Walter, Paul; Groß, Marcus; Schmid, Timo; Tzavidis, Nikos

Domain prediction with grouped income data

Journal of the Royal Statistical Society: Series A (Statistics in Society)

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Walter, Paul; Groß, Marcus; Schmid, Timo; Tzavidis, Nikos (2021) : Domain prediction with grouped income data, Journal of the Royal Statistical Society: Series A (Statistics in Society), ISSN 1467-985X, Wiley, Hoboken, NJ, Vol. 184, Iss. 4, pp. 1501-1523, https://doi.org/10.1111/rssa.12736

This Version is available at:
https://hdl.handle.net/10419/284819

Received: 29 April 2020 | Accepted: 25 June 2021

DOI: 10.1111/rssa.12736

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Domain prediction with grouped income data

Paul Walter[1] | **Marcus Groß**[1] | **Timo Schmid**[2] | **Nikos Tzavidis**[3]

[1]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany

[2]Institute of Statistics, Otto-Friedrich-Universität Bamberg, Bamberg, Germany

[3]Southampton Statistical Sciences Research Institute and Department of Social Statistics & Demography, University of Southampton, Southampton, UK

Correspondence
Timo Schmid, Institute of Statistics, Otto-Friedrich-Universität Bamberg, Bamberg, 96052, Germany.
Email: timo.schmid@uni-bamberg.de

Abstract

One popular small area estimation method for estimating poverty and inequality indicators is the empirical best predictor under the unit-level nested error regression model with a continuous dependent variable. However, parameter estimation is more challenging when the response variable is grouped due to data confidentiality concerns or concerns about survey response burden. The work in this paper proposes methodology that enables fitting a nested error regression model when the dependent variable is grouped. Model parameters are then used for small area prediction of finite population parameters of interest. Model fitting in the case of a grouped response variable is based on the use of a stochastic expectation–maximization algorithm. Since the stochastic expectation–maximization algorithm relies on the Gaussian assumptions of the unit-level error terms, adaptive transformations are incorporated for handling departures from normality. The estimation of the mean squared error of the small area parameters is facilitated by a parametric bootstrap that captures the additional uncertainty due to the grouping mechanism and the possible use of adaptive transformations. The empirical properties of the proposed methodology are assessed by using model-based simulations and its relevance is illustrated by estimating deprivation indicators for municipalities in the Mexican state of Chiapas.

KEYWORDS

data confidentiality, interval-censored data, nested error regression model, small area estimation, survey response burden

1 | INTRODUCTION

Recent applications of small area estimation (SAE) methodologies have been concerned with the estimation of area-specific income indicators, for example the median income, the head count ratio (HCR) and the Gini coefficient (Rao & Molina, 2015; Rojas-Perilla et al., 2020; Tzavidis et al., 2018). Popular SAE methods that have been used in this context include the so-called World Bank method (Elbers et al., 2003) and the empirical best predictor (EBP) method (Molina & Rao, 2010). In these papers, SAE is based on the use of a unit-level nested error regression (random effects) model estimated with income or consumption as a response variable that is measured on a continuous scale.

It is tempting for survey designers to reduce survey related costs by collecting information on income using income bands as opposed to detailed income information (Micklewright & Schnepf, 2010). Collecting data in bands may also help with reducing respondent burden, item non-response and micro-data disclosure risk. On the other hand, it is also reasonable to expect that collecting grouped data may result in a loss of information compared to collecting on a continuous scale. The impact of this loss of information on the quality of official statistics estimates is of particular importance. There are several surveys and censuses that collect grouped income data, for example, the household and land survey (HLS) of Japan (Statistics of Japan, 2013), the German Microcensus (Statistisches Bundesamt, 2018) and the censuses of Australia (Australian Bureau of Statistics, 2011), Colombia (Departamento Administrativo Nacional De Estadística, 2005), and New Zealand (Statistics New Zealand, 2013). In the United Kingdom, the Office for National Statistics experimented with the collection of grouped income data in the lead up to the 2001 census (Collins & White, 1996).

Using statistical methods for grouped data is not a problem specific to SAE. In particular, regression methods for grouped data have been studied in the econometric literature (Hsiao, 1983) but to the best of our knowledge, these methods have not been extended to include random effects. An alternative approach is to view the response as discrete and use generalized linear mixed models. For example, the response can be viewed as a multi-category or an ordinal outcome with cut-off points defined in the latter case by the grouping structure relevant to the dataset of interest. In this case one can motivate the model by assuming the existence of an underlying continuous latent variable, which although different has some similarities to the approach we propose in this paper. The emphasis in this paper is on model-based small area inference more specifically, on estimating not only linear but also non-linear indicators that are functions of the continuous (latent in the case of grouping) response variable. Hence, we propose an extension of the EBP method when the response variable is grouped. The methodology works by reversing the process of grouping, leading to an outcome measured on a continuous scale which is then used for area-specific prediction. Estimation of the parameters of the unit-level nested error regression model is implemented via a stochastic expectation–maximization (SEM) algorithm (Celeux & Diebolt, 1985). The method we propose in this paper is not only applicable in the case of SAE. The SEM algorithm can be used to estimate the model parameters when the dependent variable is grouped and interest is in using the model for drawing substantive conclusions about the relationship between the dependent variables and the explanatory variables (Walter, 2019a). The proposed methodology also allows for the use of data-driven transformations when diagnostic analyses indicate departures from the model assumptions. Using transformations as part of small-area estimation when the continuous outcome is fully available has been already proposed in the literature. In the context of area-level models, there are several papers discussing fixed transformations (e.g. Slud and Maiti (2006)) and data-driven transformations (e.g. Sugasawa and Kubokawa (2017)). Rojas-Perilla et al. (2020) presented theoretical and numerical justifications for the use of data-driven transformations with unit-level SAE models. In particular, they propose an EBP approach with data-driven transformations where the data-driven transformation parameter is estimated by likelihood-based methods.

Following Gonzalez-Manteiga et al. (2008), the estimation of the mean squared error (MSE) of the small area estimates—when the response variable is grouped—is facilitated by a parametric bootstrap.

This incorporates the additional uncertainty due to grouping of the response variable assuming that the censoring mechanism is known. The proposed method assumes that there is no measurement error in reporting the group associated with the latent continuous variable. In this paper, we develop the methodology under a two-level nested error regression model. However, an extension to three-level structures—incorporating possible cluster effects—along the lines of the methodology proposed by Marhuenda et al. (2017) is feasible. Finally, as is the case with the EBP method or the World Bank method, we assume access to micro-data for the model covariates from census or administrative data. The proposed methodology makes the use of SAE methods with grouped outcomes possible and therefore it enables survey organizations to consider collecting data in this form.

The paper is organized as follows. Section 2 presents the survey data we use in this paper and defines the indicators of interest. The EBP approach and the nested error regression model when the response variable is available on a continuous scale are discussed in Section 3. Section 4 introduces the SEM algorithm that is used for the estimation of the model parameters when the response variable is grouped. In Section 5, the EBP method with grouped data is presented. In Section 6, model-based simulations are carried out. In Section 7, the proposed methodology is used to estimate poverty and inequality indicators from grouped income data from Mexico. Finally, the main results are summarized in Section 8.

2 | ESTIMATING SMALL AREA DEPRIVATION INDICATORS FOR MUNICIPALITIES IN THE MEXICAN STATE OF CHIAPAS: DATA SOURCES AND INDICATORS

We start with an initial discussion of the data and the poverty indicators of interest before presenting the methodological details. The aim of the proposed methodology is to enable the estimation of poverty and inequality indicators from survey data with a grouped income variable. To illustrate the proposed approach in this paper we use data from Mexico.

Despite Mexico being the 15th largest economy in the world (International Monetary Fund, 2017), the fight against poverty and inequality is of great importance for the country since high poverty rates are omnipresent. During the Mexican peso crisis, extreme poverty increased from 21% in 1994 to 37% in 1996 (Pereznieto, 2010). Today, poverty rates remain at considerably high levels. According to the World Bank (2010), 33% of the population in the country experienced moderate poverty and 9% extreme poverty in 2013. This demonstrates the relevance of estimating and mapping poverty at local levels such that appropriate interventions can be designed.

The poverty indicators we are interested in include the area HCR, the area poverty gap (PGAP) as defined in Foster et al. (1984) (income deprivation), the area average household income (Mean) and the Gini coefficient (Gini) in each area i. The indicators are defined as follows:

$$\begin{aligned}
\text{HCR}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{I}(y_{ij}\le z),\\
\text{PGAP}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{z-y_{ij}}{z}\right)\mathbf{I}(y_{ij}\le z),\\
\text{Mean}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij},\\
\text{Gini}_i &= \frac{2\sum_{j=1}^{n_i} j y_{ij}}{n_i\sum_{j=1}^{n_i} y_{ij}} - \frac{(n_i+1)}{n_i},
\end{aligned}$$

where y_{ij} denotes the outcome variable, n_i is the sample size, $\mathbf{I}(\cdot)$ denotes the indicator function and z is the poverty threshold. In the simulations and application in this paper, z is set to 60% of the median of income, as defined by (Eurostat, 2014). When the income variable is measured on a continuous scale estimation of these indicators can be facilitated with standard SAE methods. However, when income is only available as grouped variable the methodology we propose in this paper can be used.

