Distr.
LIMITED
E/ESCWA/SD/2019/TP.4
21 March 2020
ORIGINAL: ENGLISH

Economic and Social Commission for Western Asia (ESCWA)

"A short guide for small area estimation in household surveys: Illustration to poverty mapping in Palestine with expenditure survey and census data"

Isabel Molina Peralta and Eduardo García Portugués

2019-08-04, v1.0

United Nations
Beirut, 2019

___________________+

Note: This document has been reproduced in the form in which it was received, without formal editing. The opinions expressed are those of the authors and do not necessarily reflect the views of ESCWA.

20-00112

Acknowledgment

This report was prepared by Isabel Molina Peralta and Eduardo García Portugués (both at the University Carlos III de Madrid) , using the data files prepared by Mr. Jawad Saleh from the Palestinian Central Bureau of Statistics (PCBS) and Mr. Nathan Reece from the Statistics Division of the Economic and Social Commission of Western Asia (ESCWA), under the supervision of Marwan Khawaja, Chief Demographic & Social Statistics Section, ESCWA Statistics Division

An earlier version of the report was presented at the “Regional Workshop on Poverty Measurement in Arab Countries”, held in Tunis - Tunisia during the period 23-25 July 2019 and the “Regional Workshop on the Use of Census Data for Development Planning and Scientific Research in Arab Countries” held in Rabat, Kingdom of Morocco during the period 1-3 October 2019. We acknowledge the helpful comments from the participants during these two workshops.

The data files used in preparing this report were provided by the Palestinian Central Bureau of Statistics (PCBS).

Table of Contents

Summary

This report introduces the problem of disaggregation of statistical data, known in the literature as the small area estimation problem, provides a description of the main methods for disaggregation of estimates and illustrates the procedures through an application to poverty mapping in Palestine combining data from a recent consumption survey and from the population census.

The document is organized as follows. Section 1 gives some motivation and introduces the main philosophy behind small area estimation. Section 2 describes the basic small area estimation methods. Section 3 applies these methods to the estimation of poverty rates and gaps in specified localities from Palestine, using data from the 2016/2017 Palestinian Expenditure Consumption Survey (PECS) and the 2017 population census, provided by the Palestinian Central Bureau of Statistics (PCBS). Finally, Section 4 provides final remarks and recommendations.

1- The disaggregation problem

1.1 Description of the problem

Government-sponsored surveys conducted by statistical institutes are designed to produce statistics at a given aggregation level; that is, for either geographical or socio-economic subdivisions of the population. However, once a survey has been conducted, with sample size established to produce reliable estimates at a given aggregation level, there is often a subsequent demand for estimates at a more disaggregated level. Satisfying this demand of producing statistical estimates for smaller subdivisions than were originally planned, without incurring in additional costs due to an increase in the survey sample size, is the objective of *small area estimation*.

Prior to the implementation of the survey, it would be possible to improve some aspects of the sampling design that avoid this problem to some extent. For example, it is possible to increase the sample sizes (with the corresponding increase in cost) in the areas where it is desired to estimate with higher precision, or to allocate the total sample size of the survey among the areas in a more efficient way. Despite the fact that there are several mechanisms to improve the sampling design and to have a sufficient minimum of data in all the subdivisions of the population, "the client will always require more than is specified at the design stage" (Fuller 1999), hence small area estimation methods are unavoidable.

In the literature, the subdivisions for which statistical data (or estimates) are desired are commonly referred to as *areas* or *domains*, irrespectively of whether they actually correspond to geographical regions or socio-economic subgroups, or crossings of both types. When estimating a particular quantity of interest (e.g., a poverty indicator) in one of these areas, we call a *direct estimator* to an estimator that uses basically the survey data from that area. The usual direct estimators are unbiased or practically unbiased with respect to the distribution of the sampling design, e.g., across all possible samples that can be drawn from the population by means of the corresponding sampling design. However, if the survey was not planned to estimate at such level of disaggregation, the sample size in some of the areas may be too small, resulting in excessively large sampling errors for the direct estimators in those areas. The areas in which this occurs, regardless of their population size, are referred to in the literature as *small areas* (with respect to the indicator of interest). As a consequence, it is not the population size of the area what gives it the adjective "small", but the poor quality of the direct estimates in these areas.

There is no universal upper limit for the sampling error of a direct estimator to consider an area as "small". Each statistical institute or international organization establishes its own limit for relative sampling error or Coefficient of Variation (CV) from which statistical data are considered unreliable and therefore not published. Some national statistical institutes agree to establish an estimator as "non-publishable" when its relative sampling error or CV exceeds 20%. Therefore, for these institutions, the areas for which direct estimates of an indicator of interest have a CV greater than 20% would be considered "small" for such indicator.

1.2 Methodologies to circumvent the disaggregation problem

Indirect estimation methods

The so-called *indirect estimation methods* do not only consider sample data relative to the area of interest, but also from other areas. These estimators use information from *auxiliary variables* that are related to the variable of interest. For example, consider that our variable of interest, let us say income, is related with education level. This relationship is considered similar for all areas and is represented through a model that links all the areas by means of common parameters. These common parameters are then estimated using the sample data from all the areas (the total sample is typically very large), providing more efficient estimators (compared to direct ones) due to the use of a greater amount of information. These estimators tend to slightly compromise the bias under the design in exchange for greatly increasing the overall efficiency of the estimator, evaluated in terms of mean squared error.

The gain in efficiency of indirect estimators with respect to direct ones tends to increase as the area sample size decreases. However, these estimators tend to improve in most of the areas, including many with large sample sizes. In fact, some indirect estimators (see Section 2.2) tend to a direct estimator as the sample size of the area increases. Therefore, indirect estimators possessing this property can be used for all areas, regardless of whether they are "small" or not, thereby reducing the importance of having a more accurate or formal definition of small area.

The simplest indirect estimators are based on unrealistic hypotheses and therefore may have considerable bias. These include the so-called *synthetic estimators*, which do not account for the heterogeneity that usually exists between areas. Well-known synthetic estimators are the *post-stratified synthetic estimator* and the *regression-synthetic estimator*. Other classical indirect estimators are composite, which are computed as a weighted average between a direct estimator and a synthetic estimator. However, these estimators have several disadvantages. One is that the weight attached to each estimator does not depend on the goodness-of-fit of the model assumed by the synthetic estimator, which means that the indirect estimator is used regardless of its precision. Moreover, the weight attached to the direct estimator is usually very close to one, which means that little information is borrowed from the other areas. More sophisticated indirect estimators, which represent better the existing between-area heterogeneity, are those based on *mixed regression models*. There are two large groups of mixed regression models that are typically used for small area estimation: models at the *area level* and models at the *unit level*. We review them next.

Area-level models

Area-level models use only aggregated data for the estimation areas. Typically, this type of data can be accessed with fewer restrictions, as aggregation avoids confidentiality issues. Widely used area-level linear mixed models are the so-called Fay–Herriot (FH) models, proposed by Fay and Herriot (1979). These models have a two-level structure:

- In the first level, the indicator of interest for the areas is assumed to be linearly related with a set of auxiliary variables at the area level, where this relationship is constant for all the areas. For example, the decrease in the area mean income due to a large proportion of unemployed people in the area, keeping other variables constant, is the same in all areas. Thus, all areas are linked through a linear regression model.
- In the second level, it is assumed that, given the true values of the area indicators of interest, the corresponding direct estimators are centered on these true values, with variances that are assumed to be known. Such variances represent the sampling errors of the direct estimators, which depend on the area sample size. Since area sample sizes are typically different, these variances vary across areas.

These models have had a well-deserved success because the resulting estimators for the areas are a weighted average between direct estimators and synthetic-regression estimators, with weights depending on the area sample size. When the synthetic model does not fit the data well (i.e., the considered auxiliary variables do not sufficiently explain the between-area heterogeneity of the indicator) or the sample size of an area is large, the estimator based on the FH model places a greater weight to the direct estimator, which is sufficiently accurate. Conversely, when the synthetic model fits well or the area sample size is small, the estimator increases the weight given to the synthetic-regression estimator. In this case, the efficiency is increased because the synthetic estimator has a common regression coefficient for all the areas, which is then estimated using data from all the areas. In addition, since direct estimators are approximately unbiased with respect to the sampling design, for areas with larger sample sizes, the estimators obtained from the FH model also preserve a small bias under the design. One challenge in FH models is to determine the values of direct estimator variances (or heteroscedastic variances of model error terms). Although, as mentioned above, these variances are assumed to be known, in practice they are replaced by estimates. Given the small sample size in some of the areas, the estimates of these variances are also very imprecise. There are smoothing methods such as the generalized variance function method, see Fay and Herriot (1979), or nonparametric estimation of these variances, see González-Manteiga et al. (2010). The estimation of these variances adds the problem of incorporate the error of estimation of these variances into the error of the final estimator.

Unit-level models

In unit-level models, as the name implies, the model is set for each population unit (superpopulation model), and therefore the fitting of these models requires microdata of the response variable and auxiliary variables. The first model of this type was proposed by Battese, Harter, and Fuller (1988) and is known as the nested error model. This is a linear regression model that includes, in addition to the individual model errors, random effects associated with the areas. The area effects represent the heterogeneity between the areas that is not explained by the available auxiliary variables. These models are widely used today when the required data is available, since they incorporate much more information than area-level models, and the true model error variances are not needed.

Since unit-level models are based on the whole sample size n, they can attain much higher efficiency than area-level models, as long as there exist unit-level auxiliary variables that are sufficiently informative about the response variable.

Best linear unbiased predictor

The assumption of a stochastic model generating the values of the variable of interest for population units makes the indicators of interest random quantities. In this context, an unbiased predictor of an indicator is one whose expectation under the model coincides with the expectation of that indicator.

When estimating linear-type indicators of the values of the variable of interest in the individuals of the population, such as means or totals, the basic models at area or individual level that are used are part of the linear mixed models that include random effects for the areas of interest. Within these models, the usual indirect estimator is the *Best Linear Unbiased Predictor* (BLUP), which is the linear combination of the observed values of the response variable for the sample units, which is unbiased under the model and minimizes the model mean squared error. The BLUP depends on the unknown model parameters, which represent the common behavior among the areas. Replacing these unknown parameters with estimators gives the *Empirical BLUP* (EBLUP). This is finally the usual estimator (or predictor) based on a model of a linear indicator in a small area.

The BLUP does not require any hypothesis of normality in the model. On the other hand, to estimate more general indicators than linear ones, the *best predictor* is the one that minimizes the mean squared error, without requiring it to be linear or unbiased. This equals the expectation under the model of the indicator to estimate, given the observed values in the sample. Under normality, the best predictor of a linear indicator that uses the weighted least squares estimator of the regression parameter matches the BLUP. In the absence of normality or when the indicator to be estimated is not linear, it is possible that the expectation that defines the best predictor cannot be computed analytically. In that case, numerical approximations are employed to approximate the best predictor.

