Small area estimation of inequality measures using mixtures of Beta

Small Area Estimation of Inequality Measures using Mixtures of Beta

Silvia De Nicolò* Maria Rosaria Ferrante Silvia Pacei

University of Bologna

Abstract

Economic inequalities referring to specific regions are crucial in deepening spatial heterogeneity. Income surveys are generally planned to produce reliable estimates at countries or macroregion levels, thus we implement a small area model for a set of inequality measures (Gini, Relative Theil and Atkinson indexes) to obtain reliable microregion estimates. Considering that inequality estimators are unit-interval defined with skewed and heavy-tailed distributions, we propose a Bayesian hierarchical model at the area level involving a Beta mixture. An application on EU-SILC data is carried out and a design-based simulation is performed. Our model outperforms in terms of bias, coverage and error the standard Beta regression model. Moreover, we extend the analysis of inequality estimators by deriving their approximate variance functions.

Keywords: Area Level Model, Beta Mixture, Beta Regression, Hierarchical Bayes, Inequality Mapping, Small Area Estimation.

1 Introduction

The issue of widespread economic inequality characterizes the current global predicament and has a central role in political and economic discourse (Piketty, 2014). The demand for inequality estimates referring to specific subpopulations is growing, policymakers and stakeholders need them in order to formulate and implement policies, distribute resources and measure the effect of policy actions at local level. Besides, such estimates may be valuable in order to further deepen some research trends in regional and inequality studies, for instance, to identify which regions constitute the drivers of national income inequality and to study spatial spillovers (Márquez et al., 2019; Moser & Schnetzer, 2017). For a recent review on spatial inequality, see Cavanaugh and Breau (2018).

Economic inequality is conventionally measured on equivalent disposable income data. Generally, such data are collected via household sample surveys which are planned for aggregates estimation at coarse geographical levels, being scarcely available at finer geographical scales. Thus, local domains fall outside the prior design plan, resulting in small-sized samples and yielding unreliable direct estimation (i.e. with large error). For instance, the Survey of Income and Living Conditions (EU-SILC), which provides information on income for the whole set

*silvia.denicolo@unibo.it

of European countries, is able to provide reliable estimates only at NUTS-2 as the maximum level of disaggregation. The problem could be overcome by increasing the survey sample size, but it is often excluded by cost–benefit analysis. A solution to cope with it is to rely on Small Area Estimation (SAE) techniques. Such techniques exploit auxiliary information to borrow strength across areas and produce area-specific estimates with an acceptable level of uncertainty. The model-based class of SAE techniques leverages hierarchical models, both at area level or unit (individual) level, producing reliable estimates when models are correctly specified and informative auxiliary variables are available without error. For a review, see Rao and Molina (2015, ch. 4) and Tzavidis et al. (2018).

In literature, only a few contributions relate to the small area estimation of inequality indicators, since in this context poverty has special consideration and inequality is treated as a minor appendix (Molina & Rao, 2010; Pratesi, 2016). Fabrizi and Trivisano (2016) deal with the Gini index estimation via area level models. Other authors propose unit level models: Tzavidis and Marchetti (2016) for Gini index and Quantile Share Ratio, Marchetti and Tzavidis (2021) for Gini and Theil indexes and Gardini et al. (2022) for the Quantile Share Ratio. All of them never treat more than two inequality measures at a time. However, inequality can be seen as a multifaceted concept, embracing diverse objective and normative assessments on the characteristics of income distribution (Cowell, 2011). It can be measured via a plethora of statistical indicators, all of them with different axiomatic properties and featuring varying sensitivities to extreme values and income transfers that could reduce inequality. Thus, the point of concurrently producing estimates of various indicators could provide a comprehensive overview of the phenomenon.

In this spirit, we propose a small area estimation strategy for a set of four income inequality measures. In addition to the Gini index, we consider the Relative Theil index, particularly appealing due to its additive decomposability property which allows expressing inequality as the sum of between and within components. Moreover, we consider the family of Atkinson measures, which stands apart from other descriptive measures by explicitly incorporating a welfare evaluation of inequality implications and enabling a complete ordering of income distributions. All the measures considered vary between 0 (case of perfect equality) and 1 (perfect inequality), having double bounded support, where degenerate cases 0 and 1 have probability very close to zero also in a small sample context.

Our approach lies within the framework of Bayesian inference of area level models. Generally, unit level proposals require the auxiliary variables to be defined for the entire population. This may be difficult to achieve as administrative archives are not easily accessible and linked to each other and to survey samples (Harmening et al., 2020). In contrast, area level models require the auxiliary information to be defined only at domain level. For this reason, an area level proposal may be valuable, being less demanding with respect to data requirements as well as incorporating design-based properties in a straightforward way.

Inequality design-based estimators have peculiar characteristics; their behaviour in complex survey small samples has been investigated showing highly skewed and heavy-tailed distributions at decreasing sample sizes, even conditionally on auxiliary variables. A Monte Carlo analysis is available in Section S.2 of the Supplementary Material. In this context, they may require a flexible regression specification. Previous proposals in this sense exploit semi-parametric specifications at the unit level (Bianchi et al., 2018; Opsomer et al., 2008), suitable data transformations (Rojas-Perilla et al., 2020) or alternative likelihood assumptions. At the area level, the inadequacy of Fay-Herriot models (Fay & Herriot, 1979) in case of skewed and heavy-tailed estimators is particularly established. The alternative likelihoods considered

are the skew-normal (Ferrante & Pacei, 2017; Ferraz & Moura, 2012), skew-t (Moura et al., 2017) and log-normal ones (Fabrizi et al., 2018; Slud & Maiti, 2006). When dealing with unit interval-defined responses, such alternative specifications may fit values outside the variable support. In this case, the small area literature at the area level gathers linear mixed models with suitable transformations and Beta regression models (Janicki, 2020). However, even the Beta distribution may fail in case of markedly skewed and heavy-tailed estimators (Bayes et al., 2012; Migliorati et al., 2018). The body of literature based on Beta regression in SAE comprises univariate proposals (Bauder et al., 2015; Fabrizi & Trivisano, 2016; Liu et al., 2007); multivariate versions (Fabrizi et al., 2011; Souza & Moura, 2016); and zero and/or one inflated extensions (Fabrizi et al., 2020; Wieczorek et al., 2012).

Our proposal involves incorporating an alternative likelihood assumption by adopting a Beta mixture-based approach, whose performances are compared with the most common Beta regression model. Specifically, we assume as sampling distribution the Flexible Beta one, proposed by Migliorati et al. (2018). The Flexible Beta distribution is a mixture of two Beta random variables, particularly interesting for the purpose of small-area estimating inequality measure due to its superior flexibility, given its four parameters structure. Indeed, the Beta distribution has good properties, being able to adapt to different shapes, but its two parameters structure hinders further flexible modelling. Eventually, we derive the approximate variance function of each inequality estimator, analyzing how their mean and variance are tied together and whether such interrelation differs among measures.

Our contribution has therefore multiple levels. On one hand, we provide a comprehensive discussion about inequality and its SAE by considering a set of multiple measures. Secondly, we deepen the analysis of inequality estimators by deriving their approximate variance functions, which may be useful for further modelling. Thirdly, our methodological proposal extends small area literature in the case of unit interval-defined, skewed and heavy-tailed estimators. Our model comes out to outperform the Beta one, both in terms of bias and error of the target estimators, avoiding to highly underestimate inequality and providing reliable estimates.

The paper is organized as follows. Inequality measures and their estimators are defined and described in Section 2, together with a proposal of sampling variance estimation. Section 3 defines the proposed Beta and Flexible Beta small area models. An application on EU-SILC income data is unraveled in Section 4 and a design-based simulation can be found in Section 5, in order to evaluate the frequentist properties of model-based estimators. Conclusions are drawn in Section 6.

2 Inequality Measures

In this section, we describe the inequality measures considered and their estimator in the complex survey case. Such estimators are known to be biased in small samples, often leading to underestimation; therefore, we adopt the bias-corrected direct estimators proposed by De Nicoló et al. (2021). Their simulation results lead us to assume that such estimators are approximately unbiased or slightly biased depending on the domain sample size. In Subsection 2.1, their variance estimation is set out and their estimates are commented.

The most famous inequality measure is the Gini index, measuring concentration in the distribution of a positive random variable; among its several equivalent definitions, we adopt the formulation of Sen (1997). Suppose we are dealing with a finite population, denoted with $\mathcal{U}$, of $N(<\infty)$ elements, and let a sample $\mathcal{S}_{iid}$ of size n be randomly drawn from $\mathcal{U}$. Let $z \in \mathbb{R}^+$

be a characteristic of interest, in our case income, which is observable for each unit in $\mathcal{S}_{iid}$. The simple random sampling (srs) estimator of the Gini index is defined as

$$G = \frac{2}{n^2}\frac{\sum_{i\in\mathcal{S}_{iid}} z_i(r_i - 1/2)}{\hat{\mu}_{srs}} - 1,$$

with r_i the rank of the i-th sampled unit and $\hat{\mu}_{srs}$ is the sample mean. Let us suppose, moreover, that a sample $\mathcal{S}$ is drawn from $\mathcal{U}$ through a complex selection scheme e.g., involving stratification and multi-stage selection, as in the case of survey data. This involves unequal inclusion probabilities across units, thus a weighted estimator should be adopted, as proposed by Langel and Tillé (2013):

$$G_w = \frac{2\sum_{i\in\mathcal{S}} w_i z_i(\hat{N}_i - w_i/2)}{\hat{N}^2\hat{\mu}} - 1,$$

with w_i denoting sampling weights attached to unit i, $\hat{N} = \sum_{k\in\mathcal{S}} w_k$, $\hat{\mu} = \sum_{k\in\mathcal{S}} w_k z_k/\hat{N}$, and $\hat{N}_i = \sum_{k\in\mathcal{S}} w_k \mathbb{1}(r_k \leq r_i)$. The notation $\mathbb{1}(A)$ defines an indicator function, assuming value 1 if A is observed and 0 otherwise. The weights could be the inverse of the inclusion probabilities or a treated and calibrated version of them. We adopted its bias-corrected version proposed by De Nicoló et al. (2021) as follows

$$G_{adj} = \frac{\tilde{n}}{\tilde{n} - 2}\left[G_w - \frac{2\hat{\gamma}}{\hat{\mu}^3}\mathbb{V}[\hat{\mu}] + \frac{2}{\hat{\mu}^2}\mathrm{Cov}[\hat{\mu}, \hat{\gamma}]\right],$$

with $\tilde{n} = \sum_{k\in\mathcal{S}} \mathbb{1}(w_k \neq 0)$ and $\hat{\gamma} = \sum_{i\in\mathcal{S}} w_i z_i(\hat{N}_i - w_i/2)/\hat{N}^2$. $\mathbb{V}[\hat{\mu}]$ and $\mathrm{Cov}[\hat{\mu}, \hat{\gamma}]$ denote variance and covariance of design estimators.

Despite its fame, the Gini index has some drawbacks. First of all, it is a stochastic dominance measure enabling only for partial ordering of probability distributions. Namely, this index is able to determine which distribution precedes the other in the ordering only among certain pairs of probability distributions. Secondly, it does not allow for decomposability into within and between components. Thirdly, it is weakly (positional) transfer sensitive which means that, in case of income transfers, the index varies depending on the donor and recipients ranks.

