Robust estimation of the Theil index and the Gini coefficient for small areas

Stefano Marchetti[1] and Nikos Tzavidis[2]

[1]University of Pisa, Department of Economia e Management, Via C. Ridolfi 10, 56124 Pisa (PI), Italy. stefano.marchetti@unipi.it

[2]University of Southampton, Social Statistics and Demography Social Sciences, Southampton SO17 1BJ, United Kingdom. n.tzavidis@soton.ac.uk

Received: date / Accepted: date

Abstract

Small area estimation is receiving considerable attention due to the high demand for small area statistics. Small area estimators of means and totals have been widely studied in the literature. Moreover, in the last years also small area estimators of quantiles and poverty indicators have been studied. In contrast, small area estimators of inequality indicators, which are often used in socio-economic studies, have received less attention. In this article we propose a robust method based on the M-quantile regression model for small area estimation of the Theil index and the Gini coefficient, two popular inequality measures. To estimate the mean squared error a non-parametric bootstrap is adopted. A robust approach is used because often inequality is measured using income or consumption data, which are often non-normal and affected by outliers. The proposed methodology is applied to income data to estimate the Theil index and the Gini coefficient for small domains in Tuscany (provinces by age groups), using survey and Census micro-data as auxiliary variables. In addition, a design-based simulation is carried out to study the behaviour of the proposed robust estimators. The performance of the bootstrap mean squared error estimator is also investigated in the simulation study.

Keywords— Small area estimation, M-quantile models, Inequality indicators

Acknowledgements— This work has been developed under the support of the project Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy (InGRID-2), grant agreement 730998.

1 Introduction

To formulate and implement policies, and allocate funds there is need for timely, reliable and disaggregated estimates of a large set of parameters, such as means, quantiles, poverty and inequality indicators. Sample surveys provide an effective way of obtaining estimates for such population characteristics. Estimation, however, can become difficult when the focus is on domains (areas) with small sample sizes. The term "small areas" is typically used to describe domains whose sample sizes are not large enough to allow for reliable direct estimation, i.e. estimation based only on the sample data from the domain (Rao and Molina, 2015). When direct estimation leads to unreliable estimates, one has to rely upon alternative model-based methods for producing small area estimates. Two approaches for model based small area estimation are based on the mixed effect models (Rao and Molina, 2015) and the M-quantile models (Chambers and Tzavidis, 2006).

Despite the fact that poverty indicators have been studied extensively under both the approaches (Molina and Rao, 2010; Marchetti et al., 2012), small area estimation of inequality indicators using the M-quantile approach has not been studied extensively. In this paper we study M-quantile small area estimators of the Theil index and the Gini coefficient. These two inequality indicators are the most commonly used by practitioners. The popularity of the Gini coefficient is mainly due to its simplicity, while the appeal of the Theil index is in its decomposability into "between" and "within" domains. The estimation of these inequality measures is challenging because of their non-linear form. Model assumptions become even more important and departure from these assumptions may have a more noticeable effect on the estimates.

Often, inequality indicators are estimated from variables that are skewed and affected by outliers, such as consumption and income. Chambers and Tzavidis (2006) and Sinha and Rao (2009) proposed robust model-based outlier robust methods for small area estimation. Chambers and Tzavidis (2006) addressed the issue of outliers robustness in small area estimation using an approach based on fitting M-quantile models (Breckling and Chambers, 1988) to the survey data, while Sinha and Rao (2009) addressed this issue from the perspective of linear mixed models. Chambers et al. (2014) defined such methods as robust projective, since they project the behavior of the robust working model of the sample onto the non-sampled part of the population. Tzavidis et al. (2010) and Chambers et al. (2014) proposed methods that allow for contributions from representative sample outliers. These methods are defined as robust predictive method, since they attempt to predict the contribution of the population outliers to target parameters. Other alternatives are possible, for example Gershunskaya and Lahiri (2010) include a modification of a classical linear mixed model assuming that the underlying distribution is a scale mixture of two normal distributions, where outliers are assumed to have a larger

variance than regular observations. The proposed estimators can be classified as robust predictive. The ELL (or World Bank) proposed by Elbers et al. (2003) and the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010) are among widely used methods for poverty mapping. These methods are based on linear mixed models, and assume normally distributed errors. When data are skewed the log transformation is commonly used to obtain approximately normally distributed model residuals. However, in some cases a log transformation may not be appropriate. Recently, Tzavidis et al. (2018) and Rojas-Perilla et al. (2020) proposed the use of data-driven power transformations in small area estimation. An alternative is to specify a model with alternative distributional assumptions to deal with skew-data. For instance Graf et al. (2019) discuss an EBP approach under a generalized beta distribution of the second kind for the errors terms and Elbers and van der Weide (2014) propose a method for estimating distribution functions using a mixture of normal distributions for the model errors. Diallo and Rao (2018) derive an EB estimator by relaxing the normality assumptions, assuming skewnormal errors. The approach we propose in this paper is based on the M-quantile model and is an alternative to estimators under the linear mixed model.

The remainder of the paper is organized as follows. Section 2 introduce the quantity of interest, which are the Theil index and the Gini coefficient, section 3 summarizes the M-quantile approach to small area estimation, section 4 introduces the small area estimators of Theil index and Gini coefficient based on M-quantile models, using a Monte Carlo approximation and a bias correction technique, i.e. the Chambers and Dunstan (1986) correction. Moreover, we discuss mean squared error estimation. Section 5 is devoted to evaluate the performance of the proposed estimators by means of Monte Carlo design-based simulations. In section 6 we present results on Gini and Theil estimates at provincial level in the Tuscany region in Italy. Section 7 summarizes the main results of the paper and puts forward ideas for further research.

2 Direct estimation of the Theil index and the Gini coefficient

Inequality measures are mainly based on non linear statistics. The most popular of these is the Gini coefficient (Gini, 1914). It has been shown to be inferior to more recently measures, such as the Zenga index (Zenga, 2007), nevertheless, it has a number of advantages over other measures, such as its simplicity, and it is still widely proposed in empirical studies.

Let i be the index for domains (or areas), $i = 1, \ldots, m$ where m is the number of domains, and let j be the index for units within the domain. We denote the population size, sample size, sampled part of the population and non sampled part of the population in area i respectively by N_i, n_i, s_i and r_i. We assume that the sum over the areas of N_i and n_i is equal to N and n respectively.

The Gini coefficient can be defined in many ways. Usually, it is defined by means of the Lorenz curve. A popular

alternative is based on the absolute value of the difference between all pairs of the target variable:

$$G_i = \frac{\Delta_i}{2\mu_i}\,, \tag{1}$$

where $\mu_i = \int y\, dF_i(y)$, $\Delta_i = \int\int |y_1 - y_2|\, dF_i(y_1)\, dF_i(y_2)$, $y \geq 0$ and y_1, y_2 are random variables with a common distribution, that is $F_i(y_1) = F_i(y_2) = F_i(y)$. Usually y represent a measure of the income or consumption. In the rest of the paper y is a continuous variable with support $(0, +\infty)$ and distribution function $F_i(y)$, where the subscript i indicates the domain.

The statistic G is equal to 1 when inequality is at its maximum and it is zero at its minimum (equal distribution).

Another popular inequality statistics is the Theil index (Theil, 1967), which belong to the family of generalized entropy measures. It can be defined as (Bourguignon, 1979; Shorrocks, 1980; Cowell and Kuga, 1981; Foster, 1983; Maasoumi, 1986)

$$T_i = \frac{\nu_i}{\mu_i} - \log(\mu_i)\,, \tag{2}$$

where $\mu_i = \int y\, dF_i(y)$, $\nu_i = \int y \log(y)\, dF_i(y)$ and $y > 0$.

The statistic T is equal 0 when all the population units share the same amount of the total of y, i.e. equal distribution, and it is equal $\log(N)$ (where N is the population size) under maximum inequality, i.e. one unit holds the total amount of y and the other units hold 0. Its popularity is mainly due to its decomposability into "between" and "within" domains. Assuming T is the Theil index for the entire population that is divided into m domains, then

$$T = \sum_{i=1}^{m} f_i T_i + \sum_{i=1}^{m} f_i \log \frac{\mu_i}{\mu},$$

where $f_i = \frac{N_i \mu_i}{N \mu}$ is the share of y in domain i, μ is the population mean of y and T_i is the Theil index in domain i. The first sum is the part that is due to inequality within domains, the second is the part that is due to differences between domains.

We now discuss direct estimation of inequality indicators for small areas (domains). Direct estimation for the Gini coefficient is not straightforward. Some popular direct estimators in the literature are known to be negatively biased in small sample (Deltas, 2003; Alfons and Templ, 2013), such as

$$\tilde{G}_i^{Dir} = \frac{2\sum_{j=1}^{n_i}\left(w_{ij}\sum_{h=1}^{j} w_{ih}\right) - \sum_{j=1}^{n_i} y_{ij} w_{ij}^2}{\sum_{j=1}^{n_i} w_{ij} \sum_{j=1}^{n_i} y_{ij} w_{ij}} - 1\,,$$

where the values $y_{ij}, j = 1, \ldots, n_i$ are assumed to be sorted in ascending order and $w_{ij}, j = 1, \ldots, n_i$ is the survey weight associated to y_{ij}.

Davidson (2009) notes that the main term in the bias of $\tilde{G}_i^{Dir}$ can be removed by a $n_i(n_i - 1)^{-1}$ multiplication,

under simple random sampling design. However, as noted by Langel and Tillé (2013) under complex sample designs the correction of Davidson (2009) is not trivial. We decide to use the following direct estimator (Langel and Tillé, 2013):

$$\hat{G}_i^{Dir} = \frac{\hat{\Delta}_i^{Dir}}{2\hat{\mu}_i^{Dir}} = \frac{\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} w_{ij}w_{ik}|y_{ij}-y_{ik}|}{N_i^2}\frac{1}{2N_i^{-1}\sum_{j=1}^{n_i} w_{ij}y_{ij}}\,, \tag{3}$$

where N_i is the population size in area i (assumed known).

