Social Indicators Research (2022) 162:643–663
https://doi.org/10.1007/s11205-021-02857-7

ORIGINAL RESEARCH

Multivariate Small Area Modelling for Measuring Micro Level Earning Inequality: Evidence from Periodic Labour Force Survey of India

Saurav Guha[1] · **Hukum Chandra**[1]

Accepted: 30 November 2021 / Published online: 6 January 2022

Abstract
The economy of India is growing continuously with its gross domestic product increasing rapidly than most of the developing countries. Nonetheless an increase in national gross domestic product is not revealing the earning parity at micro level in the country. The earning inequality in a country like India has adversely obstructed under privileged in accessing basic needs such as health and education. The Periodic labour force survey (PLFS) conducted by the National Statistical Office of India generates estimates on earning status at state and national level for both rural and urban sectors separately. However, due to a small sample size problem that leads to high sampling variability, these surveys cannot be used directly to produce reliable estimates at micro level such as district or further disaggregate levels. As earnings are often unevenly distributed among the subgroups of comparatively small areas, disaggregate level statistics are inevitably needed in the country for target specific policy planning and monitoring to reduce the earning disparity. Nonetheless, owing to unavailability of estimates at district level, the analysis and spatial mapping related to earning inequality are limited to the national and state level. As a result, the existing variability in disaggregate level earning distribution are often unavailable. This article describes multivariate small area estimation (SAE) to generate precise and representative district-wise model-based estimates of inequality in earning distribution in rural and urban areas of Uttar Pradesh state in India by linking the latest round of PLFS 2018–2019 data and the 2011 Indian Population Census data. The diagnostic measures demonstrate that the district-wise estimates of earning generated by multivariate SAE method are reliable and representative. The spatial maps produced in this analysis reveal district level inequality in earning distribution in the state of Uttar Pradesh. These disaggregate level estimates and spatial mapping of earning distribution are directly pertinent to measuring and monitoring the sustainable development goal 10 of inequality reduction within countries. These expected to offer evidence to executive policy-makers and experts for recognizing the areas demanding additional consideration. This study will definitely provide added advantage to the newly launched schemes of Government of India for fund distribution along with the better monitoring of these schemes.

Keywords Multivariate small area estimation · Earning inequality · Periodic labour force survey · NSO · Census

Extended author information available on the last page of the article

1 Introduction

The Indian economy has developed at verifiably remarkable rates and is presently one of the fastest developing economies in the world. The country has progress appreciably to reform its economy, reduction hardship and fulfilling opportunities for everyday comfort for its widespread population. This continuous economic growth has also led to poverty reduction. Regardless of this noteworthy accomplishment, the income distribution in India remains intractably unequal. 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 context 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 hugely influence the contemporary economic environment. Thus, it is immensely important to measure its short-term impact on labour market which requires 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), Govt. 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 extremely essential, the estimates of earning distribution are unachievable further down the state level in India e.g., district, block or further level of disaggregation.

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 of sustainable development goal (SDG) 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 the current situations, the growing interests of the policy makers, public agencies, scientists, government organization are focused in achieving the local (or micro) level estimates. The emphasis on disaggregate level SDG indicators by various national and international agencies has further lauded 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 & 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 SAE methodology provides a viable and cost-effective solution this problem of small sample sizes (Rao & Molina, 2015). The SAE techniques borrow strength from various external sources viz. time periods, areas etc. to obtain precise and reliable estimates.

The idea behind the SAE methodology is to link the variable under study with the auxiliary information through different statistical models which may leads to describe 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 & Herriot, 1979) is a widely accepted area level model in SAE 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 necessity of estimating correlated processes viz. poverty indicators, unemployment, etc. 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, poverty 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, 2020), 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 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 & 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 Uttar Pradesh contributing to almost 1.7 million cases (MoHFW, 2021). Uttar Pradesh, the most populous state in the country, accounts for about 17% of India's population with an area of 241 thousand square km that equals to 8% of India's total geographical area. About 29.43% 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, out of 132 countries India stands on 94th position with an overall score of 27.2 (GHI, 2020). The state of Uttar Pradesh ranks 24th out of 28 states in "zero hunger" with a score of 41 which is much lower than the national average of 47 (NITI Aayog, 2019). In addition, 54.1% of the population is in the lowest two wealth quintiles in Uttar Pradesh and it ranks last out of 28 states in "reduced inequality" parameter with a score of 41 which is much lower than the national average of 67 (NITI Aayog, 2019). Therefore, it seems rational to consider Uttar Pradesh 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 disaggregate level earning inequalities in India.

The paper 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 Uttar Pradesh. In Sect. 3, we 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 Sect. 4. 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 Uttar Pradesh. Finally, Sect. 5 summarizes the paper and provides concluding remarks.

2 Data Description

In this segment, 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 Uttar Pradesh 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. 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 revisits in urban areas has been used while there was no revisit in rural areas. A stratified multistage survey design was adopted, with the ultimate 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 & Chandra, 2021b).

The 2018–2019 PLFS data of the NSO comprised 28,132 persons in 5822 households from the rural and urban areas in 71 districts of Uttar Pradesh. For all the districts, the sample size ranges from 18 to 199 with an average of 85 for rural areas while for urban areas, it is 6–321 with an average of 57. This survey provides information on earning estimate of every person separately for rural and urban areas in Uttar Pradesh. Districts included moderately small sample sizes with an average sampling fraction of 0.000054 for rural and 0.00011 for urban areas. "On account of the constraint of small sample size, it is not possible to produce precise and reliable direct estimates at district level and subsequently leads to producing large standard errors from this survey" (Chandra et al., 2011; Rao & Molina, 2015). In this paper, we made an effort to address the problem of small sample size in achieving district-level estimates from the 2018–2019 PLFS data. The multivariate small area method has been applied to tackle this problem by including related covariates from the Population Census 2011 of India.

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 i.e., last 7 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–September followed by October–December in 2018 and January–March followed by April–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 Y1: average monthly Earning (in Rs.) of a person from employment in rural areas (hereafter denoted by Rural), Y2: average monthly Earning (in Rs.) of a person from employment in urban areas (hereafter denoted by Urban). This paper targets to estimate the inequality in average monthly earning of a person in rural and urban districts in Uttar Pradesh at small area level through joint modelling of the target variables i.e., Rural and Urban.