The Relative Theil index, instead, is additive decomposable and has the advantage to be strongly transfer-sensitive, meaning that the measure reacts to transfers depending on the donor and recipient income levels. It is an entropy-based measure and is set up as the relative formulation of the more famous Theil index (Theil, 1979), i.e. scaled on the maximum of its support ($\log n$). Its estimator in the srs case is defined as follows

$$R = \frac{1}{n\log(n)}\sum_{i\in\mathcal{S}_{iid}} \frac{z_i}{\hat{\mu}_{srs}} \log\left(\frac{z_i}{\hat{\mu}}\right). \tag{1}$$

In the complex survey case, the Horwitz-Thompson type estimator for the Theil index has been considered in its bias-adjusted formulation (Biewen & Jenkins, 2006; De Nicoló et al., 2021). This has been adapted to the relative case by replacing the superior bound of its support with its population value, in order to not induce further bias, as follows

$$T_{adj} = \frac{1}{\hat{N}}\sum_{i\in\mathcal{S}} w_i \frac{z_i}{\hat{\mu}} \log\frac{z_i}{\hat{\mu}} + \frac{\mathrm{Cov}[\hat{\mu}, \hat{\varpi}]}{\hat{\mu}^2} - \left(\frac{\hat{\varpi}}{\hat{\mu}^3} + \frac{1}{2\hat{\mu}^2}\right)\mathbb{V}[\hat{\mu}]$$

$$R_{adj} = \frac{T_{adj}}{\log N}$$

with $\hat{\varpi} = \sum_{i \in \mathcal{S}} w_i z_i \log z_i / \hat{N}$.

Another perspective on inequality is depicted by the family of Atkinson Indexes (Atkinson, 1970). They provide for an explicit value judgment by incorporating in the measurement a social welfare function, regulated by a parameter ε. Under this normative approach, the index value has clear meaning, quantifying the amount of welfare loss of the current inequality level: a value of 0.30 means that "if incomes were equally distributed then we should need only the 70% of the present national income to achieve the same level of social welfare" (Atkinson, 1970). They can be seen, therefore, as measures of distributional inefficiency. Moreover, they satisfy a multiplicative decomposition property (Lasso de La Vega & Urrutia, 2008) and may provide for a complete ranking among alternative distributions, i.e. being able to determine the ordering among every pair of distributions, and thus to establish a ranking among the full set of distributions. This happens at the expense of more stringent and subjective assumptions on the choice of the welfare utility function to adopt (Bellú & Liberati, 2006). Under a concave utility function (whose concavity level is regulated by ε), the estimator of Atkinson index in the srs case is defined as

$$A(\varepsilon \neq 1) = 1 - \frac{1}{\hat{\mu}_{srs}} \left(\frac{1}{n} \sum_{i \in \mathcal{S}_{iid}} z_i^{1-\varepsilon} \right)^{1/(1-\varepsilon)} \tag{2}$$

$$A(\varepsilon = 1) = 1 - \frac{1}{\hat{\mu}_{srs}} \exp \left\{ \frac{\sum_{i \in \mathcal{S}_{iid}} \log z_i}{n} \right\} \tag{3}$$

with $\varepsilon \geq 0$. The parameter ε denotes the level of inequality aversion: at increasing values of ε, the index becomes more sensitive to changes at the lower end of the income distribution and vice versa. We consider specifically the two indexes referring to $\varepsilon = \{0.5, 1\}$ values, which incorporate nice robustness properties to outliers, showing at the same time different sensitivities (De Nicoló et al., 2021). The estimator referred to complex survey case (Biewen & Jenkins, 2006) is

$$A_w(\varepsilon \neq 1) = 1 - \frac{1}{\hat{\mu}} \left(\frac{1}{\hat{N}} \sum_{i \in \mathcal{S}} w_i z_i^{1-\varepsilon} \right)^{1/(1-\varepsilon)}$$

$$A_w(\varepsilon = 1) = 1 - \frac{1}{\hat{\mu}} \exp \left\{ \frac{\sum_{i \in \mathcal{S}} w_i \log z_i}{\hat{N}} \right\}.$$

We adopted their bias-adjusted versions (De Nicoló et al., 2021) as follows

$$\begin{aligned} A_{adj}(\varepsilon \neq 1) &= A_w(\varepsilon) + [1 - A_w(\varepsilon)] \times \\ &\quad \times \left[\frac{\varepsilon \cdot \mathbb{V}[\hat{\varrho}]}{2(1-\varepsilon)^2} [\hat{\mu} - \hat{\mu} A_w(\varepsilon)]^{2\varepsilon - 2} + \frac{\mathbb{V}[\hat{\mu}]}{\hat{\mu}^2} - \frac{\mathrm{Cov}[\hat{\varrho}, \hat{\mu}]}{\hat{\mu}^{2-\varepsilon}(1-\varepsilon)} [1 - A_w(\varepsilon)]^{\varepsilon - 1} \right] \\ A_{adj}(\varepsilon = 1) &= A_w(\varepsilon) + [1 - A_w(\varepsilon)] \left[\frac{\mathbb{V}[\hat{\iota}]}{2} + \frac{\mathbb{V}[\hat{\mu}]}{\hat{\mu}^2} - \frac{\mathrm{Cov}[\hat{\iota}, \hat{\mu}]}{\hat{\mu}} \right] \end{aligned}$$

with $\hat{\varrho} = \sum_{i \in \mathcal{S}} w_i z_i^{1-\varepsilon} / \hat{N}$ and $\hat{\iota} = \sum_{i \in \mathcal{S}} w_i \log z_i / \hat{N}$.

2.1 Variance Estimation

The sampling variances of complex survey estimators have been estimated from the data following a two steps strategy as in Fabrizi et al. (2011). As a first step, we performed a proper

bootstrap procedure developed taking into account the complex sampling design (De Nicoló et al., 2021; Fabrizi et al., 2020), using 1,000 bootstrap samples. Secondly, the bootstrap estimates have been smoothed via a Generalized Variance Function (GVF) approach in order to mitigate the uncertainty of sampling variances induced by small sample sizes.

The definition of a GVF smoothing model needs assumptions on the shape of the variance function for such inequality estimators. Thus, in the spirit of what was done by Fabrizi and Trivisano (2016) for the Gini index, we derived the variance function of Relative Theil and Atkinson index (for any ε) under specific simplifying conditions, such as the usual log normality assumption of income variable and *srs* setting. Indeed, closed-form results for the complex survey case are hard to obtain for non-linear statistics, such as the ones we are dealing with, due to the correlation among observations. The Gini index result, as well as our following derivations for the other measures, has been directly incorporated into the GVF model.

Before moving to the variance function derivation, let us introduce the partition of the population $\mathcal{U}$ into D small domains, such as each domain d has population size N_d, with $N = \sum_{d=1}^{D} N_d$, and samples $\mathcal{S}_{iid}$ and $\mathcal{S}$ are subsequently partitioned into D subsamples of size n_d and $\tilde{n}_d$ respectively, for $d = 1, \dots, D$, with $n = \sum_{d=1}^{D} n_d$ and $\tilde{n} = \sum_{d=1}^{D} \tilde{n}_d$.

Proposition 1. *Under the assumption of log-normality of income, let us consider the j-th individual in domain d, whose level of income is z_{jd} variable. As a consequence,* $\log(z_{jd}) \sim \mathcal{N}(\mu_d, \varphi_d^2)$*, iid at varying $j = 1, \dots, n_d$. The srs estimator of Relative Theil Index* (1)*, for domain d, R_d has variance function*

$$\mathbb{V}[R_d] \cong \frac{2\theta_d^{R2}}{n_d}, \tag{4}$$

where θ_d^R denotes its population value.

Proof. The Relative Theil index is defined for a log-normal income variable as

$$\theta_d^R = \frac{1}{\log(n_d)} \left(\frac{\mathbb{E}[z \cdot \log(z)]}{\mathbb{E}[z]} - \log(\mathbb{E}[z]) \right) = \frac{\varphi_d^2}{2\log(n_d)}. \tag{5}$$

Since the moments involved in the previous expression are

$$\begin{aligned}
\mathbb{E}[z] &= \exp\left\{ \mu_d + \frac{\varphi_d^2}{2} \right\} \\
\mathbb{E}[z \cdot \log(z)] &= \int_0^{+\infty} z \log(z) \frac{1}{z\sqrt{2\pi\varphi_d^2}} \exp\left\{ -\frac{[\log(z) - \mu_d]^2}{2\varphi_d^2} \right\} dx \\
&= \int_{-\infty}^{+\infty} t \exp\{t\} \frac{1}{\sqrt{2\pi\varphi_d^2}} \exp\left\{ -\frac{(t - \mu_d)^2}{2\varphi_d^2} \right\} dt \\
&= (\varphi_d^2 + \mu_d) \exp\left\{ \mu_d + \frac{\varphi_d^2}{2} \right\}, \qquad (6)
\end{aligned}$$

where the last step (6) involves equation 3.462.6 in Gradshteyn and Ryzhik (2014). Considering that φ_d^2 is estimated by $s_d^2 = \frac{1}{n_d - 1} \sum_{j=1}^{n_d} [\log(z_{jd}) - \hat{\mu}_d]^2$ and $\mathbb{V}[s_d] \cong \frac{\varphi_d^2}{2n_d}$, by applying delta method the result follows

$$\begin{aligned}
\mathbb{V}[R_d] = \mathbb{V}\left[\frac{s_d^2}{2\log(n_d)} \right] &\cong \frac{\varphi_d^4}{2\log^2(n_d) n_d} \\
&= \frac{2\theta_d^{R2}}{n_d} \qquad (7)
\end{aligned}$$

where equation (7) is obtained by (5) considering that $\varphi_d^2 = 2\theta_d^R \log(n_d)$. □

Proposition 2. *Under Proposition 1 assumptions, the srs estimators of Atkinson index* (2) *and* (3) *for domain d, denoted with* $A_d(\varepsilon)$*, have variance function*

$$\mathbb{V}[A_d(\varepsilon)] \cong \frac{2\theta_d^A(\varepsilon)^2}{n_d} \exp\{-2\theta_d^A(\varepsilon)\}, \tag{8}$$

where $\theta_d^A(\varepsilon)$ *denotes the population value of the index.*

Proof. The population value of Atkinson index in domain d under the mentioned assumptions, for any $\varepsilon \geq 0$ and $\neq 1$, is

$$\theta_d^A(\varepsilon) = 1 - \exp\{-\varepsilon\varphi_d^2/2\}, \tag{9}$$

with φ_d^2 estimated by s_d^2. By applying the normal distribution theory, $\mathbb{V}[s_d] \cong \frac{\varphi_d^2}{2n_d}$ and using the delta method:

$$\begin{aligned}\mathbb{V}[A_d(\varepsilon)] = \mathbb{V}\left[1 - \exp\left\{-\frac{\varepsilon s_d^2}{2}\right\}\right] &\cong \frac{\varepsilon^2\varphi_d^4}{2n_d}\exp\{-\varepsilon\varphi_d^2\} \\ &\cong \frac{2\theta_d^A(\varepsilon)^2}{n_d}\exp\{-2\theta_d^A(\varepsilon)\},\end{aligned} \tag{10}$$

where equation (10) is obtained by Maclaurin expanding (9), so that $\varphi_d^2 \cong 2\theta_d^A(\varepsilon)/\varepsilon$. Note that this result can be easily generalized to the case $\varepsilon = 1$. □

Since Proposition 1 and 2 has been derived under specific simplifying assumptions, we need to ensure their validity in our context. Thus, we tested them on our bias-corrected estimators for complex survey, namely A_{adj}, G_{adj} and R_{adj}, through a Monte Carlo simulation. We used EU-SILC data, described in detail in Section 4, as synthetic population by considering 21 NUTS-2 Italian regions as domains of interest. In order to circumvent possible null or negative income values and non-robustness to outliers, we treated income data by using a semi-parametric Pareto and inverse-Pareto tail modelling procedure using the Probability Integral Transform Statistic Estimator (PITSE) proposed by Finkelstein et al. (2006) and Masseran et al. (2019). We draw 1,000 samples by mimicking EU-SILC complex scheme, stratified with two-stage selection. Then we compared Monte Carlo variances with Proposition 2 and 1 results, showing very high correlations: 0.79 for Gini index, 0.92 for Atk(1), 0.86 for Atk(0.5) and 0.99 for Relative Theil. The empirical results, as expected, clearly underestimate the variances in comparison with the Monte Carlo ones, since the effect of the design is ignored. However, the relationship is strongly linear and by fitting a regression with Monte Carlo variances as response and the empirical ones as explanatory, intercepts are zeros and slopes end up to be 2.04 for Relative Theil, 2.23 for Atk(0.5), 2.27 for Atk(1) and 4.68 for Gini index. The strong linear dependence and proportionality and, at the same time, the non-negligible underestimation, lead us to consider appropriate the implementation of a GVF model.

