Small Domain Estimation for a Brazilian Service Sector Survey

André Neves[1], Denise Silva[2] and Solange Correa[3]
[1] Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil
[2]National School of Statistical Sciences, Rio de Janeiro, Brazil
[3]University of Southampton, Southampton, United Kingdom
[2]Corresponding author: André Neves, email: andre.neves@ibge.gov.br

Abstract

Small domain 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 occurs when the survey is not designed for estimation at the required level. In order to obtain reliable estimates, additional datasets are generally brought to bear upon the process and a possible solution for this problem is the use of model based estimators. In the case of economic surveys, there are additional issues related to the asymmetry of the data and the non-linear relationship among economic variables. This paper presents the use of small area estimation methods for the Service Annual Survey conducted by the Brazilian Institute of Geography and Statistics. Due to the sampling design, sample estimates for some economic activities in the North, Northeast and Midwest regions of Brazil have low precision. To solve this problem, an area level Fay-Herriot model is used to produce model based estimates. The small domain estimation model relates operating revenues with auxiliary variables obtained from a Business Register and provides results showing improvement in precision for the majority of the domains considered.

Key Words: official statistics, business surveys, sample surveys, small area estimation.

1 – Introduction

The Brazilian Institute of Geography and Statistics (IBGE) carries out regular surveys for different sectors of the economy, including the Service Sector Annual Survey (Pesquisa Anual de Serviços – PAS) which focuses on segments of the tertiary sector (IBGE, 2010). Due to the growing demand for information with greater spatial detail and thematic range (SILVA and CLARKE, 2008), the traditional process of producing only design based estimates from this economic survey is not satisfactory. Depending on the geographic region, the survey provides information for industry service sectors at different levels of aggregation. For the South and Southeast regions, survey estimates are produced by economic activities according to disaggregation level defined by four-digit codes of the National Classification of Economic Activities. This classification was developed based on the International Standard Industrial Classification. For States in the North, Northeast, Midwest regions, the survey provides estimates by the so-called *group* (a level of disaggregation related to three-digit codes of the economic activity classification). The main objective of this paper is to apply model based methods to estimate the total operational gross revenue by areas (States) and economic activities.

2 – The Brazilian Service Sector Annual Survey

The survey encompasses a set of activities from the services sector and investigates enterprises' economic and financial characteristics such as revenue and expenses. Estimates are published for groups of economic activities according to the National Classification of Economic Activities and by States (IBGE, 2010). The sampling unit is the enterprise and the sampling frame is a business register maintained by IBGE based on administrative records. Its sample design is stratified by economic activities and geographical areas (States) and also according to the number of employees on 31st of December, as reported in the business register. The sample design involves a take-all stratum comprised of those enterprises with 20 or more

employees and enterprises that have establishments in more than one Brazilian State. The remainder of the sampling frame units (enterprises with less than 20 employees and operating in only one State) is then distributed into three other strata according to number of employees (0-4, 5-9 and 10 to 19).The final sample is comprised by those enterprises from the take-all stratum and by those that are randomly selected by simple random sampling within the strata defined by economic activities, geographical areas (States) and enterprise size (number of employees).In this work, a subset of economic activities was considered as the domains of study. The idea is to focus the analysis on activities in which the enterprises operate mainly in one State only. Table 1 displays the subset of domains considered, according to the sample design. For most of the States, direct survey estimates are solely produced by *group* (3-digit economic classification).

Table 1 - Disaggregation level of economic classification for which direct estimates are published – services considered in this study

Services	Economic Classification	
	For States in South and Southeast Regions	For Other States
Food and beverage service activities	5611-2	561
Renting of video tapes and disks	7722-5	772
Renting of clothing, jewellery and accessories	7723-3	
Teaching of art and culture	8592-9	859
Foreign language instruction	8593-7	
Activities of fitness center	9313-1	931
Washing and cleaning of textile and fur products	9601-7	960
Hairdressing and other beauty treatment	9602-5	

Source: IBGE, Service Annual Survey 2008.