3 Multivariate Small Area Modelling

In what follows, we briefly describe multivariate SAE methodology applied in the estimation of district level inequality in distribution of average monthly earning. Let us assume that the population consists M small areas or areas (districts in this analysis) and let there are R number of target variables in this study. All the way through, a subscript $m(m = 1, \ldots, M)$ is used to denote the quantities possess by small area 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 Ω_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 jth unit of the rth target variable in mth area, $r = 1, \ldots, R$, $j = 1, \ldots, N_m$ and $m = 1, \ldots, M$. The aim is to estimate small area mean $\overline{Y}_{mr} = N_m^{-1} \sum_{j \in \Omega_m} y_{mrj}$, $r = 1, \ldots, R$ and $m = 1, \ldots, M$. The traditional direct survey estimator (hereafter denoted by Direct) for $\overline{Y}_{mr}$ is given by $\bar{y}_{mr} = \sum_{j=1}^{n_m} \tilde{w}_{mrj} y_{mrj}$ with $\tilde{w}_{mrj} = w_{mrj} \Big/ \sum_{j=1}^{n_m} w_{mrj}$ where $\tilde{w}_{mrj}$ is the normalized survey weight for jth unit of the rth variable in mth area. In addition, $\tilde{w}_{mrj}$ satisfies $\sum_{j=1}^{n_m} \tilde{w}_{mrj} = 1$ with w_{mrj} being the survey weight for jth unit of the rth variable in mth 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 jth unit of the rth variable in the mth area.

Let us further assume that $y_{mr}(r = 1, \ldots, R)$ be an unbiased direct survey estimator of an unknown population parameter (e.g., the population mean) Y_{mr} of the target variable r for small area m. Let $\mathbf{x}_{mr}$ 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 informations are acquired from some available secondary sources, for example, administrative registers, the population census, etc. We denote $\mathbf{y}_m = \left(y_{1r}, ..., y_{mr}\right)^T$, a vector of direct survey estimators 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 & Herriot, 1979) used for more than one target variables is given by

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

In literature, the model in (1) is time and again referred to the multivariate form of the FH model. The MFH model in (1) consists of two stage, 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 $\mathbf{X}_m = diag\left(\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mR}\right)_{R\times p}$, a matrix of available auxiliary variables where $p = \sum_{r=1}^{R} p_r$. This model in (1) can be denoted as an area level random effect model as

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

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

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

where $\mathbf{y} = \text{col}\left(\mathbf{y}_m; 1 \le m \le M\right)$ is the vector of direct estimates of order $MR \times 1$, $\mathbf{X} = \text{col}\left(\mathbf{X}_m; 1 \le m \le M\right)$ is the matrix of known covariates of dimension $MR \times p$, $\mathbf{Z} = \text{col}'\left(\mathbf{Z}_m; 1 \le m \le M\right)$ is the $MR \times MR$ matrix of known covariates illustrating differences between the small areas, $\mathbf{u} = \text{col}\left(\mathbf{u}_m; 1 \le m \le M\right)$ is the vector of random area effects of dimension $MR \times 1$ and $\boldsymbol{\varepsilon} = \text{col}\left(\boldsymbol{\varepsilon}_m; 1 \le m \le M\right)$ is the vector of sampling errors of dimension $MR \times 1$ with $\mathbf{u} \sim N\left(0,\ \mathbf{V}_u\right)$ and $\boldsymbol{\varepsilon} \sim N\left(0,\ \mathbf{V}_\varepsilon\right)$. At large, $\mathbf{Z}$ is denoted a matrix whose mth column $\mathbf{Z}_m$, $m = 1, \ldots, M$, 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 a $MR \times MR$ diagonal matrix. Furthermore, it is supposed that the random area effects $\mathbf{u}$ are independently distributed of the sampling errors $\boldsymbol{\varepsilon}$ where $\mathbf{u} \sim N\left(0,\ \mathbf{V}_u\right)$ and $\boldsymbol{\varepsilon} \sim N\left(0,\ \mathbf{V}_\varepsilon\right)$. The random effects covariance matrix is denoted by $\mathbf{V}_u = \text{diag}\left(\mathbf{V}_{um}; 1 \le m \le M\right)$ and $\mathbf{V}_\varepsilon = \text{diag}\left(\mathbf{V}_{\varepsilon m}; 1 \le m \le M\right)$ 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 $\mathbf{V}_{u_m} = \text{diag}\left(\sigma^2_{ur}; 1 \le r \le R\right)$, $\mathbf{V}_{\varepsilon_m} = \text{diag}\left(\sigma^2_{\varepsilon mr}; 1 \le r \le R\right)$, $m = 1, \ldots, M$ and $\sigma^2_{\varepsilon mr}$'s are known for the estimator based on univariate FH model (UFH). Second estimator, denoted by MFH-1, is based on MFH model with

$\mathbf{V}_{u_m} = \text{diag}(\sigma_{ur}^2; 1 \le r \le R)$, $m = 1, \ldots, M$, 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 $\mathbf{u}_m = \left(\mathbf{u}_{m1}, \ldots, \mathbf{u}_{mR}\right)'$ is generated via a first order heteroscedastic autoregressive HAR(1) process $\mathbf{u}_{mr} = \rho \mathbf{u}_{mr-1} + \boldsymbol{\tau}_{mr}$ with $\mathbf{u}_{m0} \sim N\left(0, \sigma_0^2\right)$, $\boldsymbol{\tau}_{mr} \sim N\left(0, \sigma_r^2\right)$, $r = 1, \ldots, R$ and σ_0^2, $\mathbf{u}_{m0}$, $\boldsymbol{\tau}_{mr}$ are independent. The components of $\mathbf{V}_{u_m}$ are given by $\sigma_{umii} = \sum_{k=0}^{i} \rho^{2k} \sigma_{i-k}^2$ and $\sigma_{umij} = \sum_{k=0}^{|j-i|} \rho^{2k+|j-i|} \sigma_{|j-i|-k}^2$, $i \neq j$ and it is assumed that sampling errors are not independent with each other i.e., $\mathbf{V}_\varepsilon$ 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 $Var(\mathbf{y}) = \mathbf{V}_y = \mathbf{V}_u + \mathbf{V}_\varepsilon = diag\left(\mathbf{V}_{ym}; 1 \le m \le M\right)$, with $\mathbf{V}_u = \mathbf{Z}'\mathbf{V}_u\mathbf{Z}$ and $\mathbf{V}_{ym} = \mathbf{V}_{um} + \mathbf{V}_{\varepsilon m}$, $m = 1, \ldots, M$. Here, the covariance matrix $\mathbf{V}_y$ depends on R and $R + 1$ unknown variance component parameters given by $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_R\right)$ for UFH and MFH-1 and $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_{R+1}\right)$ for MFH-2 model respectively. The restricted maximum likelihood (REML) method is applied to estimate $\boldsymbol{\theta}$. Replacing the estimated values $\hat{\boldsymbol{\theta}}$ of parameters $\boldsymbol{\theta}$ in $\mathbf{V}_u$ to obtain $\hat{\mathbf{V}}_u = \mathbf{V}_u(\hat{\boldsymbol{\theta}})$ 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

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

Here, the empirical best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the EBLUP of $\mathbf{u}$ are obtained as $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\mathbf{V}}_y^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}_y^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \hat{\mathbf{V}}_u\mathbf{Z}'\hat{\mathbf{V}}_y^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ 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 & Chandra, 2021b). The analytical MSE estimate of EBLUP of MFH in (4) is obtained following Benavent and Morales (2016).

