STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 84–102, DOI 10.21307/stattrans-2020-032
Received – 31.01.2020; accepted – 30.06.2020

Skew normal small area time models for the Brazilian annual service sector survey

André Felipe Azevedo Neves[1], Denise Britz do Nascimento Silva[2], Fernando Antônio da Silva Moura[3]

ABSTRACT

Small domain estimation covers a set of statistical methods for estimating quantities in domains not previously considered by the sample design. In such cases, the use of a model-based approach that relates sample estimates to auxiliary variables is indicated. In this paper, we propose and evaluate skew normal small area time models for the Brazilian Annual Service Sector Survey (BASSS), carried out by the Brazilian Institute of Geography and Statistics (IBGE). The BASSS sampling plan cannot produce estimates with acceptable precision for service activities in the North, Northeast and Midwest regions of the country. Therefore, the use of small area estimation models may provide acceptable precise estimates, especially if they take into account temporal dynamics and sector similarity. Besides, skew normal models can handle business data with asymmetric distribution and the presence of outliers. We propose models with domain and time random effects on the intercept and slope. The results, based on 10-year survey data (2007-2016), show substantial improvement in the precision of the estimates, albeit with presence of some bias.

Key words: Annual Service Sector Survey, hierarchical Bayesian model.

1. Introduction

Small area estimation approaches aim at obtaining precise estimates for geographic areas or domains for which sample sizes are not sufficient to yield satisfactory precision if direct estimators are used. The issue of small area estimation can arise from the demand for information on a specific group such as when estimates for an industrial district or other restricted segment are required.

The *small area (domain) estimation* problem has received much attention in recent decades, in which Fay and Herriot (1979) and Battese, Harter and Fuller (1988) are two key papers. The first considered an area level model in which the input response variable is the direct estimate and auxiliary information comes from area level variables. Battese, Harter and Fuller (1988) proposed a unit level model with both input and auxiliary variables considered to be available at the unit sample level. The Fay-Herriot model uses data at the domain level, with greater scope for application compared to models at the

[1]National School of Statistical Sciences. Brazil. E-mail: andre.neves@ibge.gov.br.
ORCID: https://orcid.org/0000-0001-9819-2300.
[2]National School of Statistical Sciences. Brazil. E-mail: denise.silva@ibge.gov.br.
ORCID: https://orcid.org 0000-0002-5514-7558.
[3]Statistics Department of Federal University of Rio de Janeiro. Brazil. E-mail: fmoura@im.ufrj.br.
ORCID: https://orcid.org/0000-0002-3880-4675.

sampling unit level since aggregated data are more accessible and are less subjected to statistical confidentiality restrictions. However, as pointed out by Moura *et al.* (2017), the Fay–Herriot model assumes conditional normality of the direct estimator which is not suitable for fitting skewed data, particularly for domains with very small sample sizes.

Neves *et al.* (2013) developed the first small domain estimation approach for Brazilian economic surveys. The authors proposed a Fay-Herriot model for the logarithmic transformation of the variable of interest to stabilize the variance resulting from the presence of outliers. However, due to difficulties when converting the results to the original scale, a better alternative is to use an asymmetric distribution to model the direct estimator. Ferraz and Moura (2012) modeled the direct survey estimator as skew normally distributed. They successfully fitted the skew normal model to head-of-household mean income for 140 enumeration areas in the scope of an experimental Brazilian demographic census. Moura *et al.* (2017) compared different small area approaches for fitting skewed data using real business survey data. It was the first experiment in which skew normal models in a Bayesian framework were tested to produce small area estimates for the Brazilian Annual Service Sector Survey (BASSS). The main objective was to develop models for estimating service revenue totals by economic activity at levels of aggregation not planned in the BASSS sampling design.

Considering earlier research and corresponding developments, the principal aim of this work is to extend the previous skew normal models to allow sharing information from repeated surveys, such as the BASSS. We consider models to estimate *gross service revenue* totals in specific groups of economic activities (class level four-digit codes of the International Standard Industrial Classification - ISIC) for states in the Northeast region of the country since these direct survey estimates are not currently published due to small sample size and low precision (Neves, 2012).

This paper is organized as follows. Section 2 presents the Brazilian Annual Service Sector Survey and the small domain estimation problem. Section 3 introduces the skew normal models and their extensions to skew normal time models whereas Section 4 displays results and related analysis. Section 5 contains final remarks and suggestions for future research.

2. Small Area Estimation for the Brazilian Annual Service Sector Survey

Service activity comprises the production of intangible goods for immediate consumption by individuals and institutions. Activities with these characteristics include commerce, transport, advertising, information and technology activities, health and education services, tourism and hospitality, financial and insurance services, and services provided by the public sector.

Although important to the Brazilian economy, the service sector occupies a less prominent position in public since industry is considered the most dynamic and important sector. However, as all sectors are vital for the efficient integrated functioning of the economy, reliable, detailed and timely statistics about the service sector are re-

quired. The BASSS is a non-financial services survey conducted by IBGE since 1998. It investigates economic and financial variables of companies, such as *revenues, costs and expenses, inventories, wages, number of employees* and *number of establishments*. Since firms control the accounting records of all their local units (establishments), where the economic and financial results are registered, the BASSS survey unit is the enterprise – the legally constituted unit that produces services.

