sciendo

Economics and Business Review, Vol. 7 (21), No. 2, 2021: 97-114
DOI: 10.18559/ebr.2021.2.7

Small area quantile estimation based on distribution function using linear mixed models[1]

Tomasz Stachurski[2]

Abstract: In economic studies researchers are often interested in the estimation of the distribution function or certain functions of the distribution function such as quantiles. This work focuses on the estimation quantiles as inverses of the estimates of the distribution function in the presence of auxiliary information that is correlated with the study variable. In the paper a plug-in estimator of the distribution function is proposed which is used to obtain quantiles in the population and in the small areas. Performance of the proposed method is compared with other estimators of the distribution function and quantiles using the simulation study. The obtained results show that the proposed method usually has smaller relative biases and relative RMSE comparing to other methods of obtaining quantiles based on inverting the distribution function.

Keywords: quantile, distribution function, small area estimation, survey sampling, linear mixed model, Monte Carlo simulation.

JEL codes: C15, C83.

Introduction

Nowadays increasing demand for statistical data can be observed. Data are regularly collected to provide sufficient information about considered populations. Such studies are conducted both by official statistics and private research institutes. One of the most cost-effective ways of collecting data is sample surveys. A well-planned and organized sample survey allow inference about population parameters with sufficient accuracy.

Lots of the research in the area of sampling survey methodology is mainly concentrated on the estimation of the mean or the total, rather than median or other quantiles (Särndal, Swenson, & Wretman, 1992). However quantiles are

[1] Article received 18 January 2021, accepted 25 June 2021.

[2] University of Economics in Katowice, College of Management, Department of Statistics, Econometrics and Mathematics, ul. 1 Maja 50, 40-287 Katowice, tomasz.stachurski@ue.katowice.pl, ORCID: 0000-0002-5981-5306.

definitely parameters of practical interest especially in the field of economics and business. Quantiles are commonly used for instance in measuring income distribution and determining poverty lines (Osier, 2009; Vijay & Betti, 2011). Moreover median (quantile of order 0.5) rather than mean is considered to be a more appropriate measure of location for the skewed distribution that researchers often deal with in economic data such as income, expenditures, etc. (Kuk & Mak, 1989).

There are two main approaches in quantile estimation. The first is based on obtaining quantiles as inverses of the cumulative distribution function and the second approach uses direct quantile estimators such as synthetic estimators (Stachurski, 2018). This paper is concerned with estimating the distribution function and getting quantiles in particular subpopulations for which sample sizes are small. In the paper there is presented a generalization of the estimator of the distribution function proposed by Salvati, Chandra and Chambers (2012). The main goal of this paper is to obtain an estimate of the distribution function that allows the estimating of quantiles with better precision. The proposed method is based on the use of auxiliary information that is assumed to be known not only for sampled elements but also for each element of the considered population. The main emphasis of the paper is on the problem of the estimation of the quantiles in small areas. In order to estimate characteristics of small area the linear mixed model is used which belongs to the class of small area models (Rao & Molina, 2015). Not only are linear mixed models used in the small area estimation but they can be also applied for modelling longitudinal data (Mihi-Ramirez, Arteaga-Ortíz, & Ojeda-González, 2019).

The structure of the article is as follows. Section 1 presents the basic concepts of small area statistics. In Section 2 a review of different estimators of the cumulative distribution function is presented. In Section 3 the proposed estimator of the distribution function is described. Section 4 is devoted to the conducted simulation study. This section describes algorithm of the simulation study, used datasets and the model. Discussion of the obtained results is also included. The last section contains an overall summary of the obtained results and includes some limitations of the considered proposition and directions for future research.

1. Small area estimation

Small area estimation is a branch of statistics dealing with methods of gathering data and inference about distinguished subpopulations with small or even zero sample sizes. Subpopulations can be distinguished under many criteria such as: a geographical criterion (regions, municipalities, etc.), socio-demographic (e.g. cohort; Basuki, Widyanti, & Rajiani, 2021), professional status of

individuals, source of income, type of household), in the case of higher education institutions it could be the education area (Mazurek, Korzyński, & Górska, 2019) and in the case of enterprises, for instance in terms of the number of employees, the form of ownership or the type of economic activity. In order to define precisely the term "small area" it is necessary to define a direct estimator. The direct estimator is based only on information of the variable of interest from the domain of interest and the period of interest. A direct domain estimator may also use some auxiliary information. One of the most popular definition of the term "small area" is provided by Rao and Molina (2015). They used this term for "any domain for which direct estimates of adequate precision cannot be produced". The term "small" is related to the small sample size in the domain of interest.

In small area estimation it is quite common to use unit-level models (Tzavidis, Marchetti, & Chambers, 2010). It is assumed that the population Ω of size N can be divided into D mutually exclusive and exhaustive domains Ω_d of size N_d. The sample s in the dth domain is denoted by $s_d = s \cap \Omega_d$ and its size is n_d. One of the most popular models used in small area estimation is the generalized linear mixed model which involves both random and fixed effects. The model follows the assumptions (Rao & Molina, 2015):

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \\ E_\xi(\mathbf{e}) = 0 \\ E_\xi(\mathbf{v}) = 0 \\ D_\xi^2 \begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{cases} \tag{1}$$

where $\mathbf{Y}$ is a vector of the study continuous variable of size $N \times 1$ (in the model-based approach it is assumed that values of the study variable are realizations of a random variable with distribution denoted by ξ), $\mathbf{X}$ is a matrix of size $N \times p$ for fixed effects, $\mathbf{Z}$ is a matrix of size $N \times q$ for random effects, β is an unknown vector of fixed effects of size $p \times 1$, $\mathbf{v}$ is an unknown vector of random effects and $\mathbf{e}$ is a vector of random components of size $N \times 1$. It is also assumed that $\mathbf{v}$ and $\mathbf{e}$ are independently distributed with covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively.

