

Non-Normal Empirical Bayes Prediction of Local Welfare

*Chris Elbers
Roy van der Weide*



WORLD BANK GROUP

Development Economics
Development Research Group
April 2025

Abstract

Estimates of the area- and household idiosyncratic error distributions from household income and consumption regression models across 142 household surveys from 16 different countries, the type of models that underpin poverty maps, points to significant deviations from normality. Accounting for non-normality in Empirical Best estimation

of local welfare is found to increase precision relative to normal-Empirical Best estimation. Although the gains in precision range between meaningful and marginal, it is always positive. Given that non-normal-Empirical Best estimation is furthermore easy to implement, there is no downside to using it.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://www.worldbank.org/prwp>. The authors may be contacted at rvanderweide@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Non-Normal Empirical Bayes Prediction of Local Welfare

Chris Elbers and Roy van der Weide*

*Chris Elbers (c.t.m.elbers@vu.nl) is at VU University Amsterdam and the Tinbergen Institute. Roy van der Weide (rvanderweide@worldbank.org) is at the World Bank. We gratefully acknowledge financial support from VU University Amsterdam and the World Bank's Knowledge for Change Program (KCP II). A thank you also goes to Stephen Haslett, Peter Lanjouw, Isabel Molina, and David Newhouse for providing comments on earlier versions of this paper.

1 Introduction

Elbers et al. (2003; henceforward ELL) have popularized the estimation of poverty and inequality at the small area level, also referred to as “poverty mapping”.¹ Poverty here measures the share of individuals whose income or consumption falls below a given poverty line. Collecting household income and consumption data is expensive. This is particularly true for developing countries where much of the income does not come from wage employment. Consequently, income and consumption data is typically only available in the form of household income and consumption expenditure surveys. The sample size of these surveys is sufficient for the estimation of national and possibly sub-regional welfare, but too small to estimate welfare at the level of small areas. Small areas here refer to administrative areas that are below the province level, think of district or possibly even the municipality level.

Estimating poverty at the small area level is made possible by combining household survey data with unit record population census data.² The census has data on variables such as demographics, education, employment and housing, which serve as predictors of household income, but not the household income variable itself. Crucially, data on these predictors are also collected by the income survey. The income survey is used to train a household income model that is then used to predict income for every household in the population census. These predicted incomes can subsequently be aggregated to obtain estimates of poverty at the small area level. This approach has been applied to obtain maps of poverty in over 60 countries worldwide. For reviews of the literature on small area estimation, see e.g., Haslett (2013, 2016), Rao and Molina (2015), Tzavidis et al. (2018), Das and Haslett (2019), and Corral et al. (2020).

To accommodate spatial correlation in errors, a nested error structure is assumed where the total error is the sum of a household-specific error and a location error. The location error is typically modeled at the small area level. The total error observed for sampled households clearly carries information about the area random effect. Empirical Bayes (EB) estimation, also known as Empirical Best estimation, exploits this for the purpose of prediction, i.e. it uses the residuals observed in the survey to predict the location effects for sampled small areas, which in turn are used to predict local poverty rates.

Working out the conditional distribution for the location error (conditional on observed residuals) requires making distributional assumptions. It is standard in the literature on EB

¹First attempts to estimate poverty at the small area level using aggregate data, dates back to the seminal work by Fay and Herriott (1979).

²Population censuses provide complete coverage of a country and collects household and individual attributes that tend to be highly predictive of household welfare, think of household composition, education, employment, and dwelling characteristics, making it the ideal source of data for poverty mapping. An alternative source of data that similarly provides complete coverage of the country is remote-sensing data, think of night-time-lights, population density, land type and use, greenness, and local climate and pollution variables. The unit of observation at which these data can be connected to the household survey in this case would be the village level rather than the household level. For studies that have explored the use of remote-sensing data to obtain small area estimates of poverty, see e.g., Burke et al. (2021), Chi et al. (2022), Engstrom et al. (2022), Jean et al. (2016), Merfeld and Newhouse (2023), Newhouse et al. (2022), Newhouse (2023), Pokhriyal and Jacques (2017), Van der Weide et al. (2024), and the references therein. Studies vary in the resolution of imagery used, with higher resolution images allowing for a richer set of predictors that can be extracted from them (see e.g. Marx et al., 2019).

estimation to assume normally distributed errors, in which case the conditional distributions are also normal (see Molina and Rao, 2010). Assuming normality is however not without cost. Since poverty and inequality are non-linear functions of household consumption (and by extension non-linear functions of the errors), misspecification in the error distributions may introduce an error in the small area estimates of poverty.

ELL makes minimal assumptions about the errors, motivated by their empirical application where the errors did not conform to a normal distribution. Unfortunately, the non-parametric estimator for the distribution functions adopted by ELL does not lend itself for Empirical Bayes (EB) estimation. Molina and Rao (2010; henceforward MR) instead prioritizes EB estimation under the assumption of normally distributed errors. Where ELL accept a loss in precision by not implementing EB estimation, MR accept a loss in precision that might stem from a misspecification of the error distribution functions. Throughout the paper, higher precision will refer to lower Root Mean Squared Error (RMSE).

The approach put forward in this paper accommodates both EB estimation and non-normally distributed errors. We achieve this by fitting finite normal mixtures (NM) to the error distribution functions. Normal mixtures are extremely flexible; they are able to fit any well-behaved distribution function, and are ideally suited for accommodating EB estimation. With respect to the household idiosyncratic error term, we follow ELL by drawing from the empirical realizations of the idiosyncratic errors (with replacement).³ To verify what deviations from normality (or degrees of non-normality) are empirically relevant, we estimate the respective error distributions using data from 142 household surveys across 16 countries from different regions and different income groups.

Our findings are the following. An inspection of the estimated area- and idiosyncratic error distributions across the 142 household surveys reveals significant deviations from normality, especially for the household idiosyncratic component. Household errors have skewness ranging from -1.5 to 1.5 and kurtosis from 4 to 10 (the skewness of area errors ranges between -1 and 1, and kurtosis from 2 to 8, disregarding outliers). When the area error is small (2.5 percent of total error or less), the benefit of accounting for non-normality is seen to outweigh the loss stemming from foregoing Empirical Best prediction. As soon as the share of area error approaches 5 percent or higher, EB prediction dominates non-EB prediction across practically all empirically observed error distributions. Non-normal-EB prediction performs best in all cases. The improvement in performance relative to normal-EB is however comparatively marginal at higher levels of the location effect. Similarly, the improvement is marginal when compared to ELL at near-zero levels of the location effect. While the gains may range between marginal (near-zero reduction in RMSE) and meaningful (up to a 15-25 percent reduction in RMSE), they are always positive. Given that non-normal-EB prediction is easy to implement, there is no downside to using it.

Highly disaggregated estimates of poverty and inequality have inspired a range of applica-

³Another way of relaxing the normality assumption is explored in Bikauskaite et al. (2020), where a mixture is applied to the full nested error model, rather than fitting flexible mixture distributions to the respective error distributions separately.

tions. A natural use of poverty maps is for the targeting of social assistance programs, see e.g., Banerjee et al. (2025) and Smythe and Blumenstock (2022). Elbers et al. (2007) show in empirical applications to Madagascar, Cambodia, and Ecuador, that reductions in poverty achieved with uniform transfers can be realized with less than half the budget if the transfers could be targeted on the basis of district or municipality level estimates of poverty. Small area estimates are also increasingly used as regressors in empirical analysis, see e.g. Araujo et al. (2008), Baird et al. (2013), Bazzi (2017), Crost et al. (2014), Demombynes and Ozler (2005), Elbers et al. (2005), Maloney and Caicedo (2015), and Mendez and Van Patten (2022).⁴ To highlight some examples, Araujo et al. (2008), using data from Ecuador, find that villages with higher levels of inequality are less likely to invest in public goods that would benefit the poor (which they attribute to elite capture). Demombynes and Ozler (2005) find that inequality at the small area level has a positive relationship with local crime rates in an empirical application to South Africa. Finally, the multiple imputation methods underpinning small area estimation are also finding applications in the monitoring of poverty over time at a more aggregate level in contexts where household consumption surveys have been discontinued or interrupted, prompting researchers to impute household consumption data into alternative surveys, see e.g., Tarozzi (2007) and Sinha Roy and Van der Weide (2025).

The remainder of this paper is organized as follows. The problem statement is formulated in Section 2. In Section 3 we briefly summarize the two competing approach adopted in the existing literature: Non-normal-Non-EB and Normal-EB. The newly proposed approach, Non-normal-EB, is presented in Section 4. Sections 5 presents a simulation study based on empirical data from 142 household surveys across 16 countries, respectively. Finally, Section 6 concludes.

2 Problem statement: Small area estimation of welfare

Let y_{ah} measure log per capita income (or expenditure) for household h residing in area a , and let s_{ah} denote the number of household members for that same household.⁵ Suppose that at the household level the data generating process (DGP) satisfies:

$$y_{ah} = x_{ah}^T \beta + u_a + \varepsilon_{ah}, \quad (1)$$

where x_{ah} is a vector with independent variables, and where u_a and ε_{ah} are zero expectation error terms that are independent of each other.⁶ The subscripts indicate target area (or domain)

⁴Data combination models can also be applied beyond small area estimation applications, see e.g., Graham et al. (2016).