The population of enterprises in the activities showed in Table 2.1 is comprised of 276,231 enterprises out of 1,222,132 listed in the 2008 sampling frame. The sample, in turn, consists of 11,751 enterprises and 213 domains (defined by states and *class* of economic activity). In order to set up a small area estimation framework it is necessary to define the target of the estimation (the variable of interest and the geographical level or domain requirements) and also to identify the potential auxiliary data. In this paper, the variable of interest is the *gross operating revenue*. Auxiliary variables were obtained from administrative data produced by the Brazilian Ministry of Labour and Employment based on information provided compulsorily by enterprises each year: *employed persons*, *wages* and *number of establishments*. The focus of this work is to produce model based estimates by *class* (4-digit economic classification) for a subset of services in all 27 States of Brazil in order to assess the potential use of standard small area estimation methods for improving the production of official statistics from Brazilian business surveys.

3 –The Fay-Herriot Area Level Model

The well-known Fay-Herriot model (Fay and Herriot, 1979) is defined by two equations representing: the *sampling model* (3.1), that relates the direct estimator to the true parameter value, and the *binding model* (3.2) that relates the quantity of interest to the auxiliary variables for the small domains (Sample Project, 2011; Rao, 2003; Pfeffermann and Correa, 2012).

Let Y_j be the population total of y for domain j and $\tilde{Y}_j$ the corresponding direct survey estimator. The area level model is given by:

$$\tilde{Y}_j = Y_j + \varepsilon_j \quad (3.1)$$

$$Y_j = \mathbf{x}_j^t \boldsymbol{\beta} + u_j \,, \quad (3.2)$$

where: $j = 1,...,J$ are the domains; $\varepsilon_j \overset{ind}{\sim} N(0,\sigma_j^2)$ - is the sampling error related to $\tilde{Y}_j$, with mean zero and known sampling variance σ_j^2, $u_j \overset{iid}{\sim} N(0,\sigma_u^2)$ is the random effect of domain *j* with mean equal to zero and variance σ_u^2, ε_j and u_j are mutually independent. The vector of auxiliary variables for domain *j* is defined by the column vector $\mathbf{x}_j = (x_{j1},...,x_{jP})^t$ whereas $\boldsymbol{\beta} = (\beta_1,...,\beta_P)^t$ is a column vector of unknown parameters. Putting together (3.2) in (3.1), the Fay-Herriot model is defined by:

$$\tilde{Y}_j = \mathbf{x}_j^t \boldsymbol{\beta} + u_j + \varepsilon_j \quad (3.3)$$

The above formulation specifies a linear model with a random intercept that varies by small domains. As the model relates the small domain totals to a group of auxiliary variables at the domain level, this type of model is often referred to as a basic area level model.

In the case of business surveys, there are additional issues related to the asymmetry of the data and the non-linear relationship among economic variables that has to be considered. It is common to apply a logarithmic transformation to the response variable in an attempt to represent economic data by linear models. To obtain the variance estimator of the transformed variable, the Taylor's approximation (Bishop *et al*, 1975) can be used, such that $Var(g(\tilde{Y})) \approx g'[E(\tilde{Y})]^2 \cdot Var(\tilde{Y})$, where *g'* is the first order derivative of the *ln* (natural logarithm) function and, hence, $Var(g(\tilde{Y})) \approx CV^2(\tilde{Y})$. The model based estimates as well as the corresponding mean squared errors (mse) in the log-transformed scale were obtained using the well-known estimators proposed by Fay and Herriot (1979). The estimates were converted back to the original scale based on the properties of the log-normal distribution (RAO, 2003). The EBLUP estimates and corresponding mean squared errors in the original scale are given by:

$\hat{Y}_{j,EBLUP,Final} = \exp\{\hat{Y}_{j,EBLUP} + 0.5 \cdot [mse(\hat{Y}_{j,EBLUP})]\}$ and

$mse(\hat{Y}_{j,EBLUP,Final}) = \exp(\hat{Y}_{j,EBLUP})^2 \times mse(\hat{Y}_{j,EBLUP})$.

4 – Results

The Fay-Herriot estimator in equation (3.8) was used to produce model based estimates for small domains for the 2008 Brazilian Service Sector Annual Survey. The model dependent variables are the survey direct estimates of *total revenue* for the small domains defined by geographical area and economic activity as described in the previous section. The auxiliary variables used in the model are the total number of *employed persons*, *wages* and *number of establishments*. For this study in particular, graphical analysis and statistical tests provided evidence that a *double-log* model would be more suitable to model the relationship between the direct estimates and the auxiliary variables. This functional form consists in applying logarithm transformation to the dependent variable and auxiliary variables as presented in Table 2.