Without limiting the generality, the model (1) can be partition into sampled and non-sampled parts as follows (Rao & Molina, 2015):

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix} \mathbf{v} + \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix} \tag{2}$$

2. Estimation of the distribution function

The distribution function of the vector of the random variable Y at t is defined as a fraction of population elements not exceeding t (Dorfman, 2009):

$$F_N(t) = \frac{1}{N}\sum_{i\in\Omega} I(y_i \le t), \qquad -\infty < t < \infty \tag{3}$$

where $I(u)$ is the indicator function. The distribution function follows the basic properties:
- $F_N(t)$ is a monotone non-decreasing function,
- $F_N(t)$ is a step function with step size N^{-1},
- $0 \le F_N(t) \le 1$.

On the one hand estimation of the $F_N(t)$ for fixed t simplifies to the estimation of the mean of zeros and ones. However usually there is a need to estimate the $F_N(t)$ for more than one and what is more, these estimates have to be coordinated, especially when $F_N(t)$ is estimated in order to get quantiles.

Silva and Skinner (1995) and also Dorfman (2009) present a list of some criteria and desired properties of estimators of the distribution function denoted by $\hat{F}(t)$:
1. $\hat{F}(t)$ is a genuine distribution function. It means that $\hat{F}(t)$ is at least non-decreasing function and $\hat{F}(t)$ satisfies the boundary condition: $0 \le \hat{F}(t) \le 1$;
2. $\hat{F}(t)$ is simple to calculate. For instance obtaining some of estimators including estimators based on the Monte Carlo simulation is definitely time-consuming;
3. $\hat{F}(t)$ is easily invertible to get quantiles. The distribution function is a basic statistic underlying many others (Serfling, 1980). It is often estimated in order to obtain population quantiles;
4. $\hat{F}(t)$ is unbiased or asymptotically unbiased;
5. $\hat{F}(t)$ is consistent – $\hat{F}(t)$ tends to approaches to $F(t)$, as the sample size increases;
6. $\hat{F}(t)$ is an outlier-robust estimator. It is especially important in the case of economic research where outlying data are often encountered (Ren & Chambers, 2003);
7. $\hat{F}(t)$ is efficient. It signifies that the mean square error of should be less than MSE of competing estimators;
8. $\hat{F}(t)$ has a readily formulated variance and an estimator of this variance;
9. $\hat{F}(t)$ is calibrated to any ancillary variable. It is desirable for $\hat{F}(t)$ to approaches $F(t)$ as x approaches y. It means that if the study variable is replaced by one of the auxiliary variables, then the following equation should be satisfied: $\hat{F}(t) = F(t)$;
10. definition of $\hat{F}(t)$ is automatic. This implies that $\hat{F}(t)$ does not require any initial specification of a model formula or bandwidths.

However, it must be said that above criteria are not equally important. Moreover, those criteria are not compatible. For instance aiming for simplicity could negatively affect the effectiveness of an estimator.

3. Estimation of the distribution function in small areas

The paper considers mainly the problem of the estimation the distribution function in small areas. The distribution function of the vector of the random variable Y in dth domain at t is defined as a fraction of dth subpopulation elements not exceeding t (Salvati et al., 2012):

$$F_d(t) = \frac{1}{N_d} \sum_{i \in \Omega_d} I(y_i \le t), \quad -\infty < t < \infty \tag{4}$$

A broad overview of methods used to estimate the distribution function is presented by Dorfman (2009). A convenient estimator of the distribution function at t in the dth subpopulation has the following form (Dorfman, 2009):

$$\hat{F}_d^w(t) = \sum_{i \in s_d} w_i I(y_i \le t) \tag{5}$$

where weights w_i satisfies two conditions: $\wedge_i\, 0 \le w_i \le 1$ and $\sum_{i \in s} w_i = 1$. Estimator (5) follows properties 1-3 mentioned in section 2.

