REVSTAT – Statistical Journal
Volume 19, Number 3, July 2021, 399–420

AN ASYMMETRIC AREA MODEL-BASED APPROACH FOR SMALL AREA ESTIMATION APPLIED TO SURVEY DATA

Authors: Marcelo Rodríguez
– Facultad de Ciencias Básicas, Universidad Católica del Maule, Chile
mrodriguez@ucm.cl

Víctor Leiva *
– Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso, Chile
victorleivasanchez@gmail.com , www.victorleiva.cl

Mauricio Huerta
– Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso, Chile
mauricio.huerta.a@gmail.com

Camilo Lillo
– Center of Experience and Services (CES), Universidad Adolfo Ibáñez, Santiago, Chile
camilo.lillof@gmail.com

Alejandra Tapia
– Facultad de Ciencias Básicas, Universidad Católica del Maule, Chile
alejandraandreatapiasilva@gmail.com

Fabrizio Ruggeri
– CNR-IMATI, Italy
fabrizio@mi.imati.cnr.it

Received: March 2019 Revised: April 2019 Accepted: April 2019

Abstract:

- The Birnbaum–Saunders distribution is asymmetrical and has received considerable attention due to its properties and its relationship with the normal distribution. In this paper, we propose a methodology for estimating the mean of small areas based on a Birnbaum–Saunders distribution which is reparameterized in terms of its mean, similarly to the normal distribution, but in an asymmetric framework. In addition, the variance of the reparameterized Birnbaum–Saunders distribution is a function of its mean, similarly to the gamma distribution, which allows a GLM type modeling to be conducted. The Birnbaum–Saunders area model has properties that are unavailable in its competing models, as describing the mean in the original scale, unlike the existing models which employ a logarithmic transformation that reduces the test power and complicates the interpretation of results. The Birnbaum–Saunders area model can be formulated similarly as the Gaussian area model, permitting us to capture the essence of the small area estimation based on sample means and variances obtained from the areas. The methodology includes a formulation based on the Fay–Herriot model, estimation of model parameters with the maximum likelihood and Bayes empirical methods, as well as diagnostics using residuals. We illustrate the methodology with real-world survey data and compare the results with those obtained by the standard Fay–Herriot model.

Keywords:

- *empirical best linear unbiased predictor; R software; random effects; variance components.*

AMS Subject Classification:

- 62D05, 62F99, 62J99.

[*] Corresponding author.

1. INTRODUCTION

In sample surveys, it is of interest to obtain estimates for some parameters of the population from which the data are collected (Lumley and Scott, 2017 [31]). These estimates can be obtained not only for the target population, but also for sub-populations usually named small areas or domains. The small area estimation is a statistical technique used to estimate parameters in small sub-populations (Rao, 2003 [39]; Avila *et al.*, 2020 [3]), which may consist of geographical areas or socio-demographic groups, as a country, region, county, municipality or neighborhood.

Due to the high acceptance in relation to small area estimation, several models have been derived, used and analyzed. A summary of design-based small area estimation methodologies is presented in the book of Särndal *et al.* (2003) [46], whereas reviews of model-based small area estimation methodologies are provided in Ghosh and Rao (1994) [21], Rao (2003) [39], Datta (2009) [12], Lehtonen and Veijanen (2009) [25] and, more recently, in Coelho and Casimiro (2008) [9], Coelho and Pereira (2011) [10], Pereira and Coelho (2012) [36], Avila *et al.* (2020) [3] and Rueda *et al.* (2019) [43].

For small area estimation, the area model was first proposed by Fay and Herriot (1979) [19]. The Fay–Herriot (FH) model is considered as a generalization of the model formulated by Carter and Rolph (1974) [7], incorporating auxiliary variables (covariates). The FH model proposes an adaptation to the Carter-Rolph and James-Stein estimators, which was applied to income estimates in small areas during the population and housing census of the United States in 1970. The FH model assumes normality and incorporates linear regression in the context of heterogeneity of variances, so that it can be considered as a mixed model. To estimate the components of variance, different methods have been considered. Fay and Herriot (1979) [19] used weighted residual square sums and the moment method. Prasad and Rao (1990) [37] proposed an ordinary least square estimator. Datta and Lahiri (2000) [13] used the maximum likelihood (ML) and restricted maximum likelihood (REML) estimators.

When estimating means of small areas based on sampling design, there are desirable properties, such as unbiasedness and consistency, at country and region levels, but at lower levels (for example municipalities), the consistency property of the estimator is not fulfilled (Rao, 2003 [39]). Small area estimation is often based on the FH model, which allows for results in a more reliable way in order to produce statistics at lower levels than countries or regions. The FH model has good properties at low geographic levels when combining survey data with data from other sources, such as administrative or census records. In particular, the Chilean government has used the FH model since 2010 to estimate small areas (Casas-Cordero *et al.*, 2016 [8]). However, one of the drawbacks of the FH model is the assumption of normality for the response variable and random effect, because often this assumption is not fulfilled, due to asymmetry in the data distribution (Berg and Chandra, 2014 [5]). A solution to solve the problem of asymmetrical patterns present in the data is working with their log-transformations. However, data analyses performed under a wrong transformation reduces the power of the study (Huang and Qu, 2006 [22]; Dreassi *et al.*, 2014 [15]). Therefore, the research question is whether there is a gain in modifying the distributional assumption in terms of the accuracy of the estimator for producing statistics at a small area level or not.

Small-area estimation in non-normal models has been studied by few authors, even though this was postulated by Rao (2003) [39, Chap. 9] as an open problem. Fabrizi and Trivisano (2010) [18] extended the FH model assuming that the random effects follow power exponential distributions. Berg and Chandra (2014) [5] presented an empirical Bayes (EB) estimator for small area estimation based on a log-normal model and Fabrizi *et al.* (2016) [17] used the beta model for small area estimation.

The Birnbaum–Saunders (BS) distribution is asymmetrical and it has good properties (Ferreira *et al.*, 2012 [20]; Santos-Neto *et al.*, 2014 [44]; Bourguignon *et al.*, 2017 [6]). Statistical modeling based on the BS distribution has received much attention because of its relationship with the normal distribution and other properties. Rieck and Nedelman (1991) [41] were the pioneers in deriving BS regression models, whereas Villegas *et al.* (2011) [47] extended this regression model considering mixed effects and using an EB estimator to predict the random effects. Leiva *et al.* (2014) [29] and Santos-Neto *et al.* (2016) [45] focused on a reparameterized BS (RBS) distribution to model the response with no transformations following the idea of generalized linear models (McCullagh and Nelder, 1989 [32]). This modeling approach was based on fixed effects and no studies were reported using random effects. One of the parameters of the RBS distribution is its mean, such as the normal distribution, but in an asymmetric framework. In addition, the variance of the RBS distribution is a function of its mean, such as the gamma distribution. In Balakrishnan and Kundu (2019) [4] and Leiva *et al.* (2019) [27], detailed information is reviewed for these models. However, to the best of our knowledge, no area models for small area estimation based on BS, gamma and log-normal distributions have been reported in the literature.

In small area estimation, an alternative solution to solve the problem of asymmetric data is considering generalized linear models and, in particular, the RBS distribution (Leiva *et al.*, 2014 [29]). This solution provides some advantages over the log-transformation solution. First, the mean is modeled directly, making inference straightforward and avoiding the need of re-transformations back to the original scale. Second, this solution enables us to go beyond exponential family and allows some flexibility through the choice of a link function (for example, logarithmic, inverse or logit) and a distribution for the response through its mean-variance relationship. Moreover, the use of the the RBS distribution permits us to capture the essence of the small area estimation problem based on sample means and variances obtained from the areas, because it is possible to express its precision parameter as a function of these area means and variances, such as in the normal case; see Santos-Neto *et al.* (2014) [44] and Subsection 2.2 for more details about this important aspect. Therefore, the RBS distribution seems to be a good alternative to the FH type models for small area estimation.

The main objective of this work is to estimate the mean of small areas based on an RBS area model. The specific objectives are: (i) to establish an algorithm for estimating parameters from an RBS area model; (ii) to propose a residual for this model, allowing the examination of the model assumptions; and (iii) to illustrate the proposed methodology with survey data and to compare its results to the standard FH model. This methodology is implemented in the R software (`www.r-project.org` and R Core Team, 2016 [38]).

The paper is organized as follows. In Section 2, we present a background about the standard FH structure and a modeling approach based on the RBS distribution. Section 3 proposes the new RBS area model and its corresponding estimation, inference and residual analysis for its diagnostic. In Section 4, the methodology is illustrated with unpublished Chilean survey data, comparing it to a standard methodology. Section 5 gives our conclusions about this research.

2. BACKGROUND

In this section, we provide some preliminaries aspects related to the standard FH model and RBS regression.