For computing the income-defined indicators of interest, one needs to have access to grouped income data that have been equivalized to account for different household sizes. If this has not been done, the secondary analyst will need access to household composition data in order to create equivalized grouped income data. Let us assume household i reports income in [2000,3000] and consists of two adults and one child leading to a weight of 2.5, then the household has an equivalized household income in the interval [800 = 2000/2.5,1200 = 3000/2.5]. This leads to household specific intervals depending on the weight relating to the household composition and reported interval. For the application in this paper we use the 2010 equivalized household income from the ENIGH (Encuesta Nacional de Ingreso y Gasto de los Hogares) survey and a large sample of the 2010 National Population and Housing census in Mexico. Both data sets are collected by the National Institute of Statistics and Geography (INEGI, Instituto Nacional de Estadística y Geografía) and they are provided to us by the National Council for the Evaluation of Social Development Policy (CONEVAL, Consejo Nacional de Evaluación de la Política de Desarrollo Social). Both the census and survey data sets include socio-economic and regional information at household level. While the data cover all 31 states of Mexico, the application focuses on the state of Chiapas. Chiapas is one of the poorest states in Mexico with an average income of about 40% of the national median income (Levy et al., 2016). The state is located in the south of Mexico at the border to Guatemala. The survey covers 42 of the 118 municipalities in Chiapas. Hence, there are 76 out-of-sample municipalities for which no sample data are available. In order to derive reliable estimates at the level of municipality for all 118 municipalities, we rely on the use of model-based methods and auxiliary information from the census and survey data.

The sample size we analyse is n=2486 households and originally CONEVAL asked 3018 households in Chiapas. The response rate was 82%. The census sample size is N=96350 households. The regional distribution of the sample size is given in Table 1. The sample size of the in-sample municipalities varies between 13 and 651 households with a median sample size of 33 households. Since sample sizes are small in many municipalities, SAE methods can potentially improve the accuracy of direct small area estimates.

In the next section a brief introduction to the EBP method when a continuous response variable is available. Then, the newly proposed methodology when a grouped response variable is available is introduced.

3 | EMPIRICAL BEST PREDICTION METHOD

The target of inference are the small area parameters that include linear and non-linear indicators which can be expressed as functions of an income variable, for example average and median equivalized income, the HCR, the PGAP and the Gini coefficient. Since in this paper we assume the availability of

TABLE 1 Distribution of the sample and census household sizes across areas

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample	13.00	17.00	33.00	59.19	51.00	651.00
Census	82.00	399.50	617.50	816.50	839.00	7172.00

unit-level survey and census/administrative data, two methods for estimating non-linear indicators in common use are available, the World Bank method (Elbers et al., 2003) and the popular EBP method (Molina & Rao, 2010). Although our focus is on the use of the EBP method, the proposed methodology can be applied in conjunction with the World Bank method too. The EBP method makes use of unit-level nested error regression model and is summarized below. The response variable is income that is only available in the survey. The explanatory variables used for modelling the income variable are available both in the survey and in the census data sets. After the model is fitted using the survey data, the estimated model parameters are combined with census micro-data to form unit-level synthetic census predictions of the income variable. These synthetic values are then used for estimating the target parameters. Census predictions are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. Although estimation of linear and non-linear indicators can be also implemented with area-level regression models (Fabrizi & Trivisano, 2016; Schmid et al., 2017), we focus on unit-level models which can be used to produce estimates of a wide range of parameters as a by-product of fitting the model. With area-level models the focus is on one target parameter at the time. In addition, approaches to direct estimation with grouped data need to be carefully considered. Possible approaches to doing this are briefly outlined in the concluding remarks.

Consider a finite population U of size N, divided into D areas/domains. The terms areas and domains are used interchangeably in this paper. The population size of each of the D-domains $U_1, U_2, \ldots, U_D$ is given by $N_1, N_2, \ldots, N_D$. Let us for now assume that the response variable denoted by y_{ij} is measured on a continuous scale, where $j = 1, 2, \ldots, n_i$ denotes the jth unit belonging to the ith domain, with $i = 1, 2, \ldots, D$. The vector x is defined as $x_{ij}^T = (x_{1ij}, \ldots, x_{pij})$, where p denotes the number of explanatory variables. For each area i, the sample size is n_i with $n = \sum_{i=1}^{D} n_i$ and the population vector y_i for area i comprises sampled and non-sampled units $y_i^T = (y_{is}^T, y_{ir}^T)$. A nested error linear regression model is used for modelling the relationship between the variable of interest and auxiliary information with the unexplained variation being captured by the random effects term, u_i and the residuals e_{ij}. In the simplest case, a two-level nested error regression model as defined in Battese et al. (1988) is given by

$$\begin{aligned} &y_{ij} = x_{ij}^T\beta + u_i + e_{ij}, \quad (j = 1, \ldots, n_i), \quad (i = 1, \ldots, D), \\ &u_i \overset{iid}{\sim} N(0, \sigma_u^2), \\ &e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2). \end{aligned} \tag{1}$$

Assuming normality for the unit-level error terms and the domain random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. Predictions for the entire population of area i are generated from the following model,

$$\begin{aligned} &y_{ij}^* = x_{ij}^T\beta + \hat{u}_i + u_i^* + e_{ij}^*, \\ &u_i^* \overset{iid}{\sim} N(0, \sigma_u^2 \times (1 - \gamma_i)), \quad e_{ij}^* \overset{iid}{\sim} N(0, \sigma_e^2), \quad \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_i}}, \end{aligned} \tag{2}$$

where $\hat{u}_i = E(u_i \mid y_{is})$ is the conditional expectation of u_i given the sample data y_{is}. Implementation of (2) requires replacing the unknown quantities $\beta, \sigma_u, \sigma_e$, with estimates and simulating L synthetic populations of the income variable, y_{ij}^*. Linear and non-linear indicators are computed in each domain i for each replication and the estimates are averaged over the number of Monte Carlo simulations L. Following Molina and Rao (2010) and Rojas-Perilla et al. (2020) this number is usually set equal to $L = 50$ or $L = 100$ but higher numbers are also possible.

For the estimation of the unknown quantities $\beta, \sigma_u, \sigma_e$ when the response variable is grouped we propose the use of a SEM algorithm.

4 | THE NESTED ERROR LINEAR REGRESSION MODEL WITH A GROUPED RESPONSE VARIABLE

In the case of grouped data, y_{ij} is unobserved and the only observed information concerning the dependent variable is whether it falls within an interval. The continuous scale is divided into K intervals, where the k-th interval is given by (A_{k-1}, A_k). The variable $k_{ij} \in \{1, \ldots, K\}$ indicates in which of the intervals the dependent variable falls into. The first and K-th interval are allowed to be open ended, therefore $A_0 = -\infty$ and $A_K = +\infty$ are possible. Situations in which both or none of the outer intervals are open ended can also be handled by the proposed methodology. Furthermore, the interval length is allowed to be arbitrary and can vary between intervals. Since the underlying distribution of y_{ij} is unknown, the aim is to reconstruct the conditional distribution $f(y_{ij} \mid x_{ij}, k_{ij}, u_i, \theta)$, where $\theta = (\beta, \sigma_e^2, \sigma_u^2)$ are the unknown model parameters, β is a $p \times 1$ vector of regression coefficients and the random effects u_i and the unit-level error terms e_{ij} are assumed to be independent and normally distributed. Estimation methods such as maximum likelihood (ML) or restricted maximum likelihood (REML) are used for estimating θ when y_{ij} is observed on a continuous scale (Lindstrom & Bates, 1990). However, when the response variable is grouped, estimation of the parameters of interest is more challenging. The likelihood, $\prod_i \prod_j f(k_{ij} \mid x_{ij}, u_i, \theta)$, cannot be derived directly, but can be expanded to include the latent y_{ij} into $\prod_i \prod_j f(k_{ij} \mid y_{ij}, x_{ij}, u_i, \theta) \times f(y_{ij} \mid x_{ij}, u_i, \theta)$. While the second part is well known and can be maximized by the aforementioned methods, the first part, $f(k_{ij} \mid y_{ij}, x_{ij}, u_i, \theta)$, demands a more sophisticated estimation procedure as the latent part y_{ij} needs to be integrated out. In this section, an SEM algorithm for fitting the model is proposed and data-driven transformations are also considered for handling potential departures from the model assumptions. Before presenting the model and estimation method in detail, we review alternative approaches to dealing with grouped response variables and compare the SEM algorithm to alternative fitting methods.

Different approaches for dealing with grouped response variables in regression modelling that assume independent observations have been proposed in the literature. A naive approach uses ordinary least squares on the midpoints of the intervals. While this approach is easy to implement (Thompson & Nelson, 2003), it has two major drawbacks. The uncertainty associated with the value of each observation within each interval is not accounted for and dealing with open-ended intervals is not easy. Nevertheless, the naive approach can provide results of acceptable quality if the grouping is very fine (Fryer & Pethybridge, 1972). An alternative approach is to view the response as discrete and use a generalized linear mixed model. Approaches to modelling multicategory discrete outcomes have been proposed in the small area literature (Lopez-Vizcaino et al., 2015; Molina et al., 2007). Nevertheless, one difficulty with the use of discrete-type models remains. In our application we are not only interested in estimating the proportion of units in a category but also interested in estimating indicators such as the PGAP, the HCR and the Gini coefficient. Therefore, if we decide to use a discrete-type model we also need to develop a method for recovering estimates of target parameters that are usually computed by using a continuous outcome. To overcome these drawbacks, linear regression models for left-censored (Tobin, 1958), right-censored (Rosett & Nelson, 1975) and grouped (or interval-censored) (Stewart, 1983) variables have been proposed. Stewart (1983) proposes an expectation–maximization (EM) algorithm for estimating the model parameters of a linear regression model with a grouped response variable. While the original paper introducing the EM algorithm (Dempster et al., 1977) proposed maximizing the likelihood within the M-step, it mentioned that the EM algorithm can

be also used to obtain REML estimates. Literature related to this includes Kim and Taylor (1995) and Foulley et al. (2000).

