**Boletín de Estadística e Investigación Operativa**
Vol. 25, No. 3, Octubre 2009, pp. 218-225

**Estadística Oficial**

# Small area estimation of poverty indicators

**Isabel Molina**

Departamento de Estadística
Universidad Carlos III de Madrid
imolina@est-econ.uc3m.es

**Domingo Morales**

Centro de investigación Operativa
Universidad Miguel Hernández de Elche
✉ d.morales@umh.es

**Abstract**

This paper describes some of the research lines of the European project SAMPLE and presents the results of an application to the estimation of poverty indicators in Spanish provinces by the use of a Fay-Herriot model.

**Keywords:** EBLUP, Fay Herriot model, poverty indicators, small area estimation, survey on income and living conditions.
**AMS Subject classifications:** 62D05, 62J05.

## 1. Introduction

The European project SAMPLE (Small Area Methods for Poverty and Living Condition Estimates) is a research project funded by the European Commission under the Seventh Framework Programme (FP7), which started on March 2008 and finishes on March 2011. Nine partners from four different European countries including Spain are participating in the project. The Spanish partners are Universidad Carlos III de Madrid and Universidad Miguel Hernández de Elche. The aim of this project is to identify and develop new indicators and models for inequality and poverty with attention to social exclusion and deprivation, as well as to develop and implement models, measures and procedures for small area estimation of traditional and new indicators. This goal is achieved using data from the European surveys on income and living conditions with the help of local administrative databases.

This paper introduces some of the poverty indicators studied in the SAMPLE project, proposes some estimators of these indicators based on a common small area model and describes an application using data from the Spanish Survey on Income and Living Conditions (SSILC) of year 2006.

## 2. FGT family of poverty measures in small domains

Consider a finite population $P$ of size $N$ partitioned into $D$ subpopulations $P_d$, $d = 1, \ldots, D$, called hereafter domains or areas, of sizes $N_d$, $d = 1, \ldots, D$. Let $z_{dj}$ be a quantitative welfare variable, such as income, measured for individual $j$ within domain $d$, and let $z$ be a given poverty line, that is, a fixed value such that individuals with $z_{dj} < z$ are considered as under poverty. The FGT family of poverty measures, introduced by Foster, Greer and Thorbecke (1984), is

$$\bar{Y}_{\alpha;d} = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{\alpha;dj}, \quad \text{where } y_{\alpha;dj} = \left( \frac{z - z_{dj}}{z} \right)^{\alpha} I(z_{dj} < z), \tag{2.1}$$

where $I(z_{dj} < z) = 1$ if $z_{dj} < z$ and $I(z_{dj} < z) = 0$ otherwise. Observe that $\bar{Y}_{0;d}$ is the proportion of units under poverty in domain $d$. This quantity is called poverty incidence and measures the quantity of people under poverty. On the other hand, $\bar{Y}_{1;d}$ is the domain mean of relative distances to the poverty line for the people under poverty, and is called poverty gap. This quantity measures the degree of poverty of the people that are under poverty.

## 3. Direct estimators of FGT poverty measures

Let $S \subset P$ be a sample drawn from the population and let $S_d = S \cap P_d$ be the sample from domain $d$. A direct estimator of a domain parameter is an estimator that uses only the sample data from the corresponding domain. These estimators are usually based on the sampling design. A direct estimator of the total $Y_{\alpha;d} = \sum_{j=1}^{N_d} y_{\alpha;dj}$ is given by

$$\hat{Y}_{\alpha;d}^{dir} = \sum_{j \in S_d} w_{dj}\, y_{\alpha;dj}.$$

where the $w_{dj}$'s are the sampling weights attached to the units (in the case of the SSILC, they are the official calibrated sampling weights which take into account for non response). In the particular case $y_{\alpha;dj} = 1$, for all $j \in P_d$, the result is the estimated domain size

$$\hat{N}_d^{dir} = \sum_{j \in S_d} w_{dj}.$$

Using this quantity, a direct estimator of the domain mean $\bar{Y}_{\alpha;d}$ is given by

$$\bar{y}_{\alpha;d} = \hat{Y}_{\alpha;d}^{dir} / \hat{N}_d^{dir}.$$

Using the results of Särndal et al. (1992), pp. 43, 185 and 391, with the approximations $w_{dj} = 1/\pi_{dj}$ and $\pi_{di,dj} = \pi_{di}\pi_{dj}$, $i \neq j$ for the second order inclusion

probabilities, a simple estimator of the design variance of $\bar{y}_{\alpha;d}$ is given by

$$\hat{V}_\pi(\bar{y}_{\alpha;d}) = \frac{1}{\hat{N}_d^2} \sum_{j \in S_d} w_{dj}(w_{dj}-1)\left(y_{\alpha;dj} - \bar{y}_{\alpha;d}\right)^2. \tag{3.1}$$

