International Journal of Statistics and Probability; Vol. 12, No. 3; May 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Small Area Estimation of Economic Security: A Proposal

Mario Marino[1], Silvia Pacei[1]

[1] Department of Satistical Sciences, University of Bologna, Bologna, Italy

Correspondence: Silvia Pacei, Department of Statistical Sciences, University of Bologna, Bologna, Italy

Received: November 30, 2022 Accepted: April 17, 2023 Online Published: May 23, 2023

doi:10.5539/ijsp.v12n3p8 URL: https://doi.org/10.5539/ijsp.v12n3p8

Abstract

The objective of this work is to propose a small area estimation strategy for an economic security indicator. In the last decade the interest for the measurement of economic security or insecurity has grown constantly, especially since the financial crisis of 2008 and the pandemic period. In this work, economic security is measures through a longitudinal indicator that compares levels of equivalized household income over time. To solve a small area estimation problem, due to possible sample sizes too low in some areas, a small area estimation strategy is suggested to obtain reliable estimates of the indicator of interest. We consider small area models specified at area level. Besides the basic Fay-Herriot area-level model, we propose to consider some longitudinal extensions, including time-specific random effects following an AR(1) process or an MA(1) process. A simulation study based on EU-SILC data shows that all the small area models considered provide a significant efficiency gain with respect to the Horvitz-Thompson estimator, especially the small area model with MA(1) specification for random effects.

Keywords: small area estimation, economic security, Fay-Herriot model, time correlation

1. Introduction

The objective of this work is to propose a strategy to estimate an indicator of economic security for small areas, defined as geographical areas or domains for which the sample size is too low to obtain reliable estimates. The problem of small area estimation is particularly relevant when a sample survey is planned to provide reliable estimates at a specific geographical level or for specific domains, and reliable estimates for a more detailed level are demanded for some reasons. For example, the availability of reliable information at local level, may help to plan policies to improve households' well-being and reduce territorial economic inequality. This issue is particularly relevant for Italy, whose economic system is characterized by a strong territorial concentration of productive activities, with consequent effects on the territorial distribution of households' richness.

The major contributions to the small area estimation issue are, among others, those from Rao (2003), Rao and Molina (2015), Jiang and Lahiri (2006) and Pfeffermann (2013). The advantage of the small area methods suggested by these authors is that they allow to improve the reliability of "direct estimates" obtained for small area, where direct estimates are estimates obtained simply by using survey weights (for example using the Horvitz-Thompson estimator). The reliability of estimates is improved: i) by borrowing information from auxiliary variables available for the population from the Census or administrative archives; ii) linking the small areas through a model. Applied contributions to the small area estimation problem have so far focused mainly on poverty and inequality, but we think that economic insecurity/security represents a different concept, which also deserves to be investigated.

In this work we focus on economic security, the counterpart of economic insecurity. Economic security is a complex expression that carries a variety of meanings. Although there is no formal unambiguous definition, it is possible to provide a general definition by characterising it as a condition of well-being and wealth. It has strongly influenced the institutional political debate of the last decade in the West, especially since the financial crisis of 2008 (D'Ambrosio and Rohde, 2014) but even more in the last years after the economic consequences due to the pandemic period (Dvoryadkina et al., 2021).

The indicator considered in this work is the one proposed by Bossert et al. (2022), which has two previous formulations: Bossert and D'Ambrosio (2013) and D'Ambrosio and Rohde (2014). The main difference between this indicator and the previous ones (e.g., Osberg and Sharpe, 2002) is that the indicator proposed by Bossert et al. (2022) has been derived through an axiomatic approach and it benefits from some desirable properties. This indicator provides a score at individual level, based on a comparison over time of the individual's levels of an economic outcome, to capture individuals' ability to overcome an economic crisis and measure their confidence on their ability to recover after a

crisis. We consider the weighted average of the scores obtained for individuals to obtain an aggregate indicator at area level.

Moreover, there is no convergence among researchers on what is the most suitable outcome to study economic security. For example, some outcomes very commonly considered in the literature are wealth, income, consumption, but also employment, activity rate, or even the comparison of the same indicators of poverty between countries with different levels of well-being (Osberg and Sharpe, 2014).

In this work we consider small area models specified at area level. The basic small area model specified at area level was proposed by Fay and Herriot (1979). Given the nature of the indicator, we propose to consider some longitudinal extensions of the Fay-Herriot model. These longitudinal models include time-specific random effects considering autoregressive processes of order 1 (AR1; see Esteban et al., 2012) and moving average of order 1 (MA1; see Esteban et al., 2016).

We carry out a design-based simulation study to investigate the design-based properties of the estimates obtained from the small area models considered. To this purpose we use longitudinal data taken from the European Union Statistics on Income and Living Conditions (EU-SILC) referred to Italy. The results show a significant improvement in small area estimates obtained from the small area models suggested, especially in the case of MA1 model.