⁵Ideally the household income or consumption expenditure data is adjusted for spatial and temporal price differences, i.e., is measured in real terms. As this is not the focus of this study, we will be abstracting away from that, but refer the interested reader to e.g., e.g. Gibson et al. (2017) and van Veelen and van der Weide (2008).

⁶The assumed nested error structure is a common and practical way of accounting for spatial correlation between errors. The nested error structure could be extended to a two-fold nested error model, see e.g., Marhuenda et al. (2017). Alternatively one could assume a spatial auto-correlation structure, see e.g., Bell and Bockstael (2000), Pratesi and Salvati (2008), and Kelejian and Prucha (2010). In that case, the correlation between errors could be modeled as a smooth decaying function of distance rather than the step function implied by the nested error structure. The regressors included in x_{ah} typically includes interactions and transformations of selected

a and household h .⁷ Let us assume that errors are homoskedastic, so that for each household h and area a we have: $\text{var}[y_{ah}|x_{ah}] = \sigma_u^2 + \sigma_\varepsilon^2$. Throughout the paper it is assumed that consistent estimators for the variance parameters are available, which we shall denote by $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$.⁸ We will not make any assumptions about the shape of the error distributions at this point.

Let A be the total number of areas covered by the income survey and let n_a be the number of households that have been sampled in area a , so that $n = \sum_{a=1}^A n_a$ equals the total sample size of the survey. We shall denote the total household error by: $e_{ah} = y_{ah} - x_{ah}^T \beta$, and its area a average by $\bar{e}_a = \bar{y}_a - \bar{x}_a^T \beta$ where $\bar{e}_a = \sum_h e_{ah}/n_a$.

We will also use the notation $e_a = (e_{a,1}, \dots, e_{a,n_a})$ which is a vector of length n_a that contains the residuals for all households from area a . With a slight abuse of terminology we will at times refer to the errors e_{ah} and \bar{e}_a as data (as if we know the parameter vector β). Let $y_{(a)}$ and $e_{(a)}$ denote vectors of length N_a with elements y_{ah} and e_{ah} for all households from the population. Similarly, $x_{(a)}$ will denote a matrix with rows given by x_{ah}^T for all N_a households. y_a , e_a and x_a will denote the survey sample analogues. Finally, let the probability distribution functions (probability density functions) for u_a and ε_{ah} be denoted by $F_u(p(u))$ and $G_\varepsilon(p(\varepsilon))$.

The objective is to estimate:

$$E[W(y_{(a)})|x_{(a)}, y_a] = \int W(x_{(a)}\beta + e)p(e|e_a)de \quad (2)$$

$$\simeq \int \int W(x_{(a)}\beta + u + \varepsilon)p(\varepsilon)p(u|e_a)d\varepsilon du, \quad (3)$$

where $p(u|e_a)$ is the conditional probability density function for the location area error u (conditional on observed residuals) and where the welfare function W will generally be non-linear. Popular choices of W are the share of individuals whose income falls below a pre-specified poverty line (also known as the head-count poverty rate) and the Gini index of income inequality. Head-count poverty is a special case of the Foster–Greer–Thorbecke family of poverty indices, see Foster et al. (1984).

In effect EB estimates are obtained by integrating out area errors u_a using probability density $p(u_a|e_a)$, i.e. the probability density of u_a conditional on e_a observed from the survey. Non-EB estimates are obtained by using the unconditional density $p(u_a)$. The challenge is to work out $p(u_a|e_a)$ along with the marginals $p(u_a)$ and $p(\varepsilon_{ah})$ without imposing restrictions about their form so that the resulting conditional density $p(u_a|e_a)$ can also take on any form. Currently, the

covariates to account for area heterogeneities (see Tarozzi and Deaton, 2009; Tarozzi, 2011). See also Das and Chambers (2017) and Lahiri and Salvati (2023). Model selection methods such as Lasso as well as machine learning methods are also increasingly used to identify the relevant model variables (including interactions and transformations). See e.g., Corral et al. (2024) and the references therein for an overview on the use of machine learning in this context.

⁷These areas may refer to geographic areas such as districts or municipalities, but also to non-geographic domains such as ethnic groups or age groups, say.

⁸A commonly used estimator for the variance parameters from a nested error model is Henderson’s method III estimator (see Henderson, 1953; and Searle et al., 1992), which may be viewed as a method of moments estimator which does not require any assumptions about the shape of the error distributions. Alternative estimators are restricted maximum likelihood, minimum norm quadratic unbiased estimation (MINQUE; see e.g. Westfall, 1987; Searle et al., 1992), and so-called spectral decomposition estimation (see e.g. Wang and Yin, 2002; and Wu et al., 2009).

literature on EB estimation avoids this challenge by assuming normally distributed marginals, in which case the conditional distribution will be normal too. Estimation of β is obviously of primary importance when estimating $E[W]$, but in this paper we will concentrate on the error distributions and assume that β is known. Even if β is estimated perfectly, getting the error distribution wrong still has the potential to introduce a bias.

For non-linear W , the expected value $E[W]$ will be a function of the higher moments of the distributions of both u and ε . Getting these moments wrong (i.e., getting the distributions wrong) is one source of bias. Errors in the conditional density $p(u|e_a)$, stemming from misspecification of the error distributions, is another source of bias. The latter bias persists even for linear W . Recently, Skrondal and Rabe-Hesketh (2009) and McCulloch and Neuhaus (2011) have explored the magnitude of the bias that is introduced by getting the first moment of $p(u|e_a)$ wrong and found it to be modest in the case of linear W . Accordingly, our proposed estimator is anticipated to provide more value added in the case of nonlinear W .

Standard errors can be obtained by means of simulation which is ideally suited for estimating quantities that are non-linear functions of the random variables at hand, as is the case with measures of poverty and inequality. Let R denote the number of simulations. The estimator then takes the form:

$$\hat{\mu} = \frac{1}{R} \sum_{r=1}^R W\left(\tilde{y}_{(a)}^{(r)}, s_{(a)}\right), \quad (4)$$

where $\tilde{y}_{(a)}^{(r)}$ denotes the r -th simulated (or predicted) income vector with elements $\tilde{y}_{ah}^{(r)} = x_{ah}^T \tilde{\beta}^{(r)} + \tilde{u}_a^{(r)} + \tilde{\varepsilon}_{ah}^{(r)}$. With each simulation, both the model parameters $\tilde{\beta}^{(r)}$ and the errors $\tilde{u}_a^{(r)}$ and $\tilde{\varepsilon}_{ah}^{(r)}$ are drawn from their estimated distributions.⁹ In the end this gives R simulated poverty rates. The point estimates and their corresponding standard errors are obtained by computing respectively the average and the standard deviation over these simulated values.

3 Two alternative approaches

In this section we will briefly discuss the two most-commonly adopted approaches to date: (1) Empirical Best (EB) prediction advocated by Molina and Rao (2010; MR) which assumes that errors are normally distributed, and (2) the approach put forward by Elbers et al. (2003; ELL) that allows for non-normally distributed errors, but does not adopt EB prediction. Moving forward, we will also refer to these approaches as non-normal-non-EB and normal-EB, respectively.

⁹Our preferred method is to draw $\tilde{\beta}^{(r)}$ by re-estimating the model parameters using the r -th bootstrap version of the survey sample. Alternatively, $\tilde{\beta}^{(r)}$ may be drawn from its estimated asymptotic distribution. The difference between these two alternatives is expected to be modest, unless the survey sample is particularly small so that finite sample effects may play a role.

3.1 ELL: Non-normal-non-EB estimation

ELL makes minimal assumptions about the errors u_a and ε_{ah} . They proceed by first obtaining estimates of the area errors u_a and ε_{ah} , which then allows them to sample from the empirical errors \hat{u}_a and $\hat{\varepsilon}_{ah}$. u_a is estimated as the simple area average of the total residuals (appropriately re-scaled so that the sample variance of \hat{u}_a equals the estimate of σ_u^2 that corrects for the contribution of the area average of ε_{ah}). The estimate for ε_{ah} is obtained by subtracting \hat{u}_a from the total residual (re-scaled so that its sample variance matches σ_ε^2).¹⁰ Note that \hat{u}_a equals the sum of u_a and $\bar{\varepsilon}_a$. The latter term will typically be small in magnitude. It is assumed that its effect on the estimator for the distribution of u_a will be negligible.

The advantage of this approach is that it does not restrict the shape of the error distributions. A shortcoming is that it does not take full advantage of all the available data, i.e., it does not lend itself for Empirical Bayes (EB) estimation. Specifically, it does not offer a means of evaluating the conditional distribution of u_a (conditional on the mean residual for observed for area a in the event area a is included in the survey sample), i.e., $p(u|e_a)$. Working out the conditional distribution is not a trivial exercise without making further distributional assumptions. Consequently ELL decided to forego EB estimation altogether. In doing so, they have accepted a certain loss in efficiency by not fully utilizing all available information.

