UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXIV

Small Area Estimation of Inequality Measures

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Maria Rosaria Ferrante

Dottoranda: Silvia De Nicolò

28 Febbraio 2022

Abstract

The demand for income inequality estimates, referring to local areas or specific subpopulations, is growing due to their relevance for policy-making and applied research in regional and inequality studies. Income data are usually collected through household surveys where local domains fall outside the prior design plan, resulting in small-sized samples and yielding unreliable direct estimation. We cope with it by relying on Small Area Estimation (SAE) methods. Such techniques exploit auxiliary information to borrow strength across areas and produce estimates of interest with an acceptable level of uncertainty.

This dissertation aims at bringing together the world of SAE methods and the one of inequality measurement by tackling the issues of small samples bias and not well-behaved distributions of inequality estimators. Our approach lies within the framework of Bayesian inference of area-level models. Two methodological proposals have been developed in this regard: a bias correction proposal for finite population and complex survey design setting, and a small area model based on a Beta mixture, for mapping double-bounded and not-well behaved responses. The third and last part of this dissertation exposes a computational implementation, resulting in an R package, supplying a more general tool for mapping indices and proportions using Beta-based small area models. It comes equipped with a set of diagnostics and complementary tools, visualizing and exporting functions to assist the user in carrying out a complete analysis.

Sommario

La crescita della disuguaglianza economica negli ultimi 30 anni è documentata da una grande mole di informazioni statistiche. Di recente è emersa l'esigenza di utilizzare, a fini di pianificazione ed implementazione delle politiche economiche, misure di disuguaglianza a livello locale e/o per specifiche sottopopolazioni. Generalmente, le stime riferite ai parametri di disuguaglianza sono ottenute mediante informazioni raccolte in indagini campionarie. Queste non sono pianificate per la stima di indicatori a livello locale e, dunque, presentano spesso ampiezza campionaria non sufficiente ad ottenere stime affidabili con gli usuali stimatori basati sul disegno. Per far fronte a tale problema si ricorre a metodi di "stima per piccole aree". Queste tecniche integrano stime dirette ed informazioni ausiliarie al fine di produrre stime con un livello accettabile di errore.

Questa dissertazione ha lo scopo di connettere il mondo della stima per piccole aree con quello della misurazione della diseguaglianza economica. Vengono affrontate le questioni relative alla distorsione ed alle distribuzioni asimmetriche a code alte degli stimatori di diseguaglianza con riferimento a piccoli campioni. L'approccio qui adottato si inquadra nella cornice dell'inferenza Bayesiana per modelli definiti a livello di area. A questo proposito, la tesi contiene due proposte metodologiche: la prima fornisce un metodo di correzione della distorsione di alcuni stimatori di diseguaglianza nel contesto delle popolazioni finite e delle indagini complesse, la seconda definisce un modello per piccole aree basato su una mistura di distribuzioni Beta. Quest'ultima proposta si rivela particolarmente utile nel mappare indicatori definiti sull'intervallo unitario con distribuzioni marcatamente non Gaussiane. La terza ed ultima parte di questa dissertazione espone una implementazione computazionale, che costituisce un pacchetto R, per la mappatura di proporzioni ed indici definiti nell'intervallo unitario attraverso modelli per piccole aree basati sulla distribuzione Beta. Il pacchetto include una serie di specifiche diagnostiche ed alcuni strumenti complementari, al fine di assistere l'utente durante l'intera procedura di analisi.

”Chi vi credete che noi siam
Per i capelli che portiam
Noi siamo delle lucciole
Che stanno nelle tenebre”

- Franco Battiato

Ringraziamenti

Voglio ringraziare in primis Rosaria, la mia supervisor, per avermi supportato, consigliato e confortato, la sua disponibilità e pazienza sono state fondamentali. La mia coautrice Silvia Pacei ed i reviewer, Enrico Fabrizi e Nikos Tzavidis, per i suggerimenti e l'incoraggiamento. Aldo, coautore e grande amico, per gli insegnamenti e le discussioni sulla statistica e la restante ciurma di viale Filopanti: Beatrice, Rosamarie e Riccardo, per le pause pranzo e le risate.

Voglio ringraziare il mio terapista, Diego, senza il quale questa tesi non sarebbe mai stata scritta, cogliendo l'occasione per rimarcare come il problema della salute mentale dei dottorandi sia reale e necessiti di un cambiamento sistemico del mondo della ricerca.

I miei compagni di ciclo, specialmente Anna e Jacopo S., per le sofferenze e le gioie condivise. Cristina e Marco per l'avventura in via Manzoni e tutti i suoi avventori, è stato bellissimo vivere con voi. I miei amici, in particolare Davide L., Roberta e Beatrice, per essermi stati vicino. Jonathan Zenti per i suoi podcast.

Contents

List of Figures

List of Tables

Introduction

Overview

Our personal definition of inequality can teach us a lot about how we think and how we intend our societies. Due to its relational dimension, a discussion about inequality reflects a discussion about society's structure on its whole. Recently, social transformations have caused an increasing interest on economic inequality. Such interest stems not only from economic issues but also from its connection with social cohesion and issues affecting the quality of life.

In this regard, the demand for living conditions data, referring to local areas or specific subpopulations, is growing. Policy makers and stakeholders need reliable estimates at local level, in order to formulate and implement policies, to distribute resources and to measure the effect of policy actions. In addition, such estimates may be valuable to further deepen regional and inequality trends, as to identify which regions constitute the driver of national inequality and to study spatial spillovers (Cavanaugh and Breau, 2018). Income data are usually collected via household sample surveys which are planned for aggregates estimation at macro level. Thus, local domains fall outside the prior design plan, resulting in small-sized samples and yielding unreliable direct estimation. The problem could be overcome with an ex-ante adjustment of sample sizes but it is, in many practical situations, excluded by cost–benefit analysis.

A key solution to cope with it is to rely on Small Area Estimation (SAE) methods. Research in this field is accelerating in recent years, with an overwhelming diversity in new investigated problems and innovative proposed solutions (Pfeffermann, 2013). Concerning SAE of income and living conditions indicators, the poverty mapping has gained special consideration, while inequality is treated jointly as a minor appendix (Pratesi, 2016; Molina and Rao, 2010). This remarks upon a lack, given that poverty indicators provide an insight into the lower tail of the income distribution, while inequality indicators comprise by definition a full distributional evaluation. From a statistical point

of view, indeed, an inequality measure is nothing more than a measure of dispersion or heterogeneity.

In this spirit, this dissertation aims at bringing together the world of small area estimation methods and the one of inequality measurement, and makes them start a conversation. The model-based class of SAE techniques leverages hierarchical models, both at area level or unit (individual) level. Our approach lies within the framework of Bayesian inference of area-level models, being less demanding with respect to data requirements and computational issues, as well as enabling the incorporation of design-based properties.

Main contributions of the thesis

This dissertation is organized into three main parts, structured in Chapters 1 to 3. The first two chapters focus on the SAE of inequality measures. This involves two main challenges, as inequality direct estimators in small areas are known to be (a) biased and (b) unreliable due to the high variability. Bias correction and variability reduction are faced in Chapter 1 and Chapter 2, respectively. Chapter 3, on the other hand, provides a more broad computational proposal, by illustrating the `tipsae` R package. This proposal has been developed in the context of the `deep` project[1], with the final aim of implementing innovative SAE methods for mapping unit interval-defined indices and proportions. This package implements the methodological contribution provided in Chapter 2, but it goes far beyond by supplying a more general tool for handling indices and proportions using Beta-based small area models.

Chapter 1 includes a first contribution with a twofold aim. On one hand, it provides a comprehensive discussion about the behavior of inequality estimators from complex surveys in small samples, including issues generally addressed in a piecemeal manner, such as bias evaluation, robustness, sampling variance estimation and distributional analysis. On the other hand, it proposes a methodological framework for bias-correction in a finite population setting, more specifically taking into account the complex survey design. Previous proposal in this respect, only consider *iid* observations framework (Deltas, 2003; Schluter and van Garderen, 2009; Davidson, 2009; Van Ourti and Clarke, 2011). The bias-correction proposal includes a large class of inequality measures comprising the Gini Index, the Generalized Entropy and Atkinson families. Since the methodology is based on Taylor's expansions and generalized linearization method, not requiring any parametric assumption on income distribution, it comes out to be very flexible. Design-based

[1] https://povertyevidence.org/

simulation has been carried out showing a noticeable bias reduction for all measures. A bootstrap variance estimation proposal and a distributional analysis follow. Results about estimators distributions show increasing positive skewness and lepto-kurtosis at decreasing sample sizes, confirming the non-applicability of classical asymptotic results in small samples and suggesting the development of alternative methods of inference.

Chapter 2 constitutes as the core of the dissertation, setting out a small area estimation strategy for four inequality measures: Gini, Theil and two indices pertaining to the Atkinson family. All of the them have a double-bounded support defined on the unit interval. Two different bodies of literature unfold in this respect, revolving around linear mixed models with suitable transformations (Rao and Molina, 2015) and Beta regression models (Janicki, 2020). However, classical Gaussian or Beta regression options fails in case of skewed and heavy-tailed estimators (Ferrante and Pacei, 2017; Migliorati *et al.*, 2018) such as the ones of interest. Thus, our contribution extends SAE literature in case of unit interval-defined, skewed and heavy-tailed parameters by adopting a Beta-mixture approach. In addition, the methodological proposal deepen the analysis on inequality estimators by deriving their approximate variance functions. Our model comes out to outperform the Beta one, both in terms of bias and error of model-based estimators, avoiding to highly underestimate inequality and providing reliable estimates.

Chapter 3 exposes the `tipsae` R package. Among the plethora of existing packages for small area modelling, only `emdi` package (Kreutzmann *et al.*, 2019) directly accounts for unit interval responses at area-level by providing the arc-sin transformation in a Gaussian setting (Schmid *et al.*, 2017). Our package aims at filling this gap, by implementing a full set of Beta-based small area models at area-level on unit interval-defined measures. The implementation is carried out via Bayesian Hierarchical models, including Beta and its mixture extensions such as Zero and/or One Inflated Beta and Flexible Beta models. Particular dependence structures can also be modelled, including spatial and/or temporal structured priors as well as shrinkage priors for random effects. Model estimation is carried out relying on `Stan` routine (Carpenter *et al.*, 2017). Specific small-area model diagnostics are produced by ad-hoc functions, facing the most relevant aspects to deepen in a Hierarchical Bayes model, equipped with visualization tools. Moreover, variance smoothing procedures and benchmarking procedures for model-based estimates are implemented as complementary tools. A shiny app interface comes on top of the implementation.

Note that Chapters 1 to 3 have been developed as three separate papers, in joint work with M.R. Ferrante and S. Pacei (Chapters 1 and 2) and A. Gardini (Chapter 3). Therefore, there could be some repetitions among chapters and the notation, defined

differently in each chapter, could lead globally to some inconsistencies.

Chapter 1

Mind the Income Gap: Behavior of Inequality Estimators from Complex Survey Small Samples

1.1 Introduction

Income inequality measures are known to be biased in small samples (Deltas, 2003; Breunig and Hutchinson, 2008), leading usually to an underestimation. The bias depends on the variability related to the variable of interest and, for some specific measures, also on the skewness of its distribution (Breunig, 2001). In this regard, consider that income is well-known to be positively skewed.

Moreover, the magnitude of the bias varies depending on the inequality measure. This aspect deserves attention given that inequality measures values are used for comparisons across time and location. Neglecting the bias may bring out discrepancies in estimated quantities due to different sample sizes or different underlying distributions rather than being a true inequality gap (Breunig and Hutchinson, 2008).

Consider that the problem of observations scarcity may arise when dealing with inequality in specific sub-populations, such as age-sex-race groups, as well as in case of inequality mapping at great geographical levels of disaggregation. The small area estimation, therefore, could be a main field of application. In addition, the interest for reliable inequality estimates is growing due to the observed increment in gap and social exclusion among regions and to their potential contribution in planning policies and foster regional studies (Cavanaugh and Breau, 2018).

Concerning the Gini index, a large body of literature faces the bias issues, such as Jasso (1979), Lerman and Yitzhaki (1989), Deltas (2003), Davidson (2009), Van Ourti

and Clarke (2011) in *iid* samples and Fabrizi and Trivisano (2016) for the complex survey case. However, concerning alternative measures such as Atkinson Indexes and the Generalized Entropy (GE) measures, the literature on bias is very scarce, even in the *iid* case. Some contributions are provided by Giles (2005) and Schluter and van Garderen (2009) for the GE family and Breunig and Hutchinson (2008) for GE and Atkinson families. The mentioned references adopt different methodological approaches to correct or reduce bias in an *iid* context.

Income data are usually collected via specific household surveys, with a complex sampling design. Generally, their designs involve stratification and selection of sampling units in more than one stage. Thus, the survey sample selection process, together with ex-post treatment procedures such as calibration and imputation, invariably introduces a complex correlation structure in the data, which has to be taken into account. This makes the development of a theoretically valid bias correction challenging.

Furthermore, the bias issue is even exacerbated in income data applications, which are traditionally affected by the problem of extreme values (Van Kerm, 2007), since inequality measures are highly unrobust to outliers (Cowell and Victoria-Feser, 1996). It has been widely tested that those measures appear to be unrobust even to an infinitesimal amount of data contamination, especially when dealing with extreme values on the tails. This aspect depends clearly on the type of measure we are dealing with and it becomes even more cumbersome to handle in case of small samples.

The aim of this chapter is twofold. On one hand, we provide a comprehensive discussion about the behavior of inequality estimators in small samples from complex surveys, including issues generally addressed in a piecemeal manner such as bias evaluation, robustness, variance estimation and distributional analysis. This analysis could be valuable for future developments of small samples inference on such measures. On the other hand, the aim is to propose a methodological framework for bias correction in a finite population setting, more specifically taking into account the complex survey designs. The set of considered measures embraces Gini Index, Atkinson Indexes and Generalized Entropy measures, together with the Coefficient of Variation. The methodology is based on Taylor's expansions and generalized linearization method (Deville, 1999; Demnati and Rao, 2004), relying on the concept of influence functions. Any parametric assumption on income distribution is not required, providing a very flexible framework.

Our bias correction proposal is evaluated via simulations showing a noticeable bias reduction for all the measures and leading, in some cases, to approximate unbiased estimators. An in-depth analysis of measures sensitivities confirms the great impact outliers have on the magnitude of estimators bias and variance.

A preliminary definition of the considered inequality measure can be found in Section 1.2. The bias correction strategy is set out in Section 1.3 and the bias-correction estimation steps are detailed in Section 1.4. A design-based simulation study involving the European Survey of Income and Living Condition (EU-SILC) income data (Guio, 2005) is provided in Section 1.5, in order to evaluate the magnitude of the bias and the efficacy of our correction. A design-aware bootstrap for bias-corrected estimators variance is proposed in Section 1.6, while a distributional analysis involving an additional model-based simulation follows in Section 1.7. Conclusions are drawn in Section 1.8.

1.2 Inequality Measures

The most famous inequality measure is, indeed, the Gini concentration index, employed in social sciences for measuring concentration in the distribution of a positive random variable. There are several equivalent definitions of Gini index (Ceriani and Verme, 2015), we will use the formulation of Sen (1997). Suppose we have a finite population $\mathcal{U}$ of $N(<\infty)$ elements labelled as $\{1,\dots,N\}$. Let z_i be a characteristic of interest, in our case income, for the i-th unit of the finite population, where $z_i \in \mathbb{R}^+$, $\forall i = 1,\dots,N$, and a sample s_{iid} of size n_{iid} is picked through simple random sampling. The Gini estimator is defined as

$$G = \frac{2}{n_{iid}^2 \hat{\mu}} \sum_{i \in s_{iid}} n_i z_i - \frac{n_{iid}+1}{n_{iid}},$$

with n_i denoting the rank of i-th unit and $\hat{\mu}$ the sample mean.

However, the estimation of alternative measures, in addition to the Gini index, may enable a more meaningful assessment of different aspects of economic inequality. Gini index does not allow to decompose inequality into within groups and between groups components, moreover, it is positional (weakly) transfer sensitive, namely index variations depend on the ranks of the donor and recipients. Lastly, it constitutes a stochastic dominance measure, based on a partial ordering of probability distributions: two very different distributions - one having more inequality amongst the poor, the other amongst the rich can have the same index value. When the distributional dominance fails, welfare-based measures, such as Atkinson Indexes, may provide for a *complete* ranking among alternative distributions, at the expense of more stringent assumptions as to how to represent social welfare (Bellu and Liberati, 2006). Atkinson index has

support [0,1] and is defined as

$$A(\varepsilon) = \begin{cases} 1 - \frac{1}{\hat{\mu}} \left(\frac{1}{n_{iid}} \sum_{i \in s_{iid}} z_i^{1-\varepsilon} \right)^{1/(1-\varepsilon)} & \text{for } \varepsilon \neq 1 \\ 1 - \frac{1}{\hat{\mu}} \left(\prod_{i \in s_{iid}} z_i \right)^{1/n_{iid}} & \text{for } \varepsilon = 1. \end{cases}$$

The parameter ε expresses the level of inequality aversion, as ε increases, the index becomes more sensitive to changes at the lower end of the income distribution.

Besides, an additive decomposable family of inequality measure is the Generalized Entropy class. As opposed to the measures seen before, this class has the advantage to be strongly transfer-sensitive, meaning that it reacts to transfers depending on the donor and recipient income levels. It is based on the concept of entropy which applied to income distributions has the meaning of deviations from perfect equality:

$$GE(\alpha) = \begin{cases} \frac{1}{n_{iid}\alpha(\alpha-1)} \sum_{i \in s_{iid}} \left[\left(\frac{z_i}{\hat{\mu}} \right)^{\alpha} - 1 \right] & \alpha \neq 0, 1, \\ \frac{1}{n_{iid}} \sum_{i \in s_{iid}} \frac{z_i}{\hat{\mu}} \ln \frac{z_i}{\hat{\mu}} & \alpha \to 1, \\ -\frac{1}{n_{iid}} \sum_{i \in s_{iid}} \ln \frac{z_i}{\hat{\mu}} & \alpha \to 0. \end{cases}$$

Its parameter α sets the sensitivity of the index: a large α induces the index to be more sensitive to the upper tail, vice versa a small α to the lower tail. $GE(0)$ is the Mean Log Deviation, while $GE(1)$ is the more famous Theil index. Atkinson and Generalized Entropy are two interrelated parametric families of measures, as a transformation of the Atkinson Index is a member of the GE class:

$$A(\varepsilon) = 1 - [\varepsilon(\varepsilon - 1) \cdot \mathrm{GE}(1 - \varepsilon) + 1]^{1/(1-\varepsilon)}.$$

In this chapter, we consider the estimation of both classes separately, since common parameter values used in one family does not correspond deterministically to common parameter values used in the other one. Lastly, we consider as inequality measure the Coefficient of Variation, which is linked with a member of the GE family, namely GE(2) $=\mathrm{CV}^2/2$. Its square has been used in some income distribution analyses, including OECD (2011), but comparisons developed using this measure seems to be very sensitive to top outliers (Atkinson, 2015).

1.3 Bias Correction Proposal

The bias of inequality estimators in small samples can be due to the structure of inequality measures as a non-linear function of estimators. The bias can be either positive or negative, depending on the characteristics of the reference variable distribution, except for the Mean Log Deviation which has structurally negative bias; this aspect is made clearer in the following. Among the measures with non-predictable bias direction, Breunig (2001) shows that CV and $GE(2)$ bias direction is negatively related to income skewness, concerning the other measures it depends on the income dispersion. This aspect could be analyzed in-depth by imposing a distributional assumption on the income variable, but this is out of scope.

Proposition 1.1. *As stated by Breunig and Hutchinson (2008), for a subset of the considered measures, i.e. the ones belonging to the GE and Atkinson families, the relationship between the expectation of the sample measure of inequality $\hat{\theta}$ and its true population value θ is:*

$$\mathbb{E}[\hat{\theta}] = \theta + O\left(\frac{1}{n_{iid}}\right),$$

with n_{iid} denoting the sample size in the iid case.

Proof. Let us consider a sample with *iid* elements $\{z_1, \ldots, z_{n_{iid}}\}$, drawn from a population via simple random sampling, where z_i is the variable for the i-th unit with expected value μ and variance σ^2. Let us consider also $\{g(z_1), \ldots, g(z_{n_{iid}})\}$ with $g(z)$ a generic monotone transformation of the income variable, induced by $g(\cdot) : \mathbb{R}^+ \to \mathbb{R}$, that changes for each measure, having expected value γ and variance ϕ^2. Considering equation (1.2) with $\hat{\mu} = \sum_{i=1}^{n_{iid}} z_i/n_{iid}$ and $\hat{\gamma} = \sum_{i=1}^{n_{iid}} g(z_i)/n_{iid}$, we can easily obtain estimator moments as $\hat{\mu} \sim [\mu, \sigma^2/n_{iid}]$ and $\hat{\gamma} \sim [\gamma, \phi^2/n_{iid}]$. Let us consider moreover that

$$Cov[\hat{\mu}, \hat{\gamma}] = \mathbb{E}[\hat{\mu}\hat{\gamma}] - \mu\gamma = \frac{1}{n_{iid}}(\mathbb{E}[z \cdot g(z)] - \mu\gamma) = \frac{Cov[z, g(z)]}{n_{iid}}.$$

Let us define the population value of a generic inequality measure θ as $f(\mu, \gamma)$, with $f(\cdot)$ a generic twice-differentiable function. By expanding the inequality measure estimator $\hat{\theta}$ as $f(\hat{\mu}, \hat{\gamma})$, via Taylor's expansion around the population values and considering its

expected value:

$$\begin{aligned}\mathbb{E}[\hat{\theta}] &= \theta + \frac{1}{2} f_{\gamma,\gamma}(\gamma,\mu)\mathbb{V}[\hat{\gamma}] + f_{\gamma,\mu}(\gamma,\mu)Cov[\hat{\gamma},\hat{\mu}] + \frac{1}{2} f_{\mu,\mu}(\gamma,\mu)\mathbb{V}[\hat{\mu}] + O(n_{iid}^{-2}) \\ &= \theta + O(n_{iid}^{-1}) + O(n_{iid}^{-1}) + O(n_{iid}^{-1}) + O(n_{iid}^{-2}) \\ &= \theta + O(n_{iid}^{-1}), \end{aligned} \tag{1.1}$$

where $f_\gamma = \frac{\partial f(\gamma,\mu)}{\partial \gamma}$ and $f_{\gamma\mu} = \frac{\partial^2 f(\gamma,\mu)}{\partial \gamma \partial \mu}$.

□

Our bias-correction proposal constitutes a generalization of the framework of Breunig and Hutchinson (2008), developed for *iid* observations, to the finite population and full design-based setting. At the same time, we extend the proposal to a wider set of measures comprising Gini Index. We provide a closed-form bias correction for complex designs which allows us to avoid the use of resampling techniques and can be applied in a distribution-free setting at once. This generalization has been developed considering Horvitz-Thompson type estimators and Ultimate Clusters and Influence Function linearization techniques for variances and covariances estimation.

We are interested in a variety of non-linear functions of income values as inequality measures are. Let denote with s a sample of size n, drawn using a complex sampling design, and with $p(s)$ the probability of selecting the particular sample $s \subset \mathcal{U}$ out of the set of all possible samples $\mathcal{Q}$, thus $p(s) \geq 0$ and $\sum_{s\in\mathcal{Q}} p(s) = 1$. The inclusion probability of unit k is denoted with π_k, being $\pi_k = \sum_{s\in\mathcal{Q}_k} p(s)$ with $\mathcal{Q}_k$ the set of all possible samples including unit k.

We consider the generic inequality measure written as a function of the mean μ and $\gamma = \mathbb{E}[g(z)]$. The population value for the generic inequality measure is

$$\theta = f(\mu, \gamma), \tag{1.2}$$

with $f(\cdot)$ a twice-differentiable function. The related estimator in our complex survey framework is $\hat{\theta} = f(\hat{\mu}, \hat{\gamma})$ in which Horvitz-Thompson estimators of the mean and γ are plugged in, i.e.

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i z_i}{N} \quad \text{and} \quad \hat{\gamma} = \frac{\sum_{i=1}^n w_i g(z_i, w_i)}{N}. \tag{1.3}$$

where $w_i = 1/\pi_i$ or a treated and calibrated version of it and N is the population size. Note that results in this section hold also when employing Hajek type estimator, i.e. with denominator $\hat{N} = \sum_{i=1}^n w_i$, since it is approximately unbiased (Särndal *et al.*,

2003, pg. 182). Kakwani (1990) uses a similar approach to express inequality indices to derive their asymptotic standard error. By simply applying a second order Taylor's series expansion of the sample estimator around the population values and evaluating its expected value, the bias can be expressed as

$$\begin{aligned}\mathbb{E}[\hat{\theta}-\theta] =& f_\gamma(\gamma,\mu)\,\mathbb{E}[\hat{\gamma}-\gamma] + \frac{1}{2}f_{\gamma,\gamma}(\gamma,\mu)(\mathbb{V}[\hat{\gamma}]+\mathbb{E}^2[\hat{\gamma}-\gamma])+ \\ &+ f_{\gamma,\mu}(\gamma,\mu)(Cov[\hat{\gamma},\hat{\mu}]-\mu\,\mathbb{E}[\hat{\gamma}-\gamma]) + \frac{1}{2}f_{\mu,\mu}(\gamma,\mu)\mathbb{V}[\hat{\mu}] + O(n^{-2}). \qquad (1.4)\end{aligned}$$

Notice that $\hat{\mu}$ is unbiased. An alternative method for bias approximation could be the small-σ approximation described in Ullah (2004). However, this framework requires high order moments and cross-moments estimation when facing non-*iid* assumptions. This may result quite challenging in cases of multi-stage surveys and distribution-free settings.

1.3.1 Approximate Bias Corrections

In this subsection, we detail the design-based estimators for each inequality measure and we provide their explicit bias formulation based on equation (1.4), defining all the relevant quantities in Table 1.1. Let denote with n the sample size and with $\sqrt{n/(n-1)}$ the standard bias-correction adjustment for the weighted variance; $F(\cdot)$ denotes the cumulative distribution function of the variable of interest and lastly consider $\hat{N}_i = \sum_{k\in s} w_k \mathbb{1}(n_k \leq n_i)$. The notation $\mathbb{1}(A)$ defines an indicator function, assuming value 1 if A is observed and 0 otherwise.

