Spatial Disaggregation of Social Indicators: An Info-Metrics Approach

Esteban Fernandez-Vazquez^{1,2} · Alberto Diaz Dapena^{1,3} ·
Fernando Rubiera-Morollon^{1,2} · Ana Viñuela^{1,2}

Accepted: 25 July 2020 / Published online: 9 August 2020
© The Author(s) 2020

Abstract

In this paper we propose a methodology to obtain social indicators at a detailed spatial scale by combining the information contained in census and sample surveys. Similarly to previous proposals, the method proposed here estimates a model at the sample level to later project it to the census scale. The main novelties of the technique presented are that (i) the small-scale mapping produced is perfectly consistent with the aggregates -regional or national- observed in the sample, and (ii) it does not require imposing strong distributional assumptions. The methodology suggested here follows the basics presented on Golan (2018) by adapting a cross-moment constrained Generalized Maximum Entropy (GME) estimator to the spatial disaggregation problem. This procedure is compared with the equivalent methodology of Tarozi and Deaton (2009) by means of numerical experiments, providing a comparatively better performance. Additionally, the practical implementation of the methodology proposed is illustrated by estimating poverty rates for small areas for the region of Andalusia (Spain).

Keywords Small area estimation · Poverty mapping · Spatial disaggregation · Info-metrics and entropy

1 Introduction

The use of social indicators at an appropriately disaggregated geographical scale is a relevant issue, not only for research but also for policy-making because efficient policy design sometimes requires information with a high degree of spatial disaggregation. Official data, however, are not widely produced with a high detail of spatial disaggregation and are generally only available at a national or -sometimes- regional level. If the geographical location of these individual agents is observable at a highly disaggregated spatial scale, the indicators for

✉ Esteban Fernandez-Vazquez
evazquez@uniovi.es

¹ REGIOlab: Regional Economics Laboratory at the University of Oviedo, Oviedo, Spain

² Department of Applied Economics, University of Oviedo, Asturias, Spain

³ Department of Economics and Statistics, University of Leon, León, Spain

the small areas could be calculated simply by summing or averaging the individual estimates. In most cases, however, researchers have to deal with databases that might allow for a precise spatial location of the individuals, such the microdata of a Population Census, but that do not normally contain information on economic variables related to distributional issues, such as household income. On the other hand, the surveys that contain data about these indicators—such as household surveys—do not normally provide detailed information on the geographical location of the individuals surveyed. Even when they do, the small sample sizes for the sub-regions of interest make the estimates directly obtained very unstable (see Marchetti and Secondi 2017, for a recent application of this type of direct estimation).

Small Area Estimation (SAE) deals with the problem of providing reliable estimates of these indicators based on the information collected by conducting surveys in some or all areas. SAE techniques are increasingly used to provide estimates for local areas of interest (for a recent application, see Durán and Condori 2019). Generally speaking, SAE methodologies combine direct estimates with some auxiliary information. The comparative performance of different techniques in this field has been recently evaluated by Guadarrama et al. (2016). Within the group of methodologies that aim at producing social indicators for small areas, there is one in particular that has been widely applied. Specifically, the original proposal by Elbers et al. (2003) and the modification proposed later in Tarozzi and Deaton (2009) has been used by the World Bank to map poverty across small areas in poor countries and it has later been used in several developing countries (see Modrego and Berdegú, 2015) and in other more advanced economies (see Sánchez-Cantalejo et al. 2008 Melo et al. 2016; Morales et al. 2018; for recent examples).

The basic idea of this procedure consists of “projecting” predictions of the variable of interest for a household survey onto the sample of households that form the population. In a nutshell, the procedure comprises three steps:

1. From the household surveys (HS), estimate a model of your variable y of interest $y = (X)$, where X is a set of regressors also observable in the Population Census (PC).
2. Recover the set of parameters β estimated on the HS (with some degree of heterogeneity across regions or clusters of households) and take them to the PC.
3. Given the X observable in the PC and the corresponding $\hat{\beta}$, predict the figures of y for the households surveyed in the PC ($\hat{y} = X\hat{\beta}$).

The estimates produced have the advantage of a higher precision than previous methodologies, due to the large number of households in the census (see Tarozzi and Deaton, 2009). Additionally, this large number of estimates allows for studying potential differences between the individuals or households belonging to the same small area.