Further reading

This report is not meant to provide a thorough review of small area estimation methods, but just an overview of the most commonly employed techniques that are in relation with the undertaken project. For more detail on the techniques described here and for information about other techniques, we refer the interested reader to the monograph by Rao and Molina (2015), where most of the work carried out in the field up to the publication date is described.

2- Estimation methods

2.1 Direct estimators

In this section we describe the basic direct estimators for the mean of a variable Y within area d, given by

$$\overline{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}\,, \qquad (1)$$

where Y_{di} denotes the value of Y for individual i within area d.

Before reviewing the basic direct estimators, let us introduce the notation that will be employed along the document. The population U of size N is assumed to be partitioned in D subpopulations U_d of size N_d, $d = 1, \ldots, D$, with $N = \sum_{d=1}^{D} N_d$. We denote by s the sample drawn from the population U of size n, by s_d the subsample from area d of size n_d (that may be zero) and by r_d the set of elements that do not belong to the sample of the same area, $d = 1, \ldots, D$, where $n = \sum_{d=1}^{D} n_d$. Besides, we denote by π_{di} the inclusion probability of unit i in the sample from area d, by $w_{di} = \pi_{di}^{-1}$ to the corresponding sampling weight and by $\pi_{d,ij}$ to the joint inclusion probability of units i and j in the sample from area d.

Horvitz–Thompson estimator

If $\pi_{di} > 0$ for all $i = 1, \ldots, N_d$, the unbiased estimator, with respect to the sampling design, of the area mean $\overline{Y}_d$ is the well-known *Horvitz–Thompson* (HT) estimator. This estimator requires knowing the true area count N_d and the sampling weights $w_{di} = \pi_{di}^{-1}$ for the sampled units in area d. Assuming that these are known, the HT estimator of $\overline{Y}_d$ is

$$\hat{\overline{Y}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, Y_{di}. \qquad (2)$$

For the area total $Y_d = \sum_{i=1}^{N_d} Y_{di}$, the HT estimator is simply $\hat{Y}_d = \sum_{i \in s_d} w_{di}\, Y_{di}$, which does not require knowing the area count N_d.

If $\pi_{d,ij} > 0$ for all $i, j \in \{1, \ldots, N_d\}$, an unbiased estimator under the sampling design of the variance of $\overline{Y}_d$ is

$$\widehat{\text{var}}_\pi(\hat{\overline{Y}}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \frac{Y_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{Y_{di} Y_{dj}}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}. \qquad (3)$$

During the estimation phase, in many cases not all the information on the sampling design is available apart from the sampling weights w_{di}. If the second order inclusion probabilities $\pi_{d,ij}$ are not available, then the estimator (3) can not be computed. However, for sampling designs with second-order inclusion probabilities verifying $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, for $j \neq i$ (as in

Poisson sampling, where equality is satisfied), the second term of (3) is approximately zero. In addition, replacing $w_{di} = \pi_{di}^{-1}$, we get the following variance estimator, which does not depend on the second-order inclusion probabilities:

$$\widehat{\text{var}}_{\pi}(\hat{\bar{Y}}_d) = N_d^{-2} \sum_{i \in s_d} w_{di} \, (w_{di} - 1) Y_{di}^2.$$

The HT estimator (2) weights the individual observations Y_{di} using the sampling weights or inverses of the inclusion probabilities, $w_{di} = \pi_{di}^{-1}$. This protects against situations where the probability of selecting an individual is related to the value of the variable of interest (informative sampling design). For example, imagine that individuals with lower income have larger probability of appearing in the sample. Then, this type of individuals are likely to appear more often in the final sample, while those with higher incomes are likely to be scarce in the sample. This means that, if we were to estimate by giving the same weight to all the sample observations, as in the usual sample mean, we would understate the mean income. For this reason, it is necessary to reduce the weight of observations that are most likely to appear in the sample, and increase the weight to those that are least likely to appear.

Hájek estimator

Although the HT estimator is exactly unbiased with respect to the sampling design, its variance under the design can be very large when the sample size of the area d, n_d, is small. A slightly biased estimator for small n_d but with a somewhat smaller variance, and which does not require knowledge of the area size N_d to estimate the mean $\bar{Y}_d$, is the *Hájek estimator* (HA). This estimator is a weighted average of the sample observations from the area, using as weights the sampling weights, that is,

$$\hat{\bar{Y}}_d^{\text{HA}} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di} \, Y_{di}, \text{ where } \hat{N}_d = \sum_{i \in s_d} w_{di} \,. \qquad (4)$$

For the total $Y_d = \sum_{i=1}^{N_d} Y_{di}$, the Hájek estimator is $\hat{Y}_d^{\text{HA}} = N_d \hat{\bar{Y}}_d^{\text{HA}}$, which does require knowing the area count N_d.

Under the sampling design, a variance estimator of the Hájek estimator, $\hat{\bar{Y}}_d^{\text{HA}}$, is obtained using Taylor's linearization method. The resulting estimator is obtained by simply replacing Y_{di} with $\tilde{e}_{di} = Y_{di} - \hat{\bar{Y}}_d^{\text{HA}}$ in the estimator of the variance of the HT estimator of the total $\hat{Y}_d$ and dividing by $\hat{N}_d$; that is,

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{\text{HA}}) = \quad \hat{N}_d^{-2}\left\{\sum_{i\in s_d}\frac{(Y_{di}-\hat{\bar{Y}}_d^{\text{HA}})^2}{\pi_{di}^2}(1-\pi_{di}) \right.$$

$$\left. + 2\sum_{i\in s_d}\sum_{\substack{j\in s_d\\ j>i}}\frac{(Y_{di}-\hat{\bar{Y}}_d^{\text{HA}})(Y_{dj}-\hat{\bar{Y}}_d^{\text{HA}})}{\pi_{di}\pi_{dj}}\left(\frac{\pi_{d,ij}-\pi_{di}\pi_{dj}}{\pi_{d,ij}}\right)\right\},$$

assuming that $\pi_{d,ij} > 0$, for all i and j. For designs in which $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, for $j \neq i$, such as in Poisson sampling, this estimated variance reduces to

$$\widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{\text{HA}}) = \hat{N}_d^{-2}\sum_{i\in s_d} w_{di}\,(w_{di}-1)(Y_{di}-\hat{\bar{Y}}_d^{\text{HA}})^2.$$

Note that, when adding the HT direct estimators of the totals Y_d for the areas of a larger region, say for the entire population, the HT estimator of the population total $\hat{Y} = \sum_{d=1}^{D}\sum_{i\in s_d} w_{di}\,Y_{di}$ is obtained, that is,

$$\sum_{d=1}^{D}\hat{Y}_d = \hat{Y}.$$

Since at higher aggregation levels (e.g. the population level), the HT estimator is efficient, this *benchmarking* property is desirable for the area estimators. In fact, this consistency property is often required for publication of official statistical figures. However, other estimators, especially indirect ones, do not add up exactly to the considered estimator for the total population (which could be different from $\hat{Y}$). For adjustments on small area estimators to enforce this property, see Ghosh and Steorts (2013) and references therein.

2.2 Indirect estimators

As previously outlined, small area estimators based on models fall into the category of indirect estimators, as they borrow information from other areas. They do so by representing in the model the unexplained between-area heterogeneity through the random additive effects for the areas. As seen later, these random area effects provide a very good property to the estimators based on linear models: they can be written as compound estimators that tend to a direct estimator in areas with sufficient sample size. These models are significantly more realistic than the basic synthetic ones that do not include the area effects, and this translates into less biased estimators under sample design.

Fay–Herriot model

The Fay–Herriot (FH) model is a very popular area-level model that was introduced by Fay and Herriot (1979) to estimate per capita income in small places of the U.S. This model is currently used by the U.S. Census Bureau, within the Small Area Income and Poverty Estimates (SAIPE) program, to estimate proportions of poor school-age children in counties and school districts; see Bell (1997) or http://www.census.gov/hhes/www/saipe for more details.

The model links the indicators of interest for all areas δ_d, $d = 1, \dots, D$, by assuming that they are linearly related with the values of p area-level auxiliary variables $\mathbf{x}_d$, and this relationship is constant for all areas; more precisely,

$$\delta_d = \mathbf{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \qquad (5)$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, which is common for all the areas, and u_d is the regression error, also known as the *random effect* of area d. Random effects represent the heterogeneity of the indicators δ_d across the areas that is not explained by the auxiliary variables. In the simplest model, u_d are assumed to be independent and identically distributed (iid), with common (unknown) variance σ_u^2, that is, $u_d \overset{\text{iid}}{\sim} (0, \sigma_u^2)$.

The true values of δ_d are not observable, and then model (5) cannot be readily fit. One possible approach is to plug-in a direct estimator $\hat{\delta}_d^{\text{DIR}}$ of δ_d, but we must keep in mind that this estimator has sampling error. The FH model considers that $\hat{\delta}_d^{\text{DIR}}$ is unbiased under the design. In this case, we can represent the error due to sampling of this estimator through the model:

$$\hat{\delta}_d^{\text{DIR}} = \delta_d + e_d, \quad d = 1, \dots, D, \qquad (6)$$

where e_d is the sampling error in area d. Errors e_d are assumed to be independent of each other and are also independent of the area effects u_d, have zero mean and known variances ψ_d; that is, $e_d \overset{\text{i}}{\sim} (0, \psi_d)$. In practice, these variances, $\psi_d = \text{var}_\pi(\hat{\delta}_d^{\text{DIR}} | \delta_d)$, $d = 1, \dots, D$, are estimated using the survey microdata. Combining models (5) and (6), the following linear mixed model is obtained:

$$\hat{\delta}_d^{\text{DIR}} = \mathbf{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \qquad (7)$$

The BLUP of $\delta_d = \mathbf{x}_d{}'\boldsymbol{\beta} + u_d$ can be obtained simply by fitting the linear mixed model (7), that is, the BLUP under the FH model of δ_d is

$$\tilde{\delta}_d^{\text{FH}} = \mathbf{x}_d{}'\tilde{\boldsymbol{\beta}} + \tilde{u}_d, \qquad (8)$$

where $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{\text{DIR}} - \mathbf{x}_d{}'\tilde{\boldsymbol{\beta}})$ is the BLUP of u_d, being $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ and $\tilde{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$ under model (7), given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d \, \mathbf{x}_d \mathbf{x}_d{}' \right)^{-1} \sum_{d=1}^{D} \gamma_d \, \mathbf{x}_d \hat{\delta}_d^{\text{DIR}}.$$

Substituting $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{\text{DIR}} - \mathbf{x}_d{}'\tilde{\boldsymbol{\beta}})$ in (8) we can express the BLUP as a convex linear combination of the direct estimator and the synthetic regression estimator,

$$\tilde{\delta}_d^{\text{FH}} = \gamma_d \hat{\delta}_d^{\text{DIR}} + (1 - \gamma_d)\mathbf{x}_d{}'\tilde{\boldsymbol{\beta}}, \qquad (9)$$

where $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0,1)$ is the weight of the direct estimator. For an area d where the direct estimator $\hat{\delta}_d^{\text{DIR}}$ is efficient, the sample variance ψ_d will be small compared to the unexplained heterogeneity σ_u^2, hence γ_d is close to one and therefore $\tilde{\delta}_d^{\text{FH}}$ gives more weight to the direct estimator. On the other hand, in areas d where the direct estimator lacks quality due to the small sample size, ψ_d is larger than σ_u^2, then γ_d approaches zero and therefore more weight is given to the synthetic regression estimator $\mathbf{x}_d{'}\widetilde{\boldsymbol{\beta}}$, which uses data from all areas to estimate the common parameter $\boldsymbol{\beta}$. That is, $\tilde{\delta}_d^{\text{FH}}$ borrows information from the other areas through the synthetic regression estimator $\mathbf{x}_d{'}\widetilde{\boldsymbol{\beta}}$ as needed, depending on the efficiency of the direct estimator. The fact that the BLUP $\tilde{\delta}_d^{\text{FH}}$ approaches the direct estimator when the area sample size is large (ψ_d small) is a very desirable property, as it is not needed to know when an area is "small enough" to use this estimator instead of the direct estimator.

