Small Domain Estimation for a Brazilian Service Sector Survey

André Felipe Azevedo Neves
Brazilian Institute of Geography and Statistics – IBGE

Denise Britz do Nascimento Silva
National School of Statistical Sciences – ENCE - IBGE

Solange Corrêa Onel
University of Southampton

First Asian ISI Satellite Meeting on Small Area Estimation 2013

Motivation

- The Brazilian Institute of Geography and Statistics (IBGE) carries out regular business surveys, including the Service Annual Survey that focusses on segments of the tertiary sector
- The survey provides information about service sectors at different levels of aggregation according to geographic region
- Need to produce estimates for domains of study with small sample sizes (unreliable direct estimates)

Motivation

- **States of South and Southeast regions:** survey estimates produced for economic activities defined by *4-digit* codes of the National Classification of Economic Activities (ISIC)
- **States of North, Northeast and Midwest regions**: estimates provided by *group* (ISIC *3-digit codes)*
- **Objective:** to employ a model based approach to estimate total operational gross revenue by States and Economic Activities currently not published due to the survey sampling design

The Brazilian Service Sector Annual Survey

Sector of services

Economic activities related to the production of intangible goods: transportation, technical services, information services, food services, etc.

Scope of the Survey

Non-financial business services for

Coverage

All Brazilian States

Variables

Economic and financial characteristics such as revenue and expenses plus workforce composition

Survey Design

Stratified survey sampling design

- by economic activity, geographical areas (States) and also according to the number of employees
- **Small domains:** North, Northeast, Middle West and Espírito Santo States

Sampling frame

Business register based on administrative records

Sampling unit: Enterprise

Sample Design

Stratified sample design

First level Strata: defined for publication

State by Activity at 3 or 4 ISIC digits (according to Region)

- **In each first level stratum:**

  -**Take-all stratum**:

  enterprises with number of employees ≥ 20

  enterprises with number of employees < 20 but operating in more than one State

  - **Sampling stratum**:

  enterprises with number of employees < 20

Scope of the Study

Survey population: 276,231

Sample size: 11,751 enterprises and 213 domains (defined by states and ISIC codes)

Percent distribution (%)	Domain sizes	
	N	*n*
0	1	1
10	9	3
20	28	4
30	44	7
40	76	8
50	126	12
60	172	15
70	331	21
80	694	29
90	1.715	100
100	85.037	2.564

ISIC codes for which direct estimates are published

Services	Economic Classification	
	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.

Small Area Estimation Methods

- **Fay-Herriot model (1979) –** area\domain level
- **Battese at al. (1988) –** unit level
- **Kurnia at al. (2009) –** unit level log response with area level covariate

- **Target parameter:** gross operating revenue per domain

- **Auxiliary variables (from the business register):** number of employees, wages, number of establishments, indicator of one-person enterprise, indicator of enterprise operating in more than one state

Small Area Estimation Methods

- **Fay-Herriot model (1979) –** area\domain level
- **Battese at al. (1988) –** unit level
- **Kurnia at al. (2009) –** unit level log response with area level covariate

- **Target parameter:** gross operating revenue per domain

- **Auxiliary variables**: number of employees, wages, number of establishments, indicator of one-person enterprise, indicator of enterprise operating in more than one state

Fay-Herriot Area Level Model

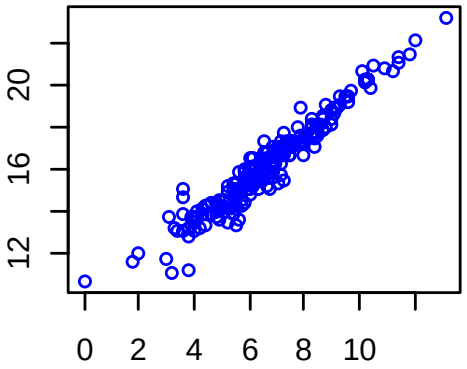
- **Response variable: log of** direct estimate of the total revenue per domain

- **Auxiliary variables: log of** (*number of employees*, *wages* and *number of establishments*)