In the following, the GVF model setting is unravelled, by considering also the Gini index variance derived by Fabrizi and Trivisano (2016) under the same assumptions of Propositions 1 and 2, defined approximately for domain d as

$$\mathbb{V}(G_d) \cong \frac{\theta_d^{G^2}(1 - \theta_d^{G^2})}{n_d}, \tag{11}$$

with θ_d^G its population value.

Consider that under a complex survey scheme, the sample $\mathcal{S}$ may be less informative than a sample of the same size $\tilde{n}_d$ under srs, being $\mathcal{S}$ affected by dependency across observations. The effective sample size, i.e. the srs equivalent sample size, is a proxy of the information carried by the sample and can be estimated for $\mathcal{S}$. Let us suppose it corresponds to $\psi_{\text{ind}} \cdot \tilde{n}_d$ for any considered index *ind*, with $\psi_{\text{ind}} > 0$ denoting a deflating factor induced by the dependence. The latter quantity can be alternatively defined as the inverse of the design effect deff_{ind}, i.e. the ratio between the design-based variance of a generic index estimator and its srs variance, which measures the amount of variance inflation induced by the complex selection process. Note that the design effect was assumed constant across areas in order to guarantee the smoothing of original estimates. Let denote with $\hat{\mathbb{V}}[\cdot]_{boot}$ a bootstrap estimator with large error, being y_d a generic inequality estimator, among $A_{adj}(\varepsilon)$, G_{adj} and R_{adj}, defined for domain d whose population values is θ_d. The numerator of its variance function is generically defined by $f(\theta_d)$. A GVF model is set up by assuming that

$$\mathbb{V}[y_d] = \frac{f(\theta_d)}{\psi_{\text{ind}} \cdot \tilde{n}_d}.$$

Therefore, we introduce the following smoothing model estimated via generalized least squares

$$\frac{f(y_d)}{\hat{\mathbb{V}}[y_d]_{boot}} = \psi_{\text{ind}} \tilde{n}_d + \epsilon_d,$$

where ϵ_d denotes zero-mean heteroskedastic residuals. The smoothed estimator comes from (8), (4), (11) by replacing θ_d with y_d and n_d with $\tilde{n}_d \cdot \hat{\psi}_{\text{ind}}$, where $\hat{\psi}_{\text{ind}}$ is the GLS estimate of ψ_{ind}. The pseudo R^2 for the smoothing models are respectively, 0.78 for Gini index, 0.72 for Atkinson ($\varepsilon = 1$) index, 0.67 for Atkinson ($\varepsilon = 0.5$) index and 0.48 for the Relative Theil index. The latter result is due to the instability of the ratio $f(y_d)/\hat{\mathbb{V}}[y_d]_{boot}$ in case of values close to zero of both the numerator and the denominator.

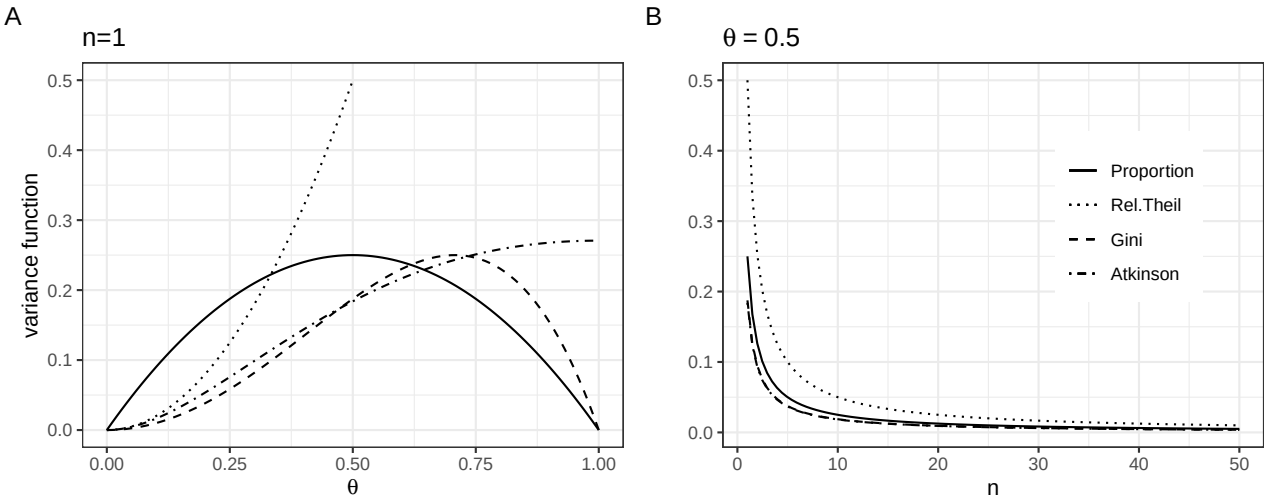

Figure 1: Variance functions with respect to θ_d ($n_d = 1$) on the left-hand side, with respect to n_d ($\theta_d = 0.5$) on the right-hand side, for each measure and considering the proportion case.

Results of Propositions 1 and 2 show a very different structure for the variance function of Atkinson and Relative Theil indexes with respect to the Gini index and proportions. A

comparison plot can be found in Figure 1 displaying each variance function with respect to θ_d and n_d. As opposed to the proportion and Gini index cases, the Atkinson and Relative Theil variance functions are both monotonically increasing, as clear from (9) and (5). Thus, greater values of θ_d correspond to greater log income dispersion, inevitably leading to an increase in index variability. The explosive trend of Relative Theil variance is related to its explosive connection with the log income population variance (5). Moreover, notice that the variance function of Atkinson index in (8) does not directly depend on its parameter ε, being fixed for the whole parametric family. This does not happen for the Generalized Entropy parametric family, as shown in Section S.1 of the Supplementary Material.

The precision estimates obtained for NUTS-3 Italian regions, employing EU-SILC data (described in Section 4), are analyzed in terms of the coefficient of variation (CV). ISTAT guidelines state that CV should not exceed 15% for domains and 18% for small domains in case of released estimates, otherwise, this serves as an indication to perform small area estimation (Eurostat, 2013). In our case, Relative Theil and Atkinson ($\varepsilon = 0.5, 1$) indexes show very similar CV distributions with medians slightly lower than 18%, ranging totally from 6% to 54%. This means that half of the domains has non-reliable estimates. On the other hand, the CV of the Gini index is significantly lower, confirming its robustness to outliers, ranging from 3% to 27%, with 7 domains having out-of-bound CV. This further motivates us to employ a small area model.

3 Small Area Models

We propose a Beta mixture model for small area estimation which from now on is named Flexible Beta (FB) model as in Migliorati et al. (2018). In order to evaluate its performance, we compare the estimation results with those obtained by the well-known Beta small area model. We start describing the Beta model in Subsection 3.1, then we set out our proposal in Subsection 3.2. Both models are completed by the prior setting described in Subsection 3.3 and the Bayesian estimation detailed in Subsection 3.4.

3.1 The Beta Model

Let us consider the Beta distribution with mean-precision parametrization (Ferrari & Cribari-Neto, 2004), such that a Beta distributed random variable is denoted with $Y \sim Beta(\mu\phi, (1-\mu)\phi)$, and has probability density function

$$f_B(y;\mu,\phi) = \frac{\Gamma[\phi]}{\Gamma[\mu\phi]\Gamma[(1-\mu)\phi]} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1.$$

Mean and variance are respectively

$$\mathbb{E}[Y] = \mu, \qquad \mathbb{V}[Y] = \frac{\mu(1-\mu)}{\phi+1}, \tag{12}$$

with $0 < \mu < 1$ and $\phi > 0$. A classical Beta small area model for y_d, denoting the direct estimator of a generic inequality measure and $\boldsymbol{x}_d$ a set of P covariates for domain d, constitutes as a hierarchical model with two levels. At the sampling level, the conditional distribution of the direct estimator is modeled as

$$y_d|\theta_d, \phi_d \overset{ind}{\sim} Beta(\theta_d\phi_d, (1-\theta_d)\phi_d), \quad \forall d.$$

In this case, $\mathbb{E}[y_d|\theta_d, \phi_d] = \theta_d$ is the target parameter and is estimated via a logit regression at the linking level, i.e. $\text{logit}(\theta_d)|\boldsymbol{\beta}, v_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + v_d$, with $v_d|\sigma_v^2 \overset{ind}{\sim} \mathcal{N}(0, \sigma_v^2)$ being an area specific random effect.

In literature, small area Beta models assume ϕ_d as known, in parallel with the known sampling variance assumption of the classical Fay-Herriot model, in order to allow identifiability. Being usually employed for proportions, this parametrization extremely simplifies the posterior geometry, given that, in this case, the variance structure is $\mathbb{V}[y_d|\theta_d, \phi_d] = \theta_d(1-\theta_d)/n_d$ under a binomial process. Thus, by combining it with (12), $\phi_d + 1$ can be seen as the effective sample size under a complex survey scheme. Its estimation is carried out considering the design effect, so that $\phi_d + 1 = \tilde{n}_d\psi_{\text{ind}} = \tilde{n}_d/\text{deff}_{\text{ind}}$.

Different approaches have been adopted in small area context for the estimation of deff_{ind}:

- estimation as a unique parameter across all areas within the hierarchical model, such as in Bauder et al. (2015) and Souza and Moura (2016) in case of proportions,
- separate estimation via a variance smoothing model of a common parameter across all areas, as in Fabrizi et al. (2011), Fabrizi et al. (2016) and Fabrizi and Trivisano (2016) similar to the one applied in Subsection 2.1,
- separate estimation of a set of parameters varying across all areas through the methodology proposed by Kish (1992), based only on design weights, as in Wieczorek et al. (2012) and Liu et al. (2007). Kalton et al. (2005) found this approximation reasonably accurate for proportions between 0.2 and 0.8.

We decided to tackle the problem from a different perspective, by assuming the sampling variance as known, rather than ϕ_d, and we estimate it separately via a two-step procedure as in Subsection 2.1. This decision has been taken for several reasons. Firstly, a direct estimation of ϕ_d within the model appears cumbersome to carry on, due to the complex structure of the variance functions defined in Propositions 1 and 2, leading to a tricky and intractable parametrization. In second place, the known variance assumption is a standard approach across different small area models (e.g. the Gaussian ones). This preserves the set of assumptions and data inputs across different models, favouring consistency of diagnostic measures, such as the goodness-of-fit ones, and allowing for performance comparison and model selection.