3.2 Normal-EB estimation

MR put forward an approach that prioritizes EB estimation. The residuals e_a (for area a covered by the survey) clearly carry information about the area random effect u_a . In the extreme where n_a tends to infinity \bar{e}_a will perfectly reveal u_a . They closely follow ELL with the notable difference that the area errors are drawn from the conditional distribution (for areas covered by the survey sample) under the assumption that the errors are normally distributed. As we have seen, under this assumption the conditional distribution of \tilde{u}_a is normal too and its parameter can be easily determined.¹¹ Where ELL accept a loss in precision by not implementing EB estimation, Molina and Rao (2010) accept a loss in precision that might stem from a misspecification of the error distribution functions.

ELL was motivated by estimating poverty and inequality in developing countries where the share of small areas (or domains) that are covered by the income surveys typically ranges between of 5 to 25 percent of all domains in the population. In this case the benefits of EB estimation may be more modest. However, in more developed countries, or countries where travel costs that are incurred when covering all small areas are manageable, income surveys often cover a much larger number of the domains. In fact, there are numerous examples where surveys cover the majority (up to 100 percent) of domains in the country. In those instances, there may be clear benefits to adopting EB estimation.

¹⁰Note that ELL allows for heteroskedasticity, so that σ_ε^2 can be household-specific.

¹¹This is arguably why errors are generally assumed normal in the literature on EB estimation.

4 Non-Normal-EB estimation

The approach presented in this paper builds on both ELL and MR. Like ELL we make no restrictive assumptions about the error distributions. Unlike ELL however, we also accommodate EB estimation. We achieve this by fitting finite normal mixtures (NM) to the error distribution functions. Normal mixtures are extremely flexible; they are able to fit any well-behaved distribution function, and are ideally suited for accommodating EB estimation. If the marginal distributions of u_a and ε_{ah} can be described by normal mixtures, then the conditional distribution too can be described by a normal mixture with known parameters (that are functions of these parameters and of the data on which is being conditioned).

There are several other studies that have explored different ways of relaxing the normality assumption in mixed linear models. Verbeke and Lesaffre (1996) is an early example that also considers normal mixtures as a “non-parametric” representation of the error distribution function. However, they impose a number of important restrictions. First, only the area random effects u_a are allowed to be non-normal; a normal-mixture is fitted to the distribution of u_a under the assumption that ε_{ah} is normally distributed. Second, it is assumed that the “component distributions” that make up the normal mixture share a common variance, which noticeably simplifies estimation but at the same time significantly limits the flexibility of the normal mixture to fit any given distribution function. This approach has also been followed by Cordy and Thomas (1997) who work with the same setup and adopt the same set of restrictions. There is another strand of the literature that permits both error terms to be non-normally distributed by imposing an alternative parametric family for the distribution functions. See for example Zhou and He (2008) who fit skewed t-distributions to the nested errors of the linear mixed model.¹² Recently, there have also been efforts to explore the impact of misspecifications in the error distributions for Empirical Bayes predictions derived from linear mixed models, see for example Skrandal and Rabe-Hesketh (2009) and McCulloch and Neuhaus (2011). They conclude that the bias is reasonably small. It should be noted however that those studies focus on prediction of the dependent variable itself, in which case the Empirical Bayes estimates of the area random effects u_a denote the only source of bias. Misspecifications in the error distributions become considerably more important when predicting nonlinear functions of the dependent variable, such as measures of poverty and inequality, as we will show in this paper.

4.1 Specification

Let us assume that F_u and G_ε can be represented by mixture distributions:

$$F_u = \sum_{i=1}^{i=m_u} \pi_i F_i \quad (5)$$

$$G_\varepsilon = \sum_{j=1}^{j=m_\varepsilon} \lambda_j G_j, \quad (6)$$

¹²Diallo and Rao (2018) relax the normality assumption by considering skew-normal errors.

where the F_i 's and G_j 's denote a basis of distribution functions which we will also refer to as components or component distribution functions. The π_i 's and λ_j 's denote unknown non-negative probabilities that satisfy $\sum_i \pi_i = 1$ and $\sum_j \lambda_j = 1$, which we will also refer to as mixing probabilities.¹³ m_u and m_ε denote the number of components used to represent F_u and G_ε , respectively. We will denote the probability density functions associated with F_i and G_j by respectively f_i and g_j .

Mixture distributions are remarkably well equipped to fit any well-behaved distribution function. For example, kernel density estimators are closely related to mixture distributions. We will be working with normal component distributions, so that the mixture distributions are normal mixtures.

Assumption 1 *The components F_i are normal distribution functions with mean μ_i and variance σ_i^2 . Similarly, components G_j are normal distribution functions with mean ν_j and variance ω_j^2 .*

Note that the modeler is at liberty to work with a different basis of component distributions. This choice does not have real implications for the ability of the mixture to fit a given distribution function.

If $p(u_a)$ and $p(\varepsilon_{ah})$ are normal-mixtures, then $p(u_a|\bar{e}_a)$ is a normal mixture too. This is a powerful result as the integral in equation (3) will generally have to be computed by simulation, while sampling from normal mixtures is straightforward. Lemma 2 shows how the parameters that define $p(u_a|\bar{e}_a)$ can be obtained as a function of the parameters of the normal-mixtures $p(u_a)$ and $p(\varepsilon_{ah})$. Implementing EB estimation is thus as easy as sampling the area errors from the normal-mixture $p(u_a|\bar{e}_a)$ whenever \bar{e}_a is observed in the survey sample.

Lemma 2 *The probability density function of u_a conditional on \bar{e}_a , which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:*

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{m_u} \sum_{j=1}^{m_\varepsilon} \sum_{k=1}^{m_\varepsilon} w_{ijk} \varphi(u_a; m_{ijk}; s_{ijk}^2), \quad (7)$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\sigma_i^2}{\sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a} \right) (\bar{e}_a - (\nu_j + \nu_k)/n_a) + \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2} \right) \mu_i \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2) \sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2} \right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \quad (8)$$

¹³An alternative specification, where a mixture is applied to the full nested error model, is explored in Bikauskaite et al. (2020). In effect, this assumes that the normal mixture distributions for the area and household errors share the same mixing probabilities, rather than specifying flexible normal mixtures for the area- and idiosyncratic error distributions separately.

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε , respectively.

If we apply the Central Limit Theorem to approximate the marginal distribution of $\bar{\varepsilon}_a$ to a normal distribution with mean zero and variance equal to σ_ε^2/n_a , the expression for $p(u_a|\bar{\varepsilon}_a)$ simplifies considerably:

$$p(u_a|\bar{\varepsilon}_a) \approx \sum_{i=1}^{i=m_u} \alpha_i \varphi \left(u_a; \gamma_{ai}\bar{\varepsilon}_a + (1 - \gamma_{ai})\mu_i; (1/\sigma_i^2 + n_a/\sigma_\varepsilon^2)^{-1} \right), \quad (9)$$

with $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha_i = \tilde{\alpha}_i / \sum_i \tilde{\alpha}_i$ with:

$$\tilde{\alpha}_i = \pi_i \varphi(\bar{\varepsilon}_a; \mu_i; \sigma_i^2 + \sigma_\varepsilon^2/n_a). \quad (10)$$

The expected value of u_a conditional on $\bar{\varepsilon}_a$, given the density function $p(u_a|\bar{\varepsilon}_a)$ from Lemma 2, is seen to solve:

$$E[u_a|\bar{\varepsilon}_a] \approx \sum_i \alpha_i(\bar{\varepsilon}_a) (\gamma_{ai}\bar{\varepsilon}_a + (1 - \gamma_{ai})\mu_i), \quad (11)$$

where $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha_i(\bar{\varepsilon}_a)$ denotes the mixing probabilities of $p(u_a|\bar{\varepsilon}_a)$.

Note that the standard assumption of normal errors is nested as a special case, where there is just one component with $\mu_i = 0$ and $\sigma_i^2 = \sigma_u^2$. The first and second moment of the normal conditional density $p(u_a|\bar{\varepsilon}_a)$ in this case are seen to solve:

$$E[u_a|\bar{\varepsilon}_a] = \gamma_a \bar{\varepsilon}_a \quad (12)$$

$$\text{var}[u_a|\bar{\varepsilon}_a] = (1 - \gamma_a)\sigma_u^2, \quad (13)$$

where $\gamma_a = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2/n_a)$ (see e.g. Molina and Rao, 2010).¹⁴ For non-normal errors, we have that $E[u_a|\bar{\varepsilon}_a]$ is generally a non-linear function of $\bar{\varepsilon}_a$, and $\text{var}[u_a|\bar{\varepsilon}_a]$ will generally be a function of $\bar{\varepsilon}_a$.

4.2 Estimation

The objective is to estimate the parameters $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$, which determine the normal mixtures (NM) $F_u(u) = \sum_{i=1}^{i=m_u} \pi_i F_i(u; \mu_i, \sigma_i^2)$ and $G_\varepsilon(\varepsilon) = \sum_{j=1}^{j=m_\varepsilon} \lambda_j G_j(\varepsilon; \nu_j, \omega_j^2)$, respectively, where m_u and m_ε denote the number of components. Estimation of NM to observed data is a routine task. A commonly used method of estimation is the Expectation-Maximization (EM) algorithm, which is an iterative method to find local maximum likelihood estimates of the parameters of interest, see e.g. Dempster et al. (1977) and McLachlan and Peel, (2000).

