sciendo

Journal of Official Statistics, Vol. 38, No. 3, 2022, pp. 823–845, http://dx.doi.org/10.2478/JOS-2022-0036

Measuring and Mapping Micro Level Earning Inequality towards Addressing the Sustainable Development Goals – A Multivariate Small Area Modelling Approach

Saurav Guha[1] *and Hukum Chandra*[1]

The earning inequality in India has unfavorably obstructed underprivileged in accessing elementary needs like health and education. Periodic labour force survey conducted by National Statistical Office of India generates estimates on earning status at national and state level for both rural and urban sectors separately. However, due to small sample size problem, these surveys cannot generate reliable estimates at micro-level viz. district or block. Thus, owing to unavailability of district-level estimates, analysis of earning inequality is restricted to the national and the state level. Therefore, the existing variability in disaggregate-level earning distribution often goes unnoticed. This article describes multivariate small area estimation method to generate precise and representative district-wise estimate of earning distribution in rural and urban areas of the Indian State of Bihar by linking Periodic labour force survey data of 2018–2019 and 2011 Population Census data of India. These disaggregate-level estimates and spatial mapping of earning distribution are essential for measuring and monitoring the goal of reduced inequalities related to the sustainable development of 2030 agenda. They expected to offer insightful information to decision-makers and policy experts for identifying the areas demanding more attention.

Key words: Multivariate small area estimation; earning inequality; SDG; periodic labour force survey; NSO; census.

1. Introduction

The economy of India has developed at verifiably notable degrees and is currently one of the fastest developing nation in terms of economy in the world. India has progressed substantially in reforming its economy, reducing the hardship and fulfilling opportunities for day to day comfort of its vast population. This unceasing economic growth has also led to significant reduction in poverty level and has big impact on the overall wellbeing of the people in this country. However, in spite of these remarkable achievement, the earning distribution in India remains obstinately uneven.

The movement of economy also directs the behavior of the labour market. The volatility in the economy, both in its inter and intra sectored linkages as well as in the framework of economic integration with rest of the world, is reflected in the domestic labour market (MoSPI 2020a). Global economic slowdown creates extreme volatility which can enormously influence the contemporary economic atmosphere. Thus, it is hugely

[1]ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi, India. Emails: saurav.iasri@gmail.com and hchandra12@gmail.com
Acknowledgments: The authors would like to acknowledge the valuable comments and suggestions of the Editor, the Associate Editor and three anonymous referees.

© Statistics Sweden

imperative to quantify its short term impact on labour market which involves the collection of labour force data at regular interval. In India, labour force participation in unorganized sector is much higher as compared to the organized sector. The frequent availability of labour force data was the need of the hour and that led to the launch of Periodic labour force survey (PLFS) in 2017 by National Statistical Office (NSO), Ministry of Statistics and Program Implementation (MoSPI), Government of India. In India, NSO is the primary body to collect PLFS data for generating estimates at the state and the national level for both rural and urban areas. The PLFS data provides estimates for a range of employment and unemployment indicators such as unemployment rate, worker population ratio, labour force participation rate, earning of different working groups. Even though being exceptionally crucial, the estimates of earning distribution are unattainable further down the state level in India for example, district, block or further level of disaggregation.

The rising earning inequality in the world indicates that the richest 10% have up to 40% of global earning whereas the poorest 10% earn only between 2% to 7% and the population growth inequality in developing countries has increased by 11% (UNDP 2015). Regardless of the achievement of Millennium Development Goals (MDGs), several objectives including equality have not been unified. As the primary focus of MDGs was on national and global averages, the growing disparities at the local level remained unchecked. The Post-2015 Development Agenda reckon with the Rio + 20 promises to "strive for a world that is just, equitable and inclusive". The Sustainable Development Goals (SDGs), also known as the Global Goals, were adopted by all United Nations Member States in 2015 as a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity by 2030 (UNDP 2015). Through the pledge to "Leave No One Behind", the SDGs ensure the addition of marginalized and nonempowered groups, and reduction of inequalities together with eradication of poverty. Extreme inequality persists in multiple extents within and between nations and it often creeps into the decision-making processes as well. According to UNCDP (2018), segments of the population that are typically excluded from meaningful participation in decision-making are unlikely to see their interests safeguarded in policy and investment decisions, as is the case for countries that are marginalized in international decision-making structures. The SDG10 among the 17 goals that constitute the 2030 Agenda for SDG aims to "reduce inequality within and among countries". The comprehensive targets to reduce inequality "within" and "among" countries emphasize the significance of inequality reduction by all means in all countries.

Inequality creates barrier to growth and development when it denies people of opportunities which in turn, lead to the state of extreme poverty. There is a growing consensus that economic growth is not sufficient to reduce poverty if it is not inclusive and if it does not involve the three dimensions of sustainable development – economic, social and environmental, (UNDP 2015). Goal 10 aims that the income growth of the bottom 40% of country's population is higher than the national average by the year 2030. The Gini coefficient of income inequality for India fell from 36.8% in 2010 to 33.6% in 2015. The Government of India prioritizes the policy of inclusion, financial empowerment and social security via major initiatives like Jan Dhan Yojona, Aadhaar etc. These comprehensive steps are in line with the SDG targets intended for achieving better equality and

encouraging the socio-economic, and political inclusion of all by 2030. In this present situation, the budding interests of the policy makers, public agencies, scientists, government organizations are focused in attaining the local (or micro) level estimates. The emphasis on disaggregate level SDG indicators by various national and international agencies has further emphasized the inevitable need of such local level estimates. These local level areas or domains, better known as small areas or small domains are formed by cross-classification of several demographic and topographic variables that includes small topographic areas (e.g., districts) or small demographic groups (e.g., land category, social groups, religion, age-sex groups) or cross classifying both (Guha and Chandra 2021a). Besides, in the existing PLFS data of NSO, the small areas or districts may have very small or even zero sample sizes which may lead to large sampling error in case of direct estimation. The small area estimation (SAE) methodology provides a viable and cost effective solution to this problem of small sample sizes (Rao and Molina 2015). The SAE techniques borrow strength from various external sources viz. time periods, areas and so on to obtain precise and reliable estimates.

The key idea behind the SAE methodology is to link the variable under study with the auxiliary information through different statistical models. This defines the model-based small area estimates corresponding to these small areas. Based on the availability of the auxiliary information, the unit level or the area level models are mostly used in SAE. The Fay–Herriot (FH) model (Fay and Herriot 1979) is a widely accepted area level model in SAE as it is better suited to account for complex survey design for point and variance estimates compared to the unit-level model and when the model covariates are available only in aggregate form. The FH model assumes the availability of area specific survey estimates and these estimates follow an area level linear mixed model with area as random effects. Application of the FH model are readily available in literature in multiple dimensions. The uncertainty of the SAE estimates was deliberated by Prasad and Rao (1990), Datta et al. (2011), Fay (1987) and Datta et al. (1991) introduced the multivariate version of the FH model while Benavent and Morales (2016) extended it further. Often, there is a need of estimating correlated processes viz. poverty indicators, unemployment, and so on. Multivariate models often allow for the correlation of several variables and usually suitable in these circumstances. Unlike the FH model, more than one variable of interest is modelled via multivariate Fay–Herriot model (MFH) by allowing for different covariance structure between the vector of the variable of interest and the random effects, see Guha and Chandra (2021b). A number of small area applications for estimating socio-economic indicators have been described in literature based on univariate FH model that ignores the correlation between the target variables, see, for example, Chandra et al. (2011), and references therein. Furthermore, surveys are generally planned to collect information on more than a few variables. In SAE problem, when the target areas comprise of insufficient sample size, taking into account of correlation between the target variables can provide an added advantage in obtaining precise and reliable small area estimates (Rao and Molina 2015). Franco and Bell (2021) also pointed out that precision in bivariate area-level models is only improved if one of the outcomes has very low variance and the correlation between the two outcomes is very strong.

According to the quarterly report of MoSPI (2020b), the unemployment rate in the age group of 15 years and above has sharply increased from 9.1% in January–March 2020 to

20.8% in April–June 2020 with the working population ratio decreased from 43.7% to 36.4%. These figures indicate the severity of the job losses and sufferings faced by the majority of the working population in the country during the first phase of the COVID-19 pandemic. Given the severe economic hardships faced by a large section of the populations during this pandemic, having precise knowledge of district-level estimates of pre-pandemic earning distribution is critical for evaluating the true impact of the disaster. India reported the second highest number of COVID-19 cases for any country in the world (>32 million) by mid-2021, with Bihar contributing to almost 8 million cases (MoHFW 2021). Bihar, the third most populous state in the country, accounts for about 8.6% of India's population with an area of 94163 square km that equals to 2.9% of India's total geographical area. About 33.74% population of the state lives below the poverty line which is higher than the national average of 21.92% (NITI Aayog 2019). According to the Global Hunger Index 2020 (GHI 2020), out of 132 countries India stands on 94th position with an overall score of 27.2. The state of Bihar ranks 27th out of 28 states in "zero hunger" with a score of 31 which is much lower than the national average of 47 (NITI Aayog 2019). In addition, 75.2% of the population is in the lowest two wealth quintiles in Bihar which is highest in India and Bihar ranks 25th out of 28 states in "reduced inequality" parameter with a score of 48 which is much lower than the national average of 67 (NITI Aayog 2019). Therefore, it seems rational to consider Bihar to generate the district level estimates of earning inequalities at rural and urban sectors using SAE techniques. To the best of our knowledge, no prior study has been done to estimate the micro level or disaggregate level earning inequalities in India.

