*Journal of Survey Statistics and Methodology (2025) **13**, 160–196*

SMALL AREA ESTIMATION OF HOUSEHOLD ECONOMIC INDICATORS UNDER UNIT-LEVEL GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE

LORENZO MORI
MARIA ROSARIA FERRANTE*

We propose a small area estimation model based on Generalized Additive Models for Location, Scale and Shape (SAE-GAMLSS) for the estimation of household economic indicators. SAE-GAMLSS relax the exponential family distributional assumption and allow each distributional parameter to depend on covariates. A bootstrap approach to estimate the MSE is proposed. The SAE-GAMLSS estimator shows a largely better performance than the well-known Empirical Best Linear Unbiased Predictor (EBLUP) under various simulated scenarios. Per-capita consumption of Italian and foreign households in Italian regions, in urban and rural areas, is estimated using SAE-GAMLSS. Results show that the well-known Italian North–South divide does not hold for foreigners.

KEY WORDS: GAMLSS; Kurtosis; Per-capita expenditure; Skewness; Urban–rural disparity.

Dr. Lorenzo Mori is with Department of Statistical Sciences P. Fortunati, University of Bologna, Via Belle Arti 41, Bologna, Italy. Maria Rosaria Ferrante is Professor in the Department of Statistical Sciences P. Fortunati, University of Bologna, Bologna, Via Belle Arti 41, Italy.
We would like to thank ISTAT for kindly providing the HBS dataset made available for research purposes in the "Elementary Data Analysis Laboratory—ADELE (ADELE LAB.)." This study was funded by the European Union—NextGenerationEU, in the framework of the "GRINS-Growing Resilient, INclusive and Sustainable project" (PNRR—M4C2—I1.3—PE00000018—CUP J33C22002910001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.
*Address correspondence to Maria Rosaria Ferrante, Department of Statistical Sciences P. Fortunati, University of Bologna, Bologna, Via Belle Arti 41, Italy; E-mail: maria.ferrante@unibo.it

https://doi.org/10.1093/jssam/smae038 Advance access publication 16 October 2024

Statement of Significance

The small area estimation (SAE) unit-level model literature usually relies on the normal distributional assumption and almost completely focuses on modeling the location parameter. In the economic context, we often observe skewed outcome variables and heterogeneity in scale, skewness, and kurtosis. This article contributes to the SAE literature by introducing a unit-level SAE model based on Generalized Additive Models for Location, Scale and Shape (GAMLSS). SAE-GAMLSS offers flexibility by allowing several distributional assumptions for the outcome variable and the possibility to model each distributional parameter in terms of covariates. This approach leads to a further reduction in estimation variability compared to the conventional SAE models.

1. INTRODUCTION

Sample surveys are designed to provide reliable estimates for the whole reference population and, occasionally, for large sub-populations. For specific sub-populations, the sample size could be too small to produce design-based (DB) estimates with an acceptable level of variability. In this case, small area estimation (SAE) models are used to estimate sub-population parameters, where a sub-population can be defined using geographical areas and/or socio-demographic categories. SAE models try to reduce estimator variability for a certain area by "borrowing strength" from neighboring areas and from auxiliary information available from administrative or census data. They can be roughly classified into two groups, area-level and unit-level models (for an overview of SAE methods, see Rao and Molina 2015). Area-level models are used when only area-level auxiliary information, generally less subject to confidentiality restrictions concerning individual information, is available. On the other hand, unit-level models take advantage of the heterogeneity of individual auxiliary information to better explain outcome variables. See Hidiroglou and You (2016) for a comparison between area- and unit-level models.

In this paper, we focus on unit-level SAE models for the estimation of per-capita amount or rates referred to economic variables. Probability distributions of income or consumption for households and of turnover, value added, and employees for firms are typically positively skewed. Moreover, economic outcome variables, besides the heterogeneity of the location parameter, considered conditional to covariates, generally show conditional heterogeneity also on scale, skewness, and kurtosis parameters. We propose a unit-level SAE model based on Generalized Additive Models for Location, Scale and Shape

(GAMLSS). In GAMLSS, several alternative distributional assumptions can be considered for the outcome variable, and each distribution parameter, not only the location one, can be modeled in terms of covariates.

Originally defined in the form of the nested error linear regression model (Battese et al. 1988), unit-level SAE models have been gradually generalized in more refined and flexible models (Rao and Molina 2015). Ghosh et al. (1998) define Generalized Linear Mixed Models (GLMM) where the distribution of the dependent variable is assumed to belong to the exponential family. Graf et al. (2019) developed a SAE model based on the flexible Generalized Beta of the second kind (GB2) distributional assumption, Lyu et al. (2020) propose a zero-inflated lognormal model and Rojas-Perilla et al. (2020) use a specific transformation of the dependent variable (i.e., Log-Shift, Box–Cox, Dual, etc.) to reach normality. Unfortunately, in the economic field, variables cannot always be brought back to normality by transformations, and distributions do not belong to those feasible with GLMM. Molina and Martìn (2018) show that, in a small area nested error linear regression model with log transformation of the outcome variable, the individual predictors obtained through the back-transformation may be severely biased. They propose optimal predictors that reduce the bias. In addition, non-parametric (Chambers and Tzavidis 2006; Opsomer et al. 2008) and semi-parametric (Rueda and Lombardia 2012) models have been proposed in SAE literature. Additionally, in the framework of the SAE normal models, Rao and Choudhry (1995), Jiang and Nguyen (2012), and Breidenbach et al. (2018) deal with heteroskedasticity. For a more recent review of SAE methods, we refer to Tzavidis et al. (2018), Sugasawa and Kubokawa (2020), and Jiang and Rao (2020).

Unit-level SAE models based on GAMLSS, to the best of our knowledge, have never been considered in the SAE context. GAMLSS, introduced by Rigby and Stasinopoulos (2005), generalize GLM and GAM, but also GAMM and GLMM models. Some unique and interesting features of GAMLSS can be suitably and profitably spent for the estimation of economic parameters in small areas. First, in GAMLSS, the response distribution is not restricted to belonging to the exponential family, whereas over one-hundred continuous and discrete distributions can be adopted for modeling the response variable. Truncated, censored, and finite mixture versions of these distributions can also be used. A family of increasingly flexible models can be defined according to the distribution that best suits the data. This opportunity is very important in the socio-economic field, where normality, although often assumed, is rare. Second, but perhaps even more relevant, not only the location parameter but also each parameter of the outcome variable distribution can be modeled as functions of the covariates and random effects. In contrast, past SAE literature focuses heavily on modeling the location parameter. The possibility to model each parameter provides the option of borrowing strength not only by the covariates explaining the location parameter, as usual in SAE, but also from the areal heterogeneity of the remaining parameters. This feature of GAMLSS

could lead to a further reduction of the estimated Mean Square Error (MSE) via the reduction of the bias introduced by the model.

The paper is organized as follows. Section 2 introduces GAMLSS. In Section 3, we extend the GAMLSS theory to the SAE framework. The prediction of general area parameters is defined, with a parametric bootstrap proposal for MSE estimation. The proposed methods are evaluated against more common competitors using model-based (MB) and DB simulation studies (Section 4). In Section 5, we specify a SAE-GAMLSS model for estimating, based on Household Budget Survey (HBS) data, the per-capita consumption of Italian and foreigner households, in the urban and rural areas of Italian regions. Section 6 summarizes the main findings and outlines further research.

2. GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE

GAMLSS assume independent observations y_i, $i = 1, \ldots, n$, from a random variable Y, with Probability Density Function (PDF) $f(Y|\boldsymbol{\theta}_i)$, conditional on $\boldsymbol{\theta}_i^T = (\theta_{i1}, \ldots, \theta_{ik}, \ldots, \theta_{ip})$, a vector of p distribution parameters, $k = 1, \ldots, p$. Under GAMLSS, each parameter can be expressed as a function of the covariates. Rigby and Stasinopoulos (2005) define the original formulation of a GAMLSS as follows. Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be the n length vector of the response variable. Let $g_k(\cdot)$ be a known monotonic link functions relating the p distribution parameters to explanatory variables by:

$$g_k(\boldsymbol{\theta}_k) = \mathbf{X}^k\boldsymbol{\beta}^k + \sum_{m=1}^{M_k} \mathbf{Z}_m^k\boldsymbol{\gamma}_m^k, \quad k = 1, \ldots, p \quad \text{and} \quad m = 1, \ldots, M_k \qquad (1)$$

where $\boldsymbol{\theta}_k^T = (\theta_{1k}, \ldots, \theta_{nk})$ is a vector of length n, $\boldsymbol{\beta}^k = (\beta_1^k, \ldots, \beta_{\rho_k}^k)^T$ is a parameter vector of length ρ_k, that is ρ_k is the number of covariates for the k^{th} parameter, $\mathbf{X}^k$ is a known design matrix of order $n \times \rho_k$, $\mathbf{Z}_m^k$ is a fixed known $n \times q_{mk}$ design matrix, γ_m^k is a q_{mk}-dimensional random effect term, and M_k is the number of additive terms for each parameter. Each parameter k in (1) allows for a combination of different types and numbers of additive random effects terms, γ_m^k, to be incorporated easily in the model. The distributional form of $f(Y|\boldsymbol{\theta}_i)$ is not restricted to a specific family. Indeed, in GAMLSS, the exponential family distributional assumption, essential in GLMs and GAMs, is relaxed and replaced by a general distribution family, including highly skewed and/or kurtotic continuous and discrete distributions. The number of parameters p, despite not being constrained, is often assumed ≤ 4 since with these limits almost every distribution could be used in the GAMLSS framework. The first two population distribution parameters are often location (μ) and scale (σ) parameters, while the remaining ones, if any, are shape parameters, i.e., skewness (ν) and kurtosis (τ). GAMLSS allow a variety of additive

smoothing terms ($\sum_{m=1}^{M_k} \mathbf{Z}_m^k \boldsymbol{\gamma}_m^k$) as P-spline, cubic splines, random effects, non-parametric random effects, and many others. We refer to $\boldsymbol{\gamma}_m^k$ always as parametric random effects as is typical in SAE models.

The estimation of the parametric part of GAMLSS is achieved by either of two different algorithmic procedures, which maximize a penalized likelihood of the data (Rigby and Stasinopoulos 2005). The first is the Newton–Raphson algorithm (RS). The second is based on the Cole and Green algorithm (CG), which is a "transformation" of the Fisher scoring algorithm. According to Rigby and Stasinopoulos (2005), a combination of the RS and CG algorithms can be adopted as well. This combination, adopted in the following, employs the RS algorithm for the first N_{RS} iterations switching later to CG and performing it for additional N_{CG} iterations. Both algorithms, for the estimation of the distribution parameters, are performed within an outer and an inner iteration (for more detail, see: Rigby and Stasinopoulos 2005; Appendix A). To conclude, a back-fitting algorithm is used to estimate $\boldsymbol{\beta}^{\boldsymbol{k}}$ and $\boldsymbol{\gamma}_{\boldsymbol{m}}^{\boldsymbol{k}}$ parameters. Crucial to the way that additive components are fitted within the GAMLSS framework is the back-fitting algorithm. The quadratic penalties in the likelihood derive from assuming a normally distributed random effect on the linear predictor. Assume in (1) the $\boldsymbol{\gamma}_m^k$ independent normally distributed, with $\boldsymbol{\gamma}_m^k \sim N_{q_{mk}}(\mathbf{0}, [\mathbf{G}_{mk}(\boldsymbol{\lambda}_{mk})]^-)$, where $[\mathbf{G}_{mk}(\boldsymbol{\lambda}_{mk})]^-$ is a generalized inverse of a $q_{mk} \times q_{mk}$ symmetric matrix, which may depend on a vector of hyper-parameters $\boldsymbol{\lambda}_{mk}$. For fixed $\boldsymbol{\lambda}_{mk}$, the $\boldsymbol{\beta}^k$ and the $\boldsymbol{\gamma}_m^k$ are estimated by maximizing the following penalized likelihood function l_p:

$$l_p = \sum_{i=1}^{n} \log\{f(y_i|\boldsymbol{\theta}_i)\} - \frac{1}{2}\sum_{k=1}^{p}\sum_{m=1}^{M_k} (\boldsymbol{\gamma}_m^k)^T \mathbf{G}_{mk} \boldsymbol{\gamma}_m^k, \quad i = 1, \ldots, n. \tag{2}$$

As noted by Rigby and Stasinopoulos (2005), the maximization of (2) is equivalent to using an empirical Bayesian argument to obtain the maximum a posteriori estimation of both the $\boldsymbol{\beta}^k$ and the $\boldsymbol{\gamma}_m^k$ assuming normal, possibly improper, priors for the $\boldsymbol{\gamma}_m^k$.