The rest of the paper is structured as follows. Section 2 presents the economic insecurity indicators employed in this work. In Sections 3 small area models considered are introduced together with the Bootstrap method used to estimates the variance of direct estimates. In Section 4 the simulation study is described, and results are presented, and Section 5 concludes.

2 The Economic Insecurity Index

Bossert and D'Ambrosio (2013, 2019) use an axiomatic approach to derive the economic insecurity index. This approach is shared with several other contributions aimed at proposing indexes in economics and social sciences. Some proprieties that a measure in this context should satisfy are defined and then the class of measure that can respect these axioms is develop (Thompson, 2001).

Given a $\mathrm{R}_{(T)}$ a $(T+1)$-dimensional Euclidean space and the vector $(T,\dots,0)$, where T is the number of past periods (lags) considered and 0 represents the current period, the economic insecurity index is defined as a function $I = \langle I^T \rangle_{T\in N}$ that for every $T \in \mathbb{N}$, gives $I^T: \mathbb{R}^{(T)} \to \mathbb{R}$. Applying this function to the stream vector of an outcome (for example income or wealth) available for an individual, $(w_0, w_1, \dots, w_T) \in \mathbb{R}^{(T)}$, a score representing the individual level of economic insecurity is obtained.

The following desirable axioms are satisfied by the index suggested by Bossert and D'Ambrosio (2019): Gain Loss monotonicity; Proximity monotonicity; Resources-variation monotonicity; Homogeneity; Translation invariance; Quasilinearity; Stationarity; Loss Priority. Gain Loss monotonicity guarantees that a lower level of insecurity is related to a stream with a gain in the earliest period compared to a stream without changes. Proximity monotonicity guarantees that a loss (or a gain) in the past has less weight than a loss (or a gain) in the present. Resources-variation monotonicity represents an extension of the two previous properties and established that a movement of the stream of the type "first down and then up", which means a loss and then straight forward a gain, produces a decrease in the insecurity score, while an increment in the score is given by a "first up then down movement". The "Homogeneity" property assures that, when the outcome stream is multiplied by a constant, then the score resulting from the index is also multiplied by the same constant. The "Translation Invariance" property assures that if the same amount is added to each outcome value of the stream, the resulting insecurity score remains the same. According to the "Quasilinearity" property, the insecurity score $I^T(w)$ corresponding to a stream $w \in \mathbb{R}^{(T)}$ can be expressed as a function of the differences $w_T - w_{T-1}$ observed for the $T-1$ most recent outcome levels. The "Stationarity" propriety guarantees that if two streams, p and q, are shifted of r periods in the past and s is assigned as outcome levels in the additional periods, the insecurity comparison associated with the two streams remains unchanged. Finally, the "Loss Priority" property establishes that, ceteris paribus, a loss has a stronger impact on individual insecurity than a gain of the same magnitude, in the same period.

The insecurity index that satisfies all these proprieties (Bossert and D'Ambrosio, 2019; Bossert et al., 2022) may be written as:

$$I^T(w) = l_0 \sum_{\substack{t\in\{1,\dots,T\}:\\ w_t > w_{t-1}}} \delta^{t-1}(w_t - w_{t-1}) + g_0 \sum_{\substack{t\in\{1,\dots,T\}:\\ w_t < w_{t-1}}} \delta^{t-1}(w_t - w_{t-1}) \quad (1)$$

where value w_t denotes the outcome level at time t. $l_0, g_0 \in \mathbb{R}_{++}$ and $\delta \in (0, g_0/l_0)$, such that $l_0 > g_0$, for all

$T \in \mathbb{N}$ and for all $w \in \mathbb{R}^{(T)}$. Coefficients l_0 and g_0 represent, respectively, the coefficient for the losses and the coefficient for the gains, chosen such as $l_0 > g_0$ (this allows the "Loss priority" propriety to be satisfied). δ represents a discount factor which is set equal for gains and losses. A higher value for the discount factor ensures a higher weight attached to the past and vice-versa.

The index is structured in two parts: the first summation captures all the periods where a loss has occurred, while the second summation regards all the periods where a gain has occurred. We choose to use for the coefficients involved in (1) the same values suggested in Bossert et al. (2022), $l_0 = 1$, $g_o = 15/16$ and $\delta = 0.9$, although the authors note that the results are quite robust with respect to different choices for these values. Moreover, in this work we choose the household equivalized income as individual outcome for the calculation of the index and, following Hacker et al. (2010), we choose to consider a security index and we take the negative value of (1).

3. Method

3.1 The Small Area Models Considered

Area level models consist of two models: a model linking the small area direct estimates to the underling parameters and a model linking the underling parameters to some auxiliary variables known for the population at small area level. Auxiliary information is known from the Census and/or administrative archives, and therefore it is free of sampling error.

The first area level model proposed in literature is the Fay-Herriot model (Fay and Herriot, 1979). This model represents the cornerstone of small area estimation and consists of two equations. The "sampling models" links the direct estimate to the underlying parameter:

$$y_d = \mu_d + e_d \qquad d = 1, \dots, D \qquad (2)$$

where y_d denotes the direct estimate of small area *d*, μ_d is the parameter of interest (in our case the average of the economic security indicator) in area *d*, and e_d represents the sampling errors independently distributed as $e_d|\mu_d \sim N(0, \sigma_d^2)$.

