refers to population older than 17 years old.
Variables 11 and 12 refer to population between five and 19 years old.
Variable 14 refers to population older than four years old.

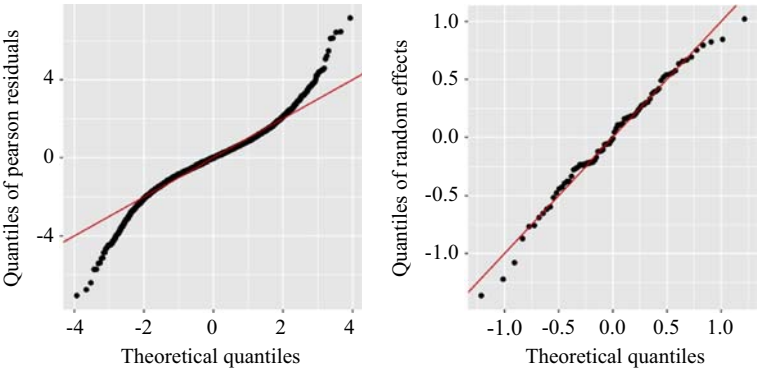


Fig. 1. Q-Q plots of the unit-level errors and the random effects.

Table 2. Summary statistics for sample and population sizes.

	Min	1st Q	Median	Mean	3rd Q	Max
Sample domains (In-sample: 100%)	12	62	103	144.37	191	877
Population domains	1705	5961	11032	15271.37	17148	84066

In order to study poverty in the three aforementioned interest groups, the two poverty lines described in Subsection 2.4 were implemented (as "customised indicators" in the R package `emdi`, for further details about this functionality see Kreutzmann et al. 2019) Descriptives of coefficients of variation (CV) for the directand model-based estimates obtained via EBP are presented in Table 3. As expected, the CVs reflect the lack of precision in the categories "extreme poor" and "poor" (not extreme) for the direct estimates. The improvement when the EBP model is conducted is clear, with a maximum CV of 29.9% and 18.2% for "extreme poor" and "poor" (not extreme) categories respectively. Notice also, that the third quartile of the CV in the category "extreme poor" is below 20%.

Figure 2 shows the proportion of households under extreme poverty based on the direct estimates and the model-based estimates obtained via the EBP model. The maps on the right side give a closer look at the metropolitan area which consists of 31 cantons, covering approximately 60% of the population. In four of these cantons, no households were identified as "extreme poor" with direct estimates (out-of-sample domains represented in

Table 3. Coefficients of variation of the direct and model-based estimates for poverty status.

	Direct			EBP		
CV	Extreme	poor	Not poor	Extreme	poor	Not poor
Min	0.068	0.071	0.033	0.085	0.045	0.012
1st Q.	0.232	0.164	0.086	0.127	0.065	0.023
Median	0.332	0.323	0.110	0.155	0.085	0.033
Mean	0.412	0.271	0.123	0.160	0.089	0.033
3rd Q.	0.524	0.234	0.145	0.181	0.106	0.040
Max	1.000	1.000	0.281	0.299	0.182	0.062

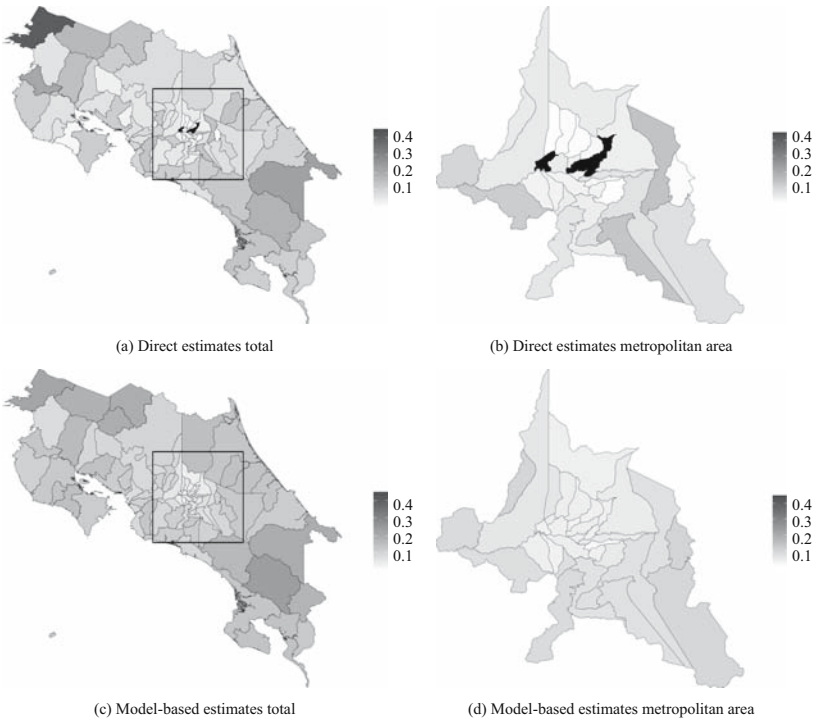


Fig. 2. Proportion of extreme poverty: direct and model-based estimates, 2011.

black in Figure 1 (a) and (b)). The results are consistent with a previous study (Méndez and Bravo 2011), where higher levels of poverty are found on the border with Nicaragua (e.g., La Cruz, located in the northwest of the country) and on the border with Panamá (e.g., Buenos Aires, and Talamanca, located in the southeast of the country).

4.2. Updated Poverty Estimates

To analyze major changes in the incidence of extreme poverty, estimates of 2011 and 2017 are compared using Z-scores:

$$Z = \frac{\text{Estimate}_{2011} - \text{Estimate}_{2017}}{\sqrt{(\text{Standard error}_{2011})^2 + (\text{Standard error}_{2017})^2}}$$

Figure 3 presents the three cantons with biggest change in this category of poverty. Among all of the 81 cantons, these are the cantons with Z-scores higher than two. Note that all of them show a reduction in the incidence of poverty between 2011 and 2017. It is also important to mention that none of the three cantons are among the poorest, which means that the biggest improvements are not observed in the areas most in need. The cantons Curridabat and Montes de Oca had the highest growth in extreme poverty (although small Z-score values: 0.2235 and 0.2052 respectively). However, both are among the cantons with the lowest incidence of extreme poverty in both years.

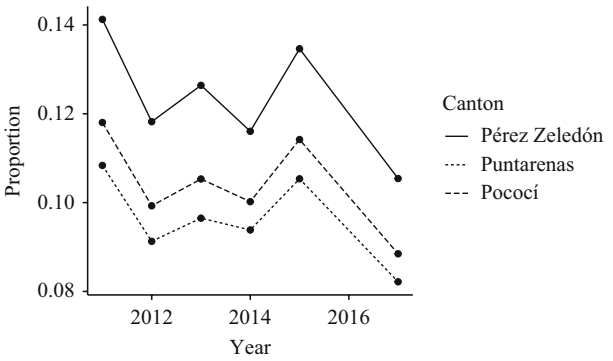


Fig. 3. Cantons with biggest change in the incidence of extreme poverty from 2011 to 2017.

Identifying the poorest cantons is also relevant in order to fight against this phenomenon in a more efficient way. Figure 4 shows the small areas with the highest incidence of extreme poverty (in proportions) in the last year (2017). Here, it is important to point out that for all the years of study (2011–2017) the same five cantons remain on this list, indicating that economic conditions of these areas have not been better in comparison with other areas in recent years.

The sources of uncertainty that were explained in Subsection 3.4, are displayed for the last year of study as a coefficient of variation in Figure 5. As expected, the category "not poor" is the one with the minimum CV and most of the values are under 20% which is considered "acceptable" according to the parameters for official publications of the National Statistical Office of Costa Rica, INEC (2015).

As explained in Subsection 2.2, the target areas in this article (81 cantons) are nested in six planning regions. Because the INEC of Costa Rica publishes official results on poverty only for these planning regions (gathered from the ENAHO), this is the only geographical level where it is possible to make the comparison with the updated estimates via EBP-MSPREE. For this reason, as a way to evaluate the updated estimates of poverty, model-based estimates of cantons are aggregated into the six planning regions and compared with

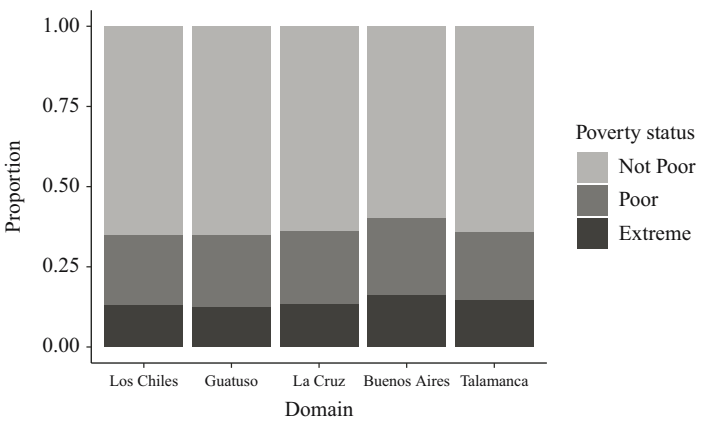


Fig. 4. Cantons with the highest incidence of extreme poverty, 2017.

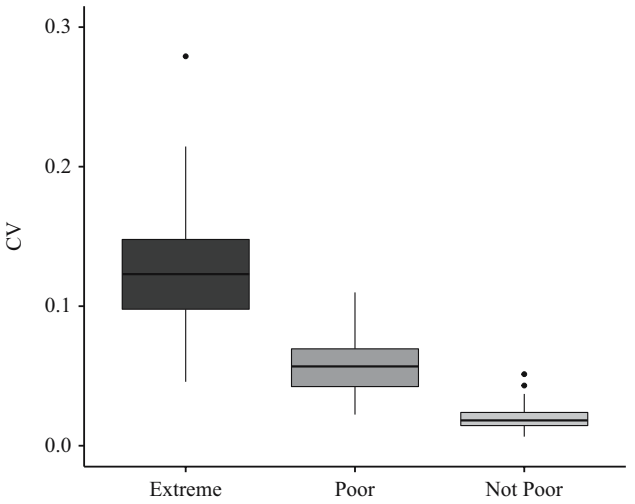


Fig. 5. Coefficients of variation for 2017.

Table 4. Direct and EBP-MSPREE estimates for poverty status (proportions) by planning region 2017.

	Direct			EBP-MSPREE		
Regions	Extreme	Poor	Not poor	Extreme	Poor	Not poor
Central	3.9	11.9	84.2	3.8	11.7	84.5
Chorotega	5.9	16.5	77.6	8.3	18.0	73.7
Pacífico Central	8.9	21.0	70.1	6.8	16.7	76.5
Brunca	10.4	19.1	70.5	10.7	20.7	68.6
Huetar Caribe	8.9	17.8	73.3	9.3	18.9	71.8
Huetar Norte	9.2	18.3	72.5	7.8	17.6	74.6
Total	5.7	14.3	80.0	5.7	14.3	80.0

the official publication. It is relevant to mention that three cantons overlap with two regions at the same time. This problem was solved by allocating the estimated counts in proportion to the respective population in each region. For a more practical comparison, proportions instead of counts are shown in Table 4. EBP-MSPREE results are satisfactory in terms of their similarity to the direct estimates. Most of the regions show close results to the published one, and the region with the highest discrepancies is the Pacífico Central. This region, however, is the domain with a smaller sample size, therefore it is expected to have less accurate results. The opposite is the case of the region Central which has the biggest sample size and results are very close to the ones in the official publication (results available in INEC 2017).

5. Design-Based Simulation

In this section, results from a design-based simulation study are presented. The objective is to evaluate the EBP-MSPREE procedure explained in Subsection 3.2 to estimate and update counts on poverty incidence in three categories: “Extreme poor”, “poor” not

extreme, and "not poor," as well as assessing the performance of the bootstrap MSE estimator described in Subsection 3.3. To conduct the evaluation, two census compositions are required. One census from time t_0 as the primary input to get EBP-MSPREE estimates but also a second census from time t_1 to compare the updated results. Due to a recent census is not available, survey data described in Subsection 2.2 is used in this experiment. Survey data set from 2011 is used as a census in t_0 (Z_{aj,t_0}) and survey data from 2012 as a census in $t_1(Z_{aj,t_1})$.

Samples are drawn from them with simple random sampling without replacement, with a sampling fraction of f = 0.2. Since survey data is used as a census, there are many domains with few observations. For this reason, the number of cantons in the simulation is reduced to $A = 23$, (instead of $A = 81$ as in the application) and the biggest domains were selected. This allows having all cells with a positive sample, in this case, with at least 15 observations in each category of poverty for all domains. For the first part of the procedure where point estimates are obtained, that is, with an EBP model, a Box-Cox transformation is chosen. Also, the income per capita is defined as the dependent variable and a reduced number of covariates are included in the model, namely: the proportion of employees in the household, the highest degree of education completed by the head of the household, zone, quantity of economically dependent members in the household, equivalized size of the household, at least one member without health insurance, and at least one member not with an educational lag.

In this simulation study, the performance of an EBP-SPREE (similarly as in Isidro et al. 2016) is compared with the proposal of this article, that is, EBP-MSPREE. A total of R = 500 Monte Carlo iterations are defined, with L = 100 Monte Carlo iterations for implementing the EBP, and B = 100 bootstrap iterations for MSE estimation. The performance of the estimated EBP-MSPREE ($\hat{Y}^{EM}_{aj,t_1}$) is evaluated with the relative bias (RB) and the root MSE (RMSE), defined as:

$$\mathrm{RB}\left(\hat{Y}^{EM}_{aj,t_1}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{Y}^{EM,r}_{aj,t_1} - Z_{aj,t_1}}{Z_{aj,t_1}}\right)$$

and,

$$\mathrm{RMSE}\left(\hat{Y}^{EM}_{aj,t_1}\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}^{EM,r}_{aj,t_1} - Z_{aj,t_1}\right)^2}, \tag{9}$$

which is treated as the empirical RMSE. A plot comparing the estimated RMSE (from Equation 8) and the empirical RMSE (from Equation 9) is used to validate the proposed MSE estimator $\widehat{\mathrm{MSE}}(\hat{Y}^{EM}_{aj,t_1})$. Relative bias and relative RMSE of the estimated RMSE for each area a and category j are also computed as follows:

$$\mathrm{rel.Bias.Est.RMSE} = \left(\frac{\mathrm{Est.RMSE} - \mathrm{Emp.RMSE}}{\mathrm{Emp.RMSE}}\right)$$

$$\text{rel.RMSE.Est.RMSE} = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\text{Est.RMSE}^{(r)} - \text{Emp.RMSE}\right)^2}}{\text{Emp.RMSE}}.$$

5.1. Results of the Design-based Simulation

Table 5 summarizes the results of the evaluation of the EBP-MSPREE estimator in comparison with a previous version, namely EBP-SPREE. The values of RB and RMSE are averaged over 23 areas for each category of poverty. When looking at the median and mean for each category of poverty, the EBP-MSPREE procedure provides a smaller RMSE than the EBP-SPREE procedure for three out of the six cases, especially in the category "not poor". The absolute value of the RB of the EBP-MSPREE procedure is considerably lower than the absolute value of the RB of EBP-SPREE in four out of the six cases. This is observed in the categories "extreme" and "not poor".

Figure 6 displays the estimated and empirical RMSE over the domains and categories of poverty for the EBP-MSPREE and the EBP-SPREE methods. Based on this figure, it is possible to conclude that the estimated RMSE tracks the empirical RMSE better for the EBP-MSPREE procedure, and this can be observed for all of the categories of poverty. A closer look at the performance of the proposed MSE is provided in Table 6. The RB-RMSE for the EBP-MSPREE indicates a moderate underestimation in the mean and the median for the "extreme" and "poor" categories and an overestimation for the median of the category "poor". In terms of RB-RMSE and RRMSE-RMSE, the results on the performance of the MSE are favorable for the EBP-MSPREE.

6. Conclusions and Further Steps

Public policies require not only accurate and reliable information for decision- making but this information should also be timely. It is common that the production of official statistics faces challenges due to limitations of resources. In this article, a methodology to obtain reliable and updated estimates in small areas is presented and exemplified with a real-world application. For many developing countries, censuses are conducted every ten years and sample sizes of annual national surveys are not big enough to provide reliable results for small areas. An additional limitation that is considered in this work is that information of interest is not present in census data, as it is required for SPREE methods. The strategy proposed considers two well-known small area estimation techniques.

Table 5. RB and RMSE for the incidence of poverty by status under different SPREE approaches.

		Extreme		Poor not extreme		Not poor	
		Median	Mean	Median	Mean	Median	Mean
RB	EBP-SPREE	0.1965	0.1279	0.0185	0.0112	-0.0219	0.0337
	EBP-MSPREE	0.0195	0.0987	0.0322	0.0144	-0.0125	0.0098
RMSE	EBP-SPREE	0.3120	0.3468	0.1274	0.1514	0.0744	0.1060
	EBP-MSPREE	0.2781	0.3538	0.1479	0.1673	0.0705	0.0985

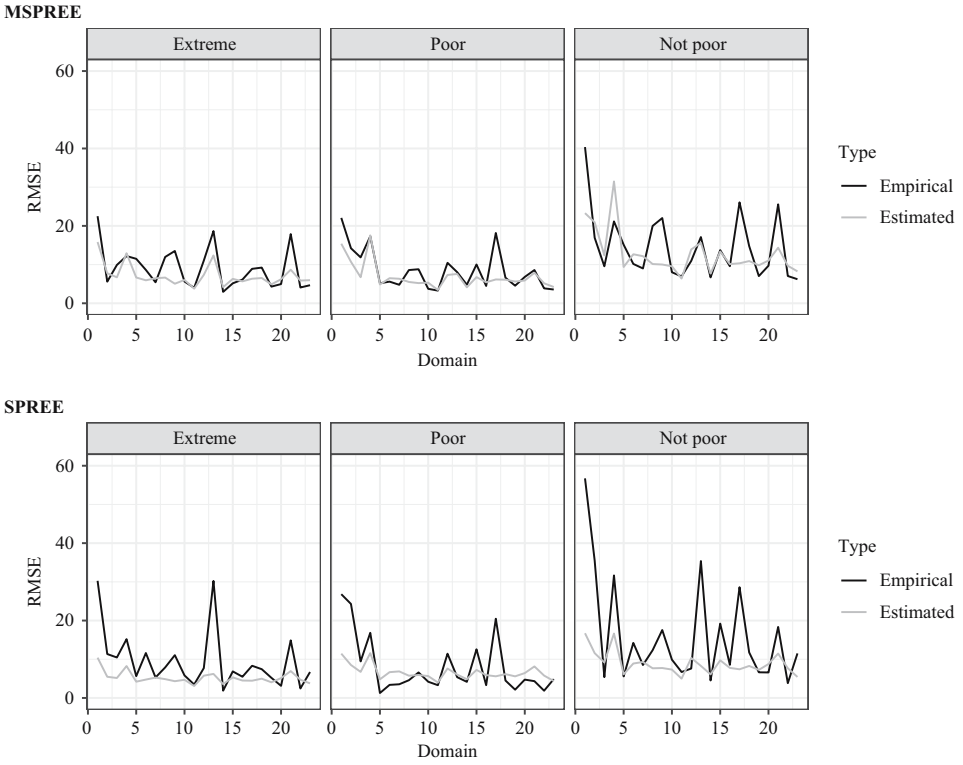


Fig. 6. Estimated and empirical domain-and category-specific RMSEs of the counts on poverty incidence.

Table 6. Performance of the MSE estimator: Mean and Median of RB and RMSE by poverty status.

		Extreme		Poor not extreme		Not poor	
		Median	Mean	Median	Mean	Median	Mean
RB-	EBP-SPREE	-0.3294	-0.2294	0.2107	0.3627	-0.2969	-0.1711
RMSE	EBP-MSPREE	-0.0752	-0.0812	-0.0765	-0.0714	0.1388	0.0080
RRMSE-	EBP-SPREE	0.4616	0.4595	0.3785	0.6769	0.3813	0.4477
RMSE	EBP-MSPREE	0.3812	0.3911	0.3586	0.3665	0.3678	0.3666

An EBP is conducted to get poverty estimates in the census data, and as a second step, the MSPREE of Luna-Hernández (2016) is applied to update the estimates in postcensal years. Based on the results of the application, it is possible to conclude that the strategy proposed delivers quality results in terms of CVs and compares favourably with direct estimates. The application showed that this methodology gives the opportunity to analyze specific groups of interest, areas, and years. For example, that the poorest cantons in Costa Rica have remained with little overall improvements for the period studied.

Although the methodology proposed allows to obtain the target estimates, there are several aspects that can be improved, especially in the uncertainty estimation. Original SPREE and two other versions (GSPREE and MSPREE) assume that the census has the

variable of interest and therefore no uncertainty from the association structure is required. However, when another small area estimation method is needed as a first step to get the census structure (in this case with an EBP), variability from it should be considered. In this methodology, a parametric bootstrap is implemented to get uncertainty from the allocation, the association structure, and the estimation of the β coefficients required in MSPREE. One potential topic for further research is to combine the MSE that is produced directly from the EBP with the one from MSPREE under an analytical approach. Furthermore, the impact that extreme values in the first part of the procedure (e.g., EBP) can have in the final updated estimates deserves also to be investigated.

SPREE methods have other disadvantages that require further study to get a more flexible technique. For example, a potential improvement in the method could be to allow updating more complex indicators or non-categorical indicators such as the Gini index or (mean and/or median) per capita income. The inclusion of associated variables as suggested by Purcell and Kish (1980) can also be beneficial in the estimation procedure, for instance, the inclusion of urbanity (urban and rural) can be relevant when working with poverty status.

The over-shrinking problem present in the context of small area estimation when the expected sample variance is smaller than the true parameter also deserves to be explored when implementing SPREE-type methods.

Understanding the benefits of SPREE-type methods in comparison with existing models in the small area estimation context requires also further research. Three alternative approaches to deal with the problem of obtaining updated counts or proportions in small domains have been identified and deserve a closer comparison with the SPREE-type methods: 1. The use of the EBP in each year of study with a final benchmark operation performed with MSPREE, 2. exploring potential advantages of using panel survey data or time-series models for example, with the extension of the Fay-Herriot model proposed by Rao and Yu (1994), or 3. implement measurement error models also in the context of area-level models (Ybarra and Lohr 2008). Finally, it is recommended to study the inclusion of non-traditional information sources (e.g., big data) as proposed in Koebe et al. (2022) since the structures of population censuses can quickly become obsolete. A clear example of this is the socio-economic effect that the COVID-19 pandemic generated in many countries, altering the living conditions of many people in a short period of time.

7. References

Agresti, A., 2002. *Categorical data analysis*. John Wiley & Sons, Inc.

Alkire, S., and J. Foster. 2007. *Counting and multidimensional poverty measures.* OPHI working article 7. Available at: https://ophi.org.uk/working-paper-number-07/ (accessed June 2021).

Berg, E., and W.A. Fuller. 2009. "A SPREE small area procedure for estimating population counts." In Proceedings of the Statistical Society of Canada. June, Vancouver. Canada. Available at: https://ssc.ca/sites/default/files/survey/documents/SSC2009_EBerg.pdf (accessed November 2023).

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. 2007. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Box, G.E., and D.R. Cox. 1964. "An analysis of transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26(2): 211–243. DOI: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

CEPAL and MIDEPLAN. 2016. *El enfoque de brechas estructurales: análisis del caso de Costa Rica. CEPAL.* Available at: https://www.cepal.org/es/publicaciones/40805-enfoque-brechas-estructurales-analisis-caso-costa-rica (accessed June 2021).

Das, S., and S. Haslett. 2019. "A comparison of methods for poverty estimation in developing countries." *International Statistical Review* 87(2): 368–392. DOI: https://doi.org/10.1111/insr.12314.

Deming, E., and F. Stephan. 1940. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *The Annals of Mathematical Statistics* 11(4): 427–444. Available at: http://www.jstor.org/stable/2235722 (accessed June 2021).

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro-level estimation of poverty and inequality." *Econometrica* 71(1): 355–364. DOI: https://doi.org/10.1111/1468-0262.00399.

Emwanu, T., J.G. Hoogeveen, and P. Okiira Okwi. 2006. "Updating poverty maps with panel data." *World Development* 34(12): 2076–2088. DOI: https://doi.org/10.1016/j.worlddev.2006.03.005.

Eurostat. 2017. *European statistics code of practice.* Available at https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142. (accessed January 2021).

Feres, J.C,. and X. Mancero. 2001. *El método de las necesidades básicas insatisfechas (NBI) y sus aplicaciones en América Latina*. CEPAL. Available at: http://hdl.handle.net/11362/4784 (accessed June 2021).

Green, A., S. Haslett, and C. Zingel. 1998. "Small area estimation given regular updates of census auxiliary variables." In Proceedings of the New Techniques and Technologies for Statistics Conference. 4–6 November, Sorrento, Italy. Available at: https://www.researchgate.net/publication/2613135_Small_Area_Estimation_Given_Regular_Updates_of_Census_Auxiliary_Variables (accessed November 2023).

Hidiroglou, M., and Z. Patak. 2009. "An application of small area estimation techniques to the canadian labour force survey." In Proceedings of the Survey Methods Section, Annual Meeting Statistical Society of Canada. June 2009. Vancouver, Canada. Available at: https://ssc.ca/sites/default/files/survey/documents/SSC2009_MHidiroglou.pdf (accessed November 2023).

INEC. 2011. *Boletín mensual. Costo de la canasta básica alimentaria,* Julio 2011. Available at: https://inec.cr/estadisticas-fuentes/estadisticas-economicas (accessed June 2021).

INEC. 2012. *Ficha Metodológica: X Censo Nacional de Población y VI de Vivienda 2011. Resultados Generales.* Available at: https://inec.cr/estadisticasfuentes/censos/ (accessed June 2021).

INEC. 2015. *Índice de Pobreza Multidimensional (IPM). Metodología.* Available at: https://www.inec.cr/metodologias (accessed June 2021).

INEC. 2017. *Encuesta Nacional de Hogares. Julio 2017. Resultados generales*. Available at: https://inec.cr/estadisticas-fuentes/encuestas/ (accessed June 2021).

INEC and CCP. 2013. *Estimaciones y Proyecciones de Población por sexo y edad 1950 – 2050*. San José: INEC. Available at: https://ccp.ucr.ac.cr/observa/CRnacional (accessed June 2021).

Isidro, M., S. Haslett, and G. Jones. 2016. "Extended structure preserving estimation for updating small area estimates of poverty." *The Annals of Applied Statistics* 10(1): 451–476. DOI: https://doi.org/10.1214/15-AOAS900.

Isidro, M.C. 2010. *Intercensal updating of small area estimates*. Ph. D. thesis, Massey University, New Zeeland. Available at: https://mro.massey.ac.nz/server/api/core/bitstreams/0b6ed6ba-b8a8-43de-8b7a-92588215e33a/content (accessed November 2023).

Jiang, J. 2007. *Linear and generalized linear mixed models and their applications.* Springer Science Business Media.

Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, and T. Schmid. 2022. "Intercensal updating using structure-preserving methods and satellite imagery." *Journal of the Royal Statistical Society: Series A (Statistics in Society):* 1–23. DOI: https://doi.org/10.1111/rssa.12802.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. 2019. "The R package emdi for estimating and mapping regionally disaggregated indicators." *Journal of Statistical Software* 91(7): 1–33. DOI: 10.18637/jss.v091.i07.

Luna-Hernández, A. 2016. *Multivariate structure preserving estimation for population compositions.* Ph. D. thesis, University of Southampton. Available at: https://eprints.soton.ac.uk/404689/1/Angela%2520Hernandez%2520Final%2520thesis.pdf (accessed November 2023).

Luna-Hernández, A., L.-C. Zhang, A. Whitworth, and K. Piller. 2015. "Small area estimates of the population distribution by ethnic group in england a proposal using structure preserving estimators." *Statistics in Transition new series* 16(4): 585–602. DOI: https://doi:10.21307/stattrans-2015-034.

Marker, D.A. 1999. "Organization of Small Area Estimators Using a Generalized Linear Regression Framework." *Journal of Official Statistics* 15(1): 1–24. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/organization-of-small-area-estimators-using-a-generalized-linear-regression-framework..pdf

Méndez, F., and O. Bravo. 2011. "Mapas de pobreza con datos censales." In proceedings: Costa Rica a la luz del Censo 2011. San José, May 2014. Costa Rica. Available at: https://admin.inec.cr/sites/default/files/media/anpoblaccenso2011-01.pdf_2_2.pdf (accessed November 2023).

Molina, I., and J. Rao. 2010. "Small area estimation of poverty indicators." *Canadian Journal of Statistics* 38(3): 369–385. DOI: https://doi.org/10.1002/cjs.10051.

Noble, A., S. Haslett, and G. Arnold. 2002. "Estimation of small areas via generalised Q10 linear models." *Journal of Official Statistics* 18(1): 45–60. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/small-area-estimation-via-generalized-linear-models.pdf.

OECD. 2016. *OECD Economic Surveys: Costa Rica 2016.* DOI: https://doi.org/10.1787/ecosurveys-cri-2016-en.

Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28: 40–68. DOI: 10.1214/12-STS395.

Pratesi, M. 2016. *Analysis of poverty data by small area estimation.* John Wiley & Sons.

Preston, S.H., P. Heuveline, and M. Guillo. 2001. *Demography: measuring and modeling population processes*. Oxford: Blackwell Publishers Ltd.

Purcell, N.J. and L. Kish. 1980. "Postcensal estimates for local areas (or domains)." *International Statistical Review* 48(1): 3–18.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/ (accessed April 2022).

Rao, J.N., and M. Yu. 1994. "Small-area estimation by combining time-series and cross-sectional data." *Canadian Journal of Statistics* 22(4): 511–528. DOI: https://doi.org/10.2307/3315407.

Rao, J.N.K. 2003. *Small area estimation*. New York: Wiley.

Rao, J.N.K., and I. Molina. 2015. *Small area estimation* (2 ed.). New York: Wiley.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis. 2020. "Data-driven transformations in small area estimation." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(1): 121–148. DOI: https://doi.org/10.1111/rssa.12488.

Sáenz, I. 2002. *Estimación de la cantidad de viviendas y consumo de agua.* Master's thesis, University of Costa Rica. Available at: https://ccp.ucr.ac.cr/documentos/bibliotecavirtual/53.pdf (accessed November 2023).

Tzavidis, N., L.-C. Zhang, A. Luna Hernandez, T. Schmid, and N. Rojas-Perilla. 2018. "From start to finish: a framework for the production of small area official statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181(4): 927–979. DOI: https://doi.org/10.1111/rssa.12364.

United Nations. 1956. *Manual III. Methods for population projections by sex and age*. Available at: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_1956_manual_iii_-_methods_for_population_projections_by_sex_and_age_0.pdf (accessed November 2023).

United Nations. 1973. *Manual VII. Methods of projecting households and families*. Available at: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_1973_manual_vii_-_methods_of_projecting_households_and_families_0.pdf (accessed November 2023).

United Nations. 2019a. *The Millennium Development Goals Report 2015*. DOI: https://doi. org/https://doi.org/10.18356/55eb9109-en.

United Nations. 2019b. *The Sustainable Development Goals Report 2019*. Available at: https://www.un-ilibrary.org/content/publication/55eb9109-en.

Ybarra, L.M., and S.L. Lohr. 2008. "Small area estimation when auxiliary information is measured with error." *Biometrika* 95(4): 919–931. DOI: https://doi.org/10.1093/biomet/asn048.

Zaloznik, M. 2011. *Iterative proportional fitting theoretical synthesis and practical limitations.* Ph. D. thesis, University of Liverpool. Available at: https://www.researchgate.net/publication/262258986_Iterative_Proportional_Fitting_-_Theoretical_Synthesis_and_Practical_Limitations (accessed November 2023).

Zhang, L.-C. and R.L. Chambers. 2004. "Small area estimates for crossclassifications." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66(2): 479–496. DOI: https://doi.org/10.1111/j.1369-7412.2004.05266.x.

Received June 2021
Revised January 2022
Accepted March 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 459–487, http://dx.doi.org/10.2478/JOS-2023-0022

Block Weighted Least Squares Estimation for Nonlinear Cost-based Split Questionnaire Design

Yang Li[1], *Le Qi*[1], *Yichen Qin*[2], *Cunjie Lin*[1], *and Yuhong Yang*[3]

In this study, we advocate a two-stage framework to deal with the issues encountered in surveys with long questionnaires. In Stage I, we propose a split questionnaire design (SQD) developed by minimizing a quadratic cost function while achieving reliability constraints on estimates of means, which effectively reduces the survey cost, alleviates the burden on the respondents, and potentially improves data quality. In Stage II, we develop a block weighted least squares (BWLS) estimator of linear regression coefficients that can be used with data obtained from the SQD obtained in Stage I. Numerical studies comparing existing methods strongly favor the proposed estimator in terms of prediction and estimation accuracy. Using the European Social Survey (ESS) data, we demonstrate that the proposed SQD can substantially reduce the survey cost and the number of questions answered by each respondent, and the proposed estimator is much more interpretable and efficient than present alternatives for the SQD data.

Key words: Block weighted least squares estimation; block-wise missing data; nonlinear cost function; split questionnaire design; large-scale survey.

1. Introduction

The European Social Survey (ESS) is an academically driven survey that has been conducted across Europe every two years since 2001. It measures moral, religious, social, economic, political attitudes and behavior patterns of various populations (Schnaudt et al. 2014). A large number of studies have been conducted based on ESS data. For example, Best and Wolf (2013) studied the attitudes toward homosexuality using a linear regression model; Davidov et al. (2018) performed a test for measurement invariance across all countries with ESS data; Vonneilich et al. (2019) discussed the effects of different aspects of social relations on educational inequalities using a linear mixed effects model. Although ESS data have generated the interest of many researchers, its collection takes a long time and requires great efforts. For example, in 2018, over 40,000 respondents were sampled from 31 countries in Europe (http://www.europeansocialsurvey.org/). Such a large-scale, hour-long, and face-to-face interview includes more than 500 questions, which requires much time and resources for respondents to complete. Furthermore, a long questionnaire often means a heavy burden for respondents, which usually results in low

[1] Renmin University of China, Center for Applied Statistics and School of Statistics, att: Cunjie Lin, 59 Zhongguancun St, Beijing 100872, China. Emails: yang.li@ruc.edu.cn, qle1999@163.com and lincunjie@ruc.edu.cn
[2] University of Cincinnati, Department of Operations, Business Analytics, and Information Systems Cincinnati, Ohio, U.S.A. Email: qinyn@ucmail.uc.edu
[3] University of Minnesota, School of Statistics Minneapolis, U.S.A. Email: yangx374@umn.edu
Acknowledgments: Li's work and Lin's work was supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD910001).

response rates, high sampling errors, and measurement errors (Early 2016; Peytchev and Peytcheva 2017; Liu and Wronski 2018).

The National Assessment of Educational Progress (NAEP) in the USA is another large-scale nationally representative survey to measure student achievements, such as what American students know and can do in various subject areas. More than 150,000 students are included in the sample and each student answers many questions as part of the assessment (Neidorf and Sheehan 2014). The details of sampling in NAEP can be found in Rust and Johnson (1992). Because NAEP is a large-scale educational assessment, it also faces the aforementioned challenges and a balanced incomplete block design is used. Typically, the design splits the pool of items into a set of blocks and the split may depend on such practical issues as the wish to offer students blocks with motivating combinations of items or to match blocks across sub-questionnaires with respect to the time needed to complete them. Then the sub-questionnaires are assigned to students in the lowest unit (usually school classes) to minimize the cluster effects involved in sampling a hierarchically structured population.

Throughout these large-scale surveys, we see that one possible approach to manage survey cost and reduce respondent burden is to use a split questionnaire design (SQD), which divides a long questionnaire into short parts or modules, with each respondent filling only one of the sub-questionnaires that include one or more modules. Along the line of SQD, Raghunathan and Grizzle (1995) proposed a survey design via matrix sampling and the resulting data were analyzed using a multiple imputation method. Adigüzel and Wedel (2008) developed an optimal design by minimizing the Kullbak-Leibler distance between the distributions of the complete questionnaire data and the split questionnaire data, and illustrated the design's efficiency in synthetic and empirical web survey data. In cognitive development research, Rhemtulla and Little (2012) described two planned missing data designs, namely the multi-form design and the two-method measurement design, which were efficient in reducing participant fatigue and improving data quality. Additionally, Andreadis and Kartsounidou (2020) presented evidence that splitting a long questionnaire into short parts contributed positively to the response rates. With the increasing interest in SQD, the optimal design obtained by minimising a proper cost function for the survey, subject to specified reliability constraints, is an important aspect for more research. For example, Chipperfield and Steel (2009, 2011) and Ioannidis et al. (2016) introduced a linear cost function, which facilitates the search for the optimal design. However, the linear form of cost function may still be insufficient to accommodate a potentially complex relationship between the number of questions and the cost of the survey.

Besides questionnaire design, estimation of regression models with SQD data is another open problem. Compared with the analysis based on a complete data set, SQD data are more challenging to analyze due to the block-wise missing structure. Although likelihood-based methods (Little and Schluchter 1985; Skinner and Coker 1996; Chipperfield et al. 2018; Chipperfield and Steel 2012) are widely used in handling missing data, they often require strong assumptions on the distribution of missing data (Dziura et al. 2013). Poor performance was observed when the distribution assumptions were violated (Lesperance and Kalbfleisch 1992). For a linear regression model or other models of interest, another possible technique is to impute the missing values before estimation (Little 1992; Cai et al. 2010; Mazumder et al. 2010; Yuan and Bentler 2010; Yang and Kim 2017). However, it is

difficult to recover the missing blocks with (multiple) imputation due to extensive missing blocks in SQD data. To address the defects of these traditional techniques for missing data, several methods were proposed that can manage multi-source/multi-modality data or fragmentary data with high missing rates. Xiang et al. (2014) presented a unified "bi-level" learning model for complete multi-source data and extended it to incomplete data, which can be generalized to other applications with block-wise missing data sources. Fang et al. (2019) developed a model averaging method for fragmentary data with missing completely at random, which fit candidate models using a vailable covariate data and combined them according to some criterion. Yu et al. (2020) proposed a direct sparse regression procedure by using covariance from multi-modality data (DISCOM) to find the optimal sparse linear prediction of a continuous response variable. However, none of these methods are designed for the block-wise missing data from a SQD. The SQD data differs from the fragmentary data in its "orderly" missing structure in the sense that there are only a few missing patterns despite the high missing rate. Thus, in addition to a well-designed split questionnaire, a good estimator for regression models based on the SQD data is also needed beyond the existing methods.

In this study, we consider a two-stage framework with respect to the aforementioned issues encountered in a large-scale survey with long questionnaires. We first focus on the design of a split questionnaire to reduce the cost of the survey while ensuring precision requirements on the estimates of the sample means of the question items, which are usually the first priority for most surveys. We then consider estimation of the linear regression model based on data obtained using a SQD. The two stages are not isolated, which means that the SQD needs to cope with both the survey cost and the subsequent estimation. Specifically, in our approach, we need a small sample size of the complete cases data obtained with full questionnaire in Stage I for correcting the possible bias of the estimate based on the SQD data in Stage II. Moreover, instead of assuming that the cost increases linearly with the number of questions, a quadratic cost function is introduced, which may be more practical in various situations. In Stage II, we propose a block weighted least squares (BWLS) estimation approach, which makes full use of all available data and has an explicit expression that is easy to implement. In particular, the proposed estimator has shown satisfactory performance in improving both estimation and prediction accuracies. Overall, this work provides a practical and useful design as well as an estimation approach for large-scale surveys in a two-stage framework.

The article is organized as follows. In Section 2, we propose a framework consisting of a SQD and an estimation of linear regression model based on the SQD data. Several numerical simulations are conducted in Section 3. In Section 4, we illustrate the performance of the proposed framework based on the ESS data in 2018, followed by a discussion in Section 5.

2. Method

2.1. Stage I: Nonlinear Cost-Based SQD

Consider a complete questionnaire containing p questions/items. The data collected for each question corresponds to one variable, which is assumed to be either continuous or

dichotomous. Suppose the p questions can be divided into m modules, which are presented by disjoined sets $M_1, ..., M_m \subseteq [p] = \{1, 2, ..., p\}$. Each module M_i corresponds to a particular aspect of the research and the number of questions in module M_i is denoted as p_i and $p = \sum_{i=1}^{m} p_i$. We assign these modules to different sub-questionnaires $Q_j, j = 1, 2, \cdots, q$. In particular, Q_1 is a full questionnaire containing all modules. Here, we use a structure matrix $\mathbf{A} = (a_{ij})_{m \times q}$ to describe the assignment, where $a_{ij} = 1$ indicates the module M_i appears in sub-questionnaire Q_j, and $a_{ij} = 0$ otherwise. Note that each module can be assigned to multiple sub-questionnaires. Such a structure matrix $\mathbf{A}$ is desired for the purpose of improving data quality and reducing survey cost.

To quantify the survey cost, we introduce the total cost function C,

$$C = \sum_{j=1}^{q} C_j = \sum_{j=1}^{q} (C_j^f + n_j \alpha_j), \tag{1}$$

where C_j is the cost of the sub-questionnaire Q_j, which consists of the fixed cost C_j^f and the varying cost $n_j \alpha_j$. The number of respondents filling the sub-questionnaire Q_j is n_j and Q_j contains $L_j = \sum_{i=1}^{m} a_{ij} p_i$ questions. Usually, C_j^f is incurred regardless of sample size, and it includes, but is not limited to, the upfront cost of implementing the survey, such as training interviewers, which is fixed for a given sub-questionnaire Q_j. The unit cost α_j is expressed as a function of L_j. Here, we use a quadratic function to describe the relationship between α_j and L_j:

$$\alpha_j = B_2 L_j^2 + B_1 L_j + B_0, j = 0, 1, ..., q,$$

where B_0, B_1, and B_2 are the coefficients, and we assume that $B_2 > 0$. If $B_1 \geq 0$, the cost function is nondecreasing with respect to L_j; If $B_1 < 0$ and $-\frac{B_1}{2B_2} \geq 2$, α_j first decreases and then increases with L_j. Here we adopt the quadratic form instead of linear form because we would like to take the difficulty in recruiting respondents into consideration. As the number of questions L_j increases, it becomes more challenging to find qualified respondents, hence increases our unit cost due to the reduced number of respondents. It is possible that the unit cost decreases at a small number of questions. This is because it is relatively easy to recruit respondents with a questionnaire of a reasonable small length. We admit the true unit cost function can be complex and even mathematically intractable. The quadratic form is only an approximation to the reality. But such a modification allows us to offer a more realistic and flexible solution to the issue. In the quadratic cost function, B_0 is the "baseline cost" when $L_j = 0$, and $\max(-B_1/2B_2, 1)$ is the critical point, beyond which the unit cost will increase with the number of questions. Generally, these parameters can be determined according to practical considerations or operational data. Figure 1 demonstrates an example of α_j with $B_2 = 0.1, B_1 = -1, B_0 = 10$. Note that when $B_2 = 0$, the cost function corresponds to the linear cost function (Ioannidis et al. 2016).

As mentioned above, the total cost C depends on the design, namely the structure matrix $\mathbf{A}$ as well as the sample size n_j of sub-questionnaire Q_j. We use $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$ to stress the relationship between C and the design, where $\boldsymbol{n}_\mathbf{A} = (n_1, ..., n_q)^\top$ is the vector of sample sizes. In this study, the goal of the SQD is to find an optimized matrix $\mathbf{A}$ and the corresponding sample sizes $\boldsymbol{n}_\mathbf{A}$ with the goal of minimizing the total cost $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$. In addition, we impose certain constraints on the minimum sample size for each module to ensure the estimation precision for each item mean μ. Let $\hat{\mu}_l$ be the sample mean of the l-th item. Then $n_i^* = \max_{l=1,\cdots,p_i} (\frac{z_{1-\alpha}^2 \hat{\sigma}_l^2}{e_l^2})$ is the minimum

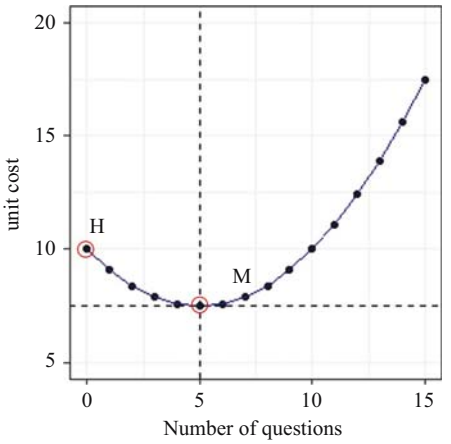


Fig. 1. The curve of cost function.

sample size for the i-th module, ensuring the estimation precision of $\hat{\mu}_l$ to achieve the absolute error of at most e_l. Here e_l is a prespecified value for precision of estimation, and $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution, $\hat{\sigma}_l^2$ is sample variance of the l-th response item in the i-th module, estimated by the historical data. A general form of the design is given in Table 1, where the first m rows and q columns with 0 or 1 denote the structure matrix $\boldsymbol{A}$. Then the optimal design can be determined by solving the optimization problem:

$$\begin{aligned} &\min_{\mathbf{A}, n_{\mathbf{A}}} \; C(\mathbf{A}, \boldsymbol{n}_{\mathbf{A}}), \\ &s.t. \quad \mathbf{A}\boldsymbol{n}_{\mathbf{A}} > \boldsymbol{n}^*, \end{aligned} \tag{2}$$

where $\boldsymbol{n}^* = (n_1^*, ..., n_m^*)^\top$.

For the optimization problem (2), we use the simulated annealing algorithm (Ioannidis et al. 2016) to obtain the structure matrix $\mathbf{A}$ and the sample sizes $\boldsymbol{n}_{\mathbf{A}}$. Different from Ioannidis et al. (2016), we require a small sample size of the full questionnaire for correcting the bias of the estimate of the linear regression model based on the SQD data in Stage II. Therefore, we fix $a_{i1} = 1$ for all $i = 1, ..., m$, and require $n_1 > p$. After obtaining

Table 1. The general form of SQD.

	Q_1	Q_2	. . .	Qq	n_i^*
M_1	1	1	. . .	1	n_1^*
M_2	1	0	. . .	0	n_2^*
⋮	⋮	⋮	⋮	⋮	⋮
M_m	1	1	. . .	0	n_m^*
n_j	n_1	n_2	. . .	n_q	-
L_j	L_1	L_2	. . .	L_q	-

the optimal design of the questionnaire, we randomly assigned different sub-questionnaires to different interviewees. The collected data is of planned missing type since we are aware of which modules are included in each sub-questionnaire before data collection.

2.2. *Stage II: Block Weighted Least Squares Estimation*

Based on the nonlinear cost SQD in Stage I, we obtained the survey data with block-wise missing structure. As shown in Figure 2, the response variable $\boldsymbol{Y}$ is fully observed, while the design matrix $\boldsymbol{X}$ is block-wise missing as n_j interviewees only fill out L_j questions in the j-th sub-questionnaire Q_j for $j > 1$. Besides the item means, we are interested in the estimation of the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n).$$

To this end, we propose a block weighted least squares procedure (BWLS) to make full use of the block-wise missing data.

Specifically, we let $r_{il}, l = 1, 2, ..., p$ be the missing indicator such that $r_{il} = 1$ means the i-th respondent answers the l-th question, and $r_{il} = 0$ means missing. Let $\Omega_j, j = 1, ..., q$, be the set of indices of questions in the j-th sub-questionnaire Q_j and $\Omega_1 = [p]$. Denote the total sample size as $n = \sum_{j=1}^{q} n_j$. For $1 \le j \le q$, let $\mathcal{S}_j = \{i : r_{il} = 1 \text{ for } l \in \Omega_j, 1 \le i \le n\}$ denote the set of respondents who have answered at least the questions in Ω_j. Hence $\mathcal{S}_1$ is the set of complete cases. In addition, we define $\mathcal{S}_j^* = \{i : r_{il} = 1 \text{ for } l \in \Omega_j, r_{il} = 0 \text{ for } l \notin \Omega_j, 1 \le i \le n\}$ as the set of respondents who have answered exactly the questions in Ω_j, that is, the sub-questionnaire Q_j. Hence we have $\mathcal{S}_j^* \subseteq \mathcal{S}_j$ for $1 \le j \le q$. Let $|\mathcal{S}|$ denote the cardinality of a set $\mathcal{S}$, then $n_j = |\mathcal{S}_j^*| \le |\mathcal{S}_j|$ for $1 \le j \le q$. Furthermore, define $\boldsymbol{Y}_{(j)} \in R^{|\mathcal{S}_j|}, \boldsymbol{X}_{(j)} \in R^{|\mathcal{S}_j| \times |\Omega j|}$, where $R^{|\mathcal{S}_j|}$ denotes the $|\mathcal{S}_j|$-dimensional real space, and also define $\mathcal{D}_j = \{\boldsymbol{Y}_{(j)}, \boldsymbol{X}_{(j)}\}$ for $j = 1, ..., q$. Assume $\tilde{\boldsymbol{X}}_{(j)} = (1, \boldsymbol{X}_{(j)})$, where 1 is a column of 1 with size of n_j. Based on the complete cases $\mathcal{D}_1$, we can get the complete case (CC) least squares estimator

$$\hat{\boldsymbol{\beta}}_{cc} = \left(\tilde{\boldsymbol{X}}_{(1)}^{\top} \tilde{\boldsymbol{X}}_{(1)}\right)^{-1} \tilde{\boldsymbol{X}}_{(1)}^{\top} \boldsymbol{Y}_{(1)}.$$

As discussed earlier, n_1 is usually set to be not too large compared to n_j for $j = 2, ..., q$ due to budget and respondent load considerations. Thus, the efficiency of $\hat{\boldsymbol{\beta}}_{cc}$ is typically

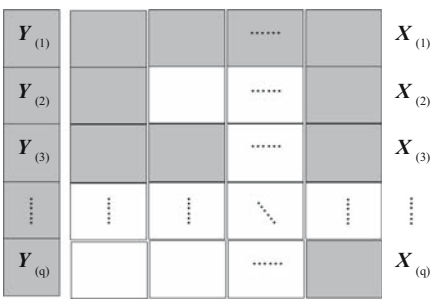


Fig. 2. An example for data structure.

unsatisfactory. Hence we propose an improved least squares estimator, with which we calculate the weights in the BWLS estimator.

Note that the covariance matrix of $\hat{\beta}_{cc}$ is $\sigma^2(\tilde{\boldsymbol{X}}_{(1)}^\top \tilde{\boldsymbol{X}}_{(1)})^{-1}$, and the sample covariance matrix $\boldsymbol{S}_{xx} = \tilde{\boldsymbol{X}}_{(1)}^\top \tilde{\boldsymbol{X}}_{(1)}/n_1$ is invertible but may be numerically unstable since n_1 is not large and p/n_1 is close to 1. Such a setting usually results in a dramatically increased estimation error, hence we use a well-conditioned covariance matrix to replace $\boldsymbol{S}_{xx}$ (Ledoit and Wolf 2004). Specifically, we consider an optimal linear shrinkage estimator for the covariance matrix, which has the form

$$\hat{\boldsymbol{\Sigma}}_{xx} = \hat{a}_1 \boldsymbol{I}_{p+1} + \hat{a}_2 \boldsymbol{S}_{xx},$$

where $\boldsymbol{I}_{p+1}$ is a $(p+1)\times(p+1)$ identity matrix, $\hat{a}_1$ and $\hat{a}_2$ are the coefficients determined by the optimization problem (4) of Ledoit and Wolf (2004). Specifically, $\hat{a}_1 = \frac{\alpha_n b_n^2}{d_n^2}$ and $\hat{a}_2 = \frac{a_n^2}{d_n^2}$, where $\alpha_n = tr(\boldsymbol{S}_{xx})/(p+1)$, $d_n^2 = \|\boldsymbol{S}_{xx} - \alpha_n \boldsymbol{I}_{p+1}\|_n^2$, $b_n^2 = \min\{d_n^2, \frac{1}{n_1^2}\sum_{i=1}^{n_1} \|\tilde{\boldsymbol{X}}_{(1),i} \tilde{\boldsymbol{X}}_{(1),i}^{\mathrm{T}} - \boldsymbol{S}_{xx}\|_n^2\}$, $a_n^2 = d_n^2 - b_n^2$, and $\tilde{\boldsymbol{X}}_{(1),i}$ is the i-th column of $\tilde{\boldsymbol{X}}_{(1)}$. Note that $\hat{\boldsymbol{\Sigma}}_{xx}$ is similar in spirit to ridge regression (Hoerl and Kennard 1970). Then the improved complete case (ICC) estimator is

$$\hat{\boldsymbol{\beta}}_{Icc} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \boldsymbol{S}_{xy},$$

where $\boldsymbol{S}_{xy} = \tilde{\boldsymbol{X}}_{(1)}^\top \boldsymbol{Y}_{(1)}$. Although the well-conditioned covariance matrix gives $\hat{\boldsymbol{\beta}}_{Icc}$ better performance, it still ignores the data from other sub-questionnaires and needs further improvement.