Direct estimators of poverty measures for Spanish provinces do not have enough precision due to the limited province sample sizes of the SSILC. For this reason, official SSILC estimates of poverty indicators at the province level are not produced. The aim of this paper is to derive indirect estimators which "borrow strength" from related provinces, based on a model linking all provinces called the Fay-Herriot model. Thus, the domains of interest in the application described in this paper are the 50 Spanish provinces along with the autonomous cities Ceuta and Melilla ($D = 52$).

## 4. Indirect estimators of FGT poverty measures based on a Fay-Herriot model

Consider a given $\alpha$ in the FGT family of poverty measures. The Fay-Herriot model (Fay & Herriot, 1979) assumes that the true means $\bar{Y}_{\alpha;d}$ are linearly related to the values of $p$ auxiliary variables measured at the area level through a linear regression model,

$$\bar{Y}_{\alpha;d} = \mathbf{x}_d\boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D, \tag{4.1}$$

where $\mathbf{x}_d$ is a vector of size $1 \times p$ containing the (fixed) values of the auxiliary variables measured are the area level, $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients and the $u_d$'s are independent with zero mean and unknown constant variance $\sigma_u^2$. This model is called linking model because it links the target parameters for all domains $\bar{Y}_{\alpha;d}$, $d = 1, \ldots, D$. Moreover, it assumes that the direct estimator of the true means, $\bar{y}_{\alpha;d}$, are design-unbiased, with

$$\bar{y}_{\alpha;d} = \bar{Y}_{\alpha;d} + e_d, \quad d = 1, \ldots, D. \tag{4.2}$$

where, the sampling errors $e_d|\bar{Y}_{\alpha;d}$, $d = 1, \ldots, D$, are independent with zero mean and heteroscedastic known variances $\sigma_d^2$, $d = 1, \ldots, D$. Model (4.2) is called sampling model. Models (4.1) and (4.2) can be put together in the form of a linear mixed model as

$$\bar{y}_{\alpha;d} = \mathbf{x}_d\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D,$$

Note that here, the direct estimators of the domain means are used as model responses. Moreover, as known error variances in the model, we can take the estimated variances of direct estimators, that is, $\sigma_d^2 = \hat{V}_\pi(\bar{y}_{\alpha;d})$, $d = 1, \ldots, D$. As auxiliary variables, we considered an intercept and the following domain

characteristics: the proportion of individuals with Spanish nationality, the proportions of individuals with ages in the intervals 16-24, 25-49, 50-64 and $\geq 65$, the proportions of individuals with an education level under primary education, no more than primary education, no more than secondary education and university level or more, and finally the proportions of individuals that are employed, unemployed and inactive.

Following the standards of the Spanish National Statistical Office, the poverty line was fixed as 60% of the median of the normalized annual net income for all Spanish households. The aim of normalizing the household income is to adjust for the varying size and composition of households. The EUROSTAT definition of the total number of normalized household members gives a weight 1.0 to the first adult in the household, 0.5 to the second and each subsequent person aged 14 and over and 0.3 to each child aged under 14 in the household. The *normalized size* of a household is the sum of the weights assigned to each person. The total number of normalized household members is then calculated as

$$H_{di} = 1 + 0.5(H_{di\geq 14} - 1) + 0.3H_{di<14}$$

where $H_{di\geq 14}$ is the number of people aged 14 and over and $H_{di<14}$ is the number of children aged under 14. The normalized net annual income of a household is obtained by dividing the net annual income by the normalized size. The Spanish poverty line in 2006 was $z = 6556.60$ euros. This is the $z$-value used in the calculation of direct estimators of poverty incidences and gaps.