The BLUP of δ_d depends on the true value of σ_u^2 of the area effects u_d. In practice, this variance is unknown, and we must estimate it. Common approaches are Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML), the latter correcting the estimator of σ_u^2 to provide a less biased estimator for finite sample sizes. Fay and Herriot (1979) also proposed a method of moments that avoids establishing the shape of the likelihood. Let $\hat{\sigma}_u^2$ be a consistent estimator of σ_u^2, like those obtained by the above fitting methods. Replacing σ_u^2 with $\hat{\sigma}_u^2$ in (8), we get the EBLUP of δ_d:

$$\hat{\delta}_d^{\text{FH}} = \hat{\gamma}_d \hat{\delta}_d^{\text{DIR}} + (1 - \hat{\gamma}_d)\mathbf{x}_d{'}\widehat{\boldsymbol{\beta}}, \qquad (10)$$

where $\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$ and $\widehat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \hat{\gamma}_d \, \mathbf{x}_d \mathbf{x}_d{'}\right)^{-1} \sum_{d=1}^{D} \hat{\gamma}_d \, \mathbf{x}_d \hat{\delta}_d^{\text{DIR}}$.

If the model parameters $\boldsymbol{\beta}$ and σ_u^2 are known, the MSE of the BLUP based on the model (7) is given by

$$\text{mse}(\tilde{\delta}_d^{\text{FH}}) = \gamma_d \psi_d \leq \psi_d = \text{var}_\pi(\hat{\delta}_d^{\text{DIR}} | \delta_d).$$

Therefore, given the true value of the indicator δ_d, if σ_u^2 and $\boldsymbol{\beta}$ are known, the BLUP under the FH model, $\tilde{\delta}_d^{\text{FH}}$, cannot be less efficient than the direct estimator. In practice, since σ_u^2 and $\boldsymbol{\beta}$ are estimated, the error due to the estimation of these two parameters is added to the MSE of the FH estimator. However, these two extra terms that are added to the MSE tend to zero as the number of areas D tends to infinity. Therefore, for a sufficient number of D areas, it is likely that the FH estimator will still improve on the direct estimator in terms of MSE. For this reason, these estimators tend to improve in most areas as long as there is a sufficient number of areas D.

Under normality of u_d and e_d, Prasad and Rao (1990) obtained a second order approximation (i.e., with error $o(D^{-1})$ when D is large) of the MSE for the FH estimator:

$$\text{mse}(\hat{\delta}_d^{FH}) = g_{d1}(\sigma_u^2) + g_{d2}(\sigma_u^2) + g_{d3}(\sigma_u^2),$$

where

$$\begin{aligned}
g_{1d}(\sigma_u^2) &= \gamma_d \psi_d,\\
g_{2d}(\sigma_u^2) &= (1-\gamma_d)^2 \mathbf{x}_d{}' \left(\sum_{d=1}^{D} \gamma_d \, \mathbf{x}_d \mathbf{x}_d{}' \right)^{-1} \mathbf{x}_d,\\
g_{3d}(\sigma_u^2) &= (1-\gamma_d)^2 (\sigma_u^2 + \psi_d^2)^{-1} \overline{\mathrm{var}}(\hat{\sigma}_u^2),
\end{aligned}$$

where $\overline{\mathrm{var}}(\hat{\sigma}_u^2)$ is the asymptotic variance of the estimator $\hat{\sigma}_u^2$ of σ_u^2, which depends on the employed estimation method.

ELL method

Elbers, Lanjouw, and Lanjouw (2003) developed a method for estimation of general indicators (henceforth referred as ELL) that is traditionally used by the World Bank to construct maps showing the regional distribution of poverty or inequality. This method was the first one designed to estimate indicators that are more complex than the average or the total, as long as they are a function of a welfare variable (usually disposable income or expenditure). This method assumes the nested error model

$$Y_{di} = \mathbf{x}_{di}{}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \qquad (11)$$

where $Y_{di} = \log(E_{di} + c)$ for $c > 0$ constant, E_{di} is the welfare variable for unit i in the area d, $u_d \overset{\text{iid}}{\sim} (0, \sigma_u^2)$, and $e_{di} \overset{\text{i}}{\sim} (0, \sigma_e^2 k_{di}^2)$, being u_d and e_{di} independent, and k_{di} known constants representing possible heteroscedasticity.

The ELL estimator of a general parameter $\delta_d = \delta_d(\mathbf{y}_d)$ under the model (11) is obtained by a bootstrap procedure. This bootstrap procedure provides a numerical approximation of the theoretical ELL estimator, which is given by the marginal expectation $\hat{\delta}_d^{\text{ELL}} = \mathbb{E}[\delta_d]$. The same bootstrap procedure is used to get an estimate of the MSE of the ELL estimator.

The bootstrap procedure works as follows:

1. From the residuals of the fitted model (11) to the data, random effects u_d^* are generated for each area $d = 1, \dots, D$, and errors e_{di}^*, for each unit $i = 1, \dots, N_d$, $d = 1, \dots, D$.

2. From the estimator $\widehat{\boldsymbol{\beta}}$ of the regression parameter $\boldsymbol{\beta}$, and using the values of the auxiliary variables for the individuals inside and outside the sample, bootstrap values of the response variable are generated for *all* the population units. This is done by the generation process

$$Y_{di}^* = \mathbf{x}_{di}{}'\widehat{\boldsymbol{\beta}} + u_d^* + e_{di}^*, \quad i = 1, \dots, N_d,\ d = 1, \dots, D.$$

3. The above step provides a census of the response variable, which can be used to estimate indicators of any kind as long as they depend only on the welfare of individuals. This generation process is repeated for $a = 1, \dots, A$, getting A full censuses. Then, for each census a, we calculate the indicator of interest $\delta_d^{*(a)} = \delta_d(\mathbf{y}_d^{*(a)})$, where $\mathbf{y}_d^{*(a)} = (Y_{d1}^{*(a)}, \dots, Y_{dN_d}^{*(a)})'$ are the values of the response variable in the area d in the bootstrap census a.

4. The ELL estimator is obtained by averaging over the A censuses:

$$\hat{\delta}_d^{\mathrm{ELL}} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{*(a)}.$$

In addition, in this method, the MSE is estimated as follows

$$\mathrm{mse}_{\mathrm{ELL}}(\hat{\delta}_d^{\mathrm{ELL}}) = \frac{1}{A}\sum_{a=1}^{A} (\,\delta_d^{*(a)} - \hat{\delta}_d^{\mathrm{ELL}})^2.$$

It is easy to see that, for areas of large population size N_d (usually the case in real applications), if we compute the ELL estimator of the area mean $\bar{Y}_d$, by averaging $\bar{Y}_d^{*(a)} \approx \bar{\mathbf{X}}_d{}'\hat{\boldsymbol{\beta}} + u_d^{*(a)}$ across the A censuses, the average of the random bootstrap effects $u_d^{*(a)}$, along the bootstrap replicates, is $A^{-1}\sum_{a=1}^{A} u_d^{*(a)} \approx \mathbb{E}[u_d] = 0$. Therefore, the ELL estimator of the area mean, $\hat{\bar{Y}}_d^{\mathrm{ELL}} = \mathbb{E}[\bar{Y}_d]$, turns out to be the synthetic-regression estimator

$$\hat{\bar{Y}}_d^{\mathrm{ELL}} = \bar{\mathbf{X}}_d{}'\hat{\boldsymbol{\beta}}.$$

The reason for this is that the marginal expectation $\mathbb{E}[\delta_d]$, unlike the conditional expectation on the sample data, does not use the sample observations and therefore adheres to the prediction obtained through the linear regression without taking into account the area effects (they vanish). Therefore, the ELL estimator has the same problems as the regression-synthetic estimator. Specifically, it can be considerably biased if the regression model without area effects does not hold, that is, if the considered auxiliary variables do not fully explain the between-area heterogeneity of the response variable.

Furthermore, in the above bootstrap procedure, unlike in usual bootstrap methods, the model is *not* fitted again, and indicators are not re-estimated with each bootstrap sample (which should be drawn from the bootstrap censuses to replicate the real world). Therefore, the real world process is not being replicated in the ELL bootstrap procedure. As a result, the MSE estimated using this method does *not correctly reproduce the error incurred in the real world*. Finally, in the original ELL method, the random effects included in the model are actually for clusters (first stage sampling units) and not for the areas of interest. If this model is considered but the available auxiliary variables do not fully explain the between-area heterogeneity, the ELL estimate of the MSE can be seriously understating the true MSE of the ELL estimator.

EB method

Molina and Rao (2010) proposed to estimate general nonlinear indicators using the *Best/Bayes predictor* (BP) based on the nested error model. This method assumes that the variables $Y_{di} = \log(E_{di} + c)$ follow the model (11) with normally distributed random effects u_d and errors e_{di}. Under this model, the variable vectors for each area, $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$, $d = 1, \dots, D$, are independent and verify $\mathbf{y}_d \overset{\mathrm{i}}{\sim} N(\boldsymbol{\mu}_d, \mathbf{V}_d)$, with mean vector

$\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta}$, where $\mathbf{X}_d = (\mathbf{x}_{d1}, \dots, \mathbf{x}_{dN_d})'$, and covariance matrix $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_d{}' + \sigma_e^2 \mathbf{A}_d$, where $\mathbf{A}_d = \text{diag}(k_{di}^2; i = 1, \dots, N_d)$.