Table 1. Disaggregation level of economic classification for which direct estimates are published and services in the scope of this study

Service	*Economic classification*	
	4-digit code (small domains)	*2-3-digit code (published estimates)*
Food and beverages	5611-2	561
Engineering and architecture	7111-4, 7112-0, 7119-7	711
Advertising	7311-4, 7312-2, 7319-0	731
Renting and leasing of personal and household goods	7722-5, 7723-3, 7729-2	772
Travel agency and tour operator activities	7911-2	79
Cleaning and pest control	8121-4, 8122-2	812
Foreign language instruction	8593-7, 8599-6	859
Creative, arts and entertainment activities	9001-9	90
Fitness centers and other physical activity providers	9313-1	931
Other personal services	9601-7, 9602-5, 9603-3	960

The survey frame is a business register comprised of administrative records with basic information about companies, such as wages, number of employees and number of establishments. The survey sample is stratified by economic activities and geographic areas (states), and also according to the number of employees. In addition, enterprises with 20 or more employees and those that operate in more than one Brazilian state are allocated in a *take-all stratum*. The survey publishes total estimates, and corresponding precision, by state and economic activity.

Here, we consider a subset of economic activities, focusing on activities in which the enterprises operate mainly in one state. Table 1 above shows the subset of domains in the scope of this study. Note that, for most of the country, direct survey estimates are only produced by group (3-digit code economic classification) due to the survey sampling design. Therefore, small domains are defined by the four-digit codes, listed in Table 1, in each of the nine Northeast Brazilian states.

Depending on the geographic region, the survey provides information at different levels of economic classification. For the South and Southeast regions, IBGE publishes *class level* data (four-digit codes) of the National Classification of Economic Activities (similar to ISIC). For the states of the North, Northeast and Midwest, survey results are only available at the *group level* (three-digit codes), therefore, at a lower level of activity breakdown (IBGE, 2018). Table 2 presents the number of enterprises and the sample sizes restricted to the services enumerated in Table 1. It also contains the number of small domains (defined by state and economic classification). We use 10-year data to develop models that can also borrow strength over time.

Table 2. Number of enterprises, sample sizes, number of domains and domain samples sizes in the scope of this study by year

Year	Population size	Sample size	Number of domains	Domain sample size	
				Median	Maximum
2007	46,056	730	81	9.0	17
2008	35,050	587	70	8.0	17
2009	37,733	637	72	9.0	16
2010	42,244	668	73	9.0	16
2011	46,501	675	74	9.0	15
2012	48,880	738	80	8.5	15
2013	48,976	665	76	8.0	16
2014	53,458	658	76	8.5	15
2015	52,019	660	80	8.0	13
2016	55,545	656	76	8.0	16

3. Skew normal small area models

Fay and Herriot (1979) developed a two-level linear model to estimate the average income per capita in small towns with less than 1,000 people in the United States. This model uses a direct estimator of the domain total and assumes residuals following a normal distribution, with zero mean and known sample variance.

The Fay-Herriot model incorporates random domain effects to capture variability between the domains that cannot be explained by fixed effects. The model is often cited in the literature of small domain estimation. Because the Fay-Herriot model uses data at the domain level, it allows a greater possibility of application when compared to unit level models considering that aggregated data are more easily accessible and are less subject to statistical confidentiality.

The basic Fay-Herriot model is defined in two stages. We denote by y_d the direct estimates of the true totals and as μ_d the response input variable of the model, where $d = 1,...,D$ are the domains of study. These estimates have a sampling error ε_d that depends on their respective sample sizes and the domain variability. Thus, the first stage model equation can be written as:

$y_d = \mu_d + \varepsilon_d$ – sampling model

$\varepsilon_d \overset{ind}{\sim} N(0, \phi_d),\ d = 1,...,D$

where ϕ_d is the sampling variance of the corresponding direct estimator, assumed known for all domains. In the second stage (linking model), the true values are assumed to be linearly related to a vector of auxiliary variables:

$\mu_d = \mathbf{x}_d^t \beta + \nu_d$ – linking model

$$v_d \overset{ind}{\sim} N(0, \sigma_0^2)$$

Errors ε_d and v_d are mutually independent. Substituting linking model equation in sampling model, we obtain:

$$y_d = \mathbf{x}_d^t \boldsymbol{\beta} + v_d + \varepsilon_d$$

Fay and Herriot (1979) assumed that the sampling variances are known and given by their respective sampling variance estimates. However, these estimates are unstable for areas with small sample sizes. There is a series of papers on joint modeling of survey-weighted estimates and sampling variances, see for example Arora and Lahiri (1997), and Gershunskaya and Savitsky (2019) for a recent discussion of this approach.

The Fay-Herriot model assumes that the sample size in each domain is large enough to apply the central limit theorem (CLT). However, in real situations, the response variable can be asymmetric, implying that assumptions of asymptotic normality are unreasonable in several domains. To overcome this problem, a response variable transformation, such as a logarithmic transformation, is commonly used. However, while the lognormal model makes the asymmetry hypothesis more plausible, an exponential function is required when estimates are converted to the original scale, increasing the variability of the estimates. Moreover, Moura *et al.* (2017) found that the lognormal model performs less well than the skew normal model in their application to BASSS data.

3.1. Skew normal model

Azzalini (1985) described the family of skew normal distributions that preserve some properties of the normal distribution except for the parameter that regulates the distribution's asymmetry. This class of distributions includes the normal distribution as a particular case and facilitates the transition from non-normality to normality. The properties of the skew normal distribution are suitable for asymmetric economic data. We adopt Azzalini's (1985) notation to describe the skew normal density function:

$$Y \sim SN(\mu, \sigma, \lambda) \Leftrightarrow f_Y(y) = \frac{2}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function, $\phi(\cdot)$ is the density function of the standard normal distribution, and the parameters μ,σ and λ are the *location*, *scale* and *asymmetry*, respectively. A particular case is the normal distribution when $\lambda = 0$. The skew normal distribution has interesting properties, some of which are shared with the normal distribution. The mean and variance of the skew normal distribution are given by:

$$E(Y) = \mu + \sigma\delta\sqrt{\frac{2}{\pi}} \quad \text{and} \quad V(Y) = \sigma^2\left\{1 - 2\pi^{-1}\delta^2\right\}$$

where δ is given by: $\delta = \lambda/\sqrt{1+\lambda^2}$.