4 Results and Discussions

4.1 Variable Selection and Model Fitting

We used Population 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 are accessible from the census data and we did some exploratory analysis prior to selection of appropriate covariates for multivariate small area model fitting. A stepwise regression is performed for choosing significant auxiliary variables based on Akaike information criterion (AIC) value. Initially, the direct estimates of two target variables i.e., Rural and Urban are plotted to get an impression about the correlation between them. From Fig. 1, it seems that these two target variables Rural and Urban are loosely correlated. Note that, for the target variable Urban, there was 05 non-sample districts for which no information is available related to earning. Consequently, we fit a FH model with the sample areas and then a synthetic estimation (see Chandra et al., 2011) is carried out to estimate the non-sample areas in urban districts. MSE of the synthetic estimates are obtained following Chandra et al., (2011).

Next, we proceed with the MFH-2 model described in the previous section using direct estimates of Rural and a combination of the direct estimates and the synthetic estimates of

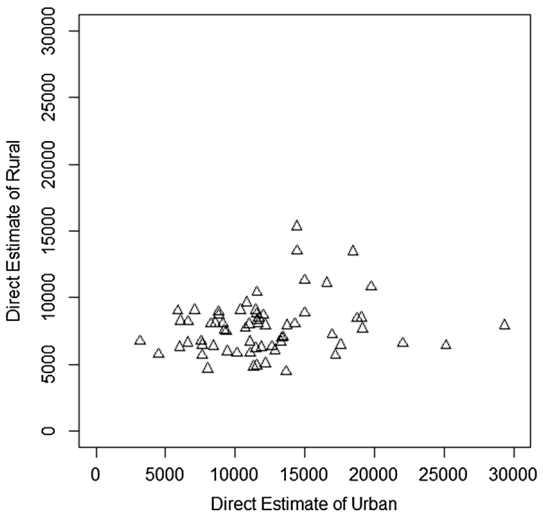

Fig. 1 Scatter plots of the direct estimates of Rural and Urban

Urban 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. main worker population (MWP), Cultivator population (CP) and marginal casual labour population (MCP) corresponding to the target variable Rural and for Urban, three significant covariates viz. literacy rate (LR), main worker population (MWP) and marginal casual labour population (MCP) are included in the model based on the AIC value. The regression parameter estimates are reported in Table 1 for the two dependent variable Rural and Urban. Observing the signs of the regression parameters estimates, it can be concluded that rural districts having lesser proportion in the covariate CP and greater proportion in MWP and MCP covariates have more earning while urban districts having greater proportion in all the three significant covariates have more earning.

The values of estimates from fitting the multivariate small area model in 2018–2019 PLFS data are described as follows. The estimate of variance component parameters for the MFH-2 model are $\hat{\sigma}^2_{u1} = 1247400$, $\hat{\sigma}^2_{u2} = 11161000$ and $\hat{\rho} = -0.3271$. 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 $t_{ij} = \hat{\sigma}^2_{ui} - \hat{\sigma}^2_{uj} \Big/ \sqrt{v_{11} + v_{22} - 2v_{12}}\ ;\ i, j = 1,\ 2,\ i \neq j$, where $v_{rs},\ r, s = 1,\ 2$ are the elements of the inverse of the matrix of Fisher information corresponding to the MFH-2 model calculated at $\hat{\boldsymbol{\theta}} = \left(\hat{\sigma}^2_{u1},\ \hat{\sigma}^2_{u2},\ \hat{\rho}\right)$. The value of test statistic is given by $t_{12} = -3.8803\ (< 0.001)$ with p-value given in parenthesis. As the value of the test statistic

Table 1 Regression parameters, Standard error and p-values for the target variables Rural and Urban

Variables	Rural				Urban			
	Intercept	MWP	CP	MCP	Intercept	LR	MWP	MCP
Estimate	5825	10,456	−34,013	13,393	−35,217	31,517	29,431	124,909
Standard error	1797	2523	12,724	6275	13,488	11,491	13,396	44,454
p-value	0.001	< 0.001	0.009	0.036	0.011	0.007	0.031	0.006

is significant at 5% level, this leads to the conclusion that variance of random area effects for Rural and Urban are significantly different. This followed by testing $H_0 : \rho = 0$ with the test statistic $t_\rho = \hat{\rho} \Big/ \sqrt{v_{33}}$ and the value of $t_\rho = -0.6208(0.2673)$, p-value is in parenthesis. This reveals that the correlation between the two target variables is not significantly different from zero and we go with MFH-1 model with diagonal covariance matrix instead of MFH-2 model. It is important to note that, although the variance of random area effects for Rural and Urban are significant at 1% level, 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 i.e., Rural and Urban for all the districts in Uttar Pradesh.

4.2 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 and Urban attained by SAE method under MFH-1 model is validated by some additional diagnostics. Corresponding to the target variable Rural and Urban, the random effects in MFH-1 model are supposed to follow a normal distribution with 0 mean and constant variance σ^2_{ur}, $r = 1, 2$. 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 and Urban are given in Fig. 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 0.138 and 0.445 for Rural and Urban respectively. Furthermore, it is evident from the Q-Q plot

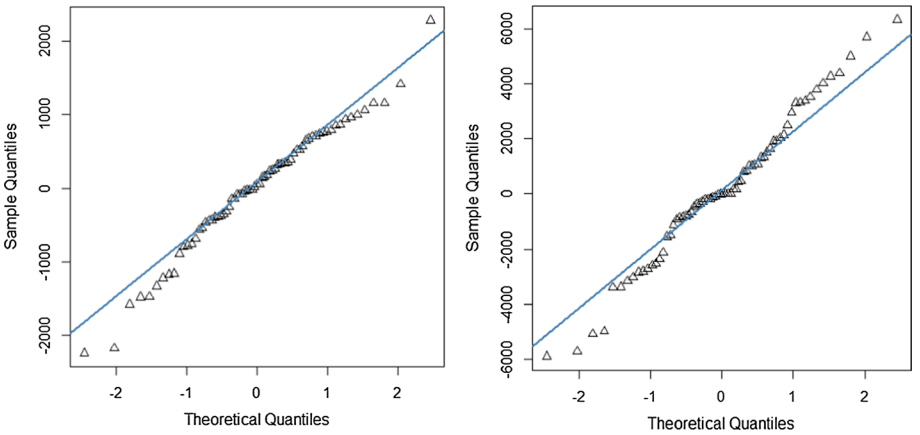

Fig. 2 Normal Q-Q plot of district-level residuals for Rural (on the left) and Urban (on the right)

in Fig. 2 that the normality assumption holds while p-values of the 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, the percentage coefficient of variation (CV) and the 95% 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, 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.