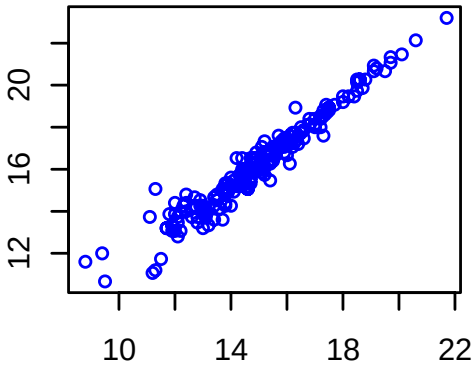
$$\left.\begin{aligned} \tilde{Y}_j &= Y_j + \varepsilon_j \\ Y_j &= \mathbf{x}_j^t \boldsymbol{\beta} + u_j \end{aligned}\right\} \quad \tilde{Y}_j = \mathbf{x}_j^t \boldsymbol{\beta} + u_j + \varepsilon_j \quad j = 1, \ldots, J$$

$$u_j \overset{iid}{\sim} N(0, \sigma_u^2) \qquad \varepsilon_j \overset{ind}{\sim} N(0, \sigma_j^2)$$

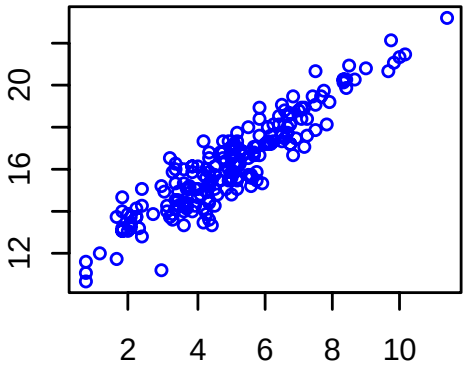
•**Response variable: log of** direct estimate of total revenue per domain



