

Small area estimation of economic inequality indices using GAMLSS in the absence of covariates

Lorenzo Mori  and Maria Rosaria Ferrante 

Department of Statistical Sciences “P. Fortunati”, University of Bologna, Via Belle Arti 41, 40100 Bologna, Italy

Address for correspondence: Lorenzo Mori, Department of Statistical Sciences “P. Fortunati”, University of Bologna, Via Belle Arti 41, 40100 Bologna, Italy. Email: lorenzo.mori7@unibo.it

Abstract

The goal of this study is to estimate inequality indices for foreigners living in Italy at the regional level, differentiating between urban, peri-urban, and rural areas. This issue requires a Small Area Estimation (SAE) model. We operate in the unit-level context that has brought forth two challenges: the identification of individual covariates and the reduction of computational time. We propose a unit-level Simplified SAE model based on Generalized Additive Models for Location, Scale, and Shape specified without covariates and able to reduce variability in comparison with the direct estimator. A non-parametric bootstrap, suitable without design information, is proposed to estimate the mean square error. The performance of the proposed model used to estimate three different inequality indices (Gini, Theil, and Atkinson) is evaluated based on design-based simulations. The results show that the proposed predictor reduces the variability of the direct estimates even when covariates are not available. This methodology has been adopted to estimate the three indices for foreigners in Italy at the regional level, distinguishing between urban, peri-urban, and rural areas, showing that relevant disparities emerge in inequality between foreigners compared to natives. These results can help to formulate place-based policies aimed at promoting equity and integration.

Keywords: auxiliary information, consumption, foreigners, non-parametric bootstrap, SAE

1 Introduction

The demographic changes in developed countries, propelled by increased immigration, raise questions about social inclusion and inter-group conflicts. Policymakers are grappling with the challenge of understanding the experiences of diverse immigrant groups and the emergence of new socio-demographic patterns of consumption inequality among various ethnicities (D’Agostino et al., 2016). As highlighted by Djajic (2003) and Barigozzi and Speciale (2011), consumption is a crucial dimension in the integration of immigrants into the host country. Higher levels of integration are achieved when immigrants’ consumption patterns become more similar to those of natives. There is a need to estimate inequalities across sub-populations defined by both social and geographical characteristics, in order to identify disparities more accurately and develop targeted interventions that address the unique challenges faced by different communities. Placing emphasis on regions rather than countries is the first step in implementing place-based policies aimed at reducing local inequalities through tailored measures (McCann & Soete, 2020).

In the last decades, the evolving landscape of urbanization and the migration from rural to urban areas has significantly reshaped the economic framework, resulting in notable shifts in family income sources (Ma & Yin, 2024). The rural economic transformation has increased farmers’ income, and this has a subsequent impact on their consumption patterns. The widening urban–rural

Received: October 28, 2024. Revised: September 1, 2025. Accepted: September 1, 2025

© The Royal Statistical Society 2025.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

consumption gap emphasizing the urgent need for comprehensive strategies to address this growing disparity.

To the best of our knowledge, no existing studies address the issue of estimating consumption inequality at the local level distinctly for native born individuals and foreigners, and also distinguishing between urban and rural areas. The literature so far has focused solely on estimating inequalities or per-capita consumption expenditure for the population as a whole (Balli & Tiezzi, 2010; Fasulo et al., 2017).

In this scenario, we aim to estimate inequality indices based on consumption expenditure for foreigners and Italians,¹ at regional level and distinguishing between urban, peri-urban, and rural areas. To achieve this, we use data from the 2019 Italian Household Budget Survey (HBS), which supply reliable estimates only for larger areas or domains. Due to the need for obtaining reliable estimates at more disaggregated areas, we have to adopt Small Area Estimation (SAE) methods. SAE techniques are recognized statistical tools for producing reliable estimates when the sample size within a specific area or domain is insufficient. SAE models aim to reduce the variability of estimates for a given area by using the collective information from neighbouring areas and auxiliary data sources that are typically free of sampling error (Rao & Molina, 2015).

Although numerous studies have examined SAE methods for inequality indices, ranging from the seminal work of Molina and Rao (2010) to more specialized contributions by Fabrizi and Trivisano (2016) and Gardini et al. (2021), our research focuses on addressing two prevalent challenges in unit-level SAE models: the scarcity of covariates and the high computational burden. To this end, we propose a unit-level SAE model that effectively reduces the variability of the Direct estimator without relying on covariates, while also requiring less computational effort compared to existing unit-level models.

Firstly, consumption expenditure is a continuous variable, defined on the positive real line, highly skewed, and potentially influenced by outliers. The standard approach of applying a log-transformation to such variables is not always effective in achieving normality and is also susceptible to back-transformation bias (Molina & Martín, 2018). The treatment of the log-transformation in SAE, in both unit- and area-level models, has been addressed by several authors. The main contributions, in addition to that of Molina and Martín (2018), are those of Slud and Maiti (2006), Chandra and Chambers (2011), Li et al. (2019), and Rojas-Perilla et al. (2020), which explore different strategies to handle the non-normality of the response variable and its implications for prediction and inference. For example, Slud and Maiti (2006) examine the problem of accurately estimating the variability of predictions, while Chandra and Chambers (2011) propose a model-based direct estimator that employs calibration techniques to define appropriate weights. Li et al. (2019) and Rojas-Perilla et al. (2020) extend these approaches to more flexible frameworks, incorporating model-based adjustments and semi-parametric techniques. These contributions underscore the importance of properly addressing the skewness and heteroscedasticity inherent in consumption data. Furthermore, normality is often assumed not only for the dependent variable but also for random effects. Outliers can violate this assumption, making an ad-hoc treatment of the outliers a crucial aspect of the analysis (Chambers et al., 2014; Jiang & Rao, 2020; Marchetti & Tzavidis, 2021; Schmid et al., 2016). Recently, numerous papers on SAE aim to show that non-Normal models lead to more accurate estimates when compared to the Normal ones (Diallo & Rao, 2018; Graf et al., 2019; Lyu et al., 2020; Sugawara & Kubokawa, 2019). However, researchers, especially those not affiliated with official statistical agencies, may encounter challenges in accessing individual-level covariates essential for non-normal models, primarily due to confidentiality restrictions.

The search for auxiliary information becomes increasingly difficult, especially for non-linear estimates. Indeed, for linear estimates with a Normal model, is sufficient to have auxiliary information at the area level. However, for non-Normal models and/or non-linear estimates, auxiliary information is required for each individual unit of the population. Individual information, even when available in national statistical institutes, is often subject to confidentiality restrictions. Moreover, the auxiliary information, when available and usable, is often demographic (age, sex, nationality) and has very little explanatory power for economic outcome variables. Several

¹ In this context, the terms ‘migrants’ and ‘foreigners’ are used synonymously, as are ‘native-born’ and ‘Italians’.

authors have examined the issue of missing covariates free from error, including [Fabrizi et al. \(2007\)](#), [Gao et al. \(2014\)](#), and [Würz et al. \(2022\)](#).

Secondly, non-Normal models often require longer computational times with respect to the Normal ones ([Molina et al., 2022](#); [Molina & Ferretti, 2012](#)). At the unit level, Monte-Carlo (MC) approximation is often used to estimate the outcome parameters and the parametric bootstrap used to obtain a measure of variability. Both of these techniques are based on computational algorithms that rely on repeated random sampling to obtain numerical results. These results improve with more replications of the algorithm and therefore with more computational effort.

Although the issue of covariates and that of computational burden may seem unrelated, their relationship is actually relevant. MC is a technique in which the value of the outcome parameters is obtained as the mean of a number of replicates. Each replicate is more computationally demanding with more population units. In unit-level SAE models, for each non-sampled unit in the MC, it is necessary to predict a value of the outcome variables based on the estimated regression parameters and the auxiliary information. In addition, the parametric bootstrap requires knowledge of the auxiliary information for the whole population to create the bootstrap population. At each loop of this bootstrap, MC estimation is required, which significantly increases computational time. For a comprehensive literature review, see [Tzavidis et al. \(2018\)](#) and [Sugasawa and Kubokawa \(2020\)](#).

The starting point of this article is the SAE model based on Generalized Additive Models for Location, Scale, and Shape (SAE-GAMLSS) proposed in [Mori and Ferrante \(2025\)](#). GAMLSS appear to align seamlessly with the objectives of SAE because of the ability to achieve substantial model fit by leveraging the most appropriate data distribution ([Rigby & Stasinopoulos, 2005](#)). The idea is that choosing the distribution that best fits the data, without introducing any source of bias through back-transformation, can lead to reliable estimates even when covariates are unavailable. GAMLSS enables the use of a wide variety of distributions. Furthermore, with GAMLSS it is possible to incorporate covariates (when available) and a distinct random effects for each parameter of the distribution and not only for the location parameter as usually done. Thus has the potential to enhance the overall fit of the distribution. In this context, the issue of missing covariates is particularly significant and evident. Our objective is to estimate non-linear indices using non-Normal models, where areas are defined based on a specific structure linked to geographical areas. The estimates are also restricted to foreigners. To the best of our knowledge, there are no publicly available covariates for this group without official statistical authorization.

We introduce a Simplified version of the SAE-GAMLSS model, referred to as Simplified SAE-GAMLSS.² The fundamental idea behind this approach is to use the closed-form expressions for the Gini, Theil, and Atkinson indices, well-known for most distributions used in economic analyses, in conjunction with GAMLSS. The challenge of limited auxiliary information is addressed by (i) selecting the distribution that best fits the data and (ii) incorporating a random effect for each distribution parameter.

The term ‘Simplified’ underscores two key advantages of this model. First, our approach eliminates the need for auxiliary information. Estimates are derived using the closed-form expressions of the indices, without relying on covariates or MC approximations. Second, we introduce a specific non-parametric bootstrap method that computes the Mean Square Error (MSE) more efficiently than the method used in SAE-GAMLSS. This simplified model offers significant advantages in scenarios where rapid estimates are required, such as in research studies or by official statistical offices. Simulations demonstrate that, even with these enhancements, the Simplified SAE-GAMLSS still meets the core objective of SAE i.e. increasing the reliability of direct estimates.

The article is structured as follows. Section 2 provides a review of the SAE-GAMLSS model and introduces the Simplified SAE-GAMLSS. Sections 3 and 4 present the results of the simulations, focusing on the reliability of estimates and computational efficiency. In Section 5, we estimate the Atkinson, Gini and Theil indices for both foreigners and Italians at the NUTS 2 (Nomenclature of Territorial Units for Statistics) level, with a specific focus on distinguishing between urban, peri-urban, and rural areas. The conclusion is provided in Section 6.

² In figures and tables, due to space reasons, we use the acronym SSAE-GAMLSS for the Simplified SAE-GAMLSS.

2 A simplified small area estimation based on GAMLSS

2.1 Generalized additive models for location, scale, and shape

GAMLSS assume independent observations y_i , $i = 1, \dots, n$, from a random variable Y , with Probability Density Function (PDF) $f(Y | \theta_i)$, conditional on $\theta_i^T = (\theta_{i1}, \dots, \theta_{ik}, \dots, \theta_{ip})$, a vector of p distribution parameters, $k = 1, \dots, p$. Under GAMLSS, each parameter can be expressed as a function of the explanatory variables. [Rigby and Stasinopoulos \(2005\)](#) define the original formulation of a GAMLSS as follows. Let $\mathbf{y}^T = (y_1, \dots, y_n)$ 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:

$$g_k(\theta_k) = \mathbf{X}^k \boldsymbol{\beta}_k + \sum_{m=1}^{M_k} \mathbf{Z}_m^k \gamma_m^k, \quad k = 1, \dots, p \quad \text{and} \quad m = 1, \dots, M_k, \quad (1)$$

where $\theta_k^T = (\theta_{1k}, \dots, \theta_{nk})$ is a vector of length n , $\boldsymbol{\beta}_k^T = (\beta_{1k}, \dots, \beta_{\rho_k k})$ is a parameter vector of length ρ_k , $\mathbf{X}^k$ 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. Let each parameter k in (1) allow for a combination of different types and numbers of additive random effects terms, γ_m^k , to be incorporated easily in the model. The functional form of $f(Y | \theta_i)$ is not bound to a specific family. In GAMLSS, it departs from the strict exponential family distribution assumption found in GLMs and GAMs and instead includes a wide range of distribution families, including highly skewed and/or kurtotic continuous and discrete distributions.

Although the number of parameters p is not subject to constraints, it is often assumed to be ≤ 4 since within these limits nearly every distribution can be accommodated within the GAMLSS framework. Typically, the first two parameters of the population distribution are the location (μ) and scale (σ) parameters, while any additional parameters, if present, are the shape of the distribution, such as skewness (v) and kurtosis (τ). The estimation of the parametric components in Model (1) is achieved through one of two distinct algorithmic procedures that maximize a penalized likelihood of the data (see [Rigby & Stasinopoulos, 2005](#)). The first method is the Newton–Raphson algorithm (RS), while the second is based on the Cole and Green algorithm (CG), which is a ‘transformation’ of the Fisher scoring algorithm. As noted by [Rigby and Stasinopoulos \(2005\)](#), a combination of the RS and CG algorithms may also be used. Both algorithms, when estimating the distribution parameters, operate within an outer and inner iteration framework (for further details, refer to [Rigby and Stasinopoulos \(2005\)](#), Appendix A or to [Rigby et al. \(2019\)](#) chapter 3). In conclusion, a back-fitting algorithm is employed to estimate the parameters $\boldsymbol{\beta}^k$ and γ_m^k . The back-fitting algorithm plays a crucial role in fitting the additive components within the GAMLSS framework. The quadratic penalties in the likelihood arise from the assumption of a normally distributed random effect on the linear predictor. In equation (1), it is assumed that γ_m^k follows a multivariate normal distribution, $\gamma_m^k \sim N_{q_{mk}}(\mathbf{0}, [\mathbf{G}_{mk}(\lambda_{mk})]^{-1})$, where $[\mathbf{G}_{mk}(\lambda_{mk})]^{-1}$ is the generalized inverse of a $q_{mk} \times q_{mk}$ symmetric matrix that may depend on a vector of hyperparameters λ_{mk} . For fixed λ_{mk} , the parameters $\boldsymbol{\beta}^k$ and γ_m^k are estimated by maximizing the following penalized likelihood function l_p :

$$l_p = \sum_{i=1}^n \log \{f(y_i | \theta_i)\} - \frac{1}{2} \sum_{k=1}^p \sum_{m=1}^{M_k} (\gamma_m^k)^T \mathbf{G}_{mk} \gamma_m^k, \quad i = 1, \dots, n. \quad (2)$$

2.2 A quick review of SAE-GAMLSS

In SAE, a finite population Ω of N units is divided into J sub-population $\Omega_1, \dots, \Omega_J$, called areas of size $N_1, \dots, N_J$, where $N = \sum_{j=1}^J N_j$. Identify with Y_{ij} the target variable and with y_{ij} the observation for $i = 1, \dots, n$ and $j = 1, \dots, J$. In the SAE context, the aim is to estimate area parameters in the form of $H_j = \zeta(Y_j)$, where $\zeta(\cdot)$ is a real measurable function and $\mathbf{Y}_j^T = (Y_{1j}, \dots, Y_{ij}, \dots, Y_{N_j j})$.