The "linking model" links the underlying parameter to some auxiliary information known at population level:

$$\mu_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d \qquad (3)$$

where vector $\mathbf{x}_d$ contains *p* auxiliary variables available from the Census or administrative archives for all the small areas, $\boldsymbol{\beta}$ is a vector of *p* regression coefficients, and u_d are model errors, assumed to be independent and identically distributed (i.i.d.) from $N(0,\sigma_u^2)$, with variance σ_u^2 unknown. Errors u_d are assumed to be independent from sampling errors e_d.

The variances of direct estimates, σ_d^2, are usually assumed to be known, and substituted with their sample estimates. In our case they are estimated using the bootstrap method and then smoothed using a Generalized Variance Function model, as described in Section 3.2.

Merging the sampling and the linking models, the extended form of the basic Fay-Herriot model is obtained:

$$y_d = \mathbf{x}_d'\boldsymbol{\beta} + u_d + e_d \qquad (4)$$

Small area estimates are obtained through the Empirical Best Linear Unbiased Prediction (EBLUP), which is a weighted combination of the direct estimator and the domain specific regression estimator (Prasad and Rao, 1990 and Datta and Lahiri, 2000). The BLUP is given by:

$$\tilde{y}_d = \mathbf{x}_d'\tilde{\boldsymbol{\beta}} + \lambda_d \cdot \left(y_d - \mathbf{x}_d'\tilde{\boldsymbol{\beta}}\right) \qquad (5)$$

$$\lambda_d = \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2} \qquad (6)$$

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d \cdot \mathbf{x}_d'}{\sigma_d^2 + \sigma_u^2}\right)^{-1} \cdot \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d \cdot \hat{y}_d}{\sigma_d^2 + \sigma_u^2}\right) \qquad (7)$$

EBLUP is simply obtained from BLUP by substituting σ_d^2 with the corresponding estimate, that can be obtained using moments, ML or REML method.

The MSE of EBLUP is obtained as follows:

$$MSE(\tilde{y}_d) = E(\tilde{y}_d - \mu_d)^2 = g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) \qquad (8)$$

where:

$$g_{1d}(\sigma_u^2) = \lambda_d \cdot \sigma_d^2 \tag{9}$$

$$g_{2d}(\sigma_u^2) = (1-\lambda_d)^2 \cdot \mathbf{x}_d' \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d \cdot \mathbf{x}_d'}{\sigma_d^2+\sigma_u^2}\right)^{-1} \mathbf{x}_d \tag{10}$$

The estimator for $MSE(\tilde{y}_d)$ depends on the method used to estimate σ_d^2. If it is estimated using moments or REML method, under regularity conditions it reduces to:

$$mse(\tilde{y}_d) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2) \tag{11}$$

where

$$g_{3d}(\sigma_u^2) = \sigma_d^4(\sigma_d^2 + \sigma_u^2)^{-3}\bar{V}(\hat{\sigma}_u^2) \tag{12}$$

and $\bar{V}(\hat{\sigma}_u^2)$ is the asymptotic variance of the $\hat{\sigma}_u^2$ estimator (Rao, 2003, Chapter 7).

Many extensions of the classical Fay-Herriot model (FH) have been developed over time. Considering the longitudinal nature of our indicator, we consider two extensions that take advantage of the availability of more waves, by adding time varying effects in the model and specifying an AR(1) or MA(1) process for them. We will call them AR1 and MA1 model respectively. The idea behind these two models is that the reliability of results can be improved by borrowing information both from space and time, using simultaneously time-varying effects and random effects. Therefore, to estimate the parameter for the last wave, previous information can be used through a longitudinal model.

Small area AR1 model (Rao and Yu, 1994; Esteban et al., 2012) can be written as follows:

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + e_{dt} \qquad d = 1,\dots,D;\ t = 1,\dots,T \tag{13}$$

where y_{dt} is the direct estimate of the parameter for area *d* and time *t*, $\mathbf{x}_{dt}$ is a *p* auxiliary variables vector for time *t*, $\boldsymbol{\beta}$ is the vector of regression coefficients, $u_{1,d}$ are area specific effects constant over time assumed i.i.d. $N(0,\sigma_1^2)$, $u_{2,dt}$ are time varying effects with common variance σ_2^2 and following an AR(1) process with parameter ρ, and e_{dt} are the sampling errors assumed to be independently distributed as $N(0,\sigma_{dt}^2)$.