4.2.1 Bias Diagnostic

The bias diagnostic measure test the validity while the precision of the model-based estimates are 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 Fig. 3 demonstrates that small area estimates are not as extreme as the direct estimates signifying the usual SAE result of diminishing greater extreme values to the average values and the value of R^2 were 0.91 and 0.94 for Rural and Urban respectively. Largely, this diagnostic specify that the small area estimates are expected to be consistent when compared with direct estimates. This is expected as the MFH estimates are

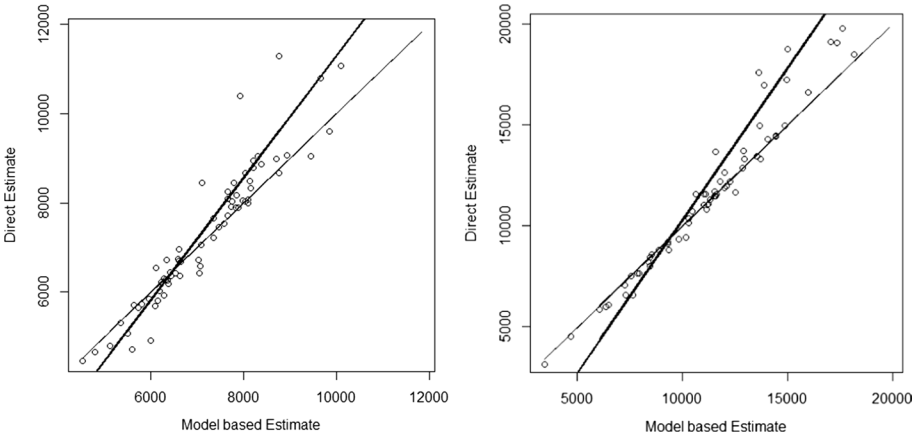

Fig. 3 Bias diagnostic plot with y = x line (thin line) and regression line (solid black line) for Rural (on the left) and Urban (on the right)

Table 2 Distribution of % CV for the direct and model-based small area estimates of Rural and Urban

Values	Rural			Urban (66 sample districts)			Urban (05 non-sample districts)		
	Direct	FH	MFH	Direct	FH	MFH	Direct	FH	MFH
Minimum	3.97	3.83	3.83	1.69	1.69	1.69	–	26.19	18.62
Q1	6.91	6.33	6.33	10.43	10.02	9.99	–	26.83	19.20
Median	9.57	8.34	8.34	13.45	12.54	12.49	–	37.66	26.93
Mean	10.97	8.38	8.39	14.60	13.07	12.98	–	37.86	27.24
Q3	12.39	9.65	9.66	17.90	15.25	15.13	–	42.63	30.66
Maximum	28.60	14.95	14.96	31.25	22.22	21.85	–	55.99	41.29

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.

4.2.2 Internal Diagnostic

Afterward, the degree of improvement in precision of model-based small area (i.e., district level) estimates of Rural and Urban 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 and Urban are given in Table 2 and the corresponding CV ratio is given in Fig. 4. District specific %CV is demonstrated in Fig. 5. The direct survey estimates possess greater CV compared to the FH and MFH estimates of Rural and Urban. It is obvious from Table 2 and Fig. 5 that direct survey estimates of Rural and Urban seem to be highly unstable. It is important to note that, for the target variable Urban, there was 05 non-sample districts. So, for Urban we first compare the performance with the sampled districts and comparison of non-sample districts are given separately. For Rural, the CV of Direct distributed from 3.98 to 28.61% with a median value of 9.58% whereas it is 3.83–14.96% with a median value of 8.34% for MFH which indicate that the MFH estimates are more strongly distributed compared to the direct estimates. Similarly for Urban, when we compare the Direct with MFH for the 66 sampled districts, the CV of Direct is distributed from 1.69 to 31.25% with a median value of 13.45% whereas it is 1.68–21.85% with a median value of 12.48% for MFH. Furthermore, when FH and MFH estimates are compared, the performance seems to be similar in case of Rural where there was no non-sampled district. However, in case of Urban with 05 non-sampled district, the CV of FH is distributed from 26.18 to 55.99% with a median value of 37.65% for the non-sampled districts while it is 18.61–41.29% with a median value of 26.93% for MFH estimates. It seems clear From Fig. 5, with decreasing sample sizes of the districts the relative performance of the MFH estimates for Rural and Urban has improved. Accordingly, these precise and reliable MFH estimates generate the district level earning estimate much better than direct and FH estimates. The 95% confidence intervals (CIs) are given in Fig. 6. The Fig. 6 indicate that the CI of MFH estimate is much tighter than the direct survey estimates.

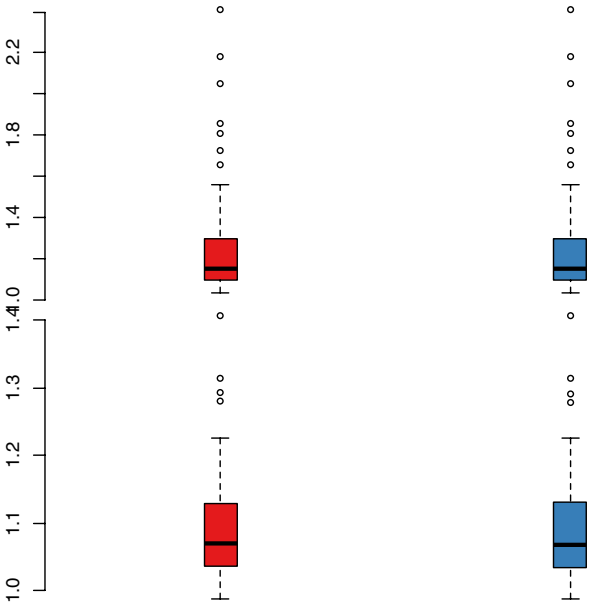

Fig. 4 CV ratio of Direct to FH (Red) and Direct to MFH (Blue) estimates for Rural (at the top) and urban (at the bottom). (Color figure online)

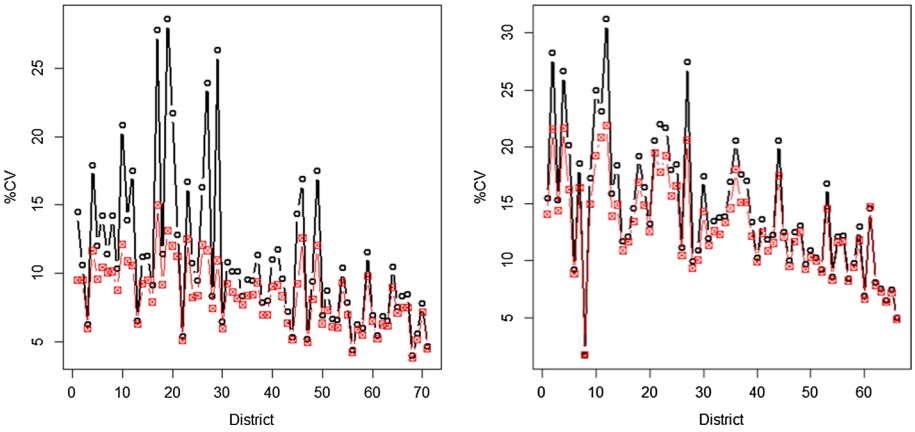

Fig. 5 District specific percentage coefficient of variation (CV) of Direct (Black, °) and MFH (Red, □) estimate for Rural (in left) and Urban (in right). Districts are arranged in increasing order of sample size. (Color figure online)

4.2.3 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 regional and state-level estimates of Rural and Urban is obtained by.