3.2 The Flexible Beta Model

The Flexible Beta distribution, introduced by Migliorati et al. (2018), is a mixture of two Beta random variables with different locations and a common dispersion parameter. Its probability density function is

$$f_{FB}(\lambda_1, \lambda_2, \phi, p) = p \cdot f_B(y; \lambda_1, \phi) + (1-p) \cdot f_B(y; \lambda_2, \phi),$$

with $0 < \lambda_2 < \lambda_1 < 1$ distinct ordered means, in order to avoid label switching problems, $0 < p < 1$ mixing coefficient and ϕ common dispersion parameter. This mixture extends the variety of shapes of Beta distribution in terms of bimodality, asymmetry and tail behavior. Besides, it ensures that each component is distinguishable, being computationally tractable.

Our small area model proposal for y_d includes, at sampling level, the Flexible Beta as likelihood assumption:

$$y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p \overset{ind}{\sim} FB(\lambda_{1d}, \lambda_{2d}, \phi_d, p), \quad \forall d.$$

In this case, the expected value and dispersion parameters of the mixture components vary across areas, while the mixing proportion p remains fixed. Let us denote with $\boldsymbol{\eta}$ the entire set of parameters, namely $\boldsymbol{\eta} = (\lambda_{1d}, \lambda_{2d}, \phi_d, p)$. In line with Migliorati et al. (2018), the parametrization considered in order to carry on estimation is

$$y_d|\boldsymbol{\eta} \sim FB(\tilde{w}_d + \lambda_{2d}, \lambda_{2d}, \phi_d, p),$$

with $\tilde{w}_d = \lambda_{1d} - \lambda_{2d} > 0$ denoting the distance between mixture components. Under such model, the expected value and variance are defined respectively as

$$\mathbb{E}[y_d|\boldsymbol{\eta}] = \theta_d = \lambda_{2d} + p \cdot \tilde{w}_d, \tag{13}$$

$$\mathbb{V}[y_d|\boldsymbol{\eta}] = \frac{\theta_d(1-\theta_d) + p(1-p)\tilde{w}_d^2\phi_d}{\phi_d + 1}. \tag{14}$$

At the linking level, we model the mean of the lowest component with a logit regression, by preserving the Gaussian random effect assumption, as follows

$$\begin{aligned} &\text{logit}(\lambda_{2d})|\boldsymbol{\beta}, v_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + v_d \\ &v_d|\sigma_v^2 \overset{ind}{\sim} \mathcal{N}(0, \sigma_v^2) \quad \forall d. \end{aligned} \tag{15}$$

As opposed to the FB regression proposed by Migliorati et al. (2018) and to the classical Beta regression, the linear predictor does not model directly the mean but rather a mixture component mean λ_{2d}, which in this case can be seen as a pure location parameter.

Being θ_d our parameter of interest, we assume it as a result of the combination of a location component and a deviation from it, caused by the intrinsic skewness of the sampling distribution, as in (13). Since we are generally dealing with right-skewed distributions, we assume the lower mixture component mean as the pure location parameter (λ_{2d}) and parameters p and $\tilde{w}_d$ as the ones able to capture such deviations. If θ_d was modelled through a logit regression, the relation among parameters would imply $\theta_d = \text{logit}^{-1}(\boldsymbol{x}_d^T\boldsymbol{\beta} + v_d)$, letting the linking level parameters masking such effect. On the other hand, in our case, we consider the location λ_{2d} to be directly modelled at the linking level, separating (13) and (15) and letting p and $\tilde{w}_d$ free to account for area-specific deviations. In this way, our location-modelling approach unleashes θ_d estimation. This is confirmed by the fact that, when θ_d is rigidly modelled through a logit regression, we notice that its estimates are basically overlapping the ones of the Beta model in Section 3.1 and estimation time is higher. The modelling of a location parameter different from the mean at the linking level is well-established in small area literature, we recall zero or zero/one inflated Beta (Fabrizi et al., 2016; Wieczorek et al., 2012), and skew-normal models (Ferrante & Pacei, 2017; Ferraz & Moura, 2012).

Similarly to the Beta model, the sampling variance $\mathbb{V}[y_d|\boldsymbol{\eta}]$ is assumed to be known and replaced by a refined estimate $\hat{\mathbb{V}}[y_d]$, as shown in Subsection 2.1. The conditioning is not emphasized on the refined estimate to underline the fact that it is not a model estimate but rather an independent one (survey estimate), treated as a known value in a small area model. As a consequence, the dispersion parameter is not directly estimated and can be obtained from (14) as

$$\phi_d|\theta_d, p, \tilde{w}_d = \frac{\theta_d(1-\theta_d) - \mathbb{V}[y_d|\boldsymbol{\eta}]}{\mathbb{V}[y_d|\boldsymbol{\eta}] - p(1-p)\tilde{w}_d^2}. \tag{16}$$

Since estimation requires a variation-independent parametrization, we decided to leave λ_{2d}, ϕ_d, and p free to assume any value of their support and to constrain $\tilde{w}_d$, as in the following Proposition.

Proposition 3. *Under FB model and the assumptions of Proposition 2, let us consider ϕ_d and its relation with the other parameters defined in (16). In order to preserve its bounded support, i.e. $\phi_d > 0$, $\tilde{w}_d$ has to be constrained such that*

$$\begin{cases} \tilde{w}_d < \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}} & \text{if} \quad \theta_d < c \\ \tilde{w}_d > \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}} & \text{if} \quad \theta_d > c \end{cases} \tag{17}$$

with c being a threshold that varies according to n_d and the measure considered: is equal to $n_d/(n_d+2)$ for Relative Theil index, $1/2 \times (\sqrt{4n_d+1}-1)$ for Gini index and does not have closed form for Atkinson index.

Proof. By imposing $\phi_d > 0$ on (16), the solution comes out to be

$$\mathbb{V}[y_d|\boldsymbol{\eta}] \in \left(\min\left\{\theta_d(1-\theta_d), p(1-p)\tilde{w}_d^2\right\}, \max\left\{\theta_d(1-\theta_d), p(1-p)\tilde{w}_d^2\right\}\right). \tag{18}$$

After splitting the problem into two cases, namely

$$\mathbb{V}[y_d|\boldsymbol{\eta}] < \theta_d(1-\theta_d) \qquad \text{and} \qquad \mathbb{V}[y_d|\boldsymbol{\eta}] > \theta_d(1-\theta_d) \tag{19}$$

we substitute $\mathbb{V}[y_d|\boldsymbol{\eta}]$ in (19) with (8), (4) and (11). The results for all three measures have respectively the following shapes $\theta_d < c$ and $\theta_d > c$, where c depends on n_d and differs for any measure. Then, we solve equation (18) for $\tilde{w}_d$ for each case obtaining (17). □

To understand the behaviour of previous constraints within our inferential problem, we evaluated them by considering the case $n_d = 2$, as a degenerate case with maximum variance. Indeed, a lower sample size does not allow for inequality measurement. We numerically derive the minimum of c, being 0.50 for the Relative Theil index, 0.84 for Atkinson indexes and 1 for the Gini index. Discarding the Gini index being always $\theta_d^G < 1$, the case $\theta_d > c$ is totally implausible for all the other income inequality measures. Indeed, those values correspond to far-fetched values of log income variable dispersion that, following (9) and (5), equals to $\varphi_d^2 > 0.69$ for Relative Theil index, $\varphi_d^2 > 3.71$ for Atkinson indexes. To be clear, a log-normal fitting on 2017 EU-SILC equivalent disposable income done by De Nicoló et al. (2021), shows $\hat{\varphi}^2 = 0.18$. Therefore, we opt to consider only the case $\theta_d < c$ in Proposition 3; the same reasoning could be easily done in case of proportions, where $\theta_d < c$ holds for any $n_d > 1$. An evaluation of the admissible support of θ_d in the considered case at varying n_d can be found in Section S.3 of the Supplementary Material.

As a consequence, the range of $\tilde{w}_d$ is defined as

$$\tilde{w}_d \in \left(0, \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}}\right\}\right), \tag{20}$$

where the upper bound of its support has a double vinculum. The first term, on λ_{2d}, can be seen as a support vinculum, since it allows θ_d to be upper bounded, i.e. $\theta_d < 1$. The second one follows from Proposition 3 and clearly takes into account the fact that the imposed sampling variance (given p) has to constrain the distance between mixture components. Thus, the distance has been modelled in line with (20) as

$$\tilde{w}_d = w \cdot \max \text{supp}\{\tilde{w}_d\} = w \cdot \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}}\right\}$$

unleashing parameter w free to vary in $(0, 1)$, being common across the areas. The underlying assumption implies that, for any given domain, different determinations of each direct estimator pertain to two latent groups, one of which displays a greater mean than the other, for any given set of covariates. Parameter w retains the meaning of distance between the regression functions of the two groups and it can be called the normalized distance (Migliorati et al., 2018).

The underlined parametrization is variation independent without penalizing the interpretability of the parameters. Moreover, the target parameter defined in (13) can be rewritten as

$$\theta_d|\lambda_{2d}, p, w = \lambda_{2d} + p \cdot w \cdot \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}}\right\}, \tag{21}$$

being a sum between the location parameter λ_{2d} and a second term depending on λ_{2d}, the sampling variance, the mixing parameter p and the normalized distance w. This shape permits to grasp an alternative interpretation of w as a factor that regulates the impact of sampling variance on θ_d (given p). Indeed, due to the different scale, usually $(1-\lambda_{2d})/p > \sqrt{\mathbb{V}[y_d|\boldsymbol{\eta}]}/\sqrt{p(1-p)}$. Note that the structure of θ_d is quite similar to its corresponding parameter in case of skew-normal likelihood (Ferrante & Pacei, 2017; Moura et al., 2017). In this case, the expected value is the sum of the location parameter and another component that depends on the known sampling variance and a skewness parameter.

Given the above considerations, as long as the sampling variances decrease, and presumably area sample sizes increase, the conditional distribution of direct estimator y_d stretches to a Beta distribution:

$$y_d|\boldsymbol{\eta} \overset{n_d \to +\infty}{\sim} Beta(\theta_d\phi_d, (1-\theta_d)\phi_d). \tag{22}$$

The expected value tends to the linear predictor and $\phi_d \to \theta_d(1-\theta_d)/\hat{\mathbb{V}}[y_d] - 1$ as in (12). Moreover, it is well known that a Beta distribution with large shape parameters, i.e. low variance, converges to a normal distribution. Therefore, it is possible to state that, as the sampling variance tends to 0, our sampling model tends asymptotically to the Gaussian one. Also when $p \to 0$, $p \to 1$ or $w \to 0$, (22) is verified, as common in degenerate mixture models (Fruhwirth-Schnatter et al., 2019).

To summarize, the parameter to be estimated in FB model are the ones related to the linear predictor $\boldsymbol{\beta}$ and the random effect variance σ_v, the mixing coefficient p and the normalized distance between mixture components w. Parameters p and w adjust the predictor depending on the magnitude of its sampling variance, guaranteeing a more flexible mean modelling as shown in (21). This can be seen as the main characteristic of the FB small area model.

3.3 Prior Distributions

The following weakly-informative priors complete the Beta model:

$$\sigma_v \sim \text{Half-}\mathcal{N}(0, \nu^2) \quad \text{and} \quad \beta_k \sim \mathcal{N}(0, \sigma^2), \quad \forall k = 1, \ldots, P \tag{23}$$

with $\sigma^2 = 10$. Considering the scale of the logit transformation, $\nu^2 = 1$ can be seen as quite a non-informative option. For the FB model, the choice of the prior includes in addition to (23)

$$p \sim \text{Unif}(0, 1) \quad \text{and} \quad w \sim \text{Unif}(0, 1). \tag{24}$$

This can be seen as a general formulation in line with Goodrich et al. (2018). However, in some specific cases e.g. when dealing with a few areas, we recommend using a slightly informative prior for the mixing coefficient to foster convergence or avoid convergence problems, such as $p \sim Beta(2, 2)$. Such an option enables to avoid the boundaries of its support while being very close to a uniform distribution.