To make full use of the data from all other sub-questionnaires, we use the weighted least squares estimation. Based on the CC data $\mathcal{D}_1$ from the full questionnaire Q_1, we can get an unbiased estimator $\hat{\boldsymbol{\beta}}_{cc}$. Meanwhile, the incomplete data $\mathcal{D}_j, j = 2, ..., q$, yield possibly biased estimators $\hat{\boldsymbol{\beta}}^{(j)} = (\tilde{\boldsymbol{X}}_{(j)}^\top \tilde{\boldsymbol{X}}_{(j)})^{-1} \tilde{\boldsymbol{X}}_{(j)}^\top \boldsymbol{Y}_{(j)}$. Then we define $\hat{\boldsymbol{\beta}}_{sub}^{(j)} = \Pi_j^\top (\tilde{\boldsymbol{X}}_{(j)}^\top \tilde{\boldsymbol{X}}_{(j)})^{-1} \tilde{\boldsymbol{X}}_{(j)}^\top \boldsymbol{Y}_{(j)}$ where Π_j is the $L_j \times p$ projection matrix such that $\Pi_j \boldsymbol{\beta} = \boldsymbol{\beta}_{(j)}$, and $\boldsymbol{\beta}_{(j)} \in R^{|\Omega_j|}$. Since $\hat{\boldsymbol{\beta}}_{sub}^{(j)}$ have different biases for different Q_js, we propose to treat samples from Q_js differently in estimating the linear regression model. Specifically, we assign different weights to $\mathcal{D}_j$s for balancing the information from various sub-questionnaires. The samples $\mathcal{S}_j$ should be assigned large weight if the sub-questionnaire Q_j provides important information in the sense that it yields a smaller residual sum of squares. With this in mind, the weight is set to be the reciprocal of the average of residual sum of squares. The j-th weight for samples $\mathcal{S}_j^*$ given by

$$w_j = \frac{|\mathcal{S}_j|}{\sum_{i\in\mathcal{S}_j} \left(Y_i - \tilde{\boldsymbol{X}}_{(j)} \Pi_j \hat{\boldsymbol{\beta}}_{Icc}\right)^2},$$

which means we compute the weights for D_j with more available cases since $\mathcal{S}_j^* \subseteq \mathcal{S}_j$. It is noticed that the denominator in w_j is the prediction error using the items in Q_j and the estimates of the associated regression coefficients are obtained from the respondents that answer the full range of questions in Q_1. Intuitively, $\frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j} (Y_i - \tilde{\boldsymbol{X}}_{(j)} \Pi_j \hat{\boldsymbol{\beta}}_{Icc})^2$ measures the missing information of the sub-questionnaire Q_j, and we should lower the weight of these samples if the corresponding sub-questionnaire misses too much information. Based

on the weights, we can get the final estimator by minimizing the weighted loss function

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{q} w_j \sum_{i \in \mathcal{S}_j^*} \left(Y_i - \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\beta} \right)^2,$$

where $\tilde{X}$ is the imputed matrix of $\boldsymbol{X}$ and replacing the missing values with the mean of each variable and $\tilde{X}_j$ is the i-th row of $\tilde{X}$. In fact, the final BWLS estimator has the explicit formulation

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\boldsymbol{X}}^\top \boldsymbol{W} \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{W} \boldsymbol{Y},$$

where $\boldsymbol{W} = \text{diag}(w_1 \boldsymbol{I}_{n_1}, ..., w_q \boldsymbol{I}_{n_q})$ is the weight matrix. When $\boldsymbol{W}$ is an identity matrix, the estimator degenerates to the OLS estimator using the mean imputed values of the missing data items. The final estimator $\hat{\beta}$ uses information from all respondents, and assigns different weights to samples according to the information loss from different sub-questionnaires, which can reduce the bias induced by the block-wise missing data while improving the CC estimator by reducing its variance. Strictly speaking, we need the assumption of missing completely at random (MCAR), since the weight involves the improved CC estimator $\hat{\beta}_{Icc}$, which may not be effective when the assumption is violated. But the numerical studies show that the BWLS estimator also has a satisfactory performance under missing at random.

In summary, in Stage I, we create a survey design by minimizing a nonlinear cost function under the constraints of the variances of item means. Based on the SQD data obtained by implementing the design in Stage I, we develop a BWLS estimator for linear regression in Stage II. We would like to note that, if the item means are the target, the best linear unbiased estimator (BLUE) for a SQD (Chipperfield and Steel 2009), combining different estimates in an optimal way, is a good choice and a more practical estimator proposed by Merkouris (2015) can also be used when the number of variables or the sizes of the samples are large.

3. Simulation

In this section, we conduct multiple sets of simulations to evaluate the performance of the proposed two-stage framework. First, we conduct a simulation based on the ESS data set to highlight the advantages of our nonlinear cost-based design. Second, we compare the BWLS estimator with alternatives based on simulated block-wise missing data and the SQD data obtained in Subsection 3.1.

3.1. Simulation for Design

In this subsection, we demonstrate the merit of the proposed nonlinear cost-based design using the ESS data set, and more information about this data set will be given later in Section Here, the response of interest is how the respondent is satisfied with the way democracy works in their country, and it is measured using an 11-point Likert scale ranging from zero to ten. Among over 500 questions, we included the 87 questions with the largest correlations with the response into the following process of SQD. These questions are divided into four modules $M_1, ..., M_4$ according to their different focuses.

For the design of the questionnaire, we consider the complete questionnaire as well as the following SQD methods:

(a). "Proposed": The nonlinear cost-based design proposed in this article with $B_2 = 1$, $B_1 = -10$ and $B_0 = 40$ and the fixed design cost $C_j^f, j = 1, ..., q$ is set to be zero.

(b). "Linear": The integrated survey design proposed in Ioannidis et al. (2016), where the matrix $\boldsymbol{A}$ is obtained based on a linear cost function. Specifically, the unit cost α_j is defined as

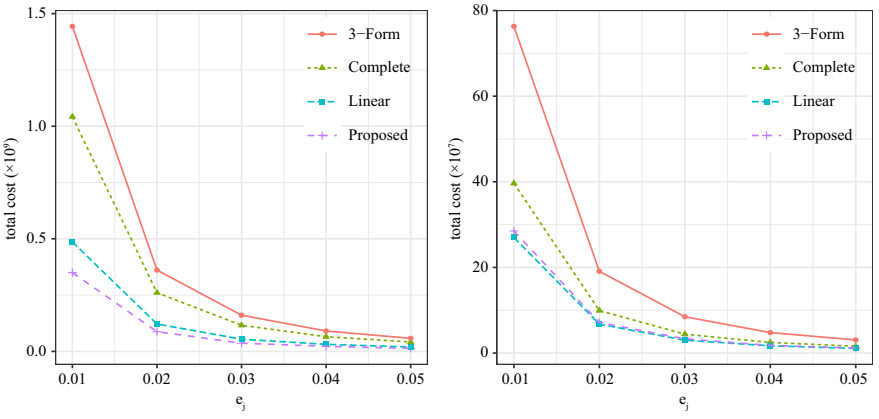
$$\alpha_j = B_1' L_j + B_0', \quad j = 0, 1, ..., q,$$

where B_1' represents for the unit cost per question and B_0' is the "baseline cost" when $L_j = 0$. The total cost $C(\mathbf{A}, \boldsymbol{n}_\mathbf{A})$ is the same as Equation (1). For fair comparison, we set $B_0 = B_0' = 40$ to ensure that the designs under linear cost function and nonlinear cost function share the same "baseline cost" when $L_j = 0$. For B_1', we first identify a possible range of it by considering that α_js under linear and non-linear cost function are equal for the number of questions $L_j = 87$. We find that the unit cost is the same as the proposed nonlinear design when $B_1' = 77$, that is $\alpha_j = 6739$ for $L_j = 87$ under both designs. Within the range of B_1', we search the optimal value from (0, 77] so that the total linear cost is minimum given the absolute error $e_l = 0.03$, and we obtained $B_1' = 29$.

(c). "3-Form": The three-form design studied in Rhemtulla and Little (2012).

Here, we consider two situations where the true unit cost function is assumed to be quadratic with $\alpha_j - L_j^2 - 10L_j + 40$ in Situation 1 and linear with $\alpha_j = 29L_j + 40$ in Situation 2, respectively. Considering that only 29 variables are included in the subsequent linear regression analysis, we set $n_1 = 39$ in the proposed design. For both "Linear" and "3-Form", the sample size of the full questionnaire Q_1 is set to be the same as the proposed design. For all designs, the first module M_1 collects information on demographic characteristics, and it appears in all sub-questionnaires. In the "3-Form" design, each sub-questionnaire is composed of the first module M_1 as well as two of the other three modules. We compare the total cost and the average number of questions answered per respondent with different absolute errors.

The results of total cost are shown in Figure 3. The total cost decreases with the increase of the allowed e_l for all designs in both situations. When the true unit cost is quadratic, the proposed design yields the lowest total cost among the four methods across all the different e_ls. The gap of the total cost between the "Linear" and the proposed design decreases with the increase of the allowed e_l. Given e_l, the total cost of "3-Form" design is larger than that of the complete questionnaire, primarily because the sample size of each sub-questionnaire must meet the most stringent requirement of modules contained in "3-Form" design, which leads to substantial extra cost. Specifically, the first module, M_1, appears in all sub-questionnaires and its required sample size is the largest among all modules for a given e_l, denoted as n_{max}. Ignoring the fixed cost C_j^f for simplicity, the total cost of the complete questionnaire is $n_{max}\alpha_1$, where α_1 is the unit cost for respondents answering the full-questionnaire with $L_1 = 87$ questions. However, in "3-Form" design, the sample size of each sub-questionnaire is also n_{max} since each sub-questionnaire includes module M_1. Then the three sub-questionnaires of sample size n_{max} and a full-questionnaire of sample size n_1 yield a total cost, $n_1\alpha_1 + n_{max}(\alpha_2 + \alpha_3 + \alpha_4)$. Here, we fix



(a) The true cost function is quadratic. (b) The true cost function is linear.

Fig. 3. The comparasion of total cost for Situation 1 (a) and Situation 2 (b).

$n_1 = 39$ for the full questionnaire to ensure the feasibility of BWLS estimator. Apparently, we have $L_2 + L_3 + L_4 > L_1$, and this is why the total cost of "3-Form" design is larger than that of the complete questionnaire.

To quantify the respondent burden, the average number of questions answered per respondent defined as $\bar{L} = \sum_{j=1}^{q} L_j n_j / \sum_{j=1}^{q} n_j$ are calculated for different methods with $e_l = 0.03$, which are "Proposed": 37.97, "Linear": 58.80, "3-Form": 62.58, and "Complete": 87, respectively. It is clear that the proposed method yields the minimum number of questions answered per respondent. We also calculate $\bar{L}$ when $e_l = 0.01, 0.02, 0.04$, and 0.05. The average number of questions answered per respondent for the complete design is always 87, which has been reduced by 53% with the proposed method on average. Moreover, the "3- Form" and "Linear" designs reduce the average number of questions by 28% and 32% on average, respectively. Overall, the proposed non-linear cost-based design outperforms the alternatives in terms of both total cost and the average number of questions answered per respondent.

For a demonstration, setting the absolute error to be 0.05, the resulting design with the proposed method is shown in Table 2. The column n^{**} is the actual sample size of module M_i based on the proposed design.

Table 2. The structure matrix.

	Q_1	Q_2	Q_3	Q_4	n_i^*	n_i^{**}
Module 1	1	1	1	1	6186	6186
Module 2	1	0	1	0	4095	4095
Module 3	1	1	0	1	2063	2130
Module 4	1	1	1	0	4329	4329
n_j	39	234	4056	1857	-	-
L_j	87	46	79	16	-	-

3.2. Simulation for Estimators

In this subsection, we compare the BWLS estimator with alternatives, including the CC estimator $\hat{\beta}_{CC}$, the improved CC estimator (I-CC) $\hat{\beta}_{ICC}$, a model average estimator (Fang et al. 2019, MA-CV), and an imputation-based estimator termed SI-LS here, which replaces missing values in a large-scale matrix using the SOFT-IMPUTE algorithm proposed in Mazumder et al. (2010) and performs a least squares estimation for $\boldsymbol{\beta}$ based on the single imputed data set.

Simulation 1. We generate data from the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is generated from $N(0, \sigma^2)$. The covariates $\boldsymbol{X}$ are generated from $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the corresponding $p \times p$ covariance matrix, with the structure as

$$\Sigma = \begin{bmatrix} \Sigma_{in} & \cdots & \Sigma_{off} \\ \vdots & \ddots & \vdots \\ \Sigma_{off} & \cdots & \Sigma_{in} \end{bmatrix}.$$

The non-diagonal and diagonal elements in $\boldsymbol{\Sigma}_{\text{in}}$ are λ_1 and one, respectively. Additionally, $\boldsymbol{\Sigma}_{off}$ is a constant matrix with element λ_2. Here, λ_1 is the correlation between questions within the same module, while λ_2 is the correlation between questions from different modules. We consider two cases of correlation structures: $\lambda_1 = 0.3$, $\lambda_2 = 0.1$ in Case 1, and $\lambda_1 = 0.5$, $\lambda_2 = 0.2$ in Case 2. The variance σ^2 takes different values such that $R^2 = Var(\boldsymbol{X}^\top \boldsymbol{\beta})/\{Var(\boldsymbol{X}^\top \boldsymbol{\beta}) + \sigma^2\} = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. We consider two settings for co-efficients with different sparsities. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(m)})$, where $\boldsymbol{\beta}^{(i)}, i = 1, ..., m$ is the coefficients of the variables in i-th module. In Setting I, $\boldsymbol{\beta}^{(1)} = (0.5, -0.5, \ 0.5, 1)$, $\boldsymbol{\beta}^{(2)} = (0.5, -0.5, -0.3, 0.5)$, $\boldsymbol{\beta}^{(3)} = (0, 0, 0, 0)$, $\boldsymbol{\beta}^{(4)} = (0, 0, 0, 0)$, $\boldsymbol{\beta}^{(5)} = (-0.5, 0.5, 1, 0.3)$, $\boldsymbol{\beta}^{(6)} = (1, 0.3, 0.3, 1)$, $\boldsymbol{\beta}^{(7)} = (-1, 1, 0.3, 0.3)$, $\boldsymbol{\beta}^{(8)} = (1, 0.5, 0.5, 0.5)$, $\boldsymbol{\beta}^{(9)} = (-0.5, 0.5, 0.3, -0.3)$ and $\boldsymbol{\beta}^{(10)} = (0.3, -0.3, 1, 0.3)$, thus M_3 and M_4 are inactive modules. In Setting II, we consider sparsity within modules, $\boldsymbol{\beta}^{(1)} = (-1, -1, 0, \ 1,)$, $\boldsymbol{\beta}^{(2)} = (0, 1, -1, 0.3)$, $\boldsymbol{\beta}^{(3)} = (0, 1, -0.5, -1)$, $\boldsymbol{\beta}^{(4)} = (0.3, 0.3, 0.3, 0)$, $\boldsymbol{\beta}^{(5)} = (0.3, \ 0.5, 1, 0)$, $\boldsymbol{\beta}^{(6)} = (-1, 0, 0, 1)$, $\boldsymbol{\beta}^{(7)} = (0.3, 0, 0, 0.3)$, $\boldsymbol{\beta}^{(8)} = (0, 0, 0.5, -0.3)$, $\boldsymbol{\beta}^{(9)} = (0.3, 0.5, 1, -1)$, and $\boldsymbol{\beta}^{(10)} = (-0.3, 1, 0.5, -0.3)$. In both settings for coefficients, we set $\beta_0 = 1$. We further consider multiple scenarios with different dimensions of covariates, number of modules, and sample sizes:

S1. $p = 20, m = 5, n_1 = 30, n_2 = ... = n_5 = 200$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(5)})$.

S2. $p = 40, m = 5, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting I for $\boldsymbol{\beta}$ (here, we merge $\boldsymbol{\beta}^{(i)}, i = 1, 2, ..., 10$ in to 5 modules pairwise).

S3. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting II for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

S4. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 200$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

S5. $p = 40, m = 10, n_1 = 45, n_2 = ... = n_{10} = 500$, with Setting I for $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(10)})$.

For each scenario, there are 2^m sub-questionnaires, of which we randomly select m sub-questionnaires. For all the settings, $B = 500$ replicates are simulated. The performances of different estimation methods are evaluated in terms of prediction error (PE) and mean square error (MSE), defined as

$$\text{PE} = \frac{1}{Bn_{test}} \sum_{b=1}^{B} \sum_{i=1}^{n_{test}} \left(\hat{Y}_i^{(b)} - E(Y_i) \right)^2, \quad \text{MSE} = \frac{1}{B(p+1)} \sum_{b=1}^{B} \sum_{l=0}^{p} \left(\hat{\beta}_i^{(b)} - \beta_l \right)^2,$$

respectively. Here $\hat{\beta}_i^{(b)}$ is the estimate of β_l, $\hat{Y}_i^{(b)}$ is the predicted value for Y_i in the b-th replicate and $E(Y_i) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. n_{test} is the sample size of the testing data set, which is set to be 200 in this simulation. Since the MSE can be decomposed into MSE = Variance + Bias^2, we have

$$\frac{1}{B(p+1)} \sum_{b=1}^{B} \sum_{l=0}^{p} (\hat{\beta}_l^{(b)} - \beta_l)^2 = \frac{1}{p+1} \sum_{l=0}^{p} \frac{1}{B} \sum_{b=1}^{B} (\hat{\beta}_l^{(b)} - \hat{\overline{\beta}}_l)^2 + \frac{1}{p+1} \sum_{l=0}^{p} (\hat{\overline{\beta}}_l - \beta_l)^2,$$

where $\hat{\overline{\beta}}_l = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_l^{(b)}$. For Case 1, the results are shown in Figures 4, 5 and 6. The results of Case 2 are shown in Figures 8, 9 and 10 in Appendix (Section 6).

Figures 4 and 5 show the comparison of the MSE and PE among different methods. We purposely truncated the figure for ease of comparison, so results with large values do not appear in certain figures. We can see the estimation and prediction accuracies gradually improve with the increase of the signal-to-noise ratio R^2. In most of the considered scenarios, the MSE of the BWLS estimator is the smallest among the five methods except for large R^2. Additionally, a larger sample size of sub-questionnaire n_i leads to a lower MSE of the proposed estimator, as is expected. The result of PE also shows the superiority of BWLS. The MSE and PE of all five estimators are increasing as the number of variables increases.

From Figure 6, we find that the variance of the BWLS estimator is significantly smaller than that of the alternatives. This is mainly due to the full utilization of the samples at hand. Moreover, when the correlation between sub-questionnaires is much smaller, the bias of the BWLS estimator is also small. Overall, I-CC outperforms the CC estimator, but is inferior to MA-CV. In addition, the performance of SI-LS is barely satisfactory with regard to MSE and PE, especially when compared with BWLS.

The results of Case 2 are displayed in the Appendix (Section 6). Figures 8, 9 and 10 demonstrate the comparison of MSE and PE among the different estimators and the decomposition of MSE. In Case 2, we can draw similar conclusions as Case 1. However, not surprisingly, BWLS performs worse in MSE and PE as λ_2 increases and the bias of BWLS is getting larger with the increasing correlation between modules. It suggests that the correlation between modules should be lower than that within modules when dividing the complete questionnaire into shorter parts.

Additionally, results about the comparison between BWLS, mean imputation, and multiple imputation (MI) are shown in Appendix (Section 6), where mean imputation replaces the missing values with the item means and multiple imputation is conducted with

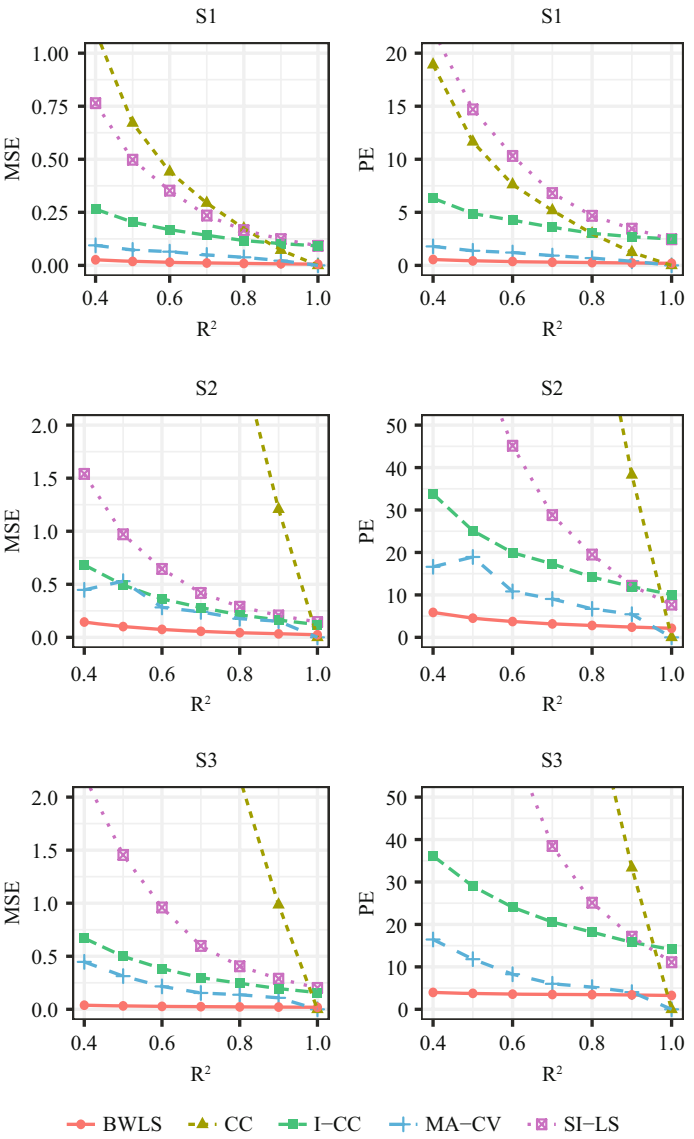


Fig. 4. The comparison of MSE and PE for different estimators in S1–S3 under case 1.

the MICE package in R. From the results, we can see that BWLS outperforms the mean imputation in most settings, especially in terms of PE. The superiority of BWLS in MSE is not that obvious, and this is mainly because the estimated weights in BWLS increase the variance of the estimator while the weights in mean imputation are constantly equal to 1. Besides, BWLS performs better than MI in terms of MSE for almost all scenarios we considered here. And BWLS also shows its superiority in terms of PE in most scenarios.

Simulation 2. Based on the design given in Table 2 with ESS data, we conduct a real data-based simulation to compare the proposed estimator BWLS with the alternatives as in

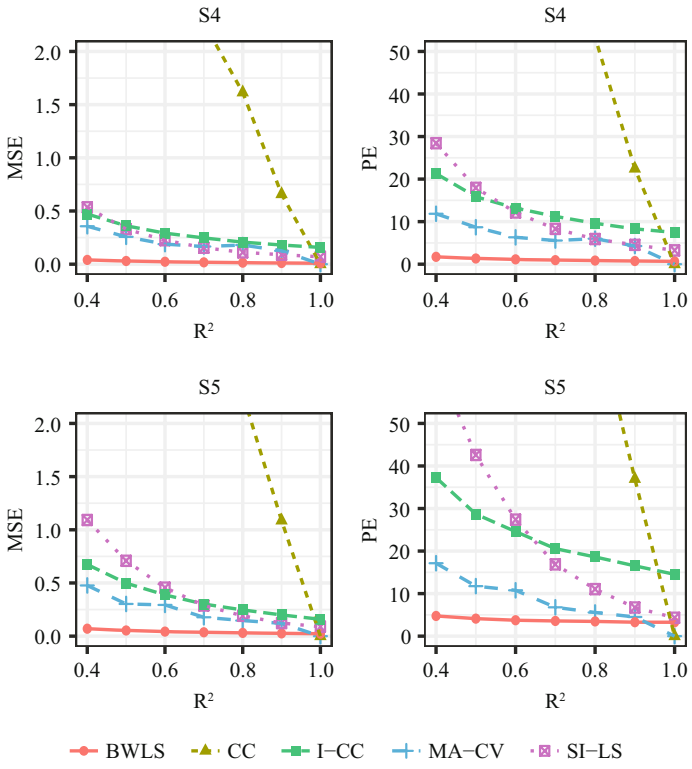


Fig. 5. The comparison of MSE and PE for different estimators in S4–S5 under case 1.

Simulation 1. We regress the response variable on each variable using the full data and select the top 29 variables with the smallest p-value for subsequent analysis. The selected variables are listed in Table 3. Based on the observations of these 29 variables, we generate Y from the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{29} X_{29} + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$. The variance of ϵ, σ^2 takes different values such that R^2 ranging from 0.4 to 1, and $\boldsymbol{\beta} = (1, -1, -1, 0, 1, 0, 1, -1, 0.3, 0, 1, -0.5, -1, 0.3, 0.3, 0.3, 0, 0.3, 0.5, 1, 0, -1, 0, 0, 1, 0.3, 0, -1, 0.3, 0.5, 0)^\top$. The estimation results based on 500 replications for BWLS, CC, I-CC, MA-CV, and SI-LS are presented in Figure 11 of Appendix (Section 6). We can draw similar conclusions as in Simulation 1 that BWLS has the best prediction and estimation performance.

4. ESS Data Analysis

The European Social Survey (ESS) was established at the National Centre for Social Research in London in 2001. Since 2003, the ESS Headquarters have been located at City, University of London in the UK. The survey measures the attitudes, beliefs, and behavior patterns of various populations, covering more than 500 questions and reaching over 40,000 respondents from more than 30 countries all over Europe. It's desirable to improve its design since such a large-scale survey requires significant resources (e.g., time and

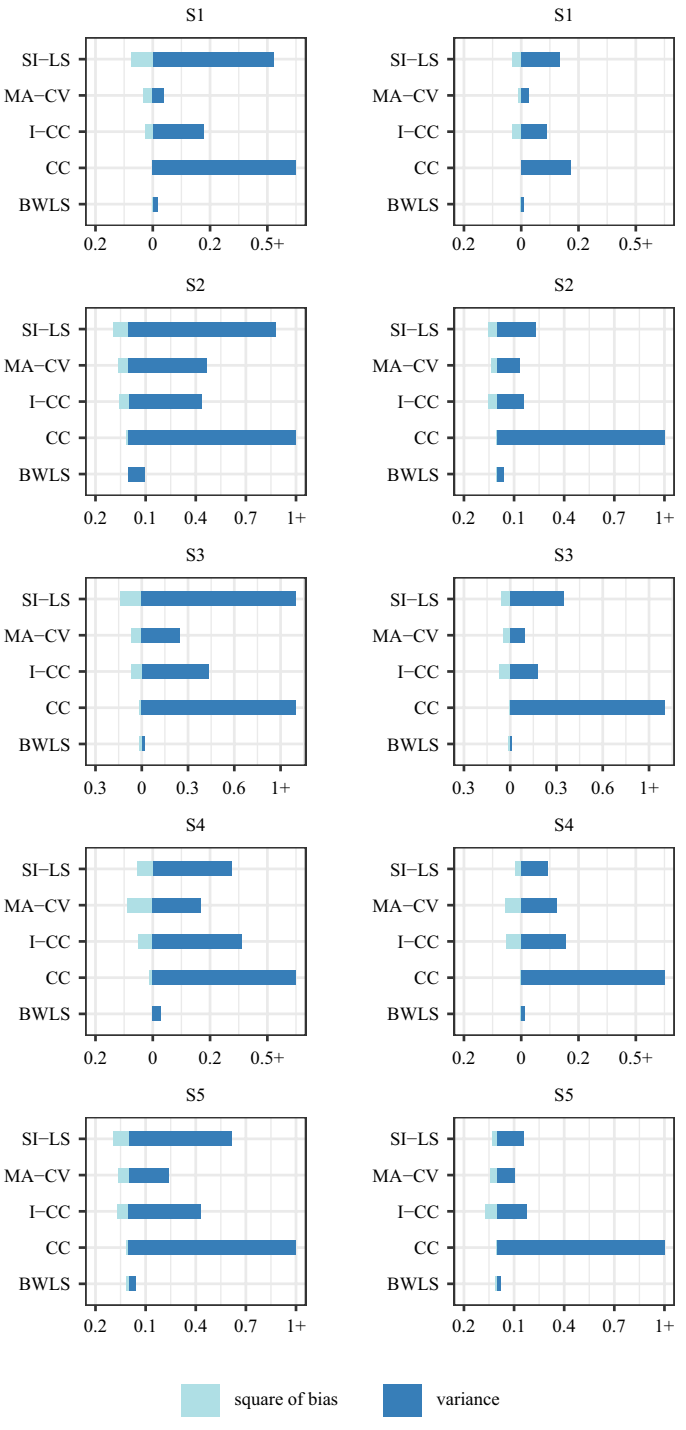


Fig. 6. The decomposition of MSE in case 1.

Table 3. The list of variables.

Module	Variable	Definition
	stfdem	How satisfied with the way democracy works in their country
	agea	Age of respondent, calculated
M_1	gndr	Gender
	eduyrs	Years of full-time education completed
	trstprl	Trust in country's parliament
	trstlgl	Trust in the legal system
	trstplc	Trust in the police
	trstprt	Trust in political parties
	trstun	Trust in the United Nations
	trstplt	Trust in politicians
	trstep	Trust in the European Parliament
M_2	stfeco	How satisfied with present state of the economy in their country
	stfedu	State of education in their country nowadays
	stfgov	How satisfied with the country's national government
	psppsgva	Political system allows people to have a say in what government does
	psppipla	Political system allows people to have an influence on politics
	frprtpl	Political system in their country ensures everyone a fair chance to participate in politics
	gvintcz	Government in their country takes into account the interests of all citizens
	poltran	Decisions in their country politics are transparent
	anvcld	Approves if a person chooses never to have children
	alvgptn	Approves if a person lives with partner they are not married to
M_3	acldnmr	Approves if a person has a child with a partner they are not married to
	aftjbyc	Approves if a person has full-time job while having children aged under three
	advcyc	Approves if a person gets divorced while having children aged under 12
	hinctnta	Household's total net income, all sources
	ifrjob	Compared with other people in their country, there is a fair chance of getting job I seek
	happy	How happy are you
M_4	recskill	Influence decision to recruit in their country: person's knowledge and skills
	ifredu	Compared to other people in their country, there is a fair chance of achieving the level of education I seek
	hincfel	Feeling about their household's income nowadays

money). Based on the proposed framework, we have finished the SQD in Subsection 3.1 and we use the SQD data obtained in Table 2 for the following regression analysis.

The ESS study concentrates on how the respondent is satisfied with the way democracy works in their country, and we choose *stfdem* as the response variable, which is measured

using an 11-point Likert scale ranging from zero to ten and treated as a continuous variable like Wu and Leung (2017). We apply the linear regression model to study the relationship between the response variable and the relevant factors in this section. As pointed out in Van Ham et al. (2017), economic outcomes and quality of governance should go a long way in explaining one's satisfaction with democracy. There are many other factors affecting popular satisfaction with the way democracy works according to some studies in political science, including the individual-level perceptions and aggregate measures based on expert and stakeholder surveys of corruption, fair and honest treatment by political officials, or impartiality. Based on the considerations above, we incorporate 29 covariates from four modules into the regression model (Table 3). Specifically, module M_1 is composed of three variables concerned with respondent background, that is age, gender, and years of education. Modules M_2, M_3, and M_4 consist of questions related to the attitude toward politics and society, marital relations and children, and the personal feelings about life, respectively. The correlation of these variables is shown in Figure 7. It illustrates the correlations between items within/between different modules. Specifically, the variables *trstprl*, *trstlgl*, ..., *poltran* are grouped into module M_2, which show obvious positive correlation. The variables *anvcld*, *alvgptn*, ..., *advcyc* are about the respondents' attitudes towards marriage and childbearing and make up the module M_3. Module M_4 are consist of variables *hinctnta*, *ifrjob*, ..., *hincfel*, which also show a positive correlation in Figure 7.

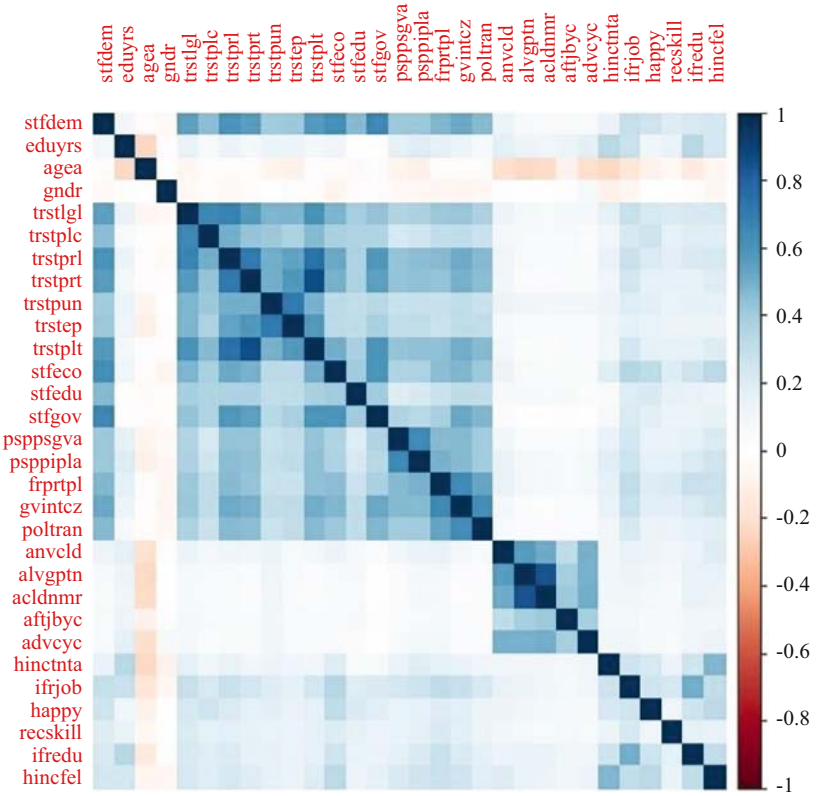


Fig. 7. The correlations between different variables.

With the proposed design, the assignment of modules in different sub-questionnaires can be obtained by minimizing the cost function as in simulation for design. The sample sizes of the four sub-questionnaires are shown in Table 2, and the total sample size is $n = 6{,}186$. After deleting the respondents with missing values, we finally obtain a complete data set containing 27,341 valid respondents, with 6,186 of them being used as the training data and the remaining 21,155 respondents being used for testing. For the training data set, the respondents are randomly assigned to one of the four sub-questionnaires. Based on the data with SQD in Table 2, we estimate the regression coefficients with BWLS, CC, I-CC, MA-CV, and SI-LS. The results based on 100 random assignments are shown in Table 4.

Table 4 shows the estimates of coefficients with training data as well as the prediction error (PE) of the testing data, with the standard deviation of the 100 replications contained in parentheses. In general, the PE of BWLS is the smallest among the five methods tested. The smaller standard error also indicates the superiority of BWLS. For interpreting the coefficients, we note that the response variable is likely to be driven by political systems' outputs and outcomes, such as economic performance (Van Ham et al. 2017). When the

Table 4. The estimates of coefficients and PE with 100 replications, where the averaged standard error is in parentheses.

Variable	CC	SI-LS	MA-CV	I-CC	BWLS
agea	0.003(0.142)	− 0.170(0.041)	0.016(0.029)	− 0.001(0.079)	0.011(0.007)
gndr	0.005(0.033)	− 0.007(0.002)	0.003(0.012)	0.002(0.019)	0.002(0.001)
eduyrs	0.031(1.156)	− 0.042(0.045)	− 0.073(0.248)	0.000(0.093)	− 0.082(0.048)
trstprl	0.127(0.362)	0.093(0.018)	0.063(0.089)	0.107(0.125)	0.092(0.018)
trstlgl	0.058(0.314)	0.013(0.018)	0.024(0.092)	0.081(0.123)	0.038(0.017)
trstplc	0.096(0.396)	0.065(0.020)	0.036(0.064)	0.070(0.143)	0.065(0.020)
trstprt	0.104(0.525)	− 0.003(0.024)	0.015(0.142)	0.065(0.133)	0.028(0.023)
trstun	− 0.007(0.412)	0.027(0.015)	0.017(0.093)	0.025(0.141)	0.034(0.015)
trstplt	0.034(0.389)	0.010(0.015)	0.002(0.086)	0.024(0.123)	0.014(0.016)
trstep	− 0.132(0.567)	− 0.015(0.022)	− 0.008(0.148)	0.015(0.128)	− 0.005(0.024)
stfeco	0.170(0.386)	0.170(0.016)	0.111(0.087)	0.161(0.141)	0.172(0.016)
stfedu	0.132(0.269)	0.089(0.016)	0.074(0.086)	0.112(0.108)	0.119(0.014)
stfgov	0.293(0.366)	0.257(0.019)	0.186(0.099)	0.251(0.124)	0.307(0.018)
psppsgva	0.163(0.819)	0.111(0.043)	0.086(0.222)	0.057(0.106)	0.098(0.040)
psppipla	− 0.060(0.923)	0.128(0.044)	− 0.007(0.284)	0.056(0.107)	0.073(0.039)
frprtpl	0.016(0.755)	0.175(0.039)	0.075(0.229)	0.081(0.118)	0.131(0.040)
gvintcz	0.181(0.812)	0.116(0.047)	0.081(0.151)	0.085(0.095)	0.108(0.049)
poltran	0.082(0.863)	0.024(0.037)	0.028(0.248)	0.066(0.107)	0.057(0.037)
anvcld	0.079(0.613)	0.074(0.051)	0.111(0.205)	0.024(0.132)	0.236(0.067)
alvgptn	0.182(1.277)	0.026(0.075)	0.055(0.427)	0.005(0.109)	0.095(0.103)
acldnmr	− 0.232(1.144)	− 0.105(0.083)	− 0.088(0.346)	− 0.001(0.096)	− 0.147(0.109)
aftjbyc	0.075(0.559)	− 0.023(0.045)	0.055(0.149)	0.007(0.126)	0.062(0.064)
advcyc	0.028(0.713)	− 0.056(0.054)	0.007(0.216)	0.008(0.120)	0.033(0.069)
hinctnta	− 0.004(0.201)	0.01(0.015)	− 0.007(0.044)	0.004(0.115)	− 0.008(0.012)
ifrjob	0.007(0.246)	0.004(0.013)	0.025(0.057)	0.015(0.122)	0.015(0.012)
happy	0.011(0.358)	0.039(0.019)	0.028(0.116)	0.037(0.123)	0.053(0.018)
recskill	0.054(0.503)	0.086(0.022)	0.063(0.145)	0.039(0.115)	0.069(0.021)
ifredu	0.032(0.290)	0.031(0.018)	0.014(0.086)	0.028(0.117)	0.011(0.012)
hincfel	0.053(0.754)	0.079(0.041)	0.064(0.175)	0.031(0.113)	0.042(0.037)
PE	12.398(6.704)	3.212(0.083)	4.150(2.102)	3.862(0.436)	2.783(0.039)

economy is doing well or at least people perceive it to be doing so, the response variable tends to increase (Van Ham et al. 2017), as the coefficients of variable *stfeco* are all positive under several estimation methods. In addition, the response variable also rises when the state apparatus, political institutions, and public officials are perceived to be transparent, impartial, and fair. Although we may expect the sign of most variables in module M_2 to be positive (Beichelt et al. 2014), we observe that CC, SI-LS, and MA-CV contain more estimated coefficients that are negative compared with I-CC and BWLS, which means that the estimators of BWLS are much more interpretable and in line with the economic significance. BWLS is likewise a more efficient estimator since it has the lowest standard deviation among the five methods.

5. Discussion

In this study, we propose a two-stage framework for a large-scale survey in order to improve its design and regression estimation. In Stage I, we create a survey design with a block-wise missing structure by minimizing a nonlinear cost function with constraint on the reliability of estimates of means. Reducing the cost and increasing the data quality can be achieved through the proposed design, as shown in our numerical studies. In Stage II, we put forward a BWLS estimator based on the SQD data obtained in Stage I under a linear regression model. Numerical studies have shown that the proposed design has better performance in terms of survey cost and respondent burden. Furthermore, the proposed estimator of BWLS leads to satisfactory estimation and prediction accuracies. In our ESS data analysis, the findings are consistent with other studies in political science.

Inspired by the two-stage framework proposed in this article, many studies can be performed in the future. First, various forms of function may be investigated to describe the total cost. Other factors such as response rates and information loss, can also be taken into account as criteria for a SQD. Also, there is an important constraint $n_1 > p$ in our framework. When it comes to the choice of n_1, we just give the guidance that n_1 is usually set to be not too large compared to n_j for $j = 2, ..., q$ due to budget and respondent load considerations.

Further research that examines how the proposed method works as $n_1 - p$ is varied is a valuable direction. Second, the number of questions included in the regression model may be increased, and data exchanges between countries or servers become much more convenient with the fast development of data science. Thus, it would be worthwhile for researchers to further explore variable selection and the corresponding distributed-computation methods in similar situations. Third, this article focuses on the situation of a single response and that its values in the data are observed. When several regression analyses are to be conducted with different response variables, it would be helpful to consider this in the survey design stage so as to ensure the collection of the information for all respondents as much as possible. If, unfortunately, some values are missing in the SQD data for the response variable being considered, we may remove those cases with missing observations or consider a suitable imputation method if available. Examination of effects of the relevant factors (e.g., nature of the missing mechanism) in the process and deriving more efficient techniques require significant future work. Fourth, the BWLS is only based on continuous or dichotomous independent variables at present, while in many other cases, multi-categorical variables are of interest as well. For example, when the impact of different levels of

educations and occupations on earnings is concerned, levels of education and occupation are usually treated as multi-categorical variables. The extension of our proposed framework to these new situations is also worthwhile to look into. This article is focused on a simple random sampling for our methodology and we only consider the influence of the design on the sample size requirements of the survey items. Extensions to complex samplings possibly with key regression parameters (in addition to the means of the survey items) to be considered in the survey design are interesting future research projects.

6. Appendix

6.1. Results for Case 2

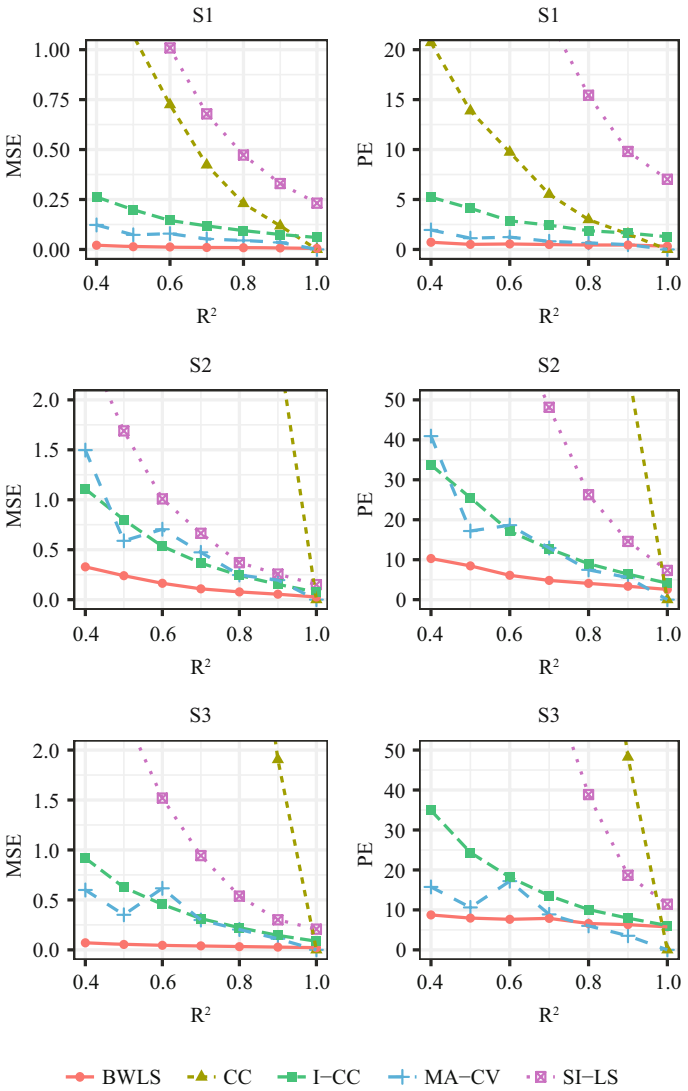


Fig. 8. The comparison of MSE and PE for different estimators in S1–S3 under case 2.

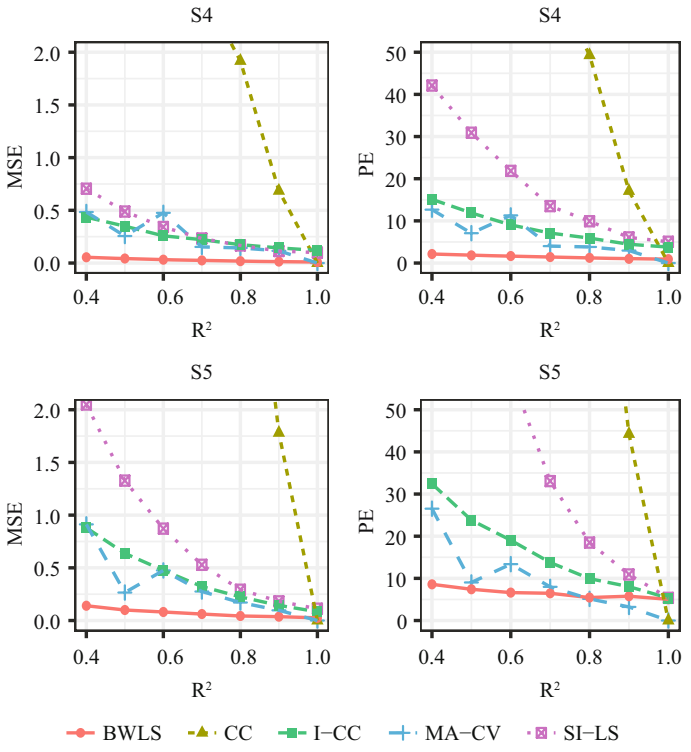


Fig. 9. The comparison of MSE and PE for different estimators in S4−S5 under case 2.

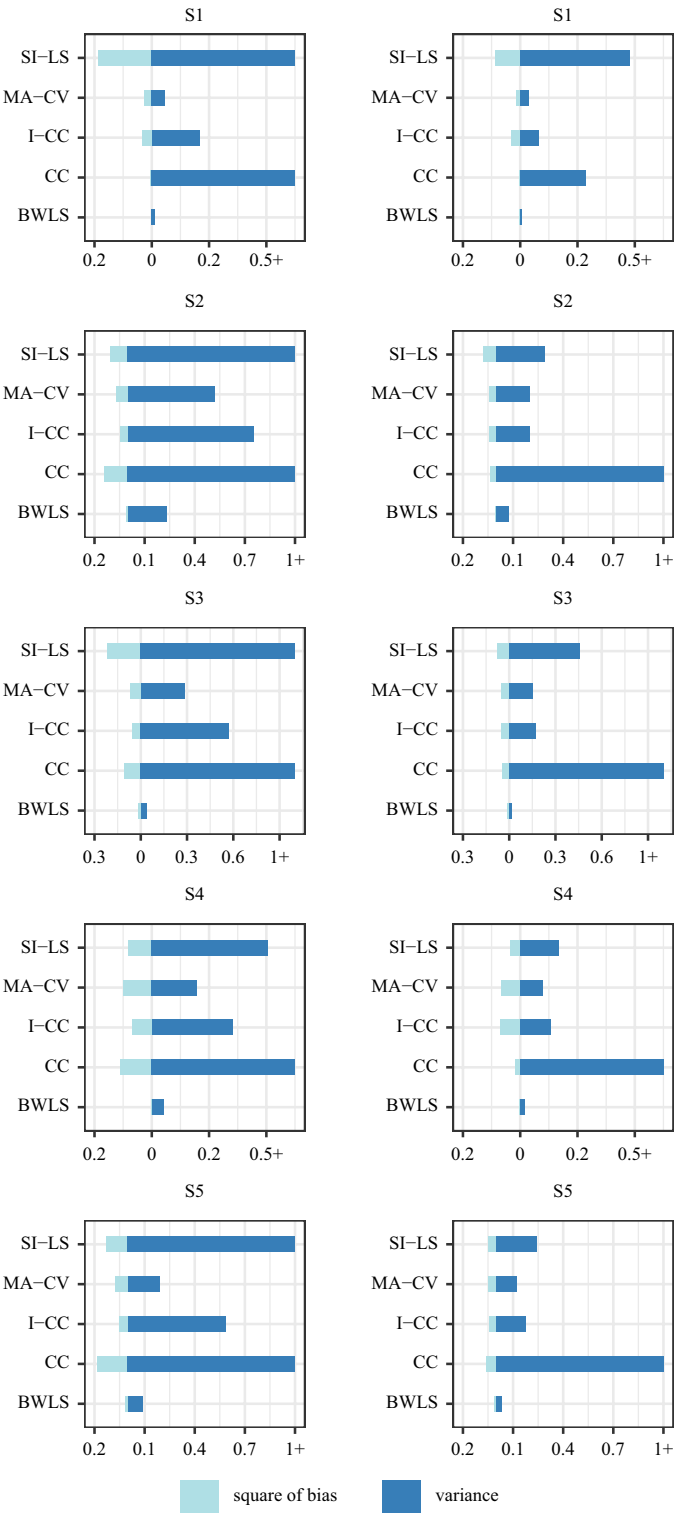
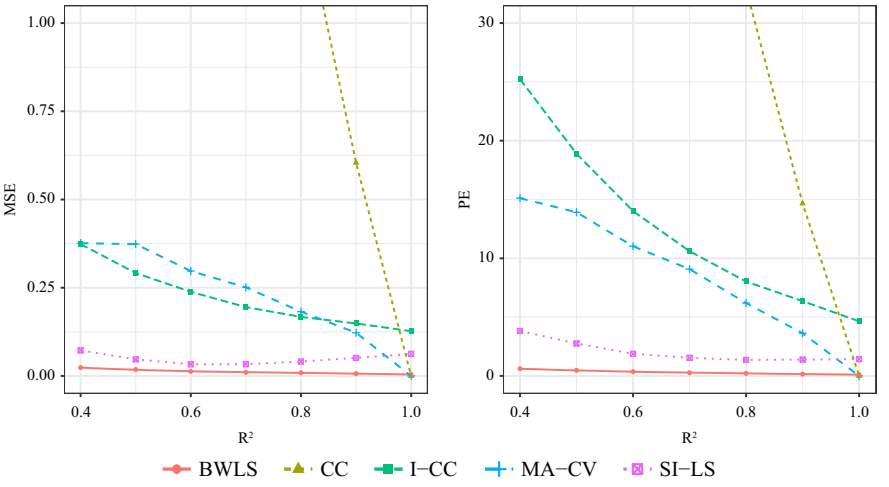
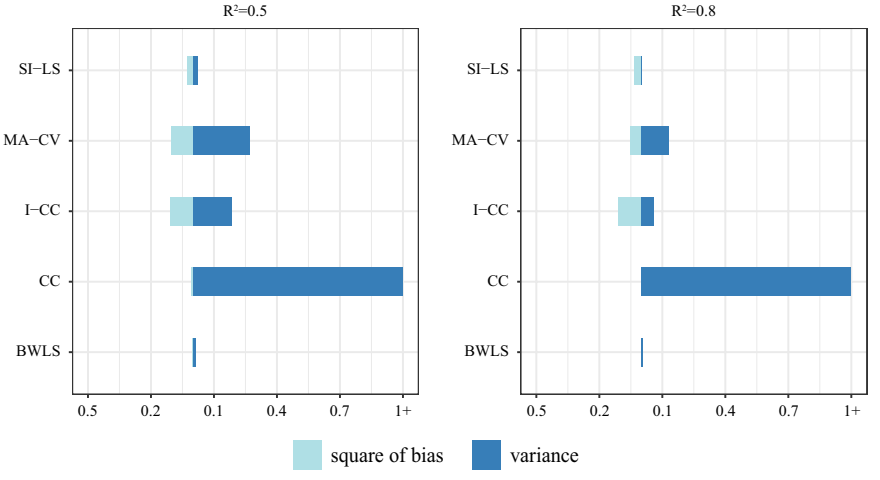


Fig. 10. The decomposition of MSE in case 2.

6.2. Results for Real Data Simulation



(a) The comparison of MSE and PE for different estimators.



(b) The decomposition of MSE.

Fig. 11. Results for real data simulation.

6.3. Results for BWLS and Mean Imputaion

We conducted a simulation study to compare the results of BWLS and mean imputation in terms of MSE and PE for Case 2. The results are presented in Figures 12 and 13

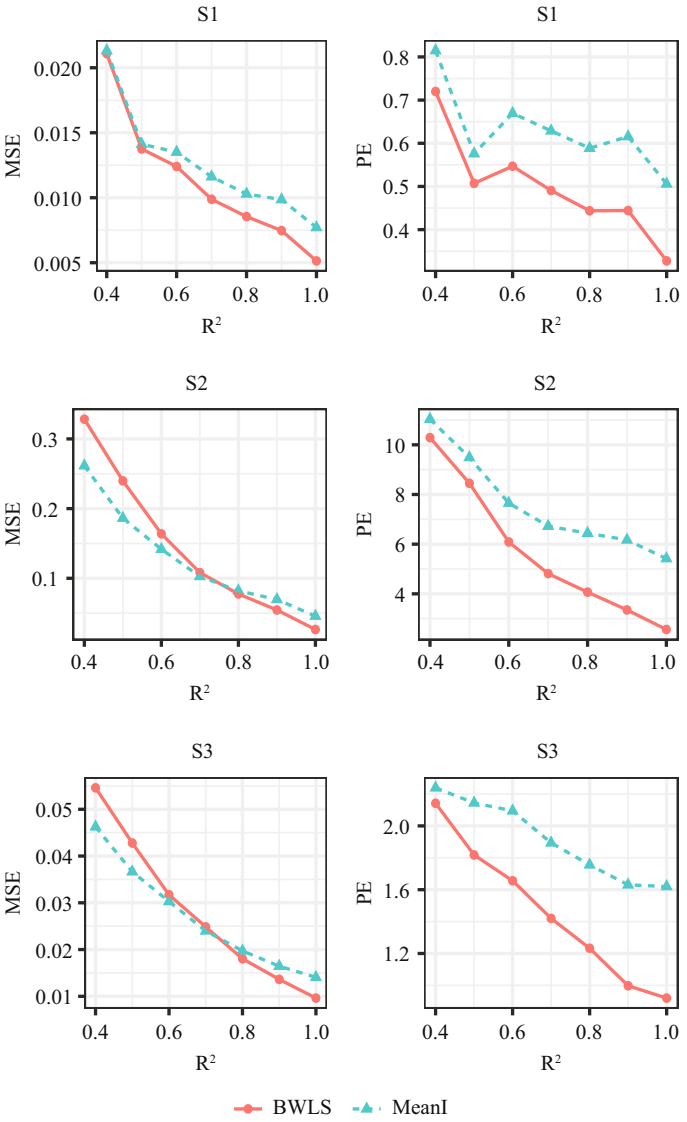


Fig. 12. The comparison of MSE and PE between BWLS and MeanI in S1–S3 under case 2.