If in estimator (5), $w_i = \dfrac{d_i}{\sum_{i \in s_d} d_i}$ is assumed, then the Hajek estimator of the distribution function at t in the dth subpopulation is obtained (Hajek, 1971; Salvati et al., 2012):

$$\hat{F}_d^{Hajek}(t) = \frac{1}{\hat{N}_d} \sum_{i \in s_d} d_i I(y_i \le t) = \frac{\sum_{i \in s_d} d_i I(y_i \le t)}{\sum_{i \in s_d} d_i} = \frac{\sum_{i \in s_d} d_i z_i}{\sum_{i \in s_d} d_i} \tag{6}$$

where weights d_i are inverses of first order inclusion probabilities: $d_i = \pi_i^{-1}$, $z_i = I(y_i \le t)$. In the case of simple random sampling without replacement the Hajek estimator reduces to the naïve estimator, given by the following formula:

$$\hat{F}_d^{nv}(t) = \frac{1}{n} \sum_{i \in s_d} I(y_i \le t) \tag{7}$$

The variance of (6) and the estimator of this variance are given by the following formulas (Särndal et al., 1992):

$$D^2\left(\hat{F}_d^{Hajek}(t)\right) = \frac{1}{N_d^2} \sum_{i \in \Omega_d} \sum_{j \in \Omega_d} (\pi_{ij} - \pi_i \pi_j) \left(\frac{z_i - F(t)}{\pi_i}\right)\left(\frac{z_j - F(t)}{\pi_j}\right) \tag{8}$$

$$\hat{D}^2\left(\hat{F}_d^{Hajek}(t)\right) = \frac{1}{\hat{N}_d^2} \sum_{i \in s_d} \sum_{j \in s_d} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right)\left(\frac{z_i - \hat{F}_d^{Hajek}(t)}{\pi_i}\right)\left(\frac{z_j - \hat{F}_d^{Hajek}(t)}{\pi_j}\right) \tag{9}$$

where the double summation in formulas (8) and (9) refers to summing elements of the second-order inclusion probabilities matrix, where those probabilities are denoted by π_{ij}.

However, the Hajek estimator given by (6) cannot be used if the sample size in the dth domain equals zero. It implies that the possibilities of application of (6) in small areas estimation are quite limited. The Hajek estimator is often used as a benchmark against which other estimators are compared and those comparisons are usually successful. Empirical results show that the Hajek estimator is substantially biased in small areas due to the fact it does not use any information about a study variable from other domains (Rueda, Martinez, Martinez, & Arcos, 2007).

Chambers and Dunstan (1986) is the groundbreaking paper in the field of the estimation of the distribution function using auxiliary information. In this paper the following heteroscedastic regression model is assumed:

$$Y_i = \beta x_i + v(x_i)\varepsilon_i \tag{10}$$

where β is unknown regression parameter, x_i is the value of the auxiliary variable, $v()$ is the known function of x and errors $\varepsilon_i \sim G(0, \sigma^2)$ are independent and follow a distribution G with mean 0 and variance σ^2. Then assuming model (10) the estimator of the distribution function at t in the population has the following formula (Chambers & Dunstan, 1986):

$$\hat{F}^{CD}(t) = \frac{1}{N}\left[\sum_{i\in s} I(y_i \le t) + \frac{1}{n}\sum_{i\in s}\sum_{j\in r}\left(\hat{\varepsilon}_i \le \frac{t - x_j^T\hat{\beta}}{v_j^{0,5}}\right)\right] \tag{11}$$

Estimator (11) satisfies properties 1, 2 and 9 (Dorfman, 2009). Chambers and Dunstan (1986) made it clear that their approach can be used also for more general model than (10). In order to use the Chambers and Dunstan (1986) approach a generalized linear mixed model given by (1) is assumed. Then the Chambers-Dunstan estimator of the distribution function at t in dth subpopulation is given by the following formula (Salvati et al., 2012):

$$\hat{F}_d^C(t) = \frac{1}{N_d}\left[\sum_{i\in s_d} I(y_i \le t) + \frac{1}{n_d}\sum_{i\in s_d}\sum_{j\in r_d}\left(\hat{y}_j^{EBLUP} + \left(y_i - \hat{y}_j^{EBLUP}\right) \le t\right)\right] \tag{12}$$

where predicted values for non-sampled elements are obtained as follows:

$$\hat{y}_j^{EBLUP} = \mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\boldsymbol{v}} \tag{13}$$

where:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^{\mathrm{T}}\mathbf{V}_{ss}^{-1}\mathbf{Y}_s \tag{14}$$

$$\hat{\boldsymbol{v}} = \boldsymbol{G}\mathbf{Z}_s^{\mathrm{T}}\mathbf{V}_{ss}^{-1}\left(\boldsymbol{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}\right) \tag{15}$$

Estimator (12) is asymptotically unbiased if the model (1) is correctly specified (Salvati et al., 2012).

Salvati and others (2012) proposed the following estimator of the distribution function in function at t in dth subpopulation:

$$\hat{F}_d^{EBPLUP}(t) = \frac{1}{N_d}\left[\sum_{i\in s_d} I(y_i \le t) + \sum_{j\in r_d} I\left(\hat{y}_j^{EBLUP} \le t\right)\right] \tag{16}$$

where $\hat{y}_j^{EBLUP}$ is given by (13).