This feature is highly appealing to the social researcher, and this is the approach that we follow in this paper. One problem derived from applying this type of estimation, however, is that the estimates are not necessarily consistent with the aggregates that are already observable: for example, the mean value of the estimates of household income produced by the techniques at household level for, say, a given region may not necessarily be equal to the mean regional household income available on the official databases.

In order to overcome this limitation, we propose an alternative methodology to adjust the estimates to official observable aggregates by incorporating into the estimation problem information on the observable aggregates. We propose to make this correction basing on the infometrics framework detailed in Golan (2018) and with a procedure similar to

the Generalized Maximum Entropy (GME) estimator presented in Bernadini-Papalia and Fernández-Vázquez (2018).

The rest of the paper is organized as follows. Section 2 describes the main characteristics of the technique proposed. In Sect. 3 we compare its performance with the original formulation by Tarozzi and Deaton (2009) we illustrate how the proposed technique can be applied to the empirical problem of recovering at risk of poverty an exclusion rates for small areas -municipalities- for a Spanish region. Finally, in Sect. 4 we draw some conclusions.

2 Spatial Disaggregation of Social Indicators Within an Infometrics Framework

The disaggregation methodology that we propose in this paper tries to exploit the information contained in the available databases while making minimal assumptions. The methodology suggested here follows the basics presented in Golan et al. (1996) and Golan (2018).

2.1 The GME Moment-Constrained Estimator

Assume that the estimation objective is to recover the cells of matrix P , with dimension $n \times J$, with a typical element p_{ij} that reflects the probability of category j for individual i , with $p_{ij} \geq 0$ and $\sum_{j=1}^J p_{ij} = 1$; $i = 1, \dots, n$. The basic idea of the infometrics framework consists of choosing as the solution the one that requires the least information (i.e., maximizes the uncertainty or entropy) of the distribution while being consistent with the observed data. Optimizing an entropy (uncertainty) function as in Eq. (1):

$$Ent(P) = - \sum_{i=1}^n \sum_{j=1}^J p_{ij} \ln(p_{ij}) \quad (1)$$

This function reaches its unconstrained maximum when the n rows in P distribute uniformly across the J categories (see Shannon, 1948 for additional details). Any additional information—observed data—is included as a constraint and it makes the solution depart from the uniform distribution.

This basic idea can be adapted to predict and model categorical variables (see Chapter 11; Golan 2018). Consider a sample consisting of n observations of basic units, which for the sake of simplicity will be assumed to be individuals but which can be easily generalized to other types of observations, such as households or firms. Let us assume that the variable of interest is a categorical indicator that reflects whether the individual belongs to a specific class. Each data point observed in the sample belongs to one single category of the J unordered alternatives and each observation is coded as a binary variable, which takes value 1 for the realized category and value 0 for the rest. For each individual $i = 1, \dots, n$, we define a variable y_{ij} that equals 1 if and only if alternative j is observed and zero otherwise. This implies that the information of the indicator of interest contained in the sample can be presented in the form of a $n \times J$ matrix as Y :

$$Y = P + U \quad (2)$$

Equation (2) indicates that the observed data in the sample are specified as a function of the unobserved probability P plus some noise contained in matrix U . Our objective is to

predict the elements of P based on the observed data, assuming that the p'_{ij} s cells in matrix P are related to a set of explanatory covariates contained in the $n \times K$ matrix X that contain K individual's characteristics. In order to accommodate these relations between P and X on a flexible way, Golan (2018) proposes the use of the cross-moments equations:¹

$$\frac{1}{n} X^T Y = \frac{1}{n} X^T [P + U] = \frac{1}{n} X^T [P + WV] \quad (3)$$

that serve as constraints in the following optimization program:

$$\underset{P, W}{\text{Max}} \text{Ent}(P, W) = - \sum_{i=1}^n \sum_{j=1}^J p_{ij} \ln(p_{ij}) - \sum_{l=1}^L \sum_{i=1}^n \sum_{j=1}^J w_{ijl} \ln(w_{ijl}) \quad (4)$$

Subject to:

$$\frac{1}{n} \sum_{i=1}^n x_{ik} y_{ij} = \frac{1}{n} \sum_{i=1}^n x_{ik} [p_{ij} + u_{ij}] = \frac{1}{n} \sum_{i=1}^n x_{ik} [p_{ij} + w_{ijl} v_l]; \quad j = 1, \dots, J \quad (5)$$

$$\sum_{j=1}^J p_{ij} = 1; \quad i = 1, \dots, n$$

$$\sum_{l=1}^L w_{ijl} = 1; \quad i = 1, \dots, n; \quad j = 1, \dots, J \quad (6)$$