The article is organized as follows. Section 2 describes the data from the 2018–2019 Periodic labour force survey of the NSO of India and the 2011 Population Census of India that will be used to estimate the district level earning distribution in rural and urban sector of the Indian State of Bihar. In this section, we also set out the theoretical background of the area level MFH model, and then discuss the different variant of this model used in estimating small area means under this model. The results obtained from the application of district-level inequalities in earning distribution along with various diagnostic measures are reported in Section 3. We also provide spatial mapping of earning distribution in this section that serves to demonstrate the degree of district-level inequalities in the distribution of earning between rural and urban sector in Bihar. Finally, Section 4 summarizes the article and provides concluding remarks.

2. Data and Model Specification

In this section, we introduce the major data sources utilized in multivariate SAE application. The 2018–2019 PLFS data of the NSO for rural and urban districts of Bihar and the data from 2011 Population Census of India are used for estimating the district level inequality in the earning distribution between rural and urban sector of the state. Afterward, we briefly describe the multivariate SAE methodology applied in the estimation of district level inequality in distribution of average monthly earning.

2.1. Description of PLFS and Population Census Data

The PLFS survey data is freely downloadable from the MoSPI , Government of India (http://mospi.nic.in/). Since 2017, NSO carries out the PLFS every year. In PLFS a

rotational panel sampling design for first visit both in rural and urban areas and three periodic revisit in urban areas has been used while there was no revisit in rural areas. A stratified multistage survey design was adopted, with the final units being households. The 2018–2019 PLFS of NSO is intended to produce precise and reliable estimates at the state and the national level for both rural and urban areas in the country. However at district level, this PLFS data cannot directly be used to generate precise and reliable estimates, since sample size within each district is not adequate to offer district-level estimates with acceptable reliability and precision. Although, district is always being a very crucial part of the planning process in the country, there are no surveys conducted to produce district level estimates in India and this leads to limit the policy interventions at the district or even further lower level (Guha and Chandra 2021b).

The 2018–2019 PLFS data of the NSO comprised 15,441 persons in 3,179 households from the rural and urban areas in 38 districts of Bihar. For all the districts, the sample size ranges from 07 to 205 with an average of 92 for rural areas while for urban areas, it is 06 to 362 with an average of 45. This survey provides information on earning estimate of every person separately for rural and urban areas in the state of Bihar. Districts included moderately small sample sizes with an average sampling fraction of 0.000046 for rural and 0.00027 for urban areas. Due to the problem of small sample size, it is barely possible to obtain precise and reliable direct estimates at district level which eventually leads to producing large standard errors from this survey, see Chandra et al. (2011) and Rao and Molina (2015). In this paper, we made an attempt to address the problem of small sample size in attaining district-level estimates from the 2018–2019 PLFS data. The multivariate small area method has been implemented to tackle this problem by including related auxiliary variables from the Population Census 2011 of India (Census 2011).

We have considered the following information on earning from employment from PLFS 2018–2019 viz. (a) self-employed persons, (b) salaried employees and regular wage earner, and (c) person working as casual labour. For salaried employees and regular wage earner in current weekly status (CWS), information on earnings in the previous calendar month was collected. For self-employed persons in CWS, information on earnings in the last 30 days from the self-employment was collected. It is important to note that average gross earnings from the self-employment activity have been calculated by excluding those self-employed persons who had reported earning as zero or not reported. For the person working as casual labour (except public works), information on earnings was collected for the casual labour work in which the person was engaged for each and every day of the reference week that is, last seven days prior to the date of the survey. For the sake of the analysis, we have transformed the daily data into monthly data for the casual labour work. The estimates in this section are derived using the data collected in the first visit schedules in the rural areas (since there was no revisit in rural areas) and for the urban areas using the data collected in the schedules of first visit and the corresponding revisits conducted during the four quarters of the survey period, viz., during July to September followed by October to December in 2018 and January to March followed by April to June in 2019. For more detailed information on the method of the data collection, readers may refer to the annual report of the PLFS 2018–2019 (MoSPI 2020a). The target variables in the 2018–2019 PLFS data are Y_1: average monthly earning (in Rs.) of a person from employment in rural areas (hereafter denoted by Rural-Earning), Y_2: average monthly

earning (in Rs.) of a person from employment in urban areas (hereafter denoted by Urban-Earning). This paper targets to estimate the inequality in average monthly earning of a person in rural and urban districts of Bihar at small area level through joint modelling of the target variables that is, Rural-Earning and Urban-Earning.

2.2. *Multivariate Small Area Modelling*

Let us consider that the population consists of M domains and let there are R target variables in this study. All the way through, a subscript $m(m = 1, \ldots, M)$ is used to denote the quantities possess by domain m and a subscript $r(r = 1, \ldots, R)$ is used to index the variable r under study. Assume a finite population Ω of size N comprises M non-overlapping domains $\Omega_m; m = 1, \ldots, M$ and a sample s of size n is drawn from Ω by any probability sampling design. We also assume that the domain size N_m is known for each domain and n_m units are selected in the sample from N_m units of m^{th} domain (hereafter denoted by small area). The population total is given by $N = \sum_{m=1}^{M} N_m$ and the corresponding sample size is $n = \sum_{m=1}^{M} n_m$. Let y_{mrj} be the value corresponding to j^{th} unit of the r^{th} target variable in m^{th} area, $r = 1, \ldots, R, j = 1, \ldots, N_m$ and $m = 1, \ldots, M$. The aim is to estimate small area mean $\bar{Y}_{mr} = N_m^{-1} \sum_{j \in \Omega_m} y_{mrj}$, $r = 1, \ldots, R$ and $m = 1, \ldots, M$. The traditional direct survey estimator (hereafter denoted by Direct) for $\bar{Y}_{mr}$ is given by $\bar{y}_{mr} = \sum_{j=1}^{n_m} \tilde{w}_{mrj} y_{mrj}$ with $\tilde{w}_{mrj} = w_{mrj} / \sum_{j=1}^{n_m} w_{mrj}$ where $\tilde{w}_{mrj}$ is the normalized survey weight for j^{th} unit of the r^{th} variable in m^{th} area. In addition, $\tilde{w}_{mrj}$ satisfies $\sum_{j=1}^{n_m} \tilde{w}_{mrj} = 1$ with w_{mrj} being the survey weight for j^{th} unit of the r^{th} variable in m^{th} area. Following Särndal et al. (1992), the estimated variance of Direct estimator is approximated by $v(\bar{y}_{mr}) = \sum_{j=1}^{n_m} \tilde{w}_{mrj}(\tilde{w}_{mrj} - 1)(y_{mrj} - \bar{y}_{mr})^2$. Under simple random sampling without replacement (SRSWOR), $w_{mrj} = 1/\pi_{mrj}$ where $\pi_{mrj} = n_m/N_m$ is the inclusion probability for j'$^{\text{th}}$ unit of the r^{th} variable in the m^{th} area.

Let us further assume that $y_{mr}(r = 1, \ldots, R)$ be an unbiased direct survey estimates of an unknown population parameter (e.g., the population mean) Y_{mr} of the target variable r for small area m. Let x_{mr} be a p_r -vector of available auxiliary variables corresponding to area m that are associated to the population mean Y_{mr} for the variable r under study. Usually, area-specific auxiliary information are acquired from some available secondary sources, for example, administrative registers, the population census, etc. We denote $\mathbf{y}_m = (y_{m1}, \ldots, y_{mr})^T$, a vector of direct survey estimates of $\boldsymbol{Y}_m$ where $\boldsymbol{Y}_m$ is the m-vector population mean of target variables. In line with Benavent and Morales (2016), an area-level FH model (Fay and Herriot 1979) used for more than one target variables is given by

$$\mathbf{y}_m = \boldsymbol{Y}_m + \boldsymbol{\varepsilon}_m \text{ and } \boldsymbol{Y}_m = \mathbf{X}_m \boldsymbol{\beta} + \mathbf{u}_m. \tag{1}$$

In literature, the model in Equation (1) is time and again referred to the multivariate form of the FH model. The MFH model in Equation (1) consists of two stages, the first one takes care of the sampling variability of the direct survey estimates $\mathbf{y}_m$ of true area means of the target variable $\boldsymbol{Y}_m$ while the second stage accounts for linking of the true area means of the target variable $\boldsymbol{Y}_m$ to $\mathbf{X}_m = diag\,(\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mR})_{R \times p}$, a matrix of available auxiliary variables where $p = \sum_{r=1}^{R} p_r$. This model in Equation (1) can be denoted as an area level random effect model as

$$\mathbf{y}_m = \mathbf{X}_m\boldsymbol{\beta} + \mathbf{u}_m + \boldsymbol{\varepsilon}_m, \quad m = 1, \ldots, M. \tag{2}$$