3.4 Model Estimation

We estimate the model by adopting a Hierarchical Bayes (HB) approach. This approach to inference has several benefits in the SAE context (Rao & Molina, 2015, section 10), as to easily manage non-Gaussian distributional assumptions and to capture the uncertainty about all target parameters through the posterior distribution.

The FB model falls within the definition of a finite mixture, thus it could be seen as an incomplete data model where the allocation of observations to each mixture component is an unknown and latent component. In this case, a Bayesian approach based on Markov Chain Monte Carlo (MCMC) techniques is particularly suitable for posterior exploration. Specifically, the fitting was carried out by implementing the no-U-turn sampler (Hoffman & Gelman, 2014), an adaptive variant of Hamiltonian Monte Carlo (HMC) algorithm via `Stan` language (Carpenter et al., 2017). The HMC exploits differential geometry properties of the posterior distribution, in order to improve MCMC efficiency (Betancourt, 2017). We performed estimation by using 4 chains, each with 5,000 iterations, discarding the first 2,000 as warm-up.

Within the HB framework, we assume a quadratic loss and define as point predictor of θ_d its posterior expected value, namely

$$\hat{\theta}_d^{HB} = \mathbb{E}[\theta_d|\text{data}] \quad \forall d, \tag{25}$$

hereafter named model-based estimate. The posterior variance of the target parameter is used to describe its uncertainty.

An important property of the Fay-Herriot model is that, under the assumption of known random effect variance in HB context, the predictors are the outcome of a shrinkage process in between the direct estimate y_d and the synthetic estimate, being bounded. Predictors tend towards y_d when sampling variance is small in comparison with model variance, and towards synthetic estimate when it goes the other way round. In case of Beta assumption, predictors are not bounded. However, Janicki (2020) proved its asymptotic behavior, showing that it tends towards y_d when sampling variance goes to zero, and towards the synthetic estimate, in this case $\text{logit}^{-1}(\boldsymbol{x}_d^T\hat{\boldsymbol{\beta}})$, when model variance goes to zero. The first property, also known as design consistency, has been proved for the Beta model also by Fabrizi et al. (2020) relying on asymptotic Gaussianity. Thus, given the asymptotical behavior of our model in (22), we can state that the design-consistency property is preserved also under the FB model.

The FB small area model can be applied to any unit interval response and its estimation has been implemented in the R package `tipsae` (De Nicoló & Gardini, 2022), together with small area specific diagnostics functions and other complementary tools.

4 Application on EU-SILC Data

We are interested in estimating inequality in Italian NUTS-3 regions by using EU-SILC data. Given the high level of uncertainty of the direct estimates, described in Subsection 2.1, we

employ small area models considering both Beta and FB likelihoods. We estimate four separate univariate models for each likelihood, referring to the four different inequality measures. A survey data description directly follows, while auxiliary variables, from other sources, are set out in Subsection 4.1. Eventually, model results are compared in Subsection 4.2.

The EU–SILC survey (Guio, 2005) collects cross-sectional and longitudinal microdata on income, poverty, social exclusion and living conditions in a timely manner. The survey is conducted in each country by the National Institute of Statistics and coordinated by Eurostat, guaranteeing consistent methodology and definitions across all EU member states. The sampling design involves a rotational panel lasting four years, where each year one-quarter of all respondents is newly introduced. As regards the Italian sample provided by ISTAT, the survey units (households) are sampled according to a complex survey scheme involving stratification and two-stage selection. The first-stage units are municipalities, stratified accordingly to the demographic size, the ones with great size are considered as self-representative units and form a take-all stratum. Within the selected primary units, households are drawn randomly as secondary sampling units.

In our case, we concentrate on the 107 NUTS-3 Italian domains by using the 2017 wave. The sample comprises 22,226 households and 48,819 corresponding individuals. The domain size ranges from a minimum of 32 to a maximum of 2,536 individuals; with 25th, 50th and 75th percentiles respectively as 196, 314, 612 (from 18 to 1,270 households; with percentiles 86, 138, 275).

4.1 Auxiliary Variables

The possible determinants of income inequality within European regions have been identified by Perugini and Martino (2008). According to them, the main ones are human capital endowment, labour market performances, economic development, industrial specialization and the demographic structure.

A small area model does not have causal inference ambitions, but rather it requires auxiliary information to be accurately known at population level. Therefore, we restrict the choice to data currently available and measured without error: census and registry office data as well as tax forms data. As a human capital endowment proxy, we calculated the ratio between the number of people aged 15–64 with a high school diploma or higher level of education, and the number of people within the same age class with compulsory education level, based on the 2011 Italian population census data. Fabrizi and Trivisano (2016) refer to this indicator as to the people-in-higher-education ratio. The demographic structure is explained by areal population density and aged dependency ratios. Moreover, as suggested by Perugini and Martino (2008), we used the percentage of resident foreigners (immigrants) and the male/female resident foreigners ratio as indirect measures of economic development.

Concerning fiscal archives data, we include average taxable income claimed by private residents, the percentage of residents aged more than 15 filling tax forms and the percentage of residents with income lower than/greater than double the national median filling tax forms. These variables measure the affluence of income earners in the area and are adopted as indirect proxies of labour market performances (Fabrizi & Trivisano, 2016).

Lastly, in order to provide strongly correlated information, we add the corresponding inequality measures calculated on a discrete scale given the income classes declared by tax forms. Note that those measures are estimated on market income (i.e. income before taxes and transfers), while our target variable is the disposable income instead (after taxes and transfers). We

		Beta	FlexBeta
Atk(0.5)	looic	-572.2	-595.7
	(se)	(18.9)	(17.4)
	acvr %	43.2	51.1
Atk(1)	looic	-404.6	-425.5
	(se)	(18.6)	(17.7)
	acvr %	41.6	46.0
Relative Theil	looic	-966.8	-992.3
	(se)	(16.9)	(15.6)
	acvr %	36.9	47.3
Gini	looic	-388.3	-392.3
	(se)	(21.3)	(20.5)
	acvr %	38.5	38.3

Table 1: `looic` and related standard error as well as `acvr` for each model and each measure

obtain raw estimates of market income inequality, legitimately greater than our response due to the redistributive power of taxes and transfers on income distribution. This happens despite the missing component of variability within income classes, not captured by our market income inequality estimators. All the auxiliary variables were standardized before being incorporated into the models, in order to harmonize the scale, and subjected to preliminary variable selection to avoid multicollinearity.

4.2 Results

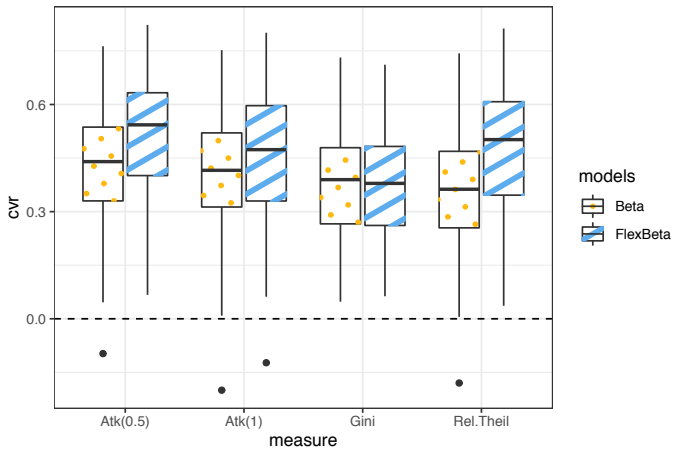

Figure 2: CV reduction for each model and each measure

The model estimation has been carried out and posterior draws have been validated through MCMC diagnostics, showing good chain mixing and quick convergence for any measure. The estimation times are between 20.62 and 26.90 CPU seconds for the Beta model and between 71.70 and 80.38 CPU seconds for the Flexible Beta model on an Intel Core i5 dual-core (1,8GHz) processor. A model comparison has been also performed through specific model diagnostics. Concerning goodness-of-fit and model comparison, the `looic` measure, based on leave-one-out cross-validation (Vehtari et al., 2017), and its standard error have been used. The `looic` is

preferred over the most classical DIC and AIC measures, since it is fully Bayesian, using the entire posterior distribution, it is invariant to parametrization and it works for singular models.

In order to evaluate model-based estimators performances in comparison with direct estimators, the Coefficient of Variation Reduction measure (Ferrante & Pacei, 2017) is used to calculate the precision improvement:

$$\texttt{cvr}_d^{HB} = 1 - \frac{\mathbb{V}[\theta_d|\text{data}]^{\frac{1}{2}} \cdot y_d}{\hat{\mathbb{V}}[y_d]^{\frac{1}{2}} \cdot \mathbb{E}[\theta_d|\text{data}]} \quad \forall d,$$

using predictors and their posterior variances for any HB model. This constitutes a frequently used measure for small area model evaluation. However, the comparison between CV of model-based and design-based estimators might sometimes be spurious since the former could be design-biased even when the model is correctly specified. A discussion about it can be found in Ferrante and Pacei (2017). Thus, we perform such comparison aware that the bias properties of our estimators have to be further deepened through the simulation study of Section 5.

Diagnostics for each model, `looic` and `cvr`, averaged among all domains (`acvr`), are displayed in Table 1, while the full distribution of `cvr` is displayed in Figure 2. Results show better goodness-of-fit and coefficient of variation reduction for the Flexible Beta model compared to the Beta one. This holds for all measures with the exception of the Gini index, where diagnostics do not vary significantly between models. Following the distributional analysis conducted in Section S.2 of the Supplementary Material, the sampling distribution of Gini index estimator does not show departures from Gaussianity in small samples, in comparison with other measures, presenting only light skewness and lepto-kurtosis. As a consequence, the employment of a mixture model seems to be irrelevant for a substantial improvement of the estimates.

The FB model allocates greater density on the right-hand tail of estimator distribution, being able to better capture it. This aspect is clear from density plots in Figure 3, displaying direct estimates versus model-based estimates of Beta and FB models in the 107 domains. Notice that any data point refers to a domain, being the expected values of different posterior distributions as in (25), thus let us consider them as global distributions not related to domain-specific posteriors. The FB model-based estimates tend to be greater than Beta model ones as clear from scatterplots in Figure 3, capturing the right-hand tail of the distribution and avoiding to underestimate inequality, as it will be more intelligible in Section 5. The Gini index case shows, again, large similarities across the two models.

Deep diving on FB estimates, Figure 4 displays posterior means distributions of mixture components expected values λ_{1d} and λ_{2d}, $\forall d$, weighted for $\mathbb{E}[p|\text{data}]$, in comparison with direct estimates and Flexible Beta model-based estimates. The posterior means of the mixing coefficient are 0.83 for both Atkinson indexes, 0.85 for Relative Theil and 0.63 for Gini index, due to its more symmetric distribution. We may notice from Figure 4 that the mixture component embracing lower inequality values has always less weight.