For a general indicator defined as a function of $\mathbf{y}_d$, that is, $\delta_d = \delta_d(\mathbf{y}_d)$, the best predictor is the one that minimizes the MSE and is given by

$$\tilde{\delta}_d^{\text{B}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{y}_{dr}}[\delta_d(\mathbf{y}_d)|\mathbf{y}_{ds}; \boldsymbol{\theta}], \qquad (12)$$

where the expectation is taken with respect to the distribution of the out-of-sample values $\mathbf{y}_{dr}$ from area d, given the sample values $\mathbf{y}_{ds}$. This conditioned distribution depends on the true value $\boldsymbol{\theta}$ of the parameters of the model for $\mathbf{y}_{ds}$. Replacing $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}$ in the best predictor (12), we get the so-called *Empirical Best/Bayes* (EB) predictor, $\hat{\delta}_d^{\text{EB}} = \tilde{\delta}_d^{B}(\widehat{\boldsymbol{\theta}})$. The usual ML and REML fitting methods based on the normal likelihood provide consistent estimators even if normality does not hold, under certain regularity conditions.

Under the nested error model (11), the distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$, required for calculation of the best predictor (12), is obtained as follows. First, we decompose the matrices $\mathbf{X}_d$ and $\mathbf{V}_d$ in the sample and the out-of-sample parts in a similar way as we have decomposed $\mathbf{y}_d$, that is,

$$\mathbf{y}_d = \begin{pmatrix}\mathbf{y}_{ds}\\ \mathbf{y}_{dr}\end{pmatrix}, \quad \mathbf{X}_d = \begin{pmatrix}\mathbf{X}_{ds}\\ \mathbf{X}_{dr}\end{pmatrix}, \quad \mathbf{V}_d = \begin{pmatrix}\mathbf{V}_{ds} & \mathbf{V}_{dsr}\\ \mathbf{V}_{drs} & \mathbf{V}_{dr}\end{pmatrix}.$$

Since $\mathbf{y}_d$ follows a normal distribution, then all the conditionals also follow a normal distribution; specifically,

$$\mathbf{y}_{dr}|\mathbf{y}_{ds} \overset{\text{i}}{\sim} N(\boldsymbol{\mu}_{dr|s}, \mathbf{V}_{dr|s}), \quad d = 1, \dots, D, \qquad (13)$$

where the conditional mean vector and the corresponding covariance matrix take the form

$$\begin{aligned}\boldsymbol{\mu}_{dr|s} &= \mathbf{X}_{dr}\boldsymbol{\beta} + \gamma_d(\overline{y}_{da} - \overline{\mathbf{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d-n_d}, \qquad (14)\\ \mathbf{V}_{dr|s} &= \sigma_u^2(1-\gamma_d)\mathbf{1}_{N_d-n_d}\mathbf{1}_{N_d-n_d}^T + \sigma_e^2\text{diag}_{i\in r_d}(k_{di}^2), \qquad (15)\end{aligned}$$

where $\mathbf{1}_k$ is a vector of ones of size k.

For indicators $\delta_d = \delta_d(\mathbf{y}_d)$ with complex shape, it might not be possible to calculate analytically the expectation defining the best predictor. In these cases, the best predictor can be approximated empirically using Monte Carlo simulation. The process is as follows:

1. Get an estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \hat{\sigma}_u^2, \sigma_e^2)'$ of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ by fitting the model (11) to the sample data $(\mathbf{y}_s, \mathbf{X}_s)$.

2. Generate, for $a = 1, \dots, A$, response variable vectors for units outside the sample of area d, $\mathbf{y}_{dr}^{(a)}$, from the distribution of $\mathbf{y}_{dr}|\mathbf{y}_{ds}$ given in (13)–(15), with $\boldsymbol{\theta}$ replaced by its estimator $\widehat{\boldsymbol{\theta}}$ obtained in Step 1.

3. Augment the generated vector $\mathbf{y}_{dr}^{(a)}$ with the sample data $\mathbf{y}_{ds}$ to form a census vector for the area d, $\mathbf{y}_d^{(a)} = (\mathbf{y}_{ds}', (\mathbf{y}_{dr}^{(a)})')'$. Using $\mathbf{y}_d^{(a)}$, compute the indicator of interest $\delta_d^{(a)} = \delta_d(\mathbf{y}_d^{(a)})$ and repeat for $a = 1, \dots, A$.

4. The Monte Carlo approximation of the EB predictor of the indicator δ_d is obtained by averaging the indicators for the A simulated censuses, that is,

$$\hat{\delta}_d^{\mathrm{EB}} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{(a)}. \qquad (16)$$

Note that, for non-linear area indicators δ_d, both ELL and EB estimators require, apart from the survey observations from the target and auxiliary variables, a census with the unit-level values of the auxiliary variables $\{\mathbf{x}_{di}; i = 1, \dots, N_d,\ d = 1, \dots, D\}$. In principle, the EB method additionally requires identifying the survey units in the census of the auxiliary variables (to identify sample and out-of-sample units). Linking the survey and census is not always possible in practice. However, in practice, the area sample size n_d is typically very small compared to the population size N_d. In this case, we can use the so-called *Census EB predictor* proposed by Correa, Molina, and Rao (2012), which avoids identifying the sample units in the census. A Monte Carlo approximation to the Census EB predictor can be obtained with the same procedure as above but replacing the out-of-sample vector $\mathbf{y}_{dr}^{(a)}$ by the full census vector $\mathbf{y}_d^{(a)}$. More precisely, by generating in Step 2 full area censuses as $\mathbf{y}_d^{(a)} = \boldsymbol{\mu}_{d|s} + v_d^{(a)}\mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_d^{(a)}$, where $\boldsymbol{\mu}_{d|s} = \mathbf{X}_d\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\mathbf{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d}$ and $\boldsymbol{\varepsilon}_d^{(a)} \sim N(\mathbf{0}_{N_d}, \sigma_e^2\mathrm{diag}_{i=1,\dots,N_d}(k_{di}^2))$. If the area sampling fraction n_d/N_d is negligible, the Census EB predictor will be approximately equal to the original EB.

In the case of complex indicators, calculating analytical approximations for the MSE of the corresponding EB predictors is complicated. Molina and Rao (2010) describe a parametric bootstrap method for estimating the MSE based on the bootstrap method for finite populations of González-Manteiga et al. (2008). This method consists on performing the following steps:

1. Fit the model to the sample data $\mathbf{y}_s = (\mathbf{y}_{1s}, \dots, \mathbf{y}_{Ds})'$, getting estimates of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$.

2. Generate the bootstrap area effects as

$$u_d^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_u^2), \quad d = 1, \dots, D.$$

3. Generate, independently of $u_1^{*(b)}, \dots, u_D^{*(b)}$, the bootstrap errors as

$$e_{di}^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_e^2), \quad i = 1, \dots, N_d,\ d = 1, \dots, D.$$

4. Generate a bootstrap population (or census) of the response variable values through the model

$$Y_{di}^{*(b)} = \mathbf{x}_{di}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i = 1, \ldots, N_d,\ d = 1, \ldots, D.$$

5. Define the census vector of the response variable for area d as $\mathbf{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \ldots, Y_{dN_d}^{*(b)})'$. Compute the indicators of interest from the bootstrap census $\delta_d^{*(b)} = \delta_d(\mathbf{y}_d^{*(b)}), d = 1, \ldots, D$.

6. For the original sample $s = s_1 \cup s_D$, let $\mathbf{y}_s^{*(b)} = ((\mathbf{y}_{1s}^{*(b)})', \ldots, (\mathbf{y}_{Ds}^{*(b)})')'$ be the vector with the bootstrap observations for the units with indexes in the sample, that is, containing the variables $Y_{di}^{*(b)}$, $i \in s_d$, $d = 1, \ldots, D$. Fit the model (11) again to the bootstrap sample data $\mathbf{y}_s^{*(b)}$ and get the EB bootstrap predictors for the area indicators of interest, $\hat{\delta}_d^{\mathrm{EB}*(b)}, d = 1, \ldots, D$.

7. Repeat Steps 2–6 for $b = 1, \ldots, B$ and obtain the real bootstrap values, $\delta_d^{*(b)}$, and the corresponding bootstrap EB predictors, $\hat{\delta}_d^{\mathrm{EB}*(b)}$, for each area $d = 1, \ldots, D$ and for each bootstrap replicate $b = 1, \ldots, B$.

8. The "naïve bootstrap" estimators of the MSE of the EB predictors $\hat{\delta}_d^{\mathrm{EB}}$ are given by

$$\mathrm{mse}_B(\hat{\delta}_d^{\mathrm{EB}}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\delta}_d^{\mathrm{EB}*(b)} - \delta_d^{*(b)}\right)^2, \quad d = 1, \ldots, D.$$

When sample units cannot be identified in the census, this bootstrap procedure can be adapted to get estimators of the MSE for the Census EB predictors. In this case, instead of generating the bootstrap censuses $\mathbf{y}_d^{*(b)}$ in Step 5 and then extracting the sample elements $\mathbf{y}_{ds}^{*(b)}$) in Step 6 (which cannot be identified in the census), we can generate the sample units $\mathbf{y}_{ds}^{*(b)}$) separately from the census, from their corresponding model using the survey design matrix $\mathbf{X}_{ds}$. The bootstrap sample data is then used to compute the Census EB predictors instead of the original EB ones.

3- Application: Poverty mapping in Palestine

In this section we apply small area estimation methods for poverty mapping in Palestine. Our data sources are the Palestinian Expenditure Consumption Survey (PECS) corresponding to 2016/17 and the Population Census of 2017. Our domains of interest are localities, which are nested within Palestinian governorates. From the 319 localities appearing in the census, only $D = 162$ are sampled in the PECS. As welfare measure E_{di}, we consider monthly expenditure per adult equivalent (in ILS), which is observed in the PECS but not in the census. The census contains several variables that are measured in a similar way in the PECS. Some of these variables are potentially related with the above welfare measure and will be used as covariates in a model for the welfare.

The indicators of interest are poverty rates and gaps. For a given locality d, the poverty rate and gap are obtained respectively taking $\alpha = 0,1$ in the following expression

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^{\alpha} I(E_{di} < z), \quad \alpha \geq 0.$$

The poverty line, measured in terms of the above welfare measure, is $z = 10{,}027$ ILS. Based on this poverty line, approximately 26% of the Palestinian population is below the line.

Concerning population and sample sizes, after removing certain records with missing data, in the census we have 4,266,953 records and in the survey, we have 18,383. With this sample size, the sampling fraction is about 43/10000. We have approximately half of the observations for each gender (9,119 for women and 9,244 for men). As to the sample size allocation by region, the West Bank is much better represented in the PECS, with 13,216 observations (out of 2,395,774) in contrast with 5,147 for Gaza (out of 1,871,179). Even if the overall sampling fraction is not small, when we disaggregate by localities and gender, the sample sizes are actually really small, see Table 1 below.

Table 1: Summary of PECS sample sizes of localities for each gender.