Design-based estimates of these area parameters are obtained using data from the sample s of size n drawn from the population Ω with sub-sample $s_j = s \cap \Omega_j$ from area j of size n_j , being $n = \sum_{j=1}^J n_j$. Following Mori and Ferrante (2025), SAE-GAMLSS are defined starting from (1) by considering area specific random effects and assuming $Z \equiv \mathbb{I}_n$.

$$\begin{cases} g_\mu(\mu_{ij}) = x_{ij}^\mu \beta^\mu + \gamma_j^\mu \\ g_\sigma(\sigma_{ij}) = x_{ij}^\sigma \beta^\sigma + \gamma_j^\sigma \\ g_v(v_{ij}) = x_{ij}^v \beta^v + \gamma_j^v \\ g_\tau(\tau_{ij}) = x_{ij}^\tau \beta^\tau + \gamma_j^\tau \end{cases}, \quad (3)$$

where x_{ij}^k represents a generic covariate used in the k th parameter of the distribution, referring to the i th unit in the j th area. In (3) random effects $\gamma_j^k \stackrel{iid}{\sim} N(0, \Psi_k)$ for $k = \mu, \sigma, v, \tau$ allow for differences among areas with reference to each parameter. The variance-covariance matrix Ψ_k involves the variance of the random effects σ_k^2 , giving $\Psi_k = \sigma_k^2 \mathbb{I}$ for independent random effects. The assumption of a distribution model $\mathcal{F}$ allows for both skewness and kurtosis in the conditional distribution of Y_{ij} :

$$Y_{ij} | \gamma_j^\mu, \gamma_j^\sigma, \gamma_j^v, \gamma_j^\tau \sim \mathcal{F}(Y_{ij} | \mu_{ij}, \sigma_{ij}, v_{ij}, \tau_{ij}), \quad i = 1, \dots, n_j \quad \text{and} \quad j = 1, \dots, J. \quad (4)$$

Estimates of the parameters can be obtained using one of the methods proposed by Rigby and Stasinopoulos (2005). Estimates lead to $\mathcal{F}(y_{ij} | \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{v}_{ij}, \hat{\tau}_{ij})$, $i \in s$ and $j = 1, \dots, J$, where $i \in s$ indicates that the parameters are estimated using the sample part of the population. In the SAE-GAMLSS, as usual for unit-level models, estimates for non-linear parameters are obtained using a MC approximation method based on predicted individual values. The MC is based on the following steps. First, the model $\mathcal{F}$ is fitted to the sample data, obtaining a consistent estimator $(\hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{v}_{ij}, \hat{\tau}_{ij}, \hat{\gamma}_j^k)$. Next, for each $i \in r$, the values of $(\hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{v}_{ij}, \hat{\tau}_{ij}, \hat{\gamma}_j^k)$ are predicted. Using these predicted values, an out-of-sample vector $^{(\ell)}\mathbf{y}_{ij}^r$ is generated for each $\ell = 1, \dots, L$, with L being large. This vector is drawn from $\mathcal{F}(y_{ij}^r | \hat{\mu}_{ij}, \hat{\sigma}_{ij}, \hat{v}_{ij}, \hat{\tau}_{ij})$ for $i \in r$. It is worth reiterating that, in order to obtain parameter estimates for units $i \in r$, one needs to know the auxiliary variables (for each parameter) for all units in the population. This requirement substantially limits the applicability of the method, while the proposed approach does not rely on such information. This clarification is particularly useful for readers who are less familiar with SAE methods. The sample data $\mathbf{y}_{ij}^s$ is then attached to this out-of-sample data, forming the area population vector $^{(\ell)}\mathbf{y}_{ij}^T = ((^{(\ell)}\mathbf{y}_{ij}^r)^T, (\mathbf{y}_{ij}^s)^T)^T$. With this combined vector, the target parameter $H_j^{(\ell)} = \zeta(^{(\ell)}\mathbf{y}_j)$ is calculated, where $^{(\ell)}\mathbf{y}_j^T = (^{(\ell)}y_{1j}, \dots, ^{(\ell)}y_{ij}, \dots, ^{(\ell)}y_{N_jj})$. The Monte Carlo approximation of $\hat{H}_j$ is then given by:

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

This MC technique involves running an algorithm a considerable number of times, often exceeding $L = 100$ iterations, which can result in significant computational time. For example, when estimating the Gini coefficient, it is necessary to obtain at least 100 different Gini coefficients. For each of these coefficients, it is essential to calculate the predicted values and, in particular, to arrange the population units on the basis of the dependent variable. The computational burden of the MC approximation increases significantly when it is included in a parametric bootstrap to estimate the MSE, a common practice in almost all unit-level models.

2.3 Prediction of area parameters under simplified SAE-GAMLSS

We propose a simplified version of the SAE-GAMLSS to estimate a general area parameter $H_j = \zeta(Y_j)$ when covariates are not available, i.e. when the matrix $x_{ij}^k = 1, \forall k$ and $\forall i$. In this specific case, equation (3) becomes:

$$\begin{cases} g_{\mu}(\mu_j) = \beta^{\mu} + \gamma_j^{\mu} \\ g_{\sigma}(\sigma_j) = \beta^{\sigma} + \gamma_j^{\sigma} \\ g_v(v_j) = \beta^v + \gamma_j^v \\ g_{\tau}(\tau_j) = \beta^{\tau} + \gamma_j^{\tau} \end{cases} \quad (5)$$

From (5), it is now clear that each parameter is only influenced by the random effects of the area. In contrast to the SAE-GAMLSS approach, in this case we employ a penalized weighted likelihood function to estimate (5) by incorporating sample weights w_{ij} . Using sample weights to estimate the model, essentially opting for a weighted penalized likelihood instead of an unweighted one, we can effectively include information from informative sampling.

Let us highlight two crucial initial statements: (i) nearly all indices can be expressed as integrals of a transformation of a PDF and (ii) when the parameters of the PDF are determined solely by an intercept and a random effect, all units within a given area share the same predicted values for the distribution parameters. In other words, the location parameter of an area, once the chosen distribution is estimated, will be equal to the intercept $\hat{\beta}^{\mu}$ plus the random effect specific to the area, and this value will be the same for all units within that area. The first statement implies that $\hat{H}_j$, the estimate of $H_j = \zeta(Y_j)$ can be obtained by solving an integral in the form of:

$$H_j = E[\zeta(Y_j)] = \int_a^b \zeta(Y_j) f(Y_j) dY_j, \quad j = 1, \dots, J, \quad (6)$$

where a and b are the limits of the integral to be chosen depending on the index. Based on the second observation, it can be gathered that all observations within the same area share identical values. Consequently, the subscript i can be omitted from (6). Notably, the Gini, Theil, and Atkinson indices possess closed-form expressions that can be derived from (6) for almost any probabilistic distribution suitable for household income/consumption. The closed-form expression has a central role in the Simplified SAE-GAMLSS. As mentioned earlier, the MC approximation finds widespread application in nearly all SAE unit-level models targeting the prediction of nonlinear values. In the particular context of missing covariates, we avoid the use of MC by employing the closed form of the selected index. Specifically, after estimating the parameters of (5) using random effects, each area exhibits distinct values for $(\mu_j, \sigma_j, v_j, \tau_j)$. These values can be directly inserted into the closed-form expression, yielding final estimates avoiding the MC algorithm and consequently resulting in a more rapid estimation. Table 1 presents the closed-form expressions for the Gini, Theil, and Atkinson indices for various commonly used distributions. The distributions considered include the Exponential, Gamma, GB2, Log-Normal, Normal, Pareto, Singh-Maddala, and Weibull distributions. All these distributions assume positive values for the variable under consideration, except for the Normal distribution, which is included here as a benchmark. As noted by Lian et al. (2024), in-fact, employing a distribution in which the variable can also take negative values may be challenging due to the definition of these inequality indices. In the SAE context, these established findings can be applied to calculate estimates for each specific area. Furthermore, estimates are feasible also for areas with no sampled units.

2.4 Non-parametric bootstrap for MSE estimation

MSE estimation is one of the main challenges in the SAE framework. Usually, the parametric bootstrap is used to obtain MSE estimates (Liu et al., 2022). The parametric bootstrap requires the construction of a bootstrap population and this can be very slow as it involves (i) predicting the parameters, (ii) drawing a pseudo-population of N units from a parametric distribution, and (iii) sampling a subset of $n \leq N$ from the bootstrap population.

Keeping in mind the results obtained by Field and Welsh (2007) and by Ng et al. (2013) for the non-parametric bootstrap with clustered data and following the idea of Rust and Rao (1996), we propose a non-parametric bootstrap to estimate the MSE that uses specific bootstrap sample weights. The use of specific sample weights is required by the weighted penalized likelihood function used to estimate GAMLSS. In the following, we distinguish individual sampled units i , $i = 1, \dots, n_j$, in area j of size n_j (with the population size of individuals equal to N and of

Table 1. Gini, Theil, and Atkinson indices for classical income/consumption distribution

	Atkinson (A_i^c)	Gini (G_i)	Theil (T_i)	References ^a
Index	$\begin{cases} 1 - \frac{1}{\mu} \left(\frac{1}{n_i} \sum_{i=1}^{n_i} y_i^{1-c} \right)^{\frac{1}{1-c}}, & 0 \leq c \neq 1 \\ 1 - \frac{1}{\mu} \left(\frac{1}{n_i} \prod_{i=1}^{n_i} y_i \right)^{\frac{1}{n_i}}, & c = 1 \end{cases}$	$\frac{1}{n} \left(n_i + 1 - 2 \frac{\sum_{i=1}^{n_i} (i-1)y_i}{\sum_{i=1}^{n_i} y_i} \right) \text{ for } y_i \leq y_{i+1}$	$\frac{1}{n_i} \sum_{i=1}^{n_i} \frac{y_i}{\mu_i} \log \left(\frac{y_i}{\mu_i} \right)$	
Dagum ^b	$1 - \frac{1}{\mu_i} [\mu_i^{1-c}]^{1/(1-c)}$	$\frac{1}{\frac{1}{v_j} B(2v_j+1/\sigma_i, 1-1/\sigma_i)} \left\{ \frac{G_1}{v_j} - \frac{G_2}{v_j+1/\sigma_i} \right\} \text{ for } \sigma_i \geq 1$	$\frac{\Psi(v_j+1/\sigma_i) - \Psi(1-1/\sigma_i)}{\sigma_i} - \frac{\log(\Gamma(v_j+1/\sigma_i)\Gamma(1-1/\sigma_i))}{\Gamma(v_j)}$	Graf and Nedyalkova (2014) and Chotikapanich et al. (2018)
Exponential	-	0.5	$1 - \mu_i^{-1}$	Lubrano (2013)
Gamma	-	$\frac{2\sigma_i^2 + 1}{\Gamma\left(\frac{2}{\sigma_i^2}\right)}$	$\Psi(1 + \sigma_i^2) - \log(\sigma_i^{-2})$	McDonald and Jensen (1979)
GB2 ^b	$1 - \frac{1}{\mu_i} [\mu_i^{1-c}]^{1/(1-c)}$	$\frac{B(2v_j+1/\sigma_i, 2\tau_j-1/\sigma_i)}{B(v_j, \tau_j)B(v_j+1/\sigma_i, \tau_j-1/\sigma_i)} \left\{ \frac{G_1}{v_j} - \frac{G_2}{v_j+1/\sigma_i} \right\} \text{ for } \tau_j \geq 1/\sigma_i$	$\frac{\Psi(v_j+1/\sigma_i) - \Psi(\tau_j-1/\sigma_i)}{\sigma_i} - \frac{\log(\Gamma(v_j+1/\sigma_i)\Gamma(\tau_j-1/\sigma_i))}{\Gamma(v_j)\Gamma(\tau_j)}$	Graf and Nedyalkova (2014) and Chotikapanich et al. (2018)
Log-Normal ^c	$1 - \exp\left(-\frac{\omega_i^2}{2}\right)$	$erf(\sigma_i/2)$	$\sigma_i^2/2$	Jobe (1989)
Normal ^d	$1 - \left(1 + \frac{(1-c)(-c)\sigma_i^2}{2\mu_i^2}\right)^{\frac{1}{1-c}}$	$0.56 \frac{\sigma_i}{\mu_i}$	$0.5 \frac{\sigma_i^2}{\mu_i^2}$	Lian et al. (2024)
Pareto ^e	$1 - \frac{\sigma_i-1}{\sigma_i} \left[\frac{\sigma_i}{\sigma_i+c-1} \right]^{1/(1-c)}$	$(2\sigma_i - 1)^{-1}$	$\log(1 - 1/\sigma_i) + (\sigma_i - 1)^{-1}$	Lubrano (2013)
Singh-Maddala ^b	$1 - \frac{1}{\mu_i} [\mu_i^{1-c}]^{1/(1-c)}$	$\frac{1}{\frac{1}{\tau_j} B(2+1/\sigma_i, 2\tau_j-1/\sigma_i)} \left\{ G_1 - \frac{G_2}{v_j+1/\sigma_i} \right\} \text{ for } \tau_j \geq 1/\sigma_i$	$\frac{\Psi(1+1/\sigma_i) - \Psi(\tau_j-1/\sigma_i)}{\sigma_i} - \frac{\log(\Gamma(1+1/\sigma_i)\Gamma(\tau_j-1/\sigma_i))}{\Gamma(\tau_j)}$	Graf and Nedyalkova (2014) and Chotikapanich et al. (2018)
Weibull	-	$1 - 2^{-1/\sigma_i}$	$\sigma_i^{-1} \Psi(1 + 1/\sigma_i) - \log(\Gamma(1 + 1/\sigma_i))$	Lubrano (2013)