The shrinking process is displayed in Figure 5 for each model. It appears distinctly that estimates are more shrunk in the case of Beta model, as highlighted by the distance between linear regression and bisector lines. Moreover, a property that should be desirable for any small area model is the consistency among direct estimates and model-based ones, namely direct estimates outliers should not be completely pushed towards the opposite tail as model estimates. This consistency property does not hold in case of Beta models, having 2 or 3 top outliers pushed towards the lowest values of the distribution. This is due to the strong impact that auxiliary variables have on the outcome and to the little flexibility of the model. On the

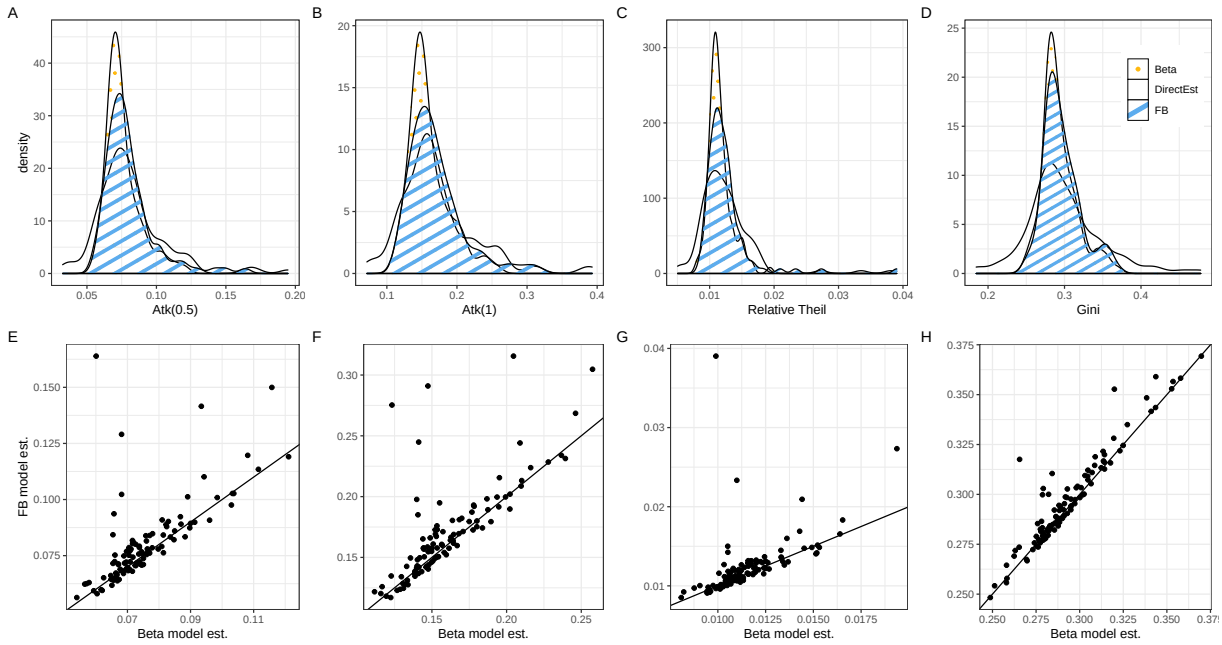

Figure 3: Densities of model-based estimates versus direct estimates and scatterplot of model-based estimates with bisector line

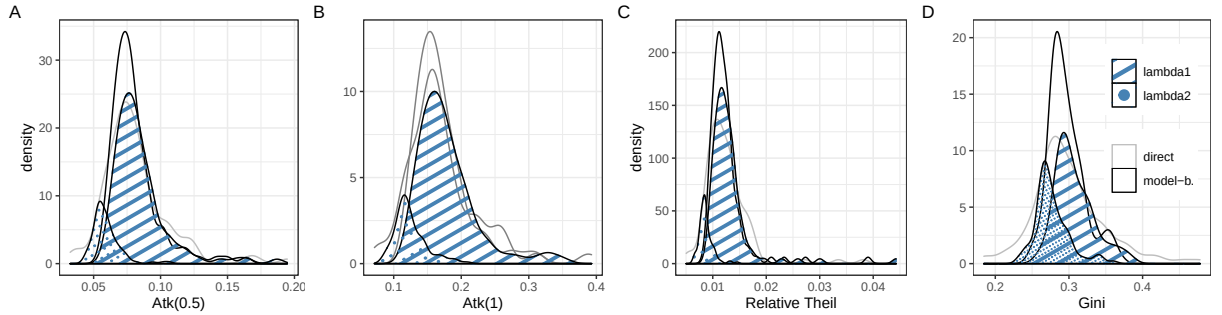

Figure 4: Posterior means distributions of the mixture components expected values λ_{1d} and λ_{2d}, $\forall d$, weighted for $\mathbb{E}[p|\text{data}]$, in comparison with direct estimates and Flexible Beta model-based estimates

contrary, the FB model keeps its model estimates consistent with their input ones, operating overall less shrinkage. Another desirable property is the design consistency, i.e. $(\hat{y}_d - \theta_d^{HB}) \to 0$ as long as $\tilde{n}_d$ increases. This property hold for all models, as clear from Figure 6. Notice that the magnitude of residuals for Beta models is relevant for domains with smaller sample sizes and strongly unbalanced on positive residuals. This makes sense, given the strong shrinkage operated on high outliers. Posterior summaries of regression coefficients in the FB model are detailed and commented in Section S.4 of the Supplementary Material.

5 Design-Based Simulation

A design-based simulation study has been carried out to evaluate the frequentist properties of the FB model-based estimators in comparison with the Beta ones. We consider the Italian EU-SILC sample as synthetic population and the 14 metropolitan cities and the remaining 21

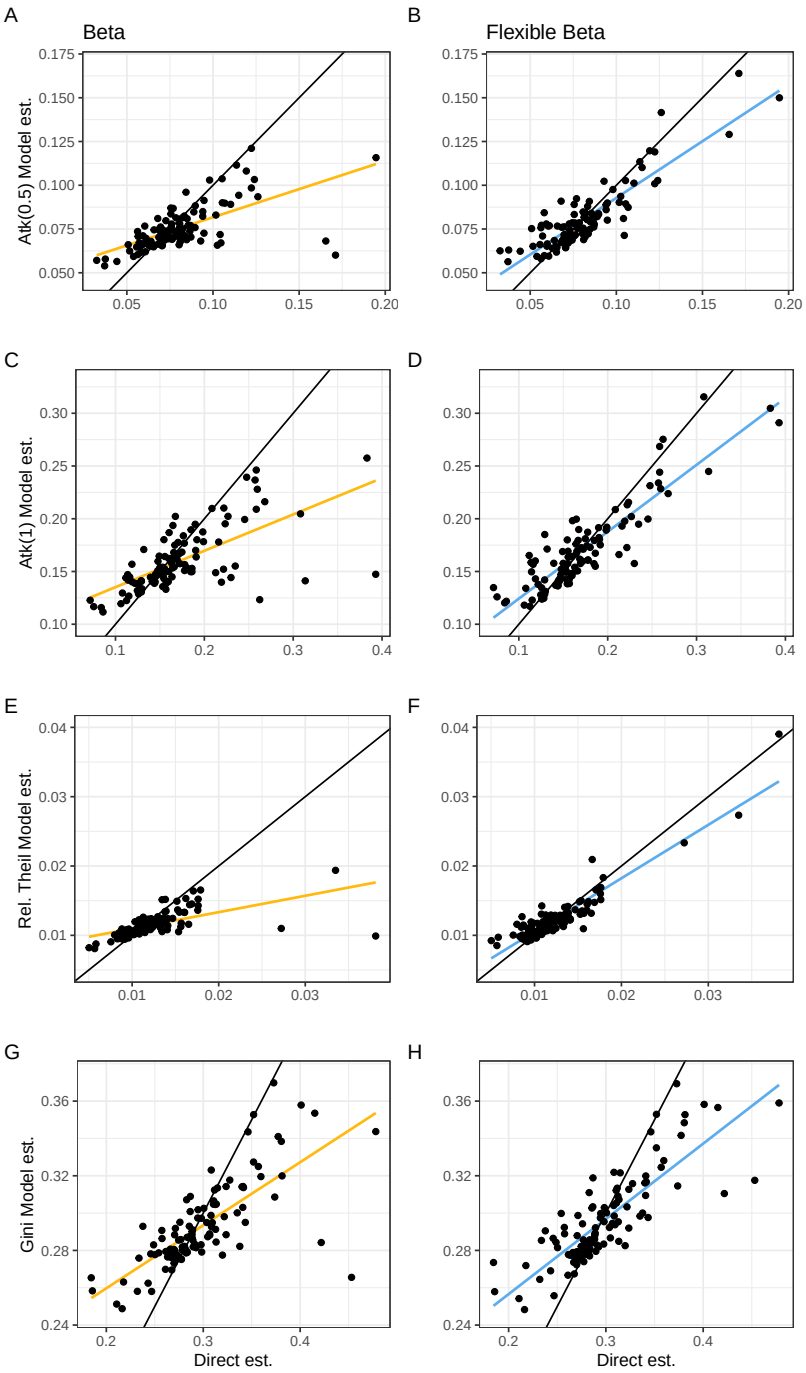

Figure 5: Direct estimates versus model-based estimates for each measure in Beta and Flexible Beta models: bisector in black, coloured linear regression line

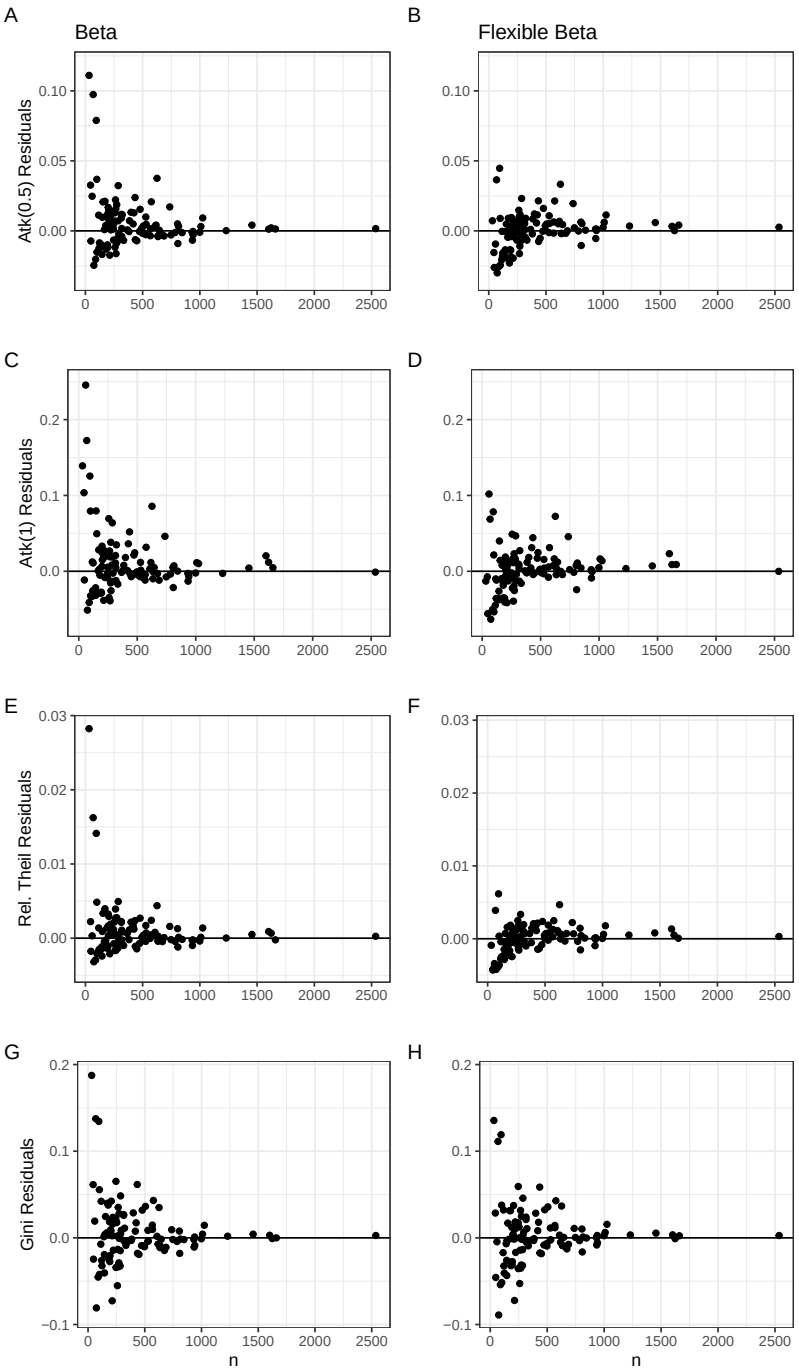

Figure 6: Design consistency check (sample sizes vs residuals) for each measure in Beta and Flexible Beta models

administrative regions as synthetic domains. In order to deal with a sufficient number of areas, assuring at the same time high variability of direct estimates (i.e. keeping synthetic population sufficiently large), we pool 2009, 2013 and 2017 EU-SILC waves as independent and separate populations with a total of 105 domains. The study is not based on generated data under some specific income distributional assumption, since the aim is to check whether this framework can work with close-to-reality income data, affected by peculiar problems, e.g. extreme values.