3. GAMLSS FOR SAE

In this section, we propose a SAE model based on GAMLSS. We first focus on the specification and estimation of the model, and then on predicting the individual values of the outcome variable for the non-sampled part of the population. In SAE, a finite population Ω with N units is distributed into J sub-population $\Omega_1, \ldots, \Omega_J$, called domains or areas of size $N_1, \ldots, N_J$, where $N = \sum_{j=1}^{J} N_j$. We identify with Y_{ij}, the target variable for unit i in group j, for $i = 1, \ldots, N_j$ and $j = 1, \ldots, J$. Following the notation used from (1) on-wards, the area-specific random effects are defined as $\mathbf{Z}_j^k \gamma_j^k$. We consider only one

additive term, that is $M_k = 1, \forall k$. Following Rigby and Stasinopoulos (2005), when random effects are considered, the matrix $\mathbf{Z}$ is defined as the identity matrix and the random effects, for a series of reasoning reported in Rigby and Stasinopoulos (2005), are estimated using an Expectation–Maximization algorithm.

SAE models aim to estimate area parameters in the form of $H_j = \zeta(\boldsymbol{Y}_j)$ where $\zeta(\cdot)$ is a real measurable function and $\boldsymbol{Y}_j^T = (Y_{1j}, \ldots, Y_{ij}, \ldots, Y_{N_j,j})$. DB estimates, for an observed variable y_{ij}, of these area parameters are obtained using data from the sample s of size n drawn from the population Ω with subsample $s_j = s \cap \Omega_j$ for area j of size n_j, being $n = \sum_{j=1}^{J} n_j$. We denote by $r_j = \Omega_j - s_j$ the sample complement from the area j. Small area models need to adopt a strategy allowing specific area variation. In this context, mixed models are particularly interesting as they involve random area-specific effects, which make it possible to add between-area heterogeneity to that introduced by covariates. We specify a GAMLSS small area model, starting from (1), by considering area-specific random effect and limiting our attention to four parameter distributions:

$$\begin{cases} g_\mu(\mu_{ij}) = x_{ij}^{\mu}\beta^{\mu} + \gamma_j^{\mu} \\ g_\sigma(\sigma_{ij}) = x_{ij}^{\sigma}\beta^{\sigma} + \gamma_j^{\sigma} \\ g_\nu(\nu_{ij}) = x_{ij}^{\nu}\beta^{\nu} + \gamma_j^{\nu} \\ g_\tau(\tau_{ij}) = x_{ij}^{\tau}\beta^{\tau} + \gamma_j^{\tau} \end{cases} \tag{3}$$

where x_{ij}^k represents a generic covariate used in the k^{th} parameter of the distribution. In (3), $\gamma_j^k \overset{iid}{\sim} N(0, \boldsymbol{\Psi}_k)$ for $k = \mu, \sigma, \nu, \tau$, allows to consider differences among areas for each parameter. Random effects are assumed to be independent (Rigby et al. 2019, Ch. 10). The variance–covariance matrix $\boldsymbol{\Psi}_k$ involves the variance of the random effects σ_k^2, for independent random effects, we have $\boldsymbol{\Psi}_k = \sigma_k^2 \mathbb{I}$. In (3), the random effects referred to $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are assumed:

$$\gamma_j^{\mu} \sim N(0, \sigma_\mu^2), \quad \gamma_j^{\sigma} \sim N(0, \sigma_\sigma^2), \quad \gamma_j^{\nu} \sim N(0, \sigma_\nu^2) \quad \text{and} \quad \gamma_j^{\tau} \sim N(0, \sigma_\tau^2).$$

The distribution model $\mathscr{F}$ allows for both skewness and kurtosis in the conditional distribution of Y_{ij}:

$$Y_{ij} | \gamma_j^{\mu}, \gamma_j^{\sigma}, \gamma_j^{\nu}, \gamma_j^{\tau} \sim \mathscr{F}(Y_{ij} | \mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij}), \quad i = 1, \ldots, n_j \quad \text{and} \quad j = 1, \ldots, J. \tag{4}$$

Assuming (4) as the correct model for the population, the estimates of the parameters can be obtained by one of the methods mentioned in Section 2, that is the RS algorithm, the CG algorithm or, as done in the following, a mixture

of both. Estimates lead to $\mathscr{F}(y_{ij}|\hat{\mu}_{ij},\hat{\sigma}_{ij},\hat{\nu}_{ij},\hat{\tau}_{ij})$, $i \in s$ and $j = 1,\ldots,J$, where $i \in s$ indicate that the parameters are estimated using the sample part of the population.

In the estimation and/or in the prediction of the model (3), we operate under the assumption of a non-informative design. Although this is a frequent assumption in the unit-level model framework, we recognize that this is a weakness of the proposed model. Some studies consider informativeness in the unit-level SAE framework (Folsom et al. 1999; You and Rao 2002; Verret et al. 2015; Kim et al. 2022; Parker et al. 2023a). The mentioned studies will serve as inspiration for the development of a future research direction on the extension of the proposed model under the assumption of an informative sampling design. For a complete review of informativeness in the SAE context, see Parker et al. (2023b).

In the next section, we assume a continuous dependent variable. What follows could be generalized, provided that appropriate covariates are available, to discrete distributions with some mathematical precautions.

3.1 Prediction of General Small Area Parameters

We propose to estimate a general small area parameter $H_j = \zeta(\boldsymbol{Y}_j)$ using a Monte Carlo approximation for an empirical plug-in predictor (EPP). Denoting the vector of estimated values of $\mu_{ij},\sigma_{ij},\nu_{ij},\tau_{ij}$ and γ_j^k as $\hat{\delta}_{ij} = (\hat{\mu}_{ij},\hat{\sigma}_{ij},\hat{\nu}_{ij},\hat{\tau}_{ij},\hat{\gamma}_j^k)^T$, where the random effects can be defined for each parameter k, define a predictor of H_j thorough the following algorithm:

(1) fit the $\mathscr{F}$ model (4) to the sample data, obtaining a consistent estimator $\hat{\delta}_{ij}$ of δ_{ij};
(2) predict, for $i \in r$, the values of $\hat{\delta}_{ij}$;
(3) using the predicted values, for each $\ell = 1,\ldots,L$, with L large, generate an out-of-sample vector ${}^{(\ell)}\mathbf{y}_{ij}^r$ from $\mathscr{F}(y_{ij}^r|\hat{\mu}_{ij},\hat{\sigma}_{ij},\hat{\nu}_{ij},\hat{\tau}_{ij})$ with $i \in r$;
(4) attach the sample data $\mathbf{y}_{ij}^s$ to the generated out-of-sample data ${}^{(\ell)}\mathbf{y}_{ij}^r$ to form the area population vector ${}^{(\ell)}\mathbf{y}_{ij}^T = (({}^{(\ell)}\mathbf{y}_{ij}^r)^T,(\mathbf{y}_{ij}^s)^T)$. With ${}^{(\ell)}\mathbf{y}_{ij}^T$ is possible to calculate the target parameter $H_j^{(\ell)} = \zeta({}^{(\ell)}\mathbf{y}_j)$ where ${}^{(\ell)}\mathbf{y}_j^T = ({}^{(\ell)}y_{1j},\ldots,{}^{(\ell)}y_{ij},\ldots,{}^{(\ell)}y_{N_j,j})$.

A MC approximation of $\hat{H}_j$ is then:

$$\hat{H}_j \approx \frac{1}{L}\sum_{i=1}^{L} H_j^{(\ell)}.$$

Let us note that the procedure described is useful when the specific parameters we are estimating do not have closed-form expressions. Below, we provide the

closed-form expressions for the predictors of the area means. Under model (4), the predictor of the area mean $\bar{Y}_j = N_j^{-1} \sum_{i=1}^{N_j} Y_{ij}$ is given by:

$$\hat{\bar{Y}}_j = N_j^{-1} \left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{y}_{ij}^r \right) \tag{5}$$

where $\hat{y}_{ij}^r$ in (5) is the predicted value for the non sample unit $i \in r_j$. Note that, if $\mathscr{F}$ is a distribution with the mean expressed as a linear function of μ only and $g_\mu(\cdot)$ is the identity link function, we have:

$$\hat{\bar{y}}_j^{\,r} = \bar{X}_j \hat{\beta}^\mu + \hat{\gamma}_j^\mu \tag{6}$$

where $\bar{X}_j$ is the matrix of the area mean of the covariates and $\hat{\gamma}_j^\mu$ is the estimate of γ_j^μ. Note that, adding a shrinkage factor and assuming a normal distribution, (6) is exactly the classical EBLUP.

As noted, we opt for an EPP, a predictor commonly used in the SAE context, especially under non-normal models; see, among others, Chandra et al. (2012), Rao and Molina (2015), Chandra et al. (2018), Chandra et al. (2019), Molina and Strzalkowska-Kominiak (2020), and references therein. An alternative approach to the EPP is the Empirical Best Predictor (EBP; Molina and Rao 2010). The EBP is the conditional mean of the small area parameter given the data and it minimizes the MSE, resulting to be the most efficient predictor. However, as noted among others by Chandra et al. (2019), computing the EBP is generally not straightforward. For non-normal models, the computation of the EBP requires the solution of multiple integrals that generally do not have a closed-form (Marino et al. 2019). Jiang (2003) points out that the EBP cannot be derived in closed-form for the exponential family, and that approximating it demands significant computational resources. This additional burden, added with those already demanded by GAMLSS, makes the computation nearly infeasible. The computational effort required by the EBP has led numerous official statistics offices to use EPP. See, for example, the UK Office for National Statistics and the Australian Bureau of Statistics (Chandra et al. 2018).

3.2 Parametric Bootstrap for MSE Estimation

An important aspect of the SAE framework is the estimation of MSE defined as $MSE_j = E[(\hat{H}_j - H_j)^2]$. The Taylor linearization MSE estimators of Prasad and Rao (1990) and González-Manteiga et al. (2007) are only appropriate for specific GAMLSS models and parameters, such as the linear mixed model for estimating a mean. Due to the generality of the model and

parameters considered, a parametric bootstrap MSE estimator is proposed as follows:

(1) fit the unit-level model with one of the allowed algorithms (i.e., RS, CG, or a combination of both, see Section 2) to obtain model parameter estimators $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Psi}}_k$ for $k = \mu, \sigma, \nu, \tau$;
(2) given the estimates obtained in step 1, for each k generate a vector $\mathbf{t}_k^*$ whose elements are J independent realizations of a *N(0,1)* variable. For each parameter k, construct the bootstrap vector $\gamma^{k*} = \hat{\boldsymbol{\Psi}}_k \mathbf{t}_k^*$ and generate:

$$\begin{cases} \hat{\mu}_{ij}^* = g_\mu^{-1}(x_{ij}^\mu \hat{\beta}^\mu + \gamma_j^{\mu*}) \\ \hat{\sigma}_{ij}^* = g_\sigma^{-1}(x_{ij}^\sigma \hat{\beta}^\sigma + \gamma_j^{\sigma*}) \\ \hat{\nu}_{ij}^* = g_\nu^{-1}(x_{ij}^\nu \hat{\beta}^\nu + \gamma_j^{\nu*}) \\ \hat{\tau}_{ij}^* = g_\tau^{-1}(x_{ij}^\tau \hat{\beta}^\tau + \gamma_j^{\tau*}) \end{cases}$$

(3) for each j and each k, generate a bootstrap population of response variables from the model $\mathscr{F}(y_{ij}|\hat{\mu}_{ij}^*, \hat{\sigma}_{ij}^*, \hat{\nu}_{ij}^*, \hat{\tau}_{ij}^*)$;
(4) let $\mathbf{Y}_j^{P_*{}^T} = (y_{j1}^*, \ldots, y_{jN_j}^{P_*})$ denote the vector of generated bootstrap response variables for area j. Calculate target quantities for the bootstrap population as $H_j^* = \zeta(\mathbf{Y}_j^{P_*})$;
(5) let $\mathbf{y}_s^*$ be the vector whose elements are the generated y_{ij}^* with indices contained in the sample s. Fit the model to the bootstrap sample data $\{(y_{ij}^*, x_{ij}; i \in s)\}$ and obtain the bootstrap model parameter estimators;
(6) obtain the bootstrap estimator of H_j through the Monte Carlo approximation, denoted $\hat{H}_j^*$;
(7) repeat steps [2–6] a large number of times B. Let $H_j^*(b)$ be the true value and $\hat{H}_j^*(b)$ the estimator obtained in b^{th} replicate of the bootstrap procedure, $b = 1, \ldots, B$.