Equations (4) to (6) depict a Generalized Maximum Entropy (GME) program, where the term $\text{Ent}(W) = - \sum_{l=1}^L \sum_{i=1}^n \sum_{j=1}^J w_{ijl} \ln(w_{ijl})$ quantifies the entropy associated with the cells u_{ij} . These elements, which reflect the noise for individual i in category j , are expressed as $u_{ij} = \sum_{l=1}^L w_{ijl} v_l$ or $U = WV$ in matrix notation where matrix V contains the set of L possible values for the error terms while matrix W denotes the corresponding (unknown) probabilities. Given the categorical nature of the data on Y , the feasible values for the error on V are intuitively bounded between -1 and 1 ; which for the sake of simplicity are set as common for all the data points.

If the GME program above is solved, estimates of $P(\hat{P})$ and $W(\hat{W})$ are recovered. While the set of constraints on Eq. (6) act just as normalization restrictions, the information contained in the sample in the form of cross moments between Y and X is given in the left hand side of Eq. (5). Without this piece of information, the solution of the GME estimator will be a uniform distribution, but this equation “pushes” the solution to be consistent with these observed moments.²

2.2 Spatial Disaggregation Based on GME

We apply a modification of the general GME estimator above as an alternative to the methods presented in Elbers et al. (2003) or Tarozzi and Deaton (2009) of predicting social

¹ Superscript T indicates transposition.

² Golan (2018) shows that under some mild assumptions the GME solution can be seen as a generalization of the classical Maximum Likelihood estimator for a logit model, yielding the same solution when $U = 0$.

indicators for small areas. The GME technique proposed follows the idea of combining household surveys with population census but exploits the information in a different way. One advantage of the method proposed here is that it combines the detailed geographical information present in population census but making it consistent with some aggregates -moments- observable in the household survey.

For the sake of clarity, let us explain the procedure by assuming that our research interest is in estimating a categorical indicator in a set of small areas $d = 1, \dots, D$. The proportion of individuals in d that fall into category j can be expressed as $\sum_{i=1}^{n_d} y_{ij}^d / n_d$ where n_d represents the number of individuals in this area d . Our problem is that the indicators y_{ij} are not directly observable in a population census. They are observable in the household survey, but we cannot identify the small area d in which the individual lives.

Following Eq. (2), our estimates of the indicator of interest $\hat{y}_{ij}^d$ are defined as $\hat{y}_{ij}^d = \hat{p}_{ij}^d + \hat{u}_{ij}^d$, where $\hat{p}_{ij}^d$ and $\hat{u}_{ij}^d$ are the solutions produced by the GME estimator and $\hat{u}_{ij}^d = \sum_{l=1}^L \hat{w}_{ijl} v_l$. These estimates will be obtained by solving the previously depicted GME program but modifying the constraints on Eq. (5) as follows:

$$\frac{1}{n} \sum_{i=1}^n x_{sik} y_{sij} = \frac{1}{N} \sum_{i=1}^N x_{cik} [p_{ij} + u_{ij}] = \frac{1}{N} \sum_{i=1}^N x_{cik} [p_{ij} + w_{ijl} v_l]; \quad j = 1, \dots, J \quad (7)$$

Or, in matrix notation:

$$\frac{1}{n} X_s^T Y_s = \frac{1}{N} X_c^T [P + U] = \frac{1}{N} X_c^T [P + WV] \quad (8)$$

In Eqs. (7) and (8) the subscripts s and c refer to moments observed in the household survey and the population census respectively. Matrix X_s -together with Y_s - is collected from the information in the sample, and it has dimension $n \times K$, while matrix X_c is observable at the population census level and has dimension $N \times K$, where N is the population size. Note that the left-hand side on these equations refers to the cross-moments between Y and the covariates in X , which are observable for the sample of households surveyed. The solutions $\hat{P}$ and $\hat{W}$ on the right-hand side must produce, at the census level, cross-moments identical to those observed in the household sample. Since these estimates are obtained with the sufficient geographical detail, it is possible to predict $\hat{y}_{ij}^d = \hat{p}_{ij}^d + \hat{u}_{ij}^d$ and make these estimates consistent with the aggregates in $\frac{1}{N} X_c^T Y_s$. This consistency can be viewed as one of the main comparative advantages of this GME technique: while other small-area estimation just make use of the available data for the small areas of interest, the GME estimator also considers the exogenous information that could come in the form of aggregates. If these aggregates contain information about the small areas, this is explicitly exploited and incorporated into the estimation problem, which can contribute to improve the accuracy of the estimates. Additionally, discrepancies between the small area estimates and the regional or national aggregates are prevented by construction.³