2.1. The Fay–Herriot model

Fay and Herriot (1979) [19] proposed their model to improve the accuracy of the estimator $Y_i = \widehat{\theta}_i$ based on the sampling design (direct estimator) used to infer on the true small area mean θ_i, for $i = 1, ..., m$, where m is the number of areas. The FH model has a hierarchical structure consisting of the following two levels:

$$\begin{aligned} &\text{Level 1. Sampling model: } Y_i|\theta_i \overset{\text{IND}}{\sim} \mathrm{N}(\theta_i, \psi_i), \text{ for } i = 1, ..., m, \\ &\text{Level 2. Linking model: } \quad \theta_i \overset{\text{IND}}{\sim} \mathrm{N}(\underline{x}_i^\top \underline{\beta}, \sigma^2), \text{ for } i = 1, ..., m, \end{aligned} \tag{2.1}$$

where “IND” denotes “independent”, ψ_i corresponds to the variance of the sampling error, $\underline{x}_i = (1, x_{1i}, ..., x_{(p-1)i})$ are the values of $p-1$ covariates for the area i, $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^\top$ is a vector of unknown regression parameters, and σ^2 is the unknown variance of the area random effect, both to be estimated. Note that Level 1 describes the variability of the direct estimator $\widehat{\theta}_i$ of θ_i attributed to the sampling, whereas Level 2 links θ_i to the vector of $p-1$ known area covariates (Jiang and Lahiri, 2006 [23]; Li and Lahiri, 2010 [30]). Mixing the components of both models at Levels 1 and 2, we get the linear mixed model

$$Y_i|\theta_i = \underline{x}_i^\top \underline{\beta} + b_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{IND}}{\sim} \mathrm{N}(0, \psi_i), \quad i = 1, ..., m, \tag{2.2}$$

where $b_i \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$ are independent and identically distributed (IID) area random effects with unknown σ^2 to be estimated from the data, whereas $\varepsilon_i \overset{\text{IND}}{\sim} \mathrm{N}(0, \psi_i)$ are the sampling errors with known variances ψ_i. Furthermore, it is assumed that b_i and ε_i are independent random variables.

We want to estimate/predict the small area mean $\theta_i = \underline{x}_i^\top \underline{\beta} + b_i$, for $i = 1, ..., m$, and to obtain an uncertainty measurement related to this estimation/prediction. Considering the model defined in (2.2), the best predictor (BP) of θ_i (Rao and Molina, 2015 [40]), which minimizes the mean squared error, may be formulated as a weighted average of the direct estimator $\widehat{\theta}_i$ and the regression-synthetic estimator $\underline{x}_i^\top \underline{\beta}$ (Rao and Molina, 2015 [40]), expressed as

$$\widehat{\theta}_i^{\text{BP}} = (1 - B_i)\widehat{\theta}_i + B_i\, \underline{x}_i^\top \underline{\beta}, \quad i = 1, ..., m, \tag{2.3}$$

with the weight $0 < B_i < 1$ defined as $B_i = \psi_i/(\sigma^2 + \psi_i)$. Observe that $(1 - B_i)$ is function of the variance ratio σ^2/ψ_i and measures the uncertainty when θ_i is estimated in relation to the total variance $\sigma^2 + \psi_i$ (Rao and Molina, 2015 [40]). In addition, the parameter σ^2 is a homogeneity measure of the areas after accounting for the values $\underline{x}_i$ of covariates. If σ^2 is known, $\underline{\beta}$ may be approximated using the standard weighted least square estimator $\underline{\widetilde{\beta}}$ (Mert, 2015 [33]). Hence, by replacing it in (2.3), we obtain the best linear unbiased prediction (BLUP) of θ_i (Rao and Molina, 2015 [40]) by

$$\widehat{\theta}_i^{\,\mathrm{BLUP}} = (1 - B_i)\widehat{\theta}_i + B_i\, \underline{x}_i^\top \underline{\widetilde{\beta}}, \quad i = 1, ..., m, \tag{2.4}$$

where

$$\underline{\widetilde{\beta}} = \frac{\sum\limits_{i=1}^{m} \underline{x}_i \widehat{\theta}_i / (\sigma^2 + \psi_i)}{\sum\limits_{i=1}^{m} \underline{x}_i \underline{x}_i^\top / (\sigma^2 + \psi_i)}. \tag{2.5}$$

The BLUP of θ_i defined by (2.4) depends on σ^2 through of $\underline{\widetilde{\beta}}$, which is unknown in practice. From (2.4), we get the empirical best linear unbiased predictor (EBLUP) of θ_i as

$$\widehat{\theta}_i^{\,\mathrm{EBLUP}} = (1 - \widehat{B}_i)\widehat{\theta}_i + \widehat{B}_i\, \underline{x}_i^\top \underline{\widetilde{\beta}}, \tag{2.6}$$

where $\widehat{B}_i$ is the estimate of $B_i = \psi_i/(\sigma^2 + \psi_i)$ when σ^2 is replaced by an estimator $\widehat{\sigma}^{\,2}$, and $\underline{\widetilde{\beta}}$ is given in (2.5). Note that the model defined in (2.2) may be rewritten as matrix by

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{I}_m \underline{b} + \underline{\varepsilon}, \tag{2.7}$$

where $\underline{Y} = (Y_1, ..., Y_m)^\top$, with $Y_i = \widehat{\theta}_i$, for $i = 1, ..., m$, $\boldsymbol{X} = (\underline{x}_1, ..., \underline{x}_m)^\top$ is of full rank, $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, $\underline{\beta}$ is given below (2.1), $\underline{b} = (b_1, ..., b_m)^\top$ and $\underline{\varepsilon} = (\varepsilon_1, ..., \varepsilon_m)^\top$. Furthermore, $\underline{b}$ and $\underline{\varepsilon}$ are independently distributed with $\underline{b} \sim \mathrm{N}_m(\underline{0}_{m\times 1}, \boldsymbol{G})$, $\underline{\varepsilon} \sim \mathrm{N}_m(\underline{0}_{m\times 1}, \boldsymbol{R})$, where $\underline{0}_{m\times 1}$ is $m \times 1$ vector of zeros, $\boldsymbol{G} = \sigma^2 \boldsymbol{I}_m$ and $\boldsymbol{R}$ is a diagonal matrix defined as $\boldsymbol{R} = \mathrm{diag}\{\psi_1, ..., \psi_m\}$. The model defined in (2.7) is a particular case of a linear mixed model with its variance-covariance matrix assuming the form $\boldsymbol{V} = \boldsymbol{G} + \boldsymbol{R}$ (Datta *et al.*, 2005 [14]).

Observe that the EBLUP given in (2.6) depends on $\widehat{\sigma}^{\,2}$, with several methods being proposed in the literature for doing this estimation (Fay and Herriot, 1979 [19]; Prasad and Rao, 1990 [37]). The ML method has been widely used in small area estimation (Jiang and Lahiri, 2006 [23]; Rao and Molina, 2015 [40]), with Datta and Lahiri (2000) [13] using it in the context of the FH model. In this case, the log-likelihood function takes the form

$$\ell(\sigma^2, \underline{\beta}; \underline{y}) = c - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\underline{y} - \boldsymbol{X}\underline{\beta})^\top \boldsymbol{V}^{-1}(\underline{y} - \boldsymbol{X}\underline{\beta}), \tag{2.8}$$

where c is a constant that is independent of σ^2 and $\underline{y}$ is the observed value of $\underline{Y}$. By taking derivatives of (2.8) with respect to $\underline{\beta}$ and σ^2, we obtain

$$\frac{\partial \ell(\sigma^2, \underline{\beta}; \underline{y})}{\partial \underline{\beta}} = \boldsymbol{X}^\top \boldsymbol{V}^{-1}\underline{y} - \boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}\underline{\beta}, \tag{2.9}$$

$$\frac{\partial \ell(\sigma^2, \underline{\beta}; \underline{y})}{\partial \sigma^2} = \frac{1}{2}(\underline{y} - \boldsymbol{X}\underline{\beta})^\top \boldsymbol{V}^{-2}(\underline{y} - \boldsymbol{X}\underline{\beta}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}), \tag{2.10}$$

where $\mathrm{tr}(A)$ is the trace of the matrix $\boldsymbol{A}$. Thus, equating (2.9) and (2.10) to zero, and solving them simultaneously with respect to σ^2 and $\underline{\beta}$, we generate the corresponding ML estimators.

2.2. Birnbaum–Saunders statistical modeling

The BS distribution can be parameterized in terms of its mean μ and precision δ from its original parameterization by $\alpha = \sqrt{2/\delta}$ and $\beta = \delta\,\mu/(\delta+1)$ (Leiva, 2016 [26]). Thus, we have $\delta = 2/\alpha^2$ and $\mu = \beta\,(1+\alpha^2/2)$, where $\delta > 0$ and $\mu > 0$ (Santos-Neto *et al.*, 2016 [45]). Hence, if $Y \sim \mathrm{RBS}(\mu,\delta)$, its probability density function (PDF) is given by

$$f(y;\mu,\delta) = \frac{\exp(\delta/2)\,\sqrt{\delta+1}}{4\sqrt{\pi\mu}\,y^{3/2}}\left(y+\frac{\delta\mu}{\delta+1}\right)\exp\left(-\frac{\delta}{4}\left(\frac{(\delta+1)y}{\delta\mu}+\frac{\delta\mu}{(\delta+1)y}\right)\right),\ y>0. \tag{2.11}$$

The RBS PDF defined in (2.11) has diverse shapes as μ changes, when δ is fixed, and similarly as δ changes when μ is fixed. Note that the μ controls the scale of the RBS distribution but it is also its mean, which may be proved because $b\,Y \sim \mathrm{RBS}(b\mu,\delta)$, with $b>0$. Notice that the parameter δ controls the shape of the RBS distribution, making it more platykurtic as δ increases. In addition, the RBS variance decreases when δ increases, converging to 5.0, as δ approaches zero, doing it to be a precision parameter, as mentioned. For more details about the graphical plots and shape analysis of the RBS distribution, see Leiva *et al.* (2014) [29], Balakrishnan and Kundu (2019) [4] and Leiva *et al.* (2019) [27].