Suppose that some data y is drawn from a NM with m components and parameter vector θ . The EM algorithm introduces a latent random variable z_{ij} that equals 1 if observation y_j has been drawn from component i , and 0 otherwise. Let the log-likelihood function that treats both

¹⁴Note that the unconditional variance solves: $\text{var}[u_a] = \text{var}[\gamma_a \bar{\varepsilon}_a] + (1 - \gamma_a)\sigma_u^2 = \gamma_a^2(\sigma_u^2 + \sigma_\varepsilon^2/n_a) + (1 - \gamma_a)\sigma_u^2$, which equals $\text{var}[u_a] = \gamma_a \sigma_u^2 + (1 - \gamma_a)\sigma_u^2 = \sigma_u^2$, since $\sigma_u^2 + \sigma_\varepsilon^2/n_a = \sigma_u^2/\gamma_a$.

y and z as data be denoted by $L(\theta; y, z)$, and its derivative with respect to θ by $L_\theta(\theta; y, z)$. The corresponding maximum-likelihood (ML) estimator solves the following moment condition:

$$E[L_\theta(\theta; y, z)] = E[E[L_\theta(\theta; y, z)|y]] = 0. \quad (14)$$

The EM algorithm essentially solves the same moment condition but iteratively:

$$E[E[L_\theta(\theta; y, z)|y, \hat{\theta}^{(k)}]] = 0, \quad (15)$$

where $\hat{\theta}^{(k)}$ denotes the iteration- k estimate for θ . Solving equation (15) yields $\hat{\theta}^{(k+1)}$. The procedure results in a non-decreasing sequence of likelihoods $L(y; \hat{\theta}^k)$. The advantage of the EM algorithm is that solving equation (15) is often considerably easier than solving equation (14). Note that the unobserved data z is integrated out conditional on the observed data y and the current estimate of θ .

5 Simulation study

5.1 Setup

To evaluate the relative performance of the different small area estimation methods in realistic settings, we implement a simulation study where the population census and household surveys are based on data from actual household income and expenditure surveys across 16 different countries and multiple time periods. Consider the DGP of equation (1) repeated below:

$$y_{ah} = x_{ah}\beta + u_a + \varepsilon_{ah}. \quad (16)$$

For explanation of symbols and assumptions see the discussion following equation (1). The DGP specifies three statistically independent determinants of a household's (log) per capita expenditure or income: a structural or 'fixed' part, $x_{ah}\beta$, an area random effect, u_a , and an household-specific random effect ε_{ah} .

5.1.1 Fixed part: $x\beta$

For the fixed part we use the expected household log per capita income from Brazil's 2010 census, as predicted from linear regression. We argue that this is more appropriate to represent the fixed part, with empirically observed cross-area variation, than making arbitrary assumptions on the distribution of $x_{ah}\beta$. More specifically we use the data for Minas Gerais from the population census. This census covers more than 10% of the population. Uniquely, the census includes a household income variable. Log per capita household income is regressed on a set of explanatory variables that are commonly included in empirical applications of small area poverty estimation: household size, education, age, housing characteristics, assets, access to services, etc. This allows us to compute $x_{ah}\hat{\beta}$ for all households in the Minas Gerais census sample. At 60% the R^2 of the regression is comparatively high but the purpose of the regression is to compile a 'fixed'

part of the DGP in equation (1), to be used in otherwise synthetic data. The Minas Gerais census sample is further limited to municipalities with more than 1000 households in the sample resulting in a dataset of 345733 households from 128 municipalities. The municipality means of $x_{ah}\hat{\beta}$ range from 5.7 to 6.67; the standard deviations range from 0.54 to 0.87.

5.1.2 Random part: u_a and ε_{ah}

The random parts, the area- and household error u_a and ε_{ah} , are based on empirical realizations of the errors derived from an inclusive set of income and expenditure regressions that spans 16 countries across multiple years, stratified by urban/rural, for a total of 142 surveys. The full list of household surveys used is documented in Appendix section 3. Taking stock of the range of error distributions observed in real data across a wide spectrum of different countries and time periods offers a first empirical contribution.

Henderson’s method III is used to decompose the total residuals in a part common to all households from a given area (the area error term u_a) and a household-specific component (the idiosyncratic error term ε_{ah}).¹⁵ We proceed by estimating non-parametric density functions for the area- and household errors, and subsequently draw u_a from the area-level density and ε_{ah} from the household-level density. Given realizations of both the fixed and random components, household log income y_{ah} is computed as:

$$y_{ah} = x_{ah}\hat{\beta} + \lambda_1 u_a + \lambda_2 \varepsilon_{ah}, \quad (17)$$

where scaling parameters $\lambda_{1,2}$ allow us to control the relative size of area and household effects in the simulations. The errors u_a and ε_{ah} in this case refer to normalized independent random variables with mean zero and unit variance, such that the location effect (share of total error variance due to the area error) equals: $s = \lambda_1^2 / (\lambda_1^2 + \lambda_2^2)$. The following shares of the location effect are considered: 0, 0.025, 0.05, 0.075, and 0.1, which represents the typical range of values observed in poverty mapping applications. A share of 0.1 would be high with the majority of empirically observed location shares ranging between 0.02 and 0.08. Keeping the size of the location effect fixed across different country-years helps isolate the role of deviations from normality as all variation observed across country-years is purely driven by variation in the degree of non-normality. The overall scale of $\lambda_1 u_a + \lambda_2 \varepsilon_{ah}$ is chosen such that a regression of y_{ah} on $x_{ah}\hat{\beta}$ is expected to result in an R^2 of 50%, a value typically found with income regressions.

After a complete synthetic census of (log) incomes has been generated for households in the Minas Gerais census using equation (17), household per capita incomes are then used to classify households as poor or non-poor. From this, we compute area-level poverty head-count rates (in terms of households). Two choices of poverty lines are considered. One that yields a headcount rate of 12.5 percent and one that yields a rate of 25 percent. The results indicate that this choice is of lesser significance. Particular simulations are defined by poverty line, variance share

¹⁵See Pérez, Peña and Molina (2011).

of the location effect, country, year and urban status.

5.1.3 Drawing of the household survey

Starting from the census dataset, we draw a balanced sample of households clustered by area (i.e., Minas Gerais municipality) with a cluster size of 20 households. A key parameter in survey sampling designs is the share of areas covered by the survey. Increasing this coverage will naturally benefit applications of Empirical Bayes estimation, but will also increase the costs of data collection (by increasing travel costs which constitutes a key component of total data collection costs). In practice this share can be as low as 10 and as high as 100 percent, with higher income countries often opting for higher coverage rates. In the original empirical application to Ecuador by ELL, the coverage rate was 17 percent (and the location effect as a share of total error is estimated at 0.02). Instead of varying the coverage rate, which would introduce another parameter, we consider the two limiting scenarios: (a) one where all areas are included in the sample (100 percent coverage), and (b) the other extreme where the number of areas covered by the survey is negligible, which is implemented by not employing EB estimation in any of the areas. By considering both extremes, we economize on the number of parameters to be varied in the simulation study.

Given that the focus of this study is on relaxing the normality assumption in applications of Empirical Bayes estimation, the sampling design is assumed to be non-informative. Allowing for informative sampling designs, building on the work by e.g., Pfeiffermann et al. (1998), Pfeiffermann et al. (2006), and Pfeiffermann and Sverchkov (2007), is a natural extension that could be considered in future work.

5.1.4 Small area estimation methods

All methods work with the same estimates of the fixed part $x\beta$.¹⁶ They only differ in how the errors u_a and ε_{ah} are imputed. These are then combined with the predicted fixed part of household log per capita income to yield household-level per capita income and area-level poverty predictions.

1. ELL: This approach is in effect non-normal-non-EB estimation. It was developed for a context where surveys provide incomplete coverage of small areas in which case the scope for applying Empirical Bayes is reduced. Area effects and household effects are simply drawn (with replacement) from the empirical estimates of u_a and ε_{ah} obtained using Henderson-method-3 to account for possible non-normality in errors.
2. Normal-EB: This method assumes that u_a and ε_{ah} are normally distributed. We assume zero mean distributions for the error components with variances as estimated by Henderson-method-3. Area effects are drawn conditionally on the residuals for survey households.

¹⁶The coefficient estimates β used to impute the fixed part are subject to sampling error due to the particular survey that has been drawn.

3. Non-normal-EB: Like ELL this method draws each household effects directly (with replacement) from the estimated idiosyncratic errors. Area effects u_a are drawn from a conditional normal mixture according to Lemma 2. Conditioning is on the estimated area component for each area.

Another candidate method is to resample total errors per area. Since there are as many as 20 households per area in the survey, it would be interesting to see if simply resampling total regression residuals per area (i.e., not attempting to split residuals in area and household components) performs well. It has the additional advantage of being robust to heteroskedasticity of error components across areas. We implemented this approach but do not report results for this method. While it outperforms ELL it is inferior to non-normal-EB, which is the main focus of this study.