^a $\Gamma(\cdot)$ is the Gamma function, $erf(\cdot)$ is the Error function and $\Psi^m(\cdot)$ is the Polygamma function of order m (for $m = 0$, $\Psi^0(\cdot) = \Psi(\cdot)$ is the Digamma function).
^b The quantity G_1 and G_2 involves the generalized hypergeometric function see: Graf and Nedyalkova (2014).
^c GAMLSS function: LOGNO.
^d If the Normal distribution is assumed, it is necessary to ensure that at least $\mu - 3\sigma > 0$, because there is a 99.86% chance that y is greater than 0 (Lian et al., 2024).
^e GAMLSS function: PARETO10.

households equal to N^H) and sample weights w_i from the sampled households. Using the superscript b for all the ‘elements’ linked to households we have: n_j^b units, $i^b = 1, \dots, n_j^b$, in area j of size $n_j^b : n_j^b \leq n_j$ and sample weights w_{i^b} . For simplicity, we denote $H_j(Y_{ij})$ as H_j and $\widehat{H}_j(Y_{ij})$ as $\widehat{H}_j$. The MSE of $\widehat{H}_j$ is defined as:

$$\text{MSE}_j = E[(\widehat{H}_j - H_j)^2] \quad \text{for } j = 1, \dots, J, \quad (7)$$

and can be estimated based on the following steps:

1. Split the data-set into j sub data-sets one for each area.
2. For each sub data-set with re-sample $n_j^b - 1$ households with replacement.
3. Create a bootstrap sample by combining the sub-datasets and refer to it as S^* . Given the replacement, adjust the household weight using a re-weighting adjustment factor to account for the number of times household i^b has been selected. The weighting adjustment factor is given by:

$$G_{i^b} = \frac{n_j^b}{n_j^b - 1} \times H_{S_{i^b}},$$

where $H_{S_{i^b}}$ denotes the number of times household i^b is selected in the resample.

4. Compute the bootstrap sampling weight as:

$$w_{i^b}^* = G_{i^b} w_{i^b}.$$

5. Account for the calibration by calibrating the weights $w_{i^b}^*$ on the known totals N^b . This leads to the bootstrap calibrated weights:

$$w_{i^b}^c = w_{i^b}^* (1 + \mathbf{1}^T \lambda^*),$$

with

$$\lambda^* = \left(\sum_{i^b \in S^*} w_{i^b}^* \right)^{-1} \left(N^b - \sum_{i^b \in S^*} w_{i^b}^* \right).$$

6. Expand the sample S^* from households to units and fit the model.
7. Obtain the bootstrap estimator of H_j through the proposed method and denote it as $\widehat{H}_j^*$.
8. Repeat steps [2–7] a large number of times B .

Let $\widehat{H}_j^*(b)$ the estimator obtained in b th replicate of the bootstrap procedure, $b = 1, \dots, B$, the bootstrap mean square-error estimator of $\widehat{H}_j$ is given by:

$$\widehat{\text{MSE}}_B(\widehat{H}_j) = \frac{1}{B-1} \sum_{b=1}^B \left[\widehat{H}_j^*(b) - \frac{1}{B} \sum_{b=1}^B \widehat{H}_j^*(b) \right]^2. \quad (8)$$

The bootstrapping procedure deserves some explanation. Firstly, we use households rather than individuals as resampling units (Bessonneau et al., 2020; Luus et al., 2012) because classical consumption and income surveys usually use households as sample units. The definition of a specific individual weight is not necessary because the most common surveys report data for all household units, i.e. within the same household, the individual weights for each unit are (often) equal to the household weight. Secondly, a correction for non-response can be applied to the weights between steps 4 and 5.

3 Design-based simulations

Design-based simulations are carried out to analyse the properties of the proposed estimators. The HBS plays the role of a pseudo-population which is kept fixed. This survey, conducted by the Italian Statistical Institute (ISTAT), is based on a sample of approximately 18,000 households, representative of all Italian households. The data are collected using a two-stage sampling design, with municipalities as the first stage and households as the second stage. We focus on the three inequality indices presented thus far, which are calculated using per-capita expenditure based on equivalent consumption, obtained through the equivalence scale officially adopted by ISTAT for consumption data: the Carbonaro equivalence scale (Cuttillo et al., 2020). In accordance with the next section, we set up a simulation based on 49 small areas (NUTS 2 cross-classified as urban, peri-urban and rural areas). The sample size ranges from a minimum of 2 to a maximum of 68. The Simplified SAE-GAMLSS is computed using the *saegamlss* package available on GitHub.

3.1 From the choice of the distribution to the models compared

Starting with the HBS data, we decided to compare the distribution presented in Table 1 (for a discussion on the distribution of consumption expenditure, among other, see: Battistin et al., 2009; Prieto-Alaiz, 2007; Ubaidillah et al., 2018). We use the Generalized Akaike Information Criterion (GAIC) to evaluate the fitting of the distributions. The Figure 1, that reports the values of the GAIC for different penalization value k , clearly shows that GB2 and Log-Normal are the distributions with the lowest GAIC values. As the penalization value increases, as seen in the zoomed-in section of Figure 1, the Log-Normal distribution emerges as the best choice. In particular, the GAIC for the Log-Normal distribution becomes lower than those of the GB2 when $k \geq 4.5$.

Furthermore, aiming to use the closed-form expressions of the inequality indices, we have analysed the significance of each distribution parameter (Table 1). To simplify, consider the Log-Normal distribution, where only the σ parameter influences the three indices. This means that only covariates (intercept in absence of covariates) and the random effect related to σ can alter and increase the accuracy of the estimates. Similarly, taking into account the roles of ν and τ in defining inequality indices with GB2, it becomes somewhat concerning that the variance of the random effect is not statistically different from zero (Table 2). Due to these considerations, we are inclined to suggest that the Log-Normal distribution might be the most appropriate choice.

Based on these preliminary results, we decided to include the following methods in the simulation study: (i) the Direct estimator (for the direct estimation of inequality indices under complex design, as well as the estimation of their variance using closed-form expressions, see: Barabesi et al. 2016, 2024; Langel & Tillé, 2013); (ii) a SAE-GAMLSS model, to assess the influence of covariates on the estimates; (iii) a simplified SAE-GAMLSS model assuming a Log-Normal distribution; (iv) a unit-level model that restores normality through the Box-Cox transformation (Kreutzmann et al., 2019; Rojas-Perilla et al., 2020; Sugawara & Kubokawa, 2019), referred to hereafter as BC-SAE; and (v) the Fast Empirical Best (FEB) estimator, proposed by Molina and Ferretti (2012) and Chandra et al. (2018), which is designed to address the same computational challenges considered in this study.³

For SAE-GAMLSS and BC-SAE we use the following covariates (known only for the synthetic population used in this simulation): sex, age, education level, and primary source of income. In SAE-GAMLSS, for the location parameter μ all estimated regression parameters are statistically different from 0, while for σ only the parameter for primary source of income is statistically different from 0. The main aims of this simulation are to evaluate if Simplified SAE-GAMLSS performs well and what is the gap, paid in terms of reliability, between model with covariates, such as SAE-GAMLSS and/or BC-SAE, and Simplified SAE-GAMLSS.

From the simulation, we expect to be able to discuss the performance of the proposed Simplified SAE-GAMLSS model for the three inequality indices (Gini, Theil, and Atkinson with an inequality

³ Here, we use a non-parametric bootstrap to estimate the MSE of the FEB, because the original parametric bootstrap defined by Molina and Ferretti (2012) uses the covariates for the whole population and it is unfeasible here.

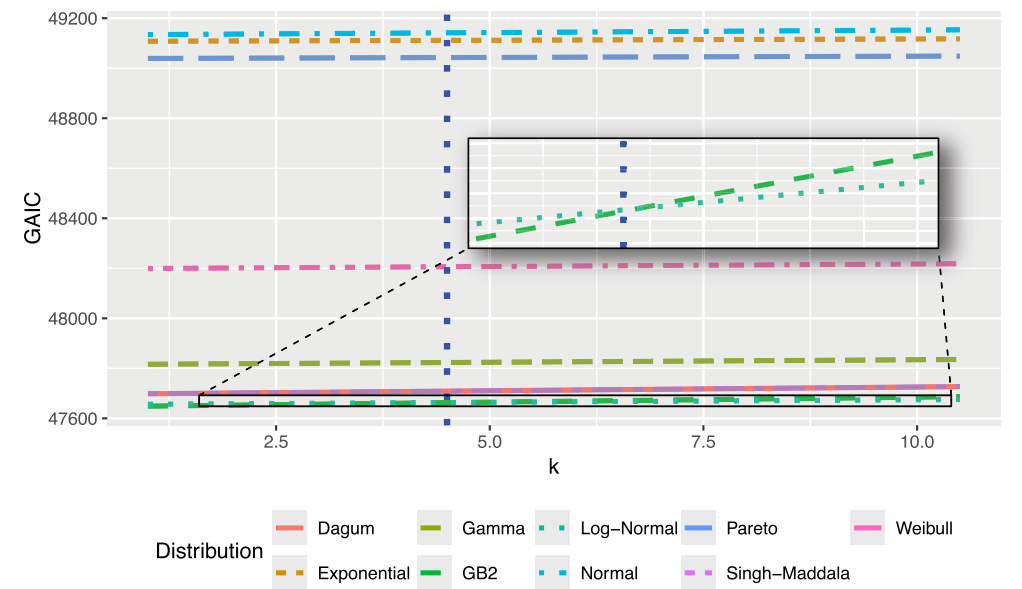


Figure 1. Generalized Akaike Criterion of selected distributions.

Table 2. Summary of regression parameter estimates. GB2 and Log-Normal distributions.

	GB2	Log-Normal
Regression parameters		
$\hat{\beta}^{\mu}$	6.255***	6.660***
$\hat{\beta}^{\sigma}$	0.750***	−0.715***
$\hat{\beta}^{\nu}$	1.267***	–
$\hat{\beta}^{\tau}$	0.590***	–
Variance of random effect		
$\hat{\sigma}_{\mu}^2$	0.249***	0.222***
$\hat{\sigma}_{\sigma}^2$	0.232***	0.223***
$\hat{\sigma}_{\nu}^2$	0.0003	–
$\hat{\sigma}_{\tau}^2$	0.0002	–

Note. Signif. codes: 0 ***0.001 **0.01 *0.05.

aversion parameter equal to 1). Additionally, we avoid using an area-level model because it presents two challenges: (i) lack of sample design information (i.e. strata and primary selection units) and (ii) direct estimates for inequality indices can be severely biased with small sample sizes, as shown by [De Nicolò et al. \(2023\)](#).

3.2 Performance measures

In order to compare results, we consider indices that assess both accuracy in terms of variability and bias. The following measures of performance averaged over areas are used: Average Relative Bias (ARB), Absolute Average Relative Bias (AARB), Average Coefficient of Variation (ACV), and Average Relative Root Mean Square Error (ARRMSE). Let us generalize the notation used in Section 2.2 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 ($t = 1, \dots, T$). We define time by the time the estimator considered whether it is model-based or design-based.

$$\begin{aligned} \text{ARB} &= J^{-1} \sum_{j=1}^J \left\{ H_j^{-1} \left(T^{-1} \sum_{t=1}^T \hat{H}_{jt} \right) - 1 \right\} \times 100 \\ \text{AARB} &= J^{-1} \sum_{j=1}^J \left| \left\{ H_j^{-1} \left(T^{-1} \sum_{t=1}^T \hat{H}_{jt} \right) - 1 \right\} \right| \times 100 \\ \text{ACV} &= J^{-1} \sum_{j=1}^J \left(T^{-1} \sum_{t=1}^T \frac{\sqrt{\widehat{\text{MSE}}_B(\hat{H}_{jt})}}{H_{jt}} \right) \times 100 \\ \text{ARRMSE} &= J^{-1} \sum_{j=1}^J \left(\sqrt{T^{-1} \sum_{t=1}^T \frac{\widehat{\text{MSE}}_B(\hat{H}_{jt})}{H_{jt}^2}} \right) \times 100, \end{aligned}$$

where $\widehat{\text{MSE}}_B(\hat{H}_{jt})$ is the bootstrap MSE defined in (8).

3.3 Simulation results

3.3.1 Model with covariates

Let us begin by comparing the three SAE models that incorporate covariates, namely BC-SAE, FEB, and SAE-GAMLSS, among themselves and with respect to the Direct estimator. Table 3 summarizes all previously reported measures of performance. The negative ARB showed by FEB can be the effect of either underestimation of variability in (both) the tail of the distribution or the back-transformation bias (Molina & Martin, 2018; Schluter, 2012). On the other hand, SAE-GAMLSS shows negative ARBs while BC-SAE exhibits a positive bias for the Gini and Atkinson indices, while showing a negative bias for the Theil index. In terms of bias, the BC-SAE outperforms the SAE-GAMLSS model for the Gini index, but performs worse for the other two indices. As expected, by leveraging covariates and information from the entire population, BC-SAE and SAE-GAMLSS achieve better performance than FEB. Model-based estimators demonstrate ARBs that, in most cases, are comparable to or even lower than those of the Direct estimator.