To estimate the parameters of the nested error regression model when the outcome is grouped, we propose the use of a SEM algorithm (Celeux & Diebolt, 1985; Celeux et al., 1996). A similar SEM algorithm is proposed in Groß et al. (2017) for kernel density estimation on aggregated data. SEM can be regarded as a middle ground between the EM and full MCMC. With the EM algorithm we alternate between calculating the expectation of the conditional distribution $f(y_{ij}\,|\,x_{ij}, k_{ij}, u_i, \theta)$ (E-Step) and obtaining θ via maximizing the complete data likelihood (M-Step). However, with fixed intervals (A_{k-1},A_k) it can be seen that a bias in θ will be introduced, for example, σ_e^2 would be underestimated as the overall variance of the expectations of y_{ij} is much smaller than that of the (latent) variable y_{ij}. SEM and full MCMC (Gelman et al., 2013) replace the E-Step by drawing from the conditional distribution of y_{ij} and therefore do not suffer from this drawback. This approach can be also viewed as part of the literature about measurement error models (Carroll et al., 2006), where the latent values y_{ij} are regarded as model parameters or partially observed data (Carpenter et al., 2012). In addition, MCMC also replaces the M-Step by sampling from the conditional distribution of θ. In summary, compared to EM, SEM avoids or reduces biases in the estimation of the parameters of interest, while compared to MCMC, SEM is considerably faster due to faster convergence because only the values y_{ij} are drawn. Using the SEM also saves time with the implementation because the users can make use of existing estimation algorithms for the M-Step, while it is easy to make draws of y_{ij}. Considering the assessment of convergence, SEM should be treated similarly to MCMC with its broad variety of convergence measures. Related to the SEM approach is also the simulated maximum likelihood (SML, Gouriéroux and Monfort, 1990) method. The SML also samples y_{ij} values but uses these samples to estimate the expectation of the density $f(k_{ij}\,|\,y_{ij}, x_{ij}, u_i, \theta) \times f(y_{ij}\,|\,x_{ij}, u_i, \theta)$ which is then maximized with respect to θ. However, SML is not unbiased, but it is consistent (Gouriéroux & Monfort, 1990), and not as straightforward to implement as SEM as one needs to deal with the numerical aspects of the optimization procedure.

Let us now consider the model we use in this paper. To reconstruct the unknown distribution $f(y_{ij}\,|\,x_{ij}, k_{ij}, u_i, \theta)$ we use the Bayes theorem and express the target distribution as follows:

$$f(y_{ij}\,|\,x_{ij}, k_{ij}, u_i, \theta) \propto f(k_{ij}\,|\,y_{ij}, x_{ij}, u_i, \theta) f(y_{ij}\,|\,x_{ij}, u_i, \theta).$$

To avoid confusion, note that in the notation we use here we treat σ_u^2 as part of θ. However, when writing the distribution of the random effects we make it explicit that this distribution depends on σ_u^2. Since $f(k_{ij}\,|\,y_{ij}, x_{ij}, u_i, \theta) = f(k_{ij}\,|\,y_{ij})$, the conditional distribution of k_{ij} is given by

$$f(k_{ij}\,|\,y_{ij}) = \begin{cases} 1 \text{ if } A_{k_{ij}-1} \leq y_{ij} \leq A_{k_{ij}}, \\ 0 \text{ else}, \end{cases}$$

and under the nested error regression model (1),

$$f(y_{ij}\,|\,x_{ij}, u_i, \theta) \sim N(x_{ij}^T\beta + u_i, \sigma_e^2), \quad f(u_i\,|\,\sigma_u^2) \sim N(0, \sigma_u^2).$$

4.1 | The SEM algorithm

Because y_{ij} and u_i are unobserved, one approach to fitting the model defined above is to use the SEM algorithm. Generally speaking, the algorithm works by replacing the unobserved response data y_{ij} in the complete data likelihood by generating pseudo samples of the unobserved response data given the

observed data and the current values of θ (S-step) and then maximizes the complete data likelihood for updating θ and the predicted random effects in the M-step. The iterations stop after $B + M$ steps.

Assuming θ is known, pseudo samples, $\tilde{y}_{ij}$, are drawn from the following conditional distribution

$$f(y_{ij} \mid x_{ij}, k_{ij}, u_i, \theta) \propto \mathbf{I}(A_{k_{ij}-1} \leq y_{ij} \leq A_{k_{ij}}) \times N(x_{ij}^T\beta + u_i, \sigma_e^2),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The conditional distribution of y_{ij} has the form of a two-sided truncated normal distribution given by

$$f(y_{ij} \mid x_{ij}, k_{ij}, u_i, \theta) = \frac{\phi\left(\frac{y_{ij}-\mu_{ij}}{\sigma_e}\right)}{\sigma_e\left(\Phi\left(\frac{A_{k_{ij}}-\mu_{ij}}{\sigma_e}\right) - \Phi\left(\frac{A_{k_{ij}-1}-\mu_{ij}}{\sigma_e}\right)\right)},$$

with $\mu_{ij} = x_{ij}^T\beta + u_i$, $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. By definition $\Phi\left(\frac{A_{k_{ij}}-\mu_{ij}}{\sigma_e}\right) = 1$ if $A_{k_{ij}} = +\infty$ and $\Phi\left(\frac{A_{k_{ij}-1}-\mu_{ij}}{\sigma_e}\right) = 0$ if $A_{k_{ij}-1} = -\infty$. For each observation with explanatory variables x_{ij} the corresponding $\tilde{y}_{ij}$ is randomly drawn from $N(x_{ij}^T\beta + u_i, \sigma_e^2)$ within the given interval $A_{k_{ij}-1} \leq y_{ij} \leq A_{k_{ij}}$. This is the S-step of the SEM algorithm. The M-step comprises fitting the nested error regression model using the newly generated $(\tilde{y}_{ij}, x_{ij})$. The steps of the SEM algorithm are as follows:

1. Estimate $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}_e^2, \widehat{\sigma}_u^2)$ and u from (1) using the midpoints of the intervals as a substitute for the unknown y_{ij}. The parameters are estimated using REML and using the empirical best linear unbiased predictor (EBLUP) for u.
2. **S-step:** For $j = 1, \ldots, n_i$ and i=1, …, D sample from the conditional distribution $f(y_{ij} \mid x_{ij}, k_{ij}, u_i, \theta)$ by drawing randomly from $N(x_{ij}^T\widehat{\beta} + \widehat{u}_i, \widehat{\sigma}_e^2)$ within the given interval $A_{k_{ij}-1} \leq y_{ij} \leq A_{k_{ij}}$ obtaining $(\tilde{y}_{ij}, x_{ij})$. The drawn pseudo $\tilde{y}_{ij}$ are used as replacement for the unknown y_{ij}.
3. **M-step:** Re-estimate the model parameters and the predicted random effects using (1) and the pseudo samples $(\tilde{y}_{ij}, x_{ij})$ from Step 2. The parameters are estimated as in Step 1.
4. Iterate Steps 2–3 $B + M$ times, with B burn-in iterations and M additional iterations.
5. Discard the burn-in iterations and estimate $\widehat{\theta}$ by averaging the derived M estimates.

For open-ended intervals $A_0 = -\infty$ and $A_K = +\infty$, the midpoints M_1 and M_K in Step 1 are computed as follows:

$$M_1 = (A_1 - \overline{D})/2,$$
$$M_K = (A_{K-1} + \overline{D})/2,$$

where

$$\overline{D} = \frac{1}{(K-2)} \sum_{k=2}^{K-1} |A_{k-1} - A_k|.$$

Note that Step 1 is purely for the initialization of the algorithm. Empirical results show that using the midpoints of the intervals as a substitute for the unknown y_{ij} and the procedure for handling open-ended intervals in the first iteration step has little impact on the estimates. Empirical results are provided in table 9 in the online supplementary material (OSM).

The proposed SEM algorithm makes repeated use of a two-sided truncated normal distribution, by drawing from $N(x_{ij}^T\widehat{\beta} + \widehat{u}_i, \widehat{\sigma}_e^2)$ within the given interval $A_{k_{ij}-1} \leq y_{ij} \leq A_{k_{ij}}$. Therefore, the performance of the SEM algorithm relies on the Gaussian assumptions of the unit-level error terms being met. To accommodate possible departures from the model assumptions, the proposed SEM algorithm is extended to allow for the use of transformations.