As regards the Gini index, we employ the alternative formulation defined by Sen (1997) and the complex survey estimator proposed by Langel and Tillé (2013). By considering γ and $\hat{\gamma}$ defined as in Table 1.1 for the Gini index, its approximate bias in small sample is

$$\begin{aligned}\mathbb{E}[\hat{G}-G] &\cong \frac{2}{\mu}\,\mathbb{E}[\hat{\gamma}-\gamma] + \frac{2\gamma}{\mu^3}\mathbb{V}[\hat{\mu}] - \frac{2}{\mu^2}(Cov[\hat{\mu},\hat{\gamma}]-\mu\,\mathbb{E}[\hat{\gamma}-\gamma]) \qquad (1.5)\\ &= \frac{4}{\mu}\,\mathbb{E}[\hat{\gamma}-\gamma] + \frac{2\gamma}{\mu^3}\mathbb{V}[\hat{\mu}] - \frac{2}{\mu^2}Cov[\hat{\mu},\hat{\gamma}],\end{aligned}$$

The derivation of the approximate bias related to the weighted estimator $\hat{\gamma}$ is not trivial. As explained by Langel and Tillé (2013), its numerator is not composed of two simple sums. Indeed the quantity $\hat{N}_k$, an estimator of the rank of unit k, is random since its value depends on the selected sample. A heuristic solution is to consider the approximate bias of the corresponding *iid* estimator, $\mathbb{E}[\hat{\gamma}-\gamma] = -1/n_{iid}(\gamma-\mu/2)$ as

derived by Davidson (2009) so that:

$$\mathbb{E}[\hat{G} - G] = \frac{-2G}{n_{iid}} + \frac{2\gamma}{\mu^3}\mathbb{V}(\hat{\mu}) - \frac{2}{\mu^2}Cov(\hat{\mu}, \hat{\gamma}). \tag{1.6}$$

This correction is in line with Davidson (2009) and Fabrizi and Trivisano (2016) proposals, whereas these are based on a first-order Taylor's expansions and thus limited to the first term of the right-hand side equation (1.5), ours extends it to a second-order expansion. This translates into the fact that, while Jasso (1979), Deltas (2003) and Davidson (2009) proposals identify the adjusted Gini in *iid* contexts as $n_{iid}(n_{iid}-1)^{-1}\hat{G}$, our correction reconsiders the shape of the adjusted estimator with a further order of approximation as

$$\frac{n_{iid}}{n_{iid}-2}(\hat{G} - a), \tag{1.7}$$

with a equals to the sum of the second and third term of (1.6).

The complex survey estimators of Atkinson and Generalized Entropy measures come from Biewen and Jenkins (2006). The bias estimation of Mean Log Deviations is consistent with Ferrante and Pacei (2019) and note that it is structurally negative. The approximate bias expressions for complex survey estimators coincide, in some specific cases, with the ones for *iid* estimators made explicit by Breunig and Hutchinson (2008) due to the invariance properties related to expected values of the sum of dependent variables. See formulas referring to Mean Log Deviation, Theil and Atkisons indexes in Table 1.1.

1.4 Bias Estimation

In this section, we detail the estimation of the approximate bias defined in Table 1.1. Such estimation is not trivial considering that the mentioned expressions depend on variances and covariances involving a non-linear statistic $\hat{\gamma}$. Thus, a linearization of $\hat{\gamma}$ is needed in order to make it tractable and carry on variance estimation. The linearization technique is described in Subsection 1.4.1, whereas $\mathbb{V}[\hat{\mu}]$, $\mathbb{V}[\hat{\gamma}]$ and $Cov[\hat{\mu}, \hat{\gamma}]$ have been estimated following Subsection 1.4.2. The estimation procedure is completed by replacing μ and γ with $\hat{\mu}$ and $\hat{\gamma}$.

Measure	$\boldsymbol{\gamma = \mathbb{E}[g(y)]}$	**Design-based Est.**	$\hat{\boldsymbol{\gamma}}$	$\boldsymbol{f(\hat{\mu}, \hat{\gamma})}$	**Approx. Bias**
Gini	$\mathbb{E}[z \cdot F(z)]$	$\frac{2\sum_{i\in s} w_i z_i (\hat{N}_i - w_i/2)}{N^2 \hat{\mu}} - 1$	$\frac{\sum_{i\in s} w_i z_i (\hat{N}_i - \frac{w_i}{2})}{N^2}$	$\frac{2\hat{\gamma}}{\hat{\mu}} - 1$	$\frac{4}{\mu}\mathbb{E}[\hat{\gamma} - \gamma] + \frac{2\gamma}{\mu^3}\mathbb{V}[\hat{\mu}] - \frac{2}{\mu^2}Cov[\hat{\mu}, \hat{\gamma}]$
GE(α) $\alpha \neq 0, 1$	$\mathbb{E}[z^\alpha]$	$\frac{1}{\alpha(\alpha-1)}\frac{n}{n-1}\left(\frac{\sum_{i\in s} w_i z_i^\alpha}{N\hat{\mu}^\alpha} - 1\right)$	$\frac{\sum_{i\in s} w_i z_i^\alpha}{N}$	$\frac{1}{\alpha(\alpha-1)}\frac{n}{n-1}\left(\frac{\hat{\gamma}}{\hat{\mu}^\alpha} - 1\right)$	$\frac{n}{n-1}\Bigg(-\frac{1}{(\alpha-1)\mu^{\alpha+1}}Cov[\hat{\gamma}, \hat{\mu}] + \frac{\alpha+1}{2(\alpha-1)}\frac{\gamma}{\mu^{\alpha+2}}\mathbb{V}[\hat{\mu}]\Bigg)$
GE(0)	$\mathbb{E}[\log z]$	$\frac{1}{N}\sum_{i\in s} w_i \log\frac{\hat{\mu}}{z_i}$	$\frac{\sum_{i\in s} w_i \log z_i}{N}$	$\log(\hat{\mu}) - \hat{\gamma}$	$-\frac{1}{2\mu^2}\mathbb{V}[\hat{\mu}]$
GE(1)	$\mathbb{E}[z(\log z)]$	$\frac{1}{N}\sum_{i\in s} w_i \frac{z_i}{\hat{\mu}} \log\frac{z_i}{\hat{\mu}}$	$\frac{\sum_{i\in s} w_i z_i \log z_i}{N}$	$\frac{\hat{\gamma}}{\hat{\mu}} - \log(\hat{\mu})$	$-\frac{Cov[\hat{\mu}, \hat{\gamma}]}{\mu^2} + \left(\frac{\gamma}{\mu^3} + \frac{1}{2\mu^2}\right)\mathbb{V}[\hat{\mu}]$
A(ε) $\varepsilon \neq 1$	$\mathbb{E}[z^{1-\varepsilon}]$	$1 - \frac{1}{\hat{\mu}}\left(\frac{1}{N}\sum_{i\in s} w_i (z_i)^{1-\varepsilon}\right)^{\frac{1}{1-\varepsilon}}$	$\frac{\sum_{i\in s} w_i z_i^{1-\varepsilon}}{N}$	$1 - \frac{\hat{\gamma}^{\frac{1}{1-\varepsilon}}}{\hat{\mu}}$	$-\frac{\varepsilon}{2(1-\varepsilon)^2}\frac{\gamma^{\frac{2\varepsilon-1}{1-\varepsilon}}}{\mu}\mathbb{V}[\hat{\gamma}] + \frac{1}{1-\varepsilon}\frac{\gamma^{\frac{\varepsilon}{1-\varepsilon}}}{\mu^2}\times$ $\times Cov[\hat{\gamma}, \hat{\mu}] - \frac{\gamma^{\frac{1}{1-\varepsilon}}}{\mu^3}\mathbb{V}[\hat{\mu}]$
A(1)	$\mathbb{E}[\log z]$	$1 - \frac{1}{\hat{\mu}}\prod_{i\in s} z_i^{w_i/N}$	$\frac{\sum_{i\in s} w_i \log z_i}{N}$	$1 - \frac{\exp\{\hat{\gamma}\}}{\hat{\mu}}$	$-\frac{\exp\{\gamma\}}{\mu}\left(\frac{\mathbb{V}[\hat{\gamma}]}{2} - \frac{Cov[\hat{\gamma}, \hat{\mu}]}{\mu} + \frac{\mathbb{V}[\hat{\mu}]}{\mu^2}\right)$
CV	$\mathbb{E}[z^2]$	$\sqrt{\frac{n}{n-1}}\left(\frac{1}{N}\sum_{i\in s} w_i \left(\frac{z_i}{\hat{\mu}}\right)^2 - 1\right)^{1/2}$	$\frac{\sum_{i\in s} w_i z_i^2}{N}$	$\sqrt{\frac{n}{n-1}}\frac{\sqrt{\hat{\gamma} - \hat{\mu}^2}}{\hat{\mu}}$	$\sqrt{\frac{n}{n-1}}\Bigg[-\frac{\mathbb{V}[\hat{\gamma}]}{8\mu(\gamma-\mu^2)^{\frac{3}{2}}} - \frac{1}{2}\left(\frac{1}{\mu^2(\gamma-\mu^2)^{\frac{1}{2}}} - \frac{1}{(\gamma-\mu^2)^{\frac{3}{2}}}\right)\times$ $\times Cov[\hat{\gamma}, \hat{\mu}] + \frac{1}{2}\Bigg(2\frac{(\gamma-\mu^2)^{\frac{1}{2}}}{\mu^3} + \frac{1}{\mu(\gamma-\mu^2)^{\frac{1}{2}}} - \frac{\mu}{(\gamma-\mu^2)^{\frac{3}{2}}}\Bigg)\mathbb{V}[\hat{\mu}]\Bigg]$

TABLE 1.1: Relevant quantities for each measures including the approximate bias.

1.4.1 Linearization of Non-Linear Estimators

The linearization technique follows the intuition of approximating a non-linear statistic with a linear function of the transformed observations. In doing so, the linear approximation can be used to measure the precision and uncertainty associated with the statistic of interest, using well-known linear estimators of variances and covariances. We apply the generalized linearization method (Deville, 1999; Demnati and Rao, 2004; Osier, 2009). This method allows encompassing more non-linear statistics than the Taylor one being, in general, more flexible and working better in case of small samples (Osier, 2009).

In particular, this procedure as stated by Antal *et al.* (2011), reconciles the two approaches introduced by Deville (1999) and Demnati and Rao (2004), both relying on the concept of influence function (Hampel, 1974). The same method has been used directly by Graf and Tillé (2014) to estimate the variance of some inequality measures estimators via linearization. Following the theoretical framework of Antal *et al.* (2011), we could say that a population parameter of interest θ in a population $\mathcal{U}$ undergoes the influence of a unit k, which depends on an infinitesimal variation in the importance assigned to the unit. Let us express the parameter as a functional $\theta = T(M)$ based on a measure $M(\cdot)$ such that

$$\begin{cases} M(z) = \tilde{w}_k \quad z = z_k \quad \forall k \in s \\ M(z) = 0 \quad \text{otherwise,} \end{cases}$$

with $\tilde{w}_k$ a generic weight associated to the unit k, in our case corresponding to $\tilde{w}_k = w_k/N$. The general measure M turns to a discrete measure, leading T into a discrete functional.

Since the functional is expressed as an explicit function of the weights assigned to each unit, the linearized variable is merely a function of the partial derivatives with respect to the weights:

$$I[T(M)]_k = v_k = \frac{\partial T(M)}{\partial \tilde{w}_k} = \frac{\partial}{\partial \tilde{w}_k} \sum_{i \in s} \tilde{w}_i g(z_i) = g(z_k). \tag{1.8}$$

The linearized variables for each inequality measure can be directly derived from (1.8) by substituting the measure-specific function $g(\cdot)$ listed in Table 1.1.

1.4.2 Estimation of Design Variances and Covariances

As regards the estimation of the design variances and covariances of linear and linearized estimators, we consider a complex survey design involving stratification and multi-stage selection, with both Self-Representing (SR) -included at the first stage with probability one - and Non Self-Representing (NSR) strata. This design is consistent with the most common income survey designs and, in general, with household surveys, whereas firm surveys are based on a single-stage stratified sample design, also a special case of the multi-stage selection.

First of all, we define the shape of the Horvitz-Thompson variance estimator in case of linear (or linearized) estimators, $\hat{\mu} = \sum_{i \in s} w_i z_i / N$ when $w_i = 1/\pi_i$, as:

$$\mathbb{V}[\hat{\mu}] = \sum_{i \in s} \frac{z_i^2}{\pi_i^2 N^2}(1 - \pi_i) + 2 \sum_{i \in s} \sum_{k \in s, i \neq k} \frac{z_i}{\pi_i} \frac{z_k}{\pi_k} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik} N^2},$$

with π_{ik}, $\forall i, k \in \mathcal{U}, i \neq k$ denoting the second-order inclusion probabilities i.e. the probability that the sample includes both i-th and k-th units. However generally (a) $w_i \neq 1/\pi_i$ and (b) π_{ik}, $\forall i, k \in \mathcal{U}, i \neq k$ are difficult to calculate under complex sampling designs.

Therefore, the variance estimator to be considered constitutes an approximation relying on simplified assumptions. Firstly, we assume that Primary Sampling Units (PSU) are sampled with replacement, and secondly we reduce multi-stage sampling into a single-stage process by relying on the Ultimate Clusters technique (Kalton, 1979). Moreover, we take into account the hybrid nature of the probability scheme, blending a variance estimator for stratified design associated with the SR strata, including a finite population correction factor, and a typical Ultimate Cluster variance estimator for multi-stage schemes associated with the NSR strata. The latter one is widely used in official statistics, see Osier *et al.* (2013) for Eurostat procedures. Therefore, considering without loss of generality a two-stage scheme, let $\hat{\mu} = \sum_h \sum_d \sum_i \tilde{w}_{hdi} z_{hdi}$ with h stratum indicator, d Primary Sampling Unit (PSU) indicator and i Secondary Sampling Unit indicator (SSU), be a linear estimator for μ, its standard error estimate is as follows:

$$\begin{aligned}\hat{\mathbb{V}}[\hat{\mu}] &= \sum_{h=1}^{H_{SR}} \mathbb{V}[\hat{\mu}_h] + \sum_{h=1}^{H_{NSR}} \mathbb{V}[\hat{\mu}_h] \\ &= \sum_{h=1}^{H_{SR}} M_h^2(1-f_h)\frac{s_h^2}{m_h} + \sum_{h=1}^{H_{NSR}} n_h s^2_{\hat{\mu}_h} \\ &= \sum_{h=1}^{H_{SR}} M_h \frac{M_h - m_h}{m_h(m_h-1)} \sum_{i=1}^{m_h}(z_{hi} - \bar{z}_h)^2 + \sum_{h=1}^{H_{NSR}} \frac{n_h}{(n_h-1)} \sum_{d=1}^{n_h}(\hat{\mu}_{hd} - \bar{\mu}_h)^2,\end{aligned} \tag{1.9}$$

with H_{SR} self-representative and H_{NSR} non self-representative strata, M_h the number of resident households in strata h, m_h the number of sample households in strata h, $f_h = m_h/M_h$ a finite population correction factor, n_h the number of PSUs in strata h. Consider, moreover, that $\bar{z}_h = \sum_{i=1}^{m_h} z_{hi}/m_h$, $\hat{\mu}_{hd} = \sum_{i=1}^{m_d} \tilde{w}_{hdi} z_{hdi}$ with i denoting the household label and m_d the number of sample households in PSU d, lastly $\bar{\mu}_h = \sum_{d=1}^{n_h} \hat{\mu}_{hd}/n_h$, with n_h being the number of PSU in stratum h. If however $n_h = 1$ for some strata, the estimator (1.9) cannot be used. A solution is to collapse strata to create "pseudo-strata" so that each pseudo-stratum has at least two PSUs. Common practice is to collapse a stratum with another one that is similar w.r.t. the target variables of the survey (Rust and Kalton, 1987).

Secondly, an estimator of $\mathbb{V}[\hat{\gamma}]$ can be obtained after the linearization of $\hat{\gamma}$, leading to $\mathbb{V}[\hat{\gamma}] \cong \mathbb{V}(\sum_{i\in s} w_i v_i/N)$, by adopting the same strategy used for $\mathbb{V}[\hat{\mu}]$ in (1.9).

Thirdly, as regards the estimation of the design covariance, let us consider that

$$Cov[\hat{\gamma}, \hat{\mu}] = \frac{1}{2}\Big(\mathbb{V}[\hat{\gamma} + \hat{\mu}] - \mathbb{V}[\hat{\gamma}] - \mathbb{V}[\hat{\mu}]\Big). \tag{1.10}$$

Thus, a possible estimator $\hat{Cov}[\hat{\gamma}, \hat{\mu}]$ would be simply obtained by plugging in the variance estimators previously mentioned, while $\mathbb{V}[\hat{\gamma} + \hat{\mu}]$ could be estimated by considering $\hat{\gamma} + \hat{\mu} = \sum_{i\in s} w_i(v_i + z_i)/N$ using (1.9).

Lastly as regards Gini index, the $\hat{\gamma}$ bias has already been adapted to the complex survey case by Fabrizi and Trivisano (2016), with a heuristic solution as follows

$$-\frac{\sum_{q\in s}(\sum_{i=1}^{r_q} w_{qi})^2}{N^2}\left(\hat{\gamma} - \frac{\hat{\mu}}{2}\right),$$

with $\sum_{i=1}^{r_q} w_{qi}$ the sum of the weights associated with the r_q individuals living in household q. However, in our case we opt for $-1/n(\hat{\gamma} - \hat{\mu}/2)$, since involving survey weights in n_{iid} estimation could dramatically induce further bias. Thus our bias estimator for

Gini index is

$$-\frac{2\hat{G}}{n} + \frac{2\hat{\gamma}}{\hat{\mu}^3}\hat{\mathbb{V}}[\hat{\mu}] - \frac{2}{\hat{\mu}^2}\hat{Cov}[\hat{\mu}, \hat{\gamma}].$$

1.5 Design-Based Simulation on Bias Correction

A first design-based simulation study has been carried out to evaluate our bias correction proposal. In this simulation, the cross-section Italian EU-SILC sample (2017 wave) has been assumed as synthetic population and the 21 NUTS-2 regions have been considered as target domains. The study is based on real income data, in order to check whether this specific framework can work with close-to-reality data, affected by peculiar problems (e.g. extreme values).

For comparison purposes, two simulation scenarios have been carried out. In the first one, the original income data are employed as synthetic population while, in the second one, an extreme values treatment is performed to circumvent non-robustness problems and the treated dataset is specified as an alternative synthetic population. Subsequently, we compare the results obtained after the treatment with the ones on original data to isolate the effect of outliers when evaluating bias-correction performances (Table 1.2).

The issue of robust estimation of economic indicators based on a semi-parametric Pareto upper tail model is well-established in literature see Brzezinski (2016) for a review and Alfons *et al.* (2013) for a specification suitable for survey data. On the contrary, the issue of robust treatment of outliers in the lower tail of income distribution appears less established (Van Kerm, 2007; Masseran *et al.*, 2019). Concerning the upper tail, we operated a semi-parametric Pareto-tail modelling procedure using the Probability Integral Transform Statistic Estimator (PITSE) proposed by Finkelstein *et al.* (2006), which blends very good performances in small samples and fast computational implementation, as suggested by Brzezinski (2016). As regards the lower tail, we used an inverse Pareto modification of the PITSE estimator, suggested by Masseran *et al.* (2019). In our simulations, the treatment has been done at a regional level to the original EU-SILC sample and the detection of outliers has been carried out following Mohd Safari *et al.* (2018) by using a Generalized Boxplot outlier detection procedure. We expect that, when outlying observations are representative, this procedure would highly bias the outcome and thus we do not recommend it.

From the assumed population, we repeatedly select 1,000 two-stage stratified samples, mimicking the sampling strategy adopted in the survey itself: in the first stage, SR strata are always included in the sample, while a stratified sample of PSU in NSR strata is

selected; in the second stage, a systematic sample of households is drawn from each PSU included in the first stage. We repeated the drawing for two different scenarios involving two different sampling rates, 1,5% and 3% respectively. Results are set out in Table 1.2 with the Average Relative Bias (ARB) and the Average Absolute Relative Error (AARE) in percentage, calculated for the 1,000 samples and the 21 regions defined as:

$$ARB = \frac{1}{21}\sum_{r=1}^{21}\frac{1}{1,000}\sum_{m=1}^{1,000}\left(\frac{\hat{\theta}_{m,r}}{\theta_r}-1\right),$$
$$AARE = \frac{1}{21}\sum_{r=1}^{21}\frac{1}{1,000}\left(\sum_{m=1}^{1,000}\left|\frac{\hat{\theta}_{m,r}}{\theta_r}-1\right|\right),$$

where θ_r is the population value for region r and $\hat{\theta}_{m,r}$ is its estimate for the generic iteration m. In our simulation setting the regional sample size ranges from 6 to 96 individuals (from 6 to 32 households) for the 1,5% sampling rate, and from 11 to 196 individuals (10 to 74 households) for the 3% sampling rate.

As clear from Table 1.2, the bias can be dramatically high for some measures estimated on non-smoothed data, due to the non-robustness properties to extreme values. It is the case of A($\varepsilon = 2$), extremely sensitive to low-income values (under 100 euro per year) which is -48% biased on average for the scenario with the smallest sample sizes. Also GE with $\alpha = 1, 2$ values are highly sensitive to high-income values being -18% and -23% biased. Moreover, the negative correlation between sample size and bias is less marked in case of non-treated data, since the bias depends more on whether a sample contains extreme values or not. On the contrary, the bias correction seems to not change in magnitude depending on the sample size and the presence of extreme values, showing good robustness properties.

Concerning extreme value treated data results, Figure 1.1 clearly illustrates the negative correlation between sample size and average relative bias in the 21 Italian regions for both the design-based estimator $\hat{\theta}$ and the bias corrected estimator $\hat{\theta}_{corr}$. The reduction of the bias provided by the correction is noticeable for all measures, leading to slightly biased estimates depending on the measure. Notice that the bias correction works well for measures not particularly sensitive to extreme observations such as Gini index, GE(0), Atk(0.5) and Atk(1). In case of CV and GE(2), the correction provides good results, but it seems, however, to not capture all the bias components. This confirms the results of Breunig (2001), suggesting that the coefficient of variation squared and GE(2) bias depends on the coefficient of skewness of the income distribution, not considered in our bias correction. Actually, a reliable estimation of that quantity, while

		CV	GE(0)	GE(1)	GE(2)	A(0.5)	A(1)	A(2)	G
without Extreme Value treatment									
1.5%									
	ARB	-18.2	-12.7	-17.5	-23.3	-15.3	-15.6	-48.0	-14.4
$\hat{\theta}$	AARE	30.0	52.9	46.4	53.5	45.9	47.4	56.8	25.6
	ARB	-12.1	-3.9	-8.7	-15.0	-6.3	-5.9	-41.6	0.4
$\hat{\theta}_{corr}$	AARE	29.3	54.4	48.0	55.7	47.5	49.4	54.6	35.5
3.0%									
	ARB	-12.7	-6.8	-10.5	-15.8	-8.7	-8.4	-38.1	-7.3
$\hat{\theta}$	AARE	24.5	39.4	36.0	46.2	34.3	35.6	49.0	17.7
	ARB	-7.8	-1.2	-3.9	-8.0	-2.5	-2.0	-32.4	0.2
	ARB ($n \geq 20$)	-8.3	-1.2	-3.6	-8.7	-2.3	-1.7	-30.0	-1.1
$\hat{\theta}_{corr}$	AARE	24.8	40.4	37.9	49.4	35.7	37.0	48.2	20.8
with Extreme Value treatment									
1.5%									
	ARB	-11.9	-13.1	-15.3	-17.0	-14.2	-14.3	-18.3	-14.2
$\hat{\theta}$	AARE	25.8	44.0	42.6	47.6	41.6	40.6	38.2	24.3
	ARB	-5.6	-3.5	-6.3	-8.7	-4.9	-4.7	-8.6	0.6
$\hat{\theta}_{corr}$	AARE	25.9	46.2	44.6	49.9	43.6	42.5	39.3	34.1
3.0%									
	ARB	-7.4	-6.4	-8.3	-10.3	-7.3	-7.1	-9.4	-7.1
$\hat{\theta}$	AARE	19.8	31.6	31.7	37.8	30.2	29.1	27.2	16.5
	ARB	-2.8	-0.6	-2.2	-3.5	-1.4	-1.1	-2.8	0.3
	ARB ($n \geq 20$)	-1.4	-0.3	-1.2	-1.5	-0.7	-0.5	-1.4	-0.9
$\hat{\theta}_{corr}$	AARE	20.4	33.0	33.3	40.2	31.7	30.4	28.4	19.5

TABLE 1.2: ARB and AARE in percentage for the 21 synthetic domains.

being straightforward in the *iid* case, appears cumbersome in case of weighted data being defined on a discrete grid of values. This leads to the non-applicability of the bias formula derived by Breunig (2001) in our case. Furthermore, the bias correction induces a slight but negligible error increase, except for the Gini index case that presents a relevant increase. This is due to the shape of the unbiased estimators, as described by (1.7), where a sum of estimators is multiplied by a factor $n/(n-2)$, which inherently inflates the variance by its square.

These results may constitute as valuable reference guideline when measuring inequality in small samples. When extreme values can be considered as a consequence of data

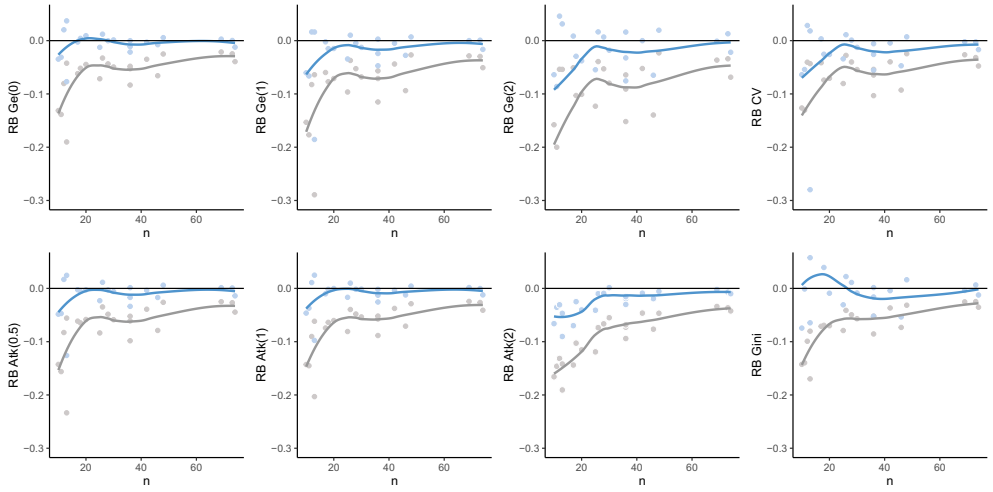

FIGURE 1.1: Relative bias of non-corrected measures (grey line), and of corrected measures (blue line) in 3% samples after extreme value treatment.

contamination, the joint application of an extreme value treatment and the bias correction may provide approximately unbiased estimates for a large class of measures. On the other hand, when extreme values constitute representative observations, it becomes necessary to restrict the attention to the most robust measures such as GE with $\alpha = 0$, Atkinson index with $\varepsilon = 1$ and Gini Index. Another important aspect to point out is that, in certain countries, the EU-SILC is based on registers that better capture top incomes, thus, a cross-country comparison of income inequality by effects on a tail-sensitive measure must be another reason for caution (Atkinson, 2015).