The bias has also been analysed in terms of the AARB. This indicator accounts for the possible compensation between negative and positive values, which may result in the ARB being close to zero even when the individual RBs are high. The values reported in Table 3 show that the Direct estimator has the highest AARB, except for the Atkinson index, for which the highest value is observed with the BC-SAE. SAE-GAMLSS consistently demonstrates the best performance across all situations.

Regarding the reliability measures (ACV, ARRMSE), Table 3 shows that, as expected, the ACV of direct estimates is the highest, and the SAE-GAMLSS and BC-SAE substantially reduce the variability of the estimates. Between SAE-GAMLSS and BC-SAE, the differences are almost negligible for the Gini and Theil indices, while they are more pronounced in favour of SAE-GAMLSS for the Atkinson index. FEB performs adequately for the Theil and Atkinson indices but worse than the Direct estimator for the Gini index.

From these initial results, it is evident that SAE-GAMLSS outperforms the other models that incorporate covariates and, as expected, also performs better than the Direct estimator. The difference between SAE-GAMLSS and BC-SAE can be mainly attributed to the greater flexibility of SAE-GAMLSS, its ability to include covariates in the scale parameter, and the avoidance of back-transformation bias. In the following analysis, we compare the Simplified SAE-GAMLSS with both the full SAE-GAMLSS model and the Direct estimator.

3.3.2 Simplified SAE-GAMLSS performance

Table 3 summarizes the performance measures, including those for the Simplified SAE-GAMLSS. SAE-GAMLSS and Simplified SAE-GAMLSS display distinct patterns in their ARBs. As previously noted, SAE-GAMLSS tends to produce negative ARBs, whereas the Simplified SAE-GAMLSS yields positive ARBs. This difference can be attributed not only to the use of covariates but also

Table 3. Measures of performance (%)

	Direct	BC-SAE	FEB	SAE-GAMLSS	SSAE-GAMLSS
Atkinson index					
ARB	−0.76	2.23	−0.83	−0.40	0.85
AARB	1.42	2.25	0.83	0.60	1.20
ACV	35.73	11.05	27.69	9.65	18.78
ARRMSE	1.92	1.12	1.61	0.61	1.40
Gini index					
ARB	−3.14	0.74	−3.21	−1.33	2.02
AARB	4.18	1.92	3.21	1.72	2.69
ACV	16.11	6.59	18.07	4.94	9.19
ARRMSE	3.81	1.32	4.42	1.35	2.77
Theil index					
ARB	−1.65	−0.89	−1.85	−0.83	1.84
AARB	3.00	1.81	1.85	1.30	2.57
ACV	40.66	11.74	27.85	10.16	20.09
ARRMSE	4.48	1.88	3.31	1.32	3.07

to the definition of the ARB itself, which is strongly influenced by the balance between overestimation and underestimation. When compared to the Direct estimator, the ARBs of the Simplified SAE-GAMLSS are, in most cases, of a similar magnitude but with the opposite sign. However, the interpretation changes notably when considering the AARBs. While the ARB values of the Simplified SAE-GAMLSS and the Direct estimator are comparable in absolute terms, the AARBs of the Direct estimator are consistently larger than those of the Simplified SAE-GAMLSS. Once again, SAE-GAMLSS demonstrates superior performance, primarily due to its use of covariates.

Figure 2 provides a visual representation of the Relative Bias (RB) for each index across the domains. The figure shows that, for each index, the Direct estimator exhibits substantial bias, particularly when the sample size is very small (De Nicolò et al., 2023). As expected, Simplified SAE-GAMLSS displays greater fluctuations in RB compared to SAE-GAMLSS. However, neither model demonstrates a consistent positive or negative trend.

With regard to the reliability measures, as expected, the Simplified SAE-GAMLSS performs worse than the full SAE-GAMLSS model, but better than the Direct estimator. Interestingly, the Simplified SAE-GAMLSS model shows the lowest ACV value for the Gini index, while higher values are observed for the Theil and Atkinson indices. Figure 3 displays the CVs for each index across the domains. First, for large sample sizes, the Simplified SAE-GAMLSS appears to be consistent, as it tends to yield CVs similar to those of the Direct estimator. Second, the differences between SAE-GAMLSS and Simplified SAE-GAMLSS are once again evident. The Simplified SAE-GAMLSS achieves a notable reduction in CVs compared to the Direct estimator, across all indices and domains. As expected, however, it still underperforms relative to the SAE-GAMLSS model.

4 MSE estimator performance

In Section 2, we presented a non-parametric bootstrap method for estimating the MSE. Here, we explore the characteristic of this estimator, considering both its bias and its impact on reducing computational time.

4.1 Bias

In order to evaluate the performance of the proposed MSE estimator, following Chandra et al. (2018), we report the Average True MSE (ATMSE), the Average Estimated MSE (AEMSE), the Average Bias of the MSE (ABMSE), and the Percentage Coverage Rate (PCR):

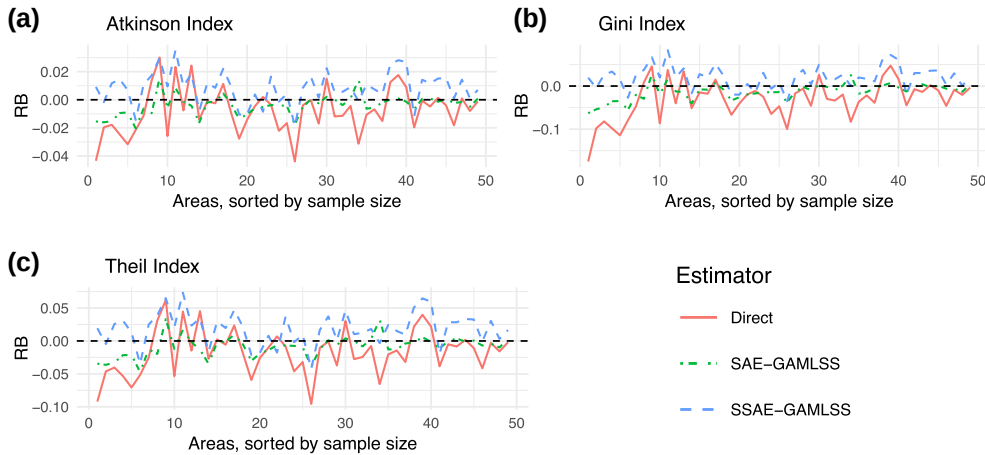


Figure 2. Relative bias for the three indices.

$$\begin{aligned}
 \text{ATMSE} &= J^{-1} \sum_{j=1}^J \left\{ T^{-1} \sum_{t=1}^T (\hat{H}_{jt} - H_{jt})^2 \right\} \times 100 \\
 \text{AEMSE} &= J^{-1} \sum_{j=1}^J \left\{ T^{-1} \sum_{t=1}^T \widehat{\text{MSE}}_B(\hat{H}_{jt}) \right\} \times 100 \\
 \text{ABMSE} &= J^{-1} \sum_{j=1}^J \left\{ T^{-1} \sum_{t=1}^T \widehat{\text{MSE}}_B(\hat{H}_{jt}) - \text{MSE}(H_j) \right\} \times 100 \\
 \text{PCR} &= J^{-1} \sum_{j=1}^J \left\{ T^{-1} \sum_{t=1}^T I \left(|\hat{H}_{jt} - H_{jt}| \leq 1.96 \sqrt{\widehat{\text{MSE}}_B(\hat{H}_{jt})} \right) \right\} \times 100,
 \end{aligned}$$

where $\text{MSE}(H_j) = T^{-1} \sum_{t=1}^T (\hat{H}_{jt} - H_{jt})^2$.

Table 4 reports the value of the indices referred to the non-parametric bootstrap. It is evident that the MSE estimators performs quite well. The ATMSE and the AEMSE are very similar, especially for the Atkinson index. In addition, the ABMSE is low, particularly for the Atkinson index, which is the only index without a negative ABMSE. Even though the PCR largely confirms the results obtained with the previous three indices, it is important to note that, for both the Gini and Theil indices, the coverage is far from the nominal 95%. It is worth noting that the assumption of normal confidence intervals, commonly adopted in SAE applications (see, among others, Chandra et al., 2018; Liu et al., 2022), could be replaced by non-parametric approaches, such as the percentile bootstrap, which is occasionally used in SAE (Chatterjee et al., 2008). Further studies are necessary to understand which type of confidence interval should be employed, and whether other bootstrap methods not yet applied in SAE, such as the bias-corrected and accelerated bootstrap method, might prove useful.

4.2 Computational times

Let us first consider the practical aspects of computation times. The results presented in the simulations were generated using a shared server equipped with 128 cores and 2TB of RAM. Naturally, the number of cores, the RAM capacity, and the type of processor used by the server have a significant impact on computation times.

Figure 4 shows the mean computation times required to estimate the MSE simultaneously for the three indices over 200 iterations for BC-SAE, FEB, SAE-GAMLSS, and Simplified SAE-GAMLSS. All models exhibit a significant reduction in computation time compared to

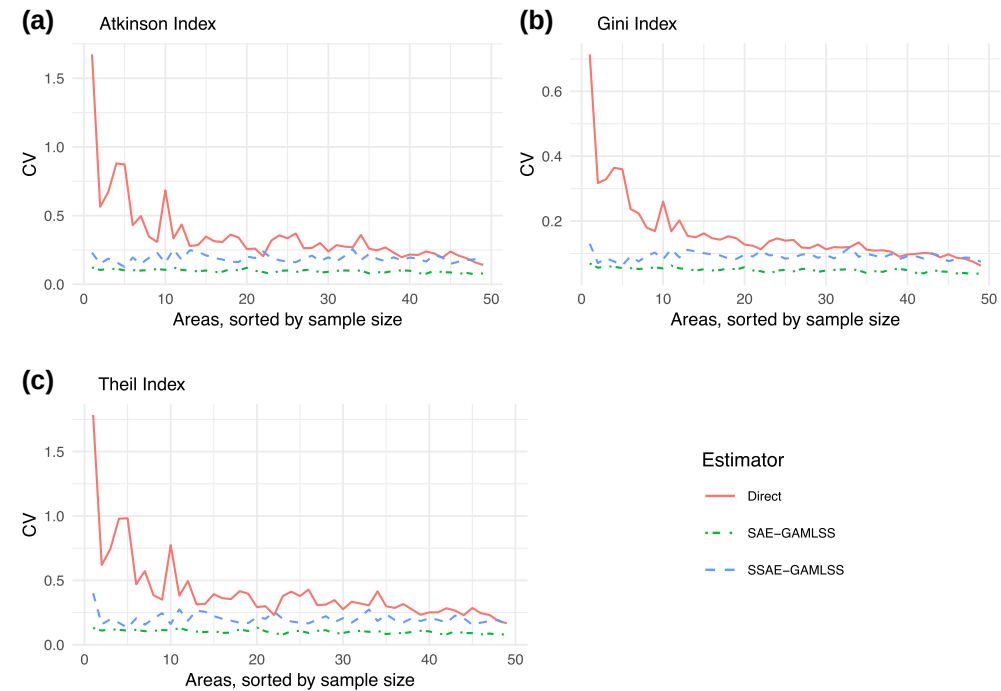


Figure 3. Coefficient of variation for the three indices.

Table 4. MSE: Measures of performance

SSAE-GAMLSS			
	Atkinson index	Gini index	Theil index
ATMSE	0.022	0.098	0.110
AEMSE	0.022	0.081	0.102
ABMSE	0.015	−0.270	−0.749
PCR	96.312	88.726	89.866

SAE-GAMLSS. Specifically, Simplified SAE-GAMLSS reduces computation time by 72 % relative to SAE-GAMLSS, FEB by 91 %, and BC-SAE by 81 %. Regarding speed, BC-SAE operates marginally faster than the Simplified SAE-GAMLSS, exhibiting a median computation time of approximately 2 min per bootstrap run for the parametric transformation, whereas the Simplified SAE-GAMLSS takes around 3 min. However, as shown in Figure 4, the latter exhibits a longer right tail in its distribution of computation times.

As emphasized by [Arora and Barak \(2009\)](#), it is crucial to assess the complexity of the algorithms used. To assess the complexity of the algorithms, we conducted a new simulation exclusively focused on computation times. In this simulation, we compare the three model-based estimators based on the pseudo-population presented in Section 3 with varying sample sizes, ranging from a minimum of 1,000 units to a maximum of 100,000. Re-sampling from the pseudo-population with replacement allowed us to achieve these sample sizes. Figure 4 shows the mean time as a function of the sample size required for each algorithm to perform a single loop of the MSE estimator.⁴ The complexity of the SAE-GAMLSS MSE estimator is approximable by a function in the form of

⁴ The x-axis is truncated to 4,000 for better visualization of all three curves.

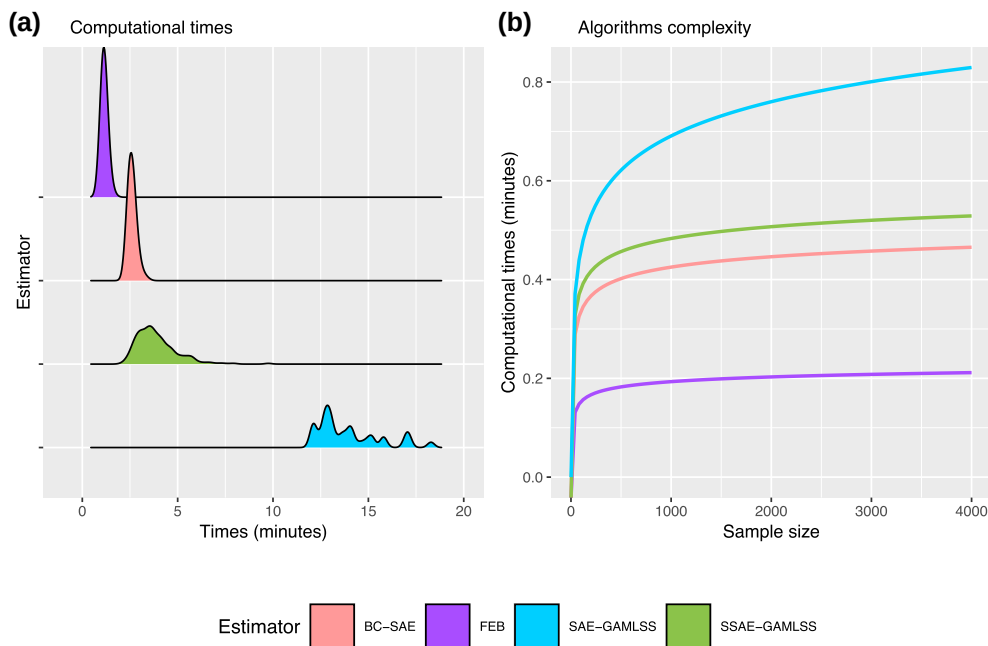


Figure 4. Computational times and algorithm complexity.