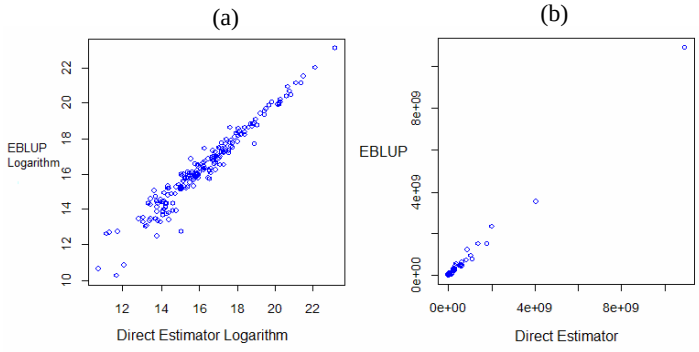
Table 2 – Regression coefficient estimates for the Fay-Herriot model

Auxiliary Variables	Estimates	Standard error	Z-value	P-value
Intercept	2.358	0.486	4.856	<0.000
Logarithm of *employed persons*	0.129	0.058	2.224	<0.030
Logarithm of *wages*	0.878	0.057	15.496	<0.000

A scatter plot of direct survey estimates (*y*-axis) against model based estimates (*x*-axis) is one method of assessing whether the relationship between the target variable and the covariates has been properly specified. The small domain estimation approach was used to overcome the problem of small sample sizes within domains. However, it is well known that, although in general more precise, the resulting model based estimates are biased. The aim of the estimation procedure adopted here is to balance the trade-off between variability and bias, producing estimates with good precision and as little bias as possible. Although the direct estimates are very variable, they are nevertheless unbiased. Thus a scatterplot of direct estimates (on the *y* axis) against model based estimates (on the *x* axis) should display a regression line close to the *y=x* line if the model based estimates are also unbiased.

Figure 1 (a) shows a scatter plot of the total revenue estimates to investigate if the model based estimates are approximately unbiased. The results show that the direct and model-based estimates do appear to track each other on the logarithm scale. The graph on the original scale (Figure1 (b)) has many overlapping data due to the asymmetric distribution of the variable *gross operating revenue*. To further evaluate the presence of bias in the model-based estimates, a regression line was fitted for the log direct and log model-based estimates, and the intercept was not significantly different from zero, indicating that there is no evidence to reject the hypothesis of lack of bias in the model-based estimates.

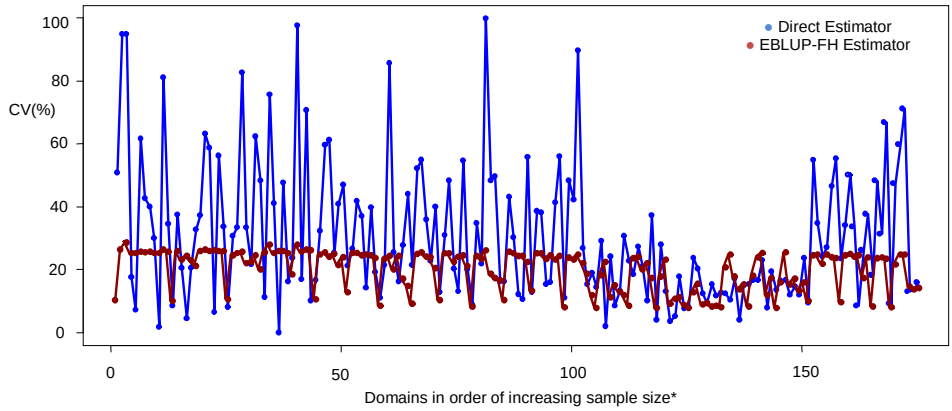
Figure1 – Scatter plots for direct estimates and model based Fay-Herriot estimates (EBLUP), in the logarithmic scale (a) and in the original scale (b):



To evaluate the accuracy of the model estimates, the coefficients of variation (CVs) were calculated for the direct and model based (EBLUP) estimates. The comparison considers the MSE estimates in the original scale. The highest CVs of the direct estimates are greatly reduced with the use of the model-based EBLUP estimator (Figure 2).

Figure 2 provides evidence that for the majority of domains, the precision of the model based figures is better than for the survey direct estimates since the small area estimation approach produces estimates with lower CVs for 83.5% of the domains considered when compared with those obtained for the direct estimates. Table 3 presents the results for the State of Piauí. A more detailed analysis can be found in Neves (2012).