1.6 Bootstrap Variance of Bias-Corrected Estimators

Concerning inequality estimators, their variance estimation may be easily carried out via linearization as seen in Section 1.3. Linearized variables for each measure could be derived from (1.8) consistently with Langel and Tillé (2013) for Gini Index and Biewen and Jenkins (2006) for Generalized Entropy and Atkinson Indexes. On the other hand, the variance of bias corrected estimators adds a new level of complexity since the estimator formula is no longer the classical one. Indeed, it comprises a bias correction component that appears cumbersome to estimate via linearization since it is inherently a result of several linearizations.

Since our main aim is to provide a general "turn-key" solution for variance estimation, we approach the issue by taking into account strategies based on resampling

methods. Specifically, we opt for a proper design-aware bootstrap procedure as developed before by Fabrizi *et al.* (2011, 2020). A comprehensive review of bootstrap methods for survey data can be found in Lahiri (2003) and an interesting comparison between variance estimation techniques for poverty and inequality measures has been carried out by De Santis *et al.* (2020). Generally, the bootstrap algorithms for complex surveys re-sample primary units within strata, given the assumption that the number of strata is large: few primary units are sampled from each stratum so that the sampling fraction at the first stage is negligible (Rao *et al.*, 1999). This assumption is, however, not satisfied in small samples. An alternative solution may be to split up primary units into multiple parts to be re-sampled, to extremes into secondary units. Thus, our strategy selects units with a stratified single-stage design with replacement from the population of households considering geographical macro-strata. After the drawing, a calibration procedure has been put in place for each bootstrap sample, to adjust weights to known gender and age classes totals similarly to the calibration procedure applied to the original sample. The algorithm is similar to the ones proposed by Fabrizi *et al.* (2011), which provides estimates close to the ones relying on linearization methods, in case of simpler parameters.

The variance estimates versus the sample sizes are displayed in Figures 1.2 and 1.3. Notice that the analysis discriminates between the behavior of a region having an upper extreme value, called "outlier region", see Figure 1.2, and the others. Specifically, bootstrap CV estimates are provided for the Italian NUTS-2 regions versus the sample sizes in Figure 1.3. It is quite interesting to notice that a small-sized extreme-valued-affected sample has always high variances estimates across all measures. However, the sensibility to small sizes changes depends on the measure analyzed. The more GE parameters α increases i.e. it becomes more sensible to upper tail values, the more the estimator variance is affected only by the presence of upper extreme values rather than by the sample size. On the other hand, the more Atkinson ε increases becoming more sensible to lower tail, the more its variances becomes negatively correlated to sample sizes.

1.7 A Distributional Analysis

Literature on the probability distribution of inequality measures in small samples is scarce. Kakwani (1990) and Thistle (1990) have studied the asymptotic distribution of Generalized Entropy measures and Atkinson Index proving their asymptotic normality. As regards the Gini index, results on asymptotic normality of the different estimators are

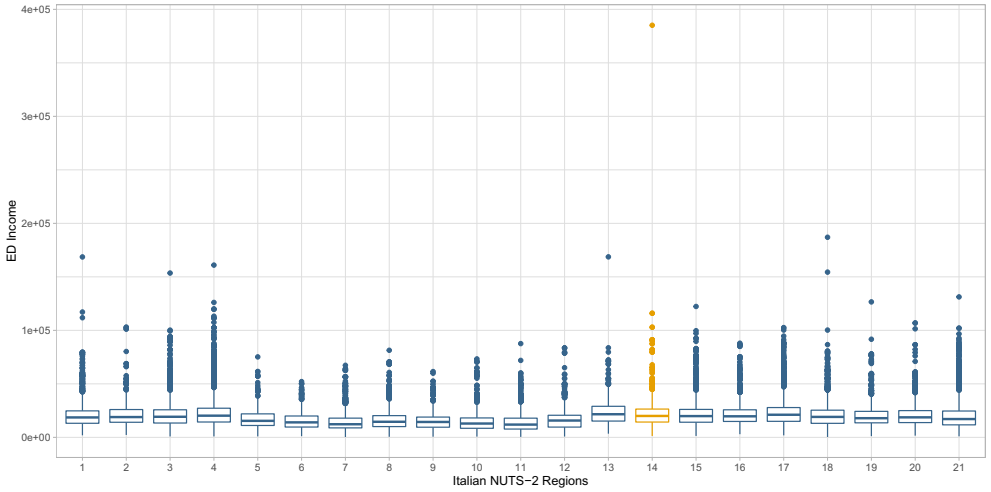

FIGURE 1.2: Income distribution for each region, outlier region in orange.

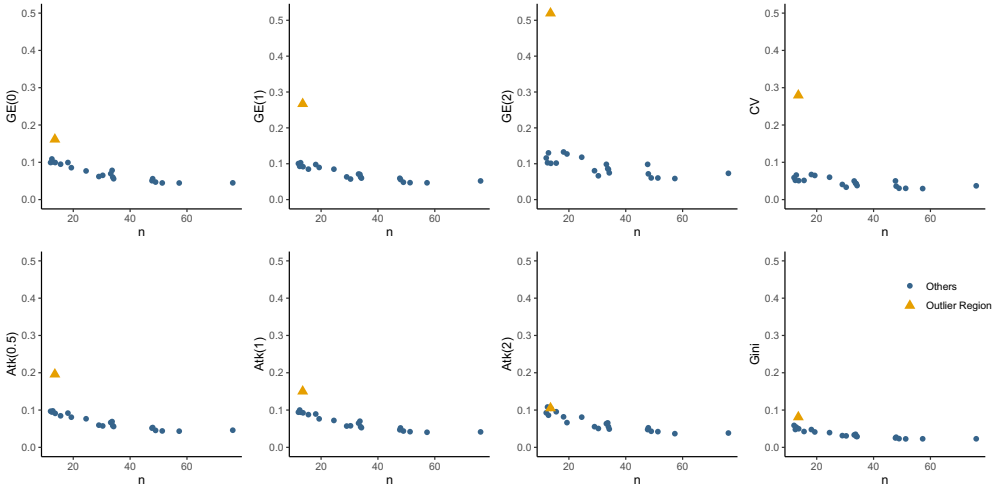

FIGURE 1.3: Bootstrap coefficient of variation for each inequality measures.

well-established (Giorgi and Gigliarano, 2017). Some interesting performance comparisons between asymptotic and bootstrap inference for inequality measures are provided by Biewen (2002) and Davidson and Flachaire (2007).

In this section, we provide a distributional analysis divided into two parts. The first one exploits empirical distributions from design-based simulation described in Section 1.5, to evaluate the skewness and heavy tails at varying sample sizes. In the second part, we performed a model-based simulation, using generated income data and unequal probability sampling in a more controlled setting. The main aim is to study the behavior of inequality measures at varying income distributional assumptions and, at the same time, to fit different parametrical densities on such distributions. We considered both mean modelling and median modelling based distributions.

		CV	GE(0)	GE(1)	GE(2)	A(0.5)	A(1)	A(2)	G
10%	$\hat{\eta}_3$	0.81	0.66	0.88	1.30	0.71	0.54	0.26	0.38
	$\hat{\eta}_4$	1.16	0.91	1.60	2.87	1.06	0.57	-0.01	0.34
5%	$\hat{\eta}_3$	1.00	0.94	1.19	1.80	0.96	0.74	0.40	0.54
	$\hat{\eta}_4$	1.88	1.73	2.81	6.09	1.81	1.01	0.08	0.56
3%	$\hat{\eta}_3$	1.00	1.07	1.27	1.90	1.04	0.83	0.49	0.58
	$\hat{\eta}_4$	1.66	1.94	2.87	6.13	1.87	1.03	0.03	0.60

TABLE 1.3: Coefficients of skewness and excess kurtosis for empirical distributions in design-based simulation.

1.7.1 Design-based Simulation

A brief analysis on the distribution of inequality measures is carried out, considering samples of increasing size, in order to evaluate how quickly their distribution tends to become symmetric. We consider regions as target domains and we keep the same simulation setting of Section 1.5 with different sampling rates, i.e. 10% (from 36 to 607 individuals, 28 to 254 households), 5% (from 16 to 337 individuals, 14 to 131 households) and 3% (from 11 to 196 individuals, 10 to 74 households). The coefficient of skewness η_3 and excess kurtosis η_4 empirical values are set out in Table 1.3 for different sampling rates. As clear from the results, the empirical distributions tend to become more positively skewed and leptokurtic at decreasing sample sizes. This is quite evident for the General Entropy measures, similarly but to a lesser extent for the other measures. The empirical distributions for each region are set out in Figure 1.4. The failing of asymptotic normality in small samples warns us to not consider parametric methods of inference based on such an assumption.

The empirical distributions from the bootstrap procedure, developed in Section 1.6, have been also considered and compared with the simulation ones, showing less asymmetry and lighter tails. This alerts us to not consider the bootstrap procedure reliable for moments of a higher order than 2 and quantiles, generally used in building confidence intervals.

1.7.2 Model-based Simulation

A model-based simulation has been also performed using the R package `simFrame` (Alfons *et al.*, 2010), to provide a more comprehensive study about the distribution of inequality estimators. The use of simulated data allows us to carry on the analysis in

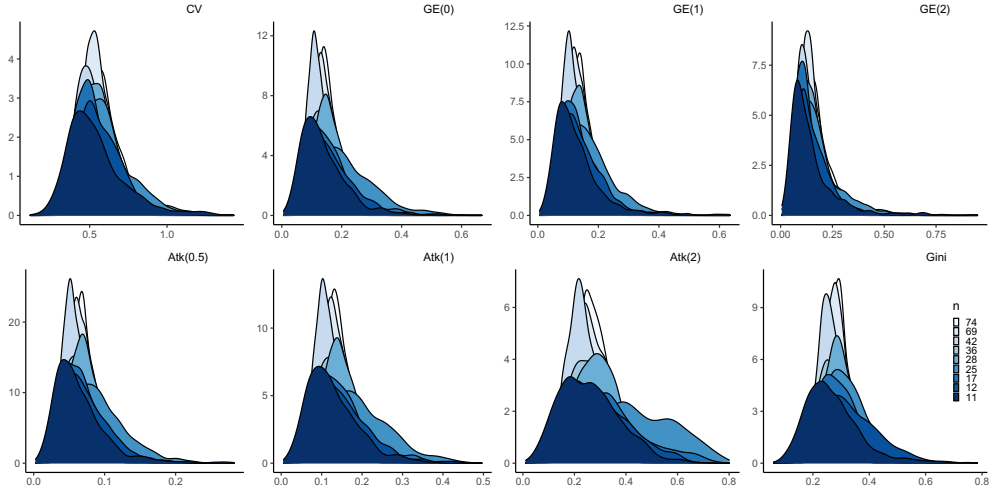

FIGURE 1.4: Empirical distributions by regions for the 3% samples from design-based simulation.

a more controlled setting, without potential confounding factors related to survey data collection dynamics.

We perform a preliminary fitting on EU-SILC income data via weighted pseudo-likelihood maximization of a plethora of suitable income distribution as Log-Normal, Dagum, Singh-Maddala and Generalized Beta of the Second Kind (Atkinson, 2015, pages 371-375), some goodness-of-fit results, such as AIC and BIC, are displayed in Table 1.4. The high flexibility of the GB2 due to its four parameters allows the best fitting in comparison to the other distributional assumptions. We decided therefore to use the best two fittings (GB2 and log-normal ones) to generate income data, in order to capture estimators behavior at varying distributional assumptions. We simulate two different finite populations with size N=10,000, from Log-Normal ($\hat{\mu} = 9.64, \hat{\sigma} = 0.43$) and GB2 ($\hat{a} = 4.11, \hat{b} = 2.16 \cdot 10^4, \hat{p} = 0.47, \hat{q} = 0.92$) with parameters resulting by pseudo-likelihood maximization.

Following the approach of Alfons *et al.* (2013), we create an auxiliary variable attached to each population unit $p_1, \ldots, p_N$ denoting probability weights to mimick a complex survey setting. It has been constructed taking $r = 100$ equally spaced values between 1 and 10 as follows:

$$p_i = \begin{cases} 1 & x_i > F_\theta^{-1}(\frac{r-1}{r}) \\ 10 - \frac{9}{r-1}j & F_\theta^{-1}(\frac{j}{r}) < x_i \leq F_\theta^{-1}(\frac{j+1}{r}) \quad \text{for any} \quad 1 \leq j < r-1 \\ 10 & x_i \leq F_\theta^{-1}(\frac{1}{r}), \end{cases}$$

with F_θ the cumulative distribution function of a Pareto distribution. From each of the

two populations, 3,000 samples are drawn for three different scenarios involving different sample sizes $n = (20, 50, 80)$.

The drawing has been performed by using Midzuno's method for unequal probability sampling Midzuno (1952) with inclusion probabilities proportional to probability weights p in line with Alfons *et al.* (2013). As a consequence, the observations with lower incomes result to have higher inclusion probabilities and in turn lower sample weights. For each sample, the bias-corrected estimators have been calculated using variance estimator under the Midzuno scheme (Narasimha Prasad and Srivenkataramana, 1980).

We transform the General Entropy measure with double-bounded support such as Theil's Index ($\alpha = 1$) and GE($\alpha = 2$) to Relative Entropy measures (RE), i.e. RE(α) = GE(α)/max supp{GE(α)} to deal with measures defined on unique support. Lastly, we consider only the measures defined on the unit interval such as Atkinson Indexes, Gini coefficient and RE, and we fit three distributions for each scenario (varying n). Figure 1.5 displays how changes in family parameter values, underlying income distribution and sample sizes, have an impact on bias-corrected estimators distributions of the Atkinson measures. On the other hand, GE measures do not show relevant changes at varying parameter values.

We consider the well-known Beta distribution, and some alternative distributions defined on the unit interval such as the Simplex distribution (Barndorff-Nielsen and Jørgensen, 1991) and L-Logistic distribution. The Simplex distribution is known to be an alternative to Beta distribution in terms of over-dispersion control (Jørgensen, 1997), robustness (Espinheira and de Oliveira Silva, 2020) and skewness modelling (Carrasco and Reid, 2021). The L-Logistic distribution (Tadikamalla and Johnson, 1982) has two parameters: the median and a shape parameter. Since the median is a natural robust measure of the centre, the median modelling may provide an interesting alternative.

Results are set out in Table 1.5 with a goodness-of-fit evaluation via AIC and BIC information criteria. Consider that all the distributions have the same number of parameters ($p = 2$). Goodness-of-fit EDF statistics are not included due to their unreliability in case of parameters estimated from the data. In all the cases, robust alternative distributions provide better performance than Beta distribution. In case of log-normal assumption, the Simplex distribution provides a greater fit. On the other hand for the GB2 assumption, the distributions of inequality estimators present large variability and very high kurtosis values, as displayed in Table 1.6, and thus a median-modelling distribution such as L-Logistic seems to work better.

	AIC	BIC
Log-Normal	25.75	43.34
Singh-Maddala	28.00	54.38
Dagum	27.76	54.14
GB2	-13.17	22.01

TABLE 1.4: Goodness-of-fit results for income distribution fitting in model-based simulation.

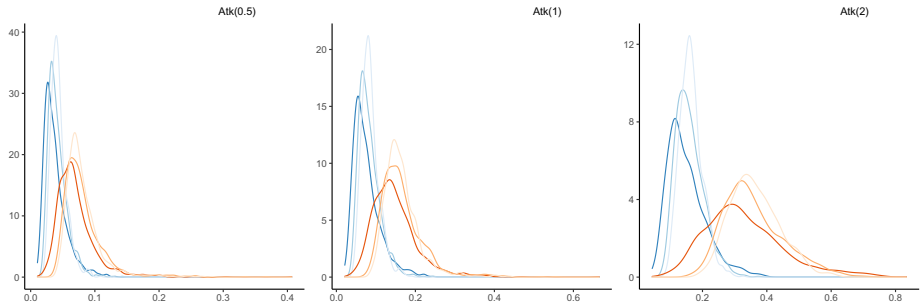

FIGURE 1.5: Atkinson distributions from Log-Normal income assumption (blue lines) and GB2 income assumption (orange lines), darker palette refers to decreasing sample sizes.

1.8 Conclusions

A strategy to correct small samples bias of inequality estimators has been proposed. A sensitivity analysis has also been conducted to study the magnitude of the correction and its sensibility to extreme values. The underlined heterogeneity of sensibilities and bias across measures can guide analysts in choosing the most suitable inequality measure depending on the context. Furthermore, this contribution may prove extremely useful in application contexts. Indeed, the well-known Gini and Theil indexes are widely applied in several fields for inequality and concentration estimation.

Generally speaking, measures that are structurally more sensible to values on the tails appear to be more biased, particularly GE($\alpha = 2$) and Atkinson($\epsilon = 2$), reaching in some cases a bias of more than -20% and more than -45%, respectively. This problem can be circumvented, under the assumption of data contamination, via a semi-parametric Pareto-based treatment of the tails, and bias can be corrected or reduced via our bias correction proposal. However, when extreme values cannot be considered as a consequence of data contamination but rather constitutes representative observations, we suggest using the most robust measure such as GE with $\alpha = 0$ and Atkinson index with $\varepsilon = 1$.

Besides, a bootstrap variance estimation proposal and a distributional evaluation

	Log-Normal Population		GB2 Population	
	AIC	BIC	AIC	BIC
Beta				
RE(1)	-8719.05	-8709.23	-7369.32	-7359.50
RE(2)	-10911.71	-10901.90	-9380.29	-9370.47
A(0.5)	-5776.14	-5766.33	-4455.21	-4445.40
A(1)	-4524.47	-4514.66	-3232.46	-3222.65
A(2)	-3452.54	-3442.72	-1891.27	-1881.45
G	-3487.33	-3477.51	-2772.32	-2762.50
Simplex				
RE(1)	-8794.90	-8785.08	-7507.13	-7497.31
RE(2)	-11019.40	-11009.59	-9601.15	-9591.33
A(0.5)	-5840.08	-5830.27	-4550.89	-4541.07
A(1)	-4576.11	-4566.29	-3280.62	-3270.81
A(2)	-3482.09	-3472.27	-1901.25	-1891.43
G	-3509.29	-3499.48	-2788.63	-2778.82
L-Logistic				
RE(1)	-8783.11	-8773.29	-7589.79	-7579.97
RE(2)	-11018.02	-11008.20	-9708.50	-9698.68
A(0.5)	-5826.16	-5816.34	-4635.19	-4625.38
A(1)	-4559.71	-4549.90	-3351.73	-3341.92
A(2)	-3459.65	-3449.83	-1906.81	-1896.99
G	-3496.99	-3487.17	-2880.50	-2870.68

TABLE 1.5: Goodness-of-fit results for estimator distributions fitting.

		RE(1)	RE(2)	A(0.5)	A(1)	A(2)	G
	$\hat{\eta_3}$	1.68	2.19	1.47	1.26	0.88	0.80
logN	$\hat{\eta_4}$	4.50	8.02	3.37	2.38	1.03	0.77
	$\hat{\eta_3}$	3.45	4.63	2.83	1.98	0.71	1.50
GB2	$\hat{\eta_4}$	19.87	32.92	14.39	7.61	0.46	4.07

TABLE 1.6: Mean values of skewness and kurtosis coefficients of the empirical distributions from model-based simulation.

of the corrected design-based estimators have been developed. Results about the inequality estimators distributions show increasing positive skewness and lepto-kurtosis at decreasing sample sizes, confirming the non-applicability of Gaussian assumption in small samples. As regards measures defined on the unit interval, the model-based simulations shows that alternative robust distributions have a better goodness-of-fit, paving the way to the development of alternative parametric methods of inference when dealing

with inequality in small samples.

A further direction of research includes the extension of this framework to other widely used inequality measures, such as those based on quintiles.

Chapter 2

Put Inequality on the Map: Small Area Model using a Beta Mixture

2.1 Introduction

The issue of widespread economic inequality characterizes the current global predicament and has a central role in political and economic discourse. 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. On the other hand, 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 driver of national income inequality and to study spatial spillovers (Moser and Schnetzer, 2017; Márquez *et al.*, 2019). 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 macro level, being rarely available at local level. 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 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 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 estimates of interest

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) 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 and Rao, 2010; Pratesi, 2016). Fabrizi and Trivisano (2016) deal with Gini index estimation, while Tzavidis and Marchetti (2016) include Gini index and Quantile Share Ratio and, lastly, Marchetti and Tzavidis (2021) consider Gini and Theil indexes. All of them mostly focus on the Gini index and never treat more than two measures at a time.

Inequality can be seen as a multifaceted concept, embracing diverse objective and subjective assessments on the characteristics of income distribution. 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 Gini index, we consider the Relative Theil index, particularly appealing due to its additive decomponsability property, which allows expressing inequality as the sum of between and within components. Moreover, we consider the entire family of Atkinson measures, which stands apart from other descriptive measures by explicitly incorporating a welfare evaluation of inequality implications and enabling for a complete ordering of income distributions. All the measures considered varies 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, being less demanding with respect to data requirements and computational issues, as well as enabling the incorporation of design-based properties.

Concerning unit interval defined responses, most of the small area literature at area-level is dedicated to proportions. This strand gathers linear mixed models with suitable transformations (Rao and Molina, 2015) and Beta regression models (Janicki, 2020). For the first body, we recall Marhuenda *et al.* (2013) and Esteban *et al.* (2012, 2020), providing Fay-Herriot model (i.e. with Gaussian assumptions) extensions to compositional, spatial and/or temporal structures. Whereas the body of literature based on

Beta regression, comprises mostly univariate proposals in Liu *et al.* (2007), Bauder *et al.* (2015) and Fabrizi and Trivisano (2016). Multivariate versions can be found in Souza and Moura (2016) by using copula functions and in Fabrizi *et al.* (2011); zero and/or one inflated extensions are adopted in Wieczorek *et al.* (2012) and Fabrizi *et al.* (2020).

Inequality design-based estimators have peculiar characteristics; their behaviour in complex survey small samples have been investigated in Chapter 1, showing highly skewed and heavy-tailed distributions at decreasing sample sizes. Besides, Gaussian or Beta regression options fail in case of skewed and heavy-tailed estimators such as the ones of interest, for the Beta case see Bayes *et al.* (2012) and Migliorati *et al.* (2018). The inadequacy of Fay-Herriot models, in case of not well-behaved distribution, is particularly established; some proposals face this issue by adopting alternative likelihoods in case of responses with infinite or positive support at area-level: we mention skew-normal (Ferraz and Moura, 2012; Ferrante and Pacei, 2017), skew-t (Moura *et al.*, 2017) and log-normal likelihoods (Slud and Maiti, 2006; Fabrizi *et al.*, 2018). Moreover, a Fay-Herriot model may fit values outside the variable support and potential transformations may affect interpretability.

Our proposal involves incorporating an alternative likelihood assumption by adopting a Beta mixture-based approach, whose performances are compared with a Beta regression proposal. Specifically, we assume as sampling distribution the Flexible Beta, 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, 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 target estimators, avoiding to highly underestimate inequality and providing reliable estimates.

The chapter is organized as follows. Inequality measures and their estimators are

defined and described in Section 2.2, together with a proposal of sampling variance estimation. Section 2.3 defines the proposed Beta and Flexible Beta small area models. An application on EU-SILC income data is unravelled in Section 2.4 and a design-based simulation can be found in Section 2.5, in order to evaluate the frequentist properties of model-based estimators. Conclusions are drawn in Section 2.6.

2.2 Inequality Measures

In this section, we describe the inequality measures considered and their estimator in complex survey case. Such estimators are known to be biased in small samples, often leading to underestimation. Therefore, we adopt the bias-corrected estimators proposed in Chapter 1. Their simulation results lead us to assume that our bias-corrected estimators are approximately unbiased or slightly biased depending on the domain sample size. Ultimately, in Subsection 2.2.1, 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 equivalent disposable income, which is observable for each unit in $\mathcal{S}_{iid}$. The *iid* Gini estimator is defined as

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

with r_i the rank of the i-th unit and $\hat{\mu}$ 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 in Chapter 1 as follows

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

with $\tilde{n} = \sum_{k\in\mathcal{S}} \mathbb{1}(w_k \neq 0)$ and $\hat{\gamma} = \sum_{i\in\mathcal{S}} w_i z_i (\hat{N}_i - w_i/2)/\hat{N}^2$.

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 i.e. scaled on the maximum of its support ($\log n$). Its estimator in the *iid* case is defined as follows

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

In the complex survey case, the Horwitz-Thompson type estimator for the Theil index has been considered in its bias-adjusted formulation of Chapter 1. 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

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

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

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. They provide for an explicit value judgement 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 (La Vega *et al.*, 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 and Liberati, 2006). Under a concave utility function (whose concavity level is regulated by ε), the estimator of Atkinson index in the *iid* case is defined as

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

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, showing at the same time different sensitivities. The estimator referred to complex survey case (Biewen and 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-corrected versions defined in Chapter 1 as follows

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

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

2.2.1 Design Variance Estimation

The sampling variances of complex survey estimators has 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 (Fabrizi *et al.*, 2020), using 1,000 bootstrap samples. Secondly, the raw estimates have been smoothed via a Generalized Variance Function (GVF) approach in order to reduce the sampling error 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. Gini index result, as well as our following derivations for the other measures, has been directly incorporated in 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 $d = 1, \ldots, D$, with $n = \sum_{d=1}^{D} n_d$ and $\tilde{n} = \sum_{d=1}^{D} \tilde{n}_d$.