4.2 | The SEM algorithm under transformations

Transformations of the outcome can be used in case of departures from the model assumptions. Broadly speaking, one can use non-adaptive or adaptive transformations. For the application in this paper that models income-type data, the logarithmic transformation is probably the one most commonly used. While the logarithmic transformation is easy to use, there is no guarantee that it will provide the best transformation for the target distribution. This is crucial in this paper since the validity of the normality assumption of the unit-level error terms cannot be tested due to the fact that the response variable is grouped. Therefore, using adaptive (data-driven) transformations instead of fixed transformations, is preferable. In addition, the logarithmic transformation can be obtained as a special case of a family of adaptive transformations. In this paper, we focus on the use of the Box-Cox transformation (Box & Cox, 1964; Draper & Cox, 1969) and its extension under the nested error regression model (Gurka et al., 2006). The Box-Cox transformation is given by

$$y_{ij}(\lambda) = \begin{cases} \dfrac{(y_{ij}+s)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s is a fixed shift parameter that ensures that $y_{ij}+s > 0$. The Box-Cox transformation depends on the transformation parameter λ that is used for transforming the data $T_\lambda(y_{ij}) = y_{ij}(\lambda)$. The aim is to find the value of λ given the data such that the assumptions about the unit-level error terms of the nested error regression model are met (Gurka et al., 2006). The implementation of data-driven transformations within the SEM algorithm is computationally intensive because the transformation parameter λ has to be estimated in each iteration step. The algorithm is structured into two parts. In Part 1 the SEM algorithm is used for finding the optimal transformation parameter, $\widehat{\lambda}^{(F)}$. In Part 2 the SEM algorithm is implemented with the estimated $\widehat{\lambda}^{(F)}$ from Part 1. The detailed steps of the SEM algorithm under transformations are given below.

Part 1

1. Define a grid g of possible values of λ. Using each value in the grid, implement the steps below.
2. Use the scaled version of the Box-Cox transformation, as defined in Rojas-Perilla et al. (2020), to transform the midpoints of each interval (A_{k-1},A_k) and fit the nested error regression model (1). Repeat the same process for each value of λ in g and select the value of $\widehat{\lambda}$ that maximizes the restricted maximum likelihood. Note that the use of the scaled version of the Box-Cox transformation, defined by

$$\frac{y_{ij}(\lambda)}{J(\lambda,y)^{\frac{1}{n}}} = y_{ij}(\lambda)\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}^{\lambda-1}\right)^{-1/n},$$

where J denotes the Jacobian, is important for estimating the transformation parameter λ. The Jacobian of the scaled Box-Cox transformation is equal to 1. This means that the scale of the likelihood is preserved independently of the transformation parameter λ in g. Thus, values of the log-likelihood function—under differently transformed $y_{ij}(\lambda)$—can be compared and the log-likelihood function simplifies to the log-likelihood function of the nested error regression model (1). For further details we refer to Rojas-Perilla et al. (2020).

3. Using the selected value of $\widehat{\lambda}$ from the previous step, fit the nested error regression model (1) to obtain $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}_e^2, \widehat{\sigma}_u^2)$ and $\widehat{u}$.
4. Generate a new pseudo sample as a proxy for the unobserved $y_{ij}(\widehat{\lambda})$. To do this, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, D$ sample from the conditional distribution $f(y_{ij}(\lambda) \mid x_{ij}, k_{ij}, u_i)$ by drawing from $N(x_{ij}^T\widehat{\beta} + \widehat{u}_i, \widehat{\sigma}_e^2)$ within the given interval $(A_{k_{ij}-1}(\widehat{\lambda}) \leq y_{ij}(\widehat{\lambda}) \leq A_{k_{ij}}(\widehat{\lambda}))$ to obtain $(\tilde{y}_{ij}(\widehat{\lambda}), x_{ij})$. Back-transform $\tilde{y}_{ij}(\widehat{\lambda})$ to the original scale $\tilde{y}_{ij}$ using the selected $\widehat{\lambda}$ from Step 2.
5. Go to Step 2 and select a new optimal $\widehat{\lambda}$ this time using the newly generated $\tilde{y}_{ij}$ from the previous step in Step 2 of the algorithm instead of the interval midpoints.
6. Iterate Steps 2-5 $B + M$ times, with B burn-in iterations and M additional iterations.
7. Discard the burn-in iterations and estimate the final $\widehat{\lambda}^{(F)}$ by averaging the M estimates of $\widehat{\lambda}$.

Part 2

8. Use $\widehat{\lambda}^{(F)}$ from Part 1 and the Box-Cox transformation to transform the midpoints and the interval bounds of each interval (A_{k-1}, A_k). Then apply the SEM algorithm as described in Section 4.1 in steps 1–5 with B burn in and M additional iterations to estimate $\widehat{\theta}$.

Figure 1 illustrates why in the case of using transformations it is important to structure the SEM algorithm in two parts, that is, finding the optimal λ first and then using the optimal λ, to estimate β. The left panel of Figure 1 plots the estimated λ for each iteration step of the algorithm for monitoring its convergence. The right panel of Figure 1 plots $\widehat{\lambda}$ against $\widehat{\beta}$ for each iteration step of Part 1. From that plot it is clear that by simply running Part 1 and averaging the M estimates of $\widehat{\beta}$ and $\widehat{\lambda}$ the averaged parameter estimates would not be the same as the parameter estimates obtained by using the value of the transformation parameter, λ, at convergence. This is the case because the relationship between $\widehat{\lambda}$ and $\widehat{\beta}$ is non-linear. In addition, as λ is different in every iteration, different iterations would be fitting models for differently defined response variables. Therefore, the SEM algorithm is divided into two parts. In Part 1 the final $\widehat{\lambda}^{(F)}$ is estimated and in Part 2 this estimate is used for estimating the parameters of the nested error regression model on the transformed scale.

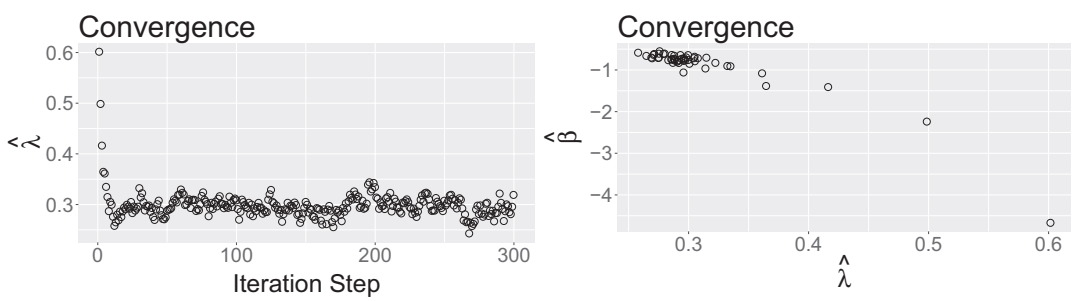

FIGURE 1 Convergence of $\widehat{\lambda}$ and $\widehat{\beta}$ using the stochastic expectation–maximization algorithm

5 | EMPIRICAL BEST PREDICTION WITH GROUPED DATA

In the presence of a grouped income variable, the EBP approach needs to be modified. In the first step, the model parameters, $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2)$, and the predicted random effects are estimated using the SEM algorithm. Note that predicted random effects are computed by using the estimated model parameters and the values of the pseudo response generated by the SEM algorithm. It is likely that when modelling an income variable the normality assumptions of the nested error regression model may not hold. In this case, a suitable transformation is needed and the SEM algorithm is implemented to estimate $\widehat{\theta}$ and $\widehat{\lambda}^{(F)}$ using the results in Section 4.2 and following the developments by Rojas-Perilla et al. (2020).

Having estimated $\widehat{\theta}$, $\widehat{\lambda}^{(F)}$ and $\widehat{u}_i$, the remaining steps of the Monte Carlo algorithm used to implement the empirical best predictor (EBP) are as follows:

1. Use the sample data and the SEM algorithm to estimate $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2)$, $\widehat{\lambda}^{(F)}$ and $\widehat{\gamma}_i = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \frac{\widehat{\sigma}_e^2}{n_i}}$.
2. For $l = 1, \ldots, L$:
   (a). Generate a synthetic population under the nested error regression model $\widehat{y}_{ij}^{*(l)}(\widehat{\lambda}^{(F)}) = x_{ij}^T\widehat{\beta} + \widehat{u}_i + u_i^{*(l)} + e_{ij}^{*(l)}$, where x_{ij} are population micro-data for unit j in area i, $u_i^{*(l)} \overset{iid}{\sim} N(0, \widehat{\sigma}_u^2(1-\widehat{\gamma}_i))$, $e_{ij}^{*(l)} \overset{iid}{\sim} N(0, \widehat{\sigma}_e^2)$ and $\widehat{u}_i$ is given by $\widehat{u}_i = E(u_i \mid y_{is})$.
   (b). Back-transform to the original scale $\widehat{y}_{ij}^{*(l)} = T^{-1}\left(\widehat{y}_{ij}^{*(l)}(\widehat{\lambda}^{(F)})\right)$.
   (c). In each area, estimate the target parameter $\widehat{I}_i^{(l)}$ using $\widehat{y}_{ij}^{*(l)}$.
3. The target parameter is estimated by averaging over the L Monte Carlo estimates $\widehat{I}_i^{(l)}$ in each area,

$$\widehat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L}\widehat{I}_i^{(l)}.$$

If the SEM algorithm is used without a transformation; $\widehat{\lambda}^{(F)}$ in Step 1 and Step 2 (a) as well as the whole Step 2 (b) (the back-transformation step) can be neglected and T is the identity function. For non-sampled areas, we cannot estimate an area random effect, hence $\widehat{u}_i$ is not available. In this case, Step 2(a) above is modified such that synthetic values of the outcome are generated as follows, $\widehat{y}_{ij}^{*(l)}(\widehat{\lambda}^{(F)}) = x_{ij}^T\widehat{\beta} + u_i^{*(l)} + e_{ij}^{*(l)}$, where the random effects are drawn from $u_i^{*(l)} \overset{iid}{\sim} N(0, \widehat{\sigma}_u^2)$ and the unit-level error terms are drawn from $e_{ij}^{*(l)} \overset{iid}{\sim} N(0, \widehat{\sigma}_e^2)$. The same applies to the case where we are working with the untransformed response variable.