Here $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_r')_{p\times 1}'$ and $\boldsymbol{\beta}_r$ is a p_r - vector of unknown fixed effect parameters. The vector of ind, random area effects $\mathbf{u}_m$ are independent and identically distributed with $\mathbf{u}_m$ *ind* $N\left(0,\ \mathbf{V}_{u_m}\right)$ while vectors of sampling errors ε_m are independent and normally distributed with $\boldsymbol{\varepsilon}_m \sim N(0, \mathbf{V}_{\varepsilon_m})$. Moreover, these two vector of errors $\mathbf{u}_m$ and $\boldsymbol{\varepsilon}_m$ are independent of each other within and between areas with $\mathbf{V}_{\varepsilon_m}$, the covariance matrices of ε_m are known while the covariance matrices of $\mathbf{u}_m$ denoted by $\mathbf{V}_{u_m}$ depend on unobservable parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R)$. Combining M-area-level models, the model in Equation (2) can be denoted in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{y} = \text{col}\ (\mathbf{y}_m; 1 \le m \le M)$ is the vector of direct estimates of order $MR \times 1$, $\mathbf{X} = \text{col}\ (\mathbf{X}_m; 1 \le m \le M)$ is the matrix of known covariates of dimension $MR \times p$, $\mathbf{u} = \text{col}\ (\mathbf{u}_m; 1 \le m \le M)$ is the vector of random area effects of dimension $MR \times 1$ and $\boldsymbol{\varepsilon} = \text{col}(\boldsymbol{\varepsilon}_m; 1 \le m \le M)$ is the vector of sampling errors of dimension $MR \times 1$ with $\mathbf{u} \sim N(0, \mathbf{V}_u)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V}_\varepsilon)$. $\mathbf{Z} = \text{col}'(\mathbf{Z}_m; 1 \le m \le M)$ is a matrix of order $MR \times MR$ whose mth column $z_m, m = 1, \ldots, M$, is an indicator variable which takes the value 1 if a unit belongs to an area m and zero otherwise. Especially, in model (3) $\mathbf{Z}$ is an $MR \times MR$ identity matrix. Furthermore, it is supposed that the random area effects u are independently distributed of the sampling errors ε where $\mathbf{u} \sim N(0, \mathbf{V}_u)$ and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{V}_\varepsilon)$. The random effects covariance matrix is denoted by $\mathbf{V}_u = \text{diag}\ (\mathbf{V}_{um}; 1 \le m \le M)$ and $\mathbf{V}_\varepsilon = \text{diag}\ (\mathbf{V}_{\varepsilon m}; 1 \le m \le M)$ is the matrix of design variances.

Next we consider three types of the model (3) to obtain model-based small area estimates. First, we take $\mathbf{V}_{u_m} = \text{diag}\ (\sigma_{ur}^2;\ 1 \le r \le R)$, $\mathbf{V}_{\varepsilon_m} = \text{diag}\ (\sigma_{\varepsilon mr}^2;\ 1 \le r \le R)$, $m = 1, \ldots, M$ and $\sigma_{\varepsilon mr}^2$'s are known for the estimator based on univariate FH model (UFH). Second estimator, denoted by MFH-1, is based on MFH model with $\mathbf{V}_{u_m} = \text{diag}$ $(\sigma_{ur}^2; 1 \le r \le \text{R})$, $m = 1, \ldots, M$, and a known but not necessarily diagonal matrix $\mathbf{V}_\varepsilon$. The third estimator, denoted by MFH-2, is also based on MFH model where the random effects $\mathbf{u}_m = (\mathbf{u}_{m1}, \ldots, \mathbf{u}_{mR})'$ is generated via a first order heteroscedastic autoregressive HAR(1) process $\mathbf{u}_{mr} = \rho\mathbf{u}_{mr-1} + \boldsymbol{\tau}_{mr}$ with $\mathbf{u}_{m0}$–$N(0, \sigma_0^2)$, $\boldsymbol{\tau}_{mr}$–$N(0, \sigma_2^2)$, $r = 1, \ldots, R$ and σ_2^2, $\mathbf{u}_{m0}$, $\boldsymbol{\tau}_{mr}$ are independent. The components of $\mathbf{V}_{u_m}$ are given by $\sigma_{umii} = \sum_{k=0}^{i} \rho^{2k}\sigma_{1-k}^2$ and $\sigma_{umij} = \sum_{k=0}^{|j-i|} \rho^{2k+|j-i|}\sigma_{|j-i|-k}^2, i \ne j$. It is further assumed that sampling errors are not independent with each other that is, $\mathbf{V}_\varepsilon$ is known but not essentially a diagonal matrix. For UFH and MFH-1 estimators, the number of unknown variance component parameters to be estimated is equal to R with $\theta_r = \sigma_{ur}^2$, $r = 1, \ldots, R$ and for MFH-2, it is $R + 1$ with $\theta_r = \sigma_{ur}^2$, $r = 1, \ldots, R$ and $\theta_{R+1} = \rho$. Under the model (3), $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $Var(\mathbf{y}) = \mathbf{V}_y = \mathbf{V}_u + \mathbf{V}_\varepsilon = diag(\mathbf{V}_{ym;} 1 \le m \le M)$, with $\mathbf{v}_u = \mathbf{z}'\mathrm{v}_u\mathbf{z}$ and $\mathbf{V}_{ym} = \mathbf{V}_{um} + \mathbf{V}_{sm}$, $m = 1, \ldots, M$. Here, the covariance matrix $\mathbf{V}_y$ depends on R and $R + 1$ unknown variance component parameters given by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_R)$ for UFH and MFH-1 and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{R+1})$ for MFH-2 model respectively. The restricted maximum likelihood (REML) method is applied to estimate θ. Replacing the estimated values $\hat{\theta}$ of parameters $\boldsymbol{\theta}$ in $\mathbf{V}_u$ to obtain $\hat{\mathbf{V}}_u = \mathbf{V}_u(\hat{\theta})$ and $\hat{\mathbf{V}}_y = \hat{\mathbf{V}}_u + \mathbf{V}_\varepsilon$, the multivariate version of empirical best linear unbiased predictors (EBLUP) of $\boldsymbol{Y}$ is defined as

$$\hat{\boldsymbol{Y}}_{\mathrm{MFH}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}. \tag{4}$$

Here, the empirical best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the EBLUP of $\mathbf{u}$ are obtained $\hat{\boldsymbol{\beta}} = (\mathbf{X}', \hat{\mathbf{V}}_y^{-1}\mathbf{X})^{-1}\ \mathbf{X}'\hat{\mathbf{V}}_y^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \hat{\mathbf{V}}_u\mathbf{Z}'\hat{\mathbf{V}}_y^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ respectively. In small area applications, the mean squared error (MSE) estimates are desirable to measure the precision of estimates and also to construct the confidence interval for the estimates (Guha and Chandra 2021b). The analytical MSE estimate of EBLUP of MFH in Equation (4) is obtained following Benavent and Morales (2016).

3. Results and Discussions

3.1. Variable Selection and Model Fitting

We used Population Census 2011 (Census 2011) data of India for selection of suitable covariates for small area modelling. As these covariates are available as counts at district level, area-level multivariate small area models were used in this analysis to obtain the small area estimates. There are almost 30 auxiliary variables accessible from the census data and we did some exploratory analysis prior to selection of appropriate covariates for multivariate small area model fitting. Principal Component Analysis (PCA) was also applied to obtain combined scores for some particular sets of auxiliary variables. The PCA variables (i.e., combined scores obtained using PCA) in place of Population Census variables are frequently used in SAE method since maximum variation is explained with dimension reduction of auxiliary variables, see, for example Johnson et al. (2010) and Chandra et al. (2011). In particular, we did PCA independently on three sets of auxiliary variables for both Rural-Earning and Urban-Earning and all measured at district level. For Rural-Earning and Urban-Earning, these three sets of auxiliary variables are noted as PC1, PC2 and PC3 below. The first set (PC1) included the literacy rates by gender and proportions of worker population by gender, the second set (PC2) included the proportions of main worker by gender, proportions of main cultivator by gender and proportions of main agricultural labourers by gender and the third set (PC3) included the proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. Table 1 describes the proportion of explained variability by these principal components in all the three sets of auxiliary variables for both the target variables. This followed by a stepwise regression for choosing significant auxiliary variables based on Akaike information criterion value.

Initially, the direct estimates of two target variables that is, Rural-Earning and Urban-Earning are plotted to get an impression about the correlation between them. From Figure 1, it seems that these two target variables Rural-Earning and Urban-Earning are loosely correlated. Here it should be noted that, for Urban-Earning, there was 05 non-sample districts and for Rural-Earning, there was one non-sample district. Consequently, we fit a FH model separately to both the target variables with the sample areas and then a synthetic estimation (see Chandra et al. 2011) is carried out to estimate the non-sample areas. MSE of the synthetic estimates are obtained following Chandra et al. (2011). In most of the cases, the estimates of error variance are extremely noisy and need to be smoothed before being used

Table 1. Principal components and explained variations for the target variable rural-earning and urban-earning.