Moreover, Salvati and others (2012) proposed to use empirical best predictor under Molina and Rao (2010) approach. Molina and Rao (2010) consider the problem of prediction of any function of the study variable $\mathbf{Y}$ denoted by $\theta(\mathbf{Y})$ or shortly as θ. The best predictor (BP) is the predictor which minimizes the mean square error: $MSE(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2$. Hence, the best predictor is given by the following formula:

$$\hat{\theta}^{BP} = E(\theta \mid \mathbf{Y}_s) \tag{17}$$

$\hat{\theta}^{BP}$ can be obtained provided that conditional distribution of $\mathbf{Y}_r|\mathbf{Y}_s$ is known. The conditional distribution of $\mathbf{Y}_r|\mathbf{Y}_s$, which is unknown in practice, depends on vector of unknown parameters $\boldsymbol{\gamma} = [\boldsymbol{\beta}^T, \boldsymbol{\delta}^T]$. If those parameters are replaced by their estimates, the empirical best predictor (EBP) denoted by $\hat{\theta}^{EBP}$ is obtained. Molina and Rao (2010) proposed to calculate $\hat{\theta}^{EBP}$ using the following Monte Carlo approximation:

1. estimate the vector of parameters $\boldsymbol{\gamma}$ using sample data $\mathbf{Y}_s$. As a result vector $\hat{\boldsymbol{\gamma}}$ is obtained;
2. assuming the conditional distribution $\mathbf{Y}_r|\mathbf{Y}_s$ is known, generate L vectors denoted by $\mathbf{Y}_r^{(l)}$ (where $L = 1, 2, \ldots, L$), but $\boldsymbol{\gamma}$ is replaced by $\hat{\boldsymbol{\gamma}}$;
3. form the population vectors denoted by $\mathbf{Y}^{(l)}$, where $\mathbf{Y}^{(l)} = [\mathbf{Y}_s, \mathbf{Y}_r^{(l)}]$ and $l = 1, 2, \ldots, L$;
4. calculate empirical best predictor as follows:

$$\hat{\theta}^{EBP} = \frac{1}{L}\sum_{l=1}^{L}\theta\left(\mathbf{Y}^{(l)}\right) \tag{18}$$

However, generating multivariate normal vectors of $\mathbf{Y}_r$ is time-consuming, that is why Molina and Rao (2010) present also a fast algorithm of comput-

ing EBP for a special case of the linear mixed model called nested error linear mixed model. It is based on the generation of independent univariate normal variables. Moreover, vector of the study variable **Y** can be replaced by a transformed vector: $\mathbf{Y} = \mathbf{T}(\ddot{\mathbf{Y}})$. Then assumptions of the model (1) are made not for the study variable **Y** but for the study variable after transformation $T(.)$. Then the best predictor can be written as follows:

$$\hat{\theta}^{BP} = E\left(\theta\left(T^{-1}(\mathbf{Y})\right)\middle|\ \mathbf{Y}_s\right) \tag{19}$$

Salvati and others (2012) assumed that and then the estimator of the distribution function in dth subpopulation is given by the following formula:

$$\hat{F}_d^{EBP.MR}(t) = \frac{1}{N_d}\left[\sum_{i\in s_d} I\left(y_i \le t\right) + \sum_{j\in r_d} E\left[\, I(\hat{y}_j \le t)\,|\,\boldsymbol{Y}_s\,\right]\right] \tag{20}$$

Salvati and others (2012) applied Molina and Rao (2010) approach to log-transformed data and then the obtained results were transformed into original scale using exponential back-transformation.

4. Proposed estimator of the distribution function

In the paper the plug-in estimator of the distribution function is proposed, which is inspired by the paper of Salvati and others (2012). Chwila and Żądło (2020) consider a plug-in predictor of an any parametric function $\theta = \theta\left(T^{-1}(Y)\right) = \theta\left(T^{-1}[\mathbf{Y}_s^{\mathrm{T}}\ \ \mathbf{Y}_r^{\mathrm{T}}]^T\right)$ defined as:

$$\hat{\theta} = \theta\left(T^{-1}[\mathbf{Y}_s^{\mathrm{T}}\ \ \hat{\mathbf{Y}}_r^{\mathrm{T}}]^T\right) \tag{21}$$

where $T^{-1}(.)$ is a back-transformation and $\hat{\mathbf{Y}}_R$ is a vector of fitted values of the model for non-sampled elements where the dependent variable of the model is the study variable after transformation $T(.)$.

The proposed estimator of the distribution function based on a plug-in predictor is given by the following formula:

$$\hat{F}_d^{PLUG-IN}(t) = \frac{1}{N_d}\left[\sum_{i\in s_d} I\left(T^{-1}(y_i) \le t\right) + \sum_{j\in r_d} I\left(T^{-1}(\hat{y}_j^{EBLUP}) \le t\right)\right] \tag{22}$$

where $\hat{y}_j^{EBLUP}$ is given by (13). The proposed estimator given by (22) is a generalization of estimator (16) presented in Salvati and others (2012) for non-linear mixed models which can be transformed into a linear mixed model.

5. Quantile estimation

The distribution function is a basic statistic that plays an important role in the statistical inference. It can be also used to get quantiles. The τth quantile denoted by q_τ is defined as follows (Dorfman, 2009):

$$q_\tau = \min\left\{t : F(t) \geq \tau\right\} \tag{23}$$

The quantile of order τ is such a value of a variable which divides the whole frequency distribution into two parts such that at least $\tau \cdot 100\%$ of total number units are not greater than q_τ and simultaneously at least $(1 - q_\tau) \cdot 100\%$ units are not less than q_τ.