Mean squared error estimation is a crucial step in SAE. Complications arise due to the complexity of non-linear indicators which make the development of analytic MSE estimators difficult. For the EBP, Molina & Rao, (2010) propose a parametric bootstrap MSE estimator under the nested error regression model. The use of bootstrap under the EBP approach with data-driven transformations has been discussed by Rojas-Perilla et al. (2020). The authors propose an approach to accounting for the additional uncertainty due to the estimation of the transformation parameter. A parametric bootstrap is also used when working with a grouped outcome. However, there are two additional sources of variability we need to account for. One is the uncertainty due to the estimation of the transformation parameter and the second is the uncertainty resulting from working with limited information due to grouping. The bootstrap MSE assumes that the mechanism used to group the response variable is known. Denoting by k the bootstrap iteration, the bootstrap MSE estimator below is presented in the more general case where a transformation of the response variable is used.

1. (a). Using the sample estimates, $\widehat{\beta}, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2, \widehat{\lambda}^{(F)}$, at convergence of the SEM algorithm, generate $u_i^{*(k)} \overset{iid}{\sim} N(0, \widehat{\sigma}_u^2)$ and $e_{ij}^{*(k)} \overset{iid}{\sim} N(0, \widehat{\sigma}_e^2)$ and a bootstrap superpopulation $\widehat{y}_{ij}^{*(k)}(\widehat{\lambda}^{(F)}) = x_{ij}^T \widehat{\beta} + u_i^{*(k)} + e_{ij}^{*(k)}$, where x_{ij} are population micro-data for unit j in area i.
   (b). Back-transform $\widehat{y}_{ij}^{*(k)} = T^{-1}\left(\widehat{y}_{ij}^{*y(k)}(\widehat{\lambda}^{(F)})\right)$ to the original scale and compute the population value of the target parameter in area i and bootstrap iterations k, $I_{i,k}$.
   (c). Select a bootstrap sample using a simple random sampling with replacement from each area that respects the area-specific sample sizes of the original sample.
   (d). Using the known censoring mechanism and the bootstrap sample data, create the grouped response variable.
   (e). Use the SEM algorithm with the current bootstrap sample for deriving EBP estimates of the target parameters. In this case where a transformation is used this consists of using Parts 1 and 2 from Section 4.2 and the EBP algorithm under a transformation described in Section 5.
   (f). Obtain EBP estimates of the target parameter in area i and bootstrap iteration k, $\widehat{I}_{i,k}^{EBP}$.
2. Using a total of K bootstrap samples, the MSE estimator is computed as follows:

$$\widehat{\mathrm{MSE}}\left(\widehat{I}_i^{EBP}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{I}_{i,k}^{EBP} - I_{i,k}\right)^2.$$

6 | MODEL-BASED SIMULATIONS

This section presents model-based simulation results for assessing the performance of the proposed methodology for estimating poverty and inequality indicators introduced in Section 2. In particular, we assess the performance of point estimators and of corresponding MSE estimators. In order to evaluate the properties of estimators of the model parameters obtained with the proposed methodology we have conducted additional simulation studies that are presented in Section 2 in the OSM.

Three population models (Normal, Log-scale and Pareto)—in line with the scenarios considered by Rojas-Perilla et al. (2020)—are used for generating the simulated data. Details about the data generation mechanisms and the corresponding conditional R_c^2 (Nakagawa & Schielzeth, 2013) are outlined in tables 1 and 2 in the OSM. The normal scenario (in Section 6.1) is used for evaluating the performance of the EBP approach under grouping of the response variable when the model assumptions are met. The log-scale scenario (in Section 6.2) attempts to mimic the distribution of an income variable we might work with in practice. In addition, we also assess the properties of the proposed bootstrap MSE estimator. The Pareto scenario attempts to mimic an observed income distribution and illustrates the performance of the SEM Box-Cox algorithm under model misspecification. The results from this simulation study are available in the OSM.

For the normality-based scenario, we use two different grouping mechanisms, referred to as normal scenario 1 (with 14 income groups) and normal scenario 2 (with 7 income groups) (see tables 4 and 5 of the OSM). This allows us to explore the impact of the number of groups on the performance of the small area estimators which is of interest for survey practitioners.

In each Monte Carlo run, a finite population U of size $N = 10000$ is generated and is partitioned into $D = 50$ areas each with size $N_i = 200$. From the finite population we select a sample using an unbalanced design with area-specific sample sizes n_i ranging between $8 \leq n_i \leq 29$. The total sample size is $n = 921$. In total we run 200 Monte Carlo iterations with the number of Monte Carlo iterations for implementing the EBP set to $L = 200$ and the number of bootstrap iterations for MSE estimation set to $K = 200$.

We compare the EBP under the model that assumes that the continuous response variable is available (abbreviated below by LME) to the EBP when only the grouped variable is available and the use

of the SEM algorithm is necessary (abbreviated below by SEM). For both model-based scenarios we further compare the standard EBP when a Box-Cox transformation is used (LME Box-Cox) to the EBP–SEM approach when a Box-Cox transformation is used (SEM Box-Cox). This allows us to examine how well the parameter of the Box-Cox transformation, λ, is estimated when we only have access to the grouped response. For assessing the use of a fixed transformation, the standard EBP as well as the EBP with grouped data is used with a logarithmic transformation (LME Log, SEM Log). The SEM algorithm uses 40 burn-in iterations and 200 additional iterations. Note that the estimators above are available in the packages *emdi* (Kreutzmann et al., 2019) and *smicd* (Walter, 2019b) in R.

The performance of point estimates is assessed by computing the area-specific empirical root mean squared error $RMSE\left(\hat{I}_i^{EBP}\right) = \left[\frac{1}{200}\sum_{m=1}^{200}\left(\hat{I}_i^{EBP(m)} - I_i^{(m)}\right)^2\right]^{1/2}$, where m denotes the Monte Carlo iteration, $\hat{I}_i^{EBP}$ is the estimated indicator using one of the above-mentioned methods and I_i is the true population value. Tables are used to report the mean and median over areas of the RMSE. The proposed MSE estimators are evaluated by the relative bias and the relative RMSE for each area i. We treat the empirical root MSE as the true MSE.

6.1 | Results: Normality-based scenarios

Table 2 presents a summary of the results for normal scenario 1 (14 intervals) and normal scenario 2 (7 intervals) using the SEM method, the SEM Box-Cox method, the LME and LME Box-Cox methods. In figures 1 and 2 in the OSM the estimated density of the population y_{ij} values is plotted against the estimated densities of $\hat{y}_{ij}^{*(l)}$ using the different estimation methods from one arbitrarily chosen simulation run. For normal scenario 2, we also have included the regression on the midpoints method (denoted by MID) as a *naive* competitor. The results show that the performance of the EBPs using the SEM algorithm a) outperforms the estimates obtained using midpoint regression and b) is close to the performance of the EBPs when the continuous outcome is fully available. As expected, when using the fully available continuous outcome the EBP estimates are more efficient (lower RMSE) than the SEM-based estimates. However, despite working with the grouped outcome, the increase in RMSE (reduction in efficiency) is not dramatic which demonstrates that the SEM algorithm works well. In line with the theory, the results also show that as the number of classes used to discretize the continuous outcome reduces (from 14 to 7 groups), the RMSE of the SEM-based estimates increases. This is reasonable as in this case the information available is reduced. Nevertheless, even in the case of scenario 2 we would argue that the performance of the SEM-based estimates is reasonable. Our view is based on the fact that seven groups present a rather extreme scenario in real applications.

The performance of the estimates using the SEM and SEM Box-Cox methods is very similar. In the case of the normal-based scenarios, this is expected since the data-driven transformation parameter, λ, is estimated to be close to one, which is equivalent to using no transformation. This is confirmed by looking at the estimation of λ in table 3 in the OSM. Hence, the structure of the SEM algorithm in two parts works as expected and the Box-Cox transformation adapts well to the shape of the data distribution, even though only the grouped information is used for estimating λ.

The MSE results for the different indicators are summarized in Table 3. Overall, the relative bias and relative RMSE of the estimated RMSE are low. In particular, for most scenarios and target parameters the relative bias is below 10% and for a few scenarios somewhat above 10%. The relative RMSE also shows that the bootstrap estimator is stable. In the OSM, we also present domain-specific coverage rate plots of the confidence intervals for the different indicators. In particular, figures 4-6

TABLE 2 Performance of the estimated empirical best predictors (EBPs) in terms of RMSE over areas

	Mean		HCR		PGAP		Gini	
Indicator:	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Normal scenario 1 (14 intervals)								
RMSE								
LME	201.482	212.450	0.033	0.035	0.014	0.015	0.013	0.014
LME Box-Cox	201.675	212.466	0.033	0.035	0.014	0.016	0.013	0.014
SEM	203.783	217.075	0.034	0.036	0.014	0.016	0.013	0.014
SEM Box-Cox	204.604	217.335	0.034	0.036	0.014	0.017	0.013	0.015
Normal scenario 2 (7 intervals)								
RMSE								
MID	258.199	264.024	0.045	0.045	0.027	0.028	0.018	0.019
LME	200.725	212.405	0.033	0.035	0.014	0.015	0.013	0.014
LME Box-Cox	201.422	212.502	0.033	0.035	0.014	0.016	0.013	0.014
SEM	216.780	225.692	0.035	0.038	0.015	0.017	0.014	0.015
SEM Box-Cox	215.324	225.897	0.035	0.037	0.016	0.018	0.014	0.016
Log-scale scenario (7 intervals)								
RMSE								
LME Log	994.586	988.374	0.063	0.065	0.039	0.040	0.035	0.034
LME Box-Cox	995.068	992.021	0.063	0.065	0.040	0.040	0.035	0.034
SEM Log	1046.724	1030.190	0.066	0.068	0.041	0.042	0.035	0.035
SEM Box-Cox	1043.407	1040.646	0.066	0.068	0.040	0.042	0.037	0.037

(in the OSM) show the coverage rates of 95% confidence intervals constructed by using the estimated bootstrap MSEs. We observe that the coverage rates of the EBP estimates using the SEM Box-Cox method are close to that of the EBP when the continuous outcome is fully available.