For direct estimation of the Theil index we use the estimator proposed in Davidson and Flachaire (2007), here adapted to account for the use of a complex sampling design:

$$\hat{T}_i^{Dir} = \frac{\hat{\nu}_i^{Dir}}{\hat{\mu}_i^{Dir}} - \log(\hat{\mu}_i^{Dir}), \tag{4}$$

where $\hat{\mu_i}^{Dir} = N_i^{-1}\sum_{j=1}^{n_i} y_{ij}w_{ij}$, $\hat{\nu}_i^{Dir} = N_i^{-1}\sum_{j=1}^{n_i} y_{ij}\log(y_{ij})w_{ij}$. The direct estimator we use is biased for small samples because $\hat{\nu}_i^{Dir}/\hat{\mu}_i^{Dir}$ is a biased ratio estimator of ν_i/μ_i, though it should be consistent for large samples. Nevertheless, we decided to use estimators (3) and (4) because their forms are suitable for applying the Chambers and Dunstan (1986) correction.

Although variance estimation of direct estimates is not of interest in this work, it can be shown that an asymptotic variance estimator of $\hat{T}_i^{Dir}$ (under simple random sampling) can be derived using the Delta method. However, Davidson and Flachaire (2007) notes that this variance estimator leads to inference that is not accurate even in large sample. The same result applies to standard bootstrap variance estimation. Variance estimation of the Theil index is also discussed, among others, in Mills and Zandvakili (1997).

Variance estimation of (3) is not straightforward, even under assumption of log-normality of the target. Asymptotic estimators of the variance have been proposed for example by Battacharya (2007), while bootstrap techniques are discussed for example in Mills and Zandvakili (1997); Alfons and Templ (2013). A literature review about the variance estimation of the Gini coefficient is in Langel and Tillé (2013).

3 Outlier robust small area estimation using M-quantiles

3.1 M-quantile approach to small area

A robust approach to small area estimation is based on the use of the quantile/M-quantile regression model (Chambers and Tzavidis, 2006).

In what follows we assume that a vector of p auxiliary variable $\boldsymbol{x}_{ij}$ is known for each population unit j in small area $i = 1, \ldots, m$ and that values of the variable of interest y are available from a sample that includes units from all the small areas of interest. We further assume that conditional on covariate information for example, design variables, the

sampling design is ignorable.

The M-quantile of order $q \in (0,1)$ of the conditional density of y given the set of covariates $\boldsymbol{x}$, $f(y|\boldsymbol{x})$, is defined as the solution $Q_y(q|\boldsymbol{x},\psi)$ of an estimating equation $\int \psi_q\{y - Q_y(q|\boldsymbol{x},\psi)\}f(y|\boldsymbol{x})\,dy = 0$, where ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. In particular, a linear M-quantile regression model for y_{ij} given $\boldsymbol{x}_{ij}$ is one where we assume that

$$Q_y(q|\boldsymbol{x}_{ij},\psi) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q). \tag{5}$$

That is, we allow a different set of p regression parameters for each value of $q \in (0,1)$. The estimator of $\boldsymbol{\beta}_\psi(q)$ can be obtained by solving

$$\sum_{i=1}^{m}\sum_{j\in s_i}\psi_q(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q))\boldsymbol{x}_{ij} = \mathbf{0}$$

with respect to $\boldsymbol{\beta}_\psi(q)$, assuming that

$$\psi_q(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q)) = 2\psi\{s^{-1}(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q))\} \times \{qI(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q) > 0) + (1-q)I(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q) \le 0)\},$$

where s is a suitable robust estimate of scale, e.g. the MAD estimate $s = median|y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_\psi(q)|/0.6745$. A popular choice for the influence function is the Huber, $\psi(u) = uI(|u| \le c) + c\,\mathrm{sgn}(u)I(|u| > c)$ (Chambers and Tzavidis, 2006). However, alternative influence functions are also possible. Provided that the tuning constant c is strictly greater than zero, estimates of $\boldsymbol{\beta}_\psi(q)$ are obtained using iterative weighted least squares (IWLS).

Chambers and Tzavidis (2006) extended the use of M-quantile regression models to small area estimation. They characterized the conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit j in area i this coefficient is the value q_{ij} such that $Q_y(q_{ij}|\boldsymbol{x}_{ij},\psi) = y_{ij}$. The M-quantile coefficients are determined at the population level. Consequently, if a hierarchical (grouping/clustering) structure does explain part of the variability in the population data, then we expect units within clusters to have similar M-quantile coefficients.

When the conditional M-quantiles are assumed to follow the linear model (5), with $\boldsymbol{\beta}_\psi(q)$ a sufficiently smooth function of q, Chambers and Tzavidis (2006) define a naive estimator of the mean, i.e. $\hat{\mu}_i^{naive} = N_i^{-1}\big\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i}\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)\big\}$, where $\hat{\theta}_i$ is an estimate of the average value of the M-quantile coefficients of the units in area i. See Chambers and Tzavidis (2006) for further details on the estimation of the M-quantile coefficients at unit level and for the computation of the small area M-quantile coefficients. Bianchi et al. (2018) proposed a test statistic for testing how close the domain-specific quantile coefficients are to 0.5, which is used in the application.

The M-quantile small area model can be more formally defined as follows:

$$y_{ij} = \boldsymbol{x}_{ij}^T \beta_\psi(\theta_i) + \epsilon_{ij}\,, \tag{6}$$

where $\beta_\psi(\theta_i)$ is the unknown vector of M-quantile regression parameters for the unknown area-specific M-quantile coefficient θ_i, and ϵ_{ij} is the unit level random error term with distribution function for which no explicit parametric assumptions are being made. The unknown parameters $\beta_\psi(\theta_i)$ and θ_i are estimated as mentioned from sample data, the model residuals are then $e_{ij} = y_{ij} - \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$.

3.2 Bias correction

A robust projective estimator (naive estimator, e.g. $\hat{\mu}_i^{naive}$) assumes that all the non-sampled units follow the (robustly fitted) working model. However, in practice we should expect that there will be outliers not only in the sample, but also among the non-sampled units. Hence, using the M-quantile predictions for the out-of-sample units directly leads to a biased estimator of the small area target parameter. This is linked to the idea of representative and non-representative outliers described in Chambers (1986) and Chambers et al. (2014). Using the ideas in Chambers (1986), Tzavidis et al. (2010) substitute a consistent estimator of the distribution function, using the approach of Chambers and Dunstan (1986), to derive a version of the M-quantile estimator adjusted for bias also referred to as robust-predictive estimator. In particular, Tzavidis et al. (2010) define the Chambers-Dunstan (CD) estimator of the small area distribution function as

$$\hat{F}_i^{CD}(t) = N_i^{-1}\Big[\sum_{j\in s_i} I(y_{ij} \le t) + n_i^{-1} \sum_{k\in r_i}\sum_{j\in s_i} I(\boldsymbol{x}_{ik}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + e_{ij} \le t)\Big]. \tag{7}$$

Estimates of θ_i and $\boldsymbol{\beta}_\psi(\theta_i)$ are obtained following Chambers and Tzavidis (2006).

By using the Chambers-Dunstan estimator of the small area distribution function one can define a general framework for small area estimation that allows for the estimation of small area averages, quantiles, non-linear indicators for example, the Gini coefficient and the Theil index. For example the M-quantile CD-based estimator of the average of y in small area i is defined as

$$\begin{aligned}\hat{\mu}_i^{CD} &= \int_{-\infty}^{+\infty} y \; \mathrm{d}\hat{F}_i^{CD}(y) \\ &= N_i^{-1}\Big[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij} + (1-f_i)\sum_{j\in s_i} e_{ij}\Big].\end{aligned} \tag{8}$$

where $f_i = n_i N_i^{-1}$ is the sampling fraction in area i and $\hat{y}_{ij} = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)$, $j \in r_i$ (Tzavidis et al., 2010). The bias correction is the third addend in (8), and means that this estimator has higher variability than the naive M-quantile estimator.

Nevertheless, because of its bias robust properties, (8) is usually preferred, over the naive M-quantile estimator, in practice.

Similarly, Tzavidis et al. (2010) use the CD estimator of the small area distribution function to propose an estimator of the small area quantiles, and Marchetti et al. (2012) discuss estimation of the Foster et al. (1984) poverty measures.

4 M-quantile model-based estimation of the Theil index and the Gini coefficient

In this section we describe the methodology for estimating the Theil index and the Gini coefficient for small areas using the M-quantile approach. We derive these estimators using the bias correction introduced by Chambers and Dunstan (1986) and extended to the small area framework by Tzavidis et al. (2010). We start by describing the small area estimator of the Theil index and then the Gini coefficient. The Monte-Carlo version of these estimators is also considered at the end of the section.

4.1 Small area estimation for the Theil index

To estimate T at the small area level we plug-in the CD estimator of the distribution function (7) in (2). Therefore, the small area estimator of the Theil index can be written as

$$\hat{T}_i^{CD} = \frac{\hat{\nu}_i^{CD}}{\hat{\mu}_i^{CD}} - \log(\hat{\mu}_i^{CD}) \tag{9}$$

where $\hat{\mu}_i^{CD} = \int y\, d\hat{F}_i^{CD}(y)$, $\hat{\nu}_i^{CD} = \int y\log(y)\, d\hat{F}_i^{CD}(y)$. As an alternative, ν_i^{CD} can also be estimated using a transformed variable $z = y\log(y)$, therefore $\hat{\nu}_i^{CD} = \int z\, d\hat{F}_i^{CD}(z)$. Using first order Taylor expansion we can show that (9) is unbiased, assuming model-unbiasedness of $\hat{\mu}_i^{CD}$ and $\hat{\nu}_i^{CD}$ (see (17) and (18) in the Appendix). The estimators $\hat{\mu}_i^{CD}$ and $\hat{\nu}_i^{CD}$ can be assumed model-unbiased because $\hat{F}_i^{CD}(t)$ is model-unbiased for $F_i(t)$ under some reasonable conditions specified in Chambers and Dunstan (1986); Wu and Sitter (2001).