Note that the random variables Y and Z with RBS and standard normal distributions, respectively, are related by

$$\begin{aligned} Y &= \frac{\delta\,\mu}{\delta+1}\left(\frac{Z}{\sqrt{2\,\delta}}+\sqrt{\left(\frac{Z}{\sqrt{2\,\delta}}\right)^2+1}\right)^2, \\ Z &= \sqrt{\frac{\delta}{2}}\left(\sqrt{\frac{(\delta+1)\,Y}{\mu\,\delta}}-\sqrt{\frac{\mu\,\delta}{(\delta+1)\,Y}}\right). \end{aligned} \tag{2.12}$$

Thus, from (2.12), the cumulative distribution function (CDF) and the quantile function (QF) of $Y \sim \mathrm{RBS}(\mu,\delta)$ are defined respectively as

$$\begin{aligned} F(y;\mu,\delta) &= \Phi\left(\sqrt{\frac{\delta}{2}}\left(\sqrt{\frac{(\delta+1)\,y}{\mu\,\delta}}-\sqrt{\frac{\mu\,\delta}{(\delta+1)\,y}}\right)\right), \quad y>0, \\ y(q;\mu,\delta) &= F^{-1}(q) = \frac{\delta\,\mu}{\delta+1}\left(\frac{z(q)}{\sqrt{2\,\delta}}+\sqrt{\left(\frac{z(q)}{\sqrt{2\,\delta}}\right)^2+1}\right)^2, \quad 0<q<1, \end{aligned} \tag{2.13}$$

where Φ and z are the standard normal CDF and QF, respectively, whereas F^{-1} is the inverse function of the RBS CDF. The mean and variance of $Y \sim \mathrm{RBS}(\mu,\delta)$ are given by $\mathrm{E}[Y]=\mu$ and $\mathrm{Var}[Y]=\psi=\mu^2(2\delta+5)/(\delta+1)^2$, respectively. Note the similarity of the variances of the RBS and gamma distributions, which allows the RBS distribution to model data analogously as in generalized linear models (Leiva *et al.*, 2014 [29]). Note also that, as mentioned, the RBS distribution has the mean as one of its parameters, which is an advantage on the gamma distribution. Note that, in small area estimation, one has available the sample mean and variance of each area, which is a natural aspect under normality. However, in the case of the RBS distribution, it is characterized by the mean (as in the normal case) but also by a precision parameter δ, which is different from the variance of the normal case. Santos-Neto *et al.* (2014) [44] proposed a moment estimator of δ through

$$\widehat{\delta} = \frac{\overline{Y}-S^2+\sqrt{\overline{Y}^4+3\overline{Y}^2S^2}}{S^2}, \tag{2.14}$$

where $\overline{Y}$ and S^2 represent the mean and sample variance of the random variable Y, respectively. Thus, (2.14) allows us to see the problem under the RBS perspective such as the normal framework.

Rieck and Nedelman (1991) [41] defined that if $Y \sim \mathrm{BS}(\alpha, \beta)$, then $Z = \log(Y)$ follows a logarithmic BS distribution with shape parameter α and location parameter $\gamma = \log(\beta) \in \mathbb{R}$. In this regression model, the original response must be transformed to a logarithmic scale. Thus, although in this scale the mean $\gamma = \log(\beta)$ is modeled, in the natural scale $\beta = \exp(\gamma)$ is modeled, which in the BS case corresponds to the median. Leiva *et al.* (2014) [29] introduced a new approach for BS modeling, generalizing the existing works on the topic. In the estimation process, they considered $Y_1, ..., Y_m$ as independent $\mathrm{RBS}(\mu_i, \delta)$ distributed random variables, for $i = 1, ..., m$. Then, the authors defined a statistical model based on the systematic component $\mu_i = g^{-1}(\underline{x}_i^\top \underline{\beta})$, where g^{-1} is the inverse function of the link function g, $\underline{\beta}$ is a vector of unknown parameters to be estimated, and $\underline{x}_i$ represents the values of the covariates. For the vector of parameters $(\underline{\beta}^\top, \delta)^\top$, simplifying the notation according to $\ell(\underline{\beta}, \delta; \underline{y}) = \ell(\underline{\beta}, \delta)$, $\ell_i(\mu_i, \delta; y_i) = \ell_i(\mu_i, \delta)$, and by using this same simplified notation from now on, the log-likelihood function of the model is given by $\ell(\underline{\beta}, \delta) = \sum_{i=1}^m \ell_i(\mu_i, \delta)$, where

$$\ell_i(\mu_i, \delta) = \frac{\delta}{2} - \frac{\log(16\pi)}{2} - \frac{1}{2} \log\left(\frac{(\delta+1)y_i^3\mu_i}{(\delta y_i + y_i + \delta\mu_i)^2}\right) - \frac{y_i(\delta+1)}{4\mu_i} - \frac{\delta^2\mu_i}{4(\delta+1)y_i}.$$

The score functions with first derivatives of β_l, for $l = 0, 1..., p-1$, and δ are respectively given by $\dot{\ell}_{\beta_l} = \partial\ell(\underline{\beta}, \delta)/\partial\beta_l$ and $\dot{\ell}_\delta = \partial\ell(\underline{\beta}, \delta)/\partial\delta$. Thus, the score vector is $\underline{\dot{\ell}}_{\beta,\delta} = (\dot{\ell}_{\underline{\beta}}^\top, \dot{\ell}_\delta)^\top$; see details in Leiva *et al.* (2014) [29]. To estimate the model parameters by the ML method, the equation $\underline{\dot{\ell}}_{\beta,\delta} = \underline{0}_{p\times 1}$ must be solved. However, no closed-form expressions for these estimates are available. Then, an iterative approach is needed, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm; see details in Nocedal and Wright (1999) [35]. This iterative approach is used for solving unconstrained non-linear optimization problems, belonging to the class of quasi-Newton methods.

3. THE NEW STATISTICAL MODEL

In this section, we propose a methodology based on the FH model and the RBS regression model. The methodology considers the formulation of the new RBS area model, the estimation algorithm and inference for the population mean, as well as a residual analysis for model diagnostics. The standard FH model defined in (2.1) assumes normality for random effects and errors. In this case, the EB estimator and the EBLUP coincide. Note that the distribution of the direct domain mean estimator comes from the survey design, which from design-based theory is known to be approximately normal (for large enough samples). The normal approximation is not necessarily good in small areas with very small sample sizes. We consider the RBS distribution to model small area mean, whereas the random effect distribution is also assumed RBS for computational and theoretical convenience. When non-normality is assumed in the response or in the random effects, Rao (2003) [39] proposed to use the EB estimator.

3.1. Formulation

Such as in the standard model defined in (2.1), the proposed model consists of the two following levels:

$$\begin{aligned}&\text{Level 1. Sampling model: } Y_i|\theta_i \overset{\text{IND}}{\sim} \text{RBS}(\theta_i, \delta_i), \text{ for } i = 1, \ldots, m,\\ &\text{Level 2. Linking model: } \quad \theta_i \overset{\text{IND}}{\sim} \text{RBS}(g^{-1}(\underline{x}_i^\top \underline{\beta}), \kappa), \text{ for } i = 1, \ldots, m,\end{aligned} \tag{3.1}$$

where θ_i is the mean of the area i, g^{-1} is the inverse of the link function g, $\underline{\beta}$ and $\underline{x}_i$ are as defined in (2.1), whereas κ is the unknown precision parameter of the area random effect to be estimated. Note that δ_i depends on known variances ψ_i of the area i which are related according to the results proposed by Santos-Neto *et al.* (2014) [44], from where the empirical relationship is given in (2.14). Therefore, from this relationship, we have

$$\delta_i = \frac{\theta_i - \psi_i + \sqrt{\theta_i^4 + 3\theta_i^2\psi_i}}{\psi_i}, \quad i = 1, \ldots, m. \tag{3.2}$$

Thus, from (3.2), we put the model proposed in (3.1) in a small area framework.

The proposed BS area models have properties that are unavailable in the models of this type existing in the literature. Specifically, the BS area models considered in this work allow us to describe the mean of the data in their original scale, unlike the existing models, which employ a logarithmic transformation of the data, provoking a possible reduction of the power of the study and difficulties of interpretation. In addition, these BS area models can be formulated in a similar form as the normal area models, permitting us to capture the essence of the small area estimation problem based on sample means and variances obtained from the areas.