³ The solutions of the GME estimator are compared with those obtained by applying the technique proposed in Tarozi and Deaton (2009) by means of numerical simulations on the [Appendix](#).

3 An Empirical Application: Estimating Poverty and Exclusion Rates for Small Areas in Andalusia (Spain), 2011

3.1 Data

This section illustrates how the GME estimator works with a real-world example. It will be applied to predict at-risk of poverty and exclusion (AROPE) rates for small areas (municipalities) in a Spanish region. Following the definition given by Eurostat, for an individual to be classified as AROPE, it is necessary to belong to a household for which one or more of the following three conditions holds:

- i. Having a disposable income below 60% of the national median,
- ii. Suffering severe material deprivation, or,
- iii. Living in a household with a low work intensity.

This empirical example will be based on the data collected for 2011 in the Social Inclusion and Living Conditions (EU-SILC) survey. This survey provides information for the EU member states on issues related to income distribution, poverty or, more in general, living conditions and it is the most commonly used database for distributional issues. The EU-SILC survey is rich on information about several characteristics of individuals and households.⁴ However, due to confidentiality issues, the EU-SILC does not provide detailed information on the geographical location of the individuals surveyed: depending on the specific EU country, information on location is only released at the level of NUTS1 or NUTS 2 regions. In our case, we will focus our analysis on the Spanish region of Andalusia. This NUTS 2 region is interesting for a number of reasons, the most important being that it presented the highest poverty rate in Spain on the year studied here, and one of the highest in the EU. Moreover, the region experienced large unemployment rates and low educational levels (measured as the percentage of population with a third-level degree) compared with other Spanish regions.

Additionally, this region contains large metropolitan areas -the urban agglomeration of Seville is larger than 1.5 million—with a relatively high concentration of manufacturing and service activities, together with many depopulated rural areas mainly specialized in agricultural and other low-value added activities. Consequently, one would expect to find a sizable internal heterogeneity in the poverty indicators across the municipalities of the region. However, this is not possible to study by using only the information contained in the EU-SILC.

In order to overcome the problems generated by the lack of data availability, we will make use of the GME estimator proposed, and the information contained in the sample of households in the EU-SILC will be combined with the sample of households surveyed in Andalusia in the Spanish Population Census conducted on 2011. While the EU-SILC surveyed slightly more than 3,500 adult individuals in the region, the sample in the Population Census contained information on more than 50,000, but not on the variables related to income or deprivation that would allow an individual to be classified as AROPE or not. The Census survey also provides information on the geographical location at the scale of municipality, although only those municipalities larger than 20,000 inhabitants can be

⁴ See <https://ec.europa.eu/eurostat/web/income-and-living-conditions/overview> for details.

Table 1 Mean values. Andalusia 2011

Variable	EU-SILC 2011 (X_s)	Population Census 2011 (X_c)
Poor	37.80%	
Age	49.45	46.69
National	87.91%	91.86%
Rural	35.16%	32.29%
College	21.24%	17.22%
Worker	35.54%	39.78%
Unemployed	20.90%	25.10%
Sample size	$n = 3548$	$N = 52,490$

Note: Data taken from the EU-SILC and the Population Census 2011. All the variables, excepting Age (years) are given in percentage. The people interviewed are 18 or older. Poor is the percentage of adults classified as AROPE. Female is the proportion of women in the sample. National is the percentage of people with the Spanish nationality. Rural is the percentage of people living in a municipality with a low population density. College is the percentage of people with a college degree. Unemployed and Worker are the proportions of individuals looking actively for a job and having a job respectively

properly identified, and the rest are aggregated because of privacy issues. Nevertheless, this spatial scale allows us to identify larger urban settlements and differentiate them from rural areas.