6.2 | Results: Log-scale scenario

In this section, we present results when the assumptions of the nested error regression model are not met. This is the case for the log-scale scenario. For this scenario, the response variable is grouped in seven intervals, hence a fairly extreme censoring mechanism is evaluated. The distribution of the response variable using one arbitrarily chosen Monte Carlo population can be seen in table 6 of the OSM. The results in Table 2 show that the performance of the estimates using the SEM Box-Cox and the SEM Log methods is close to the performance of the estimates using the LME Box-Cox and to the LME Log methods that assume that the continuous outcome variable is available. As expected, some accuracy in estimation is compromised when working with the grouped outcome. However, the SEM-based estimates remain competitive when compared to the estimates obtained by assuming that full information for the response variable is available. This is also confirmed by looking at how the SEM-based methods recover the true population density in figure 3 in the OSM.

The use of the Box-Cox transformation appears to work well. Under this scenario, the transformation parameter λ should be estimated to be close to zero. This is confirmed by examining the

TABLE 3 Performance of the bootstrap root mean squared error (MSE) estimator over areas

	Mean		HCR		PGAP		Gini	
Indicator:	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Normal scenario 1 (14 intervals)								
rel.Bias[%]								
SEM	7.37	7.05	5.91	5.07	2.31	3.07	3.90	3.88
SEM Box-Cox	7.56	7.33	5.61	5.47	−6.86	−5.58	−3.67	−4.41
rel.RMSE[%]								
SEM	9.50	10.50	10.59	11.38	12.05	13.34	8.68	9.87
SEM Box-Cox	9.92	10.85	10.78	11.42	13.03	14.01	8.81	10.53
Normal scenario 2 (7 intervals)								
rel.Bias[%]								
SEM	5.30	5.84	4.71	3.65	−0.18	0.30	2.24	1.94
SEM Box-Cox	5.46	6.10	4.59	3.91	−15.77	−14.82	−10.22	−11.67
rel.RMSE[%]								
SEM	8.59	9.91	10.22	10.98	12.22	13.29	8.84	9.59
SEM Box-Cox	8.82	10.30	10.30	11.07	19.27	19.16	12.09	14.79
Log-scale scenario (7 intervals)								
rel.Bias[%]								
SEM Log	7.22	6.56	6.73	7.58	6.74	7.13	0.88	0.77
SEM Box-Cox	13.17	26.00	6.78	7.65	7.10	7.57	6.54	6.40
rel.RMSE[%]								
SEM Log	33.49	34.78	13.19	14.25	21.02	21.63	7.95	8.36
SEM Box-Cox	42.19	60.85	13.23	14.36	21.33	21.93	16.49	17.05

estimation results of λ in table 3 in the OSM. Finally, Table 3 shows that the proposed bootstrap MSE estimator has reasonably low relative bias. As expected, the MSE under the Box-Cox version of the SEM is somewhat more unstable than the corresponding MSE for the Log SEM. This is due to the fact that in the case of the Box-Cox method the transformation parameter is estimated with each bootstrap sample whereas for the Log method the transformation is held fixed.

In order to further evaluate the impact of the grouping on the performance of the SEM estimators, we report as part of the OSM two additional Log-scale scenarios. For the first one we used 14 equally spaced intervals leading to a large proportion of observations in the upper open-ended interval. For the second scenario we increased the interval size with increasing y values. The results are reported in table 12 in the OSM.

7 | ESTIMATING SMALL AREA DEPRIVATION INDICATORS FOR MUNICIPALITIES IN THE MEXICAN STATE OF CHIAPAS: AN APPLICATION OF THE SEM BOX-COX METHOD

In our application the response variable, equivalized household income, is measured on a continuous scale. In order to assess the performance of the proposed methodology, we group equivalized

TABLE 4 Variables used in the nested error regression working model

Variable type	Description
Response variable:	Grouped equivalized household labour income
Auxiliary variables:	Value of all household goods
	Value of household communication equipment
	Share of employees in the household
	Educational level of head of household
	Social class of head of household
	Municipalities of Chiapas

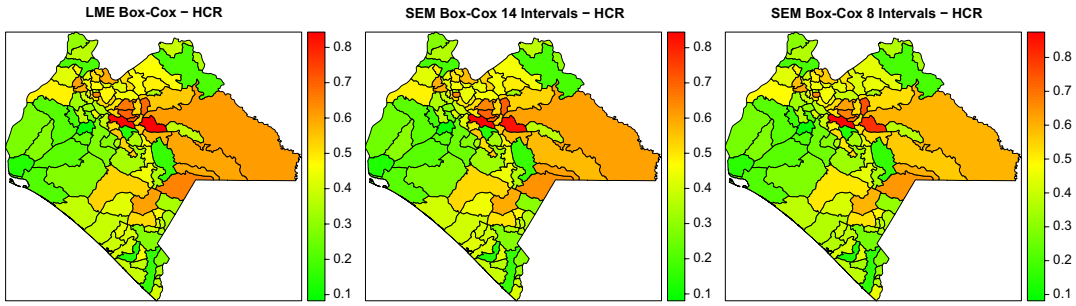

FIGURE 2 Estimated head count ratio (HCR) for municipalities in the state of Chiapas based on different estimation methods. The empirical best predictor (EBP) under the model with continuous response and Box-Cox transformation is abbreviated by LME Box-Cox, the EBP under the model with grouped data and Box-Cox transformation is abbreviated by SEM Box-Cox [Colour figure can be viewed at wileyonlinelibrary.com]

household income to 14 and 8 intervals. The distribution of the grouped equivalized household income is presented in tables 16 and 17 of the OSM. The variables in Table 4 were identified as possible covariates that predict equivalized household labour income well. The variables in the working model are selected by using the coefficient of determination proposed by Nakagawa and Schielzeth (2013). The conditional R_c^2, interpreted as the variance explained by the whole model, is $R_{c,lme}^2 = 0.61$ when estimating the model with the observed continuous response variable on the transformed scale (Box-Cox transformation). When estimating the model with a grouped response variable on the transformed scale (Box-Cox transformation) using the SEM algorithm the $R_{c,sem(14)}^2$ is 0.61 and $R_{c,sem(8)}^2$ is 0.62 for the 14 and 8 interval scenario respectively.

The Box-Cox transformation is used as the preferred transformation method because it is data-driven. This is crucial when working with grouped data as response variable, because the normality assumption of the residuals cannot be checked. The estimated transformation parameters are $\widehat{\lambda}_{lme} = 0.16$ for the continuous response, $\widehat{\lambda}_{sem(14)}^{(F)} = 0.18$ and $\widehat{\lambda}_{sem(8)}^{(F)} = 0.17$ for the 14 and 8 grouping scenarios respectively. The results indicate that the use of a logarithmic transformation or the use of the untransformed response variable may lead to erroneous results. Rojas-Perilla et al. (2020) and Tzavidis et al. (2018) show that even if λ is estimated to be close to 0 the EBP estimates using the Box-Cox transformation may outperform the EBP estimates using the logarithmic transformation.

Estimates of the mean equivalized household labour income, HCR and PGAP for each of the 118 municipalities are obtained by using the SEM Box-Cox method based on 14 and 8 intervals, and by using LME Box-Cox based on the observed continuous response variable. The mean and median

TABLE 5 Point estimates and corresponding relative efficiencies of the point estimates (EFF) over municipalities using the stochastic expectation–maximization (SEM) Box-Cox algorithm

	Mean		HCR		PGAP		Gini	
Box-Cox	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Point est. LME	814.4	872.6	0.426	0.432	0.220	0.233	0.535	0.539
Point est. SEM 14 intervals	812.1	870.8	0.426	0.427	0.224	0.233	0.531	0.536
EFF	0.983	0.995	1.018	1.024	1.022	1.035	1.057	1.064
Point est. SEM 8 intervals	810.5	863.9	0.421	0.428	0.219	0.232	0.529	0.534
EFF	1.028	1.013	1.029	1.038	1.043	1.046	1.071	1.064

The empirical best predictor (EBP) under the model with continuous response and Box-Cox transformation is abbreviated by LME Box-Cox, the EBP under the model with grouped data and Box-Cox transformation is abbreviated by SEM Box-Cox. HCR stands for the head count ratio and PGAP for the poverty gap