The bootstrap MSE estimator of $\hat{H}_j$ is given by:

$$\hat{MSE}_B(\hat{H}_j) = B^{-1} \sum_{b=1}^{B} \left[\hat{H}_j^*(b) - H_j^*(b)\right]^2. \tag{7}$$

3.3 Model Selection for GAMLSS

Here, we highlight some inferential aspects related to GAMLSS. We focus on model selection within the GAMLSS framework in the context of SAE, as we believe this topic deserves further attention. For a complete overview of the

inferential framework in GAMLSS, see Kneib (2013), Kneib (2020), and Wood (2020). Let $\mathscr{M} = \{\mathscr{D}, \mathscr{G}, \mathscr{T}, \mathscr{L}\}$ represent a GAMLSS as defined in (1), where the components of $\mathscr{M}$ represent: $\mathscr{D}$ the set of the possible distribution of the response variable, $\mathscr{G}$ the set of link functions, $\mathscr{T}$ the set of the terms appearing in all the predictors, and $\mathscr{L}$ the set of the smoothing hyperparameters. In GAMLSS, the model selection operates starting from $\mathscr{D}$ till $\mathscr{L}$. Here, we focus only on the choice within the sets $\mathscr{D}$ and $\mathscr{T}$ that are peculiar for SAE (for the selection referred to the other components, see Rigby et al. 2019, Ch. 11). For a parametric statistical model $\mathscr{M}$ and within a likelihood-based inferential procedure, the distribution $\mathscr{D}$ is selected by the global deviance $GDEV = -2l(\hat{\boldsymbol{\theta}})$. Let $\mathscr{M}_0$ and $\mathscr{M}_1$ be nested statistical models with fitted global deviances $GDEV_0$ and $GDEV_1$ and degrees of freedom df_0 and df_1, respectively. $\mathscr{M}_0$ and $\mathscr{M}_1$ may be compared using the generalized likelihood ratio test statistic which, under certain conditions, is asymptotically χ^2_d distributed with $d = df_0 - df_1$. To compare non-nested GAMLSS models, the generalized Akaike information criterion (GAIC) could be used to penalize over-fitting. This is obtained by adding to the fitted deviance, a penalty for each effective degree of freedom used in the model. Rigby and Stasinopoulos (2005) suggest starting with a penalty equal to $\sqrt{\log(n)}$ if $n \geq 1000$ and to use different values of the penalty with the aim of investigating the sensitivity or robustness of model selection. In short, given that GAMLSS, unlike the classical regression models, consider more than one parameter to be explained by covariates, the selection of $\mathscr{D}$ is done using GDEV instead of the deviance and there is the necessity to use a specific penalty.

Another technique that can be used to select the best distribution within the set $\mathscr{D}$ is the K-fold cross-validation. This method can be employed as a substitute for GAIC in evaluating (nested or non-nested) GAMLSS models, especially in the context of predictive modeling (Rigby et al. 2019, Ch. 11). The goodness-of-fit in the K-fold cross-validation when used with GAMLSS is evaluated with the GDEV. The model with the smallest overall value of the GDEV is the best for prediction purposes. Cross-validation is sometimes used in SAE (Pfeffermann and Correa 2012; Steorts et al. 2020; Ren et al. 2022). However, its application in the GAMLSS framework can be challenging due to well-documented computational complexities (Nti et al. 2021). At the end of this section, we suggest a strategy to use the cross-validation with a reduced computational effort integrating it with the selection within the set $\mathscr{T}$ as well.

The selection within the set $\mathscr{T}$ requires specific procedures that could be seen as an extension and a generalization of the usual ones. There are four possible procedures: (i) the criterion-based methods, based on the GAIC criterion (Rigby et al. 2019, Ch. 11), (ii) the regularization method, which relies on the ridge and lasso regression concept and that can also be extended to the GAMLSS framework, (iii) the boosting algorithm discussed by Hofner et al. (2014), emerging from the field of supervised machine learning and, nowadays, often applied as a flexible alternative to estimate and select predictor,

(iv) the dimension-reduction method, which uses a principal component analysis to choose the covariates. In particular, focusing on the criterion-based methods, Rigby and Stasinopoulos (2005) propose two different strategies: (i) the "fixed order procedure" in which the selection take place in the following order: first μ followed by σ, then ν and finally τ, (ii) the "fixed terms procedure," which forces all the distribution parameters to have the same terms, then it is not possible to use different predictors depending on the parameters.

To conclude, we would like to highlight a potential synergistic application between the mentioned techniques. To mitigate computational efforts, we suggest the following steps. First, employ the GAIC method to narrow down potential distributions. Second, for the selected distributions use the "fixed order procedure" to determine covariates for each parameter and, third, finalize the distribution selection based on K-fold cross-validation results. See Appendix B in the supplementary data online for a step-by-step definition of the model selection procedure.

4. MB AND DB SIMULATIONS

We study the characteristics and performance of small area predictors based on GAMLSS under different scenarios with MB and DB simulations. We consider as parameters of interest averages or proportions, which, for income and consumption variables, could be per-capita income or consumption, poverty rates, budget shares, and so on. We comparatively evaluate the performance of the GAMLSS estimator and of the well-known unit-level Battese–Harter–Fuller (BHF, Battese et al. 1988) model. We refer to the BHF estimator as the EBLUP. When the dependent variable is not normal, we use a BHF with Box–Cox transformation (Rojas-Perilla et al. 2020; Würz et al. 2022). MB estimators are compared with the Horvitz—Thompson direct estimator. The estimated MSE is computed by approximating the second order inclusion probability, following Molina and Marhuenda (2015). For every simulation, we define a different distribution model $\mathscr{F}$ using, for each parameter of the chosen distribution, the link function that minimizes the GAIC and a set of different covariates. The code used for the simulations is publicly available on GitHub: *saegamlss*.

4.1 MB Simulations

MB simulations are frequently used to determine estimator characteristics when it is hard to achieve analytic results on estimator properties. In the following, we perform simulations based on four different models and, eventually, different outcome parameters. We refer to them, by considering the underlying distributional assumptions, as *(A) MB Normal*, *(B) MB*

Heteroskedastic Normal, *(C) MB Log-Normal*, and *(D) MB Dagum*. In all simulations, the population Ω is fixed to 50,000 with a number of areas $J = 50$ and a population of size $N_j = 1{,}000$. The n_j units, where n_j varies among areas, are sampled based on a stratified random design without replacement. The sample size goes from a minimum of four units to a maximum of sixty-one. In all scenarios, the covariates are kept fixed and sampled from the standard normal distribution (Liu et al. 2022). All the models are chosen to reproduce realistic situations either in the type of distribution or in the single coefficients. They are chosen in order to reproduce the corresponding estimates, that is the lognormal parameter values are based on those estimated for the Italian household consumption expenditure. The GAMLSS are estimated through a mixed algorithm with a number of loops equal to $N_{RS} = 250$ and $N_{CG} = 250$ and the bootstrap for MSE estimation is performed with $B = 200$ iterations. For each simulation, results are obtained on the basis of $T = 500$ runs. The other parameters, which have to be fixed to perform the algorithms, are reported in the appropriate sections.

(A) *MB Normal*: Normal model for the estimation of the mean. A GAMLSS, where normality is assumed and only μ is defined in terms of covariates, is equivalent to a linear mixed-effects model. In this simulation, the model is defined as:

$$\mu_{ij} = \beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \gamma_j^{\mu} + \varepsilon_{ij}$$

Data are generated by assuming $\beta_0^{\mu} = 100$ and $\beta_1^{\mu} = 4$, while the value of the variance of γ and ε changes from one scenario to another as follows:

$$\textbf{Scenario 1} \quad \gamma_j^{\mu} \sim N(0, 4^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, 20^2)$$

$$\textbf{Scenario 2} \quad \gamma_j^{\mu} \sim N(0, 6^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, 22^2)$$

$$\textbf{Scenario 3} \quad \gamma_j^{\mu} \sim N(0, 8^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, 24^2).$$

We compare GAMLSS estimators obtained under the Normal distribution with EBLUP estimators expecting the same results, both referring to estimates and relative MSE.

(B) *MB Heteroskedastic Normal*: Normal model for the estimation of the mean with heteroskedastic data. In GAMLSS, both location and scale are modeled depending on covariates. We expect that the estimator could reduce the MSE compared with EBLUP MSE. As well summarized by Breidenbach et al. (2018), how to handle heteroskedasticity in SAE is still an active field of research. Our interest here is to verify if this model is able to reduce the variability and if the MSE estimator we propose in Section 3.2 is able to catch the real path of the MSE. In this simulation, we use two covariates, $\mathbf{X}_1$ and $\mathbf{X}_2$, where the $\mathbf{X}_2$ is used to add

heteroskedasticity to our data. The GAMLSS predictor is based on a normal distribution with:

$$\begin{cases} \mu_{ij} = \beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \beta_2^{\mu} x_{2ij} + \gamma_j^{\mu} + \varepsilon_{ij} \\ \sigma_{ij} = \exp(\beta_0^{\sigma} + \beta_1^{\sigma} x_{2ij} + \gamma_j^{\sigma}) \end{cases}$$

The data generation process we use is different and is based on

$$\mu_{ij} = 100 + 10x_{1ij} + 8x_{2ij} + \gamma_j^{\mu} + w_{ij}\varepsilon_{ij}$$

The term w_{ij}, useful to reproduce heteroskedasticity (Ramirez-Aldana and Naranjo 2021), is defined in terms of unit and area-level components as follows: $w_{ij} = \gamma_j^{\sigma} + 0.1x_{2ij}$ with $\gamma_j^{\mu} \sim N(0, 6^2)$ and $\varepsilon_{ij} \sim N(0, 22^2)$, while the value of the variance of γ_j^{σ} changes from one scenario to another as follows:

$$\begin{aligned} &\textbf{Scenario 1} \quad \gamma_j^{\sigma} \sim N(0, 0.8^2) \\ &\textbf{Scenario 2} \quad \gamma_j^{\sigma} \sim N(0, 1^2) \\ &\textbf{Scenario 3} \quad \gamma_j^{\sigma} \sim N(0, 1.2^2). \end{aligned}$$

We compare GAMLSS with a classical EBLUP in order to verify if our model reduces the overestimation of the MSE typical of the EBLUP.

(C) *MB Log-Normal*: lognormal model, commonly adopted in literature to model consumption expenditure (Battistin et al. 2009). We focus on the estimation of the mean where both location and scale parameters are modeled depending on covariates, that is our dependent variable $y \sim Log-Normal(\mu, \sigma)$ where both μ and σ depend on covariates and random effects. Note that in SAE, positively skewed variables are usually modeled through a normal model with transformation leading the variable to normality. The model is:

$$\begin{cases} \mu_{ij} = \exp(\beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \gamma_j^{\mu}) \\ \sigma_{ij} = \exp(\beta_0^{\sigma} + \beta_1^{\sigma} x_{2ij} + \gamma_j^{\sigma}) \end{cases}$$

Data are generated assuming $\beta_0^{\mu} = 7$, $\beta_1^{\mu} = 1$, $\beta_0^{\sigma} = -2$ and $\beta_1^{\sigma} = 0.5$, the random effects are sampled by:

$$\begin{aligned} &\textbf{Scenario 1} \quad \gamma_j^{\mu} \sim N(0, 0.6^2) \quad \text{and} \quad \gamma_j^{\sigma} \sim N(0, 0.2^2) \\ &\textbf{Scenario 2} \quad \gamma_j^{\mu} \sim N(0, 0.4^2) \quad \text{and} \quad \gamma_j^{\sigma} \sim N(0, 0.3^2) \\ &\textbf{Scenario 3} \quad \gamma_j^{\mu} \sim N(0, 0.6^2) \quad \text{and} \quad \gamma_j^{\sigma} \sim N(0, 0.4^2). \end{aligned}$$

We compare GAMLSS estimators obtained under the lognormal distribution with Box–Cox EBLUP estimators.