Proposition 2.1. *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 $j = 1, \ldots, n_d$. The simple random sampling (srs) estimator of Atkinson index for domain d, denoted with $A_d(\varepsilon)$, has variance function*

$$\mathbb{V}[A_d(\varepsilon)] \cong \frac{2\theta_d^A(\varepsilon)^2}{n_d} \exp\{-2\theta_d^A(\varepsilon)\}, \tag{2.1}$$

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

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

$$\theta_d^A(\varepsilon) = 1 - \exp\{-\varepsilon\varphi_d^2/2\}, \tag{2.2}$$

with φ_d^2 estimated by $s_d^2 = \frac{1}{n_d-1}\sum_{j=1}^{n_d}[\log(z_{jd}) - \hat{\mu}_d]^2$. By applying the normal distribution theory, $\mathbb{V}[s_d] \cong \frac{\varphi_d^2}{2n_d}$ and using the delta method:

$$\begin{aligned}\mathbb{V}[A_d(\varepsilon)] = \mathbb{V}\left[1 - \exp\left\{-\frac{\varepsilon s_d^2}{2}\right\}\right] &\cong \frac{\varepsilon^2\varphi_d^4}{2n_d}\exp\{-\varepsilon\varphi_d^2\} \\ &\cong \frac{2\theta_d^A(\varepsilon)^2}{n_d}\exp\{-2\theta_d^A(\varepsilon)\}, \end{aligned} \tag{2.3}$$

where equation (2.3) is obtained by McLaurin expanding (2.2), so that $\varphi_d^2 \cong 2\theta_d^A(\varepsilon)/\varepsilon$. Note that this result can be easily generalized to the case $\varepsilon = 1$. □

Proposition 2.2. *Under Proposition 2.1 assumptions, the srs estimator of Relative Theil Index, for domain d, R_d has variance function*

$$\mathbb{V}[R_d] \cong \frac{2\theta_d^{R2}}{n_d}, \tag{2.4}$$

where θ_d^R denotes its population value.

Proof. Similarly to Proposition 2.1 proof, the Relative Theil index is defined in a log-normal income population as

$$\theta_d^R = \frac{1}{\log(n_d)}\left(\frac{\mathbb{E}[z\cdot\log(z)]}{\mathbb{E}[z]} - \log(\mathbb{E}[z])\right) = \frac{\varphi_d^2}{2\log(n_d)}. \tag{2.5}$$

Since the moments involved in the previous expression are

$$\mathbb{E}[z] = \exp\left\{\mu_d + \frac{\varphi_d^2}{2}\right\} \tag{2.6}$$

$$\begin{aligned}\mathbb{E}[z\cdot\log(z)] &= \int_0^{+\infty} z\log(z)\frac{1}{z\sqrt{2\pi\varphi_d^2}}\exp\left\{-\frac{[\log(z)-\mu_d]^2}{2\varphi_d^2}\right\}dx \\ &= \int_{-\infty}^{+\infty} t\exp\{t\}\frac{1}{\sqrt{2\pi\varphi_d^2}}\exp\left\{-\frac{(t-\mu_d)^2}{2\varphi_d^2}\right\}dt \\ &= (\varphi_d^2+\mu_d)\exp\left\{\mu_d + \frac{\varphi_d^2}{2}\right\}, \end{aligned} \tag{2.7}$$

where the last step (2.7) involves equation 3.462.6 in Gradshteyn and Ryzhik (2014). Considering that φ_d^2 is estimated by s_d^2 and $\mathbb{V}[s_d] \cong \frac{\varphi_d^2}{2n_d}$, by applying delta method the result follows

$$\begin{aligned}\mathbb{V}[R_d] = \mathbb{V}\left[\frac{s_d^2}{2\log(n_d)}\right] &\cong \frac{\varphi_d^4}{2\log^2(n_d)n_d} \\ &= \frac{2\theta_d^{R2}}{n_d} \end{aligned} \tag{2.8}$$

where equation (2.8) is obtained by (2.5) considering that $\varphi_d^2 = 2\theta_d^R \log(n_d)$. □

Since Proposition 2.1 and 2.2 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 2.4, as synthetic population by considering 21 NUTS-2 Italian regions as domains of interest. In order to circumvent non-robustness, 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 2.1 and 2.2 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, leads 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 2.1 and 2.2, defined approximately for domain d as

$$\mathbb{V}(G_d) \cong \frac{\theta_d^{G^2}(1-\theta_d^{G^2})}{n_d}, \tag{2.9}$$

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. Let denote with $\hat{\mathbb{V}}[\cdot]_{boot}$

the raw 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 (2.1), (2.4), (2.9) 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. 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.

Results of Propositions 2.1 and 2.2 show a very different structure for the variance function of Atkinson and Relative Theil indexes with respect to the Gini index and to the proportion ones. A comparison plot can be found in the Appendix. As opposed to the proportion and Gini index cases, the Atkinson and Relative Theil variance functions are both monotonically increasing, as clear from (2.2) and (2.5). Thus, higher values of θ_d correspond to higher log income dispersion, inevitably leading to an increase of the index variability. The explosive trend of Relative Theil variance is related to its explosive connection with log income population variance (2.5). Moreover, notice that the variance function of Atkinson index in (2.1) 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 the Appendix.

The precision estimates obtained for NUTS-3 Italian regions, employing EU-SILC data (described in Section 2.4), 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). Let us consider it as a rule of thumb, since not all the domains are equivalent, being associated with different population rates. However, 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 Gini index is significantly lower, confirming the great robustness properties of the index, ranging from 3% to 27%, with 7 domains having out-of-bound CV. This motivates us to employ a small area model.

2.3 Small Area Models

We propose a Beta mixture model for small area estimation by adopting a Bayesian framework, from now on we name it 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 one in Subsection 2.3.1, then we set out our proposal in Subsection 2.3.2, models are completed by the prior setting in Subsection 2.3.3 and their Bayesian estimation is detailed in Subsection 2.3.4.

2.3.1 The Beta Model

Let us consider the Beta distribution with mean-precision parametrization (Ferrari and Cribari-Neto, 2004), such that a generic random variable Beta distributed 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

$$\mathbb{E}[Y] = \mu, \qquad \mathbb{V}[Y] = \frac{\mu(1-\mu)}{\phi+1}, \tag{2.10}$$

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. The sampling level models the conditional distribution of the direct estimator 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, a small area Beta model uses to 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 (2.10), $\phi_d + 1$ 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, 2016) and Fabrizi and Trivisano (2016) similar to the one applied in Subsection 2.2.1,
- 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 2.2.1. 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 2.1 and 2.2, leading to a tricky and intractable parametrization. In second place, the known variance assumption is a standard approach across different small area models. 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.

2.3.2 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 p.d.f. 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 (Migliorati *et al.*, 2018).

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

$$\mathbb{E}[y_d|\boldsymbol{\eta}] = \theta_d = \lambda_{2d} + p \cdot \tilde{w}_d, \tag{2.11}$$

$$\mathbb{V}[y_d|\boldsymbol{\eta}] = \frac{\theta_d(1 - \theta_d) + p(1 - p)\tilde{w}_d^2\phi_d}{\phi_d + 1}. \tag{2.12}$$

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

$$\begin{aligned} &\text{logit}(\lambda_{2d})|\boldsymbol{\beta}, v_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + v_d \\ &v_d|\sigma_v^2 \overset{ind}{\sim} \mathcal{N}(0, \sigma_v^2) \quad \forall d. \end{aligned} \tag{2.13}$$

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 (2.11). 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 (2.11) and (2.13) 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, its estimate is basically overlapping the one of the Beta model in Section 2.3.1 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 (Wieczorek *et al.*, 2012; Fabrizi *et al.*, 2016), and skew-normal models (Ferraz and Moura, 2012; Ferrante and Pacei, 2017).

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 2.2.1. 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 given value in a small area model. As a consequence, the dispersion parameter is not directly estimated and can be obtained from (2.12) as

$$\phi_d|\theta_d, p, \tilde{w}_d = \frac{\theta_d(1-\theta_d) - \mathbb{V}[y_d|\boldsymbol{\eta}]}{\mathbb{V}[y_d|\boldsymbol{\eta}] - p(1-p)\tilde{w}_d^2}. \tag{2.14}$$

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 2.3. *Under FB model and the assumptions of Proposition 2.1, let us consider ϕ_d and its relation with the other parameters defined in (2.14). In order to preserve its bounded support, i.e. $\phi_d > 0$, $\tilde{w}_d$ has to be constrained such that*

$$\begin{cases} \tilde{w}_d < \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}} & \text{if} \quad \theta_d < c \\ \tilde{w}_d > \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}} & \text{if} \quad \theta_d > c \end{cases} \tag{2.15}$$

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 (2.14), the solution comes out to be

$$\mathbb{V}[y_d|\boldsymbol{\eta}] \in \left(\min\left\{\theta_d(1-\theta_d), p(1-p)\tilde{w}_d^2\right\}, \max\left\{\theta_d(1-\theta_d), p(1-p)\tilde{w}_d^2\right\}\right). \tag{2.16}$$

Without loss of generality, we consider the case

$$\mathbb{V}[y_d|\boldsymbol{\eta}] < \theta_d(1-\theta_d), \tag{2.17}$$

substituting $\mathbb{V}[y_d|\boldsymbol{\eta}]$ in (2.17) with (2.1), (2.4) and (2.9), we obtain three inequalities on θ_d. The generic result for each inequality is denoted with $\theta_d < c$, where c depends on n_d and differs for any measure. After splitting the problem into two cases, namely $\theta_d < c$ and $\theta_d > c$, we solve equation (2.16) for $\tilde{w}_d$ obtaining (2.15). □

To understand the behaviour of the previous constrains within our inferential problem, we evaluated them 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 Relative Theil index, 0.84 for Atkinson indexes and 1 for Gini index. Discarding Gini index being always $\theta_d^G < 1$, observing $\theta_d > c$ is totally implausible for the considered income inequality measures. Indeed, those values correspond to far-fetched values of log income variable dispersion that, following (2.2) and (2.5), 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 in Chapter 1, shows $\hat{\varphi}^2 = 0.18$. Therefore, we opt to consider only the case $\theta_d < c$ in Proposition 2.3; the same reasoning could be easily done in case of proportions, where $\theta_d < c$ holds for any $n_d > 1$.

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

$$\tilde{w}_d \in \left(0, \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}}\right\}\right), \tag{2.18}$$

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 2.3 and clearly takes into account the fact that the imposed sampling variance (given p) has to constrain the distance between the mixture

components. Thus, following (2.18), the distance has been modelled 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, the 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 (2.11) can be rewritten as

$$\theta_d|\lambda_{2d}, p, w = \lambda_{2d} + p \cdot w \cdot \min \left\{ \frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}[y_d|\boldsymbol{\eta}]}{p(1-p)}} \right\}, \tag{2.19}$$

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 (Moura *et al.*, 2017; Ferrante and Pacei, 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:

$$y_d|\boldsymbol{\eta} \overset{n_d \to +\infty}{\sim} Beta(\theta_d\phi_d, (1-\theta_d)\phi_d). \tag{2.20}$$

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 (2.10). 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 $w \to 0$, (2.20) 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 and shrinkage. This can be seen as the main characteristic of the FB small area model.

2.3.3 Prior Distributions

The following weakly-informative priors complete the Beta model:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.21}$$

$$\sigma_v \sim \text{Half-}\mathcal{N}(0, \nu^2), \tag{2.22}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal $10 \times \mathbf{1_p}$. Considering the scale of the logit transformation, $\nu^2 = 1$ can be seen as quite a non-informative option. As regards FB model, the prior choice includes, in addition to (2.21) and (2.22),

$$p \sim \text{Unif}(0, 1) \quad \text{and} \quad w \sim \text{Unif}(0, 1). \tag{2.23}$$

In order to foster convergence or avoid convergence problems, in some specific cases, e.g. when dealing with a few areas, we recommend using a slightly informative prior for the mixing coefficient such as $p \sim Beta(2, 2)$, being able to avoid the boundaries of its support but still being very close to a uniform distribution.

2.3.4 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 and 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 *et al.*, 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

$$\hat{\theta}_d^{HB} = \mathbb{E}[\theta_d|\text{data}] \quad \forall d, \tag{2.24}$$

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 bound. 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 (2.20), we can state that the design-consistency property is preserved also under the FB model.

2.4 Application on EU-SILC Data

We are interested in estimating inequality in Italian NUTS-3 regions using EU-SILC data. Given the high level of uncertainty of the direct estimates, described in Subsection 2.2.1, 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 2.4.1. Eventually, model results are compared in Subsection 2.4.2.

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).

2.4.1 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 and industrial specialization and demographic structure.

A small area model does not have causal inference ambitions, but rather it requires auxiliary information to be accurately known, without error, at population level. Therefore, we must restrict the choice to accessible data: census and registry office data as well as tax forms data, publicly available. 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 included average taxable income claimed by private residents, percentage of residents aged more than 15 filling tax forms and percentage of residents with income lower than/greater than double 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 and Trivisano, 2016).

		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 2.1: The **looic** and related standard error as well as average **cvr** for each model and each measure.

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 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 in the models, in order to harmonize the scale, and subjected to preliminary variable selection to avoid multicollinearity.

2.4.2 Results

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. A models 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, is invariant to parametrization and works for singular models.

In order to evaluate model-based estimators performances in comparison with direct estimators, the Coefficient of Variation Reduction measure (Ferrante and Pacei, 2017)

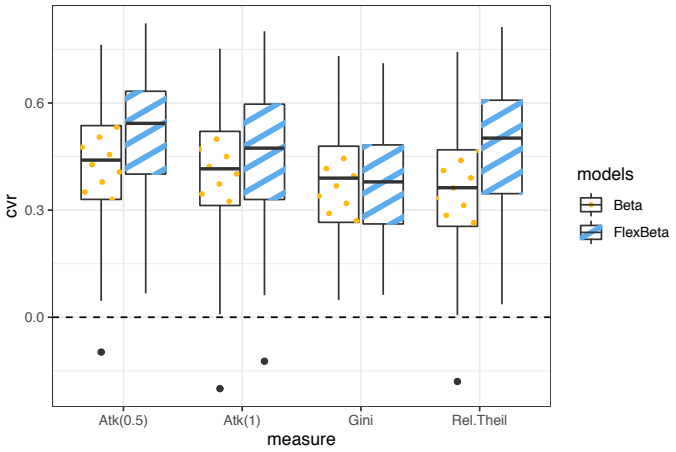

FIGURE 2.1: Coefficient of variation reduction for each model and each measure.

is used to calculate the precision improvement:

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

using predictors and their posterior variances for any HB model. This constitutes a frequently used measure for small area model evaluation. However, 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 (Ferrante and Pacei, 2017).

Diagnostics for each model, **looic** and **cvr**, averaged among all domains (**acvr**), are displayed in Table 2.1, while the full distribution of **cvr** is displayed in Figure 2.1. Results show better goodness-of-fit and coefficient of variation reduction for the Flexible Beta model with respect to the Beta one. This holds for all measures with the exception of Gini index, where diagnostics do not vary significantly between models. The main point is that, following the distributional analysis of Chapter 1, the Gini index estimator has the most well-behaved distribution 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 any 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 2.2, 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 (3.2), thus let us consider them as global

distributions, not related with domain-specific posteriors. The FB model-based estimates tend to be greater than Beta model ones as clear from scatterplots in Figure 2.2, capturing the right-hand tail of the distribution and avoiding underestimating inequality, as it will be more intelligible in Section 2.5. The Gini index case shows, again, large similarities across the two models.

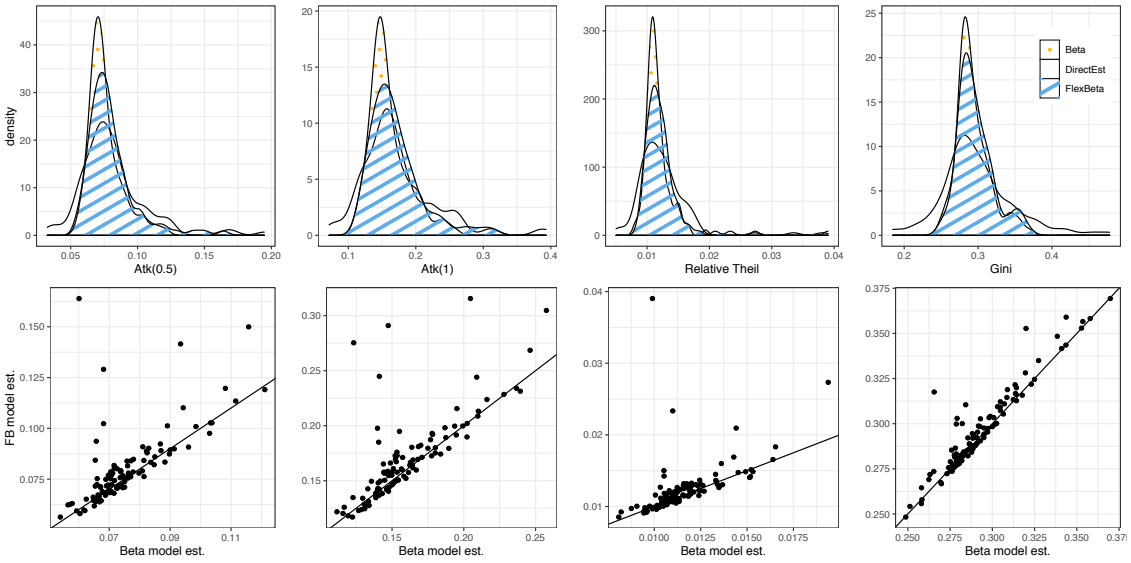

FIGURE 2.2: Densities of model-based estimates versus direct estimates and scatterplot of model-based estimates with bisector line

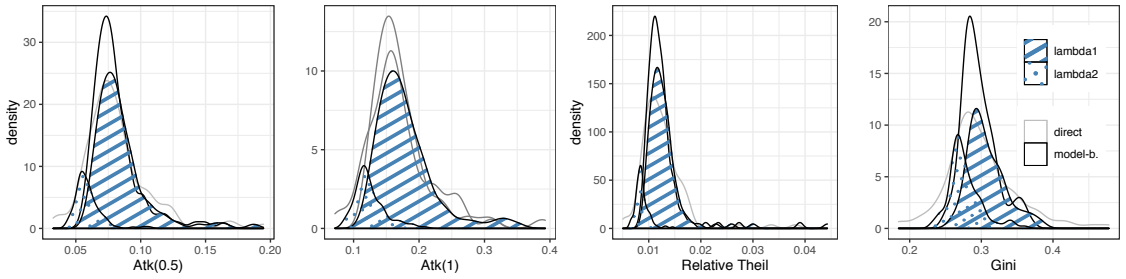

FIGURE 2.3: Posterior means distributions of the mixture components expected values, weighted for $\mathbb{E}[p|\text{data}]$, in comparison with direct estimates and Flexible Beta model-based estimates

Deep diving on FB estimates, Figure 2.3 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. Notice that for all measures except for the Gini index, the second mixture component, embracing lower inequality values, has less weight, and helps to model inequality estimators by shifting the first component (and overall) mode towards the centre of the distribution.

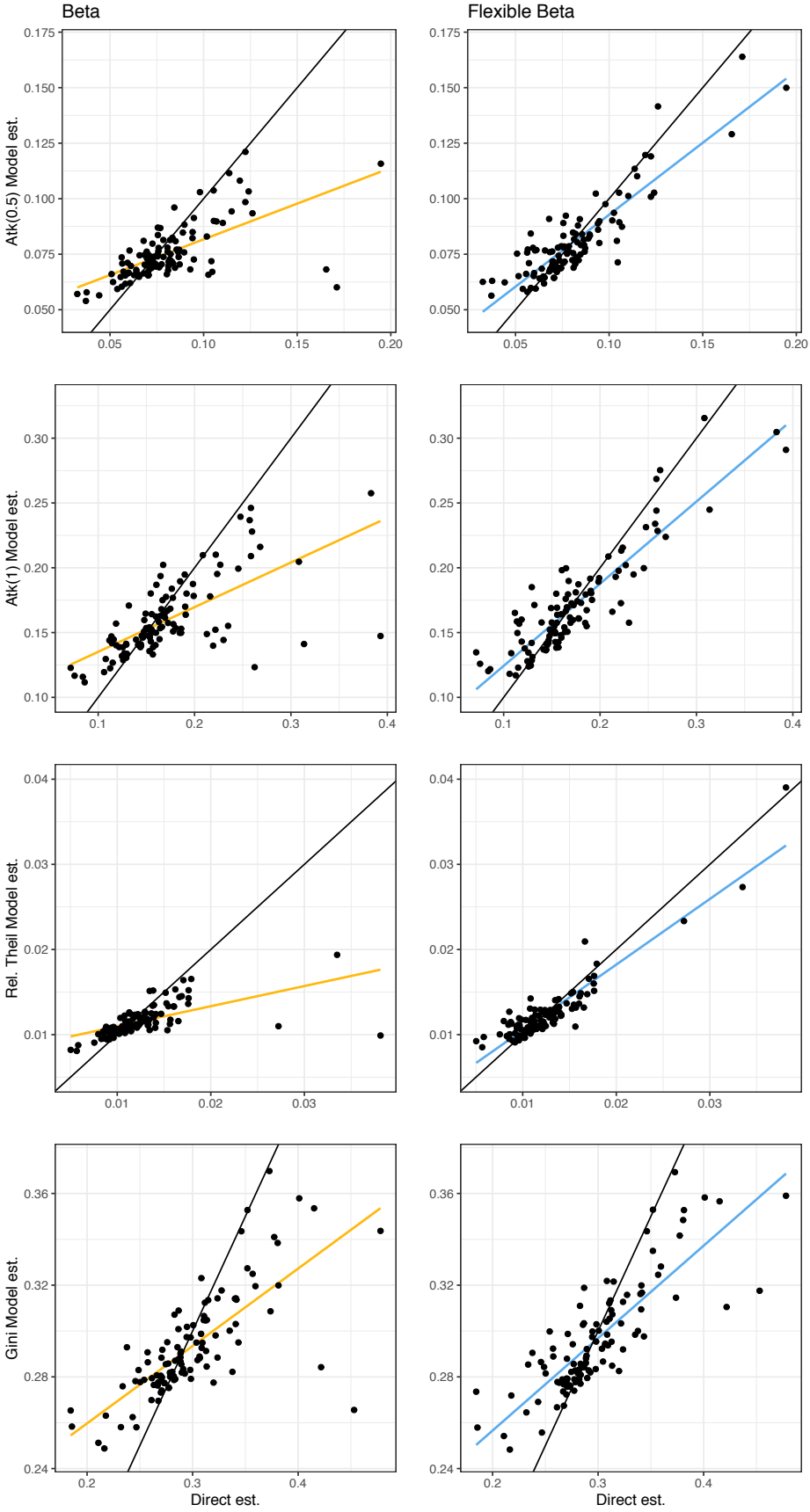

FIGURE 2.4: Shrinking process for each measure in Beta and Flexible Beta models: bisector in black, coloured linear regression line

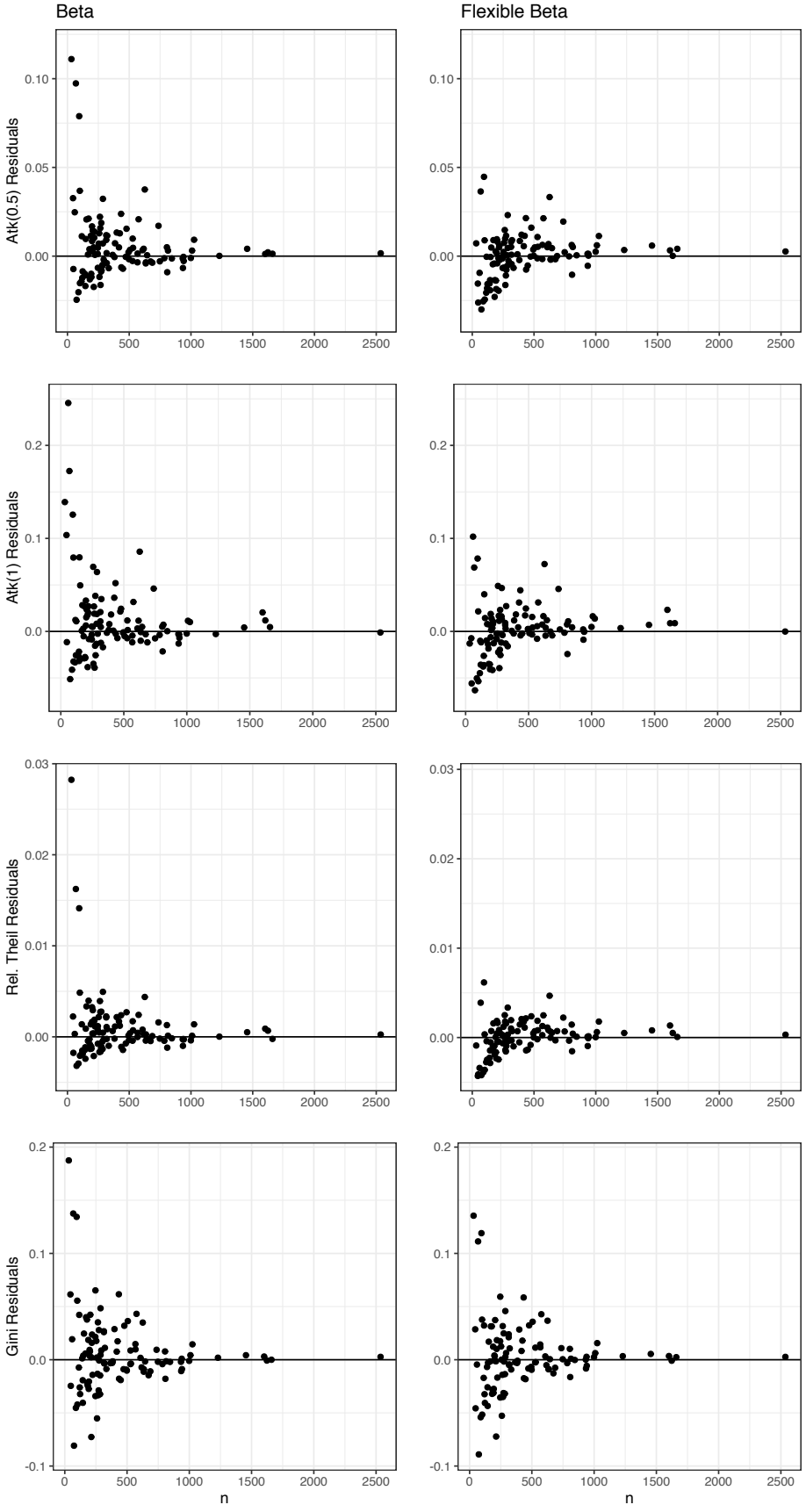

FIGURE 2.5: Design consistency check for each measure in Beta and Flexible Beta models

The shrinking process is displayed in Figure 2.4 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 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 2.5. 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.