When $\sigma_u^2$ is known, the best linear unbiased predictor (BLUP) of the true mean $\bar{Y}_{\alpha;d} = \mathbf{x}_d\boldsymbol{\beta} + u_d$ is given by

$$\tilde{\bar{Y}}_{\alpha;d}^{blup} = \gamma_d\, \bar{y}_{\alpha;d} + (1 - \gamma_d)\, \mathbf{x}_d\tilde{\boldsymbol{\beta}} \tag{4.3}$$

where $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_d^2)$ is the proportion of variance due to $u_d$ (accounting for between area variation) and $\tilde{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$, given by

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d \mathbf{x}_d'\mathbf{x}_d\right)^{-1} \sum_{d=1}^{D} \gamma_d \mathbf{x}_d'\bar{y}_{\alpha;d}.$$

Observe that the BLUP in (4.3) approaches the direct estimator of the domain when the variance $\sigma_u^2$ is large relative to the total variance, and it is close to the synthetic regression estimator when this variance is small. Moreover, it is easy to see that the mean square error of $\tilde{\bar{Y}}_{\alpha;d}^{blup}$ fulfills the inequality

$$MSE(\tilde{\bar{Y}}_{\alpha;d}^{blup}) = \gamma_d\, \sigma_d^2 \leq \sigma_d^2 =: g_{1d}(\sigma_u^2).$$

This means that the BLUP based on the Fay-Herriot model is always at least as efficient as the direct estimator, gaining more efficiency for the areas with larger

sampling variances $\sigma_d^2$.

When $\sigma_u^2$ in unknown, it can be estimated by different methods, such as maximum likelihood (ML) or restricted maximum likelihood (REML). The latter method accounts for the degrees of freedom due to the estimation of $\boldsymbol{\beta}$ and, for this reason, it gives a less biased estimator of $\sigma_u^2$. Replacing $\sigma_u^2$ by an estimator $\hat{\sigma}_u^2$ in (4.3), we obtain the empirical BLUP (EBLUP) of the domain mean $\bar{Y}_{\alpha;d}$. Prasad and Rao (1990) obtained the following asymptotic approximation of the MSE of the EBLUP,

$$MSE(\hat{\bar{Y}}_{\alpha;d}^{eblup}) \approx g_{1d}(\sigma_u^2) + g_{2d}(\sigma_u^2) + g_{3d}(\sigma_u^2),$$

where

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

Here $var(\hat{\sigma}_u^2)$ is the variance of the estimator of $\sigma_u^2$. When $\hat{\sigma}_u^2$ is the REML estimator of $\sigma_u^2$, $var(\hat{\sigma}_u^2)$ can be approximated by the inverse of the Fisher amount of information obtained from the restricted likelihood. In this case, an approximately unbiased estimator of the MSE is given by

$$mse(\hat{\bar{Y}}_{\alpha;d}^{eblup}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2\, g_{3d}(\hat{\sigma}_u^2).$$

## 5. Application results

Figures 1 and 2 show respectively cartograms of the estimated poverty incidence and poverty gap, $\hat{\bar{Y}}_{0;d}^{eblup}$ and $\hat{\bar{Y}}_{1;d}^{eblup}$, in percentage. In these cartograms, the Spanish provinces are colored according to a classification in 5 intervals for the values of the estimates. Observe that the Spanish provinces with smaller poverty incidences are mainly in the north-east of Spain along with Madrid and Guadalajara in the center, while the larger values are for those provinces in the south-west of Spain. The provinces with a critical level of poverty incidence (over 30%) are Cuenca, Ciudad Real, Badajoz, Jaén and Cádiz.

Concerning the poverty gap, which measures the degree of poverty for people under the poverty line, the conclusions obtained from the map are similar. Observe that Badajoz is the only province in which the people under poverty have a mean income of over 12.5% under the poverty line. It is followed by Zamora and then by Almería, Granada, Jaén and Cádiz.

Finally, Figures 3 and 4 plot the estimated coefficients of variation (CVs) of the EBLUPs and the direct estimators (sorted by the values of the CV of direct estimators). Observe that there is an overall clear gain of precision when using the EBLUPs based on the Fay-Herriot model instead of direct estimators.

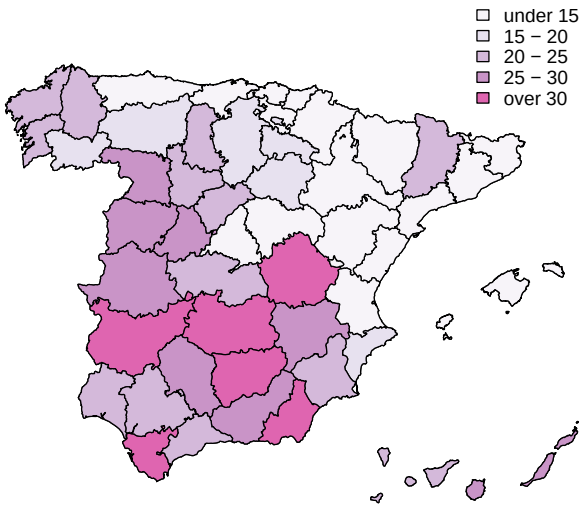


Figure 1: Estimated poverty incidences in Spanish provinces.