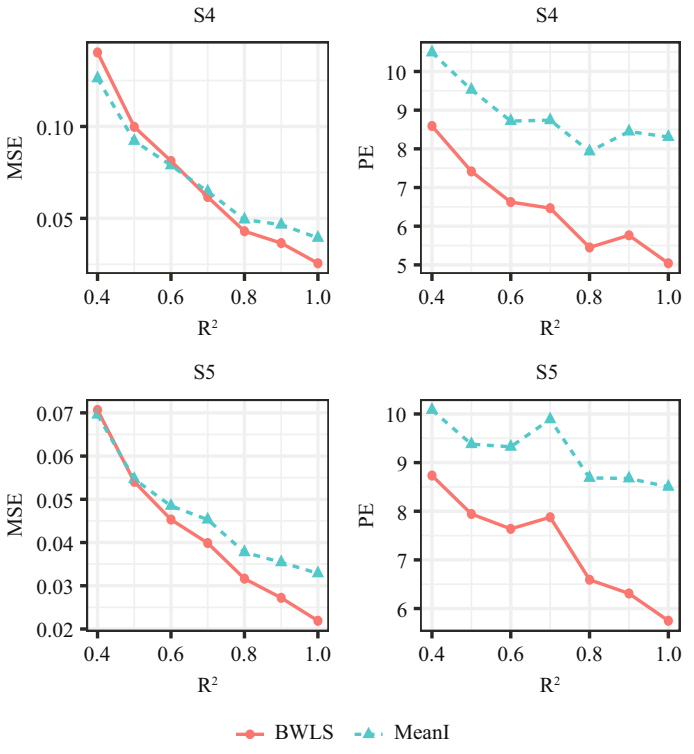


Fig. 13. The comparison of MSE and PE between BWLS and MeanI in S4–S5 under case 1.

6.4. Results for BWLS and Multiple Imputation

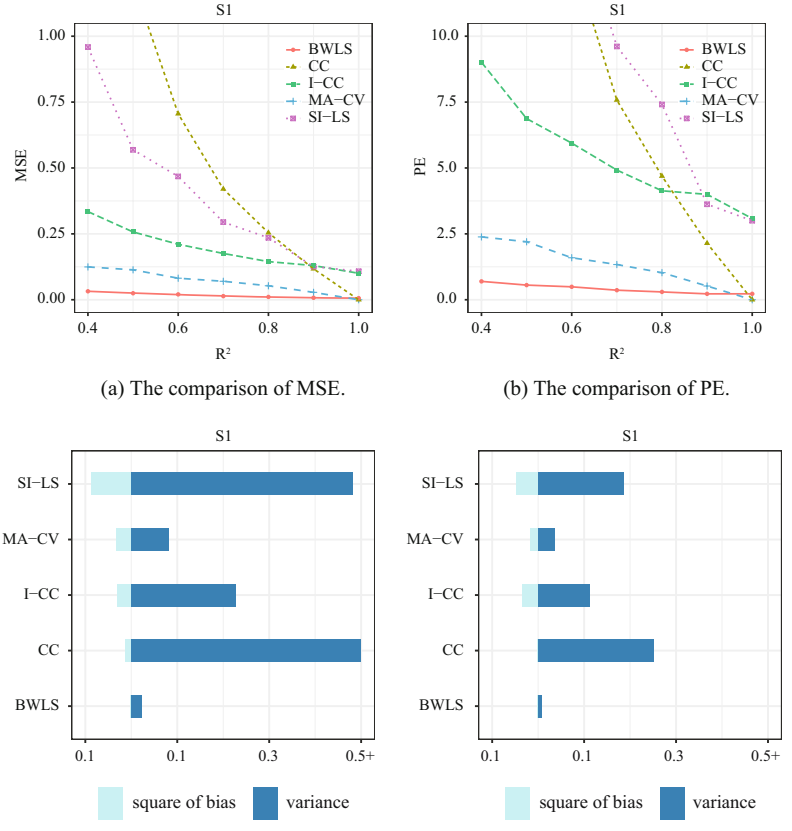
We conducted a simulation study to compare the results of BWLS and multiple imputation (MI). Here, we present the results in Table 5 of Scenarios S1-S4 with correlation structure in Case 1 ($\lambda_1 = 0.3$ and $\lambda_2 = 0.1$) considered in Simulation 1.

Table 5. Simulation results for BWLS and MI.

	R_2	*S*1		*S*2		*S*3		*S*4	
		BWLS	MI	BWLS	MI	BWLS	MI	BWLS	MI
MSE	0.4	0.0263	0.0490	0.1445	0.3860	0.0401	0.0853	0.0703	0.1132
	0.5	0.0187	0.0346	0.1014	0.2765	0.0300	0.0625	0.0534	0.0848
	0.6	0.0150	0.0274	0.0748	0.2088	0.0228	0.0494	0.0427	0.0626
	0.7	0.0112	0.0187	0.0568	0.1399	0.0172	0.0357	0.0365	0.0477
	0.8	0.0091	0.0139	0.0434	0.0952	0.0144	0.0274	0.0307	0.0337
	0.9	0.0074	0.0086	0.0327	0.0543	0.0111	0.0181	0.0261	0.0216
PE	0.4	0.5529	0.8364	5.8703	12.3396	1.7240	2.7974	4.7488	3.8173
	0.5	0.4219	0.5868	4.5167	8.7446	1.3387	2.0661	4.1224	2.7836
	0.6	0.3679	0.4685	3.7295	6.5603	1.1338	1.6330	3.7433	2.0942
	0.7	0.2926	0.3161	3.1695	4.3525	0.9484	1.1789	3.5797	1.5803
	0.8	0.2605	0.2357	2.7958	3.0022	0.8511	0.9038	3.4674	1.1129
	0.9	0.2291	0.1466	2.3951	1.6991	0.7417	0.6005	3.2825	0.7123

6.5. Results Under MAR

Based on the settings in Scenario S1 with Case 1 in Simulation 1, we generate one more variable, namely, $p = 21$. Let X_{i1} be observed for all respondents, and the remaining 20 variables are evenly divided into five modules. For generating MAR data, let $\delta_{ij} = 1$ if the i-th respondent answers the j-th sub-questionnaire and otherwise $\delta_{ij} = 0, i = 1, 2, ..., n$; $j = 1, 2, ..., q$. Here, we set $\delta_{ij} = I(\tau_{j-1} < X_{i1} <_{\tau_j}), i = 1, 2, ..., n$, where τ_j is (j/q)-th quantile of X_{i1}, which means that the missing probability depends on the value of Xi_1. The results are shown in Figure 14. It is observed that the performance of BWLS is superior to the competing methods in terms of MSE and PE, which is consistent with the results under missing completely at random. The conclusion is similar under other settings in the article and we omit the details here due to space limitation.



(a) The comparison of MSE.

(b) The comparison of PE.

(c) The decomposition of MSE under $R^2 = 0.5$.

(d) The decomposition of MSE under $R^2 = 0.8$.

Fig. 14. Simulation results for S1 with case 1 under MAR.

7. References

Adigüzel, F., and M. Wedel. 2008. "Split questionnaire design for massive surveys." *Journal of Marketing Research* 45(5): 608–617. DOI: https://doi.org/10.1509/jmkr.45.5.608.

Andreadis, I., and E. Kartsounidou. 2020. "The impact of splitting a long onlin equestionnaire on data quality." *Survey Research Methods* 14(1): 31–42. DOI: https://doi.org/10. 18148/srm/2020.v14i1.7294.

Beichelt, T., I. Hahn, F. Schimmelfennig, and S. Worschech. 2014. *Civil society and democracy promotion*. Springer. DOI: https://doi.org/10.1057/9781137291097.

Best, H. and, C. Wolf. 2013. *The SAGE handbook of regression analysis and causal inference*. Sage. DOI: https://doi.org/10.4135/9781446288146.

Cai, J., E.J. Candès, and Z. Shen. 2010. "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization* 20(4): 1956–1982. DOI: https://doi.org/10.1137/080738970.

Chipperfield, J.O., M.L. Barr, and D.G. Steel. 2018. "Split questionnaire designs: collecting only the data that you need through MCAR and MAR designs." *Journal of Applied Statistics* 45(8): 1465–1475. DOI: https://doi.org/10.1080/02664763.2017.1375085.

Chipperfield, J.O., and D.G. Steel. 2009. Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics* 25(2): 227–244. Available at: https://ro.uow.edu.au/infopapers/3334/.

Chipperfield, J.O., and D.G. Steel. 2011. "Efficiency of split questionnaire surveys." *Journal of Statistical Planning and Inference* 141(5): 1925–1932. DOI: https://doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., and D.G. Steel. 2012. "Multivariate random effect models with complete and incomplete data." *Journal of Multivariate Analysis* 109: 146–155. DOI: https://doi.org/10.1016/j.jmva.2012.02.014.

Davidov, E., J. Cieciuch, and P. Schmidt. 2018. "The cross-country measurement comparability in the immigration module of the European Social Survey 2014-15." 12(1): 15–27. DOI: https://doi.org/10.18148/srm/2018.v12i1.7212.

Dziura, J.D., L.A. Post, Q. Zhao, Z. Fu, and P. Peduzzi. 2013. "Strategies for dealing with missing data in clinical trials: from design to analysis." *The Yale journal of biology and medicine* 86(3): 343. Available at: https://pubmed.ncbi.nlm.nih.gov/24058309/.

Early, K. 2016. *Dynamic question ordering: obtaining useful information while reducing user burden proposal*. Ph. D. thesis, Carnegie Mellon University Pittsburgh, PA. DOI: https://doi.org/10.1184/R1/6716123.v1.

Fang, F., W. Lan, J. Tong, and J. Shao. 2019. "Model averaging for prediction with fragmentary data." *Journal of Business & Economic Statistics* 37(3): 517–527. DOI: https://doi.org/10.1080/07350015.2017.1383263.

Hoerl, A.E. and R.W. Kennard (1970). "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics* 12(1): 55–67. DOI: *Technometrics* 12(1): 55–67. DOI: https://doi.org/10.1080/00401706.1970.10488634.

Ioannidis, E., T. Merkouris, L.-C. Zhang, M. Karlberg, M. Petrakos, F. Reis, and P. Stavropoulos. 2016. "On a Mmodular Approach to the Design of Integrated Social

Surveys." *Journal of Official Statistics* 32(2): 259–286. DOI: https://doi.org/10.1515/-jos-2016-0013.

Ledoit, O., and M. Wolf. 2004. "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis* 88(2): 365–411. DOI: https://doi.org/10.1016/S0047-259X(03)00096-4.

Lesperance, M.L., and J.D. Kalbfleisch. 1992. "An algorithm for computing the nonparametric MLE of a mixing distribution." *Journal of the American Statistical Association* 87(417): 120–126. DOI: https://doi.org/10.1080/01621459.1992.10475182.

Little, R.J. 1992. "Regression with missing X's: a review." *Journal of the American statistical association* 87(420): 1227–1237. DOI: https://doi.org/10.1080/01621459.1992.10476282.

Little, R.J., and M.D. Schluchter. 1985. "Maximum likelihood estimation for mixed continuous and categorical data with missing values." *Biometrika* 72(3): 497–512. DOI: https://doi.org/10.1093/biomet/72.3.497.

Liu, M., and L. Wronski. 2018. "Examining completion rates in web surveys via over 25,000 real-world surveys." *Social Science Computer Review* 36(1): 116–124. DOI: https://doi.org/10.1177/0894439317695581.

Mazumder, R., T. Hastie, and R. 2010. "Spectral regularization algorithms for learning large incomplete matrices." *Journal of Machine Learning Research* 11: 2287–2322. DOI: https://dl.acm.org/doi/10.5555/1756006.1859931.

Merkouris, T. 2015. "An efficient estimation method for matrix survey sampling." *Survey Methodology* 41(1): 237–262. DOI: https://www150.statcan.gc.ca/n1/en/catalogue/ 12-001-X201500114174.

Neidorf, T., and M. Sheehan. 2014. "National Assessment of Educational Progress (NAEP)." In *Encyclopedia of Science Education*., edited by R. Gunstone. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-007-6165-0_67-2.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of measurement error due to survey length: evaluation of the split questionnaire design approach." *Survey Research Methods* 11(4): 361–368. DOI: https://doi.org/10.18148/srm/2017.v11i4.7145.

Raghunathan, T.E., and J.E. Grizzle. 1995. "A split questionnaire survey design." *Journal of the American Statistical Association* 90(429): 54–63. DOI: https://doi.org/10.1080/01621459.1995.10476488.

Rhemtulla, M., and T.D. Little. 2012. "Planned missing data designs for research in cognitive development." *Journal of Cognition and Development* 13(4): 425–438. DOI: https://doi.org/10.1080/15248372.2012.717340.

Rust, K.F., and E.G. Johnson. 1992. "Chapter 2: Sampling and weighting in the national assessment." *Journal of Educational Statistics* 17(2): 111–129. DOI: https://doi.org/10.3102/10769986017002111.

Schnaudt, C., M. Weinhardt, R. Fitzgerald, and S. Liebig. 2014. "The European Social Survey: contents, design, and research potential." *Journal of Contextual Economics: Schmollers Jahrbuch* 134(4): 487–506. DOI: https://doi.org/10.3790/schm.134.4.487.

Skinner, C.J., and O. Coker. 1996. "Regression analysis for complex survey data with missing values of a covariate." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159(2): 265–274. DOI: https://doi.org/10.2307/2983173.

Van Ham, C., J.J. Thomassen, K. Aarts, and R.B. Andeweg. 2017. *Myth and reality of the legitimacy crisis: explaining trends and cross-national differences in established democracies*. Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198793717.001.0001.

Vonneilich, N., D. Lüdecke, and O. von dem Knesebeck. 2019. "Educational inequalities in self-rated health and social relationships–analyses based on the European Social Survey 2002–2016." *Social Science and Medicine*. DOI: https://doi. org/10.1016/j.-socscimed.2019.112379.

Wu, H., and S.-O. Leung. 2017. "Can likert scales be treated as interval scales? – a simulation study." *Journal of Social Service Research* 43(4): 527–532. DOI: https://doi.org/10.1080/01488376.2017.1329775.

Xiang, S., L. Yuan, W. Fan, Y. Wang, P.M. Thompson, and J. Ye. 2014. "Bi-level multi-source learning for heterogeneous block-wise missing data." *NeuroImage* 102: 192–206. DOI: https://doi.org/10.1016/j.neuroimage.2013.08.015.

Yang, S., and J.K. Kim 2017. "A semiparametric inference to regression analysis with missing covariates in survey data." *Statistica Sinica* 27: 261–285. DOI: https://doi.org/10.5705/ss.2014.174.

Yu, G., Q. Li, D. Shen, and Y. Liu. 2020. "Optimal sparse linear prediction for blockmissing multi-modality data without imputation." *Journal of the American Statistical Association* 115(531): 1406–1419. DOI: https://doi.org/10.1080/01621459.2019.1632079.

Yuan, K.-H., and P.M. Bentler. 2010. "Consistency of normal-distribution-based pseudo maximum likelihood estimates when data are missing at random." *American Statistician* 64(3): 263–267. DOI: https://doi.org/10.1198/tast.2010.09203.

Received September 2021
Revised August 2022
Accepted March 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 489–505, http://dx.doi.org/10.2478/JOS-2023-0023

Answering Current Challenges of and Changes in Producing Official Time Use Statistics Using the Data Collection Platform MOTUS

Joeri Minnen[1], *Sven Rymenants*[1], *Ignace Glorieux*[2], *and Theun Pieter van Tienoven*[2]

The modernization of the production of official statistics faces challenges related to technological developments, budget cuts, and growing privacy concerns. At the same time, there is a need for shareable and scalable platforms to support comparable data, leading to several online data collection strategies being rolled out. Time Use Surveys (TUS) are particularly affected by these challenges and needs as they (while producing rich data) are complex, time-intensive studies (because they include multiple tasks and are administered at the household level). This article introduces the Modular Online Time Use Survey (MOTUS) data collection platform and explains how it accommodates the challenges of and changes in the production of a TUS that is carried out in line with the Harmonized European Time Use Survey guidelines. It argues that MOTUS supports a shift in the methodological paradigm of conducting TUS by being timelier and more cost efficient, by lowering respondent burden, and by improving the reliability of the data collected. Importantly, the modular structure allows MOTUS to be easily deployed for various TUS configurations. Moreover, this versatile structure allows comparable, complex diary surveys (such as the household budget survey) to be performed on the same platform and with the same applications.

Key words: Time-use survey; data collection platform; cost efficiency; data quality, respondent burden.

1. Introduction

Today, National Statistical Institutes (NSIs) face challenges and changes in the way they produce official statistics (Radermacher 2020). On the one hand, technological developments create the opportunity for paradigm shifts in methodology (Ashofteh and Bravo 2021). On the other hand, modern societal changes and challenges create new user demands for high-quality data and statistics (Cai and Zhu 2015). Taken together with the budgetary restrictions in place, this results in a large pressure to shift to online data collection and to connect data collection environments with other data sources that bring valuable information to specific statistical domains (Ricciato et al. 2020). This digital transformation rapidly changes the context and needs, and it also leads to growing privacy and data security concerns and suspicion towards official statistics (Keusch et al. 2019; Ricciato et al. 2020). Amidst these challenges and changes, modernisation initiatives should

[1]hbits, Witte Patersstraat 4, 1040 Etterbeek, Brussels, Belgium. Emails: joeri.minnen@hbits.io and sven.rymenants@hbits.io
[2]Research Group TOR, Sociology Department, Vrije Universiteit, Pleinlaan 2, 1050 Brussels, Belgium. Emails: ignace.glorieux@vub.be and t.p.van.tienoven@vub.be
Acknowledgments: This research is part of the project Software Outreach and Redefinition to Collect E-data Through MOTUS (SOURCE TM; Grant nr. 847218-BE-2018-TUS) funded by EUROSTAT and executed by Statistics Belgium (Statbel) in collaboration with the German Federal Statistical Office (Destatis) and hbits.

be supported by trustable, shareable, and scalable processes considering "smart" ways to collect data (Bruno et al. 2022; Ricciato et al. 2019). These processes are assumed to lead to cost reductions for the statistical offices and to lower the respondent burden (Salemik et al. 2020). In addition, these processes must remain standardized for reasons of comparability, yet flexible and agile enough to meet (country) specific needs and allow statistics to be disseminated quickly. At the same time, these processes must not compromise on the quality and reliability of the collected data (Salgado et al. 2018; Stodden 2014).

At the European level, the European Statistical System (ESS), which is a partnership between Eurostat and the NSIs of the EU and EFTA countries, aims at enhancing the strengths (such as comparability) of harmonised statistical methods and reversing the trend of a gradual disintegration of the data collection process stemming from NSIs facing declining participation rates and increasing difficulties in organising data collections (and thereby jeopardizing the quality and reliability of the statistics). At the same time, the ESS foresees to jump on the bandwagon of the process of digitalisation, growing smartphone usage (Keusch et al. 2019) and the availability of 4G and 5G networks (Gohar and Nencioni 2021). New technologies should improve respondent responsiveness by using new tools, integrating new data flows by connecting data sources, and help NSIs become more efficient by defining data collection platforms. The goal is to better capture and disseminate the perspective of households (Carletto et al. 2022).

The Time Use Survey (TUS) is one of the European surveys that are substantially affected by the challenges of and changes in the way NSIs produce statistics, but at the same time would substantially benefit from new technological developments. Against that backdrop, this contribution aims to answer whether the Modular Online Time Use Survey (MOTUS) data collection platform is able to tackle these challenges and align with these changes. Official TUSs face numerous challenges, such as the need to replace the expensive and laborious paper-and-pencil method by a digitalized method with smart ways to reduce respondent burden amid the absence of updated guidelines to harmonize digitalized TUS across NSIs. Many of these challenges relate to the principles of the European Statistics Code of Practice (Eurostat 2018). The central question this contribution addresses is: can MOTUS improve on respondent burden (principle 9), cost efficiency (principle 10) and quality such as accuracy and reliability (principle 12), and timeliness and punctuality (principle 13) in producing official TUS statistics?

In answering this question, we consider respondent burden as a *perceived* burden, which results from low motivation, the complexity of the tasks at hand, and the challenging effort to complete the survey (Yan et al. 2019). Furthermore, we consider the timeliness, the accuracy and reliability of the (intermediate statistics production steps as well as the final) time use statistics as quality indicators. We assume that the accuracy and reliability of statistics can be gained by reducing human data entry errors, by reducing the respondent recall error, and by supporting respondents with real-time prompts during the data collection process. In what follows, we evaluate MOTUS in terms of expected improvements in costs, respondent burden, and quality compared to the current best practice of paper-and-pencil TUS for different phases of the GSBPM, Generic Statistical Business Process Model (Kuonen and Loison 2019).

2. Background

2.1. Time Use Surveys

A TUS collects data on daily life. They are a way to picture the "many interesting patterns of social life [that] are associated with the temporal distribution of human activities, with the regularities in their timing, duration, frequency, and sequential order" (Szalai 1972, 1). Respondents use a log or a time use diary of at least twenty-four consecutive hours to self-report their daily behaviour in a chronological and open-ended fashion on an activity-to-activity basis (Pronovost 1989; Robinson 1999). In the time use diary, respondents specify – for each new activity – the start and end time as well as some contextual information like the place of occurrence and the possible presence of others. This not only makes time use diaries capable of simultaneously collecting data on the duration, timing, tempo, and sequence of activities (Zerubavel 1982) but it also reduces respondent errors related to self-reporting of activities in daily life compared to other survey methods (Lavrakas 2008). Respondent errors related to understanding the concept (of the question asked) are reduced because respondents are not directly queried but use their own wording to describe their activities. However, insufficient detail in verbatim activity descriptions complicates posterior activity coding (Chenu 2004). Recall biases are reduced because respondents are asked to register their activities in close to real time, resulting in multiple registration moments per day. Other biases such as social desirability biases or confirmation biases are reduced because time diaries do not focus on a particular activity, activities chronologically follow each other (i.e., the ending time of one activity is the start time of the next activity), and activity durations are restricted to 24 hours a day (Te Braak et al. 2022b).

As TUS is a source for official statistics on which policymakers rely, and as it can further enhance the understanding of daily life, initiatives have been taken around the world to harmonize the production of time use data (Robinson and Godbey 1997). One of the most extensive harmonization processes was carried out by Eurostat and resulted in the guidelines on Harmonised European Time Use Surveys (HETUS) for these surveys conducted by NSIs (Eurostat (2020), referred to as "the guidelines" below). The guidelines (which include sample design harmonization and standardization, mode and methodology design, activity coding, data coding, weights, and metadata) have been used by nearly 20 European NSIs in two HETUS rounds between 1998 and 2015.

The TUSs are not merely a European matter. Since 2003, BLS, the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics 2023) collects yearly waves of the American Time Use Survey (ATUS) to support policy research related to household production, health and safety, and family and work-life balance. Similarly, and often with support of the International Labor Organization (ILO), numerous countries outside Europe use time use statistics to gain valuable insights on household production and gender (in)equality (United Nations 2016).

The major strength of TUSs is capturing detailed information of daily activities in a chronological and contextualised way. Yet this strength is also its weakness, both at the organisational "back office", as well as at the participation environment or "front office". From an organisational point of view, these surveys are costly, mainly due to postage, printing, and personnel costs resulting from multiple interviewer visits to the household

and data entry from paper time use diaries. Regarding the latter, the large number of offline manual operations increase the risk of errors. Additionally, fieldwork periods typically run for 12 months to capture seasonal variations. From the respondent point of view, the burden to complete such a survey is relatively high, because household members complete multiple questionnaires and keep track of their daily time use in paper time use diaries.

2.2. A HETUS Based TUS

To address the central question whether MOTUS can improve respondent burden, cost efficiency, accuracy and reliability, and timeliness and punctuality in producing official TUS statistics, we consider the guidelines to be the benchmark. As the HETUS is a household survey, sampling is carried out at the household level. The identified head of each participating household will complete a grid that records the relationships between all persons in the household (i.e., the household grid, see Eurostat 2020, 33) and a household questionnaire. Additionally, all eligible household members (i.e., aged ten and above) will complete an individual questionnaire. Currently, this is (most frequently) done via Computer Assistant Personal Interviews (CAPI), which implies an interviewer visit – at which the interviewer also leaves behind two paper time use diaries per eligible household member with the dates on which both time use diaries must be completed. One diary concerns a weekday, and one diary concerns a weekend day (the same two days for all household members). The interviewer might also leave behind a drop-off questionnaire, which is to be completed by all eligible household members after the time use diaries. At a prearranged date, the interviewer returns to check and collect the time use diaries and the drop-off questionnaire. At the NSIs, the paper time use diaries and drop-off questionnaires are entered into a database, often using parallel data entry to prevent input and coding errors.

3. Modular Online Time Use Survey

3.1. Introducing MOTUS

To counter the high costs of conducting TUSs and to lower the respondent burden, while maintaining reliable and quality output on daily life, scholars and NSIs started to experiment with conducting these surveys through web- and mobile applications (Bonke and Fallesen 2010; Fernee and Sonck 2013; Sonck and Fernee 2013; Sullivan et al. 2020), with the first applications coming into circulation around 2010. The first version of MOTUS was rolled out in 2012.

Figure 1 shows the platform architecture of MOTUS. The MOTUS data collection platform consists of a front office as well as a back office. The front office relates to the collection tool or application, with which the users can interact via a user interface (UI) and which delivers, through its functionalities, a user experience (UX). The MOTUS application is available as a web version for browsers (https://app.motusresearch.io) and in iOS and Android mobile versions for smartphones and tablets. The purpose of the application is to make it easier for the respondent to carry out all tasks of a (time use or other) survey.

The back office serves to build a survey, to facilitate data collection and monitoring, and to process the data. To this end, the back office, which is accessible via a web environment,

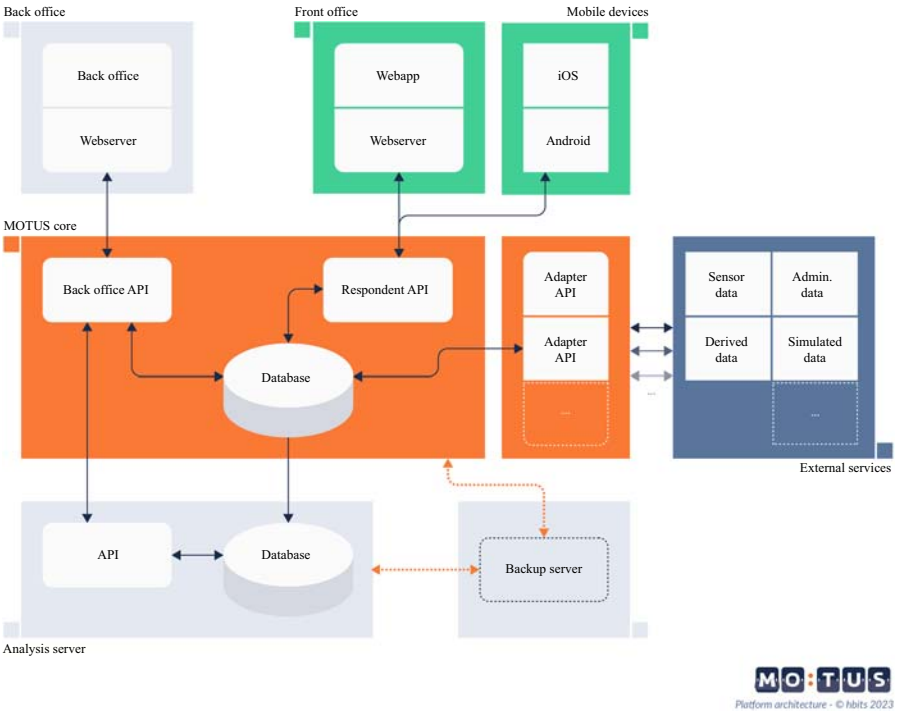


Fig. 1. Overview of the MOTUS platform architecture.

contains several *builders*. Both the front office and back office connect to the MOTUS core ("the core") through Application Programming Interfaces (APIs). The core holds the database with all information required to build a survey and collect data. A separate analysis server holds a replica of the database from the core and facilitates the processing of information in the back office. The back-up server is a replica of the core and analysis servers. Adapter APIs serve to adapt external information so that it can be processed in the core, thereby allowing the ingestion of, for example, passive data gathered via integrated sensors or connected devices, administrative/secondary data available via external data sources, or other processed data. For reasons of optimization, data security and privacy, these data are handled and organised in an anonymized way in stand-alone microservices. All input provided by the user is sent encrypted via an https communication to the server and is immediately propagated to all devices of the user via the respondent API. As a result, the MOTUS web and mobile applications can be used interchangeably.

3.2. Building TUS With MOTUS

To enhance the comparability of official TUSs in Europe, the design hereof in MOTUS is largely informed by the guidelines, which are regularly updated (Eurostat 2020). In the current situation, these guidelines provide a good starting point to include online applications and data collection platforms, while considering an online first approach which still ensures comparability with paper diaries (Vassilev et al. 2020). At the same

time, new applications and platforms and the options to implement smart solutions will produce possibilities that most likely impact the TUS design.

MOTUS supports building a HETUS guided TUS (see Subsection 2.2) and currently features nine builders – eight of which are relevant for a TUS, while the ninth builder offers future possibilities (see Subsection 4.1). All builders contribute in varying degrees to the lowering of the respondent burden, the cost reduction, the improvement of the accuracy and reliability of the data, and the increased timeliness and punctuality. Table 1 provides an overview of the builders in relation to the GSBPM build phase and the improvements that they bring in relation to current TUS practices.

3.2.1. Collection instruments

The *survey builder* serves to create online questionnaires based on all common question types with all common functionalities (e.g., answer-based routing, piping). This builder allows sharing previous questionnaires over studies. For a TUS the survey builder would be used to construct a household questionnaire, two individual questionnaires (i.e., one before and one after the time use diaries) and context questionnaires. Context questionnaires are linked to activities that are registered in the time use diary and can gauge where the activity took place (or what mode of transport used in case of travel), with whom the activity was undertaken, and if any information or communication technology was used during the activity. Obviously, online questionnaires are timelier and more punctual as well as more cost efficient because they (can) eliminate interview and data entry processes (as data are already digitized, which also eliminates human data entry errors). They also contribute to accuracy and reliability because conditions (e.g., mandatory questions) and restrictions (e.g., an answer cannot exceed a certain value) can be defined.

The *diary builder* sets up the time use diary. At the core of the time use diary is the Online Activity Classification List (OACL) that respondents use to register their daily life. The OACL is derived from the Activity Classification List (ACL) as described in the guidelines. In MOTUS, an OACL is created as a tree structure with up to three levels and as many activities or activity categories in any given level as needed. In MOTUS, a (different) context questionnaire (as created in the survey builder) can be attached to each specified activity. The diary builder contains a repository with previous OACLs for reuse.

The use of OACLs presents a major improvement. Firstly, it is cost-efficient because there is no need to assign actual activity codes to written, verbatim activities. MOTUS can present the OACL to the respondents as a collapsible tree structure, and/or as a searchable list, and/or as a list of favourites. The searchable list is very similar to the traditional verbal description, with the difference that respondents are shown the activities that match their description and thus code their description themselves. Since respondents do this straight away, this also improves the accuracy and reliability as well as timeliness and punctuality. For the searchable list to work, an unlimited number of search tags can be assigned to each of the activities at the most granular level of the tree structure in the diary builder. For the favourite list to work, respondents need to star activities. The different options of selecting activities in the time use diary are also likely to lower the respondent burden, accuracy and reliability as relevant response alternatives are suggested. To handle the situation when an activity cannot be found, OACLs might contain the option to describe activities in the respondents' own words. The search terms used, and the finally selected activity are stored in the background to

Table 1. Overview of the MOTUS builders in relation to the GSBPM build phase and improvements (by CoP principle) to TUS practices

Subphase	Supported phase	Builder	TUS elements	Improvement in CoP principles			
				Respondent burden	Cost efficiency	Timeliness & Punctuality	Accuracy & Reliability
3.1 Reuse or build collection instruments	4. Collect	Survey builder	– Household questionnaire – Individual questionnaires – Activity context questionnaires	–	•	•	○
		Diary builder	– Online Activity Classification List – Activity selection settings – Diary settings	•	•	•	•
		Grid builder	– Household members relationships – Eligibility criteria for participation	–	•	○	•
3.2 Reuse or build processing and analysis components	5. Process 6. Analyse	R builder	– Activation report (for reminders) – Finalisation report (for remuneration) – Quality and cleaning criteria – Dashboard progress report	–	•	•	•
3.3 Reuse or build dissemination components	7. Disseminate	R builder	– Data export	–	•	•	•
3.4 Configure workflow	4. Collect	Communication builder	– Communication content – Communication type	○	○	○	○
		Translation builder	– Set up multiple languages	○	–	–	○
		Invitation builder	– Manage respondent inflow	–	○	○	○
		Research builder	– Create workflow	○	•	•	•

progressively improve the efficiency of the search algorithm during the course of the survey. Secondly, paper-and-pencil questionnaires are limited regarding the context questions and these questions cannot vary per activity in the ACL. In contrast, OACLs can lower the respondent burden as (for instance) not all context questions need to be asked.

Next to the activity list, the diary builder also allows the survey manager to set a large array of time use diary parameters. These include the granularity of the time intervals (e.g., continuously or in whole minute intervals), the diary period and diary period calculation, the start and assignment of focus periods (i.e., the day or days for which the time use diary needs to be completed), and the (length of the) learning period. For a HETUS based TUS, the granularity would be set at ten-minute intervals, while the the focus days are a function of an algorithm that ensures an equal dispersion of starting days across the week and assigns one weekday and one weekend day to all eligible individuals of the household. Controlling the time use diary parameters brings a substantial improvement to accuracy and reliability. The major disadvantage of drop-off paper time use diaries is the lack of control over and insight in what happens between dropping off the diaries and collecting them (Te Braak et al. 2022a). The diary builder allows the survey manager to set, monitor, and adjust the time use diary during the fieldwork.

The *grid builder* is used when the unit of participation is not the individual but a group or, in this case, a household. In a TUS, the reference person of the household composes a household grid by adding household members, providing relevant information (e.g., at least date of birth), specifying relationships (e.g., mother-daughter, partners, siblings ...), and answering questions about household members less than ten years old (e.g., about day care arrangements). Based on this information, household members are checked for their eligibility (according to the criteria set out in the grid builder) to take part in the survey. If the reference person provides group members' email addresses, all group members who are eligible to participate will receive an invitation via email with their initially assigned personal credentials. An online household grid has the same cost and time benefits as online survey questionnaires.

In a HETUS based TUS, participation needs to be coordinated, because of synchronous time use diary registration by all household members. In MOTUS, this is achieved by all group members enter a virtual waiting room. Once all eligible members have entered the waiting room, a subsequent, synchronized task can be assigned. In other words, only when all eligible household members completed their previous task(s), they can proceed to the time use diary task. Optionally, the reference person can manually request the next task if waiting for other group members is deemed to be futile. The cost reductions are obvious because of the elimination of the interviewer and the fully automated process of completing the household grid, checking of eligibility, and distributing individual questionnaires and time diaries. This also improves accuracy and reliability as well as timeliness and punctuality. However, as the household grid still needs to be completed by the head of the household, the respondent burden is not decreased. Nevertheless, accuracy and reliability will improve if a waiting room is used because it allows the household members' time diaries to be truly synchronized; something which cannot be guaranteed (or even assessed) when the traditional method (dropping off paper-and-pencil time use diaries for pick-up at a later moment in time) is used.

3.2.2. Processing, analysis, and dissemination components

It is necessary to set up several processes (in addition to the collection instruments) that support the collection, the analysis, and the dissemination of the statistics. Many of the processing components are part of the MOTUS architecture (see Subsection 3.1), but some processes are built in the *R builder*.

Firstly, the R builder contains the `motusr` package which allows the creation of closing criteria settings or quality assessment of the time use diary. These thresholds or quality criteria relate to the amount of undefined time, the variance and number of different of activities logged, the prevalence of activities which start or end at the top of the hour, and the registration of certain activities, such as sleeping, eating, drinking, and travelling in case of changing locality (Juster 1986). Feedback on data quality can be presented to the respondent purely informatively via onscreen messages or lead to an explicit request to the respondent to adjust the registration in the diary as a requirement to proceed or end the time diary stage. The `motusr` package is currently under development and not yet listed on CRAN.

Secondly, the R builder periodically performs calculations on live data on the MOTUS server to check for changes and to update the outputs. These calculations feed into a dashboard that allows progress monitoring. Finally, the R builder facilitates the construction, labelling, and exporting of (including para- and metadata and Universally Unique Identifier (UUID) keys to merge different databases) in various formats.

In addition to making the fieldwork timelier and more cost efficient, the various automated processes outlined above also improve the accuracy and reliability of the data.

3.2.3. Configure workflow

The collection instruments and processes need to be brought together to form a workflow and are linked through communication. All communication is defined in the *communication builder* and, in the absence of an interviewer and except for initial postal invitations when an email address is not yet available, there are four ways of communicating throughout the data collection process: email, push messages, and static pages. Push messages include real-time prompts that remind respondents of their survey tasks and support respondents registration process by, for example, suggesting relevant response alternatives. This improves the accuracy and reliability. Additionally, if studies need to be conducted in multiple languages, all elements (i.e., collection instruments and communication) can be translated in the *translation builder*. The translation builder supports the xliff format (an XML variant) which allows translations to be done externally and imported into MOTUS. Furthermore, the *invitation builder* manages how respondents enter the workflow. There are different invitation strategies, ranging from voluntarily registering on the MOTUS webpage (possibly following advertising through various channels), via receiving a letter with login details, to uploading a list of potential respondents in advance. For a TUS that follows the guidelines, NSIs typically draw their sample from a national population register wherein no email address information is available. In this case, the invitation builder generates usernames and temporary passwords which are printed in the invitation letters that are send to the sampled households. Invitation letters contain both a QR-code and a fully written web link

directing respondents to the MOTUS website. Once respondents use the login credentials to participate, MOTUS will ask them to provide an email address for further communication throughout the survey.

While all the collection instruments and communications are created in their respective builders, the *research builder* sets up the overall collection process workflow. The workflow brings all instruments them together and places them in a linear order based on the different stages a respondent must go through to successfully participate in a survey. As these stages typically consist of tasks to be performed (collection instruments to be completed or communications to be read), they may also be referred to as "tasks".

Moving through stages is based on actions governed by conditions that are defined in the research builder. The conditions can be based on the completion of tasks or can be time based (e.g., sending a reminder after 24 hours of inactivity). Actions are communicated to the respondent by means of communications that are created in the communication builder. Additionally, communication criteria can be defined as a function of the progress within a stage.

For a TUS that follows the guidelines, the workflow is complex. It starts with sampling household members that will receive credentials to log in to MOTUS and complete the tasks of filling out a household questionnaire and composing the household grid. Thereafter, all eligible household members will be invited via email to carry out several tasks in MOTUS: completion of a first individual questionnaire, completion of two focus days in the time use diary, and completion of a second individual questionnaire. Actions involve numerous communications, for example, on what task needs to be completed next, reminders to complete certain tasks, or instructions on how to record an online time use diary.

To demonstrate how this works in practice, Figure 2 gives an example of a simplified workflow of a TUS that involves an individual pre-questionnaire and a two-day time use diary. Each box defines a stage and includes the title of the stage, a short description of the stage, and the option (for the survey manager) to edit or delete the stage. Within each stage, different actions are defined (the dark coloured bars), such as communicating, proceeding to the next task, or closing the survey participation for the respondent after a predefined period of inactivity.

The communication builder improves cost efficiency, timeliness and punctuality since communication is created online and sent to respondents through automated processes. Since the transmission of communications is conditional, it is tailored to the respondent and might increase the involvement of the respondent. In turn, this might lower their burden and therefore improve the accuracy and reliability of the data. The translation builder cannot alter the translation costs. The major advantage, though, is that respondents can easily switch between languages, which again might increase their involvement and lower their burden, especially in countries such as Belgium with multiple official languages. In a TUS, the initial invitation comes in the form of a paper letter, so the improvement provided by the invitation builder is limited at first. However, in case information is provided by the head of household, the eligible household members are invited via e-mail, which is cost and time efficient. Additionally, automated processes for assigning credentials and linking these to UUIDs leaves less room for error which improves accuracy and reliability.

The research builder improves current practices of TUS substantially because it allows building the complete workflow in an online environment and as a fully automated process. It enhances cost efficiency, timeliness and punctuality, while also improves the

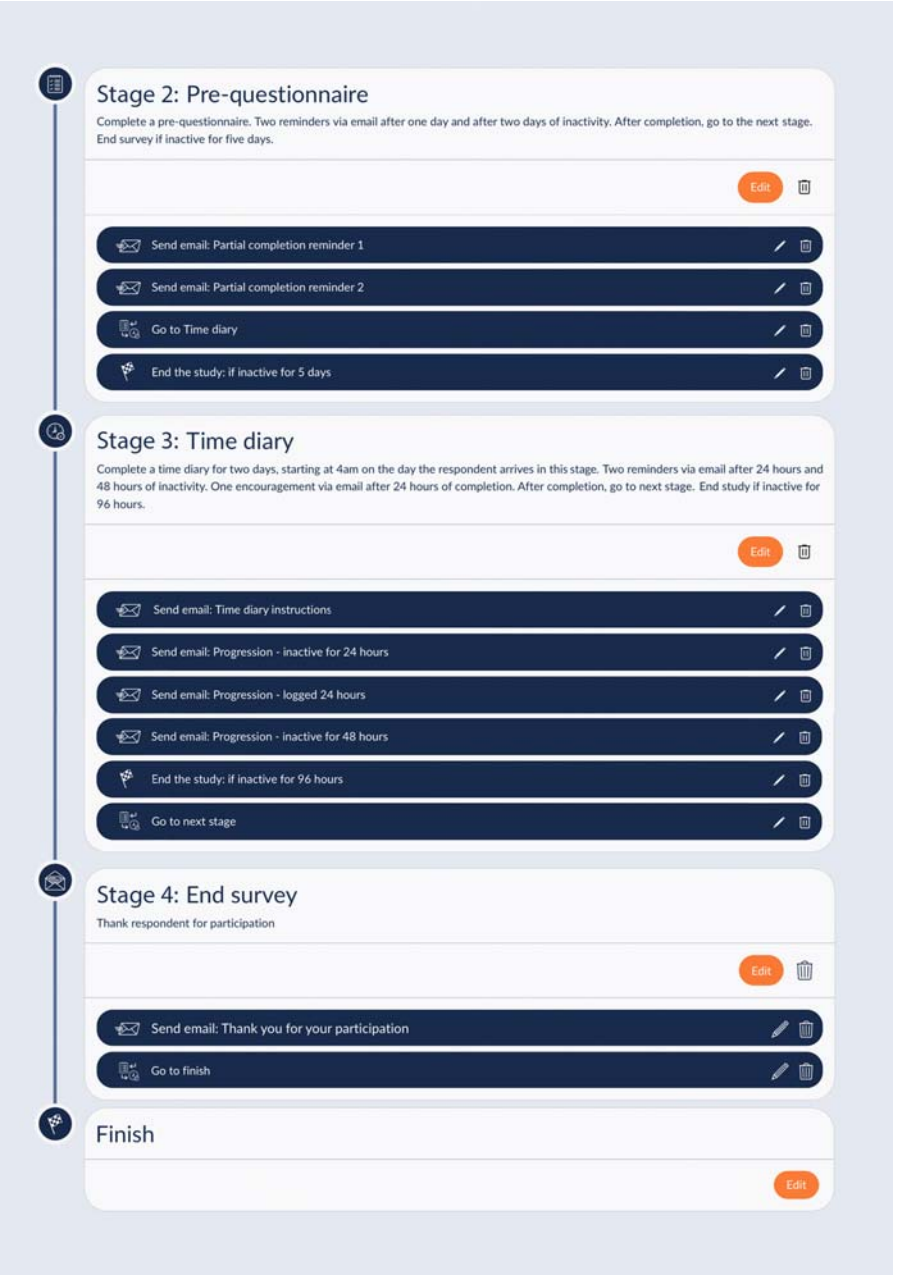


Fig. 2. Simplified workflow of a TUS on the MOTUS platform.
Note. Stage 1 (not pictured) involves the activation of the MOTUS account. The simplified workflow involves an individual pre-questionnaire (Stage 2) and a two-day time use diary (Stage 3).

accuracy and reliability of the data as it allows a more accurate and complete follow-up of respondents as they progress through the various stages. Although this closer follow-up cannot reduce the number of tasks involved, the communication between tasks might lower the respondent burden as it creates a sense of being supported.

3.3. The Generic Statistical Business Process Model

Building (or reusing) the designed collection instruments and processes is central to any statistical production process and part of the GSBPM. The GSBPM serves as a framework to describe and define the business processes involved to produce official statistics in a standardised way. It started as a joint effort of the United Nations Economic Commission for Europe (UNECE), Eurostat and the Organisation for Economic Cooperation and Development (OECD). The GSBPM is based on the business model of Statistics New Zealand (Kuonen and Loison 2019). Describing the business process of the production of official statistics using the GSBPM as the reference model allows NSIs to communicate these processes more easily.

The GSBPM is considered a non-linear process model and is aimed to apply to any data production (e.g., surveys, censuses, administrative registers). It serves as a reference model, which does not prevent NSIs from arriving at national versions of the GSBPM based on organisation-specific adaptations, combining phases, or a sequential reassessment to make it a linear process description (Ahmad and Koh 2011).

As shown in the first three columns of Table 1, each of the builders discussed above refers to one or more of the subphases of the build phase (i.e., GSBPM phase 3), while also supporting one or more other process phases (i.e., GSBPM phases 4 to 7). This highlights the non-linear sequence of the different phases of the GSBPM and the importance of iterative processes to support, evaluate and inform different phases and sub-phases.

4. Discussion

4.1. Wider Applications

The MOTUS applications (mobile and web) are not single purpose applications aimed at conducting a particular survey (or supporting a single area of statistics, such as time-use statistics). Instead, the MOTUS front office applications serve as a host for any survey that is defined in the back office. This modular capacity of MOTUS is based on the different builders that can be defined and put into a workflow for every different survey created in MOTUS. As such, MOTUS works particularly well for complex studies that are a sequence of multiples tasks (e.g., questionnaire and diary) or studies that link survey elements with data from other, external services (e.g., geolocation data). The Household Budget Survey (HBS) is an example of a complex survey with challenges comparable to those of the TUS. Like the TUS, it is also sampled at the household level and consists of recording data in a diary over time (in the the case of the HBS, this concerns purchases by household members over a period of at least 15 days). The HBS also includes completing a household grid and questionnaire. Given these major similarities and the modular approach of MOTUS, the project CRŒSS (Minnen et al. 2022) upgraded MOTUS to a platform that also can offer HBS studies. This was done by extending the diary builder, which can now also use the Classification of Individual Consumption by Purpose (COICOP) codes (instead of the ACL code used in a TUS). The adjustments achieved uniformity of the front office in the sense that the UI/UX is the same for TUS and HBS. This also holds for the back office. The MOTUS platform can now organize both a TUS and an HBS on the same platform and with the same applications. At the same time,

MOTUS uses container technology to make the platform available as an ESS platform. Each Docker container is a separate part of the MOTUS platform, as shown in Figure 1, with its software dependencies. How and where the containers are used is the responsibility of an NSI. It is recommended to use Kubernetes to deploy the containers on ISO/IEC 27001 certified infrastructure. This setup brings natural security barriers and also provides tools for scalability and high availability.

4.2. Smart Data Collection

Another future challenge of (digital) data collection concerns "smart" ways of collecting data, from which time use surveys could benefit (Zeni et al. 2020). "Smart" refers to data collection that combines passive or sensor data from personal smart devices (e.g., GPS, accelerometer) with active data explicitly provided by the respondent (e.g., responses to queries). Here, "passive" refers to the respondent not actively providing input (Ricciato et al. 2020).

MOTUS interprets the "smart" concept in a very broad sense, noting that data collection can be smart not only in the way it uses or processes already available data, but also be smart in the way it supports respondents to participate in surveys. MOTUS therefore continues to develop and add builders with new possibilities to the back office. One such builder is the *event builder*. Events follow the if-this-than-that (ITTT) approach and are thus triggers that are pulled if a certain condition is met. These conditions and the actions they initiate are defined in the event builder and are available from microservices that collect sensor data and are connected by an adapter API to communicate with MOTUS. These events can on the one hand ask the respondents to perform a specific action (e.g., answering a short questionnaire), or on the other hand show tentative entries in the respondents' diary, which they can commit and as such can reduce the registration burden and increase the quality of the registration. For example, if the GPS coordinates correspond with respondents' working address, working activities might be suggested in their time use diary.

The inclusion of smart data requires a data collection platform that is able to communicate with different other environments or standalone microservices (Ricciato et al. 2020). As shown in Figure 1, the MOTUS platform architecture allows these external smart data sources to communicate with the core via so-called adapter APIs. An example is the connection to the GeoService that collects geolocation data points from the respondents' smartphones. Particularly in complex studies such as TUS and HBS, the inclusion of sensor data, or administrative data in line with the Only Once Principle (OOP), should result in increased response rates, lower time investments of respondents as data providers, a further reduction of survey costs, and an increase in the accuracy and reliability of the data.

4.3. Para- and Metadata

The wealth of para- and metadata captured by MOTUS can provide insights into a lot of processes that have remained hidden from view in the traditional paper-and-pencil TUSs. For example, who actually completes the time use diaries? Each household members by themself? Or one person for all? We can only guess how this might have affected the intra-household correlation of the time use diary registration. Similarly, when were time diaries completed? Throughout the day? At the end of the day? Or just before the interviewer

came to pick up the diaries? Again, we can only guess how this might have affected the reliability of the time use diary registration in the past. Furthermore, if respondents drop out during the fieldwork, all information prior to drop-out remains available in the database of the server running the survey. Unless the respondents exercise their rights as defined in the General Data Protection Regulation (http://data.europa.eu/eli/reg/2016/679/2016-05-04) to delete all stored information. This might be useful to evaluate the dropout. On the negative side, it is yet to become known what all this will bring to light in terms of accuracy and reliability. On the positive side, at least then we know – and may be able to compensate for it.

4.4. Communication

One of the future challenges of online research, and especially with surveys like TUSs, are the multiple and complex tasks respondents must complete. The absence of face-to-face contact puts substantial pressure on online communication and gives rise to questions such as how much to communicate, by which means and in which wordings – and whether the communication should be differentiated by background characteristics. Options for respondents to switch on or off optional communication, such as reminders, suggestions, tips and tricks and select preferred media channels (e.g., email, text message, on screen notifications) could further tailor the user experience to the respondent and increase the feeling of being supported and decrease the potential challenging effort to complete the survey (Yan et al. 2019).

4.5. Conclusion

TUSs have a history of collecting data that can produce reliable and widely applicable statistics and indicators. However, the implementation of a (HETUS based) TUS is based on a complex sequence of household and individual level questionnaires and time use diaries on two different days of the week. A paper-and-pencil version comes with high postal and printing costs and with substantial cost and time investments in multiple interventions from interviewers and coders. These surveys also imply a relative high participation burden and thus a risk for accuracy and reliability. The modernization of TUSs, driven by current and future technological developments, involves more than just translating the current paper and pen-based version into a digital format. It requires a shift in the methodological paradigm of doing these surveys and an overhaul of the business processes for producing official time use statistics.

This contribution introduced MOTUS not only as an online TUS, but as a provider for the collection of these surveys by breaking down all elements of conducting an online TUS into modular builders that are congruent with and supportive to several subphases of the GSBPM. It showed that MOTUS stands for a modern approach to surveys in general and to complex surveys (such as the TUS and the HBS) in particular. The MOTUS builders inform the design phase, enable the build phase, and facilitate the collect, process, analyse, and disseminate phases of the GSBPM. It also showed that MOTUS makes it possible for modern, online data collections to provide a partial answer to recent challenges by lowering the respondent burden, by being more cost efficient, and by providing timelier, more punctual, more accurate and more reliability official statistics. MOTUS has already

partly proven itself in the past for TUS both for a population sampled TUS (see Minnen et al. 2014) and for several target sampled TUSs (see, for example, Te Braak et al. 2022a). Future challenges include further applications and use of MOTUS for TUS and other surveys in different statistical domains (e.g., the HBS – for which first steps have been taken as described in Subsection 4.1 above) and collecting feedback for adjustments and improvements. These applications and subsequent evaluations will continue to cement and expand the potential of MOTUS to meet current challenges of and changes in producing official statistics based on complex surveys.

5. References

Ahmad, N., and S.-H. Koh. 2011. *Incorporating estimates of household production of non-market services into international comparisons of material well-being. UNECE Working Paper* No. 42. STD/DOC(2011)7. DOI: https://doi.org/10.1787/5kg3h0jgk87g-en.

Ashofteh, A., and J.M. Bravo. 2021. "Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems." *Statistical Journal of the IAOS* 37(3): 771–789. DOI: https://doi.org/10.3233/SJI-210841.

Bonke, J., and P. Fallesen. 2010. "The impact of incentives and interview methods on response quantity and quality in diary-and booklet-based surveys." *Survey Research Methods* 4(2): 91–101. DOI: https://doi.org/10.18148/srm/2010.v4i2.3614.

Bruno, M., F. Inglese, and G. Ruocco. 2022. "Trusted Smart Surveys: Architectural and Methodological Challenges Related to New Data Sources." In *Studies in Theoretical and Applied Statistics*, edited by N. Salvati, C. Perna, S. Marchetti,and R. Chambers, Springer Proceedings in Mathematics & Statistics, 406. DOI: https://doi.org/10.1007/978-3-031-16609-9_31.

Cai, L, and Y. Zhu. 2015. "The challenges of data quality and data quality assessment in the big data era." *Data science journal* 14. DOI: http://doi.org/10.5334/dsj-2015-002.

Carletto, C., H. Chen, T. Kilic, and F. Perucci. 2022. "Positioning household surveys for the next decade." *Statistical Journal of the IAOS* 38(3): 923–946. DOI: https://10.3233/SJI-220042.

Chenu, A. 2004. "Prendre la mesure du travail." In *Pour une histoire des sciences sociales. Hommage à Pierre Bourdieu*, edited by J. Heilbron, R. Lenoir and G.D. Sapiro: 281–304. Paris: Fayard.

Eurostat. 2018. Eurostat, *European statistics code of practice: for the national statistical authorities and Eurostat (EU statistical authority).* Luxembourg: Publications Office of the European Union. DOI: https://doi.org/10.2785/798269:

Eurostat. 2020. *Harmonised European Time Use Surveys (HETUS) – 2018 guidelines – Re-edition*. Luxembourg: Publications Office of the European Union. DOI: https://doi.org/10.2785/926903.

Fernee, H., and N. Sonck. 2013. "Is everyone able to use a smartphone in survey research?" *Survey Practice* 6(4): 2884. DOI: https://doi.org/10.29115/SP-2013-0020.

Gohar, A., and G. Nencioni. 2021. "The role of 5G technologies in a smart city: The case for intelligent transportation system." *Sustainability* 13(9): 5188. DOI: https://doi.org/10.3390/su13095188.