3.2. EB estimation and quadrature methods

We consider the EB approach to estimate the small area mean. First, by considering the PDF given in (2.11), we obtain the marginal PDF from the conditional (sampling model) and prior (linking model) distributions. Second, we estimate the parameters $\underline{\beta}$ and κ based on the corresponding marginal likelihood function. Third, we obtain the posterior distribution by plugging it in the estimated value of $\underline{\lambda} = (\underline{\beta}^\top, \kappa)^\top$. Fourth, we find the EB estimator of the conditional expectation of a small area mean given the observed data with respect to the RBS area model. In order to calculate this expected value, we use the posterior distribution presented in (3.13). The EB approach described above is detailed in Algorithm 1.

Algorithm 1 – Empirical Bayes approach

```
1: Establish the conditional PDF of $Y_i$ given $\theta_i$, denoted by $f(y_i|\theta_i)$, for $i = 1, ..., m$.
2: Indicate the prior distribution $\pi(\theta_i; \underline{\lambda})$, for $i = 1, ..., m$.
3: Obtain the marginal PDF
       $m(y_i; \underline{\lambda}) = \int_{\mathrm{R}_{\theta_i}} f(y_i|\theta_i)\pi(\theta_i; \underline{\lambda})\, \mathrm{d}\theta_i, \quad i = 1, ..., m,$
   recalling that $\mathrm{R}_{\theta_i}$ is the parameter space of $\theta_i$.
4: Estimate the model parameter $\underline{\lambda}$ by maximizing the marginal likelihood function
       $L(\underline{\lambda}) = \prod_{i=1}^{m} \int_{\mathrm{R}_{\theta_i}} f(y_i|\theta_i)\pi(\theta_i; \underline{\lambda})\, \mathrm{d}\theta_i.$
5: Calculate the posterior distribution
       $\pi(\theta_i|y_i; \widehat{\underline{\lambda}}) = \frac{f(y_i|\theta_i)\pi(\theta_i; \widehat{\underline{\lambda}})}{\int_{\mathrm{R}_{\theta_i}} f(y_i|\theta_i)\pi(\theta_i; \widehat{\underline{\lambda}})\, \mathrm{d}\theta_i}, \quad i = 1, ..., m,$
   to make inferences about $\theta_i$, where $\widehat{\underline{\lambda}}$ is an estimator of $\underline{\lambda}$.
6: Determine the EB estimator of $\theta_i$ using
       $\widetilde{\theta}_i^{\mathrm{EB}} = \mathrm{E}(\theta_i|y_i; \widehat{\underline{\lambda}}) = \frac{\int_{\mathrm{R}_{\theta_i}} \theta_i f(y_i|\theta_i)\pi(\theta_i; \widehat{\underline{\lambda}})\, \mathrm{d}\theta_i}{\int_{\mathrm{R}_{\theta_i}} f(y_i|\theta_i)\pi(\theta_i; \widehat{\underline{\lambda}})\, \mathrm{d}\theta_i}, \quad i = 1, ..., m.$
```

The conditional PDF (sampling model), for $i = 1, ..., m$, is given by

$$f(y_i|\theta_i) = \frac{\exp(\delta_i/2)\sqrt{\delta_i + 1}}{4\sqrt{\pi\theta_i}\, y_i^{3/2}} \left(y_i + \frac{\delta_i\theta_i}{\delta_i + 1}\right) \exp\left(-\frac{\delta_i}{4}\left(\frac{y_i(\delta_i + 1)}{\delta_i\theta_i} + \frac{\delta_i\theta_i}{y_i(\delta_i + 1)}\right)\right), \tag{3.3}$$

whereas the prior distribution, for $i = 1, ..., m$, is defined as

$$\pi(\theta_i; \underline{\lambda}) = \frac{\exp(\kappa/2)\sqrt{\kappa + 1}}{4\sqrt{\pi\, g^{-1}(\underline{x}_i^\top \underline{\beta})\, \theta_i^{3/2}}} \left(\theta_i + \frac{\kappa g^{-1}(\underline{x}_i^\top \underline{\beta})}{\kappa + 1}\right) \exp\left(-\frac{\kappa}{4}\left(\frac{\theta_i(\kappa + 1)}{\kappa g^{-1}(\underline{x}_i^\top \underline{\beta})} + \frac{\kappa g^{-1}(\underline{x}_i^\top \underline{\beta})}{\theta_i(\kappa + 1)}\right)\right). \tag{3.4}$$

Based on (3.3) and (3.4), the marginal PDF is obtained as

$$m(y_i; \underline{\lambda}) = \int_0^\infty f(y_i|\theta_i)\pi(\theta_i; \underline{\lambda})\, \mathrm{d}\theta_i, \quad i = 1, ..., m. \tag{3.5}$$

In order to calculate the integral given in (3.5), a Gaussian quadrature can be used. A quadrature rule is an approximation of the definite integral of a function, usually stated as a weighted sum of values at specified points within the domain of integration, which is conventionally taken as $[-1, 1]$. Thus, this rule may be stated as

$$\int_{-1}^{1} f(u)\, \mathrm{d}u = \sum_{j=1}^{n} w_j f(u_j). \tag{3.6}$$

Observe that the Gaussian quadrature given in (3.6) only produces good results if the function f is well approximated by a polynomial function within the range $[-1, 1]$. Then, the integration problem presented in (3.5) can be expressed in a more general way by introducing a positive weight function ω into the integrand, and allowing an interval other than $[-1, 1]$. In this way, the problem reduces to calculating

$$\int_a^b \omega(u)\, f(u)\, \mathrm{d}u, \tag{3.7}$$

for some choices of a, b and ω. Note that if $a = -1$, $b = 1$ and $\omega(u) = 1$, the integral given in (3.7) is the same as that given in (3.6). Some particular cases of the Gaussian quadrature are presented in Table 1.

Table 1: Intervals and forms for $\omega(u)$ of some Gaussian quadratures corresponding to the indicated orthogonal polynomial.

Interval	$\omega(u)$	Orthogonal polynomial
$[-1, 1]$	1	Legendre
$(-1, 1)$	$(1-u)^\alpha(1+u)^\beta, \quad \alpha, \beta > -1$	Jacobi
$(-1, 1)$	$1/\sqrt{1-u^2}$	Chebyshev
$[0, \infty)$	$\exp(-u)$	Laguerre
$(-\infty, \infty)$	$\exp(-u^2)$	Hermite

Note that the Gauss–Laguerre (GL) quadrature is an extension of the Gaussian quadrature method over the interval $[0, \infty)$ to approximate the integral obtained in (3.5) (Abramowitz and Stegun, 1972 [1]). Therefore, we approximate the marginal PDF presented in (3.5) by the GL quadrature by means of

$$m(y_i; \underline{\beta}, \kappa) = \sum_{j=1}^{n} w_j f(y_i|\theta_{ij})\pi(\theta_{ij}; \underline{\lambda}) \exp(\theta_{ij}), \quad i = 1, ..., m, \tag{3.8}$$

where n is the number of quadrature points, m is the number of areas, θ_{ij} is the jth root of the Laguerre polynomial in the area i given by

$$L_n(\theta_{ij}) = \sum_{r=0}^{n} \binom{n}{r} \frac{(-1)^r}{r!} \theta_{ij}^r,$$

and the weight w_j is given by

$$w_j = \frac{\theta_{ij}}{(n+1)^2 (L_{n+1}(\theta_{ij}))^2}, \quad i = 1, ..., m, \quad j = 1, ..., n.$$

3.3. ML estimation and Fisher information

Once the marginal PDF presented in (3.5) is approximated by the GL quadrature, we can approximate the corresponding likelihood function to estimate the parameters of the model defined in (3.1) with the ML method. Recalling that $\underline{\lambda} = (\underline{\beta}^\top, \kappa)^\top$, the marginal likelihood function is given by

$$L(\underline{\lambda}) = \prod_{i=1}^{m} m(y_i; \underline{\lambda}).$$

Therefore, the corresponding log-likelihood function approximated by the GL quadrature is given by

$$\ell(\underline{\lambda}) = \sum_{i=1}^{m} \log\left(\sum_{j=1}^{n} w_j f(y_i|\theta_{ij})\pi(\theta_{ij}; \underline{\lambda}) \exp(\theta_{ij})\right). \tag{3.9}$$

The respective score vector, obtained by differentiating (3.9) with respect to $\underline{\lambda}$, is established as

$$\underline{\dot{\ell}}(\underline{\lambda}) = \frac{\partial \ell(\underline{\lambda})}{\partial \underline{\lambda}} = (\underline{\dot{\ell}}_{\underline{\beta}}(\underline{\lambda})^\top, \dot{\ell}_\kappa(\underline{\lambda}))^\top.$$

The ML estimates of $\underline{\beta}$ and κ, $\widehat{\underline{\beta}}$ and $\widehat{\kappa}$ namely, respectively, are the solution to the system of equations given by $\underline{\dot{\ell}}_{\underline{\beta}}(\underline{\lambda}) = \underline{0}_{p\times 1}$ and $\dot{\ell}_\kappa(\underline{\lambda}) = 0$. Since the corresponding ML estimates cannot be expressed in a closed form, we compute them by maximizing the log-likelihood function defined in (3.9) numerically with the BFGS algorithm. As starting values, the estimates obtained under an RBS regression model can be considered.