$\log(n)$, a relatively fast algorithm. On the other hand, the complexity of Simplified SAE-GAMLSS is approximable by a function in the form of $\log(\log(n))$, which is even faster than $\log(n)$. In summary, the Simplified SAE-GAMLSS remains slightly slower than FEB and BC-SAE but is considerably faster than SAE-GAMLSS.

5 Mapping inequality heterogeneity for foreigners and Italians

In this section, we use the Simplified SAE-GAMLSS to estimate the three inequality indices for foreigners and Italian citizens in 49 different areas. The areas are defined as the Italian NUTS 2 (20 administrative Regions), each of which is further subdivided into three subgroups: urban, peri-urban and rural (Cuttillo et al., 2020). Urban areas are metropolitan centres, peri-urban areas are cities with more than 50,000 inhabitants and those included in the metropolitan areas but outside the centre. All other areas are rural. Not all NUTS 2 have all three groupings. The migrants in question are regular migrants who already have a residence permit. Those in an irregular status, those waiting for a permit or those who own property in Italy without having a residence are not included in the sample and are therefore not considered in this discussion. The dependent variable is the equivalized consumption expenditure recorded in the HBS.

Table 5 presents the CVs of the direct estimates. According to Statistics-Canada (2007), the estimates for foreigners, having high CVs, are often affected by limited precision, whereas those for Italians are generally deemed reliable. As shown in Table 5, for foreigners, at least 25% of the estimates are fully reliable for Atkinson and Theil, and nearly 75% are reliable for Gini when considering a $CV \leq 16\%$ as the benchmark for full reliability. If we also include estimates with $16\% < CV < 33\%$, which are publishable with a warning, the proportion of usable estimates further increases. Nevertheless, for foreigners the limited sample size still poses challenges, making SAE methods valuable to improve the efficiency of the estimates, even when some are already considered reliable. By contrast, for Italians the HBS sample size supports reliable estimation of the indices.

Following Mori and Ferrante (2025), we perform a three-step procedure to select the distribution that best fits the data. Mori and Ferrante (2025) proceed as follows: (i) employ the GAIC method to narrow down potential distributions, (ii) determine covariates and random effects for each parameter in the selected distributions, and (iii) finalize the distribution selection based

Table 5. Summary of coefficient of variation (%) for direct estimates for the three inequality indices over the 49 areas

Index	Min.	25° Perc.	Median	Mean	75° Perc.	Max.
Foreigners						
Atkinson	8.1	14.6	20.4	25.2	31.4	74.4
Gini	4.4	7.9	11.8	15.4	17.7	55.2
Theil	8.8	15.8	22.7	26.9	32.3	70.0
Italians						
Atkinson	2.8	4.7	5.6	6.3	7.2	13.7
Gini	1.5	2.5	2.9	3.4	3.9	8.0
Theil	3.3	5.7	6.6	7.3	8.7	15.9

on K-fold cross-validation results. In this case, step (ii) is necessary only for selecting the random effects. The second point is that, while similar, our analysis differs slightly from the one in Section 3.1. In that section, we select the distribution that best fits the entire dataset for simulation purposes. Here, instead, we aim to predict only foreigners using the SAE model. Therefore, we have to follow the procedure outlined above to select the distribution that best fits this specific sub-sample of the HBS.

We begin by comparing several potential alternative distributions (see Table 1 for the potential distributions) based on the GAIC, using penalty values of $k \approx 4$ and $k \approx 6$. These values of k are chosen in accordance with Figure 1. The two distributions with the lowest GAIC values are the Log-Normal and GB2. Specifically, the GB2 distribution has the lowest GAIC ($\approx 45,933$) for $k = 4$, while the Log-Normal distribution has the lowest GAIC for $k = 6$ ($\approx 45,944$). After incorporating random effects for each parameter, we find that only those associated with the location and scale parameters are statistically different from zero. Based on this, we conduct K-fold cross-validation with $K = 7$. This choice follows Nti et al. (2021), who argue that $K = 7$ represents a suitable compromise between $K = 5$ and $K = 10$, providing faster results than the latter while ensuring greater accuracy than the former. The results suggest that the Log-Normal distribution performs best, with a global deviance of 42,450, compared to a global deviance of 42,842 for the GB2 distribution.

Table 6 reports the estimates of the regression parameters, the variance values of the random effects, and the diagnostic quantile residuals. All parameters are statistically different from 0. It is important to note here that the random effect of σ has a variance statistically different from 0. Only because of the random effects associated with this parameter, the estimates are different from one area to another. To conclude, looking at the quantile residuals diagnostic, the model seems to perform quite well.

5.1 Simplified SAE-GAMLSS estimates for foreigners

Table 7 shows the CVs for the three indices obtained with the Simplified SAE-GAMLSS estimator, obtained assuming the Log-Normal distribution, and the Average Coefficient of Variation Reduction (ACVR).⁵ The reduction in the CVs is consistent across all indices: approximately 24% for Atkinson and Theil, and more than 30% for Gini. Moreover, Figure 5 displays the CVs through box-plots and line-plots for each area and for each index. From the box-plots, it is evident that the Simplified SAE-GAMLSS, even without covariates, is able to outperform the Direct estimator. The line-plots highlight that in each area and for each index, the CV is lower for the Simplified SAE-GAMLSS than for the Direct estimator. As noted in the simulation, here again, the difference between the CV of direct estimates and the CV of Simplified

⁵ ACVR is defined as: $ACVR = \frac{1}{J} \sum_{j=1}^J \left(1 - \frac{CV(\widehat{H}_{SSAE-GAMLSS}^{SSAE-GAMLSS})}{CV(\widehat{H}_{dir}^{dir})} \right) \times 100$. Where the superscript SAE-GAMLSS or dir

indicates that the consider quantity is the CV computed for SAE-GAMLSS or Direct estimator.

Table 6. Summary of simplified SAE-GAMLSS regression and quantile residuals diagnostic.

Regression parameters	
$\hat{\beta}^{\mu}$	6.660***
$\hat{\beta}^{\sigma}$	−0.715***
Variance of random effect	
$\hat{\sigma}_{\mu}^2$	0.222***
$\hat{\sigma}_{\sigma}^2$	0.223***
Quantile residuals diagnostic	
Mean	−0.0011
Variance	1.000
Skewness	0.304
Kurtosis	3.1

Note. Signif. codes: 0 ***0.001 **0.01 *0.05.

Table 7. Summary of coefficient of variation for simplified SAE-GAMLSS estimates for the three inequality indices over the 49 areas (foreigners, %)

Index	Min.	25° Perc.	Median	Mean	75° Perc.	Max.	ACVR
Atkinson	4.0	11.9	14.8	17.5	21.9	35.5	24.3
Gini	3.8	6.3	7.9	9.5	11.9	23.01	30.6
Theil	4.0	12.4	15.2	18.8	23.00	40.28	24.7

SAE-GAMLSS is lower, the higher the sample size, and the difference becomes negligible for high n_i . The line-plots, also have two horizontal lines which, according to [Statistics-Canada \(2007\)](#), delineate three areas of the plot: (i) reliable estimates under the solid line ($CV \leq 16.6\%$), (ii) quite reliable estimates to be used with warning between the solid and the dashed line ($16.6 < CV \leq 33.3\%$), and (iii) unreliable estimates above the dashed line ($CV > 33.3\%$). Overall, by considering all the indices, using the Direct estimator results in 17% of the estimates being unreliable and 38% of the estimates requiring a warning. By using Simplified SAE-GAMLSS, these two values, the percentage of areas with unreliable estimates and those with ‘warning’ estimates, decrease to 7% and 20%, respectively. In particular, the figure clearly shows the Gini index yields the best results. In fact, when using Simplified SAE-GAMLSS, only one area (the one with the lowest sample size) exceeds the 16.6% threshold.

5.2 A comparison of inequality between Italians and foreigners

[Djajic \(2003\)](#) and [Barigozzi and Speciale \(2011\)](#) highlight the significance of consumption patterns as a crucial aspect of immigrants’ integration into their host community. As immigrants become more closely aligned with the native population, their integration in terms of consumption can unlock various market and social prospects for them. Extending this consideration to the inequality measures obtained by consumption data, we compare the three inequality indices for foreigners with those of Italians.

[Figures 6, 7, and 8](#) show the choropleth maps of the estimates of the Atkinson, Gini, and relative Theil indices for Italians and for foreigners. In particular, we report the relative Theil index computed by dividing the Theil index by its maximum possible value ($\log(N_i)$), where N_i is the sample size in each area.⁶ This allows for more meaningful comparisons across areas and with other inequality indices. [Tables A1, A2, and A3](#) in Appendix A report the estimated values and CVs of the

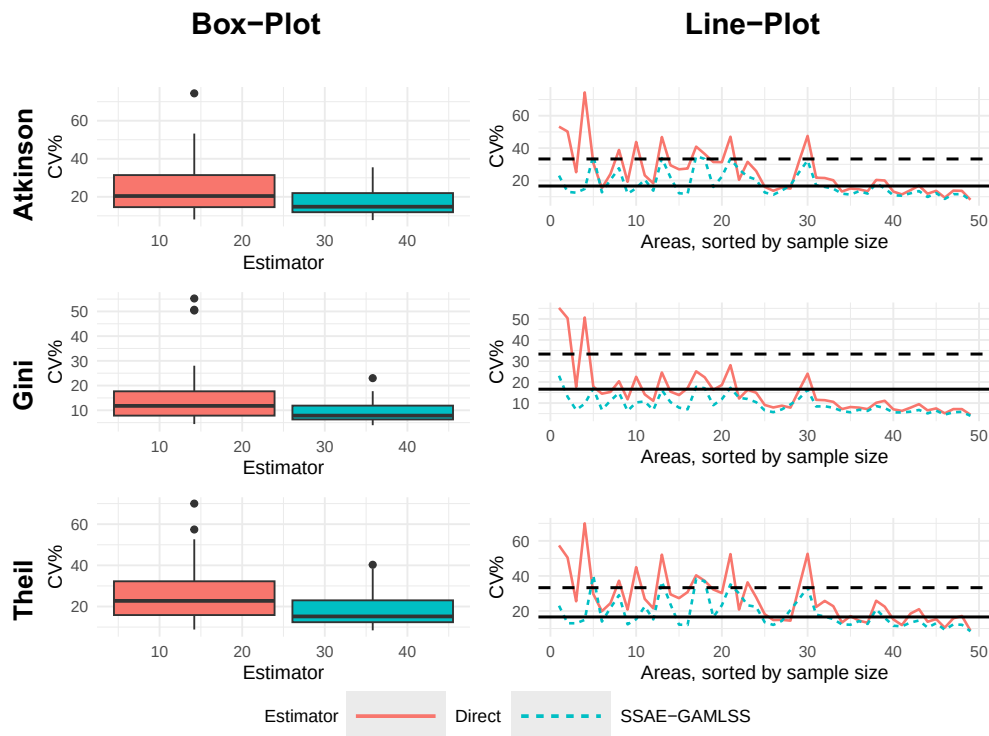


Figure 5. Comparison of the coefficient of variation. The horizontal lines delineate three areas of the plot: (i) reliable estimates under the solid line ($CV \leq 16.6\%$), (ii) estimates with warning between the solid and the dashed line ($16.6 < CV \leq 33.3\%$), and (iii) unreliable estimates above the dashed line ($CV > 33.3\%$) (Statistics-Canada, 2007).

indices for all areas. Table A4, also in Appendix A, presents the estimates of the relative Theil index. Since this index is a rescaled version of the original Theil index, the CVs remain unchanged and are therefore the same as those reported in Table A3.

The range for foreigners for the three indices is, respectively, 0.05–0.30, 0.18–0.45, and 0.00–0.03 while for Italians it is 0.07–0.17, 0.22–0.34, and 0.00–0.02. The heterogeneity between areas is more evident when considering the urban, peri-urban, and rural areas. Specifically, although the minimum value in these three areas remains almost unchanged, the maximum values of the indices undergo considerable variation. For example, for the Atkinson index, in urban areas, the maximum for foreigners (Italians) is around 0.30 (0.17), decreasing to 0.19 (0.14) in peri-urban areas and reaching a value of 0.16 (0.15) in rural areas. For the Gini index the same quantities values: 0.45, 0.35, and 0.32 for foreigners and 0.34, 0.31, and 0.31 for Italians. The relative Theil index has smaller values given that are narrowed down by the logarithm of the sample size.

In general, inequality, as measured by all indices, tends to be higher in urban areas compared to the other two types of areas. The disparity between foreigners and Italians is particularly pronounced in urban settings. Specifically, all indices indicate that inequality tends to be greater among foreigners than among Italians in the northern regions of Italy, whereas in the southern regions, it is higher among Italians. For instance, the metropolitan area of Turin records the highest values across all three indices (Atkinson ≈ 0.30 , Gini ≈ 0.45 , and relative Theil ≈ 0.03). In contrast, for Italians, the Gini index is approximately 0.25, while the other two indices remain below 0.15. However, in the southern regions, the estimates reveal a markedly different trend, where inequality is often greater among Italians than among foreigners.