$\hat{Y}_i = \sum_{j=1}^{M} N_j \hat{Y}_{ij} \Big/ \sum_{j=1}^{M} N_j$, $i = 1,\ 2$ and $j = 1, \ldots, M$,where $\hat{Y}_{ij}$ denote the MFH estimate of Rural and Urban for $i = 1,\ 2$ and district j and the population size is N_j corresponding to the jth district. The districts are classified in four regions i.e. Central, Southern, Eastern

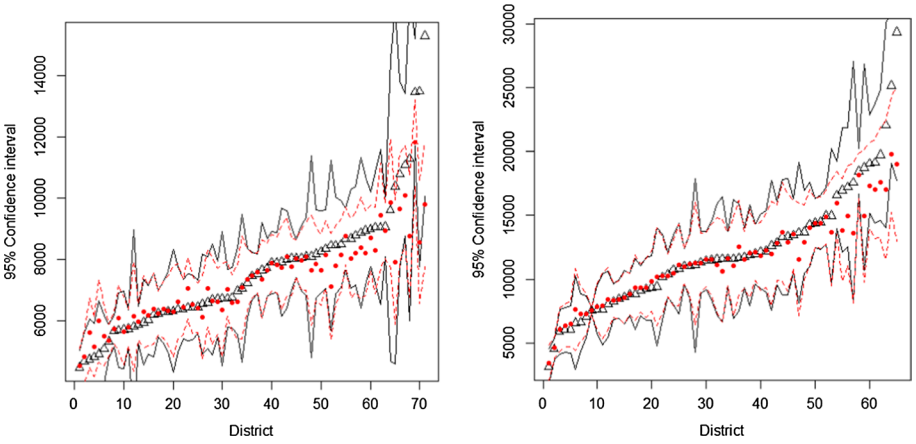

Fig. 6 District-wise 95% nominal confidence interval for the Direct (Black) and MFH (Red) estimates for Rural (on the left) and urban (on the right). Districts are arranged in increasing order of direct estimates. (Color figure online)

Table 3 Aggregated estimates of Rural and Urban obtained from Direct and MFH. Estimates are aggregated over 71 districts for rural areas and 66 districts for urban areas at the state and regional levels

Region	Rural		Urban	
	Direct	MFH	Direct	MFH
State	7609	7351	13,893	13,215
Eastern	6980	6789	16,812	14,935
Western	8468	8208	13,066	12,692
Central	7416	6907	13,152	13,068
Southern	6549	6593	14,201	12,653

Table 4 Distribution of CV of 71 districts in rural areas and 66 districts in urban areas

Group	%CV	Rural		Urban	
		Direct	MFH	Direct	MFH
1	≤ 5	03	04	02	02
2	5.01–10	33	50	12	15
3	10.01–15	23	17	25	31
4	15.01–20	06	0	15	13
5	≥ 20	06	0	12	05

and Western regions and we studied the aggregation property. The state and the regional-level estimates of Rural and Urban are reported in Table 3. 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 5 Distribution of earning range of 71 districts in rural and urban areas based on MFH estimates

Group	Earning range (in Rs.)	Rural	Urban
1	≤ 5000	02	02
2	5001–7500	36	05
3	7501–10,000	31	15
4	10,001–15,000	02	39
5	$\geq 15{,}000$	0	10

4.3 Spatial Distribution of Earning Inequality

Tables 4 and 5 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 and Urban are given in Tables 6 and 7. The spatial maps of earning (in Rs) by districts for both rural and urban areas are produced for the district level estimates generated by MFH method. Figure 7 displays spatial maps of the MFH estimates of earning for Rural and Urban areas of Uttar Pradesh. 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, western region in Uttar Pradesh has lower earning followed by the central and eastern region. For urban areas, the lower earning level exist in central region followed by the eastern and western region. The average monthly earning is ranging from Rs. 4518 to 11,827 in rural areas whereas it is Rs. 3424 to Rs. 19,809 in urban areas. This clearly indicate that there is a huge difference in average monthly earning between rural and urban areas in Uttar Pradesh. Moreover, from Table 4 and Fig. 7, MFH estimates also reveal that number of districts having average monthly earning of Rs. 10,000 or more is only 02 in rural areas while it is 46 for urban areas. In case of lower earning level, 38 districts in rural areas showing average monthly earning of Rs. 7500 or less whereas it is only 07 districts for urban areas. Table 4 further described that almost 97% of rural areas possess an average monthly earning of Rs. 7500 or less however nearly 89% of urban areas hold an an average monthly earning of more than Rs. 7500. Taken together, it is evident from these results that the degree of earning inequality between rural and urban districts is extremely severe and clearly visible. The difference in earning in rural and urban areas of Uttar Pradesh can be obtained from Tables 6 and 7 and we may conclude that out of 71 districts in Uttar Pradesh, 06 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 06 districts are truly unstable with higher CV percentage. Districts viz. Jyotiba Phule Nagar, Kannauj, Etawah, Chitrakoot, Fatehpur and Faizabad 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 for which sample sizes are nearly tends to zero. These spatial maps and results provide useful information to policymakers in effective policy formulation and financial planning.

Table 6 Direct and MFH estimates along with 95% confidence interval (95% CI) and percentage coefficient of variation (CV) of the target variable Rural by District in Uttar Pradesh