	Min	1st Qu	Median	Mean	3rd Qu.	Max
Women	14	26	35	56.29	61.5	405
Men	13	28	36	57.06	63.0	464

Note that the poverty rate is a proportion. Consider the simple case of estimating a population proportion p under simple random sampling with the sample proportion $\hat{p}$. With a true proportion around $p = 0.26$ (poverty rate at the national level), the minimum sample size that we need to have an estimated coefficient of variation, $\mathrm{CV}(\hat{p}) = \sqrt{(1-p)/(pn)}$, below 20%, is $n = 71$. According to Table 1, at least three quarters of the localities have a CV exceeding 20% for the two genders. This means that direct estimators are not reliable for all these localities. In fact, we computed Hájek estimator $\hat{F}_{\alpha d}^{\mathrm{HA}}$ of $F_{\alpha d}$ for each locality d as in (4) of Section 3.1, noticing that our indicators can be expressed as area means $F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}$ of the variables

$$F_{\alpha,di} = \left(\frac{z - E_{di}}{z}\right)^{\alpha} I(E_{di} < z), \quad i = 1, \dots, N_d.$$

The resulting direct estimates of poverty rates and gaps take the value zero (because of zero individuals with welfare below the poverty line z) for 32 localities when looking at men and 29 for women. Thus, for those localities, direct estimates do not really make sense.

The area-level Fay–Herriot model described in Section 3.2.1 uses as response variables the direct estimators, which are zero for many localities. For those localities, the estimates of the sampling error variances ψ_d become also zero, which could be interpreted as no error, while these estimators are obviously subject to very large sampling error due to the small sample sizes. As a consequence, we get $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) = 1$ and the resulting EBLUPs based on this model, given in (9), reduce to the corresponding direct estimators for those localities, which make no sense and have also misleading error measures. In fact, even if we wished to estimate only for the localities where the direct estimates are not zero, the usual estimates of the MSEs of the EBLUPs based on the FH model require normality, and in this case the shape of the histograms of direct estimators (not shown here for brevity) is fairly different from the normal distribution. Consequently, these estimators are not recommended in this application.

The mentioned problems of area-level models in this application and the very rich census microdata that are available make the unit-level models the most suitable small area techniques in this application. For this reason, we apply the Census EB method based on the unit-level nested error model given in (11), to find estimators of our poverty indicators for each locality, $F_{\alpha d}$, for $\alpha = 0,1$, $d = 1, \dots, D$. As response variable in the nested error model, we take the transformed welfare variables $Y_{di} = T(E_{di}) = \log(E_{di} + c)$, for $c = 1000$. The distribution of the transformed welfare measures really looks like normal, see Figure 1 for women. For men the plot looks practically identical.

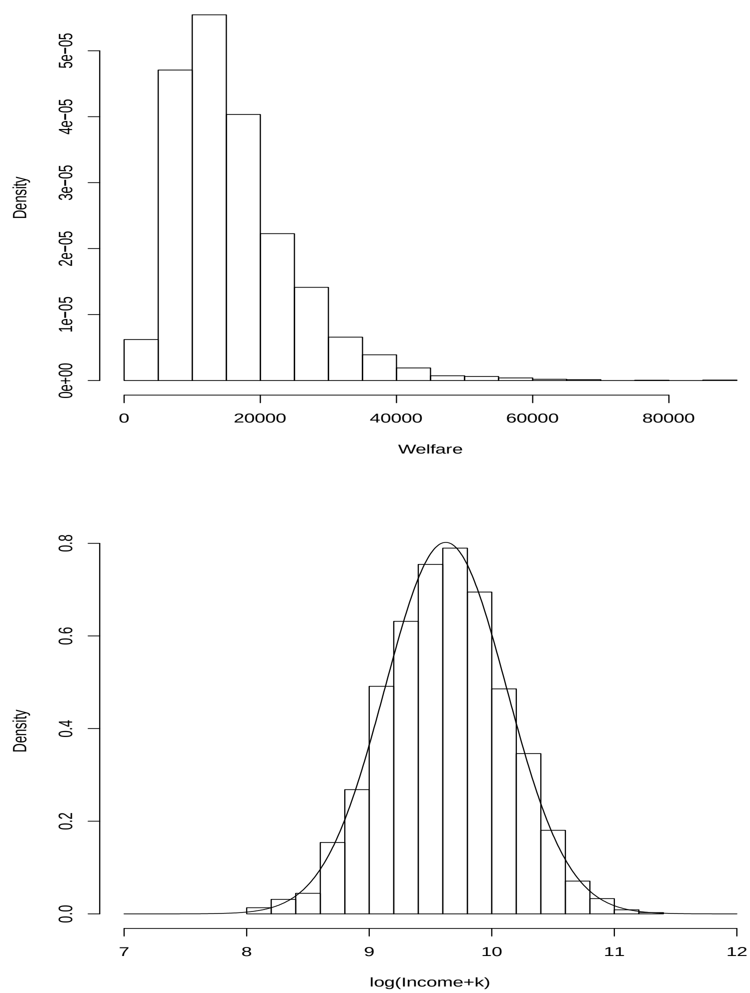


Figure 1: Histogram of welfare measure in the original scale (up) and after shift and log transformation (down), for women.

As explanatory variables in the nested error model, we considered location descriptors, household characteristics, attributes of the household head, dwelling characteristics, and types of supplies and amenities. Concretely, we included indicators of region (Gaza, West Bank) and of type of locality (rural and urban, camp). Concerning the household characteristics, we included size, proportion of females and employed ratio. The considered attributes of the household head were indicators of being unemployed, of ever been employed in Israel/settlement, of ever been employed for the national government, of refugee status, of having some difficulty, of never attended school and of having education level higher than secondary. Dwelling characteristics included type of dwelling (villa, separate room, other), type of tenure (rented, other) and number of rooms. Finally, within commodities or supplies, we included the type of water, waste and heating systems. Finally, we included indicators of owing washing machine, freezer, microwave, dishwasher, LED/LCD TV, electricity fan, air conditioning, central heating, solar boiler, phone line, home library, computer, iPad/tablet, and smartphone. Separate models were fitted for each gender, using exactly the same covariates for both.

After fitting the models by REML, practically all the categories of the considered explanatory variables were significant for both genders and the R^2 in the linear model for each gender was over 53% for both models. Let us now check the usual model assumptions. Unit level residuals from the model show a normal distribution, see Figure 2 for men (plots are almost identical for women). Moreover, a scatterplot of model residuals versus predicted values, shown in Figure 3 for women, shows no pattern at all, which gives no indication against the linearity model assumption (similarly for men). Finally, the histogram and qq-normal plot of area-level residuals (predicted locality effects) displayed in Figure 4 neither show serious departure from the normality assumption of the area effects.

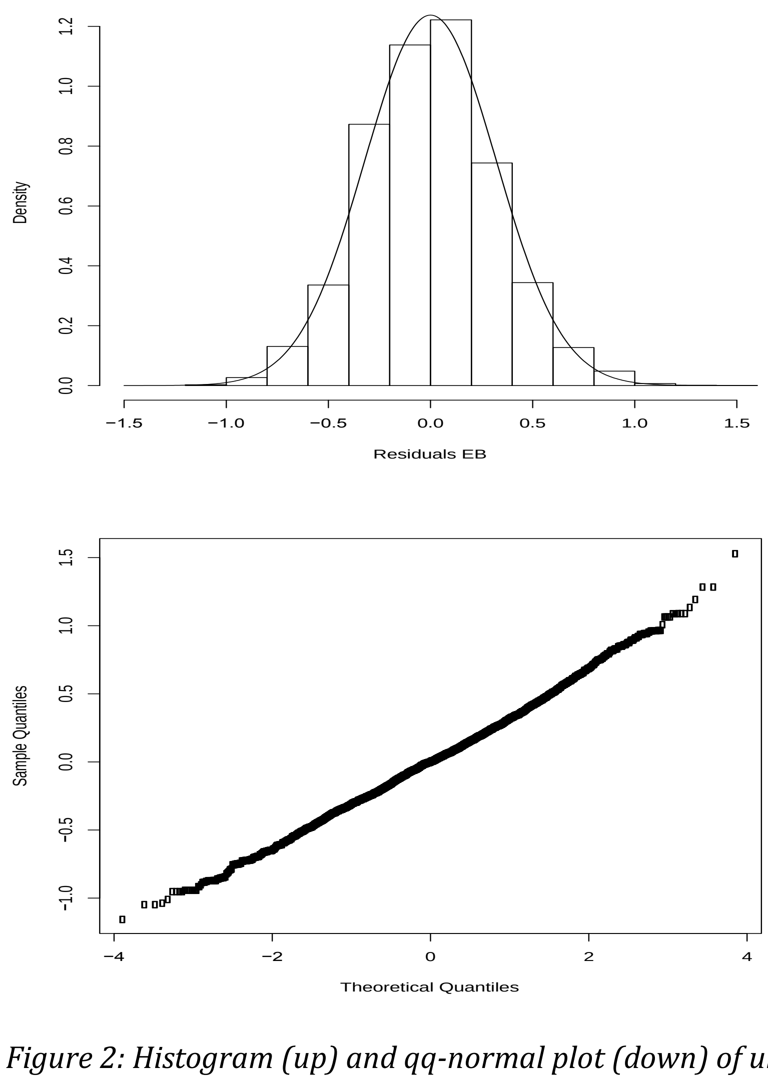


Figure 2: Histogram (up) and qq-normal plot (down) of unit level residuals for men.

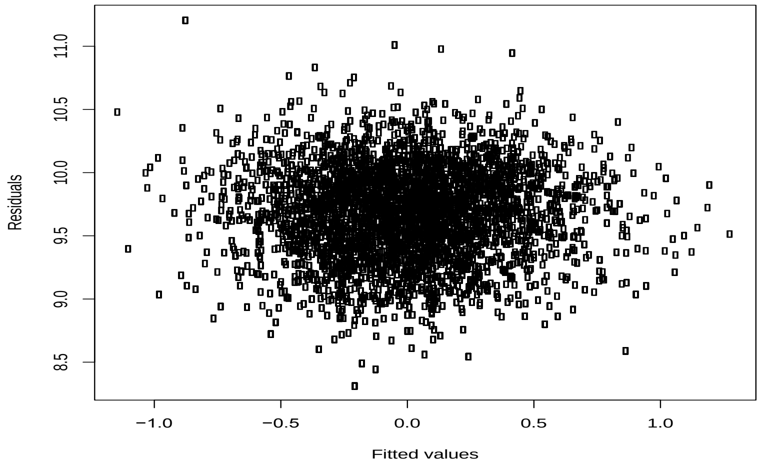


Figure 3: Unit level residuals versus fitted values in the model for women.