(D) *MB Dagum*: Dagum model, commonly used to model income distribution. We focus on the estimation of the poverty rate by modeling all three Dagum distribution parameters depending on covariates, that is our dependent variable $y \sim Dagum(\mu, \sigma, \nu)$ where all the three parameters depend on covariates and/or random effects (Rigby et al. 2019). In this simulation, unlike the previous ones, in the three scenarios, not only the value of the variance of the random effects changes but also the definition of $\boldsymbol{\sigma}$ and $\boldsymbol{\nu}$, as follows:

$$\textbf{Scenario 1}\begin{cases}\mu_{ij} = \exp(\beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \gamma_j^{\mu}), & \text{with} \quad \gamma_j^{\mu} \sim N(0, 0.15^2)\\ \sigma_{ij} = 3.4\\ \nu_{ij} = \exp(\beta_0^{\nu} + \beta_1^{\nu} x_{2ij})\end{cases}$$

$$\textbf{Scenario 2}\begin{cases}\mu_{ij} = \exp(\beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \gamma_j^{\mu}), & \text{with} \quad \gamma_j^{\mu} \sim N(0, 0.15^2)\\ \sigma_{ij} = \exp(\beta_0^{\sigma} + \beta_1^{\sigma} x_{2ij} + \gamma_j^{\sigma}), & \text{with} \quad \gamma_j^{\sigma} \sim N(0, 0.1^2)\\ \nu_{ij} = 0.6\end{cases}$$

$$\textbf{Scenario 3}\begin{cases}\mu_{ij} = \exp(\beta_0^{\mu} + \beta_1^{\mu} x_{1ij} + \gamma_j^{\mu}), & \text{with} \quad \gamma_j^{\mu} \sim N(0, 0.15^2)\\ \sigma_{ij} = \exp(\beta_0^{\sigma} + \beta_1^{\sigma} x_{2ij} + \gamma_j^{\sigma}), & \text{with} \quad \gamma_j^{\sigma} \sim N(0, 0.1^2)\\ \nu_{ij} = \exp(\beta_0^{\nu} + \beta_1^{\nu} x_{2ij}).\end{cases}$$

Data are generated assuming $\beta_0^{\mu} = 3$, $\beta_1^{\mu} = 1.5$, $\beta_0^{\sigma} = 1.2$, $\beta_1^{\sigma} = 0.1$, $\beta_0^{\nu} = -0.4$ and $\beta_1^{\nu} = 0.1$. We compare GAMLSS estimators obtained under Dagum distribution with Box–Cox EBLUP estimators.

4.2 DB Simulations

DB simulations are carried out to analyze the properties of the estimators when the data-generation process is unknown. In this case, an economic survey plays the role of a pseudo-population that is kept fixed: the Italian HBS. A second DB simulation based on the Italian Statistics on Income and Living Conditions is reported in Appendix A in the supplementary data online. For realistic simulations, we consider only covariates known in the population.

Moreover, for each model, we select only significant covariates. For the model selection within GAMLSS, we use the procedure described in Section 3.3 and detailed in Appendix B in the supplementary data online. For EBLUP, a classical Akaike step-wise procedure is used. Note that the model selection within GAMLSS involves three steps, where both distribution and covariates are selected. In contrast, for EBLUP, the parameters of the Box–Cox transformation are computed at each run of the simulation, that is, they are data-driven (Rojas-Perilla et al. 2020). The simulation assumes:

(A) *DB consumption data*: in this simulation, we refer to data collected in 2019 HBS. This survey, conducted by the Italian Statistical Institute (ISTAT), is based on a sample of approximately 18,000 households, which is representative of all Italian households. Data are collected on the basis of a two-stage sample design where the first stage refers to municipalities and the second stage to households. We target the per-capita expenditure based on equivalent consumption obtained through the Carbonaro equivalence scale (Cutillo et al. 2020). Given that the data disclosure policies of ISTAT cannot release geo-referenced small area data, that is, for municipality or provinces (ISTAT 2022), we consider administrative provinces as small areas and mimic small areas similar to provinces as follows. For each region, we set a number of clusters equal to the number of provinces and we use the amount of Waste Tax paid by each household as a variable of clustering, since the amount of this tax is decided on the basis of municipal and provincial area specific quotas. By performing a different k-means cluster analysis for each region, we are able to preserve the heterogeneity between areas obtaining a cluster per-capita equivalized mean expenditure that goes from 297.8 to 1270.9. The samples are drawn by stratified sampling with clustered areas acting as strata, and with simple random sampling of size $n_j = N_j/20$ $(j = 1, \ldots, 107)$ within each small area (for summaries of population and sample size see table 1). The model used is a lognormal distribution defined as follows:

$$\begin{cases} \mu_{ij} = \exp(\beta_0^{\mu} + \beta_1^{\mu} Age_{ij} + \beta_2^{\mu} Citizenship_{ij} + \gamma_j^{\mu}) \\ \sigma_{ij} = \exp(\beta_0^{\sigma} + \beta_1^{\sigma} Age_{ij} + \beta_2^{\sigma} Citizenship_{ij} + \gamma_j^{\sigma}) \end{cases}$$

where *Age* is the age of the respondent in 15 different classes, while *Citizenship* is a dummy variable that is 0 if the respondent is Italian, 1 if otherwise. The presented GAMLSS based on a lognormal distribution is compared with an EBLUP defined as:

$$\mu_{ij} = \beta_0^{\mu} + \beta_1^{\mu} Age_{ij} + \beta_2^{\mu} Citizenship_{ij} + \gamma_j^{\mu}$$

where normality is reached due to the Box–Cox transformation.

Table 1. Summaries of Population and Sample Size in Design-Based Simulation

(A) DB consumption data

	Min.	Q1	Median	Mean	Q3	Max.
N_j	37	198	328	400	518	1921
n_j	2	10	16	20	26	96

4.3 Measures of Performance

In order to compare results, we consider indicators evaluating both accuracy in terms of variability and bias. The following measures of performance averaged on areas are used: average relative bias (ARB) and average absolute relative bias (AARB). Let us generalize the notation used in Section 3 and define H_j as the true general area parameter in the j^{th} area and $\hat{H}_{jt}$ its estimated value in the t^{th} replication. The following measures are defined independently of the estimator considered. We use a specific superscript to distinguish, whether necessary, between MB and DB estimators.

$$ARB = \frac{1}{J}\sum_{j=1}^{J}\left\{H_j^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{H}_{jt}\right)-1\right\}\times 100,$$

$$AARB = \frac{1}{J}\sum_{j=1}^{J}\left\{\left|H_j^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\hat{H}_{jt}\right)-1\right|\right\}\times 100$$

The variability of the estimates is evaluated using the Average True MSE (ATMSE):

$$ATMSE = \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{1}{T}\sum_{t=1}^{T}(\hat{H}_{jt}-H_{jt})^2\right\}.$$

To evaluate the performance of the proposed MSE estimator, the ATMSE is compared with the Average Bootstrap MSE (ABMSE), and the Percentage Coverage Rate (PCR) is also used for evaluation:

$$ABMSE = \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{1}{T}\sum_{t=1}^{T}\hat{MSE}_B(\hat{H}_{jt})\right\} \quad \text{and}$$

$$PCR = \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{1}{T}\sum_{t=1}^{T}I(|\hat{H}_{jt}-H_{jt}|\le 1.96\sqrt{\hat{MSE}_B(\hat{H}_{jt})})\right\}\times 100.$$

where $\hat{MSE}_B(\hat{H}_{jt})$ is the bootstrap MSE defined in (7).

4.4 Discussion of Simulation Results

4.4.1 MB simulations.

The results of measures of performance for MB simulations are reported in table 2. Figure 6, available in Appendix A in the supplementary data online, shows area-specific relative bias. As expected, differences between GAMLSS and EBLUP are almost fully negligible in terms of bias when data are normal[1] (simulations *(A) MB Normal* and *(B) MB Heteroskedastic Normal*). Additionally, results are almost equal among scenarios and some negligible differences between GAMLSS and EBLUP are given by the different method for estimating parameters and random effects. The slight increase in precision measured on average by ARB and AARB is to be sought in the outer and inner iteration of the algorithm used by GAMLSS in which μ and σ are estimated simultaneously. Figure 6a in Appendix A in the supplementary data online, shows that the bias for the GAMLSS and EBLUP are almost identical for both homoskedastic and heteroskedastic data. What is important to note is the difference in the MSE estimation between simulation *(A) MB Normal* and *(B) MB Heteroskedastic Normal*. The values of ATMSE and ABMSE for GAMLSS are very similar in both the simulations and in all scenarios. Meanwhile, the values of PCR (approximately 94 percent) seem to confirm the accuracy of the proposed bootstrap estimator. In *(A) MB Normal*, results for EBLUP and for GAMLSS are almost identical. In *(B) MB Heteroskedastic Normal*, EBLUP shows an overestimation of the MSE, leading to a higher PCR. Area-specific MSE estimation of GAMLSS and EBLUP estimators is compared with the true MSE in figures 1a and b. Results for simulation *(A) MB Normal* clearly show that the MSE estimator substantially replicates the EBLUP MSE obtained by using the well-known EBLUP MSE estimator (Rao and Molina 2015). In simulation *(B) MB Heteroskedastic Normal*, the results on empirical MSE are completely different. Figure 1b shows how the MSE GAMLSS estimators correctly estimate the MSE, significantly reducing the over-estimation that is typical of the EBLUP estimators. The reduction is more marked as the variance of the random effects increases, that is, moving from the first to the third scenario. In short, the GAMLSS estimator performs better in terms of reliability compared to EBLUP and GAMLSS modeling σ is better able to estimate MSE. Through a specific area random effect used for σ, GAMLSS is able to reproduce the real variability in the data and to suitably use it in the bootstrap MSE estimation.

With regard to simulation *(C) MB Log-Normal*, from table 2 arises that the ARBs increase slightly if we switch from the direct estimator to MB ones, as expected. However, GAMLSS show, in each scenario, a lower ARB than that

1. In order to reduce the bias of $\hat{\sigma}$, we use a correction factor as suggested by Nelder (2006). This is needed to take into account the error term present in μ due to the identity link function used for this parameter.

Table 2. Measures of Performance: Model-Based Simulations

	Scenario 1			Scenario 2			Scenario 3		
	Direct	GAMLSS	EBLUP	Direct	GAMLSS	EBLUP	Direct	GAMLSS	EBLUP
(A) MB Normal									
ARB	0.04	0.11	0.11	–0.09	0.01	0.03	–0.02	0.13	0.14
AARB	3.52	2.35	2.33	3.84	2.88	2.85	4.19	3.41	3.39
ATMSE	19.91	6.51	6.53	26.34	13.46	13.56	31.18	18.30	18.70
ABMSE	–	6.06	6.02	–	13.44	13.41	–	18.24	18.32
PCR	–	92.95	92.94	–	94.00	93.86	–	94.20	94.21
(B) MB Heteroskedastic Normal									
ARB	0.01	0.08	0.11	–0.02	0.07	0.11	–0.06	0.09	0.11
AARB	3.50	1.94	2.31	3.95	2.23	2.73	4.51	2.47	3.09
ATMSE	24.90	7.72	7.70	33.99	10.63	10.65	40.12	12.66	12.64
ABMSE	–	7.95	10.25	–	10.41	13.58	–	12.48	16.73
PCR	–	91.90	93.64	–	90.84	93.43	–	96.74	93.22
(C) MB Log-Normal									
ARB	0.02	–0.06	0.41	–0.04	–0.09	0.26	–0.07	0.01	0.17
AARB	1.96	1.88	3.64	1.97	2.07	3. 73	1.95	2.07	3.63
ATMSE	38887.72	4019.78	8389.60	27898.37	3510.56	5600.96	39501.54	4625.86	12077.40
ABMSE	–	4023.38	7286.64	–	3526.45	10876.71	–	5061.31	13034.36
PCR	–	91.76	89.94	–	91.57	90.51	–	92.03	86.74

Continued

Table 2. Measures of Performance: Model-Based Simulations

	Scenario 1			Scenario 2			Scenario 3		
	Direct	GAMLSS	EBLUP	Direct	GAMLSS	EBLUP	Direct	GAMLSS	EBLUP
(D) MB Dagum									
ARB	0.13	0.32	–5.34	0.17	0.33	–6.01	–0.05	0.20	–5.83
AARB	1.80	4.54	7.23	1.81	4.86	7.98	1.81	4.59	7.34
ATMSE	0.0066	0.0004	0.001	0.0075	0.0004	0.002	0.0076	0.0005	0.001
ABMSE	–	0.0005	0.001	–	0.0005	0.0006	–	0.0005	0.001
PCR	–	92.03	86.91	–	91.75	81.86	–	92.84	90.68

NOTE.—*(A) MB Normal:* random effect and error term variances for the location parameter increase for Scenarios 1–3.
(B) MB Heteroskedastic Normal: random effect variance for the scale parameter increases for Scenarios 1–3.
(C) MB Log-Normal: random effect variances for the location and scale parameter change.
(D) MB Dagum: in Scenario 1 σ is constant, μ and ν are defined using covariates; in Scenario 2 ν is constant, μ and σ are defined using covariates; in Scenario 3, all the parameters are defined using covariates.