Shifting the focus to peri-urban areas, all the indices, highlight that differences between the two groups are less pronounced. Among foreigners, the three indices follow a similar path from

⁶ The sample size N_j for each area can be obtained from the official ISTAT website.

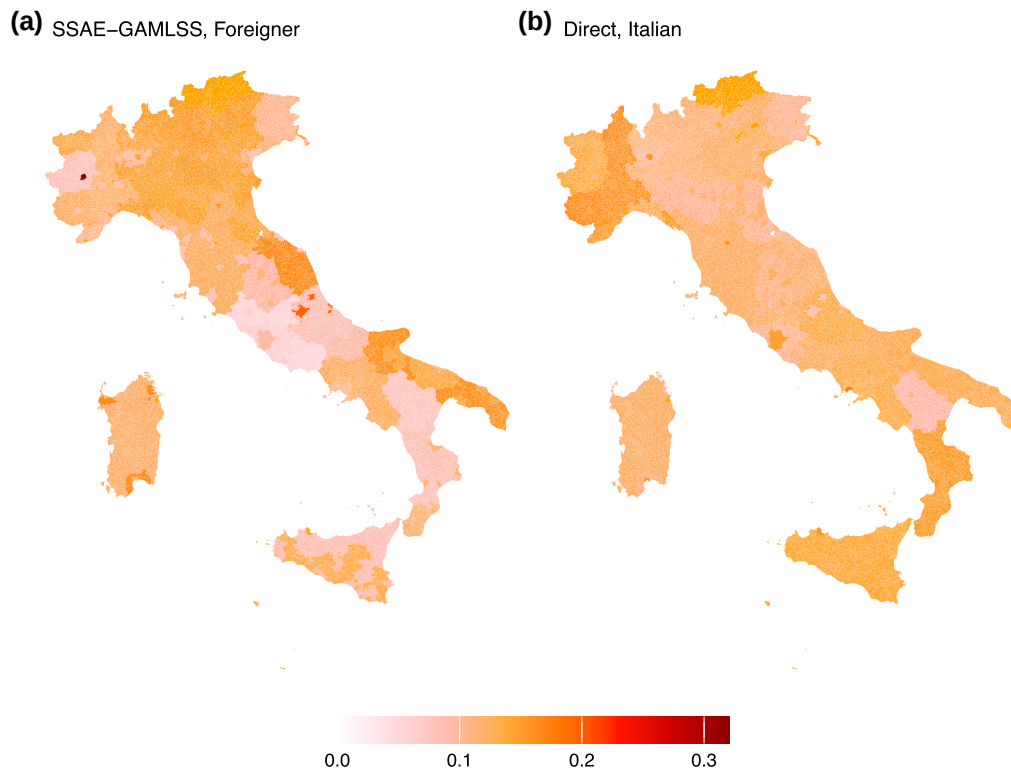


Figure 6. Estimates of the Atkinson index for Foreigners and Italians.

North to South as observed in urban areas. Furthermore, the gap between the two groups is narrower.

In rural areas, indices are nearly identical to those in peri-urban areas for Italians. However, for foreigners, inequalities slightly increase. There is no clear order of magnitude between rural and peri-urban areas; in some regions, inequality is greater in peri-urban areas, while in others, it is higher in rural areas.

In summary, inequality is generally higher in urban areas, especially among foreigners, with the North showing greater disparities for foreigners and the South for Italians. In peri-urban and rural areas, differences between groups are less pronounced, though slight increases in inequality among foreigners are observed in rural settings.

Let us analyse some specific areas, beginning with the two regions of Apulia and Campania in southern Italy. The urban areas of Bari and Naples consistently exhibit the lowest levels of inequality across all indices for foreigners, with disparities increasing in peri-urban areas and further escalating in rural areas. In rural areas, the Atkinson and relative Theil indices are substantially higher than in their urban counterparts. The Gini index in Bari (Apulia) is six percentage points lower than in the rural areas of the same region.

For Italians, the Atkinson index indicates very similar levels of inequality across the areas. Meanwhile, in Apulia, the Gini and relative Theil indices show nearly identical values in the three areas, whereas in Campania, higher inequality is observed in the urban area.

This disparity can be explained by the significant reliance on foreign labour, which is frequently employed in manual tasks, particularly in the rural areas of these two regions. This is especially evident in roles such as seasonal fruit and vegetable harvesting. Conversely, the metropolitan areas of these regions host two of the largest ports in the Mediterranean. In these areas, foreign workers tend to hold more skilled positions, which are consequently better compensated. This, in turn, has a noticeable impact on consumption inequality.

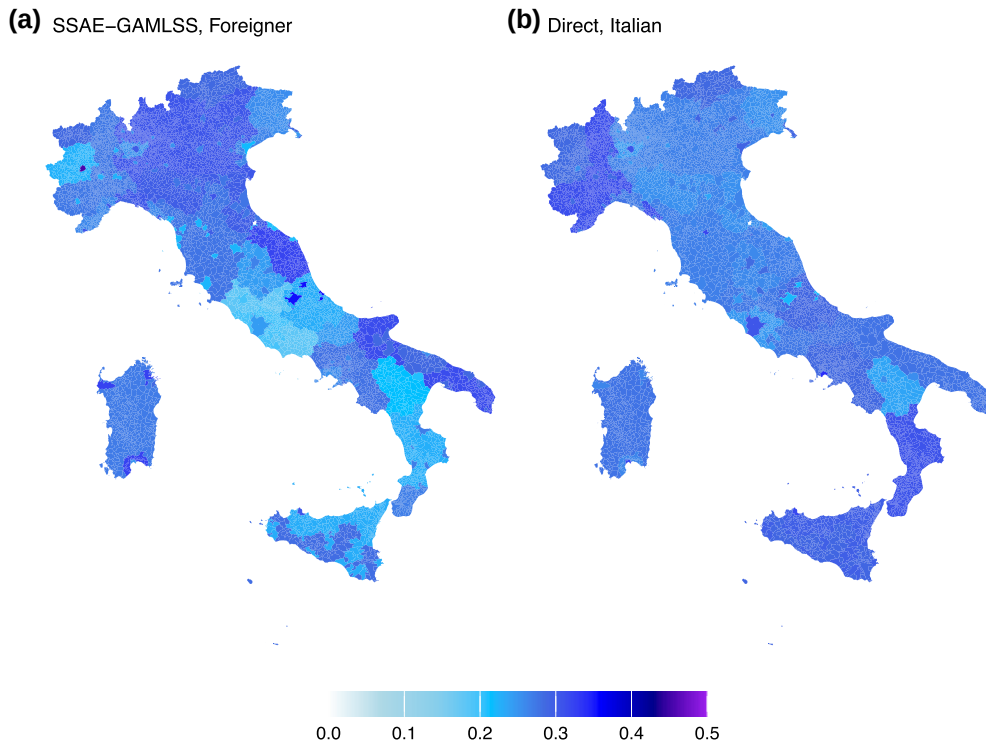


Figure 7. Estimates of the Gini index for Foreigners and Italians.

In the southern part of Italy, Sicily warrants special attention. This region exhibits an intriguing pattern of inequality among foreigners, with relatively low inequality levels in peri-urban areas (approximately 0.08 for the Atkinson, 0.0071 for the relative Theil indices, and 0.23 for the Gini index). However, two important considerations should be noted. First, due to its geographical location, Sicily is often chosen by migrants as an initial point of arrival, though it is not necessarily their preferred region for long-term settlement. Second, Sicily is unique in being the only region with more than one metropolitan city, hosting three in total. The presence of these three metropolitan areas considerably extends the urban space, resulting in a vast peri-urban zone and leaving only a limited portion of the territory classified as rural. For Italians, the Atkinson index shows lower values in rural areas, increases in peri-urban zones, and rises further in urban areas.

Finally, we focus on the analysis of the Lombardy region in northern Italy. In the metropolitan area of Milan, the Atkinson index stands at 0.13 for foreigners and 0.15 for Italians, while the Gini index is 0.30 and 0.31, respectively. While inequality decreases for both groups in peri-urban areas, the trend diverges in rural areas. Among foreigners, the Atkinson in rural areas is approximately 0.15, while the Gini index reaches 0.31, consistently higher than in urban areas and also exceeding the corresponding inequality indices for Italians in rural areas. An examination of the relative Theil index yields a similar conclusion: inequality declines for both foreigners and Italians in peri-urban areas, rising modestly in rural areas and more substantially in urban areas.

These are just a few examples of the conclusions that can be drawn from these estimates, which cannot be fully grasped without taking into account SAE.

6 Conclusion and policy implication

This article presents a Simplified variant of the SAE-GAMLSS unit-level model. Within this framework, the article outlines a method for generating small area predictions for three inequality

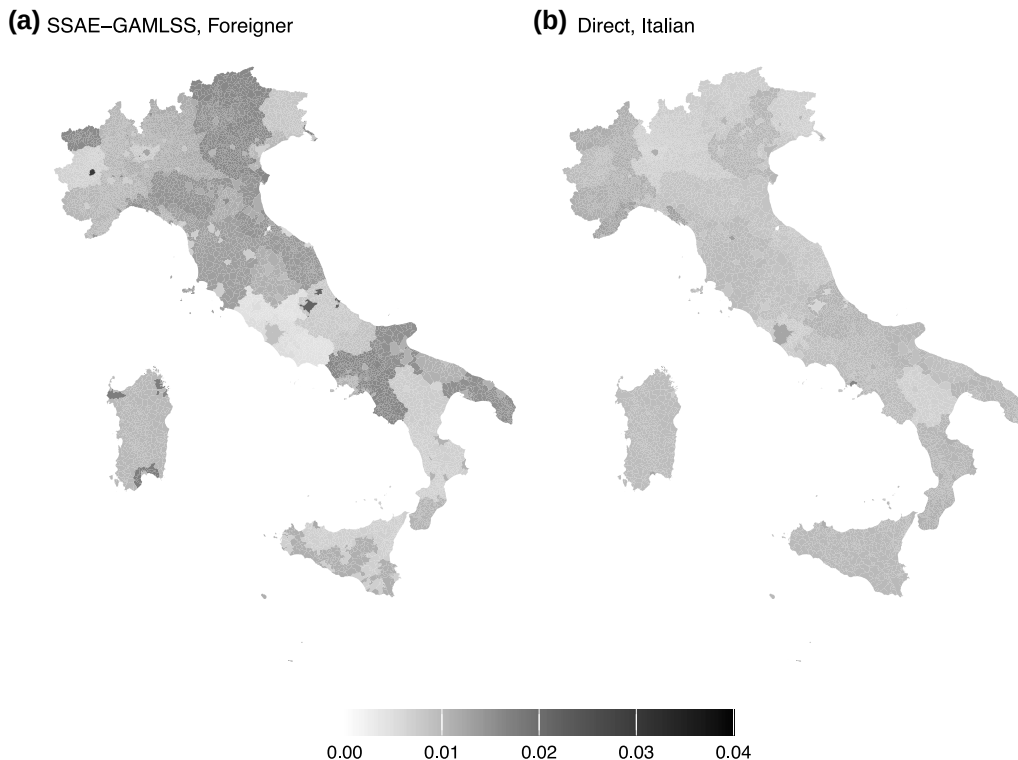


Figure 8. Estimates of the Relative Theil index for Foreigners and Italians.

indices: Gini, Theil, and Atkinson. In addition, the article explores the application of a non-parametric bootstrap approach to estimate the MSE associated with these predictions.

The Simplified SAE-GAMLSS proves particularly valuable in two distinct scenarios: (i) when researchers or official statistics lack access to covariate data due to privacy constraints and (ii) when researchers seek to obtain estimates while minimizing computational complexity and effort.

Our results indicate that the proposed estimator effectively reduces the variability of the Direct estimator. Furthermore, when comparing the Simplified SAE-GAMLSS to the SAE-GAMLSS, the former exhibits a satisfactory performance, as expected, given that it does not incorporate covariates. A particularly interesting development for future research would be to compare the Simplified SAE-GAMLSS with other non-Normal models, such as [Diallo and Rao \(2018\)](#), which assumes a Skew-Normal distribution, or with non-parametric and semi-parametric models, such as those proposed by [Chambers and Tzavidis \(2006\)](#); [Opsomer et al. \(2008\)](#). Furthermore, more in-depth analyses on the use of non-normal bootstrap confidence intervals are necessary.

Leveraging the Simplified SAE-GAMLSS, we proceed to estimate the three inequality indices for both Italian and foreign citizens across urban, peri-urban, and rural areas within administrative Regions. The resulting maps reveal substantial heterogeneity among the defined domains. The estimates of the Atkinson index for urban, peri-urban, and rural areas, obtained in the article, can be used to formulate location-specific policies aimed at promoting equity.

Regions generally show that migrants have higher levels of inequality than the native population, especially in urban areas. The aim of reducing this gap can be an important policy objective to resolve tensions between groups and to promote integration between them.

From an applied point of view the areas that deserves more attention are certainly those in the South of Italy, in particular the rural area of Apulia. These areas are sadly notorious for the exploitation of migrants in the agricultural sector ([Caterina & Berti, 2015](#)).

A medical reason could serve as a suggestion for future research directions. According to [Hamilton and Kawachi \(2013\)](#), there is evidence that the degree of economic inequality within

an area has a detrimental effect on health at the individual level. Building on this study, and recognizing that the Italian healthcare system is decentralized, with significant autonomy granted to the Regions, a potential area for future research is to examine changes in health expenditure by analysing consumption data and testing the association with inequality indices. An additional study could examine whether economic migration within Italy, generally from areas with a higher Atkinson index to those with a lower index, leads to improvements not only in economic but also in health outcomes.

Acknowledgments

We would like to thank ISTAT for kindly providing the HBS dataset. 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’.

Conflicts of interest: Authors declare no conflicts of interest.

Funding

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.