Variable	Set	Principal component	Proportion of variance	Cumulative proportion
	PC1	R-PC11	0.958	0.958
		R-PC12	0.041	1.000
Rural-earning	PC2	R-PC21	0.953	0.953
		R-PC22	0.027	0.980
	PC3	R-PC31	0.963	0.963
		R-PC32	0.037	1.000
	PC1	U-PC11	0.915	0.915
		U-PC12	0.085	1.000
Urban-earning	PC2	U-PC21	0.714	0.714
		U-PC22	0.217	0.931
	PC3	U-PC31	0.807	0.807
		U-PC32	0.193	1.000

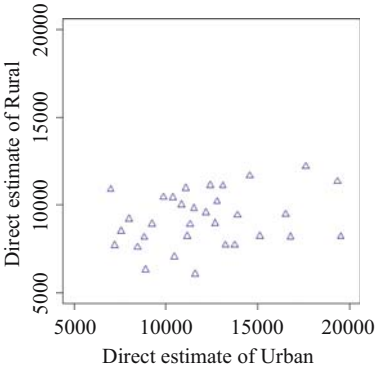


Fig. 1. Scatter plot of the direct estimates of rural-earning and urban-earning.

in the analysis. In line with Schall (1991), we applied a generalized variance function (GVF) method for smoothing the estimated error variance of the direct estimates. The details of the GVF method is described in the supplementary materials. Next, we proceed with the MFH-2 model described in the previous section using a combination of the direct estimates and the synthetic estimates corresponding to the sample and the non-sample areas as the input of the two target variables and some selected covariates from the census data as suitable auxiliary variables. Finally, three significant covariates viz. proportions of marginal cultivator population (MCP), proportions of marginal agriculture labour population (MALP) and R-PC31 corresponding to the target variable Rural-Earning and for Urban-Earning, four significant covariates viz. proportions of main worker population (MWP), proportions of marginal cultivator population (MCP), U-PC21 and U-PC31 are included in the model. The regression parameter estimates are reported in Table 2 for the two dependent variable Rural-Earning and Urban-Earning. Observing the signs of the regression parameters estimates, it can be concluded that rural districts having lesser proportion in all the three covariates have more earning while urban districts having greater proportion U-PC21 and U-PC31 and lesser proportion in MWP and MCP have more earning.

Table 2. Regression parameters, standard error and p-values for the target variables rural-earning and urban-earing.

Variable	Parameter	Estimate	Standard error	P-value
Rural-Earning	Intercept	21,906	4,835	< 0.001
	MALP	-14,987	5,553	< 0.01
	R-PC31	-5,429	1,907	< 0.01
Urban-Earning	Intercept	45,139	12,169	< 0.001
	MWP	-38,588	14,644	< 0.01
	MCP	-45,798	19,477	0.018
	U-PC31	7,699	3,499	0.021

The values of the estimates from fitting the multivariate small area model using 2018–2019 PLFS data are described as follows. The estimate of variance component parameters for the MFH-2 model are $\hat{\sigma}_{u1} = 1317$, $\hat{\sigma}_{u2} = 2448$ and $\hat{\rho} = 0.478$. We also test the null hypothesis $H_0 : \sigma^2_{ui} = \sigma^2_{uj},\ i,j = 1,2,\ i \neq j$ against the alternative hypothesis $H_1 : \sigma^2_{ui} \neq \sigma^2_{uj}$. The test statistic is given by $t_{ij} = \hat{\sigma}^2_{ui} - \hat{\sigma}^2_{uj}\ /\sqrt{v_{11} + v_{22} - 2v_{12}};\ i,j = 1,2,\ i \neq j$, where $v_{rs}, r,s = 1,2$ are the elements of the inverse of the matrix of Fisher information corresponding to MFH-2 model calculated at $\hat{\theta} = \left(\hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}, \hat{\rho}\right)$. The value of the test statistic is given by $t_{12} = -2.056$ (0.039) with p-value given in parenthesis. As the value of the test statistic is significant at 5% level, this leads to the conclusion that variance of random area effects for Rural-Earning and Urban-Earning are significantly different. This followed by testing $H_0 : \rho = 0$ with the test statistic $t_\rho = \hat{\rho}/\sqrt{v_{33}}$ and the value of $t_p = 1.071(0.284)$, p-value is in parenthesis. This reveals that the correlation between the two target variables is not significantly different from zero and we proceed with MFH-1 model with diagonal covariance matrix of random effects instead of MFH-2 model. It is important to note that, although the variance of random area effects for Rural-Earning and Urban-Earning are significant at 5% level, they are not significant at 1% level and also the correlation between the two target variables is not significantly different from zero. This leads to the almost identical results in univariate and multivariate estimates which established the idea reported in Franco and Bell (2021) that that precision in multivariate area-level models is only improved if one of the outcomes has very low variance and the correlation between the two outcomes is very strong. Finally, the MFH-1 model is applied with all the significant auxiliary variables to obtain the earnings estimates that is, Rural-Earning and Urban-Earning for all the districts in Bihar.

3.2. Diagnostic Measures

In what follows, we described some standard diagnostic measures to examine the model assumptions and inspect the reliability and validity of the generated estimates through MFH method. In line with Brown et al. (2001), two forms of diagnostics viz. (a) the model diagnostics, and (b) the multivariate SAE diagnostics are employed to endorse the model assumptions. The reliability of the model-based estimates of Rural-Earning and Urban-Earning attained by SAE method under MFH-1 model is validated by some additional diagnostics. Corresponding to the target variable Rural-Earning and Urban-Earning, the random effects in MFH-1 model are supposed to follow a normal distribution with zero

mean and constant variance σ_{ur}^2, $r = 1, 2$. If the underlying model assumptions hold, the residuals are supposed to be distributed randomly around zero. We used the normal probability (Q-Q) plots to examine the normality assumption. Q-Q plots of district level residuals corresponding to the two target variables Rural-Earning and Urban- Earning are given in Figure 2. In addition, we also examined the normality assumption of the random area effects via Shapiro-Wilk test and the p-values of the test are 0.321 and 0.457 for Rural-Earning and Urban-Earning respectively. Furthermore, it is evident from the Q-Q plot in Figure 2 that the normality assumption holds while p-values of the Shapiro-Wilk test are greater than 0.05 and both of these evidences taken together indicate that the district wise random area effects are likely to be distributed normally. Next we evaluated the validity and the reliability of the small area estimates by some frequently used diagnostics. In line with Brown et al. (2001) and Chandra et al. (2011), model-based small area estimates should be (1) consistent with unbiased direct survey estimates and (2) more efficient than direct estimates in terms of MSE. The subsequent measures e.g., the bias diagnostic, goodness of fit (GoF) diagnostic, the percentage coefficient of variation (CV) and the 95 percentage confidence interval (CI) are selected. Later, we classified the measurements of CV and CI as internal diagnostic measures as these indicate the efficiency of the small area estimates. Moreover, for benchmarking, a calibration diagnostic is also applied in which the model-based small area estimates are combined to an upper level so that these estimates can be compared with direct estimates at that higher level and we classified this as an external diagnostic measure. It is important to note that in this case, the direct estimates are in survey weighted form.

3.2.1. Bias Diagnostic

The bias diagnostic measure tests the validity while the precision of the model-based estimates is examined by the CI and CV. Following Chandra et al. (2011), the bias diagnostic is performed. Being unbiased of the population values, the regression of the direct estimates on the true population values likely to be linear with the identity line. If the model-based estimates are close to these true values of the population, the regression of direct estimates on model-based small area estimates expected to be similar. Consequently, we plotted the direct estimates and model-based estimates in the *y* and *x-axis* respectively and examined the departure of the small area estimates from the regression line fitted values. The plot given in Figure 3 demonstrates that small area

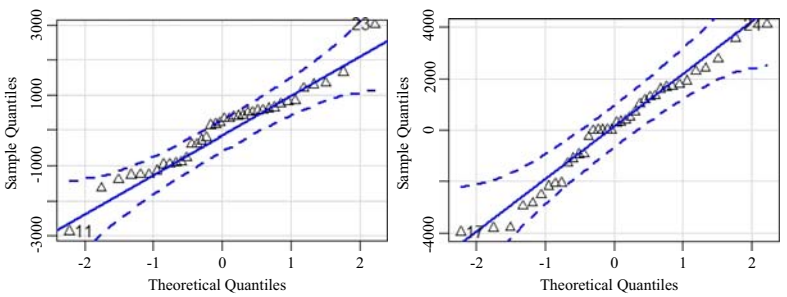


Fig. 2. Normal Q-Q plot of district-level residuals for rural (on the left) and urban (on the right).

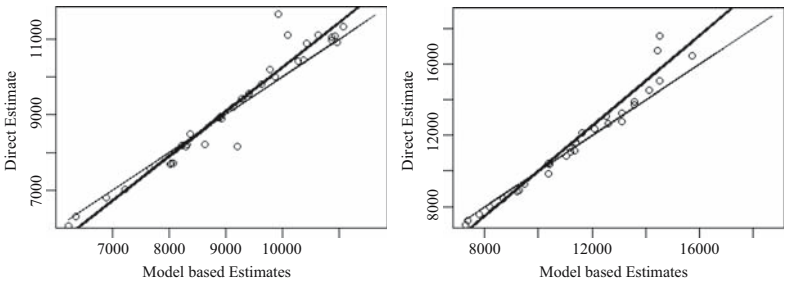


Fig. 3. Bias diagnostic plot with $y = x$ line (thin line) and regression line (solid black line) for rural (on the left) and urban (on the right).

estimates are not as extreme as the direct estimates signifying the usual SAE result of diminishing greater extreme values to the average values. Largely, this diagnostic specify that the small area estimates are expected to be consistent and approximately design unbiased when compared with direct estimates. This is expected as the MFH estimates are realization of random variables and so the regression of the direct estimates on the MFH estimates is unbiased for a test of common expected values. The Goodness of Fit (GoF) diagnostic provides such test. The statistic for GoF diagnostic is obtained using the Wald statistic for MFH estimate,

$$W = \sum_m \left[\frac{(direct\ estimate_m - MFH\ estimate_m)^2}{\hat{\text{v}}\text{ar}(direct\ estimate_m) + \hat{MSE}(MFH\ estimate_m)} \right]$$

The test statistic value is compared against the value from a chi-square distribution with M degrees of freedom. The value of chi square statistic with $M = 37$ degrees of freedom for rural districts is 52.19 and for urban districts is 47.40 with $M = 33$ degrees of freedom at 5% level of significance. We obtained the value of Wald statistic for rural districts as $W = 3.69$ and for urban districts it is $W = 5.10$. A smaller value than the table value indicates no significant difference between the direct and the model-based MFH estimates. This diagnostic results clearly indicate that the MFH estimates are consistent with the direct survey estimates.