We already introduced the CD-based estimator of the small area mean $\hat{\mu}_i^{CD}$ in (8). Noting that

$$\int_{-\infty}^{+\infty} g(t)\, d\hat{F}_i^{CD}(t) = N_i^{-1}\Big\{ \sum_{j\in s_i} g(y_{ij}) + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} g(\hat{y}_{ik} + e_{ij}) \Big\},$$

we can obtain the CD-based estimator of $\nu = g(y) = y\log(y)$ as follows,

$$\hat{\nu}_i^{CD} = N_i^{-1}\Big\{ \sum_{j\in s_i} y_{ij}\log(y_{ij}) + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} (\hat{y}_{ik} + e_{ij})\log(\hat{y}_{ik} + e_{ij}) \Big\}, \tag{10}$$

where $\hat{y}_{ik}$ is the predicted value of y_{ik} for the out of sample unit $k \in r_i$ using model (6) and $e_{ij}, j = 1, \ldots, n_i$ are the model residuals in area i. Alternatively, ν_i can also be estimated as

$$\hat{\nu}_i^{CD} = N_i^{-1}\Big\{ \sum_{j \in s_i} z_{ij} + \sum_{k \in r_i} \hat{z}_{ik} + (N_i/n_i - 1) \sum_{j \in s_i} e_{ij}^z \Big\}, \tag{11}$$

where $z_{ij} = y_{ij}\log(y_{ij})$, $\hat{z}_{ik} = \hat{y}_{ik}\log(\hat{y}_{ik})$ and e_{ij}^zs are residuals obtained from the M-quantile small area model in (6) where y_{ij} is replaced by z_{ij}, see (15) and (16) in the appendix for further details. Empirically equations (10) and (11) are equivalent and give the same results, however it is difficult to show this algebraically. However, (11) is computationally faster because it doesn't involve the double summation present in (10).

4.2 Small area estimation for the Gini coefficient

To estimate the Gini coefficient we adopt the same strategy used before for the Theil index. Therefore, we plug-in the distribution function estimator (7) in (1) leading to the following small area estimator

$$\hat{G}_i^{CD} = \frac{\hat{\Delta}_i^{CD}}{2\hat{\mu}_i^{CD}}\,, \tag{12}$$

where $\hat{\Delta}_i^{CD} = \int\int |y_1 - y_2|\, d\hat{F}_i^{CD}(y_1)\, d\hat{F}_i^{CD}(y_2)$. Assuming $\hat{F}_i^{CD}$ is model-unbiased for F_i then using first-order Taylor expansion we can show that the estimator (12) is approximately model-unbiased (see (20) and (21) in the Appendix).

Estimator $\hat{\mu}_i^{CD}$ is that of equation (8). Estimator $\hat{\Delta}_i^{CD}$ is obtained as follows (see (19) in the Appendix)

$$\begin{aligned}\hat{\Delta}_i^{CD} &= \int\int |t_1 - t_2|\, d\hat{F}_i^{CD}(t_1)\, d\hat{F}_i^{CD}(t_2) \\ &= N_i^{-2}\Big\{ \sum_{j \in s_i}\sum_{l \in s_i} |y_{ij} - y_{il}| + n_i^{-2} \sum_{j \in s_i}\sum_{k \in r_i}\sum_{l \in s_i}\sum_{h \in r_i} |\hat{y}_{ik} + e_{ij} - (\hat{y}_{ih} + e_{il})| \Big\}. \end{aligned} \tag{13}$$

Computing the quadruple summation in (13) is computationally intensive when the population area size is large (for example greater than 5000 units). In the R language (R Development Core Team, 2013) the use of arrays to speed up the computation is possible. As an alternative, we wrote a C function that can be called in R through a dynamic library, which uses a nested "for" to compute the quadruple summation in reasonable time also for large population domain sizes. The R-code is available in the supplementary materials. The required computational time is discussed in the application Section.

4.3 Small area estimation based on Monte Carlo approximation

It is important to mention that small area target parameters can alternatively be estimated by approximating the distribution of the unknown quantity $y_{ik}, k \in r$ by means of Monte-Carlo simulations. Let δ_i be a parameter of interest

in area i that depends from a vector of known constants $\boldsymbol{c} = \{c_1, c_2, \ldots\}$:

$$\delta_i = \delta_i(\boldsymbol{c}) = h(y_{ij} \cup y_{ik}, \boldsymbol{c}) \quad j \in s_i, k \in r_i \,,$$

where h is a function of the target variable y and the vector of known constants $\boldsymbol{c}$. Let $\boldsymbol{y}_s = \{y_j, j \in s\}$ be the vector of sample observations, which obey a superpopulation model, and let $\boldsymbol{t}$ be the vector of unknown parameters of the superpopulation model. A predictor of δ_i can be obtained by preserving the values corresponding to the sample units and predicting those corresponding to non sampled units:

$$\hat{\delta}_i = h(y_{ij} \cup E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}], \boldsymbol{c}) \,,$$

where $\hat{\boldsymbol{t}}$ is a consistent estimator of $\boldsymbol{t}$ and $E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}] = \hat{y}_{ik}$ an unknown quantity that can be approximated by using of Monte Carlo simulation. It is important to note that if $E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}]$ depends on $\boldsymbol{x}_{ij}$ then the covariate values need to be known for all the units in the population. This is comparable to other methodologies that use unit-level models to estimate domain-specific non-linear indicators, see for example the EBP and ELL methods.

When we use the M-quantile model to estimate δ_i the Monte Carlo approximation can be obtained as follows:

1. Fit the M-quantile small area model using the sample values $\boldsymbol{y}_s$ and obtain estimates $\hat{\boldsymbol{t}} = \{\hat{\theta}_i, \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)\}$.
2. Generate an out of sample vector of size $N_i - n_i$ using

$$y^*_{ik} = \boldsymbol{x}^T_{ik}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + e^*_{ik}, k \in r_i, i = 1, \ldots, m \,,$$

   where $e^*_{ik}, k \in r_i, i = 1, \ldots, m$ is drawn from the empirical distribution function of the M-quantile model residuals (residuals can be drawn either from the domain (area) i residuals or from all the residuals).
3. Repeat the process L times. Each time combine the sample data $y_{ij}, j \in s_i$ and out of sample data $y^*_{ik}, k \in r_i$ for computing $\hat{\delta}^{(l)}_i$.
4. Average the results over L simulations to obtain an estimate of δ_i, $\hat{\delta}_i = L^{-1}\sum_{l=1}^{L}\hat{\delta}^{(l)}_i$.

Further discussion on this Monte Carlo approach can be found in Marchetti et al. (2012). Usually in real applications linkage between sampled units and population units is not possible, that is the set r is unknown. In this case the prediction is carried out for all the units in the population $U_i = \{s_i, r_i\}$, then $\hat{\delta}_i = h(E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}], \boldsymbol{c}), k \in U_i$. When the sampling fraction is very small $h(E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}], \boldsymbol{c}), k \in U_i$ and $h(y_{ij} \cup E[y_{ik}|\boldsymbol{y}_s; \hat{\boldsymbol{t}}], \boldsymbol{c}), k \in r_i$ are very similar.

Setting

$$h(y_1, \ldots, y_{n_i}) = \frac{n_i^{-1}(n_i - 1)^{-1}\sum_{j=1}^{n_i}\sum_{l=1}^{n_i}|y_{ij} - y_{il}|}{n_i^{-1}\sum_{j=1}^{n_i} y_{ij}}$$

we obtain the Gini coefficient MC estimator, and setting

$$h(y_1, \ldots, y_{n_i}) = \frac{n_i^{-1} \sum_{j=1}^{n_i} y_{ij} \log y_{ij}}{n_i^{-1} \sum_{j=1}^{n_i} y_{ij}} - \log n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$$

we obtain the Theil index MC estimator.

The M-quantile MC estimators mimic the Elbers et al. (2003) approach, however, it is challenging to theoretically justify this method therefore, statistical properties are shown via simulations. In contrast, for the M-quantile CD estimators the theoretical background is better understood (Tzavidis et al., 2010).

4.4 MSE estimation

MSE estimation for M-quantile small area estimators is widely discussed for linear statistics, such as means and totals (Chambers et al., 2014). Less research is available for non-linear statistics. An MSE estimator based on a non-parametric bootstrap scheme for small area estimators under the M-quantile model that can be used also with non-linear statistics is extensively discussed in Marchetti et al. (2012). More details on the non-parametric bootstrap approach for finite population can also be found, among others, in Lombardía et al. (2003).

Starting from a random sample s selected from a finite population U without replacement, we fit the M-quantile small area model (6), and we obtain estimates $\hat{\boldsymbol{t}} = \{\hat{\theta}_i,\ \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i)\}$ and residuals $e_{ij}, i = 1, \ldots, m; j \in s_i$. The bootstrap MSE estimates can be obtained as follows:

1 Given an estimator $\hat{G}(u)$ of the distribution of the residuals $G(u) = P(e \leq u)$, a bootstrap population, consistent with the M-quantile small area model can be generated by sampling from $\hat{G}(u)$ to obtain e_{ij}^*:

$$y_{ij}^* = \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_i) + e_{ij}^*.$$

For defining $\hat{G}(u)$ we consider two approaches: (a) sampling from the empirical distribution function of the model residuals or (b) sampling from a smoothed distribution function of the model residuals. For each of the two above mentioned approaches, sampling can be done in two ways: (i) by sampling from the distribution of all residuals without conditioning on the small area (unconditional approach) or (ii) by sampling from the distribution of the residuals within small area i (conditional approach). These methods are described in detail in Tzavidis et al. (2010).