2.5 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 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 synthetic subpopulation, $S = 1,000$ samples have been extracted by mimicking complex EU-SILC design, with stratification, multi-stage selection and 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 2.2.1 (Finkelstein *et al.*, 2006; Masseran *et al.*, 2019). The drawn 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 2.4.1. 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 2.4 regards the prior of mixing coefficient p in (2.23), 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:

$$
\begin{aligned}
\text{RB}(\hat{\theta}_d^{HB}) &= \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\theta}_{ds}^{HB}}{\theta_d} - 1\right), \\
\text{ARB}(\hat{\theta}_d^{HB}) &= \left|\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\theta}_{ds}^{HB}}{\theta_d} - 1\right)\right|, \\
\text{MSE}(\hat{\theta}_d^{HB}) &= \frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_{ds}^{HB} - \theta_d)^2, \\
\text{RMSE}(\hat{\theta}_d^{HB}) &= \frac{\text{MSE}(\hat{\theta}_d^{HB})}{\theta_d^2}, \\
\text{AEFF}(\hat{\theta}^{HB}) &= \sqrt{\frac{\sum_{d=1}^{D}\text{MSE}(y_d)}{\sum_{d=1}^{D}\text{MSE}(\hat{\theta}_d^{HB})}}.
\end{aligned}
$$

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 2.2. 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 the cases. Considering bias and variance components of the error, the first one shows a noticeable decrease in the case of FB estimators, as confirmed by ARB and RB values in Table 2.2, 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 2.6. 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 2.2, 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 2.6, 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. Moreover, notice that, except for the Gini index, the MSE of Beta estimators does not decrease at increasing sampling rate, having substantially identical distribution for 3% and 5% scenarios, whereas the MSE of FB estimators diminishes at increasing rates. We can deduce that the first one has the bias as predominant component of the MSE, being invariant to sampling rate. On the other hand, the second one is dominated by the variance which, in a finite population context, highly depends on the sampling rate.

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 2.7. 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.

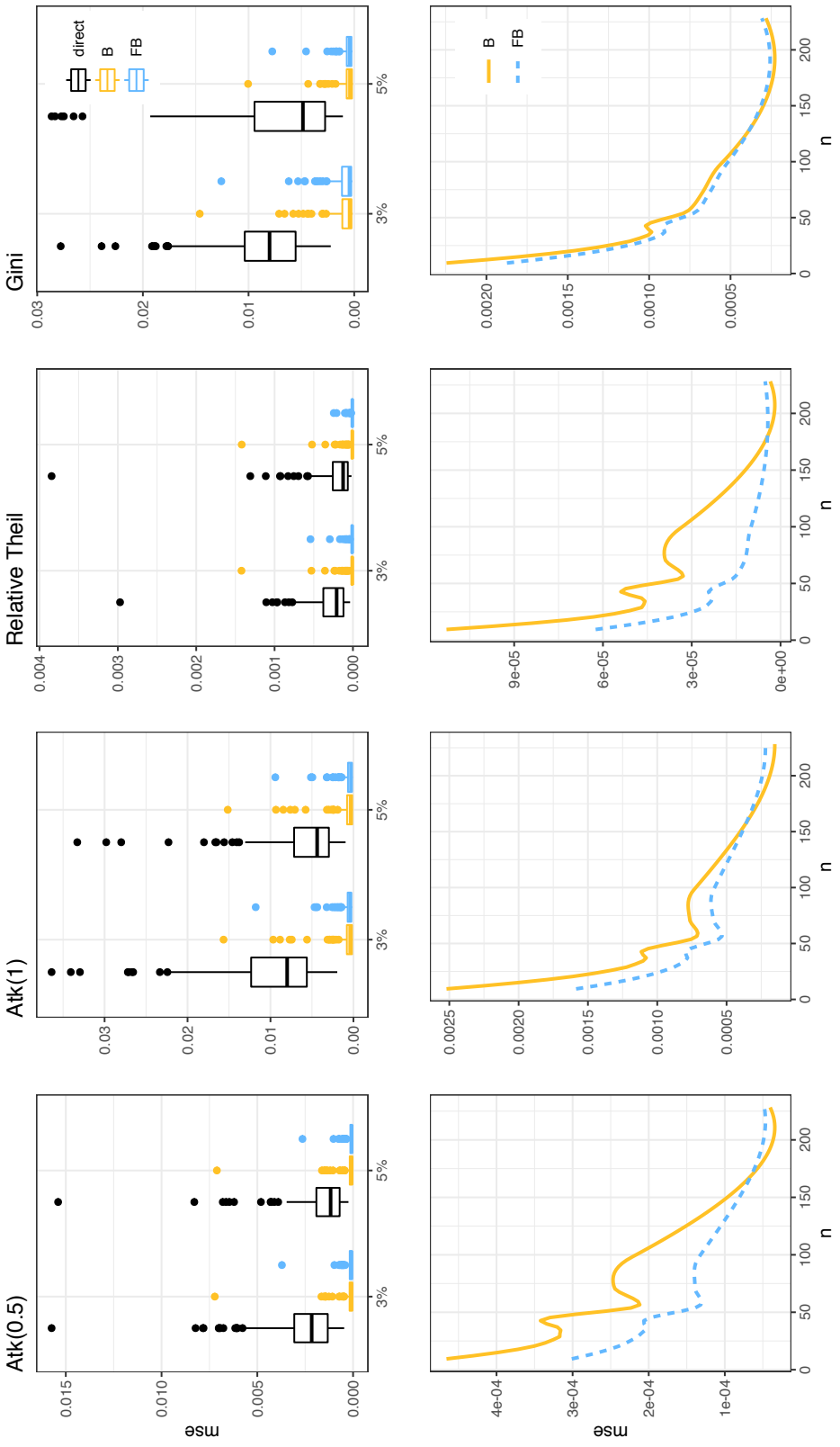

FIGURE 2.6: MSE for each area. Plots on the top row show direct estimators values versus model-based estimators ones, while bottom row plots zoom-in model-based results versus sample sizes

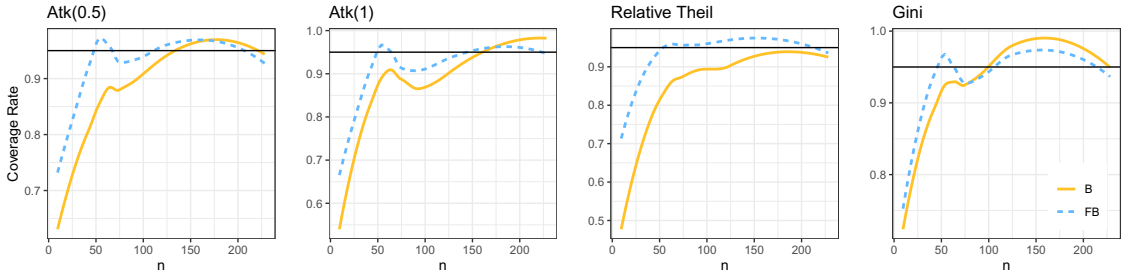

FIGURE 2.7: Coverage rate versus sample sizes in 5% simulation samples, black line fixed at the nominal level 0.95

The similarity among Beta and FB model-based estimates, in case of Gini index is confirmed also in simulation. 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.

2.6 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 of 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 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

parameters. 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.

Our findings provide a basis for further research focused on multiple directions. Firstly, it would be worthwhile to further study the shrinking process in the FB model, by investigating the asymptotic behaviour of its estimator when model variance tends to zero. Secondly, 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.

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

TABLE 2.2: ARB, RB, RMSE, AEFF jointly with the coverage of direct estimators (de), and model-based estimators concerning Beta (B) model and Flexible Beta (FB) model.

Chapter 3

The tipsae package: Tools for mapping Indices and Proportions in Small Area Estimation

3.1 Introduction

The growing demand for timely and reliable statistical estimates leads to extensive exploitation of survey data at an increasingly greater level of disaggregation. However, domains or areas of study are often different from the ones for which the survey was originally planned, leading to possibly unreliable direct estimates due to observations-poor samples. Small Area Estimation (SAE) tackles this problem by providing a set of indirect estimation techniques, relying on external information, which borrow strength across areas and increase the efficiency of the estimates. Indirect estimators based on explicit regression models are labelled model-based estimators and assume a relationship between the target variable and explanatory variables, which remains constant across areas. Classical small area models embrace two basic linear mixed models: the Fay–Herriot model and the Battese-Harter-Fuller model, which are foundational for the strand of area-level models and unit-level models, respectively (Rao and Molina, 2015). While the former relates area-specific target quantities to area covariates, the second one relates individual observations of the underlying variables of interest to individual covariates.

Hereafter, we focus on area-level models due to their practical convenience. They require, in fact, only data aggregated at the area-level, avoiding both computational and data disclosure issues. In area-level contexts, a well-established body of literature is concentrated on Gaussian models. However, many quantities of interest have specific

features that are not considered in the Gaussian setting and need to be accounted for, such as a bounded or double bounded support and distributions markedly skewed or heavy-tailed. Specifically, we focus on unit interval responses, common in SAE modelling because of the growing need for rates and proportions releases in official statistics, such as Head-Count Ratio for poverty mapping or Health Insurance Coverage rates. Not to mention the treatment of other measures of interest defined in (0,1) or [0,1], such as some inequality measures (e.g., Gini index).

In this regard, two different bodies of literature revolve around linear mixed models with suitable transformations (Rao and Molina, 2015) and Beta regression models (Janicki, 2020). For the first approach, we recall the works by Marhuenda *et al.* (2013, 2014), Morales *et al.* (2015), and Esteban *et al.* (2012, 2020) that provide Fay-Herriot extensions to deal with proportions. The second strand focus on classical Beta regression, both in the univariate case (Liu *et al.*, 2007; Bauder *et al.*, 2015; Fabrizi and Trivisano, 2016; Giovinazzi and Cocchi, 2021) and in the multivariate ones (Fabrizi *et al.*, 2011; Souza and Moura, 2016), considering also zero and/or one inflated extensions (Wieczorek *et al.*, 2012; Fabrizi *et al.*, 2016, 2020). Lastly, a Beta mixture approach in SAE has been proposed in Chapter 2.

By considering the SAE field as a whole, there is a clear imbalance between a plethora of methodological proposals defined in academic literature and the tight circle of methods actually used in official statistics and applied researches. A bridge-building process between methodological and applied fields is needed, involving collaboration, dissemination, and development of user-friendly tools to facilitate tough steps. With the latter aim, several routines for SAE have been released by developer teams of `R`, `SAS`, `SPSS`, and `STATA`. Our focus is on `R` routines (R Core Team, 2021) due to flexibility and availability reasons as well as for the equipment of complementary tools. Several `R` packages have been developed to implement SAE tools, and in the following, we attempt to provide a clear overview focusing on model-based methods.

In general, the most complete released packages are:

- **sae** (Molina and Marhuenda, 2015). It implements a wide range of small area methods from a frequentist perspective, including both area-level and unit-level models.

- **emdi** (Kreutzmann *et al.*, 2019). It allows making inference on both area-level and unit-level models in a frequentist framework, providing model diagnostics, plots, and exporting tools.

- **mcmcsae** (Boonstra, 2021). It comprises hierarchical area and unit-level models estimated via Markov Chain Monte Carlo (MCMC) simulation, allowing for spatial and temporal dependencies. It includes different prior settings, model diagnostics, and posterior predictive checks functions.

Among listed packages, only the **emdi** package directly accounts for unit interval responses at area-level by providing the arc-sin transformation in a Gaussian setting (Schmid *et al.*, 2017). Thus, while a Fay-Herriot model for unit interval responses may be implemented via existing packages, Beta-based small area models lack proper implementations.

The **tipsae** package aims at filling this gap by implementing Beta-based small area models specified at the area-level on measures that can assume values in $(0, 1)$, $[0, 1)$, $(0, 1]$, and $[0, 1]$ intervals. We decided to operate in a Bayesian fashion in order to exploit the advantages brought by approaching this inferential framework via MCMC methods. For instance, it is possible to easily manage non-Gaussian assumptions, incorporate structured random effects, obtain straightforward estimates for out-of-sample areas, and capture the uncertainty about all target parameters through posterior inference. Nowadays, several tools are available to implement Bayesian models with probabilistic languages: our choice falls on `Stan` (Carpenter *et al.*, 2017), that can be easily employed to fit statistical models within `R` packages thanks to the tools provided by the **rstantools** package (Gabry *et al.*, 2020).

The main features of the **tipsae** package are listed in the following:

- It includes a variety of area-level models based on the Beta likelihood. Besides the standard Beta-regression model, Zero and/or One Inflated Beta (ZOIB) and Flexible Beta models can be chosen. Moreover, particular dependence structures can be modelled, including spatial and/or temporal random effects.

- It implements an efficient Hamiltonian Monte Carlo (HMC) fitting algorithm and customized parallel computing imported from **rstan** (Stan Development Team, 2020). We also tested other languages that build MCMC samplers, and `Stan` turned out to be the most efficient one for Beta regression models, which are particularly tricky to handle due to the non-orthogonality between location and scale parameters.

- The `stanfit S4` object produced by the **rstan** package can be exploited to check convergence, monitor sampler diagnostics, and, lastly, perform an exhaustive posterior analysis, relying on existing tools such as **loo** (Vehtari *et al.*, 2020) and

bayesplot (Gabry and Mahr, 2021) packages. In this way, users familiar with posterior predictive checks can carefully assess the model performance.

- Specific diagnostics for small area models are produced by ad-hoc functions, facing the most relevant aspects to deepen within the SAE framework. We implemented both visualization tools for graphical assessments and functions that easily export the final results. Moreover, variance smoothing routines and benchmarking procedures are also provided, remarking that, to the best of our knowledge, the first tool is not available in any existing SAE package.

- To further facilitate the workflow for non-expert users of `R`, a Shiny application (Chang *et al.*, 2021) with an intuitive graphical user interface can be launched through the `runShiny_tipsae()` function. The application assists the user in carrying out a complete SAE analysis, exploiting all the main features of the **tipsae** package.

The package is freely available at the repository `https://github.com/silviadenicolo/tipsae` and can be installed through the following command.

```
R> devtools::install_github("silviadenicolo/tipsae")
```

The chapter is organized as follows: covered models and implemented methodology are set out in Section 3.2, the datasets made available in the package are presented in Section 3.3, while Section 3.4 provides a step-by-step description of inputs and outputs of the available functions. Lastly, Section 3.5 contains some concluding remarks, discussing possible extensions that could be supplied.

3.2 Methodology

In this section, the theory behind the statistical methods implemented in the **tipsae** package is summarized. The main aspects are those related to the area-level models for indices and proportions that can be estimated using the function `fit_sae()`.

From now on, we consider a finite population of size N that is partitioned into D small areas having sizes $N_1, \ldots, N_D$. We are interested in estimating a generic measure defined on the unit interval that we denote as $\theta_d,\ d = 1, \ldots, D$. To this aim, a random sample of size n is drawn from the whole population using a possibly complex survey design, obtaining sub-samples of sizes $n_1, \ldots, n_D$, specified for each domain. Among them, we define the first $\tilde{D}$ domains, with $\tilde{D} \leq D$ as the ones actually observed, i.e., with $n_d > 0$. The observations recorded at the individual level are aggregated to produce the direct

estimates y_d, that are stored in the vector $\boldsymbol{y}$ and are the observed determinations of the direct estimator Y_d for a quantity of interest θ_d, with $d = 1, \dots \tilde{D}$. The Bayesian area-level model is specified for Y_d, including also a set of auxiliary variables $\mathbf{x}_d$, which are assumed to be available for each domain.

The details about the statistical models that can be set through the argument `likelihood` are discussed in Section 3.2.1. Furthermore, a small area model usually includes also random effects in the linear predictor. The random effect part, hereafter indicated with e_d, can incorporate either a temporal and/or a spatial dependency structure, as will be discussed in Section 3.2.2, devoted to the prior specification settings. In addition, different prior assumptions can be specified for the unstructured random effects, allowing for robust and shrinking priors.

In small area models, the dispersion parameters are generally assumed as given and previously estimated from the data. Separate estimation could involve a smoothing procedure to refine the sampling variances estimates and reduce their errors. Section 3.2.3 describes the proposed algorithms to carry out this step if required. Eventually, Section 3.2.4 outlines the main aspects of posterior inference: we will mainly focus on the out-of-sample treatment, diagnostics, and goodness-of-fit tools employed to validate or select the models and, lastly, the benchmarking procedures complementing SAE analysis.

3.2.1 Area-Level Models: Likelihoods

The statistical models available in **tipsae** are set out in the following subsections, whereas a comprehensive overview of the key quantities under each model is provided in Table 3.1. In particular, we specify the response support, the conditional expectation, constituting the predictor for θ_d, the conditional variance, allowed parametrizations, and the out-of-sample predictor (denoted with θ_d^{oos}). From now on, $\boldsymbol{\eta}$ indicates the vector of all the model parameters.

3.2.1.1 The Beta Model

Let us consider the mean-precision parametrization of the Beta random variable (Ferrari and Cribari-Neto, 2004): in this case, if $Y \sim \text{Beta}(\mu\phi, (1-\mu)\phi)$, then its probability density function is

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

Model	Support	$\theta_d = \mathbb{E}[Y_d\|\boldsymbol{\eta}]$	$\mathbb{V}[Y_d\|\boldsymbol{\eta}]$	Admitted `type_disp`	Predictor for θ_d^{oos}
`"beta"`	$(0;1)$	μ_d	$\frac{\mu_d(1-\mu_d)}{\phi_d+1}$	Both	$\text{logit}^{-1}(\boldsymbol{x}_d^T\boldsymbol{\beta}+e_d)$
`"flexbeta"`	$(0;1)$	$p\lambda_{1d}+(1-p)\lambda_{2d}$	$\frac{\theta_d(1-\theta_d)+p(1-p)\phi_d(\lambda_{1d}-\lambda_{2d})^2}{\phi_d+1}$	`"var"`	-
`"Infbeta0"`	$[0;1)$	$(1-p_d^z)\mu_d$	$(1-p_d^0)\left[\frac{\mu_d(1-\mu_d)}{\phi_d+1}+p_d^0\mu_d^2\right]$	`"neff"`	$(1-p_d^0)\text{logit}^{-1}(\boldsymbol{x}_d^T\boldsymbol{\beta}+e_d)$
`"Infbeta1"`	$(0;1]$	$p_d^o+(1-p_d^o)\mu_d$	$(1-p_d^1)\left[\frac{\mu_d(1-\mu_d)}{\phi_d+1}+p_d^1(1-\mu_d)^2\right]$	`"neff"`	$p_d^1+(1-p_d^1)\text{logit}^{-1}(\boldsymbol{x}_d^T\boldsymbol{\beta}+e_d)$
`"Infbeta01"`	$[0;1]$	$p_d^o+(1-p_d^z-p_d^o)\mu_d$	$p_d^o(1-\zeta_d)+(1-\alpha_d)\times$ $\times\left[\frac{\mu_d(1-\mu_d)}{\phi_d+1}+\alpha_d(\zeta_d-\mu_d)^2\right]$	`"neff"`	$p_d^1+(1-p_d^0-p_d^1)\times$ $\times\text{logit}^{-1}(\boldsymbol{x}_d^T\boldsymbol{\beta}+e_d)$

TABLE 3.1: Relevant quantities for each model implemented in **tipsae**.

where $\mu \in (0,1)$ is the location parameter and $\phi \in (0,+\infty)$ is the dispersion one. In SAE context, the Beta regression area-level model is usually specified as

$$\begin{aligned} Y_d|\mu_d,\phi_d &\overset{ind}{\sim} \text{Beta}\left(\mu_d\phi_d,(1-\mu_d)\phi_d\right), \\ \text{logit}\left(\mu_d\right) &= \mathbf{x}_d^T\boldsymbol{\beta}+e_d, \quad d=1,\dots,D; \end{aligned}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and ϕ_d is the area specific dispersion parameter, usually assumed to be known to guarantee identifiability. Recalling the expression of $\mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]$ from Table 3.1, it can be shown that, when the target response is a proportion, the parameter ϕ_d is related to the effective sample size, i.e., the corresponding sample size under simple random sampling (Janicki, 2020). For a more complete explanation of those aspects, we refer to the discussion in Section 3.2.3. On the other hand, if a generic indicator (e.g., Gini index) is considered, the meaning of ϕ_d becomes less clear. For this reason, we let the user specify the model parametrization (argument `type_disp`), choosing between:

- `"neff"` option, namely an estimate of the effective sample size ϕ_d+1 is provided;
- `"var"` option, in which an estimate of the sampling variance of the direct estimator i.e., $\widehat{\mathbb{V}}[Y_d]$, is used. In this case, the parameters ϕ_d are retrieved using the relations in Table 3.1, replacing $\mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]$ with $\widehat{\mathbb{V}}[Y_d]$, and substantially changing model parameterization.

3.2.1.2 The Flexible Beta Model

When the distribution of the response is characterized by heavy tails and/or high skewness, the standard Beta regression could fail in properly modelling Y_d (Bayes *et al.*, 2012; Migliorati *et al.*, 2018). To improve the model performances in these conditions, the standard Beta distribution can be replaced by the Flexible Beta distribution. The Flexible Beta small area model has been proposed in Chapter 2. It is defined as a mixture of two Beta random variables having a common dispersion parameter ϕ_d:

$$\begin{aligned} Y_d|\lambda_{1d},\lambda_{2d},\phi_d,p &\overset{ind}{\sim} p\ \text{Beta}\left(\lambda_{1d}\phi_d,(1-\lambda_{1d})\phi_d\right)+ \\ &\quad +(1-p)\ \text{Beta}\left(\lambda_{2d}\phi_d,(1-\lambda_{2d})\phi_d\right), \\ \text{logit}\left(\lambda_{2d}\right) &= \mathbf{x}_d^T\boldsymbol{\beta}+e_d, \qquad d=1,\dots,D. \end{aligned}$$

In this case, only the direct estimator variance (i.e., `disp_type = "var"`) can be used as input to determine the dispersion parameter of the model. Therefore, ϕ_d is expressed

as a function of the sampling variances and other model parameters (see the expression of $\mathbb{V}[Y_d|\boldsymbol{\eta}]$ in Table 3.1). The Flexible Beta distribution is characterized by four parameters: this enhances the model flexibility, if compared to the standard Beta distribution, leading to better performances in modelling not well-behaved measures and, consequently, reducing the bias of model-based estimators.

3.2.1.3 The Zero-One Inflated Beta Model

The supports of Beta and Flexible Beta models do not include the extremes 0 and 1. However, in some applications, zero and one values are observed, and a model able to encompass them is required. Therefore, following Wieczorek *et al.* (2012), we include in the package the ZOIB model, specified as:

$$
\begin{aligned}
Y_d|\mu_d, \phi_d, p_d^z, p_d^o &\overset{ind}{\sim} p_d^z \mathbb{1}\{Y_d = 0\} + p_d^o \mathbb{1}\{Y_d = 1\} + \\
&\quad + (1 - p_d^z - p_d^o)\text{Beta}\left(\mu_d\phi_d, (1-\mu_d)\phi_d\right)\mathbb{1}\{0 < Y_d < 1\} \\
\text{logit}\left(p_d^z\right) &= \mathbf{x}_d^T\boldsymbol{\beta}_p^z, \quad \text{logit}\left(p_d^o\right) = \mathbf{x}_d^T\boldsymbol{\beta}_p^o, \\
\text{logit}\left(\mu_d\right) &= \mathbf{x}_d^T\boldsymbol{\beta} + e_d, \quad d = 1, \ldots, D;
\end{aligned}
$$

where p_d^z and p_d^o denote the probabilities of observing zero and one values, respectively. They are modelled by means of a logit regression model having coefficients $\boldsymbol{\beta}_p^z$ and $\boldsymbol{\beta}_p^o$. The notation $\mathbb{1}\{A\}$ defines the indicator function that assumes value 1 if the event A is observed, and 0 otherwise. The user can specify a model that accounts both for zeroes and ones setting `likelihood = "Infbeta01"`; however, simpler versions inflating only the ones or the zeroes are also available (`"Infbeta1"` and `"Infbeta0"`, respectively). Relevant quantities for each version of the ZOIB model are listed in Table 3.1, having defined $\alpha_d = p_d^z + p_d^o$ and $\zeta_d = p_d^o/\alpha_d$. For further details, see Ospina and Ferrari (2010).

3.2.2 Prior distributions

To facilitate practitioners, standard wide-range prior distributions are assumed for the parameters included in the model. Starting from the priors for the regression coefficients, we decided to follow the default prior specification strategy of the popular **rstanarm** package (Goodrich *et al.*, 2020). Firstly, auxiliary variables are standardized in order to avoid issues related to possibly different magnitudes. Thus, posterior results for the regression coefficients must be interpreted accordingly. A weakly informative

prior for the intercept β_0 is specified:

$$\beta_0 \sim \mathcal{N}(0, 2.5^2),$$

and independent normal priors are also assigned to the coefficients related to standardized covariates:

$$\beta_j \overset{ind}{\sim} \mathcal{N}(0, 2.5^2),\ j = 1, \ldots, p.$$

Note that the same prior setting is also assumed for coefficients $\boldsymbol{\beta}_p^z$ and $\boldsymbol{\beta}_p^o$ involved in ZOIB models.

As regards the Flexible Beta model, we additionally specify the following priors for the mixing probability p and the differences between the means of mixture components:

$$\begin{aligned} & p \sim \text{Beta}(2, 2), \\ & \lambda_{1d} - \lambda_{2d} | p, \lambda_{2d} \sim \text{Unif}\left(0, \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}(Y_d|\boldsymbol{\eta})}{p(1-p)}}\right\}\right), \end{aligned}$$

following equations (2.18) and (2.23).

The priors for the random effects are discussed in the following: the case of unstructured random effects is faced in Section 3.2.2.1, spatially structured random effects are described in Section 3.2.2.2, and temporal random effects in Section 3.2.2.3.

3.2.2.1 Unstructured Random Effects

The basic assumption on the random effect is $e_d = v_d$, where v_d is an unstructured area-specific random effect accounting for deviations from the synthetic predictor. We propose three different strategies to specify its prior distribution, that can be chosen through the `prior_reff` argument of `fit_sae()`. Firstly, a zero-mean normal prior with scale σ_v is considered (`"normal"` option, default), putting a half-normal prior for σ_v, in line with Gelman (2006):

$$\begin{aligned} & v_d | \sigma_v \overset{ind}{\sim} \mathcal{N}\left(0, \sigma_v^2\right),\ d = 1, \ldots, D; \\ & \sigma_v \sim \text{Half-}\mathcal{N}(0, 2.5^2). \end{aligned}$$

The choice of such half-normal prior is usually weakly informative if compared to the scale of the random effects.