In matrix notation, model (13) is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e} \tag{14}$$

where $\mathbf{y} = col_{1\le d\le D}(\mathbf{y}_d)$, $\mathbf{y}_d = col_{1\le t\le T}(y_{dt})$, $\mathbf{u}_1 = col_{1\le d\le D}(u_{1,d})$, $\mathbf{u}_2 = col_{1\le d\le D}(\mathbf{u}_{2,d})$, $(\mathbf{u}_{2,d}) = col_{1\le t\le T}(u_{2,dt})$, $\mathbf{e} = col_{1\le d\le D}(\mathbf{e}_d)$, $\mathbf{e}_d = col_{1\le t\le T}(e_{dt})$, $\mathbf{X} = col_{1\le d\le D}(\mathbf{X}_d)$, $\mathbf{X}_d = col_{1\le t\le T}(\mathbf{x}_{dt})$, $\mathbf{x}_{dt} = col_{1\le j\le p}(x_{dtj})$, $\boldsymbol{\beta} = col_{1\le j\le p}(\beta_j)$, $\mathbf{Z}_1 = diag_{1\le d\le D}(\mathbf{1}_T)$, $\mathbf{Z}_2 = \mathbf{I}_{D\cdot T}$. It is assumed that $\mathbf{u}_1 \sim N(\mathbf{0},\mathbf{V}_{u_1})$, $\mathbf{u_2} \sim N(\mathbf{0},\mathbf{V}_{u_2})$ and $\mathbf{e} \sim N(\mathbf{0},\mathbf{V}_e)$, where $\mathbf{V}_{u_1} = \sigma_1^2\mathbf{I}_D, \mathbf{V}_{u_2} = \sigma_2^2\Omega(\rho), \Omega(\rho) = diag_{1\le d\le D}(\Omega_d(\rho)), \mathbf{V}_e = diag_{1\le d\le D}(\mathbf{V}_{ed})$, $\mathbf{V}_{ed} = diag_{1\le t\le T}(\sigma_{dt}^2)$,

$$\Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & & & \rho^{T-2} \\ \vdots & & \ddots & \vdots & \\ \rho^{T-2} & & & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}_{T\times T} \tag{15}$$

The BLUPs for $\boldsymbol{\beta}$ and $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')$ for this model are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, where

$$\mathbf{V} = var(\mathbf{y}) = \sigma_1^2\mathbf{Z}_1\mathbf{Z}_1' + \sigma_2^2 diag_{1\le d\le D}(\Omega_d(\rho)) + \mathbf{V}_e = diag_{1\le d\le D}(\sigma_1^2\mathbf{1}_T\mathbf{1}_T' + \sigma_2^2\Omega_d(\rho) + \mathbf{V}_{ed}) = diag_{1\le d\le D}(\mathbf{V}_d). \tag{16}$$

Unknown parameters $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\beta})$ may be estimated using restricted maximum likelihood method and substituted in the following equation to obtain the EBLUP for the population mean:

$$\hat{\bar{Y}}_{dt}^{eblup} = \hat{\mu}_{dt} = \mathbf{x}_{dt}\hat{\boldsymbol{\beta}} + \hat{\mu}_{1,d} + \hat{\mu}_{2,dt} \tag{17}$$

The MSE of such EBLUP is given by:

$$MSE\left(\hat{\bar{Y}}_{dt}^{eblup}\right) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta}) + g_3(\boldsymbol{\theta}) \tag{18}$$

where:

$g_1(\boldsymbol{\theta}) = \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a}$

$g_2(\boldsymbol{\theta}) = [\mathbf{a}'\mathbf{X} - \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{V}_e^{-1}\mathbf{X}]\mathbf{Q}[\mathbf{X}'\mathbf{a} - \mathbf{X}'\mathbf{V}_e^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a}]$

$g_3(\boldsymbol{\theta}) \approx tr\left\{(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')'E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\right]\right\}$

and $\mathbf{a} = col_{1\le d\le D}(\delta_{dl}\mathbf{a}_l)$, $\mathbf{a}_l = col_{1\le k\le T}(\delta_{tk})$ and δ_{ij} is Kronecker's delta taking the value 1 if $i = j$ and 0 otherwise, $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{T} = \mathbf{V}_u - \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{V}_u$, $\mathbf{b}' = \mathbf{a}'\mathbf{Z}\mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}$.

Straightforward algebra to obtain the estimator of $MSE\left(\hat{\bar{Y}}_{dt}^{eblup}\right)$, $mse\left(\hat{\bar{Y}}_{dt}^{eblup}\right) = g_1\left(\hat{\boldsymbol{\theta}}\right) + g_2\left(\hat{\boldsymbol{\theta}}\right) + 2g_3\left(\hat{\boldsymbol{\theta}}\right)$, are reported in Esteban at al. (2012).