Along with weighted direct (Hájek) estimates, since data does not indicate any model departure, Census EB predictors of poverty rates and gaps were computed based on the considered nested-error model. For the considered poverty indicators $\delta_d = F_{\alpha d}$ for $\alpha = 0{,}1$, Census EB predictors can be computed analytically as

$$\hat{F}_{\alpha d}^{\text{CEB}} = \frac{1}{N_d}\sum_{i=1}^{N_d} \hat{F}_{\alpha,di}\,,$$

where $\hat{F}_{\alpha,di} = \mathbb{E}[F_{\alpha,di}|\mathbf{y}_{ds};\boldsymbol{\theta}]$ is the expectation of $F_{\alpha,di}$ with respect to the distribution of $Y_{di}|\mathbf{y}_{ds}$, given by

$$Y_{di}|\mathbf{y}_{ds} \sim N(\mu_{di|s}, \sigma_{di|s}^2),$$

with conditional mean and variance given by

$$\begin{aligned} \mu_{di|s} &= \mathbf{x}_{di}{}'\boldsymbol{\beta} + \gamma_d(\overline{y}_{da} - \overline{\mathbf{x}}_{da}^T\boldsymbol{\beta}), \\ \sigma_{di|s}^2 &= \sigma_u^2(1-\gamma_d) + \sigma_e^2 k_{di}^2. \end{aligned}$$

For $\alpha = 0{,}1$, the expectations are respectively given by

$$\begin{aligned} \hat{F}_{0,di} &= \Phi(\alpha_{di}), \\ \hat{F}_{1,di} &= \Phi(\alpha_{di})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{di|s} + \frac{\sigma_{di|s}^2}{2}\right)\frac{\Phi(\alpha_{di} - \sigma_{di|s})}{\Phi(\alpha_{di})} - c\right]\right\}, \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of a standard normal random variable and $\alpha_{di} = [\log(z+c) - \mu_{di|s}]/\sigma_{di|s}$, for $\mu_{di|s}$ and $\sigma_{di|s}^2$ defined above.

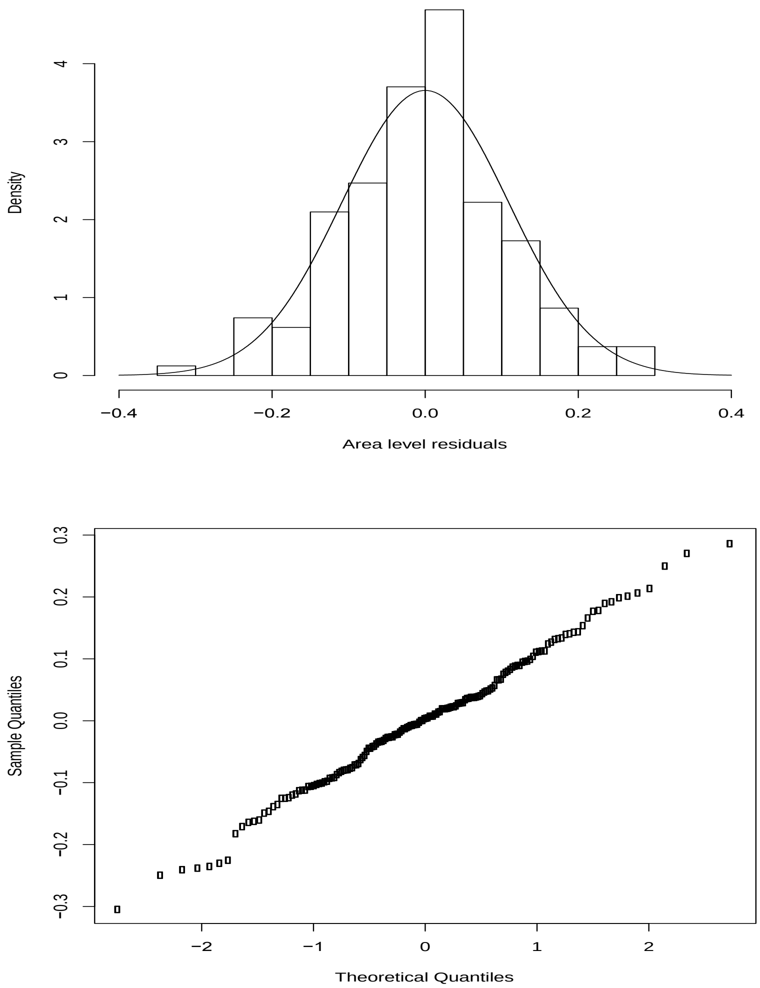


Figure 4: Histogram (up) and qq-normal plot (down) of area-level residuals (predicted locality effects) for men.

Let us now compare the resulting direct and Census EB estimates. Figure 5 displays Census EB estimates against direct estimates of locality poverty rates (up) and poverty gaps (down), for women. First of all, we can see the unreasonable zero values of the direct estimates for many localities. Census EB estimates do not take the value zero for any of the localities. On the other hand, we know that direct estimates are approximately unbiased for the localities with large sample sizes, when averaging across all the possible samples drawn with the same sampling mechanism. These plots show Census EB estimates for the sampled localities distributed around the line of equality with direct estimates, which suggests no serious systematic design bias of Census EB estimates. However, in the plot for the poverty rates (up), there appear a couple of points on the right that are further apart from the rest, indicating a much larger direct estimate of poverty rate than the corresponding Census EB estimate. Since in applications with real data the true values cannot be known, it is impossible to know whether direct estimates are overstating or Census EB estimates are understating the poverty rate for those localities. For men, conclusions are basically the same. Even if the models for both genders seem to fit very well in this application, perhaps model refinements can still be found.

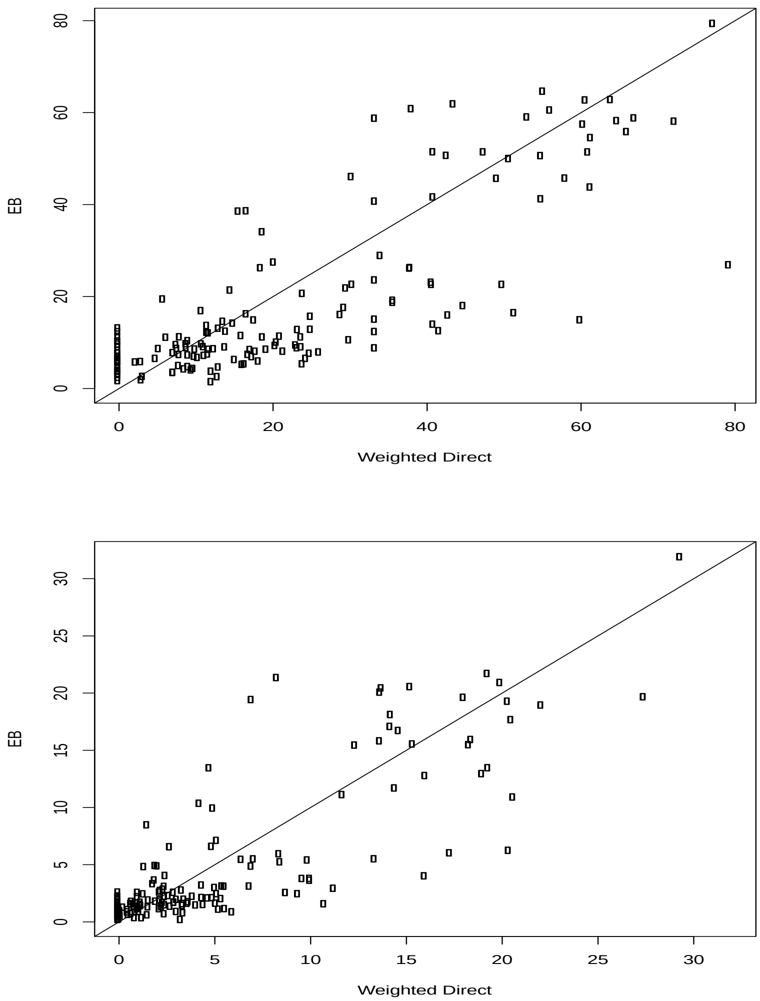
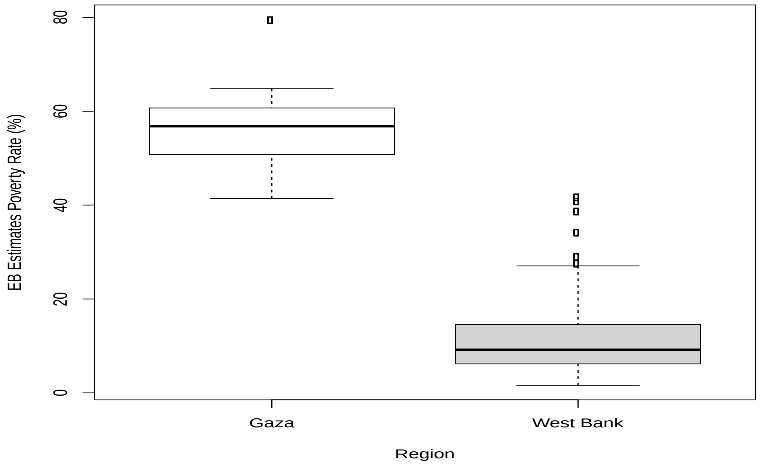


Figure 5: Census EB estimates versus weighted direct ones for locality percent poverty rates (up) and poverty gaps (down).

Now it is important to mention the huge differences in the estimated poverty indicators by region. Figure 6 shows boxplots of the locality poverty rates (up) and gaps (down) for women, for Gaza and West Bank. The median EB estimate of poverty rate for Gaza is about 55%, compared to 8.3% for West Bank. For the poverty gap, the median EB estimate is 17.4% for Gaza and 1.5% for West Bank. This great differences in the estimates leads us to analyze the remaining results separately for each region.



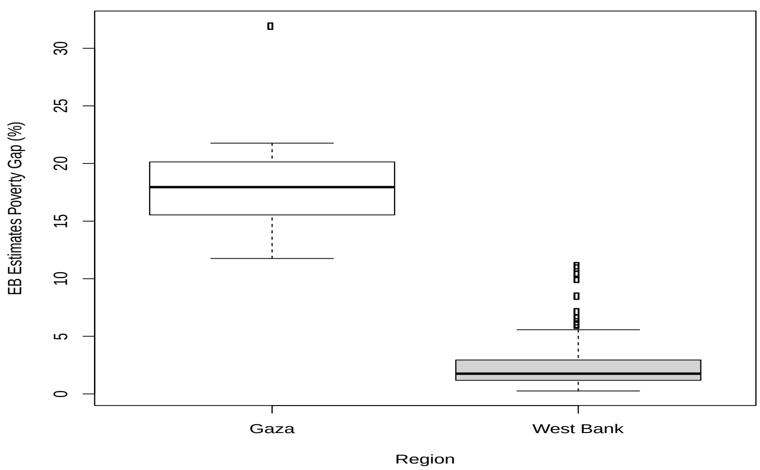


Figure 6: Boxplots of Census EB estimates of locality percent poverty rates (up) and percent poverty gaps (down).

Let us now compare the quality of Census EB and weighted direct estimates. In order to have comparable accuracy measures, we estimated the MSE for both types of estimators using exactly the same procedure. Concretely, we applied the parametric bootstrap method described in Section 3.2.3. Since previous poverty estimates were given in percentage, MSE estimates are now given $\times 10^4$. Figure 7 displays boxplots of the estimated MSEs for the two types of estimators, for women (up) and men (down) for the sampled localities. In the case of women, the median of the estimated MSEs of direct estimates is 47, compared with 6.7 for Census EB. For men, the median estimated MSEs are 45.8 and 5.5 for direct and Census EB respectively. If we look at the estimated MSEs for each locality (see Figure 8), the average percent decrease in estimated MSE of Census EB estimates with respect to direct estimators is 84% in the case of poverty rate. Moreover, there is gain in efficiency of Census EB estimates in all the localities except for one (although for that locality there is practically no loss). For poverty gaps, the gains are even larger on average.