Log number of employees



Log total wages



Log number of establishments

Results – Fay-Herriot Model

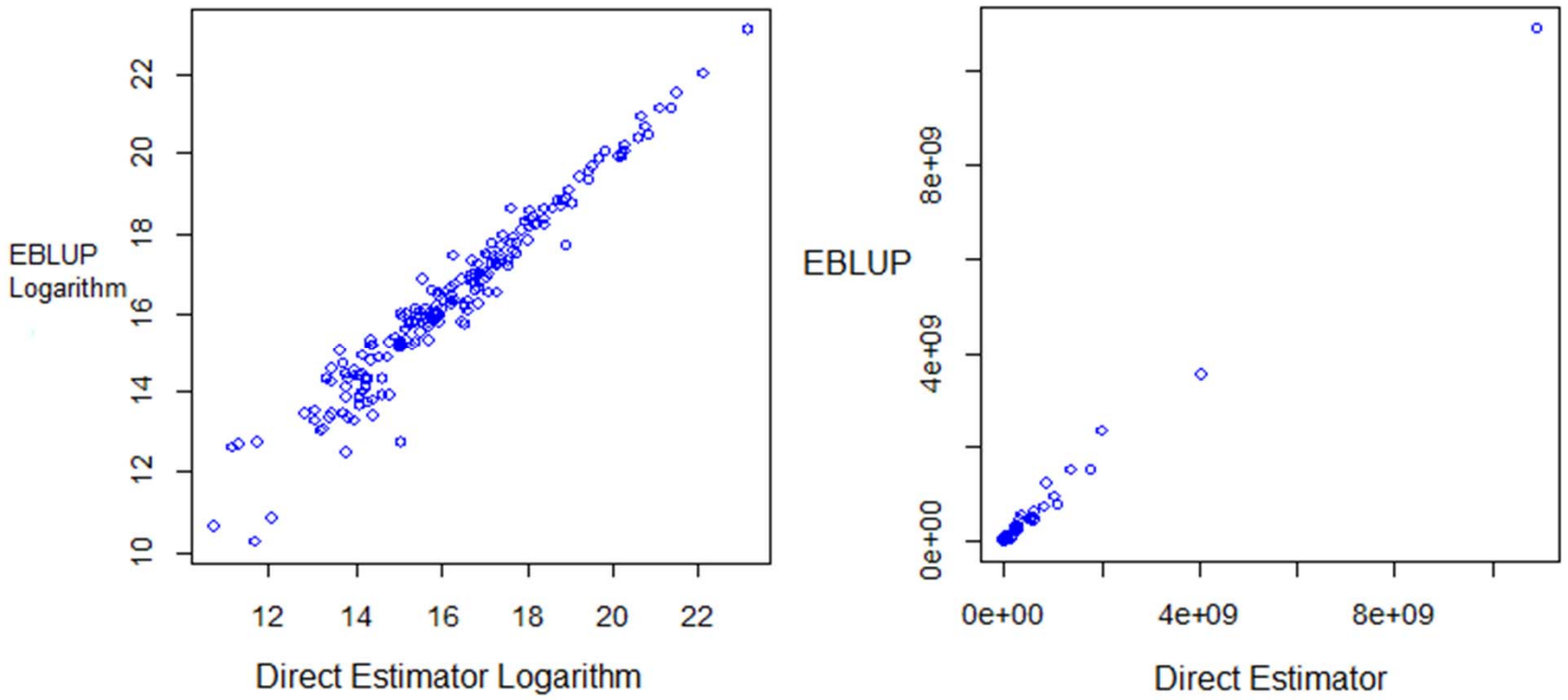
Coefficient Estimates

Auxiliary Variables	Estimates	Standard error	P-value
Intercept	2.358	0.486	<0.000
Logarithm of *number of employees*	0.129	0.058	<0.030
Logarithm of *wages*	0.878	0.057	<0.000

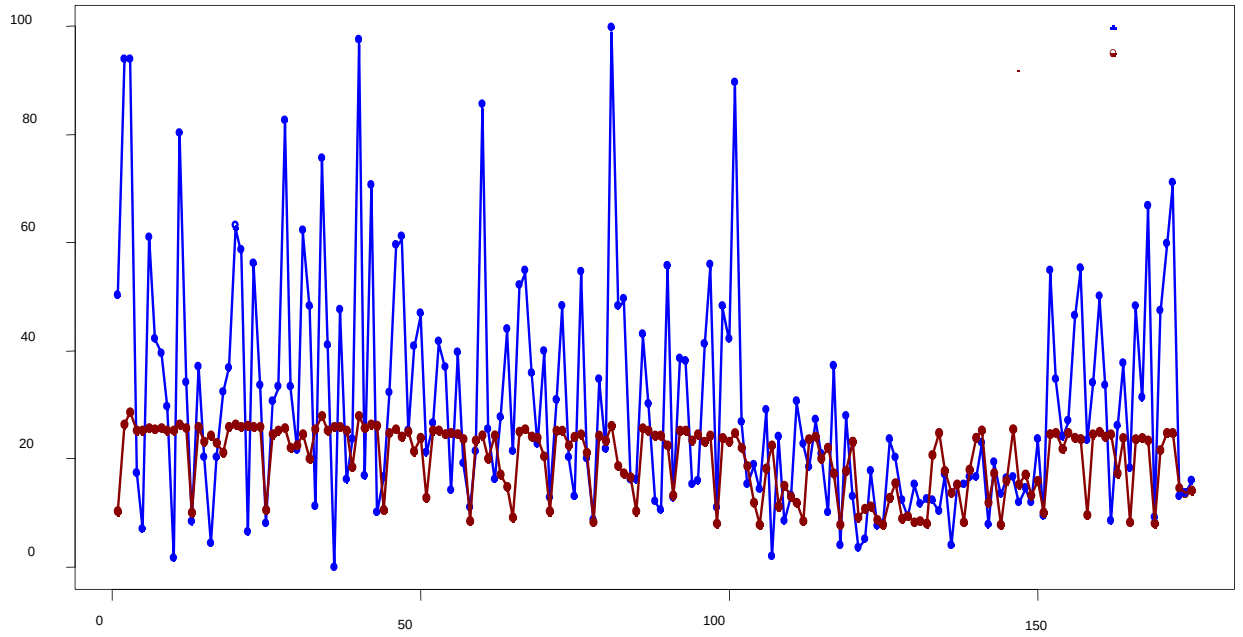
$R^2 \geq 0.90$ for linear regression model