5.1.5 Estimation of normal mixture distributions

Obtaining estimates of the conditional normal mixtures for the area errors used in the non-normal-EB method involves two steps. First, fit a normal mixture to the estimated area errors u_a (i.e., estimate the unconditional area error distribution). Second, apply Lemma 2, which uses the estimated unconditional normal-mixture as an input, to obtain estimates of the conditional normal mixtures for the area errors. For the estimation of the unconditional normal mixture distributions to the area errors, we adopt the EM algorithm. Identifying the number of components can require experimentation. To avoid this we have automated the search for mixtures, which involves fitting normal mixtures with 1, 2, and 3 components, as follows:¹⁷

- a. As initial values use for the means 0 in the case of 1 component, -1 and 1 in the case of 2 components, and -1, 0, and 1 in the case of 3 components
- b. Run the EM algorithm until convergence or stop after 1000 iterations
- c. Integrate absolute difference between the estimated density function and the kernel density function of the normal-mixture approximation
- d. Choose the normal mixture with the smallest difference

This automated procedure facilitates application across the large number of different surveys. Although this yields satisfactory approximations it can be argued that better approximations can be achieved with more sophisticated strategies (or tailoring specification to dataset at hand). In that sense whatever advantage our simulation method has over the other approaches can be considered a lower bound. Alternatively, one could consider restricting the variance of the components distributions to be the same across components and offset this restriction by increasing the number of components (akin to fitting a kernel density). The results obtained with this choice of normal mixture estimator (not reported here) are practically identical.

¹⁷Note that 1 component corresponds to a normal distribution.

5.1.6 Performance metric

All methods impute log per capita incomes at the household level, which translates to area-level head-count rates and, by comparison with the ‘actual’ head-count rates from the synthetic census, an overall Mean Squared Error (MSE). This study focuses on overall aggregate performance of the methods. A single survey is drawn from the synthetic census for each imputation, which is repeated 100 times.

For each area, the 100 imputations yield an average simulated head-count poverty rate, $\hat{\mu}_a$, which is compared to the ‘true’ poverty rate μ_a from the synthetic census. Our main performance indicator is the MSE of $\hat{\mu}_a$. Let n_a be the number of households in area a and $N = \sum_a n_a$, then:

$$MSE = \frac{1}{N} \sum_a n_a (\hat{\mu}_a - \mu_a)^2. \quad (18)$$

A version of the MSE where areas are not population weighted is also implemented. This yields nearly identical results.

5.2 Empirically observed error distributions

Let us first take stock of the empirical distributions for the area and household errors that are observed across the 142 different surveys (from 16 different counties spanning multiple years). The degree of non-normality is evaluated through estimates of the skewness and kurtosis. Figure 1 plots the kernel densities for the estimated skewness and kurtosis of u_a and ε_{ah} , respectively. The estimated moments point to significant deviations from normality (when skewness is zero and kurtosis 3), especially for the household idiosyncratic component. The skewness of area errors ranges between -1 and 1, and kurtosis from 2 to 8, disregarding outliers. Household errors have skewness ranging from -1.5 to 1.5 and kurtosis from 4 to 10.

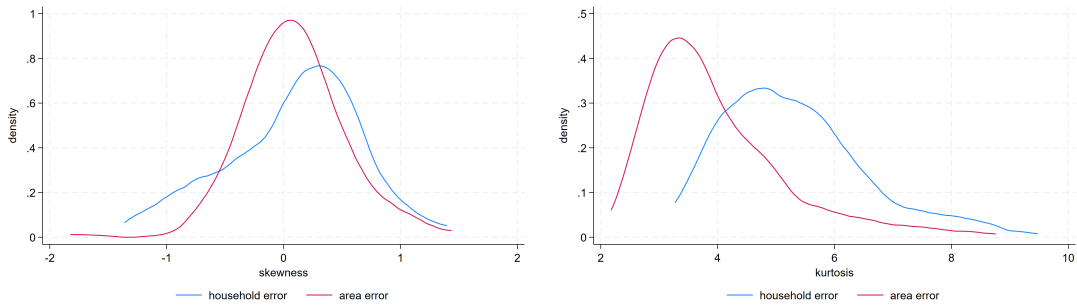


Figure 1: Kernel density for skewness (left) and kurtosis (right)

The empirical relationship between the moments for the area- and the household-error components are shown in Figure 2. It can be seen that the two moments exhibit an U-shaped relationship, with minimum kurtosis if skewness is around zero.

We can similarly verify whether the moments of the area error are correlated with the moments of the household error, i.e., whether a skewed (or heavy tailed) household error tends

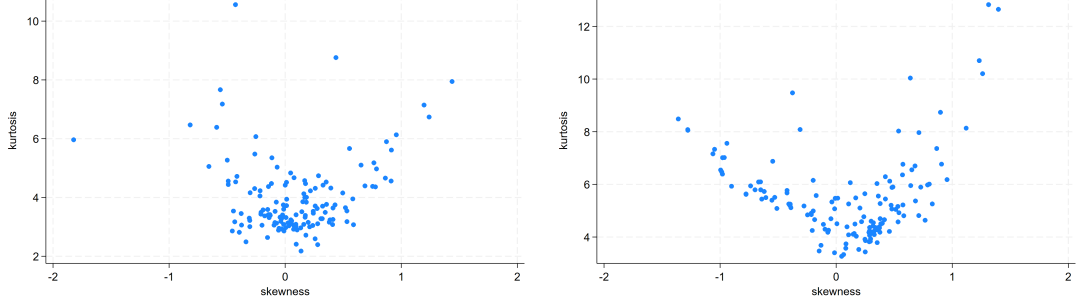


Figure 2: Kurtosis versus skewness for area errors (left) and idiosyncratic errors (right)

to coincide with a skewed (or heavy tailed) area error. Figure 3 confirms that skewness of household and area errors show a positive association. In summary, for the majority of surveys, the empirically observed error distributions show significant deviations from normality.¹⁸

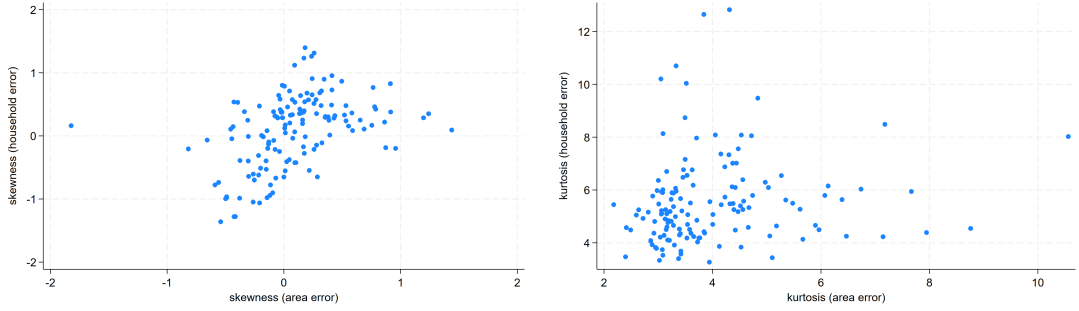


Figure 3: Household error versus area error moments: skewness (left) and kurtosis (right)

5.3 Small area estimation results

Normal-EB is expected to do well when the location effect is comparatively large and the degree of non-normality is small, while ELL is expected to do well when the location effect is small and the degree of non-normality is large. Non-normal-EB is expected to do well in either scenario. Our simulations confirm these expectations. By considering different values of the location effect and varying degrees of non-normality, we can assess at what point ELL and normal-EB are at par.

Figure 4 plots the Root Mean Squared Error (RMSE) against the degree of non-normality of errors as measured by skewness of the idiosyncratic error in the case of a 12.5 headcount poverty rate.¹⁹ The graphs are smoothed summaries of simulation results obtained by fitting kernel

¹⁸A Shapiro-Wilk normality test is rejected in all cases for the household error components and in 72% of the cases for the area components ($p=0.05$).

¹⁹Given that the moments of the individual error and moments of the area error are highly correlated, our figures effectively show how RMSE varies with non-normality in errors (without isolating non-normality in individual errors from non-normality in area errors). Having said that, our results show what gains in RMSE one can reasonably expect in practical settings using empirically observed error distributions across a wide range of countries, where non-normality in individual errors and area errors are seen to co-vary.

regressions. Note that normal and ELL represent the counter-parts of normal-EB and non-normal-EB for areas that are not covered by the survey. Estimates of the RMSE corresponding to a headcount poverty rate of 25 percent are presented in Appendix section 4.

Our results indicate that ELL outperforms normal-EB when the location effect is near-zero and errors are non-normally distributed, while normal-EB is seen to significantly outperforms ELL when the location effect is comparatively large (non-normality of errors notwithstanding). The two approaches are found to be at par for location effects in the range of 2.5 – 5 percent (with upper range corresponding to cases with higher degrees of non-normality observed in empirical data). ELL’s advantage dissipates quickly when location effects are more important. Figure 4 shows this for a location variance share of as little as 2.5%. Also, note that (as one would expect) RMSE is very sensitive to location effects, increasing more than threefold across the different choices of the location effect shares. Finally, the results obtained for the poverty that yields a headcount rate of 25 percent, which are qualitatively identical, are included in Appendix section 4.