2 According to point 1 choose one approach from (a) or (b) and one from (i) or (ii), and generate B bootstrap populations.

3 From *each* of the B bootstrap population draw L samples using simple random sample – of size n_i – within areas.

4 Using the L samples compute the estimates of the Theil index and the Gini coefficient according to the methods proposed in section 4.

5 Let $\hat{\tau}_i$ be the the estimated small area parameter (from the original sample), τ_i^{*b} be the small area parameter (true value) of the bth bootstrap population, $\hat{\tau}_i^{*bl}$ be the small area parameter estimated by using the l sample from the b bootstrap population. The bootstrap estimator of the bias and the variance of $\hat{\tau}_i$ are defined respectively by

$$\hat{B}(\hat{\tau}_i) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}(\hat{\tau}_i^{*bl} - \tau_i^{*b})\,,$$
$$\hat{V}(\hat{\tau}_i) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}(\hat{\tau}_i^{*bl} - \bar{\hat{\tau}}_i^{*b})^2\,,$$

where $\bar{\hat{\tau}}_i^{*b} = L^{-1}\sum_{l=1}^{L}\hat{\tau}_i^{*bl}$. The bootstrap MSE estimator of the estimated small area parameter is finally defined as

$$\widehat{MSE}(\hat{\tau}_i) = \hat{V}(\hat{\tau}_i) + \hat{B}(\hat{\tau}_i)^2. \tag{14}$$

Bootstrapping in the presence of outlier contamination is a challenging problem. The properties of the proposed bootstrap MSE are examined in Section 5.2. The issue of bootstrapping in the presence of outlier contamination is discussed in Schmid et al. (2016), but further research on bootstrap MSE estimation in the presence of contamination is needed. A promising approach to tackling this problem is offered by the more recent work in Dongomo-Jiongo and Nguimkeu (2018). The authors propose to generate bootstrap populations by using the non-robust mixed model fit. Although this idea can be applied to the M-quantile predictors, this extension is not immediately applicable and will be considered in future work.

To estimate MSE of (9) and (12) one can also attempt to use a Taylor linearization. However, using simulations, which are not reported here, we have noted that this approximation is not accurate to the desired order, and hence not reliable. The reason is that Taylor expansions are asymptotic results and depend on having a sufficient sample size to work well, while in the small area estimation framework a number of areas are expected to have small sample sizes. Moreover, the Taylor-linearized MSE for the Theil index is the same to the one obtained by the delta method in Davidson and Flachaire (2007), which they prove not to be accurate even for a large sample. It is worth noting that MSE estimation for such indicators is very difficult, in particular for small samples. Therefore, it may be reasonable to expect poor performance of MSE estimators. Future work will consider a bootstrap bias correction for the linearized MSE estimator.

5 Design-based evaluation of the proposed estimators

In this section we use design-based Monte-Carlo simulations to study the performance of the proposed small area robust estimators of the Theil index and the Gini coefficient. Moreover, we also evaluate the performance of the bootstrap MSE estimator of these.

The population underpinning the design-based simulation is based on the data used in the application in section 6. Our target domains are the same used in the application. The population for the design-based simulation has been obtained by fitting a mixed effects model to the EU-SILC data, and then predicting the target using the Census data.

We fit a linear mixed model (random intercept) on the EU-SILC data using the household equivalised income as target variable and as auxiliary variables *owners* (proportion of households who hold their house), *work status* (a binary variable indicating if the head of the household works), *gender* (a binary variable indicating the gender of the head of the household), *education* (number of year of education of the head of the household), *household size* (number of household members), which are common between Census and EU-SILC.

Then, we generate the target value for all the population units using the Census auxiliary variables and the model estimates adding variability by sampling from model-level one and two residuals. The resulting synthetic target value has a distribution similar to that observed in the EU-SILC data, as shown in figure 1. We refer to the generated synthetic equivalised household income and the auxiliary variables as synthetic population.

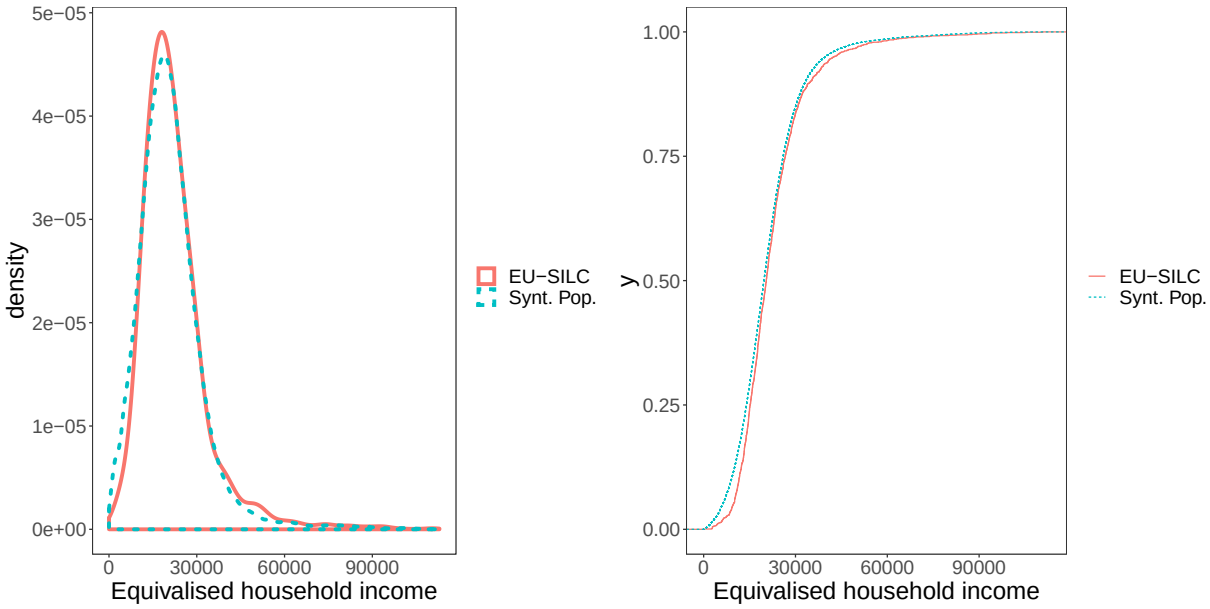

Figure 1: Density estimates of the household equivalised income from the EU-SILC (solid line) and the synthetic population (dashed line).

From the synthetic population we draw 1000 samples with a design similar to that of the EU-SILC survey in Italy in 2008. The survey design of the EU-SILC in Italy is a two stage stratified sample with a rotating panel, for details see `http://siqual.istat.it/SIQual/visualizza.do?id=5000170&refresh=true&language=IT`. Applying this design to each sample leads to a different sample size, which varies between 1277 to 1704 households, with an average of 1472 (the actual EU-SILC 2008 sample size is 1495). The average sample size across the domains varies from a minimum of 4.9 to a maximum of 204.4, with a mean of 49.1 and a median of 40.

For each sample we estimate the Theil index and the Gini coefficient at the domain level (province by age class) using the M-quantile CD and MC estimators. To compare the results of the proposed estimators we use as benchmark the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010). This method is based on a linear mixed model and requires a transformation of the response variable to obtain an approximate normal distribution of the model error terms. We first tried to use the log scale, but the results were unsatisfactory. Therefore, we decided to use a data-driven Box-Cox transformation (Box and Cox, 1964; Rojas-Perilla et al., 2020). We apply this data-driven transformation in each Monte Carlo replication using the R package `emdi` (Kreutzmann et al., 2019). For comparing the EBP and M-quantile estimators we also fit the M-quantile model using the same Box-Cox transformation as in the case of the EBP, even though we acknowledge that the best transformation for the EBP it is not necessarily the best transformation for the MQ model.

Usually, in applications to real cases it is not possible to link the sampled units with the population units, and then obtain the set r of the non sampled units. We replicate this situation in this design-based simulation. Estimators are then modified accordingly, see (22), (23) and (24) in the appendix.

5.1 Discussion about point estimation

In table 1 we present results for comparing the M-quantile MC estimator and the EBP estimator. For the M-quantile MC estimator we produce results by using a model that is estimated both with the untransformed income data and the transformed income data. The EBP estimates are produced by using a mixed model fitted to the transformed income following the methodology described in Rojas-Perilla et al. (2020). At this point it is important to clarify the following points.

Although the EBP results on the untransformed scale have been produced, we have decided not to report these because the mixed model assumptions are not satisfied on this scale. The results are available from the authors. Overall, the results from using the EBP on the untransformed scale show that estimates of the Theil index and the Gini coefficient have very large relative bias compared to M-quantile MC estimates on the same scale. This provides evidence for the robustness properties of the M-quantile estimators.

The results in table 1 also show that the M-quantile MC estimator on the transformed scale (using the same trans-

formation parameter as the one for the EBP) competes very well with the EBP on the same scale. Here, we acknowledge that using a transformation in conjunction with the M-quantile estimator is not done in an optimal way -as in the case with EBP- and should be used only for initial comparisons of the results on the transformed scale. More research is needed for developing data-driven transformations for the M-quantile methods.

Generally speaking, these results show that the M-quantile-based methods perform well both on the untransformed and transformed scales. Using a transformation appears to improve the results of the M-quantile MC further but as we mentioned above this requires additional research. The EBP method is only considered on the transformed scale for the reasons we described above. These results illustrate the robustness properties of the M-quantile-based methods.