The second derivatives of $\ell(\underline{\lambda})$ defined in (3.9), with respect to $\underline{\beta}$ and κ, are expressed as

$$\frac{\partial^2 \ell(\underline{\lambda})}{\partial \beta_l \partial \beta_k}, \frac{\partial^2 \ell(\underline{\lambda})}{\partial \beta_l \partial \kappa}, \frac{\partial^2 \ell(\underline{\lambda})}{\partial \kappa^2}, \quad l = 0, 1, ..., p-1.$$

Consequently, the corresponding Hessian matrix is given by

$$\ddot{\boldsymbol{\ell}}(\underline{\lambda}) = \begin{pmatrix} \dfrac{\partial^2 \ell(\underline{\lambda})}{\partial \underline{\beta} \partial \underline{\beta}^\top} & \dfrac{\partial^2 \ell(\underline{\lambda})}{\partial \underline{\beta} \partial \kappa} \\ \dfrac{\partial^2 \ell(\underline{\lambda})}{\partial \kappa \partial \underline{\beta}^\top} & \dfrac{\partial^2 \ell(\underline{\lambda})}{\partial \kappa^2} \end{pmatrix}.$$

In addition, the expected Fisher information matrix is obtained as

$$\boldsymbol{K}(\underline{\lambda}) = -\mathrm{E}[\ddot{\boldsymbol{\ell}}(\underline{\lambda})]. \tag{3.10}$$

3.4. Inference

Regularity conditions (see Cox and Hinkley, 1974 [11]) must be fulfilled for an RBS area model if its parameters are within the parameter space. Then, the ML estimator $\widehat{\underline{\lambda}}$ is consistent and follows an asymptotic joint distribution, which is normal with asymptotic mean $\underline{\lambda}$, and an asymptotic variance-covariance matrix $\boldsymbol{\Sigma}(\underline{\lambda})$. Thus, as $m \to \infty$ and recalling that $\underline{\lambda} = (\underline{\beta}^\top, \kappa)^\top$, we have

$$\sqrt{n}\,(\widehat{\underline{\lambda}} - \underline{\lambda}) \quad \overset{\mathrm{D}}{\to} \quad \mathrm{N}_{p+1}(\underline{0}_{(p+1)\times 1}, \boldsymbol{\Sigma}(\underline{\lambda})), \tag{3.11}$$

where $\overset{\mathrm{D}}{\to}$ denotes convergence in distribution. Note that if $\boldsymbol{J}(\underline{\lambda}) = \lim_{n\to\infty}(1/n)\boldsymbol{K}(\underline{\lambda})$ exists and is non-singular, with $\boldsymbol{K}(\underline{\lambda})$ being the expected Fisher information matrix given in (3.10), then $\boldsymbol{\Sigma}(\underline{\lambda}) = \boldsymbol{J}(\underline{\lambda})^{-1}$. The diagonal elements of $\boldsymbol{K}(\underline{\lambda})^{-1}$, $k_{ll}^{-1}(\underline{\lambda})$ namely, may be used for approximating the corresponding asymptotic standard errors (SEs), that is, by using

$$\mathrm{SE}[\widehat{\underline{\lambda}}_l] = \sqrt{k_{ll}^{-1}(\underline{\lambda})}, \quad l = 1, ..., p+1. \tag{3.12}$$

Note that $\widehat{\boldsymbol{K}}(\underline{\lambda})^{-1} = \boldsymbol{K}(\widehat{\underline{\lambda}})^{-1}$ is a consistent estimator of $\boldsymbol{\Sigma}(\underline{\lambda})$ and then the associated asymptotic SEs given in (3.12) may be estimated as $\widehat{\mathrm{SE}}[\widehat{\underline{\lambda}}_l] = (k_{ll}^{-1}(\widehat{\underline{\lambda}}))^{1/2}$, for $l = 1, ..., p+1$. Asymptotic inference on parameters can be conducted using (3.11) and (3.12).

3.5. Estimating the small area mean and bootstrapping

To estimate a small area mean, we use the posterior PDF evaluated at the ML estimates given by

$$\pi(\theta_i | y_i; \widehat{\underline{\beta}}, \widehat{\kappa}) = \frac{f(y_i|\theta_i)\pi(\theta_i; \widehat{\underline{\beta}}, \widehat{\kappa})}{m(y_i; \widehat{\underline{\beta}}, \widehat{\kappa})}, \quad i = 1, ..., m, \tag{3.13}$$

where $m(y_i; \widehat{\underline{\beta}}, \widehat{\kappa})$ is presented in (3.8), and $\widehat{\underline{\beta}}, \widehat{\kappa}$ are the corresponding ML estimates. Therefore, the EB estimator for the mean of an RBS area model, based on the GL quadrature, is given by

$$\widetilde{\theta}_i^{\mathrm{EB}} = \mathrm{E}(\theta_i | y_i; \widehat{\underline{\beta}}, \widehat{\kappa}) = \frac{\sum_{j=1}^n w_j \theta_{ij} f(y_i|\theta_{ij})\pi(\theta_{ij}; \widehat{\underline{\beta}}, \widehat{\kappa}) \exp(\theta_{ij})}{\sum_{j=1}^n w_j f(y_i|\theta_{ij})\pi(\theta_{ij}; \widehat{\underline{\beta}}, \widehat{\kappa}) \exp(\theta_{ij})}, \quad i = 1, ..., m. \tag{3.14}$$

Suppose that we have a random sample from an unknown distribution function F, and we want to make statistical inference about a parameter θ_i, for $i = 1, ..., m$. Bootstrapping is a non-parametric approach which relies upon the assumption that the current sample is representative of the population, and therefore, the empirical CDF $\widehat{F}$ is a non-parametric estimate of the population CDF F. From the sample, the statistic of interest, $\widetilde{\theta}_i^{\mathrm{EB}}$ namely, can be calculated as an empirical estimate of the true parameter. To measure the accuracy of the estimator, a bootstrapped SE, defined as

$$\mathrm{SE}(\widetilde{\theta}_i^{\mathrm{EB}}) = \sqrt{\mathrm{Var}(\widetilde{\theta}_i^{\mathrm{EB}})}, \quad i = 1, ..., m,$$

can be calculated; see Algorithm 2.

Algorithm 2 – Bootstrap standard error

```
1: Collect a random sample of size $m$ with replacement (bootstrap sample) from a matrix of data with $m$ rows corresponding to the areas and three columns related to the response $Y_i = \widehat{\theta}_i$, which is based on the sampling design used to estimate the true small area mean $\theta_i$, the variance of the sampling error $\psi_i$, and the covariates $\underline{x}_i$, for $i = 1, ..., m$.
2: Fit an RBS area model with the bootstrap sample of Step 1 and compute the statistic of interest $\widetilde{\theta}_i^{\mathrm{EB}}$, for $i = 1, ..., m$.
3: Repeat Steps 1–2 a large number of times (for example, $\mathrm{B} = 10,000$) and compute B bootstrap values of $\widetilde{\theta}_i^{\mathrm{EB}}$, which forms its empirical sampling distribution.
4: Calculate the sample standard deviation (SD) of the B bootstrap values of $\widetilde{\theta}_i^{\mathrm{EB}}$, which allows us to obtain the bootstrap SE of $\widetilde{\theta}_i^{\mathrm{EB}}$, for $i = 1, ..., m$.
```

3.6. Model selection

Models are often compared using selection measures as the log-likelihood function or Akaike information (AIC) and Bayesian information (BIC) criteria. Note that AIC and BIC are defined as

$$\mathrm{AIC} = -2\ell(\widehat{\underline{\lambda}}) + 2(p+1), \quad \mathrm{BIC} = -2\ell(\widehat{\underline{\lambda}}) + (p+1)\log(m), \tag{3.15}$$

where ℓ is the corresponding log-likelihood function given in (3.9), $p+1$ is the number of parameters and m the number of areas. AIC and BIC correspond to the log-likelihood function plus a component penalizing such a function, as the model has more parameters making it more complex. A model with a smaller AIC or BIC is better than another competing model (Ferreira *et al.*, 2012 [20]).

3.7. Diagnostic analysis

Residuals are frequently used to validate the assumptions of statistical models and may also be employed as tools for model selection. Based on Nobre and da Motta-Singer (2007) [34], we define a conditional residual which follows a standard normal distribution and accommodates the extra source of variability present in linear mixed models as $r_i^{(\mathrm{C})} = y_i - \widetilde{\theta}_i^{\mathrm{EB}}$, where $\widetilde{\theta}_i^{\mathrm{EB}}$ is given in (3.14) and y_i is an observed value of Y_i. We consider the randomized quantile (RQ) residual proposed by Dunn and Smyth (1996) [16], which is useful for asymmetric distributions. We use an index plot of the conditional RQ residual to verify homoscedasticity, whereas the distributional assumption is analyzed by simulated envelopes (Atkinson, 1985 [2]). For the RBS area model proposed in this work, the conditional RQ residual is defined as

$$r_i^{\mathrm{RQ(C)}} = \phi^{-1}(F(y_i; \widetilde{\theta}_i^{\mathrm{EB}}, \widehat{\kappa})) \quad i = 1, ..., m, \tag{3.16}$$

where F is the RBS CDF defined in (2.13). As F is continuous, then $F(Y_i)$ is uniformly distributed on the unit interval. In order to verify the normality of the conditional RQ residual based on the RBS area model, we utilize a theoretical quantile versus empirical quantiles (QQ) plot with simulated envelopes proposed by Atkinson (1985) [2]; see Algorithm 3.