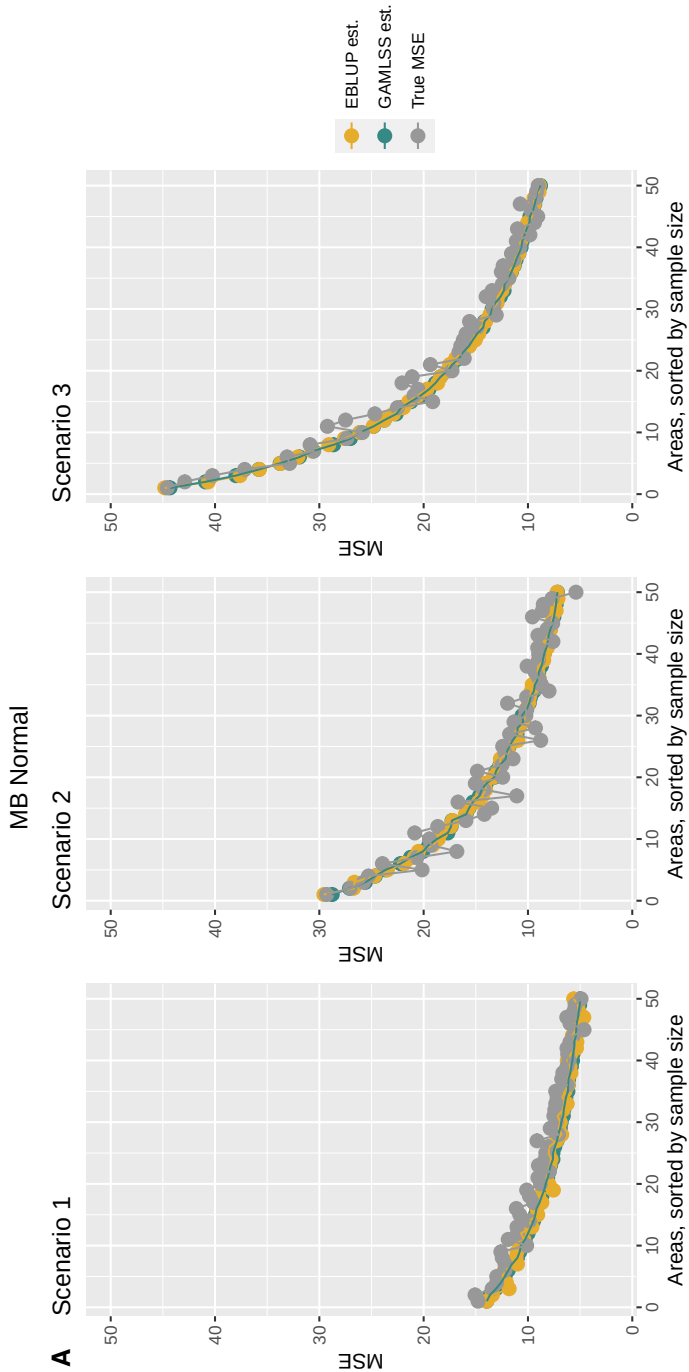

Figure 1. MSE estimation: model-based simulations.

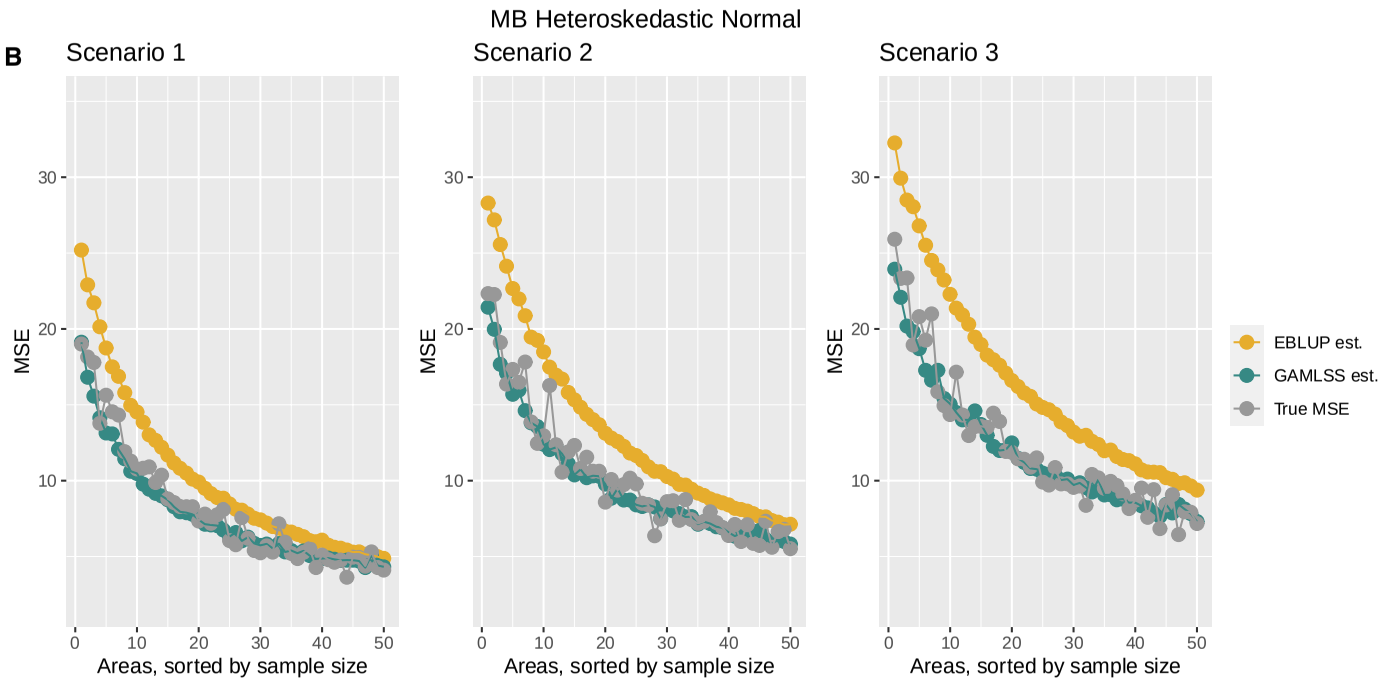

Figure 1. Continued.

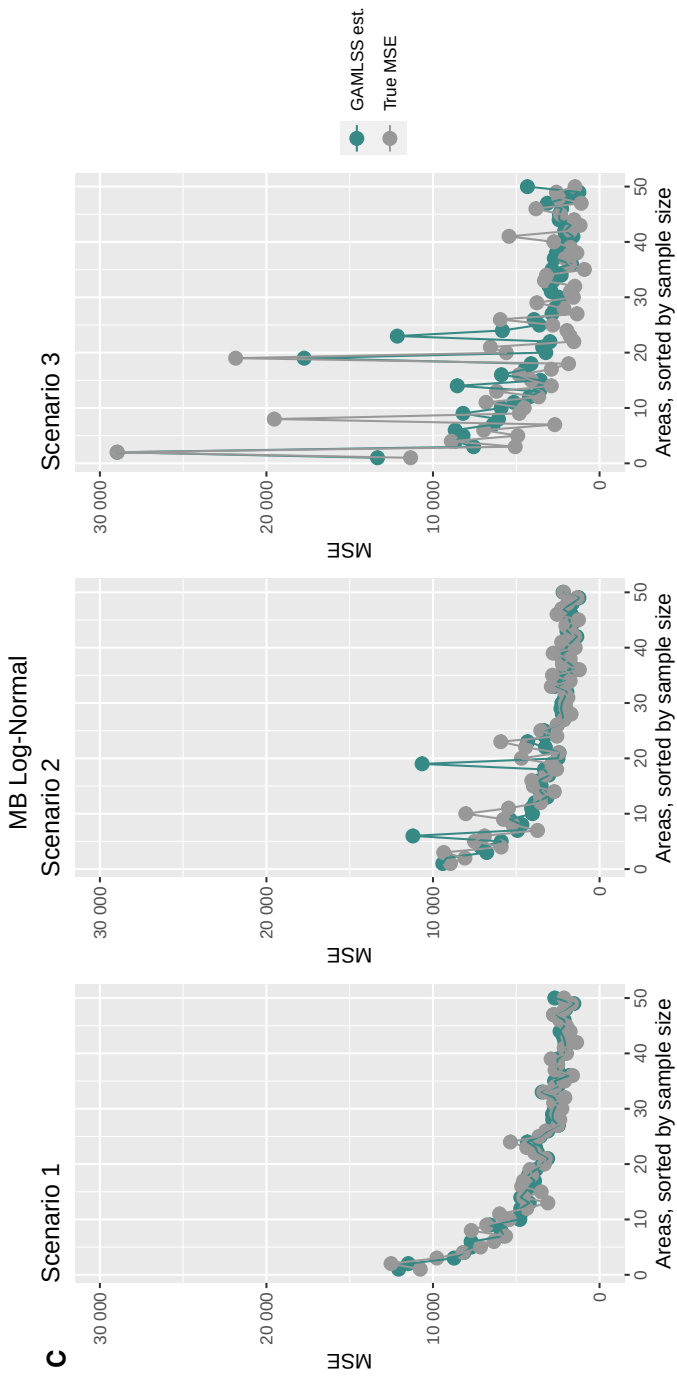

Figure 1. Continued.

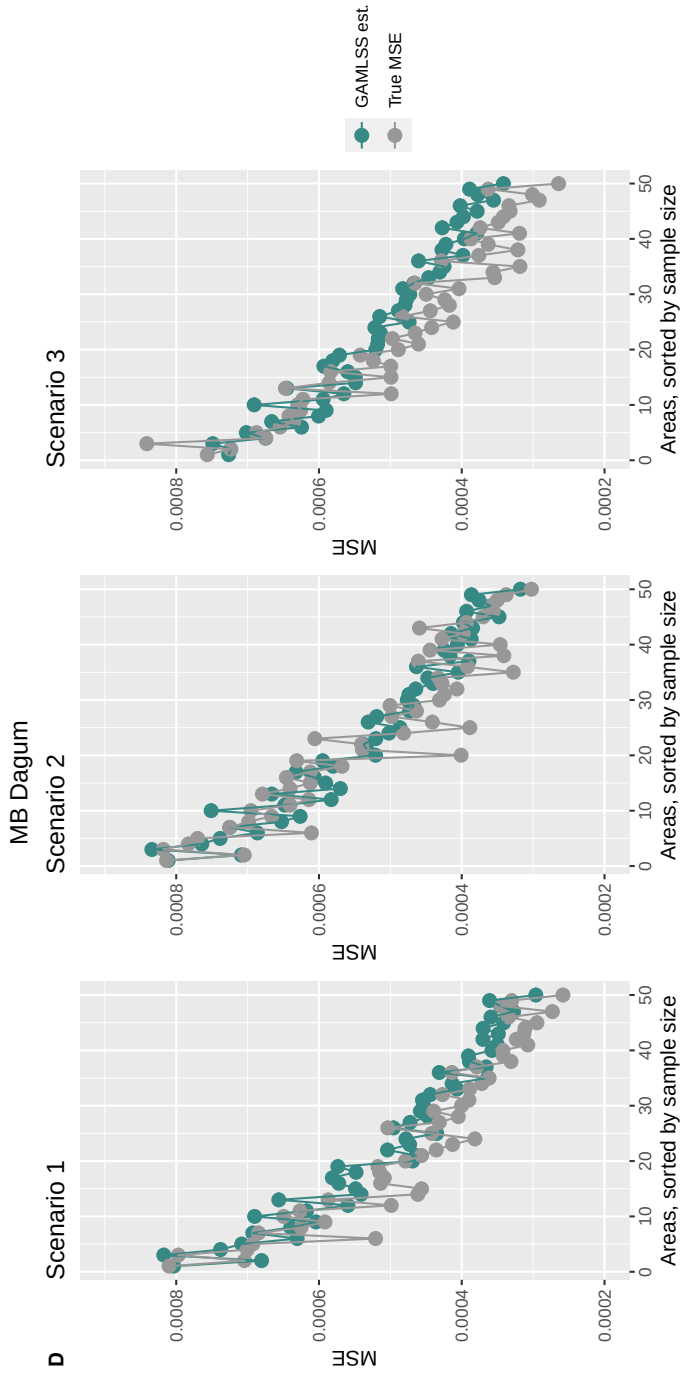

Figure 1. Continued.