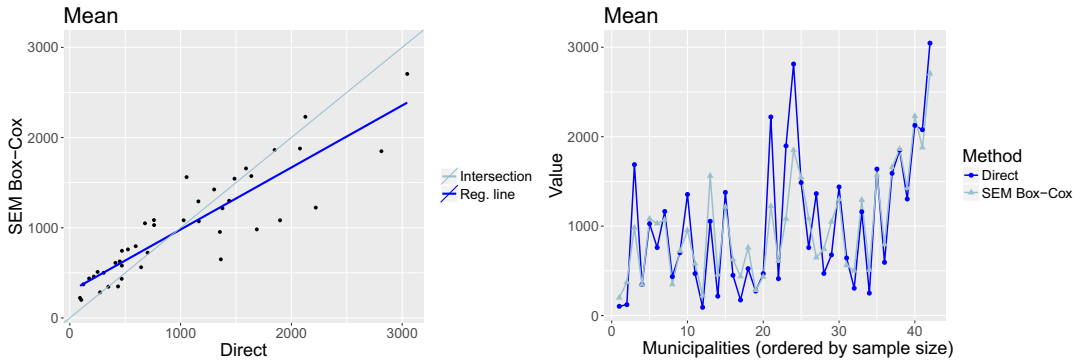

FIGURE 3 The left panel shows a scatter plot and the right panel a line plot of the direct and the model-based estimates for each in-sample domain (municipality). The empirical best predictor under the model with grouped data and Box-Cox transformation is abbreviated by SEM Box-Cox [Colour figure can be viewed at wileyonlinelibrary.com]

averaged over all municipalities are given in Table 5 and plotted in Figures 2 and 4. The results show that the point estimates from all three estimation methods are very similar. Interval censoring does not appear to impact significantly on the estimation results. Additionally, the relative efficiencies of the estimators (EFF) defined as $EFF(\widehat{I}_i^{EBP}) = RMSE_{sem}(\widehat{I}_i^{EBP})/RMSE_{lme}(\widehat{I}_i^{EBP})$ is reported in the Table 5. It is notable that the efficiency loss is small even when the response variable is grouped to only eight intervals. In the 14 interval scenario the point estimates of the mean are even more efficient, but this result is only due to the Monte Carlo variability. The spatial distributions of the HCR in municipalities in Chiapas are shown in Figure 2 for all three estimation methods. The figure supports the previously mentioned results that the estimates obtained by using the different methods are very close.

A possible way to further validate the estimation results is by comparing the direct estimates, where available, to the model-based estimates. In Figure 3 the direct estimates of the mean (based on the observed continuous data) are compared to the model-based estimates (SEM Box-Cox) of the mean using a grouped response variable with 14 intervals. As expected, the left panel shows a positive linear correlation between the estimates. However, there is a disparity between the intersection line (the identity) and the regression line. As anticipated, the model-based estimates are less extreme compared to the direct estimates

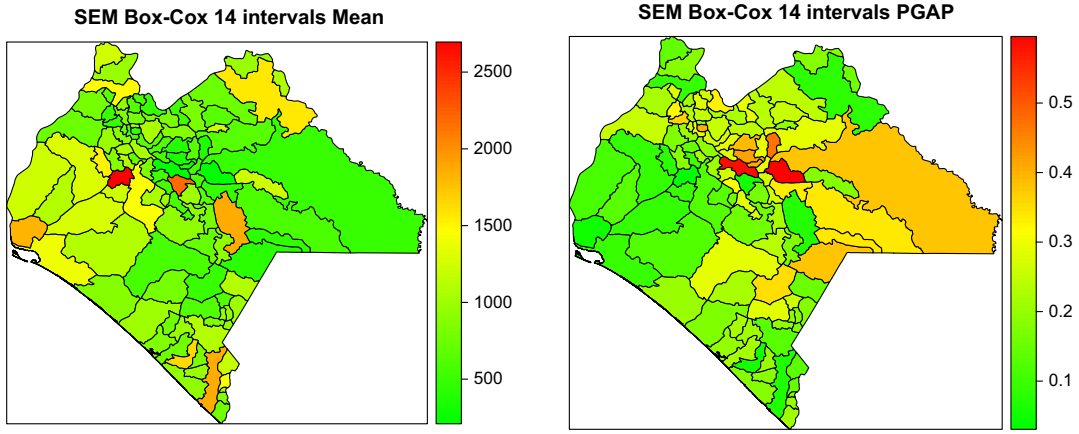

FIGURE 4 Estimated mean and poverty gap for municipalities in the state of Chiapas [Colour figure can be viewed at wileyonlinelibrary.com]

for municipalities with very small and very high mean estimates. The right panel plots the value of the estimates for both estimation methods for each in-sample domain. The pattern shows that as the sample size increases the direct estimates and the SEM Box-Cox estimates are almost identical. Figure 4 presents municipal estimates of mean income and PGAP for the SEM Box-Cox algorithm based on 14 intervals. The plots for the other estimation methods are omitted because the results are comparable. We observe that municipalities in the middle and in the east of Chiapas exhibit high rates of HCRs and PGAPs and low levels of mean equivalized household labour income and are thus more adversely affected by poverty. These regions are characterized by high mountains, the Chiapas Highlands and a large concentration of indigenous population. There are, however, two regions in the centre of the state with relatively high mean income and low rates of poverty. These are the regions where the capital Tuxtla Gutiérrez and the larger city San Cristóbal de las Casas are located. Also the coastal region—especially in the south—where the most important city economically Tapachula is located, is less affected by poverty. The analysis shows that even though Chiapas is one of the poorest states in Mexico, there are spatial variations between the municipalities. These differences can be revealed by using SAE methods designed for grouped data. The proposed SEM Box-Cox method is, to the best of our knowledge, the first approach that allows the use of the popular EBP method in conjunction with a grouped response variable. This enables the estimation of spatially disaggregated target indicators with small sample sizes when confidentiality restrictions or decisions about the survey design require the use of relatively limited information for the response variable.

8 | CONCLUDING REMARKS

The paper proposes SAE methodology when working with a response variable that is grouped. The novel aspects of the paper include the estimation of a nested error regression model when the response is grouped, the estimation both of linear and non-linear indicators for small areas, the use of data-driven transformation with the nested error regression model and the estimation of the MSE of the small area target parameters that accounts for the fact that we are working with limited information compared to standard small area models.

The proposed methods are evaluated using model-based simulations under different scenarios for the unit-level error terms. The results show that the proposed methods work well and in most scenarios the loss of accuracy is small when compared to the use of EBPs that are estimated by assuming the availability of full information for the response variable. As expected, the loss of accuracy also depends on the number of intervals used for grouping the data and the proposed methodology appears to work well even when the number of groups used is fairly small. The results also show that the use of an adaptive transformation works satisfactorily and the transformation parameter is estimated well in the presence of limited information for the response variable. Finally, the proposed MSE estimator appears to capture the different sources of variability and appropriately tracks the empirical MSE.

The new methodology is used to estimate disaggregated poverty and inequality indicators for municipalities in Chiapas, a southern state of Mexico, using grouped income banded in 8 and 14 intervals. In order to evaluate the proposed methodology estimates of the target parameters are also obtained when income is fully available, that is, not grouped. The Box-Cox transformation is applied to ensure that the model assumptions are met. The estimates from the continuous and grouped responses are very close, indicating the validity of the proposed methodology. The plotted poverty maps enable policy makers to get a spatial overview of the distribution of poverty in Chiapas and to target poorer regions more precisely.

The proposed methodology for estimating non-linear indicators with grouped response data requires access to unit-level census micro-data for the covariates. Access to such data may be very

challenging due to confidentiality constraints. Although the proposed methodology assumes access to unit-level census micro-data, it is important to discuss briefly an alternative when such data are not available. An alternative approach would be to use area-level models which are based on direct estimates of the linear or non-linear indicator of interest. Methods for direct estimation with grouped data can be mainly categorized in three groups: (1) Direct estimation based on the midpoints of the intervals, (2) parametric (Chen, 2017; Reed & Wu, 2008) and (3) non-parametric (Kakwani & Podder, 2008) modelling of the distribution function. How these different direct estimation methods can be combined with area-level models is an open area for further research.

The proposed model-based small area methodology does not incorporate survey weights in the estimation. Conventionally, SAE methods are model based and in most cases the survey weights are not used in model fitting. However, not including the survey weights carries risks. One example is when the assumption of a non-informative sample selection mechanism does not hold, even after conditioning on auxiliary variables, hence wrongly assuming that the model for the sample also holds for the population. An approach to accounting for the survey weights in EBP was recently proposed by Guadarrama et al. (2018). Although not implemented in this paper, the pseudo EBP approach can be adapted to the setting of the present paper. Doing so requires fitting the nested error regression model and estimating the fixed effects and the variance components in each step of the SEM algorithm by using the methods in You and Rao (2002).

The issue of missing data in SAE has received some attention in the literature. Similarly to the use of survey weights, most small area literature assumes a missing at random mechanism hence after conditioning on covariates, the probability to respond is assumed not to depend on the response. This is the assumption we are making in the present paper. Provided that the survey weights adjust for non-response, one approach to account for non-response is by incorporating the weights in model fitting as proposed by Guadarrama et al. (2018). In a recent paper Sverchkov and Pfeffermann (2018) proposed an alternative approach to modelling non-missing at random mechanisms in SAE. However, to the best of our knowledge this has not been extended yet for estimating the more general parameters that are of interest in this paper.