Our estimates also confirm that non-normal-EB records the lowest RMSE in all cases. The difference (in RMSE) when compared to normal-EB is however comparatively marginal at higher levels of the location effect. Similarly, the difference is marginal when compared to ELL at near-zero levels of the location effect. By the same token, there is no downside to adopting non-normal-EB. The gains of non-normal-EB over normal-EB and ELL are between marginal and notable, but never negative.

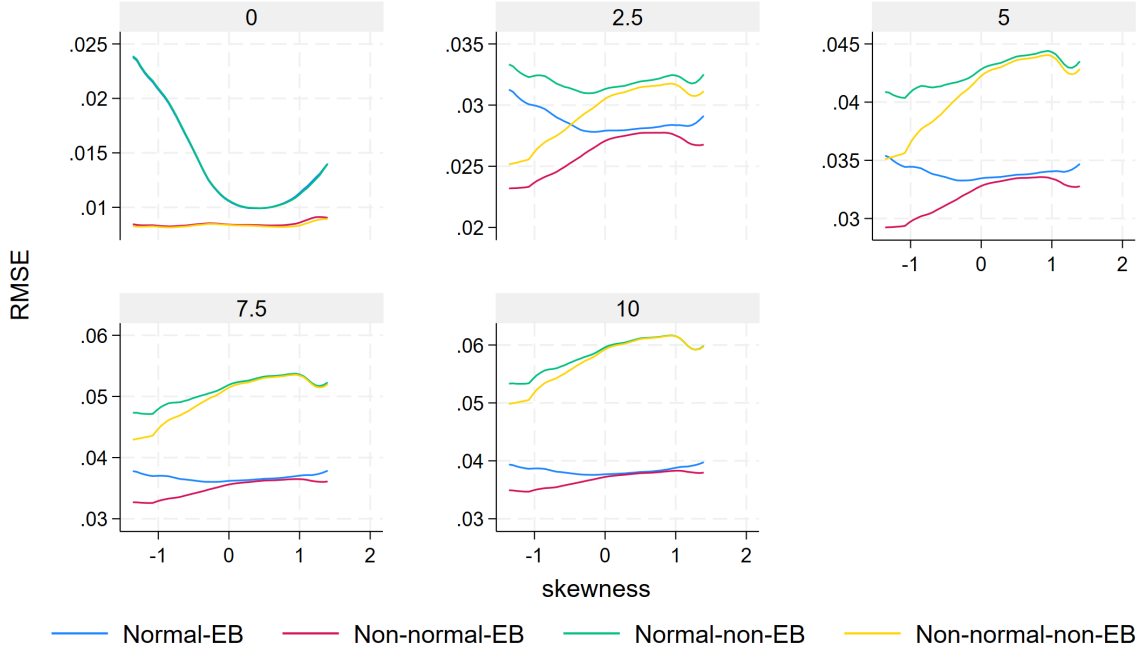


Figure 4: RMSE versus skewness for different magnitudes of the location effect (as a percentage of total error)

One notable feature shown in Figure 4 is that negative skewness tends to result in sizable differences in RMSE, more than a positive skewness does. With kurtosis we see the difference increase if kurtosis is increasing. Such patterns are driven by the step nature of the headcount as a function of income on the one hand, and on the other hand by the distribution of $x\beta$, the poverty line and the actual skewness (and other moments) of the error distributions used in the simulations. There is therefore no reason to expect that for e.g. skewness the pattern should be symmetric.

To explore the different effects of positive and negative skewness we let skewness range from -0.453 to 0.453 by mixing a density estimator $p_\varepsilon(\varepsilon)$ of the individual residuals from Bangladesh (rural Bangladesh survey from 2000) with a normal distribution as follows:

$$p_{mix}(\varepsilon; \lambda) = |\lambda|p_\varepsilon(\text{sign}(\lambda)\varepsilon) + (1 - |\lambda|)p_{norm}(\varepsilon), \quad (19)$$

and letting λ range between -1 (corresponding to skewness -0.453) and +1 (skewness +0.453). The average difference (weighted by $x\beta$ frequencies) between skewed and normal poverty probabilities is graphed in Figure 5 (left panel), showing indeed a larger difference for negative skewness. The overall squared difference is pictured in Figure 5 (right panel), with the same conclusion.

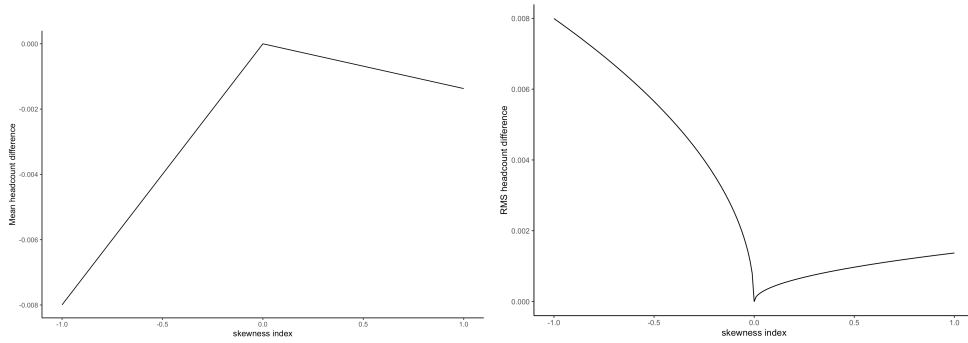


Figure 5: Effect of skewness on expected headcount poverty

6 Concluding remarks

Normal-EB prediction is expected to do well relative to ELL (non-normal-non-EB prediction) when: (i) the area error makes up a significant share of total error, (ii) the errors exhibit modest deviations from normality, and (iii) the household survey covers a large share of the total number of target areas. Using a large-scale simulation study based on actual household survey data from 142 different survey rounds across 16 different countries, thereby working with empirically relevant error distributions, we can confirm this prediction. In cases where the location effect makes up 5 percent or more of total error, normal-EB will outperform ELL (i.e., achieve lower RMSE) for all empirically observed degrees of non-normality. In those cases where the location effect is comparatively small and the deviations from normality are more

meaningful however, the ordering may be reversed.

The non-normal-EB approach put forward in this paper is expected to do well in either of the different scenarios described above. This too is confirmed in the simulation study. Depending on the scenario, the increase in precision of non-normal-EB relative to both normal-EB and ELL can range from meaningful (up to a 15-25 percent reduction in RMSE) to negligible (near-zero reduction in RMSE), but it is always positive. Given that the proposed non-normal-EB approach is furthermore easy to implement, there is no downside to using it. Note that the approach is equally relevant for survey-to-survey prediction, where similar improvements in precision may be expected.

An alternative means for handling non-normality in errors is to consider transformations of the dependent variable such that the transformed variable is more normally distributed, see for example Rojas Perilla et al. (2020) and the references therein. An investigation into combining this approach with the one presented in this paper is something that could be explored in future research. Another potential avenue for future research is to extend non-normal-EB estimation to non-linear models such as logistic regression models (see e.g., Farrell, 1997) and multiple regression contexts (see e.g., Isaki, 1990).

References

- Araujo, M., Ferreira, F., Lanjouw, P. and Ozler, B. (2008). Local inequality and project choice: Theory and evidence from ecuador. *Journal of Public Economics*, **92**, 1022–1046.
- Baird, S., McIntosh, C. and Ozler, B. (2013). The regressive demands of demand-driven development. *Journal of Public Economics*, **106**, 27–41.
- Banerjee, A., Hanna, R., Olken, B. and Lisker, D. (2025). Social ptotection in the developing world. *Journal of Economic Literature*.
- Bazzi, S. (2017). Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, **9**, 219–255.
- Bell, K. and Bockstael, N. (2000). Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *Review of Economics and Statistics*, **82**, 72–82.
- Bikauskaite, A., Molina, I. and Morales, D. (2020). Multivariate mixture model for small area estimation of poverty indicators. *Journal of the Royal Statistical Society Series A*, **185**, S724–S755.
- Burke, M., Driscoll, A., Lobell, D. and Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, **371**.
- Chi, G., Fang, H., Chatterjee, S. and Blumenstock, J. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, **119**, number 3, 1–11.

