A Fay–Herriot Model for Estimating the Proportion of Households in Poverty in Brazilian Municipalities

Quintaes, Viviane, Hansen, Nícia, Silva, Denise e Pessoa, Djalma
Brazilian Institute of Geography and Statistics - IBGE, Coordination of Methods and Quality - COMEQ
Avenida República do Chile , 500, 10º floor
Rio de Janeiro (20031-170), Brazil
E-mail: viviane.quintaes@ibge.gov.br, nicia.brendolin@ibge.gov.br, denise.silva@ibge.gov.br

Silva, Pedro
Brazilian Institute of Geography and Statistics - IBGE, National School of Statistical Sciences - ENCE
Rua André Cavalcanti, 106
Rio de Janeiro (20231-050), Brazil
E-mail: pedro-luis.silva@ibge.gov.br

Introduction

The National Statistical Institutes face the challenge of producing comprehensive, accurate and reliable information under financial and time constraints. The pressure for reducing sample sizes and respondent burden prompted the need for the use of methods to produce small area statistics from combined data sources. Small area estimation covers a variety of methods used to produce survey-based estimates for geographical areas or domains of study in which the sample sizes are too small to provide reliable direct estimates. Thus, small areas or domains are those for which the sample is very small, common in studies that are designed to produce accurate estimates at national, regional or state level.

The estimation methodology implemented in this paper is based on the use of a regression model with random effects that relates the sample estimates to auxiliary data obtained from census and administrative records available for the small areas (or domains) of interest. The resulting predicted values provide estimates, called model-based estimates, for the small areas based on the model.

This work is part of a project aiming at developing small area estimation models and procedures to allow publication of statistical outputs for areas or domains that are not currently considered for publication due to the low precision of the estimates (or lack of sample observations in some of these areas or domains). IBGE has already published a first set of poverty estimates based on the so-called World Bank Method (Elbers, Lanjouw, and Lanjouw , 2002). The focus of this work is to test some well-known small area estimation models that have been used successfully by other National Statistical Institutes (such as SAIPE1 from US Bureau of Census) and compare results with those obtained under the World Bank Method.

[1] Small Area Income and Poverty Estimates (http://www.census.gov//did/www/saipe/).

The Fay-Herriot Model

The model proposed by Fay-Herriot (FH) is a linear model with random area effects and it was developed to predict per capita income for areas in the United States with population of fewer than 1000 persons. This model is useful in cases where the auxiliary data are available at the area level or when is not possible to link the information of the sample units with census data and administrative records. It links the parameter of interest $\overline{Y}_d$ in the area d to the auxiliary variables $\boldsymbol{x}_d = [x_{d1} \;\; \dots \;\; x_{dp}]^t$ through the linear model: $\theta_d = \boldsymbol{x}^t_d \boldsymbol{\beta} + u_d$ with $d = 1, \cdots, D$ where $\theta_d = g(\overline{Y}_d)$ and $u_d \overset{i.i.d.}{\sim} N(0, \sigma^2_d)$. The random effect u_d aims at capturing the variation between areas not explained by the auxiliary information. Inferences about the small areas are made based on assumption that the population model applies to the $q < D$ areas. In addition to this, the model $\hat{\theta}_d = \theta_d + e_d$ with $d = 1, \cdots, q$ and $\hat{\theta}_d = g(\hat{\overline{Y}}_d)$ indicates how the sample estimates $\hat{\overline{Y}}_d$ are related to the unknown population value and sampling error $e_d \overset{indep.}{\sim} N(0, \sigma^2_D)$. Combining sampling and population models we obtain $\hat{\theta}_d = \boldsymbol{x}^t_d \boldsymbol{\beta} + u_d + e_d$ with $d = 1, \cdots, q$.

Model estimation is carried out using data from sampled areas and the model-based estimates for the small areas are given by $\tilde{\theta}_d = \boldsymbol{x}^t_d \hat{\boldsymbol{\beta}} + \hat{u}_d$ or by the synthetic estimator $\tilde{\tilde{\theta}}_d = \boldsymbol{x}^t_d \hat{\boldsymbol{\beta}}$ when there is no sample in a specific area. More information about the FH estimator and its corresponding mean square error can be found in Rao (2003).