3.2.2. Internal Diagnostic

Afterward, the degree of improvement in precision of model-based multivariate small area (i.e., district level) estimates of Rural-Earning and Urban-Earning are examined against the FH and direct survey estimates. Typically small area estimates having smaller CVs are likely to be reliable. The summary of %CVs of the Direct, FH and MFH estimates of Rural-Earning and Urban-Earning are given in Table 3. District specific %CV is demonstrated in Figure 4. The direct survey estimates possess greater CV compared to the FH and MFH estimates of Rural-Earning and Urban-Earning. It is obvious from Table 3 and Figure 4 that direct survey estimates of Rural-Earning and Urban- Earning seem to be highly unstable. It is important to note that, for the target variable Urban- Earning, there was 05 non-sample districts and for Rural-Earning, there was one non-sample district. So for Urban-Earning we first compare the performance with the sampled districts and comparison of non-sample districts are given separately. For Rural-Earning, the CV of

Table 3. Distribution of % CV for the direct and model-based small area estimates of rural-earning and urban-earning.

Values	Rural-earning (37 sample districts)			Urban-earning (33 sample districts)			Urban-earning (05 non-sample districts)		
	Direct	FH	MFH	Direct	FH	MFH	Direct	FH	MFH
Minimum	3.22	3.18	3.18	5.10	5.06	5.06	-	19.46	14.70
Q1	4.74	4.61	4.61	8.64	8.23	8.18	-	23.62	16.95
Median	6.86	6.34	6.34	9.57	9.02	8.93	-	23.68	18.09
Mean	7.77	6.57	6.56	10.37	9.48	9.38	-	23.36	17.23
Q3	8.60	7.93	7.92	10.69	10.22	10.16	-	24.61	18.19
Maximum	21.60	11.81	11.77	24.22	17.08	16.46	-	25.41	18.24

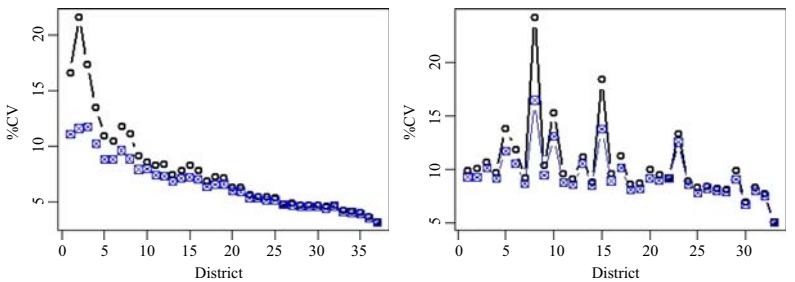


Fig. 4. District specific percentage coefficient of variation (CV) of direct (black, °) and MFH (blue, □) estimate for rural-earning (on the left) and urban-earning (on the right). Districts are arranged in increasing order of sample size.

Direct distributed form 3.22% to 21.60% with a median value of 6.86% whereas it is 3.18% to 11.77% with a median value of 6.34% for MFH which indicate that the MFH estimates are more strongly distributed compared to the direct estimates as the distribution of CVs from MFH estimates are much narrower than that of the direct estimates. Similarly for Urban-Earning, when we compare the Direct with MFH for the 33 sampled districts, the CV of Direct is distributed form 5.10% to 24.22% with a median value of 9.57% whereas it is 5.06% to 16.46% with a median value of 8.93% for MFH. It is important to note that, for sample districts the performance of FH and MFH are almost identical which in turn validate the point about the usefulness of MFH model discussed earlier in the model fitting section. Furthermore, when the performance of the FH and MFH estimates are compared for the non-sampled areas, there is fair improvement in both the rural and urban areas. For the single non-sample district of Rural-Earning, the CV of FH is 15.27% while it is 11.03% for MFH estimate and for Urban-Earning with 05 non-sampled district, the CV of FH is distributed from 19.46% to 25.41% with a median value of 23.68% whereas it is 14.70% to 18.24% with a median value of 18.09% for MFH estimates. As we have mentioned that, the input for the target variables in the multivariate model is a combination of the direct estimate and the synthetic estimate, the MFH model is utilized the extra information from the synthetic estimate apart from the information used by the FH model. It seems clear From Figure 4, with decreasing sample sizes of the districts the relative performance of the MFH estimates for Rural-Earning and Urban-Earning has improved. Accordingly, these precise and reliable MFH estimates generate the district level earning

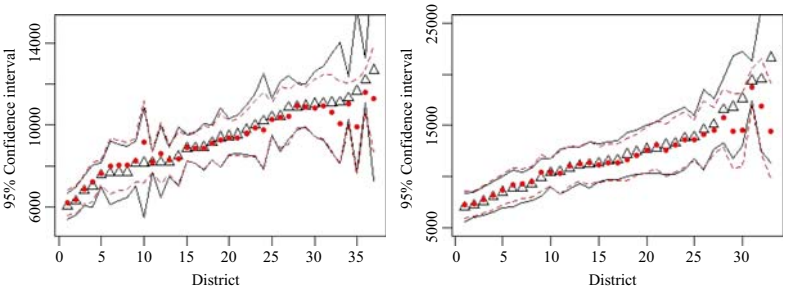


Fig. 5. District-wise 95% nominal confidence interval for direct (black) and MFH (red) estimates for rural-earning (on the left) and urban-earning (on the right). Districts are arranged in increasing order of direct estimates.

estimate much better than direct and FH estimates. The 95% confidence intervals (CIs) are given in Figure 5. The Figure 5 indicates that the CI of MFH estimate is much tighter than the direct survey estimates.

3.2.3. External Diagnostic

The aggregation property of the MFH based district-level SAE estimates at higher aggregation level (viz. state and regional level) are examined. The divisional and state-level estimates of Rural-Earning and Urban-Earning is obtained by

$$\hat{Y}_i = \sum_{j=1}^{M} N_{ij}\hat{Y}_{ij} / \sum_{j=1}^{M} N_{ij},\ i = 1, 2 \text{ and } j = 1, \ldots, M,$$

where $\hat{Y}_{ij}$ denote the MFH estimate of Rural-Earning and Urban-Earning for $i = 1, 2$ and district j and the population size, corresponding to the variable i and district j is N_{ij}. The districts of Bihar are classified in nine divisions and we studied the aggregation property. The state and the divisional-level estimates of Rural-Earning and Urban-Earning are reported in Table 4. While comparing the small area estimates against the direct estimates, it seems that in both the state and the regional level these small area estimates are close enough to the direct estimates.

Table 4. Aggregated estimates of rural-earning and urban-earning obtained from Direct and MFH. Estimates are aggregated over 37 districts for rural areas and 33 districts for urban areas at the state and divisional levels.

Division	Rural-earning		Urban-earning	
	Direct	MFH	Direct	MFH
State	9,149	9,201	14,478	13,971
Patna	9,709	9,841	11,247	11,567
Tirhut	8,754	9,147	11,413	11,365
Saran	8,080	8,180	12,643	11,970
Darbhanga	8,283	8,803	17,342	13,784
Kosi	8,148	8,289	13,740	13,397
Purnia	8,570	8,837	12,562	12,269
Bhagalpur	12,150	10,790	17,552	14,527
Munger	10,710	9,820	17,745	17,188
Magadh	10,722	10,745	11,174	11,369

3.3. Spatial Distribution of Earning Inequality

Table 5 and 6 report the distribution of CV and earning range across all the districts respectively. The district specific direct survey estimates and MFH estimates together with the 95% CI and CV for Rural-Earning and Urban-Earning are given in Table 7 and 8. The spatial maps of earning distribution (in Rs.) by districts for both rural and urban areas are produced for the district level estimates generated by MFH method. Figure 6 displays spatial maps of the MFH estimates of earning for rural and urban areas of Bihar. These spatial mapping assist in describing the magnitude of inequality in earning distribution between the district of rural and urban areas of the state. In case of rural areas, central part of Bihar has lower earning followed by the eastern and western part. For urban areas, the lower earning level exists in eastern part followed by the western and central part. The average monthly earning is ranging from Rs. 6,231 to Rs. 11,634 in rural areas whereas it is Rs. 7,279 to Rs. 18,727 in urban areas. This clearly indicates that there is a huge difference in average monthly earning between rural and urban areas in Bihar. From Table 5, the distribution of CVs indicate the precision of the MFH estimates compared to the direct estimates. In rural areas of Bihar, 08 districts have more than 10% CV of direct estimates while it is only 04 districts for FH and MFH estimates and in case of urban sector, 12 districts have more than 10% CV whereas it is only 09 districts for FH and MFH estimates. Based on the results obtained in Table 5, we further described the distribution of earning range based on precise MFH estimates in Table 6. From Table 6 and Figure 6, MFH estimates also reveal that number of districts having average monthly earning more than Rs. 10,000 is only 12 in rural areas while it is 30 for urban areas. In case of lower earning level, 26 districts in rural areas have average monthly earning of Rs. 10,000 or less whereas it is only 08 districts for urban areas. Table 6 further described that all of the rural areas possess an average monthly earning of Rs. 12,500 or less. However, nearly 40% of

Table 5. Distribution of CV of 37 districts in rural areas and 33 districts in urban areas.