- Cordy, C. and Thomas, D. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association*, **92**, 1459–1465.
- Corral, P., Henderson, H. and Segovia, S. (2024). Poverty mapping in the age of machine learning. *Journal of Development Economics*.
- Corral, P., Molina, I. and Nguyen, M. (2020). Pull your small area estimates up by the bootstraps. *World Bank Policy Research Working Paper*, **9256**.
- Croft, B., Felter, J. and Johnston, P. (2014). Aid under fire: Development projects and civil conflict. *American Economic Review*, **104**, 1833–1856.
- Das, S. and Chambers, R. (2017). Robust mean-squared error estimation for poverty estimates based on the method of elbers, lanjouw and lanjouw. *Journal of the Royal Statistical Society Series A*, **180**, 1137–1161.
- Das, S. and Haslett, S. (2019). A comparison of methods for poverty estimation in developing countries. *International Statistical Review*, **87**, 368–392.
- Demombynes, G. and Ozler, B. (2005). Crime and local inequality in south africa. *Journal of Development Economics*, **76**, 265–292.
- Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Serie B*, **39**, 1–38.
- Diallo, M. and Rao, J. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, **45**, 1092–1116.
- Elbers, C., Fujii, T., Lanjouw, P., Ozler, B. and Yin, W. (2007). Poverty alleviation through geographic targeting: How much does disaggregation help? *Journal of Development Economics*, **83**, number 1, 198–213.
- Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.
- Elbers, C., Lanjouw, J. and Lanjouw, P. (2005). Crime and local inequality in south africa. *Journal of Economic Geography*, **5**, 101–118.
- Engstrom, R., Hersh, J. and Newhouse, D. (2022). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. *World Bank Economic Review*, **36**, 382–412.
- Farrell, P., MacGibbon, B. and Tomberlin, T. (1997). Empirical bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, **15**, 101–108.

- Fay, R. and 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.
- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.
- Gibson, J., Le, T. and Kim, B. (2017). Prices, engel curves, and time-space deflation: Impacts on poverty and inequality in vietnam. *World Bank Economic Review*, **31**, 504–530.
- Graham, B., Campos de Xavier Pinto, C. and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, **34**, 288–301.
- Haslett, S. (2013). Small area estimation of poverty using the ell/povmap method, and its alternatives. *Poverty and Social Exclusion: Chapter 12* 224–245.
- Haslett, S. (2016). Small area estimation using both survey and census unit record data. *Handbook on Analysis of Poverty Data by Small Area Estimation: Chapter 18* 328–348.
- Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–253.
- Isaki, C. (1990). Small-area estimation of economic statistics. *Journal of Business and Economic Statistics*, **8**, 435–441.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, **353**, 790–794.
- Kelejian, H. and Prucha, I. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, **157**, 53–67.
- Lahiri, P. and Salvati, N. (2023). A nested error regression model with high-dimensional parameter for small area estimation. *Journal of the Royal Statistical Society Series B*, **85**, 212–239.
- Maloney, W. and Caicedo, F. (2015). The persistence of (subnational) fortune. *Economic Journal*, **126**, 2363–2401.
- Marhuenda, Y., Molina, I., Morales, D. and Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **180**, number 4, 1111–1136.
- Marx, B., Stoker, T. and Suri, T. (2019). There is no free house: Ethnic patronage in a kenyan slum. *American Economic Journal: Applied Economics*, **11**, 36–70.
- McCulloch, C. and Neuhaus, J. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, **26**, 388–402.

- McLachlan, G. and Peel, D. (2000). *Finite Mixture Model*. New York: John Wiley & Sons.
- Mendez, E. and Van Patten, D. (2022). Multinationals, monopsony, and local development: Evidence from the united fruit company. *Econometrica*, **90**, 2685–2721.
- Merfeld, J. and Newhouse, D. (2023). Improving estimates of mean welfare and uncertainty in developing countries. *World Bank Policy Research Working Paper*, **10348**.
- Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38**, 369–385.
- Newhouse, D. (2023). Small area estimation of poverty and wealth using geospatial data: What have we learned so far? *Calcutta Statistical Association Bulletin*, **76**, 7–32.
- Newhouse, D., Merfeld, J., Ramakrishnan, A., Swartz, T. and Lahiri, P. (2022). Small area estimation of monetary poverty in mexico using satellite imagery and machine learning. *World Bank Policy Research Working Paper No. 10175*.
- Perez, B., Pena, D. and Molina, I. (2011). Robust henderson iii estimators of variance components in the nested error model. *Universidad Carlos III De Madrid Working Paper*, **11-43-32**.
- Pfeffermann, D., Da Silva Moura, F. and Do Nascimento Silva, P. (2006). Multi-level modelling under informative sampling. *Biometrika*, **93**, 943–959.
- Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.
- Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, **102**, 1427–1439.
- Pokhriyal, N. and Jacques, D. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*.
- Pratesi, M. and Salvati, N. (2008). Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17**, 113–141.
- Rao, J. and Molina, I. (2015). *Small area estimation*. New York: John Wiley and Sons.
- Rojas Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A*, **183**, 121–148.
- Searle, S., Casella, G. and McCulloch, C. (1992). *Variance components*. New York: Wiley.
- Sinha Roy, S. and van der Weide, R. (2025). Estimating poverty for india after 2011 using private-sector survey data. *Journal of Development Economics*, **172**.

- Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society*, **172**, 659–687.
- Smythe, I. and Blumenstock, J. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, **119**, number 32, 1–10.
- Tarozzi, A. (2007). Calculating comparable statistics from incomparable surveys, with an application to poverty in india. *Journal of Business and Economic Statistics*, **25**, 314–336.
- Tarozzi, A. (2011). Can census data alone signal heterogeneity in the estimation of poverty maps? *Journal of Development Economics*, **95**, 170–185.
- Tarozzi, A. and Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics*, **91**, 773–792.
- Tzavidis, N., Zhang, L., Luna, A., Schmid, T. and Rojas Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society Series A*, **181**, 927–979.
- Van der Weide, R., Blankespoor, B., Elbers, C. and Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? an application to malawi. *Journal of Development Economics*, **171**.
- Van Veelen, M. and van der Weide, R. (2008). A note on different approaches to index number theory. *American Economic Review*, **98**, 1722–1730.
- Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association*, **91**, 217–221.
- Wang, S. and Yin, S. (2002). A new estimate of the parameters in linear mixed models. *Science in China Series A: Mathematics*, **45**, 1301–1311.
- Westfall, P. (1987). A comparison of variance component estimates for arbitrary underlying distributions. *Journal of the American Statistical Association*, **82**, 866–874.
- Wu, M., Yu, K. and Liu, A. (2009). Estimation of variance components in the mixed effects models: A comparison between analysis of variance and spectral decomposition. *Journal of Statistical Planning and Inference*, **139**, 3962–3973.
- Zhou, T. and He, X. (2008). Three-step estimation in linear mixed models with skew-t distributions. *Journal of Statistical Planning and Inference*, **138**, 1542–1555.

Appendix 1 A lemma and corollary that are used in the proofs of Lemmas 2

Lemma 3 *Let v_1 and v_2 be two independent normal random variates, with $Ev_i = \mu_i$, $\text{var } v_i = \sigma_i^2$ for $i = 1, 2$. Let $v_3 = v_1 + v_2$. Denote the normal density function with expectation μ and variance σ^2 by $\varphi(\cdot; \mu, \sigma)$. Then for the joint density $p(v_1, v_3)$ and the conditional density $p(v_1|v_3)$ we have*

$$\begin{aligned} p(v_1, v_3) &= \varphi(v_1; \mu_1, \sigma_1) \varphi(v_3 - v_1; \mu_2, \sigma_2) \\ &= p(v_1|v_3) p(v_3) \\ &= \varphi\left(v_1; \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(v_3 - \mu_2) + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1, \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \\ &\quad \times \varphi\left(v_3; \mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right). \end{aligned}$$

Proof This can be directly verified. Alternatively, note that both v_3 and $v_1|v_3$ are normally distributed. The distribution of $v_1|v_3$ has expectation

$$E(v_1|v_3) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(v_3 - \mu_2) + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1$$

and variance

$$\text{var}(v_1|v_3) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

□

Corollary 4 *Let x and y be independent random variables distributed as normal mixtures, and let $q = x + y$. Then (i) q is also distributed as a normal mixture, and (ii) the conditional distributions $p(x|q)$ and $p(y|q)$ are also normal mixtures. Moreover, the mixing parameters of $p(q)$, $p(x|q)$ and $p(y|q)$ are closed expressions in terms of the distributional parameters of x and y . In particular, if the density function of x is*

$$f(x) = \sum_{k=1}^a \pi_k \varphi(x; \mu_k, \sigma_k),$$

(with $\pi_k > 0$ and $\sum_k \pi_k = 1$) and the density function of y is

$$g(y) = \sum_{j=1}^b \lambda_j \varphi(y; \nu_j, \tau_j),$$

(with $\lambda_j > 0$ and $\sum_j \lambda_j = 1$) then the conditional distribution of x , given q is

$$h(x|q) = \sum_{jk} \kappa_{jk} \varphi(x; m_{jk}, s_{jk}),$$

where the mixing probabilities κ_{jk} are proportional to

$$\kappa_{jk} \propto \pi_k \lambda_j \varphi\left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2}\right)$$

and

$$\begin{aligned} m_{jk} &= \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2}(q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2}\mu_k \\ s_{jk}^2 &= \frac{\sigma_k^2 \tau_j^2}{\sigma_k^2 + \tau_j^2} \end{aligned}$$

Proof The random variable x can be expressed as

$$x = \sum_{k=1}^a z_k x_k,$$

where the x_k are independent (latent) random variables, and also independent of (z_1, \dots, z_a) . x_k is normally distributed according to component distribution k , with density $\varphi(x_k; \mu_k, \sigma_k)$. The z_k s are indicator random variables taking values 0 and 1 with probability π_k . One and only one indicator takes the value 1, the others are 0.²⁰ Likewise y can be expressed as

$$y = \sum_{j=1}^b w_j y_j,$$

where the (w_j, y_j) independent of the (z_k, x_k) . Note that $z_k = z_k \sum_j w_j$ and $w_j = w_j \sum_k z_k$, so that

$$q = x + y = \sum_k z_k x_k + \sum_j w_j y_j = \sum_{jk} z_k w_j (x_k + y_j).$$

Part (i) of the Corollary now immediately follows; the mixing probabilities for q are $P[z_k w_j = 1] = \pi_k \lambda_j$ and the component distribution for component (k, j) has expectation $\mu_k + \nu_j$ and variance $\sigma_k^2 + \tau_j^2$.