Data availability

The data supporting the findings of this study are available from the Italian National Institute of Statistics (ISTAT). Access to these microdata is restricted and requires authorization from ISTAT. Researchers may obtain access by submitting a formal request directly to ISTAT, subject to the institute’s terms and conditions. All code used for the analyses in this study is publicly available at <https://github.com/llorenxo/saegamlss>. The repository contains a detailed README describing the structure of the package and the workflow, from the creation of artificial populations to model selection and the estimation of small area indicators and their variability. Functions are fully documented with help files and examples, which illustrate their applicability and provide guidance for replication and further use.

Appendix

Table A1. Estimates of Atkinson index and coefficients of variation

Area	Foreigners	CV Foreigners	Italians	CV Italians
Piedmont (urban, Turin)	0.30	13.00	0.13	6.75
Piedmont (peri-urban)	0.08	23.00	0.10	6.27
Piedmont (rural)	0.12	12.47	0.15	6.08
Aosta Valley (rural)	0.13	14.80	0.13	4.59
Lombardy (urban, Milan)	0.13	20.38	0.15	8.06
Lombardy (peri-urban)	0.10	35.51	0.14	4.28
Lombardy (rural)	0.14	12.71	0.11	2.79

(continued)

Table A1. Continued

Area	Foreigners	CV Foreigners	Italians	CV Italians
Emilia Romagna (urban, Bologna)	0.14	15.21	0.08	10.73
Emilia Romagna (peri-urban)	0.12	11.84	0.11	5.60
Emilia Romagna (rural)	0.14	27.59	0.10	4.74
Friuli Venezia Giulia (peri-urban)	0.09	13.72	0.13	7.66
Friuli Venezia Giulia (rural)	0.10	20.60	0.10	5.51
Lazio (urban, Rome)	0.09	12.13	0.14	3.59
Lazio (peri-urban)	0.06	21.96	0.10	5.83
Lazio (rural)	0.05	33.86	0.12	4.73
Liguria (urban, Genoa)	0.13	33.30	0.12	5.21
Liguria (peri-urban)	0.14	35.04	0.12	7.48
Liguria (rural)	0.10	11.80	0.15	8.35
Marche (peri-urban)	0.07	16.02	0.10	11.53
Marche (rural)	0.16	22.40	0.11	5.51
Molise (rural)	0.09	33.49	0.12	3.78
Apulia (urban, Bari)	0.10	20.58	0.11	9.31
Apulia (peri-urban)	0.13	26.85	0.12	5.89
Apulia (rural)	0.15	22.41	0.12	4.87
Sardinia (urban, Cagliari, and Sassari)	0.10	13.78	0.16	9.22
Sardinia (peri-urban)	0.16	11.23	0.11	7.49
Sardinia (rural)	0.11	12.61	0.12	6.18
Sicily (urban, Palermo, Catania, and Messina)	0.14	32.41	0.15	5.37
Sicily (peri-urban)	0.08	25.02	0.13	4.99
Sicily (rural)	0.12	17.13	0.13	4.43
Tuscany (urban, Florence)	0.14	14.93	0.16	7.21
Tuscany (peri-urban)	0.08	16.05	0.11	6.10
Tuscany (rural)	0.12	16.93	0.12	5.59
Trentino South-Tyrol (peri-urban)	0.11	11.30	0.11	3.57
Trentino South-Tyrol (rural)	0.10	11.90	0.11	5.23
Umbria (peri-urban)	0.11	11.91	0.10	5.45
Umbria (rural)	0.09	13.40	0.09	6.09
Veneto (urban, Venice)	0.07	10.87	0.13	13.74
Veneto (peri-urban)	0.11	15.11	0.11	5.84
Veneto (rural)	0.14	18.31	0.12	4.07
Abruzzo (peri-urban)	0.16	12.12	0.07	11.46
Abruzzo (rural)	0.08	10.60	0.13	5.31
Basilicata (peri-urban)	0.07	9.81	0.11	10.24
Basilicata (rural)	0.07	13.45	0.09	5.02
Calabria (peri-urban)	0.11	8.63	0.14	6.22
Calabria (rural)	0.08	12.28	0.14	3.90
Campania (urban, Naples)	0.10	4.00	0.17	7.22
Campania (peri-urban)	0.12	11.54	0.15	4.26
Campania (rural)	0.13	11.58	0.15	3.62

Note. For the Urban areas has been reported the names of the metropolitan cities as well.

Table A2. Estimates of Gini index and coefficients of variation

Area	Foreigners	CV Foreigners	Italians	CV Italians
Piedmont (urban, Turin)	0.45	2.02	0.25	3.50
Piedmont (peri-urban)	0.23	2.82	0.28	3.35
Piedmont (rural)	0.27	6.72	0.31	3.37
Aosta Valley (rural)	0.29	9.68	0.29	2.49
Lombardy (urban, Milan)	0.30	11.10	0.31	4.48
Lombardy (peri-urban)	0.25	16.98	0.25	2.28
Lombardy (rural)	0.31	6.82	0.27	1.51
Emilia Romagna (urban, Bologna)	0.31	10.31	0.23	5.94
Emilia Romagna (peri-urban)	0.28	6.29	0.26	2.75
Emilia Romagna (rural)	0.30	14.91	0.25	2.48
Friuli Venezia Giulia (peri-urban)	0.25	6.72	0.29	4.20
Friuli Venezia Giulia (rural)	0.26	10.76	0.25	2.86
Lazio (urban, Rome)	0.24	7.84	0.31	1.96
Lazio (peri-urban)	0.20	10.47	0.25	2.98
Lazio (rural)	0.18	16.16	0.27	2.51
Liguria (urban, Genoa)	0.29	17.05	0.28	2.71
Liguria (peri-urban)	0.31	17.73	0.27	4.38
Liguria (rural)	0.26	7.03	0.31	4.61
Marche (peri-urban)	0.21	8.98	0.24	6.00
Marche (rural)	0.32	11.89	0.27	2.93
Molise (rural)	0.24	17.30	0.28	2.04
Apulia (urban, Bari)	0.25	10.41	0.28	4.99
Apulia (peri-urban)	0.29	12.50	0.28	2.76
Apulia (rural)	0.31	11.94	0.28	2.57
Sardinia (urban, Cagliari, and Sassari)	0.25	7.01	0.32	5.15
Sardinia (peri-urban)	0.32	5.68	0.26	3.95
Sardinia (rural)	0.27	6.54	0.28	3.25
Sicily (urban, Palermo, Catania, and Messina)	0.30	16.36	0.32	2.91
Sicily (peri-urban)	0.23	12.08	0.30	2.67
Sicily (rural)	0.28	9.35	0.29	2.41
Tuscany (urban, Florence)	0.30	7.77	0.32	3.91
Tuscany (peri-urban)	0.22	8.50	0.26	3.14
Tuscany (rural)	0.28	8.39	0.27	2.72
Trentino South-Tyrol (peri-urban)	0.27	6.24	0.29	2.80
Trentino South-Tyrol (rural)	0.31	5.61	0.27	1.92
Umbria (peri-urban)	0.27	6.26	0.28	2.96
Umbria (rural)	0.24	6.83	0.26	3.17
Veneto (urban, Venice)	0.22	5.57	0.29	8.03
Veneto (peri-urban)	0.26	7.64	0.26	3.14
Veneto (rural)	0.30	8.59	0.27	2.10
Abruzzo (peri-urban)	0.35	5.90	0.22	6.03
Abruzzo (rural)	0.23	5.55	0.29	2.92
Basilicata (peri-urban)	0.21	5.11	0.26	5.35

(continued)

Table A2. Continued

Area	Foreigners	CV Foreigners	Italians	CV Italians
Basilicata (rural)	0.21	6.74	0.24	2.68
Calabria (peri-urban)	0.27	4.51	0.31	3.43
Calabria (rural)	0.23	6.40	0.31	2.12
Campania (urban, Naples)	0.25	3.95	0.34	3.94
Campania (peri-urban)	0.26	5.82	0.29	2.25
Campania (rural)	0.28	5.64	0.29	1.96

Note. For the Urban areas has been reported the names of the metropolitan cities as well.

Table A3. Estimates of Theil index and coefficients of variation

Area	Foreigners	CV Foreigners	Italians	CV Italians
Piedmont (urban, Turin)	0.36	4.21	0.11	8.52
Piedmont (peri-urban)	0.08	5.79	0.13	6.86
Piedmont (rural)	0.12	13.00	0.17	7.23
Aosta Valley (rural)	0.13	14.93	0.14	5.40
Lombardy (urban, Milan)	0.15	21.72	0.16	9.10
Lombardy (peri-urban)	0.10	40.28	0.10	4.79
Lombardy (rural)	0.15	13.96	0.12	3.27
Emilia Romagna (urban, Bologna)	0.16	15.31	0.09	11.17
Emilia Romagna (peri-urban)	0.12	12.37	0.12	5.64
Emilia Romagna (rural)	0.15	29.01	0.11	5.89
Friuli Venezia Giulia (peri-urban)	0.10	15.19	0.14	9.47
Friuli Venezia Giulia (rural)	0.11	22.57	0.11	6.59
Lazio (urban, Rome)	0.10	12.23	0.15	4.42
Lazio (peri-urban)	0.06	23.89	0.10	6.47
Lazio (rural)	0.05	36.77	0.12	5.75
Liguria (urban, Genoa)	0.14	36.98	0.13	6.59
Liguria (peri-urban)	0.15	38.29	0.12	8.69
Liguria (rural)	0.11	12.01	0.17	10.67
Marche (peri-urban)	0.07	16.55	0.10	15.83
Marche (rural)	0.17	23.30	0.12	6.39
Molise (rural)	0.10	35.16	0.13	4.37
Apulia (urban, Bari)	0.10	22.38	0.12	10.14
Apulia (peri-urban)	0.14	30.02	0.13	6.01
Apulia (rural)	0.16	23.27	0.13	5.95
Sardinia (urban, Cagliari, and Sassari)	0.10	14.38	0.17	10.22
Sardinia (peri-urban)	0.17	12.06	0.11	8.90
Sardinia (rural)	0.12	13.57	0.13	7.87
Sicily (urban, Palermo, Catania, and Messina)	0.15	33.45	0.17	6.87
Sicily (peri-urban)	0.08	26.76	0.14	5.44
Sicily (rural)	0.13	20.61	0.14	5.68

(continued)

Table A3. Continued

Area	Foreigners	CV Foreigners	Italians	CV Italians
Tuscany (urban, Florence)	0.15	15.36	0.17	8.08
Tuscany (peri-urban)	0.08	16.84	0.11	6.93
Tuscany (rural)	0.13	17.91	0.12	6.10
Trentino South-Tyrol (peri-urban)	0.12	12.32	0.14	6.26
Trentino South-Tyrol (rural)	0.15	12.18	0.12	4.03
Umbria (peri-urban)	0.11	12.60	0.13	6.09
Umbria (rural)	0.10	14.23	0.11	7.41
Veneto (urban, Venice)	0.08	11.59	0.14	15.64
Veneto (peri-urban)	0.11	16.07	0.12	6.65
Veneto (rural)	0.15	21.00	0.12	5.00
Abruzzo (peri-urban)	0.21	13.42	0.08	13.33
Abruzzo (rural)	0.08	11.22	0.14	6.41
Basilicata (peri-urban)	0.07	10.53	0.11	10.42
Basilicata (rural)	0.07	14.58	0.10	5.61
Calabria (peri-urban)	0.12	9.28	0.15	7.21
Calabria (rural)	0.08	13.03	0.15	4.42
Campania (urban, Naples)	0.10	8.42	0.20	9.53
Campania (peri-urban)	0.11	12.15	0.13	4.43
Campania (rural)	0.13	12.35	0.14	4.33

Note. For the Urban areas has been reported the names of the metropolitan cities as well.

Table A4. Estimates of Relative Theil index and coefficients of variation

Area	Foreigners	CV Foreigners	Italians	CV Italians
Piedmont (urban, Turin)	0.0302	4.21	0.0083	8.52
Piedmont (peri-urban)	0.0070	5.79	0.0097	6.86
Piedmont (rural)	0.0103	13.00	0.0114	7.23
Aosta Valley (rural)	0.0159	14.93	0.0117	5.40
Lombardy (urban, Milan)	0.0125	21.72	0.0114	9.10
Lombardy (peri-urban)	0.0080	40.28	0.0067	4.79
Lombardy (rural)	0.0116	13.96	0.0076	3.27
Emilia Romagna (urban, Bologna)	0.0152	15.31	0.0074	11.17
Emilia Romagna (peri-urban)	0.0115	12.37	0.0087	5.64
Emilia Romagna (rural)	0.0147	29.01	0.0087	5.89
Friuli Venezia Giulia (peri-urban)	0.0082	15.19	0.0104	9.47
Friuli Venezia Giulia (rural)	0.0084	22.57	0.0072	6.59
Lazio (urban, Rome)	0.0091	12.23	0.0124	4.42
Lazio (peri-urban)	0.0054	23.89	0.0075	6.47
Lazio (rural)	0.0047	36.77	0.0091	5.75
Liguria (urban, Genova)	0.0149	36.98	0.0103	6.59
Liguria (peri-urban)	0.0142	38.29	0.0091	8.69

(continued)

Table A4. Continued

Area	Foreigners	CV Foreigners	Italians	CV Italians
Liguria (rural)	0.0102	12.01	0.0131	10.67
Marche (peri-urban)	0.0057	16.55	0.0072	15.83
Marche (rural)	0.0139	23.30	0.0083	6.39
Molise (rural)	0.0088	35.16	0.0099	4.37
Apulia (urban, Bari)	0.0088	22.38	0.0084	10.14
Apulia (peri-urban)	0.0110	30.02	0.0092	6.01
Apulia (rural)	0.0152	23.27	0.0104	5.95
Sardinia (urban, Cagliari, and Sassari)	0.0093	14.38	0.0128	10.22
Sardinia (peri-urban)	0.0176	12.06	0.0094	8.90
Sardinia (rural)	0.0101	13.57	0.0096	7.87
Sicily (urban, Palermo, Catania, and Messina)	0.0119	33.45	0.0114	6.87
Sicily (peri-urban)	0.0071	26.76	0.0100	5.44
Sicily (rural)	0.0112	20.61	0.0100	5.68
Tuscany (urban, Florence)	0.0141	15.36	0.0135	8.08
Tuscany (peri-urban)	0.0073	16.84	0.0083	6.93
Tuscany (rural)	0.0133	17.91	0.0097	6.10
Trentino South-Tyrol (peri-urban)	0.0114	12.32	0.0104	6.26
Trentino South-Tyrol (rural)	0.0150	12.18	0.0082	4.03
Umbria (peri-urban)	0.0095	12.60	0.0087	6.09
Umbria (rural)	0.0113	14.23	0.0090	7.41
Veneto (urban, Venice)	0.0071	11.59	0.0102	15.64
Veneto (peri-urban)	0.0102	16.07	0.0080	6.65
Veneto (rural)	0.0161	21.00	0.0104	5.00
Abruzzo (peri-urban)	0.0218	13.42	0.0062	13.33
Abruzzo (rural)	0.0079	11.22	0.0108	6.41
Basilicata (peri-urban)	0.0063	10.53	0.0077	10.42
Basilicata (rural)	0.0074	14.57	0.0070	5.61
Calabria (peri-urban)	0.0108	9.28	0.0112	7.21
Calabria (rural)	0.0071	13.02	0.0104	4.42
Campania (urban, Naples)	0.0110	8.42	0.0164	9.53
Campania (peri-urban)	0.0109	12.15	0.0106	4.43
Campania (rural)	0.0162	12.35	0.0097	4.33

Note. For the Urban areas has been reported the names of the metropolitan cities as well.