Juster, F.T. 1986. "Response errors in the measurement of time use." *Journal of the American Statistical Association* 81(394): 390–402. DOI: https://doi.org/10.1080/01621459.1986.10478283.

Keusch, F., B. Struminskaya, C. Antoun, M.P. Couper, and F. Kreuter. 2019. "Willingness to participate in passive mobile data collection." *Public Opinion Quarterly* 83 (S1): 210–235. DOI: https://doi.org/10.1093/poq/nfz007.

Kuonen, D., and B. Loison. 2019. "Production processes of official statistics and analytics processes augmented by trusted smart statistics: Friends or foes?" *Statistical Journal of the IAOS* 35(4): 615–622. DOI: https://doi.org/10.3233/SJI-190530.

Lavrakas, P.J. 2008. *Encyclopedia of survey research methods*. Sage Publications.

Minnen, J., I. Glorieux, T.P. van Tienoven, S. Daniels, D. Weenas, J. Deyaert, S. van den Bogaert, and S. Rymenants 2014. "Modular Online Time Use Survey (MOTUS)-Translating an existing method in the 21st century." *Electronic International Journal of Time Use Research* 11(1): 73–93. DOI: https://dx.doi.org/10.13085/eIJTUR.11.1.73-93.

Minnen, J., J. Olsen, and K. Sabbe. 2022. *CRŒSS: Establishing a Cross-domain data collection platform for the ESS (European Statistical System)*. Brussels and Bonn: Statistics Belgium, Destatis, hbits CV and Vrije Universiteit Brussel. Available at: https://torvub.be/torwebdat/publications/t2023_13.pdf.

Pronovost, G. 1989. "The sociology of time." *Sociologie Contemporaine (La)* 37(3): 1–124.

Radermacher, W.J. 2020. *Official Statistics 4.0. Verified Facts for People in the 21st Century*. Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-030-31492-7.

Ricciato, F., A. Wirthmann, K. Giannakouris, and M. Skaliotis. 2019. "Trusted smart statistics: Motivations and principles." *Statistical Journal of the IAOS* 35(4): 589–603. DOI: https://doi.org/10.3233/SJI-190584.

Ricciato, F., A. Wirthmann, and M. Hahn. 2020. "Trusted Smart Statistics: How new data will change official statistics." *Data and Policy* 2. DOI: https://doi.org/10.1017/dap.2020.7.

Robinson, J.p. 1999. "The time diary method. Structure and uses." In *Time use research in the social sciences*, edited by W.E. Pentland, A.S. Harvey, M.P. Lawton, and M.A. McColl: 47–89. New York: Kluwer Academic/Plenum Publishers. DOI: https://doi.org/10.1007/0-306-47155-8_3.

Robinson, J.P., and G. Godbey. 1997. *Time for life: The surprising ways Americans use their time*. Pennsylvania: Penn State Press.

Salemink, I., S. Dufour, and M. van der Steen. 2020. "A vision on future advanced data collection." *Statistical Journal of the IAOS* 36 (3): 685–699. DOI: https://10.3233/SJI-200658.

Salgado, D., M.E. Esteban, M. Novás, S. Sadaña, and L. Sanguiao. 2018. "Data Organisation and Process Design Based on Functional Modularity for a Standard Production Process." *Journal of Official Statistics* 34(4): 811–833. DOI: https://doi.org/10.2478/jos-2018-0041.

Sonck, N., and H. Fernee. 2013. *Using smartphones in survey research: a multifunctional tool.* The Hague: The Netherlands Institute for Social Research.

Stodden, V. 2014. "The reproducible research movement in statistics." *Statistical Journal of the IAOS* 30(2): 91–93. DOI: https://doi.org/10.3233/SJI-140818.

Sullivan, O., J. Gershuny, A. Sevilla, P. Walthery, and M. Vega-Rapun. 2020. "Time use diary design for our times-an overview, presenting a Click-and-Drag Diary Instrument

(CaDDI) for online application." *Journal of Time Use Research* 10. DOI: https://doi.org/10.32797/jtur-2020-1.

Szalai, A. 1972. *The use of time: Daily activities of urban and suburban populations in twelve countries*. The Hague: Mouton.

Te Braak, P., F. van Droogenbroeck, J. Minnen, T.P. van Tienoven, and I. Glorieux. 2022a. "Teachers' working time from time-use data: Consequences of the invalidity of survey questions for teachers, researchers, and policy.? *Teaching and Teacher Education* 109: 103536. DOI: https://doi.org/10.1016/j.tate.2021.103536.

Te Braak, P., T.P. van Tienoven, J. Minnen, and I. Glorieux. 2022b. "Bias in estimated working hours in time use diary research: The effect of cyclical work time patterns on postponing designated registration days.? *Time and Society* 31(4): 508–534. DOI: https://doi.org/10.1177/0961463X221111948.

United Nations. 2016. *Integrating a Gender Perspective into Statistics*. *Studies in Methods, Series F No. 111*. New York: United Nations Publication. Available at: https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/gender/Integrating-a-Gender-Perspective-into-Statistics-E.pdf.

U.S. Bureau of Labor Statistics. 2023. *American Time Use Survey User's Guide. Understanding ATUS 2003 to 2023*. Available at: https://www.bls.gov/tus/atususersguide.pdf.

Vassilev, G., W. King, S. Wallace, and J. White. 2020. *Modernization of the Production of Time-use Statistics.* UK: UK Office for National Statistics. https://unstats.un.org/unsd/-statcom/53rd-session/documents/BG-3h-Modernization_UN_EG_TUS2021_FINAL_-SENT_rev-E.pdf.

Yan, T., S. Fricker and S. Tsai. 2019. "Response burden: What is it and what predicts it?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P. Beatty, D. Collins, L. Kaye, J.L. Padilla, G. Willis and A. Wilmot: 193–212. New Jersey: John Wiley & Sons. DOI: https://doi.org/10.1002/9781119263685.

Zeni, M., I. Bison, F. Reis, B. Gauckler, and F. Giunchiglia. 2020. "Improving Time Use Measurement with Personal Big Data Collection – The Experience of the European Big Data Hackathon 2019." *Journal of Official Statistics* 37(2): 341–365. DOI: https://doi.org/10.2478/jos-2021-0015.

Zerubavel, E. 1982. *Hidden rhythms: Schedules and calendars in social life*. Chicago: The University of Chicago Press.

Received September 2021
Revised November 2022
Accepted June 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 507–533, http://dx.doi.org/10.2478/JOS-2023-0024

Small Area with Multiply Imputed Survey Data

Marina Runge[1] *and Timo Schmid*[2]

In this article, we propose a framework for small area estimation with multiply imputed survey data. Many statistical surveys suffer from (a) high nonresponse rates due to sensitive questions and response burden and (b) too small sample sizes to allow for reliable estimates on (unplanned) disaggregated levels due to budget constraints. One way to deal with missing values is to replace them by several plausible/imputed values based on a model. Small area estimation, such as the model by Fay and Herriot, is applied to estimate regionally disaggregated indicators when direct estimates are imprecise. The framework presented tackles simultaneously multiply imputed values and imprecise direct estimates. In particular, we extend the general class of transformed Fay-Herriot models to account for the additional uncertainty from multiple imputation. We derive three special cases of the Fay-Herriot model with particular transformations and provide point and mean squared error estimators. Depending on the case, the mean squared error is estimated by analytic solutions or resampling methods. Comprehensive simulations in a controlled environment show that the proposed methodology leads to reliable and precise results in terms of bias and mean squared error. The methodology is illustrated by a real data example using European wealth data.

Key words: Fay-Herriot model; mean squared error; multiple imputation; nonresponse; survey statistics.

1. Motivation

Financial reports based on asset data can provide insights into a wide range of issues of major importance for political decisions and can help in the precise allocation of funds. In addition, wealth data can give an overview of the distribution of assets and liabilities, which can be highly relevant for financial stability and play a central role in assessing inequality. For this reason, survey data on wealth are of particular importance. Since questions about assets and income are sensitive issues, such surveys often suffer from high item nonresponse (Riphahn and Serfling 2005). For example, the Household Finance and Consumption Survey (HFCS) reports for France item nonresponse rates of nearly 30% for value of saving accounts and largest mortgage on household main residence and almost 80% for current value of household main residence (HFCN 2020a).

Listwise deletion, retaining only records with no items missing, leads to a loss of information, and the remaining units in this dataset are not a good representation of the population, which can lead to biased estimates. Missing values are a problem because the

[1] Institute of Statistics and Econometrics, Freie Universität Berlin, Garystraße 21, 14195 Berlin, Germany, Email: marina.runge@fu-berlin.de
[2] Institute of Statistics, Otto-Friedrich-Universität Bamberg, Feldkirchenstraße 21, 96045 Bamberg, Germany, Email: timo.schmid@uni-bamberg.de
Acknowledgments: The authors appreciate gratefully the support of the German Research Foundation within the TESAP project (grant number: 281573942). This article uses data from the Eurosystem Household Finance and Consumption Survey (HFCS). The results published and the related observations and analysis may not correspond to results or analysis of the data producers. Finally, the authors are indebted to the Editor-in-Chief, Associate Editor and the referees for comments that significantly improved the article.

incomplete data do not have the regular (matrix) form needed in almost any statistical method, and therefore handling missing values is necessary. In the literature there are various approaches for dealing with missing data in studies, such as in Rubin (1987) or Longford (2005). Van Buuren (2018) gives an extended overview of approaches to handling and imputing of missing data. Rubin (1976) formulated for the first time the concept of missing data mechanisms by using the indicators of the missing values as random variables and posited a model for them. Methods for missing data are generally based on the assumption that the probability of the missing data does not depend on the missing values after conditioning on the observed values (MAR). To obtain valid statistical inferences, appropriate assumptions about the mechanism of missing values must be made (Van Buuren 2018). Two approaches to handling incomplete data are single imputation, where each missing value is imputed once, and multiple imputation (MI), where the missing values are replaced by a small number of plausible values. The advantage of MI is that it reflects the uncertainty of missing data, which is then taken into account in the estimation. There are several surveys of income and wealth data where MI is used, including the Consumer Expenditure Survey, where the income variable is imputed five times (Fisher 2006), and the HFCS, where also five imputations of the data sets are provided to the user (HFCN 2020a).

Of particular interest may be subpopulations of households, either regionally disaggregated or sociodemographic such as households with particular composition (of ages, gender, labor market status, or educational levels). Various political decisions or global events, such as the financial crisis of 2007/2008 or the COVID-19 pandemic in 2020/2021, may affect these subgroups, usually referred to as areas or domains, to varying degrees. Some of these domains may be represented by very few units in the sample and direct estimators (based only on these subjects) result in a large variance. This issue may be solved by small area estimation (SAE) methods. The model-based estimators used in SAE supplement information from other areas and other data sources. Pfeffermann (2013), Rao and Molina (2015) and Jiang and Rao (2020) give compact overviews and Tzavidis et al. (2018) propose a general framework for the production of small area statistics. SAE methods can be distinguished in unit-level (e.g., Battese et al. 1988) and area-level (Fay and Herriot 1979) models. Unit-level models have the greater information content, but can only be used when unit-level covariate data are available. In addition, area-level models are often used because they are better suited to account for complex survey designs for point and variance estimates. Therefore, we focus on the Fay-Herriot model in this article. The Fay-Herriot model can be applied to transformed direct estimators to attain normality of the error terms or to ensure that the resulting estimates are within an appropriate range. Slud and Maiti (2006) and Chandra et al. (2017) study the log-transformed Fay-Herriot model and Sugasawa and Kubokawa (2017) consider a general parametric transformation of the response values. Schmid et al. (2017) use an arcsine transformation to estimate literacy rates of Senegal and Casas-Cordero et al. (2016) to estimate poverty rates of Chile.

In the context of SAE, nonresponse rates in combination with small sample sizes could have significant influence on the estimates especially with sensitive data such as income and wealth data. The investigation of the integration of the imputation uncertainty into small area estimators has received some attention. Among the publications are, for example, Longford (2004), who uses a multiple hot-deck imputation method in the UK

Labour Force Survey to estimate unemployment rates using a small area multivariate shrinkage method. Longford (2005) presents methods for dealing with incomplete data and making inferences using small area estimation methods. An approach to modeling the non-missing at random mechanism in SAE under informative sampling and nonresponse can be found in Sverchkov and Pfeffermann (2018). Kreutzmann et al. (2022) and Bijlsma et al. (2020) use a Fay-Herriot model with pooled direct estimators after multiple imputation and take into account the additional uncertainty due to the missing values in the sampling variance. However, both ignore the additional uncertainty in the regression-synthetic part of the model. We extend this approach to address the latter problem in addition to extending the methodology to ratios.

We present an approach in which we combine MI with the transformed Fay-Herriot model. We take the multiply imputed values of the missing values as given by the data provider. To account for the additional uncertainty from imputation, pooled components of the direct estimator are used, as well as pooled components of the regression-synthetic part of the Fay-Herriot model. In particular, the components (direct and regression-synthetic part) are combined for a given transformation in such a way that the resulting MI adjusted model has the known structure of Fay-Herriot models. This approach exploits the existing knowledge about transformations, back-transformations and mean squared error (MSE) approximations of the transformed Fay-Herriot model. We apply the general approach to three special cases relevant to practice and additionally discuss MSE estimators for these special cases:

1. For the general Fay-Herriot model for a mean value, we adapt the Prasad-Rao MSE estimator (Prasad and Rao 1990) to account for the uncertainty owing to missing values.
2. If the distribution of the target indicator is right-skewed, a log transformation can be used. For this case, we use the adapted Prasad-Rao MSE estimator and apply a back-transformation similar to that presented in Rao and Molina (2015).
3. For the Fay-Herriot model for a ratio with an arcsine transformation, we use insights from Hadam et al. (2023) for the back-transformation of the point estimator, as well as for a parametric bootstrap MSE estimator that can reflect the uncertainty due to the missing values.

The validity of the presented point estimators is demonstrated for the three cases outlined above in a simulation study. It is also shown that the additional uncertainty caused by the missing values is accounted for by the proposed MSE estimators.

The article is structured as follows. Sections 2, 3, and 4 describe the statistical methodology. In Section 2, the transformed Fay-Herriot model is presented, which serves as the basis for the combination with MI. Section 3 describes how the direct and regression-synthetic components of the transformed Fay-Herriot model are combined after MI, which leads to a MI adjusted Fay-Herriot model. In Section 4, we consider three common special cases of the model from Section 3 and present associated uncertainty measures. The proposed methodology is evaluated in simulation experiments in Section 5 and then applied to HFCS data in Section 6. Section 7 summarizes the main findings, discusses limitations of the approach and outlines further research potential.

2. Transformed Fay-Herriot Model

In the following the transformed Fay-Herriot model is introduced, where the transformation is described by a known function h. Let N be the size of a finite population which is partitioned into $d = 1, ..., D$ domains and n the sample size with $i = 1, ..., n_d$ units per domain so that $n = \sum_{d=1}^{D} n_d$. The Fay-Herriot model involves in the first stage a sampling model in which it is supposed that the direct estimator consists of the true domain-specific population indicator θ_d and a sampling error e_d:

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \qquad e_d \overset{ind}{\sim} N\left(0, \sigma_{e_d}^2\right).$$

It is assumed that the sampling errors e_d are independently normally distributed with known variance $\sigma_{e_d}^2$. Although the sampling variances $\sigma_{e_d}^2$ are assumed to be known, in practice they are estimated by unit-level data (Rivest and Vandal 2002; Wang and Fuller 2003: You and Chapman 2006). Another unit-level approach to address the problem of unknown sampling variances is proposed by Maiti et al. (2014) and Sugasawa et al. (2017) by shrinking and simultaneous modeling of small area means and variances. When the indicator of interest is a mean value, a domain specific direct estimator is the weighted average of the sampled values:

$$\hat{\theta}_d^{Dir} = \frac{\sum_{i=1}^{n_d} w_{id} y_{id}}{\sum_{i=1}^{n_d} w_{id}}.$$

The incorporation of sampling weights w_{id} makes the point estimator design unbiased. Note that the population and the outcomes y_{id} are assumed to be fixed, and the sampling mechanism is the only source of uncertainty. The sampling weights reflect a complex design in the estimation of the associated variance. The second stage of the Fay-Herriot model is a linking model, which links covariate information to the population indicator. x_d is a $p \times 1$ vector with area-level population covariates and β is the corresponding $p \times 1$ vector with regression coefficients. υ_d are normally distributed domain specific random effects:

$$\theta_d = x_d^T \beta + \upsilon_d, \qquad \upsilon_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right). \tag{1}$$

Combining the sampling and the linking model results in:

$$\hat{\theta}_d^{Dir} = x_d^T \beta + \upsilon_d + e_d, \ \upsilon_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right), \ e_d \overset{ind}{\sim} N\left(0, \sigma_{ed}^2\right). \tag{2}$$

If a smooth and monotone transformation function h is applied to the direct estimator, $\hat{\theta}_d^{Dir}$ is replaced by $\hat{\theta}_d^{Dir*} := h\left(\hat{\theta}_d^{Dir}\right)$ in Equation (2) and we want to predict $h^{-1}(\theta_d)$. The transformed Fay-Herriot model is then defined, for example, as in Sugasawa and Kubokawa (2017):

$$h\left(\hat{\theta}_d^{Dir}\right) = x_d^T \beta + \upsilon_d + e_d, \qquad \upsilon_d \overset{iid}{\sim} N\left(0, \sigma_v^2\right), \qquad e_d \overset{ind}{\sim} N\left(0, \sigma_{ed}^{2*}\right). \tag{3}$$

In the following, * always refers to the transformed scale of the direct estimator, its variance and the Fay-Herriot estimator presented at the end of this section. The model parameters, the model variance σ_v^2 and the regression coefficients β are not known and

must be estimated. There are various methods to obtain estimates of σ_v^2, for example, restricted maximum likelihood (REML), maximum likelihood (ML) and the FH method-of-moments. More details on the estimation methods of the model variance can be found in Chapter 6 in Rao and Molina (2015). A drawback of ML is that it does not account for the loss in degrees of freedom arising from the estimation of the regression coefficients β (Rao and Molina 2015). Therefore, we use in this article the REML method. The regression coefficients β and the random effects v_d are estimated by:

$$\hat{\beta} = \hat{\beta}(\hat{\sigma}_v^2) = \left(\sum_{d=1}^{D} \frac{x_d x_d^T}{\sigma_{e_d}^{2^*} + \hat{\sigma}_v^2} \right)^{-1} \left(\sum_{d=1}^{D} \frac{x_d \hat{\theta}_d^{Dir^*}}{\sigma_{e_d}^{2^*} + \hat{\sigma}_v^2} \right), \tag{4}$$

$$\hat{v}_d = \frac{\hat{\sigma}_v^2}{\sigma_{e_d}^{2^*} + \hat{\sigma}_v^2} \left(\hat{\theta}_d^{Dir^*} - x_d^T \hat{\beta} \right). \tag{5}$$

Plugging those predictors into Equation (1) leads to the empirical best linear unbiased predictor (EBLUP), that is, the transformed Fay-Herriot estimator:

$$\hat{\theta}_d^{FH*} = x_d^T \hat{\beta} + \hat{v}_d. \tag{6}$$

This estimator can be expressed as a convex combination of the direct estimator and the regressionsynthetic component, resulting in an optimal combination of the two components. If the variance of the direct estimator is large, more weight is given to the synthetic component, and vice versa:

$$\hat{\theta}_d^{FH^*} = \hat{\gamma}_d \hat{\theta}_d^{Dir^*} + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_{e_d}^{2^*} + \hat{\sigma}_v^2}. \tag{7}$$

At this point $\hat{\theta}_d^{FH^*}$ is still on the transformed scale and has to be transformed to the original scale to obtain $\hat{\theta}_d^{FH^*}$.

3. Combining Transformed Fay-Herriot Models after Multiple Imputation

An often applied technique to handle missing values is MI, where the missing values are replaced by several plausible values. To obtain these values, an imputation model is required. It is not sufficient to generate only one imputation, since the imputation is treated as if it were true, and the uncertainties arising from the nonresponse are ignored. On the contrary, a large number of imputations is usually not necessary, and M between 5 and 20 is sufficient, but it may be advantageous to choose a higher value (20–100) if the non-response is high and there is a large uncertainty about the estimand (Van Buuren 2018). The procedure for MI involves two steps: the imputation step and the analysis step. In the former, the imputer, usually the data provider, generates the M replicate completions of the survey data using a suitable imputation model and provides them to the analyst. In the second step, the analyst applies a statistical model suitable for the complete data separately to each imputed data set. The focus of this article is on the latter. If θ is the indicator of interest and $\hat{\theta}$ its estimator, the analysis model is calculated with each imputed data set, so we obtain $\hat{\theta}_m$ and $\widehat{\text{Var}}(\hat{\theta}_m)$ for $m = 1, ..., M$. The results are then combined with the application of pooling rules developed by Rubin (1987) for point estimates and their

variances, which include the additional variability and uncertainty induced by the missing data. Rubin's rules (RR) are defined as follows. The pooled estimator of θ is the mean value of the M estimators:

$$\hat{\theta}^{RR} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m. \tag{8}$$

The variance of the pooled estimator $\hat{\sigma}^{2RR}$ is composed by the mean value of the individual variances of each estimator (within-variance) and the variance between the M estimates (between-variance) with an correction due to the finite sample size:

$$\hat{\sigma}^{2^{RR}} = \widehat{\mathrm{Var}}\left(\hat{\theta}^{RR}\right) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathrm{Var}}\left(\hat{\theta}_m\right) + \frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}^{RR}\right)^2. \tag{9}$$

In the next sections, we describe how the combining rules are applied to the components of the transformed Fay-Herriot model from Section 2.

3.1. Component Pooling

With the M multiply imputed sampling values $y_{id,m}$ of each unit $i = 1, ..., n_d$ and domains $d = 1, ..., D$, the transformed direct estimators $\hat{\theta}^{Dir^*}_{d,m} = h\left(\hat{\theta}^{Dir}_{d,m}\right)$ of the target indicator and their corresponding sample variances $\sigma^{2^*}_{e_{d,m}}$ are calculated for each domain $d = 1, ..., D$ and $m = 1, ..., M$. Rubin's rules are based on asymptotic theory, and the resulting combined estimate is more accurate if the distribution of the indicator of interest is better approximated by the normal distribution (Rubin 1987). Van Buuren (2018) states that to promote approximate normality, target indicators can be transformed, then pooled and back-transformed. Therefore, the M direct estimators $\hat{\theta}^{Dir^*}_{d,m}$ and their variances $\sigma^{2^*}_{e_{d,m}}$ are pooled on the transformed scale and substituted in Equations (8) and (9). Kreutzmann et al. (2022) present a Fay-Herriot estimator which uses pooled direct components on the original scale, which are substituted in the (log transformed) Fay-Herriot model. We extend this approach and transform the direct components of each imputed data set to estimate the regression-synthetic components. This allows the uncertainty of the missing values to be included not only in the direct components, but also in those of the linking model. The model components of the linking model are estimated for each imputed data set. The estimated variances of the random effects $\hat{v}_{d,m}$ are combined according to Rubin's rule:

$$W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2_{v_m} \quad \text{and} \quad B_d = \frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{v}_{d,m} - \frac{1}{M}\sum_{m=1}^{M}\hat{v}_{d,m}\right)^2. \tag{10}$$

The mean squared distance of the random effects of the domains of the M imputed data sets and the pooled random effects per domain is different between the areas. In order to guarantee that the random effects have a common variance, further pooling has to be applied. Therefore, the mean value of the between variance is taken. Together with Equation (10) this leads to the pooled model variance:

$$\hat{\sigma}^{2^{RR}}_v = W + \frac{1}{D}\sum_{d=1}^{D}B_d. \tag{11}$$

The pooled model variance $\hat{\sigma}_v^{2^{RR}}$ and the pooled direct components are now used to obtain MI adjusted estimates of the regression coefficients and random effects $\hat{\sigma}_v^{2^{RR}}$, $\hat{\sigma}_{e_d}^{2^{RR^*}}$ and $\hat{\theta}_d^{Dir.RR^*}$ are inserted into Equation (4) to obtain the MI adjusted regression coefficients $\hat{\beta}$ and then together into Equation (5) to obtain the MI adjusted random effects $\hat{v}_d$.

3.2. MI Adjusted Fay-Herriot Model

The pooled direct components together with the pooled and MI adjusted regression-synthetic parts of the model lead to the MI adjusted Fay-Herriot model, which preserves the structure of the transformed Fay-Herriot model. The area-level population auxiliary information x_d, obtained from external sources, such as the census, is fixed and complete as in Equation (1). The model can be written analogously to Equation (3) with pooled direct components and the pooled model variance. Using the estimators of unknown model parameters as elaborated in Subsection 3.1 leads to the proposed FH.MI estimator $\hat{\theta}_d^{FH.MI^*}$, which can be written analogously to Equation (7) with $\hat{\theta}_{e_d}^{Dir.RR^*}$, $\hat{\sigma}_{e_d}^{2^{RR^*}}$ and $\hat{\sigma}_v^{2^{RR}}$ plugged in:

$$\hat{\theta}_d^{FH.MI^*} = \hat{\gamma}_d \hat{\theta}_d^{Dir.RR^*} + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_v^{2^{RR}}}{\hat{\sigma}_{e_d}^{2^{RR^*}} + \hat{\sigma}_v^{2^{RR}}}. \tag{12}$$

The presented $\hat{\theta}_d^{FH.MI^*}$ estimator preserves the representations of the Fay-Herriot estimator. As $\hat{\theta}_d^{FH.MI^*}$ is on the transformed scale, a suitable back transformation depending on h has to be applied to obtain $\hat{\theta}_d^{FH.MI}$.

Small area estimators with multiply imputed data can be derived in two ways: 1. Fit the Fay-Herriot model to each of the M imputed data sets and combine the Fay-Herriot estimators with Rubin's rule. 2. Estimate the direct and the regression synthetic components M times and combine them using Rubin's rules as described in Subsection 3.1 and then estimate the shrinkage estimator in Equation (12). The advantage of the first approach is that it is simple. However, it loses the structure of the Fay-Herriot model and the representation of the estimator as a weighted combination of the direct and regression synthetic components. In addition, it is unclear how the uncertainty of the M Fay-Herriot estimators is combined, since Rubin's rule is commonly used for variances and it is unclear how this rule can be applied to the MSE. The advantage of the second (the proposed) approach and the resulting FH.MI estimator is that the model structure of the Fay-Herriot model is preserved, the interpretability of the components is maintained, and the existing knowledge about MSE estimators is directly transferable and extensible. The estimator of the first approach is used as a benchmark in the model-based simulation study in Section 5 and is denoted by FH.RR.

4. MI Adjusted Fay-Herriot Estimators with Uncertainty Measures

In the following sections, we focus on three special cases of the transformed MI adjusted Fay-Herriot estimator (12). For each case we specify the FH.MI point estimator and an associated MSE estimator.

4.1. Estimator for a Mean

The (population) mean of a quantity of interest for domain d is estimated by the weighted sample average per imputed data set m:

$$\hat{\theta}_{d,m}^{Dir} = \frac{\sum_{i=1}^{n_d} w_{id} y_{id,m}}{\sum_{i=1}^{n_d} w_{id}} \quad \text{for} \quad d = 1, .., D \quad \text{and} \quad m = 1, ..., M. \tag{13}$$

If no transformation is required for the direct estimator, $\hat{\theta}_d^{FH.MI^*}$ is on the original scale such that $\hat{\theta}_d^{FH.MI} = \hat{\theta}_d^{FH.MI^*}$. With the pooled and MI adjusted estimators presented in Section 3, the FH.MI estimator $\hat{\theta}_d^{FH.MI}$ can be calculated according to Equation (12). As a measure of uncertainty which captures the additional uncertainty due to multiple imputation, we adapt the MSE estimator of Prasad and Rao (1990) in the following. The second-order approximation of the MSE of $\hat{\theta}_d^{FH}$ is given by:

$$\text{MSE}\left(\hat{\theta}_d^{FH}\right) \approx g_{1d}\left(\sigma_v^2\right) + g_{2d}\left(\sigma_v^2\right) + g_{3d}\left(\sigma_v^2\right).$$

The first component g_{1d} is based on the prediction of the random effects and g_{2d} reflects the variability arising from the estimation of the regression coefficients. g_{1d} and g_{2d} are independent of the estimation method of the model variance σ_v^2, whereas, g_{3d} reflects the uncertainty caused by the estimation of σ_v^2 and depends on the estimation method through its asymptotic variance $\bar{V}\left(\hat{\sigma}_v^2\right)$ (as $D \to \infty$) (see e.g., Rao and Molina 2015). According to Prasad and Rao (1990) a second-order unbiased estimator of MSE ($\hat{\theta}_d^{FH}$) is:

$$\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH}\right) = g_{1d}\left(\hat{\sigma}_v^2\right) + g_{2d}\left(\hat{\sigma}_v^2\right) + 2g_{3d}\left(\hat{\sigma}_v^2\right).$$

The components of the Prasad-Rao estimator using REML are defined as follows:

$$g_{1d}\left(\hat{\sigma}_v^2\right) = \hat{\gamma}_d^2 \sigma_{e_d}^2, \tag{14}$$

$$g_{2d}\left(\hat{\sigma}_v^2\right) = (1 - \hat{\gamma}_d)^2 x_d^T \left\{ \sum_{d=1}^{D} \frac{x_d x_d^T}{\sigma_{e_d}^2 + \hat{\sigma}_v^2} \right\}^{-1} x_d, \tag{15}$$

$$g_{3d}\left(\hat{\sigma}_v^2\right) = (\sigma_{e_d}^2)^2 (\sigma_{e_d}^2 + \hat{\sigma}_v^2)^{-3} \bar{V}(\hat{\sigma}_v^2), \tag{16}$$

$$\bar{V}(\hat{\sigma}_v^2) = 2\left\{ \sum_{d=1}^{D} \frac{1}{\left(\sigma_{e_d}^2 + \hat{\sigma}_v^2\right)^2} \right\}^{-1}.$$

In the same way as in Subsection 3.1, where we obtain M estimates of the model variance, that is, $\hat{\sigma}_{v_m}^2$ for $m = 1, ..., M$, we obtain M corresponding asymptotic ($D \to \infty$) variances $\bar{V}_m(\hat{\sigma}_{v_m}^2)$ for $m = 1, ..., M$. To adjust the MSE estimator for this additional uncertainty, the asymptotic variances are pooled with Rubin's rule for variances (9):

$$\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right) = \frac{1}{M}\sum_{m=1}^{M}\bar{V}_m\left(\hat{\sigma}_{v_m}^2\right) + \frac{M+1}{M}\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\sigma}_{v_m}^2 - \hat{\sigma}_v^{2^{RR}}\right)^2$$

$$\text{with}\quad \bar{V}\left(\hat{\sigma}_{v_m}^2\right) = 2\left\{\sum_{d=1}^{D}\frac{1}{\left(\sigma_{e_{d,m}}^2 + \hat{\sigma}_{v_m}^2\right)^2}\right\}^{-1} \quad \text{for} \quad m = 1, ..., M. \tag{17}$$

Using $\hat{\sigma}_v^{2^{RR}}$ and $\sigma_{e_d}^{2^{RR}}$ in Equations (14), (15), and (16) together with the pooled asymptotic variance (17) takes into account the uncertainty about the missing values. Note that instead of plugging the pooled variance terms into the asymptotic variance formula, the pooled asymptotic variance $\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right)$ is used, introducing an additional term into the estimator due to the between-variation. This leads to the proposed MSE estimator for $\hat{\theta}_d^{FH.MI}$, which captures the uncertainty due to missing values:

$$\hat{\mathrm{V}}\left(\hat{\theta}_d^{FH.MI}\right) = g_{1d}\left(\hat{\sigma}_v^{2^{RR}}\right) + g_{2d}\left(\hat{\sigma}_v^{2^{RR}}\right) + 2\left(\sigma_{e_d}^{2^{RR}}\right)^2\left(\sigma_{e_d}^{2^{RR}} + \hat{\sigma}_v^{2^{RR}}\right)^{-3}\bar{V}^{RR}\left(\hat{\sigma}_v^{2^{RR}}\right). \tag{18}$$

4.2. Estimator for a log Mean

Domain specific mean values of income and wealth data are often skewed to the right, or the relationship with the auxiliary information may be non-linear. In such a case, the linear Fay-Herriot model (Subsection 4.1) may be more appropriate for the log-transformed direct estimator. Using the direct estimator from Equation (13) and $h\colon z \rightarrow log(z)$ the direct components of the model for the M imputed data sets are:

$$\hat{\theta}_{d,m}^{Dir^*} = \log\left(\hat{\theta}_{d,m}^{Dir}\right) \quad \text{with variances} \quad \sigma_{e_{d,m}}^{2^*} \approx \left(\hat{\theta}_{d,m}^{Dir}\right)^{-2}\sigma_{e_{d,m}}^2$$

$$\text{for} \quad d = 1, .., D, \quad m = 1, ..., M.$$

Using a Taylor expansion for moments, the sample variance, that is, the variance of the direct estimator, can be moved to the logarithmic scale. Although this is an approximation for large samples, it is used in SAE as in Neves et al. (2013). Citro and Kalton (2000) use the same approximation with a minor modification based on the properties of the log-normal distribution, while noting that the results do not differ considerably. Calculating the direct and the regression-synthetic components as described in Subsection 3.1 with $h\colon z \rightarrow log(z)$ and together with Equation (12) leads to the Fay-Herriot-MI estimator $\hat{\theta}_d^{FH.MI^*}$, which is still on the log-scale. The estimates can be transformed back to the original scale by several methods. Slud and Maiti (2006) present a bias-correction under a log-transformed Fay-Herriot model and propose a corresponding estimator for the MSE. Chandra et al. (2017) extend this estimator by an additional bias correction that accounts for the sampling variation of the estimator. These methods can be applied only to observed/sampled areas. We apply a method that is suitable even for domains/areas with no observations. To obtain the point estimator on the original scale, properties of the log-normal distribution are used and the back-transformation for the MSE estimator is based on a Taylor expansion similar to that presented in Rao and Molina (2015). A short

derivation can be found in the Appendix (Section 8). The back-transformation is defined as follows:

$$\hat{\theta}_d^{FH.MI} = exp\left\{\hat{\theta}_d^{FH.MI^*} + 0.5\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI^*}\right)\right\},$$

$$\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI}\right) = exp\left\{\hat{\theta}_d^{FH.MI^*} + 0.5\widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI^*}\right)\right\}^2 \widehat{\text{MSE}}\left(\hat{\theta}_d^{FH.MI^*}\right).$$

$\widehat{\text{MSE}}(\hat{\theta}_d^{FH.MI^*})$ denotes at this point the adapted Prasad-Rao MSE estimator defined in Equation (18).

4.3. *Estimator for an arcesine Ratio*

The Fay-Herriot model is widely used for estimating poverty or literacy rates with high regional resolution. In order to guarantee that the estimated rates are between 0 and 1 suitable transformations are frequently used. The arcsine transformation h: $z \rightarrow sin^{-1}(\sqrt{z})$, of which the inverse maps its values to $[0, 1]$, is commonly used. Schmid et al. (2017) compared in a design-based simulation the arcsine transformation with an estimator based on a normal-logistic distribution. Both estimators provided very similar results regarding bias and root mean squared error (RMSE). We concentrate on the arcsine transformation because, unlike the logit, it is well defined even at zero and unity. The arcsine transformation is applied to the direct ratio estimators of the M imputed data sets:

$$\hat{\theta}_{d,m}^{Dir^*} = sin^{-1}\left(\sqrt{\hat{\theta}_{d,m}^{Dir}}\right) \quad \text{with variances} \quad \sigma_{e_{d,m}}^{2^*} = \sigma_{e_d}^{2^*} = \frac{1}{4\tilde{n}_d} \quad \text{for} \quad m = 1, .., M.$$

The effective sample size of domain d is denoted by $\tilde{n}_d$, which takes into account the sampling design effect (Jiang et al. 2001). The approximation of the sampling error variance on the transformed scale is based on a Taylor expansion for moments like in Jiang et al. (2001). The combined point estimator $\hat{\theta}_d^{Dir.RR^*}$ and its variance $\hat{\sigma}_{e_d}^{2^{RR^*}}$ are calculated by applying Rubin's rules presented in Equations (8) and (9). The components of the regression-synthetic part of the model are calculated as described in Subsection 3.1 with the pooled direct components on the transformed scale. Afterwards $\hat{\theta}_d^{FH.MI^*}$ can be calculated as in Equation (12). The resulting estimator $\hat{\theta}_d^{FH.MI^*}$ is on a $sin^{-1}(\sqrt{})$-scale and needs to be transferred to the original scale. A naive back-transformation is the inverse h^{-1}, which introduces a bias for non-linear h. For this reason, for common transformations bias-corrected back-transformations are proposed, such as in Hadam et al. (2023) for the arcsine transformation which is a special case of Sugasawa and Kubokawa (2017), who present an asymptotically unbiased back-transformation for a general parametric transformation. We apply the bias-corrected back-transformation following Hadam et al. (2023), using the normal distribution of the transformed estimator and the expected value (E) of a transformed variable:

$$\hat{\theta}_d^{FH.MI} = \mathrm{E}\left[sin^2\left(\hat{\theta}_d^{FH.MI^*}\right)\right] = \int_{-\infty}^{\infty} sin^2(t) f_{\hat{\theta}_d^{FH.MI^*}}(t)dt$$

$$= \int_{-\infty}^{\infty} sin^2(t) \frac{1}{\sqrt{2\pi \frac{\hat{\sigma}_v^{2RR}\sigma_{e_d}^{2RR^*}}{\hat{\sigma}_v^{2RR} + \sigma_{e_d}^{2RR^*}}}} exp\left\{ - \frac{\left(t - \hat{\theta}_d^{FH.MI^*}\right)^2}{2\frac{\hat{\sigma}_v^{2RR}\sigma_{e_d}^{2RR^*}}{\hat{\sigma}_v^{2RR} + \sigma_{e_d}^{2RR^*}}} \right\} dt. \quad (19)$$

The integral in Equation (19) must be solved by numerical integration methods. The MSE of $\hat{\theta}_d^{FH.MI}$ is approximated with a parametric bootstrap procedure analogue to Hadam et al. (2023) based on Gonzalez-Manteiga et al. (2005). The bootstrap procedure comprises the following steps:

1. Estimate the regression-synthetic components $\hat{\beta}$ and $\hat{\sigma}_v^{2RR}$ analogously to Subsection 3.1 using the pooled direct components $\hat{\theta}_d^{Dir.RR^*}$ and $\hat{\sigma}_{e_d}^{2RR^*}$ on the arcsine scale.
2. For b = 1,..., B
   (a) Generate sampling errors $e_d^{(b)} \overset{ind}{\sim} N\left(0, \hat{\sigma}_{e_d}^{2RR^*}\right)$ and random effects $v_d^{(b)} \overset{ind}{\sim} N\left(0, \hat{\sigma}_v^{2RR}\right)$,
   (b) Simulate a bootstrap sample $\hat{\theta}_d^{Dir^*(b)} = x_d^T\hat{\beta} + v_d^{(b)} + e_d^{(b)}$,
   (c) Calculate the true bootstrap population indicator $\theta_d^{*(b)} = x_d^T\hat{\beta} + v_d^{(b)}$ on the transformed scale and back-transform with $\theta_d^{(b)} = sin^2\left(\theta_d^{*(b)}\right)$,
   (d) Calculate the bootstrap estimator of the model variance $\hat{\sigma}_v^{2(b)}$ using $\hat{\theta}_d^{Dir^*(b)}$ and $\hat{\sigma}_{e_d}^{2RR^*}$,
   (e) Using $\hat{\sigma}_v^{2(b)}$ and $\hat{\theta}_d^{Dir^*(b)}$, calculate bootstrap estimators of the regression coefficients $\hat{\beta}^{(b)}$and estimate the random effects $\hat{v}_d^{(b)}$, and
   (f) Determine the bootstrap estimator $\hat{\theta}_d^{FH.MI^*(b)}$ with Equation (12) by using the estimates from the step before and back-transform to the original scale applying (19) to obtain $\hat{\theta}_d^{FH.MI^*(b)}$.
3. Estimate the MSE:

$$\widehat{\mathrm{MSE}}\left(\hat{\theta}_d^{FH.MI}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_d^{FH.MI(b)} - \theta_d^{(b)}\right)^2.$$

The pooled sampling and model variances, which account for the additional uncertainty about the missing values, are used in the initialization of the bootstrap method. Hence, the extra uncertainty induced by the missing data is accounted for by the bootstrap MSE estimator.

5. Simulation Study

In this section, we investigate the behaviour of the estimators proposed in Sections 3 and 4 by simulation studies with suitable data models. The population is repeatedly generated according to an underlying model. With each simulation run, a sample is taken from the generated population, to which the methods are then applied. We evaluate the performance in terms of bias and RMSE of the proposed point estimators and the inflation of RMSE arising from MI.

5.1. Data Generation

The simulation setup and data models are chosen to be consistent with those of Kreutzmann et al. (2022). For the simulations, finite populations of size $N = 60{,}000$ with $D = 100$ domains are generated so that in each domain the population size N_d is between 200 and 1,000 for $d = 1,..., D$. The samples were drawn via stratified random sampling, where the strata represent the domains. To have rather small and large domains in the samples, sample sizes n_d lie within a range of 8 and 145, so that the total sample size is $n = 5{,}961$. To apply the transformations discussed in the special cases in Section 4, appropriate data models are chosen. In the standard case, a normal data model is used, where no transformation to the direct estimator of a mean value is necessary. Right-skewed log-normal data is generated when investigating the proposed method with a log transformation like in Subsection 4.2. In many applications, the indicator of interest is a ratio. In order to construct a ratio that is used in real data applications, a wealth ratio is calculated. In publications of the Federal Statistical Office (see e.g., Destatis 2018) it is derived by taking the percentage of households with a household income above the 200% median household income. As data model for the ratio the log-scale data is also used. The unit-level data models and scenarios are described in detail in Table 1. The shapes of the distribution for one selected population can be found in Figure 5 in the Appendix (Subsection 8.2). With a sample at the unit-level, the missing data is generated.

As mentioned in Section 1, MAR is often plausible and assumed in most programs for handling missing data. Therefore, in the simulation, missing values are generated using the fully observed additional variable x, from the data models in Table 1. The MAR mechanism is implemented as follows:

$$y_{id} = \begin{cases} y_{\text{missing}}, & x_{id} \leq x_q \\ y_{id}, & \text{otherwise.} \end{cases} \tag{20}$$

x_q is the q-quantile of the auxiliary information x from the sample. This results in a non-response rate of $q \cdot 100\%$ by definition of the q-quantile. For the selected data models, the implemented MAR mechanism leads to missing values in the upper ends of the distribution. When it comes to sensible data as wealth related data, item nonresponse rates can be very high. For example, the Household Finance and Consumption Network (HFCN) reports for 2017 (HFCN 2020a) nonresponse rates for the value of savings account between 18% in Belgium and 64% in Finland. Therefore, it is reasonable to investigate the proposed methods under varying $q \in \{0.1, 0.3, 0.5\}$ to obtain nonresponse rates of 10%, 30% and 50%. A two-level normal model is used as an imputation model for the missing y_{id} values, which is implemented in the R-package `mice` (Van Buuren and Groothius-Oudshoorn 2011). The x serve as covariate information and v_d as area-specific random effects, so that

Table 1. Overview of unit-level data models in model-based simulation, $i = 1,..., N$, $d = 1,..., D$.

Setting	y_{id}	x_{id}	μ_d	v_d	e_{id}
Mean	$250000 - 400x_{\text{id}} + v_d + e_{id}$	$N(\mu_d; 150^2)$	$U[-150, 150]$	$N(0, 25000^2)$	$N(0, 50000^2)$
Logmean	$exp(15 - x_{id} + v_d + e_{id})$	$N(\mu_{\text{d}}; 1)$	$U[3, 5]$	$N(0, 0.4^2)$	$N(0, 0.6^2)$
Ratio	$exp(15 - x_{\text{id}} + v_d + e_{id})$	$N(\mu_{\text{d}}; 1)$	$U[3, 5]$	$N(0, 0.4^2)$	$N(0, 0.6^2)$

the clustering is incorporated in the imputation model. According to Van Buuren (2018), between five and 20 imputed values are often sufficient for each missing observation. The HFCN delivers five imputed values per missing observation, hence in the simulation we set $M = 5$. In the log-scale setting the data was log transformed prior to the imputation to achieve normality and back transformed with the inverse afterwards. After imputation, the data is still on a unit-level and has to be aggregated on an area-level according to the indicator of interest of the setting. Then the appropriate FH.MI estimators given in Section 3 with the special cases in Section 4 are calculated. Table 2 provides an overview showing for each setting the direct estimator, the transformation used, and the section of the corresponding FH.MI model for the special case. In Table 2, I denotes an indicator function that is 1 if the condition is true and 0 otherwise; $\tilde{Y}$ denotes the population median of y.

Each setting, including the generation of the population according to the data model, the sampling, the missing data generating process, the multiple imputation and the application of the MI adjusted FH estimators is repeated $R = 500$ times. The steps of the simulation can be summarized as follows: We generate the population according to a data model in Table 1. Next a stratified random sample is selected. Then missing values are generated according to Equation (20) and imputed to create M copies of the data. Using the M data sets the direct estimators are calculated according to Table 2 and x_{id} are aggregated to a domain level by taking the mean per domain. Afterwards the indicator of interest and its MSE are estimated by applying the methods described in Sections 3 and 4.

5.2. *Performance of Point Estimators*

In the simulation we assess the performance of six point estimators in the *mean* and *log mean* setting and five in the *ratio* setting. For each setting direct, (Direct) and Fay-Herriot (FH) estimators are calculated before deletion on the aggregated sample, that is, the steps of deleting and imputing are omitted. In the case of the FH estimator, the transformation corresponding to the setting is applied so that the Fay-Herriot estimator introduced in Section 2 is calculated. The FH estimator before deletion serves as the gold standard in this simulation. In addition, we compare the performance of the proposed FH.MI estimators with the pooled Fay-Herriot estimator (FH.RR) mentioned in Section 3 and with the estimator proposed by Kreutzmann et al. (2022) denoted by FH.DirectRR. They consider the estimator under a normal and log-normal setting for a mean value, and so we also examine this estimator only under these settings. Furthermore, with Rubin's rule combined direct estimators (Direct.RR) are calculated to show the efficiency gain of the Fay-Herriot estimators with good covariate information after MI. All estimators are implemented in the statistical programming language `R` (R Core Team 2020) and for the

Table 2. Overview of settings.

Setting	$\hat{\theta}_d^{Dir}$	$h\left(\hat{\theta}_d^{Dir}\right)$	FH.MI model
Mean	$\frac{1}{n_d}\sum_{i=1}^{n_d} y_{id}$	$\hat{\theta}_d^{Dir}$	4.1
Log mean	$\frac{1}{n_d}\sum_{i=1}^{n_d} y_{id}$	$log\left(\hat{\theta}_d^{Dir}\right)$	4.2
Ratio	$\frac{1}{n_d}\sum_{i=1}^{n_d} I(y_{id} > 2{\cdot}\tilde{Y})$	$sin^{-1}\left(\sqrt{\hat{\theta}_d^{Dir}}\right)$	4.3

standard area-level models and its components the package `emdi` (Kreutzmann et al. 2019) was used. The code can be obtained from the authors on request. To evaluate and compare the performance of the estimators, the following quality measures are calculated using the R Monte-Carlo replications. $\hat{\theta}_{d_r}$ denotes the estimator of the target indicator in domain d and replication r, θ_{d_r} is the true value of the indicator:

$$\begin{aligned} \text{Bias}(\hat{\theta}_d) &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{d_r} - \theta_{d_r}), \quad \text{rel. Bias}(\hat{\theta}_d) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\theta}_{d_r} - \theta_{d_r}}{\theta_{d_r}}\right), \\ \text{RMSE}(\hat{\theta}_d) &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{d_r} - \theta_{d_r})^2}, \quad \text{RRMSE}(\hat{\theta}_d) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\theta}_{d_r} - \theta_{d_r}}{\theta_{d_r}}\right)^2}. \end{aligned} \tag{21}$$

We want to evaluate the performance of the introduced methodology in terms of bias and RMSE. For the *mean* and *log mean* setting we consider the relative bias and the RRMSE. For the *ratio* setting the bias and RMSE are taken into account since the indicator itself is already on a relative scale. The median and mean values over domains of the bias and RMSE values for different nonresponse rates are presented in Table 3. The direct estimators (Direct.RR) remain unbiased after multiple imputation in the *mean* and *ratio* setting as before deletion (Direct) and almost unbiased in the *log mean* setting. The small bias could be introduced by the inverse back-transformation after applying the imputation model. Compared to the combined direct estimators (Direct.RR) and the model-based estimators before deletion (FH), the model-based estimators FH.MI, FH.RR and FH.DirectRR remain also unbiased in the *mean* and *ratio* setting and the results of the model-based estimators are comparable. Only in the *log mean* setting does the FH.MI estimator, like the other two model-based estimators, suffer from a small bias that increases slightly with higher nonresponse rates. Again this bias could be due to the inverse back-transformation in the imputation process. In terms of efficiency, we see that the RRMSE/RMSE are the smallest before deletion and increase with higher nonresponse rates for each estimator in each setting, reflecting the additional uncertainty about missing values. Within each setting and nonresponse rate the order of the RRMSE/RMSE is as expected: the RRMSE/RMSE of the direct estimators is always higher than that of the proposed FH.MI estimator, which shows that the introduced methodology behaves the same way as in cases without missing values (i.e., before deletion). The RRMSE/RMSE of the FH.MI and the FH.RR are almost identical, which indicates that the proposed methodology leads to reasonable results and is similar to the more straightforward approach of combining the Fay-Herriot estimators. The proposed FH.MI estimator is at least as efficient as the FH.DirectRR estimator. In the *log mean* setting, the *superefficiency* of imputation, when more information is used than in the analysis model (Rubin 1996), can be observed. At a nonresponse rate of 10%, Direct.RR is slightly more efficient than the direct estimator before deletion (Direct). All summed up, the results confirm our expectations. The presented FH.MI estimators lead to plausible results regarding bias and efficiency in the investigated settings, in which the imputation models follow the data structure of the generated population and thus fit the data.

Table 3. Relative bias and RRMSE for mean and log mean, bias and RMSE for ratio.

Nonresponse rate		Before deletion		10%		30%		50%	
	Estimator	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Mean									
(rel.) Bias [%]	Direct	0.0464	0.0149						
	Direct.RR			0.0254	0.0198	0.0390	0.0092	0.0862	0.0290
	FH	0.2390	0.1812						
	FH.Direct.RR			0.2291	0.1691	0.2536	0.1872	0.3082	0.2583
	FH.MI			0.2245	0.1615	0.2355	0.1761	0.2704	0.2186
	FH.RR			0.2171	0.1568	0.2195	0.1639	0.2554	0.1840
Log mean									
(rel.) Bias [%]	Direct	-0.2191	-0.0318						
	Direct.RR			0.1548	0.0342	1.1479	0.8566	2.8100	2.2903
	FH	-0.8797	-0.6057						
	FH.Direct.RR			0.0191	0.2091	1.4284	1.4639	3.1864	2.9609
	FH.MI			-0.2772	-0.1096	0.8383	0.8272	2.4169	2.3568
	FH.RR			-0.6948	-0.4258	0.0216	0.2115	1.3747	1.4485
Ratio									
Bias	Direct	-0.0004	0.0000						
	Direct RR			-0.0003	0.0000	-0.0000	0.0005	0.0009	0.0007
	FH	-0.0027	-0.0022						
	FH.MI			-0.0016	-0.0010	0.0012	0.0012	0.0011	0.0016
	FH.RR			-0.0026	-0.0021	-0.0024	-0.0018	-0.0015	-0.0009
Mean									
RRMSE [%]	Direct	5.0318	4.2722						
	Direct.RR			5.1345	4.4849	5.5337	4.7889	6.1003	5.4419
	FH	4.4300	3.9609						
	FH.Direct.RR			4.5470	4.1570	4.9845	4.5694	5.6775	5.3471
	FH.MI			4.5444	4.1524	4.9643	4.5509	5.6018	5.2498
	FH.RR			4.5386	4.1385	4.9517	4.5388	5.5741	5.1978
Log mean									
RRMSE [%]	Direct	25.5219	23.0001						
	Direct.RR			24.8037	22.0991	26.3014	23.1315	29.1076	26.1128
	FH	20.7739	20.0316						
	FH.Direct.RR			21.916	21.4101	23.8789	22.5175	27.0243	25.9548
	FH.MI			21.3353	20.6552	22.7919	21.3174	25.4294	23.9328
	FH.RR			20.7741	19.9455	22.1078	20.6367	24.7177	23.3957
Ratio									
RMSE	Direct	0.0655	0.0563						
	Direct.RR			0.0655	0.0565	0.0663	0.0565	0.0702	0.0617
	FH	0.0539	0.0506						
	FH.MI			0.0544	0.0510	0.0572	0.0533	0.0636	0.0607
	FH.RR			0.0541	0.0507	0.0564	0.0524	0.0624	0.0590

5.3. Performance of Uncertainty Measures

We now move on to the performance of the three proposed MSE estimators of the FH.MI estimator, each corresponding to one setting. In the case of the *mean* and *log mean* setting, we evaluate the adapted analytical Prasad-Rao estimator as described in Subsections 4.1 and 4.2 with a back-transformation when the log transformation is used. In the *ratio* setting the parametric bootstrap estimator from Subsection 4.3 with $B = 500$ replications is evaluated. Performance is evaluated by looking at the relative bias of the MSE estimator defined as followed:

$$\text{RBRMSE}(\hat{\theta}_d) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{\text{MSE}}_{dr} - RMSE(\hat{\theta}_d)}}{\text{RMSE}(\hat{\theta}_d)}.$$

Table 4 shows the median and mean values over the domains of the RBRMSE. We see a slight underestimation in the *mean* setting with an increasing effect at higher nonresponse rates. On the other hand, in the *log mean* setting the true RMSE is slightly overestimated at a lower nonresponse rate of 10% and minimally underestimated at a higher nonresponse rate of 50%. Nevertheless, the values are all close to zero. In the *ratio* setting, the bias of the bootstrap RMSE estimator is close to zero at 10% nonresponse rate. At 30% and 50% it increases and reaches almost identical values, but still at a tolerable level. In all three settings the additional uncertainty of the FH.MI estimator can be satisfactorily addressed and the bias is within an acceptable range. To have a closer look on the performance of the adapted Prasad-Rao MSE estimator the estimated and true RMSE values per domain are plotted in Figure 1 for the *mean* setting. First we observe that within each nonresponse rate the estimated RMSE decreases with higher sample size, which is in line with the behaviour of the true RMSE. Secondly, we see that per domain the estimated RMSE values increase with increasing nonresponse rates, which is consistent with the expected behaviour. At a nonresponse rate of 10% and 30%, the estimated RMSE tracks very well the behaviour of the true RMSE. With a higher nonresponse rate of 50% we see that there are underestimations in some areas, but overall the uncertainty is well accounted for. The proposed methods are good at capturing the additional variation due to the missing observations and imputation and also provide a realistic estimate of the uncertainty of the FH.MI estimator in our settings.