A Model for Estimating the Proportion of Households in Poverty

The survey data used in this study comes from the Brazilian Family Expenditure Survey - Pesquisa de Orçamentos Familiares - POF 2002-2003 (POF, IBGE, 2004) and the auxiliary information was obtained from: the 2000 Demographic Census , an inquiry that collects administrative data from Brazilian municipalities (Pesquisa de Informações Básicas Municipais - 2002 (MUNIC, IBGE, 2005), the School Census - Censo Escolar 2003 (INEP, 2004) and from a system of municipal development indicators maintained by an association of Rio de Janeiro enterprises - Sistema Firjan – IFDM 2000 (2000). The database contains estimates of the proportion of households in poverty for the municipalities of Minas Gerais (MG) state calculated based on the household per capita income. The sample sizes are small (from 6 to 75 households2) and 75% of the sampled municipalities have less than 21 households in the sample. Besides, the survey was not designed to provide accurate estimates for municipalities and in only 151 from the 853 municipalities of MG there are households in the sample.

The auxiliary variables are available for all municipalities in the state of interest and the model fitted with survey data allows the estimation of a poverty measure for those municipalities that do not have any sampled household, assuming that the statistical model that links the estimates with the auxiliary information is also valid in this case. Although, in general, the small area model-based estimates are more precise than the direct sample estimates, they are subjected to bias. The

[2] Belo Horizonte (Minas Gerais capital city) has a sample size of 336 households.

selection of an appropriate model is vital in the estimation process. In this paper, the response is the estimated proportion of households in poverty calculated based on the household income. A household is considered in poverty when its per capita income is below a poverty line3.

Results and Discussion

Table 1 presents the selected variables included in the model and the estimates of $\hat{\boldsymbol{\beta}}$ obtained by the FH estimator. The model selection procedure was carried out fitting a single level linear model (without area random effects (because the database only contained records at municipal level). For the final single area level, an R^2 = 0,47 was obtained. The FH estimator, however, incorporates the random area effect (municipality).

Table 1: Estimates of the fitted model

Covariates	**Coefficient Estimates FH model**
Intercept	**-0,035**
FIRJAN index: employment and income	**-0,215**
2000 Demographic Census 2000 Covariates	
Log of Proportion of Households with no Bathroom	**0,934**
Log of Proportion of Households Served by Urban Sanitation	**0,407**
Log of Proportion of Households with income between 5 and 10 Brazilian Minimum Wages[4]	**-0,990**
Failure Rate in Junior School5	**0,007**

Figure 1 displays the scatterplot of the direct survey estimates against the model-based estimates obtained by FH method for the 151 municipalities of Minas Gerais in which there are some sampled households. The graphic that shows the line $y = x$ (in red) constitutes a bias diagnostic based on the assumption that the direct estimates of the small areas values are unbiased. Although the direct estimates are very variable, they are nevertheless Unbiased, thus a plot of direct estimates (on the y axis) and model-based estimates (on the x axis) should display direct estimates randomly scattered about the model estimates. From Figure 1, it can be noted that the direct estimates are randomly distributed around the model-based FH estimates and they are close to the line in red indicating no evidence of bias.

[3] The development of an absolute poverty line followed the steps: (1) conversion the amount of food acquired by households in terms of energy (kcal) using tables of food composition, (2) estimation of per capita household energy consumption, (3) calculation of nutritional requirements following the FAO / WHO / UN recommendations , (4) definition of the reference group, (5) calculation of the food poverty line, (6) calculation of the poverty line.

[4] Minimum Wage Value in 2011: US$ 330.

[5] For 6-14 years old children.

Figure 1: Estimates of proportion of households in poverty by municipalities of Minas Gerais (151 municipalities that have sample in the Family Expenditure Survey 2002/2003).