From each domain, $S = 1,000$ samples have been repeatedly extracted by mimicking complex EU-SILC design, with stratification, multi-stage selection and the distinction between self-representative and non-self-representative strata. We adopted three different scenarios for simulation, the first two involve different sampling rates, respectively 3% and 5%. Note that observations included in 3% samples have been selected from those of 5% samples to attenuate the effect of sampling variability. The third scenario consists of running the simulation on 5% samples with smoothed income data, where an extreme values treatment (`evt`) has been performed as previously described in Subsection 2.1 (Finkelstein et al., 2006; Masseran et al., 2019). The drawing is done at the household level, with a total of extracted individuals in each domain ranging respectively from 9 to 115 (median equals 31) for 3% and from 9 to 228 (median equals 63) for 5%.

Bias-corrected inequality estimators have been calculated for any extracted sample, and a suitable set of covariates have been selected among the ones cited in Subsection 4.1. Covariates have been calculated at the corresponding geographical detail for the 105 synthetic domains. At a lower level of geographical disaggregation, such as in this case (NUTS-2 regions), the correlation among covariates and between covariates and response is stronger in comparison with our application setting (NUTS-3 regions). This is coherent since raw estimates are measured at macro level, inducing less error. For any iteration $s = 1, \dots, S$, Beta and FB model has been estimated with a fixed set of covariates. The sole distinction with the setting adopted in Section 4 regards the prior of mixing coefficient p in (24), substituted with $p \sim \text{Beta}(2, 2)$, in order to speed up convergence and save computational time.

Considering the generic model-based estimate at iteration s for domain d as $\hat{\theta}_{ds}^{HB}$ and the corresponding population value θ_d, we define Relative Bias (RB), Absolute Relative Bias (ARB), Mean Squared Error (MSE), Relative Mean Squared Error (RMSE) and Average Effect (AEFF) as follows:

$$\text{RB}(\hat{\theta}_d^{HB}) = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\theta}_{ds}^{HB}}{\theta_d} - 1\right),$$

$$\text{ARB}(\hat{\theta}_d^{HB}) = \left|\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\theta}_{ds}^{HB}}{\theta_d} - 1\right)\right|,$$

$$\text{MSE}(\hat{\theta}_d^{HB}) = \frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_{ds}^{HB} - \theta_d)^2,$$

$$\text{RMSE}(\hat{\theta}_d^{HB}) = \frac{\text{MSE}(\hat{\theta}_d^{HB})}{\theta_d^2},$$

$$\text{AEFF}(\hat{\theta}^{HB}) = \sqrt{\frac{\sum_{d=1}^{D}\text{MSE}(y_d)}{\sum_{d=1}^{D}\text{MSE}(\hat{\theta}_d^{HB})}}.$$

Lastly, we consider the frequentist coverage of credible intervals defined by the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior of θ_d,

$$\text{Coverage}_{1-\alpha}(\hat{\theta}_d^{HB}) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}(\theta_d \in [Q_{\alpha/2}[\theta_{ds}|\text{data}], Q_{1-\alpha/2}[\theta_{ds}|\text{data}]]),$$

where $Q_\pi[\theta_{ds}|\text{data}]$ denotes the posterior quantile of order π of θ_{ds}. The nominal coverage probability $1-\alpha$ is chosen to be equal to 0.95.

Simulations results are fully described for any setting in Table 2. RB, ARB, RMSE and Coverage are reported on average over the 105 simulations domains, showing that FB estimators outperform the Beta ones.

Focusing on estimates reliability, both Beta and Flexible Beta models perform significantly better than direct estimators: RMSE and AEFF show a great error reduction for all measures. Among them, the FB estimators perform better than the Beta ones in all cases. Moreover, the FB estimators show a noticeable bias decrease compared to the Beta one, as clear from ARB and RB values in Table 2, concerning both magnitude and direction. This confirms the clues of inequality underestimation under a Beta model, notwithstanding the measure adopted, and shows that the FB model consistently reduces this underestimation. The bias improvement is at the expense of a slight variance increase, but the bias-variance trade-off favors the FB model, as notable from RMSE and AEFF.

The full distribution of MSEs of direct and model-based estimators related to the different domains is depicted by boxplots in Figure 7. Firstly, all distributions show heavy right-tails, with several outlier domains having great error levels. Again, the error reduction induced by both small area models is noticeable, allowing estimators to borrow strengths across areas. Specifically, while RMSEs, displayed in Table 2, indicate on average a moderate error improvement for the FB model, the full distributions show a great reduction in case of outlier domains. This reduction, as clear from the bottom row plots of Figure 7, takes place in case of domains with the smallest sized samples. The greatest MSE reduction regards the Relative Theil Index, the lowest the Gini index.

Flexible Beta models produce credible intervals that exhibit a noticeable better performance in terms of coverage, in some cases outperforming Beta intervals coverage by more than 10% on average. Its trend over the sample sizes is displayed in Figure 8. While FB coverage rates converge to their nominal level in correspondence to 50 individual-sized samples, the Beta ones converge near 100 sized samples in case of Gini Index, near 130/150 in case of Atkinsons and Relative Theil Indexes.

The similarity among Beta and FB model-based estimates for the Gini index is confirmed also in the simulation setting. Nevertheless, the FB model has higher coverage in case of small samples. Concerning different simulation settings, the Relative Theil direct estimators show noticeably high bias in case of extreme value treated setting in comparison with other settings. This could be due to a failure of preliminary bias correction on direct estimators; indeed, input estimators does not satisfy unbiasedness, not allowing for comparability. The extremely low coverage of both model-based estimators is indeed justified by the high bias. Generally speaking, the performance gap between Beta and FB models increases at decreasing sampling rates/sizes and with no smoothed data. This denotes a progressive failure of the Beta model under the mentioned conditions, as clear from how fast its diagnostics get worse at varying settings.

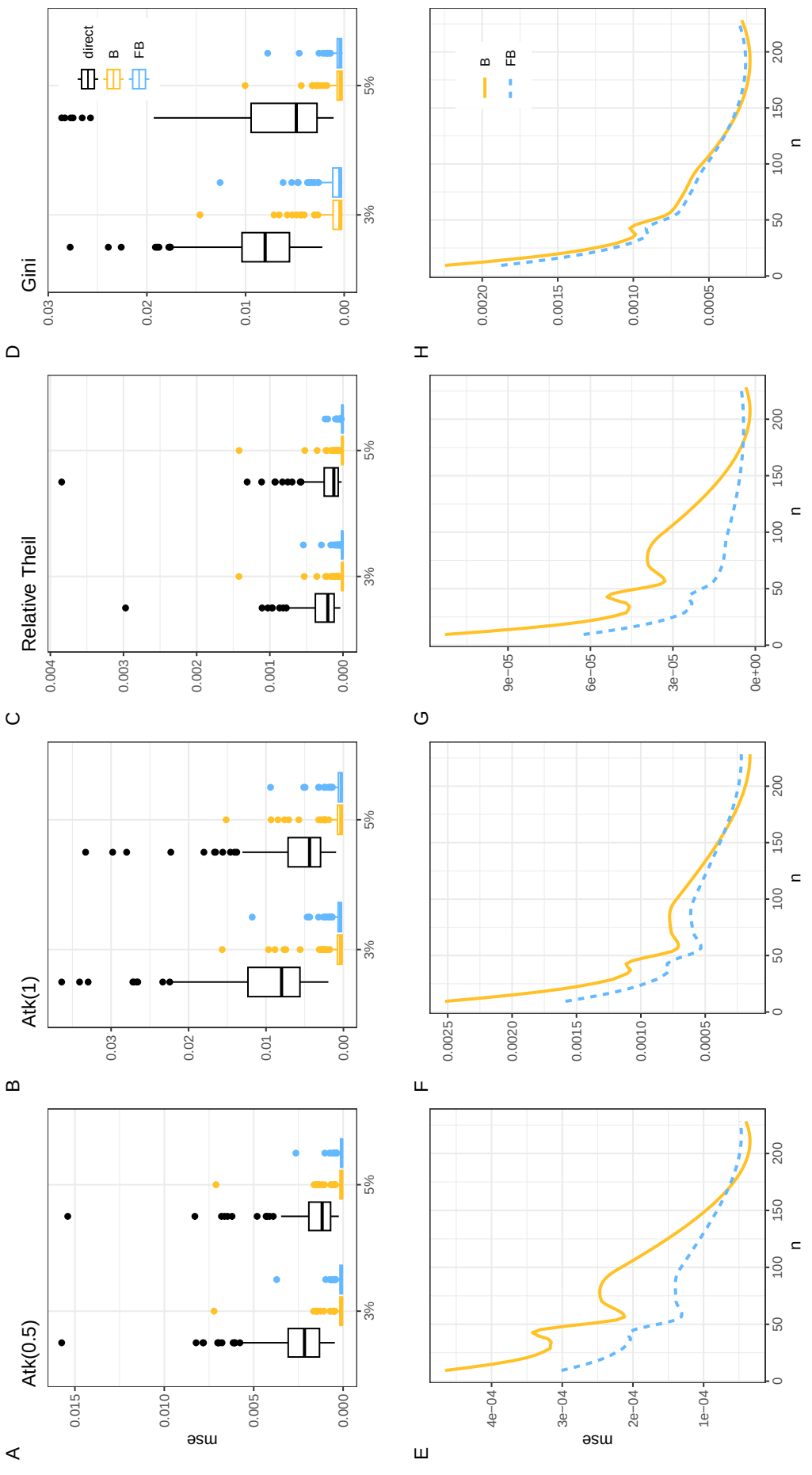

Figure 7: MSE for each area. Plots on the top row show direct estimators values versus model-based estimators ones, while bottom row plots show model-based estimates MSE versus the sample sizes, depicted through smoothed lines

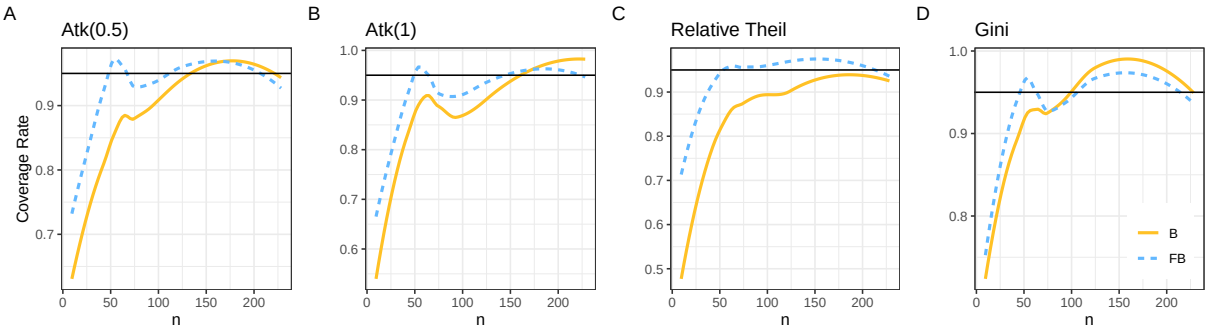

Figure 8: Coverage rate versus sample sizes for each measure and each model in 5% simulation samples depicted through smoothed lines, black line fixed at the nominal level 0.95

6 Conclusions

The reduction of inequalities, both within and between countries, is a prerequisite for achieving the Sustainable Development Goals of the 2030 Agenda for Sustainable Development, adopted by all United Nations Member States. At regional level, an increase in disparities within regions has been observed, whereas regional disparities between European countries are gradually decreasing. Several low-growth regions exist within EU member states and, among the richest states, it is possible to find areas characterised by high levels of poverty and inequality (often post-industrial or rural ones). In this context, inequality indicators at a regional-level breakdown would allow us to shift towards a more comprehensive and multifaceted view of territorial convergence in the EU and to better understand causal mechanisms, essential for regional-targeted policies.