Ferraz and Moura (2012) proposed the following model, here named Model 1, whose joint distribution of the direct estimator y_d and its sample variance estimator $\hat{\phi}_d$ are described in the following expressions:

$$
\begin{aligned}
y_d|\mu_d,\lambda,n_d,\phi_d &\sim SN(\mu_d,\sqrt{\phi}_d,\lambda/\sqrt{n_d}) \\
\hat{\phi}_d|n_d,\phi_d &\sim Ga\left\{\frac{1}{2}(n_d-1),\frac{1}{2}(n_d-1)\phi_d^{-1}\right\},\ d=1,...,D, \\
\phi_d^{-1}|a_\phi,b_\phi &\sim Ga(a_\phi,b_\phi) \\
\mu_d|\beta,\sigma_0^2 &\sim N(\mathbf{x}_d^t\beta,\sigma_0^2)
\end{aligned} \tag{1}
$$

where D is the number of small domains and n_d is the sample size in the d^{th} domain from a population of N_d units. They assume that the parameters ϕ_d, $d=1,\ldots,D$ are conditionally independent, following each an inverse-gamma distribution $\phi_d^{-1} \sim Ga(a_\phi,b_\phi)$, with unknown common hyperparameters a_ϕ and b_ϕ.

For BASSS survey data, μ_d can be written as a linear function of area-level auxiliary variables with unknown fixed coefficient and a random small domain effect β_{0d}, i.e., $\mu_d=\beta_0+\beta_{0d}+\beta_1 x_d$ where: i) the parameter β_0 is the global intercept; ii) β_{0d} is an intercept that varies by domain; iii) and β_1 is the slope. The auxiliary variable x_d is the total wage by domain, which comes from the business register used as the BASSS sample frame.

As the sample size grows, the skew normal distribution converges to the normal with mean μ_d and variance ϕ_d. Our main parameter of interest is $\theta_d^{sn}=E_d^{sn}(y_d)$, the expected value of y_d in the skew normal model, given by:
$\theta_d^{sn}=\mu_d+\delta_d\sqrt{2\phi_d/\pi}$ where $\delta_d=\lambda_d/\sqrt{1+\lambda_d^2}=\lambda/\sqrt{n_d+\lambda^2}$, with $\lambda_d=\lambda/\sqrt{n_d}$.

The sampling variance estimator $\hat{\phi}_d$ is assumed to be unbiased, providing information about the scale parameter ϕ_d. To borrow strength over domains, the model is completed through a hierarchical structure with respect to the parameters β_{0d} and ϕ_d. The parameters β_{0d} are hypothetically independent and distributed as $\beta_{0d} \sim N(0,\sigma_0^2)$.

The Ferraz and Moura model described by the equations in (1) is complemented by assigning a proper and independent prior distribution to the hyperparameters. When modeling the BASSS survey data, we assigned the following priors to these hyperarameters: $\beta=(\beta_0,\beta_1)^t \sim N_2(\mathbf{0},\Omega_\beta)$, $a_\phi \sim Ga(a,b)$, $b_\phi \sim Ga(c,d)$. To obtain relatively vague prior distributions, we set $\Omega_\beta=1000\mathbf{I}_2$, where $\mathbf{I}_2$ is an identity matrix of order 2 and $a=b=c=d=0.01$.

It is worth noting that Ferraz and Moura (2012) considered $\sigma_0^{-2} \sim Ga(a_0,a_0)$, with $a_0=0.01$. Since we experienced difficulties in fitting some models with this prior, we follow Gelman (2006) and placed a relative vague uniform prior on σ_0, i.e., $\sigma_0 \sim U(0,100)$.

The selection of a prior distribution to the λ parameter must be done carefully. Ferraz and Moura (2012), using results obtained in Sugden *et al.* (2000), proposed a normal distribution for the parameter λ, centered close to zero and with standard deviation given by $\sigma_\lambda=5.5a_\gamma/2.576$, where a_γ is an initial suggested value or estimate of the γ

asymmetry coefficient. For BASSS survey data, we estimated a_γ= 4.7. Therefore, the prior for γ was fixed at $\lambda \sim N(0, 100)$.

3.2. Skew Normal Time Models

In this section, we propose to generalize the skew normal model by introducing an extra random time effect (Models 2, 3 and 4). Models 2 to 4, showed in this section, take into account information from domains over time. As mentioned in Section 2, the BASSS data used here cover a 10-year period from 2007 to 2016. The models are developed to estimate the total *gross revenue* from services for 2016, the final year of this series. Therefore, Model 2 is written as:

$$y_{dt}|\mu_{dt}, \lambda, n_{dt}, \phi_d \sim SN(\mu_{dt}, \sqrt{\phi}_d, \lambda/\sqrt{n}_{dt})$$

$$\hat{\phi}_{dt}|n_{dt}, \phi_d \sim Ga\left\{\frac{1}{2}(n_{dt}-1), \frac{1}{2}(n_{dt}-1)\phi_d^{-1}\right\}$$

$$\phi_d^{-1}|a_\phi, b_\phi \sim Ga(a_\phi, b_\phi)$$

$$\mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt}$$

where $d = 1, ..., D$ denotes the domains of study in all years $t = 1, ..., T$ and n_{dt} is the sample size in the d^{th} domain in year t from the population of N_{dt} units. Note that μ_{dt} can be written as a linear function of area-level auxiliary variables with unknown fixed coefficients, a random small domain effect β_{0d} and a random time effect ζ_{0t}. Because the sample size for each domain does not vary much over the years, we assume that the true sampling variance of the direct estimator is constant over time.