Quantiles can be obtained by estimating direct estimators (see e.g. Dorfman, 2009) or through inverting estimates of the distribution function. The procedure of achieving quantiles based on estimates of distribution functions is as follows (Dorfman, 2009):

1. obtaining a grid of values $\hat{F}_\star(t_v)$ for $t_1 < t_2 < \cdots < t_v < \cdots < t_{v^*}$ so that values of $\hat{F}_\star(t_v)$ be close to order τ of estimated quantile;
2. the τth quantile is the smallest value that satisfies $q_\tau^* = \min\left\{t : \hat{F}_\star(t_v) \geq \tau\right\}$:

In step 1 the distribution function is not calculated for each value of between minimum and maximum of the study variable. Instead some numerical methods are used. In the paper the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, denoted by L-BFGS-B was used (Byrd, Lu, Nocedal, & Zhu, 1995). L-BFGS-B algorithm extends the limited memory BFGS algorithm belonging to the quasi-Newton methods. It allows nonlinear optimization problems with a simple bound on the variables to be solved. The L-BFGS-B algorithm is implemented in package GoFKernel in R in the function inverse (McGrath, Sohn, Steele, & Benedetti, 2019; R Core Team, 2020).

6. Simulation study

In order to verify the properties of quantile estimators considered in paper the simulation study was conducted using R (R Core Team, 2020). The simulation study allows the illustration of the performance of the quantiles' estimators in the small areas. Their performance is evaluated by computing empirical relative biases:

$$rB = \frac{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_d^b - \theta_d\right)}{\theta_d} \cdot 100\% \tag{24}$$

where $\hat{\theta}_d^b$ is the quantile of estimator of the study in dth domain, θ_d denotes the true value of the quantile in dth subpopulations and B stands for number of iterations in the Monte Carlo simulation. In order to evaluate the estimation accuracy, the empirical relative root mean square errors were computed:

$$rRMSE = \frac{\sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_d^b - \theta_d\right)^2}}{\theta_d} \cdot 100\% \tag{26}$$

In the design-based simulation study, the MU284 population was used. It is the population consisting of 284 Swedish municipalities where the study variable is revenues from municipal taxation in 1985 (in millions of kronor) and the auxiliary variable is the number of municipal employees in 1984 (Särndal et al., 1992). This population was also considered in other simulation studies in which the performance of the quantiles estimators obtained as inverses of estimates of the distribution function were studied (Rueda & Arcos, 2001; Berger & Muñoz, 2015). The population of municipalities was originally divided into eight regions. The division into domains in the simulation study is based on the division into regions. However, region Seven with the smallest number of municipalities was combined with region Eight in order to avoid zero domain sample sizes which makes it impossible to obtain direct estimators of distribution function (which are used as a benchmark in the simulation study). In each of iterations in the simulation study the sample was selected using stratified sampling without replacement where $n = 42 \approx 15\%\ N$. In the simulation the model with random effects for domains was used. Normality of the variable under study was verified using Shapiro-Wilk test ($p = 0.28$).

In the simulation study five quantiles are considered: 0.1 quantile, 0.25 quantile (first quartile), 0.5 quantile (median), 0.75 quantile (third quartile) and 0.9 quantile. In the simulation study there were studied 8 estimators listed in Table 1.

Table 2 presents the values of the empirical relative biases of the considered in the paper estimators of quantiles. In the case of the estimation quantiles in the population, the lowest relative biases are obtained for the EBP.MR_DF and EBP.MR_Q methods which are based on Monte Carlo approximation. Slightly higher values were obtained for the proposed in the paper PLUG-IN_DF method. When it comes to the estimation quantiles in small areas, the best results were obtained again for the EBP.MR_DF and EBP.MR_Q methods. However, for the other methods considered in the analysis the values of empirical relative biases are quite similar. It is worth noting that the plug-in estimator of the distribution function proposed in the paper (PLUG_IN_DF) has slightly lower relative biases compared to the EBLUP. The highest values of the empirical relative biases were obtained for the Hajek estimator. It is a direct estimator

Table 1. Description of the estimators considered in the simulation study

Estimator	Description
HAJEK	Quantiles are obtained as the inverses of the Hajek distribution function estimator given by (6)
CD	Quantiles are obtained as the inverses of the Chambers-Dunstan distribution function estimator given by (12)
EBP.MR_DF	Quantiles are obtained as the inverses of the distribution function estimator given by (20). Number of iterations in Monte Carlo approximation of the empirical best predictor equals L =50
EBP.MR_Q	Quantiles are obtained directly using the Molina and Rao (2010) empirical best predictor given by (18) where the function $\theta(.)$ is τth quantile. Number of iterations in the Monte Carlo approximation of the empirical best predictor equals L = 50
PLUG-IN_Q	Quantiles are obtained directly using a plug-in predictor given by (21) where the function $\theta(.)$ is τth quantile and the model is assumed for the original data without any transformation
PLUG-IN_Q_T	Quantiles are obtained directly using a plug-in predictor given by (21) where the function $\theta(.)$ is τth quantile and the model is assumed for the data after logarithmic transformation
EBLUP	Quantiles are obtained as the inverses of the distribution function estimator given by (16)
PLUG-IN_DF	Quantiles are obtained as the inverses of the proposed distribution function estimator given by (22)

Source: Own elaboration.

and it is not recommended for purposes of small area estimation (analogues results were obtained by Berger & Muñoz, 2015).