Bias Diagnostic

Direct and model based Fay-Herriot (uncalibrated) estimates - logarithmic and original scales



Estimated CV% of Direct and Model Based Estimates of Total Operating Revenue



Results for State of Piauí

Activities	Direct Estimates	CV	FH Estimates	CV
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 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

Comments – Area Level Model

- Results showed considerable reduction on the estimated CVs for 83% of domain (when comparing model based and direct estimator)

- Promising results that encourage further research

- However…
  - ✓ evidence of non normality of the residuals
  - ✓ when testing $\tilde{\bar{Y}}_j = \alpha + \beta \cdot \hat{\bar{Y}}_{j,EBLUP}$ there is evidence to reject the hypothesis $H_o : \alpha = 0$

Unit Level Model – Results

Auxiliary Variables	Estimates	Standard Error	t-value	P-value
Intercept	25.953	0.195	132.9	<0,000
Log number of employees	0.184	0.014	12.7	<0,000
Log of wages	0.847	0.012	69.3	<0,000
Log of number of establishments	0.061	0.016	3.9	<0,000
Enterprise operates in more than one state	0.157	0.057	2.7	<0,007
One-person enterprise	-0.236	0.020	-12.0	<0,000
Null numbers of employees	-2.887	0.245	-11.8	<0,000
Total wages equal zero	-20.630	0.317	-65.1	<0,000

$R^2 = 0.73$ for linear regression VP=0.11

Problems:

- Many enterprises with zero value for number of employees and wages and even revenue

Unit Level Model – Results

- **Estimated CVs were reduced for 85.6% of the domains**
- However…estimates differ greatly from direct estimates
  - ➢ strong evidence of underestimation in large domains in which the results of the direct estimates are reliable

% Difference between EBLUP and Direct estimator

Percentile	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Difference	-100,0	-84,8	-79,1	-74,0	-69,6	-63,8	-57,7	-49,8	-41,2	-25,5	150,9

- This may suggest that unit level model based estimates are biased
- Unit level model may fail due to the non-inclusion of sampling weights (very large values or less than 1)

Conclusions

- First initiative to use small area estimation approach to Brazilian business survey data
- The overall performance of the Fay-Herriot model was very good showing lower coefficients of variation for the model based estimators for most of domains
- However, statistical tests showed that the model residuals do not meet the assumption of normality
- The unit level estimator produced estimates with low CVs compared to the direct estimates ones.
  - results were very discrepant in comparison to direct estimates

Futre work

Employ models that account for skewed distributions or mixture models that account for data with many zero values

http://www.ence.ibge.gov.br/web/ence/mestrado/dissertacoes/2012

Estimação em Pequenos Domínios Aplicada à Pesquisa Anual de Serviços 2008.