District	Sample size	Direct				MFH			
		Estimate	95% CI		CV	Estimate	95% CI		CV
			Lower	Upper			Lower	Upper	
Saharanpur	60	9607	4650	14,564	26.32	9850	7739	11,960	10.93
Muzaffarnagar	92	9052	6504	11,600	14.36	9437	7730	11,144	9.23
Bijnor	132	8665	7646	9684	6.00	8759	7824	9695	5.45
Moradabad	181	9060	8076	10,044	5.54	8933	8030	9835	5.15
Rampur	80	8081	6812	9350	8.01	8091	6989	9194	6.95
Jyotiba Phule Nr	104	8987	7771	10,203	6.90	8695	7621	9770	6.31
Meerut	56	11,080	8750	13,410	10.73	10,089	8465	11,713	8.21
Baghpat	18	10,790	7731	13,849	14.47	9659	7859	11,460	9.51
Ghaziabad	107	13,461	11,703	15,219	6.66	11,827	10,424	13,230	6.05
Gautam B. Nr	45	15,313	10,053	20,573	17.52	9792	7760	11,824	10.59
Bulandshahr	176	8032	7406	8658	3.98	8039	7435	8642	3.83
Aligarh	108	7893	6874	8912	6.58	7879	6950	8808	6.02
Hathras	49	7907	6144	9670	11.38	7734	6343	9126	9.18
Mathura	79	8329	7053	9605	7.82	8152	7045	9260	6.93
Agra	65	8001	6412	9590	10.13	8085	6794	9376	8.14
Firozabad	122	7526	6878	8174	4.39	7569	6946	8193	4.20
Etah	68	8676	7060	10,292	9.50	8035	6717	9353	8.37
Mainpuri	199	6366	5391	7341	7.81	6436	5535	7338	7.15
Budaun	118	7224	6109	8339	7.88	7349	6352	8347	6.93
Bareilly	32	4903	3180	6626	17.92	5988	4623	7352	11.63
Pilibhit	122	7721	6774	8668	6.26	7650	6769	8531	5.88
Shahjahanpur	95	5850	5255	6445	5.19	5960	5383	6536	4.94
Kheri	199	5652	5139	6165	4.63	5726	5226	6226	4.46
Sitapur	154	6267	5246	7288	8.31	6363	5432	7295	7.47
Hardoi	145	8026	7002	9050	6.51	7746	6814	8678	6.14
Unnao	54	13,501	7762	19,240	21.69	8546	6536	10,555	12.00
Lucknow	117	6968	5549	8387	10.39	6614	5409	7818	9.29
Rae Bareli	42	8862	6395	11,329	14.20	8369	6708	10,031	10.13
Farrukhabad	43	9037	7213	10,861	10.30	8295	6873	9716	8.74
Kannauj	48	8173	6715	9631	9.10	7842	6627	9058	7.91
Etawah	63	7894	6897	8891	6.44	7819	6909	8730	5.94
Auraiya	55	7450	6662	8238	5.39	7462	6719	8205	5.08
Kanpur Dehat	58	11,295	5994	16,596	23.95	8764	6760	10,767	11.67
Kanpur Nagar	73	6182	4809	7555	11.33	6385	5217	7552	9.33
Jalaun	87	6017	4888	7146	9.58	6176	5170	7183	8.31
Jhansi	46	5067	3957	6177	11.18	5490	4495	6485	9.24
Lalitpur	32	5935	4538	7332	12.01	6288	5108	7467	9.57
Hamirpur	43	8085	4785	11,385	20.82	7654	5841	9468	12.09
Mahoba	34	4722	3407	6037	14.21	5594	4453	6735	10.41
Banda	59	6429	5379	7479	8.33	6525	5576	7474	7.42
Chitrakoot	66	8245	6902	9588	8.31	7652	6497	8807	7.70

Table 6 (continued)

District	Sample size	Direct				MFH			
		Estimate	95% CI		CV	Estimate	95% CI		CV
			Lower	Upper			Lower	Upper	
Fatehpur	87	6716	5771	7661	7.18	7022	6146	7897	6.36
Pratapgarh	165	8488	7078	9898	8.48	8132	6936	9328	7.51
Kaushambi	100	8780	7162	10,398	9.40	8214	6912	9516	8.09
Allahabad	64	6287	4959	7615	10.78	6322	5182	7463	9.20
BaraBanki	57	6309	4294	8324	16.30	6288	4801	7776	12.07
Faizabad	140	5730	5119	6341	5.44	5795	5205	6385	5.19
Ambedkar Nr	136	6435	5562	7308	6.92	6410	5595	7225	6.49
Sultanpur	56	6694	5456	7932	9.43	6613	5532	7694	8.34
Bahraich	45	4458	3888	5028	6.53	4518	3964	5072	6.26
Shrawasti	18	4653	3687	5619	10.59	4801	3910	5692	9.47
Balrampur	54	6545	4905	8185	12.78	6102	4761	7444	11.21
Gonda	146	7647	6077	9217	10.48	7353	6062	8644	8.96
Siddharthnagar	148	6743	5746	7740	7.55	6582	5665	7498	7.11
Basti	93	8451	5654	11,248	16.89	7107	5357	8857	12.57
Sant Kabir Nr	134	4803	3714	5892	11.56	5125	4146	6105	9.75
Mahrajganj	48	5808	2644	8972	27.79	6146	4344	7949	14.96
Gorakhpur	82	5685	4459	6911	11.01	6091	5011	7170	9.05
Kushinagar	101	7053	4635	9471	17.49	7082	5410	8754	12.04
Deoria	140	8048	6970	9126	6.84	7977	7001	8953	6.24
Azamgarh	72	8452	6887	10,017	9.45	7796	6506	9086	8.44
Mau	105	6419	5325	7513	8.70	7041	6038	8044	7.27
Ballia	84	6586	5075	8097	11.70	7058	5795	8320	9.13
Jaunpur	49	10,391	4564	16,218	28.61	7913	5881	9944	13.10
Ghazipur	43	6685	4869	8501	13.86	6644	5228	8059	10.87
Chandauli	64	8937	7163	10,711	10.13	8208	6820	9595	8.63
Varanasi	23	6227	5469	6985	6.21	6230	5504	6957	5.95
Bhadohi	39	5699	4425	6973	11.41	5641	4529	6754	10.06
Mirzapur	55	6713	4521	8905	16.66	6337	4784	7890	12.51
Sonbhadra	87	5315	4763	5867	5.30	5341	4804	5877	5.13
Kanshiram Nr	47	6362	4957	7767	11.27	6619	5390	7848	9.47

Nr- Nagar

5 Conclusion

According to UNDP (2015), earning inequality has increased by 11% in developing countries during 1990–2010 and 27.5% of the population in India are multidimensionally poor. World Bank (2020) reported that almost 21% of the India's population living in extreme poverty in 2017 while India accounts for 17.8% of the world population (World Bank, 2019) which reveals that India's share of the world's extreme poor population is greater than its share of the world population. Like all the major developed countries around the world, the COVID-19 pandemic has also hit the India's economy

Table 7 Direct and MFH estimates along with 95% confidence interval (95% CI) and percentage coefficient of variation (CV) of the target variable Urban by District in Uttar Pradesh