Small area MA1 model may be written using formulas already seen for model AR1, apart for the time correlation matrix $\Omega(\rho)$ that is substituted by $\Omega(\vartheta) = diag_{1\le d\le D}\left(\Omega_d(\vartheta)\right)$, where ϑ represents the parameter of the MA(1) process assumed for time varying effects. Matrix $\Omega_d(\vartheta)$ is given by:

$$\Omega_d(\vartheta) = \begin{bmatrix} 1+\vartheta^2 & -\vartheta & \cdots & 0 & 0 \\ -\vartheta & 1+\vartheta^2 & \cdots & & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & & \cdots & 1+\vartheta^2 & -\vartheta \\ 0 & 0 & & -\vartheta & 1+\vartheta^2 \end{bmatrix}_{T\times T} \tag{19}$$

EBLUP for the small area population mean and its MSE may be obtained as in equation (17) and (18), by estimating the unknown parameters by REML. Moreover, Esteban et al. (2016) proposed a Bootstrap strategy to estimate $MSE\left(\hat{\bar{Y}}_{dt}^{eblup}\right)$ for MA1 model.

In this work, Fay-Herriot, AR1 and MA1 models are estimated using R packages "sae" and "saery", and unknown parameters are estimated using REML method.

3.2 Bootstrap Variance Estimation

As it is customary, we assume that the variances of the direct estimates are known, and we substitute them with their respective estimates. Given the complex structure of the indicator considered, the estimates of these variances are obtained using a Bootstrap strategy, that is by repeatedly selecting *B* random samples with replacement from the survey sample by small area, calculating the weighting average of the economic security index for each replication by small area, and then calculating the variance of these estimates by small area. In this work we set *B*=1,000.

Nevertheless, Bootstrap variances calculated for small areas may be highly instable estimates, given the small number of sampling units available for small areas. For this reason, these variances are usually smoothed using a model. The procedure we used to this purpose belongs to the Generalized Variance Function approach (GVF) (Wolter, 2007). An important aspect to emphasize is that, although this methodology is widely used for the estimation of variances in household surveys, it is primarily a practical methodology and, overall, a general theory supporting the use of one specific model over another has not been developed yet. GVF allows to select a model that enables to link the direct estimates to the estimates of their variances.

4. Results of the Simulation Study

4.1 Data Used

A simulation is performed to understand the properties of the small area estimators suggested, considering their application to income survey data. A design-based simulation is chosen. The advantage of design-based simulation is that the estimator properties are evaluated under the randomised distribution, i.e. the distribution over all possible samples that could be selected from the population of interest under the sampling design. This allows to have a more realistic view of the small area estimation problem considered. In contrast, using model-based methods, inference is made with respect to the underlying models, that are always approximations. On the other hand, design-based simulations allow the robustness of model-based estimation methods to be assessed against misspecification, by

repeatedly sampling from a realistic population. The advantages between choosing a design-based or model-based simulation have been widely discussed, see for example Salvati at al. (2010), Pfeffermann (2013) and Warnholz and Schmid (2016).

In this simulation study, EU-SILC survey data are used. EU-SILC represents the main source of data for the periodic reports of the European Union on the socio-economic conditions and the spread of poverty in the member countries, with indicators focused on income and social exclusion, in a multidimensional approach to the problem of poverty and material deprivation. The sampling strategy adopted in EU-SILC is complex. The sample consists in a rotating panel in which in each successive year of the survey only a portion of the sample is replaced, so that each unit is expected to participate in the survey for 4 years. The sample of households that are introduced in each successive year is selected according to a two-stage stratified sampling scheme, a sampling strategy often used for household surveys. In the case of Italy, the first-stage units are municipalities, stratified by region and population size. In the second stage, households are randomly selected from the municipalities selected in the first stage.

In this simulation the sample is used as the synthetic population, and subdivided into 20 regions that represent the small areas of reference. Due to the lack of information on the first-stage sampling and on the region (NUTS2), we subdivide households within each territorial repartition (NUTS1) into a number of clusters equal to the number of regions in the territorial repartition (from a minimum of 2 regions to a maximum of 6 regions per territorial repartition). Clusters are defined according to the results of a k-means non-hierarchical cluster analysis carried out in each territorial repartition by considering two variables, the equivalized tax on income and social insurance contributions and the equivalized regular tax on wealth. The reason for this choice is that in Italy taxes have a small regional component. In the algorithm, the number of cluster is set equal to the number of regions in the territorial repartition. Clusters obtained are than ranked according to the average of the equivalized taxes, and matched with the regions having the same rank within the territorial repartition according to individual taxed paid on average in the population. We consider this approximation acceptable in the context of a simulation study. The resulting 20 clusters are considered target small areas from which to select random samples.

We set $T = 2$ and consider data referred to the following waves: 2014 (2012, 2013, 2014), 2015 (2013, 2014, 2015) and 2016 (2014, 2015, 2016). We repeatedly select 1,000 random sample. Simple random samples are selected from each area. Samples were drawn without replacement considering a 15% sampling rate.

4.2 Results

The size of samples repeatedly selected from the simulated population ranges from a minimum of 18 households in the smallest area to a maximum of 261 households in the largest one. 81 households are selected on average from the areas.