Group	%CV	Rural-earning			Urban-earning		
		Direct	FH	MFH	Direct	FH	MFH
1	$< = 5$	12	12	12	0	0	0
2	5.01–10	17	21	21	21	24	24
3	10.01–15	05	04	04	09	08	08
4	15.01–20	02	0	0	02	01	01
5	$> = 20$	01	0	0	01	0	0

Table 6. Distribution of earning range of 38 districts in rural and urban areas based on MFH estimates.

Group	Earning range (in Rs.)	Rural-earning	Urban-earning
1	$< = 7{,}500$	04	02
2	7,501–10,000	22	06
3	10,001–12,500	12	15
4	12,501–15,000	0	12
5	$> = 15{,}000$	0	03

Table 7. Direct and MFH estimates along with 95% confidence interval (95 % CI) and percentage coefficient of variation (CV) of the target variable rural-earning by district in Bihar.

		District				MFH			
			95% CI				95% CI		
District	Sample size	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Pashchim Champaran	156	8,912	8,188	9,636	4.15	8,899	8,192	9,606	4.05
Purba Champaran	154	10,423	9,565	11,281	4.20	10,275	9,451	11,099	4.09
Sheohar	67	7,038	6,012	8,064	7.44	7,234	6,262	8,207	6.86
Sitamarhi	163	7,588	6,986	8,190	4.05	7,676	7,086	8,266	3.92
Madhubani	56	7,720	6,463	8,977	8.31	8,050	6,882	9,218	7.40
Supaul	34	10,198	7,843	12,553	11.78	9,777	7,932	11,623	9.63
Araria	140	8,223	7,486	8,960	4.57	8,326	7,607	9,044	4.40
Kishanganj	136	9,192	8,347	10,037	4.69	9,124	8,312	9,936	4.54
Purnia	80	8,160	7,063	9,257	6.86	8,298	7,266	9,329	6.34
Katihar	55	8,499	7,066	9,932	8.60	8,382	7,075	9,688	7.95
Madhepura	114	6,053	5,409	6,697	5.43	6,231	5,599	6,863	5.17
Saharsa	135	9,459	8,583	10,335	4.72	9,389	8,548	10,231	4.57
Darbhanga	107	6,806	6,057	7,555	5.62	6,902	6,174	7,630	5.38
Muzaffarpur	92	8,952	7,838	10,066	6.35	8,915	7,869	9,960	5.98
Gopalganj	131	9,559	8,650	10,468	4.85	9,411	8,535	10,287	4.75
Siwan	205	8,207	7,689	8,725	3.22	8,223	7,709	8,736	3.19
Saran	112	6,303	5,627	6,979	5.47	6,368	5,704	7,032	5.32
Vaishali	68	9,414	7,965	10,863	7.85	9,288	7,990	10,585	7.13
Samastipur	83	10,881	9,349	12,413	7.18	10,427	9,077	11,777	6.61
Begusarai	7	8,175	5,511	10,839	16.63	9,198	7,194	11,202	11.12
Khagaria	28	11,121	8,171	14,071	13.53	10,086	8,063	12,109	10.24
Bhagalpur	16	11,673	7,698	15,648	17.37	9,922	7,633	12,210	11.77
Banka	140	12,201	11,078	13,324	4.70	11,634	10,582	12,686	4.61
Munger	137	11,340	10,286	12,394	4.74	11,065	10,068	12,062	4.60
Lakhisarai	95	9,805	8,592	11,018	6.31	9,623	8,503	10,743	5.94
Sheikhpura	65	10,432	8,720	12,144	8.37	10,374	8,889	11,859	7.30
Nalanda	68	11,089	9,279	12,899	8.33	10,931	9,378	12,484	7.25
Patna	133	10,958	9,915	12,001	4.86	10,869	9,880	11,858	4.64
Bhojpur	46	7,715	6,333	9,097	9.14	8,072	6,819	9,326	7.92
Buxar	178	8,897	8,259	9,535	3.66	8,921	8,297	9,545	3.57
Kaimur (Bhabua)	33	7,687	6,103	9,271	10.52	8,021	6,636	9,405	8.81
Rohtas	71	10,009	8,481	11,537	7.79	9,865	8,513	11,216	6.99
Aurangabad	32	8,212	6,456	9,968	10.91	8,636	7,139	10,134	8.85
Gaya	10	12,699	7,322	18,076	21.60	11,283	8,716	13,850	11.61
Nawada	82	11,051	9,473	12,629	7.29	10,862	9,459	12,265	6.59
Jamui	118	10,906	9,750	12,062	5.41	10,962	9,864	12,060	5.11
Jehanabad	42	11,112	8,688	13,536	11.13	10,622	8,777	12,467	8.86
			Non-sample District						
Arwal	0	-	-	-	-	8,981	7,039	10,924	11.03

Table 8. Direct and MFH estimates along with 95% confidence interval (95 % CI) and percentage coefficient of variation (CV) of the target variable urban-earning by district in Bihar.

		District				MFH			
			95% CI				95% CI		
District	Sample size	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Pashchim Champaran	29	9,251	7,679	10,823	8.67	9,458	7,941	10,975	8.18
Purba Champaran	89	10,384	8,971	11,797	6.94	10,404	9,031	11,777	6.73
Sitamarhi	8	10,463	8,271	12,655	10.69	10,367	8,302	12,432	10.16
Madhubani	18	8,468	6,957	9,979	9.10	8,702	7,240	10,165	8.57
Araria	24	13,744	10,708	16,780	11.27	13,598	10,888	16,308	10.17
Kishanganj	32	12,790	10,276	15,304	10.03	13,119	10,761	15,477	9.17
Purnia	102	15,108	12,654	17,562	8.29	14,518	12,249	16,787	7.97
Katihar	9	7,988	6,472	9,504	9.68	8,192	6,722	9,661	9.15
Madhepura	7	8,831	7,082	10,580	10.10	9,199	7,529	10,870	9.26
Saharsa	15	7,573	6,212	8,934	9.17	7,788	6,460	9,115	8.70
Darbhanga	52	11,591	9,676	13,506	8.43	11,392	9,555	13,229	8.23
Muzaffarpur	83	16,513	13,325	19,701	9.85	15,723	12,934	18,512	9.05
Siwan	15	21,609	11,351	31,867	24.22	14,446	9,785	19,107	16.46
Saran	39	12,670	10,300	15,040	9.55	12,605	10,398	14,812	8.93
Vaishali	41	12,167	9,978	14,356	9.18	11,627	9,549	13,705	9.12
Samastipur	22	19,518	12,485	26,551	18.38	16,926	12,366	21,486	13.74
Begusarai	28	8,904	7,396	10,412	8.64	9,272	7,807	10,737	8.06
Bhagalpur	102	13,891	11,791	15,991	7.71	13,572	11,581	15,563	7.48
Banka	6	6,995	5,641	8,349	9.88	7,279	5,952	8,605	9.30
Munger	16	16,782	11,753	21,811	15.29	14,440	10,746	18,134	13.05
Lakhisarai	50	12,407	10,250	14,564	8.87	12,107	10,082	14,132	8.53
Sheikhpura	9	14,565	10,622	18,508	13.81	14,120	10,878	17,362	11.72
Nalanda	42	17,610	13,010	22,210	13.33	14,535	10,964	18,106	12.53
Patna	362	19,354	17,418	21,290	5.10	18,727	16,868	20,586	5.07
Bhojpur	59	11,530	9,690	13,370	8.14	11,390	9,624	13,156	7.91
Buxar	17	9,859	8,009	11,709	9.58	10,389	8,612	12,166	8.73
Kaimur (Bhabua)	20	13,101	10,241	15,961	11.14	12,540	9,960	15,120	10.50
Rohtas	51	11,070	9,275	12,865	8.27	11,195	9,480	12,910	7.82
Aurangabad	14	13,251	10,169	16,333	11.87	13,128	10,420	15,836	10.52
Gaya	54	11,317	9,488	13,146	8.25	11,159	9,412	12,906	7.99
Nawada	21	7,214	5,968	8,460	8.81	7,346	6,119	8,572	8.52
Jamui	22	10,847	8,797	12,897	9.64	11,037	9,109	12,965	8.91
Jehanabad	15	11,174	8,907	13,441	10.35	11,369	9,257	13,481	9.48
			Non-sample Districts						
Sheohar	0	-	-	-	-	13,923	9,912	17,934	14.70
Supaul	0	-	-	-	-	11,443	7,641	15,245	16.95
Gopalganj	0	-	-	-	-	11,487	7,379	15,595	18.24
Khagaria	0	-	-	-	-	10,665	6,862	14,468	18.19
Arwal	0	-	-	-	-	11,011	7,107	14,915	18.09