Table 4. Relative bias (%) of estimated RMSE (RBRMSE) of FH.MI.

Nonresponse rate	10%		30%		50%	
	Mean	Median	Mean	Median	Mean	Median
Mean	-1.4198	-1.8291	-3.4427	-3.1477	-6.9352	-6.9390
Log mean	2.5119	2.5719	1.8185	2.6527	-4.0788	-3.2214
Ratio	2.9396	3.1787	8.7815	9.0866	8.1231	8.278

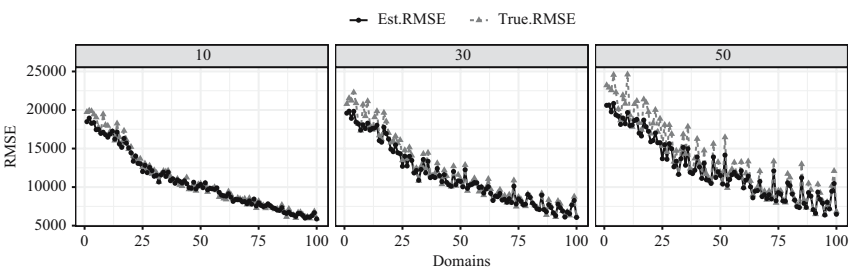


Fig. 1. RMSE of FH.MI estimator per domain for mean setting and varying nonresponse rates. Domains are ordered by increasing sample size.

6. Application to Eurosystem's HFCS

In the following, we provide an example of how the proposed framework can be used for surveys with multiply imputed data in combination with small area methods. The purpose is to show a possible application with the HFCS data for scientists or institutions from relevant research areas rather than to discuss the estimates for each country. The HFCS is a large-scale survey of the financial and consumption situation of European households. The first wave was carried out in 2010 in 15 countries of the European Union (EU). The HFCS contains household data on both economic and demographic variables such as income, wealth, private pension, employment and consumption characteristics (HFCN 2020a). So far three waves have been carried out, the last of which was collected in 2017 and released in March 2020. For the application the third wave is considered. The sample contains about 91,200 households in 22 countries of the EU, between 1,000 and 14,000 households per country. The HFCS is a joint project of several national statistical institutes, Eurosystem national central banks (NCB) and three noneuro area NCBs (Poland, Hungary, Croatia). For these countries, all values are converted into euros by the HFCN (HFCN 2020a). The HFCN asked very sensitive questions, so the item nonresponse rate is high. Missing values in the HFCS data were iteratively and sequentially imputed. The variables are imputed along a path of imputation models. Each model is run several times, and the imputed values from the previous round are treated as given in the subsequent iteration (HFCN 2020a). For each missing observation the HFCS data set contains $M = 5$ imputed values. For more information on the imputation method see HFCN (2020a). Of interest for this application is the value of the household's bonds, which is part of the household's assets and therefore relevant when considering the distribution of wealth. The HFCN reports conditional medians for the value of bonds per EU country (HFCN 2020b). The values are calculated conditioned on households that have bonds; households with no bonds are discarded from the analysis. This results in partly very small sample sizes even on a country level, so that for some countries with fewer than 25 observations direct estimates are not reported by the HFCN. Furthermore, the rate of collected values differs between the countries. Since some households do not even indicate whether they own bonds or not, these values are also imputed by the HFCN. Therefore, the sample size per country, that is, the number of households with bonds and the collected rate for these households, may differ slightly among the five imputed data sets provided by the HFCN. We calculate the sample sizes and collection rates based on the first imputed data sets. An overview of the sample sizes per country and the collected rates are given in Table 5. As dependent variable we choose the

Table 5. Summary of EU-countries sample sizes, collected rates and auxiliary variables.

	Min	1stQ	Median	Mean	3rdQ	Max
Sample size	2.00	12.25	61.50	148.73	209.50	832.00
Collected rate	0.04	0.49	0.66	0.61	0.81	1.00
Total receipts from taxes and social contributions (% of GDP)	23.20	32.83	36.90	36.84	41.85	48.10
Final consumption expenditure (Current prices, EUR per capita)	5630	11710	17170	20424	29258	48140

mean value of bonds in thousand of euros (TEUR) on a country level, resulting in $D = 22$ domains. In 2017 the EU consisted of 28 member states. Six EU members are not included in the HFCS as their noneuro area NCBs do not participate. These domains are considered as out-of-sample (OOS) and model-based estimates are provided in the application. The direct estimators of the mean value of bonds for each imputed data set $\hat{\theta}^{Dir}_{d,m}, d = 1,...,22, m = 1,...,5$ are calculated according to Equation (13) using the sampling weights provided by the data provider, which corrects for potential bias due the sampling design and unit nonresponse. The variances $\sigma^2_{e_{d,m}}$ are estimated with a bootstrap method following the instructions by HFCN (2020a) using the provided replicate weights derived by the Rao-Wu rescaled bootstrap method. As a result we obtain $M = 5$ replicates of direct estimators and their variances, which are then pooled according to Sections 3 and 4.

6.1. *Model Selection and Validation*

To obtain auxiliary information from additional sources needed for the Fay-Herriot models, country-level data were collected from Eurostat, the statistical office of the EU and the European Commission. Within this set, data such as real estate data, unemployment rates, age dependency ratios, national accounts and tax aggregates from 2011 and 2017 were collected. The sources and years of this supplemental information are shown in Table 6 in the Appendix (Subsection 8.2). Due to the small number of domains, variables that were not available for the entire set of domains were excluded. The remaining auxiliary information includes variables such as the old, youth and age dependency ratio, the unemployment rate, the ratio of taxes to GDP, final consumption expenditure, the share of consumption expenditure on GDP, GDP at market prices and a variable indicating whether the country has a wealth tax. In addition, the number of covariates in the model is severely limited by the small number of domains, which is why we restricted the model to two possible auxiliary variables. In the context of area-level data, Han (2013) transferred the conditional Akaike information in linear mixed models from Vaida and Blanchard (2005) to a conditional Akaike information criterion for Fay-Herriot models. Marhuenda et al. (2014) examine this criterion among Kullback symmetric divergence criterion (KIC) and propose a bootstrap variant of the KIC (KICb2) especially developed for FH models. They conclude that KICb2 criterion is one of the best model selection criteria for Fay-Herriot models. Therefore, in this application the preselection of variables was performed using the KICb2 criterion. Model selection was carried out for each of the five imputed data sets, with no particular difference in the results. A union of two auxiliary variables was selected for the final model, as shown in Table 5. To obtain a model-based estimator of the mean value of household bonds, the estimator from Subsection 4.1 is calculated with the auxiliary information in Table 5. The model variances $\sigma^{2^{RR}}_v$ are calculated for the MI-adjusted Fay-Herriot model on the original scale using the REML method. The distributional assumptions of the model presented in Section 3 are checked by the Shapiro-Wilk test applied to the residuals and the random effects. For the MI-adjusted Fay-Herriot model for a normal mean, the p-values of the tests for the standardized residuals and the random effect are 0.223 and 0.965, respectively. Therefore, the normality assumptions for both error terms cannot be rejected at a 5% significance level. Consequently, all further considerations and results are based on the

MI-adjusted Fay-Herriot model for a mean value as presented in Subsection 4.1. The explanatory power of the model is assessed using the modified R^2 for Fay-Herriot models according to Lahiri and Suntornchost (2015) and we obtain a value of 45%. Due to the low number of domains, it is not possible to include more auxiliary variables to potentially increase explanatory power. We obtain positive estimated regression coefficients for both auxiliary variables. The impact on the tax-to-GDP ratio seems reasonable, given that tax contributions include taxes on wealth (at least in some countries) and that high tax revenues from income could indicate a high level of capital assets. The relationship between consumption and wealth is not independent of income, because if income is higher than consumption, the rest can be invested, and if consumption cannot be covered by income, there is nothing left to invest. Nevertheless, with the given data, the model also shows a positive effect for consumption.

6.2. Small Area Estimates

The estimates of the mean value of bonds on a country level are calculated using the FH.MI estimator for a mean value and to estimate the MSE the MI adapted Prasad-Rao estimator is applied as described in Subsection 4.1. To compare the model-based estimators with a direct estimator, the direct estimators and their variance estimates are computed for each imputed data set as described above and pooled using Rubin's rule in Equation (8) (Direct.RR). The point estimates of the model-based estimators (FH.MI) should be consistent with the unbiased estimates of the direct estimator, but be more precise. Figure 2 compares the direct and the model-based point estimates for the 22 in-sample domains and additionally reports the estimates for the six OOS EU countries. Due to the guidelines of the data provider, the direct estimates for domains with less than 25 observations are not reported. We observe that, for countries with large sample sizes, the direct and model-based estimates are almost identical, consistent with the expectation that high weight is given to the direct estimator when precision is high. An exception is Belgium (BE), where the sample size is rather high, but the shrinkage to the mean quite strong. For most of the direct estimates, which tend to be high, we see that the model-based estimates are smaller, showing the shrinkage effect to the mean of the model-based estimates. (see summary statistics of point estimates in Table 7 in the Appendix (Subsection 8.2)). Possibly due to the low number of covariates very little shrinkage takes place for some countries with small sample sizes (GR, SI, LI). The model-based point

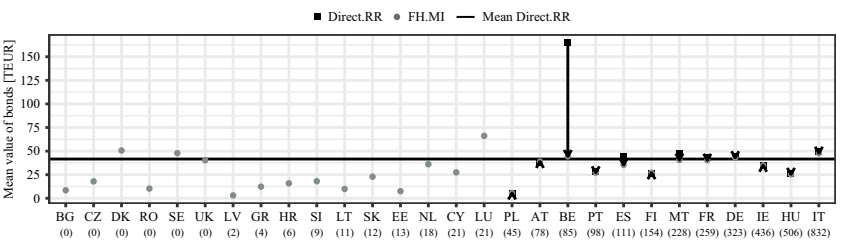


Fig. 2. Direct and model-based estimates for the mean value of bonds, own estimations. Domains are ordered by increasing sample size, sample sizes in brackets. Direct estimates for domains with less than 25 observations are not reported.

estimators are furthermore reported in the map in Figure 3. The highest values are estimated for Luxembourg (LU), followed by Denmark (DK) (OOS) and Sweden (SE) (OOS). For eastern European countries, the estimates are rather low, followed by southern European countries. The estimated model-based values range from EUR 3,000 to EUR 66,000 (cf. Table 7 in the Appendix (Subsection 8.2)), which seems plausible given the median values reported by the HFCN (HFCN 2020b) between EUR 2,000 and EUR 25,000, considering that the distribution at the household level tends to be right skewed and therefore the mean values should be higher than the median values. Figure 4 shows the coefficients of variation (CV) for the direct and model-based estimates. We see that the model-based estimator is at least as efficient as the direct estimator. The CVs of the model-based estimators are mostly significantly smaller than those of the direct estimators, with the effect decreasing with increasing sample size. For large sample sizes, the gain is barely noticeable, but this is consistent with the expected behavior that the direct estimator is sufficiently accurate in this case. For some domains, such as Croatia (HR) and Cyprus (CY), the CV is almost halved. Due to the relatively small domain size of $D = 22$ and hence the limitation to the number of covariates in the model, the efficiency gain is limited.

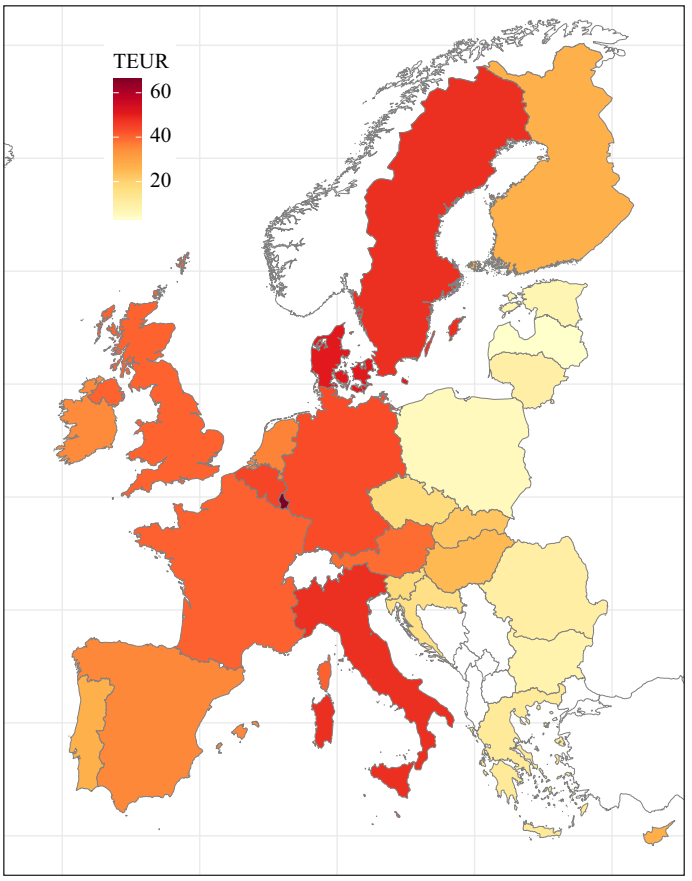


Fig. 3. Map of model-based FH.MI estimates for mean value of bonds, own estimations. Non-EU countries in 2017 are colored in white.

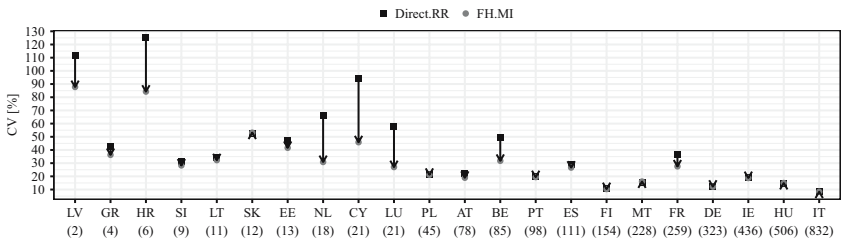


Fig. 4. CVs of direct and model-based estimates, own estimations. Domains are ordered by increasing sample size, sample sizes in brackets.

A summary of the distribution of the point estimators and CVs from Figures 2 and 4 can be found in Table 7 in the Appendix (Subsection 8.2).

7. Concluding Remarks

In this article, we derive small area indicators based on multiply imputed survey data and present uncertainty measures for common cases that capture the additional uncertainty. We present the transformed Fay-Herriot model calculated on each imputed data set. We then combine the components into a MI adjusted Fay-Herriot model that retains the model structure of the Fay-Herriot model. With this approach, results that exist for the Fay-Herriot model regarding transformations, back-transformations and MSE estimators can be extended. It is a general approach that can be applied to any indicator with a given transformation and an appropriate back-transformation. We discuss common special cases of the model (mean, log mean, arcsine ratio). For these special cases we propose MSE estimators. For the mean and logarithmic mean, we present an analytical adaption of the Prasad-Rao estimator and, for the arcsine ratio, we use a bootstrap estimator. We demonstrate in simulation studies that the resulting FH.MI point estimators lead to valid results in terms of bias and RMSE in the given settings and under different nonresponse rates and that the proposed MSE estimators are able to capture the additional imputation uncertainty and lead to good uncertainty measures. We carried out an application using the proposed framework to obtain estimates for European household assets.

A limitation of the proposed approach is that it is not as straightforward for the user as it would be if only the Fay-Herriot estimators were estimated for each imputed data set and the mean value calculated. But, as mentioned above, it is not clear how the variance pooling rules can be applied to the MSE. This could be part of further research. To facilitate the application, it is planned to provide an R-package with the methodology presented. Other open research questions are the extension from a cross-sectional to a longitudinal analysis to provide stable estimates across panel waves (i.e., over time) when multiple imputations are performed and sample sizes are small. If the underlying data structure is a panel survey and individuals or households are observed over multiple time periods, the Fay-Herriot model can be adapted to consider the correlation of the same observations over time. To borrow strength for domain estimates, Rao and Yu (1994) propose a model with auto-correlated random effects and assume an autoregressive process of first order. In addition to the temporal Fay-Herriot models, a multivariate approach could serve the requirement to consider the temporal dimension in the data.

In the multivariate Fay-Herriot model (Benavent and Morales 2016) the domain indicators are estimated simultaneously for the different panel waves. In this way, correlations for both error terms can be considered. These models have not yet been investigated in combination with multiple imputation. The approach in this article could be extended to include correlations over time to ensure reliable estimates over time based on multiply imputed survey data. Since asset values are usually highly skewed, more robust indicators such as the median or other quantiles could be estimated instead of the mean. Therefore, the estimation of small area medians using the Fay-Herriot model would be interesting for future research.

8. Appendix

8.1. MSE Back-Transformation for a Log Mean

Let $\mu = exp(\theta)$ be the true indicator value and $\hat{\theta}$ be an estimate for θ. Furthermore, $\hat{\mu}$ is an estimator for μ with $\hat{\mu} = g(\hat{\theta})$, where g is a continuously differentiable function. For

$$g(\hat{\theta}) = exp\{\hat{\theta} + 0.5\widehat{\text{MSE}}(\hat{\theta})\}$$

an approximation of $\text{MSE}(\hat{\mu})$ using a Taylor expansion can be derived as follows:

$$\begin{aligned}
\text{MSE}(g(\hat{\theta})) &= \text{Var}(g(\hat{\theta})) + \text{Bias}^2(g(\hat{\theta})) \\
&= \text{E}[g(\hat{\theta})^2] - \text{E}[g(\hat{\theta})]^2 + \text{E}[g(\hat{\theta}) - g(\theta)]^2 \\
&\approx \text{E}[\{g(\theta) + g'(\theta)(\hat{\theta} - \theta)\}^2] - \text{E}[\{g(\theta) + g'(\theta)(\hat{\theta} - \theta)\}]^2 + \text{E}[g'(\theta)(\hat{\theta} - \theta)]^2 \\
&= g'(\theta)^2\{\text{E}[\hat{\theta}^2] - \text{E}[\hat{\theta}]^2\} + g'(\theta)^2\text{E}[\hat{\theta} - \theta]^2 \\
&= g'(\theta)^2\{\text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})\} = g'(\theta)^2\text{MSE}(\hat{\theta}).
\end{aligned}$$

A estimator of $\text{MSE}(\hat{\mu})$ is then obtained by

$$\widehat{\text{MSE}}(\hat{\mu}) = \widehat{\text{MSE}}(g(\hat{\theta})) = g'(\hat{\theta})^2\widehat{\text{MSE}}(\hat{\theta}) = exp\{\hat{\theta} + 0.5\widehat{\text{MSE}}(\hat{\theta})\}^2\widehat{\text{MSE}}(\hat{\theta})\}.$$

8.2. Plots and Tables (Figure 5 and Tables 6–7)

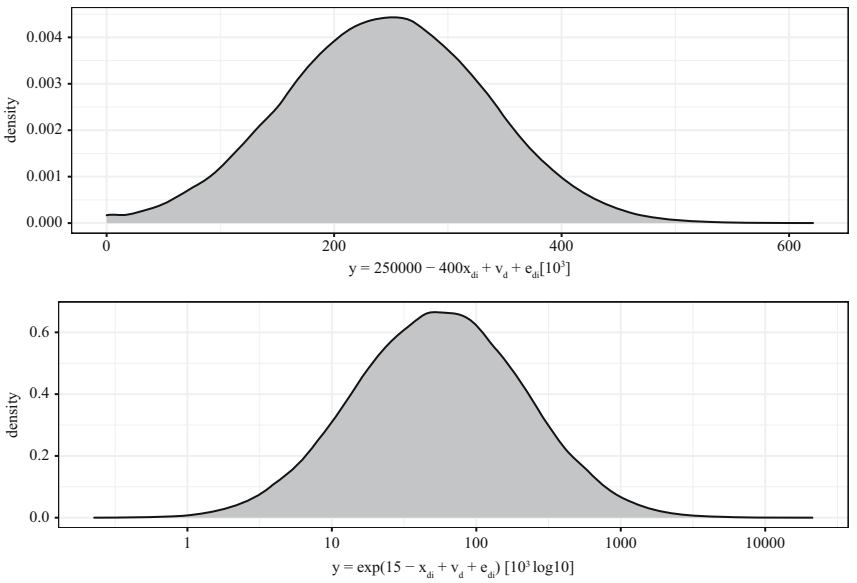


Fig. 5. Density of population target variable of one replication.

Table 6. Source and year of auxiliary information.

	Year	Source
Private households by type, tenure status (Real estate)	2011	Eurostat (2011b)
Dwellings by occupancy status, type of building (Real estate)	2011	Eurostat (2011a)
Age, Old, Young-age dependency ratios	2017	Eurostat (2017d)
Unemployment rate	2017	Eurostat (2017a)
Tax to GDP ratio	2017	Eurostat (2017c)
Final consumption expenditure	2017	Eurostat (2017b)
GDP at market prices	2017	Eurostat (2017b)
Share of consumption expenditure on GDP	2017	Eurostat (2017b)
Indicator for presence of wealth tax	2017	European Commission (2017)

Table 7. Summary of point estimators and CVs for mean value of bonds (TEUR).

Estimator		Min	1stQ	Median	Mean	3rdQ	Max
Direct.RR	Point est.	2.5	19.6	36.2	41.6	49.0	165.5
FH.MI		3.1	16.4	27.3	28.6	40.0	66.2
Direct.RR	CV [%]	8.3	19.3	32.8	41.9	51.7	125.0
FH.MI		8.3	18.8	27.3	31.5	35.3	87.8

9. References

Battese, G., Harter, R., and W. Fuller. 1988. "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of The American Statistical Association* 83: 28–36. DOI: https://doi.org/10.1080/01621459.1988.10478561.

Benavent, R., and D. Morales. 2016. " Multivariate Fay Herriot models for small area estimation." *Computational Statistics and Data Analysis* 94: 372–390. DOI: https://doi.org/10.1016/j.csda.2015.07.013.

Bijlsma, I., Brakel, J., R. van der Velden, and J. Allen. 2020. "Literacy Levels at a Detailed Regional Level: An Application Using Dutch Data." *Journal of Official Statistics* 36: 251–274. DOI: https://doi.org/10.10.2478/jos-2020-0014.

Casas-Cordero, C., J. Encina, and P. Lahiri. 2016. *Poverty Mapping for the Chilean Comunas*: 379–404. DOI: https://doi.org/10.10.1002/9781118814963.ch20.

Chandra, H., K. Aditya, and S. Kumar. 2017. "Small area estimation under a log transformed area level model." *Journal of Statistical Theory and Practice* 12: 497–505. DOI: https://doi.org/10.10.1080/15598608.2017.1415174.

Citro, C.F., and G. Kalton. 2000. *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. Washington, DC: The National Academies Press. DOI: https://doi.org/10.10.17226/10046.

Destatis. 2018. *Wirtschaftsrechnungen Einkommens-und Verbrauchsstichsprobe Einkommensverteilung in Deutschland 2013*. Statistisches Bundesamt (Destatis). Available at: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Einkommen-Einnahmen-Ausgaben/Publikationen/Downloads-Einkommen/einkommensverteilung-2152606139004.pdf?__blob=publicationFile.

European Commission. 2017. *Taxation and customs, taxes in europe database v3*. Available at: https://ec.europa.eu/taxation_customs/tedb/taxSearch.html (accessed February 2023).

Eurostat. 2011a. *Conventional dwellings by occupancy status, type of building and nuts 3 region*. Available at: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = cens_11dwob_r3&lang = en (accessed April 2021).

Eurostat. 2011b. *Private households by type, tenure status and nuts 2 region*. Available at: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = cens_11htts_r2&lang = en (accessed April 2021).

Eurostat. 2017a. *Harmonised unemployment rates*. Available at: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = ei_lmhr_m&lang = en (accessed April 2021).

Eurostat. 2017b. *Main gdp aggregates per capita*. Available at: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = nama_10_pc&lang = en (accessed April 2021).

Eurostat. 2017c. *Main national accounts tax aggregates*. Available at: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = gov_10a_taxag&lang = en (accessed April 2021).

Eurostat. 2017d. *Population: Structure indicators*. Available at: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset = demo_pjanind (accessed April 2021).

Fay, R.E., and R.A. 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: https://doi.org/10.10.2307/2286322.

Fisher, J. 2006. *Income imputation and the analysis of expenditure data in the consumer expenditure survey*. U.S. Bureau of Labor Statistics, Working Articles. Available at: https://www.bls.gov/opub/mlr/2006/11/art2full.pdf (accessed October 2020).

Gonzalez-Manteiga, W., Lombardia, M.J., I. Molina, D. Morales, and L. Santamaria. 2005. "Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model." *Computational Statistics and Data Analysis* 52: 5242–5252. DOI: https://doi.org/10.10.1016/j.csda.2008.04.031.

Hadam, S., N. Würz., A.-K. Kreutzmann, and T. Schmid. 2023. "Estimating regional unemployment with mobile network data for functional urban areas in Germany." *Statistical Methods & Applications* DOI: https://doi.org/10.1007/s10260-023-00722-0.

Han, B. 2013. "Conditional Akaike information criterion in the Fay-Herriot model." *Statistical Methodology* 11: 53–67. DOI: https://doi.org/10.10.1016/j.stamet.2012.09.002.

HFCN. 2020a. *The household finance and consumption survey: Methodological report for the 2017 wave*. Household Finance and Consumption Network. Available at: https://www.ecb.europa.eu/pub/pdf/scpsps/ecb.sps35~b9b07dc66d.en.pdf (accessed April 2020).

HFCN. 2020b. *The household finance and consumption survey wave 2017 statistical tables*. Household Finance and Consumption Network. Available at: https://www.ecb.europa.eu/home/pdf/research/hfcn/HFCS_Statistical_Tables_Wave_2017_May2021.pdf?ca15e575b6b7765dad1147e7a3dba728 (accessed April 2020).

Jiang, J., Lahiri, P., S.-M. Wan, and C.-H. Wu. 2001. "Jackknifing in the Fay-Herriot model with an example. "In Proceedings og the seminar." Funding Opportunity in Survey Research, Washington D.C., U.S. Bureau of Labor Statistics. Available at: https://citeseerx.ist.psu.edu/document?repid = rep1&type = pdf&doi = 70c03a8b572a00931735409d7173b087645006b9 (accessed April 2020).

Jiang, J., and J. Rao. 2020. "Robust small area estimation: An overview." *Annual Review of Statistics and Its Application* 7: 337–360. DOI: https://doi.org/10.10.1146/annurev-statistics-031219-041212.

Kreutzmann, A.-K., P. Marek, M. Runge, N. Salvati, and T. Schmid. 2022. "The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany." *Journal of Applied Statistics* 49: 3278–3299. DOI: https://doi.org/10.10.1080/02664763.2021.1941805.

Kreutzmann, A.-K., Pannier, S., N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. 2019. "The R package emdi for estimating and mapping regionally disaggregated indicators." *Journal of Statistical Software* 91: 1–33. DOI: https://doi.org/10.10.18637/jss.v091.i07.

Lahiri, P. and J.B. Suntornchost. 2015. "Variable selection for linear mixed models with applications in small area estimation." *The Indian Journal of Statistics* 77: 312–320. DOI: https://doi.org/10.10.1007/s13571-015-0096-0.

Longford, N. 2004. "Missing data and small area estimation in the uk labour force survey." *Journal of the Royal Statistical Society* 167: 341–373. DOI: https://doi.org/10.10.1046/j.1467-985X.2003.00728.x.

Longford, N.T. 2005. *Missing data and small-area estimation*. Springer, London.

Maiti, T., H. Ren, and S. Sinha. 2014. "Prediction error of small area predictors shrinking both means and variances." *Scandinavian Journal of Statistics* 41: 775–790. DOI: https://doi.org/10.10.1111/sjos.12061.

Marhuenda, Y., D. Morales, and M. Pardo. 2014."Information criteria for Fay-Herriot model selection." *Computational Statistics and Data Analysis* 70: 268–280. DOI: https://doi.org/10.10.1016/j.csda.2013.09.016.

Neves, A., D. Silva, and S. Correa. 2013. "Small domain estimation for the brazilian service sector survey." *Estadistica* 65: 13–37. Available at: https://www.statistics.gov.hk/wsc/CPS003-P7-S.pdf (accessed November 2019).

Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28: 40–68. DOI: https://doi.org/10.10.1214/12-STS395.

Prasad, N., and J. Rao. 1990. "The estimation of the mean squared error of small-area estimators." *Journal of the American Statistical Association* 85: 163–171. DOI: https://doi.org/10.10.2307/2289539.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Rao, J., and M. Yu. 1994. "Small area estimation by combining time series and cross sectional data." *Canadian Journal of Statistics* 22: 511–528. DOI: https://doi.org/10.10.2307/3315407.

Rao, J.N.K. and Molina. 2015. *Small Area Estimation*. John Wiley and Sons, Hoboken.

Riphahn, R. and O. Serfling. 2005. "Item non-response on income and wealth questions." *Empirical Economics* 30: 521–538. DOI: https://doi.org/10.10.1007/s00181-005-0247-7.

Rivest, L.-P., and N. Vandal. 2002. "Mean squared error estimation for small areas when the small area variances are estimated." In Proceedings of International Conference of Recent Advanced Survey Sampling, edited by J. Rao, July 10–13, Ottawa, Canada: 197–206. Available at: https://www.mat.ulaval.ca/fileadmin/mat/documents/lrivest/-Publications/64-RivestVandal2003.pdf (accessed October 2020).

Rubin, D.B. 1976. "Inference and missing data." *Biometrika* 63: 163–171. DOI: https://doi.org/10.10.2307/2335739.

Rubin, D.B. 1987. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, Hoboken.

Rubin, D.B. 1996. "Multiple imputation after 18 + years." *Journal of the American Statistical Society* 91: 473–489. DOI: https://doi.org/10.10.2307/2291635.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski. 2017. "Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal." *Journal of the Royal Statistical Society* 180: 1163–1190. DOI: https://doi.org/10.10.1111/rssa.12305.

Slud, E., and T. Maiti. 2006. "Mean-squared error estimation in transformed Fay-Herriot models." *Journal of the Royal Statistical Society* 68: 239–257. DOI: https://doi.org/10.10.1111/j.1467-9868.2006.00542.x.

Sugasawa, S., and T. Kubokawa. 2017. "Transforming response values in small area prediction." *Computational Statistics and Data Analysis* 114: 47–60. DOI: https://doi.org/10.10.1016/j.csda.2017.03.017.

Sugasawa, S., H. Tamae, and T. Kubokawa. 2017. "Bayesian estimators for small area models shrinking both means and variances." *Scandinavian Journal of Statistics* 44: 150–167. DOI: https://doi.org/10.10.1111/sjos.12246.

Sverchkov, M., and D. Pfaeffermann. 2018. "Small area estimation under informative sampling and not missing at random non-response." *Journal of the Royal Statistical Society*: 981–1008. DOI: https://doi.org/10.10.1111/rssa.12362.

Tzavidis, N., l.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla. 2018. "From start to finish: a framework for the production of small area official statistics." *Journal of the Royal Statistical Society* 181: 927–979. DOI: https://doi.org/10.10.1111/rssa.12364.

Vaida, F., and S. Blanchard. 2005. "Conditional akaike information for mixed-effects models." *Biometrika* 92: 351–370.

Van Buuren, S. 2018. *Flexible Imputation of Missing Data*. Second Edition. Chapman and Hall/CRC.

Van Buuren, S., and K. Groothius-Oudshoorn. 2011. "mice: Multivariate imputation by chained equation in R." *Journal of Statistical Software* 45: 1–67. DOI: https://doi.org/10.10.18637/jss.v045.i03.

Wang, J., and W.A. Fuller. 2003. "The mean squared error of small area predictors constructed with estimated area variances." *Journal of the American Statistical Association* 98: 716–723. DOI: https://doi.org/10.10.1198/016214503000000620.

You, Y., and B. Chapman. 2006. "Small area estimation using area level models and estimated sampling variances." *Survey Methodology* 32: 97–103. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263 (accessed March 2020).

Received November 2022
Revised March 2023
Accepted June 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 535–570, http://dx.doi.org/10.2478/JOS-2023-0025

Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics

Milena Suarez Castillo[1], *Francois Sémécurbe*[1], *Cezary Ziemlicki*[2], *Haixuan Xavier Tao*[1], *and Tom Seimandi*[1]

Mobile network data records are promising for measuring temporal changes in present populations. This promise has been boosted since high-frequency passively-collected signaling data became available. Its temporal event rate is considerably higher than that of Call Detail Records – on which most of the previous literature is based. Yet, we show it remains a challenge to produce statistics consistent over time, robust to changes in the "measuring instruments" and conveying spatial uncertainty to the end user. In this article, we propose a methodology to estimate – consistently over several months – hourly population presence over France based on signaling data spatially merged with fine-grained official population counts. We draw particular attention to consistency at several spatial scales and over time and to spatial mapping reflecting spatial accuracy. We compare the results with external references and discuss the challenges which remain. We argue data fusion approaches between fine-grained official statistics data sets and mobile network data, spatially merged to preserve privacy, are promising for future methodologies.

Key words: Big data; High frequency statistics; dynamic population mapping; spatial accuracy.

1. Introduction

Mobile network data has the potential to significantly enhance population statistics by increasing their levels of spatiotemporal details and their timeliness. Several international initiatives aimed at incorporating this new data source into the production of official statistics, such as the United Nations Big Data dedicated Task Team (U.N. Global Working Group 2019), the European Statistical System working groups (ESSnet Big Data 2021), in addition to national initiatives (Statistics Netherlands 2020; Coudin et al. 2021). Yet, despite the new interest in timely measuring dynamic population during the Covid crisis, producing statistics from mobile network data has remained a great challenge for

[1] INSEE, SSP Lab, 88 avenue Verdier, Montrouge, Île-de-France, 92120, France. Email: milena.suarez-castillo@sante.gouv.fr, francois.semecurbe@agriculture.gouv.fr, tao.xavier@outlook.com,and tom.seimandi@insee.fr

[2] Orange Labs R&D Châtillon, Chatillon, Île-de-France, France. Email: cezary.ziemlicki@orange.com

Acknowledgments: Since 2016, INSEE, Eurostat and Orange Labs have been collaborating on exploring the usefulness of Mobile Phone Data for official statistics. In this context, we have been able to benefit from the work of a French national collaborative research project (ANR Cancan†) which collected three months of mobile signaling data in 2019. The raw data was deleted after twelve months. The approach described in this document only required exchanges of anonymous aggregates between INSEE and ORANGE. Processing of individual data was performed by each data owner. However, methods and algorithms starting from the raw data were developed jointly and transparently allowing both parties to evaluate and validate the outputs. We are grateful to Elise Coudin, Mathilde Poulhes, Etienne Come, Angelo Furno, Latifa Oukhellou and Zbigniew Smoreda for insightful comments, and to European Statistical System collegues for great discussions, in particular Fabio Ricciato, David Salgado and May Offermans. We also thank the audience at the BDES 2020 and at the NTTS 2021 conference. We also thank Romain Avouac and Vincent Attia who first explored the data, and the TIC survey team.

official statistics. Lack of access to real data, privacy protection and models to efficiently cooperate with private data holders have been long discussed challenges. The need to invest in a transparent methodology may be one of the distinctive features of official statistics relative to private producers who already disseminate statistical products worldwide.

Historically, a large body of research provided insights on human mobility by relying on Call Detail Records (CDR), data generated by the user when actively communicating through their cell phone (texting, calling...) and collected for billing purposes (Blondel et al. 2015). The reliance on user activity and the low time event rate has proven an obstacle for inferring from CDR present populations fitted for official statistics purposes (Sakarovitch et al. 2018; Vanhoof et al. 2018). Today in France and generally in Europe, private actors producing population statistics from their networks instead use re-purposed signaling data as base material when available. This new generation of mobile network data is characterized by a much greater spatiotemporal event rate. To ensure the centralized network knows about the mobile's state and location to reach it efficiently, passive communications are pervasive with the device when switched on, and are increasingly collected by Mobile Network Operators (MNOs) for network optimization and monitoring purposes. For statistical purposes, both types of data are however similar in the unitary information they contain: records convey *non-continuous* information on the proximate *radio cell* ($\approx$ antennas) mobile *devices* are wirelessly communicating with. Thus, three main dimensions of uncertainty should be addressed to infer from MNOs records information on population presence: temporal, spatial, and population coverage uncertainties (Ricciato et al. 2020).

To be disseminated as official statistics, dynamic population counts should be steadily comparable over time, thats is, as insensitive as possible to network-related and user behavioral effects or MNOs client turnover. They should also be consistent with other approaches to estimate population counts when other high-quality data sets are available. Indeed, if many case studies providing dynamic maps of mobile phone users exist, focusing only on mobile phone users with-out any extrapolation to the total population is of limited interest for official population statistics (Panczak et al. 2020). Mobile network data represent an opportunity for developing countries where alternative data sets are scarce and high-quality census data are not always available. In these countries, static population estimates can be vital for development and infrastructure planning (Salat et al. 2020). In developed countries, dynamic population counts, within typical days or over the year, are not part of official statistics production today. Yet population is a highly temporally dynamic variable, with considerable shifts in its distribution occurring in daily cycles. Today the majority of published indices are monthly or quarterly and correspond to the night-time population distribution. Hence, all applied sciences and policy support that require spatially detailed information on population distribution must rely on a fractional and static representation of reality (Batista e Silva et al. 2020). This is a major issue for planning of infrastructure and transport for instance, as well as assessment of human exposure to natural, environmental, epidemiological, and technological hazards (Freire and Aubrecht 2012; Panczak et al. 2020).

The literature on addressing spatial uncertainty is insightful. Location at radio-cell level is imprecise and depends on the network's local density. Voronoi tessellations are

commonly used models for radio-cell coverage as they require limited information (the cell tower physical coordinates) and are easy to implement by partitioning space. Yet, they are imprecise and do not correspond to the way the network functions (Sakarovitch et al. 2018; Ricciato et al. 2017). Probabilistic approaches accounting for overlapping coverage of cells are more realistic and thus preferable in a context where the mapping choice entails large discrepancies in outputs (Tennekes and Gootzen 2022; Ricciato et al. 2020; Ricciato and Coluccia 2021; Salgado et al. 2021). As for temporal uncertainty, interpolation techniques were explored for mobility analysis when targeted time granularity is high or when working with sparse Call Detail Records data (Hoteit et al. 2016; Chen et al. 2019; Bonnetain et al. 2021). Typically, CDR records a few events per device per day. In turn, signaling data ensures a very large detection, even at night – but the literature is still in its infancy due to lack of access. The availability of signaling data may allow for simpler strategies for the use case of dynamic population measurement and a reassessment of this dimension. Finally, only a specific fraction of the population is observed, and thus scaling is needed. To link mobile network data to other data sources, such as census data, home detection algorithms aim at identifying where the user lives, to rescale mobile-phone counts by residency locations, for example, as in Fekih et al. (2021). In the absence of a home detection step, cruder alternatives exist, using counts of mobile-phone users at night or census counts as rescaling factors. Home detection can be considered a data augmentation technique and rescaling can be stratified across more device characteristics to improve representativeness. Data fusion approaches – ensuring coherency across several sources with distinct strengths and weaknesses – can enhance mobile network data using additional sources. In particular, combining other data sources also allows to estimate daytime and nighttime populations. For instance, Batista e Silva et al. (2020) combine official statistics (residents, employees, students, tourist, and other population counts and estimated flows between region) with geospatial data to produce and validate a European data set of population grids taking into account intraday and monthly population variations, called "ENACT" (Schiavina et al. 2020).

In this article, we focus on estimating hourly population counts at a fine spatial scale over several months over France's metropolitan territory. Our approach is based on spatially merging passively-collected cellular network signaling data with fine-grained official population counts by place of residence. The structure of this article is as follows. In Section 2, we define the theoretical framework and draw particular attention to various dimensions of uncertainty observed in the data. In Section 3, we describe our estimation method, driven by consistency constraints at several spatial scales and over time. We require that the dynamic population counts stratified by living places are locally consistent with the official number of residents. This requires estimating a residency for each device and allows to break down dynamic population counts by place of residence. We focus on building dynamic flows for French residents only (more precisely residents of metropolitan France) and exclude from the onset the existence of inbound and outbound trips, absent reliable sources on daily population flows in and out of the country. We design the spatial mapping of population counts to reflect spatial accuracy by building an adapted quadtree grid – independent of administrative borders. In Section 4, we compare the results with the ENACT dynamic population counts. Finally, in Section 5, we discuss challenges that remain and argue that data fusion approaches between fine-grained official

statistics data sets and mobile network data, spatially merged to preserve privacy, are promising for future methodologies. We review how this work relates to recent initiatives in the European Statistical System to build a reference methodological framework for mobile network data integration into official statistics production (Salgado et al. 2021; Ricciato et al. 2020; Statistics Netherlands 2020).

2. Statistics of Interest, Theoretical Framework, and Challenges Deserving Particular Attention

2.1. Data

Three months of raw signaling data from Orange clients were collected from March 16th to June 15th, 2019. These data included all Orange client device interactions (active, such as text or calls, and passive, such as hand-overs between antennas) with the Orange metropolitan France 2G, 3G, and 4G networks. All personal information was removed and device identifiers were pseudonymized before storing the data, which were erased twelve months after collection. Data are collected through probes positioned on the Orange networks for monitoring performance. Events are information exchanges between the devices and radio cells, base station antennas which communicate wirelessly with mobile devices, hereafter, cell or radio cell. Users of the 4G networks generate on average about a thousand events per day, while users of the 3G and 2G networks generate respectively about 50 and 20 events per day.

Orange provides two types of radio cell location information. First, weekly extractions of a cell registry covering the data collection period were performed. The cell registry exists for network maintenance purposes and sees regular entries and exits following network life. The extracted data were limited to cell tower coordinates and cell technology type. Second, Orange Fluxvision provided a cell-specific coverage map, which modelizes the network coverage as of February 2019. The latter is static and is obtained from a radio-propagation model taking into account network specificities, local topography and land use, and device diversities through simulations. Applications for these data include helping to provide emergency call locations.

The official source which serves as a reference for localized population counts is fiscal data from 2016 (Filosofi for Fichier Localisé Social et Fiscal), obtained from merging fiscal income tax files (at the individual and household levels) and local residence tax files. It includes households with at least one income declaration and liable for the local residence tax (all owners, tenants, or occupants on a free basis of a piece of real estate). The situation is evaluated at the address on the first of January. Filosofi excludes the homeless population and people living in collective housing (retirement homes, young worker homes, prison, university residences, etc.). The native individual data is geolocalized at the tax address and can be aggregated over our regular grid of interest covering metropolitan France. As this data records tax addresses, it tends to be of lower quality to determine residency for young adults when they are fiscally attached to the home address of their parents. Yet, it is the most granular source of resident population localization available for France (Lamarche and Lollivier 2021). We note that the forthcoming gridded populations from the EU 2021 censuses could be used in future development as an alternative.

Finally, the TIC survey (Enquête sur les technologies de l'information et de la communication) is a survey conducted by Insee every year, with the objective to collect information on the usage of information and communication technology in households. In 2021, 31,000 different households were interrogated across metropolitan and overseas France (excluding Mayotte). A question was specifically added asking for the MNO of the surveyed individuals. This allows us to get a sense of how representative Orange subscribers are of French residents in 2021, which is close to the study period.

We provide the definitions of the populations we use to construct our estimates in Table 1.

2.2. *Notations*

Let us introduce the notations summarized in Table 2 . The device set $\{d \in D\}$ is defined as all devices (as identified by their International Mobile Subscriber Identity) appearing at least 30 distinct days over Orange 2G, 3G, and 4G networks within the three-month time window T, identified as an Orange client (based on the Mobile Network Code of their IMSI) and as a mobile phone (based on their Type Allocation Code and an external register of known mobile phones TAC). The hourly time grid of interest $\{t \in T\}$ contains each hour in the three consecutive months. P is the total population of interest, which is assumed constant over T. D_t denotes devices observed during t and D_l devices observed at date l. The grid $\{i \in I\}$ on which we want to estimate present populations is made up of tiles, and $\{j \in J\}$ is the set of cells in the MNO network.

We would be in an ideal setting for statistical purposes if the number of observed devices were constantly equal to the size of the target population, that is $D_t \Leftrightarrow P$. In this case, a count of all devices in location i at hour t would provide our target statistics $u_{i,t}$, the

Table 1. Population definitions by data source.

	Source	Period (location)	Population
Residents P	Filosofi (tax data)	01-01-2016	Households with an income declaration and liable for residency taxation
Devices D	Orange Probes	16/03 – 15/06/2019	Users of a mobile phone, with an Orange contract, interacting at least 30 distinct days with the MNO network

Table 2. Notations.

$d \in D$	Devices in the scope, detected on the network over the period of interest
$t \in T$	Hourly time grid of interest (several weeks)
P	The target population set
D_t	Set of devices in the scope, detected on the network during t
D_l	Set of devices in the scope, detected on the network at date l
$i \in I$	Tiles that cover the territory of interest
$j \in J$	Cells of the MNO network
$u_{i,t}$	Population count in location i at time t (target statistics)

present population for this location and hour. In fact, several dimensions of uncertainty in the data must be dealt with to compute valid estimates for present populations. We detail these dimensions now.

2.3. *Temporal Uncertainty: Only Counting Active Mobile Devices Leads to Unreasonable Variations in Aggregates*

The first uncertainty observed in the data is the high temporal variation in counts of active devices. This variation confounds many mechanisms unrelated to population variation and is actually a major stylized fact of mobile network data. The presence of a given device may be very sporadic in the collected data, *both within and across days*. Although we improve time event rate by relying on both passive and active data, this remains a crucial issue when estimating population variation.

Within a given day, the behaviors of mobile phone users and network coverage are the main reasons for the observed variation: users may choose to shut down their phones, run out of battery, or out of signal. Network coverage is uneven across space – and can not be ignored in a country such as France where low-density areas still host a large fraction of the population. Across days, we may expect a large role for client turnover, the telecom market in France being competitive and changing operators being increasingly easy for customers. For instance, over the first semester of 2018, four million mobile phone numbers were kept while changing MNO. Telco market dynamics are also at play: the sim cards number quarterly growth was more than 200,000 in the second quarter of 2019, over our period of interest (for all MNOs, excluding M2M).We also have to take into account failure in data collection of the richer passive data – as probes may punctually malfunction, without a too high cost for the MNO as it does not affect the network communications but is rather a monitoring tool. Investing in their continuous reliability is not a priority. In addition to these mechanisms entailing undesirable user disappearance, users may disappear due to outbound trips – but for now, they are indistinguishable from the former, although of interest for present population estimation. Characterizing places where outbound trips originate (airports, train stations, borders) could be considered to discriminate absence from the territory from other phenomena.

Figure 1 illustrates the issue on the data set. The first panel represents the ratio of observed devices $\frac{|D_l|}{|D|}$ for each date l. Over a long period, aggregate variations in the percentage of users observed on a given date vary considerably: by several percentage points regularly, occasionally by more than ten percentage points. Underlying causes of the aggregate variations remain ultimately speculative (reported data collection failure, outbound trips e.g., on bank holidays, etc.). While we can detect relatively easily unreasonable variations at the aggregate level, we may suspect that the same issues go undetected at local levels. Within a day, the hourly detection rate varies between 70 and almost 90% (second panel of Figure 1), a rate considerably higher than for CDR data. For instance, Orange CDR data from 2007 recorded only a few events per device per day and a percentage of observed devices which goes from less than five at nighttime to about 50% in the late afternoon (Galiana et al. 2020a). Variations seem highly driven by devices disconnecting at nighttime. Although the detection rate is arguably rather high, only counting active mobile devices would lead to unreasonable temporal variations in aggregates.

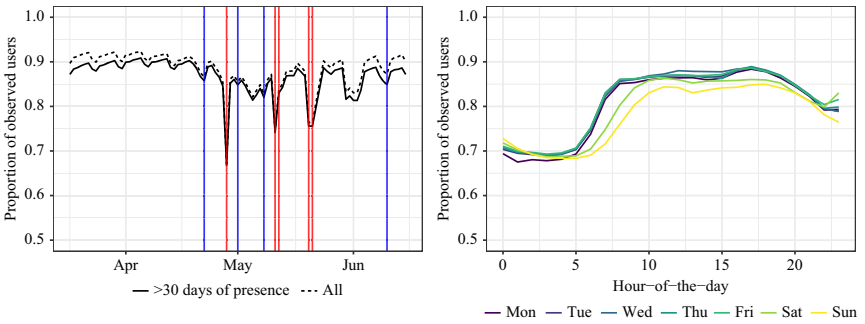


Fig. 1. Proportion of observed devices by date and hour of the day.
Note: Left: % among all Orange devices. We distinguish all users (dotted line) and users present for at least 30 days. We represent dates with reported data collection issues (in red) and within-week bank holidays (in blue). Right: % among devices appearing from March 16th to March 31st, 2019. Scope: Orange metropolitan France network, Orange-client devices identified as mobile phones.

A simple method to deal with this source of inconsistency is to account for all devices across the entire time grid by adopting a device-centric view and interpolating device trajectories when unobserved – building a panel of device trajectories. We detail this method in Section 3.

2.4. Spatial Uncertainty: Mapping Presence Over the Network in Space

Recorded events are located at the level of radio cells, not on the grid of interest I on which we estimate present populations. Spatial information on devices over time must be mapped onto the grid of interest. In general, radio cells overlap which means that a mobile phone in tile i at time t can potentially be detected in several cells, as pictured in Figure 2.

Let us define coverage probability matrix A, such that A_{ji} represents the probability of being detected at cell j while being in tile i:

$$A_{ji} = \mathbb{P}\{\text{device detected in cell } j \mid \text{device in tile } i\}.$$

Orange Fluxvision provided an instantiated cell-specific coverage map matrix A, which modelizes the network and devices population as of February 2019. Note that without this

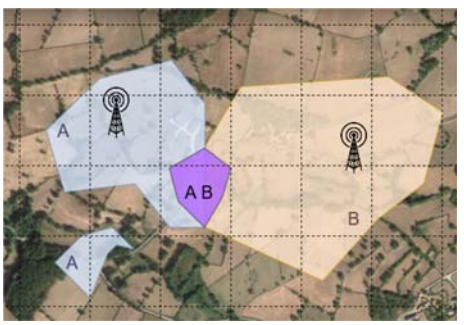


Fig. 2. Illustration of the relationship between tiles and radio cells.
Note: The grid of interest is made up of tiles represented by dotted squares. In this illustration two radio cells A and B are pictured. In general a cell covers more than one single tile, and cells overlap, meaning that a single tile is covered by more than one cell.

information coming from the MNO itself one can use a Voronoï tessellation of the cell plan to model radio-cell coverage or design a more complex model like in Tennekes and Gootzen (2022). Figure 3 illustrates how the coverage probability matrix encodes the device connexion information from tiles to cells, and the spatial mapping task which aim at locating devices connected to cells into tiles. Note: Left: 4 tiles are represented by rectangles and 3 cells by ellipses. A mobile phone in tile i has probability A_{ji} to be detected in cell j. For instance, the mobile phone in tile 2 has probability A_{12} to be detected in cell 1 and probability A_{22} to be detected in cell 2, with $A_{12} + A_{22} = 1$. Right: The phone detected in cell 2 is mapped to the tile grid using spatial mapping Q. It has probability Q_{22} to be located in tile 2 and probability Q_{32} to be located in tile 3. Using a uniform prior, we have $Q_{22} = A_{22}/(A_{22} + A_{23})$ and $Q_{32} = 1 - Q_{22}$.

The dimensionality of matrix A is high when considering the entire French territory (approximately 55 million tiles with sides of 100 meters for several hundred thousand cells). If we take this matrix as a good approximation of the ground truth, radio cells are massively overlapping with each other over the territory and the signal is quite diluted, in the sense that the cell-tiles mapping is highly non-exclusive. On average over France in the coverage map, there are 33 cells with positive coverage per 100-meter tile. However, most of these links are really low: if we restrict to links with significant coverage (say $A_{ji} > 0.1$) there are only 2.4 cells per tile and 16% of tiles without any linked cell. The first panel of Figure 4 illustrates the distribution of areas covered (resp. significantly covered) by cells. The median cell covers 25 tiles significantly (1706 tiles with non-zero coverage). The second panel of Figure 4 illustrates the number of cells per tile. The median tile is covered by 30 distinct cells, but significantly by two cells.

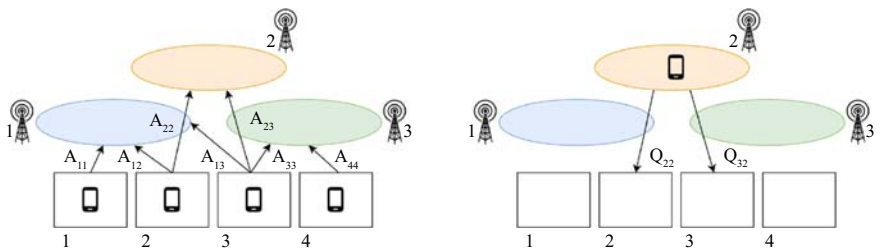


Fig. 3. Coverage probability matrix and spatial mapping.

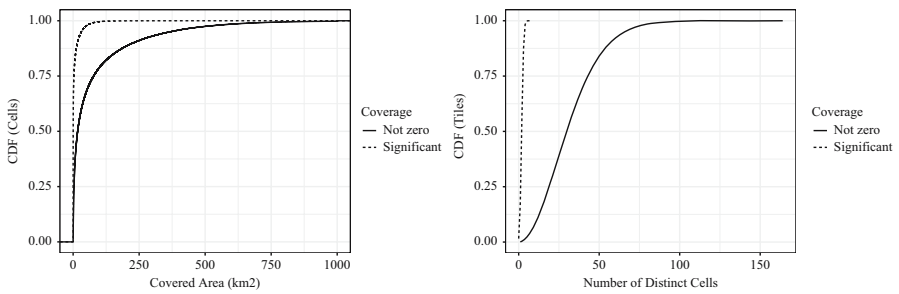


Fig. 4. Distributions of cells coverage areas and of cells per tiles.
Note: Here, a tile i is said covered (significantly covered) by a cell i if $A_{j,i} > 0$ ($A_{j,i} > 0.1$).