In this study, we propose a SAE model that aims at obtaining reliable estimates of the most common inequality indicators: the Gini index, the Relative Theil index and two Atkinson indexes defined for two different values of the inequality aversion parameter. By considering that inequality estimators are unit-interval defined, skewed and heavy-tailed, we propose a FB small area model. The results are really encouraging as the estimates we obtain outperform in different ways the most common Beta small area model, generally used for parameters defined on the unit interval. We may define the FB model as a good general-purpose model for unit interval data as the two-component mixture is able to capture skewness and heavy tails in a satisfying way, guaranteeing at the same time analytic tractability.

Our findings provide a basis for further research focused on multiple directions. An inequality mapping based on obtained estimates could be used to single out at first a territorial disparity analysis within the country, complementing the usual consideration of disparity between countries. Furthermore, the set of inequality measures should be properly complemented by quantile-based inequality indexes that, by focusing on distribution tails, are able to capture different aspects of the income distribution with respect to concentration indexes. Quantile-based indexes are not defined on the unit interval support and thus their estimation has to rely on different likelihood assumptions. Lastly, given the longitudinal nature of the EU-SILC survey, a model extension in this sense naturally follows, by considering subsequent waves at the same time.

Acknowledgements

The work of Silvia De Nicolò was partially funded by the ALMA IDEA 2022 grant (title: "Social exclusion and territorial disparities: poverty and inequality mapping through advanced methods of small area estimation", project J45F21002000001), part of the European Union - NextGenerationEU funding.

References

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*(3), 244–263.

Bauder, M., Luery, D., & Szelepka, S. (2015). *Small Area Estimation of Health Insurance Coverage in 2010-2013* (tech. rep.). Washington, DC: US Census Bureau.

Bayes, C. L., Bazán, J. L., & García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, *7*(4), 841–866.

Bellú, L., & Liberati, P. (2006). *Policy Impacts on Inequality, Welfare based Measures of Inequality: The Atkinson Index* (tech. rep.).

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv Preprint arXiv:1701.02434*.

Bianchi, A., Fabrizi, E., Salvati, N., & Tzavidis, N. (2018). Estimation and testing in m-quantile regression with applications to small area estimation. *International Statistical Review*, *86*(3), 541–570.

Biewen, M., & Jenkins, S. P. (2006). Variance estimation for Generalized Entropy and Atkinson inequality indices: the complex survey data case. *Oxford Bulletin of Economics and Statistics*, *68*(3), 371–383.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Cavanaugh, A., & Breau, S. (2018). Locating geographies of inequality: publication trends across OECD countries. *Regional Studies*, *52*(9), 1225–1236.

Cowell, F. (2011). *Measuring inequality*. Oxford University Press.

De Nicoló, S., Ferrante, M. R., & Pacei, S. (2021). Mind the income gap: Bias correction of inequality estimators in small-sized samples. *arXiv Preprint arXiv:2107.08950v3*. https://arxiv.org/abs/2107.08950

De Nicoló, S., & Gardini, A. (2022). *`tipsae`: Tools for Handling Indices and Proportions in Small Area Estimation* [R package version 0.0.6]. https://cran.r-project.org/web/packages/tipsae/index.html

Eurostat. (2013). *Handbook on Precision Requirements and Variance Estimation for ESS Households Surveys*. Publications Office of the European Union, Luxembourg.

Fabrizi, E., Ferrante, M. R., Pacei, S., & Trivisano, C. (2011). Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics & Data Analysis*, *55*(4), 1736–1747.

Fabrizi, E., Ferrante, M. R., & Trivisano, C. (2016). Bayesian Beta regression models for the estimation of poverty and inequality parameters in small areas. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 299–314). John Wiley; Son.

Fabrizi, E., Ferrante, M. R., & Trivisano, C. (2018). Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(4), 861–879.

Fabrizi, E., Ferrante, M. R., & Trivisano, C. (2020). A functional approach to small area estimation of the Relative Median Poverty Gap. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 1273–1291.

Fabrizi, E., & Trivisano, C. (2016). Small area estimation of the Gini concentration coefficient. *Computational Statistics & Data Analysis*, *99*, 223–234.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277.

Ferrante, M. R., & Pacei, S. (2017). Small domain estimation of business statistics by using multivariate skew normal models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1057–1088.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, *31*(7), 799–815.

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

Finkelstein, M., Tucker, H. G., & Alan Veeh, J. (2006). Pareto tail index estimation revisited. *North American Actuarial Journal*, *10*(1), 1–10.

Fruhwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of Mixture Analysis*. CRC press.

Gardini, A., Fabrizi, E., & Trivisano, C. (2022). Poverty and inequality mapping based on a unit-level log-normal mixture model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). Bayesian applied regression modeling via stan. *R package version 2.21*, *1*.

Gradshteyn, I. S., & Ryzhik, I. M. (2014). *Table of integrals, series, and products*. Academic press.

Guio, A.-C. (2005). Income poverty and social exclusion in the EU25. *Statistics in focus: population and social conditions*, (13), 1–7.

Harmening, S., Kreutzmann, A.-K., Schmidt, S., Salvati, N., & Schmid, T. (2020). A framework for producing small area estimates based on area-level models in r. http://mirror.epn.edu.ec/CRAN/web/packages/emdi/vignettes/vignette_fh.pdf

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Janicki, R. (2020). Properties of the Beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods*, *49*(9), 2264–2284.

Kalton, G., Brick, J., & Le, T. (2005). *Estimating Components of Design Effects for Use in Sample Design* (tech. rep.).

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, *8*(2), 183.

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*(2), 521–540.

Lasso de La Vega, C., & Urrutia, A. (2008). The Extended Atkinson family: The class of multiplicatively decomposable inequality measures, and some new graphical procedures for analysts. *Journal of Economic Inequality*, *6*(2).

Liu, B., Lahiri, P., & Kalton, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. *Proceedings of the American Statistical Association, Survey Research Section*, 3181–3186.

Marchetti, S., & Tzavidis, N. (2021). Robust estimation of the Theil index and the Gini coefficient for small areas. *Journal of Official Statistics*.

Márquez, M. A., Lasarte, E., & Lufin, M. (2019). The role of neighborhood in the analysis of spatial economic inequality. *Social Indicators Research*, *141*(1), 245–273.

Masseran, N., Yee, L. H., Safari, M. A. M., & Ibrahim, K. (2019). Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics*, *7*(3), 70–77.

Migliorati, S., Di Brisco, A. M., & Ongaro, A. (2018). A new regression model for bounded responses. *Bayesian Analysis*, *13*(3), 845–872.

Molina, I., & Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385.

Moser, M., & Schnetzer, M. (2017). The income-inequality nexus in a developed country: Small-scale regional evidence from Austria. *Regional Studies*, *51*(3), 454–466.

Moura, F. A., Neves, A. F., & Silva, D. B. d. N. (2017). Small area models for skewed brazilian business survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1039–1055.

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*(1), 265–286.

Perugini, C., & Martino, G. (2008). Income inequality within European regions: Determinants and effects on growth. *Review of Income and Wealth*, *54*(3), 373–406.

Piketty, T. (2014). Capital in the twenty-first century. *Capital in the twenty-first century*. Harvard University Press.

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. John Wiley & Sons.

Rao, J. N., & Molina, I. (2015). *Small-Area Estimation*. Wiley Series in Survey Methodology.

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.

Sen, A. (1997). *On Economic Inequality*. Oxford University Press.

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.

Souza, D. F., & Moura, F. A. (2016). Multivariate Beta regression with application in small area estimation. *Journal of Official Statistics*, *32*(3), 747–768.

Theil, H. (1979). World income inequality and its components. *Economics Letters*, *2*(1), 99–102.

Tzavidis, N., & Marchetti, S. (2016). Robust domain estimation of income-based inequality indicators. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 171–186). Wiley Online Library.

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.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Wieczorek, J., Nugent, C., & Hawala, S. (2012). A Bayesian zero-one inflated Beta model for small area shrinkage estimation. *JSM Proceedings*, 3896–3910.

Measure	Scenario		ARB%	RB%	RMSE%	AEFF	Cov. 95%
Atk(1)	evt						
		direct	1.44	-0.99	13.99		
		Beta	9.90	-3.06	2.10	2.21	87.00
		FB	8.78	-0.22	1.81	2.65	89.70
	5%						
		direct	2.00	-1.42	19.48		
		Beta	11.15	-2.46	2.44	2.46	84.98
		FB	9.38	-0.12	2.04	2.97	90.28
	3%						
		direct	3.89	-3.68	32.77		
		Beta	11.49	-2.17	2.61	3.07	85.82
		FB	9.12	-0.74	2.33	3.54	91.50
Atk(0.5)	evt						
		direct	1.69	-1.24	15.87		
		Beta	9.56	-2.75	1.96	2.50	87.64
		FB	8.41	-0.23	1.67	2.95	91.19
	5%						
		direct	2.40	-1.92	21.90		
		Beta	10.85	-2.41	2.35	2.60	85.25
		FB	8.85	-0.04	1.85	3.34	92.07
	3%						
		direct	4.78	-4.59	34.72		
		Beta	11.06	-2.42	2.47	3.14	86.30
		FB	8.67	-1.28	2.08	3.79	93.01
Relative Theil	evt						
		direct	9.69	-5.14	18.73		
		Beta	12.55	-6.52	3.59	1.94	70.72
		FB	9.71	-3.72	3.31	2.22	73.31
	5%						
		direct	3.63	-3.23	27.87		
		Beta	12.30	-5.02	3.21	2.36	81.79
		FB	7.69	-0.80	1.90	3.75	92.84
	3%						
		direct	7.48	-7.35	39.90		
		Beta	12.71	-4.82	3.38	2.62	80.76
		FB	8.32	-3.33	2.55	3.44	90.68
Gini	evt						
		direct	2.58	-1.13	4.75		
		Beta	4.60	-1.06	0.53	2.85	91.50
		FB	4.43	-0.28	0.50	3.01	93.23
	5%						
		direct	3.29	-0.10	7.43		
		Beta	5.10	-0.58	0.69	3.16	91.23
		FB	4.90	0.68	0.67	3.33	93.04
	3%						
		direct	6.20	-5.99	9.02		
		Beta	6.21	-5.19	0.99	2.75	89.84
		FB	5.79	-4.80	0.92	2.89	92.09

Table 2: Absolute Relative Bias, Relative Bias, Relative Mean Squared Error, Average Effect, jointly with the 95% coverage on average of direct estimators and model-based estimators related to both Beta and FB models for extreme value treated data (evt), 5% and 3% sampling rates. Quantities denoted with the bar are averaged among all areas.