Our estimation objective will be to recover AROPE rates at municipal level by applying the GME program maximizing Eq. (4) subject to the set of normalizing constraints on (6) and the moment constraints on Eq. (7). In our problem, matrix Y_s contains the binary variable of interest in the EUSILC sample where $y_i = 1$ if an individual can be classified as AROPE and zero otherwise; while X_s contains a set of covariates observable in the same EUSILC sample that are also observable at census level, i.e., it is possible to observe the equivalent matrix X_c . For the sake of simplicity in this empirical example, we consider as covariates only one continuous variable ("Age") and its square, plus the binary variables "National", "Rural", "College", "Unemployed" and "Worker". Table 1 presents a summary of these variables in both databases:

Additionally, the sample correlations found between Y_s and X_s are reported in the first column on Table 2:

In the sample surveyed in the EUSILC, the indicator of risk of poverty and exclusion is negatively correlated with the age of the individuals. As expected, situations of AROPE seems to be negatively correlated with the fact of having a job, holding a college degree and having Spanish nationality, while it is positively correlated with the fact of living in a rural environment and being unemployed. Given the observed correlation between the covariates chosen and the indicator of interest, our GME estimator is designed to "project" onto the set of individuals surveyed in the Population Census estimates that are consistent with the cross moments contained in the matrix $\frac{1}{n} X_s^T Y_s$. These cross-moments are also reported on Table 4 (second column).⁵

⁵ Given the nature of the variables considered, these cross-moments can be interpreted as joint probabilities for the binary covariates. For example, they indicate that 13.91% of the individuals sampled in the EUSILC were unemployed and classified as AROPE simultaneously. Regarding the interpretation for the continuous

Table 2 Sample correlation and cross moments. Andalusia 2011, EUSILC sample

	Correlation with AROPE	Cross moments with AROPE ($X_s^T Y_s$)
Age	-0.1226**	17.4879
Age ²	-0.1222**	927.9262
National	-0.0967**	0.3021
Rural	0.0617**	0.1444
College	-0.1958**	0.0474
Worker	-0.2370**	0.0758
Unemployed	0.2887**	0.1391

Note: Data taken from the EU-SILC 2011. Two stars (**) indicate that they are significantly different from zero at 5%

3.2 Results: The Spatial Distribution of Rates of AROPE Population in Andalusia

From these data, we have applied a moment-constrained GME program including as constraints the sample information on $X_s^T Y_s$. In this example, we followed the same procedure for the reparameterization of the error term as applied in the numerical experiments and we set wide error supports with $L = 3$ points as $v_l = (-1, 0, 1)$. The solution of this program predicts the probability of being classified as AROPE with the individuals sampled in the Population Census calculated as $\hat{y}_i^d = \hat{p}_i^d + \hat{u}_i^d$, where the superscript d refers to the small area – a specific municipality or a group of municipalities- where the individual lives. Having recovered the $\hat{y}_i^d$, local estimates of AROPE rates can be calculated as $\sum_{i=1}^{N_d} \hat{y}_i^d / N_d$, where N_d represents the population size on this area d .⁶ These local estimates allows for studying the spatial heterogeneity of the incidence of AROPE across the municipalities of the region. In total, we managed to recover estimates for 115 small areas and Fig. 1 visually represents these estimates:

The estimates produced show sizable heterogeneity across the Andalusian municipalities. Approximately 10% of the small areas identified presented an AROPE rate smaller than 30%, being close to the national AROPE for Spain in that year (26%), while the estimate for the top 10% was higher than 50%. The figures represented in Fig. 1 also allow some geographical patterns to be identified. First, the map shows that the areas located in the South-west of the region generally presented higher values. Additionally, and as the correlation calculated in the EUSILC survey suggested, the greater AROPE rates are mainly concentrated in rural areas and are comparatively smaller in the urban agglomerations. Figure 1 displays the names of the municipalities that concentrate the five largest metropolitan areas of Andalusia, which accounted for approximately 4 million inhabitants and almost 50% of the total population of the region. In general, the estimates of poverty

Footnote 5 (continued)

covariates (“Age” and its square), their cross-moments can be easily transformed into conditioned means by simply dividing by the AROPE rate. For example, the figures in Table 4 indicate that the mean age of the individuals classified as AROPE was $46.26 = 17.4879/0.3780$.

⁶ In this estimation exercise we have made use of the sampling weights reported both in the EUSILC and the Population Census, although we opted for not mentioning this explicitly in the description of the GME estimator for simplicity.