In practice, the location accuracy we can expect when mapping in space an event located at a given cell relies on the extent of the cell-covered area. If taken as the ground truth, the estimation A encodes a relatively low network precision which is highly heterogeneous over space. The method to map cells to tiles based on A as precisely as possible while managing dimensionality is detailed in Section 3.

2.5. Coverage Uncertainty: Mapping Devices to the Population

The third dimension of uncertainty in the data is linked to population coverage. By definition, we do not observe the target population (all French residents) but a selected subset through their device(s). It is a subset to the extent that we have a reliable method to exclude devices which are not used by human beings (M2M, IoT), and that the remaining Orange devices identified as mobile phones indeed belong to French residents who carry a single Orange device along. We maintain this assumption throughout this work, although some devices may not be carried out by French residents in the fiscal sense. Present population estimates relying on device counts are consistent under the assumption that the mobility and presence patterns we observe for the selected subset of the target population can be extrapolated to the target population.

However, active Orange devices are not exactly representative of the target population. Orange Fluxvision estimated that the national market share of Orange was around 37% (29% for individual consumers) in 2020 (Insee 2020). The TIC survey from 2021 indicates that biases first include an over-representation of the elderly (from 55 to 84 years old) amongst Orange subscribers in comparison with the general population, and an under-representation of younger individuals (in particular from 15 to 24) as shown in Figure 5. The youngest individuals interrogated in the TIC survey are 15 years old: children younger than 15 are also naturally under-represented in our signaling data, as many do not possess a mobile phone. Individuals belonging to households with lower incomes are also less likely to be Orange mobile users. Finally, non-French citizens living in France are also under-represented amongst Orange subscribers.

A milder assumption can be formulated when some characteristics of the devices are observed: mobility and presence patterns of unobserved residents can be extrapolated from mobility and presence patterns of persons carrying an Orange mobile device when they share these characteristics. High-frequency signaling data fittingly allows observing one such characteristic – the home environment.

A related question is then whether the places of residence of Orange clients are representative of places of residence in France. Locally, we may derive a ratio between residents and MNO-detected residents. We expect that the lower this ratio is, the better the local representativity, as we observe more devices per resident. This ratio informs on differential local representativity but also divulgates Orange market shares and is thus not reported on a map. In practice, it is highly heterogeneous over space. The local representativity tends to deteriorate in poor neighborhoods in some urban areas. As an illustration, Figure 6 represents how this ratio distribution evolves by municipality median disposable income. At the municipality level, there is no clear correlation between disposable income and local representativity as captured by the ratio between residents and MNO-detected residents, except at the lower hand of disposable incomes. If the

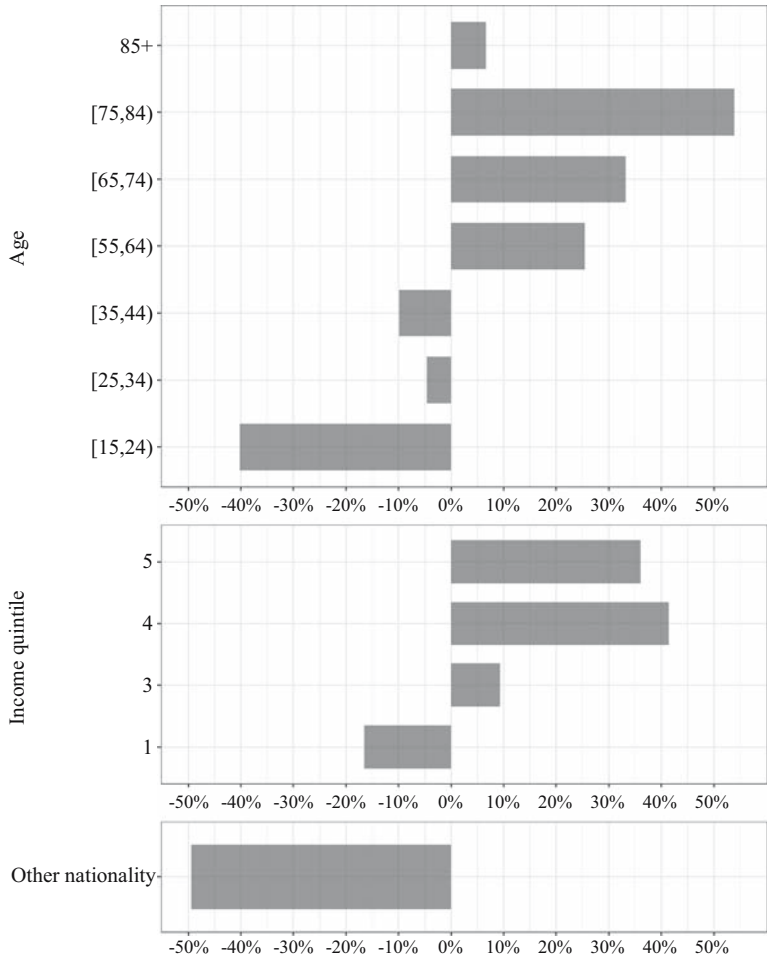


Fig. 5. Probability of being an Orange subscriber as a function of age, income quintile and nationality on the TIC survey data, modeled with logistic regression.
Note: The odd ratios minus 1 as percentages are displayed. Reference modalities are "from 45 to 54" for the age variable, "2" for the income quintile and "French" for the nationality.

median ratio is about 0.33 detected resident per resident, the first decile (D1) of the ratio is 0.16 while the ninth decile (D9) is 0.58. A large heterogeneity may also exist at lower spatial scales.

On top of users' socio-economic characteristics which may differ from one MNO to another, mobile network data better captures the behavior of active users who benefit from good network coverage. In this case, raw mobile network activity data under-represents the population from less dense areas where the network is less developed.

3. Measuring Present Population

We now present the methodology we use to compute hourly present population estimates over metropolitan France using three months of signaling data from the Orange network and the geography of French residents from fiscal data. Formally, we aim at computing estimates $\hat{u}_{i,t}$

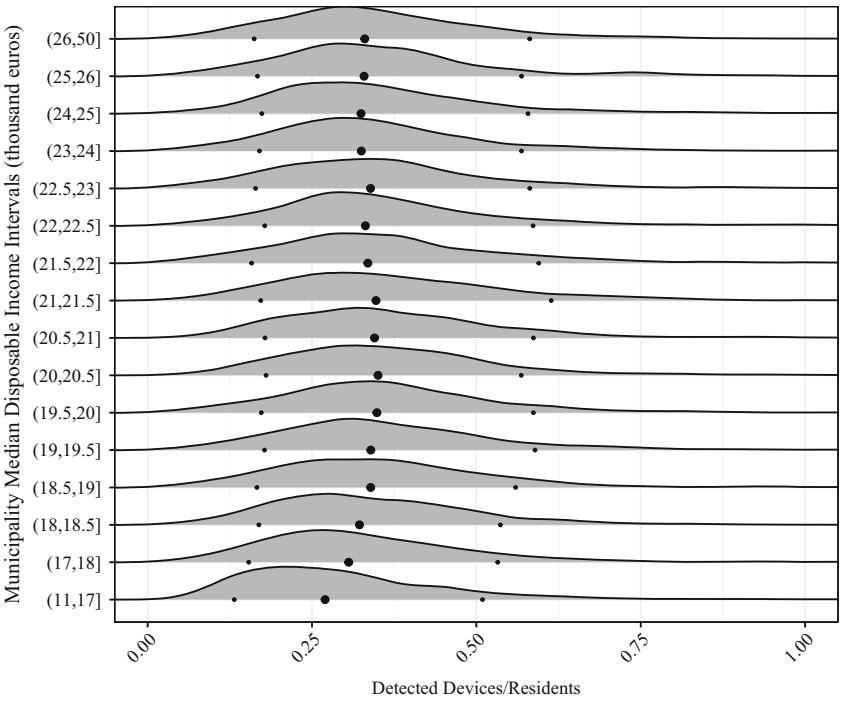


Fig. 6. Detected residents per actual residents and Municipality-level Disposable Income.
Note: Distribution (D1, Median, D9) of the municipality-level ratio by municipality median disposable income range.

giving the distribution of France metropolitan residents at hours t in T over several months and over tiles $i \in I$. As an intermediate output, we aim at estimating $\hat{u}_{i,t,r}$ giving the distribution of France metropolitan residents in location r present in tile i at hour t. Thus, $\hat{u}_{i,t} = \sum_r \hat{u}_{i,t,r}$.

Minimal consistency constraints. For consistency, we require that the present population estimate total matches an external official source: $\forall t \in T, \sum_i \hat{u}_{i,t} = P$. We will further require that we have as many residents of r contributing to $\hat{u}$ as there are residents of r in the official source to balance our estimates across residencies: $\forall t \in T, \sum_i \hat{u}_{i,t,r} = P_r$. This requires estimating a residency for each device. It allows to break down population presence by place of residence. We note that we exclude from the onset the existence of inbound and outbound trips, due to the absence of reliable sources on daily population flows in and out of the country.

As we do not restrict the analysis to a sub-period or part of the territory, we favored a simple method to deal with the high dimensionality of the data.

Table 3. Additional notations.

$j_{d,t} \in J$	Estimated *presence cell* for device d during time interval t
P_r	Official residents in place r
D_r	Devices with estimated residency in r
$\mathbf{m}_{r,t} \in \mathbb{N}^{\lvert J \rvert}$	Devices count in *presence cells* at time t with residency r
$\mathbf{u}^0 \in \mathbb{N}^{\lvert I \rvert}$	Population count from official source over tiles
$\hat{\mathbf{u}}_{i,t} \in \mathbb{R}^{\lvert I \rvert}$	Estimated population count at time t over tiles
$\hat{\mathbf{u}}_{i,t} \in \mathbb{E}^{\lvert I \rvert}$	Estimated population count at time t who are resident in r over tiles

3.1. Overview of the Method

Our population estimation relies on several modules, the critical ones being the construction of a device presence panel and residency-based weighting.

Device presence panel. This first module's role is to bypass the temporally sporadic presence of users over the network, by interpolating the trajectory of each device before any aggregation. In practice, we define and estimate for each hour and each device a *presence cell* $j_{d,t}$ – whether or not the device was observed during t. The presence cell $j_{d,t}$ represents the cell where the device d would mostly connect during the time interval t if it was active.

Residency characterization. This second module's role is to estimate the residency r of each device d at an adapted geographical level to be defined. We denote D_r the set of devices with residency in place r, which can be compared to the set of residents in the official source, P_r.

We can then define presence over cells of devices residing in r, denoted $\mathbf{m}_{r,t} \in \mathbb{R}^{|J|}$ with

$$m_{j,r,t} = \sum_{d \in D_r} 1\{j_{d,t} = j\}.$$

$m_{j,r,t}$ is the count of devices which are resident in r and are considered present in cell $j \in J$.

Residency-based weighting. This third module's role is to extrapolate the number of devices to an estimate of the present population. If we assume that the sample D_r has been randomly chosen among P_r with sampling rate $\frac{|D_r|}{|P_r|} = \frac{1}{w_r}$, a valid estimation of the expected presence of residents of r over the network cells is $w_r \times \mathbf{m}_{r,t}$. We extrapolate the presence patterns of D_r to P_r. It amounts to applying a rescaling factor $|P_r|$ to the density of resident devices. The pseudo-weight $w_r = \frac{|P_r|}{|D_r|}$ is the ratio of residents from the official sources to the resident devices in place r. Instead of counting for 1 person, each device in the scope will participate in the aggregate with weight $w_d = w_{r(d)}$. Of course, this approach is valid if D_r is indeed close to a random sample drawn from P_r. If we could add additional inferred characteristics to the devices, we could improve this stage by stratifying weights beyond residency.

Spatial Mapping. This fourth module's role is to transform an active device at the cell level into an active device at the tile level. We do it by defining a linear spatial mapping Q: $\mathbb{R}^{|J|} \to \mathbb{R}^{|I|}$ which distributes a vector of presence over the network cells in the tiles with $\sum_i Q_{ij} = 1$, by specifying

$$Q_{ij} = \mathbb{P}\{\text{device mapped to tile } i \mid \text{device connected to cell } j\}.$$

Then, $Q\mathbf{m}_{r,t} \in \mathbb{R}^{|I|}$ represents the presence over tiles of devices which reside in r, at time interval t.

Presence Estimation. We estimate the presence over tiles of the residents of place r denoted $\hat{\mathbf{u}}_{r,t}$ by attributing to residents of r the presence distribution of devices who are resident in r. That is,

$$\hat{\mathbf{u}}_{t,r} = w_r Q \mathbf{m}_{r,t} \tag{1}$$

We finally estimate the total population presence over tiles with

$$\hat{\mathbf{u}}_t = \sum_r \hat{\mathbf{u}}_{t,r}$$

These definitions enforce our minimal consistency constraints. Note that $\hat{u}_t$ can be written as a reweighted projection of device-level trajectories:

$$\hat{\mathbf{u}}_{i,t} = \sum_{j \in J} Q_{ij} \sum_{d \in D} w_d 1\{j_{d,t} = j\} \tag{2}$$

3.2. Implementation

The raw signaling data contain about 20 billion events per date, totalizing 130 Tb of parquet files. We handled them using the big data framework Spark on an HDFS infrastructure at the MNO office. The Spark cluster was configured to stop any job lasting more than 24 hours. Given that the cluster was shared with other projects, the estimated resources available for our project were 300 CPUs and 1.2 Tb of RAM. We queried the data through PySpark, the Python API for Spark. We did not use any algorithm that could not be simply implemented using the PySpark API functions on the raw signaling data, which limited device-level algorithms. Even with simple algorithms, the whole-network longitudinal analysis entailing device-level sorting over a long period was challenging given the available resources.

3.2.1. Device-level Simplifications to Manage Dimensionality

Given these constraints, the following simplifications drove the choice of the method presented in Subsection 3.1 and were adopted in the implementation:

1. **For device presence, the location information is restricted to one cell per hour**. It seems a reasonable simplification for official statistics low-frequency purposes (at most, presence per hour). It stabilizes by design the oscillation phenomenon by which a motionless device may switch cells for network-related reasons (Katsikouli et al. 2019). However, if it is probably a good approximation for motionless devices, it is not for non-stationary devices visiting a high number of distant cells during an hour. Therefore, we oversimplify the presence of non-stationary devices by attributing them to one cell on their path.
2. **The location information was kept at the cell level for all device-level calculations**. We were only allowed to export aggregates to the national statistical institute premises, as opposed to device-level data. Hence, the spatial mapping was performed on aggregates only. This avoids bottleneck operations on $|D| \times |I|$-sized matrices during calculations (such as operations involving A). On one hand, this leaves the spatial mapping from cells to tiles easily adjustable downstream and leaves room for comparison between several choices of spatial mapping, which has been shown to matter a lot (Ricciato et al. 2020). On the other hand, as the spatial links between cells are never considered at the device level, some loss of spatial information is likely.
   In particular, the location information was kept at the cell level for device-level residency characterization. To bridge the tiles resident counts and the cells resident devices so as to build weights with counts in the same geography, we can map residents to cells or resident devices to tiles. Instead of affecting to each device a home tile i, we distribute each resident in a home cell j, the cell they would have probabilistically connected to if they were Orange users, using the probability distribution encoded in A. This choice avoids resorting to a home-cell spatial mapping $\mathbb{R}^{|J|} \rightarrow \mathbb{R}^{|I|}$ which requires dealing with

an additional probabilistic error. This approach is not straightforward in a multi-operator scenario where there is in general a specific radio cell network for each MNO.

While these simplifications are acceptable given our constraints, they are not necessarily recommended for future standard methodologies. Figure 7 summarizes the different steps of the method now described in more depth.

3.2.2. Device Presence Panel

We first filter devices identified as mobile phones (to filter M2M). Note that about 20% of daily unique identifiers are filtered as they can not be identified as mobile phones based on their TAC. We expect a large part of these excluded devices to be carried out by machines rather than persons (M2M and IoT devices, such as cameras, vehicles, alarms, sensors, etc.). We retain devices that are present at least 30 distinct days out of the three months so as to ensure relative stability of the scope (e.g., to filter movements due to client turnover as they are irrelevant to inform on total counts). Here, a user is present on a given day if there is at least one radio-cell event associated with its IMSI on this day. Figure 18 in the Appendix (Section 7) gives the distribution of the lengths of observations (in days) over the period. We keep 87 % of devices when filtering those present at least 30 days out of the three months. 67 % of the devices are observed for 80 days or more, and 9% are observed for ten days or less.

We attribute to each device appearing on date l a presence cell $j_{d,t}$ for each hour interval from 5 a.m. on date l to 5 a.m. the next day $l+1$ as follows. When the device appears during an hour interval t, the presence cell at t is taken as the cell recording the most events during this hour. When the device does not appear during an hour interval t, we consider as candidates all cells with an event from the device on hour t' from 0 a.m. on date l to 5 a.m. the next day. We extended the search window to 29 hours for helping interpolation in the early morning with nighttime observations. We rank the candidate cells for hour t first by the time interval distance between hour t and the hour of the observed event t' and

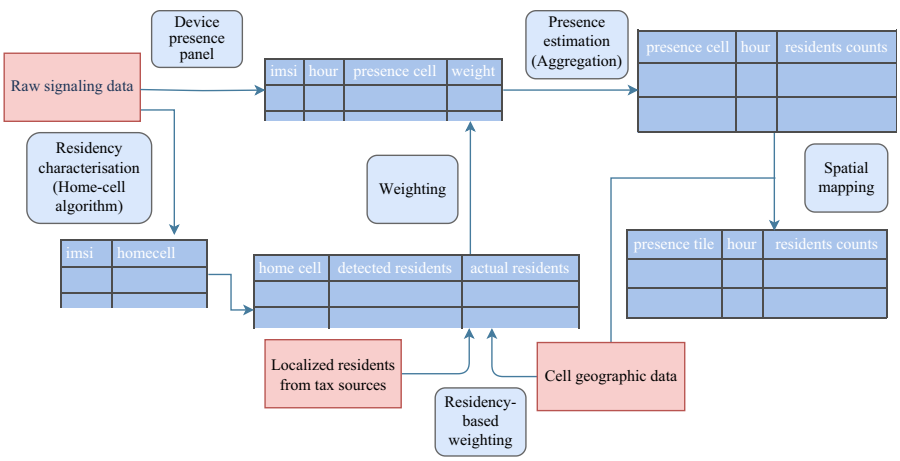


Fig. 7. Overview of the method implementation.
Note: Red boxes represent source data sets, dark blue boxes represent produced data sets and the key variables. Arrows represent transforms, which pertain to a given module referenced in light blue boxes.

second by the number of signalling events generated by the device at t'. We retain the cell ranked first. This cell selection based on the number of events is also a simplification as generating a higher number of events in a certain cell for a given hour does not automatically indicate the device spent more time in that cell than in others. One could choose the presence cell using more complex rules based on the distribution of signaling events on a sub-hour level.

Figure 8 presents the mean interpolation error, defined as the absolute difference between the reference hour t and the hour of actual observation, used to estimate the *presence cell*. It is on average lower than one hour but varies following the daily user behaviors and technology. It is particularly high when the presence cell is a 2G or 3G cell.

3.2.3. Residency Characterization

We can draw a picture of the typical day (including weekends and holidays) of (over-15) persons in France using time use surveys. On average in 2010, 8:30 hours are spent sleeping, 1:02 hours for washing/health care, 2:13 hours for eating, 4:04 hours for leisure (including 2:06 hours spent watching tv), 3:10 hours for domestic work, 2:51 hours working or studying and 0:24 minutes of work-home commute (Ricroch and Roumier 2011). This average includes unemployed, retirees, and housewives and may be heterogeneous across population types. However, for the "average" person, the vast majority of the time is spent at home.

For the characterization of residency, we therefore choose the cell with the most time spent rather than favoring a heuristic using time spent at night, although we run it as an alternative. Observation at night suffers from undersampling (Figure 1). In addition, we do not want to assume or constrain where the population is at night but rather deduce it from the data. For instance, we note that 1.8 million employees work at nighttime (8 p.m. to 5 a.m.) more than half of their working hours a given month (Létroublon and Daniel 2018).

To estimate the residency of each device, we thus want to identify recurring points of presence. A device is present at cell j in hour t as soon as an event is recorded at this cell during the time interval t. For this task, a device is therefore counted as present in all cells which have detected it at some point – even for a single event. This is therefore distinct from the presence cell as defined to track the longitudinal presence of devices. Here, all

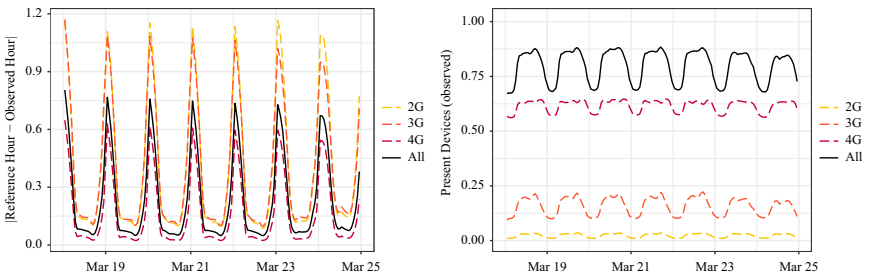


Fig. 8. Active devices and Interpolation in Time.
Note: The right panel presents the proportion of active devices among all devices in dotted lines, in total and by cell technology. The left panel presents the mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation.

events are used and unobserved periods are not inferred. Then, the max-presence cell is the cell where the device has been recorded present the most. For simplicity in computation and scalability, we kept the analysis at the cell level but note that this approximation could probably be improved by considering several cells and their geography. It turns out that we find evidence of at least one strong "anchor point" for most of the devices in the scope. Figure 9 represents the distribution of the number of distinct hours of presence in the max-presence cell when the latter is defined over two weeks. 75% of the devices in the scope are observed at least 27% of the hours over the period in the same max-presence cell. Overall, signaling data prove very promising for pinpointing anchor points. Note that for this step, we did not interpolate device trajectories over an unobserved time period. If we define residency as the place with the most time spent, surely longitudinal signaling data offer many perspectives.

This step requires us to use all events and to sort the longitudinal data by device pseudoidentifier to be able to rank cells. To derive the max-presence cell over three months, we run a max-presence cell algorithm by two-weeks windows and kept per device only ten max-presence cell candidates. Filtering the least likely candidate cells allowed to keep the computational burden manageable. Finally, we define the home cell as the max-presence cell over the pooled max-cell candidates. At this stage, this step is highly stylized from a methodological point of view but benefits from the richness of the data over a long period.

Note that the international definition of population recommended in official statistics is based on the concept of 'usual residence', which is the place of daily rest assessed over a period of at least 12 months. We only had three months of data for this work, but the same methodology could essentially be applied with a longer observation window to produce comparable population estimates, consistent with the international definition of population.

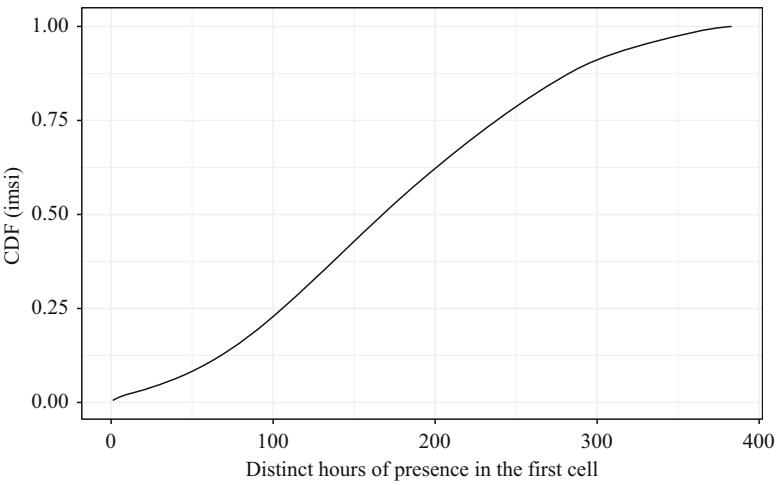


Fig. 9. Distinct hours of presence in the max-presence cell.
Note: The max-presence cell is defined as the cell recording the highest number of distinct hours of presence over a two-week time period. One event within the hour is enough to consider presence in this cell at hour *t*. Scope: Orange metropolitan France network, Orange-client devices identified as mobile phones, from March 16th to March 31st, 2019 (384 hours).

3.2.4. Residency-Based Weighting

In this step, we map French residents over home cells using realistic information on the coverage of each tile of 100 meters by Orange cells as provided by Orange (matrix A). Let us denote $\mathbf{u}^0$ the resident counts over grid I estimated from 2016 fiscal data. If all French residents were Orange clients, active and at home, we would expect to observe over the Orange network cells the following counts:

$$A\mathbf{u}^0$$

We define places of residency r as home cells or groups of contiguous home cells gathering at least 20 detected resident devices. This allows us to adapt our definition of residency to the MNOs' varying local market shares, having enough devices per residency place while keeping the place of residence relatively precise. We start from all home cells, search for the closest home cells for home cells with less than 20 detected resident devices, group both, and iterate until all groups of contiguous home cells reach the condition. This ensures a minimal size for D_r while keeping a high level of disaggregation in r. In turn, $P_r = \sum_{j \in r} [A\mathbf{u}^0]_j$ is the expected number of residents in the home-cell group r. Figure 10 presents at level r (group of contiguous home cells) and at the municipality level the number of detected residents against actual residents, hinting at weight heterogeneity.

We define weights with the ratio of actual residents P_r divided by the network-detected residents D_r at the level of contiguous groups of home cells r which contain at least twenty detected resident devices. We end by trimming weights at their 2nd and 98th percentiles – weights fall in [0.07, 53.5].

Ideally, D_r and P_r should be consistently defined and both tend to match the concept of "usual residence" to avoid bias.

3.2.5. Spatial Mapping

We use the coverage probability matrix A provided by Orange Fluxvision that modelizes the network as of February 2019. A_{ji} represents the probability of being detected at cell j while being in tile i:

$$A_{ji} = \mathbb{P}\{\text{device detected in cell } j \mid \text{device in tile } i\}.$$

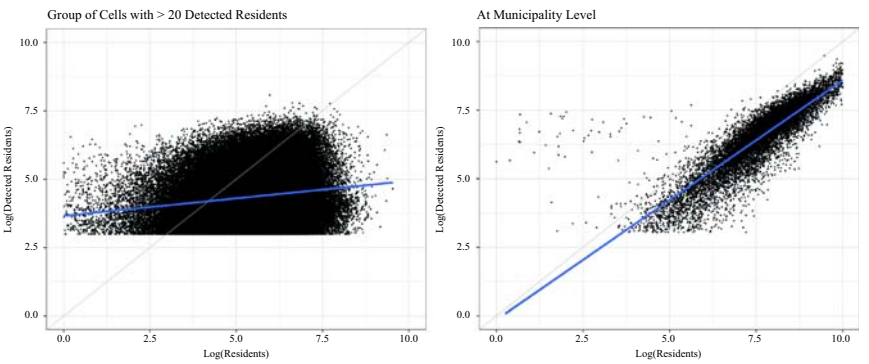


Fig. 10. Detected residents and actual residents, by aggregation level.

In particular $\sum_j A_{ji} = 1$. From a vector of presence over all tiles $\mathbf{u_t}$, we expect to observe on network cells $\mathbb{E}[\mathbf{m_t}] = A\mathbf{u_t}$ translating the presence of devices (we here assume $D \Leftrightarrow P$ for clarity of exposition). The estimate $\hat{\mathbf{u}}_t$ can be written in general as $\hat{\mathbf{u}}_t = g(A, \mathbf{m}_t)$ where g is a chosen *spatial mapping* (Ricciato et al. 2020). In this work, we focus on a linear estimator $\hat{\mathbf{u}}_t = Q\mathbf{m}_t$ (see Figure 3). Although any spatial mapping could be used, for our empirical results we follow Tennekes and Gootzen (2022) who suggest deducing Q from A using Bayes' rule by introducing a prior that reflects where the population is most likely located (e.g., based on land-use). In the results presented here, we use a simpler uniform prior. Specifically,

$$Q_{ij} = \frac{A_{ji}}{\sum_{i'} A_{ji'}}.$$

In addition, we propose a general framework to evaluate the location estimation accuracy of cellular network events. This evaluation combined with a quadtree algorithm enables us to build an adaptive spatial grid featuring small tiles for high accuracy areas and large tiles for low accuracy areas. The spatial accuracy is embedded within the dissemination grid.

Estimating accuracy locally. The accuracy of the linear estimator Q can be approached locally by defining the probability to localize in i a device that is in i_0 and connects to the network probabilistically through A.

$$N_{i,i_0} = \mathbb{P}\{\text{device mapped to tile } i | \text{device in tile } i_0\}$$

Formally, $N = QA$. A good estimator Q should lead to a high N_{i_0,i_0} probability (correct mapping), or at least a high probability of tiles i in the neighborhood of i_0. With previous notations, if we take the example of localizing a single device d which is in i_0, that is $\mathbf{u}_t = 1_{i_0}$, $N1_{i_0}$ can be interpreted as $\mathbb{E}[\hat{\mathbf{u}}_t | \text{device in tile } i_0]$. N encodes the spatial error by integrating the uncertainty from A and Q.

Embedding accuracy within dissemination. We build a quadtree that directly embeds the calculated spatial accuracy by gathering tiles until the probability of correct location in the macro tile I_0 (group of tiles) is higher than a threshold: $N_{I_0,I_0} > s$. We derive present population estimates within this reduced spatial grid, which visually provide a clear idea of the achievable accuracy (Figure 11). In what follows, the tile grid $i \in I$ should be understood as this quadtree-derived grid for $s = 5\%$, referred to as the dissemination grid. Table 11 describes the link between tile size and resident density in further detail.

3.2.6. Presence Estimation (Aggregation)

In practice, the set of devices present a given day varies (Figure 1). We denote this set $D_l \subset D$ for each date l. We only interpolate device-level trajectories within days. To respect our consistency constraint, we finally define $w_{d,l} = w_{r(d)} \times \frac{|D_r|}{|D_r \cap D_l|}$. If all detected residents of r are here on date l, their weights are set to w_r. If some are missing, their mass is transferred to the remaining residents' devices.

Precisely, our final estimator writes, for the chosen spatial mapping Q:

$$\hat{u}_{it} = \sum_j Q_{ij} \sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j\}$$

To illustrate the difference with counts of active devices, we reproduce Figure 8 by reweighting each device with $w_{d,l}$ in Figure 12. The mean interpolation error increases –

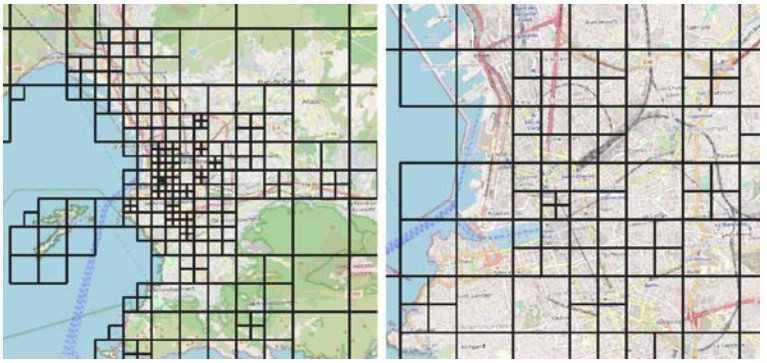


Fig. 11. Reduced grid with a threshold $s = 1\%$. The larger the tiles, the less accurate the spatial mapping is locally.
Note: The grid was based on the Orange matrix **A** and a uniform prior.

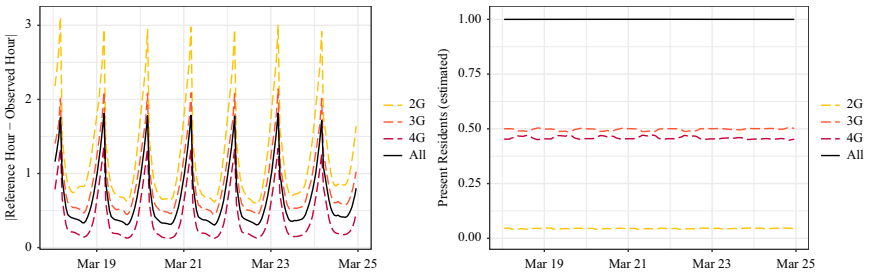


Fig. 12. Estimated present population and interpolation in time.
Note: The right panel presents the proportion of estimated present residents among all residents, in total (by assumption, all residents are represented) and by cell technology. The left panel presents the weighted mean interpolation error defined as the absolute difference between the reference hour t and the hour of actual observation. Weights are $w_{d,l}$.

showing that devices with high interpolation error have been upweighted. This suggests that this weighting scheme corrects for the under-representation of users with limited network access or limited network interactions. By construction, the total number of present residents is constant. Estimated presence over the 3G network is now comparable to estimated presence over the 4G network: $\sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \& j \in 3G\} \approx \sum_{d \in D_l} w_{d,l} 1\{j_{d,t} = j \& j \in 4G\}$.

3.2.7. Code Availability

The code implementing these modules is available on github at https://github.com/InseeFrLab/presentpop.

4. Results and Comparison with External Sources

The clear advantages of present population estimates derived from mobile network data are their timeliness, their granularity in time, and (relatively) in space. We first provide a rapid overview of the hourly and weekly fine-grained patterns which can be uncovered. This dynamic nature and this spatial extent are rarely achievable with other sources. We

then compare some present population estimate snapshots to other external, more static, sources of population density.

4.1. Daily and Weekly Cycle, Local and National Variations

Figure 13 illustrates the daily variations of present population density at the national level and in Paris. To represent the 24-hour *within-day* variations, we subtract from the hourly density its 24 hours mean. The 24-hour variation features the daily pendulum movement of suburban commuters: while the present population tends to be higher in the periphery of urban areas at night ($\geq$ 100 persons per square kilometer compared to average), at 9 a.m. the population in these peripheries decreases for the benefit of urban centers until to the evening. At a finer scale in the Paris surroundings, the variation of present population density discriminates places mainly characterized by their economic, leisure, and touristic activities from more residential areas, and shows the attractiveness of a multi-polarized center.

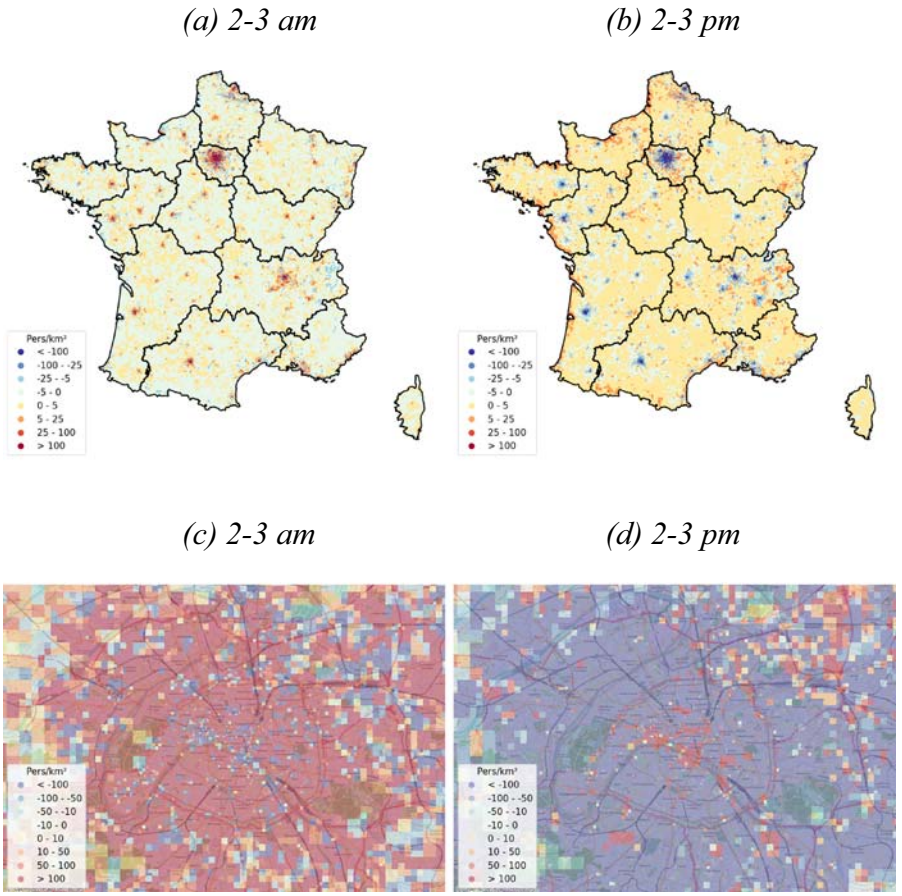


Fig. 13. Variation of present population density (persons per square kilometer) at daytime and nighttime at national level and in Paris.
Note: To represent the 24-hour *within-day* variations, we substract from the hourly density its 24 hours mean (Wednesday, March the 20th).

Figure 14 illustrates the weekly variations of present population density at the national level and in Paris. To represent the seven days *within-week* variations, we subtract from the present population density at 3 to 4 a.m. its seven days average. The variation of nighttime present population density within the week discriminates the nights from Friday to Saturday and from Saturday to Sunday, where some locations in coastal areas and the mountains fill up. In Paris, the present population is overall lower during the nights from Friday to Saturday and Saturday to Sunday. However, we observe some nighttime excess in the present population in some places on these nights, probably reflecting nightlife activity or touristic overnight stays. The animated and colored version of Figures 13 and 14 is available at https://github.com/InseeFrLab/presentpop.

4.2. Comparison with External Sources

Measuring the quality of present population statistics is difficult as there is no source of truth. But, we can assess how comparable our estimates are to other high-quality

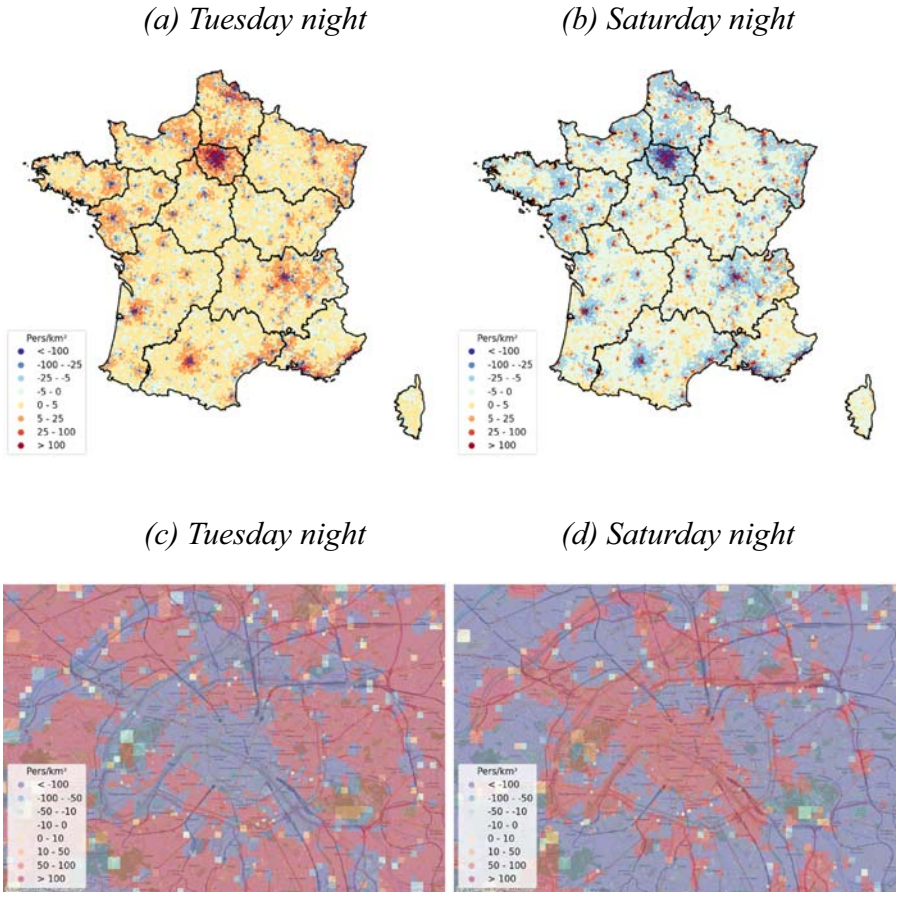


Fig. 14. Variation of present population density (persons per square kilometer), Tuesday or Saturday at nighttime, at national level and in Paris.
Note: To represent the seven days *within-week* variations, we substract from the present population density at 3 to 4 a.m. its seven days average (Monday 18th to the Sunday 24th of March 2019).

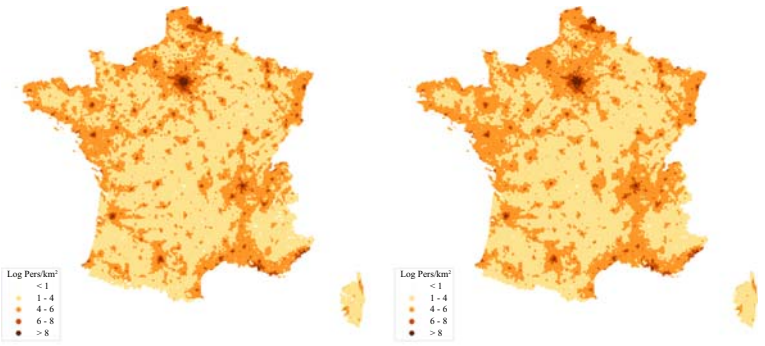
population measures. We here choose two points of comparison: residents geolocalized at their tax address and day and nighttime population density estimation from Batista e Silva et al. (2020).

We have considered comparisons with estimates of other external actors like Meta's high-resolution population density mapsand WorldPop population distributions. Indeed, the general idea of the methodology behind these estimations is – just like in our case – to calibrate population counts originating from new data sources (in the case of Meta and WorldPop, mostly satellite imagery) with spatially detailed population census data. However, we believe that a comparison with each of these two data sets is not entirely relevant. Indeed, their estimates are static and do not vary as a function of time, which is precisely the aspect where mobile phone data make a difference. The main interest of these static estimates lies in countries in which fine-grained census or fiscal data does not exist at a national level. This is not the case for the methodology outlined in this article, designed to estimate dynamic present populations. We still include a static comparison with the resident populations u^0 computed according to geolocalized fiscal data from 2016, which should be more accurate than Meta or WorldPop estimates in a country like France. Note that we also used this data source to build our weights.

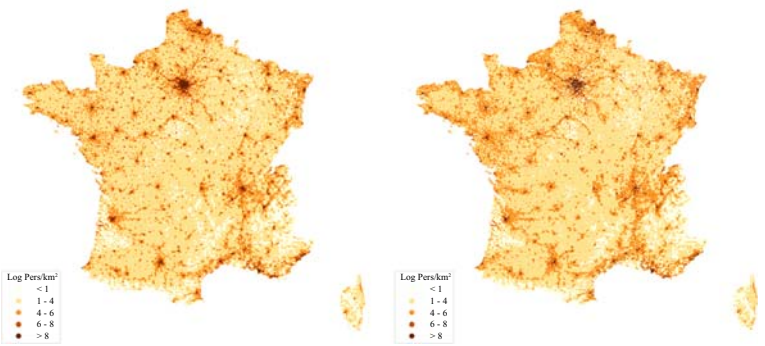
The second comparison source is the ENACT database which uses data fusion to estimate the population at daytime and nighttime at the European level on a 1km grid. Batista e Silva et al. (2020) use a top-down approach disaggregating NUTS3 population counts based on the assumed places of activities of 16 different population groups, using land use information. In contrast to our estimation, foreign tourists' presence is estimated and contributes to the population density.

We compare population densities in our dissemination grid. Figure 15 presents maps over France and Figure 16 a focus on the Paris area. At the country scale, the present population estimate respects the distribution of the population found in the other sources. However, the present population estimate shows a dilution of the population mass in space. Around dense urban areas, we observe a halo of presence absent from other sources. The first reason for this is probably the lack of accuracy of the cell-level localization. Another reason could be that by definition, the external sources considered here locate the population in buildings. Adopting a land-use prior in the spatial mapping task (Avouac et al. 2019) could therefore partially close the gap between our estimate and external sources and in any case improve its quality. However, this may create bias, in particular during daytime and weekends. We here chose a static spatial mapping, that is, independent of t. Another way forward could be to experiment with other estimators g, where $\hat{\mathbf{u}}_t = g(A, \mathbf{m}_t)$. On simulated data, Ricciato and Coluccia (2021) show that some estimators (called *data-first, ML/EM)* may be able to reduce the dilution effect compared to our linear spatial mapping with a uniform prior (*Simple Bayes-rule estimator* in the terminology of the article).

At the level of the Paris area, the structure of the present population distribution differs strongly during the day from during the night. The present population at 3 a.m. on a weekday (f) tends to offer a smoothed but quite similar version to the high-resolution image of the resident population (h). We report a contextual map of the Paris region in the Appendix (Section 7). For instance, the large parks in the east and the west (Boulogne and Vincennes) display a non-null density at nighttime according to our present population estimates, most likely due to a lack of accuracy in the spatial mapping. The population



(a) Present Population at 3p.m. *(b) Present Population at 3a.m.*



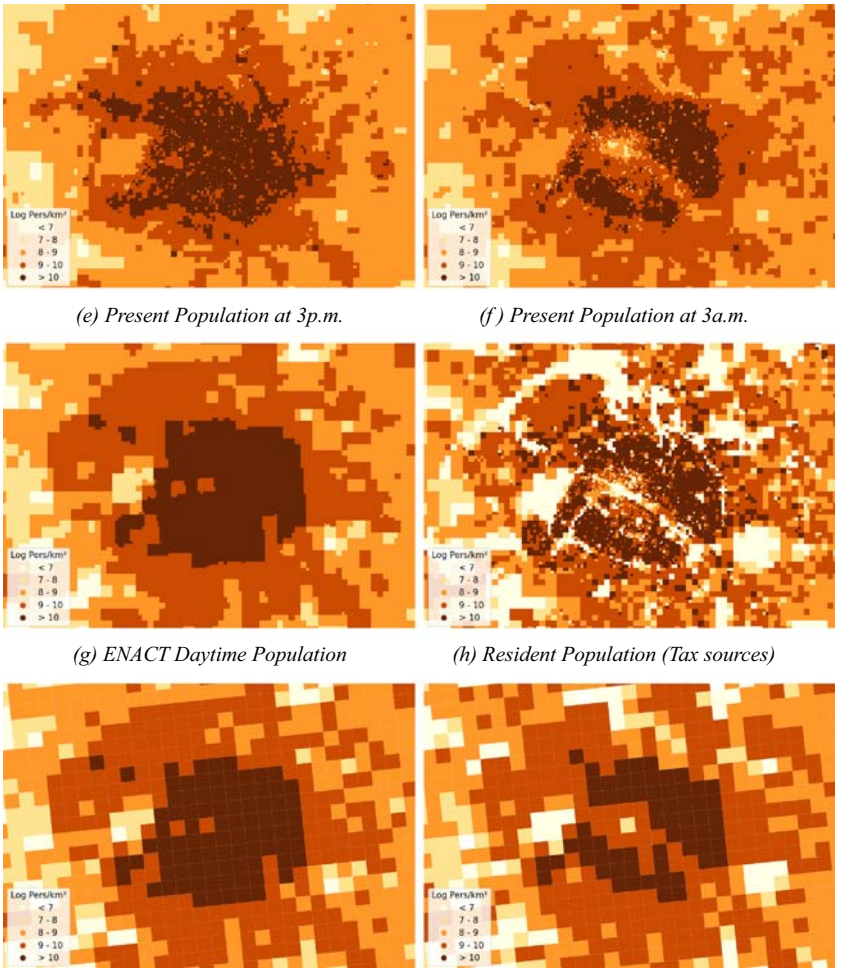
(c) ENACT Daytime Population *(d) Resident Population (Tax sources)*

Fig. 15. Population Densities in the Dissemination Grid (a-d). Present Population: March 2019 (week day). ENACT: March 2011. Resident Population: 2016.

variation from nighttime to daytime is similar if we consider either present populations as estimated from mobile network data from (f) to (e) or from disaggregation of NUTS3 official sources counts as obtained in ENACT data, from (j) to (i). For instance, the core center near the *Seine* river and the *Défense* neighborhood attract population during the day, as predicted from activity-related presence with the ENACT methodology. Figure 20 in the Appendix (Section 7) presents the differences between the present population statistics and both external sources over France. It makes the tendency of mobile network data estimates to create halos around dense areas clearer. It shows that at night in dense areas such as Paris, except for particular places such as parks, the error resembles a white noise (no tendencies over space to either overestimate or under-estimate the population compared to the resident population). During the day, it tends to offer even more contrast between places' densities than the ENACT estimates.

In addition, Table 4 reports comparisons along three metrics: correlation, rank correlation and allocation accuracy. Allocation accuracy can be interpreted as the percentage of population density allocated in the same tiles in both sources and is defined as:

$$\mathrm{AA}\left(\rho^1, \rho^0\right) = 1 - \sum_i \frac{\frac{1}{2} \times \left|\rho_i^0 - \rho_i^1\right|}{\sum_i \rho_i^0}.$$



(e) Present Population at 3p.m. *(f) Present Population at 3a.m.*

(g) ENACT Daytime Population *(h) Resident Population (Tax sources)*

(i) ENACT Daytime Population *(j) ENACT Nighttime Population*

Fig. 16. Population Densities in the Dissemination Grid (e-h). Present Population: March 2019 (week day). ENACT: March 2011, daytime (original grid 1km^2 in (i-j)). Resident Population: 2016.

The present population at nighttime is as close to the resident population as is the ENACT estimation during the night (slightly closer according to correlation and allocation accuracy – and farther according to rank correlation). Both present population measures get more distant from the resident population during the day, although the night/day difference is more pronounced in the MNO-derived estimation. Finally, the MNO-derived presence estimation and the ENACT presence estimation are closer to each other during both day and night than they are to the resident population.

Overall, at nighttime, MNO and ENACT densities fall in the same metrics range when compared with the resident population and are aligned with each other. If we had used a prior based on land uses for spatial mapping Q, we would probably be even closer to the ENACT estimation – which by definition follows land use.

Table 5 reports correlations with external sources by urban area type. Consistency between sources is highest in urban areas of intermediate size (between 200,000 and

Table 4. Population density comparisons. Note: All densities are computed in our dissemination grid, using proportional area estimation for ENACT data and direct calculations for resident density from tax files. ENACT: March 2011. Present population: five working days in March 2019. Resident population: 2016. Daytime (resp. nighttime) estimates are taken at 3 p.m. (resp. 3 a.m.).

	Allocation accuracy	Correlation	Rank correlation
Residents			
Present at daytime	0.57	0.56	0.73
Present at nighttime	0.74	0.83	0.75
ENACT – Day	0.66	0.77	0.78
ENACT – Night	0.71	0.78	0.81
ENACT – Day			
Present at daytime	0.74	0.75	0.89
ENACT – Night			
Present at nighttime	0.79	0.87	0.88

Table 5. Population density correlations by urban area type. Note: All densities are computed in our dissemination grid, using proportional area estimation for ENACT data and direct calculations for resident density from tax files. ENACT: March 2011. Present Population: March 2019 (five working days). Resident Population: 2016. Daytime (resp. nighttime) estimates are taken at 3 p.m. (resp. 3 a.m.).

	Paris urban area	Other urban areas $\geq$ 700,000 inhabitants	Urban areas 200,000–700,000 inhabitants
Residents			
Present at daytime	0.42	0.62	0.74
Present at nighttime	0.77	0.81	0.86
ENACT – Day	0.69	0.75	0.77
ENACT – Night	0.70	0.77	0.79
ENACT – Day			
Present at daytime	0.70	0.76	0.76
ENACT – Night			
Present at nighttime	0.81	0.84	0.83
	Urban areas 50,000–200,000 inhabitants	Urban areas $\leq$ 50,000 inhabitants	Municipalities outside of cities attraction
Residents			
Present at daytime	0.76	0.69	0.58
Present at nighttime	0.81	0.73	0.62
ENACT – Day	0.75	0.71	0.69
ENACT – Night	0.79	0.73	0.74
ENACT – Day			
Present at daytime	0.70	0.65	0.52
ENACT – Night			
Present at nighttime	0.76	0.70	0.59

700,000 inhabitants) and at night. The Paris urban area stands out as where the correlation between present population at daytime and resident population is the lowest (0.42) and is the most in contrast with the correlation between present population at nighttime and resident population (0.77). The correlation between MNO and ENACT densities is particularly low for the smaller urban areas and municipalities outside of cities attraction. This reflects that our methodology probably suffers in areas with bad network coverage where the spatial mapping is imprecise.

5. Discussion

5.1. Other Works on the Present Population Estimation for Official Statistics

These experimental present population estimates were built with knowledge and inspiration from a number of existing works (Salgado et al. 2021; Statistics Netherlands 2020; Ricciato et al. 2020; Ricciato and Coluccia 2021). However, we found that off-the-shelf solutions were never fully applicable to our case, and resorted to a number of simplifications which we discuss here.

Salgado et al. (2021) propose an ambitious bayesian general framework for producing statistics from mobile network data, based on a Hidden Markov Model (HMM) modelization for device-level trajectories. In contrast, we do not rely in this work on any inference framework to quantify the uncertainties in our final estimates. Of course, this is a downside, but the level of computational complexity entailed by resorting to this modelization seemed prohibitive in our context. Static spatial mapping on cell-level aggregates, as opposed to the dynamical spatial mapping at the device level delivered by the HMM model, was chosen to avoid the computational burden. One significant advantage of our method is that it gives the ability to vary the spatial mapping after the computationally intensive aggregation step, as various spatial mappings have been shown to provide quite different results. On a similar signaling data set restricted to a large urban area, Bonnetain et al. (2019) resort to a hidden Markov chain modelization for map-matching device trajectories on the transportation network, but simplify the cell spatial coverage information to a Voronoï tesselation. The computational complexity of resorting to simple temporal interpolation has nothing to do with setting up a HMM estimation in a high-dimensional states space (up to 55 million tiles), emissions probabilities (connecting these millions of tiles to hundreds of thousands of cells), and devices (about twenty millions three-months trajectories). Although the problem is in theory parallelizable at the device level, the single device problem can be quickly high dimensional in space and time.