of EBLUP. Moving on to the estimates' reliability, ATMSEs show that MB estimators perform better than the direct estimator and GAMLSS performs better than EBLUP. Figure 6c, in Appendix A in the supplementary data online, reports area-specific bias. From figure 6c, we can appreciate how GAMLSS reduce the bias with respect to the EBLUP. As noted by Molina and Martìn (2018), the use of log-transformation in models where only the location parameter is defined through covariates, also employed by Molina and Rao (2010) among others, could add to the estimates also a back-transformation error. The main differences between a SAE model based on GAMLSS and the models proposed by Molina and Martìn (2018) and Molina and Rao (2010), lie in the flexibility of incorporating covariates for all parameters and in the use of suitable lognormal distributional assumption. SAE model based on GAMLSS assumes a lognormal model and incorporate covariates into both the location parameter (μ) and the scale parameter (σ). In contrast, Molina and Rao (2010) revert to normality with a log-transformation and use covariates only on the location parameters. Molina and Martìn (2018), while assuming the same model, correct the estimates for back-transformation errors but do not have the opportunity to include covariates in all parameters of the distribution. The values of ATMSE, ABMSE, and PCR for both GAMLSS and EBLUP shows, for GAMLSS, the accuracy of the proposed bootstrap MSE estimator in terms of MSE estimate. The PCR is a bit lower than 95 percent. Again, figure 1c compares the path of the real and empirical MSE for the GAMLSS, showing how well the latter approximates the true MSE.

On simulation *(D) MB Dagum*, we note that at first GAMLSS sharply reduces the bias with respect to the EBLUP. Secondly, differences in the ATMSEs increase when both parameters depend on covariates, that is the difference between the ATMSE of GAMLSS and EBLUP is higher in the third scenario than in the first two. Figure 6d reports area-specific bias (Appendix A in the supplementary data online). As previously seen for area-specific results, we can appreciate the reduction of relative bias if a GAMLSS model is used instead of an EBLUP. As noted by Schluter (2012), the high negative values of the EBLUP estimator bias are due to the capacity of a SAE model that assumes normal distribution to estimate well the central trend, and to underestimate value on the tails of the distribution. When normality is assumed, the predicted values are compressed around the mean with thin tails leading to underestimate inequality indices. This problem, however, is not evident in the MSE, given the high capacity of the normal distribution to tend to the mean values. To conclude, the values of ATMSE and ABMSE, reported in table 2 and figure 1d, which compares the path of the real and empirical MSE for the GAMLSS, show the accuracy of the proposed MSE estimator even in this simulation. The values of the ATMSE, ABMSE, and PCR for the EBLUP show how this can perform very poorly if the distribution is not normal.

4.4.2 DB simulation.

Table 3 shows that for the DB simulation (*(A) DB consumption data)* GAMLSS clearly reduces the ARB if compared to the EBLUP (see also figure 2). The value of the ATMSE is lower for GAMLSS than for EBLUP and the direct estimator. The values of the ABMSE and PCR for both GAMLSS and EBLUP are satisfying and in line with those obtained for MB simulations with acceptable PCR. To conclude, from figure 2, we can also appreciate that the MSE procedure defined for GAMLSS well approximates the true MSE in both simulations.

Table 3. Measures of Performance: Design-Based Simulation

(A) DB consumption data

	Direct	GAMLSS	EBLUP
ARB	–0.07	0.62	1.07
AARB	19.51	13.06	13.66
ATMSE	31947.81	11964.91	13408.81
ABMSE	–	10803.83	10984.75
PCR	–	91.20	88.41

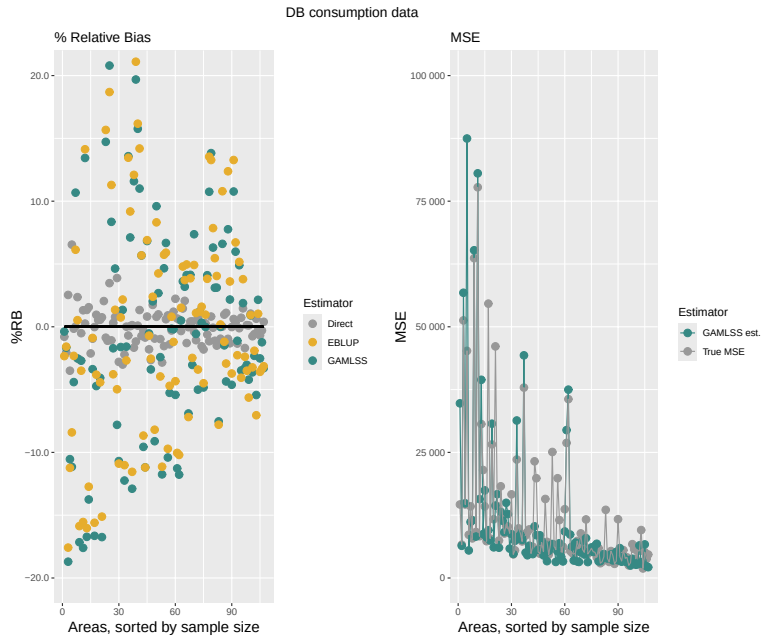

Figure 2. Area-specific relative bias and MSE: design-based simulation.

Before moving away from simulations, let us highlight the last two points. First, all the performance indices presented in this paragraph have also been analyzed by grouping n_j into three categories (i.e., $n_j \leq 15$, $15 < n_j \leq 35$, and $35 < n_j \leq max(n_j)$). Results do not indicate any emerging notable trends beyond the known one, that is the increase in ARB as the sample size decreases. Results are available upon request. Secondly, we briefly address the computational burden for both MB and DB simulations. It is evident that simulation *(A) MB Normal* can be seen only as an exercise to verify if the two models give the same results. Once this has been verified, and given that EBLUP is faster than GAMLSS, there is no reason to use GAMLSS. For all the other simulations, although GAMLSS times are higher, in our opinion, the improvements that GAMLSS determine in measures of performance justify their use. Moreover, using appropriate shrewdness makes it possible to reduce computational times using parallelization and/or specific options of the GAMLSS functions, that is, *start.from()* (Stasinopoulos et al. 2017).

5. ESTIMATION OF PER-CAPITA CONSUMPTION FOR ITALIAN AND FOREIGN CITIZENS IN REGIONAL RURAL AND URBAN AREAS

Here, we use the proposed SAE model based on GAMLSS (now on SAE-GAMLSS) to estimate the per-capita equivalized consumption expenditure (PCE) of Italian and foreign citizens in the various regions of Italy, distinguishing between urban and rural areas.[2] The reduction of the gap in the expenditure between the two groups should be a central point for every politician that aims to reduce difference and social tension between immigrants and natives. As noted by Djajic (2003) and Barigozzi and Speciale (2011), consumption is an important dimension in the integration of immigrants in the host country. As immigrants appear to be more similar to natives, integration in terms of consumption can open up a wide range of market and social opportunities for them. We consider data from the 2019 Italian HBS. This survey is the main data source used to produce regular statistics on household expenditure in Italy. It gathers information on expenditure as well as on potential auxiliary variables such as age groups and sex. We would like to point out here that the only covariates that we are able to use are the ones known in the population and available in the "Adele LAB." of ISTAT. The sample sizes allow us to obtain reliable estimates of mean expenditure at a regional level. Due to small sample sizes, estimates are unreliable for foreigners with reference to regional urban and rural areas, where the minimum sample size is equal to two

2. A person is defined as living in an urban area if he/she lives in a community with more than 50,000 citizens (Dijkstra and Poelman 2014). Two Italian regions (Aosta Valley and Molise) do not have any community with a populations higher than this value.

with a median number of sampled units of only fifty-four and direct DB estimators for 25 percent of the areas have a coefficient of variation (CV) higher than about 21 percent. We need to restore estimates with SAE methods. In order to select, among a certain number of parametric distributions, the one that best fit our data, we use the function *fitDist* of the GAMLSS package (Rigby et al. 2019, Ch. 6) that use the penalized maximum likelihood. The final marginal distribution is selected based on the GAIC. Following Section 3.3, we decide to use a GAIC with a penalty equal to $\sqrt{\log(n)} \simeq 2$ and to also report the Bayesian information criterion (BIC). We will now move on to the following distributions: the gamma distribution used by Battese and Bonyhady (1981), the Dagum used by Prieto-Alaiz (2007), the lognormal used by Battistin et al. (2009), and, to conclude, the Student-*t* used by Ubaidillah et al. (2018). Alongside these distributions that have already been used by some authors to model the PCE, we also decide to include the following distributions: normal, skew-normal, skew-*t*, GB2, and Singh–Maddala. Note that the GB2 distribution includes as special cases the Singh–Maddala and the Dagum distributions. Table 4 summarizes the GAIC and the BIC of the selected distributions. The two criteria give different results: the minimum GAIC is reached with the GB2 distribution while the minimum BIC with the lognormal one. Following Section 3.3, for those two distributions, after selecting the appropriate covariates for each parameter, we use the K-fold cross-validation, with seven-folds and repeating the loop 200 times (for the choice of the number of

Table 4. GAIC and BIC of Selected Distributions

Dist.	GAIC	BIC
Dagum	45,883	45,881
Gamma	46,096	46,108
GB2	**45,859**	45,882
Log-Normal	45,883	**45,864**
Normal	47,538	47,526
Singh-Maddala	45,887	45,868
Skew-Normal	47,546	47,582
Skew-*t*	45,909	45,891
Student-*t*	46,581	46,563

NOTE.—In bold the lowest values of GAIC and BIC.

Table 5. Summary of the Quantile Residuals

Mean = –0.001	Coef. of skewness = 0.297
Variance = 1.000	Coef. of kurtosis = 3.018

Table 6. Model Estimates

Fitted method					
	Estimate	Std. Error	*t* value	Pr(>|t|)	
Mu link function: log					
Mu coefficients:					
Intercept	6.94	0.02	344.71	0.00	***
Sex (women)	0.04	0.02	2.47	0.01	*
Age 18–24	0.17	0.03	5.00	0.00	***
Age 25–29	0.19	0.03	5.10	0.00	***
Age 30–34	0.22	0.04	6.63	0.00	***
Age 35–39	0.21	0.03	6.27	0.00	***
Age 50–44	0.23	0.03	6.65	0.00	***
Age 45–49	0.26	0.04	7.53	0.00	***
Age 50–54	0.28	0.03	8.04	0.00	***
Age 55–59	0.23	0.04	5.64	0.00	***
Age 60–64	0.21	0.06	3.56	0.00	***
Age 65–69	0.45	0.08	5.35	0.00	***
Age 70–74	0.34	0.10	3.33	0.00	***
Age 75+	0.27	0.11	2.37	0.01	*
Sigma link function: log					
Sigma coefficients:					
Intercept	–0.78	0.03	–28.60	0.00	***
Age 18–24	–0.06	0.05	–1.29	0.25	
Age 25–29	–0.07	0.06	–0.13	0.90	
Age 30–34	0.10	0.05	2.12	0.03	*
Age 35–39	0.09	0.05	1.88	0.06	.
Age 50–44	0.15	0.05	3.04	0.00	***
Age 45–49	0.09	0.05	1.62	0.09	.
Age 50–54	0.04	0.05	0.71	0.48	
Age 55–59	0.01	0.06	0.23	0.81	
Age 60–64	0.15	0.08	1.96	0.04	*
Age 65–69	0.31	0.09	3.19	0.00	***
Age 70–74	–0.01	0.16	–0.07	0.95	
Age 75+	0.33	0.13	2.51	0.01	*
Significant codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1

NOTE.—No. of observations in the fit: 2,854.
Degrees of freedom for the fit: 89.84.
Residual degree of freedom: 2,764.15.