When covariates have poor explanatory power, in some domains, it is possible to

observe large deviations of the predicted value from the observed one, requiring more flexible handling of random effect through a robust prior. Among those proposed in the literature, we implement the one introduced by Figueroa-Zúñiga *et al.* (2013), and previously considered in the small area framework by Fabrizi *et al.* (2016). It consists of a Student's t prior with exponential hyperprior for degrees of freedom ν and half-normal hyperprior for the scale σ_v (`"t"` option):

$$\begin{aligned} v_d|\nu,\sigma_v &\overset{ind}{\sim} t\left(\nu,0,\sigma_v\right),\ d=1,\dots,D;\\ \nu &\sim \text{Exponential}(0.1);\\ \sigma_v &\sim \text{Half-}\mathcal{N}(0,2.5^2). \end{aligned}$$

The notation $t\left(\nu,0,\sigma_v\right)$ indicates a Student's t distribution with ν degrees of freedom, location parameter equal to 0, and scale σ_v.

In other cases, the variability of the small area parameters may not require the inclusion of a random effect term in presence of very informative covariates (Datta *et al.*, 2011b). Therefore, the variance gamma shrinkage prior introduced by Brown and Griffin (2010) and implemented in a small area application by Fabrizi *et al.* (2018) is included as a prior choice for v_d (`"VG"` option). This option enables for shrinking to 0 the random effects related to a subset of the areas by mimicking the behaviour of a spike-and-slab prior. Following Fabrizi *et al.* (2018), we propose a general hyperparameters choice that induces a prior variance of the random effects equal to 0.5:

$$\begin{aligned} v_d|\psi_d,\lambda &\overset{ind}{\sim} \mathcal{N}\left(0,\frac{\psi_d}{\lambda}\right),\ d=1,\dots,D;\\ \psi_d &\overset{ind}{\sim} \text{Gamma}(0.5,1),\ d=1,\dots,D;\\ \lambda &\sim \text{Gamma}(2,1). \end{aligned}$$

It can be noted that the independent ψ_d are local scales, whereas λ is a global precision hyperparameter.

3.2.2.2 Spatially Structured Random Effects

Setting the argument `spatial_error` equal to `TRUE`, we let the user add a spatially structured effect s_d to the linear predictor, leading to the formulation $e_d = v_d + s_d$. For the vector $\boldsymbol{s} = (s_1,\dots,s_D)$, we assume an intrinsic conditional autoregressive (ICAR) prior (Besag *et al.*, 1991), i.e., an improper prior proportional to:

$$\boldsymbol{s}|\sigma_s \propto \exp\left\{-\frac{1}{2\sigma_s^2}\boldsymbol{s}^T\tilde{\mathbf{K}}_s^-\boldsymbol{s}\right\},$$

where $\tilde{\mathbf{K}}^{-}$ is the generalized inverse of a singular precision matrix. To describe its structure, we first define $\mathbf{K} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix containing the number of connections for each area and $\mathbf{W}$ is the adjacency matrix (the generic entry $[w]_{ij}$ is 1 if area i and j are adjacent and 0 otherwise). Following Freni-Sterrantino *et al.* (2018), the actual precision matrix $\tilde{\mathbf{K}}$ is obtained with a scaling procedure aimed at reducing the impact of the structure on the prior variability, keeping into consideration the possible presence of $G \geq 1$ disconnected graphs in the model (e.g., islands). Note that $G - 1$ dummy variables are added to the linear predictor in order to obtain island-specific means, placing a sum-to-zero constraint on the random effects related to the same island. Islands defined by singleton areas are also allowed, even if they do not constitute a graph counted in G. Lastly, a half-normal prior is fixed for the hyperparameter σ_s. For further details on the implementation of ICAR priors in `Stan`, see Morris *et al.* (2019).

To include a spatially structured random effect, an object of class `SpatialPolygonsDataFrame` (from the **sp** package, Bivand *et al.*, 2013) is required as input of the `spatial_df` argument, carefully checking that the order of its rows and the order of the `data` input are coherent.

3.2.2.3 Temporally Structured Random Effects

If multiple observations of the target indicator are available for different time periods, a suitable model can be specified, in order to borrow strength from time repetitions. In this framework, a second subscript must be added in the notation: Y_{dt} indicates the direct estimator for area d at time $t = 1, \ldots, T$, whereas e_{dt} is the random effect component in the linear predictor. The user can choose to add a temporal random effect u_{dt} to the unstructured one ($e_{dt} = v_d + u_{dt}$) setting `spatial_error = TRUE`. If both temporal and spatial random effects are declared in `fit_sae()`, then a spatio-temporal model is fitted, removing the unstructured random effect ($e_{dt} = s_d + u_{dt}$).

As prior for the sequence of random effects $\{u_{dt}\}_t$, we specify a random walk prior of order 1, assuming independence among the areas (Rao and Molina, 2015). It represents a flexible prior that can be defined recursively as:

$$u_{dt}|u_{d,t-1}, \sigma_u \sim \mathcal{N}\left(u_{d,t-1}, \sigma_u^2\right), \ t = 2, \ldots, T;$$

implicitly assuming a uniform improper prior on u_{d1}. Sum-to-zero constraints are placed for each area-specific time sequences, to guarantee the identifiability of all the parameters in the linear predictor. Even then, a half-normal prior is fixed for the hyperparameter σ_u and the contribution of the correlation structure to the prior variability is mitigated

by adopting a scaling procedure (Riebler *et al.*, 2016).

3.2.3 Data Pre-Processing

Before stepping into estimation, we propose an elective function for refining raw variance estimates, which are inputs of our models. It can be useful both for reducing their sampling error and estimating the effective sample size parameter $\phi_d + 1$. The `smoothing()` function implements three methods, all yielding refined estimates of either variance or ϕ_d+1, to account for indicators with different variance functions. The output estimates are ready to be used as known parameters in an area-level model, and they need to be added to the analysed `data.frame`.

Let us consider that, under simple random sampling, a general variance function has the following structure:

$$\mathbb{V}_{\text{srs}}\left[Y_d\right] = \frac{f(\theta_d)}{n_d},$$

where n_d is the sample size. Note that if the target quantity is a proportion, then $f(\theta_d) = \theta_d(1-\theta_d)$. However, when dealing with complex survey designs, the selection process invariably introduces a correlation structure in the data. In this way, the information actually available may be lower than the one provided by a sample of the same size under simple random sampling. In order to formalize this concept, we need to introduce the effective sample size $\tilde{n}_d$. It can be estimated as $\tilde{n}_d = n_d/\text{deff}$, where deff is the design effect, defined as the ratio between the complex design-based variance $\mathbb{V}_{\text{cd}}\left[Y_d\right]$ and $\mathbb{V}_{\text{srs}}\left[Y_d\right]$. Clearly, under simple random sampling $\tilde{n}_d$ equals n_d.

All the three implemented methods enable the estimation of the effective sample sizes, whereas `"ols"` and `"gls"` also perform a variance smoothing procedure. The argument `method` allows to choose among:

- `"kish"`, implementing an area-specific design effect estimation proposed by Kish (1992). It employs solely the design weights and requires an additional data frame as input of the `survey_data` argument, whose structure is specified in Subsection 3.4.3. The specific design effect is estimated as:

$$\text{deff}_d = n_d \cdot \sum_{h \in d} \frac{W_{dh}^2}{n_{dh}}$$

  where h refers to a generic sampling unit in area d (e.g., the household). Indicating with subscript c the generic individual in sampling unit h, we define

$W_{dh} = \hat{N}_{dh}/\hat{N}_d$, $\hat{N}_{dh} = \sum_{c \in h} w_{dhc}$, $\hat{N}_d = \sum_{h \in d} w_{dh}$ and $n_d = \sum_{h \in d} n_{dh}$. We denote with w_{dh} and n_{dh} the design weight and the sample size of unit h in area d, respectively; while w_{dhc} is the individual design weight. Thus, the design-based variance can be defined as

$$\mathbb{V}_{cd}\left[Y_d\right] = \frac{f(\theta_d)}{n_d}\text{deff}_d, \tag{3.1}$$

while $\phi_d + 1 = \tilde{n}_d = n_d/\text{deff}_d$. This method has already been used in small area context by Wieczorek and Hawala (2011) and Liu *et al.* (2007). Kalton *et al.* (2005) found this approximation accurate for proportions ranging between 0.2 and 0.8.

- `"ols"`, implementing a variance smoothing model using a Generalized Variance Function approach, as in Fabrizi *et al.* (2011) and Fabrizi and Trivisano (2016). Considering the design-based variance as

$$\mathbb{V}_{cd}\left[Y_d\right] = \frac{f(\theta_d)}{n_d}\text{deff},$$

the smoothing procedure is based on the assumption that the design effect does not vary across areas. By assuming $\hat{\mathbb{V}}_{raw}[Y_d]$ as a raw estimator of complex survey variance with large error, let us specify the following smoothing equation:

$$\frac{f(Y_d)}{\hat{\mathbb{V}}_{raw}[Y_d]} = \psi n_d + \epsilon_d,$$

where $\psi = 1/\text{deff}$ and ϵ_d are zero-mean and homoscedastic residuals. The model is estimated using ordinary least squares via the `gls()` function from **nlme** package (Pinheiro *et al.*, 2021), providing the smoothed dispersion parameters defined as $\hat{\phi}_d + 1 = \hat{\psi} n_d$ and the refined estimate as $\hat{\mathbb{V}}[Y_d] = f(y_d)/(\hat{\phi}_d + 1)$.

- `"gls"`, extending the `"ols"` method in case of heteroskedasticity of the error component ϵ_d of equation (3.1). The default method assumes only heteroskedastic error with a power variance function on absolute fitted values (see Pinheiro *et al.*, 2021, for further details).

3.2.4 Posterior Inference

We are interested in making posterior inference on θ_d. Since we are not dealing with conjugate models, not even conditionally, the posterior inference is carried out through MCMC draws. As a point estimate, the optimal Bayes estimator of θ_d under quadratic

loss is considered, i.e., the posterior mean. We indicate it with the notation:

$$\hat{\theta}_d^{HB} = \mathbb{E}[\theta_d|\boldsymbol{y}] \quad d = 1, \dots D. \tag{3.2}$$

The point estimates can be complemented with uncertain measures like the posterior standard deviation and credible intervals, determined by the quantiles of the posterior distribution. The generic method `summary()` applied on as `S3` object of class `fitsae` produces by default point estimates (posterior mean and median) and credible intervals (at 95% and 50% levels) for predictors, basic model parameters, and random effects.

3.2.4.1 Out-of-Sample Treatment

The package provides an automatic out-of-sample prediction. This feature is available for all considered `likelihood`, except for Flexible Beta, since in this specific case, θ_d depends on its sampling variance, which is not available in case of out-of-samples.

Recalling that θ_d^{oos}, $d = \tilde{D}, \dots, D$ denotes the out-of-sample target quantity, their predictors are reported in Table 3.1. Note that they depend on e_d: when spatial and temporal dependencies are defined, s_d and u_{dt} gain information from the assumed correlation structure, whereas v_d is always drawn from a zero-mean distribution, contributing only to the posterior variability of θ_d^{oos}. Exploiting the MCMC estimation framework, it is possible to obtain a sample from the posterior of θ_d^{oos} by combining the samples drawn from the posterior of the involved parameters. Eventually, the point predictor defined in (3.2) holds also for out-of-sample observations, together with the other posterior summaries.

3.2.4.2 Diagnostics and Goodness-of-fit Tools

The method `summary()` returns, in addition, goodness-of-fit and model validation diagnostics, as well as SAE-specific diagnostics. In the following, we provide a brief theoretical overview of such measures.

One of the main advantages of estimating models within the Bayesian framework is the plethora of tools that allow investigating model performances. Among the most relevant ones, we can find those relying on the posterior predictive distribution, that we denote with $Y_d^{\bullet}|\boldsymbol{y},\ d = 1, \dots, D$. Area-specific Bayesian P-values (BP_d) under the following discrepancy measure (You and Rao, 2002; Fabrizi *et al.*, 2011) are computed:

$$BP_d = \mathbb{P}\left[Y_d^{\bullet} > y_d|\boldsymbol{y}\right], \quad d = 1, \dots, D. \tag{3.3}$$

In absence of systematic deviations, the expected Bayesian p-value is 0.5, whereas values near 0 or 1 highlight issues of over-estimation and under-estimation, respectively.

Information criteria are widely used in Bayesian inference to compare models with different specifications, e.g., diverse distributional assumptions, random effects structures, or covariates. Following Vehtari *et al.* (2017), we consider the approximate leave-one-out cross-validation information criterion (LOOIC) computed using Pareto-smoothed importance sampling. It can be retrieved through the **loo** package and is provided together with the approximate standard errors for estimated predictive errors.

Stepping into SAE-specific diagnostics, the standard deviation reduction (SDR_d) indicator is commonly used to assess the decrease of uncertainty associated with the employment of a small area model. It is obtained evaluating

$$SDR_d = 1 - \sqrt{\frac{\mathbb{V}\left[\theta_d|\boldsymbol{y}\right]}{\mathbb{E}[\mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]|\boldsymbol{y}]}}, \quad d = 1, \ldots, D, \tag{3.4}$$

where the denominator is defined in this way when `type_disp = "neff"`, taking into account the fact that $\mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]$ has a posterior distribution to be summarized. Conversely, if `type_disp = "var"`, the denominator is replaced by $\hat{\mathbb{V}}\left[Y_d\right]$. This diagnostic has to be considered with caution when performing model selection since it does not account for the design bias of different model-based estimators, which could be relevant even when the model is correct.

A measure of distance between direct and model-based estimators may be useful to detect when model and data evidence differ significantly and for models comparison purposes. We adopted a normalized Euclidean distance weighted for the standard deviations of direct estimators, similar to the one mentioned by Morales *et al.* (2015):

$$\left(\frac{1}{\tilde{D}}\sum_{d=1}^{\tilde{D}}\frac{(y_d - \hat{\theta}_d^{HB})^2}{\mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]^{1/2}}\right)^{1/2}.$$

On the same wave, a Confidence Interval Rate (CIR) can be set up, by constructing the confidence intervals for direct estimators, and by counting the times model-based estimators fall within such intervals:

$$CIR = \frac{1}{\tilde{D}}\sum_{d=1}^{\tilde{D}}\mathbb{1}\{\hat{\theta}_d^{HB} \in (y_d - 1.96 \times \mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]^{1/2}, y_d + 1.96 \times \mathbb{V}\left[Y_d|\boldsymbol{\eta}\right]^{1/2})\},$$

using a Gaussian approximation and a confidence interval level of 95%.

Lastly, the Shrinking Bound Rate (SBR) is computed:

$$SBR = \frac{1}{\tilde{D}} \sum_{d=1}^{\tilde{D}} \mathbb{1}\{\hat{\theta}_d^{HB} \in (p_d^*, y_d)\}, \tag{3.5}$$

where $p_d^* = \exp(\mathbf{x}_d^T \boldsymbol{\beta}) / \left[1 + \exp(\mathbf{x}_d^T \boldsymbol{\beta})\right]$ is the synthetic estimate of θ_d. In fact, in the standard Fay-Herriot model, the shrinking process is clearly identified by the shape of the best linear unbiased predictor, for known values of β and σ_v^2 such as,

$$\gamma_d y_d + (1 - \gamma_d) p_d^* \quad \text{with} \quad \gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \mathbb{V}[Y_d|\boldsymbol{\eta}]}.$$

Beta regression models do not provide a closed form predictor, since the conditional distribution of θ_d, $\forall d = 1, \ldots, D$ does not belong to a standard family. Janicki (2020) shows that, in a Beta regression model with standard diffuse priors, $\hat{\theta}_d^{HB}$ converges to the direct estimate as $\mathbb{V}[Y_d|\eta] \to 0$ and the synthetic estimates as $\sigma_v^2 \to 0$. The first property has also been proved by Fabrizi *et al.* (2020). However, $\hat{\theta}_d^{HB}$ is not bounded by its convergence limits, conjecturing $Y_d < \hat{\theta}_d^{HB} < p_d^*$ will hold only for $\mathbb{V}[Y_d|\boldsymbol{\eta}]$ sufficiently small (Janicki, 2020). Thus, checking whether model estimates fit inside the bound, could yield important insights into the shrinking process and estimators consistency.

3.2.4.3 Benchmarking Procedure

The `benchmark()` function gives the chance to perform a benchmarking procedure on model-based estimates. The need for benchmarking arises since model-based estimates may widely differ from direct estimates and, consequently, model estimates aggregates may widely differ from corresponding direct estimates. However, latter quantities refer to a larger geographical area or a larger socio-demographic group whose target domains are a subset of, and, therefore, are considered to be reliable. This feature may introduce drawbacks in many situations (e.g., when small area estimates are used to allocate funding), and exact benchmarking is required to avoid surpluses or shortfalls (Zhang and Bryant, 2020). When adopting a benchmarking approach, model-based estimates are constrained to direct estimates of supra-domain sets.

Existing methods generally address the benchmark issue as a constraint to be imposed. The difference in between various methods is about the way such constraints are interpreted and incorporated in the estimation. Some methods estimate the small area models and then modify the resulting point estimators to satisfy the benchmarking constraints as a two-step procedure (Datta *et al.*, 2011a). Other methods treat benchmarks as vincula on the underlying small area parameters or on their point estimators which

are directly incorporated into the probabilistic structure of the small area model, either in a revised likelihood or in the prior distributions (Pfeffermann *et al.*, 2014; Ranalli *et al.*, 2018; Zhang and Bryant, 2020). By considering only Bayesian benchmarking methods, when models yield a full posterior distribution for all unknown quantities after benchmarking, they end up being categorized as fully Bayesian. While methods such as the one we adopted, which derive posterior distributions without benchmarking and separately benchmark point estimates, do not fall within this definition. An up-to-date review can be found in Zhang and Bryant (2020).

In our non-fully Bayesian approach, widely explained by Datta *et al.* (2011a), point estimates from a Bayesian model, estimated via the `fit_sae()` function, are adjusted to obtain a new set of estimates that satisfies the constraints. Benchmarking could solely target the point estimators (single benchmarking) or, alternatively, also the ensemble variability (double benchmarking). Furthermore, an estimate of the overall posterior risk is provided, aggregated for all areas. This value is only yielded when in-sample areas are treated and a single benchmarking is performed.

The considered benchmarking procedures require the definition of a set of area-specific weights, which in the case of proportions are defined as $w_d = N_d / \sum_{j=1}^{D} N_j$, where N_d is the population size for area d. The benchmark is indicated with B, and it could be the reliable direct estimate referring to a larger area or a prespecified value from another data source or, eventually, $B = \sum_{d=1}^{D} w_d Y_d$, if the aim is to perform internal benchmarking. The function allows performing three different benchmarking methods, according to the argument `method`.

- The `"ratio"` method provides benchmarked estimates $\hat{\theta}_d^{BM}$ that minimize the posterior expectation of the weighted squared error loss. The benchmarked estimates are

$$\hat{\theta}_d^{BM} = \hat{\theta}_d^{HB} + \frac{B - \sum_d w_d \hat{\theta}_d^{HB}}{s} r_d, \tag{3.6}$$

  where $r_d = \hat{\theta}_d^{HB}$, and $s = \sum_d w_d \hat{\theta}_d^{HB}$. Datta *et al.* (2011a) provide also the posterior risk for the whole set of benchmarked estimates:

$$\sum_d \frac{w_d}{r_d} \left[\mathbb{V}[\theta_d | \boldsymbol{y}] + \frac{\left(B - \sum_d w_d Y_d\right)^2}{s^2} r_d^2 \right]. \tag{3.7}$$

- The `"raking"` method provides the benchmarked estimate in (3.6) and the posterior risk (3.7) with $r_d = 1$ and $s = 1$.

- The `"double"` method extends this procedure accounting for a further benchmark on the weighted ensemble variability. The simultaneous constraints are $\sum_d w_d \hat{\theta}_d^{BM} = B$ and $\sum_d w_d (\hat{\theta}_d^{BM} - B)^2 = H$, where H is a prespecified value of the estimators variability taken from other sources. The expression of the resulting benchmarked estimate is:

$$\hat{\theta}_d^{BM} = B + \sqrt{\frac{H}{\sum_d w_d \left(\hat{\theta}_d^{HB} - \sum_d w_d \hat{\theta}_d^{HB}\right)^2}} \left(\hat{\theta}_d^{HB} - \sum_d w_d \hat{\theta}_d^{HB}\right).$$

Note that the benchmarking procedure can be performed in case of temporal or spatio-temporal models by specifying multiple time-period benchmarks.

3.3 Datasets

In SAE field, data typically come from multiple sources. Direct estimators and their sampling variances typically result from survey data, aggregated at area-level, while covariates come from census and/or administrative/register sources. As a consequence, explanatory variables, aggregated at area level, are required to be defined at population level i.e., without error, and potentially correlated with the target variable. In order to outline the workflow of **tipsae** package, its functions are illustrated in Section 3.4 and applied to an example dataset, released within the package. The whole dataset is named `emilia` and consists of a panel on poverty mapping concerning 38 health districts within the Emilia-Romagna region, located in North-East of Italy, with annual observations recorded from 2014 to 2018. We built it starting from model-based estimates and related CV freely available on Emilia-Romagna region website [1]. Since it is used for illustrative purposes only, such estimates are assumed to be unreliable direct estimates, requiring a SAE procedure.

We considered the Head-Count Ratio estimates as direct (`$hcr`) and its associated variance as sampling variance (`$vars`). A fake standardized covariate `$x` has been generated. We also provide area sample sizes (`$n`), population sizes (`$pop`), province identification (`$prov`), years (`$year`) and health district name (`$id`). The `emilia` dataset can be loaded as follows.

```
R> library(tipsae)
R> data("emilia")
R> head(emilia)
```

[1] https://statistica.regione.emilia-romagna.it/documentazione/pubblicazioni/documenti_catalogati/stima-poverta-2009-2018-distretti-sociosanitari-province-emilia-romagna

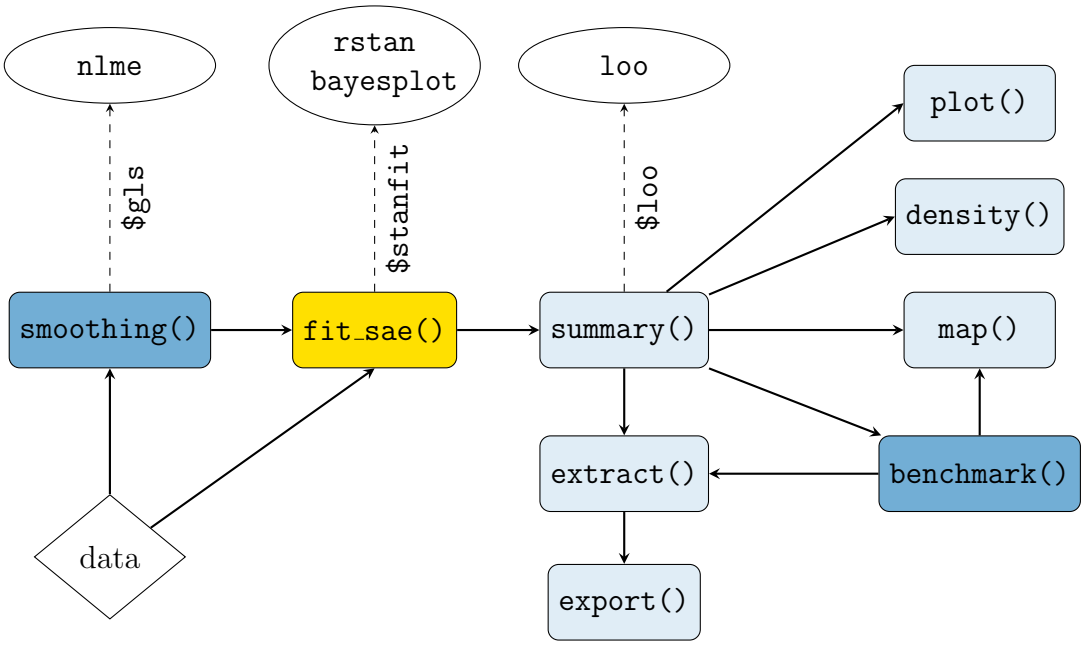

FIGURE 3.1: Flowchart that describe the structure of the tools implemented in the **tipsae** package.

```
                    id prov year    hcr         vars   n       x    pop
1 CASALECCHIO DI RENO   BO 2014 0.0404 9.090478e-05  42 -0.2624 108261
2   CITTA' DI BOLOGNA   BO 2014 0.0825 6.404001e-05 285 -0.0008 371151
3               IMOLA   BO 2014 0.1033 3.120275e-04  49 -0.0522 130007
4         PIANURA EST   BO 2014 0.0633 1.025764e-04 190 -0.4007 154213
5       PIANURA OVEST   BO 2014 0.0625 1.562500e-04  10 -0.2277  80951
6      PORRETTA TERME   BO 2014 0.1276 6.643609e-04  26 -0.4434  56428
```

A cross-sectional subset concerning a single year (2016) is taken from `emilia`, for non-temporal models illustration purpose: it is named `emilia_cs` and can be loaded as follows.

```
R> data("emilia_cs")
```

3.4 Workflow

In this section, a typical flow of a SAE analysis is outlined with step-by-step instructions, showing the potentials of **tipsae** tools. As illustrated with a flowchart in Figure 3.1, the package is structured into three parts that relate to: model building and fitting (•, Section 3.4.1), diagnostics and results displaying (•, Section 3.4.2), and complementary tools for SAE analysis (•, Section 3.4.3). Figure 3.1 displays also the possible connections with external functions, drawn with dashed arrows, useful to further exploit the produced objects.

3.4.1 Model Building and Fitting

The first step of the workflow represents the core of our package, concerning the estimation of models with the diverse extensions and parametrizations defined in Section 3.2. The sole function `fit_sae()` allows users to construct personalized models and fit them using `Stan` routines, called up through the `sampling` function of **rstan** package. It also allows customized parallel computing when the model runs on multiple chains. A simple parallelization can be set out using the following command, which imposes a number of R processes equal to the number of CPU cores.

```
R> options(mc.cores=parallel::detectCores())
```

It is also possible to change the default options for parallelization using the function `setDefaultClusterOptions()` from **parallel** package. For further details, see **rstan** guidelines.