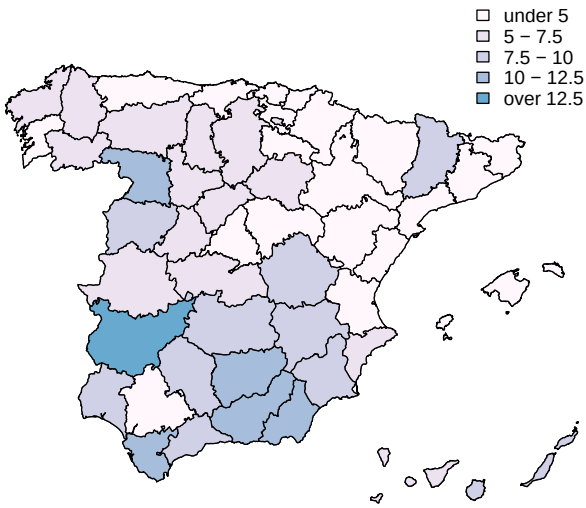


Figure 2: Estimated poverty gaps in Spanish provinces.

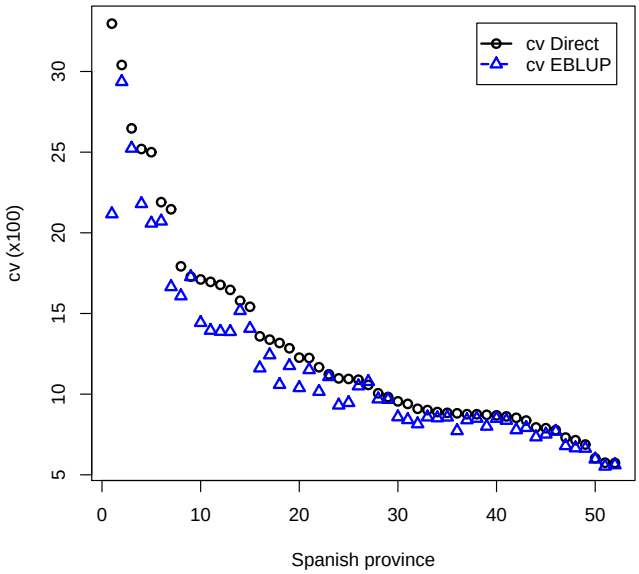


Figure 3: CVs of direct and EBLUP estimators of poverty incidence in Spanish provinces, sorted by increasing order of CVs of direct estimators.

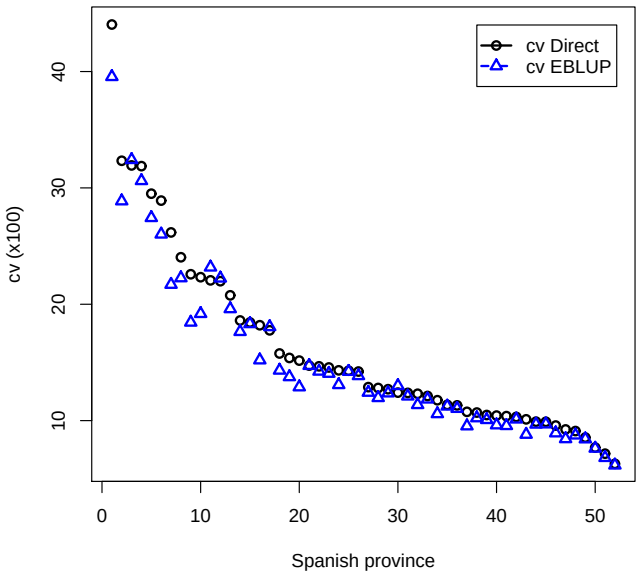


Figure 4: CVs of direct and EBLUP estimators of poverty gap in Spanish provinces, sorted by increasing order of CVs of direct estimators.

## References

[1] Fay, R.E. and 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**, 269-277.

[2] Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica*, **52**, 761–766.

[3] Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

[4] Särndal, C.E., Swensson, B. and Wretman J. (1992) *Model assisted survey sampling*, Springer-Verlag.

### About the authors

**Isabel Molina** is assistant professor in the Departamento de Estadística of Universidad Carlos III of Madrid. Her main research interests are small area estimation and mixed models.

**Domingo Morales** is full professor in the Departamento de Estadística, Matemática Aplicada e Informática and senior researcher in the Instituto Universitario Centro de Investigación Operativa. His main research interests are small area estimation and statistical information theory.