The distributions of the inverse of scale parameter ϕ_d^{-1}, as well as the parameters a_ϕ and b_ϕ are the same as in Model 1. The distributions of the random coefficients under the influence of their respective random effects are defined by:

$$\beta_{0d} \sim N\left(0, \sigma_0^2\right) \text{ and } \zeta_{0t} \sim N\left(0, \sigma_{\zeta_0}^2\right)$$

We assigned a uniform prior distribution to the standard deviations σ_0 and σ_{ζ_0}. As discussed in Gelman (2006), the use of this prior guarantees a proper posterior density as well as other desirable properties. Thus, the relatively vague uniform priors for the standard deviations of both domain and time random effects on the intercept are:

$$\sigma_0 \sim U\left(0, 100\right) \text{ and } \sigma_{\zeta_0} \sim U\left(0, 100\right)$$

In addition, the following constraints are imposed to ensure identifiability of the parameters:

$$\beta_{01} = -\sum_{d=2}^{D} \beta_{0d} \text{ and } \zeta_{01} = -\sum_{t=2}^{T} \zeta_{0t}$$

3.3. Skew Normal Model with Random Effects on the Intercept and Slope

Following Moura and Holt (1999), Model 3 includes domain and time random effects on the intercept and a domain random effect on the slope, whereas Model 4 considers domain and time random effects on both intercept and slope:

$$\text{Model 3: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + \beta_{1d} x_{dt}$$
$$\text{Model 4: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + (\beta_{1d} + \zeta_{1t}) x_{dt}$$

As in Model 2, independent uniform priors with mean 50 are assigned to the standard deviations of both domain and time random effects, as follows:

σ_0^2 – variance of the domain random effect on the intercept,
σ_1^2 – variance of the domain random effect on the slope,
$\sigma_{\zeta_0}^2$ – variance of the time random effect on the intercept,
$\sigma_{\zeta_1}^2$ – variance of the time random effect on the slope.

The identifiability constraints are given by:

$$\beta_{01} = -\sum_{d=2}^{D} \beta_{0d},\ \zeta_{01} = -\sum_{t=2}^{T} \zeta_{0t},\ \beta_{11} = -\sum_{d=2}^{D} \beta_{1d} \text{ and } \zeta_{11} = -\sum_{t=2}^{T} \zeta_{1t}$$

3.4. Skew normal model with random walk effect

Rao and Yu (1994) proposed an extension of the Fay-Herriot model to handle cross-sectional and time-series data, see also Molina and Rao (2015) for further explanation and extensions. Unlike Rao and Yu (1994), Datta *et al.* (1999) employed a Bayesian method to implement a time series cross-sectional model with random walk component to estimate unemployment rates of U.S. states. Since it is reasonable to suppose influence of lag random effects when working with economic data, we also considered another model that includes an addictive random lag term effect of first order:

$$\mu_{dt} = \beta_0 + \beta_d + \beta_{0d,t} + \beta_1 x_{dt}$$

where $\beta_d \sim N(0, \sigma_0^2)$ and $\beta_{0d,t} \sim N(\beta_{0d,t-1}, \sigma_{\xi_0}^2)$ and they are all assumed independent. In Bayesian framework, it is also needed to assign prior distributions to $\beta_{0d,0}$ for $d = 1, \ldots, D$. We considered $\beta_{0d,0} \sim N(0, 100),\ \forall d$ and independently distributed. The other model components are analogously defined as the previous models. We named this Model 5 as "Skew normal model with random walk effect".

Therefore, the linear functions of area-level auxiliary variables for all five models are:

$$\text{Model 1: } \mu_d = \beta_0 + \beta_{0d} + \beta_1 x_d$$
$$\text{Model 2: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt}$$
$$\text{Model 3: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + \beta_{1d} x_{dt}$$

$$\text{Model 4: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + (\beta_{1d} + \zeta_{1t}) x_{dt}$$
$$\text{Model 5: } \mu_{dt} = \beta_0 + \beta_d + \beta_{0d,t} + \beta_1 x_{dt}$$

The models are evaluated in Section 4. Model 1 is fitted based on 2016 survey data (direct estimates of total gross service revenue by domain) whereas Models 2 to 5 take into account 10-year (2007–2016) data. Model comparisons are carried out considering direct and model-based estimates for 2016.

4. Results

Parameter and small domain estimates for the models defined in Section 3.1 to 3.4 (Tables 3 and 4) were obtained via MCMC (*Markov chain Monte Carlo*). All results correspond to $100,000$ MCMC sweeps, after a burn-in of $50,000$ iterations. The chain was subsequently thinned by taking every 5^{th} sample value. The Gelman and Rubin (1992) statistics are less than 1.05 for all estimated coefficients and fitted models, showing convergence of chains. Computational details of how to implement MCMC estimation procedure and corresponding Winbugs code are displayed in the Appendix. It also contains the full conditionals of the model described by the equations in (1) as in Ferraz and Moura (2012).

The auxiliary information, such as number of employees, total wages and number of establishments, were obtained from the business register used as the BASSS sampling frame. Model selection procedures showed that simultaneous inclusion of those variables was not adequate since they are highly correlated. Taking into account economic analysis, total wages was chosen as the only explanatory variable for the small area estimation models.

Both response (total gross service revenue) and auxiliary variables (total wages) are expressed in millions of Brazilian currency (Reais-R$). The estimated *wages* coefficients are positive, as anticipated, since the total revenue per domain might increase with the investment in the labor factor. The estimates of the asymmetry coefficient are positive for all models in accordance with the usual pattern of economic data (positively asymmetrical distribution). Nevertheless, the estimated values of this coefficient in Models 2 to 5 are about half of the value in Model 1.

When estimates are compared, the highlight is the variance reduction of the domain random effect on the intercept in the presence of random effects on the slope in Models 3 or 4. Also, the domain random effect on the intercept in Model 2 is considerably greater (4.884) than the time random effect (0.267). Similarly, in Models 3 and 4, the domain random effects have higher coefficients than the estimates of time random effects. In addition, the posterior mean for the intercept parameter in Model 5 exceeds more than twice the estimated values for other models.