One of the strengths of the HMM model is to probabilistically recover trajectories when the device is unobserved from future and past observations. Given Figure 1, it is a guarantee against network and behavioral effects which seems highly desirable. We see this figure as urging for longitudinal views to derive sensible statistics. We therefore resorted to a simple interpolation method. Only a few works mention interpolation as a key feature for deriving reliable present population statistics, whereas we tend to consider interpolation as essential for sustainable and comparable-in-time statistics. Ricciato et al. (2020) point it has a promising line for future research. Interpolation techniques were explored mostly for mobility analysis. The issues of data time sparsity and sensitivity to user behavior are generic, and we show that they apply as well when working with

signaling data on a present population estimation use case. Up to now, existing literature derived snapshots of "dynamic population" (e.g., within a given day) due to the lack of access to longitudinal data for research. Throughout the lockdown period imposed in France during the Covid-19 crisis, some methods used by MNO showed sensitivity to changes in behavior (increased usage of the phone, change in the timing of usage). Roughly, the increased presence over the network translated into more detected devices and therefore to unexpected large mechanical increases in present population estimates. The French National Statistical Institute computed estimates based on reprocessed multi-MNO data (Galiana et al. 2020b; Coudin et al. 2021). In addition, interpolation has a positive effect on representativeness if the extent of network detection is correlated with socio-economic background: less active users (or users having access to a less performant mobile phone or local network) contribute to aggregates more equally after interpolation relatively to more active users. Network usage has indeed been shown related to individual characteristics, such as gender (Jahani et al. 2017).

CBS (Statistics Netherlands) has recently published a report on its methodology (Statistics Netherlands 2020) which shares similarities with our implementation. Working on signaling data of one MNO in the Netherlands, they integrate a device presence estimation with a residency detection module. They perform a home-cell detection step, where the computation barrier appears to be important as well – and based on a similar heuristic. Their calibration step is based on rescaling the estimated number of active devices to the number of local residents independently of their places of presence. Implicitly, the minimal consistency constraints are therefore the same as the ones we impose. One difference is that Statistics Netherlands (2020) does not interpolate device-level trajectories. Only the calibration step ensures the consistency constraints.

Ricciato et al. (2020) stress the practical importance of the geolocation step. In this work as here, the geolocation step is performed after cell-level aggregation. Ricciato and Coluccia (2021) propose several classes of estimators based on matrix A and cell-level aggregates which could be considered in place of our bayesian spatial mapping, for instance, to deliver confidence intervals and improve spatial accuracy. In this article, the adopted linear estimator corresponds to the "Simple Bayes-rule estimator" described in Ricciato and Coluccia (2021) applied to reweighted device counts. Our framework, which introduces device-level weights and a home estimation, could easily be integrated with the other estimators introduced in Ricciato and Coluccia (2021) by replacing the simple count vector c with a reweighed version.

5.2. Discussion on Device-Level Weights

We argue that residency detection is not only useful for statistical filtering, but also to balance the estimates across residencies to correct for unbalanced representativeness (as illustrated in Figures 6 and 10). Imposing a constraint of equality at a local level between "usual" residents detected from mobile-phone data and actual residents as estimated from official sources appears milder than an alternative that would consist of equalizing resident population P_r to population present at night: $u_{r,t_0} = P_r$. A significant part of the population spends nights regularly outside of its main residency or works at night. These atypical location behaviors are better captured by MNO data than by traditional sources. Figure 14

shows how the present population at night can vary greatly during the week – and it is also even more the case for holidays and bank holidays. Note that if the pseudo-weights are based only on residency location, they could be based on as many characteristics as we can accurately recover at the device level and for which we have an external population-level estimate (see the discussion on pseudo-weights in Beresewicz et al. (2018)). This framework could be promising for future combinations of sources aiming at facing selectivity issues to derive representative statistics. To support this statement, we report in Table 6 how the consistency between our estimate and ENACT estimate drops when we do not use individual residency-based weights but simply multiply the device density by an hourly constant to match the total population size. This consistency decreases noticeably both in terms of allocation accuracy and correlation, at night and daytime. Rank correlation comparisons are similar, suggesting that rank ordering across tiles is rather close when using a constant weighting scheme.

Using device-level weights based on individual characteristics may lift part of the concern of using a single MNO data set. If mobility patterns of the MNO customers are similar to that of the resident populations which share these same characteristics, the estimation would be unbiased. Of course, the larger and the more representative the mobile phone users used in the analysis, the better will be the quality of the estimates. Therefore, multiple MNOs data sets could be fruitfully employed with a similar methodology. Device-level weights are obtained from merging MNO data and reliable and localized estimates of the resident populations. The latter is more and more commonly produced for a large number of countries. For instance in Europe, population density have been disseminated on a common European 1 km^2 reference grid in 2011 as a prototype (GEOSTAT-1) and a legal act was adopted for Census 2021 for an harmonised publication of key census topics on an EU-wide 1 km^2 grid by the European Statistical System.

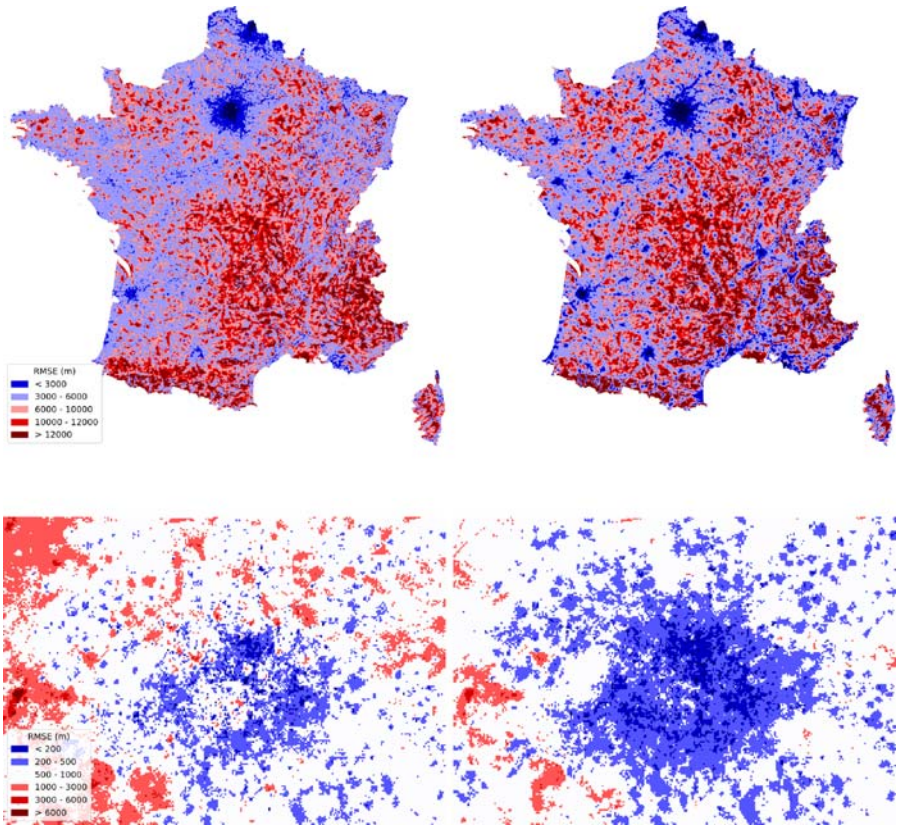
5.3. *Quantifying Uncertainty*

This work does not provide an end-to-end variance estimator for our final estimates but analyzes uncertainties at various stages while building the estimates, which will

Table 6. Comparisons between the ENACT estimate and both our present population estimate as well as globally reweighted device densities. Note: All densities are computed in our dissemination grid, using proportional area estimation for ENACT data and direct calculations for resident density from tax files. ENACT: March 2011. Present population and rescaled device densities: Five working days in March 2019. Resident Population: 2016. Rescaled devices estimates are obtained when omitting reweighting by residency and rescaling device densities hour by hour with a single scalar weight. Daytime (resp. nighttime) estimates are taken at 3 p.m. (resp. 3 a.m.).

	Allocation accuracy	Correlation	Rank correlation
ENACT – Day			
Present at daytime	0.74	0.75	0.89
Rescaled devices at daytime	0.68	0.66	0.89
ENACT – Night			
Present at nighttime	0.79	0.87	0.88
Rescaled devices at nighttime	0.73	0.78	0.88

hopefully help future theoretical development. We estimate the spatial mapping accuracy locally through the probability to localize in i a device that is in i_0, N_{i,i_0}, where $N = QA$. With previous notations, if we take the example of localizing a single device d which is in i_0, that is $\mathbf{u}_t = 1_{i_0}, N1_{i_0}$ can be interpreted as $\mathbb{E}[\hat{\mathbf{u}}_t|\text{device in tile } i_0]$. Thus, $\|N1_{i_0} - 1_{i_0}\|$ provides a local evaluation of the spatial accuracy: it is the root mean square error of the spatialisation task of a device which is in i_0. We provide in Figure 17 maps of this spatial uncertainty based on this framework, with Q a spatial mapping base on a uniform or a resident density prior. Dense areas have a small spatial error compared to rural areas, as could be expected from the network density. In addition, using a prior helps in reducing this spatial uncertainty. As for the temporal uncertainties, with the large time event rate, we find that interpolation errors are limited for hourly estimates of present population (Figure 12). Finally, part of population coverage uncertainties may be captured through the weights $w_r = \frac{P_r}{D_r}$, as this ratio informs on differential local representativity. Although informative, a map of w_r also divulgates Orange market shares and is thus not reported.



(a) Uniform Prior *(b) Resident Density Prior*

Fig. 17. RMSE of the spatial mapping, with a linear spatial mapping with an uniform or resident density prior.

5.4. Other Limitations

In addition to the limitations discussed above, some issues were left aside in this work.

First, we hypothesize that a person carries a single device and we did not attempt to filter out potential additional devices. The issue is probably less salient than when relying on multiple MNO data, but even in a single MNO framework, identifying from the data devices owned by the same person can be a complex task. In our work, without additional filtering, the hourly population is biased toward the mobility of persons carrying multiple devices with an Orange contract which seemlingly represent multiple persons.

Second, our methodology should be adapted next to border areas – as we only observe the Orange network on the French territory. We here detail the consequences of daily worker flows, but note that similar issue could be described for touristic outflows and location where devices are last seen over the French territory (such as airport). Take the example of daily outflows of cross-border workers leaving the French territory in the morning and coming back at night. Due to the interpolation method (device presence panel step), the trajectories of devices carried by these workers will stop and remain at the border in the meantime. Thus, present population will be overestimated in tiles next to the border during the day. As for cross-border inflows, the presence of foreign residents carrying an Orange device with a French Orange contract will be interpolated to cells next to the border at night. In addition, their attributed residence will most likely be their workplace cells. This complex subject should be examined in detail in future work.

6. Conclusion

We derived a prototype for an experimental present population statistic for France. Daily and weekly cycles, both at the local and national levels offer unprecedented insight into population dynamics in France. However, to develop a full-fledged methodology, access is of utmost importance. A legal basis for processing MNO data for official statistics under due privacy protection, as well as the cooperation of MNOs, is of primary importance. It is all the more challenging today that this work tends to demonstrate that some form of access to longitudinal individual data would be needed for deriving reliable population estimates, even aside from an interest in mobility analysis (interpolation to avoid being plagued with activity bias, home detection or device-level characterization to ensure representativeness, deduplication...). Data management from the MNO side seems decisive for the quality of the final statistics (network topology modelization and cell register management, filtering of IoT and M2M, reports on data collection failure...). A combination of sources from the MNO and the NSI sides at a fine-grained level (e.g., build weights for representativeness) seems very promising. Data fusion approaches, merging high-quality population and mobile network data sets could be considered and studied in the future to make the most of both sources. Nonetheless, the combination of sources requires cooperation, privacy-preserving transfer of information, and a transparent sharing of computations – which has been at this stage possible only within research projects, if at all, in European countries.

7. Appendix

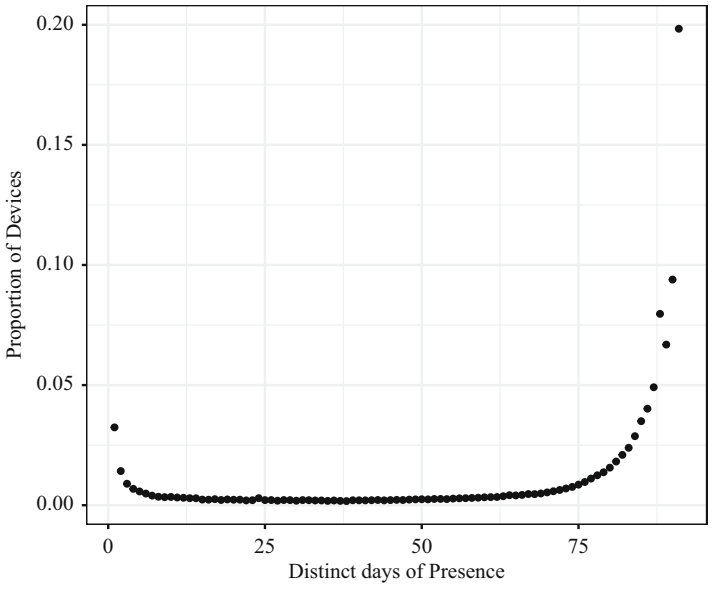


Fig. 18. Distinct dates of observation over the period.

Table 7. Dissemination grid structure and average resident density. In our dissemination grid, presented in the section "Embedding accuracy within dissemination", and illustrated in Figure 11, there are 4,062 tiles of 200 meters square. The average resident density in these tiles is 15,810 persons per square kilometer (source: Filosofi).

Tile side (meters)	Resident density (persons/km^2)	Number of tiles
100	22970	2816
200	15810	4062
400	5720	8796
800	1670	12721
1600	280	20743
3200	60	28339
6400	30	5024
12800	10	2

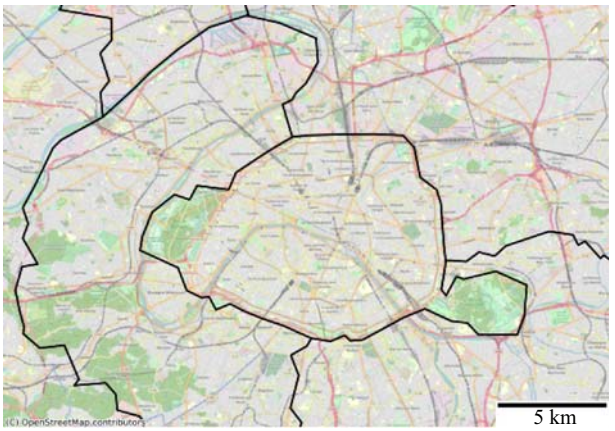
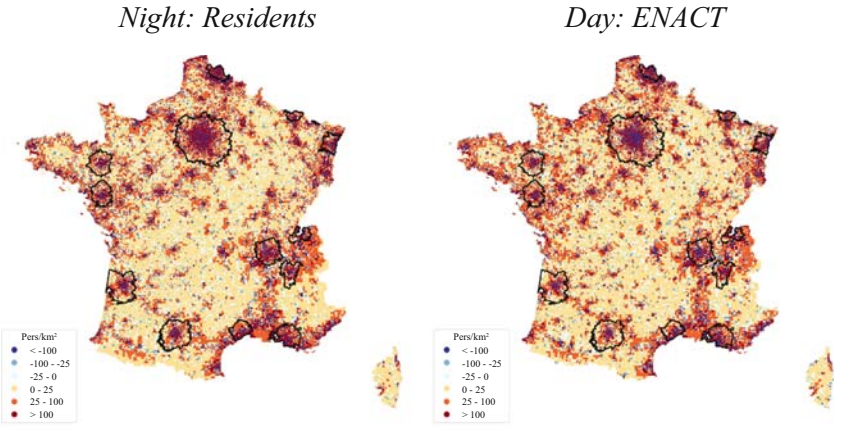
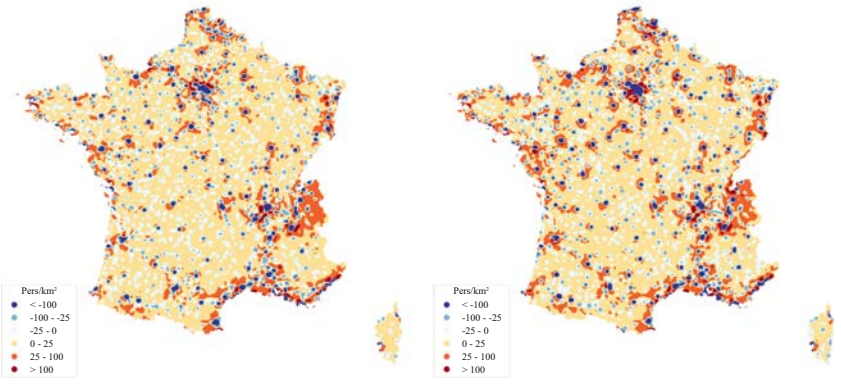


Fig. 19. Paris area context, with Département administrative boundaries.



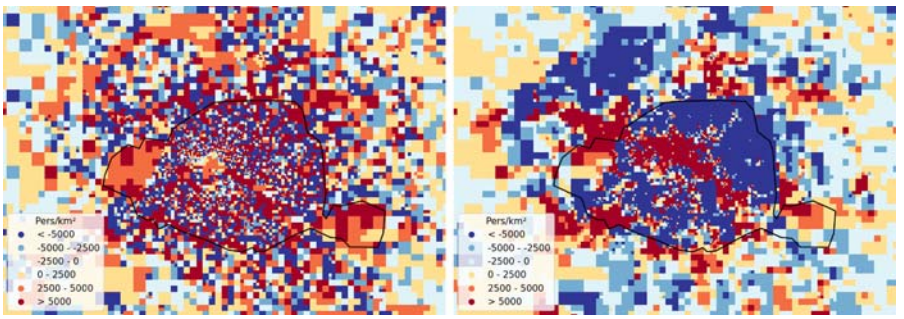
(a) Differences Present Population - External Source Population



(b) Smoothed Differences Present Population - External Source Population

Fig. 20. Differences between present population at 3 a.m. and resident population (left) and between present population at 3 p.m. and ENACT day time population (right). The first panel shows urban attraction areas with more than 700,000 inhabitants.

Night: Residents *Day: ENACT*



(c) Differences Present Population - External Source Population

Fig. 21. Differences between present population at 3 a.m. and resident population (left) and between present population at 3 p.m. and ENACT day time population (right).

Table 8. Notations used throughout the document.

$d \in D$	Devices in the scope, detected on the network over the period of interest
$t \in T$	Hourly time grid of interest (several weeks)
P	The target population set
D_t	Set of devices in the scope, detected on the network during t
D_l	Set of devices in the scope, detected on the network at date l
$i \in I$	Tiles that cover the territory of interest
$j \in J$	Cells of the MNO network
$j_{d,t} \in J$	Estimated *presence cell* for device d during time interval t
P_r	Official residents in place r
D_r	Devices with estimated residency in r
$\mathbf{u}_t \; in \mathbb{R}^{\lvert I \rvert}$	Population counts at time t over tiles (target statistics)
$\mathbf{m}_{r,t} \in \mathbb{N}^{\lvert J \rvert}$	Devices count in *presence cells* at time t with residency r
$\mathbf{u}_0 \in \mathbb{N}^{\lvert I \rvert}$	Population count from official source over tiles
$\hat{\mathbf{u}}_r \in \mathbb{R}^{\lvert I \rvert}$	Estimated population count at time t over tiles
$\hat{\mathbf{u}}_{r,t} \in \mathbb{R}^{\lvert I \rvert}$	Estimated population count at time t who are resident in r over titles
w_r	Pseudo-weight of devices residing in place r
$w_{r,l}$	Final weight of devices in place r on date l
A	Coverage probability matrix
$Q : \mathbb{R}^{\lvert J \rvert} \in \mathbb{R}^{\lvert I \rvert}$	Linear spatial mapping from cells to tiles
N_{i,i_0}	Probability that a device in tile i_0 is mapped in tile i

8. References

Avouac, R., B. Sakarovitch, and Z. Smoreda. 2019. "A Bayesian Approach to Improve the Estimation of Population using Mobile Phone Data." Research Workshop on Digital Demography in the Era of Big Data, Seville, Spain, 6–7 June 2019. Available at: https://iussp.org/en/digital - demography - era - big - data (accessed March 2023).

Batista e Silva, F., S. Freire, M. Schiavina, K. Rosina, M.A. Marìn-Herrera, L. Ziemba, M. Craglia, E. Koomen, and C. Lavalle. 2020. "Uncovering Temporal Changes in Europe's Population Density Patterns using a Data Fusion Approach." *Nature communications* 11(1): 1–11. DOI: https://doi.org/10.1038/s41467-020-18344-5.

Beresewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio, and M. Karlberg. 2018. *An Overview of Methods for Treating Selectivity in Big Data Sources*. Technical report. Eurostat Statistical Working Article. Available at: https://doi.org/10.2785/312232.

Blondel, V.D, A. Decuyper, and G. Krings. 2015. "A Survey of Results on Mobile Phone Data Sets Analysis." *EPJ data science* 4(1): 10. DOI: https://doi.org/10.1140/epjds/s13688-015-0046-0.

Bonnetain, L., A. Furno, N.-E. el Faouzi, M. Fiore, R. Stanica, Z. Smoreda, and C. Ziemlicki. 2021. "TRANSIT: Fine-Grained Human Mobility Trajectory Inference at Scale with Mobile Network Signaling Data." *Transportation Research Part C: Emerging Technologies* 130: 103257. DOI: https://doi.org/10.1016/j.trc.2021.103257.

Bonnetain, L., A. Furno, J. Krug, and N.-E. el Faouzi. 2019. "Can We Map-Match Individual Cellular Network Signaling Trajectories in Urban Environents? Data-Driven Study." *Transportation Research Record* 2673(7): 74–88. DOI: https://doi.org/10.1177/0361198119847472.

Chen, G., A. Carneiro Viana, M. Fiore, and C. Sarraute. 2019. "Complete Trajectory Reconstruction from Sparse Mobile Phone Data." *EPJ Data Science* 8(1): 30. DOI: https://doi.org/10.1140/epjds/s13688-019-0206-8.

Coudin, E., M. Poulhes, and M. Suarez Castillo. 2021. "The French Official Statistics Strategy: Combining Signaling Data from Various Mobile Network Operators for Documenting COVID-19 Crisis Effects on Population Movements and Economic Outlook." *Data & Policy* 3. DOI: https://doi.org/10.1017/dap. 2021.1.

ESSnet Big Data, II. 2021. *Work Package I – Mobile Networks Data Deliverables*. Available at: https://ec.europa.eu/eurostat/cros/content/wpi-milestones-and-deliverables_en. (accessed March 2023).

Fekih, M., T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, and S. Galland. 2021. "A Data-Driven Approach for Origin – Destination Matrix Construction from Cellular Network Signalling Data: a Case Study of Lyon Region (France)." *Transportation* 48(4): 1671–1702. DOI: https://doi.org/10.1007/s11116-020-10108-w.

Freire, S., and C. Aubrecht. 2012. "Integrating population dynamics into mapping human exposure to seismic hazard." *Natural Hazards and Earth System Sciences* 12(11): 3533–3543. DOI: https://doi.org/10.5194/nhess-12-3533-2012.

Galiana, L., B. Sakarovitch, F. Sémécurbe, and Z. Smoreda. 2020. *Residential Segregation, Daytime Segregation and Spatial Frictions: an Analysis from Mobile Phone Data, Insee Working article G2020-12*. Available at: https://www.insee.fr/en/-statistiques/4925202.

Galiana, L., M. Suarez Castillo, F. Sémécurbe, E. Coudin, and M.-P. de Bellefon. 2020b. *Retour Partiel des Mouvements de Population avec le Déconfinement, Insee Analyses 54*. Available at: https://www.insee.fr/fr/statistiques/4635407.

Hoteit, S., G. Chen, A. Viana, and M. Fiore. 2016. "Filling the Gaps: On the Completion of Sparse Call Detail Records for Mobility Analysis." In Proceedings of the eleventh ACM workshop on challenged networks: 45–50. DOI: https://doi.org/10.1145/2979683.2979685.

Insee. 2020. *Population présente sur le territoire avant et aprés le début du confinement – Premiers résultats (communiqué de presse)*. Available at: https://www.insee.fr/fr/information/4477356 (accessed March 2023).

Jahani, E., P. Sundsoy, J. Bjelland, L. Bengtsson, A. Pentland, and Y.-A. de Montjoye. 2017. "Improving Official Statistics in Emerging Markets using Machine Learning and Mobile Phone Data." *EPJ Data Science* 6(1): 3. DOI: https://doi.org/10.1140/epjds/s13688-017-0099-3.

Katsikouli, P., M. Fiore, A. Furno, and R. Stanica. 2019. "Characterizing and Removing Oscillations in Mobile Phone Location Data." In 2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM): 1–9. IEEE. DOI: https://doi.org/10.1109/WoWMoM.2019.8793034.

Lamarche, P., and S. Lollivier. 2021. Fidéli, l'intégration des sources fiscales dans les données sociales *Courrier des statistiques* 6. Available at: https://www.insee.fr/fr/information/5398683 (accessed March 2023).

Léetroublon, C., and C. Daniel. 2018. *Le Travail en Horaires Atypiques: Quels Salariés pour quelle Organisation du Temps de Travail? Dares Analyses 2018-30*. Available at: https://dares.travail-emploi.gouv.fr/publications/le-travail-en-horaires-atypiques. (accessed March 2023).

Panczak, R., E. Charles-Edwards, and J. Corcoran. 2020. "Estimating Temporary Populations: a Systematic Review of the Empirical Literature." *Humanities and Social Sciences Communications* 6(1): 1–10. DOI: https://doi.org/10.1057/s41599-020-0455-y.

Ricciato, F., and A. 2021. "On the Estimation of Spatial Density from Mobile Network Operator Data." *IEEE Transactions on Mobile Computing*, 22(6): 1. DOI: https://doi.org/10.1109/TMC.2021.3134561.

Ricciato, F., G. Lanzieri, A. Wirthmann, and G. Seynaeve. 2020. "Towards a Methodological Framework For Estimating Present Population Density from Mobile Network Operator Data." *Pervasive and Mobile Computing* 101263. DOI: https://doi.org/10.1016/j.pmcj.2020.101263.

Ricciato, F., P. Widhalm, F. Pantisano, and M. Craglia. 2017. "Beyond the "Single-Operator, CDR-only Paradigm: An Interoperable Framework for Mobile Phone Network Data Analyses and Population Density Estimation." *Pervasive and Mobile Computing* 35: 65–82. DOI: https://doi.org/10.1016/j.pmcj.2016.04.009.

Ricroch, L., and B. Roumier. 2011. *Depuis 11 Ans, Moins de Ŝaches Ménagères, Plus d'Internet. Insee premiére 1377*. Available at: https://www.insee.fr/fr/statistiques/1281050. (accessed March 2023).

Sakarovitch, B., M.-P. de Bellefon, P. Givord, and M. Vanhoof. 2018. "Estimating the Residential Population from Mobile Phone Data, an Initial Exploration." *Economie et Statistique* 505(1): 109–132. DOI: https://doi.org/10.24187/ecostat.2018.505d.1968.

Salat, H., Z. Smoreda, and M. Schläpfer. 2020. "A Method to Estimate Population Densities and Electricity Consumption from Mobile Phone Data in Developing Countries." *PLOS ONE* 15(6): 1–11. DOI: https://doi.org/10.1371/journal.pone.0235224.

Salgado, D., L. Sanguiao, B. Oancea, S. Barragàn, and M. Necula. 2021. "An End-to-End Statistical Process with Mobile Network Data for Official Statistics." *EPJ Data Science* 10(1): 1–46. DOI: https://doi.org/10.1140/epjds/s13688-021-00275-w.

Schiavina, M., S. Freire, K. Rosina, L. Ziemba, M. Marin Herrera, M. Craglia, C. Lavalle, T. Kemper, and F. Batista. 2020. *ENACT-POP R2020A-ENACT 2011 Population Grid*.

European Commission, Joint Research Centre (JRC). DOI: https://doi.org/10.2905/-BE02937C-5A08-4732-A24A-03E0A48BDCDA.

Statistics Netherlands, C.B.S. 2020. *Estimating Hourly Population Flows in the Netherlands*. Avalialble at: https://www.cbs.nl/en-gb/longread/diversen/2020/pilot-study-mobile-phone-meta-data-records-introduction-to-the-research-method (accessed March 2023).

Tennekes, M., and Y. Gootzen. 2022. "Bayesian location estimation of mobile devices using a signal strength model." *Journal of Spatial Information Science* 25: 29–66. DOI: https://doi.org/10.5311/JOSIS.2022.25.166.

U.N. Global Working Group. 2019. *Handbook on the Use of Mobile Phone Data for Official Statistics*. Technical report. United Nations. Available at: https://unstats.un.org/bigdata/task-teams/mobile-phone/MPD%20Handbook%2020191004.pdf.

Vanhoof, M., F. Reis, T. Ploetz, and Z. Smoreda. 2018. "Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics." *Journal of Official Statistics* 34(4): 935–960. DOI: https://doi.org/10.2478/jos-2018-0046.

Received March 2022
Revised August 2022
Accepted March 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 571–590, http://dx.doi.org/10.2478/JOS-2023-0026

Application of Sampling Variance Smoothing Methods for Small Area Proportion Estimation

Yong You[1] *and Mike Hidiroglou*[1]

Sampling variance smoothing is an important topic in small area estimation. In this article, we propose sampling variance smoothing methods for small area proportion estimation. In particular, we consider the generalized variance function and design effect methods for sampling variance smoothing. We evaluate and compare the smoothed sampling variances and small area estimates based on the smoothed variance estimates through analysis of survey data from Statistics Canada. The results from real data analysis and simulation study indicate that the proposed sampling variance smoothing methods perform very well for small area estimation.

Key words: Coefficient of variation; design effect; generalized variance function; log-linear model; relative error.

1. Introduction

Small area estimation has become very popular and important in both public and private agencies due to the growing demand for reliable small domain estimates. Small area estimation is based on models that borrow strength across areas and combine different sources of information in order to obtain reliable estimates. In this article, we focus on area level models that are based on direct survey estimates aggregated from the unit level data and area level auxiliary variables. Various area level models have been proposed in the literature to improve the precision of the direct survey estimates: a good summary of these methods is discussed in Rao and Molina (2015). The Fay-Herriot model (Fay and Herriot 1979) is a basic area level model that is widely used in practice. The Fay-Herriot model has two components, namely, a sampling model for the direct survey estimates and a linking model for the small area parameters of interest. The sampling model assumes that there exists a direct survey estimator y_i, which is usually design unbiased, for the small area parameter θ_i such that

$$y_i = \theta_i + e_i, \quad i = 1, ..., m, \tag{1}$$

where e_i is the sampling error associated with the direct estimator y_i and m is the number of small areas. It is customary in practice to assume that the e_i's are independently distributed normal random variables with mean $E(e_i) = 0$ and sampling variance $Var(e_i) = \sigma_i^2$. The linking model assumes that the small area parameter of interest θ_i is related to area level auxiliary variables $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})'$ through a linear regression model

$$\theta_i = \boldsymbol{x}_i{}'\boldsymbol{\delta} + v_i, \quad i = 1, ..., m, \tag{2}$$

[1] Statistics Canada, Ottawa, K1A 0T6, Canada; Emails: yongyou@statcan.gc.ca and hidirog@yahoo.ca.
Acknowledgments: We would like to thank the Editors-in-Chief, one Associate Editor and four referees for their constructive comments and suggestions that help us to improve the article substantially

where $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)'$ is a $p \times 1$ vector of regression coefficients, and v_i's are area-specific random effects assumed to be independent and identically distributed with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$.

The assumption of normality for v_i is generally also included. The model variance σ_v^2 is unknown and needs to be estimated from the data. For the Fay-Herriot model, the sampling variance σ_i^2 is assumed to be known in model (1). As this is a very strong assumption, a smoothing or modeling approach is usually used to estimate σ_i^2. The sampling variance can be smoothed or can be modeled directly as in Wang and Fuller (2003), You and Chapman (2006), Maples et al. (2009), Dass et al. (2012), Sugasawa and Kubokawa (2017), Sugasawa et al. (2017), Ghosh et al. (2018), and so on. It is also shown in You (2021) that the smoothing approach can provide more efficient and accurate model-based estimates than the modeling approach for small areas under the hierarchical Bayes framework. Lesage et al. (2021) also have some discussions on the sampling variance smoothing for the Fay-Herriot model.

The objective of this article is to compare different methods that smooth the direct estimates of the sampling variances for proportions in small area estimation using the Fay-Herriot model. Let $\hat{p}_{iw}$ be the direct design-based estimator for the proportion p_i for a given characteristic in the i-th area. Applying the Fay-Herriot model to $\hat{p}_{iw}$, we have

$$\hat{p}_{iw} = p_i + e_i, \tag{3}$$

where the sampling variance $Var(e_i) = \sigma_i^2$ is unknown. Now let $\hat{V}_i$ be a direct sampling variance estimator for σ_i^2 obtained from the survey data. Usually some of the $\hat{V}_i$'s are very unstable due to small sample sizes. We, therefore, need to smooth the sampling variance estimate, $\tilde{V}_i$, and then treat the resulting smoothed variance estimate $\tilde{V}_i$ in the sampling model (3) as known.

In this article, we compare two smoothing methods. One method is based on the generalized variance function (GVF, see, e.g., Wolter 2007), and the other one is based on design effects (DEFF). We then propose an average smoothed (ASM) variance estimator that combines the GVF and DEFF smoothed estimators. The main purpose of the article is to promote the proposed GVF and DEFF methods. The ASM is used as an additional choice as it pools the GVF and DEFF estimates by taking their average. Recently Hirose et al. (2023) proposed a variance stabilizing transformation for small area estimation of proportions. They used the arc-sin transformation approach to construct a Fay-Herriot model with known sampling variance. Note that their transformation approach is applied to data assumed to have a binomial distribution, and this implies simple random sampling (SRS). However, many surveys are based on various complex sampling designs. The sampling variance smoothing approach can be applied to any estimator of proportions that is based on complex designs. We re-estimate the variance via a smoothing process that involves all small area estimates of the sampling variances.

There are many applications of the GVF in small area estimation, see, for example, the early work of Dick (1995) and the recent application in Hidiroglou et al. (2019). The DEFF can also be used in variance modeling and smoothing for small area estimation. For example, You (2008) used the smoothed design effects over time to obtain the smoothed variance and covariance matrices. Liu et al. (2014) also applied area level models to proportions using design effects for the sampling variance smoothing and modeling.

In this article, we provide a general method to compute the design effect and propose a smoothed variance estimator based on the average design effects over areas. We will also show that the DEFF-smoothed variance estimator and the GVF-smoothed variance estimator are roughly equivalent under certain conditions. We illustrate the smoothing methods via various survey data sets and a simulation study.

The article is organized as follows. In section 2, we propose several sampling variance smoothing procedures that include the GVF and DEFF methods. In section 3, we apply the proposed smoothing methods to two Statistics Canada survey data and compare the smoothing variances. In section 4, we compare the model-based estimates based on different smoothed sampling variance estimates using the Canadian Labor Force Survey (LFS) survey data. Section 5 is a simulation study that compares proposed variance smoothing estimators. In section 6, we offer some concluding remarks and suggestions.

2. Sampling Variance Smoothing Methods

2.1. Smoothing Using Log-Linear Models

In this section, we will construct a GVF model to obtain smoothed sampling variances. This procedure is widely used in practice to model the variance. We apply a log-linear regression model to the direct sampling variance $\tilde{V}_i$ using the sample size n_i as the auxiliary variable in the model as follows:

$$log(\hat{V}_i) = \beta_0 + \beta_1 log(n_i) + \varepsilon_i, \quad i = 1, ..., m, \tag{4}$$

where the model error term is $\varepsilon_i \sim N(0, \tau^2)$, and the model error variance τ^2 is unknown. Note that the proposed regression model (4) is equivalent to the following model:

$$log(\hat{V}_i) = \beta_0 + \beta_1 log(n_i^{-1}) + \varepsilon_i, \quad i = 1, ..., m, \tag{5}$$

where $log(n_i^{-1})$ is used as the auxiliary variable. The proposed GVF models (4) or (5) are the same models used in You (2021) for the hierarchical Bayes (HB) modeling of sampling variance. This GVF model also extends the model proposed by Souza et al. (2009) for sampling variances by using $log(n_i^{-1})$ and adding a normal random effect (ε_i) to the regression part in the model.

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the ordinary least square estimators of the regression coefficients β_0 and β_1. A naïve GVF-smoothed estimator of the sampling variance is obtained by taking the exponential of the fitted value:

$$\tilde{V}_i^{naive} = exp(\hat{\beta}_0 + \hat{\beta}_1 log(n_i)). \tag{6}$$

Dick (1995) used the naïve smoothed estimator $\tilde{V}_i^{naive}$ in the application of census undercoverage small area estimation. As noted by Rivest and Belmonte (2000), the naïve smoothed estimator $\tilde{V}_i^{naive}$ underestimates the sampling variance. This can be seen as follows. If Y is a log-normal random variable with mean μ and variance τ^2, the mean of Y is $E(Y) = exp(\mu)exp(\tau^2/2)$. It follows that the smoothed estimator $\tilde{V}_i^{naive}$ underestimates the true values by ignoring the second term $exp(\tau^2/2)$ in the mean of the log-normal random variable. Denote as $\hat{\omega}_{RB} = exp(\hat{\tau}^2/2)$ the Rivest and Belmonte (2000) correction,

where $\hat{\tau}^2$ is the estimated residual variance of the proposed log-linear regression model (4). Then a GVF-smoothed estimator, denoted as $\tilde{V}_i^{GVF.RB}$, is given by

$$\tilde{V}_i^{GVF.RB} = \tilde{V}_i^{naive} \cdot \hat{\omega}_{RB} = \tilde{V}_i^{naive} \cdot exp(\hat{\tau}^2/2). \tag{7}$$

The naïve GVF estimator $\tilde{V}_i^{naive}$ in Equation (6) underestimates the sampling variance by $exp(\hat{\tau}^2/2)$. This term is always greater than 1, and sometimes it can be large, depending on the value of $\hat{\tau}^2$.

Hidiroglou et al (2019) proposed another correction term for the naïve estimator $\tilde{V}_i^{naive}$. Let $\tilde{V}^{naive}$ be the sum of the naïve smoothed variance estimators, that is, $\tilde{V}_i^{naive} = \Sigma_{i=1}^m \tilde{V}_i^{naive}$, and $\tilde{V}^{total}$ be the sum of the direct sampling variances, that is, $\tilde{V}^{total} = \Sigma_{i=1}^m \tilde{V}_i$. Following Hidiroglou et al. (2019), we define a correction term, $\hat{\omega}_{HBY} = \tilde{V}^{total}/\tilde{V}^{naive}$, named as the Hidiroglou, Beaumont, and Yung (HBY) correction term. This leads to a second GVF-smoothed variance estimator, denoted as $\tilde{V}_i^{GVF.HBY}$. It is given by

$$\tilde{V}_i^{GVF.HBY} = \tilde{V}_i^{naive} \cdot \hat{\omega}_{HBY} = \tilde{V}_i^{naive} \cdot \frac{\tilde{V}^{total}}{\tilde{V}^{naive}} \tag{8}$$

Note that $\hat{\omega}_{HBY}$ is obtained as an alternative estimator to $exp(\tau^2/2)$ using the method of moments (Beaumont and Bocci 2016). This avoids the sensitivity of the GVF model to deviations from the normality assumption of ε_i in model (4). The HBY correction term is also equivalent to the so-called smearing estimator, see Duan (1983). A nice property of $\tilde{V}_i^{GVF.HBY}$ is that the average of the smooth variance estimates is equal to the average of the direct sampling variance estimates, that is,

$$\frac{1}{m}\sum_{i=1}^{m} \tilde{V}_i^{GVF.HBY} = \frac{1}{m}\sum_{i=1}^{m} \hat{V}_i.$$

This property may ensure that the smoothing procedure does not systematically overestimate or underestimate the sampling variances.

2.2. *Smoothing Using Design Effects*

Let $\hat{p}_{iw}$ be the direct design-based estimate for a proportion p_i and $\hat{V}_i$ the corresponding direct sampling variance under complex design for the i-th small area. Then the estimated design effect can be approximately computed as

$$def\ f_i = \frac{\hat{V}_i}{\hat{p}_{iw}(1-\hat{p}_{iw})/n_i + \hat{V}_i/n_i}, \tag{9}$$

where n_i is the sample size of the i-th small area; see Gambino (2009, 143), Remark iii for a more detailed discussion of the special case 0-1 variables. Noting that the $def\ f_i$ in Equation (9) is not equal to 1 under simple random sampling design, we modify the $def\ f_i$ by multiplying it by a correction term $(n_i+1)/n_i$:

$$def\ f_i = \frac{\hat{V}_i}{\hat{p}_{iw}(1-\hat{p}_{iw})/n_i + \hat{V}_i/n_i} \cdot \frac{n_i+1}{n_i}. \tag{10}$$

Using Equation (10), we can re-write the design-based sampling variance $\hat{V}_i$ as

$$\hat{V}_i = def\ f_i.\frac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i}.\left(1+\frac{1-deff_i}{n_i}\right)^{-1}. \tag{11}$$

If the sample size n_i is large, the term $(1 - def\ f_i)/n_i$ may be negligible in Equation (11) so that the Equation (11) reduces to

$$\hat{V}_i = def\ f_i.\frac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i}. \tag{12}$$

Equation (12) is used, for example, in Liu et al. (2014) for sampling variance smoothing and modeling. However, in small area estimation, n_i can be very small, and the term $(1 - def\ f_i)/n_i$ may not be negligible.

We can compute all the design effects $def\ f_i$'s using Equation (10) for all areas, and the average value over all areas, thereby obtaining a smoothed design effect $\overline{def\ f} = \frac{1}{m}\Sigma_{i=1}^{m} def\ f_i$. The average proportion estimate over all areas is given by $\bar{p}_w = \frac{1}{m}\Sigma_{i=1}^{m}\hat{p}_{iw}$. Replacing the $def f_i$ by $\overline{def\ f}$ and $\hat{p}_{iw}$ by $\bar{p}_w$ in Equation (11), a DEFF-smoothed estimator of the sampling variance for proportion estimate $\hat{p}_{iw}$ is:

$$\tilde{V}_i^{DEFF} = \overline{def\ f}.\frac{\bar{p}_w(1-\bar{p}_w)}{n_i}.\left(1+\frac{1-\overline{def\ f}}{n_i}\right)^{-1}. \tag{13}$$

If the sample size n_i is large, then the term $\left(1-\overline{def\ f}\right)/n_i$ in $\tilde{V}_i^{DEFF}$ can be negligible. The smoothed variance $\tilde{V}_i^{deff}$ can then be simplified to

$$\tilde{V}_i^{def\ f} = \overline{deff}.\frac{\bar{p}_w(1-\bar{p}_w)}{n_i} \tag{14}$$

2.3. *Comparing the GVF and DEFF Smoothing*

We now show the similarity between the GVF-estimators and the DEFF-estimator $\tilde{V}_i^{DEFF}$ under certain conditions. Using $\tilde{V}_i^{GVF.RB}$ as an illustration, we can express this term as:

$$\begin{aligned}\tilde{V}_i^{GVF.RB} &= exp\left(\hat{\beta}_0+\hat{\beta}_1\cdot log(n_i)\right)\cdot exp(\frac{\hat{\tau}^2}{2})\\ &= exp(\hat{\beta}_0+\frac{\hat{\tau}^2}{2})\cdot exp\left(\hat{\beta}_1\cdot log(n_i)\right) = C_0\cdot exp(log(n_i)^{\hat{\beta}_1})\\ &= C_0\cdot n_i^{\hat{\beta}_1}\end{aligned}$$

where $C_0 = exp(\hat{\beta}_0 + \frac{\hat{\tau}^2}{2})$ is a constant. If the value of the regression coefficient $\hat{\beta}_1$ is close to -1, then the GVF-estimator $\tilde{V}_i^{GVF.RB}$ can be approximately expressed as $\tilde{V}_i^{GVF.RB} \approx C_0/n_i$.

The DEFF-estimator $\tilde{V}_i^{DEFF}$ can be rewritten as follows:

$$\tilde{V}_i^{DEFF} = \overline{def\ f}\cdot\frac{\bar{p}_w(1-\bar{p}_w)}{n_i}\cdot\left(1+\frac{1-\overline{def\ f}}{n_i}\right)^{-1}$$

$$= \frac{C_1}{n_i} \cdot \left(\frac{n_i + 1 - \overline{def\ f}}{n_i} \right)^{-1} = \frac{C_1}{n_i + 1 - \overline{def\ f}}$$

$$\approx \frac{C_1}{n_i},$$

where $C_1 = \overline{def\ f} \cdot \bar{p}_w(1 - \bar{p}_w)$ is a constant. Both the GVF-estimator $\tilde{V}_i^{GVF.RB}$ and the DEFF-estimator $\tilde{V}_i^{DEFF}$ are proportional to n_i^{-1} if the regression coefficient $\hat{\beta}_1$ is close to -1 in the GVF regression model. Given this condition, both the GVF and DEFF smoothed variances should perform similarly. Hirose et al. (2023) used the arc-sin transformation for binomial samples to construct the Fay-Herriot model with a fixed known variance of $1/4n_i$, which on the other hand, shows that the sampling variance is proportional to $1/n_i$. Their result for variance estimation via a binomial data transformation is consistent with our proposed GVF and DEFF smoothing variance.

In practical applications, we can combine the GVF and DEFF smoothed variance estimators by averaging them. We denote this averaged estimator as $\tilde{V}_i^{ASM} = (\tilde{V}_i^{GVF.RB} + \tilde{V}_i^{GVF.HBY} + \tilde{V}_i^{DEFF})/3$, where ASM stands for average smoothed. This simple data pooling method can provide additional smoothing to obtain the final smoothed variance estimate. As we will see in the LFS small area application (Section 4) and a simulation study (Section 5), the average smoothed estimator $\tilde{V}_i^{ASM}$ can perform very well and lead to large bias and CV reductions for small area estimates.

3. Application of Sampling Variance Smoothing

In this section, we will compare the GVF-estimators and DEFF-estimator by analysing two survey data sets. These data sets have information about the disease rate estimates of the Canadian Community Health Survey (CCHS) and adult disability rate estimates from the Participation and Activity Limitation Survey (PALS). Estimated variances for these two surveys are computed via the Rao-Wu bootstrap procedure. This procedure constructs bootstrap weights that reflect the sample details: see Rao and Wu (1988) or Rao et al. (1992) for details on how the bootstrap weights are computed.

3.1. CCHS Application

The CCHS is a federal survey conducted by Statistics Canada. The primary objective of CCHS is to provide timely and reliable estimates of health determinants, health status, and health system utilization across Canada. It is a cross-sectional survey that is carried out on a two-year collection cycle. The first year of the survey cycle "x.1" targets individuals aged 12 or older who are living in private dwellings, and it is a general population health survey with a large sample (130,000 persons) designed to provide reliable estimates at the health region, provincial and national levels. The second year of the survey cycle "x.2" has a smaller sample (30,000 persons) allocated based on provincial sample buy-ins and is designed to provide provincial and national level results on specifically focused health topics. Cycle "x.1" of the CCHS collected data corresponds to 136 health regions in the ten provinces and three territories. It primarily used two sampling frames. The first one, used

as the primary frame, was based on the area frame designed for the Canadian Labour Force Survey, and within the area frame, a multistage stratified cluster design was used to sample dwellings. The second frame consists of a list of telephone numbers. Random digit dialing methodology is used in some of the health regions for cost reasons. Following You and Zhou (2011), we use a small data set from Cycle 1.1 containing the estimates of asthma rates for 20 health regions in the province of British Columbia (BC) to demonstrate the sampling variance smoothing. In our data analysis, we use direct point estimates and direct sampling variance estimates to obtain the smoothed sampling variance estimates. Details of the methodology for the CCHS are given in Béland (2002).

For the CCHS data set, Figure 1 shows the plot of the log sampling variance $log(\tilde{V}_i)$ vs log sample size $log(n_i)$ with the fitted regression line. The GVF model fitting is very good as shown in Figure 1. The estimated regression parameters with standard errors (in parentheses) in the log-linear regression model (4) are obtained as $\hat{\beta}_0 = -2.861$ (1.321) and $\hat{\beta}_1 = -0.926$ (0.208). The residual correction term $exp(\hat{\tau}^2/2)$ is equal to 1.029. The HBY correction term in (8) is obtained as $\hat{\omega}_{HBY} = \hat{V}^{total}/\tilde{V}^{native} = 1.031$. Since these two correction terms are almost identical for this data set, so the two GVF estimators $\tilde{V}_i^{GVF.RB}$ and $\tilde{V}_i^{GVF.HBY}$ are almost the same. Since the correction term is close to 1, the naïve estimator $\tilde{V}_i^{native}$ just slightly underestimates the sampling variances.

Recall that in section 2, we claimed that the DEFF-estimator and the GVF-estimator should be approximately equivalent if the regression coefficient β_1 was close to -1. In this application, the estimated coefficient is $\hat{\beta}_1 = -0.926$. We would therefore expect $\tilde{V}_i^{DEFF}$, $\tilde{V}_i^{GVF.RB}$ and $\tilde{V}_i^{GVF.HBY}$ to be similar as well. Figure 2 compares the direct and smoothed sampling variance estimates sorted by the corresponding sample size (from small to large). For a more detailed comparison, we use the standard deviation as the squared root of the variance and plot the smoothed deviation with the direct deviation. In this application, all three smoothed variance estimators $\tilde{V}_i^{GVF.RB}$, $\tilde{V}_i^{GVF.HBY}$ and $\tilde{V}_i^{DEFF}$ and are almost identical and perform almost the same. The three smoothed estimators perform as expected and lead to smoothed sampling variances. For areas with large sample sizes, as expected, the smoothed variances and the direct variances are close to one another, and the smoothing method hardly modifies the direct estimate of sampling variance for large sample sizes.

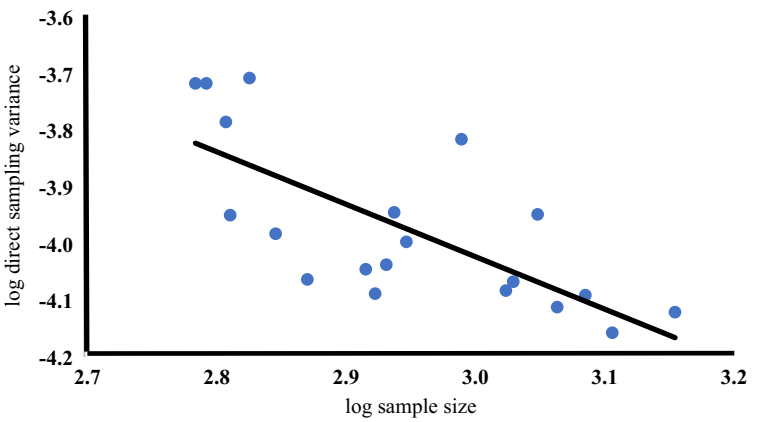


Fig. 1. Log direct sampling variance vs log sample size (BC health data).

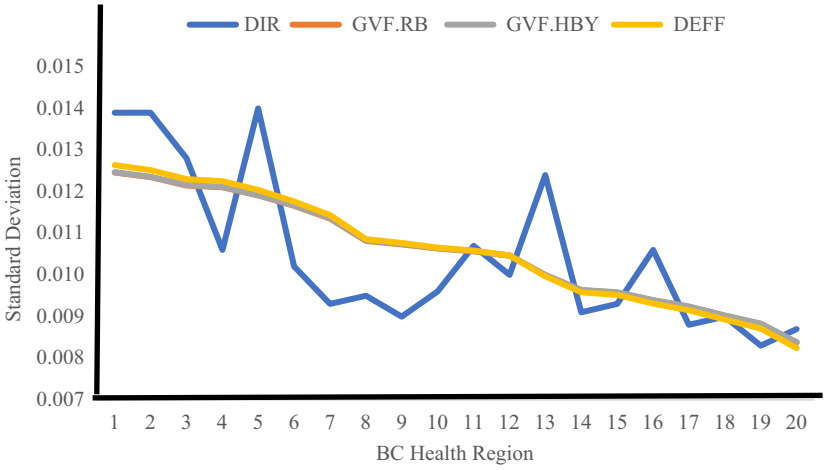


Fig. 2. Comparison of direct and smoothed deviation (BC health data).

3.2. PALS Application

The Participation and Activity Limitation Survey (PALS) is a post census survey that collects information about persons with disabilities whose everyday activities are limited because of a health-related condition or problem. This nationwide survey provides key information on the prevalence of different types of disabilities, on support provided to people with disabilities, on their labour force profile, their income, and their participation in society. The PALS sample was 48,000, consisting of approximately 39,000 adults and 9,000 children. The sample was selected using a two-phase stratified design where in the first phase, a Census questionnaire was distributed to approximately one out of five persons, and in the second phase, a stratified sample was selected based on characteristics from the first phase. However, the number of respondents to the survey does not allow for accurate direct estimates at the sub-provincial level. Following the demands to that effect which were expressed by many provincial governments as well as municipalities, Statistics Canada has put in place a model-based approach to small area estimation for the disability count and rate. Following Bizier et al. (2009), we consider the data of adult disability rate estimates for 116 small areas across Canada. These 116 areas include metropolitan areas, census agglomerations, and other sub-provincial areas. The survey took place between November 2006 and February 2007. Details of the methodology for PALS are given in Langlet et al. (2003).

For the PALS data set, Figure 3 shows the plot of log direct sampling variance vs log sample size. It is very clear from Figure 3 that the linear regression GVF model (4) is suitable for the PALS data. The estimated regression parameters with standard errors in the log-linear regression model (4) are obtained as $\hat{\beta}_0 = -3.033$ (0.263) and $\hat{\beta}_1$ -1.029 (0.056). The residual correction term $exp(\hat{\tau}^2/2)$ is equal to 1.334. The HBY correction term is $\hat{\omega}_{HBY} = \hat{V}^{total}/\tilde{V}^{navie} = 1.163$. Since the estimated coefficient is $\hat{\beta}_1 = -1.029$, we would expect the DEFF and GVF estimators to perform similarly. The naïve estimator $\tilde{V}_i^{navie}$ will underestimate the sampling variances, and estimates $\tilde{V}_i^{GVF.RB}$ will be slightly larger than $\tilde{V}_i^{GVF.HBY}$ in this example.