For each random sample of households and for each year, we calculate the weighted average (Horvitz-Thompson estimator) of the security scores obtained for the individuals in the households selected within each area. These averages represent the direct estimates. Then the Bootstrap technique and GVF procedure discussed in Section 2 are applied to estimate the variance of direct estimates. In particular, we use as inputs for the GVF procedure the weighted average of the economic security scores calculated for small area (direct estimates) and their respective Bootstrap variances. To select a unique function for every simulated sample, we apply the GVF procedure to the average of the direct area estimates and the average of the Bootstrap area variances obtained for the 1,000 simulated samples. Then we apply the same function to smooth the variances of direct estimates for all simulated samples.

After comparing several specifications for the smoothing model, we select the following model on the basis of AIC and BIC criteria:

$$log\big(cv^2(y_{dt})\big) = \beta_0 + \beta_1 log(y_{dt})$$

To compute this model, we used the ReGenesees package from R, which includes computational algorithms for GVF. Through this model we find the parameters we need to predict the estimated coefficients of variation that will allow us to obtain the smoothed variances, that is:

$$cv^2_{dt,(SMT)}(y_{dt}) = \sqrt{\big(\hat{\sigma}^2_{(SMT)}/2\big) \cdot e^{\hat{\beta}_0 + \hat{\beta}_1 log(y_{dt})}}$$

where $\hat{\sigma}^2_{(SMT)}$ is the variance of the model.

In the linking model of the small area models we include auxiliary variables calculated from data related to the populations of the Italian regions available from the National Statistical Institute and from fiscal archives. The model should be parsimonious, but in order to better explain the dependent variable, a number of regressors have to be selected in this case. This is also due to the nature of our indicator, which has a very high variability. The following auxiliary variables are chosen, by fitting a regression model for the averages of the regional direct estimates obtained from the simulated sample, and using a stepwise regression: average income from building (earned for a building), average

amount of retirement income, average individual income (calculated on the number of income earners), proportion of the working-age population, proportion of the working-age population calculated for foreigners, percentage of graduates, the ratio between the number of individuals aged 0-14 year-olds and the population between 15 and 64 years old.

The EBLUPs derived from Fay-Herriot, AR1 and MA1 models are obtained using the methodologies seen in Section 3. The following performance measures are calculated to compare the performance of the small area estimators proposed:

$$ARB = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{1000}\sum_{s=1}^{1000}\left(\frac{\hat{Y}_{ds}}{Y_d}-1\right)\right|$$

where *ARB* means average absolute relative bias and it is a measure of the bias of an estimator (see Rao, 2003). In this formula *d* denotes the area, while *s* = 1, ...,1000 denotes the sample, $\hat{Y}_{ds}$ is the estimate for the *d*-th area and the *s*-th sample (Direct, Fay-Herriot, AR1 or MA1) and Y_d is the parameter in population for the *d*-th domain. Then we measure the accuracy of estimates considering:

$$AMSE = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{1000}\sum_{s=1}^{1000}\left(\hat{Y}_{ds}-Y_d\right)^2$$

where $AMSE$ is the average mean-squared error of an estimator (Direct, Fay-Herriot, AR1 or MA1). Finally:

$$AEFF(St) = \sqrt{AMSE(Dir)/AMSE(St)},$$

where *St* denotes the small area estimator (FH, AR1 or MA1), measures the gain in efficiency provided on average by the small area estimator. AMSE and AEFF measure the accuracy of an estimator in terms of its mean square error.

Results are reported in Table 1. They clearly show the gain in efficiency provided by the small area estimators. MA1 model performs better than both FH and AR1 models in terms of bias. The ARB of the direct estimator is very close to zero, as expected. On the other hand, the ARB of the Fay-Herriot model is slightly lower than that of the AR1, but higher than that of the MA1, therefore MA1 must be preferred in terms of bias.

All small area models provide significantly lower value for AMSE than the direct estimator. The most efficient estimates are produced by MA1 model, that provides a gain in efficiency of 130% with respect to the direct estimator. It is followed by AR1 and then by FH model. It is possible to note that, even though AR1 model provides more biased estimates than FH model, it overall provides more efficient estimates than FH model. Therefore, the simulation highlights that the best model in our case is MA1, but AR1 also appears to provide an overall efficiency gain compared to the FH, although it results more biased.

Table 1. Average relative bias, average mean-squared errors, and average relative efficiency

	Direct	FH	AR1	MA1
ARB (%)	0,00	99,39	112,33	43,97
AMSE	5,81	2,77	2,53	1,09
AEFF (%)	-	144,94	151,51	230,62

Source: Our elaboration on the simulated population.

We focus on the best performing small area model, MA1, and we carry out a graphical analysis to understand the properties of this estimator better. Figure 1 highlights that as the sample size increases, the absolute relative bias decreases. It can be noticed that the highest values for the bias correspond to the areas with the lowest sample size, and that the bias decreases as the sample size increases. This highlights that the small area estimator based on the MA1 model tends to be asymptotically design-unbiased and consistent. In Figure 2 the square root of the ratio between *MSE*(*Dir*) and *MSE*(*MA*1) is plotted against the sample size, to show the efficiency gain provided by the small area estimator with respect to the direct estimator for different sample sizes. This gain seems to be particularly pronounced for those areas with low values of the sample size, although, in general, even in the largest areas, MA1 model provides a noticeable, although slightly less pronounced, gain in efficiency. Finally, Figure 3 compares the square root of the MSE of MA1 estimates with that of direct estimates. All observations are above the diagonal, thus highlighting that MA1 estimates are more reliable than direct estimates in all small areas.