Finally, the M-quantile CD (the results are available from the authors) estimator performs similarly in terms of relative bias to the M-quantile MC on the raw scale (6.2% average relative bias for the Theil index and 2.4% for the Gini coefficient). In terms of relative root MSE the M-quantile CD shows more variability than the M-quantile MC for the Theil index (average relative MSE of 50.2%) and competes well with the M-quantile MC for the Gini coefficient (average relative root MSE of 22.1%). Moreover, further improvement of the M-quantile CD estimators could be obtained using an influence function for the residuals in (9) and (7) as suggested in Chambers et al. (2014).

The M-quantile MC and CD both provide an alternative to the EBP in those cases where the mixed model assumptions are not met. Although the theory of the M-quantile CD is better understood, the M-quantile MC is computationally simpler and faster to implement. For these reasons practitioners may prefer this approach.

Table 1: Design-based simulation results. Average and median of the relative bias (%) and relative empirical root MSE.

	Transform	Theil		Gini	
		Relative bias %			
		Median	Average	Median	Average
M-quantile MC	No	-1.8	8.4	-1.9	1.7
M-quantile MC	Box-Cox	9.4	9.8	5.3	6.2
EBP	Box-Cox	3.9	5.2	7.0	6.4
		Relative root mean squared error %			
M-quantile MC	No	25.1	31.6	20.8	21.2
M-quantile MC	Box-Cox	20.8	20.7	10.7	13.2
EBP	Box-Cox	22.4	23.0	11.7	12.8

5.2 Empirical evaluation of the mean squared error estimator

As concerns the estimation of the MSE, we evaluate the bootstrap estimator (14) using the same data as in the design-based simulation, but limited number of runs, equal to 250, given the high computational time required. We use 1

bootstrap population ($B = 1$) from which we draw 100 bootstrap samples. We draw residuals from the smooth error distribution function unconditionally to the areas (for further details on this technique see Marchetti et al. (2012)).

Due to the long computational time required, we select a sub-set of the population, namely, the provinces of Pisa, Lucca and Massa, which correspond to the North-West of Tuscany. Therefore, there are a total of nine domains, three age groups by three provinces. We study the performance of the bootstrap estimator (14) by computing the relative bias (RB)

$$RB(\widehat{MSE}(\hat{\tau}_i)) = H^{-1} \sum_{h=1}^{H} \frac{\widehat{MSE}(\hat{\tau}_{i,h}) - MSE(\hat{\tau}_i)}{MSE(\hat{\tau}_i)},$$

where $\widehat{MSE}(\hat{\tau}_{i,h})$ is the MSE bootstrap estimate of the target parameter $\hat{\tau}_{i,h}$ in area i and simulation h and $MSE(\hat{\tau}_i)$ is the empirical MSE of estimator $\hat{\tau}_i$ (which we consider as the "true" MSE) computed over 1000 Monte Carlo simulations. We also show a summary of empirical MSEs and estimated MSEs for checking if the bootstrap estimator tracks well the empirical (true) MSE over domains.

The results are summarized in table 2 and figure 2. Table 2 shows the average and median across the 9 small domains of the relative bias (RB) of the bootstrap MSE estimator for the Theil index and Gini coefficient M-quantile CD and MC estimators. The average RB is around 7% for the Theil index, while for the Gini coefficient M-quantile CD estimator is -15.4%. The average RB of the Gini coefficient M-quantile MC estimator is about 35%. This high value is mainly due to a high bias in three areas (indeed, the median RB is about 0%), where the presence of big outliers affects the MC method. Looking at both the median and the average of the relative bias (RB) of the M-quantile MC we can see that the distribution of the RBs is skewed both for the Theil index and the Gini coefficient. Also the RB related to the M-quantile CD of the Theil index is skewed, while it seems not to be skewed for the Gini coefficient. However, given the small number of areas used in the simulation due to computational time, it is hard to properly assess the quality of the proposed bootstrap MSE estimators. Considering the limited number of bootstrap populations generated the performance of the MSE estimator is judged to be acceptable for practical purposes. Moreover, since the values of the root MSE are small, a small difference has a big impact in relative terms. We also studied the convergence of the bootstrap MSE estimator for the M-quantile MC. More specifically, we computed the median of the difference between the estimated MSE and the "true" (empirical) MSE while increasing the number of bootstrap replications. The results seem to indicate a small negative biased value for the Theil index, which remains constant after 50 bootstrap replications and a bias that tends to zero for the Gini coefficient as the number of boostrap iterations increase. The results reported here are from a design-based simulation that uses real data. Model-based simulations assessing the properties of the bootstrap MSE estimator (not reported here but available upon request to the authors) show markedly better results.

Table 2: Design-based simulation bootstrap MSE estimator results. Average and median across domains of relative bias (%) of the bootstrap MSE estimator.

	Theil		Gini	
	Median	Average	Median	Average
M-quantile CD	-24.3	6.7	-19.0	-15.4
M-quantile MC	-22.3	7.6	-0.6	35.2

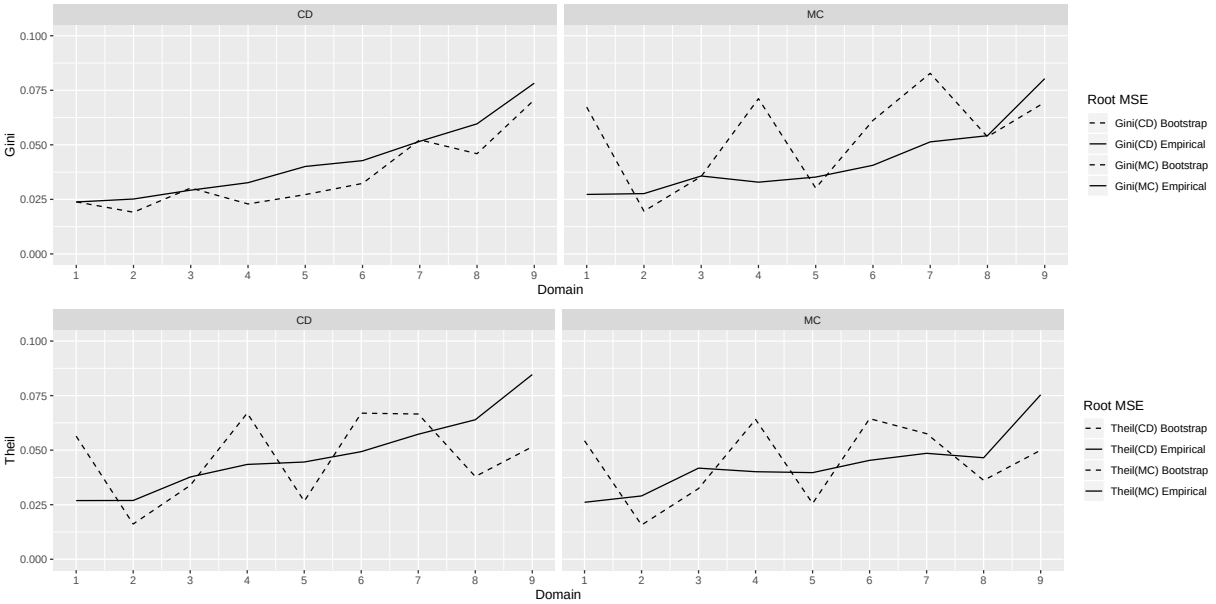

Figure 2: Design-based simulation bootstrap MSE estimator results. Empirical (true) root MSE and estimated root MSE.

From the results in figure 2 we can see that the estimated root MSE tracks reasonably the empirical MSE both for the M-quantile CD and MC estimators of Theil index and Gini coefficient.

6 Estimating the Gini coefficient and the Theil index for small domains in Tuscany

In this section we present an application of the proposed methodology, to EU-SILC (Statistics on Income and Living Conditions) data from Italy. A short description of the design was given in section 5.

The aim is to study the differences in the inequality, if any, among age groups within provinces and provinces within age groups. The domains are defined by the cross-classification of provinces in Tuscany by the age class of the head of the household, leading to a total of 30 domains (10 provinces $\times$ 3 age categories). The age of the head of the household has been divided into three categories, "up to 34", "35-64", "65 and above". This classification comes from the age classes used by ISTAT in some labor force statistics reports, for example in `https://www.istat.it/it/files/2017/07/CS_Occupati-e-disoccupati_giugno_2017.pdf`. To evaluate inequality we estimate both the Gini coefficient (1) and the Theil index (2) to see whether or not they result in estimates of inequality that point in the same direction.

Throughout the paper we refer to the age class "up to 34" as *Young*, "35-64" as *Worker* and "65 and above" as *Aged*. The domain-specific sample size varies between 4 households (Young in Grosseto) and 207 households (Worker in Firenze) with an average sample size across domains of 46.9. The population size is about 1.39 millions households, it varies between 7329 (Young in Massa) and 201019 (Worker in Firenze) with an average of 46280 households per domain. The sampling fraction across domains is between 0.05% (Young in Grosseto) and 0.22% (Young in Pistoia), with an average of about 0.11%, which approximately correspond to the overall sampling fraction in the EU-SILC in Italy.