Algorithm 3 – Goodness of fit to any distribution based on QQ plots with simulated envelopes.

```
Collect data $y_1, ..., y_m$.
Obtain the empirical quantiles $y_{i:m}$ as observed order statistics for $i = 1, ..., m$ from $y_1, ..., y_m$.
Estimate the parameters of the model by $\widehat{\lambda}$ with $y_1, ..., y_m$.
Compute $w_{i:m} = (i - 0.5)/m$, for $i = 1, ..., m$.
Calculate the theoretical quantiles $t_{i:m} = F^{-1}(w_{i:m})$, where $F^{-1}$ is the inverse function of the CDF $F$.
Draw the QQ plot with points $y_{i:m}$ versus $t_{i:m}$, for $i = 1, ..., m$.
Specify an $\alpha$ level for the simulated envelopes.
Generate $s$ samples of size $m$ from a distribution with CDF $F$ and estimated parameters $\widehat{\lambda}$.
Construct envelopes with limits given by $l_i = y_{i:m}(\alpha/2)$ and $u_i = y_{1:m}(1 - \alpha/2)$ for $i = 1, ..., m$.
Establish that the assumed distribution is adequate if all the points are inside of the envelope, otherwise it is not adequate.
```

4. SURVEY DATA ANALYSYS

In this section, we provide an illustrative example with a Chilean survey data set for analysis of service quality. Also, we compare the results obtained with the proposed methodology to a standard methodology based on the normal distribution.

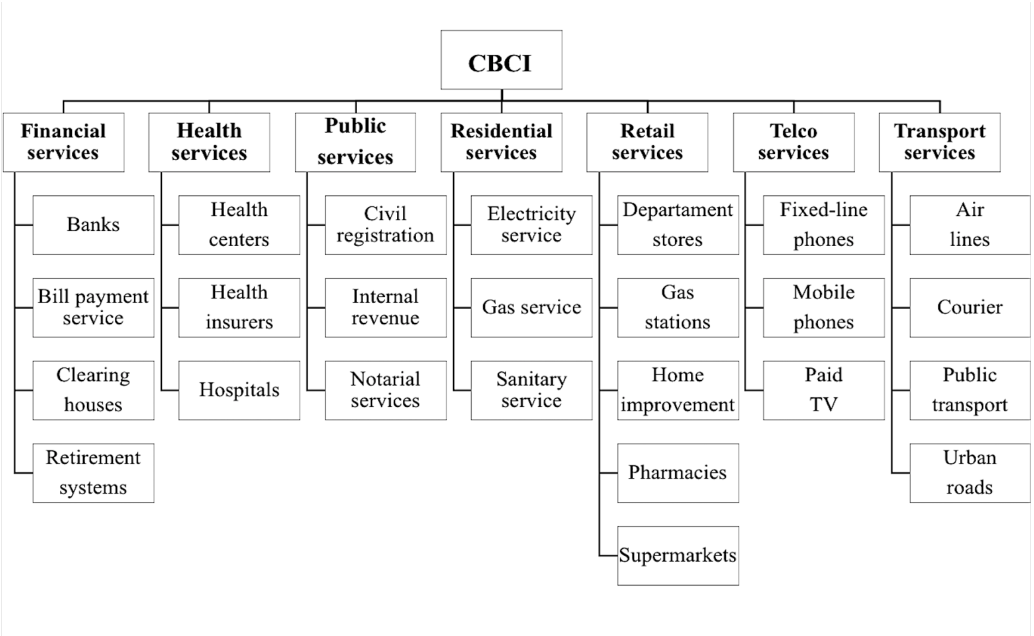

Figure 1: Structure of Chilean industries and sectors used to calculate the CBCI in 2017.

4.1. The data set

The data set under analysis was collected between January-2017 and November-2017 in 34 of 52 municipalities located at the Metropolitan region of Chile. In this data set, the response is the Chilean business confidence index (CBCI). This index is built from a sample survey which measures the confidence of customers towards the service provided by diverse companies. The CBCI is calculated by the Center of Experiences and Services (CES) of the Adolfo Ibáñez University (UAI), CES-UAI in short; see `http://www.ces-uai.cl` and more details of the CBCI in Leiva *et al.* (2018) [28]. Figure 1 shows the industrial sectors that allow us to estimate the CBCI. In this study, we consider as covariate the urban life quality index (ULQI) which allows us to model the CBCI. This covariate is obtained from the Institute of Urban and Territorial Studies of the Pontifical Catholic University of Chile (`http://fadeu.uc.cl`). The data set used in this illustration is presented in Table 2.

Table 2: CBCI (with variance and size sample) and UQLI values for the indicated municipality.

Municipality ID	$Y_i\mid\theta_i$	ψ_i	n_i	x_i
1. Pedro Aguirre Cerda (PC)	30.11	83.93	382	26.45
2. Conchalí (CO)	30.32	81.32	508	30.74
3. Quinta Normal (QN)	31.17	82.77	401	30.18
4. Lo Espejo (LE)	31.49	82.69	416	24.11
5. Cerro Navia (CN)	31.80	82.34	522	26.98
6. La Granja (LG)	32.23	78.28	453	33.98
7. Renca (RN)	32.63	83.67	472	36.42
8. Independencia (IN)	34.41	80.64	529	30.05
9. Estación Central (EC)	34.81	81.91	497	33.41
10. Lo Prado (LP)	34.81	83.05	451	30.09
11. San Ramón (SR)	35.63	84.88	394	35.53
12. Quilicura (QU)	37.13	83.31	505	39.70
13. El Bosque (EB)	37.25	80.58	502	28.10
14. Pudahuel (PU)	37.28	80.74	566	36.27
15. Puente Alto (PA)	37.87	79.54	676	36.92
16. Huechuraba (HU)	38.46	78.78	559	37.26
17. La Pintana (LA)	38.99	79.32	477	24.29
18. San Joaquín (SJ)	39.18	79.05	462	38.29
19. La Cisterna (LC)	39.23	80.12	418	32.89
20. Recoleta (RE)	40.00	79.11	520	32.36
21. Cerrillos (CE)	42.25	79.10	426	32.65
22. San Miguel (SM)	42.66	78.59	511	43.42
23. Maipú (MP)	43.50	78.39	1016	46.43
24. San Bernardo (SB)	43.91	76.56	608	28.93
25. Santiago (SA)	44.00	78.14	759	40.55
26. Peñalolen (PE)	48.54	75.99	789	38.83
27. La Florida (LF)	49.22	74.69	963	38.95
28. Macul (MA)	49.50	79.59	605	47.87
29. La Reina (LR)	51.82	74.49	716	52.45
30. Ñuñoa (NU)	52.14	73.89	980	54.27
31. Lo Barnechea (LB)	56.08	73.62	658	57.67
32. Vitacura (VI)	65.60	72.21	643	57.93
33. Providencia (PR)	71.10	68.81	928	59.96
34. Las Condes (LN)	73.60	72.58	1099	63.61

4.2. Exploratory data analysis

Table 3 provides a descriptive summary of the CBCI in the different municipalities of the Chilean Metropolitan region, which includes $\overline{y}$, median (MD), SD, coefficients of variation (CV) of skewness (CS) and of kurtosis (CK), as well as the minimum ($y_{(1)}$) and maximum ($y_{(m)}$) values. Figure 2 presents the histogram, adjusted box-plot and standard box-plot of the CBCI, as well as the scatter-plot between CBCI and UQLI. Figure 3 displays the map of the municipalities (with their abbreviations detailed in Table 3) located in the Chilean Metropolitan region with their corresponding CBCI colored in gray according to an intensity related to the value of this index.

Table 3: Descriptive statistics for the CBCI in municipalities of the Chilean Metropolitan region.

$y_{(1)}$	MD	$\overline{y}$	$y_{(m)}$	SD	CV	CS	CK
30.11	39.09	42.32	73.6	11.12	26.27	1.36	4.33

Based on Figure 2 and Table 3, we conduct an exploratory data analysis (EDA). First, from Figure 2 (left and center), note that the CBCI follows a positive skew (asymmetric) distribution (CS $>$ 0). We use an adjusted boxplot for asymmetric data (see Rousseeuw *et al.*, 2016 [42]), from which we conclude that there are no atypical data. In addition, Figure 2 (right) presents a linear or logarithmic relationship between CBCI and UQLI. Furthermore, a non-constant variance is detected by this scatter-plot. Supported by this EDA, the RBS area model proposed in this work seems to be a good candidate to describe the data set under study.