The second aspect of the conducted simulation study was the analysis of accuracy of the considered methods. The obtained results are displayed in Table 3. Quantiles obtained as the inverses of the Hajek estimator of the distribution function are the least accurate (analogue results were obtained by Berger & Muñoz, 2015). The rRMSE are values in the order of several dozen percent. The best accuracy is obtained for the EBP.MR_Q estimator, PLUG-IN_Q_T and proposed in the paper – the PLUG-IN_DF estimator.

Compared to results obtained in the other simulation studies (Salvati et al., 2012) there is the same conclusion that EBP.MR has usually better properties than EBLUP. It is also confirmed in this paper. However, it is quite interesting that for every quantile order considered in the simulation study there are no clear differences in relative biases or relative RMSEs. In other papers, e.g. (Salvati et al., 2012) it can be observed that for left-tailed quantiles often huge values of rB or rRMSE are obtained.

Table 2. Simulation relative biases of the estimators of quantiles in the population and seven domains (percentage-wise)

Domain ID	τth quantile	HAJEK	CD	EBP. MR_DF	EBP. MR_Q	PLUG–IN_Q	PLUG–IN_Q_T	EBLUP	PLUG–IN_DF
1st domain	0.1	25.17	–19.18	–10.05	–7.05	–4.39	–6.93	–8.26	**–5.66**
	0.25	24.55	3.10	4.91	7.36	7.66	5.61	4.04	**5.99**
	0.5	5.83	–4.49	–2.33	–2.27	–1.98	–2.89	–2.82	**–1.98**
	0.75	–1.71	10.58	7.68	5.76	4.97	5.08	14.09	**13.76**
	0.9	–11.36	8.43	9.23	7.31	5.57	6.19	8.50	**7.81**
2nd domain	0.1	5.75	–8.16	–9.05	–6.59	–5.77	–3.46	–9.33	**–11.69**
	0.25	6.96	–3.51	–5.64	–4.60	–4.17	–3.58	–4.55	**–5.10**
	0.5	20.00	–0.22	–3.73	–3.73	–5.90	–6.30	–6.18	**–5.56**
	0.75	69.86	9.13	7.26	5.67	3.26	2.40	5.04	**6.44**
	0.9	–8.83	8.27	3.49	0.04	–1.04	–0.73	10.24	**9.78**
3rd domain	0.1	12.70	–16.11	–6.91	–4.79	–3.49	–5.09	–6.09	**–4.52**
	0.25	8.59	–11.36	–7.44	–6.78	–5.05	–6.53	–7.24	**–5.71**
	0.5	19.58	–2.65	0.92	0.87	–1.03	–2.16	–1.94	**–0.79**
	0.75	24.71	1.29	3.33	2.45	–0.55	–1.06	2.30	**3.08**
	0.9	0.01	7.37	10.63	4.73	1.69	2.19	5.58	**5.03**
4th domain	0.1	4.48	–13.14	–6.09	–4.74	–2.49	–3.98	–4.86	**–3.29**
	0.25	9.50	–3.81	–2.37	–1.91	–4.56	–5.56	–5.56	**–4.56**
	0.5	9.33	–2.36	–0.91	–0.81	0.21	–1.21	–1.21	**0.21**
	0.75	19.13	6.80	5.33	4.75	5.07	4.14	4.14	**5.07**
	0.9	11.76	25.24	19.32	8.96	1.93	1.75	26.14	**26.06**

5th domain	0.1	7.17	4.34	6.54	8.45	11.54	12.75	8.98	**7.40**
	0.25	3.34	–1.77	–3.22	–2.84	–1.98	–1.82	–4.67	**–5.29**
	0.5	24.24	7.44	1.84	1.43	–3.01	–2.80	–2.80	**–3.01**
	0.75	–1.37	–3.02	–2.78	–3.27	–2.53	–2.73	–2.10	**–1.90**
	0.9	16.61	8.21	4.63	3.44	5.45	5.68	8.85	**8.56**
6th domain	0.1	8.65	0.73	–4.54	–2.38	–2.13	0.47	0.47	**–2.13**
	0.25	3.80	–0.10	–5.39	–4.71	–2.42	–2.38	–2.38	**–2.42**
	0.5	18.49	6.33	1.52	1.38	0.20	–0.40	–0.40	**0.20**
	0.75	9.39	0.26	–0.72	–1.45	0.19	–1.25	–1.25	**0.19**
	0.9	13.08	–5.14	–2.21	–5.85	–9.24	–10.48	–10.48	**–9.24**
7th domain	0.1	12.59	–22.13	–3.74	–1.67	–0.64	–1.90	–4.06	**–2.62**
	0.25	6.66	–5.54	4.87	6.54	11.38	10.90	8.94	**9.42**
	0.5	50.19	18.33	16.22	16.25	14.37	16.25	17.36	**15.63**
	0.75	19.24	5.78	2.66	2.08	2.65	6.83	8.21	**3.96**
	0.9	2.70	3.58	0.14	–2.80	–4.60	0.65	3.68	**–1.75**
The population	0.1	1.96	0.46	0.25	0.40	0.98	1.12	1.12	**–0.44**
	0.25	3.51	0.57	–0.36	–0.36	–2.24	–2.54	–2.54	**0.98**
	0.5	0.59	–2.37	–1.66	–1.85	–3.72	–4.19	–4.19	**–2.24**
	0.75	0.04	3.37	3.00	2.34	1.30	2.48	2.48	**–3.72**
	0.9	1.96	0.46	0.25	0.40	0.98	1.12	1.12	**1.30**

Source: Own elaboration.