The outcome we model is the household equivalised disposable income which is available for each sampled household from the EU-SILC survey 2008. The household equivalised disposable income corresponds to the total household net income (the sum of households' member income after tax payments and social transfers, including pensions) divided by the equivalised household size, which gives a weight of 1.0 to the first adult, 0.5 to other persons aged 14 or over who are living in the household and 0.3 to each child aged less than 14. The explanatory variables are the marital status of the head of the household (four categories, single, married, divorced and widow), the employment status of the head of the household (working/not working), the years of education of the head of the household, the mean house surface (in square meters) at municipality level (LAU 2 level) and the number of household members. These covariates are available both from the EU-SILC and from the Population Census of Italy in 2001. Although the 2008 EU-SILC data were collected seven years after the Census, the 2001–2007 period (2008 EU-SILC data refers to 2007 income) was one of relatively slow growth and low inflation in Italy, therefore, it is reasonable to assume that there was relatively little

changes in the considered period. It is also important to mention that EU-SILC and Census datasets are confidential. The datasets were provided by ISTAT, the Italian National Institute of Statistics, to the researchers of the SAMPLE project (`http://www.sample-project.eu`) and were analyzed by respecting the confidentiality restrictions.

Figure 3 shows box-plots of the household equivalised income in each of the 30 domains. The box-plots highlight the asymmetry of the income distribution. The box-plots are ordered (ascending) according to the estimated average of the equivalised households income. We can see that, in general and as expected, Young and Aged groups have a lower income than the Worker group, with some exceptions like the Young group in Lucca which has a rather high income while the Worker group in Massa has a low income.

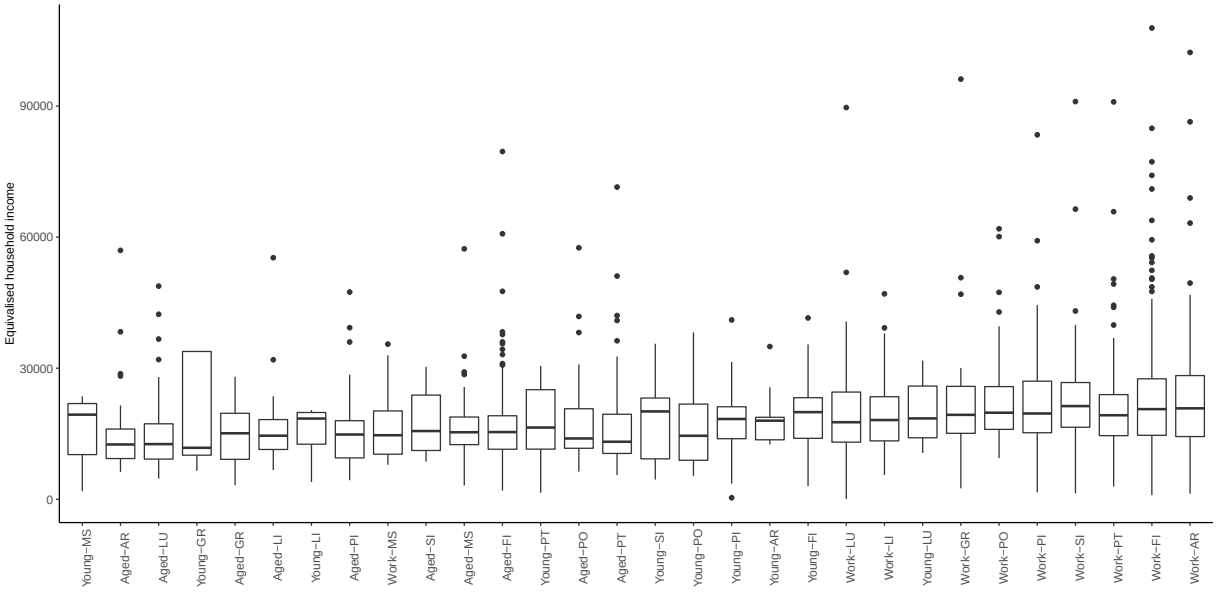

Figure 3: Box-plots of equivalised income by province and age class. Domains are ordered by average income.

Figure 4 shows normal probability plots of level one and level two residuals obtained by fitting a two-level random effects model to the EU-SILC data both on the original scale outcomes (top) and log scale outcomes (bottom). Households are the level one units and the 30 domains define the level two units. Figure 4 suggests departures from the normality assumptions of level one errors, also for the log scale model. The use of the Shapiro and Wilk (1965) test statistic confirms that the hypothesis of normally distributed level one residuals, both when using the original and log-transformed income variable, is rejected. It may be appropriate in this case to use a small area estimation approach that imposes less strict parametric assumptions and it is robust to outliers.

Using the test statistic proposed by Bianchi et al. (2018) we test how close the domain-specific quantile coefficients are to 0.5. This test statistic is trying to emulate the test for the statistical significance of the random effects variance under

the nested regression model. If the test statistic indicates statistically significant differences in the domain M-quantile coefficients, then the model that allows for domain-specific M-quantile coefficients should be preferred to a model that assumes a common M-quantile coefficient leading to a synthetic estimator. The Bianchi et al. (2018) test statistic has been applied to our data. The value of the test statistic is equal to 62.146 and the p-value is equal to 0.000331. The results show that for this application the domain M-quantile coefficients are statistically different from 0.5 and as a result using an M-quantile model with domain-specific M-quantile coefficients should be preferred in this case.

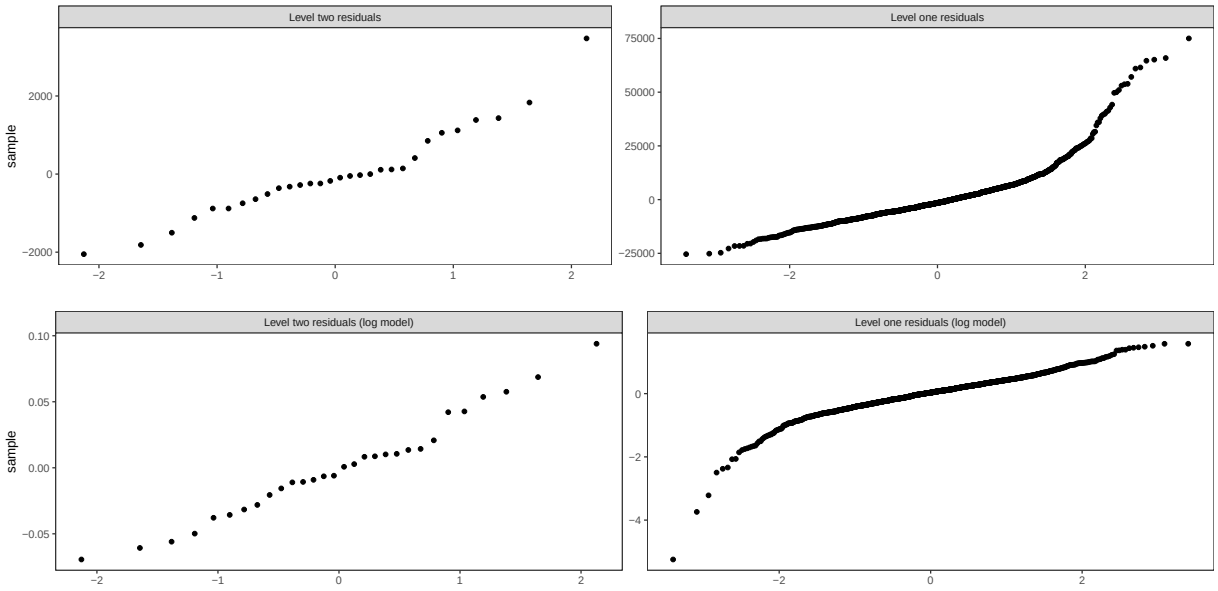

Figure 4: Q-Q plots of level one and two residuals, row scale (top) and log scale (bottom).

We estimate the Theil index and the Gini coefficient using direct, M-quantile CD and MC estimators (for M-quantile CD estimators we use (22), (23) and (24) in the Appendix because it is not possible to link the sampled units with the population units). Comparing these three different point estimates within each domain we observe that the M-quantile CD and MC estimates follow the same trend as the direct ones. The point estimates are shown in figure 5.

Small area estimates of the Theil index and the Gini coefficient obtained by the M-quantile MC approach are summarised in table 3. Both indices vary between provinces within each age group, and also vary between age groups within each province. In particular, the between province variation of the point estimates of the Theil index within the age groups is lower for the Aged group compared to the Young and Worker groups. The between province variation of the point estimates of the Gini coefficients is lower for the Aged group compared to the Worker group, which is lower than the Young group. Moreover, according to both inequality indicators the Young group shows a lower inequality compared to Worker and Aged groups. The same conclusions are reached by looking at the M-quantile CD estimates.