The noticeable reduction of the intercept domain random effect variance from Model 1 to Model 3 suggests the need for a domain random effect on the slope indicating that the relation between direct estimates and auxiliary variables is not the same for all domains.

Table 3. Summary of hyperparameters' posterior distributions – Models 1, 2 and 5 - domain and time random effects on the intercept

Parameter	Model 1				Model 2				Model 5			
	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%
β_0	3.404	1.592	0.334	6.619	3.085	0.460	2.207	4.007	8.185	2.783	2.434	13.510
β_1	1.953	0.159	1.660	2.300	2.379	0.075	2.234	2.526	2.561	0.099	2.372	2.760
λ	9.174	5.195	2.922	22.390	3.505	0.319	2.912	4.161	4.424	0.486	3.555	5.445
σ_0	5.452	1.589	2.370	8.799	4.884	0.721	3.666	6.473	2.442	1.920	0.109	6.936
σ_{ζ_0}	-	-	-	-	0.267	0.225	0.008	0.836	-	-	-	-
σ_{ξ_0}	-	-	-	-	-	-	-	-	2.411	0.344	1.758	3.111
a_ϕ	0.321	0.044	0.240	0.414	0.302	0.013	0.277	0.329	0.304	0.013	0.278	0.331
b_ϕ	14.267	4.334	7.193	23.990	4.210	0.430	3.411	5.096	4.054	0.416	3.288	4.914

Table 4. Summary of hyperparameters' posterior distributions – Models 3 and 4 - Domain and time random effects on the intercept and on the slope

Parameter	Model 3				Model 4			
	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%
β_0	1.757	0.469	0.902	2.737	2.365	0.519	1.422	3.432
β_1	2.959	0.174	2.619	3.302	2.707	0.182	2.351	3.073
λ	4.066	0.369	3.379	4.838	4.489	0.427	3.711	5.383
σ_0	1.583	0.446	0.765	2.498	1.846	0.456	1.003	2.800
σ_1	1.548	0.190	1.196	1.939	1.474	0.188	1.124	1.860
σ_{ζ_0}	0.598	0.367	0.040	1.435	0.222	0.205	0.007	0.760
σ_{ζ_1}	-	-	-	-	0.389	0.133	0.204	0.718
a_ϕ	0.304	0.013	0.278	0.331	0.304	0.013	0.278	0.331
b_ϕ	4.12	0.421	3.355	5.003	4.122	0.419	3.354	4.991

4.1. Model Comparison

Table 5 presents the deviance information criterion (DIC), the posterior mean of the deviance ($\bar{D}$) and the effect number of parameters (pD) for Models 1 to 5. Note that $DIC = \bar{D} + pD$, see Spiegelhalter *et al.* (2002) for further details about the meaning of these measures. Because the data are formed by the joint pairs $(y_d, \hat{\phi}_d)$, $d = 1, .., D$, all these measurements can be calculated separately and overall values, as presented in Table 5, were obtained by summation. The model with the smallest DIC should be the one that would best jointly predict a replicate data set of y_d and $\hat{\phi}_d$. It can be seen that Model 1 (with domain and time effects in the intercept) seems to fit the service revenue data better than its counterparts. However, the performance of Models 3, 4 and 5 is similar.

Table 5. Model selection – Deviance Information Criterion (DIC)

Model	*DIC*	*pD*	$\bar{D}$
Model 1	1,636.3	145.0	1,491.3
Model 2	1,705.5	103.7	1,601.8
Model 3	1,661.5	115.1	1,546.4
Model 4	1,655.9	119.9	1,536.0
Model 5	1,664.7	119.4	1,545.3

The posterior predictive p-values (Meng, 1994), given by $P(y_d^{rep} > y_d|Data)$, where y_d^{rep} is a predictive value of the observed y_d under the considered model, were also calculated for all models with 2016 data. Values around 0.5 indicate that the distributions of the replicate and the actual values are close. Figure 1 displays the boxplots of the posterior predictive p-values for all models. According to Figure 1, model 5 seems to fit best the 2016 BASSS data. Additional information on precision and bias of small domain estimates follows next to enhance the analysis.

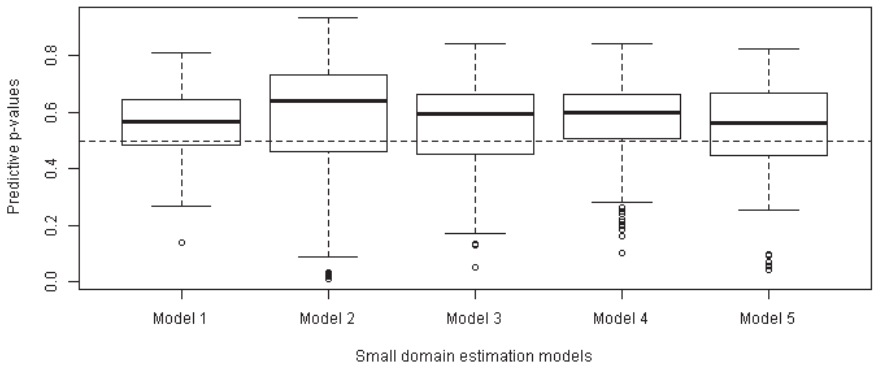

Figure 1 - Posterior predictive p-values of model-based estimates

Model-based estimates are biased, although more precise in general. The estimation procedure aims to balance the trade-off between variance and bias, producing estimates with good precision and little bias as possible. To compare the model performances, precision of estimates and relative differences of the model-based and direct estimates are presented. Figure 2 displays the improvement in coefficients of variation (CVs) for model-based estimates in relation to the direct estimates. Model 1 reduces the coefficients of variation of the small domain estimates with respect to the direct ones in 93.7% of the cases and Models 2 to 5 produce estimates with better precision for all domains. There is evidence that Model 2 provides estimates with better precision than the others. Nevertheless, considering that National Statistical Institutes may suppress the publication of estimates with CV greater than 20% as a quality threshold, Models 2,

3 and 4 do not differ in this aspect. Model 2 has 92.1% of domain estimates with CV below the threshold. This is achieved for 90.8% of the domains in the case of Models 3 and 4, but in only 81.6% of Model 5 estimates.