Fig. 1 AROPE rates at municipal level. Andalusia 2011

rates around these areas are comparatively smaller than in other zones of the region. These results are consistent with previous research that found higher persistence of poverty on rural areas for the US (Partridge and Rickman 2008) or for the EU (Weziak-Białowolska 2016).

For the specific case of Spain (Sánchez-Cantalejo et al. 2008) estimated a multidimensional poverty indicators basing on the 2001 Census data. Although the results are not directly comparable (the reference years are different and they use an alternative definition of the poverty indicator) the spatial patterns found in their study for the case of Andalusia seem to be similar to those estimated in this paper. Although their study comprehends all the Spanish regions, for the specific case of Andalusia on 2001 their estimates suggest a similar geographical concentration of poverty for rural areas on Andalusia or, at least, a comparatively smaller incidence on Andalusian urban agglomerations (see Sánchez-Cantalejo et al. 2008, Figs. 2 and 3, page 268).

4 Concluding Remarks

This paper develops and applies an estimation procedure based on the info-metrics approach to estimate social indicators for small areas that are consistent with observable aggregate information. In this regard, the technique proposed here can be seen as a spatial disaggregation technique. More specifically, we propose a modified version of the cross-moment constrained GME estimator for categorical variables developed in Golan (2018), which we adapt to be applied in the field of small area estimation. Although this is not the first attempt to use this type of methodology for small area estimation problems (see Bernadini-Papalia and Fernández-Vázquez 2018), the novelty of our proposal is that we do not use only aggregates of the variable of interest as constraints, but also the information contained in its cross-moments with potential covariates can be exploited.

The formulation presented here is also inspired in the idea of the proposals made originally by Elbers et al. (2003) and the subsequent modification by Tarozzi and Deaton (2009) to map poverty for small areas. Our estimator is evaluated through numerical experiments, revealing a comparatively better performance than the previous proposals which attempt to make a similar analysis. We also illustrate, by means of an empirical illustration, how the proposed procedure can be implanted to solve a real-world problem and we have applied it to derive At Risk of Poverty and Exclusion (AROPE) rates for small geographical areas. More specifically, we estimate these rates for the municipalities of the region of Andalusia (Spain) for 2011, combining microdata from the EUSILC database and the Spanish Population Census produced that year.

The numerical simulations together with the empirical example illustrate how this methodology can be helpful for studying the spatial heterogeneity of social indicators across small areas. However, further work should be done to explore its performance when the variable of interest is not categorical but continuous, or it has the form of count data. Future research should also explicitly address potential processes of spatial dependence between the geographical locations of interest.

One innovation of the technique presented here it is that it does not require strong distributional assumptions and that it combines information from two different sources of information, while maintaining the consistency of the estimates produced with national and regional aggregates. One potential limitation associated to this characteristic is that the proposed GME technique has comparatively larger requirements of information if compared to other methodologies, since it bases on combining two different sources of data. Alternatives based on just exploiting observable information -as, for example, the Principal Components approach applied in (Sánchez-Cantalejo et al. 2008)– are less demanding on information, but they do not necessarily produce small area estimates consistent with the national or regional figures that are usually estimated by official statistical agencies.

Acknowledgement This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 726950.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix: Numerical Experiments

This appendix conducts a Monte Carlo experiment that evaluates the GME procedure proposed. For the sake of simplicity, and also in order to facilitate comparison with the original formulation in Tarozzi and Deaton (2009), we replicate the general characteristics of the Monte Carlo simulations conducted in that paper (see Tarozzi and Deaton 2009, pages 781–784).

The idea of this simulation is to generate “true” but not completely observable data for a set of households and small spatial areas. The population simulated consists of $N = 15,000$ households distributed uniformly across 150 small (local) areas. The values of a continuous variable (z) are generated as:

$$z_i^d = 20 + x_i^d + e^d + \varepsilon_i^d \quad (9)$$

For simplicity, we assume that our categorical indicator of interest y_{ij} is just a binary variable defined as:⁷

$$y_i = 1 (z_i < z^*) \quad (10)$$

where z^* is a given threshold and zero otherwise. In their paper, Tarozzi and Deaton (2009) use this data generation process to simulate poverty rates for the local areas as the simple head-count indicators where a household is classified as “poor” if the household income is below a given poverty line z^* .