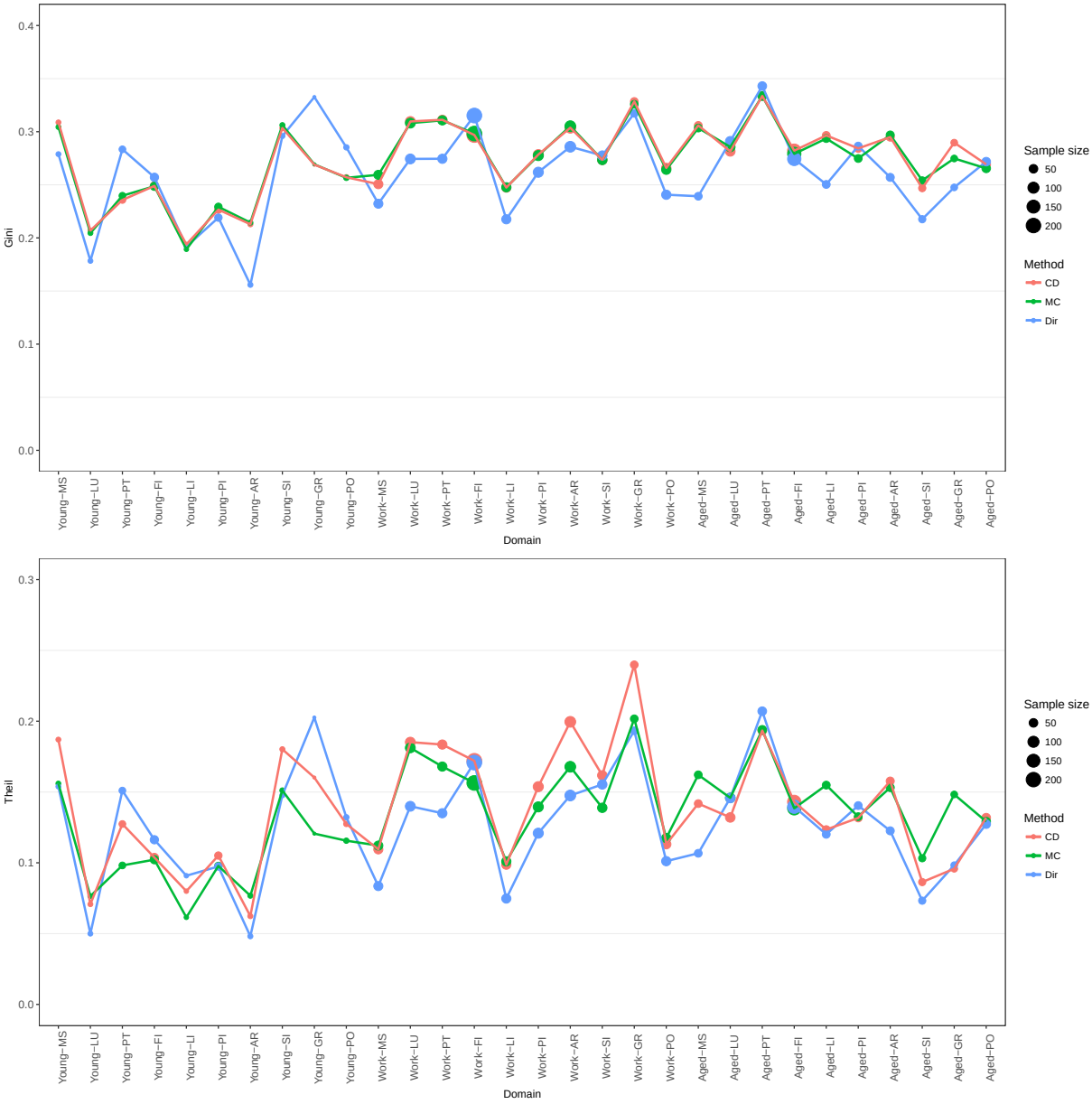

Figure 5: Point estimates of the Gini coefficient estimates (upper plot) and Theil index estimates (lower plot).

Finally, even though the two indices are not directly comparable, we can say that both the Gini coefficient and the Theil index show similar levels of inequality.

Table 3: Small area estimates of Theil index and Gini coefficient (M-quantile MC approach) by provinces and age groups.

	Theil MC			Gini MC		
	Young	Work	Aged	Young	Work	Aged
MS	0.156	0.112	0.162	0.305	0.259	0.304
LU	0.076	0.181	0.146	0.205	0.308	0.286
PT	0.098	0.168	0.194	0.239	0.311	0.334
FI	0.102	0.156	0.138	0.249	0.298	0.279
LI	0.062	0.101	0.155	0.190	0.248	0.294
PI	0.098	0.139	0.132	0.229	0.277	0.274
AR	0.077	0.168	0.153	0.215	0.305	0.297
SI	0.151	0.139	0.103	0.306	0.273	0.254
GR	0.120	0.202	0.148	0.269	0.326	0.275
PO	0.116	0.117	0.129	0.257	0.264	0.266

The results of table 3 seem reasonable in the level and in the direction among the age groups. One result that can be highlighted is the remarkable difference of the level of inequality between the Work and the Aged group in the province of Grosseto (GR) and Livorno (LI). These two results are somehow unexpected. Indeed, we can accept a small reduction or increase of the inequality between Worker and Aged group, but not as big as for the Grosseto and Livorno cases. Moreover, Grosseto and Livorno are quite similar provinces for many aspects, from an economic point of view Grosseto and Livorno are among the medium-income provinces in Italy. Nevertheless, we observed an increase in the inequality of about 20 percentage points of the Gini coefficient and of about 50 percentage points of the Theil index in Livorno and a decrease of about 15 percentage points of the Gini coefficient and about 27 percentage points of the Theil index in Grosseto. We think that these figures need to be further investigated, making use of other indicators – such as poverty indexes, income/consumption distributions, GDP level, etc. These estimates should help socio-economic analysts to better describe local phenomena.

Estimates of the MSE for the M-quantile CD and MC estimates have been obtained using $B = 50$ bootstrap populations and $L = 100$ bootstrap samples (from each population, for a total of 5000 samples). The residuals to generate the populations have been drawn from a smooth distribution unconditional to the areas both for the CD and MC estimators. The choice of the number of bootstrap populations and bootstrap samples has been discussed in Marchetti et al. (2012). The bootstrap resampling scheme we propose is time consuming, however, non-optimized R code run on 2.6GHz quad-core Intel Core i7 took about 260 minutes for the Gini M-quantile CD, 280 minutes for the Theil M-quantile CD and 1100 minutes for the Gini and Theil M-quantile MC. Therefore, we judge the method to be feasible for many applications. Estimates of the standard error of the direct estimates of the Gini coefficient and the

Theil index have been obtained by bootstrap techniques. In particular, we obtained the standard error estimates of the Gini coefficient direct estimates using the bootstrap method proposed by Alfons and Templ (2013), available in the R package "laeken" (R Development Core Team, 2013; Alfons and Templ, 2013), and the standard error estimates of the Theil index direct estimates using the semiparametric bootstrap method proposed by Davidson and Flachaire (2007). The estimated variability of direct, M-quantile CD and MC estimates are summarized in table 4. Both the proposed small area estimators show a gain in efficiency compared to the direct estimators.

Table 4: Estimated MSE summarized across domains.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
T^{CD}	0.008	0.016	0.024	0.027	0.034	0.072
T^{MC}	0.006	0.012	0.018	0.020	0.024	0.055
T^{Dir}	0.043	0.080	0.099	0.097	0.109	0.146
G^{CD}	0.010	0.017	0.022	0.027	0.032	0.069
G^{MC}	0.008	0.015	0.020	0.024	0.028	0.065
G^{Dir}	0.017	0.031	0.035	0.040	0.044	0.093

7 Conclusions

In this work we presented robust small area estimators based on the M-quantile regression model for the Theil index and the Gini coefficient, two popular inequality measures. M-quantile based estimators are robust versus outliers, which occur frequently on income and consumption data that are often used in socio-economic studies to compute inequality measures. For both the measures of interest we presented two estimating approaches: one based on the Monte Carlo approach and one based on the Chambers and Dunstan (1986) distribution function estimator extended for M-quantile models. The proposed estimators have been applied to EU-SILC data from Tuscany (an Italian region) combined with population Census micro data. The aim of the application was to compare the two inequality measures for provinces by age groups (30 domains in total). Results show that the two inequality indicators go to the same direction, pointing out different levels of inequality among provinces within age groups and vice versa. Moreover, we showed that the proposed methods succeed in improving the estimation efficiency compared to direct estimation. Finally, we evaluated the statistical properties of the proposed estimators as well as their bootstrap mean squared error estimators by means of a design-based Monte Carlo simulation. The proposed methodologies to estimate the Theil index and the Gini coefficient for small domains under a robust framework can be applied widely. The possibility to obtain sound estimates of inequality at a low aggregation level, breaking down domains and geographical areas, provides a valuable tool for socio-economic studies.

Future works may focus on analytic mean squared error estimation of the proposed estimators, and bootstrap based confidence intervals.

Appendix

Theil index

The M-quantile CD estimator of the Theil index in area i is defined as $\hat{T}_i = \frac{\hat{\nu}_i^{CD}}{\hat{\mu}_i^{CD}} - \log \hat{\mu}_i^{CD}$, $\hat{\mu}_i^{CD}$ is derived in (8). In what follows we show how to obtain $\hat{\nu}_i^{CD}$. First, an estimator of $E[g(y)]$ using the CD approach is:

$$\begin{aligned}
E[g(y)] &= \int_{-\infty}^{+\infty} g(t)\, d\hat{F}_i^{CD}(t) \\
&= N_i^{-1} \int_{-\infty}^{+\infty} g(t)\, d\Big\{ \sum_{j\in s_i} I(y_{ij} \le t) + n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} I(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j\in s_i} \int_{-\infty}^{+\infty} g(t)\, dI(y_{ij} \le t) \\
&+ n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} \int_{-\infty}^{+\infty} g(t)\, dI(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j\in s_i} g(y_{ij}) + n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} g(\hat{y}_{ki} + e_{ij}) \Big\}. \qquad (15)
\end{aligned}$$

Then, the CD estimator of $\nu_i = E[y \log(y)]$ follows directly

$$\begin{aligned}
\hat{\nu}_i^{CD} &= \int_{-\infty}^{+\infty} t \log(t)\, d\hat{F}_i^{CD}(t) \\
&= N_i^{-1} \int_{-\infty}^{+\infty} t \log(t)\, d\Big\{ \sum_{j\in s_i} I(y_{ij} \le t) + n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} I(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j\in s_i} \int_{-\infty}^{+\infty} t \log(t)\, dI(y_{ij} \le t) \\
&+ n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} \int_{-\infty}^{+\infty} t \log(t)\, dI(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j\in s_i} y_{ij} \log(y_{ij}) + n_i^{-1} \sum_{j\in s_i} \sum_{k\in r_i} (\hat{y}_{ik} + e_{ij}) \log(\hat{y}_{ik} + e_{ij}) \Big\}. \qquad (16)
\end{aligned}$$