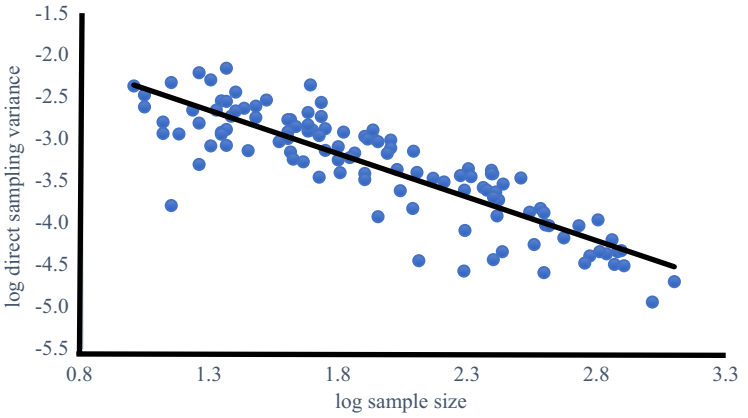


Fig. 3. Log direct sampling variance vs log sample size (PALS data).

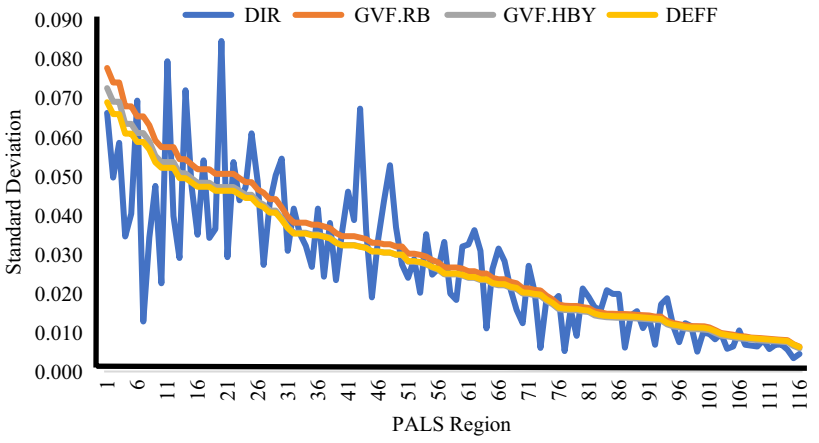


Fig. 4. Comparison of direct and smoothed deviation (PALS data).

Figure 4 compares the direct and smoothed sampling variances for the PALS data. The sampling variance estimates are sorted by the corresponding sample size from small to large. It is clear from Figure 4 that the direct sampling variance estimates have large variations when the sample size is small. The GVF and DEFF smoothed estimators perform similarly and lead to smoothed variance estimates. When the sample size is large, the direct and the three smoothed estimates are about the same as expected. When the sample size is small, the smoothed variances could be different. In this case, we could use the average ASM estimator $\tilde{V}_i^{ASM} = (\tilde{V}_i^{GVF.RB} + \tilde{V}_i^{GVF.HBY} + \tilde{V}_i^{DEFF})/3$ as a simple data pooling method to obtain the final smoothed variance estimate.

4. LFS Small Area Estimation Using Smoothed Sampling Variances

In this section, we apply the variance smoothing methods to the Canadian Labour Force Survey (LFS) data and compare the small area estimates based on the smoothed sampling variances. In the previous section, we displayed the performance of variance smoothing using GVF and DEFF as applied to the CCHS and PALS surveys. For the LFS, we will

exhibit the performance of GVF and DEFF, and the improvement of model-based small area estimates that use the smoothed sampling variances. The small area estimates of the LFS will be compared to the corresponding census values for the same reference month (May 2016).

The LFS produces monthly estimates of the unemployment rate at the national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Census Agglomerations (CAs) across Canada. Details of the methodology of the LFS are given in Methodology of the Canadian Labour Force Survey (2017). The estimated variances for the LFS are also computed through the Rao-Wu bootstrap procedure. For some sub-provincial areas, the direct estimates are not reliable because the sample sizes in some areas are quite small. Small area estimation, as applied to the LFS, usually estimates unemployment rates for local sub-provincial areas such as CMA/CAs using small area models. These models are discussed in Hidiroglou et al. (2019), Lesage et al. (2021), You et al. (2003), and You (2008, 2021).

We apply the Fay-Herriot model given by (1) and (2) to the May 2016 unemployment rate estimates at the CMA/CA level. There are 128 CMA/CAs (areas) in our study: three of these areas have a sample size smaller than or equal to 10, 10 have a sample size smaller than 30, 33 have a sample size smaller than 60, and 59 have a sample size smaller than 120, representing almost 50% of the areas in the study. In contrast, there are also 13 of the 128 areas with a sample size larger than 1,000, including some large cities such as Toronto, Montreal, and Vancouver. The Median sample size of all 128 areas is 129. We used four smoothed variance estimators in the LFS application, namely, $\tilde{V}_i^{GVF.RB}$, $\tilde{V}_i^{GVF.HBY}$, $\tilde{V}_i^{DEFF}$ and the average smoothed estimator $\tilde{V}_i^{ASM} = (\tilde{V}_i^{GVF.RB} + \tilde{V}_i^{GVF.HBY} + \tilde{V}_i^{DEFF})/3$. Figure 5 shows the plot of log direct sampling variance vs log sample size for the LFS data. The GVF model fitting is very good as shown in this figure, except for one or two outliers.

We first obtain the smoothed sampling variances for all the 128 CMA/CAs using the proposed $\tilde{V}_i^{GVF.RB}$, $\tilde{V}_i^{GVF.HBY}$, $\tilde{V}_i^{DEFF}$ and $\tilde{V}_i^{ASM}$. For the GVF model (4), the regression estimates with standard errors are $\hat{\beta}_0 = -3.194$ (0.306) and $\hat{\beta}_1 = -0.901$ (0.058). The RB residual correction term $exp(\hat{\tau}^2/2)$ is equal to 1.467 and the HBY correction term is

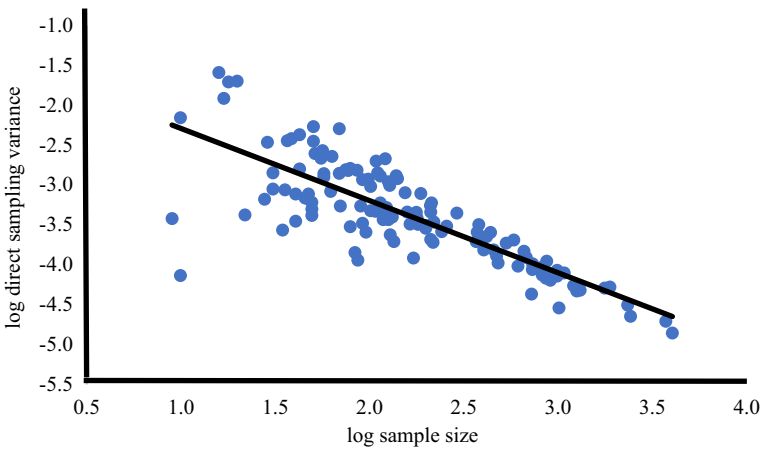


Fig. 5. Log direct sampling variance vs log sample size (LFS data).

$\hat{\omega}_{HBY} = \hat{V}^{total}/\tilde{V}^{naive} = 1.786$. Since the regression coefficient $\hat{\beta}_1 = -0.901$ is close to -1, and the difference between the two correction terms is not large, we could expect the GVF and DEFF lead to similarly smoothed sampling variances for the LFS data. Figure 6 shows the GVF and DEFF smoothed estimators perform very similarly and all lead to smoothed variance estimates.

We applied the empirical best linear unbiased prediction (EBLUP) approach in the LFS application to obtain the model-based estimates. The details of the EBLUP estimator and related mean squared error (MSE) estimation based on the Fay-Herriot model with REML method to estimate the model variance can be found, for example, in Rao and Molina (2015) and You (2021). Local area employment insurance monthly beneficiary rate is used as an auxiliary variable x_i in the linking model (2) as in Hidiroglou et al. (2019) and You (2008, 2021), The resulting linking model (2) is specified as $\theta_i = \delta_1 + x_i\delta_2 + v_i$, and $v_i \sim (0, \sigma_v^2)$. The model-based estimates and the direct estimates are compared with the census estimates to evaluate the effects of sampling variance smoothing. We applied the Fay- Herriot model to the 128 CMA/CA LFS unemployment rate data with the four different smoothed sampling variances and obtained the corresponding EBLUP estimates. The small area EBLUP estimates are compared via the absolute relative error (ARE) of the direct and EBLUP estimates with respect to the census estimates for each CMA/CA as follows:

$$\text{ARE}_i = \left| \frac{\theta_i^{Census} - \theta_i^{Est}}{\theta_i^{Census}} \right|,$$

where θ_i^{Est} is the direct or the EBLUP estimate and θ_i^{Census} is the corresponding census value of the LFS unemployment rate. It is a common practice to evaluate the model-based estimates with the census values, for example, as in Hidiroglou et al. (2019) and You (2021). We then take the average of AREs over CMA/CAs by different subgroups with respect to the sample size, as in Hidiroglou et al. (2019). Table 1 presents the estimates of

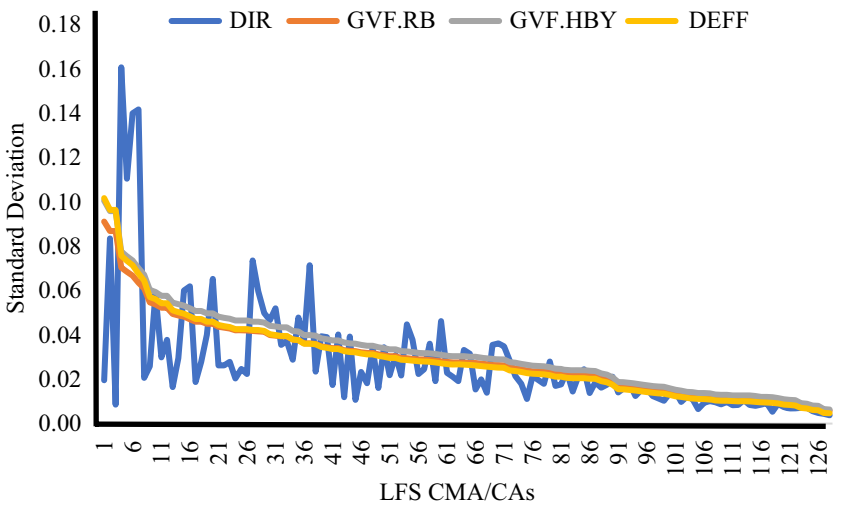


Fig. 6. Comparison of direct and smoothed deviation (LFS data).

the regression parameters and model variance in the linking model for the LFS application with different smoothed input sampling variances.

From Table 1, the estimates of the regression coefficients δ_1 and δ_2 are very similar for the different input smoothed sampling variances. The model variance estimate is slightly smaller using the GVF.HBY sampling variance in our application. Using the ASM sampling variance, the estimates of δ_1, δ_2 and σ_v^2 are quite good and reasonable as compared to the estimates that use the GVF or DEFF sampling variances.

Table 2 presents the average ARE for the direct LFS and EBLUP estimators based on different input sampling variance estimates. For comparison, we also used the direct sampling variance as input sampling variance in the Fay-Herriot model. For example, for EBLUP(DIR) the direct (DIR) sampling variance estimate is used in the Fay-Herriot model. EBLUP(GVF.RB) means that the smoothed sampling variance estimate $\tilde{V}_i^{GVF.RB}$ (GVF.RB) is used, etc. It is clear from Table 2 that the EBLUP estimates substantially improve the direct estimates by reducing the ARE. Even with the use of the direct sampling variance estimates, EBLUP(DIR) results in much smaller ARE than the direct survey estimator. However, by using the smoothed sampling variance estimates, EBLUP performs substantially much better than the direct estimator. The AREs are reduced for each area group, and consequently over all the areas. In general, all the EBLUPs with the four smoothed sampling variances perform very similarly. Amongst the EBLUP estimators using the smoothed sampling variances, EBLUP(GVF.HBY) has a slightly larger ARE than the others, and the EBLUP(DEFF) has a slightly smaller ARE. The respective AREs of EBLUP(GVF.RB), EBLUP(GVF.HBY) and EBLUP(DEFF) over all the 128 CMA/CAs are 0.138, 0.144, and 0.135. EBLUP(DEFF) performs the best in terms

Table 1. Estimates of regression parameter and model variance in the Fay-Herriot model.

Parameters	GVF.RB	GVF.HBY	DEFF	ASM
δ_1	4.884	4.916	4.875	4.892
δ_2	0.796	0.788	0.796	0.793
σ_v^2	0.551	0.269	0.836	0.532

Table 2. Comparison of ARE for EBLUP estimates based on the different input sampling variances.

CMA/CAs	Direct LFS	EBLUP (DIR)	EBLUP (GVF. RB)	EBLUP (GVF.HBY)	EBLUP (DEFF)	EBLUP (ASM)
25 smallest areas (sample size less than 50)	0.489	0.279	0.181	0.184	0.180	0.182
Next 25 smallest areas (sample size 50 to 100)	0.338	0.214	0.146	0.147	0.146	0.146
Next 25 smallest areas (sample size 100 to 180)	0.276	0.198	0.138	0.143	0.134	0.138
Next 25 smallest areas (sample size 180 to 550)	0.198	0.161	0.134	0.141	0.130	0.135
28 largest areas (sample size 550 and over)	0.132	0.125	0.099	0.108	0.091	0.099
Overall areas	0.283	0.194	0.138	0.144	0.135	0.139

of relative error. For the average smoothed sampling variance $\tilde{V}_i^{ASM}$ used in the Fay-Herriot model, the EBLUP(ASM) has an overall ARE value of 0.139, which is between the ARE values of EBLUPs using GVF and DEFF. The EBLUP(ASM) performs very well.

In terms of the overall average CV, EBLUP also reduces the CV substantially over the direct estimator. The direct LFS estimator has an average CV of 39.4%, EBLUP(DIR) has an average CV of 24.5%.The EBLUP(GVF.RB) has an average CV of 10.3%, EBLUP(GVF.HBY) has a slightly smaller average CV of 8.2%, and EBLUP(DEFF) has the average CV value 11.8%. The EBLUP(ASM) has an average CV of 10.2%. This means that using smoothed sampling variances substantially reduces the CV for EBLUPs, and once more the CV for EBLUP(ASM) is between the CV values of EBLUPs that use the GVF and DEFF variances.

EBLUP(ASM) has a smaller ARE value than either EBLUP(GVF.RB) or EBLUP(GVF.HBY). It also has a smaller CV than EBLUP(GVF.RB) and EBLUP(DEFF). The use of the averaged smoothed sampling variances $\tilde{V}_i^{ASM}$ in the model allows us to achieve a balanced reduction for both ARE and CV in our application. By comparing the ARE and CV for the EBLUP estimates, it is clear that the average smoothed estimator $\tilde{V}_i^{ASM}$ performs very well.

Lesage et al. (2021) considered the following smoothing model, denoted as the LBB model, for sampling variance smoothing:

$$log(\hat{V}_i) = \beta_0 + \beta_1 log(z_i) + \beta_2 log(1 - z_i) + \beta_3 log(n_i) + \varepsilon_i, i = 1, ..., m, \qquad (15)$$

where z_i is the employment insurance beneficiary rate used in the Fay-Herriot model as an auxiliary variable to obtain the EBLUP estimators. By applying the LBB smoothing model (15) to the 128 area sampling variance data, we have the following regression estimates $\hat{\beta}_0 = -4.443, \hat{\beta}_1 = -0.486$, $\hat{\beta}_2 = -29.139$ and $\hat{\beta}_3 = -0.886$. The residual correction term $\hat{\omega}_{RB} = exp\left(\hat{\tau}^2/2\right)$ is equal to 1.461 and the HBY correction term is $\hat{\omega}_{HBY} = \hat{V}^{total}/\tilde{V}^{naive} = 1.782$. We denote $\tilde{V}_i^{LBB.RB}$ as the smoothed variance estimator based on the LBB model (15) and formula (7) with a correction term $\hat{\omega}_{RB} = 1.461$. Similarly, let $\tilde{V}_i^{LBB.HBY}$ be the smoothed variance estimator based on the LBB model (15) using formula (8) with a correction term $\hat{\omega}_{HBY} = 1.782$. We now compare the EBLUP estimates based on the LBB smoothing model and the proposed smoothing method. In particular, we compare the proposed EBLUP(ASM) to EBLUP estimates using $\tilde{V}_i^{LBB.RB}$ and $\tilde{V}_i^{LBB.HBY}$, e.g., EBLUP(LBB.RB) and EBLUP(LBB.HBY).

Table 3 presents the average ARE to compare the effects of variance smoothing using the ASM and the LBB procedures. It is clear from Table 3 that all EBLUP estimates perform very well and improve the direct survey estimates by substantially reducing the ARE with respect to the census values. EBLUP(ASM) and EBLUP(LBB.RB) perform almost the same, and EBLUP(LBB.HBY) has slightly larger ARE, same as the performance of EBLUP(GVF.HBY) in Table 2. EBLUP(LBB.HBY) and EBLUP(GVF.HBY) perform almost identically by comparing the results in Table 2 and Table 3. In terms of CV, EBLUP(LBB.RB) and EBLUP(ASM) have the same average CV 10.2%, and EBLUP(LBB.HBY) has the same average CV 8.2% as EBLUP(GVF.HBY).

Table 3. Comparison of ARE for EBLUP estimates based on the different GVF models and smoothed sampling variances.

CMA/CAs	Direct LFS	EBLUP (ASM)	EBLUP (LBB.RB)	EBLUP (LBB.HBY)
25 smallest areas (sample size less than 50)	0.489	0.182	0.181	0.183
Next 25 smallest areas (sample size 50 to 100)	0.338	0.146	0.144	0.145
Next 25 smallest areas (sample size 100 to 180)	0.276	0.138	0.137	0.142
Next 25 smallest areas (sample size 180 to 550)	0.198	0.135	0.135	0.141
28 largest areas (sample size 550 and over)	0.132	0.099	0.099	0.108
Overall areas	0.283	0.139	0.138	0.143

The LFS small area application shows that the proposed GVF model (4) and the proposed sampling variance smoothing methods GVF, DEFF and ASM perform very well by comparing the EBLUP estimates with the census values and other GVF smoothing model for LFS application, for example, Lesage et al. (2021).

5. Simulation Study

In this section, we conduct a simulation study to verify and evaluate the proposed GVF, DEFF, and ASM smoothing estimators for small area estimation with data generated from different mechanisms. Following Lesage et al. (2021), we used the LFS data studied in Section 4 to generate the simulated data. We considered m = 128 areas in the simulation study. Let θ_i be the simulated true parameter of interest. The θ_i's were generated as $\theta_i = \gamma_0 + \gamma_1 z_i + \nu_i$, where z_i is the LFS beneficiary rate used in Section 4, $\gamma_0 = 0.05$, $\gamma_1 = 0.88$, and ν_i was generated from $N(0, \sigma_v^2)$, and $\sigma_v^2 = 4.78653e-05$. The values of γ_0, γ_1 and σ_v^2 were obtained from the EBLUP estimation of LFS application in Section 4 with the ASM smoothed sampling variances as input data. To generate the direct estimate $\hat{\theta}_i$ for the parameter θ_i and the corresponding direct sampling variance, we considered two approaches. The first one generated $\hat{\theta}_i$ from a binomial distribution, that is, $\hat{\theta}_i = n_i^{-1}$ Binomial(n_i, θ_i): Lesage et al. (2021) used the same simulation setup. The direct variance estimator was then computed as $\hat{V}_i = (n_i - 1)^{-1} \hat{\theta}_i(1 - \hat{\theta}_i)$. We denote this simulation setup as LBB setup. The second method generated the data directly from the Fay-Herriot model using the sampling variance modeling given by You and Chapman (2006) and You et al. (2013). The direct estimate $\hat{\theta}_i$ is generated as $\hat{\theta}_i = n_i + e_i$, where $e_i = \mathrm{N}(0, \sigma_i^2)$ and the sampling variance σ_i^2 is obtained from Section 4 using the average smoothed sampling variance ASM. The sampling variance σ_i^2 is treated as the true sampling variance in the simulation. The direct sampling variance estimate $\hat{V}_i$ is generated using $\hat{V}_i = (d_i)^{-1}\sigma_1^2\chi_{d_i}^2$, where $d_i = n_i - 1$, as of Rivest and Vandal (2002), Wang and Fuller (2003) and You (2021). We denote this simulation setup as FHM (Fay-Herriot modeling) setup.

We generated 5,000 samples for the LBB and FHM simulation setup respectively. Under the LBB simulation setup, the average estimated design effects $\overline{def\ f}$ is 1.0131 and

Table 4. Comparison of ARE, CV and coverage rate (CR) for simulation study.

	LBB simulation setup			FHM simulation setup		
	ARE	CV	CR	ARE	CV	CR
Direct estimator	0.2553	36.57%		0.3112	78.95%	
EBLUP(DIR)	0.0829	10.12%	94.06%	0.0872	10.26%	93.92%
EBLUP(GVF.RB)	0.0713	6.38%	94.05%	0.0776	8.54%	93.71%
EBLUP(GVF.HBY)	0.0716	6.13%	93.11%	0.0776	8.61%	93.85%
EBLUP(DEFF)	0.0714	6.28%	93.73%	0.0765	8.37%	93.21%
EBLUP(ASM)	0.0715	6.22%	93.68%	0.0769	8.27%	93.36%

the estimated $\hat{\beta}_1 =$ -1.0081. For simulation under the FHM setup, the average estimated design effects $\overline{def\ f}$ is 2.6679 and the estimated $\hat{\beta}_1 =$ -0.7302. Table 4 presents the ARE comparison results to the average CV for the direct estimator and EBLUPs that use different smoothed sampling variances. Following Hidiroglou and You (2016), we also compute and compare the confidence interval coverage rate (CR) of the EBLUPs. The 95% confidence interval of the EBLUP estimator is obtained as EBLUP ± 1.96 $\sqrt{mse(\text{EBLUP})}$. Table 4 also reports the confidence interval coverage rate for EBLUPs over the 5,000 simulated samples.

Under the LBB simulation setup, the ARE of the direct estimator is 0.2553 with an average CV of 36.57%. As expected, the EBLUP has a much smaller ARE and CV. EBLUP(DIR) has an ARE of 0.0829 with an average CV of 10.12%, whereas the EBLUPs using smoothed sampling variances have even smaller AREs and CVs. The GVF, DEFF and ASM lead to almost the same ARE which is around 0.0715, and GVF.HBY has slightly smaller CV compared with GVF.RB or DEFF as in the real LFS application.

Under the FHM simulation setup, again, the proposed GVF, DEFF and ASM lead to smaller AREs and CVs than either the direct estimator or the EBLUP that uses direct sampling variances. Using the ASM variance estimates leads to a smaller CV than using the GVF and DEFF as shown in Table 4 under the FHM simulation setup. In general the ASM leads to a balanced ARE and CV in the simulation study, which is the same as in the LFS application, which indicates that using ASM as a data pooling to average the GVF and DEFF is useful in practice for the sampling variance smoothing. Note that using the ASM is suggested, but it is also a matter of choice. For confidence intervals, the EBLUPs all provide similar and reasonable coverage of around 93–94%. In summary, as shown in the simulation results reported in Table 4, the proposed GVF, DEFF and ASM smoothing models and methods all perform similarly and very well.

6. Conclusions and Suggestions

In this article, we have proposed sampling variance smoothing estimators using the generalized variance function method and smoothed design effect method for small area estimation. The proposed smoothing models and methods only require the use of the sample size in the model and the computation of design effects. The proposed estimators $\tilde{V}_i^{GVF.RB}$, $\tilde{V}_i^{GVF.HBY}$ and $\tilde{V}_i^{DEFF}$ usually result in similar smoothed variance estimates.

In practical applications, we may use the average smoothed estimator $\tilde{V}_i^{ASM}$ as a data pooling procedure to obtain the final smoothed variance estimate. The use of ASM may be

particularly useful and may have more advantages when the GVF and DEFF lead to different smoothing estimates. The proposed smoothing model and method can be easily implemented in practice. The proposed smoothing methods simplify the smoothing procedure for practical users as they do not need other complicated GVF models or auxiliary variables for the sampling variance modeling. The small area estimation results in our LFS application in Section 4 and the simulation study in Section 5 both indicate that the proposed GVF, DEFF and ASM perform very well.

It may also be possible in practice to consider a weighted average of the three variance estimators estimators $\tilde{V}_i^{GVF.RB}$, $\tilde{V}_i^{GVF.HBY}$ and $\tilde{V}_i^{DEFF}$. For example, if DEFF leads to a larger CV for the final small area EBLUP estimates, and GVF leads to a smaller CV for the final EBLUP estimates, then we may want to reduce the weights for DEFF and put more weights on GVF. The weights could be proportional to the inverse of the average CV of the small area estimates corresponding to the smoothed sampling variances, as the idea of small area estimation is to achieve the CV reduction and to obtain reliable small area estimates as shown in our data analysis and simulation study. From Section 4, EBLUP(GVF.RB) has an average CV of 10.3%, EBLUP(GVF.HBY) has a slightly smaller average CV of 8.2%, and EBLUP(DEFF) has an average CV value of 11.8%. Then we may construct a weighted ASM (WASM) smoothed sampling variance as follows:

$$\tilde{V}_i^{WASM} = (\tilde{V}_i^{GVF.RB} + 1.2* \tilde{V}_i^{GVF.HBY} + 0.8* \tilde{V}_i^{DEFF})/3.$$

This weighted smoothed sampling variance $\tilde{V}_i^{WASM}$ puts more weight on $\tilde{V}_i^{GVF.HBY}$ and less weight on $\tilde{V}_i^{DEFF}$ according to the inverse of the CVs of the corresponding EBLUPs. We then obtained the EBLUP(WASM) using the weighted $\tilde{V}_i^{WASM}$, and compared the results with the EBLUP(ASM). The ARE for EBLUP(ASM) is 0.1392, whereas the ARE for EBLUP(WASM) is 0.1398. The average CV for EBLUP(ASM) is 10.2%, whereas the average CV for EBLUP(WASM) is 9.96%. Thus by putting more weight on GVF.HBY and less weight on DEFF, the EBLUP(WASM) has a slightly smaller CV than EBLUP(ASM) as expected, and using the WASM leads to a very tiny increase in ARE. The difference between using ASM and WASM in the LFS application is very small and could be ignored. In general, the simple average ASM should provide adequate additional smoothing by combining the GVF and DEFF smoothed variances.

For future work, we may consider more auxiliary variables in the GVF modeling for both the real data analysis and simulation study. It would be interesting to study the relationship between the GVF and DEFF estimators if other auxiliary variables including design variables and/or proxies are available and used in the GVF models. Note that the GVF models need the sample size as an auxiliary variable to make it compatible with the smoothed DEFF procedure. Bertarelli et al. (2018) used a log CV modelling for proportions instead of the variance. It would also be interesting to compare the log CV modelling with the GVF modelling on log variance through real data analysis and simulation study. For the averaged smoothing using both GVF and DEFF, more combinations or choices of weights could be considered and evaluated via additional simulation studies.

In practice, for areas with larger sample sizes, it may also be fine to use the direct variances by assuming that the direct variance estimates are stable enough when the

sample size is large. This is a reasonable procedure to follow, because as shown in Figures 2, 4, and 6 in our applications, the direct estimate and smoothed estimate are usually very similar for areas with larger sample sizes. It is possible to set a minima threshold in terms of a minimum sample size, say n_{min}, to decide on the choice between smoothed or direct variance estimates. If the sample size of an area is smaller than n_{min}, then the preferred choice would be to use smoothed variance estimates. If on the other hand, the sample size of an area is greater than n_{min}, then the preferred choice would be the direct variance estimates. For example, Hidiroglou et al. (2019) used direct variance estimates in the LFS small area estimation for the areas with large sample sizes. As a rule of thumb, they set $n_{min} = 400$ in their application and used the direct sampling variance estimate when the sample size was greater than 400. However, care should be taken when choosing the value for n_{min}, as more comparisons and studies may be needed in practice to evaluate the results. In general, we recommend using the smoothed variance estimates for all areas, as it is quite simple to apply, and can avoid the additional problem of settling a value for n_{min}.

Finally, we make some recommendations for total variance smoothing using the proposed methods. For the estimation of totals, we can modify the proposed method as follows: If $\hat{y}_{iw}$ is an estimator of a total, and if we know the corresponding population size N_i, we can transform the total estimator $\hat{y}_{iw}$ to a rate (proportion) estimator by dividing it by N_i, that is, $\hat{p}_{iw} = \hat{y}_{iw}/N_i$, and the corresponding sampling variance is $\hat{V}_i(\hat{p}_{iw}) = \hat{V}_i(\hat{y}_{iw})/N_i^2$. We can apply the proposed smoothing methods for proportions to get smoothed variance estimates $\tilde{V}_i(\hat{p}_{iw})$. Then the smoothed sampling variance for the total can be simply obtained as $\tilde{V}_i(\hat{y}_{iw}) = \tilde{V}_i(\hat{p}_{iw}) \cdot N_i^2$. It is a good idea in practice to transform the estimates of totals into estimates of proportions. The reason for this is that the transformed proportion estimates are more easily modeled than estimates of totals. The total estimates can be obtained by transforming back the model-based rate estimates.

7. References

Beaumont, J.-F., and C. Bocci. 2016. "Small area estimation in the Labour Force Survey." Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada, Ottawa, Canada.

Béland, Y. 2002. *Canadian Community Health Survey Methodological overview.* Health Reports, 13(3): Catalogue 82–003. Statistics Canada. Available at: https://www150.statcan.gc.ca/n1/pub/82-003-x/2001003/article/6099-eng.pdf (accessed March 2023).

Bertarelli, G., M.G. Ranalli, F. Bartolucci, M. D'Alò and F. Solari. 2018. "Small area estimation for unemployment using latent Markov models." *Survey Methodology* 44: 167–192. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2018002/article/54956-eng.htm (accessed March 2023).

Bizier, V., Y. You, L. Veilleux, and C. Grondin. 2009. "Model-based Approach to Small Area Estimation of Disability Count and Rate Using Data from the 2006 Participation and Activity Limitation Survey." In Proceedings of Section on Survey Research Methods, Joint Statistical Meeting 2009. Available at: http://www.asasrms.org/Proceedings/y2009/Files/303608.pdf (accessed March 2023).

Dass, S.C., T. Maiti, H. Ren, and S. Sinha. 2012. "Confidence interval estimation of small area parameters shrinking both means and variances." *Survey Methodology* 38: 173–187. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11756-eng.pdf (accessed March 2023).

Dick, P. 1995. "Modelling net undercoverage in the 1991 Canadian Census." *Survey Methodology* 21 (1): 45–54. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1995001/article/14411-eng.pdf (accessed March 2023).

Duan, N. 1983. "Smearing estimate: a nonparametric retransformation method." *Journal of the American Statistical Association* 78(383): 605–610. DOI: http://dx.doi.org/10.1080/01621459.1983.10478017.

Fay, R.E., and R.A. 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.

Gambino, J.G. 2009. "Design effect caveats." *The American Statistician* 63: 141–146. DOI: http://dx.doi.org/10.1198/tast.2009.0028.

Ghosh, M., J. Myung, and F.A.S. Moura. 2018. "Robust Bayesian small area estimation." *Survey Methodology* 44: 101–115. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2018001/article/54959-eng.htm (accessed March 2023).

Hidiroglou, M.A., J.-F. Beaumont, and W. Yung. 2019. "Development of a small area estimation system at Statistics Canada." *Survey Methodology* 45: 101–126. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2019001/article/00009-eng.htm (accessed March 2023).

Hidiroglou, M., and Y. You. 2016. "Comparison of unit level and area level small area estimators." *Survey Methodology* 42: 41–46. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2016001/article/14540-eng.pdf (accessed March 2023).

Hirose, M.Y., M. Ghosh, and T. Ghosh. 2023. "Arc-sin transformation for binomial sample proportions in small area estimation." *Statistica Sinica* 33: 1–23. DOI: http://dx.doi.org/10.5705/ss.202020.0446.

Langlet, É.R., D. Faucher, and E. Lesage. 2003. "An application of the bootstrap variance estimation method to the Canadian participation and activity limitation survey." In Proceedings of the Joint Statistical Meetings, August 3–7, San Francisco, Califonia, U.S.A. American Statistical Association: 2299–2306. Available at: http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000873.pdf (accessed March 2023).

Lesage, E., J.-F. Beaumont, and C. Bocci. 2021. "Two local diagnostics to evaluate the efficiency of the empirical best predictor under the Fay-Herriot model." *Survey Methodology* 47: 279–297. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2021002/article/00001-eng.htm (accessed March 2023).

Liu, B., P. Lahiri, and G. Kalton. 2014. "Hierarchical Bayes modeling of survey weighted small area proportions." *Survey Methodology* 40: 1–13. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2014001/article/14030-eng.pdf (accessed March 2023).

Maples, J., W. Bell, and E. Huang. 2009. "Small area variance modeling with application to county poverty estimates from the American community survey." In Proceedings of the Section on Survey Research Methods, August 1–6, Washington. D.C., U.S.A. American Statistical Association: 5056–5067. Available at: http://www.asasrms.org/Proceedings/y2009/Files/305528.pdf. (accessed March 2023).

Methodology of the Canadian Labour Force Survey (2017). Statistics Canada: 71-526-X, ISBN 978-0-660-24068-8. Available at: https://www150.statcan.gc.ca/n1/pub/71-526-x/71-526-x2017001-eng.htm (accessed March 2023).

Rivest, L.P., and E. Belmonte. 2000. "A conditional mean squared error of small area estimators." *Survey Methodology* 26: 67–78. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2000001/article/5179-eng.pdf (accessed March 2023).

Rivest, L.P., and N. Vandal. 2002. "Mean squared error estimation for small areas when the small area variances are estimated." In Proceedings of the International Conference on Recent Advances in Survey Sampling, July 1013, 2002, Ottawa, Canada. Available at: https://www.mat.ulaval.ca/fileadmin/mat/documents/lrivest/Publications/64-Rivest-Vandal2003.pdf (accessed March 2023).

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation*, 2nd Edition. John Wiley & Sons, New York. DOI: http://dx.doi.org/10.1002/9781118735855.

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference With Complex Survey Data," *Journal of the American Statistical Association* 83: 231–241. DOI: http://dx.doi.org/10.1080/01621459.1988.10478591.

Rao, J.N.K., C.F.J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18: 209–219. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14486-eng.pdf (accessed March 2023).

Souza, D.F., F.A.S. Moura, and H.S. Migon. 2009. "Small area population prediction via hierarchical models." *Survey Methodology* 35: 203–214. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2009002/article/11042-eng.pdf (accessed March 2023).

Sugasawa, S., and T. Kubokawa. 2017. "Heteroscedastic nested error regression models with variance functions." *Statistica Sinica* 27: 1101–1123. DOI: http://dx.doi.org/10.5705/ss.202015.0318.

Sugasawa, S., H. Tamae, and T. Kubokawa. 2017. "Bayesian estimators for small area models shrinking both means and variances." *Scandinavian Journal of Statistics* 44: 150–167. DOI: http://dx.doi.org/10.1111/sjos.12246.

Wang, J., and W.A. Fuller. 2003. "The mean square error of small area predictors constructed with estimated area variances." *Journal of the American Statistical Association* 98: 716–723. DOI: https://doi.org/10.1198/016214503000000620.

Wolter, K.M. 2007. "Generalized Variance Functions." In: *Introduction to Variance Estimation. Statistics for Social and Behavioral Sciences*. Springer, New York, NY. DOI: http://dx.doi.org/10.1007/978-0-387-35099-8_7.

You, Y. 2008. "An integrated modeling approach to unemployment rate estimation for sub- provincial areas of Canada." *Survey Methodology* 34(1): 19–27. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10614-eng.pdf (accessed March 2023).

You, Y. 2021. "Small area estimation using Fay-Herriot model with sampling variance smoothing and modeling." *Survey Methodology* 47: 361–370. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2021002/article/00007-eng.htm (accessed March 2023).

You, Y., and B. Chapman. 2006. "Small area estimation using area level models and estimated sampling variances." *Survey Methodology* 32: 97–103. Available at: https://

www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263 (accessed March 2023).

You, Yong, J.N.K. Rao, and J. Gambino. 2003. "Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach." *Survey Methodology* 29: 25–32. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20030016602 https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019263 (accessed March 2023).

You, Y., J.N.K. Rao, and M. Hidiroglou. 2013. "On the performance of self-benchmarked small area estimators under the Fay-Herriot area level model." *Survey Methodology* 39: 217–229. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-eng.pdf (accessed March 2023).

You, Y., and Q.M. Zhou. 2011. "Hierarchical Bayes small area estimation under a spatial model with application to health survey data." *Survey Methodology* 37: 25–37. Available at:https://www150.statcan.gc.ca/n1/pub/12-001-x/2011001/article/11445-eng.pdf (accessed March 2023).

Received September 2022
Revised December 2022
Accepted March 2023

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 591–595, http://dx.doi.org/10.2478/JOS-2023-0027

Book Review

Maria del Mar Rueda Garcia[1]

Silvia Biffignandi and Jelke Bethlehem. *Handbook of Web Surveys, 2nd edition*. 2021 Wiley, ISBN: 978-1-119-37168-7, 624 pps.

A look at non-probability surveys

Survey methodologies are currently in flux due to social and technological changes that have led to a significant increase in refusals to participate and difficulties in accessing individuals to interview. These problems may provoke significant biases and compromise the validity of the results obtained. However, the development of new technologies has facilitated the emergence of new data acquisition techniques, such as web surveys, that present great advantages in terms of speed in obtaining data, reduced costs and the possibility of accessing specific population sectors. Although web surveys can be probabilistic, in practice many operate via self-selection and the principles of probability sampling are not applied.

The analysis of three databases that compile surveys carried out during the first year of the COVID-19 pandemic has made it possible to quantify this trend towards non-probability sampling designs. The latter represent 38% of the 63 surveys included in Oxford Supertracker, a global directory that compiles the most significant efforts to obtain information on the social and policy-related impacts of the pandemic (Daly et al. 2020). This figure rises to 92% in a review of surveys related to COVID-19 in Spain (Sánchez-Cantalejo et al. 2023). Finally, according to a tracker of studies on Covid-19 carried out in the field of social sciences (Matias and Levitt 2023), 73% used a web survey as their main mode of data collection, and of these, 90% used non-probability designs.

Although some authors believe that probability surveys might soon be phased out from the production of official statistics, others, such as Beaumont (2020), argue that the time has not yet come for this change because the alternatives are not reliable and generalisable enough to eliminate the use of probabilistic surveys without severely sacrificing the quality of the estimates obtained. This opinion is in line with Cornesse et al. (2020), who summarised the empirical evidence on the accuracy of probability and nonprobability sample surveys and recommended that probability sample surveys should continue to be used, at least for the present.

Non-probability surveys, although they usually have large sample sizes, can present important selection and coverage problems (Beaumont and Rao 2021), as in most cases the sample generation process is unknown, which can compromise the generalisation of the

[1] Maria del Mar Rueda Garcia, University of Granada, Department of Statistics and O.R., Granada, 18071, Spain. Email: mrueda@ugr.es
Acknowledgments: The author are grateful to Ministerio de Ciencia e Innovación (Spain) by Project PID2019-106861RB-I00/AEI/10.13039/501100011033].

results to the population under study. Accordingly, the use of non-probability surveys must be accompanied by a significant effort to detect and correct biases in the sample, once the raw data have been obtained.

Powerful new methodologies have been developed to infer parameters using data from non-probability samples, and this research has been reviewed by Buelens et al. (2018), Rao (2022), Valliant (2020) and Yang and Kim (2020), among others. The methods considered include propensity score adjustment (Lee and Valliant 2009), propensity-adjusted probability prediction (Elliott and Valliant 2017), inverse sampling (Kim and Wang 2019), mass imputation (or statistical matching) (Rivers 2007), doubly robust methods (Chen et al. 2019), kernel smoothing methods (Kern et al. 2021), superpopulation modelling (Buelens et al. 2018) and combinations of these techniques (Castro-Martín et al. 2021; Liu and Valliant 2023).

In this field, an immense number of papers have been presented in recent years, reflecting the importance assigned to this question. Moreover, specialised sessions have been held in many statistics and survey-oriented congresses, and special issues have been published by journals (for example, Survey Methodology 48(2)). In this recent publication, Wu (2022) summarised the state of the literature on the analysis of non-probability survey data. Further comments on this article contributed to its interest and topicality.

In another noteworthy review, Kalton (2023) presented an overview of the history of the use of probability and non-probability sampling, from the birth of surveying to the present day. In his words, "This is an exciting and challenging time for survey methodologists".

In practice, however, the majority of non-probability surveys do not apply these error correction techniques, making at most only basic adjustments by post-stratification or calibration with sociodemographic variables. Thus, Sanchez-Cantalejo et al. (2023) found no survey that used any of the aforementioned new methods to adjust for self-selection bias.

Nevertheless, it is necessary to highlight the importance of correcting self-selection biases, and to make modern techniques better known among survey designers and practitioners. The *Handbook of Web Surveys* represents a good means of introducing these methodologies, since it is aimed at a broad sector, including researchers and end-users, and provides basic knowledge on this topic.

In its 14 chapters, the Handbook provides a comprehensive description of the field of web surveys and associated problems. The following chapters are especially valuable regarding non-probability surveys.

Chapter 11 examines the estimation problems arising from the self-selection of sample elements in a survey. After introducing the problem of sample representativeness and providing some examples from real surveys, the basic concepts of estimation in probability samples are presented and explicit expressions are obtained for the biases of the sample means and the factors that may influence these biases. This chapter also discusses the use of reweighting techniques to reduce (but not necessarily eliminate) self-selection bias.

Chapter 12 details various weighting adjustment techniques, including post-stratification, generalised regression, the raking ratio method and calibration estimation. These techniques are well-known means of dealing with non-response or undercoverage.

Their effectiveness in reducing self-selection bias is examined in two practical situations: when the population distribution of the auxiliary variables is known, and when it must be estimated from a reference sample.

Chapter 13 is more innovative and introduces the concept of response probabilities, describing how they can be estimated through response propensities under a *Missing at Random* pattern. Logit, probit and linear models are considered for the propensities, and the advantages and disadvantages of each are considered. Two approaches based on these response propensities are then described: inverse propensity weighting and response propensity stratification. The chapter also includes a simulation study in which these two approaches are compared against the base procedure, when no correction is applied.

In my opinion, the authors should have made some reference (without entering into exhaustive detail) to other, more current, methodologies, at least mentioning *statistical matching,* a technique that has been known for decades and which in many cases has proven to be more appropriate than inverse propensity weighting and response propensity stratification. I would also have liked to see a real application in this chapter, in addition to the simulation study.

Other questions that might usefully have been addressed include the selection of auxiliary variables (Ferri-García and Rueda 2022) and the use of classification and regression machine learning techniques (Buelens et al. 2018; Ferri-García and Rueda 2018; Kern et al. 2021), which are often used as an alternative to generalised linear models. However, this omission can be considered normal since the Handbook is intended to be a generic introduction to web surveys, not a specific manual on non-probability surveys.

In conclusion, this Handbook provides a comprehensive insight into the potential of web surveys for data collection, highlighting the problems of coverage and self-selection bias that can arise with this type of survey. As an introduction to the problem of estimation with data obtained in non-probability surveys, this text is especially useful for researchers who are not specialists in sampling theory, but require web surveys as a means to obtain information.

References

Beaumont, J.F. 2020. "Are probability surveys bound to disappear for the production of official statistics?" *Survey Methodology* 46: 1–29. Available at: http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-eng.htm.

Beaumont, J.F., and J.N.K. Rao. 2021. "Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies?" *The Survey Statistician* 83: 11–22. DOI: https://api.semanticscholar.org/CorpusID:231751185.

Buelens, B., J. Burger, J.A. van den Brakel. 2018. "Comparing Inference Methods for Non-probability Samples." *International Statistical Review* 86: 322–343. DOI: https://doi.org/10.1111/insr.12253.

Castro-Martín, L.; M.D.M. Rueda, R. Ferri-García. 2021. "Combining statistical matching and propensity score adjustment for inference from non-probability surveys." *Journal of Computational and Applied Mathematics* 404: 113414. DOI: https://doi.org/10.1016/j.cam.2021.113414.

Chen, Y., P. Li., C. Wu. 2019. "Doubly Robust Inference With Nonprobability Survey Samples." *Journal of the American Statistical Association* 115:(532): 2011–2021. DOI: https://doi.org/10.1080/01621459.2019.1677241.

Cornesse, C, A. Blom, D. Dutwin, J. Krosnick, E. de Leeuw, S. Legleye, J. Pasek, D. Pennay, B. Phillips, J. Sakshaug, B. Struminskaya, and A. Wenz. 2020. "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." *Journal of Survey Statistics and Methodology* 8(1): 4–36. DOI: https://doi.org/10.1093/jssam/smz041.

Daly, M., B. Ebbinghaus, L. Lehner, M. Naczyk, and T. Vlandas. 2020. *Oxford Supertracker: The Global Directory for COVID Policy Trackers and Surveys.* Department of Social Policy and Intervention. Available at: https://supertracker.spi.ox.ac.uk/ (accessed September 2023).

Elliott, M.R., and R. Valliant. 2017. "Inference for nonprobability samples." *Statistical Science* 32(2): 249–264. DOI: https://doi.org/10.1214/16-STS598.

Ferri-García, R., and M.M. Rueda. 2018. "Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys." *SORT* 42(2): 159–182. DOI: https://doi.org/10.2436/20.8080.02.73.

Ferri-García, R., and M.M. Rueda. 2022. "Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys." *Statistical Papers* 63: 1829–1881. DOI: https://doi.org/10.1007/s00362-022-01296-x.

Kalton, G. 2023. "Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day." *Statistics in Transition* 24(3): 1–22. DOI: https://doi.org/10.59170/stattrans-2023-032.

Kern, C. Y. Li, and L. Wang. 2021. "Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples." *Journal of Survey Statistics and Methodology* 9(5): 1088–111. DOI: https://doi.org/10.1093/jssam/smaa028.

Kim, J.K., and Z. Wang. 2019. "Sampling techniques for big data analysis." *International Statistical Review* 87: S177–S191. DOI: https://doi.org/10.1111/insr.12290.

Lee, S., and R. Valliant. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods and Research* 37: 319–343. DOI: https://doi.org/10.1177/0049124108329643.

Liu, Z., and R. Valliant. 2023. "Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration." *Journal of Official Statistics* 39:1: 45–78. DOI: http://dx.doi.org/10.2478/JOS-2023-0003.

Matias, J. N., and A. Leavitt. 2023. *COVID-19 social science research tracker*. GitHub Repository. Available at: https://github.com/natematias/covid-19-social-science-research (accessed June 2023).

Rao, J.N.K. 2022. "On Making Valid Inferences by Integrating Data from Surveys and Other Sources." *Sankhya B* 83: 242–272. DOI: https://doi.org/10.1007/s13571-020-00227-w.

Rivers D. 2007. "Sampling for Web Surveys." In Proceedings of the Joint Statistical Meetings, Salt Lake City, UT, USA, 29 July–2 August. DOI: https://doi.org/10.1002/9781119371717.ch4.

Sánchez-Cantalejo, C., D. Yucumá, M.M. Rueda, A. Orly, E. Martín, C. Higeras, and A. Cabrera-León. 2023. "Scoping Review of the methodology of large health surveys

conducted in Spain early on the COVID-19 pandemic." *Frontiers in Public Health*11. DOI: https://doi.org/10.3389/fpubh.2023.1217519.

Valliant, R. 2020. "Comparing alternatives for estimation from nonprobability samples." *Journal of Survey Statistics and Methodology* 8(2): 231–263. DOI: https://doi.org/10.1093/jssam/smz003.

Wu, C. 2022. "Statistical inference with non-probability survey samples." *Survey Methodology Statistics Canada* 48(2). Available at: http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm.

Yang, S., and Kim, J.K. 2020. "Statistical data integration in survey sampling: A review." *Japanese Journal of Statistics and Data Science* 3: 625–650. Available at: https://link.springer.com/article/10.1007/s42081-020-00093-w.

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 597–600, http://dx.doi.org/10.2478/JOS-2023-0028

Editorial Collaborators

The editors wish to thank the following referees and guest editors of theme issues who have generously given their time and skills to the Journal of Official Statistics during the period 1 October 2022 to 30 September 2023. An asterisk indicates that the referee served more than once during the period.

Abe, Naohito, Hitotsubashi Daigaku, Institute of Economic Research, Tokyo, Japan
Ardilly, Pascal, INSEE, Lyon, France
Arias-Salazar, Alejandra, Free University of Berlin, Berlin, Germany
Avella, Marco, Columbia University, New York, U.S.A.
Axelson Martin, Statistics Sweden, Örebro, Sweden*
Bacchini, Fabio, ISTAT, Rome, Italy
Balk, Bert, Amersfoort, 3818 GC, the Netherlands*
Bauder, Donald, U.S. Census Bureau, Washington D.C., U.S.A.
Bell, William, U.S. Census Bureau, Washington D.C., U.S.A.*
Benassi, Federico, University of Naples Federico II, Naples, Italy
Benedetti, Ilaria, University of Tuscia, Viterbo, Italy
Bentley, Alan, Statistics New Zealand, Wellington, New Zealand
Beręsewicz, Maciej, Poznań University of Economics and Business, Poznań, Poland
Beresovsky, Vladislav, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.
Berg, Emily, Iowa State University of Science and Technology, Ames, Iowa, U.S.A.*
Bethlehem, Jelke, Leiden University, Hazerswoude-Rijndijk, the Netherlands
Białek, Jacek, University of Lodz, Lodz, Poland*
Biffignandi, Silvia, University of Bergamo, Bergamo, Italy
Bocci, Cynthia, Statistics Canada, Ottawa, Ontario, Canada
Böhning, Dankmar, University of Southampton, Southampton, UK
Bollinger, Christopher, University of Kentucky, Lexington, Kentucky, U.S.A.
Bonnéry, Daniel, University of Cambridge, Cambridge, UK
Bryant, John, Bayesian Demography Limited, Christchurch, New Zealand*
Bycroft, Christine, Statistics New Zealand, Christchurch, New Zealand
Cervera Ferri, Jose, DevStat Servicios de Consultoría Estadística, Valencia, Spain
Chen, Baoline Bureau of Economic Analysis, Suitland, Maryland, U.S.A.
Chen, Lu, National Institute of Statistical Sciences, Washington D.C., U.S.A.
Chipperfield, James, Australian Bureau of Statistics, Belconnen, Australia
Christian, Leah, NORC, University of Chicago, Chicago, Illinois, U.S.A.
Clark, Stephen, University of Leeds, Leeds, UK*
Coffey, Stephanie, U.S. Census Bureau, Suitland, Maryland, U.S.A.
Czajka, John, Mathematica Policy Research, Inc., Washington D.C., U.S.A.
Dambon, Jakob, Swiss Re, Zurich, Switzerland

Lugtig, Peter, Utrecht University, Utrecht, the Netherlands
Malmros, Jens, Stockholm University, Stockholm, Sweden
Mariyah, Siti, STIS Polytechnic of Statistics, Jakarta, Indonesia*
Marker, David, Westat, Columbia, Maryland, U.S.A.
Maślankowski, Jacek, University of Gdansk, Sopot, Poland*
McClain, Colleen, Pew Research Center, Washington D.C., U.S.A.
McCool, Danielle, Utrecht University, Utrecht, the Netherlands
Mitra, Pratik, Reserve Bank of India, Mumbai, India
Moauro, Filippo, ISTAT, Rome, Italy*
Nedyalkova, Desislava, Swiss Federal Statistical Office, Neuchâtel, Switzerland
Newhouse, David, World Bank Group, Washington. D.C., U.S.A.
Novás Maria, National Statistical Institute, Madrid, Spain*
Ollech, Daniel, Deutsche Bundesbank, Frankfurt am Main, Germany
Osier, Guillaume, STATEC, Luxembourg, Luxembourg
Paiva, Thais, University of Minas Gerais, Belo Horizonte, Brazil
Parker, Paul, University of California, Santa Cruz, California, U.S.A.
Pfeffermann, Danny, University of Southampton, Southampton, UK
Polidoro, Federico, Italian National Institute of Statistics, Rome, Italy
Rezaei, Ghahroodi, Zahra, University of Tehran, Tehran, Iran
Ricciato, Fabio, Eurostat, Luxembourg, Luxembourg*
Rivest, Louis-Paul, University of Laval, Quebec, Canada*
Rocci, Fabiana, ISTAT, Rome, Italy
Rojas-Perilla, Natalia, United Arab Emirates University, Al Ain, United Arab Emirates*
Rothbaum, Jonathan, U.S. Census Bureau, Washington D.C., U.S.A.*
Ruggles, Steven, University of Minnesota, Minneapolis, Minnesota, U.S.A.*
Schechter,Susan, NORC, University of Chicago, Bethesda, Maryland, U.S.A.
Scholtus, Sander, Statistics Netherlands, The Hague, the Netherlands.*
Schonlau, Matthias, University of Waterloo, Waterloo, Ontario, Canada
Shabbir, Javid, Quaid-i-Azam University, Islamabad, Pakistan
Sharygin, Ethan, Portland State University, Portland, U.S.A.
Shimizu, Chihiro, The University of Tokyo, Kunitachi, Chiba, Japan
Shonosuke, Sugasawa, University of Tokyo, Tokyo, Japan
Smith, Alan, The Financial Times, London, UK
Smith, James, UWE Bristol, University of The West of England, Bristol, UK
Smith, Paul, University of Southampton, Southampton, UK
Smith, Peter, University of Southampton, Southampton, UK*
Souza, Debora, IBGE, Rio de Janeiro, Brazil
Spoorenberg, Thomas, United Nations, New York, New York, U.S.A.*
Steel, David, University of Wollongong, Wollongong, Australia*
Steurer, Miriam, University of Graz, Graz, Austria
Stokes, Chase, UC Berkeley, Berkeley, California, U.S.A.
Tayman, Jeff, UC San Diego, San Diego, California, U.S.A.*
Thibaudeau, Yves, U.S. Census Bureau, Washington D.C., U.S.A.
Tillé, Yves, University of Neuchâtel, Neuchâtel, Switzerland*
Van Berkel, Kees, Statistics Netherlands, Heerlen, the Netherlands

sciendo

Journal of Official Statistics, Vol. 39, No. 4, 2023, pp. 601–603, http://dx.doi.org/10.2478/JOS-2023-0029

Index to Volume 39, 2023

Contents of Volume 39, Numbers 1–4

Author Index