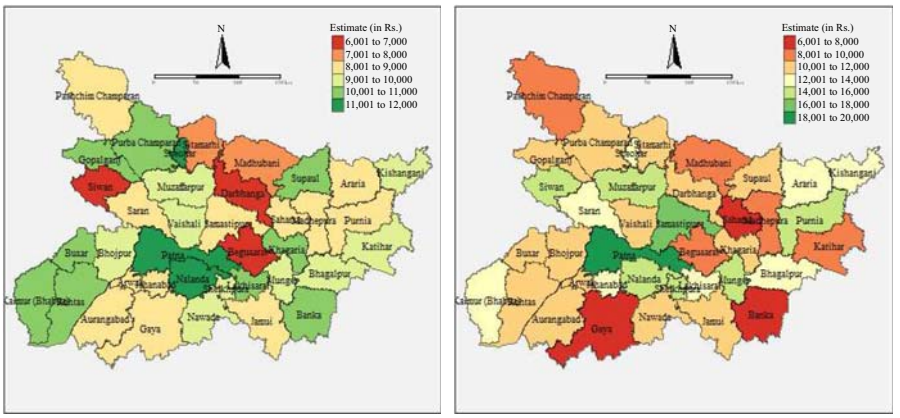


Fig. 6. Model-based MFH estimates displaying the spatial distribution of earning inequality by district between rural (on the left) and urban (on the right) districts in Bihar.

urban areas hold an average monthly earning of more than Rs. 12,500. Taken together, it is evident from these results that the degree of inequality in earning distribution between rural and urban districts is extremely severe. The difference in average monthly earning between rural and urban areas of Bihar can be obtained from Table 7–Table 8 and we may conclude that out of 38 districts in Bihar, 04 districts in rural areas are having earning higher than urban areas. But this seems not be the case as the direct estimates for urban areas in these 04 districts are truly unstable with higher CV percentage. Districts viz. Katihar, Saharsa, Banka and Gaya indicate higher level of earning in rural areas compared to urban. Sample sizes for rural areas in these districts also indicate that these particular districts covered more rural parts than the urban areas. These spatial maps and results provide useful information to policymakers in effective policy formulation and financial planning.

3.4. Distribution of Average Monthly Earning and Earning Inequality

In what follows, we tried to get a closer look into the degree of inequality between rural and urban areas. Figure 7 demonstrates two bivariate plots showing the spatial distribution of average monthly earning and Theil index of inequality by district in rural and urban areas in Bihar. We have used the Theil index to measure the degree of earning inequality as this index is perhaps best suited to measure the economic inequality. The U.S. Census Bureau (2016) mentioned that "the Theil index measures an entropic "distance" the population is away from the "ideal" egalitarian state of everyone having the same income." We have calculated the Theil index separately for Rural-Earning and Urban-Earning as

$$T_i = \sum_{j=1}^{M} \frac{\hat{Y}_{ij}}{\hat{Y}_i} T_{ij} + \sum_{j=1}^{M} \frac{\hat{Y}_{ij}}{\hat{Y}_i} \ln\left(\frac{\hat{Y}_{ij}/\hat{Y}_i}{N_{ij}/N_i}\right),$$

where $\hat{Y}_{ij}$ denote the MFH estimate of Rural-Earning and Urban-Earning for $i = 1, 2$ and district j and the population size corresponding to the variable i and district j is N_{ij} with $N_i = \sum_{j=1}^{M} N_{ij}$. Here $\hat{Y}_i$ is defined in Subsection 3.2.3 and T_{ij} is the Theil index

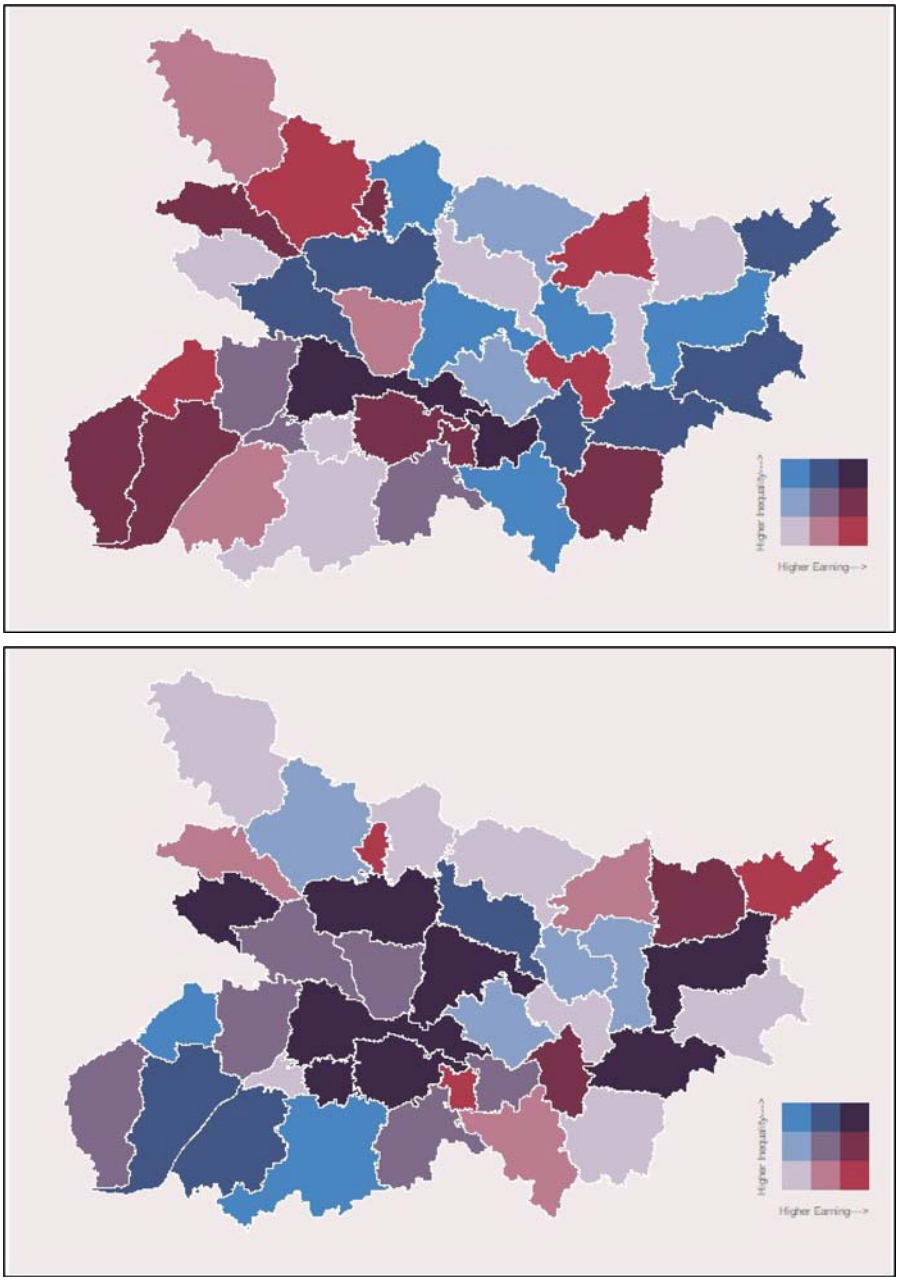


Fig. 7. Bivariate plot showing the spatial distribution of average monthly earning and Theil index of earning inequality by district in rural (at the top) and urban (at the bottom) districts in Bihar.

corresponding to the variable *i* and district *j* obtained from the original survey data. From Figure 7, the areas in grey indicate low earning and low inequality, blue areas show high earning and low inequality, areas in violate indicate high earning and high inequality and lastly the red areas specify low earning and high inequality. These bivariate spatial maps in Figure 7, also reveal that degree of inequality is increased with earning level and

consequently urban areas indicate greater degree of inequality than its counterpart. These bivariate maps offer worthy evidence to administrators and decision makers in prioritizing the areas more attention.

4. Conclusion

According to UNDP (2015), during 1990–2010 earning inequality has increased by 11% in developing countries and in India, 27.5% of the population are multidimensionally poor. World Bank (2020) mentioned that India accounts for 139 million of the total 689 million people (20.17%) living in extreme poverty in 2017 while its population being 17.8% of the world population (World Bank 2019) which indicates that India's share of the world's extreme poor population is higher than its share of the world population. Similar to the major developed economies around the globe, the COVID-19 pandemic has also hit the India's economy very hard with the loss of millions of jobs, causing in considerably reduced household incomes and extreme poverty. Due to the socioeconomic and health crisis in this pandemic, India's economy has experienced the biggest annual contraction of 7.3% in its gross domestic product (MoSPI 2021) since independence.

As part of the 2030 Agenda for sustainable development, the target 1 of the 10th goal aims that "By 2030, progressively achieve and sustain income growth of the bottom 40% of the population at a rate higher than the national average". From Table S2 in the online supplementary material, we can dig deeper to see the past pattern of earning inequality in India. The share in total earning of the top 1% declined gradually from 13% (1961) to 6.9% (1981) but after the major economic reforms in 1990s, it is continuously increasing from 10.4% (1991) to 21.7% (2019). The share in total earning of the bottom 50% remained fairly constant from 19.3% in 1961 to 21.4% in 1981. Afterward, it started declining from 19.8% in 1991 to 13.1% in 2019. To implement the agenda of sustainable development, India currently lacks the critically essential disaggregate level measures and maps of localized earning inequality.