K-fold, see Nti et al. 2021). The GAMLSS based on the lognormal distribution is the one with the smallest overall value of the GDEV ($\simeq 42496$). The one based on the GB2 distribution has a GDEV about to $\simeq 42528$. Based on those

Table 7. Summary of the Random Effects Standard Deviations

Parameter	Estimate	Std. Error	t value	Pr($>\|t\|$)	
σ_μ	0.22	0.04	5.28	$<5e-7$	***
σ_σ	0.21	0.04	4.53	0.00	***
Significant codes:	0^{***}	0.001^{**}	0.01^{*}	$0.05^{\cdot}$	0.1

results and on the residual diagnostic reported on table 5, we conclude that lognormal stands as the most suitable distribution for our predictive objective.

The selected GAMLSS is based on a lognormal distribution, with *Sex* (coded as 0 for men and 1 for women) and 14 *Age* classes serving as covariates for μ, while 14 *Age* classes are considered for σ. Both parameters are also defined by a random effect. The model is:

$$\begin{cases} \mu_{ij} = \exp(\beta_0^\mu + \beta_1^\mu \text{Sex}_{ij} + \beta_1^\mu \text{Age}_{ij} + \gamma_j^\mu) \\ \sigma_{ij} = \exp(\beta_0^\sigma + \beta_1^\sigma \text{Age}_{ij} + \gamma_j^\sigma) \end{cases}$$

where *Sex* and *Age* are factors. Table 6 reports the estimated coefficients for μ and σ, while table 7 is about the variance of the random effects. What is important to note, with regard to the SAE framework, is that the possibility to define each parameter in terms of covariates, improves the model. Additionally, the variances of the random effects are statistically different from 0.

SAE-GAMLSS estimates of PCE for the selected domains are summarized in figure 3, while in table 8 summary statistics on CV are reported for direct and SAE-GAMLSS estimators. The MB estimated PCE has a very similar path for the unbiased DB estimators and the SAE-GAMLSS estimator noticeably reduces the CVs. The dashed red line in figure 3 is the threshold beyond which the estimates would be un-publishable according to the criteria adopted by, among others (Statistics-Canada 2007). More than half the areas have a CV higher than this line with the direct estimators reducing only to two if we use the GAMLSS-based estimator.

To obtain reliable estimates of the PCE for Italian citizens, it is sufficient to use direct estimators since the mean CV is about 4.30. Figure 4 reports the 95 percent confidence intervals of the estimated values for Italian (direct estimator) and foreign citizens (SAE-GAMLSS estimator) divided into urban and rural areas. From this figure, it is possible to appreciate that in almost every area, the confidence intervals of the SAE-GAMLSS do not overlap with the ones of Italian citizens. PCE for foreigners is markedly lower than the ones of Italians. This result cannot be considered significant based on direct estimates where confidence intervals overlap.

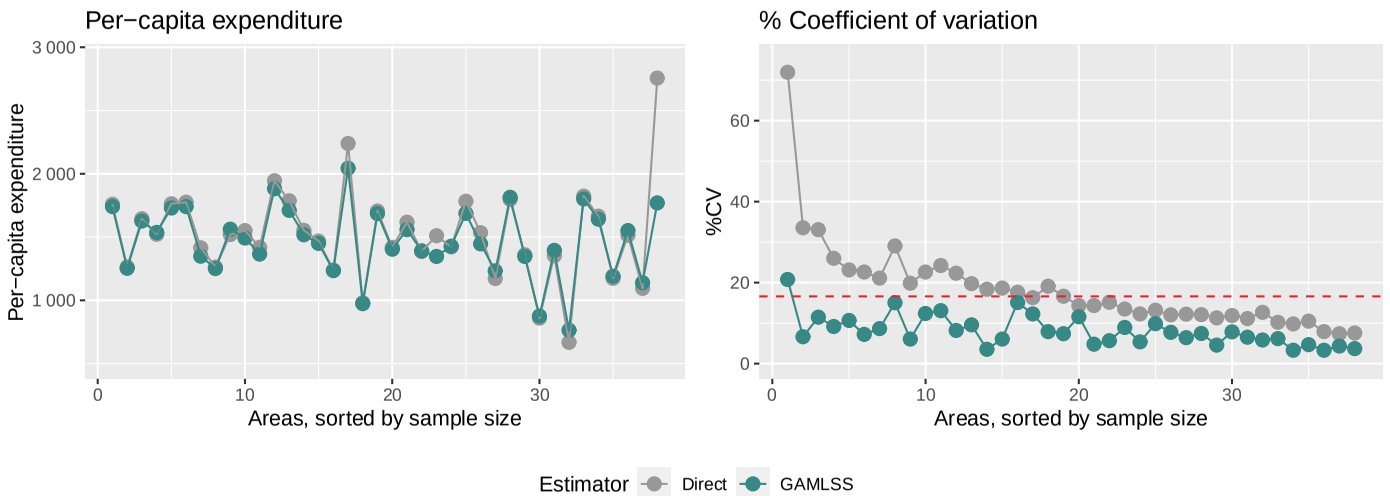

Figure 3. Estimated per-capita expenditure of foreigners and coefficient of variation. Dashed line: threshold beyond which estimates would be un-publishable (CV$\simeq$ 16%).

Table 8. Summary of Coefficient of Variation

	Min.	Q1	Median	Mean	Q3	Max.
Direct	7.07	11.48	15.27	18.18	20.99	70.77
GAMLSS	3.27	5.70	7.41	8.13	9.77	20.76

Figure 5 maps the PCE for Italian and foreign citizens in each regional urban and rural area. The well-known North–South divide is recognizable with reference to Italian citizens, whereas for foreigners, the difference between North and South, even if present, is much less marked. Those results are also confirmed by figure 4. Note that in this figure, the x-axis follows the enumeration of the regions in figure 5. That is to say, approximately, the first half of figure 4 represents the regions in the North and Center of Italy, while the second half represents the ones in the Islands and in the South. Looking at this figure, the descending trends, for of the estimates of Italians moving from North to South in both rural and urban areas are clear. The same trend is not present when looking at foreigners. A Kruskal–Wallis rank sum test confirms those results having p-values $\leq .05$ only for Italians in both urban and rural areas.

As regards the single regions, Liguria and Tuscany show the largest difference between the two groups based on citizenship, at the rural and urban level, respectively. Interestingly, and perhaps unexpectedly, Apulia and Basilicata are characterized by a relative homogeneity among the four considered groups. These considerations are only few examples based on results otherwise unobtainable without SAE models. Reliable MB SAE could help policy makers to define place-based policies aimed at reducing differences between Italian and foreign citizens, in this way encouraging the integration of the two groups and reducing potential social tensions.

6. CONCLUDING REMARKS

This paper proposes a new unit-level SAE model based on GAMLSS. Under this model, we present a method of obtaining small area predictions for various economic indicators on household income and consumption, and a bootstrap approach to estimate their MSE. To avoid model misspecification, we also suggest a step-by-step procedure for model selection that in the GAMLSS framework can refer to a high number of possible distributions. Additionally, in future work, an extension of the benchmarking method recently proposed by Gutiérrez et al. (2022) might be considered to further avoid misspecification. The performances of the different estimators are compared through MB and DB simulations. Our results provide evidence that the

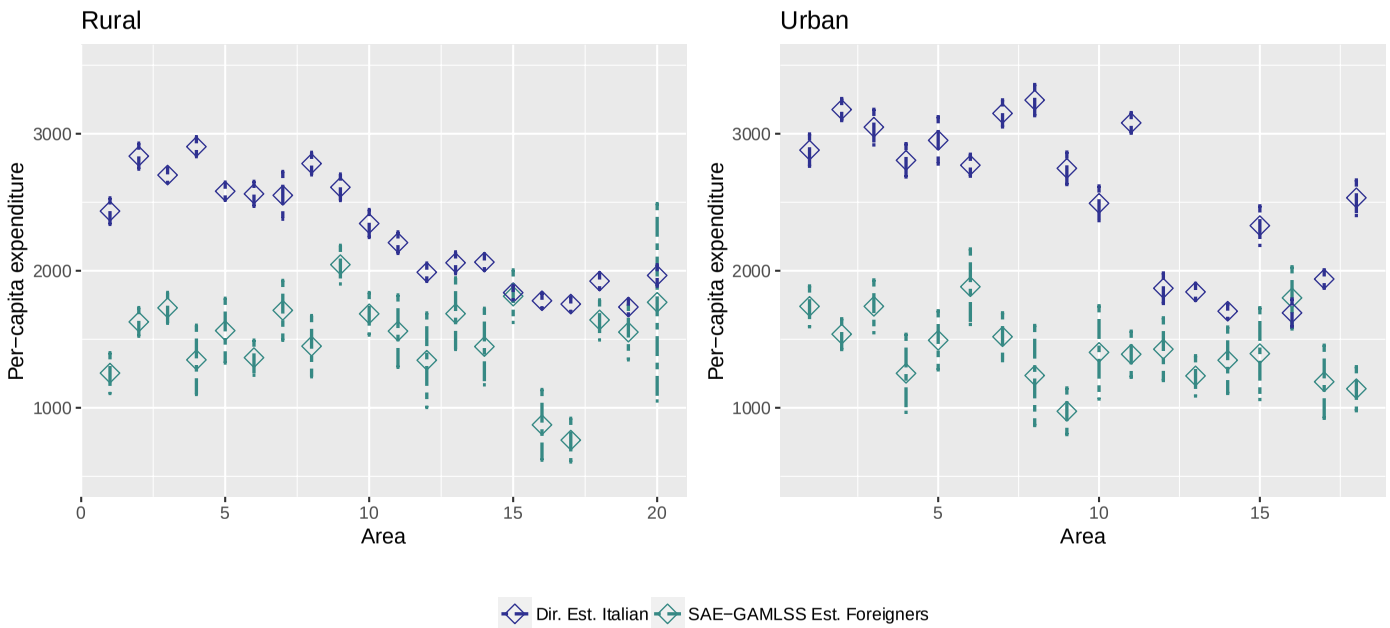

Figure 4. Confidence intervals of estimated values.

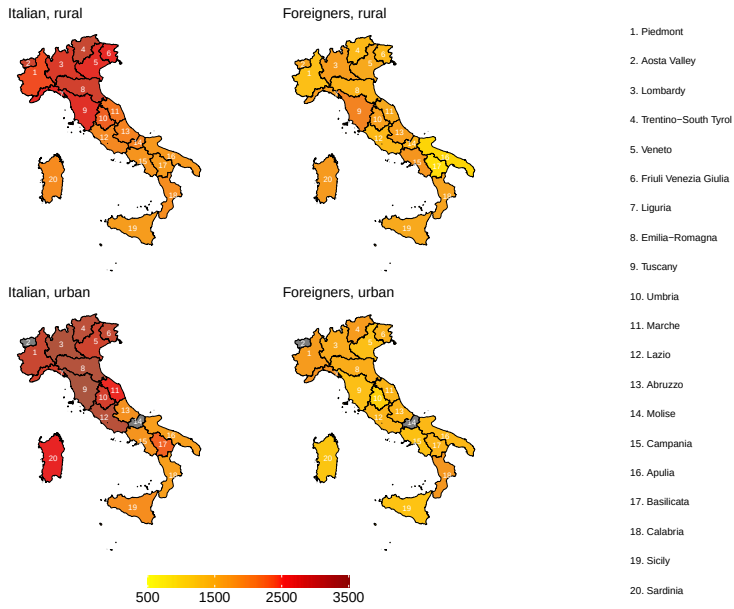

Figure 5. Estimated per-capita expenditure for Italian and foreign citizens (Aosta Valley and Molise do not have ubran areas).

proposed small area predictors perform very well in terms of variability reduction with respect to the well-known EBLUP estimator. If data are normal and homoskedastic, SAE-GAMLSS and EBLUP gives almost the same results. However, in case of heteroskedasticity, the SAE-GAMLSS also allows us to model σ depending on covariates, and the related predictor clearly outperforms EBLUP. In the presence of variables with a skewed distribution, when either a lognormal distributional assumption is adopted or when a three-parameter Dagum distribution is assumed, the SAE-GAMLSS largely improve estimates. Finally, the strategy to estimate the MSE of SAE estimators is proven to be able to estimate well the true MSE. Based on the SAE-GAMLSS, we estimate the per-capita expenditure of Italian and foreign citizens, in rural and urban areas within administrative regions. The obtained maps allow us to enlighten a noticeable heterogeneity among defined domains, thus allowing the definition of place-based policies aiming to equity.