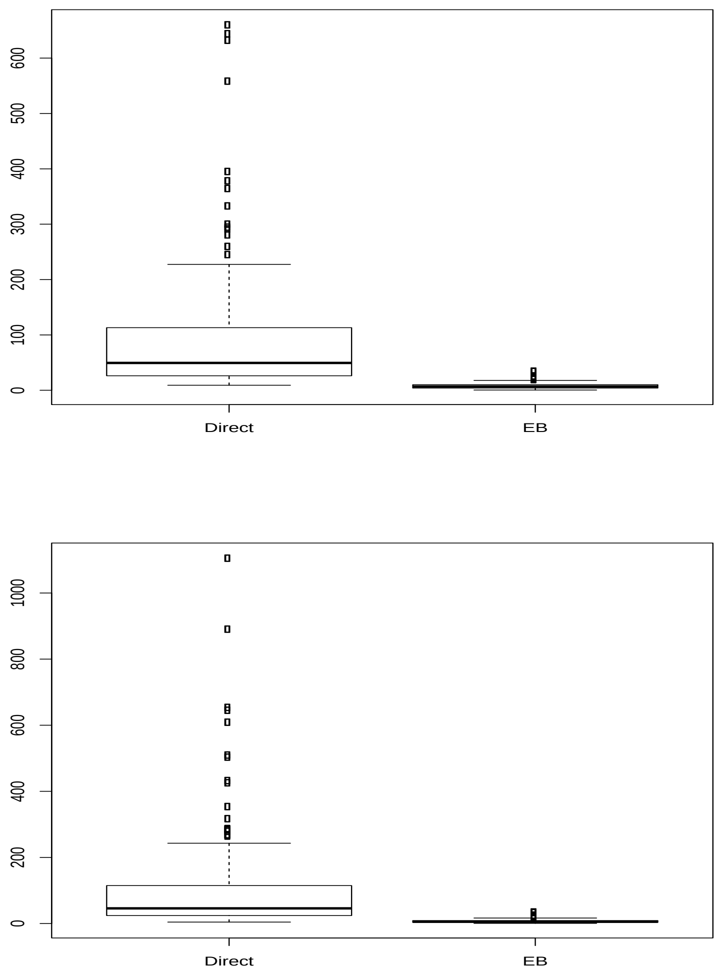
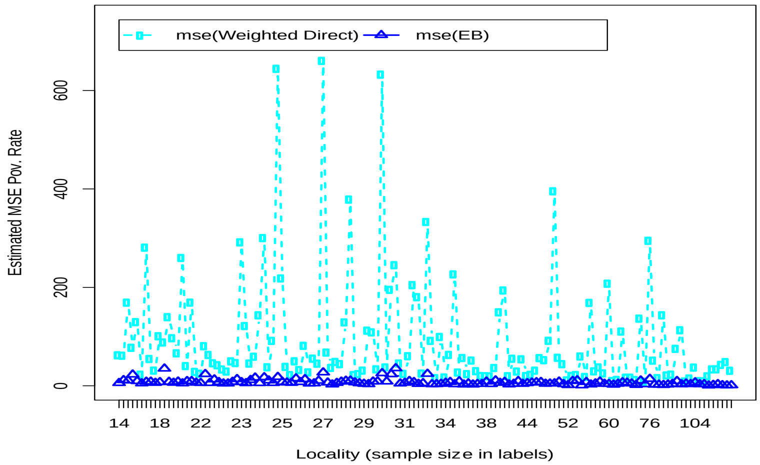


Figure 7: Boxplots of estimated MSEs of weighted direct and Census EB estimates of locality poverty rates for women (up) and men (down).



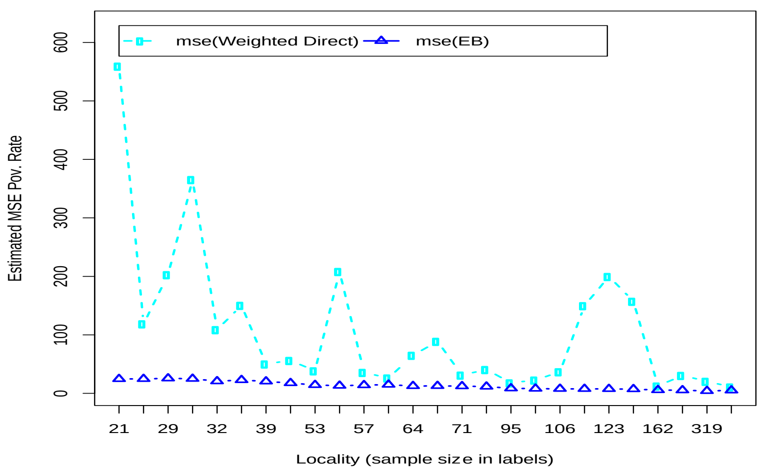
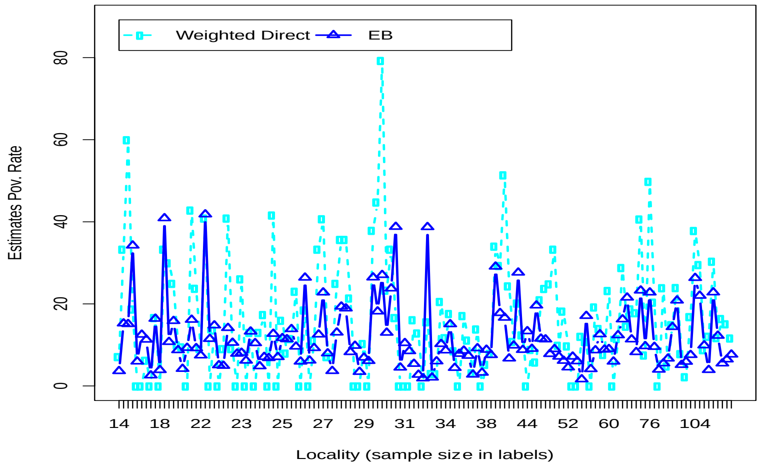


Figure 8: Estimated MSEs of weighted direct and EB estimators of poverty rates for each sampled locality in West Bank (up) and Gaza (down), for women. Localities are sorted from smaller to larger sample sizes, with locality sample sizes indicated in the x-axis labels.

Let us now compare Census EB and direct estimates for each sample locality. Figure 9 shows the two types of estimates of poverty rate for the sampled localities in West Bank (up) and Gaza (down), with localities (in x-axis) sorted from smaller to larger sample sizes, and with sample sizes indicated in the x-axis labels. See the much greater instability of direct estimates compared with Census EB ones in the two regions. We can also see the zero direct estimates for many localities. Summarizing, Census EB estimators are much more efficient than direct estimators for practically all the localities, more stable across localities, never take unreasonable zero values and seem not be seriously biased across the possible samples.



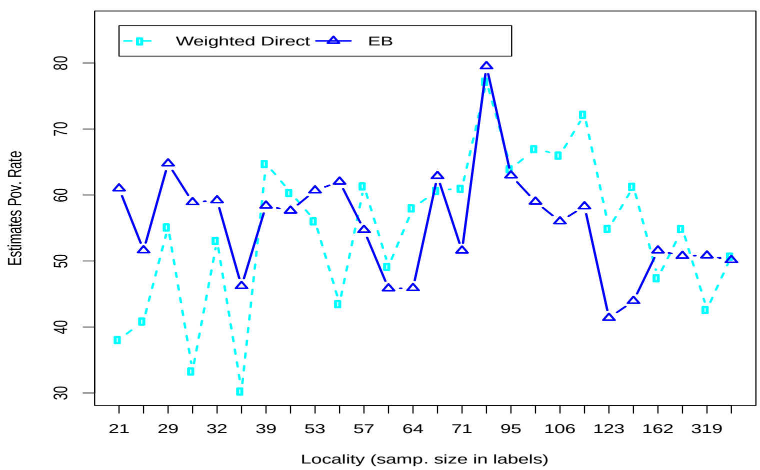


Figure 9: Weighted direct and Census EB estimates of poverty rates for each sampled locality in West Bank (up) and Gaza (down), for women. Localities are sorted from smaller to larger sample sizes, with locality sample sizes indicated in the x-axis labels.

Next, we compare the Census EB estimates for men and women. Looking at Figure 10, showing estimated poverty rates for West Bank (up) and Gaza (down), we can see only slight differences between estimates for the two genders. Although about 70% of the localities in West Bank have larger poverty rates for women (in 30% they are larger for men), Gaza does not practically show gender differences. Conclusions are similar for poverty gaps (Figure 11).

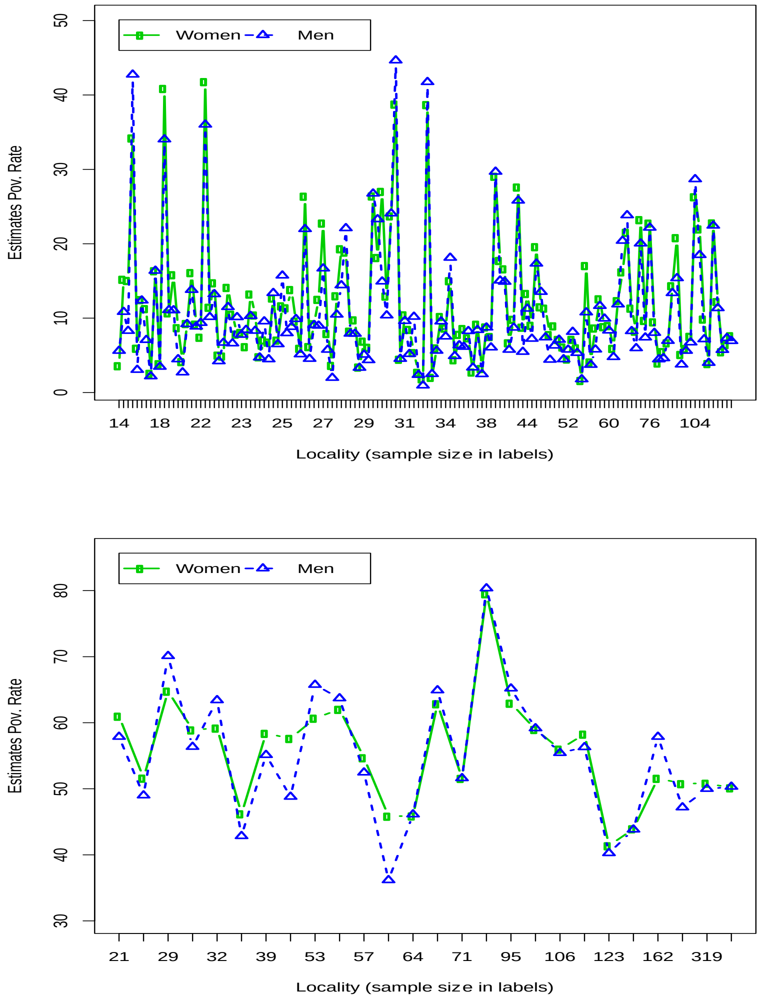
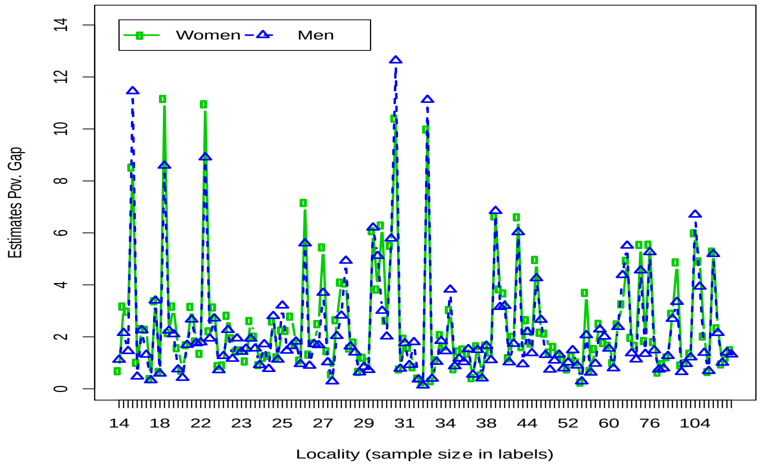