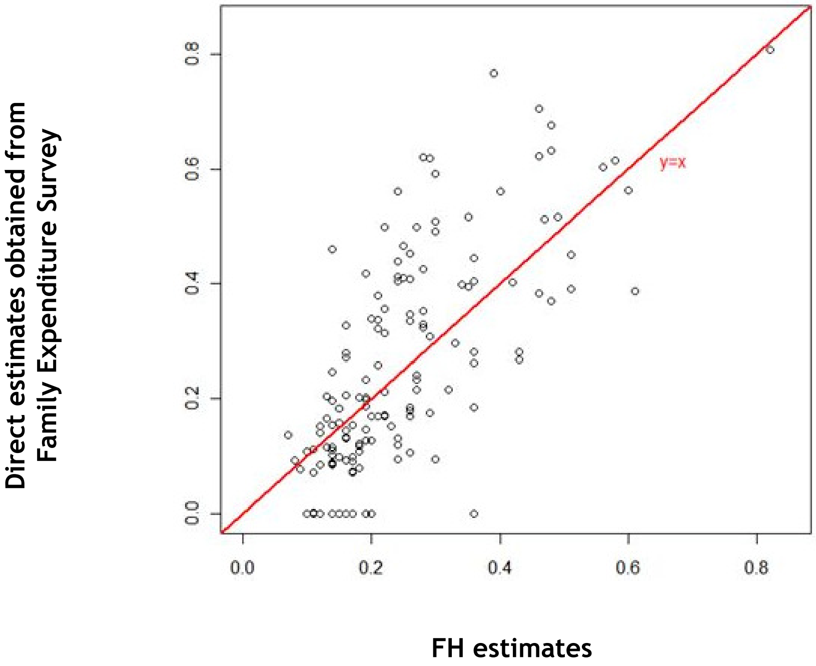
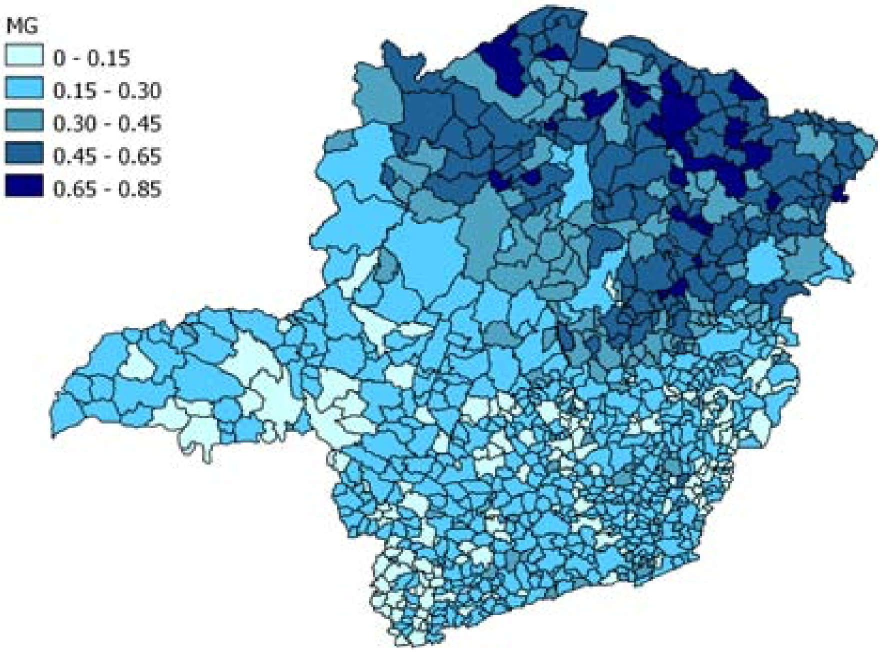


Figure 2 shows the synthetic estimates obtained by FH method for the proportion of households in poverty for the municipalities of MG. Vale do Jequitinhonha, a well known region with high incidence of poverty in the north of the state is among the darkest areas of the map. This region is known to have low socioeconomic indexes.