References

- Arora S., & Barak B. (2009). *Computational complexity: A modern approach*. Cambridge University Press.
- Balli F., & Tiezzi S. (2010). Equivalence scales, the cost of children and household consumption patterns in Italy. *Review of Economics of the Household*, 8(4), 527–549. <https://doi.org/10.1007/s11150-009-9068-3>
- Barabesi L., Crescenzi F., & Mori L. (2025). Theil index estimation by means of the influence function with an application to income surveys. *Journal of the Royal Statistical Society: Series A*, 1–23. <https://doi.org/10.1093/jrssa/qnaf143>
- Barabesi L., Diana G., & Perri P. F. (2016). Linearization of inequality indices in the design-based framework. *Statistics*, 50(5), 1161–1172. <https://doi.org/10.1080/02331888.2015.1135924>
- Barigozzi M., & Speciale B. (2011). Immigrants' legal status, permanence in the destination country and the distribution of consumption expenditure. *Applied Economics Letters*, 18(14), 1341–1347. <https://doi.org/10.1080/13504851.2010.537618>

- Battistin E., Blundell R., & Lewbel A. (2009). Why is consumption more log-normal than income? Gibrat's law revisited. *Journal of Political Economy*, 117(6), 1140–1154. <https://doi.org/10.1086/648995>
- Bessonneau P., Brilhaut G., Chauvet G., & Garcia C. (2020). *With-replacement bootstrap variance estimation for household surveys: principles, examples and implementation*. Working Paper, Statistic Canada.
- Caterina F. G., & Berti F. (2015). *The exploitation of migrant workers in the Italian agricultural sector: a case study from a Ghanaian community of day-labourers in northern Apulia* (Technical Report). International Centre for Rural Health of the University Hospital San Paolo of Milan and Brescia.
- Chambers R., Chandra H., Salvati N., & Tzavidis N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 76, 47–69. <https://doi.org/10.1111/rssb.12019>
- Chambers R., & Tzavidis N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268. <https://doi.org/10.1093/biomet/93.2.255>
- Chandra H., & Chambers R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, 37, 39–51. <https://hdl.handle.net/10779/uow.27776043.v1>
- Chandra H., Kumar S., & Aditya K. (2018). Small area estimation of proportions with different levels of auxiliary data. *Biometrical Journal*, 60(2), 395–415. <https://doi.org/10.1002/bimj.v60.2>
- Chatterjee S., Lahiri P., & Li H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, 36(3), 1221–1245. <https://doi.org/10.1214/07-AOS512>
- Chotikapanich D., Griffiths W. E., Hajargasht G., Karunaratne W., & Rao D. P. (2018). Using the GB2 income distribution. *Econometrics*, 6(2), 21. <https://doi.org/10.3390/econometrics6020021>
- Cuttillo A., Arigoni I., Valdoni L., & De Martino V. (2020). *Indagine sulle spese delle famiglie. periodo di riferimento: Anno 2019. aspetti metodologici dell'indagine* (Technical Report).
- D'Agostino A., Regoli A., Cornelio G., & Berti F. (2016). Studying income inequality of immigrant communities in Italy. *Social Indicators Research*, 127(1), 83–100. <https://doi.org/10.1007/s11205-015-0954-1>
- De Nicolò S., Ferrante M. R., & Pacei S. (2023). Small-sample bias correction of inequality estimators in complex surveys. *Journal of Official Statistics*. <https://doi.org/10.1177/0282423X241244>
- Diallo M. S., & Rao J. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, 1092–1116. <https://doi.org/10.1111/sjso.12336>
- Djajic S. (2003). Assimilation of immigrants: Implications for human capital accumulation of the second generation. *Journal of Population Economics*, 16, 831–845. <https://doi.org/10.1007/s00148-003-0162-1>
- Fabrizi E., Ferrante M. R., & Pacei S. (2007). Small area estimation of average household income based on unit level models for panel data. *Survey Methodology*, 33, 187–198. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2007002/article/10496-eng.pdf>
- Fabrizi E., & Trivisano C. (2016). Small area estimation of the Gini concentration coefficient. *Computational Statistics & Data Analysis*, 99, 223–234. <https://doi.org/10.1016/j.csda.2016.01.010>
- Fasulo A., Guandalini A., & Terribili M. D. (2017). Google trends for nowcasting quarterly household consumption expenditure. *Rivista Italiana Di Economia, Demografia E Statistica*, 71, 5.
- Field C. A., & Welsh A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 69, 369–390. <https://doi.org/10.1111/j.1467-9868.2007.00593.x>
- Gao X., He C., & Sun D. (2014). Adjusting nonresponse bias in small area estimation without covariates via a Bayesian spatial model. *Statistics and its Interface*, 7(4), 517–530. <https://doi.org/10.4310/SII.2014.v7.n4.a8>
- Gardini A., Trivisano C., & Fabrizi E. (2021). Bayesian analysis of anova and mixed models on the log-transformed response variable. *Psychometrika*, 86(2), 619–641. <https://doi.org/10.1007/s11336-021-09769-y>
- Graf M., Marin J. M., & Molina I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *Test*, 28(2), 565–597. <https://doi.org/10.1007/s11749-018-0594-2>
- Graf M., & Nedyalkova D. (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth*, 60(4), 821–842. <https://doi.org/10.1111/roiw.2014.60.issue-4>
- Hamilton T. G., & Kawachi I. (2013). Changes in income inequality and the health of immigrants. *Social Science & Medicine*, 80(6), 57–66. <https://doi.org/10.1016/j.socscimed.2012.10.002>
- Jiang J., & Rao J. S. (2020). Robust small area estimation: An overview. *Annual Review of Statistics and Its Application*, 7(1), 337–360. <https://doi.org/10.1146/statistics.2020.7.issue-1>
- Jobe J. M. (1989). Lognormal distributions: Theory and applications.
- Kreutzmann A.-K., Pannier S., Rojas-Perilla N., Schmid T., Templ M., & Tzavidis N. (2019). The R package EMDI for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7), 1–33. <https://doi.org/10.18637/jss.v091.i07>
- Langel M., & Tillé Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 176, 521–540. <https://doi.org/10.1111/j.1467-985X.2012.01048.x>

- Li H., Liu Y., & Zhang R. (2019). Small area estimation under transformed nested-error regression models. *Statistical Papers*, 60(4), 1397–1418. <https://doi.org/10.1007/s00362-017-0879-7>
- Lian M., Chen L., Hui C., Zhu F., & Shi P. (2024). On the relationship between the GINI coefficient and skewness. *Ecology and Evolution*, 14(12), e70637. <https://doi.org/10.1002/ece3.v14.12>
- Liu Y., Liu X., Pan Y., Jiang J., & Xiao P. (2022). An empirical comparison of various MSPE estimators and associated prediction intervals for small area means. *Journal of Statistical Computation and Simulation*, 93(10), 1–27. <https://doi.org/10.1080/00949655.2022.2144854>
- Lubrano M. (2013). The econometrics of inequality and poverty. Lorenz curves, the Gini coefficient and parametric distributions. <https://perso.amse-aixmarseille.fr/lubrano/cours/Lecture-4.pdf>
- Luus R., Neethling A., & De Wet T. (2012). Effectiveness of weighting and bootstrap in the estimation of welfare indices under complex sampling: Theory and methods. *South African Statistical Journal = Suid-Afrikaanse Statistiese Tydskrif*, 46, 85–114. <https://hdl.handle.net/10520/EJC99143>
- Lyu X., Berg E. J., & Hofmann H. (2020). Empirical Bayes small area prediction under a zero-inflated log-normal model with correlated random area effects. *Biometrical Journal*, 62(8), 1859–1878. <https://doi.org/10.1002/bimj.v62.8>
- Ma R., & Yin S. (2024). Rural economic transformation and household consumption structure: An empirical study in the context of urbanization. *Journal of the Knowledge Economy*, 15(4), 18286–18302. <https://doi.org/10.1007/s13132-024-01815-1>
- Marchetti S., & Tzavidis N. (2021). Robust estimation of the theil index and the gini coefficient for small areas. *Journal of Official Statistics*, 37(4), 955–979. <https://doi.org/10.2478/jos-2021-0041>
- McCann P., & Soete L. (2020). *Place-based innovation for sustainability* (Technical Report). Publications Office of the European Union: Luxembourg.
- McDonald J. B., & Jensen B. C. (1979). An analysis of some properties of alternative measures of income inequality based on the gamma distribution function. *Journal of the American Statistical Association*, 74(368), 856–860. <https://doi.org/10.1080/01621459.1979.10481042>
- Molina I., Corral P., & Nguyen M. (2022). Estimation of poverty and inequality in small areas: Review and discussion. *TEST*, 31(4), 1143–1166. <https://doi.org/10.1007/s11749-022-00822-1>
- Molina I., & Ferretti C. (2012). Fast EB method for estimating complex poverty indicators in large populations. *Journal of the Indian Society of Agricultural Statistics*, 66, 105–120. <https://doi.org/10.1007/978-3-540-22118-0>
- Molina I., & Martin N. (2018). Empirical best prediction under a nested error model with log transformation. *Annals of Statistics*, 46, 1961–1993. <https://doi.org/10.1214/17-AOS1608>
- Molina I., & Rao J. N. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3), 369–385. <https://doi.org/10.1002/cjs.v38:3>
- Mori L., & Ferrante M. R. (2025). Small area estimation of household economic indicators under unit-level generalized additive models for location, scale and shape. *Journal of Survey Statistics and Methodology*, 13(1), 160–196. <https://doi.org/10.1093/jssam/smae038>
- Ng E. S.-W., Grieve R., & Carpenter J. R. (2013). Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data. *The Stata Journal*, 13(1), 141–164. <https://doi.org/10.1177/1536867X1301300111>
- Nti I. K., Nyarko-Boateng O., & Aning J. (2021). Performance of machine learning algorithms with different K values in K-fold cross-validation. *I.J. Information Technology and Computer Science*, 6, 61–71. <https://doi.org/10.5815/ijitcs.2021.06.05>
- Opsomer J. D., Claeskens G., Ranalli M. G., Kauermann G., & 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. <https://doi.org/10.1111/j.1467-9868.2007.00635.x>
- Prieto-Alaiz M. (2007). [https://doi.org/10.1016/S1049-2585\(06\)14003-X](https://doi.org/10.1016/S1049-2585(06)14003-X)
- Rao J. N., & Molina I. (2015). *Small area estimation*. John Wiley & Sons.
- Rigby R. A., & 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. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- Rigby R. A., Stasinopoulos D. M., Heller G. Z., & De Bastiani F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R* (1st ed.). Chapman and Hall.
- Rojas-Perilla N., Pannier S., Schmid T., & Tzavidis N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1), 121–148. <https://doi.org/10.1111/rssa.12488>
- Rust K. F., & Rao J. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3), 283–310. <https://doi.org/10.1177/096228029600500305>
- Schluter C. (2012). On the problem of inference for inequality measures for heavy-tailed distributions. *The Econometrics Journal*, 15(1), 125–153. <https://doi.org/10.1111/j.1368-423X.2011.00356.x>

- Schmid T., Tzavidis N., Münnich R., & Chambers R. (2016). Outlier robust small-area estimation under spatial correlation. *Scandinavian Journal of Statistics*, 43, 806–826. <https://doi.org/10.1111/sjos.12205>
- Slud E. V., & Maiti T. (2006). Mean-squared error estimation in transformed fay-herriot models. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 68(2), 239–257. <https://doi.org/10.1111/j.1467-9868.2006.00542.x>
- 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) Statistics-Canada.
- Sugasawa S., & Kubokawa T. (2019). Adaptively transformed mixed-model prediction of general finite-population parameters. *Scandinavian Journal of Statistics*, 46(4), 1025–1046. <https://doi.org/10.1111/sjos.v46.4>
- Sugasawa S., & Kubokawa T. (2020). Small area estimation with mixed models: A review. *Japanese Journal of Statistics and Data Science*, 3(2), 693–720. <https://doi.org/10.1007/s42081-020-00076-x>
- Tzavidis N., Zhang L.-C., Luna A., Schmid T., & 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(4), 927–979. <https://doi.org/10.1111/rssa.12364>
- Ubaidillah A., Notodiputro K. A., Kurnia A., & 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(2), 57–68. <https://doi.org/10.12988/ams.2018.712361>
- Würz N., Schmid T., & 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(4), 1679–1706. <https://doi.org/10.1111/rssa.12913>