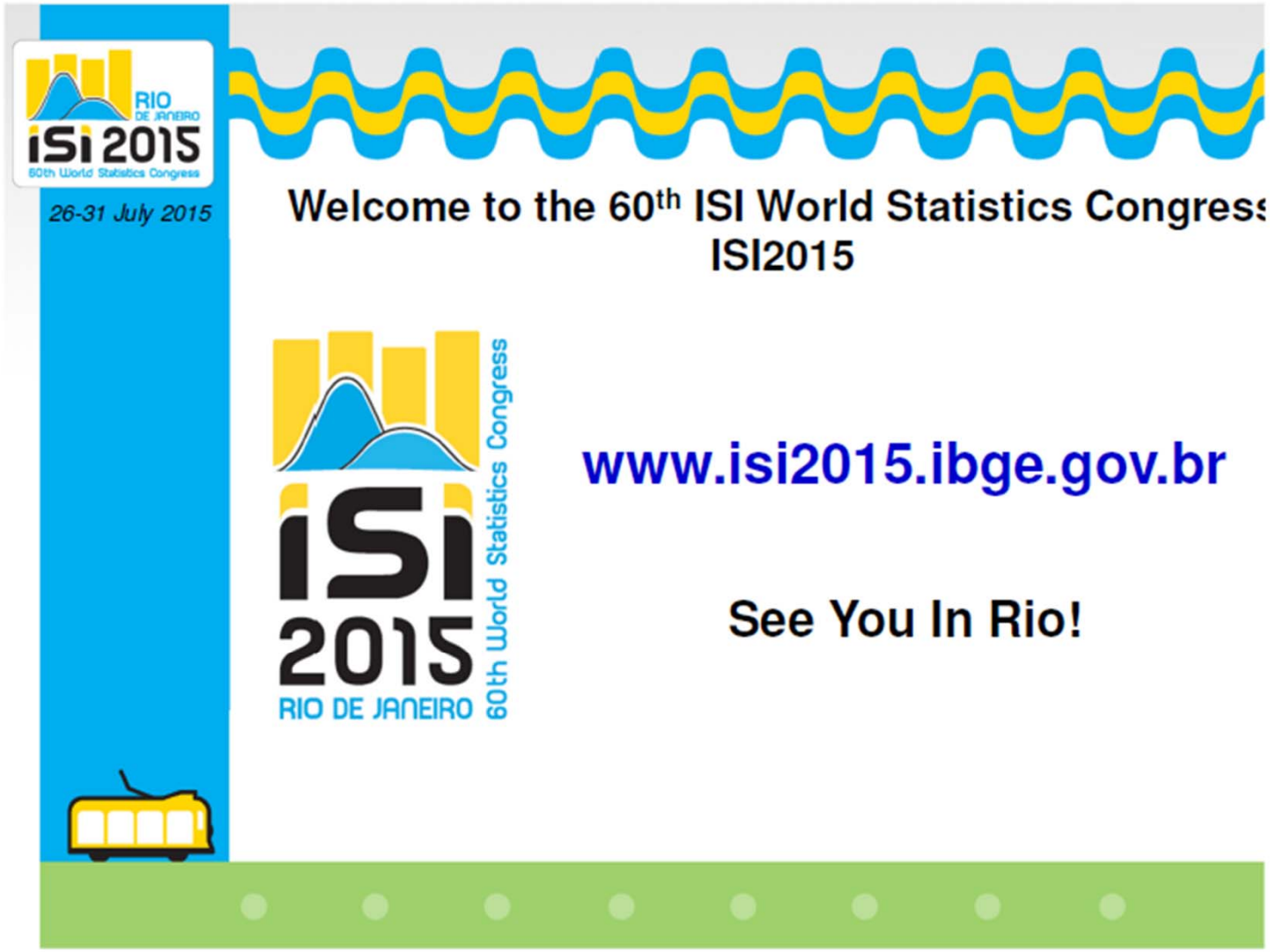
Autor(a): André Felipe Azevedo Neves

Orientador(a): Denise Britz do Nascimento Silva

Co-orientador(a): Solange Corrêa Onel

Resumo Texto completo

English version: 6 pages paper for WSC2013 – Hong Kong


RIO DE JANEIRO
ISI 2015
60th World Statistics Congress
26-31 July 2015
Welcome to the 60th ISI World Statistics Congress
ISI2015
ISI 2015
RIO DE JANEIRO
60th World Statistics Congress
www.isi2015.ibge.gov.br
See You In Rio!

Bibliography

BATTESE, G.E.; HARTER, R.M. FULLER, W.A. *An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data*. Journal of the American Statistical Association, vol.83, núm.401 (mar.1988), pág. 28-36.

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.*

Bibliography

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. *Small Domain Estimation for the 2008 Service Sector Survey. Master dissertation in Population Studies and Social Researches* (originally published in Portuguese)*. National School of Statistical Sciences of IBGE. 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-roject.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.*