In Eq. (9), subscript d stands for the local areas and i for the households, e^d is an area-specific shifter that distributes as $e^d \sim N(0, 0.01)$ and ε_i^d is a disturbance that distributes normally with mean zero and variance $\sigma^2(x)$, where:

$$\sigma^2(x) = \frac{\exp \left[0.5x_i^d - 0.01(x_i^d)^2 \right]}{1 + \exp \left[0.5x_i^d - 0.01(x_i^d)^2 \right]} \quad (11)$$

The predictor x_i^d is generated as $x_i^d = 5 + r^d h_i^d + g^d$, where $r^d, g^d \sim U(0, 1)$ and $r^d \perp g^d$ and $h_i^d \sim N(0, 1)$.

We assume that we sample $n^d = 10$ household on every local area, which makes a total sample size of $n = 1,500$. For this sample, we replicate Tarozzi and Deaton (2009) (TD) by estimating probit models where the dependent variable is y_i and taking as matrix of covariates X the values of variable x , its square, $\sin(x)$, $\cos(x)$, $\sin(2x)$ and $\cos(2x)$.⁸ Once these equations are estimated, the parameter estimates are projected onto the whole population of $N = 15,000$ households to predict the values of z and y . Note that this population plays the role of the census, where the geographical location of the household (each local area d) is observable. In the numerical experiment conducted they use three different poverty lines by setting $z^* = 24, 25, 26$. Once the predictions of $\hat{y}_i^d$ for the whole population are calculated, this allows us to produce estimates of some average for the indicator on each d .

In order to evaluate their method, Tarozzi and Deaton (2009) compare the true values of y_i^d and the predictions obtained by their TD estimator, finding that it produced smaller bias and Square Mean Errors (SME) in comparative terms to Elbers et al. (2003). Following this idea, we have extended this numerical experiment by including the proposed GME procedure in the comparison. In the Monte Carlo experiment conducted, we apply the GME moment-constrained estimator including equations such as (8) as constraints on each simulation drawn. Note that we take the same $n \times 7$ matrix X_s as the TD estimator to calculate the cross-moments $\frac{1}{n} X_s^T Y_s$ in the household survey that act as restrictions. Our GME estimates at the census level are $\frac{1}{N} X_c^T [\hat{P} + \hat{U}]$, where matrix X_c has dimension $N \times 7$. Regarding the specification of the support vectors for the error term contained in matrix V , we opted for wide error supports centered on zero with $L = 3$ points as $v_l = (-1, 0, 1)$.

As in Tarozzi and Deaton (2009), we take 250 replications in the experiment, and a summary of the results that report the bias and the root of the MSE is presented in Table 3:

⁷ Since the indicator of interest is binary, we simplify the notation and we do not make use of the sub index j in this and the following section.

⁸ Note that this implies the estimation of seven parameters because an intercept is considered as well.

Table 3 Simulation results. 250 simulation draws. Normal incomes

	Poverty line		
<i>Tarozzi & Deaton</i>	24	25	26
Bias	−0.002	−0.004	−0.002
RMSE	0.006	0.0012	0.0012
<i>GME</i>	24	25	26
Bias	−0.003	−0.005	−0.002
RMSE	0.009	0.0013	0.0012

Note: Bias is defined as the mean difference along the 250 replications of the true and estimated values of the variable of interest. RMSE stands for the square root of the mean square error along the 250 replications. See Tarozzi and Deaton (2009), page 782, for further details

Table 4 Simulation results. 250 simulation draws. Log-normal incomes

	Poverty line		
<i>Tarozzi & Deaton</i>	24	25	26
Bias	−0.148	−0.187	−0.210
RMSE	0.149	0.188	0.212
<i>GME</i>	24	25	26
Bias	0.002	0.003	0.003
RMSE	0.009	0.009	0.009

Note. Bias is defined as the mean difference along the 250 replications of the true and estimated values of the variable of interest. RMSE stands for the square root of the mean square error along the 250 replications. See Tarozzi and Deaton, page 782, for further details (2009)

The results of the numerical simulation suggest that the GME strategy produces equivalent results to those obtained by the TD estimator, both in terms of bias and RMSE. However, the results of the experiment are to some extent conditioned by the idea of normality present on the probit estimation, which is on line with the conditions of the data generation process since both e^d and ϵ_i^d (and x_i^d as well, but only partially) are simulated to distribute normally. We have modified the conditions of the data generating process by generating $\ln(z_i^d) \sim N(\mu_z, \sigma_z)$ where μ_z and σ_z are respectively the empirical mean and standard deviation along the 250 simulation draws of the variable z generated before. With this modification we aim at generating values of the household income following a log-normal distribution, as empirical datasets usually suggest.