There are other areas that warrant further investigation. First of all, a possible extension of the SAE-GAMLSS is to account for the assumption of sample informativeness. The results obtained by Verret et al. (2015) and Kim et al. (2022) could be profitably used for this purpose. In particular, the method proposed by Verret et al. (2015), which integrates sample weights as covariates, could serve as a starting point for the development of a future research direction considering informativeness in the SAE-GAMLSS. However, in the

context of GAMLSS, the approach proposed by Verret et al. (2015) needs to be adapted to models where link functions are often different from the identity one as done, for example, by Vandendijck et al. (2016). Second, the implementation of multivariate SAE-GAMLSS to estimate related outcomes could be of great interest in the economic framework (i.e., turnover and number of employees for firms, income, and consumption for households). Third, it could be interesting to take advantage of the dependence of some non-linear indicators on the parameters' distribution (i.e., the Gini index under the lognormal assumption is completely defined by the scale parameter). In this case, the distribution parameters can be modeled in terms of covariates, and then the outcome prediction can be obtained in a closed-form through the mentioned relationship.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Barigozzi, M., and Speciale, B. (2011), "Immigrants' Legal Status, Permanence in the Destination Country and the Distribution of Consumption Expenditure," *Applied Economics Letters*, 18, 1341–1347.

Battese, G. E., and Bonyhady, B. P. (1981), "Estimation of Household Expenditure Functions: An Application of a Class of Heteroscedastic Regression Models," *Economic Record*, 57, 80–85.

Battese, G. E., Harter, R. M., and 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*, 83, 28–36.

Battistin, E., Blundell, R., and Lewbel, A. (2009), "Why is Consumption More Log-Normal than Income? Gibrat's Law Revisited," *Journal of Political Economy*, 117, 1140–1154.

Breidenbach, J., Magnussen, S., Rahlf, J., and Astrup, R. (2018), "Unit-Level and Area-Level Small Area Estimation under Heteroscedasticity Using Digital Aerial Photogrammetry Data," *Remote Sensing of Environment*, 212, 199–211.

Chambers, R., and Tzavidis, N. (2006), "M-Quantile Models for Small Area Estimation," *Biometrika*, 93, 255–268.

Chandra, H., Chambers, R., and Salvati, N. (2012), "Small Area Estimation of Proportions in Business Surveys," *Journal of Statistical Computation and Simulation*, 82, 783–795.

Chandra, H., Chambers, R. L., and Salvati, N. (2019), "Small Area Estimation of Survey Weighted Counts under Aggregated Level Spatial Model," *Survey Methodology*, 45, 31–59.

Chandra, H., Kumar, S., and Aditya, K. (2018), "Small Area Estimation of Proportions with Different Levels of Auxiliary Data," *Biometrical Journal*, 60, 395–415.

Cutillo, A., Arigoni, I., Valdoni, L., and De Martino, V. (2020), "Indagine sulle spese delle famiglie. periodo di riferimento: anno 2019. Aspetti metodologici." Technical report, Rome, Italy.

Dijkstra, L., and Poelman, H. (2014), "A harmonised definition of cities and rural areas: the new degree of urbanisation. European Comission Directorate-General for Regional and Urban Policy." Regional Working Paper 2014, Bruxelles, Belgium.

Djajic, S. (2003), "Assimilation of Immigrants: Implications for Human Capital Accumulation of the Second Generation," *Journal of Population Economics*, 16, 831–845.

Folsom, R., Shah, B. and Vaish, A. (1999), "Substance Abuse in States: A Methodological Report on Model Based Estimates From the 1994-1996 National Household Surveys on Drug Abuse," in Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria, VA, pp. 371–375.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998), "Generalized Linear Models for Small-Area Estimation," *Journal of the American Statistical Association*, 93, 273–282.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2007), "Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model," *Computational Statistics & Data Analysis*, 51, 2720–2733.

Graf, M., Marin, J. M., and Molina, I. (2019), "A Generalized Mixed Model for Skewed Distributions Applied to Small Area Estimation," *TEST*, 28, 565–597.

Gutiérrez, A., Mancero, X., and Guerrero, S. (2022), "Poverty Mapping in Latin America: ECLAC Experiences on Small Area Estimation," *Statistical Journal of the IAOS*, 38, 1021–1033.

Hidiroglou, M. A., and You, Y. (2016), "Comparison of Unit Level and Area Level Small Area Estimators," *Survey Methodology*, 42, 41–61.

Hofner, B., Mayr, A., and Schmid, M. (2014), "gamboostlss: An R Package for Model Building and Variable Selection in the gamlss Framework," arXiv:1407.1774.

ISTAT (2022), "Il laboratorio per l'analisi dei dati elementari ADELE," Technical report, Rome, Italy: ISTAT.

Jiang, J. (2003), "Empirical Best Prediction for Small-Area Inference Based on Generalized Linear Mixed Models," *Journal of Statistical Planning and Inference*, 111, 117–127.

Jiang, J., and Nguyen, T. (2012), "Small Area Estimation via Heteroscedastic Nested-Error Regression," *Canadian Journal of Statistics*, 40, 588–603.

Jiang, J., and Rao, J. S. (2020), "Robust Small Area Estimation: An Overview," *Annual Review of Statistics and Its Application*, 7, 337–360.

Kim, J. K., Rao, J., and Kwon, Y. (2022), "Analysis of Clustered Survey Data Based on Two-Stage Informative Sampling and Associated Two-Level Models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185, 1522–1540.

Kneib, T. (2013), "Beyond Mean Regression," *Statistical Modelling*, 13, 275–303.

———. (2020), "Comments on: Inference and Computation with Generalized Additive Models and Their Extensions," *TEST*, 29, 351–353.

Liu, Y., Liu, X., Pan, Y., Jiang, J., and Xiao, P. (2022), "An Empirical Comparison of Various MSPE Estimators and Associated Prediction Intervals for Small Area Means," *Journal of Statistical Computation and Simulation, Pages*, 93, 1532–1558.

Lyu, X., Berg, E. J., and Hofmann, H. (2020), "Empirical Bayes Small Area Prediction under a Zero Inflated Log-Normal Model with Correlated Random Area Effects," *Biometrical Journal*, 62, 1859–1878.

Marino, M. F., Ranalli, M. G., Salvati, N., and Alfo, M. (2019), "Semiparametric Empirical Best Prediction for Small Area Estimation of Unemployment Indicators," *The Annals of Applied Statistics*, 13, 1166–1197.

Molina, I., and Marhuenda, Y. (2015), "Sae: An R Package for Small Area Estimation," *R Journal*, 7, 81.

Molina, I., and Martìn, N. (2018), "Empirical Best Prediction under a Nested Error Model with Log Transformation," *The Annals of Statistics*, 46, 1961–1993.

Molina, I., and Rao, J. (2010), "Small Area Estimation of Poverty Indicators," *Canadian Journal of Statistics*, 38, 369–385.

Molina, I., and Strzalkowska-Kominiak, E. (2020), "Estimation of Proportions in Small Areas: Application to the Labour Force Using the Swiss Census Structural Survey," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183, 281–310.

Nelder, J. A. (2006), "Contribution to the Discussion of Rigby and Stasinopoulos. *Generalized Additive Models for Location, Scale and Shape*," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 547–548.

Nti, I. K., Nyarko-Boateng, O., and Aning, J. (2021), "Performance of Machine Learning Algorithms with Different K Values in K-Fold Cross-Validation," *I.J. Information Technology and Computer Science*, 13, 61–71.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008), "Non-Parametric Small Area Estimation Using Penalized Spline Regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 265–286.

Parker, P. A., Janicki, R., and Holan, S. H. (2023a), "Comparison of Unit-Level Small Area Estimation Modeling Approaches for Survey Data under Informative Sampling," *Journal of Survey Statistics and Methodology*, 11, 858–872.

———. (2023b), "A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation under Informative Sampling," *Journal of Survey Statistics and Methodology*, 11, 829–857.

Pfeffermann, D., and Correa, S. (2012), "Empirical Bootstrap Bias Correction and Estimation of Prediction Mean Square Error in Small Area Estimation," *Biometrika*, 99, 457–472.

Prasad, N. N., and Rao, J. N. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163–171.

Prieto-Alaiz, M. (2007), "Spanish Economic Inequality and Gender: A Parametric Lorenz Dominance Approach," *Emerald Group Publishing Limited*, 14, 49–70.

Ramirez-Aldana, R., and Naranjo, L. (2021), "Random Intercept and Linear Mixed Models Including Heteroscedasticity in a Logarithmic Scale: Correction Terms and Prediction in the Original Scale," *PloS One*, 16, e0249910.

Rao, J., and Choudhry, G. (1995), "Small Area Estimation: Overview and Empirical Study," in *Business Survey Methods*, eds. B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott, New York: John Wiley & Sons, Inc.

Rao, J., and Molina, I. (2015), *Small Area Estimation* (2nd ed.), Hoboken, NJ: John Wiley & Sons, Inc.

Ren, W., Li, J., Erciulescu, A., Krenzke, T., and Mohadjer, L. (2022), "A Variable Selection Method for Small Area Estimation Modeling of the Proficiency of Adult Competency," *Stats*, 5, 689–713.

Rigby, R. A., and Stasinopoulos, D. M. (2005), "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554.

Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019), *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R* (1st ed.), New York, NJ: Chapman and Hall.

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2020), "Data-Driven Transformations in Small Area Estimation," *Journal of the Royal Statistical Society: Series A (Statistics in Society*, 183, 121–148.

Rueda, C., and Lombardia, M. (2012), "Small Area Semiparametric Additive Monotone Models," *Statistical Modelling*, 12, 527–549.

Schluter, C. (2012), "On the Problem of Inference for Inequality Measures for Heavy-Tailed Distributions," *The Econometrics Journal*, 15, 125–153.

Stasinopoulos, D. M., Rigby, R., Heller, G., Voudouris, V., and De Bastiani, F. (2017), *Flexible Regression and Smoothing: Using GAMLSS in R.*, New York, NJ: Chapman and Hall CRC.

Statistics-Canada (2007), "2005 Survey of Financial Security—Public Use Microdata File, User Guide," Published by Authority of the Minister Responsible for Statistics Canada. Technical report, Ottawa, ON, Canada: Statistics-Canada.

Steorts, R., Schmid, T., and Tzavidis, N. (2020), "Smoothing and Benchmarking for Small Area Estimation," *International Statistical Review*, 88, 580–598.

Sugasawa, S., and Kubokawa, T. (2020), "Small Area Estimation with Mixed Models: A Review," *Japanese Journal of Statistics and Data Science*, 3, 693–720.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., and Rojas-Perilla, N. (2018), "From Start to Finish: A Framework for the Production of Small Area Official Statistics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 927–979.

Ubaidillah, A., Notodiputo, K. A., Kurnia, A., and Mangku, I. W. (2018), "A Comparative Study of Robust t Linear Mixed Models with Application to Household Consumption per Capita Expenditure Data," *Applied Mathematical Sciences*, 12, 57–68.

Vandendijck, Y., Faes, C., Kirby, R., Lawson, A., and Hens, N. (2016), "Model-Based Inference for Small Area Estimation with Sampling Weights," *Spatial Statistics*, 18, 455–473.

Verret, F., Rao, J. N., and Hidiroglou, M. A. (2015), "Model-Based Small Area Estimation under Informative Sampling," *Survey Methodology*, 41, 333–348.

Wood, S. N. (2020), "Inference and Computation with Generalized Additive Models and Their Extensions," *Test*, 29, 307–339.

Würz, N., Schmid, T., and Tzavidis, N. (2022), "Estimating Regional Income Indicators under Transformations and Access to Limited Population Auxiliary Information," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,185, 1679–1706.

You, Y., and Rao, J. (2002), "A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights," *Canadian Journal of Statistics*, 30, 431–439.

Journal of Survey Statistics and Methodology, 2025, 13, 160–196
https://doi.org/10.1093/jssam/smae038
Article