In this study, the MFH model and its corresponding empirical best linear unbiased predictor for estimating small area means along with MSE estimation are summarized. Then, we applied this method in the 2018–2019 PLFS data of NSO, Govt. of India to produce the district level estimates and spatial mapping of earning inequalities in rural and urban areas of the state of Bihar in India. For the selection of suitable covariates for MFH model, data from 2011 Population Census of India are used and we applied PCA followed by a stepwise regression technique for selecting significant covariates. Guha and Chandra (2021b) described that multivariate modelling approach incorporating the correlation between the target variables often leads to efficient estimation of small area parameters of correlated measures. However, following Franco and Bell (2021) it should also be noted that the gain in precision for multivariate models can only be realized if one of the variables under study has low variance and the target variables are strongly correlated. Otherwise, both FH and MFH models may produce identical results. In this analysis, both the target variables viz. Rural-Earning and Urban-Earning are jointly modelled via MFH model and the gain is achieved in terms of efficiency (i.e., MSE) and CV in estimating the district level estimates of rural and urban earning inequalities. These estimates related to earning inequalities across the state of Bihar can assist in initiating the dialogue about the

drivers of earning inequalities in this state. Various diagnostic methods were used to assess the model-based MFH estimates and it also reveals significant gains in efficiency in producing district level estimates of earning which consequently measures the distribution of inequality that persist between rural and urban areas in the state. Moreover, the spatial maps so obtained show the evidence of unequal earning distribution and uneven distribution of inequality across the districts of rural and urban areas in Bihar. In rural areas, districts such as Patna, Lakhisarai exhibit higher level of earning and very high level of inequality whereas in urban areas, districts like Patna, Nalanda, Muzaffarpur, Samastipur, Purnia, Bhagalpur revealed higher level of earning as well as very high level of inequality in the state. When we tried to get a closer look on the degree of inequality between the rural and urban areas, we found that 08 districts in urban areas indicate very high level of inequality compared to only 02 districts in rural areas which eventually shows that though the earning level is much higher in urban areas, the degree of inequality is also far greater than the rural areas in Bihar.

This analysis undoubtedly established the advantages of SAE approach to deal with the problem of small sample sizes in obtaining precise and cost effective disaggregate or local level estimates along with the confidence intervals from existing survey data, (Guha and Chandra 2021b). In addition, this study also reveals the advantages of MFH over FH model in case of nonsample districts. This analysis established the fact that many areas in rural sector of Bihar possess very low level of earning compared to the urban sector while the degree of inequality is much higher in urban areas. The NSO surveys of Government of India are intended for obtaining state and national level estimates and these surveys do not reveal the real situation at the micro level (for example block or district level). However, substantial importance is given on micro level planning by the Government of India for realizing a stable economic development together with earning generation. For definite planning and development in a country, district is always an important unit and thus availability of district-level data and statistics are much vital for planning and monitoring of policy action plans. These cost effective and precise model-based estimates together with spatial maps may be useful for various Ministries and Departments of Government of India along with International organizations for effective policy planning and monitoring related to sustainable development goal ten – reduced inequalities. This study can be useful in obtaining the district level estimates and examine the inequality in earning distribution in the remaining parts of the country. Moreover, as earning data are generally skewed in nature, authors are working on this problem in multivariate SAE framework.

5. References

Benavent, R., and D. Morales. 2016. "Multivariate Fay-Herriot models for small area estimation". *Computational Statistics and Data Analysis* 94: 372–390. DOI: https://doi.org/10.1016/j.csda.2015.07.013.

Brown, G., R. Chambers, P. Heady, and D. Heasman. 2001. "Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS". In Proceedings of Statistics Canada Symposium 2001. Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Ottawa: Statistics Canada. Available

at: https:// www150.statcan.gc.ca/n1/pub/11-522-x/2001001/session6/6247-eng.pdf (accessed April 2021).

Census. 2011. Primary Census Abstracts, Registrar General of India, Ministry of Home Affairs, Government of India. Available at: https://www.censusindia.gov.in/2011-Common/CensusData2011.html (accessed April 2021).

Chandra, H., N. Salvati, and U.C. Sud. 2011. "Disaggregate-level Estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique". *Journal of Applied Statistics* 38(11): 2413–2432. DOI: https://doi.org/10.1080/02664763.2011.559202.

Datta, G.S., R.E., Fay, and M. Ghosh. 1991. "Hierarchical and empirical Bayes multivariate analysis in small area estimation". In Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, January, Washington, D.C., U.S.A.: 63–79 Available at: https://play.google.com/books/reader?id = Lr3rAAAAMAAJ&pg = GBS.PA64&hl = en (accessed April 2021).

Datta, G., T. Kubokawa, I. Molina, and J.N.K. Rao. 2011. "Estimation of mean squared error of model-based small area estimators". *TEST* 20(2): 367–388. DOI: https://doi.org/10.1007/s11749-010-0206-2.

Fay, R.E., and R. Herriot. 1979. "Estimates of income for small places: an application of James stein procedures to census data". *Journal of the American Statistical Association* 74: 269–277.DOI: https://doi.org/10.2307/2286322.

Fay, R.E. 1987. "Application of multivariate regression of small domain estimation". In *Small Area Statistics,* edited by R. Platek, J.N.K. Rao, C.E. Särndal, and M.P. Singh: 91–102. New York: Wiley. DOI: https://doi.org/10.1002/bimj.4710300737.

Franco, C., and W.R. Bell. 2021. "Using American community survey data to improve estimates from smaller U.S. surveys through bivariate small area estimation models". *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smaa040.

GHI. 2020. *Global Hunger Index 2020*. Available at: https://www.globalhungerindex.org/results.html (accessed September 2021).

Guha, S., and H. Chandra. 2021a. "Measuring and Mapping Disaggregate Level Disparities in Food Consumption and Nutritional Status via Multivariate Small Area Modelling". *Social Indicators Research* 154(2): 623–646. DOI: https://doi.org/10.1007/s11205-020-02573-8.

Guha, S., and H. Chandra. 2021b. "Measuring disaggregate level food insecurity via multivariate small area modelling: evidence from rural districts of Uttar Pradesh, India". *Food Security* 13(3): 597–615. DOI: https://doi.org/10.1007/s12571-021-01143-1.

Johnson, F.A., H. Chandra, J. Brown, and S. Padmadas. 2010. "Estimating district-level births attended by skilled attendants in ghana using demographic health survey and census data: an application of small area estimation technique". *Journal of Official Statistics* 26(2): 341–359. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/district-level-estimates-of-institutional-births-in-ghana-application-of-small-area-estimation-technique-using-census-and-dhs-data.pdf

MoHFW. 2021. *COVID-19 State wise Status*. Available at: https://www.mohfw.gov.in (accessed September 2021).

MoSPI. 2020a. *Annual Report: PLFS, 2018–19*. Government of India, Ministry of Statistics and ProgrammeImplementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/publicationreports/Annual_Report_PLFS_2018_19_HL.pdf (accessed April 2021).

MoSPI. 2020b. *Quarterly Bulletin, PLFS: April–June* 2020. Government of India, Ministry of Statistics and Programme Implementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/publication_reports/PLFS_Quarterly_Bulletin_April-June_2020.pdf (accessed September 2021).

MoSPI. 2021. Press note: *On estimates of gross domestic product for the first quarter (april − june) 2021–2022*. Government of India, Ministry of Statistics and Programme Implementation, New Delhi. Available at: http://mospi.nic.in/sites/default/files/press-release/PRESS_NOTE-Q1_2021-22.pdf (accessed September 2021).

NITI Aayog. 2019. *SDG India Index*. Available at: https://sdgindiaindex.niti.gov.in (accessed September 2021).

Prasad, N.G.N., and J.N.K. Rao. 1990. "The estimation of the mean squared error of small-area estimators". *Journal of the American Statistical Association* 85: 163–171. DOI: https://doi.org/10.2307/2289539.

Rao, J.N.K., and I. Molina. 2015. Small area estimation. New York: Wiley. DOI: https://doi.org/10.1002/9781118735855.

Särndal, C.E., B. Swensson, and J.H. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer-Verlag.

Schall, R. 1991. "Estimation in generalized linear models with random effects". *Biometrika* 78(4): 719–727. DOI: https://doi.org/10.2307/2336923.

UNCDP. 2018. *Official Records of the Economic and Social Council, 2018*. Available at: https://undocs.org/pdf?symbol = en/E/2018/33 (accessed September 2021).

UNDP. 2015. *Sustainable Development Goals*. Available at: https://www.undp.org (accessed April 2021).

U.S. Census Bureau. 2016. *Theil Index*. Available at: https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/theil-index.html (accessed September 2021).

World Bank. 2019. Open Data. Available at: https://www.worldbank.org/en/home (accessed September 2021).

World Bank. 2020. Open Data. Available at: https://www.worldbank.org/en/home (accessed September 2021).

Received March 2021
Revised September 2021
Accepted January 2022