Figure 10: Census EB estimates of poverty rates for each sampled locality in West Bank (up) and Gaza (down), for men and women. Localities are sorted from smaller to larger sample sizes, with locality sample sizes indicated in the x-axis labels.



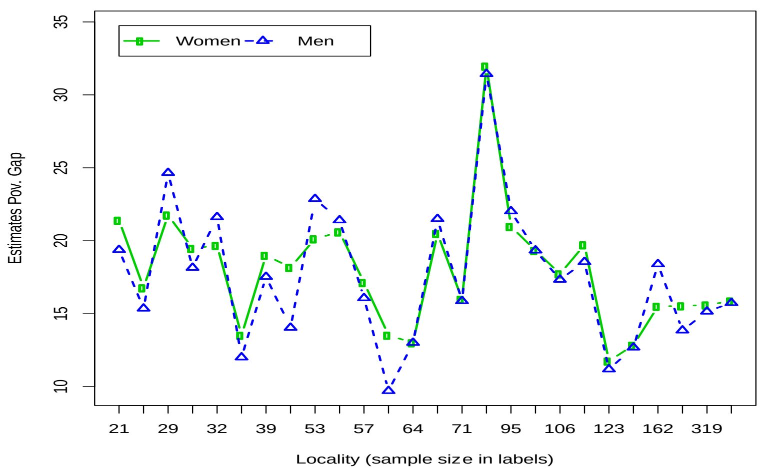


Figure 11: Census EB estimates of poverty gaps for each sampled locality in West Bank (up) and Gaza (down), for men and women. Localities are sorted from smaller to larger sample sizes, with locality sample sizes indicated in the x-axis labels.

Finally, the following tables summarize the Census EB predictor model as fitted by REML.

Coeffficients	Value	Std.Error	t-value	p-value
(Intercept)	9.129654	0.07032788	129.81557	0.0000***
regionWest Bank	0.163741	0.02971331	5.51069	0.0000***
loctype2.f3	-0.124346	0.03326557	-3.73797	0.0003***
hhsize_ae	-0.125092	0.00332163	-37.65991	0.0000***
femalep	-0.104234	0.02347222	-4.44075	0.0000***
head_age	0.010015	0.00218386	4.58577	0.0000***
head_age2	-0.000089	0.00002159	-4.13964	0.0000***
head_refugstatRegistered refugee	-0.009893	0.00939727	-1.05277	0.2925
head_refugstatUn-registered refugee	0.100793	0.03590492	2.80721	0.0050**
head_diffYes	-0.148510	0.01854243	-8.00918	0.0000***
head_neverschoolYes	-0.060934	0.02491440	-2.44574	0.0145*
head_secondaboveYes	0.038375	0.00939718	4.08363	0.0000***
employed_ratio	0.220275	0.02344003	9.39738	0.0000***
head_unemployed.f1	-0.094549	0.02014352	-4.69376	0.0000***
head_employisrasett.f1	0.104872	0.01103657	9.50224	0.0000***
head_employnatgov.f1	0.023796	0.01004289	2.36947	0.0178*
dwelltype2.fSeparate room	-0.300266	0.19087919	-1.57307	0.1157
dwelltype2.fVilla	0.241097	0.04346100	5.54744	0.0000***
tenure2.fRented	-0.060497	0.01626722	-3.71894	0.0002***
rooms	0.050280	0.00362796	13.85913	0.0000***
water_bottled	0.219500	0.02850772	7.69967	0.0000***
wasteOthers	0.083603	0.03113443	2.68523	0.0073***
wasteThrowing in the container	0.070630	0.03035760	2.32659	0.0200*
wasteThrowing outside	0.108581	0.04704006	2.30827	0.0210*
heating2.fDiesel	0.482247	0.09170628	5.25860	0.0000***
heating2.fElectricity	-0.053199	0.02067120	-2.57358	0.0101*
heating2.fGas	-0.044206	0.02161235	-2.04540	0.0408*
heating2.fNot available	-0.140460	0.02233715	-6.28818	0.0000***
heating2.fOther	0.160004	0.09648764	1.65828	0.0973.
heating2.fWood	-0.057447	0.02137587	-2.68747	0.0072**
freezer_ysno.f1	0.075935	0.01149020	6.60871	0.0000***
microwave_ysno.f1	0.027830	0.00863602	3.22259	0.0013**
dishwasher_ysno.f1	0.060347	0.02141698	2.81773	0.0048**
tv_ledlcd_ysno.f1	0.120884	0.00843869	14.32496	0.0000***

electric_fan_ysno.f1	0.070604	0.01205076	5.85887	0.0000***
air_conditioner_ysno.f1	0.130964	0.00989567	13.23450	0.0000***
central_heating_ysno.f1	-0.070338	0.04254488	-1.65326	0.0983.
solar_boiler_ysno.f1	0.029272	0.00827522	3.53728	0.0004***
phone_line_ysno.f1	0.075816	0.00866141	8.75331	0.0000***
home_library_ysno.f1	0.073284	0.01172368	6.25092	0.0000***
computer_ysno.f1	0.067749	0.00828415	8.17817	0.0000***
ipad_tablet_ysno.f1	0.080382	0.00912558	8.80841	0.0000***
smartphone_ysno.f1	0.120630	0.01124397	10.72840	0.0000***
washing_machine_ysno.f1	0.033221	0.01861235	1.78488	0.0743.
Observations	9,244			
Log-likelihood	-3,027.861			
AIC	6,147.722			
BIC	6,475.562			
$\hat{\sigma}_u$	0.1203215			
$\hat{\sigma}_\varepsilon$	0.3254558			

4- Final remarks and recommendations

Direct estimators are (at least approximately) unbiased under the sampling replication mechanism and do not require model assumptions. This is great when estimating in subpopulations with large sample sizes. However, for domains or areas with small sample size, they can take unreasonable values and can be highly unstable, possibly leading to serious changes in the estimates from one period to the next. Thus, they are of limited practical usefulness for small areas.

On the other hand, small area estimators based on models are obtained using the n observations from all of the areas or domains in the survey, where n is typically much larger than the target area sample size n_i. For this reason, model-based estimators are much more efficient. Unit-level models allow to estimate general monetary poverty indicators (perhaps several indicators using the same fitted model) and disaggregate the estimates at any other disaggregation level. In our application, the obtained Census EB estimates based on the nested error model take reasonable values for all the localities of interest, seem to be absent of serious systematic design bias and have smaller estimated MSEs than the corresponding direct estimators for nearly all the localities, with an average MSE reduction of 84% for poverty rates and even greater reduction for poverty gaps.

Model-based procedures require a thorough checking of the model assumptions for the actual data. In our application with Palestinian data, the considered model seems to fit really well the available data. However, slight model variations can be further explored, such as changing some of the covariates or trying with different groupings of the categories of these covariates. We have considered separate models for the two genders and included fixed region effects. However, gender differences are pretty small, whereas region differences are substantial. Separate modeling for the two regions but common for the two genders (so that sample sizes are not further reduced) could be also explored in this application. The types of models that one can consider in a given data application depends mainly on the shape of the auxiliary information and on the mathematical expressions of our target indicators. Of course, for the same available data sources and indicators, possibly different types of modelling can be applied. Although estimates derived from different models obviously cannot coincide exactly owing to the error inherent with any statistical figure, if the considered models fit well the available data, the corresponding estimates should agree to some extent.

In the application, we have estimated poverty rates and gaps, but any other poverty indicator based on the same welfare measure can be estimated without much more effort based on the same fitted model. However, indicators that depend on several characteristics together, such as multidimensional poverty indicators, have rarely been treated in the small area estimation literature and deserve special attention. Efficient estimation methods depending on several models at the same time (for the different variables involved in the indicator) or based on multivariate models that need to be developed for this problem.

In our application, we have derived estimates and their corresponding MSEs only for the sampled localities, i.e., those appearing in the PECS. Even if models allow to produce estimates also for non-sampled areas, we do not recommend it since model checking is not possible for those localities (they could be outlying). For fixed values of the model covariates, design biases of model-based estimators tend to increase as the sample size decreases, so in those localities this design bias could be significant.

References

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83: 28–36.

Bell, W. 1997. "Models for County and State Poverty Estimates." Statistical Research Division, U.S. Census Bureau.

Correa, L., I. Molina, and J.N.K. Rao. 2012. "Comparison of Methods for Estimation of Poverty Indicators in Small Areas." Unpublished report.

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro-Level Estimation of Poverty and Inequality." *Econometrica* 71: 355–364.

Fay, R.E., and R.A. Herriot. 1979. "Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–277.

Fuller, W.A. 1999. "Environmental Surveys over Time." *Journal of Agricultural, Biological and Environmental Statistics* 4: 331–345.

Ghosh, M., and R.C. Steorts. 2013. "Two-Stage Benchmarking as Applied to Small Area Estimation." *TEST* 22: 670–687.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2010. "Small Area Estimation Under Fay-Herriot Models with Nonparametric Estimation of Heteroscedasticity." *Statistical Modelling* 10: 215–239.

González-Manteiga, W., M. J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2008. "Bootstrap Mean Squared Error of a Small-Area EBLUP." *Journal of Statistical Computation and Simulation* 75: 443–462.

Molina, I., and J.N.K. Rao. 2010. "Small Area Estimation of Poverty Indicators." *The Canadian Journal of Statistics* 38: 369–385.

Prasad, N.G.N., and J.N.K. Rao. 1990. "The Estimation of the Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association* 85: 163–171.

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation*. Second edition. Hoboken, NJ: Wiley.