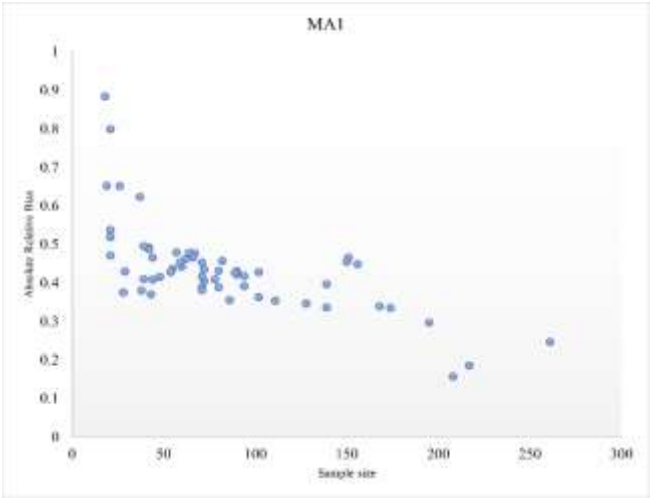

Figure 1. Absolute relative bias of MA1 estimates plotted against the domain sample size

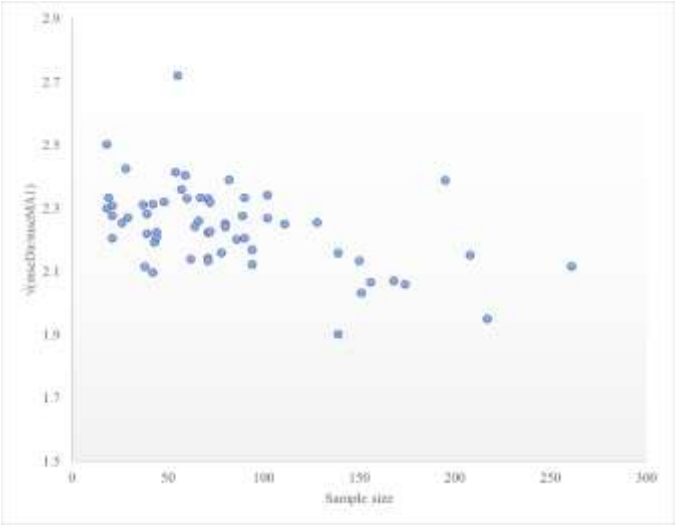

Figure 2. Comparison between $\sqrt{(\mathrm{MSE(Dir)/MSE(MA1)})}$ and the small area sample size

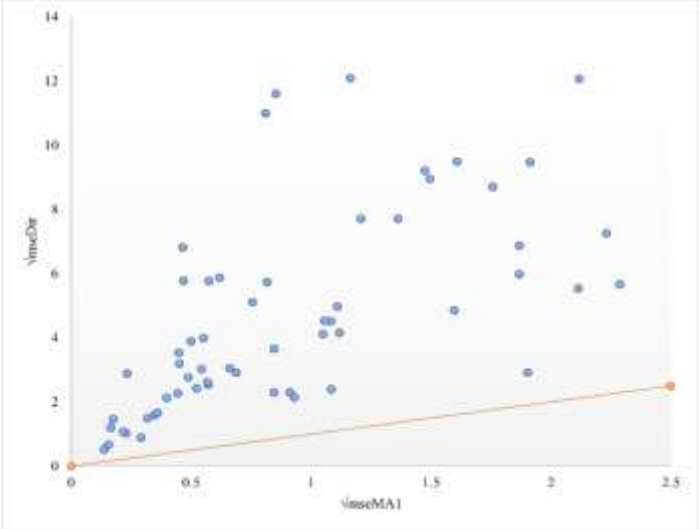

Figure 3. Comparison between $\sqrt{(\mathrm{MSE(MA1)})}$ and $\sqrt{(\mathrm{MSE(Dir)})}$

5. Conclusions

In this work a strategy for the small area estimation of economic security is proposed. To this purpose an indicator obtained by summing absolute differences between levels of equivalent household income in consecutive years, is applied to data taken from EU-SILC sample survey carried out for Italy from 2014 and 2016. The target parameter is the small area average of the individual economic security score. To improve the reliability of direct estimates that can be obtained for small areas, small area models specified at area level are considered. This indicator has been used here for the first time in a small area estimation context, whereas poverty and inequality indicators have usually been considered in the small area literature so far. In addition to the basic Fay-Herriot model, given the nature of the indicator, we proposed to consider some longitudinal extensions of the Fay-Herriot model, specifically two models including temporal random effects. In the first longitudinal model temporal random effects follow an autoregressive process of order 1, in the second one they follow a moving average process of order 1. Thus, we try to improve the reliability of estimates by borrowing information, not only from auxiliary variables available from administrative archives, but also from temporal correlation. The variances of small area direct estimates, to be included in the small area models, are estimated by using a bootstrap methodology, and then smoothing using the GVF method.