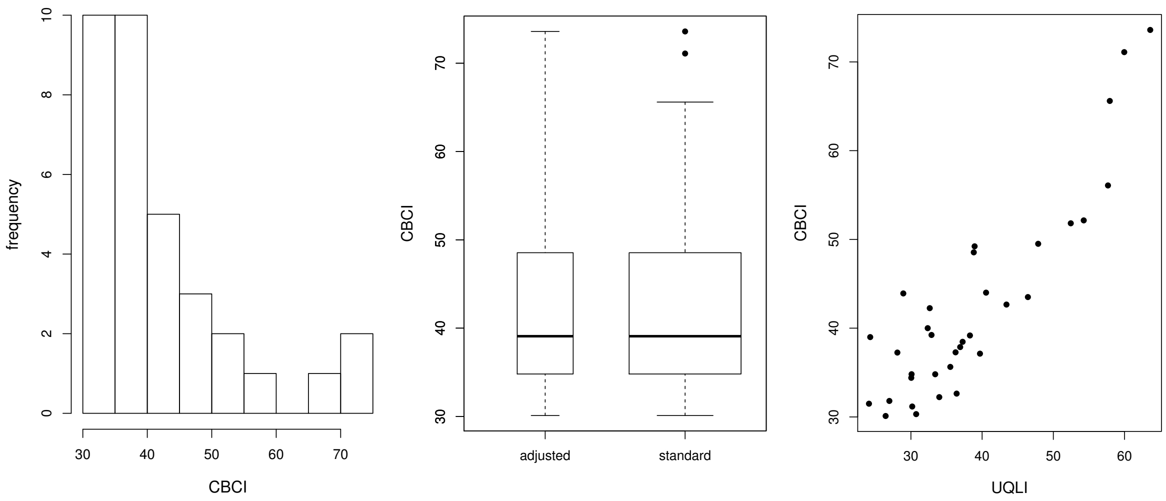

Figure 2: Histogram (left) and box-plot (center) of CBCI, and scatterplot between CBCI and UQLI (right).

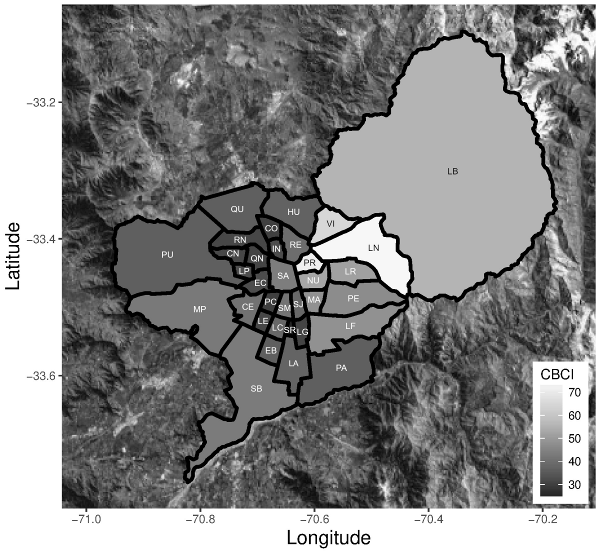

Figure 3: Map with CBCI of the indicated municipalities located at the Chilean Metropolitan region.

4.3. Modeling, estimation and inference

Based on the EDA above performed, we use the RBS area model defined in (3.1), with $i = 1, ..., 34$. In addition, δ_i can be obtained from (3.2) as $\delta_i = (y_i - \psi_i + (y_i^4 + 3y_i^2\psi_i)^{1/2})/\psi_i$, for $i = 1, ..., 34$, where ψ_i is the known variance of the municipality i. RBS area models with identity and logarithmic link functions, in short log-RBS, defined in (3.1) are compared to FH models with these same link functions. We use naive model selection tools such as AIC and BIC given in (3.15). Based on the values of AIC and BIC reported in Table 4, note that the RBS area model with logarithmic link function is the best one among the competing models to fit Chilean survey data. Once the RBS area model with logarithmic link function is selected, we estimate its parameters and the SE of the EB estimator using bootstrapping, denoted by $\widehat{\mathrm{SE}}(\widetilde{\theta}_i^{\mathrm{EB}}) = (\widehat{\mathrm{Var}}(\widetilde{\theta}_i^{\mathrm{EB}}))^{1/2}$; see Algorithm 2. Table 5 presents the values for the response variables $(Y_i|\theta_i)$, EB estimates $(\widetilde{\theta}_i^{\mathrm{EB}})$, estimated SE $(\widehat{\mathrm{SE}}(\widetilde{\theta}_i^{\mathrm{EB}}))$ and lower limit (LL) and upper limit (UP) of the 95% bootstrap confidence interval for $\widetilde{\theta}_i^{\mathrm{EB}}$.

Table 4: AIC and BIC values for the listed model and link by municipality ID with CBCI-UQLI data.

Criteria	RBS-log	RBS-identity	Normal-log	Normal-identity
$\ell(\widehat{\underline{\lambda}})$	−119.807	−129.750	−130.250	−129.750
AIC	247.614	253.601	264.501	267.501
BIC	250.194	256.188	265.079	270.081

The ML estimates of the parameters β_0, β_1 and κ of the model given in (3.1) using a logarithmic link function, with the estimated SEs in parenthesis, are: $\widehat{\beta}_0 = 4.027(0.237)$, $\widehat{\beta}_1 = 0.063(0.006)$ and $\widehat{\kappa} = 163.505(6.401)$. From this information, note that all coefficients are significant at 5% based on the normal approximation of the distribution of the ML estimators.

Table 5: Estimates, SEs and 95% confidence intervals for the area small mean based on the RBS area model with logarithm link function using CBCI and UQLI data.

ID	$\widetilde{\theta}_i^{\text{EB}}$	$\widehat{\text{SE}}(\widetilde{\theta}_i^{\text{EB}})$	LL	UL	ID	$\widetilde{\theta}_i^{\text{EB}}$	$\widehat{\text{SE}}(\widetilde{\theta}_i^{\text{EB}})$	LL	UL
PC	30.59	1.26	28.11	33.06	SJ	39.16	1.19	36.81	41.50
CO	31.38	2.39	26.68	36.08	LC	38.94	0.99	36.99	40.88
QN	32.05	1.90	28.32	35.77	RE	39.46	1.42	36.66	42.25
LE	31.68	0.86	29.99	33.36	CE	41.50	2.18	37.22	45.77
CN	32.17	0.89	30.41	33.92	SM	43.03	1.49	40.09	45.96
LG	33.17	2.64	27.98	38.35	MP	43.32	2.20	39.00	47.63
RN	33.77	3.28	27.32	40.21	SB	43.34	3.85	35.78	50.89
IN	34.74	0.74	33.27	36.20	SA	43.85	0.58	42.70	44.99
EC	35.44	1.40	32.68	38.19	PE	48.68	2.62	43.54	53.81
LP	35.09	0.64	33.83	36.34	LF	48.68	2.77	43.24	54.11
SR	36.21	1.71	32.85	39.56	MA	48.68	0.77	47.15	50.20
QU	37.13	2.48	32.26	41.99	LR	52.50	1.25	50.04	54.95
EB	36.42	1.62	33.24	39.59	NU	52.49	1.71	49.13	55.84
PU	37.00	1.27	34.49	39.50	LB	56.37	1.63	53.16	59.57
PA	37.28	1.26	34.80	39.75	VI	65.71	2.16	61.45	69.96
HU	37.80	1.13	35.57	40.02	PR	71.44	3.17	65.22	77.65
LA	37.59	3.25	31.22	43.96	LN	73.87	2.85	68.27	79.47

4.4. Diagnostics and model checking

Based on Figure 4, we evaluate the assumptions of the RBS area model with logarithm link function by an analysis of the conditional QR residual defined in (3.16) based on Chilean service quality data. This figure shows on the left an index plot of the conditional RQ residual by municipality, whereas on the right, a QQ plot with simulated envelopes for this residual is sketched. Note that outliers are not detected in these figures. In addition, since in the RBS model the variance is a function of its mean, the RBS area model manages well the problem of non-constant variance detected in the EDA. Also, note that the simulated envelopes for the conditional RQ residual verify the distributional assumption for the RBS area model and the absence of outlying observations. Therefore, based on this residual analysis and such as conjectured in our EDA, the RBS area model with logarithm link function is an excellent formulation for describing the Chilean service quality data analyzed in this study.

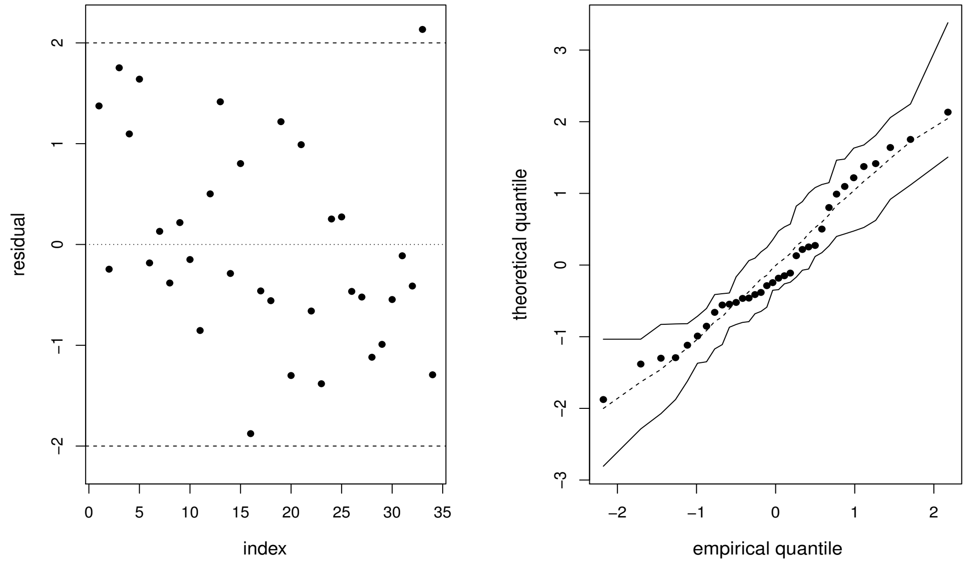

Figure 4: Index plot (left) of the conditional RQ residual and QQ plot with simulated envelopes (right) with CBCI-UQLI data.