Figure.2 –Direct and model based estimates of total gross revenue and corresponding precision estimates for specific domains



*This numeration is the list of domains. About domains size distribution, see Neves, 2012.

Table 3 – Estimates for the domains and their coefficients of variation

Domains		Direct Estimation	CV	EBLUP-FH Estimation	CV
State of Piauí	Food and beverage service activities	77.438.734	21,1	85.121.570	2,9
	Renting of video tapes and disks	1.128.425	26,6	2.138.840	4,9
	Renting of clothing, jewellery and accessories	1.296.512	41,7	1.832.639	3,6
	Teaching of art and culture	3.189.312	37,0	3.644.969	2,8
	Foreign language Instruction	2.555.536	14,1	2.968.606	3,6
	Activities of fitness center	3.083.838	39,7	4.924.355	2,2
	Washing and cleaning of textile and fur products	6.257.175	19,1	9.991.417	1,1

Source: IBGE, PAS 2008.

The Shapiro-Wilk test provides the *W* test statistic to assess whether the standardized residuals $\tilde{\xi}_j = (\varepsilon_j + u_j)/\sqrt{\sigma_j^2 + \sigma_u^2}$ follow a standard normal distribution. Small values of *W* are evidence against the hypothesis of normality. The *W* statistic is given by:

$$W = \frac{\left(\sum_{j=1}^{J} a_j \cdot \tilde{\xi}_j\right)^2}{\sum_{j=1}^{J} (\tilde{\xi}_j - \bar{\tilde{\xi}})^2}, \qquad (4.1)$$

where $\tilde{\xi}_j$ are the ordered values of the standard residuals, $\bar{\tilde{\xi}}_j$ is the corresponding mean and a_j are generated from a normal distribution with the means, variances and covariances of the order statistics for a sample of size *J* (Hidiroglou, 2011). In this study we obtained $W = 0.766$ and *p-value* $= 2.4 \times 10^{-6}$, which yields to rejection of the hypothesis of normality of the standardized residuals.

5 – Conclusions

In this study we have considered the Fay-Herriot model, a procedure for small area estimation commonly cited in the literature. The overall performance of the Fay-Herriot model was very good, showing lower CVs for the model based estimators for 83% of the domains considered when compared to the CVs obtained for the direct estimates. However, the statistical tests showed that the model residuals do not meet the assumption of normality, which indicates the need for additional study in this direction. This article applies a well-known method for small domain estimation with application to a real economic survey data conducted by IBGE. The main objective of this article is to take another step in the direction of development and application of

small area estimation approaches to IBGE's economic surveys. The promising results found in this study along with its practical importance and relatively new initiative of application to Brazilian economic survey data make this topic an area that certainly deserve further research.

BIBLIOGRAPHY

BISHOP, Y.M.M; FIENBERG, S.E.; HOLLAND, P. W. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge-Massachussets, London-England, 1975.

FAY, R. E., HERRIOT, R. A. *Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data*. Journal of the American Statistical Association, Vol. 74, n° 366. Jun/79, p.269-277.

PFEFFERMANN, D., CORREA, S. *Empirical Bootstrap Bias Correction and Estimation of Prediction Mean Square Error in Small Area Estimation*. Biometrika, Vol. 99, n° 2. April/2012, p.457-472.

IBGE. *Pesquisa Anual de Serviços 2008*. Diretoria de Pesquisas, Coordenação de Serviços e Comércio, 2010.

HIDIROGLOU, M. A. *Small area estimation – Fay-Herriot Area Level Model with EBLUP Estimation* (*methodology specifications*). Methodology Software Library, 11/07/2011.

NEVES, A. F. A. *Estimação em Pequenos Domínios Aplicada à Pesquisa Anual de Serviços 2008*. Dissertação do Mestrado em Estudos Populacionais e Pesquisas Sociais. Escola Nacional de Ciências Estatísticas. Rio de Janeiro, Jul/2012.

RAO, J.N.K. *Small Area Estimation*. New York, Wiley, 2003.

SAMPLE Project. *Software Beta on Small Area Estimation*. Deliverable number 13. Link: www.sample-project.eu/images/stories/docs/samplewp2d13_softbeta.pdf

SILVA, D. B. N; CLARKE, P. *Some Initiatives on Combining Data to Support Small Area Statistics and Analytical Requirements at ONS-UK*. Paper presented at the IAOS 2008 Conference on Reshaping Official Statistics.