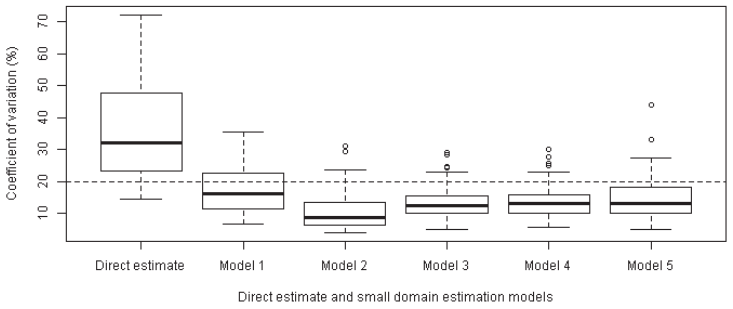

Figure 2 - Coefficients of variation of direct and model-based estimates

The analysis of the relative differences of model-based and the direct estimates $\left(\frac{Model-Direct}{Direct}\ \%\right)$ allows investigating the presence of bias. Relative differences for Models 3 and 4 that incorporate random slopes are closer to zero compared to those from models with random intercept only, as illustrated in Figure 3. In addition, the symmetric distribution for Model 5 relative differences, centered at zero, is good evidence against bias.

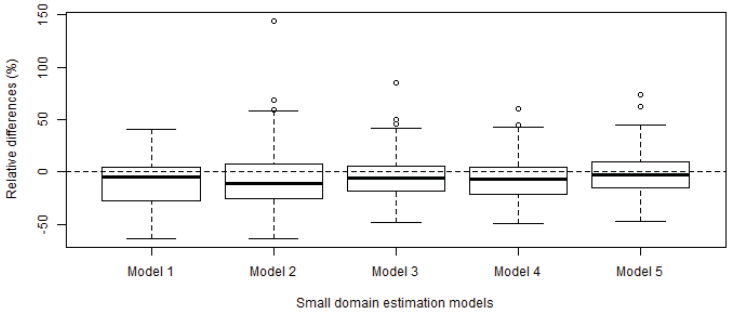

Figure 3 - Relative differences of model-based and direct estimates

The deviance information criterion and the posterior predictive p-values, together with precision and bias of small domain estimates, show that Models 3 and 4 exhibit similar performance. Results for Model 5, with comparable *DIC* value, indicate a slight improvement on the bias, but a disadvantage regarding the precision of estimates. However, Model 5 presents the best performance with respect to the predictive p-value statistics.

Considering all these measures when comparing Models 3, 4 and 5 and the quality threshold for the precision of estimates, the random walk time model (Model 5) can be recommended to produce small domain estimates for the service sector survey.

4.2. Model Diagnostics

We carried out an analysis of the standard residuals, $r_d = \frac{(y_d - \mu_d)}{\sqrt{\phi_d}}$. Since the parameters μ_d and ϕ_d are unknown, they were replaced by their respective posterior means to obtain the $\hat{r}_d$ statistics. According to Genton (2004), if y_d is skew normal distributed, the statistics $\hat{r}_d^2$ is approximately χ_1^2. Figure 4 exhibits residual plots for the application of Model 5 to the BASSS data. The histogram of the $\hat{r}_d$ statistics shows that they have positive skewness. QQ-plots and corresponding envelopes are also presented with lines for the 5^{th} percentile, the mean and the 95^{th} percentile of each observation based on the estimates of squared standard residuals, $\hat{r}_d^2$. The random variable $\hat{r}_d^2$ also enables marginal model checking and detection of outlying observations. The simulated envelope graph plotted to validate the skew-normal Model 5 indicates a few points outside the confidence bounds.

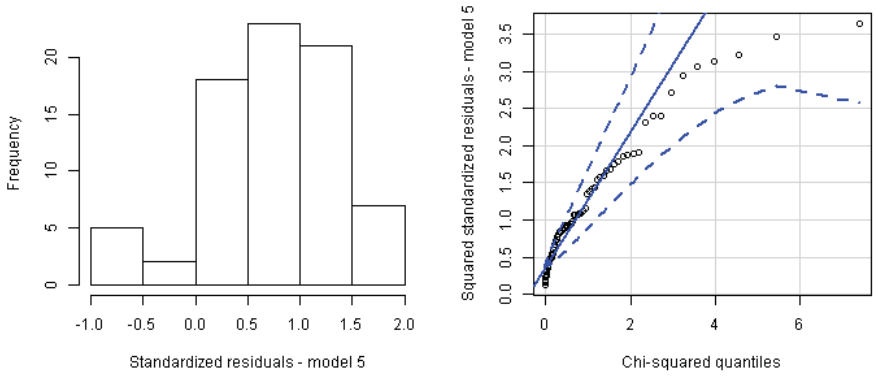

Figure 4 - Histogram and qqplot - Model 5 residuals

We also investigated the relationship between the relative differences and the domain sample sizes (Figure 5) for Model 5 estimates. Although the domain sample sizes are all very modest, with maximum value 16, large relative differences are associated with the smallest sample sizes. The negative relative difference of almost 40% for a sample size of 16 enterprises deserves mention. It refers to a domain whose economic activity is coded as 9001 - *Performing arts, shows and complementary activities*, with unstable demand since these services are not essential and, therefore, subject to income fluctuations and seasonality. Other domains with a sample size greater than 10 for which the relative differences are beyond the limits of 20% are related to economic activity 9313 - *Fitness activities,* which are constantly changing and very diverse (currently the traditional gym

centers coexist with other smaller businesses such as Pilates studios and the services of personal trainers).