Figure 2: Model–based estimates for the proportion of households in poverty for Municipalities in Minas Gerais – Brazil. Estimates obtained by FH method (synthetic estimator).



Due to the lack of sample in many municipalities, it is noteworthy that synthetic estimator was used in 702 of them consequently affecting the quality of the estimates. It would, therefore, be interesting to consider the use of different geographical boundaries to define the areas of interest in a future work. In addition, these results are being compared to those obtained based on the World Bank methodology and confidence intervals for the estimates will be produced.

Concluding Remarks

The results of this study demonstrate the feasibility of producing small area estimates based on sample surveys of IBGE. However, there is the need for a careful evaluation of the definition of areas of interest, since the lack of sample in several municipalities imposes the use of synthetic estimators. It is important to mention that the process of choosing the auxiliary variables is crucial, since the small area estimation methods are mainly based on the development of a regression model. Thus the availability and choice of auxiliary data have a direct effect on the quality of the estimates.

REFERENCES

Elbers, C., Lanjouw, J.O. and Lanjow, P. Micro-level estimation of welfare. Policy Research Working Paper n. 2911. World Bank, Washington, 2002.

Fay, R.E. and Herriot, R.A. Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association. V. 74, p. 269-277, 1979.

IBGE. Mapa de pobreza e desigualdade: municípios brasileiros 2003. Rio de Janeiro, 2008.

IBGE. Perfil dos municípios brasileiros: pesquisa de informações básicas municipais – gestão pública 2002. Rio de Janeiro, 2005. 120 p.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Sinopse estatística da educação básica: censo escolar 2003. Brasília, 2004.

IPEA/IBGE (Instituto de Pesquisa Econômica Aplicada / Fundação Instituto Brasileiro de Geografia e Estatística). Dimensionando a pobreza no Brasil: uma proposta metodológica (in press).

National Statistical Service. A guide to small area estimation. Austrália, 2006. Available at: < http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas +Estimates?OpenDocument >

Pessoa, D.G.C., Corrêa, S. and Quintaes, V.C.C. (2010). Comparação de Estimadores da Proporção de Domicílios Pobres em Minas Gerais Obtidos por Estimação Direta, Modelo de Fay-Herriot e Método ELL. Rio de Janeiro: IBGE, Coordenação de Métodos e Qualidade, 6 p.

Rao, J.R.K. Small Area Estimation. New York: Wiley, 2003.

Sistema FIRJAN – Federação das Indústrias do Estado do Rio de Janeiro. Índice FIRJAN de desenvolvimento municipal 2000. Available at: http://www.firjan.org.br >.

ABSTRACT

The Brazilian Institute of Geography and Statistics (IBGE) faces the challenge of producing comprehensive, accurate and reliable information under financial and time constraints. The pressure for reducing sample sizes and respondent burden reveals increases the need for methods to produce small area statistics from combined data sources. Small area estimation covers a variety of methods used to produce survey-based estimates for geographical areas or domains of study in which the sample sizes are too small to provide reliable direct estimates. This work presents a first attempt to implement an area level Fay-Herriot model to estimate the proportion of households in poverty for municipalities of Minas Gerais state combining survey data from the Family Expenditure Survey with Census and administrative data. This work is part of a project aiming at developing small area estimation models and procedures to allow publication of statistical outputs for areas or domains that are not currently considered for publication due to the low precision of the estimates (or lack of sample observations in some of these areas or domains). IBGE has already published a first set of poverty estimates based on the so-called World Bank Method (Elbers, Lanjouw, and Lanjouw , 2002). The focus of this work is to test some well-known small area estimation models that have been used successfully by other National Statistical Institutes and compare results with those obtained under the World Bank Method. The paper describes the steps taken for developing this area level model. Results obtained so far proved the feasibility of the area level approach and the vital role of good quality auxiliary data but also indicated the need to revise the choice of target areas for estimation in order to avoid the use of synthetic estimator to obtain estimates for large numbers of areas with no sample observations.

Key-words: Fay-Herriot model, small area estimation, poverty, Brazilian Family Expenditure Survey