Current research focuses on extending the SEM method for fitting nested error regression models for more complex structures, for example models with random coefficients. In future research, we also plan to focus on the case where grouping also affects some of the auxiliary variables. This is a more challenging problem but perhaps more realistic if interest is in protecting data confidentiality. Finally, there are three further aspects that we do not discuss in this paper and are left open for future research. First, the proposed methodology does not adjust for the effects of heaping. This can be resolved by following the methods proposed by Groß and Rendtel (2016). Second, we also acknowledge that another type of measurement error may exist when respondents report their income in the wrong interval. However, this is a type of misclassification error that cannot be solved unless we are willing to impose additional assumptions or use results from a validation sample, which can be treated as a gold standard. Third, the proposed methodology assumes normality for the random effects. The assumption may be relaxed by leaving the distribution of the random effects unspecified and use non-parametric methods (Marino et al., 2019). This leads to a discrete mixture distribution which avoids the need to impose parametric assumptions like normality for the random effects. However, the extension to the case of grouped response variables is a topic for further research.

ACKNOWLEDGEMENTS

The research has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 730998, InGRID-2, Integrating Research Infrastructure

for European expertise on Inclusive Growth from data to policy. Furthermore, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council and funding under the Data and Evidence to End Extreme Poverty (DEEP) research programme. DEEP is a consortium of the Universities of Cornell, Copenhagen, and Southampton led by Oxford Policy Management, in partnership with the World Bank's Development Data Group and funded by the UK Foreign, Commonwealth & Development Office. The authors are grateful to the National Council for the Evaluation of Social Development Policy (CONEVAL, Consejo Nacional de Evaluación de la Política de Desarrollo Social) for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods. Finally, the authors are indebted to the Joint Editor, Associate Editor and two referees for comments that significantly improved the paper.

ORCID


Timo Schmid http://orcid.org/0000-0002-7217-2501

REFERENCES

Australian Bureau of Statistics (2011) Census household form. Available from https://unstats.un.org/unsd/demographic/sources/census/quest/AUS2011en.pdf. Accessed data 2018-04-05.

Battese, G.E., Harter, R.M. & Fuller, W.A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401), 28–36.

Box, G.E.P. & Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2), 211–252.

Carpenter, J.R., Goldstein, H. & Kenward, G.M. (2012) *Statistical modelling of partially observed data using multiple imputation: principles and practice*, Netherlands, Dordrecht: Springer, 15–31.

Carroll, R.J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006) *Measurement error in nonlinear models: A modern perspective*. Boca Raton: CRC Press.

Celeux, G. & Diebolt, J. (1985) The sem algorithm: A probalistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly,* 2:73–82.

Celeux, G., Chauveau, D. & Diebolt, J. (1996) Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4), 287–314.

Chen, Y.T. (2017) A unified approach to estimating and testing income distributions with grouped data. *Journal of Business & Economic Statistics*, 36(3), 1–18.

Collins, D. & White, A. (1996) In search of an income question for the 2001 census. *Survey Methodology Bulletin*, 39(7), 2–10.

Dempster, A., Laird, N. & Rubin, D. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–38.

Departamento Administrativo Nacional De Estadística (2005) Censo general 2005. Available from https://www.dane.gov.co/files/censos/libroCenso2005nacional.pdf?&. Accessed date 2018-04-05.

Draper, N.R. & Cox, D.R. (1969) On distributions and their transformation to normality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(3), 472–476.

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

Eurostat (2014) Statistics explained: At-risk-of-poverty rate. Available from http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:At-risk-of-poverty_rate. Accessed date 2018-05-30.

Fabrizi, E. & Trivisano, C. (2016) Small area estimation of the gini concentration coefficient. *Computational Statistics & Data Analysis,* 99:223–234.

Foster, J., Greer, J. & Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, 52(3), 761–766.

Foulley, J.-L., Jaffrézic, F. & Robert-Granié, C. (2000) Em-reml estimation of covariance parameters in gaussian mixed models for longitudinal data analysis. *Genetics Selection Evolution*, 32(2), 129.

Fryer, J.G. & Pethybridge, R.J. (1972) Maximum likelihood estimation of a linear regression function with grouped data. *Journal of the Royal Statistical Society: Series C*, 21(2), 142–154.

Gelman, A., Carlin, J. B., Stern, H.S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013) *Bayesian data analysis*. Boca Raton: CRC Press.

Gonzalez-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. & Santamaria, L. (2008) Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics & Data Analysis*, 52 (12), 5242–5252.

Gouriéroux, C. & Monfort, A. (1990) Simulation based inference in models with heterogeneity. *Annals of Economics and Statistics*, 20–21, 69–107.

Groß, M. & Rendtel, U. (2016) Kernel density estimation for heaped data. *Journal of Survey Statistics and Methodology*, 4(3), 339–361.

Groß, M., Rendtel, U., Schmid, T., Schmon, S. & Tzavidis, N. (2017) Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. *Journal of the Royal Statistical Society: Series A*, 180(1), 161–183.

Guadarrama, M., Molina, I. & Rao, J. (2018) Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121, 20–40.

Gurka, M.J., Edwards, L.J., Muller, K.E. & Kupper, L. (2006) Extending the box-cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A*, 169(2), 273–288.

Hsiao, C. (1983) *Studies in econometrics, time series and multivariate statistics, chapter regression analysis with a categorized explanatory variable*, Cambridge, Massachusetts: Academic Press, 93–129.

International Monetary Fund (2017) World economic outlook database. Available from http://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx. Accessed date 2017-10-14.

Kakwani, N.C. & Podder, N. (2008) Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations Lorenz curve and associated inequality measures from grouped observations. In: Chotikapanich, D. (Ed.) *Modeling income distributions and lorenz curves*. New York: Springer, pp. 57–70.

Kim, D.K. & Taylor, J.M. (1995) The restricted em algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association*, 90(430), 708–716.

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

Levy, D., Hausmann, R., Santos, M. A., Espinoza, L. & Flores, M. (2016) Why is chiapas poor? Center for International Development at Harvard University Working Paper, (300)

Lindstrom, M. J. & Bates, D. M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3), 673–687.

Lopez-Vizcaino, E., Lombardia, M. J. & Morales, D. (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society: Series A*, 178(3), 535–565.

Marhuenda, Y., Molina, I., Morales, D. & Rao, J. N. K. (2017) Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, 180(4), 1111–1136.

Marino, M. F., Ranalli, M. G., Salvati, N. & Alfò, M. (2019) Semiparametric empirical best prediction for small area estimation of unemployment indicators. *Annals of Applied Statistics*, 13(2), 1166–1197.

Micklewright, J. & Schnepf, S. (2010) How reliable are income data collected with a single question? *Journal of the Royal Statistical Society: Series A,* 173(2), 409–429.

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

Molina, I., Saei, A. & Jose Lombardia, M. (2007) Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A*, 170(4), 975–1000.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133–142.

Pereznieto, P. (2010) The case of mexico's 1995 peso crisis and argentina's 2002 convertibility crisis: Including children in policy responses to previous economic crises. *UNICEF: Social and economic policy*.

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

Reed, W.J. & Wu, F. (2008) New four- and five-parameter models for income distributions. In: Chotikapanich, D. (Ed.) *Modeling income distributions and lorenz curves*. New York: Springer, pp. 211–224.

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

Rosett, R.N. & Nelson, F.D. (1975) Estimation of the two-limit probit regression model. *Econometrica*, 43(1), 141–146.

Schmid, T., Bruckschen, F., Salvati, N. & Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A*, 180(4), 1163–1190.

Slud, E. & Maiti, T. (2006) Mean-squared error estimation in transformed fay-herriot models. *Journal of the Royal Statistical Society: Series B*, 68(2), 239–257.

Statistics New Zealand (2013) New Zealand census of population and dwellings. Available from https://unstats.un.org/unsd/demographic/sources/census/quest/NZL2013enIn.pdf. Accessed date 2018-05-13.

Statistics of Japan (2013) Survey results of the housing and land survey. Available from https://www.stat.go.jp/english/data/jyutaku/results.html. Accessed date 2021-01-13.

Statistisches Bundesamt (2018) Der Mikrozensus stellt sich vor. Available from https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Mikrozensus.html. Accessed date 2020-12-20.

Stewart, M. (1983) On least square estimation when the dependent varaible is grouped. *The Review of Economic Studies*, 50(4), 737–753.

Sugasawa, S. & Kubokawa, T. (2017) Transforming response values in small area prediction. *Computational Statistics and Data Analysis,* 114:47–60.

Sverchkov, M. & Pfeffermann, D. (2018) Small area estimation under informative sampling and not missing at random non-response. *Journal of the Royal Statistical Society: Series A*, 181(4), 981–1008.

Thompson, M.L. & Nelson, K. (2003) Linear regression with type i interval- and leftcensored response data. *Environmental and Ecological Statistics*, 10(2), 221–230.

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. & 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(4), 927–979.

Walter, P. (2019a) A selection of statistical methods for interval-censored data with applications to the german microcensus. Available from https://refubium.fu-berlin.de/bitstream/handle/fub188/23841/Dissertation_Paul_Walter.pdf?sequence=3&isAllowed=y.

Walter, P. (2019b) smicd: Statistical Methods for Interval Censored Data. R package version 1.0.3.

World Bank (2010) Poverty & equity data portal. Available from http://povertydata.worldbank.org/poverty/country/MEX/. Accessed date 2017-10-14.

You, Y. & Rao, J.N.K. (2002) A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 30(3), 431–439.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Walter P, Groß M, Schmid T, Tzavidis N. Domain prediction with grouped income data. *J R Stat Soc Series A*. 2021;184:1501–1523. https://doi.org/10.1111/rssa.12736