District	Sample Size	Direct				MFH			
		Estimate	95% CI		CV	Estimate	95% CI		CV
			Lower	Upper			Lower	Upper	
Saharanpur	91	10,799	7241	14,357	16.81	11,130	7945	14,315	14.60
Muzaffarnagar	54	11,461	9205	13,717	10.04	11,551	9395	13,707	9.52
Bijnor	118	8806	7120	10,492	9.77	8868	7227	10,509	9.44
Moradabad	117	10,372	8660	12,084	8.42	10,282	8616	11,948	8.26
Rampur	52	8587	6482	10,692	12.51	8539	6510	10,569	12.13
Jyotiba Phule Nr	7	5875	4109	7641	15.34	6072	4356	7788	14.42
Meerut	174	16,585	13,955	19,215	8.09	15,977	13,513	18,441	7.87
Baghpat	19	19,743	14,620	24,866	13.24	17,593	13,248	21,938	12.60
Ghaziabad	321	18,458	16,670	20,246	4.94	18,141	16,406	19,876	4.88
Gautam B. Nr	129	14,404	12,466	16,342	6.86	14,421	12,548	16,294	6.63
Bulandshahr	102	8241	6851	9631	8.60	8357	6993	9721	8.33
Aligarh	111	12,176	9275	15,077	12.16	11,798	9118	14,478	11.59
Hathras	16	13,697	10,444	16,950	12.12	12,879	9925	15,833	11.70
Mathura	72	11,569	9358	13,780	9.75	11,590	9483	13,697	9.28
Agra	207	11,015	9605	12,425	6.53	11,008	9626	12,390	6.41
Firozabad	31	9226	6788	11,664	13.48	9317	7014	11,619	12.61
Etah	29	11,984	9655	14,313	9.92	12,086	9876	14,296	9.33
Mainpuri	85	8442	6912	9972	9.24	8433	6934	9932	9.07
Budaun	154	16,973	12,115	21,831	14.60	13,854	9854	17,854	14.73
Bareilly	21	11,565	6650	16,480	21.68	10,641	6625	14,657	19.25
Pilibhit	64	10,726	7970	13,482	13.11	10,460	7891	13,029	12.53
Shahjahanpur	36	10,147	6760	13,534	17.03	10,269	7228	13,310	15.11
Kheri	46	7645	5809	9481	12.25	7843	6067	9619	11.55
Sitapur	17	6017	4292	7742	14.63	6368	4693	8044	13.42
Hardoi	72	9109	7156	11,062	10.94	9318	7436	11,199	10.30
Unnao	265	14,455	12,338	16,572	7.47	14,411	12,382	16,440	7.18
Lucknow	34	13,319	8737	17,901	17.55	12,913	9082	16,744	15.14
Rae Bareli	25	14,961	9677	20,245	18.02	13,645	9455	17,835	15.67
Farrukhabad	30	7062	5402	8722	11.99	7264	5648	8880	11.35
Kannauj	9	6595	4360	8830	17.29	7278	5146	9410	14.95
Etawah	8	29,341	17,762	40,920	20.14	19,027	12,958	25,096	16.27
Auraiya	31	9330	6806	11,854	13.80	9830	7455	12,204	12.33
Kanpur Dehat	204	14,958	12,751	17,165	7.53	14,869	12,759	16,979	7.24
Kanpur Nagar	44	11,479	8411	14,547	13.64	11,518	8710	14,326	12.44
Jalaun	125	12,838	9558	16,118	13.04	12,848	9875	15,821	11.81
Jhansi	41	12,193	8986	15,400	13.42	12,261	9343	15,179	12.14
Lalitpur	9	9426	4816	14,036	24.95	10,162	6325	13,999	19.27
Hamirpur	11	11,655	7456	15,854	18.38	12,524	8859	16,189	14.93
Mahoba	7	49,472	23,630	75,314	26.65	16,119	9278	22,960	21.65
Banda	27	17,574	8123	27,025	27.44	13,618	8125	19,111	20.58
Chitrakoot	6	6075	4230	7920	15.50	6468	4680	8256	14.10

Table 7 (continued)

District	Sample Size	Direct				MFH			
		Estimate	95% CI		CV	Estimate	95% CI		CV
			Lower	Upper			Lower	Upper	
Fatehpur	9	3154	1726	4582	23.11	3425	2024	4825	20.86
Pratapgarh	73	19,038	15,203	22,873	10.28	17,330	13,944	20,716	9.97
Kaushambi	17	11,556	7200	15,912	19.23	11,110	7426	14,794	16.92
Allahabad	32	11,887	7940	15,834	16.94	12,011	8570	15,452	14.62
BaraBanki	51	12,629	7537	17,721	20.57	11,986	7877	16,095	17.49
Faizabad	8	4520	3705	5335	9.20	4667	3858	5477	8.85
Ambedkar Nr	26	7646	4882	10,410	18.44	7920	5349	10,491	16.56
Sultanpur	12	11,037	8493	13,581	11.76	11,162	8773	13,551	10.92
Bahraich	19	13,642	8143	19,141	20.56	11,560	7145	15,975	19.49
Shrawasti	18	8011	5432	10,590	16.42	8444	5974	10,915	14.93
Balrampur	8	22,036	14,026	30,046	18.55	16,994	11,529	22,459	16.41
Gonda	61	19,114	14,415	23,813	12.54	17,053	13,145	20,961	11.69
Siddharthnagar	8	7529	7280	7778	1.69	7533	7284	7782	1.69
Basti	19	18,739	10,656	26,822	22.01	14,968	9758	20,178	17.76
Sant Kabir Nr	26	11,280	8817	13,743	11.14	11,293	8971	13,615	10.49
Mahrajganj	9	11,088	4297	17,879	31.25	11,228	6419	16,037	21.85
Gorakhpur	31	17,207	12,516	21,898	13.91	14,963	11,044	18,882	13.36
Kushinagar	29	13,440	10,569	16,311	10.90	13,513	10,850	16,176	10.06
Deoria	45	14,287	10,966	17,608	11.86	14,028	11,029	17,027	10.91
Azamgarh	41	11,708	9354	14,062	10.26	11,538	9300	13,776	9.90
Mau	114	25,131	19,113	31,149	12.22	19,809	15,265	24,353	11.70
Ballia	6	6582	2935	10,229	28.27	7623	4396	10,849	21.59
Jaunpur	33	11,571	6905	16,237	20.57	11,021	7123	14,919	18.05
Ghazipur	29	13,295	8749	17,841	17.45	13,710	9860	17,560	14.33
Chandauli	10	8828	6074	11,582	15.92	9359	6799	11,920	13.96
Varanasi	–	–	–	–	–	13,229	8402	18,056	18.62
Bhadohi	–	–	–	–	–	6204	1183	11,225	41.29
Mirzapur	–	–	–	–	–	8133	3246	13,021	30.66
Sonbhadra	–	–	–	–	–	12,916	8055	17,777	19.20
Kanshiram Nr	–	–	–	–	–	9204	4346	14,062	26.93

Nr- Nagar

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 first target of the 10th goal is "By 2030, progressively achieve and sustain income growth of the bottom 40% of the population at a rate higher than the national average". After the major economic reforms in India during 1990s, the share in total earning of the top 1% is continuously increasing while the share in total earning of the bottom 50% started declining. To implement the agenda of sustainable