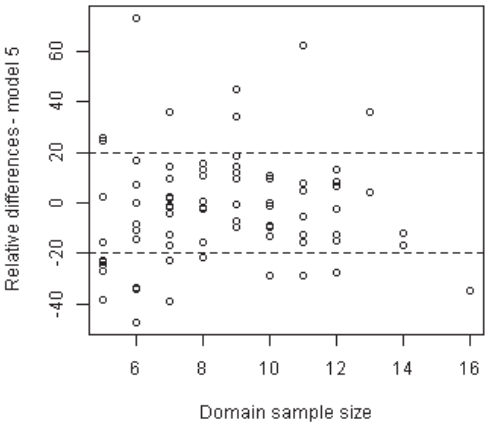

Figure 5 - Relative differences (%) by domain sample sizes - Model 5

5. Conclusions

The small domain estimation models proposed in this article showed good performance in improving the precision of estimates of *gross service revenue* by state and economic activity in the Brazilian Annual Service Sector Survey. The use of skew normal models leads to estimates with much better precision than the direct estimates. Moreover, for most domains, the coefficients of variation are below 20%, which could allow their publication. The skew normal time models with domain and time random effects on the intercept and slope exhibit promising performance. However, the presence of bias is still noted. This is better in Model 5 (Skew normal model with random walk effect), which shows some balance between estimates that exceed or not the direct estimates. Nevertheless, even considering the modest domain sample sizes, there are some domains for which values of relative differences are too high. Thus, despite the relevant gains in precision, the issue of controlling bias requires additional studies. It is important to highlight that this work was carried out using real survey data, focusing on the production of official statistics. Future work is planned to investigate new models to overcome the difficult problem of borrowing strength from domains associated with similar economic activities.

Acknowledgements

This research was supported by IBGE, the National School of Statistical Sciences (ENCE) and Federal University of Rio de Janeiro (UFRJ).

REFERENCES

ARORA, V., LAHIRI, P., (1997). On the superiority of the Bayesian methods over the BLUP in small area estimation problems. *Statistica Sinica* 7, pp. 1053–1063.

AZZALINI, A., (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

BATTESE, G. E., HARTER, R. M., FULLER, W.A., (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, Mar/1988, Vol.83, 401, pp. 28–36.

DATTA, G. S, LAHIRI, P., MAITI, T. and LU, K. L., (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, 94, pp. 1074–1082.

FAY, R. E., 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*, Vol. 74, 366, pp. 269–277.

FERRAZ, V. R. S., MOURA, F. A. S., (2012). Small area estimation using skew normal models. *Computational Statistics & Data Analysis* 56(10), pp. 2864–2874.

GELMAN, A., (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* 3, pp. 515–534.

GELMAN, A., RUBIN, D. B., (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7(4), pp. 457–472.

GENTON, M. G., (2004). *Skew-elliptical distributions and their applications: a journey beyond normality*. Chapman & Hall/CRC.

GERSHUNSKAYA, J., SAVITSKY, T. D., (2019). Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, 8, 2, pp. 181–205.

IBGE – Instituto Brasileiro de Geografia e Estatística. Pesquisa Anual de Serviços 2016. Diretoria de Pesquisas, Coordenação de Serviços e Comércio. Rio de Janeiro, 2018.

MENG, X.-L., (1994). Posterior predictive p-values. *Annals of Statistics*, 22, pp. 1142–1160.

MOURA, F. A. S., HOLT, D., (1999). Small area estimation using multilevel models. *Survey Methodology*, June 1999, 73, Vol. 25, 1, pp. 73–80, Statistics Canada, Catalogue No. 12-001.

MOURA, F. A. S., NEVES, A. F. A., SILVA, D. B. N., (2017). Small area models for skewed Brazilian business survey data. *Journal of Royal Statistical Society*, 180,Part 4, pp. 1039–1055, serie A.

NEVES, A. F. A., (2012). Small domain estimation applied to Annual Service Sector Survey 2008. *Master dissertation of National School of Statistical Sciences (originally in Portuguese)*. Rio de Janeiro, jul/2012.

NEVES, A. F. A.,SILVA, D. B. N.,CORRÊA, S. T., (2013). Small domain estimation for the Brazilian Service Sector Survey. *Estadística, 65, 185, pp. 13–37, Instituto Interamericano de Estadística)*.

RAO, J. N. K., MOLINA, I., (2015). *Small area estimation*, 2nd ed., New York, Wiley.

RAO, J. N. K., YU, M., (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. *The Canadian Journal of Statistics*, 22, 4, pp. 511–528.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., LINDE, A. V., (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, B 64, Part 4, pp. 583–639.

SUGDEN, R., SMITH, T., JONES, R., (2000). Cochran's rule for simple random sampling. *Journal of the Royal Statistical Society*: Series B 62, pp. 787–793.

COMPUTATIONAL APPENDIX

Stochastic representation

Samples from skew normal density can be generated using the following stochastic representation:

$$y_d|\eta_d,\mu_d,\lambda,\phi_d^2 \sim N(\mu_d+\phi_d\delta_d\eta_d,\phi_d^2(1-\delta_d^2)) \text{ and } \eta_d \sim HN\,(0,1),\ d=1,...,D$$

where $HN(a,b)$ denotes a half-normal distribution with location and scale parameters a and b, respectively. This stochastic representation is useful for implementing the skew normal distribution in statistical packages, such as *WinBUGS* (Spiegelhalter *et al.*, 2002).