The performance of the models proposed is evaluated though a simulation study, based on EU-SILC data. Results obtained from the simulation study highlight that the best performing small area model, both in terms of bias and overall efficiency, is the MA1 model. Moreover, the simulation study provides further evidence of some properties of the estimators: indeed, all models perform better than direct estimator in terms of mean square error. Furthermore, the graphical analysis shows that the small area estimator based on best performing model, MA1 model, tends to be asymptotically design-unbiased and consistent, and that the efficiency gain is relevant for all areas.

However, this work suffers from some limitations. The available auxiliary variables are not highly correlated with the target variable. The variability of the economic security indicator considered in this work makes particularly difficult to find suitable covariates for the small area models. This problem is often encountered in small area application for poverty and inequality indicators. Finally, a further limitation is the low number of waves available. In fact, with a higher number of waves available, the longitudinal models considered may provide more precise estimates for the economic security indicator.

Disclaimer

This paper is based on data from Eurostat, EU Statistics on Income and Living Conditions (2014, 2015, 2016). The responsibility for all conclusions drawn from the data lies entirely with the authors.

References

Bossert, W., & D'Ambrosio, C. (2013). Measuring economic insecurity. *International Economic Review*, *54*(3), 1017-1030.

Bossert, W., & D'Ambrosio, C. (2019). Economic insecurity and variations in resources. Working Papers 422, ECINEQ, *Society for the Study of Economic Inequality*.

Bossert, W., Clark, A. E., D'Ambrosio, C., & Lepinteur, A. (2022). Economic insecurity and political preferences, *Oxford Economic Papers*, 1–26. https://doi.org/10.1093/oep/gpac037

D'Ambrosio, C., & Rohde, N. (2014). The distribution of economic insecurity: Italy and the us over the great recession. *Review of Income and Wealth*, *60*, S33-S52. https://doi.org/10.1111/roiw.12039

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, 613-627.

Dvoryadkina, E., Guseynly, K., & Sobyanin, A. (2021). Economic security of the region's periphery in conditions of digitalization, urbanization, and covid-19, in SHS Web of Conferences, vol. 93. EDP Sciences. https://doi.org/10.1051/shsconf/20219305019

Esteban, M. D., Morales, D., Pérez, A., & Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics & Data Analysis*, *56*(10), 2840-2855. https://doi.org/10.1016/j.csda.2011.10.015

Esteban, M. D., Morales, D., & Pérez, A. (2016). Area-level spatio-temporal small area estimation models, Analysis of Poverty Data by Small Area Estimation, 205-226. https://doi.org/10.1002/9781118814963.ch11

Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data, *Journal of the American Statistical Association*, *74*(366a), 269-277.

https://doi.org/10.1080/01621459.1979.10482505

Hacker, J. S., Huber, G. A., Nichols, A., Rehm, P., Schlesinger, M., Valletta, R., & Craig, S. (2014). The economic security index: A new measure for research and policy analysis, *Review of Income and Wealth, 60*, S5-S32, 2014. https://doi.org/10.1111/roiw.12053

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation, *Test, 15*(1), 1-96. https://doi.org/10.1007/BF02595419

Osberg, L., & Sharpe, A. (2002). An index of economic well-being for selected OECD countries, *Review of Income and Wealth, 48*(3), 291-316. https://doi.org/10.1111/1475-4991.00056

Pfeffermann, D. (2013). New important developments in small area estimation, *Statistical Science, 28*(1), 40-68. https://doi.org/10.1214/12-STS395

Prasad N. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association, 85*(409), 163-171. https://doi.org/10.1080/01621459.1990.10475320

Rao, J. N. K. (2003). *Small area estimation*. John Wiley & Sons. https://doi.org/10.1002/0471722189

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons, Inc., Hoboken. https://doi.org/10.1002/9781118735855

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

Salvati, N., Chandra, H., Ranalli, M. G., & Chambers, R. (2010). Small area estimation using a nonparametric model-based direct estimator, *Computational Statistics & Data Analysis, 54*(9), 2159-2171. https://doi.org/10.1016/j.csda.2010.03.023

Thomson, W. (2001). On the Axiomatic Method and Its Recent Applications to Game Theory and Resource Allocation. *Social Choice and Welfare, 18*, 327-386. https://doi.org/10.1007/s003550100106

Warnholz, S., & Schmid, T. (2016). Simulation tools for small area estimation: Introducing the r-package saesim, *Austrian Journal of Statistics, 45*(1), 55-69. https://doi.org/10.17713/ajs.v45i1.89

Wolter, K. M. (2007). *Introduction to variance estimation*. Springer, vol. 53.