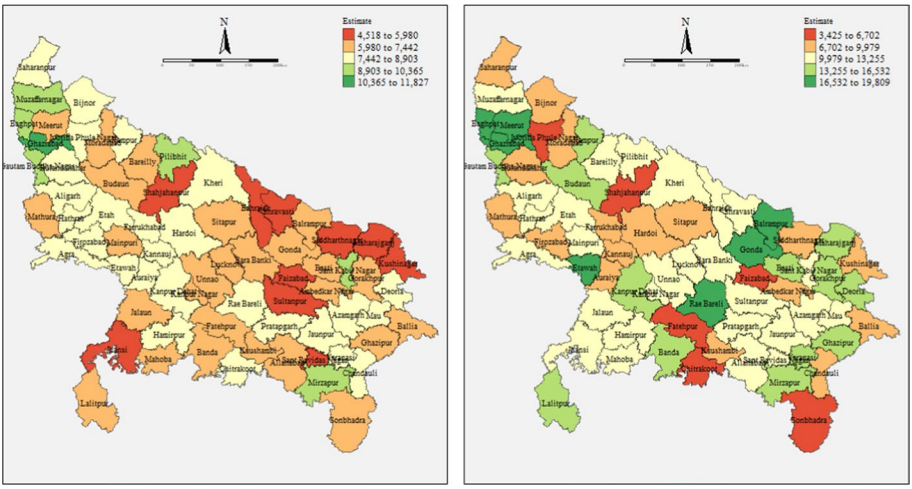

Fig. 7 Model-based MFH estimates displaying the spatial distribution of earning inequality by District between rural (on the left) and urban (on the right) areas in Uttar Pradesh

development, India currently lacks the critically essential disaggregate level measures and maps of localized earning inequality.

At the outset of this paper, the multivariate Fay–sHarriot model (MFH) and its corresponding empirical best linear unbiased predictor are summarized. Then we applied this method in the 2018–2019 PLFS data of NSO, Govt. of India to produce the model-based estimate and spatial mapping of earning inequalities in rural and urban areas of Uttar Pradesh. For selection of suitable covariates, data from 2011 Population Census of India are used and we applied stepwise regression technique for choosing significant covariates. "Efficient estimation of correlated measures like food insecurity, nutritional consumption disparities are often required multivariate modelling approach which takes into account for the correlation between the target variables" (Guha & Chandra, 2021b). In this analysis, both the target variables viz. Rural and Urban are jointly modelled via MFH model and the gain is achieved in terms of MSE and CV for the target variables Rural and Urban. These estimates related to earning inequalities across the state of Uttar Pradesh can assist in motivating 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 persist between rural and urban areas in the state. Moreover, the spatial maps so obtained show the evidence of unequal earning distribution across the districts of rural and urban areas in Uttar Pradesh. Districts such as Shajahanpur, Faizabad, Sonbhadra exhibit lower level of earning to a great extent and demonstrate very high-level inequality whereas districts like Gonda, Ghaziabad, Etawah revealed high level of earning in the state.

This study indisputably recognized the benefits of SAE technique to tackle the small sample size problem when we want to obtain cost effective and precise disaggregate level estimates together with the confidence intervals from the existing PLFS data. In addition, this study also reveals the advantages of MFH over FH model in case of non-sample districts. This analysis also established the fact that many areas in rural sector of Uttar Pradesh possess very low level of earning compared to the urban sector and

the earning gap is clearly visible from this study. 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). 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 purview and thus availability of district-level data and statistics are very 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 and Ministries and Departments in Government of India along with international organizations for effective policy planning and monitoring related to sustainable development goal 10—reduced inequalities. This study can assist 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 to handle this problem in multivariate SAE framework.

Acknowledgements The authors would like to acknowledge the valuable comments and suggestions of the Editor and two anonymous referees. The corresponding author would also like to acknowledge the efforts made by the co-author, the Late Dr. Hukum Chandra in finalizing the article. He left for his heavenly abode before the final acceptance of this article.

Declarations

Conflict of interest The authors declared that they have no conflict of interest.

References

Benavent, R., & Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis, 94*, 372–390.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (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*.

Census. (2011). *Primary census abstracts, registrar general of India, ministry of home affairs, government of India*. http://www.censusindia.gov.in/2011census/population_enumeration.html.

Chandra, H., Salvati, N., & Sud, U. C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal Applied Statistics, 38*(11), 2413–2432.

Chandra, H., Aditya, K., Gupta, S., Guha, S., & Verma, B. (2020). Food and nutrition in indo gangetic plain region-a disaggregate level analysis. *Current Science, 119*(11), 1783–1788.

Datta, G.S., Fay, R.E., & Ghosh, M. (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, Washington, DC, pp. 63–79.

Datta, G., Kubokawa, T., Molina, I., & Rao, J. N. K. (2011). Estimation of mean squared error of model-based small area estimators. *TEST, 20*(2), 367–388.

Fay, R. E., & Herriot, R. (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.

Fay, R. E. (1987). Application of multivariate regression of small domain estimation. In R. Platek, J. N. K. Rao, C. E. Särndal, & M. P. Singh (Eds.), *Small area statistics* (pp. 91–102). Wiley.

Franco, C., & Bell, W. R. (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*. https://doi.org/10.1093/jssam/smaa040

GHI (2020). *Global Hunger Index 2020*. https://www.globalhungerindex.org/results.html
Guha, S., & Chandra, H. (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. https://doi.org/10.1007/s11205-020-02573-8
Guha, S., & Chandra, H. (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. https://doi.org/10.1007/s11205-020-02573-8.
MoHFW, Government of India (2021). *COVID-19 statewise status*, https://www.mohfw.gov.in.
MoSPI, Government of India. (2020). *Annual report: PLFS, 2018–2019*. Ministry of statistics and programme implementation.
MoSPI, Government of India. (2020). *Quarterly bulletin, PLFS: April–June 2020*. Ministry of Statistics and Programme Implementation.
MoSPI, Government of India. (2021). *Press note on provisional estimates of annual national income 2020–21 and quarterly estimates of gross domestic product for the fourth quarter (Q4) Of 2020–21*. New Delhi: Ministry of Statistics and Programme Implementation.
NITI Aayog. (2019). *SDG India Index*. https://sdgindiaindex.niti.gov.in
Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association, 85*, 163–171.
Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Wiley.
Särndal, C. E., Swensson, B., & Wretman, J. H. (1992). *Model Assisted survey sampling*. New York, USA: Springer-Verlag.
UNDP. (2015). *Sustainable development goals*. https://www.undp.org/content/undp/en/home/sustainable-development-goals.html.
World Bank (2019). Available at https://data.worldbank.org.
World Bank (2020). Available at https://data.worldbank.org.

Authors and Affiliations

Saurav Guha[1] · **Hukum Chandra**[1]

✉ Saurav Guha
saurav.iasri@gmail.com

Hukum Chandra
hchandra12@gmail.com

[1] ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi, India