Full conditional distributions for Model 1 as in Ferraz and Moura (2012)

$$\begin{aligned}
\pi(\sigma_0^2) &\sim IG\left[a_0+\frac{D}{2},a_0+\frac{1}{2}\sum_{d=1}^{D}(\mu_d-\mathbf{x}_d^t\beta)^2\right],\\
\pi(\beta) &\sim N\left(\left[\sigma_0^2\Omega_\beta^{-1}+\sum_{d=1}^{D}\mathbf{x}_d\mathbf{x}_d^t\right]^{-1}\sum_{d=1}^{D}\mathbf{x}_d\mu_d,\left[\sigma_0^2\Omega_\beta^{-1}+\sum_{d=1}^{D}\mathbf{x}_d\mathbf{x}_d^t\right]^{-1}\right),\\
\pi(\mu_d) &\sim N\left[\left(\frac{(y_d-\sqrt{\phi_d}w_d\delta_d)}{\phi_d(1-\delta_d^2)}+\frac{\mathbf{x}_d^t\beta}{\sigma_0^2}\right)\left(\frac{1}{\phi_d(1-\delta_d^2)}+\frac{1}{\sigma_0^2}\right)^{-1},\left(\frac{1}{\phi_d(1-\delta_d^2)}+\frac{1}{\sigma_0^2}\right)^{-1}\right],\\
\pi(W_d) &\sim N\left[\left(\frac{\delta_d(y_d-\mu_d)}{\sqrt{\phi_d}(1-\delta_d^2)}\right)\left(1+\frac{\delta_d^2}{(1-\delta_d^2)}\right)^{-1},\left(1+\frac{\delta_d^2}{(1-\delta_d^2)}\right)^{-1}\right]I_{(w_d>0)},
\end{aligned}$$

where the symbol $Y\sim IG(a,b)$ generically denotes that Y is inverse gamma distributed, that is, $Y^{-1}\sim Ga(a,b)$, and $N(a,b)I_{(w_d>0)}$ denotes a truncated normal distribution with parameters a and b.

There are no closed forms for the full conditional distributions of ϕ_d, a_ϕ, b_ϕ and λ. Nevertheless, Gibbs sampling with Metropolis-Hasting steps can be used to sample from them. The transition distribution for λ may be normal with the variance tuned for appropriate chain movements. The proposed distributions for ϕ_d, a_ϕ, b_ϕ can be gamma with the mean and variance updated with chain movement.

WinBUGS code

model
{
# Model 5
# Prior distributions
$aphi \sim dgamma(0.01, 0.01)$
$bphi \sim dgamma(0.01, 0.01)$
$beta0 \sim dnorm(0, 0.001)$
$beta1 \sim dnorm(0, 0.001)$
$sigmad0 \sim dunif(0, 100)$
$sigmadt0 \sim dunif(0, 100)$
$Lambda \sim dnorm(0, 0.01)$

Function of the hyperparameters
$sigma2d0 \leftarrow pow(sigmad0, 2)$
$taud0 \leftarrow 1/sigma2d0$
$sigma2dt0 \leftarrow pow(sigmadt0, 2)$
$taudt0 \leftarrow 1/sigma2dt0$

Model 5 description
$for(d\ in\ 1 : Ntot)\{$
$ytot[d] \sim dnorm(mus[d], taus[d])$
$mus[d] \leftarrow mu[d] + delta[d] * sqrt(1/invphi[d]) * t[d]$
$mu[d] \leftarrow beta0 + bd0[domid[d]] + bdt0[domid[d], timeid[d]]$
$+ beta1 * saltot[d]$
$delta[d] \leftarrow lambda[d]/(sqrt(1 + pow(lambda[d], 2)))$
$lambda[d] \leftarrow Lambda/sqrt(n[d])$
$t[d] \leftarrow dnorm(0, 1)I(0,)$
$thetasn[d] \leftarrow mu[d] + sqrt(2/3.14159265359) * delta[d] * sqrt(1/invphi[d])$
$as[d] \leftarrow (n[d] - 1)/2$
$bs[d] \leftarrow (n[d] - 1) * invphi[d]/2$
$phiest[d] \sim dgamma(as[d], bs[d])$
$phi[d] \leftarrow 1/invphi[d]$
$invphi[d] \sim dgamma(aphi, bphi)$
$taus[d] \leftarrow invphi[d] * (1/(1 - pow(delta[d], 2)))$
Standardized residuals
$res[d] \leftarrow (ytot[d] - mu[d]) * sqrt(invphi[d])$
Squared standardized residuals
$dest[d] \leftarrow pow((ytot[d] - mu[d]), 2) * invphi[d]$

DIC calculation
$D1[d] \leftarrow 1.837877 - log(taus[d]) + taus[d] * (pow(ytot[d] - mus[d], 2))$
$D2[d] \leftarrow -2 * as[d] * log(bs[d]) - 2 * (as[d] - 1) * log(phiest[d]) +$

$2 * bs[d] * phiest[d] + 2 * loggam(as[d])$
$D[d] \leftarrow D1[d] + D2[d]$

# Random walk
# Distributions of coefficients
}
$for(j\ in\ 1 : Ndom)\{$
$bd0[j] \sim dnorm(0, taud0)$
}
$for(l\ in\ 1 : Ndom)\{$
$for(k\ in\ 2 : Ntime)\{$
$bdt0[l,k] \sim dnorm(bdt0[l,k-1], taudt0)$
}
}
$for(l\ in\ 1 : Ndom)\{$
$bdt0[l,1] \sim dnorm(bdt0f[l], taudt0)$
}
$for(m\ in\ 1 : Ndom)\{$
$bdt0f[m] \sim dnorm(0, 0.001)$
}
# predictive p-value
$for(i\ in\ ii : ie)\{$
$ypred[i] \sim dnorm(mus[i], taus[i])$
$ppred[i] \leftarrow step(ytot[i] - ypred[i])$
}
}