Table 3. Simulation relative root mean square errors of the estimators of quantiles in the population and seven domains (percentage-wise)

Domain ID	τth quantile	HAJEK	CD	EBP. MR_DF	EBP. MR_Q	PLUG–IN_Q	PLUG–IN_Q_T	EBLUP	PLUG–IN_DF
1st domain	0.1	79.22	26.26	10.71	8.08	5.29	11.00	11.96	**6.43**
	0.25	64.84	12.12	6.99	8.94	8.92	9.08	8.61	**7.89**
	0.5	37.99	7.79	4.11	4.04	3.65	5.23	5.29	**3.77**
	0.75	29.50	12.43	8.63	7.01	6.30	6.59	15.78	**15.39**
	0.9	28.09	9.60	10.44	8.68	7.18	7.87	10.12	**9.38**
2nd domain	0.1	45.18	16.67	9.60	7.47	6.56	8.92	12.69	**12.18**
	0.25	44.67	8.44	6.32	5.42	4.81	6.44	7.03	**5.68**
	0.5	96.37	9.60	4.77	4.77	6.81	7.50	7.50	**6.85**
	0.75	163.71	12.80	7.93	6.44	4.19	4.06	6.45	**6.89**
	0.9	52.15	9.44	6.48	5.14	3.80	4.34	11.73	**11.09**
3rd domain	0.1	63.92	28.75	8.04	6.49	6.70	10.45	11.17	**7.51**
	0.25	53.39	15.09	8.05	7.40	5.70	8.72	9.29	**6.31**
	0.5	67.06	6.51	3.05	2.80	2.47	3.80	3.99	**2.71**
	0.75	83.79	4.00	4.16	3.52	2.16	2.35	4.40	**4.90**
	0.9	60.33	8.40	11.56	7.26	3.11	3.91	6.56	**5.74**
4th domain	0.1	33.49	19.61	6.65	5.43	3.43	7.06	7.66	**4.00**
	0.25	35.90	8.40	3.25	3.00	5.21	7.11	7.11	**5.21**
	0.5	48.66	5.36	2.72	2.69	2.74	3.16	3.16	**2.74**
	0.75	70.99	7.84	6.61	6.01	6.97	5.99	5.99	**6.97**
	0.9	65.77	25.69	19.70	12.19	3.25	3.43	26.49	**26.33**

5th domain	0.1	30.45	15.54	7.79	9.47	12.43	15.90	13.23	**8.57**
	0.25	31.90	8.61	4.14	3.85	3.62	5.75	7.57	**6.69**
	0.5	62.02	11.44	3.41	3.20	4.02	4.70	4.70	**4.02**
	0.75	39.12	4.21	4.04	4.34	3.53	3.58	3.15	**3.10**
	0.9	53.35	9.14	6.11	5.16	6.59	6.90	9.85	**9.47**
6th domain	0.1	40.75	13.17	5.62	4.38	3.61	9.59	9.59	**3.61**
	0.25	35.13	8.85	6.37	5.78	3.85	6.81	6.81	**3.85**
	0.5	56.05	10.64	3.19	3.22	2.69	4.45	4.45	**2.69**
	0.75	59.00	3.85	3.08	3.47	3.04	3.16	3.16	**3.04**
	0.9	60.41	7.28	4.01	7.52	10.43	11.74	11.74	**10.43**
7th domain	0.1	56.23	32.75	5.81	4.90	4.28	18.00	18.50	**5.00**
	0.25	46.62	16.64	7.34	8.44	12.67	15.23	13.85	**11.46**
	0.5	112.37	23.01	17.28	17.32	15.81	18.50	19.61	**17.45**
	0.75	82.03	6.97	5.52	4.94	5.02	7.53	8.89	**6.04**
	0.9	54.08	5.41	4.63	5.75	6.40	3.43	5.16	**4.83**
The population	0.1	14.69	11.49	3.29	3.04	3.00	7.45	7.45	**3.00**
	0.25	11.40	4.32	1.89	1.94	2.15	4.89	4.89	**2.15**
	0.5	19.35	2.23	1.55	1.58	2.97	3.47	3.47	**2.97**
	0.75	20.01	3.30	2.36	2.48	4.05	4.67	4.67	**4.05**
	0.9	26.84	4.90	3.93	3.42	2.96	4.23	4.23	**2.96**

Source: Own elaboration.

Conclusions

In the paper the problem of quantile estimation is considered. The performance of the proposed method, which is a generalization of EBLUP presented in (Salvati et al., 2012) was compared with other methods of quantile estimation. There are two kinds of competing estimators. The first group is based on inversing the estimates of the distribution function and the second approach is about obtaining quantiles directly using empirical best predictors or a plug-in predictor. The proposed method performs usually better in terms of relative bias and relative RMSE than other estimators based on the estimates of the distribution function. At the same time it covers non-linear models such as the EBP approach but it is faster because it does not require the use of Monte Carlo algorithms to compute estimates. This property is important both for practitioners in the data production process as well as for theoreticians who would like to assess their properties in extensive simulation studies.