Let us show that (9) is unbiased by using a first order Taylor expansion. Consider that $\hat{T}_i^{CD}$ is a function of the random variables (estimators) $\hat{\mu}_i^{CD}$ and $\hat{\nu}_i^{CD}$, and let us write $\hat{T}_i^{CD} = g(\hat{\nu}_i^{CD}, \hat{\mu}_i^{CD})$. Now let us expand function g using a first order Taylor series around point (ν_i, μ_i)

$$g(\hat{\nu}_i^{CD}, \hat{\mu}_i^{CD}) = \frac{\nu_i}{\mu_i} - \log(\mu_i) + \frac{1}{\mu_i}(\hat{\nu}_{ii}^{CD} - \nu_i) - \frac{\nu_i}{\mu_i^2}(\hat{\mu}_i^{CD} - \mu_i) - \frac{1}{\mu_i}(\hat{\mu}_i^{CD} - \mu_i) + O(n^{-1}). \qquad (17)$$

If $\hat{\nu}_i^{CD}$ and $\hat{\mu}_i^{CD}$ are model-unbiased estimators of the parameters ν_i and μ_i the expectation of $g(\hat{\nu}_i^{CD}, \hat{\mu}_i^{CD})$ is

$$E[\hat{T}_i^{CD}] = E[g(\hat{\nu}_i^{CD}, \hat{\mu}_i^{CD})] \approx E\Big[\frac{\nu_i}{\mu_i} - \log(\mu_i) + \frac{1}{\mu_i}(\hat{\nu}_i^{CD} - \nu_i) - \frac{\nu_i}{\mu_i^2}(\hat{\mu}_i^{CD} - \mu_i) - \frac{1}{\mu_i}(\hat{\mu}_i^{CD} - \mu_i)\Big]$$
$$= \frac{\nu_i}{\mu_i} - \log(\mu_i) = T_i\,. \tag{18}$$

Gini coefficient

The estimator $\hat{\Delta}_i^{CD}$ used in (12) is derived as follows

$$\begin{aligned}
\hat{\Delta}_i^{CD} &= \int\int |t_1 - t_2|\, d\hat{F}_i^{CD}(t_1)\, d\hat{F}_i^{CD}(t_2) \\
&= \int N_i^{-1} \int |t_1 - t_2|\, d\Big\{ \sum_{j\in s_i} I(y_{ij} \le t_1) \\
&+ n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} I(\hat{y}_{ik} + e_{ij} \le t_1)\Big\}\, d\hat{F}_i^{CD}(t_2) \\
&= \int N^{-1}\Big\{ \sum_{j\in s_i} \int |t_1 - t_2|\, dI(y_{ij} \le t_1) \\
&+ n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} \int |t_1 - t_2|\, dI(\hat{y}_{ik} + e_{ij} \le t_1)\Big\}\, d\hat{F}_i^{CD}(t_2) \\
&= \int N_i^{-1}\Big\{ \sum_{j\in s_i} |y_{ij} - t_2| + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} |\hat{y}_{ik} + e_{ij} - t_2|\Big\}\, d\hat{F}_i^{CD}(t_2) \\
&= N_i^{-2} \int \Big\{ \sum_{j\in s_i} |y_{ij} - t_2| + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} |\hat{y}_{ik} + e_{ij} - t_2|\Big\} \\
&\times d\Big\{ \sum_{j\in s_i} I(y_{ij} \le t_2) + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} I(\hat{y}_{ik} + e_{ij} \le t_2)\Big\} \\
&= N_i^{-2}\Big\{ \sum_{j\in s_i}\sum_{l\in s_i} |y_{ij} - y_{il}| + n_i^{-2} \sum_{j\in s_i}\sum_{k\in r_i}\sum_{l\in s_i}\sum_{h\in r_i} |\hat{y}_{ik} + e_{ij} - (\hat{y}_{ih} + e_{il})|\Big\}\,.
\end{aligned} \tag{19}$$

Let us show that (12) is unbiased by using a first order Taylor expansion. Consider that $\hat{G}_i^{CD}$ is a function of the random variables (estimators) $\hat{\mu}_i^{CD}$ and $\hat{\Delta}_i^{CD}$, and let us write $\hat{G}_i^{CD} = g(\hat{\Delta}_i^{CD}, \hat{\mu}_i^{CD})$. Now let us expand function g using a first order Taylor series around point (Δ_i, μ_i):

$$g(\hat{\Delta}_i, \hat{\mu}_i) = \frac{\Delta_i}{2\mu_i} + \frac{1}{2\mu_i}(\hat{\Delta}_i - \Delta_i) - \frac{\Delta_i}{2\mu_i^2}(\hat{\mu}_i - \mu_i) + O(n^{-1})\,, \tag{20}$$

then let compute the expectation of $\hat{G}_i^{CD} = g(\hat{\Delta}_i, \hat{\mu}_i)$ under the assumptions that $\hat{\Delta}_i$ and $\hat{\mu}_i$ are model-unbiased

$$E[\hat{G}_i^{CD}] = E[g(\hat{\Delta}_i, \hat{\mu}_i)] \approx E\Big[\frac{\Delta_i}{2\mu_i} + \frac{1}{2\mu_i}(\hat{\Delta}_i - \Delta_i) - \frac{\Delta_i}{2\mu_i^2}(\hat{\mu}_i - \mu_i)\Big] = \frac{\Delta_i}{2\mu_i} = G_i\,. \tag{21}$$

Estimator when linkage between sampled units and population units is not possible

When linkage between sampled units and population units is not possible, that is the set r is unidentifiable, then the prediction is carried out for all the units in the population $U_i = \{s_i \cup r_i\}$. Then the estimators of μ_i, ν_i and Δ_i are as follows

$$\hat{\mu}_i^{CD} = N_i^{-1}\Big[\sum_{j\in U_i} \hat{y}_{ij} + (1-f_i)\sum_{j\in s_i} e_{ij}\Big] \tag{22}$$

$$\hat{\nu}_i^{CD} = N_i^{-1}\Big\{n_i^{-1}\sum_{j\in s_i}\sum_{k\in U_i}(\hat{y}_{ik}+e_{ij})\log(\hat{y}_{ik}+e_{ij})\Big\} \tag{23}$$

$$\hat{\Delta}_i^{CD} = N_i^{-2}\Big\{n_i^{-2}\sum_{j\in s_i}\sum_{k\in U_i}\sum_{l\in s_i}\sum_{h\in U_i}|\hat{y}_{ik}+e_{ij}-(\hat{y}_{ih}+e_{il})|\Big\}. \tag{24}$$

References

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The r package laeken. *Journal of Statistical Software 54*(15), 1–25.

Battacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics 137*, 674–707.

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

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica 42*, 27–41.

Box, G. and D. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B 27*(2), 211–252.

Breckling, J. and R. Chambers (1988). M-quantiles. *Biometrika 75*(4), 761–771.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 47–69.

Chambers, R. and Dunstan (1986). Estimating distribution function from survey data. *Biometrika 73*, 597–604.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. *Biometrika 93*(2), 255–68.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association 81*(396), 1063–1069.

Cowell, F. and K. Kuga (1981). Inequality measurement: An axiomatic approach. *Journal of Economic Theory 15*, 287–305.

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

Davidson, R. and E. Flachaire (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics 141*(1), 141–66.

Deltas, G. (2003). The small-samples bias of the gini coefficient: results and implications for empirical research. *The Review of Economics and Statistics 85*, 226–34.

Diallo, M. S. and J. N. K. Rao (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics 45*(4), 1092–1116.

Dongomo-Jiongo, V. and P. Nguimkeu (2018). Bootstrapping mean squared errors of robust small-area estimators: Application to the method-of-payments data. Technical report, Staff Working Paper 18-28, Bank of Canada.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica 71*(1), 355–364.

Elbers, C. and R. van der Weide (2014). *Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality*. The World Bank.

Foster, J. (1983). An axiomatic characteriazation of the tehil measure of income inequality. *Journal of Economic Theory 31*, 105–121.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. *Econometrica 52*, 761–766.

Gershunskaya, J. and P. Lahiri (2010). Robust small area estimation using a mixture model. In *Proceedings of the Joint Statistical Meeting 2010*.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. In *Atti del Regio Istituto Veneto di Scienze Lettere ed Arti*.

Graf, M., J. M. Marín, and I. Molina (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST 28*(2), 565–597.

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

Langel, M. and Y. Tillé (2013). Variance estimation of the gini index: revisiting a result several time published. *Journal of the Royal Statistical Society Series A 7*, 521–40.

Lombardía, M., W. González-Manteiga, and J. Prada-Sánchez (2003). Bootstrapping the chambers-dunstan estimate of finite population distribution function. *Journal of Statistical Planning and Inference 116*, 367–388.

Maasoumi, E. (1986). The measurement and decomposition of multi-dimensional inequality. *Econometrica 54*, 991–97.

Marchetti, S., N. Tzavidis, and M. Pratesi (2012). Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics and Data Analysis 56*(10), 2889–2902.

Mills, J. and S. Zandvakili (1997). Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics 12*(2), 133–50.

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

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 183*(1), 121–148.

Schmid, T., N. Tzavidis, R. Münnich, and R. L. Chambers (2016). Outlier robust small area estimation under spatial correlation. *Scandinavian Journal of Statistics 43*(3), 806–826.

Shapiro, S. and M. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 67*, 215–216.

Shorrocks, A. (1980). The class of additively decomposable inequality measures. *Econometrica 48*, 613–625.

Sinha, S. and J. Rao (2009). Robust small area estimation. *The Canadian Journal of Statistics 37*(3), 381–399.

Theil, H. (1967). *Economics and Information Theory*. Chicago: Rand McNally and Company.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small area means and quantiles. *Australian and New Zeland Journal of Statistics 52*(2), 167–186.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (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.

Wu, C. and R. Sitter (2001). Variance estimator for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics 29*.

Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica e Applicazioni 4*, 3–27.