Once this modification has been introduced in the data generating process, we repeated the exercise and compare the performance of both TD and GME estimators. Results are reported in Table 4.

We can see how the deviations both in terms of bias and RMSE are now considerably smaller in the case of the GME estimator, irrespective of the specification of the poverty line. This is a consequence of the greater flexibility of the GME procedure, which only requires minimal assumptions on the distribution of the error terms u_i . In the specific case of our experiments, we only impose a symmetric distribution around

zero with bounds at -1 and 1 . Additionally, the GME estimator exploits the information contained in the household surveys and “pushes” the estimates at the census level to be consistent with the aggregates (i.e., the cross moments $\frac{1}{n}X_s^T Y_s$) at the sample level. If the information contained on these aggregates in the sample are reliable estimates of the equivalent aggregates in the population census (i.e., the unobservable $\frac{1}{N}X_c^T Y_c$), this information helps the estimator to produce more accurate estimates.

References

- Bernadini-Papalia, R., & Fernández-Vázquez, E. (2018). Information theoretic methods in small domain estimation. *Econometric Reviews*, 37(4), 347–359.
- Durán, R. J., & Condori, M. Á. (2019). Deprivation index for small areas based on census data in Argentina. *Social Indicators Research*, 141(1), 331–363. <https://doi.org/10.1007/s11205-017-1827-6>.
- Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), 355–364.
- European Union Statistics on Income and Living Conditions. (2011). *Eurostat*, 1. <https://doi.org/10.2907/EUSILC2004-2018V.1>.
- Golan, A., Judge, G., & Miller, D. (1996). *A maximum entropy econometrics: Robust estimation with limited data*. New York: Wiley.
- Golan, A. (2018). *Foundations of info-metrics. modeling, inference, and imperfect information*. Oxford: Oxford University Press.
- Guadarrama, M., Molina, I., & Rao, J. N. K. (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition New series and Survey Methodology*, 17(1), 41–66.
- Marchetti, S., & Secondi, L. (2017). Estimates of household consumption expenditure at provincial level in Italy by using small area estimation methods: “Real” comparisons using purchasing power parities. *Social Indicators Research*, 131(1), 215–234. <https://doi.org/10.1007/s11205-016-1230-8>.
- Melo, P. C., Copus, A., & Coombes, M. (2016). Modelling small area at-risk-of-poverty rates for the UK using the World Bank methodology and the EU-SILC. *Applied Spatial Analysis and Policy*, 9(1), 97–117. <https://doi.org/10.1007/s12061-014-9132-0>.
- Modrego, F., & Berdegú, J. A. (2015). A Large-scale mapping of territorial development dynamics in Latin America. *World Development*, 73, 11–31. <https://doi.org/10.1016/j.worlddev.2014.12.015>.
- Morales, D., del Rueda, M., M., & Esteban, D. (2018). Model-assisted estimation of small area poverty measures: an application within the Valencia region in Spain. *Social Indicators Research*, 138(3), 873–900. <https://doi.org/10.1007/s11205-017-1678-1>.
- Partridge, M. D., & Rickman, D. S. (2008). Distance from urban agglomeration economies and rural poverty. *Journal of Regional Science*, 48(2), 285–310. <https://doi.org/10.1111/j.1467-9787.2008.00552.x>.
- Population and Housing Census. (2011). *INE*. <https://www.ine.es>.
- Sánchez-Cantalejo, C., Ocana-Riola, R., & Fernández-Ajuria, A. (2008). Deprivation index for small areas in Spain. *Social Indicators Research*, 89(2), 259–273. <https://doi.org/10.1007/s11205-007-9114-6>.
- Shannon, C. E. (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27, 379–423.
- Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The Review of Economics and Statistics*, 91(4), 773–792. <https://doi.org/10.1162/rest.91.4.773>.
- Weziak-Białowolska, D. (2016). Spatial variation in EU poverty with respect to health, education and living standards. *Social Indicators Research*, 125(2), 451–479. <https://doi.org/10.1007/s11205-014-0848-7>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.