5. CONCLUSIONS

The Birnbaum–Saunders area models proposed in this article have properties that are unavailable in the models of this type existing in the literature. Some of these properties are quite needed for describing small areas problems. Specifically, the Birnbaum–Saunders area models considered in this work allow us to describe the mean of the data in their original scale, unlike the existing models, which employ a logarithmic transformation of the data with the consequent problems. In addition, these Birnbaum–Saunders area models can be formulated in a similar form as the normal area models, permitting capturing the essence of the small area estimation problem based on sample means and variances obtained from the areas. Furthermore, the Birnbaum–Saunders area models considered in this study assume a link function, which enables for different structures present in the data. The proposed methodology allowed us to find the estimator of the small area mean based on the empirical Bayes estimator using Gaussian quadrature methods. We also considered a residual to evaluate the model assumptions and atypical data. Finally, we performed a statistical modeling for small area estimation with unpublished Chilean survey data by using the new approach proposed in the article, which have shown the applicability and scope of our proposal. The methodology introduced in this article has been implemented in the R software.

ACKNOWLEDGMENTS

The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript. This work was supported partially by FONDECYT, grant number 1200525 (V. Leiva, M. Huerta) and fellowship "Becas-Conicyt" (M. Rodríguez) from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge and Innovation.

REFERENCES

[1] Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, US.

[2] Atkinson, A. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press, Oxford, UK.

[3] Avila, J.L.; Huerta, M.; Leiva, V.; Riquelme, M. and Trujillo, L. (2020). The Fay–Herriot model in small area estimation: EM algorithm and application to official data, *REVSTAT*, **18**(5), 613–635.

[4] Balakrishnan, N. and Kundu, D. (2019). Birnbaum–Saunders distribution: A review of models, analysis and applications, *Applied Stochastic Models in Business and Industry*, **35**, 4–49.

[5] Berg, E. and Chandra, H. (2014). Small area prediction for a unit-level lognormal model, *Computational Statistics and Data Analysis*, **78**, 159–175.

[6] Bourguignon, M., Leao, J., Leiva, V., and Santos-Neto, M. (2017). The transmuted Birnbaum–Saunders distribution, *REVSTAT*, **5**, 601–628.

[7] Carter, G. and Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities, *Journal of the American Statistical Association*, **69**, 880–885.

[8] Casas-Cordero, C.; Encina, J. and Lahiri, P. (2016). *Poverty mapping for the Chilean comunas.* In "Analysis of Poverty Data by Small Area Estimation" (M. Pratesi, Ed.), vol. 20, pp. 379–404, Wiley, Chichester, UK.

[9] Coelho, P.S. and Casimiro, F. (2008). Post Enumeration Survey of the 2001 Portuguese population and housing censuses, *REVSTAT*, **6**, 231–252.

[10] Coelho, P.S. and Pereira, L.N. (2011). A spatial unit level model for small area estimation, *REVSTAT*, **9**, 155–180.

[11] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall, London, UK.

[12] Datta, G.S. (2009). *Model-based approach of small area estimation.* In "Handbook of Statistics. Sample Surveys: Inference and Analysis" (D. Pfeffermann and C.R. Rao, Eds.), vol. 29B, pp. 251–288, Elsevier, Oxford, UK.

[13] Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, **10**, 613–627.

[14] Datta, G.S.; Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**, 183–196.

[15] Dreassi, E.; Petrucci, A. and Rocco, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany, *Biometrical Journal*, **56**, 141–156.

[16] Dunn, P. and Smyth, G. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 236–244.

[17] Fabrizi, E.; Ferrante, M.R. and Trivisano, C. (2016). *Bayesian beta regression model for the estimation of poverty and inequality parameters in small area.* In "Analysis of Poverty Data by Small Area Estimation" (M. Pratesi, Ed.), pp. 299–314, Wiley, Chichester, UK.

[18] Fabrizi, E. and Trivisano, C. (2010). Robust linear mixed models for small area estimation, *Journal of Statistical Planning and Inference*, **140**, 433–443.

[19] 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.

[20] Ferreira, M.; Gomes, M.I. and Leiva, V. (2012). On an extreme value version of the Birnbaum–Saunders distribution, *REVSTAT*, **10**, 181–210.

[21] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal with discussion, *Statistical Science*, **9**, 55–76.

[22] Huang, S. and Qu, Y. (2006). The loss in power when the test of differential expression is performed under a wrong scale, *Journal of Computational Biology*, **13**, 786–797.

[23] Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation, *TEST*, **15**, 1–96.

[24] Kiaer, A.N. (1895). Observations et expériences concernant les dénombrements représentatifs, *Bulletin of the International Statistical Institute*, **9**, 176–183.

[25] Lehtonen, R. and Veijanen, A. (2009). *Design-based methods of estimation for domains and small areas.* In "Handbook of Statistics. Sample Surveys: Inference and Analysis" (D. Pefeffermann and C.R. Rao, Eds.), vol. 29B, pp. 219–249, Elsevier, Oxford, UK.

[26] Leiva, V. (2016). *The Birnbaum–Saunders Distribution*, Academic Press, New York, US.

[27] Leiva, V.; Lillo, C.; Gomes, M.I. and Ferreira, M. (2019). Discussion of Birnbaum–Saunders distribution: A review of models, analysis, and applications and a novel financial extreme value data analytics from natural disasters, *Applied Stochastic Models in Business and Industry*, **35**, 90–95.

[28] Leiva, V.; Lillo, C. and Morrás, R. (2018). *On a business confidence index and its data analytics: A Chilean case.* In "Recent Studies on Risk Analysis and Statistical Modeling" (T. Oliveira, C. Kitsos, A. Oliveira and L.M. Grilo, Eds.), pp. 67–85, Springer, Switzerland.

[29] Leiva, V.; Santos-Neto, M.; Cysneiros, F.J.A. and Barros, M. (2014). Birnbaum–Saunders statistical modelling: A new approach, *Statistical Modelling*, **14**, 21–48.

[30] Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems, *Journal of Multivariate Analysis*, **101**, 882–892.

[31] Lumley, T. and Scott, A. (2017). Fitting regression models to survey data, *Statistical Science*, **32**, 265–278.

[32] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Chapman and Hall, London, UK.

[33] Mert, Y. (2015). Generalized least squares and weighted least squares estimation methods for distributional parameters, *REVSTAT*, **13**, 263–282.

[34] Nobre, J. and Singer, J. (2007). Residual analysis for linear mixed models, *Biometrical Journal*, **49**, 863–875.

[35] Nocedal, J. and Wright, S. (1999). *Numerical Optimization*, Springer, New York, US.

[36] Pereira, L.N. and Coelho, P.S. (2012). Small area estimation using a spatio-temporal linear mixed model, *REVSTAT*, **10**, 285–308.

[37] 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.

[38] R Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[39] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley, New Yersey, US.

[40] Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley, New Yersey, US.

[41] Rieck, J.R. and Nedelman, J.R. (1991). A log-linear model for the Birnbaum–Saunders distribution, *Technometrics*, **3**, 51–60.

[42] Rousseeuw, P.J.; Croux, C.; Todorov, V.; Ruckstuhl, A.; Salibian-Barrera, M.; Verbeke, T.; Koller, M. and Maechler, M. (2016). *robustbase: Basic robust statistics*, R package version 0.92-6.

[43] Rueda, M.M.; Arcos, A.; Molina, D. and Trujillo, M. (2019). Model-assisted and model-calibrated estimation for class frequencies with ordinal outcomes, *REVSTAT*, **16**(3), 323–348.

[44] Santos-Neto, M.; Cysneiros, F.J.A.; Leiva, V. and Barros, M. (2014). A reparameterized Birnbaum–Saunders distribution and its moments, estimation and applications, *REVSTAT*, **12**, 247–272.

[45] Santos-Neto, M.; Cysneiros, F.J.A.; Leiva, V. and Barros, M. (2016). Reparameterized Birnbaum–Saunders regression models with varying precision, *Electronic Journal of Statistics*, **10**, 2825–2855.

[46] Särndal, C.E.; Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*, Springer, New York, US.

[47] Villegas, C.; Paula, G.A. and Leiva, V. (2011). Birnbaum–Saunders mixed models for censored reliability data analysis, *IEEE Transactions on Reliability*, **60**, 748–758.