A complete list of the input arguments of the `fit_sae()` function is specified in Table 3.2, and a first example of model fitting on the `emilia_cs` dataset is provided. Firstly, we consider model default options: a Beta likelihood and a Gaussian prior for unstructured random effects. Since `emilia_cs` dataset contains the sampling variance as a measure of dispersion, `disp_direct` must be fixed equal to `"var"`, setting a mean-variance parametrization. Moreover, argument `domains_size` has to be specified for having visual design consistency diagnostics in the subsequent plotting function.

The estimation can be done in practice by running the `fit_sae()` function as follows. For the sake of reproducibility, we set `seed=0`.

```
R> fit_beta <- fit_sae(formula_fixed = hcr ~ x,
+                      data = emilia_cs,
+                      domains = "id",
+                      type_disp = "var",
+                      disp_direct = "vars",
+                      domain_size = "n",
+                      seed = 0)
```

Note that further arguments, concerning `sampling` function options, can be additionally specified. In particular, we mention those related to HMC algorithm setting such as `iter`, allowing to set the number of iterations per chain (default equal to 2000), `warmup`, determining the number of iterations per chain to be discarded as warm-up period (default `iter/2`), `chains`, fixing the number of independent Markov chains (default 4).

Different models can be estimated relying on diverse assumptions, being subsequently compared with each other. For example, we assume a Flexible Beta likelihood and a

Argument	Short description	Default
`formula_fixed`	`formula` object specifying the fixed regression part.	-
`data`	`data.frame` containing all relevant quantities.	-
`domains`	`data` column name displaying domains names. If `NULL` (default) the domains are denoted with a progressive number.	`NULL`
`type_disp`	Parametrization of the dispersion parameter. The choices are variance (`"var"`) or $\phi_d + 1$ (`"neff"`) parameter.	`"neff"`
`disp_direct`	`data` column name displaying given values of sampling dispersion for each domain.	-
`domain_size`	`data` column name indicating domain sizes (optional).	`NULL`
`likelihood`	Sampling likelihood to be used. The choices are `"beta"`, `"flexbeta"`, `"Infbeta0"`, `"Infbeta1"` and `"Infbeta01"`.	`"beta"`
`prior_reff`	Prior distribution of the unstructured random effect. The choices are: `"normal"`, `"t"`, `"VG"`.	`"normal"`
`spatial_error`	Logical indicating whether to include a spatially structured random effect.	`FALSE`
`spatial_df`	Object of class `SpatialPolygonsDataFrame` with the shapefile of the studied region. Required if `spatial_error = TRUE`.	`NULL`
`temporal_error`	Logical indicating whether to include a temporally structured random effect.	`FALSE`
`temporal_variable`	`data` column name indicating temporal variable. Required if `temporal_error = TRUE`.	`NULL`
`adapt_delta`	HMC option: target average proposal acceptance probability. See `Stan` documentation.	`0.95`
`max_treedepth`	HMC option: maximum allowed tree depth for each transition. See `Stan` documentation.	`10`
`init`	HMC option: initial values setting. The choices are: `"0"`, `"random"`, or manual setup via list or function. See `Stan` documentation.	`"0"`
`...`	Further inputs for the `sampling` function.	

TABLE 3.2: Input arguments for function `fit_sae`.

Position	Name	Short description
1	`model_settings`	List summarizing all the assumptions of the model: sampling likelihood, presence of intercept, dispersion parametrization, random effects priors and possible structures.
2	`data_obj`	List containing input objects including in-sample and out-of-sample relevant quantities.
3	`stanfit`	`stanfit` object, outcome of `sampling` function containing full posterior draws. For details, see `rstan` documentation.
4	`pars_interest`	Vector containing the names of parameters whose posterior samples are stored.
5	`call`	Image of the function call that produced the `fit_sae` object.

TABLE 3.3: Components of `fitsae` objects.

variance gamma shrinking prior for the unstructured random effect, in order to propose a more flexible model for the data. Given the increasing complexity of model assumptions, more HMC iterations are required, together with a higher proposal acceptance probability (`adapt_delta`).

```
R> fit_FB <- fit_sae(formula_fixed = hcr ~ x,
+                    data = emilia_cs,
+                    domains = "id",
+                    type_disp = "var",
+                    disp_direct = "vars",
+                    domain_size = "n",
+                    likelihood = "flexbeta",
+                    prior_reff = "VG",
+                    adapt_delta = 0.99,
+                    iter = 8000,
+                    seed = 0)

Warnings:
1: There were 10 divergent transitions after warmup. See
http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
to find out why this is a problem and how to eliminate them.
2: Examine the pairs() plot to diagnose sampling problems
```

The `fit_sae()` function returns an `S3` object of class `fitsae`, being a list of relevant items that are listed in Table 3.3. The core element is the `$stanfit` object, incorporating

posterior draws and raw MCMC information to be extracted, whereas the remaining elements only provide details about the function call and model settings.

3.4.2 Diagnostics and Results Displaying

After the MCMC drawing, a careful check on algorithm convergence is required, in order to validate posterior results. With this aim, our suggestion is to exploit the plethora of diagnostic methods implemented for `stanfit` objects within the **bayesplot** package. For example, the following code generates the trace-plots related to the `fit_beta` model, as in Figure 3.2, useful to visually inspect the convergence of the chains to a unique stationary distribution.

```
R> library(bayesplot)
R> post_beta <- as.array(fit_beta$stanfit, pars = c("beta0","beta"))
R> mcmc_trace(x = post_beta)
```

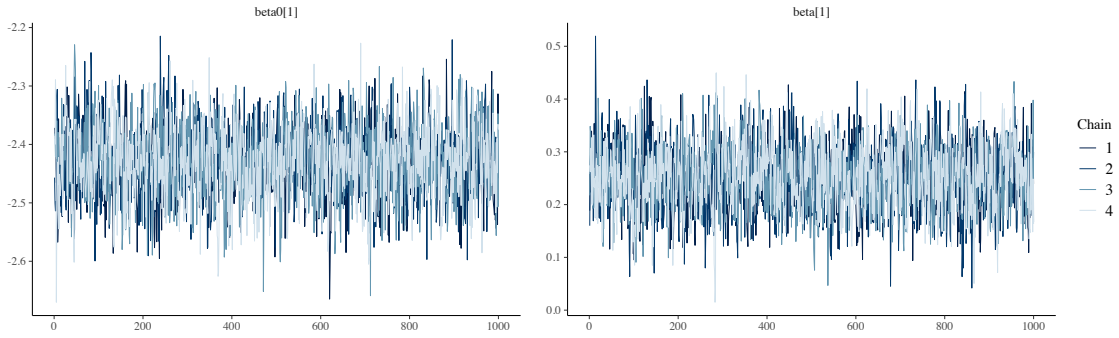

FIGURE 3.2: Traceplots of the parameters β_0 and β_1 of the Beta regression model.

The `stanfit` object also provides useful visual diagnostics to deepen the warnings printed by `Stan`, such as those about the maximum tree depth and divergent transitions after the warm-up period.

However, small area diagnostics are required at this stage, in order to check whether results meet specific properties which turn out to be desirable in such context. Peculiar diagnostic measures can be obtained through `summary()` method applied on `fitsae` objects. Besides the printed output, the method produces an object of class `summary_fitsae` which contains relevant information for posterior inference. Argument `probs` allow specifying the quantiles of interest to be visualized as posterior summary measures. The logical argument `compute_loo` allows deciding whether `loo` information criterion should be computed or not.

```
R> summ_beta <- summary(fit_beta)
Warnings:
```

```
Some Pareto k diagnostic values are too high.
See help('pareto-k-diagnostic') for details.

R> summ_beta
Summary for the SAE model call:
 fit_sae(formula_fixed = hcr ~ x, domains = "id", disp_direct = "vars",
    type_disp = "var", domain_size = "n", data = emilia_cs, seed = 0)

----- S.D. of the random effects: posterior summaries -----

         mean    sd  2.5%  25%   50%   75% 97.5%
sigma_v 0.267 0.055 0.168 0.23 0.263 0.299 0.388

----- Fixed effects coefficients: posterior summaries -----

              mean    sd   2.5%    25%    50%    75%  97.5%
(Intercept) -2.428 0.060 -2.550 -2.467 -2.428 -2.387 -2.309
x            0.253 0.061  0.135  0.213  0.253  0.293  0.372

--------------- Model diagnostics summaries ---------------

                   Min. 1st Qu. Median  Mean 3rd Qu.  Max.
Residuals        -0.016  -0.004  0.002 0.004   0.011 0.032
S.D. Reduction   -0.100   0.197  0.254 0.240   0.318 0.390
Bayesian p-value  0.172   0.339  0.459 0.461   0.555 0.785

Shrinkage Bound Rate: 100 %

LOO Information Criterion:
          Estimate    SE
 elpd_loo   87.719 3.629
 p_loo      17.787 2.546
 looic    -175.439 7.259
```

If printed, the produced summary displays:

- Posterior summaries about the fixed effect coefficients and the scale parameters related to unstructured and possible structured random effects.
- Model diagnostics summaries of (a) model residuals; (b) standard deviation reductions computed using (3.4); (c) Bayesian P-values obtained approximating the (3.3) with the MCMC samples.
- Shrinking Bound Rate, defined in (3.5).
- LOO information criteria and related diagnostics from the `loo` package.

3.4.2.1 What Can Be Accidentally Done with a `summary_fitsae` Object

The `summary_fitsae` object contains additional valuable elements for further exploration. For instance, the `$loo` element consists of the whole object of class `loo` which may be employed in external functions, such as the ones provided by **loo** package e.g. for model comparison, as follows.

```
R> summ_FB <- summary(fit_FB)
Warnings:
Some Pareto k diagnostic values are too high.
See help('pareto-k-diagnostic') for details.

R> library(loo)
R> loo_compare(list("beta" = summ_beta$loo, "flexbeta" = summ_FB$loo))
         elpd_diff se_diff
flexbeta  0.0       0.0
beta     -6.9       2.9
```

The output shows that the Flexible Beta model has a significantly higher expected log pointwise predictive density for a new dataset, gaining in prediction power with respect to the default model.

Another element that can be employed in external functions to assess model goodness of fit is `$y_rep`, an array with values generated from the posterior predictive distribution, enabling the implementation of posterior predictive checks through the **bayesplot** package. The observed data, required for the checks, can be extracted through `$direct_est` element. The following code allows comparing the empirical densities of generated samples under the considered models, reported in Figure 3.3.

```
R> library(ggplot2)
R> ppc_dens_overlay(y = summ_beta$direct_est, yrep = summ_beta$y_rep[1:100,]) +
+    ggtitle("Beta likelihood")
R> ppc_dens_overlay(y = summ_FB$direct_est, yrep = summ_FB$y_rep[1:100,]) +
+    ggtitle("Flexible Beta likelihood")
```

Lastly, all the posterior summaries related to random effects are stored in the `$raneff` element, being a list of `data.frame` objects, one for each type: `$unstructured`, `$temporal`, and `$spatial`. Such outputs may be exploited to produce meaningful plots, e.g., the caterpillar plot of Figure 3.4, created via the following code.

```
R> ggplot(summ_beta$raneff$unstructured, aes(x = reorder(Domains, mean))) +
+    geom_point(aes(y = mean)) +
+    geom_linerange(aes(ymin = `2.5%`, ymax = `97.5%`)) +
+    geom_hline(yintercept = 0, lty = 2) +
+    ylab("Random effect") + xlab("") +
+    theme_bw(base_size = 12) +
+    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

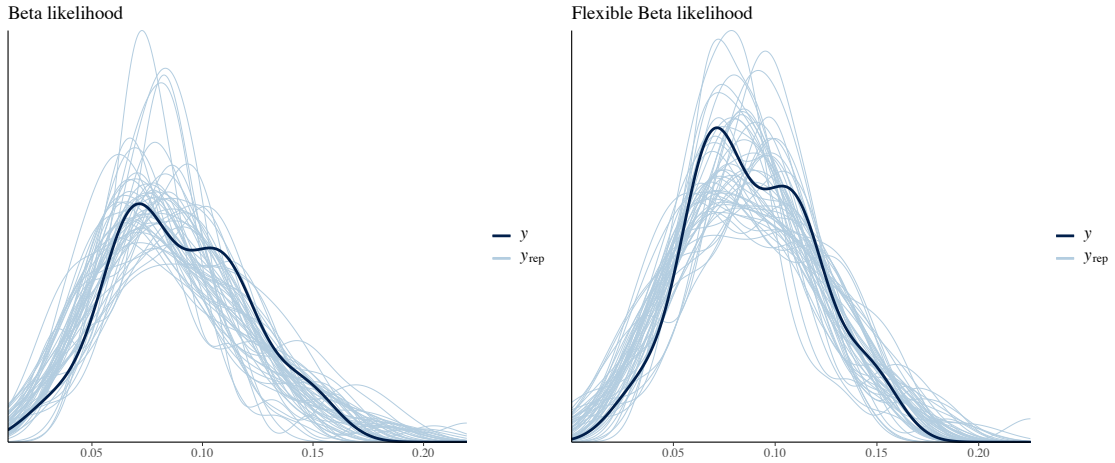

FIGURE 3.3: The empirical densities from posterior predictive samples (y_{rep}) versus the observed data one (y).

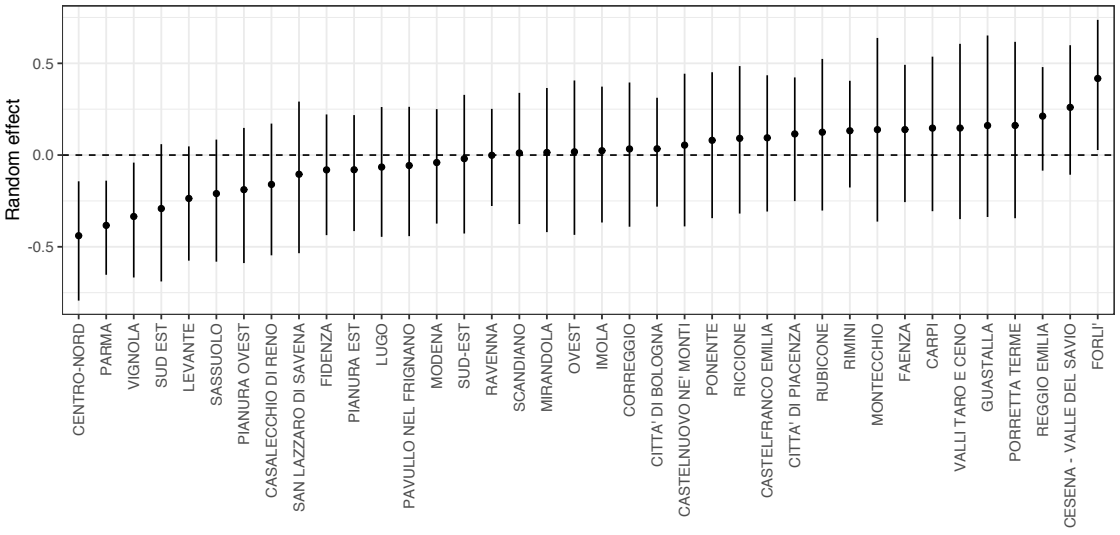

FIGURE 3.4: Caterpillar plot of unstructured random effects from Beta regression model.

3.4.2.2 Ad-Hoc Plot Functions

Our package comes equipped with ad-hoc functions for visual diagnostic tools. The `S3` object `summary_fitsae` can be used as input for `plot()` and `density()` visual methods as well as for `map()` function.

The generic method `plot()` provides, in a grid (default) or sequence, (a) a scatterplot of direct estimates versus model-based estimates, visually capturing the shrinking process, (b) a Bayesian P-values histogram, (c) a boxplot of standard deviation reduction values, and, if areas sample sizes are provided as input in `fit_sae()`, (d) a scatterplot of model residuals versus sample sizes, in order to check for design-consistency i.e., as long

as sizes increase residuals should converge to zero. The following code line produces Figure 3.5.

```
R> plot(summ_beta)
```

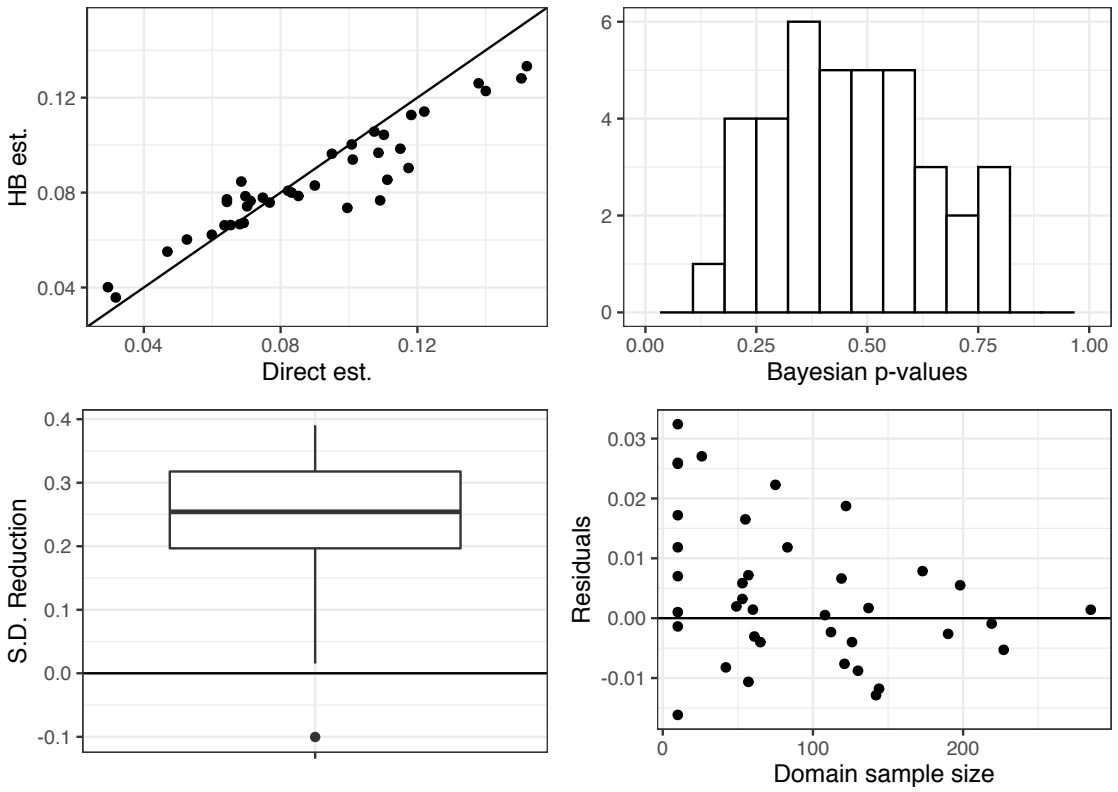

FIGURE 3.5: `plot()` method visual outcome.

The method `density()` provides, in a grid (default) or sequence, the density plot of direct estimates versus HB model estimates and the density plot of standardized posterior means of the random effects versus standard normal, in order to check for Gaussian assumption. Figure 3.6 is produced as the output of the following command.

```
R> density(summ_beta)
```

Lastly, the `map()` function enables the investigation of the analysed phenomenon by accounting for its geographical dimension, if it exists. More in detail, a `SpatialPolygonsDataFrame` object from the **sp** package should be provided as input in `spatial_df` argument. The `spatial_id_domains` argument must receive as input the name of `spatial_df` variable containing area denominations, in order to correctly match the areas. If such names match the ones provided through the original dataset, no extra arguments are required. Otherwise, the `match_names` argument should receive an encoding two-columns `data.frame`: the first with the original data coding (`domains`) and

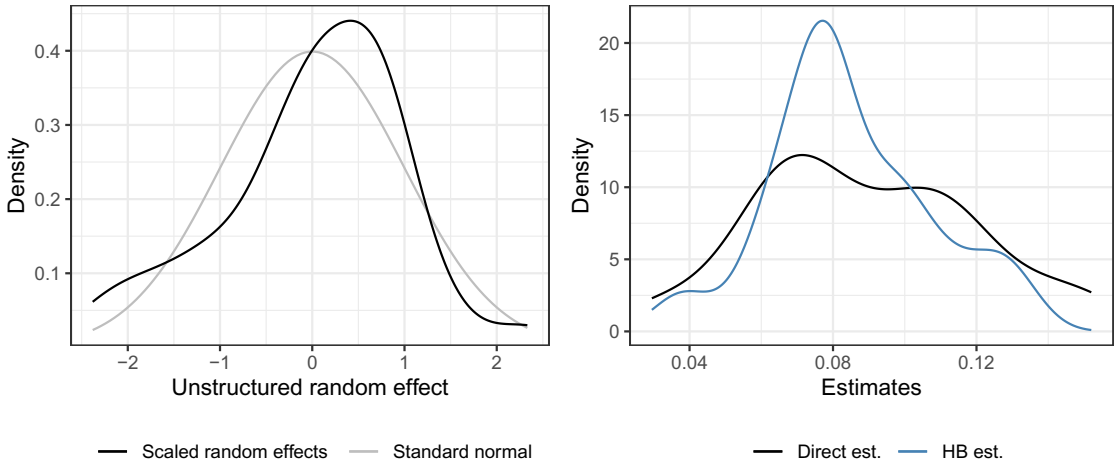

FIGURE 3.6: `density()` method visual outcome.

the second one with corresponding `spatial_df` object labels. The feature to be displayed on the map can be defined in `quantity` argument, choosing among HB model estimates `HB_est`, direct estimates `Direct_est`, posterior standard deviations `SD`, and benchmarked estimates `Bench_est` when a `benchmark_fitsae` class object is given as input (see Section 3.4.3). The following code loads the Emilia-Romagna health districts shapefile and produces the maps in Figure 3.7, with model-based estimates and their posterior standard deviations.

```
R> data("emilia_shp")
R> map(x = summ_beta,
+      spatial_df = emilia_shp,
+      spatial_id_domains = "NAME_DISTRICT")
R> map(x = summ_beta,
+      spatial_df = emilia_shp,
+      quantity = "SD",
+      spatial_id_domains = "NAME_DISTRICT")
```

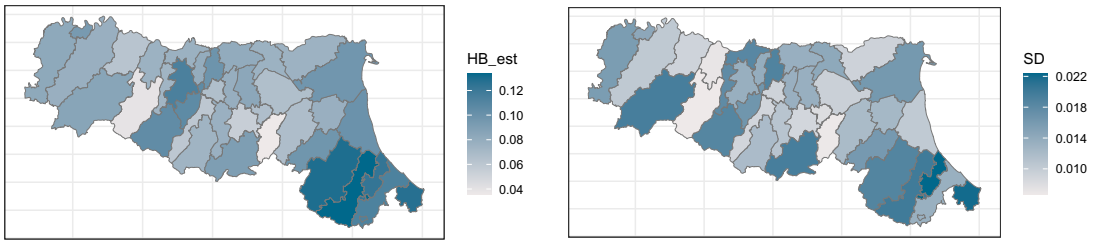

FIGURE 3.7: `map()` function visual outcome.

3.4.2.3 Take-Home Function

Lastly, `summary_fitsae` object provides target parameters posterior and model-based estimates, visually accessible through the function `extract()` as follows.

```
R> HB_estimates <- extract(summ_beta)
R> head(HB_estimates$in_sample)
               Domains Direct est.    HB est.          sd       2.5%
1 CASALECCHIO DI RENO      0.0469 0.05511808 0.009143088 0.03696825
2   CITTA' DI BOLOGNA      0.0681 0.06668932 0.008168604 0.05040972
3               IMOLA      0.0692 0.06723489 0.011631916 0.04467289
4         PIANURA EST      0.0636 0.06621838 0.009554082 0.04787149
5       PIANURA OVEST      0.0685 0.08465454 0.013695722 0.05789687
6      PORRETTA TERME      0.1174 0.09035749 0.019929132 0.05449263
         25%        50%        75%      97.5%
1 0.04897258 0.05504724 0.06121161 0.07330119
2 0.06132649 0.06695005 0.07234576 0.08205602
3 0.05913559 0.06721644 0.07501379 0.09011058
4 0.05944194 0.06623004 0.07286028 0.08465238
5 0.07552132 0.08449351 0.09385837 0.11194273
6 0.07645193 0.08868046 0.10297773 0.13293759
```

The function returns an object of class `estimates_fitsae`, being a list of two data frames, distinguishing between `$in_sample` and `$out_of_sample` areas, which gathers domains name, direct and HB estimates, as well as posterior summaries of parameters θ_d, $\forall d$.

A function for exporting such results in csv format is directly accessible, with name `export()`. This function requires an `estimate_fitsae` object and a character string naming the output file (argument `file`). It is also possible to indicate whether to export both in and out of sample areas results (default, `type="all"`), or only in or out of sample areas, (`"in"` or `"out"`, respectively), as follows.

```
R> export(HB_estimates,
+         file = "results.csv",
+         type = "all")
```

Additional arguments of `write.csv()` function from **utils** package can be further indicated.

3.4.3 Complementary Tools

Complementary tools for small-area analysis provided by the package are the smoothing and benchmarking functions. The `smoothing()` function allows for data pre-processing of sampling variance estimates and retrieving effective sample sizes, as described in

Section 3.2.3. After its usage, output results have to be incorporated in the dataset used as input of the `fit_sae` function. The `smoothing()` function requires as input the `data` including the direct estimates, whose variable name has to be specified in `direct_estimates` argument, the method to be used among `"ols"`, `"gls"` and `"kish"` (`method`), and the specification of a variance function $f(\theta)$, through `var_function` argument. The default option (`NULL`) for $f(\theta)$ match the proportion case, being equal to $\theta(1-\theta)$, while for other measures it can widely differ, for instance, the Gini index variance can be approximated to $f(\theta) = \theta^2(1-\theta^2)$ (Fabrizi and Trivisano, 2016) and therefore the following object has to be provided in `var_function` argument:

```
R> gini_variance <- function(x){ x^2 * (1 - x^2) }
```

If method `"ols"` or `"gls"` is chosen, the function requires the raw variance estimates (`raw_variance` argument), areas sample sizes (`areas_sample_sizes`), and, possibly, additional covariates (`additional_covariates`), all of them being column names of the `data.frame` provided to the `data` argument. On the other hand, method `"kish"` requires the domain names (`area_id`, as column name in `data`) and the specification of an additional dataset (`survey_data`), defined at sampling unit level (e.g., households). Such dataset must include sampling weights (`weights`), unit sizes (`sizes`) and domain names (`survey_area_id`), in order to allow for matching. The output is an object of `smoothing_fitsae` class, being a list of vectors including dispersion parameters estimates: both the variance and $\hat{\phi}_d + 1$. If `"ols"` or `"gls"` method has been selected, the list incorporates also an object of class `gls` from **nlme** package, ready to be further explored through **nlme** additional tools.