For the conditional distribution $p(x|q)$ we have

$$\begin{aligned} p(x|q) &\propto \\ p(x, q) &= \sum_{jk} \pi_k \lambda_j \varphi(x; \mu_k, \sigma_k) \varphi(q - x; \nu_j, \tau_j) \\ &= \sum_{jk} \pi_k \lambda_j \varphi\left(x; \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2}(q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2}\mu_k, \frac{\sigma_k \tau_j}{\sqrt{\sigma_k^2 + \tau_j^2}}\right) \\ &\quad \times \varphi\left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2}\right). \end{aligned}$$

The last equality follows from the Lemma.

²⁰In other words, the (z_1, \dots, z_a) have a multinomial distribution with class probabilities π_k and a single draw.

The conditional distribution $p(x|q)$ is therefore again a normal mixture with mixing probabilities for the (k, j) component proportional to

$$\pi_k \lambda_j \varphi \left(q; \mu_k + \nu_j, \sqrt{\sigma_k^2 + \tau_j^2} \right)$$

and corresponding component distribution

$$\varphi \left(x; \frac{\sigma_k^2}{\sigma_k^2 + \tau_j^2} (q - \nu_j) + \frac{\tau_j^2}{\sigma_k^2 + \tau_j^2} \mu_k, \frac{\sigma_k \tau_j}{\sqrt{\sigma_k^2 + \tau_j^2}} \right).$$

The distribution of $p(y|q)$ follows in exactly the same fashion. \square

Appendix 2 Proofs

Lemma 2: The probability density function of u_a conditional on \bar{e}_a , which we denote by $p(u_a|\bar{e}_a)$, is a normal-mixture with known parameters:

$$p(u_a|\bar{e}_a) = \sum_{i=1}^{i=m_u} \sum_{j=1}^{j=m_\varepsilon} \sum_{k=1}^{k=m_\varepsilon} w_{ijk} \varphi(u_a; m_{ijk}; s_{ijk}^2), \quad (20)$$

where:

$$\begin{aligned} m_{ijk} &= \left(\frac{\sigma_i^2}{\sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a} \right) (\bar{e}_a - (\nu_j + \nu_k)/n_a) + \left(\frac{\omega_j^2 + \omega_k^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2} \right) \mu_i \\ s_{ijk}^2 &= \left(\frac{(\omega_j^2 + \omega_k^2) \sigma_i^2}{\omega_j^2 + \omega_k^2 + n_a \sigma_i^2} \right), \end{aligned}$$

and where $w_{ijk} = \tilde{w}_{ijk} / \sum_{ijk} \tilde{w}_{ijk}$ with:

$$\tilde{w}_{ijk} = \pi_i \lambda_j \lambda_k \varphi(\bar{e}_a; \mu_i + (\nu_j + \nu_k)/n_a; \sigma_i^2 + (\omega_j^2 + \omega_k^2)/n_a), \quad (21)$$

where φ denotes the normal probability density function, and where $(\pi_i, \mu_i, \sigma_i^2)$ and $(\lambda_j, \nu_j, \omega_j^2)$ denote the parameters associated with the normal-mixture distributions F_u and G_ε , respectively.

Proof The result stated in this lemma follows from a direct application of Corollary 4. \square

Appendix 3 Household surveys used for random components

3.1 Data sources

The household income and expenditure survey data used to obtain estimates of the area and household level random components across a wide range of countries and years are listed in Table 1. For each of the included countries, the table documents both the specific survey and

years used. These data correspond to the official nationally representative household surveys that the World Bank also uses to estimate poverty levels and consumption growth for the respective countries.

Table 1: List of countries' survey(s) and rounds

Country	Survey(s)	Years/rounds
Bangladesh	Household Income and Expenditure Survey (HIES)	2000, 2005, 2010 and 2016
Brazil	Pesquisa Nacional por Amostra de Domicílios (PNAD)	2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2011, 2012, 2013, 2014, 2015, 2016 and 2017
Chile	Encuesta de Caracterización Socioeconómica Nacional (CASEN)	2000, 2006, 2011, 2013, 2015 and 2017
Colombia	Encuesta Continua de Hogares (ECH); and Gran Encuesta Integrada de Hogares (GEIH)	ECH: 2001, 2002, 2003, 2004 and 2005; GEIH: 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016 and 2017
Ethiopia	Household Income, Consumption and Expenditure Survey (HICES)	2010 and 2015
Ghana	Ghana Living Standards Survey (GLSS)	2005, 2012 and 2016
India	National Sample Survey (NSS)	1993, 2004, 2009 and 2011
Kazakhstan	Household Budget Survey (HBS)	2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016 and 2017
Kenya	Integrated Household Budget Survey (IHBS)	2005 and 2015
Mexico	Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)	2000, 2002, 2004, 2005, 2006, 2008, 2010, 2012, 2014 and 2016
Pakistan	Pakistan Integrated Household Survey (PIHS); and Pakistan Social and Living Standards Measurement Survey (PSLM)	PIHS: 2001; PSLM: 2004, 2005, 2007, 2010, 2011, 2013 and 2015
Philippines	Family Income and Expenditure Survey (FIES)	2006, 2009, 2012 and 2015
Poland	Household Budget Survey (HBS)	2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015 and 2016
Rwanda	Integrated Household Living Conditions Survey (EICV)	2010, 2013 and 2016
South Africa	Income and Expenditure Survey (IES); and Living Conditions Survey (LCS)	IES: 2010; LCS: 2014
Sri Lanka	Household Income and Expenditure Survey (HIES)	2002, 2006, 2009, 2012 and 2016

3.1.1 Harmonization

The survey data are obtained from the Global Monitoring Database (GMD). The GMD is a global comparable micro database hosted by the World Bank that harmonizes household survey data across selected countries and years. The list of harmonized variables includes household income and consumption expenditures, demographics, educational attainment, employment, ownership of durable assets, and housing among others.

Appendix 4 RMSE for headcount rate of 25 percent

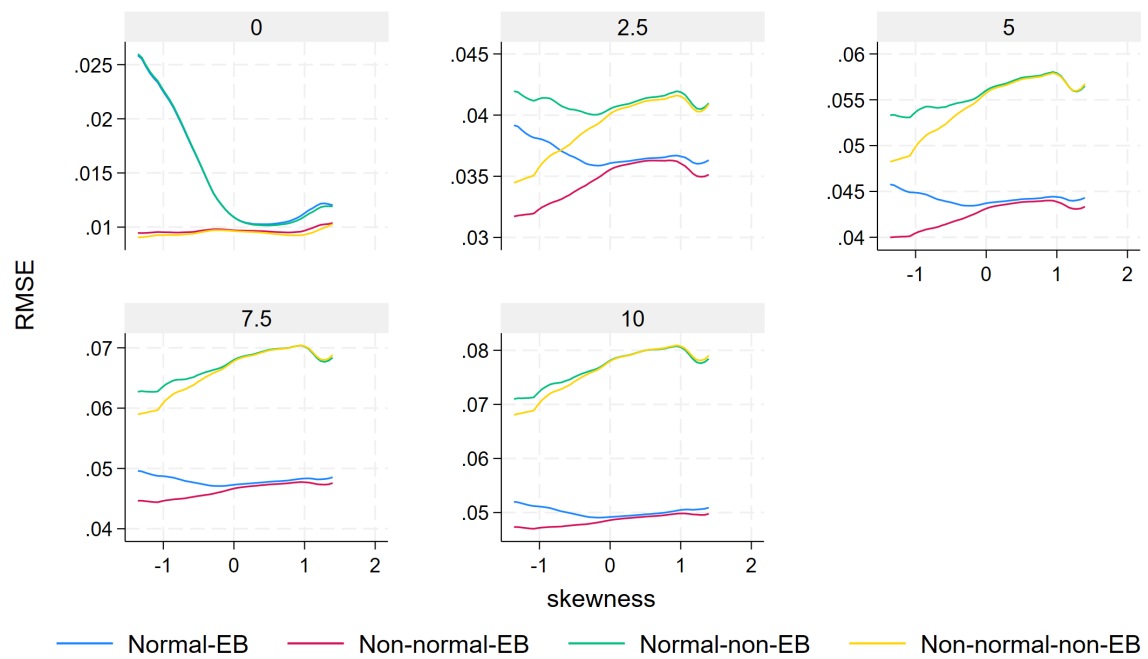


Figure 1: RMSE versus skewness for different magnitudes of the location effect (as a percentage of total error)