The results of the conducted analysis constitute a recommendation for practitioners who deal with the issue of quantile estimation in their research. The issue is particularly important for institutions and organizations dealing with the measurement of poverty and social exclusion, as many poverty measures are based on quantiles, for instance some of the Laeken Indicators (Beil, Kolb, & Münnich, 2011). Moreover, the proposed approach allows for the estimation of quantiles in selected subpopulations with a small or even zero sample size. Quantiles can also be used in other areas especially to estimate the average level for highly asymmetric distributions—which is quite common in economic research.

It needs to be highlighted that the in the conducted analysis only design-based properties of the proposed method are considered. The impact of model misspecification on the accuracy of the presented method has not been taken into account, which may be an area of further research in the future. Future research should also examine the problem of the estimation the variance of the presented quantile estimation method. It is important from the practical point of view because the estimation of the estimator's variance allows the calculation of the standards errors of quantile estimators.

References

Basuki, Widyanti, R., & Rajiani, I. (2021). Nascent entrepreneurs of millennial generations in the emerging market of Indonesia. *Entrepreneurial Business and Economics Review*, *9*(2), 151-165.

Beil, S., Kolb, J. P., & Münnich, R. (2011). *Policy use of Laeken indicators*. (Proceedings of the New Techniques and Technologies for Statistics 2011). Brussels, Belgium. https://doi.org/10.13140/2.1.4027.7764

Berger, Y. G., & Muñoz, J. F. (2015). On estimating quantiles using auxiliary information. *Journal of Official Statistics*, *31*(1), 101-119. https://doi.org/10.1515/JOS-2015-0005

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190-1208. https://doi.org/10.1137/0916069

Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, *73*(3), 597-604. https://doi.org/10.2307/2336524

Chwila, A., & Żądło, T. (2020). On the choice of the number of Monte Carlo iterations and bootstrap replicates in Empirical Best Prediction. *Statistics in Transition*, *21*(2), 35-60. https://doi.org/10.21307/stattrans-2020-013

Dorfman, A. H. (2009). Inference on distribution functions and quantiles. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics. Volume 29B Sample Surveys: Inference and Analysis* (pp. 371-395). Amsterdam: Elsevier.

Hajek, J. (1971) Comment on an essay on the logical foundations of survey sampling by Basu D. In V.P. Godambe & D.A. Sprott (Eds.), *Foundations of statistical inference* (pp. 36). Holt: Rinehart and Winston.

Kuk, A. Y. C., & Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society: Series B (Methodological)*, *51*(2), 261-269. https://doi.org/10.1111/j.2517-6161.1989.tb01763.x

Mazurek, G., Korzyński, P., & Górska, A. (2019). Social media in the marketing of higher education institutions in Poland: Preliminary empirical studies. *Entrepreneurial Business and Economics Review*, *7*(1), 117-133. https://doi.org/10.15678/EBER.2019.070107

McGrath, S., Sohn, H., Steele, R., & Benedetti, A. (2019). Meta-analysis of the difference of medians. *Biometrical Journal*, *69*(2), 69-98.

Mihi-Ramirez, A., Arteaga-Ortíz, J., & Ojeda-González, S. (2019). The international movements of capital and labour: A study of foreign direct investment and migration flows. *Entrepreneurial Business and Economics Review*, *7*(3), 143-160. https://doi.org/10.15678/EBER.2019.070308

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, *38*, 369–385. https://doi.org/10.1002/cjs.10051

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Journal of the European Survey Research Association*, *3*, 167-195. https://doi.org/10.18148/srm/2009.v3i3.369

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. Hoboken: John Wiley & Sons.

Ren, R., & Chambers, R. (2003). *Outlier robust imputation of survey data via reverse calibration*. (S3RI Methodology Working Papers No. M03/19). Southampton: Southampton Statistical Sciences Research Institute.

Rueda, M., & Arcos, A. (2001). On estimating the median from survey data using multiple auxiliary information. *Metrika*, *54*, 59-76. https://doi.org/10.1007/s001840100116

Rueda, M., Martinez, S., Martinez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, *137*(2), 435-448. https://doi.org/10.1016/j.jspi.2005.12.011

Salvati, N., Chandra, H., & Chambers, R. (2012). Model-based direct estimation of small-area distributions. *Australian & New Zealand Journal of Statistics*, *54*(1), 103--123. https://doi.org/10.1111/j.1467-842X.2012.00658.x

Särndal, C. E., Swenson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons.

Silva, N., & Skinner, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics*, *11*(3), 277-294.

Stachurski, T. (2018). A simulation analysis of the accuracy of median estimators for different sampling designs. In L. Vachova, V. Kratochvil (Eds.), *Conference proceedings: 36th International Conference Mathematical Methods in Economics. MME 2018* (pp. 50-514). Praha: MatfyzPress.

Tzavidis, N., Marchetti, S., & Chambers, R. (2010). Robust estimation of small area means and quantiles. *Australian & New Zealand Journal of Statistics*, *52*(2), 167--186. https://doi.org/10.1111/j.1467-842X.2010.00572.x

Vijay, V., & Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, *38*(8), 1549-1576. https://doi.org/10.1080/02664763.2010.515674