```
R> smoo <- smoothing (emilia_cs,
+                     direct_estimates = "hcr",
+                     area_id = "id",
+                     raw_variance = "vars",
+                     areas_sample_sizes = "n",
+                     var_function = NULL,
+                     method = "ols")
R> smoo
Proportions variance function specified.
Generalized least squares fit by REML
  Model: as.formula(paste0("y ~ -1", str))
  Data: regdata
       AIC      BIC    logLik
  481.1331 484.3549 -238.5666

Coefficients:
     Value Std.Error  t-value p-value
n 2.888026 0.1825709 15.81866       0
```

```
Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.5003066 -0.3865189  0.4100642  0.7766002  3.1200482

Residual standard error: 127.9598
Degrees of freedom: 38 total; 37 residual

R> emilia_cs$smoo_phi <- smoo$phi
R> emilia_cs$smoo_vars <- smoo$vars
```

The `benchmark()` function implements benchmarking procedures, described in Section 3.2.4.3, on model-based estimates provided by indicating a `summary_fitsae` object, given a vector of areas weights (`share`), in our case the population shares, a benchmark value (`bench`), and a method among `"raking"`, `"ratio"` and `"double"` (`method`). When the double benchmarking method is selected, the user must also indicate a second benchmark through the `H` argument, corresponding to the ensemble variability. The output is an object of class `benchmark_fitsae`, being a list including the vector of benchmark estimates, the posterior risk, and relevant information about the call. A `benchmark_fitsae` object may be used as input for `map()` function, in order to spatially display benchmarked estimates, `extract()` or `export()` functions. The first option is included in the following code, whose visual output is in Figure 3.8.

```
R> shares <-  emilia_cs$pop / sum(emilia_cs$pop)
R> bmk <- benchmark(summ_beta,
+                   bench = 0.13,
+                   share = shares,
+                   method = "raking")
R> map(x = bmk,
+      spatial_df = emilia_shp,
+      spatial_id_domains = "NAME_DISTRICT")
```

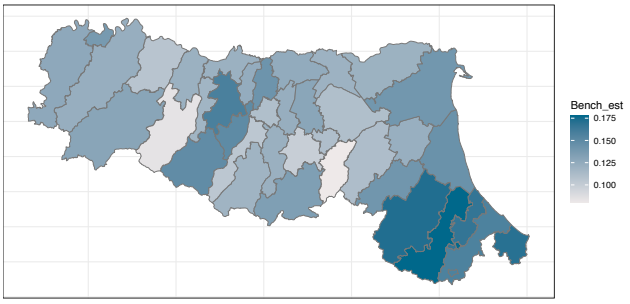

FIGURE 3.8: Benchmarked estimates plotted through `map()` function.

Benchmarking can be done on the whole set of areas (default option) or even on a subset of them. In the latter case, the vector containing the names of the considered areas has to be indicated through the `areas` argument. Moreover, the function automatically takes out-of-sample estimates if they are involved in the benchmarking procedure. Benchmark estimates and posterior risk are stored within an object of class `benchmark_fitsae` and can be accessed through `$bench_est` and `$post_risk` elements, as shown below.

```
R> subset <- c("RIMINI", "RICCIONE", "RUBICONE",
+              "CESENA - VALLE DEL SAVIO")
R> pop <- emilia_cs$pop[emilia_cs$id %in% subset]
R> shares_subset <- pop/sum(pop)
R> bmk_subset <- benchmark(summ_beta,
+                          bench = 0.13,
+                          share = shares_subset,
+                          method = "raking",
+                          areas = subset)
R> bmk_subset$bench_est
[1] 0.1416332 0.1311741 0.1344299 0.1210718
R> bmk_subset$post_risk
[1] 0.0003528578
```

For temporal models, a benchmark can be specified only for one time period at a time, indicated in the `time` argument.

3.4.4 Spatio-Temporal Examples

As explained in Section 3.2, it is possible to fit models that incorporate a spatial dependency structure, a temporal dependency structure or even both of them. The first extension, useful when the domains of interest are geographical entities, relaxes the assumption of spatial independence. Commonly, the boundaries across areas are arbitrarily set, and thus it can be reasonable to assume that the quantities of interest belonging to neighbouring areas are correlated. This can happen when dealing with data where the spatial dimension is relevant, e.g., agricultural, environmental, economic and epidemiological analyses. A spatial extension can be implemented through the `fit_sae()` function by switching to `TRUE` the `spatial_error` argument and supplying an object of class `SpatialPolygonsDataFrame` in `spatial_df` argument, being careful to include it ordered as the `data` object.

When dealing with panel data, such as the `emilia` dataset, a temporal dependency structure has to be taken into account due to the presence of repeated measures across time. It is possible to implement a temporal model by switching to `TRUE` the

`temporal_error` argument and by providing the name of the dataset temporal variable in `temporal_variable` argument.

Note that if a spatio-temporal model is required, the domain records should keep the same ordering within each recorded time in the `data` object. Hence, it is possible to re-order the shapefile accordingly, using the following commands.

```
R> data("emilia")
R> data("emilia_shp")
R> emilia_shp_ord <- emilia_shp[match(unique(emilia$id),
+                                     emilia_shp$NAME_DISTRICT),]
```

The following code allows estimating a spatio-temporal model under a Beta likelihood. In presence of structured random effects within the model, our suggestion is to increase the `max_treedepth` argument above 10, to improve the mixing of the HMC algorithm. After estimating the model, the `fitsae` object can be explored through `summary()` method.

```
R> fit_ST <- fit_sae(formula_fixed = hcr ~ x,
+                     domains = "id",
+                     disp_direct = "vars",
+                     type_disp = "var",
+                     domain_size = "n",
+                     data = emilia,
+                     spatial_error = TRUE,
+                     spatial_df = emilia_shp_ord,
+                     temporal_error = TRUE,
+                     temporal_variable = "year",
+                     max_treedepth = 15,
+                     seed = 0)
R> summ_ST <- summary(fit_ST)
R> summ_ST

Summary for the SAE model call:
 fit_sae(formula_fixed = hcr ~ x, domains = "id", disp_direct = "vars",
    type_disp = "var", domain_size = "n", data = emilia,
    spatial_error = TRUE, spatial_df = emilia_shp_ord,
    temporal_error = TRUE,  temporal_variable = "year",
    max_treedepth = 15, seed = 0, iter = 2000)

----- S.D. of the random effects: posterior summaries -----

         mean    sd  2.5%   25%   50%  75% 97.5%
sigma_t 0.104 0.022 0.061 0.090 0.104 0.12 0.148
sigma_s 0.297 0.055 0.205 0.259 0.292 0.33 0.418

----- Fixed effects coefficients: posterior summaries -----
```

```
              mean    sd   2.5%    25%    50%    75%  97.5%
(Intercept) -2.273 0.016 -2.305 -2.284 -2.273 -2.261 -2.242
x            0.123 0.020  0.084  0.109  0.123  0.136  0.163

--------------- Model diagnostics summaries ---------------

                   Min. 1st Qu. Median  Mean 3rd Qu.  Max.
Residuals        -0.024  -0.006  0.001 0.002   0.009 0.036
S.D. Reduction    0.113   0.372  0.456 0.445   0.517 0.677
Bayesian p-value  0.076   0.314  0.465 0.475   0.608 0.978

Shrinkage Bound Rate: 100 %

LOO Information Criterion:
          Estimate     SE
 elpd_loo  484.699  8.332
 p_loo      47.063  4.892
 looic    -969.398 16.664
```

In case of temporal or spatio-temporal object, it is possible to select the year of interest for map plotting via `map()` or when performing benchmarking as follows:

```
R> shares <- aggregate(emilia$pop, list(emilia$year),
+                      function(x) x/sum(x))
R> shares <- as.vector(t(shares[,-1]))
R> bmk_st <- benchmark(summ_ST,
+           bench = 0.09,
+           share = shares[1:38],
+           method = "raking",
+           time = "2014")
```

3.5 Conclusions and Future Developments

The **tipsae** package is a dedicated tool for mapping proportions and indicators defined on the unit interval, widely used to measure, for instance, unemployment, educational attainment and also disease prevalence. To the best of our knowledge, it is the first package implementing Beta-based small area methods, particularly indicated for unit interval responses. Such methods, developed within a Bayesian framework, come equipped with a set of diagnostics and complementary tools, visualizing and exporting functions. The features of the **tipsae** package assist the user in carrying out a complete SAE analysis through the entire process of estimation, validation and results presentation, making the application of Bayesian algorithms and complex SAE methods

straightforward. A Shiny application with a user-friendly interface can be launched to further simplify the process.

Additional features to be integrated in future releases could be, firstly, the implementation of shrinking priors for the regression coefficients, useful for variable selection when several covariates are employed. Secondly, the Beta zero and/or one inflated version already implemented could fail when very few zero or one values are observed. Thus, a possible extension could comprise further flexible alternatives. Lastly, other directions may focus on model extensions for variance shrinking (You and Chapman, 2006; Sugasawa *et al.*, 2017), able to relax the assumption of known dispersion parameter, and for covariates measured with error (Arima *et al.*, 2015).

Appendix

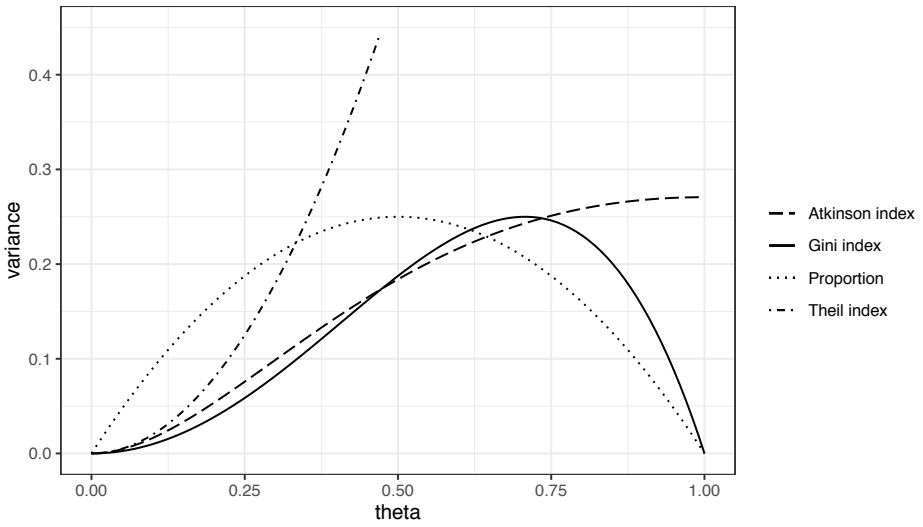

FIGURE A.1: Variance functions with $n_d = 1$ for each measure in comparison with the proportion case

The following proposition derives the approximate variance function for the entire family of Relative Entropy measure, for $\alpha \neq 0, 1$. When $\alpha = 1$, the resulting measure coincides with the Relative Theil index, whose variance function has been obtained in Proposition 2.2 .

Proposition A.1. Under Proposition 2.1 assumptions, the srs estimator of Relative Entropy Index $R_d(\alpha)$ with $\alpha \neq 0, 1$, for domain d, has variance function

$$\mathbb{V}[R_d(\alpha)] \cong \frac{2\theta_d^R(\alpha)^2}{n_d} \exp\{2\theta_d^R(\alpha)(n_d^{\alpha-1} - 1)\},$$

with $\theta_d^R(\alpha)$ its population value.

Proof. Similarly as Propositions 2.1 and 2.2 proofs, the population value of the Relative Entropy index $\theta_d^R(\alpha)$ $\forall \alpha \neq 0, 1$ under log-normal income assumption is defined as

$$\theta_d^R(\alpha) = \frac{1}{n_d^{\alpha-1} - 1}\left(\frac{\mathbb{E}[z^\alpha]}{\mathbb{E}[z]^\alpha} - 1\right) = \frac{1}{n_d^{\alpha-1} - 1}\left(\exp\left\{\frac{\varphi_d^2}{2}\alpha(\alpha - 1)\right\} - 1\right), \qquad \text{(A.1)}$$

with φ_d^2 estimated by $s_d^2 = \frac{1}{n_d-1}\sum_{j=1}^{n_d}[\log(x_{jd}) - \hat{\mu}_d]^2$. By applying the normal distribution theory, $V(s_d) \cong \frac{\phi_d^2}{2n_d}$ and using the delta method:

$$\begin{aligned}
\mathbb{V}[R_d(\alpha)] &= \mathbb{V}\left[\frac{1}{n_d^{\alpha-1}-1}\left(\exp\left\{\frac{s_d^2\alpha(\alpha-1)}{2}\right\}-1\right)\right] \\
&\cong \frac{\varphi_d^4}{2n_d}\exp\{\varphi_d^2\alpha(\alpha-1)\}\frac{\alpha^2(\alpha-1)^2}{(n_d^{\alpha-1}-1)^2} \\
&\cong \frac{2\theta_d^R(\alpha)^2}{n_d}\exp\{2\theta_d^R(\alpha)(n_d^{\alpha-1}-1)\}, \qquad \text{(A.2)}
\end{aligned}$$

where (A.2) is obtained by McLaurin expanding (A.1) so that

$$\varphi_d^2 \cong \frac{2\theta_d^R(\alpha)(n_d^{\alpha-1}-1)}{\alpha(\alpha-1)}.$$

□

Bibliography

Alfons, A., Templ, M. and Filzmoser, P. (2010) An object-oriented framework for statistical simulation: The R package simframe. *Journal of Statistical Software* **37**(3), 1–36.

Alfons, A., Templ, M. and Filzmoser, P. (2013) Robust estimation of economic indicators from survey samples based on Pareto tail modelling. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **62**(2), 271–286.

Antal, E., Langel, M. and Tillé, Y. (2011) Variance estimation of inequality indices in complex sampling designs. In *Proceedings 58th World Statistical Congress.*

Arima, S., Datta, G. S. and Liseo, B. (2015) Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics* **42**(2), 518–529.

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

Atkinson, A. B. (2015) *Handbook of Income Distribution.* Volume 2. ISBN 9780444594303.

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991) Some parametric models on the simplex. *Journal of Multivariate Analysis* **39**(1), 106–116.

Bauder, M., Luery, D. and Szelepka, S. (2015) Small area estimation of health insurance coverage in 2010-2013. Technical report.

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

Bellú, L. and Liberati, P. (2006) Policy impacts on inequality: Welfare based measures of inequality the atkinson index. Technical report.

Bellu, L. G. and Liberati, P. (2006) Social welfare, social welfare functions and inequality aversion. *Food and Agriculture Organization of the United Nations* .

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.

Betancourt, M. (2017) A conceptual introduction to hamiltonian monte carlo. *arXiv Preprint arXiv:1701.02434* .

Biewen, M. (2002) Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics* **108**(2), 317–342.

Biewen, M. and 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.

Bivand, R. S., Pebesma, E. and Gomez-Rubio, V. (2013) *Applied spatial data analysis with R, Second edition*. Springer, NY.

Boonstra, H. J. (2021) *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. R package version 0.7.0.

Breunig, R. (2001) An almost unbiased estimator of the coefficient of variation. *Economics Letters* **70**(1), 15–19.

Breunig, R. and Hutchinson, D. L.-a. (2008) Small sample bias corrections for inequality indices. In *New Econometric Modelling Research*, ed. W. N. Toggins, pp. pp. 61–83. Nova Science Publishers.

Brown, P. J. and Griffin, J. E. (2010) Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis* **5**(1), 171–188.

Brzezinski, M. (2016) Robust estimation of the Pareto tail index: a Monte Carlo analysis. *Empirical Economics* **51**(1), 1–30.

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

Carrasco, J. M. and Reid, N. (2021) Simplex regression models with measurement error. *Communications in Statistics: Simulation and Computation* **50**(11), 3420–3435.

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

Ceriani, L. and Verme, P. (2015) Individual Diversity and the Gini Decomposition. *Social Indicators Research* **121**(3), 637–646.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. and Borges, B. (2021) shiny: Web application framework for r. R package version 1.6.0.

Cowell, F. A. and Victoria-Feser, M.-P. (1996) Robustness properties of inequality measures. *Econometrica: Journal of the Econometric Society* **64**(1), 77–101.

Datta, G. S., Ghosh, M., Steorts, R. and Maples, J. (2011a) Bayesian benchmarking with applications to small area estimation. *Test* **20**(3), 574–588.

Datta, G. S., Hall, P. and Mandal, A. (2011b) Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association* **106**(493), 362–374.

Davidson, R. (2009) Reliable inference for the Gini index. *Journal of Econometrics* **150**(1), 30–40.

Davidson, R. and Flachaire, E. (2007) Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* **141**(1), 141–166.

De Santis, R., Barabesi, L. and Betti, G. (2020) Variance estimation techniques for poverty and inequality measures from complex surveys: a simulation study. Technical report, Department of Economics, University of Siena.

Deltas, G. (2003) The small-sample bias of the Gini coefficient: Results and implications for empirical research. *Review of Economics and Statistics* **85**(1), 226–234.

Demnati, A. and Rao, J. (2004) Linearization variance estimators for survey data. *Survey Methodology* **30**(1), 17–26.

Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**(2), 193–204.

Espinheira, P. L. and de Oliveira Silva, A. (2020) Residual and influence analysis to a general class of simplex regression. *Test* **29**(2), 523–552.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D. and Pérez, A. (2020) Small area estimation of proportions under area-level compositional mixed models. *Test* **29**(3), 793–818.

Esteban, M. D., Morales, D., Pérez, A. and Santamaría, L. (2012) Small area estimation of poverty proportions under area-level time models. *Computational Statistics & Data Analysis* **56**(10), 2840–2855.

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. and 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. and Trivisano, C. (2016) Bayesian beta regression models for the estimation of poverty and inequality parameters in small areas. In *Analysis of Poverty Data by Small Area Estimation*, ed. M. Pratesi, chapter 16, pp. 299–314. John Wiley and Son.

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

Ferrante, M. R. and 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.

Ferrante, M. R. and Pacei, S. (2019) Small Area Estimation of Entropy Inequality Measures: a Comparison Between Alternative Distribution Models. In *JSM Proceedings, Survey Research Methods Section*, pp. 1642–1651. American Statistical Association.

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

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

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. (2013) Mixed beta regression: A bayesian perspective. *Computational Statistics & Data Analysis* **61**, 137–147.

Finkelstein, M., Tucker, H. G. and Alan Veeh, J. (2006) Pareto Tail Index Estimation Revisited. *North American Actuarial Journal* **10**(1), 1–10.

Freni-Sterrantino, A., Ventrucci, M. and Rue, H. (2018) A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and Spatio-Temporal Epidemiology* **26**, 25–34.

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

Gabry, J., Goodrich, B. and Lysy, M. (2020) *rstantools: Tools for Developing R Packages Interfacing with 'Stan'*. R package version 2.1.1.

Gabry, J. and Mahr, T. (2021) bayesplot: Plotting for bayesian models. R package version 1.8.1.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* **1**(3), 515–534.

Giles, D. E. (2005) The Bias of Inequality Measures in Very Small Samples : Some Analytic Results. *Econometrics Working Papers 0514*, (250).

Giorgi, G. M. and Gigliarano, C. (2017) The gini concentration index: a review of the inference literature. *Journal of Economic Surveys* **31**(4), 1130–1148.

Giovinazzi, F. and Cocchi, D. (2021) Social integration of second generation students in the italian school system. *Social Indicators Research* Forthcoming.

Goodrich, B., Gabry, J., Ali, I. and Brilleman, S. (2020) rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1.

Gradshteyn, I. S. and Ryzhik, I. M. (2014) *Table of Integrals, Series, and Products*. Academic press.

Graf, E. and Tillé, Y. (2014) Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology* **40**(1), 61–79.

Guio, A.-C. (2005) Income poverty and social exclusion in the eu25. *Statistics in Focus: Population and Social Conditions* (13), 1–7.

Hampel, F. R. (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**(346), 383–393.

Hoffman, M. D., Gelman, A. *et al.* (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **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.

Jasso, G. (1979) On gini's mean difference and gini's index of concentration. *American Sociological Review* **44**(5), 867–870.

Jørgensen, B. (1997) *The Theory of Dispersion Models*. Volume 41.

Kakwani, N. (1990) Large sample distribution of several inequality measures: with application to cote d'ivoire. In *Contributions to Econometric Theory and Application*, pp. 50–81. Springer.

Kalton, G. (1979) Ultimate cluster sampling. *Journal of the Royal Statistical Society: Series A (General)* **142**(2), 210–222.

Kalton, G., Brick, J. and Le, T. (2005) Estimating components of design effects for use in sample design. Technical report.

Kish, L. (1992) Weighting for unequal p_i. *Journal of Official Statistics* **8**(2), 183–200.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. and Tzavidis, N. (2019) The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software* **91**(7), 1–33.

La Vega, D., Lasso, C. and 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**, 211–225.

Lahiri, P. (2003) On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**(2), 199–210.

Langel, M. and 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.

Lerman, R. I. and Yitzhaki, S. (1989) Improving the accuracy of estimates of gini coefficients. *Journal of Econometrics* **42**(1), 43–47.

Liu, B., Lahiri, P. and Kalton, G. (2007) Hierarchical bayes modeling of survey-weighted small area proportions. In *JSM Proceedings, Survey Research Section*, pp. 3181–3186.

Marchetti, S. and Tzavidis, N. (2021) Robust estimation of the theil index and the gini coeffient for small areas. *Journal of Official Statistics* **37**(4), 955–979.

Marhuenda, Y., Molina, I. and Morales, D. (2013) Small area estimation with spatio-temporal fay–herriot models. *Computational Statistics & Data Analysis* **58**, 308–325.

Marhuenda, Y., Morales, D. and del Carmen Pardo, M. (2014) Information criteria for fay–herriot model selection. *Computational Statistics & Data Analysis* **70**, 268–280.

Márquez, M. A., Lasarte, E. and 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. and Ibrahim, K. (2019) Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics* **7**(3), 70–77.

Midzuno, H. (1952) On the sampling system with probability proportionate to sum of sizes. *Annals of Institute of Statistical Mathematics* **3**(2), 99–108.

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

Mohd Safari, M. A., Masseran, N. and Ibrahim, K. (2018) Outliers detection for Pareto distributed data. *AIP Conference Proceedings* **1940**.

Molina, I. and Marhuenda, Y. (2015) sae: An r package for small area estimation. *The R Journal* **7**(1), 81–97.

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

Morales, D., Pagliarella, M. C. and Salvatore, R. (2015) Small area estimation of poverty indicators under partitioned area-level time models. *SORT* **39**(1), 19–34.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A. and DiMaggio, C. (2019) Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and Spatio-Temporal Epidemiology* **31**, 100301.

Moser, M. and 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. and 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.

Narasimha Prasad, N. and Srivenkataramana, T. (1980) A modification to the horvitz-thompson estimator under the midzuno sampling scheme. *Biometrika* **67**(3), 709–711.

OECD (2011) *Divided we stand: Why inequality keeps rising*. OECD Publishing Paris.

Osier, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods* **3**(3), 167–195.

Osier, G., Berger, Y. G. and Goedeme, T. (2013) Standard error estimation for the eusilc indicators of poverty and social exclusion. *Eurostat Methodologies and Working Papers Series* .

Ospina, R. and Ferrari, S. L. (2010) Inflated beta distributions. *Statistical Papers* **51**(1), 111–126.

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

Pfeffermann, D. (2013) New important developments in small area estimation. *Statistical Science* **28**(1), 40–68.

Pfeffermann, D., Sikov, A. and Tiller, R. (2014) Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test* **23**(4), 631–666.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2021) nlme: Linear and nonlinear mixed effects models. R package version 3.1-152.

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

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranalli, M. G., Montanari, G. E. and Vicarelli, C. (2018) Estimation of small area counts with the benchmarking property. *Metron* **76**(3), 349–378.

Rao, J. N., Chaudhuri, A., Eltinge, J., Fay, R. E., Ghosh, J., Ghosh, M., Lahiri, P. and Pfeffermann, D. (1999) Some current trends in sample survey theory and methods (with discussion). *Sankhyā: The Indian Journal of Statistics, Series B* **61**(1), 1–57.

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

Riebler, A., Sørbye, S. H., Simpson, D. and Rue, H. (2016) An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research* **25**(4), 1145–1165.

Rust, K. and Kalton, G. (1987) Strategies for collapsing strata for variance estimation. *Journal of Official Statistics* **3**(1), 69.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003) *Model assisted survey sampling*. Springer Science & Business Media.

Schluter, C. and van Garderen, K. J. (2009) Edgeworth expansions and normalizing transforms for inequality measures. *Journal of Econometrics* **150**(1), 16–29.

Schmid, T., Bruckschen, F., Salvati, N. and Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**(4), 1163–1190.

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

Slud, E. V. and 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. and Moura, F. A. (2016) Multivariate beta regression with application in small area estimation. *Journal of Official Statistics* **32**(3), 747–768.

Stan Development Team (2020) RStan: the R interface to Stan. R package version 2.21.2.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017) Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics* **44**(1), 150–167.

Tadikamalla, P. R. and Johnson, N. L. (1982) Systems of frequency curves generated by transformations of logistic variables. *Biometrika* **69**(2), 461–465.

Thistle, P. (1990) Large Sample Properties of Two Inequality Indices. *Econometrica* **58**(3), 725–728.

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

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. and Rojas-Perilla, N. (2018) From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181**(4), 927–979.

Ullah, A. (2004) *Finite Sample Econometrics*. Volume 53. ISBN 9788578110796.

Van Kerm, P. (2007) Extreme incomes and the estimation of poverty and inequality indicators from eu-silc. *IRISS Working Paper Series* .

Van Ourti, T. and Clarke, P. (2011) A simple correction to remove the bias of the gini coefficient due to grouping. *Review of Economics and Statistics* **93**(3), 982–994.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T. and Gelman, A. (2020) loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.4.1.

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

Wieczorek, J. and Hawala, S. (2011) A bayesian zero-one inflated beta model for estimating poverty in us counties. In *JSM Proceedings, Section on Survey Research Methods*. American Statistical Association.

Wieczorek, J., Nugent, C. and Hawala, S. (2012) A Bayesian Zero-One Inflated Beta Model for Small Area Shrinkage Estimation. In *JSM Proceedings*, pp. 3896–3910. American Statistical Association.

You, Y. and Chapman, B. (2006) Small area estimation using area level models and estimated sampling variances. *Survey Methodology* **32**(1), 97–103.

You, Y. and Rao, J. (2002) Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics* **30**(1), 3–15.

Zhang, J. L. and Bryant, J. (2020) Fully bayesian benchmarking of small area estimation models. *Journal of Official Statistics* **36**(1), 197–223.